ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Εθνικόν και Καποδιστριακόν
Πανεπιστήμιον Αθηνών
ΙΔΡΥΘΕΝ ΤΟ 1837

ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Εθνικό και Καποδιστριακό
Πανεπιστήμιο Αθηνών

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΒΙΟΛΟΓΙΑΣ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
«ΒΙΟΠΛΗΡΟΦΟΡΙΚΗ-ΥΠΟΛΟΓΙΣΤΙΚΗ ΒΙΟΛΟΓΙΑ»

Δ Ι Π Λ Ω Μ Α Τ Ι Κ Η Ε Ρ Γ Α Σ Ι Α

«Biological Data Mining and its Applications in Healthcare»

Φανή Τσιτούρα

Πτυχιούχος Τμήματος Βιολογίας, Εθνικού και Καποδιστριακού Πανεπιστημίου Αθηνών (ΕΚΠΑ)

ΑΘΗΝΑ 2021

HELLENIC REPUBLIC
National and Kapodistrian
University of Athens
EST. 1837

HELLENIC REPUBLIC
National and Kapodistrian
University of Athens

SCHOOL OF SCIENCE
DEPARTMENT OF BIOLOGY

MASTER IN
«BIOINFORMATICS-COMPUTATIONAL BIOLOGY»

Master Diploma Thesis

«Biological Data Mining and its Applications in Healthcare»

FANI TSITOURA

DEGREE OF BIOLOGY
NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

A T H E N S 2 0 2 1

ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Εθνικόν και Καποδιστριακόν
Πανεπιστήμιον Αθηνών
ΙΔΡΥΘΕΝ ΤΟ 1837

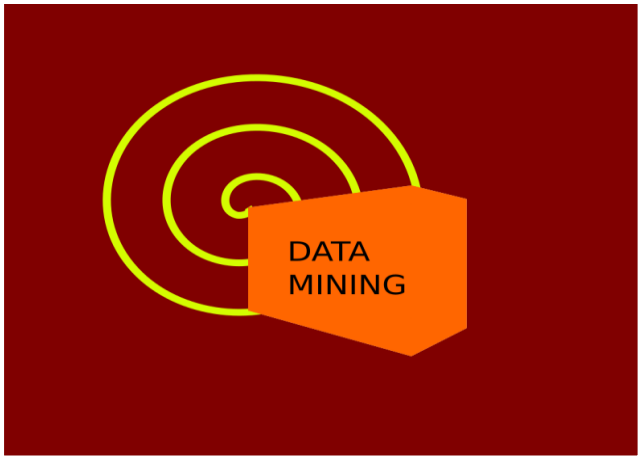
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Εθνικό και Καποδιστριακό
Πανεπιστήμιο Αθηνών

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΒΙΟΛΟΓΙΑΣ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
«ΒΙΟΠΛΗΡΟΦΟΡΙΚΗ-ΥΠΟΛΟΓΙΣΤΙΚΗ ΒΙΟΛΟΓΙΑ»

Δ Ι Π Λ Ω Μ Α Τ Ι Κ Η Ε Ρ Γ Α Σ Ι Α

«Εξόρυξη Βιολογικών Δεδομένων και οι Εφαρμογές της στην Υγειονομική Περίθαλψη»



Τριμελής εξεταστική επιτροπή

Αναπληρωτής Καθηγητής Μιχαήλ Φιλιππάκης (Επιβλέπων)
Τμήμα Ψηφιακών Συστημάτων, Πανεπιστήμιο Πειραιώς

Καθηγητής Ιωάννης Τρουγκάκος
Τομέας Βιολογίας Κυττάρου και Βιοφυσικής,
Τμήμα Βιολογίας, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών

Αναπληρώτρια Καθηγήτρια Βασιλική Οικονομίδου
Τομέας Βιολογίας Κυττάρου και Βιοφυσικής,
Τμήμα Βιολογίας, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών

Πίνακας Περιεχομένων

1. Περίληψη

*Η **εξόρυξη δεδομένων (data mining)** αφορά την χρήση πληροφοριακού συστήματος βασισμένου σε υπολογιστή (Computer-Based Information System, CBIS) με νέες τεχνικές, για εξαγωγή γνώσεων από δεδομένα* (Vlahos et al., 2004). Συνδυάζει στατιστική, μηχανική μάθηση και τεχνητή νοημοσύνη για να φέρει σε πέρας αναλύσεις των δεδομένων. Όμως τα **βιολογικά / κλινικά δεδομένα** που είναι συνήθως άμεσα διαθέσιμα δεν πληρούν τις προϋποθέσεις για καλές αναλύσεις.
Στην εργασία αυτή πραγματοποιήθηκε εξόρυξη δεδομένων (*data mining*) σε Βιολογικό Σύνολο Δεδομένων για ασθενείς με **Ηπατοκυτταρικό καρκίνωμα (*HCC-Hepatocellular Carcinoma*)**, που ήταν διαθέσιμο στο ψηφιακό αποθετήριο δεδομένων UCI. Η συλλογή του συνόλου δεδομένων, είχε γίνει με βάση τις οδηγίες κλινικής πράξης του Ευρωπαϊκού Συνδέσμου για την μελέτη του Ήπατος – Ευρωπαϊκό Οργανισμό για Έρευνα και Θεραπεία του Καρκίνου (European Association for the Study of the Liver - European Organisation for Research and Treatment of Cancer, EASL-EORTC). Το συγκεκριμένο σύνολο δεδομένων επιλέγχθηκε λόγω του ότι ήταν **δωρεάν**, **άμεσα διαθέσιμο** και **συνδυάζει**, αρκετές **κακές ποιότητες που συναντώνται σε σύνολα βιολογικών / κλινικών δεδομένων** (**μη ισορροπημένη εκπροσώπηση όλων των τύπων ασθενών που εξετάζονται** και **ελλειπούσες τιμές**). Είναι ένα **καλό παράδειγμα ρεαλιστικού συνόλου**, πάνω στο οποίο αξιολογήθηκε η αποτελεσματικότητα διαφόρων μεθόδων εξόρυξης δεδομένων, ώστε τα ευρήματα να έχουν **γενικότερο αντίκτυπο** και πάνω σε αντίστοιχης, κακής ποιότητας δεδομένων, αναλύσεις (και για **άλλες ασθένειες**).
Επιπλέον της εξόρυξης δεδομένων έγινε προετοιμασία του συνόλου δεδομένων (*data preparation*), γνωστή και ως **προεπεξεργασία** ή καθαρισμός (*preprocessing / cleaning*), κάτι που υπάγεται στο υπερσύνολο της Εξόρυξης Δεδομένων, την διαδικασία **Ανακάλυψης Γνώσεων από Βάσεις Δεδομένων (*KDD-Knowledge discovery from databases*)**. Στο στάδιο αυτό, δοκιμάστηκε και μία **νέα μέθοδος συμπλήρωσης κενών τιμών**.
Αναζητήθηκαν οι **στρατηγικές προεπεξεργασίας των δεδομένων και μέθοδοι ταξινόμησης που κατηγοριοποιούν καλύτερα τους ασθενείς με βάση την μεταβλητή-στόχο** (για την οποία έχουμε πρότερη γνώση), την **επιβίωση στο 1 έτος**. Οι ανωτέρω πορείες **αξιολογήθηκαν με χρήση περισσότερων και ποικιλέστερων μετρικών αξιολόγησης της ταξινόμησης** από ό,τι συμβαίνει συνήθως στην βιβλιογραφία, ώστε να έχουμε **πληρέστερη και σφαιρικότερη** αξιολόγηση. Μετά την επεξεργασία των αποτελεσμάτων διερευνήθηκαν τα είδη **εφαρμογών** που προέκυψαν και αφορούν την **Υγειονομική Περίθαλψη**, τα εξής:

- **κατηγοριοποίηση των ασθενών με Ηπατοκυτταρικό καρκίνωμα (HCC)** στην καθημερινή κλινική πράξη (*για επιβίωση στο 1 έτος*), με βάση τα κλινικά ή βιολογικά χαρακτηριστικά που εξετάζονται από το σύνολο δεδομένων που αναλύσαμε,
  και συνακόλουθη
- **διευκόλυνση λήψης κλινικών αποφάσεων / ρίσκων** εξαρτώμενων από τον κίνδυνο που έχουν οι ασθενείς από την ασθένεια (*μη υπερθεραπεία με επικίδυνες αγωγές σε ασθενείς χαμηλού κινδύνου κατάληξης στο 1 έτος, μη υποθεραπεία σε ασθενείς υψηλού κινδύνου*)
- τα **διαγράμματα αξιολόγησης** των τακτικών που ακολουθήθηκαν, είναι διαθέσιμα ώστε **επιστήμονες της υγείας να μπορούν να διαλέξουν την τακτική προεπεξεργασίας-ταξινόμησης** που κρίνουν αποτελεσματικότερη για **σκοπούς δικών τους αναλύσεων** (*εάν επιθυμούν να χρησιμοποιήσουν κάποια από όσες δοκιμάστηκαν στην παρούσα εργασία*).

1. Abstract

Data mining *is the process implemented in a computer-based information system (CBIB) with new techniques in order to discover knowledge from data* (Vlahos et al., 2004). Data mining combines statistics, machine learning and artificial intelligence during data analysis. However, the **biological / clinical data** that are usually readily available do not qualify for good data mining results. We performed data mining on a Biological dataset of Patients diagnosed with **Hepatocellular Carcinoma (HCC)**, that was available at the UCI data mining repository. The collection of the data set was based on the clinical practice guidelines of the European Association for the Study of the Liver - European Organisation for Research and Treatment of Cancer, EASL-EORTC. The data set was selected since it was **free**, **readily available** and combined several **poor qualities found in biological / clinical data sets** (**unbalanced representation of all patient types examined** and **missing values**). It serves as a **good example** of a **real biological / clinical set**, upon which the effectiveness of various data mining methods has been evaluated, so that the findings can be of a more **general impact** on various analyzes of poor data quality (even for **other types of disease**). Apart from data mining, the data set was prepared as well, also known as **preprocessing** / cleaning, which is part of the **KDD-Knowledge discovery from databases process**. In addition, a **novel method of missing values imputation** was tested.
We sought **data preprocessing strategies** and **classification methods that best categorize patients based on the target variable survival after 1 year** (of which we have solid knowledge). The above preprocessing and classification ways were **evaluated** in detail, using **numerous and more varied classification metrics** compared to what is often the case in the literature. This provides **complete and sperical** evaluation. We noted the emerging **applications** that were related to Healthcare:

- **categorization of patients with HCC** (survival after 1 year), based on the clinical or biological characteristics examined by the set we analyzed, and consequent
- **facilitation of clinical decision-making / risk-undertaking** that is dependent to patient risk (no over-treatment with high-risk treatments for low-risk patients, no under-treatment for high-risk patients)
- The **evaluation charts** of the tactics followed are available so that **health professionals can choose the pre-processing / classification tactic they deem most effective for their own analysis purposes** (in case they wish to use any of the ones tested in the current work).

2. Σκοπός

Η διπλωματική εργασία συνίσταται από δύο αλληλένδετα μέρη, το βιολογικό και το υπολογιστικό.

Επιλέχθηκε σύνολο δεδομένων που περιελάμβανε 49 βιολογικά-κλινικά γνωρίσματα ασθενών με Ηπατοκυτταρικό καρκίνο. Το σύνολο δεδομένων είχε και πληροφορία για το αν οι ασθενείς επιβίωσαν ή όχι μετά την πάροδο ενός έτους. Η επιλογή της προαναφερθείσας ασθένειας έγινε μετά από παρακολούθηση ορισμένων Ιατρικών Συνεδρίων, τα οποία αναφέρονται στα συνοδευτικά αρχεία της εργασίας, όπου καταγράφηκαν ενδεικτικές ασθένειες που είχαν περισσότερο ερευνητικό ενδιαφέρον και έχρηζαν περαιτέρω μελέτης. Το **Ηπατοκυτταρικό καρκίνωμα** (Hepatocellular Carcinoma – HCC) ήταν μία ασθένεια εξ' αυτών. Διαλέξαμε την συγκεκριμένη ασθένεια επειδή συγκεντρώνει τα εξής ιδιαίτερα **χαρακτηριστικά**, που αναλύονται στην Βιβλιογραφική ανασκόπηση:

- Έχει ακόμη μεγάλο ποσοστό θνητότητας και υποτροπής
- Προκύπτει από ποικίλα αίτια σχετιζόμενα με ασθένειες του Ήπατος
- Προκύπτει με διαφορετική συχνότητα σε διάφορες κατηγορίες ατόμων
- Ερευνάται ο καλύτερος χαρακτηρισμός του σε υποκατηγορίες

Σκοπός ήταν να γίνει **εξόρυξη δεδομένων στην πράξη, με αποδοτικότερο τρόπο, κάτω από κακές στατιστικά συνθήκες** που έχουν να κάνουν με την κακή ποιότητα σου συνόλου δεδομένων, καθώς και η **αξιοποίηση περισσότερων μεθόδων αξιολόγησης της ποιότητας της ανάλυσης** από ό,τι συνήθως συμβαίνει στην βιβλιογραφία -ώστε να υπάρχει **σφαιρικότερη εικόνα** για την αξιολόγηση. Παράλληλα στόχος ήταν να δοκιμαστούν και **νέες μέθοδοι προεπεξεργασίας των δεδομένων πριν την ανάλυση**, καθώς και η εκτίμηση της **συμβολής** των μεθόδων **αυτών στην βελτίωση της ποιότητας της ανάλυσης**. Το τελικό **βιολογικό ερώτημα** με το οποίο ασχοληθήκαμε αφορούσε **την ταξινόμηση ασθενών διαγνωσμένων με Ηπατοκυτταρικό καρκίνωμα σε** δύο κατηγορίες: **όσους αναμένεται να επιβιώσουν μετά το ένα έτος και όσους αναμένεται να αποβιώσουν μέχρι τότε.** Η χρησιμότητά του σχετίζεται με την εκτίμηση του κινδύνου να μην επιβιώσουν μετά το ένα έτος ασθενείς που μόλις διαγνώστηκαν με Ηπατοκυτταρικό καρκίνωμα.

Επιθυμητό ήταν να επιλεγεί για την ανάλυση **σύνολο δεδομένων ευρέως διαθέσιμο και αντιπροσωπευτικό ενός ρεαλιστικού συνόλου κλινικών / βιολογικών δεδομένων που συλλέγεται συνήθως στην πράξη** λόγω των συνήθων προβλημάτων που εμποδίζουν να γίνει ιδανική συλλογή. Επειδή - μεταξύ άλλων - το σύνολο που επιλέξαμε συνδυάζει, αρκετές κακές ποιότητες που συναντώνται συχνά σε σύνολα βιολογικών / κλινικών δεδομένων (μη ισορροπημένη εκπροσώπηση όλων των τύπων ασθενών που αφορά και τιμές δεδομένων που δεν έχουν συμπληρωθεί), επιλέχθηκε.

Ο δεύτερος στόχος ήταν η **εξερεύνηση τεχνικών που έχουν την δυνατότητα να δώσουν βελτιωμένη ποιότητα ανάλυση**ς κατά την εξόρυξη δεδομένων που πραγματοποιήθηκε, **ξεπερνώντας τους περιορισμούς που αναγκαστικά επάγονται από την κακή ποιότητα των δεδομένων**. Οι αναλύσεις έγιναν **παραδειγματικά**, ώστε τα ευρήματα να έχουν **γενικότερο** αντίκτυπο και πάνω σε αντίστοιχης, κακής ποιότητας δεδομένων, αναλύσεις (ακόμη και για άλλες ασθένειες).

Ακόμη ένας σκοπός ήταν η αναζήτηση στην βιβλιογραφία για τις **δυνατές εφαρμογές που προκύπτουν συνήθως** από εξορύξεις βιολογικών δεδομένων.

Τελευταίος σκοπός ήταν η διερεύνηση του αν εξάγονται βάσει των πιο πάνω εφαρμογές **στο πεδίο της Υγειονομικής Περίθαλψης**, οι οποίες καταγράφηκαν και αναλύονται παρακάτω.

3. Εισαγωγή-Βιβλιογραφική Ανασκόπηση

Οι εξελίξεις στον τομέα της τεχνολογίας έχουν επιτρέψει την διενέργεια μεγάλου όγκου πειραμάτων ***μοριακής βιολογίας*** σε διάφορους τομείς και με υψηλή ακρίβεια. Η συσσώρευση βιολογικών δεδομένων γίνεται με εκρηκτικούς ρυθμούς (Tzanis et al., 2005). Αναφέρονται ενδεικτικά πειράματα όπως αυτά της αλληλούχισης βιομορίων, του υβριδισμού νουκλεϊκών οξέων (πχ. πειράματα με μικροσυστοιχίες, SAGE), πρωτεομικής ανάλυσης, εύρεσης τριδιάστατων βιομοριακών δομών και αλληλούχισης γονιδιωμάτων (λχ. το Human Genome Project).

Επόμενη ήταν η ανάγκη για αποθήκευση των δεδομένων των αποτελεσμάτων με οργανωμένο τρόπο, σε νέες βιολογικές βάσεις δεδομένων. Υπήρξε επίσης συσσώρευση νέων εγγραφών στις ήδη υπάρχουσες βιολογικές βάσεις δεδομένων καθώς και αύξηση της πολυπλοκότητας των εγγραφών τους (Brusic and Zeleznikow, 1999). Όπως είναι αυτονόητο απαιτείται η χρήση υπολογιστών για αποθήκευση, συντήρηση και ανάλυση αυτών των δεδομένων (Tzanis et al., 2005).

Η επεξεργασία των περιεχομένων τέτοιων ψηφιακών αποθετηρίων έχει σκοπό την εξαγωγή χρήσιμης γνώσης από αυτά. Αυτή η διαδικασία λέγεται και Ανακάλυψη Γνώσης από Βάσεις Δεδομένων (*Knowledge Discovery from Databases - KDD*). Η Ανακάλυψη Γνώσης από Βάσεις Δεδομένων (*KDD*) γίνεται συμπληρωματικά του εργαστηριακού πειράματος και επιταχύνει την βιολογική έρευνα (Brusic and Zeleznikow, 1999). Η εξόρυξη δεδομένων (*data mining*) -με την οποία και θα ασχοληθούμε- αποτελλεί ένα από τα στάδια της Ανακάλυψης Γνώσης από Βάσεις Δεδομένων (*KDD*).

Ο όρος «εξόρυξη δεδομένων» (*data mining*) έχει επικρατήσει να χρησιμοποιείται στην διεθνή βιβλιογραφία σε εναλλαγή με τον όρο «Ανακάλυψη Γνώσης από Βάσεις Δεδομένων» (*KDD*) (Tzanis et al., 2005), κάτι που δεν είναι όμως δόκιμο, μιας και η εξόρυξη δεδομένων αφορά μόνο ένα μέρος της Ανακάλυψης Γνώσεων από Βάσεις Δεδομένων. Παρά το ότι κατά τις αναλύσεις εξάγεται γνώση από τα δεδομένα και όχι δεδομένα καθεαυτά, χάρην συντομίας και λόγω επικράτησης του όρου «εξόρυξη δεδομένων» στην βιβλιογραφία στην παρούσα εργασία θα γίνει χρήση του όρου αυτού, έχοντας όμως διευκρινίσει την θέση μας ώς προς την διαφωνία που υπάρχει στην βιβλιογραφία για τον όρο.

Υπάρχουν διάφορες τεχνικές που αξιοποιούνται για την Εξόρυξη Δεδομένων. Αυτές συμπεριλαμβάνουν μεθόδους σχετικές με γενίκευση (*generalization*), χαρακτηρισμό (*characterization*), κατηγοριοποίηση (*classification*), εύρεση συναθροίσεων (*clustering*), εύρεση συσχετίσεων (*association*), εξέλιξης (*evolution*), ομοιότητας μοτίβων (*pattern matching*), οπτικοποίησης δεδομένων (*data visualization*) και εξόρυξη που καθοδηγείται από μετα-κανόνες (*meta-rule guided mining*) (Jothi et al., 2015). Τα δεδομένα υπόκεινται σε στατιστική ανάλυση, ενώ αρκετές φορές αποτελούν και «τροφή» σε προγράμματα τεχνιτής νοημοσύνης (επιβλεπόμενα ή όχι – *supervised* ή *unsupervised*), τα οποία εκπαιδεύονται με βάση τα δεδομένα για την δημιουργία μοντέλων περιγραφής τους ή και πρόγνωσης ανάλογα με τα εκάστοτε ερευνητικά ενδιαφέροντα.

Στον τομέα της Εξόρυξης Δεδομένων (*data mining*) απασχολείται πληθώρα επιστημόνων. Απαιτείται καλή γνώση του επιστημονικού αντικειμένου από το οποίο προέρχονται τα δεδομένα, της στατιστικής καθώς και ερευνητική ειδίκευση, με γνώσεις εξόρυξης δεδομένων και πληροφορικής (Koh and Tan, 2005).

Η Εξόρυξη Δεδομένων (*data mining*) χρησιμοποιείται ήδη στον τομέα της Υγειονομικής Περίθαλψης για σκοπούς τόσο Ιατρικούς, όσο και διοικητικούς ή οικονομικούς. Σαν πρώτη ύλη χρησιμοποιούνται δεδομένα οικονομικής φύσης, γραμματειακής φύσης, αλλά και ιατρικής φύσης, όπως καταγραφές ιατρικών ιστορικών ή κλινικής παρακολούθησης ασθενών και ιατρικές σημειώσεις. Είναι σαφές ότι τα πρώτα δεν άπτονται του αντικειμένου της εργασίας αυτής, ενώ τα τελευταία θα ήταν μία χρήσιμη πηγή δεδομένων.

Τα είδη αναλύσεων των δεδομένων που υπάρχουν είναι η Περιγραφική Ανάλυση (*Descriptive Analysis*), Προγνωστική Ανάλυση (*Predictive Analysis*) και η Καθοδηγητική (*Prescriptive*) Ανάλυση. Η πρώτη είναι η σύνοψη του συνόλου δεδομένων. Η δεύτερη αφορά την δημιουργία στατιστικών μοντέλων βάσει εκπαιδευτικού διαθέσιμου συνόλου δεδομένων με στόχο την πρόγνωση ιδιοτήτων που αφορούν τις νέες εγγραφές δεδομένων που θα υπάρξουν (Mohammed et al., 2014). Η τελευταία εξετάζει την συμπεριφορά του μοντέλου που δημιουργήθηκε από το σύνολο δεδομένων κάτω από το καθεστώς διαφορετικών σεναρίων - χρήσιμη κυρίως σε προβλήματα βελτιστοποίησης και όχι ιδιαίτερα σε τομείς όπως αυτόν της ανάλυσης κλινικών δεδομένων (Mohammed et al., 2014).

3.1 Πηγές Βιολογικών Δεδομένων

Οι Βιολογικές Βάσεις Δεδομένων περιλαμβάνουν μεγάλη ποικιλία τύπων δεδομένων και συχνά είναι δομημένες σχεσιακά (Page and Craven, 2003). Τέτοια βιοπληροφορικά δεδομένα μπορεί να αφορούν αλληλουχίες γονιδιωμάτων, μικροσυστοιχίες DNA, δεδομένα από SAGE (σειριακή ανάλυση γονιδιακής έκφρασης-*serial analysis of gene expression*), δομές πρωτεϊνών και μικρών μορίων, διαμοριακές αλληλεπιδράσεις, Γονιδιακές Οντολογίες (*GO*), μονοπάτια ασθενειών κ.ά. (Tzanis et al., 2005)

Βρίσκει κανείς ωστόσο και Βιολογικά Σύνολα Δεδομένων που μπορεί να προέρχονται από άλλες πηγές. Κάποιες μπορεί να πλησιάζουν την ιατρική συνιστώσα λ.χ. δεδομένα από Ηλεκτρονικά Ιατρικά Αρχεία (*electronic medical record -EMR*), βιοϊατρικά (*biomedical*) πχ. δεδομένα από απεικονιστικές ιατρικές μεθόδους, βιομετρικά (*biometric*), από πληροφοριακά Συστήματα Εργαστηριακών Πειραμάτων μεγάλης κλίμακας (*large-scale laboratory information system data-LIS*) αλλά και από δοκιμαστικές χρήσεις (*test utilisation*) (Mohammed et al., 2014).

Άλλες συνήθεις πηγές δεδομένων μπορεί να είναι οι εκθέσεις παρακολούθησης ασθενών και άλλα καθοριστικά ευρήματα (Jothi et al., 2015) όταν αναφερόμαστε κυρίως στα ιατρικά δεδομένα. Υπάρχουν και δεδομένα από καταγραφή Βιοϊατρικών σημάτων (*πχ. από βιοαισθητήρες, ηλεκτρο-εγκεφαλογράφημα κά*) αλλά και εικόνων (*πχ. Ιστολογικές εικόνες ή και από MRI*) (Mohammed et al., 2014).

Το διαδίκτυο είναι σημαντική πηγή για τέτοια σύνολα δεδομένων, μιας και οι περισσότερες προαναφερθείσες κατηγορίες δεδομένων είναι αποθηκευμένες σε βάσεις υποστηριζόμενες από διαδικτυακή πλατφόρμα για πραγματοποίηση ερωτημάτων, αναζήτησης και σε κάποιες περιπτώσεις αναλύσεων βασιζόμενες σε αλγορίθμους τεχνητής νοημοσύνης, στατιστικής σε πραγματικό χρόνο από τους χρήστες. Τέτοιες βάσεις (*βιολογικές και κλινικές*) συχνά είναι δημόσιες, ενώ υπάρχουν και κάποιες που χρησιμοποιούνται από βιομηχανίες με την πρόσβαση να μην είναι ελεύθερη σε όλες (Tzanis et al., 2005).

Υπάρχουν ωστόσο προβλήματα που αφορούν την δυνατότητα εύρεσης και την επεξεργασίας δεδομένων που αφορούν την υγεία. Τέτοια είναι:

Η διασπορά τους σε διάφορες δομές όπως πχ. οι δομές υγειονομικής περίθαλψης, τα εργαστήρια, οι πωλητές δεδομένων, οι οικονομικές και ρυθμιστικές πηγές και η έλλειψη ψηφιοποίησης -κυρίως σε υγειονομικά δεδομένα (Mohammed et al., 2014). Επίσης συχνά τα σύνολα των δεδομένων είναι δυναμικά (Holzinger and Jurisica, 2014) και μεταβάλλονται συν τω χρόνω όσο η καταγραφή γίνεται πληρέστερη.

Η πρόσβαση στα δεδομένα δεν είναι πάντα ελεύθερη αλλά, ακόμη και όταν είναι ελεύθερη, βρίσκουμε ότι οι πηγές είναι μη συνδεδεμένες και συνήθως παρατηρείται έλλειψη κάποιου κοινού προτύπου οργάνωσης των αρχείων δεδομένων (Holzinger and Jurisica, 2014), αν και αυτό το πρόβλημα έχει ξεπεραστεί σε μεγάλα διεθνή project (πχ. Genbank-DDBJ-ENA). Άλλες περιπτώσεις αφορούν αρχεία με (παντελή) έλλειψη δόμησης-οργάνωσης ή ελλειπή και κατεστραμμένα αρχεία και ασυνέπειες στα δεδομένα, με μεγάλο όγκο, πολυπλοκότητα, ετερογένεια και κακό μαθηματικό χαρακτηρισμό (Koh and Tan, 2005).

Υπάρχουν – πολύ ευλόγως – ηθικά, νομικά και κοινωνικά ζητήματα πχ. ιδιοκτησία δεδομένων, ζητήματα απορρήτου (Koh and Tan, 2005) τα οποία χρειάζεται οι ερευνητές να έχουν υπ’ όψην.

Συχνά τα δεδομένα είναι ποικίλα, περίπλοκα και περιλαμβάνουν μεγάλο αριθμό διαστάσεων. (Holzinger and Jurisica, 2014)

Η εξονυχιστική ανάλυση δεδομένων αναπόφευκτα αποφέρει την εύρεση μοτίβων που προέκυψαν από τύχη (λόγω της φύσης τους), για αυτό και είναι απαραίτητη η σαφής στόχευση και θέση επιστημονικών ερωτημάτων για την ανάλυση δεδομένων, γιατί αλλιώς θα καταλήξουμε σε ψάρεμα δεδομένων/εκβάθυνση δεδομένων (*data fishing/dredging*) (Koh and Tan, 2005).

Συνεπώς η διαδικασία της ανάλυσης των δεδομένων είναι Χρονο- και κοστο-βόρα, απαιτεί αφοσιωμένη εργασία, και πόρους (Koh and Tan, 2005). Συγχρόνως προκύπτει η ανάγκη για δημιουργία μηχανισμού ενοποίησης και συντήρησης δεδομένων εντός “αποθηκών” δεδομένων (*data warehouse*) (Mohammed et al., 2014).

Με την Βιοπληροφορική -μεταξύ άλλων- επιτυγχάνεται και η οργάνωση των δεδομένων με τρόπο ώστε να υπάρχει πρόσβαση στην υπάρχουσα πληροφορία από τους ερευνητές, αλλά και να υπάρχει η δυνατότητα υποβολής νέων δεδομένων. Παράλληλα, αναπτύσσονται εργαλεία που βοηθούν στην ανάλυση των δεδομένων (Tzanis et al., 2005). Επίσης οι περισσότερες μεγάλες Βάσεις Βιολογικών Δεδομένων έχουν αναπτύξει μεθόδους για επίλυση των προβλημάτων συμβατότητας και συγχρονισμού τους, καθώς και της επικύρωσης και κωδικοποίησης των καταχωρήσεων που φέρουν.

Συγχρόνως οι Επιστήμονες του αντικειμένου της Βιοπληροφορικής πραγματοποιούν αναλύσεις σε σύνολα δεδομένων που είναι διαθέσιμα προκειμένου να εξάγουν από αυτά χρήσιμα συμπεράσματα. Τα σύνολα δεδομένων που χρησιμοποιούν επιλέγουν συνήθως να είναι καλής ποιότητας.

Όμως μέχρι να επιτευχθούν οι πιο πάνω στόχοι και να έχουμε καλή ποιότητα σχεδόν σε όλα τα διαθέσιμα σύνολα δεδομένων, θα συνεχίσει να υπάρχει πίεση για ανάλυση σε συνόλα δεδομένων που είναι άμεσα διαθέσιμα και συχνά είναι μικρά ή ατελή, ειδικά όταν αφορούν σπάνιες παθήσεις. Οι αποσπασματικές καταγραφές (κενές τιμές, λίγες καταγραφές), με κακή ποιότητα και η ανισσορροπία (έλλειψη ίσης εκπροσώπησης όλων των κατηγοριών ασθενών) στα σύνολα δεδομένων επιδεινώνουν την ποιότητα της ταξινόμησης που προκύπτει από τις εκάστοτε διαθέσιμες μεθόδους.

3.2 Εξόρυξη Δεδομένων (Data mining)

Όπως αναφέρθηκε και προηγουμένως, η εξόρυξη δεδομένων (ή ανακάλυψη γνώσης) είναι διεπιστημονική διαδικασία που περιλαμβάνει μηχανική μάθηση, στατιστική, τεχνητή νοημοσύνη, βάσεις δεδομένων, αναγνώριση μοτίβων και οπτικοποίηση δεδομένων (li, 2013), ενώ έχει συμβολές και από την ανάκτηση πληροφοριών και τα παράλληλα και κατανεμημένα υπολογιστικά συστήματα και τα συστήματα προεπεξεργασίας-αποθήκευσης δεδομένων (data warehousing) (Tzanis et al., 2005).

Παρότι υπάρχουν έτοιμες σουΐτες για διεξαγωγή αναλύσεων εξόρυξης δεδομένων (data mining) ακόμα και από άτομα που δεν έχουν γνώσεις προγραμματισμού, η γνώση γραφής κώδικα (scripting) σε διάφορες γλώσσες προγραμματισμού -όπως για παράδειγμα η Python και η R (R Core Team, 2018) – δίνει πολλαπλάσιες δυνατότητες για πιο ευέλικτη, ταχύτερη υπολογιστικά και αποδωτική ανάλυση, καθώς και για ποικιλία στο ρεπερτόριο τεχνικών ανάλυσης που μπορούν να εφαρμοστούν στα δεδομένα. Όσο ο όγκος των δεδομένων αυξάνει και πλησιάζει την τάξη των μεγάλων δεδομένων (big data), τόσο αυξάνεται και η ανάγκη για βελτιστοποίηση της ταχύτητας υπολογισμών και υιοθετούνται πιο αποτελεσματικά υπολογιστικά συστήματα και λογισμικά.

Μιας και οι Οντολογίες είναι αρκετά διαδεδομένες στις μέρες μας, είναι χρήσιμο να αναφερθεί και η μετα-εξόρυξη (meta-mining), που προσεγγίζει την εξόρυξη δεδομένων όχι απλώς αναζητώντας μοτίβα στα πρωτογενή δεδομένα, αλλά ενσωματώνοντας και πληροφορίες που παρέχονται από τη δομή της οντολογίας αναπαράστασης (Scheuermann et al., 2009). Η μετα-εξόρυξη (meta-mining) αντιπροσωπεύει την εξόρυξη και ανάλυση ολοκληρωμένων συνόλων γνώσεων που προέρχονται από πολλαπλές, συχνά διαφορετικές, πηγές και μπορεί να χρησιμοποιηθεί για ένα ευρύ φάσμα διαφορετικών δραστηριοτήτων εξόρυξης δεδομένων, όπως ευρετηρίαση και ανάκτηση δεδομένων και πληροφοριών, χαρτογράφηση μεταξύ οντολογιών, ενσωμάτωση δεδομένων, ανταλλαγή δεδομένων, σημασιολογική διαλειτουργικότητα, επιλογή και συγκέντρωση δεδομένων, υποστήριξη αποφάσεων, επεξεργασία φυσικής γλώσσας, εφαρμογές και ανακάλυψη γνώσεων (Scheuermann et al., 2009).

Μεγάλης σημασίας είναι και η εξόρυξη γνώσης από κείμενα (πχ. Επιστημονικά άρθρα) μέσω της εξόρυξης κειμένου (text-mining), αλλά δεν γίνεται εκτενής αναφορά στην παρούσα εργασία για αυτό, δεδομένου του ότι διαφεύγει των ορίων της θεματικής που έχει οριστεί.

Οι (Holzinger and Jurisica, 2014) προτείνουν την Ανακάλυψη Γνώσης από Αλληλεπίδραση Ανθρώπου-Υπολογιστή και Εξόρυξη Δεδομένων (HCI-KDD, Human-Computer Interaction Knowledge Discovery & Data Mining) *σαν λύση στο πρόβλημα της δημιουργίας και εφαρμογής νέων μεθόδων, αλγορίθμων και εργαλείων για συμπλήρωση, σύντηξη, προ-επεξεργασία, χαρτογράφηση, ανάλυση και ερμηνεία πολύπλοκων βιοϊατρικών δεδομένων, τον καιρό που η νέα βιολογική και ιατρική πληροφορία έρχεται με ρυθμούς μεγαλύτερους από αυτούς που οι επιστήμονες μπορούν να διαχειριστούν μόνοι τους. Εστιάζουν στην μετάβαση από ανεξάρτητα συστήματα σε ενοποιημένα και διαδραστικά και την ανάγκη της ανάλυσης με χρήση μεθόδων που βασίζονται σε γράφους για δεδομένα, καθώς και την ανάγκη για χρήση Οντολογιών και πτυχών του Σημασιολογικού Ιστού (Semantic Web). Αναφέρουν την σημασία της διαδραστικής εξόρυξης μοτίβων (pattern mining), της ανάλυσης δικτύων για συνδυασμό δεδομένων από διαφορετικά πειράματα, την συμπλήρωση*

αποκλίνοντων συνόλων δεδομένων, από διαφορετικά πειράματα και πηγές καθώς και νέων μεθόδων που αφορούν δεδομένα από τον κλάδο απεικονιστικών εξετάσεων. Γίνεται επίσης αναφορά και στην Οπτική Εξόρυξη Δεδομένων (Visual Data Mining) σαν πιθανή λύση -παρά τις δυσκολίες της- για μεγάλα σετ δεδομένων, με πολλές διαστάσεις.

3.3 Σχετικές εργασίες

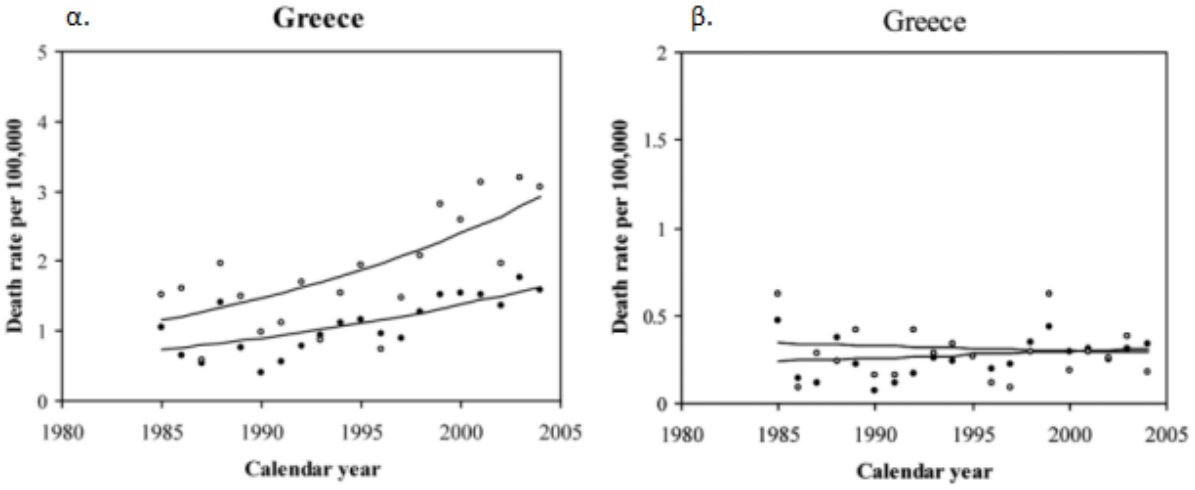
Οι αρθρογράφοι που κατέθεσαν το σύνολο με τα δεδομένα που χρησιμοποιήθηκαν στην παρούσα εργασία χρησιμοποίησαν σαν μέθοδο προεπεξεργασίας των δεδομένων, για συμπλήρωση των τιμών που έλειπαν στα δεδομένα την *μέθοδο Κ-κοντινότερων γειτόνων, με χρήση της Ετερογενούς Ευκλείδειας – Μετρικής επικάλυψης (Heterogenous Euclidean – Overlap Metric - HEOM) σαν μέτρο της απόστασης. Η HEOM είναι κατάλληλη για συνεχείς και διακριτές μεταβλητές, αλλά και για ελλειπή δεδομένα. Δημιούργησαν συνθετικά σύνολα δεδομένων με βάση το αρχικό, προκειμένου να μπορέσουν να εκπροσωπηθούν με στατιστική σημαντικότητα και υποομάδες των δεδομένων που είναι σε μειοψηφία και δεν εκπροσωπούνται με τις κλασικές τεχνικές. Χρησιμοποίησαν τον αλγόριθμο SMOTE (Τεχνική συνθετικής υπερδειγματοληψίας μειοψηφιών - synthetic minority over-sampling technique) για την δημιουργία νέου συνθετικού συνόλου. Επίσης δημιούργησαν και δικό τους αλγόριθμο για δημιουργία συνθετικών συνόλων που αυξάνει την εκπροσώπηση των μειονοτήτων στα δεδομένα, με χρήση μη επιβλεπόμενης μεθόδου και συγκεκριμένα μιας μεθόδου εύρεσης συστάδων (clustering). Χρησιμοποίησαν την μη επιβλεπόμενη μέθοδο εύρεσης συστάδων εντός των δεδομένων (clustering) K-means. Υπολόγισαν το βέλτιστο αριθμό Κ- κεντροειδών με χρήση της στατιστικής GAP. Στη συνέχεια δημιούργησαν ένα σύνολο δεδομένων για κάθε συστάδα που ενισχύεται. Ενώσαν όλα τα παραγώμενα σύνολα σε ένα ενιαίο, αντιπροσωπευτικό και ισορροπημένο σύνολο. Στη συνέχεια συνδύασαν κάθε ένα από τα ενισχυμένα σύνολα με το ενιαίο και πραγματοποίησαν ταξινόμηση με λογιστική παλλινδρόμηση ισάριθμες φορές. Τα τελικά αποτελέσματα της ταξινόμησης για κάθε καταχώρηση λήφθησαν με πλειοψηφική ψήφο των επιμέρους ταξινομήσεων*. (Santos et al., 2015) Στην παρούσα εργασία δοκιμάστηκαν διαφορετικοί από τους πιο πάνω τρόποι συμπλήρωσης κενών τιμών, που αναλύονται πιο κάτω, αλλά και επιπλέον μέθοδοι ταξινόμησης.

Σε έρευνα των (Tzanis et al., 2006) χρησιμοποιήθηκαν νέες μεθόδοι εξόρυξης δεδομένων (data mining) για την έρευση των θέσεων εκκίνησης της μετάφρασης (TIS) σε μόρια mRNA πυρηνικών γονιδίων. Εκ πρώτης όψης δεν σχετίζεται με την παρούσα εργασία, ωστόσο οι συγκεκριμένοι ερευνητές χρησιμοποίησαν ταξινομητές των κατηγοριών Δενδρών Αποφάσεων, Πινάκων Αποφάσεων, Naive Bayes, SVM, Νευρωνικό Δίκτυο Perceptron (πολλών στρωμάτων) και Κ-κοντινότερων γειτόνων και για αυτό αναφέρονται εδώ. Επίσης χρησιμοποίησαν τον αλγόριθμο RIPPER και πρότειναν δύο αλγορίθμους που αξιοποιούν την απλή πλειοψηφική ψηφορία των επιμέρους ταξινομητών προκειμένου να ταξινομήσουν μία νέα καταχώρηση είτε την ανισοβαρή πλειοψηφική ψηφορία. Αξιολόγηση των αποτελεσμάτων γίνεται με την μέθοδο του leave-one-out, με χρήση των 9/10 του σετ για εκπαίδευση των μοντέλων ταξινόμησης και το 1/10 του σετ για τεστ και επανάληψη της διαδικασίας αυτής 10 φορές. Αυτή ήταν η μέθοδος αξιολόγησης που χρησιμοποιήθηκε και στην παρούσα εργασία.

3.4 Ηπατοκυτταρικό Καρκίνωμα (Hepatocellular carcinoma – HCC)

Στην παρούσα εργασία κάναμε εξόρυξη δεδομένων σε σύνολο δεδομένων που σχετίζεται με το Ηπατοκυτταρικό καρκίνωμα. *Το ηπατοκυτταρικό καρκίνωμα είναι ο κυριότερος ιστολογικός τύπος που συναντούμε στα πρωτοπαθή κακοήθη νεοπλάσματα του ήπατος.* (Janevska, Chaloska-Ivanova and Janevski, 2015) *Το ήπαρ είναι επίσης ένα από τα πιο κοινά σημεία δευτεροπαθών όγκων και σημαντικό είναι να διευκρινίσουμε ότι η διάκριση μεταξύ πρωτοπαθούς και μεταστατικού καρκίνου του ήπατος μπορεί να γίνεται σωστά ή λανθασμένα και ανάλογα με το ισχύον πρωτόκολλο σε κάθε χώρα ή χρονική περίοδο.* (Bosetti et al., 2008) Το Ηπατοκυτταρικό καρκίνωμα -με το οποίο και ασχοληθήκαμε- είναι ο *πέμπτος κοινότερος καρκίνος παγκοσμίως και η τρίτη πιο κοινή αιτία θανάτου από καρκίνο.* (Ada Hamosh, George E. Tiller, Cassandra L. Kniffin, Paul J. Converse, Victor A. McKusick, John A. Phillips, III, 2020) Είναι επίσης *η πιο κοινή αιτία θανάτου σε άτομα με κίρρωση.* ('Hepatocellular carcinoma', 2021)

Ενδεικτικά, παλιότερα στην *Ελλάδα το συνολικό -τυποποιημένο για την ηλικία- ποσοστό θνησιμότητας από HCC ανά 100.000 άνδρες - γυναίκες, είχε αυξηθεί κατά 101.3% για την περίοδο 2000-2004 (σε σχέση με την περίοδο 1990-94) σε άρρενες ασθενείς, ενώ οι θήλεις ασθενείς παρουσίαζαν αύξηση κατά 76.5%. Τα αντίστοιχα ποσοστά ανά 100.000 άνδρες ή γυναίκες ηλικίας 45-59 ετών ήταν 124.0% για τους άρρενες ασθενείς και -3.7% για τις θήλεις ασθενείς. Το ετήσιο ποσοστό μεταβολής θνησιμότητας μεταξύ 1985-2004 ήταν 5.0 για άρρενες ασθενείς και για θήλεις -1.0.* (Bosetti et al., 2008) Το μοτίβο αυτό φαίνεται στην ακόλουθη εικόνα, που επίσης προέρχεται από την εργασία των (Bosetti et al., 2008) και αναπαριστά τα αποτελέσματα της *συνδυασμένης ανάλυσης που πραγματοποίησαν για τη θνησιμότητα από Ηπατοκυτταρικό καρκίνο (HCC) σε γυναίκες και άντρες όλων των ηλικιών (γεμάτοι κύκλοι στο διάγραμμα) και ηλικίας 45-59 ετών (κενοί κύκλοι στο διάγραμμα) σε ευρωπαϊκές χώρες από 1970-2002.* Το α. αφορά άρρενες ασθενείς και το β. θήλεις ασθενείς.



3.4.1 Παράγοντες κινδύνου για Ηπατοκυτταρικό καρκίνωμα

Ποσοτικά πρωτεωμικά δεδομένα δείχνουν ότι υπάρχει ετερογένεια στο ηπατοκυτταρικό καρκίνωμα πρώιμου σταδίου. Οι ασθενείς διακρίθηκαν σε υποτύπους S-I, S-II και S-III, καθένας με διαφορετική κλινική έκβαση. Ο S-III, με διαταραγμένη την ομοιόσταση χοληστερόλης, σχετίζεται με το χαμηλότερο συνολικό ποσοστό επιβίωσης και τον μεγαλύτερο κίνδυνο κακής πρόγνωσης μετά την χειρουργική επέμβαση πρώτης γραμμής." (Ada Hamosh, George E. Tiller, Cassandra L. Kniffin, Paul J. Converse, Victor A. McKusick, John A. Phillips, III, 2020)

Οι κύριοι παράγοντες κινδύνου για την ανάπτυξη του Ηπατοκυτταρικού καρκινώματος είναι η *μόλυνση από τον ιό της ηπατίτιδας Β (HBV), από τον ιό της ηπατίτιδας C (HCV), η παρατεταμένη έκθεση σε αφλατοξίνη μέσω διατροφής και η αλκοολική (ή λόγω άλλων αιτιών) κίρρωση*". (Ada Hamosh, George E. Tiller, Cassandra L. Kniffin, Paul J. Converse, Victor A. McKusick, John A. Phillips, III, 2020) *Παρουσία κίρρωσης, υπάρχει κίνδυνος ανάπτυξης Ηπατοκυτταρικού καρκινώματος (1-2% ετησίως).* (Hübscher, 2011) *Το Ηπατοκυτταρικό καρκίνωμα εμφανίζεται σε πλαίσιο χρόνιας ηπατικής φλεγμονής και επιπλέον παράγοντες κινδύνου είναι και η έκθεση σε τοξίνες όπως το αλκοόλ, ή τα αλκαλοειδή πυρρολιζιδίνης, καθώς και ασθένειες όπως η αιμοχρωμάτωση, η ανεπάρκεια άλφα 1-αντιθρυψίνης* ('Hepatocellular carcinoma', 2021) *καθώς και η ασθένεια Wilson εφόσον υπάρχει κίρρωση.* ('EASL-EORTC clinical practice guidelines: management of hepatocellular carcinoma.', 2012)

Επιπλέον παράγοντες κινδύνου για την ανάπτυξη και έκβαση του Ηπατοκυτταρικού καρκινώματος είναι η διαταρραχή της ομοιόστασης σιδήρου -οφειλόμενη σε διάφορους παράγοντες. Η Εργασία των (Sung and Bae, 2014) συνοψίζει πολύ εύστοχα το ζήτημα και μεταξύ άλλων αναφέρει ότι η *κυτταρική ομοιόσταση σιδήρου είναι σημαντικός ρυθμιστής της παραγωγής Ελεύθερων ριζών οξυγόνου (ROS) των οποίων το επίπεδο καθορίζει την κυτταρική δυνατότητα μετανάστευσης. Το ηπατικό προκαρκινογόνο p-διμεθυλαμινοαζοβενζίνη (p-DAB) αυξάνει τις ελεύθερες ρίζες προκαλώντας υπερβολική αιμόλυση και επιταχύνοντας την εναπόθεση σιδήρου στο ήπαρ αυξάνοντας την οξειδωτική βλάβη. Η βαριά αλυσίδα φερριτίνης (FHC, με λειτουργία φερροξειδάσης) είναι από τις σημαντικά τροποποιημένες πρωτεΐνες στην κατάσταση της -επαγόμενης από τον αυξητικό παράγοντα μετασχηματισμού β1-επιθηλιακής-μεσεγχυματικής μετάπτωσης (EMT) σε ηπατοκύτταρα ποντικού και κατ' επέκταση στην κυτταρική μετανάστευση. Η κληρονομική αιμοχρωμάτωση (HH) οφείλεται σε γενετικές βλάβες γονιδίων βιοχημικών μονοπατιών απορρόφησης, μεταφοράς και αποθήκευσης σιδήρου, πχ. γονίδιο HFE, γονίδιο HAMP (αυτομικροβιακό πεπτίδιο εψιδίνης) ή στο HFE2 (γνωστό ως HJV), γονίδιο μεταφοράσης R2 (TFR2), γονίδιο σιδηροπορτίνης (φερροπορτίνης-FPN, SLC40A1).* (Sung and Bae, 2014)

Μετά τον εκτεταμένο εμβολιασμό για τον ιό της ηπατίτιδας Β (HBV) και την εμφάνιση αντιϊικών φαρμάκων άμεσης δράσης για τη λοίμωξη από HCV, αναδείχθηκαν και άλλοι παράγοντες ως κύριοι παγκόσμιοι παράγοντες κινδύνου για εμφάνιση Ηπατοκυτταρικού καρκινώματος (HCC), πχ. η Μη αλκοολική λιπώδης ηπατική νόσος (Non-alcoholic fatty liver disease-NAFLD) και συναφείς καταστάσεις όπως ο διαβήτης και η παχυσαρκία (Martínez-Chantar, Avila and Lu, 2020) *Στους παράγοντες κινδύνου για ανάπτυξη Ηπατοκυτταρικού Καρκινώματος (HCC) ανήκουν συνεπώς μεταβολικό σύνδρομο, καθώς και η μη-αλκοολική στεατοηπατίτιδα (NASH).* ('Hepatocellular carcinoma', 2021) *Αλλά και σε ενήλικες με χρόνια λοίμωξη από τον ιό της Ηπατίτιδας Β (HBV), ο Διαβήτης αποτελεί σημαντικό παράγοντα κινδύνου για ανάπτυξη Ηπατοκυτταρικού καρκίνου (HCC)* (Campbell et al., 2021) Από το βιβλίο των (Xu and Yu, 2017, p. 20) βρέθηκε σύνοψη παραγόντων που *συμβάλουν στην ανάπτυξη NASH και ηπατικού καρκινώματος. Η ισομορφή JNK1 του γονιδίου JNK προάγει ηπατική στεάτωση και φλεγμονή και συμβάλει στην ενεργοποίηση του συμπλέγματος mTOR1. Η ενεργοποίηση της JNK1 οδηγεί σε φωσφορυλίωση του IRS-1 και συμβάλει στην επαγόμενη από την παχυσαρκία αντίσταση στην ινσουλίνη, την αυξημένη ηπατική φλεγμονή, την ίνωση και την απόπτωση. Η ισομορφή JNK2 από την άλλη αναστέλλει τον κυτταρικό θάνατο.* (Xu and Yu, 2017, p. 20)

Σύνοψη για τον ρόλο του αλκοόλ στον Ηπατοκυτταρικό καρκίνωμα βρίσκουμε στην εργασία των (Hübscher, 2011) Η αυξημένη *πρόσληψη αλκοόλ έχει αναγνωριστεί ως παράγοντας κινδύνου για καρκίνο και σε άλλα όργανα, λόγω επιδράσεών του που προάγουν και την ηπατική καρκινογένεση, συμπεριλαμβανομένου του οξειδωτικού στρες, των βλαβών στο DNA και της παραγωγής καρκινογόνων μεταβολιτών όπως η ακεταλδεΰδη.* (Hübscher, 2018) *Έως και 20% των Ηπατοκυτταρικών καρκινωμάτων που σχετίζονται με το αλκοόλ μπορεί να προκύψει σε υπόβαθρο προκιρρωτικής ηπατικής νόσου. Εκτός από αυτά, το αλκοόλ είναι και πολύ σημαντικός παράγοντας κινδύνου κίρρωσης ήπατος.* (Hübscher, 2011) *Ασθενείς με λιπώδη ηπατική νόσο έχουν αυξημένο κίνδυνο ανάπτυξης Ηπατοκυτταρικού καρκίνου με χαρακτηριστικά στεατοηπατίτιδας, -αν και αυτό αφορά συνήθως ασθενείς με NAFLD. Υπήρξε μελέτη όπου υψηλότερη συχνότητα Ηπατοκυτταρικού καρκίνου υπήρχε σε ασθενείς με Αλκοολική ηπατική νόσο (Alcoholic liver disease-ALD) (24%) σε σύγκριση με ασθενείς με μη λιπώδη κίρρωση του ήπατος.* (Hübscher, 2018)

Η *παρατεταμένη έκθεση σε CCl_4 προκαλεί λιπώδες ήπαρ, ίνωση, μέσω της πυροδότησης έντονου οξειδωτικού στρες (με υπεροξείδωση των λιπιδίων στα ηπατικά παρεγχυματικά κύτταρα), και θεωρείται ακόμα μία από τις αιτίες της ηπατικής καρκινογένεσης. Από του στόματος χορήγηση αποξηραμένης σκόνης ολόκληρου του φυτού (0,50 g/kg β. β.) του Ocimum sanctum L. Syn (Holy Basil ή Tulsi) σε εναιώρηση σε νερό για επτά συνεχόμενες ημέρες, και επίσης, τα φυτοχημικά ουρσολικό οξύ και ευγενόλη έχουν ηπατοπροστατευτική δράση έναντι της τοξικότητας από CCl_4.* (Baliga et al., 2013)

Στην εργασία των (Sung and Bae, 2014) βρίσκουμε ορισμένους από τους *φαρμακευτικούς παράγοντες που έχουν σχετιστεί με ηπατοκαρκινογένεση. Το Thorotrast είναι ένας ραδιοσκιαγραφικός παράγοντας στην ιατρική ακτινογραφία που ήταν σε χρήση τις δεκαετίες του 1930 και του 1940, έως και την δεκαετία του 1950 σε ορισμένες χώρες* ('Thorotrast', 2021). Έχει προκαλέσει διάφορους τύπους *κακοηθών όγκων μεταξύ αυτών και ηπατοκυτταρικά καρκινώματα, λόγω βλαβών που μπορεί να οφείλονται στη ραδιενέργεια και όχι στη χημική τοξικότητά του* (Sung and Bae, 2014). Έχει *παρατηρηθεί ανάπτυξη ηπατοκυτταρικού καρκινώματος λόγω μακροπρόθεσμης λήψης αναβολικών στεροειδών καθώς και αντισυλληπτικών στεροειδών*. (Sung and Bae, 2014) *Άλλοι φαρμακευτικοί παράγοντες που έχουν σχετιστεί με ηπατοκαρκινογένεση είναι η ταμοξιφαίνη (Tamoxifen) - μη στεροειδές φάρμακο τριφαινυλινικής δομής, που εμφανίζει σύνθετο φάσμα αντι-οιστρογονικών και οιστρογονικών φαρμακολογικών επιδράσεων στους διάφορους ιστούς* ('Ταμοξιφαίνη', Γαληνός Οδηγός Φαρμάκων) - *και η δαναζόλη (danazol) -ανδρογόνο με ήπια δράση που χρησιμοποιείται για την θεραπεία της ενδομητρίωσης, για την ινοκυστική μαστοπάθεια και προφυλακτικώς για το κληρονομικό αγγειοοίδημα* ('Δαναζόλη', Γαληνός Οδηγός Φαρμάκων). (Sung and Bae, 2014)

Σε υπόβαθρο χρόνια Ηπατίτιδας, η λοίμωξη με τον ιό της ανθρώπινης Ανοσοανεπάρκειας (HIV) είναι επίσης επιβαρυντική ('EASL-EORTC clinical practice guidelines: management of hepatocellular carcinoma.', 2012)

Επίσης, το *Ηπατοκυτταρικό καρκίνωμα είναι περισσότερο από τέσσερις φορές πιο συχνό στους άντρες από ότι στις γυναίκες*. ('Hepatocellular carcinoma', 2021)

Στην *Αφρική και την Ανατολική Ασία (χώρες που είναι ενδημικές της Ηπατίτιδας Β), το μεγαλύτερο ποσοστό Ηπατοκυτταρικών καρκίνων αποδίδεται στην ηπατίτιδα Β (60%), ενώ στον «ανεπτυγμένο δυτικό κόσμο» μόνο το 20%. Στον Δυτικό κόσμο η χρόνια ηπατίτιδα C είναι κύριος παράγοντας κινδύνου*. ('EASL-EORTC clinical practice guidelines: management of hepatocellular carcinoma.', 2012) Στην *Ασία και την υποσαχάρια Αφρική, πολλά άτομα μολύνονται με τον ιό της Ηπατίτιδας Β από τη γέννηση με χαμηλότερο ποσοστό επιβίωσης μετά τη θεραπεία* ('Hepatocellular carcinoma', 2021) αν αναπτύξουν Ηπατοκυτταρικό καρκίνωμα.

*Χαρακτηριστικά που μας πληροφορούν για το πόσο σοβαρή είναι η κατάσταση της ηπατικής νόσου (αριθμός αιμοπεταλίων < 100*103, παρουσία κιρσών οισοφάγου – οισοφαγικών αλλοιώσεων (esophageal varices)), αλλά και η μεγάλη ηλικία συσχετίζονται με την ανάπτυξη HCC σε ασθενείς με κίρρωση. Επιπλέον των προαναφερθέντων παραγόντων κινδύνου αναφέρουμε και την αυξημένη πυλαία πίεση, την Οροθετικότητα για το αντιγόνο του ιού της ηπατίτιδας Β (HBeAg), το υψηλό ιικό φορτίο, τον γονότυπο C (Genotype C), την δυσκαμψία του ήπατος, το κάπνισμα.* ('EASL-EORTC clinical practice guidelines: management of hepatocellular carcinoma.', 2012) Από (ZIMMERMANN, 2007) πληροφορούμαστε για κάποιες *ηπατικές βλάβες που έχουν εντοπιστεί μέχρι στιγμής και υποστηρίζεται από Κλινικά, Μοριακά, Μορφολογικά και Ανοσοϊστοχημικά δεδομένα ότι είναι πρόδρομες Ηπατοκυτταρικού καρκινώματος. Αυτές περιλαμβάνουν την ηπατική κυτταρική δυσπλασία (LCD) -το να ανευρίσκονται μορφολογικά άτυπα ηπατοκύτταρα σε ένα δομικά φυσιολογικό ήπαρ ή μέσα σε κιρρωτικά οζίδια- και ένα ευρύ φάσμα περισσότερο ή λιγότερο άτυπων οζιδιακών αλλοιώσεων που σχετίζονται ή δεν σχετίζονται με κίρρωση του ήπατος (δυσπλαστικές εστίες ή οζίδια*). (ZIMMERMANN, 2007) Σε τέτοια κύτταρα, χαρακτηριστικά όπως *αυξημένη αναλογία πυρηνικής πυκνότητας (nuclear density), καθαρή κυτταρική αλλαγή (clear cell change), μικροκυτταρική δυσπλασία (small cell dysplasia) και λιπώδης αλλαγή (fatty*

change), συνδέονται με υψηλό κίνδυνο εξέλιξης σε ηπατοκυτταρικού καρκίνου (HCC). (ZIMMERMANN, 2007)

Ως προς το γενετικό προφίλ των ασθενών έχει υπάρξει ένδειξη ότι ο *πολυμορφισμός 61*G στο EGF (rs4444903) είναι παράγοντας κινδύνου για ηπατοκαρκινογένεση ενώ το αλληλόμορφο EGF 61*A είναι προστατευτικός παράγοντας*. (Zhong et al., 2012) *Μερικά HCC φιλοξενούν μετάλλαξη στο μονοπάτι Wnt/β-κατενίνης.* (Colombo, Sangiovanni and Lencioni, 2018) Τα *επίπεδα του ελεύθερου κυτταρικού DNA (cell-free DNA, ctDNA) συσχετίζονται με το στάδιο της νόσου*. (Howell et al., 2019) Στην *ανάλυση υγρών βιοψιών που πραγματοποίησε η πιο πάνω ερευνητική ομάδα εντοπίστηκαν συχνότερα μεταλλάξεις στο ARID1A (11,7%), ακολουθούμενο από το CTNNB1 (7,8%) και το TP53 (7,8%). Ωστόσο το 71% των ασθενών είχαν επιπλέον μεταλλάξεις στο DNA του ιστού του Ηπατοκαρκινώματος HCC που δεν ανιχνεύθηκαν σε αντίστοιχο ctDNA*. (Howell et al., 2019) Επίσης το *GSK3β έχει βρεθεί ότι είναι θετικός ρυθμιστής της ογκογένεσης στα μοντέλα ηπατικής καρκινογένεσης ποντικών*. (Bosso and Al-Mulla, 2020, p. 3) Σύμφωνα με την OMIM, "*υπάρχουν σωματικές μεταλλάξεις σε αρκετά διαφορετικά γονίδια, ταυτοποιημένες στο ηπατοκυτταρικό καρκίνωμα (HCC) και το ηπατοβλάστωμα, όπως πχ. η TP53 (191170), η MET (164860), η CTNNB1 (116806), η PIK3CA (171834), η AXIN1 (603816) και η APC (611731). Επίσης, το οικογενειακό ηπατικό αδένωμα συνδέεται μερικές φορές με το ηπατοκυτταρικό καρκίνωμα.*" (Ada Hamosh, George E. Tiller, Cassandra L. Kniffin, Paul J. Converse, Victor A. McKusick, John A. Phillips, III, 2020)

Η ερευνητική ομάδα των (Karageorgos et al., 2017) μελέτησε την *αιτιολογική εξέλιξη των περιστατικών κίρρωσης και ηπατοκυτταρικού καρκινώματος, σε διάστημα 25 ετών (1990-2014) στην Κρήτη. Μελέτησαν 812 περιπτώσεις κίρρωσης (561 άνδρες, διάμεση ηλικία 69 έτη) και 321 περιπτώσεις ηπατοκυτταρικού καρκινώματος (234 άνδρες, διάμεση ηλικία 70 ετών). Υπήρξε αύξηση στη συχνότητα εμφάνισης ηπατοκυτταρικού καρκινώματος από 2004-2014, ενώ η συχνότητα της κίρρωσης έμεινε σταθερή. Παρατηρήθηκε τετραπλάσια μείωση στη συχνότητα εμφάνισης κίρρωσης λόγω ηπατίτιδας C, (από την πρώτη στην τρίτη θέση ως παράγοντας κινδύνου για κίρρωση). Το αλκοόλ έγινε ο πρώτος παράγοντας κινδύνου για κίρρωση (1990-94: 36,1%, 2010-14: 52,3%) και καρκίνωμα. Η πιο απότομη αύξηση στη συχνότητα εμφάνισης κίρρωσης και καρκινώματος ήταν λόγω της μη αλκοολικής λιπώδους νόσου του ήπατος.* (Karageorgos et al., 2017) Η εικόνα που ακολουθεί προέρχεται από την μελέτη των (Karageorgos et al., 2017).

Στους άξονες ψ βλέπουμε την εμφάνιση νέων περιστατικών ανά 100000 άτομα του πληθυσμού. Στους άξονες χ βλέπουμε την χρονική εξέλιξη. Όπου HCV αφορά την Ηπατίτιδα C, όπου HBV αφορά Ηπατίτιδα B, όπου ALC αφορά σχέτιση με αλκοόλ, όπου NASH αφορά μη αλκοολική στεατοηπατίτιδα και όπου ALC+VIR αφορά σχέση με αλκοόλ και ιογενή ηπατίτιδα.

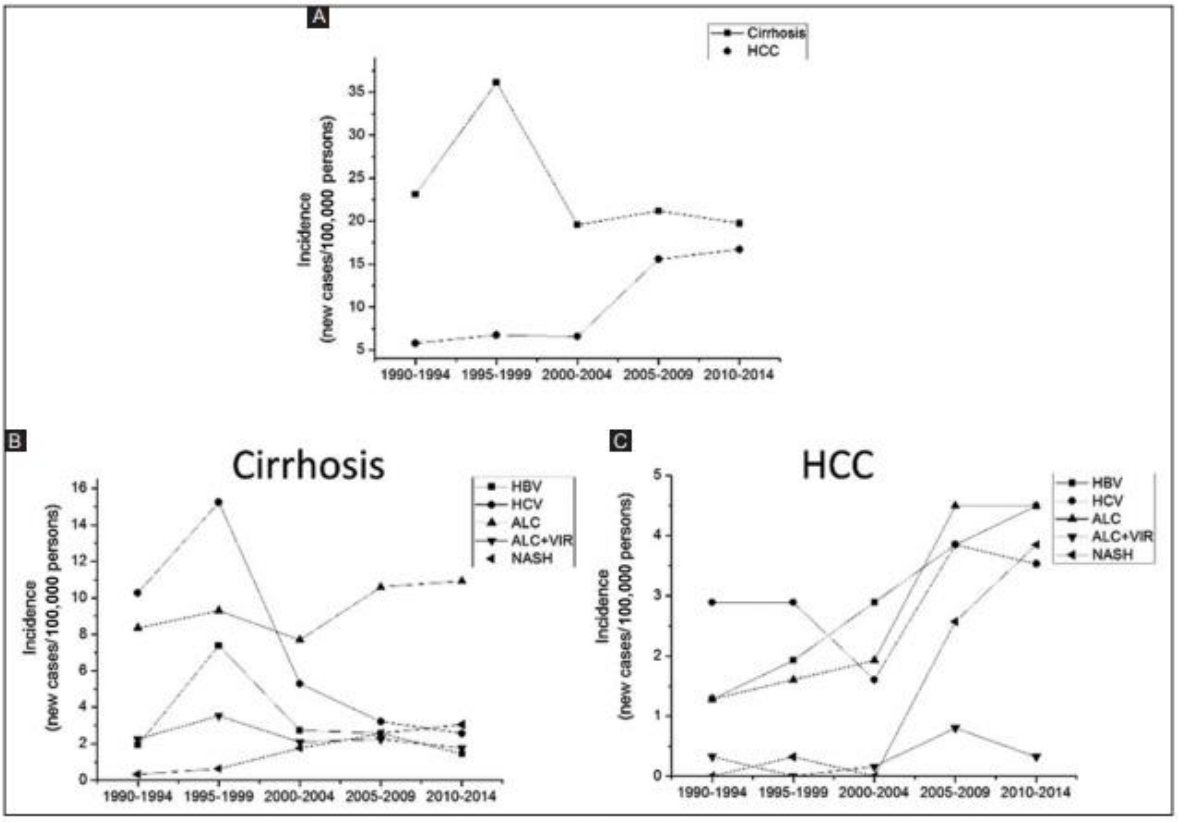


Figure 1 (A) Overall incidence of cirrhosis and hepatocellular carcinoma (HCC) over 25 years. (B) Etiology-based comparison of the changes in the incidence of new cases of cirrhosis. (C) Etiology-based comparison of the changes in the incidence of new cases of HCC over the 25 years of the study

Υπάρχει ανάγκη για να αποκτηθεί καλύτερη γνώση του μεταβολισμού των καρκινικών κυττάρων, της επιγενετικής του Ηπατοκυτταρικού καρκινώματος, των βλαστικών κυττάρων του Ηπατοκυτταρικού καρκίνου, αλλά και της ανοσολογίας του και του ηπατικού περιβάλλοντος, ώστε να ανοίξουν νέοι δρόμοι για καλύτερη θεραπευτική παρέμβαση. Πιθανόν και η περαιτέρω διερεύνηση του ρόλου του μικροβιώματος του εντέρου και του άξονα εντέρου-ήπατος στην ηπατοκαρκινογένεση να συμβάλει στην διάγνωση και θεραπεία του ηπατοκυτταρικού καρκίνου (Martínez-Chantar, Avila and Lu, 2020). Το σύνολο δεδομένων που χρησιμοποιήσαμε θα μπορούσαμε να πούμε ότι προσεγγίζει το θέμα από την πλευρά του Ηπατικού περιβάλλοντος παρά από τις υπόλοιπες.

Υπάρχουν και περιπτώσεις που θεωρείται πως υπάρχει κληρονομική προδιάθεση για Ηπατοκυτταρικό καρκίνωμα. Σύμφωνα με την OMIM *το ηπατοβλάστωμα αποτελεί το 1 έως 2% των κακοηθών νεοπλασμάτων της παιδικής ηλικίας, που συμβαίνουν συχνότερα κάτω των 3 ετών και έχει περιγραφεί σε αδέλφια. Πιστεύεται ότι προέρχεται από μη διαφοροποιημένα ηπατοκύτταρα. Επιπλέον, έχει καταγραφεί πρωτογενής καρκίνος του ήπατος σε 3 αδέλφια χωρίς αναγνωρισμένη προϋπάρχουσα ηπατική νόσο. Έχουν επίσης καταγραφεί και 2 ενήλικα αδέλφια που πέθαναν από πρωτοπαθές ηπατοκυτταρικό καρκίνωμα και είχαν μικροδονική κίρρωση με χαρακτηριστικά υποξείας προοδευτικής ιογενούς ηπατίτιδας. Το*

αντιγόνο της Αυστραλίας βρέθηκε στον αδελφό στον οποίο έγινε σχετική εξέταση. Ο πατέρας τους είχε πεθάνει πολύ νωρίτερα από ηπατοκυτταρικό καρκίνωμα. Υπάρχει και περιγραφή του καρκίνου του ήπατος ως επιπλοκή της ηπατίτιδας γιγαντιαίων κυττάρων (βρεφικής ηλικίας). Το οικογενειακό καρκίνωμα των ηπατικών κυττάρων μπορεί επίσης να εξηγείται λόγω τυροσιναιμίας (276700).” (Ada Hamosh, George E. Tiller, Cassandra L. Kniffin, Paul J. Converse, Victor A. McKusick, John A. Phillips, III, 2020)

Οι περισσότερες περιπτώσεις Ηπατοκυτταρικού καρκινώματος (HCC) εμφανίζονται σε άτομα που έχουν ήδη σημεία και συμπτώματα χρόνιας ηπατικής νόσου. Μπορεί να εμφανιστούν είτε με επιδείνωση των συμπτωμάτων είτε μπορεί να λείπουν συμπτώματα τη στιγμή της ανίχνευσης του καρκίνου. Συμπτώματα που σχετίζονται περισσότερο με την ηπατική νόσο περιλαμβάνουν ίκτερο, κοιλιακό πρήξιμο λόγω υγρού στην κοιλιακή κοιλότητα, εύκολο μώλωπα από ανωμαλίες πήξης του αίματος, απώλεια όρεξης, ακούσια απώλεια βάρους, κοιλιακό άλγος, ναυτία, έμετο, ή αίσθημα κόπωσης. (‘Hepatocellular carcinoma’, 2021)

Οι παρεμβάσεις πρόληψης του Ηπατοκυτταρτικού καρκίνου περιλαμβάνουν τον εμβολιασμό για τον ιό της ηπατίτιδας Β (HBV) σε ενδημικές περιοχές, την εκρίζωση του ιό της ηπατίτιδας C (HCV) με αντιιικά φάρμακα άμεσης δράσης, την προώθηση υγιεινής διατροφής και μείωσης βάρους, τη βελτίωση του ελέγχου του διαβήτη και την αποφυγή υπερκατανάλωσης αλκοόλ. (Martínez-Chantar, Avila and Lu, 2020)

Η θεραπεία και η πρόγνωση του Ηπατοκυτταρικού καρκινώματος (HCC) ποικίλλουν ανάλογα με τις ιδιαιτερότητες της ιστολογίας του όγκου, το μέγεθος, το πόσο έχει εξαπλωθεί ο καρκίνος και τη συνολική υγεία του ασθενούς. (‘Hepatocellular carcinoma’, 2021)

Σύμφωνα με τους (Liu et al., 2016) *ανεξάρτητοι προγνωστικοί παράγοντες δυσμενούς έκβασης ασθενών με Ηπατοκυτταρικό καρκίνωμα (στους 17 μήνες) ήταν η λευκωματίνη ορού <3,5 g/dl, η χολερυθρίνη ⩾1 mg/dl, η κρεατινίνη ⩾1 mg/dl, η άλφα-εμβρυϊκή πρωτεΐνη ⩾20 ng/ml, η αλκαλική φωσφατάση ⩾200 IU/L, η πολλαπλή παρουσία οζιδίων όγκου, το μέγιστο μέγεθος όγκου >5 cm, η παρουσία αγγειακής διήθησης, η παρουσία εξωηπατικής μετάστασης και η κακή κατάσταση απόδοσης (όλα p <0,001). Σημαντικές διαφορές στην επιβίωση βρέθηκαν σε όλα τα στάδια των 11 συστημάτων κατηγοριοποίησης του Ηπατοκυτταρικού καρκινώματος εκτός από το σταδίο IV και V της «κατηγοριοποίησης Χονγκ Κονγκ για τον Καρκίνο του ήπατος», την «Ολοκληρωμένη Σταδιοποίηση της Ιαπωνίας» (για σκορ 4 και 5) και το «σκορ του Τόκιο» (5 έως 8).*

4. Περιγραφή του χρησιμοποιηθέντος συνόλου δεδομένων ασθενών με Ηπατοκυτταρικό καρκίνωμα (HCC)

Το σύνολο δεδομένων που χρησιμοποιήθηκε στην παρούσα εργασία είναι κατατεθημένο στο διαδικτυακό αποθετήριο δεδομένων του UCI με επισήμανση ότι ενδείκνυται για Εργασία Ταξινόμησης (*Classification*) και ένδειξη ότι αφορά πολλές μεταβλητές (Multivariate). Σημειώνεται πως οι μεταβλητές που περιέχει είναι είτε ακέραιοι είτε πραγματικοί αριθμοί. Τα γνωρίσματα-μεταβλητές (στήλες) κάθε καταχώρησης είναι 49 και το σύνολο των καταχωρήσεων (Γραμμών) είναι 165. Η συλλογή και δωρεά του σετ έγινε στις 2017-11-29 από τα εξής μέλη της ακαδημαϊκής κοινότητας: Miriam Seoane Santos, Pedro Henriques Abreu, Armando Carvalho και Adélia Simão". (Miriam Seoane Santos, Pedro Henriques Abreu, Armando Carvalho, Adélia Simão, 2017) (Santos et al., 2015)

Άλλες παρατηρήσεις που αφορούν τα δεδομένα είναι οι εξής:

Το σύνολο δεδομένων διαθέτει 23 ποσοτικά και 26 ποιοτικά βιολογικά-κλινικά χαρακτηριστικά/μεταβλητές (στήλες) που σχετίζονται με την κλινική εικόνα των ασθενών. Οι ασθενείς προέρχονται από Πανεπιστημιακό Νοσοκομείο της Πορτογαλίας. Αυτά τα χαρακτηριστικά είναι καταγεγραμένα για τους ασθενείς που περιλαμβάνονται στο σύνολο δεδομένων. Πρόκειται για *δημογραφικά χαρακτηριστικά, παράγοντες κινδύνου, εργαστηριακές μετρήσεις και πληροφορίες για την συνολική επιβίωση των ασθενών που είχαν διαγνωσθεί με HCC*. (Miriam Seoane Santos, Pedro Henriques Abreu, Armando Carvalho, Adélia Simão, 2017) Τα συγκεκριμένα χαρακτηριστικά επιλέχθηκαν από την ερευνητική ομάδα που δώρησε το σύνολο δεδομένων που εξετάζουμε, σύμφωνα με τις Οδηγίες Κλινικών Πρακτικών της EASL-EORTC (Ευρωπαϊκός Σύνδεσμος για τη Μελέτη του Ήπατος - Ευρωπαϊκός Οργανισμός Έρευνας και Θεραπείας του Καρκίνου - European Association for the Study of the Liver - European Organisation for Research and Treatment of Cancer). Δεν αναφέρεται η θεραπευτική πρακτική που ακολουθήθηκε σε έκαστο ασθενή, παράγοντας που αν ήταν γνωστός πιθανόν να ήταν επίσης χρήσιμος στην ανάλυση σχετικά με την επιβίωση.

Περιέχεται πληροφορία και για την έκβαση των ασθενών. Η *μεταβλητή-στόχος είναι η επιβίωση στο 1 έτος. Η κωδικοποίηση ήταν η εξής: 0 (πεθαίνει) και 1 (ζει). Είναι σημαντικό να αναφερθεί ότι υπάρχει ανισορροπία στην Κλάση αυτή (63 καταγραφές αφορούν την κατηγορία 0 και 102 την κατηγορία 1)*. (Miriam Seoane Santos, Pedro Henriques Abreu, Armando Carvalho, Adélia Simão, 2017) (Santos et al., 2015)

Επίσης, δεν είναι πλήρεις οι καταγραφές όλων των ασθενών. Το 10,22% των δεδομένων λείπουν.

Σε αυτό το σύνολο δεδομένων πραγματοποιήθηκε Περιγραφική Ανάλυση, καθώς και Προγνωστική Ανάλυση. Στην ενότητα Αποτελέσματα φαίνονται όλα τα σχετικά γραφήματα που προέκυψαν από τις ανωτέρω αναλύσεις, καθώς και σύντομος σχολιασμός τους, ο οποίος συνεχίζεται και στην Συζήτηση.

5. Υλικά - Μεθοδολογία

5.1 Υλικά & Λογισμικό

Το σύνολο των εργασιών έγινε στον προσωπικό υπολογιστή SAMSUNG [NP300E5C-A01GR] με λειτουργικό σύστημα Windows ‘08, της εκπονούσας την διπλωματική, μιας και ο όγκος των δεδομένων ήταν μικρός, με συνέπεια να μην είναι αναγκαία η εκμετάλευση των επεξεργαστικών πλεονεκτημάτων άλλων -ταχύτερων- συστημάτων στην πραγματοποίηση των υπολογισμών. Λόγω βλάβης που παρουσίασε ο ανωτέρω υπολογιστής, η κατασκευή κάποιων εκ των διαγραμμάτων με τα αποτελέσματα πραγματοποιήθηκε σε άλλον προσωπικό υπολογιστή τύπου laptop ACER [Aspire 5742 series, MODEL NO PEW71] με λειτουργικό σύστημα Manjaro linux 21.0.6.

Για την ανάλυση των δεδομένων χρησιμοποιήθηκαν εργαλεία και βιβλιοθήκες της γλώσσας προγραμματισμού Python 3 (Van Rossum and Drake, 2009). Συγκεκριμένα, εκτός των βασικών βιβλιοθηκών της Python χρησιμοποιήθηκαν οι εξής: os, warnings, math, numpy (Harris et al., 2020), pandas (McKinney and others, 2010), sklearn (Pedregosa et al., 2011), imblearn (Lemaître et al., 2017) , matplotlib (Hunter, 2007), seaborn (Waskom et al., 2017). Για την ευκολότερη εγκατάστασή της Python σε λειτουργικό σύστημα των Windows 8.0, όπου και πραγματοποιήθηκε το σύνολο των εργασιών, χρησιμοποιήθηκε η διανομή ανοιχτού κώδικα της γλώσσας προγραμματισμού Python 3 που περιέχεται στο Anaconda, μια διανομή που προορίζεται και για επιστημονικούς υπολογισμούς (*Anaconda Software Distribution*, 2020). Η διαδικασία της στατιστικής ανάλυσης των δεδομένων έγινε στο διαδραστικό υπολογιστικό περιβάλλον που βασίζεται στον Ιστό, Jupyter Notebook (Kluyver et al., 2016), που συνδυάζει την δυνατότητα υλοποίησης αλγορίθμων σε Python και την δημιουργία σημειώσεων σε Markdown. (Matt Cone, n.d.)

5.2 Προεπεξεργασία Δεδομένων (Preprocessing)

Η διαδικασία Προεπεξεργασίας των δεδομένων περιγράφεται ακολούθως.

5.2.1 Συμπλήρωση Κενών Τιμών στα Δεδομένα (Missing Data Imputation)

Κάθε χαρακτηριστικό/γνώρισμα/μεταβλητή (στήλη) ελέγθηκε για ελλείψεις τιμών στις καταχωρήσεις του. Στην παρούσα ανάλυση επιλέχθηκε να συμπληρωθούν οι τιμές που λείπουν με τρεις διαφορετικές προσεγγίσεις:

1. Η ***πρώτη τακτική*** ήταν με χρήση μίας πεπερασμένης τιμής, η οποία βρίσκεται αρκετές τάξεις μεγέθους έξω από το εύρος των τιμών του εκάστοτε γνωρίσματος. Η επιλογή της ακριβούς τιμής παρότι πληρούσε τα πιο πάνω χαρακτηριστικά είναι αυθαίρετη. Αυτή η προσέγγιση έγινε χωρίς να υπάρχει κάποια προσδοκία για τα αποτελέσματά της και παρατείθεται μόνο για λόγους σύγκρισης. Στο εξής θα αναφερόμαστε σε αυτή για συντομία ως «εμπειρική συμπλήρωση».

Για τις μεταβλητές που αντιστοιχούσαν σε ποιοτικά, μη διατακτικά γνωρίσματα, που συμβολίζονται με ακέραιες τιμές πχ. 0 και 1, επιλέχθηκε η τιμή ***-99999*** για συμπλήρωση των κενών. Τέτοια γνωρίσματα ήταν:

το φύλο ("Gender"), η παρουσία συμπτωμάτων ("Symptoms"), η κατανάλωση αλκοόλ ("Alcohol"), η παρουσία του επιφανειακού αντιγόνου της Ηπατίτιδας Β ("HbsAg"), ή του αντιγόνου e της Ηπατίτιδας Β ("HbeAg"), ή του βασικού αντισώματος για την Ηπατίτιδα Β ("HbcAb"), ή του αντισώματος για τον ιό της Ηπατίτιδας C ("HCVAb"), η ύπαρξη κίρρωσης ("Cirrhosis"), η σχέση με ενδημική χώρα ("Endemic"), το κάπνισμα ("Smoking"), η ύπαρξη διαβήτη στον ασθενή ("Diabetes"), η ύπαρξη παχυσαρκίας ("Obesity"), η ύπαρξη αιμοχρωμάτωσης ("Hemochro"), η ύπαρξη αρτηριακής υπέρτασης ("AHT"), η χρόνια νεφρική ανεπάρκεια ("CRI"), η μόλυνση με τον ιό HIV ("HIV"), η μη-αλκοολική στεατοηπατίτιδα ("NASH"), οι οισοφαγικές αλλοιώσεις ("Varices"), η Σπληνομεγαλία ("Spleno"), η ύπαρξη υπέρτασης πυλαίας φλέβας ("PHT"), η θρόμβωση πυλαίας φλέβας ("PVT"), η μετάσταση ("Metastasis"), και το ακτινολογικό σήμα ("Hallmark").

Τα γνωρίσματα που χρειάστηκαν συμπλήρωση, καθώς τους έλλειπαν τιμές ήταν τα: "Symptoms","HbsAg", "HbeAg", "HbcAb", "HCVAb", "Endemic", "Smoking", "Diabetes", "Obesity", "Hemochro", "AHT", "CRI", "HIV", "NASH", "Varices", "Spleno", "PHT", "PVT", "Metastasis", "Hallmark".

Στα ποιοτικά, διατακτικά γνωρίσματα, που ήταν: η κατάσταση απόκρισης ("PS"), ο βαθμός εγκεφαλοπάθειας ("Encephalopathy") και ο βαθμός ανάπτυξης ασκιτών (“Ascites“) ακολουθήθηκε η ίδια ακριβώς στρατηγική και χρησιμοποιήθηκε η τιμή -99999 για να καλύψει τις κενές τιμές. Τα γνωρίσματα που χρειάστηκαν συμπλήρωση, καθώς τους έλλειπαν τιμές, ήταν τα: "Encephalopathy" και “Ascites“.

Στα διακριτά τακτικά γνωρίσματα, όπως η Ηλικία (“Age”), και ο αριθμός των Οζιδίων (“Nodules”), ακολουθήθηκε επίσης η ίδια ακριβώς στρατηγική (επιλέχθηκε η τιμή -99999). Στην συγκεκριμένη κατηγορία χρειάστηκε να συμπληρωθούν τιμές μόνο στο γνώρισμα “Nodules”, καθώς μόνο εκεί έλλειπαν.

Το ίδιο συνέβη και με τα περισσότερα συνεχή τακτικά γνωρίσματα: Γραμμάρια αλκοόλ/μέρα ("Grams/day"), Πακέτα τσιγάρων/έτος ("Packs/year"), Διεθνής Κανονικοποιημένος Λόγος ("INR"), α-εμβρυϊκή πρωτεΐνη (ng/mL) ("AFP"), Αιμοσφαιρίνη (g/dL) ("Hemoglobin"), Μέσος όγκος του όγκου (fl) ("MCV"), Λευκοκύτταρα (G/L) ("Leucocytes"), Αιμοπετάλια (G/L) ("Platelets"), Αλβουμίνη (mg/dL) ("Albumin"), Συνολική Χολεριθρίνη (mg/dL) ("Total Bil"), Τρανσαμινάση αλανίνης (U/L) ("ALT"), Τρανσαμινάση ασπαρτικού (U/L) ("AST"), γ-γλουταμυλ-τρανσφεράση (U/L) ("GGT"), αλκαλική φωσφατάση (U/L) ("ALP"), Συνολικές Πρωτεΐνες (g/dL) ("TP"), Κρεατινίνη (mg/dL) ("Creatinine"), Κύρια διάσταση του οζιδίου (cm) ("Major Dim"), Άμεση χολερυθρίνη (mg/dL) ("Dir. Bil"), Σίδηρος (mcg/dL) ("Iron"), Κορερσμός Οξυγόνου (%) ("Sat"), Φεριτίνη (ng/mL) ("Ferritin"). Συμπλήρωση τιμών (επιλέχθηκε η τιμή -99999) χρειάστηκαν τα: "Grams/day", "Packs/year", "INR", "Hemoglobin", "MCV", "Leucocytes", "Albumin", "Total Bil", "ALT", "AST", "GGT", "ALP", "TP", "Creatinine", "Major Dim", "Dir. Bil", "Iron", "Sat", "Ferritin".

Η παρούσα κατηγορία διαφέρει καθότι υπήρξαν δύο εξαιρέσεις. Για το γνώρισμα α-εμβρυϊκή πρωτεΐνη (ng/mL) η τιμή που χρησιμοποιήθηκε για να καλύψει τα κενά ήταν η ***-1000000000*** και στα Αιμοπετάλια: ***-5000000***. Οι τιμές αυτές επιλέχθηκαν με κριτήριο να είναι αρκετές τάξεις μεγέθους εκτός του εύρους τιμών του εκάστοτε γνωρίσματος. Ωστόσο η ακριβής τιμή και εδώ αποφασίστηκε επίσης αυθαίρετα.

2. Η ***δεύτερη τακτική*** ακολουθεί την ίδια φιλοσοφία με την πρώτη, με πιο φορμαλιστικό και επαναλήψιμο τρόπο. Γίνεται χρήση της συνάρτησης ***y= - R*(log10(R)+100),*** όπου R: εύρος τιμών για έκαστο γνώρισμα, για να βρεθεί η τιμή που θα συμπληρώσει τα κενά ανά γνώρισμα. Η συγκεκριμένη συνάρτηση επιλέχθηκε λόγω του ότι αποδίδει τιμές που βρίσκονται αρκετές τάξεις μεγέθους κάτω από το εύρος τιμών του γνωρίσματος. Προτιμήθηκε έναντι της y= - log10(R) γιατί ήταν επιθυμητή μία αναβαθμονόμηση στην λογαριθμική κλίμακα. Επίσης η συνάρτηση που επιλέχθηκε υπερείχε σε σχέση με την y= - R*(log10(R)) λόγω του ότι διασφάλιζε ότι θα αποδωθεί τιμή αρκετών τάξεων μεγέθους κάτω από το εύρος, ακόμη και αν το R ήταν ίσο με 1. Για R=1 (γεγονός αρκετά συχνό στο συγκεκριμένο σύνολο δεδομένων),

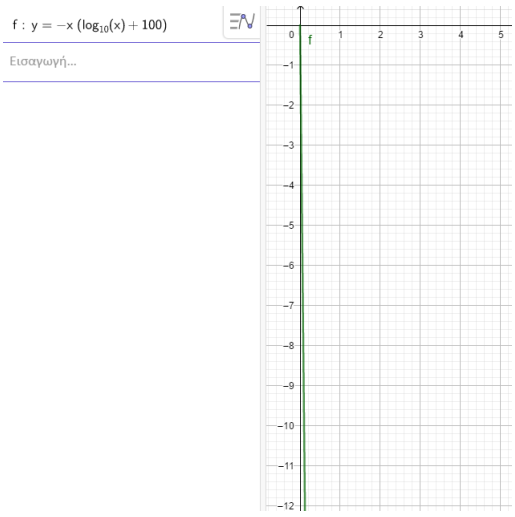
   η y= - R*(log10(R))

   θα έδινε y=0. Ακολουθεί στο Σχήμα 1 η γραφική παράσταση της

   y= - R*(log10(R)+100),

   που έγινε με το πρόγραμμα GeoGebra (Hohenwarter et al., 2013). Στο εξής θα αναφερόμαστε σε αυτή για συντομία ως «συμπλήρωση με την δοκιμαζόμενη συνάρτηση».

Σχήμα 1: Γραφική παράστασης εξίσωσης y = -R(log10(R)+100)*



3. Η ***τρίτη τακτική*** ήταν να συμπληρωθούν τα κενά μέσω Επαναληπτικού καταλογισμού (Iterative imputation), με την μέθοδο ***IterativeImputer()***, της βιλβιοθήκης της Python sklearn. Η μέθοδος αυτή μοντελοποιεί κάθε χαρακτηριστικό με τιμές που λείπουν ως συνάρτηση των άλλων χαρακτηριστικών και χρησιμοποιεί αυτήν την εκτίμηση για συμπλήρωση τιμών. Αυτό γίνεται επαναληπτικά, με αλλαγή κάθε φορά στο ποιά θα είναι η μεταβλητή της οποίας οι αρχικά κενές τιμές θα συμπληρώνονται. Σε κάθε βήμα, μια στήλη χαρακτηριστικών ορίζεται ως έξοδος y και οι άλλες στήλες χαρακτηριστικών αντιμετωπίζονται ως είσοδοι X. Εκπαιδεύεται μοντέλο γραμμικής παλλινδρόμησης (X, y), με είσοδο τα γνωστά y. Στη συνέχεια, το μοντέλο παλινδρόμησης χρησιμοποιείται για να προβλέψει τις τιμές που λείπουν από το y. Αυτό γίνεται για κάθε χαρακτηριστικό του συνόλου δεδομένων, με επαναληπτικό τρόπο και στη συνέχεια επαναλαμβάνεται για γύρους συμπλήρωσης (max_iter) κατά τη διάρκεια των οποίων η διαφορά μεταξύ τιμών που χρησιμοποιήθηκαν για συμπλήρωση στον κάθε προηγούμενο γύρο σε σχέση με τις τιμές που χρησιμοποιήθηκαν στον κάθε τρέχοντα γύρο συγκλίνει στο 0. Εν τέλει τα αποτελέσματα του τελικού γύρου συμπλήρωσης επιστρέφονται. Στο εξής θα αναφερόμαστε σε αυτή για συντομία ως «συμπλήρωση με Iterative imputation».

Κατόπιν, έγιναν κατάλληλες ενέργειες ώστε οι τύποι δεδομένων κάθε γνωρίσματος να είναι οι αναμενόμενοι βάσει της περιγραφής που συνόδευε το σύνολο δεδομένων (ακέραιοι ή δεκαδικοί, integer ή float). Επιπλέον έγινε ανάγνωση των δεδομένων, σαν μέτρο επιπλέον ελέγχου. Σε αυτό βοήθησε και το γεγονός ότι οι καταχωρήσεις ήταν λίγες και το σετ δεδομένων μικρό, διαφορετικά κάτι τέτοιο δεν θα ήταν εφικτό. Τα δεδομένα επιθεωρήθηκαν μηχανικά και γραφικά για την εύρεση του εύρους τους, για ύπαρξη λογικών ή τυπογραφικών λαθών, πχ. Ηλικία: 200. Τα γραφήματα που αφορούν την Περιγραφική Ανάλυση (Descriptive analysis) του συνόλου δεδομένων παρουσιάζονται στο ***Παράρτημα Ι***.

5.2.2 Εξισορρόπηση κλάσεων του συνόλου δεδομένων

Στην συνέχεια είχαμε τις ακόλουθες διακριτές προσεγγίσεις ως προς την εξισορρόπηση της εκπροσώπησης των κλάσεων στο σύνολο δεδομένων. (Costa et al., 2020)

I. Στην πρώτη προσέγγιση **δεν έγινε εξισορρόπηση**. Στο εξής θα αναφερόμαστε σε αυτή για συντομία ως «χωρίς εξισορρόπηση».

II. Στην δεύτερη προσέγγιση έγινε **υποδειγματοληψία** της πλειοψηφούσας κλάσης, ώστε το τελικό υποσύνολο να περιέχει 50% της μειοψηφούσας κλάσης και 50% της πλειοψηφούσας. Στο εξής θα αναφερόμαστε σε αυτή για συντομία ως «υποδειγματοληψία» ή «undersampling».

III. Στην τρίτη προσέγγιση έγινε χρήση της **παραμέτρου class_weights** κατά την εκπαίδευση των μοντέλων ταξινόμησης. Έκαστος συνδυασμός από βάρη επιλέχθηκε ανάμεσα σε άλλους 20 δυνατούς συνδυασμούς, έπειτα από grid search που έγινε στο διάστημα (0.00001, 0.99999), με κριτήριο την βέλτιστη τιμή του Σκορ-f1 (f1-score) που προέκυπτε από κάθε συνδυασμό βαρών ανά ταξινομική μέθοδο και training set. Η παράμετρος class_weights δεν ήταν διαθέσιμη σε όλες τις μεθόδους ταξινόμησης που δοκιμάστηκαν, συνεπώς δοκιμάστηκε μόνο σε όσες μεθόδους υπήρχε σαν επιλογή. Αυτές ήταν οι εξής: Support Vector Machines, Logistic Regression, Decision Trees, Random Forest. Στο εξής θα αναφερόμαστε σε αυτή για συντομία ως «class_weights».

IV. Στην τέταρτη προσέγγιση έγινε **υπερδειγματοληψία** της μειοψηφούσας κλάσης (στο εξής αναφέρεται ως *oversampling* για οικονομία χώρου), ώστε να φτάσει στο 50% των δεδομένων, **μέσω της μεθόδου SMOTE**. (Chawla et al., 2002) Στην συγκεκριμένη στρατηγική δημιουργούνται συνθετικά δεδομένα που αντιστοιχούν στην μειοψηφούσα κλάση, έως ότου τα ποσοστά των δύο κλάσεων γίνουν ίσα. Στο εξής θα αναφερόμαστε σε αυτή για συντομία ως «υπερδειγματοληψία SMOTE» ή «SMOTE oversampling».

V. Στην πέμπτη προσέγγιση έγινε **τυχαία υπερδειγματοληψία** (random oversampling) της μειοψηφούσας κλάσης, ώστε να φτάσει στο 50% των δεδομένων. Σε αυτή την στρατηγική διπλασιάζονται τυχαία κάποιες καταγραφές που ανήκουν στην μειοψηφούσα κλάση έως ότου τα ποσοστά των δύο κλάσεων γίνουν ίσα. Στο εξής θα αναφερόμαστε σε αυτή για συντομία ως «τυχαία υπερδειγματοληψία» ή «random oversampling».

Τέλος, για κάθε μη διατακτικό, ποιοτικό γνώρισμα που κωδικοποιείται με ακέραιο αριθμό, δημιουργήθηκαν κάποια ακόμη βοηθητικά γνώρισματα. Σκοπός ήταν η **προσομοίωση διανυσματικής αναπαράστασης του γνωρίσματος**, σε δυαδικό σύστημα και η αποφυγή τροφοδοσίας των μοντέλων μηχανικής μάθησης με ψευδή στοιχεία (artifacts) που οφείλονται σε αριθμητικές σχέσεις των τιμών που αντιπροσωπεύουν την κάθε κατηγορία. Το πλήθος των βοηθητικών γνωρισμάτων που αντιπροσώπευαν κάθε τέτοιο γνώρισμα ήταν ίσο με το πλήθος των δυνατών τιμών που μπορεί να πάρει το γνώρισμα.

Στην προεπεξεργασία της υποδειγματοληψίας (Undersampling) αναμένεται να χαθεί πληροφορία. Ο λόγος είναι ότι απορρίπτονται στοχαστικά καταγραφές που ανήκουν στην πλειοψηφούσα κλάση. Είναι πιθανό η κατανομή των δεδομένων της (πρώην)

πλειοψηφούσας κλάσης να μην αντιπροσωπεύει την αρχική της κατανομή στο προκύπτον από την διαδικασία σύνολο δεδομένων. Αν λ.χ. τύχει να αλλάξει η δομή της κατανομής μπορεί να χαθεί μέρος της ποικιλομορφίας που χαρακτηρίζει το υπόβαθρο της κλάσης.

Στην προεπεξεργασία υπερδειγματοληψίας SMOTE (SMOTE-oversampling) αναμένεται να γίνει είσοδος θορύβου στα δεδομένα, αλλά και αύξηση της αυθαιρεσίας λόγω του ότι τα συνθετικά δεδομένα είναι πλασματικά και όχι πραγματικές καταγραφές. Ωστόσο η σύνθεση δεν γίνεται εντελώς αυθαίρετα, αλλά επιχειρεί να βρίσκεται κοντά στις πραγματικές καταγραφές, σύμφωνα με την μέθοδο SMOTE. Η μέθοδος SMOTE επιλέγει τυχαία μια καταγραφή που ανήκει στην μειοψηφούσα κλάση τυχαία και βρίσκει τους k κοντινότερους γείτονες που ανήκουν στην ίδια κλάση. Επιλέγεται τυχαία ένας από τους κ κοντινότερους γείτονες β τυχαία και συνδέονται το α και β με ένα τμήμα γραμμής ορισμένο στον πολυδιάστατο χώρο των γνωρισμάτων (έχει τόσες διαστάσεις όσα και τα γνωρίσματα). Ο κυρτός συνδυασμός των δύο επιλεγμένων καταγραφών α και β είναι η νέα, συνθετική καταγραφή.

Μειονέκτημα της πορείας της τυχαίας υπερδειγματοληψίας (random-oversampling) αποτελλεί το γεγονός ότι διπλασιάζονται καταγραφές που ανήκουν στην μειοψηφική κλάση, με αποτέλεσμα να αυξάνεται η πιθανότητα να βρεθούν δύο πανομοιότυπες καταγραφές στο εκπαιδευτικό σετ και στο σετ ελέγχου. Αναμένεται αυξημένη πλασματική απόδωση του συνόλου των ταξινομητών για την συγκεκριμένη πορεία προεπεξεργασίας δεδομένων, λόγω της υπερπροσαρμογής (overfitting) που πιθανότατα θα γίνει.

5.3 Διαχωρισμός Δεδομένων, Σύνολο Εκπαίδευσης-Ελέγχου (Set Split)

Ο στοχαστικός διαχωρισμός του συνόλου δεδομένων σε μικρότερα σύνολα που προορίζονται για εκπαίδευση αξιοποιήθηκε ως εξής. Μέσω Στρωματοποιημένης (Stratified) διασταυρωμένης επικύρωσης (cross-validation) 10-μερών (10-fold) δημιουργήθηκαν 10 υποσύνολα των δεδομένων, με εκπροσώπηση κάθε κλάσης σε έκαστο αντίστοιχη με την συνολική εκπροσώπησή της στο σύνολο των δεδομένων. Πραγματοποιήθηκαν εκπαιδεύσεις μοντέλων μηχανικής μάθησης σε 9 υποσύνολα κάθε φορά, με το υπολειπόμενο υποσύνολο να χρησιμοποιείται σαν σύνολο ελέγχου. Η διαδικασία αυτή έγινε 10 φορές, με χρήση ενός διαφορετικού υποσυνόλου σαν σύνολο ελέγχου ανά επανάληψη και την χρήση των υπολοίπων σαν σύνολα εκπαίδευσης. Σε κάθε ένα σύνολο εκπαίδευσης περιέχονταν το 90% των δεδομένων και σε κάθε ένα σύνολο ελέγχου το 10%.

5.4 Εκπαίδευση Μοντέλων (Model Training)

Στο βήμα της εκπαίδευσης των Μοντέλων έγινε δοκιμή δημιουργίας διαφορετικών μοντέλων ταξινόμησης των δεδομένων. Οι μέθοδοι ταξινόμησης που χρησιμοποιήθηκαν (σε διάφορες παραλλαγές τους) ήταν οι εξής: Μηχανές Υποστήριξης Διανυσμάτων (Support Vector Machines) [Για μετάφραση του όρου βλ. (Σκούρα Αγγελική, ceid.upatras)], Λογιστική Παλλινδρόμηση (Logistic Regression), Δέντρο Απόφασης (Decision Tree) [Για μετάφραση του όρου βλ. (Σκούρα Αγγελική, ceid.upatras)], Τυχαίο Δάσος (Random Forest), Νευρωνικά Δίκτυα (Neural Networks), Κ-κοντινότεροι Γείτονες (K-Nearest Neighbours), Αφελείς Μπεϋσιανοί Κατηγοριοποιητές (Naive Bayes Classifiers: Bernoulli, Gaussian) [Για μετάφραση του όρου βλ. (ΒΑΣΙΛΙΚΗ ΠΑΠΑΘΑΝΑΣΙΟΥ, 2019)]

5.4.1 Μηχανές Υποστήριξης Διανυσμάτων (Support Vector Machines - SVM)

Χρησιμοποιήθηκαν οι προεπιλεγμένες παράμετροι SVC(kernel='rbf')
Ο πυρήνας rbf σε δύο δείγματα x και x΄, που αντιπροσωπεύονται ως διανύσματα χαρακτηριστικών σε κάποιο χώρο εισόδου, ορίζεται ως:

$K(x,x') = exp(-\gamma*||x-x'||^2)$,

Όπου: $||x-x'||^2$ *είναι η τετραγωνισμένη ευκλείδεια απόσταση μεταξύ των δύο διανυσμάτων χαρακτηριστικών,*

$\gamma = 1/(2*\sigma^2)$ *και*

σ: μια ελεύθερη παράμετρος.

Η τιμή του πυρήνα RBF μειώνεται με την απόσταση και κυμαίνεται μεταξύ 0 (στο όριο) και 1 (όταν x = x΄), συνεπώς ερμηνεύεται ως μέτρο ομοιότητας. Ο χώρος χαρακτηριστικών του πυρήνα έχει άπειρο αριθμό διαστάσεων. ("Radial basis function kernel," n.d.)

5.4.2 Λογιστική Παλινδρόμηση (Logistic regression, LR1-6)

Το λογιστικό μοντέλο (ή το μοντέλο logit) χρησιμοποιείται για να μοντελοποιήσει την πιθανότητα ύπαρξης μιας συγκεκριμένης κλάσης ή συμβάντος, όπως π.χ. 0 ή 1 (στην περίπτωσή μας). ("Logistic regression," n.d.)

Χρησιμοποιήθηκαν οι ακόλουθες παραλλαγές:

```
LogisticRegression(LogisticRegression(penalty='elasticnet',solver='saga',max_iter=10000, l1_ratio=0.5))
```

- *saga: Ο λύτης 'Saga' είναι μια παραλλαγή του λύτη 'Sag' και χρησιμοποιείται με l1 κανονικοποίηση. Είναι λύτης αρκετά αποδωτικός χρονικά και συνήθως χρησιμοποιείται για πολύ μεγάλα σύνολα δεδομένων.*
- *ποινή l1: υποστηρίζεται από τους λύτες 'liblinear', 'saga'*
- *l1_ratio=0.5*
  εφόσον ο λύτης είναι 'Saga' και η ποινή 'elasticnet', αυτή η παράμετρος μπορεί να προσφέρει περαιτέρω βελτιστοποίηση. Αν l1_ratio = 0, η ποινή είναι l2, αν l1_ratio = 1 η ποινή είναι l1, αν 0 <l1_ratio <1, η ποινή θα είναι ένας συνδυασμός l1 & l2 και ο λόγος l1_ratio θα καθορίσει το βάρος του l1 στο συνδυασμό. ("Logistic Regression Optimization Parameters Explained," n.d.)

```
LogisticRegression(penalty='l2',solver='saga', max_iter=10000)
```

- *l2 ποινή: υποστηρίζεται από τους λύτες 'cg', 'sag', 'saga', 'lbfgs'* ("Logistic Regression Optimization Parameters Explained," n.d.)

```
LogisticRegression(penalty='l2',solver='sag', max_iter=10000)
```

- sag: *Στοχαστική κάθοδος μέσης κλίσης (Stochastic Average Gradient Descent). Πιο αποτελεσματικός λύτης με μεγάλα σύνολα δεδομένων* ("Logistic Regression Optimization Parameters Explained," n.d.)

```
LogisticRegression(penalty='l2',solver='lbfgs', max_iter=10000)
```

- *lbfgs: περιορισμένης μνήμης 'BFGS'. Είναι λύτης που υπολογίζει μόνο μια προσέγγιση του 'Hessian' με βάση την κλίση που το καθιστά υπολογιστικά πιο αποτελεσματικό. Η χρήση της μνήμης είναι περιορισμένη σε σύγκριση με τα κανονικά bfgs, συνεπώς απορρίπτει τις προηγούμενες κλίσεις και συσσωρεύει μόνο νέες κλίσεις, λόγω περιορισμού μνήμης*. ("Logistic Regression Optimization Parameters Explained," n.d.)

```
LogisticRegression(penalty='l2',solver='lbfgs', max_iter=890)
```

- Επιλέχθηκε μικρότερος αριθμός επαναλήψεων για να αποφευχθεί η υπερπροσαμογή (overfitting) (με βάση τον αριθμό των επαναλήψεων που εξακολουθούσαν να δίνουν εξίσου καλή ακρίβεια με την προηγούμενη περίπτωση, για προεπεξεργασία (preprocessing) χωρίς εξισορρόπηση των δεδομένων και εμπειρική συμπλήρωση κενών (περίπτωση 1).

```
LogisticRegression(penalty='l2', solver='newton-cg', max_iter=100000)
```

- *newton-cg: Λύτης που υπολογίζει ρητά τον 'Hessian' και μπορεί να είναι υπολογιστικά ακριβός σε υψηλές διαστάσεις* ("Logistic Regression Optimization Parameters Explained," n.d.)

5.4.3 Δέντρα αποφάσεων (Decision trees, DT1-4)

Πρόκειται για ταξινομητή που αποδίδει καλά ακόμη και σε μη ισορροπημένα σύνολα δεδομένων. Ο διαχωρισμός των κλάδων γίνεται με βάση την ελαχιστοποίηση ενός κριτηρίου. Πρόκειται για τα κριτήρια 'gini', 'entropy' στην περίπτωσή μας.

*Το **Κριτήριο Gini** είναι μέτρο έλλειψης αμοιγίας. Αφορά το πόσο συχνά ένα τυχαία επιλεγμένο στοιχείο από το σύνολο θα χαρακτηριζόταν εσφαλμένα αν επισημαινόταν τυχαία, σύμφωνα με την κατανομή των αποδιδόμενων ετικετών κλάσεων στο υποσύνολο.* (Rahul Agarwal, 2019) *Ενδείκνυται για συνεχή χαρακτηριστικά και για ελαχιστοποίηση της λανθασμένης ταξινόμησης και τείνει να βρει τη μεγαλύτερη τάξη* (Gary Sieling, 2014) Η συνάρτηση υπολογισμού του είναι η: $I_G(n) = 1 - \sum_{i=1}^{J} (p_i)^2$, όπου

J: αριθμός των κλάσεων που υπάρχουν στον κόμβο,

p: η κατανομή της κλάσης στον κόμβο.

Τελικά επιλέγεται μία διάσπαση σε κλάδους τέτοια που να μας παρέχει τη χαμηλότερη σταθμισμένη έλλειψη αμοιγίας στους θυγατρικούς κόμβους. (Rahul Agarwal, 2019)

*Το **κριτήριο εντροπία (entropy**) είναι μέτρο της τυχαιότητας στο σύστημα* (Rahul Agarwal, 2019) *Ενδείκνυται για χαρακτηριστικά που εμφανίζονται σε τάξεις (π.χ. χρώματα) και για διερευνητική ανάλυση. Τείνει να βρει ομάδες τάξεων που αποτελούν το ~ 50% των δεδομένων, είναι λίγο πιο αργή στον υπολογισμό*. (Gary Sieling, 2014) Η συνάρτηση υπολογισμού της είναι η:

$Entropy = \sum_{i=1}^{C} -p_i * \log_2(p_i)$, όπου

C: αριθμός των κλάσεων που υπάρχουν στον κόμβο,

p: κατανομή της κλάσης στον κόμβο (Rahul Agarwal, 2019)

Δοκιμάστηκαν οι παραλλαγές:

`DecisionTreeClassifier(criterion='gini',max_leaf_nodes=10,random_state=0)`

- random_state=0
  Εάν δεν είναι επιθυμητή η τυχαιότητα σχετικά με τη δημιουργία του μοντέλου (για λόγους επαναληψιμότητας), επιλέγεται ένας αριθμός -στην περίπτωσή μας το 0- και το χρησιμοποιούμε σε όλα τα παραδείγματα, διαφορετικά τίθεται random_state= None (Pedregosa et al., 2011)
- max_leaf_nodes=10
  ο μέγιστος αριθμός φύλλων ανά κόμβο είναι ίσος με 10

`DecisionTreeClassifier(criterion='gini',max_leaf_nodes=11,random_state=0)`

- max_leaf_nodes=11

`DecisionTreeClassifier(criterion='gini',max_leaf_nodes=12,random_state=0)`

- max_leaf_nodes=12

`DecisionTreeClassifier(criterion=' entropy',max_leaf_nodes=6, random_state=0)`

- max_leaf_nodes=6

5.4.4 Τυχαία Δάση (Random Forest, RF1-2)

Πρόκειται για ταξινομητή που αποδίδει καλά ακόμη και σε μη ισορροπημένα σύνολα δεδομένων. Οι παράμετροι που υπάρχουν κι εδώ είναι αντίστοιχες των Δενδρών Αποφάσεων. Δοκιμάστηκαν οι παραλαγές:

```
RandomForestClassifier(criterion='entropy',max_leaf_nodes=2, bootstrap=True)

RandomForestClassifier(criterion='gini',max_leaf_nodes=3, bootstrap=True)
```

5.4.5 Νευρωνικά Δίκτυα (Neural Networks, NN1-10)

- *hidden_layer_sizes: Το ν-στό στοιχείο αντιπροσωπεύει τον αριθμό των νευρώνων στο ν-στό κρυφό στρώμα.*
- *activation='logistic': λογιστική σιγμοειδής συνάρτηση, επιστρέφει f(x)=1 / (1 + exp(-x))*
- *solver='lbfgs':*
  βελτιστοποιητής της οικογένειας των οιωνεί-Νευτώνιων μεθόδων
- *learning_rate='adaptive': διατηρεί τον ρυθμό μάθησης στο μοντέλο ταξινόμησης ίσο με τον αρχικό ('learning_rate_init') για όσο η απώλεια εκπαίδευσης συνεχίζει να μειώνεται. Κάθε φορά που δύο διαδοχικές εποχές αποτυγχάνουν να μειώσουν την απώλεια εκπαίδευσης για τουλάχιστον όσο το tol (ανεκτικότητα) ή αποτυγχάνουν να αυξήσουν το validation score για τουλάχιστον όσο το tol αν το 'early_stopping' είναι ενεργό, ο τρέχων ρυθμός μάθησης διαιρείται με 5.*
- *random_state=1: Καθορίζει τη δημιουργία τυχαίων αριθμών για βάρη και αρχικοποίηση μεροληψίας. Αν τεθεί ίσο με έναν int δίνει αναπαραγώγιμα αποτελέσματα σε πολλές κλήσεις της μεθόδου.*
- *max_iter=1000: ο μέγιστος αριθμός επαναλήψεων τέθηκε ίσος με 1000 και αυξήθηκε, όταν αποτύγχανε η σύγκλιση, στο 2000.* (Pedregosa et al., 2011)

Δοκιμάστηκαν οι εξής παραλλαγές:

```
MLPClassifier(hidden_layer_sizes=(50,25),activation='logistic',solver='lbfgs',learning_rate='adaptive',random_state=1,max_iter=1000)

MLPClassifier(hidden_layer_sizes=(35,10),activation='logistic',solver='lbfgs',learning_rate='adaptive',random_state=1,max_iter=1000)

MLPClassifier(hidden_layer_sizes=(25,8),activation='logistic',solver='lbfgs',learning_rate='adaptive',random_state=1,max_iter=1000)

MLPClassifier(hidden_layer_sizes=(35,8),activation='logistic',solver='lbfgs',learning_rate='adaptive',random_state=1,max_iter=1000)

MLPClassifier(hidden_layer_sizes=(25,25),activation='logistic',solver='lbfgs',learning_rate='adaptive',random_state=1,max_iter=1000)

MLPClassifier(hidden_layer_sizes=(35,15),activation='logistic',solver='lbfgs',learning_rate='adaptive',random_state=1,max_iter=1000)

MLPClassifier(hidden_layer_sizes=(35,9),activation='logistic',solver='lbfgs',learning_rate='adaptive',random_state=1,max_iter=1000)

MLPClassifier(hidden_layer_sizes=(35,11),activation='logistic',solver='lbfgs',learning_rate='adaptive',random_state=1,max_iter=1000)

MLPClassifier(hidden_layer_sizes=(36,10),activation='logistic',solver='lbfgs',learning_rate='adaptive',random_state=1,max_iter=2000)

MLPClassifier(hidden_layer_sizes=(34,10),activation='logistic',solver='lbfgs',learning_rate='adaptive',random_state=1, max_iter=2000)
```

5.4.6 Κ-Κοντινότεροι Γείτονες (K-Nearest Neighbours, KNN1-6)

Η ταξινόμηση της καταγραφής υπολογίζεται από που ανήκουν (σε ποια κλάση) οι k κοντινότεροι γειτόνές της κατά πλειοψηφία. Δοκιμάστηκαν οι εξής παραλλαγές:

```
KNeighborsClassifier(n_neighbors=6)
KNeighborsClassifier(n_neighbors=11)
KNeighborsClassifier(n_neighbors=14)
KNeighborsClassifier(n_neighbors=16)
KNeighborsClassifier(n_neighbors=22)
KNeighborsClassifier(n_neighbors=48)
```

Αναφορικά με την επιλογή της τιμής του k στην μέθοδο K-NN, έγινε πρώτα υπολογισμός του Ποσοστού Σφάλματος (Error Rate) για διαφορετικά k, σε σετ στο οποίο έγινε προεπεξεργασία (preprocessing) κατά την πορεία (εμπειρική συμπλήρωση, έλλειψη εξισορρόπησης) και τελικά επιλέχθηκαν αυτά με το μικρότερο σφάλμα (K=6, 11, 14, 16, 22, 48) για να γίνει ταξινόμηση. Τα σχετικά με το Error Rate διαγράμματα υπάρχουν στο *Παράρτημα ΙΙ*.

Το KNN θεωρείται "τεμπέλης" αλγόριθμος. Δεν χρειάζεται σημεία εκπαίδευσης για τη δημιουργία μοντέλου. Τα δεδομένα εκπαίδευσης χρησιμοποιούνται και στη φάση των ελέγχων. Αυτό καθιστά την εκπαίδευση γρηγορότερη και τη φάση ελέγχων πιο αργή και δαπανηρή σε χρόνο και μνήμη. (Avinash Navlani, 2018)

5.4.7 Αφελείς Μπεϋσιανοί Κατηγοριοποιητές (Naive Bayes Classifiers- BNB1-2, GNB)

Δοκιμάστηκαν οι κατηγοριοποιητές Bernoulli Naive Bayes και Gaussian Naive Bayes. Η υπόθεση στο Bernoulli Naive Bayes είναι ότι το χαρακτηριστικό είναι δυαδικό (0 ή 1).

```
BernoulliNB(binarize = True)
```

Με την παράμετρο ‘binarize’ ορίζεται το κατώφλι για τη δυαδικότητα χαρακτηριστικών του δείγματος. Η δυαδικοποίηση αφορά χαρτογράφηση στα Booleans. Εάν η τιμή του είναι ‘None’, η είσοδος αποτελείται από δυαδικά διανύσματα.

```
BernoulliNB(binarize = 100)
```

Στο Gaussian Naive Bayes υπάρχει υπόθεση ότι τα δεδομένα συντίθενται από συνδυασμό κανονικών κατανομών (τύπου Gauss).

```
GaussianNB()
```

Οι πρότερες πιθανότητες (priors) προσαρμόζονται σύμφωνα με τα δεδομένα, λόγω του ότι δεν τις ορίσαμε. (Pedregosa et al., 2011)

Στο ***Παράρτημα ΙΙΙ*** υπάρχει πίνακας όπου καταγράφεται κάθε παραλλαγή στις παραμέτρους, που δοκιμάστηκε για έκαστη μέθοδο Ταξινόμησης που αναφέρεται πιο πάνω.

5.5 Αξιολόγηση Ταξινόμησης - Classification Evaluation

Συμπεριλήφθησαν περισσότερες μετρικές (metrics) αξιολόγησης της ταξινόμησης από ό,τι ήθισται σε αντίστοιχες εργασίες, ώστε να έχουμε πληρέστερη εικόνα για την ποιότητα της ταξινόμησης και να μπορούμε να την αξιολογίσουμε με αξιόπιστο τρόπο. Ενδεικτικά αναφέρουμε ότι συνήθως στην βιβλιογραφία γίνεται αξιολόγηση ταξινόμησης με χρήση από μίας έως πέντε μετρικών. Συνήθως ανάμεσά τους είναι η

μετρική ακρίβεια ή ορθότητα (accuracy) και το σκορ-f1 (f1-score). Ωστόσο είναι προβληματικό το να βασιζόμαστε αποκλειστικά σε αυτές τις λίγες μετρικές, γιατί κάθε μία αξιολογεί την ταξινόμηση με βάση άλλο κριτήριο και χρειάζεται να δούμε μία ολόπλευρη αξιολόγηση για να μπορούμε να βγάλουμε χρήσιμα συμπεράσματα. Έγινε χρήση μετρικών ("Evaluation of binary classifiers," n.d.) *όπως:*

a. η ακρίβεια/ορθότητα (**accuracy**): **(TP + TN) / (TP + TN + FR + FN)**
Ωστόσο η μετρική ακρίβεια/ορθότητα (accuracy) μιας μεθόδου δεν αρκεί για να έχουμε ένα καλό προγνωστικά μοντέλο. Αν υποθέσουμε ότι έχουμε να κάνουμε με μία ασθένεια που πλήττει το 0.1% του πληθυσμού, τότε το να προγνώσουμε ότι μία νέα καταγραφή ασθενούς αφορά υγιές άτομο, μας δίνει ακρίβεια 0.9999.("Lecture 13: Classification," 2016). Χρειάζεται προσοχή γιατί και η μετρική θετική προγνωστική αξία (precision) πολλές φορές μεταφράζεται ως ακρίβεια. Πιθανώς η μετάφραση ορθότητα να έλυνε το πρόβλημα και να ήταν πιο εύστοχη εφόσον η μετρική αυτή αφορά το πόσο «σωστή» είναι μία ταξινόμηση. Η μετρική ακρίβεια ή ορθότητα (accuracy) αν και ευρέως χρησιμοποιούμενη σε προβλήματα Ταξινόμησης (Classification) **δεν δίνει βάσιμα αποτελέσματα σε περιπτώσεις που το σύνολο εκπαίδευσης των μοντέλων ταξινόμησης δεν είναι ισορροπημένο ως προς την μεταβλητή στόχο**. Η μετρική αυτή έχει συμπεριληφθεί στην εργασία λόγω του ότι χρησιμοποιείται κατά κανόνα και η παράληψή της δεν θα επέτρεπε την σύγκριση με άλλες εργασίες. Χρειάζεται οι τιμές της, με εύρος τιμών [0,1], να είναι όσο πιο κοντά στην μονάδα γίνεται για να έχουμε μία καλή -ως προς την μετρική αυτή- ταξινόμηση.

b. η *ευαισθησία (recall ή **sensitivity** ή true positive rate-TPR):* **TP / (TP + FN)**,
Πρόκειται για μετρική που δείχνει πόσες καταγραφές από το σύνολο ελέγχου της ταξινόμησης ταξινομήθηκαν ως *θετικές* (ότι ανήκουν δηλαδή στην Class '1') *αληθώς* (ενώ δηλαδή όντως ανήκουν εκεί), προς το σύνολο των καταγραφών που πράγματι ανήκουν στην Κατηγορία ασθενών (Class) '1' -δηλαδή όσους δεν ζουν μετά το ένα έτος- (είτε ταξινομήθηκαν ορθά είτε όχι). Σε σύνολα δεδομένων στα οποία πλειοψηφούν οι καταγραφές της Κατηγορίας (Class) '1' (όπως και στο δικό μας), αν δεν γίνει εξισορρόπηση ως προς την μεταβλητή στόχο, η ευαισθησία τείνει να είναι υψηλή αφού τα μοντέλα ταξινόμησης μαθαίνουν να ταξινομούν σωστότερα την πλειοψηφούσα Class '1' και λιγότερο καλά την μειοψηφούσα Class '0' (στην περίπτωσή μας). Και σα αυτή την μετρική με εύρος τιμών [0,1], οι τιμές που πλησιάζουν την μονάδα δείχουν καλή ευαισθησία στην ταξινόμηση.

c. η *θετική προγνωστική αξία (**precision**, positive predictive value-PPV)*:
TP / (TP + FP),
ο λόγος αληθώς θετικών προς το σύνολο όσων καταγραφών ταξινομήθηκαν ως θετικές (και είτε πράγματι είναι θετικές, είτε όχι). Σε σύνολα δεδομένων μη ισορροπημένα, όπως το δικό μας, αν δεν γίνει εξισορρόπηση, αναμένεται να έχουμε υψηλές τιμές θετικής προγνωστικής αξίας, αν και όχι τόσο ψηλές όσο της ευαισθησίας. Αυτό σημαίνει ότι ακόμα και οι καταγραφές που ανήκουν στην μειοψηφούσα Κατηγορία ασθενών (Class) ('0' στην περίπτωσή μας) ταξινομούνται ως να ανήκουν στην πλειοψηφούσα (Class '1' στην περίπτωσή μας) λόγω του ότι τα μοντέλα εκπαιδεύονται σε μη ισορροπημένο σύνολο εκπαίδευσης ως προς την μεταβλητή στόχο και συνακόλουθα μεροληπτούν (bias) υπέρ της πλειοψηφούσας Κατηγορίας ασθενών (Class) «στέλνοντας» εκεί πολύ περισσότερες καταγραφές από όσες πράγματι ανήκουν εκεί. Και εδώ το εύρος τιμών είναι [0,1] και οι τιμές που πλησιάζουν την μονάδα δείχνουν ότι έχουμε καλή θετική προγνωστική αξία.

d. η *ειδικότητα (**specificity** ή selectivity ή true negative rate-TNR):*
TN / (TN + FP),
Η μετρική ειδικότητα είναι ο λόγος των αληθώς αρνητικών καταγραφών - όσων δηλαδή ταξινομήθηκαν να ανήκουν στην Κατηγορία ασθενών (Class) ‘0’ ορθά- διά το σύνολο των καταγραφών του συνόλου δεδομένων ελέγχου που όντως ανήκουν στην Κατηγορία ασθενών (Class) ‘0’. Αν δεν γίνει εξισορρόπηση ως προς την μεταβλητή στόχο κατά την προεπεξεργασία των δεδομένων και πλειοψηφεί η Κατηγορία (Class) ‘1’ στο σύνολο εκπαίδευσης η ειδικότητα αναμένεται να είναι χαμηλή και ενδεικτική κακής ποιότητας ταξινόμησης για την Κατηγορία (Class) ‘0’. Αυτό προκύπτει από το γεγονός ότι τα μοντέλα ταξινόμησης -όπως αναφέρθηκε και πιο πάνω- μεροληπτούν υπέρ της Κατηγορίας (Class) ‘1’. Έτσι ταξινομούν ως θετικές (‘1’) καταγραφές που είναι αρνητικές (‘0’) και τα ψευδώς θετικά αυξάνονται, αυξάνοντας τον παρονομαστή της ειδικότητας. Καλή ειδικότητα σε μία ταξινόμηση έχουμε για τιμές που πλησιάζουν την μονάδα και το εύρος τιμών είναι [0,1].

e. η *αρνητική προγνωστική αξία (**negative predictive value**-NPV):* **TN / (TN + FN)**
Η μετρική αυτή είναι ο λόγος των αληθώς αρνητικών διά το σύνολο των ταξινομηθέντων ως αρνητικά (είτε είναι αρνητικά, είτε όχι). Σε σύνολα δεδομένων μη ισορροπημένα κατά τον τρόπο που είναι και το δικό μας είναι πιθανό -ελλείψει εξισορρόπησης ως προς την μεταβλητή στόχο- να έχουμε μεγάλη αρνητική προγνωστική αξία. Δεν πρέπει να αξιολογηθεί όμως αυτό το αποτέλεσμα ως απαραίτητα καλό, μιας και όταν η πλειοψηφία των καταγραφών ταξινομείται μεροληπτικά σαν θετική, αναμένεται να έχουμε μικρό αριθμό ψευδώς αρνητικών. Γενικά οι τιμές, με εύρος [0,1], που πλησιάζουν την μονάδα δείχνουν μία καλή αρνητική προγνωστική αξία.

f. *ψευδώς αρνητικό ποσοστό (false negative rate-FNR ή **miss rate**):*
FN / (FN + TP)
Είναι ο ρυθμός εμφάνισης αρνητικών αποτελεσμάτων δοκιμών σε εκείνους που έχουν το χαρακτηριστικό ή την ασθένεια για την οποία δοκιμάζονται. (“false-negative rate,” 2009) Στην περίπτωσή μας είναι ο λόγος των ψευδώς αρνητικών προς το σύνολο των πράγματι θετικών (όσων ανήκουν στην Κατηγορία ασθενών (Class) ‘1’). Σε σύνολα δεδομένων όπως το δικό μας - ελλείψει ταξινόμησης- συνήθως το ψευδώς αρνητικό ποσοστό είναι χαμηλό. Αυτό δείχνει ότι η Κατηγορία ασθενών (Class) ‘1’ ταξινομείται σχετικά καλά. Το εύρος τιμών της είναι [0,1], και οι τιμές της μετρικής αυτής που πλησιάζουν στο 0 δείχνουν ότι υπάρχει καλό ψευδώς αρνητικό ποσοστό.

g. *Ψευδώς θετικό ποσοστό (false positive rate-FPR ή **fall-out***):
FP / (FP + TN)
Είναι ο λόγος μεταξύ των ψευδώς θετικών προς τον συνολικό αριθμό των αρνητικών (ανεξάρτητα από την ταξινόμηση) (FP/N). Σε σύνολα δεδομένων όπως το δικό μας αναμένεται να υπάρχει υψηλό ψευδώς θετικό ποσοστό. Αυτό οφείλεται γιατί πολλές καταγραφές που ανήκουν στην Κατηγορία ασθενών (Class) ‘0’ τοποθετούνται στην ‘1’ κατά την ταξινόμηση από τα μοντέλα ταξινόμησης που μεροληπτούν υπέρ της ‘1’. Το εύρος τιμών της είναι [0,1] και τιμές ψευδώς θετικού ποσοστού που πλησιάζουν το 0 δείχνουν ότι έχουμε καλή ταξινόμηση της Κατηγορίας ασθενών (Class) ‘0’.

h. *ποσοστό ψευδούς παράλειψης (**false omission rate**-FOR)*: **FN / (FN + TN)**
Η μετρική αυτή είναι ο λόγος των ψευδώς αρνητικών προς το σύνολο όσων ταξινομήθηκαν ως αρνητικά (είτε είναι, είτε όχι). Σε σύνολα δεδομένων όπως το δικό μας αναμένεται να έχουμε χαμηλό ποσοστό ψευδούς παράλειψης αν δεν προηγηθεί εξισορρόπηση του συνόλου δεδομένων. Δεν αρκεί όμως η συγκεκριμένη μετρική για να μας δείξει ότι έχουμε καλή ταξινόμηση, καθώς ένα μοντέλο που μεροληπτεί υπέρ της Κατηγορίας ασθενών (Class) '1' στην περίπτωσή μας θα έδινε πράγματι χαμηλό ποσοστό ψευδούς παράλειψης. Γενικά τιμή της μετρικής αυτής που τείνει στο 0 είναι ενδεικτική καλής ταξινόμησης και το εύρος τιμών είναι [0,1].

i. *κατώφλι επικράτησης (**prevalence threshold**-PT):*

$$\mathrm{PT} = \frac{\sqrt{\mathrm{TPR}(-\mathrm{TNR}+1)}+\mathrm{TNR}-1}{(\mathrm{TPR}+\mathrm{TNR}-1)} = \frac{\sqrt{\mathrm{FPR}}}{\sqrt{\mathrm{TPR}}+\sqrt{\mathrm{FPR}}}$$

("Evaluation of binary classifiers," n.d.)
Είναι ένα σημείο τοπικής ακραίας καμπυλότητας που ορίζεται ως συνάρτηση της ευαισθησίας και της ειδικότητας, πέρα από το οποίο ο ρυθμός μεταβολής της θετικής προγνωστικής αξίας ενός τεστ πέφτει απότομα σε σχέση με το ϕ (Συντελεστής Phi). *Όσο μεγαλύτερη είναι η επιφάνεια κάτω από την καμπύλη ROC (AUC,* βλ. παρακάτω*), τόσο χαμηλότερο είναι το κατώφλι επικράτησης και αντίστροφα. Η ερμηνεία των προγνωστικών τιμών γίνεται στο επίπεδο ενός μόνο αποτελέσματος δοκιμής, μεταξύ ατόμων στα οποία δεν έχει γίνει ακόμη διάγνωση και των οποίων η τελική διαγνωστική κατάσταση είναι άγνωστη*. (Balayla, 2020) Υπολογίζεται από το ψευδώς θετικό ποσοστό και την ευαισθησία. Σε σύνολα δεδομένων όπως το δικό μας – ελείψει εξισορρόπησης- αναμένεται το ψευδώς θετικό ποσοστό και η ευαισθησία να είναι υψηλά. Συνεπώς το κατώφλι επικράτησης θα είναι κοντά στο 0,5. Το εύρος τιμών του είναι [0,1] και τιμές όσο πιο χαμηλές είναι ενδεικτικές καλύτερης ταξινόμησης.

j. *βαθμολογία απειλής ή κρίσιμος δείκτης επιτυχίας (**threat score** ή critical success index):* **TP / (TP + FN + FP)**
Το εύρος τιμών της είναι [0-1] (όπου 1 = τέλεια πρόβλεψη). Δεν είναι αμερόληπτο, δίνει χαμηλότερες βαθμολογίες για σπανιότερα γεγονότα. (Larner, 2021) Είναι ο λόγος των αληθώς θετικών προς το σύνολο όσων όντως είναι θετικά και όσων ταξινομήθηκαν ψευδώς ως θετικά. Σε σύνολα δεδομένων όπως το παρόν, αν δεν προηγηθεί εξισορρόπηση, αναμένεται να έχουμε λίγες καταγραφές που ανήκουν στην μειοψηφική Κατηγορία ασθενών (Class) '0', άρα και τα ψευδώς θετικά (που θα είναι σχετικά αυξημένα) θα είναι λίγα σε σχέση με την πλειοψηφούσα Κατηγορία ασθενών (Class) '1'. Έτσι η μετρική αναμένεται να λαμβάνει υψηλές τιμές αλλά όχι εύκολα τιμή ίση με το 1, αφού ο παρονομαστής της δεν θα είναι πολύ μεγαλύτερος του αριθμητή.

k. *Εξισορροπημένη ακρίβεια (**balanced accuracy**)*: **(TPR + TNR) / 2**
Ο μέσος όρος ευαισθησίας που λαμβάνεται από κάθε τάξη. Η καλύτερη τιμή είναι 1 και η χειρότερη τιμή είναι 0. (Pedregosa et al., 2011) Στην περίπτωσή μας αναμένεται μέτρια εξισορροπημένη ακρίβεια αν δεν προηγηθεί εξισορρόπηση του συνόλου δεδομένων ως προς την μεταβλητή στόχο, καθώς η ευαισθησία για την Κατηγορία ασθενών (Class) '1' αναμένεται να είναι υψηλή, ενώ για την '0' χαμηλή. Το εύρος τιμών της είναι [0,1] και τιμή κοντά στο 1 είναι ενδεικτική καλής ταξινόμησης και των δύο κατηγοριών.

l. *Σκορ-f1 (**f1-score**):* **(2 * TP) / (2 * TP + FP + FN)**
Είναι ο αρμονικός μέσος θετικής προγνωστικής αξίας και ευαισθησίας. Το εύρος τιμών του είναι [0,1] και τιμές κοντά στο 1 είναι δείχνουν καλή ποιότητα ταξινόμησης και για τις δύο κατηγορίες. Ωστόσο δεν είναι συγκρίσιμη η τιμή της σαν μετρική αν αφορά διαφορετικά σύνολα δεδομένων με διαφορετικές αναλογίες σε Κατηγορίες ασθενών (Class) '0', '1' στο καθένα. Σε σύνολα δεδομένων όπως στο δικό μας -αν δεν προηγηθεί εξισορρόπηση ως προς την μεταβλητή στόχο- αναμένεται να έχουμε μέτρια προς καλή τιμή σε αυτή την μετρική, μιας και υ ευαισθησία αναμένεται να είναι πολύ υψηλή και να έλκει τον αρμονικό μέσο προς τα πάνω -παρά την σχετικά χαμηλότερη θετική προγνωστική αξία.

m. *Συντελεστής συσχέτισης Matthews (Matthews correlation coefficient-**MCC**):*

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}}$$

("Evaluation of binary classifiers," n.d.)
Είναι ισορροπημένο μέτρο, μπορεί να χρησιμοποιηθεί και για Κατηγορίες (Class) με διαφορά μεγέθους. Είναι συντελεστής συσχέτισης μεταξύ των παρατηρούμενων και των προβλεπόμενων ταξινομήσεων. Επιστρέφει μια τιμή μεταξύ −1 και +1, όπου +1: τέλεια πρόβλεψη, 0: όχι καλύτερη από την τυχαία και −1: πλήρης διαφωνία μεταξύ πρόβλεψης και παρατήρησης. Αν δεν ισούται με −1, 0 ή +1, δεν είναι αξιόπιστος δείκτης για το πόσο παρόμοιος είναι ένας προγνωστικός παράγοντας με την τυχαία εικασία, επειδή εξαρτάται από το σύνολο δεδομένων. ("Matthews correlation coefficient," n.d.)

n. *Δείκτης Fowlkes-Mallows (Fowlkes-Mallows index -**FM**):*

$$\mathrm{FM} = \sqrt{\frac{TP}{TP + FP} \times \frac{TP}{TP + FN}} = \sqrt{PPV \times TPR}$$

("Evaluation of binary classifiers," n.d.)
Είναι ο γεωμετρικός μέσος θετικής προγνωστικής αξίας και ευαισθησίας. Το εύρος τιμών του είναι [0,1] Αν είναι ίσο με 0, όλα τα στοιχεία έχουν ταξινομηθεί εσφαλμένα, ενώ αν είναι 1, όλα τα στοιχεία έχουν ταξινομηθεί σωστά. Σε σύνολα δεδομένων όπου υπερεκπροσωπείται η Κατηγορία ασθενών (Class) '1', όπως στο δικό μας, αναμένεται -ελλείψει εξισορρόπησης- να έχουμε ψηλή τιμή, αλλά όχι τέλεια, μιας και τα ψευδώς θετικά θα είναι αρκετά.

o. *Πληροφοριακότητα (**Informedness**):* TPR + TNR – 1
Κυμαίνεται από [-1,1], όπου 0: ένα διαγνωστικό τεστ δίνει την ίδια αναλογία θετικών αποτελεσμάτων για ομάδες που είναι και για ομάδες που δεν είναι «θετικές» (άχρηστο τεστ), 1: δεν υπάρχουν ψευδώς θετικά ή ψευδώς αρνητικά, δηλαδή (τέλειο τεστ). Δίνει ίσο βάρος στις ψευδώς θετικές και ψευδώς αρνητικές τιμές, οπότε όλες οι δοκιμές με την ίδια τιμή του δείκτη δίνουν την ίδια αναλογία συνολικών λανθασμένων αποτελεσμάτων. Είναι τεχνικά δυνατό να ληφθεί μια τιμή μικρότερη από μηδέν από αυτήν την εξίσωση, αν π.χ. υπάρξουν μόνο Ψευδώς Θετικά (ΨΘ,FP) και Ψευδώς Αρνητικά (ΨΑ, FN), μια τιμή μικρότερη από μηδέν δείχνει απλώς ότι οι θετικές και αρνητικές ετικέτες έχουν αλλάξει. Αφού διορθωθούν οι ετικέτες, το αποτέλεσμα θα είναι στη περιοχή 0-1. ("Youden's J statistic," n.d.) *Είναι ένα μέτρο του πόσο ενημερωμένο είναι το μοντέλο ταξινόμησης για τις*

καταγραφές που είναι θετικές και αρνητικές ('Informedness and Markedness Alternatives to Recall and Precision as Evaluation Measures', 2018). Οι μεγαλύτερες τιμές της μετρικής αυτής δείχνουν καλύτερη ποιότητα ταξινόμησης.

p. *Μέτρο του αξιοσημείωτου (**markedness**): PPV* + NPV - 1
Η συγκεκριμένη μετρική δεν έχει κάποια επίσημη ελληνική μετάφραση -με βάση την αναζήτηση που κάναμε. Μετρά το πόσο μια μεταβλητή είναι προγνωστική ή πιθανή αιτία μιας άλλης. Είναι επίσης γνωστή ως Δp (deltaP) σε απλές περιπτώσεις δύο επιλογών. ("Markedness," n.d.) Αφορά το πόσο αξιοσημείωτο είναι κάτι σε σχέση με το συνηθισμένο/φυσικό. *Είναι ένα μέτρο αξιοπιστίας θετικών και αρνητικών προβλέψεων από το μοντέλο ταξινόμησης.* ('Informedness and Markedness Alternatives to Recall and Precision as Evaluation Measures', 2018) Επιχειρήθηκε να γίνει μία πρόταση μετάφρασής της εδώ. *Είναι ευαίσθητη στις αλλαγές δεδομένων οπότε δεν είναι κατάλληλη για μη ισορροπημένα δεδομένα. Αυτό συμβαίνει επειδή εξαρτάται από τις θετική προγνωστική αξία (PPV) και αρνητική προγνωστική αξία (NPV), οι οποίες είναι ευαίσθητες στις αλλαγές κατανομών δεδομένων.* (Tharwat, 2021) Το εύρος τιμών της είναι [-1,1].

q. Δημιουργήθηκε *Πίνακας Σύγχυσης* (***confusion matrix***)
(βλ.αρχεία αναλύσεων) για κάθε μοντέλο,

r. η *καμπύλη Χαρακτηριστικού λειτουργίας δέκτη (Receiver operating characteristic-**ROC**),*
μια γραφική παράσταση που απεικονίζει τη διαγνωστική ικανότητα ενός μοντέλου δυαδικού ταξινομητή καθώς το όριο διάκρισής του (discrimination threshold) ποικίλλει. Δημιουργείται με τη σχεδίαση του ποσοστού αληθώς θετικών ή ευαισθησία (sensitivity, TPR) έναντι του ποσοστού ψευδώς θετικών (1-specificity, FPR) σε διάφορες ρυθμίσεις κατωφλίου. ('Receiver operating characteristic', 2021)

s. η *επιφάνεια κάτω από την καμπύλη ROC (Area Under the Curve–**AUC**),*
το μέτρο της ικανότητας ενός ταξινομητή να διακρίνει μεταξύ Κατηγοριών και χρησιμοποιείται ως σύνοψη της καμπύλης ROC. Όσο υψηλότερη είναι η AUC, τόσο καλύτερη είναι η απόδοση του μοντέλου στη διάκριση μεταξύ των θετικών και αρνητικών κατηγοριών. (Aniruddha Bhandari, 2020)

t. η *τιμή κ του Cohen (**Cohen's kappa** value ή kappa statistic)*:

$$\kappa = \frac{2 \times (TP \times TN - FN \times FP)}{(TP + FP) \times (FP + TN) + (TP + FN) \times (FN + TN)}$$

Η μετρική αυτή στην συγκεκριμένη *περίπτωση μετρά την αξιοπιστία του ταξινομητή για ποιοτικά (κατηγορικά) στοιχεία (συμφωνία μεταξύ ταξινομητή και της στέρεης γνώσης που έχουμε για τα δεδομένα). Πιστεύεται ότι είναι ένα πιο ισχυρό μέτρο από τον απλό υπολογισμό του ποσοστού συμφωνίας, καθώς λαμβάνει υπόψη την πιθανότητα η συμφωνία να συμβεί τυχαία. Όμως έχει τάση να λαμβάνει τις συχνότητες των παρατηρούμενων κλάσεων ως δεδομένες, γεγονός που μπορεί να το κάνει αναξιόπιστο για τη μέτρηση της συμφωνίας με τη στέρεη γνώση που έχουμε σχετικά με το σετ ελέγχου, σε καταστάσεις όπως η διάγνωση σπάνιων ασθενειών. Σε αυτές τις καταστάσεις, το κ τείνει να υποτιμά τη συμφωνία για τη σπάνια κατηγορία. Για το λόγο αυτό, το κ θεωρείται υπερβολικά συντηρητικό μέτρο συμφωνίας. Αμφισβητείται επίσης ο ισχυρισμός ότι το κ "λαμβάνει υπόψη" την τυχαία συμφωνία. Για να γίνει αυτό αποτελεσματικά θα απαιτούσε ένα σαφές μοντέλο*

> *του τρόπου με τον οποίο η πιθανότητα επηρεάζει τις αποφάσεις ταξινόμησης. Αντ' αυτού όταν δεν υπάρχει απόλυτη βεβαιότητα, οι ταξινομητές θεωρείται ότι μαντεύουν ένα πολύ εξωπραγματικό σενάριο*. ("Cohen's kappa," n.d.)

Στα Αποτελέσματα και συγκεκριμένα στα *Παραρτήματα IV a, b* ακολουθεί η παρουσίαση των γραφημάτων όπου αντιπαραβάλλονται οι τιμές των πιο πάνω μέτρων αξιολόγησης της ταξινόμησης για τις διαφορετικές κατηγορίες μοντέλων ταξινόμησης που δοκιμάστηκαν και για τις διαφορετικές πορείες προεπεξεργασίας δεομένων (preprocessing) που ακολουθήθηκαν. Οι καμπύλες ROC και οι πίνακες περιέχονται στα αρχεία των αναλύσεων που κατατίθενται μαζί με την παρούσα εργασία, αλλά για λόγους οικονομίας χώρου δεν περιέχονται στο αρχείο του κειμένου.

5.6 Περιορισμοί – Αδύναμα σημεία της Έρευνας

Ακολουθούν ιδέες που θα μπορούσαν να δοκιμαστούν, αλλά διαφεύγουν από τους διδακτικούς και χρονικούς στόχους που εκ των πραγμάτων τίθενται κατά την εκπόνηση διπλωματικής εργασίας.

Σε μελλοντική πραγματοποίηση των πειραμάτων θα μπορούσε να αναζητηθεί κάποιος βελτιωμένος τρόπος καθορισμού της τιμής κάλυψης των κενών καταγραφών.

Χρήσιμη μελλοντική προσθήκη θα ήταν η πραγματοποίηση Ανάλυσης κύριας συνιστώσας (Principal component analysis-PCA) και κανονικοποίησης των δεδομένων πριν τις αναλύσεις για αποφυγή της υπερπροσαρμογής (overfitting). Ωστόσο, στόχος στην παρούσα εργασία ήταν να υπάρξει για αρχή το βέλτιστο δυνατό αποτέλεσμα με τις λιγότερες δυνατές τροποποιήσεις πάνω στο αρχικό σύνολο των δεδομένων, μιας και εστίαση είχε δωθεί στην συμπλήρωση των κενών και στην εξισορρόπηση των κλάσεων στο σύνολο των δεδομένων.

Όσον αφορά την μέθοδο ταξινόμησης K-NN, το μέσο σφάλμα (Mean error) έχει προϋπολογιστεί για την προεπεξεργασία (εμπειρική συμπλήρωση, χωρίς εξισορρόπηση) για επιλεγμένο εύρος τιμών του K. Τελικά έχουν δοκιμαστεί τα K που έδιναν το μικρότερο σφάλμα. Ωστόσο χρήσιμη προσθήκη θα ήταν να γίνει ο ίδιος υπολογισμός για κάθε διαφορετική προεπεξεργασία, ώστε να βρούμε τα βέλτιστα K για κάθε μία. Δεν παραβλέπεται όμως το γεγονός ότι αν οι βέλτιστες τιμές των K διέφεραν μεταξύ προεπεξεργασιών, θα υπήρχε πρόβλημα με την μεταξύ τους σύγκριση.

Είναι επίσης χρήσιμη η αξιοποίηση μη επιβλεπόμενων μεθόδων εύρεσης συστάδων εντός των δεδομένων (clustering). Συνηθέστερη είναι η Ομαδοποίηση K-μέσων (K-means) [για μετάφραση όρου βλ. (‘Ομαδοποίηση K-μέσων’, 2019)]. Κάτι τέτοιο δεν πραγματοποιήθηκε στην εργασία αυτή λόγω χρονοδιαγράμματος περάτωσής της.

Μία πιθανή βελτίωση θα μπορούσε να είναι η δοκιμή κάθε δυνατού συνδυασμού για τον αριθμό των νευρώνων του 1ου ενδιάμεσου στρώματος και του 2ου, πχ. (1 έως 50 , 1 έως 50) ανά κατηγορία προεπεξεργασίας δεδομένων. Κατόπιν θα μπορούσαν να επιλεγούν τα αντίστοιχα μοντέλα που έχουν τον συνδυασμό με την καλύτερη ταξινομική απόδωση για σύγκριση με τις άλλες ταξινομικές μεθόδους.

Χρήσιμη θα ήταν επίσης η γραφική παράσταση των κατανομών κάθε ενός από τα πιο πάνω μέτρα με την χρήση λχ. Θηκογράμματος (boxplot) αντί της παράθεσης των μέσων όρων τους με τις τυπικές αποκλίσεις τους σαν γραμμές σφαλμάτων. Κάτι τέτοιο δεν έγινε στην παρούσα εργασία λόγω έλλειψης χρόνου για νέα πραγματοποίηση όλων των αναλύσεων εκ νέου, δεδομένου του ότι κάποιες αναλύσεις απαιτούσαν ημέρες χρόνου εκτέλεσης (run-time). Δεν είχε προβλεφθεί εξ’ αρχής η πιο πάνω ιδέα ώστε να έχουν αποθηκευθεί όλες οι τιμές των μέτρων κατά την κάθε διασταυρούμενη αξιολόγηση 10-μερών (10-fold cross-validation). Έτσι είχαν αποθηκευτεί οι μέσοι όροι και οι τυπικές αποκλίσεις των μετρικών για κάθε ανάλυση και δημιουργήθηκαν τα γραφήματα που περιλαμβάνονται στην εργασία.

Κάτι που θα μπορούσε να βελτιωθεί είναι να δοκιμαστεί να γίνει συμπλήρωση με Iterative imputation με διαφορετικές παραμέτρους, πχ. Περισσότερες επεναλήψεις υπολογισμού των τιμών που είναι προς συμπλήρωση.

6. Αποτελέσματα

Ακολουθεί παρουσίαση των αποτελεσμάτων της Περιγραφικής Ανάλυσης, καθώς και της Προγνωστικής Ανάλυσης που πραγματοποιήθηκε.

6.1 Περιγραφική Ανάλυση (Predictive Analysis)

Ανακεφαλαιώνοντας αναφέρουμε ότι για να καταστεί σαφέστερο το σύνολο δεδομένων που χρησιμοποιήθηκε στην εργασία έγινε πρώτα περιγραφική ανάλυσή του, στην οποία κατασκευάστηκαν αντίστοιχα γραφήματα κατά περίπτωση.

Για τα ποιοτικά (μη διατακτικά και διατακτικά) χαρακτηριστικά που περιέχονταν στο σύνολο δεδομένων δημιουργήθηκαν ραβδογράματα. Στα ραβδογράμματα που ακολουθούν φαίνεται το πλήθος των καταγραφών που αντιστοιχούν σε κάθε δυνατή τιμή του κάθε χαρακτηριστικού που περιλαμβάνει το σύνολο δεδομένων που χρησιμοποιήσαμε. Μάλιστα αυτές μοιράζονται και με βάση την τελική έκβαση του ασθενούς. Για ‘Class’ = 0 έχουμε τους ασθενείς που επιβίωσαν μετά το ένα έτος, ενώ για ‘Class’ = 1 έχουμε τους ασθενείς που δεν επιβίωσαν μετά το ένα έτος. Έτσι διακρίνεται και η διαφορετική έκβαση των ασθενών ανάλογα με την τιμή που παίρνουν στο κάθε γνώρισμα.

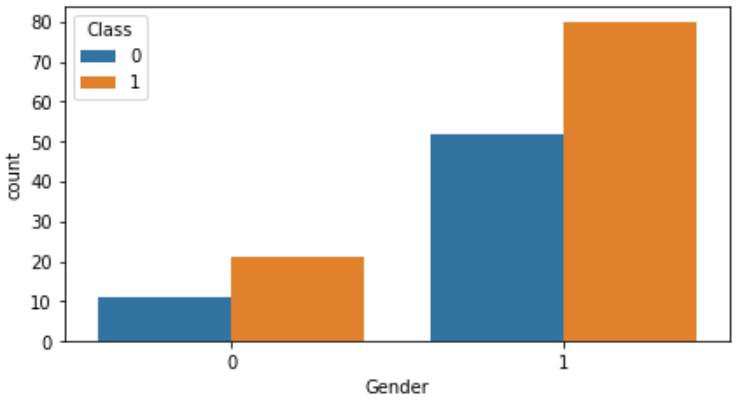
Στα ραβδογράμματα έχουμε συμπεριλάβει μία στήλη με το σύμβολο ‘?’ για να φαίνεται πόσες είναι οι καταγραφές που δεν έχουν καταγεγραμένο το χαρακτηριστικό που αφορά το εκάστοτε ραβδόγραμμα.

Για τα χαρακτηριστικά που περιέχονται στο σύνολο δεδομένων και είναι ποσοτικά, συνεχή ή διακριτά, κατασκευάστηκαν διαγράμματα κατανομών συχνοτήτων.

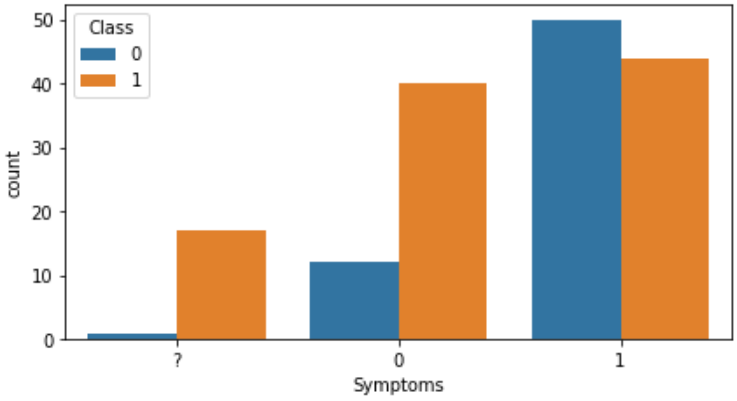
Όλα τα γραφήματα, καθώς και σύντομος σχολιασμός τους ακολουθούν στις επόμενες σελίδες.

Τα ραβδογράμματα για κάθε χαρακτηριστικό είναι τα εξής:

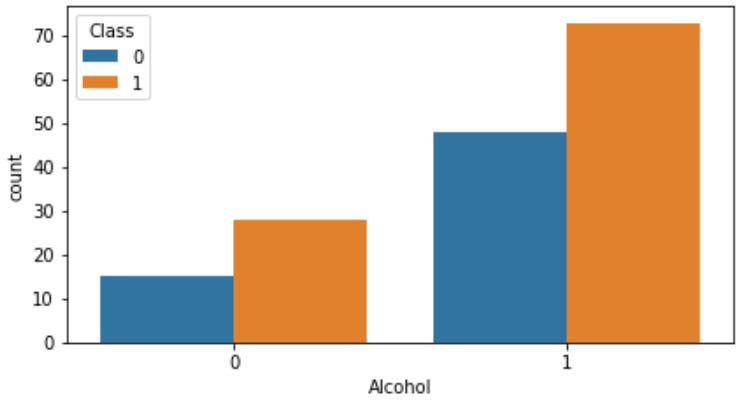
- το φύλο (Gender) (αν πρόκειται για άνδρα -1- ή γυναίκα -0-). Οι άντρες υπερεκπροσωπούνται, έχοντας την πλειοψηφία των καταγραφών, αυτό συμβαδίζει με το γεγονός ότι ο Ηπατοκυτταρικός καρκίνος είναι 4-5 φορές συχνότερος σε άρρενες ασθενείς από ό,τι σε θήλεις, όπως είδαμε και στην Βιβλιογραφική Ανασκόπηση. Μία πρώτη παρατήρηση είναι ότι ο λόγος των αντρών που πεθαίνουν μετά το ένα έτος, προς αυτούς που ζουν μετά το ένα έτος είναι μεγαλύτερος σε σχέση με τον αντίστοιχο λόγο για τις γυναίκες.
[*Σημ: Συνήθως –πιο δόκιμα– το φύλο αναφέρεται ως sex, αφού μιλάμε για βιολογικό φύλο*]



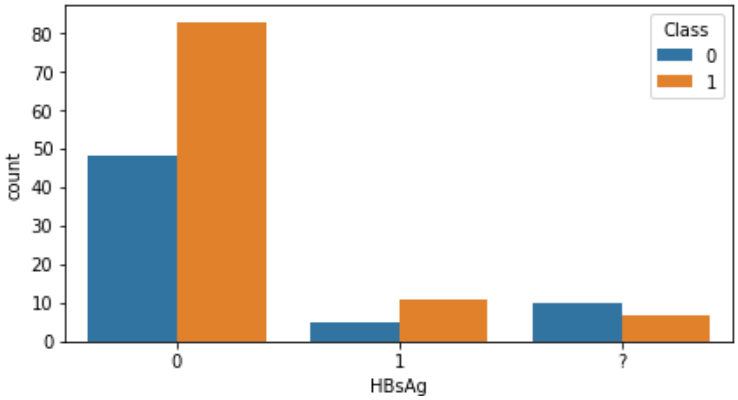
- η παρουσία συμπτωμάτων (αν υπάρχουν συμπτώματα -1- ή όχι -0-, Symptoms) **Παρατηρούμε ότι εδώ λείπουν πολλές καταγραφές για όσους ασθενείς που έχουν αποβιώσει μετά το ένα έτος (σε αντίθεση με όσους ζουν).** Μπορεί αυτό να οφείλεται στο γεγονός ότι η καταγραφή δεν έγινε έγκαιρα ώστε να μπορέσουν να καταγραφούν και όσοι αποβίωσαν νωρίς. Παρατηρούμε ότι για όσους έζησαν μετά το ένα έτος η συντριπτική πλειοψηφία ήταν συμπτωματική. Αυτό θα μπορούσε να εξηγηθεί, δεδομένου του ότι η παρουσία συμπτωμάτων συνιστά ένα δείκτη κακής υγείας που πυροδοτεί την δρομολόγηση ιατρικών εξετάσεων για να διερευνηθεί το αίτιο των συμπτωμάτων και η -κατ' επέκταση- έγκαιρη διάγνωση και πρόσβαση σε θεραπεία. Επίσης, τα άτομα που έχουν συμπτώματα που σχετίζονται με κίρρωση είναι υπό τακτικότερη ιατρική παρακολούθηση βάσει των ισχύουσων ως τότε οδηγιών κλινικής πράξης ('EASL-EORTC clinical practice guidelines: management of hepatocellular carcinoma.', 2012). Για τα άτομα που δεν επιβίωσαν μετά το ένα έτος ο αριθμός των συμπτωματικών σε σχέση με τα μη συμπτωματικά είναι σχετικά κοντά - λαμβάνοντας όμως υπ' όψη και το ότι αν οι τιμές που έλειπαν ήταν καταγεγραμένες ίσως να υπήρχε μία παραπάνω διαφοροποίηση ή και εξίσωση.



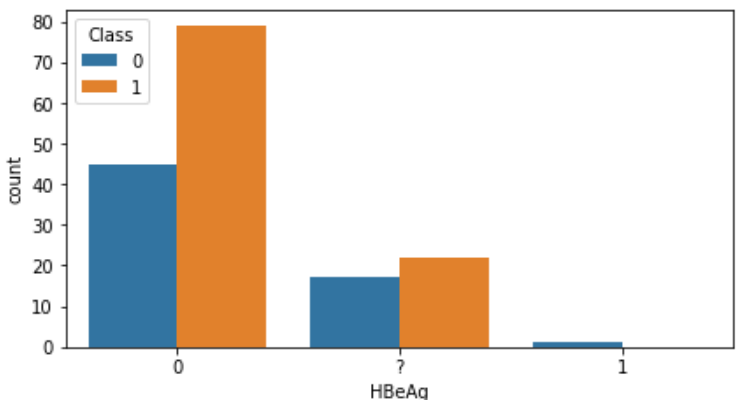
- η κατανάλωση αλκοόλ -1- ή όχι -0- (Alcohol),
  Επιβεβαιώνεται ότι για όσους κάνουν συστηματική κατανάλωση αλκοόλ ο λόγος των αποβιώσαντων προς τους επιβίωσαντες μετά το ένα έτος είναι μεγαλύτερος. Λίγο μικρότερος είναι για όσους δεν κάνουν κατανάλωση αλκοόλ. Αυτό αναμένεται όπως είδαμε και στην Εισαγωγή.



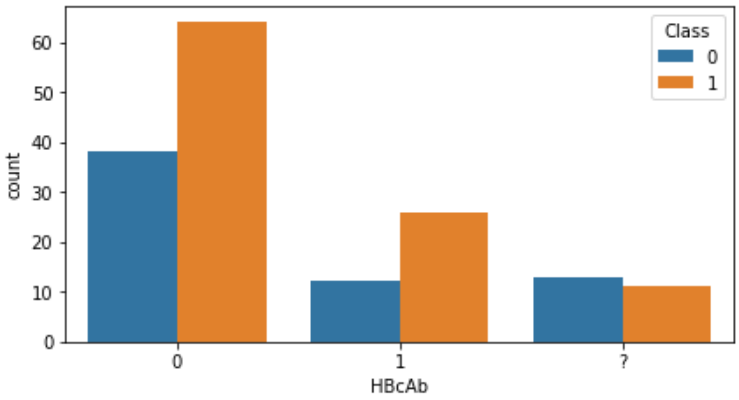
- Η παρουσία επιφανειακού αντιγόνου της Ηπατίτιδας Β -1- ή όχι -0- (HBsAg),
  Η πλειοψηφία δεν έχει παρουσία του HBsAg. Για αυτά τα άτομα λίγο λιγότεροι από τα 2/3 δεν επιβιώνουν μετά το πρώτο έτος. Παρόμοιος λόγος παρατηρείται και για όσους έχουν παρουσία του HBsAg. Ωστόσο υπάρχει αριθμός καταγραφών για τις οποίες η παρουσία του HBsAg δεν είναι καταγεγραμένη σχεδόν ίσο με τις καταγραφές που έχουν παρουσία του HBsAg. Δεν μπορούν λοιπόν να βγουν ιδιαίτερα ασφαλή συμπεράσματα.



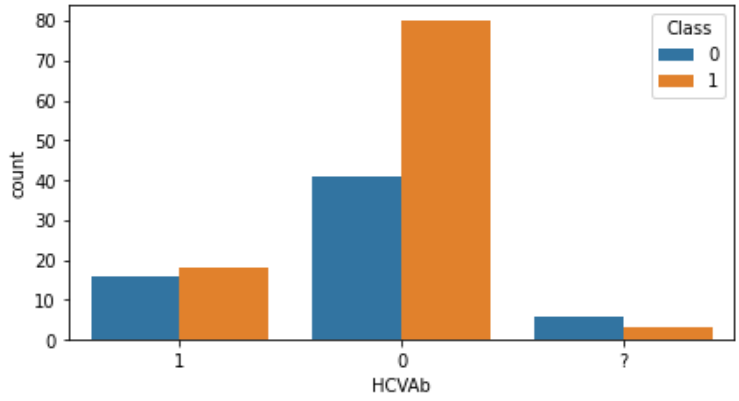
- Η παρουσία του αντιγόνου e της Ηπατίτιδας Β -1- ή όχι -0- (HBeAg),
  Ως προς το συγκεκριμένο χαρακτηριστικό λείπουν πολλές τιμές. Για ασθενείς που δεν έχουν παρουσία του HBeAg, λίγο λιγότεροι από τα 2/3 δεν επιβιώνουν μετά το 1 έτος. Ο λόγος αυτός διαφέρει για όσους διαθέτουν το HBeAg, αλλά σε καμία περίπτωση δεν μπορεί να εξαχθεί κάποιο ασφαλές συμπέρασμα -εφόσον πολύ λίγες καταγραφές έχουν το HBeAg. Γνωρίζουμε από την Εισαγωγή ότι η Οροθετικότητα για το HBeAg αποτελεί παράγοντα κινδύνου για ανάπτυξη Ηπατοκυτταρικού καρκινώματος.



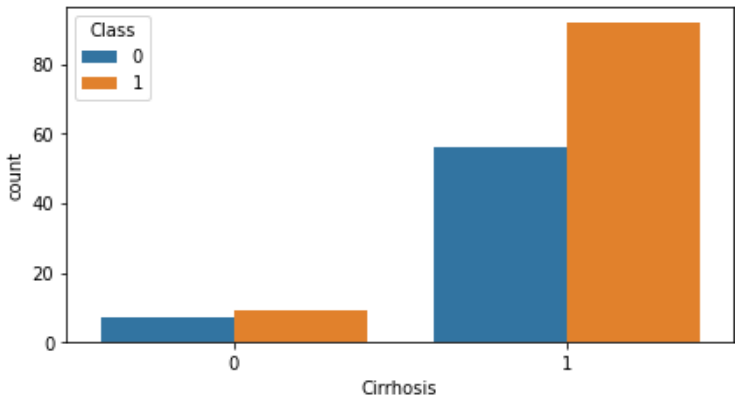
- Η παρουσία του βασικού αντισώματος για την Ηπατίτιδα Β ή όχι (HbcAb),
  Στο συγκεκριμένο χαρακτηριστικό υπάρχουν πολλές καταγραφές από τις οποίες λείπει η τιμή. Μάλιστα λείπει εξίσου σε όσους έζησαν μετά το ένα έτος και σε όσους δεν έζησαν. Για ασθενείς με παρουσία του HbcAb ο λόγος αποβιώσαντων / επιβιώσταντες είναι λίγο μικρότερος από ό,τι για ασθενείς με απουσία του HbcAb. Ωστόσο και εδώ η εξαγωγή κάποιου συμπεράσματος είναι επισφαλής, μιας και οι καταγραφές που λείπουν είναι αρκετές για αλλάξουν κατά πολύ τις αναλογίες σε περίπτωση που συμπληρωθούν.



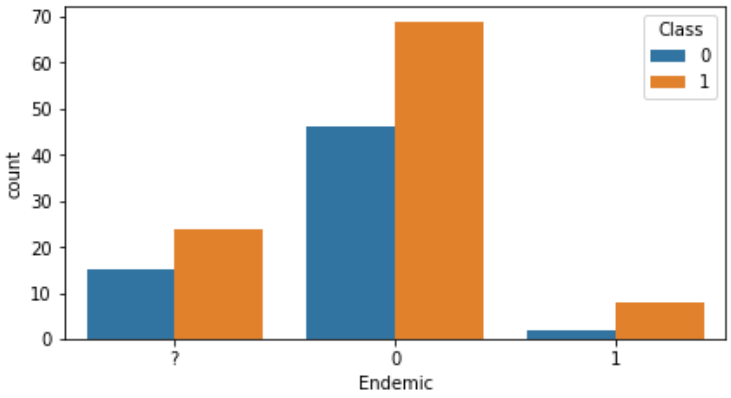
- Η παρουσία του αντισώματος για τον ιό της Ηπατίτιδας C ή όχι (HCVAb),
  Εδώ οι καταγραφές που δεν είναι συμπληρωμένες είναι σχετικά λίγες και θα μπορούσαμε με μεγαλύτερη ασφάλεια να πούμε ότι για απουσία του HCVAb ο λόγος αποβιώσαντων/επιβιώσαντες είναι περί το 2, ενώ για παρουσία του HCVAb είναι περί του 1. Από αυτό φαίνεται να είναι διπλάσια η πιθανότητα να πεθάνει κάποιος ασθενής του συνόλου δεδομένων που μελετάμε αν δεν έχει HCVAb σε σχέση με το να έχει HCVAb.



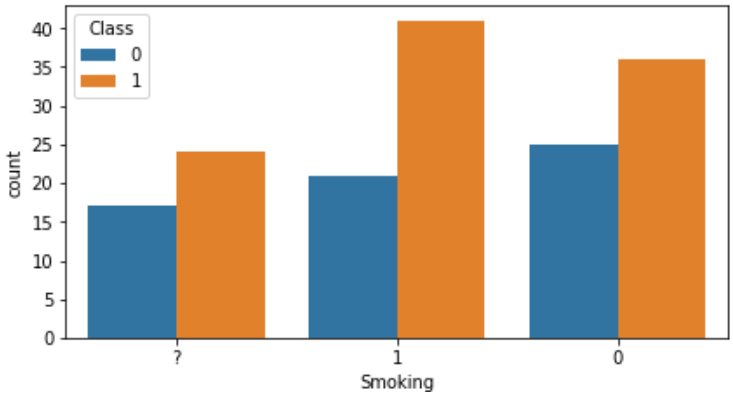
- η ύπαρξη κίρρωσης του ήπατος ή όχι (Cirrhosis),
  Για ασθενείς με κίρρωση του ήπατος ο λόγος αποβιώσαντων/επιβιώσαντες μετά το ένα έτος είναι κοντά στο 1,5. Αντίθετα για ασθενείς χωρίς κίρρωση του ήπατος είναι στο 1. Με άλλα λόγια οι ασθενείς που έχουν κίρρωση του ήπατος έχουν μεγαλύτερη πιθανότητα να αποβιώσουν μετά το ένα έτος, σε αντίθεση με όσους δεν έχουν κίρρωση. Κάτι τέτοιο είναι αναμενόμενο, όπως είδαμε και στην Εισαγωγή.



- η συσχέτιση με ενδημική χώρα ή όχι (Endemic)
  Για το παρόν χαρακτηριστικό λείπουν πολλές καταγραφές, σχεδόν διπλάσιες από αυτές που αφορούν συσχέτιση με ενδημική χώρα. **Λείπουν λίγο συχνότερα σε ασθενείς που δεν επιβίωσαν.** Η αυξημένη πλειοψηφία των καταγραφών όμως αφορά μη συσχέτιση με ενδημικές χώρες. Παρόλα αυτά δεν αρκεί για να μπορούμε να εξάγουμε ιδιαίτερα ασφαλή συμπεράσματα για το συγκεκριμένο χαρακτηριστικό. Με μία πρώτη ματιά φαίνεται οι αποβιώσαντες να είναι αισθητά περισσότεροι σε σχέση με τους επιβιώσαντες όταν υπάρχει συσχέτιση με ενδημική χώρα. Κάτι τέτοιο όπως είδαμε στην Εισαγωγή δεν αποτελεί έκπληξη.

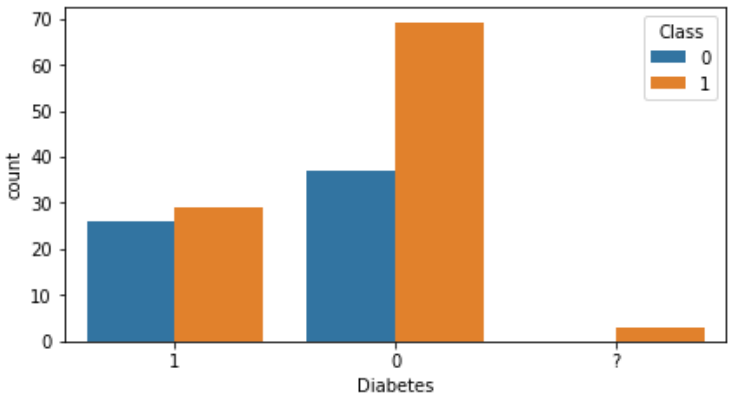


- το κάπνισμα (εάν πρόκειται για καπνιστή ή όχι) (Smoking), Και στο παρόν χαρακτηριστικό **λείπουν πολλές καταγραφές, λίγο πιο συχνά σε αποβιώσαντες, από ό,τι σε επιβιώσαντες.** Κάτι τέτοιο θα μπορούσε να καταγραφεί σχετικά εύκολα και είναι εύκολο να ξεκινήσει να καταγράφεται συστηματικά στο μέλλον. Βλέπουμε ότι για όσους ασθενείς έχει γίνει καταγραφή οι καπνιστές έχουν συχνότερα κακή έκβαση από ό,τι όσοι δεν καπνίζουν. Παρόλα αυτά δεν μπορούμε να εξάγουμε αυτό το συμπέρασμα με ασφάλεια από το συγκεκριμένο σετ μόνο -λόγω μεγάλου ποσοστού καταγραφών που λείπουν. Γνωρίζουμε πάντως ότι αυτό συμβαδίζει και με την υπάρχουσα γνώση για το θέμα αυτό (βλ. Εισαγωγή)

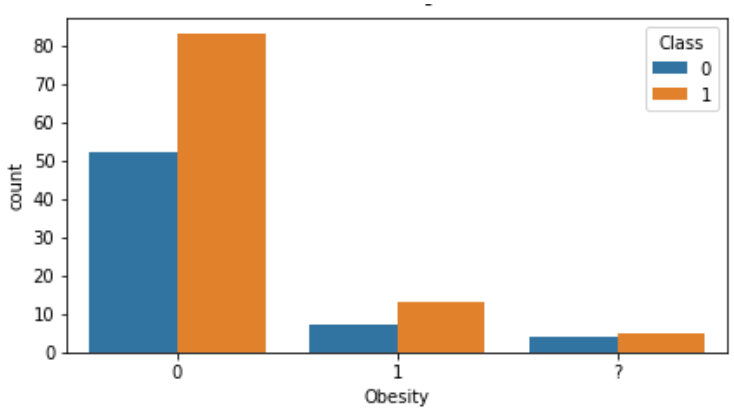


- το αν έχει διαγνωσθεί ή όχι με διαβήτη η/ο ασθενής (Diabetes),

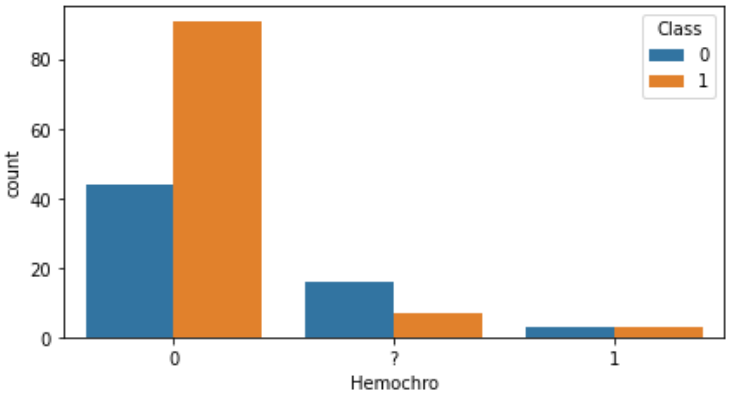
Σε περιπτώσεις διαβητικών ασθενών η αναλογία αποβιώσαντων προς επιβιώσαντες είναι κοντά στο 1. Αντίθετα για μη διαβητικούς η ίδια αναλογία είναι κοντά στο 2. Φαίνεται εδώ ότι έχουμε μεγαλύτερη πιθανότητα για κακή έκβαση στους μη διαβητικούς. Σε μία άμεση σύγκριση με τις Οδηγίες (‘EASL-EORTC clinical practice guidelines: management of hepatocellular carcinoma.’, 2012) για τους παράγοντες κινδύνου Ηπατοκυτταρικού καρκίνου βλέπουμε ότι το πιο πάνω δεν συνάδει με αυτά που συμπεριλαμβάνονται στο φύλλο οδηγιών. Ωστόσο πιθανώς να υπάρχουν κι άλλοι παράγοντες που πρέπει να ληφθούν υπ’ όψη πέρα από τους παράγοντες κινδύνου για την εκδήλωση ηπατοκυτταρικού καρινώματος. Δεν αποκλείεται και να υπάρχει μη αντιπροσωπευτικό δείγμα ασθενών στο σύνολο που διαθέτουμε.



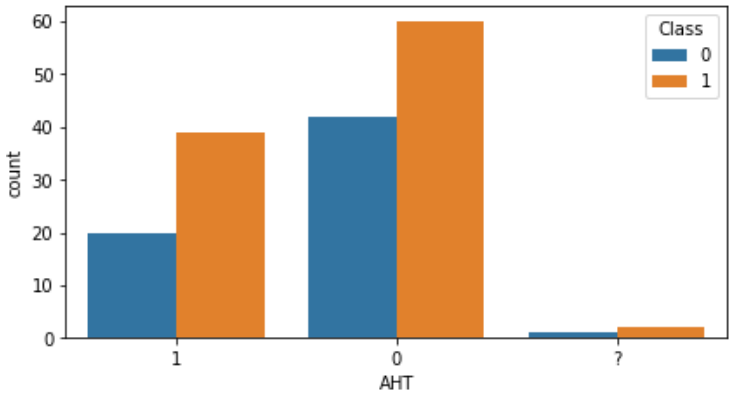
- το αν είναι ή όχι παχύσαρκος/η ο/η ασθενής (Obesity),
  Ο αριθμός των καταγεγραμένα παχόσαρκων ασθενών είναι λίγος σε σχέση με τους μη παχύσαρκους. Ωστόσο υπάρχουν και σχετικά αρκετές καταγραφές που δεν είναι συμπληρωμένες με ίδια αναλογία ως προς την μεταβλητή στόχο. Με κατάλληλη συμπλήρωση αρκούν για να αλλάξουν τις αναλογίες αποβιώσαντων / επιβιώσαντων και τα συμπεράσματα που βγαίνουν από αυτές. Συνεπώς δεν μπορούμε με ασφάλεια να βγάλουμε κάποιο συμπέρασμα. Ωστόσο όπως βλέπουμε και στην εισαγωγή η παχυσαρκία είναι παράγοντας κινδύνου για την ανάπτυξη ηπατοκυτταρικού καρκινώματος.



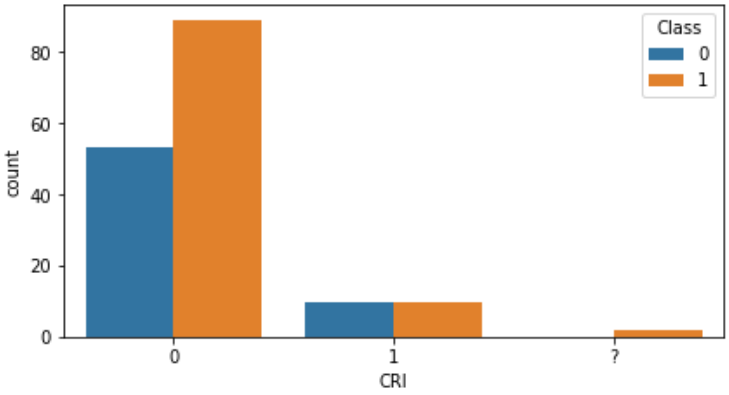
- η παρουσία ή όχι αιμοχρωμάτωσης στην/ον ασθενή (Hemochro), Υπάρχουν πολύ λίγοι ασθενείς για τους οποίους έχει καταγραφεί αιμοχρωμάτωση, καθώς και αρκετές καταγραφές που δεν είναι συμπληρωμένες ως προς αυτό το χαρακτηριστικό. Συνεπώς δεν βγαίνουν συμπεράσματα. **Είναι πιο συχνή η έλλειψη καταγραφή του χαρακτηριστικού σε επιβιώσαντες**, οπότε είναι κάτι που θα μπορούσε να δρομολογηθεί. Η αιμοχρωμάτωση είναι αναγνωρισμένη σαν ασθένεια που αποτελεί παράγοντα κινδύνου για εκδήλωση ηπατοκυτταρικού καρκινώματος.



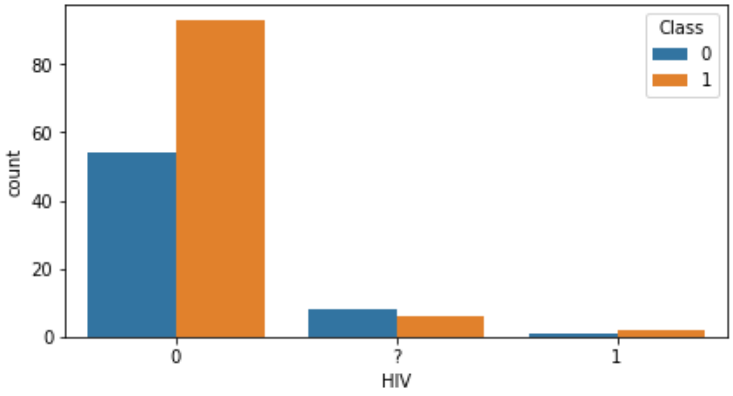
- η ύπαρξη ή όχι αρτηριακής υπέρτασης στην/ον ασθενή (AHT),
  Παρατηρούμε ότι για ασθενείς με αρτηριακή υπέρταση ο λόγος αποβιώσαντων/επιβιώσαντες στο ένα έτος είναι μεγαλύτερος από ό,τι για ασθενείς χωρίς αρτηριακή υπέρταση. Άρα η **αρτηριακή υπέρταση μπορεί να είναι δείκτης κακής πρόγνωσης**, αλλά έχουμε κατά νου ότι η συσχέτιση **δεν συνεπάγεται αναγκαστικά αιτιολογική σχέση**. Είναι πιθανόν να οφείλεται η ύπαρξη υψηλής αρτηριακής πίεσης λόγω παρουσίας άλλων κλινικών γνωρισμάτων των ασθενών που είναι παράγοντες κινδύνου λχ. Στην παχυσαρκία.



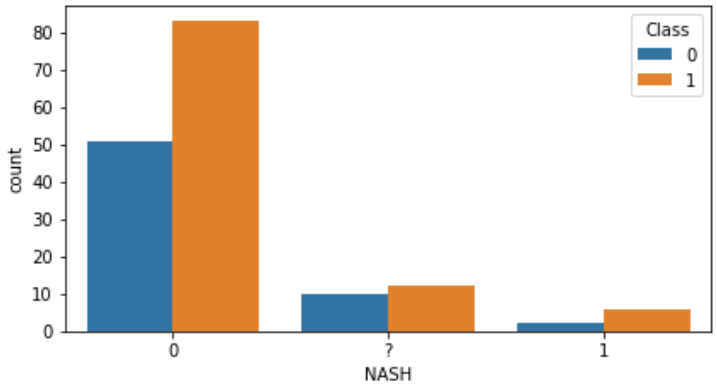
- η ύπαρξη ή όχι χρόνιας νεφρικής ανεπάρκειας στην/ον ασθενή (CRI),
  οι ασθενείς χωρίς χρόνια νεφρική ανεπάρκεια έχουν μεγαλύτερο λόγο αποβιώσαντων/επιβιώσαντες στο 1 έτος σε σχέση με τους ασθενείς με νεφρική ανεπάρκεια. Ωστόσο υπάρχουν λίγοι ασθενείς με νεφρική ανεπάρκεια και ίσως να μην εκπροσωπούνται αρκετά καλά στο δείγμα.



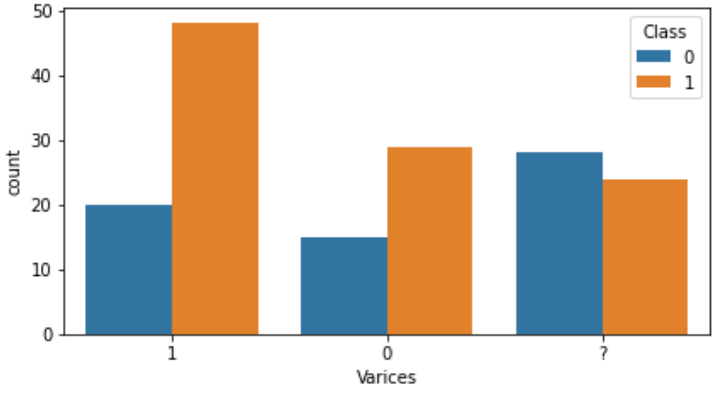
- η μόλυνση από τον ιό της Ανθρώπινης Ανοσοανεπάρκειας στην/ον ασθενή ή η έλειψή της (HIV),
  Έχουμε πολύ λίγες καταγραφές με HIV, ενώ πάνω από διπλάσιες από αυτές είναι μη καταγεγραμένες ως προς το χαρακτηριστικό αυτό. Συνεπώς δεν βγάζουμε ασφαλή συμπεράσματα. Ωστόσο η οροθετικότητα για HIV είναι παράγοντας κινδύνου όπως είδαμε στην εισαγωγή.



- η ύπαρξη ή όχι μη-αλκοολικής στεατοηπατίτιδας (NASH),
  Κάτι παρόμοιο ισχύει όπως και με το χαρακτηριστικό HIV. Η ύπαρξη μη-αλκοολικής στεατοηπατίτιδας ωστόσο είναι αναγνωρισμένος παράγοντας κινδύνου για εκδήλωση ηπατοκυταρρικού καρκινώματος, όπως είδαμε στην εισαγωγή. Δεν μπορούν να βγουν ιδιαίτερα ασφαλή συμπεράσματα από το σύνολο που έχουμε, αλλά βλέπουμε ότι ο λόγος αποβιώσαντων / επιβιώσαντες μετά το ένα έτος είναι μεγαλύτερος για άτομα με μη-αλκοολική στεατοηπατίτιδα από ό,τι για άτομα χωρίς.

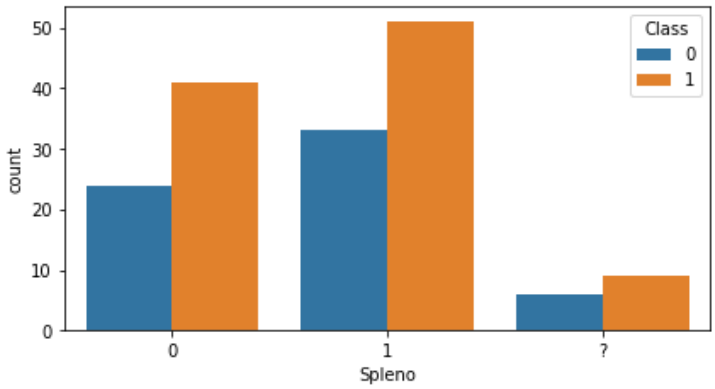


- η ύπαρξη ή όχι οισοφαγικών αλλοιώσεων (Varices),
  Και σε αυτό το χαρακτηριστικό οι καταγραφές με τιμές μη συμπληρωμένες είναι πάρα πολλές. Δεν μπορούν συνεπώς να εξαχθούν χρήσιμα συμπεράσματα. Το χαρακτηριστικό αυτό δεν είναι συμπληρωμένο σχεδόν εξίσου, αλλά λίγο περισσότερο σε επιβιώσαντες μετά το ένα έτος. Ωστόσο από την πρότερη γνώση που υπάρχει όσον αφορά τους παράγοντες κινδύνου ηια εκδήλωση ηπατοκυτταρικού καρκινώματος (βλ. Εισαγωγή) αναφέρουμε ότι η παρουσία οισοφαγικών αλλοιώσεων είναι παράγοντας κινδύνου. Ο λόγος αποβιώσαντων / επιβιώσαντες μετά το ένα έτος είναι μεγαλύτερος παρουσία οισοφαγικών αλλοιώσεων, από ό,τι απουσία τους.



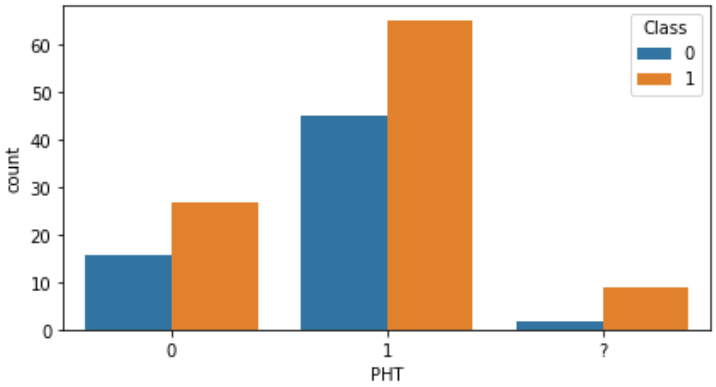
- η ύπαρξη ή όχι Σπληνομεγαλίας (Spleno),

Αναφορικά με την Σπληνομεγαλία βλέπουμε ότι και πάλι οι καταγραφές που είναι μη συμπληρωμένες αρκούν να αλλάξουν τις αναλογίες αποβιώσαντων/επιβιώσαντες μετά το ένα έτος. Ως εκ τούτου, ούτε εδώ μπορούμε να βγάλουμε κάποιο χρήσιμο συμπέρασμα, χωρίς να συμβουλευθούμε το ('EASL-EORTC clinical practice guidelines: management of hepatocellular carcinoma.', 2012). Βάσει των (Wu et al., 2012), *οι ασθενείς με σπληνομεγαλία έχουν σχετικά φτωχότερο λειτουργικό απόθεμα ήπατος από εκείνους με φυσιολογικό όγκο σπλήνας. Η σπληνομεγαλία είναι ένας ανεξάρτητος παράγοντας κινδύνου που προβλέπει τη συνολική επιβίωση για ασθενείς με μικρό HCC που υποβάλλονται σε RFA*. Παρατηρούμε ότι οι καταγραφές που δεν είναι συμπληρωμένες διαταρράσουν την αναμενόμενη κατανομή.



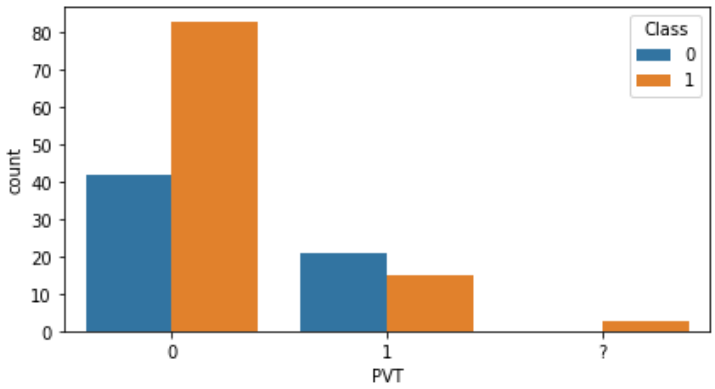
- η ύπαρξη ή όχι υπέρτασης πυλαίας φλέβας (PHT),

Δεν φαίνεται να υπάρχει μεγάλη διαφορά στον λόγο αποβιώσαντων/επιβιώσαντων μετά το ένα έτος ως προς αυτό το χαρακτηριστικό, όμως βλέπουμε ότι υπάρχουν καταγραφές μη συμπληρωμένες που θα μπορούσαν να μεταβάλλουν κάπως τις αναλογίες. **Πιο συχνά λείπει η καταγραφή του χαρακτηριστικού αυτού στους αποβιώσαντες μετά το ένα έτος. Θα ήταν χρήσιμο να δρομολογηθεί η καταγραφή του νωρίτερα σε επόμενες συλλογές δεδομένων, ώστε να μην μένει μη συμπληρωμένο.** Βάσει των ('EASL-EORTC clinical practice guidelines: management of hepatocellular carcinoma.', 2012) πρόκειται για παράγοντα κινδύνου για την εμφάνιση Ηπατοκυτταρικού καρκινώματος.

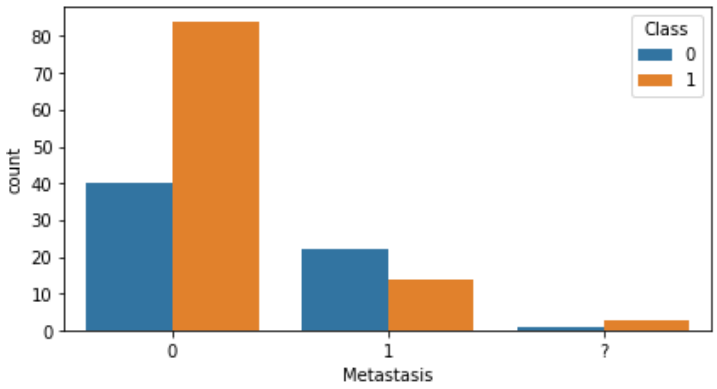


Από την εργασία των (Moriguchi, Furuta and Itoh, 2017) βλέπουμε ότι ο *θρόμβος όγκου της πυλαίας φλέβας (PVTT) εμφανίζεται συχνά με την εξέλιξη του ηπατοκυτταρικού καρκινώματος (HCC) και είναι ένας σημαντικός παράγοντας στον καθορισμό της πρόγνωσης του HCC. Σε πολλές περιπτώσεις HCC με προχωρημένο PVTT, η θεραπεία είναι δύσκολη επειδή ο όγκος έχει σημαντική επέκταση στο ήπαρ και η πυλαία υπέρταση είναι μια συχνή επιπλοκή.*

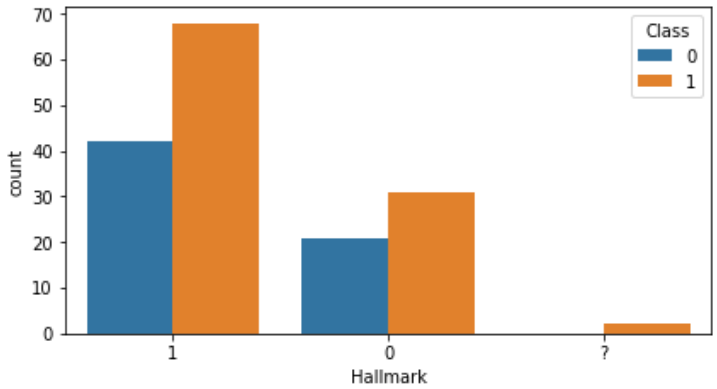
- η ύπαρξη ή όχι θρόμβωσης πυλαίας φλέβας (PVT),
  Ως προς το συγκεκριμένο χαρακτηριστικό λείπουν πολύ λίγες καταγραφές (μόνο από αποβιώσαντες). Για ασθενείς χωρίς θρόμβωση πυλαίας φλέβας ο λόγος αποβιώσαντων /επιβιώσαντων είναι κοντά στο 2. Αντίθετα για ασθενείς με θρόμβωση πυλαίας φλέβας ο ίδιος λόγος είναι λίγο μικρότερος του 1. Κρατάμε βέβαια με την επιφύλαξη ως προς το πόσο θα μπορούσαν να μεταβληθούν οι αναλογίες αν είχαν συμπληρωθεί πλήρως όλες οι καταγραφές και δεν έλειπαν τιμές. Παρόλα αυτά βλέπουμε ότι για απουσία θρόμβωσης οι ασθενείς έχουν χειρότερη έκβαση στο σετ που έχουμε.
  .



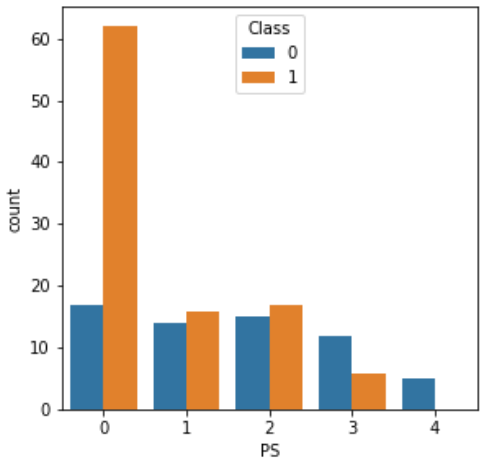
- η ύπαρξη ή όχι μετάστασης (Metastasis),
  Για τους ασθενείς χωρίς μετάσταση, ~ 2/3 πεθαίνουν μετά το ένα έτος. Για τους ασθενείς με μετάσταση αυτοί που ζουν είναι περισσότεροι. Κάτι τέτοιο είναι παράδοξο. Πρέπει να σημειωθεί ότι υπάρχουν διάφοροι τύποι μεταστάσεων και θα ήταν χρησιμότερο να γίνει μελέτη αφού πρώτα γίνει ο διαχωρισμός τους.



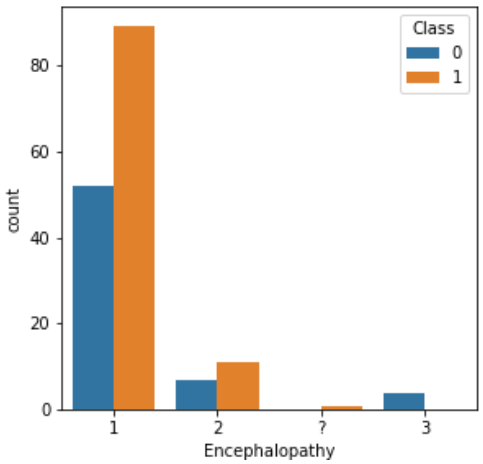
- το ακτινολογικό σήμα (Hallmark),
  Ως προς την ύπαρξη χαρακτηριστικού ακτινολογικού σήματος (υπεραγγειακό στην αρτηριακή φάση με έκπλυση στην πυλαία φλεβική ή καθυστερημένες φάσεις) (‘EASL-EORTC clinical practice guidelines: management of hepatocellular carcinoma.’, 2012) στην διάγνωση, ο λόγος αποβιώσαντων / επιβιώσαντες παρουσία του είναι αυξημένος σε σχέση με απουσία του.



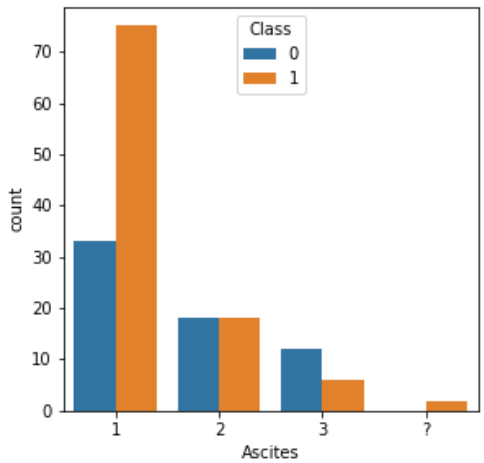
- η κατάσταση απόκρισης (PS),
  Παρατηρούμε ότι όσο καλύτερη είναι η κατάσταση απόκρισης στην θεραπεία, τόσο μικραίνει ο λόγος αποβιώσαντων / επιβιώσαντες μετά το ένα έτος. Βλέπουμε μεγάλη διαφορά του λόγου για PS = 3, PS = 4. Κάτι τέτοιο αναμένεται.



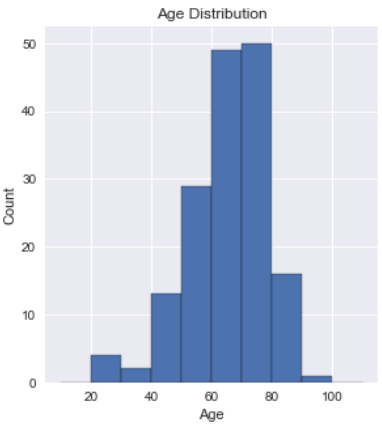
- ο βαθμός εγκεφαλοπάθειας (Encephalopathy),
  Παρατηρούμε ότι ο λόγος αποβιώσαντων / επιβιώσαντες είναι αυξημένος για βαθμό εγκεφαλοπάθειας = 1, ενώ μειώνεται όσο αυξάνει ο βαθμός εγκεφαλοπάθειας. Από την εργασία των (Yoneyama et al., 2004) βλέπουμε ότι η *ύπαρξη ηπατικής εγκαφαλοπάθειας είναι ενδεικτική της ανάπτυξης κίρρωσης του ήπατος*. Όπως είδαμε στην εισαγωγή η ύπαρξη κίρρωσης είναι πολύ σημαντικός παράγοντας κινδύνου για την εκδήλωση ηπατοκυτταρικού καρκίνου.

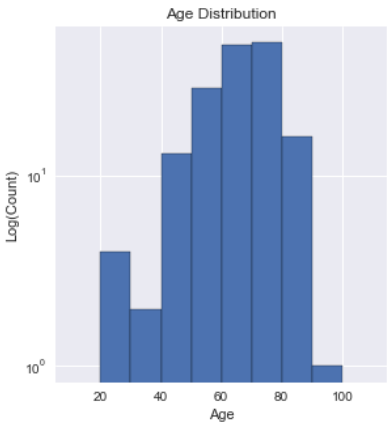


- ο βαθμός ανάπτυξης ασκίτη (Ascites),
  Όσο ο βαθμός ανάπτυξης ασκίτη είναι μικρός, τόσο αυξημένος είναι ο λόγος αποβιώσαντων / επιβιώσαντες. *Ο ασκίτης εμφανίζεται συχνά σε ασθενείς με HCC και σχετίζεται τόσο με καρκινικούς παράγοντες όσο και με παράγοντες κίρρωσης και με μειωμένη μακροχρόνια επιβίωση.* (Hsu et al., 2013) *Η σοβαρότητά του συσχετίζεται σημαντικά με την υπερχολερυθριναιμία, την υπολευκωματιναιμία, την υπονατριαιμία, την παράταση του χρόνου προθρομβίνης (PT) και τη* ***νεφρική ανεπάρκεια****.* (Hsu et al., 2013) *Μεγάλο φορτίο όγκου και συχνότερη αγγειακή διήθηση παρατηρήθηκαν συχνά σε ασθενείς με πιο σοβαρό ασκίτη. Ο ασκίτης είναι αναγνωρισμένος ανεξάρτητος προγνωστικός παράγοντας με 80-94% αυξημένο κίνδυνο θνησιμότητας*. (Hsu et al., 2013)



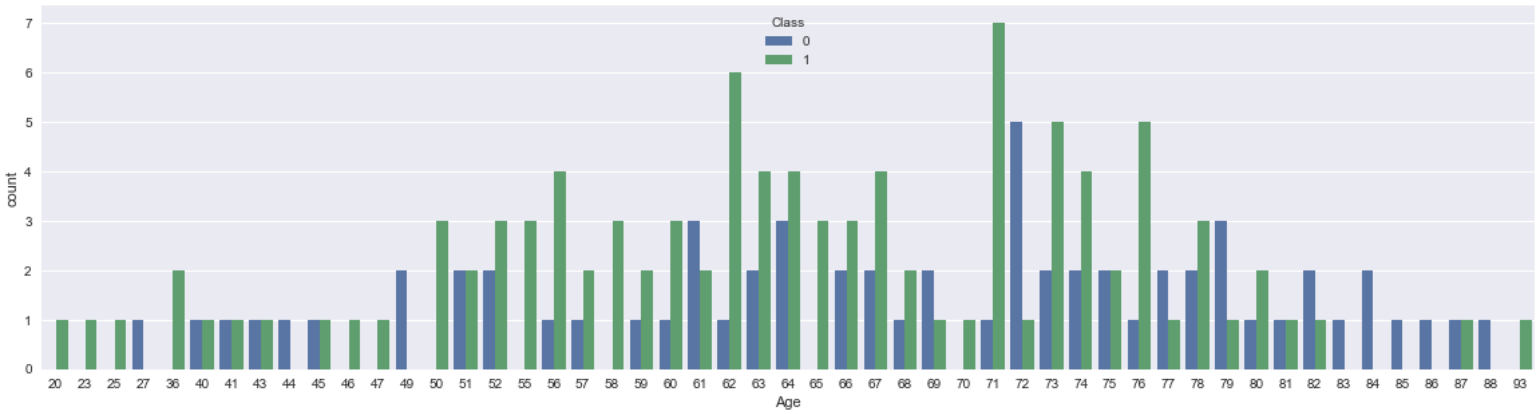
- η Ηλικία κατά την διάγνωση (σε έτη) (Age),





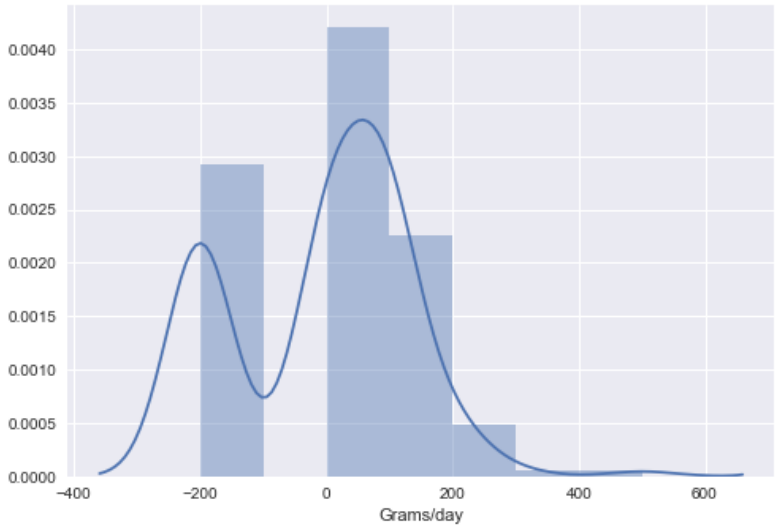
Το αριστερό διάγραμμα στον άξονα ψ έχει τον αριθμό των ατόμων που ανήκουν σε έκαστη ηλικιακή κλάση. Το δεξί διάγραμμα έχει στον άξονα ψ τον λογάριθμο του αριθμού αυτού. Η κατανομή συχνοτήτων είναι σχεδόν κανονική.
Ωστόσο βλέπουμε ότι δημιουργείται «ώμος» στην ηλικιακή κλάση [20,30).
Αυτός ο ώμος είναι καλύτερα ορατός στο δεξί διάγραμμα.

Το ακόλουθο σύνθετο ραβδόγραμμα παρουσιάζει τον αριθμό των καταγραφών που υπάρχουν για κάθε ηλικιακή τιμή, χωρισμένες ανάλογα με την έκβασή τους μετά το ένα έτος. Με πράσινο παρουσιάζονται οι αποβιώσαντες και με μπλε οι επιβιώσαντες. Παρατηρούμε ότι για ηλικίες κάτω των 40 οι αποβιώσαντες υπερισχύουν, ενώ για ηλικίες άνω των 80 οι επιβιώσαντες είναι αυτοί που είναι περισσότεροι. Προφανώς κάτι τέτοιο μπορεί να σχετίζεται με ετερογένεια στην ασθένεια, στην επιθετικότητά της, καθώς και στο αν έγινε εγκαίρως αντιληπτή σε κάποιο έλεγχο ρουτίνας, ή όταν πια είχε χαθεί πολύτιμος χρόνος.

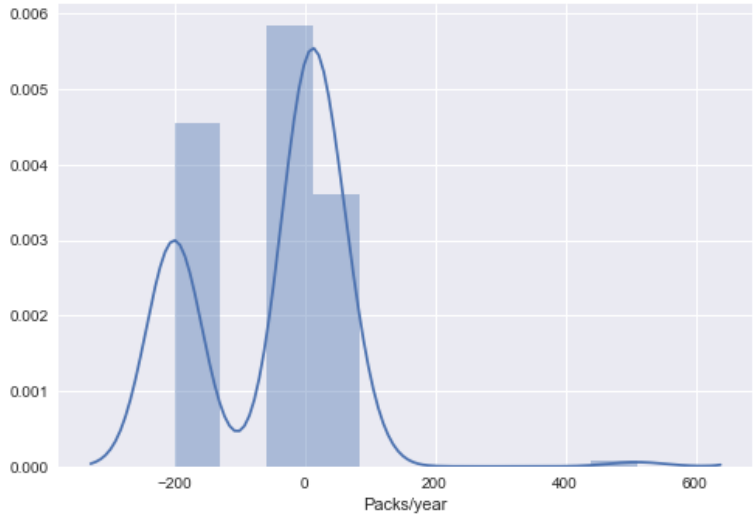


Όπως πληροφορούμαστε από την ομάδα των (Wang et al., 2015*), η ηλικία είναι ένας πολύπλοκος προγνωστικός παράγοντας στο HCC και μπορεί να διαδραματίσει παράδοξο ρόλο στην πρόγνωση των ασθενών με HCC. Οι νεότεροι ασθενείς στο κέντρο τους έτειναν να παρουσιάζουν πιο επιθετικούς όγκους και είχαν μεγαλύτερο κίνδυνο υποτροπής.* (Wang et al., 2015) Κάτι τέτοιο συνάδει και με τα στοιχεία που παρουσιάζονται εδώ.

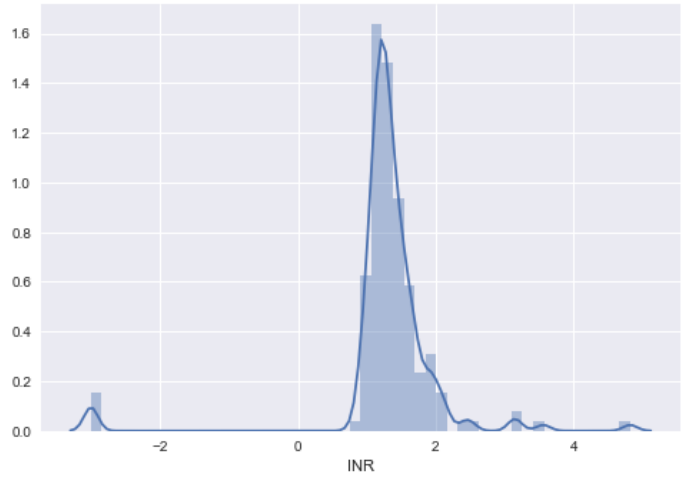
- τα Γραμμάρια αλκοόλ/μέρα (Grams/day),
  Η κατανομή συχνοτήτων παρατηρούμε ότι κάνει ένα μικρό ώμο δεξιά, για γραμμάρια >400 / μέρα. Για λόγους γραφικής αναπαράστασης και του ύψους των τιμών που λείπουν, επιλέχθηκε (μόνο για το διάγραμμα να τους δωθεί «προσωρινή» τιμή = -200, η οποία και δεν υιοθετήθηκε στις αναλύσεις. Φαίνεται ότι λείπουν πολλά δεδομένα ως προς το συγκεκριμένο χαρακτηριστικό. *Οι ασθενείς με ηπατοκυτταρικό καρκίνωμα που σχετίζεται με το αλκοόλ φάνηκε να έχουν χειρότερη πρόγνωση από εκείνους με νόσο που δεν σχετίζεται με το αλκοόλ. Ωστόσο, η συνολική επιβίωση δεν διέφερε μεταξύ των ομάδων όταν το στάδιο του όγκου κατά τη διάγνωση ήταν το ίδιο, γεγονός που υποδηλώνει ότι το στάδιο της διάγνωσης είναι ένας σημαντικός οδηγός επιβίωσης*. (Charlotte E. Costentin, 2018)



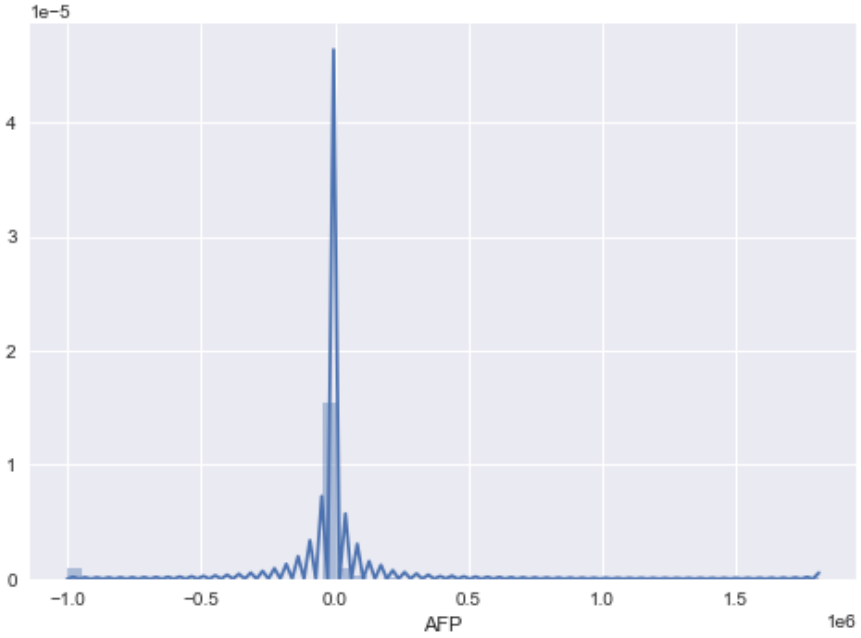
- τα Πακέτα τσιγάρων/έτος (Packs/year),
  Το ποσοστό των καταγραφών που λείπουν (-200) είναι αρκετά υψηλό. Η πλειοψηφία των καταγραφών βρίσκεται στο διάστημα 0-200 πακέτων/έτος (με φθίνουσα καθώς ανεβαίνει ο αριθμός των πακέτων. Υπάρχει όμως και εδώ μία ομάδα που ξεχωρίζει και φτάνει στο διάστημα 400-600. Το κάπνισμα επηρεάζει τη μακροπρόθεσμη έκβαση σε ασθενείς με Ηπατοκυτταρικό καρκίνωμα. Προκαλεί *δυσμενείς επιπτώσεις στο ήπαρ: άμεσες και έμμεσες τοξικές επιδράσεις, ανοσολογικές επιδράσεις (κατασταλτικές) και ογκογόνες επιδράσεις. Το κάπνισμα είναι κύρια πηγή 4-αμινοδιφαινυλίου, ενός ηπατικού καρκινογόνου που έχει ενοχοποιηθεί ως αιτιολογικός παράγοντας κινδύνου για HCC. Το κάπνισμα παράγει χημικές ουσίες με κυτταροτοξικό δυναμικό που αυξάνουν τη νεκροφλεγμονή και την ίνωση. Το κάπνισμα αυξάνει την παραγωγή προφλεγμονωδών κυτοκινών (IL-1, IL-6 και TNF-α) που θα μπορούσαν να εμπλέκονται στη βλάβη των ηπατικών κυττάρων. Συμβάλλει έμμεσα στην υπερφόρτωση σιδήρου και διαταρραζή της ομοιόστασής του, καθώς και στην υπερβολική παραγωγή ουρικού οξέος*. (El-Zayadi, 2006) *Το βαρύ κάπνισμα (PY ≥ 20) ήταν ο πιο σημαντικός παράγοντας που σχετίζεται με την υποτροπή HCC που σχετίζεται με τον HBV μετά από θεραπευτική χειρουργική εκτομή (p = 0,001), ακολουθούμενη από θεραπεία κατά του HBV (p<0,01), τρέχον κάπνισμα (p = 0,028), χειρουργική περιθώριο <1 cm (p = 0,048) και μετάγγιση αίματος >600 ml (p = 0,028)*. (Zhang et al., 2014)



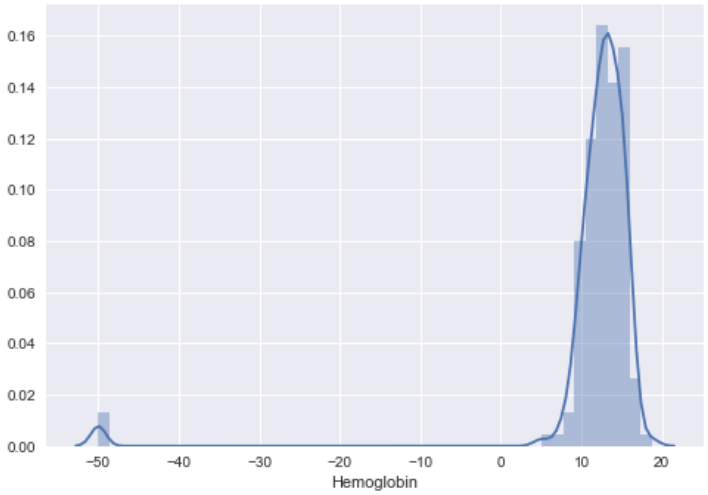
- ο Διεθνής Κανονικοποιημένος Λόγος (INR),
  μια δοκιμή που χρησιμοποιείται για να μετρήσει πόσο γρήγορα το αίμα σχηματίζει θρόμβο, σε σύγκριση με τον κανονικό χρόνο πήξης. (U.S. Department of Veterans Affairs, INR (international normalized ratio) Hepatitis C for Patients) Παρατηρούμε ότι κι εδώ υπάρχει «ώμος» στην δεξιά πλευρά της κατανομής κοντά στο INR = 2. Επιπλέον υπάρχουν και καταγραφές για INR >2, έως και λίγο πριν το 6. Οι τιμές που λείπουν έχουν παρασταθεί γραφικά με την τιμή -3.



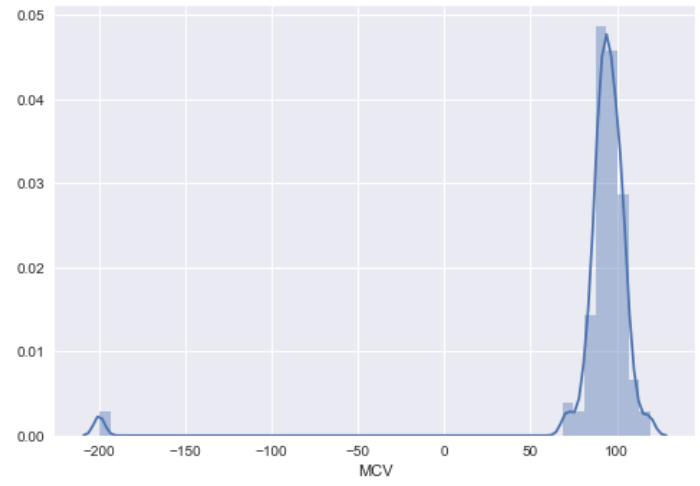
- η α-εμβρυϊκή πρωτεΐνη (ng/mL) (AFP),
  Οι τιμές που λείπουν έχουν παρασταθεί γραφικά με αρνητική τιμή AFP. Βλέπουμε ότι η κατανομή φθίνει, όσο αυξάνεται η AFP, θυμίζοντας σχηματικά υπερβολή, ενώ έχει -προφανώς- μόνο θετικές τιμές. *Η άλφα-εμβρυοπρωτεΐνη (AFP) είναι ο σημαντικότερος διαγνωστικός και προγνωστικός δείκτης του ηπατοκυτταρικού καρκινώματος (HCC). Η AFP-θετικό HCC μπορεί εύκολα να διαγνωστεί με βάση το επίπεδο AFP στον ορό και τα τυπικά απεικονιστικά χαρακτηριστικά, αλλά ένας αριθμός ασθενών με HCC είναι "αρνητικοί" (AFP < 20 ng/mL) για AFP*. (Liu et al., 2021)



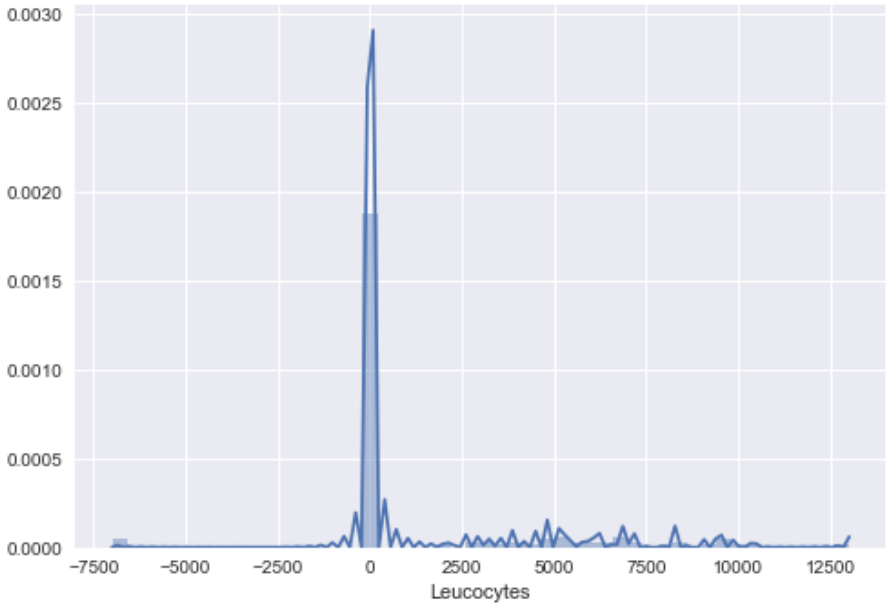
- η Αιμοσφαιρίνη (g/dL) (Hemoglobin),
  Η αιμοσφαιρίνη έχει μικρό «ώμο» στα αριστερά. Πιθανόν θα μπορούσαμε να πούμε ότι έχει λόξωση (skewed). Οι τιμές που λείπουν έχουν αναπαρασταθεί με αιμοσφαιρίνη = -50. Από την εργασία των (Finkelmeier et al., 2013) *η αναιμία είναι μια κοινή επιπλοκή σε διάφορους τύπους καρκίνου συμπεριλαμβανομένου του ηπατοκυτταρικού καρκινώματος (HCC) και τα χαμηλά επίπεδα Hb συσχετίστηκαν με τη θνησιμότητα ανεξάρτητα από το στάδιο του όγκου, την ηλικία, το φύλο και τα επίπεδα της C-αντιδρώσας πρωτεΐνης σε ένα πολυπαραγοντικό μοντέλο παλινδρόμησης Cox. Η αναιμία θεωρείται ως παράγοντας κινδύνου για θνησιμότητα σε ασθενείς με HCC*. Είναι πιθανό ο ώμος που παρουσιάζεται στην καμπύλη να αντιστοιχεί σε αναιμικούς ασθενείς.



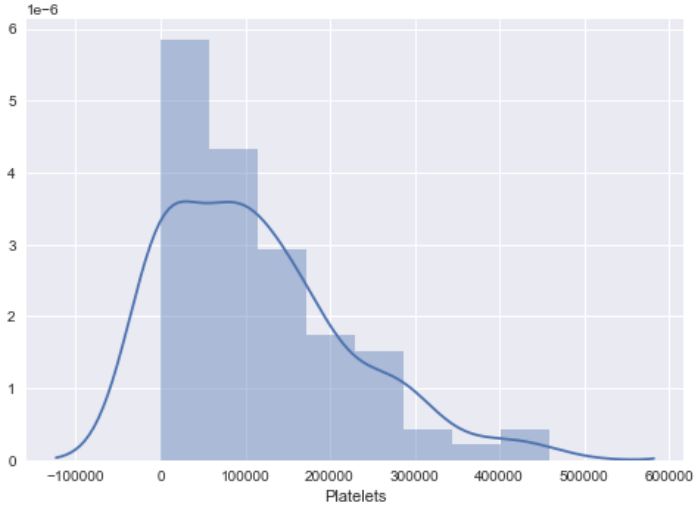
- ο Μέσος σωματικός όγκος (fl) (MCV),
  Παρόμοιο μοτίβο με την αιμοσφαιρίνη παρουσιάζει και η κατανομή για το χαρακτηριστικό μέσος όγκος του όγκου. Οι τιμές που λείπουν πήραν (για την γραφική αναπαράστασή τους) την προσωρινή τιμή MCV = -200. Στην εργασία των (Yoon et al., 2016) *το αυξημένο επίπεδο MCV σε μη αναιμικά άτομα χωρίς καρκίνο συσχετίστηκε με αυξημένη θνησιμότητα από όλες τις αιτίες τόσο σε άνδρες όσο και σε γυναίκες και με θνησιμότητα από καρκίνο, ιδιαίτερα με θνησιμότητα από καρκίνο του ήπατος στους άνδρες*.



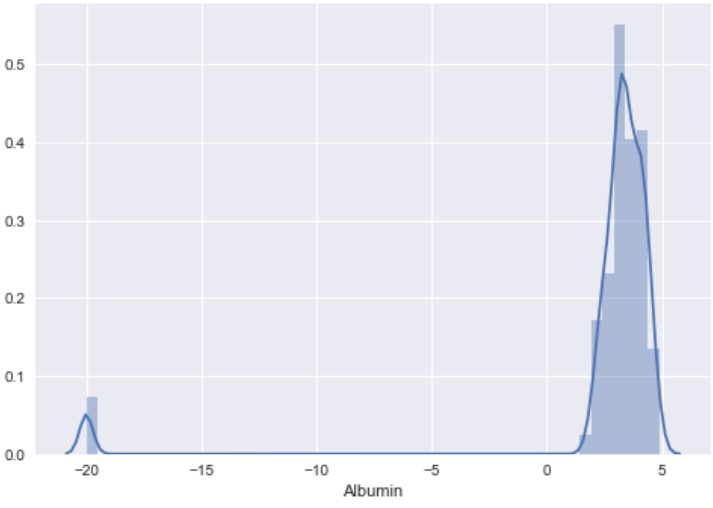
- τα Λευκοκύτταρα (G/L) (Leucocytes),
  Βάσει των (Liu et al., 2012) *το μήκος των τελομερών στα λευκοκυττάρα προβλέπει τη συνολική επιβίωση στο ηπατοκυτταρικό καρκίνωμα που αντιμετωπίζεται με διαρτηριακό χημειοεμβολισμό*. Οι (Lee et al., 2018) εντόπισαν *οκτώ μονονουκλεοτιδικούς πολυμορφισμούς ενός νουκλεοτιδίου (SNPs) στην γονιδιακή περιοχή του ανθρώπινου αντιγόνου λευκοκυττάρου (HLA), που συσχετίστηκαν με HCC, πλησίον του HLA-DQB1 (HLA-DQB1*03:01, DQB1*06:02). Το DQB1*03:01 έδειξε προστατευτικά αποτελέσματα μόνο σε ασθενείς με γονότυπο 1 HCV. Το DQB1*06:02 προσέφερε κίνδυνο HCC μόνο σε ασθενείς με γονότυπους HCV non-1. Το HLA-DRB1*15:01, το οποίο βρίσκεται σε ανισορροπία σύνδεσης με το DQB1*06:02, αύξησε επίσης τον κίνδυνο HCC*. (Lee et al., 2018) *Η λευκοκυτταρική χημειοταξίνη 2 (LECT2) ρυθμίζεται από τη β-κατενίνη στο HCC τόσο σε ποντίκια όσο και σε άνδρες, αλλά η LECT2 του ορού αντανακλά τη δραστηριότητα της β-κατενίνης μόνο σε ποντικούς. Η LECT2 ορού θα μπορούσε να είναι ένας πιθανός βιοδείκτης του HCC σε ασθενείς*. (Okabe et al., 2014) *Αλλά και γενικά η εγγύτητα των Τ-λεμφοκυττάρων και των Β-λεμφοκυττάρων που διεισδύουν στον όγκο υποδηλώνει μια λειτουργική αλληλεπίδραση μεταξύ τους που συνδέεται με ενισχυμένη τοπική ανοσολογική ενεργοποίηση και συμβάλλει στην καλύτερη πρόγνωση για ασθενείς με HCC* (Garnelo et al., 2017) Στο δείγμα η πλειοψηφία των καταγραφών έχει χαμηλές τιμές λευκοκυττάρων, ωστόσο υπάρχει μία αξιοσημείωτη ουρά προς τα δεξιά, τέτοια που κάνει το σχήμα της κατανομής να θυμίζει οπτικά υπερβολή. Οι τιμές που λείπουν έχουν παρασταθεί γραφικά με αρνητική τιμή Leucocytes = -7500. Φαίνεται καθαρά η διαφοροποίηση που υπάρχει μεταξύ των ασθενών ως προς το πλήθος των λευκοκυττάρων τους.



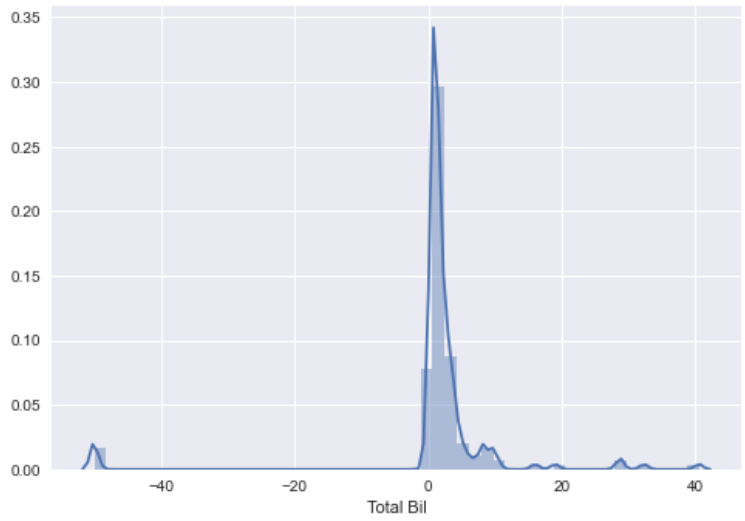
- τα Αιμοπετάλια (G/L) (Platelets),
  Τα αιμοπετάλια προάγουν τον πολλαπλασιασμό και την εισβολή των κυττάρων HCC, αλλά η εμπλοκή τους υπερβαίνει την άμεση επίδραση στα καρκινικά κύτταρα, καθώς είναι γνωστό ότι παίζουν ρόλο στην προ-ινωδογόνο σηματοδότηση και την ηπατική ανοσοαπόκριση, καθώς και στη μεσολάβηση των αλληλεπιδράσεων μεταξύ αυτών των παραγόντων στο στρώμα. Τα αιμοπετάλια έχει επίσης παίζουν καθοριστικό ρόλο στην αναγέννηση του ήπατος μετά από βλάβη. (Pavlovic et al., 2019) Έτσι σχετίζονται με την έκβαση των ασθενών με Ηπατοκυτταρικό καρκίνωμα. Οι καταγραφές με λίγα αιμοπετάλια είναι οι περισσότερες και καθώς αυξάνει ο αριθμός των αιμοπεταλίων μειώνονται. Η κατανομή φαίνεται να έχει «ώμο» περί τα > 400000 Αιμοπετάλια και περί τα >30000.



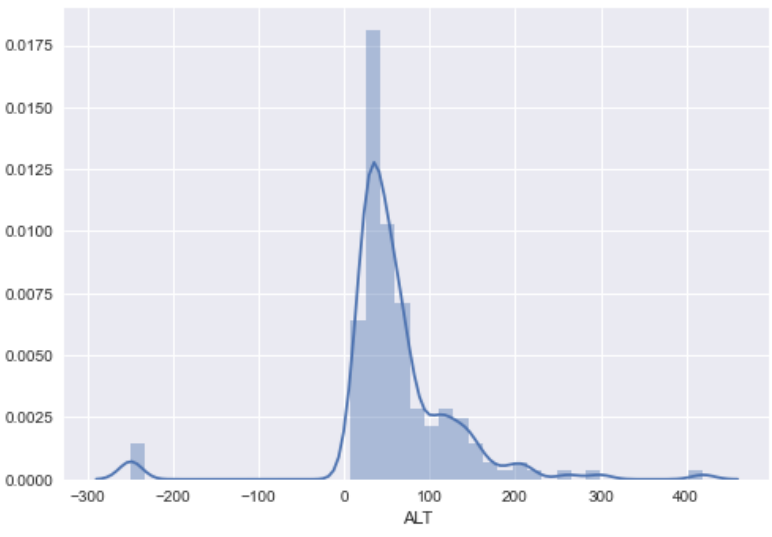
- η Λευκωματίνη (mg/dL) (Albumin),
  Η λευκωματίνη έχει παρόμοια κατανομή με την Αιμοσφαιρίνη και τον Μέσο όγκο του όγκου. Οι τιμές που λείπουν έχουν πάρει για το διάγραμμα προσωρινή τιμή Albumin = -20. Σύμφωνα με τους (Carr and Guerra, 2017) τα *χαμηλά επίπεδα λευκωματίνης ορού συσχετίζονται με αυξημένες μετρήσεις παραμέτρων της επιθετικότητας του HCC, πέραν του ρόλου τους στην παρακολούθηση της συστηματικής φλεγμονής. Η αναλογία αλβουμίνης προς αλκαλική φωσφατάση (AAPR) είναι ανεξάρτητος προγνωστικός δείκτης (από απλή, χαμηλού κόστους εξέταση ρουτίνας αίματος) για ασθενείς με HCC για συνολική και χωρίς νόσο επιβίωση (ανεξάρτητα από τις θεραπευτικές επιλογές)* (Chan et al., 2015)



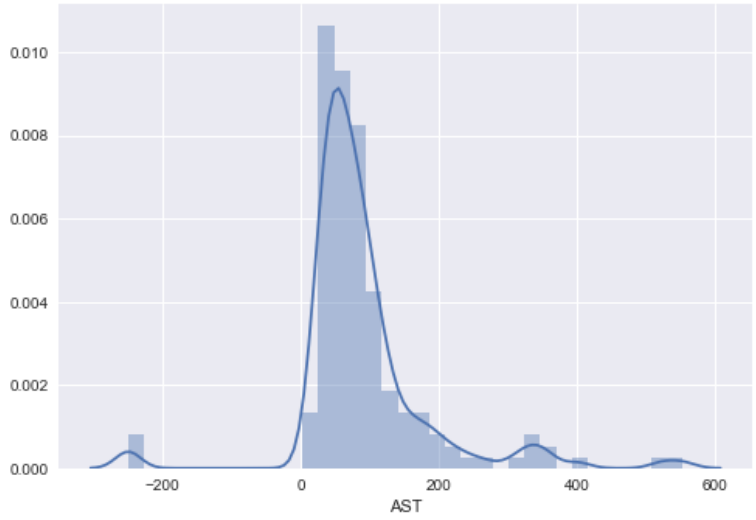
- η Συνολική Χολεριθρίνη (mg/dL) (Total Bil),
  Οι τιμές που λείπουν αναπαριστώνται προσωρινά στο διάγραμμα με αρνητική τιμή Χολερυθρίνης. Βλέπουμε ότι η κατανομή είναι ασύμμετρη, με κορυφή προς τα αριστερά και ουρά στα δεξιά (για τιμές από ~10 έως ~40). Εκτιμάται ότι αφορούν διαφορετικές κατηγορίες συστάδων ασθενών ως προς την έκβαση. *Ο λόγος λευκωματίνης-χολερυθρίνης (ALBI) στην αρχή, αλλά και η αλλαγή του ALBI κατά τη διάρκεια της θεραπείας θα μπορούσαν να προβλέψουν την πρόγνωση ασθενών με προχωρημένο HCC που έλαβαν sorafenib*. (Kuo et al., 2017) *Επίσης αυξημένα επίπεδα χολερυθρίνης συσχετίστηκαν με υψηλότερα επίπεδα α-εμβρυϊκής πρωτεΐνης (AFP), αυξημένη θρόμβωση της πυλαίας φλέβας (PVT), πολυεστιακότητα και χαμηλότερη επιβίωση, σε ασθενείς με μικρούς και μεγαλύτερους όγκους*. (Carr et al., 2014)



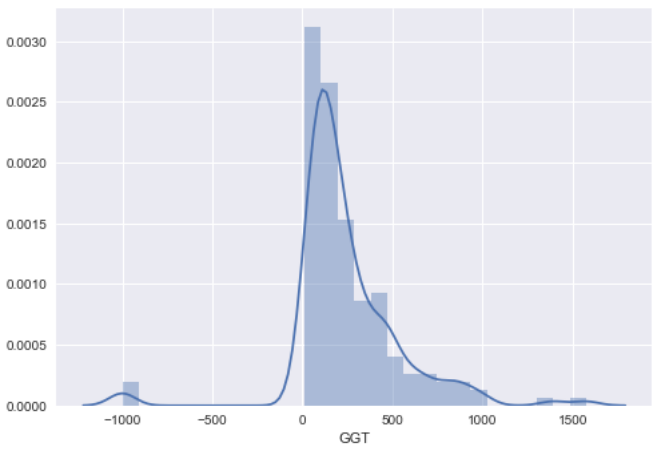
- η Τρανσαμινάση αλανίνης (U/L) (ALT),
  Η αναλογία ουδετερόφιλων προς λεμφοκύτταρα (NLR) και η αναλογία αμινοτρανσφεράσης ασπαρτικού προς αμινοτρανσφεράση αλανίνης (AAR) είναι βιοδείκτες της ηπατικής ίνωσης και κίρρωσης. Όταν συνδυάζονται για την παραγωγή ενός δείκτη με βάση τη φλεγμονή και τη βαθμολογία ίνωσης, δημιουργούν ανεξάρτητο δείκτη κακής πρόγνωσης σε ασθενείς με HCC που λαμβάνουν διαιαρτηριακό χημειοεμβολισμό. (Liu et al., 2017) Παρατηρούμε ότι για ALT > 100 υπάρχει ευδιάκριτος ώμος στην κατανομή, ενώ υπάρχουν και μερικές καταγραφές με τιμές που είναι >200 μέχρι και λίγο άνω του 400. Οι τιμές που λείπουν αναπαριστώνται προσωρινά στο διάγραμμα με αρνητική τιμή.



- η Τρανσαμινάση ασπαρτικού (U/L) (AST),
  Παρόμοιος ώμος παρατηρείται κι εδώ, για AST >300 έως 400, επίσης υπάρχουν λίγες καταγραφές που είναι λίγο μικρότερες του 600. Οι τιμές που λείπουν αναπαριστώνται προσωρινά στο διάγραμμα με αρνητική τιμή. *Η αναλογία ασπαρτικής αμινοτρανσφεράσης (AST) προς λεμφοκύτταρα (ALRI) > 86,3 συσχετίστηκε σε μεγάλο βαθμό με υψηλότερα ποσοστά Child-Pugh B&C, θρόμβωση όγκου της πυλαίας φλέβας (PVTT) και ασκίτη. Το προγνωστικό νομογράφημα συμπεριλαμβανομένου του ALRI ήταν το καλύτερο στην πρόβλεψη της επιβίωσης 3 μηνών, 6 μηνών, 1 έτους, 2 ετών και συνολικής επιβίωσης μεταξύ των Σύστημα σταδιοποίησης όγκου-κόμβου-μετάστασης (TNM), ALRI, ALRI-TNM νομογράμματων.* (Zhao et al., 2019)

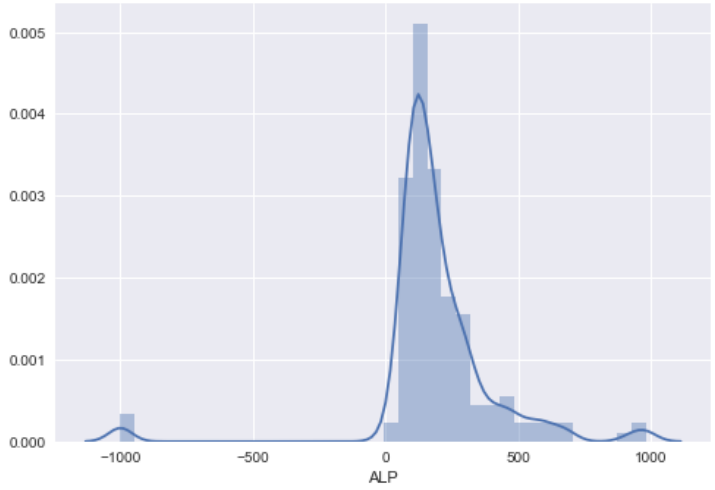


- η γ-γλουταμυλ-τρανσφεράση (U/L) (GGT),
  Σύμφωνα με τους (Liao et al., 2019) *είναι ένα ένζυμο που δρα ως δείκτης για διάφορες μορφές καρκίνου όπως το καρκίνωμα των νεφρών, ο καρκίνος του μαστού, ο καρκίνος του πνεύμονα και ο καρκίνος των ωοθηκών, και τα αυξημένα επίπεδά της μπορεί να είναι ένας προγνωστικός δείκτης για την υποτροπή του HCC μετά από ηπατεκτομή. Η αυξημένη τιμή του λόγου γάμμα-γλουταμυλ τρανπεπτιδάσης προς τον αριθμό των λεμφοκυττάρων (GLR) είχε κακή συνολική επιβίωση (OS) και επιβίωση χωρίς εξέλιξη (PFS) ασθενών Ηπατοκυτταρικού καρκινώματος με μέγεθος όγκου ≤ 5 cm.* (Liao et al., 2019) Κι εδώ παρατηρούμε την ύπαρξη ώμου για GGT ~750 και έως το 1000. Υπάρχουν και λίγες καταγραφές που φτάνουν ως το 1500. Οι τιμές που λείπουν αναπαριστώνται προσωρινά στο διάγραμμα με αρνητική τιμή.



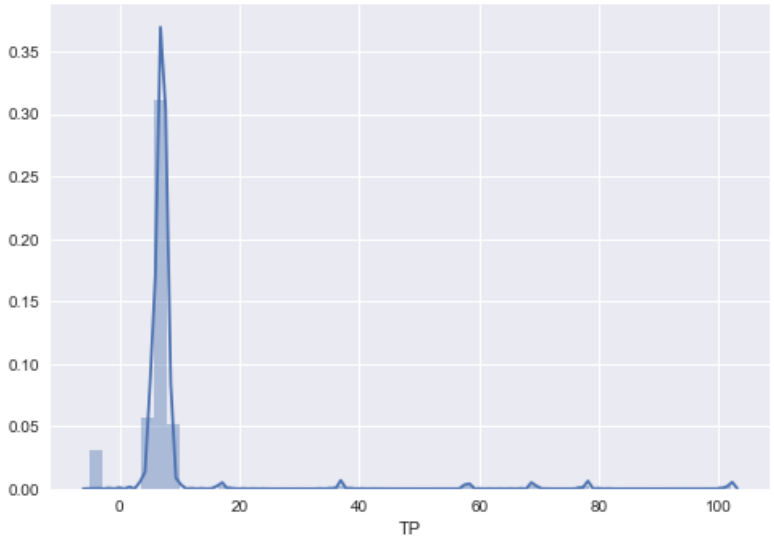
- η αλκαλική φωσφατάση (U/L) (ALP),

Ομοίως κι εδώ, αν και δεν είναι ιδιαίτερα διακριτός ο ώμος παρατηρείται, υπάρχει λόξωση της κατανομής, και για ALP >~300 έως και ~ 700 παρατηρείται πλάτωμα. Υπάρχουν και κάποιες καταγραφές που προσεγγίζουν και την τιμή 1000. Οι τιμές που λείπουν αναπαριστώνται προσωρινά στο διάγραμμα με αρνητική τιμή. *Υπάρχει στενή συσχέτιση μεταξύ του υψηλού επιπέδου ALP πριν από τη θεραπεία και της κακής επιβίωσης σε ασθενείς με HCC, υποδεικνύοντας ότι η ALP μπορεί να χρησιμοποιηθεί ως βιοδείκτης για την πρόγνωση*. (Sun, Chen and Li, 2020)

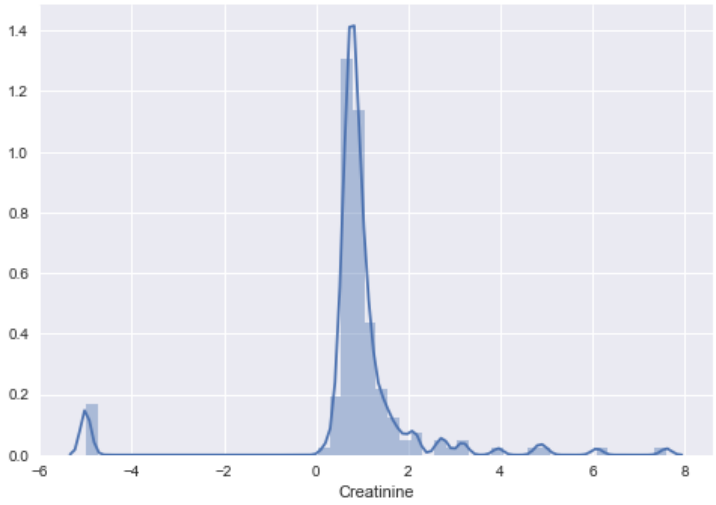


- οι Σύνολο Πρωτεϊνών (g/dL) (TP),

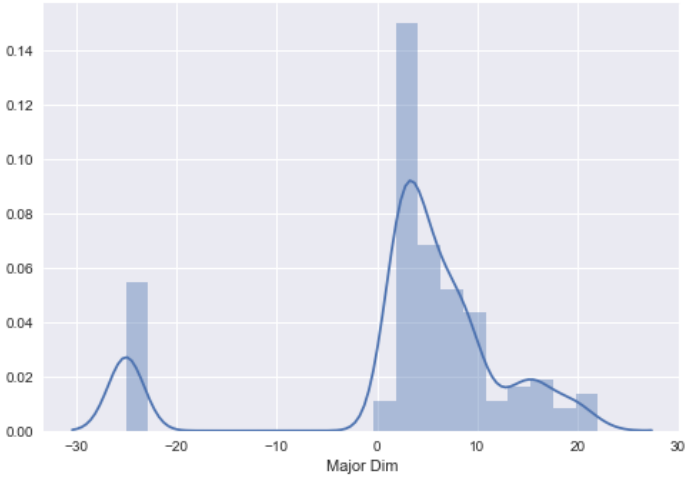
Οι τιμές που λείπουν αναπαριστώνται προσωρινά στο διάγραμμα με αρνητική τιμή. Παρατηρούμε ότι η κατανομή μοιάζει συμμετρική, αν και υπάρχουν λίγες καταγραφές που μπορεί να λάβουν σποραδικές τιμές έως και >100.



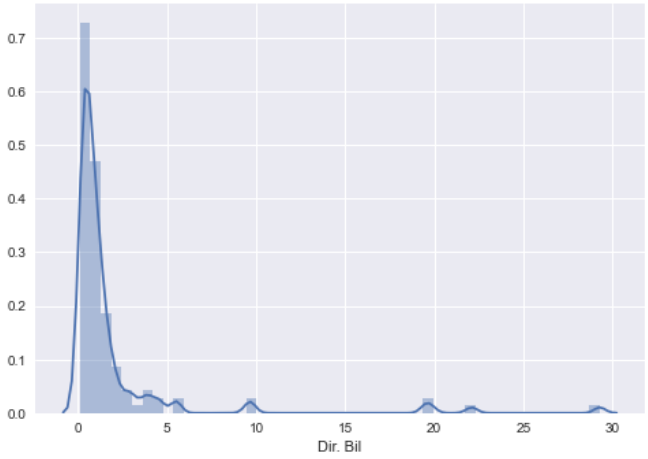
- η Κρεατινίνη (mg/dL) (Creatinine),
  Οι τιμές που λείπουν αναπαριστώνται προσωρινά στο διάγραμμα με αρνητική τιμή. Παρατηρούμε ότι οι τιμές Κρεατινίνης από 2 και άνω δεν φθίνουν με τον ίδιο ρυθμό, αλλά πιο αργά, δημιουργώντας λόξωση στην κατανομή. Επίσης υπάρχουν σποραδικά λίγες καταγραφές στο διάστημα 3 έως 8. Σύμφωνα με τους (Yaghi et al., 2006) *η αυξημένη κρεατινίνη είναι ένας από τους Ανεξάρτητους προγνωστικούς παράγοντες της πρώιμης θνησιμότητας στο HCC*.

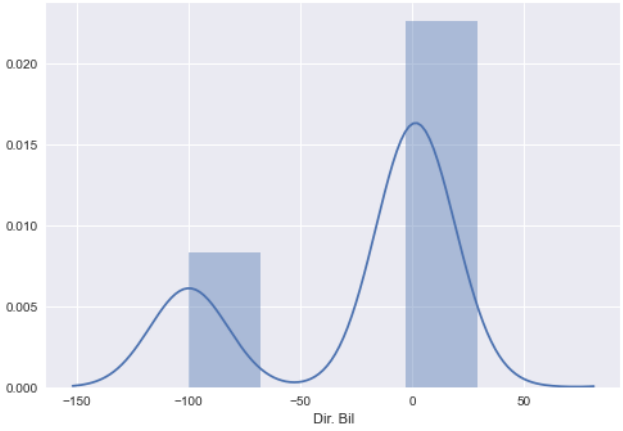


- η Κύρια διάσταση του οζιδίου (cm) (Major Dim),
  Οι τιμές που λείπουν αναπαριστώνται προσωρινά στο διάγραμμα με αρνητική τιμή. Εδώ ο ώμος που παρατηρούσαμε στα άλλα διαγράμματα ξεχωρίζει περισσότερο και θυμίζει ακόμα μία κορυφή. Έχουμε λοιπόν δύο κατανομές κατά πάσα πιθανότητα που κατά την σύνθεσή τους δημιουργούν μία δικόρυφη. Η δεύτερη φαίνεται να έχει κορυφή για Κύρια διάσταση του οζιδίου ~15 και μεγαλύτερη τυπική απόκλιση από την πρώτη. Αναμένουμε ότι όσο μεγαλύτερο είναι το οζίδιο τόσο μεγαλύτερος θα είναι ο κίνδυνος να αποβιώσει το άτομο που νοσεί, ωστόσο εμπλέκονται κι άλλοι παράγοντες με εξίσου σημαντική επίδραση, οπότε κάτι τέτοιο δεν είναι απόλυτο. *Το μέγεθος του όγκου συσχετίζεται με την επιβίωση, είναι προγνωστικός παράγοντας μακροπρόθεσμης επιβίωσης και υποτροπής όγκου σε ασθενείς με ηπατοκυτταρικό καρκίνωμα μετά από ηπατεκτομή*. (Lau et al., 1998)

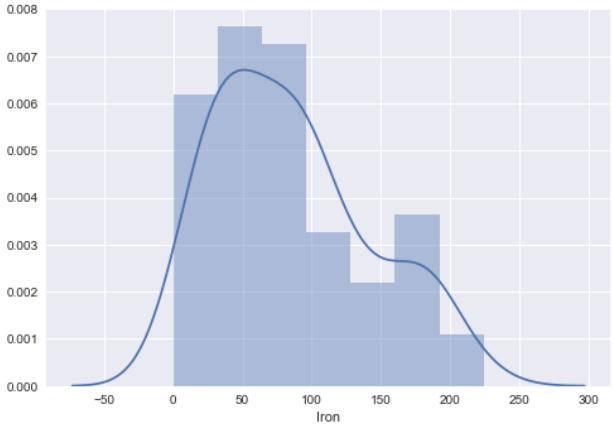


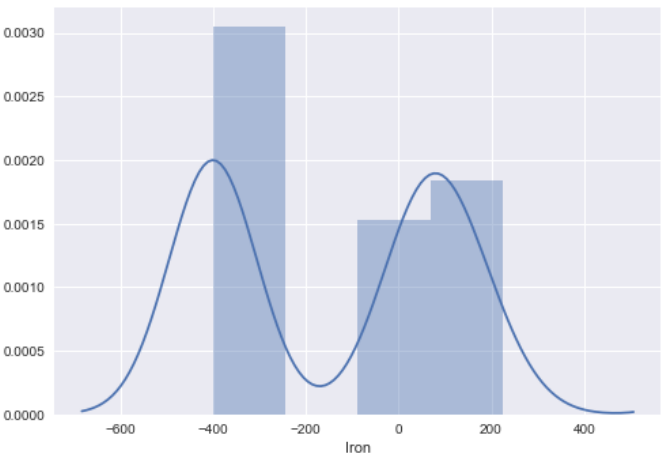
- η Άμεση χολερυθρίνη (mg/dL) (Dir. Bil),
  Οι τιμές που λείπουν αναπαριστώνται προσωρινά στο διάγραμμα με αρνητική τιμή. Για λόγους καλύτερης αυκρίνιας το αριστερό διάγραμμα δεν συμπεριλαμβάνει τις τιμές που λείπουν. *Ο βαθμός λευκωματίνης-χολερυθρίνης βρέθηκε να είναι ανώτερος για τη διάκριση ασθενών με καλύτερη ηπατική λειτουργία. μπορεί να είναι ένα καλύτερο σύστημα συνολικής προγνωστικής βαθμολόγησης για την πρόβλεψη της επιβίωσης των Ιαπώνων ασθενών με HCC*. (Hiraoka et al., 2016) Παρατηρούμε ότι κι εδώ υπάρχει ώμος για Άμεση χολερυθρίνη > 2.5 και υπάρχουν σποραδικά καταγραφές στο διάστημα (5,10). Οι τιμές που λείπουν είναι σχεδόν οι μισές όσων έχουν καταγραφεί.





- ο Σίδηρος (mcg/dL) (Iron),
  Οι τιμές που λείπουν αναπαριστώνται προσωρινά στο διάγραμμα με αρνητική τιμή. Στο αριστερό διάγραμμα -για λόγους καλύτερης ευκρίνειας- δεν συμπεριλαμβάνονται οι τιμές που λείπουν. Κι εδώ υπάρχει ξεκάθαρα ώμος για Σίδηρο > 150. Επίσης παρατηρούμε ότι οι τιμές που λείπουν είναι σχεδόν τόσες όσο κι αυτές που είναι καταγεγραμένες. Κάτι τέτοιο μειώνει την αξιοπιστία των συμπερασμάτων που μπορεί να εξαχθούν. Πρόκειται για γνώρισμα που σχετίζεται με την ομοιόσταση του σιδήρου, όπως αναλύεται και στην εισαγωγή η διαταρραχή του είναι ανάμεσα στους παράγοντες κινδύνου για ανάπτυξη ηπατοκυτταρικού καρκινώματος. Συγκεκριμένα, η ύπαρξη μεγάλης ποσότητας ελεύθερου σιδήρου είναι ενδεικτική μεγαλύτερου κινδύνου.

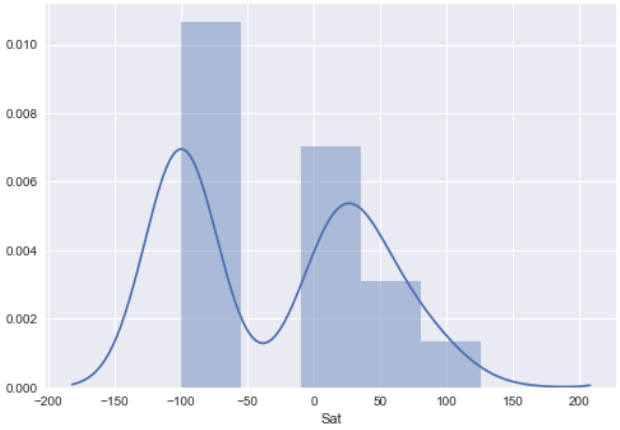




Σύμφωνα με τους (Li et al., 2020) *οι ασθενείς με χαμηλό προεγχειρητικό επίπεδο σιδήρου στον ορό είχαν χειρότερη μετεγχειρητική επιβίωση και υψηλότερο ποσοστό υποτροπής στο HCC. Ο προεγχειρητικός σίδηρος ορού είναι ένας ανεξάρτητος προγνωστικός παράγοντας των ασθενών με HCC*.

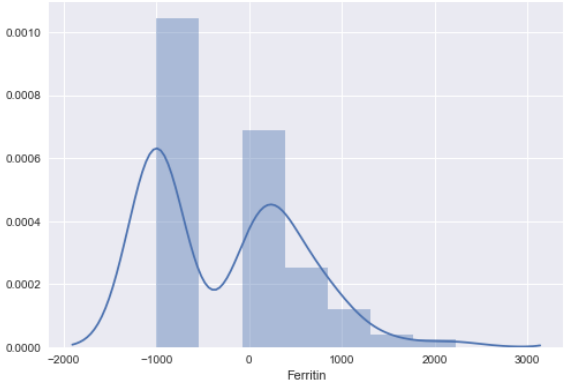
- ο Κορεσμός Οξυγόνου (%) (Sat),

Οι τιμές που λείπουν αναπαριστώνται προσωρινά στο διάγραμμα με αρνητική τιμή. Παρατηρούμε ότι λείπουν πολλές τιμές σε σχέση με όσες έχουν καταγραφεί, συνεπώς τα συμπεράσματα που μπορεί να βγουν δεν είναι εξίσου αξιόπιστα με περιπτώσεις όπου δεν λείπουν πολλά δεδομένα. Κι εδώ θα μποεούσαμε να πούμε ότι υπάρχει μία πολύ μικρή μείωση του ρυθμού μεταβολής κλίσης της καμπύλης της κατανομής για Κορερσμό Οξυγόνου >75.



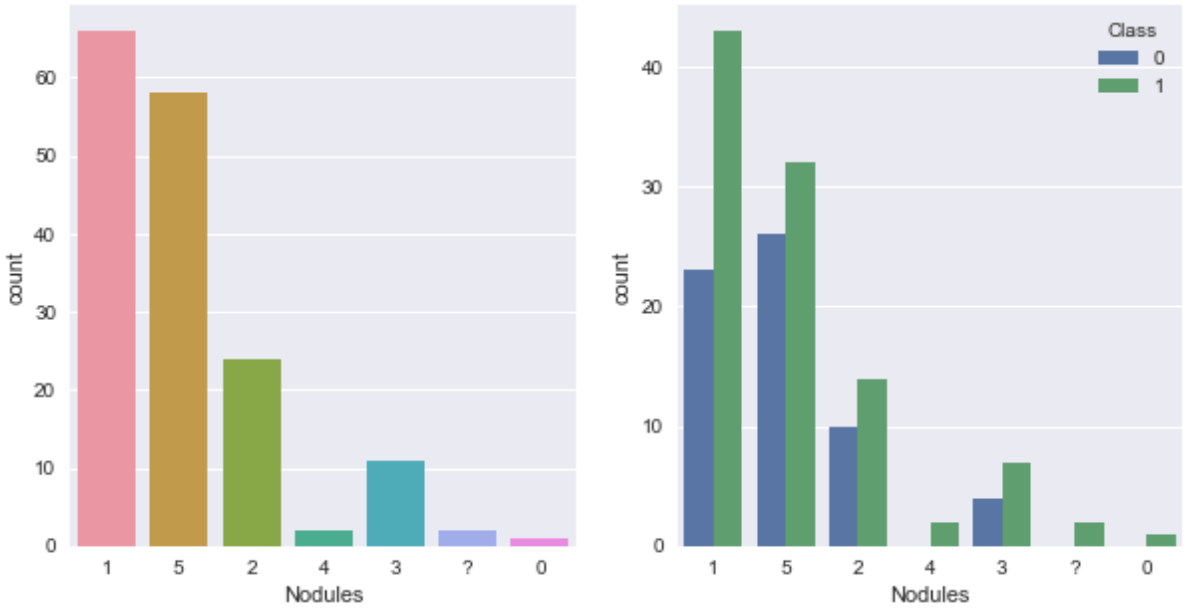
- η Φεριτίνη (ng/mL) (Ferritin)

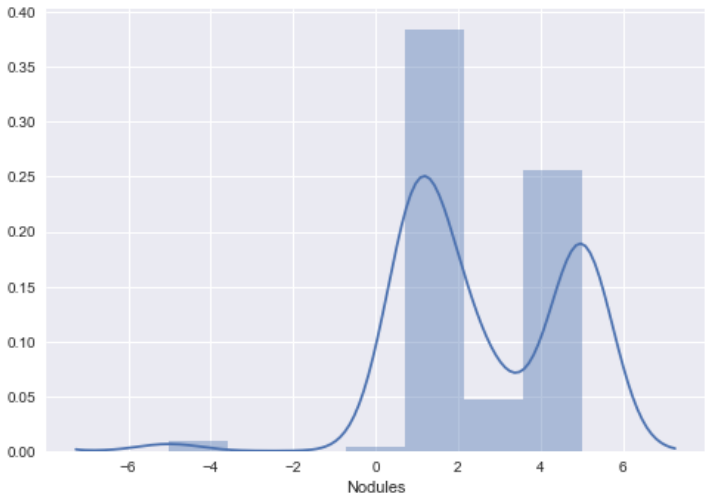
Οι τιμές που λείπουν αναπαριστώνται προσωρινά στο διάγραμμα με αρνητική τιμή. Παρατηρούμε ότι κι εδώ είναι πολλές. Κι εδώ θα μποεούσαμε να πούμε ότι υπάρχει μία πολύ μικρή μείωση του ρυθμού μεταβολής κλίσης της καμπύλης της κατανομής για Φεριτίνη λίγο πριν το 1000 και ξανά για Φεριτίνη ~2000. Πρόκειται για γνώρισμα που σχετίζεται με την ομοιόσταση του σιδήρου, όπως αναλύεται και στην εισαγωγή η διαταρραχή του είναι ανάμεσα στους παράγοντες κινδύνου για ανάπτυξη ηπατοκυτταρικού καρκινώματος. *Τα επίπεδα φερριτίνης ορού δεν επηρεάζουν την πρόγνωση ασθενών με HCC που υποβάλλονται σε ραδιοκαυτηρίαση (radiofrequency ablation – RFA)* σύμφωνα με τους (Uchino et al., 2018). *Η αναλογία φερριτίνης προς τρανσφερίνη ορού (FTR) >7,7 σχετίζεται με καλύτερη συνολική επιβίωση και ξεπέρασε τις επιδόσεις του συστήματος σταδιοποίησης της Βαρκελώνης*. (Vohra et al., 2020) *Ο λόγος φερριτίνης-τρανσφερρίνης (FTR), τη στιγμή της διάγνωσης με HCC πρόβλεψε τη θνησιμότητα στη κοόρτη των ασθενών των* (Vohra et al., 2021), *με βέλτιστο διαχωρισμό στην τιμή 7,7 σε ομάδες υψηλού και χαμηλού κινδύνου*.



- ο αριθμός των Οζιδίων (Nodules),

Οι τιμές που λείπουν αναπαριστώνται προσωρινά στο διάγραμμα με αρνητική τιμή. Παρατηρούμε ότι στο σύνθετο ραβδόγραμμα ως προς την έβαση των ασθενών μετά το ένα έτος (δεξί διάγραμμα), ο λόγος αποβιώσαντων / επιβιώσαντες είναι μεγαλύτερος για αριθμό Οζιδίων = 4, 0 (αν και έχουν πολύ λίγες καταγραφές) και ακολουθεί για 1 και 3, ενώ είναι μικρότερος για αριθμό Οζιδίων = 5 και 2. Λείπουν και λίγες καταγραφές που παρουσιάζονται με '?'. Θα μπορούσαμε να πούμε ότι κι εδώ η κατανομή είναι δικόρυφη (βλ. κατανομή συχνοτήτων). *Ο αριθμός των οζιδίων ήταν ένας από τους 5 ανεξάρτητους προγνωστικούς παράγοντες που συσχετίζονταν με την επιβίωση βάσει μοντέλου παλινδρόμησης* των (Adhoute et al., 2016). *Η μελέτη αφορούσε κοόρτες ασθενών με Καρκίνο του Ήπατος, σταδίου C, από Κλινική Βαρκελώνης (BCLC, C) με HCC, που έλαβαν θεραπεία με sorafenib*. (Adhoute et al., 2016)





6.2 Προγνωστική Ανάλυση (Predictive Analysis)

Τα γραφήματα με τις τιμές των μετρικών που χρησιμοποιήθηκαν για αξιολόγηση της ταξινόμησης παρουσιάζονται στο *Παράρτημα IV*. Λόγω πλήθους των διαγραμμάτων έχουν τοποθετηθεί σε σμίκρυνση, αλλά με καλή ανάλυση, ώστε μόλις γίνει μεγέθυνση της σελίδας να φαίνεται ξεκάθαρα ο τίτλος κάθε γραφήματος, καθώς και οι ετικέτες που αφορούν τον κάθε άξονά του.

Στο ***Παράρτημα IVa***, παρουσιάζονται ραβδογράμματα (bar charts) για τις τιμές μίας μετρικής και μίας πορείας προεπεξεργασίας (preprocessing) ανά γράφημα, όπου συγκρίνονται ως προς την απόδωσή τους οι διάφορες ταξινομικές μέθοδοι. Ως άνω και κάτω σφάλματα (error bars) έχουν τοποθετηθεί αντίστοιχες τυπικές αποκλίσεις (standard deviation).

Στο ***Παράρτημα IVb***, παρουσιάζονται ραβδογράμματα (bar charts) για τις τιμές μίας μετρικής και μίας μεθόδου ταξινόμησης ανά γράφημα, όπου συγκρίνονται οι διαφορετικές πορείες προεπεξεργασίας που δοκιμάστηκαν. Σαν άνω και κάτω σφάλματα (error bars) έχουν τοποθετηθεί με όμοιο τρόπο οι αντίστοιχες τυπικές αποκλίσεις (standard deviation).

Έχει συνειδητά επιλεγεί η κλίμακα του άξονα y στα ραβδογράμματα να έχει τα όρια (-1,1.1) (ακόμα και σε μετρικές που δεν υπήρχαν αρνητικές τιμές) προκειμένου να έχουν ίδια κλίμακα στον άξονα των y οι στήλες μεταξύ διαφορετικών γραφημάτων και σαν συνέπεια αυτού να είναι συγκρίσημα τα γραφήματα των ίδιων μετρικών μεταξύ τους. Θεωρήθηκε σημαντικό να μπορεί και οπτικά να παρατηρηθεί η διαφορά μεταξύ των στηλών που οφείλεται στην διαφορά της τιμής των μετρικών. Η έλλειψη κάποιων από τις στήλες σε κάποια από τα ραβδογράμματα οφείλεται στο ότι κατά τους υπολογισμούς της μετρικής προέκυψε απροσδιοριστία (συνήθως επρόκειτο για διαίρεση με 0) και ως εκ τούτου η τιμή που ανακτήθηκε ήταν not a number (nan). Το γεγονός ότι ο άξονας των x τέμνει τον άξονα των y σε σημείο διαφορετικό του 0 είναι επίσης συνειδητό, προκειμένου να είναι εύκολα παρατηρήσιμο το πού προέκυψε απροσδιοριστία και να είναι διακριτό από τις περιπτώσεις εκείνες όπου οι στήλες διαθέτουν πραγματική τιμή, η οποία είναι όμως ίση με το 0 και θα ταυτίζοταν με τον άξονα των x αν αυτός έτεμνε τον άξονα των y στο σημείο (y=0).

Ο αριθμός των γραφημάτων είναι τέτοιος που καθιστά απαγορευτική την ανάλυση κάθε ενός από αυτά ξεχωριστά. Στο κείμενο παρουσιάζεται μία σύνοψη των γραφημάτων και αναλύονται τα γραφήματα με το περισσότερο ενδιαφέρον.

Σημειώνεται ότι υπήρχε πρόθεση να γίνει και υπολογισμός της μετρικής ποσοστού ψευδούς ανακάλυψης (false discovery rate), αλλά δεν συμπεριλαμβάνεται στα αποτελέσματα λόγω του ότι είχε γραφεί λάθος ο τύπος υπολογισμού της στον κώδικα, με συνέπεια αντί για αυτήν να υπολογίζεται ξανά η μετρική αρνητική προγνωστική αξία (negative predictive value). Αυτό έγινε αντιληπτό αφού είχαν ολοκληρωθεί οι αναλύσεις κατά τον σχετικό έλεγχο, με συνέπεια να μην υπάρχει χρόνος για εκ νέου υπολογισμό της.

Στην λεκτική περιγραφή των αποτελεσμάτων έχει δωθεί μικρότερο μέγεθος γραμματοσειράς προκειμένου να ξεχωρίζει από την σύνοψη και τον σχολιασμό. Για κάθε περιγραφή έχουν ληφθεί υπ'όψη μεγάλος αριθμός ραβδογραμμάτων, το οποία βρίσκονται στο Παράρτημα και είναι αδύνατο να τοποθετηθούν στο κείμενο λόγω έλειψης χώρου. Θα γίνεται όμως αναφορά στο ποιά είναι αυτά τα ραβδογράμματα ανά παράγραφο, προκειμένου να είναι δυνατή η εξέτασή τους.

Στην εργασία των δωρητών του συνόλου δεδομένων που χρησιμοποιήσαμε (Santos et al., 2015) οι βέλτιστες τιμές μετρικών ήταν, για νευρωνικό δίκτυο (NN):
Ορθότητα-ακρίβεια (Accuracy): 0.730, τυπική απόκλιση (STD): 0.014,
Επιφάνεια κάτω από την καμπύλη ROC (AUC): 0.673, τυπική απόκλιση (STD): 0.012,
f-σκορ (f-measure): 0.652, τυπική απόκλιση (STD): 0.016
Για λογιστική παλλινδρόμηση (LR):
Ορθότητα-ακρίβεια (Accuracy): 0.752, τυπική απόκλιση (STD): 0.011,
Επιφάνεια κάτω από την καμπύλη ROC(AUC):0.700, τυπική απόκλιση(STD):0.015,
f-σκορ (f-measure): 0.665, τυπική απόκλιση (STD): 0.018
Δεδομένου του ότι ενδιαφέρον έχει κάποια απόδοση εξίσου καλή ή και καλύτερη από την ανωτέρω ενδεικτική απόδοση, μας ενδιαφέρουν ακρίβειες μεγαλύτερες ή περί του 0.752 για την Λογιστική Παλλινδρόμηση και του 0.730 για τα Νευρωνικά Δίκτυα. Αντίστοιχα για την AUC οι τιμές είναι 0.700 και 0.673. Για το 0.018 (f-score) οι τιμές είναι 0.665 και 0.652 αντίστοιχα, αλλά δεν είναι δόκιμη η σύγκριση για λόγους που αναλύονται ακολούθως.

Η μετρική ακρίβεια ή ορθότητα (accuracy)

Στο σύνολο όπου έγινε Iterative imputation και Υποδειγματοληψία (undersampling), οι LR5, LR6 έχουν ακρίβεια κοντά στην τιμή στόχο, με τυπική τους απόκλιση αρκετά μεγαλύτερη.
Κάτι αντίστοιχο συμβαίνει και για Iterative imputation και εξισορρόπηση με την παράμετρο class_weights, στον LR5. Ομοίως και για το σύνολο όπου έγινε Iterative imputation και τυχαία υπερδειγματοληψία (random-oversampling), ωστόσο κάτι τέτοιο αναμένεται.
Όσον αφορά τα Νευρωνικά Δίκτυα, η μέση ακρίβεια πλησιάζει την τιμή στόχο για τα σετ όπου έγινε εμπειρική συμπλήρωση κενών και τυχαία υπερδειγματοληψία (random-oversampling) για τους ταξινομητές NN2-9, με μερικούς να την ξεπερνούν λίγο (αλλά έχουν μεγαλύτερη τυπική απόκλιση). Αντίστοιχα και για το σύνολο όπου χρησιμοποιήθηκε η συνάρτηση που δοκιμάσαμε και τυχαία υπερδειγματοληψία (random-oversampling) (ο NN6 πλησιάζει την τιμή στόχο, αλλά έχει μεγαλύτερη τυπική απόκλιση). Ομοίως και για iterative imputation με τυχαία υπερδειγματοληψία (random-oversampling) (ο NN9 πλησιάζει, αλλά έχει μεγαλύτερη τυπική απόκλιση).

Η μετρική AUC

Για εμπειρική συμπλήρωση κενών, χωρίς εξισορρόπηση ξεχωρίζουν οι Παραλλαγές Λογιστικής Παλλινδρόμησης νο5 (στο εξής LR5), και νο 6 (στο εξής LR6) (κοντά στην τιμή στόχο), αλλά η τυπική απόκλιση είναι και εδώ μεγαλύτερη.
Ομοίως για iterative imputation, χωρίς εξισορρόπηση ή με Υποδειγματοληψία (undersampling) ή με την παράμετρο class_weights. Ομοίως και για εμπειρική συμπλήρωση κενών και Υποδειγματοληψία (undersampling) ή class_weights (ξεπερνούν οι LR5, LR6 τις αντίστοιχη τιμή στόχο). Το ίδιο συμβαίνει και για εμπειρική συμπλήρωση κενών ή iterative imputation και υπερδειγματοληψία με την μέθοδο SMOTE (SMOTE-oversampling) και για εμπειρική συμπλήρωση κενών ή iterative imputation και τυχαία υπερδειγματοληψία (random oversampling).
Στο σύνολο όπου έγινε iterative imputation και δεν εξισορροπήθηκαν οι κλάσεις ο ταξινομητής NN10 πλησιάζει την τιμή στόχο (αλλά δυστυχώς κι εδώ έχει μεγαλύτερη τυπική απόκλιση). Ομοίως για την περίπτωση εμπειρικής συμπλήρωσης κενών και υποδειγματοληψίας (undersampling) (για τον ταξινομητή NN9). Το ίδιο συμβαίνει και για iterative imputation και υπερδειγματοληψία μέσω SMOTE (SMOTE-oversampling) για τον NN9 (και ο NN10 ακολουθεί κοντά).
Για εμπειρική συμπλήρωση κενών και τυχαία υπερδειγματοληψία (random-oversampling) όλοι οι ταξιομητές NN ξεπερνούν την τιμή στόχο, (με μεγαλύτερη τυπική απόκλιση), αλλά αναμένεται κάτι τέτοιο και αξιολογείται ως πλασματικό -δεδομένης της πιθανής παρουσίας ίδιων καταγραφών που ανήκουν στην μειοψηφούσα κλάση τόσο στο εκπαιδευτικό σετ, όσο και στο σετ ελέγχου. Για την δοκιμαζόμενη συνάρτηση και τυχαία υπερδειγματοληψία (random-oversampling) κάποιοι ταξινομητές NN ξεπερνούν σε μικρότερο βαθμό την τιμή στόχο και κάποιοι πλησιάζουν (με μεγαλύτερη τυπική απόκλιση), αλλά αναμένεται. Η συνάρτηση τα πάει χειρότερα από τις άλλες μεθόδους συμπλήρωσης κενών.
Σε σύνολο με συμπλήρωση κενών με iterative imputation και εξισορρόπηση με τυχαία υπερδειγματοληψία (random-oversampling) κάποιοι ταξινομητές ξεπερνούν κι αυτοί, αλλά σε μικρότερο βαθμό (NN2-10) (με μεγαλύτερη τυπική απόκλιση), την τιμή στόχο, αλλά κι αυτό αναμένεται και δεν αξιολογείται σαν ιδιαίτερα σημαντικό εύρημα. Θα μπορούσαμε να έχουμε μία αξιόπιστη αξιολόγηση των συγκεκριμένων μοντέλων αν εξασφαλιζόταν ότι το σετ ελέγχου θα ήταν διαφορετικό από το σετ εκπαίδευσης.
Ενδιαφέρον έχει η περίπτωση (Iterative imputation, SMOTE-υπερδειγματοληψία), (Iterative imputation, υποδειγματοληψία), (εμπειρική συμπλήρωση, υποδειγματοληψία), όπου απέδωσαν καλά τόσο οι LR5,LR6 όσο και ο NN9. Ενδιαφέρον επίσης έχει η περίπτωση (Iterative imputation, χωρίς

εξισορρόπηση) όπου απέδωσαν καλά οι LR5,LR6 και ο NN10. Σε γενικές γραμμές οι NN απέδωσαν λιγότερο καλά από τους καλύτερους LR ταξινομητές, κάτι που όμως αναμενόταν. Γενικά ξεχωρίζουν οι LR5, LR6 από τους υπόλοιπους ταξινομητές Λογιστικής παλλινδρόμησης, ως προς τις συγκεκριμένες μετρικές αξιολόγησης.

Συμπεράσματα σύγκρισης

Παρατηρείται ότι οι ταξινομητές NN απέδωσαν καλά ως προς την ακρίβεια ή ορθότητα (accuracy) μόνο όπου είχε γίνει τυχαία υπερδειγματοληψία (random-oversampling), συνεπώς δεν θεωρείται αυτή η απόδωσή τους αξιόλογο εύρημα. Οι ταξινομητές της λογιστικής παλλινδρόμησης LR5, LR6 απέδωσαν καλύτερα από τους υπόλοιπους συνολικά, ωστόσο κι αυτοί είχαν ενδιαφέρουσα ακρίβεια μόνο όταν τα κενά συμπληρώθηκαν μέσω Iterative imputation. Αυτό θεωρείται πιο ενδιαφέρον εύρημα.

Όσον αφορά την μετρική AUC ξεχωρίζουν πάλι οι ταξινομητές LR5, LR6 και NN9, NN10, με τους δεύτερους να υστερούν σε γενικές γραμμές έναντι των πρώτων. Στις περισσότερες περιπτώσεις τυχαίας υπερδειγματοληψίας αρκετοί ταξινομητές NN ξεπερνούν την τιμή στόχο (αν και με μεγαλύτερη τυπική απόκλιση), αλλά και πάλι αναμένεται κάτι τέτοιο και αξιολογείται ως μη σημαντικό -δεδομένης της πιθανής παρουσίας ίδιων καταγραφών που ανήκουν στην μειοψηφούσα κλαση τόσο στο εκπαιδευτικό σετ, όσο και στο σετ ελέγχου που οφείλεται στην τυχαία υπερδειγματοληψία.

Ως προς το Σκορ-f1 (f1-score) το αποτέλεσμα είναι αρκετά καλύτερο συνολικά. Ακόμη και για περιπτώσεις εμπειρικής συμπλήρωσης κενών με έλλειψη εξισορρόπησης του σετ οι ταξινομητές αποδίδουν αρκετά καλά, ξεπερνώντας τις αντίστοιχες τιμές στόχους ή βρισκόμενοι περί αυτών. Η μετρική του Σκορ-f1 (f1-score) αποτελεί τον αρμονικό μέσο ευαισθησίας και θετική προγνωστική αξία (precision). Δεν πρέπει να παραπλανηθούμε από την υψηλή ευαισθησία, η οποία τείνει να είναι υψηλή στις περισσότερες ταξινομήσεις όταν δεν υπάρχει εξισορρόπηση. Σε αυτές τις περιπτώσεις το μοντέλο ταξινόμησης εκπαιδεύεται σε μη αντιπροσπευτικό και για τις δύο κλάσεις δείγμα, με συνέπεια να μεροληπτεί υπέρ της πλειοψηφούσας. Η υψηλή ευαισθησία όπου υπάρχει αυξάνει και το Σκορ-f1 (f1-score).

Υπάρχει ακόμη ένα ***πρόβλημα με την σύγκριση διαφορετικών σετ δεδομένων στα οποία υπάρχουν διαφορετικές αναλογίες κλάσεων, μιας και το Σκορ-f1 (f1-score) επιρρεάζεται από την ανισορροπία των κλάσεων.*** Συνεπώς τα ανωτέρω ευρήματα αξιολογούνται ως μη σημαντικά. ***Δόκιμη σύγκριση θα μπορούσε να γίνει μόνο μεταξύ σετ στα οποία ο λόγος μεταξύ των δύο κλάσεων είναι ίδιος*** (λχ. σε αυτά όπου έχει γίνει εξισορρόπηση προς τον ίδιο τελικό λόγο κλάσεων). (“F-score,” n.d.) Σε αυτό το πνεύμα αξιοσημείωτο είναι ότι ως προς το Σκορ-f1 (f1-score) εμφανίζεται να αποδίδει καλά και ο LR4 (μαζί με τους LR5,LR6) για τις περιπτώσεις εμπειρικής συμπλήρωσης κενών συνδυασμένες με υπερδειγματοληψία (μέσω SMOTE ή τυχαία). Για Iterative imputation συνδυασμένο με υπερδειγματοληψία μέσω SMOTE μαζί με τους υπόλοιπους ταξινομητές αποδίδουν καλά και οι NN3, NN4, NN9.

Η παρούσα εργασία συμπεριλαμβάνει περισσότερες ταξινομικές μεθόδους και μετρικές αξιολόγησης της ποιότητας ταξινόμησης, οι οποίες θα συζητηθούν στην συνέχεια. Η απόδωση ορισμένων ξεπέρασε κατά πολύ τις ανωτέρω προσδοκίες. Ακολουθεί συνοπτική περιγραφή των γραφημάτων. Η πληροφορία που αποκομιζόταν από το σύνολο των διαγραμμάτων φαινόταν και στο γράφημα του μέτρου του αξιοσημείωτου (markedness), συνεπώς για περιπτώσεις που οι διαφορές ήταν δύσκολο να εκτιμηθούν εμπιστευθήκαμε την μετρική αυτή, καθώς και τις μετρικές Πληροφοριακότητα (Informedness), Συντελεστής συσχέτισης Matthews (MCC). Επίσης, σημαντικό για να κριθεί μία απόδοση ως καλή ήταν να μην

παρουσιάζει τιμές άνω του μετρίου για τις μετρικές ψευδώς θετικό ποσοστό (fall-out), ψευδώς αρνητικό ποσοστό (Miss rate) και false omision rate.

SVM

Φτωχή απόδωση δείχνει να έχει γενικά η μέθοδος αυτή, με εξαίρεση το Σκορ-f1 (f1-score), το ψευδώς αρνητικό ποσοστό (Miss rate) και την ευαισθησία της. Η ευαισθησία είναι ικανοποιητική για εξισορρόπηση με class_weights. Επίσης η ευαισθησία είναι καλή για Iterative imputation και υπερδειγματοληψία ή υποδειγματοληψία. Το πολύ υψηλό ψευδώς θετικό ποσοστό (fall-out) σχεδόν σε κάθε κατηγορία προεπεξεργασίας είναι ενδεικτικό κακής ποιότητας ταξινόμησης. Εξαίρεση, με μικρό δείκτη ψευδώς θετικού ποσοστού (fall-out) έχουμε για τη δοκιμαζόμενη συνάρτηση και υποδειγματοληψία ή SMOTE-υπερδειγματοληψία. Στα ανωτέρω έχουμε υψηλό ποσοστό ψευδώς αρνητικών (false-negative rate) (και με αρκετά μεγάλη τυπική απόκλιση). Περίεργο είναι το γεγονός ότι η τυχαία υπερδειγματοληψία έχει επίσης υψηλό δείκτη ψευδώς θετικού ποσοστού (fall-out) (κάτι που δεν θα έπρεπε να συμβαίνει αν υπήρχαν συγχρόνως στο εκπαιδευτικό και στο σετ ελέγχου ίδιες καταγραφές). Όσον αφορά το ποσοστό ψευδούς παράλειψης (false omission rate) (FN/N) παρατηρούμε ότι η συνάρτηση που δοκιμάστηκε ή το iterative imputation συνδυαστικά με υποδειγματοληψία φτάνει αρκετά ψηλά, κάτι που δεν είναι καλό. Το ψευδώς αρνητικό ποσοστό (Miss rate) είναι μικρό για τα μη ισορροπημένα σετ, κάτι το αναμενόμενο. Εντούτοις είναι χαμηλό και για εξισορροπημένα σετ με class_weights. Για iterative imputation, χαμηλό ψευδώς αρνητικό ποσοστό (Miss rate) έχουν και τα εξισορροπημένα σετ με υποδειγματοληψία ή SMOTE-υπερδειγματοληψία, κάτι που αξιολογείται ως χρήσιμο. Χαμηλό ψευδώς αρνητικό ποσοστό (Miss rate) έχει και η τυχαία υπερδειγματοληψία. Η αρνητική Προγνωστική Αξία (Negative Predictive Value) είναι μέτρια προς καλή για iterative imputation και υποδειγματοληψία ή τυχαία υπερδειγματοληψία (αν και το τελευταίο δεν αξιολογείται ως σημαντικό). Η θετική προγνωστική αξία (precision) δεν είναι αρκετά καλή για καμία περίπτωση προεπεξεργασίας (<0.7). Το κατώφλι επικράτησης (prevalence threshold) δεν είναι αρκετά χαμηλό (50%), αφού ούτε και οι AUC είναι πολύ ψηλές. Η ειδικότητα (specificity) είναι μέτρια προς κακή για τις περισσότερες περιπτώσεις, με την συνάρτηση μαζί με υποδειγματοληψία να έχει λίγο καλύτερη μέση ειδικότητα, και τυπική απόκλιση πολύ μεγάλη, ομοίως και για (συνάρτηση, SMOTE-υπερδειγματοληψία). Λίγο μικρή τυπική απόκλιση υπαρχει για (Iterative imputation, τυχαία υπερδειγματοληψία), (Iterative imputation, SMOTE-υπερδειγματοληψία), (Iterative imputation, υποδειγματοληψία), αν και η μέση ειδικότητα είναι μέτρια προς κακή. Αξιόλογο είναι το γεγονός ότι η βαθμολογία απειλής (threat score) για class_weights ίδια με αυτή των μη εξισορροπημένων σετ (που αναμένεται να έχουν το μέγιστο λόγω μεροληψίας υπέρ της πλειοψηφούσας κλάσης). Λίγο λιγότερο αποδίδει και η συμπλήρωση κενών με iterative imputation ανεξαρτήτως τρόπου εξισορρόπησης.

Καλύτερα αποτελέσματα έχουμε για Iterative imputation,τυχαία υπερδειγματοληψία.

LR1, LR2, LR3, LR4

Η ακρίβεια ή ορθότητα (accuracy) δεν ξεπερνά το 0,75 για καμία προεπεξεργασία, είναι μέτρια. Καλή ευαισθησία έχουν οι εξισορροπήσεις με class_weights. Η καλή ευαισθησία για (iterative imputation, χωρίς εξισορρόπηση) είναι αναμενόμενη λόγω μεροληψίας που οφείλεται στην ανισορροπία κλάσεων. Αξιοσημείωτη είναι η κακή ευαισθησία για εμπειρική συμπλήρωση ή για την δοκιμαζόμενη συνάρτηση και έλλειψη εξισορρόπησης, γεγονός αρνητικό. Το αντίστροφο συμβαίνει για την ειδικότητα (specificity). Η τιμή κ του Cohen (Cohen's kappa), η πληροφοριακότητα (Informedness), το μέτρο του αξιοσημείωτου (markedness), ο Συντελεστής συσχέτισης Matthews (MCC), είναι χαμηλά γενικά, κάτι που αξιολογείται αρνητικά για τους συγκεκριμένους ταξινομητές. Το Σκορ-f1 (f1-score) για την εξισορρόπηση με class_weights είναι καλύτερο από ό,τι για άλλες εξισορροπήσεις. Η εξισορροπημένη ακρίβεια (balanced accuracy) είναι μέτρια (50%). Η AUC είναι μέτρια προς καλή γενικά, με καλύτερες τις περιπτώσεις όπου έγινε υποδειγματοληψία, τυχαία υπερδειγματοληψία, ή χρησιμοποιήθηκε η συνάρτηση που δοκιμάζεται μαζί με SMOTE-υπερδειγματοληψία ή class_weights. Υψηλό ψευδώς θετικό ποσοστό (fall-out) έχουμε για class_weights (κάτι που δεν είναι καλό), επίσης μέτριο ψευδώς θετικό ποσοστό (fall-out) παρουσιαζεται για την συνάρτηση που δοκιμάστηκε. Χαμηλό ψευδώς θετικό ποσοστό (fall-out) έχει η υποδειγματοληψία και η υπερδειγματοληψία, κάτι το καλό. Το ποσοστό ψευδούς παράλειψης (false omission rate) είναι μέτριο στις περισσότερες προεπεξεργασίες. Ο Δείκτης Fowlkes-Mallows (FM) είναι μέτριος προς καλός για class_weights, ενώ μέτρια προς καλά τα πάει και η συνάρτηση που δοκιμάστηκε (ομοίως και για περιπτώσεις μη εξισορρόπησης, αλλά αυτό αναμένεται λόγω της ανισορροπίας κλάσεων). Το ψευδώς αρνητικό ποσοστό (Miss rate) είναι χαμηλό (όπως και χρειάζεται) για Iterative imputation χωρίς εξισορρόπηση και για class_weights. Η αρνητική προγνωστική αξία (negative predictive value) είναι μέτρια γενικά, δεν μπορεί να θεωρηθεί καλή. Η θετική προγνωστική αξία (precision), είναι μέτρια προς καλή γενικά, με σταθερή απόδοση για εξισορρόπηση με class_weights, ή για Iterative imputation συνδυασμένο με έλλειψη εξισορρόπησης. Καλύτερα τα πήγε και η δοκιμαζόμενη συνάρτηση, με την τυπική απόκλιση όμως να είναι μεγαλύτερη. Το κατώφλι επικράτησης (prevalence threshold) είναι έως μέτριο προς χαμηλό σε ορισμένες περιπτώσεις (γεγονός που φαίνεται και από τις αντίστοιχες AUC). Για εξισορρόπηση με

class_weights, ή όπου έγινε Iterative imputation συνδυασμένο με έλλειψη εξισορρόπησης, το κατώφλι επικράτησης είναι σταθερά μέτριο. Λίγο μικρότερο, αλλά με μεγαλύτερη τυπική απόκλιση είναι για την συνάρτηση που δοκιμάζεται. Αυτό δεν είναι κακό, ούτε όμως και πολύ αξιόλογο εύρημα. Η βαθμολογία απειλής (threat score) είναι χαμηλή έως μέτρια (για class_weights). Σε γενικές γραμμές δεν πρόκειται για ταξινομήτριες που θα ήταν χρήσιμο να επιλεγούν σε πραγματικές συνθήκες για τις κατηγορίες προεπεξεργασίας που δοκιμάστηκαν.

Καλύτερα αποτελέσματα έχουμε για (δοκιμαζόμενη συνάρτηση, χωρίς εξισορρόπηση) και ακολουθούν (δοκιμαζόμενη συνάρτηση, SMOTE-υπερδειγματοληψία), (δοκιμαζόμενη συνάρτηση, τυχαία υπερδειγματοληψία).

LR5

Η ακρίβεια ή ορθότητα (accuracy) παρουσιάζει το εξής μοτίβο: για την συνάρτηση που δοκιμάζεται είναι μικρότερη από ό,τι για εμπειρική συμπλήρωση, η οποία με την σειρά της είναι μικρότερη από την συμπλήρωση με Iterative imputation (η οποία και πετυχαίνει την ψηλότερη ακρίβεια ή ορθότητα (accuracy), πλησιάζοντας το 0,75). Η AUC και η εξισορροπημένη ακρίβεια (balanced accuracy) και η τιμή κ του Cohen (Cohen's kappa) παρουσιάζουν το ίδιο μοτίβο επίσης. Στην τιμή κ του Cohen (Cohen's kappa) οι συνδυασμοί (Iterative imputation, υποδειγματοληψία), και όπου έχει γίνει τυχαία υπερδειγματοληψία (με εξαίρεση την δοκιμαζόμενη συνάρτηση που τα έχει πάει χαρακτηριστικά άσχημα) πλησιάζουν μέτριες τιμές. Στο Σκορ-f1 (f1-score) το μοτίβο διατηρείται, αλλά για class_weights, είναι πολύ μικρότερες οι διαφορές, καθώς πλησιάζουν μεταξύ τους προς πάνω. Ομοίως και για το Δείκτη Fowlkes-Mallows (FM). Στο ψευδώς θετικό ποσοστό (fall-out), το προηγούμενο μοτίβο αντιστρέφεται όπως αναμενόταν. Χαμηλότερο ψευδώς θετικό ποσοστό (fall-out) υπάρχει για υποδειγματοληψία (για Iterative imputation έχει το χαμηλότερο) και για (εμπειρική συμπλήρωση, τυχαία υπερδειγματοληψία). Σχετικά μέτριο προς χαμηλό ψευδώς θετικό ποσοστό (fall-out) έχουν οι συνδυασμοί (συνάρτηση, υποδειγματοληψία), καθώς και όπου έγινε SMOTE-υπερδειγματοληψία ή τυχαία υπερδειγματοληψία. Χαρακτηριστικά υψηλό (άρα και κακό εύρημα) είναι το ψευδώς θετικό ποσοστό (fall-out) για τον συνδυασμό (συνάρτηση, class_weights). Αντίστοιχα ευρήματα έχουμε γενικά και για το ποσοστό ψευδούς παράλειψης (false omission rate), με τους συνδυασμούς (iterative imputation, υποδειγματοληψία), (iterative imputation, χωρίς εξισορρόπηση), (iterative imputation, τυχαία υπερδειγματοληψία) -όπου ίσως να έχει γίνει υπερπροσαρμογή στα δεδομένα, (εμπειρική συμπλήρωση, τυχαία υπερδειγματοληψία) -όπου ίσως να έχει γίνει υπερπροσαρμογή, (iterative imputation, SMOTE-υπερδειγματοληψία), (iterative imputation, class_weights) -όπου όμως έχουμε μεγάλη τυπική απόκλιση, να αξιολογούνται ως πιο ενδιαφέροντες. Κακό (ψηλό) ποσοστό ψευδούς παράλειψης (false omission rate) έχουμε για τον συνδυασμό (εμπειρική συμπλήρωση, class_weights). Η Πληροφοριακότητα (Informedness) είναι καλύτερη –μέτρια- για (iterative imputation, υποδειγματοληψία), (iterative imputation, τυχαία υπερδειγματοληψία)- όπου όμως ίσως έχει γίνει υπερπροσαρμογή, (εμπειρική συμπλήρωση, τυχαία υπερδειγματοληψία) - όπου όμως ίσως έχει γίνει υπερπροσαρμογή, (iterative imputation, SMOTE-υπερδειγματοληψία). Χαρακτηριστικά κακή είναι για την την δοκιμαζόμενη συνάρτηση, ενώ συγκριτικά η εμπειρική συμπλήρωση τα έχει πάει αρκετά καλά, υστερόντας λίγο σε σχέση με το iterative imputation. Το ίδιο παρατηρείται και για το Μέτρο του αξιοσημείωτου (markedness), ενώ για (iterative imputation, class_weights) έχει αντίστοιχη τιμή, αλλά μεγάλη τυπική απόκλιση. Αντίστοιχα συμβαίνει και για το Συντελεστή συσχέτισης Matthews (MCC). Το ανεστραμένο μοτίβο της ακρίβειας ή ορθότητας (accuracy) παρατηρείται και για το ψευδώς αρνητικό ποσοστό (Miss rate). Χαμηλότερο ψευδώς αρνητικό ποσοστό (Miss rate) έχουμε για (iterative imputation, class_weights) και γενικά για class_weights, καθώς και για την χρήση του iterative imputation για συμπλήρωση κενών χωρίς εξισορρόπηση (με τις διάφορες τεχνικές εξισορρόπησης να ακολουθούν). Η αρνητική προγνωστική αξία (negative predictive value) είναι μέτρια προς γενικά. Ενδιαφέρον έχει η περίπτωση (Iterative imputation, υποδειγματοληψία). Καλή θετική προγνωστική αξία (precision) έχουν σχεδόν όλες οι προεπεξεργασίες, με εξαίρεση αυτές όπου χρησιμοποιείται η συνάρτηση που δοκιμάζεται. Το κατώφλι επικράτησης (prevalence threshold) είναι μέτριο προς χαμηλό (χαμηλότερες τιμές είναι επιθυμητές). Για (iterative imputation, υποδειγματοληψία), (εμπειρική συμπλήρωση, υποδειγματοληψία), (εμπειρική συμπλήρωση, τυχαία υπερδειγματοληψία), (iterative imputation, τυχαία υπερδειγματοληψία), (iterative imputation, SMOTE-υπερδειγματοληψία) έχουμε τις καλύτερες τιμές, όμως δεν πεφτουν κάτω από 0.25. Η ευαισθησία είναι μέτρια προς καλή πλειοψηφικά, ξεχωρίζουν οι (iterative imputation, class_weights), (δοκιμαζόμενη συνάρτηση, class_weights), (iterative imputation, χωρίς εξισορρόπηση), (εμπειρική συμπλήρωση, class_weights). Η ειδικότητα (specificity) είναι μέτρια προς καλή, με καλύτερες τιμές για (iterative imputation, υποδειγματοληψία), (εμπειρική συμπλήρωση, τυχαία υπερδειγματοληψία), (εμπειρική συμπλήρωση, υποδειγματοληψία), (iterative imputation, τυχαία υπερδειγματοληψία). Η βαθμολογία απειλής (threat score) είναι λίγο άνω του μετρίου. Ξεχωρίζουν οι περιπτώσεις (iterative imputation, class_weights), (iterative imputation, χωρίς εξισορρόπηση), (iterative imputation, SMOTE-υπερδειγματοληψία) και ακολουθούν οι υπόλοιπες για class_weights, οι υπόλοιπες συμπληρώσεις με iterative imputation και οι υπόλοιπες εμπειρικές συμπληρώσεις.

Καλύτερα αποτελέσματα έχουμε για (Iterative imputation, υποδειγματοληψία), (εμπειρική συμπλήρωση, τυχαία υπερδειγματοληψία), (iterative imputation, τυχαία

υπερδειγματοληψία), ενώ ακολουθούν κοντά και άλλα (βλ. Πληροφοριακότητα - informedness).

LR6

Για την ακρίβεια ή ορθότητα (accuracy) τα ενδιεφέροντα αποτελέσματα είναι αντίστοιχα με την LR5 γενικά. Το ίδιο ισχύει και για την AUC, μόνο που για Iterative imputation, υποδειγματοληψία η AUC είναι λίγο μικρότερη από της LR5. Η εξισορροπημένη ακρίβεια (balanced accuracy) είναι επίσης παρόμοια με της LR5, το ίδιο και η τιμή κ του Cohen (Cohen's kappa), με την διαφορά ότι υπάρχει μικρή βελτίωση για τις περιπτώσεις που πλησίαζε τις μέτριες τιμές. Το Σκορ-f1 (f1-score) κι εδώ έχει μικρές διαφορές από την LR5. Ομοίως και για το ψευδώς θετικό ποσοστό (fall-out), με την διαφορά ότι εδώ ο συνδυασμός (εμπειρική συμπλήρωση, τυχαία-υπερδειγματοληψία) έχει το χαμηλότερο και ότι οι προεπεξεργασίες με SMOTE είναι λίγο χαμηλότερα, ενώ αυτές με υποδειγματοληψία λίγο ψηλότερα από ό,τι στην LR5. Οι διαφορές είναι όμως αρκετά μικρές. Ο Δείκτης Fowlkes-Mallows (FM) είναι επίσης παρόμοιος με της LR5. Το ποσοστό ψευδούς παράλειψης (false omission rate) έχει αντίστοιχο μοτίβο με της LR5, με την διαφορά ότι εδώ λόγω απροσδιοριστίας δεν ορίζεται η μετρική για την προεπεξεργασία (εμπειρική συμπλήρωση, class_weights). Έχουμε σχετικά μικρή βελτίωση ως προς την Πληροφοριακότητα (Informedness) και το Συντελεστή συσχέτισης Matthews (MCC) σε σχέση με την LR5, με εξαίρεση τον συνδυασμό (iterative imputation, class_weights). Το Μέτρο του αξιοσημείωτου (markedness) παρουσιάζει επίσης πολύ μικρή βελτίωση σε σχέση με την LR5. Για το ψευδώς αρνητικό ποσοστό (Miss rate) τα αποτελέσματα είναι αντίστοιχα με της LR5, αλλά για (iterative imputation, υποδειγματοληψία), το ψευδώς αρνητικό ποσοστό (Miss rate) είναι μικρότερο συγκριτικά με την LR5. Παρόμοια αποτελέσματα με την LR5 έχουμε και για την αρνητική προγνωστική αξία (negative predictive value) και την θετική προγνωστική αξία (precision). Τα αποτελέσματα είναι παρόμοια με την LR5 για το κατώφλι επικράτησης (prevalence threshold). Η ευαισθησία είναι παρόμοια με της LR5, με διαφορά ότι εδώ ξεχωρίζουν και οι (iterative imputation, υποδειγματοληψία), (iterative imputation, τυχαία υπερδειγματοληψία), (εμπειρική συμπλήρωση, SMOTE-υπερδειγματοληψία), (iterative imputation, SMOTE- υπερδειγματοληψία). Η ειδικότητα (specificity) είναι αντίστοιχη με της LR5, με την διαφορά ότι για (iterative imputation, υποδειγματοληψία) είναι πιο μικρή. Η βαθμολογία απειλής (threat score) είναι επίσης παρόμοια με της LR5, αλλά εδώ οι διαφορές μεταξύ των περιπτώσεων είναι μικρότερες και είναι πιο κοντά στο μέτριο.

Καλύτερα αποτελέσματα έχουμε για (Iterative imputation, υποδειγματοληψία), (εμπειρική συμπλήρωση, τυχαία υπερδειγματοληψία), (iterative imputation, τυχαία υπερδειγματοληψία), (iterative imputation, SMOTE-υπερδειγματοληψία). Ενώ ακολουθούν κοντά κι άλλα.

DT1, DT2

Η ακρίβεια ή ορθότητα (accuracy) και η εξισορροπημένη ακρίβεια (balanced accuracy) είναι μέτρια προς καλή για κάθε προεπεξεργασία, ξεχωρίζουν λίγο οι περιπτώσεις τυχαίας υπερδειγματοληψίας. Ομοίως και η AUC, μόνο που εδώ ξεχωρίζει και η περίπτωση (iterative imputation, χωρίς εξισορρόπηση). Το Σκορ-f1 (f1-score) είναι κι αυτό μέτριο προς καλό, αλλά οι περιπτώσεις υποδειγματοληψίας έχουν λίγο χειρότερες τιμές (άνω του μετρίου), ομοίως και για το Δείκτη Fowlkes-Mallows (FM). Το ψευδώς θετικό ποσοστό (fall-out) είναι χαμηλό γενικά, με υψηλότερο να είναι για περιπτώσεις εξισορρόπησης με class_weights. Η τυχαία υπερδειγματοληψία έχει χαμηλές τιμές FM, αλλά ενδέχεται να οφείλεται σε υπερπροσαρμογή. Μέτριες προς χαμηλες τιμές έχουμε για την υποδειγματοληψία και τη SMOTE-υπερδειγματοληψία. Το ποσοστό ψευδούς παράλειψης (false omission rate) είναι μέτριο προς χαμηλό. Για τυχαία υπερδειγματοληψία είναι χαμηλό αλλά πιθανόν να οφείλεται σε υπερπροσαρμογή. Το DT1 έχει ελάχιστα υψηλότερες τιμές από τα υπόλοιπα. Η Πληροφοριακότητα (Informedness) είναι μέτρια προς χαμηλή και ξεχωρίζουν οι προεπεξεργασίες με τυχαίας υπερδειγματοληψίας. Για iterative imputation έχουμε καλύτερη Πληροφοριακότητα (Informedness) από ό,τι για τις άλλες στρατηγικές συμπλήρωσης κενών, με εξαίρεση τις περιπτώσεις όπου έγινε εξισορρόπηση με SMOTE (αν και η διαφορά είναι μικρή). Για εξισορρόπηση με class_weights η Πληροφοριακότητα (Informedness) δεν ήταν καλή. Η τιμή κ του Cohen (Cohen's kappa) είναι γενικά χαμηλή, με μέτριες τιμές μόνο για τυχαία υπερδειγματοληψία, κάτι που μπορεί και να οφείλεται σε υπερπροσαρμογή, αλλά γενικά παρουσιάζει αντίστοιχο μοτίβο με την Πληροφοριακότητα (Informedness). Το ίδιο ισχύει και για το Μέτρο του αξιοσημείωτου (markedness) και το MCC. Η αρνητική προγνωστική αξία (negative predictive value) είναι μέτρια προς καλή γενικά, με τις περιπτώσεις τυχαίας υπερδειγματοληψίας να ξεχωρίζουν, τις περιπτώσεις έλλειψης εξισορρόπησης και υποδειγματοληψίας να είναι λίγο άνω του μετρίου και τις περιπτώσεις εξισορρόπησης με class_weights ή SMOTE-υπερδειγματοληψία να είναι μέτριες. Το ψευδώς αρνητικό ποσοστό (Miss rate) είναι πιο χαμηλό για class_weights ή για έλλειψη εξισορρόπησης (κάτι που είναι αναμενόμενο), ενώ ακολουθούν με χαμηλό ψευδώς αρνητικό ποσοστό (Miss rate) οι εξισορροπήσεις με υπερδειγματοληψία. Για την υποδειγματοληψία έχουμε χαμηλό προς μέτριο ψευδώς αρνητικό ποσοστό (Miss rate) (κάτι που δεν είναι πολύ καλό). Η θετική προγνωστική αξία (precision) είναι μέτρια προς καλή γενικά. Καλύτερη είναι για την τυχαία υπερδειγματοληψία, που

μπορεί να οφείλεται σε υπερπροσαρμογή. Το κατώφλι επικράτησης (prevalence threshold) είναι χαμηλότερο για τυχαία υπερδειγματοληψία και χαμηλό προς μέτριο για τις υπόλοιπες περιπτώσεις. Ελάχιστα χαμηλότερο είναι για iterative imputation με υποδειγματοληψία / class_weights. Η ευαισθησία (sensitivity) είναι μέτρια προς καλή γενικά, ξεχωρίζουν οι περιπτώσεις χωρίς εξισορρόπηση / με εξισορρόπηση με class_weights, ακολουθούν οι εξισορροπήσεις με υπερδειγματοληψία και τελευταία είναι η εξισορρόπηση με υποδειγματοληψία. Η ειδικότητα (specificity) είναι καλή για τυχαία υπερδειγματοληψία, άνω του μετρίου για υποδειγματοληψία, μέτρια για SMOTE-υπερδειγματοληψία / έλειψη εξισορρόπησης, μέτρια προς χαμηλή για class_weights. Η βαθμολογία απειλής (threat score) είναι μέτρια γενικά, λίγο άνω του μετρίου για έλλειψη εξισορρόπησης ή για class_weights και λίγο κάτω του μετρίου για υποδειγματοληψία. Κάτι τέτοιο δεν αποτελεί έκπληξη, αφού δεν είναι αμερόληπτη, δίνοντας χαμηλότερες βαθμολογίες για σπανιότερα γεγονότα.

Καλύτερα αποτελέσματα έχουμε για (iterative imputation, τυχαία υπερδειγματοληψία), (εμπειρική συμπλήρωση, τυχαία υπερδειγματοληψία), (συνάρτηση, τυχαία υπερδειγματοληψία).

DT3

Ισχύουν ό,τι και για τα DT1, DT2. Το DT3 έχει χαμηλότερο ψευδώς αρνητικό ποσοστό (Miss rate) από τα υπόλοιπα DT, γεγονός που έχει ιδιαίτερο ενδιαφέρον για τις περιπτώσεις SMOTE-υπερδειγματοληψίας. Για Iterative imputation δεν είναι πολύ χαμηλότερη η τιμή, ενώ για τυχαία υπερδειγματοληψία αυξάνει αισθητά.

Καλύτερα αποτελέσματα έχουμε για (iterative imputation, τυχαία υπερδειγματοληψία), (εμπειρική συμπλήρωση, τυχαία υπερδειγματοληψία), (συνάρτηση, τυχαία υπερδειγματοληψία).

DT4

Οι τιμές για ακρίβεια ή ορθότητα (accuracy) και η AUC, η εξισορροπημένη ακρίβεια (balanced accuracy), το Σκορ-f1 (f1-score) είναι μέτρια προς καλά για κάθε προεπεξεργασία, είναι ελάχιστα μικρότερα από των προηγούμενων DT, αλλά παρόμοια. Ο Δείκτης Fowlkes-Mallows (FM) είναι αντίστοιχος με των προηγούμενων DT, αλλά χαμηλότερος, πχ. η τυχαία υπερδειγματοληψία δεν ξεχωρίζει. Το ψευδώς θετικό ποσοστό (fall-out) είναι υψηλότερο από τα προηγούμενα DT (πχ. για class_weights, τυχαία υπερδειγματοληψία) αλλά για περιπτώσεις iterative imputation είναι πιο χαμηλό. Επίσης χαμηλότερο είναι και για (εμπειρική συμπλήρωση, SMOTE-υπερδειγματοληψία). Το ποσοστό ψευδούς παράλειψης (false omission rate) είναι αντίστοιχο με των προηγούμενων DT, αλλά εδώ για τυχαία υπερδειγματοληψία οι τιμές αν και κοντα στο χαμηλό είναι λίγο υψηλότερες από τα προηγούμενα DT. Η Πληροφοριακότητα (Informedness), η τιμή κ του Cohen (Cohen's kappa), το Μέτρο του αξιοσημείωτου (markedness), ο Συντελεστής συσχέτισης Matthews (MCC) είναι αντίστοιχα ή λίγο χαμηλότερα με των προηγούμενων DT, με την class_weights να τα πηγαίνει αισθητά χειρότερα. Η αρνητική προγνωστική αξία (negative predictive value) είναι αντίστοιχη με των προηγούμενων DT, αλλά εδώ είναι λίγο χαμηλότερη. Το ψευδώς αρνητικό ποσοστό (Miss rate) είναι γενικά ψηλότερο από τα προηγούμενα DT (όχι καλό εύρημα), με εξαίρεση την εξισορρόπηση με class_weights όπου μένει ίδιο ή και μειώνεται. Η θετική προγνωστική αξία (precision) είναι μέτρια προς καλή γενικά. Λίγο καλύτερη είναι για υπερδειγματοληψία (SMOTE / τυχαία), ή για έλλειψη εξισορρόπησης. Το κατώφλι επικράτησης (prevalence threshold) είναι χαμηλό (θετικό εύρημα) για υπερδειγματοληψία και για iterative imputation με υποδειγματοληψία (δυστυχώς με μεγάλη τυπική απόκλιση) και χαμηλό προς μέτριο για τις υπόλοιπες περιπτώσεις. Η ευαισθησία (sensitivity) είναι παρόμοια και λίγο χειρότερη από τα προηγούμενα DT, με εξαίρεση την εξισορρόπηση με class_weights, όπου είναι σταθερή προς καλύτερη. Η ειδικότητα (specificity) είναι παρόμοια με αυτή των προηγούμενων DT, αλλά λίγο βελτιωμένη (με εξαίρεση τις περιπτώσεις class_weights, τυχαίας υπερδειγματοληψίας). Η βαθμολογία απειλής (threat score) έχει αντίστοιχο μοτίβο με των προηγούμενων DT, αλλά είναι λίγο μικρότερες οι τιμές της.

Καλύτερα αποτελέσματα έχουμε για τυχαία υπερδειγματοληψία, (iterative imputation, χωρίς εξισορρόπηση), (iterative imputation, υποδειγματοληψία).

RF1, RF2

Οι τιμές που έχουν για τις διάφορες μετρικές είναι παρόμοιες μεταξύ τους. Το RF2 φαίνεται να αποδίδει λίγο καλύτερα από το RF1 απουσία εξισορρόπησης και για (εμπειρική συμπλήρωση, υποδειγματοληψία). Η ακρίβεια ή ορθότητα (accuracy), είναι μέτρια προς καλή περί το 0,75, με τις (συνάρτηση, υποδειγματοληψία), (συνάρτηση, class_weights), τυχαία υπερδειγματοληψία, SMOTE-υπερδειγματοληψία να παρουσιάζουν ενδιαφέρον. Η AUC είναι καλή προς πολύ καλή (>0,75) για κάθε περίπτωση. Ξεχωρίζουν οι περιπτώσεις τυχαίας υπερδειγματοληψίας - που μπορεί να

σχετίζεται με υπερπροσαρμογή - η έλλειψη εξισορρόπησης (που εξηγείται μιας και η χρήση ταξινομητών που βασίζονται σε δέντρα χρησιμοποιείται ούτως ή άλλως για αντιμετώπιση του προβλήματος που προκύπτει από την ανισορροπία των κλάσεων. Πολύ καλά ως προς αυτή τη μετρική τα πηγαίνουν και οι κατηγορίες που εξισορροπήθηκαν με class_weights, υποδειγματοληψία, SMOTE-υπερδειγματοληψία. Η εξισορροπημένη ακρίβεια (balanced-accuracy) είναι μέτρια προς καλή για τις περιπτώσεις όπου έγινε εξισορρόπηση. Για απουσία εξισορρόπησης καλύτερα τα πηγαίνει το RF2, ενώ το RF1 δίνει μέτριες τιμές. Ξεχωρίζει η περίπτωση (συνάρτηση, υποδειγματοληψία), οι υπόλοιπες περιπτώσεις υποδειγματοληψίας και υπερδειγματοληψίας. Ακολουθουν οι class_weights με τον συνδυασμό (συνάρτηση, class_weights) να τα πηγαίνει σχετικά καλύτερα. Το Σκορ-f1 (f1-score) έχει τιμές περί το 0,75 γενικά, με το RF1 να τα πηγαίνει ελαφρώς καλύτερα από το RF2 σε κάποιες περιπτώσεις (πχ. για υποδειγματοληψία). Η έλλειψη εξισορρόπησης ξεχωρίζει, ακολουθούν οι περιπτώσεις εξισορρόπησης με class_weights και SMOTE-υπερδειγματοληψία. Το μικρότερο ψευδώς θετικό ποσοστό (fall-out) εμφανίζεται για υποδειγματοληψία (ιδιαίτερα για συμπλήρωση με την συνάρτηση που δοκιμάζεται), για τυχαία υπερδειγματοληψία (που μπορεί να σχετίζεται με υπερπροσαρμογή) και για SMOTE-υπερδειγματοληψία. Χειρότερο (ψηλότερο) ψευδώς θετικό ποσοστό (fall-out) έχει η περίπτωση έλλειψης εξισορρόπησης (με χειρότερο το RF1). Το ποσοστό ψευδούς παράλειψης (false omission rate) είναι (καλώς) χαμηλό (περί το 0,25). Χαμηλότερο είναι για περιπτώσεις τυχαίας υπερδειγματοληψίας, για (συνάρτηση, υποδειγματοληψία) και γενικά για υποδειγματοληψία, (συνάρτηση, class_weights). Ο Δείκτης Fowlkes-Mallows (FM) είναι καλός (περί το 0,75) γενικά. Λίγο χειρότερο είναι για υποδειγματοληψία (ιδιαίτερα στο RF2 για εμπειρική συμπλήρωση και iterative imputation). Οι καλύτερες τιμές της Πληροφοριακότητας (Informedness) είναι περί του μετρίου για υποδειγματοληψία, τυχαία υπερδειγματοληψία και λίγο κάτω από το μέτριο για SMOTE-υπερδειγματοληψία. Η τιμή κ του Cohen (Cohen's kappa) πλησιάζει το μέτριο για τυχαία υπερδειγματοληψία, ακολουθεί η SMOTE-υπερδειγματοληψία, η υποδειγματοληψία (για το RF1 λίγο καλύτερο), όπου ο συνδυασμός (συνάρτηση, υποδειγματοληψία) πλησιάζει το μέτριο αλλά έχει μεγάλη τυπική απόκλιση. Το Μέτρο του αξιοσημείωτου (markedness) και ο Συντελεστής συσχέτισης Matthews (MCC) είναι περί του μετρίου, με την τυχαία υπερδειγματοληψία να πηγαίνει καλύτερα από το SMOTE και την υποδειγματοληψία και τους συνδυασμούς (συνάρτηση, υποδειγματοληψία), (συνάρτηση, class_weights) να ξεχωρίζουν με τιμές λίγο άνω του μετρίου αλλά μεγάλες τυπικές αποκλίσεις. Το χαμηλότερο ψευδώς αρνητικό ποσοστό (Miss rate) (προς ελάχιστο) εμφανίζεται για έλλειψη εξισορρόπησης και για εξισορρόπηση με class_weights. Χαμηλό είναι για υπερδειγματοληψία, (συνάρτηση, υποδειγματοληψία) -αλλά με μεγάλη τυπική απόκλιση. Η αρνητική προγνωστική αξία (negative predictive value) είναι καλή (περί το 0,75) και ξεχωρίζει για (συνάρτηση, class_weights), (συνάρτηση, υποδειγματοληψία), την τυχαία υπερδειγματοληψία, ενώ αντίστοιχα καλή είναι και για έλλειψη εξισορρόπησης, αλλά έχει πολύ μεγάλη τυπική απόκλιση. Η θετική προγνωστική αξία (precision) είναι μέτρια προς καλή. Καλή (περί το 0,75) είναι για SMOTE-υπερδειγματοληψία, (συνάρτηση, υποδειγματοληψία) και ακολουθούν οι υπόλοιπες υποδειγματοληψίες και τυχαίες υπερδειγματοληψίες. Για έλλειψη εξισορρόπησης το RF2 είναι λίγο καλύτερο. Το χαμηλότερο κατώφλι επικράτησης (prevalence threshold) υπάρχει για τυχαία υπερδειγματοληψία, SMOTE-υπερδειγματοληψία, (συνάρτηση, υποδειγματοληψία), κοντά στο 0,25. Η υψηλότερη (καλή προς πολύ καλή) ευαισθησία (sensitivity), υπάρχει για έλλειψη εξισορρόπησης (κάτι που είναι αναμενόμενο, αφού η κλάση 1 είναι η πλειοψηφούσα), καθώς και για class_weights. Ακολουθούν με τιμές περί του 0,75 οι περιπτώσεις υπερδειγματοληψίας, (συνάρτηση, υποδειγματοληψία) -με μεγαλη τυπική απόκλιση- και ύστερα οι υπόλοιπες υποδειγματοληψίες. Η καλύτερες ειδικότητες, εμφανίζονται με τιμές περί το 0,75 για (συνάρτηση, υποδειγματοληψία), τυχαία υπερδειγματοληψία, SMOTE-υπερδειγματοληψία, υποδειγματοληψία. Για class_weights οι τιμές είναι μέτριες και για έλλειψη εξισορρόπησης είναι χαμηλές προς το ελάχιστο (RF1). Η βαθμολογία απειλής (threat score) είναι μέτρια προς καλή. Κοντά στο μέτριο είναι οι τιμες για τις περιπτώσεις υποδειγματοληψίας, πλην της (συνάρτηση, υποδειγματοληψία), όπου η τιμή είναι αντίστοιχα καλή με τις άλλες κατηγορίες, αλλά έχει μεγάλη τυπική απόκλιση. Από ό,τι φαίνεται ο συνδυασμός (συνάρτηση, υποδειγματοληψία) φαίνεται να συμβιβάζει την μέτρια προς καλή ταξινόμηση και των δύο κλάσεων.

Καλύτερα αποτελέσματα είχαμε για (συνάρτηση, υποδειγματοληψία), τυχαία υπερδειγματοληψία και με σχετικά μικρή διαφορά SMOTE-υπερδειγματοληψία, οι υπόλοιπες υποδειγματοληψίες, (συνάρτηση, class_weights).

Η επιλογή εξισορρόπησης μέσω της παραμέτρου class_weights δεν υπήρχε στις ακόλουθες μεθόδους ταξινόμησης, συνεπώς δεν αναφέρεται. Πιθανώς να υπήρχε καλή μέση τιμή των μετρικών για χρήση των class_weights, λόγω του ότι λάμβανε υπ'όψη μόνο τις πιο πάνω μεθόδους -που είχαν καλή σχετικά απόδωση. Ιδιαίτερα τα RF ανέβαζαν την μέση τιμή. Οι μέθοδοι που έπονται δεν είχαν τόσο μεγάλη επιρροή από την συμμετοχή των RF στην διαμόρφωση του μέσου όρου των μετρικών τους, λόγω της ύπαρξης και άλλων ταξινομητριών που τον διαμορφώνουν, οι οποίες είχαν ενδεχομένως μικρότερη απόδωση.

NN1-10

Η ακρίβεια ή ορθότητα (accuracy) είναι μέτρια προς καλή γενικά. Στα NN3-8, NN10, είναι καλή για την περίπτωση (εμπειρική εξισορρόπηση, τυχαία υπερδειγματοληψία). Στο NN5-7 είναι καλή και η (συνάρτηση, τυχαία υπερδειγματοληψία). Στο NN9 όλες οι περιπτώσεις τυχαίας υπερδειγματοληψίας, είναι κοντά στο να θεωρούνται καλές. Η τυχαία υπερδειγματοληψία είναι επιρρεπής στην υπερπροσαρμογή, οπότε αυτά τα ευρήματα δεν αξιολογούνται σαν εξαιρετικά ενδιαφέροντα. Παρομοίως και οι AUC είναι άνω του μετρίου, με εξαίρεση τις περιπτώσεις όπου έχει γίνει τυχαία υπερδειγματοληψία όπου έχουν λίγο υψηλότερες τιμές. Το NN9 έχει πολύ ελαφρά ψηλότερες τιμές από τα υπόλοιπα NN, καθώς και το NN10 για τις περιπτώσεις όπου έχει γίνει Iterative imputation. Αντίστοιχα ευρήματα έχουμε και για την εξισορροπημένη ακρίβεια (balanced accuracy). Στο NN9 οι περιπτώσεις όπου έχει γίνει iterative imputation τα πηγαίνουν ελαφρώς καλύτερα σε σύγκριση με τις άλλες μεθόδους συμπλήρωσης κενών (για κάθε περίπτωση εξισορρόπησης). Ενδιαφέρον έχει ότι στο NN9 εμφανίζεται λίγο πίσω από τις πρώτες τιμές η τιμή των (iterative imputation, SMOTE-υπερδειγματοληψία), (iterative imputation, υποδειγματοληψία). Το Σκορ-f1 (f1-score) είναι άνω του μετρίου προς καλό, με ελαφρώς καλύτερη την τυχαία υπερδειγματοληψία, ενώ ακολουθεί η SMOTE-υπερδειγματοληψία και η υποδειγματοληψία με μικρή διαφορά. Η διαφορά τους είναι πολύ μικρή στο NN1. Ενδιαφέρον παρουσιάζει το γεγονός ότι στα NN3-4 το Σκορ-f1 (f1-score) είναι ελαφρά καλύτερο από τα προηγούμενα και ανάμεσα στις πρώτες θέσεις είναι μαζί με την τυχαία υπερδειγματοληψία και τα (iterative imputation, SMOTE-υπερδειγματοληψία), (συνάρτηση, SMOTE-υπερδειγματοληψία). Κάτι αντίστοιχο συμβαίνει και στα NN7, NN9-10, όπου και η (εμπειρική συμπλήρωση, SMOTE-υπερδειγματοληψία) εμφανίζεται επίσης καλύτερη. Τα NN4, NN9 έχουν το περισσότερο ενδιαφέρον. Χαμηλό ψευδώς θετικό ποσοστό (fall-out) έχουμε μόνο για (εμπειρική συμπλήρωση, τυχαία υπερδειγματοληψία) και ακολουθεί η (συνάρτηση, τυχαία υπερδειγματοληψία) και η (iterative imputation, τυχαία υπερδειγματοληψία). Ενδιαφέρον έχουν τα NN7 όπου χαμηλώνουν ελαφρά όλες οι τιμές της μετρικής αυτής, το NN9, όπου είναι μέτρια προς χαμηλή και η (iterative imputation, υποδειγματοληψία) -αλλά με μεγάλη τυπική απόκλιση- και το NN3 όπου το ίδιο συμβαίνει με την (συνάρτηση, υποδειγματοληψία). Το ποσοστό ψευδούς παράλειψης (false omission rate) είναι χαμηλότερο για την τυχαία υπερδειγματοληψία, ακολουθεί η υποδειγματοληψία και η SMOTE-υπερδειγματοληψία. Ενδιαφέρον έχουν τα NN9-10, όπου για (iterative imputation, υποδειγματοληψία) η τιμή είναι μέτρια προς χαμηλή, αλλά δυστυχώς με μεγάλη τυπική απόκλιση. Ο Δείκτης Fowlkes-Mallows (FM) είναι ελαφρά καλύτερος για έλλειψη εξισορρόπησης και τυχαία υπερδειγματοληψία, έπεται η SMOTE-υπερδειγματοληψία και η υποδειγματοληψία. Ξεχωρίζει στο NN9 και στο NN4 η περίπτωση (iterative imputation, SMOTE-υπερδειγματοληψία). Ενδιαφέρον έχει το NN1, όπου το SMOTE τα πηγαίνει καλύτερα από την τυχαία υπερδειγματοληψία. Αντίστοιχα κυμαίνεται και η Πληροφοριακότητα (Informedness), αλλά οι τιμές της είναι χαμηλές γενικά. Άνω του χαμηλού είναι οι τιμές για (εμπειρική συμπλήρωση, τυχαία υπερδειγματοληψία), βλ. NN4, NN6-NN8, NN10. Ενδιαφέρον έχει το NN9, όπου οι περιπτώσεις (iterative imputation, υποδειγματοληψία), (iterative imputation, SMOTE-υπερδειγματοληψία), (iterative imputation, τυχαία υπερδειγματοληψία) βελτιώνονται αισθητά. Η τιμή κ του Cohen (Cohen's kappa) παρουσιάζει αντίστοιχο μοτίβο, παίρνοντας την ψηλότερη τιμή για NN6 και (εμπειρική συμπλήρωση, τυχαία υπερδειγματοληψία). Όπως και για την Πληροφοριακότητα (Informedness) το NN9 έχει ενδιαφέρον. Το Μέτρο του αξιοσημείωτου (markedness) έχει παρόμοιο μοτίβο με την Πληροφοριακότητα (Informedness). Ενδιαφέρον έχει το γεγονός ότι για (iterative imputation, χωρίς εξισορρόπηση) η τιμή του είναι χαμηλή, αλλά σταθερή (βέβαια έχει μεγάλη τυπική απόκλιση). Στα NN2-5 και NN9 χαμηλή τιμή παρουσιάζει και η (iterative imputation, SMOTE-υπερδειγματοληψία). Στο NN9 χαμηλή τιμή παρουσιάζει και η (iterative imputation, υποδειγματοληψία). Ο συντελεστής συσχέτισης Matthews (MCC) έχει παρόμοιο μοτίβο με την Πληροφοριακότητα (Informedness). Καλύτερη τιμή (λίγο άνω του χαμηλού) έχει για τυχαία υπερδειγματοληψία γενικά, με την περίπτωση της εμπειρικής συμπλήρωσης να έχει τα πρωτεία, με τιμές λίγο άνω του μετρίου. Παρόλα αυτά οι τιμές του δεν είναι καλές γενικά. Γενικά συμπεραίνουμε ότι εάν επρόκειτο να χρησιμοποιηθεί κάποια μορφή προεπεξεργασίας πλην της τυχαίας υπερδειγματοληψίας και ήταν αναγκαίο να ταξινομήσουμε με NN, καλύτερη πιθανότητα να αποδώσει έστω και λίγο θα είχε το NN9 και ίσως τα NN3, NN4, NN1. Η βαθμολογία απειλής (threat score) τείνει να είναι καλύτερη για iterative imputation ανά κατηγορία εξισορρόπησης, με εξαίρεση την τυχαία υπερδειγματοληψία, όπου επικρατεί εννίοτε και η εμπειρική συμπλήρωση. Κάτι τέτοιο πιθανόν να είναι ένδειξη ότι για τυχαία υπερδειγματοληψία έχουμε υπερπροσαρμογή στα δεδομένα. Γενικά καλύτερη βαθμολογία απειλής (threat score) έχει η έλλειψη εξισορρόπησης, η τυχαία υπερδειγματοληψία, ακολουθούν οι SMOTE-υπερδειγματοληψία και η υποδειγματοληψία. Οι τιμές της είναι περί του μετρίου. Η καλύτερη απόδωση της έλλειψης εξισορρόπησης στην μετρική αυτή αποδίδεται στην ανισορροπία των κλάσεων. Το NN9 έχει καλύτερη τιμή για (εμπειρική συμπλήρωση, SMOTE-υπερδειγματοληψία). Το κατώφλι επικράτησης (prevalence-threshold) έχει το αντίστροφο μοτίβο της βαθμολογίας απειλής (threat-score). Το ποσοστό ψευδούς παράλειψης (false omission rate) είναι γενικά καλύτερο για τυχαία υπερδειγματοληψία, ακολουθεί η υποδειγματοληψία (όπου για iterative imputation έχει εξίσου καλές τιμές στα NN9-10) και μετά η SMOTE-υπερδειγματοληψία. Τελευταία είναι η έλλειψη εξισορρόπησης. Το ψευδώς αρνητικό ποσοστό (Miss rate) είναι ιδιαίτερα χαμηλό για (iterative imputation, χωρίς εξισορρόπηση), κάτι το αναμενόμενο. Επίσης είναι καλύτερο για iterative imputation στην πλειοψηφία των περιπτώσεων. Για (iterative imputation, υποδειγματοληψία) έχουμε χαρακτηριστικά μεγάλη τυπική απόκλιση γενικά, αλλά σχετικά καλές τιμές για NN1, NN4, NN6. Για τυχαία υπερδειγματοληψία έχουμε τις καλύτερες τιμές μετά την

έλλειψη εξισορρόπησης, σχεδόν για κάθε NN (ιδιαίτερα για iterative imputation). Η SMOTE-υπερδειγματοληψία ακουλουθεί με το iterative imputation να έχει αντίστοιχα καλές τιμές για NN3, NN4, NN7. Η αρνητική προγνωστική αξία (negative predictive value) είναι καλύτερη για τυχαία υπερδειγματοληψία, ενώ ακολουθεί η υποδειγματοληψία (ειδικά για iterative imputation οι τιμές είναι καλές) και έπειτα βρίσκεται η SMOTE-υπερδειγματοληψία. Χαρακτηριστικά παρόμοια με την (iterative imputation, υποδειγματοληψία) τα πηγαίνει και η (iterative imputation, χωρίς εξισορρόπηση), αλλά με μεγάλη τυπική απόκλιση. Η θετική προγνωστική αξία (precision) είναι μέτρια προς καλή γενικά και είναι χαρακτηριστικά καλή για το συνδυασμό (εμπειρική συμπλήρωση, τυχαία υπερδειγματοληψία). Η έλλειψη εξισορρόπησης φαίνεται να τα πηγαίνει παρόμοια με τη SMOTE-υπερδειγματοληψία. Η υποδειγματοληψία πετυχαίνει τιμές άνω του μετρίου, με εξαίρεση το (iterative imputation, υποδειγματοληψία), που συμβαδίζει με τις άλλες κατηγορίες στο NN9. Η ειδικότητα (specificity) είναι χαρακτηριστικά υψηλή για τυχαία υπερδειγματοληψία (ιδιαίτερα για εμπειρική συμπλήρωση). Στα NN10,NN6-8 οι περιπτώσεις SMOTE-υπερδειγματοληψία, υποδειγματοληψία φαίνεται να έχουν αντίστοιχα αποτελέσματα. Στο NN1-4, NN9 η υποδειγματοληψία υπερτερεί ελαφρώς. Στο NN5 το SMOTE υπερτερεί. Για έλλειψη εξισορρόπησης οι τιμές δεν είναι αξιόλογες. Η ευαισθησία (sensitivity) είναι καλύτερη για το iterative imputation σε σχέση με άλλες μεθόδους συμπλήρωσης κενών. Η ευαισθησία είναι μέτρια προς καλή γενικά. Είναι καλύτερη για περιπτώσεις έλλειψης εξισορρόπησης, όπως και αναμενόταν. Ακολουθούν με μικρές διαφορές οι υπερδειγματοληψίες και έπειτα οι υποδειγματοληψίες.

Στο NN1 τα πάει καλύτερα η (εμπειρική συμπλήρωση, τυχαία υπερδειγματοληψία), τα υπόλοιπες τυχαίες υπερδειγματοληψίες και ακολουθεί η (εμπειρική συμπλήρωση, SMOTE-υπερδειγματοληψία), η (iterative imputation, υποδειγματοληψία) κι άλλες. Στο NN2, τα πάει καλύτερα η (εμπειρική συμπλήρωση, τυχαία υπερδειγματοληψία), οι υπόλοιπες τυχαίες υπερδειγματοληψίες και ακολουθεί η (iterative imputation, υποδειγματοληψία) και η (iterative imputation, SMOTE-υπερδειγματοληψία). Στο NN3 καλύτερα τα πάει και εδώ η (εμπειρική συμπλήρωση, τυχαία υπερδειγματοληψία), οι υπόλοιπες τυχαίες υπερδειγματοληψίες και ακολουθεί η (iterative imputation, SMOTE-υπερδειγματοληψία), η (συνάρτηση, υποδειγματοληψία) και η (iterative imputation, υποδειγματοληψία). Στο NN4 καλύτερα τα πάει και εδώ η (εμπειρική συμπλήρωση, τυχαία υπερδειγματοληψία), οι υπόλοιπες τυχαίες υπερδειγματοληψίες και ακολουθεί η (συνάρτηση, SMOTE-υπερδειγματοληψία) και η (iterative imputation, SMOTE-υπερδειγματοληψία). Τα NN5-8 τα πάνε καλύτερα αντίστοιχα για (εμπειρική συμπλήρωση, τυχαία υπερδειγματοληψία) και για τις υπόλοιπες τυχαίες υπερδειγματοληψίες. Το NN9 καλύτερα τα πάει για (εμπειρική συμπλήρωση, τυχαία υπερδειγματοληψία), οι υπόλοιπες τυχαίες υπερδειγματοληψίες και το Iterative imputation μαζί με τη SMOTE-υπερδειγματοληψία / υποδειγματοληψία. Το NN10 καλύτερα τα πάει για (εμπειρική συμπλήρωση, τυχαία υπερδειγματοληψία) και για τις υπόλοιπες τυχαίες υπερδειγματοληψίες.

KNN1-6

Η ακρίβεια ή ορθότητα (accuracy) είναι μέτρια προς καλή, γενικά, αλλά για KNN1 είναι σχετικά χαμηλότερη. Η εξισορροπημένη ακρίβεια (balanced accuracy) είναι πανομοιότυπου μοτίβου με την ακρίβεια ή ορθότητα (accuracy), αλλά έχει λίγο χαμηλότερες τιμές (άνω του μετρίου). Χαμηλότερο ψευδώς θετικό ποσοστό (fall-out) έχουμε για περιπτώσεις εμπειρικής συμπλήρωσης κενών και για εξισορρόπηση (καλύτερα τα πηγαίνει η τυχαία υπερδειγματοληψία, ακολουθεί η υποδειγματοληψία και η SMOTE-υπερδειγματοληψία). Η Πληροφοριακότητα (Informedness) είναι χαμηλή προς μέτρια στις καλύτερες δυνατές τιμές που παίρνουν τα KNN. Οι υπόλοιπες τιμές είναι πολύ μικρές. Γενικά η βαθμολογία απειλής (threat score) δεν είναι καλή για τις μεθόδους με εξισορρόπηση που δεν είναι επιρρεπείς στην υπερπροσαρμογή στα δεδομένα (overfitting). Βέβαια, σαν μετρική δεν επιρρεάζεται από την ανισορροπία, για αυτό και δίνει καλές τιμές για έλλειψη εξισορρόπησης γενικά.

KNN1

Η ακρίβεια ή ορθότητα (accuracy) είναι καλύτερη για τυχαία υπερδειγματοληψία, καθώς και για τον συνδυασμό (εμπειρική συμπλήρωση, υποδειγματοληψία). Ακολουθούν οι περιπτώσεις έλλειψης εξισορρόπησης (για iterative imputation ή εμπειρική συμπλήρωση) και (iterative imputation,

υποδειγματοληψία), (εμπειρική συμπλήρωση, SMOTE-υπερδειγματοληψία). Η AUC είναι καλύτερη για τυχαία υπερδειγματοληψία, ενώ η εμπειρική συμπλήρωση κενών δείχνει να τα πηγαίνει καλύτερα από τις άλλες δύο. Για το Σκορ-f1 (f1-score) όλες οι τιμές σε περιπτώσεις όπου έχει γίνει εξισορρόπηση είναι άνω του μετρίου, με εξαίρεση για την περίπτωση (συνάρτηση, υποδειγματοληψία). Για το Δείκτη Fowlkes-Mallows (FM) έχουμε μέτριες προς καλές, παρόμοιες τιμές γενικά. Οι εξαιρέσεις είναι ότι καλές τιμές έχουμε για (iterative imputation, χωρίς εξισορρόπηση) και μέτριες για (συνάρτηση, υποδειγματοληψία). Χαμηλό ποσοστό ψευδούς παράλειψης (false omission rate) έχουμε για τυχαία υπερδειγματοληψία και ακουλουθούν η εμπειρική συμπλήρωση ή iterative imputation με υποδειγματοληψία. Η Πληροφοριακότητα (Informedness), ο Συντελεστής συσχέτισης Matthews (MCC) και η τιμή κ του Cohen (Cohen's kappa) είναι χαμηλή προς μέτρια για τυχαία υπερδειγματοληψία, (εμπειρική συμπλήρωση, υποδειγματοληψία). Το Μέτρο του αξιοσημείωτου (markedness) είναι χαμηλό προς μέτριο για (εμπειρική συμπλήρωση, υποδειγματοληψία), (εμπειρική συμπλήρωση, τυχαία υπερδειγματοληψία) και ακολουθούν οι υπόλοιπες τυχαίες υπερδειγματοληψίες και η (εμπειρική συμπλήρωση, SMOTE-υπερδειγματοληψία). Μέτρια προς καλή αρνητική προγνωστική αξία (negative predictive value) έχουμε για τυχαία υπερδειγματοληψία, (εμπειρική συμπλήρωση, υποδειγματοληψία), (iterative imputation, υποδειγματοληψία). Η θετική προγνωστική αξία (precision) είναι καλή για εμπειρική συμπλήρωση κενών συνδυασμένη με εξισορρόπηση και ακολουθούν οι υπόλοιπες περιπτώσεις με μέτριες προς καλές τιμές. Καλή ευαισθησία (sensitivity) έχουμε για (iterative imputation, χωρίς εξισορρόπηση) και χαμηλή για (συνάρτηση, υποδειγματοληψία), κάτι που αναμέναμε. Καλή ειδικότητα (specificity) έχουμε για εμπειρική συμπλήρωση κενών με εξισορρόπηση και για τυχαία υπερδειγματοληψία. Καλή προς πολύ καλή είναι η ευαισθησία για (εμπειρική συμπλήρωση, τυχαία υπερδειγματοληψία). Για τις υπόλοιπες περιπτώσεις εξισορρόπησης οι τιμές της ευαισθησίας είναι μέτριες προς καλές και για έλλειψη εξισορρόπησης είναι μέτριες μόνο για την εμπειρική συμπλήρωση και χαμηλές γενικά. Για τις υπόλοιπες περιπτώσεις οι τιμές περί του μετρίου. Η βαθμολογία απειλής (threat score) είναι κάτω του μετρίου με εξαίρεση την (iterative imputation, χωρίς εξισορρόπηση). Χαμηλό (καλώς) ψευδώς αρνητικό ποσοστό (Miss rate) έχουμε μόνο για (εμπειρική συμπλήρωση, χωρίς εξισορρόπηση). Το κατώφλι επικράτησης (prevalence threshold) είναι χαμηλό (καλώς), αλλά με μεγάλη τυπική απόκλιση, μόνο για την εμπειρική συμπλήρωση συνδυασμένη με εξισορρόπηση.

Καλύτερα τα πηγαίνουν τα (Iterative imputation, τυχαία υπερδειγματοληψία), (εμπειρική συμπλήρωση, υποδειγματοληψία) και οι υπόλοιπες τυχαίες υπερδειγματοληψίες.

KNN2

Η ακρίβεια ή ορθότητα (accuracy) είναι καλύτερη για την τυχαία υπερδειγματοληψία (ιδιαίτερα για iterative imputation) και για έλλειψη εξισορρόπησης. Για την AUC ισχύει ό,τι και για το KNN1. Η εξισορροπημένη ακρίβεια (balanced accuracy) είναι λίγο καλύτερη σε σχέση με το KNN1. Η εξισορροπημένη ακρίβεια (balanced accuracy) είναι ελαφρώς καλύτερη για την τυχαία υπερδειγματοληψία, ενώ η εμπειρική συμπλήρωση κενών δείχνει να τα πηγαίνει καλύτερα από τις άλλες δύο γενικά. Καλύτερο Σκορ-f1 (f1-score) έχουμε για iterative imputation, υπερδειγματοληψία (SMOTE / τυχαία). Ο Δείκτης Fowlkes-Mallows (FM) είναι γενικά μέτριος προς καλός. Ξεχωρίζουν οι περιπτώσεις έλλεψης εξισορρόπησης (αναμενόμενο), (iterative imputation, υπερδειγματοληψία). Χαμηλό ποσοστό ψευδούς παράλειψης (false omission rate) έχουμε όπως και στην KNN1 για τυχαία υπερδειγματοληψία (για iterative imputation είναι η χαμηλότερη τιμή) και ακολουθούν οι υποδειγματοληψίες και η (iterative imputation, χωρίς εξισορόπηση) -η οποία όμως έχει πολύ μεγάλη τυπική απόκλιση. Η Πληροφοριακότητα (Informedness), ο Συντελεστής συσχέτισης Matthews (MCC) και η τιμή κ του Cohen (Cohen's kappa) είναι χαμηλές προς μέτριες για (iterative imputation, τυχαία υπερδειγματοληψία) και ακολουθούν οι υπόλοιπες τυχαίες υπερδειγματοληψίες. Το Μέτρο του αξιοσημείωτου (markedness) είναι χαμηλό προς μέτριο για (iterative imputation, τυχαία υπερδειγματοληψία) και για (εμπειρική συμπλήρωση, τυχαία υπερδειγματοληψία). Χαμηλό (καλώς) ψευδώς αρνητικό ποσοστό (Miss rate) έχουμε μόνο για έλλειψη εξισορρόπησης και ακολουθεί η (iterative imputation, τυχαία υπερδειγματοληψία). Καλή αρνητική προγνωστική αξία (negative predictive value) έχουμε για (iterative imputation, τυχαία υπερδειγματοληψία). Μέτρια προς καλή είναι για τις υπόλοιπες τυχαίες υπερδειγματοληψίες, υποδειγματοληψίες, για την (iterative imputation, έλλειψη εξισορρόπησης) - αλλά η τελευταία έχει μεγάλη τυπική απόκλιση. Η θετική προγνωστική αξία (precision) είναι καλή για (εμπειρική συμπλήρωση,SMOTE-υπερδειγματοληψία) και ακολουθούν οι (iterative imputation, SMOTE-υπερδειγματοληψία), η τυχαία υπερδειγματοληψία με εμπειρική συμπλήρωση ή iterative imputation και η (εμπειρική συμπλήρωση, χωρίς εξισορρόπηση). Ακολουθούν οι υπόλοιπες περιπτώσεις με μέτριες προς καλές τιμές. Το κατώφλι επικράτησης (prevalence threshold) είναι χαμηλό προς μέτριο γενικά. Καλή προς πολύ καλή ευαισθησία (sensitivity) έχουμε για έλλειψη εξισορρόπησης και καλή για Iterative imputation γενικά. Για τις υπόλοιπες περιπτώσεις οι τιμές είναι άνω του μετρίου προς καλές. Το μοτίβο της ειδικότητας (specificity) είναι πανομοιότυπο με αυτό στο KNN1, αλλά οι τιμες είναι μικρότερες γενικά. Η βαθμολογία απειλής (threat score) είναι περί του μετρίου με εξαίρεση τις περιπτώσεις χωρίς εξισορρόπηση και τις (iterative imputation, τυχαία υπερδειγματοληψία), (iterative imputation, SMOTE-υπερδειγματοληψία).

Καλύτερα τα πηγαίνουν οι: (Iterative imputation, τυχαία υπερδειγματοληψία), (εμπειρική συμπλήρωση, τυχαία υπερδειγματοληψία), (δοκιμαζόμενη συνάρτηση, τυχαία υπερδειγματοληψία), (εμπειρική συμπλήρωση, υποδειγματοληψία), οι υπόλοιπες υποδειγματοληψίες και SMOTE-υπερδειγματοληψίες (αλλα όχι η δοκιμαζόμενη συνάρτηση, SMOTE-υπερδειγματοληψία).

KNN3

Ισχύουν ό,τι και για την KNN2. Καλύτερο Σκορ-f1 (f1-score) έχουμε για (iterative imputation, τυχαία υπερδειγματοληψία). Χαμηλό ποσοστό ψευδούς παράλειψης (false omission rate) έχουμε για (iterative imputation, τυχαία υπερδειγματοληψία) και ακολουθούν οι υπόλοιπες τυχαίες υπερδειγματοληψίες και οι (εμπειρική συμπλήρωση, υποδειγματοληψία), (συνάρτηση, υποδειγματοληψία). Ο Δείκτης Fowlkes-Mallows (FM) είναι γενικά μέτριος προς καλός. Ξεχωρίζουν οι περιπτώσεις έλλεψης εξισορρόπησης (αναμενόμενο), (iterative imputation, τυχαία υπερδειγματοληψία). Η Πληροφοριακότητα (Informedness), ο Συντελεστής συσχέτισης Matthews (MCC), το Μέτρο του αξιοσημείωτου (markedness) και η τιμή κ του Cohen (Cohen's kappa) είναι χαμηλές προς μέτριες για (iterative imputation, τυχαία υπερδειγματοληψία) και ακολουθούν οι υπόλοιπες τυχαίες υπερδειγματοληψίες και η (εμπειρική συμπλήρωση, υποδειγματοληψία). Χαμηλό (καλώς) ψευδώς αρνητικό ποσοστό (Miss rate) έχουμε μόνο για έλλειψη εξισορρόπησης και ακολουθεί η (iterative imputation, τυχαία υπερδειγματοληψία). Καλή αρνητική προγνωστική αξία (negative predictive value) έχουμε για (iterative imputation, τυχαία υπερδειγματοληψία). Μέτρια προς καλή είναι για τις υπόλοιπες τυχαίες υπερδειγματοληψίες, τις υποδειγματοληψίες, την (iterative imputation, έλλειψη εξισορρόπησης) - αλλά η τελευταία έχει μεγάλη τυπική απόκλιση. Η θετική προγνωστική αξία (precision) είναι καλή για (εμπειρική συμπλήρωση,SMOTE-υπερδειγματοληψία) και μέτρια για (iterative imputation, υποδειγματοληψία) και λίγο άνω του μετρίου για (δοκιμαζόμενη συνάρτηση, υποδειγματοληψία). Οι υπόλοιπες περιπτώσεις έχουν μέτριες προς καλές τιμές. Το κατώφλι επικράτησης (prevalence threshold) είναι χαμηλό προς μέτριο γενικά. Καλή προς πολύ καλή ευαισθησία (sensitivity) έχουμε για έλλειψη εξισορρόπησης και καλή για (Iterative imputation, τυχαία υπερδειγματοληψία). Για τις υπόλοιπες περιπτώσεις οι τιμές ευαισθησίας είναι άνω του μετρίου. Το μοτίβο της ειδικότητας (specificity) μοιάζει με αυτό στο KNN1, αλλά εδώ ισοψηφεί με την πρώτη και η (δοκιμαζόμενη συνάρτηση, τυχαία υπερδειγματοληψία). Η βαθμολογία απειλής (threat score) είναι κάτω του μετρίου με εξαίρεση τις περιπτώσεις χωρίς εξισορρόπηση και την (iterative imputation, τυχαία υπερδειγματοληψία).

Καλύτερα τα πηγαίνουν τα (Iterative imputation, τυχαία υπερδειγματοληψία), οι υπόλοιπες τυχαίες υπερδειγματοληψίες και η εμπειρική συμπλήρωση μαζί με υποδειγματοληψία. Ακολουθεί η εμπειρική συμπλήρωση μαζί με SMOTE-υπερδειγματοληψία.

KNN4

Ισχύει ό,τι και για την KNN2. Για το Σκορ-f1 (f1-score) ισχύει ο,τι και για την KNN3. Ως προς το ψευδώς θετικό ποσοστό (fall-out) ισχύει ό,τι και για τα προηγούμενα KNN, αλλά το πιο χαμηλό είναι για (συνάρτηση, τυχαία υπερδειγματοληψία). Ο Δείκτης Fowlkes-Mallows (FM) έχει μικρές διαφορές με τον αντίστοιχο του KNN3. Για το ποσοστό ψευδούς παράλειψης (false omission rate) ισχύει ό,τι και στην KNN3 ως προς τα ενδιαφέροντα μέρη, με την διαφορά ότι η έλλειψη εξισορρόπησης έχει κι αυτή μέτριες προς χαμηλές τιμές πίσω από τις προηγούμενες. Η Πληροφοριακότητα (Informedness) είναι χαμηλή προς μέτρια για (iterative imputation, τυχαία υπερδειγματοληψία) και ακολουθούν οι υπόλοιπες τυχαίες υπερδειγματοληψίες και η (εμπειρική συμπλήρωση, υποδειγματοληψία). Η τιμή κ του Cohen (Cohen's kappa) είναι χαμηλή προς μέτρια για (iterative imputation, τυχαία υπερδειγματοληψία) και ακολουθούν οι υπόλοιπες τυχαίες υπερδειγματοληψίες. Το Μέτρο του αξιοσημείωτου (markedness) είναι χαμηλό προς μέτριο για (iterative imputation, τυχαία υπερδειγματοληψία) και ακολουθούν οι υπόλοιπες τυχαίες υπερδειγματοληψίες. Ακολουθούν με λίγο πάνω από χαμηλές τιμές οι περιπτώσεις όπου έχει γίνει εμπειρικη συμπλήρωση η (εμπειρική συμπλήρωση, υποδειγματοληψία) και η (iterative imputation, χωρίς εξισορρόπηση) - αλλά με πολύ μεγάλη τυπική απόκλιση. Ο Συντελεστής συσχέτισης Matthews (MCC) είναι επίσης χαμηλός προς μέτριος για τυχαία υπερδειγματοληψία. Χαμηλό (καλώς) ψευδώς αρνητικό ποσοστό (Miss rate) έχουμε μόνο για έλλειψη εξισορρόπησης και ακολουθεί η (iterative imputation, τυχαία υπερδειγματοληψία). Καλή αρνητική προγνωστική αξία (negative predictive value) έχουμε για (iterative imputation, τυχαία υπερδειγματοληψία). Μέτρια προς καλή είναι για τις υπόλοιπες τυχαίες υπερδειγματοληψίες, υποδειγματοληψίες, την (iterative imputation, έλλειψη εξισορρόπησης), (εμπειρική συμπλήρωση, έλλειψη εξισορρόπησης) - αλλά οι τελευταίες έχουν μεγάλη τυπική απόκλιση. Η θετική προγνωστική αξία (precision) είναι καλή για (εμπειρική συμπλήρωση, SMOTE-υπερδειγματοληψία) και λίγο άνω του μετρίου για (iterative imputation, υποδειγματοληψία), (δοκιμαζόμενη συνάρτηση, υποδειγματοληψία) και (δοκιμαζόμενη συνάρτηση, τυχαία υπερδειγματοληψία). Οι υπόλοιπες περιπτώσεις έχουν μέτριες προς καλές τιμές. Το κατώφλι επικράτησης (prevalence threshold) είναι χαμηλό προς μέτριο γενικά. Η ευαισθησία (sensitivity) είναι -όπως και στο KNN3- πολύ καλή για έλλειψη εξισορρόπησης και καλή για (Iterative imputation,

τυχαία υπερδειγματοληψία). Για τις υπόλοιπες περιπτώσεις οι τιμές της είναι άνω του μετρίου. Το μοτίβο της ειδικότητας (specificity) μοιάζει με της KNN3, αλλά εδώ είναι πρώτη η (δοκιμαζόμενη συνάρτηση, τυχαία υπερδειγματοληψία). Η βαθμολογία απειλής (threat score) είναι αντίστοιχη με της KNN3.

Καλύτερα τα πηγαίνει η τυχαία υπερδειγματοληψία και ακολουθεί η (εμπειρική συμπλήρωση, υποδειγματοληψία).

KNN5

Για την ακρίβεια ή ορθότητα (accuracy) ισχύει ό,τι και για την KNN2, αλλά ο συνδυασμός (δοκιμαζόμενη συνάρτηση, τυχαία υπερδειγματοληψία) δεν έχει τόσο καλή τιμή, σε αντίθεση με τον (εμπειρική συμπλήρωση, υποδειγματοληψία), που τα πάει αντίστοιχα καλά με τους πρώτους. Η AUC είναι όπως και των προηγουμένων, αλλά έχει λίγο ψηλότερες τιμές για όλες τις προεπεξεργασίες πλην της τυχαίας υπερδειγματοληψίας και του (iterative imputation, υποδειγματοληψία). Για την εξισορροπημένη ακρίβεια (balanced accuracy) ξεχωρίζουν οι περιπτώσεις εμπειρικής συμπλήρωσης και iterative imputation συνδυασμένες με τυχαία υπερδειγματοληψία, η (εμπειρική συμπλήρωση, υποδειγματοληψία), η (εμπειρική συμπλήρωση, SMOTE-υπερδειγματοληψία). Καλύτερο Σκορ-f1 (f1-score) έχουμε για (iterative imputation, τυχαία υπερδειγματοληψία), (εμπειρική συμπλήρωση, τυχαία υπερδειγματοληψία). Ο Δείκτης Fowlkes-Mallows (FM) έχει μικρές διαφορές με τον αντίστοιχο του KNN3. Ως προς το ψευδώς θετικό ποσοστό (fall-out) ισχύει ό,τι και για τα προηγούμενα KNN, αλλά τα πιο χαμηλά είναι για τις υπερδειγματοληψίες. Χαμηλό ποσοστό ψευδούς παράλειψης (false omission rate) έχουμε καλώς για την τυχαία υπερδειγματοληψία (για iterative imputation το χαμηλότερο) και ακολουθούν οι υποδειγματοληψίες και η έλλειψη εξισορρόπησης. Η Πληροφοριακότητα (Informedness), ο Συντελεστής συσχέτισης Matthews (MCC) και η τιμή κ του Cohen (Cohen's kappa) είναι χαμηλές προς μέτριες για (iterative imputation, τυχαία υπερδειγματοληψία), (εμπειρική συμπλήρωση, τυχαία υπερδειγματοληψία) και (εμπειρική συμπλήρωση, υποδειγματοληψία). Το Μέτρο του αξιοσημείωτου (markedness) είναι επίσης χαμηλό προς μέτριο για (iterative imputation, τυχαία υπερδειγματοληψία), (εμπειρική συμπλήρωση, τυχαία υπερδειγματοληψία), (εμπειρική συμπλήρωση, υποδειγματοληψία) και για έλλειψη εξισορρόπησης όπου τα κενά έχουν συμπληρωθεί με την δοκιμαζόμενη συνάρτηση ή με iterative imputation -αλλά οι τυπικές αποκλίσεις των τελευταίων είναι πολύ μεγάλες για να τα λάβουμε σοβαρά υπ'όψη. Χαμηλό (καλώς) ψευδώς αρνητικό ποσοστό (Miss rate) έχουμε μόνο για έλλειψη εξισορρόπησης και ακολουθεί η (iterative imputation, τυχαία υπερδειγματοληψία). Καλή αρνητική προγνωστική αξία (negative predictive value) έχουμε για (iterative imputation, τυχαία υπερδειγματοληψία), (εμπειρική συμπλήρωση, τυχαία υπερδειγματοληψία). Μέτρια προς καλή είναι η αρνητική προγνωστική αξία για τις υπόλοιπες τυχαίες υπερδειγματοληψίες, τις υποδειγματοληψίες και για έλλειψη εξισορρόπησης -αλλά η τελευταία έχει μεγάλη τυπική απόκλιση. Η θετική προγνωστική αξία (precision) είναι καλή για (εμπειρική συμπλήρωση,SMOTE-υπερδειγματοληψία) και λίγο άνω του μετρίου για υποδειγματοληψία και για (συνάρτηση, τυχαία υπερδειγματοληψία). Οι υπόλοιπες περιπτώσεις έχουν μέτριες προς καλές τιμές. Το κατώφλι επικράτησης (prevalence threshold) είναι χαμηλό προς μέτριο γενικά, με την (εμπειρική συμπλήρωση, SMOTE-υπερδειγματοληψία) να έχει την χαμηλότερη τιμή, αλλά μεγάλη τυπική απόκλιση. Καλή προς πολύ καλή ευαισθησία (sensitivity) έχουμε για έλλειψη εξισορρόπησης και καλή για (Iterative imputation, τυχαία υπερδειγματοληψία). Γενικά καλύτερα τα πηγαίνει το iterative imputation. Για τις υπόλοιπες περιπτώσεις οι τιμές είναι άνω του μετρίου προς καλές. Το μοτίβο της ειδικότητα (specificity) μοιάζει με της KNN2. Η βαθμολογία απειλής (threat score) είναι αντίστοιχη με της KNN2.

Καλύτερα τα πηγαίνει κατά σειρά η (Iterative imputation, τυχαία υπερδειγματοληψία) και η εμπειρική συμπλήρωση μαζί με τυχαία υπερδειγματοληψία ή υποδειγματοληψία ή SMOTE-υπερδειγματοληψία.

KNN6

Η ακρίβεια ή ορθότητα (accuracy) παρουσιάζει το πιο πάνω μοτίβο, αισθητά εξασθενημένο, μιας και οι διαφορές είναι πολύ μικρές. Η AUC έχει μέτρια προς καλή τιμή για όλες τις προεπεξεργασίες. Η εξισορροπημένη ακρίβεια (balanced accuracy) είναι σαν της KNN5, με τις τιμές που ξεχωρίζουν να είναι ελαφρώς ανεβασμένες, καθώς και τις τιμές για έλλειψη εξισορρόπησης. Για το Σκορ-f1 (f1-score) όλες οι τιμές σε περιπτώσεις όπου έχει γίνει εξισορρόπηση είναι μέτριες προς καλές με εξαίρεση την (δοκιμαζόμενη συνάρτηση, υποδειγματοληψία). Ο Δείκτης Fowlkes-Mallows (FM) έχει καλύτερες τιμές για έλλειψη εξισορρόπησης, μέτριες προς καλές για κάθε περίπτωση εξισορρόπησης και μέτριες για (συνάρτηση, υποδειγματοληψία). Ως προς το ψευδώς θετικό ποσοστό (fall-out), καλύτερη τιμή έχουμε για (εμπειρική συμπλήρωση, SMOTE-υπερδειγματοληψία), (iterative imputation, τυχαία υπερδειγματοληψία), (συνάρτηση, υποδειγματοληψία). Χαμηλό ποσοστό ψευδούς παράλειψης (false omission rate) έχουμε για (iterative imputation, υποδειγματοληψία), (εμπειρική συμπλήρωση, τυχαία υπερδειγματοληψία), (iterative imputation, τυχαία υπερδειγματοληψία). Η Πληροφοριακότητα (Informedness) και η τιμή κ του Cohen (Cohen's kappa) είναι χαμηλές για (iterative imputation, τυχαία υπερδειγματοληψία), (εμπειρική συμπλήρωση, τυχαία υπερδειγματοληψία) και (εμπειρική συμπλήρωση, SMOTE-υπερδειγματοληψία, ακολουθεί η

(εμπειρική συμπλήρωση, υποδειγματοληψία) -με μεγαλη τυπική απόκλιση. Κάτι παρόμοιο έχουμε και για το Μέτρο του αξιοσημείωτου (markedness) και το Συντελεστή συσχέτισης Matthews (MCC). Χαμηλό (καλώς) ψευδώς αρνητικό ποσοστό (Miss rate) έχουμε μόνο για έλλειψη εξισορρόπησης και ακολουθεί η (iterative imputation, υποδειγματοληψία) και η (iterative imputation, τυχαία υπερδειγματοληψία). Μέτρια προς καλή αρνητική προγνωστική αξία (negative predictive value) έχουμε για (iterative imputation, τυχαία υπερδειγματοληψία), (εμπειρική συμπλήρωση, τυχαία υπερδειγματοληψία), (iterative imputation, υποδειγματοληψία). Ακολουθούν οι υπόλοιπες τυχαίες υπερδειγματοληψίες και οι υποδειγματοληψίες. Η θετική προγνωστική αξία (precision) είναι καλή για (εμπειρική συμπλήρωση,SMOTE-υπερδειγματοληψία) και ακολουθούν οι υπόλοιπες SMOTE-υπερδειγματοληψίες. Το κατώφλι επικράτησης (prevalence threshold) είναι χαμηλό προς μέτριο γενικά, με την (εμπειρική συμπλήρωση, SMOTE-υπερδειγματοληψία) να έχει την χαμηλότερη τιμή, αλλά και μεγάλη τυπική απόκλιση. Καλή προς πολύ καλή ευαισθησία (sensitivity) έχουμε για έλλειψη εξισορρόπησης και καλή για (Iterative imputation, υποδειγαμτοληψία), (συνάρτηση, τυχαία υπερδειγματοληψία). Για τις υπόλοιπες περιπτώσεις οι τιμές είναι άνω του μετρίου προς καλές. Η ειδικότητα (specificity) είναι μέτρια προς καλή για (iterative imputation, τυχαία υπερδειγματοληψία), (εμπειρική συμπλήρωση, τυχαία υπερδειγματοληψία), (εμπειρική συμπλήρωση, SMOTE-υπερδειγματοληψία), (συνάρτηση, υποδειγματοληψία). Η ειδικότητα είναι κακή για έλλειψη εξισορρόπησης και είναι χαμηλή για (iterative imputation, υποδειγματοληψία) και (συνάρτηση, τυχαία υπερδειγματοληψία). Η βαθμολογία απειλής (threat score) είναι περί του μετρίου με εξαίρεση τις περιπτώσεις χωρίς εξισορρόπηση. Χαμηλότερη βαθμολογία απειλής (threat score) έχουμε για (δοκιμαζόμενη συνάρτηση, υποδειγματοληψία).

Καλύτερα τα πηγαίνει η (Iterative imputation, τυχαία υπερδειγματοληψία), εμπειρική συμπλήρωση μαζί με SMOTE-υπερδειγματοληψία ή τυχαία υπερδειγματοληψία και η (Iterative imputation, υποδειγματοληψία).

BernNB1-2

Το BernNB2 δείχνει να τα πηγαίνει ελαφρώς καλύτερα από το BernNB1 για κάθε μετρική. Μεταξύ των δύο σίγουρα θα επιλέγαμε το πρώτο. Η ακρίβεια ή ορθότητα (accuracy) είναι μέτρια προς καλή σε όλες τις περιπτώσεις, με το BernNB2 να έχει εμφανώς καλύτερες τιμές. Η AUC είναι καλή για το BernNB2 (περί το 0,75) και μέτρια προς καλή για το BernNB1. Ο συνδυασμός (δοκιμαζόμενη συνάρτηση, υποδειγματοληψία) τα πηγαίνει καλύτερα από τις υπόλοιπες περιπτώσεις υποδειγματοληψίας. Το BernNB2 έχει μέτρια προς καλή εξισορροπημένη ακρίβεια (balanced accuracy) για όλες τις κατηγορίες, ενώ το BernNB1, μέτρια. Το Σκορ-f1 (f1-score) παρουσιάζει αντίστοιχη διαφορά για BernNB1, BernNB2. Το Σκορ-f1 είναι λίγο καλύτερο για τις περιπτώσεις εξισορρόπησης με class_weights, (iterative imputation, τυχαία υπερδειγματοληψία), τυχαία υπερδειγματοληψία γενικά, (iterative imputation, υποδειγματοληψία), υποδειγματοληψία γενικά. Χαμηλότερο (καλώς) ψευδώς θετικό ποσοστό (fall-out) έχει το BernNB2, συγκεκριμένα για τις περιπτώσεις εξισορρόπησης τυχαία υπερδειγματοληψία, υποδειγματοληψία και SMOTE-υπερδειγματοληψία. Ως προς το ποσοστό ψευδούς παράλειψης (false omission rate), η έλλειψη εξισορρόπησης δεν φαίνεται να αποδίδει χειρότερα από ό,τι η ύπαρξη εξισορρόπησης. Και πάλι η διαφορά μεταξύ των δύο ταξινομητριών είναι εμφανής. Η εξισορρόπηση με υποδειγματοληψία ή με SMOTE-υπερδειγματοληψία αποδίδει καλά και στις δύο ταξινομήτριες (χαμηλή τιμή). Ο Δείκτης Fowlkes-Mallows (FM) είναι υψηλότερος για έλλειψη εξισορρόπησης (περί το 0,75). Ακολουθούν οι SMOTE-υπερδειγματοληψίες, οι τυχαίες υπερδειγματοληψίες και οι υποδειγματοληψίες. Η συμπλήρωση κενών με iterative imputation αποδίδει ελαφρώς καλύτερα συγκριτικά με τις άλλες τακτικές συμπλήρωσης κενών. Η Πληροφοριακότητα (Informedness) είναι χαμηλή σε γενικές γραμμές. Ξεχωρίζει η περίπτωση (iterative imputation, τυχαία υπερδειγματοληψία) και ακολουθεί με χαμηλότερες τιμές η (iterative imputation, SMOTE-υπερδειγματοληψία). Ωστόσο οι τυπικές αποκλίσεις τους είναι μεγάλες. Παρόμοια συμβαίνει και για την τιμή κ του Cohen (Cohen's kappa), το Μέτρο του αξιοσημείωτου (markedness), το Συντελεστή συσχέτισης Matthews (MCC). Η αρνητική προγνωστική αξία (negative predictive value) είναι γενικά μέτρια προς καλή. Είναι καλύτερη για έλλειψη εξισορρόπησης -γεγονός αναμενόμενο και ακολουθούν οι περιπτώσεις τυχαία υπερδειγματοληψία, υποδειγματοληψία, η SMOTE-υπερδειγματοληψία, με το iterative imputation να τα πηγαίνει σχετικά καλύτερα σε αρκετές περιπτώσεις. Χαμηλότερο ψευδώς αρνητικό ποσοστό (Miss rate) έχουμε για έλλειψη εξισορρόπησης, γεγονός που αναμένεται, αφού η Κατηγορία ασθενών Class ‘1’ είναι η πλειοψηφούσα και το να προγνωστεί ως μηδενική (Class ‘0’) μία δική της καταχώρηση είναι δυσκολότερο όταν το μοντέλο μεροληπτεί υπέρ της - όντας εκπαιδευμένο σε ένα σετ όπου αυτή πλειοψηφεί. Ωστόσο για τις περιπτώσεις που έχουμε εξισορρόπηση -μιας και αυτό είναι επιθυμητό- χαμηλότερο ψευδώς αρνητικό ποσοστό (Miss rate) έχουμε για (iterative imputation, τυχαία υπερδειγματοληψία) -που μπορεί να οφείλεται σε υπερπροσαρμογή στα δεδομένα, για (iterative imputation, SMOTE-υπερδειγματοληψία) και ακολουθούν κοντά οι υπόλοιπες περιπτώσεις υπερδειγματοληψίας και η (iterative imputation, υποδειγματοληψία). Καλύτερη θετική προγνωστική αξία (precision) έχουμε για SMOTE-υπερδειγματοληψία και έλλειψη εξισορρόπησης, ακολουθούν οι τυχαίες υπερδειγματοληψίες και οι υποδειγματοληψίες. Στο BernNB1 το iterative imputation τα πάει ελαφρώς χειρότερα από τις άλλες τακτικές συμπληρωσης κενών, όπως και στο BernNB2 τα πηγαίνει αντίστοιχα καλά. Χαμηλότερο κατώφλι επικράτησης (prevalence threshold) έχουμε στο BernNB2, για (iterative imputation, τυχαία υπερδειγματοληψία), (iterative imputation, υποδειγματοληψία) και ακολουθούν οι υπόλοιπες υποδειγματοληψίες, οι τυχαίες υπερδειγματοληψίες, οι SMOTE-

υπερδειγματοληψίες. Η ευαισθησία (sensitivity) είναι άνω του 0,75 για έλλειψη εξισορρόπησης (κάτι που είναι αναμενόμενο) και ακολουθούν οι συνδυασμοί iterative imputation με υπερδειγματοληψία, οι υπόλοιπες υπερδειγματοληψίες και οι υποδειγματοληψίες. Η ειδικότητα (specificity) είναι μέτρια προς καλή για όλες τις περιπτώσεις εξισορρόπησης. Η καλύτερη βαθμολογία απειλής (threat score) εμφανίζεται για έλλειψη εξισορρόπησης, γεγονός αναμενόμενο, αφού σαν μετρική επιρρεάζεται από την ανισορροπία στα δεδομένα και ακολουθούν η (iterative imputation, SMOTE-υπερδειγματοληψία), οι υπόλοιπες SMOTE-υπερδειγματοληψίες, η (iterative imputation, τυχαία υπερδειγματοληψία), οι υπόλοιπες τυχαίες υπερδειγματοληψίες και οι υποδειγματοληψίες.

Στο BernNB1 καλύτερα τα πηγαίνει η εμπειρική συμπλήρωση με SMOTE-υπερδειγματοληψία, και γενικά η υποδειγματοληψία. Ακολουθούν κοντά όλες οι υπόλοιπες εξισορροπήσεις. Για έλλειψη εξισορρόπησης έχουμε μεροληψία υπέρ της πλειοψηφούσας κλάσης. Στο BernNB2 καλύτερα τα πηγαίνει το Iterative imputation με τυχαία υπερδειγματοληψία / SMOTE-υπερδειγματοληψία. Ακολουθεί κοντά η δοκιμαζόμενη συνάρτηση με τυχαία υπερδειγματοληψία / SMOTE-υπερδειγματοληψία και το Iterative imputation με υποδειγματοληψία. Ακολουθούν οι υπόλοιπες περιπτώσεις, εκτός της έλλειψης εξισορρόπησης, που μεροληπτεί εμφανώς.

GNB

Η ακρίβεια ή ορθότητα (accuracy) είναι καλύτερη για (iterative imputation, υποδειγματοληψία), (εμπειρική συμπλήρωση, τυχαία υπερδειγματοληψία), (iterative imputation, τυχαία υπερδειγματοληψία) και την άλλη τυχαία υπερδειγματοληψία, τις υπόλοιπες υποδειγματοληψίες. Η AUC είναι γενικά μέτρια προς καλή και είναι καλύτερη για iterative imputation. Ακολουθεί η εμπειρική συμπλήρωση και η δοκιμαζόμενη συνάρτηση. Η εξισορροπημένη ακρίβεια (balanced accuracy) είναι μέτρια προς καλή. Ξεχωρίζουν οι iterative imputation συνδυασμένη με υποδειγματοληψία ή χωρίς εξισορρόπηση ή με τυχαία υπερδειγματοληψία, καθώς και η (εμπειρική συμπλήρωση, τυχαία υπερδειγματοληψία) και η εμπειρική συμπλήρωση συνδυασμένη με έλλειψη εξισορρόπησης ή SMOTE-υπερδειγματοληψία. Κατά τον υπολογισμό του Σκορ-f1 (f1-score) έχει προκύψει απροσδιοριστία στις περισσοτερες περιπτώσεις, αλλά βλέπουμε ότι σχετικά καλό είναι για (iterative imputation, υποδειγματοληψία). Το χαμηλοτερο ψευδώς θετικό ποσοστό (fall-out) υπάρχει για χρήση της δοκιμαζόμενης συνάρτησης για συμπλήρωση κενών και για (εμπειρική συμπλήρωση, τυχαία υπερδειγματοληψία). Το συγκεκριμένο εύρημα αξιολογείται ως ιδιαίτερα ενδιαφέρον. Το ποσοστό ψευδούς παράλειψης (false omission rate) είναι μέτριο, αλλά χαμηλότερο για τυχαία υπερδειγματοληψία και υποδειγματοληψία. Κατά τον υπολογισμό του Δείκτη Fowlkes-Mallows (FM) έχει προκύψει απροσδιοριστία και έτσι δεν έχουμε τιμές για κάθε περίπτωση. Ξεχωρίζει ο συνδυασμός (iterative imputation, υποδειγματοληψία) που βρίσκεται κοντά στο 0,75. Ωστόσο η Πληροφοριακότητα (Informedness) και η τιμή κ του Cohen (Cohen's kappa) είναι σχετικά χαμηλή προς ελάχιστη κατά περιπτώσεις, αντίστοιχα και για το Συντελεστή συσχέτισης Matthews (MCC). Λίγο κάτω του μετρίου είναι το Μέτρο του αξιοσημείωτου (markedness) για (iterative imputation, τυχαία υπερδειγματοληψία), αλλά με μεγάλη τυπική απόκλιση. Ωστόσο κι εδώ είχαμε πολλές απροσδιοριστίες κατά τους υπολογισμούς. Χαμηλό ψευδώς αρνητικό ποσοστό (Miss rate) έχουμε καλώς για (iterative imputation, υποδειγματοληψία), ενώ τα υπόλοιπα είναι πολύ ψηλά. Η αρνητική προγνωστική αξία (negative predictive value) είναι μέτρια γενικά. Καλή προς πολύ καλή θετική προγνωστική αξία (precision) έχουμε για έλλειψη εξισορρόπησης με iterative imputation ή εμπειρική συμπλήρωση, για (εμπειρική συμπλήρωση, SMOTE-υπερδειγματοληψία), (iterative imputation, τυχαία υπερδειγματοληψία). Το αντίστροφο ακριβώς μοτίβο έχουμε -καλώς- για την μετρική κατώφλι επικράτησης (prevalence threshold). Υψηλή ευαισθησία (sensitivity) έχουμε μόνο για (iterative imputation, υποδειγματοληψία), ενώ για τις υπόλοιπες περιπτώσεις είναι κάτω του μετρίου προς ελάχιστη. Η χρήση του iterative imputation δείχνει να αποδίδει σχετικά καλύτερα, ενώ έπεται η εμπειρική συμπλήρωση. Αξιοσημείωτη είναι η πολύ καλή ειδικότητα (specificity) όλων σχεδόν των μεθόδων, με το προηγούμενο μοτίβο να εμφανίζεται ανεστραμένο. Η καλύτερη βαθμολογία απειλής (threat score) εμφανίζεται για (iterative imputation, υποδειγματοληψία), με μοτίβο αντίστοιχο με της ευαισθησία (sensitivity), αλλά με μικρότερες τιμές.

Καλύτερα τα πηγαίνει το Iterative imputation μαζί με υποδειγματοληψία (μεγάλη τυπική απόκλιση) / τυχαία υπερδειγματοληψία / χωρίς εξισορρόπηση. Επίσης καλά τα πηγαίνει η εμπειρική συμπλήρωση μαζί με τυχαία υπερδειγματοληψία / SMOTE-υπερδειγματοληψία / έλλειψη εξισορρόπησης.

Εμπειρική συμπλήρωση, χωρίς εξισορρόπηση

Η μετρική AUC είναι άνω του καλού για RF1-2, BNB2, καλή για LR5-6. Η μετρική τιμή κ του Cohen (Cohen's kappa) είναι άνω του χαμηλού για BNB2, LR5, R6, DT1-3, RF2. Η μετρική Μέτρο του αξιοσημείωτου (markedness) είναι μέτρια προς χαμηλή για BNB2, LR5 και άνω του χαμηλού για LR6, DT1-3, KNN4, GNB, KNN5, BNB1. Η μετρική Πληροφοριακότητα (Informedness) είναι άνω του χαμηλού για LR5, BNB2, LR6, DT1-3, RF2. Η μετρική Δείκτης Fowlkes-Mallows (FM) είναι καλή για SVM, LR5-6, DT1-3, RF1-2, KNN2-6, BNB2. Η μετρική Συντελεστής συσχέτισης Matthews (MCC) είναι μέτρια προς χαμηλή για BNB2, LR5 και άνω του χαμηλού για LR6, DT1-3. Η μετρική Σκορ-f1 (f1-score) είναι καλή για SVM, LR5-6, DT1-3, RF1-2, KNN2-6 και BNB2. Η μετρική εξισορροπημένη ακρίβεια (balanced accuracy) είναι μέτρια προς καλή για LR5-6, BNB2, DT1-3, RF2, NN5, KNN4, BNB1 και GNB. Η μετρική βαθμολογία απειλής (threat score) είναι μέτρια προς καλή για RF2, SVM, BNB2, KNN2, KNN6. Η μετρική κατώφλι επικράτησης (prevalence threshold) είναι χαμηλή για GNB. Η μετρική ποσοστό ψευδούς παράλειψης (false omission rate) είναι άνω του χαμηλού για BNB2. Η μετρική ψευδώς θετικό ποσοστό (Fall out) είναι ελάχιστη για GNB και άνω του χαμηλού για LR1-4. Η μετρική ψευδώς αρνητικό ποσοστό (Miss rate) είναι χαμηλή προς ελάχιστη για KNN2, KNN5, KNN4, BNB2, KNN3, χαμηλή για DT1-3 και ελάχιστη για SVM, RF1, KNN6, RF2. Η μετρική αρνητική προγνωστική αξία (negative predictive value) είναι μέτρια προς καλή για BNB2, LR5, DT1-2, KNN4-5. Η μετρική ακρίβεια (Precision) είναι άνω του καλού για GNB και καλή για LR5-6, DT1-3, BNB2, NN5. Η μετρική ειδικότητα (specificity) είναι άριστη για GNB, μέτρια προς καλή για LR1-4, LR5-6, NN5. Η μετρική ευαισθησία (sensitivity) είναι καλή προς πολύ καλή για RF2, KNN1, KNN5, KNN4, BNB2, KNN3 και άριστη για SVM, RF1, NN6. Η μετρική ακρίβεια ή ορθότητα (accuracy) είναι σχεδόν καλή για BNB2, RF2, LR5, DT1-3, KNN2, KNN4, LR6, KNN3, KNN5 και BNB1.

Καλύτερα πηγαίνουν οι LR5, BNB2 και ακολουθούν οι LR6, DT1-3.

Δοκιμαζόμενη συνάρτηση, χωρίς εξισορρόπηση

Η μετρική AUC είναι άνω του καλού για RF1-2 και καλή για BNB2. Η τιμή κ του Cohen (Cohen's kappa) είναι άνω του χαμηλού για BNB2 και χαμηλή για DT1-3. Η μετρική Μέτρο του αξιοσημείωτου (markedness) είναι κάτω του μετρίου για BNB2 και RF2 (μεγάλη τυπική απόκλιση). Η μετρική Πληροφοριακότητα (Informedness) είναι άνω του χαμηλού για BNB2, DT1-3 και NN1. Η μετρική Δείκτης Fowlkes-Mallows (FM) είναι καλή για RF1-2, SVM, BNB2, KNN2-6, DT1-3, NN3, NN6 και NN10. Η μετρική Συντελεστής συσχέτισης Matthews (MCC) είναι μέτρια προς χαμηλή για BNB2, DT1-3 και RF2. Η μετρική Σκορ-f1 (f1-score) είναι καλή για RF1-2, SVM, BNB2, KNN2-6, DT1-3, NN3, NN6 και NN10. Η μετρική εξισορροπημένη ακρίβεια (balanced accuracy) είναι μέτρια προς καλή για BNB2, DT1-3, LR1-6, RF2, NN1, NN3, NN6, NN10 και BNB1. Η μετρική βαθμολογία απειλής (threat score) είναι μέτρια προς καλή για RF2, BNB2, RF1, DT1-3, NN3 και KNN2-6. Η μετρική κατώφλι επικράτησης (prevalence threshold) είναι κάτω του μετρίου για LR1-4 και DT2-4. Η μετρική ποσοστό ψευδούς παράλειψης (false omission rate) είναι χαμηλή για RF2 και άνω του χαμηλού για KNN5, BNB2. Η μετρική ψευδώς θετικό ποσοστό (Fall out) είναι ελάχιστη για GNB και κάτω του μετρίου για LR1-3, DT4. Η μετρική ψευδώς αρνητικό ποσοστό (Miss rate) είναι χαμηλή για DT1-3, NN1, NN3, NN6, NN10, BNB1, ελάχιστη για SVM, RF1, RF2, KNN6, KNN5, KNN2 και χαμηλή προς ελάχιστη για KNN3-4, BNB2. Η μετρική αρνητική προγνωστική αξία (negative predictive value) είναι μέτρια προς καλή για KNN5, BNB2, DT1-2 και καλή για RF2 (μεγάλη τυπική απόκλιση). Η μετρική ακρίβεια (Precision) είναι σχεδόν καλή για DT1-3, BNB2, LR1-6, NN1, NN3, NN10 και BNB1. Η μετρική ειδικότητα (specificity) είναι άριστη για GNB, άνω του μετρίου για LR1-6, DT1-4 και μέτρια για NN1, NN10, BNB1, BNB2. Η μετρική ευαισθησία (sensitivity) είναι καλή για DT1-3, NN3, NN6, NN10, NN1, BNB1, άριστη για SVM, RF1 και πολύ καλή προς άριστη για RF2, KNN5-6, KNN2, KNN1. Η μετρική ακρίβεια ή ορθότητα (accuracy) είναι μέτρια προς καλή για BNB2, RF2, DT1-3, RF1, NN1, NN3, NN6, NN10, KNN1-6, BNB1, LR1-6 και SVM.

Καλύτερα τα πηγαίνουν τα BNB2, DT1-3, RF2, NN1, NN3 και NN10.

Iterative imputation, χωρίς εξισορρόπηση

Η μετρική AUC είναι άνω του καλού για RF1-2 και καλή για LR5-6, BNB2, DT4. Η τιμή κ του Cohen (Cohen's kappa) είναι χαμηλή προς μέτρια για LR5-6 και DT1-4. Η μετρική Μέτρο του αξιοσημείωτου (markedness) είναι σχεδόν μέτρια για LR5-6 και χαμηλή προς μέτρια για DT1-4, NN7-10, KNN5, BNB2, GNB. Η μετρική Πληροφοριακότητα (Informedness) είναι χαμηλή προς μέτρια για LR5-6 και για DT1-4. Η μετρική Δείκτης Fowlkes-Mallows (FM) είναι καλή για όλες τις μεθόδους ταξινόμησης εκτός του GNB. Η μετρική Συντελεστής συσχέτισης Matthews (MCC) είναι χαμηλή προς μέτρια για LR5-6, DT1-4 και BNB2. Η μετρική Σκορ-f1 (f1-score) είναι καλή για όλες τις μεθόδους ταξινόμησης εκτός του GNB. Η μετρική εξισορροπημένη ακρίβεια (balanced accuracy) είναι σχεδόν καλή για LR5-6 και DT1-4. Η μετρική βαθμολογία απειλής (threat score) είναι μέτρια προς καλή για SVM, LR5-6, RF1-2, NN6-8, KNN2, KNN4-6 και BNB2. Η μετρική κατώφλι επικράτησης (prevalence threshold) είναι μέτρια προς χαμηλή για DT4, LR5-6, DT1-3 και χαμηλή για GNB. Η μετρική ποσοστό ψευδούς παράλειψης (false omission rate) είναι μέτρια προς χαμηλή για NN9, KNN5, BNB2, NN7, KNN2 και

άνω του χαμηλού για LR5-6. Η μετρική ψευδώς θετικό ποσοστό (Fall out) είναι χαμηλή για GNB και μέτρια προς χαμηλή για DT1-3. Η μετρική ψευδώς αρνητικό ποσοστό (Miss rate) είναι χαμηλή προς ελάχιστη για RF2, LR1-4, LR6, NN1-2, NN6-8, KNN2-5, BNB1-2, LR5, ελάχιστη για SVM, RF1, KNN6, άνω του χαμηλού για DT1-3 και χαμηλή για NN3, NN9-10, KNN1. Η μετρική αρνητική προγνωστική αξία (negative predictive value) είναι μέτρια προς καλή για LR5-6, NN9, BNB2, KNN5, KNN2, NN7-8 και DT1-3. Η μετρική θετική προγνωστική αξία (precision) είναι άνω του καλού για GNB και καλή για LR5-6, DT4, DT1-3, BNB2. Η μετρική ειδικότητα (specificity) είναι μέτρια προς καλή για DT4, DT1-3, καλή για GNB. Η μετρική ευαισθησία (sensitivity) είναι άριστη για SVM, RF1, KNN6 και καλή προς πολύ καλή για RF2, KNN2, LR1-4, LR5, NN1, NN2, NN6-8, KNN3-5, BNB1-2. Η μετρική ακρίβεια ή ορθότητα (accuracy) είναι καλή για LR5-6.

Καλύτερα τα πηγαίνουν οι LR5-6, τα DT1-4 και ακολουθούν τα NN7-10, KNN5, BNB2, GNB.

Εμπειρική Συμπλήρωση, υποδειγματοληψία

Η μετρική AUC είναι άνω του καλού για RF1-2 και καλή για LR5-6, BNB2. Η τιμή κ του Cohen (Cohen's kappa) είναι κάτω του μετρίου για RF1. Η μετρική Μέτρο του αξιοσημείωτου (markedness) είναι μέτρια για RF1 και κάτω του μετρίου για LR6, RF2, KNN1. Η μετρική Πληροφοριακότητα (Informedness) είναι μέτρια για RF1. Η μετρική Δείκτης Fowlkes-Mallows (FM) είναι καλή για RF1 και μέτρια προς καλή για LR6, LR5, RF2, KNN5, KNN4, DT3, BNB2. Η μετρική Συντελεστής συσχέτισης Matthews (MCC) είναι μέτρια για RF1. Η μετρική Σκορ-f1 (f1-score) είναι καλή για RF1, μέτρια προς καλή για LR6, RF2, SVM, KNN6, DT3, KNN5 και BNB2. Η μετρική εξισορροπημένη ακρίβεια (balanced accuracy) είναι καλή για RF1, μέτρια προς καλή για LR6, RF2, LR5, KNN1, KNN5, BNB2, DT1-3, KNN2-4 και KNN5-6. Η μετρική βαθμολογία απειλής (threat score) είναι μέτρια προς καλή για RF1, μέτρια για LR6, RF2, LR5, KNN6, BNB2 και KNN5. Η μετρική κατώφλι επικράτησης (prevalence threshold) είναι άνω του χαμηλού για LR6, KNN1 και μέτρια προς χαμηλή για SVM, RF1, KNN3, BNB2. Η μετρική ποσοστό ψευδούς παράλειψης (false omission rate) είναι χαμηλή για RF1 και μέτρια προς χαμηλή για RF2, LR6, LR5, DT1-3, KNN1-5, BNB2. Η μετρική ψευδώς θετικό ποσοστό (Fall out) είναι χαμηλή προς ελάχιστη για LR1-4, GNB και χαμηλή για KNN1, LR6, LR5, DT4, RF1-2. Η μετρική ψευδώς αρνητικό ποσοστό (Miss rate) είναι χαμηλή για RF1 και άνω του χαμηλού για KNN6, RF2. Η μετρική αρνητική προγνωστική αξία (negative predictive value) είναι καλή για RF1 και μέτρια προς καλή για LR6, RF2, LR5, BNB2, KNN1, KNN3-5. Η μετρική θετική προγνωστική αξία (precision) είναι καλή για KNN1, LR6, RF1 και LR5. Η μετρική ειδικότητα (specificity) είναι καλή προς πολύ καλή για LR1-4, GNB, KNN1 και καλή για LR5, LR6, DT4, RF1-2, KNN3, KNN5. Η μετρική ευαισθησία (sensitivity) είναι καλή για RF1 και μέτρια προς καλή για KNN6, RF2, LR6, SVM, LR5, DT1-3, KNN2, KNN5-6, BNB2. Η μετρική ακρίβεια ή ορθότητα (accuracy) είναι καλή για RF1 και μέτρια προς καλή για LR6, LR5, RF2, KNN1, KNN5, BNB2, KNN3, DT1-3.

Καλύτερα τα πηγαίνουν οι RF1 και ακολουθούν οι LR6, RF2, KNN1.

Δοκιμαζόμενη συνάρτηση, Υποδειγματοληψία

Η μετρική AUC είναι άνω του καλού για RF1, RF2 και καλή για BNB2. Η τιμή κ του Cohen (Cohen's kappa) είναι άνω του μετρίου για RF1 και RF2. Η μετρική Markedness είναι άνω του μετρίου για RF1 και RF2. Η μετρική Πληροφοριακότητα (Informedness) είναι άνω του μετρίου για RF1, RF2. Η μετρική Δείκτης Fowlkes-Mallows (FM) είναι καλή για RF1 και RF2. Η μετρική Συντελεστής συσχέτισης Matthews (MCC) είναι άνω του μετρίου για RF1 και RF2. Η μετρική Σκορ-f1 (f1-score) είναι καλή για RF1, RF2. Η μετρική εξισορροπημένη ακρίβεια (balanced accuracy) είναι καλή για RF1, RF2. Η μετρική βαθμολογία απειλής (threat score) είναι μέτρια προς καλή για RF1, RF2. Η μετρική κατώφλι επικράτησης (prevalence threshold) είναι μέτρια προς χαμηλή για RF1 και χαμηλή για RF2. Η μετρική ποσοστό ψευδούς παράλειψης (false omission rate) είναι χαμηλή για RF1-2. Η μετρική ψευδώς θετικό ποσοστό (Fall out) είναι ελάχιστη για GNB, χαμηλή για RF2, RF1. Η μετρική ψευδώς αρνητικό ποσοστό (Miss rate) είναι χαμηλή για RF1-2. Η μετρική αρνητική προγνωστική αξία (negative predictive value) είναι καλή για RF1-2. Η μετρική θετική προγνωστική αξία (precision) είναι καλή για RF2, RF1. Η μετρική ειδικότητα (specificity) είναι άριστη για GNB και καλή για RF2, RF1, SVM. Η μετρική ευαισθησία (sensitivity) είναι καλή για RF1-2. Η μετρική ακρίβεια ή ορθότητα (accuracy) είναι καλή για RF2 και RF1.

Καλύτερα τα πηγαίνουν τα RF2, RF1.

Iterative imputation, Υποδειγματοληψία

Η μετρική AUC είναι άνω του καλού για RF1-2, LR5 και καλή για LR6, GNB. Η τιμή κ του Cohen (Cohen's kappa) είναι μέτρια για LR5-6 και κάτω του μετρίου για RF1-2. Η μετρική Μέτρο του αξιοσημείωτου (markedness) είναι μέτρια για LR5-6 και κάτω του μετρίου για RF1-2. Η μετρική Πληροφοριακότητα (Informedness) είναι μέτρια για LR5-6 και κάτω του μετρίου για RF1-2. Η μετρική Δείκτης Fowlkes-Mallows (FM) είναι καλή για LR6, RF1-2, GNB, LR5 και KNN6. Η μετρική Συντελεστής συσχέτισης Matthews (MCC) είναι μέτρια για LR5-6 και κάτω του μετρίου για RF1-2. Η μετρική Σκορ-f1 (f1-score) είναι καλή για LR6, RF1-2, LR5 και GNB. Η μετρική εξισορροπημένη ακρίβεια (balanced accuracy) είναι καλή για LR5-6 και RF1-2. Η μετρική βαθμολογία απειλής (threat score) είναι μέτρια προς καλή για LR6, LR5, RF1-2, GNB και μέτρια για KNN6, SVM, DT1-3. Η μετρική κατώφλι επικράτησης (prevalence threshold) είναι άνω του χαμηλού για LR5, DT4 και μέτρια προς χαμηλή για LR6, BNB2, RF2, RF1, DT1-3, NN7, NN9. Η μετρική ποσοστό ψευδούς παράλειψης (false omission rate) είναι χαμηλή για LR6, LR5 και RF1-2. Η μετρική ψευδώς θετικό ποσοστό (Fall out) είναι χαμηλή προς ελάχιστη για LR1-4 και χαμηλή για LR5, DT4, RF2. Η μετρική ψευδώς αρνητικό ποσοστό (Miss rate) είναι χαμηλή για LR6, RF1, χαμηλή προς ελάχιστη για SVM, KNN6, GNB και άνω του χαμηλού για NN6, NN5. Η μετρική αρνητική προγνωστική αξία (negative predictive value) είναι καλή για LR6, LR5, RF1-2 και KNN6. Η μετρική θετική προγνωστική αξία (precision) είναι καλή για LR5, DT4, RF2, LR6 και RF1. Η μετρική ειδικότητα (specificity) είναι καλή προς πολύ καλή για LR1-4 και καλή για LR5, DT4, RF2. Η μετρική ευαισθησία (sensitivity) είναι καλή για LR6, RF1, NN6 και άνω του καλού για SVM, KNN6, GNB. Η μετρική ακρίβεια ή ορθότητα (accuracy) είναι καλή για LR5-6 και RF1-2.

Συνολικά καλύτερα τα πηγαίνουν τα LR6, LR5, RF1-2.

Εμπειρική συμπλήρωση, class weights

Η μετρική AUC είναι άνω του καλού για RF1-2. Η τιμή κ του Cohen (Cohen's kappa) είναι άνω του χαμηλού για RF2, RF1. Η μετρική Μέτρο του αξιοσημείωτου (markedness) είναι μέτρια για RF2 και κάτω του μετρίου για RF1. Η μετρική Πληροφοριακότητα (Informedness) είναι χαμηλή για RF2, RF1, LR5 και DT2. Η μετρική Δείκτης Fowlkes-Mallows (FM) είναι καλή για όλες τις μεθόδους ταξινόμησης. Η μετρική Συντελεστής συσχέτισης Matthews (MCC) είναι άνω του χαμηλού για RF2, RF1 και χαμηλή για LR5, DT2. Η μετρική Σκορ-f1 (f1-score) είναι καλή για όλες τις μεθόδους ταξινόμησης. Η μετρική εξισορροπημένη ακρίβεια (balanced accuracy) είναι μέτρια προς καλή για RF2, RF1, LR5 και DT2. Η μετρική βαθμολογία απειλής (threat score) είναι μέτρια προς καλή για όλες τις μεθόδους ταξινόμησης. Η μετρική κατώφλι επικράτησης (prevalence threshold) είναι μέτρια για όλες τις μεθόδους ταξινόμησης. Η μετρική ποσοστό ψευδούς παράλειψης (false omission rate) είναι χαμηλή για RF2 και άνω του χαμηλού για RF1. Η μετρική ψευδώς θετικό ποσοστό (Fall out) είναι άνω του μετρίου για LR5, DT2. Η μετρική ψευδώς αρνητικό ποσοστό (Miss rate) είναι ελάχιστη για SVM, LR1-4 και χαμηλή προς ελάχιστη για RF2, RF1. Η μετρική αρνητική προγνωστική αξία (negative predictive value) είναι καλή για RF2 και RF1. Η μετρική θετική προγνωστική αξία (precision) είναι σχεδόν καλή για LR5, RF2, RF1 και DT1-3. Η μετρική ειδικότητα (specificity) είναι σχεδόν μέτρια για LR5 και DT2. Η μετρική ευαισθησία (sensitivity) είναι άριστη για SVM, LR1-4 και καλή προς πολύ καλή για RF2, RF1. Η μετρική ακρίβεια ή ορθότητα (accuracy) είναι μέτρια προς καλή για όλες τις μεθόδους ταξινόμησης.

Γενικά καλύτερα τα πηγαίνουν οι RF2, RF1 και ακολουθούν οι LR5, DT2.

Δοκιμαζόμενη συνάρτηση, class weights

Η μετρική AUC είναι άνω του καλού για RF1-2. Η τιμή κ του Cohen (Cohen's kappa) είναι κάτω του μετρίου για RF1-2. Η μετρική Μέτρο του αξιοσημείωτου (markedness) είναι μέτρια προς καλή για RF1-2. Η μετρική Πληροφοριακότητα (Informedness) είναι κάτω του μετρίου για RF1-2. Η μετρική Δείκτης Fowlkes-Mallows (FM) είναι καλή για όλες τις μεθόδους ταξινόμησης. Η μετρική Συντελεστής συσχέτισης Matthews (MCC) είναι μέτρια για RF1-2. Η μετρική Σκορ-f1 (f1-score) είναι καλή για όλες τις μεθόδους ταξινόμησης. Η μετρική εξισορροπημένη ακρίβεια (balanced accuracy) είναι μέτρια προς καλή για RF1-2, DT2, DT1 και DT3. Η μετρική βαθμολογία απειλής (threat score) είναι μέτρια προς καλή για RF2, RF1, SVM και DT2. Η μετρική κατώφλι επικράτησης (prevalence threshold) είναι μέτρια προς χαμηλή για RF1-2. Η μετρική ποσοστό ψευδούς παράλειψης (false omission rate) είναι χαμηλή προς ελάχιστη για RF2, RF1. Η μετρική ψευδώς θετικό ποστοστό (Fall out) είναι μέτρια για RF2 και RF1. Η μετρική ψευδώς αρνητικό ποσοστό (Miss rate) είναι ελάχιστη για SVM και χαμηλή προς ελάχιστη για RF2, RF1, LR5-6. Η μετρική αρνητική προγνωστική αξία (negative predictive value) είναι καλή προς πολύ καλή για RF2 και RF1. Η μετρική θετική προγνωστική αξία (precision) είναι καλή για RF2, RF1. Η μετρική ειδικότητα (specificity) είναι μέτρια για RF1-2. Η μετρική ευαισθησία (sensitivity) είναι άριστη για SVM και καλή προς πολύ καλή για RF2, RF1, LR5-6. Η μετρική ακρίβεια ή ορθότητα (accuracy) είναι καλή για RF1-2.

Καλύτερα τα πηγαίνουν οι RF1-2.

Iterative imputation, class weights

Η μετρική AUC είναι καλή για RF1-2 και LR5-6. Η τιμή κ του Cohen (Cohen's kappa) είναι χαμηλή προς μέτρια για LR5. Η μετρική Μέτρο του αξιοσημείωτου (markedness) είναι μέτρια για LR5-6. Η μετρική Πληροφοριακότητα (Informedness) είναι χαμηλή προς μέτρια για LR5 και άνω του χαμηλού για LR6, RF1, DT2-3. Η μετρική Δείκτης Fowlkes-Mallows (FM) είναι καλή για όλες τις μεθόδους ταξινόμησης. Η μετρική Συντελεστής συσχέτισης Matthews (MCC) είναι κάτω του μετρίου για LR5 και χαμηλή προς μέτρια για LR6, DT2-3. Η μετρική Σκορ-f1 (f1-score) είναι καλή για όλες τις μεθόδους ταξινόμησης. Η μετρική εξισορροπημένη ακρίβεια (balanced accuracy) είναι μέτρια προς καλή για LR5, LR6, DT1-3, RF1-2. Η μετρική βαθμολογία απειλής (threat score) είναι μέτρια προς καλή για όλες τις μεθόδους ταξινόμησης. Η μετρική κατώφλι επικράτησης (prevalence threshold) είναι μέτρια για όλες τις μεθόδους ταξινόμησης (DT2, LR5 λίγο χαμηλήτερα). Η μετρική ποσοστό ψευδούς παράλειψης (false omission rate) είναι χαμηλή για LR5-6 και χαμηλή προς μέτρια για DT1-3. Η μετρική ψευδώς θετικό ποσοστό (Fall out) είναι μέτρια για DT1-3, άνω του μετρίου για LR5-6 και μέτρια προς υψηλή για RF1-2. Η μετρική ψευδώς αρνητικό ποσοστό (Miss rate) είναι ελάχιστη για SVM, LR1-4, χαμηλή προς ελάχιστη για LR5, RF1-2, LR6, DT4 και χαμηλή για DT1-3. Η μετρική αρνητική προγνωστική αξία (negative predictive value) είναι μέτρια προς καλή για DT1-3 και καλή για LR5-6. Η μετρική θετική προγνωστική αξία (precision) είναι καλή για LR5, DT2, LR6, DT1, DT3, RF1, RF3. Η μετρική ειδικότητα (specificity) είναι μέτρια για DT1-3, LR5-6. Η μετρική ευαισθησία (sensitivity) είναι πολύ καλή για SVM, LR1-4, καλή προς πολύ καλή για LR5, RF1, LR6, DT4 και καλή για DT1-3. Η μετρική ακρίβεια ή ορθότητα (accuracy) είναι καλή για LR5, RF1 και μέτρια προς καλή για LR6, DT1-3, RF2.

Συνολικά, καλύτερα τα πηγαίνουν οι ταξινομήτριες LR5-6.

Εμπειρική συμπλήρωση, SMOTE-υπερδειγματοληψία

Η μετρική AUC είναι άνω του καλού για RF1-2 και καλή για BNB2 και LR5-6. Η τιμή κ του Cohen (Cohen's kappa) είναι κάτω του μετρίου για RF1-2. Η μετρική Μέτρο του αξιοσημείωτου (markedness) είναι μέτρια για RF1-2. Η μετρική Πληροφοριακότητα (Informedness) είναι μέτρια για RF1-2. Η μετρική Δείκτης Fowlkes-Mallows (FM) είναι καλή για LR6, RF1-2, BNB2. Η μετρική Συντελεστής συσχέτισης Matthews (MCC) είναι μέτρια για RF2 και RF1. Η μετρική Σκορ-f1 (f1-score) είναι καλή για RF2, RF1, LR6, BNB2 και LR5. Η μετρική εξισορροπημένη ακρίβεια (balanced accuracy) είναι καλή για RF1-2. Η μετρική βαθμολογία απειλής (threat score) είναι μέτρια προς καλή για RF1-2, LR6 και BNB2. Η μετρική κατώφλι επικράτησης (prevalence threshold) είναι χαμηλή για GNB και άνω του χαμηλού για RF2, KNN5, KNN1, RF1. Η μετρική ποσοστό ψευδούς παράλειψης (false omission rate) είναι άνω του χαμηλού για RF2, RF1. Η μετρική ψευδώς θετικό ποσοστό (Fall out) είναι ελάχιστη για GNB, χαμηλή προς ελάχιστη για LR1-4 και χαμηλή για DT4, KNN1, KNN5-6, RF1-2. Η μετρική ψευδώς αρνητικό ποσοστό (Miss rate) είναι χαμηλή για RF2, RF1, LR6 και άνω του χαμηλού για BNB2, LR5. Η μετρική αρνητική προγνωστική αξία (negative predictive value) είναι μέτρια προς καλή για RF2, RF1, BNB2 και LR6. Η μετρική θετική προγνωστική αξία (precision) είναι άνω του καλού για GNB, RF2, RF1, LR6, KNN1 και καλή για KNN5-6, BNB1-2, DT1-4. Η μετρική ειδικότητα (specificity) είναι άριστη για GNB, καλή προς πολύ καλή για LR1-4 και καλή για KNN1, DT4, KNN4-5, RF1-2. Η μετρική ευαισθησία (sensitivity) είναι καλή για RF2, RF1 και LR6. Η μετρική ακρίβεια ή ορθότητα (accuracy) είναι καλή για RF2, RF1, LR6 και BNB2.

Συνοπτικά, καλύτερα τα πηγαίνουν οι μέθοδοι ταξινόμησης RF1-2, ενώ οι υπόλοιπες μέθοδοι μεροληπτούν περισσότερο υπέρ της μίας ή της άλλης κλάσης.

Δοκιμαζόμενη συνάρτηση, SMOTE-υπερδειγματοληψία

Η μετρική AUC είναι άνω του καλού για RF1-2 και καλή για BNB2. Η μετρική Cohen's K είναι μέτρια για RF1 και κάτω του μετρίου για RF2. Η μετρική Μέτρο του αξιοσημείωτου (markedness) είναι μέτρια για RF1 και κάτω του μετρίου για RF2. Η μετρική Πληροφοριακότητα (Informedness) είναι μέτρια για RF1 και κάτω του μετρίου για RF2. Η μετρική Δείκτης Fowlkes-Mallows (FM) είναι καλή για RF1, RF2 και BNB2. Η μετρική Συντελεστής συσχέτισης Matthews (MCC) είναι μέτρια για RF1 και κάτω του μετρίου για RF2. Η μετρική Σκορ-f1 (f1-score) είναι καλή για RF1, RF2, BNB2. Η μετρική εξισορροπημένη ακρίβεια (balanced accuracy) είναι καλή για RF1-2. Η μετρική βαθμολογία απειλής (threat score) είναι μέτρια προς καλή για RF1, RF2, BNB2. Η μετρική κατώφλι επικράτησης (prevalence threshold) είναι μέτρια προς χαμηλή για RF1-2. Η μετρική ποσοστό ψευδούς παράλειψης (false omission rate) είναι χαμηλή για RF1-2. Η μετρική ψευδώς θετικό ποσοστό (Fall out) είναι ελάχιστη για GNB και άνω του χαμηλού για RF1-2. Η μετρική ψευδώς αρνητικό ποσοστό (Miss rate) είναι χαμηλή για RF1-2 και άνω του χαμηλού για BNB2. Η μετρική αρνητική προγνωστική αξία (negative predictive value) είναι μέτρια προς καλή για RF1, RF2, BNB2. Η μετρική θετική προγνωστική αξία (precision) είναι άνω του καλού για RF1-2 και καλή για BNB2. Η μετρική ειδικότητα (specificity) είναι άριστη για GNB και μέτρια προς καλή για RF1-2, DT4, BNB2, LR1-4, NN7. Η μετρική ευαισθησία (sensitivity) είναι καλή για RF1-2, BNB2. Η μετρική ακρίβεια ή ορθότητα (accuracy) είναι καλή για RF1-2.

Γενικά καλύτερα τα πηγαίνουν στις περισσότερες μετρικές τα RF1-2 και ακολουθεί, λίγο χειρότερα το BNB2.

Iterative imputation, SMOTE-υπερδειγματοληψία

Η AUC είναι άνω του καλού για RF1-2 και καλή για LR5-6, BNB2. Η τιμή κ του Cohen (Cohen's kappa) είναι κάτω του μετρίου για RF1-2, LR5-6, BNB2. Η μετρική Μέτρο του αξιοσημείωτου (markedness) είναι μέτρια για RF2, LR6 και κάτω του μετρίου για RF1, LR5. Η μετρική Πληροφοριακότητα (Informedness) είναι κάτω του μετρίου για RF2, RF1, LR6, BNB2, LR5. Η μετρική Δείκτης Fowlkes-Mallows (FM) είναι καλή για SVM, LR5-6, RF1-2, BNB2, NN4, NN9. Η μετρική Συντελεστής συσχέτισης Matthews (MCC) είναι μέτρια για RF2 και κάτω του μετρίου για RF1, BNB2, LR6, LR5. Η μετρική Σκορ-f1 (f1-score) είναι καλή για RF1-2, LR5-6, BNB2, SVM και NN4. Η μετρική εξισορροπημένη ακρίβεια (balanced accuracy) είναι καλή για RF2, RF1, LR6, LR5 και BNB2. Η μετρική βαθμολογία απειλής (threat score) είναι καλή προς πολύ καλή για LR4 και μέτρια προς καλή για RF1-2, LR5-6, BNB2, SVM. Η μετρική κατώφλι επικράτησης (prevalence threshold) είναι μέτρια προς χαμηλή για RF2, RF1 και LR6. Η μετρική ποσοστό ψευδούς παράλειψης (false omission rate) είναι άνω του χαμηλού για RF2, LR6, LR5, RF1, BNB2. Η μετρική ψευδώς θετικό ποσοστό (fall-out) είναι κάτω του χαμηλού για LR1-4, GNB και άνω του χαμηλού για DT4, RF2, RF1, NN5, BNB2. Η μετρική ψευδώς αρνητικό ποσοστό (Miss rate) είναι χαμηλή προς ελάχιστη για SVM και χαμηλή για LR5-6, RF1-2, NN4, GNB2. Η μετρική αρνητική προγνωστική αξία (negative predictive value) είναι μέτρια προς καλή για RF2, LR6, BNB2, LR5, RF1, NN9 και NN3-6. Η μετρική θετική προγνωστική αξία (precision) είναι άνω του καλού για RF2, RF1, LR6, LR5, BNB2, NN9 και καλή για DT4, NN5. Η μετρική ειδικότητα (specificity) είναι καλή προς πολύ καλή για LR1-4, GNB και ακολουθούν τα DT4, RF1-2, LR6 και NN5. Η μετρική ευαισθησία (sensitivity) είναι άνω του καλού για SVM και καλή για LR5, BNB2, LR6, RF1-2, NN4. Η μετρική ακρίβεια ή ορθότητα (accuracy) είναι καλή για RF1-2, BNB2, LR6, LR5.

Γενικά καλύτερα στις περισσότερες μετρικές τα πηγαίνουν οι RF2, RF1, LR6, LR5, BNB2.

Εμπειρική συμπλήρωση, τυχαία υπερδειγματοληψία

Η μετρική ακρίβεια ή ορθότητα (accuracy) είναι καλή για LR5-6, DT1-3, RF1-2, NN2-8, NN10. Η μετρική εξισορροπημένη ακρίβεια (balanced accuracy) είναι καλή για NN6, NN10, RF1-2, NN7-8, DT2, DT1, DT3, NN2-5. Η μετρική AUC είναι καλή προς πολύ καλή για RF1-2, NN6, άνω του καλού για NN10, NN3-9, LR5-6, DT1-3, καλή για NN1-2, BNB2. Η μετρική Σκορ-f1 (f1-score) είναι καλή για RF1, NN6, RF2, DT2, NN7, NN10. Η μετρική ψευδώς θετικό ποσοστό (fall-out) είναι ελάχιστή για GNB, χαμηλή προς ελάχιστη για NN10, KNN1, LR1-4, NN6, NN2-4 και κάτω του χαμηλού για DT1-4, RF1-2, NN1, NN5 και NN9. Η μετρική ποσοστό ψευδούς παράλειψης (false omission rate) είναι χαμηλή για RF1-2, DT2-3, NN6, DT1, NN7, NN10. Η μετρική Δείκτης Fowlkes-Mallows (FM) είναι καλή για NN6, RF1-2, NN10, NN7, DT1-3, NN3-5, NN8. Η μετρική Πληροφοριακότητα (Informedness) είναι άνω του μετρίου για NN6, NN10, RF1-2 και μέτρια για NN7, NN8, DT2, DT1, DT3, NN2-5. Η τιμή κ του Cohen (Cohen's kappa) είναι άνω του μετρίου για NN6, NN10, RF1-2, NN7, DT2, NN8 και μέτρια για DT1, DT3. Η μετρική Μέτρο του αξιοσημείωτου (markedness) είναι μέτρια προς καλή για NN10, NN6, άνω του μετρίου για NN8, RF1-2, NN7, NN4, DT2, DT1, DT3, NN3 και μέτρια για NN2, NN5, NN9. Η μετρική Συντελεστής συσχέτισης Matthews (MCC) είναι καλή για NN10, NN6, RF1-2, NN7-8, DT22, DT3, NN4, DT1, NN3 και NN5. Η μετρική βαθμολογία απειλής (threat score) είναι μέτρια προς καλή για NN6, RF1-2, NN7, NN10, NN8, DT2, DT1, DT3, LR5-6, NN3-4, NN2, BNB2, KNN5 και μέτρια για NN1, NN9. Η μετρική κατώφλι επικράτησης (prevalence-threshold) είναι χαμηλή για NN10, NN8, KNN2, NN4 και NN6. Η μετρική αρνητική προγνωστική αξία (negative predictive value) είναι καλή για DT1-3, RF1-2, NN5-7, NN10, LR5. Η μετρική θετική προγνωστική αξία (precision) είναι καλή προς πολύ καλή για NN10, NN8, NN6, NN4, NN3, RF2, DT2, RF1, DT1, NN2, DT3, KNN1, NN7, NN9 και καλή για NN5, NN1, DT4, LR5-6. Η μετρική ειδικότητα (specificity) είναι άριστη για GNB, καλή προς πολύ καλή για NN10, KNN1, LR1-4, NN8, NN6, NN4, NN2, NN7, NN9, NN3, RF2, DT2, DT1, DT3, RF1, NN1, NN5 και καλή για LR5-6, KNN2-4. Η μετρική ευαισθησία (sensitivity) είναι καλή προς πολύ καλή για SVM (μεγάλη τυπική απόκλιση) και καλή για RF1. Η μετρική ψευδώς αρνητικό ποσοστό (Miss rate) είναι χαμηλή προς ελάχιστη για SVM (μεγάλη τυπική απόκλιση) και χαμηλή για RF1-2, BNB2.

Γενικά καλύτερα τα πηγαίνουν οι μέθοδοι NN6-8, NN10, ενώ ακολουθούν κοντά τα υπόλοιπα NN, μαζί με τα DT1-3, RF1-2, LR5-6.

Δοκιμαζόμενη συνάρτηση, τυχαία υπερδειγματοληψία

Η AUC είναι καλή προς πολύ καλή για RF1-2 και καλή για DT1-3, NN6, NN3, NN7, NN9, BNB2. Η τιμή κ του Cohen (Cohen's kappa) είναι μέτρια για RF1-2, DT1-3 (μεγάλη τυπική απόκλιση). Το Μέτρο του αξιοσημείωτου (markedness) είναι άνω του μετρίου για DT1-3, RF1-2. Η Πληροφοριακότητα (Informedness) είναι μέτρια για DT1-3, RF1-2. Ο Δείκτης Fowlkes-Mallows (FM) είναι καλός για RF1-2, DT1-3. Ο Συντελεστής συσχέτισης Matthews (MCC) είναι μέτριος για DT2, RF2, RF1, DT1 και DT3. Το Σκορ-f1 (f1-score) είναι καλό για RF2, RF1, DT2, DT1 και DT3. Η εξισορροπημένη ακρίβεια (balanced accuracy) είναι καλή για RF2, RF1, DT2, DT1 και DT3. Η βαθμολογία απειλής (threat score) είναι μέτρια προς καλή για RF2, RF1, DT2, DT1, DT3 και μέτρια για NN6, NN3, BNB2. Το κατώφλι επικράτησης (prevalence threshold) είναι χαμηλό (καλώς) για DT2, DT1, RF2, DT4, NN9, NN7 και RF1. Το ποσοστό ψευδούς παράλειψης (false omission rate)

είναι χαμηλό για RF2, RF1, DT2, DT3 και DT1. Το ψευδώς θετικό ποσοστό (fall-out) είναι κάτω του χαμηλού για NN7, KNN4, DT1, DT2, DT3. Το ψευδώς αρνητικό ποσοστό (Miss rate) είναι χαμηλό προς πολύ χαμηλό για SVM (μεγάλη τυπική απόκλιση) και χαμηλό για RF2, KNN6, RF1. Η αρνητική προγνωστική αξία (negative predictive value) είναι καλή για RF2, RF1, DT3, DT2 και DT1. Η θετική προγνωστική αξία (precision) είναι άνω του καλού για DT2, DT1, NN7, RF2, RF1, DT3 και καλή για DT4. Η ειδικότητα (specificity) είναι άριστη για GNB, άνω του καλού για NN7, KNN4, DT1, DT2, DT3, DT4, RF1, NN4, NN3, NN5, NN6, RF2, NN1. Η ευαισθησία (Sensitivity) είναι καλή προς πολύ καλή για SVM (μεγάλη τυπική απόκλιση) και καλή για RF2, RF1, KNN6. Η ακρίβεια ή ορθότητα (accuracy) είναι καλή για DT2, RF2, RF1, DT1 και DT3.

Συμπερασματικά καλύτερα τα πηγαίνουν συνολικά οι RF1-2, DT1-3.

Iterative imputation, τυχαία υπερδειγματοληψία

Στην AUC ξεχωρίζουν τα RF με καλές προς πολύ καλές τιμές και με καλές τιμές τα DT1, DT, LR5-6, NN2, NN9-10, BNB1. Η τιμή κ του Cohen (Cohen's kappa) είναι μέτρια προς καλή για DT2-4, RF1-2 και μέτρια για DT1, BNB1. Το Μέτρο του αξιοσημείωτου (markedness) είναι μέτριο προς καλό για DT1-3, RF1-2, LR6 και καλό για BNB1. Η Πληροφοριακότητα (Informedness) είναι μέτρια προς καλή για DT1-3 και RF1-2. Ο Δείκτης Fowlkes-Mallows (FM) είναι καλός για LR5-6, DT1-3, RF1-2, KNN2, KNN4-5 και BNB2. Ο Συντελεστής συσχέτισης Matthews (MCC) είναι μέτριος προς καλός για DT1-3, RF1-2 και μέτριος για LR6, BNB2. Το Σκορ-f1 (f1-score) είναι καλό για RF1-2, DT1-3, LR5-6, KNN2, KNN4 και BNB2. Η εξισορροπημένη ακρίβεια (balanced accuracy) είναι καλή για DT1-3, RF1-2 και BNB2. Η βαθμολογία απειλής (threat score) είναι μέτρια προς καλή για SVM, LR5-6, DT1-3, RF1-2, KNN2, KNN4, KNN5, KNN3, NN6, NN3, NN7, NN9 και BNB2. Το κατώφλι επικράτησης (prevalence-threshold) είναι χαμηλό (καλώς) για GNB (μεγάλη τυπική απόκλιση) και DT1-3, ενώ ακολουθούν κοντά τα RF1-2. Το ποσοστό ψευδούς παράλειψης (false omission rate) είναι χαμηλό (καλώς) για LR6, RF2, SVM (μεγάλη τυπική απόκλιση), RF1, KNN2, NN7, LR5, DT1-3. Το ψευδώς θετικό ποσοστό (fall-out) είναι πολύ χαμηλό (καλώς) για LR4, LR1-3, GNB, DT1-3 και χαμηλό για RF2,1, KNN1, NN10. Το ψευδώς αρνητικό ποσοστό (Miss rate) είναι πολύ χαμηλό (καλώς) για SVM, KNN2, KNN5, LR6, NN7, LR5, BNB2, NN1-7 και KNN2-5. Η αρνητική προγνωστική αξία (negative predictive value) είναι καλή για SVM, LR5-6, RF1-2, NN7, KNN2 και ακολουθούν τα DT1-3, KNN4-5 και BNB2. Η θετική προγνωστική αξία (precision) είναι άνω του καλού για GNB, DT1-3, RF2-1 και καλή για NN9-10, LR6, DT4. Η ειδικότητα (specificity) είναι καλή προς πολύ καλή για LR4, LR1-3, GNB, DT1-3 και καλή για RF2-1, KNN1, DT4, NN10. Η ευαισθησία (sensitivity) είναι καλή προς πολύ καλή για SVM και καλή για KNN2, KNN5, NN1, NN7, NN3, RF1-2, NN2, NN4, NN6, KNN3. Η ακρίβεια ή ορθότητα (accuracy) είναι καλή για DT1-3, RF2-1, BNB2, LR6 και KNN2.

Συμπερασματικά καλύτερα τα πηγαίνουν οι LR5, LR6, DT1-3, RF1-2, BNB2 για τις περισσότερες μετρικές. Τα DT1-3 τείνουν να ταξινομούν ελαφρά καλύτερα την κλάση 0, ενώ το αντίστροφο συμβαίνει για τα άλλα.

6.3 Σχολιασμός Αποτελεσμάτων

Γενικές παρατηρήσεις

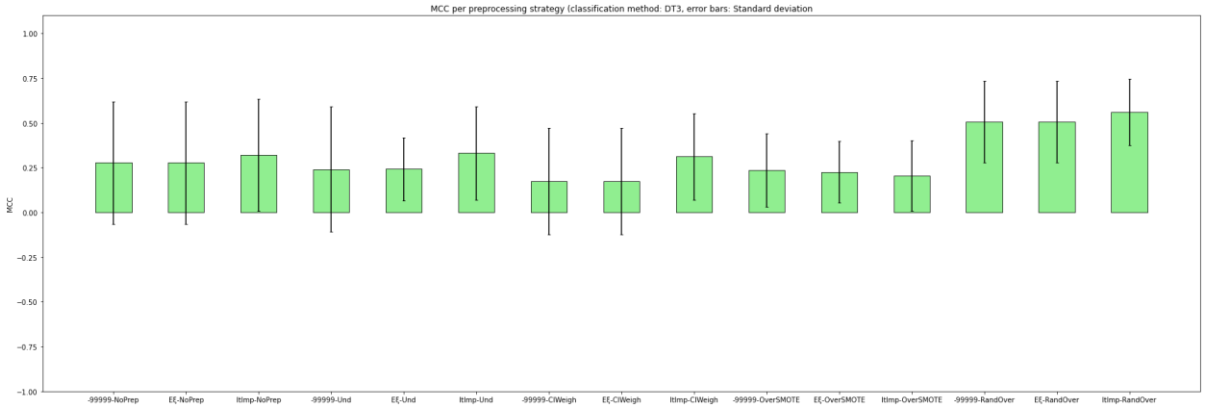
Δεν παραβλέπουμε το γεγονός ότι οι τυπικές αποκλίσεις είναι σχετικά μεγάλες αρκετά συχνά. Κάτι τέτοιο μειώνει την αξιοπιστία των αποτελεσμάτων όπου παρατηρείται.

Είναι επίσης πιθανό να προέκυψε υπερπροσαρμογή στα δεδομένα (overfitting) σε αρκετές περιπτώσεις (KNN, NN, RF, τυχαία υπερδειγματοληψία, SMOTE-υπερδειγματοληψία).

Έχοντας αυτά κατά νου μπορούμε να επιλέξουμε τρόπο με τον οποίο θα γίνει μία ανάλυση σε νέα κλινικά δεδομένα. Παρουσιάζονται οι ενδιαφέρουσες αποδώσεις ανά περίπτωση προεπεξεργασίας και κατά μέθοδο ταξινόμησης των δεδομένων σε κατηγορίες ασθενών που επιβιώνουν μετά το 1 έτος και ασθενών που δεν επιβιώνουν. Η αναφορά γίνεται προκειμένου να έχουμε το περιθώριο επιλογής της καταλληλότερης για αναάλυση σε νέα κλινικά δεδομένα.

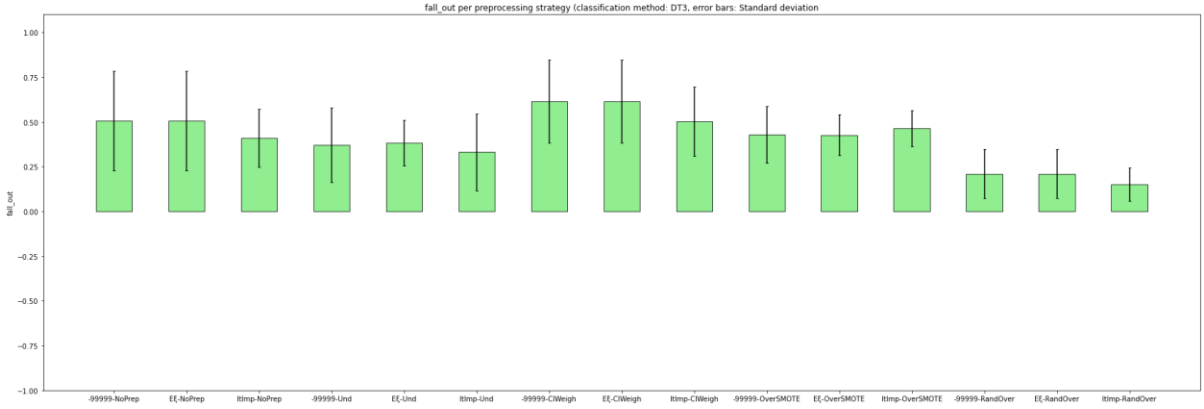
Σύγκριση πορειών προεπεξεργασιών

Η εξισσορόπηση με τυχαία υπερδειγματοληψία έχει αξιολογηθεί καλά, όπως αναμενόταν (βλ. Εικόνα 1, ενδεικτικά). Είναι πιθανό η καλή της απόδοση σε αρκετές περιπτώσεις να οφείλεται σε υπερπροσαρμογή στα δεδομένα (overfitting).

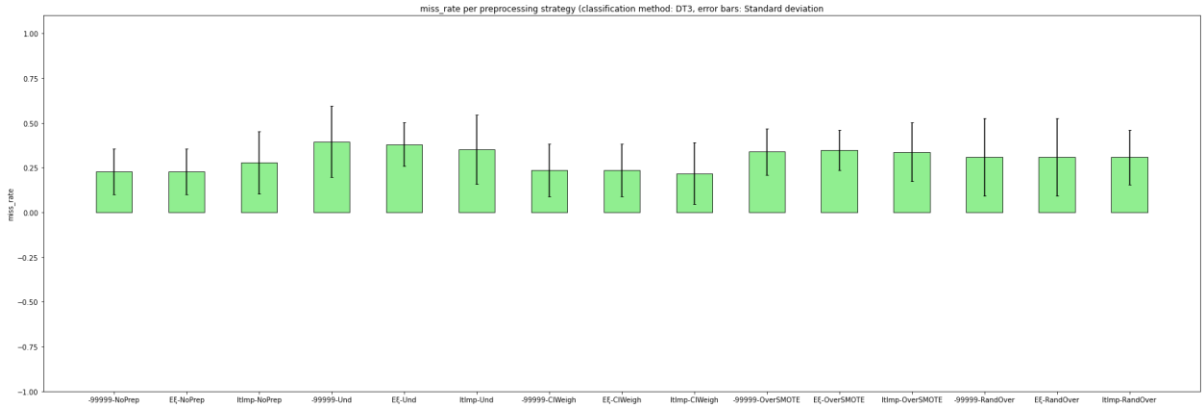


Εικόνα 1 Μετρική MCC για διάφορα σύνολα δεδομένων. Ταξινομήτρια μέθοδος είναι η DT3. Κατά σειρά από αριστερά προς τα δεξιά έχουμε (εμπειρική συμπλήρωση, χωρίς εξισορρόπηση), (συνάρτηση, χωρίς εξισορρόπηση), (iterative imputation, χωρίς εξισορρόπηση), (εμπειρική συμπλήρωση, υποδειγματοληψία), (συνάρτηση, υποδειγματοληψία), (iterative imputation, υποδειγματοληψία), (εμπειρική συμπλήρωση, class weights), (συνάρτηση, class weights), (iterative imputation, class weights), (εμπειρική συμπλήρωση, SMOTE-υπερδειγματοληψία), (συνάρτηση, SMOTE-υπερδειγματοληψία), (iterative imputation, SMOTE-υπερδειγματοληψία), (εμπειρική συμπλήρωση, τυχαία υπερδειγματοληψία), (συνάρτηση, τυχαία υπερδειγματοληψία), (iterative imputation, τυχαία υπερδειγματοληψία).

Δεν μπορούμε όμως να παραβλέψουμε ότι η τυχαία υπερδειγματοληψία φαίνεται να κάνει καλή ταξινόμηση και για τις δύο κλάσεις και να μην μεροληπτεί έντονα (βλ. Εικόνα 2,3 ενδεικτικά).

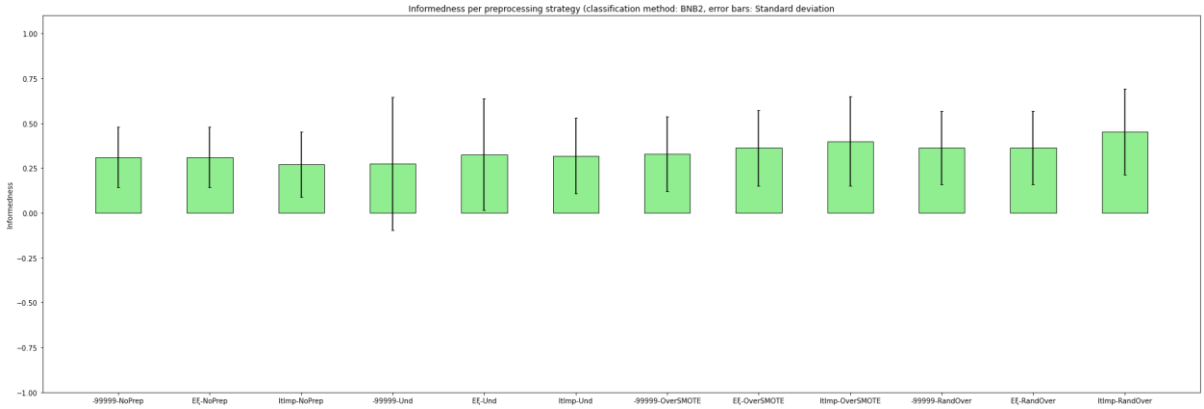


Εικόνα 2 Μετρική ψευδώς θετικό ποσοστό (fall-out) για διάφορες προεπεξεργασίες (οι ράβδοι έχουν την ίδια σειρά με Εικ1). Ταξινομήτρια DT3. Οι τιμές για τυχαία υπερδειγματοληψία (3 τελευταίες στήλες) είναι χαμηλές, όπως θα έπρεπε να είναι για αυτή την μετρική.



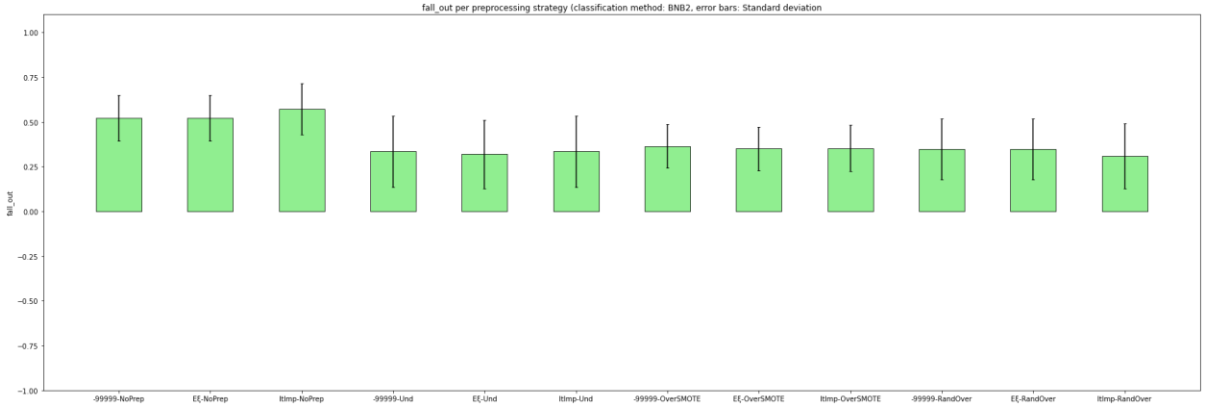
Εικόνα 3 Μετρική ψευδώς αρνητικό ποσοστό (Miss rate) για διάφορες προεπεξεργασίες δεδομένων. Ταξινομήτρια DT3. Οι τιμες για τυχαία υπερδειγματοληψία (3 τελευταίες στήλες) είναι σχετικά χαμηλές (αν και όχι κατώτατες). Σε αυτή την μετρική καλό είναι να έχουμε μικρές τιμές.

Σε αρκετές περιπτώσεις απόδοση όχι ιδιαίτερα χειρότερη από τις άλλες φαίνεται να έχει συνολικά και η έλλειψη εξισορρόπησης (βλ. Εικόνα 4, ενδεικτικά).



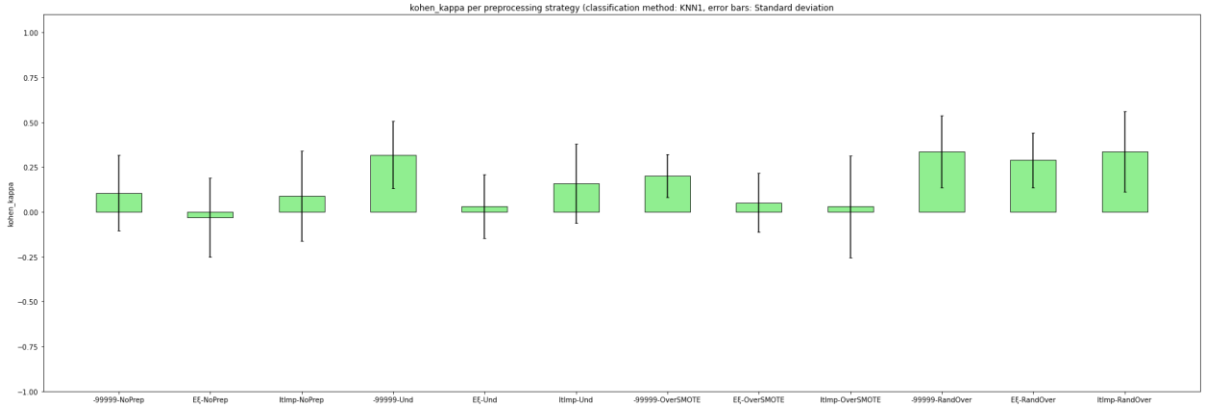
Εικόνα 4 Μετρική Πληροφοριακότητα (Informedness) για διάφορες προεπεξεργασίες. Ταξινομήτρια BNB2. Οι 3 πρώτες στήλες (έλλειψη εξισορρόπησης) δεν έχουν μεγάλη διαφορά με τις άλλες.

Ωστόσο αυτό συνήθως συνοδεύεται από μεροληψία υπέρ της πλειοψηφούσας κλάσης (κάτι που αποδεικνύεται και από την χαμηλή ειδικότητα (specificity), το υψηλό ψευδώς θετικό ποσοστό (fall-out) κοκ, (βλ Εικόνα 5).



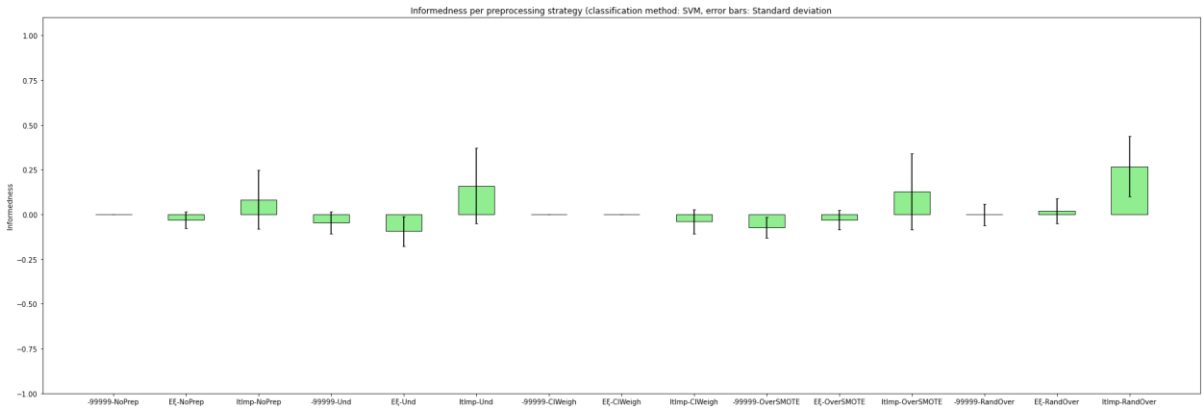
Εικόνα 5 Μετρική ψευδώς θετικό ποσοστό (fall-out) για διάφορες προεπεξεργασίες. Ταξινομήτρια BNB2. Σε αυτή την μετρική καλό είναι οι τιμές να είναι χαμηλές. Οι 3 πρώτες στήλες (έλλειψη εξισορρόπησης) έχουν μεγάλη διαφορά με τις άλλες.

Η καλή απόδοση της εμπειρικής συμπλήρωσης (βλ. Εικόνα 6) –η οποία συμβαίνει και σε περιπτώσεις μη εξισορρόπησης- μπορεί να οφείλεται είτε σε υπερπροσαρμογή στα δεδομένα (overfitting) (όπως πχ. για τυχαία υπερδειγματοληψία βλ. Εικόνα 13) είτε σε ιδιάζον της πλεονέκτημα. Όταν για παράδειγμα οι τιμές στα δεδομένα τείνουν να λείπουν πιο συχνά στην μία Κλάση από ό,τι στην άλλη, η συμπλήρωση των κενών με τόσο ακραίες τιμές όπως στην περίπτωση της εμπειρικής συμπλήρωσης (ή της συνάρτησης που δοκιμάσαμε) μπορεί να βοηθήσει την ταξινόμηση, μιας και διαφοροποιεί ακόμα περισσότερο τις καταγραφές που ανήκουν σε διαφορετικές κλάσεις. Χρειάζεται όμως περισσότερη διερεύνηση σε μελλοντική εργασία για να απαντήσουμε τί από τα δύο συμβαίνει εδώ.



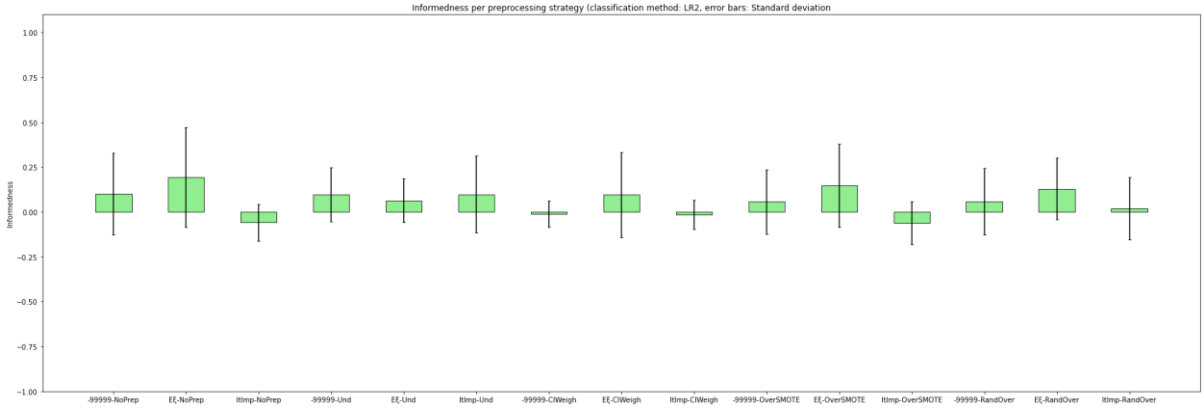
Εικόνα 6 τιμή κ του Cohen (Cohen's kappa) για διάφορες προεπεξεργασίες. Ταξινομήτρια KNN1

Ο Αλγόριθμος **SVM** αποδίδει καλύτερα σε σύνολα δεδομένων όπου έχει γίνει Iterative imputation. Κάτι τέτοιο αναμένεται για λόγο συνυφασμένο με τον ορισμό του αλγορίθμου. Η συμπλήρωση των κενών με τιμές απομακρυσμένες από το αρχικό εύρος τιμών (πχ. για εμπειρική συμπλήρωση ή συνάρτηση) καθιστά δύσκολο τον σωστό διαχωρισμό των δεδομένων με ένα υπερεπίπεδο όπως γίνεται μέσω του SVM. *Τα Support Vector Machines και άλλα μοντέλα που χρησιμοποιούν τον πυρήνα rbf* (που επιλέξαμε) *δεν κλιμακώνονται καλά σε μεγάλο αριθμό δειγμάτων εκπαίδευσης ή μεγάλο αριθμό χαρακτηριστικών στο χώρο εισόδου*. (“Radial basis function kernel,” n.d.) Συνεπώς εξηγείται η μη καλή απόδωσή του.



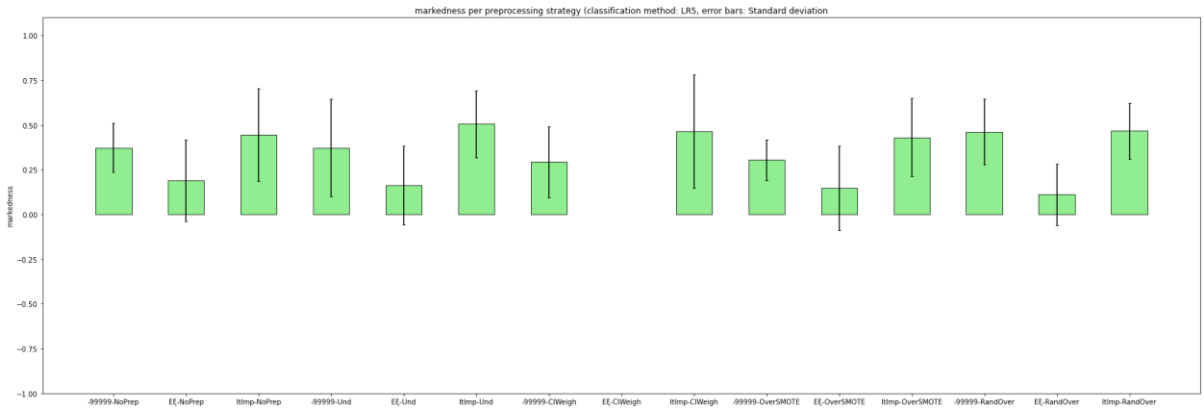
Εικόνα 7 Μετρική Πληροφοριακότητα (Informedness) για διάφορες προεπεξεργασίες. Ταξινομήτρια SVM.

Οι **LR1-4**, που μεροληπτούν έντονα υπέρ της κλάσης (0), αποδίδουν καλύτερα για (δοκιμαζόμενη συνάρτηση, χωρίς εξισορρόπηση) και ακολουθούν (δοκιμαζόμενη συνάρτηση, SMOTE υπερδειγματοληψία), (δοκιμαζόμενη συνάρτηση, τυχαία υπερδειγματοληψία). Η αποτελεσματικότητά τους είναι μάλιστα καλύτερη αν δεν γίνει καν εξισορρόπηση, βλ Εικόνα 8. Ωστόσο δεν πρόκειται για ιδιαίτερα αξιόπιστες μεθόδους (τουλάχιστον για τις συγκεκριμένες πορείες προεπεξεργασίας).



Εικόνα 8 Μετρική Πληροφοριακότητα (Informedness), για διάφορες προεπεξεργασίες. Ταξινομήτρια LR2.

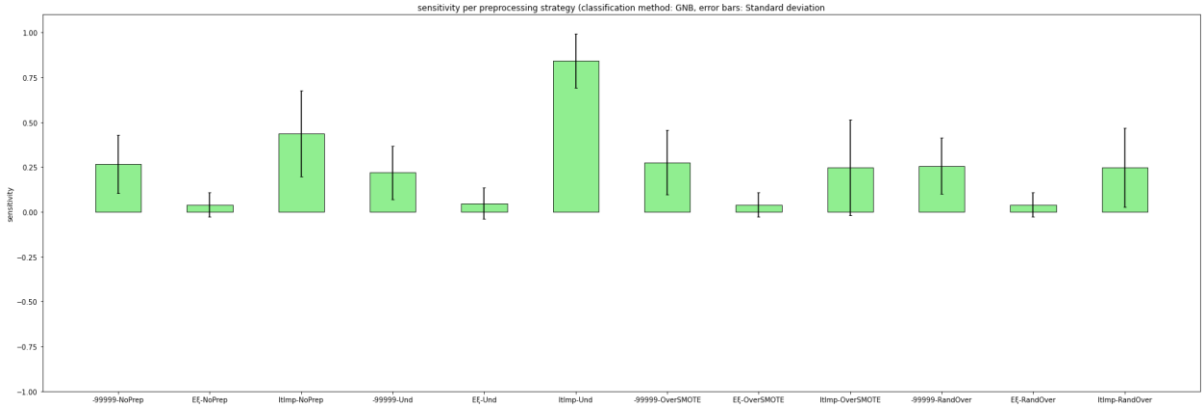
Αντίθετα οι **LR5-6**, που μεροληπτούν λιγότερο από τις προηγούμενες, αποδίδουν καλά και για χρήση του Iterative imputation μαζί με κάποια μέθοδο εξισορρόπησης. Χρήσιμο είναι το γεγονός ότι αποδίδουν καλύτερα και για υποδειγματοληψία, κάτι που μειώνει την πιθανότητα το αποτέλεσμα να οφείλεται σε υπερπροσαρμογή στα δεδομένα (κάτι που ειναι πιο πιθανό στην υπερδειγματοληψία). Αποδίδουν επίσης καλά για εμπειρική συμπλήρωση μαζί με τυχαία υπερδειγματοληψία, ένας συνδυασμός που γενικά παρατηρούμε πως αποδίδει καλά, μιας και συνδυάζει τα προαναφερθέντα χαρακτηριστικά/πλεονεκτήματα της εμπειρικής συμπλήρωσης και της τυχαίας υπερδειγματοληψίας, βλ Εικόνα 9, 13.



Εικόνα 9 Μέτρο του αξιοσημείωτου (markedness) για διάφορες προεπεξεργασίες. Ταξινομήτρια LR5.

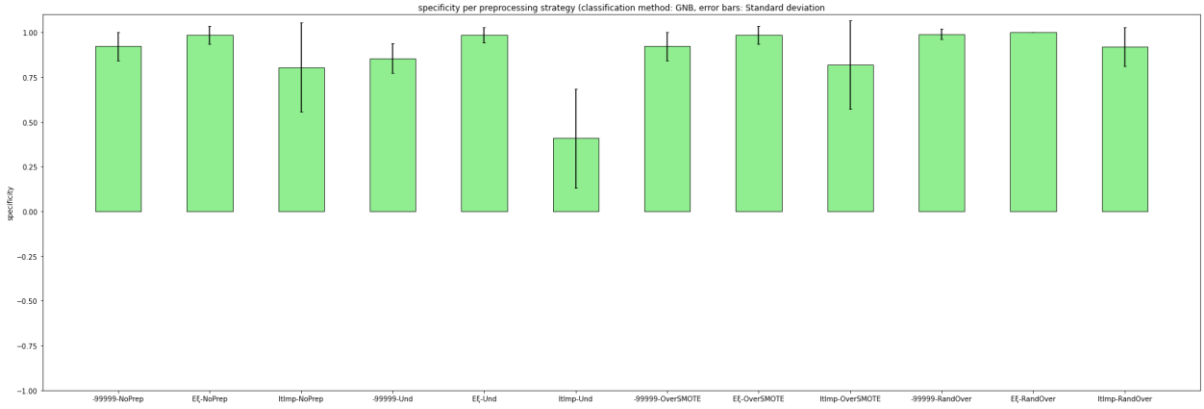
Τα **DT1-3** αποδίδουν καλύτερα για τυχαία υπερδειγματοληψία (iterative imputation, εμπειρική συμπλήρωση, συνάρτηση) βλ Εικόνα 1. Αντίστοιχα συμβαίνει και για το **DT4**, μόνο που αποδίδει και για (iterative imputation, χωρίς εξισορρόπηση / υποδειγματοληψία). Ίσως αυτή η απόδοση των DT να οφείλεται σε υπερπροσαρμογή (overfitting) (βλ. μέγιστο αριθμό κλάδων ανά κόμβο).

Επειδή η κλάση 1 είναι η πλειοψηφούσα κλάση (όταν δεν έχουμε εξισορρόπηση), η ειδικότητα (specificity) είναι πολύ σημαντικό να είναι υψηλή προκειμένου να βρεθούν κατά την ταξινόμηση όσο περισσότεροι οι ασθενείς της μειοψηφούσας κλάσης. Σημειώνουμε ότι ο ταξινομητής **GNB** έχει αξιοσημείωτη ειδικότητα σχεδόν σε κάθε κατηγορία προεπεξεργασίας, παρόλα αυτά η ευαισθησία (sensitivity) του είναι κακή (χαμηλή) σε γενικές γραμμές.



Εικόνα 10 Μετρική ευαισθησία (sensitivity) για διάφορες προεπεξεργασίες. Ταξινομήτρια GNB.

Παρατηρούμε ότι η ευαισθησία για (iterative imputation, υποδειγματοληψία) είναι κατ' εξαίρεση καλή, βλ. Εικόνα 10, αλλά η αντίστοιχη ειδικότητα είναι κακή, Εικόνα 11.



Εικόνα 11 Μετρική ειδικότητα (specificity), Ταξινομήτρια GNB

Τα **RF1-2** αποδίδουν καλύτερα για (συνάρτηση, υποδειγματοληψία), τυχαία υπερδειγματοληψία και με σχετικά μικρή διαφορά για SMOTE-υπερδειγματοληψία, τις υπόλοιπες υποδειγματοληψίες και τη (συνάρτηση, class_weights). Η συνάρτηση που δοκιμάζεται συνδυαστικά με την υποδειγματοληψία ξεπερνά την εμπειρική συμπλήρωση στην αντίστοιχη κατηγορία. Γενικά τα RF έχουν από τις καλύτερες αποδόσεις μεταξύ των ταξινομητριών.

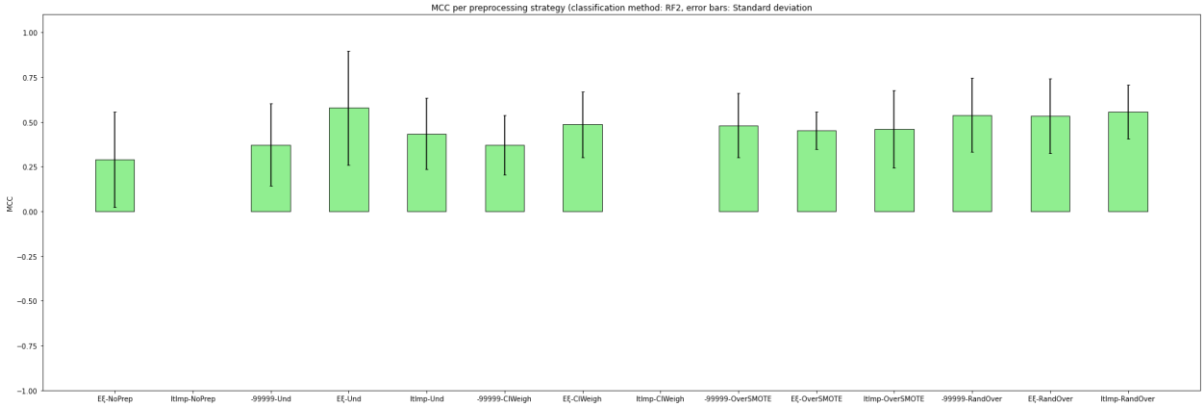


Figure 12 Μετρική Συντελεστής συσχέτισης Matthews (MCC), ταξινομήτρια RF2

Για τα **NN** βλέπουμε ότι τα πάει ιδιαίτερα καλύτερα, σε κάθε NN, η (εμπειρική συμπλήρωση, τυχαία υπερδειγματοληψία), ακολουθούν οι υπόλοιπες τυχαίες υπερδειγματοληψίες και έπονται τα iterative imputation με υποδειγματοληψία, SMOTE-υπερδειγματοληψία και έλλειψη εξισορρόπησης.

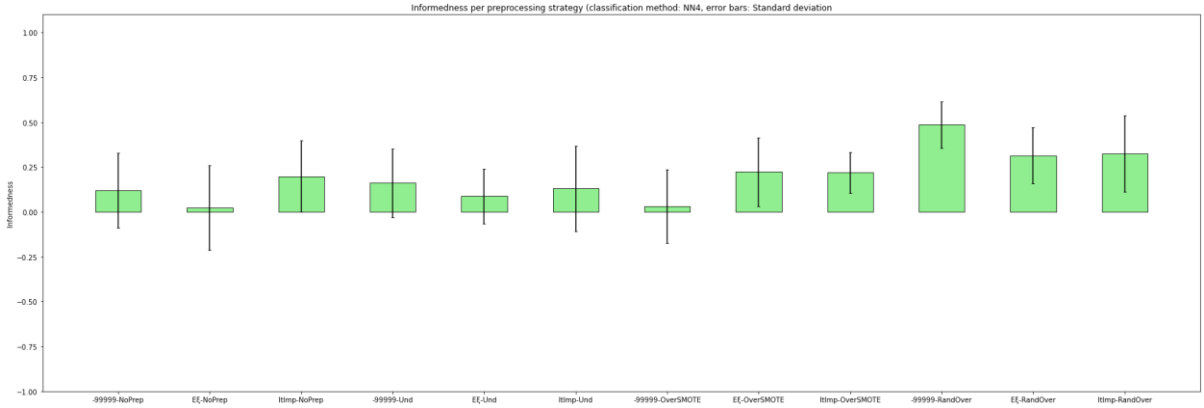
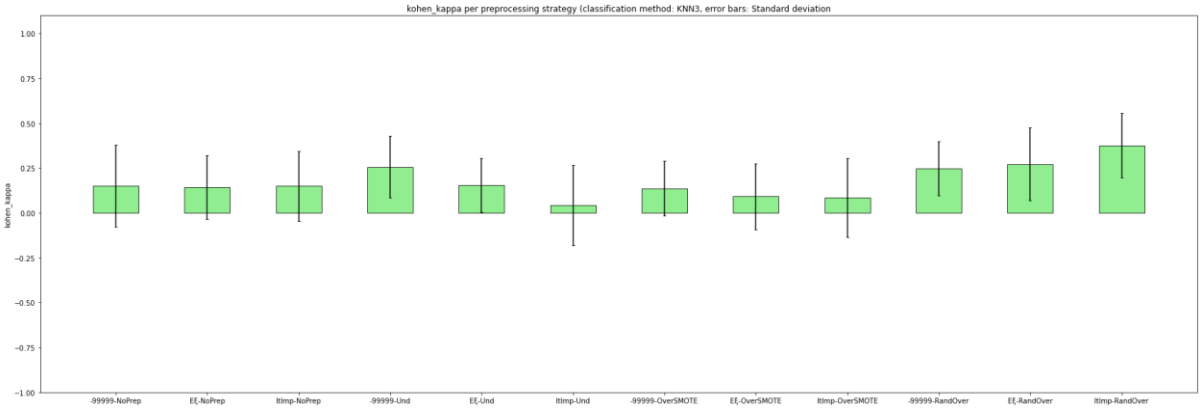


Figure 13 Πληροφοριακότητα (Informedness)*, Ταξινομήτρια NN4*

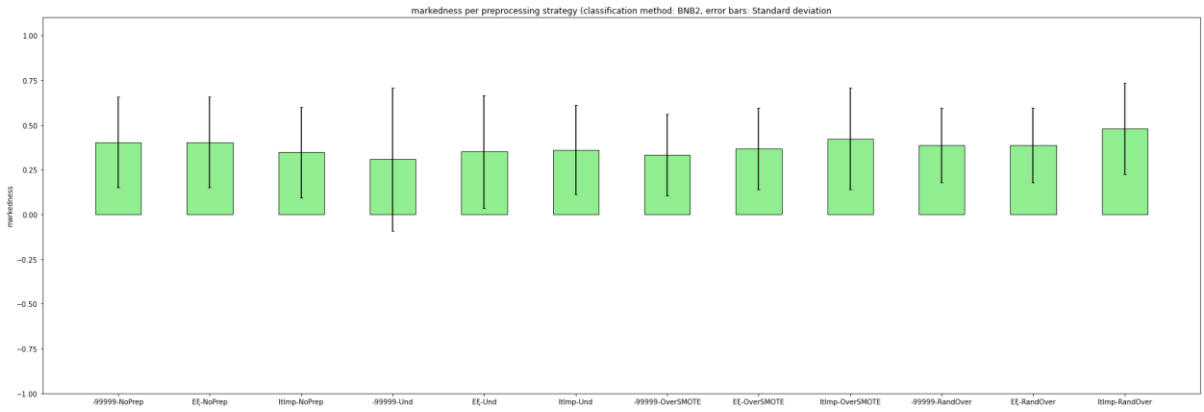
Στα KNN2,4 καλά αποδίδει ο συνδυασμός (Iterative imputation, τυχαία υπερδειγματοληψία), βλ. Εικ. 14, κάτι το αναμενόμενο δεδομένου του ότι οι άλλες μέθοδοι συμπλήρωσης κενών έδιναν τιμές μη γειτονικές, που δυσκόλευαν τον αλγόριθμο αυτό. Επίσης καλά τα πηγαίνει και η τυχαία υπερδειγματοληψία (random-oversampling) γενικά στα **KNN**. *Όταν ο αριθμός των γνωρισμάτων αυξάνεται απαιτούνται περισσότερα δεδομένα και δημιουργείται πρόβλημα υπερπροσαρμογής. Για να αποφευχθεί η υπερβολική προσαρμογή, τα απαραίτητα δεδομένα θα πρέπει να αυξηθούν εκθετικά καθώς αυξάνεται ο αριθμός των διαστάσεων.* (Avinash Navlani, 2018) Στην περίπτωσή μας αυτό δεν συμβαίνει, άρα αναμένεται ότι η απόδοση του συγκεκριμένου αλγορίθμου μπορεί να οφείλεται σε υπερπροσαρμογή. *Σε μεγάλες διαστάσεις η Ευκλείδεια απόσταση δεν είναι πλέον χρήσιμη και είναι ευαίσθητη σε μεγέθη. Τα γνωρίσματα με μεγάλα μεγέθη θα έχουν μεγαλύτερο βάρος από τα γνωρίσματα με μικρά μεγέθη.* (Avinash Navlani, 2018) Εξηγείται συνεπώς η μέτρια απόδωσή του, σε περιπτώσεις ισορροπίας κλάσεων, αλλά και σε περιπτώσεις ανισορροπίας (την ώρα που άλλοι αλγόριθμοι αποτυγχάνουν). *Το KNN μπορεί να είναι χρήσιμο και σε περίπτωση μη γραμμικών δεδομένων. Αποδίδει καλύτερα με μικρότερο αριθμό γνωρισμάτων.* (Avinash Navlani, 2018)



Εικόνα 14 τιμή κ του Cohen (Cohen's kappa), Ταξινομήτρια KNN3

Η KNN είναι μη παραμετρική μέθοδος, άρα δεν κάνει υπόθεση για υποκείμενη κατανομή των δεδομένων. Αυτό θεωρείται πλεονέκτημα λόγω του ότι τα περισσότερα σύνολα δεδομένων του πραγματικού κόσμου δεν ακολουθούν τις μαθηματικές θεωρητικές υποθέσεις(Avinash Navlani, 2018) και στην συγκεκριμένη περίπτωση πράγματι η πλήρης κατανομή των δεδομένων δεν είναι γνωστή.

Στα **BernNB1-2** υπάρχει μικρή διαφορά για τις διάφορες προεπεξεργασίες. Δεν ξεχωρίζει μόνο η τυχαία υπερδειγματοληψία (random-oversampling). Ειδικά για το BNB2, έχουμε μία σταθερά μέτρια απόδοση ανεξαρτήτως προεπεξεργασίας, γεγονός ενδιαφέρον. Εν μέρει εξηγείται και από το ότι το πρόβλημα ταξινόμησης που έχουμε αφορά δυαδική ταξινόμηση.

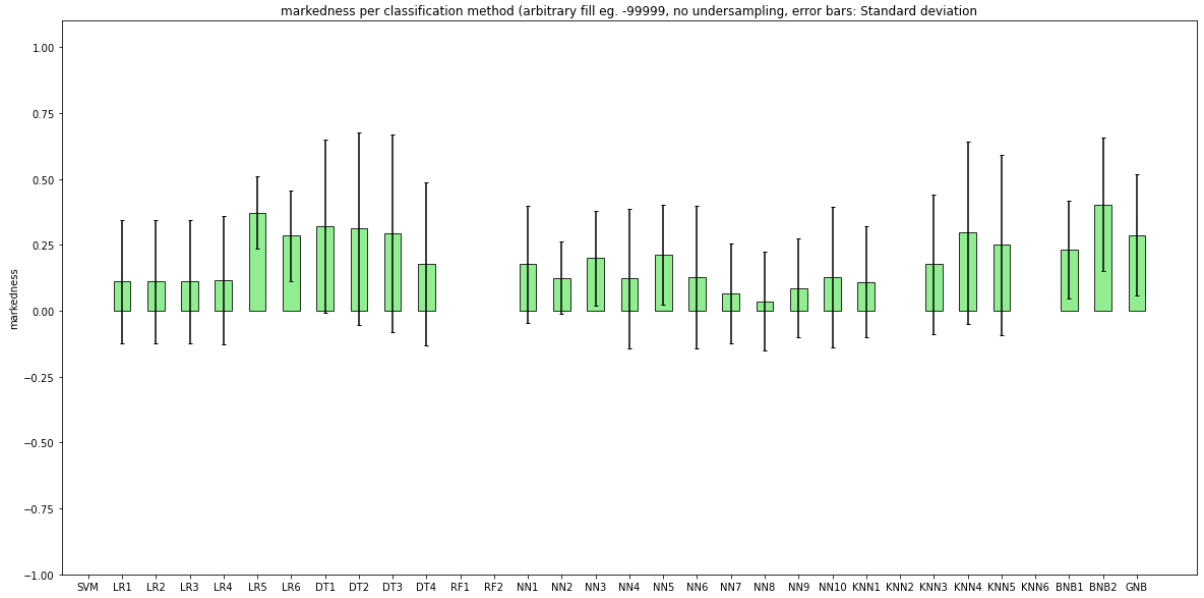


Εικόνα 15 Μέτρο του αξιοσημείωτου (markedness), Ταξινομήτρια BNB2

Είναι αναμενόμενο τα δύο μοντέλα της μεθόδου ταξινόμησης Τυχαίων Δασών (Random Forest-RF) να αποδίδουν καλύτερα από τα Δέντρα αποφάσεων (Decision Trees-DT), μιας και η πρώτη μέθοδος αποτελεί «concensus» μέθοδο της δεύτερης. Στην μέθοδο ταξινόμησης Τυχαίων Δασών απαιτούνται πολλά Δέντρα αποφάσεων, μετά από αξιολόγηση των οποίων και επιλογή των καλύτερων ανά περίπτωση, θα προκύψει το τελικό αποτέλεσμα της Ταξινόμησης. Ωστόσο υπάρχουν περιπτώσεις που τα DT τα πηγαίνουν αισθητά καλύτερα από τα RF. (Βλ. Εικόνα ii)

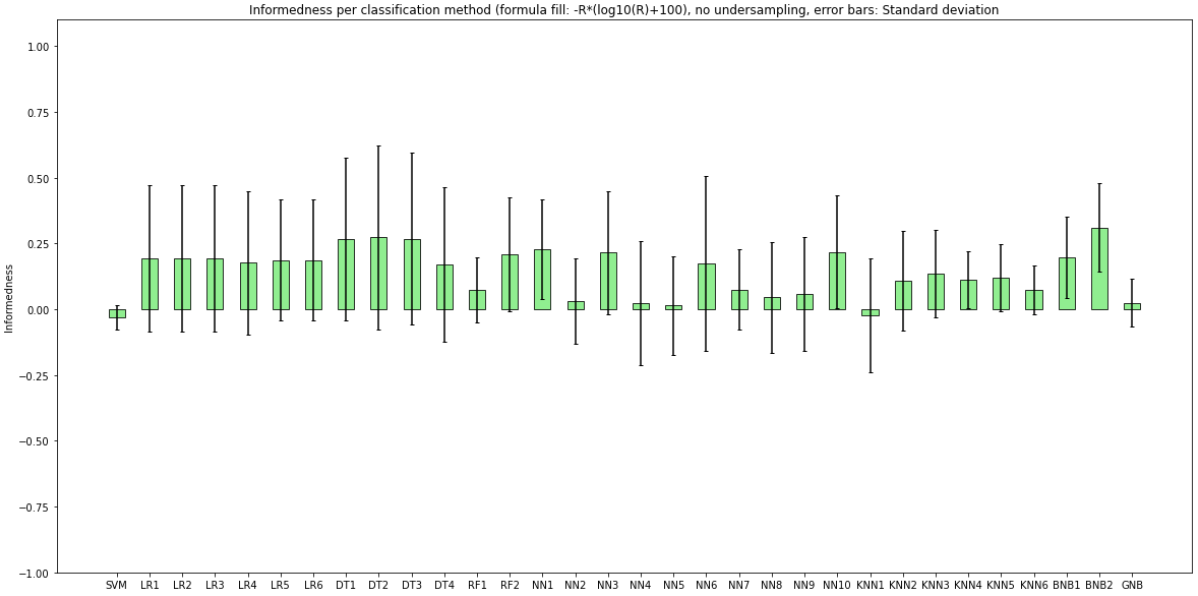
Σύγκριση μεθόδων ταξινόμησης

Στην περίπτωση προεπεξεργασίας όπου έγινε **εμπειρική συμπλήρωση κενών, χωρίς να γίνει εξισορρόπηση** έχουμε καλύτερη απόδοση για LR5 και BNB2 και ακολουθούν οι LR6, DT1-3, RF2 (Εικόνα i)



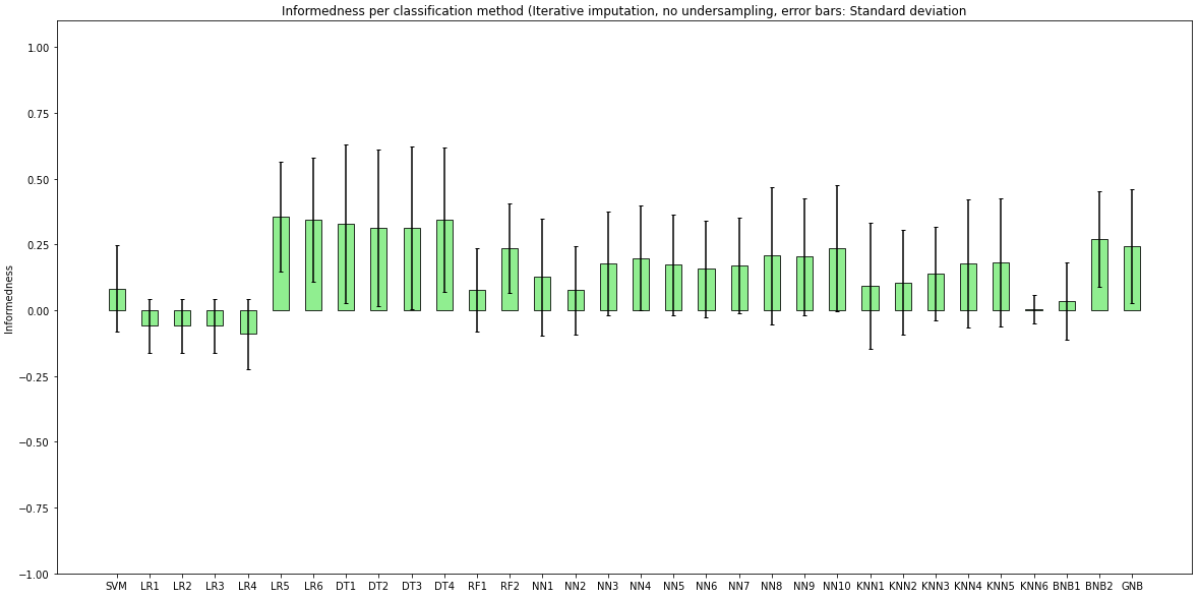
Εικόνα i Μέτρο του αξιοσημείωτου (markedness) για εμπειρική συμπλήρωση κενών και έλλειψη εξισορρόπησης. Παρατηρείται ότι ξεχωρίζει σε απόδοση η μέθοδος BNB2 και LR5.

Για συμπλήρωση κενών με την **συνάρτηση που δοκιμάσαμε και έλλειψη εξισορρόπησης** έχουμε καλή απόδοση για BNB2, DT1-3, RF2, NN1, NN3, NN10. (Εικόνα ii)



Εικόνα ii Μετρική Πληροφοριακότητα (Informedness). Συμπλήρωση κενών με την συνάρτηση που δοκιμάσαμε και έλλειψη εξισορρόπησης

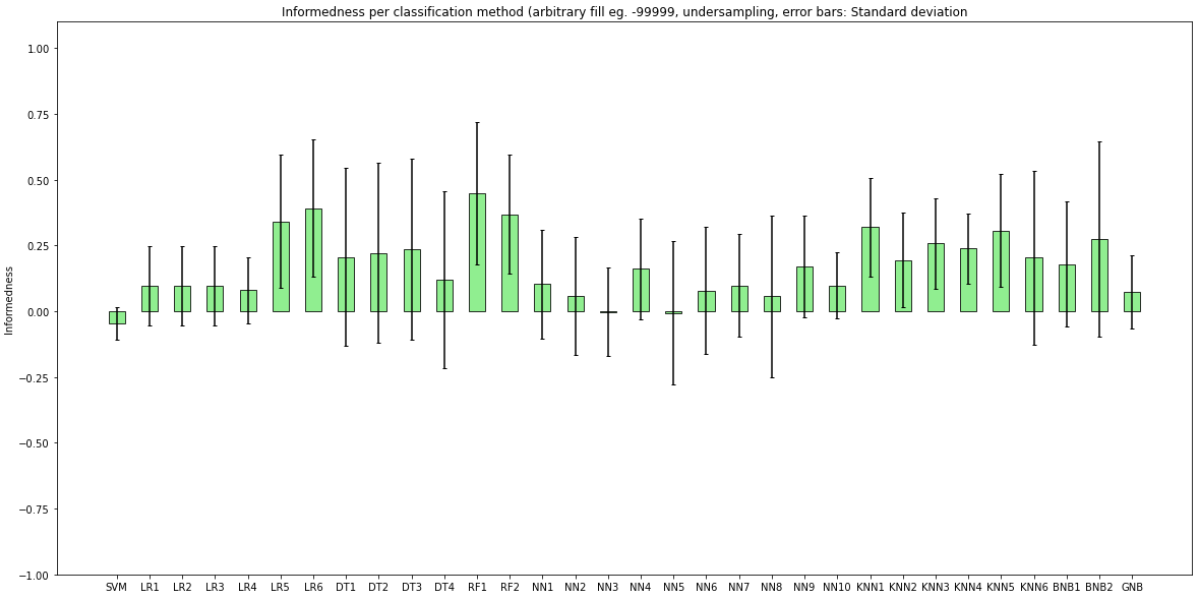
Για **Iterative imputation, χωρίς εξισορρόπηση** έχουμε καλή απόδοση για LR5-6, τα DT1-4 και ακολουθούν τα NN7-10, KNN5, BNB2, GNB (Εικόνα iii).



Εικόνα iii Μετρική Πληροφοριακότητα (Informedness) για iterative imputation, έλλειψη εξισορρόπησης

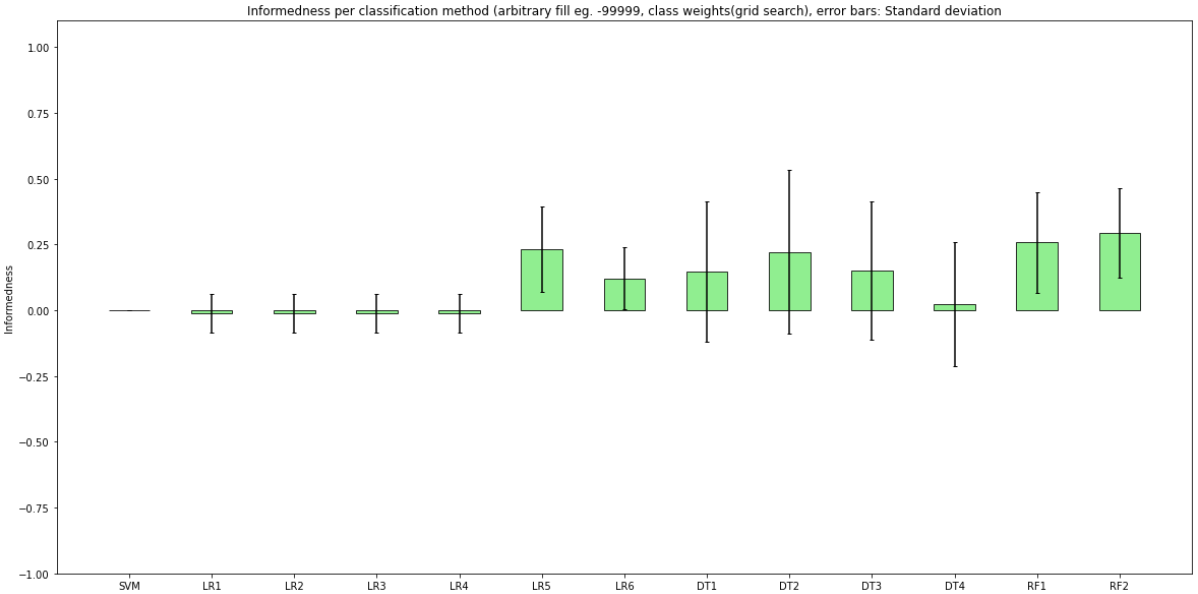
Γενικά **ελλείψει εξισορρόπησης** καλύτερη απόδοση έχει το BNB2, τα LT5-6, τα DT και κάποια NN και το RF2.

Για εξισορρόπηση με **Υποδειγματοληψία (Undersampling) και εμπειρική συμπλήρωση** καλές είναι οι RF1 και ακολουθούν οι LR6, RF2, KNN1. Για συμπλήρωση με την **συνάρτηση** καλύτερες είναι οι ταξινομήτριες μέθοδοι RF2, RF1. Για **iterative imputation** καλύτερες είναι οι LR6, LR5, RF1-2. Συνολικά για εξισορρόπηση με Υποδειγματοληψία (Undersampling) καλύτερα τα πηγαίνουν τα RF και LR6. (Εικόνα iv)



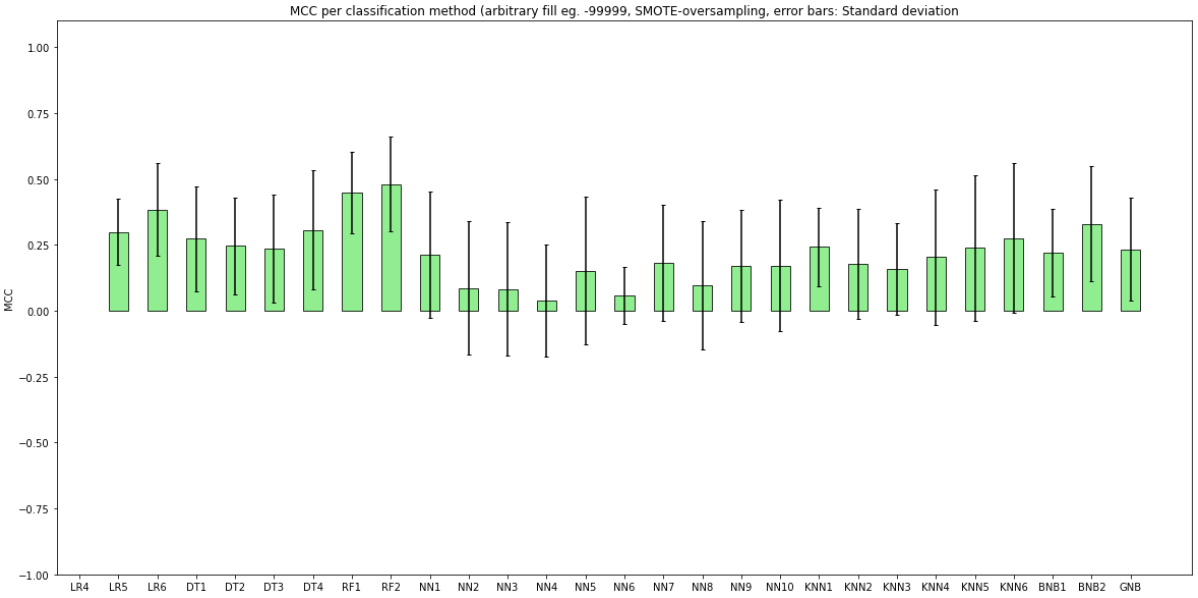
Εικόνα iv Μετρική Πληροφοριακότητα (Informedness) για εμπειρική συμπλήρωση, υποδειγματοληψία

Για εξισορρόπηση με class weights και εμπειρική συμπλήρωση καλύτερη απόδοση έχουν οι RF2, RF1 και ακολουθούν οι LR5, DT2. Για συμπλήρωση με τη δική μας συνάρτηση καλύτερες είναι οι RF1-2 και για iterative imputation και οι LR5-6. Συνολικά τα πάνε καλά οι RF, DT1-3.



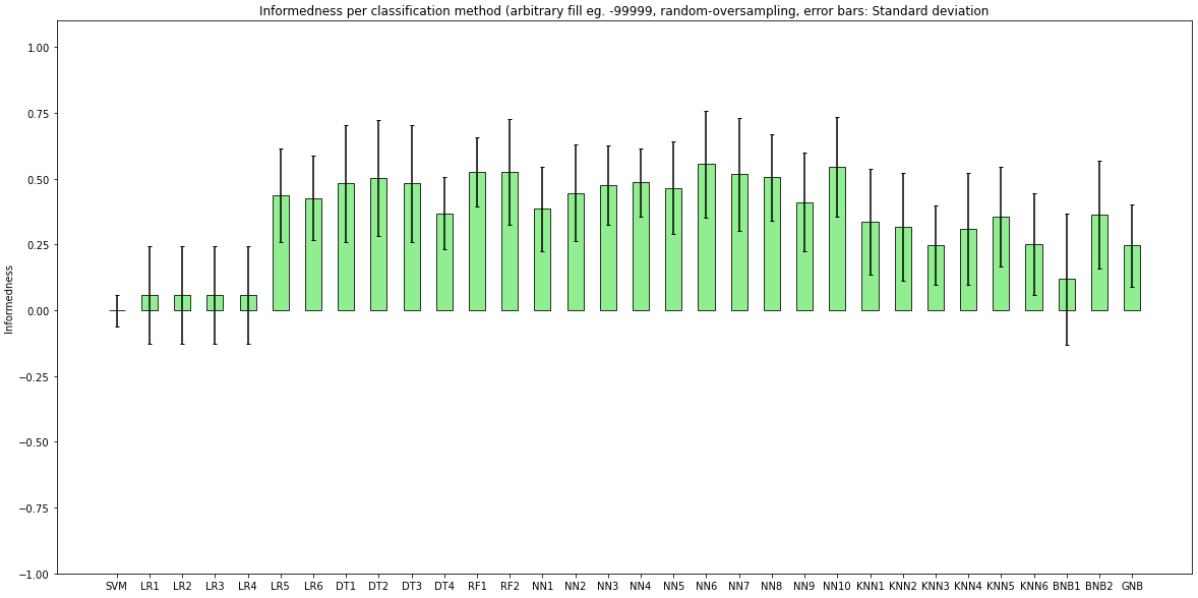
Εικόνα v Μετρική Πληροφοριακότητα (Informedness). Εμπειρική συμπλήρωση, class weights

Για εξισορρόπηση με Υπερδειγματοληψία με τη μέθοδο SMOTE (SMOTE-oversampling) καλύτερα τα πάνε οι RF1-2, ενώ οι υπόλοιπες μέθοδοι μεροληπτούν περισσότερο υπέρ της μίας ή της άλλης κλάσης (επιβιώσαντες στο 1 έτος ή μη). Ακολουθεί, λίγο χειρότερα το BNB2 και η LR6. Τα KNN αποδίδουν λιγότερο, μαζί με τα DT (Εικ. vi)



Εικόνα vi Μετρική Συντελεστής συσχέτισης Matthews (MCC) Εμπειρική συμπλήρωση, SMOTE-υπερδειγματοληψία

Για εξισορρόπηση με τυχαία υπερδειγματοληψία (random-oversampling) και εμπειρική συμπλήρωση κενών έχουμε καλύτερα αποτελέσματα για NN6-8, NN10, ενώ ακολουθούν κοντά τα υπόλοιπα NN, μαζί με τα DT1-3, RF1-2 και LR5-6. Στην συγκεκριμένη προεπεξεργασία τα NN είχαν την καλύτερή τους απόδοση, κάτι που φαίνεται μιας και πλησιάζουν τα RF & DT.

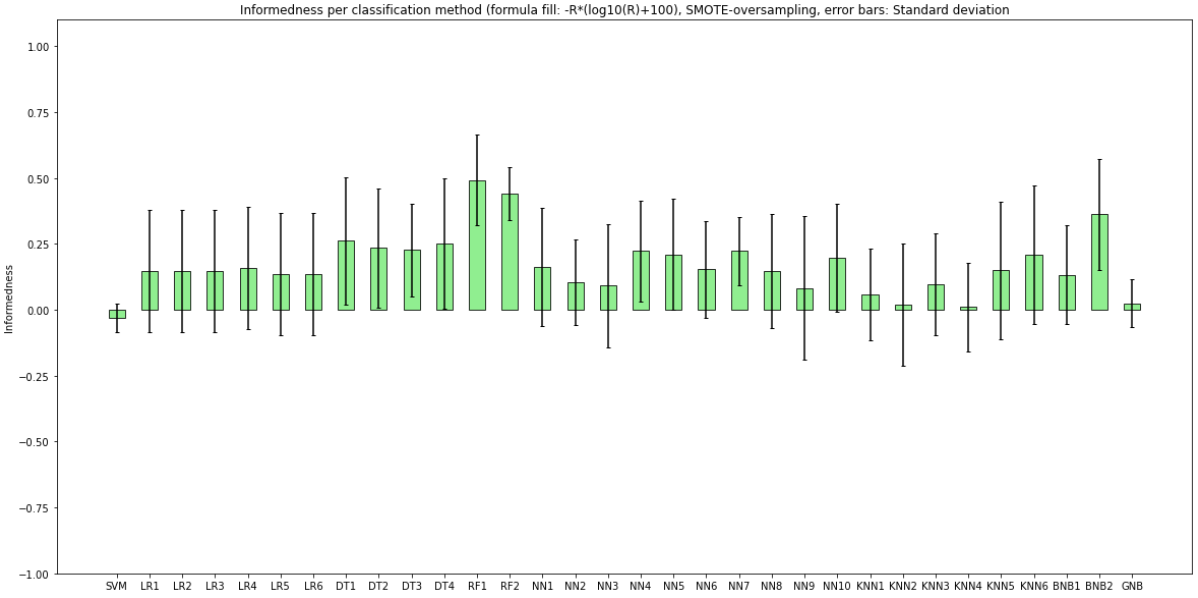


Εικόνα vii Μετρική Πληροφοριακότητα (Informedness). Εμπειρική συμπλήρωση, τυχαία υπερδειγματοληψία

Γενικά στην τυχαία υπερδειγματοληψία (random-oversampling) οι RF1-2, DT1-3, κάποια NN(6-8, 10) και οι LR5, LR6, BNB2 τα πάνε καλά .

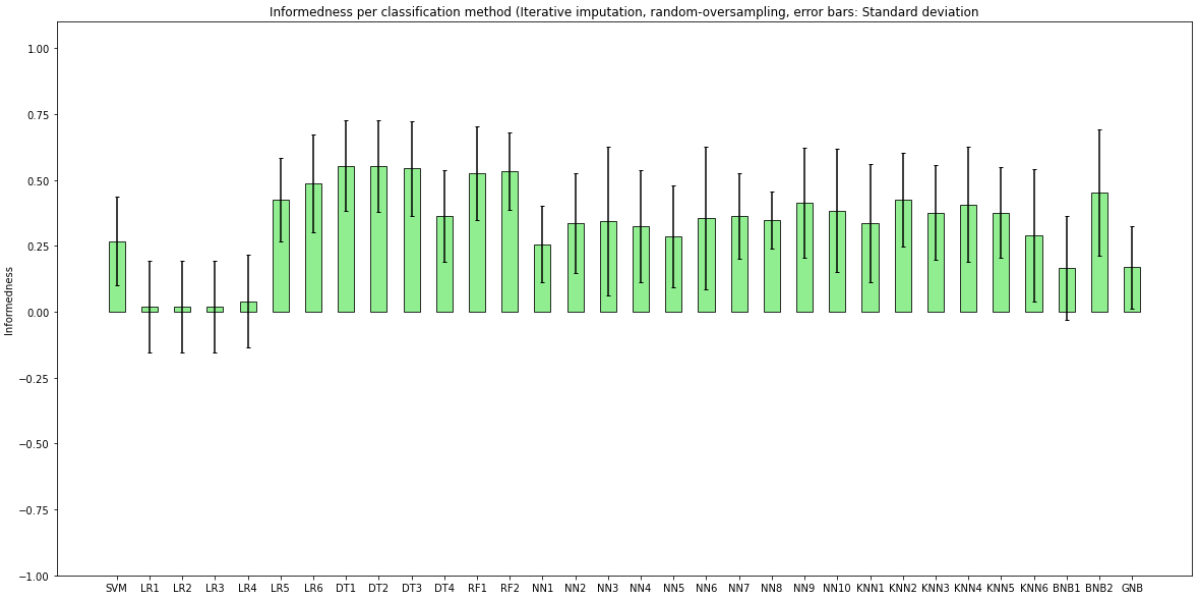
Για εμπειρική συμπλήρωση καλύτερα τα πάνε οι LR5-6, RF1-2, DT1-3, BNB2.

Για την συνάρτηση που δοκιμάστηκε καλύτερα τα πηγαίνουν οι RF1-2, BNB2 και λιγότερο τα DT1-3. Εξαίρεση αποτελεί η τυχαία υπερδειγματοληψία, όπου βελτιώνονται τα NN.



Εικόνα viii Πληροφοριακότητα (Informedness). Συνάρτηση, SMOTE-υπερδειγματοληψία.

Για συμπλήρωση με Iterative imputation καλύτερα τα πηγαίνουν οι LR5-6, RF1-2 και ακολουθούν τα DT1-3 και BNB2 κά. Εξαίρεση έχουμε για τον συνδυασμό με την τυχαία υπερδειγματοληψία, βλ. Εικόνα xi. Πιθανώς εδώ να κάνουν υπερπροσαρμογή στα δεδομένα (overfitting) κάποιες μέθοδοι που βελτιώθηκαν.



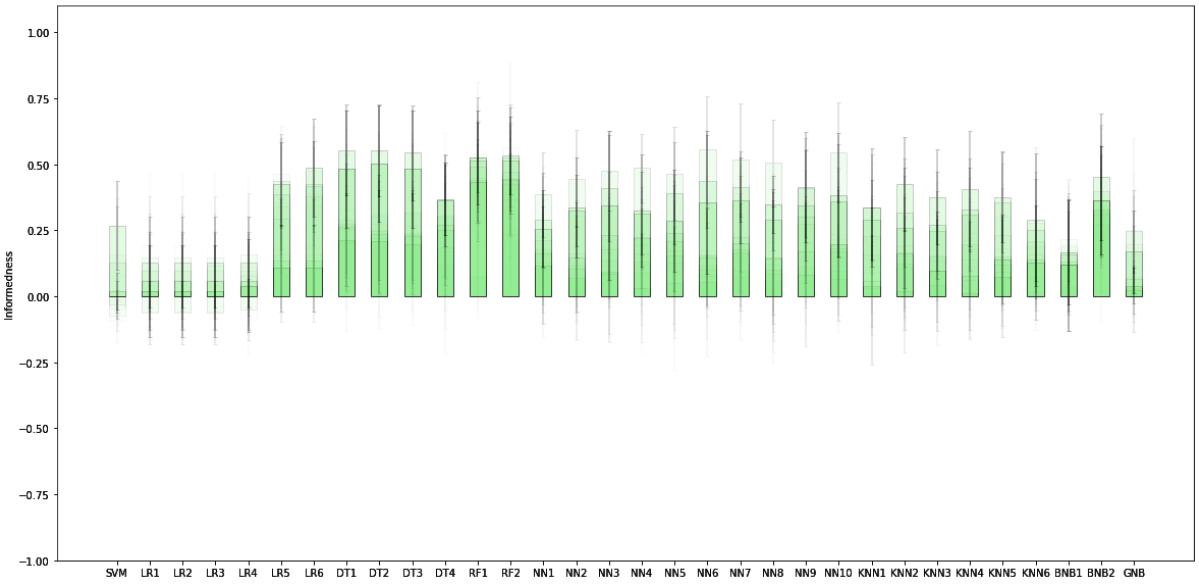
Εικόνα ix Μετρική Πληροφοριακότητα(Informedness). Iterative imputation,τυχαία υπερδειγματοληψία

Η τυχαία υπερδειγματοληψία ως τρόπος εξισορρόπησης του συνόλου δεδομένων, τα έχει πάει καλά γενικά χωρίς να μεροληπτεί ιδιαίτερα. Ωστόσο είναι πολύ πιθανό να υπάρχει υπερπροσαρμογή στα δεδομένα, μιας και αυτή η μέθοδος εξισορρόπησης είναι επιρρεπής στην υπερπροσαρμογή. Χαρακτηριστικά καλά τα πηγαίνει όταν συνδυάζεται με εμπειρική συμπλήρωση των κενών τιμών στα δεδομένα. Αυτό μπορεί να είναι είτε μία ακόμη ένδειξη υπερπροσαρμογής στα δεδομένα είτε ιδιάζον πλεονέκτημα του συνδυασμού αυτού, κάτι που χρειάζεται περεταίρω διερεύνηση. Η υποδειγματοληψία ήταν και αυτή αξιόλογη μέθοδος εξισορρόπησης, που παρουσίασε ενδιαφέρον καθώς είναι μη επιρρεπής στην υπερπροσαρμογή στα δεδομένα, με απόδοση καλή σε αρκετές περιπτώσεις. Η υπερδειγματοληψία με SMOTE είναι επίσης αξιόλογη μέθοδος εξισορρόπησης, όμως δεν φτάνει συνήθως την απόδοση της τυχαίας υπερδειγματοληψίας. Παρόλα αυτά η γενική επίδοσή της δεν ήταν κακή. Η χρήση της παραμέτρου class_weights για εξισορρόπηση της συνεισφοράς των καταγραφών των δύο κατηγοριών των ασθενών στην εκπαίδευση των ταξινομικών μοντέλων δεν ανταποκρίθηκε στις προσδοκίες που υπήρχαν, μιας και δεν είχε αντίστοιχα σταθερά καλή απόδοση για αυτήν, σε σύγκριση με τις υπόλοιπες μεθόδους εξισορρόπησης. Ωστόσο και σε αυτήν υπήρξαν συνδυασμοί προεπεξεργασίας και μεθόδων ταξινόμησης με αποδεκτή απόδοση. Η έλλειψη εξισορρόπησης δεν προτείνεται να επιλεγεί λόγω της έντονης μεροληψίας που παρουσιάζει.

Ενδιαφέρον παρουσίασε σε ορισμένες περιπτώσεις η σχετικά μέτρια απόδοση για συμπλήρωση των κενών στα δεδομένα με εμπειρική συμπλήρωση, ωστόσο δεν είναι σκόπιμο να επιλεγεί. Η δοκιμαζόμενη συνάρτηση σαν μέθοδος συμπλήρωσης των κενών στα δεδομένα δεν βοήθησε σημαντικά στην βελτίωση της απόδοση της ταξινόμησης, πέρα από λίγες περιπτώσεις όπου τα μοντέλα που προέκυψαν ήταν πολύ μεροληπτικά (LR1-4). Πιο αξιόπιστη μέθοδος συμπλήρωσης κενών φαίνεται το Iterative imputation, το οποίο αν και σε ορισμένες φορές δεν βοηθούσε στην επίτευξη της καλύτερης απόδοσης ταξινόμησης, είχε μεγαλύτερη αξιοπιστία, μιας και σε συνδυασμό με κάποια μέθοδο εξισορρόπησης του συνόλου δεδομένων βοηθούσε στην καλή απόδοση των ταξινομητριών μεθόδων, ακόμη και για υποδειγματοληψία της πλειοψηφούσας κατηγορίας ασθενών σαν μέθοδο εξισορρόπησης των δεδομένων, μέθοδο που δεν είναι τόσο επιρρεπής στην υπερπροσαρμογή.

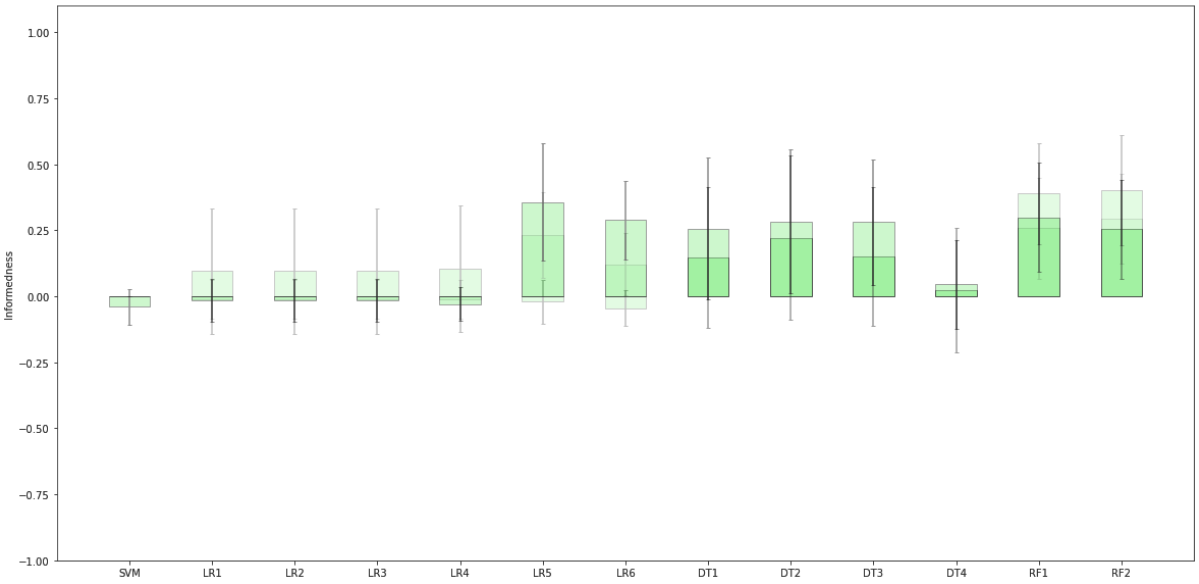
Ορισμένες από τις ταξινομήτριες μέθοδους που δοκιμάστηκαν σε συνδυασμό με προεπεξεργασίες που επίσης δοκιμάστηκαν ξεπέρασαν σε απόδοση άλλες μεθόδους που είχαν προταθεί στην βιβλιογραφία, ως προς την ακρίβεια ή ορθότητα (accuracy) και την εξισορροπημένη ακρίβεια (balanced accuracy), όπως αναφέραμε και στην αρχή της ενότητας Αποτελέσματα. Ωστόσο δεν είναι δόκιμη μία τέτοια σύγκριση από την στιγμή που δεν έχουμε εικόνα και για τις τιμές που έχουν πετύχει ως προς άλλες μετρικές. Η συμπερίληψη μετρικών που αξιολογούν από διαφορετικές οπτικές την ταξινόμηση είναι κάτι που επιλέξαμε να κάνουμε λόγω του ότι σπάνια έχουμε σφαιρική εικόνα για την συνολική ποιότητα μιας ταξινόμησης στις περισσότερες αναλύσεις της βιβλιογραφίας. Μία σύγκριση βασισμένη μόνο στις μετρικές ορθότητα και εξισορροπημένη ακρίβεια μπορεί να παραπλανούσε και να εμφάνιζε κάποιο δικό μας μοντέλο σαν καλύτερο, χωρίς απαραίτητα να είναι ή και το αντίθετο, μιας και για την συνολική αξιολόγηση μιας ταξινόμησης θα χρειαζόταν να έχουμε γνώση και για την τιμή άλλων μετρικών που έχουμε συμπεριλάβει, όπως λχ. της Πληροφοριακότητας. Ο λόγος είναι ότι χρειάζεται να διερευνηθεί αν η φαινομενικά καλύτερη απόδοση που είχαμε βρει οφείλεται πράγματι σε καλύτερη ταξινόμηση και όχι στις αδυναμίες των μετρικών αξιολόγησης ορθότητα και εξισορροπημένη ακρίβεια, που επηρεάζονται μεταξύ άλλων και από την

ανισορροπία στα δεδομένα. Θα ήταν χρήσιμο σε μελλοντική ανάλυση να επαναληφθούν τα πειράματα των ερευνητικών ομάδων που είχαν αντίστοιχη απόδοση και να υπολογιστούν όλες οι μετρικές που συμπεριλάβαμε.



Εικόνα 10 Μετρική Πληροφοριακότητα (informedness) Κολάζ των γραφημάτων για κάθε προεπεξεργασία πλην του class_weights

Όπως βλέπουμε και στην Εικόνα 10 οι ταξινομήτριες μέθοδοι που είναι καλύτερες συνολικά είναι οι RF1, RF2, ενώ ιδιαίτερο ενδιαφέρον έχει η σταθερά μέτρια απόδοση του BNB2, με τα DT4, LR5-6 και DT1-3 και DT4 να έχουν καλή απόδοση για ορισμένες προεπεξεργασίας και όχι για όλες. Αυτό το μοτίβο γενικά επιβεβαιώμενται και για την περίπτωση χρήσης της παραμέτρου class_weights (βλ. Εικ. 11). Τα NN έχουν μεγάλη διακύμανση ανάλογα με το ποια θα είναι η προεπεξεργασία που επιλέγεται.



Εικόνα 11 Μετρική Πληροφοριακότητα κολάζ γραφημάτων για τις περιπτώσεις προεπεξεργασίας με χρήση της παραμέτρου class_weights

7. Συζήτηση

7.1 Συνήθεις Εφαρμογές της Εξόρυξης Δεδομένων στην Υγειονομική Περίθαλψη

Η Εξόρυξη Δεδομένων (*Data Mining*) γενικώς έχει πληθώρα εφαρμογών στην Υγειονομική Περίθαλψη. Βοηθά στην δημιουργία μοντέλων περίθαλψης με επίκεντρο τον ασθενή. Οι νέες γνώσεις βοηθούν την δόμηση και καθογήση της εξατομικευμένης υγειονομικής περίθαλψης (*personalised healthcare*) (*με κατάλληλη διαχείριση ασθενών και δημιουργία σχεδίων για την ευεξία τους*) (Chawla and Davis, 2013).

Οι εφαρμογές μπορεί να προκύπτουν είτε άμεσα ή και έμμεσα. Ένας έμμεσος τρόπος είναι η χρήση της εξόρυξης δεδομένων για πρόγνωση ιδιοτήτων των πρωτεϊνών πχ. ενεργά κέντρα, σημεία τροποποιήσης, κυτταρικό εντοπισμό, σταθερότητα, σφαιρικότητα, σχήμα, αυτοτελείς δομικές-λειτουργικές περιοχές (*domains*), δευτεροταγή δομή, και αλληλεπιδράσεις (Tzanis et al., 2005). Σε περίπτωση που αυτές οι πρωτείνες σχετίζονται με την βιολογική αιτιολογία κάποιας ασθένειας ή αποτελούν φαρμακευτικό στόχο, τότε μέσω της νέας γνώσης που αποκτάται κατορθώνουμε να καταννοήσουμε καλύτερα τον βιολογικό μηχανισμό που προκαλεί την αντίστοιχη παθολογική κατάσταση και να προσανατολίσουμε την έρευνα για νέα φάρμακα προκειμένου να ανακαλύψουμε αποτελεσματικότερη θεραπεία ή/και τρόπους πρόληψης.

Άλλος έμμεσος τρόπος είναι η πρόγνωση μέσω εξόρυξης δεδομένων των θέσεων επί της αλληλουχίας των νουκλεϊκών οξέων (*DNA, RNA*) που έχουν συγκεκριμένη λειτουργία, λ.χ. εύρεση του κωδικονίου έναρξης ενός mRNA ή των θέσεων συρραφής (*splice sites*) ή και των θέσεων αναγνώρισης από ορισμένα miRNAs. Όταν η γνώση που αποκτούμε αφορά θέσεις του γονιδιώματος που έχουν συσχετισθεί με ασθένειες, τότε έχουμε επίσης έμμεσες εφαρμογές και στην Υγειονομική Περίθαλψη.

Η εξόρυξη δεδομένων (*data mining*) χρησιμοποιείται και άμεσα, για τον καθορισμό υποομάδων ασθενών που πάσχουν από σύνδρομα ή ασθένειες με μεγάλη ετερογένεια, με διαφορετικές πιθανές αιτίες, διαφορετικά συμπώματα ή/και απόκριση σε θεραπευτικές αγωγές. Η δημιουργία προφίλ ασθενών μπορεί να γίνει μέσω εύρεσης συστάδων ασθενών (*clustering*) (Koh and Tan, 2005).

Και ο τομέας της διάγνωσης επιχειρείται να γίνει από αυτούς που κατ' εξοχήν δέχονται παρεμβάσεις βασιζόμενες στην γνώση που αποκτήθηκε μέσω εξόρυξης δεδομένων και Ανακάλυψης Γνώσεων από Βάσεις Δεδομένων (*KDD*). Με αυτόν τον τρόπο δύνανται να δωθούν συστάσεις για διάγνωση των νέων ασθενών, συμπληρωματικά προς την πείρα του ιατρικού προσωπικού και να αποφευχθούν λάθη που οφείλονται σε παράβλεψη λόγω ανθρώπινου παράγοντα (Chawla and Davis, 2013).

Μέσω της Εξόρυξης Δεδομένων (Data Mining) μπορεί να εκτιμηθεί και η πιθανότητα ανάπτυξης αντίστασης στην θεραπεία από τους ασθενείς (Chawla and Davis, 2013). Η πιθανότητα να παρουσιαστούν παρενέργειες ή/και συννοσηρότητα είναι μία παράμετρος που το ιατρικό προσωπικό θα επιθυμούσε να γνωρίζει και είναι επίσης δυνατό να εκτιμηθεί μέσω εξόρυξης δεδομένων (*data mining*) (Chawla and Davis, 2013).

Η επαναστόχευση φαρμάκων (*drug repurposing*), μπορεί να βοηθηθεί από την εξόρυξη δεδομένων, για αντιστοίχιση των κατάλληλων μορίων με πιθανολογούμενη

φαρμακευτική δράση με ορισμένες ομάδες ασθενών - της ίδιας ή διαφορετικής από την αρχική στοχευόμενη ασθένεια – οι οποίες διαθέτουν το βιολογικό υπόβαθρο για να επωφεληθούν από αυτά χωρίς να κινδυνεύσουν. Έτσι δύναται να ξεκινήσει έρευνα για το αν πράγματι αποτελούν ασφαλή και στοχευμένη θεραπεία του νέου στόχου.

Η εύρεση ομοιοτήτων μεταξύ μη διαγνωσμένων ασθενών και γνωστών περιπτώσεων ασθενών με σπάνια νοσήματα (Chawla and Davis, 2013) μπορεί να γίνει μέσω εξόρυξης δεδομένων (*data mining*), να επιταχύνει την διάγνωση μιας σπάνιας νόσου και να διευκολύνει την μετέπειτα θεραπευτική παρακολούθηση.

Οι εφαρμογές της Εξόρυξης Δεδομένων (*data mining*) στην Υγειονομική περίθαλψη επεκτείνονται και στο διοικητικό κομμάτι της υγείας, κάτι που δεν σχετίζεται με την παρούσα εργασία. Προϋπόθεση βέβαια είναι να μην γίνονται εκπτώσεις στην ποιότητα της περίθαλψης και να μην μπαίνουν περιορισμοί στην πρόσβαση των περιθαλπωμένων σε αντίστοιχες υπηρεσίες. Σημαντικό επίσης είναι οι τεχνικές της Εξόρυξης Δεδομένων (*data mining*) να χρησιμοποιούνται με στόχο την βελτίωση της κατάστασης υγείας των ατόμων. Η κατάχρηση των τεχνικών της εξόρυξης δεδομένων για την μείωση της περίθαλψης ακόμη και σε περιπτώσεις που αυτό παραβαίνει όρους ιατρικής δεοντολογίας ή βιοηθικής, έχει συνέπειες καταστροφικές για τους δυνητικούς αποδέκτες της ιατροφαρμακευτικής περίθαλψης και ως εκ τούτου, δεν είναι θεμιτή ή αποδεκτή.

Μέσω εξόρυξης δεδομένων (*data mining*) γίνεται να πετύχουμε εντοπισμό χρόνιων ασθενών και ασθενών υψηλού κινδύνου προκειμένου να γίνει έγκαιρη παρέμβαση και εκπαίδευση των ασθενών για καλύτερη διαχείριση της υγείας τους (Koh and Tan, 2005). Τέτοιου είδους γνώση που προκύπτει μέσω τεχνικών εξόρυξης δεδομένων (*data mining*), μπορεί να ενισχύσει την προληπτική ιατρική, την ενεργή διαχείριση ασθενειών, να ενδυναμώσει τους ασθενείς, να ωθήσει σε αλλαγές στον τρόπο ζωής τους που μπορούν να συμβάλλουν στην πρόληψη κακής εξέλιξης της υγείας τους και έτσι να μείωσει τα ποσοστά επανεισαγωγών (*re-admission*) σε δομές υγείας (Chawla and Davis, 2013). Κάτι τέτοιο βέβαια γίνεται μόνο εντός των πλαισίων της διαρκούς πρόσβασής τους σε υπηρεσίες υγείας ακόμη και αν για οποιοδήποτε λόγο η πρόληψη αποτύχει ή δεν δυνηθεί να υπάρξει. Επίσης, στα πιο πάνω πλαίσια – και με καλή πρόληψη - μειώνεται η διάρκεια εισαγωγών, αποφεύγονται οι επιπλοκές, υιοθετούνται οι καλύτερες θεραπευτικές πορείες, βελτιώνεται η έκβαση των ασθενών και πληροφορείται καλύτερα το ιατρικό προσωπικό (Koh and Tan, 2005).

Πέραν από τα πιο πάνω υπάρχουν και εφαρμογές πιο απρόσμενες όπως λχ. ο εντοπισμός επιθέσεων βιο-τρομοκρατών, ο έλεγχος των ενδονοσοκομειακών μολύνσεων/λοιμώξεων αλλά και πιο επίκαιρες, όπως η εκκίνηση συναγερμού στην έναρξη επιδημιών με βάση συστήματα που βασίζονται σε μελέτη συμπτωμάτων του πληθυσμού των ασθενών, αλλά και η χαρτογράφηση της πορείας μιας επιδημίας (Koh and Tan, 2005).

Η Ιατρική Ακριβείας (*Precision Medicine*) θα είναι από τους κερδισμένους τομείς σε περίπτωση της χρήσης των τεχνικών εξόρυξης δεδομένων (*data mining*) για αποσαφήνιση της ετερογένειας του πληθυσμού των ασθενών που έχουν διαγνωστεί από την ίδια νόσο (Baytas et al., 2017). Θα βρεθούμε πιο κοντά στην ανακάλυψη των μονοπατιών προόδου της κάθε υποομάδας ασθενών και στην δημιουργία θεραπευτικών σχημάτων εξειδικευμένων στην κατάσταση υγείας έκαστης, άρα και πιο κοντά στην Ιατρική Ακριβείας (Baytas et al., 2017). Η εξατομίκευση σε επίπεδο ατόμου είναι αδύνατη σύμφωνα με την εργασία των (Williams and Moore, 2015) ωστόσο με την χρήση αρκετά μεγάλου όγκου δεδομένων και τις κατάλληλες αναλύσεις των περιβαλλοντικών, γενετικών και επιγενετικών δεδομένων που

αφορούν ασθενείς εκτιμούν πως θα γίνει εφικτός ο ορισμός ικανά μικρών ομάδων ασθενών με μικρές αποκλίσεις εντός της κάθε ομάδας, ώστε να αυξηθεί σημαντικά η ακρίβεια στην Ιατρική.

Αλλά και η κοινή Ιατρική (*average medicine*) θα επωφεληθεί από την εξόρυξης δεδομένων (*data mining*). Χρειάζεται όμως πιο ενδελεχής ενασχόληση με αυτό. Πολύ συχνά αναλύσεις δεδομένων ως τώρα είχαν βασιστεί στην σύμβαση της ύπαρξης της μέσης αλληλομορφικής επίδρασης (average allelic effect) που συναντάται στην ποσοτική γενετική, την διερεύνηση δηλαδή της ύπαρξης συσχέτισης μεταξύ της κατοχής ενός αλληλομόρφου έναντι ενός άλλου (σε ένα συγκεκριμένο γονιδιακό τόπο) και των αντίστοιχων διαφορών στον φαινότυπο (συνήθως γύρω από το δίπολο παθολογικό-φυσιολογικό) που παρουσίαζε έκαστη περίπτωση (Williams and Moore, 2015). Μια τέτοια θεώρηση υπεραπλουστεύει το βιολογικό υπόβαθρο μιας νόσου που μπορεί λ.χ. να είναι πολυγονιδιακή ή να εμφανίζει φαινόμενα επίστασης, ή ποικίλους διεισδυτικότητας ή να έχει πολλαπλά “παθολογικά” και “φυσιολογικά” αλληλόμορφα, ακόμη και να συμμετέχουν επιγενετικές τροποποιήσεις στο τελικό αποτέλεσμα. Ακόμα και έτσι, εντός κάποιων πλαισίων εμπιστοσύνης είναι δυνατό η εξόρυξη δεδομένων να αποβεί χρήσιμη για την εξαγωγή γνώσης σε ορισμένες τέτοιες περιπτώσεις ερευνών. Το ιδανικό βέβαια είναι να μην αρκεστεί η έρευνα σε απλουστευτικές παραδοχές.

7.2 Εφαρμογές της παρούσας εργασίας στην Υγειονομική Περίθαλψη – Συζήτηση & Συμπεράσματα

Δεδομένου του ότι μιλάμε για πρόγνωση της μοιραίας έκβασης της ασθένειας του Ηπατοκυτταρικού καρκινώματος, μας ενδιαφέρει από την σκοπιά της Υγειονομικής Περίθαλψης να γνωρίζουμε ποιοί ασθενείς διατρέχουν μεγαλύτερο κίνδυνο να μην επιβιώσουν μετά το 1 έτος, ώστε να μην τους υποθεραπεύσουμε. Επίσης έχει σημασία να μην υπερθεραπεύσουμε (αν πρόκειται για θεραπείες μεγάλης τοξικότητας) ασθενείς που δεν κινδυνεύουν να αποβιώσουν στο 1 έτος. Για αυτό τον λόγο ανάλογα με το ζητούμενο κλινικά, χρειάζεται να συμβουλευθούμε και αντίστοιχη μέθοδο ταξινόμησης-προεπεξεργασίας δεδομένων από αυτές που αξιολογήσαμε, η οποία να μπορεί να ταξινομήσει εκ των προτέρων όσο πιο σωστά γίνεται ένα νέο άτομο που έχει την συγκεκριμένη ασθένεια σε μία από τις δύο κατηγορίες: του επιβιώσαντες μετά το 1 έτος και του μη επιβιώσαντες. Λόγω του μεγάλου πλήθους τους, οι καταλληλότερες κατά περίπτωση αναφέρονται στην Ενότητα *Σχολιασμός Αποτελεσμάτων*.

Στην παρούσα εργασία δοκιμάστηκαν διαφορετικές πορείες προεπεξεργασίας των δεδομένων, με ένα μη ισορροπημένο σύνολο δεδομένων, με πολλές ελλείπουσες τιμές. Στα σύνολα δεδομένων που προέκυψαν από τις αυτές τις προεπεξεργασίες κλήθηκαν να κάνουν ταξινόμηση διάφορες ταξινομήτριες μέθοδοι. Η απόδωσή τους αξιολογήθηκε κατά περίπτωση. Τα ραβδογράμματα που προέκυψαν κατά την αξιολόγηση της κάθε ταξινόμησης μπορούν να χρησιμεύσουν στην Υγειονομική Περίθαλψη ως βάση αναφοράς για επιλογή της καταλληλότερης πορείας προεπεξεργασίας και ταξινόμησης σε καθημερινές κλινικές αποφάσεις –ανάλογα με το κλινικά ζητούμενο όπως αναφέραμε και πιο πριν.

Συνεπώς με χρήση του παρόντος σύνολου δεδομένων και των αντίστοιχων μοντέλων ταξινόμησης που είναι κατάλληλα για την κάθε περίπτωση ασθενούς, μπορεί να ταξινομηθεί νέος ασθενής με Ηπατοκυτταρικό καρκίνωμα (HCC), για τον οποίο έχουν καταγραφεί τα ίδια χαρακτηριστικά με αυτά που υπάρχουν στο σετ και του οποίου η έκβαση είναι άγνωστη προς το παρόν.

Ένα πλεονέκτημα του συνόλου δεδομένων που χρησιμοποιήθηκε είναι ότι συμπεριλαμβάνει στις καταγραφές των ασθενών με Ηπατοκυτταρικό καρκίνωμα και τις τιμές που έχουν για πληθώρα παραγόντων που έχουν αποδεδειγμένη προγνωστική αξία αναφορικά με την έκβασή τους και όχι απλά και μόνο τους παράγοντες κινδύνου για την εμφάνιση Ηπατοκυτταρικού καρκινώματος. Πολλοί από αυτούς τους παράγοντες πρόσφατα ανακαλύφθηκε ότι έχουν σχέση και με την έκβαση των ασθενών, όπως φαίνεται και στην ανωτέρω ενότητα *Περιγραφή του Συνόλου δεδομένων*.

Η εμπιστοσύνη που θα δωθεί στο αποτέλεσμα της ταξινόμησης που έγινε από τις μεθόδους που δοκιμάσαμε όμως δεν θα πρέπει σε καμία περίπτωση να είναι τυφλή, αλλά να λειτουργεί βοηθητικά. Σε κάθε περίπτωση αναγκαία είναι η αναζήτηση καλύτερων τρόπων ταξινόμησης και προεπεξεργασίας, μιας και οι καλύτερες που είχαμε δοκιμάσει πραγματοποιούν καλή ταξινόμηση, αλλά όχι άριστη, όπως θα χρειαζόταν για να φτάσουν στο επίπεδο που είναι επιθυμητό ώστε να δίνουν ιδιαίτερα αξιόπιστη πληροφόρηση για την καθημερινή κλινική χρήση. Συγχρόνως υπάρχουν ενδείξεις ότι σε αρκετές από τις περιπτώσεις ταξινόμησης είναι πιθανό να έχει γίνει υπερπροσαρμογή στα δεδομένα, συνεπώς χρειάζεται περεταίρω αξιολόγηση.

Η πληροφορία που δίνει το αποτέλεσμα της ταξινόμησης με τις μεθόδους που δοκιμάσαμε (το εάν δηλαδή κάποιος ασθενής που διαγνώστηκε πρόσφατα με Ηπατοκυτταρικό καρκίνωμα κινδυνεύει να αποβιώσει στο ένα έτος ή όχι) μπορεί να

χρησιμοποιηθεί στην Υγειονομική Περίθαλψη συμπληρωματικά προς άλλες πληροφορίες από το θεράπων Υγεινομικό προσωπικό, για την λήψη των κατάλληλων θεραπευτικών επιλογών και πιθανή ανάληψη ρίσκων. Ένα παράδειγμα είναι η έγκαιρη προετοιμασία για την γραφειοκρατεία που απαιτεί η συμμετοχή σε πειραματικές θεραπείες, προκειμένου το άτομο που νοσεί να μπορέσει -αν το επιθυμήσει- να συμμετέχει σε περίπτωση αποτυχίας των συμβατικών πορειών θεραπείας- και να μην προβεί μοιραία μία γραφειοκρατική καθυστέρηση, την στιγμή που μπορούμε να προγνώσουμε σε ένα βαθμό –με τις πιο πάνω μεθοδους- αν αναμένεται να υπάρξει μικρό προσδόκιμο ζωής με τις υπάρχουσες θεραπευτικές αγωγές.

Σε περιπτώσεις που -από τις αναλύσεις που γίναν με την βοήθεια των μεθόδων που δοκιμάσαμε στην εργασία- αναμένεται μη επιβίωση των ασθενών μετά το ένα έτος, αναγκαία είναι η έγκαιρη προετοιμασία για την απόδωση επιπλέον φροντίδας (ανακουφιστική ή και θεραπευτική) ώστε να δημιουργηθούν οι συνθήκες για την βελτίωση της ποιότητας ζωής του ασθενή και ίσως και του προσδόκιμού του.

Αυτονόητο είναι ότι ακόμη και όσα άτομα δεν αναμένεται να έχουν μικρό προσδόκιμο χρειάζεται να λαμβάνουν την καλύτερη φροντίδα που μπορεί να προσφέρει η επιστήμη. Ωστόσο αν πρόκειται να γίνει λήψη πολύ τοξικής θεραπείας, τότε χρήσιμο θα ήταν να είναι γνωστό το αποτέλεσμα της ταξινόμησης έκαστου ασθενούς εκ των προτέρων, προκειμένου να προσμετρηθεί στην απόφαση. Δεν είναι επιθυμητή η υπερθεραπεία ασθενών που ούτως ή άλλως δεν διέτρεχαν μεγάλο κίνδυνο να αποβιώσουν, ειδικά όταν η ίδια η διαδικασία της θεραπευτικής αγωγής μπορεί να θέσει σε κίνδυνο την ζωή του ασθενούς.

Χρειάζεται επίσης προσοχή στην επιλογή της κατάλληλης μεθόδου ταξινόμησης και προεπεξεργασίας, ώστε να χρησιμοποιηθεί στην κλινική πράξη μία μέθοδος (από όσες δοκιμάσαμε ή και διαφορετική) που να δίνει αξιόπιστα αποτελέσματα, που δίνει δηλαδή όσο πιο έγκυρες ταξινομήσεις. Αν πχ. μια μέθοδος έχει μέγιστη ευαισθησία (sensitivity) και αρκετά μικρή ειδικότητα (specificity) μεροληπτεί και ταξινομεί σωστά μόνο την μία κατηγορία ασθενών (τους αποβιώσαντες). Κάτι τέτοιο σημαίνει ότι δεν χρησιμεύει καθόλου. Χρειάζεται να επιλεγεί μία μέθοδος που να συμβιβάζει την καλή ταξινόμηση και των δύο κατηγοριών ασθενών (επιβιώσαντες μετά το 1 έτος και μη). Αυτό είναι χρήσιμο ώστε να μπορέσουμε να είμαστε σίγουροι για το ποιοί ασθενείς μπορούν με βεβαιότητα να γνωρίζουν ότι θα ζήσουν μετά το διάστημα που ορίζεται από τους συλλογείς του Συνόλου δεδομένων και να ληφθούν οι κατάλληλες κλινικές αποφάσεις.

Επίσης, η εργασία χρησίμευσε και ως δοκιμή νέων μεθόδων προεπεξεργασίας, προκειμένου να εμπλουτιστούν οι γνώσεις που έχουμε όσον αφορά το θεωρητικό τμήμα της ανάλυσης, μιας και δοκιμάστηκε η απόδοση της συμπλήρωσης κενών με την συνάρτηση που προτάθηκε, κάτι που εξ’ όσων γνωρίζουμε είναι πρωτότυπο στην βιβλιογραφία. Η αξιολόγησή τους γίνεται στην ενότητα *Σχολιασμός Αποτελεσμάτων*.

Τέλος, με αφορμή την σύγκριση της απόδωσης διαφόρων μεθόδων προεπεξεργασίας και ταξινομικών μεθόδων σε ένα σύνολο δεδομένων που είχε χαρακτηριστικά κακή ποιότητα, παρόμοια με αυτή άλλων βιολογικών / κλινικών συνόλων δεδομένων συμβάλουμε -στο μέτρο των δυνατοτήτων μας- στην συζήτηση γύρω από την συνολική βελτίωση της απόδωσης για εξορύξεις δεδομένων που γίνονται σε βιολογικά / κλινικά δεδομένα, κάτι που επίσης έχει έμμεση εφαρμογή στην Υγειονομική Περίθαλψη. Η συμπερίληψη μετρικών αξιολόγησης των ταξινομήσεων που εκτιμούν συνολικά και σφαιρικά την ποιότητα μιας ταξινόμησης, κάτι που γίνεται σπάνια στην βιβλιογραφία με τέτοια πληρότητα, βοηθά πέραν των πιο πάνω και στην εύκολη μετέπειτα πραγματοποίηση συγκρίσεων με την παρούσα εξόρυξη δεδομένων και από άλλα μέλη της επιστημονικής κοινότητας, κάτι που χρειάζεται να συμβαίνει.

8. Βιβλιογραφία - Δικτυογραφία

Ada Hamosh, George E. Tiller, Cassandra L. Kniffin, Paul J. Converse, Victor A. McKusick, John A. Phillips, III, 2020. # 114550 HEPATOCELLULAR CARCINOMA / HCC / CANCER, HEPATOCELLULAR / LIVER CANCER / LIVER CELL CARCINOMA; LCC / HEPATOMA. OMIM Online Mendel. Inherit. Man® Online Cat. Hum. Genes Genet. Disord.

Adhoute, X. et al. (2016) 'Prognosis of advanced hepatocellular carcinoma: a new stratification of Barcelona Clinic Liver Cancer stage C: results from a French multicenter study.', European journal of gastroenterology & hepatology, 28(4), pp. 433–440. doi:10.1097/MEG.0000000000000558.

Anaconda Software Distribution, 2020. , Anaconda Documentation. Anaconda Inc.

Aniruddha Bhandari (2020) 'AUC-ROC Curve – The Star Performer!', 16 June. Available at: https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/.

Avinash Navlani, 2018. KNN Classification using Scikit-learn. DATACAMP.

Balayla, J., 2020. Prevalence threshold (ϕe) and the geometry of screening curves. PloS One 15, e0240215. https://doi.org/10.1371/journal.pone.0240215

Baliga, M.S. et al. (2013) 'Chapter 21 - Gastrointestinal and Hepatoprotective Effects of Ocimum sanctum L. Syn (Holy Basil or Tulsi): Validation of the Ethnomedicinal Observation', in Watson, R.R. and Preedy, V.R. (eds) Bioactive Food as Dietary Interventions for Liver and Gastrointestinal Disease. San Diego: Academic Press, pp. 325–335. doi:https://doi.org/10.1016/B978-0-12-397154-8.00039-7.

Baytas, I.M., Xiao, C., Zhang, X., Wang, F., Jain, A.K., Zhou, J., 2017. Patient Subtyping via Time-Aware LSTM Networks, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '17. Association for Computing Machinery, New York, NY, USA, pp. 65–74. https://doi.org/10.1145/3097983.3097997

Bosetti, C. et al. (2008) 'Trends in mortality from hepatocellular carcinoma in Europe, 1980-2004', Hepatology, 48(1), pp. 137–145. doi:10.1002/hep.22312.

Bosso, M. and Al-Mulla, F. (2020) 'Chapter 10 - RKIP & GSK3β: The interaction of two intracellular signaling network regulators and their role in cancer', in Bonavida, B. and Baritaki, S. (eds) Prognostic and Therapeutic Applications of RKIP in Cancer. Academic Press, pp. 147–173. doi:https://doi.org/10.1016/B978-0-12-819612-0.00010-9.

Brusic, V., Zeleznikow, J., 1999. Knowledge discovery and data mining in biological databases. Knowl. Eng. Rev. 14. https://doi.org/10.1017/S0269888999003069

Campbell, C. et al. (2021) 'Risk factors for the development of hepatocellular carcinoma (HCC) in chronic hepatitis B virus (HBV) infection: a systematic review and meta-analysis.', Journal of viral hepatitis, 28(3), pp. 493–507. doi:10.1111/jvh.13452.

Carr, B.I. et al. (2014) 'Association of abnormal plasma bilirubin with aggressive hepatocellular carcinoma phenotype.', Seminars in oncology, 41(2), pp. 252–258. doi:10.1053/j.seminoncol.2014.03.006.

Carr, B.I. and Guerra, V. (2017) 'Serum albumin levels in relation to tumor parameters in hepatocellular carcinoma patients.', The International journal of biological markers, 32(4), pp. e391–e396. doi:10.5301/ijbm.5000300.

Chan, A.W.H. et al. (2015) 'Albumin-to-Alkaline Phosphatase Ratio: A Novel Prognostic Index for Hepatocellular Carcinoma', Disease Markers. Edited by S. Theocharis, 2015, p. 564057. doi:10.1155/2015/564057.

Charlotte E. Costentin (2018) 'Alcohol-related liver cancer has worse prognosis', Healio, 28 March. Available at: https://www.healio.com/news/hematology-oncology/20180328/alcoholrelated-liver-cancer-has-worse-prognosis.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: Synthetic Minority over-Sampling Technique. J Artif Int Res 16, 321–357.

Chawla, N.V., Davis, D.A., 2013. Bringing Big Data to Personalized Healthcare: A Patient-Centered Framework. J. Gen. Intern. Med. 28, 660–665. https://doi.org/10.1007/s11606-013-2455-8

Cohen's kappa, n.d. . Wikipedia Free Encycl.

Colombo, M., Sangiovanni, A. and Lencioni, R. (2018) '49 - Benign Liver Tumors', in Sanyal, A.J. et al. (eds) Zakim and Boyer's Hepatology (Seventh Edition). Seventh Edition. Philadelphia: Elsevier, pp. 720-735.e4. doi:https://doi.org/10.1016/B978-0-323-37591-7.00049-5.

Costa, A., Santos, M., Soares, C., Henriques Abreu, P., 2020. Analysis of Imbalance Strategies Recommendation using a Meta-Learning Approach.

'EASL-EORTC clinical practice guidelines: management of hepatocellular carcinoma.' (2012) Journal of hepatology, 56(4), pp. 908–943. doi:10.1016/j.jhep.2011.12.001.

El-Zayadi, A.-R. (2006) 'Heavy smoking and liver', World journal of gastroenterology, 12(38), pp. 6098–6101. doi:10.3748/wjg.v12.i38.6098.

Evaluation of binary classifiers, n.d. . Wikipedia Free Encycl.

false-negative rate, 2009. . FREE Dict. FARLEX, Medical Dictionary.

Finkelmeier, F. et al. (2013) 'Single measurement of hemoglobin predicts outcome of HCC patients', Medical Oncology, 31(1), p. 806. doi:10.1007/s12032-013-0806-2.

F-score, n.d. Wikipedia Free Encycl.

Garnelo, M. et al. (2017) 'Interaction between tumour-infiltrating B cells and T cells controls the progression of hepatocellular carcinoma.', Gut, 66(2), pp. 342–351. doi:10.1136/gutjnl-2015-310814.

Gary Sieling, 2014. Decision Trees: "Gini" vs. "Entropy" criteria. Gary Sieling Softw. Archit. URL https://www.garysieling.com/blog/sklearn-gini-vs-entropy-criteria/

Harris, C.R., Millman, K.J., Walt, S.J. van der, Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., Kern, R., Picus, M., Hoyer, S., Kerkwijk, M.H. van, Brett, M., Haldane, A., Río, J.F. del, Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., Oliphant, T.E., 2020. Array programming with NumPy. Nature 585, 357–362. https://doi.org/10.1038/s41586-020-2649-2

Henriques Abreu, P., Santos, M., Henriques Abreu, M., Aveleira Andrade, B., Silva, D., 2016. Predicting Breast Cancer Recurrence Using Machine Learning Techniques: A Systematic Review. ACM Comput. Surv. 49, 1–40. https://doi.org/10.1145/2988544

'Hepatocellular carcinoma' (2021) Wikipedia, the free encyclopedia. Available at: https://en.wikipedia.org/wiki/Hepatocellular_carcinoma.

Hiraoka, A. et al. (2016) 'Usefulness of albumin-bilirubin grade for evaluation of prognosis of 2584 Japanese patients with hepatocellular carcinoma.', Journal of gastroenterology and hepatology, 31(5), pp. 1031–1036. doi:10.1111/jgh.13250.

Hohenwarter, M., Borcherds, M., Ancsin, G., Bencze, B., Blossier, M., Delobelle, A., Denizet, C., Éliás, J., Fekete, Á., Gál, L., Konečný, Z., Kovács, Z., Lizelfelner, S., Parisse, B., Sturr, G., 2013. GeoGebra 4.4 [WWW Document].

Holzinger, A., Jurisica, I., 2014. Knowledge Discovery and Data Mining in Biomedical Informatics: The Future Is in Integrative, Interactive Machine Learning Solutions. pp. 1–18. https://doi.org/10.1007/978-3-662-43968-5_1

Howell, J. et al. (2019) 'Identification of mutations in circulating cell-free tumour DNA as a biomarker in hepatocellular carcinoma', European Journal of Cancer, 116, pp. 56–66. doi:https://doi.org/10.1016/j.ejca.2019.04.014.

Hsu, C.-Y. et al. (2013) 'Ascites in patients with hepatocellular carcinoma: prevalence, associated factors, prognostic impact, and staging strategy', Hepatology International, 7(1), pp. 188–198. doi:10.1007/s12072-011-9338-z.

Hübscher, S.G. (2011) '30 - Alcohol-Induced Liver Disease', in Saxena, R. (ed.) Practical Hepatic Pathology: A Diagnostic Approach. Saint Louis: W.B. Saunders, pp. 417–433. doi:https://doi.org/10.1016/B978-0-443-06803-4.00030-7.

Hübscher, S.G. (2018) '24 - Alcohol-Induced Liver Disease', in Saxena, R. (ed.) Practical Hepatic Pathology: a Diagnostic Approach (Second Edition). Second Edition. Philadelphia: Elsevier (Pattern Recognition), pp. 371–390. doi:https://doi.org/10.1016/B978-0-323-42873-6.00024-X.

Hunter, J.D., 2007. Matplotlib: A 2D graphics environment. Comput. Sci. Eng. 9, 90–95.

'Informedness and Markedness Alternatives to Recall and Precision as Evaluation Measures' (2018) Venkatesh-Prasad Ranganath, 3 October. Available at: https://rvprasad.medium.com/informedness-and-markedness-20e3f54d63bc.

Janevska, D., Chaloska-Ivanova, V. and Janevski, V. (2015) 'Hepatocellular Carcinoma: Risk Factors, Diagnosis and Treatment', Open Access Macedonian Journal of Medical Sciences, 3. doi:10.3889/oamjms.2015.111.

Jothi, N., Rashid, N.A., Husain, W., 2015. Data Mining in Healthcare – A Review. Procedia Comput. Sci. 72, 306–313. https://doi.org/10.1016/j.procs.2015.12.145

Karageorgos, S.A. et al. (2017) 'Long-term change in incidence and risk factors of cirrhosis and hepatocellular carcinoma in Crete, Greece: a 25-year study.', Annals of gastroenterology, 30(3), pp. 357–363. doi:10.20524/aog.2017.0135.

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., Willing, C., 2016. Jupyter Notebooks – a publishing format for reproducible computational workflows, in: Loizides, F., Schmidt, B. (Eds.), Positioning and Power in Academic Publishing: Players, Agents and Agendas. IOS Press, pp. 87–90.

Koh, H.C., Tan, G., 2005. Data mining applications in healthcare. J. Healthc. Inf. Manag. JHIM 19, 64–72.

Kuo, Y.-H. et al. (2017) 'Albumin-Bilirubin grade predicts prognosis of HCC patients with sorafenib use.', Journal of gastroenterology and hepatology, 32(12), pp. 1975–1981. doi:10.1111/jgh.13783.

Larner, A., 2021. Assessing cognitive screeners with the critical success index. Prog. Neurol. Psychiatry 25, 33–37. https://doi.org/10.1002/pnp.719

Last, F., Douzas, G., Bação, F., 2017. Oversampling for Imbalanced Learning Based on K-Means and SMOTE.

Lau, H. et al. (1998) 'Long term prognosis after hepatectomy for hepatocellular carcinoma', Cancer, 83(11), pp. 2302–2311. doi:10.1002/(SICI)1097-0142(19981201)83:11<2302::AID-CNCR9>3.0.CO;2-1.

Lecture 13: Classification, 2016.

Lee, M.-H. et al. (2018) 'Human leukocyte antigen variants and risk of hepatocellular carcinoma modified by hepatitis C virus genotypes: A genome-wide association study', Hepatology, 67(2), pp. 651–661. doi:10.1002/hep.29531.

Lemaître, G., Nogueira, F., Aridas, C.K., 2017. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. J. Mach. Learn. Res. 18, 1–5.

li, X., 2013. Biological Data Mining and its Applications in Healthcare.

Li, Y. et al. (2020) 'Correlation Analysis Between Preoperative Serum Iron Level and Prognosis as Well as Recurrence of HCC After Radical Resection.', Cancer management and research, 12, pp. 31–41. doi:10.2147/CMAR.S227418.

Liao, M. et al. (2019) 'Prognostic Value of Gamma-Glutamyl Transpeptidase to Lymphocyte Count Ratio in Patients With Single Tumor Size ≤ 5 cm Hepatocellular Carcinoma After Radical Resection', Frontiers in Oncology, 9, p. 347. doi:10.3389/fonc.2019.00347.

Liu, C. et al. (2017) 'Neutrophil-to-lymphocyte and aspartate-to-alanine aminotransferase ratios predict hepatocellular carcinoma prognosis after transarterial embolization.', Medicine, 96(45), p. e8512. doi:10.1097/MD.0000000000008512.

Liu, H.-Q. et al. (2012) 'Leukocyte telomere length predicts overall survival in hepatocellular carcinoma treated with transarterial chemoembolization.', Carcinogenesis, 33(5), pp. 1040–1045. doi:10.1093/carcin/bgs098.

Liu, P.-H. et al. (2016) 'Prognosis of hepatocellular carcinoma: Assessment of eleven staging systems', Journal of Hepatology, 64(3), pp. 601–608. doi:10.1016/j.jhep.2015.10.029.

Liu, Z. et al. (2021) 'Investigation of Potential Molecular Biomarkers for Diagnosis and Prognosis of AFP-Negative HCC. Int J Gen Med. 2021;14:4369-4380', Int J Gen Med. 2021 [Preprint]. doi:https://doi.org/10.2147/IJGM.S323868.

Logistic regression, n.d. . Wikipedia Free Encycl.

Logistic Regression Optimization Parameters Explained, n.d. https://holypython.com/log-reg/logistic-regression-optimization-%20parameters/

Martínez-Chantar, M.L., Avila, M.A. and Lu, S.C. (2020) 'Hepatocellular Carcinoma: Updates in Pathogenesis, Detection and Treatment', Cancers, 12(10). doi:10.3390/cancers12102729.

Matt Cone, n.d. The Markdown Guide.

Matthews correlation coefficient, n.d. . Wikipedia Free Encycl.

Markedness, n.d. Wikipedia Free Encycl.

McKinney, W., others, 2010. Data structures for statistical computing in python, in: Proceedings of the 9th Python in Science Conference. Austin, TX, pp. 51–56.

Miriam Seoane Santos, Pedro Henriques Abreu, Armando Carvalho, Adélia Simão, 2017. HCC Survival Data Set. UCI Mach. Learn. Repos. Cent. Mach. Learn. Intell. Syst.

Mohammed, E.A., Far, B.H., Naugler, C., 2014. Applications of the MapReduce programming framework to clinical big data analysis: current landscape and future trends. BioData Min. 7, 22–22. https://doi.org/10.1186/1756-0381-7-22

Moriguchi, M., Furuta, M. and Itoh, Y. (2017) 'A Review of Non-operative Treatments for Hepatocellular Carcinoma with Advanced Portal Vein Tumor Thrombus', Journal of Clinical and Translational Hepatology. Xia & He Publishing, pp. 1–7. doi:10.14218/jcth.2016.00075.

Okabe, H. et al. (2014) 'Role of Leukocyte Cell-Derived Chemotaxin 2 as a Biomarker in Hepatocellular Carcinoma', PLOS ONE, 9(6), p. e98817. doi:10.1371/journal.pone.0098817.

Page, D., Craven, M., 2003. Biological applications of multi-relational data mining. SIGKDD Explor. 5, 69–79. https://doi.org/10.1145/959242.959250

Pavlovic, N. et al. (2019) 'Platelets as Key Factors in Hepatocellular Carcinoma.', Cancers, 11(7). doi:10.3390/cancers11071022.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., others, 2011. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Radial basis function kernel, n.d. . Wikipedia Free Encycl.

Rahul Agarwal, 2019. The Simple Math behind 3 Decision Tree Splitting criterions. towardsdatascience.com

R Core Team, 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

'Receiver operating characteristic' (2021) Wikipedia, the free encyclopedia. Available at: https://en.wikipedia.org/wiki/Receiver_operating_characteristic.

Santos, M., Henriques Abreu, P., García-Laencina, P., Simao, A., Carvalho, A., 2015. A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients. J. Biomed. Inform. 58, 49–59. https://doi.org/10.1016/j.jbi.2015.09.012

Santos, M., Soares, J., Henriques Abreu, P., Araujo, H., Santos, J., 2017. Influence of Data Distribution in Missing Data Imputation. https://doi.org/10.1007/978-3-319-59758-4_33

Scheuermann, R., Kong, M., Dahlke, C., Cai, J., Lee, J., Qian, Y., Squires, R., Dunn, P., Wiser, J., Hagler, H., Smith, B., Karp, D., 2009. Ontology-Based Knowledge Representation of Experiment Metadata in Biological Data Mining. Biol. Data Min. https://doi.org/10.1201/9781420086850.ch21

Sun, P., Chen, S. and Li, Y. (2020) 'The association between pretreatment serum alkaline phosphatase and prognosis in hepatocellular carcinoma: A meta-analysis.', Medicine, 99(11), p. e19438. doi:10.1097/MD.0000000000019438.

Sung, M.-K. and Bae, Y.-J. (2014) 'Chapter 13 - Iron, Oxidative Stress, and Cancer', in Preedy, V. (ed.) Cancer. San Diego: Academic Press, pp. 139–149. doi:https://doi.org/10.1016/B978-0-12-405205-5.00013-1.

Tharwat, A., 2021. Classification assessment methods. Appl. Comput. Inform. 17, 168–192. https://doi.org/10.1016/j.aci.2018.08.003

'Thorotrast' (2021) Wikipedia, the free encyclopedia. Available at: https://en.wikipedia.org/wiki/Thorotrast.

Tzanis, G., Berberidis, C., Vlahavas, I., 2005. Biological data mining. Encycl. Database Technol. Appl. 35–41. https://doi.org/10.4018/978-1-59140-560-3.ch007

Tzanis, G., Berberidis, C., Vlahavas, I., 2006. A Novel Data Mining Approach for the Accurate Prediction of Translation Initiation Sites. pp. 92–103. https://doi.org/10.1007/11946465_9

Uchino, K. et al. (2018) 'Serum levels of ferritin do not affect the prognosis of patients with hepatocellular carcinoma undergoing radiofrequency ablation.', PloS one, 13(7), p. e0200943. doi:10.1371/journal.pone.0200943.

U.S. Department of Veterans Affairs (INR (international normalized ratio) Hepatitis C for Patients) 'INR (international normalized ratio) Hepatitis C for Patients', U.S. Department of Veterans Affairs. Available at: https://www.hepatitis.va.gov/hcv/patient/diagnosis/labtests-INR.asp.

Van Rossum, G., Drake, F.L., 2009. Python 3 Reference Manual. CreateSpace, Scotts Valley, CA.

Vlahos, G.E., Ferratt, T.W., Knoepfle, G., 2004. The use of computer-based information systems by German managers to support decision making. Inf. Manage. 41, 763–779. https://doi.org/10.1016/j.im.2003.06.003

Vohra, I. et al. (2020) 'Prognostic Utility of Ferritin Transferrin Ratio in Hepatocellular Carcinoma', in.

Vohra, I. et al. (2021) 'Evaluation of Ferritin and Transferrin Ratio as a Prognostic Marker for Hepatocellular Carcinoma', Journal of Gastrointestinal Cancer, 52(1), pp. 201–206. doi:10.1007/s12029-020-00373-4.

Wang, P. et al. (2015) 'Impact of age on the prognosis after liver transplantation for patients with hepatocellular carcinoma: a single-center experience', OncoTargets and therapy, 8, pp. 3775–3781. doi:10.2147/OTT.S93939.

Waskom, M., Botvinnik, O., O'Kane, D., Hobson, P., Lukauskas, S., Gemperline, D.C., Augspurger, T., Halchenko, Y., Cole, J.B., Warmenhoven, J., Ruiter, J. de, Pye, C., Hoyer, S., Vanderplas, J., Villalba, S., Kunter, G., Quintero, E., Bachant, P., Martin, M., Meyer, K., Miles, A., Ram, Y., Yarkoni, T., Williams, M.L., Evans, C., Fitzgerald, C., Brian, Fonnesbeck, C., Lee, A., Qalieh, A., 2017. mwaskom/seaborn: v0.8.1 (September 2017). Zenodo. https://doi.org/10.5281/zenodo.883859

Williams, S., Moore, J., 2015. Lumping versus splitting: The need for biological data mining in precision medicine. BioData Min. 8, 16. https://doi.org/10.1186/s13040-015-0049-1

Wu, W.-C. et al. (2012) 'Prognostic significance of computed tomography scan-derived splenic volume in hepatocellular carcinoma treated with radiofrequency ablation.', Journal of clinical gastroenterology, 46(9), pp. 789–795. doi:10.1097/MCG.0b013e31825ceeb5.

Xu, W. and Yu, J. (2017) 'Chapter 20 - Obesity and Hepatocellular Carcinoma', in Muriel, P. (ed.) Liver Pathophysiology. Boston: Academic Press, pp. 267–277. doi:https://doi.org/10.1016/B978-0-12-804274-8.00020-5.

Yaghi, C. et al. (2006) 'Hepatocellular carcinoma in Lebanon: Etiology and prognostic factors associated with short-term survival', World journal of gastroenterology, 12(22), pp. 3575–3580. doi:10.3748/wjg.v12.i22.3575.

Yoneyama, K. et al. (2004) 'Prognostic index of cirrhotic patients with hepatic encephalopathy with and without hepatocellular carcinoma.', Digestive diseases and sciences, 49(7–8), pp. 1174–1180. doi:10.1023/b:ddas.0000037808.44897.8a.

Yoon, H.-J. et al. (2016) 'Mean corpuscular volume levels and all-cause and liver cancer mortality.', Clinical chemistry and laboratory medicine, 54(7), pp. 1247–1257. doi:10.1515/cclm-2015-0786.

Youden's J statistic, n.d. . Wikipedia Free Encycl.

Zhang, X.-F. et al. (2014) 'Impact of Cigarette Smoking on Outcome of Hepatocellular Carcinoma after Surgery in Patients with Hepatitis B', PLOS ONE, 9(1), p. e85077. doi:10.1371/journal.pone.0085077.

Zhao, L.-Y. et al. (2019) 'The Prognostic Value of aspartate aminotransferase to lymphocyte ratio and systemic immune-inflammation index for Overall Survival of Hepatocellular Carcinoma Patients Treated with palliative Treatments', J Cancer, 10, pp. 2299–2311. doi:10.7150/jca.30663.

Zhong, J.-H. et al. (2012) 'Epidermal growth factor gene polymorphism and risk of hepatocellular carcinoma: a meta-analysis', PLoS One, 7(3), p. e32159. doi:10.1371/journal.pone.0032159.

ZIMMERMANN, A. (2007) 'Chapter 69 - Tumors of the Liver—Pathologic Aspects', in Blumgart, L.H. et al. (eds) Surgery of the Liver, Biliary Tract and Pancreas (Fourth Edition). Fourth Edition. Philadelphia: W.B. Saunders, pp. 1085–1130. doi:https://doi.org/10.1016/B978-1-4160-3256-4.50082-X.

ΒΑΣΙΛΙΚΗ ΠΑΠΑΘΑΝΑΣΙΟΥ (2019) Ανάλυση Κλινικών Δεδομένων με χρήση των Αλγορίθμων Μηχανικής Μάθησης για την πρόβλεψη του καρκίνου. ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ. Available at: https://dione.lib.unipi.gr/xmlui/bitstream/handle/unipi/12166/Papathanasiou_1613.pdf?sequence=1&isAllowed=y.

'Δαναζόλη' (Γαληνός Οδηγός Φαρμάκων) Γαληνός Οδηγός Φαρμάκων. Available at: https://www.galinos.gr/web/drugs/main/substances/danazol/marketing.

'Ομαδοποίηση Κ-μέσων' (2019) Βικιπαίδεια, την ελεύθερη εγκυκλοπαίδεια. Available at: https://el.wikipedia.org/wiki/%CE%9F%CE%BC%CE%B1%CE%B4%CE%BF%CF%80%CE%BF

%CE%AF%CE%B7%CF%83%CE%B7_%CE%9A-%CE%BC%CE%AD%CF%83%CF%89%CE%BD.
Σκούρα Αγγελική (ceid.upatras) ‘Κατηγοριοποίηση - Εξόρυξη Δεδομένων και Αλγόριθμοι Μάθησης (3ο Φροντιστήριο)’.
‘Ταμοξιφαίνη’ (Γαληνός Οδηγός Φαρμάκων) Γαληνός Οδηγός Φαρμάκων. Available at: https://www.galinos.gr/web/drugs/main/substances/tamoxifen.

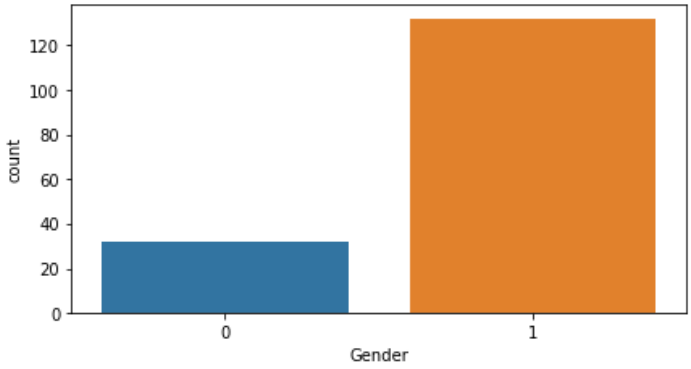
9. Ευχαριστίες

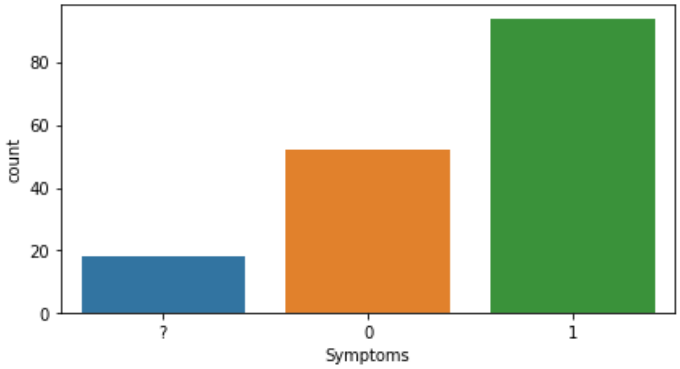
Ευχαριστώ τον Επιβλέποντα Αναπληρωτή Καθηγητή, κο Φιλιππάκη Μιχαήλ, για την καθοδήγηση, τις χρήσιμες υποδείξεις και την συνεργασία που υπήρξε κατά την εκπόνηση της παρούσας διπλωματικής εργασίας, παρά τις αντικειμενικά δύσκολες συνθήκες που προέκυψαν λόγω της τρέχουσας πανδημίας του SARS-Cov2.
Ευχαριστώ επίσης όλα τα μέλη της Τριμελούς Επιτροπής για τις χρήσιμες παρατηρήσεις τους για διορθώσεις πάνω στο κείμενο της εργασίας και ιδιαίτερα την Αναπληρώτρια Καθηγήτρια κα Βασιλική Οικονομίδου.
Ευχαριστώ την οικογένειά μου, τους συγκατοίκους μου και όλα τα κοντινά μου πρόσωπα για την ψυχολογική συνδρομή τους καθ' όλη τη διάρκεια εκπόνησης της εργασίας, την στήριξή τους και την μακρινή ή κοντινή αντίστοιχα παρέα τους.
Επίσης ευχαριστώ τους συναδέλφους και προϊσταμένους μου για την διευκόλυνση μου σχετικά με το εβδομαδιαίο πρόγραμμα απασχόλησής μου, κατά τον σχεδιασμό και την εκπόνηση των πειραμάτων, αλλά και κατά τη συγγραφή του τελικού κειμένου που βρίσκεται στα χέρια σας.

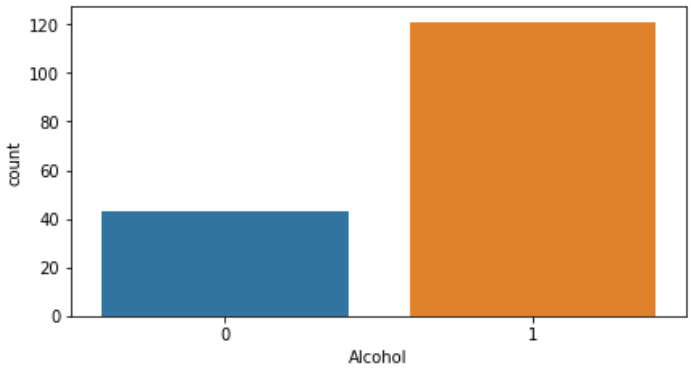
10. Παραρτήματα

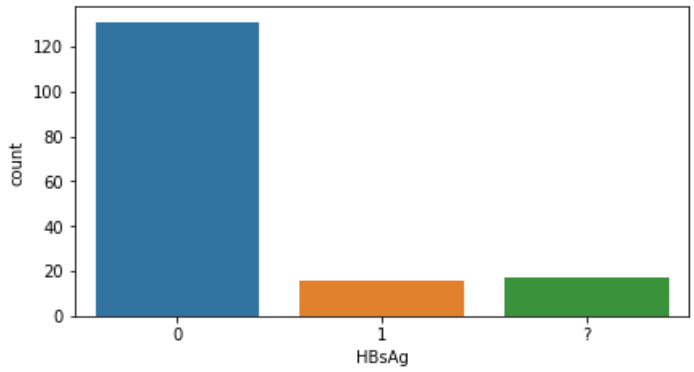
10.1 Παράρτημα Ι – Περιγραφική Ανάλυση (Descriptive Analysis)

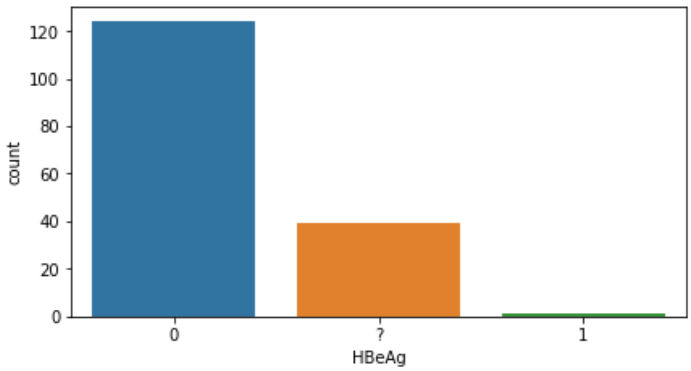
Ακολουθούν ραβδογράμματα για κάθε χαρακτηριστικό (λεγόμενο και γνώρισμα ή μεταβλητή) που περιέχεται στο σύνολο δεδομένων που εξετάσαμε. Στον οριζόντιο άξονα παρουσιάζονται οι δυνατές τιμές του χαρακτηριστικού. Στον κατακόρυφο άξονα παρουσιάζεται ο αριθμός των εμφανίσεων κάθε τιμής στο σύνολο δεδομένων.

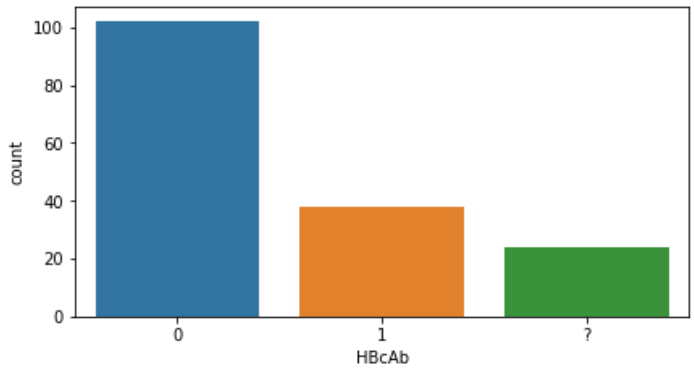


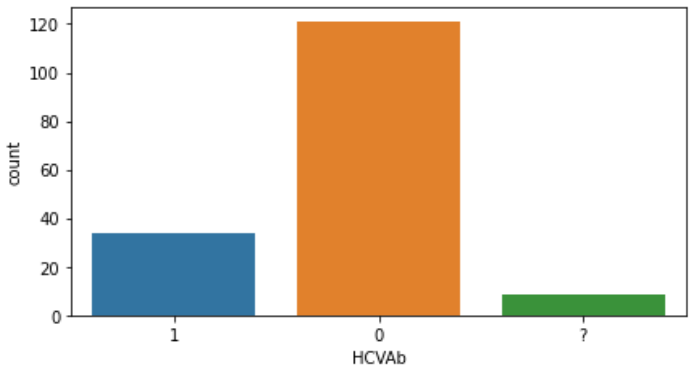


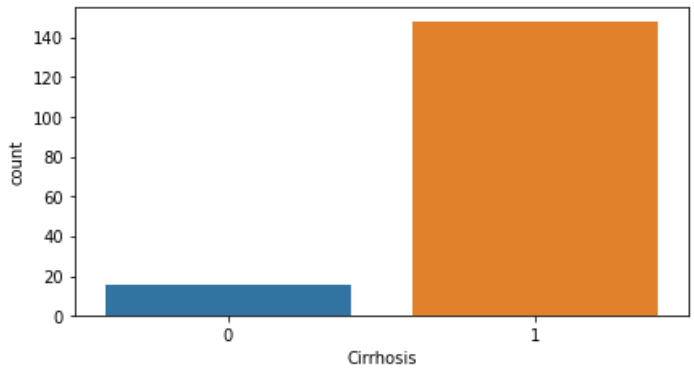


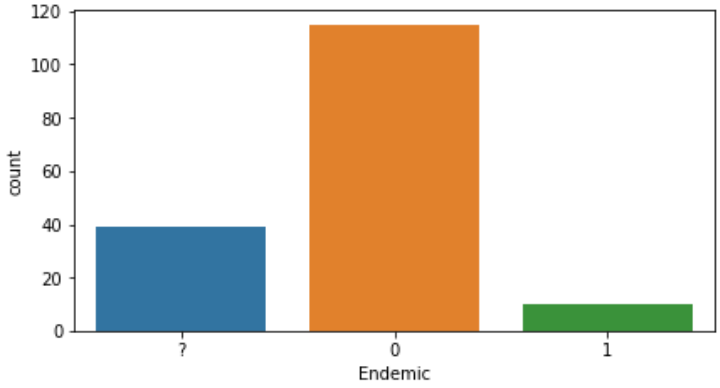


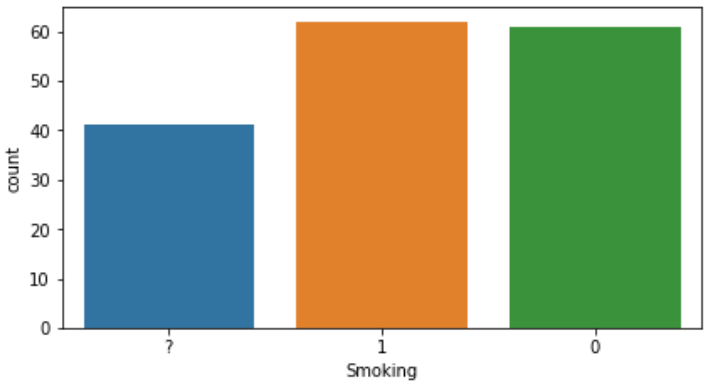


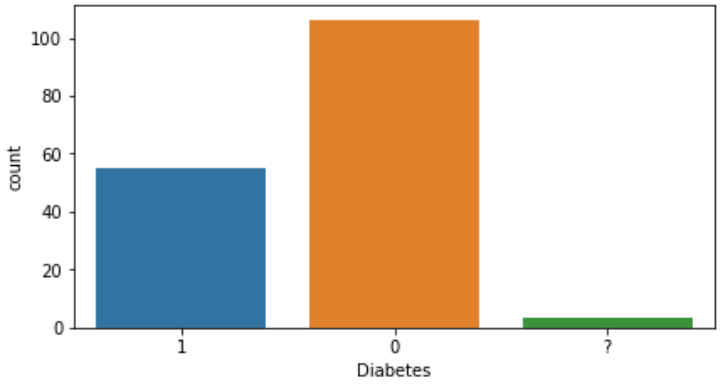


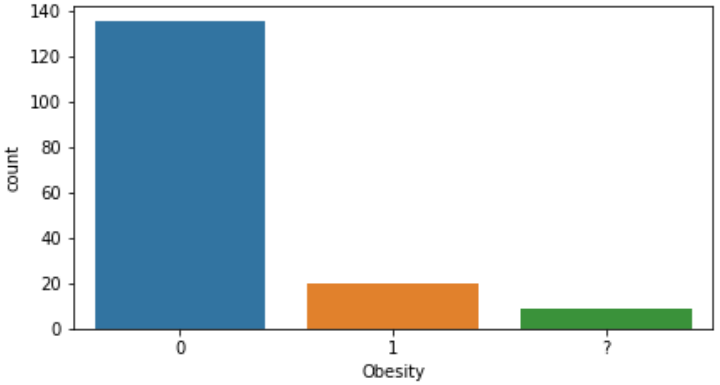


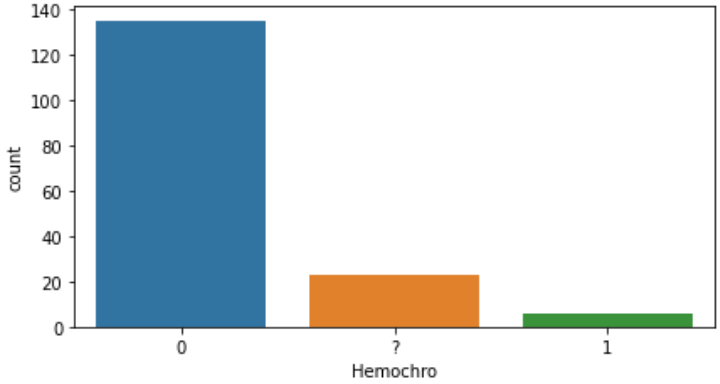


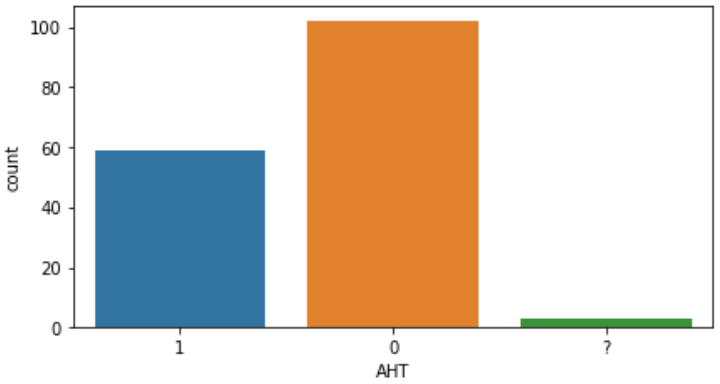

count
Endemic
?
0
1

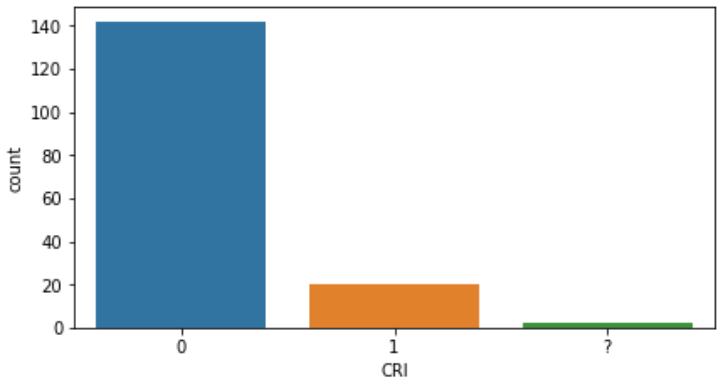

count
Smoking
?
1
0

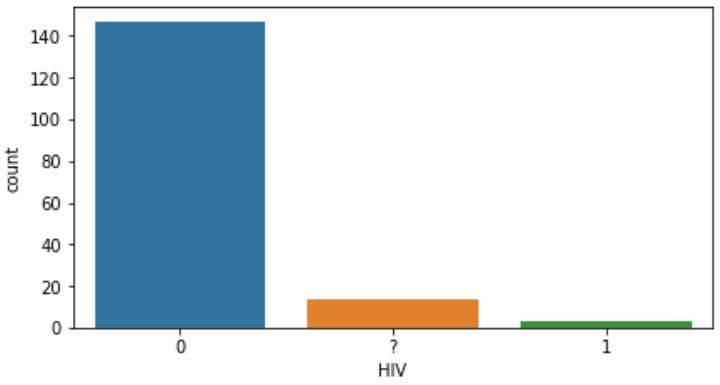

count
Diabetes
1
0
?

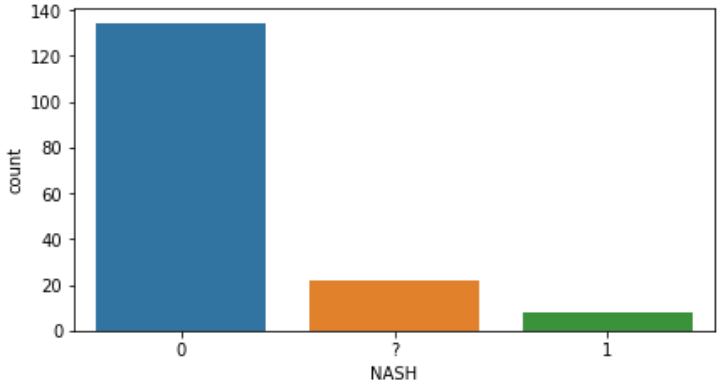

count
Obesity
0
1
?

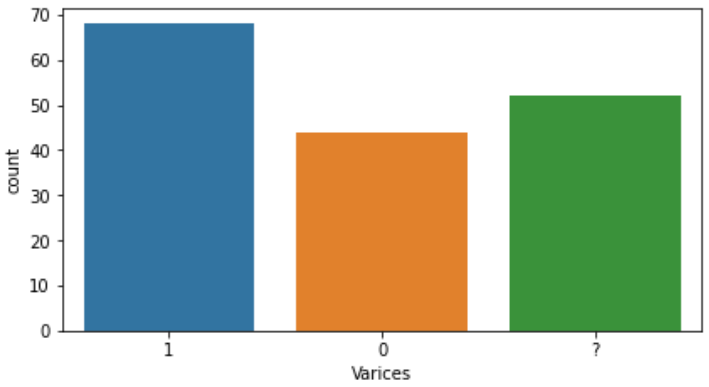

count
Hemochro
0
?
1

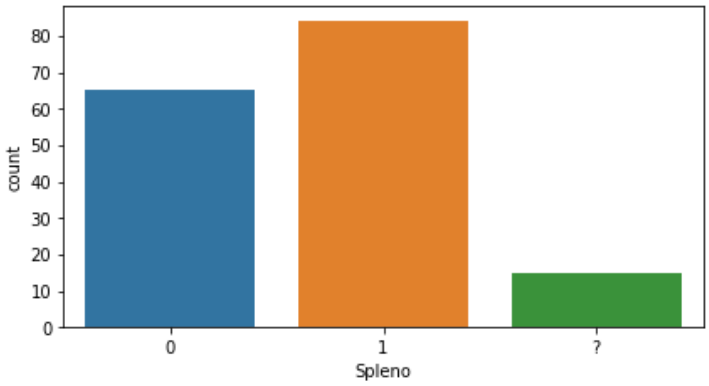

count
AHT
1
0
?

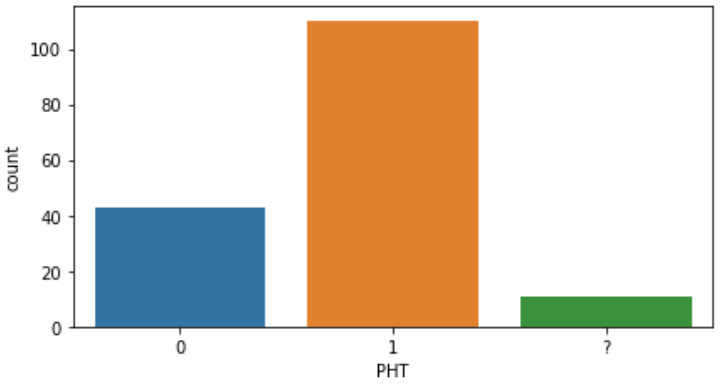

count
CRI
0
1
?

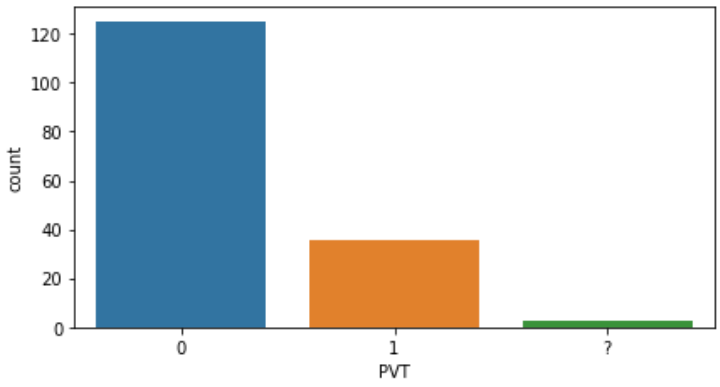

count
HIV
0
?
1

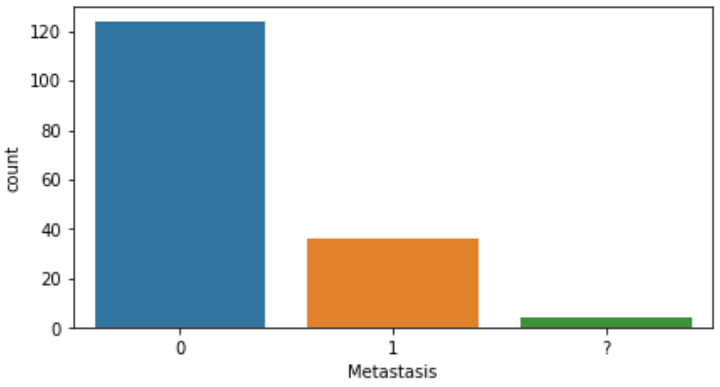

count
NASH
0
?
1

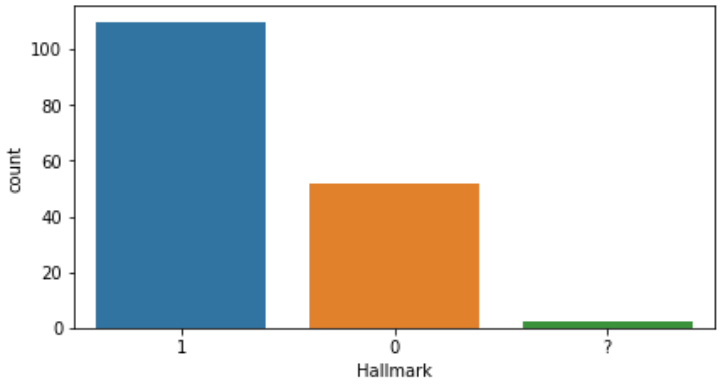

count
Varices
1
0
?

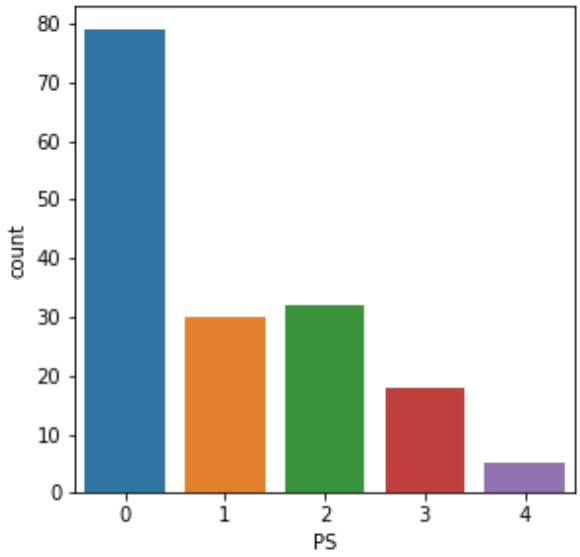

count
Spleno
0
1
?

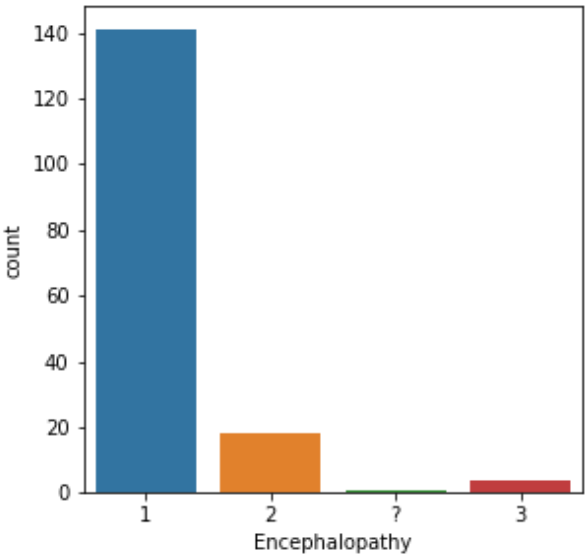

count
PHT
0
1
?

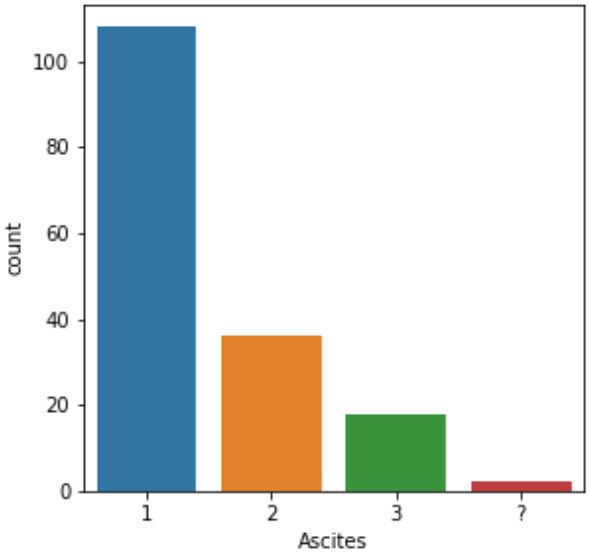

count
PVT
0
1
?

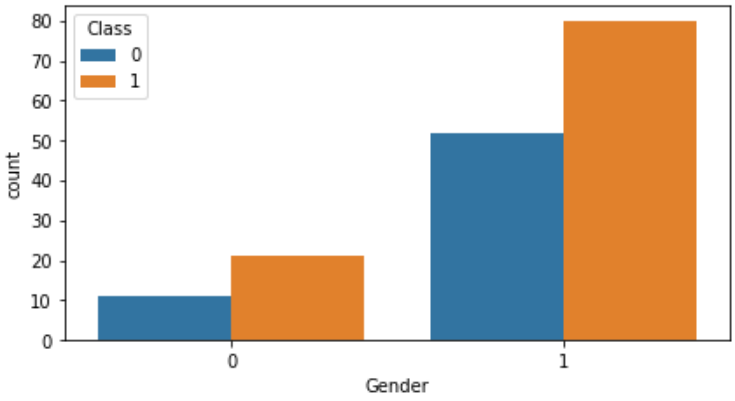

count
Metastasis
0
1
?

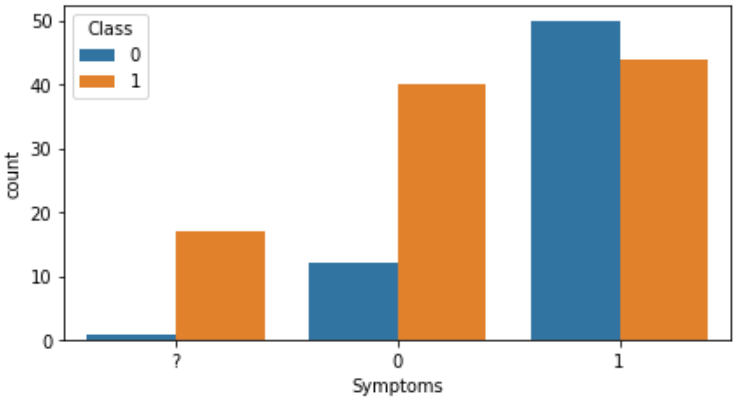

count
Hallmark
1
0
?

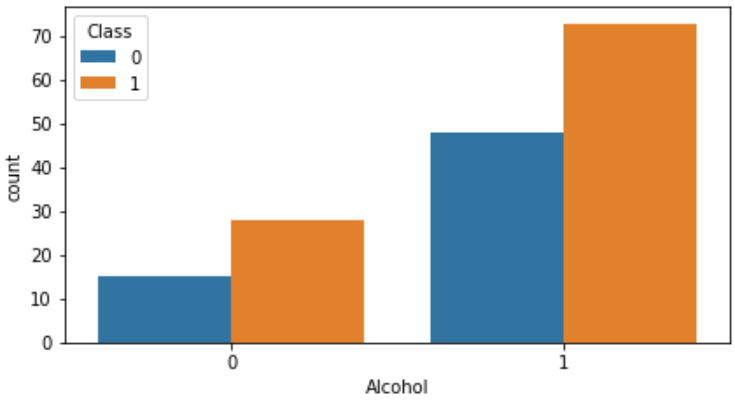

count
PS
0
1
2
3
4

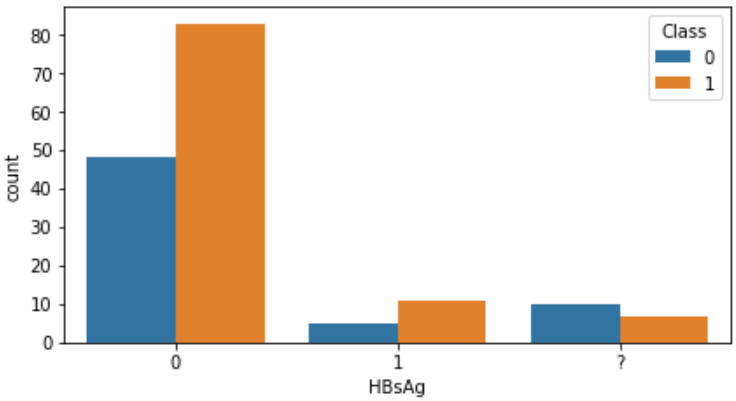

count
Encephalopathy
1
2
?
3

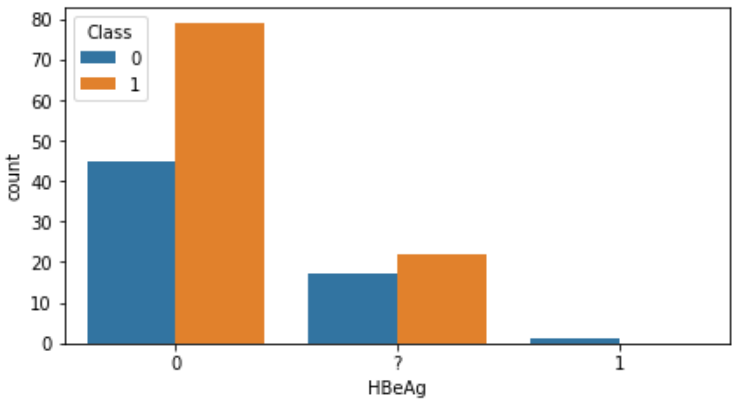

count
Ascites
1
2
3
?

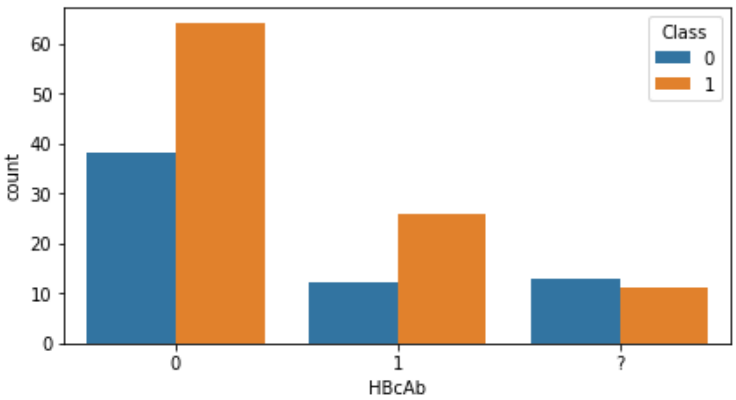

Class
0
1
count
Gender
0
1

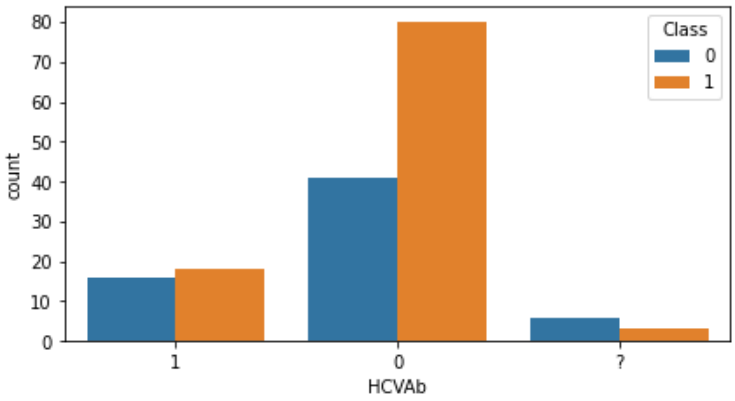

Class
0
1
count
Symptoms
?
0
1

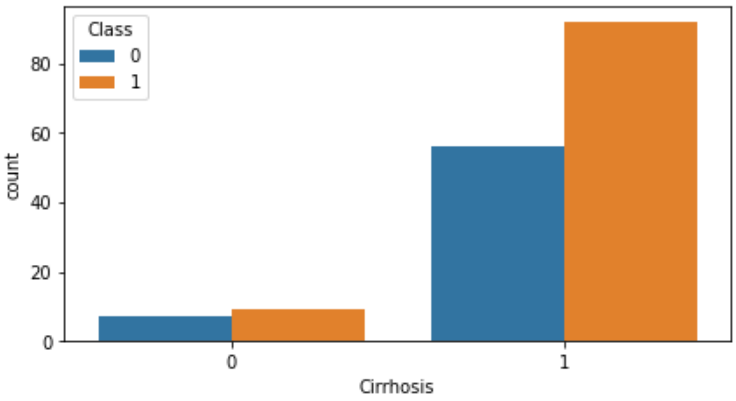

Class
0
1
count
Alcohol
0
1

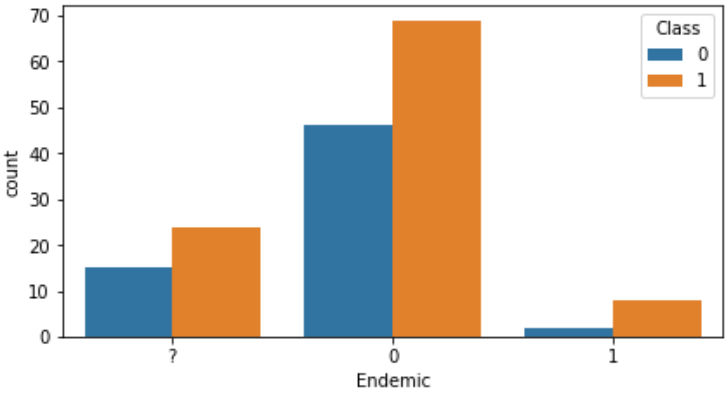

Class
0
1
count
HBsAg
0
1
?

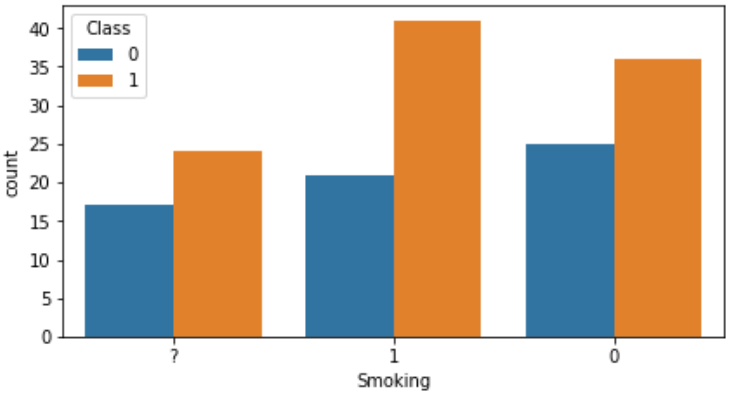

Class
0
1
count
HBeAg
0
?
1

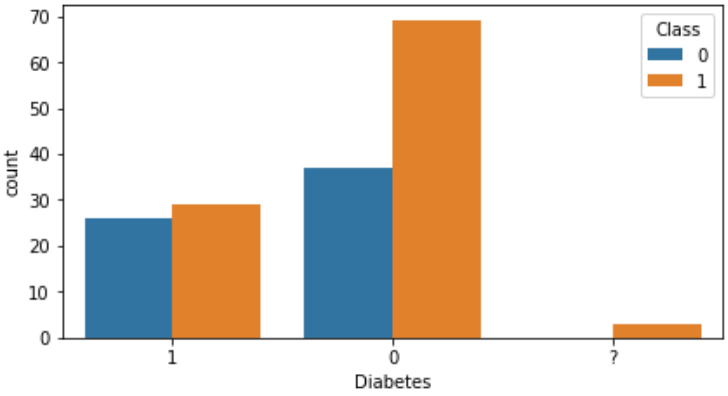

Class
0
1
count
HBcAb
0
1
?

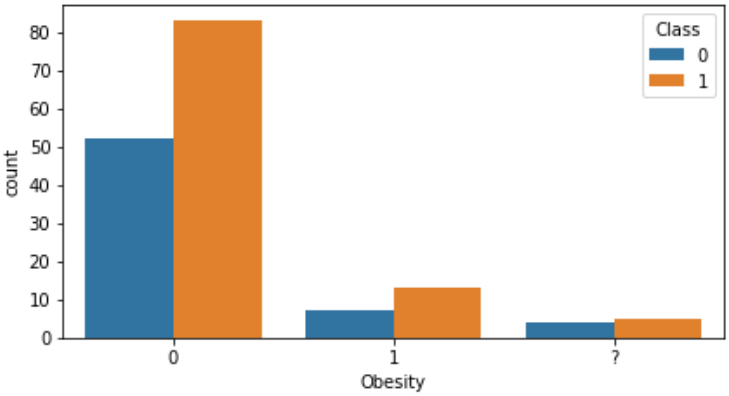

Class
0
1
count
HCVAb
1
0
?

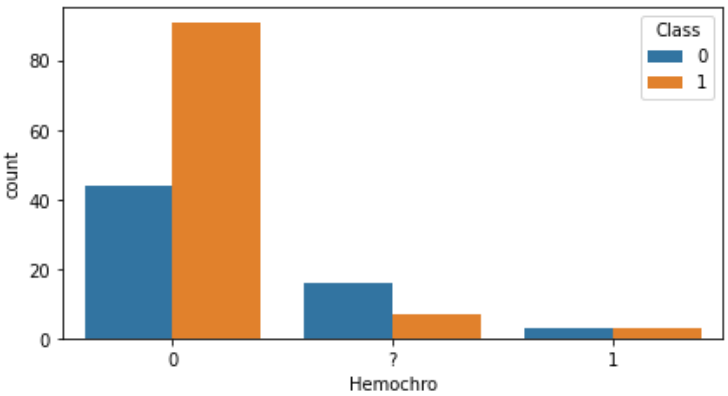

Class
0
1
count
Cirrhosis
0
1

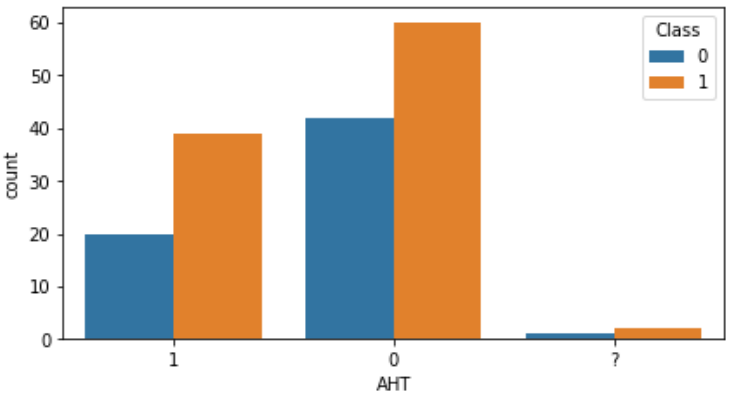

Class
0
1
count
Endemic
?
0
1

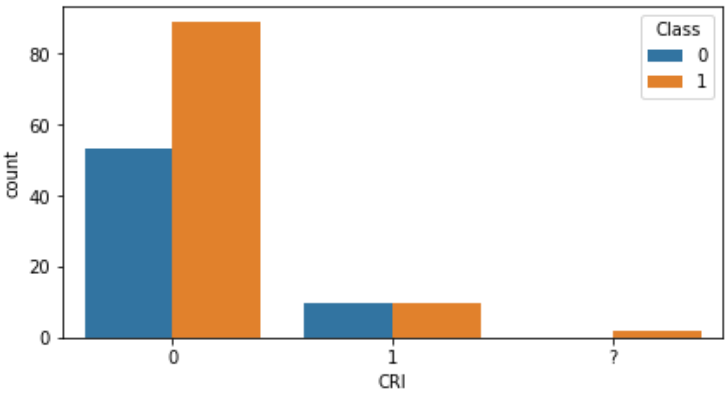

Class
0
1
count
Smoking
?
1
0

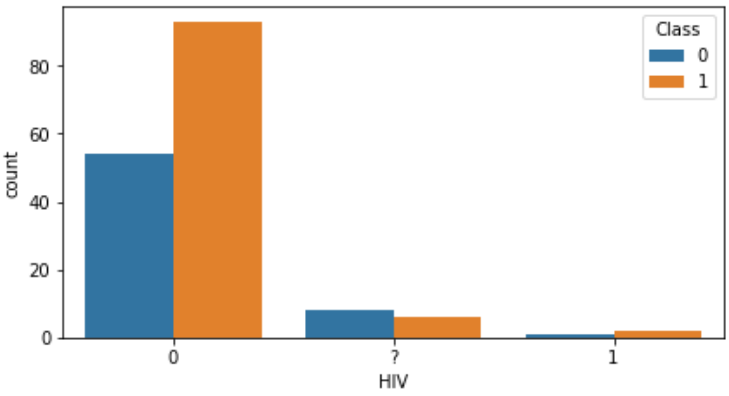

Class
0
1
count
Diabetes
1
0
?

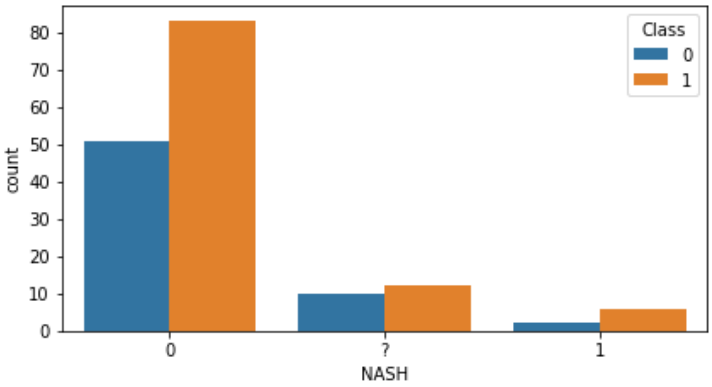

Class
0
1
count
Obesity
0
1
?

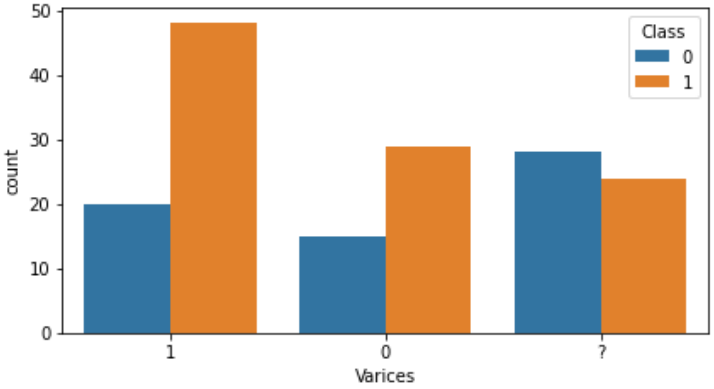

Class
0
1
count
Hemochro
0
?
1

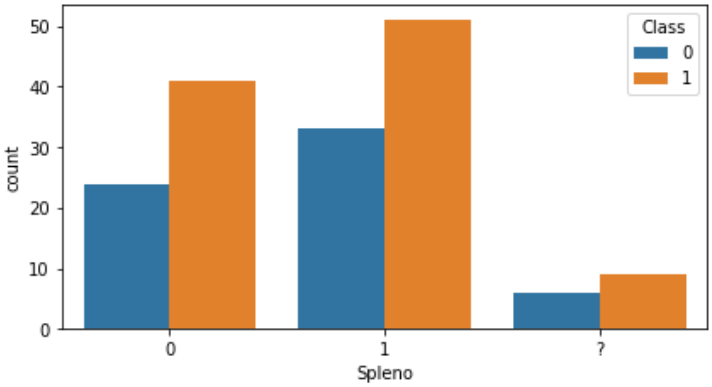

Class
0
1
count
AHT
1
0
?

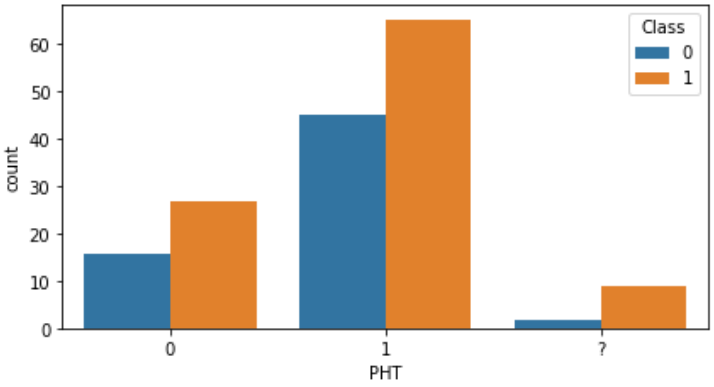

Class
0
1
count
CRI
0
1
?

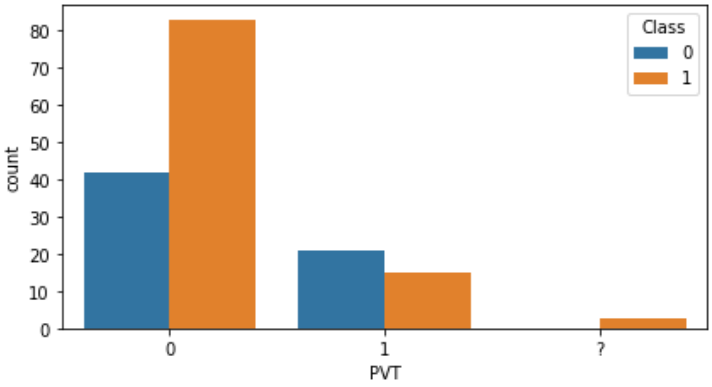

Class
0
1
count
HIV
0
?
1

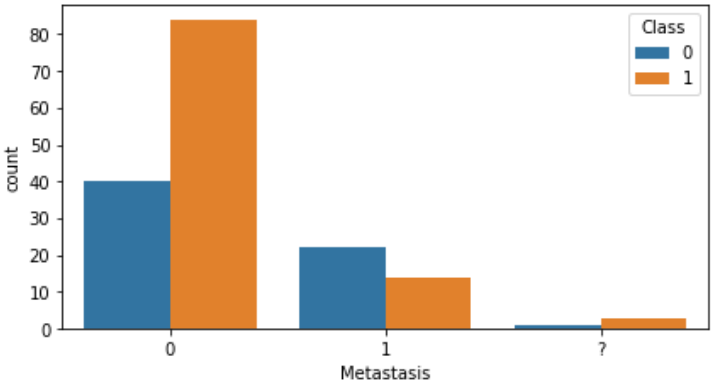

Class
0
1
count
NASH
0
?
1

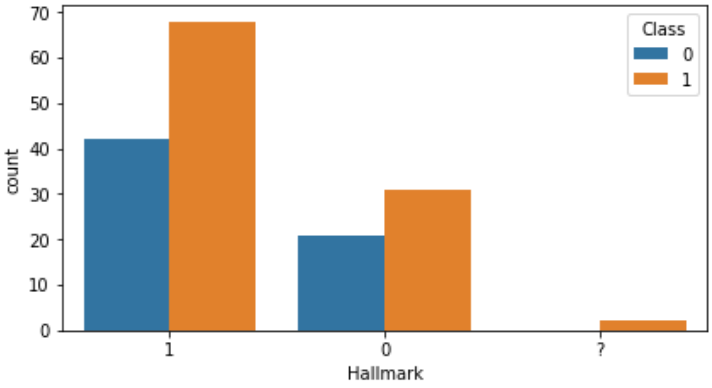

Class
0
1
count
Varices
1
0
?

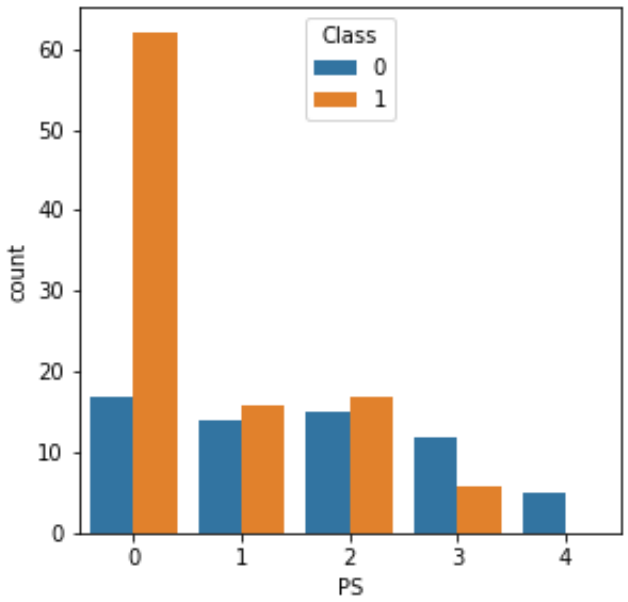

Class
0
1
count
Spleno
0
1
?

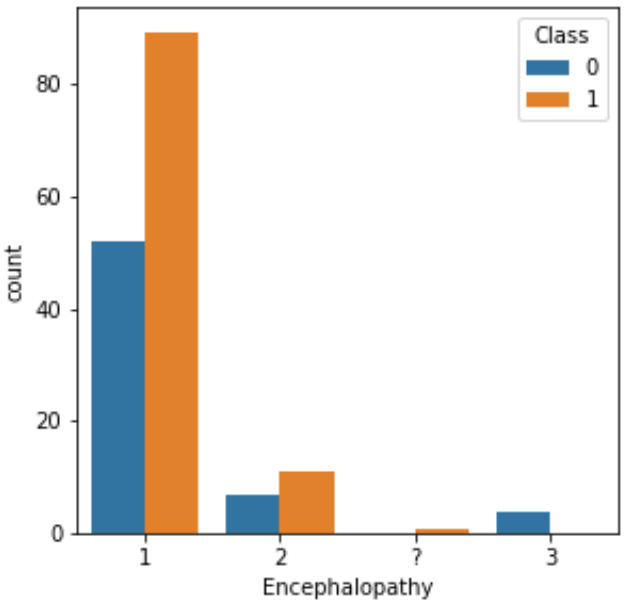

Class
0
1
count
PHT
0
1
?

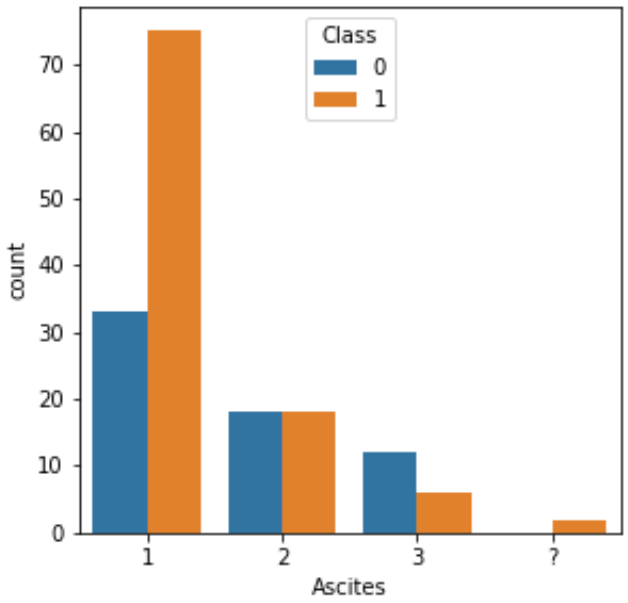

Class
0
1
count
PVT
0
1
?


Class
0
1
count
Metastasis
0
1
?


Class
0
1
count
Hallmark
1
0
?

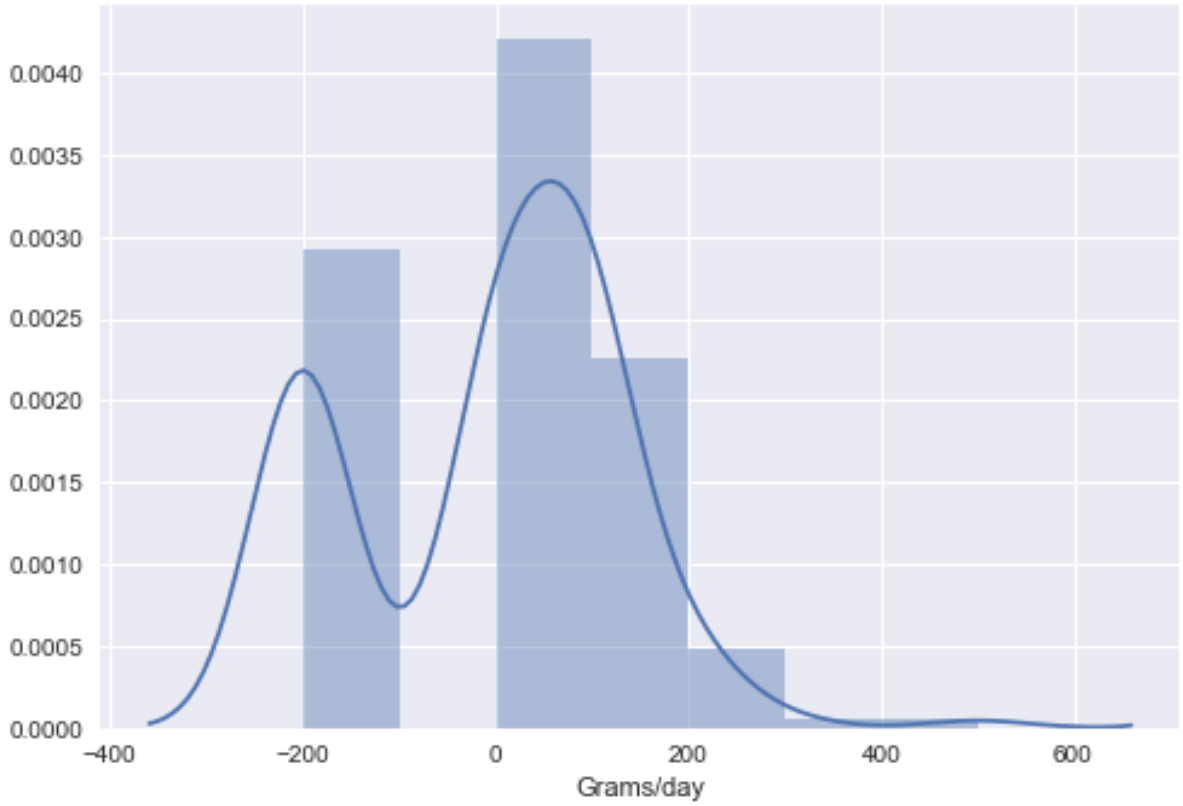




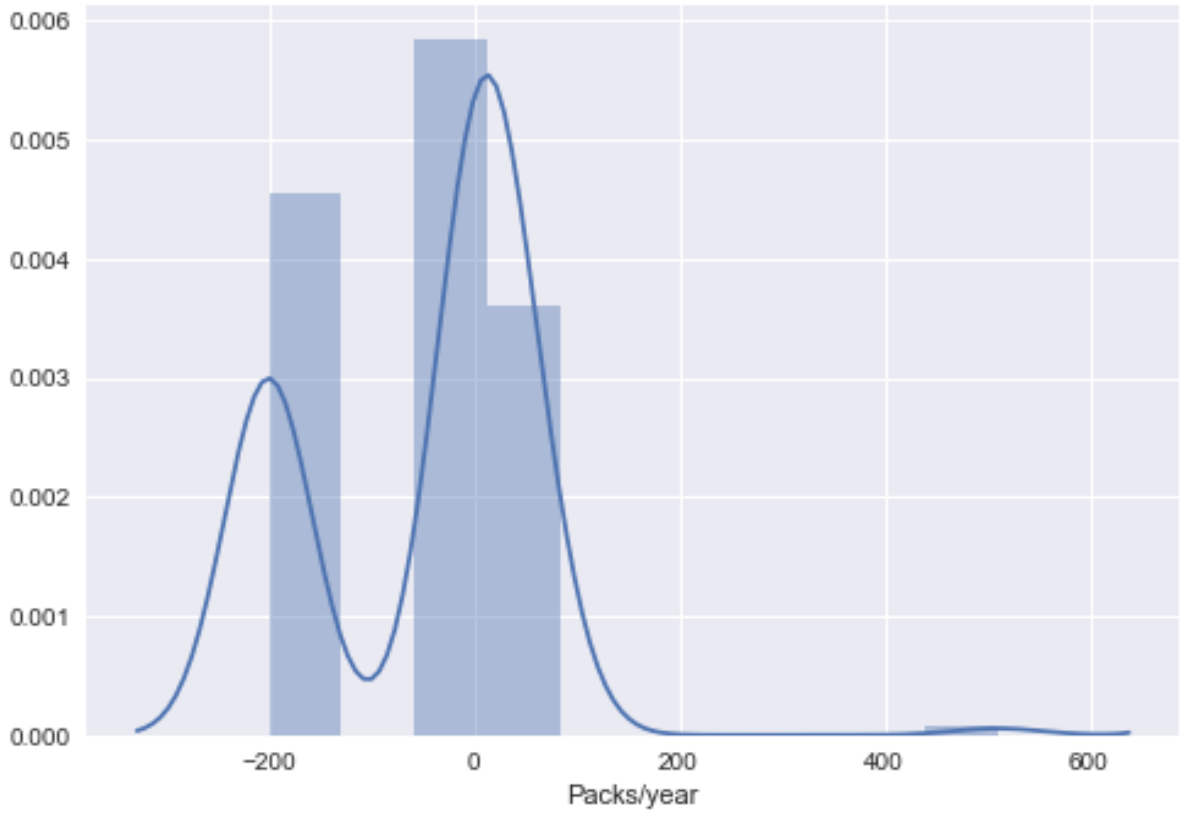


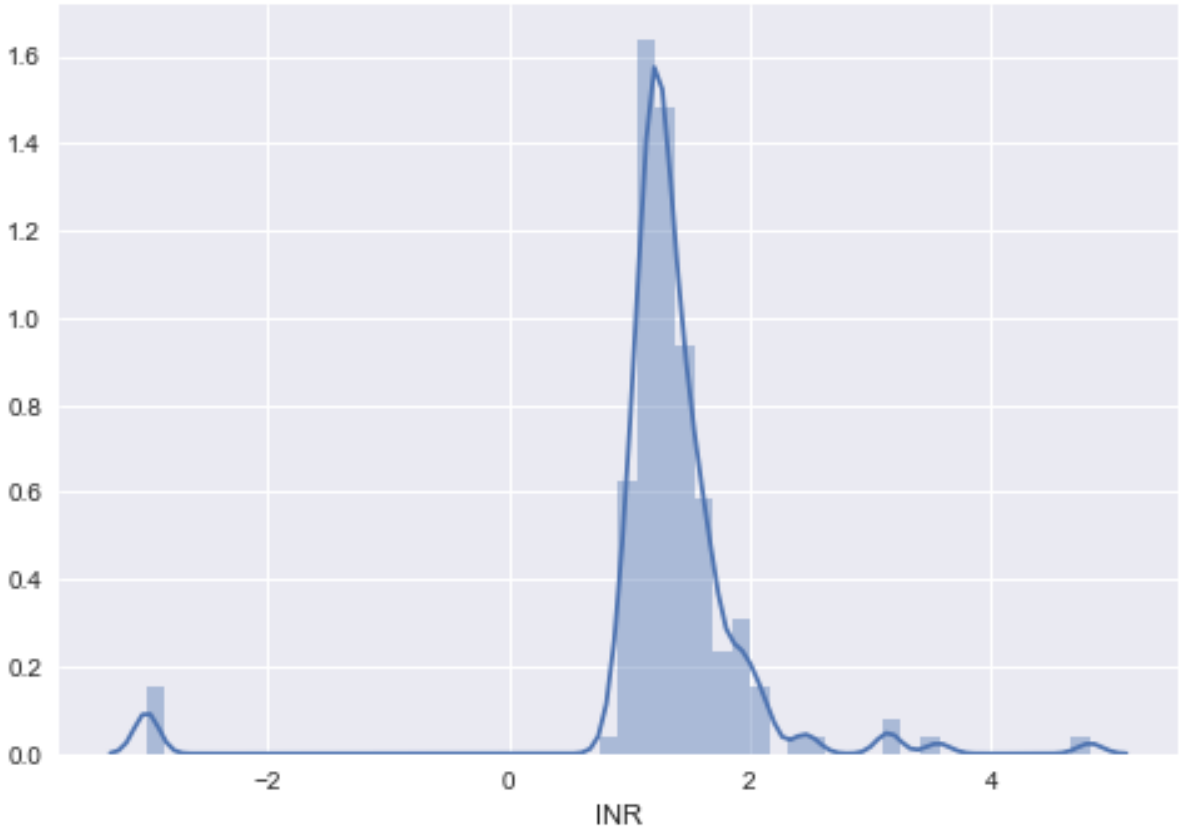
Η δωρεά του σετ πραγματοποιήθηκε από τα εξής επιστημονικά πρόσωπα της ακαδημαϊκής κοινότητας:

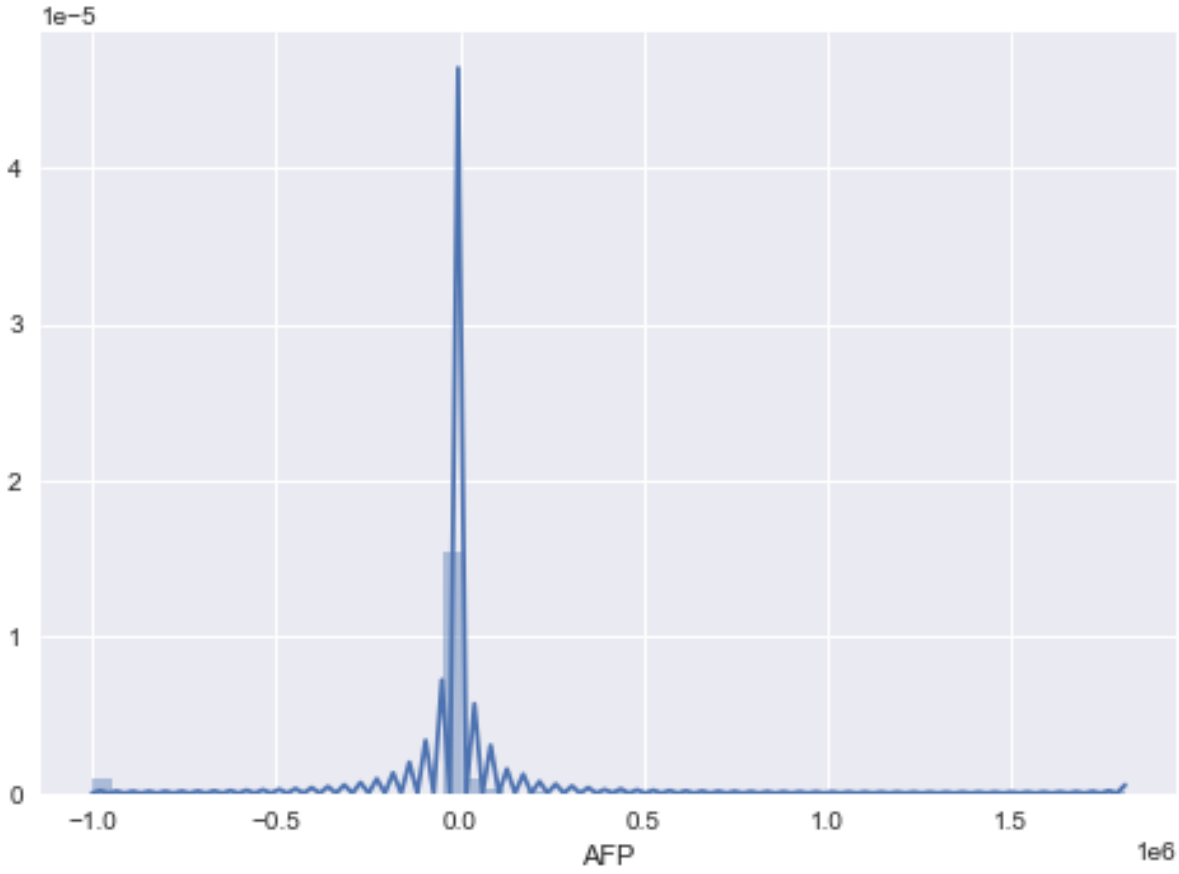
- Miriam Seoane Santos, Department of Informatics Engineering, Faculty of Sciences and Technology, University of Coimbra (miriams@student.dei.uc.pt),
- Pedro Henriques Abreu, Department of Informatics Engineering, Faculty of Sciences and Technology, University of Coimbra (pha@dei.uc.pt),
- Armando Carvalho, Internal Medicine Service, Hospital and University Centre of Coimbra (aspcarvalho@gmail.com),
- Adélia Simão, Internal Medicine Service, Hospital and University Centre of Coimbra (adeliasimao@gmail.com)

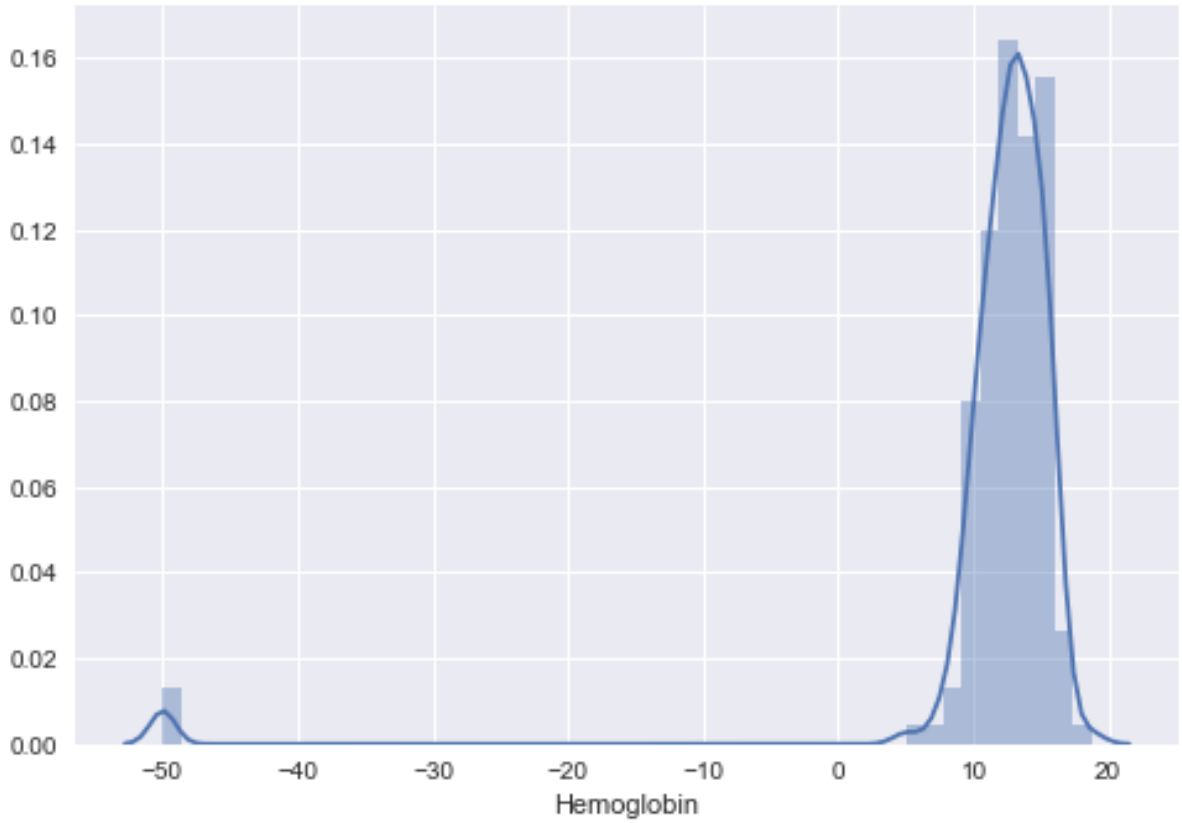


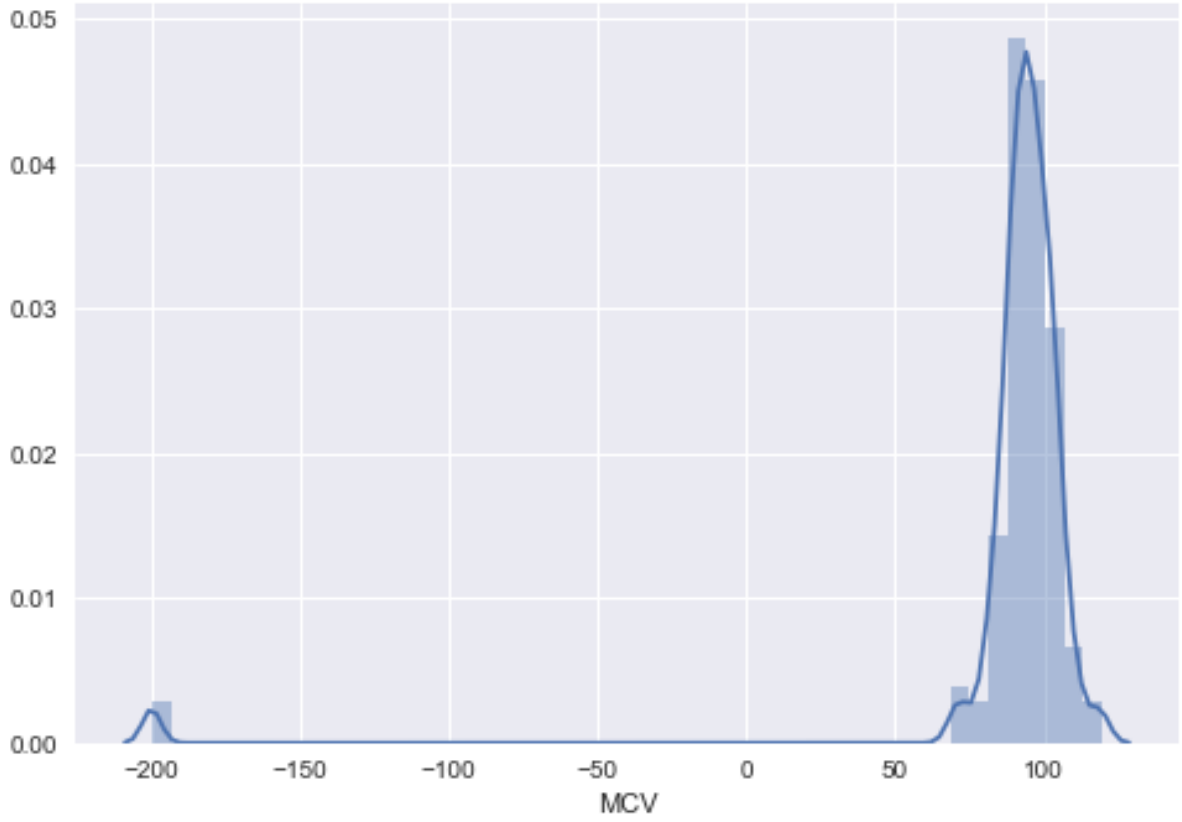
Σχήμα 2: Οι αρνητικές τιμές αντιπροσωπεύουν τις κενές τιμές (για γραφικούς λόγους και μόνο τους δώθηκε στο γράφημα η τιμή -200, δεν χρησιμοποιήθηκε όμως αυτή στην ανάλυση. Ομοίως και για τις ακόλουθες εικόνες για τις όποιες αρνητικές τιμές).

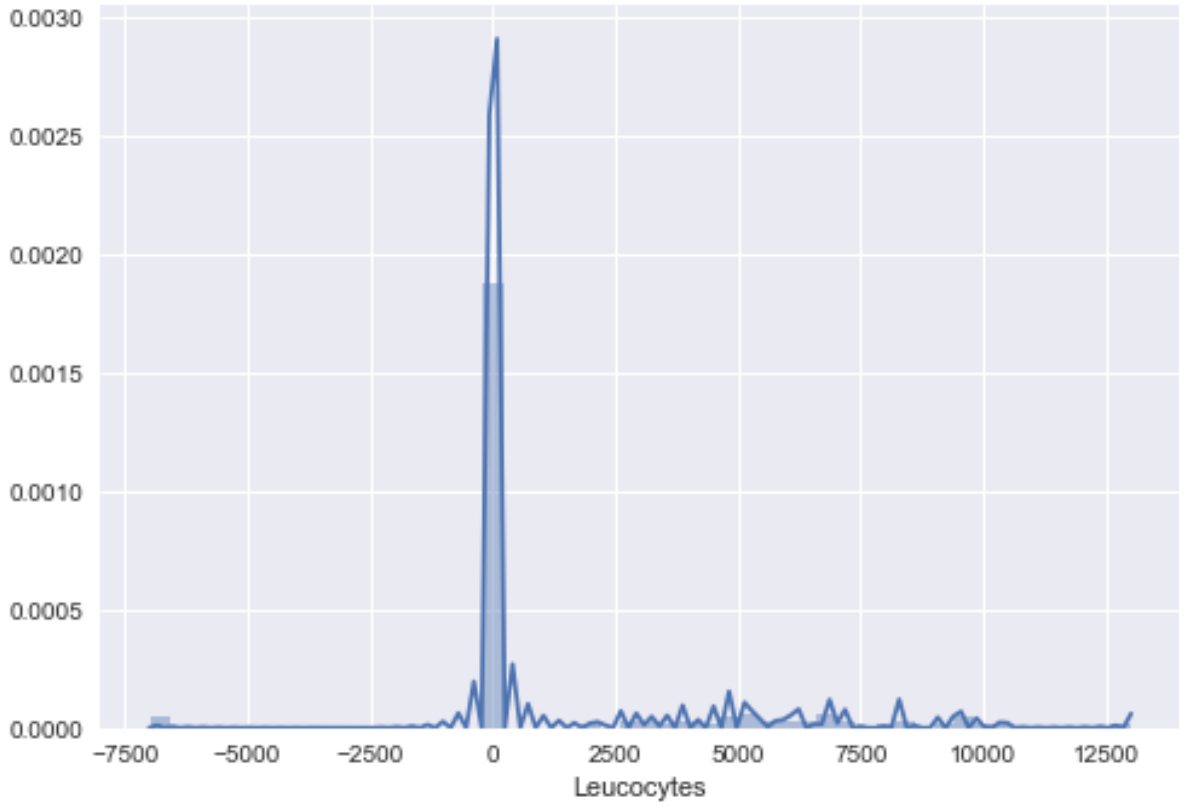

0.006
0.005
0.004
0.003
0.002
0.001
0.000
−200
0
200
400
600
Packs/year

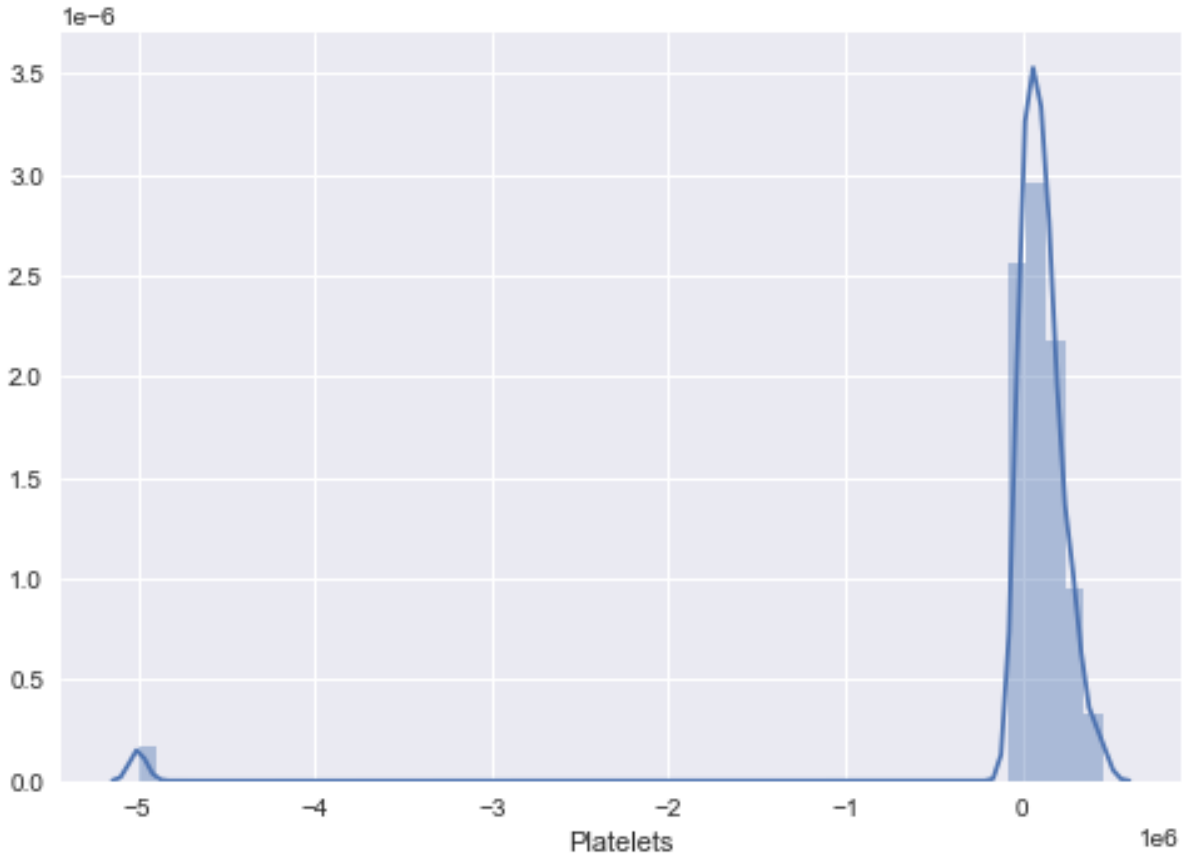

1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−2
0
2
4
INR

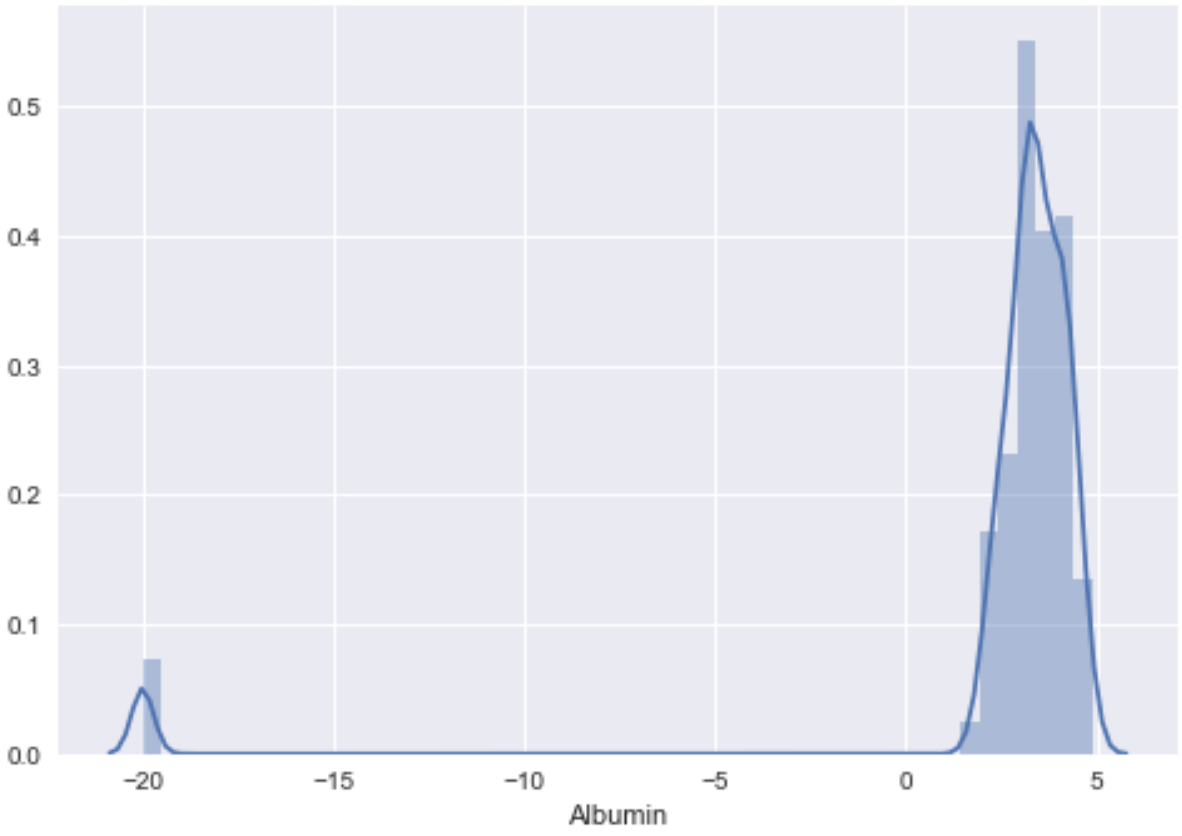

1e−5
4
3
2
1
0
−1.0
−0.5
0.0
0.5
1.0
1.5
AFP
1e6

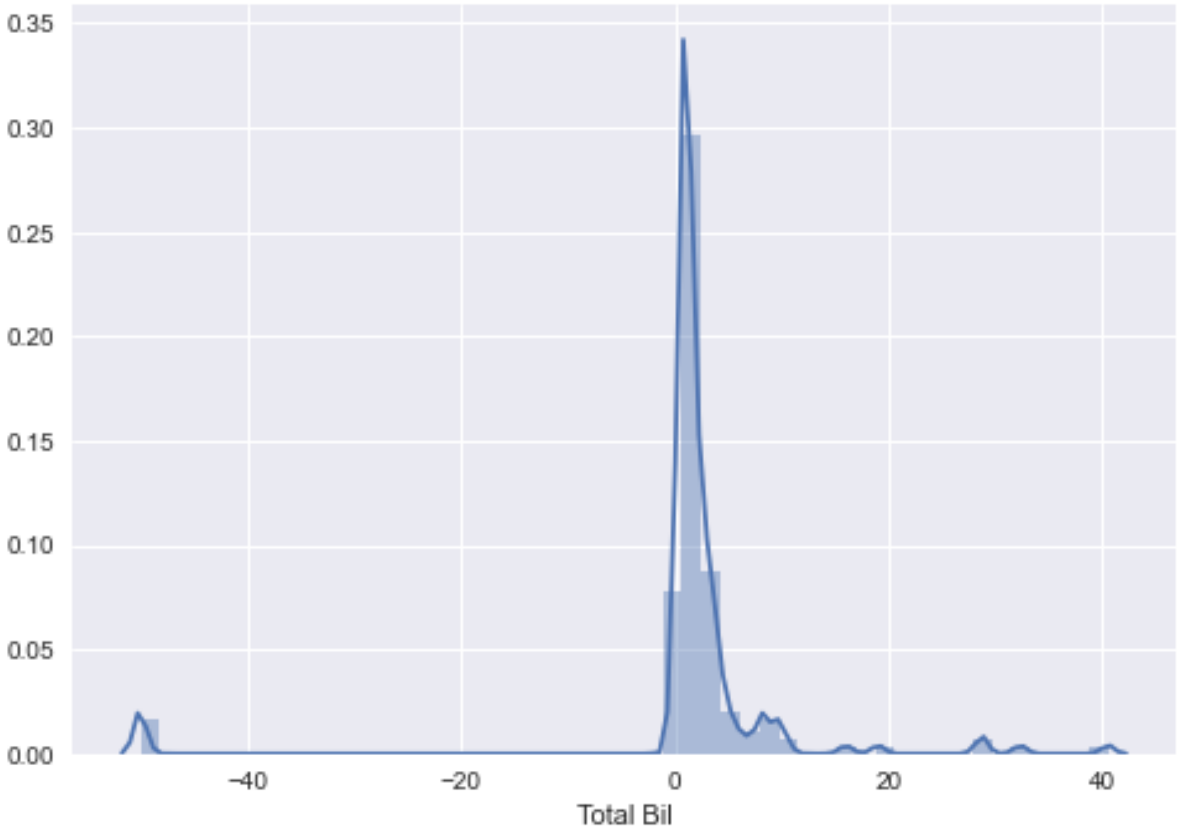

0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
−50
−40
−30
−20
−10
0
10
20
Hemoglobin

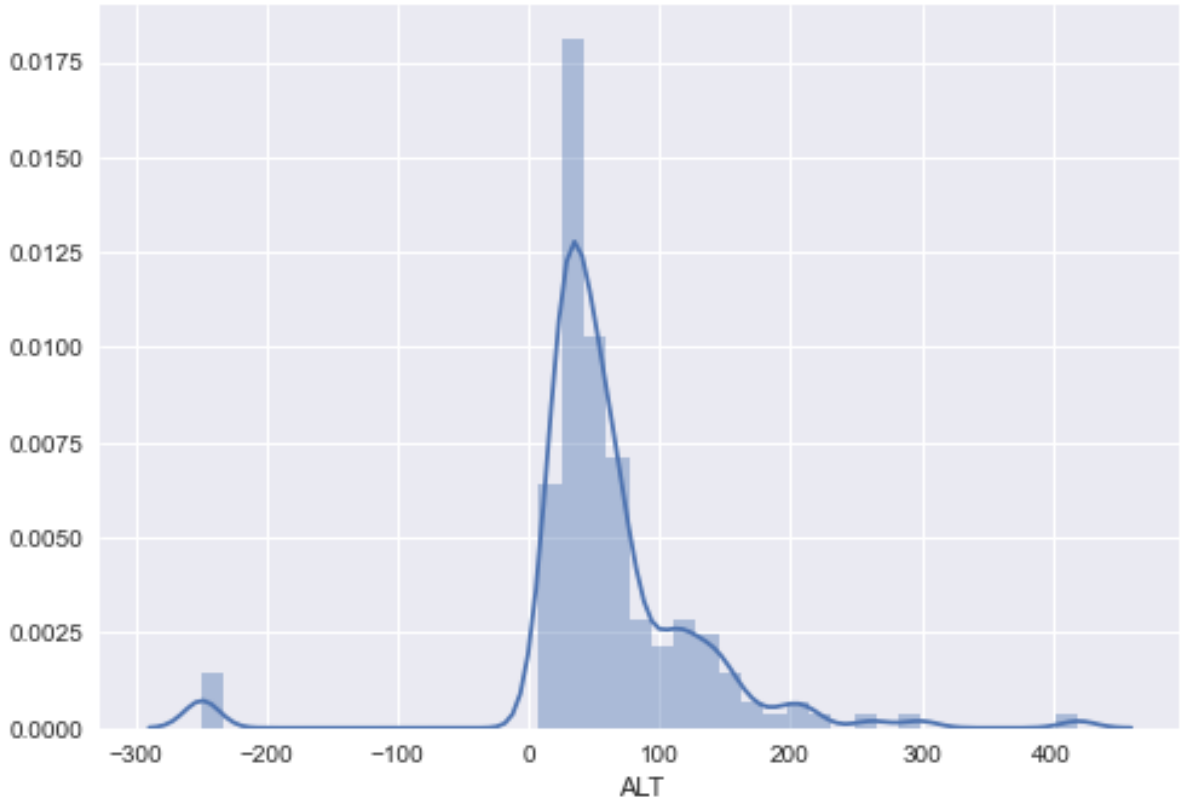

0.05
0.04
0.03
0.02
0.01
0.00
−200
−150
−100
−50
0
50
100
MCV

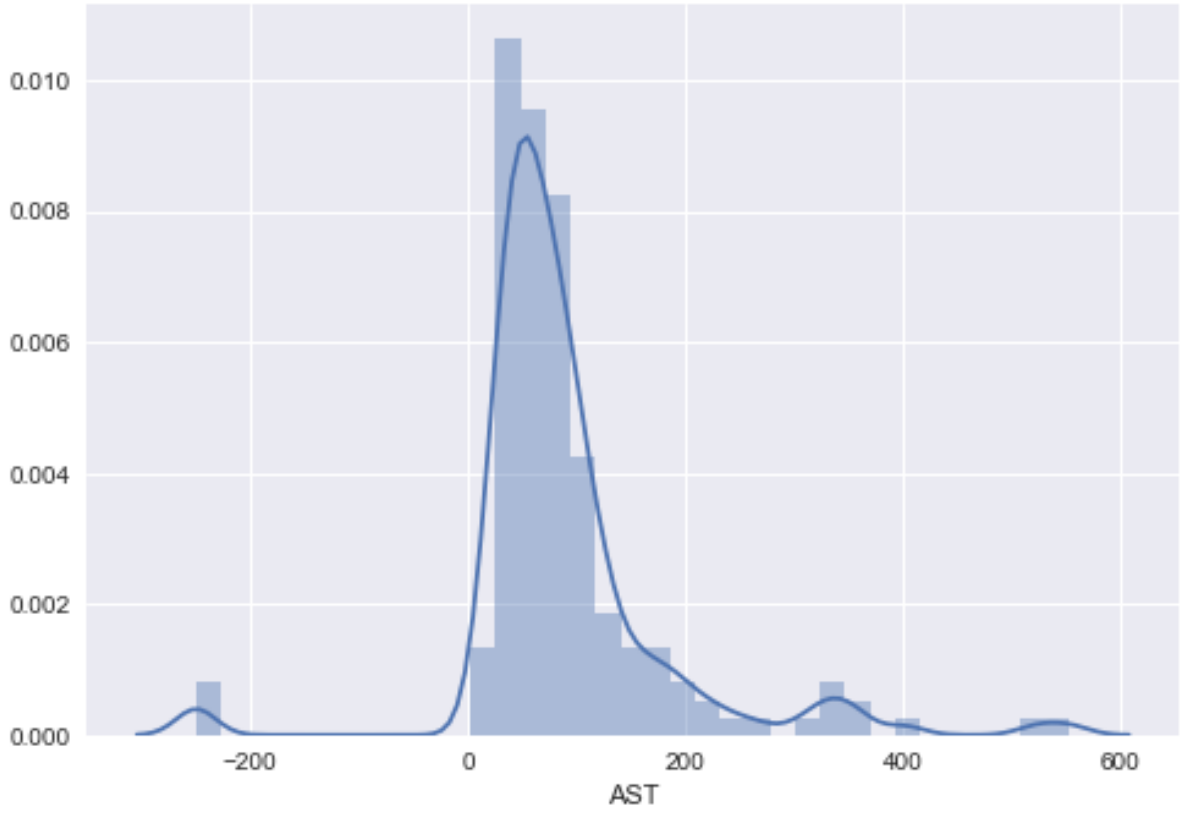

0.0030
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
−7500
−5000
−2500
0
2500
5000
7500
10000
12500
Leucocytes

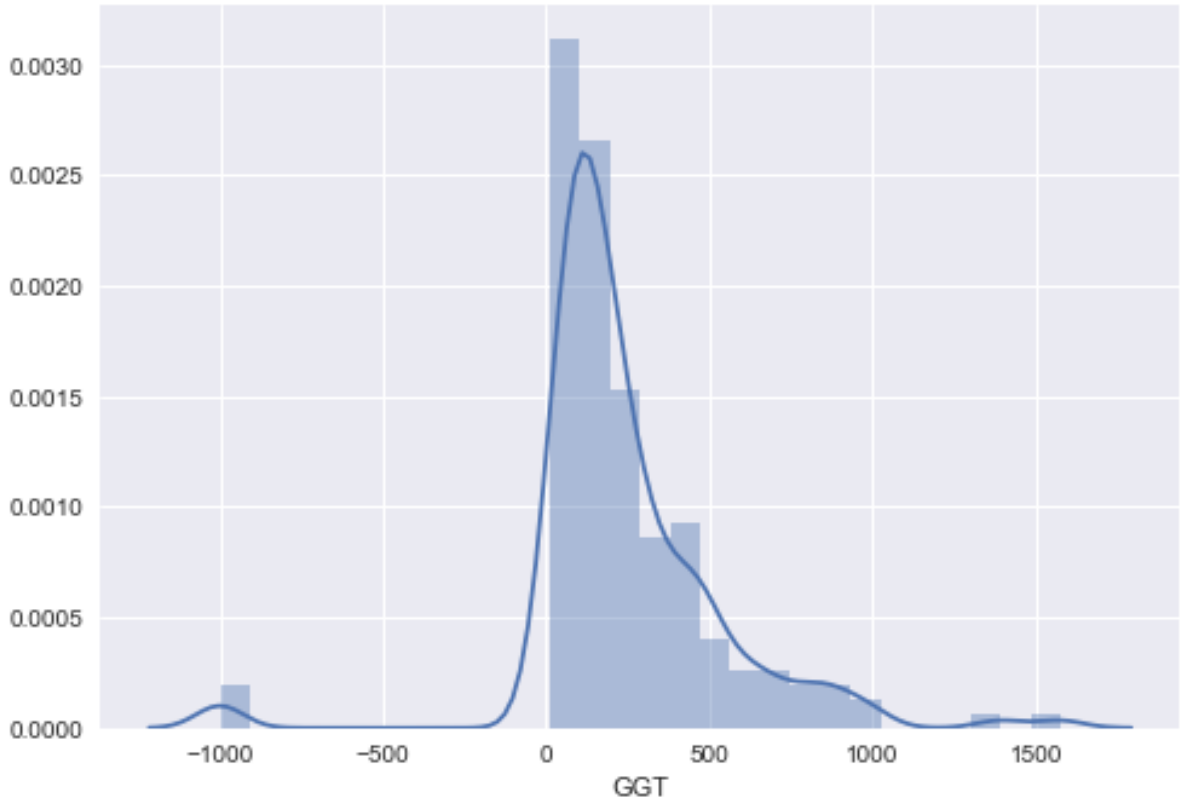

1e−6
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
−5
−4
−3
−2
−1
0
Platelets
1e6

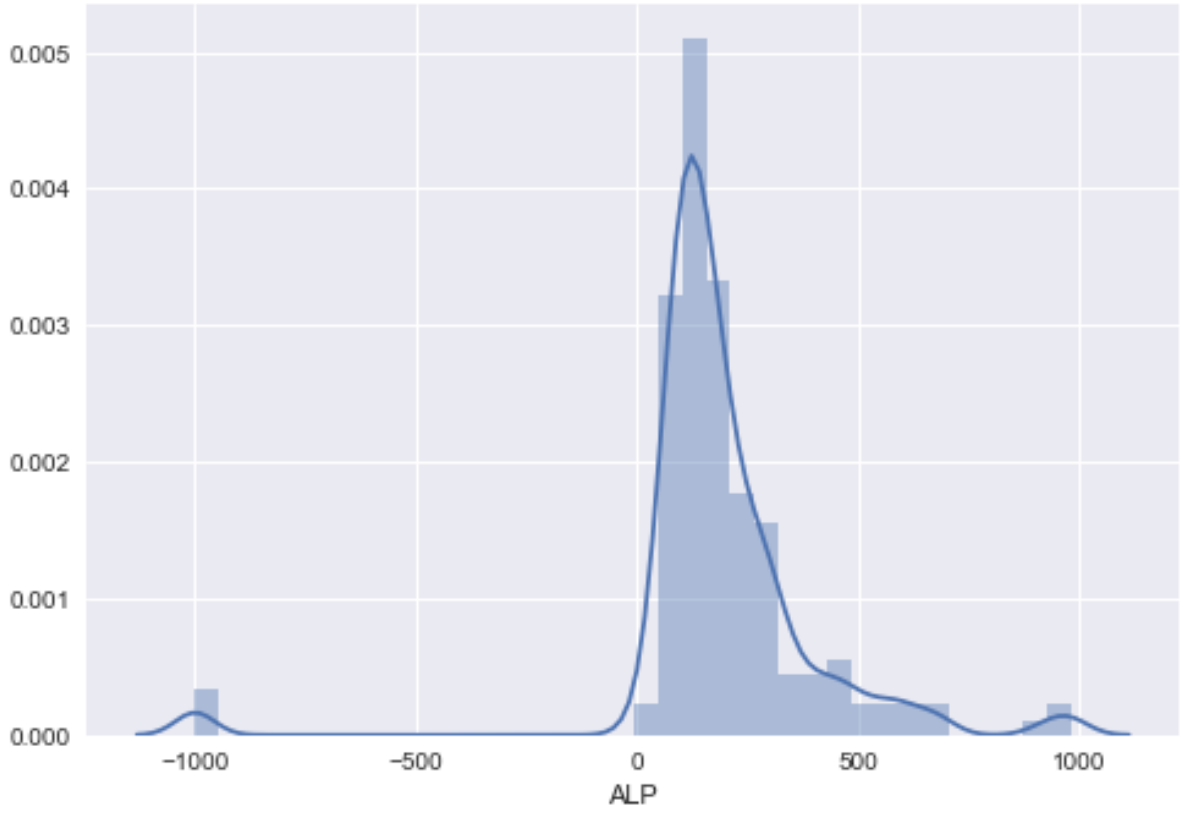

0.5
0.4
0.3
0.2
0.1
0.0
−20
−15
−10
−5
0
5
Albumin

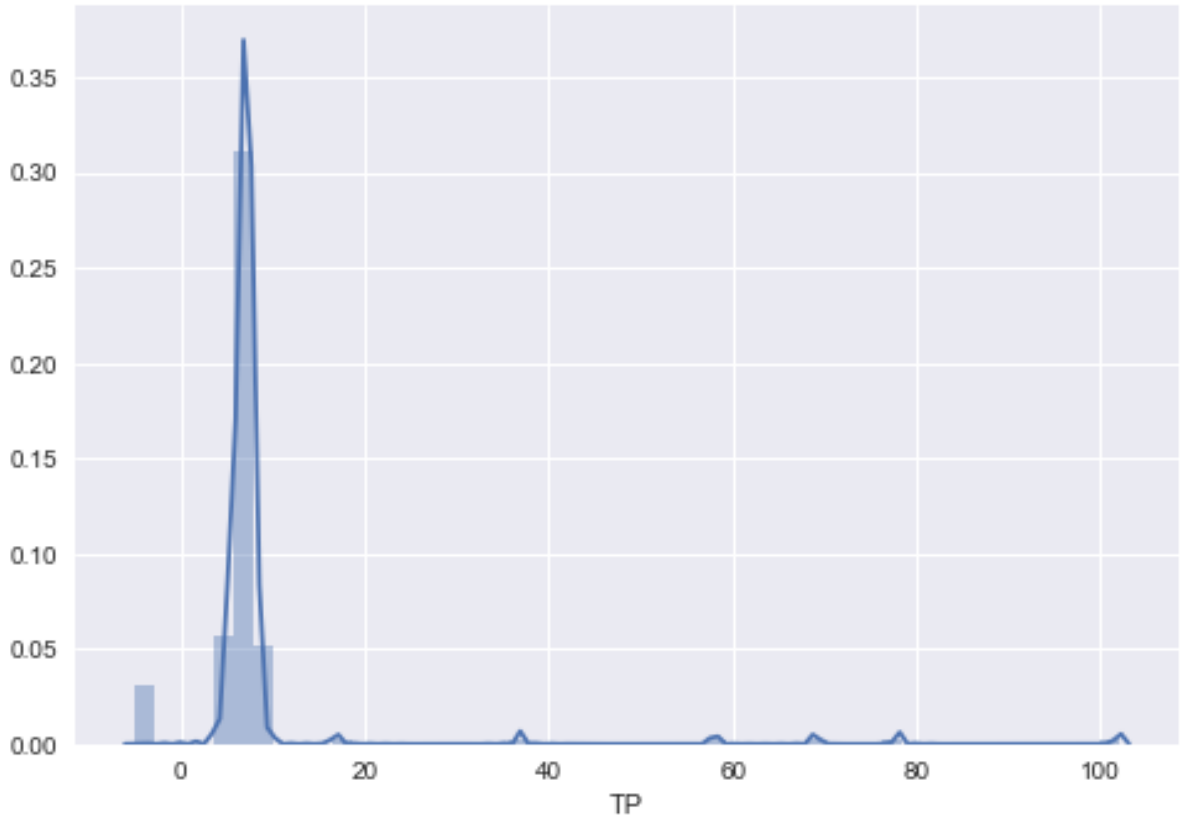

0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−40
−20
0
20
40
Total Bil

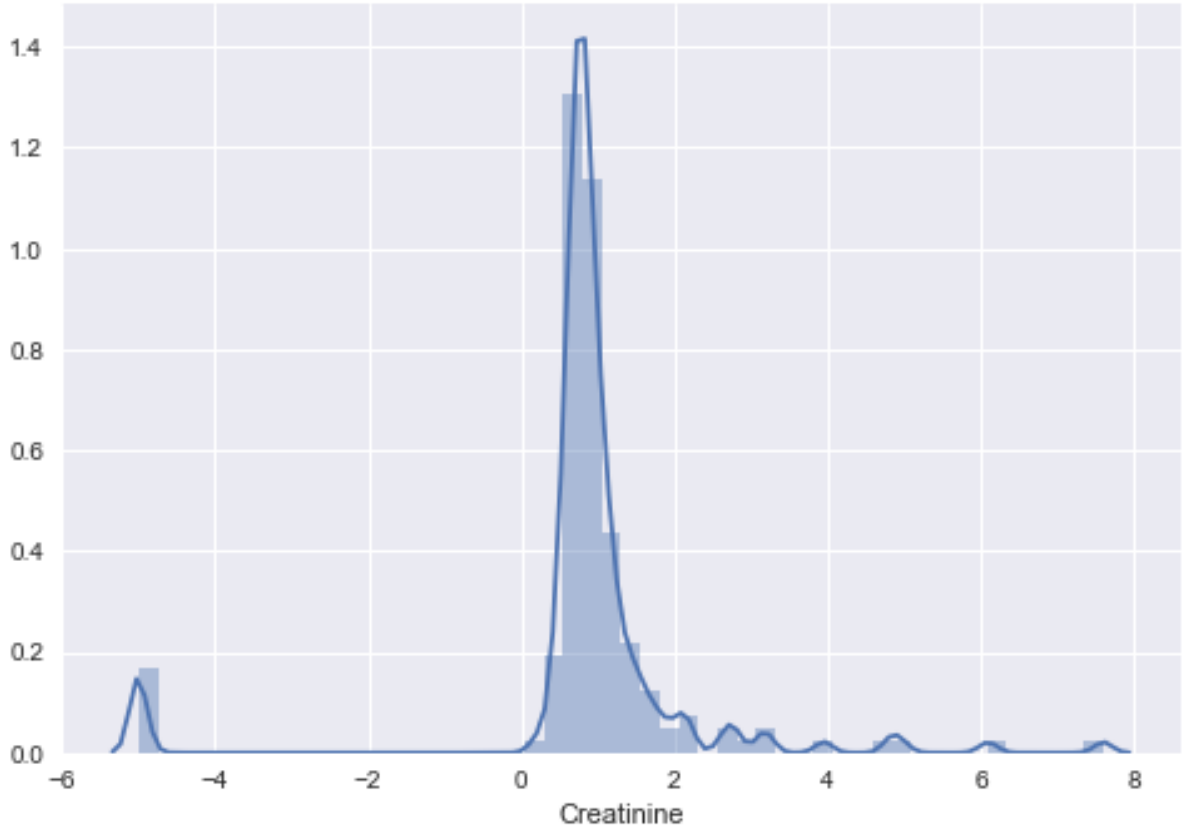

0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
−300
−200
−100
0
100
200
300
400
ALT

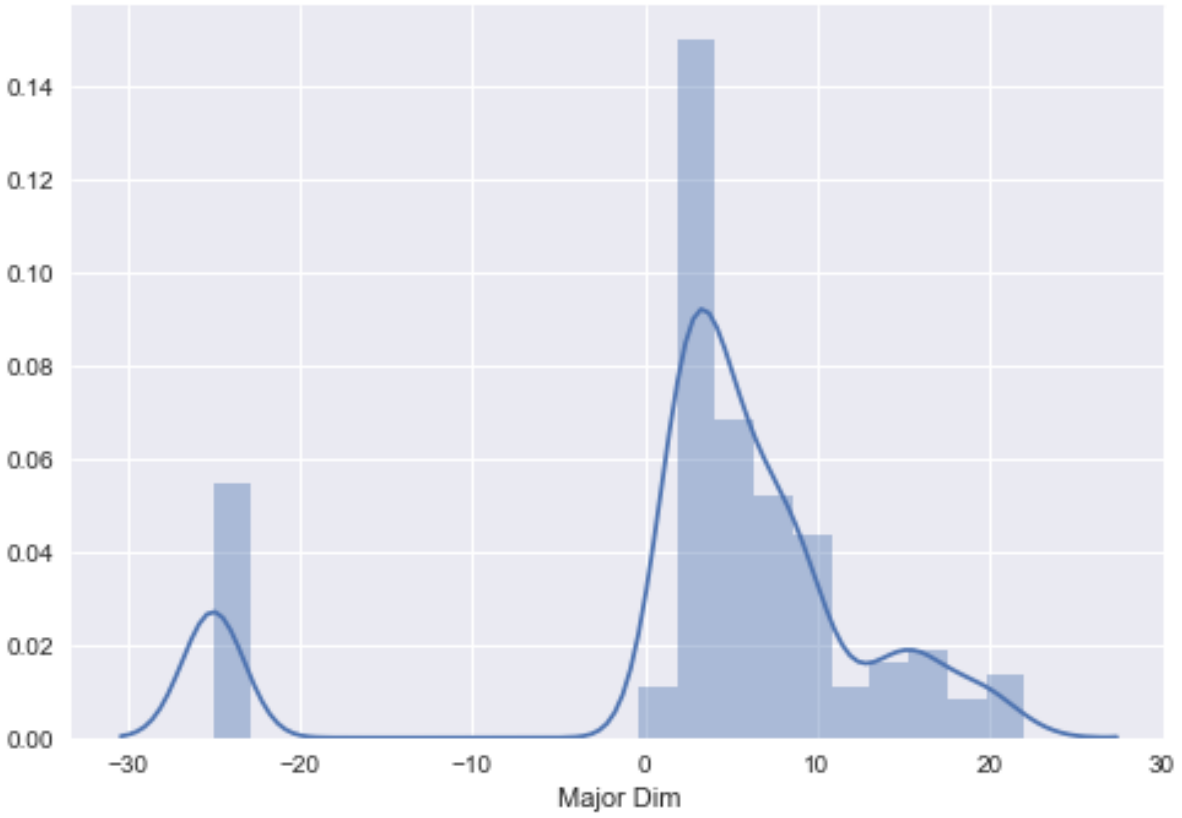

0.010
0.008
0.006
0.004
0.002
0.000
−200
0
200
400
600
AST

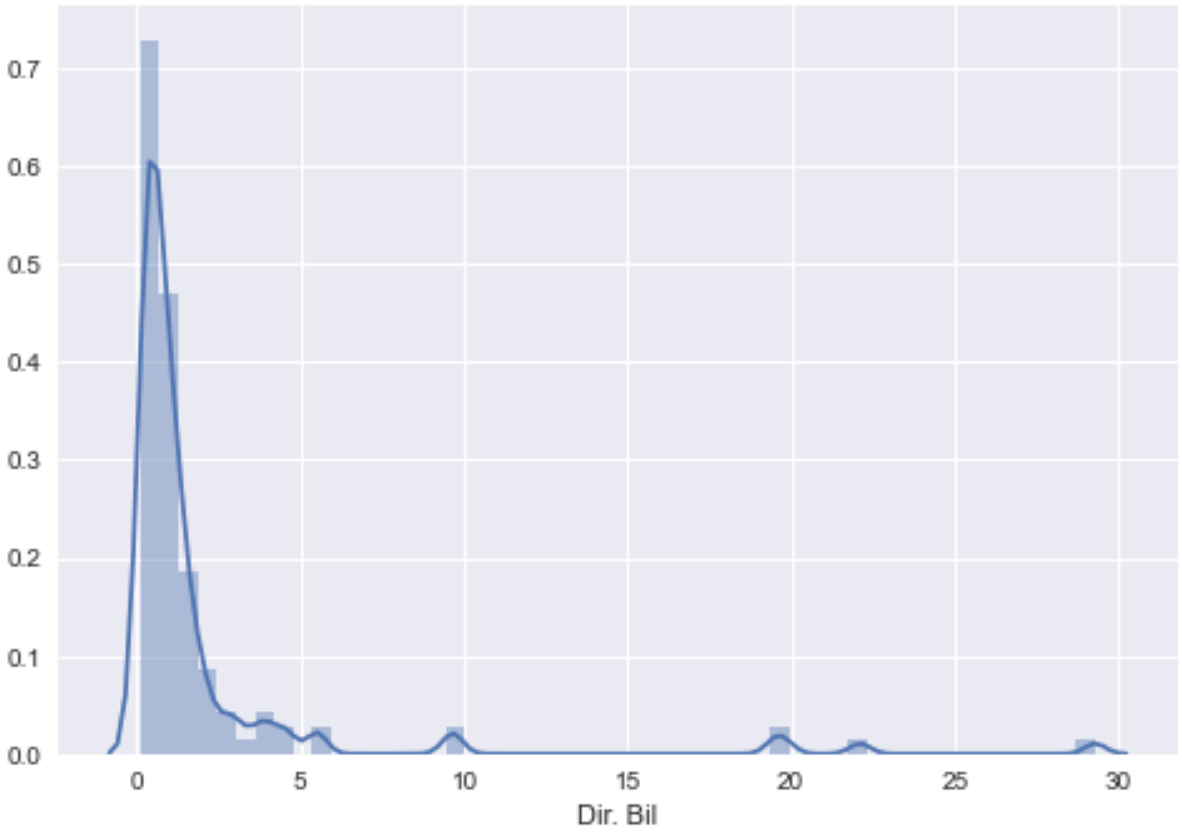

0.0030
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
−1000
−500
0
500
1000
1500
GGT

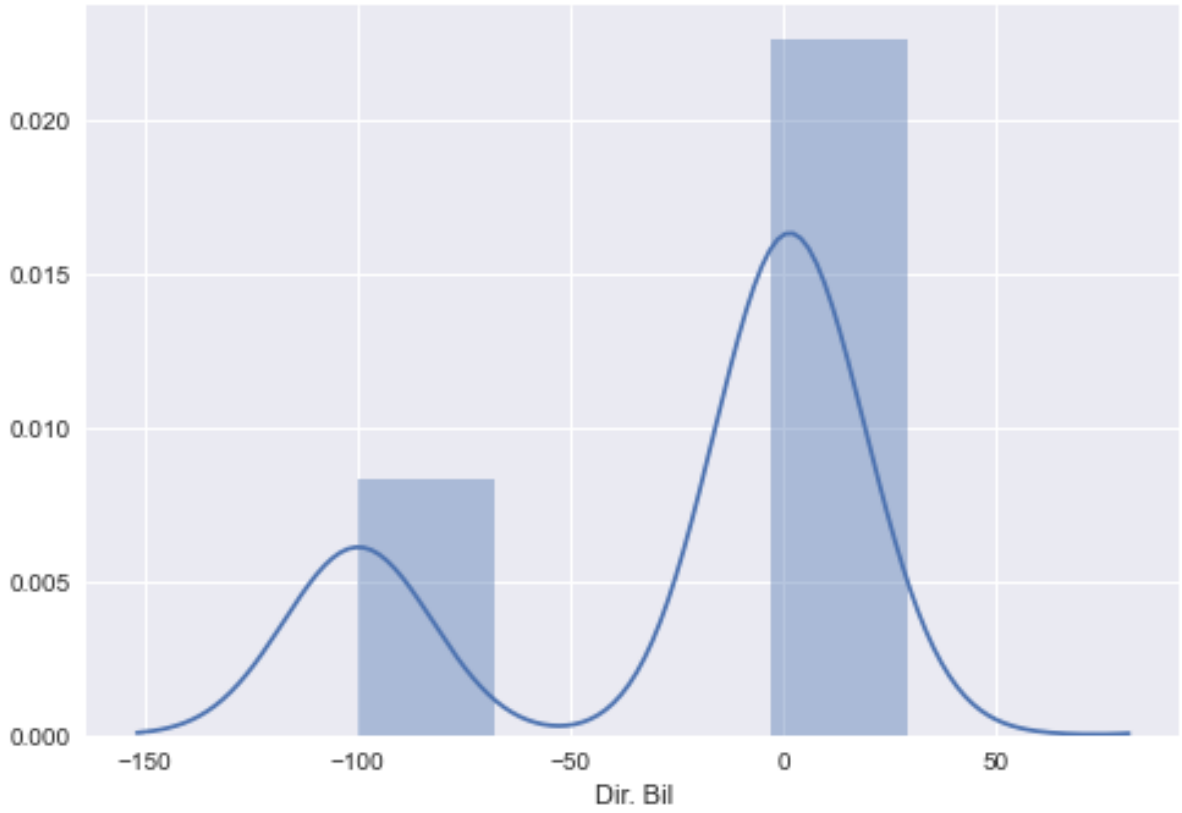

0.005
0.004
0.003
0.002
0.001
0.000
−1000
−500
0
500
1000
ALP

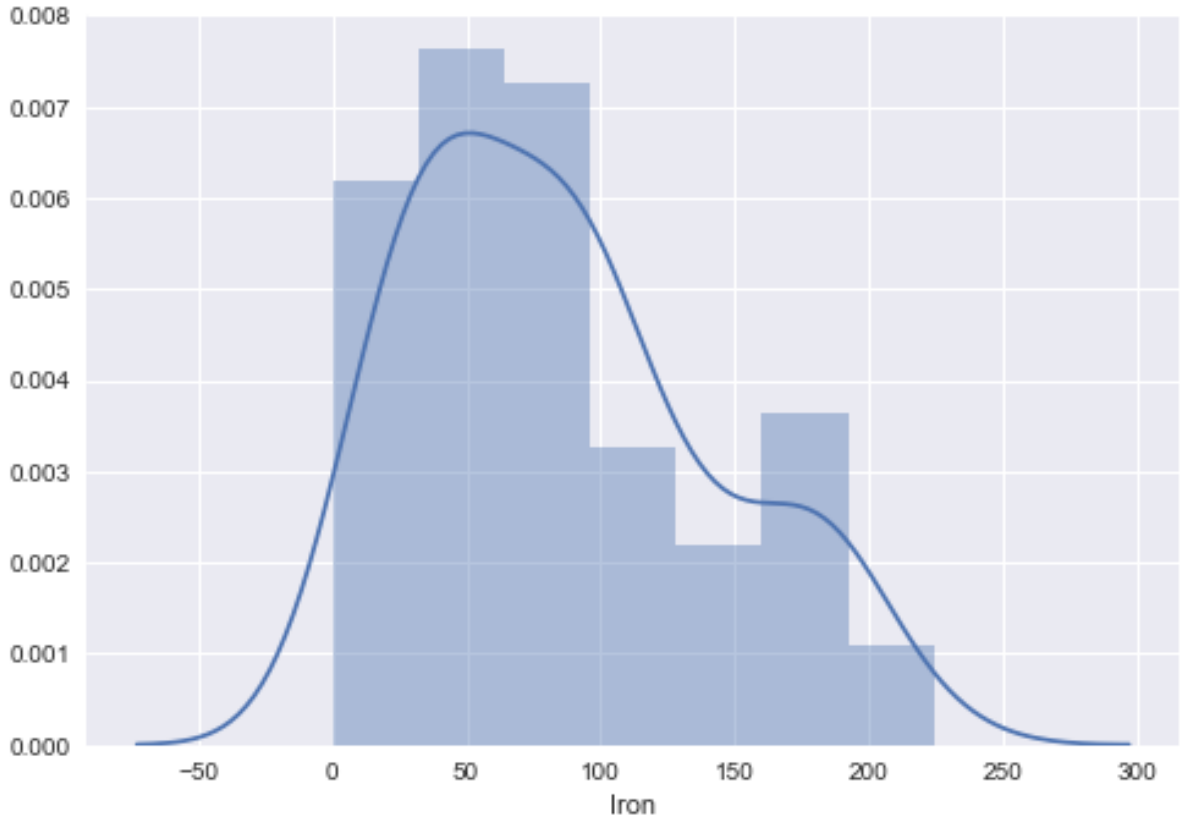

0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
20
40
60
80
100
TP

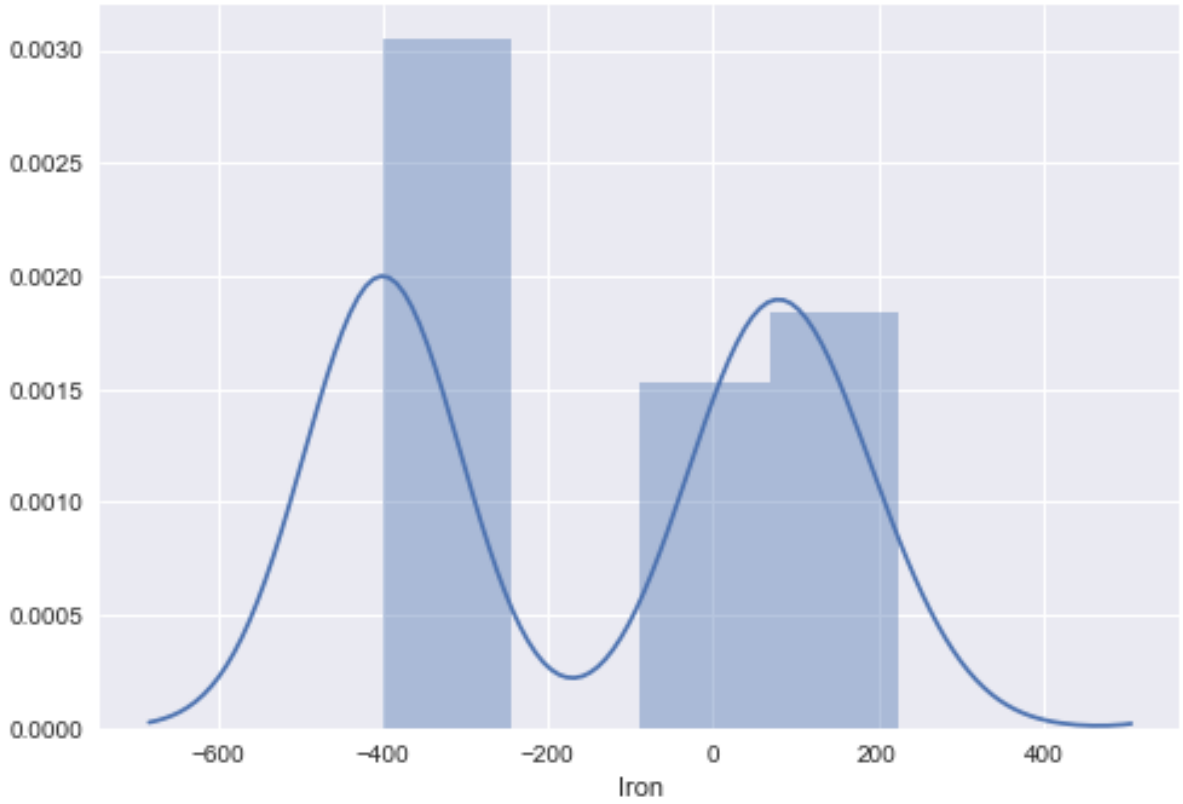

1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−6
−4
−2
0
2
4
6
8
Creatinine

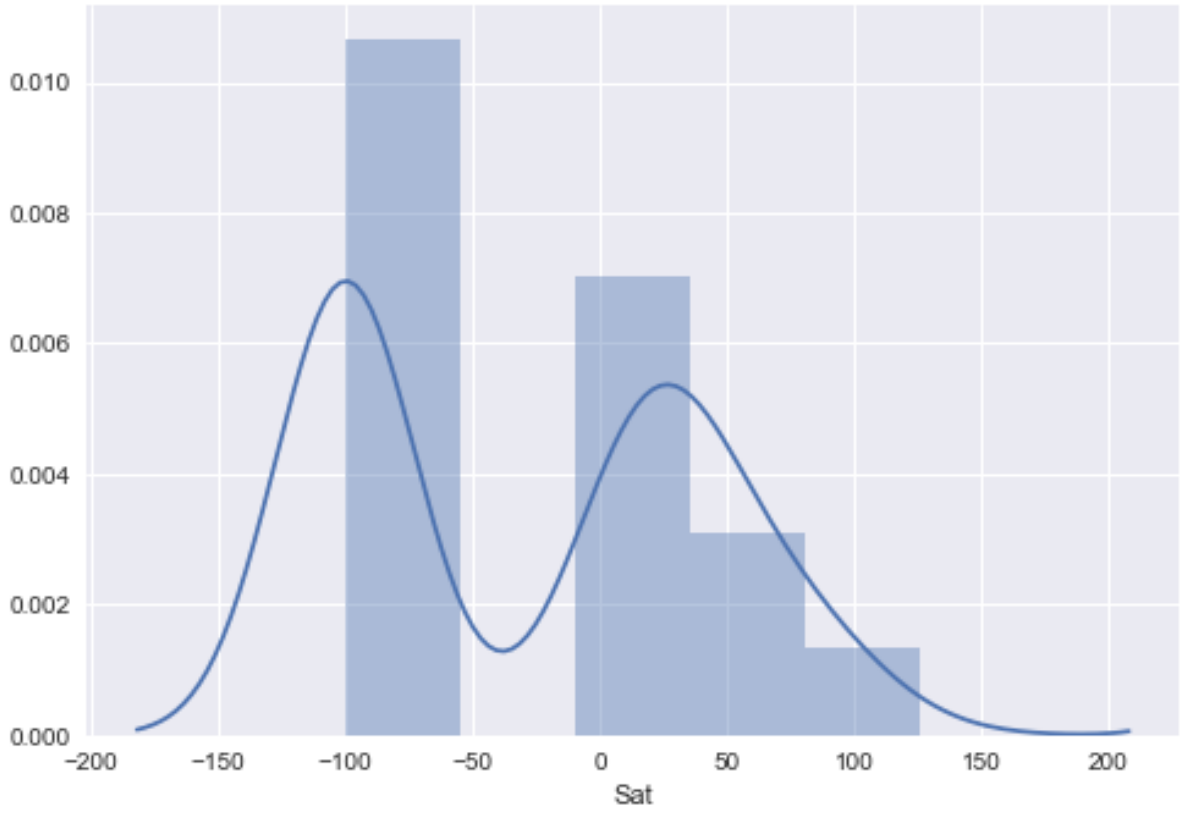

0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
−30
−20
−10
0
10
20
30
Major Dim

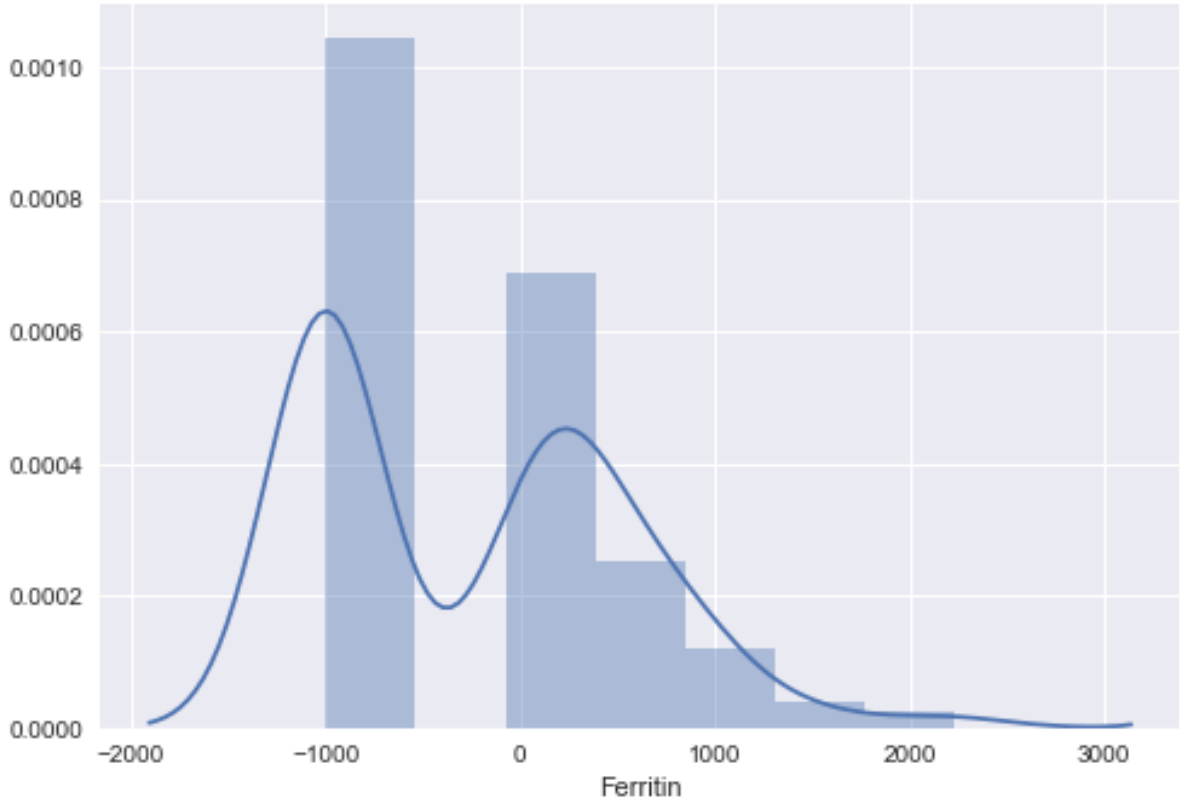

0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
5
10
15
20
25
30
Dir. Bil


0.020
0.015
0.010
0.005
0.000
−150
−100
−50
0
50
Dir. Bil


0.008
0.007
0.006
0.005
0.004
0.003
0.002
0.001
0.000
−50
0
50
100
150
200
250
300
Iron

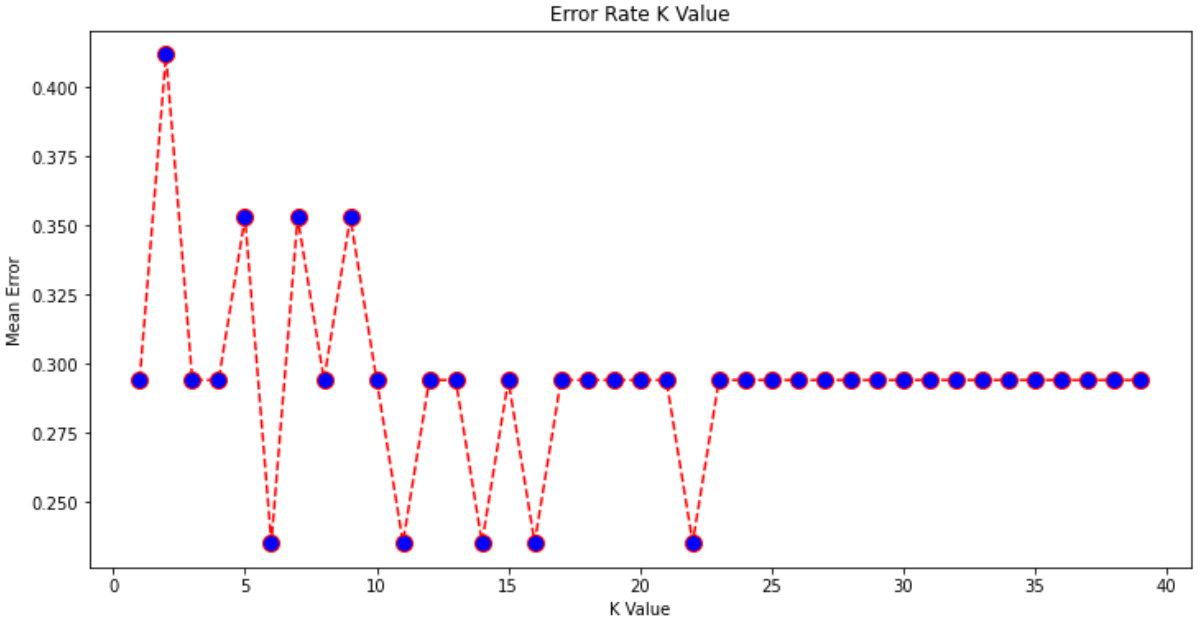

0.0030
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
−600
−400
−200
0
200
400
Iron

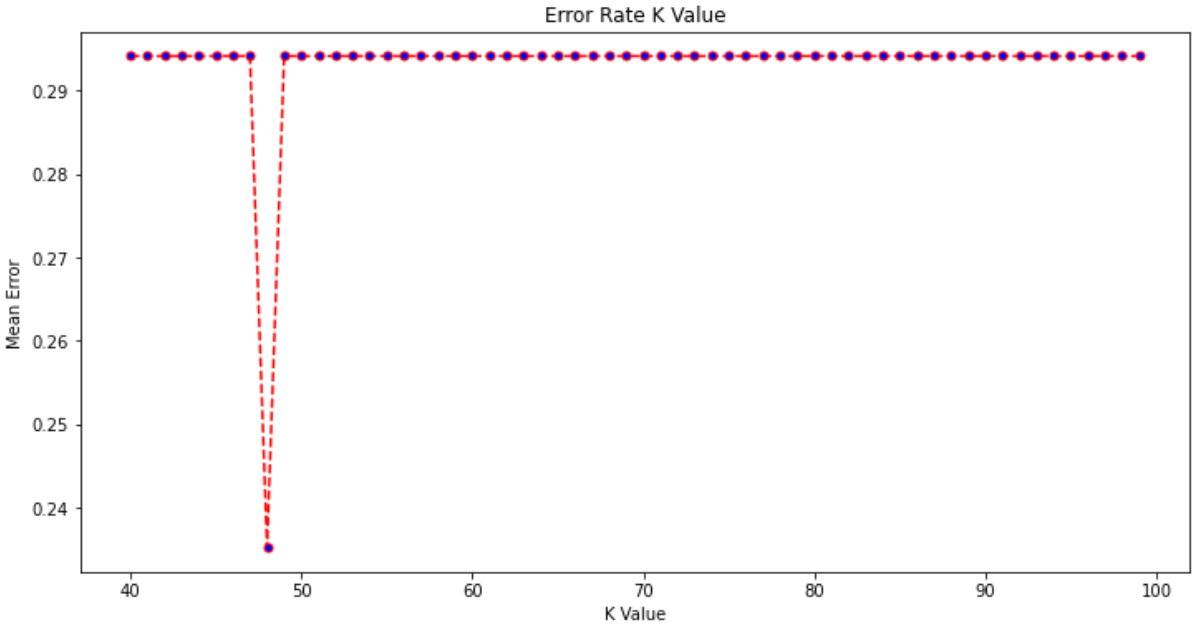

0.010
0.008
0.006
0.004
0.002
0.000
−200
−150
−100
−50
0
50
100
150
200
Sat


0.0010
0.0008
0.0006
0.0004
0.0002
0.0000
−2000
−1000
0
1000
2000
3000
Ferritin

10.2 Παράρτημα II

Γραφήματα με το Ποσοστό σφάλματος Error Rate των K-NN, υπολογισμένο για την προεπεξεργασία (εμπειρική συμπλήρωση, χωρίς εξισορρόπηση), όπου K-value: αριθμός των K-κοντινότερων γειτόνων που λαμβάνονται υπ'όψη.





10.3 Παράρτημα III

Ακολουθεί Πίνακας με τις παραλλαγές ταξινομητών που δοκιμάστηκαν ανά μέθοδο:

Μέθοδος	***Ταξινομητής (και παράμετροι)***
Support Vector Machines	SVC(kernel='rbf')
Logistic Regression	LogisticRegression(penalty='elasticnet', solver='saga', max_iter=10000, l1_ratio=0.5)
	LogisticRegression(penalty='l2', solver='saga', max_iter=10000)
	LogisticRegression(penalty='l1', solver='saga', max_iter=10000)
	LogisticRegression(penalty='l2', solver='sag', max_iter=10000)
	LogisticRegression(penalty='l2', solver='lbfgs', max_iter=10000)
	LogisticRegression(penalty='l2', solver='lbfgs', max_iter=890)
	LogisticRegression(penalty='l2', solver='newton-cg', max_iter=100000)
Decision Tree	DecisionTreeClassifier(criterion='gini',max_leaf_nodes=10, random_state=0)
	DecisionTreeClassifier(criterion='gini',max_leaf_nodes=11, random_state=0)
	DecisionTreeClassifier(criterion='gini',max_leaf_nodes=12, random_state=0)
	DecisionTreeClassifier(criterion='entropy', max_leaf_nodes=6, random_state=0)
Random Forest	RandomForestClassifier(criterion='entropy', max_leaf_nodes=2, bootstrap=True)
	RandomForestClassifier(criterion='gini', max_leaf_nodes=3, bootstrap=True)
Neural Network	MLPClassifier(hidden_layer_sizes=(50,25), activation='logistic', solver='lbfgs', learning_rate='adaptive', random_state=1, max_iter=1000)
	MLPClassifier(hidden_layer_sizes=(35,10), activation='logistic', solver='lbfgs', learning_rate='adaptive', random_state=1, max_iter=1000)
	MLPClassifier(hidden_layer_sizes=(25,8), activation='logistic', solver='lbfgs', learning_rate='adaptive', random_state=1, max_iter=1000)
	MLPClassifier(hidden_layer_sizes=(35,8), activation='logistic', solver='lbfgs', learning_rate='adaptive', random_state=1, max_iter=1000)
	MLPClassifier(hidden_layer_sizes=(25,25), activation='logistic', solver='lbfgs', learning_rate='adaptive',

	random_state=1, max_iter=1000)
	MLPClassifier(hidden_layer_sizes=(35,15), activation='logistic', solver='lbfgs', learning_rate='adaptive', random_state=1, max_iter=1000)
	MLPClassifier(hidden_layer_sizes=(35,9), activation='logistic', solver='lbfgs', learning_rate='adaptive', random_state=1, max_iter=1000)
	MLPClassifier(hidden_layer_sizes=(35,11), activation='logistic', solver='lbfgs', learning_rate='adaptive', random_state=1, max_iter=1000)
	MLPClassifier(hidden_layer_sizes=(36,10), activation='logistic', solver='lbfgs', learning_rate='adaptive', random_state=1, max_iter=2000)
	MLPClassifier(hidden_layer_sizes=(34,10), activation='logistic', solver='lbfgs', learning_rate='adaptive', random_state=1, max_iter=2000)
K-Nearest Neighbours	KNeighborsClassifier(n_neighbors=5)
	KNeighborsClassifier(n_neighbors=6)
	KNeighborsClassifier(n_neighbors=11)
	KNeighborsClassifier(n_neighbors=14)
	KNeighborsClassifier(n_neighbors=16)
	KNeighborsClassifier(n_neighbors=22)
	KNeighborsClassifier(n_neighbors=48)
Naive Bayes	BernoulliNB(binarize = True)
	BernoulliNB(binarize = 100)
	GaussianNB()

10.4 Προγνωστική Ανάλυση (Predictive Analysis)

Υπόμνημα:

Ακολουθούν ***ραβδογράμματα***.

Κατακόρυφοι άξονες:
εκάστοτε εξεταζόμενη ***μετρική***

Οριζόντιοι άξονες:
κάθε μία από τις δοκιμασθείσες ***πορείες προεπεξεργασίας*** ή ***μεθόδους ταξινόμησης***.

Όριο σφάλματος (αναφέρεται ως “error bars”) : η ***τυπική απόκλιση***

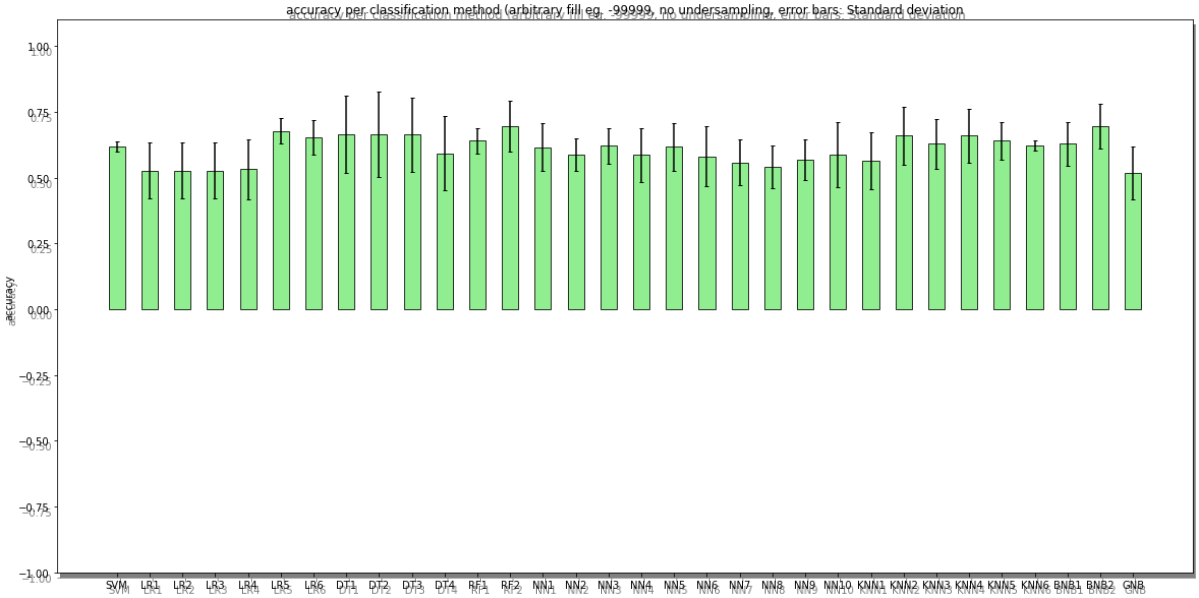
Τίτλος-λεζάντα κάθε γραφήματος στην ***πάνω*** πλευρά.
[ο αριθμός των διαγραμμάτων είναι τέτοιος που καθιστά απαγορευτικό το να τοποθετηθεί ξεχωριστή λεζάντα σε κάθε ένα γράφημα (από άποψη χώρου, μνήμης και επεξεργαστικής δυνατότητας του κειμενογράφου)]

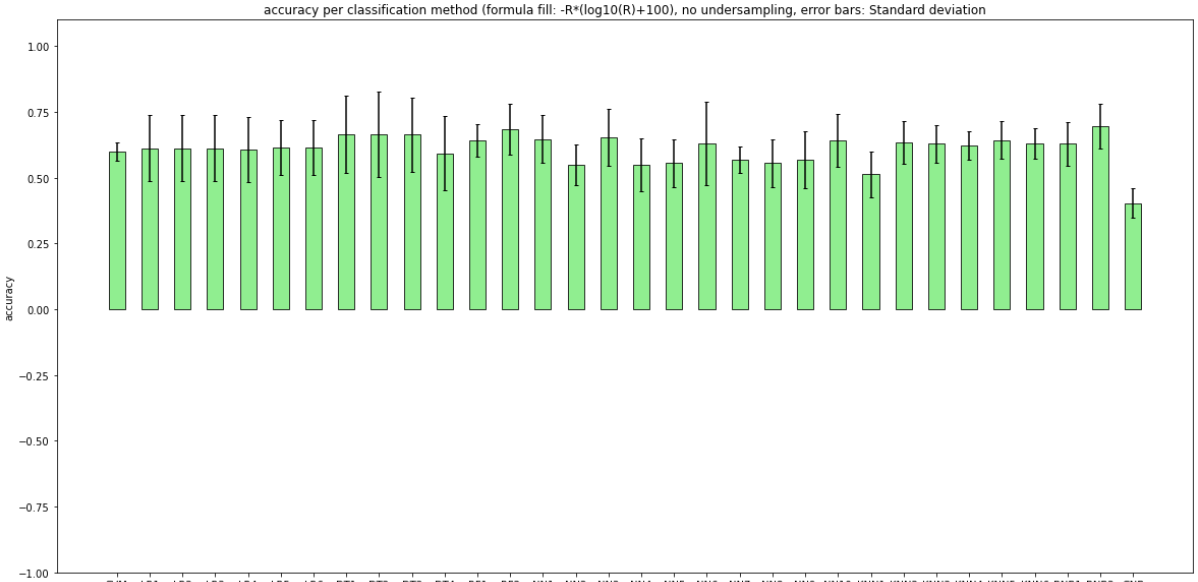
Για να είναι ευδιάκριτα τα γραφήματα και οι λεζάντες τους ***συστήνεται μεγέθυνση***, και έχει προνοηθεί να είναι καλή η ανάλυσή τους ακόμη και με μεγάλη μεγέθυνση. Τα ραβδογράμματα έχουν σμικρυνθεί προκειμένου να χωρούν 3 ανά σελίδα.

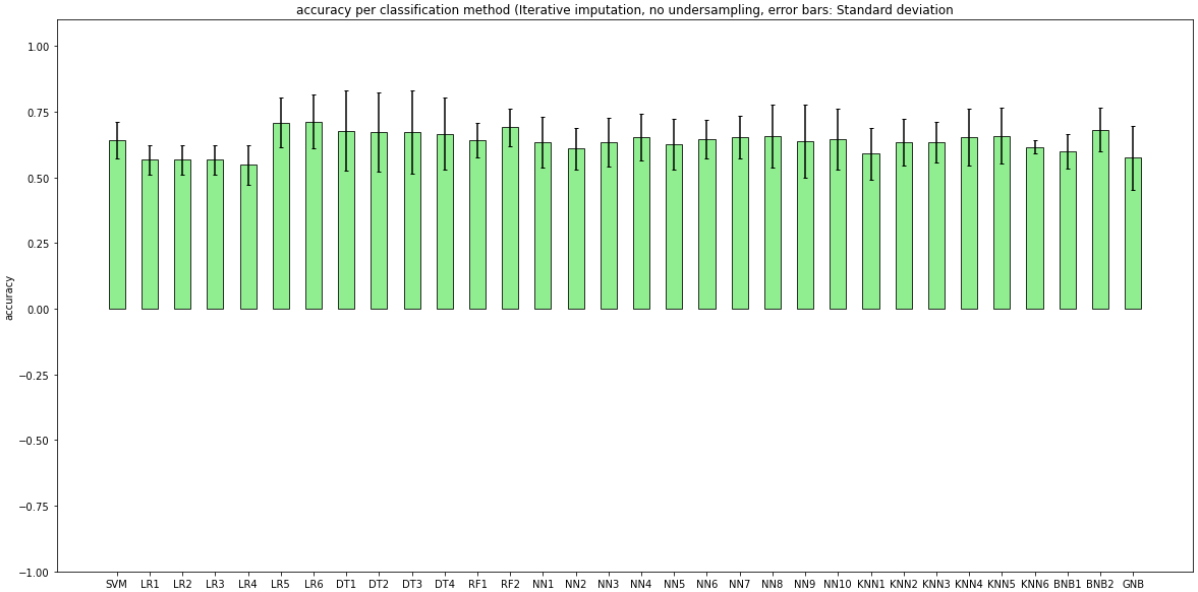
[Σημείωση: όπου υπάρχει η συντομογραφική ορολογία «*no undersampling» αναφέρεται στην έλλειψη εξισορρόπησης*. Όπου υπάρχει η συντομογραφική ορολογία “*formula fill” αναφέρεται στην συνάρτηση που δοκιμάστηκε για την συμπλήρωση των κενών* στα δεδομένα. Όπου υπάρχει η συντομογραφική ορολογία “*arbitrary fill” αναφέρεται στην εμπειρική συμπλήρωση κενών που δοκιμάστηκε.*]

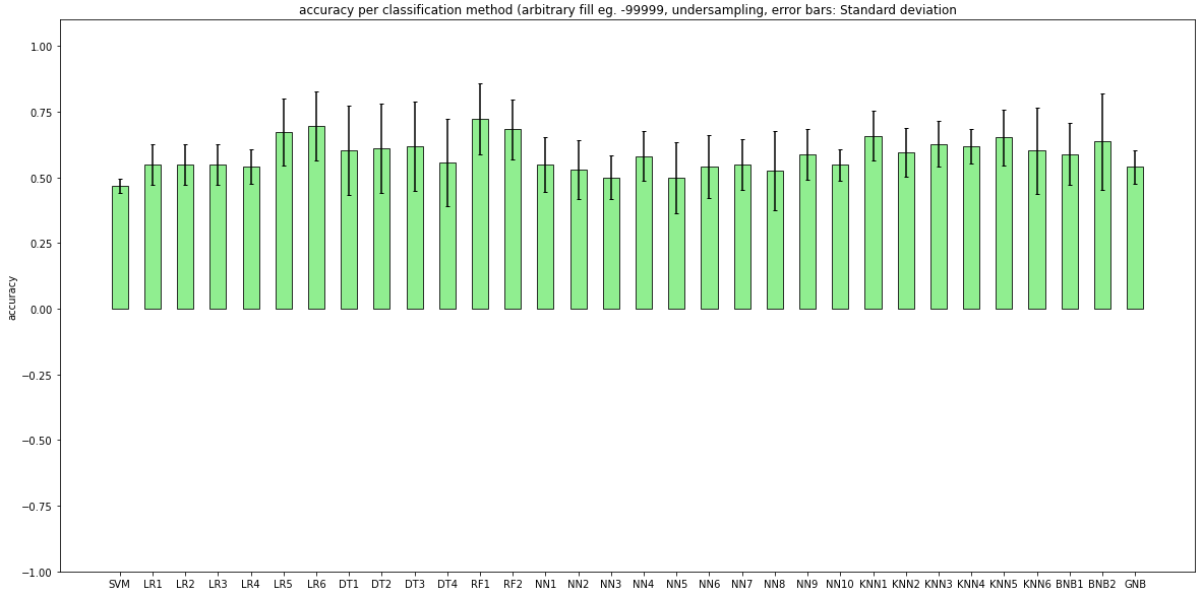
Τα γραφήματα έχουν όλες τις λεζάντες στα αγγλικά λόγω του ότι έτσι θα είναι ευκολότερο να γίνουν αντιληπτά και από άτομα που δεν είναι ελληνόφωνα.

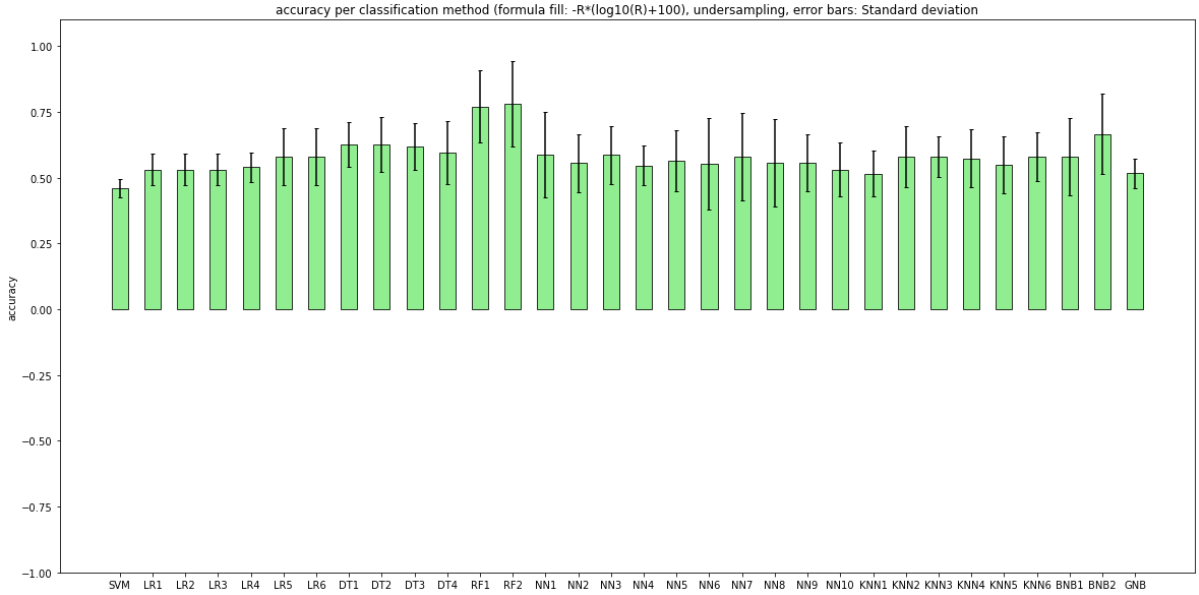
10.4.1 Παράρτημα IV a

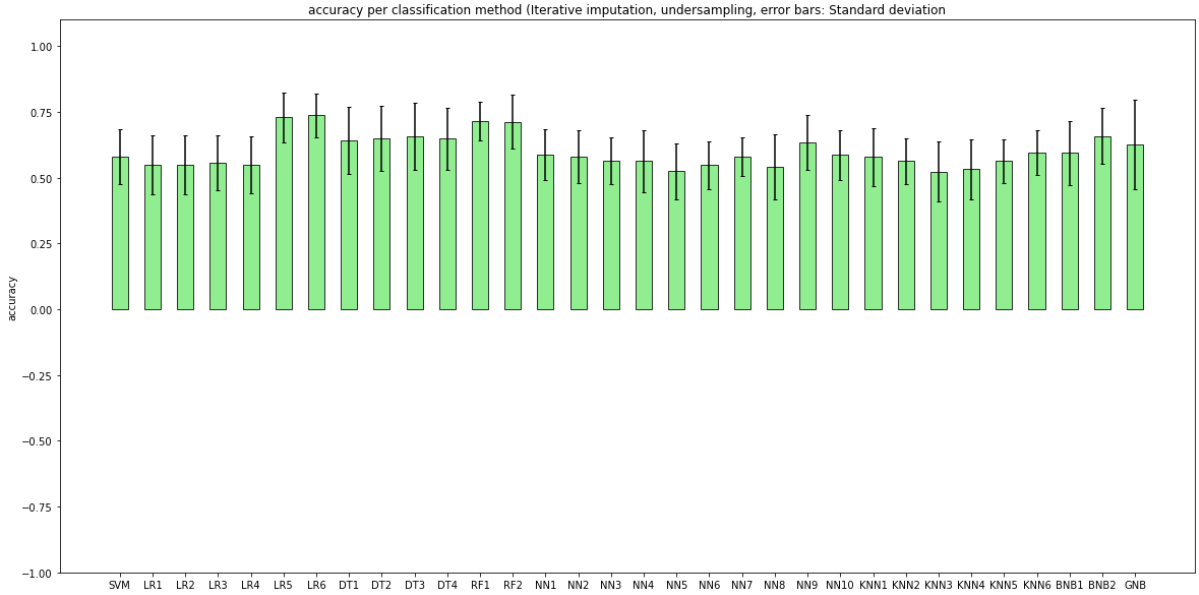


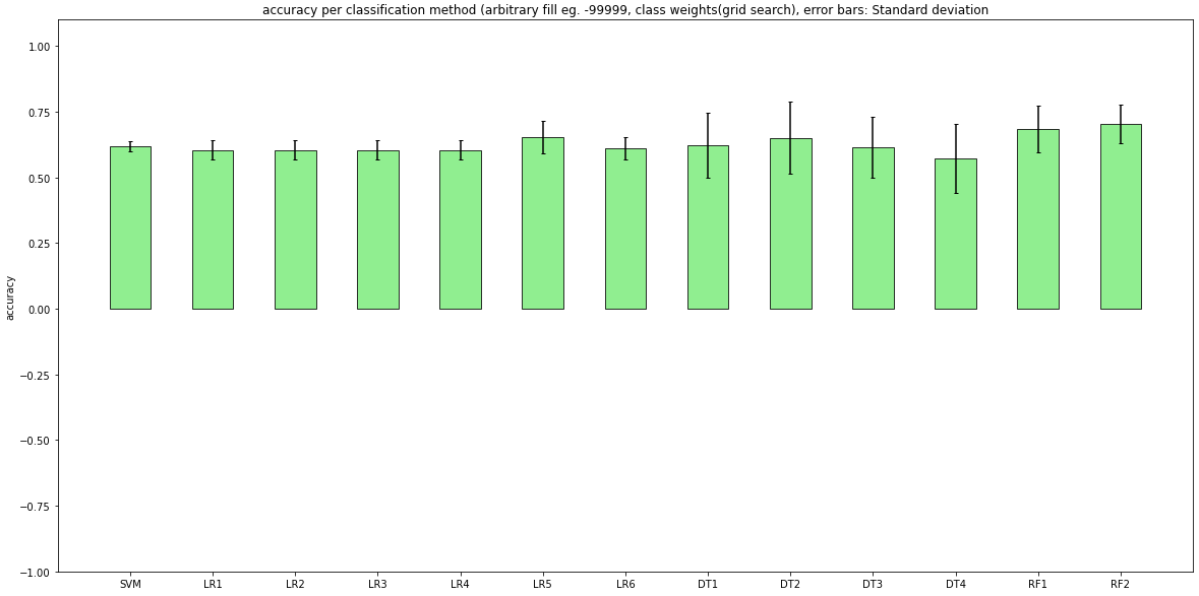


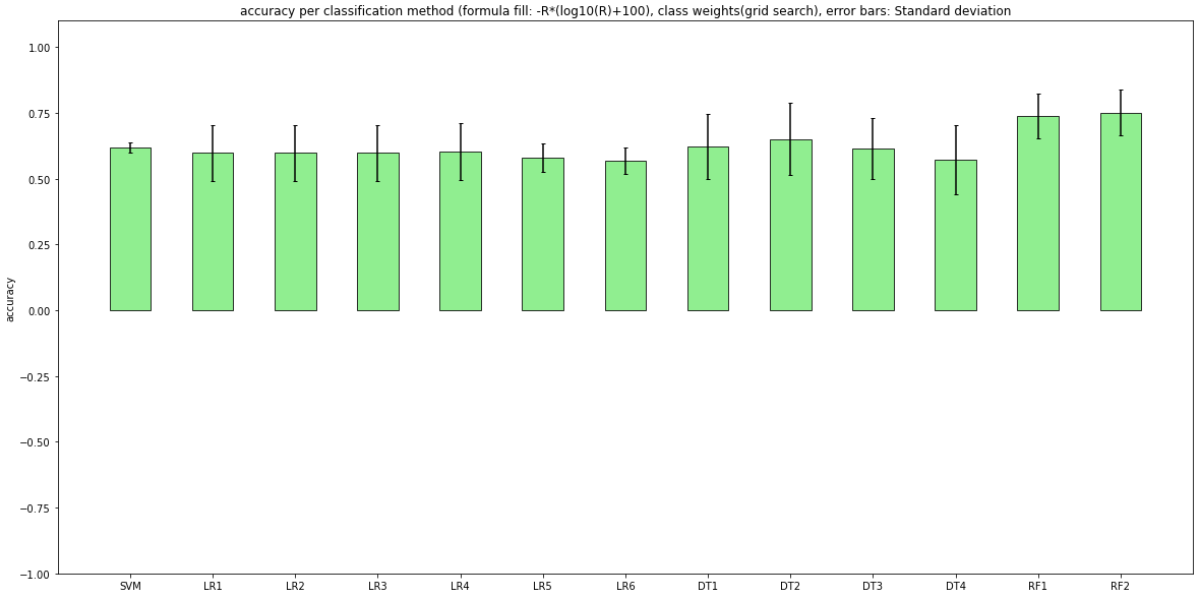


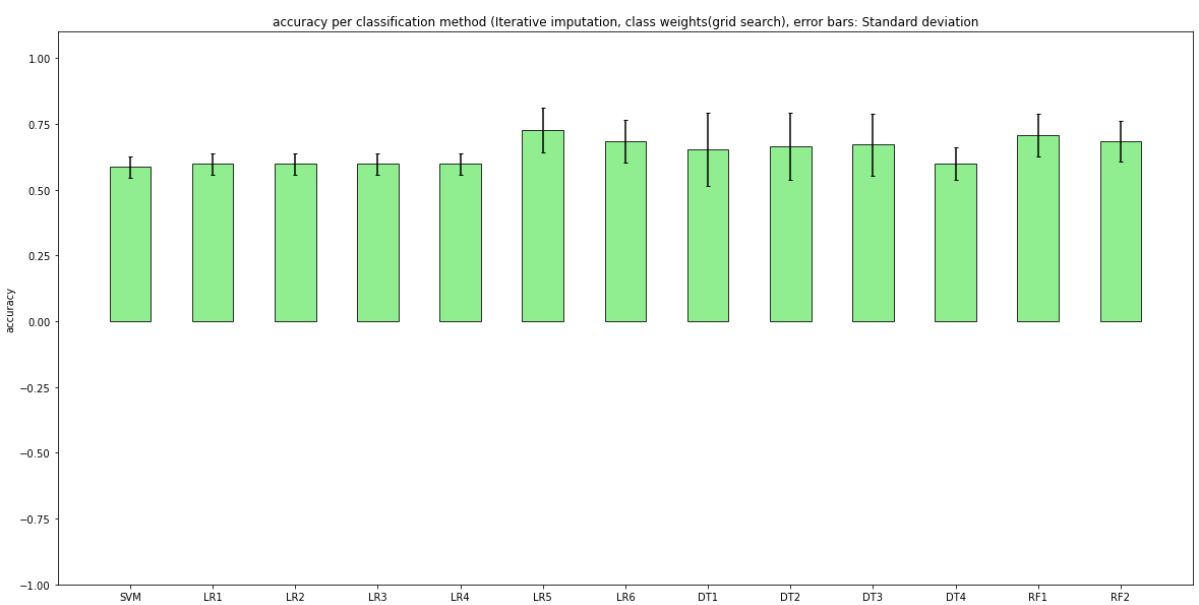

accuracy per classification method (arbitrary fill eg. -99999, undersampling, error bars: Standard deviation
accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

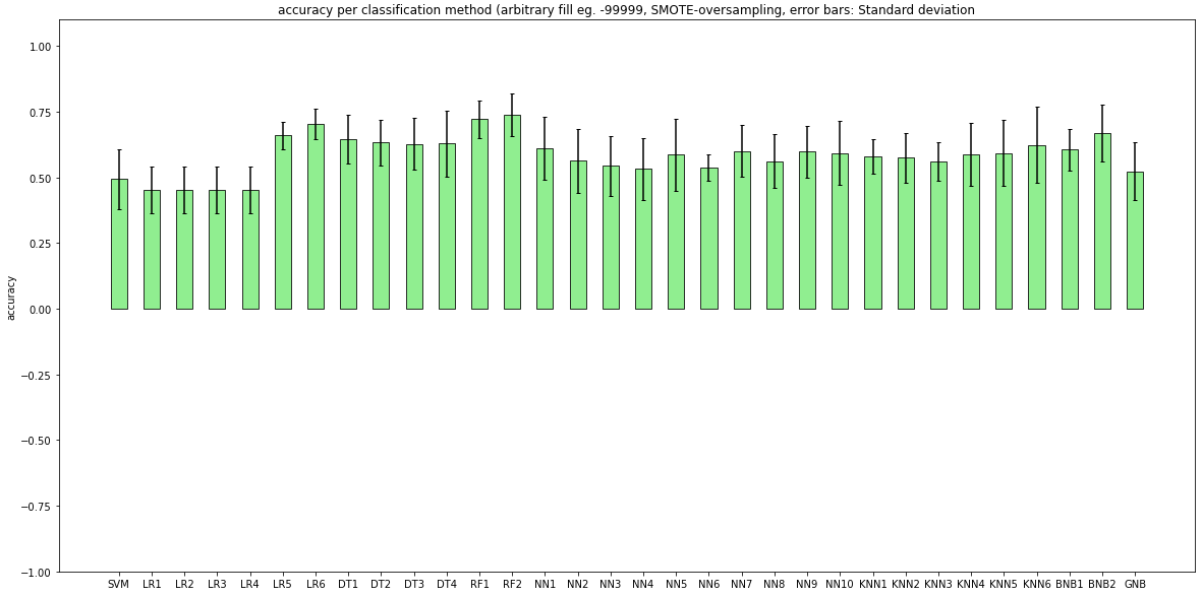

accuracy per classification method (formula fill: -R*(log10(R)+100), undersampling, error bars: Standard deviation
accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

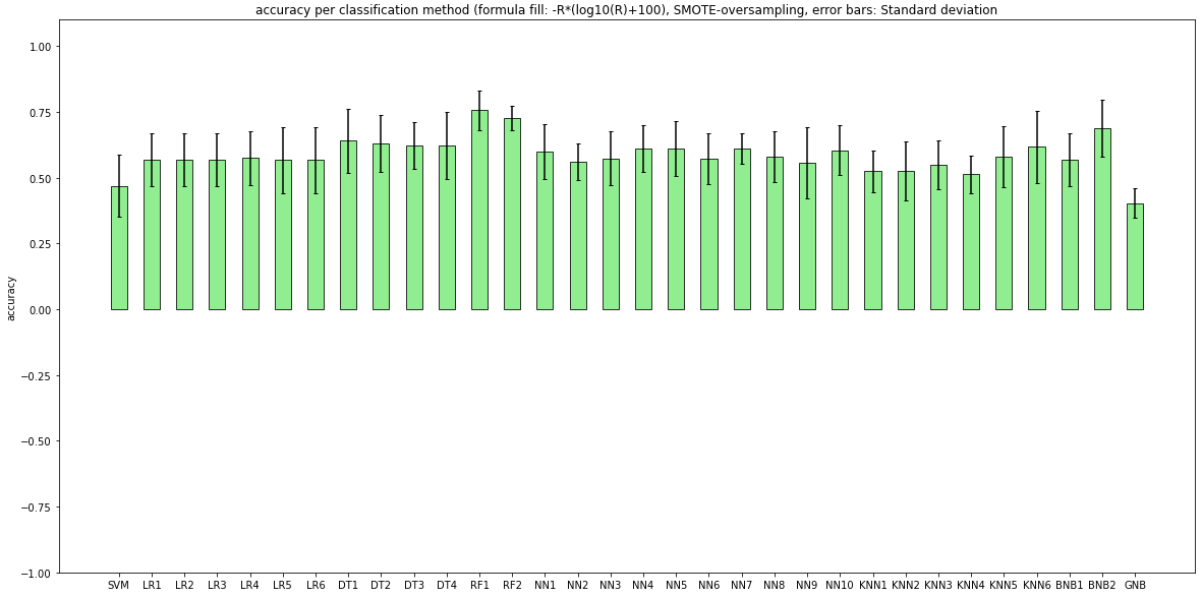

accuracy per classification method (Iterative imputation, undersampling, error bars: Standard deviation
accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

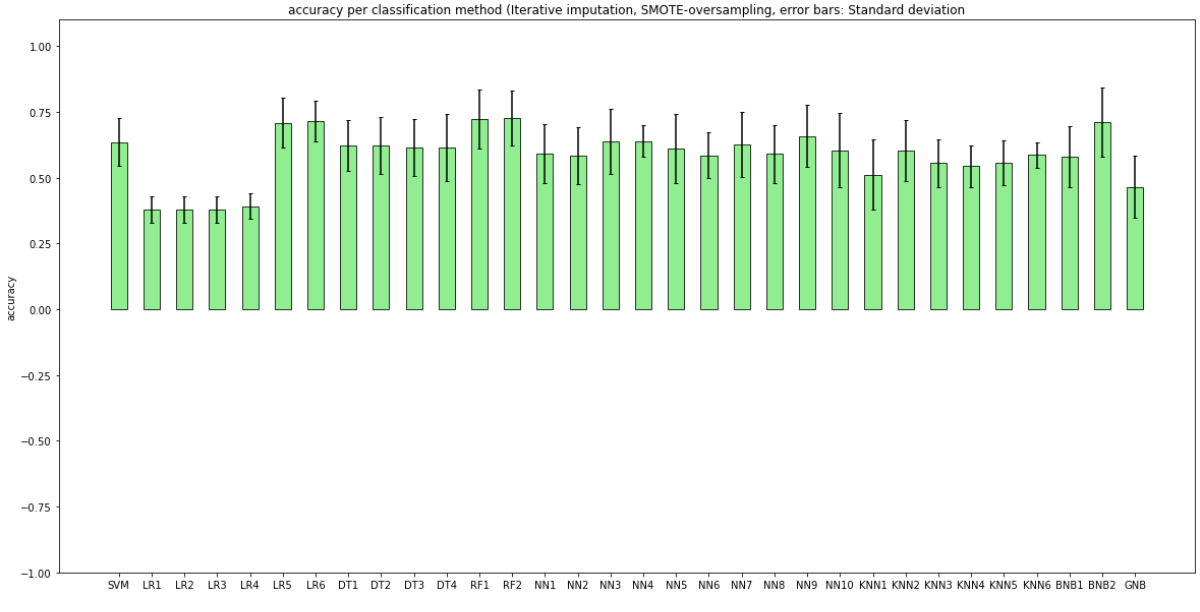

accuracy per classification method (arbitrary fill eg. -99999, class weights(grid search), error bars: Standard deviation
accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM
LR1
LR2
LR3
LR4
LR5
LR6
DT1
DT2
DT3
DT4
RF1
RF2

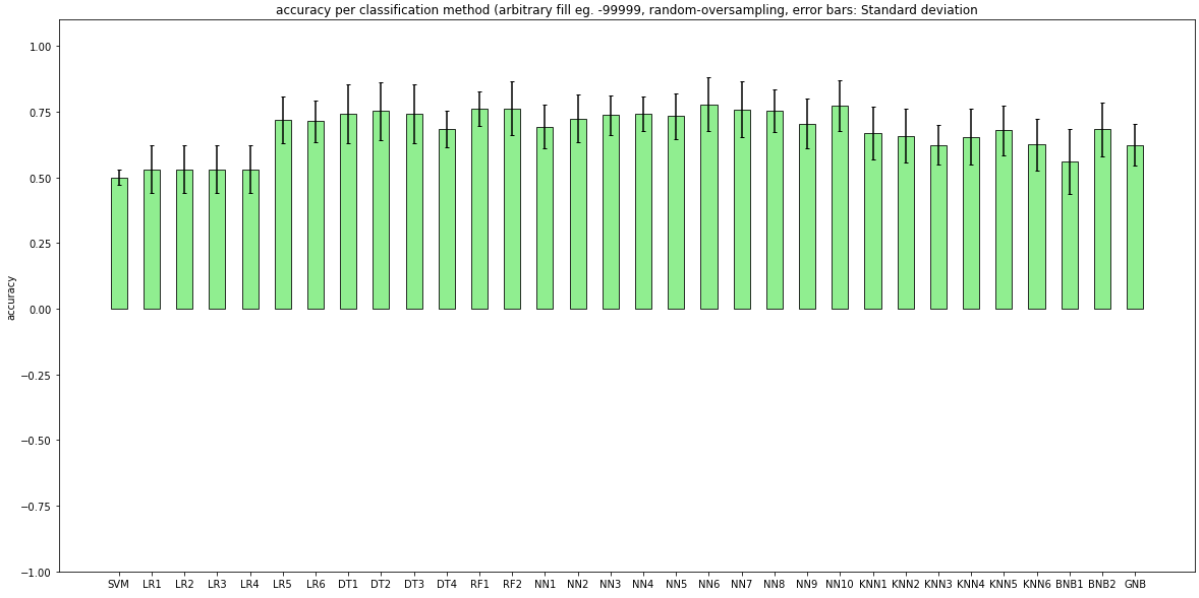

accuracy per classification method (formula fill: -R*(log10(R)+100), class weights(grid search), error bars: Standard deviation
accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM
LR1
LR2
LR3
LR4
LR5
LR6
DT1
DT2
DT3
DT4
RF1
RF2

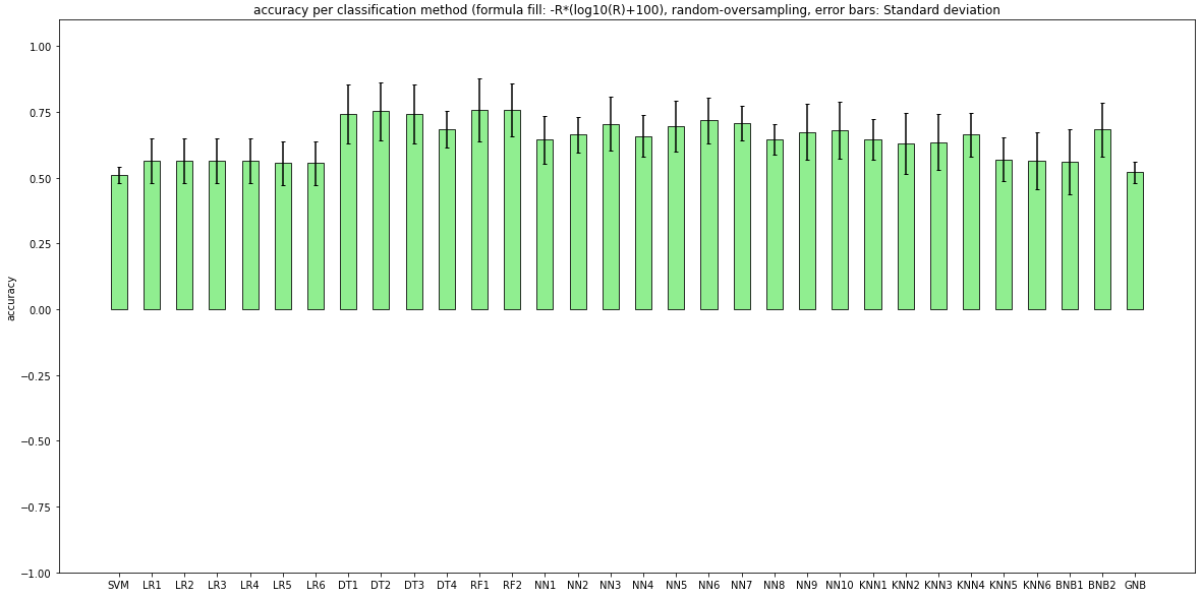

accuracy per classification method (Iterative imputation, class weights(grid search), error bars: Standard deviation
accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM
LR1
LR2
LR3
LR4
LR5
LR6
DT1
DT2
DT3
DT4
RF1
RF2

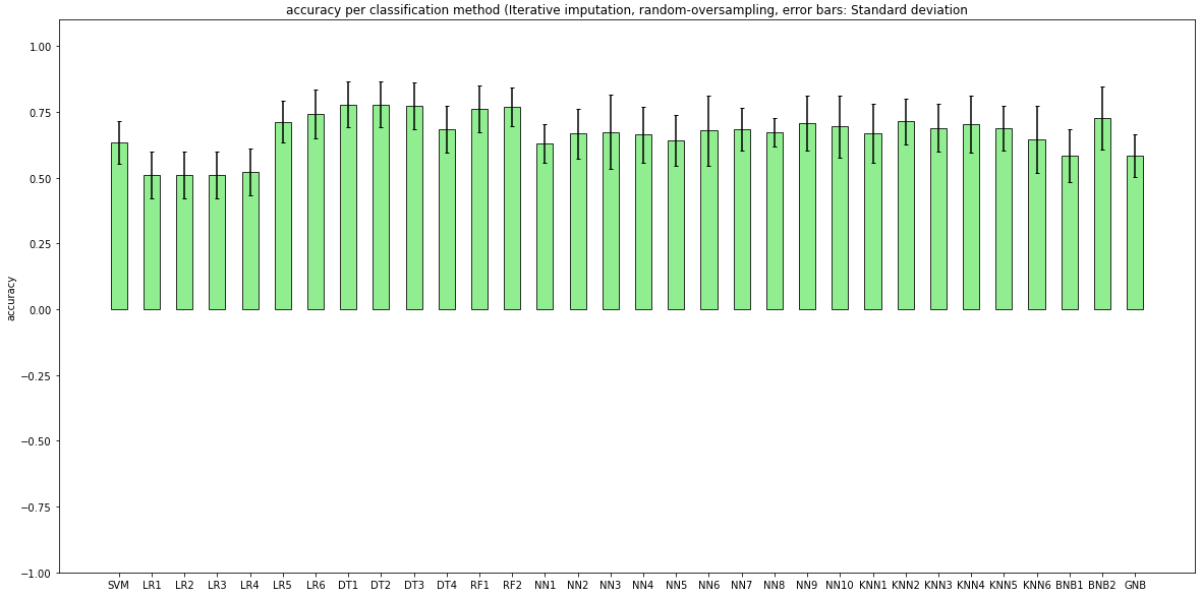

accuracy per classification method (arbitrary fill eg. -99999, SMOTE-oversampling, error bars: Standard deviation
accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

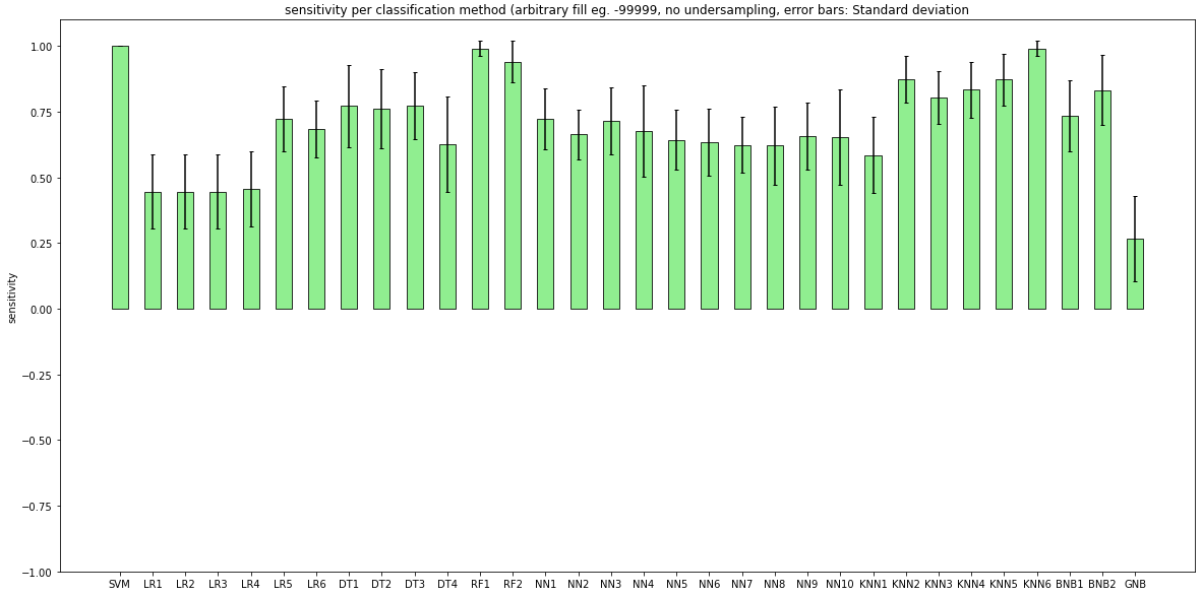

accuracy per classification method (formula fill: -R*(log10(R)+100), SMOTE-oversampling, error bars: Standard deviation
accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

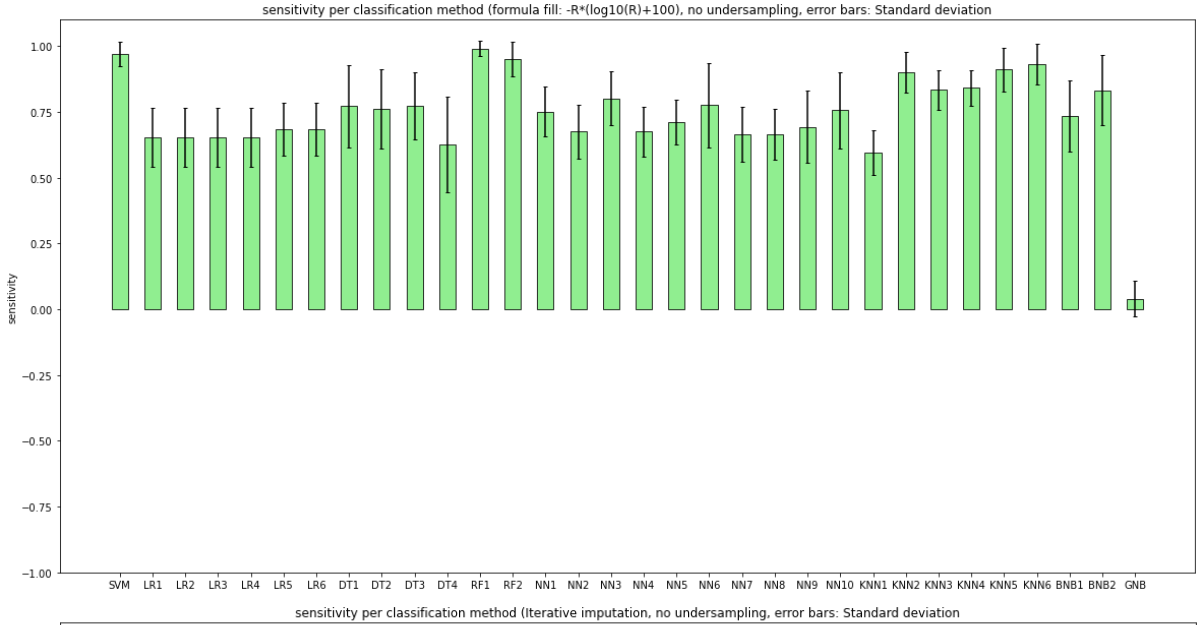

accuracy per classification method (Iterative imputation, SMOTE-oversampling, error bars: Standard deviation
accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

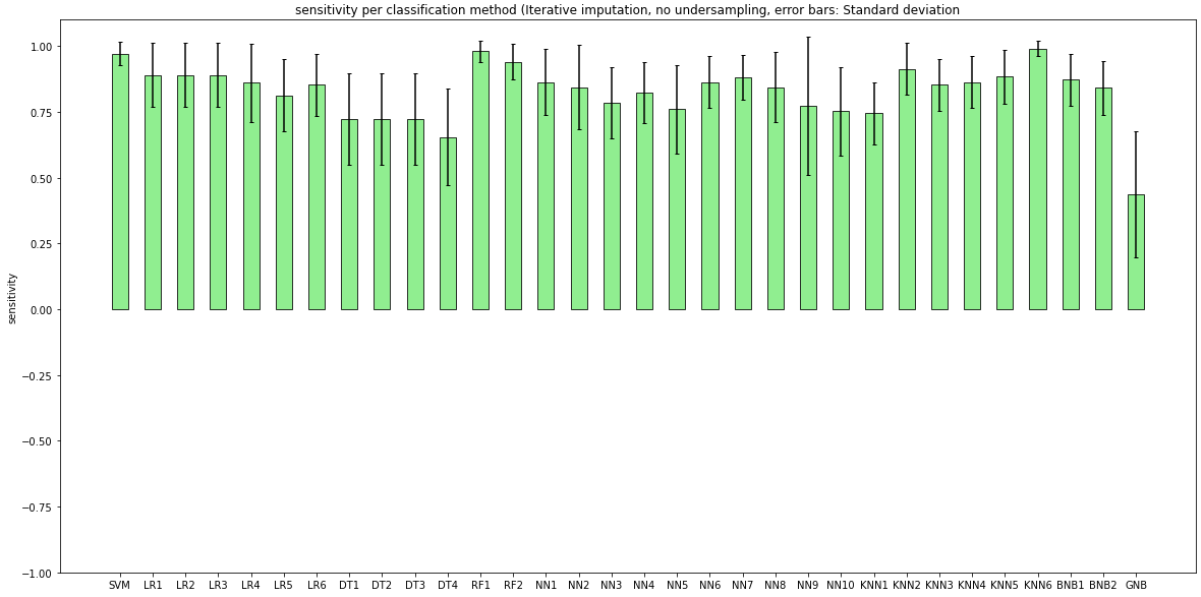

accuracy per classification method (arbitrary fill eg. -99999, random-oversampling, error bars: Standard deviation
accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

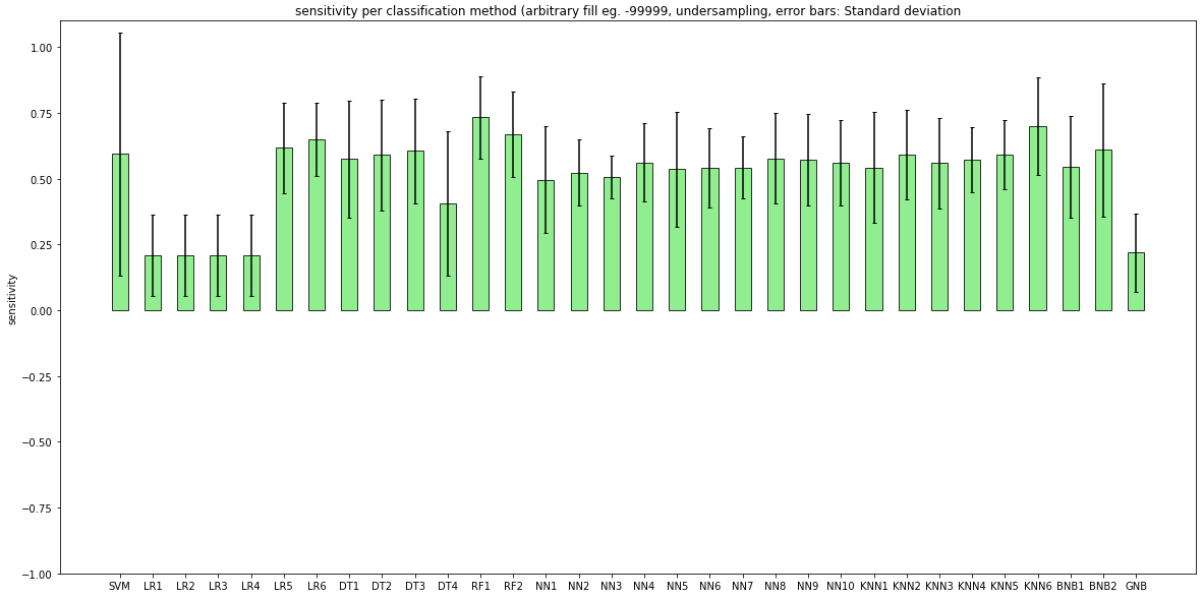

accuracy per classification method (formula fill: -R*(log10(R)+100), random-oversampling, error bars: Standard deviation
accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

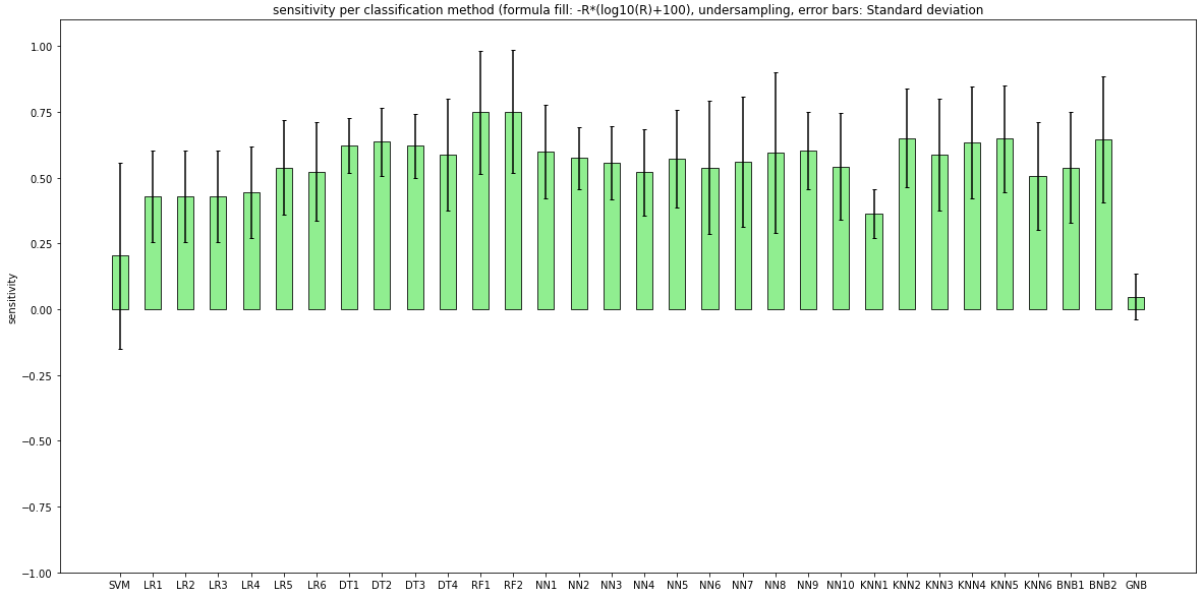

accuracy per classification method (Iterative imputation, random-oversampling, error bars: Standard deviation
accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

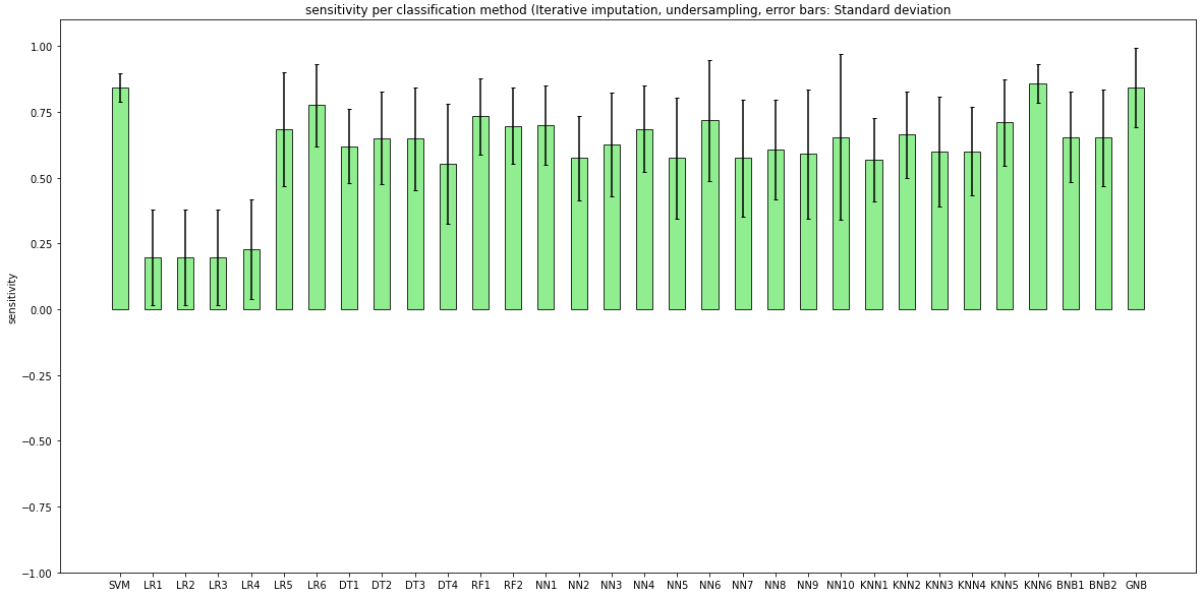

sensitivity per classification method (arbitrary fill eg. -99999, no undersampling, error bars: Standard deviation
sensitivity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

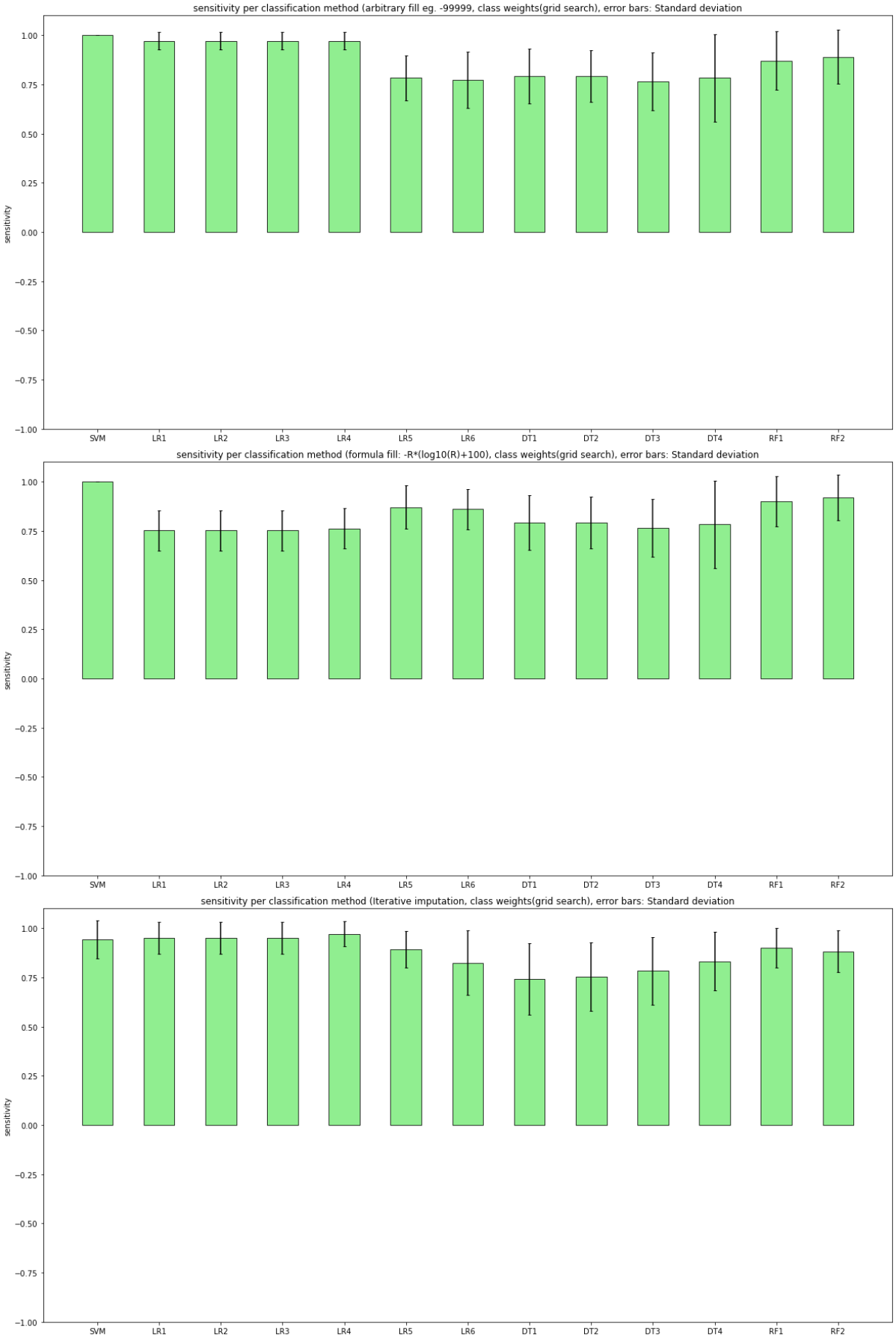

sensitivity per classification method (formula fill: -R*(log10(R)+100), no undersampling, error bars: Standard deviation
sensitivity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

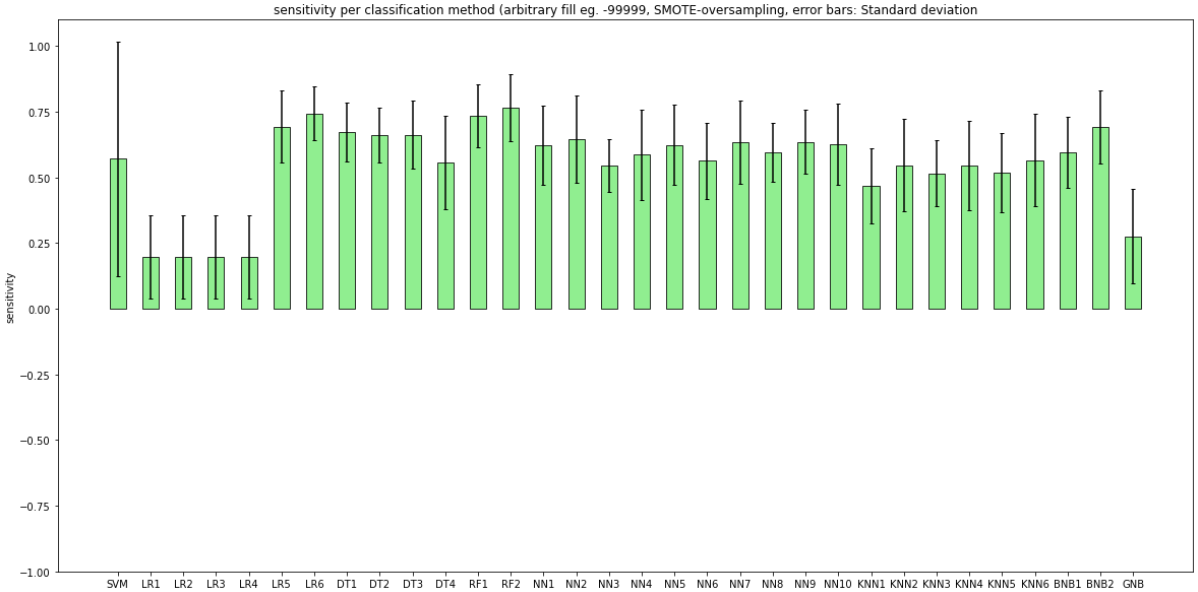

sensitivity per classification method (Iterative imputation, no undersampling, error bars: Standard deviation
sensitivity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

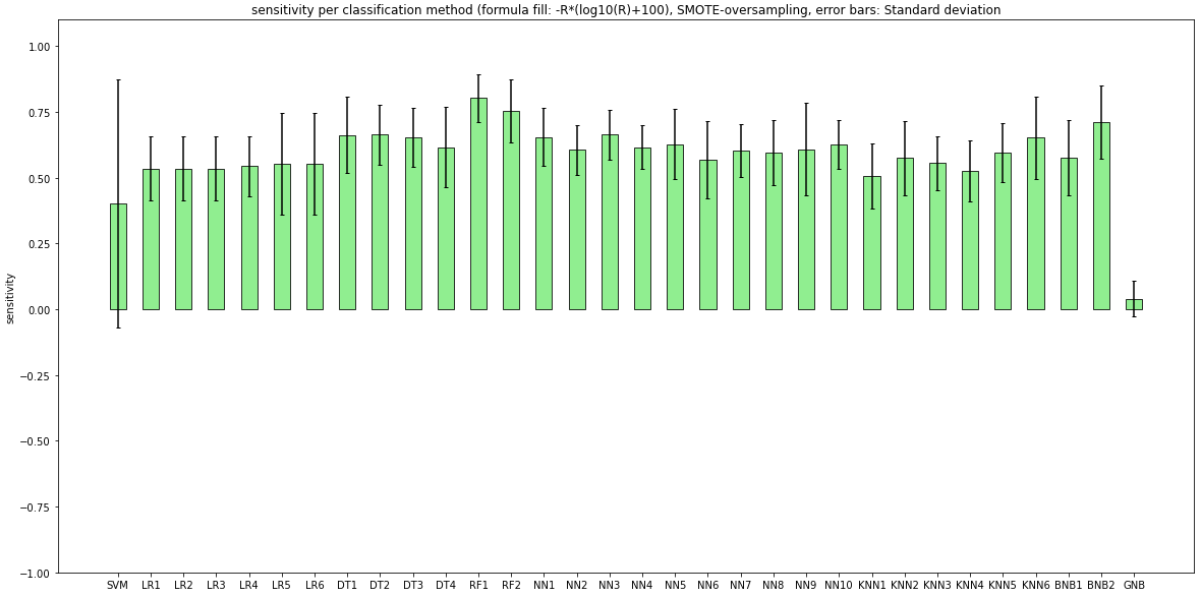

sensitivity per classification method (arbitrary fill eg. -99999, undersampling, error bars: Standard deviation
sensitivity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

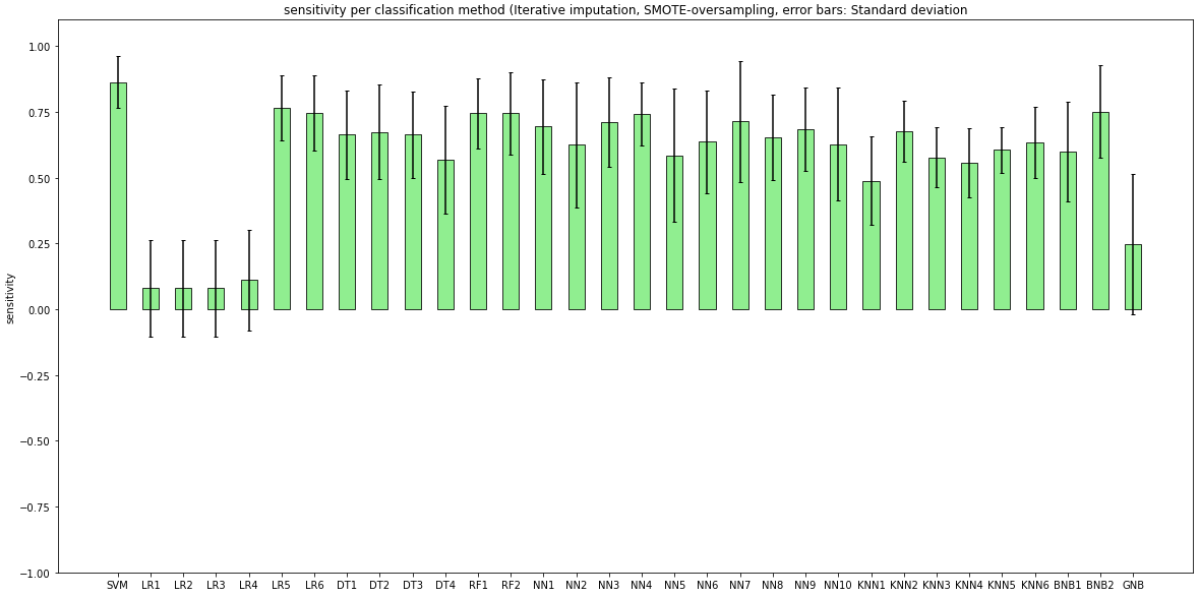

sensitivity per classification method (formula fill: -R*(log10(R)+100), undersampling, error bars: Standard deviation
sensitivity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

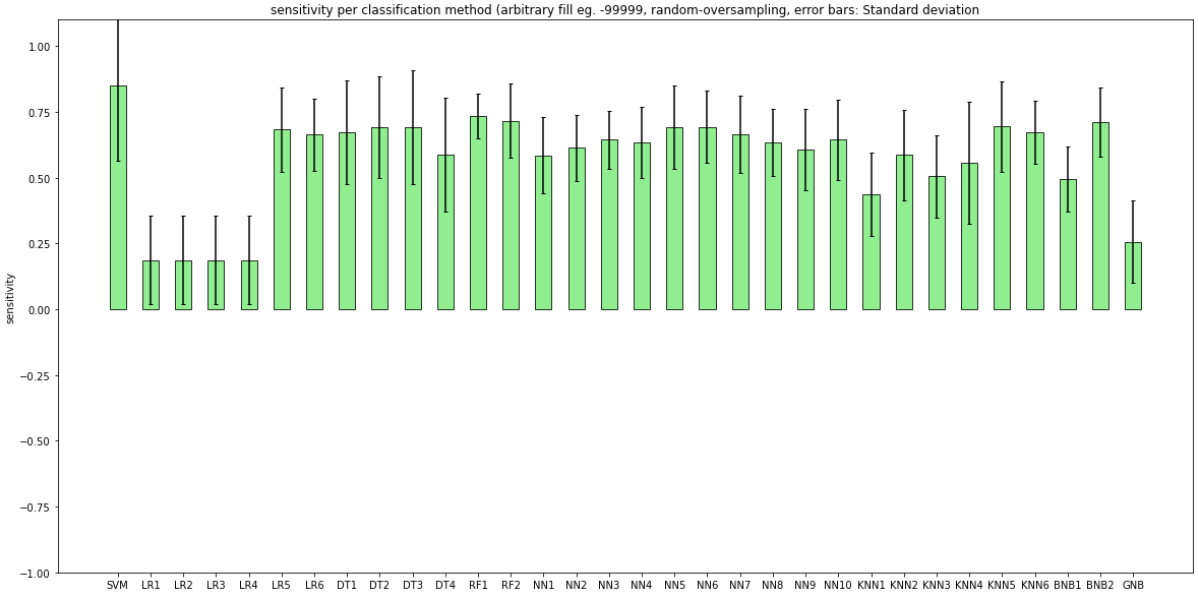

sensitivity per classification method (Iterative imputation, undersampling, error bars: Standard deviation
sensitivity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

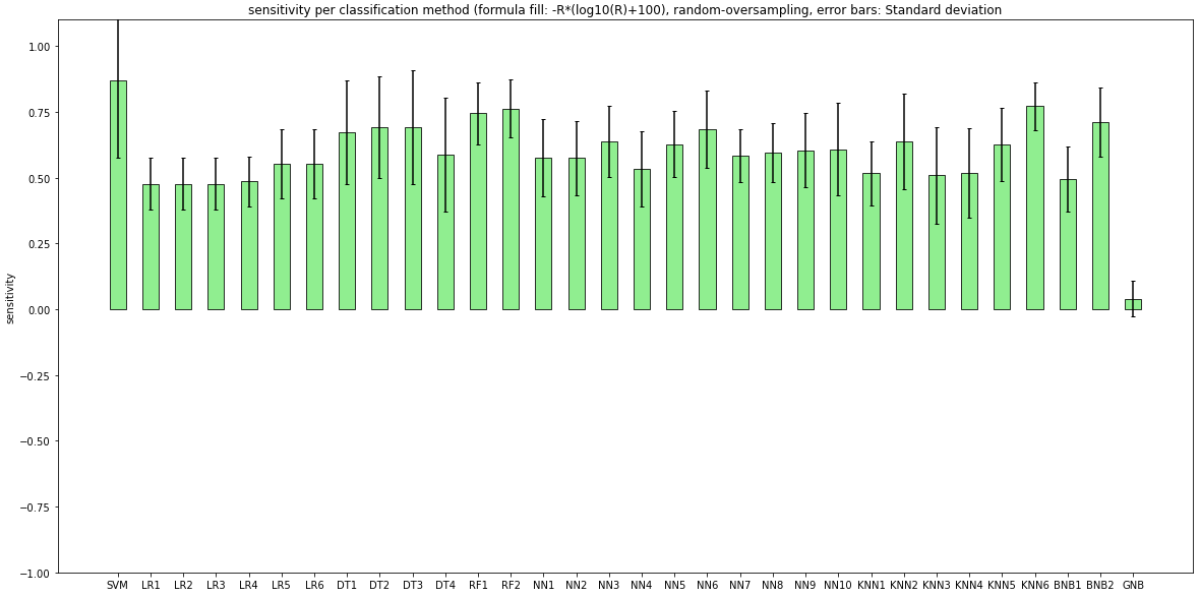

sensitivity per classification method (arbitrary fill eg. -99999, class weights(grid search), error bars: Standard deviation
sensitivity per classification method (formula fill: -R*(log10(R)+100), class weights(grid search), error bars: Standard deviation
sensitivity per classification method (Iterative imputation, class weights(grid search), error bars: Standard deviation
sensitivity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM
LR1
LR2
LR3
LR4
LR5
LR6
DT1
DT2
DT3
DT4
RF1
RF2

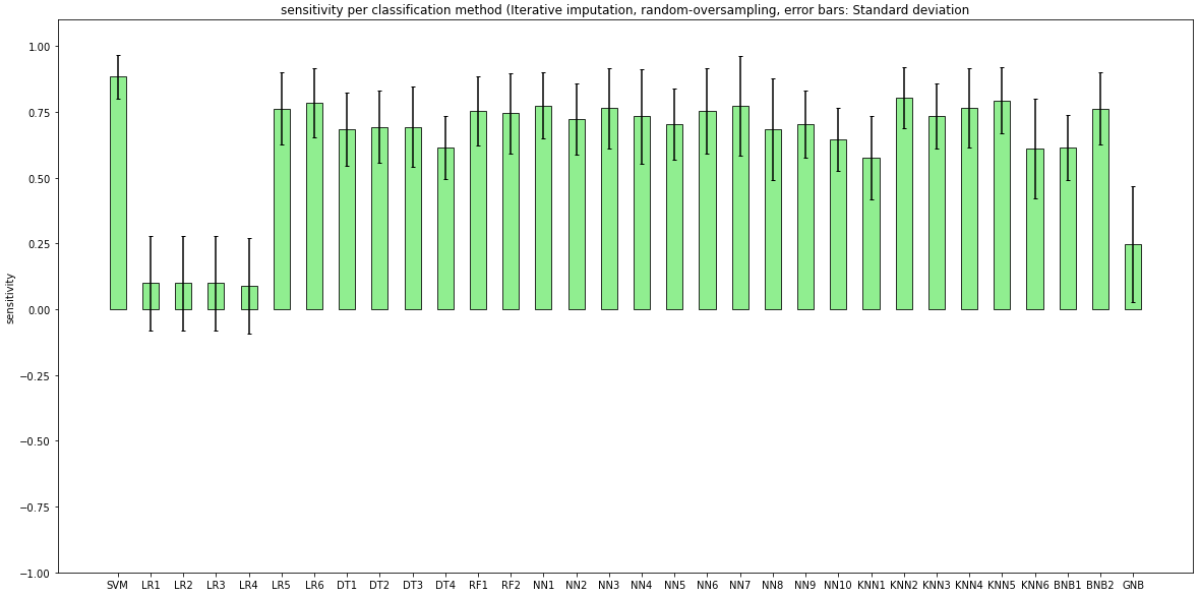

sensitivity per classification method (arbitrary fill eg. -99999, SMOTE-oversampling, error bars: Standard deviation
sensitivity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

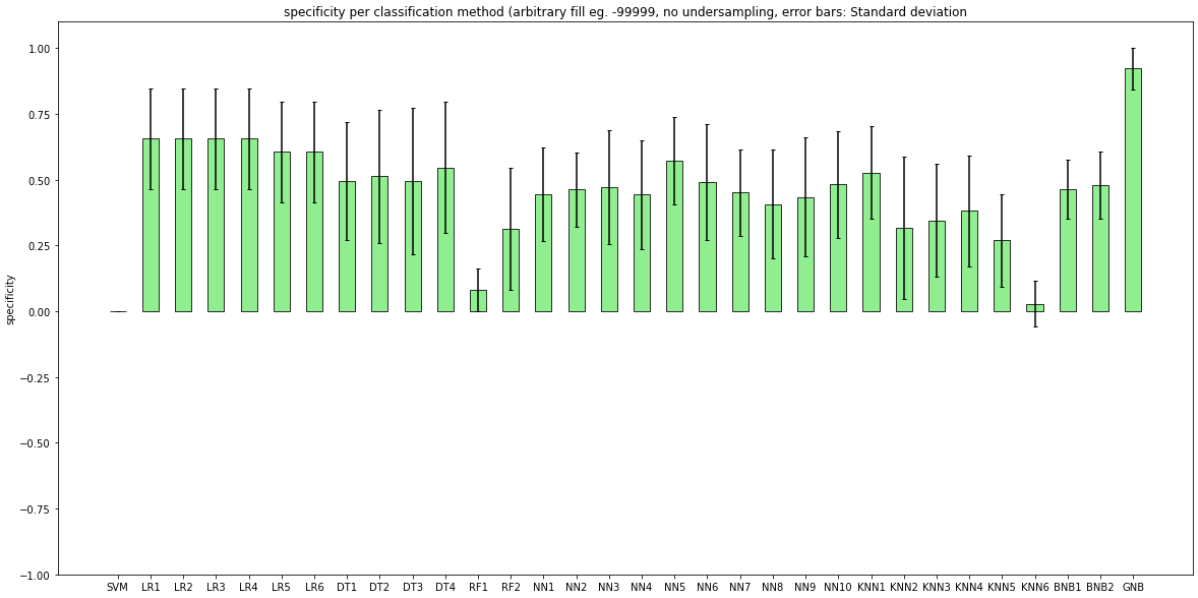

sensitivity per classification method (formula fill: -R*(log10(R)+100), SMOTE-oversampling, error bars: Standard deviation
sensitivity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

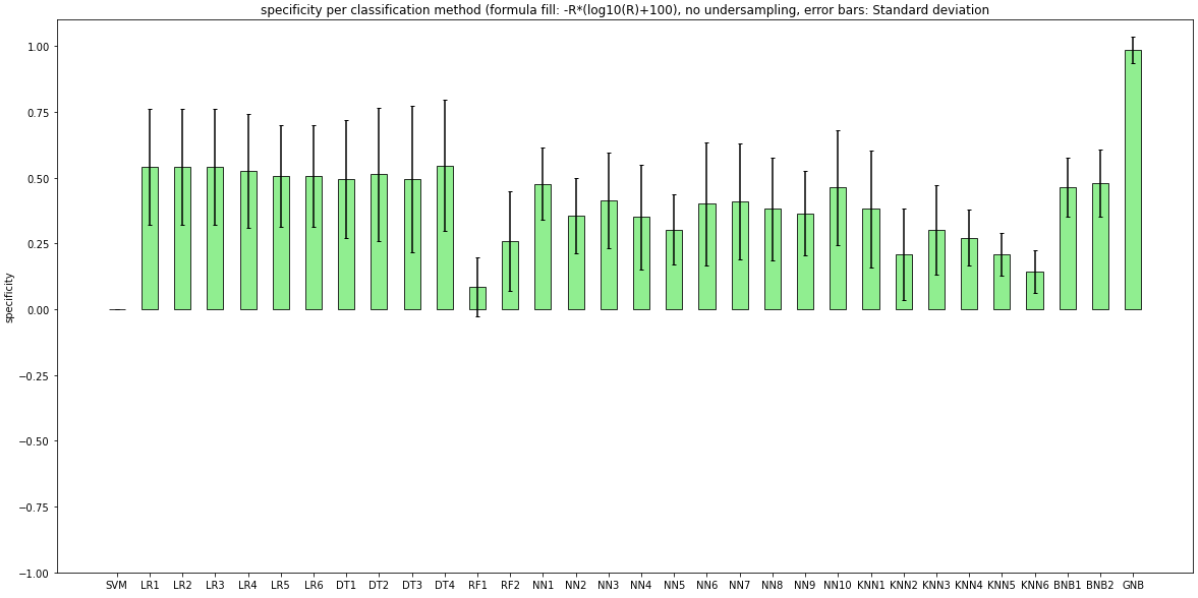

sensitivity per classification method (Iterative imputation, SMOTE-oversampling, error bars: Standard deviation
sensitivity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

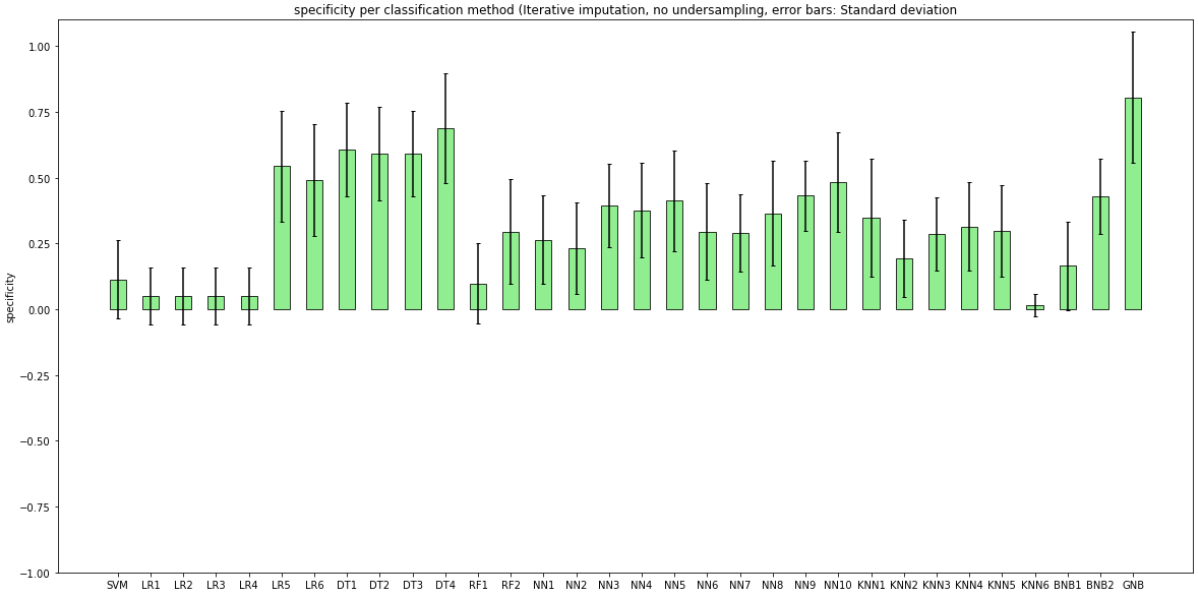

sensitivity per classification method (arbitrary fill eg. -99999, random-oversampling, error bars: Standard deviation
sensitivity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

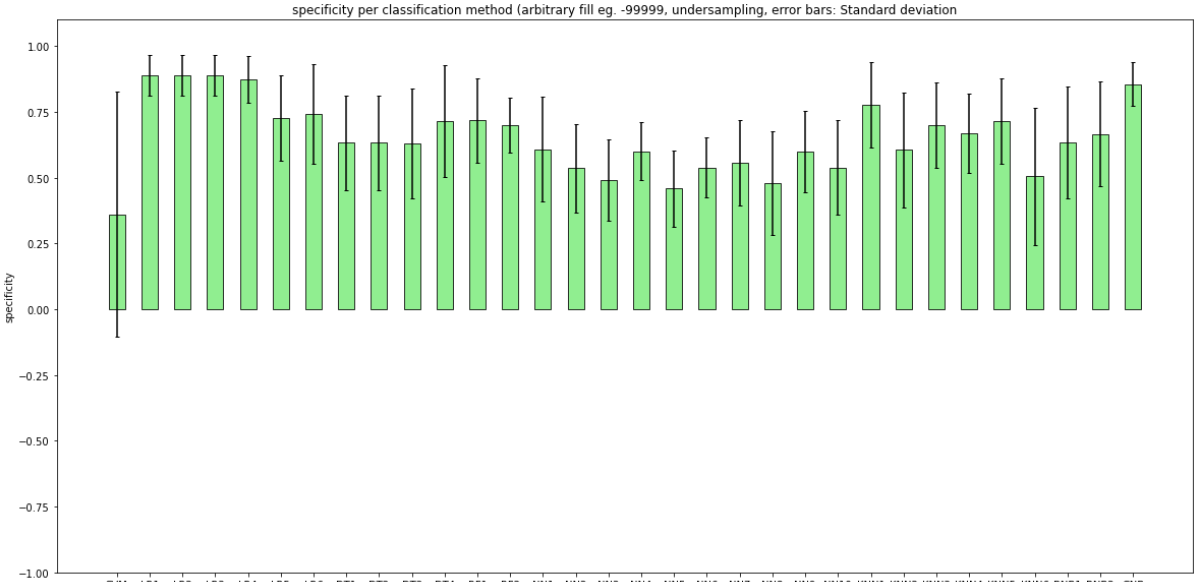

sensitivity per classification method (formula fill: -R*(log10(R)+100), random-oversampling, error bars: Standard deviation
sensitivity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

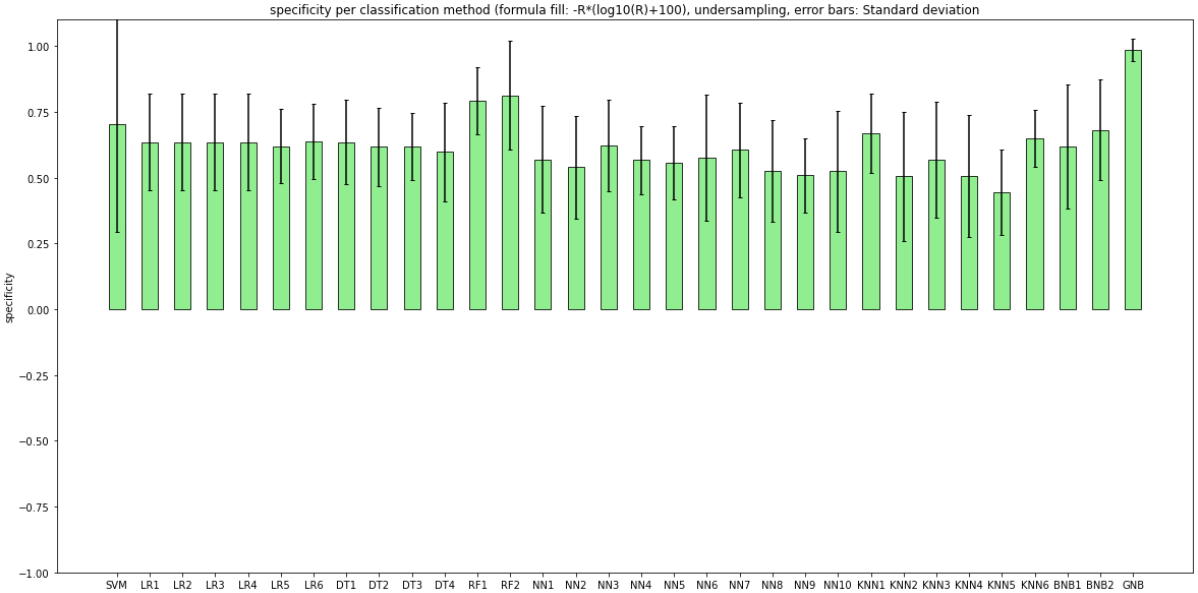

sensitivity per classification method (Iterative imputation, random-oversampling, error bars: Standard deviation
sensitivity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

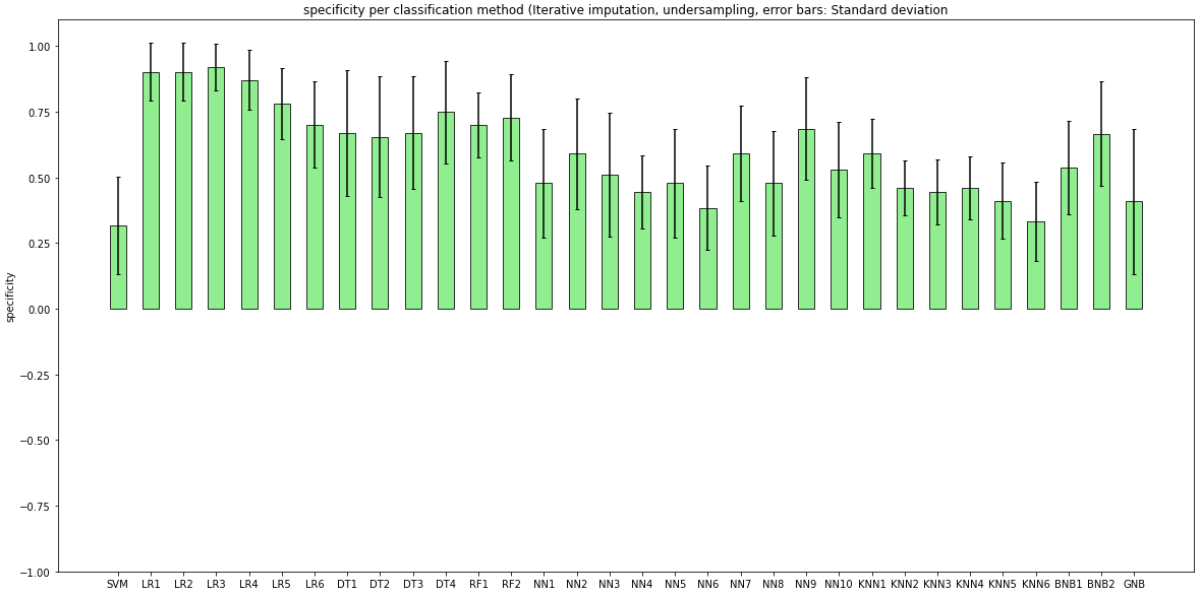

specificity per classification method (arbitrary fill eg. -99999, no undersampling, error bars: Standard deviation
specificity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

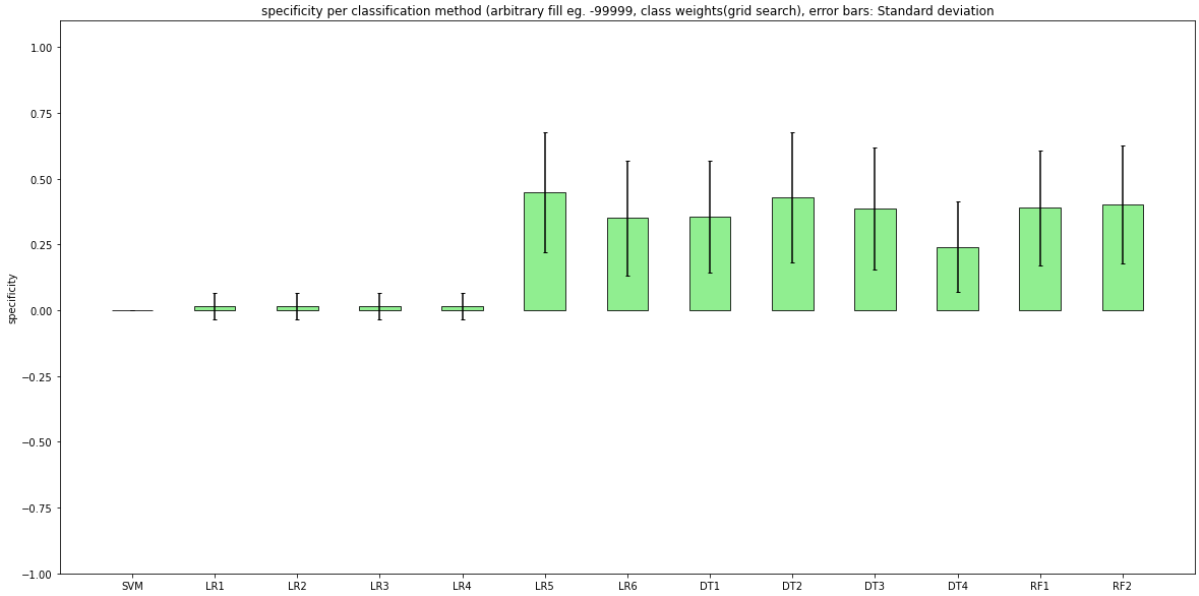

specificity per classification method (formula fill: -R*(log10(R)+100), no undersampling, error bars: Standard deviation
specificity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

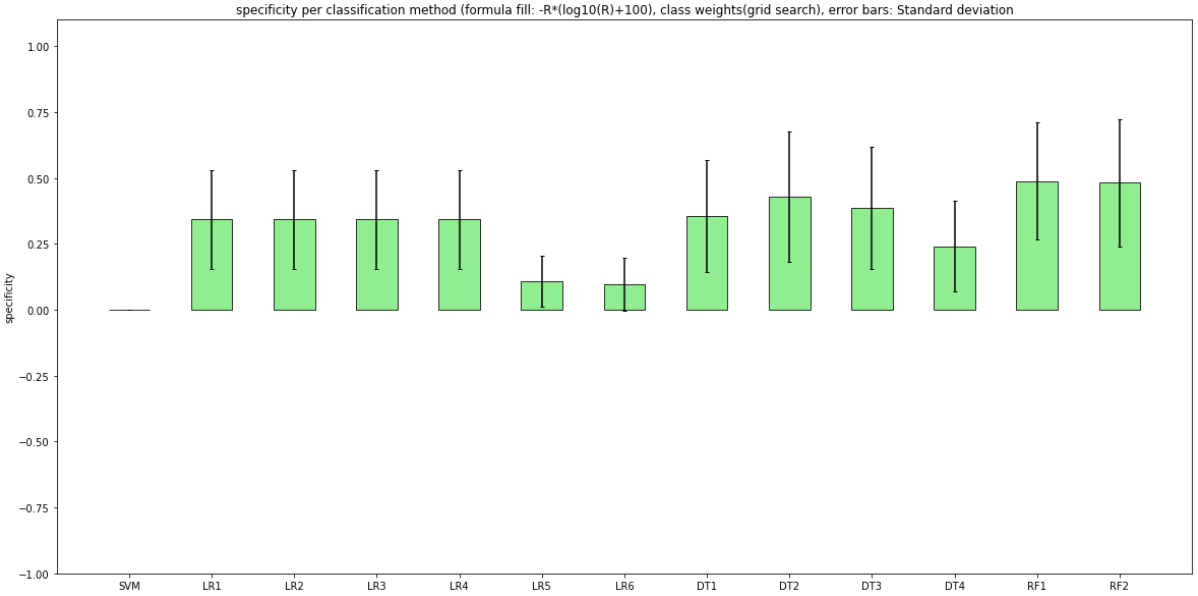

specificity per classification method (Iterative imputation, no undersampling, error bars: Standard deviation
specificity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

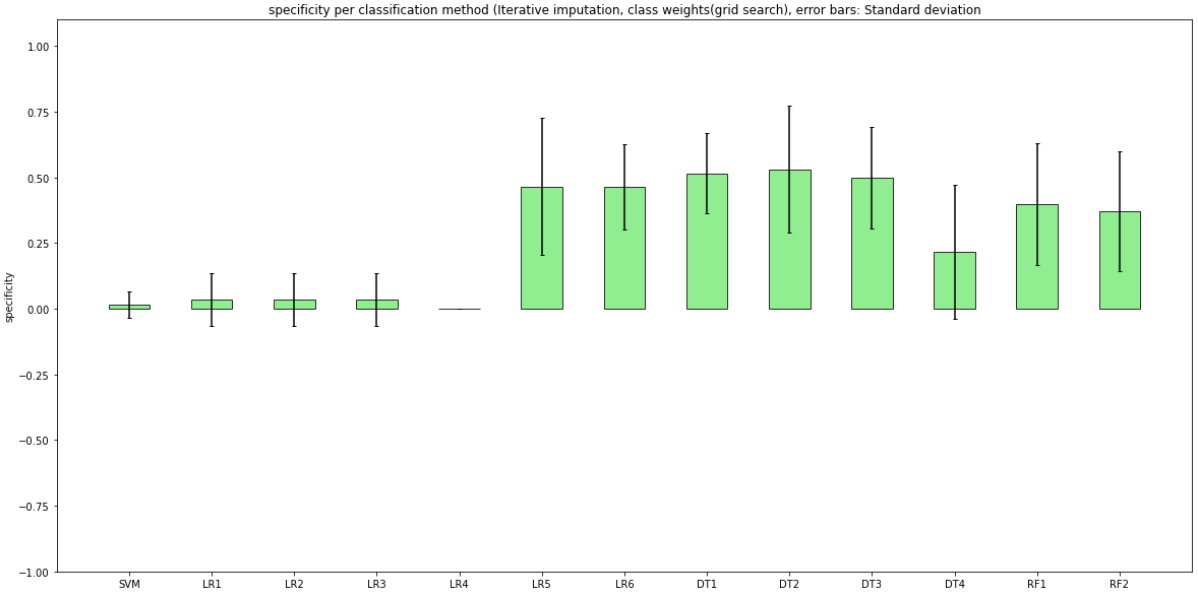

specificity per classification method (arbitrary fill eg. -99999, undersampling, error bars: Standard deviation
specificity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

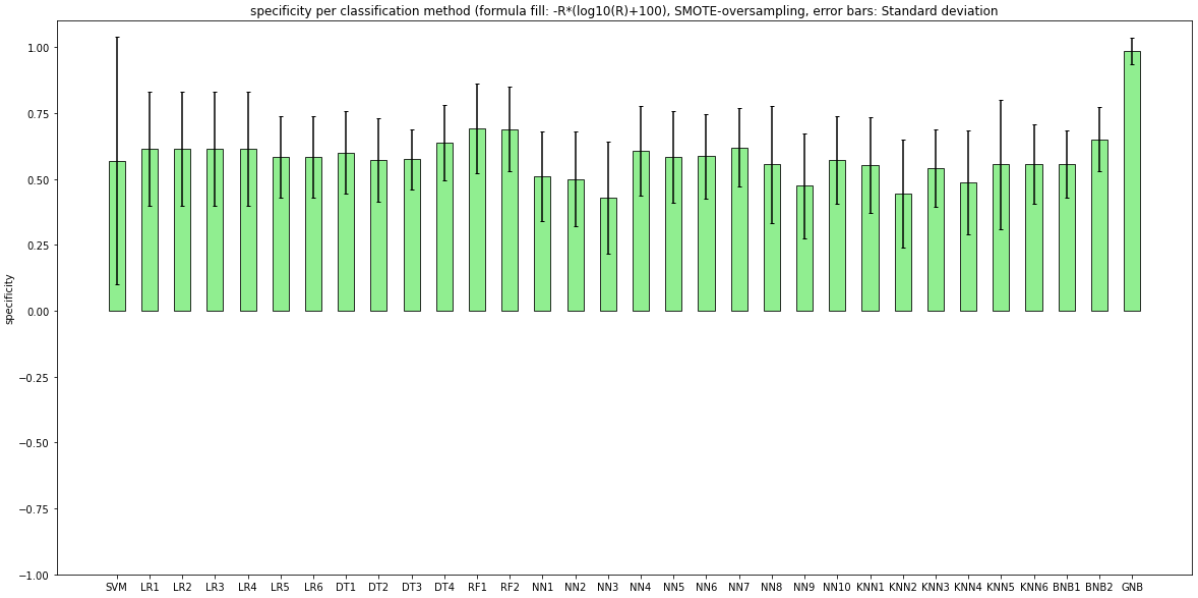

specificity per classification method (formula fill: -R*(log10(R)+100), undersampling, error bars: Standard deviation
specificity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

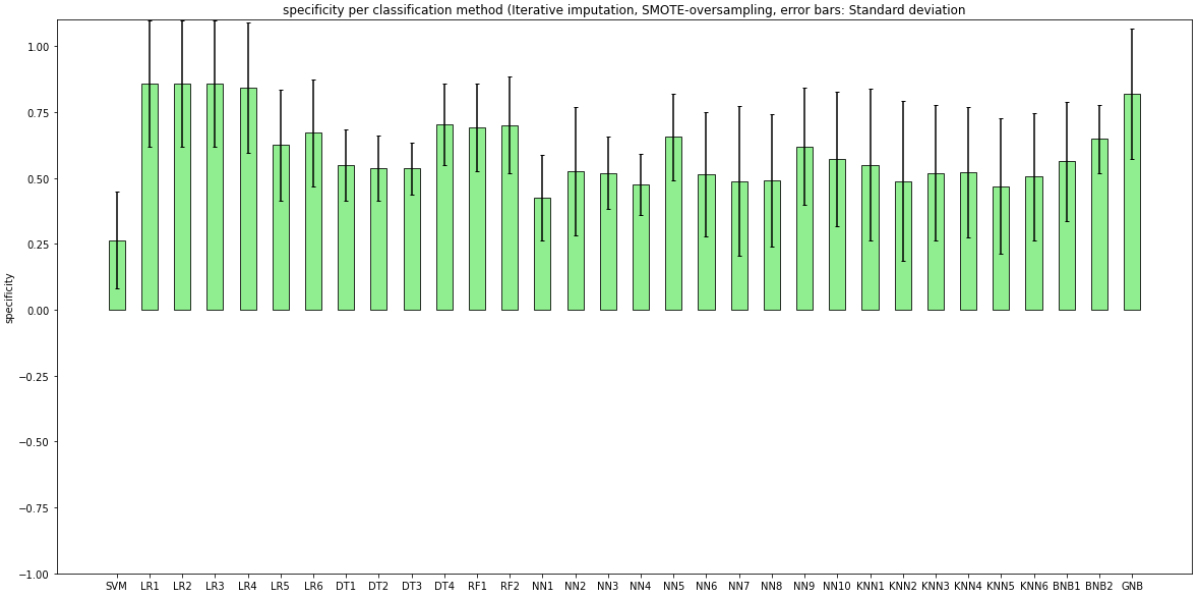

specificity per classification method (Iterative imputation, undersampling, error bars: Standard deviation
specificity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

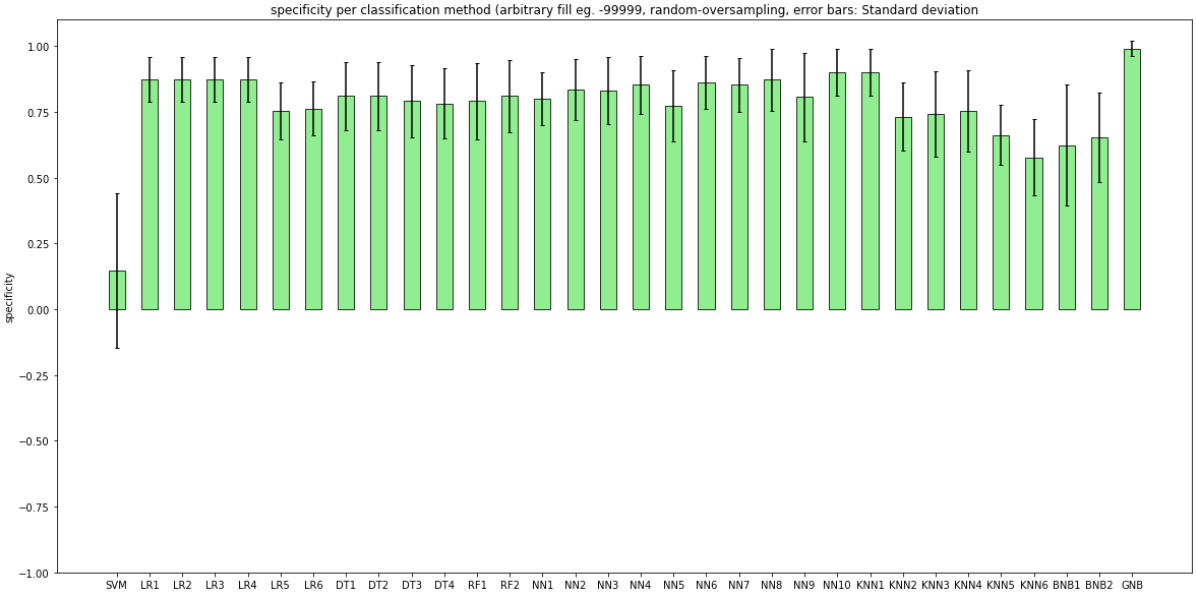

specificity per classification method (arbitrary fill eg. -99999, class weights(grid search), error bars: Standard deviation
specificity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM
LR1
LR2
LR3
LR4
LR5
LR6
DT1
DT2
DT3
DT4
RF1
RF2

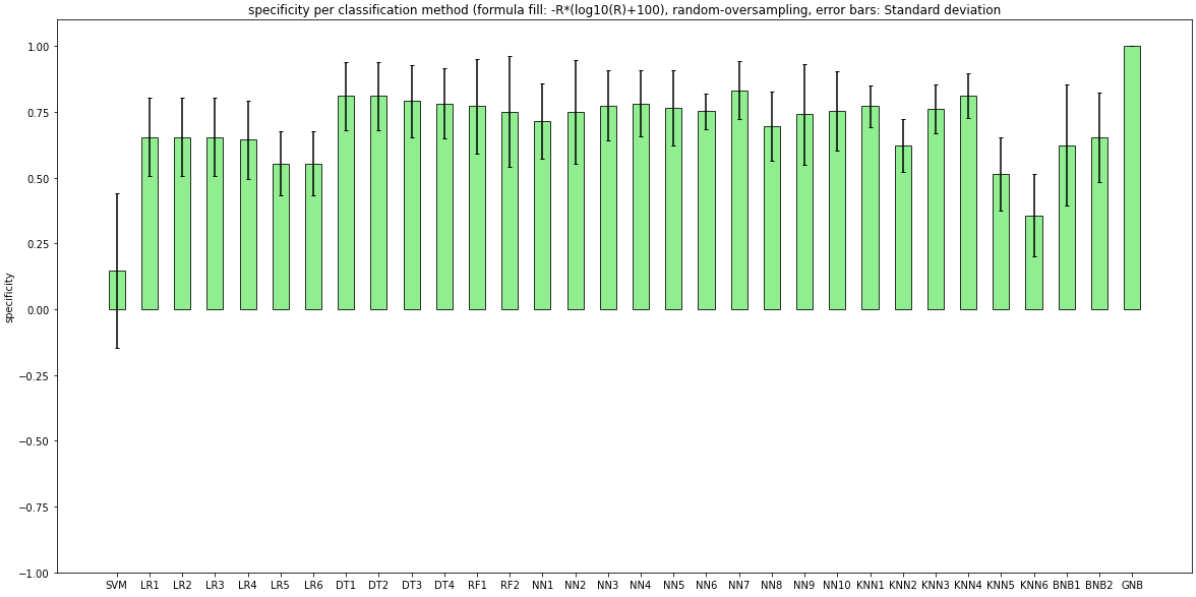

specificity per classification method (formula fill: -R*(log10(R)+100), class weights(grid search), error bars: Standard deviation
specificity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM
LR1
LR2
LR3
LR4
LR5
LR6
DT1
DT2
DT3
DT4
RF1
RF2

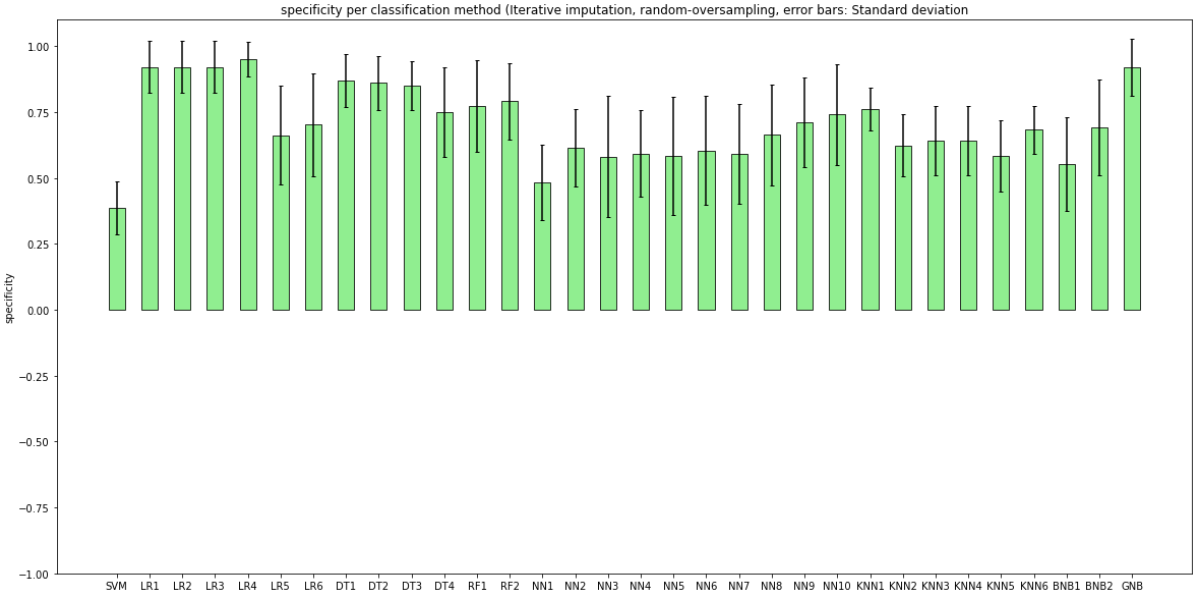

specificity per classification method (Iterative imputation, class weights(grid search), error bars: Standard deviation
specificity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM
LR1
LR2
LR3
LR4
LR5
LR6
DT1
DT2
DT3
DT4
RF1
RF2

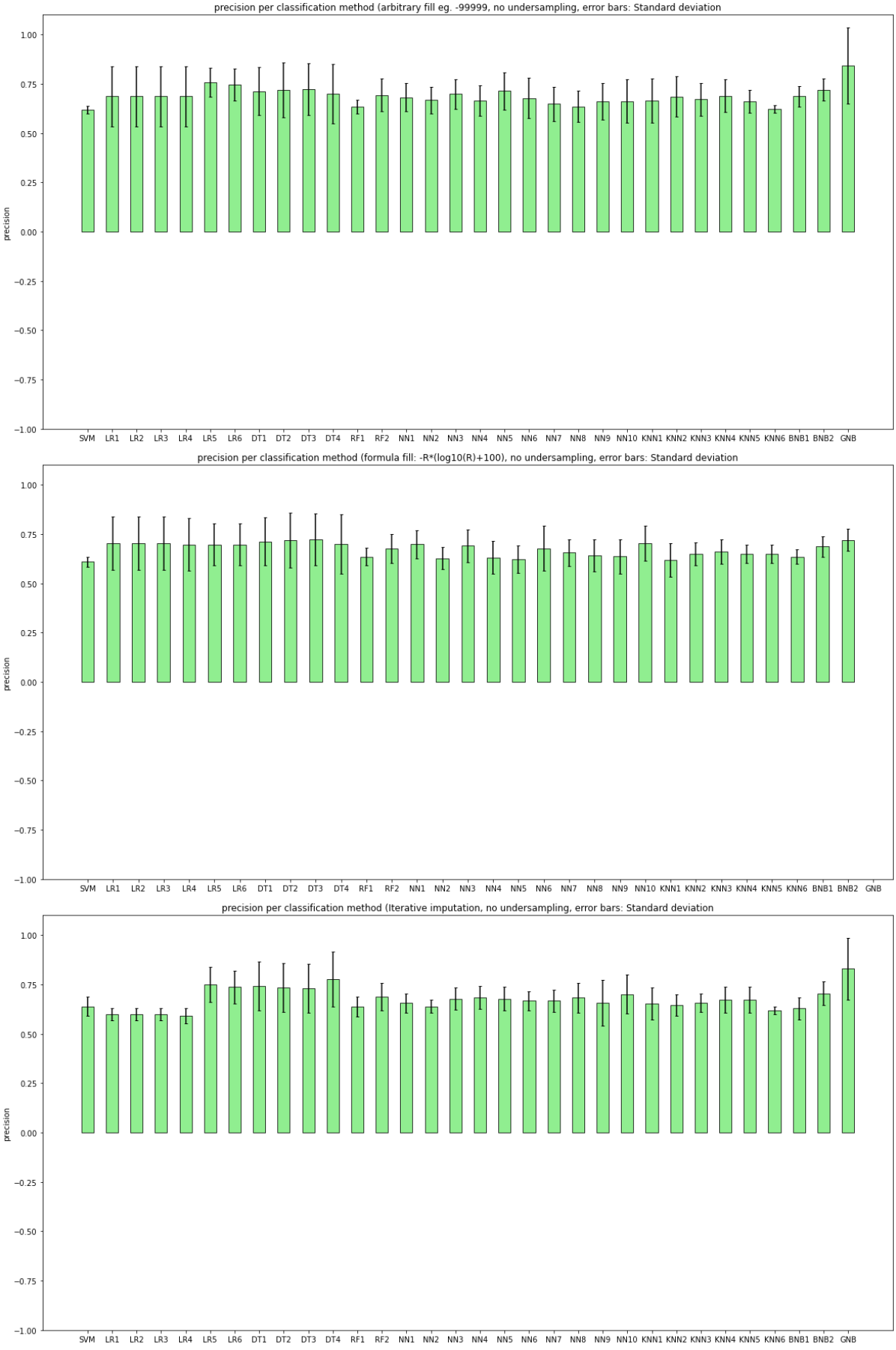

specificity per classification method (formula fill: -R*(log10(R)+100), SMOTE-oversampling, error bars: Standard deviation
specificity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

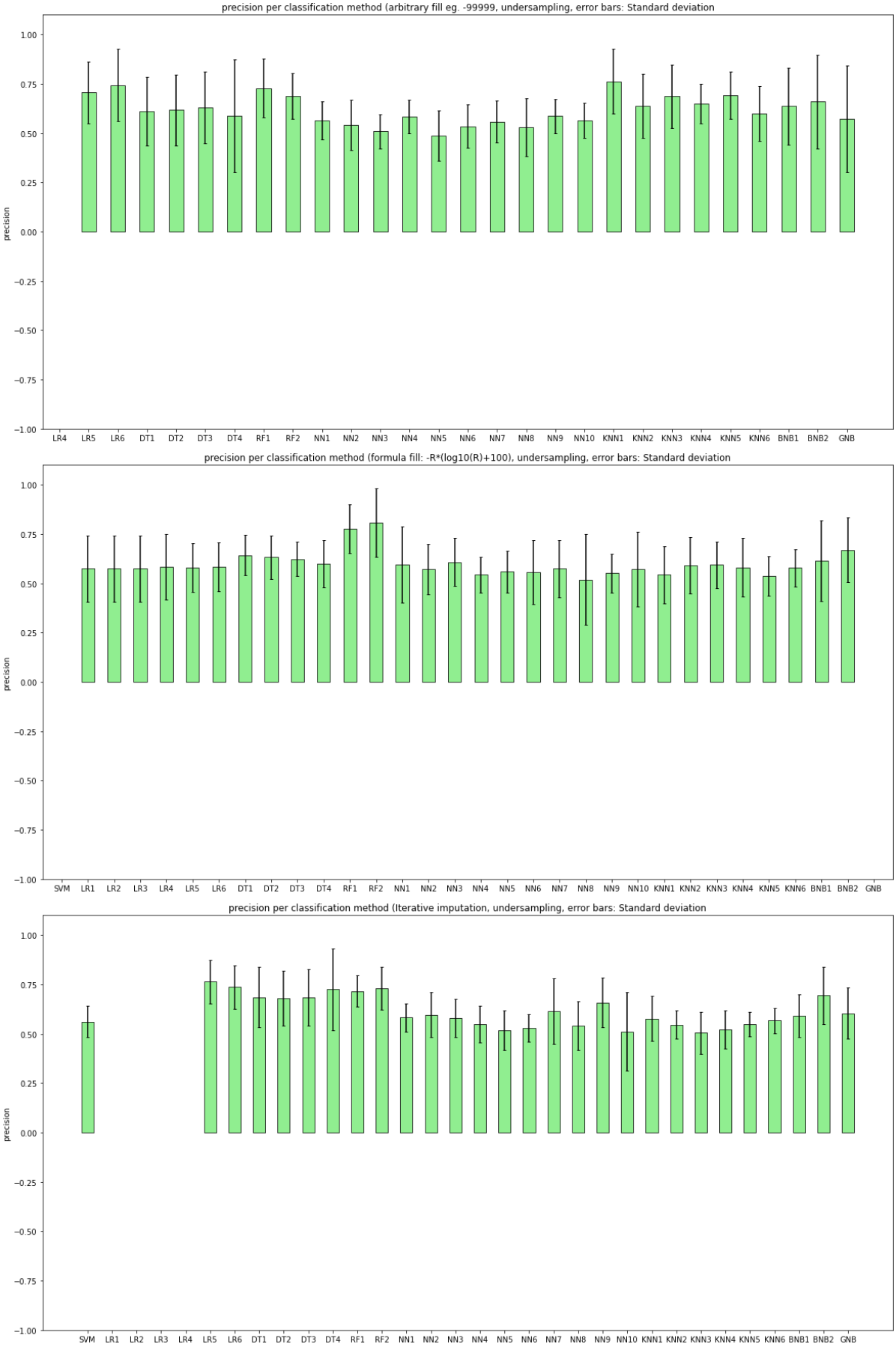

specificity per classification method (Iterative imputation, SMOTE-oversampling, error bars: Standard deviation
specificity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

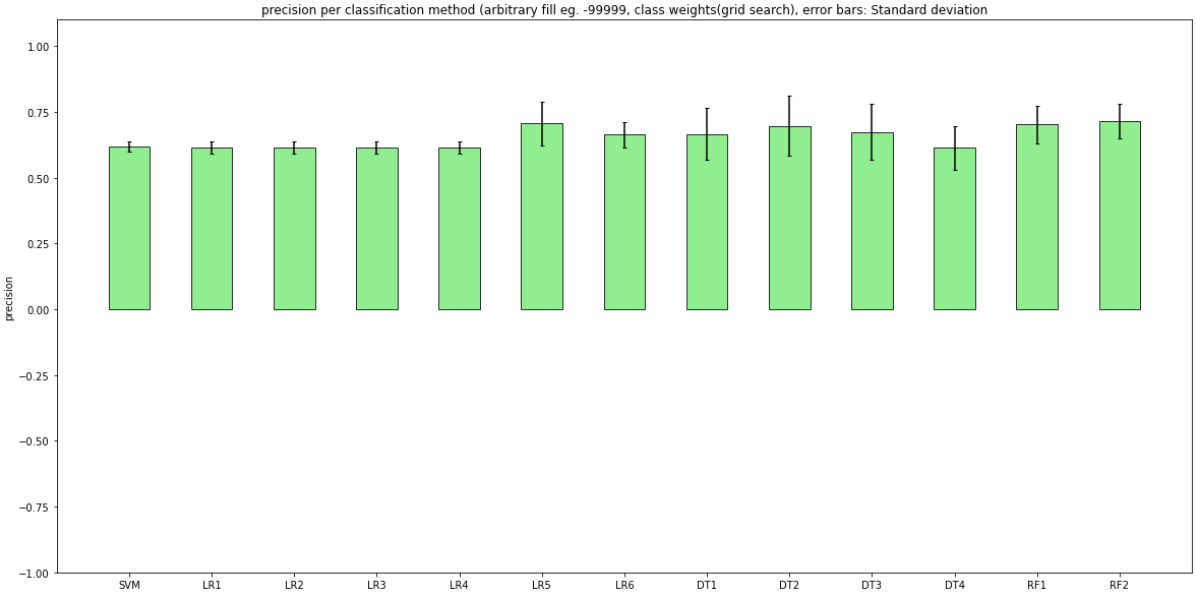

specificity per classification method (arbitrary fill eg. -99999, random-oversampling, error bars: Standard deviation
specificity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

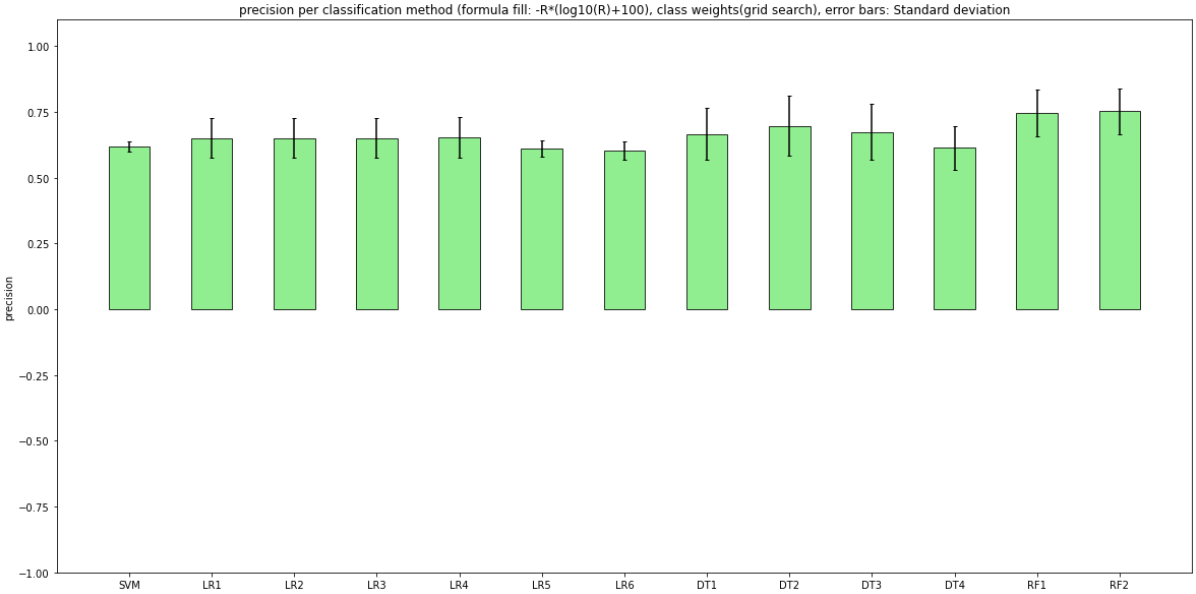

specificity per classification method (formula fill: -R*(log10(R)+100), random-oversampling, error bars: Standard deviation
specificity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

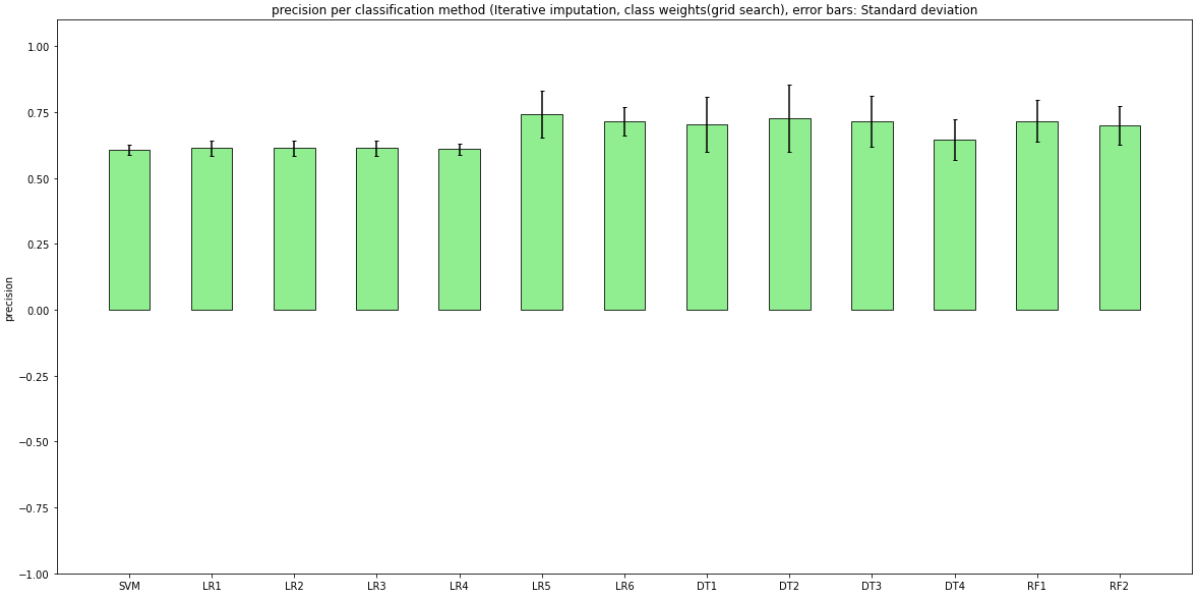

specificity per classification method (Iterative imputation, random-oversampling, error bars: Standard deviation
specificity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

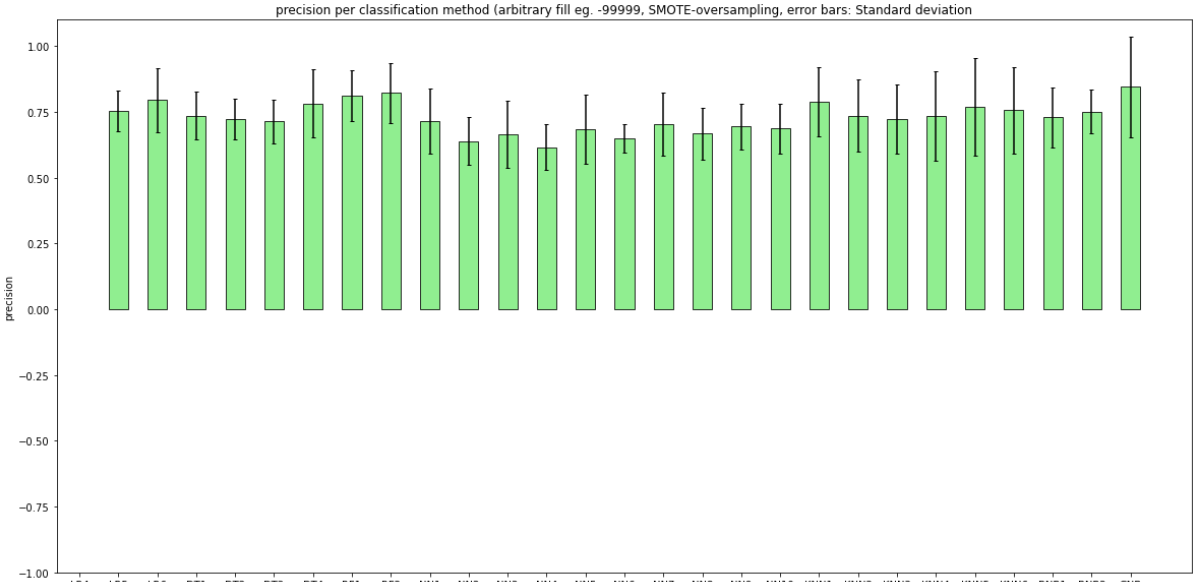

precision per classification method (arbitrary fill eg. -99999, no undersampling, error bars: Standard deviation
precision per classification method (formula fill: -R*(log10(R)+100), no undersampling, error bars: Standard deviation
precision per classification method (Iterative imputation, no undersampling, error bars: Standard deviation
precision
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

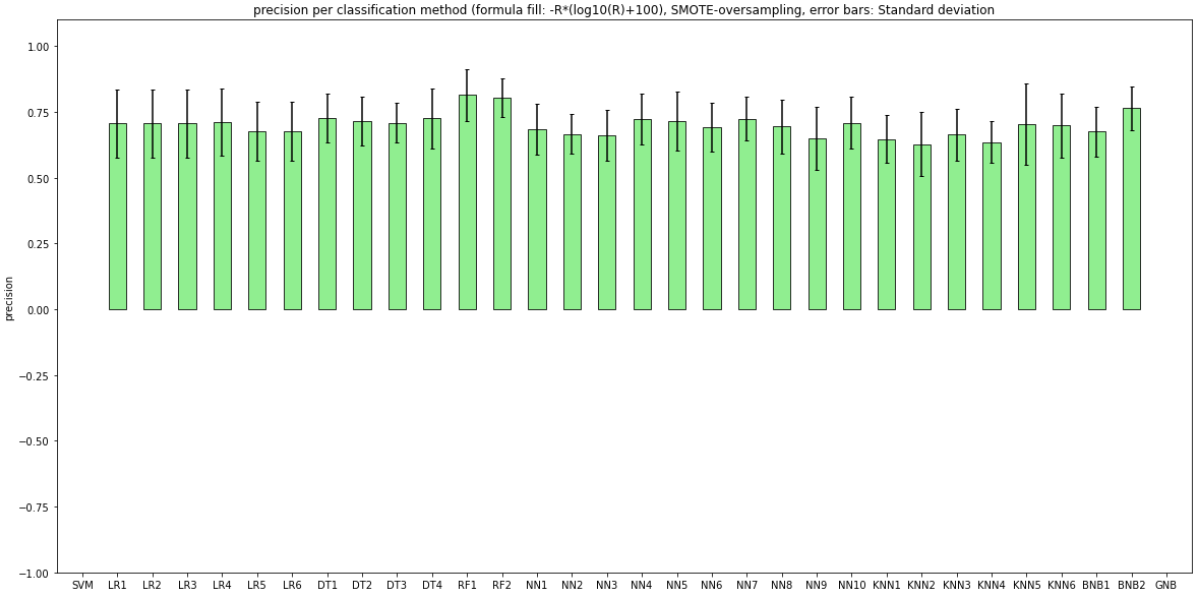

precision per classification method (arbitrary fill eg. -99999, undersampling, error bars: Standard deviation
precision
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB
precision per classification method (formula fill: -R*(log10(R)+100), undersampling, error bars: Standard deviation
precision
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB
precision per classification method (Iterative imputation, undersampling, error bars: Standard deviation
precision
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

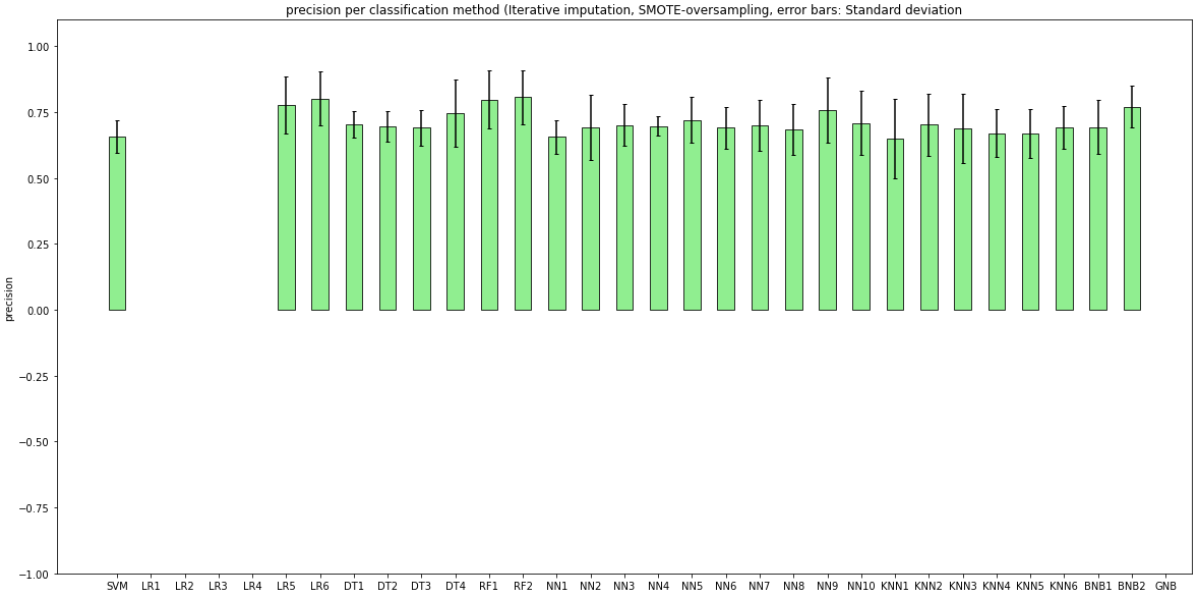

precision per classification method (arbitrary fill eg. -99999, class weights(grid search), error bars: Standard deviation
precision
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM
LR1
LR2
LR3
LR4
LR5
LR6
DT1
DT2
DT3
DT4
RF1
RF2

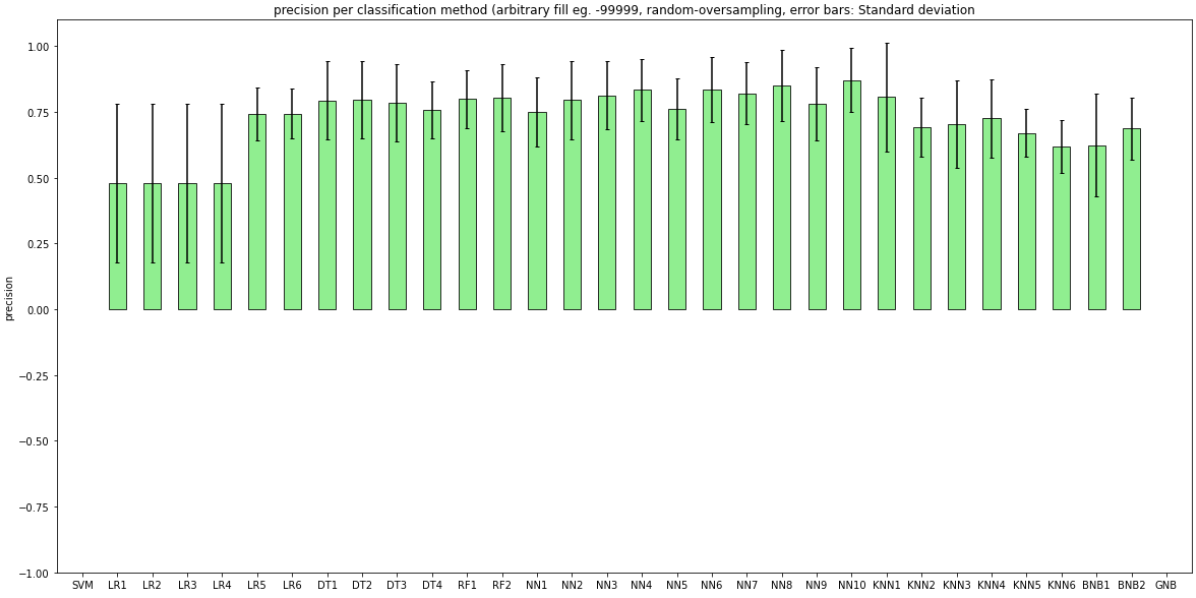

precision per classification method (formula fill: -R*(log10(R)+100), class weights(grid search), error bars: Standard deviation
precision
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM
LR1
LR2
LR3
LR4
LR5
LR6
DT1
DT2
DT3
DT4
RF1
RF2

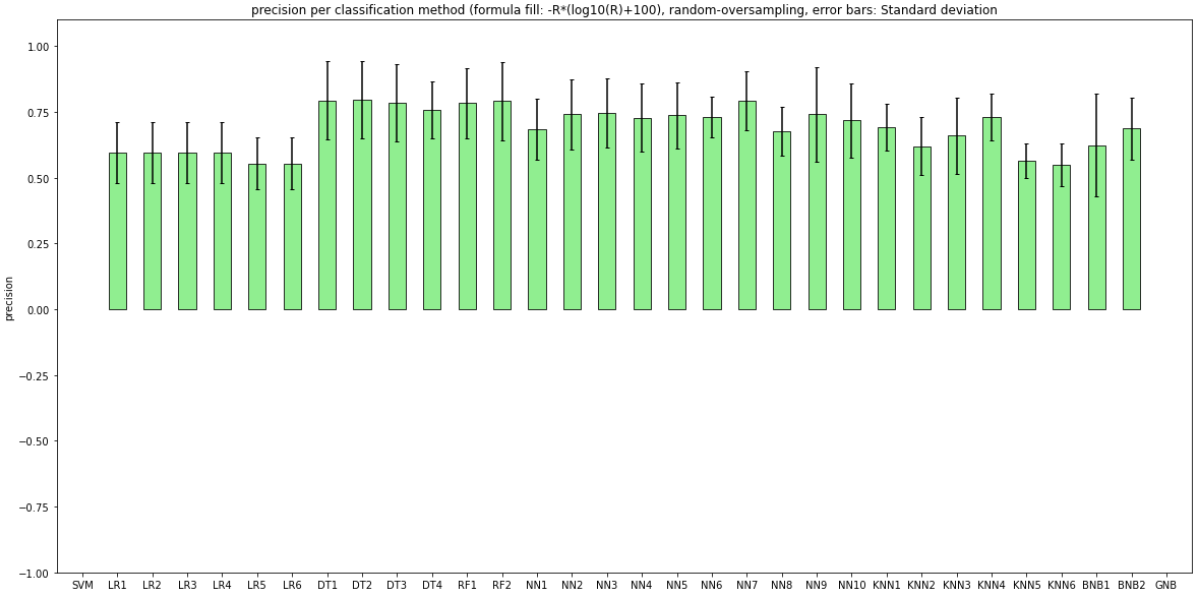

precision per classification method (Iterative imputation, class weights(grid search), error bars: Standard deviation
precision
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM
LR1
LR2
LR3
LR4
LR5
LR6
DT1
DT2
DT3
DT4
RF1
RF2

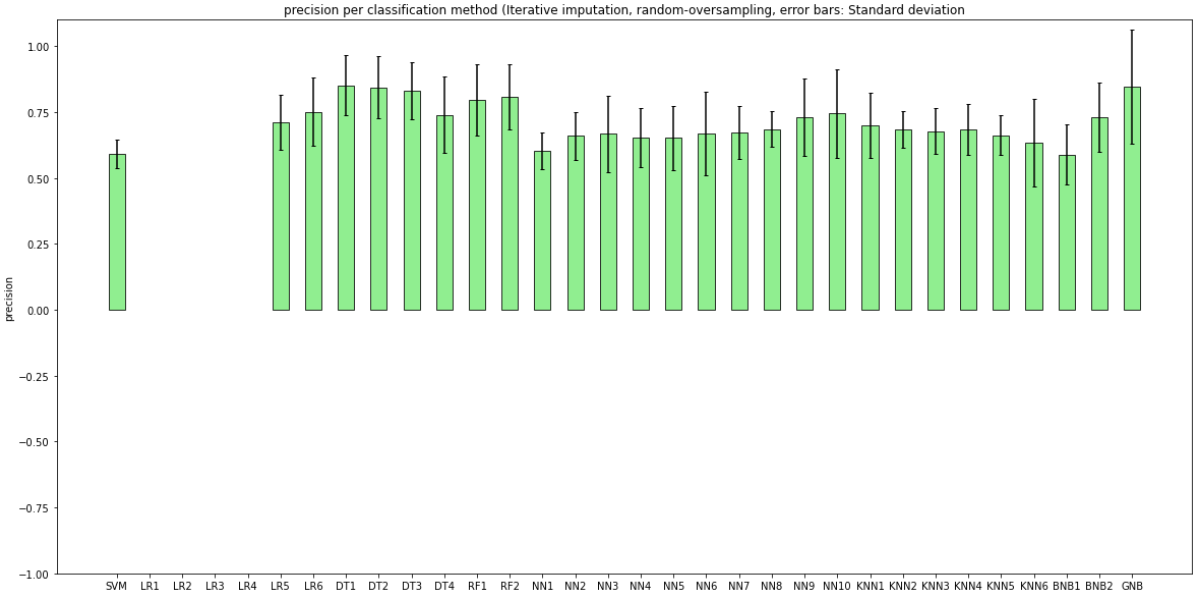

precision per classification method (arbitrary fill eg. -99999, SMOTE-oversampling, error bars: Standard deviation
precision
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

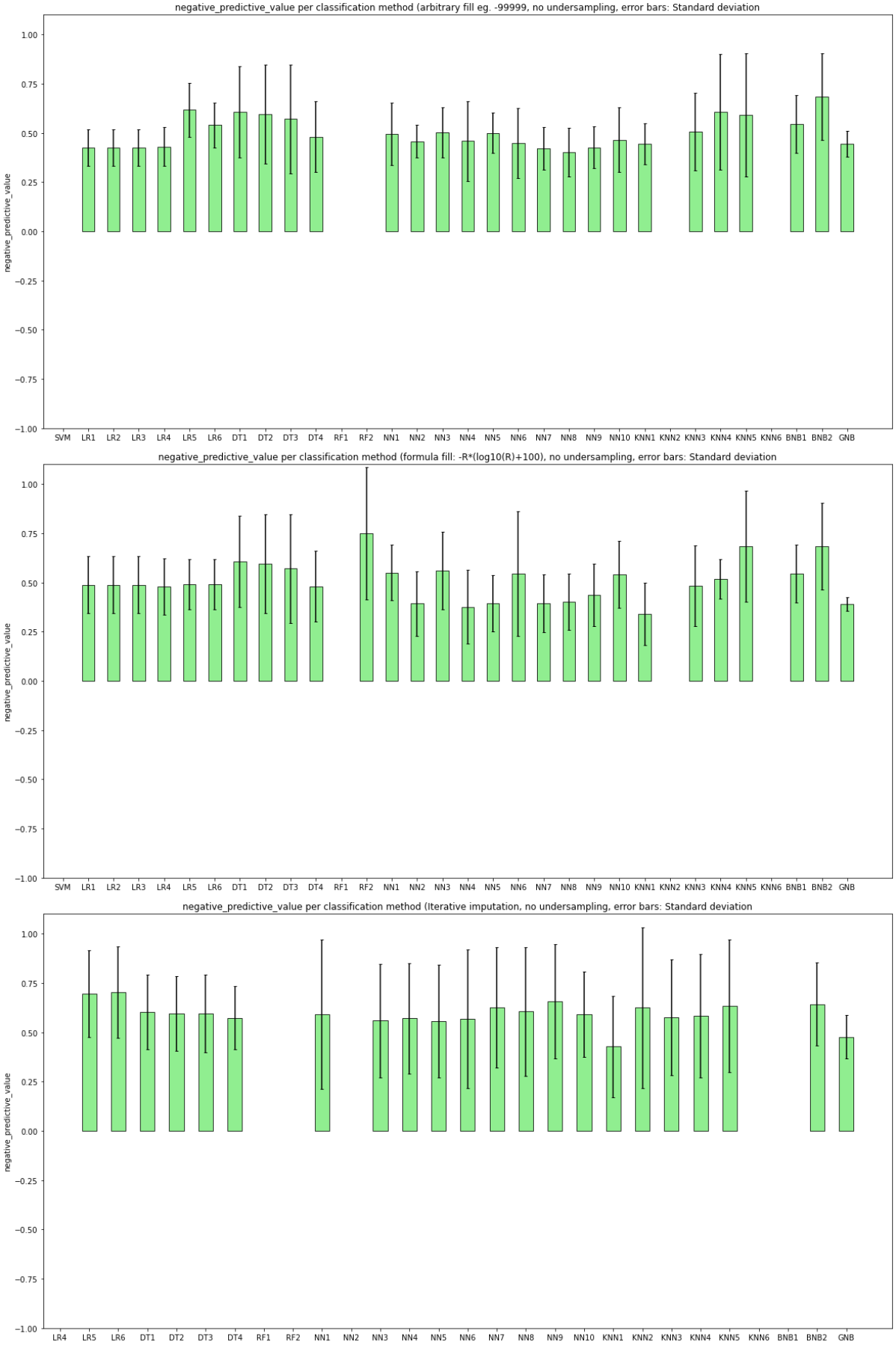

precision per classification method (formula fill: -R*(log10(R)+100), SMOTE-oversampling, error bars: Standard deviation
precision
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

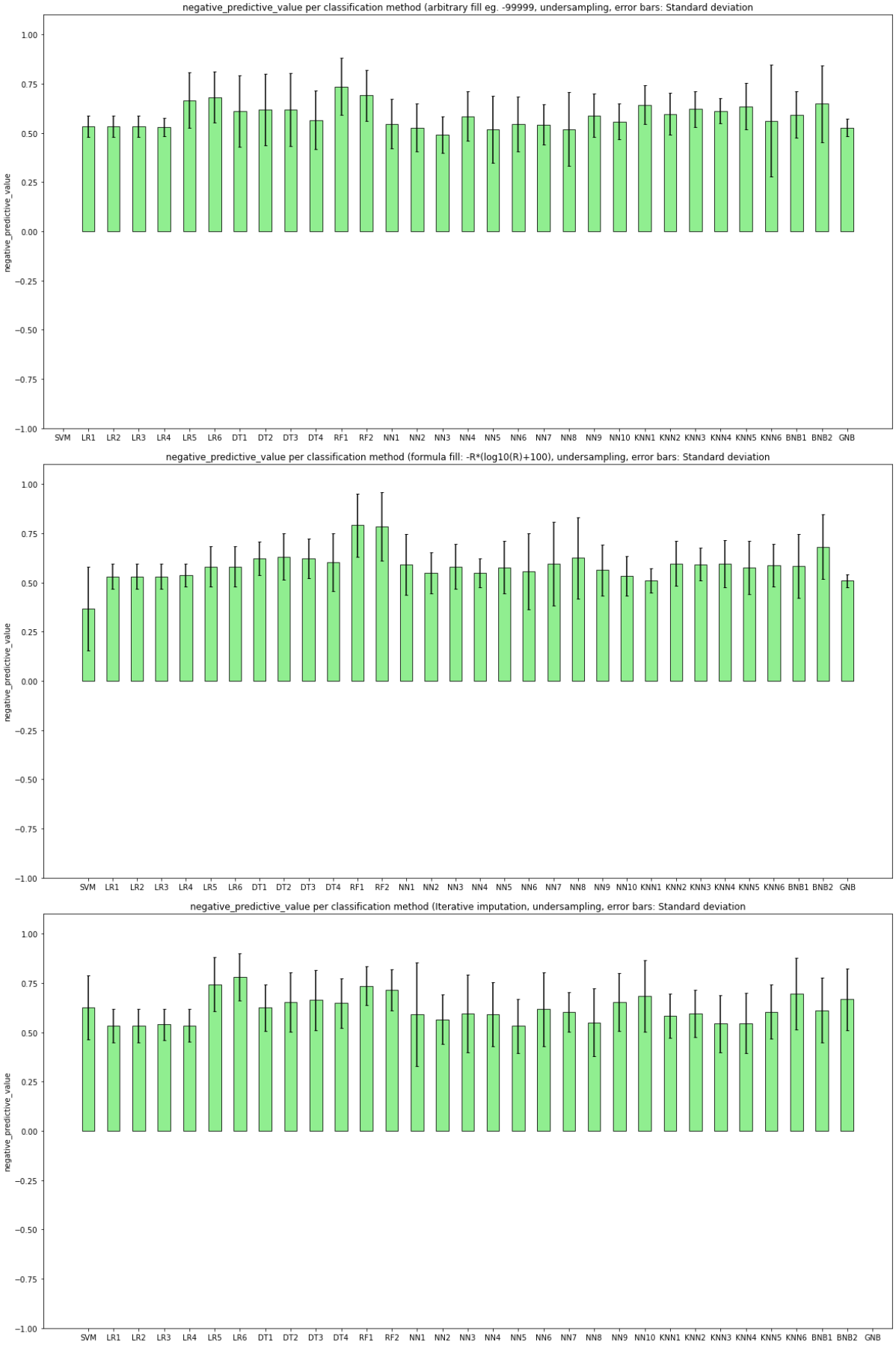

precision per classification method (Iterative imputation, SMOTE-oversampling, error bars: Standard deviation
precision
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

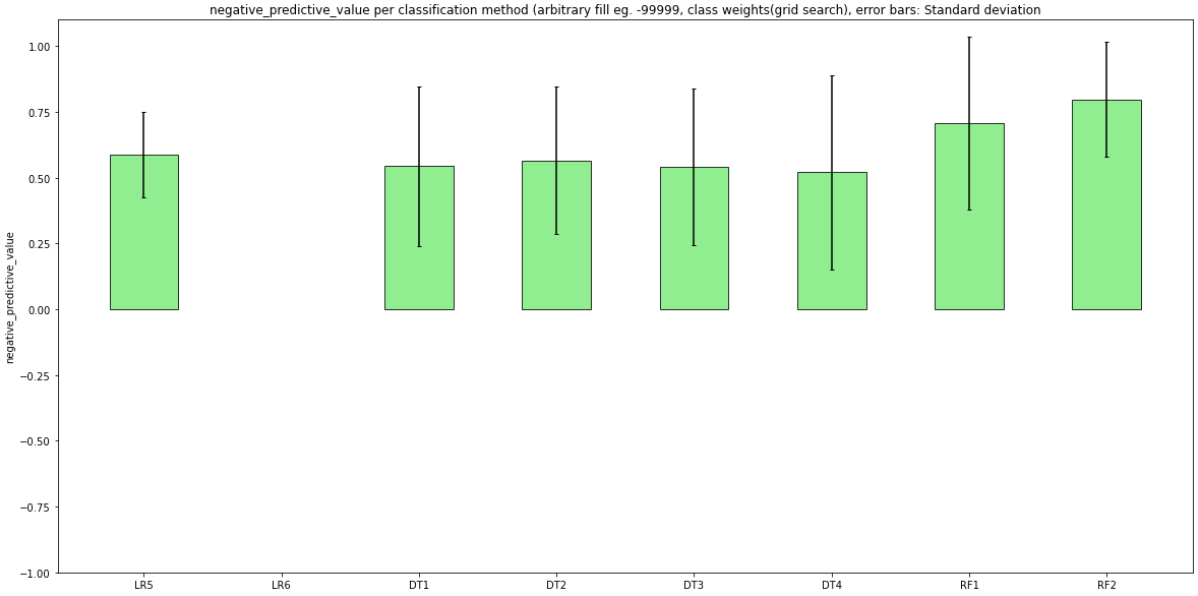

precision per classification method (arbitrary fill eg. -99999, random-oversampling, error bars: Standard deviation
precision
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

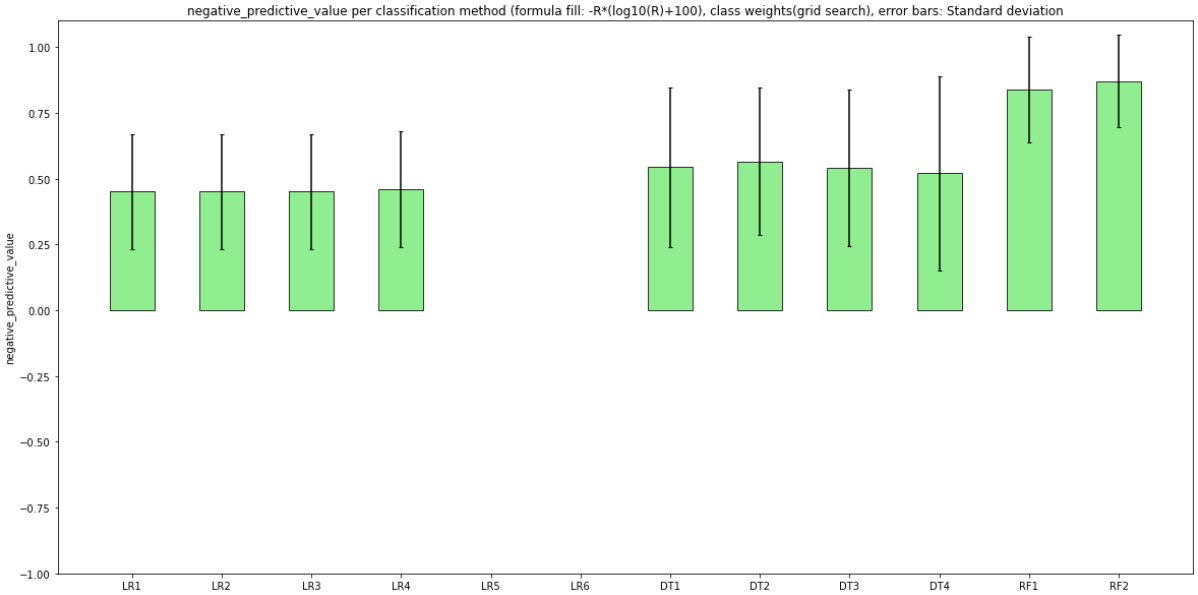

precision per classification method (formula fill: -R*(log10(R)+100), random-oversampling, error bars: Standard deviation
precision
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

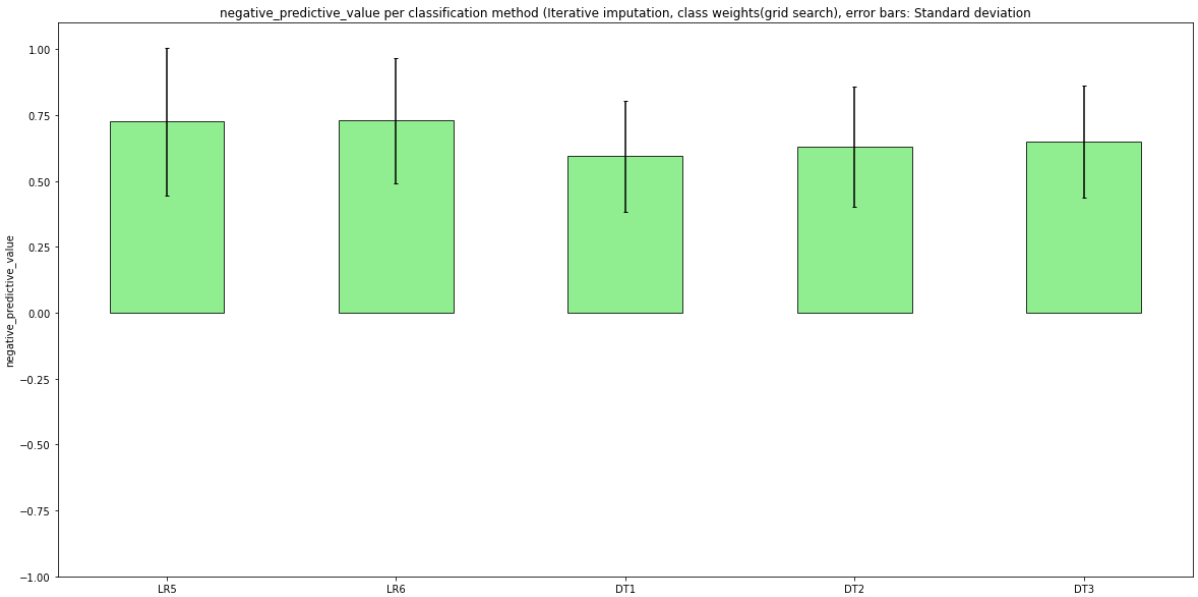

precision per classification method (Iterative imputation, random-oversampling, error bars: Standard deviation
precision
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

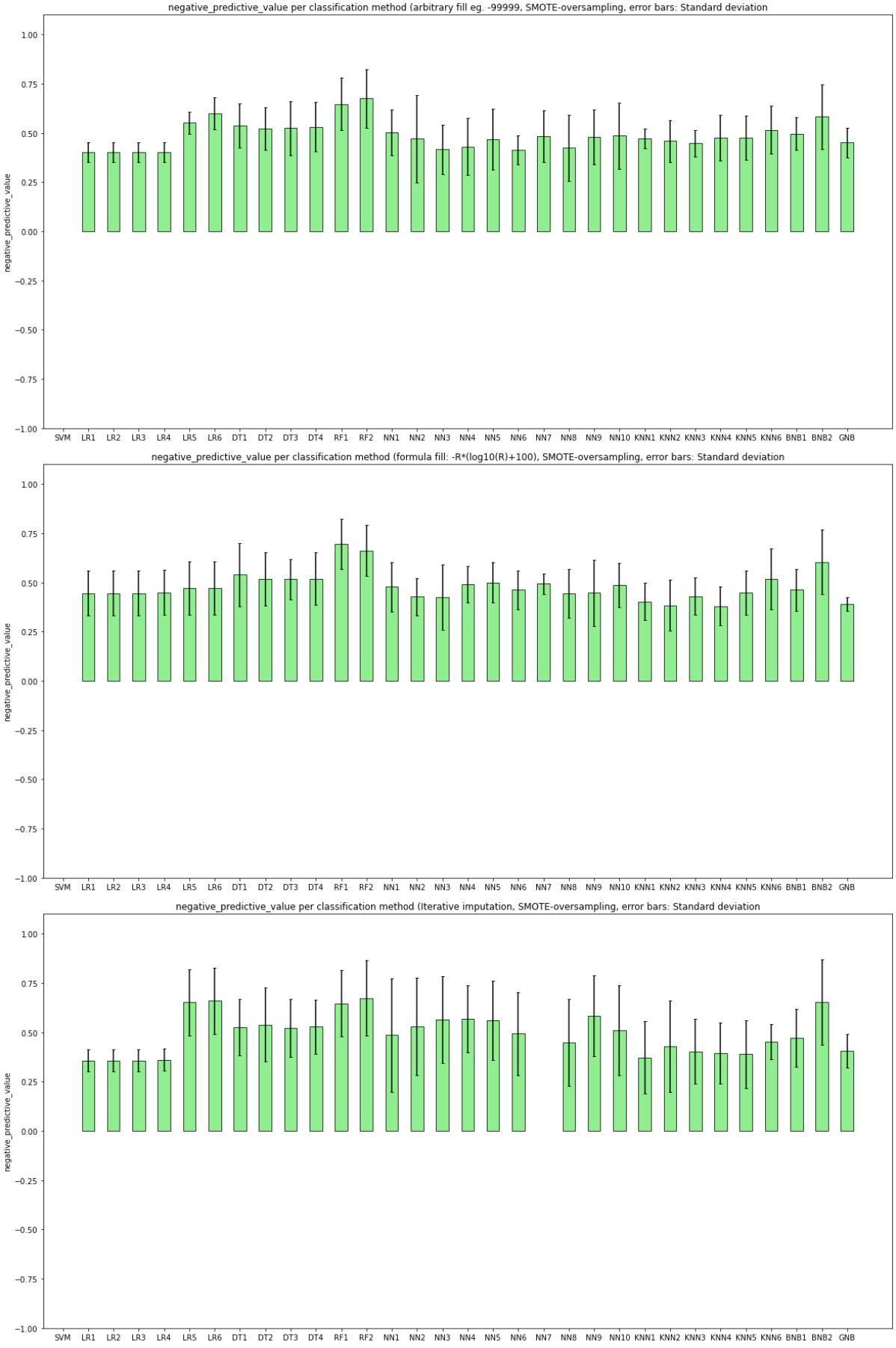

negative_predictive_value per classification method (arbitrary fill eg. -99999, no undersampling, error bars: Standard deviation
negative_predictive_value
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB
negative_predictive_value per classification method (formula fill: -R*(log10(R)+100), no undersampling, error bars: Standard deviation
negative_predictive_value
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB
negative_predictive_value per classification method (Iterative imputation, no undersampling, error bars: Standard deviation
negative_predictive_value
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

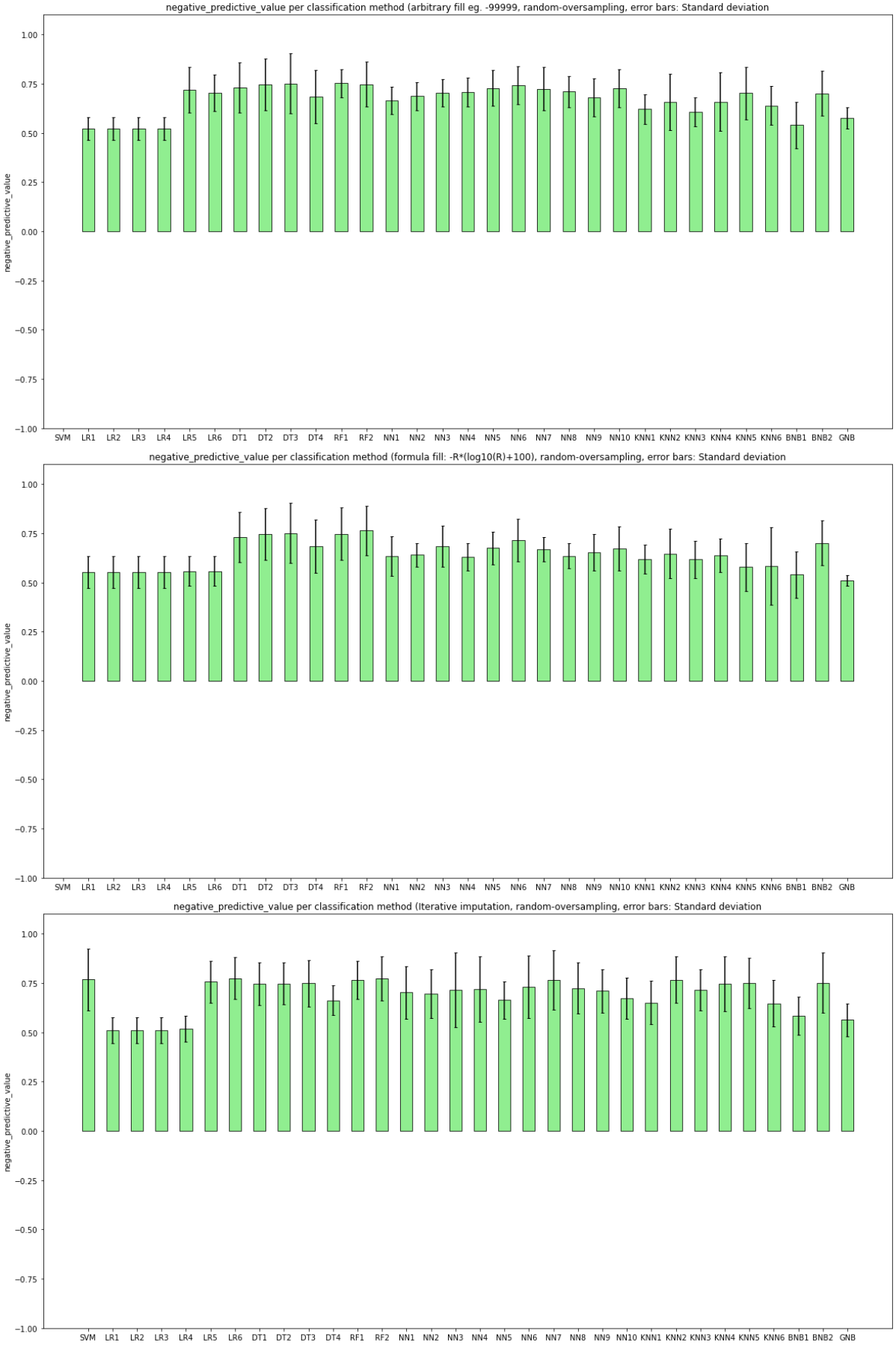

negative_predictive_value per classification method (arbitrary fill eg. -99999, undersampling, error bars: Standard deviation
negative_predictive_value per classification method (formula fill: -R*(log10(R)+100), undersampling, error bars: Standard deviation
negative_predictive_value per classification method (Iterative imputation, undersampling, error bars: Standard deviation
negative_predictive_value
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

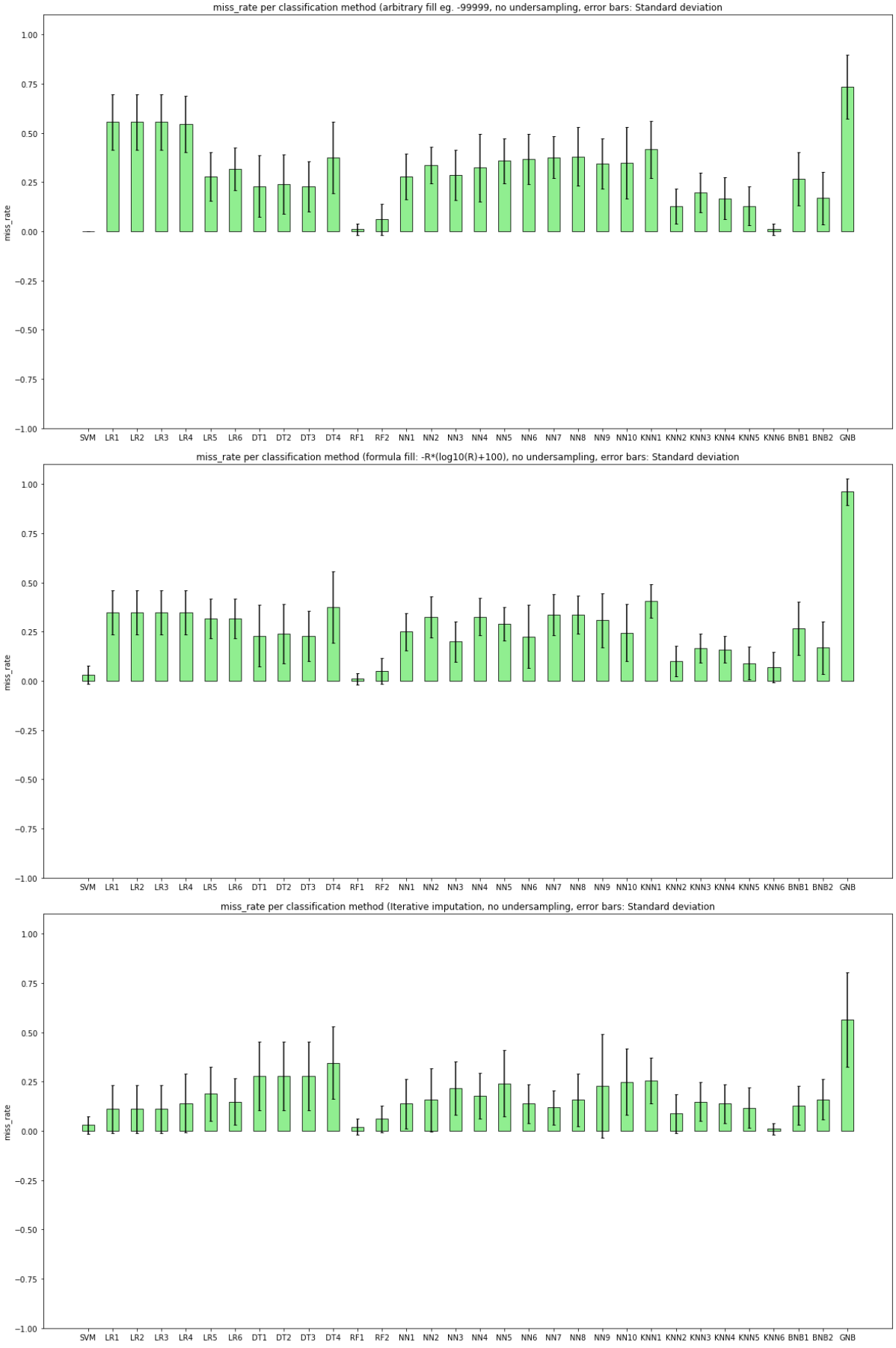

negative_predictive_value per classification method (arbitrary fill eg. -99999, class weights(grid search), error bars: Standard deviation
negative_predictive_value
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
LR5
LR6
DT1
DT2
DT3
DT4
RF1
RF2

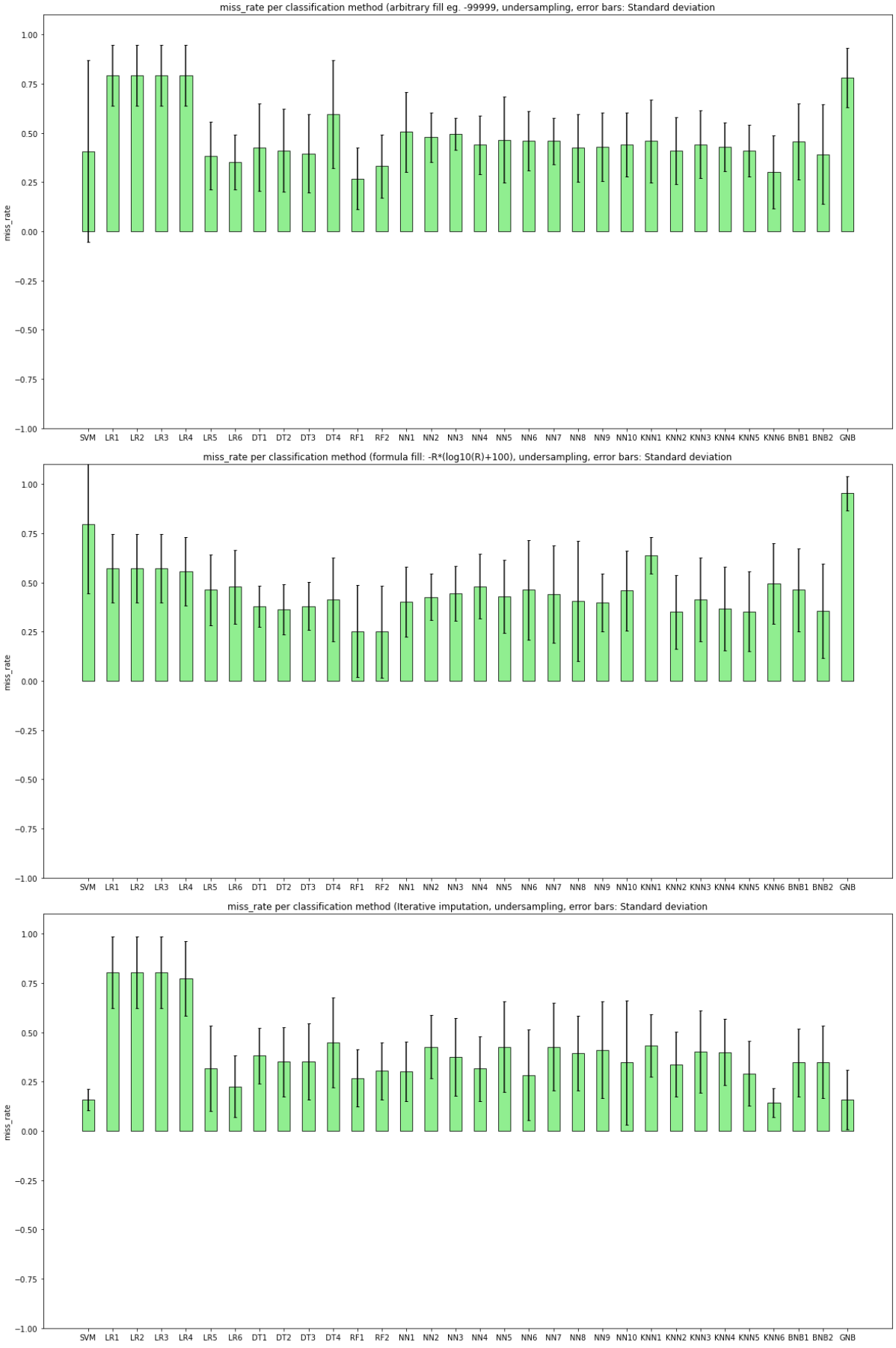

negative_predictive_value per classification method (formula fill: -R*(log10(R)+100), class weights(grid search), error bars: Standard deviation
negative_predictive_value
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
LR1
LR2
LR3
LR4
LR5
LR6
DT1
DT2
DT3
DT4
RF1
RF2

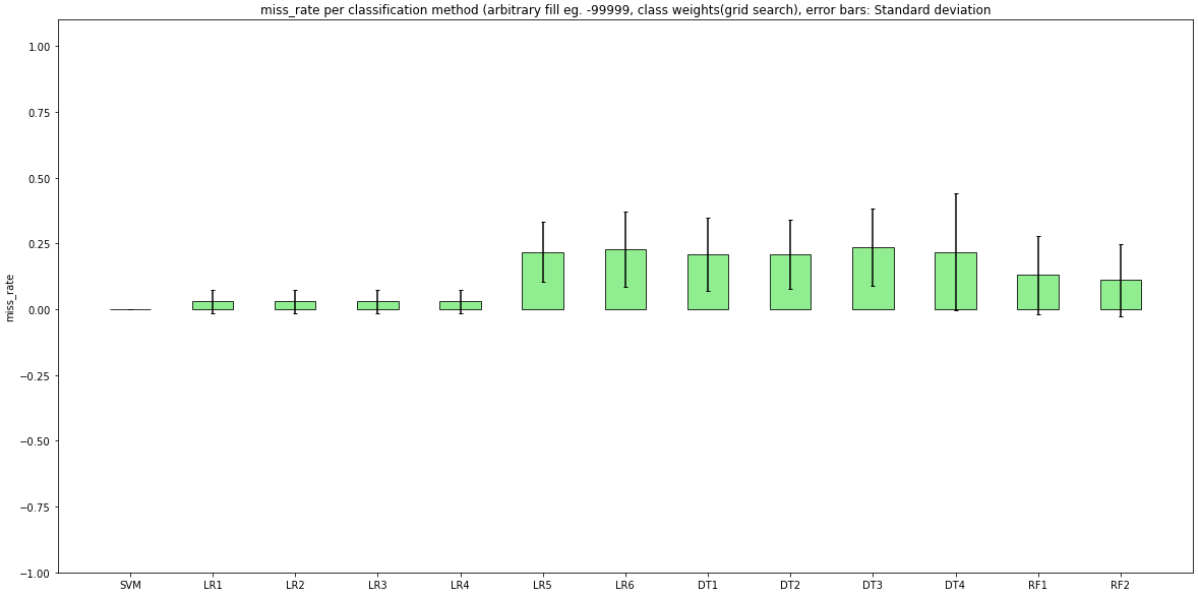

negative_predictive_value per classification method (Iterative imputation, class weights(grid search), error bars: Standard deviation
negative_predictive_value
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
LR5
LR6
DT1
DT2
DT3

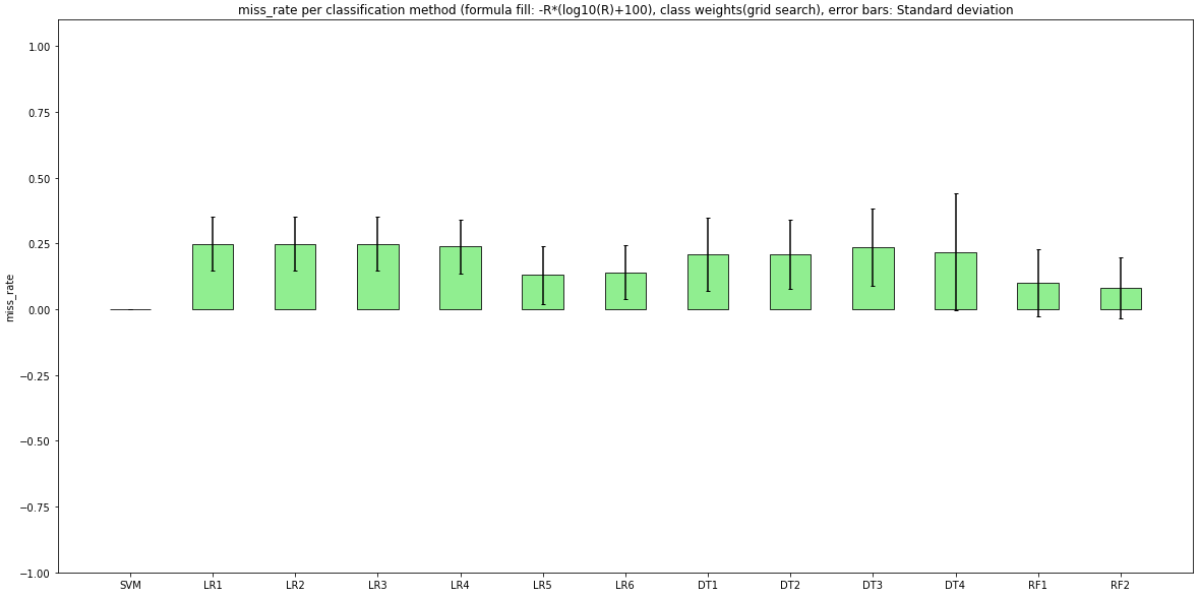

negative_predictive_value per classification method (arbitrary fill eg. -99999, SMOTE-oversampling, error bars: Standard deviation
negative_predictive_value per classification method (formula fill: -R*(log10(R)+100), SMOTE-oversampling, error bars: Standard deviation
negative_predictive_value per classification method (Iterative imputation, SMOTE-oversampling, error bars: Standard deviation
negative_predictive_value
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

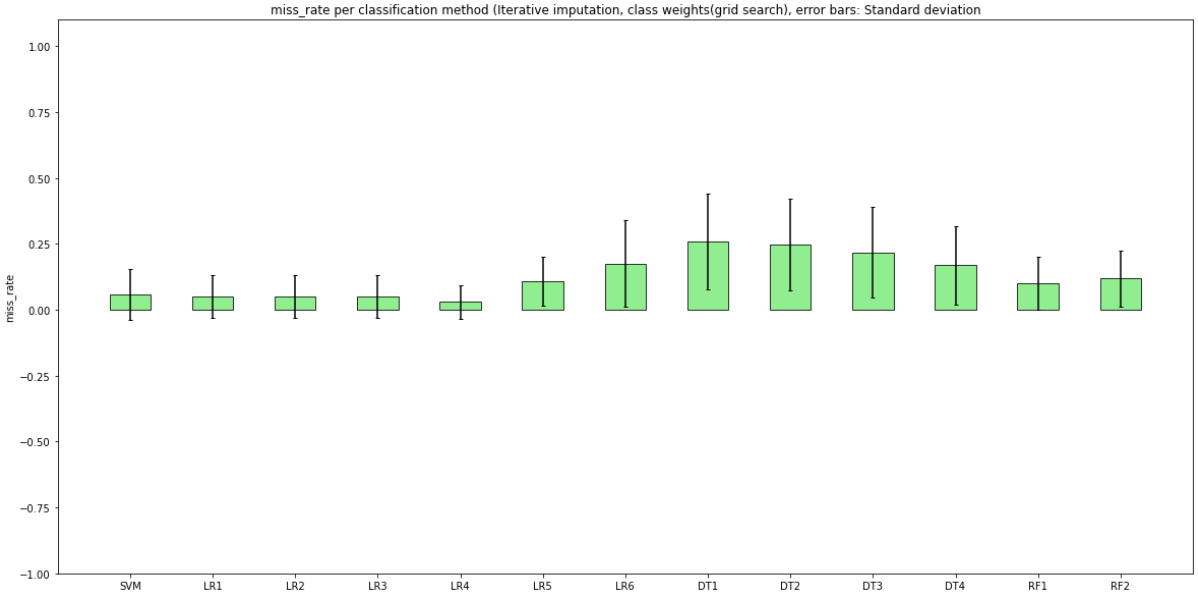

negative_predictive_value per classification method (arbitrary fill eg. -99999, random-oversampling, error bars: Standard deviation
negative_predictive_value per classification method (formula fill: -R*(log10(R)+100), random-oversampling, error bars: Standard deviation
negative_predictive_value per classification method (Iterative imputation, random-oversampling, error bars: Standard deviation
negative_predictive_value
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

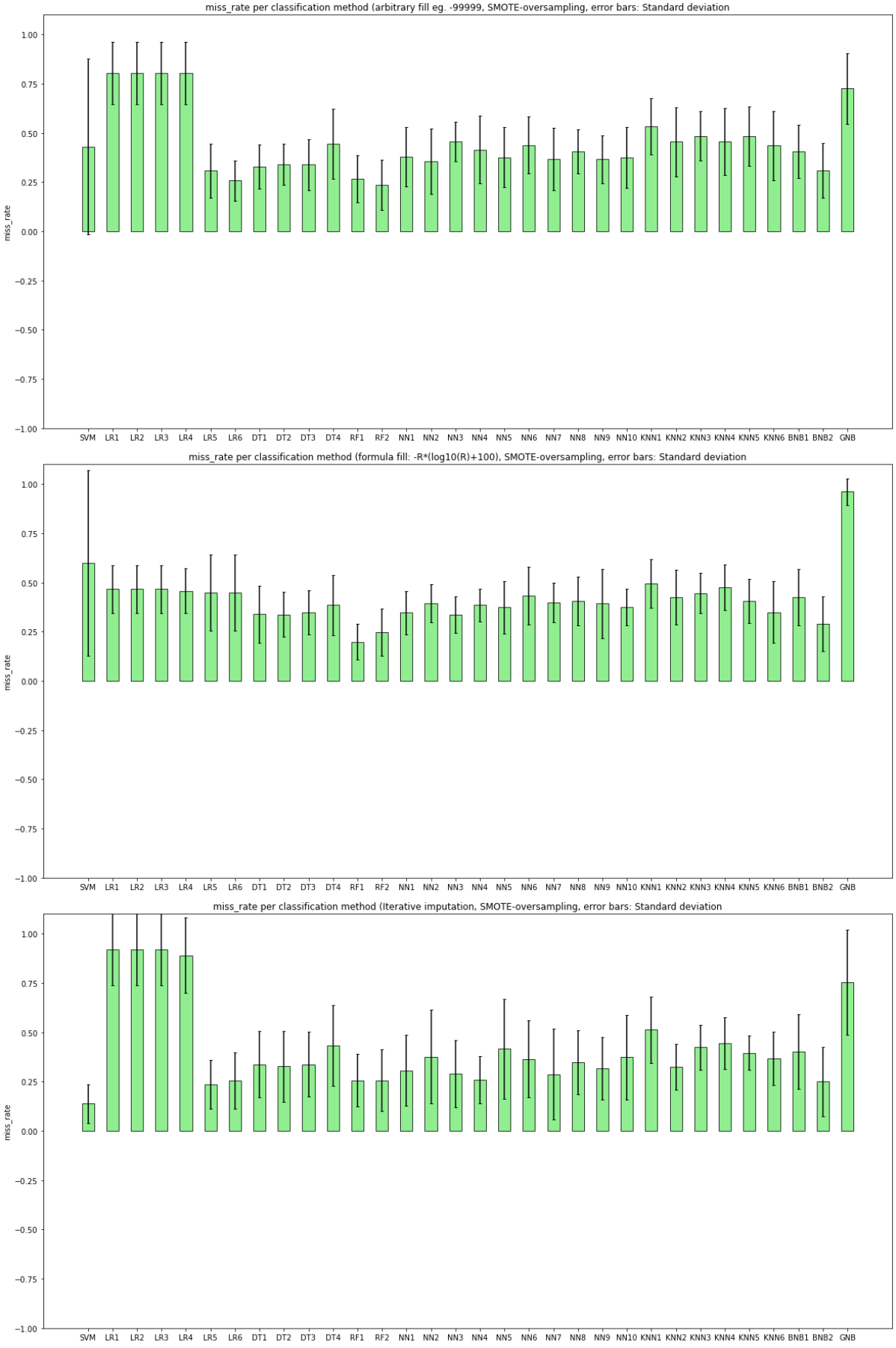

miss_rate per classification method (arbitrary fill eg. -99999, no undersampling, error bars: Standard deviation
miss_rate per classification method (formula fill: -R*(log10(R)+100), no undersampling, error bars: Standard deviation
miss_rate per classification method (Iterative imputation, no undersampling, error bars: Standard deviation
miss_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

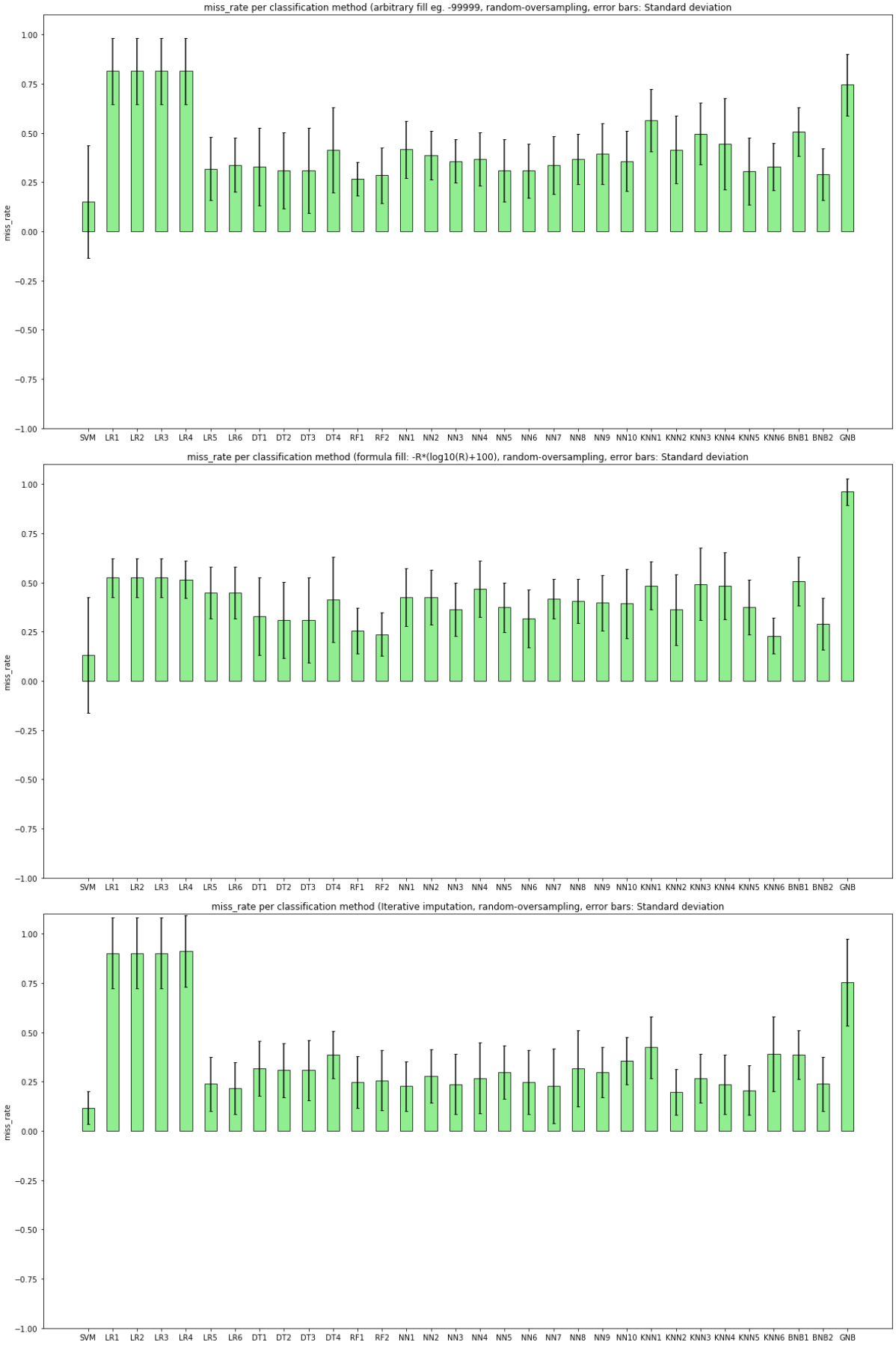

miss_rate per classification method (arbitrary fill eg. -99999, undersampling, error bars: Standard deviation
miss_rate per classification method (formula fill: -R*(log10(R)+100), undersampling, error bars: Standard deviation
miss_rate per classification method (Iterative imputation, undersampling, error bars: Standard deviation
miss_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

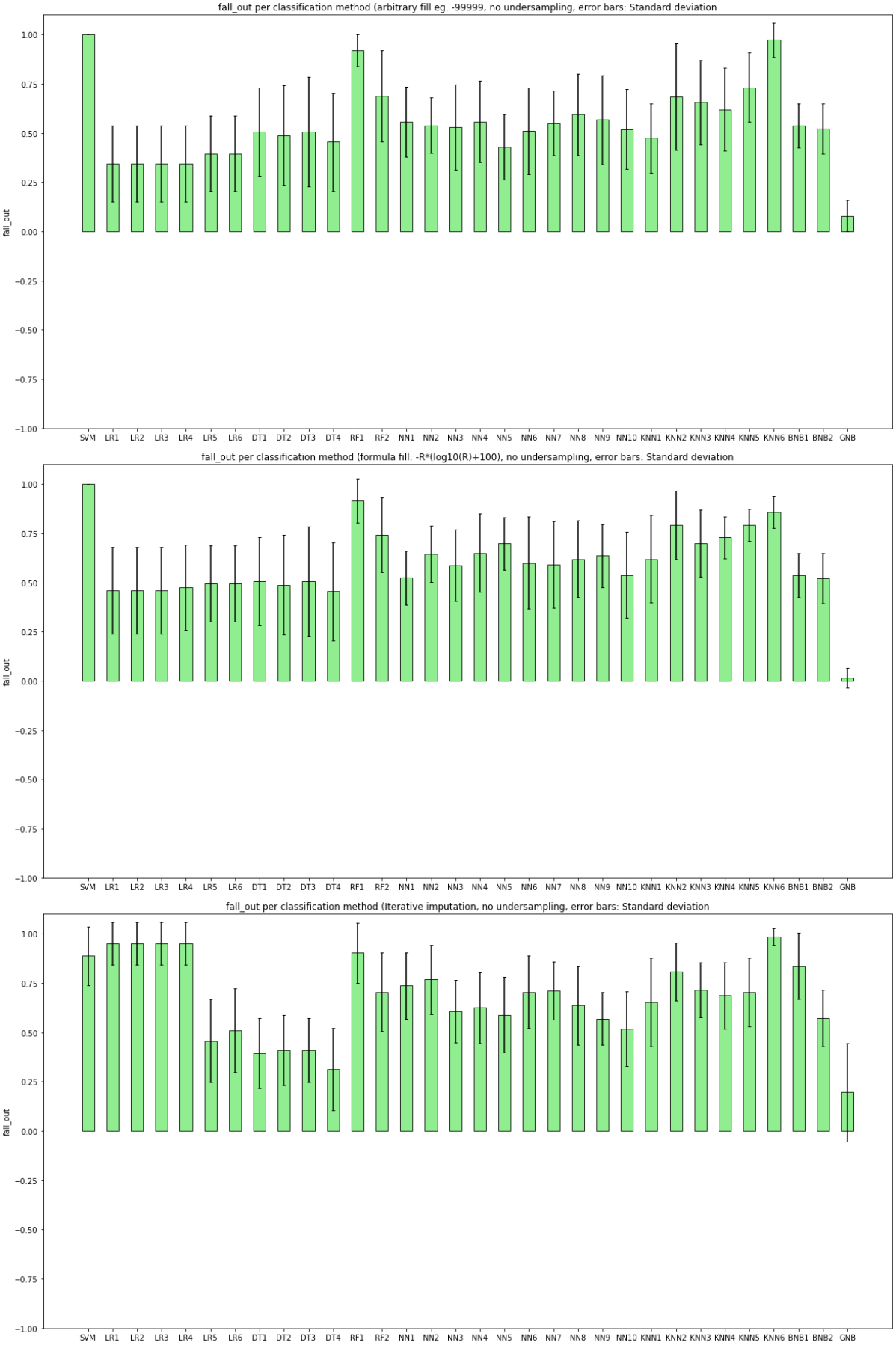

miss_rate per classification method (arbitrary fill eg. -99999, class weights(grid search), error bars: Standard deviation
miss_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM
LR1
LR2
LR3
LR4
LR5
LR6
DT1
DT2
DT3
DT4
RF1
RF2

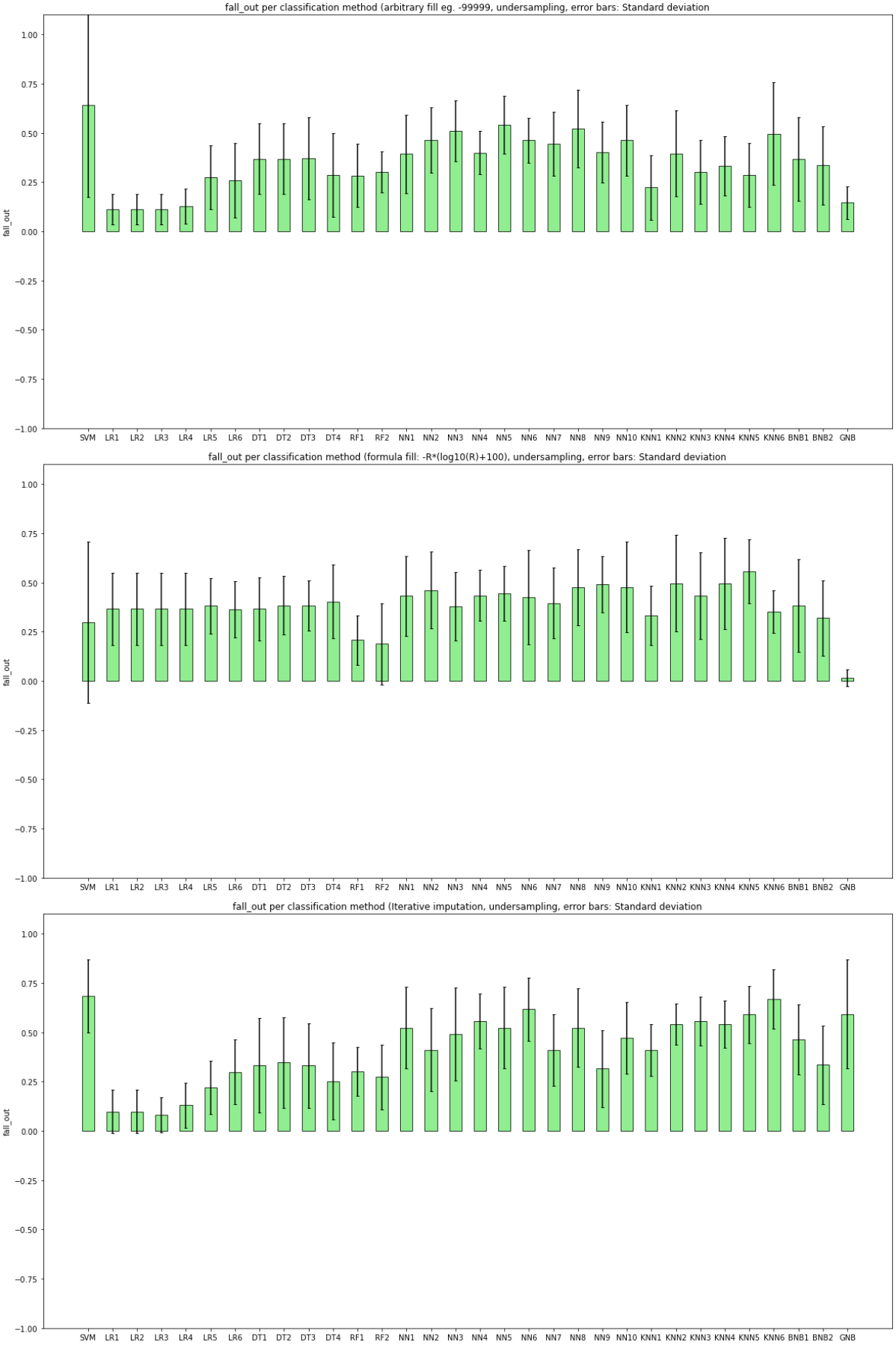

miss_rate per classification method (formula fill: -R*(log10(R)+100), class weights(grid search), error bars: Standard deviation
miss_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM
LR1
LR2
LR3
LR4
LR5
LR6
DT1
DT2
DT3
DT4
RF1
RF2

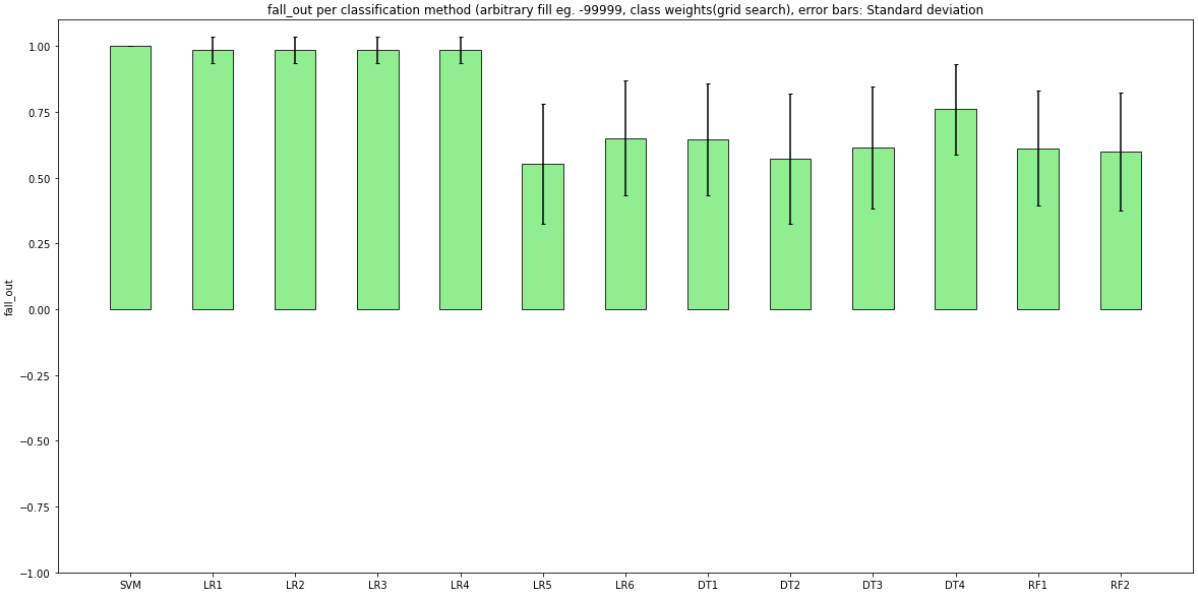

miss_rate per classification method (Iterative imputation, class weights(grid search), error bars: Standard deviation
miss_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM
LR1
LR2
LR3
LR4
LR5
LR6
DT1
DT2
DT3
DT4
RF1
RF2

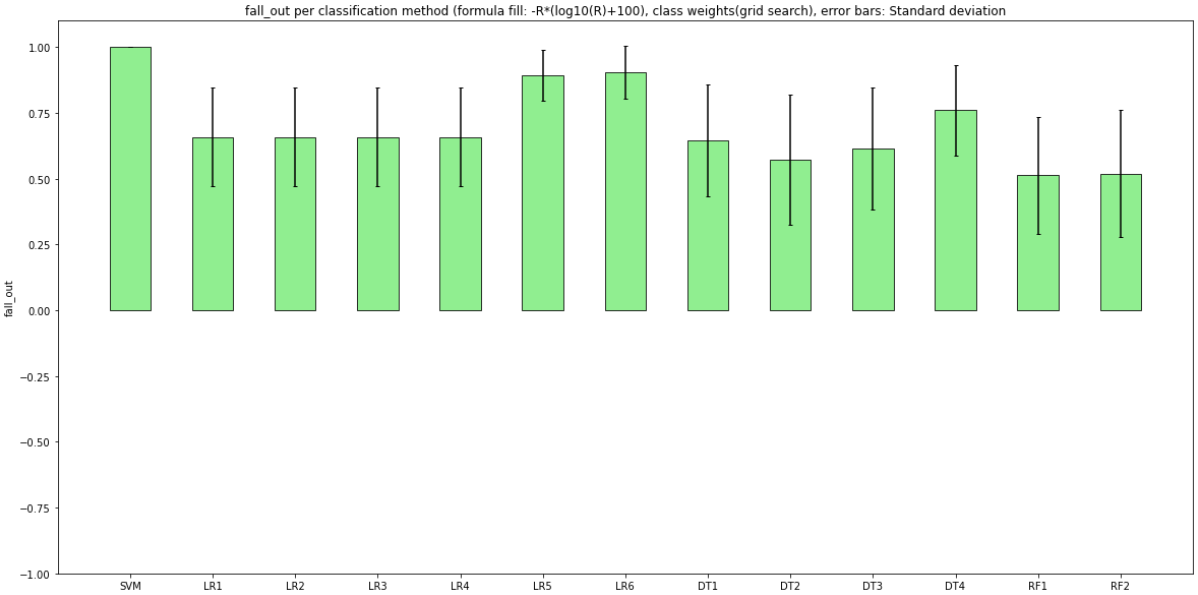

miss_rate per classification method (arbitrary fill eg. -99999, SMOTE-oversampling, error bars: Standard deviation
miss_rate per classification method (formula fill: -R*(log10(R)+100), SMOTE-oversampling, error bars: Standard deviation
miss_rate per classification method (Iterative imputation, SMOTE-oversampling, error bars: Standard deviation
miss_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

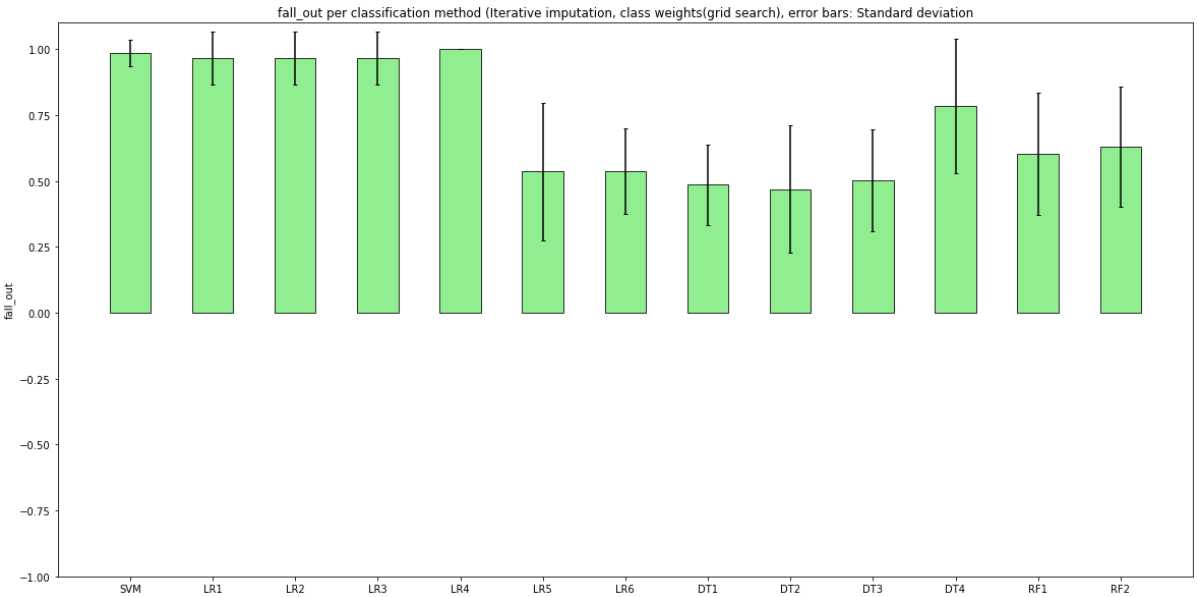

miss_rate per classification method (arbitrary fill eg. -99999, random-oversampling, error bars: Standard deviation
miss_rate per classification method (formula fill: -R*(log10(R)+100), random-oversampling, error bars: Standard deviation
miss_rate per classification method (Iterative imputation, random-oversampling, error bars: Standard deviation
miss_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

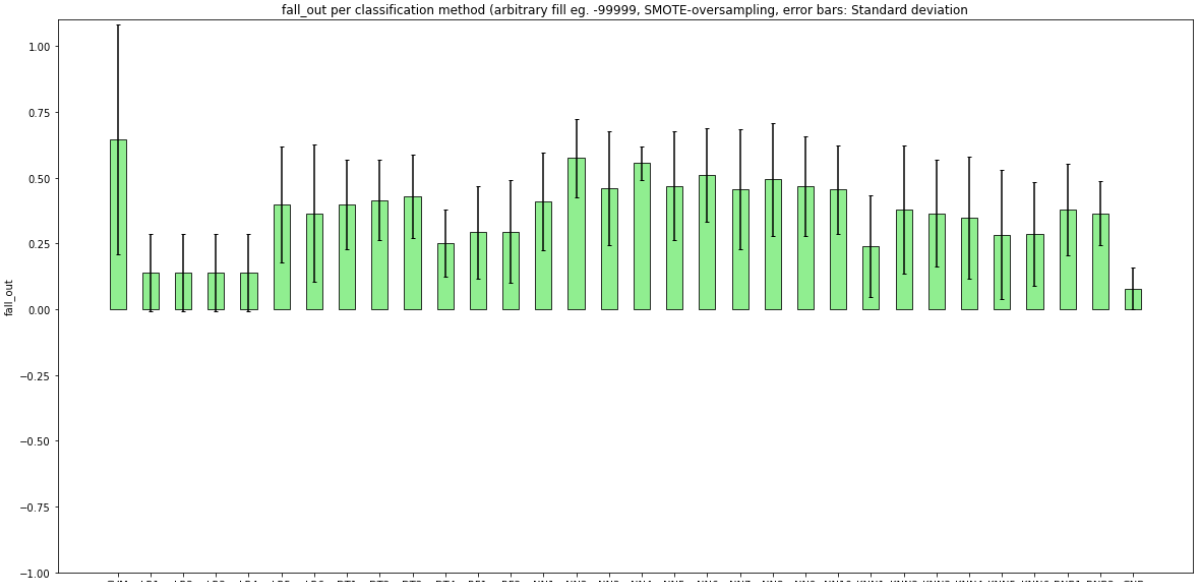

fall_out per classification method (arbitrary fill eg. -99999, no undersampling, error bars: Standard deviation
fall_out per classification method (formula fill: -R*(log10(R)+100), no undersampling, error bars: Standard deviation
fall_out per classification method (Iterative imputation, no undersampling, error bars: Standard deviation
fall_out
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

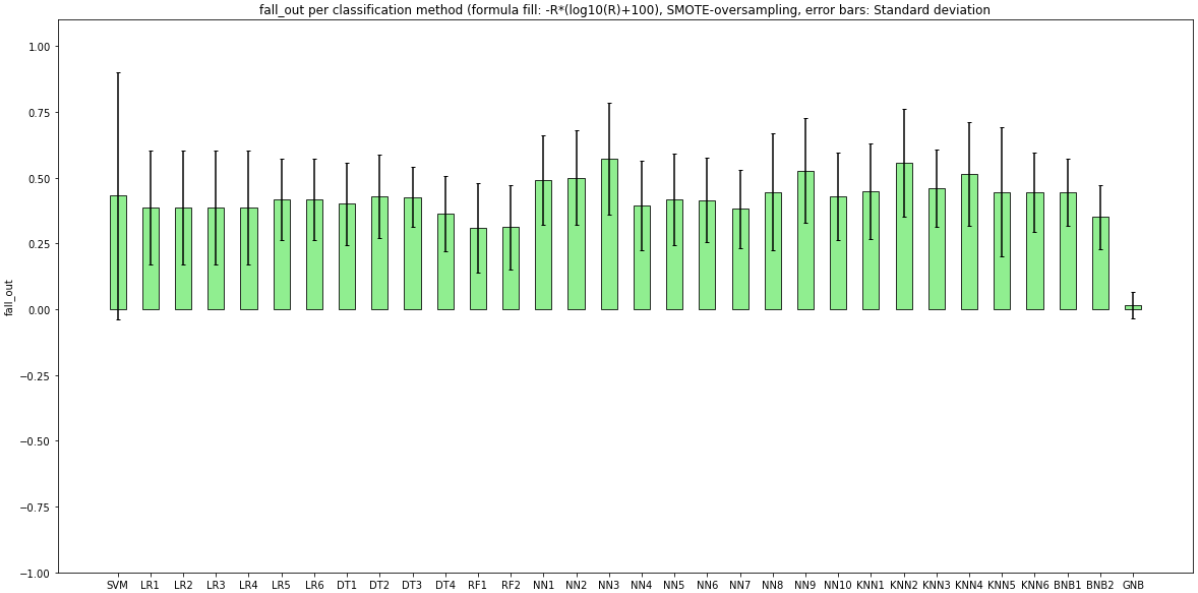

fall_out per classification method (arbitrary fill eg. -99999, undersampling, error bars: Standard deviation
fall_out per classification method (formula fill: -R*(log10(R)+100), undersampling, error bars: Standard deviation
fall_out per classification method (Iterative imputation, undersampling, error bars: Standard deviation
fall_out
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

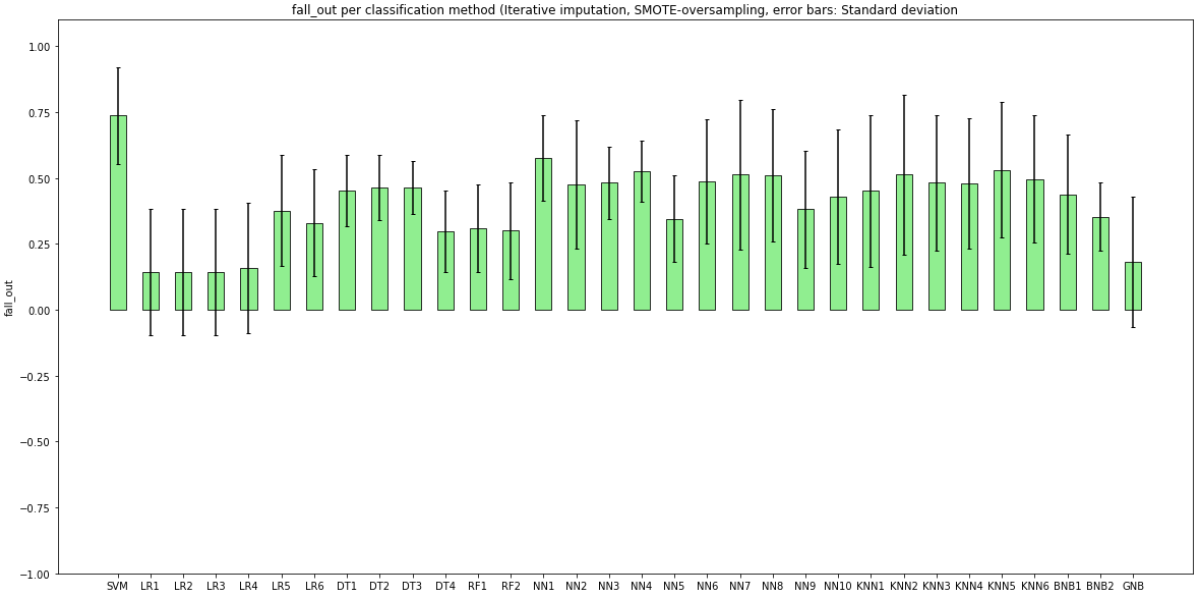

fall_out per classification method (arbitrary fill eg. -99999, class weights(grid search), error bars: Standard deviation
fall_out
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM
LR1
LR2
LR3
LR4
LR5
LR6
DT1
DT2
DT3
DT4
RF1
RF2

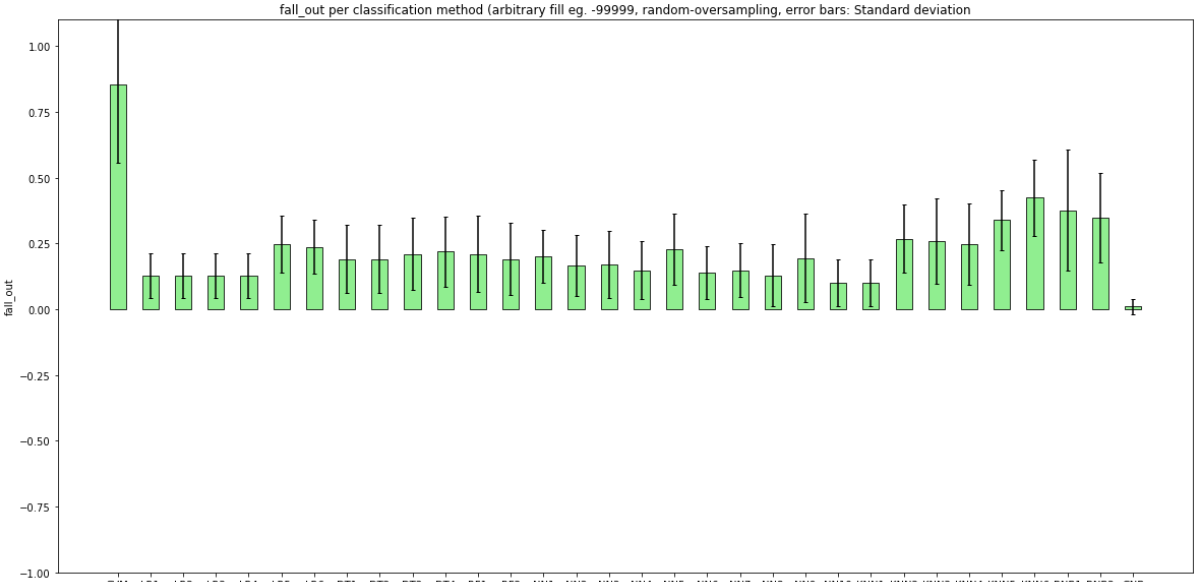

fall_out per classification method (formula fill: -R*(log10(R)+100), class weights(grid search), error bars: Standard deviation
fall_out
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM
LR1
LR2
LR3
LR4
LR5
LR6
DT1
DT2
DT3
DT4
RF1
RF2

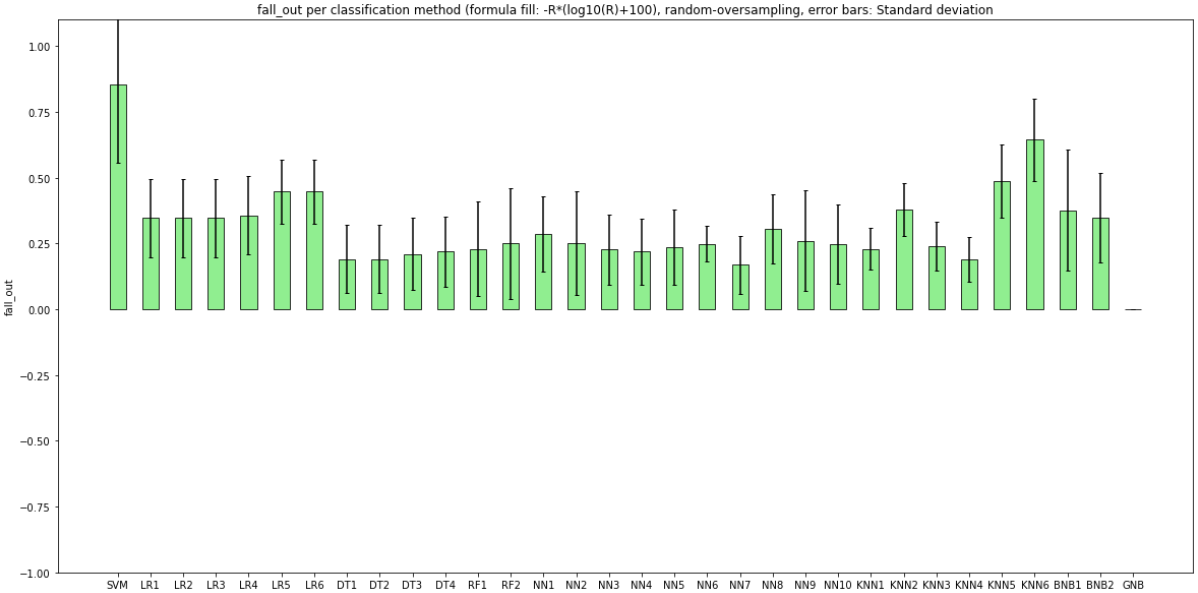

fall_out per classification method (Iterative imputation, class weights(grid search), error bars: Standard deviation
fall_out
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM
LR1
LR2
LR3
LR4
LR5
LR6
DT1
DT2
DT3
DT4
RF1
RF2

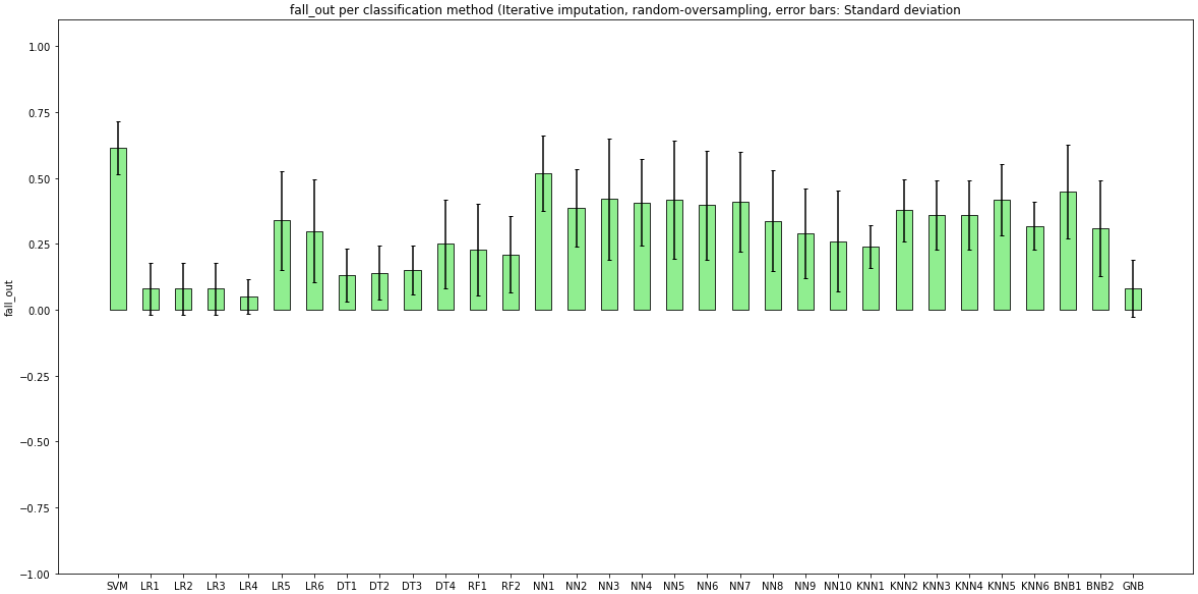

fall_out per classification method (arbitrary fill eg. -99999, SMOTE-oversampling, error bars: Standard deviation
fall_out
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB


fall_out per classification method (formula fill: -R*(log10(R)+100), SMOTE-oversampling, error bars: Standard deviation
fall_out
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

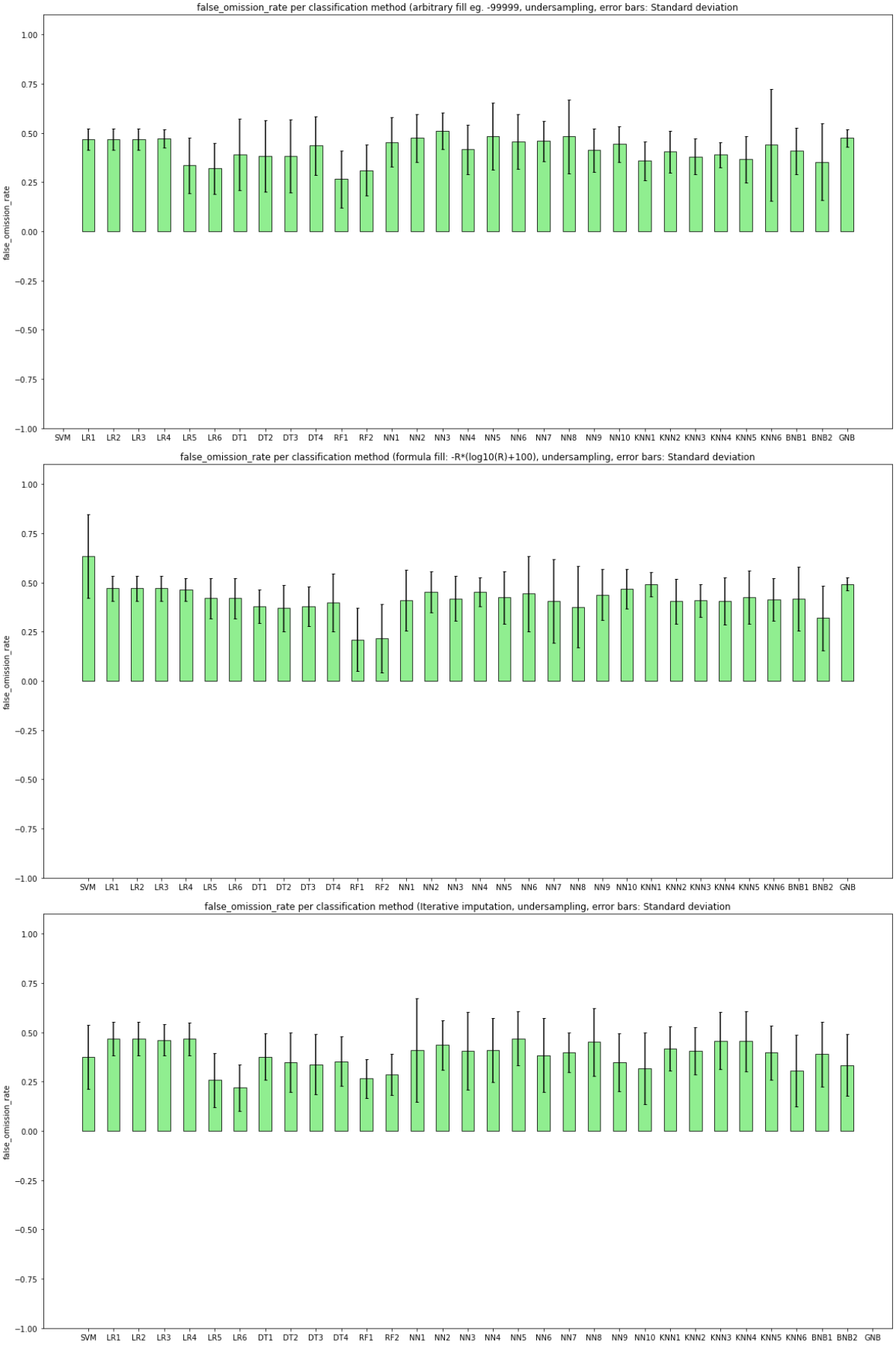

fall_out per classification method (Iterative imputation, SMOTE-oversampling, error bars: Standard deviation
fall_out
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

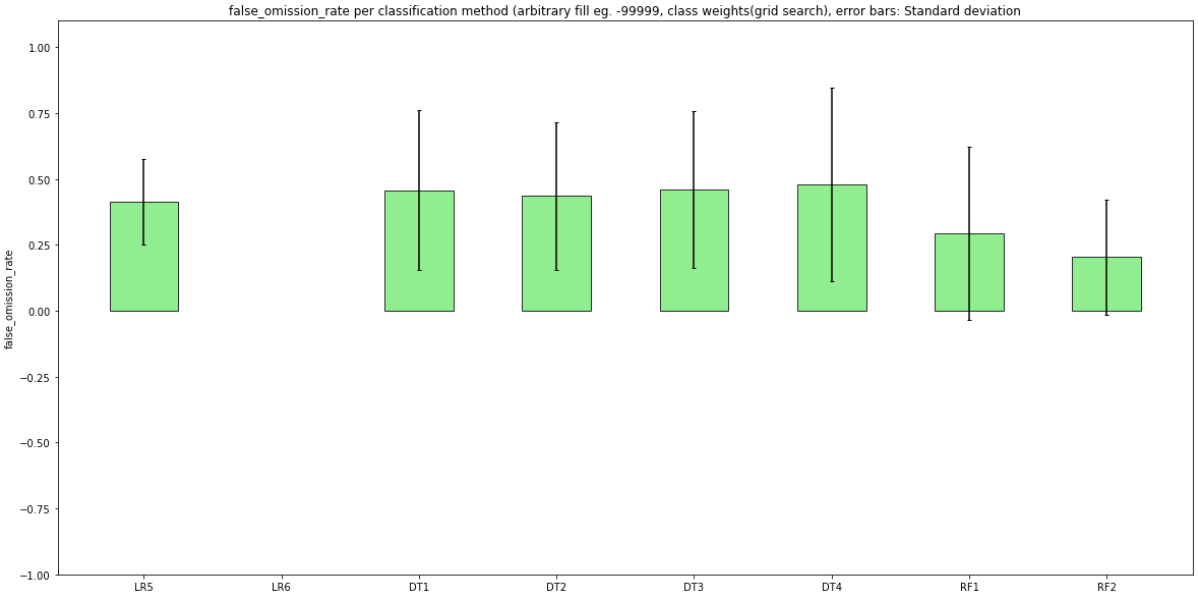

fall_out per classification method (arbitrary fill eg. -99999, random-oversampling, error bars: Standard deviation
fall_out
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

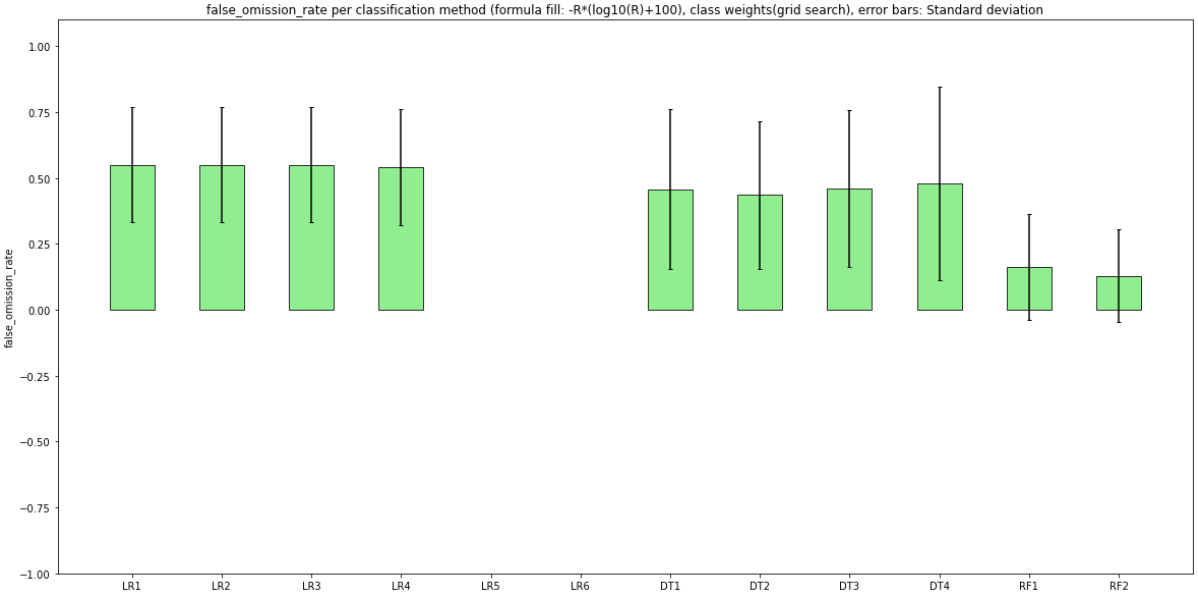

fall_out per classification method (formula fill: -R*(log10(R)+100), random-oversampling, error bars: Standard deviation
fall_out
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

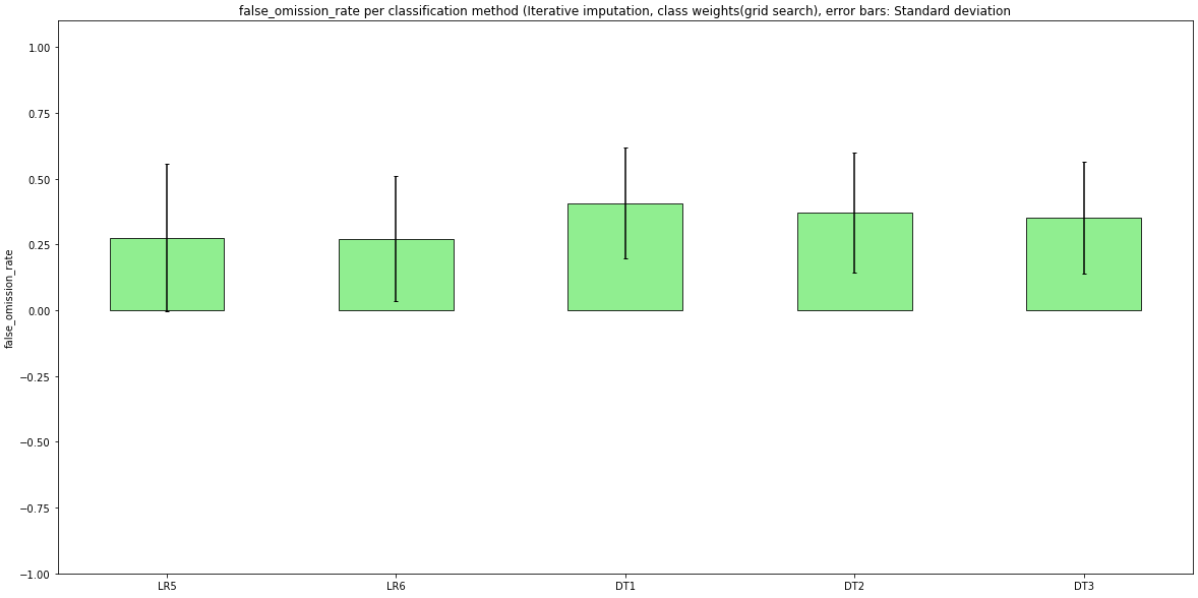

fall_out per classification method (Iterative imputation, random-oversampling, error bars: Standard deviation
fall_out
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

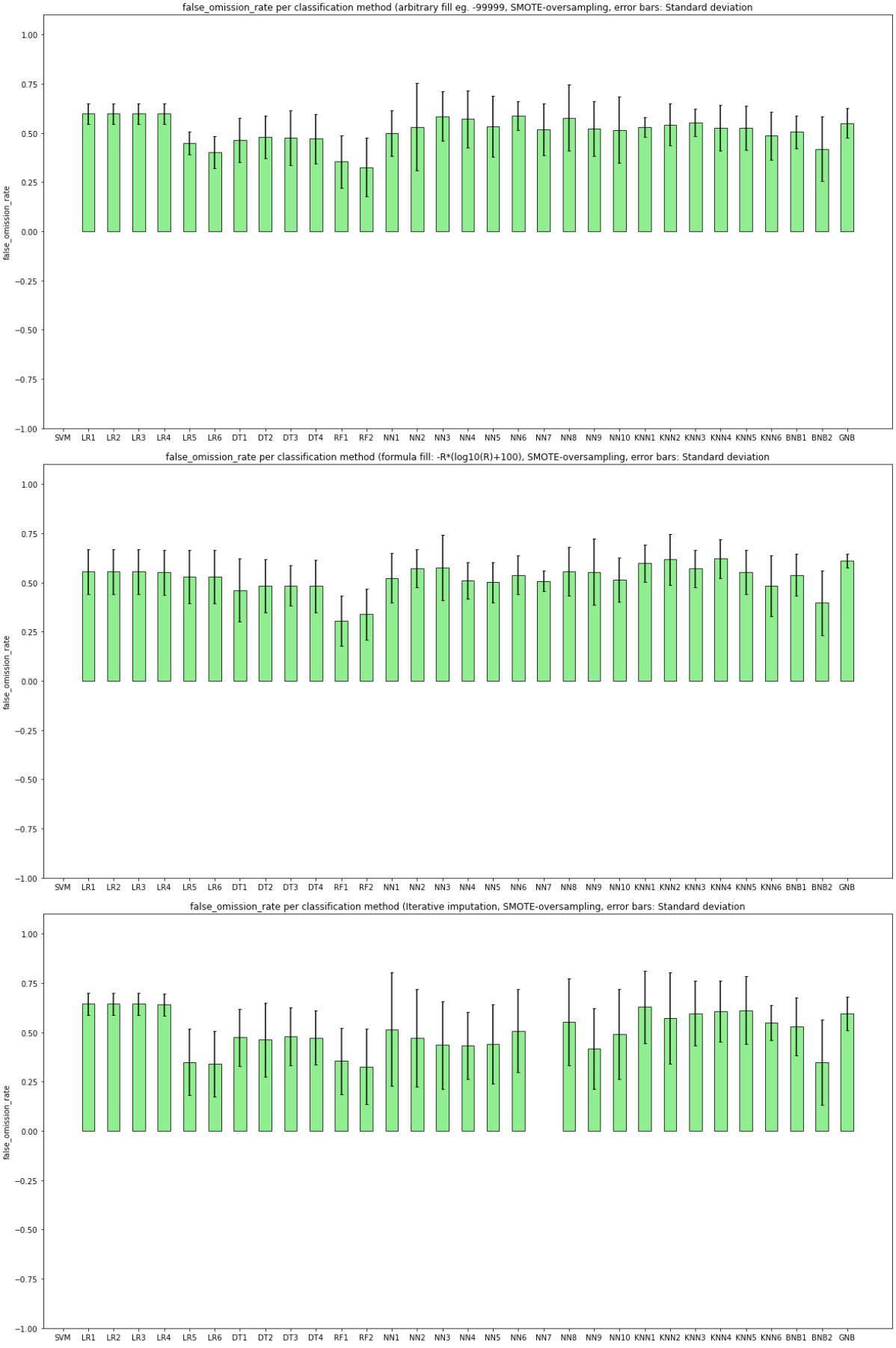

false_omission_rate per classification method (arbitrary fill eg. -99999, no undersampling, error bars: Standard deviation
false_omission_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB
false_omission_rate per classification method (formula fill: -R*(log10(R)+100), no undersampling, error bars: Standard deviation
false_omission_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB
false_omission_rate per classification method (Iterative imputation, no undersampling, error bars: Standard deviation
false_omission_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

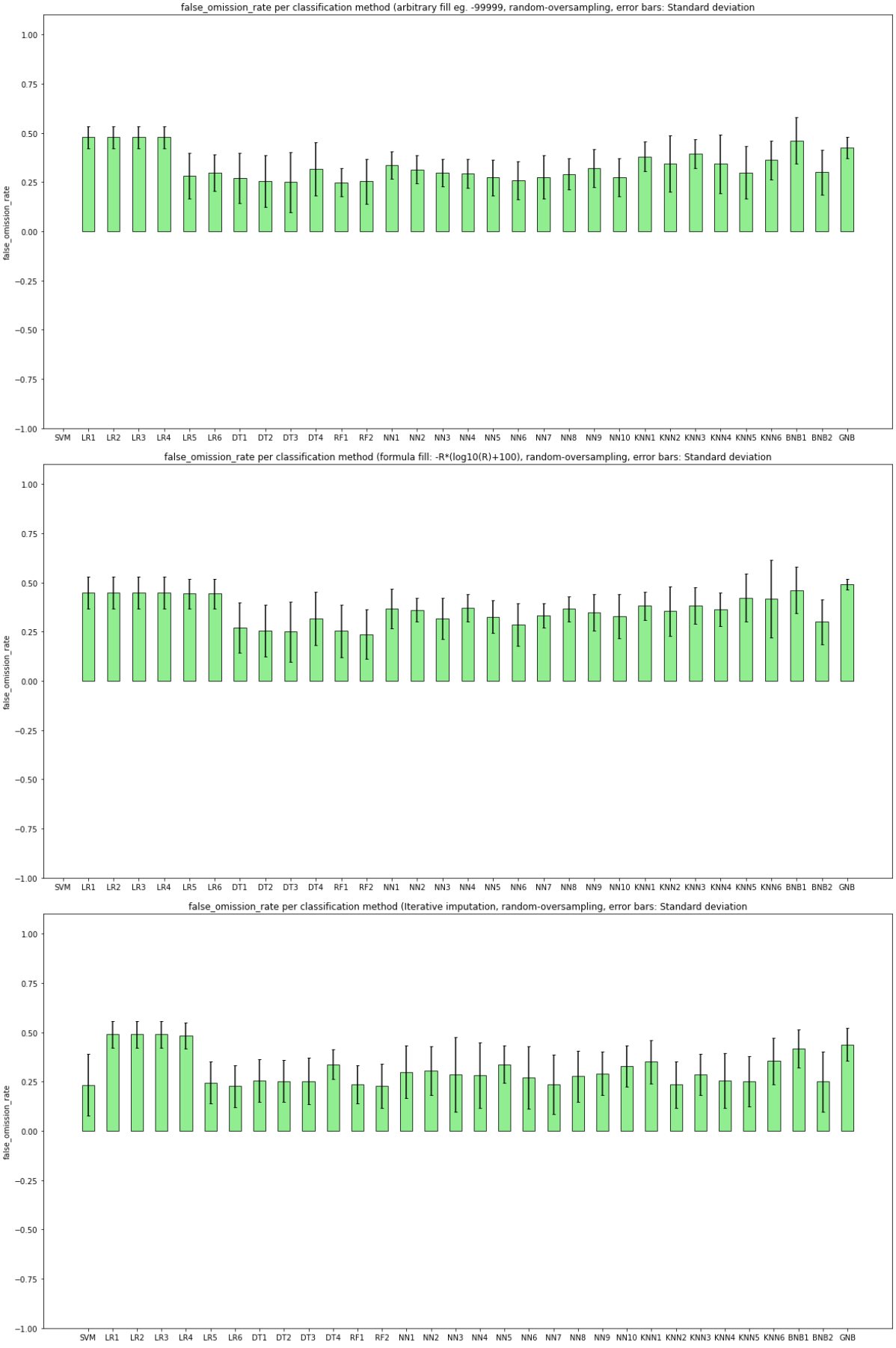

false_omission_rate per classification method (arbitrary fill eg. -99999, undersampling, error bars: Standard deviation
false_omission_rate per classification method (formula fill: -R*(log10(R)+100), undersampling, error bars: Standard deviation
false_omission_rate per classification method (Iterative imputation, undersampling, error bars: Standard deviation
false_omission_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

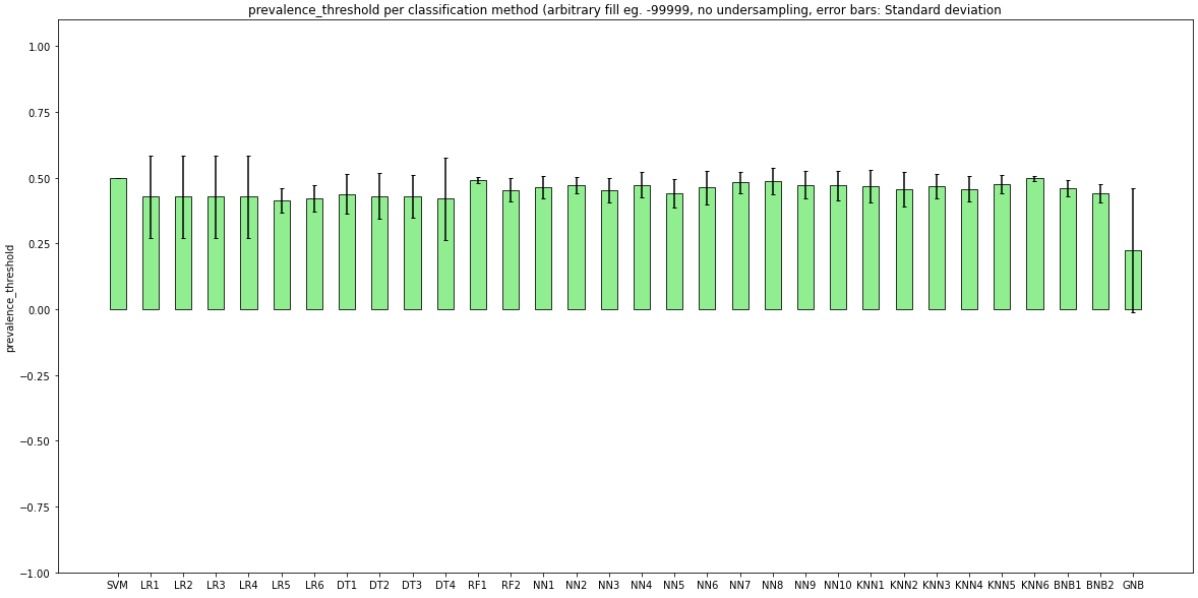

false_omission_rate per classification method (arbitrary fill eg. -99999, class weights(grid search), error bars: Standard deviation
false_omission_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
LR5
LR6
DT1
DT2
DT3
DT4
RF1
RF2

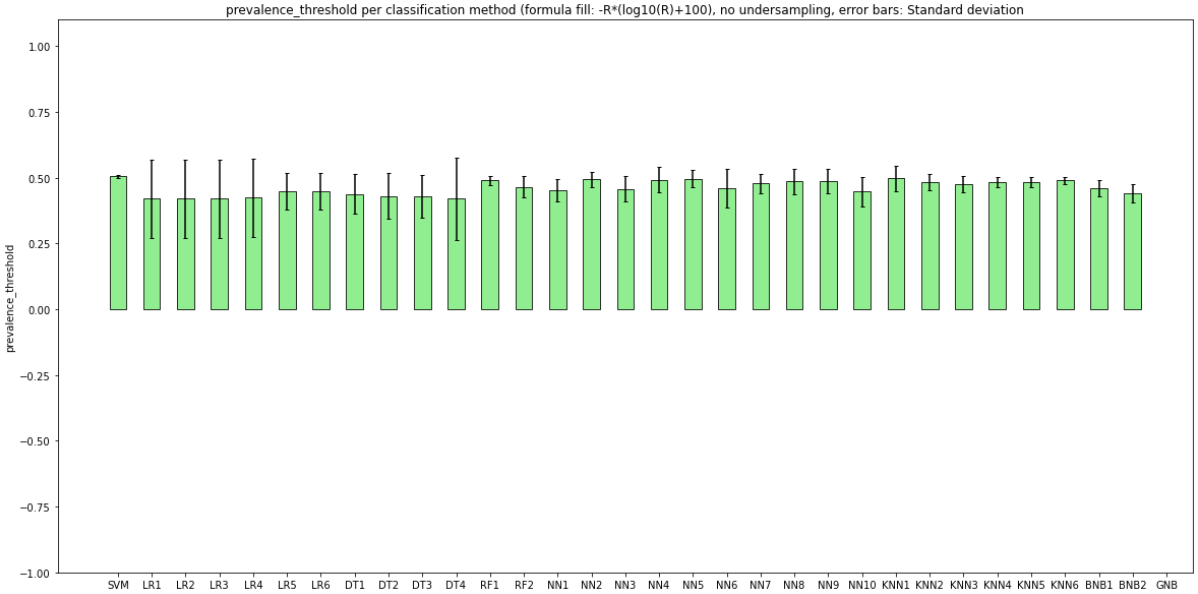

false_omission_rate per classification method (formula fill: -R*(log10(R)+100), class weights(grid search), error bars: Standard deviation
false_omission_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
LR1
LR2
LR3
LR4
LR5
LR6
DT1
DT2
DT3
DT4
RF1
RF2

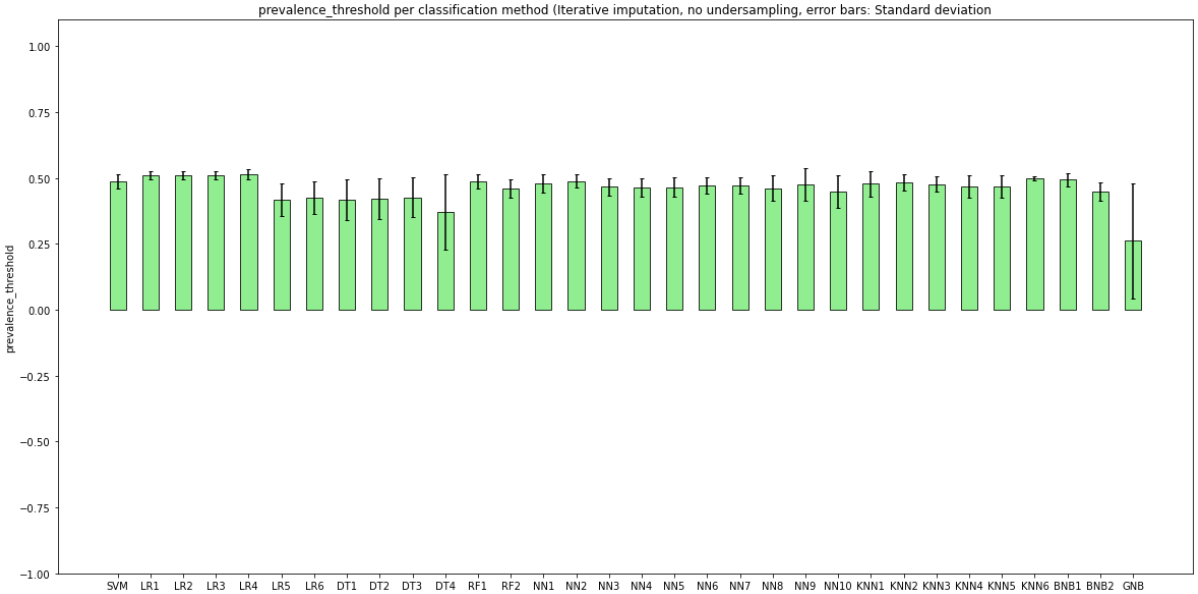

false_omission_rate per classification method (Iterative imputation, class weights(grid search), error bars: Standard deviation
false_omission_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
LR5
LR6
DT1
DT2
DT3

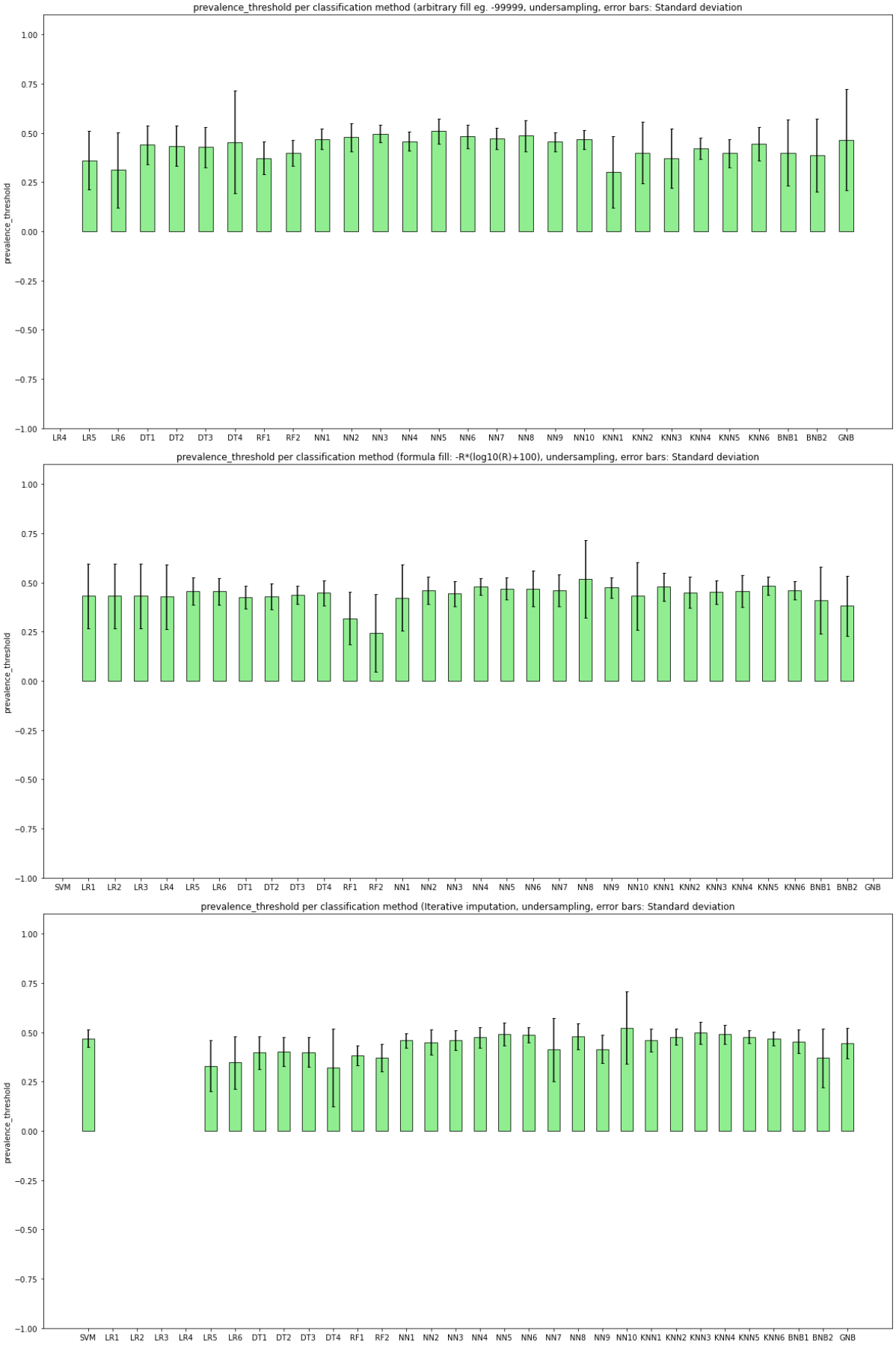

false_omission_rate per classification method (arbitrary fill eg. -99999, SMOTE-oversampling, error bars: Standard deviation
false_omission_rate per classification method (formula fill: -R*(log10(R)+100), SMOTE-oversampling, error bars: Standard deviation
false_omission_rate per classification method (Iterative imputation, SMOTE-oversampling, error bars: Standard deviation
false_omission_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

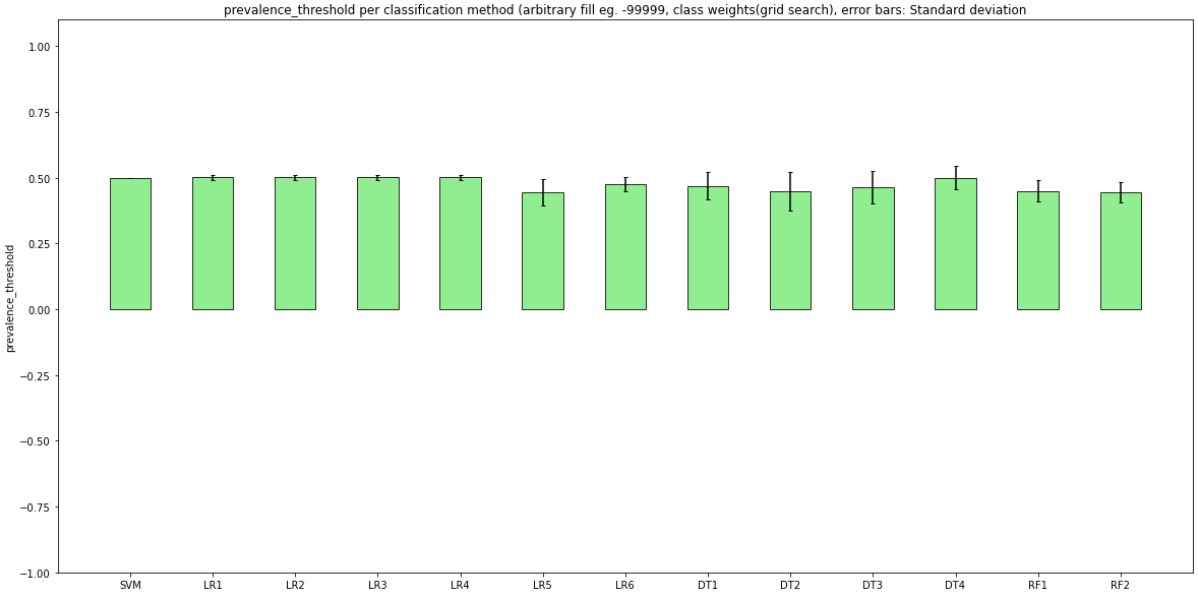

false_omission_rate per classification method (arbitrary fill eg. -99999, random-oversampling, error bars: Standard deviation
false_omission_rate per classification method (formula fill: -R*(log10(R)+100), random-oversampling, error bars: Standard deviation
false_omission_rate per classification method (Iterative imputation, random-oversampling, error bars: Standard deviation
false_omission_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

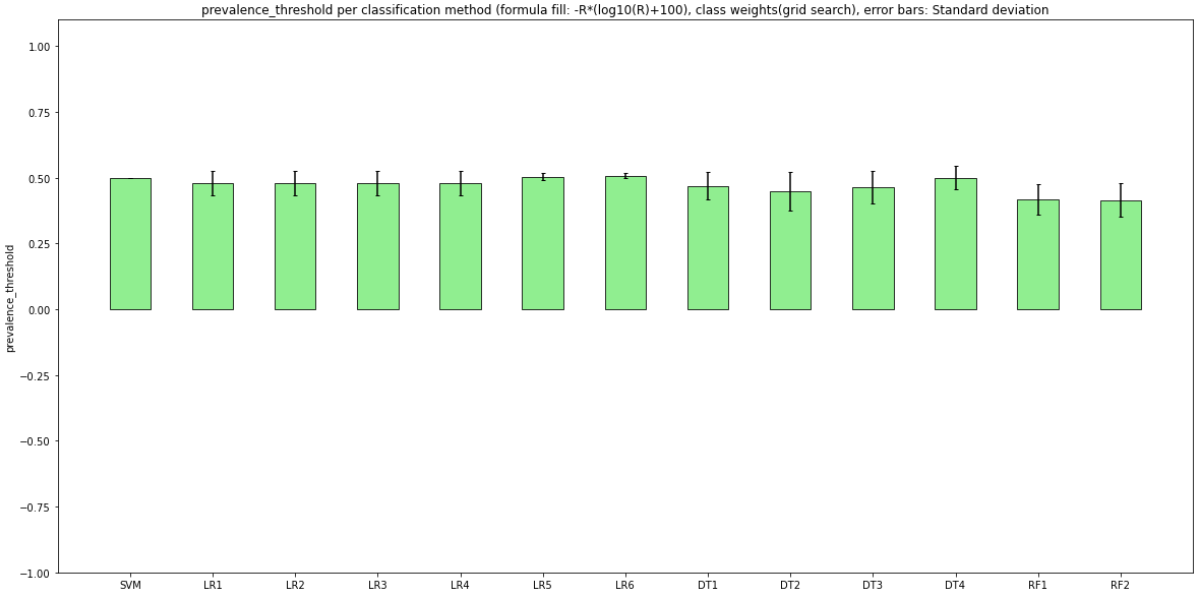

prevalence_threshold per classification method (arbitrary fill eg. -99999, no undersampling, error bars: Standard deviation
prevalence_threshold
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

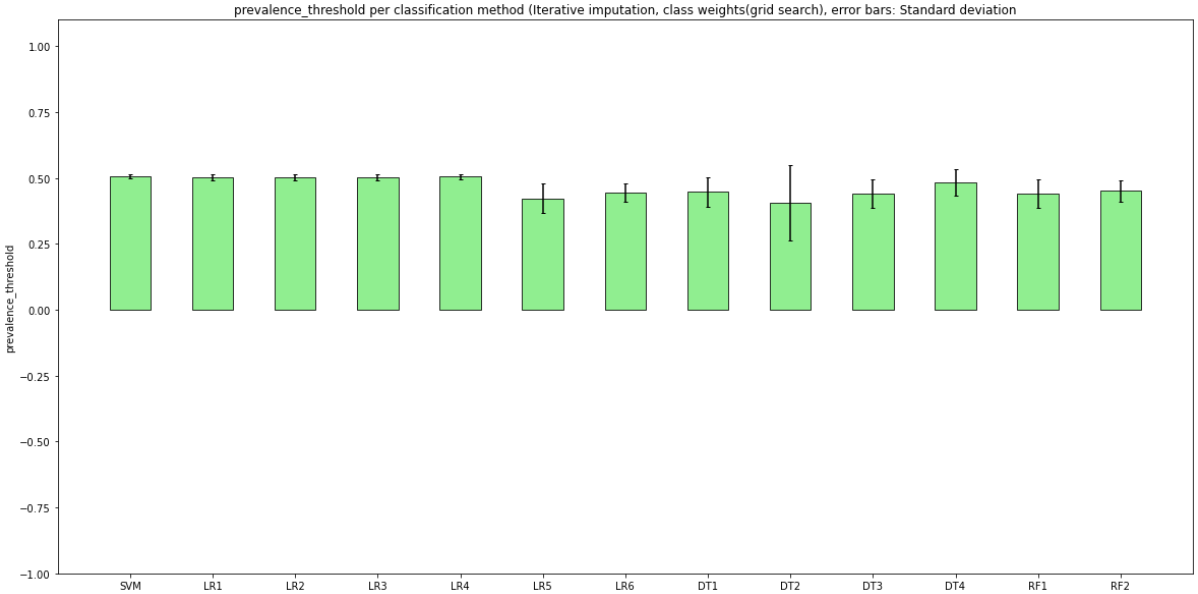

prevalence_threshold per classification method (formula fill: -R*(log10(R)+100), no undersampling, error bars: Standard deviation
prevalence_threshold
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

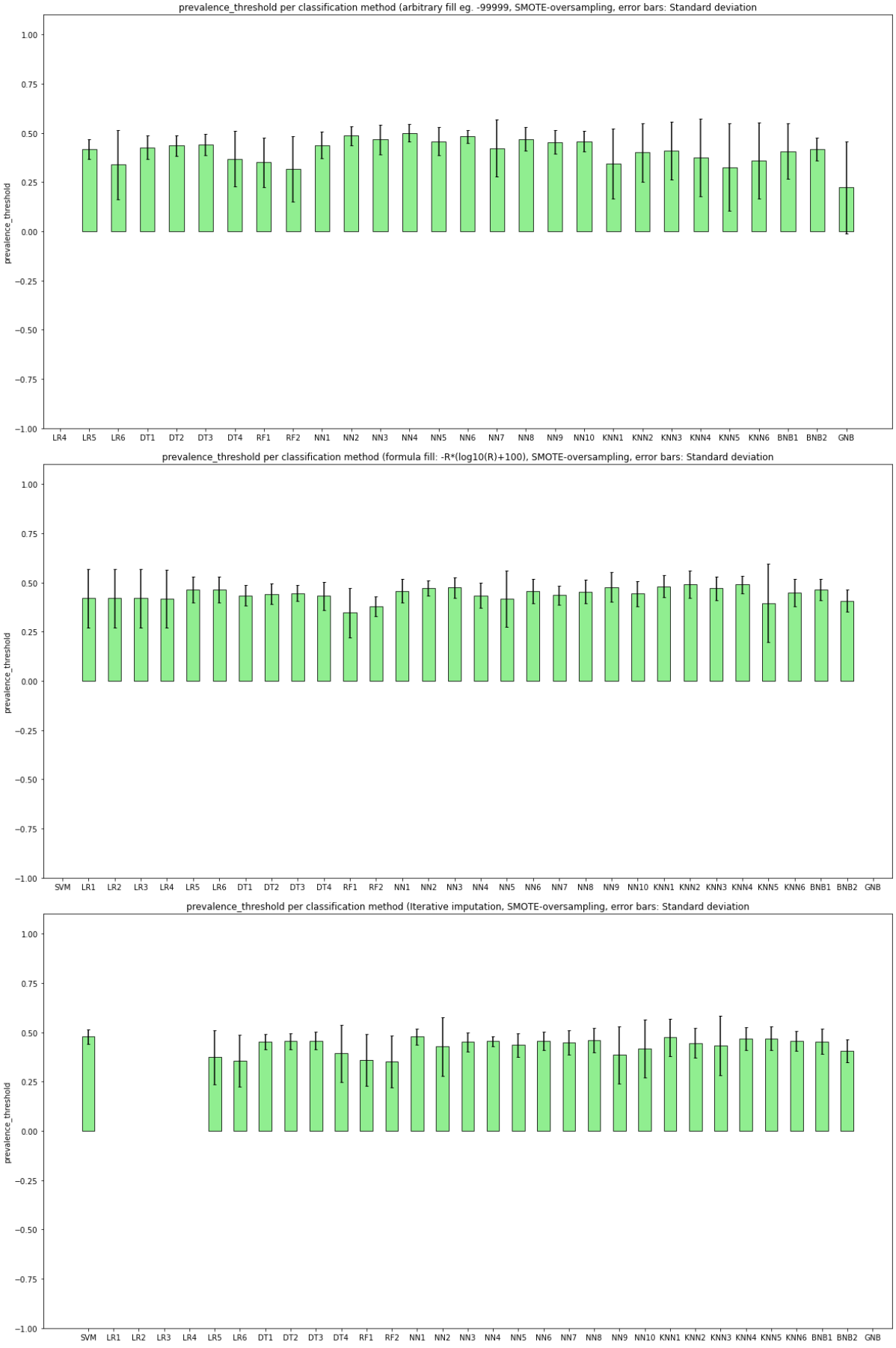

prevalence_threshold per classification method (Iterative imputation, no undersampling, error bars: Standard deviation
prevalence_threshold
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

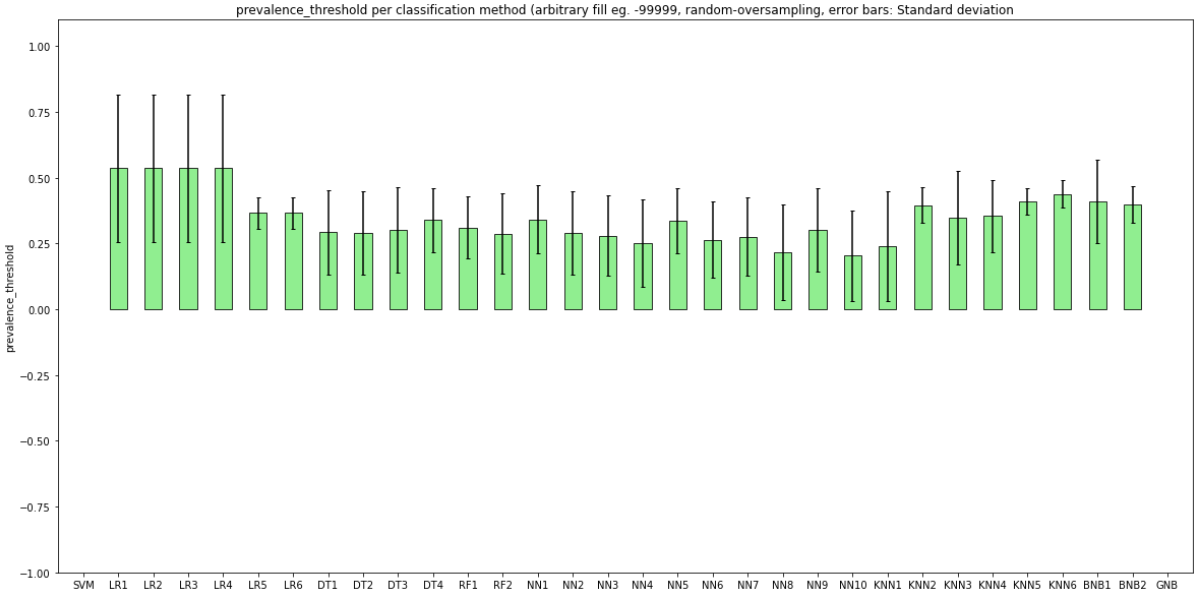

prevalence_threshold per classification method (arbitrary fill eg. -99999, undersampling, error bars: Standard deviation
prevalence_threshold per classification method (formula fill: -R*(log10(R)+100), undersampling, error bars: Standard deviation
prevalence_threshold per classification method (Iterative imputation, undersampling, error bars: Standard deviation
prevalence_threshold
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

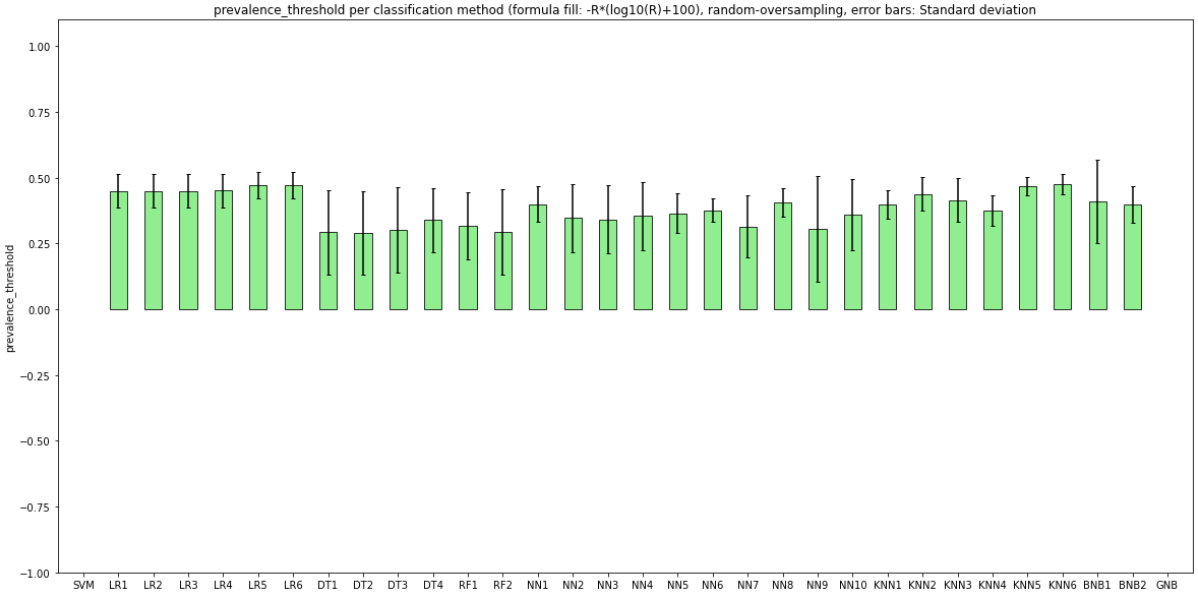

prevalence_threshold per classification method (arbitrary fill eg. -99999, class weights(grid search), error bars: Standard deviation
prevalence_threshold
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM
LR1
LR2
LR3
LR4
LR5
LR6
DT1
DT2
DT3
DT4
RF1
RF2

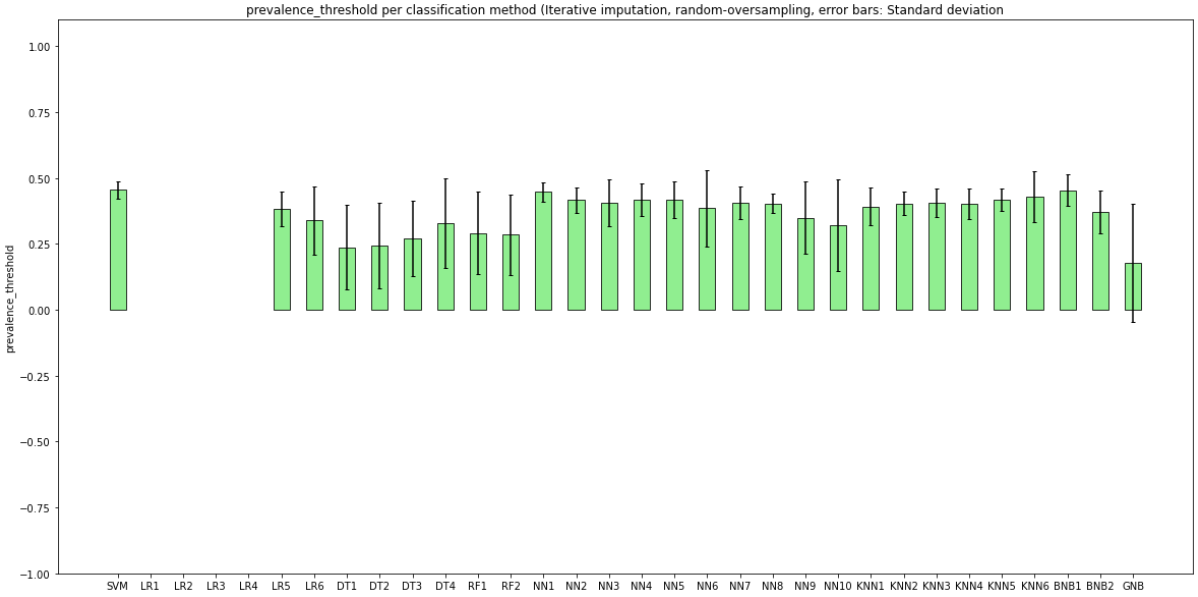

prevalence_threshold per classification method (formula fill: -R*(log10(R)+100), class weights(grid search), error bars: Standard deviation
prevalence_threshold
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM
LR1
LR2
LR3
LR4
LR5
LR6
DT1
DT2
DT3
DT4
RF1
RF2

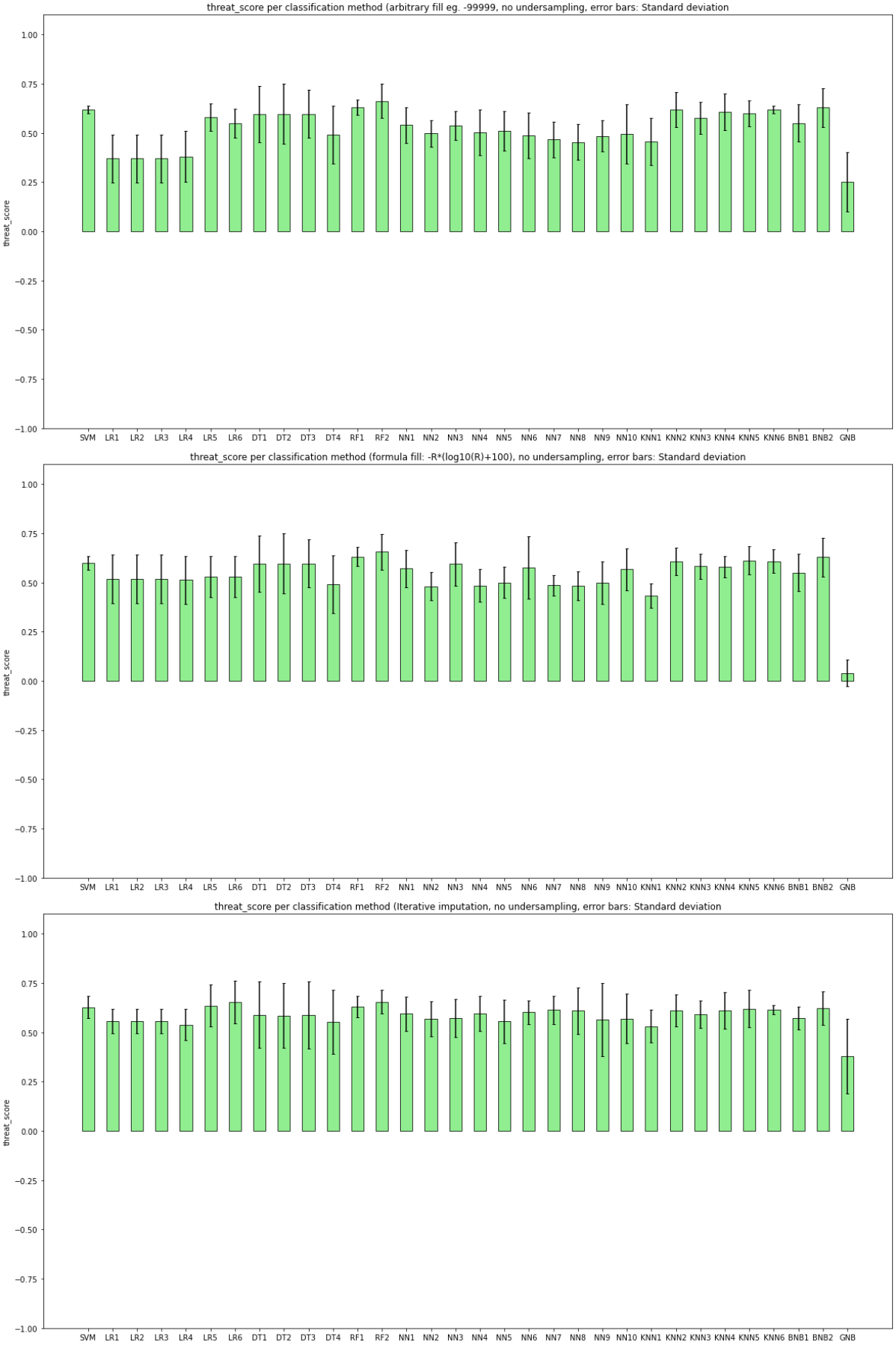

prevalence_threshold per classification method (Iterative imputation, class weights(grid search), error bars: Standard deviation
prevalence_threshold
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM
LR1
LR2
LR3
LR4
LR5
LR6
DT1
DT2
DT3
DT4
RF1
RF2

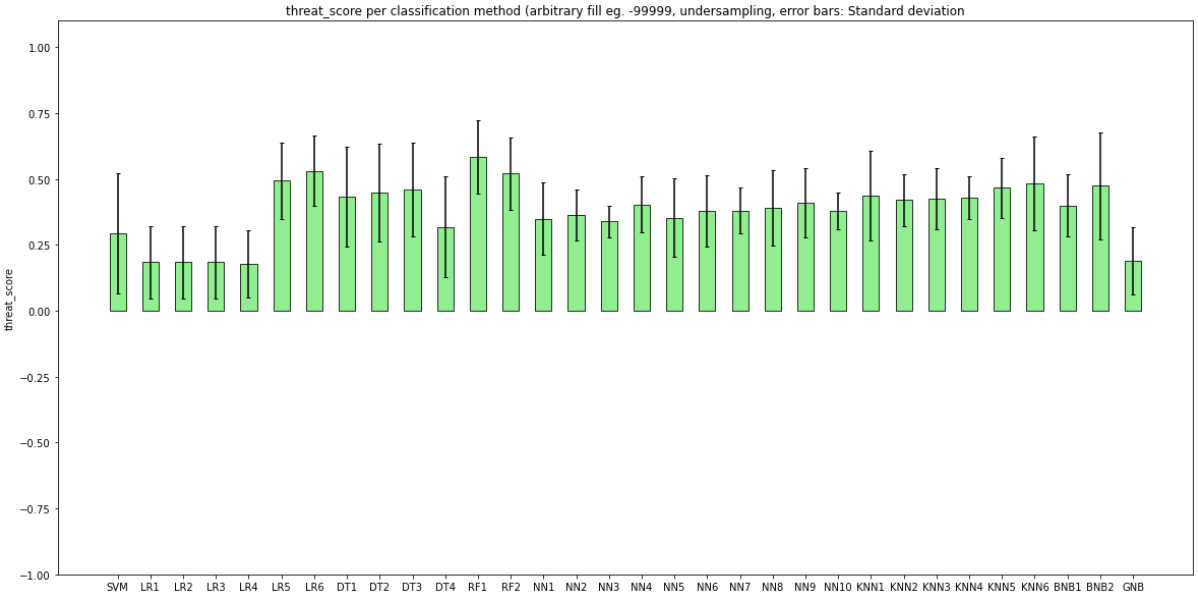

prevalence_threshold per classification method (arbitrary fill eg. -99999, SMOTE-oversampling, error bars: Standard deviation
prevalence_threshold per classification method (formula fill: -R*(log10(R)+100), SMOTE-oversampling, error bars: Standard deviation
prevalence_threshold per classification method (Iterative imputation, SMOTE-oversampling, error bars: Standard deviation
prevalence_threshold
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

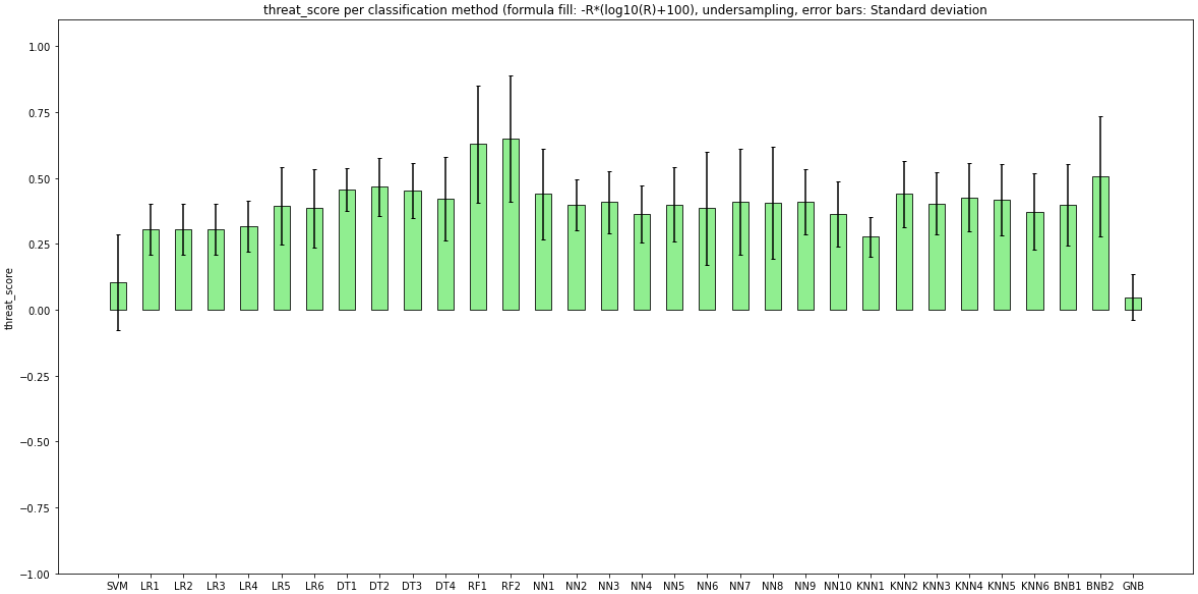

prevalence_threshold per classification method (arbitrary fill eg. -99999, random-oversampling, error bars: Standard deviation
prevalence_threshold
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

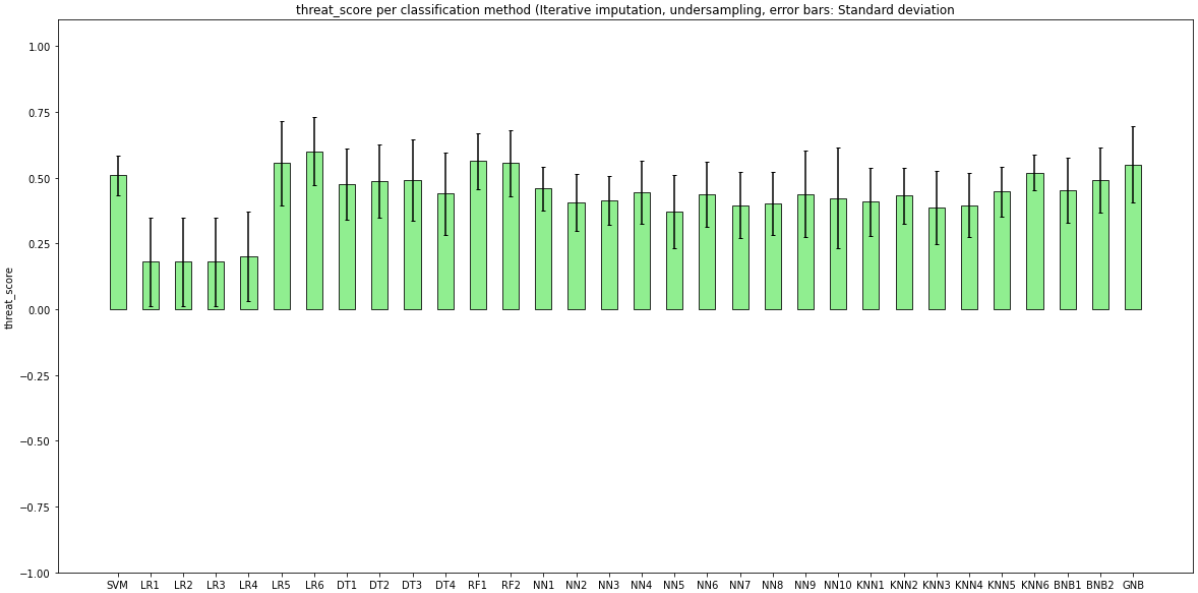

prevalence_threshold per classification method (formula fill: -R*(log10(R)+100), random-oversampling, error bars: Standard deviation
prevalence_threshold
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

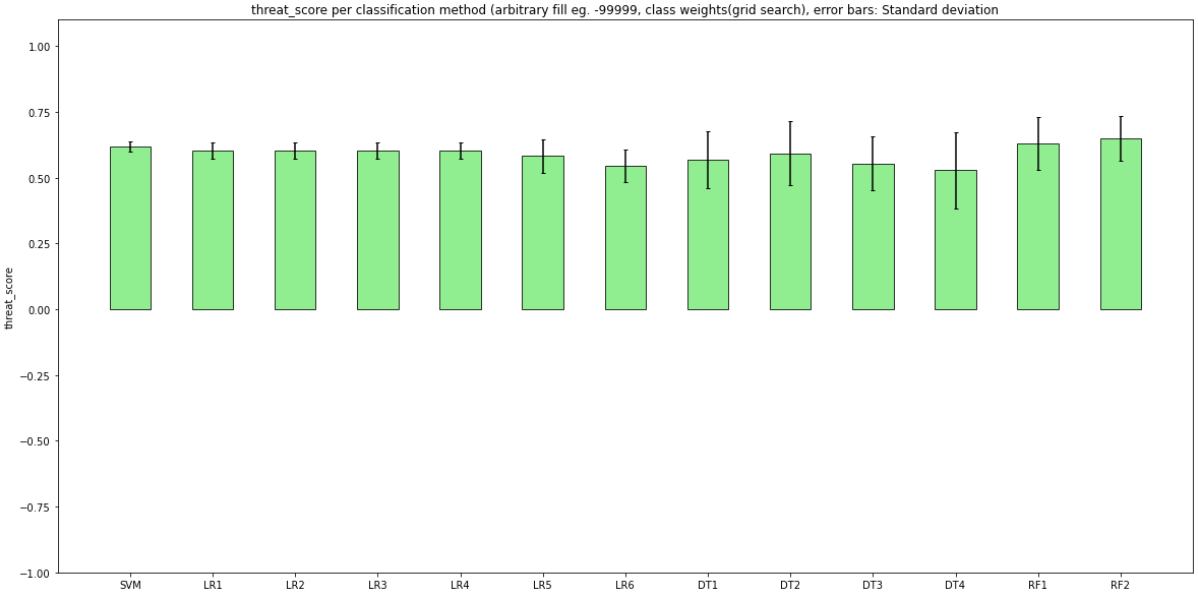

prevalence_threshold per classification method (Iterative imputation, random-oversampling, error bars: Standard deviation
prevalence_threshold
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

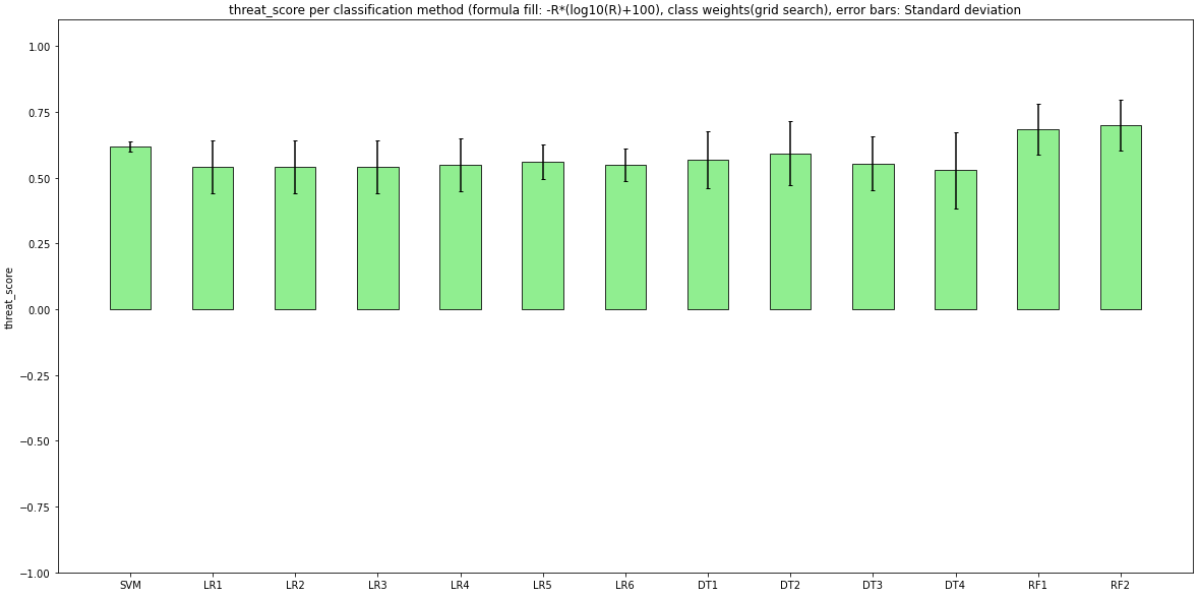

threat_score per classification method (arbitrary fill eg. -99999, no undersampling, error bars: Standard deviation
threat_score per classification method (formula fill: -R*(log10(R)+100), no undersampling, error bars: Standard deviation
threat_score per classification method (Iterative imputation, no undersampling, error bars: Standard deviation
threat_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

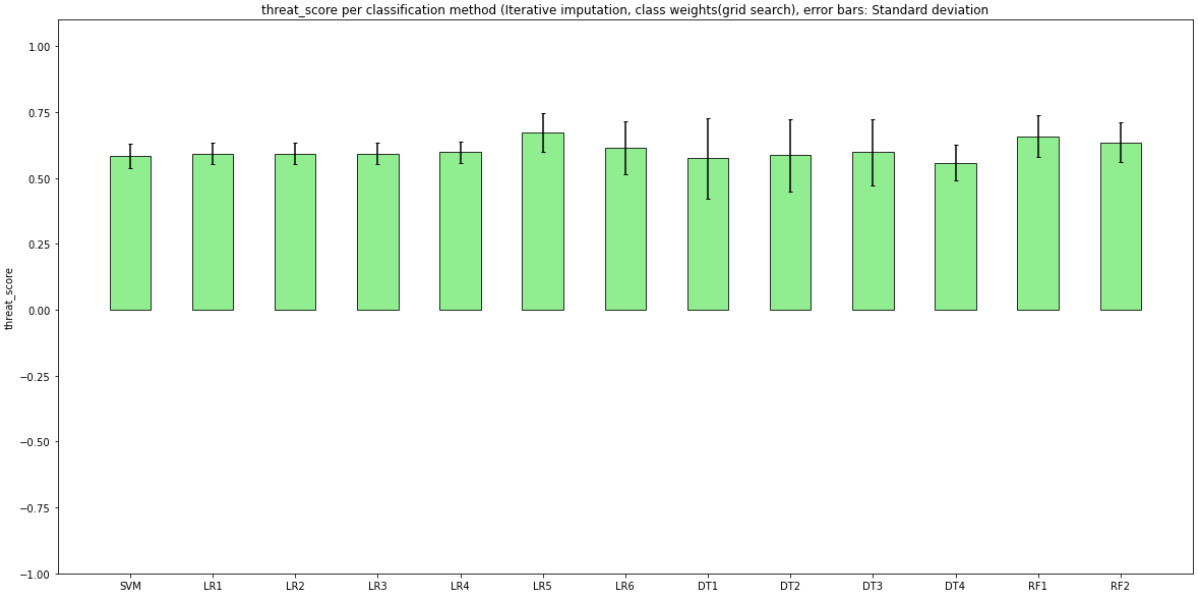

threat_score per classification method (arbitrary fill eg. -99999, undersampling, error bars: Standard deviation
threat_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

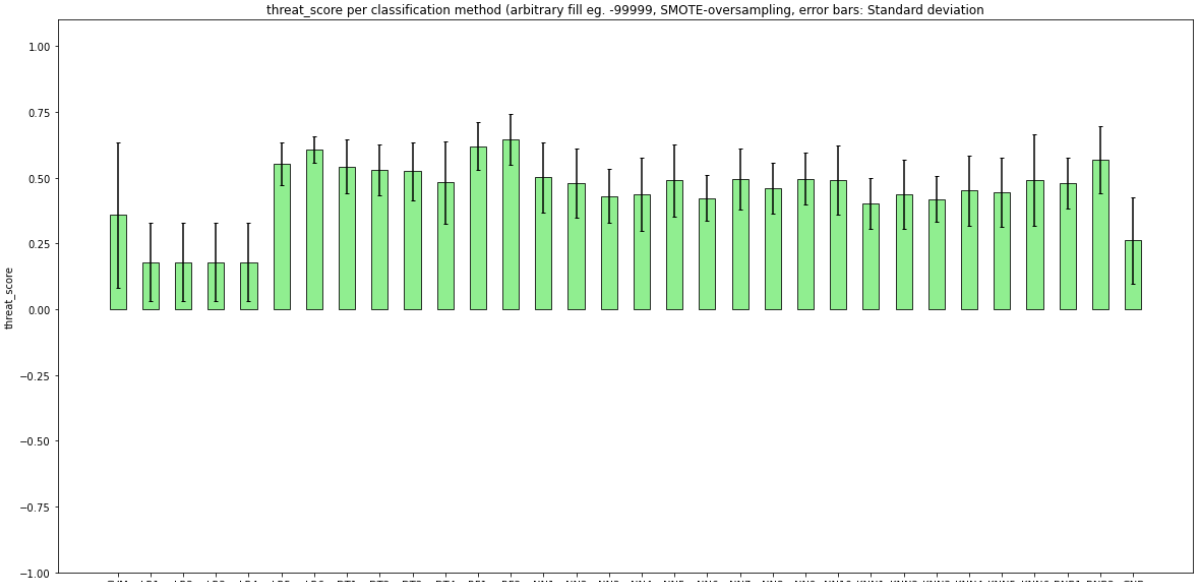

threat_score per classification method (formula fill: -R*(log10(R)+100), undersampling, error bars: Standard deviation
threat_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

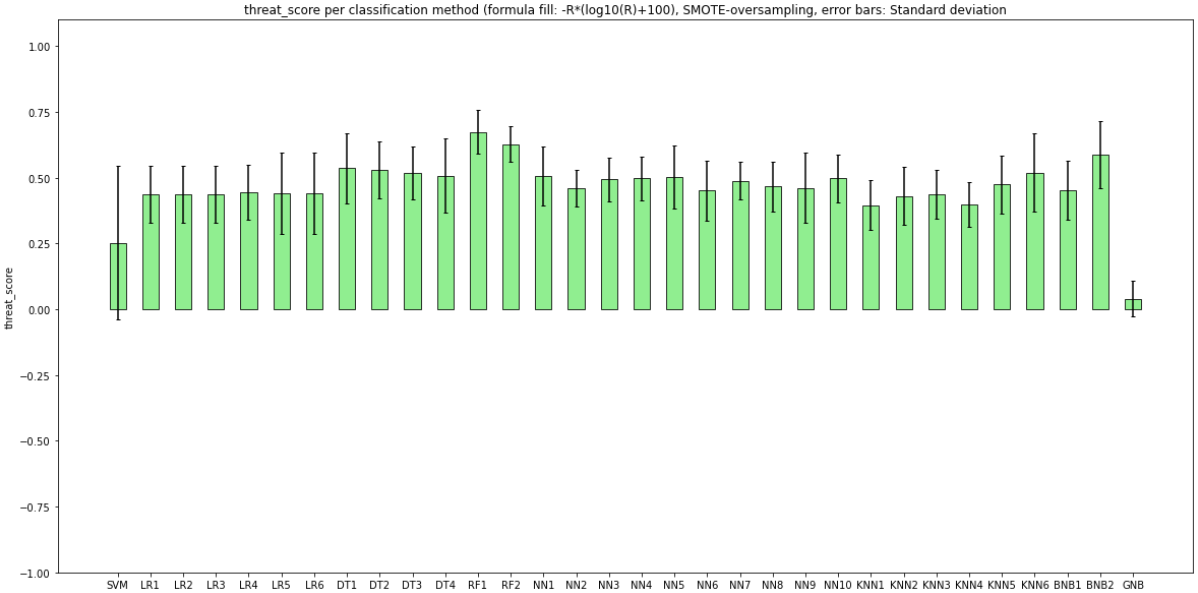

threat_score per classification method (Iterative imputation, undersampling, error bars: Standard deviation
threat_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

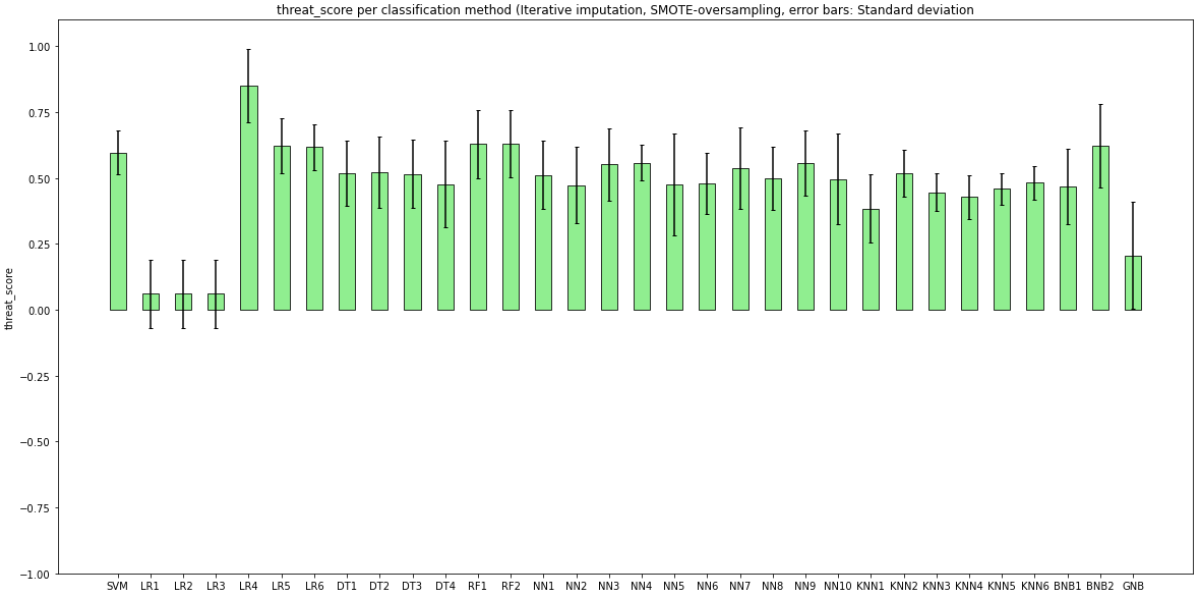

threat_score per classification method (arbitrary fill eg. -99999, class weights(grid search), error bars: Standard deviation
threat_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM
LR1
LR2
LR3
LR4
LR5
LR6
DT1
DT2
DT3
DT4
RF1
RF2

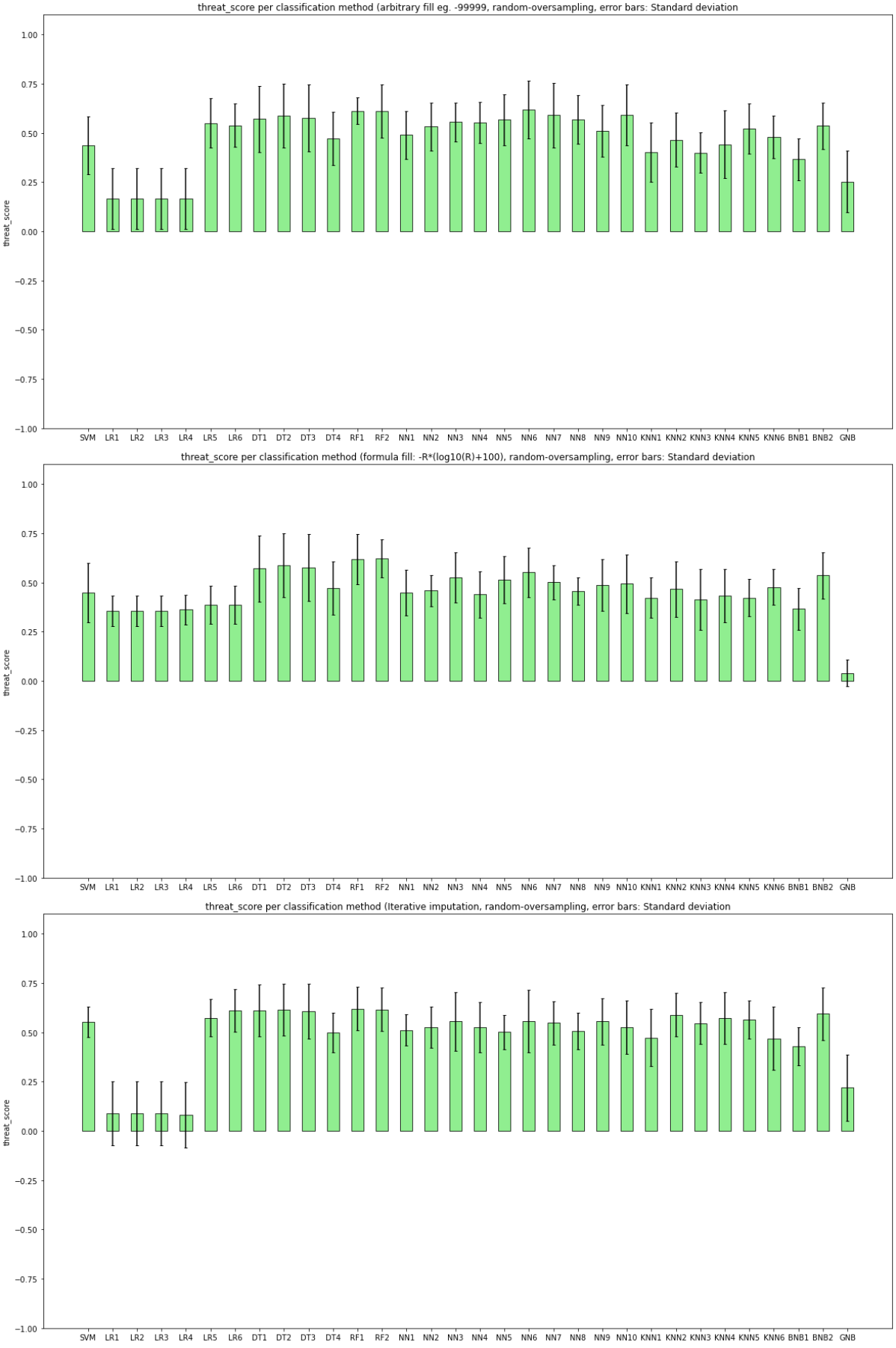

threat_score per classification method (formula fill: -R*(log10(R)+100), class weights(grid search), error bars: Standard deviation
threat_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM
LR1
LR2
LR3
LR4
LR5
LR6
DT1
DT2
DT3
DT4
RF1
RF2

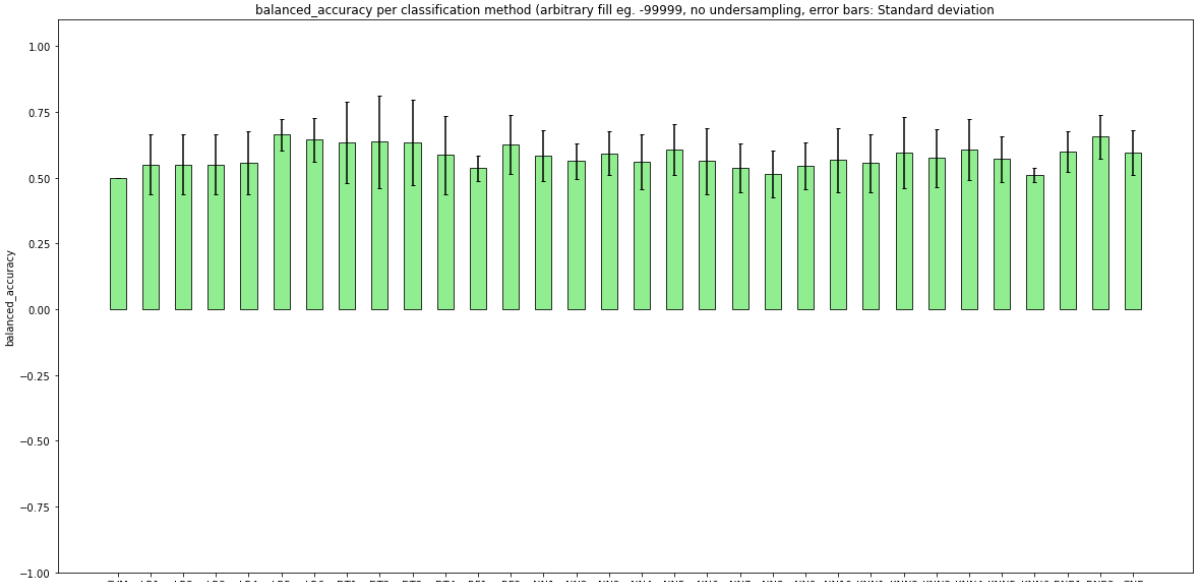

threat_score per classification method (Iterative imputation, class weights(grid search), error bars: Standard deviation
threat_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM
LR1
LR2
LR3
LR4
LR5
LR6
DT1
DT2
DT3
DT4
RF1
RF2

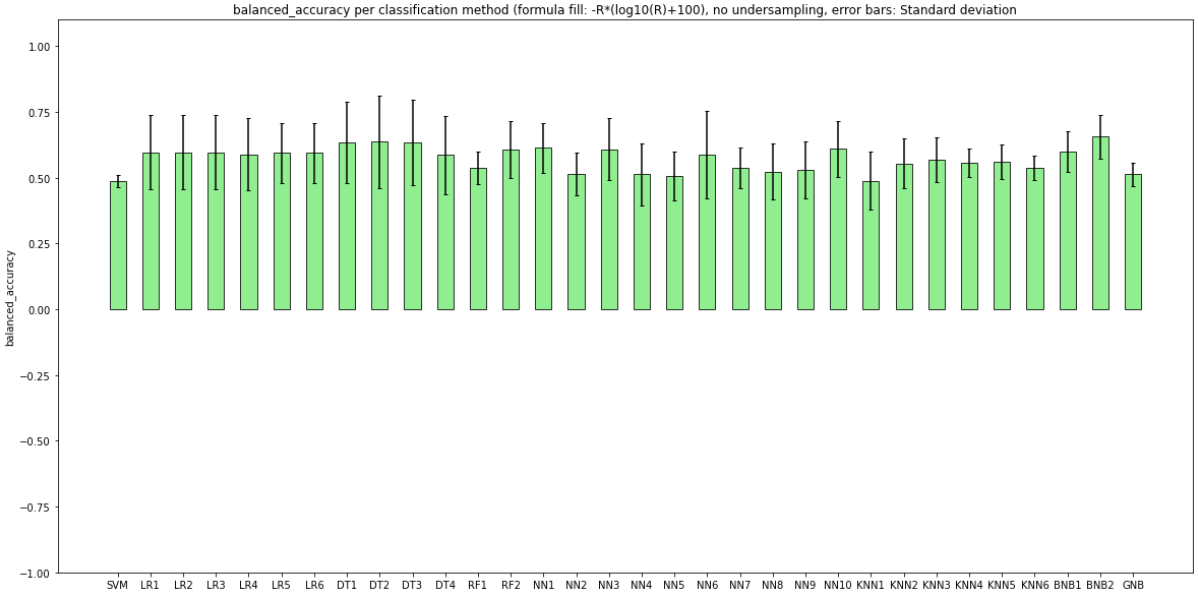

threat_score per classification method (arbitrary fill eg. -99999, SMOTE-oversampling, error bars: Standard deviation
threat_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

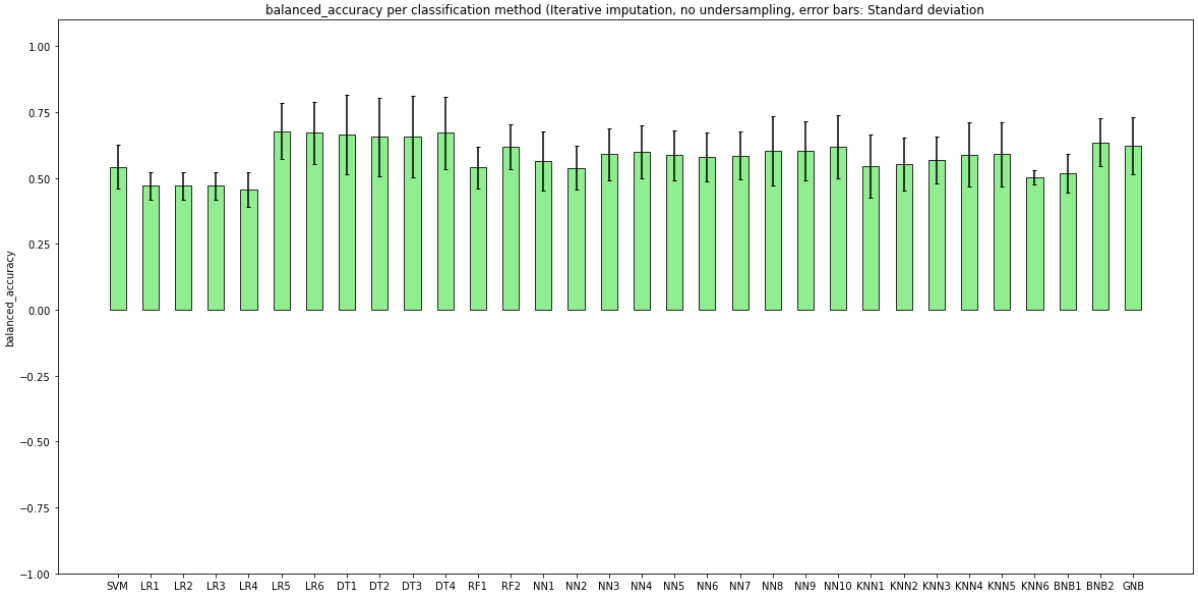

threat_score per classification method (formula fill: -R*(log10(R)+100), SMOTE-oversampling, error bars: Standard deviation
threat_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

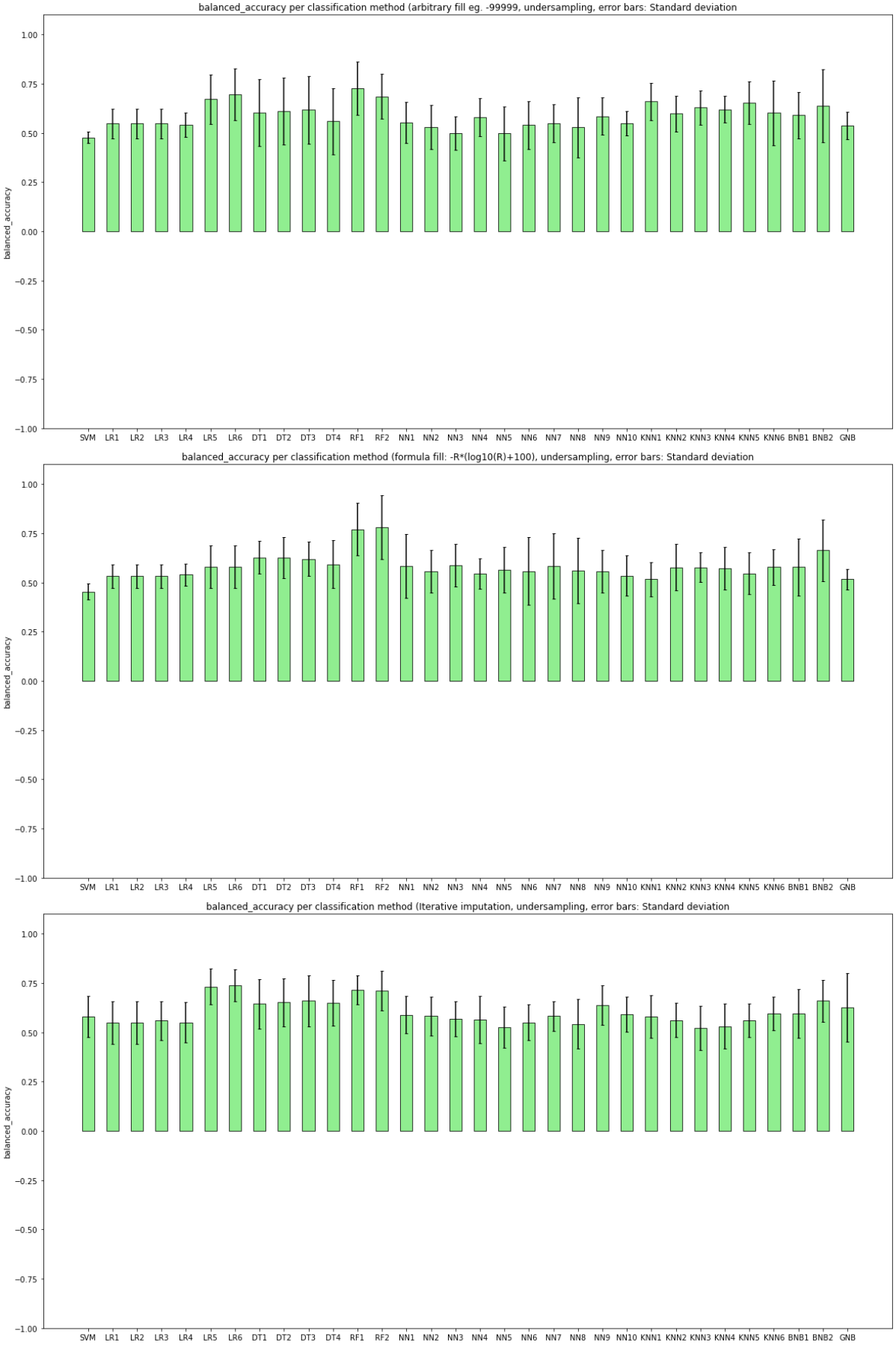

threat_score per classification method (Iterative imputation, SMOTE-oversampling, error bars: Standard deviation
threat_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

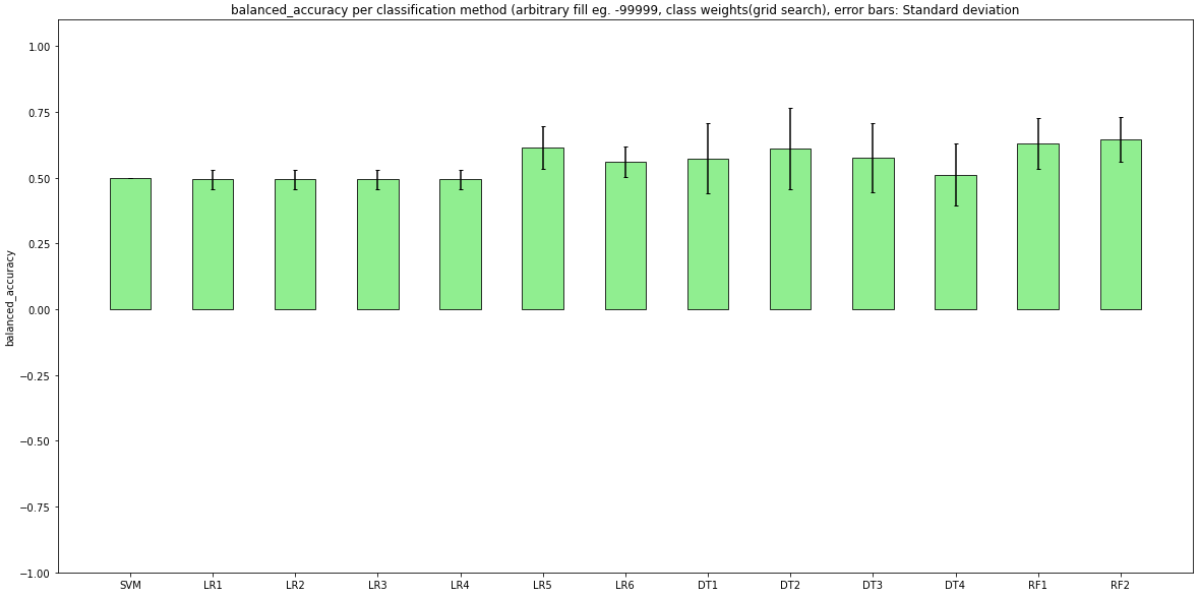

threat_score per classification method (arbitrary fill eg. -99999, random-oversampling, error bars: Standard deviation
threat_score per classification method (formula fill: -R*(log10(R)+100), random-oversampling, error bars: Standard deviation
threat_score per classification method (Iterative imputation, random-oversampling, error bars: Standard deviation
threat_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

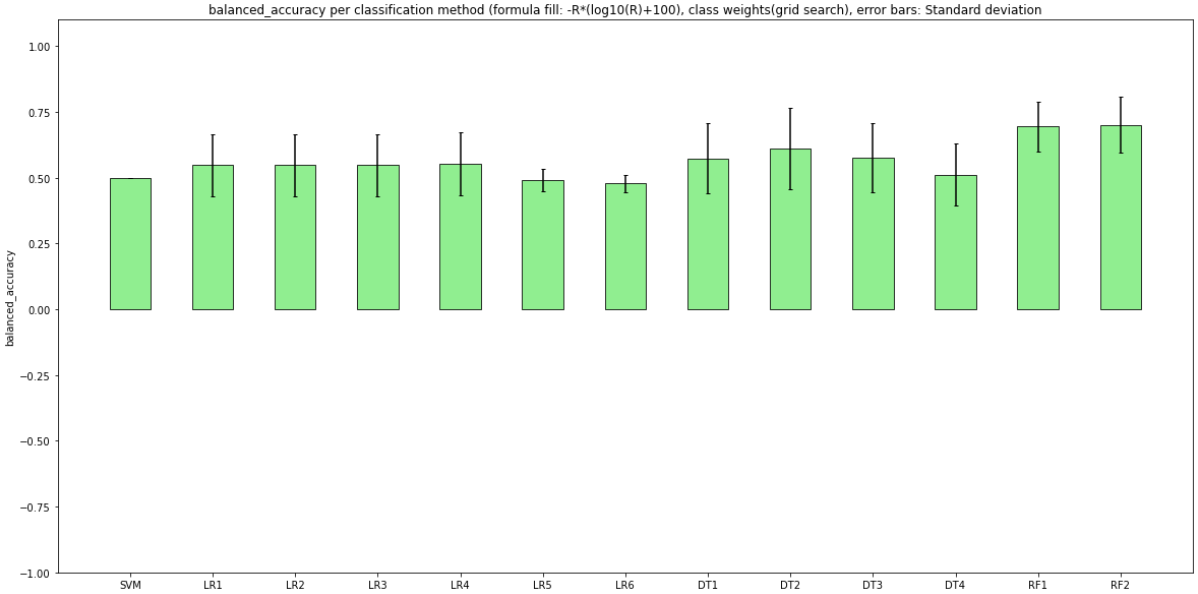

balanced_accuracy per classification method (arbitrary fill eg. -99999, no undersampling, error bars: Standard deviation
balanced_accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

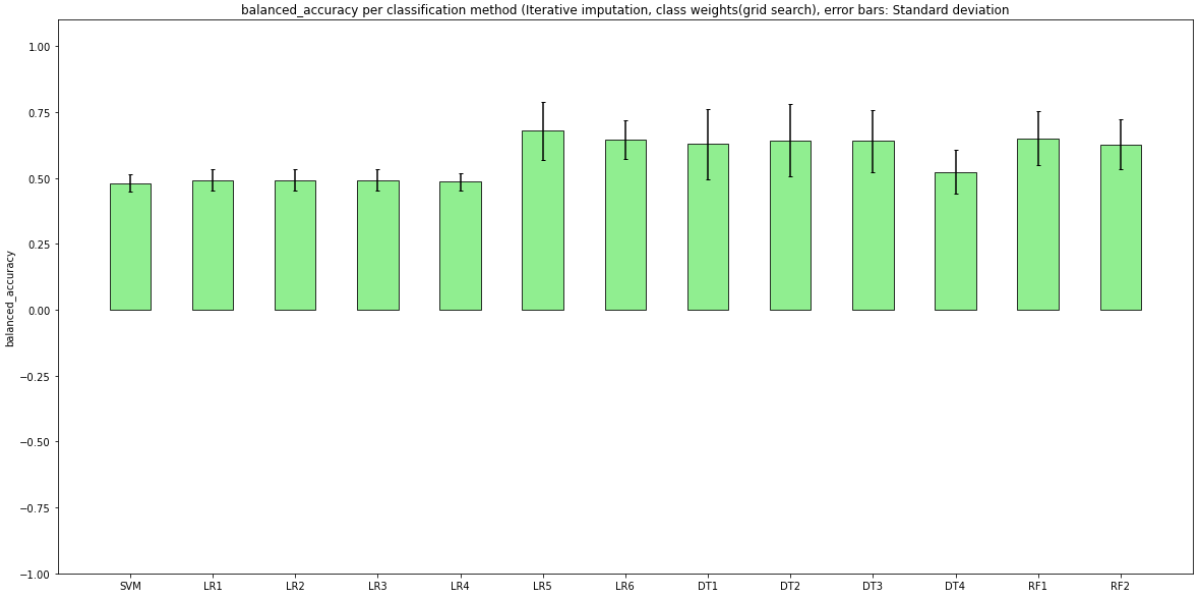

balanced_accuracy per classification method (formula fill: -R*(log10(R)+100), no undersampling, error bars: Standard deviation
balanced_accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

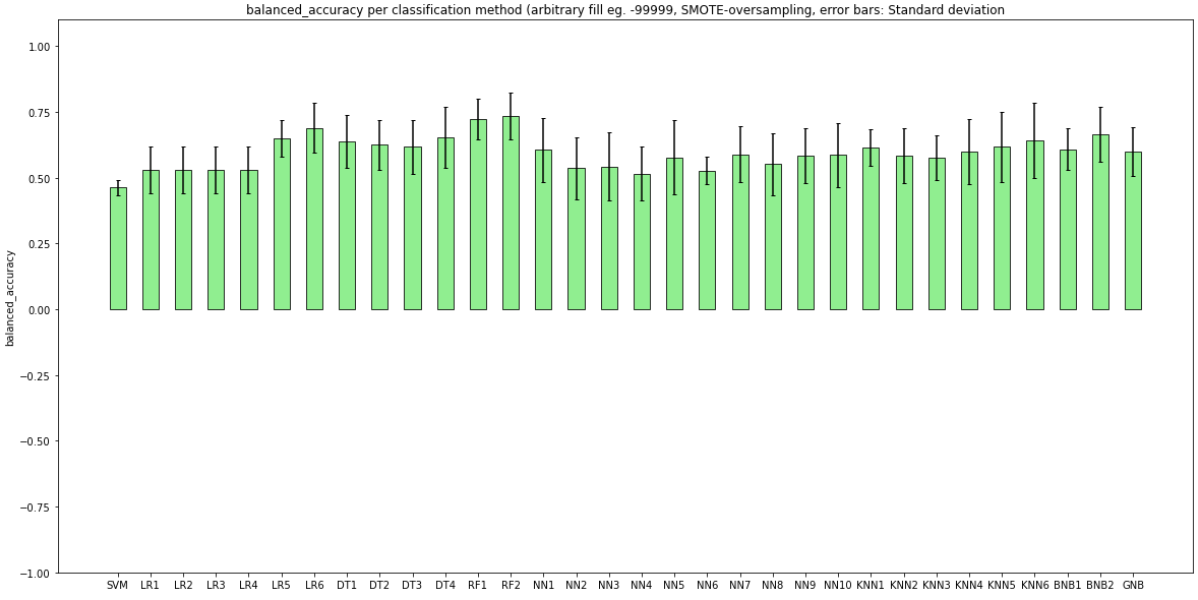

balanced_accuracy per classification method (Iterative imputation, no undersampling, error bars: Standard deviation
balanced_accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

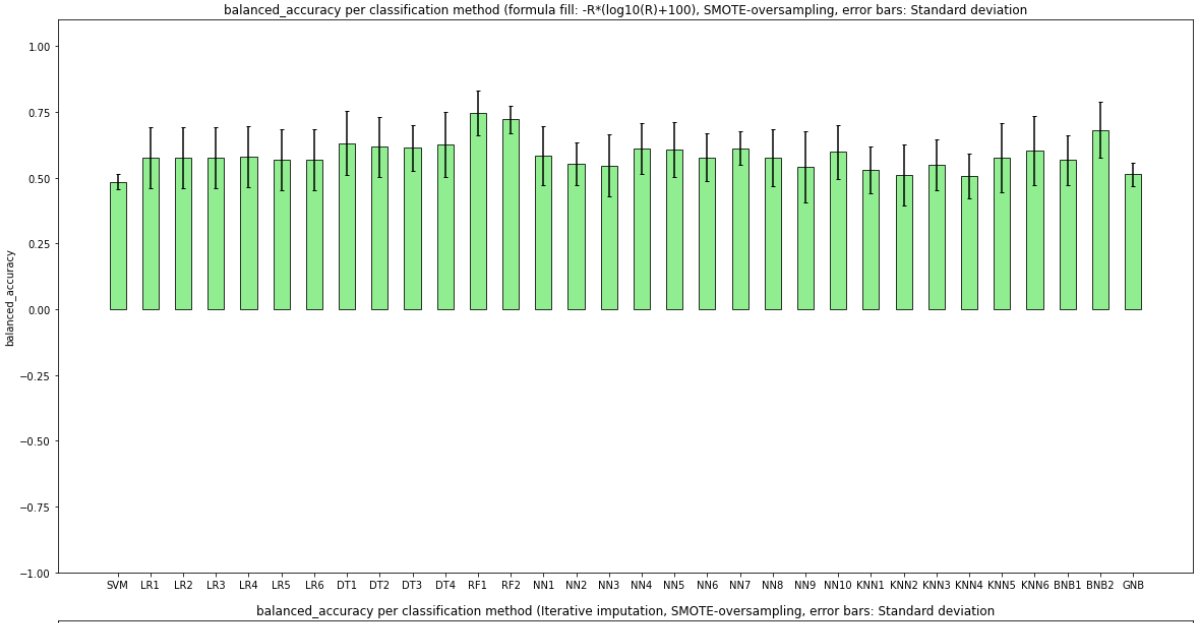

balanced_accuracy per classification method (arbitrary fill eg. -99999, undersampling, error bars: Standard deviation
balanced_accuracy per classification method (formula fill: -R*(log10(R)+100), undersampling, error bars: Standard deviation
balanced_accuracy per classification method (Iterative imputation, undersampling, error bars: Standard deviation
balanced_accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

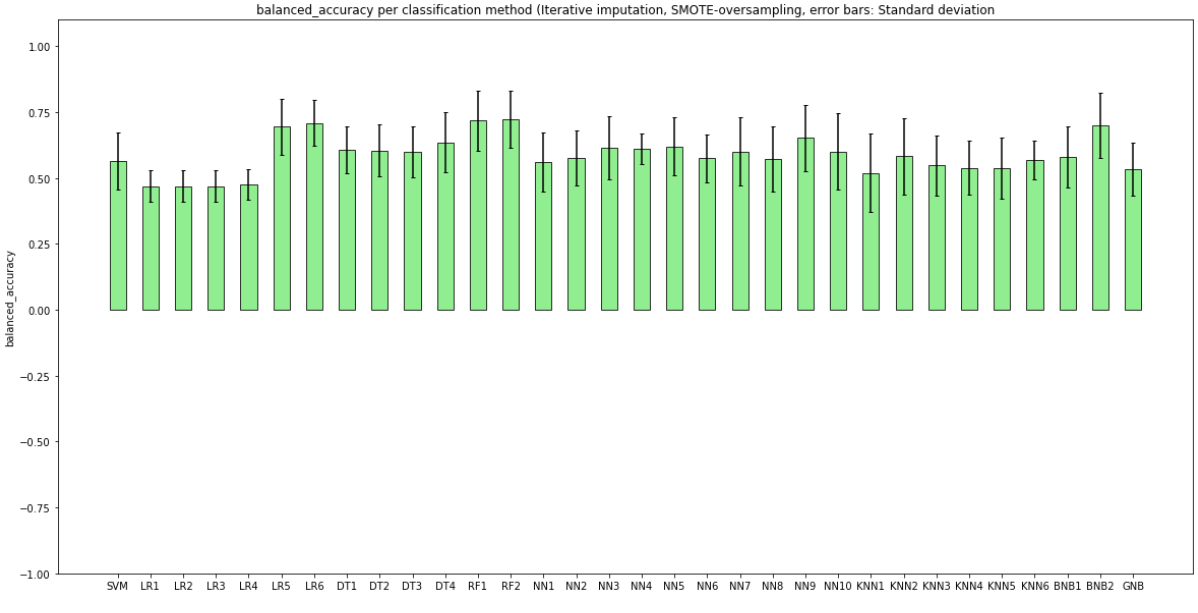

balanced_accuracy per classification method (arbitrary fill eg. -99999, class weights(grid search), error bars: Standard deviation
balanced_accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM
LR1
LR2
LR3
LR4
LR5
LR6
DT1
DT2
DT3
DT4
RF1
RF2

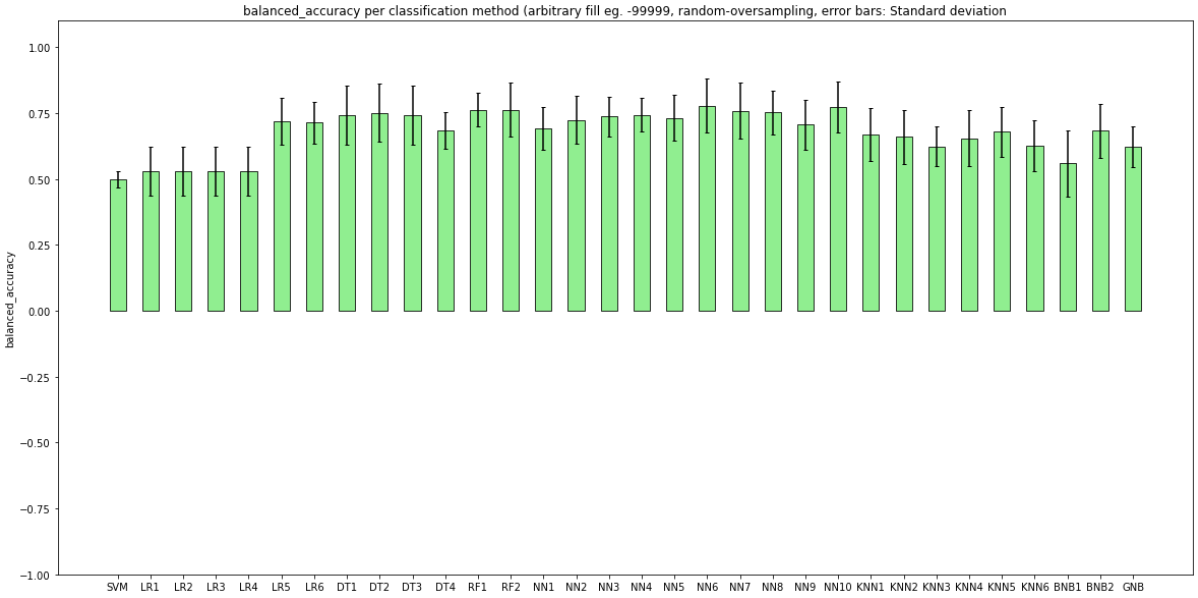

balanced_accuracy per classification method (formula fill: -R*(log10(R)+100), class weights(grid search), error bars: Standard deviation
balanced_accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM
LR1
LR2
LR3
LR4
LR5
LR6
DT1
DT2
DT3
DT4
RF1
RF2

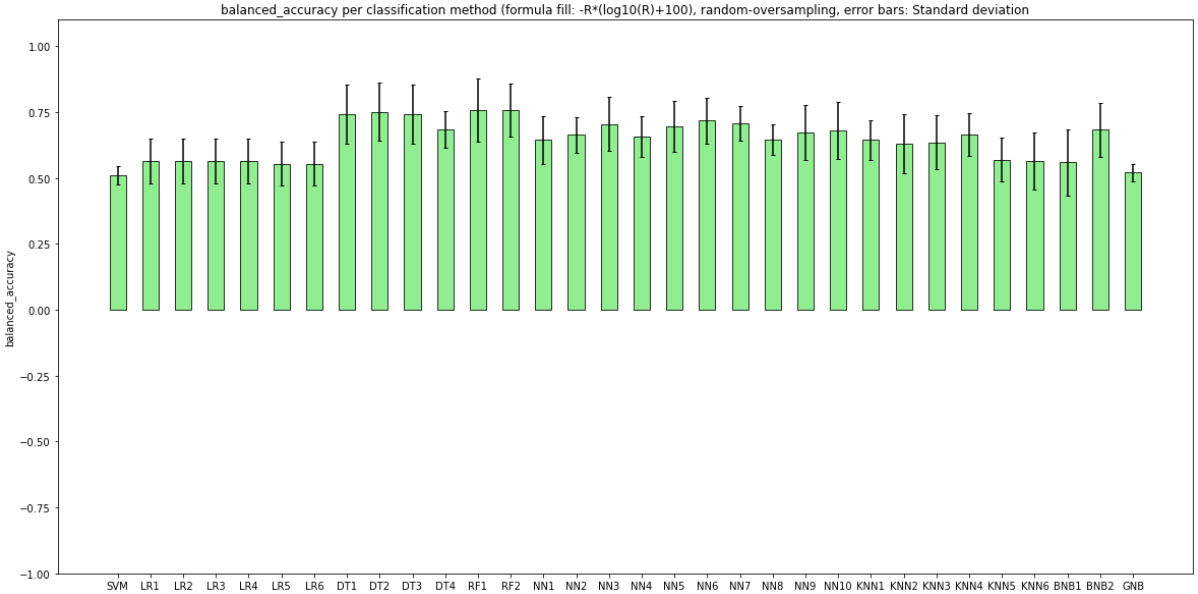

balanced_accuracy per classification method (Iterative imputation, class weights(grid search), error bars: Standard deviation
balanced_accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM
LR1
LR2
LR3
LR4
LR5
LR6
DT1
DT2
DT3
DT4
RF1
RF2

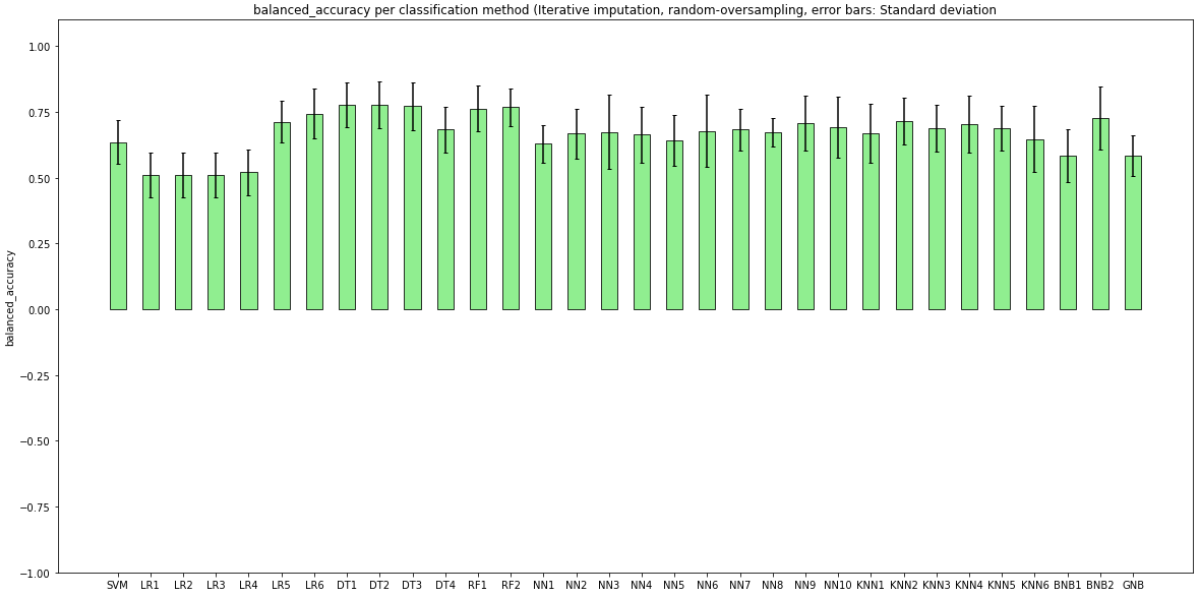

balanced_accuracy per classification method (arbitrary fill eg. -99999, SMOTE-oversampling, error bars: Standard deviation
balanced_accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

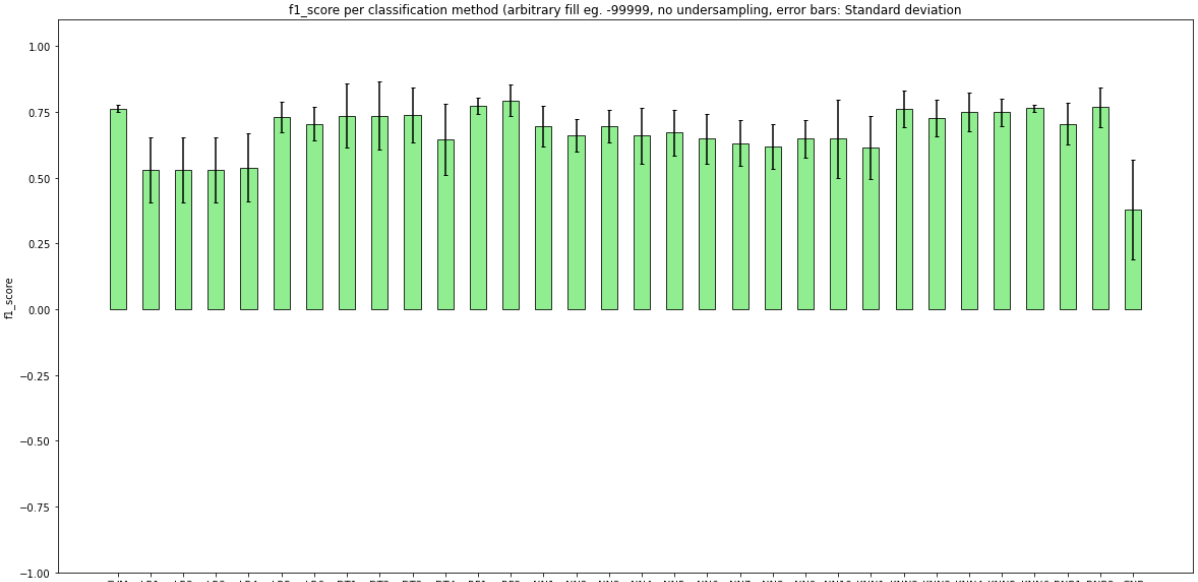

balanced_accuracy per classification method (formula fill: -R*(log10(R)+100), SMOTE-oversampling, error bars: Standard deviation
balanced_accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

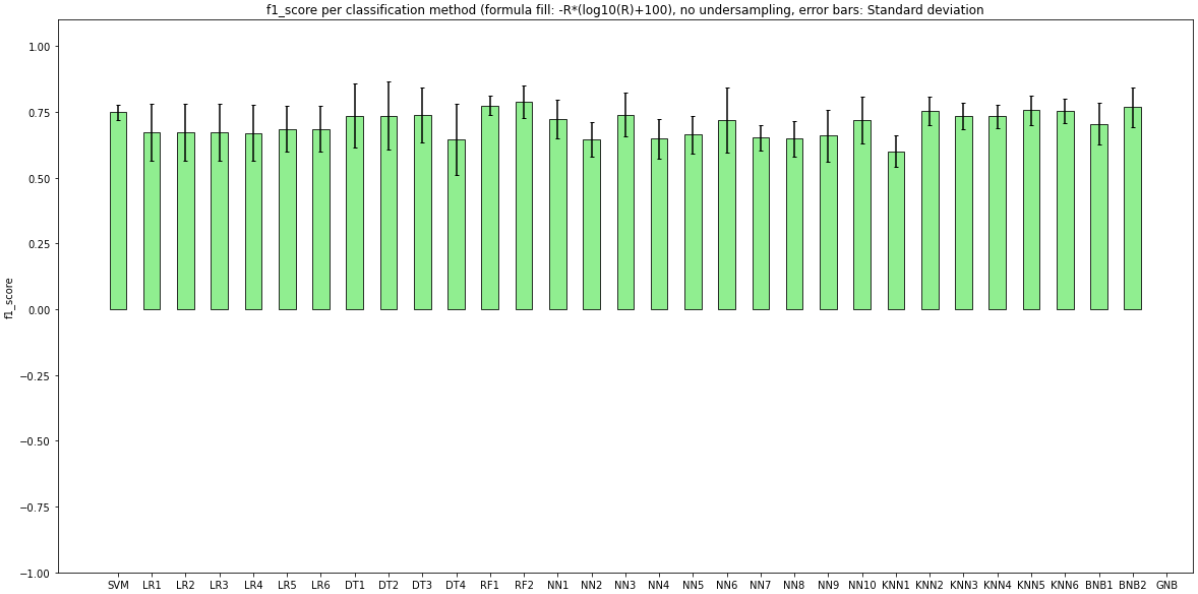

balanced_accuracy per classification method (Iterative imputation, SMOTE-oversampling, error bars: Standard deviation
balanced_accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

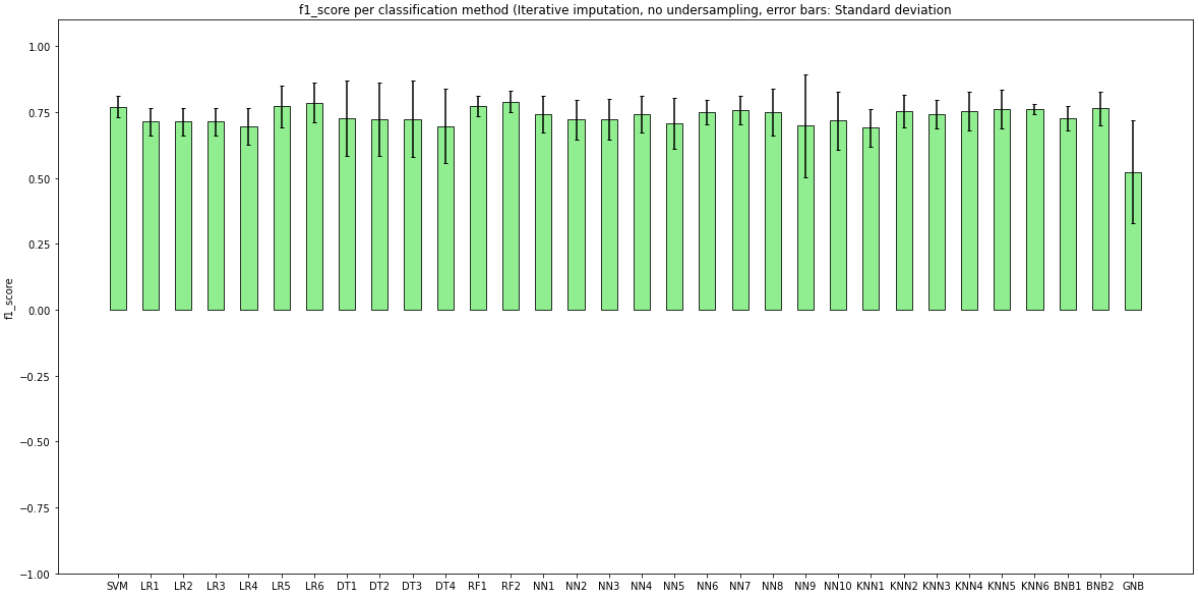

balanced_accuracy per classification method (arbitrary fill eg. -99999, random-oversampling, error bars: Standard deviation
balanced_accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

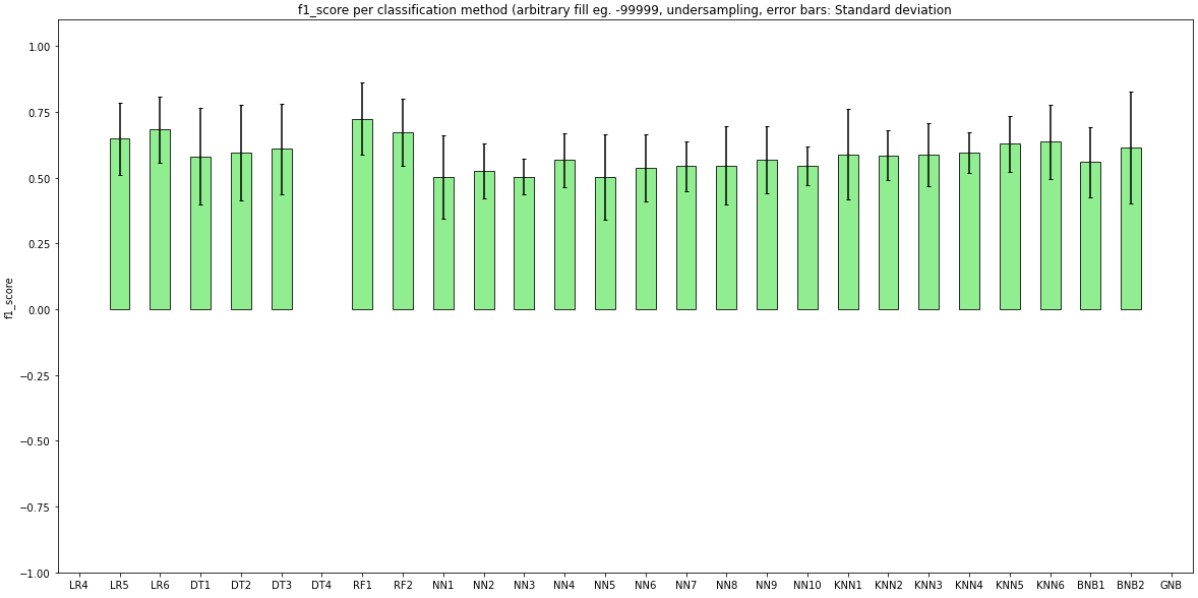

balanced_accuracy per classification method (formula fill: -R*(log10(R)+100), random-oversampling, error bars: Standard deviation
balanced_accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

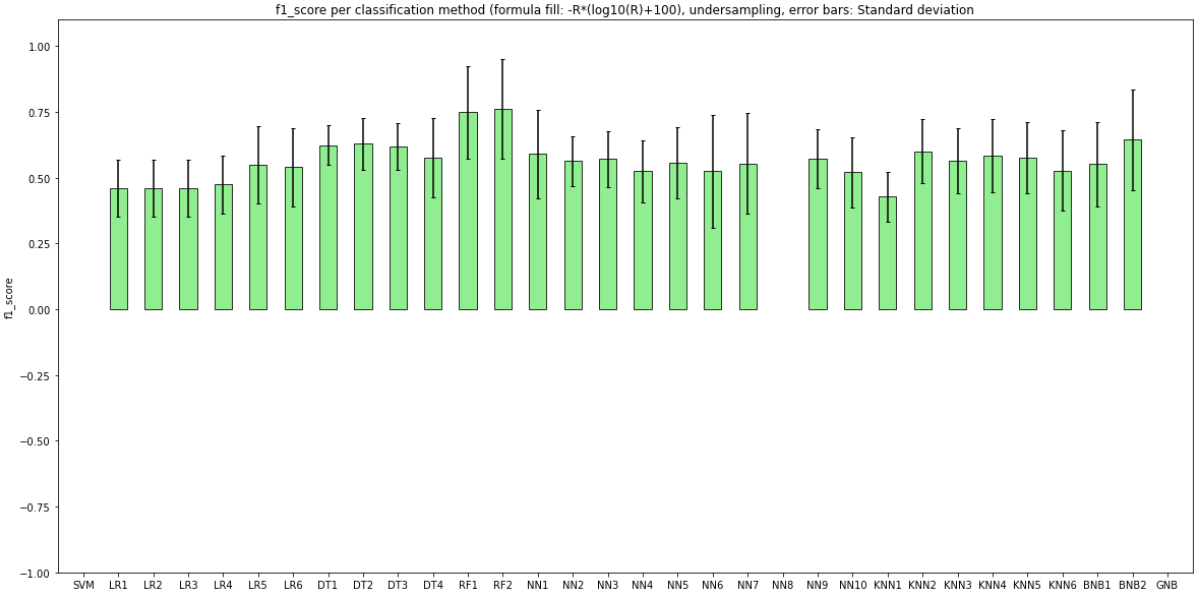

balanced_accuracy per classification method (Iterative imputation, random-oversampling, error bars: Standard deviation
balanced_accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

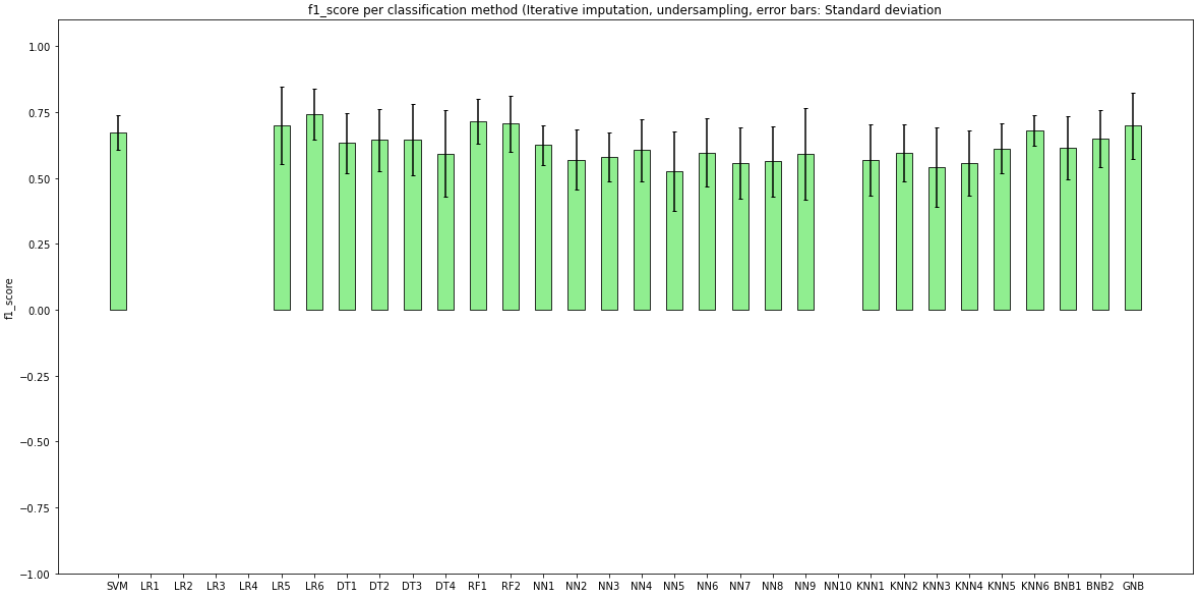

f1_score per classification method (arbitrary fill eg. -99999, no undersampling, error bars: Standard deviation
f1_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

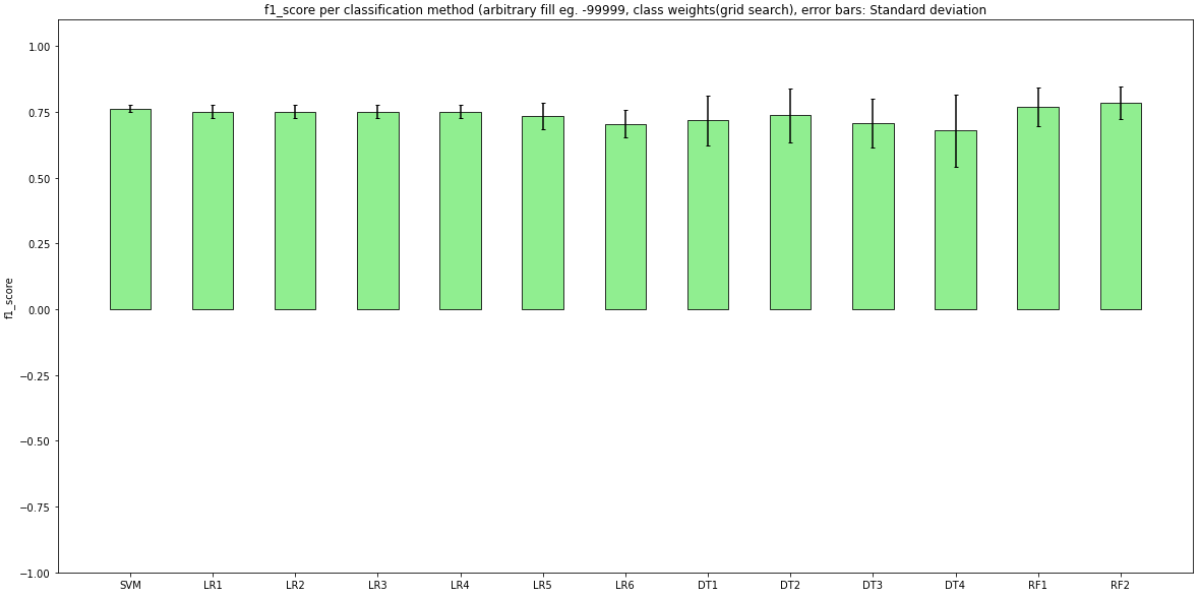

f1_score per classification method (formula fill: -R*(log10(R)+100), no undersampling, error bars: Standard deviation
f1_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

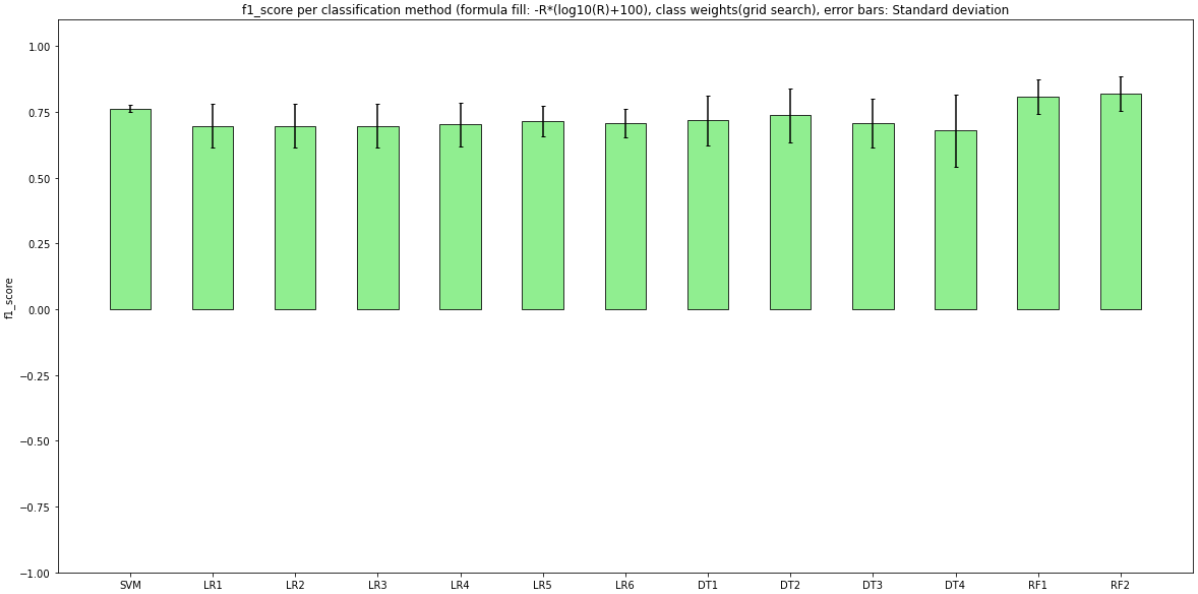

f1_score per classification method (Iterative imputation, no undersampling, error bars: Standard deviation
f1_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

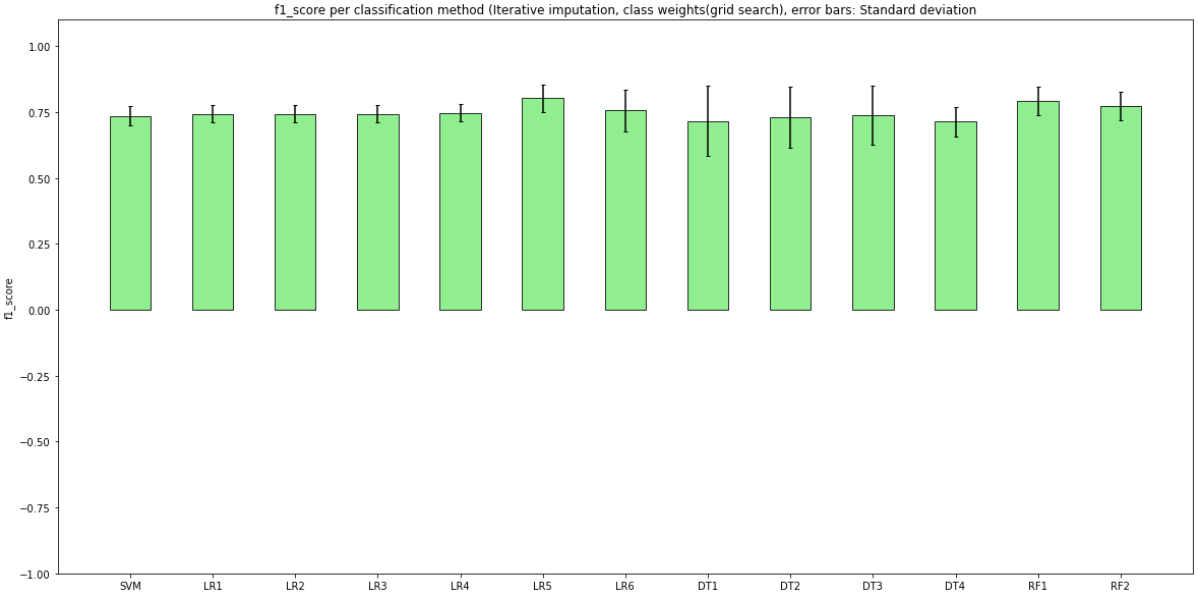

f1_score per classification method (arbitrary fill eg. -99999, undersampling, error bars: Standard deviation
f1_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

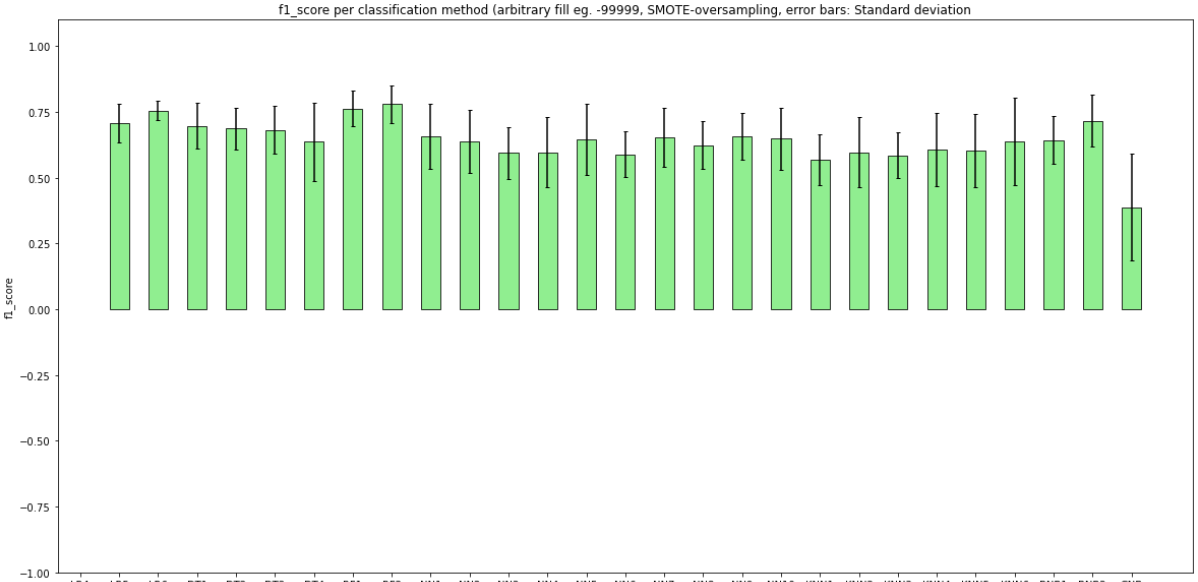

f1_score per classification method (formula fill: -R*(log10(R)+100), undersampling, error bars: Standard deviation
f1_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

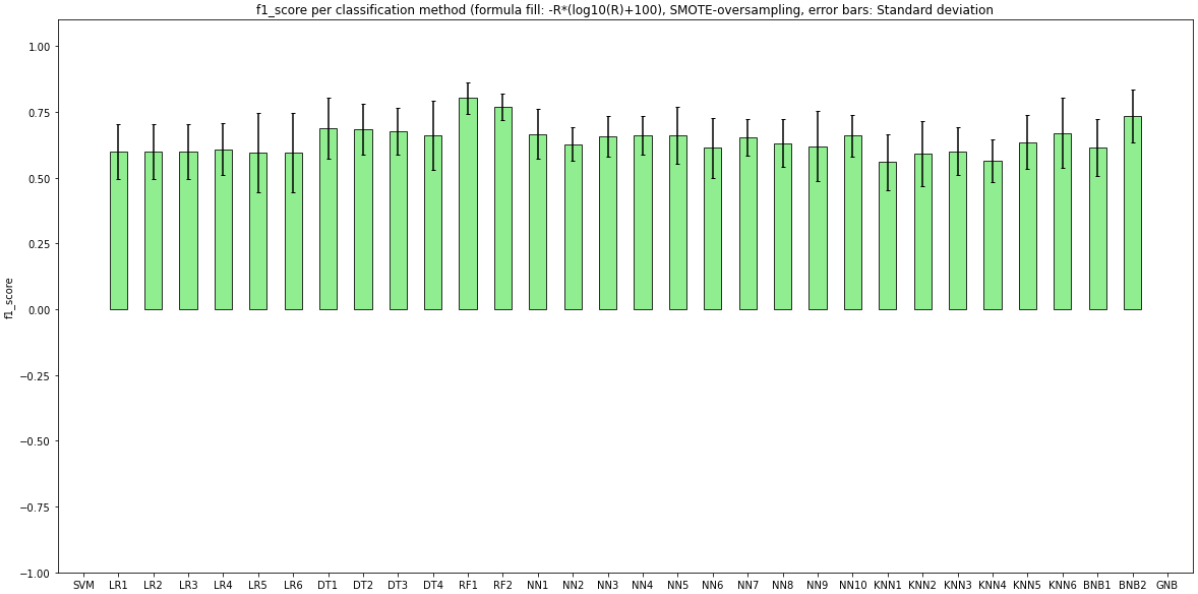

f1_score per classification method (Iterative imputation, undersampling, error bars: Standard deviation
f1_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

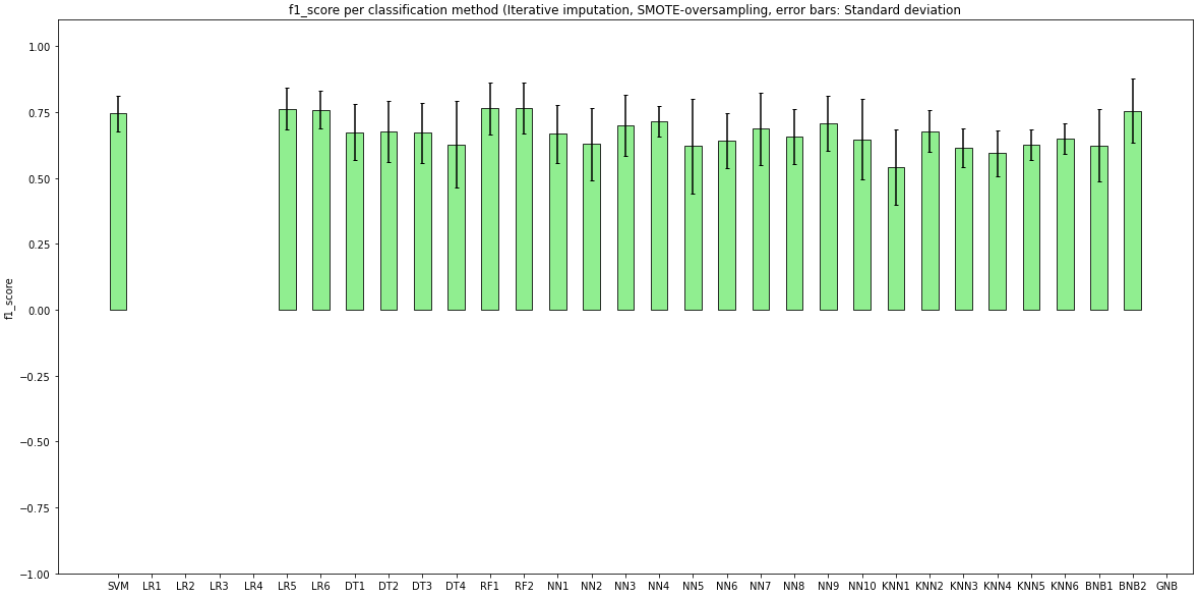

f1_score per classification method (arbitrary fill eg. -99999, class weights(grid search), error bars: Standard deviation
f1_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM
LR1
LR2
LR3
LR4
LR5
LR6
DT1
DT2
DT3
DT4
RF1
RF2

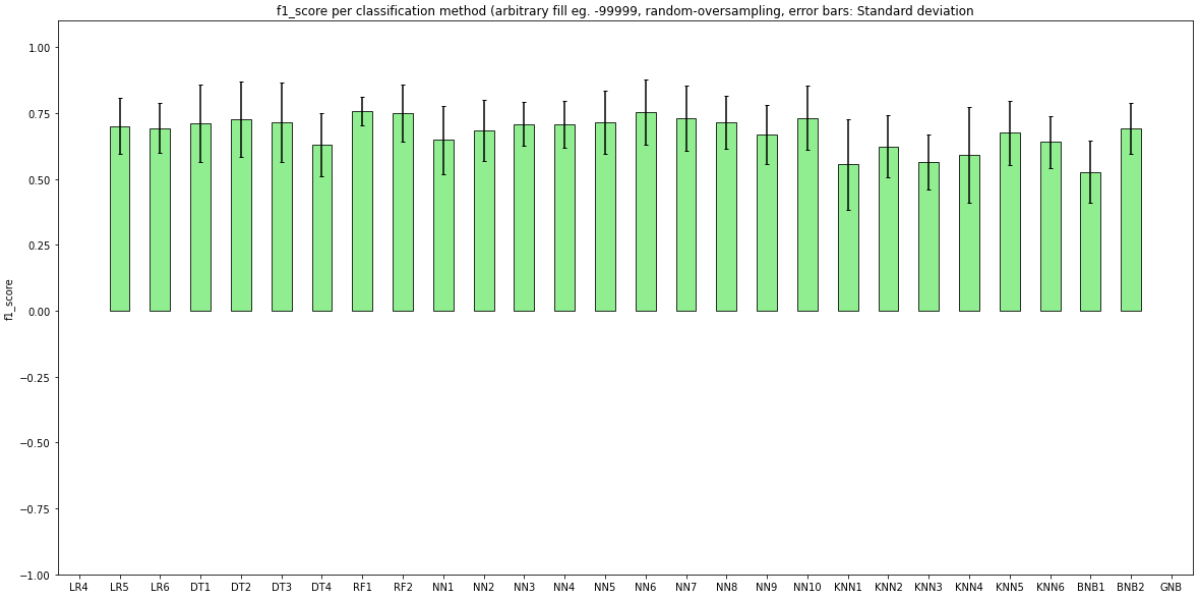

f1_score per classification method (formula fill: -R*(log10(R)+100), class weights(grid search), error bars: Standard deviation
f1_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM
LR1
LR2
LR3
LR4
LR5
LR6
DT1
DT2
DT3
DT4
RF1
RF2

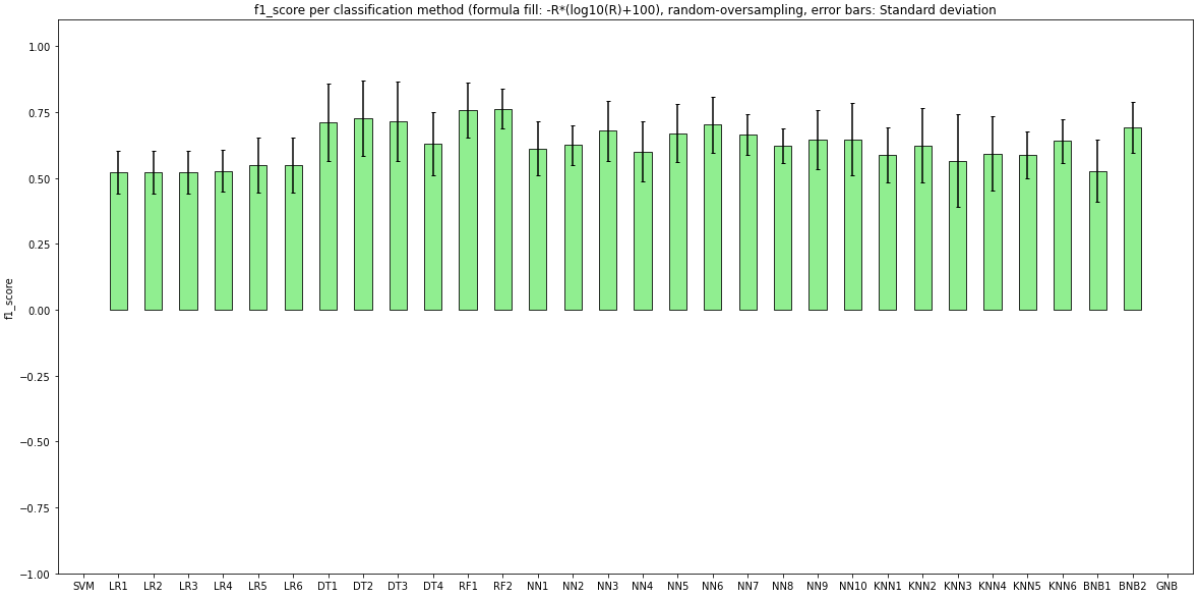

f1_score per classification method (Iterative imputation, class weights(grid search), error bars: Standard deviation
f1_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM
LR1
LR2
LR3
LR4
LR5
LR6
DT1
DT2
DT3
DT4
RF1
RF2

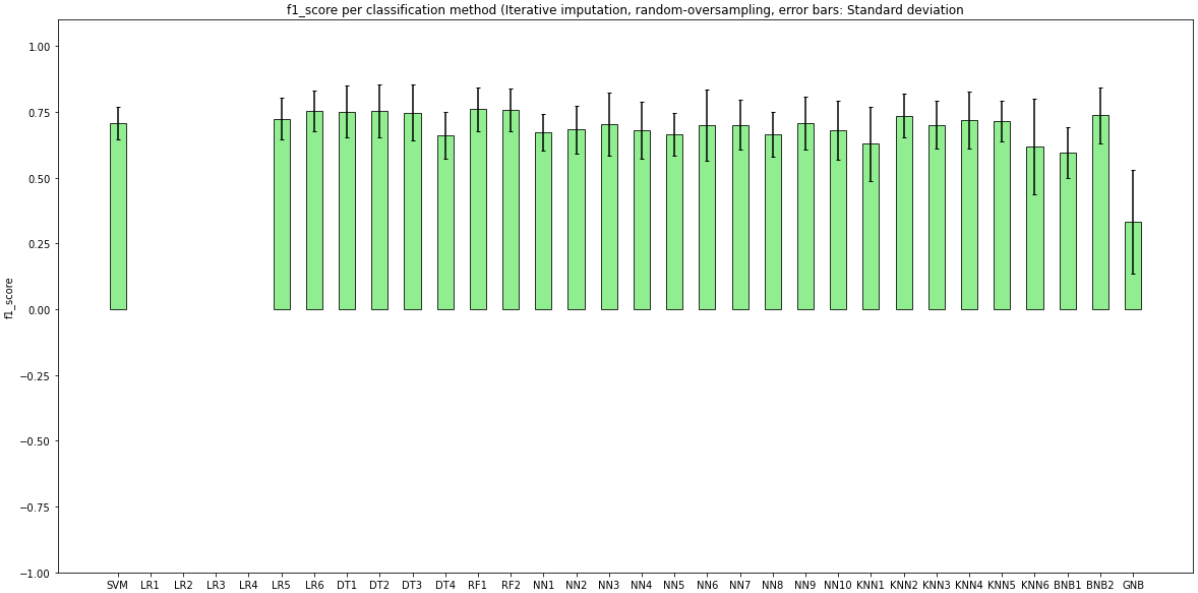

f1_score per classification method (arbitrary fill eg. -99999, SMOTE-oversampling, error bars: Standard deviation
f1_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

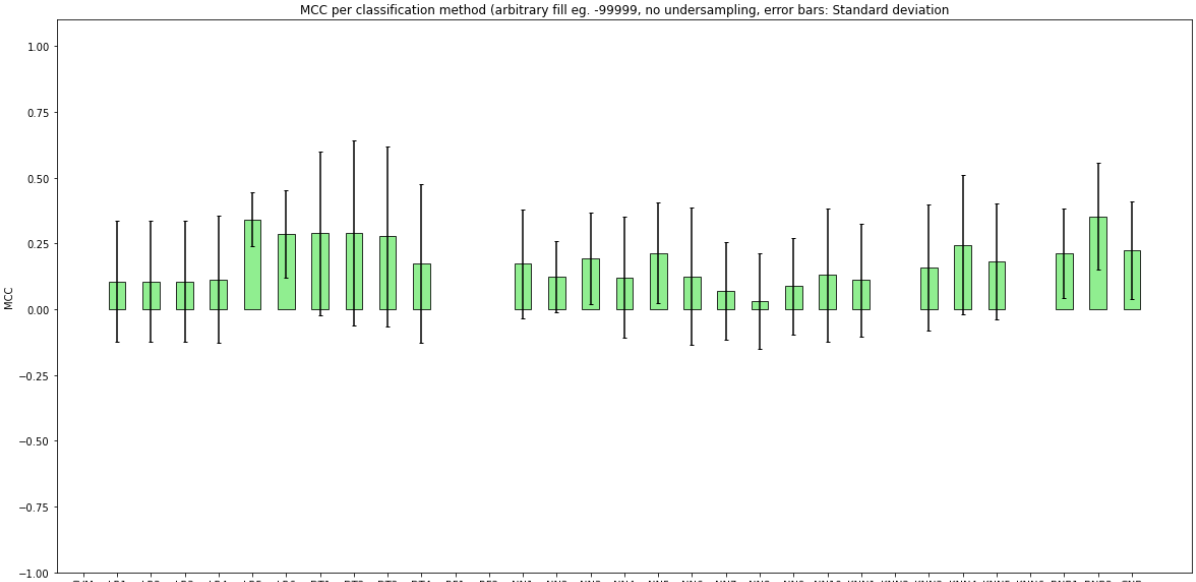

f1_score per classification method (formula fill: -R*(log10(R)+100), SMOTE-oversampling, error bars: Standard deviation
f1_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

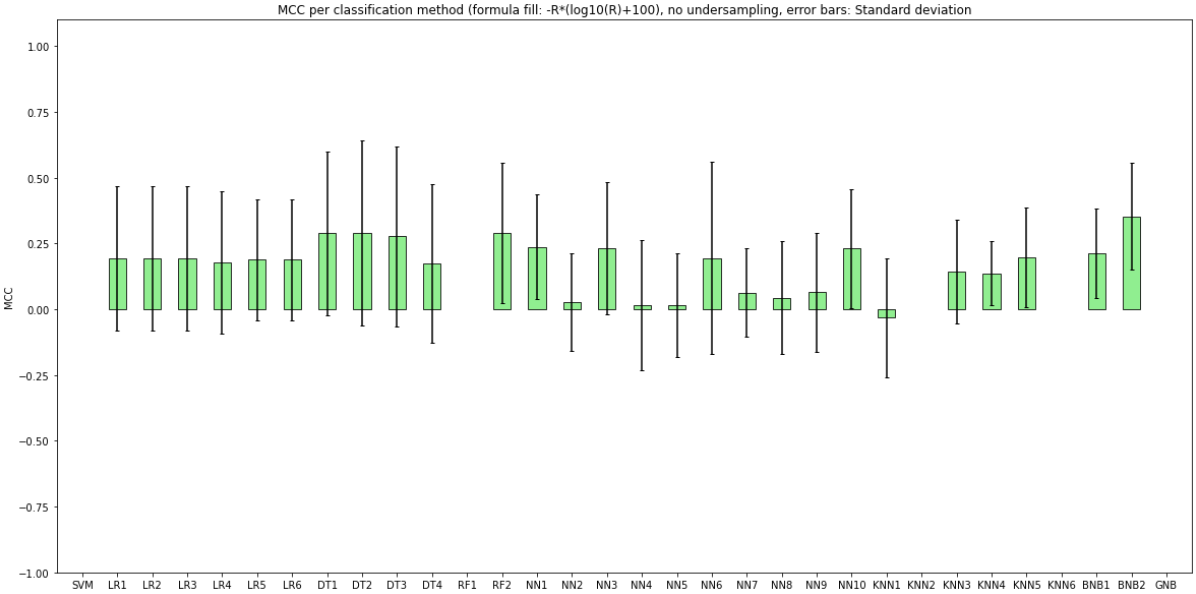

f1_score per classification method (Iterative imputation, SMOTE-oversampling, error bars: Standard deviation
f1_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

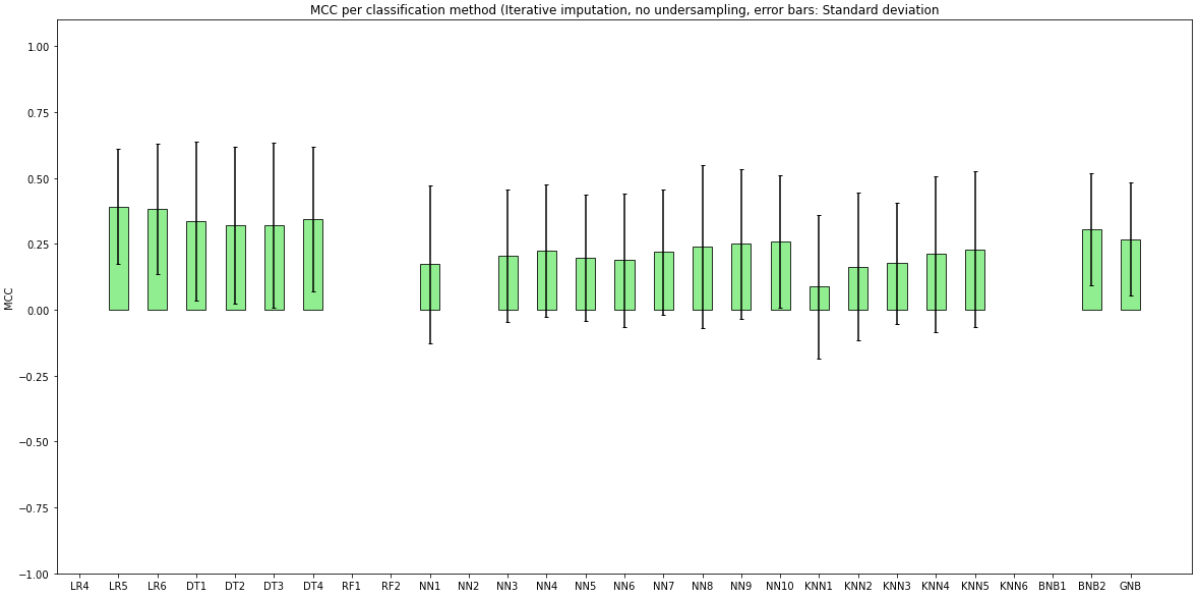

f1_score per classification method (arbitrary fill eg. -99999, random-oversampling, error bars: Standard deviation
f1_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

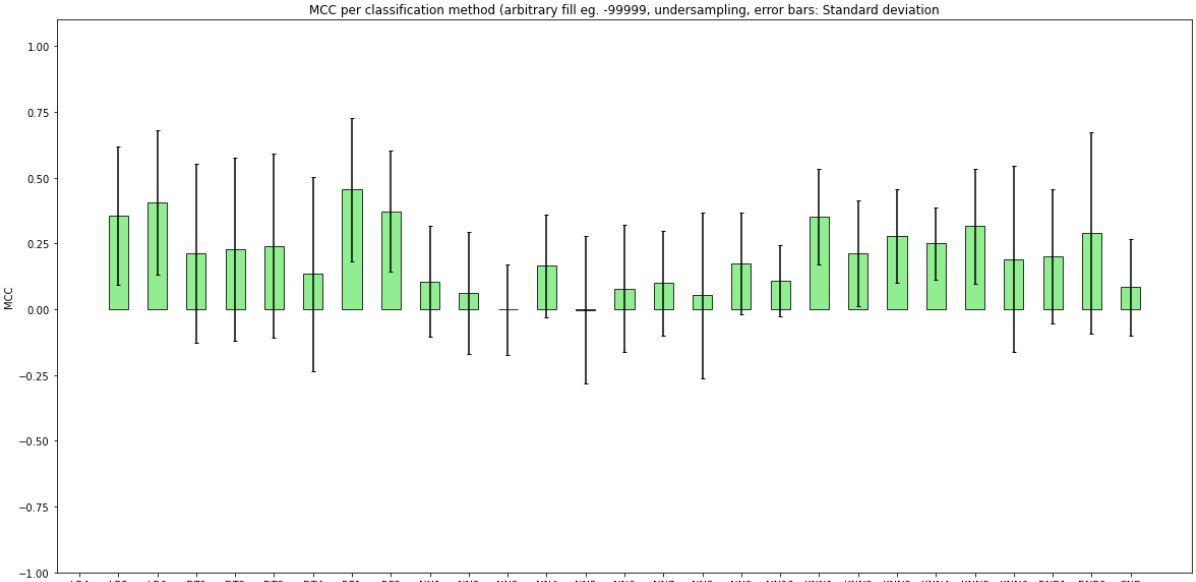

f1_score per classification method (formula fill: -R*(log10(R)+100), random-oversampling, error bars: Standard deviation
f1_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

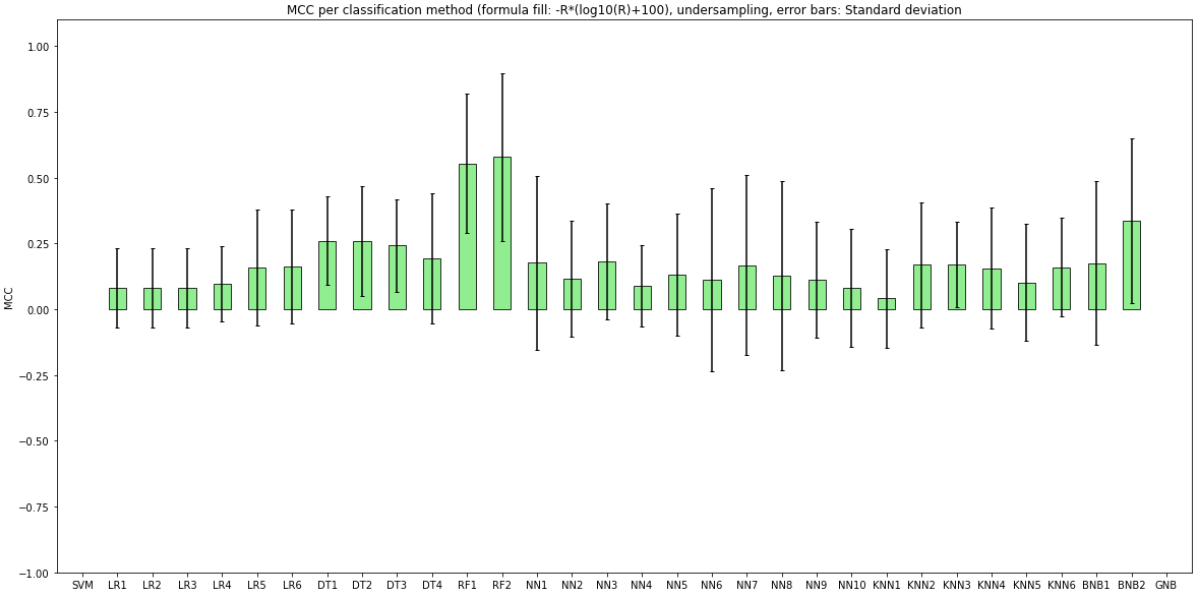

f1_score per classification method (Iterative imputation, random-oversampling, error bars: Standard deviation
f1_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

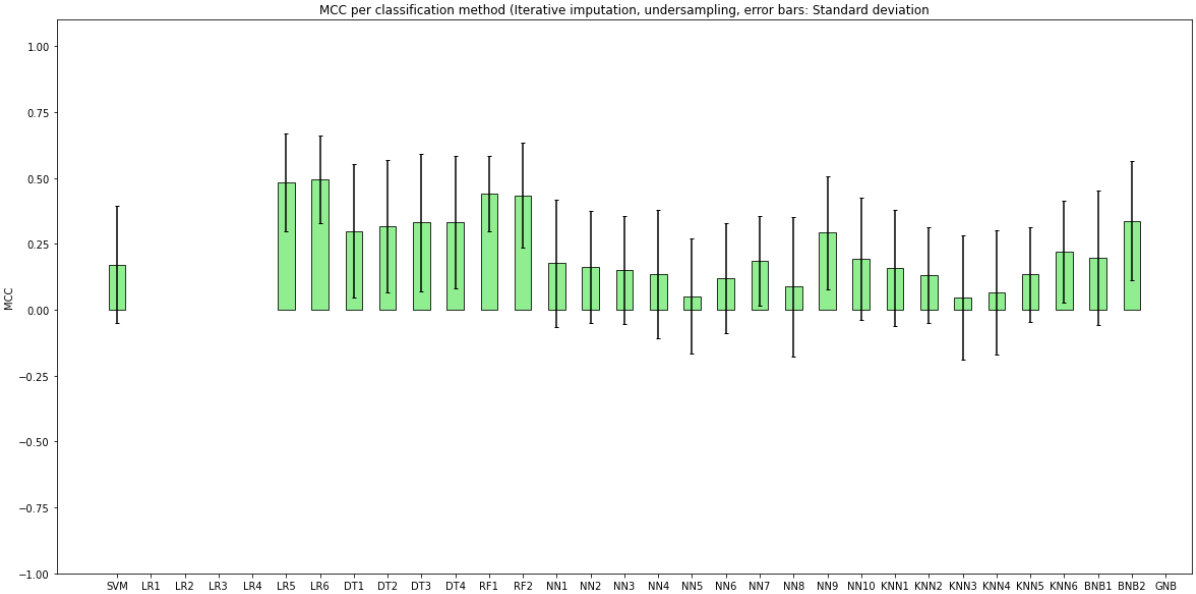

MCC per classification method (arbitrary fill eg. -99999, no undersampling, error bars: Standard deviation
MCC
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

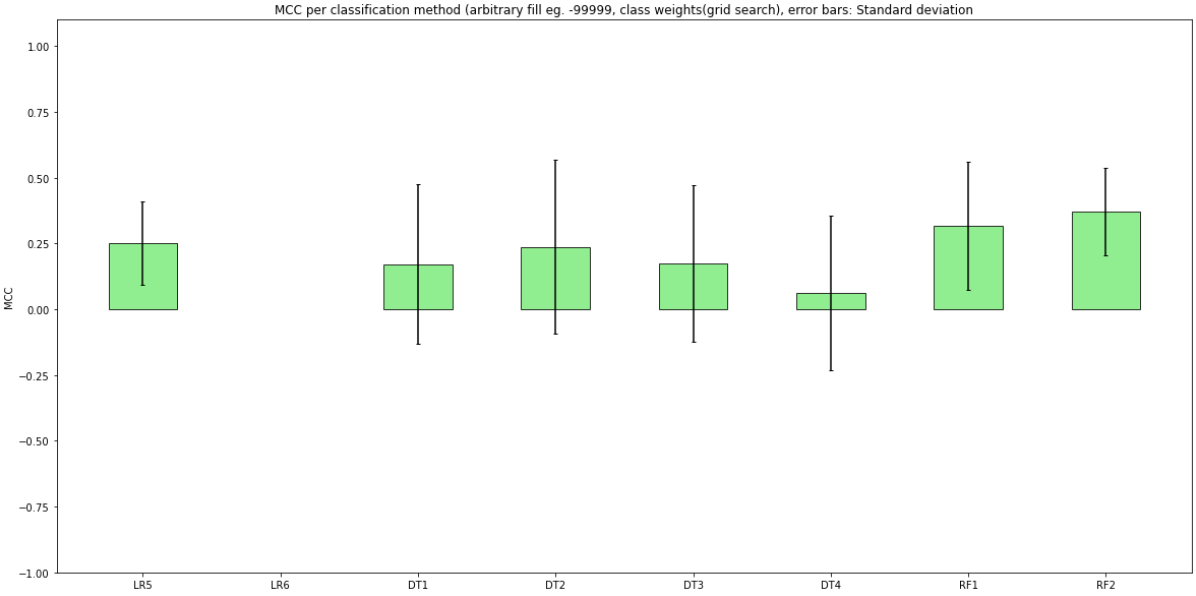

MCC per classification method (formula fill: -R*(log10(R)+100), no undersampling, error bars: Standard deviation
MCC
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

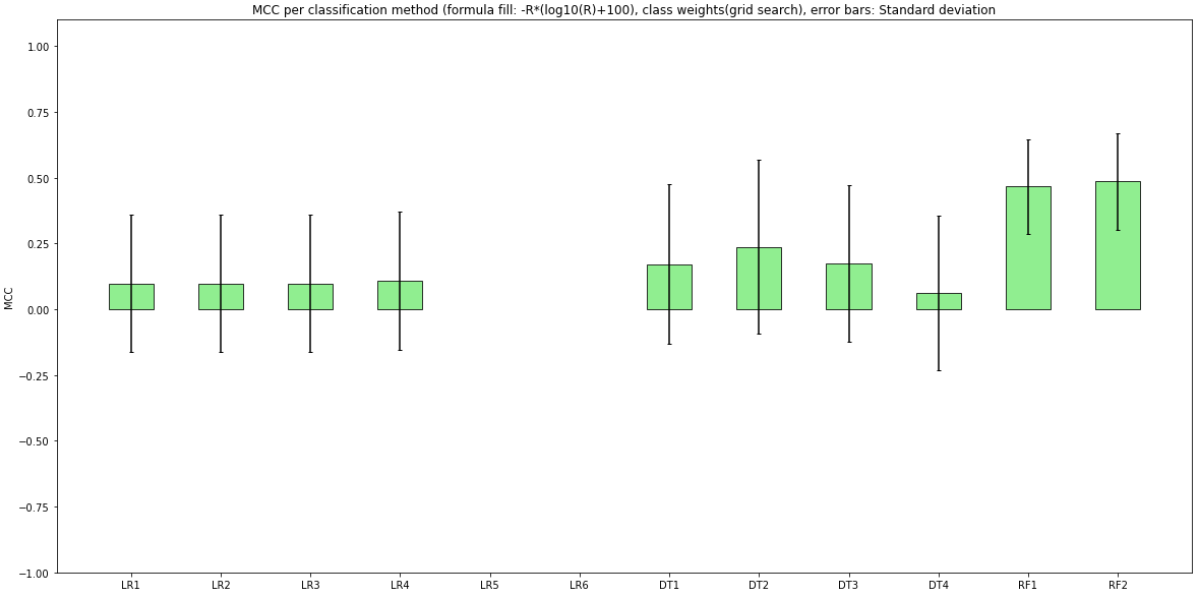

MCC per classification method (Iterative imputation, no undersampling, error bars: Standard deviation
MCC
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

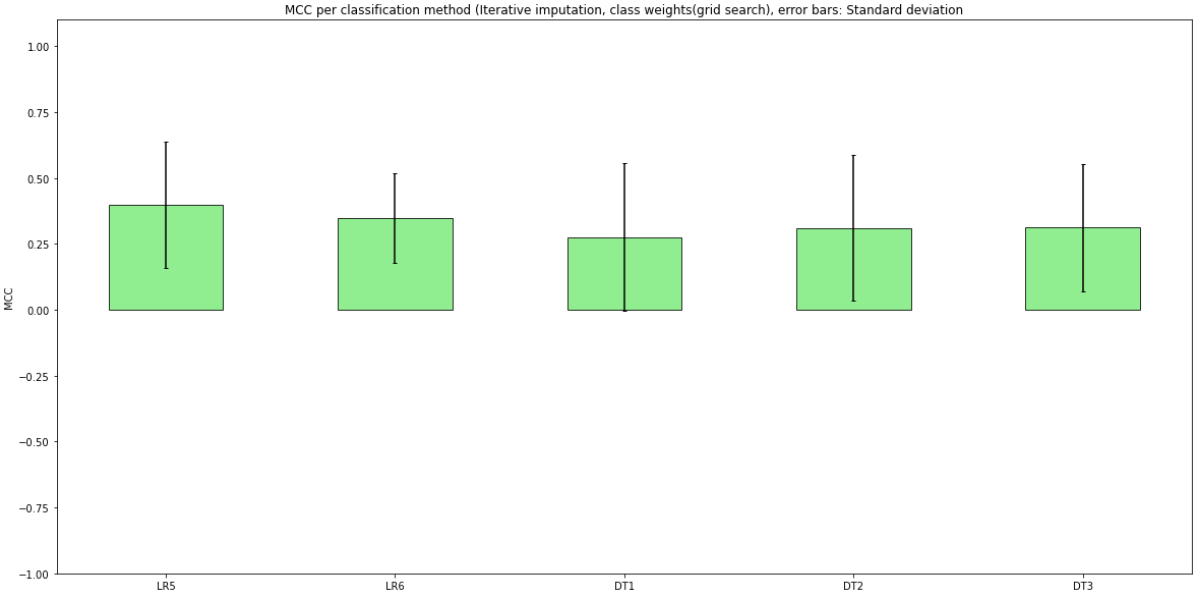

MCC per classification method (arbitrary fill eg. -99999, undersampling, error bars: Standard deviation
MCC
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

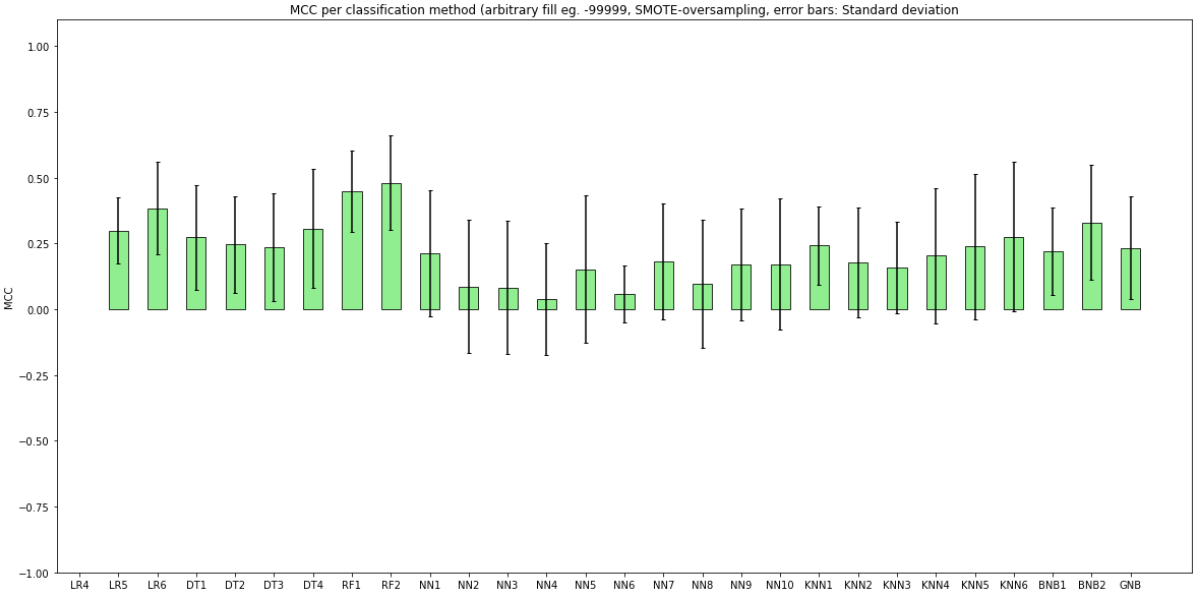

MCC per classification method (formula fill: -R*(log10(R)+100), undersampling, error bars: Standard deviation
MCC
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

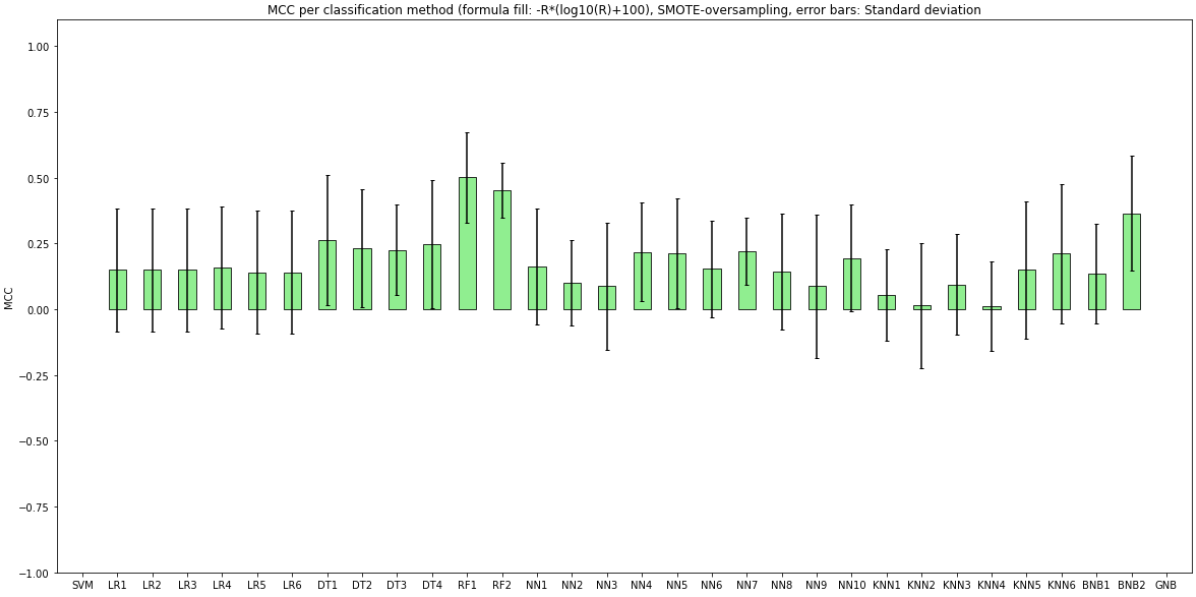

MCC per classification method (Iterative imputation, undersampling, error bars: Standard deviation
MCC
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

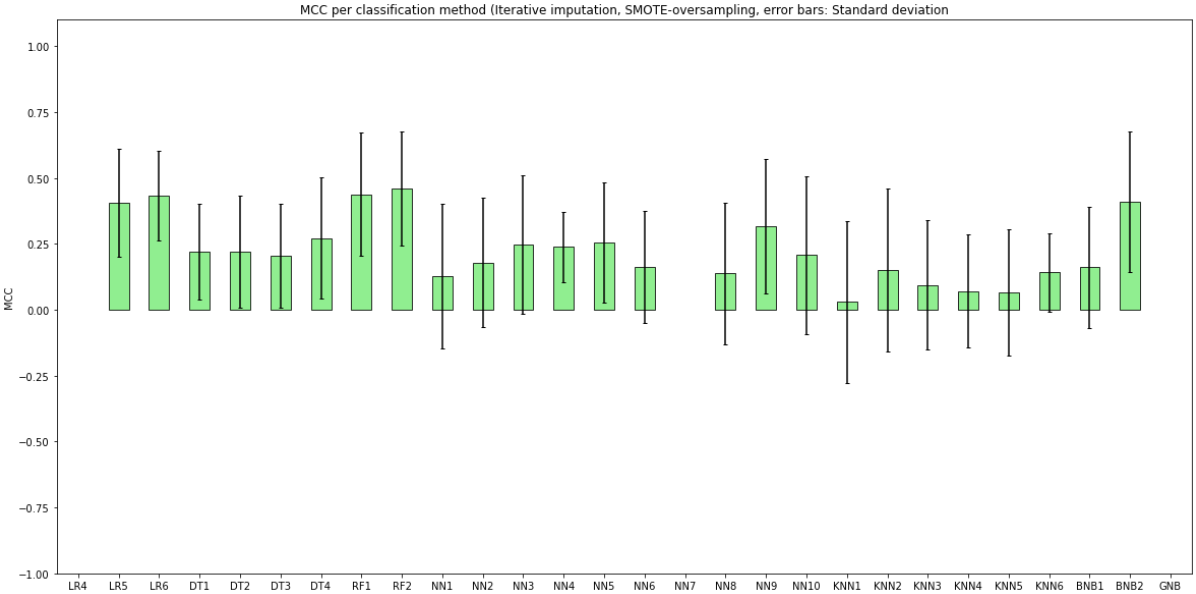

MCC per classification method (arbitrary fill eg. -99999, class weights(grid search), error bars: Standard deviation
MCC
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
LR5
LR6
DT1
DT2
DT3
DT4
RF1
RF2

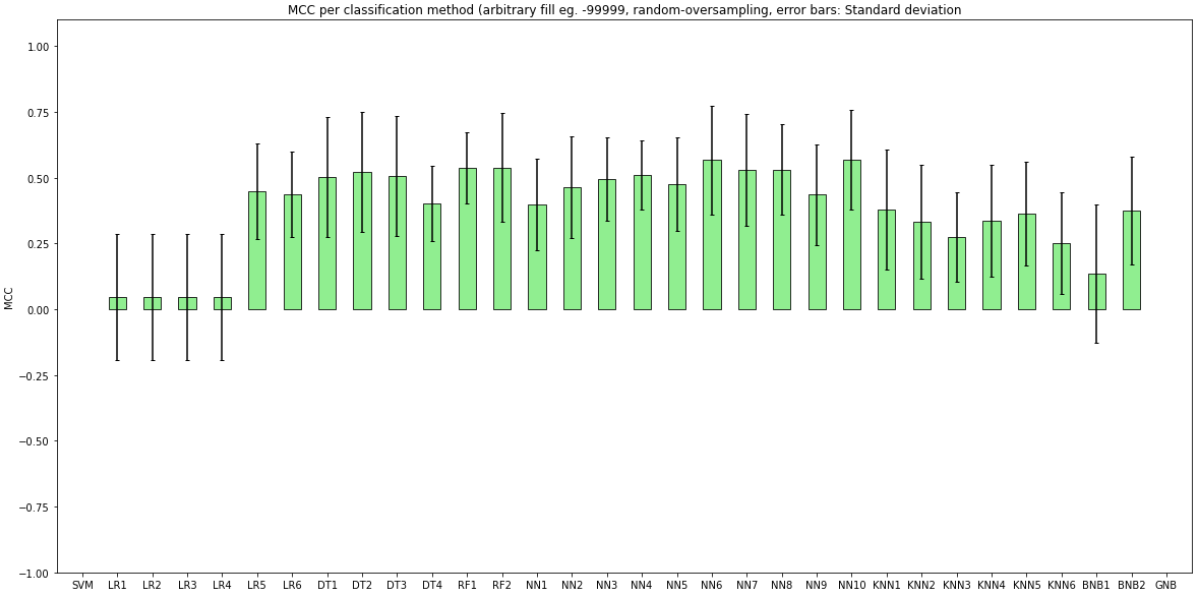

MCC per classification method (formula fill: -R*(log10(R)+100), class weights(grid search), error bars: Standard deviation
MCC
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
LR1
LR2
LR3
LR4
LR5
LR6
DT1
DT2
DT3
DT4
RF1
RF2

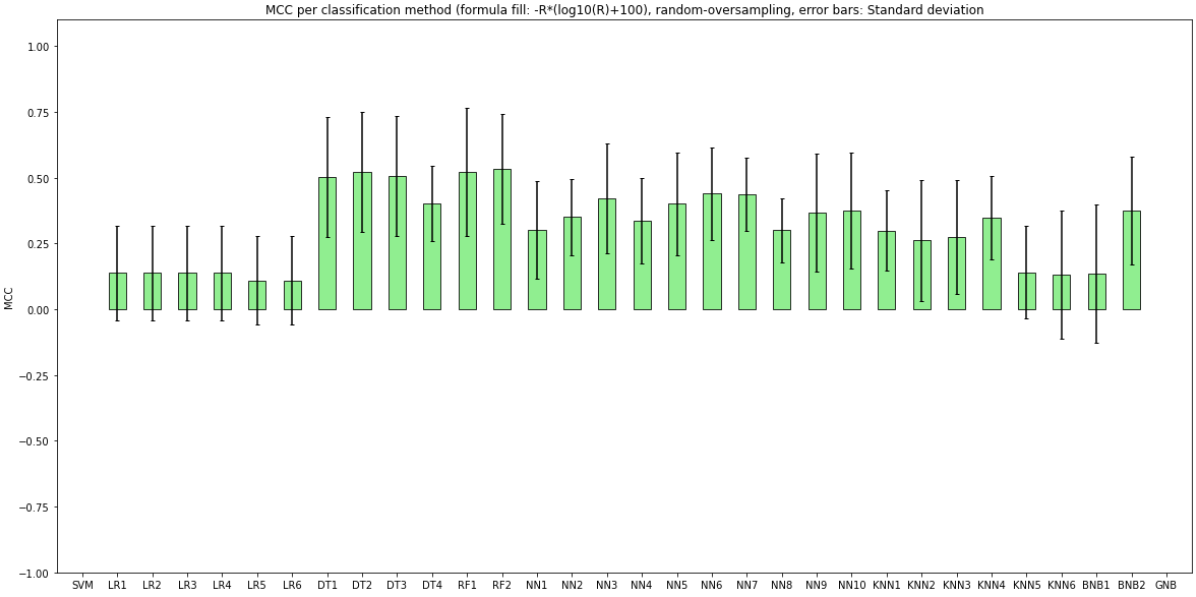

MCC per classification method (Iterative imputation, class weights(grid search), error bars: Standard deviation
MCC
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
LR5
LR6
DT1
DT2
DT3

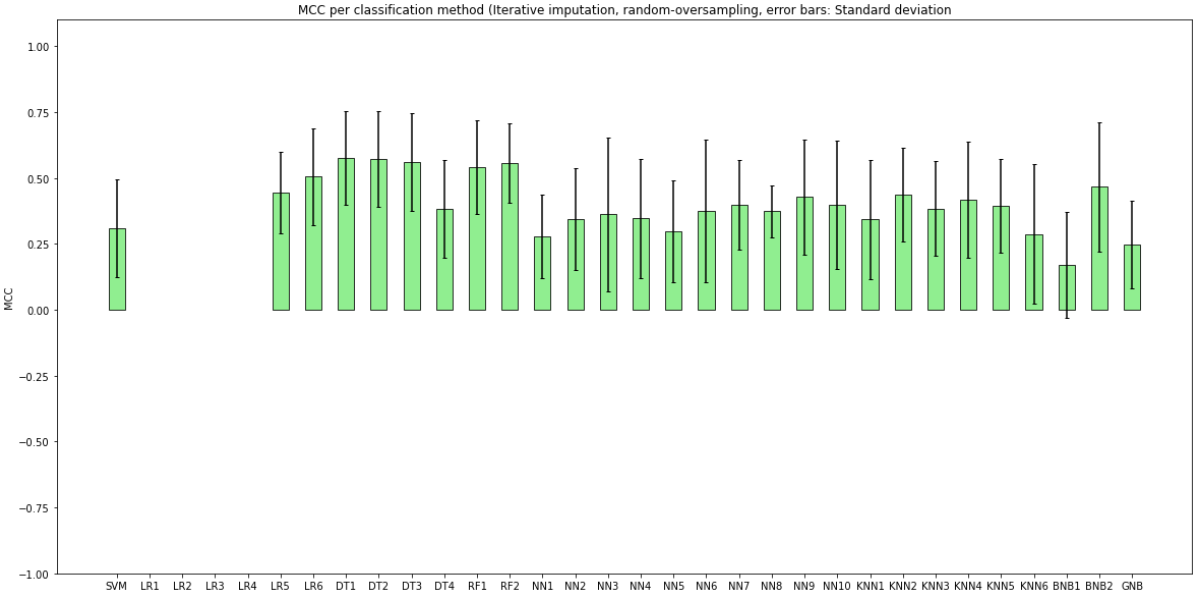

MCC per classification method (arbitrary fill eg. -99999, SMOTE-oversampling, error bars: Standard deviation
MCC
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

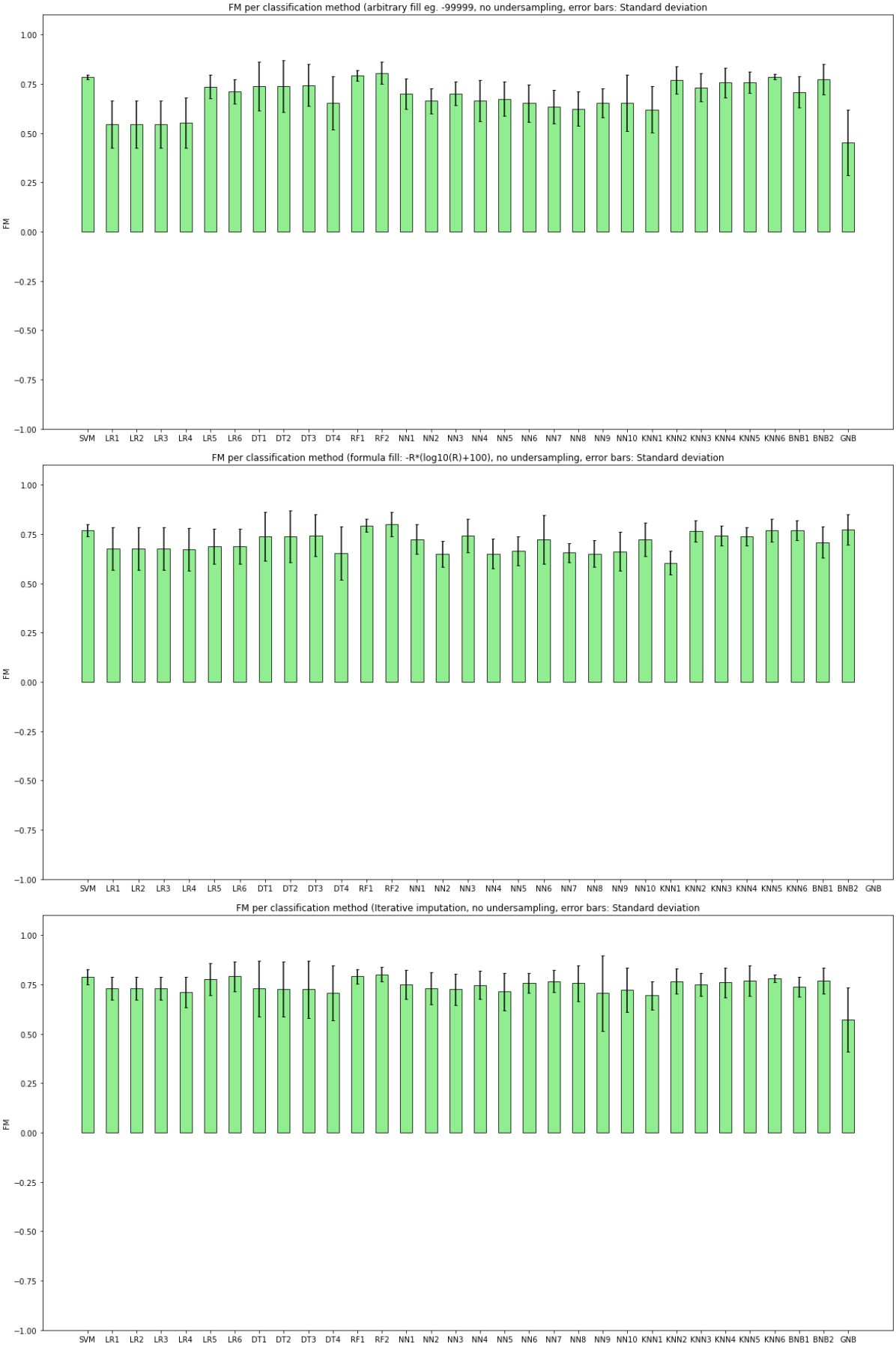

MCC per classification method (formula fill: -R*(log10(R)+100), SMOTE-oversampling, error bars: Standard deviation
MCC
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

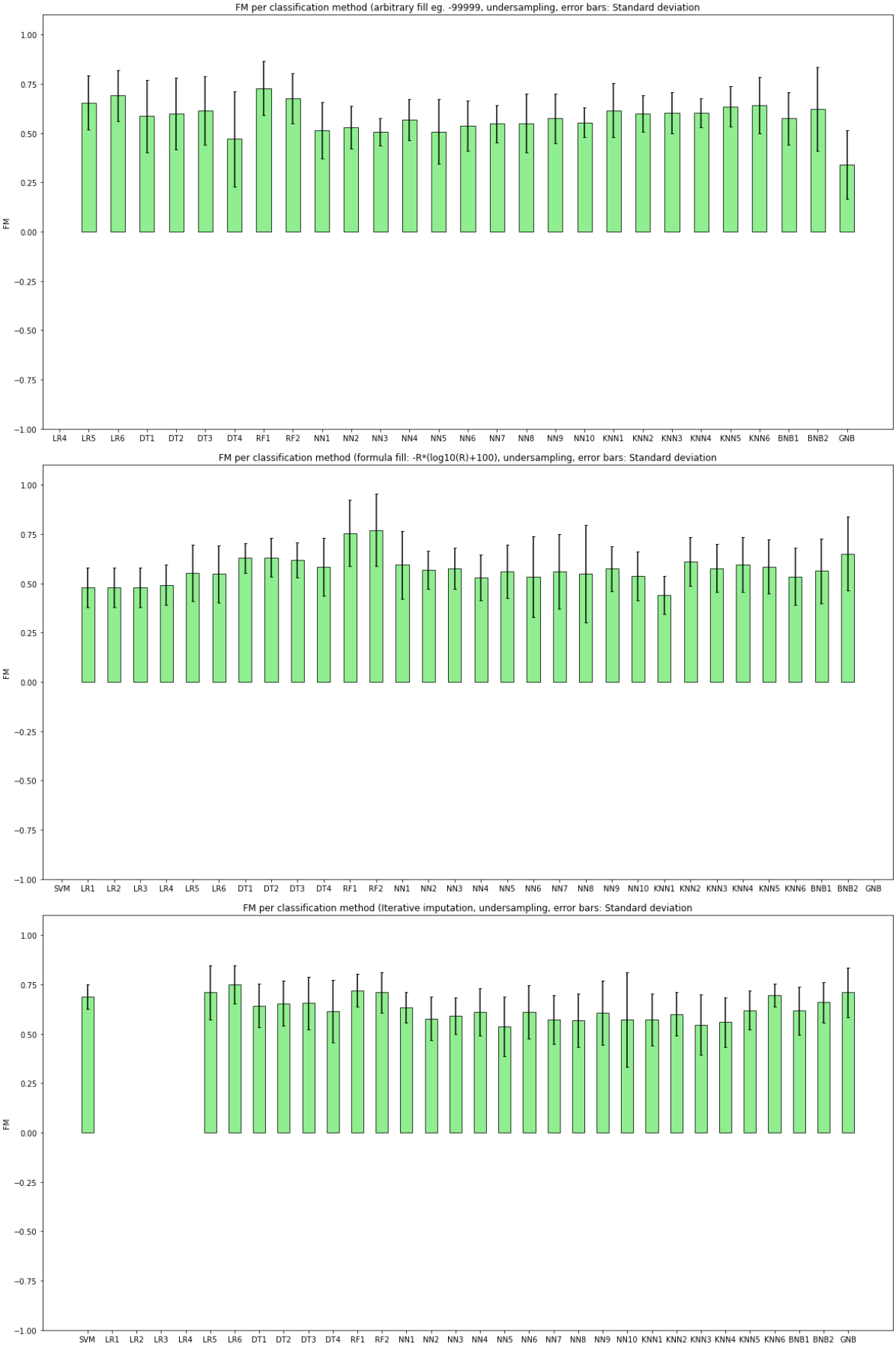

MCC per classification method (Iterative imputation, SMOTE-oversampling, error bars: Standard deviation
MCC
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

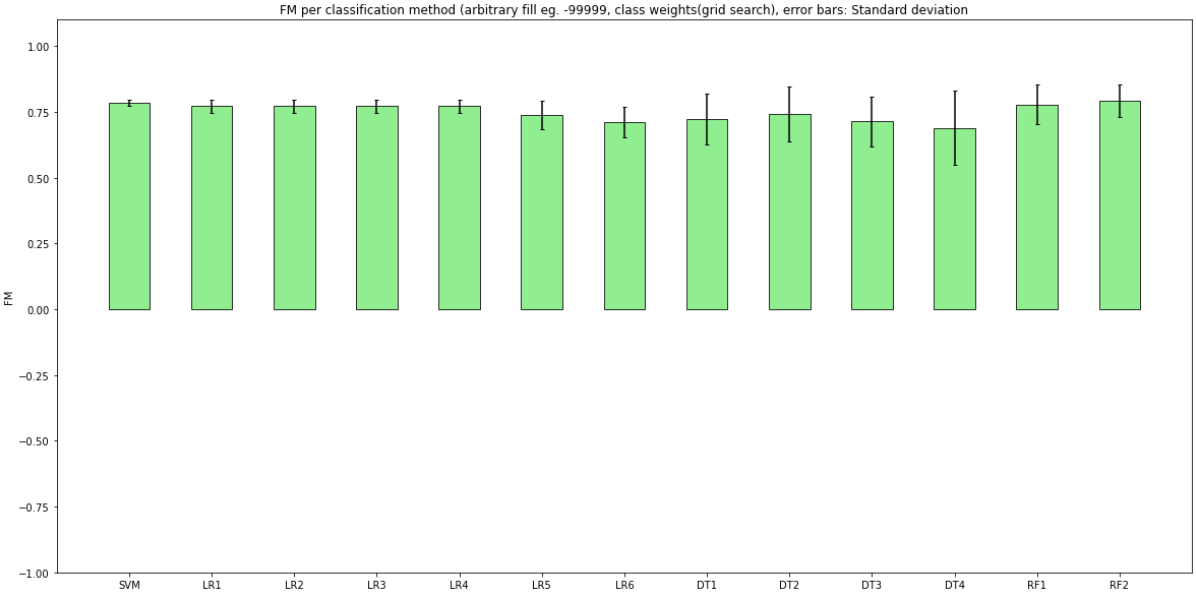

MCC per classification method (arbitrary fill eg. -99999, random-oversampling, error bars: Standard deviation
MCC
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

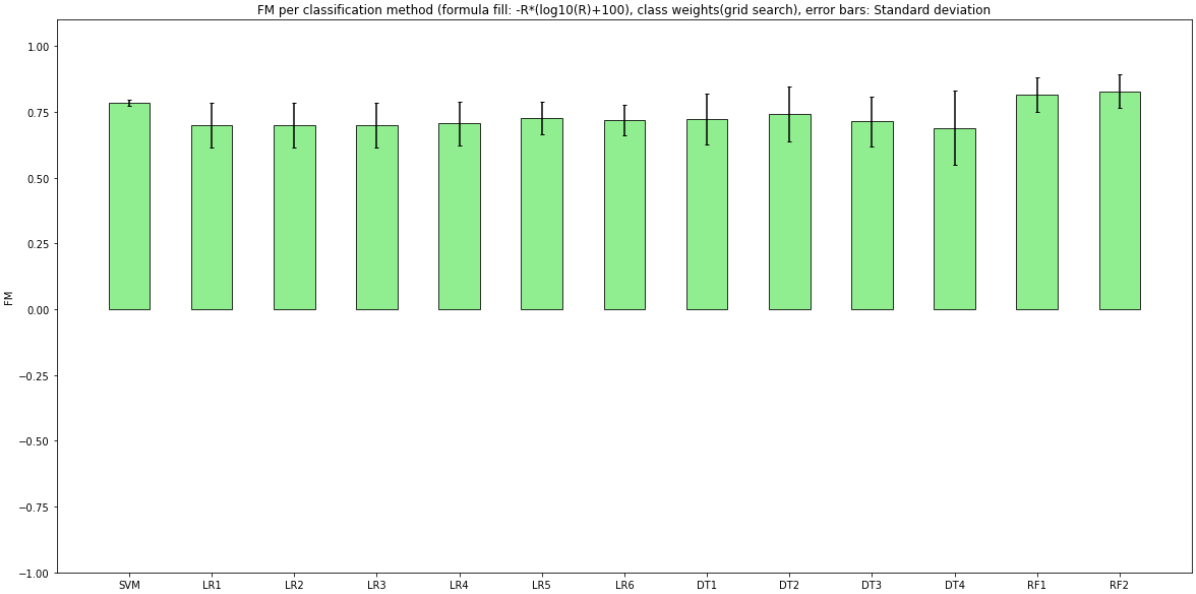

MCC per classification method (formula fill: -R*(log10(R)+100), random-oversampling, error bars: Standard deviation
MCC
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

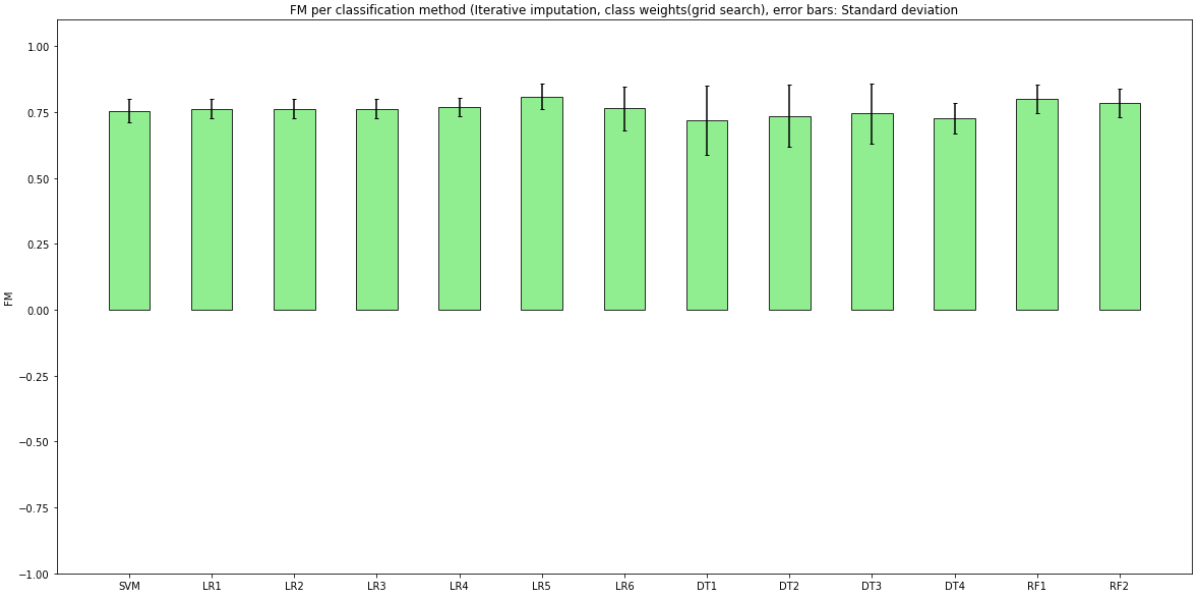

MCC per classification method (Iterative imputation, random-oversampling, error bars: Standard deviation
MCC
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

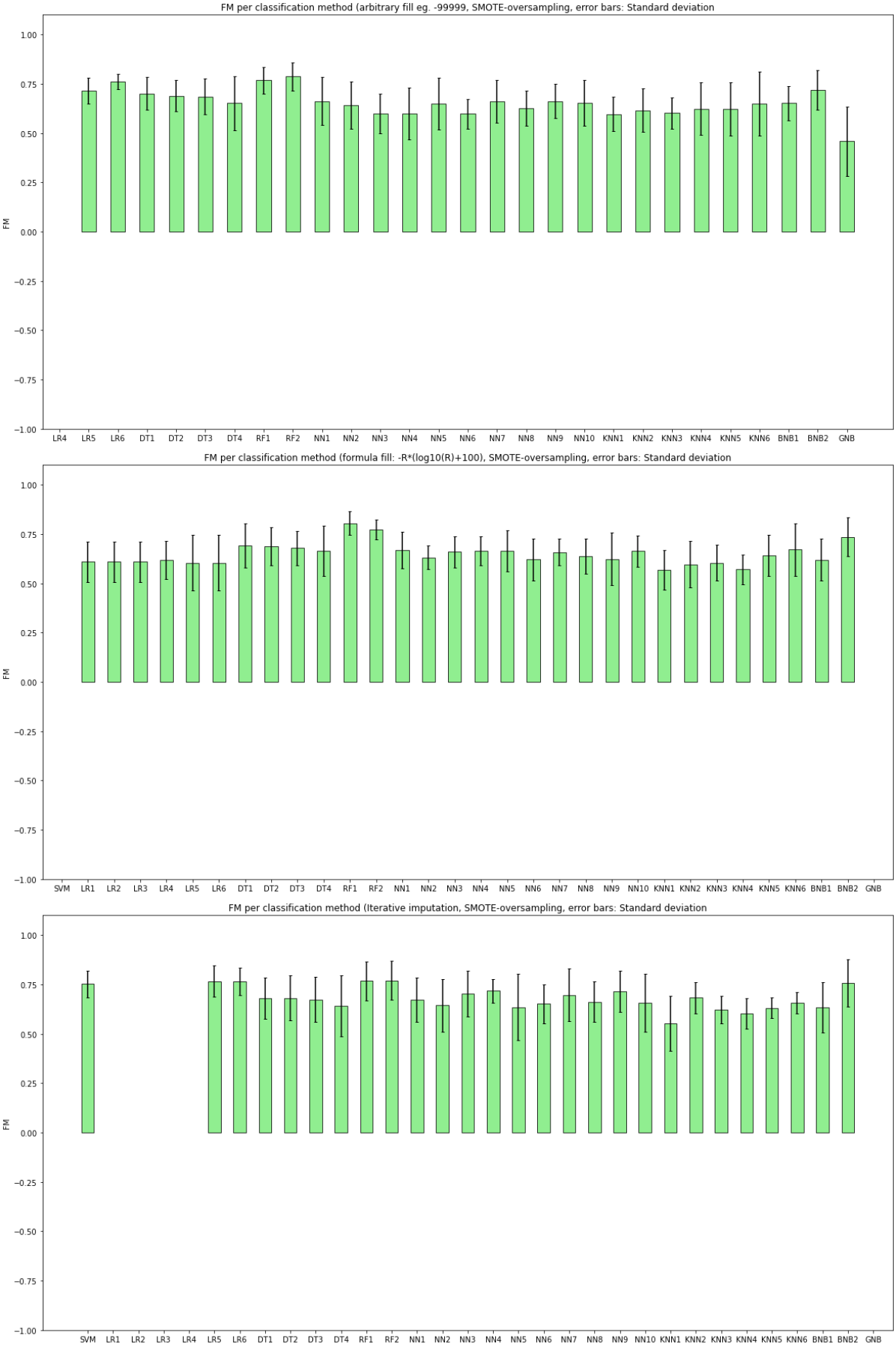

FM per classification method (arbitrary fill eg. -99999, no undersampling, error bars: Standard deviation
FM per classification method (formula fill: -R*(log10(R)+100), no undersampling, error bars: Standard deviation
FM per classification method (Iterative imputation, no undersampling, error bars: Standard deviation
FM
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

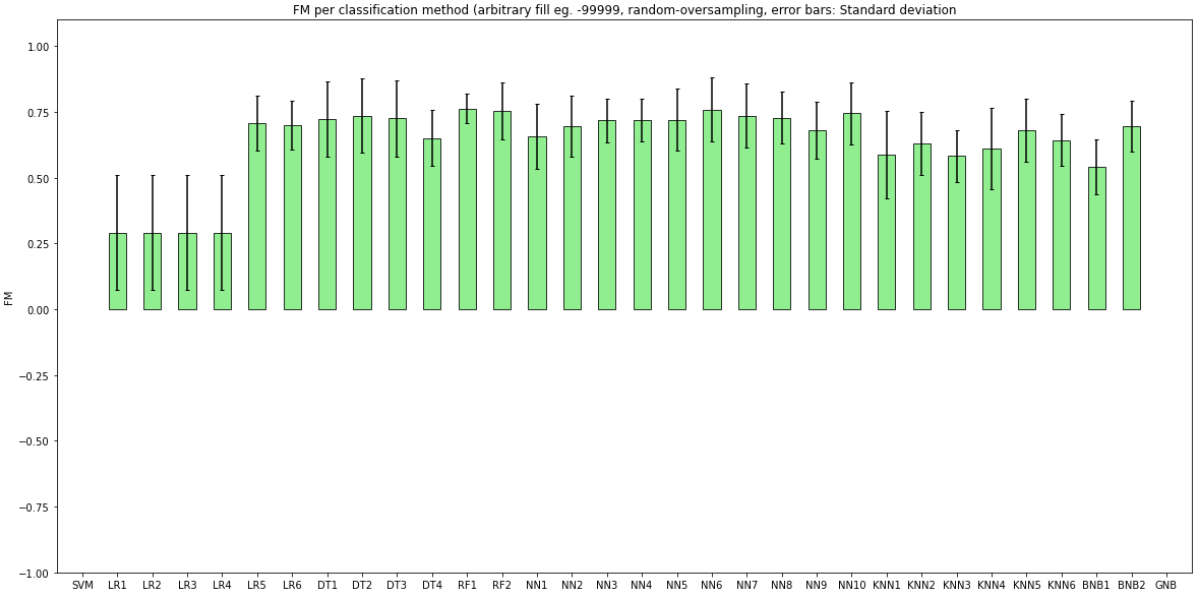

FM per classification method (arbitrary fill eg. -99999, undersampling, error bars: Standard deviation
FM
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB
FM per classification method (formula fill: -R*(log10(R)+100), undersampling, error bars: Standard deviation
FM
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB
FM per classification method (Iterative imputation, undersampling, error bars: Standard deviation
FM
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

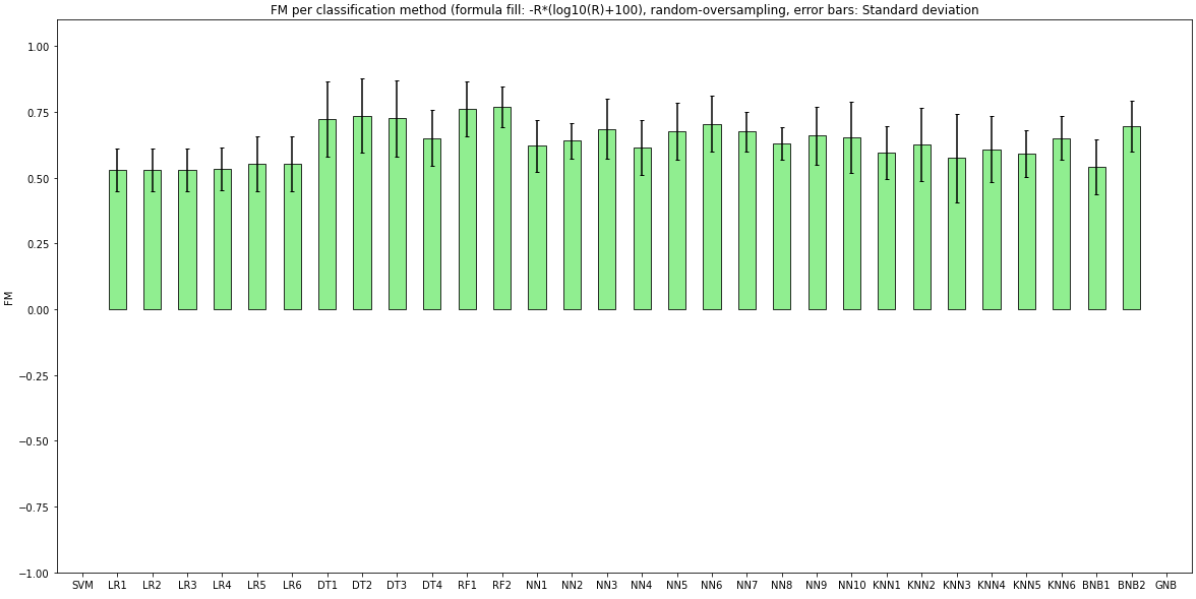

FM per classification method (arbitrary fill eg. -99999, class weights(grid search), error bars: Standard deviation
FM
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM
LR1
LR2
LR3
LR4
LR5
LR6
DT1
DT2
DT3
DT4
RF1
RF2

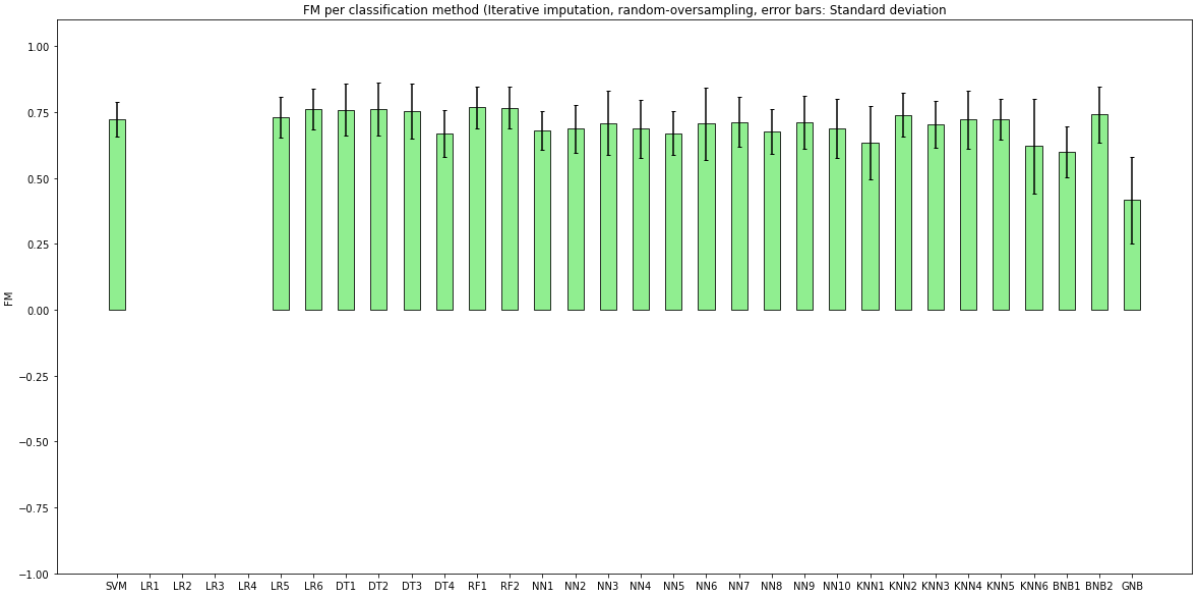

FM per classification method (formula fill: -R*(log10(R)+100), class weights(grid search), error bars: Standard deviation
FM
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM
LR1
LR2
LR3
LR4
LR5
LR6
DT1
DT2
DT3
DT4
RF1
RF2

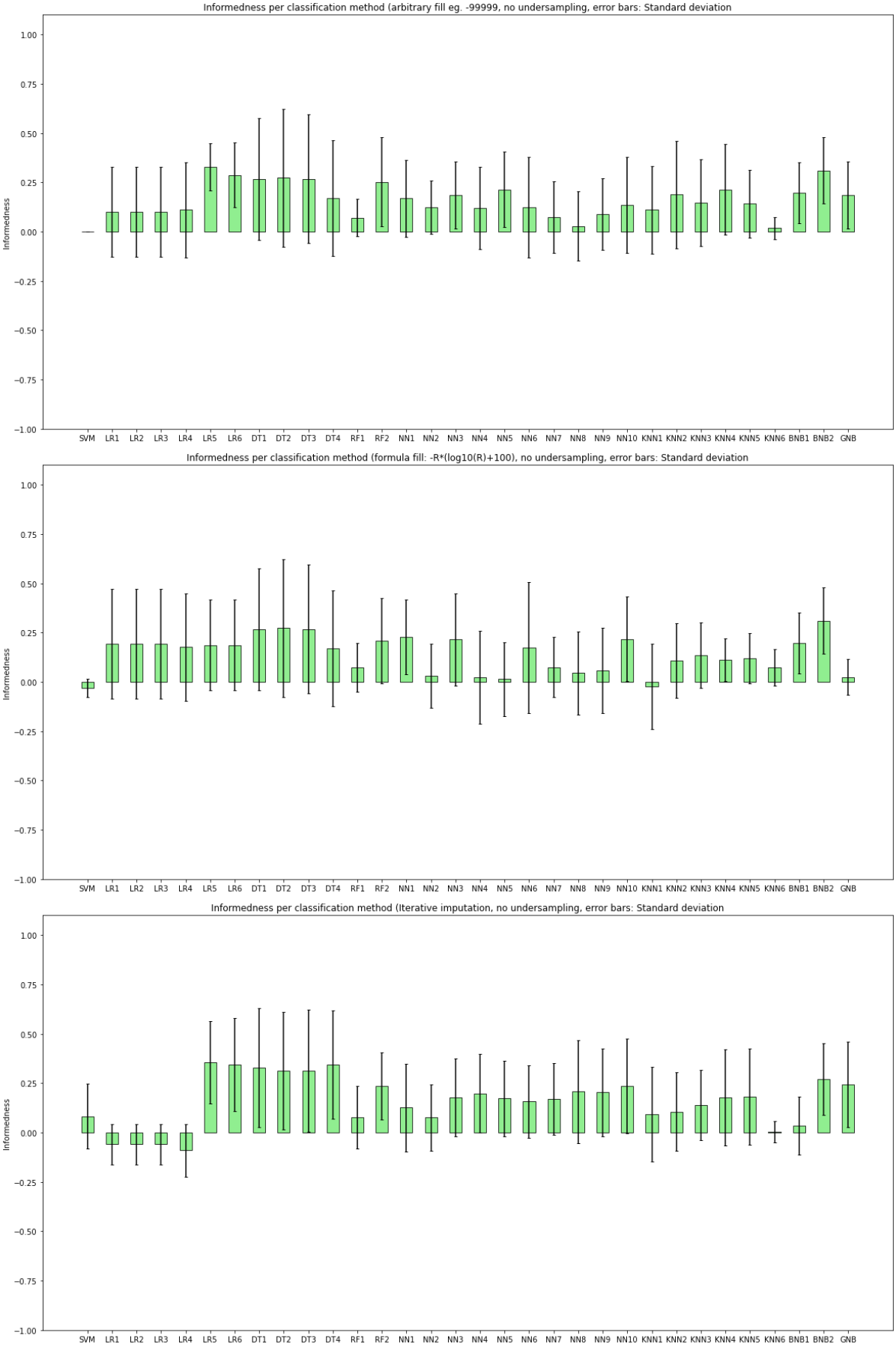

FM per classification method (Iterative imputation, class weights(grid search), error bars: Standard deviation
FM
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM
LR1
LR2
LR3
LR4
LR5
LR6
DT1
DT2
DT3
DT4
RF1
RF2

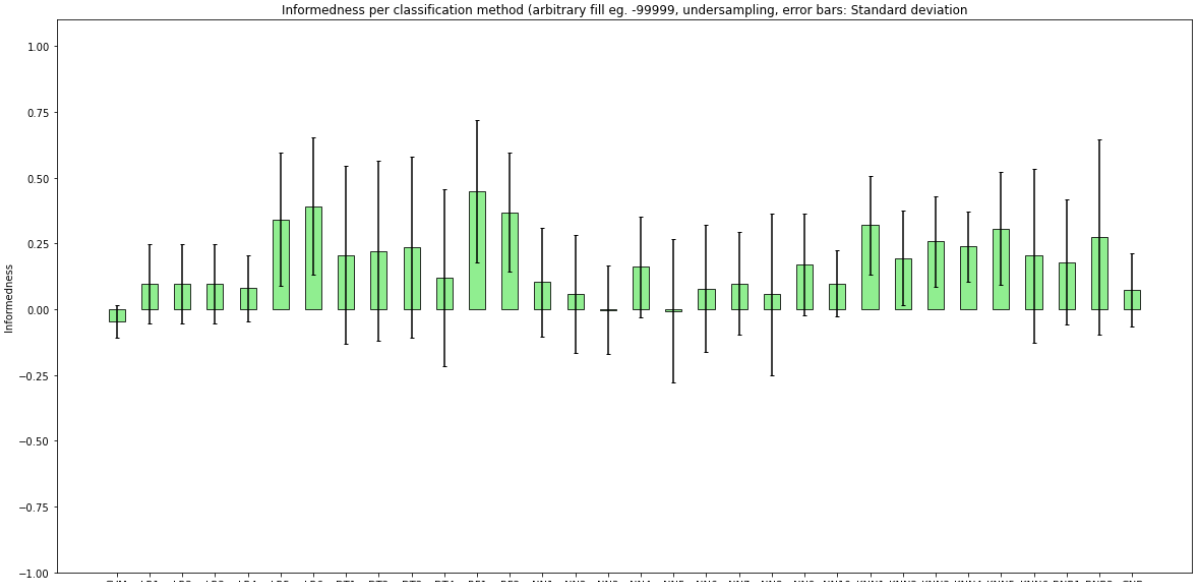

FM per classification method (arbitrary fill eg. -99999, SMOTE-oversampling, error bars: Standard deviation
FM
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB
FM per classification method (formula fill: -R*(log10(R)+100), SMOTE-oversampling, error bars: Standard deviation
FM
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB
FM per classification method (Iterative imputation, SMOTE-oversampling, error bars: Standard deviation
FM
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

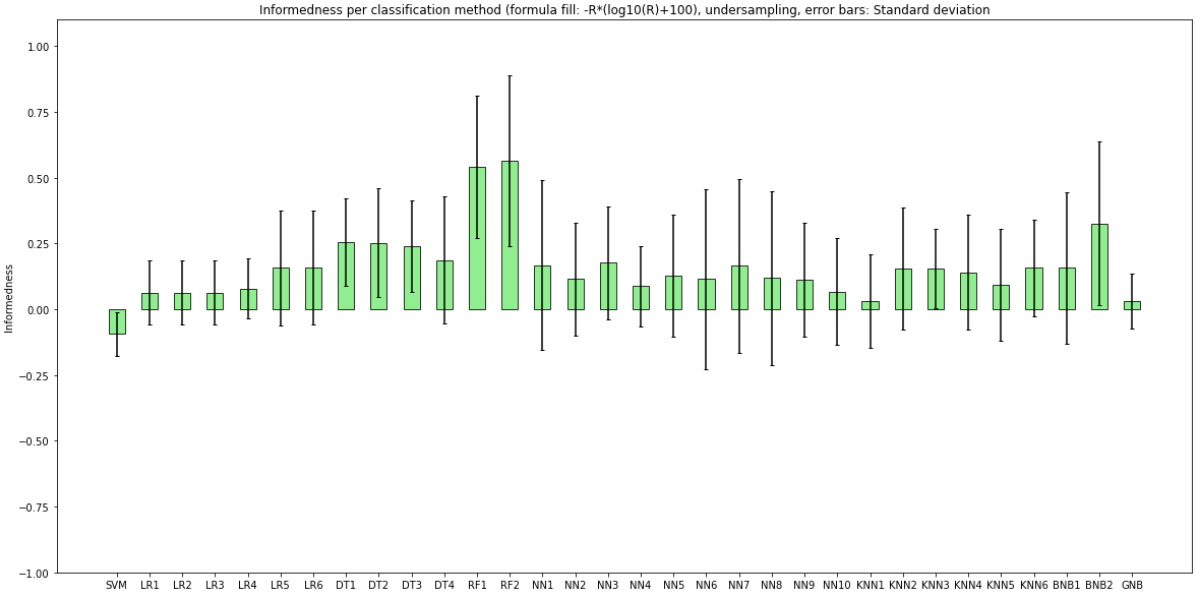

FM per classification method (arbitrary fill eg. -99999, random-oversampling, error bars: Standard deviation
FM
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

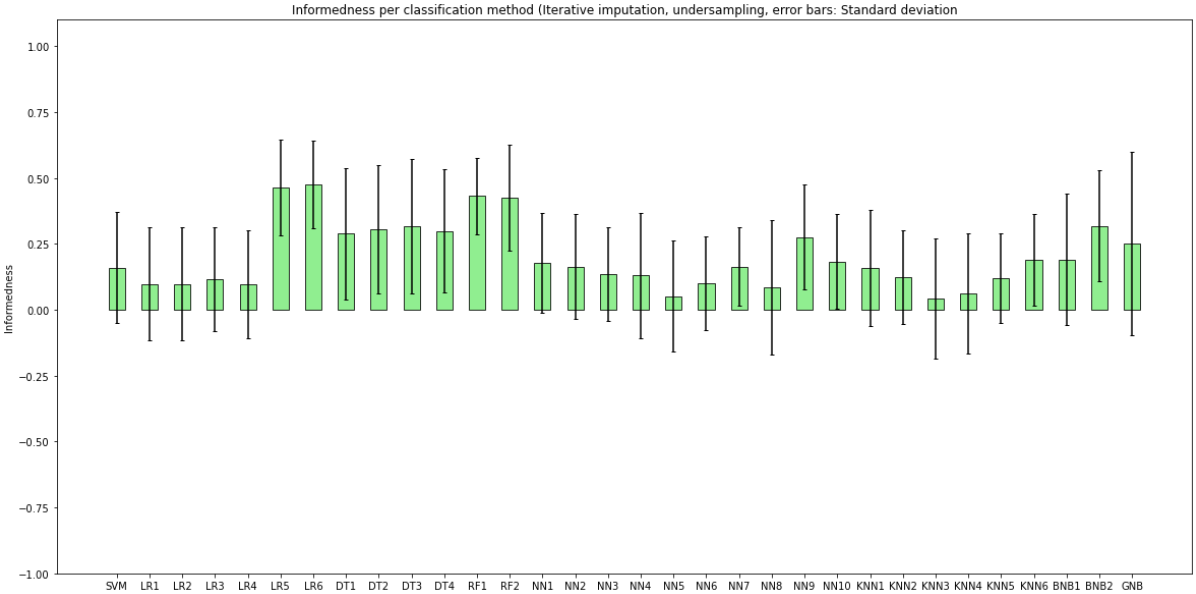

FM per classification method (formula fill: -R*(log10(R)+100), random-oversampling, error bars: Standard deviation
FM
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

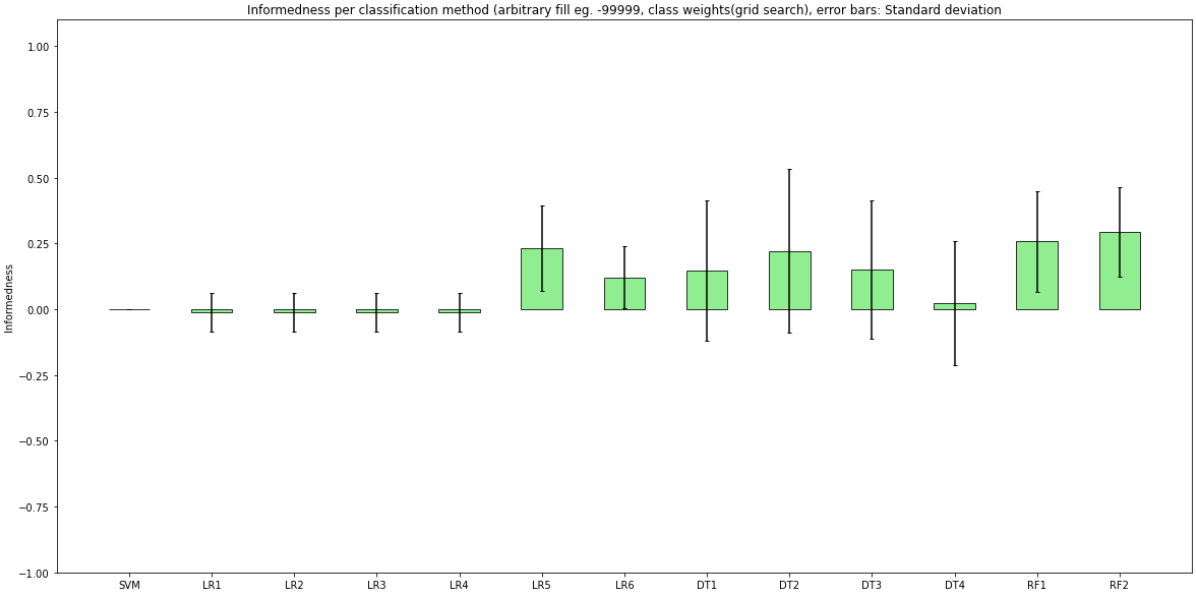

FM per classification method (Iterative imputation, random-oversampling, error bars: Standard deviation
FM
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

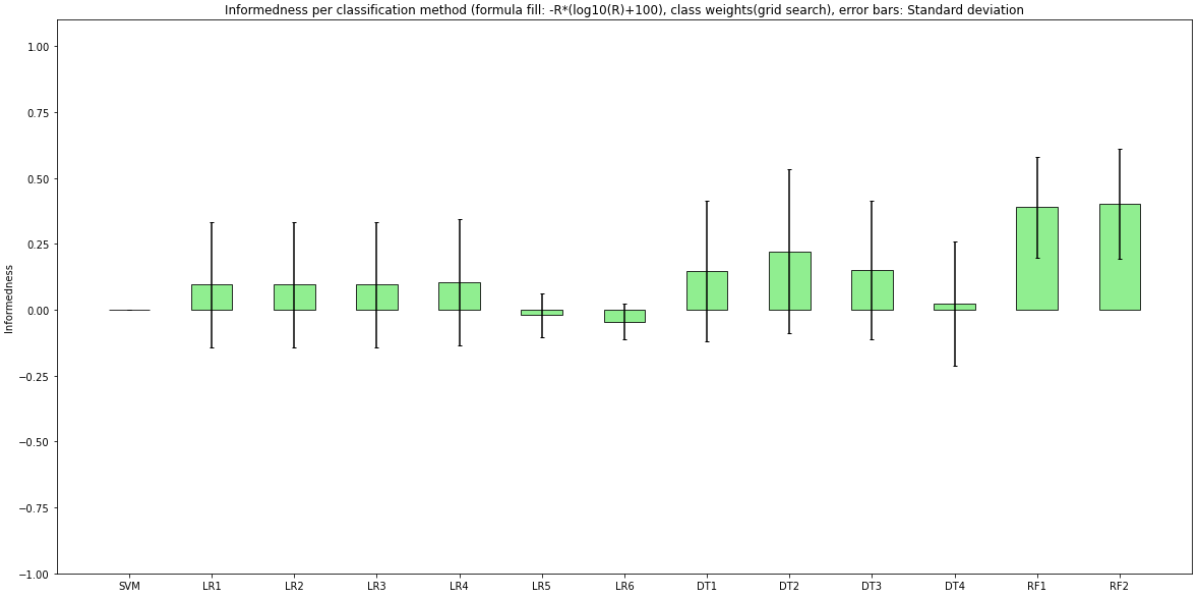

Informedness per classification method (arbitrary fill eg. -99999, no undersampling, error bars: Standard deviation
Informedness per classification method (formula fill: -R*(log10(R)+100), no undersampling, error bars: Standard deviation
Informedness per classification method (Iterative imputation, no undersampling, error bars: Standard deviation
Informedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

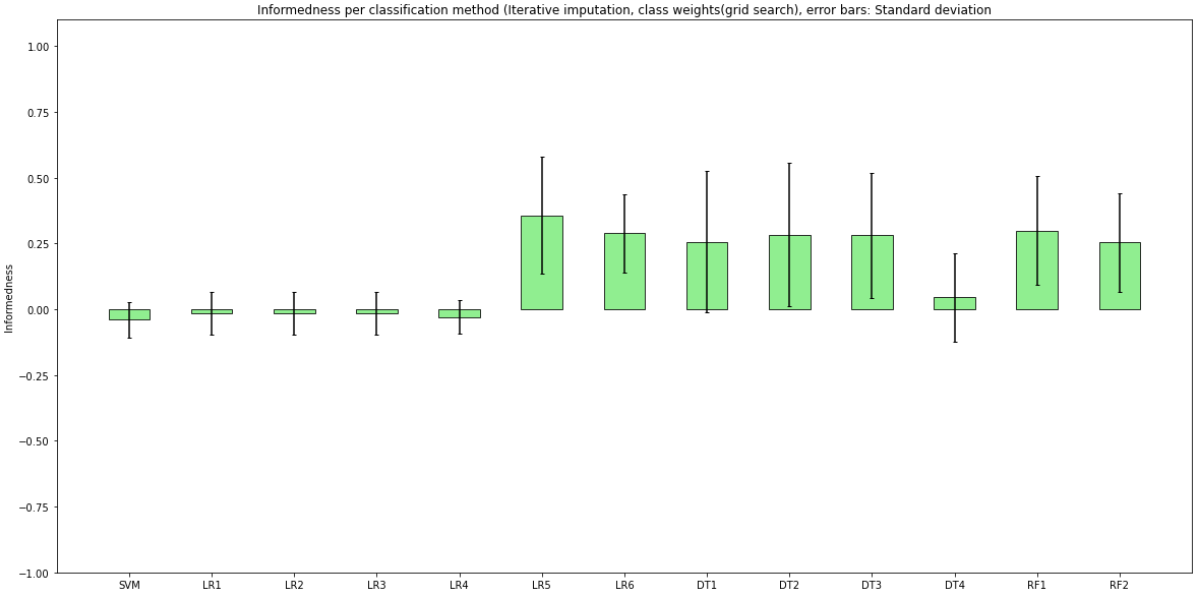

Informedness per classification method (arbitrary fill eg. -99999, undersampling, error bars: Standard deviation
Informedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

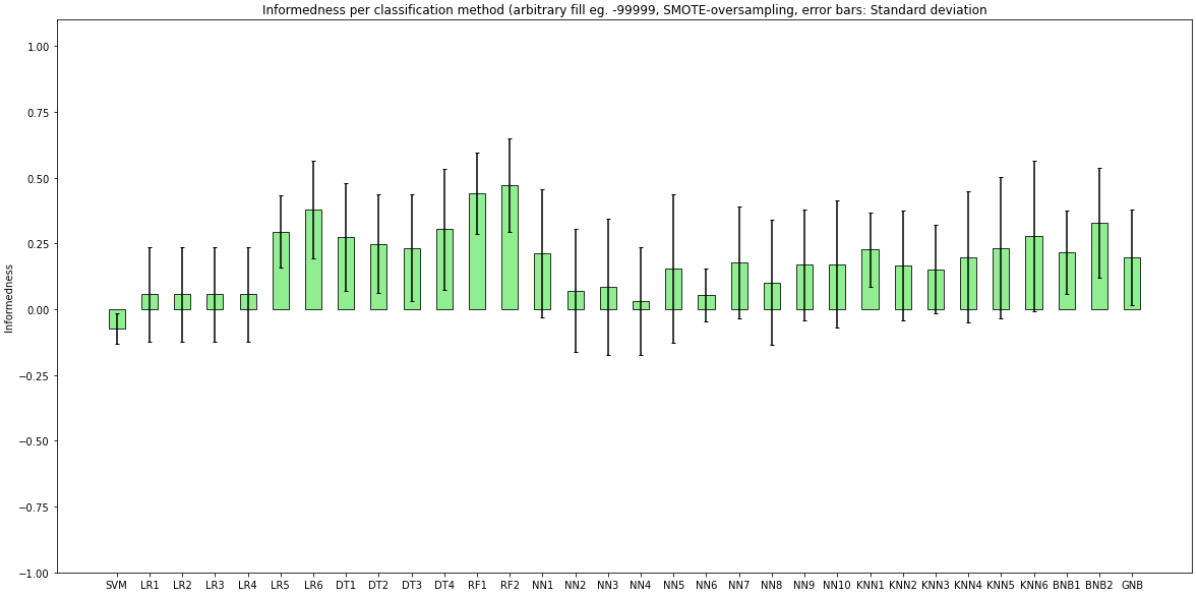

Informedness per classification method (formula fill: -R*(log10(R)+100), undersampling, error bars: Standard deviation
Informedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

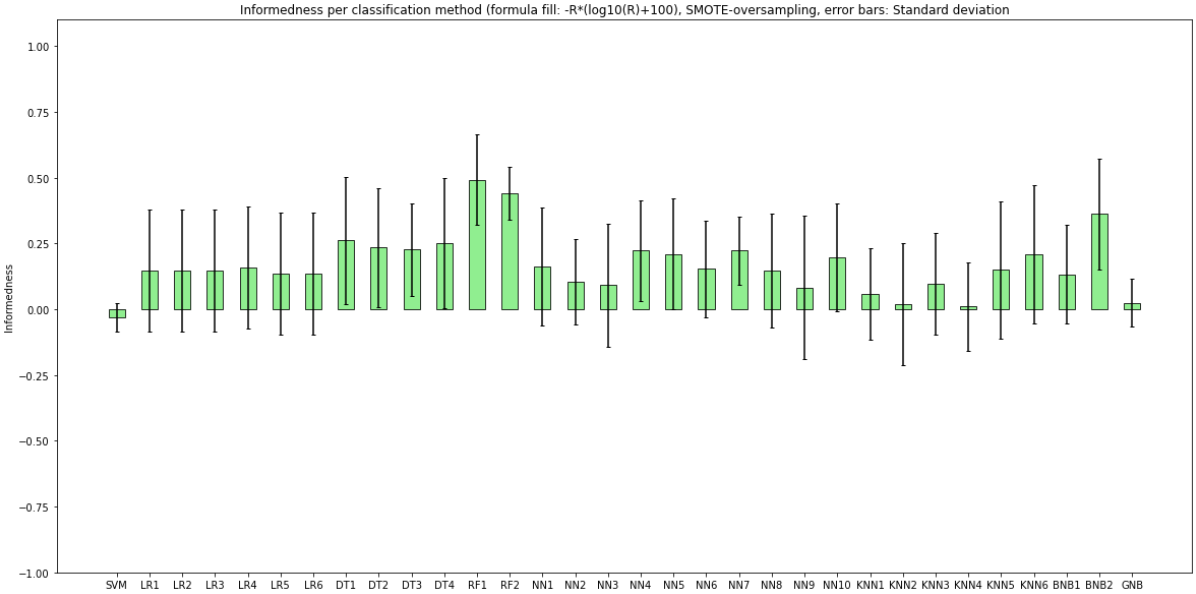

Informedness per classification method (Iterative imputation, undersampling, error bars: Standard deviation
Informedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

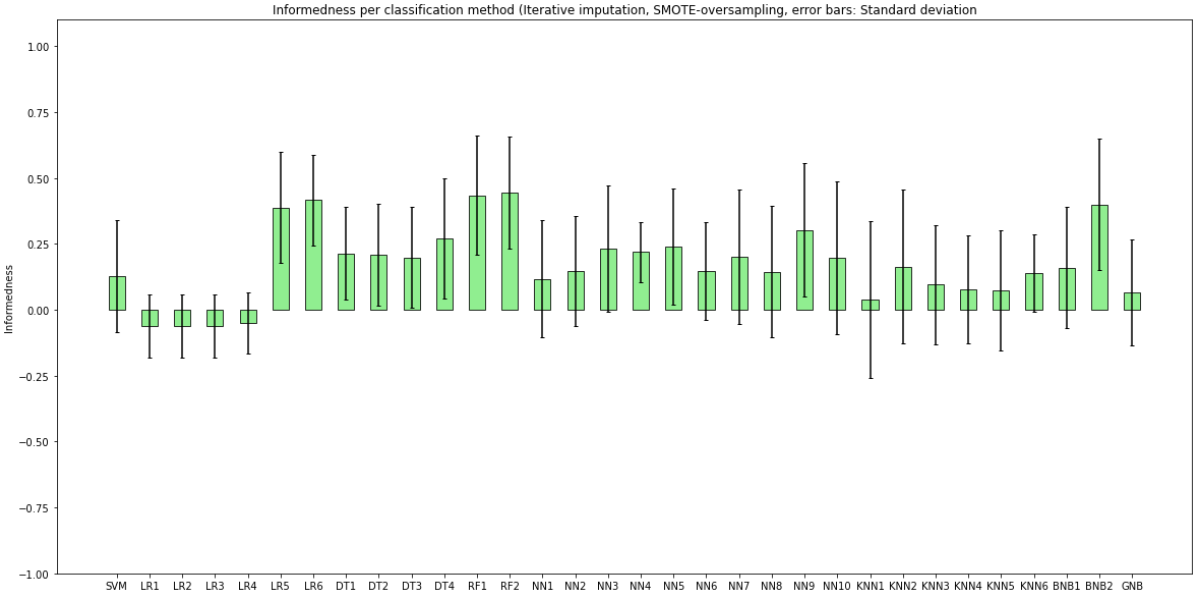

Informedness per classification method (arbitrary fill eg. -99999, class weights(grid search), error bars: Standard deviation
Informedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM
LR1
LR2
LR3
LR4
LR5
LR6
DT1
DT2
DT3
DT4
RF1
RF2

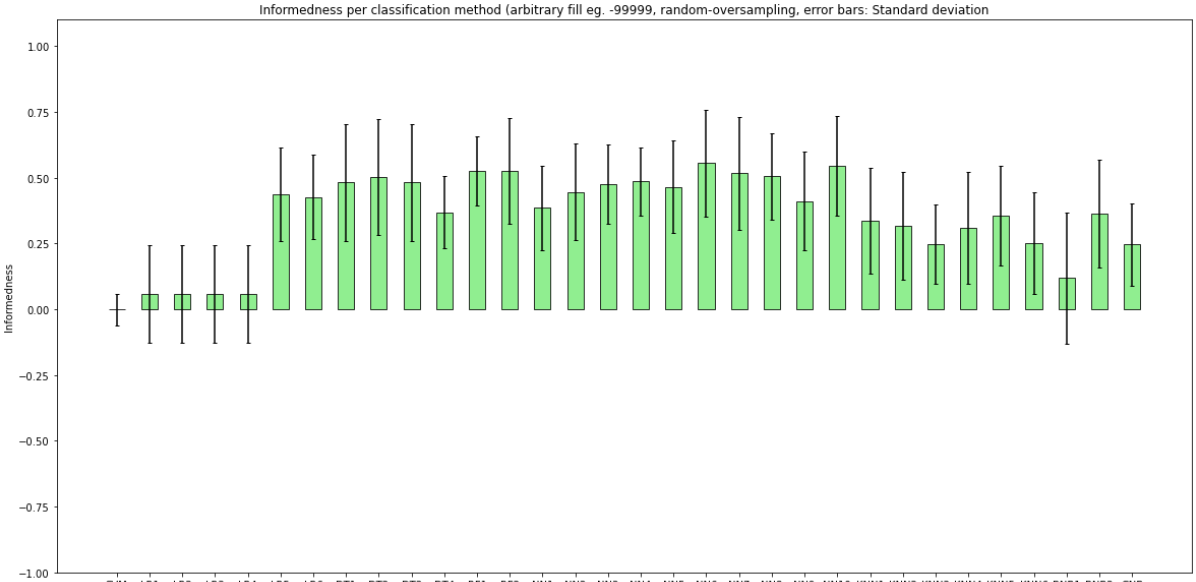

Informedness per classification method (formula fill: -R*(log10(R)+100), class weights(grid search), error bars: Standard deviation
Informedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM
LR1
LR2
LR3
LR4
LR5
LR6
DT1
DT2
DT3
DT4
RF1
RF2

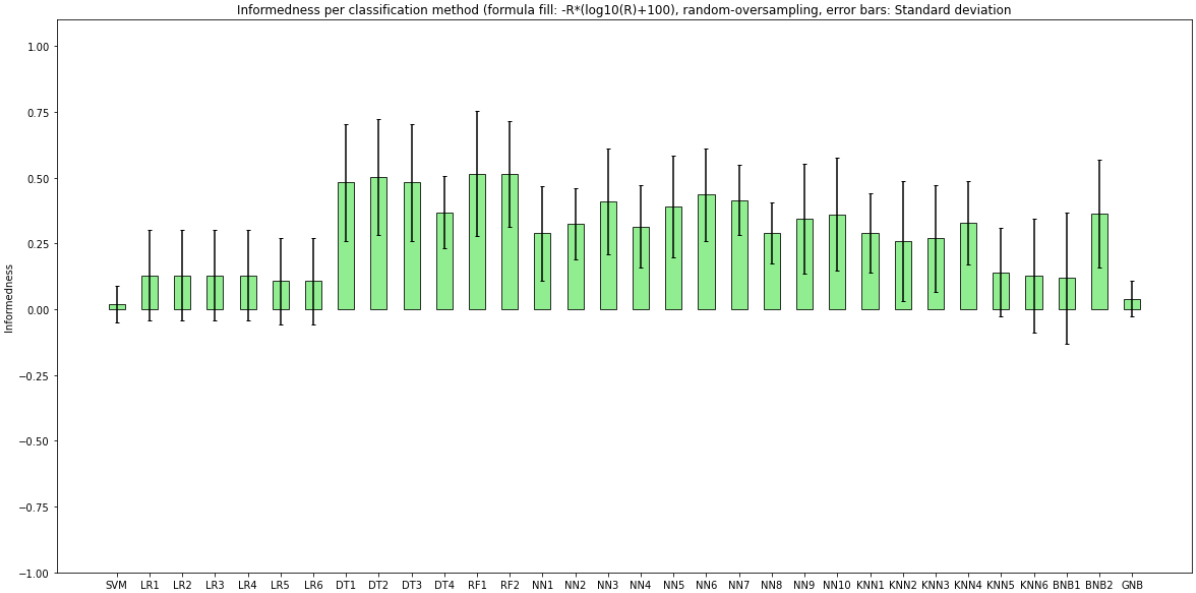

Informedness per classification method (Iterative imputation, class weights(grid search), error bars: Standard deviation
Informedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM
LR1
LR2
LR3
LR4
LR5
LR6
DT1
DT2
DT3
DT4
RF1
RF2

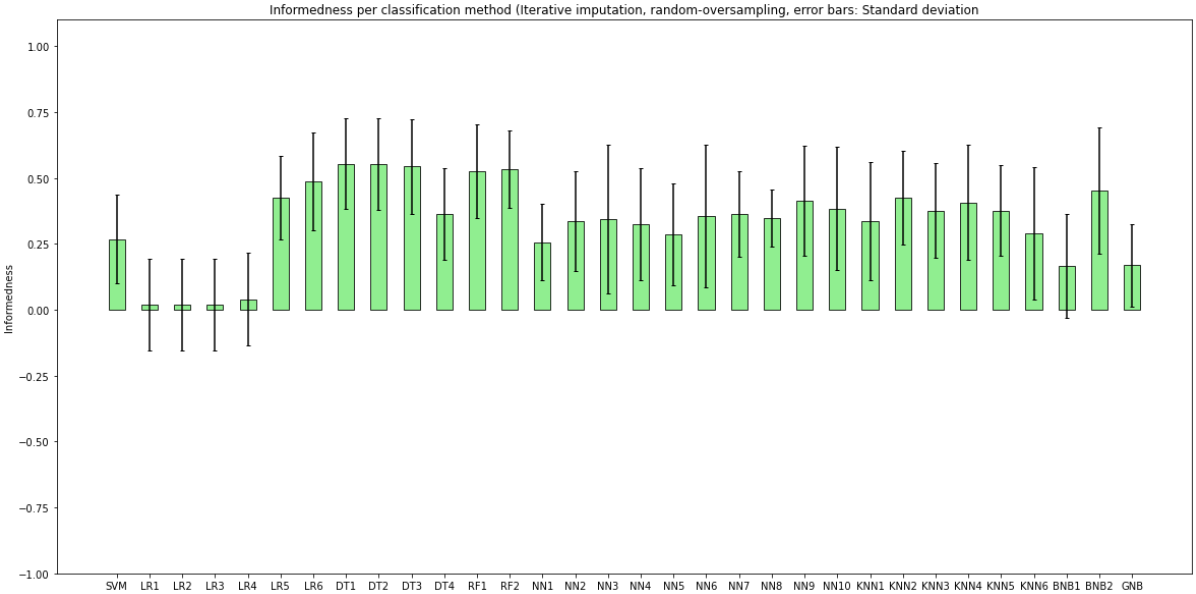

Informedness per classification method (arbitrary fill eg. -99999, SMOTE-oversampling, error bars: Standard deviation
Informedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

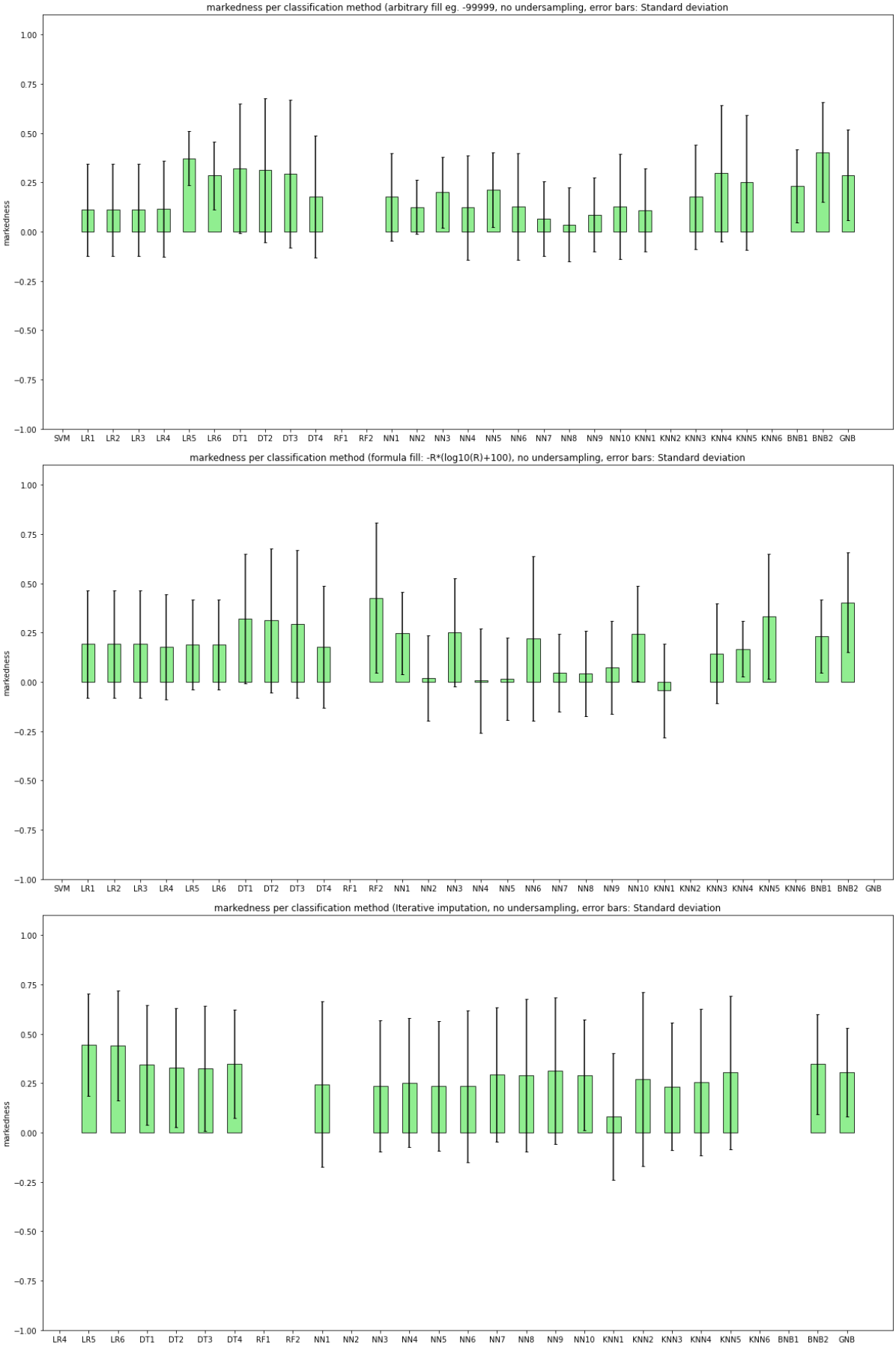

Informedness per classification method (formula fill: -R*(log10(R)+100), SMOTE-oversampling, error bars: Standard deviation
Informedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

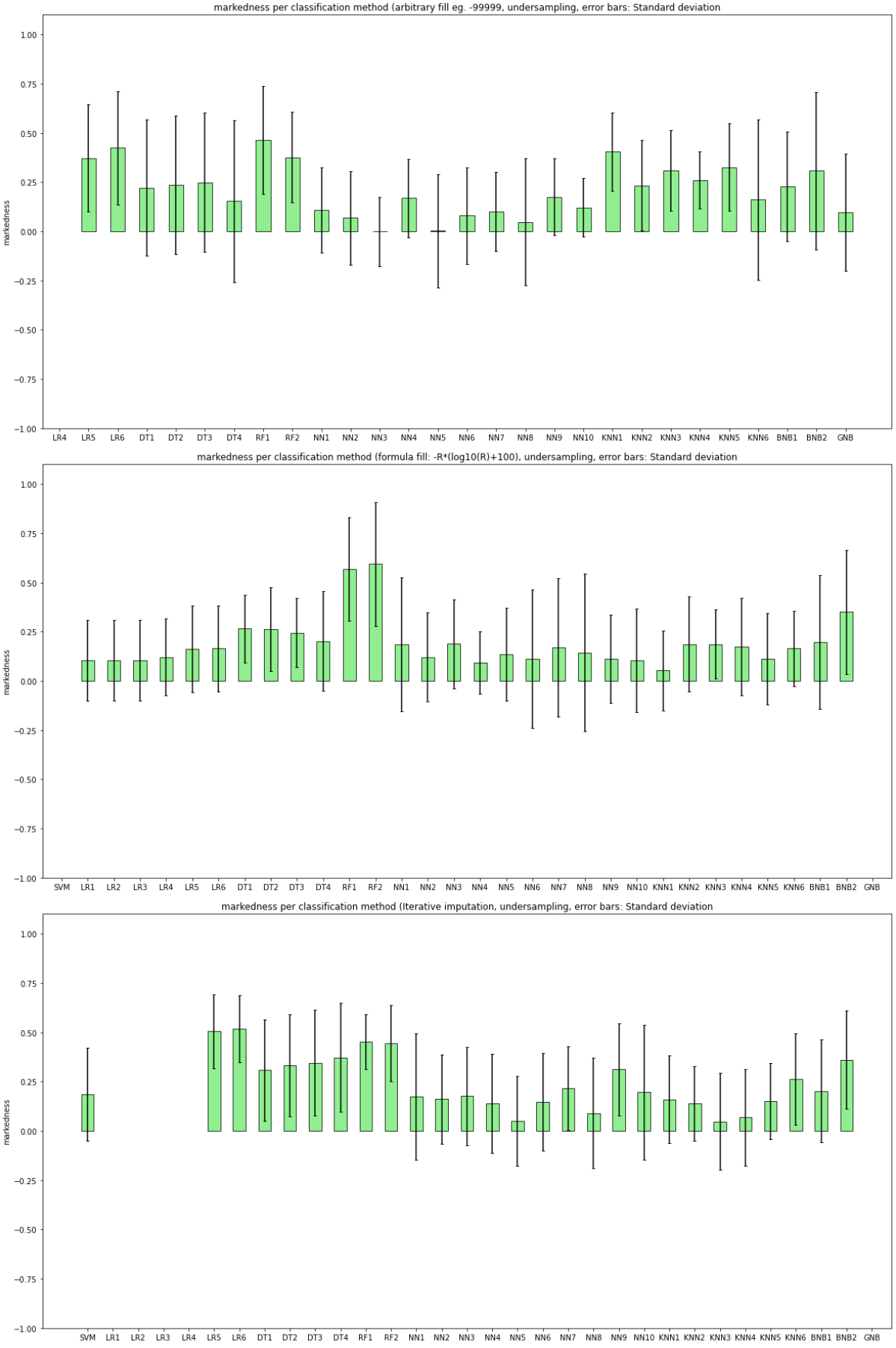

Informedness per classification method (Iterative imputation, SMOTE-oversampling, error bars: Standard deviation
Informedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

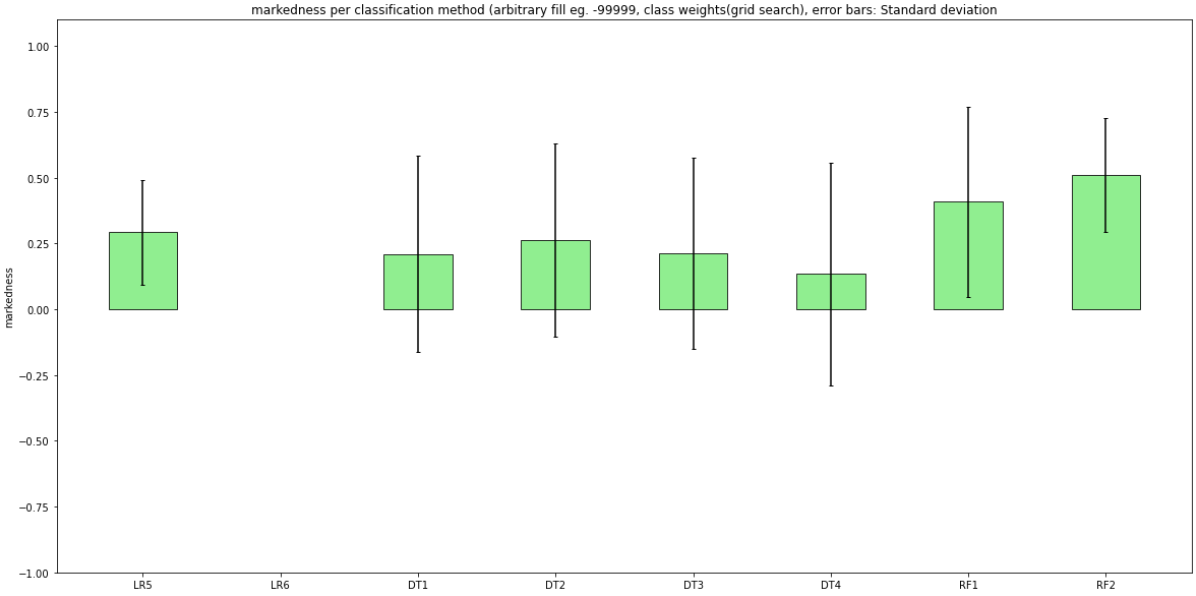

Informedness per classification method (arbitrary fill eg. -99999, random-oversampling, error bars: Standard deviation
Informedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

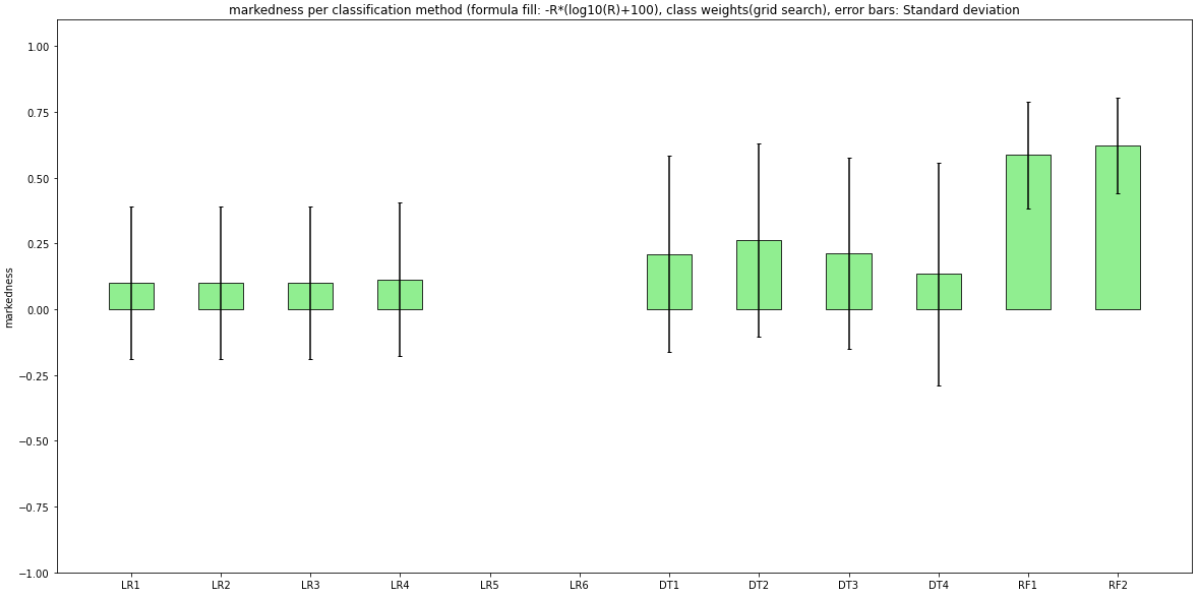

Informedness per classification method (formula fill: -R*(log10(R)+100), random-oversampling, error bars: Standard deviation
Informedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

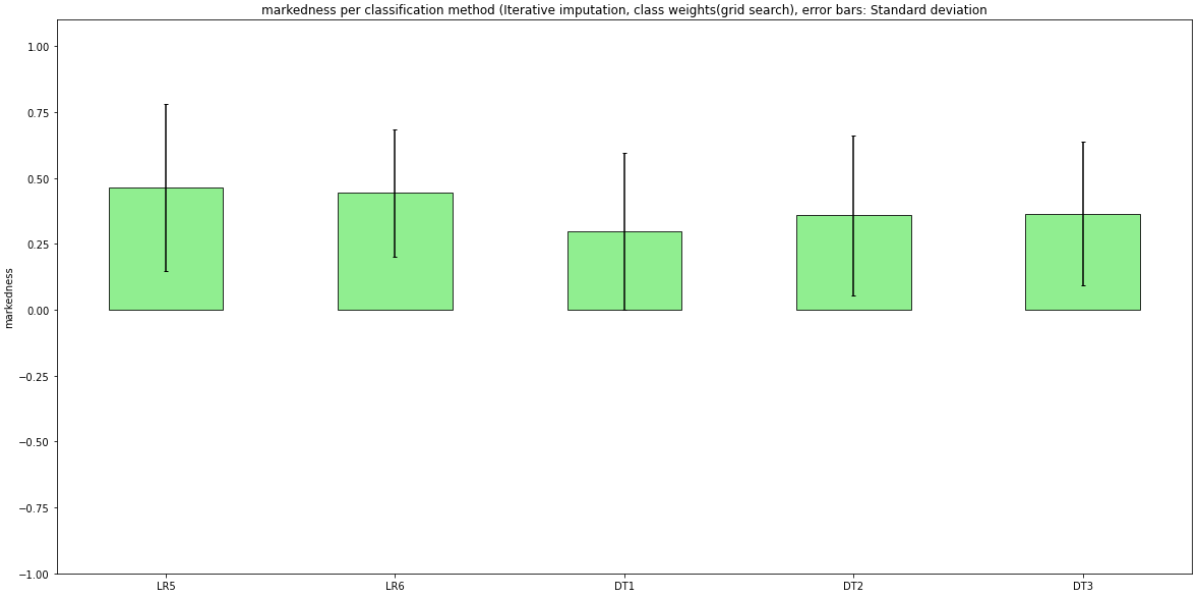

Informedness per classification method (Iterative imputation, random-oversampling, error bars: Standard deviation
Informedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

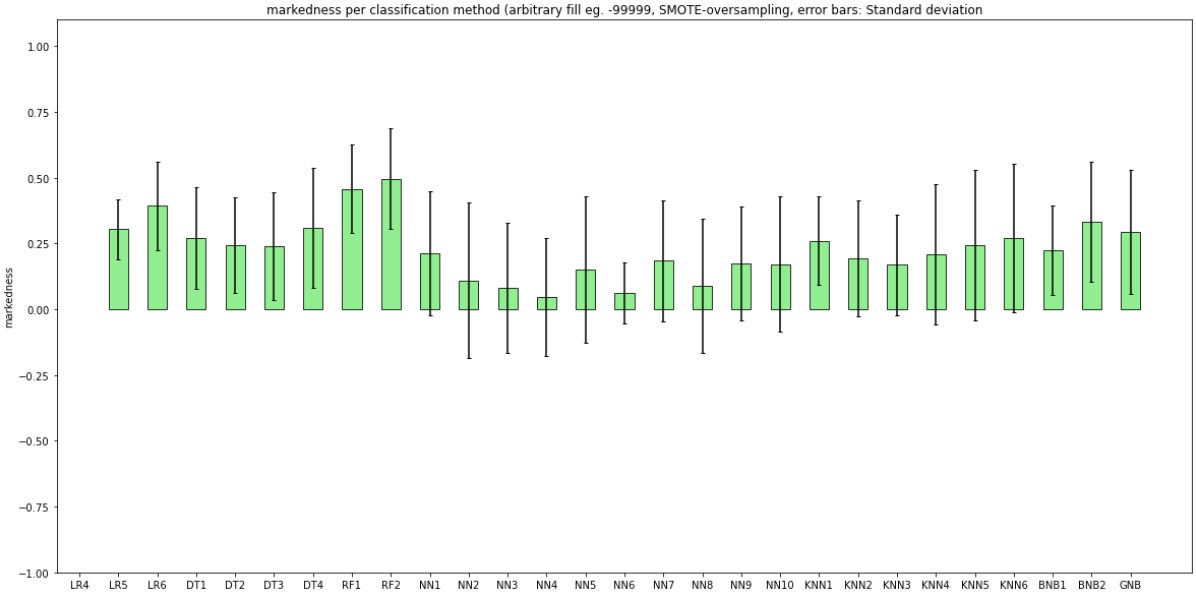

markedness per classification method (arbitrary fill eg. -99999, no undersampling, error bars: Standard deviation
markedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB
markedness per classification method (formula fill: -R*(log10(R)+100), no undersampling, error bars: Standard deviation
markedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB
markedness per classification method (Iterative imputation, no undersampling, error bars: Standard deviation
markedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

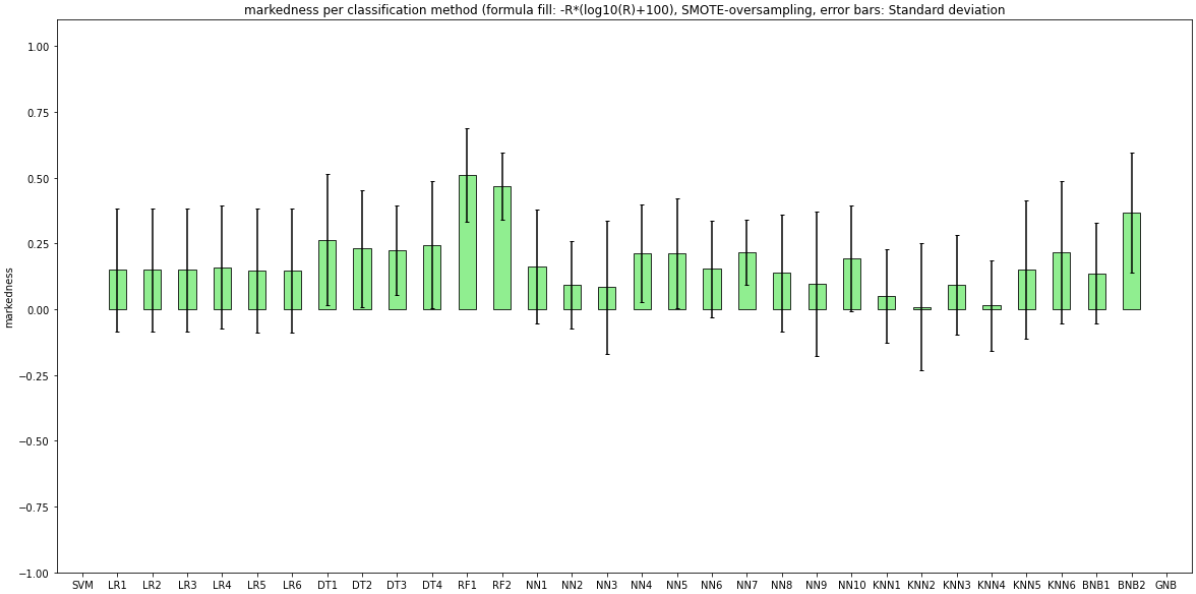

markedness per classification method (arbitrary fill eg. -99999, undersampling, error bars: Standard deviation
markedness per classification method (formula fill: -R*(log10(R)+100), undersampling, error bars: Standard deviation
markedness per classification method (Iterative imputation, undersampling, error bars: Standard deviation
markedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

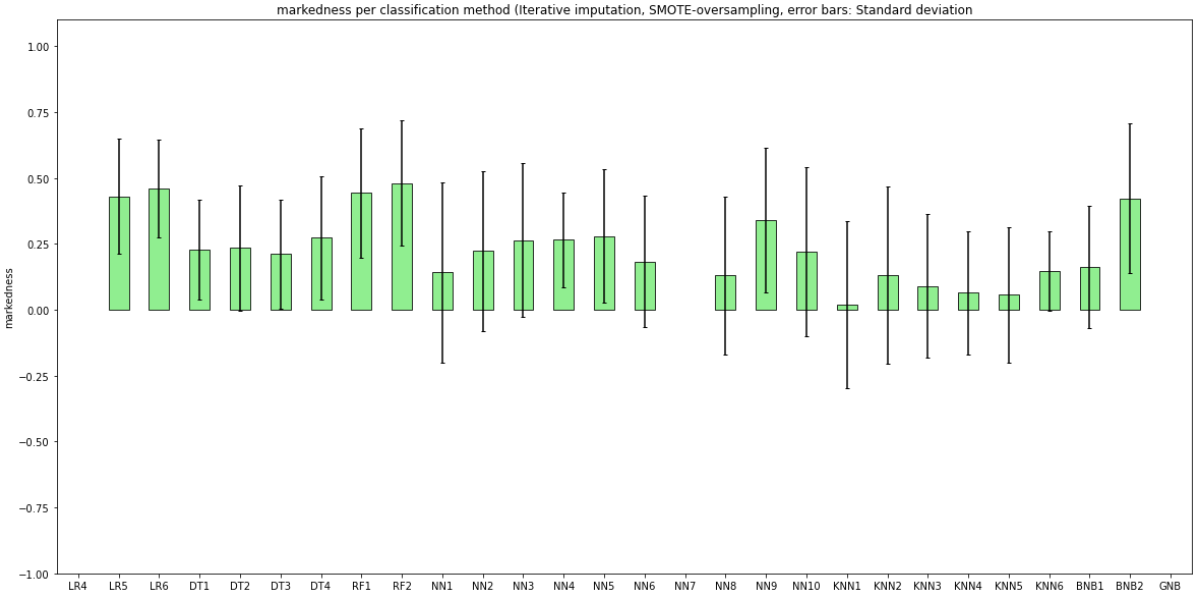

markedness per classification method (arbitrary fill eg. -99999, class weights(grid search), error bars: Standard deviation
markedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
LR5
LR6
DT1
DT2
DT3
DT4
RF1
RF2

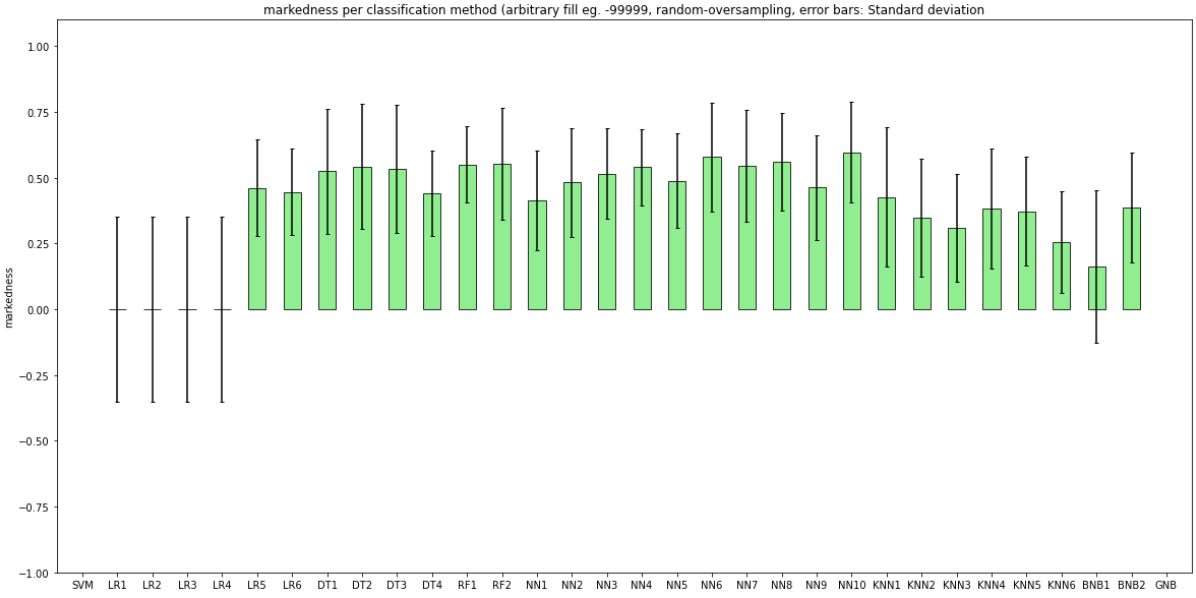

markedness per classification method (formula fill: -R*(log10(R)+100), class weights(grid search), error bars: Standard deviation
markedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
LR1
LR2
LR3
LR4
LR5
LR6
DT1
DT2
DT3
DT4
RF1
RF2

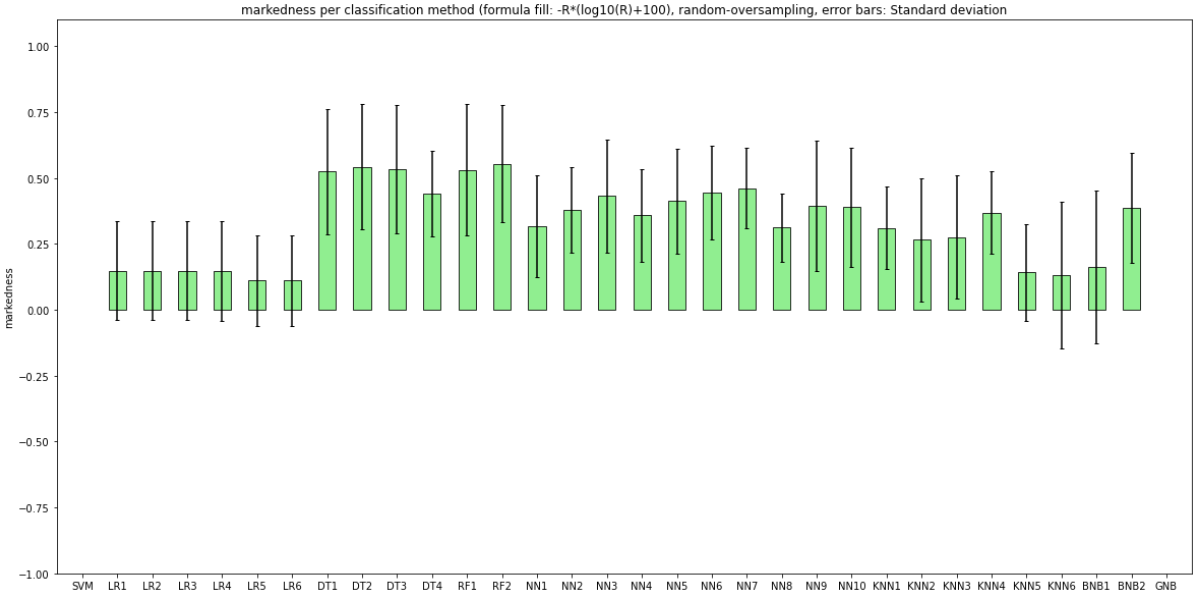

markedness per classification method (Iterative imputation, class weights(grid search), error bars: Standard deviation
markedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
LR5
LR6
DT1
DT2
DT3

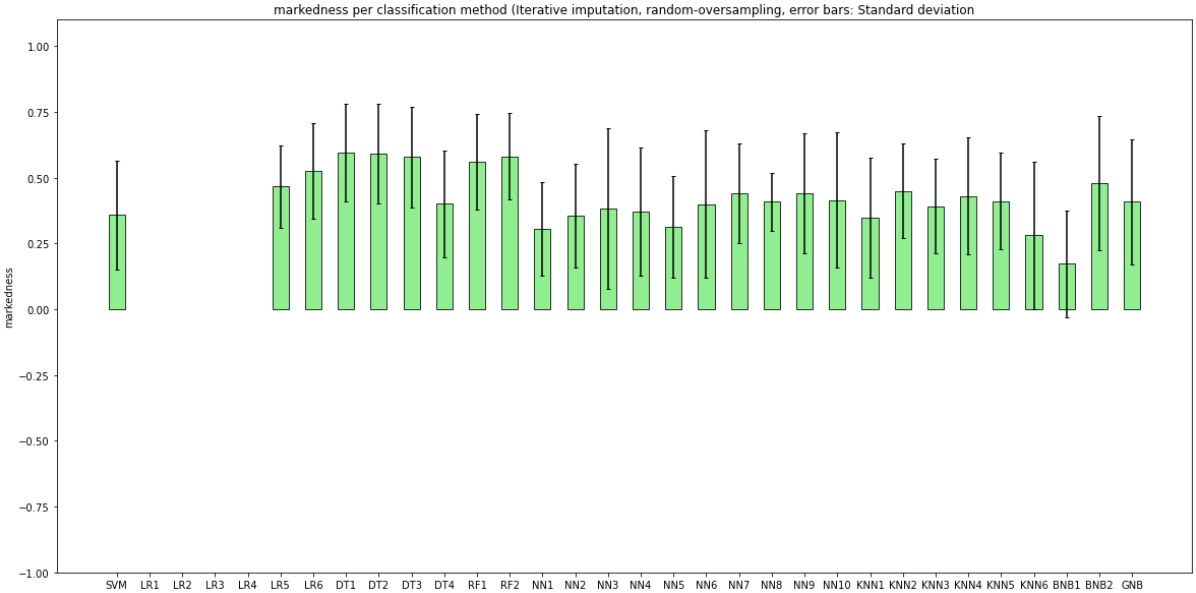

markedness per classification method (arbitrary fill eg. -99999, SMOTE-oversampling, error bars: Standard deviation
markedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

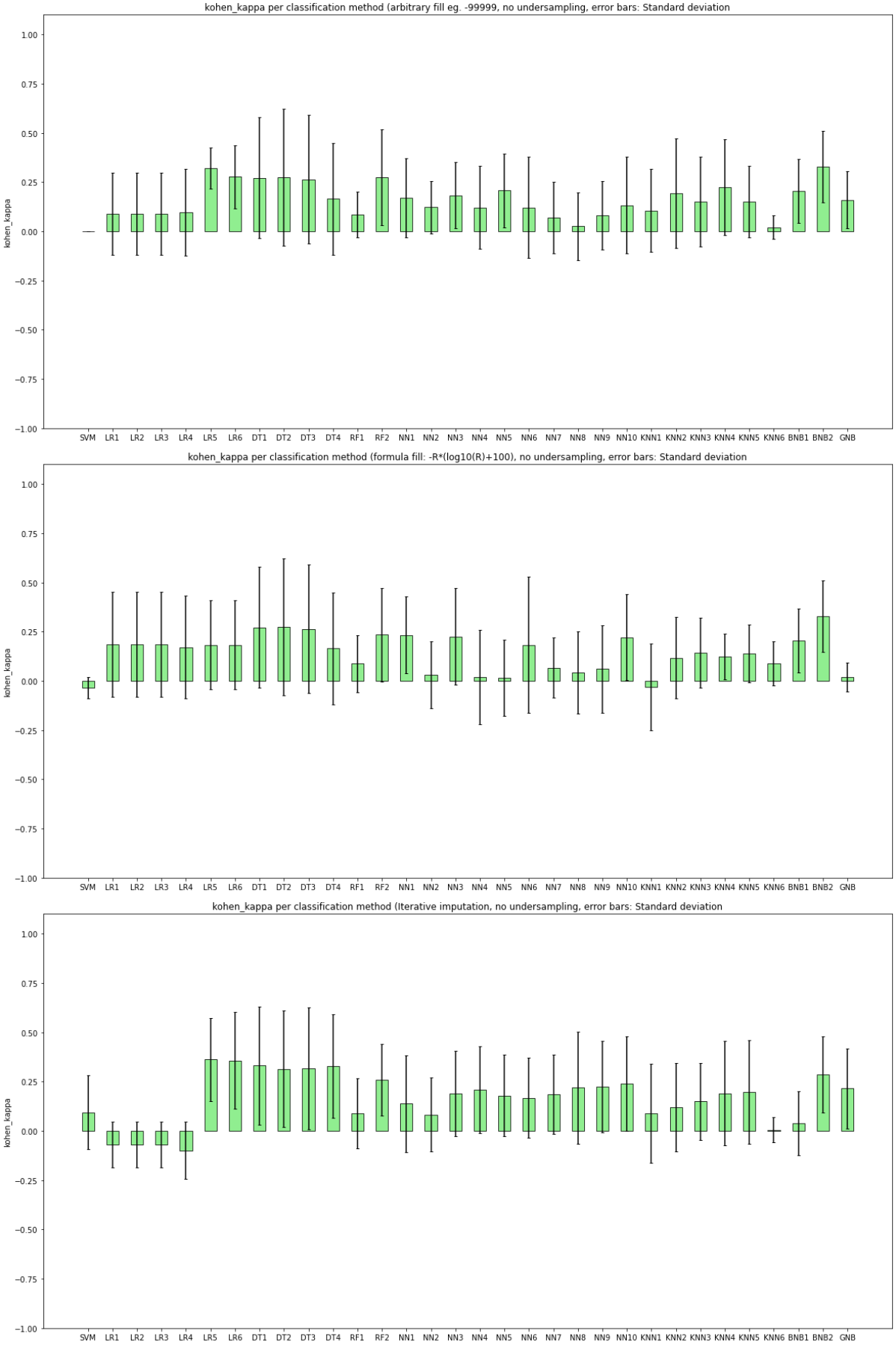

markedness per classification method (formula fill: -R*(log10(R)+100), SMOTE-oversampling, error bars: Standard deviation
markedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

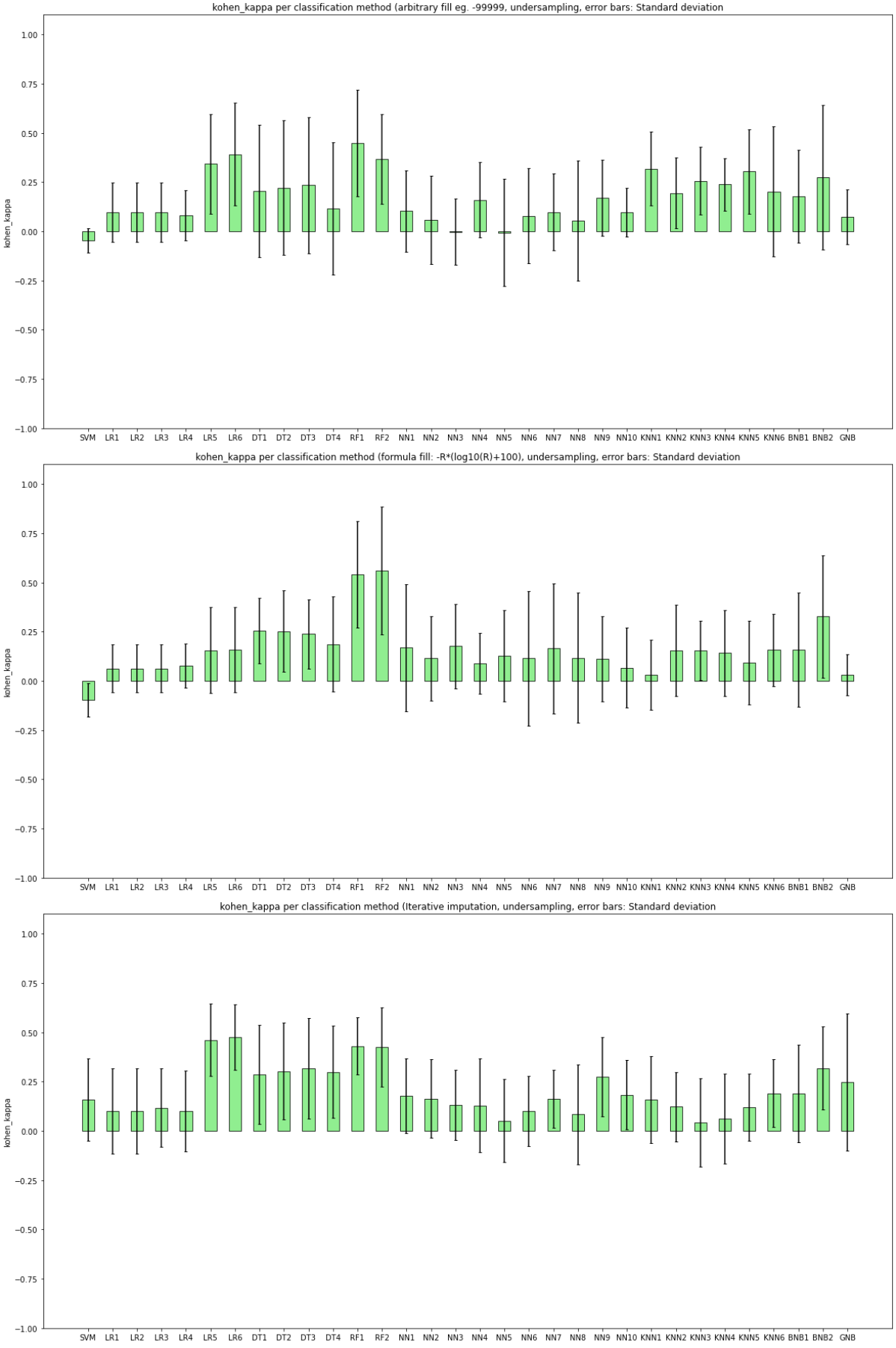

markedness per classification method (Iterative imputation, SMOTE-oversampling, error bars: Standard deviation
markedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

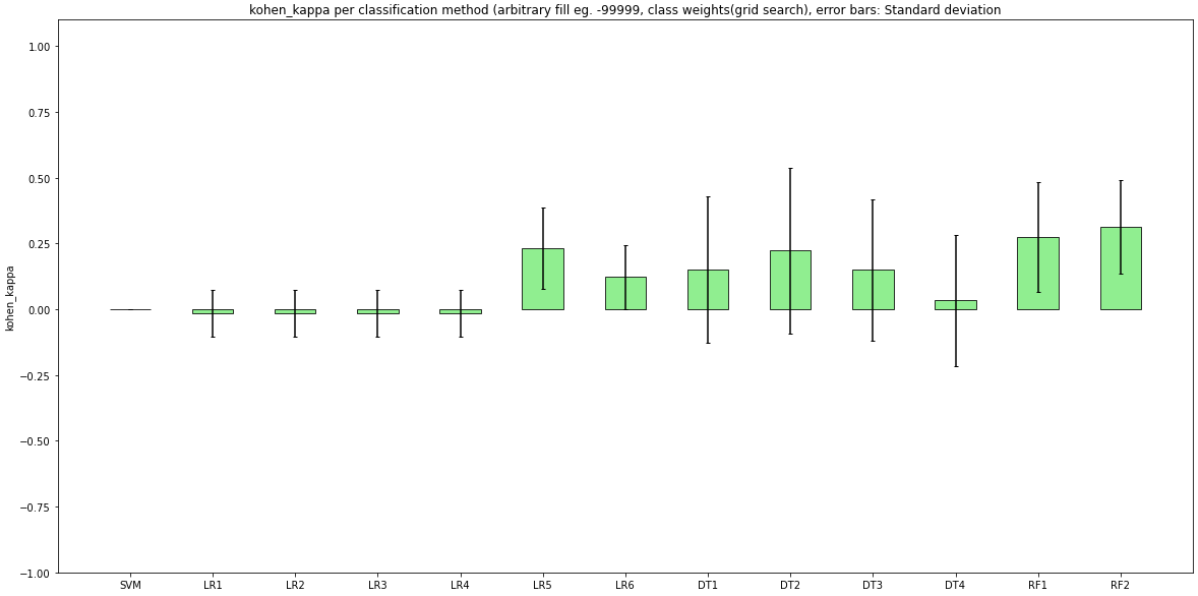

markedness per classification method (arbitrary fill eg. -99999, random-oversampling, error bars: Standard deviation
markedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

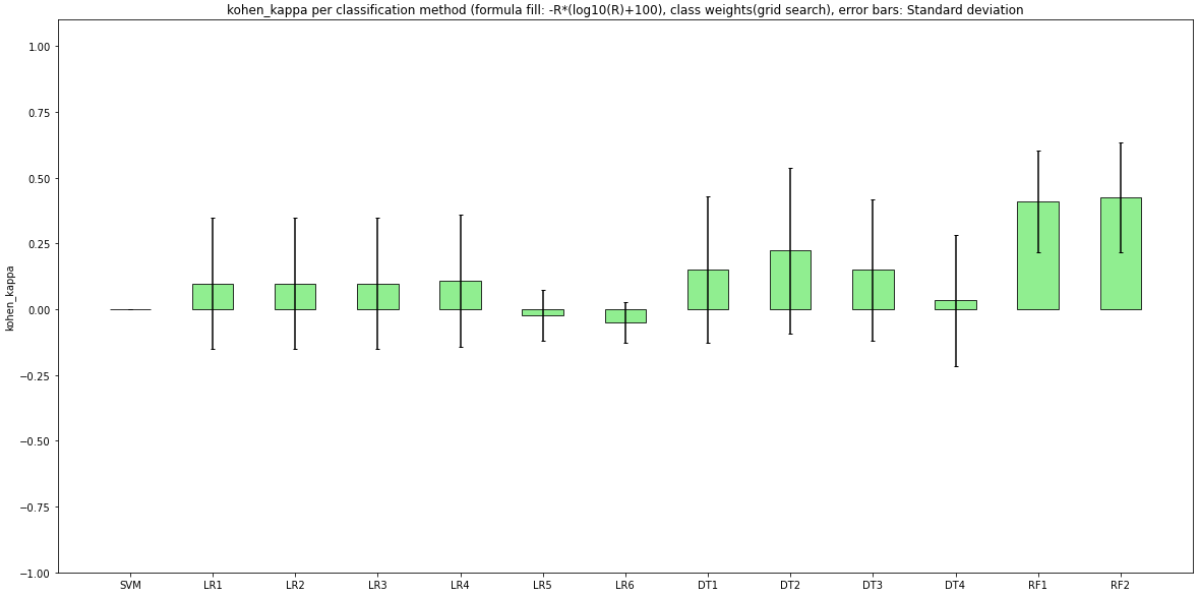

markedness per classification method (formula fill: -R*(log10(R)+100), random-oversampling, error bars: Standard deviation
markedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

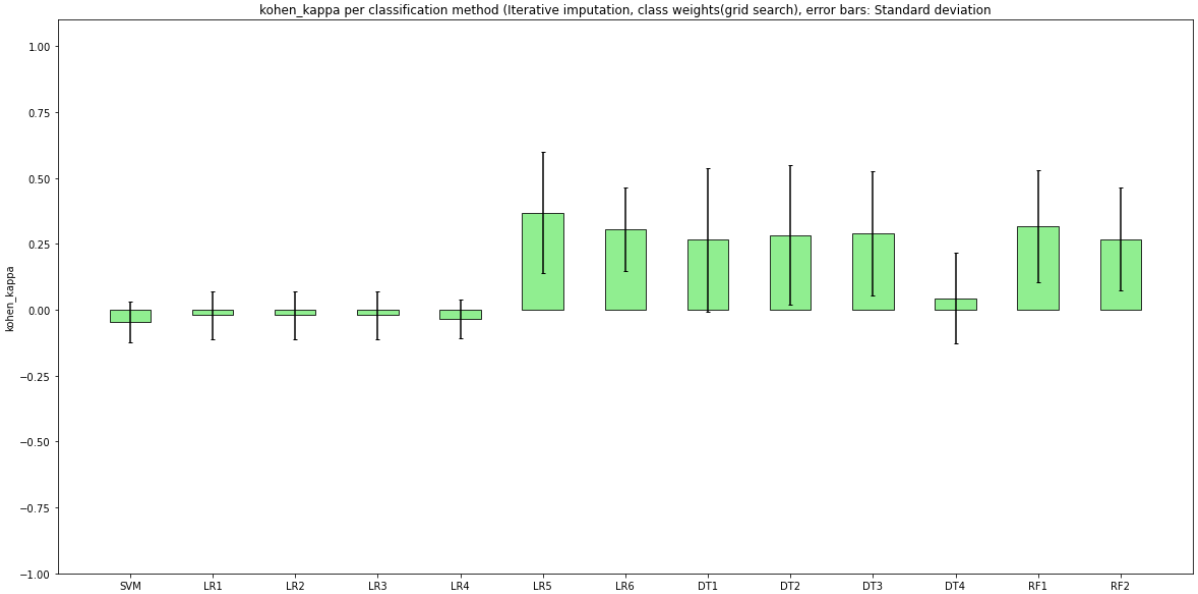

markedness per classification method (Iterative imputation, random-oversampling, error bars: Standard deviation
markedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

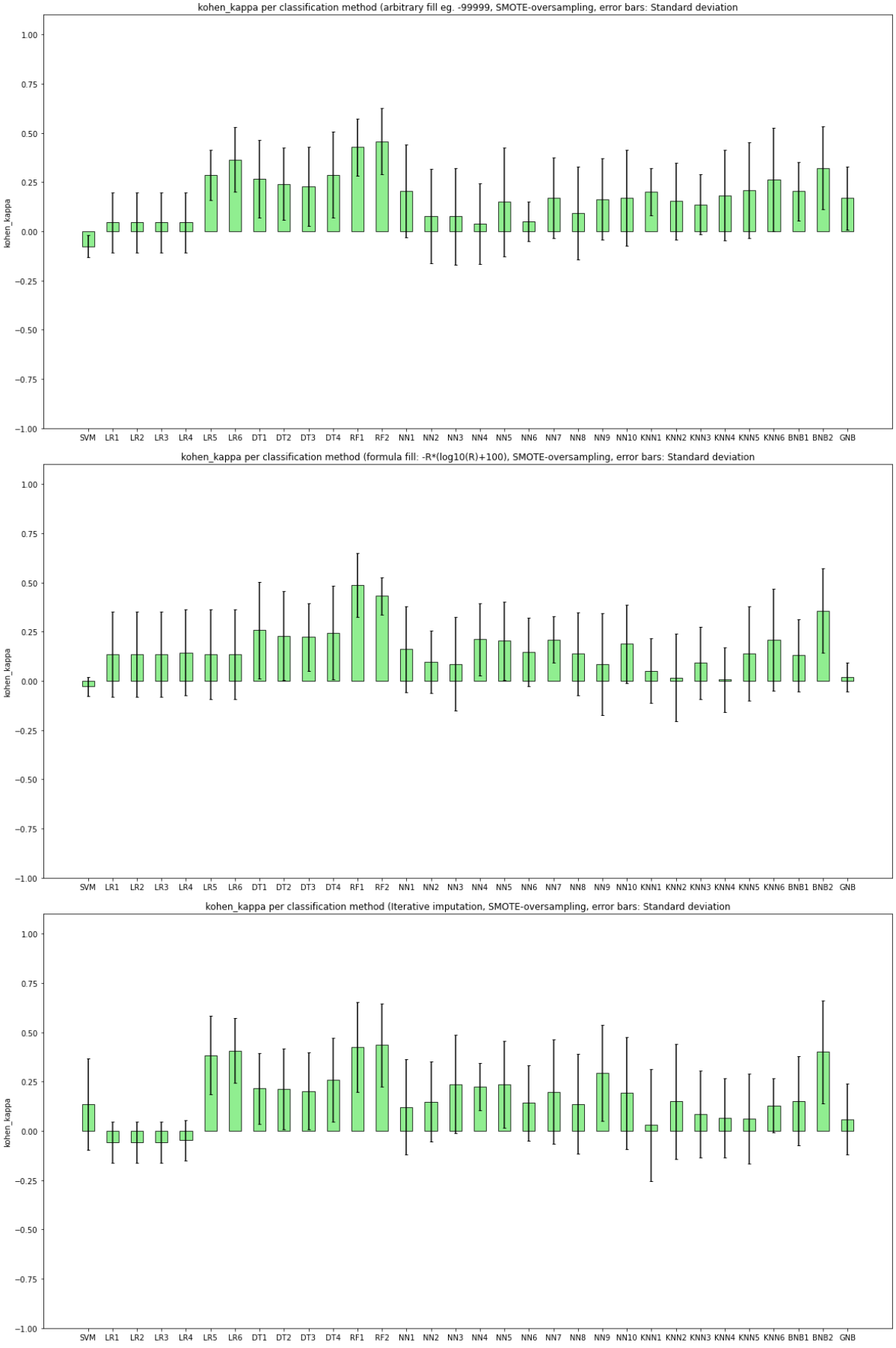

kohen_kappa per classification method (arbitrary fill eg. -99999, no undersampling, error bars: Standard deviation
kohen_kappa per classification method (formula fill: -R*(log10(R)+100), no undersampling, error bars: Standard deviation
kohen_kappa per classification method (Iterative imputation, no undersampling, error bars: Standard deviation
kohen_kappa
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

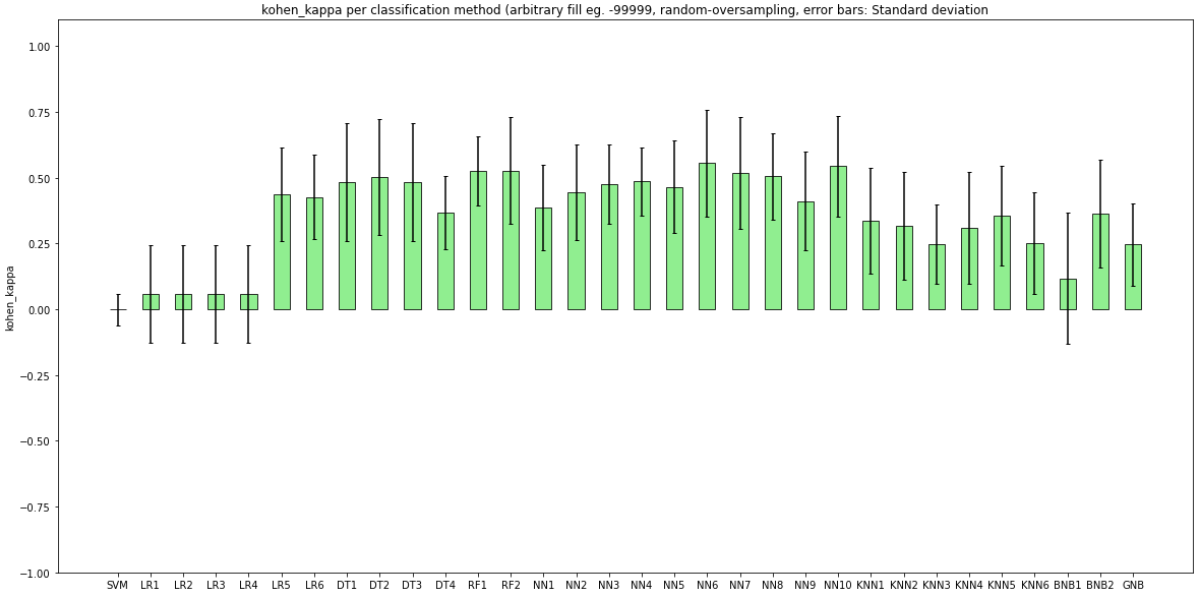

kohen_kappa per classification method (arbitrary fill eg. -99999, undersampling, error bars: Standard deviation
kohen_kappa per classification method (formula fill: -R*(log10(R)+100), undersampling, error bars: Standard deviation
kohen_kappa per classification method (Iterative imputation, undersampling, error bars: Standard deviation
kohen_kappa
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

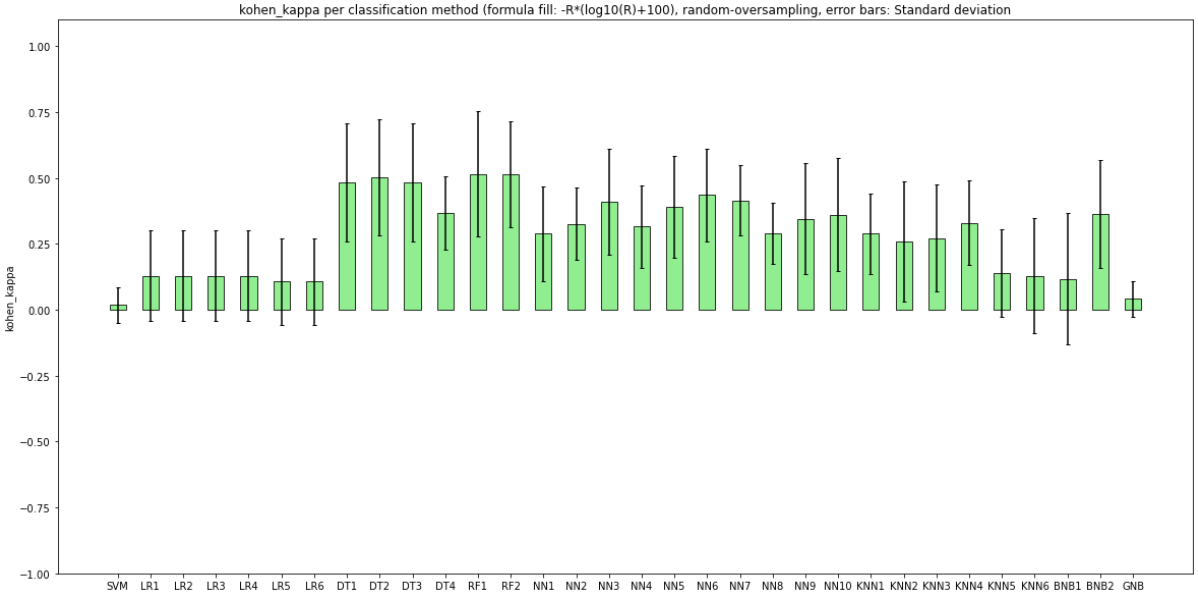

kohen_kappa per classification method (arbitrary fill eg. -99999, class weights(grid search), error bars: Standard deviation
kohen_kappa
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM
LR1
LR2
LR3
LR4
LR5
LR6
DT1
DT2
DT3
DT4
RF1
RF2

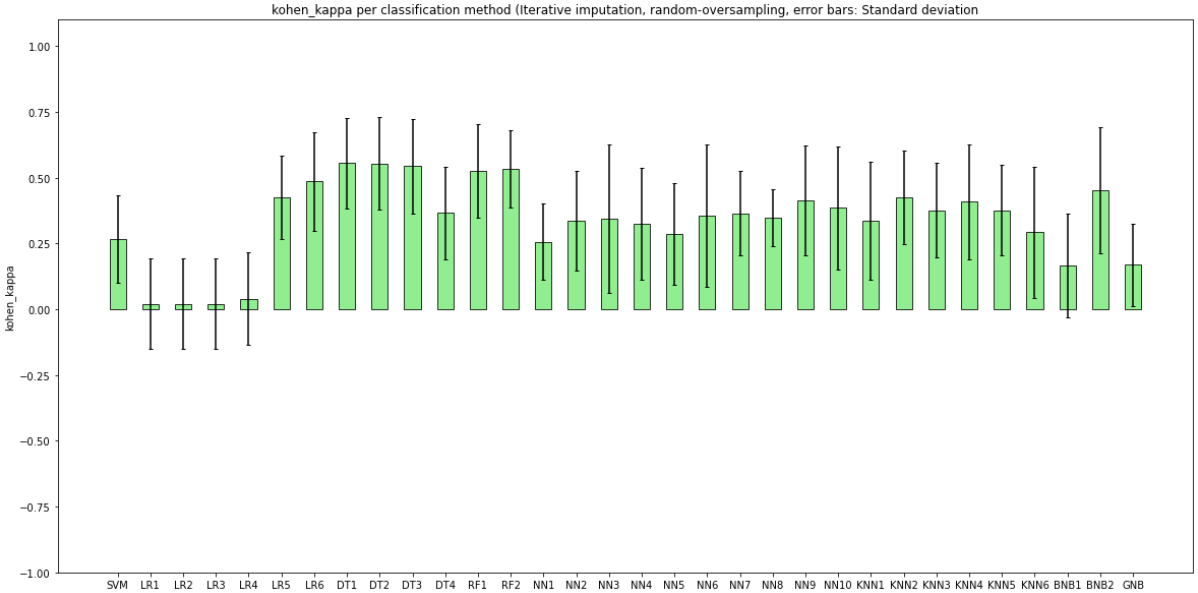

kohen_kappa per classification method (formula fill: -R*(log10(R)+100), class weights(grid search), error bars: Standard deviation
kohen_kappa
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM
LR1
LR2
LR3
LR4
LR5
LR6
DT1
DT2
DT3
DT4
RF1
RF2

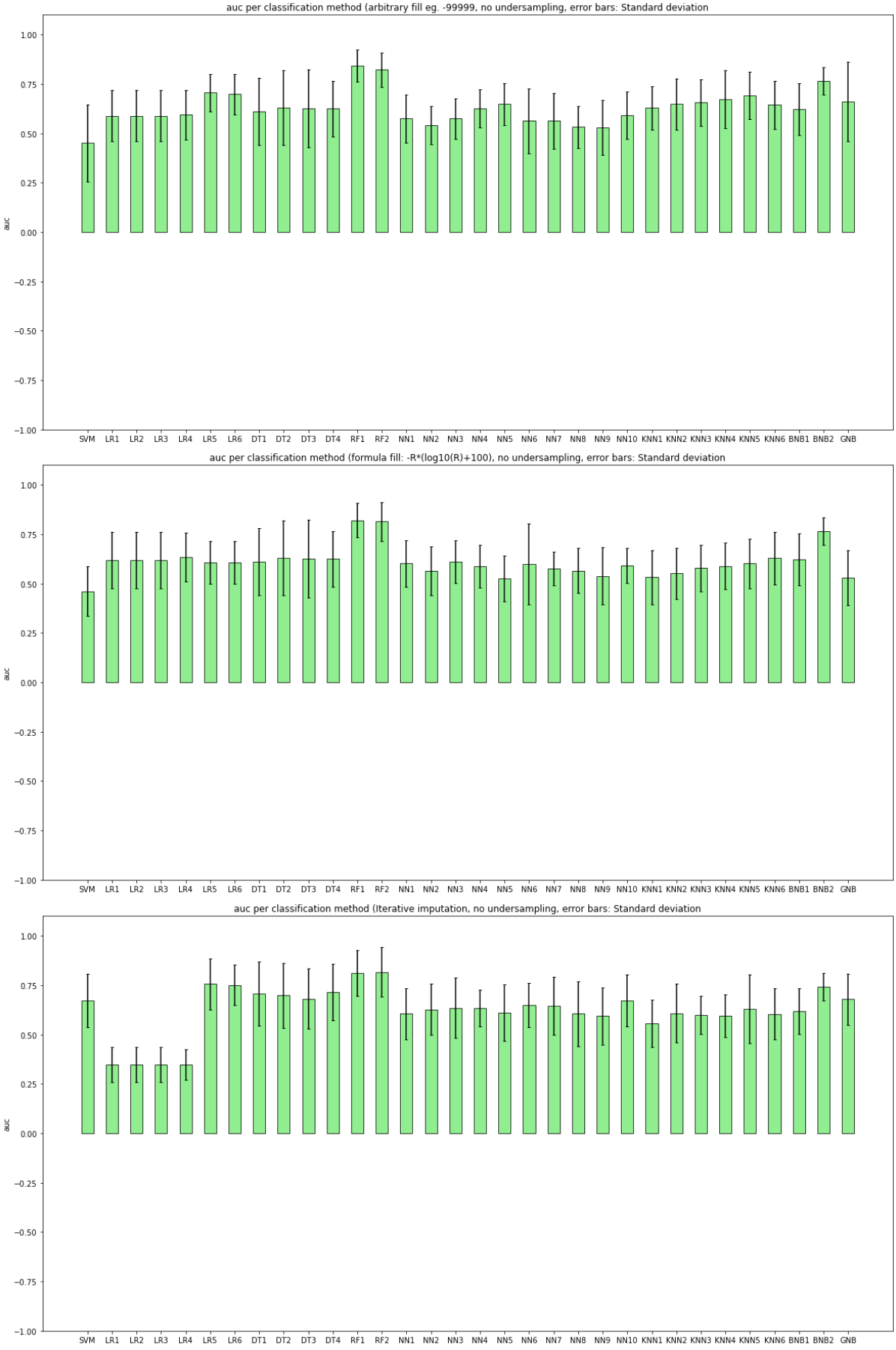

kohen_kappa per classification method (Iterative imputation, class weights(grid search), error bars: Standard deviation
kohen_kappa
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM
LR1
LR2
LR3
LR4
LR5
LR6
DT1
DT2
DT3
DT4
RF1
RF2

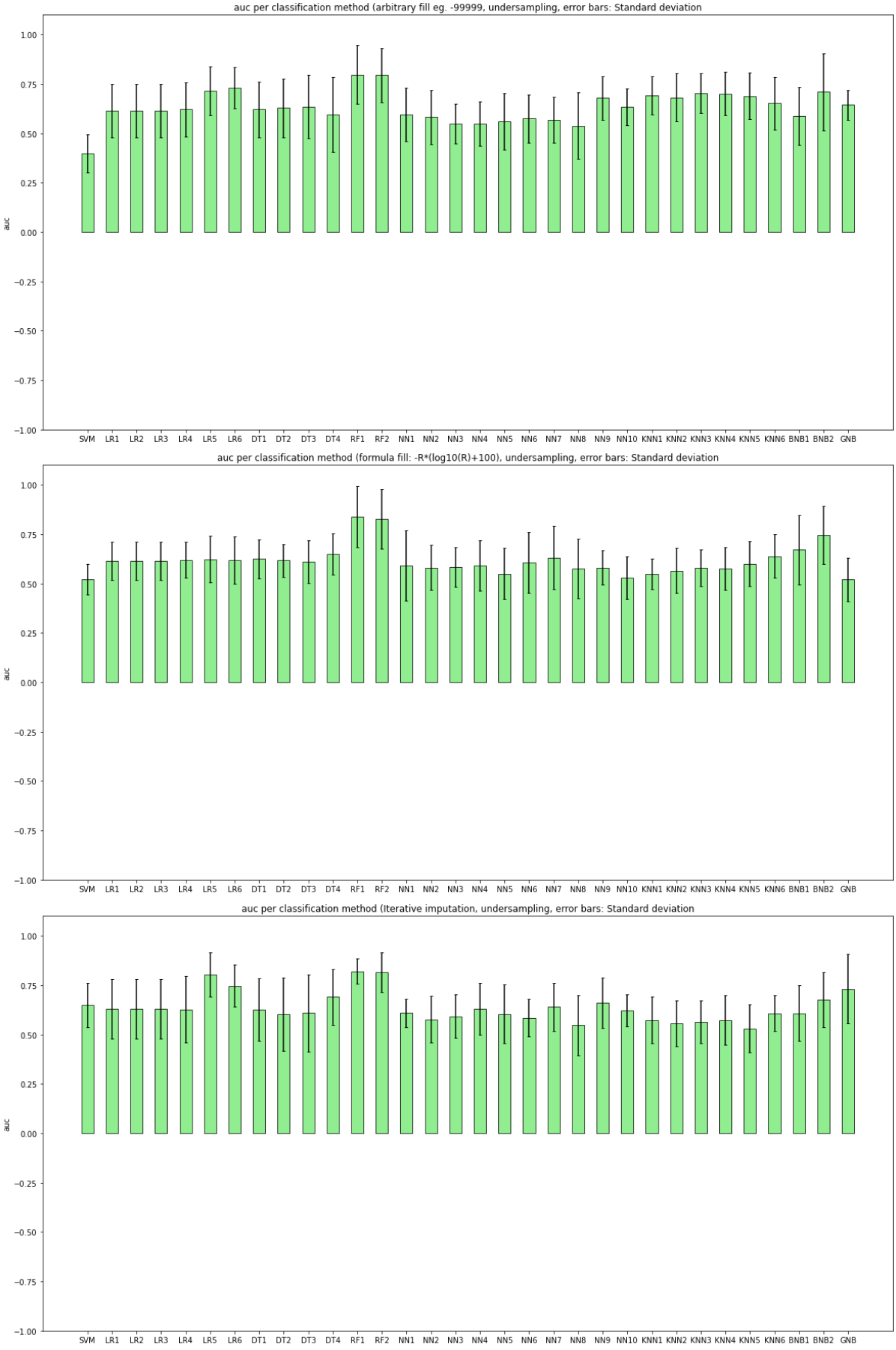

kohen_kappa per classification method (arbitrary fill eg. -99999, SMOTE-oversampling, error bars: Standard deviation
kohen_kappa per classification method (formula fill: -R*(log10(R)+100), SMOTE-oversampling, error bars: Standard deviation
kohen_kappa per classification method (Iterative imputation, SMOTE-oversampling, error bars: Standard deviation
kohen_kappa
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

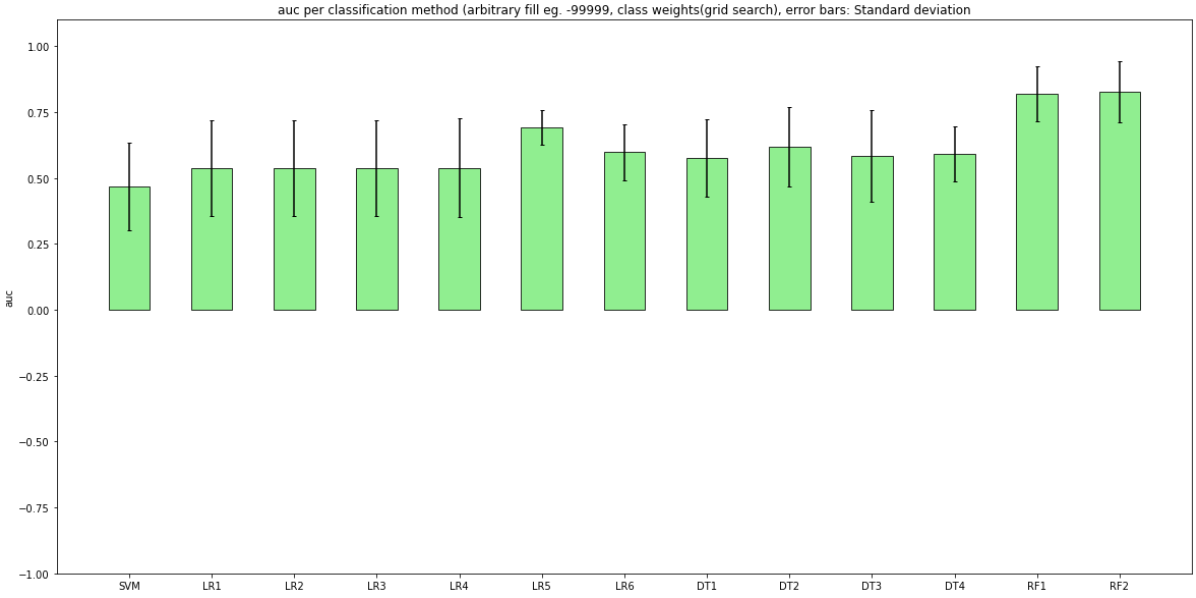

kohen_kappa per classification method (arbitrary fill eg. -99999, random-oversampling, error bars: Standard deviation
kohen_kappa
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

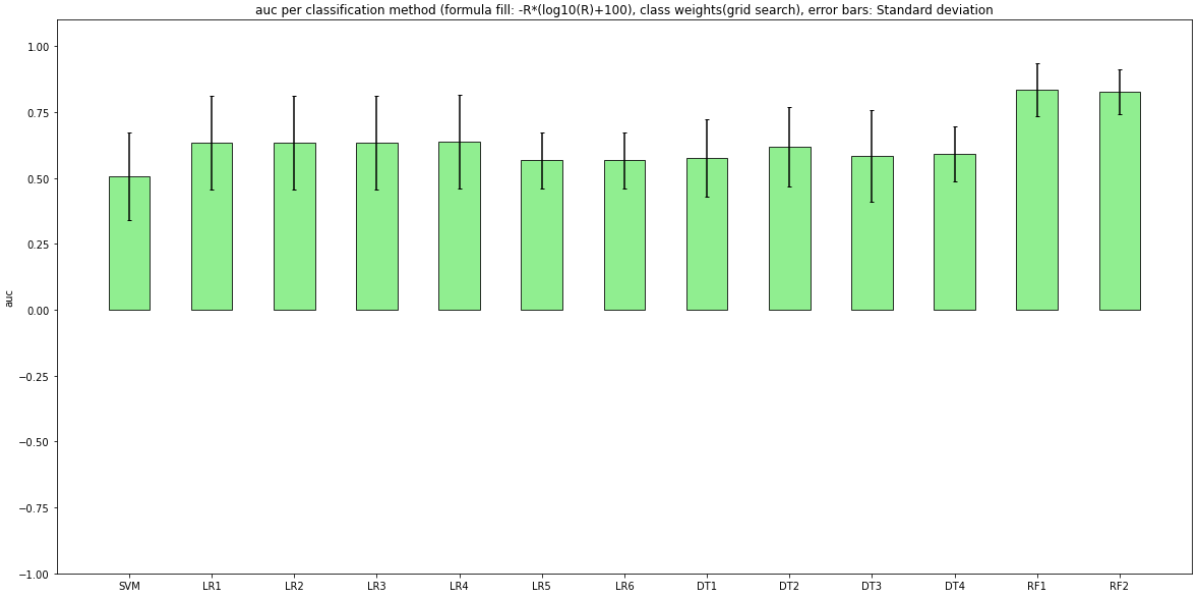

kohen_kappa per classification method (formula fill: -R*(log10(R)+100), random-oversampling, error bars: Standard deviation
kohen_kappa
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

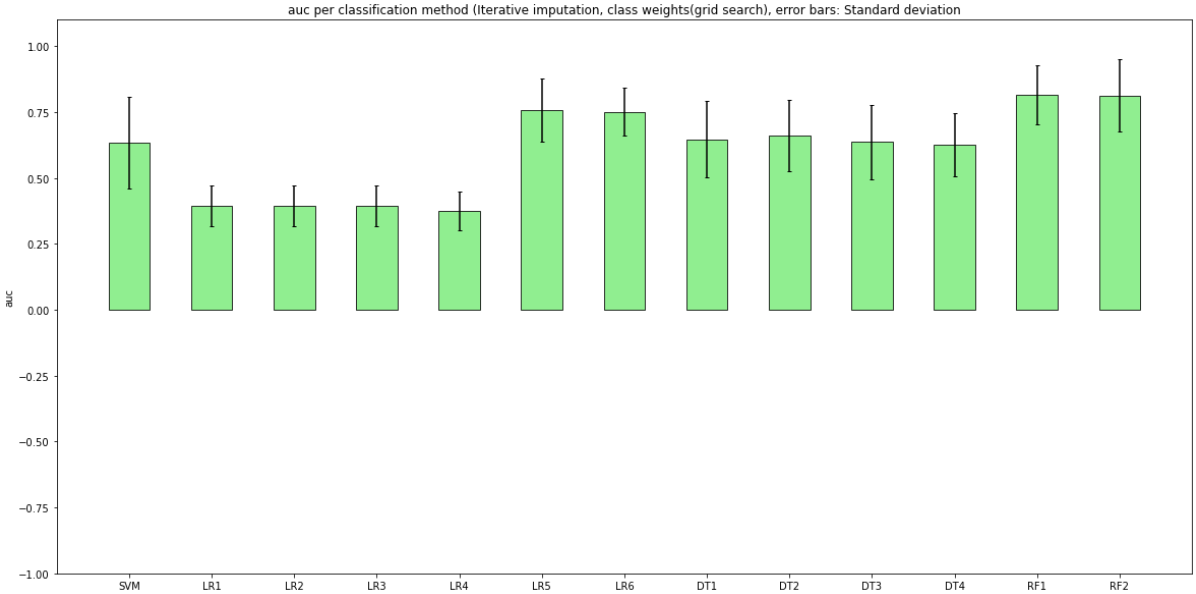

kohen_kappa per classification method (Iterative imputation, random-oversampling, error bars: Standard deviation
kohen_kappa
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

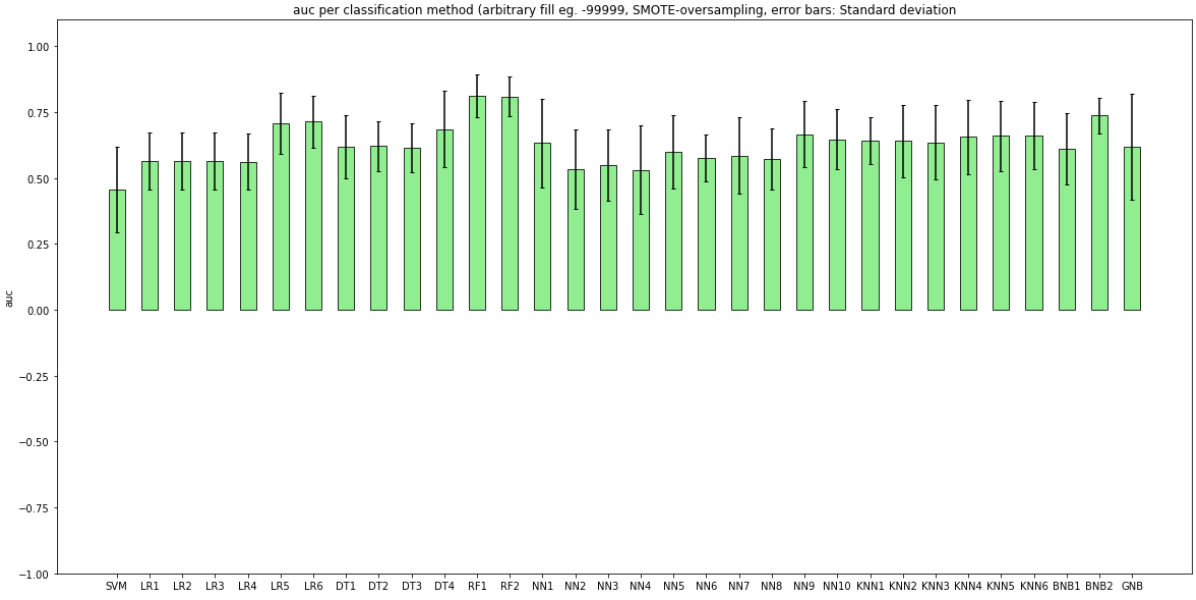

auc per classification method (arbitrary fill eg. -99999, no undersampling, error bars: Standard deviation
auc per classification method (formula fill: -R*(log10(R)+100), no undersampling, error bars: Standard deviation
auc per classification method (Iterative imputation, no undersampling, error bars: Standard deviation
auc
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

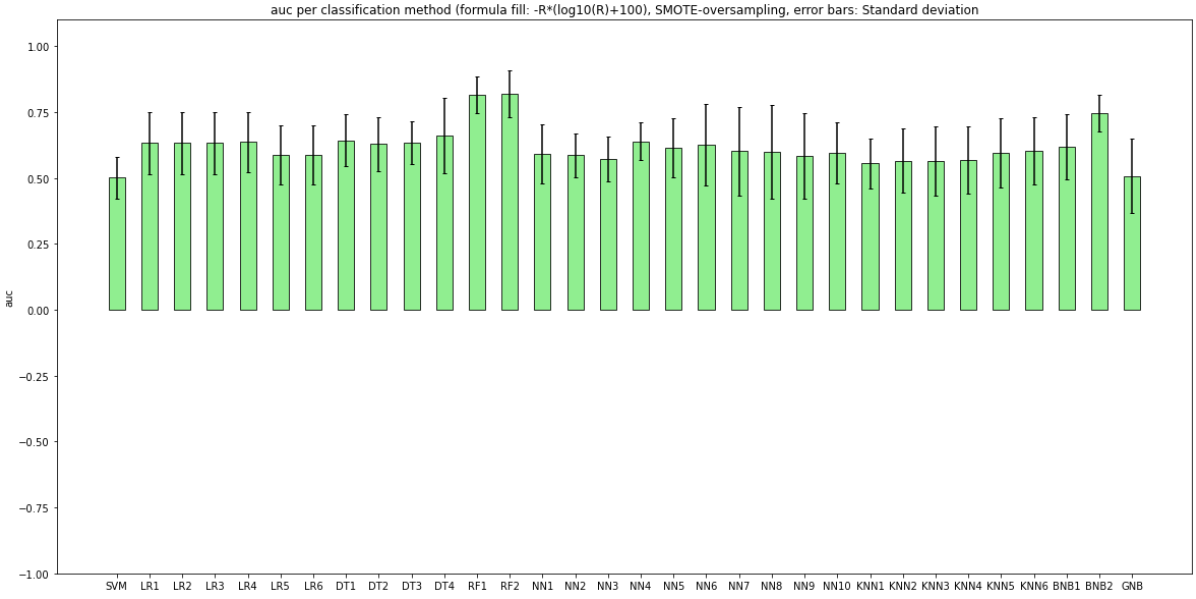

auc per classification method (arbitrary fill eg. -99999, undersampling, error bars: Standard deviation
auc per classification method (formula fill: -R*(log10(R)+100), undersampling, error bars: Standard deviation
auc per classification method (Iterative imputation, undersampling, error bars: Standard deviation
auc
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

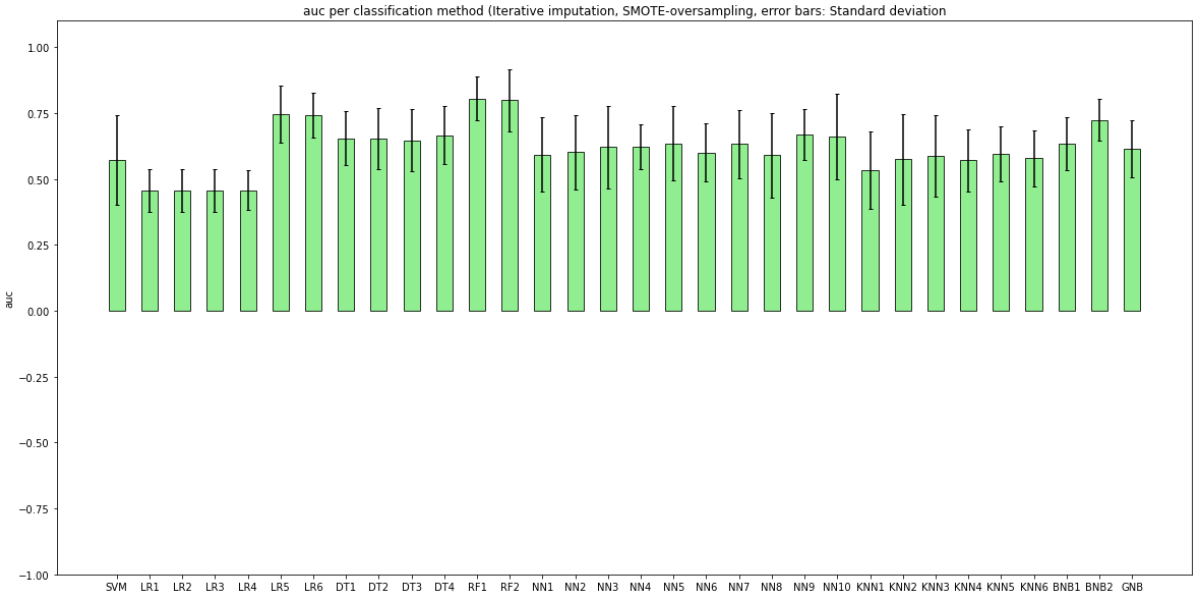

auc per classification method (arbitrary fill eg. -99999, class weights(grid search), error bars: Standard deviation
auc
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM
LR1
LR2
LR3
LR4
LR5
LR6
DT1
DT2
DT3
DT4
RF1
RF2

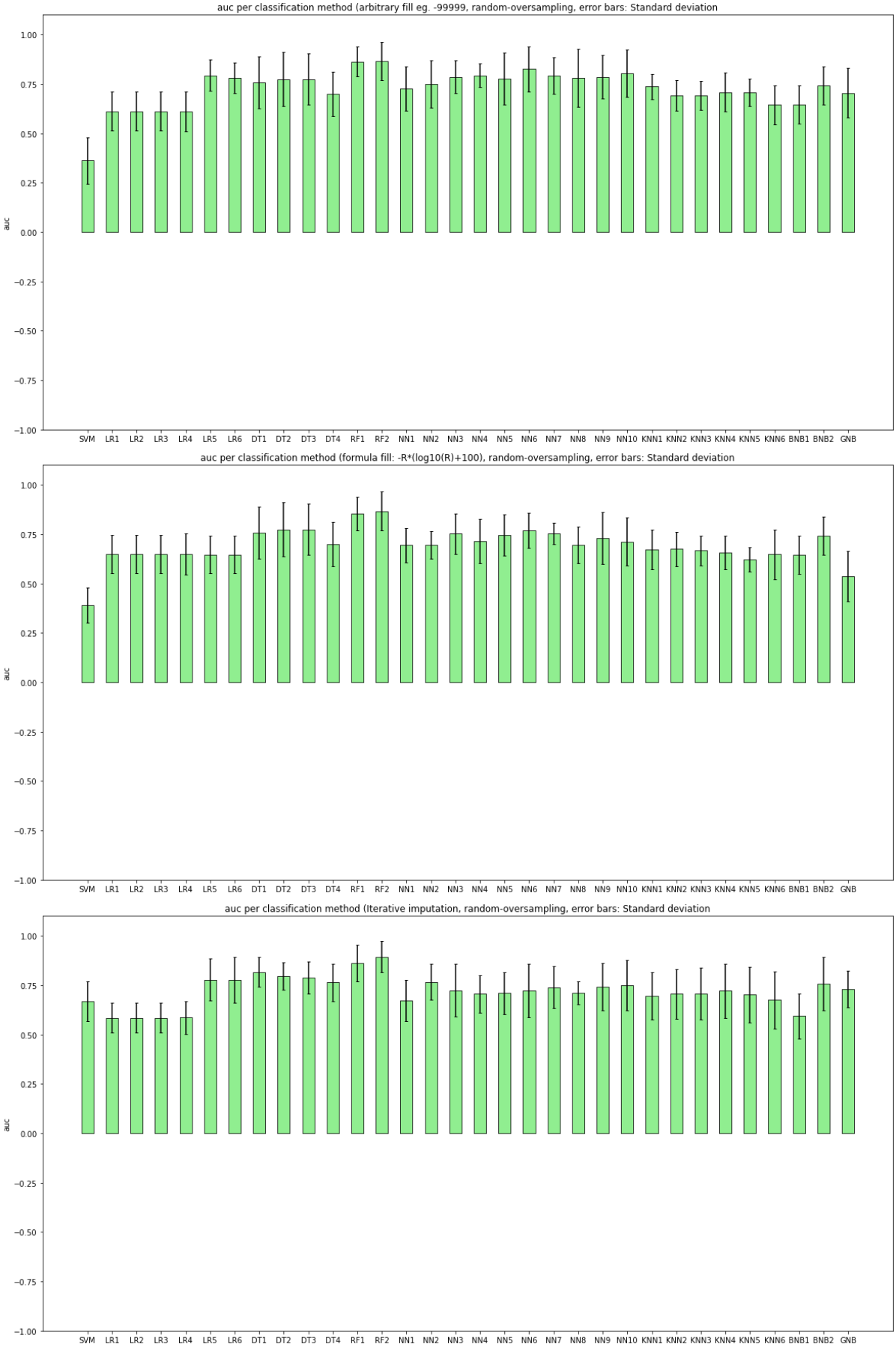

auc per classification method (formula fill: -R*(log10(R)+100), class weights(grid search), error bars: Standard deviation
auc
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM
LR1
LR2
LR3
LR4
LR5
LR6
DT1
DT2
DT3
DT4
RF1
RF2


auc per classification method (Iterative imputation, class weights(grid search), error bars: Standard deviation
auc
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM
LR1
LR2
LR3
LR4
LR5
LR6
DT1
DT2
DT3
DT4
RF1
RF2

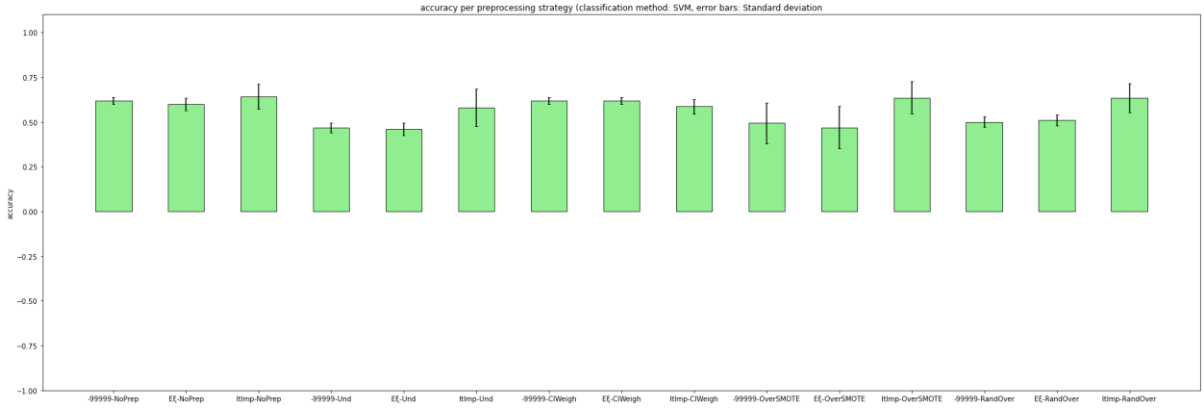

auc per classification method (arbitrary fill eg. -99999, SMOTE-oversampling, error bars: Standard deviation
auc
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

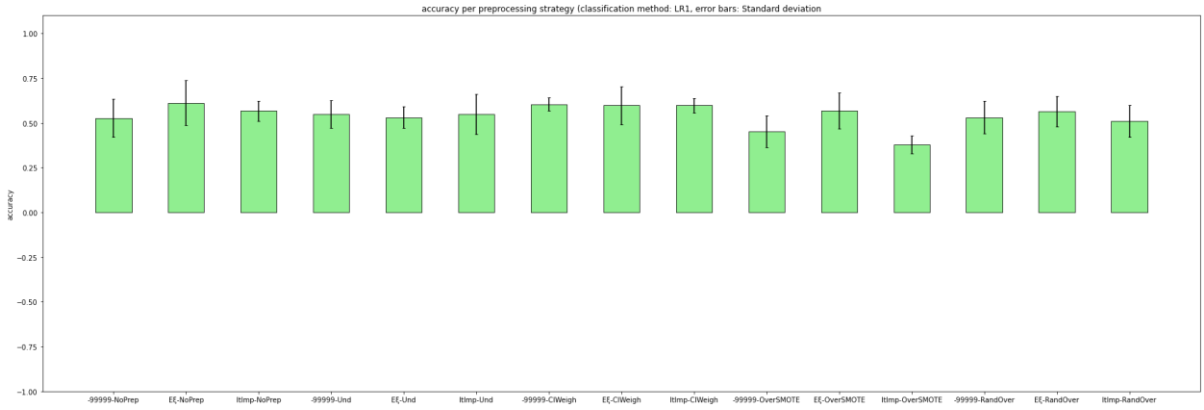

auc per classification method (formula fill: -R*(log10(R)+100), SMOTE-oversampling, error bars: Standard deviation
auc
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

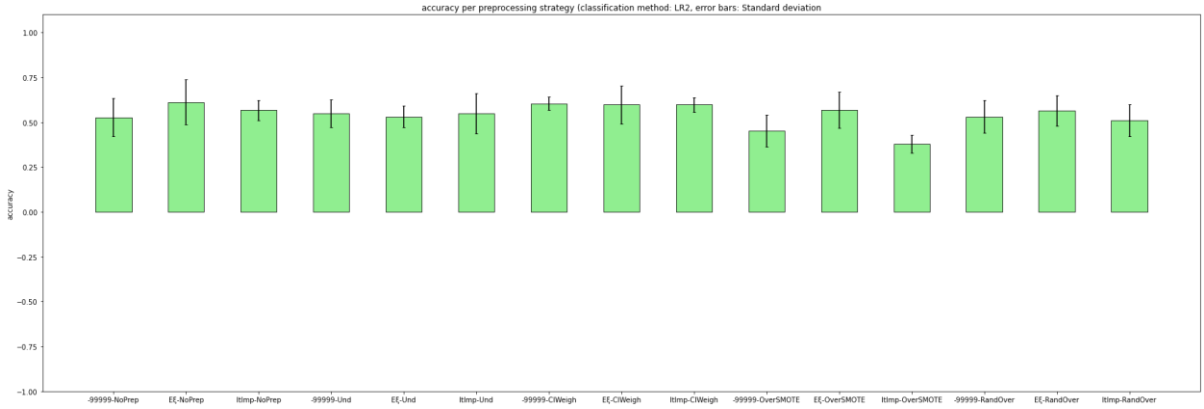

auc per classification method (Iterative imputation, SMOTE-oversampling, error bars: Standard deviation
auc
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

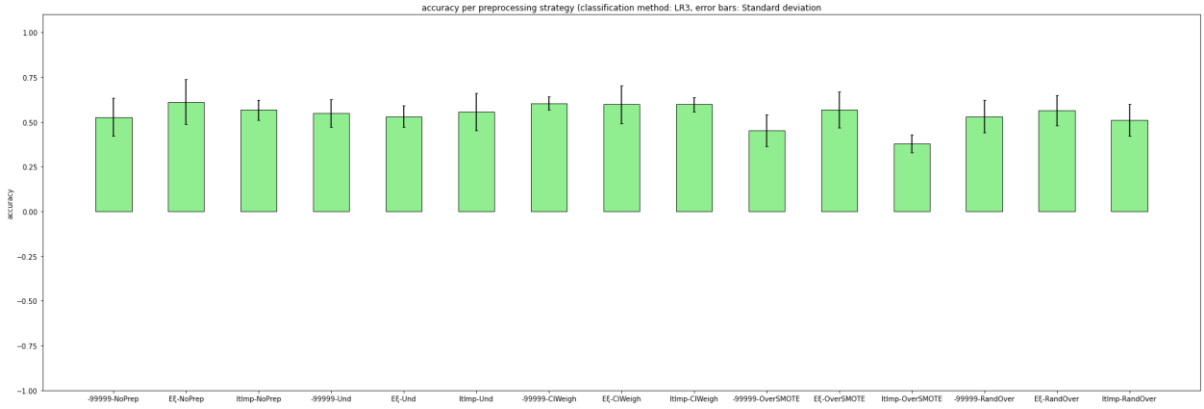

auc per classification method (arbitrary fill eg. -99999, random-oversampling, error bars: Standard deviation
auc per classification method (formula fill: -R*(log10(R)+100), random-oversampling, error bars: Standard deviation
auc per classification method (Iterative imputation, random-oversampling, error bars: Standard deviation
auc
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
SVM LR1 LR2 LR3 LR4 LR5 LR6 DT1 DT2 DT3 DT4 RF1 RF2 NN1 NN2 NN3 NN4 NN5 NN6 NN7 NN8 NN9 NN10 KNN1 KNN2 KNN3 KNN4 KNN5 KNN6 BNB1 BNB2 GNB

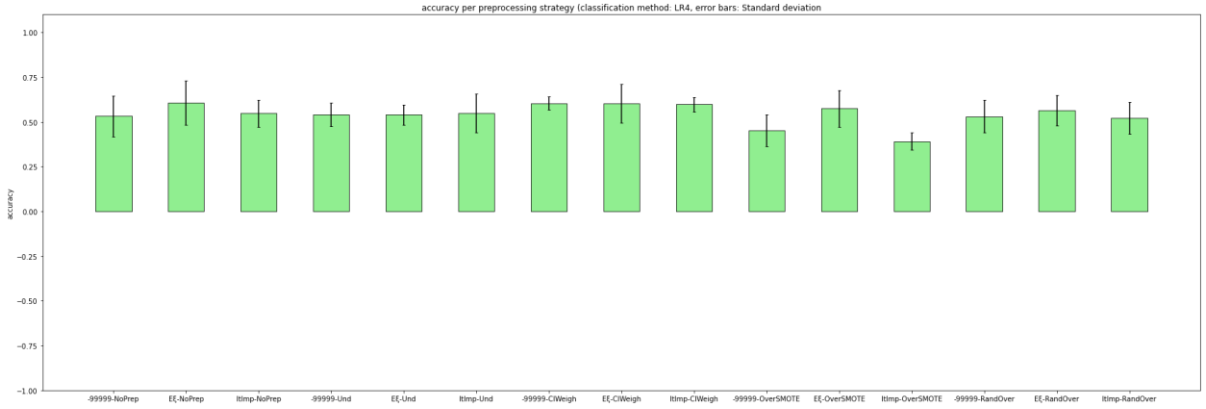
10.4.2 Παράρτημα IVb

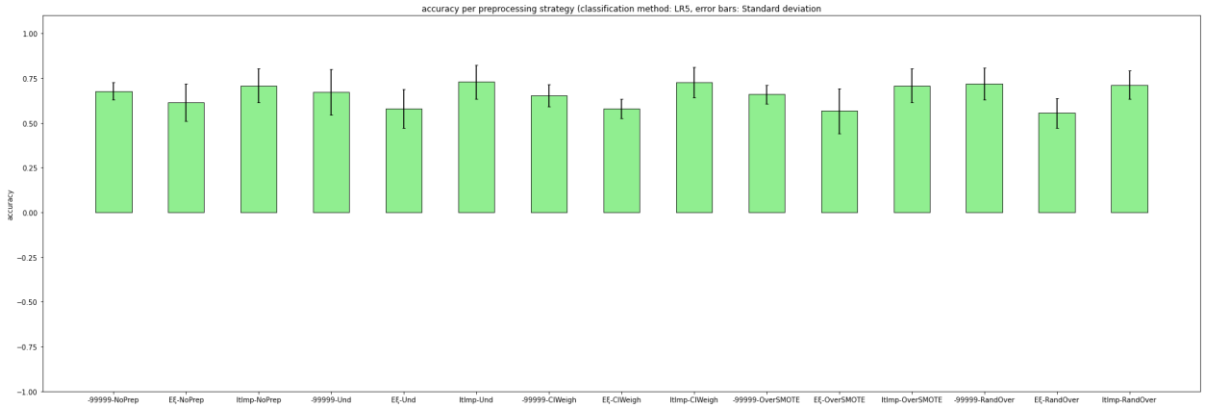


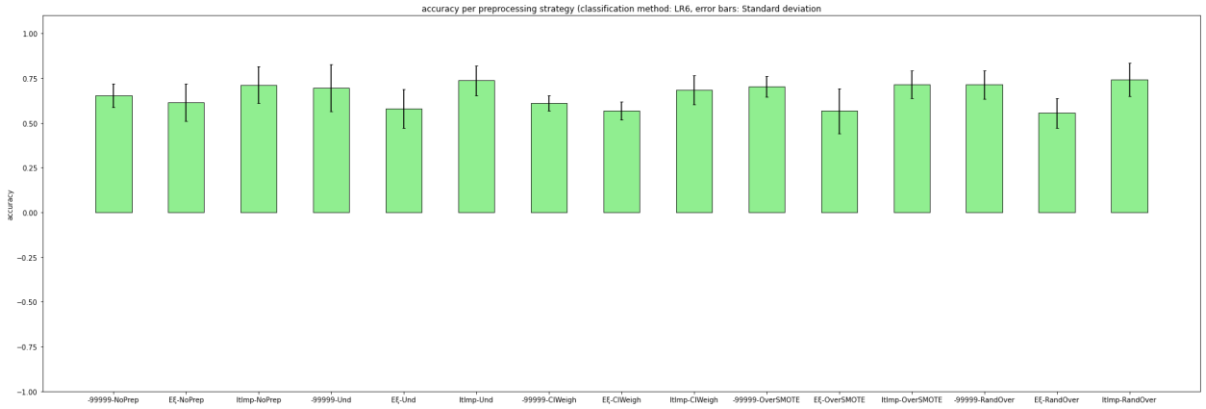


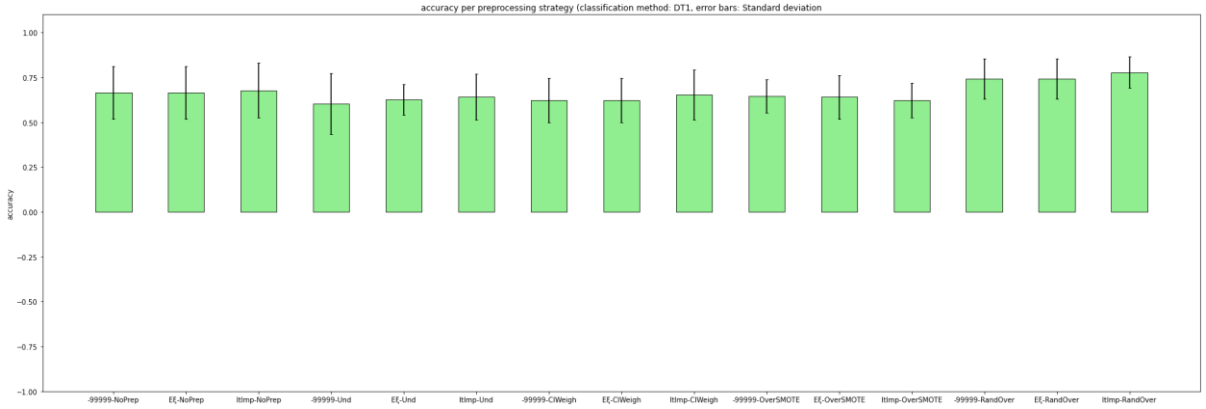


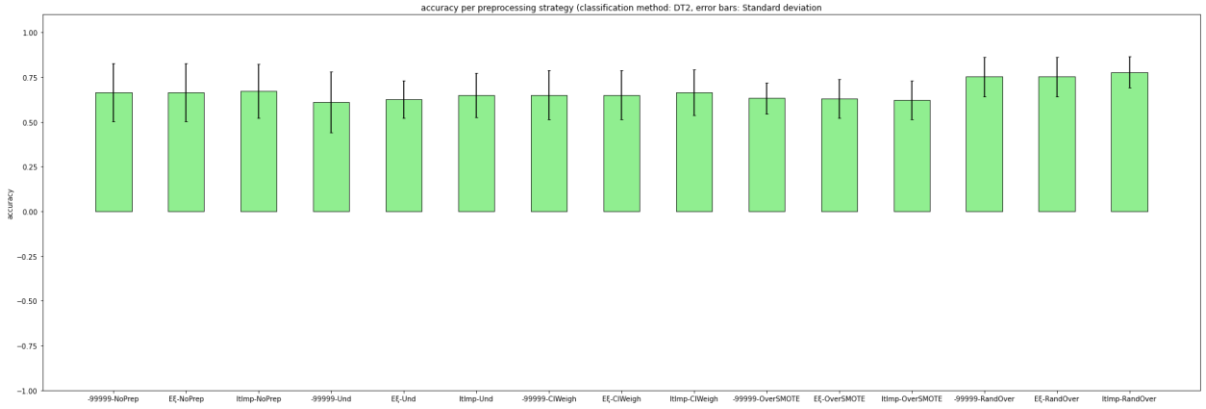


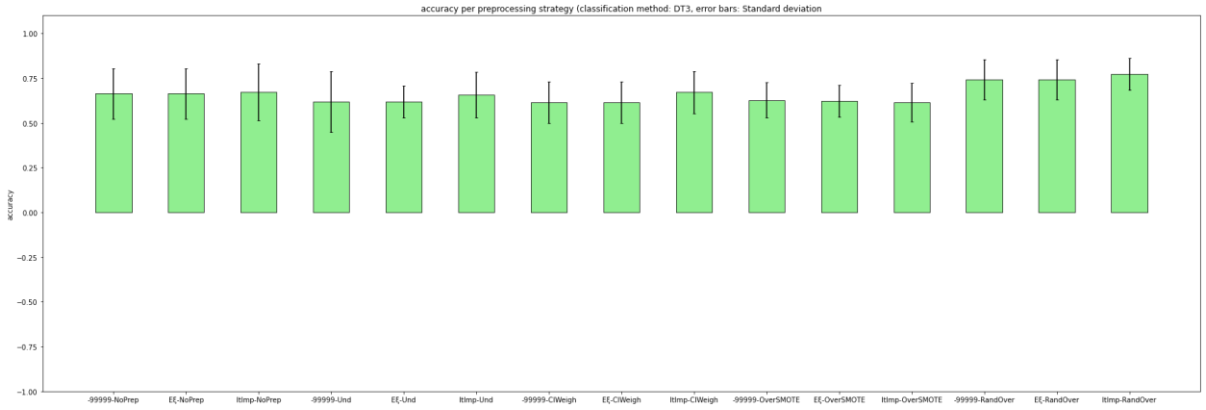

accuracy per preprocessing strategy (classification method: LR4, error bars: Standard deviation
accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

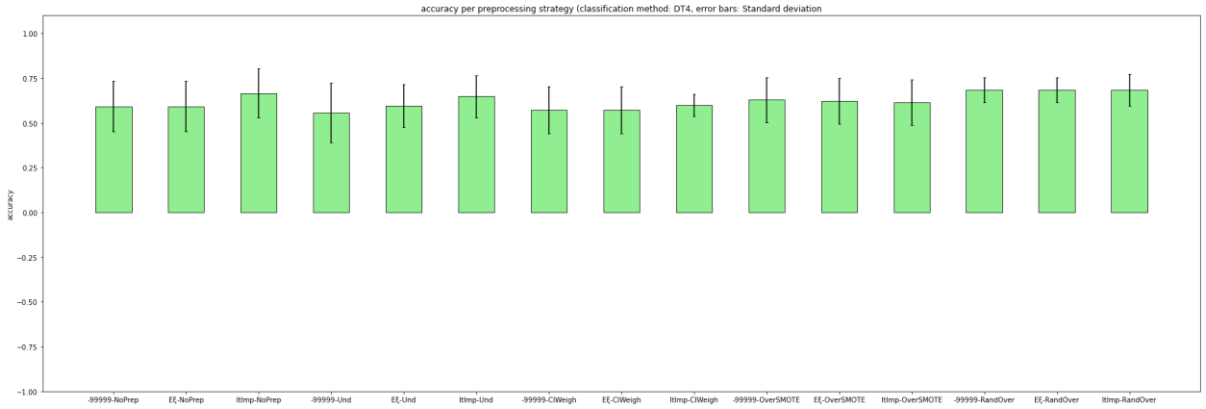

accuracy per preprocessing strategy (classification method: LR5, error bars: Standard deviation
accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

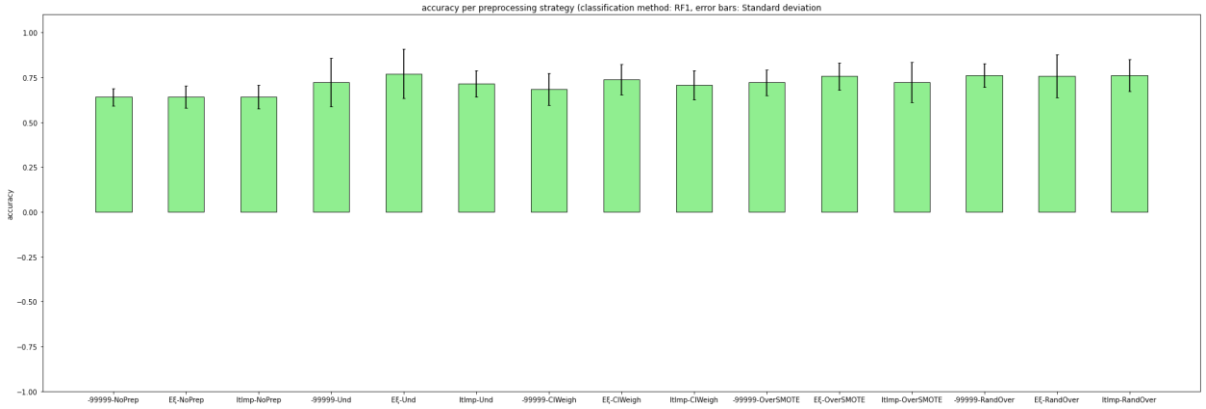

accuracy per preprocessing strategy (classification method: LR6, error bars: Standard deviation
accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

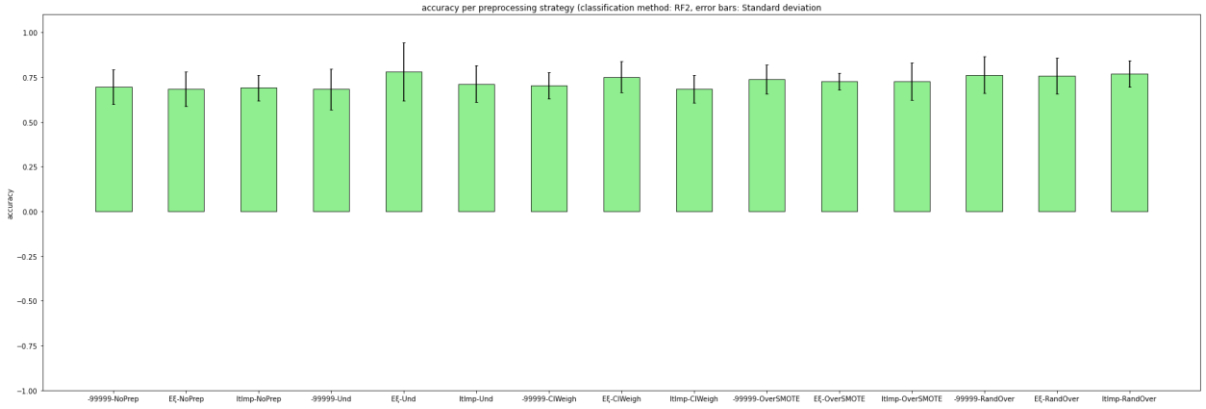

accuracy per preprocessing strategy (classification method: DT1, error bars: Standard deviation
accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

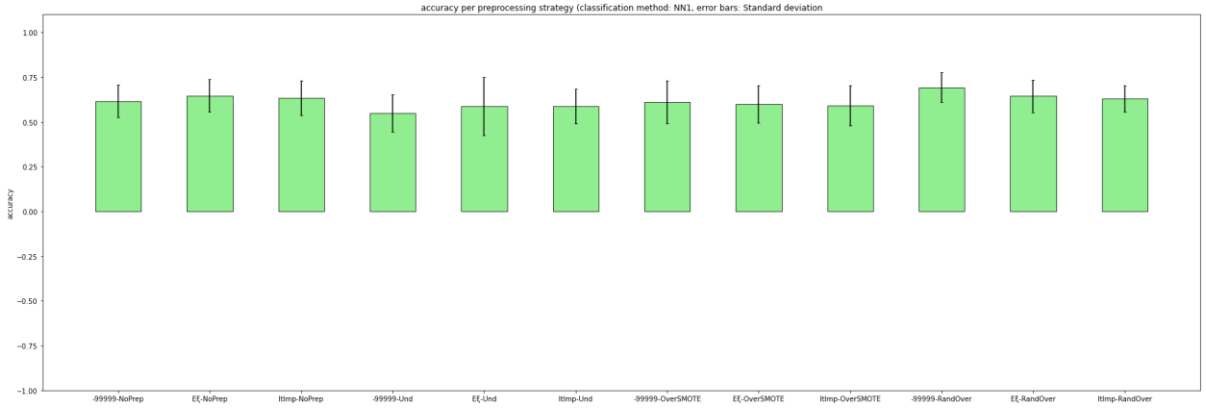

accuracy per preprocessing strategy (classification method: DT2, error bars: Standard deviation
accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

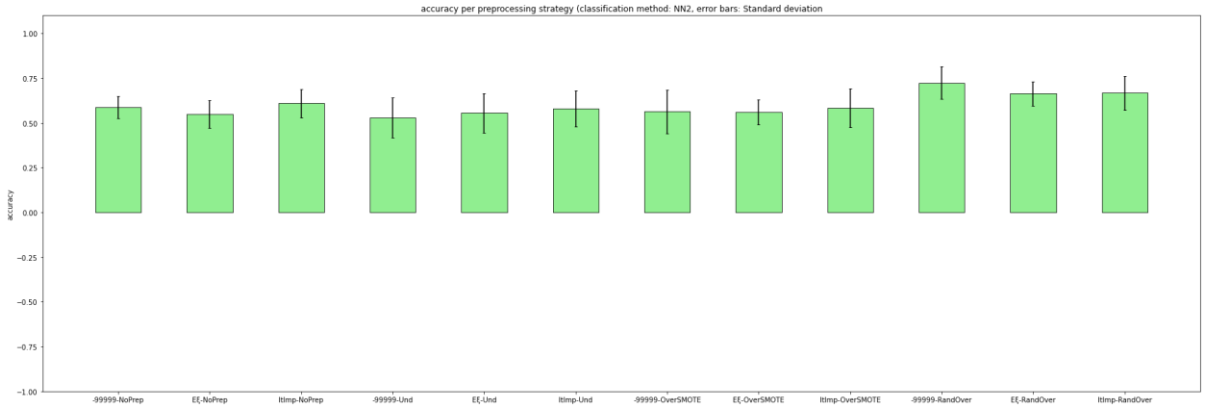

accuracy per preprocessing strategy (classification method: DT3, error bars: Standard deviation
accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

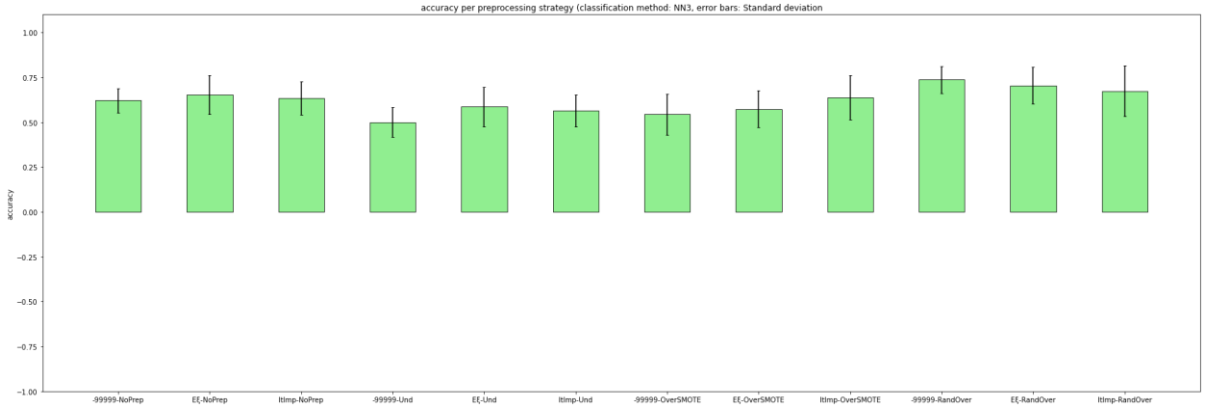

accuracy per preprocessing strategy (classification method: DT4, error bars: Standard deviation
accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

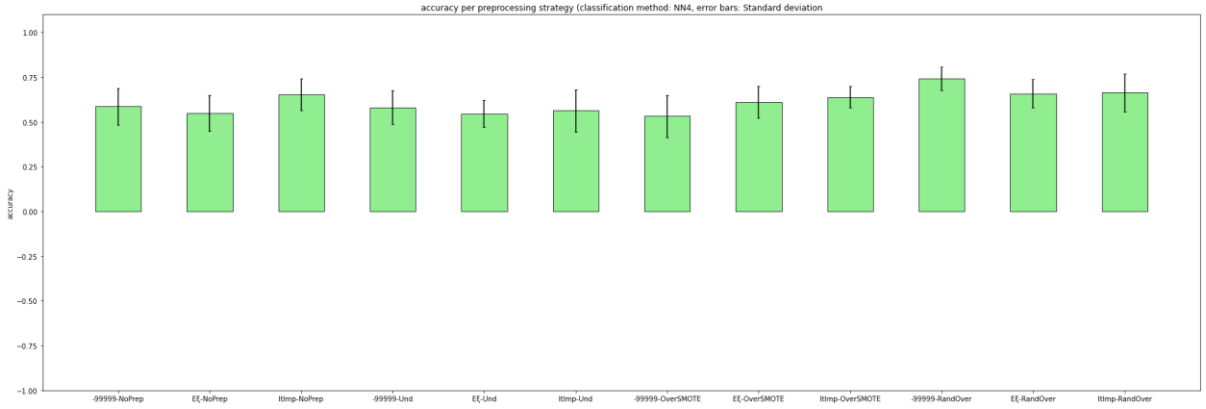

accuracy per preprocessing strategy (classification method: RF1, error bars: Standard deviation
accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

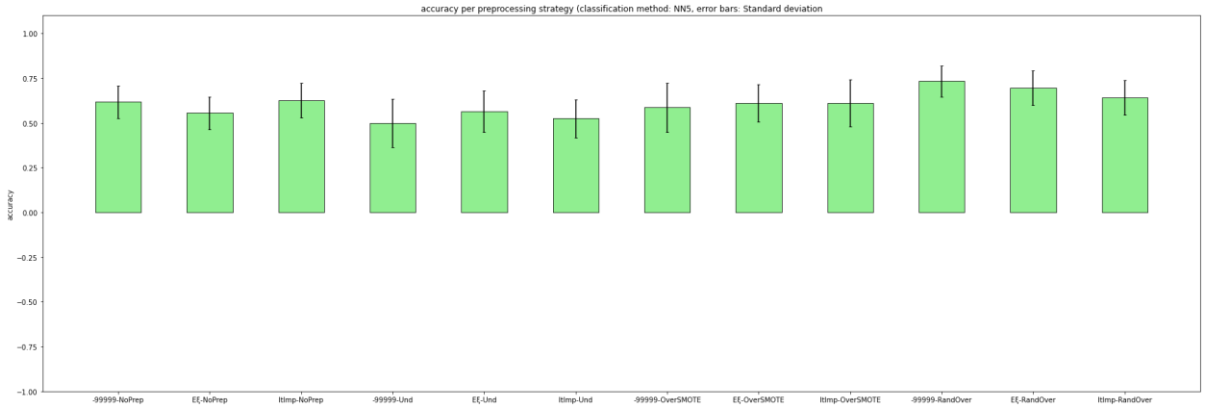

accuracy per preprocessing strategy (classification method: RF2, error bars: Standard deviation
accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

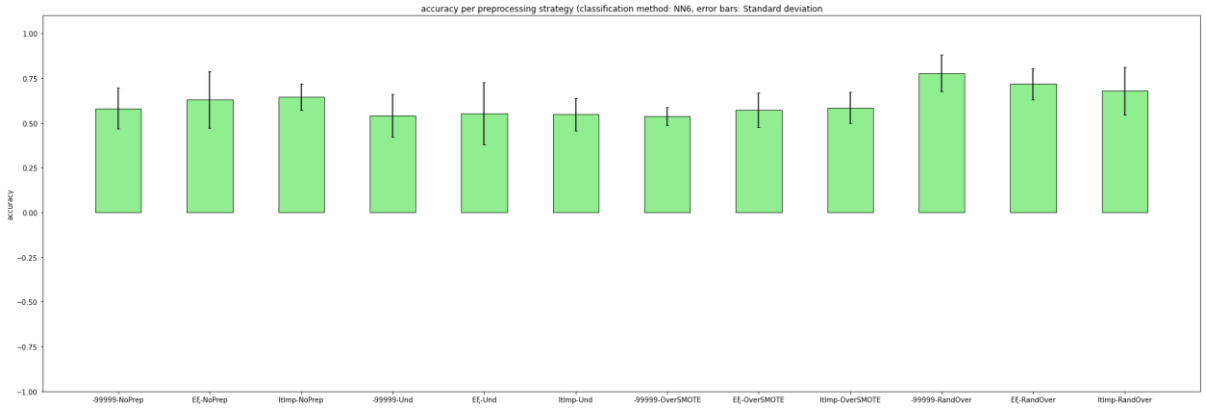

accuracy per preprocessing strategy (classification method: NN1, error bars: Standard deviation
accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

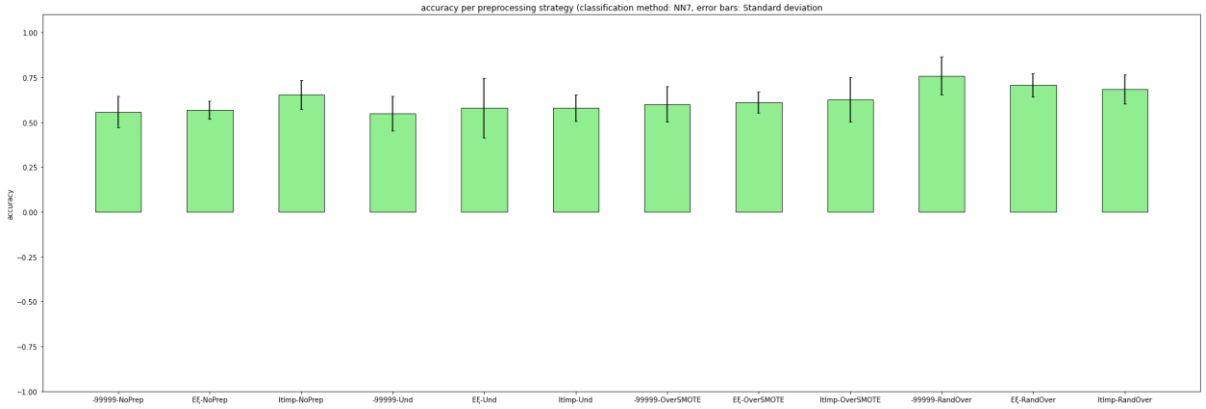

accuracy per preprocessing strategy (classification method: NN2, error bars: Standard deviation
accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

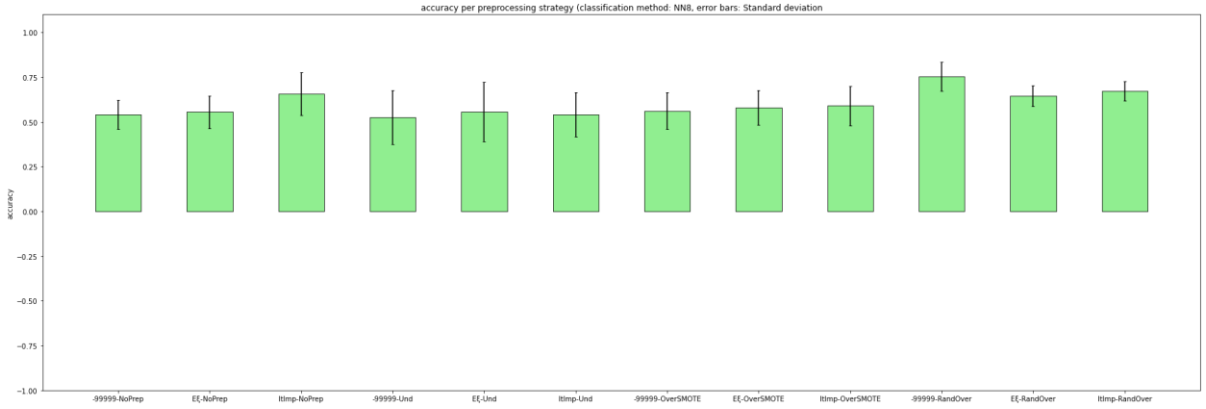

accuracy per preprocessing strategy (classification method: NN3, error bars: Standard deviation
accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

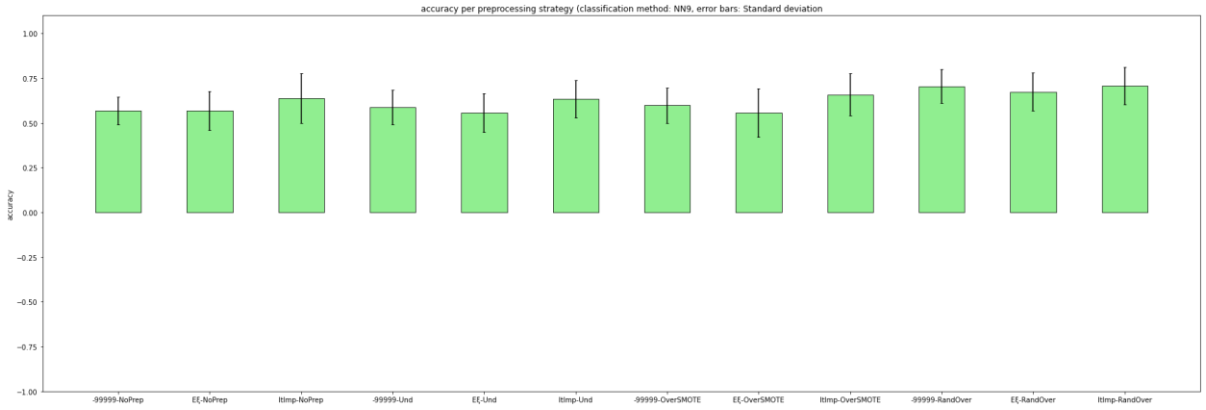

accuracy per preprocessing strategy (classification method: NN4, error bars: Standard deviation
accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

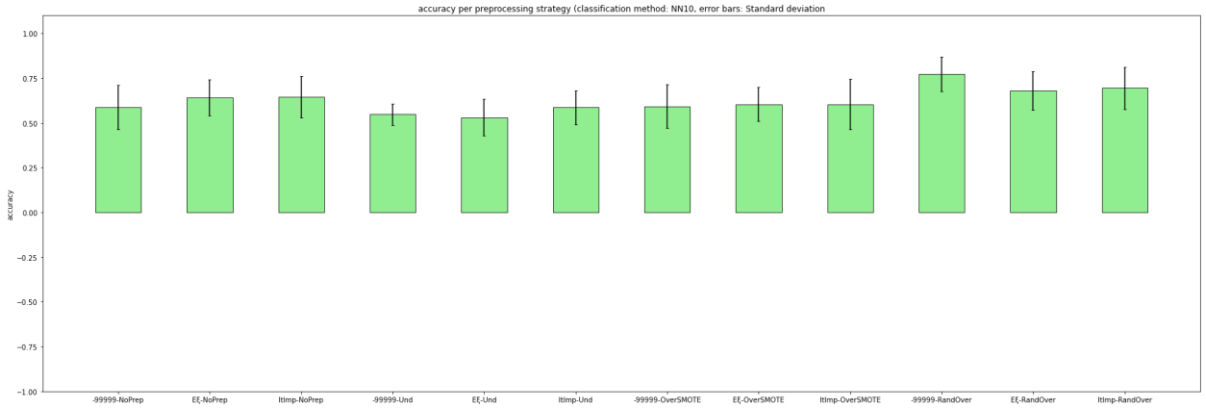

accuracy per preprocessing strategy (classification method: NN5, error bars: Standard deviation
accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

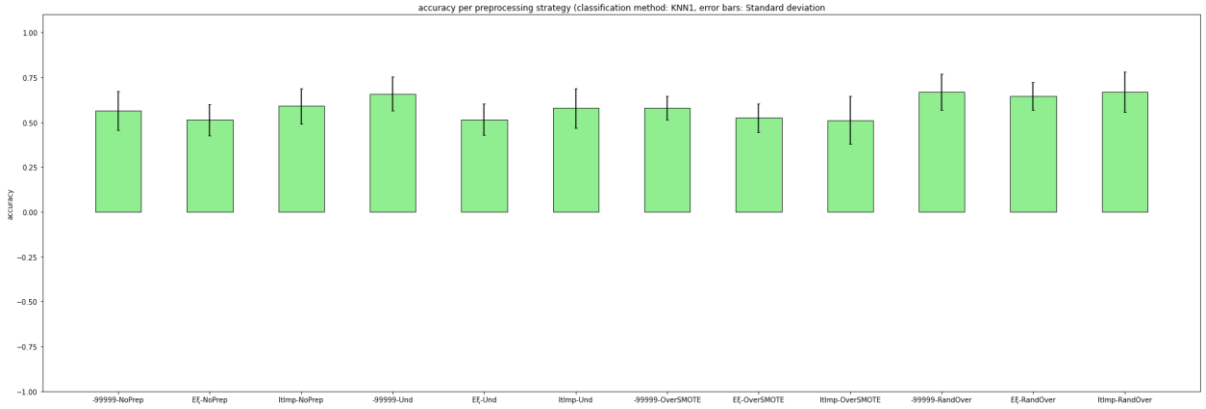

accuracy per preprocessing strategy (classification method: NN6, error bars: Standard deviation
accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

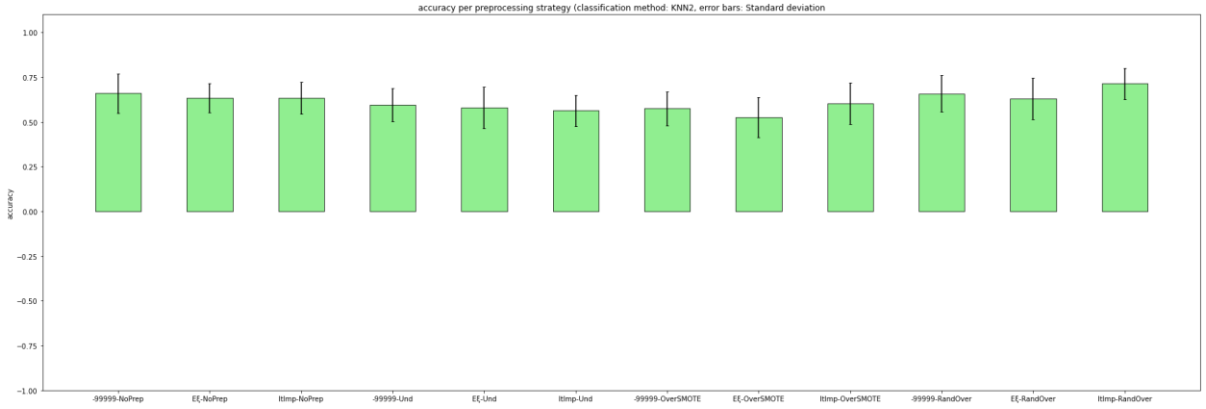

accuracy per preprocessing strategy (classification method: NN7, error bars: Standard deviation
accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

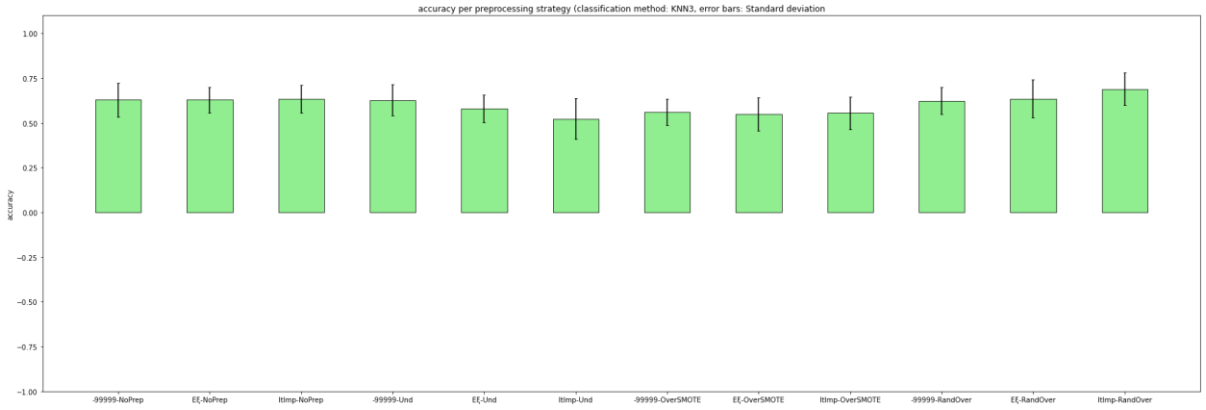

accuracy per preprocessing strategy (classification method: NN8, error bars: Standard deviation
accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

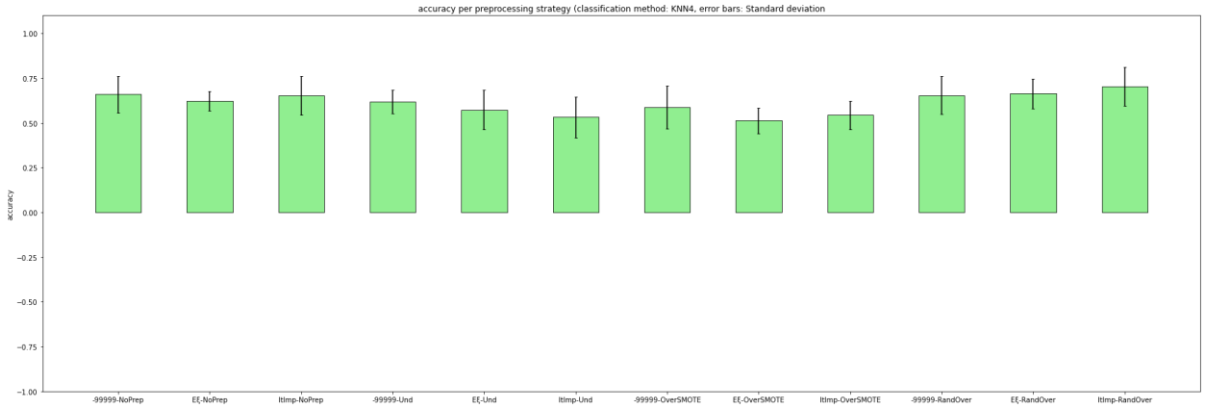

accuracy per preprocessing strategy (classification method: NN9, error bars: Standard deviation
accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

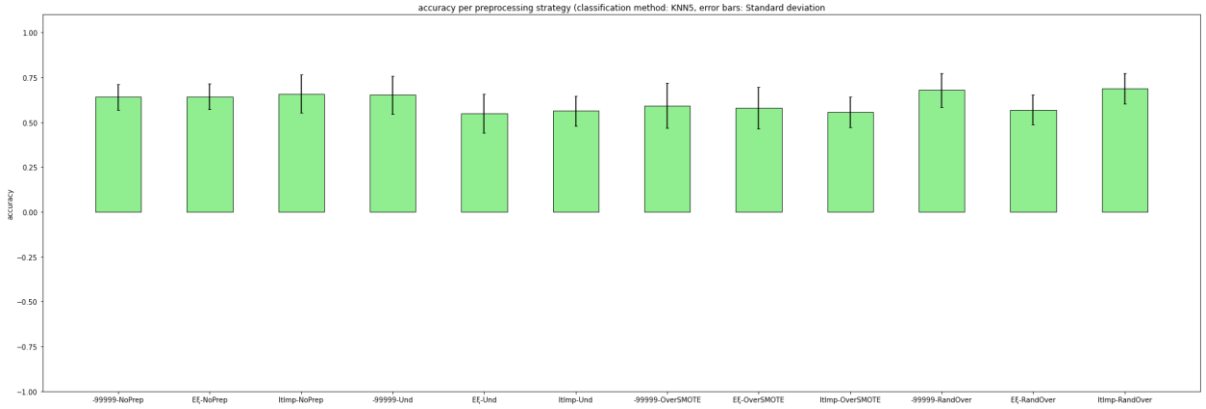

accuracy per preprocessing strategy (classification method: NN10, error bars: Standard deviation
accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

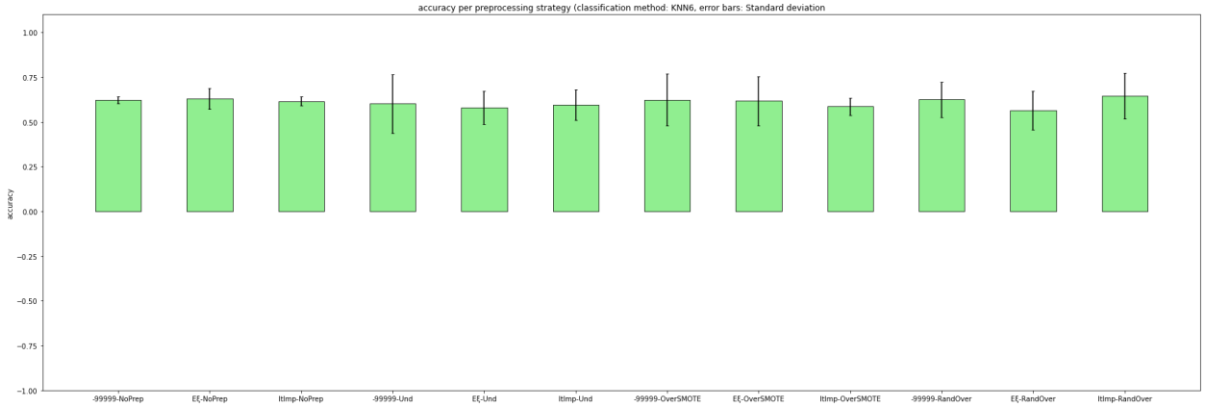

accuracy per preprocessing strategy (classification method: KNN1, error bars: Standard deviation
accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

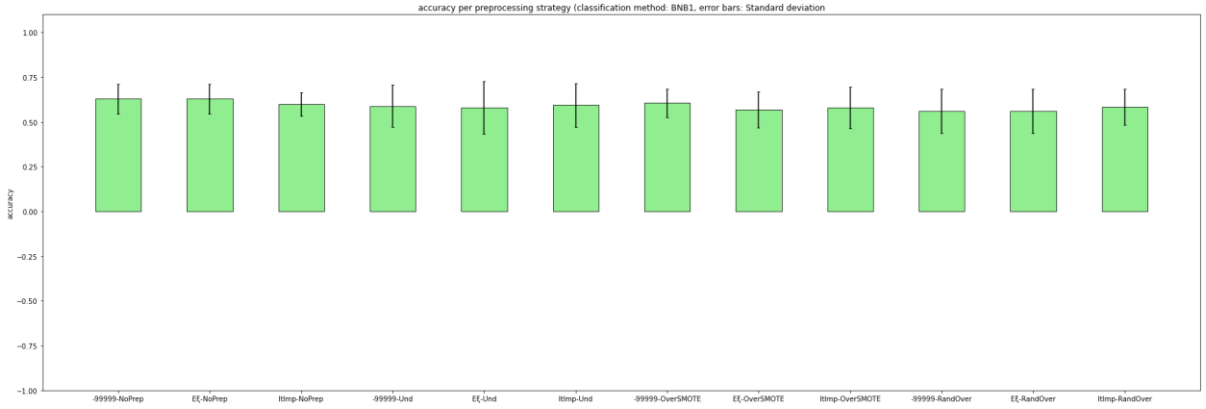

accuracy per preprocessing strategy (classification method: KNN2, error bars: Standard deviation
accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

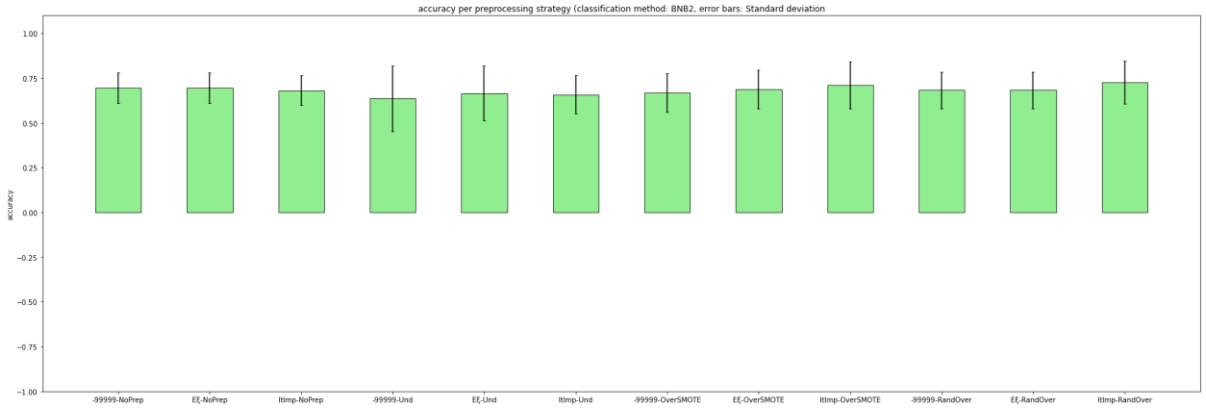

accuracy per preprocessing strategy (classification method: KNN3, error bars: Standard deviation
accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

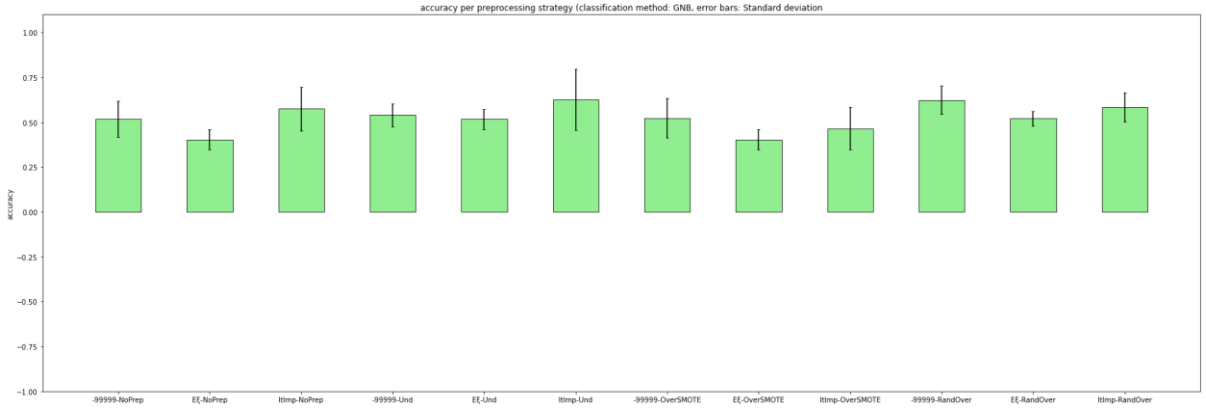

accuracy per preprocessing strategy (classification method: KNN4, error bars: Standard deviation
accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

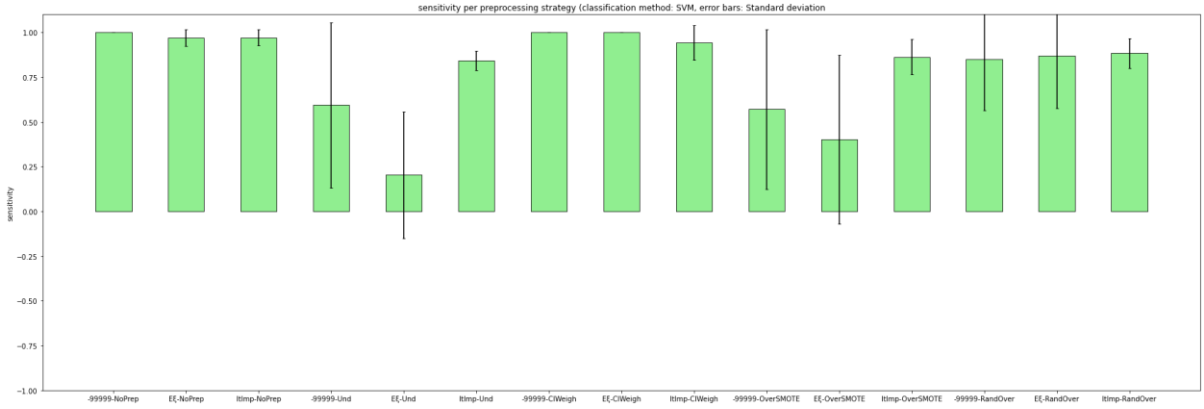

accuracy per preprocessing strategy (classification method: KNN5, error bars: Standard deviation
accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

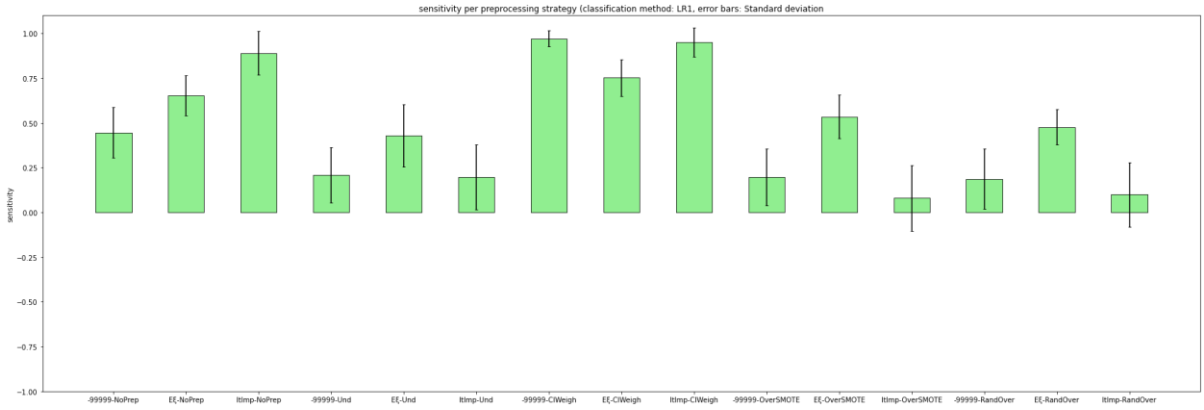

accuracy per preprocessing strategy (classification method: KNN6, error bars: Standard deviation
accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

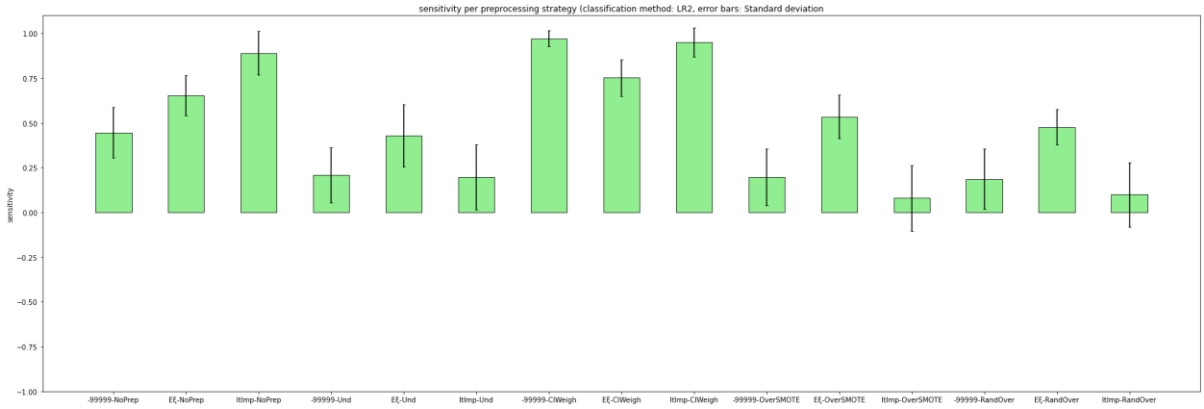

accuracy per preprocessing strategy (classification method: BNB1, error bars: Standard deviation
accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

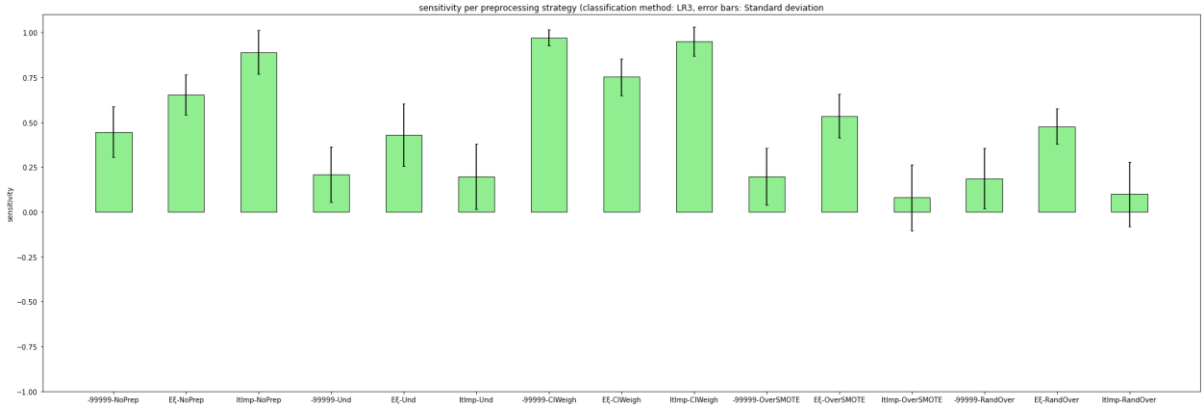

accuracy per preprocessing strategy (classification method: BNB2, error bars: Standard deviation
accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

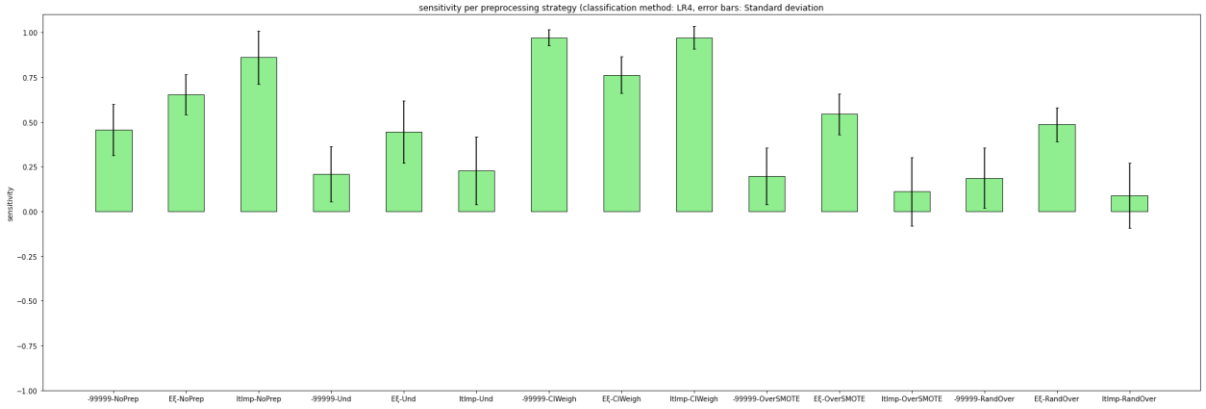

accuracy per preprocessing strategy (classification method: GNB, error bars: Standard deviation
accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

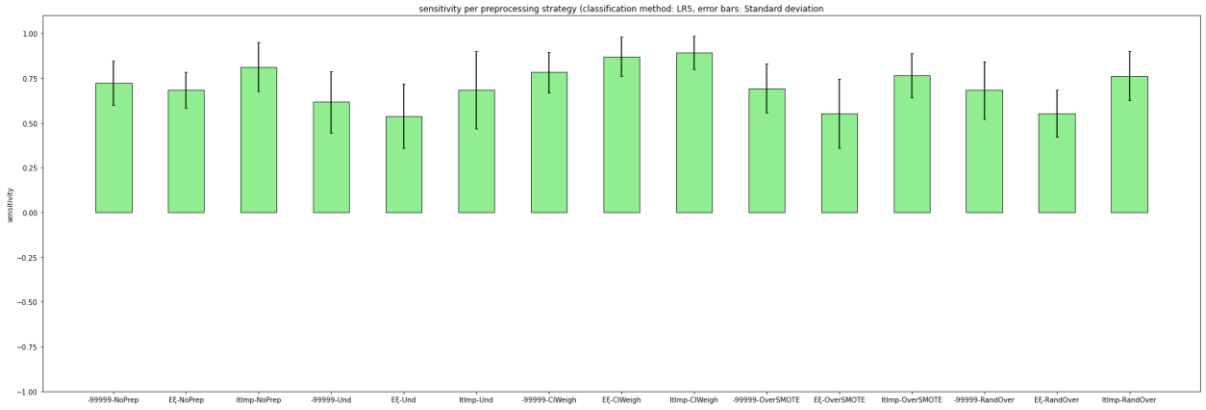

sensitivity per preprocessing strategy (classification method: SVM, error bars: Standard deviation
sensitivity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

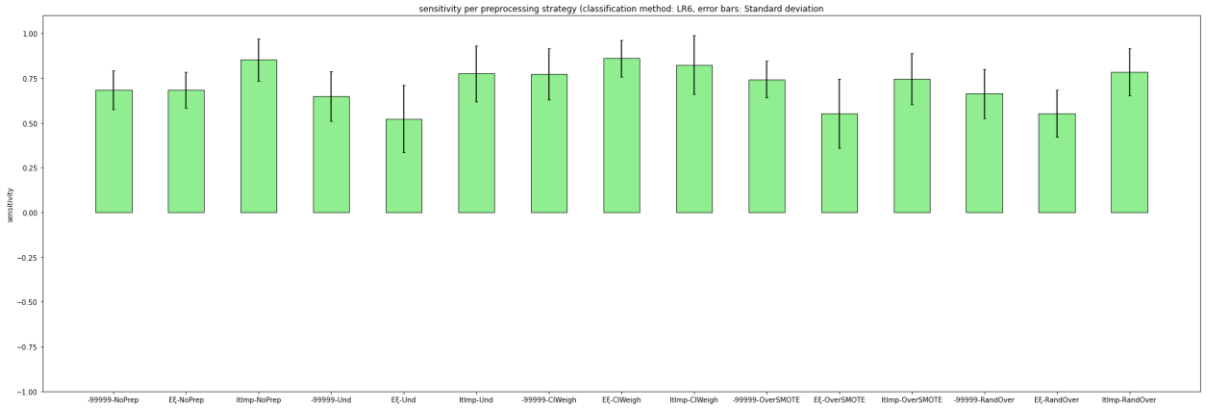

sensitivity per preprocessing strategy (classification method: LR1, error bars: Standard deviation
sensitivity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

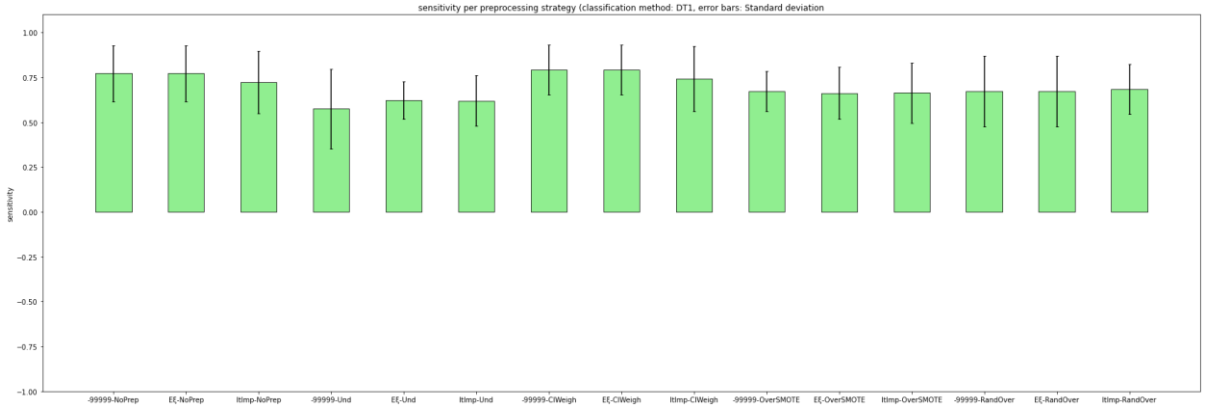

sensitivity per preprocessing strategy (classification method: LR2, error bars: Standard deviation
sensitivity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

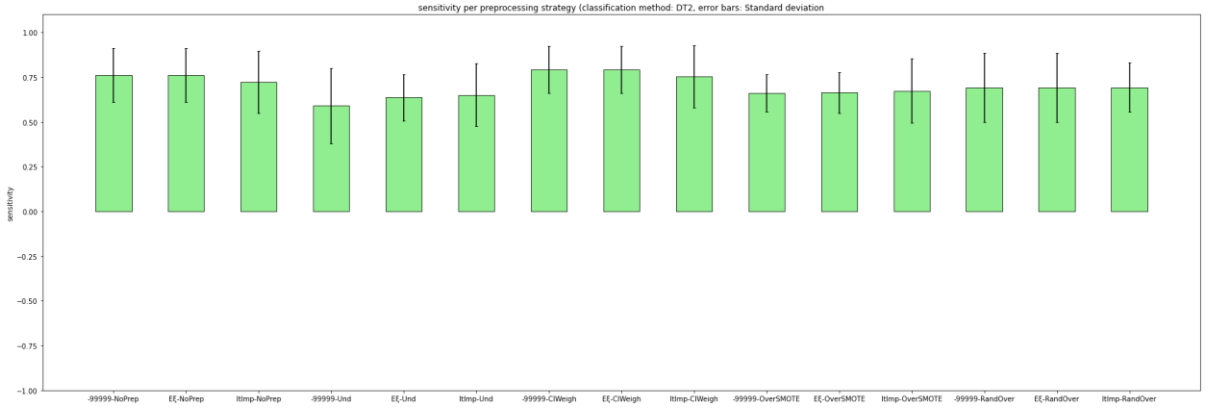

sensitivity per preprocessing strategy (classification method: LR3, error bars: Standard deviation
sensitivity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

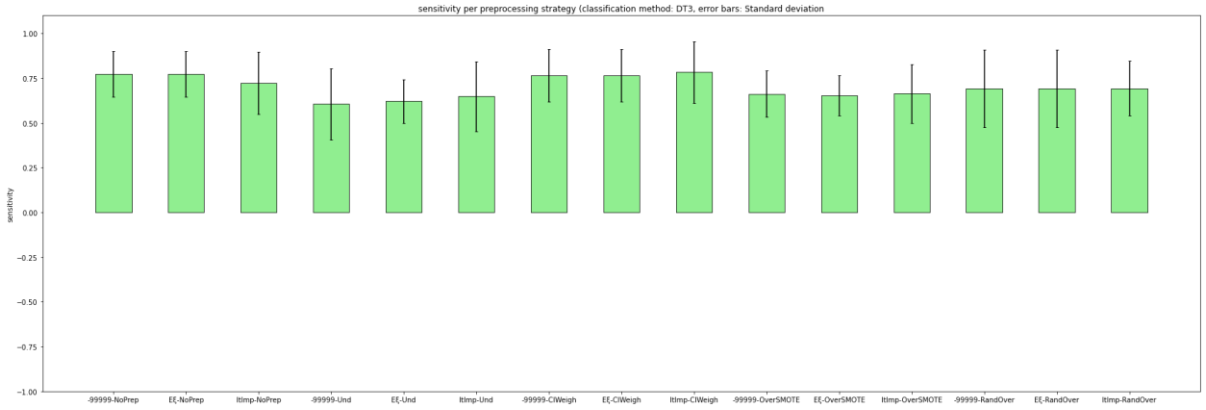

sensitivity per preprocessing strategy (classification method: LR4, error bars: Standard deviation
sensitivity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

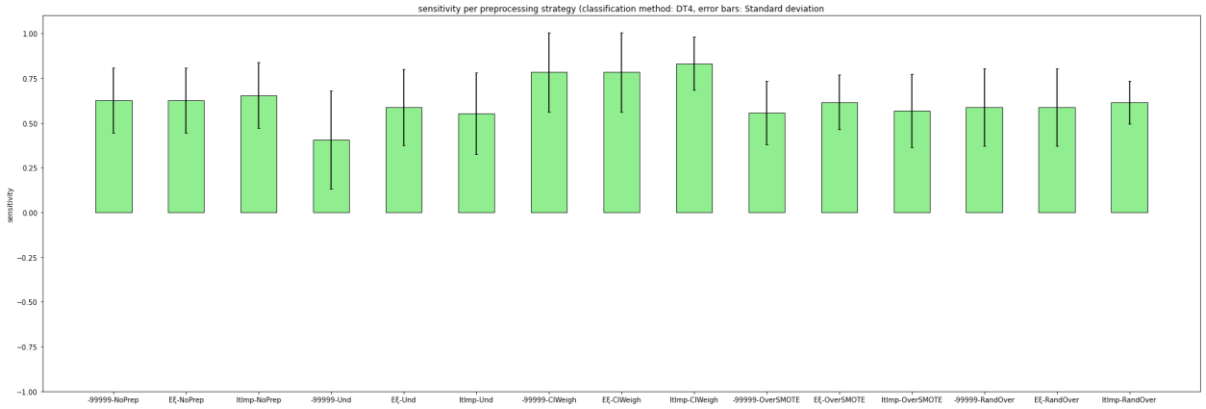

sensitivity per preprocessing strategy (classification method: LR5, error bars: Standard deviation
sensitivity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

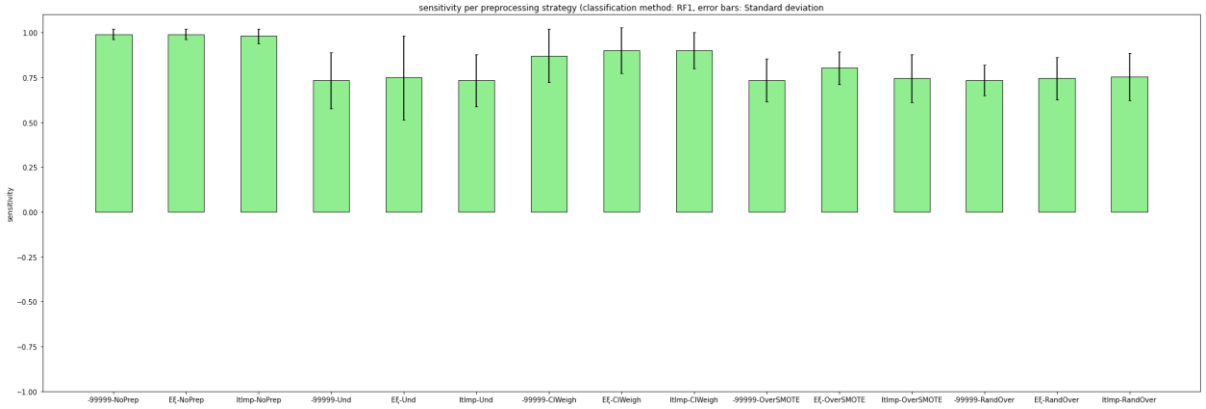

sensitivity per preprocessing strategy (classification method: LR6, error bars: Standard deviation
sensitivity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

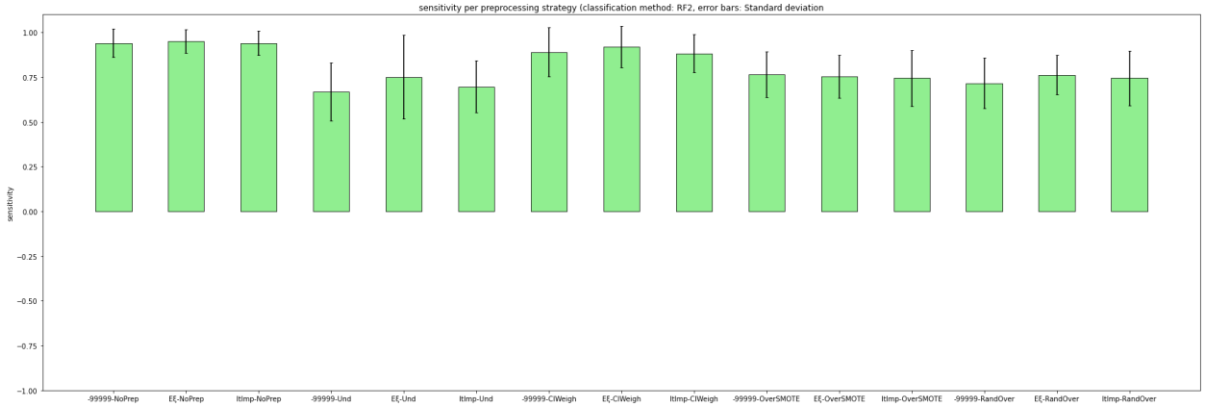

sensitivity per preprocessing strategy (classification method: DT1, error bars: Standard deviation
sensitivity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

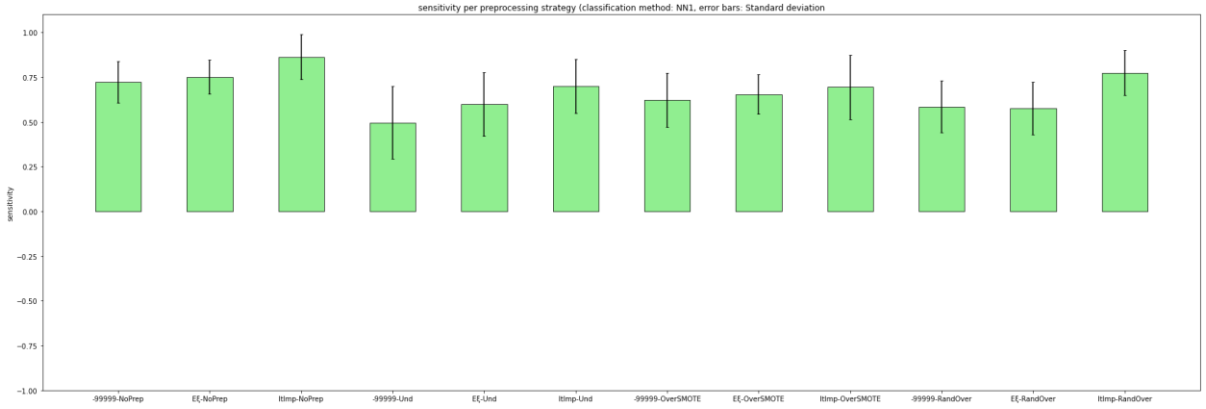

sensitivity per preprocessing strategy (classification method: DT2, error bars: Standard deviation
sensitivity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

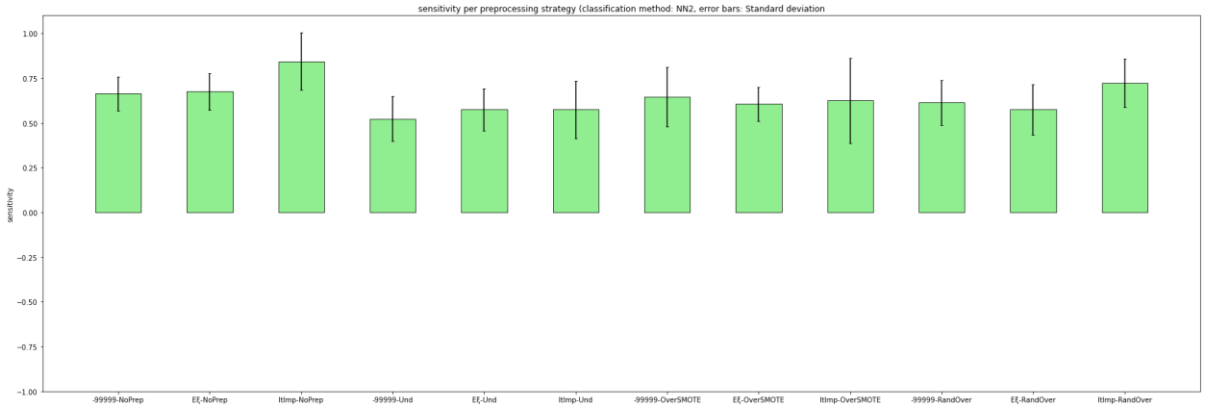

sensitivity per preprocessing strategy (classification method: DT3, error bars: Standard deviation
sensitivity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

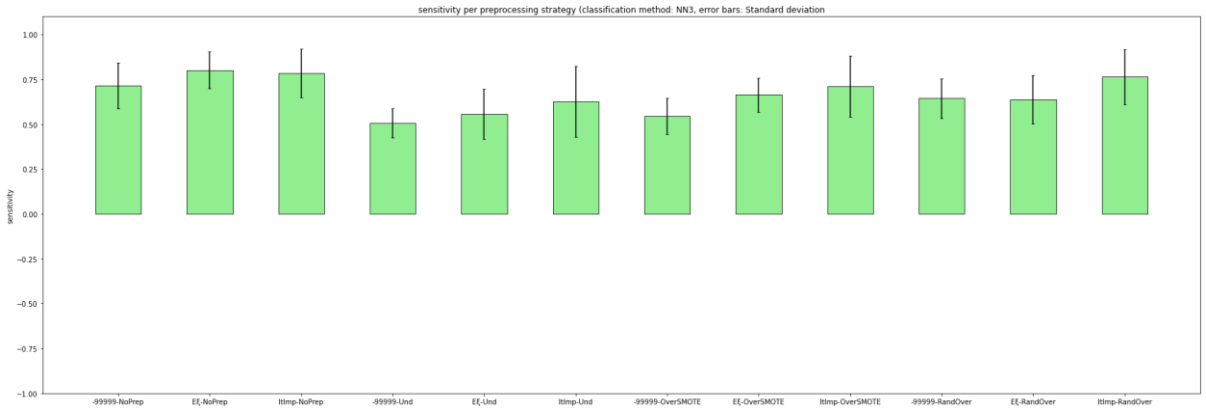

sensitivity per preprocessing strategy (classification method: DT4, error bars: Standard deviation
sensitivity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

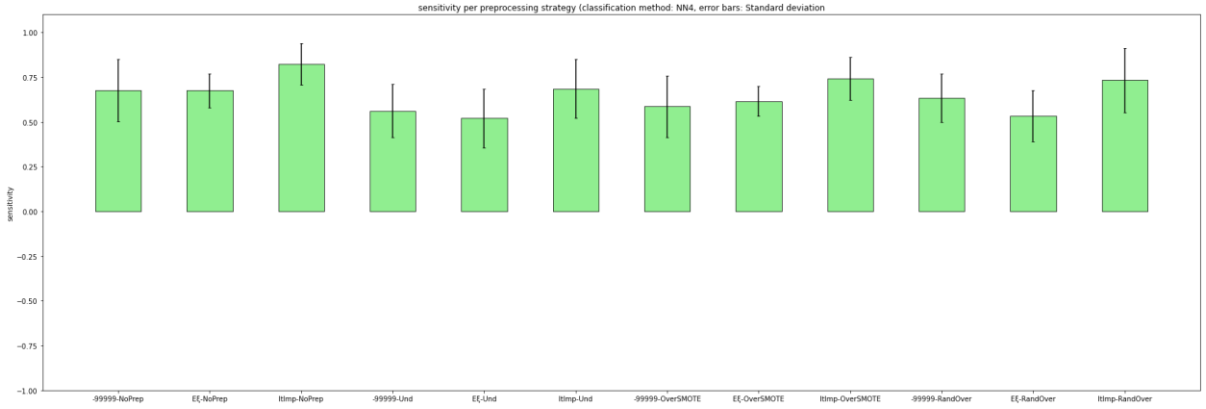

sensitivity per preprocessing strategy (classification method: RF1, error bars: Standard deviation
sensitivity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

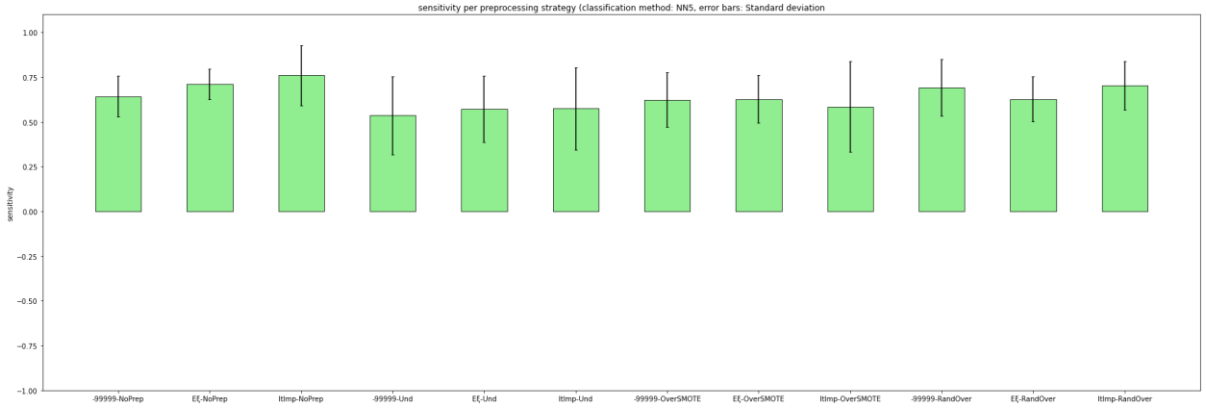

sensitivity per preprocessing strategy (classification method: RF2, error bars: Standard deviation
sensitivity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

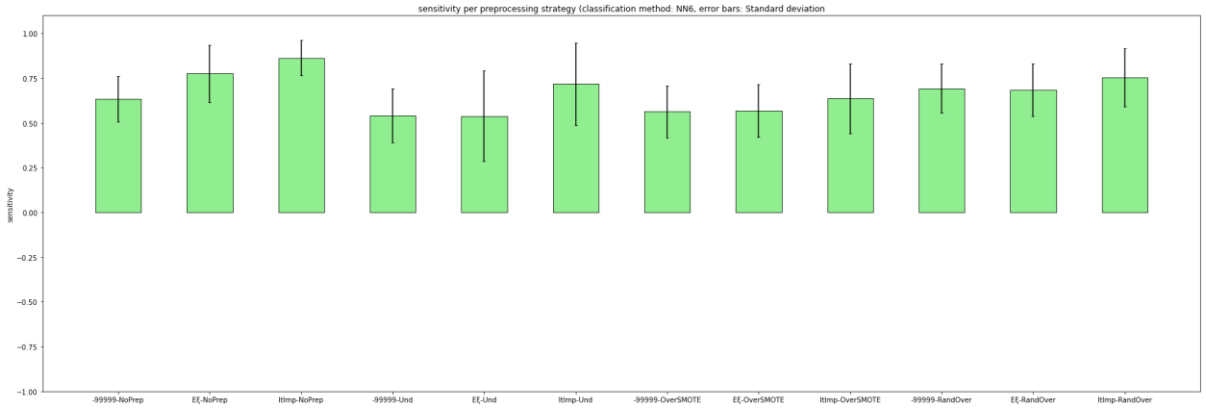

sensitivity per preprocessing strategy (classification method: NN1, error bars: Standard deviation
sensitivity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

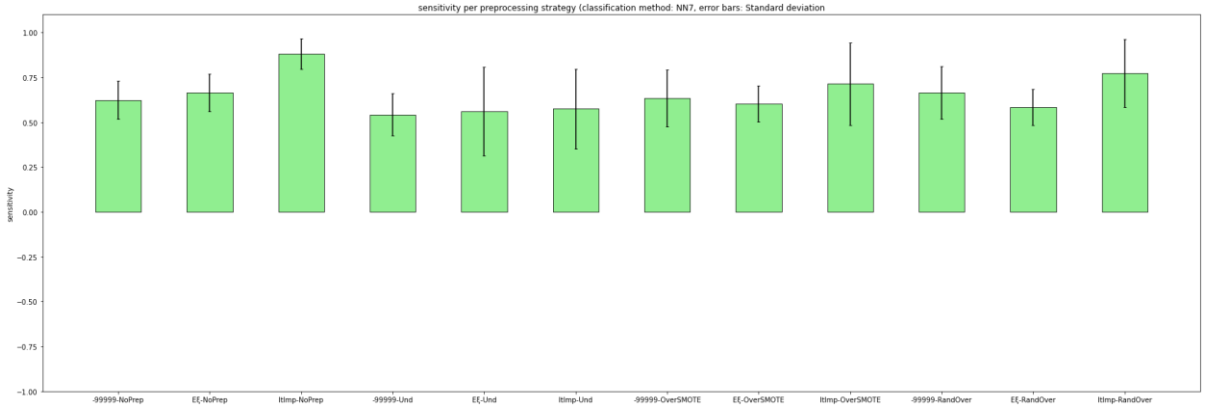

sensitivity per preprocessing strategy (classification method: NN2, error bars: Standard deviation
sensitivity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

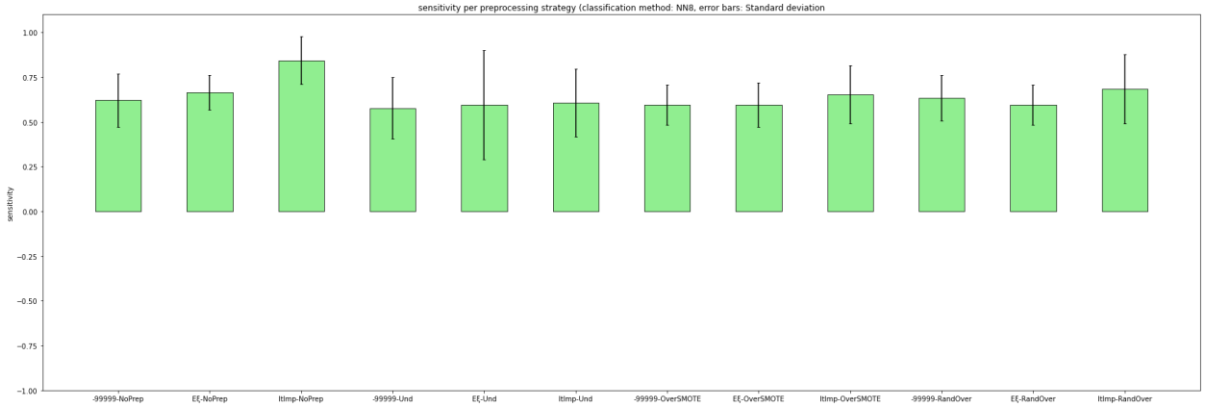

sensitivity per preprocessing strategy (classification method: NN3, error bars: Standard deviation
sensitivity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

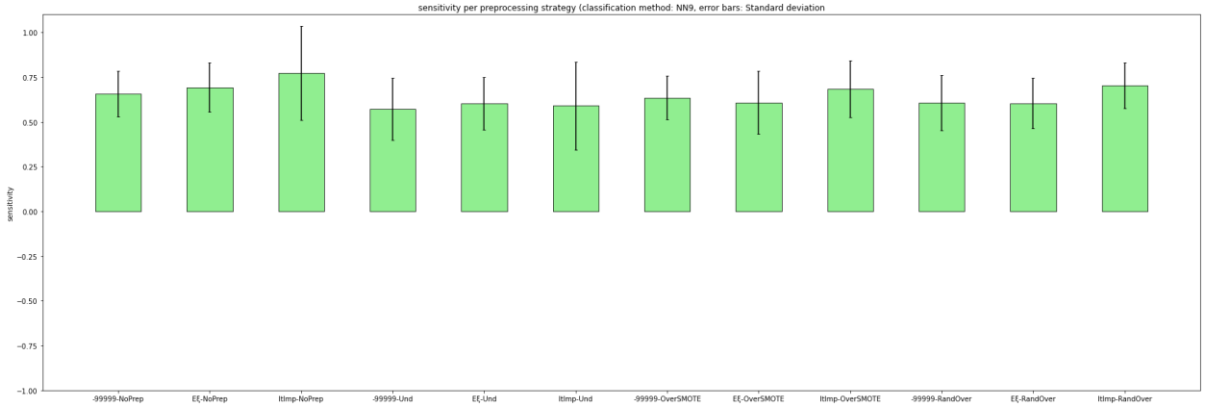

sensitivity per preprocessing strategy (classification method: NN4, error bars: Standard deviation
sensitivity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

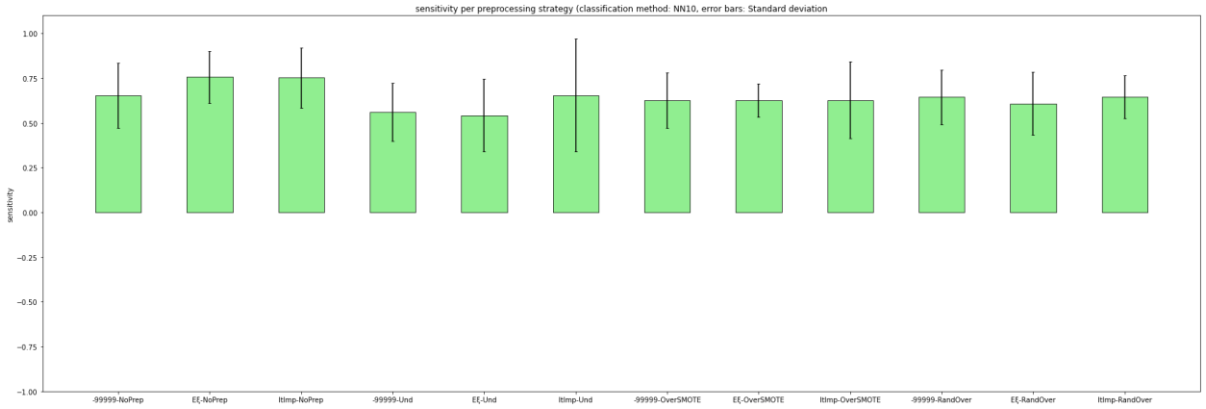

sensitivity per preprocessing strategy (classification method: NN5, error bars: Standard deviation
sensitivity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

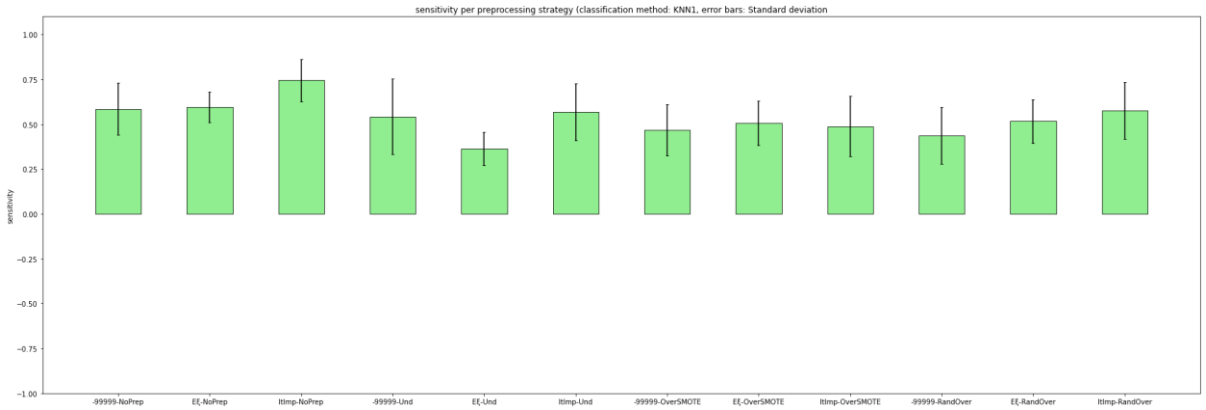

sensitivity per preprocessing strategy (classification method: NN6, error bars: Standard deviation
sensitivity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

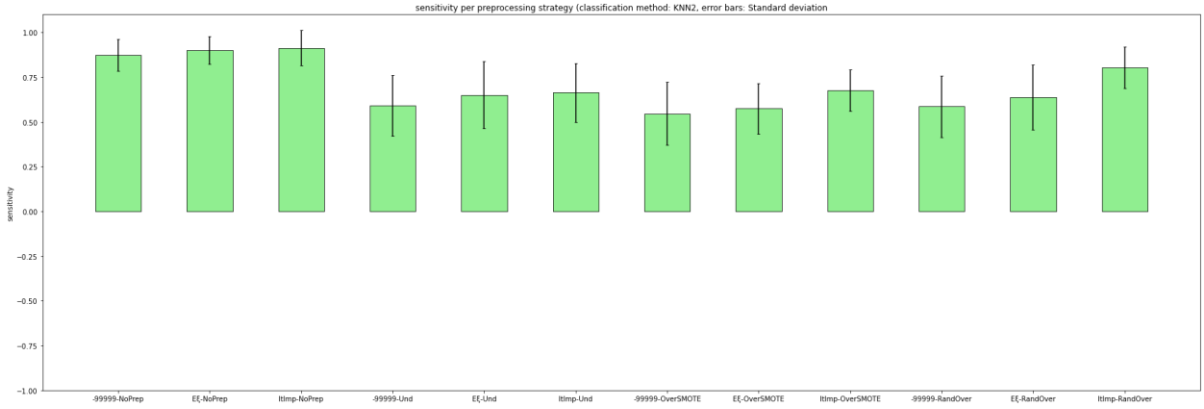

sensitivity per preprocessing strategy (classification method: NN7, error bars: Standard deviation
sensitivity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

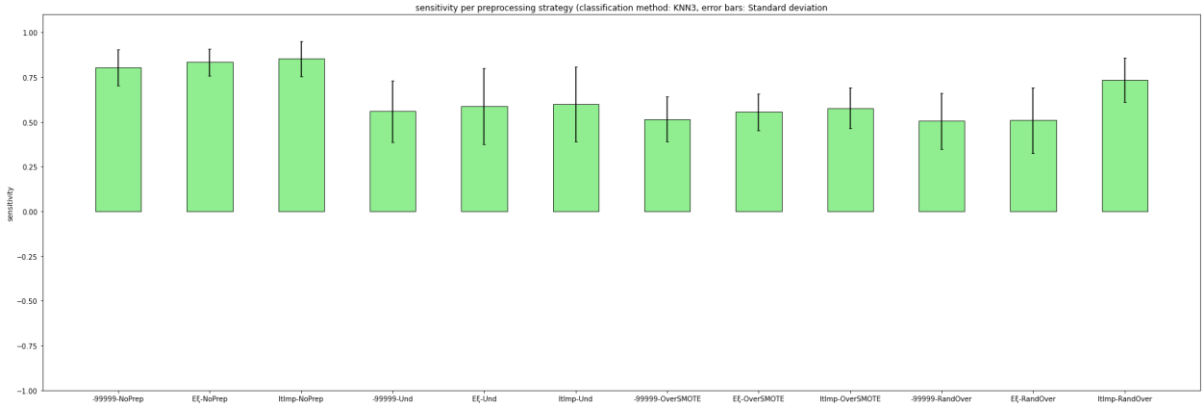

sensitivity per preprocessing strategy (classification method: NN8, error bars: Standard deviation
sensitivity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

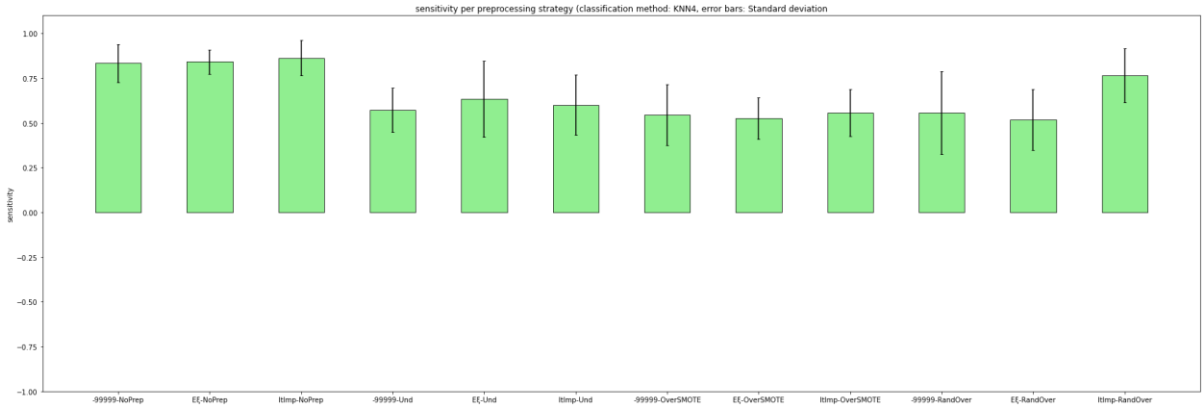

sensitivity per preprocessing strategy (classification method: NN9, error bars: Standard deviation
sensitivity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

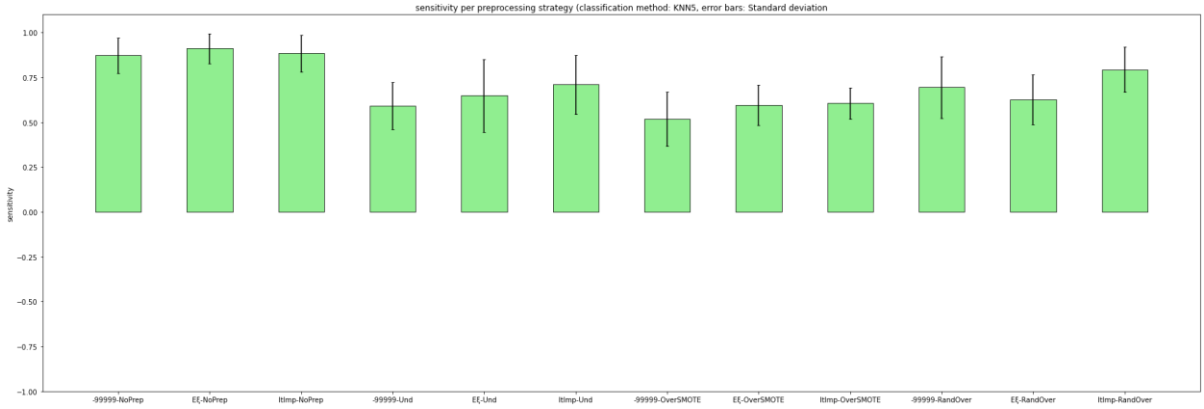

sensitivity per preprocessing strategy (classification method: NN10, error bars: Standard deviation
sensitivity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

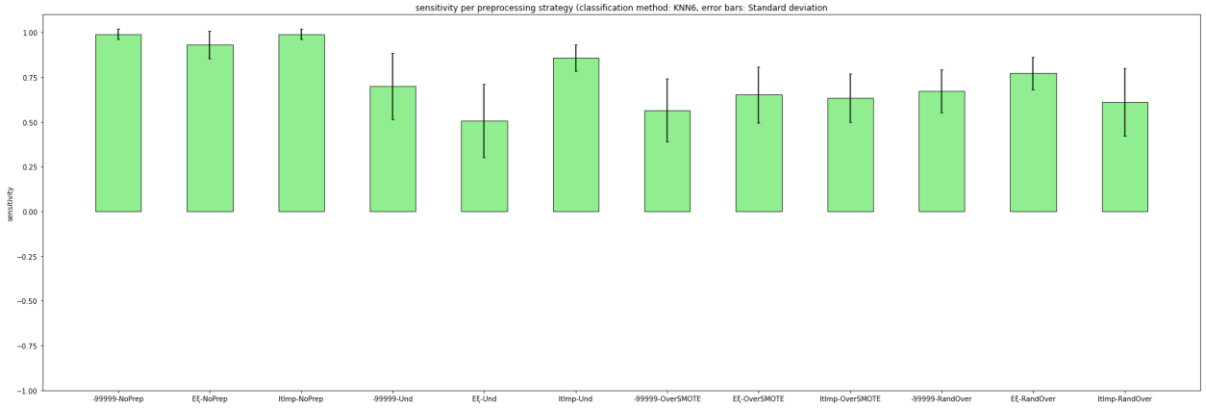

sensitivity per preprocessing strategy (classification method: KNN1, error bars: Standard deviation
sensitivity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

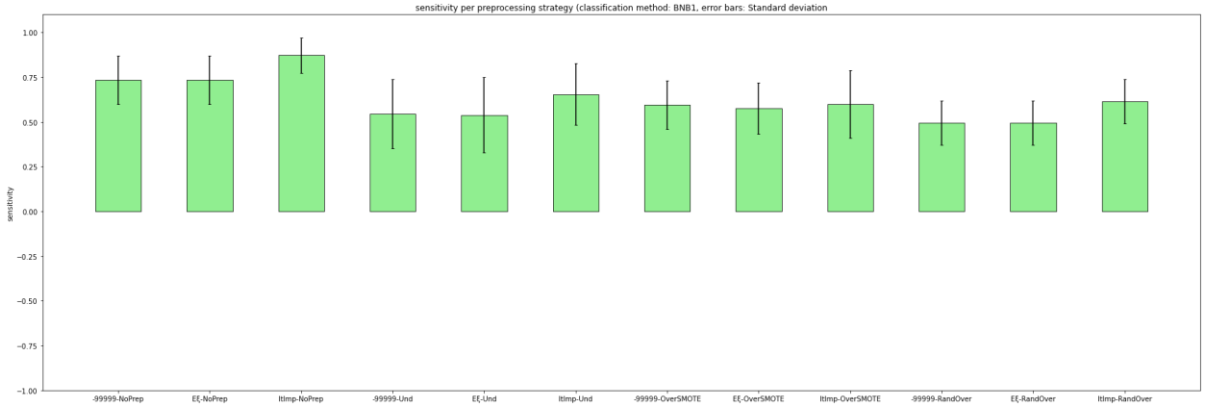

sensitivity per preprocessing strategy (classification method: KNN2, error bars: Standard deviation
sensitivity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

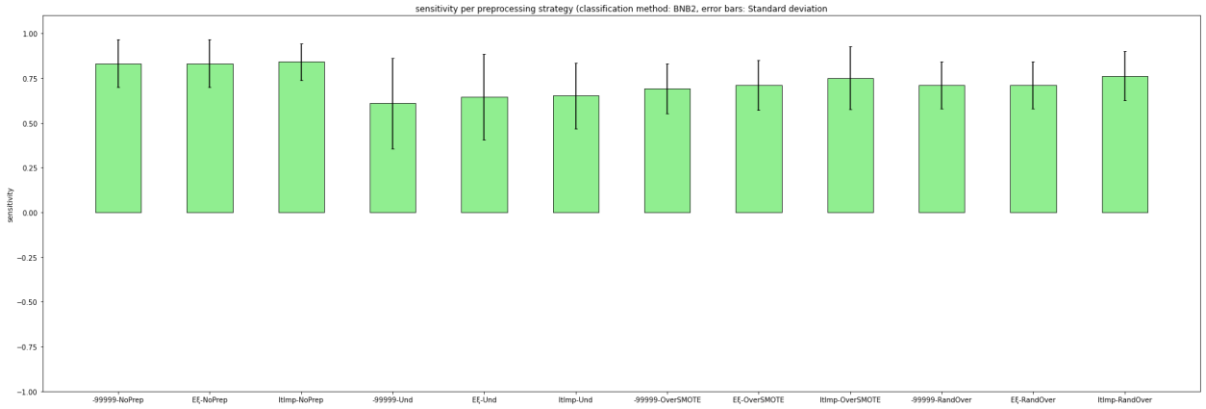

sensitivity per preprocessing strategy (classification method: KNN3, error bars: Standard deviation
sensitivity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

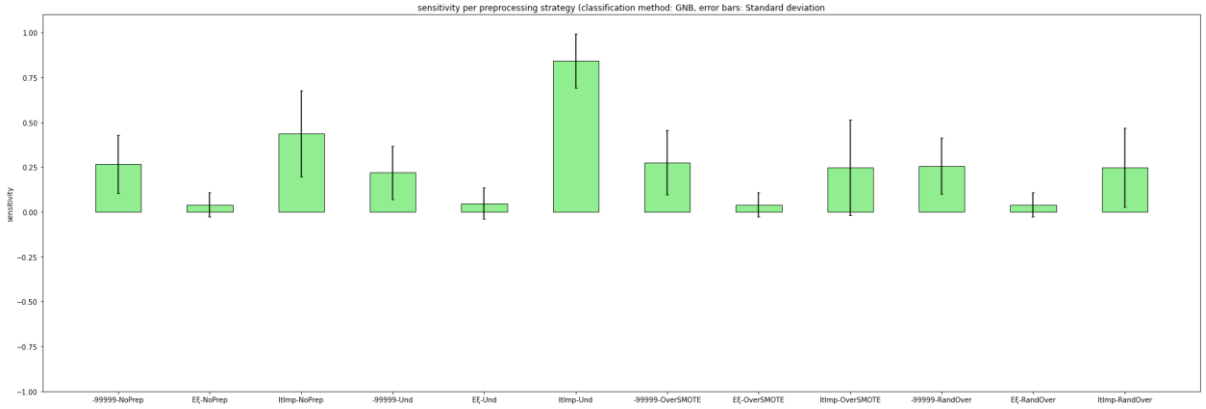

sensitivity per preprocessing strategy (classification method: KNN4, error bars: Standard deviation
sensitivity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

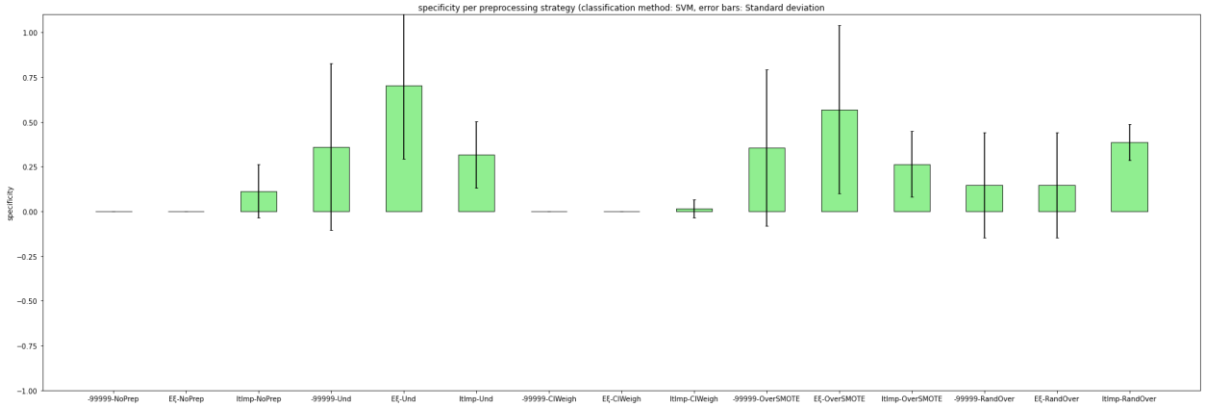

sensitivity per preprocessing strategy (classification method: KNN5, error bars: Standard deviation
sensitivity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

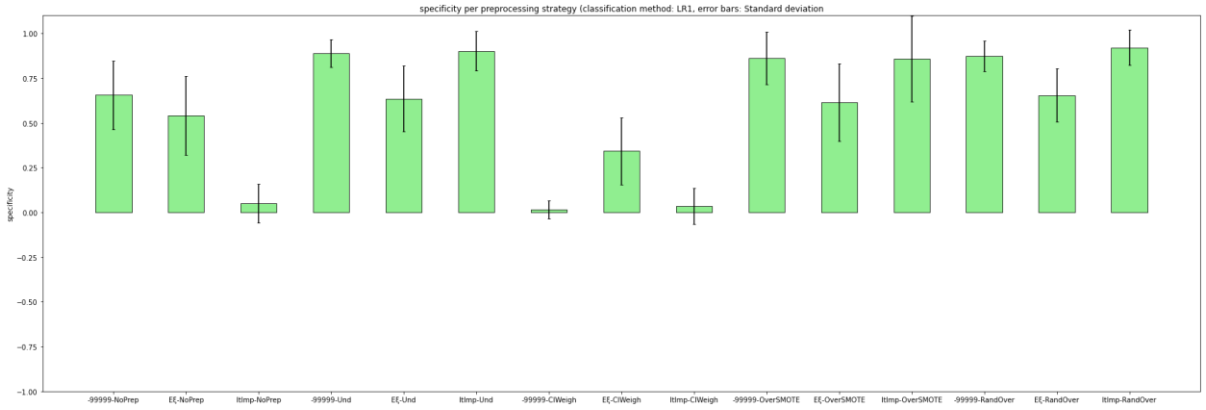

sensitivity per preprocessing strategy (classification method: KNN6, error bars: Standard deviation
sensitivity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

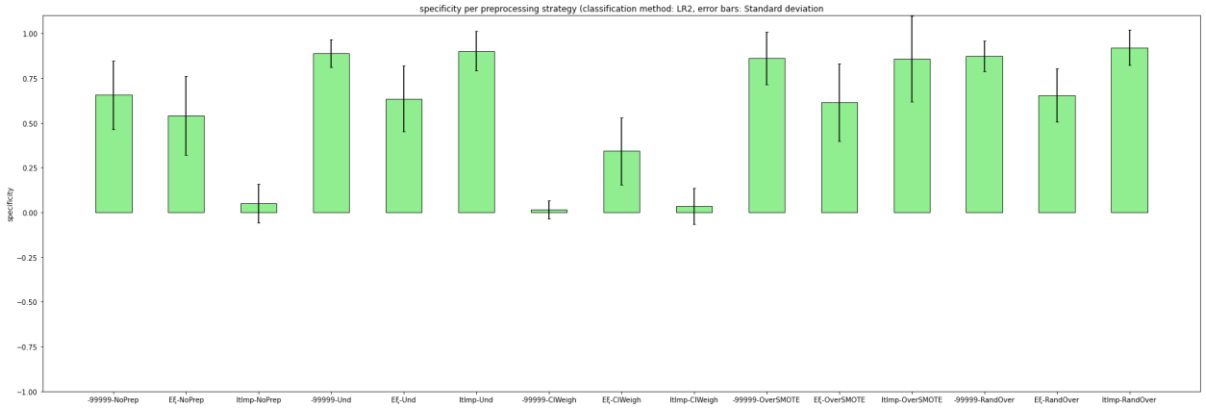

sensitivity per preprocessing strategy (classification method: BNB1, error bars: Standard deviation
sensitivity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

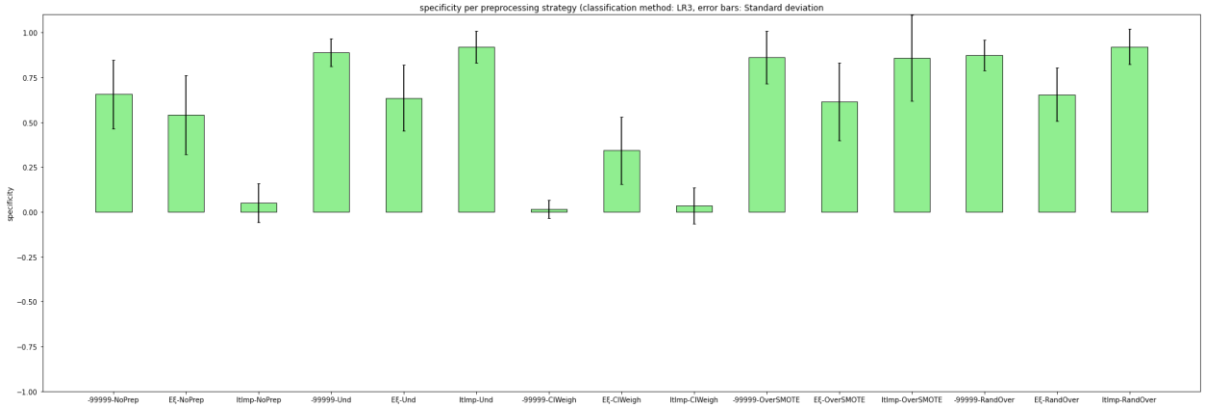

sensitivity per preprocessing strategy (classification method: BNB2, error bars: Standard deviation
sensitivity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

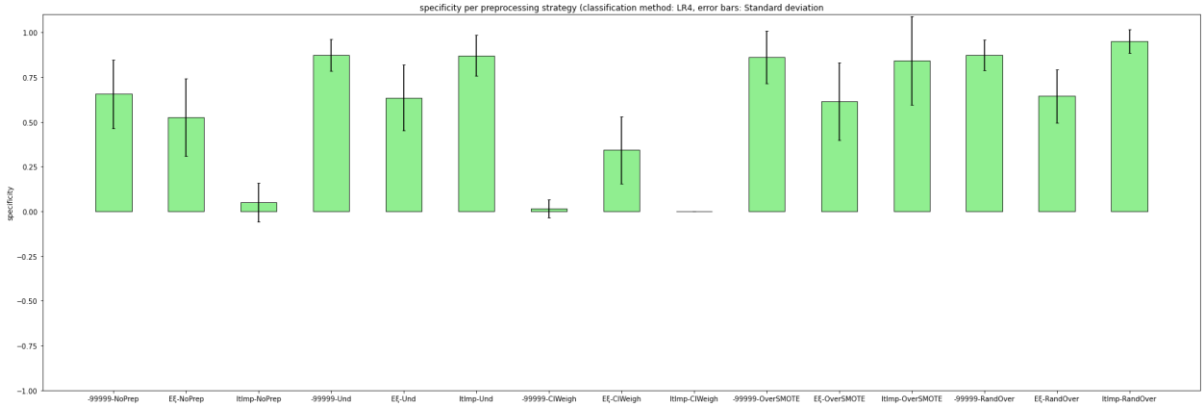

sensitivity per preprocessing strategy (classification method: GNB, error bars: Standard deviation
sensitivity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

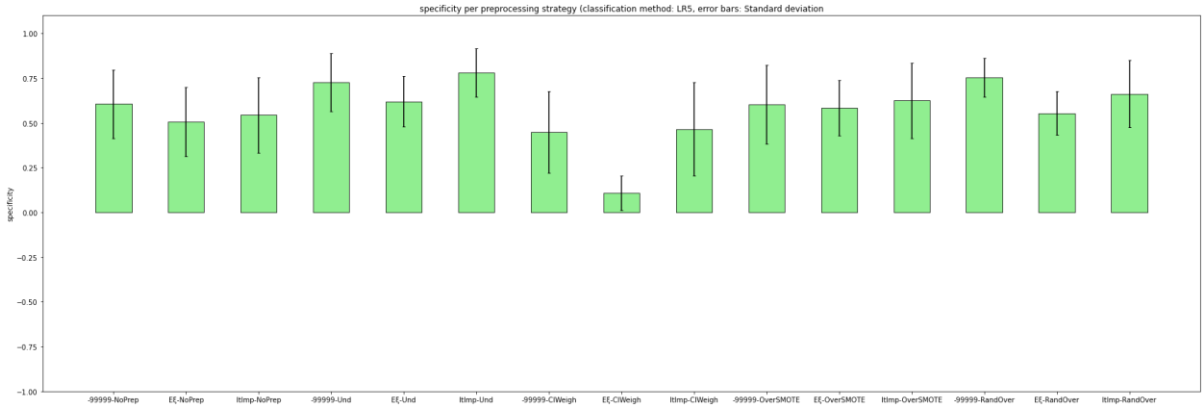

specificity per preprocessing strategy (classification method: SVM, error bars: Standard deviation
specificity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

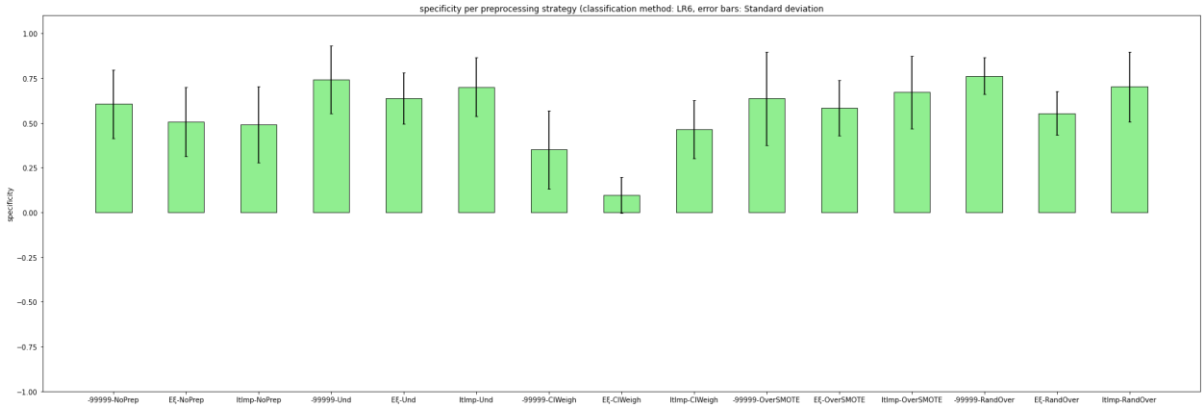

specificity per preprocessing strategy (classification method: LR1, error bars: Standard deviation
specificity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

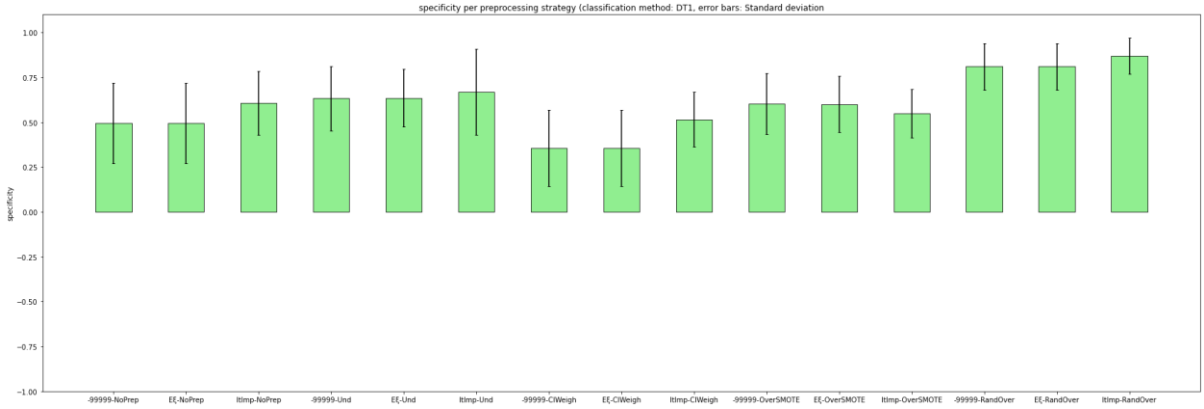

specificity per preprocessing strategy (classification method: LR2, error bars: Standard deviation
specificity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

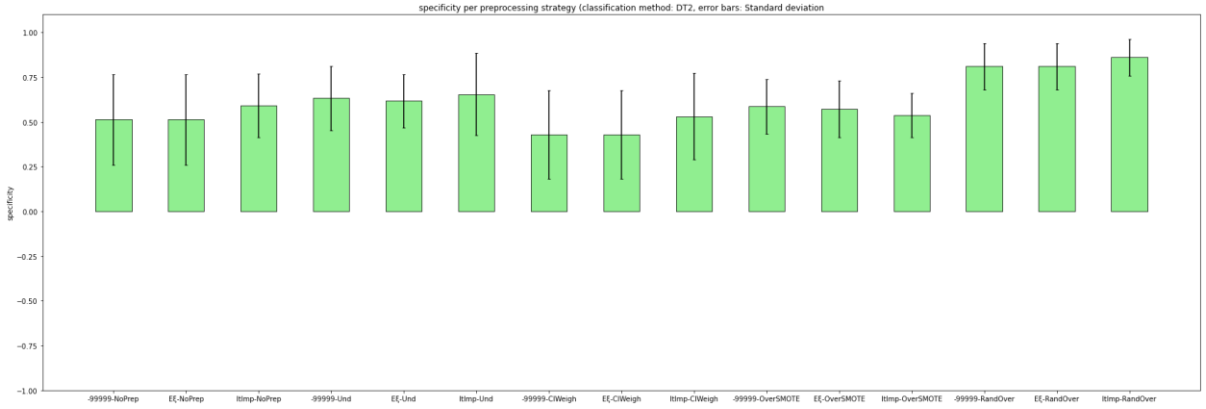

specificity per preprocessing strategy (classification method: LR3, error bars: Standard deviation
specificity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

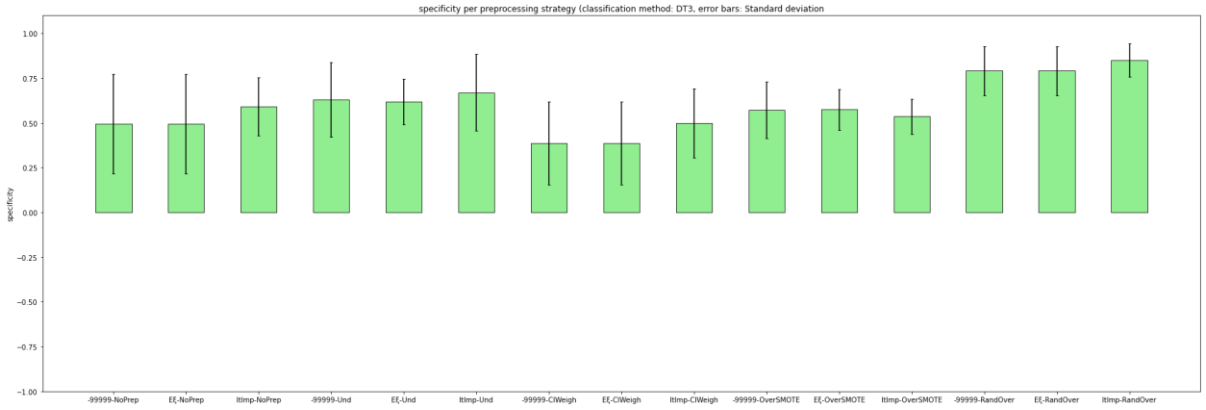

specificity per preprocessing strategy (classification method: LR4, error bars: Standard deviation
specificity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

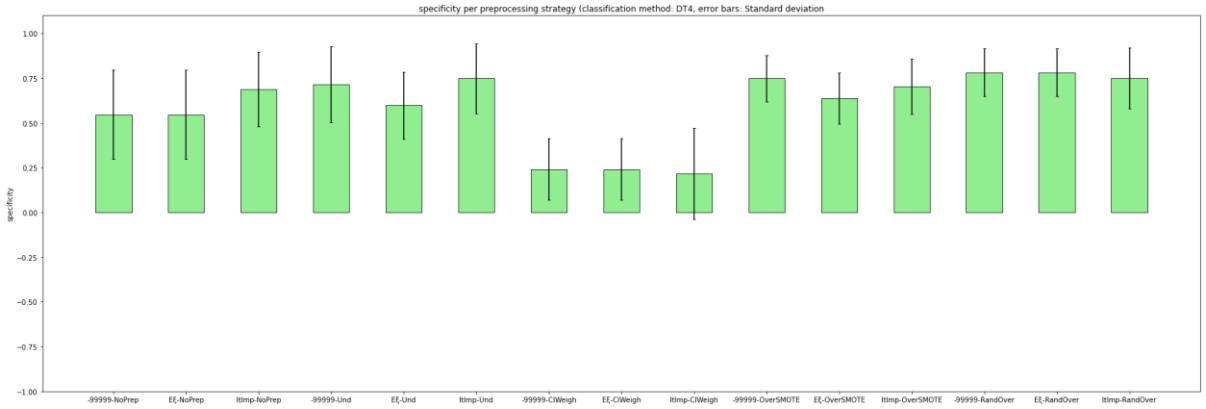

specificity per preprocessing strategy (classification method: LR5, error bars: Standard deviation
specificity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

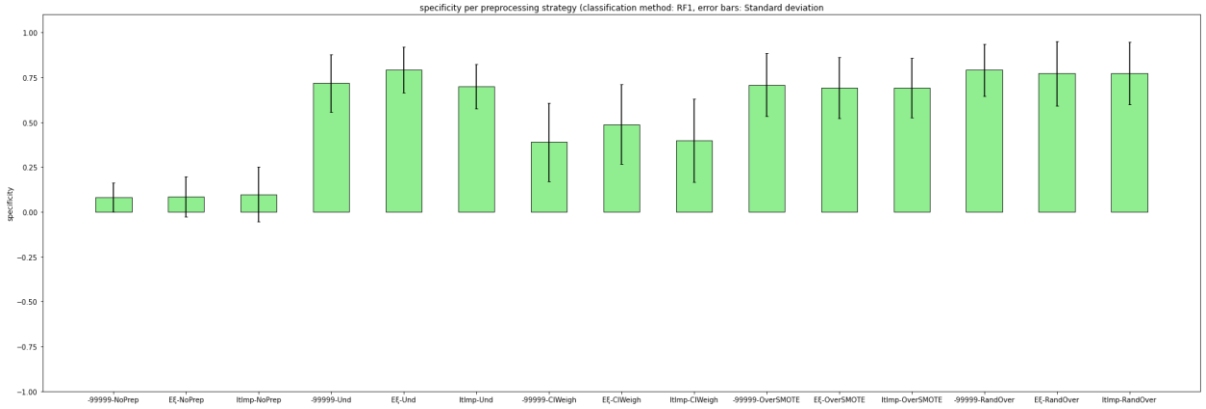

specificity per preprocessing strategy (classification method: LR6, error bars: Standard deviation
specificity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

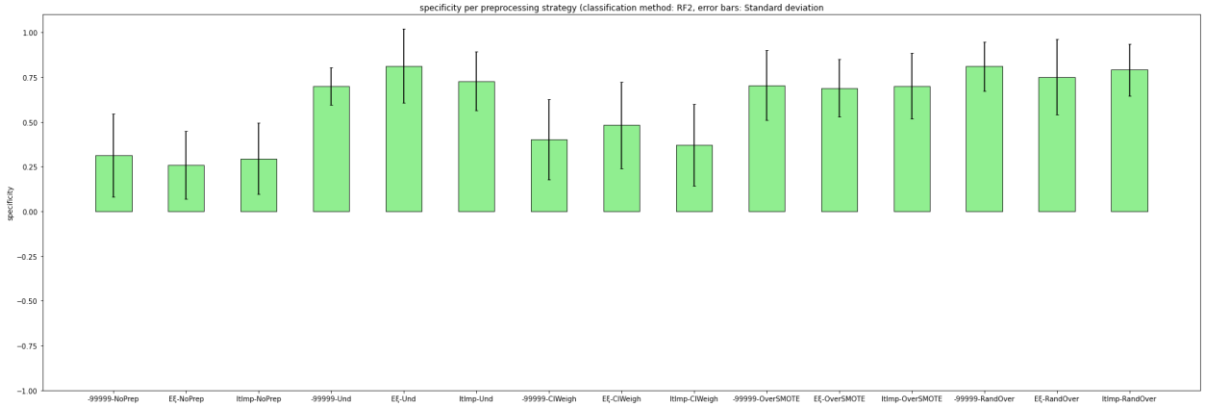

specificity per preprocessing strategy (classification method: DT1, error bars: Standard deviation
specificity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

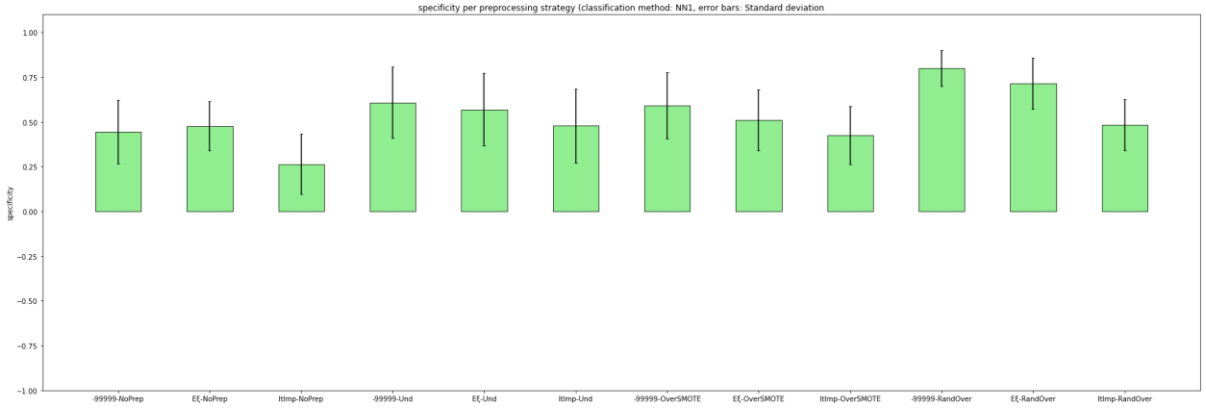

specificity per preprocessing strategy (classification method: DT2, error bars: Standard deviation
specificity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

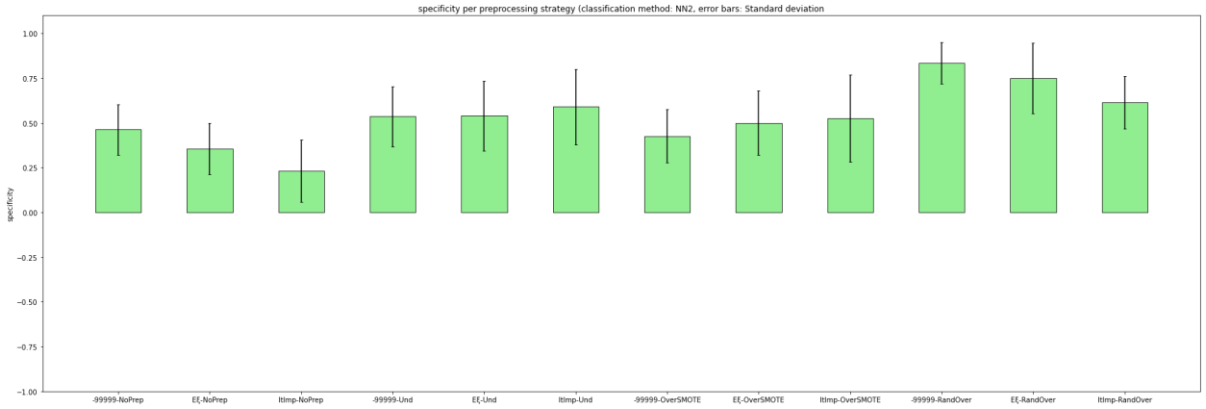

specificity per preprocessing strategy (classification method: DT3, error bars: Standard deviation
specificity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

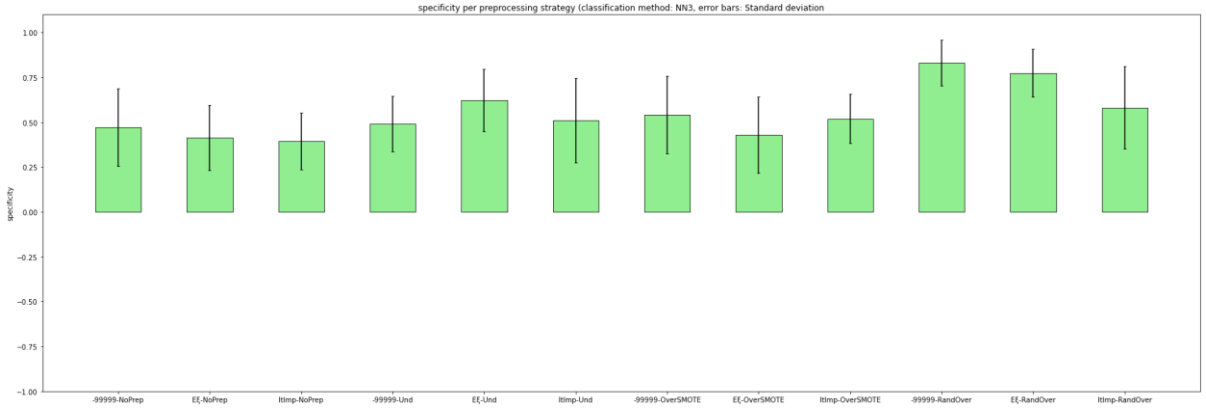

specificity per preprocessing strategy (classification method: DT4, error bars: Standard deviation
specificity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

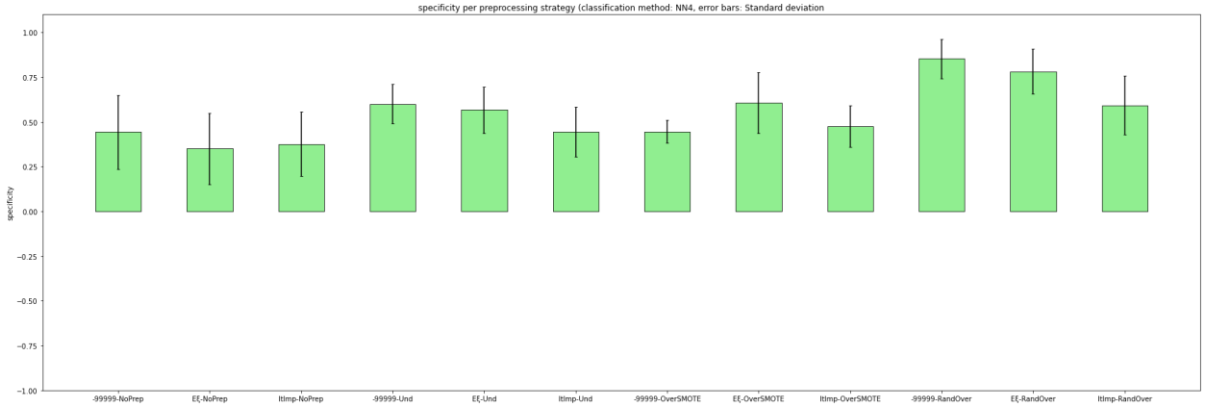

specificity per preprocessing strategy (classification method: RF1, error bars: Standard deviation
specificity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

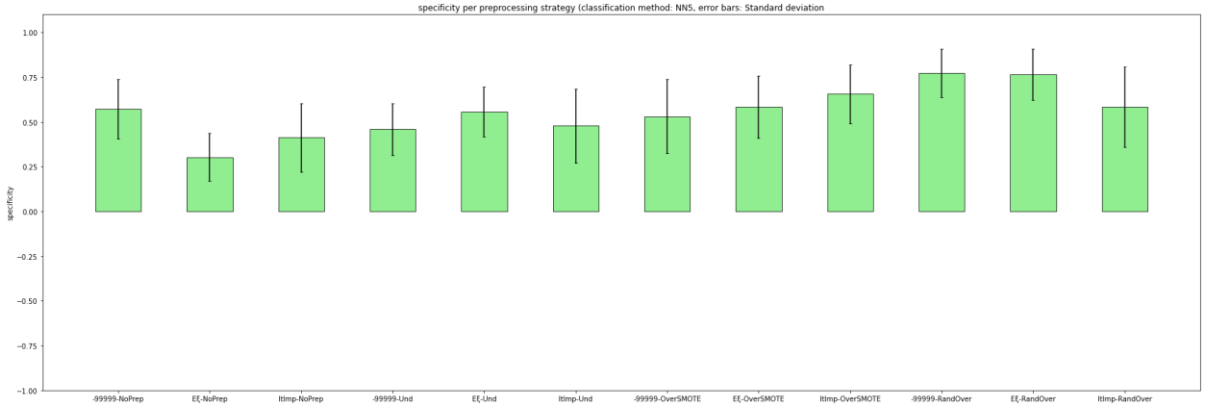

specificity per preprocessing strategy (classification method: RF2, error bars: Standard deviation
specificity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

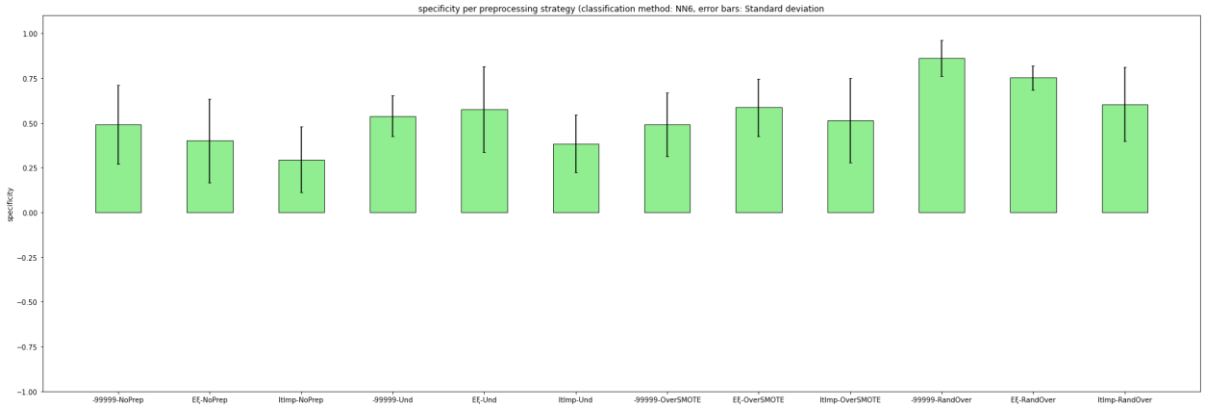

specificity per preprocessing strategy (classification method: NN1, error bars: Standard deviation
specificity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

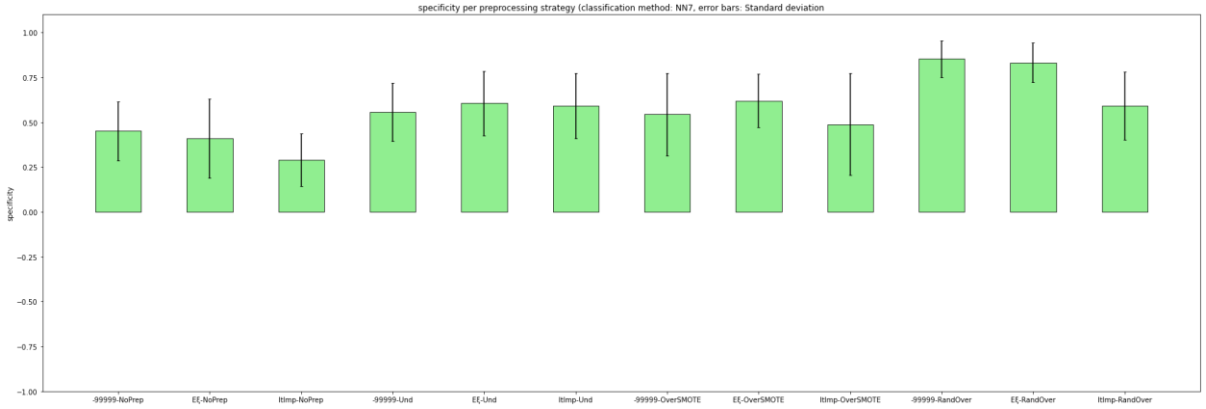

specificity per preprocessing strategy (classification method: NN2, error bars: Standard deviation
specificity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

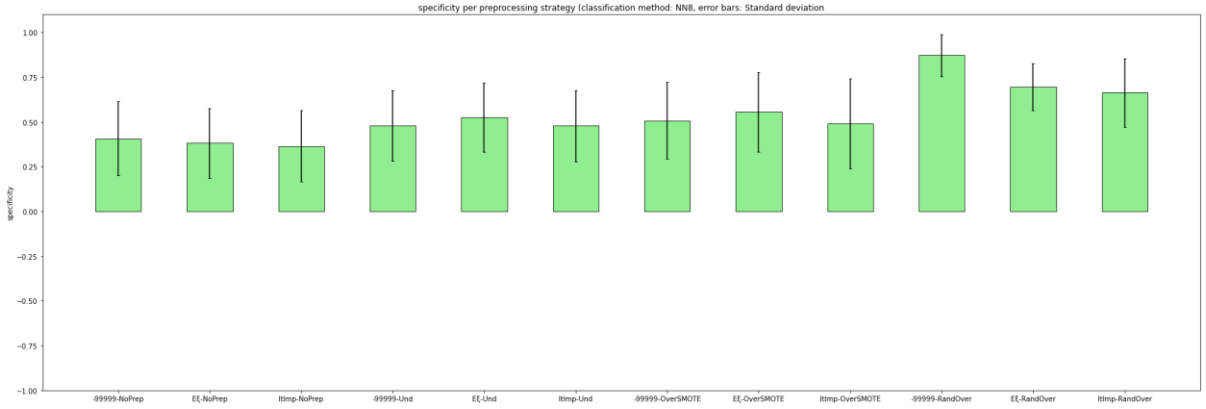

specificity per preprocessing strategy (classification method: NN3, error bars: Standard deviation
specificity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

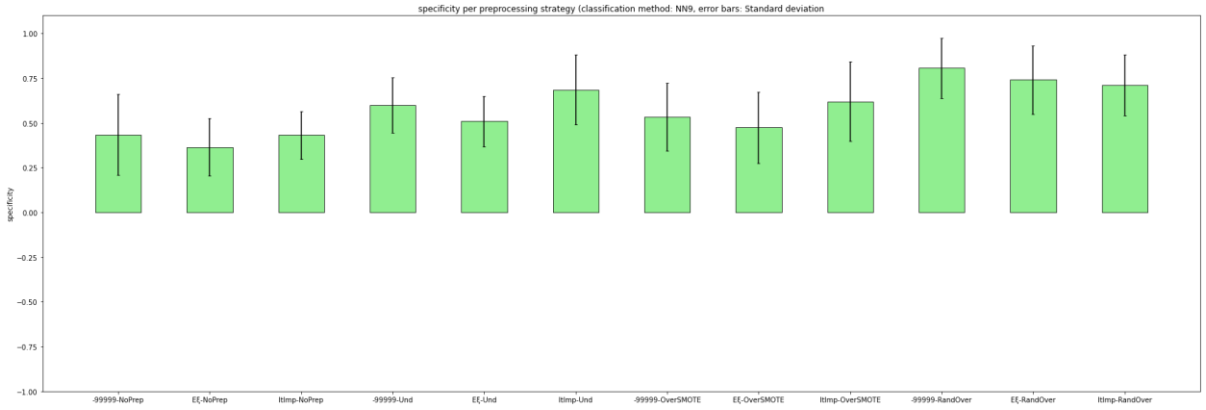

specificity per preprocessing strategy (classification method: NN4, error bars: Standard deviation
specificity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

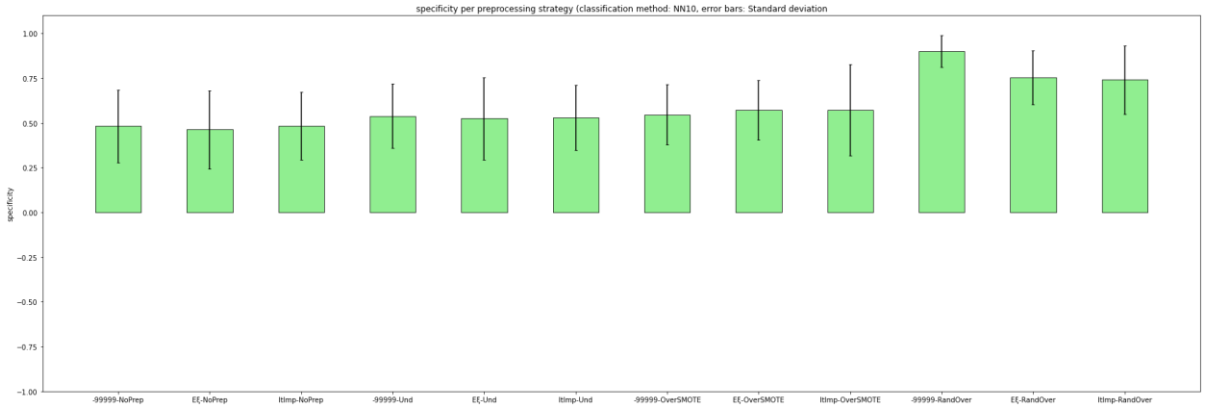

specificity per preprocessing strategy (classification method: NN5, error bars: Standard deviation
specificity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

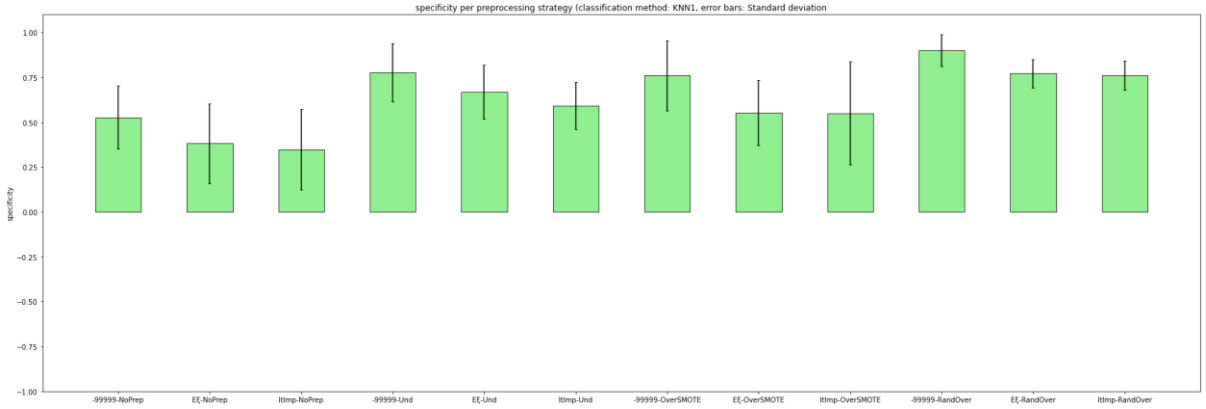

specificity per preprocessing strategy (classification method: NN6, error bars: Standard deviation
specificity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

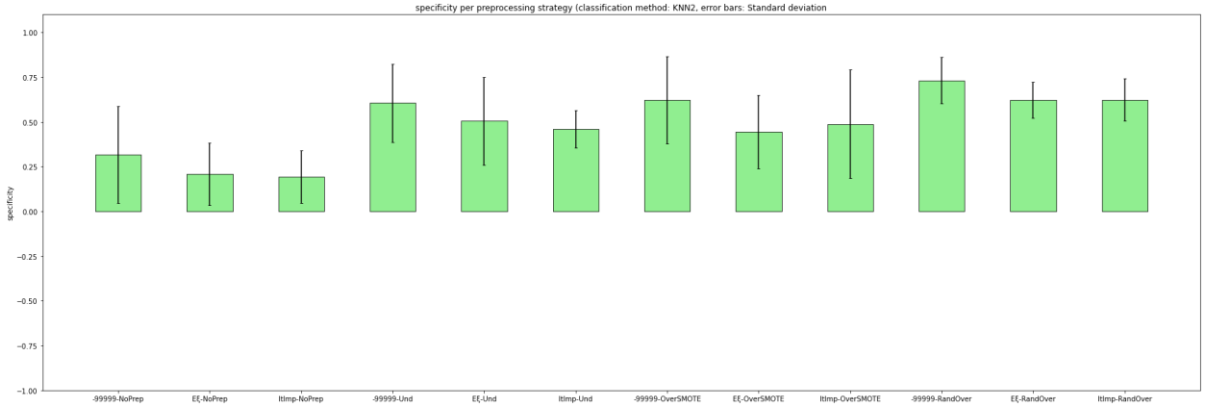

specificity per preprocessing strategy (classification method: NN7, error bars: Standard deviation
specificity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

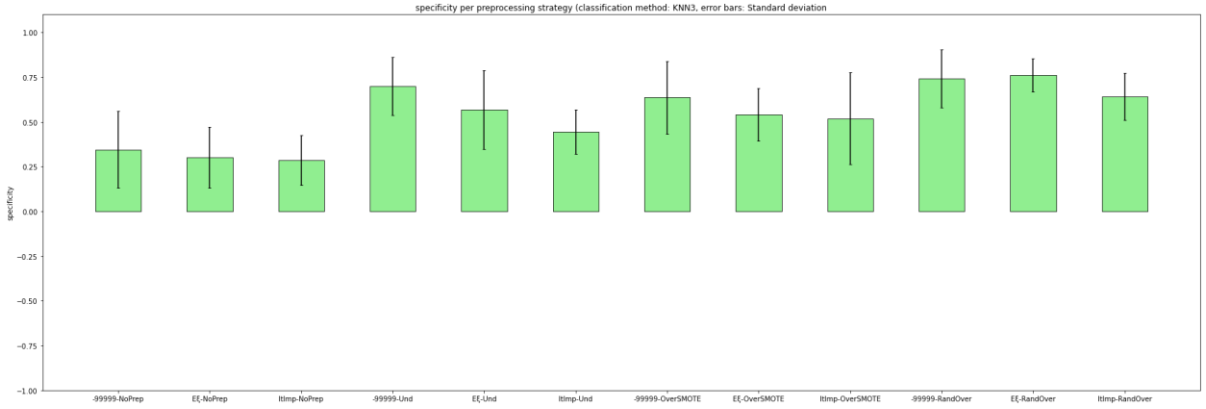

specificity per preprocessing strategy (classification method: NN8, error bars: Standard deviation
specificity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

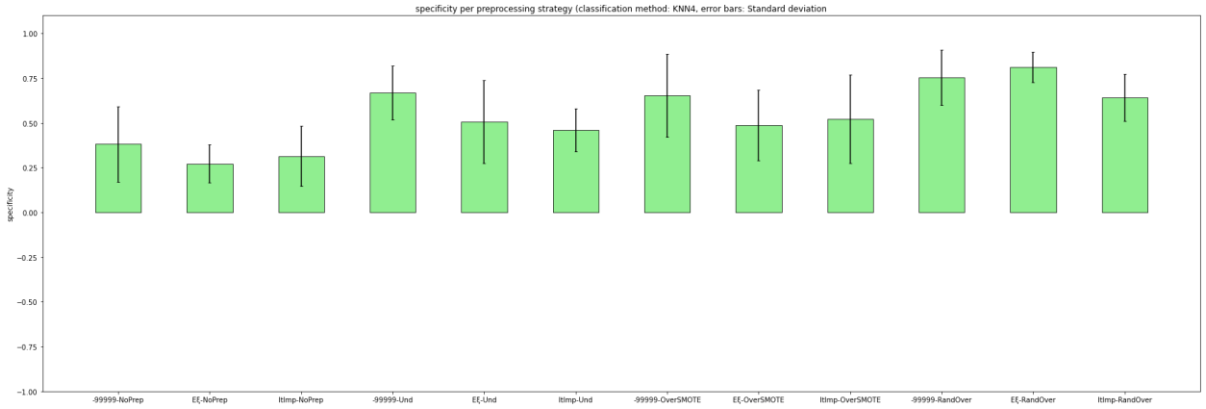

specificity per preprocessing strategy (classification method: NN9, error bars: Standard deviation
specificity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

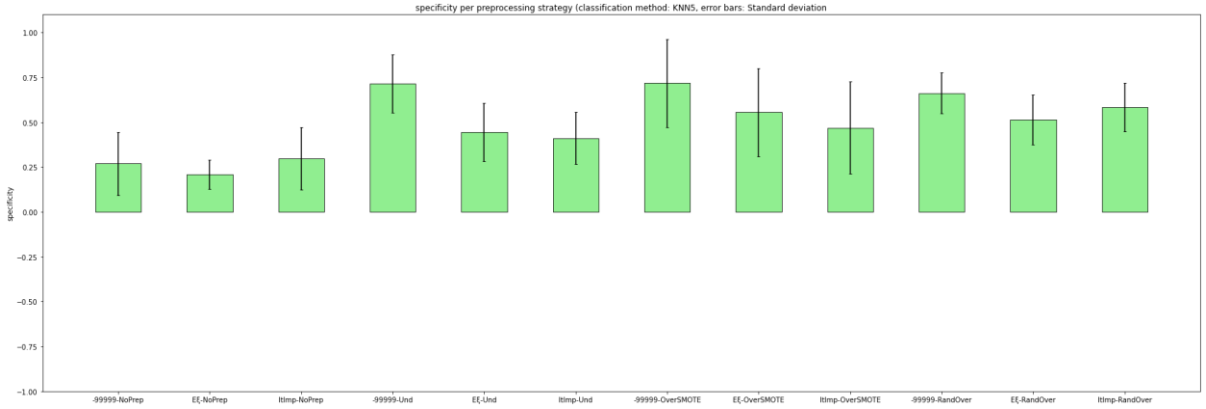

specificity per preprocessing strategy (classification method: NN10, error bars: Standard deviation
specificity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

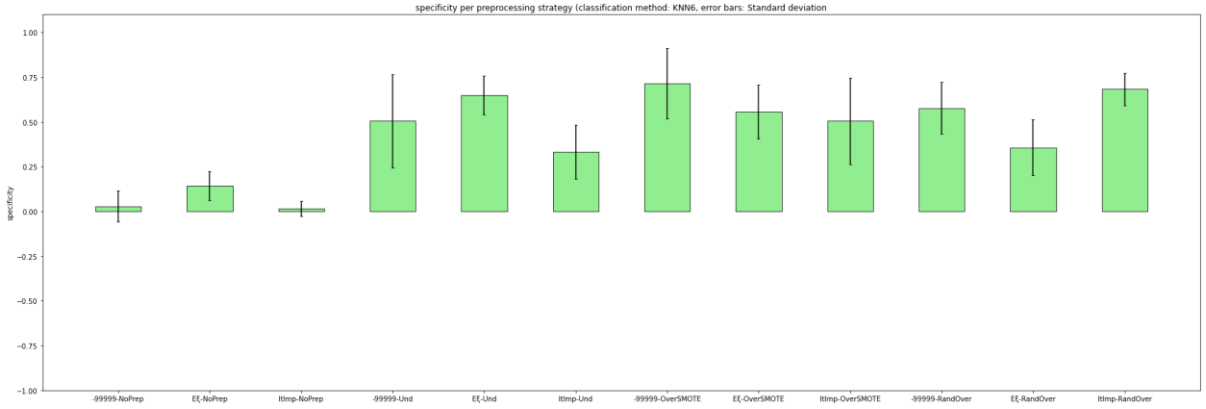

specificity per preprocessing strategy (classification method: KNN1, error bars: Standard deviation
specificity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

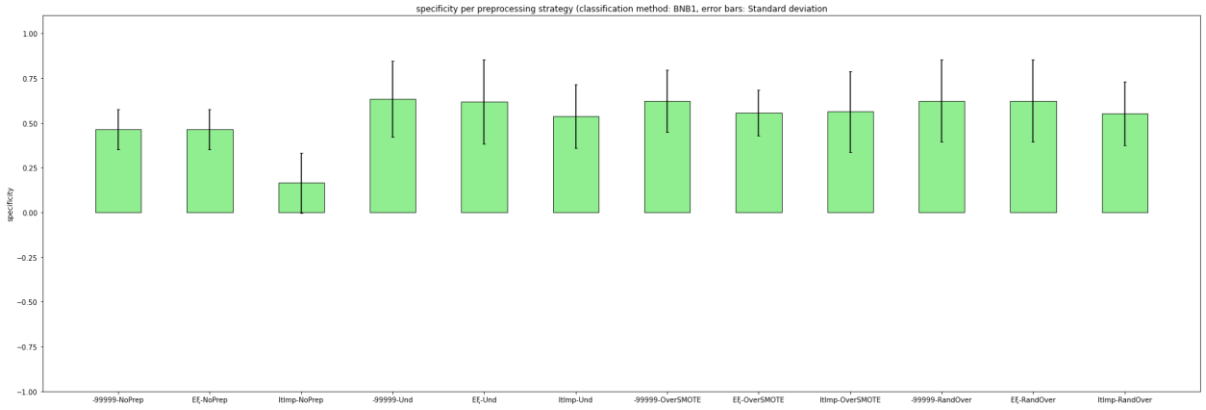

specificity per preprocessing strategy (classification method: KNN2, error bars: Standard deviation
specificity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

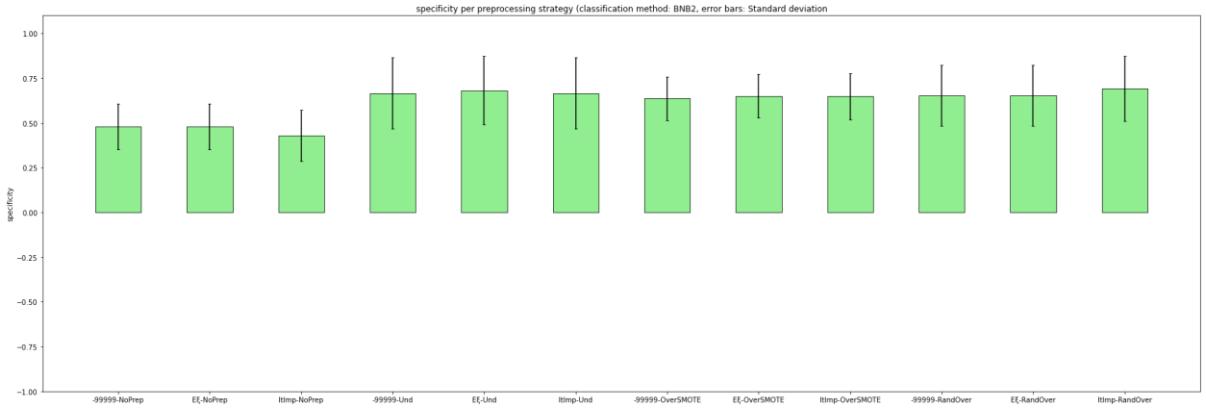

specificity per preprocessing strategy (classification method: KNN3, error bars: Standard deviation
specificity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

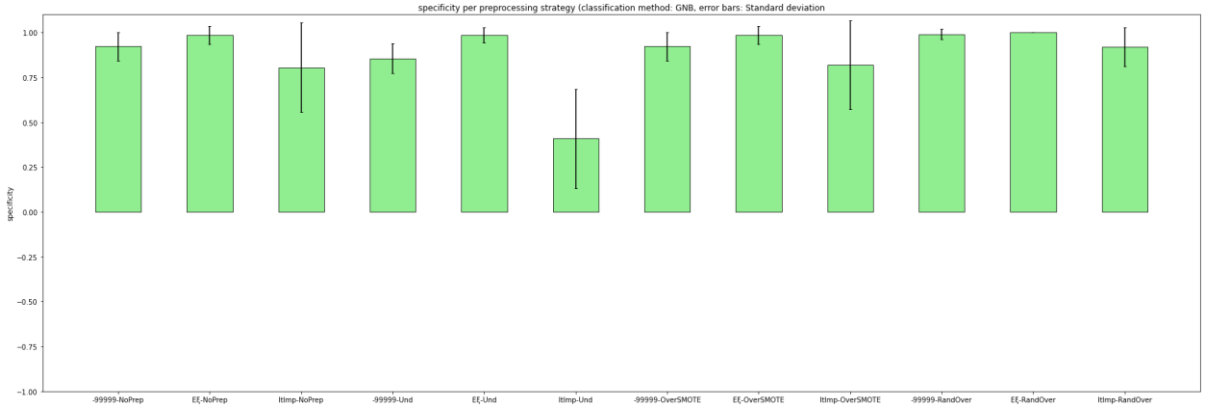

specificity per preprocessing strategy (classification method: KNN4, error bars: Standard deviation
specificity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

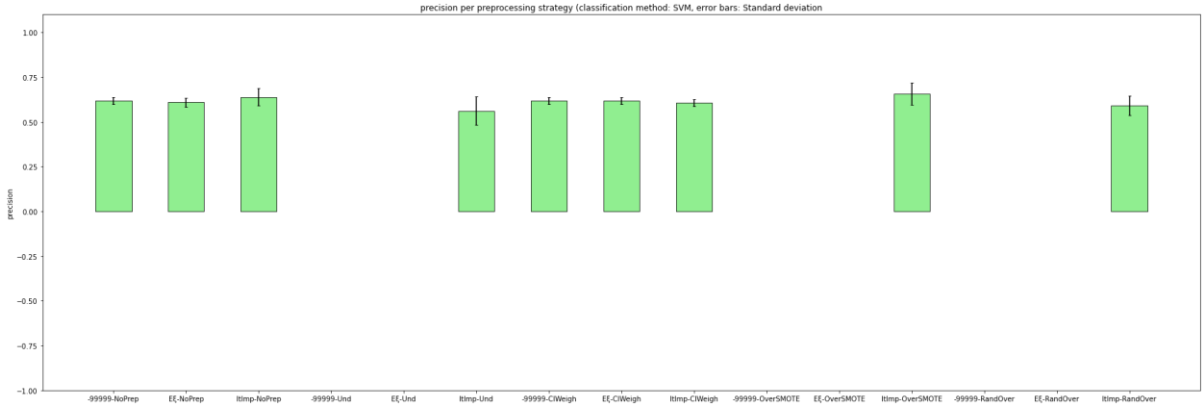

specificity per preprocessing strategy (classification method: KNN5, error bars: Standard deviation
specificity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

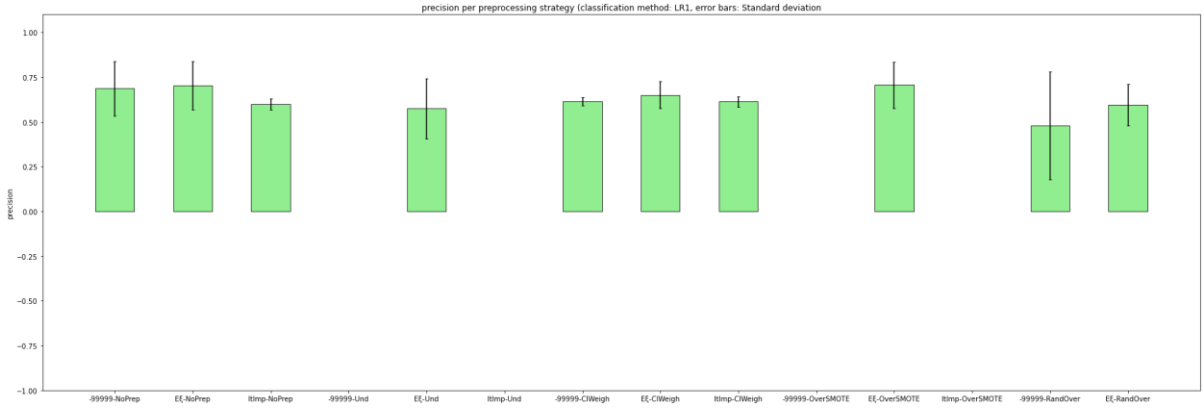

specificity per preprocessing strategy (classification method: KNN6, error bars: Standard deviation
specificity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

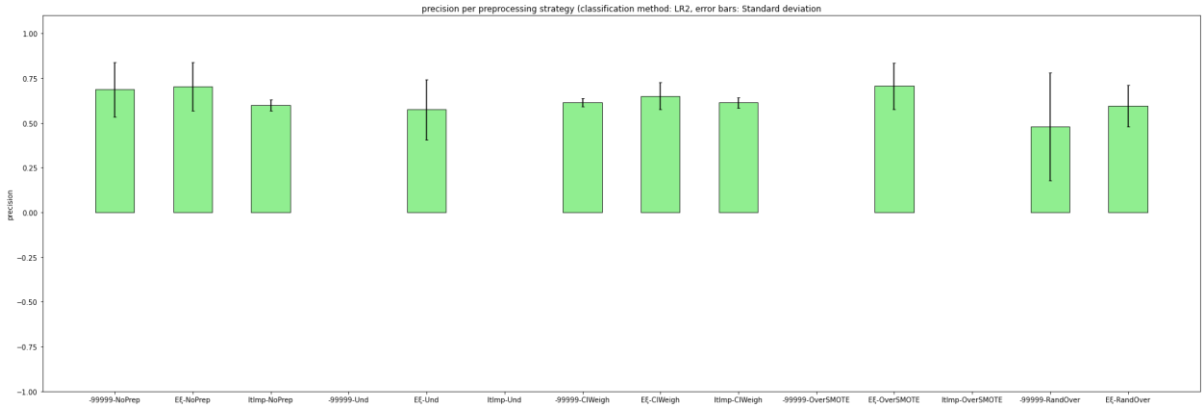

specificity per preprocessing strategy (classification method: BNB1, error bars: Standard deviation
specificity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

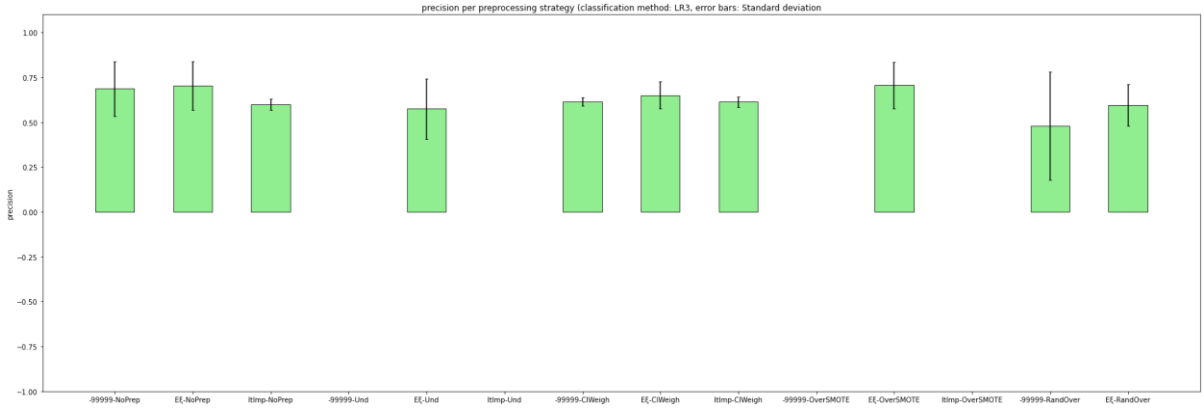

specificity per preprocessing strategy (classification method: BNB2, error bars: Standard deviation
specificity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

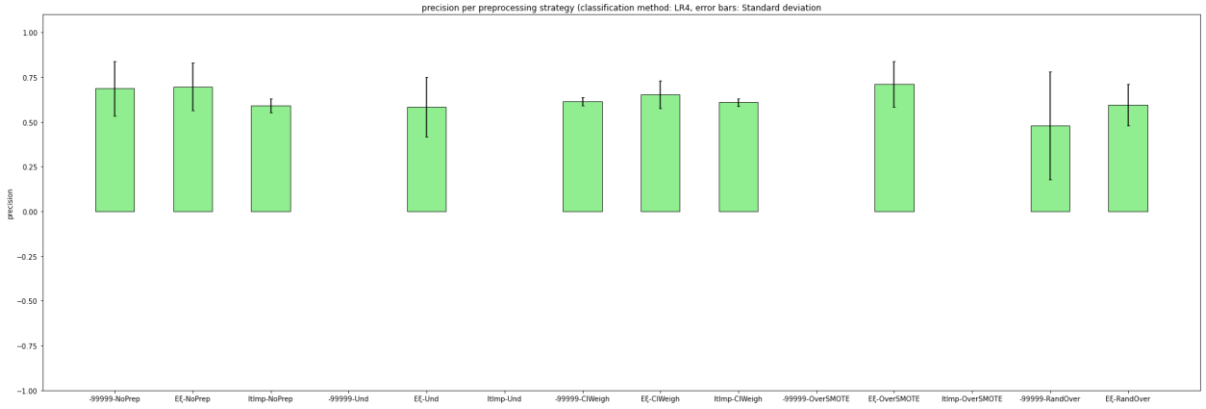

specificity per preprocessing strategy (classification method: GNB, error bars: Standard deviation
specificity
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

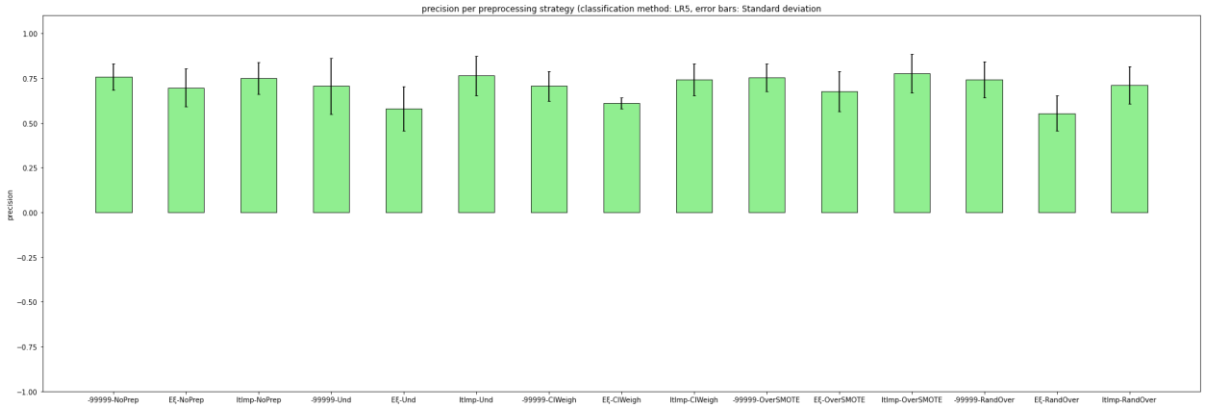

precision per preprocessing strategy (classification method: SVM, error bars: Standard deviation
precision
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

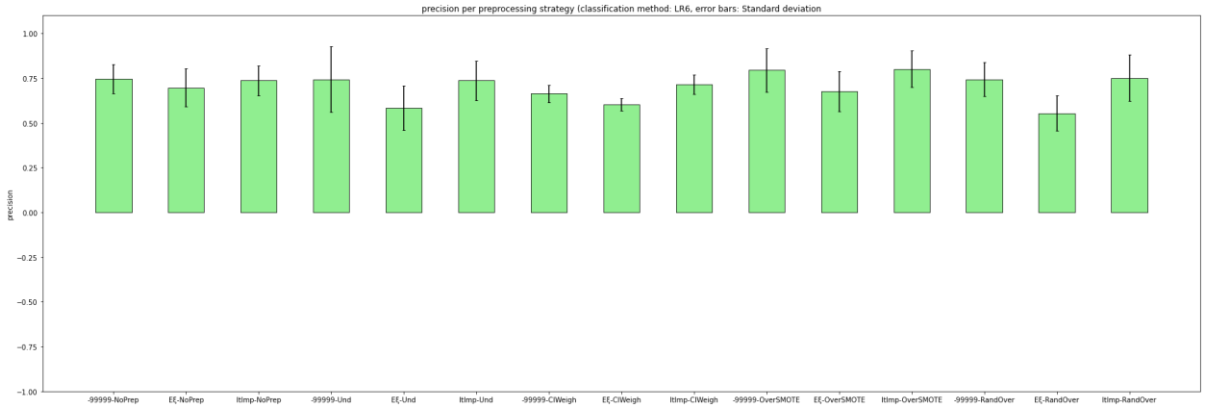

precision per preprocessing strategy (classification method: LR1, error bars: Standard deviation
precision
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver

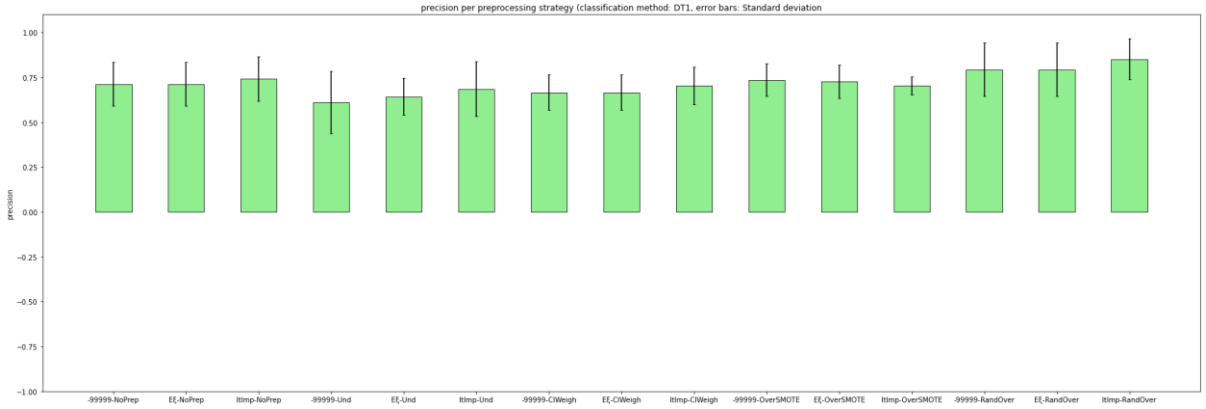

precision per preprocessing strategy (classification method: LR2, error bars: Standard deviation
precision
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver

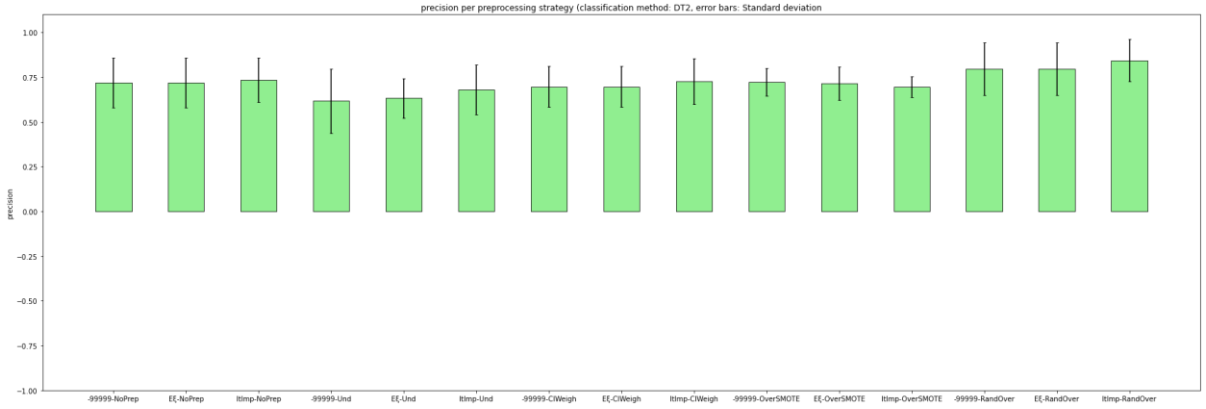

precision per preprocessing strategy (classification method: LR3, error bars: Standard deviation
precision
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver

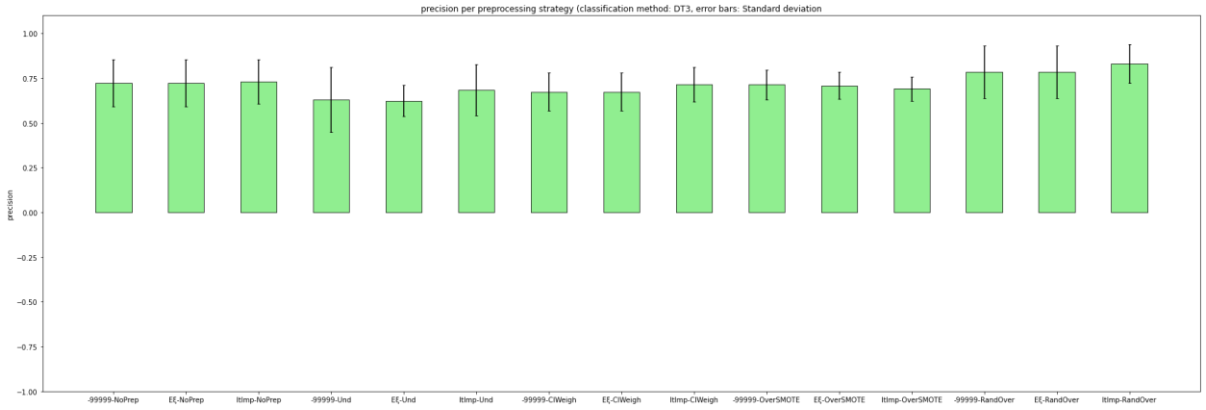

precision per preprocessing strategy (classification method: LR4, error bars: Standard deviation
precision
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver

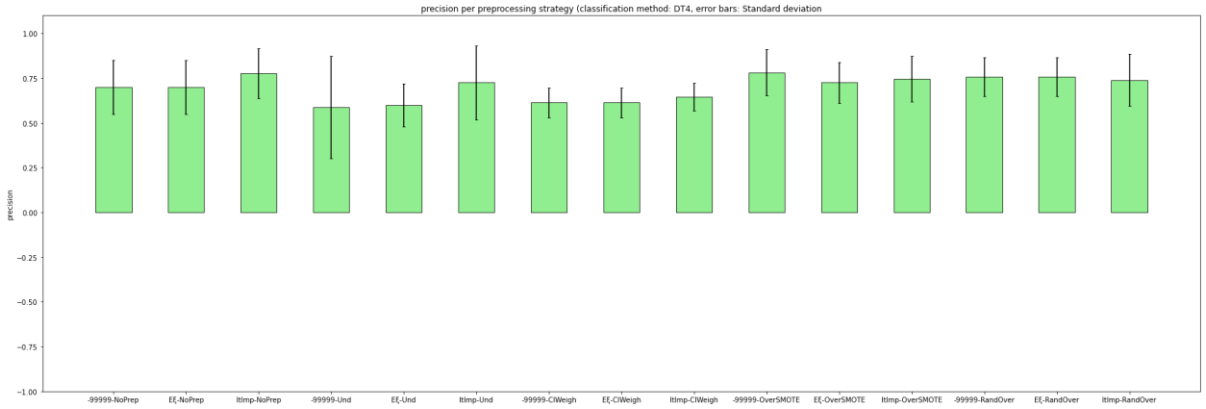

precision per preprocessing strategy (classification method: LR5, error bars: Standard deviation
precision
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

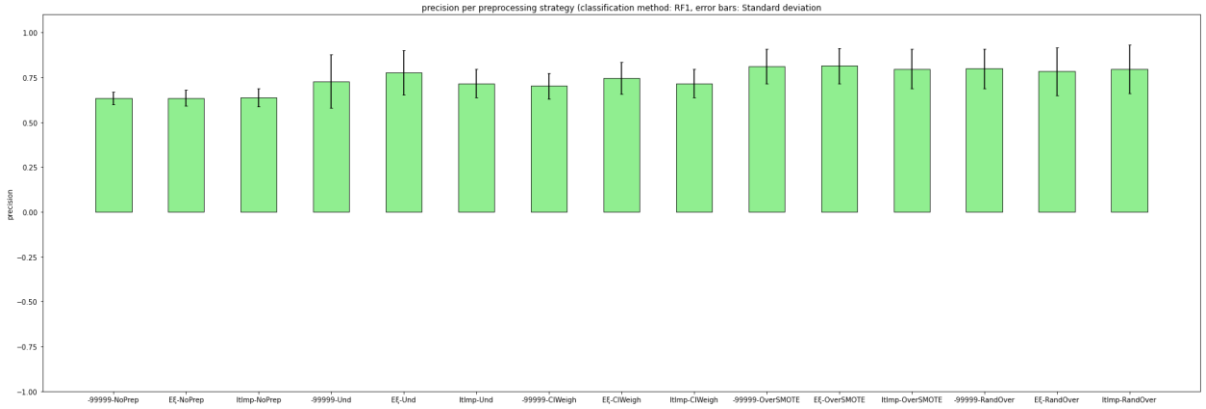

precision per preprocessing strategy (classification method: LR6, error bars: Standard deviation
precision
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

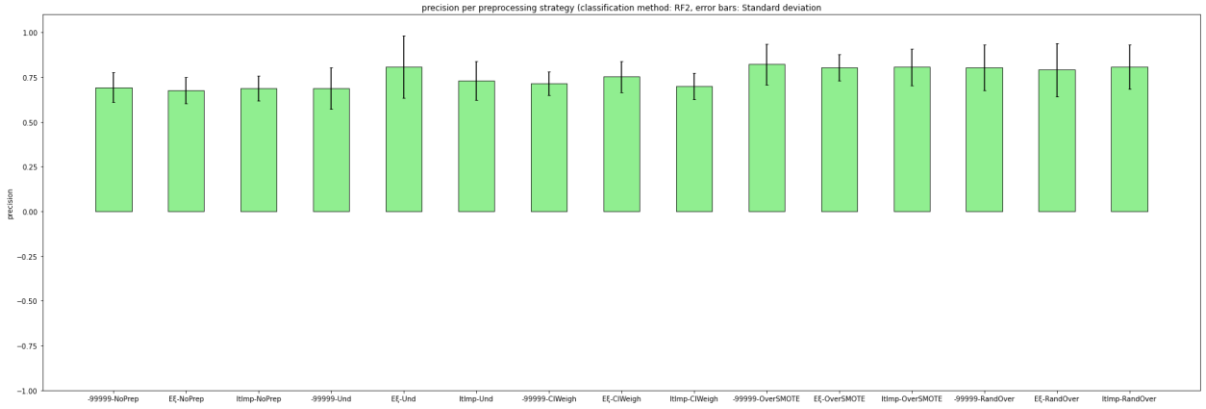

precision per preprocessing strategy (classification method: DT1, error bars: Standard deviation
precision
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

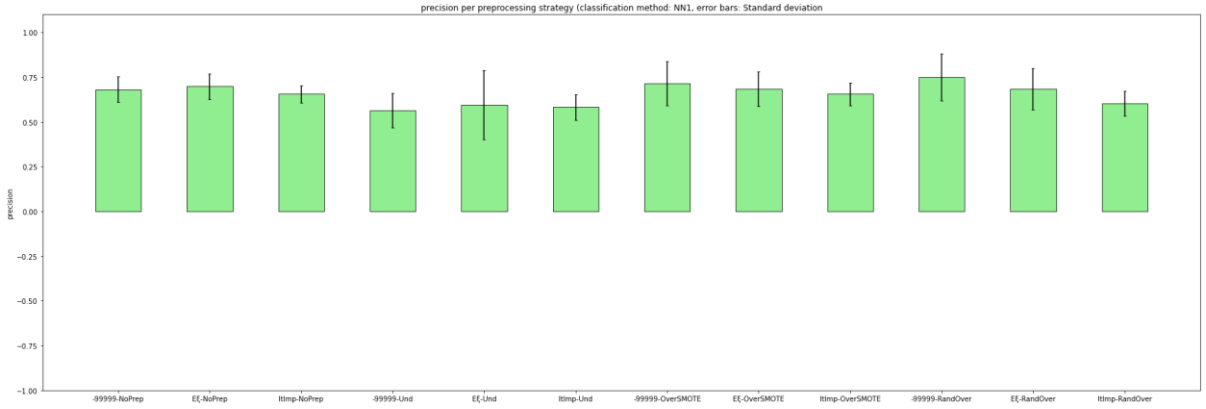

precision per preprocessing strategy (classification method: DT2, error bars: Standard deviation
precision
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

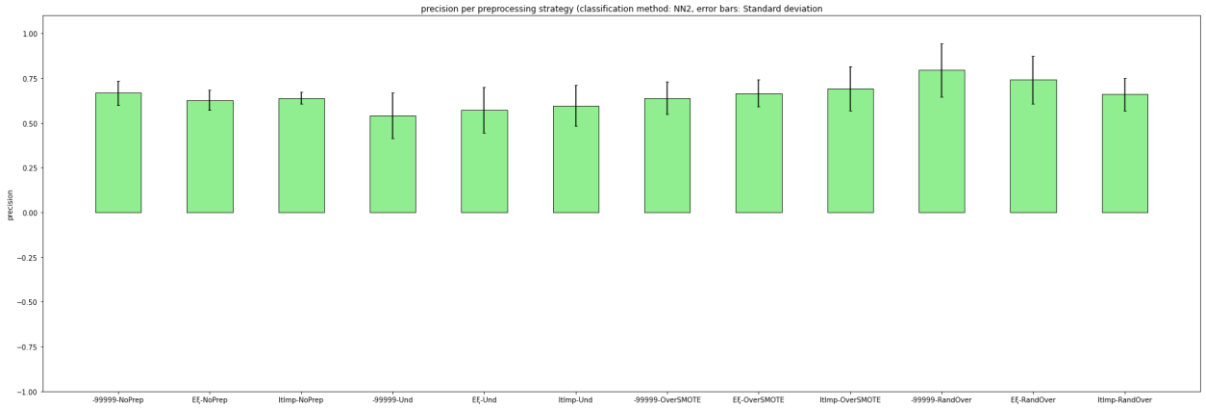

precision per preprocessing strategy (classification method: DT3, error bars: Standard deviation
precision
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

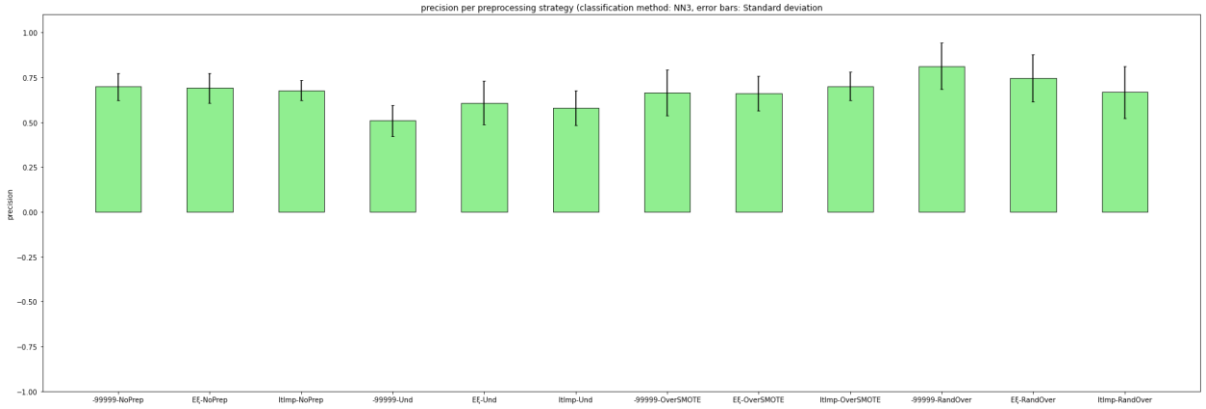

precision per preprocessing strategy (classification method: DT4, error bars: Standard deviation
precision
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

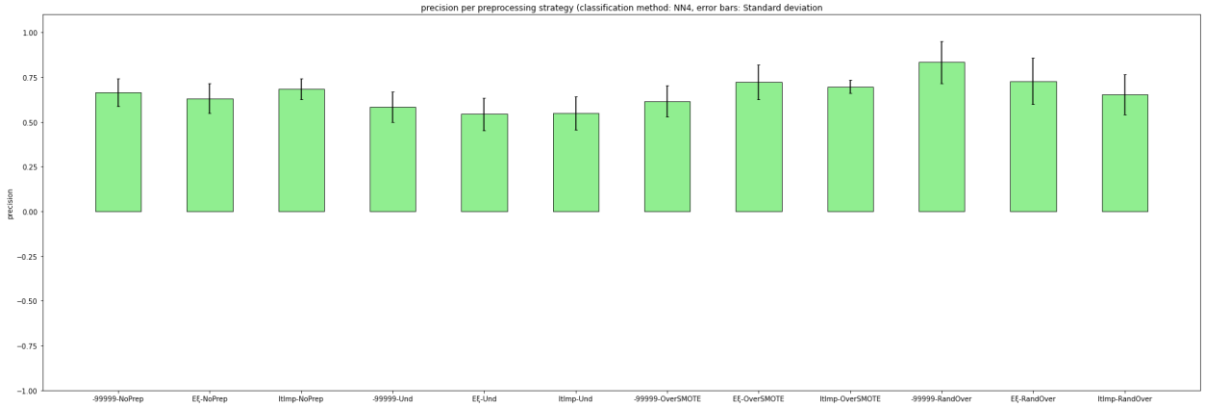

precision per preprocessing strategy (classification method: RF1, error bars: Standard deviation
precision
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

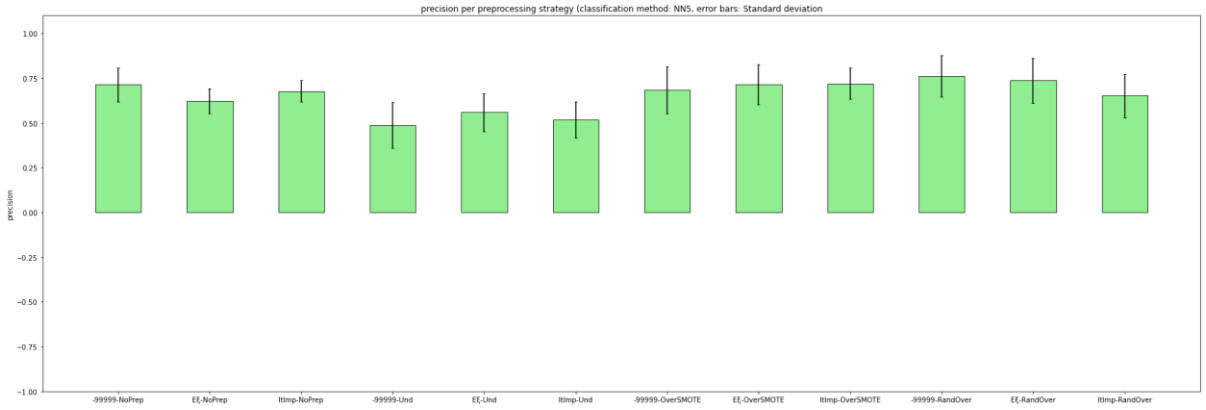

precision per preprocessing strategy (classification method: RF2, error bars: Standard deviation
precision
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

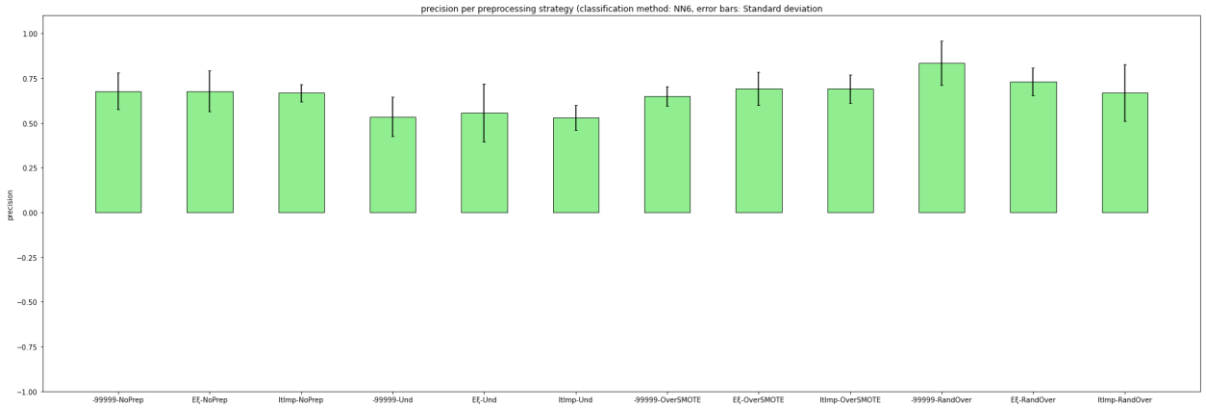

precision per preprocessing strategy (classification method: NN1, error bars: Standard deviation
precision
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

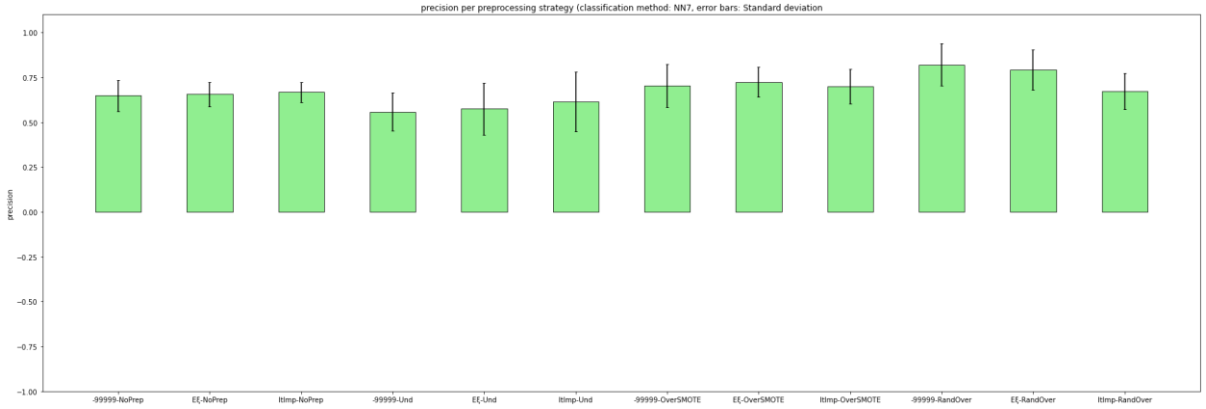

precision per preprocessing strategy (classification method: NN2, error bars: Standard deviation
precision
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

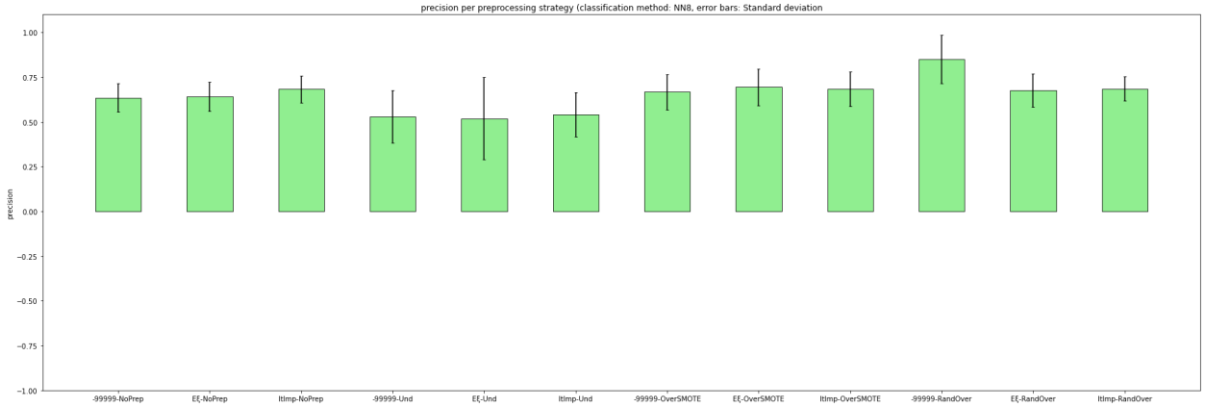

precision per preprocessing strategy (classification method: NN3, error bars: Standard deviation
precision
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

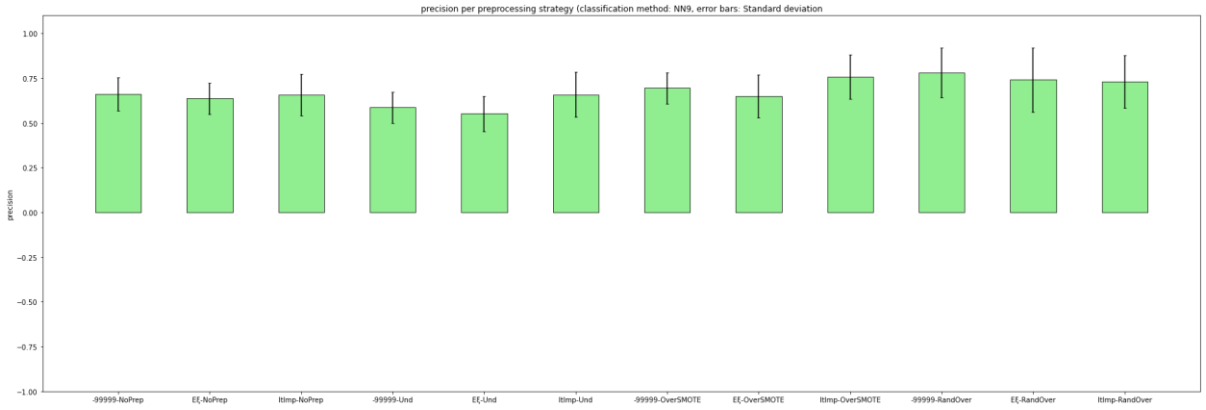

precision per preprocessing strategy (classification method: NN4, error bars: Standard deviation
precision
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

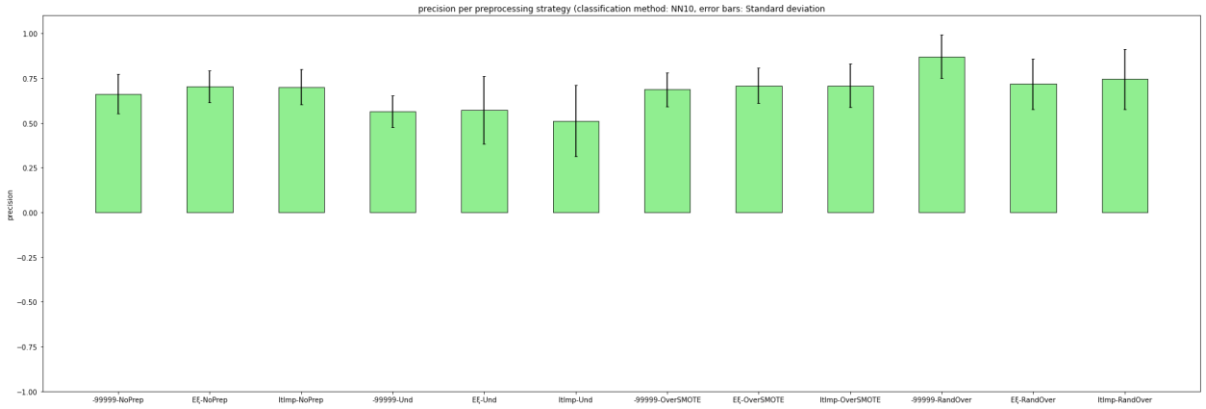

precision per preprocessing strategy (classification method: NN5, error bars: Standard deviation
precision
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

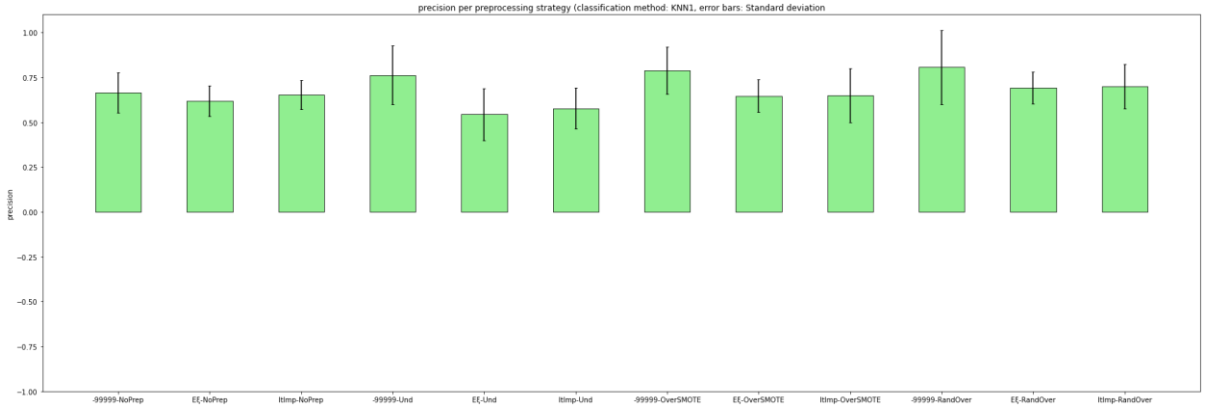

precision per preprocessing strategy (classification method: NN6, error bars: Standard deviation
precision
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

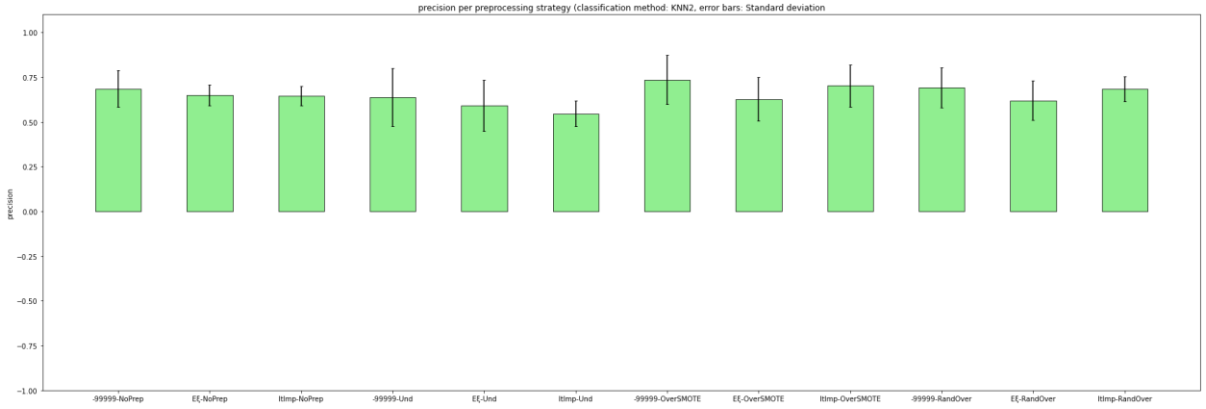

precision per preprocessing strategy (classification method: NN7, error bars: Standard deviation
precision
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

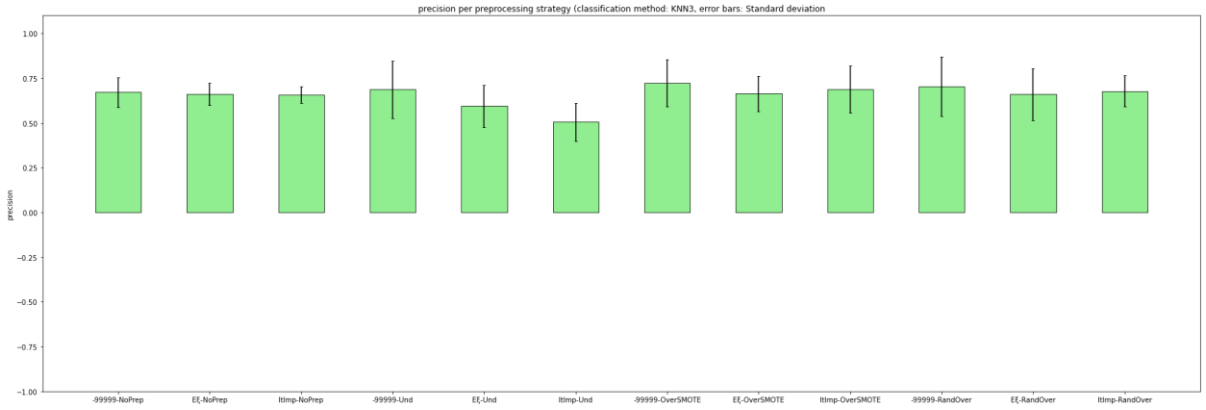

precision per preprocessing strategy (classification method: NN8, error bars: Standard deviation
precision
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

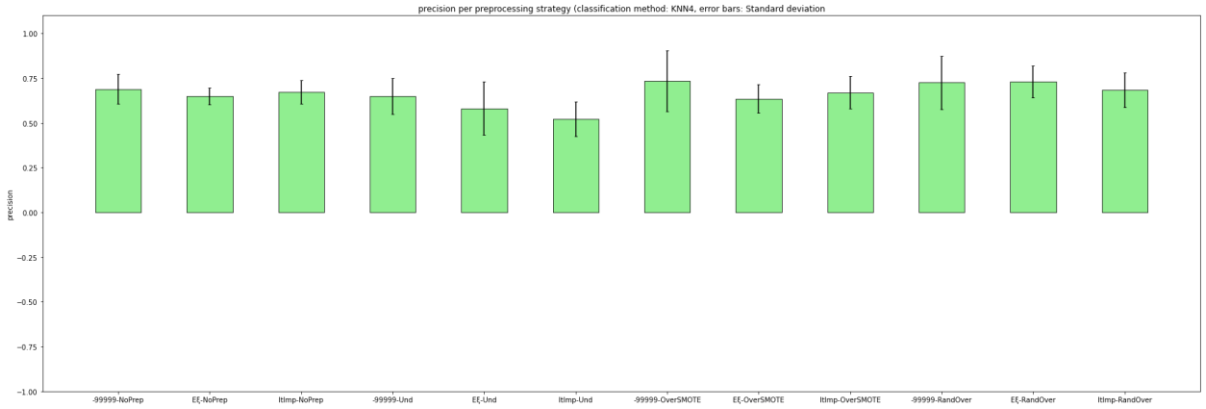

precision per preprocessing strategy (classification method: NN9, error bars: Standard deviation
precision
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

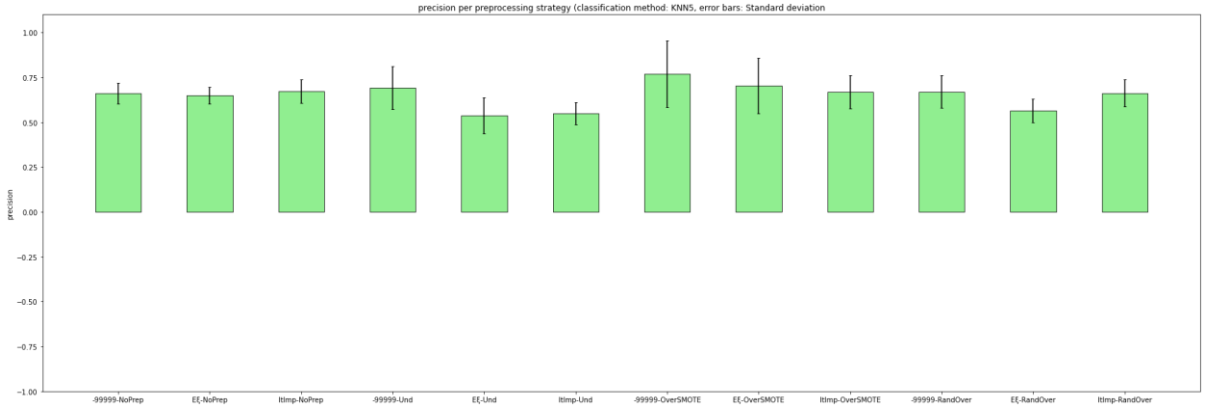

precision per preprocessing strategy (classification method: NN10, error bars: Standard deviation
precision
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

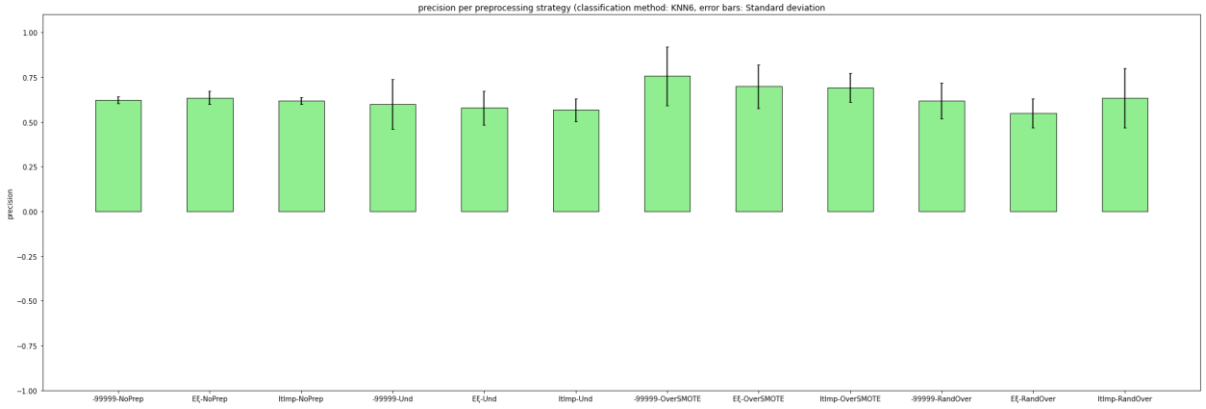

precision per preprocessing strategy (classification method: KNN1, error bars: Standard deviation
precision
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

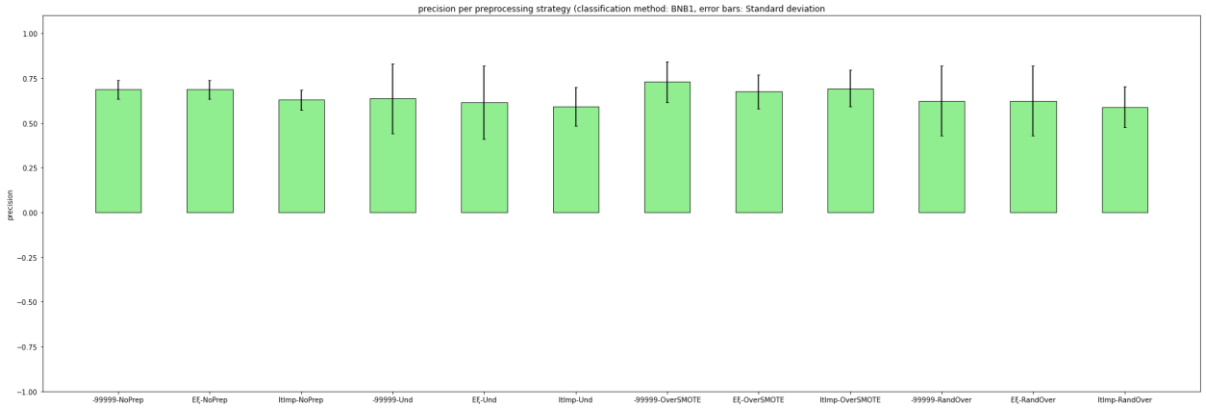

precision per preprocessing strategy (classification method: KNN2, error bars: Standard deviation
precision
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

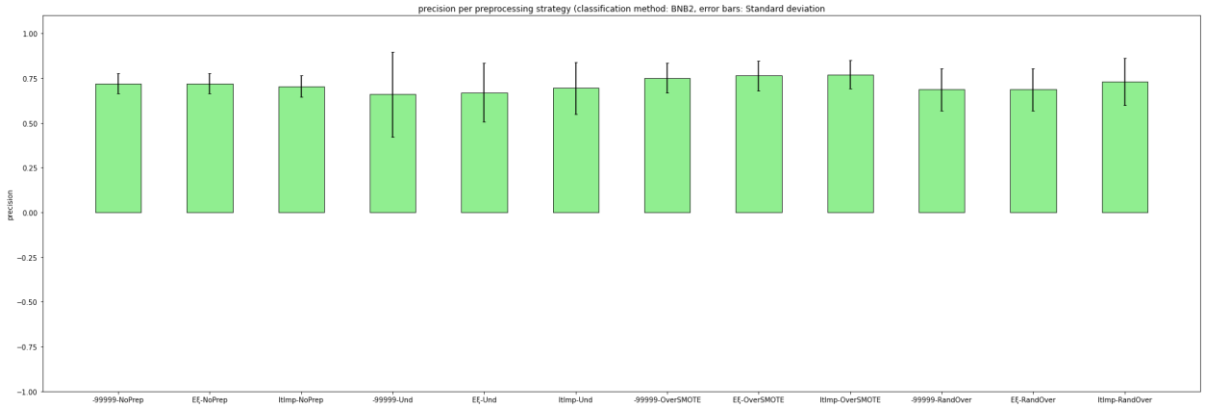

precision per preprocessing strategy (classification method: KNN3, error bars: Standard deviation
precision
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

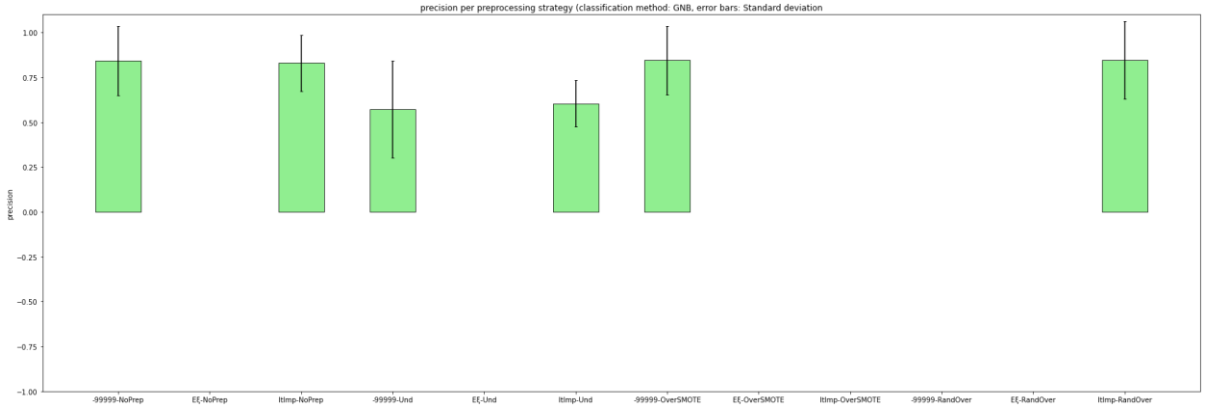

precision per preprocessing strategy (classification method: KNN4, error bars: Standard deviation
precision
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

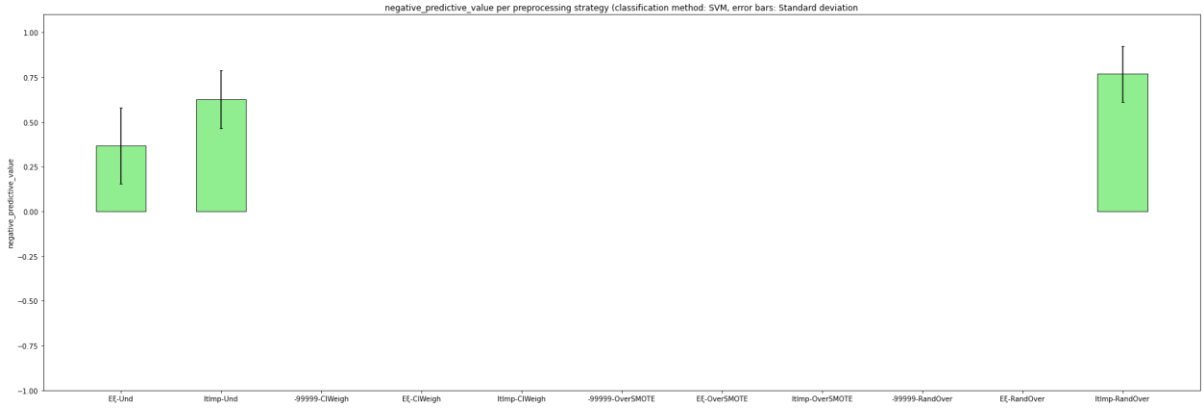

precision per preprocessing strategy (classification method: KNN5, error bars: Standard deviation
precision
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

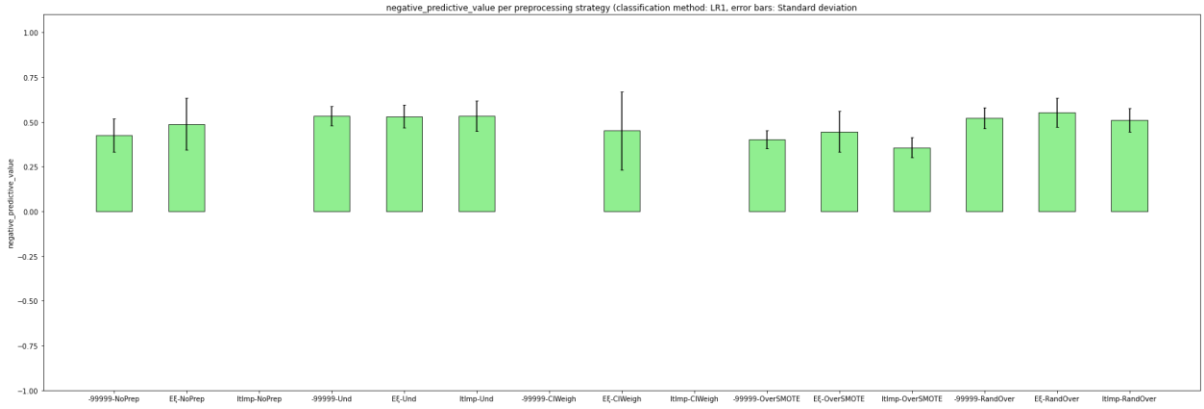

precision per preprocessing strategy (classification method: KNN6, error bars: Standard deviation
precision
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

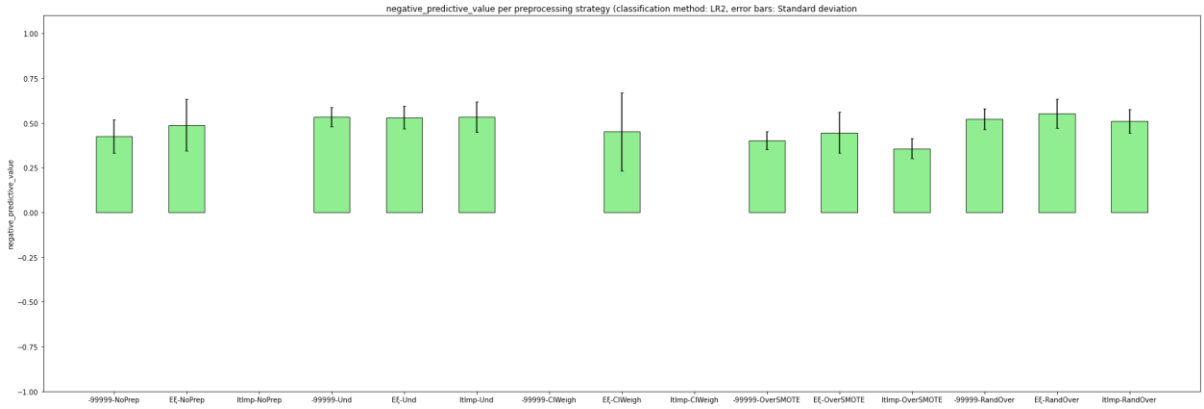

precision per preprocessing strategy (classification method: BNB1, error bars: Standard deviation
precision
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

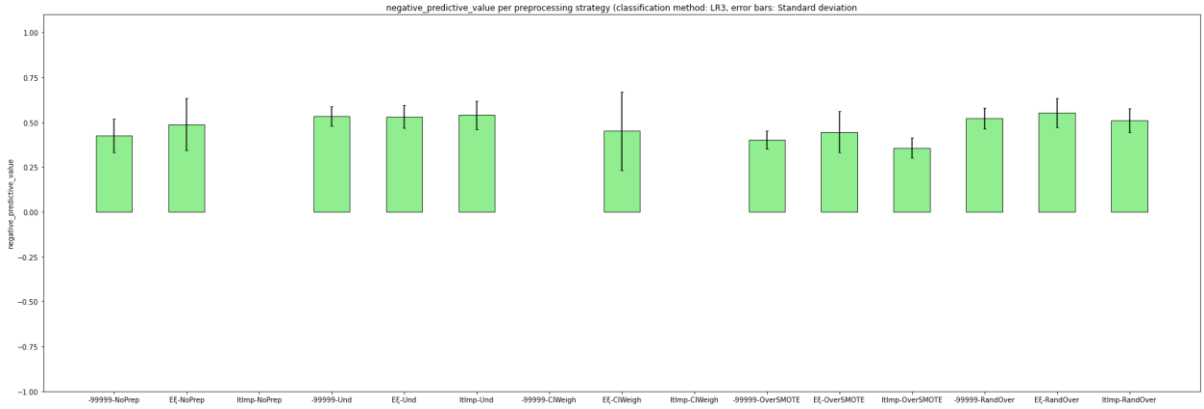

precision per preprocessing strategy (classification method: BNB2, error bars: Standard deviation
precision
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

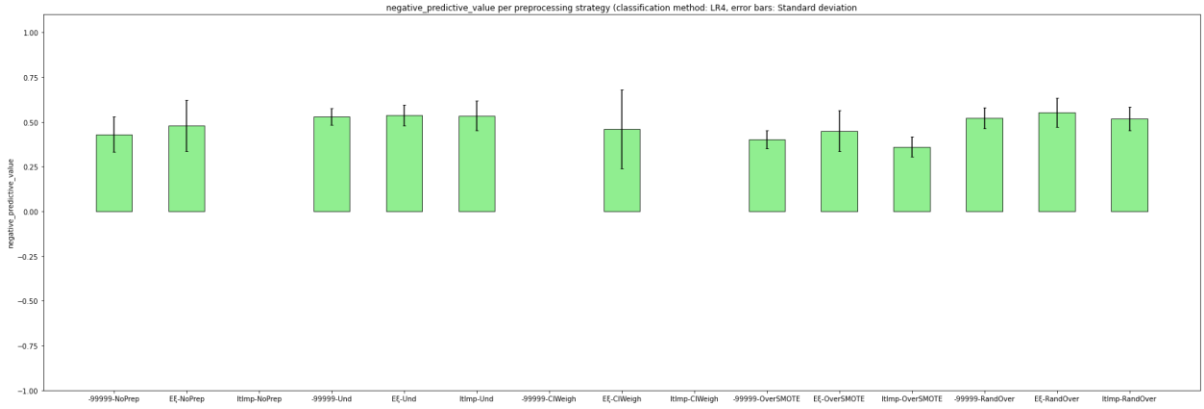

precision per preprocessing strategy (classification method: GNB, error bars: Standard deviation
precision
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

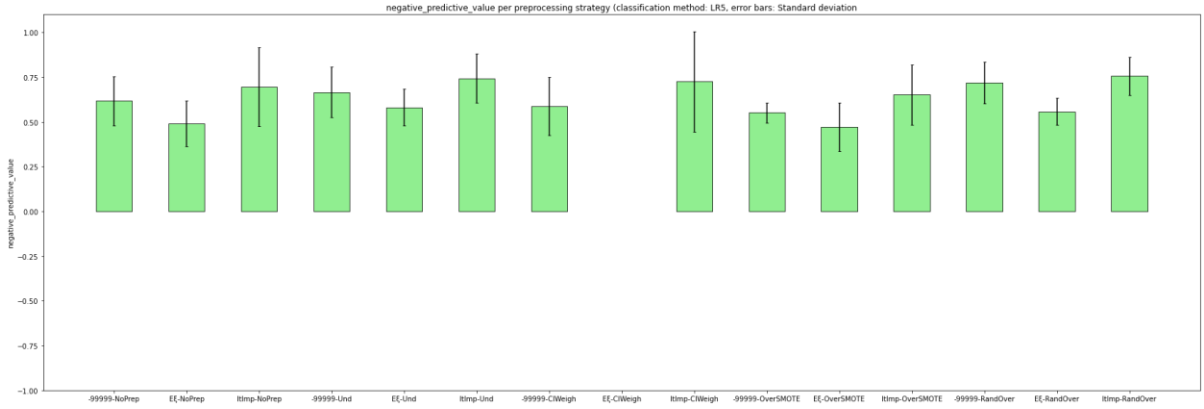

negative_predictive_value per preprocessing strategy (classification method: SVM, error bars: Standard deviation
negative_predictive_value
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

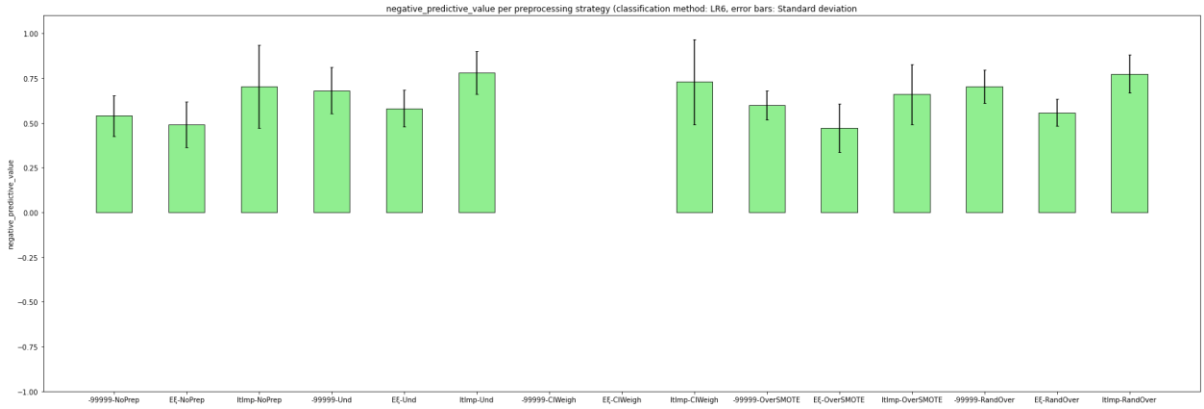

negative_predictive_value per preprocessing strategy (classification method: LR1, error bars: Standard deviation
negative_predictive_value
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

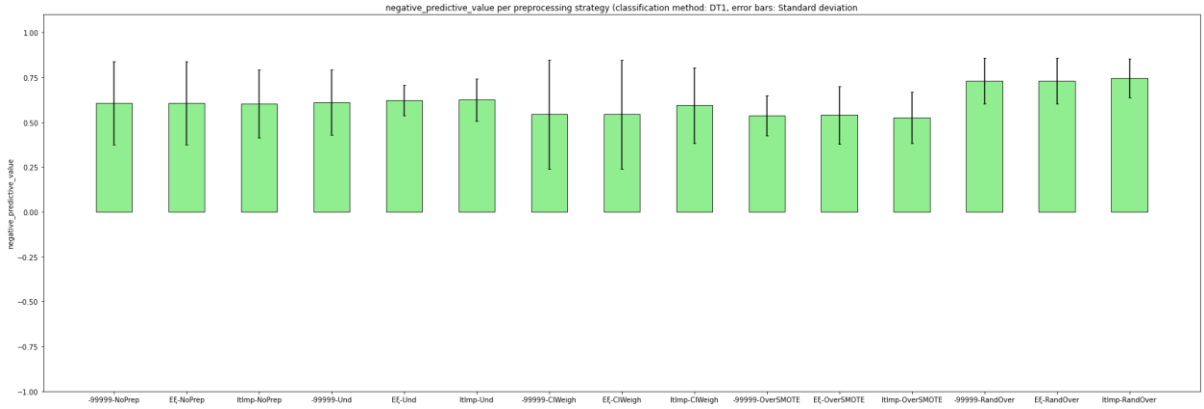

negative_predictive_value per preprocessing strategy (classification method: LR2, error bars: Standard deviation
negative_predictive_value
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

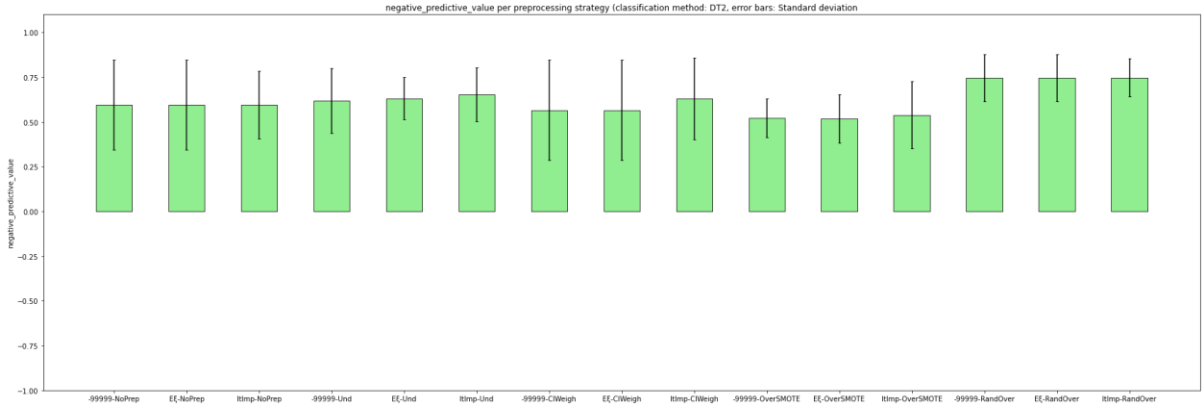

negative_predictive_value per preprocessing strategy (classification method: LR3, error bars: Standard deviation
negative_predictive_value
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

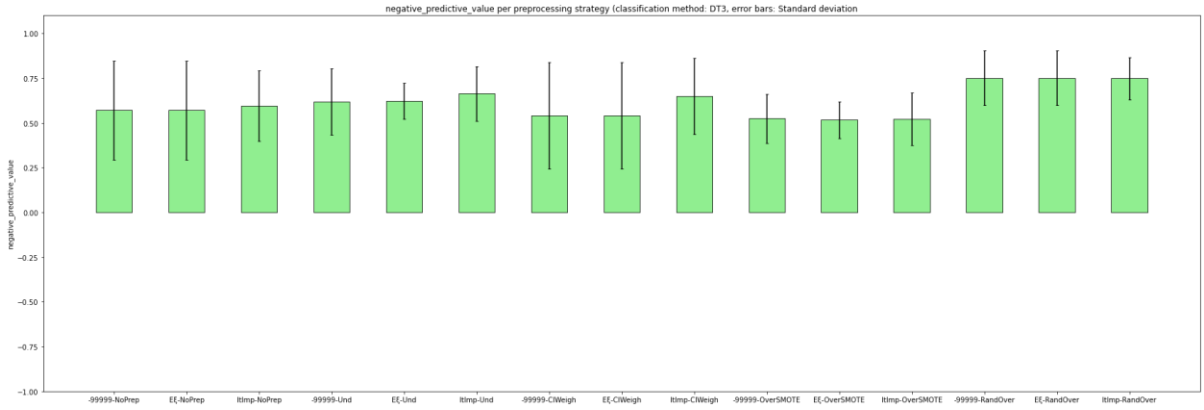

negative_predictive_value per preprocessing strategy (classification method: LR4, error bars: Standard deviation
negative_predictive_value
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

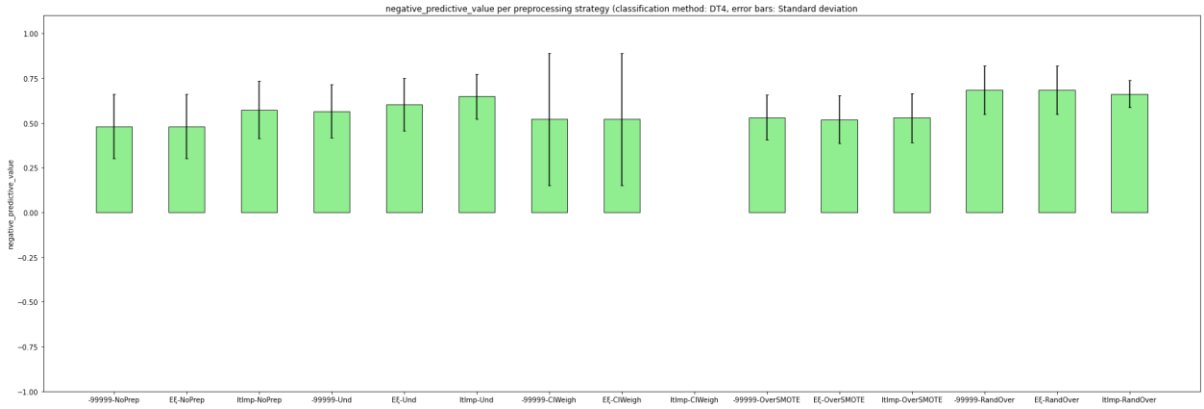

negative_predictive_value per preprocessing strategy (classification method: LR5, error bars: Standard deviation
negative_predictive_value
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

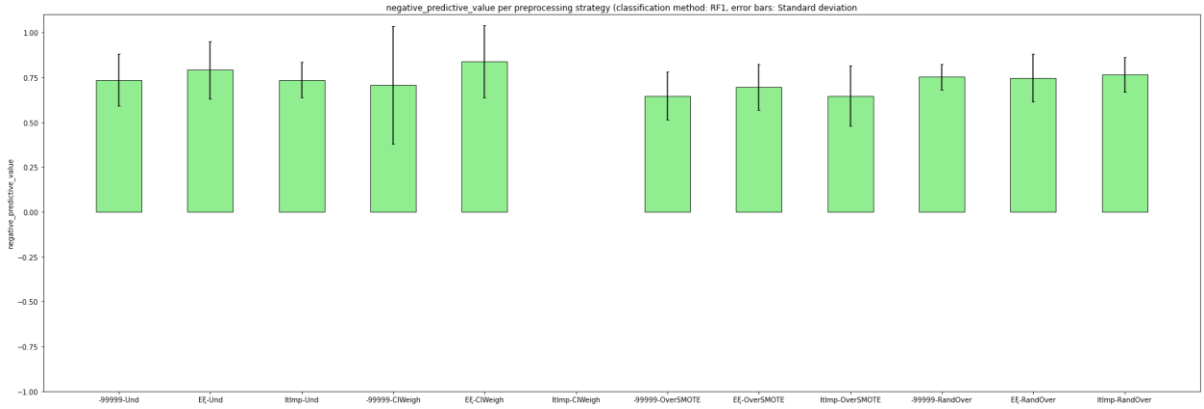

negative_predictive_value per preprocessing strategy (classification method: LR6, error bars: Standard deviation
negative_predictive_value
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

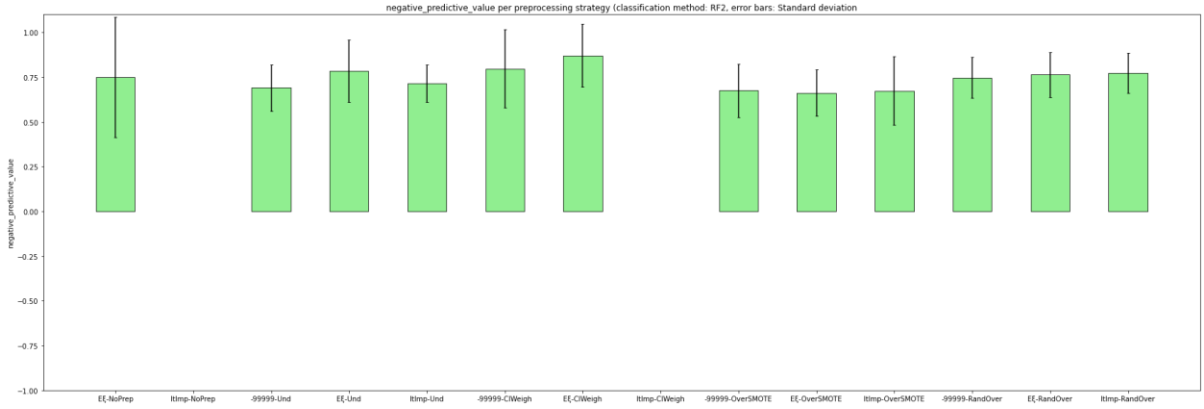

negative_predictive_value per preprocessing strategy (classification method: DT1, error bars: Standard deviation
negative_predictive_value
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

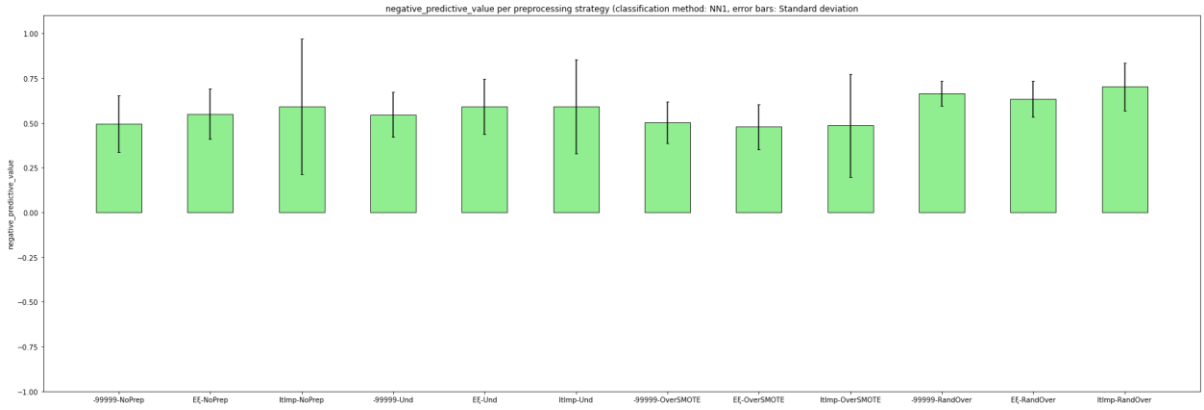

negative_predictive_value per preprocessing strategy (classification method: DT2, error bars: Standard deviation
negative_predictive_value
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-ClWeigh
EE-ClWeigh
ItImp-ClWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

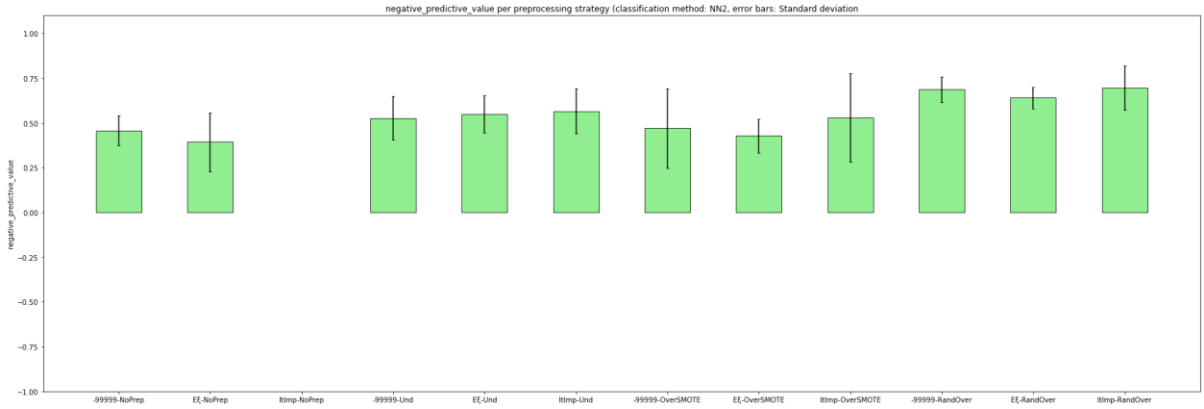

negative_predictive_value per preprocessing strategy (classification method: DT3, error bars: Standard deviation
negative_predictive_value
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-ClWeigh
EE-ClWeigh
ItImp-ClWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

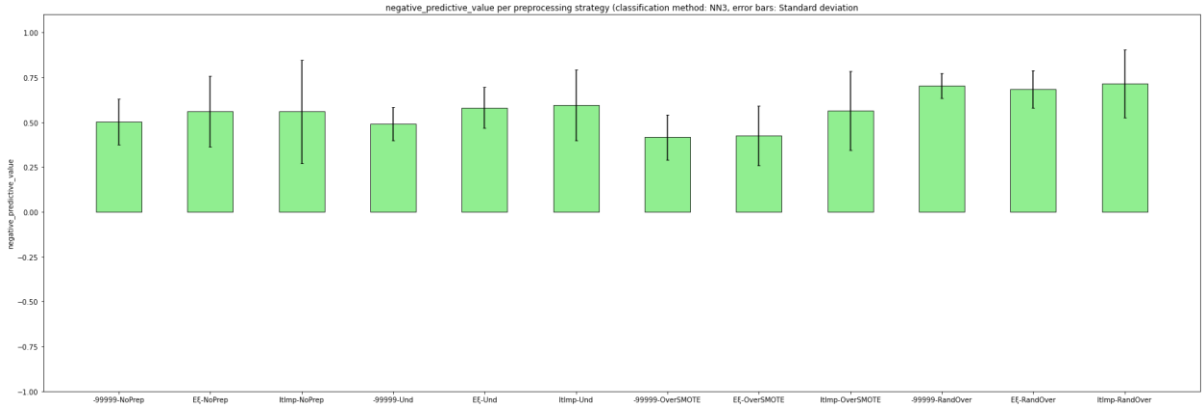

negative_predictive_value per preprocessing strategy (classification method: DT4, error bars: Standard deviation
negative_predictive_value
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-ClWeigh
EE-ClWeigh
ItImp-ClWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

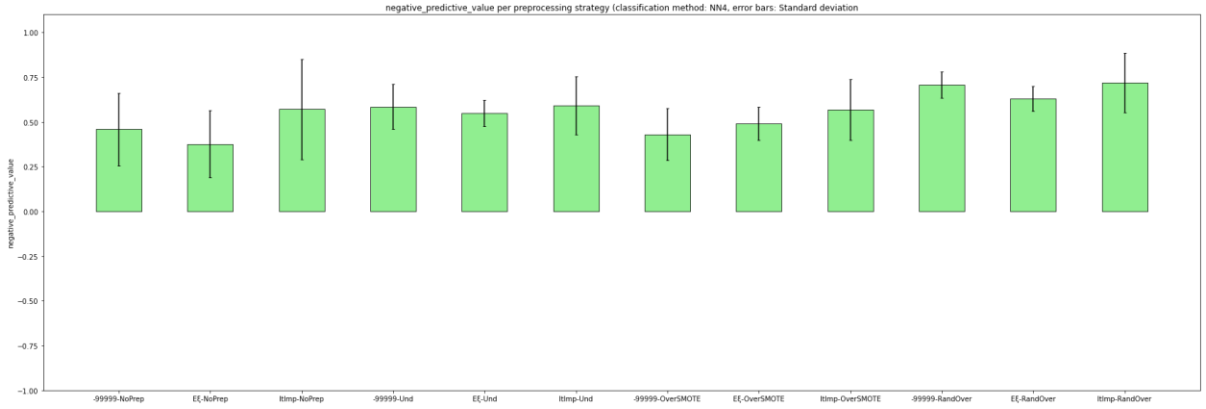

negative_predictive_value per preprocessing strategy (classification method: RF1, error bars: Standard deviation
negative_predictive_value
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-Und
EE-Und
ItImp-Und
-99999-ClWeigh
EE-ClWeigh
ItImp-ClWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

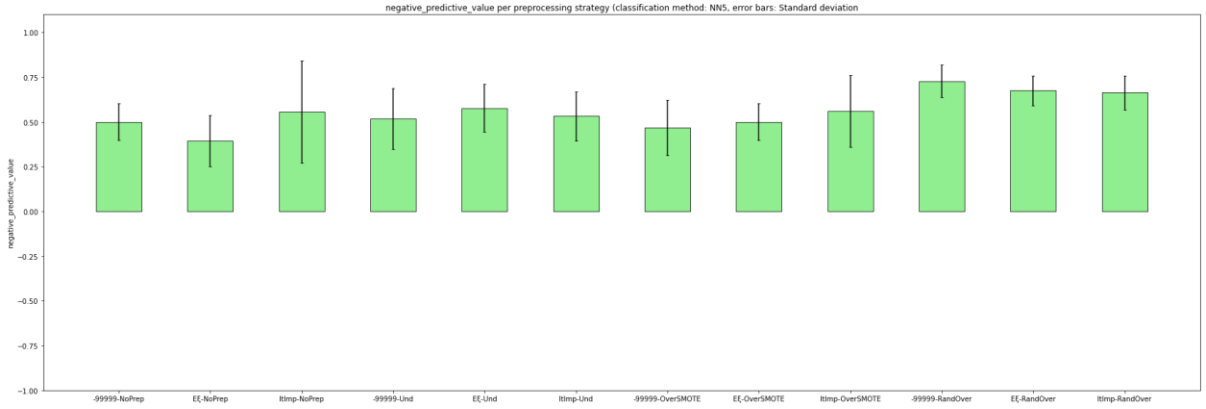

negative_predictive_value per preprocessing strategy (classification method: RF2, error bars: Standard deviation
negative_predictive_value
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-ClWeigh
EE-ClWeigh
ItImp-ClWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

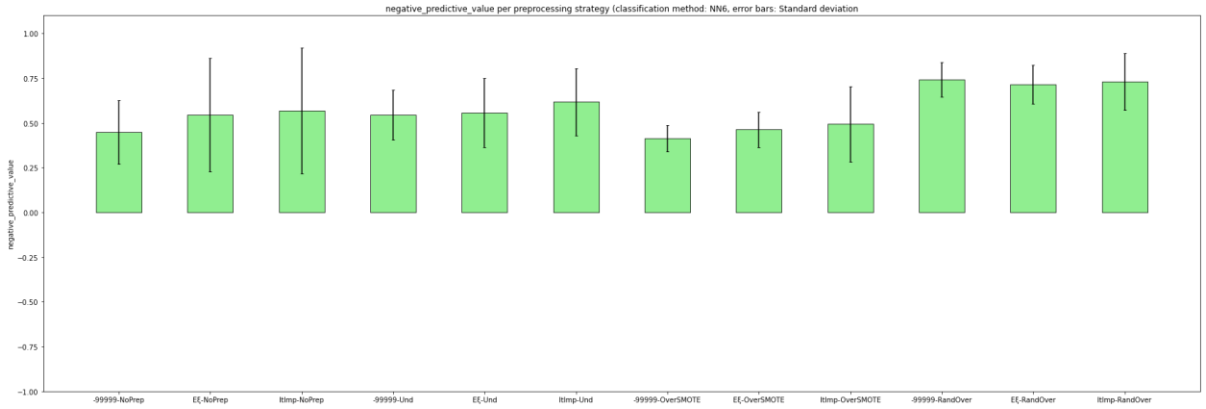

negative_predictive_value per preprocessing strategy (classification method: NN1, error bars: Standard deviation
negative_predictive_value
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

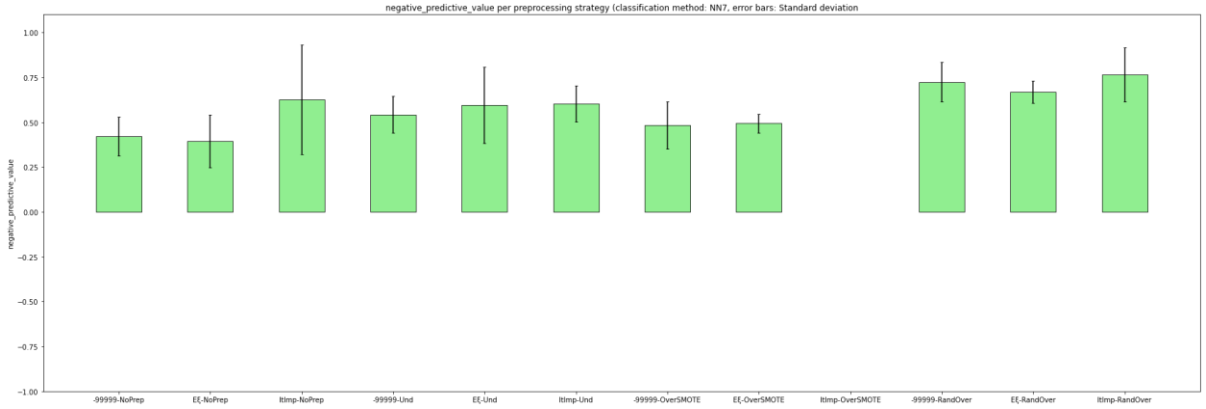

negative_predictive_value per preprocessing strategy (classification method: NN2, error bars: Standard deviation
negative_predictive_value
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

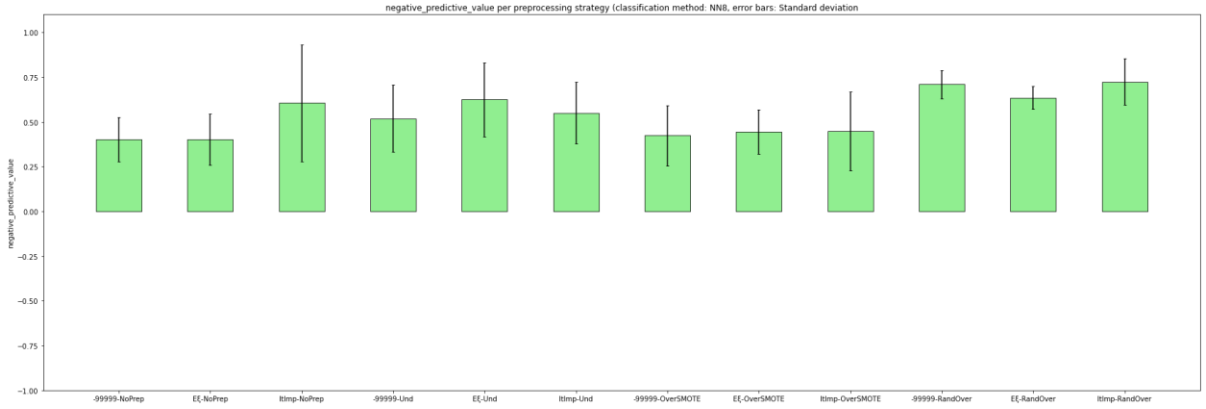

negative_predictive_value per preprocessing strategy (classification method: NN3, error bars: Standard deviation
negative_predictive_value
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

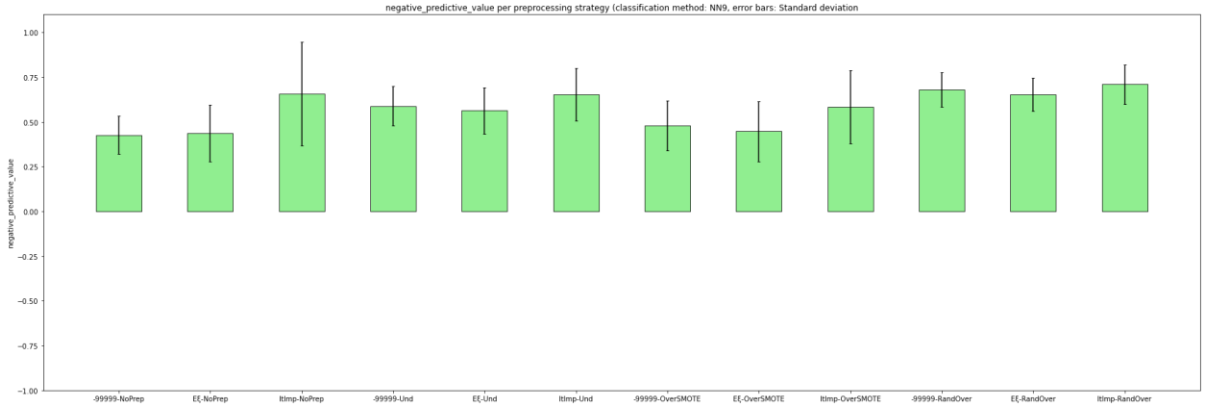

negative_predictive_value per preprocessing strategy (classification method: NN4, error bars: Standard deviation
negative_predictive_value
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

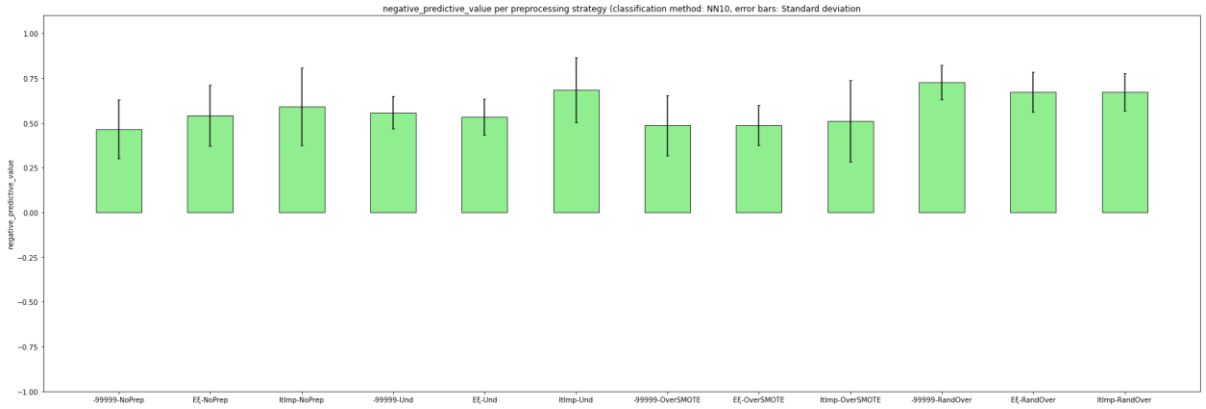

negative_predictive_value per preprocessing strategy (classification method: NN5, error bars: Standard deviation
negative_predictive_value
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

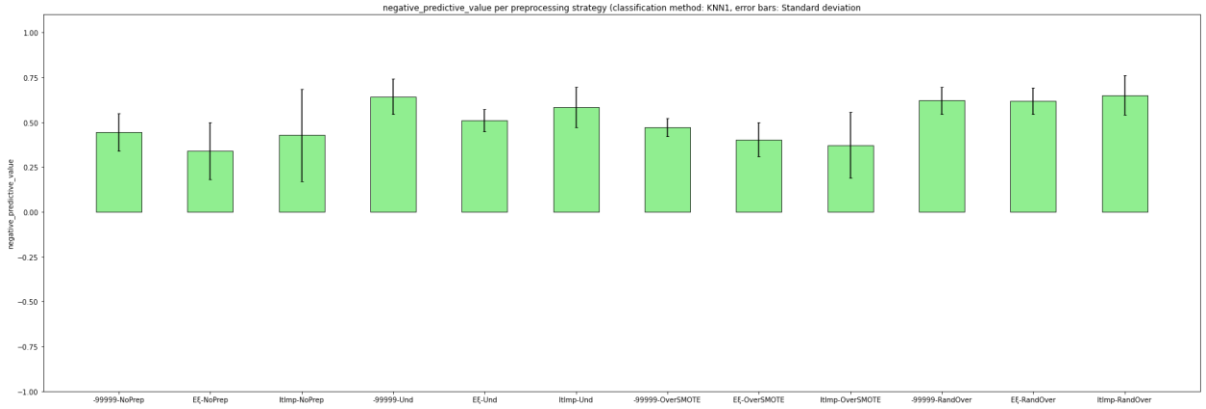

negative_predictive_value per preprocessing strategy (classification method: NN6, error bars: Standard deviation
negative_predictive_value
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

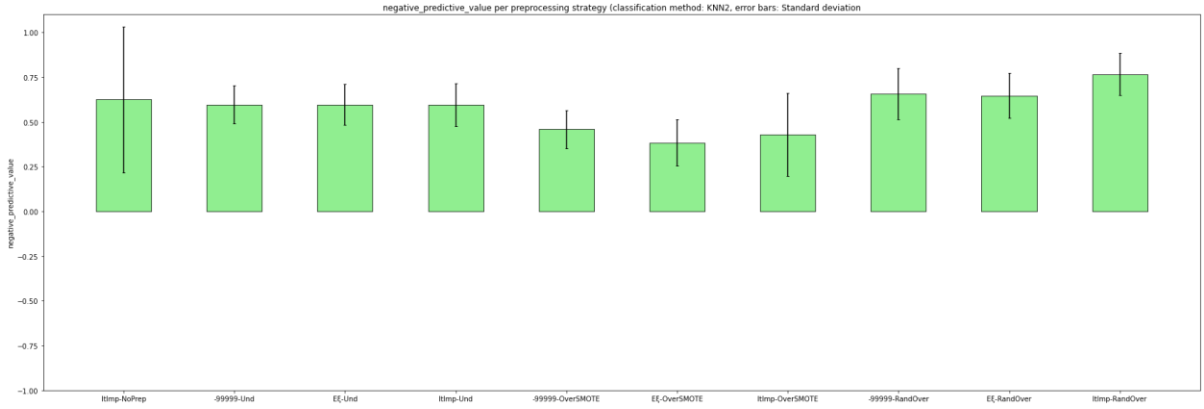

negative_predictive_value per preprocessing strategy (classification method: NN7, error bars: Standard deviation
negative_predictive_value
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

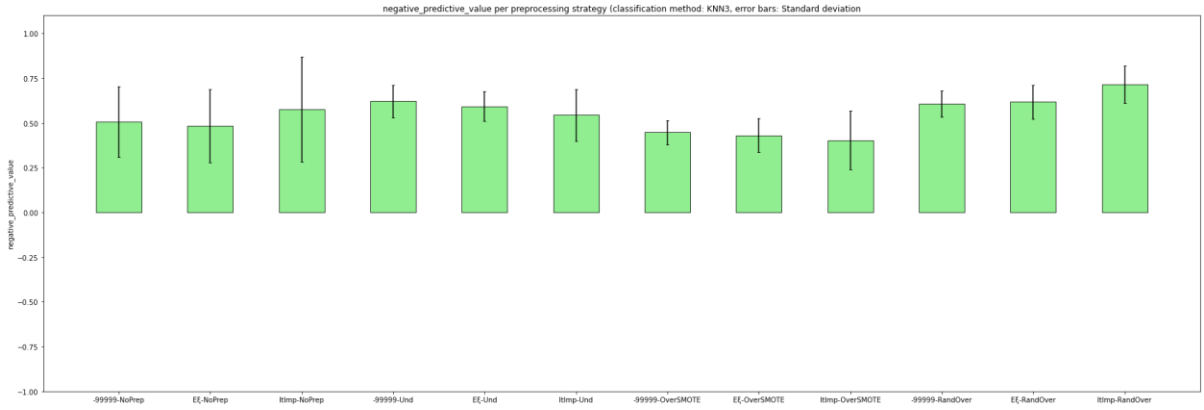

negative_predictive_value per preprocessing strategy (classification method: NN8, error bars: Standard deviation
negative_predictive_value
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

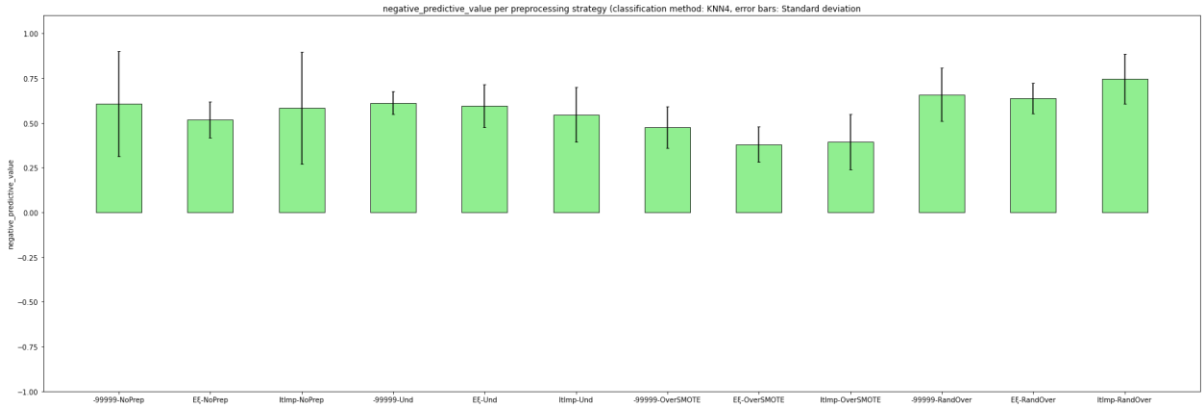

negative_predictive_value per preprocessing strategy (classification method: NN9, error bars: Standard deviation
negative_predictive_value
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

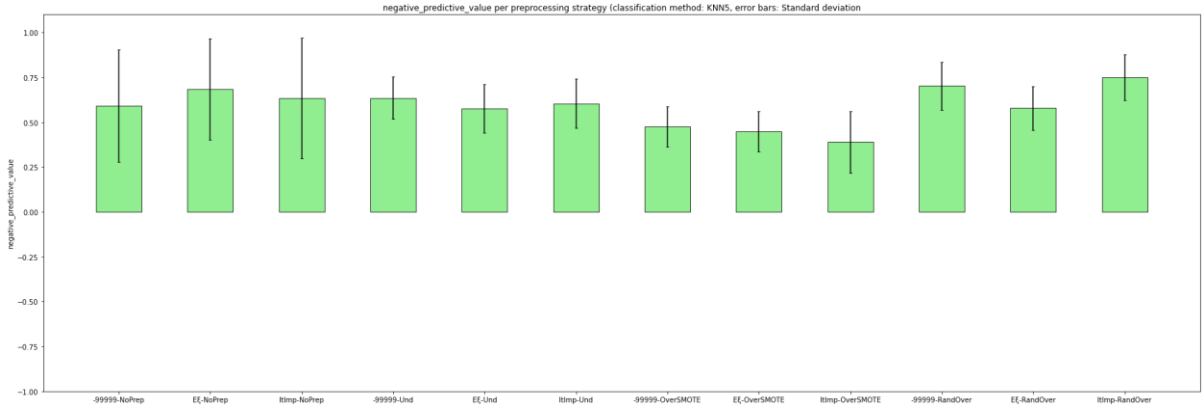

negative_predictive_value per preprocessing strategy (classification method: NN10, error bars: Standard deviation
negative_predictive_value
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

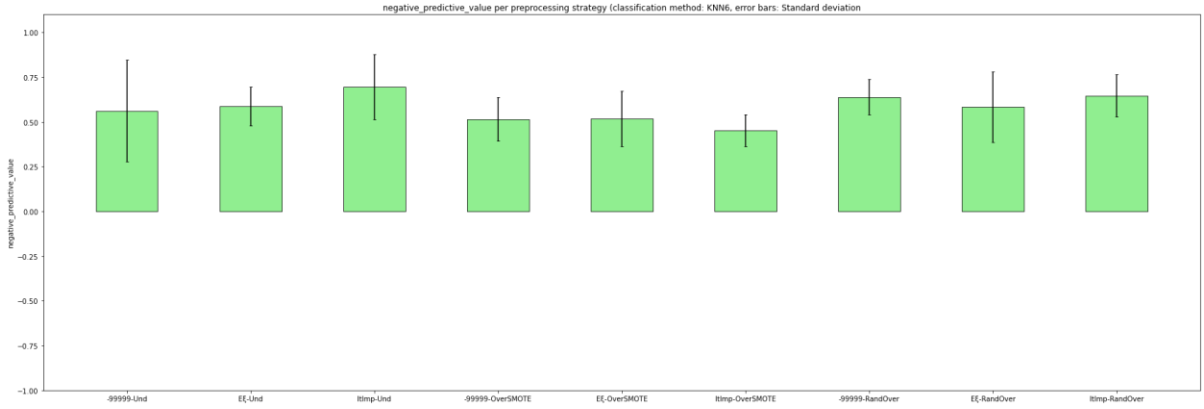

negative_predictive_value per preprocessing strategy (classification method: KNN1, error bars: Standard deviation
negative_predictive_value
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

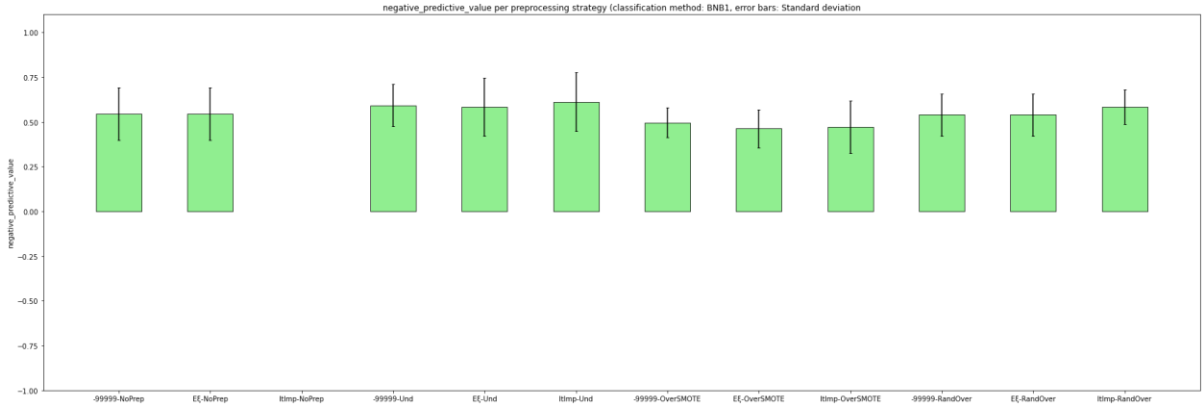

negative_predictive_value per preprocessing strategy (classification method: KNN2, error bars: Standard deviation
negative_predictive_value
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

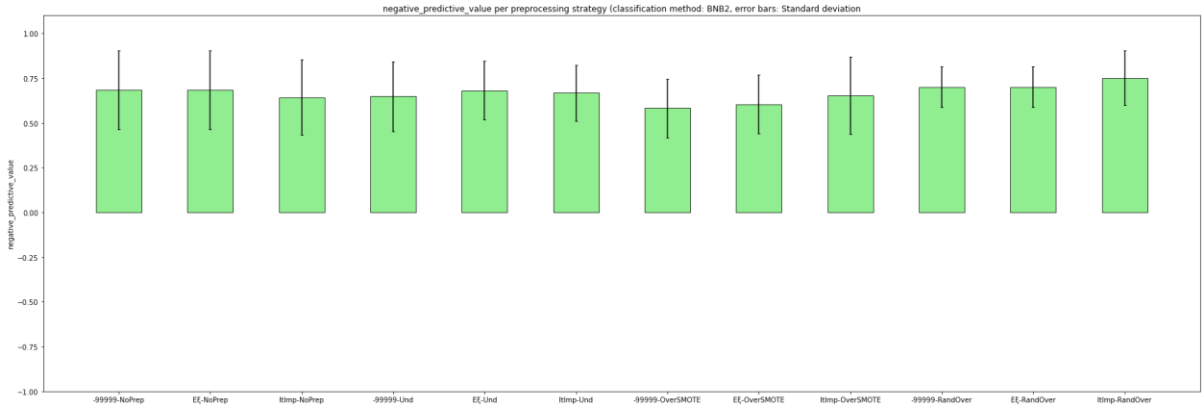

negative_predictive_value per preprocessing strategy (classification method: KNN3, error bars: Standard deviation
negative_predictive_value
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

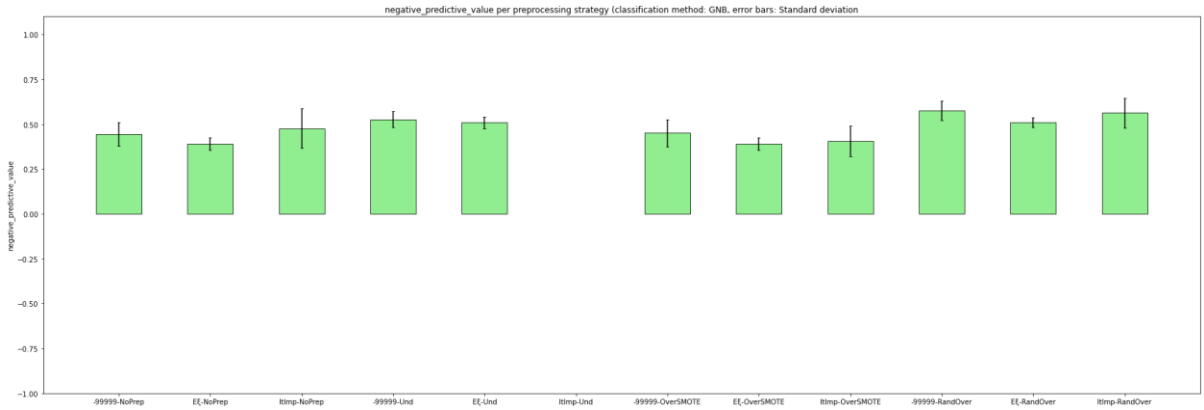

negative_predictive_value per preprocessing strategy (classification method: KNN4, error bars: Standard deviation
negative_predictive_value
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

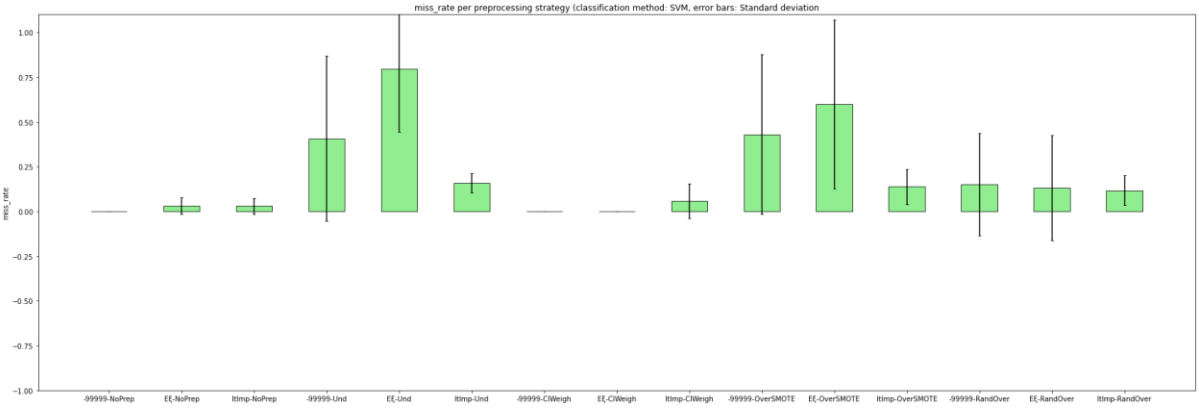

negative_predictive_value per preprocessing strategy (classification method: KNN5, error bars: Standard deviation
negative_predictive_value
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

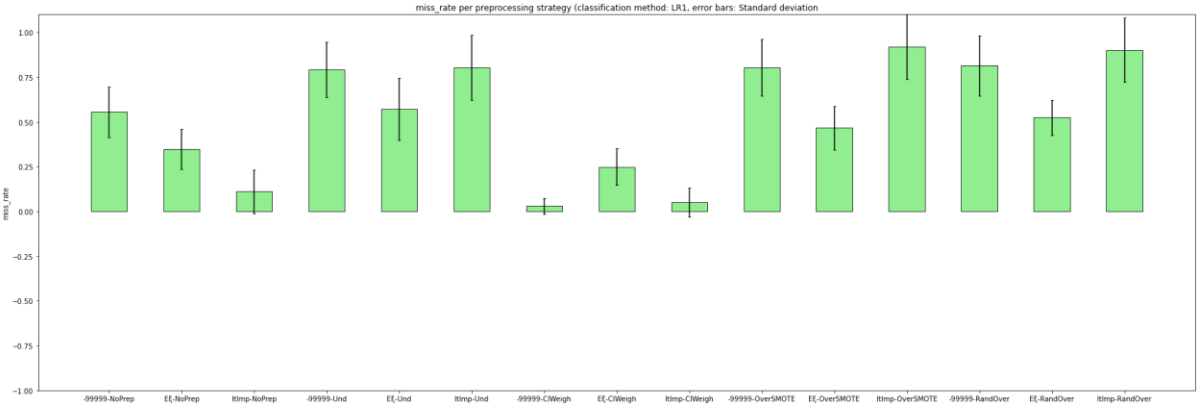

negative_predictive_value per preprocessing strategy (classification method: KNN6, error bars: Standard deviation
negative_predictive_value
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

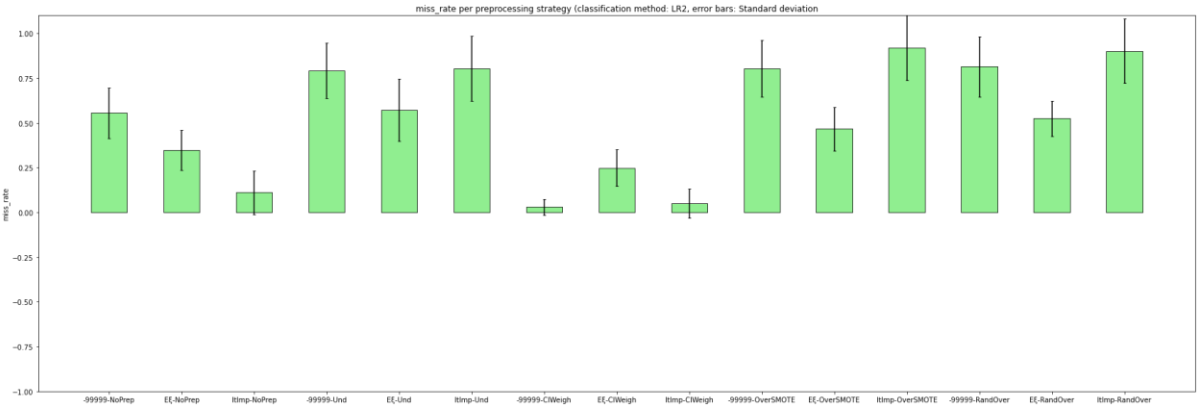

negative_predictive_value per preprocessing strategy (classification method: BNB1, error bars: Standard deviation
negative_predictive_value
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

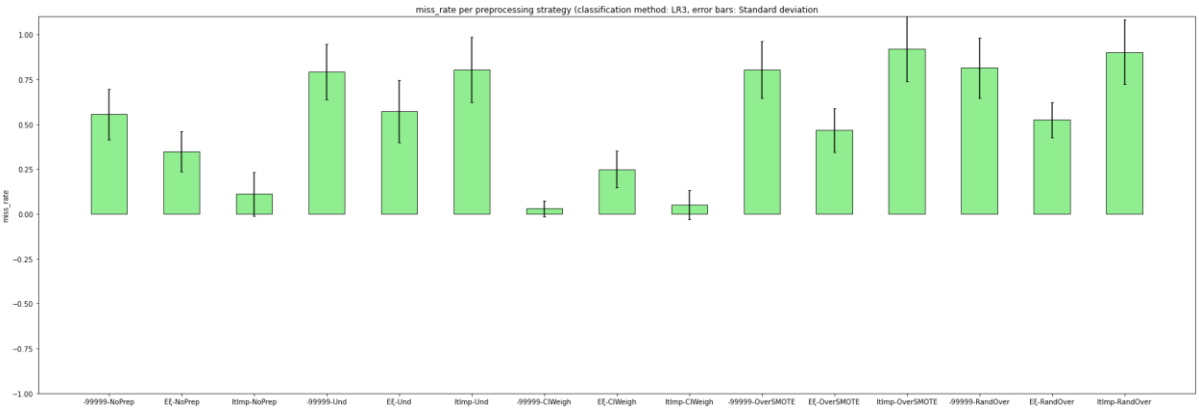

negative_predictive_value per preprocessing strategy (classification method: BNB2, error bars: Standard deviation
negative_predictive_value
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

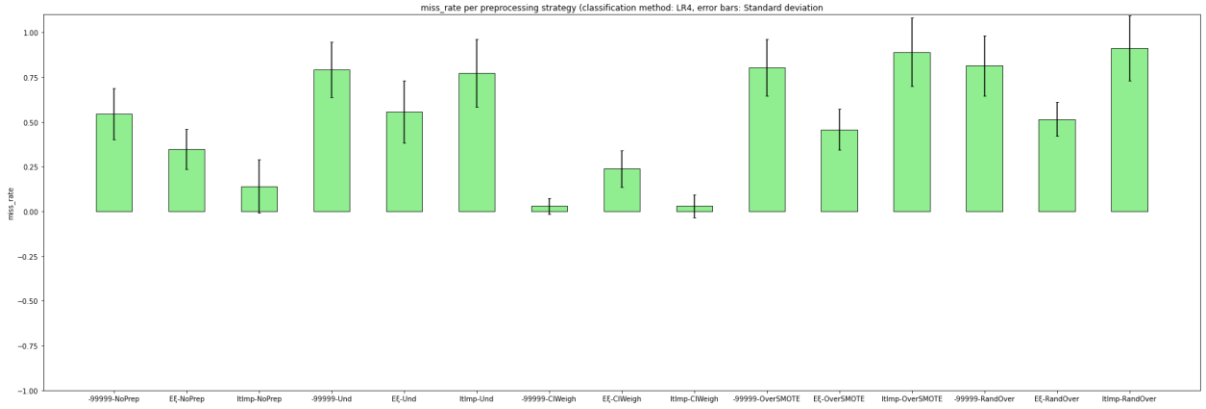

negative_predictive_value per preprocessing strategy (classification method: GNB, error bars: Standard deviation
negative_predictive_value
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

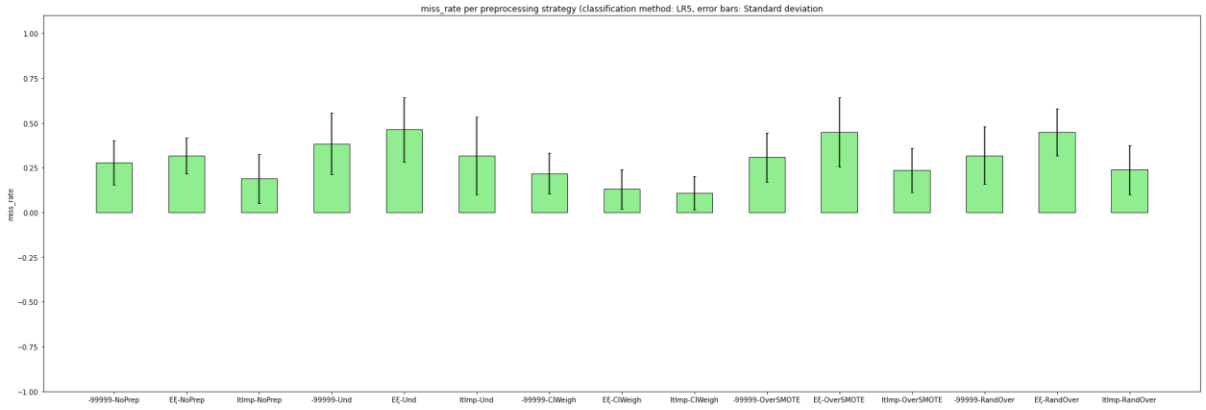

miss_rate per preprocessing strategy (classification method: SVM, error bars: Standard deviation
miss_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

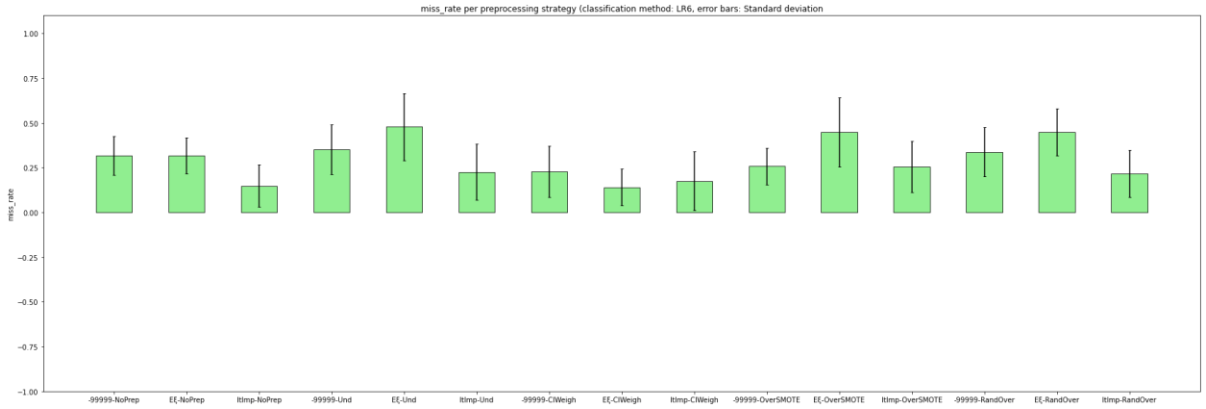

miss_rate per preprocessing strategy (classification method: LR1, error bars: Standard deviation
miss_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

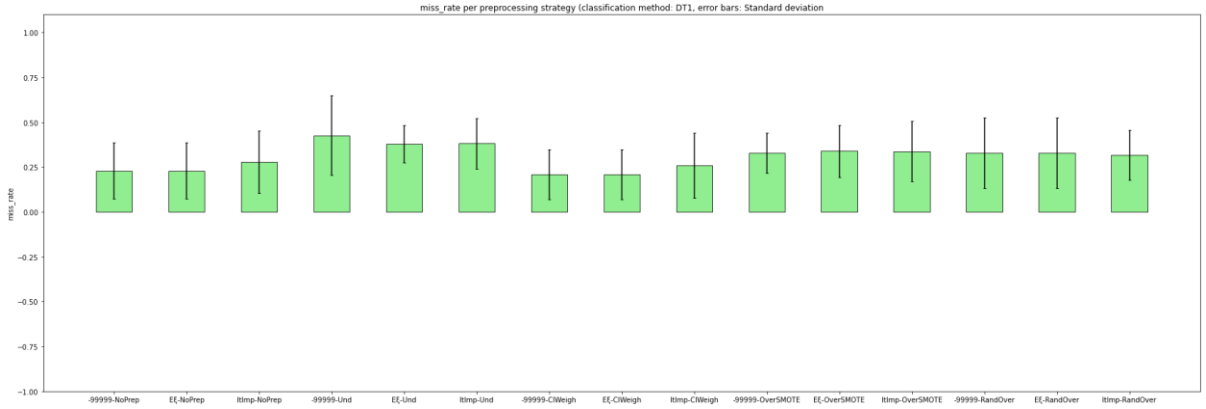

miss_rate per preprocessing strategy (classification method: LR2, error bars: Standard deviation
miss_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

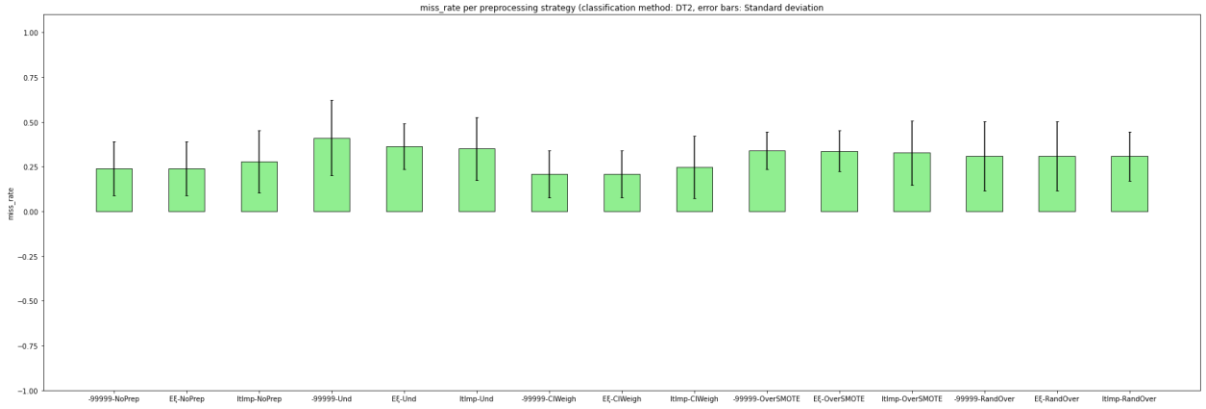

miss_rate per preprocessing strategy (classification method: LR3, error bars: Standard deviation
miss_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

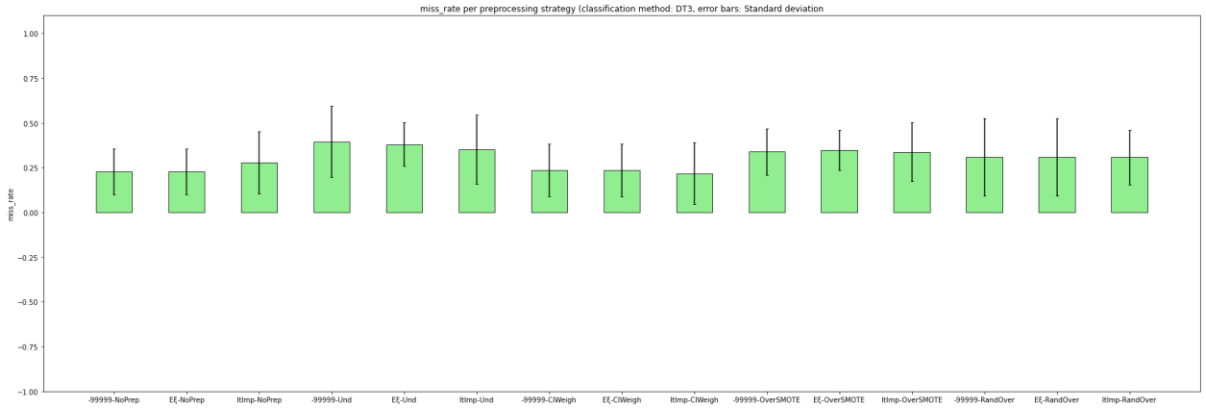

miss_rate per preprocessing strategy (classification method: LR4, error bars: Standard deviation
miss_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

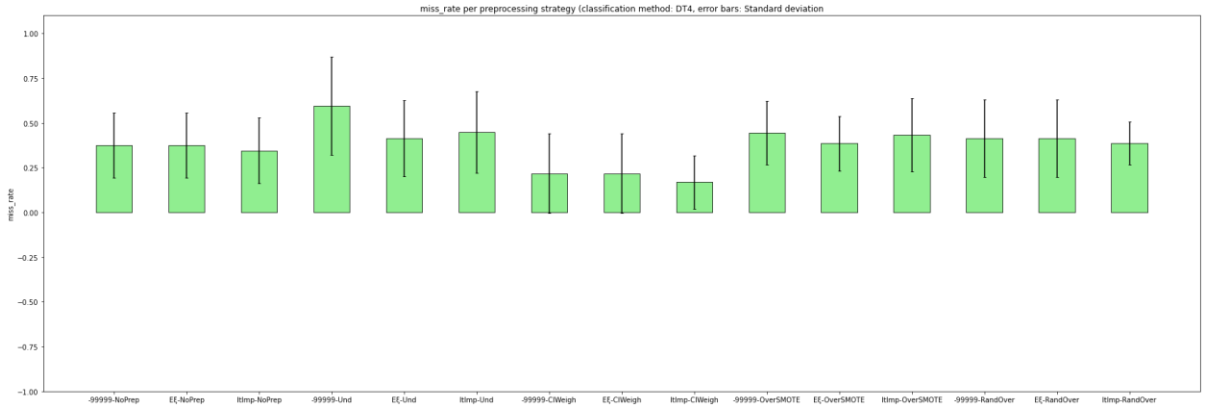

miss_rate per preprocessing strategy (classification method: LR5, error bars: Standard deviation
miss_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

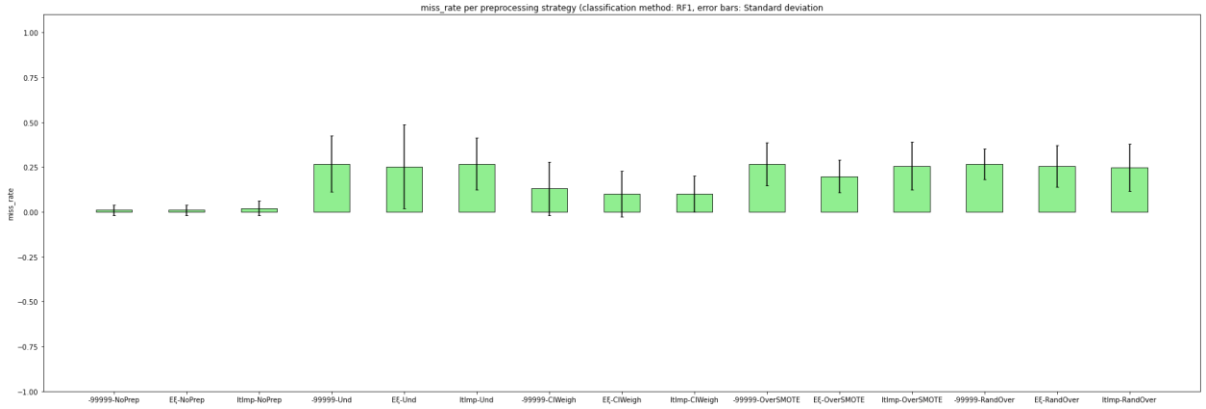

miss_rate per preprocessing strategy (classification method: LR6, error bars: Standard deviation
miss_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

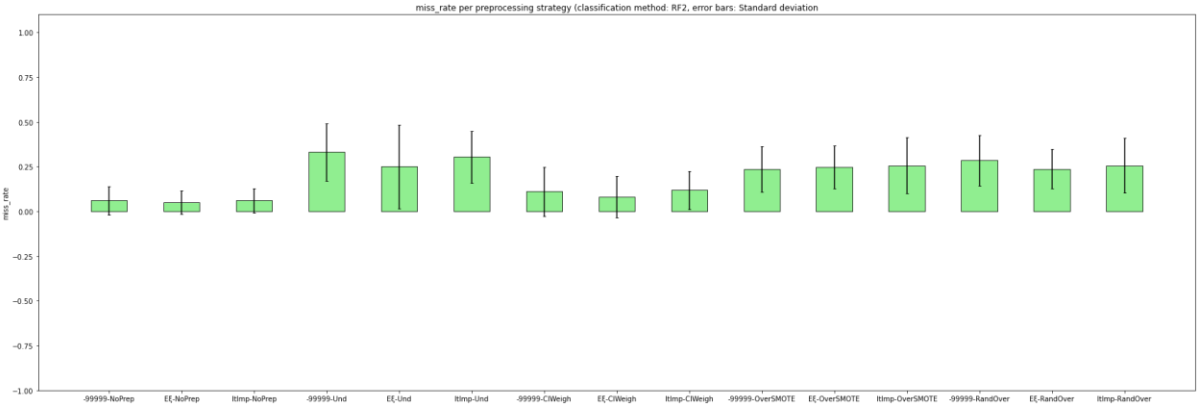

miss_rate per preprocessing strategy (classification method: DT1, error bars: Standard deviation
miss_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

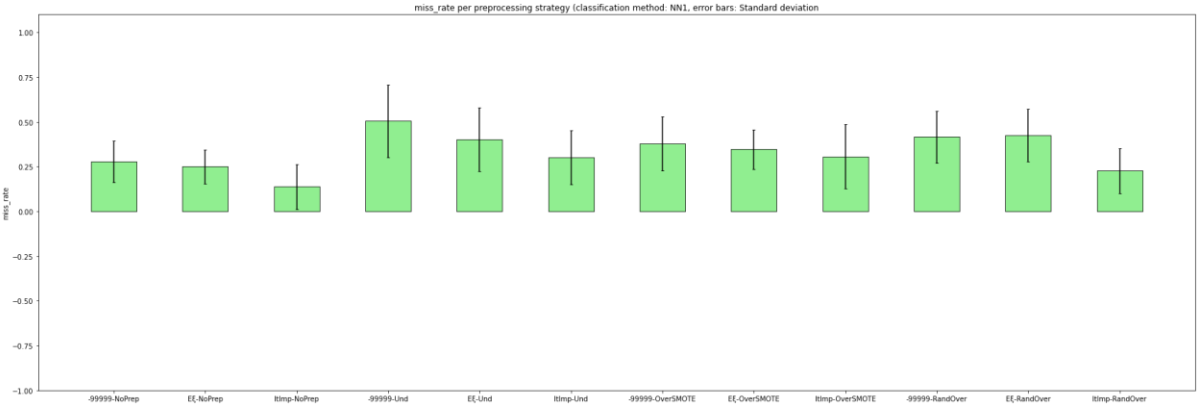

miss_rate per preprocessing strategy (classification method: DT2, error bars: Standard deviation
miss_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

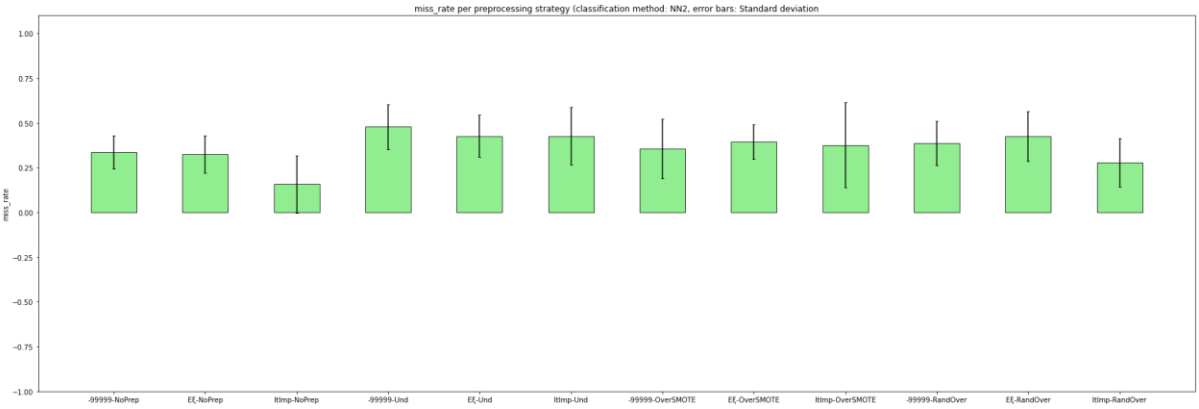

miss_rate per preprocessing strategy (classification method: DT3, error bars: Standard deviation
miss_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

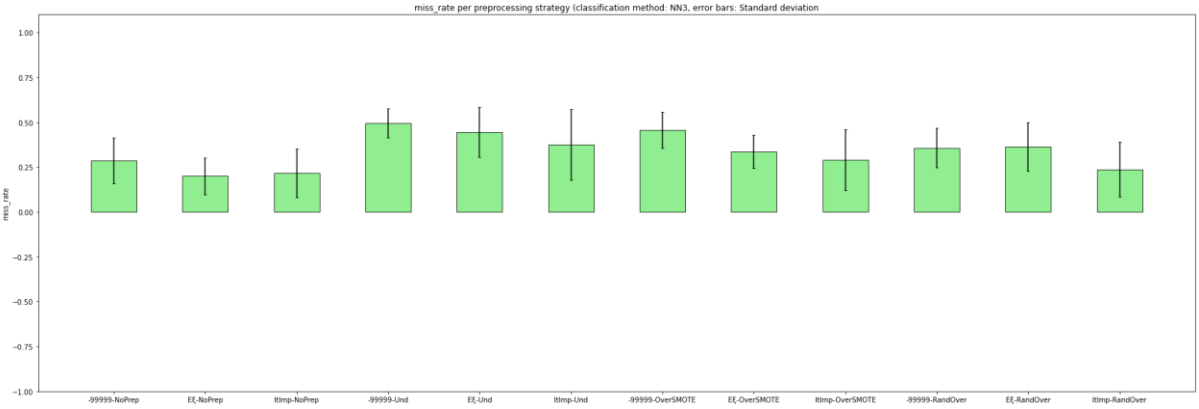

miss_rate per preprocessing strategy (classification method: DT4, error bars: Standard deviation
miss_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

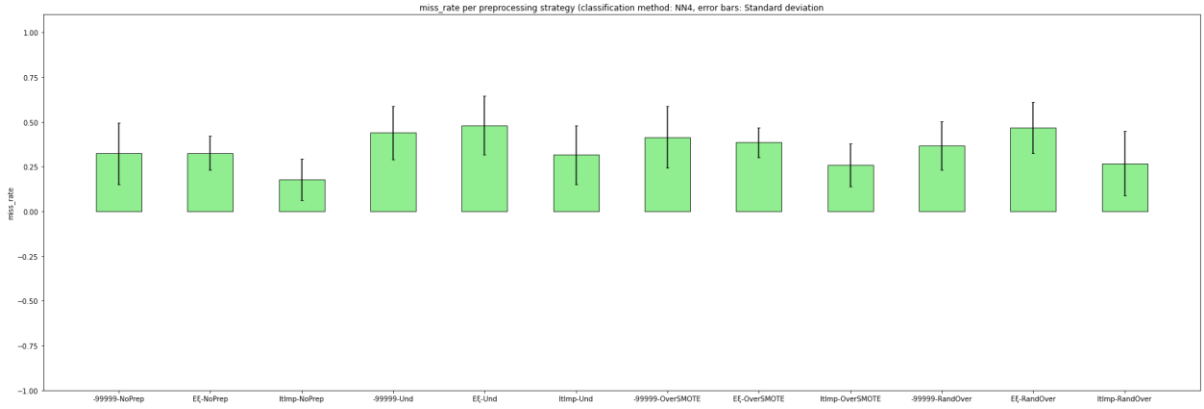

miss_rate per preprocessing strategy (classification method: RF1, error bars: Standard deviation
miss_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

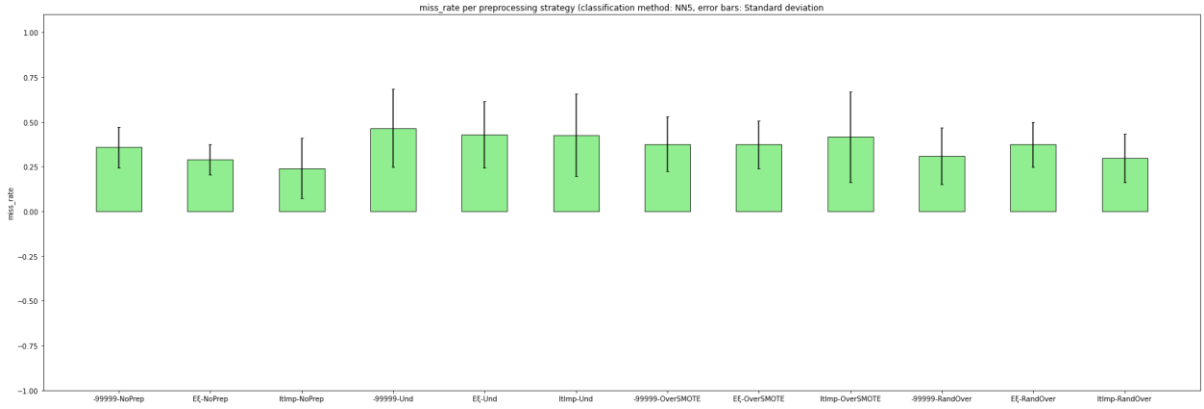

miss_rate per preprocessing strategy (classification method: RF2, error bars: Standard deviation
miss_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

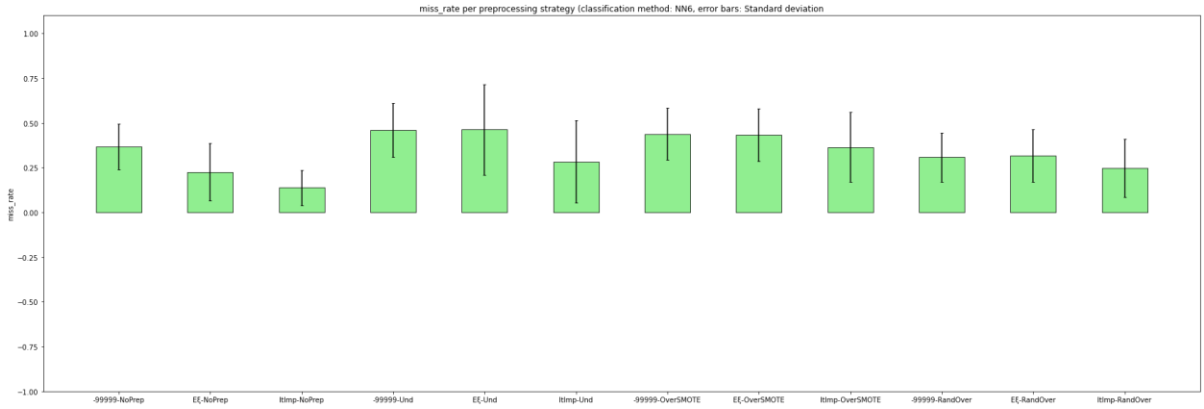

miss_rate per preprocessing strategy (classification method: NN1, error bars: Standard deviation
miss_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

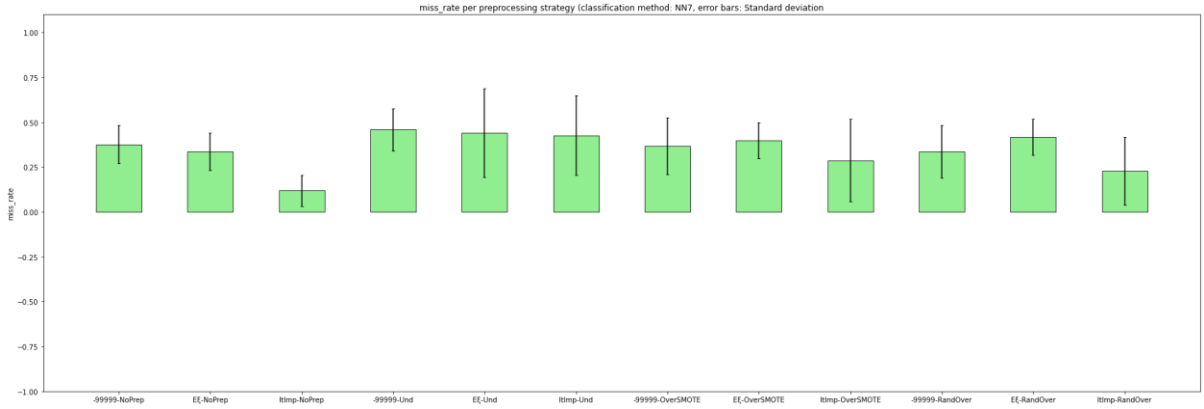

miss_rate per preprocessing strategy (classification method: NN2, error bars: Standard deviation
miss_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

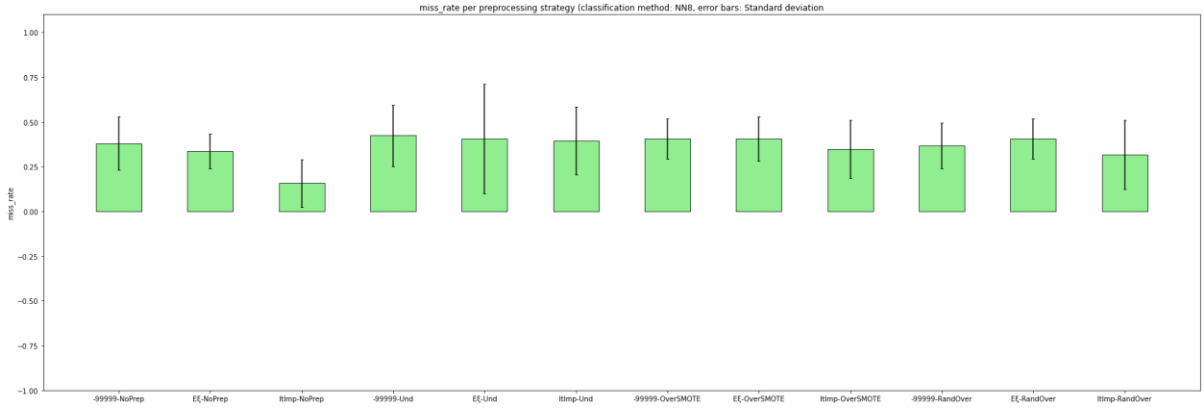

miss_rate per preprocessing strategy (classification method: NN3, error bars: Standard deviation
miss_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

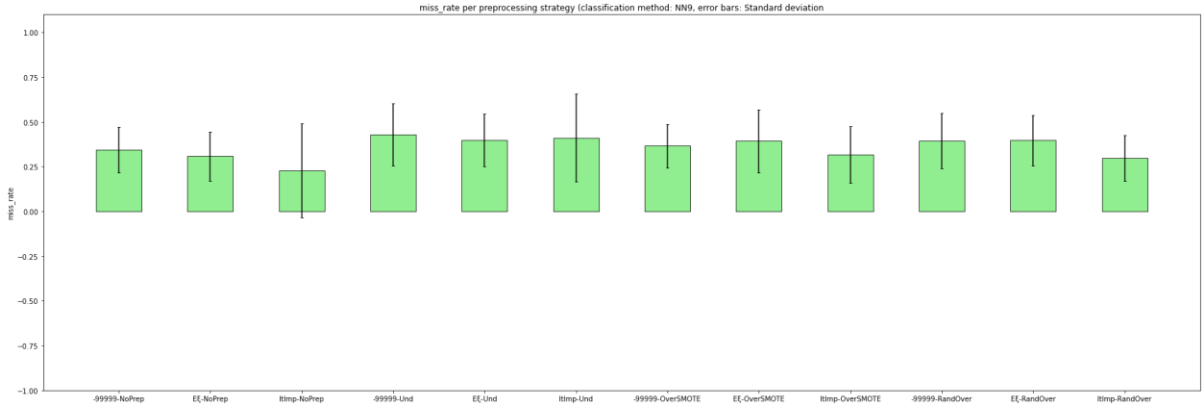

miss_rate per preprocessing strategy (classification method: NN4, error bars: Standard deviation
miss_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

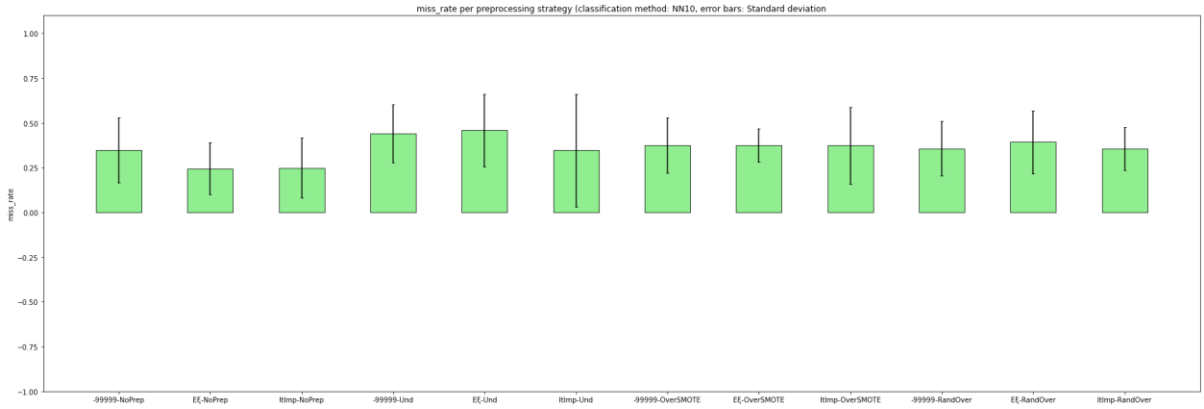

miss_rate per preprocessing strategy (classification method: NN5, error bars: Standard deviation
miss_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

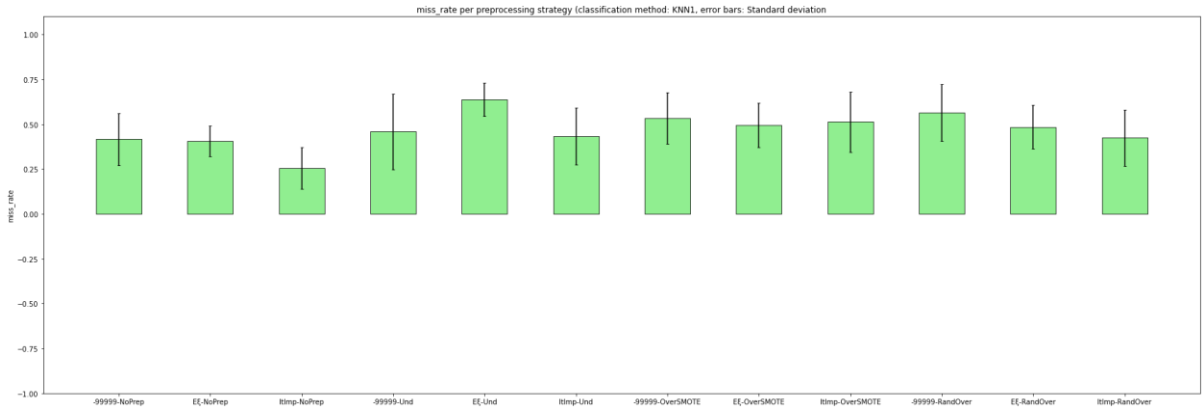

miss_rate per preprocessing strategy (classification method: NN6, error bars: Standard deviation
miss_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

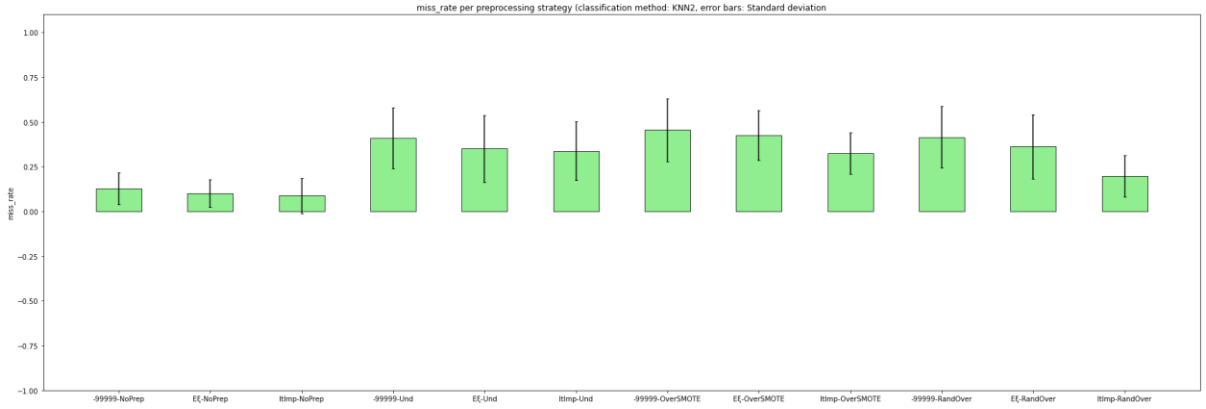

miss_rate per preprocessing strategy (classification method: NN7, error bars: Standard deviation
miss_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

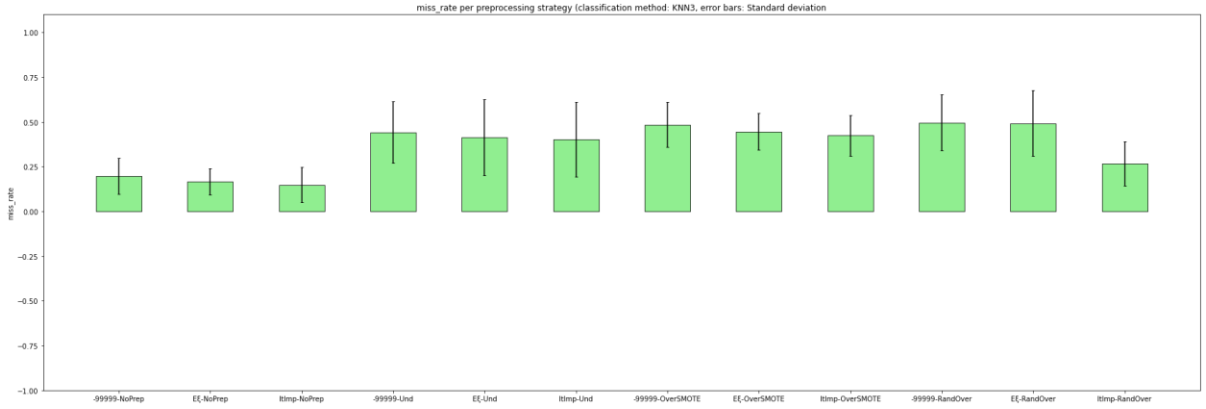

miss_rate per preprocessing strategy (classification method: NN8, error bars: Standard deviation
miss_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

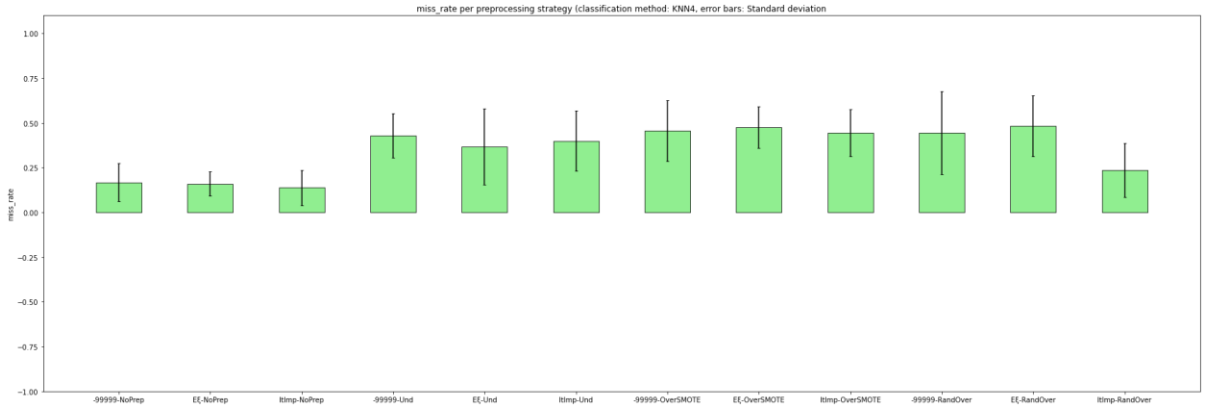

miss_rate per preprocessing strategy (classification method: NN9, error bars: Standard deviation
miss_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

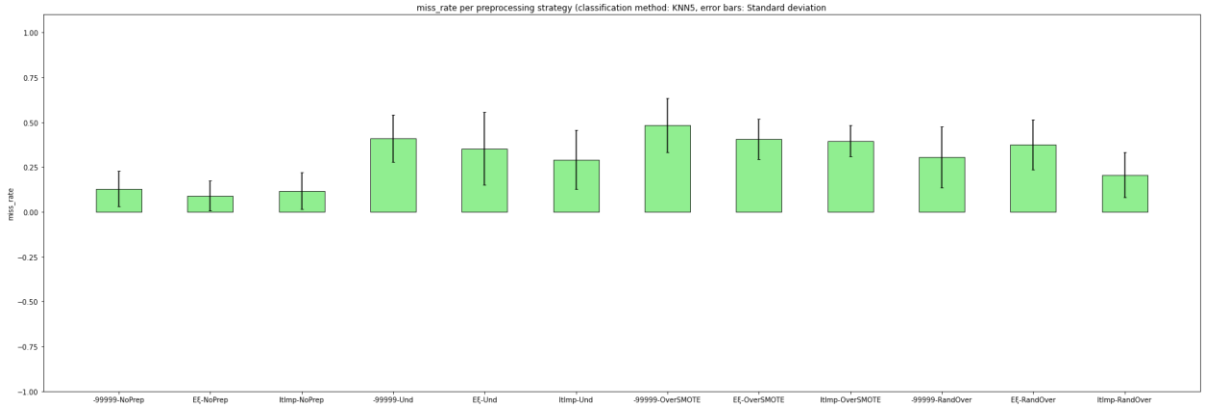

miss_rate per preprocessing strategy (classification method: NN10, error bars: Standard deviation
miss_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

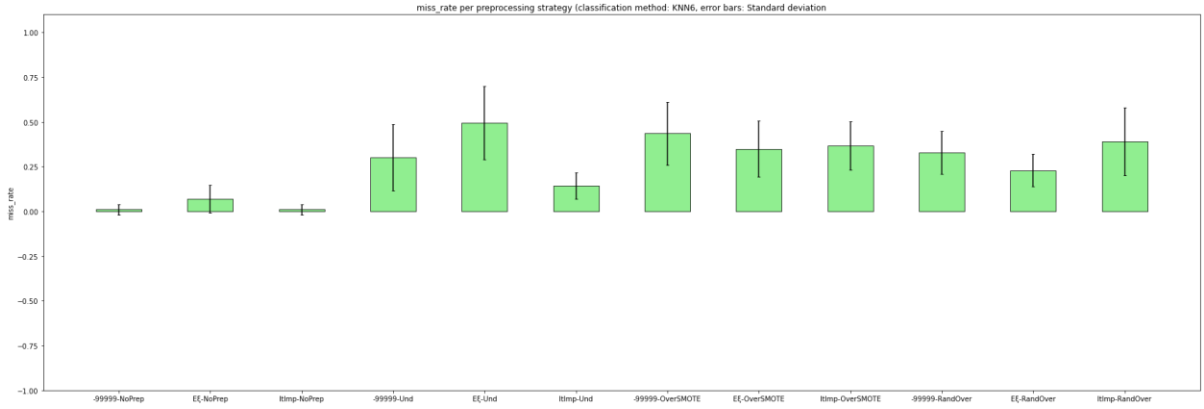

miss_rate per preprocessing strategy (classification method: KNN1, error bars: Standard deviation
miss_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

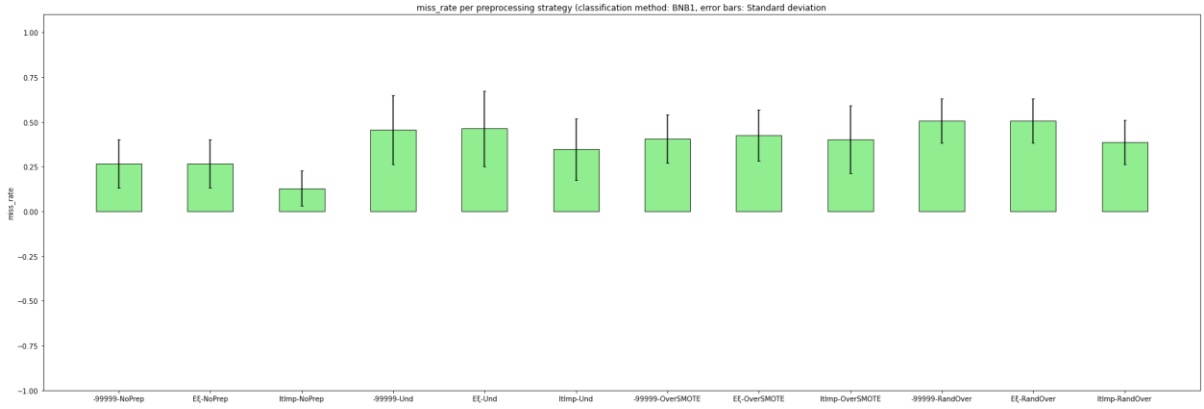

miss_rate per preprocessing strategy (classification method: KNN2, error bars: Standard deviation
miss_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

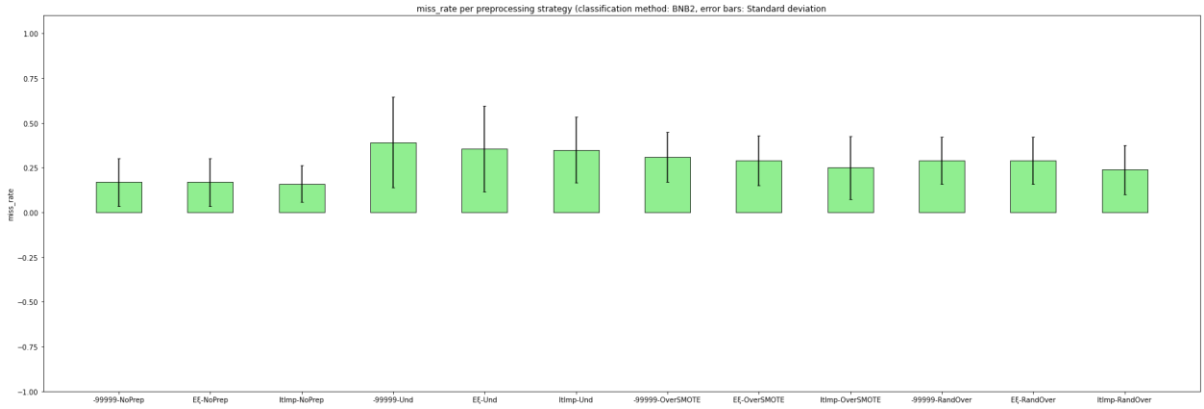

miss_rate per preprocessing strategy (classification method: KNN3, error bars: Standard deviation
miss_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

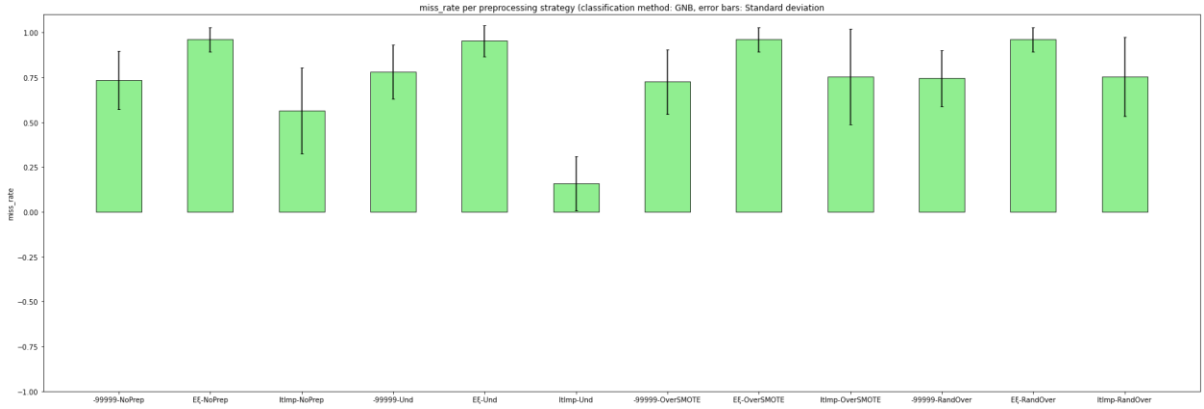

miss_rate per preprocessing strategy (classification method: KNN4, error bars: Standard deviation
miss_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

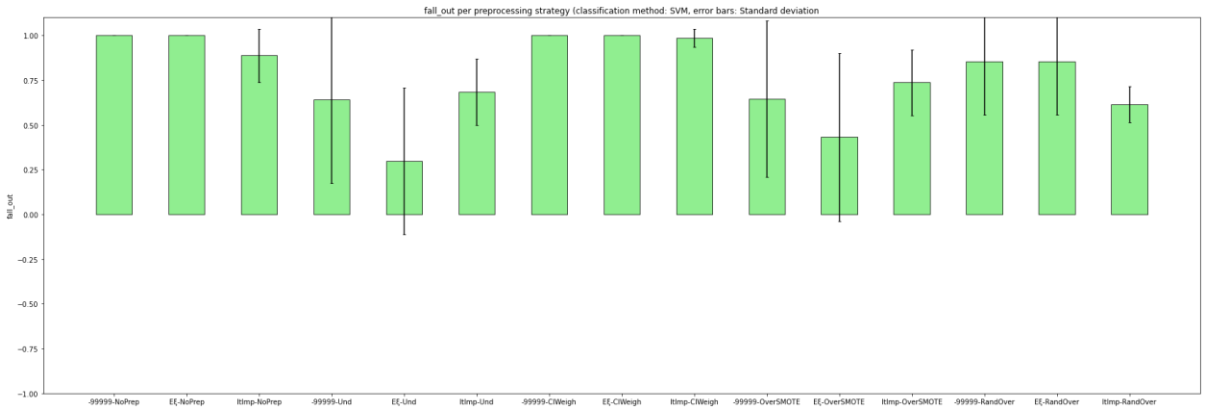

miss_rate per preprocessing strategy (classification method: KNN5, error bars: Standard deviation
miss_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

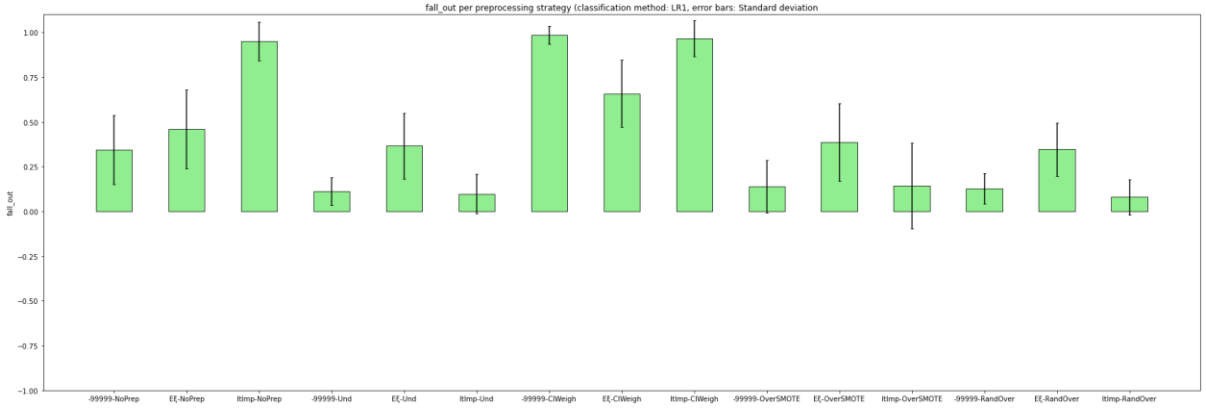

miss_rate per preprocessing strategy (classification method: KNN6, error bars: Standard deviation
miss_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

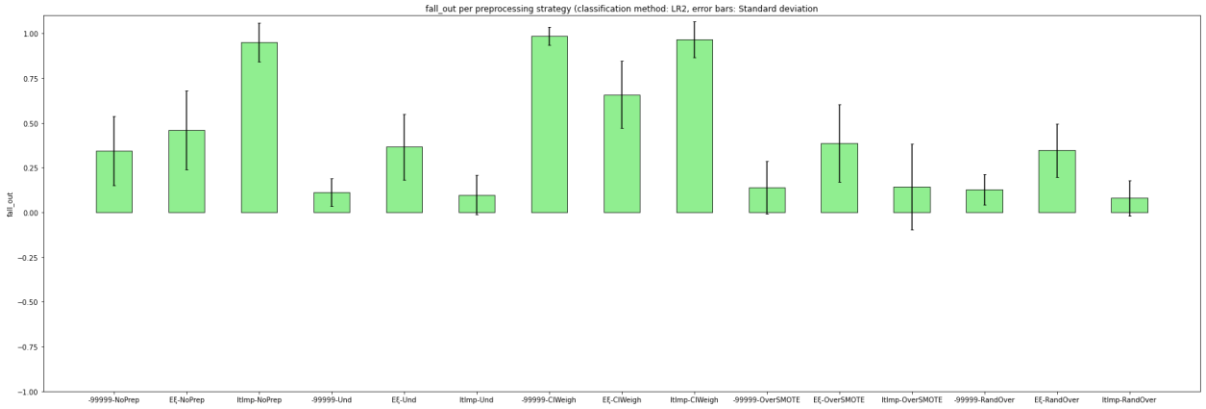

miss_rate per preprocessing strategy (classification method: BNB1, error bars: Standard deviation
miss_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

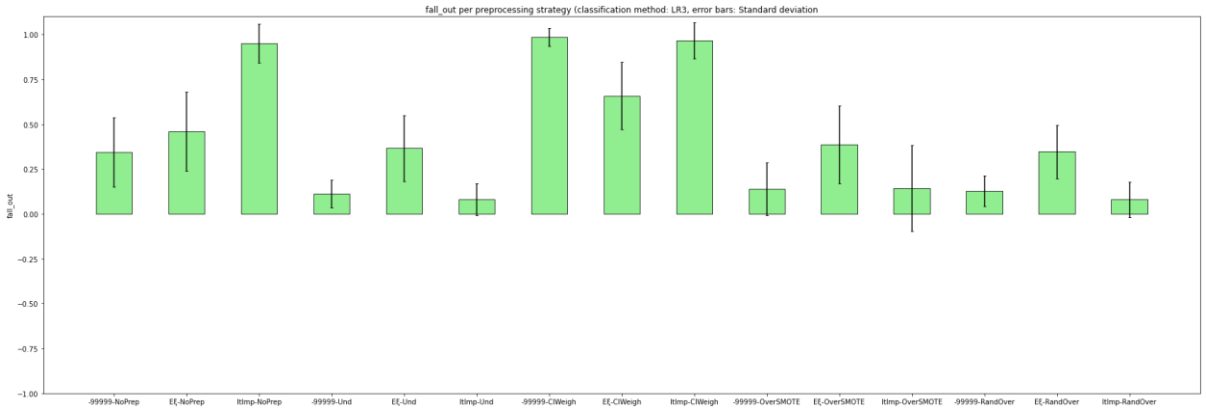

miss_rate per preprocessing strategy (classification method: BNB2, error bars: Standard deviation
miss_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

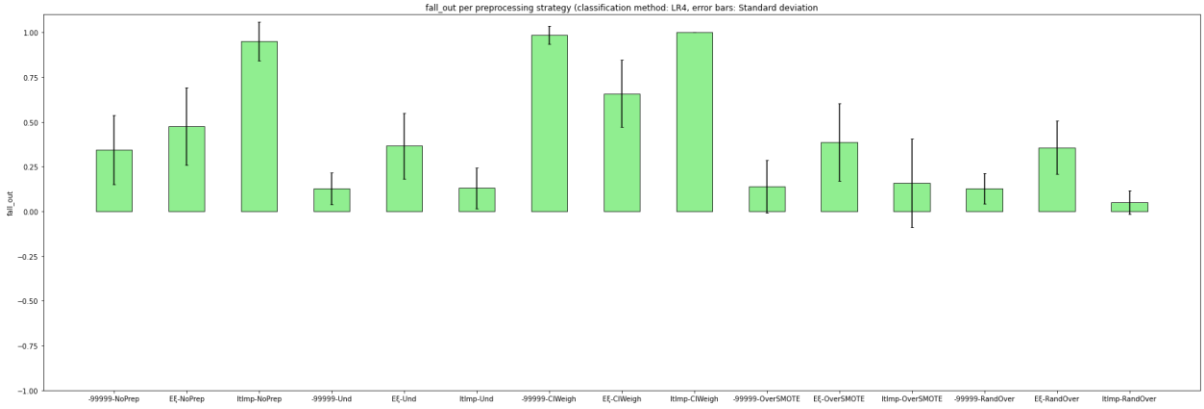

miss_rate per preprocessing strategy (classification method: GNB, error bars: Standard deviation
miss_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

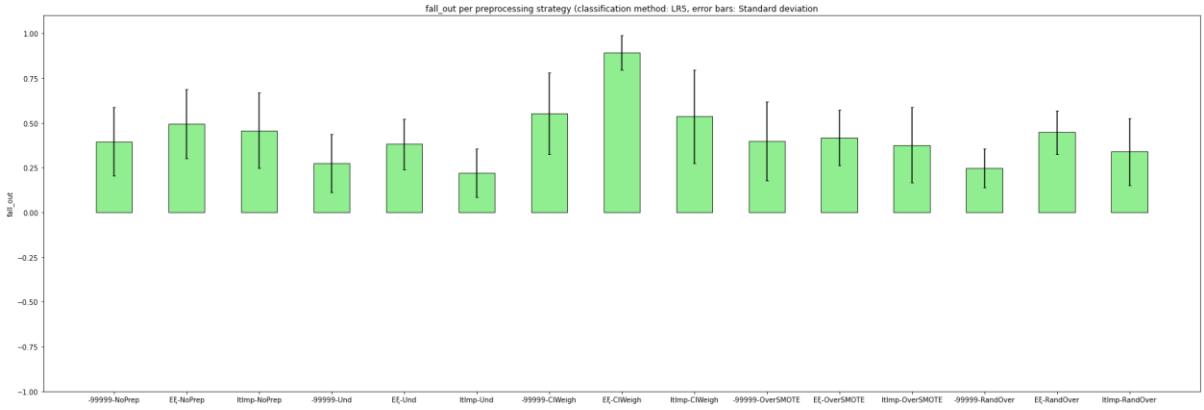

fall_out per preprocessing strategy (classification method: SVM, error bars: Standard deviation
fall_out
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

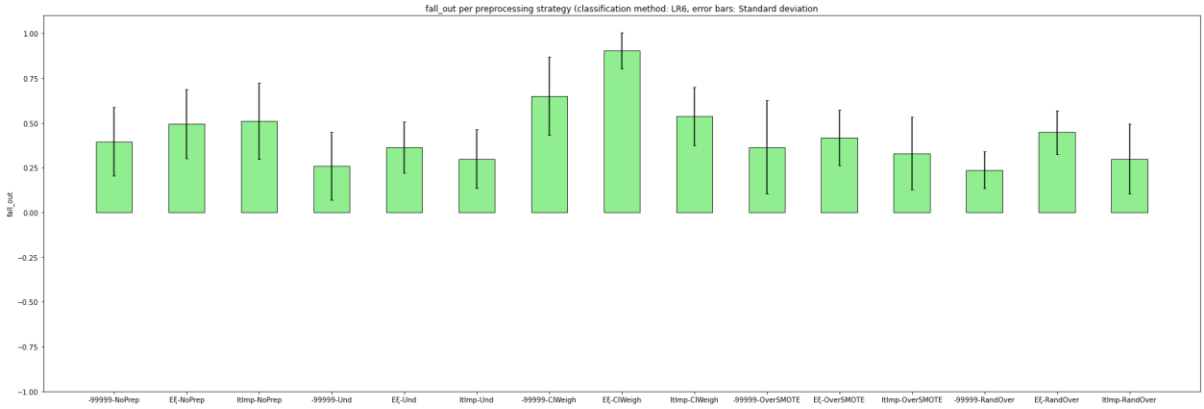

fall_out per preprocessing strategy (classification method: LR1, error bars: Standard deviation
fall_out
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

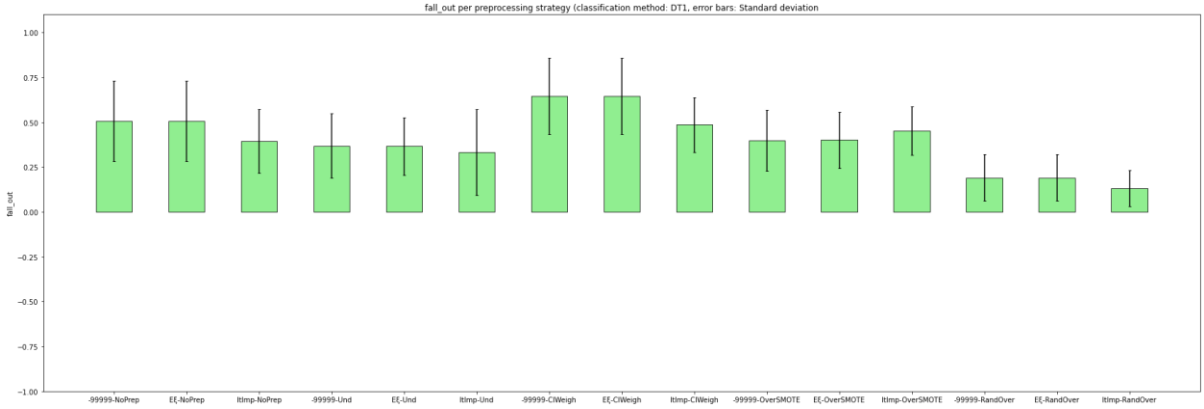

fall_out per preprocessing strategy (classification method: LR2, error bars: Standard deviation
fall_out
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

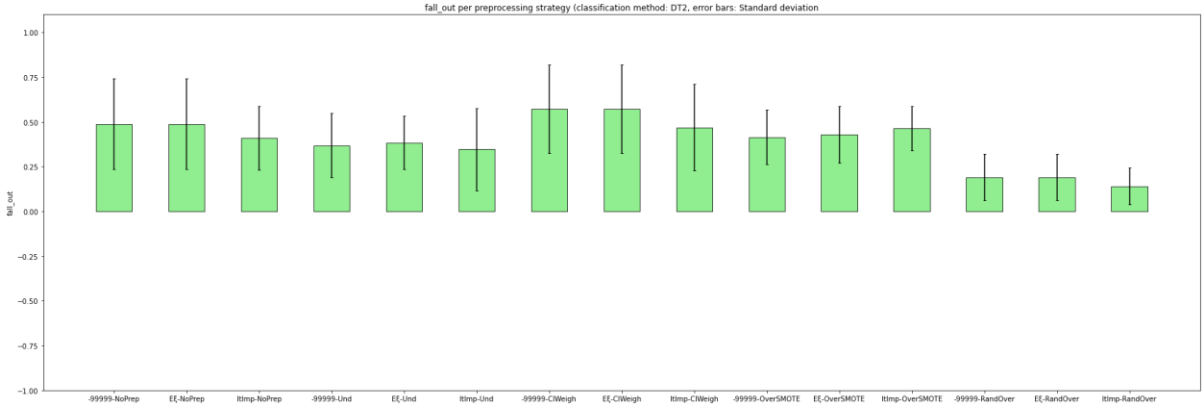

fall_out per preprocessing strategy (classification method: LR3, error bars: Standard deviation
fall_out
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

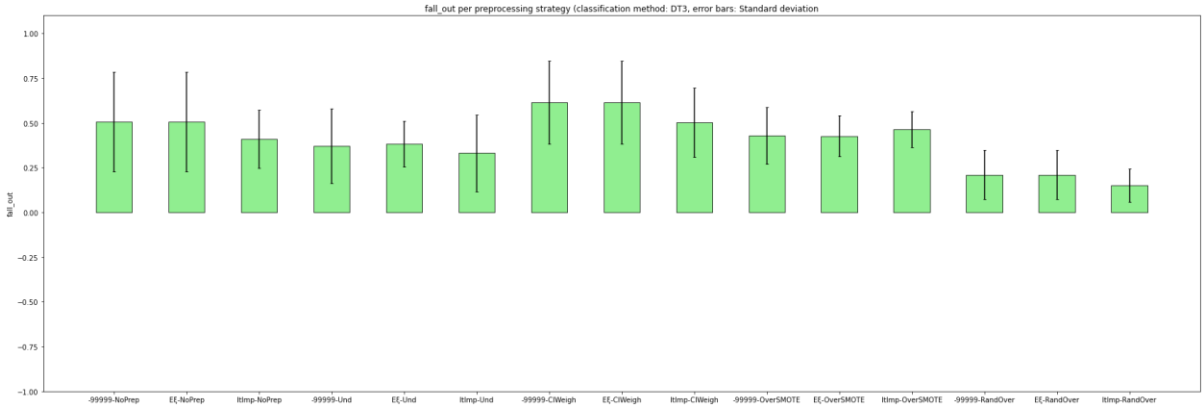

fall_out per preprocessing strategy (classification method: LR4, error bars: Standard deviation
fall_out
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

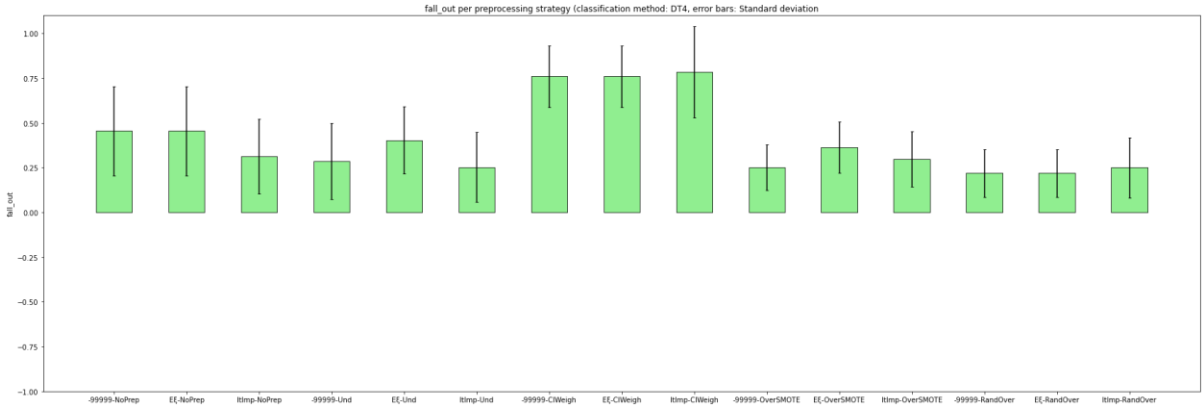

fall_out per preprocessing strategy (classification method: LR5, error bars: Standard deviation
fall_out
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

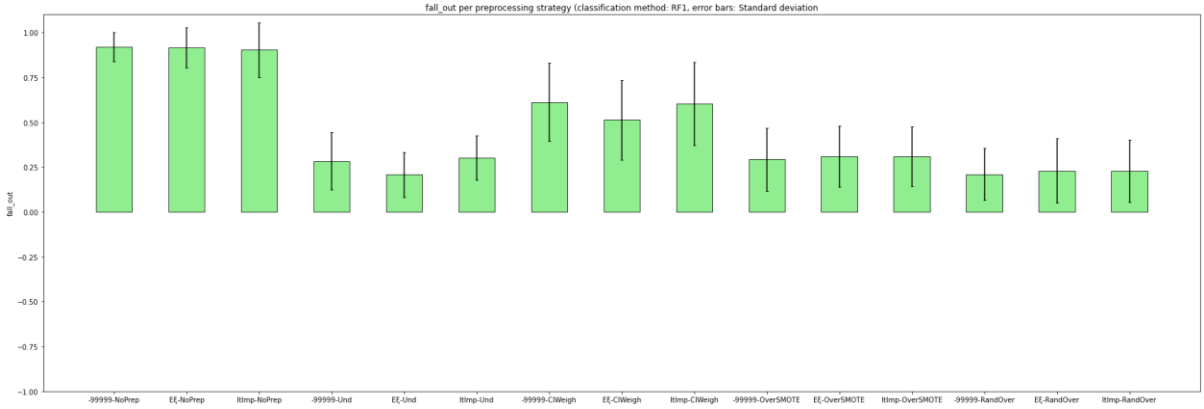

fall_out per preprocessing strategy (classification method: LR6, error bars: Standard deviation
fall_out
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

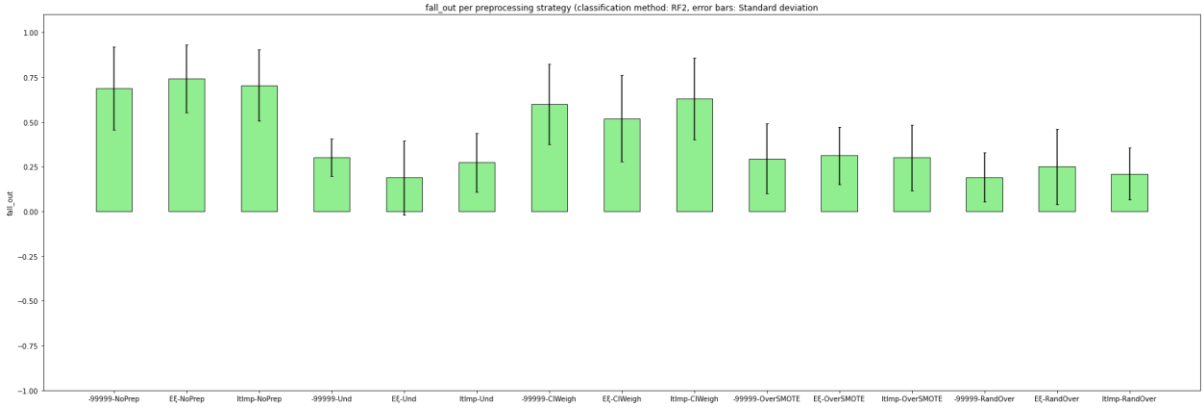

fall_out per preprocessing strategy (classification method: DT1, error bars: Standard deviation
fall_out
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

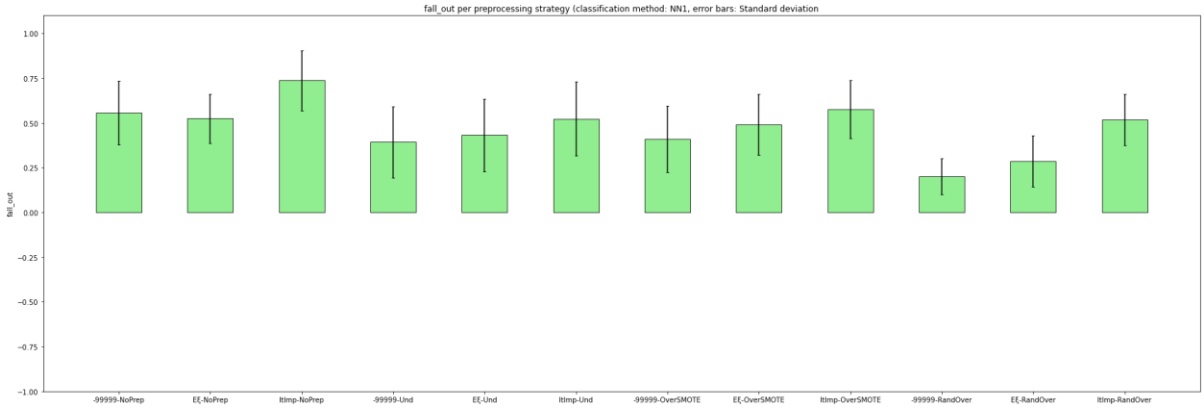

fall_out per preprocessing strategy (classification method: DT2, error bars: Standard deviation
fall_out
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

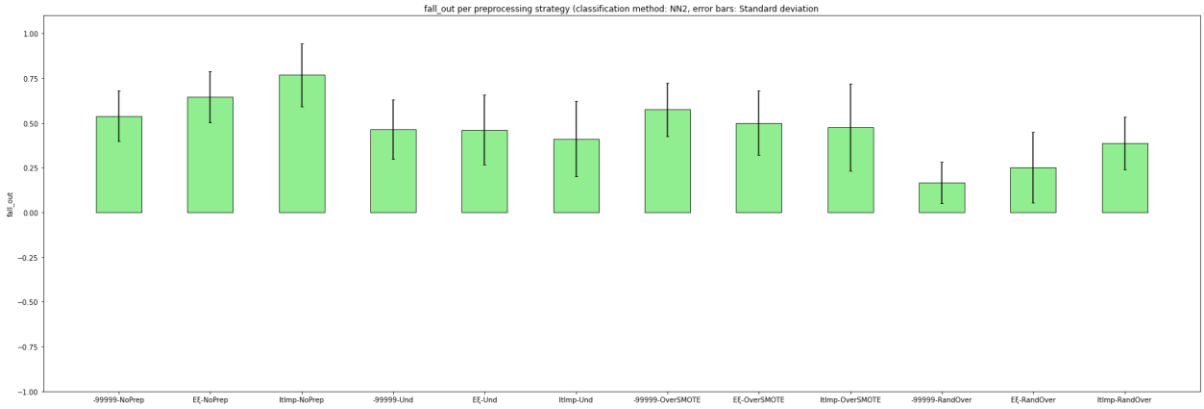

fall_out per preprocessing strategy (classification method: DT3, error bars: Standard deviation
fall_out
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

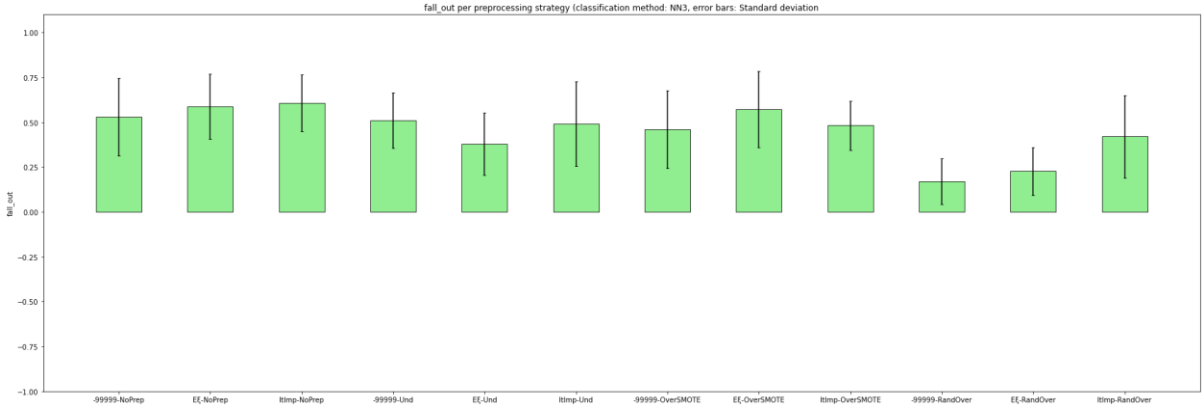

fall_out per preprocessing strategy (classification method: DT4, error bars: Standard deviation
fall_out
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

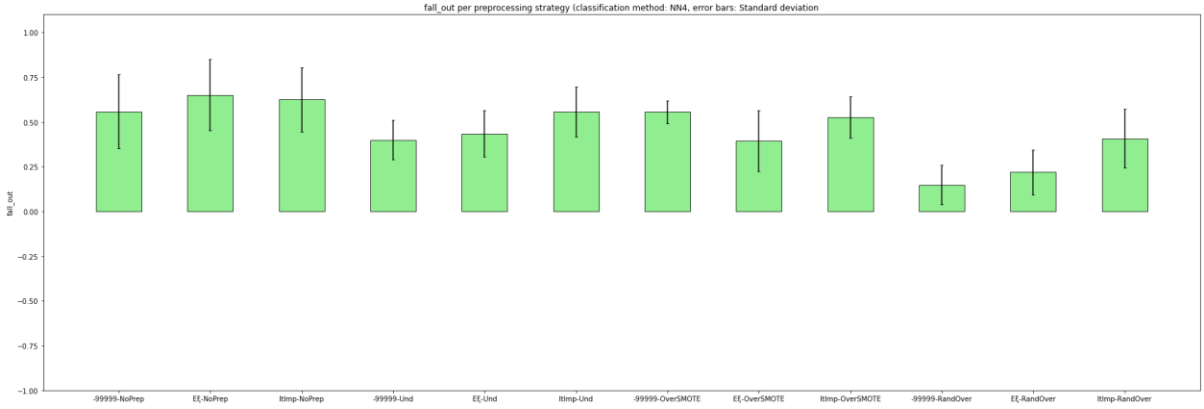

fall_out per preprocessing strategy (classification method: RF1, error bars: Standard deviation
fall_out
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

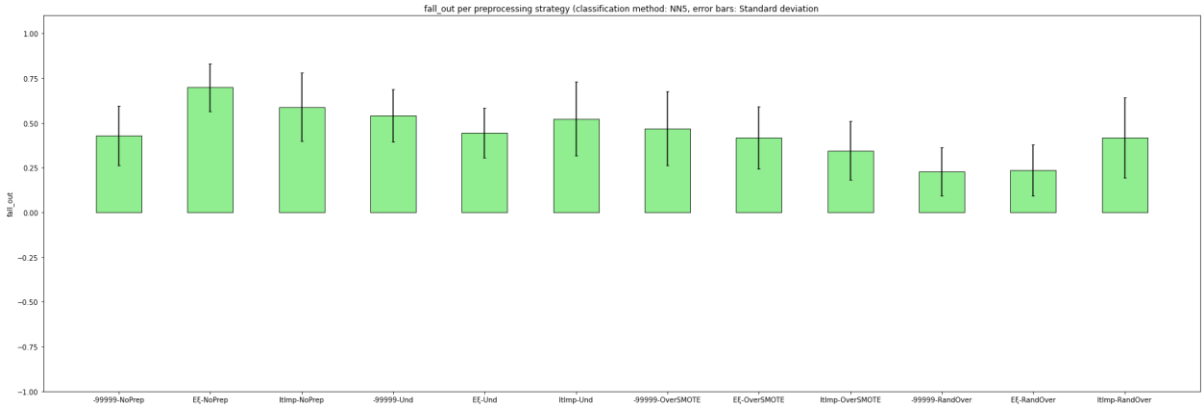

fall_out per preprocessing strategy (classification method: RF2, error bars: Standard deviation
fall_out
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

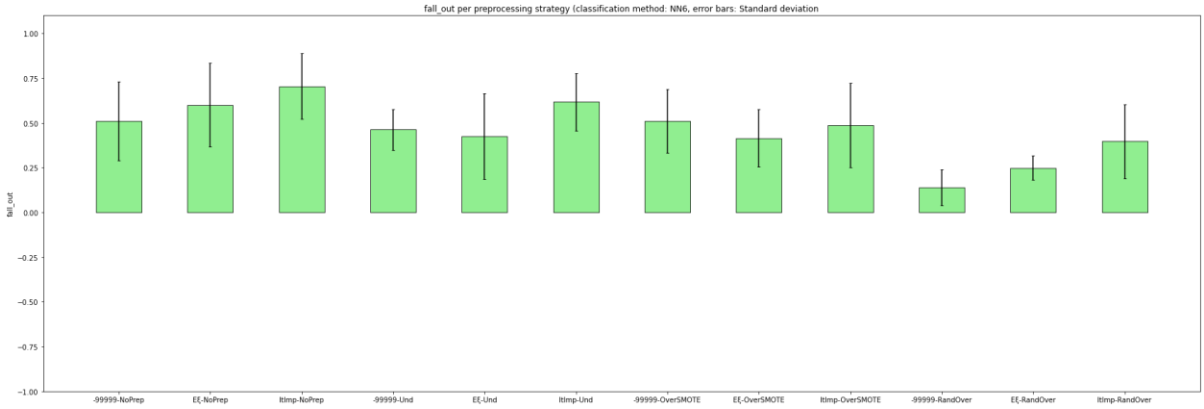

fall_out per preprocessing strategy (classification method: NN1, error bars: Standard deviation
fall_out
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

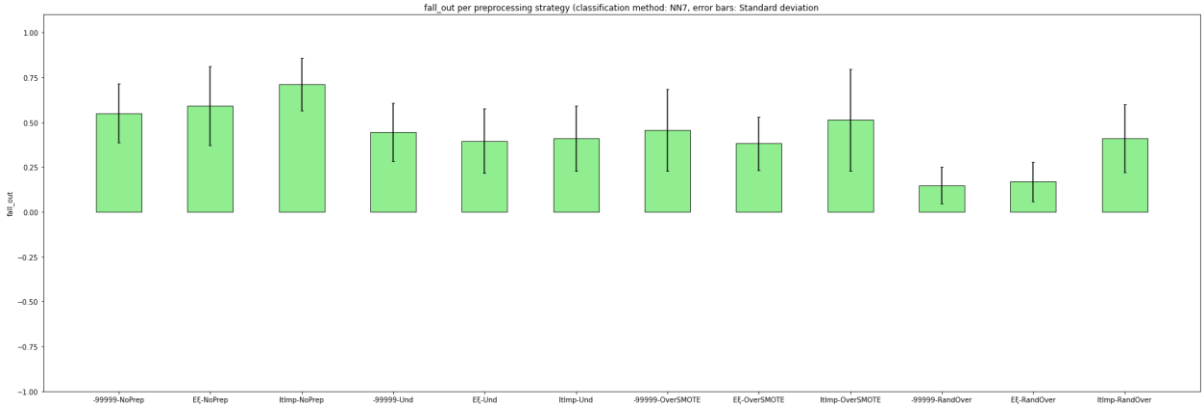

fall_out per preprocessing strategy (classification method: NN2, error bars: Standard deviation
fall_out
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

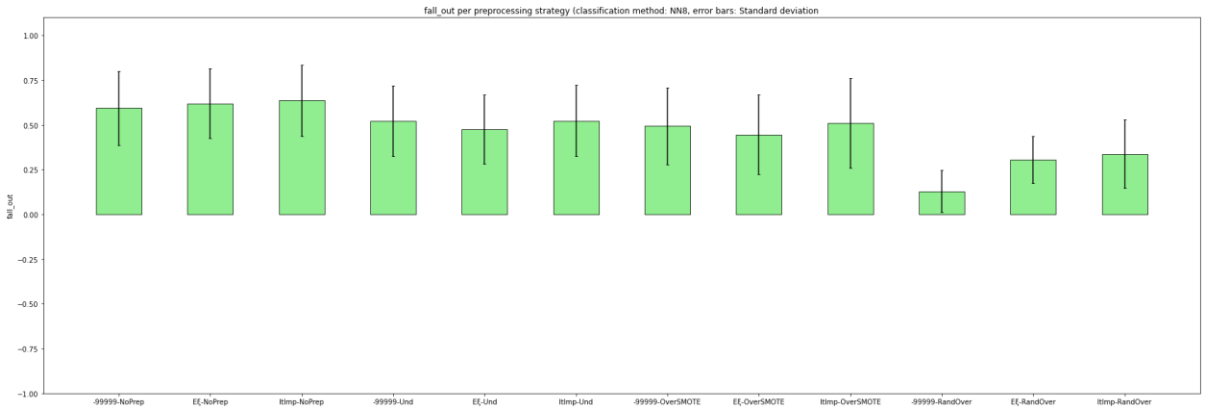

fall_out per preprocessing strategy (classification method: NN3, error bars: Standard deviation
fall_out
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

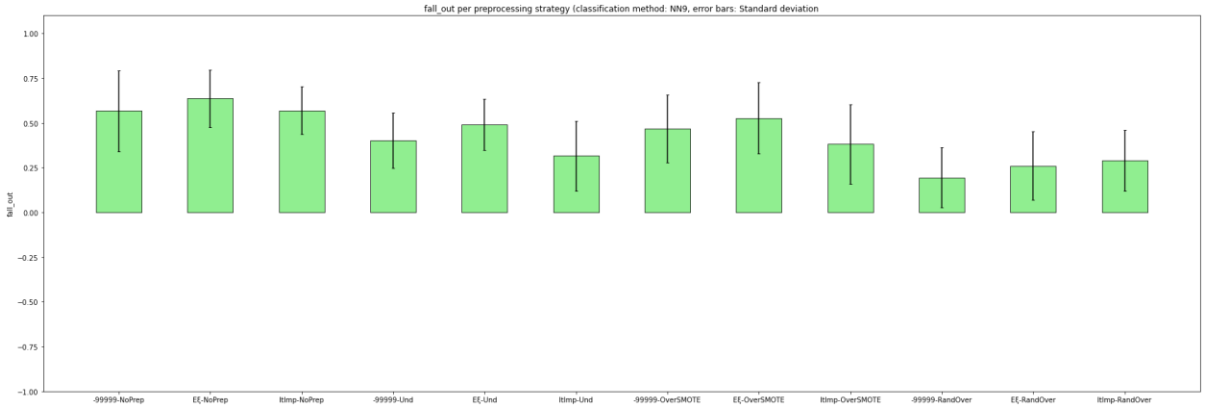

fall_out per preprocessing strategy (classification method: NN4, error bars: Standard deviation
fall_out
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

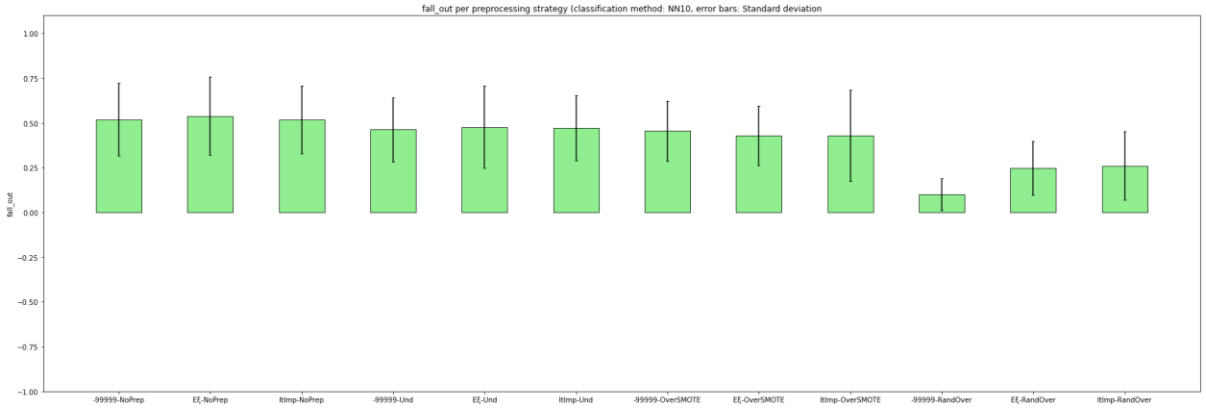

fall_out per preprocessing strategy (classification method: NN5, error bars: Standard deviation
fall_out
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

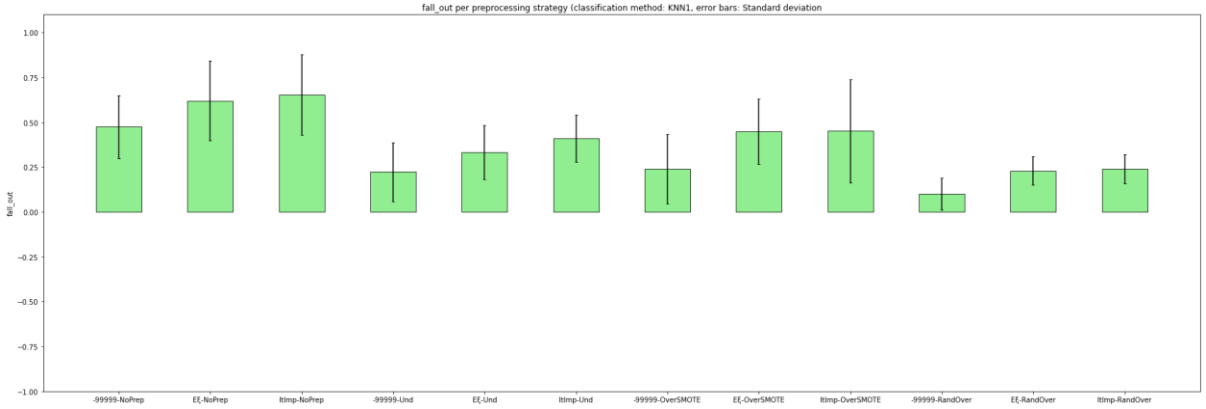

fall_out per preprocessing strategy (classification method: NN6, error bars: Standard deviation
fall_out
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

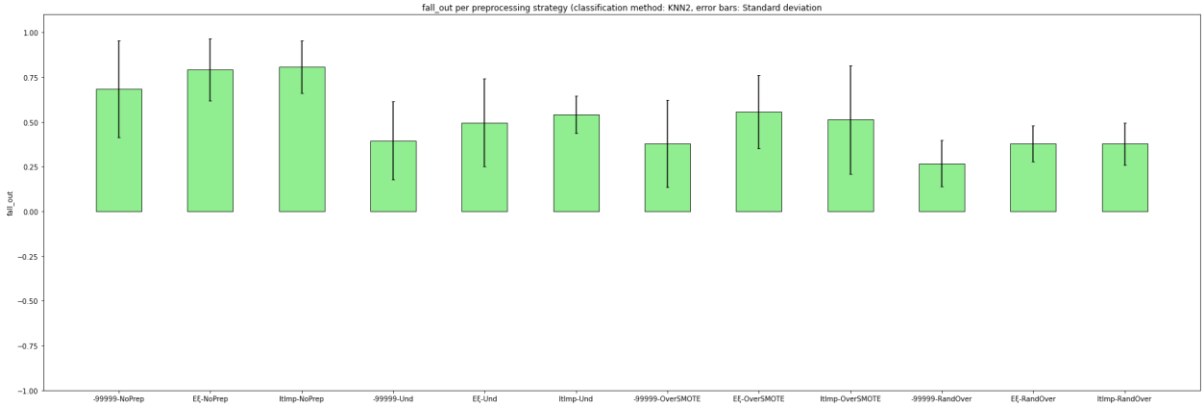

fall_out per preprocessing strategy (classification method: NN7, error bars: Standard deviation
fall_out
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

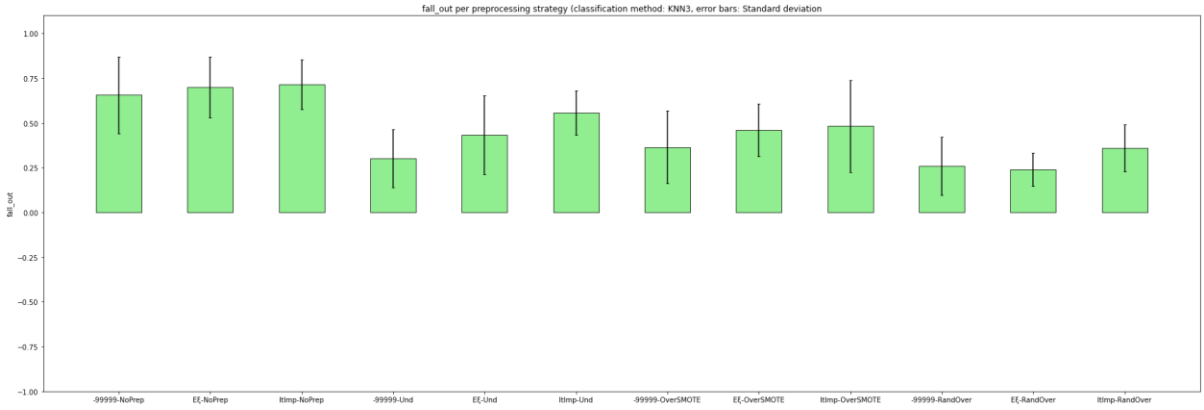

fall_out per preprocessing strategy (classification method: NN8, error bars: Standard deviation
fall_out
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

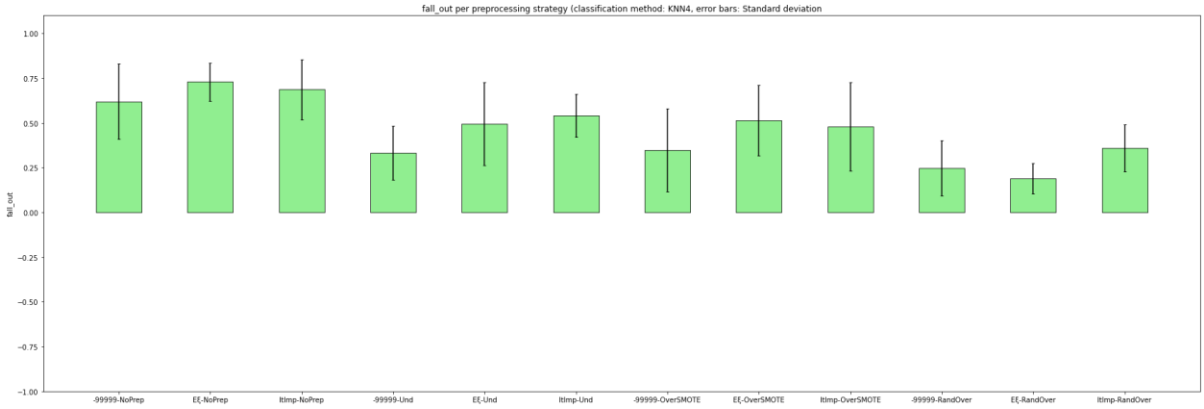

fall_out per preprocessing strategy (classification method: NN9, error bars: Standard deviation
fall_out
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

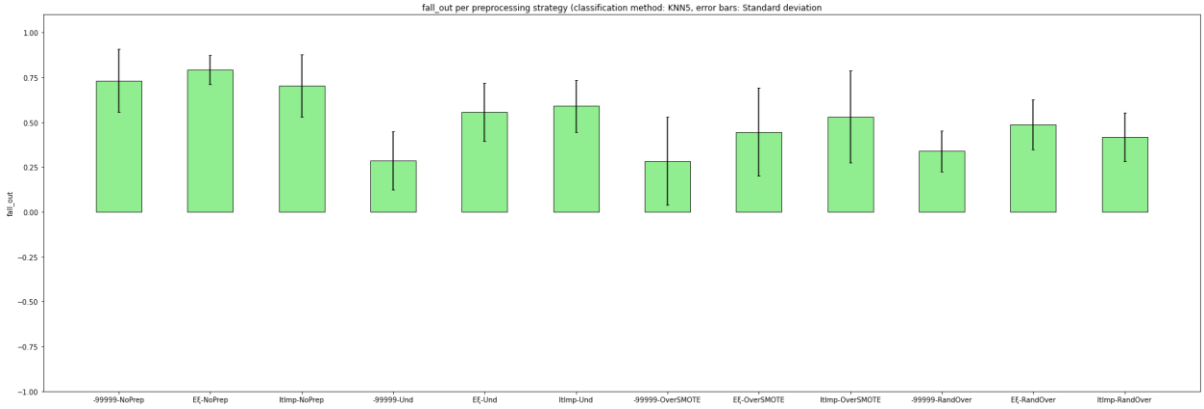

fall_out per preprocessing strategy (classification method: NN10, error bars: Standard deviation
fall_out
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

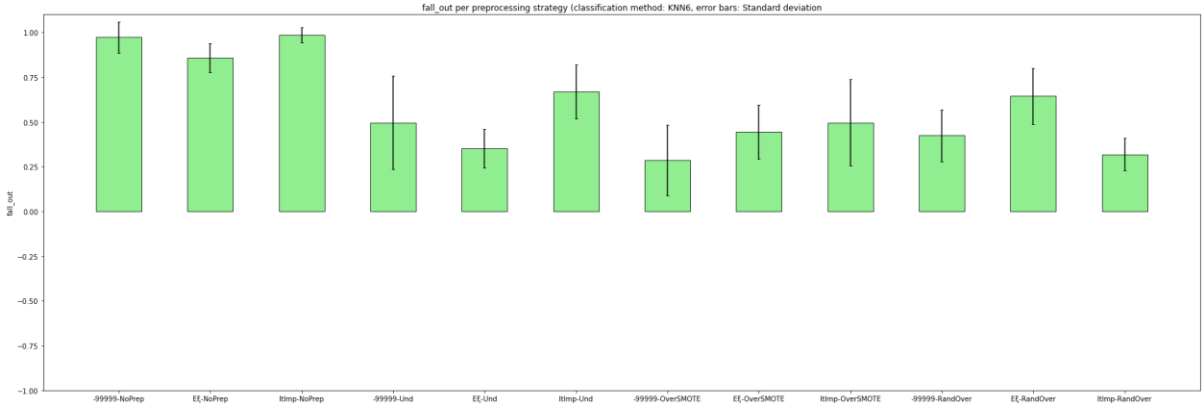

fall_out per preprocessing strategy (classification method: KNN1, error bars: Standard deviation
fall_out
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

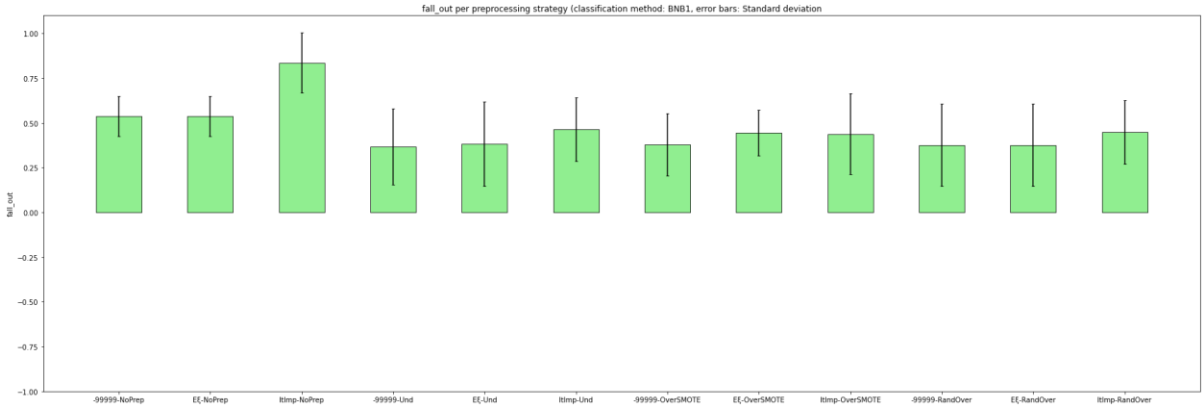

fall_out per preprocessing strategy (classification method: KNN2, error bars: Standard deviation
fall_out
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

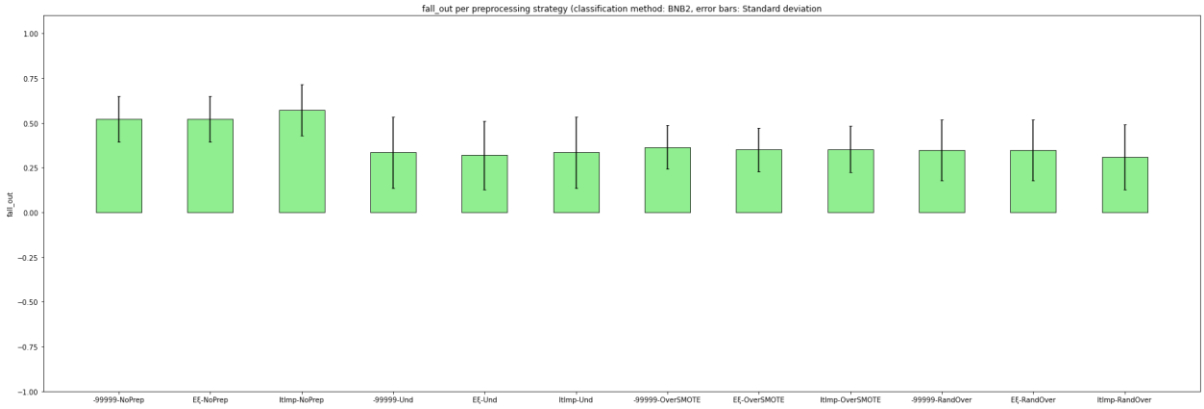

fall_out per preprocessing strategy (classification method: KNN3, error bars: Standard deviation
fall_out
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

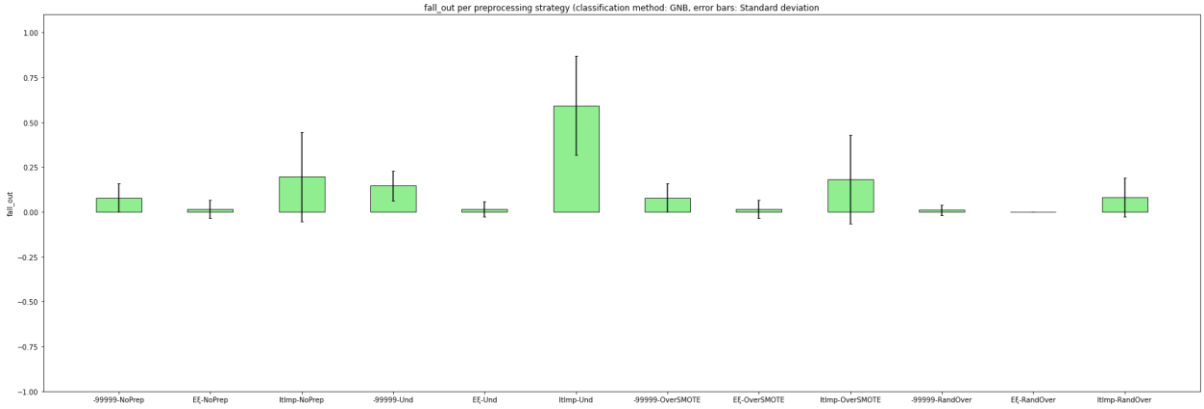

fall_out per preprocessing strategy (classification method: KNN4, error bars: Standard deviation
fall_out
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

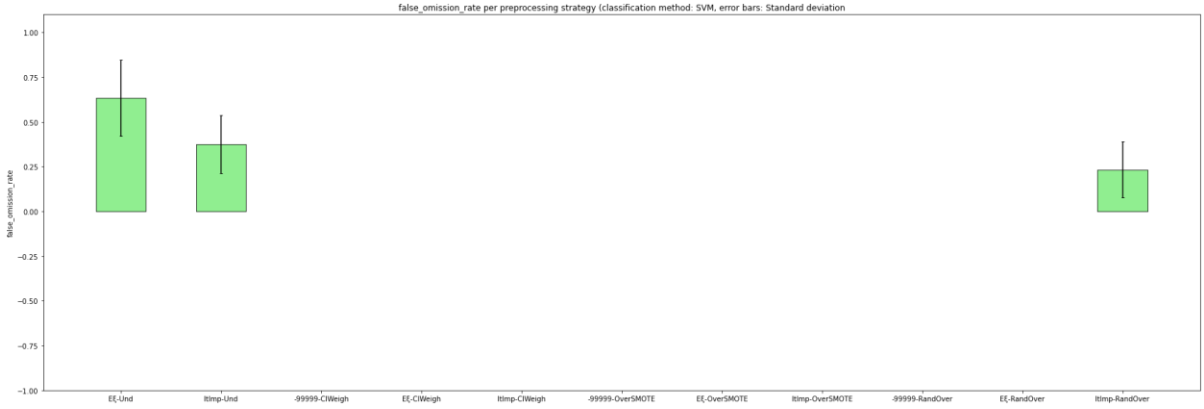

fall_out per preprocessing strategy (classification method: KNN5, error bars: Standard deviation
fall_out
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

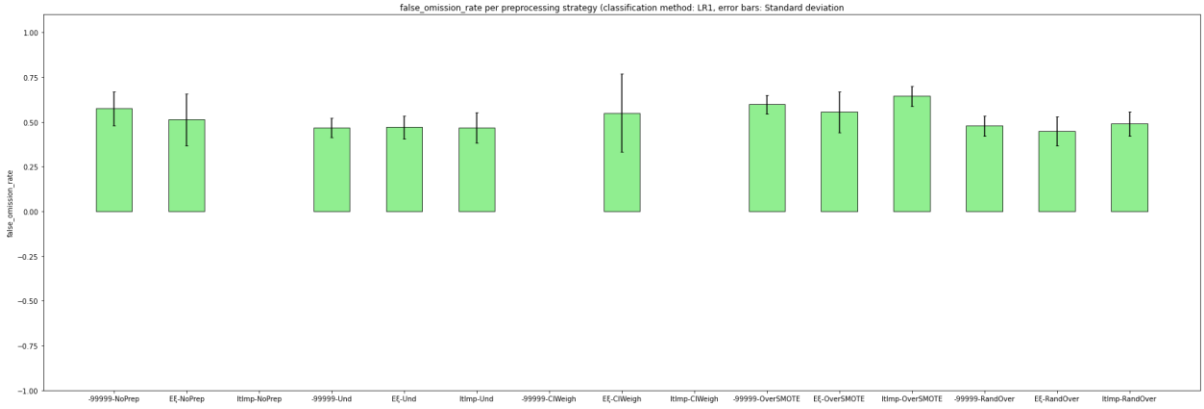

fall_out per preprocessing strategy (classification method: KNN6, error bars: Standard deviation
fall_out
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

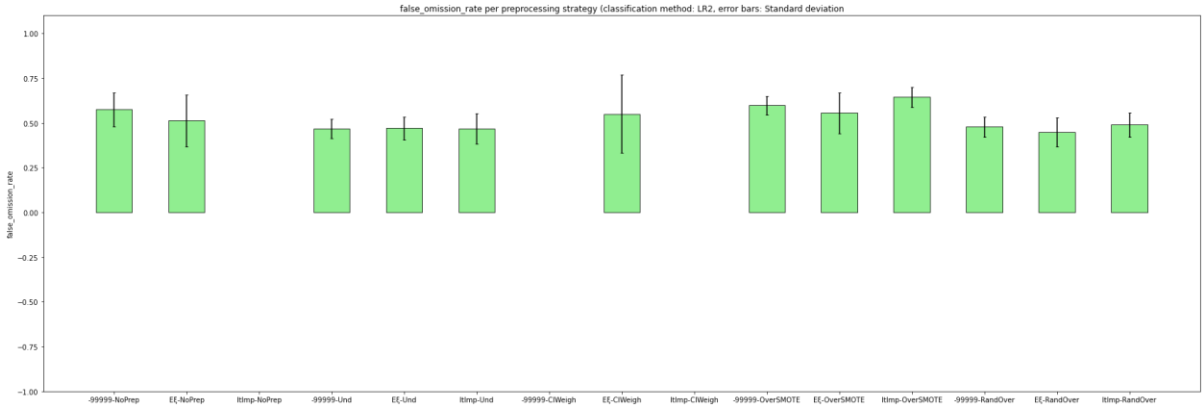

fall_out per preprocessing strategy (classification method: BNB1, error bars: Standard deviation
fall_out
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

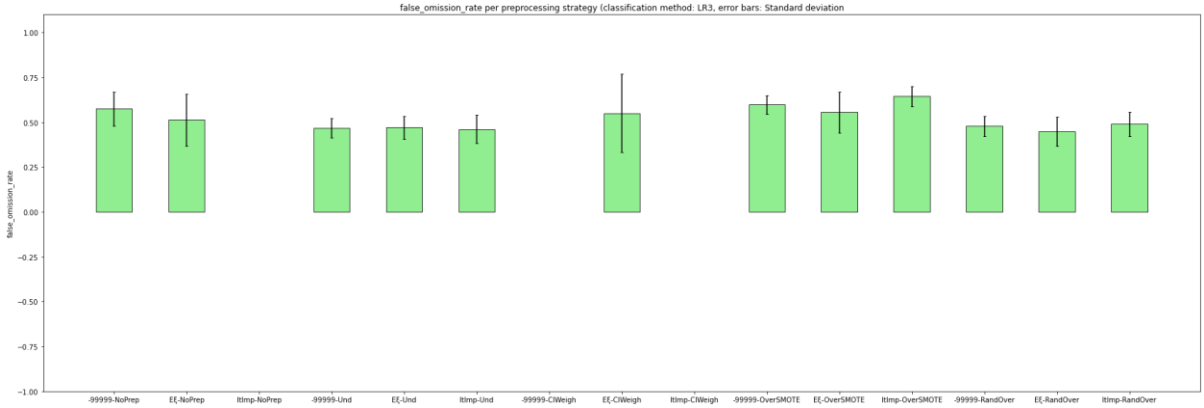

fall_out per preprocessing strategy (classification method: BNB2, error bars: Standard deviation
fall_out
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

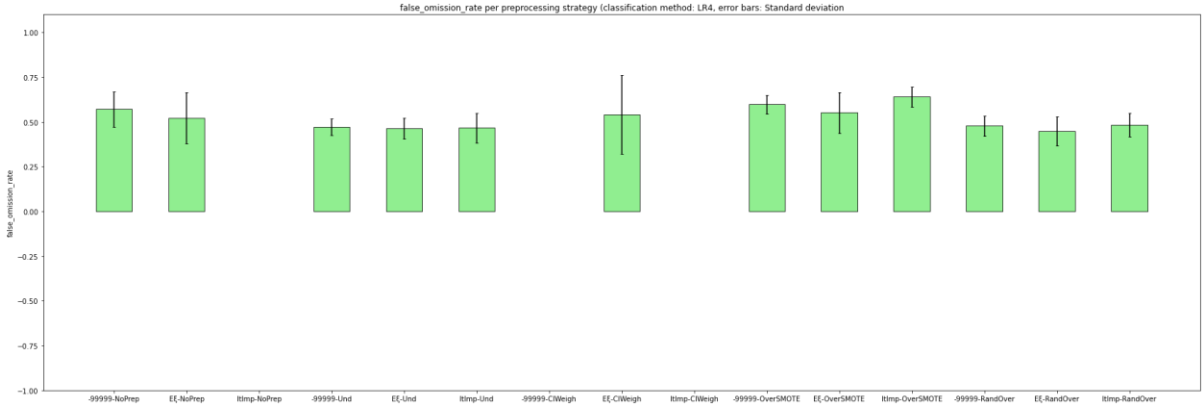

fall_out per preprocessing strategy (classification method: GNB, error bars: Standard deviation
fall_out
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

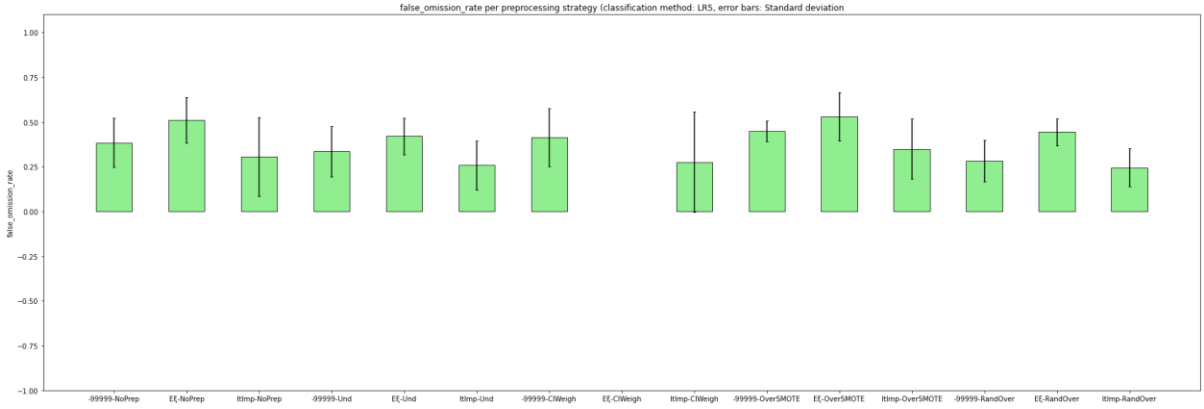

false_omission_rate per preprocessing strategy (classification method: SVM, error bars: Standard deviation
false_omission_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

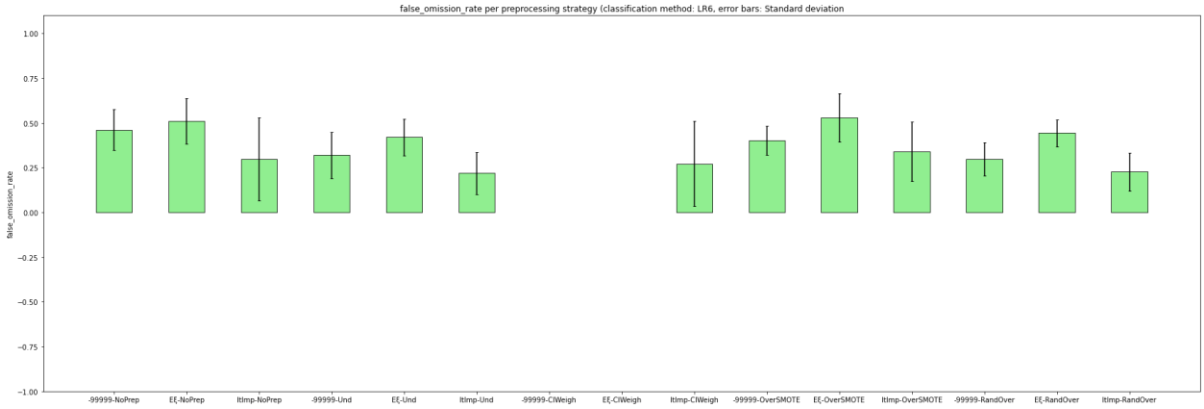

false_omission_rate per preprocessing strategy (classification method: LR1, error bars: Standard deviation
false_omission_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

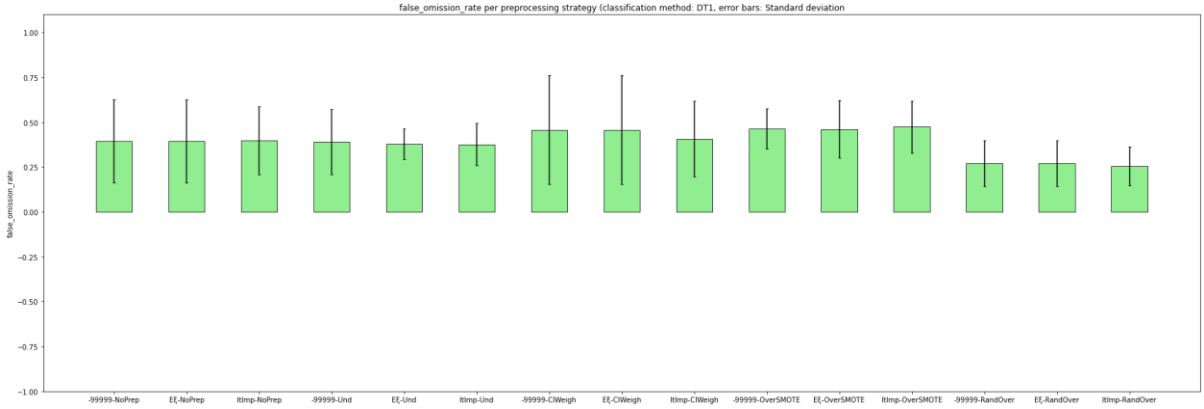

false_omission_rate per preprocessing strategy (classification method: LR2, error bars: Standard deviation
false_omission_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

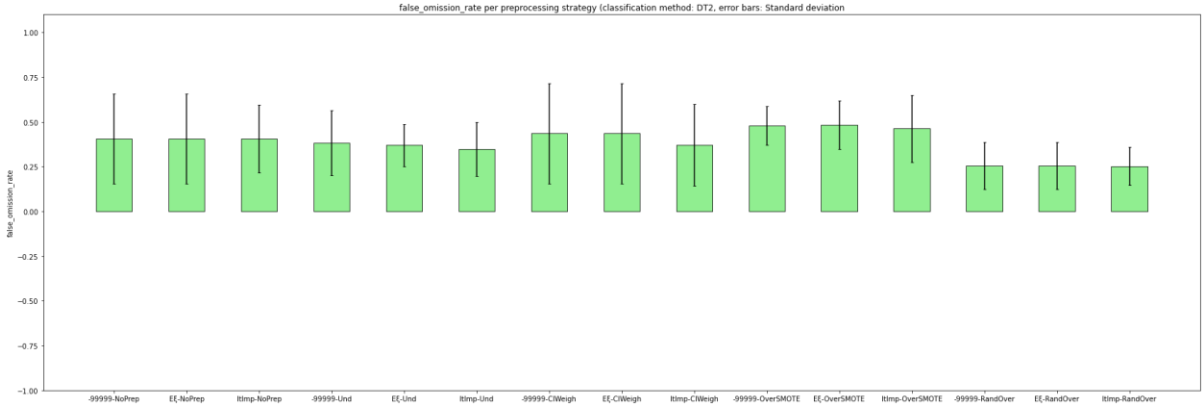

false_omission_rate per preprocessing strategy (classification method: LR3, error bars: Standard deviation
false_omission_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

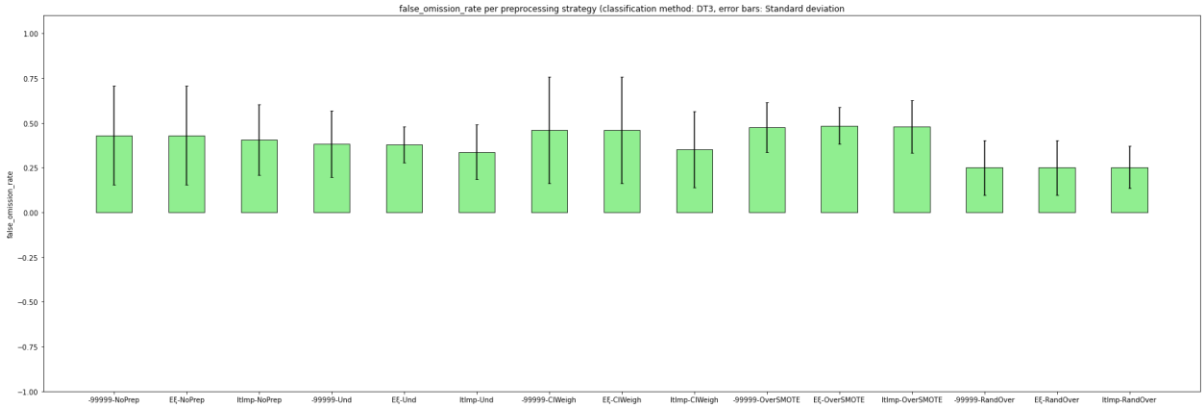

false_omission_rate per preprocessing strategy (classification method: LR4, error bars: Standard deviation
false_omission_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

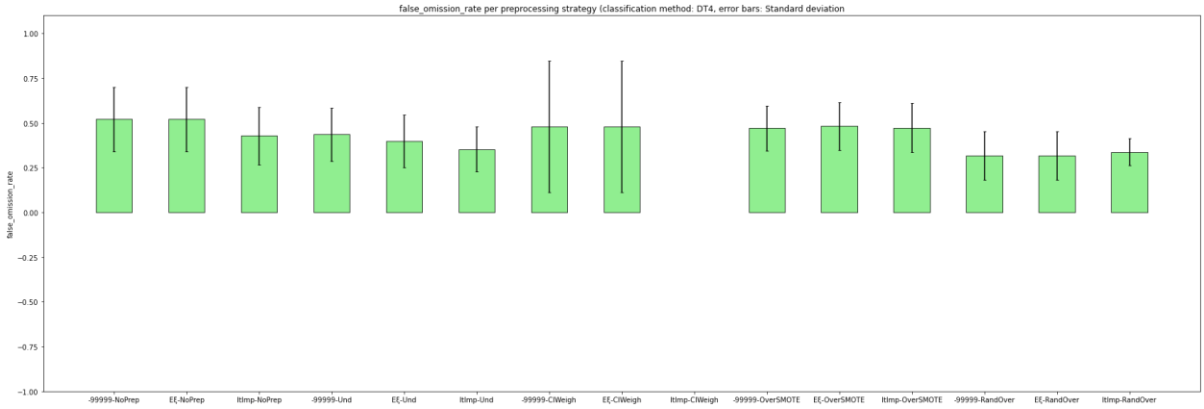

false_omission_rate per preprocessing strategy (classification method: LR5, error bars: Standard deviation
false_omission_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

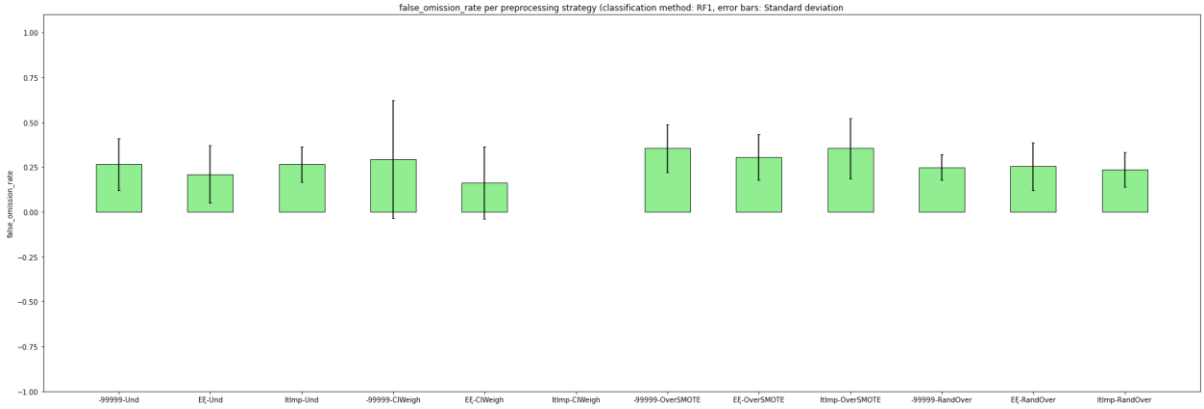

false_omission_rate per preprocessing strategy (classification method: LR6, error bars: Standard deviation
false_omission_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

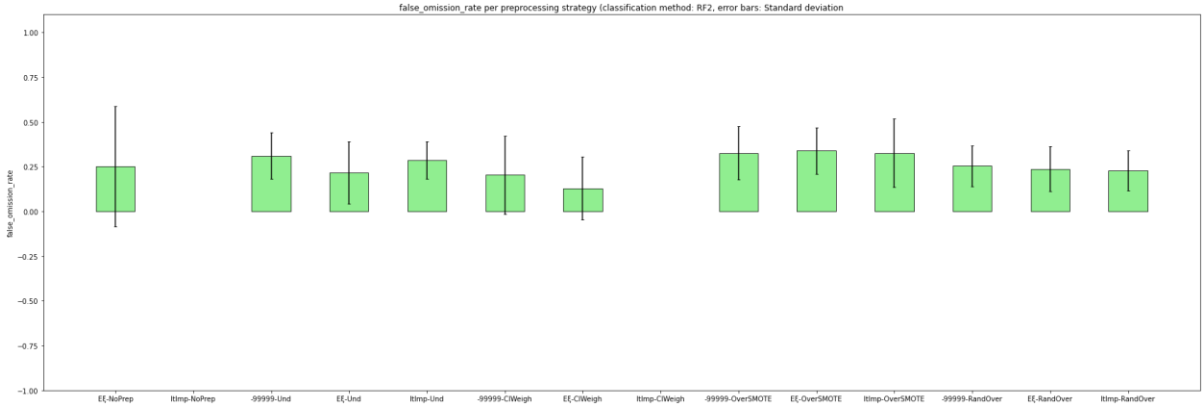

false_omission_rate per preprocessing strategy (classification method: DT1, error bars: Standard deviation
false_omission_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

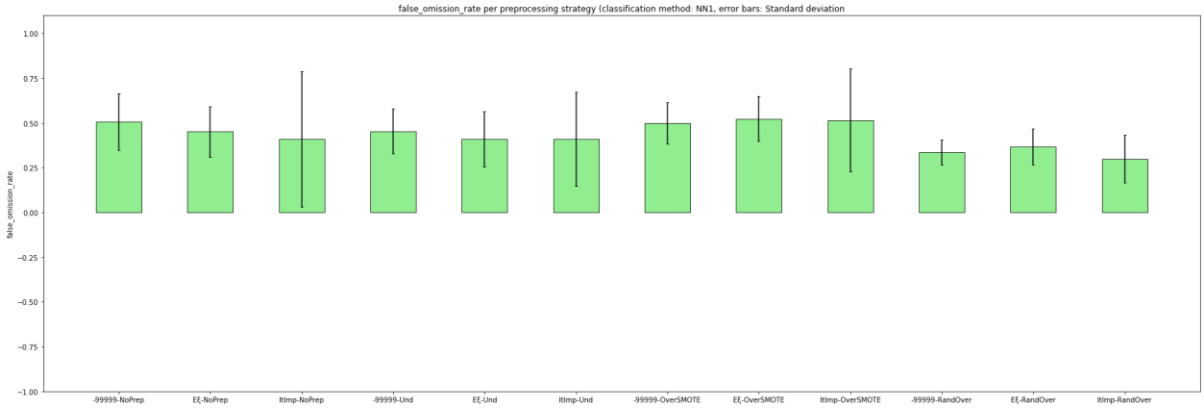

false_omission_rate per preprocessing strategy (classification method: DT2, error bars: Standard deviation
false_omission_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-ClWeigh
EE-ClWeigh
ItImp-ClWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

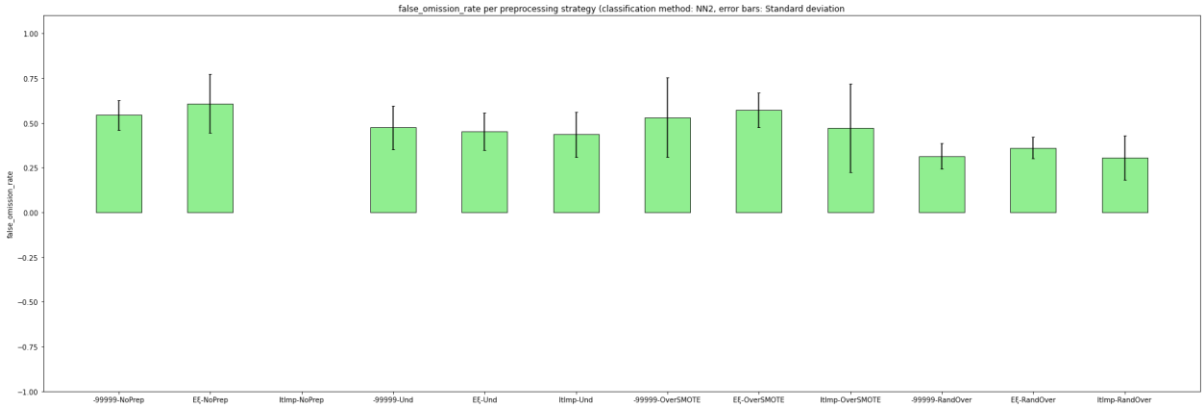

false_omission_rate per preprocessing strategy (classification method: DT3, error bars: Standard deviation
false_omission_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-ClWeigh
EE-ClWeigh
ItImp-ClWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

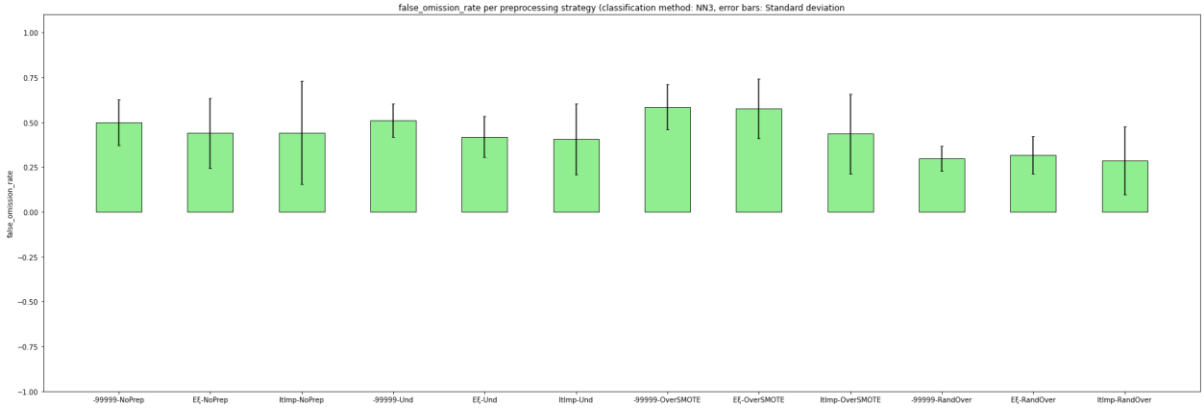

false_omission_rate per preprocessing strategy (classification method: DT4, error bars: Standard deviation
false_omission_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-ClWeigh
EE-ClWeigh
ItImp-ClWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

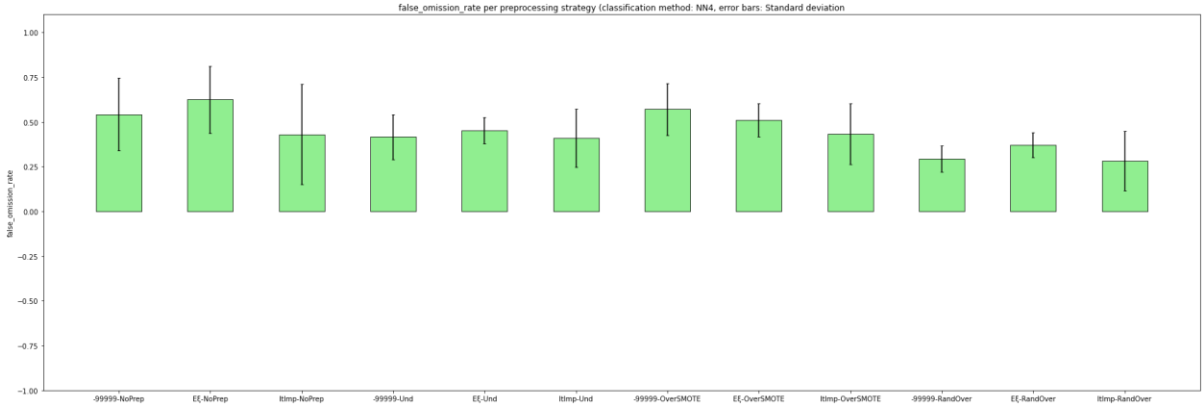

false_omission_rate per preprocessing strategy (classification method: RF1, error bars: Standard deviation
false_omission_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-Und
EE-Und
ItImp-Und
-99999-ClWeigh
EE-ClWeigh
ItImp-ClWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

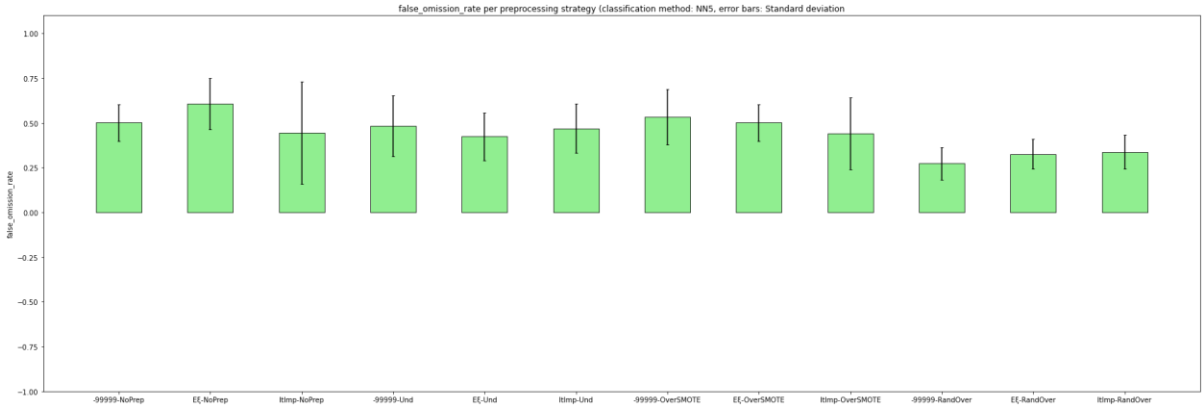

false_omission_rate per preprocessing strategy (classification method: RF2, error bars: Standard deviation
false_omission_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CtWeigh
EE-CtWeigh
ItImp-CtWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

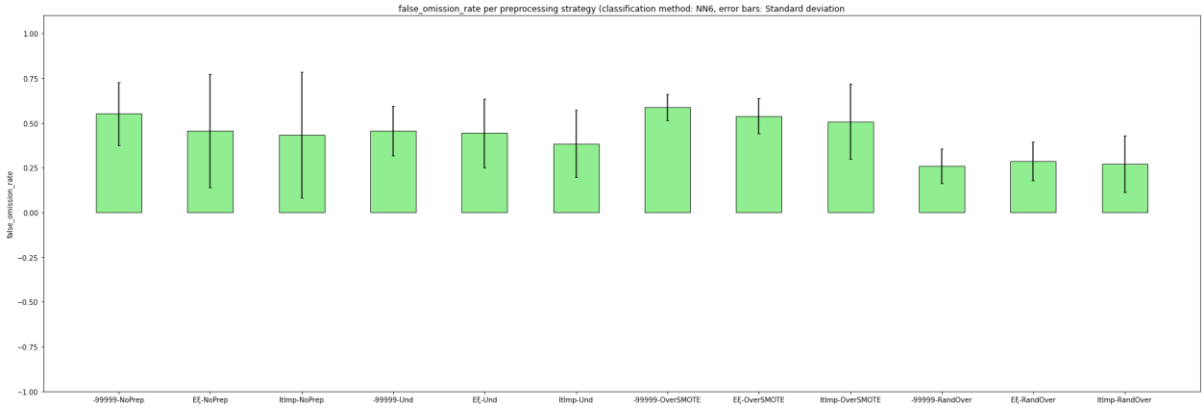

false_omission_rate per preprocessing strategy (classification method: NN1, error bars: Standard deviation
false_omission_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

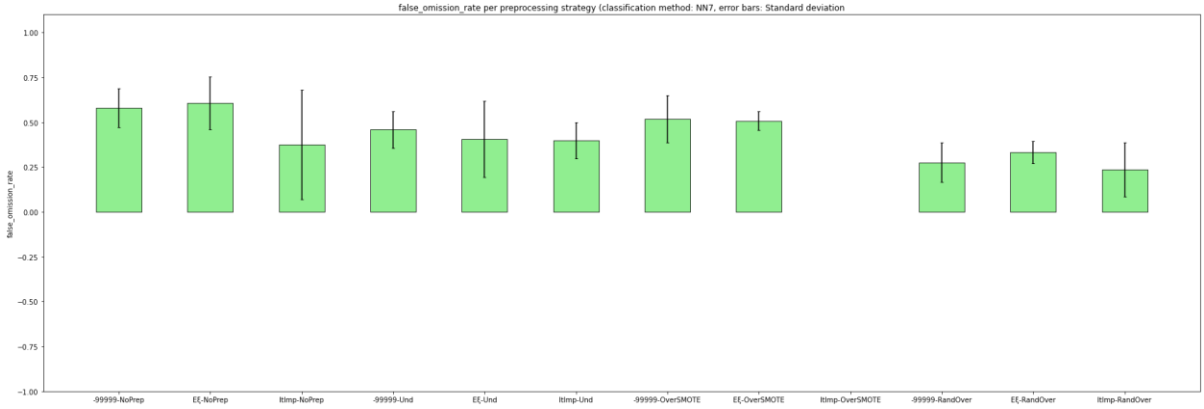

false_omission_rate per preprocessing strategy (classification method: NN2, error bars: Standard deviation
false_omission_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

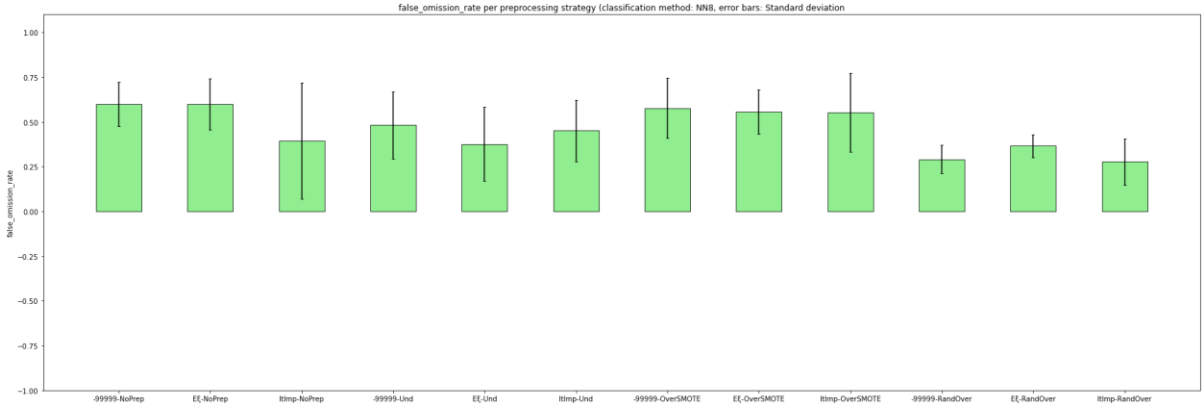

false_omission_rate per preprocessing strategy (classification method: NN3, error bars: Standard deviation
false_omission_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

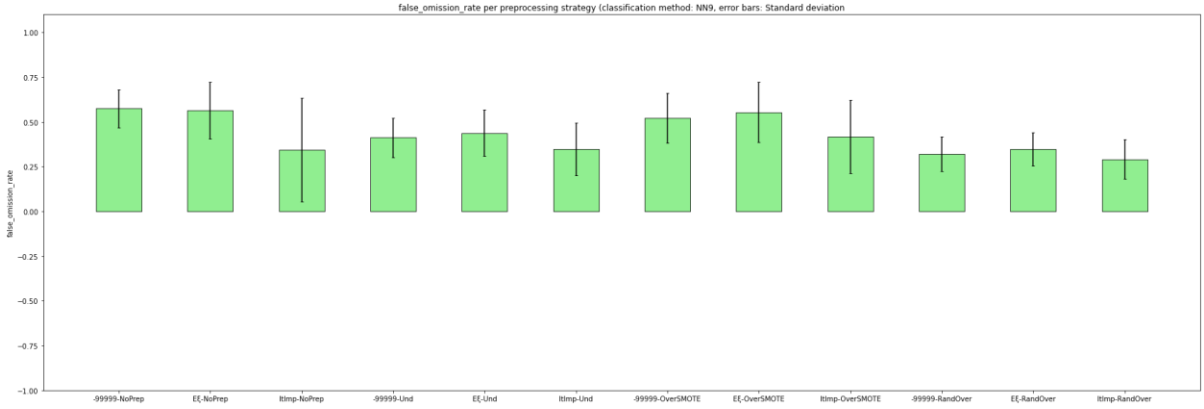

false_omission_rate per preprocessing strategy (classification method: NN4, error bars: Standard deviation
false_omission_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

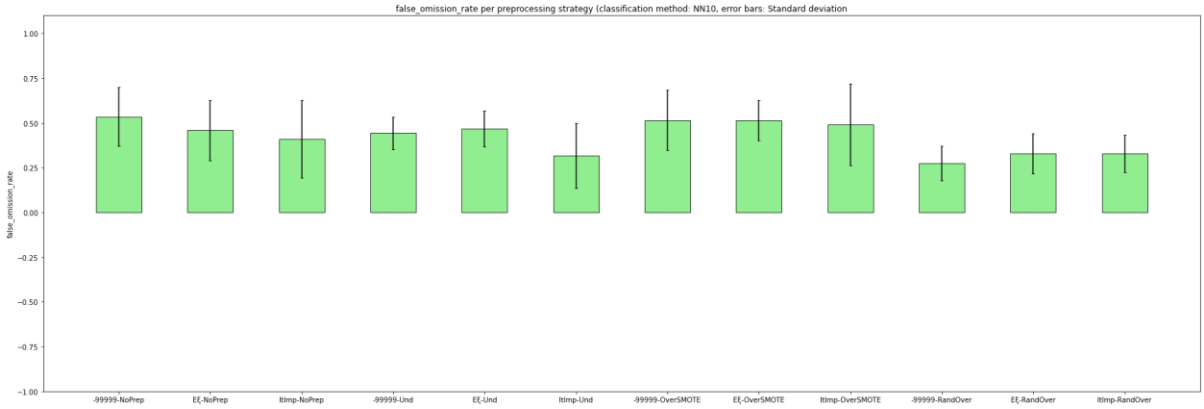

false_omission_rate per preprocessing strategy (classification method: NN5, error bars: Standard deviation
false_omission_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

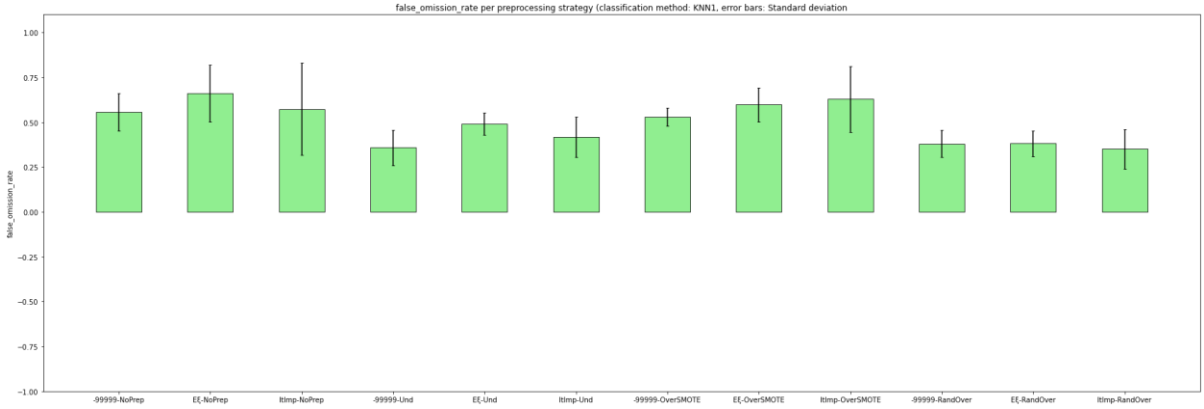

false_omission_rate per preprocessing strategy (classification method: NN6, error bars: Standard deviation
false_omission_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

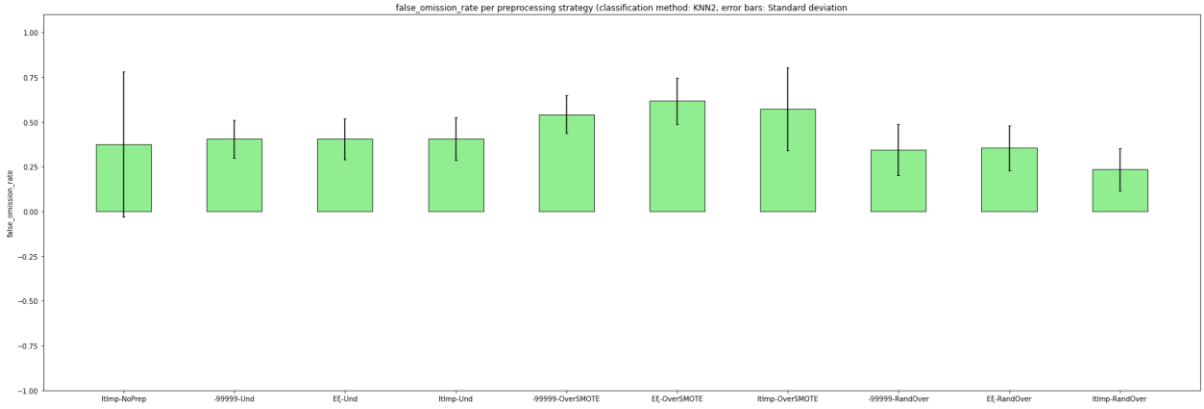

false_omission_rate per preprocessing strategy (classification method: NN7, error bars: Standard deviation
false_omission_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

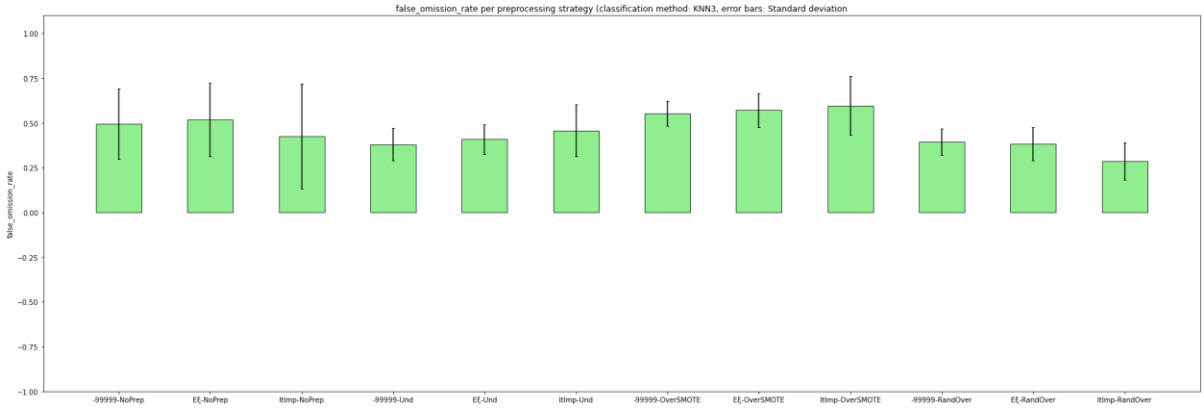

false_omission_rate per preprocessing strategy (classification method: NN8, error bars: Standard deviation
false_omission_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

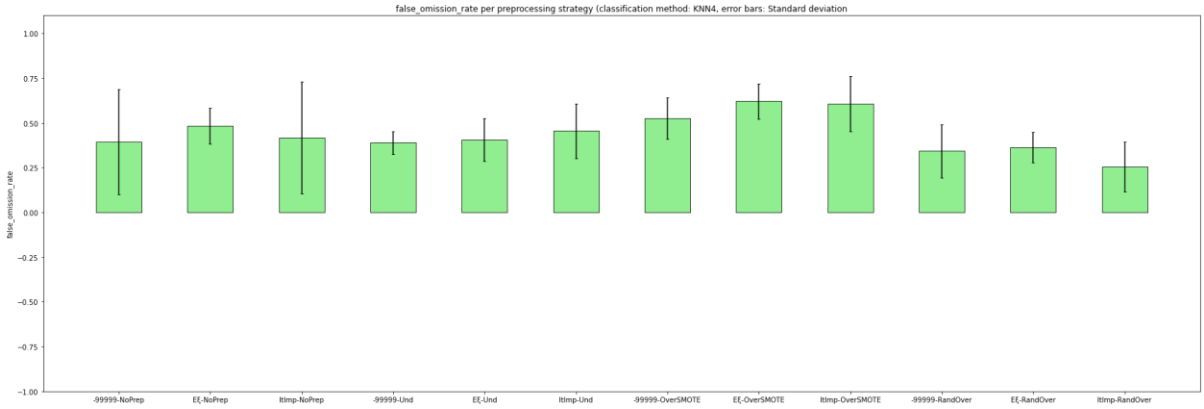

false_omission_rate per preprocessing strategy (classification method: NN9, error bars: Standard deviation
false_omission_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

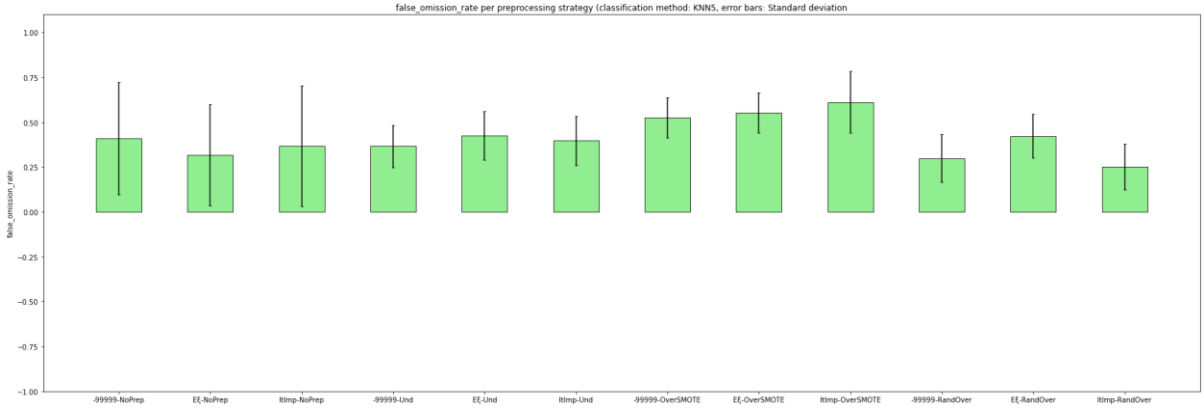

false_omission_rate per preprocessing strategy (classification method: NN10, error bars: Standard deviation
false_omission_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

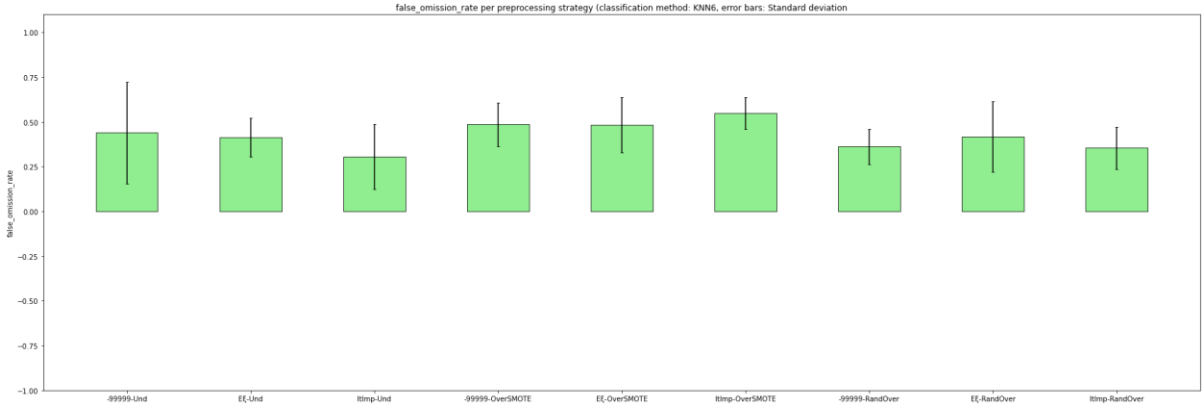

false_omission_rate per preprocessing strategy (classification method: KNN1, error bars: Standard deviation
false_omission_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

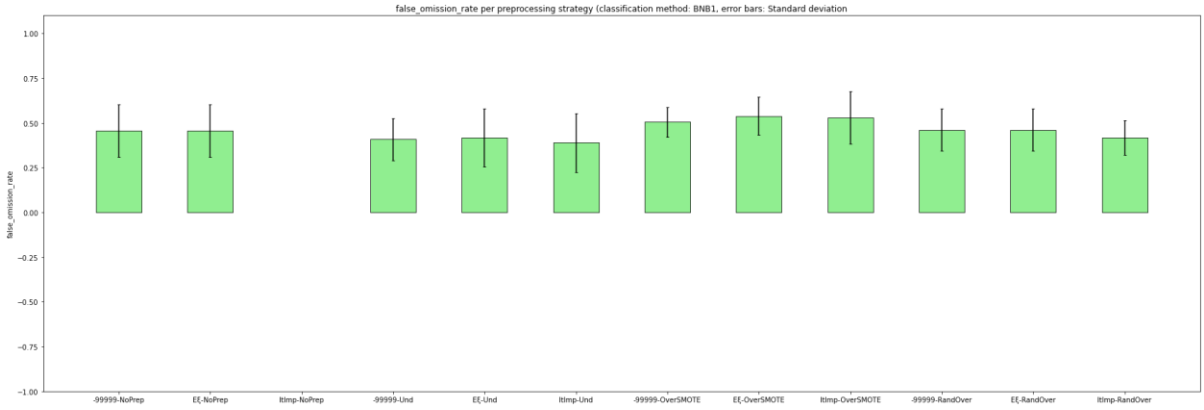

false_omission_rate per preprocessing strategy (classification method: KNN2, error bars: Standard deviation
false_omission_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

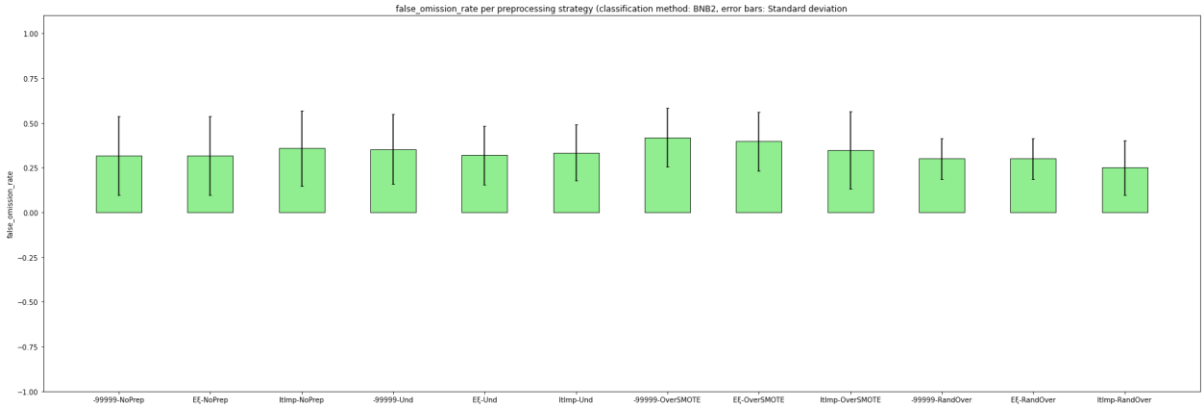

false_omission_rate per preprocessing strategy (classification method: KNN3, error bars: Standard deviation
false_omission_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

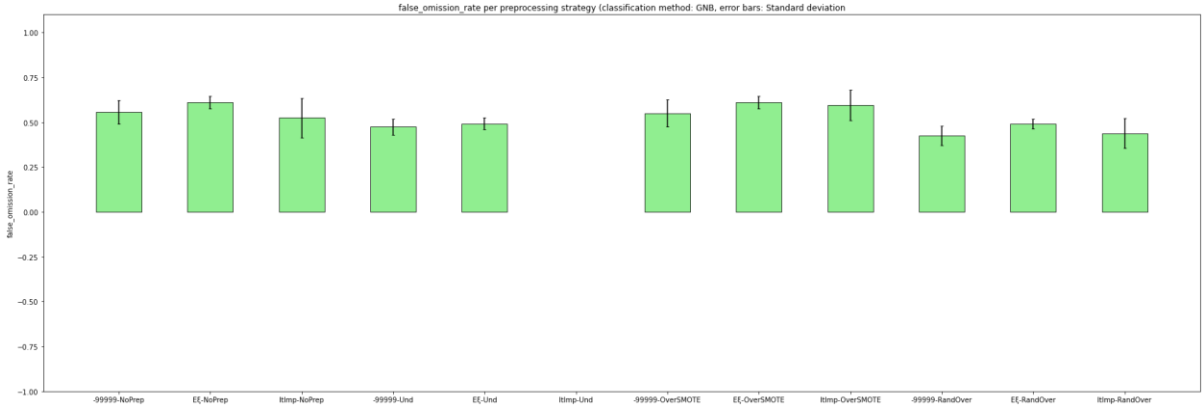

false_omission_rate per preprocessing strategy (classification method: KNN4, error bars: Standard deviation
false_omission_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

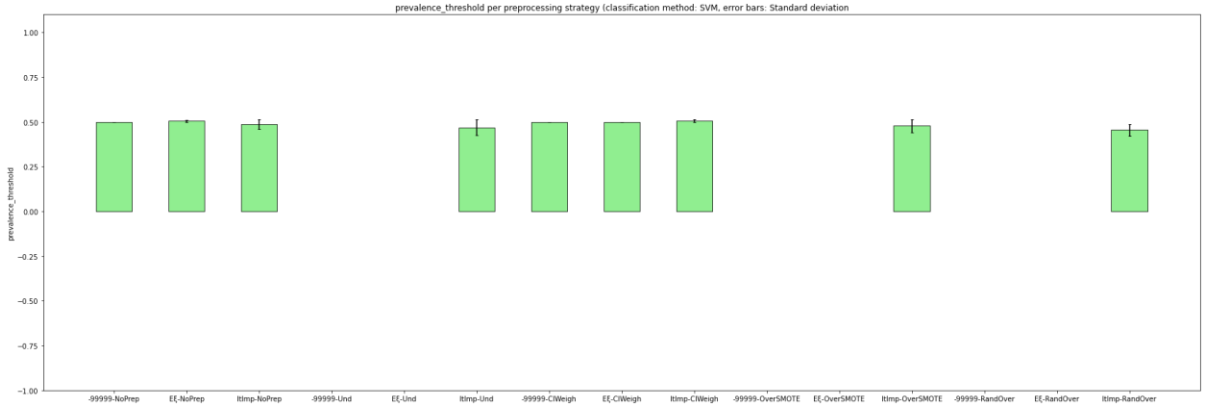

false_omission_rate per preprocessing strategy (classification method: KNN5, error bars: Standard deviation
false_omission_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

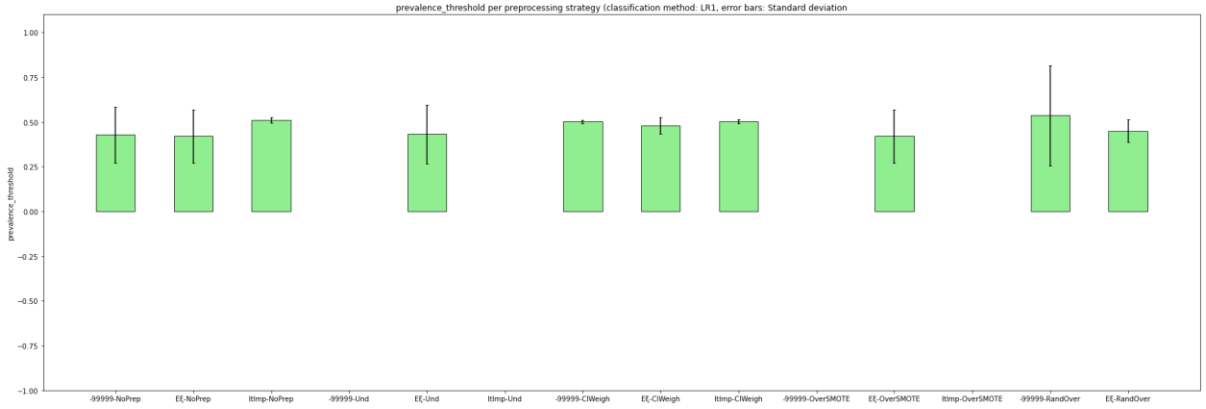

false_omission_rate per preprocessing strategy (classification method: KNN6, error bars: Standard deviation
false_omission_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

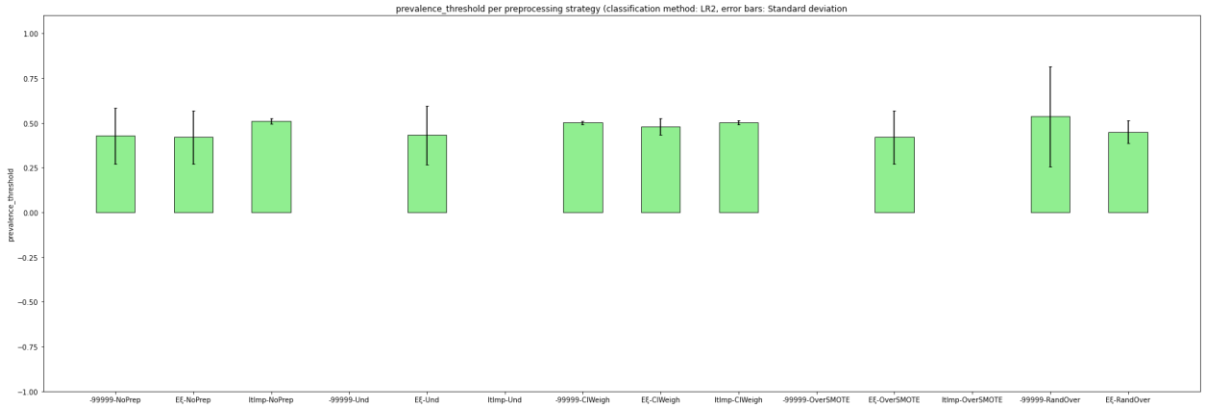

false_omission_rate per preprocessing strategy (classification method: BNB1, error bars: Standard deviation
false_omission_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

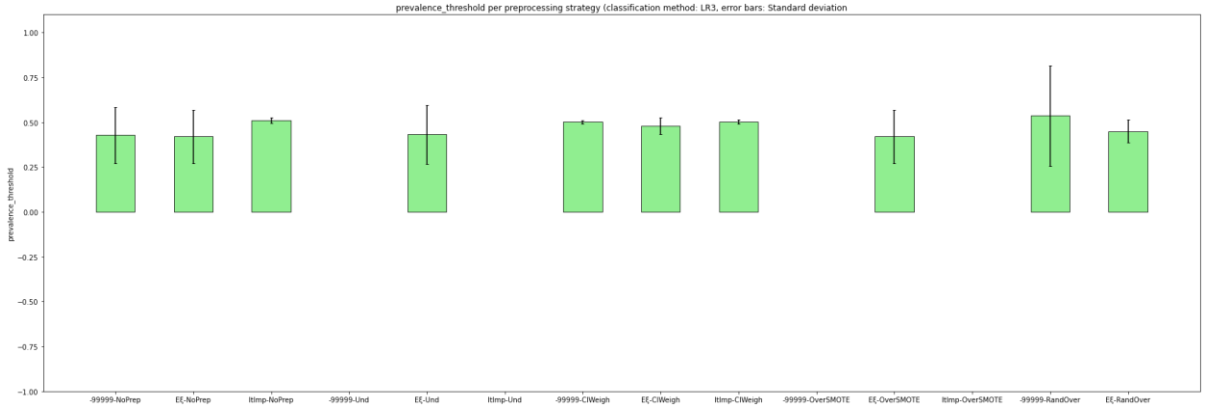

false_omission_rate per preprocessing strategy (classification method: BNB2, error bars: Standard deviation
false_omission_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

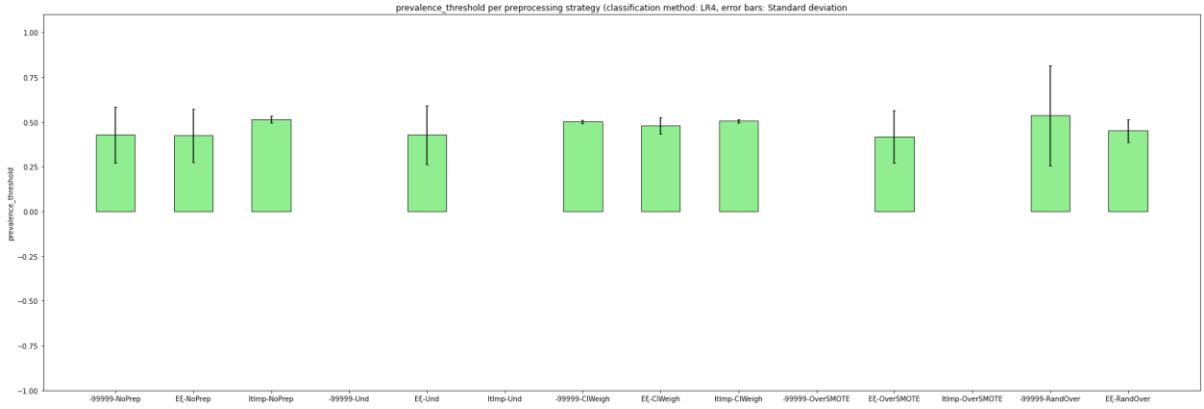

false_omission_rate per preprocessing strategy (classification method: GNB, error bars: Standard deviation
false_omission_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

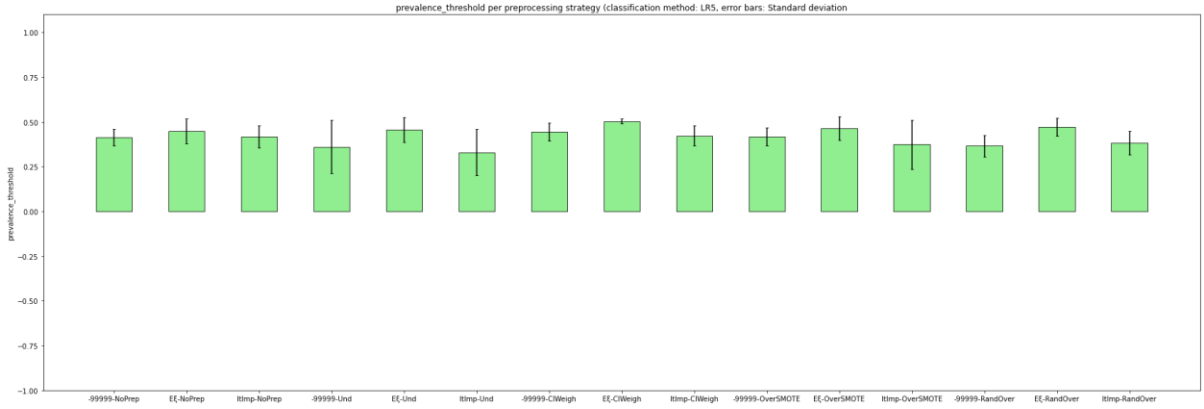

prevalence_threshold per preprocessing strategy (classification method: SVM, error bars: Standard deviation
prevalence_threshold
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

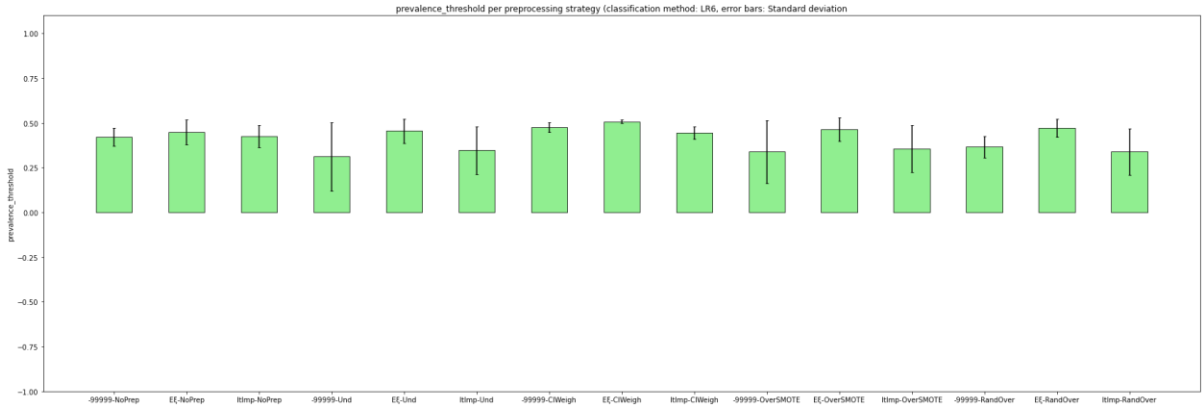

prevalence_threshold per preprocessing strategy (classification method: LR1, error bars: Standard deviation
prevalence_threshold
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver

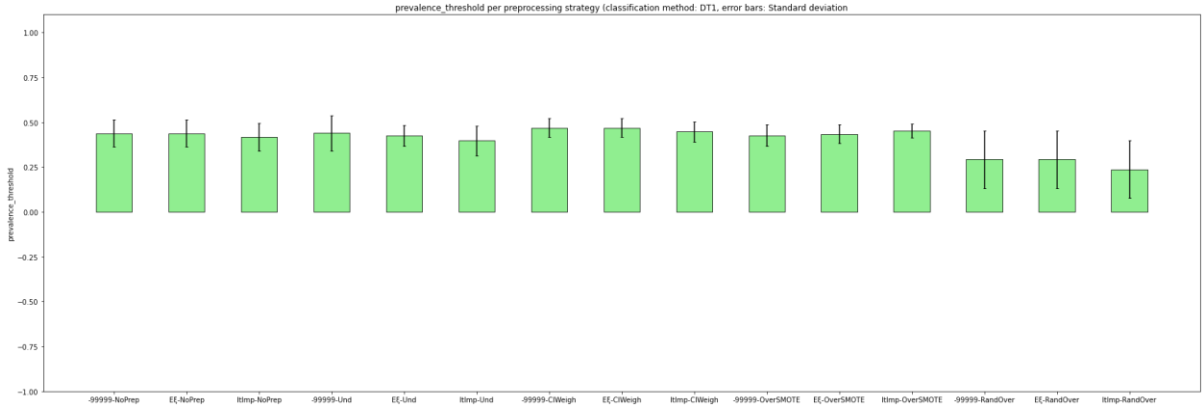

prevalence_threshold per preprocessing strategy (classification method: LR2, error bars: Standard deviation
prevalence_threshold
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver

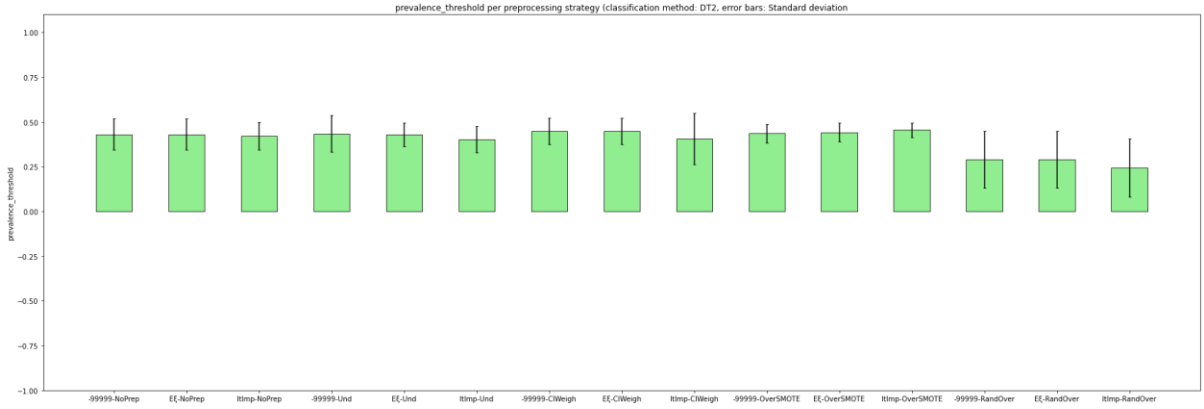

prevalence_threshold per preprocessing strategy (classification method: LR3, error bars: Standard deviation
prevalence_threshold
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver

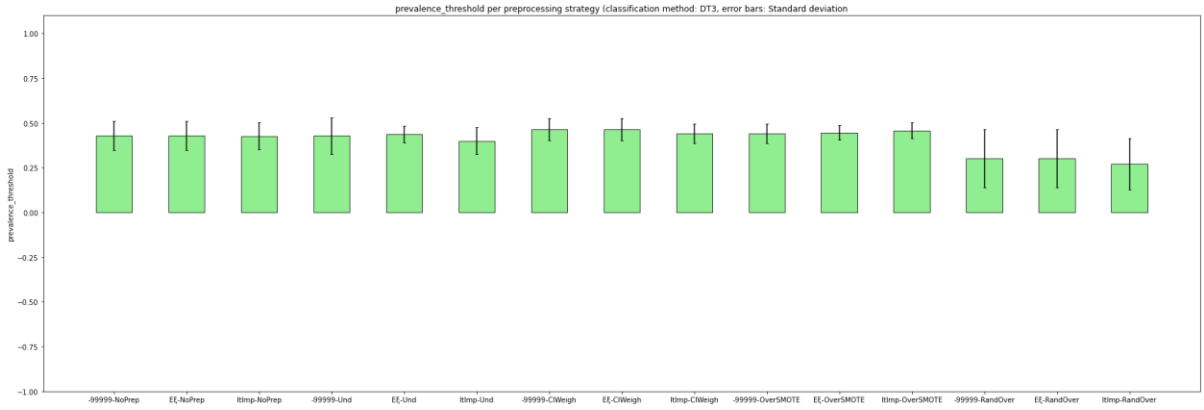

prevalence_threshold per preprocessing strategy (classification method: LR4, error bars: Standard deviation
prevalence_threshold
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver

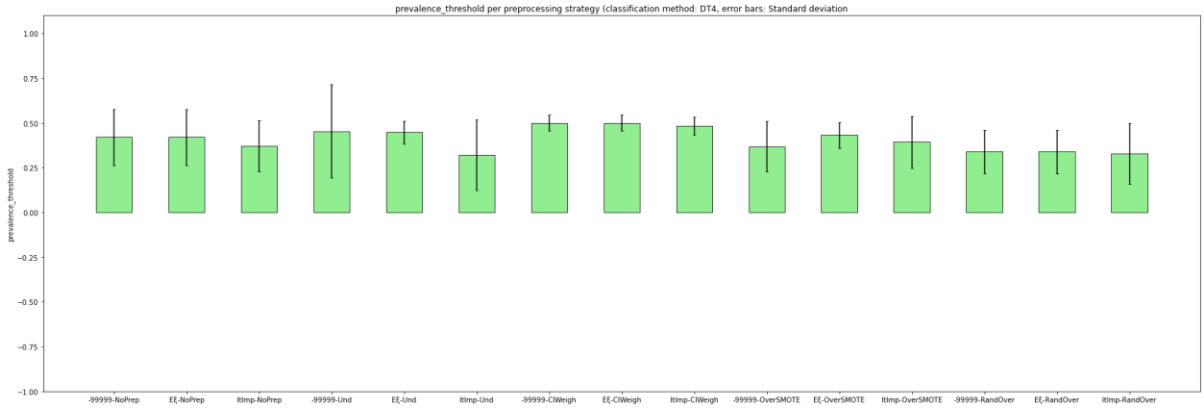

prevalence_threshold per preprocessing strategy (classification method: LR5, error bars: Standard deviation
prevalence_threshold
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

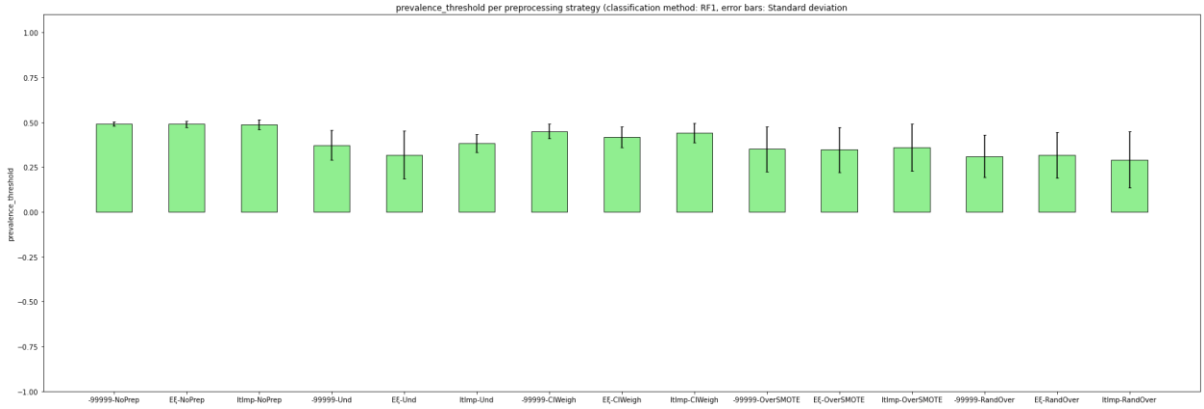

prevalence_threshold per preprocessing strategy (classification method: LR6, error bars: Standard deviation
prevalence_threshold
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

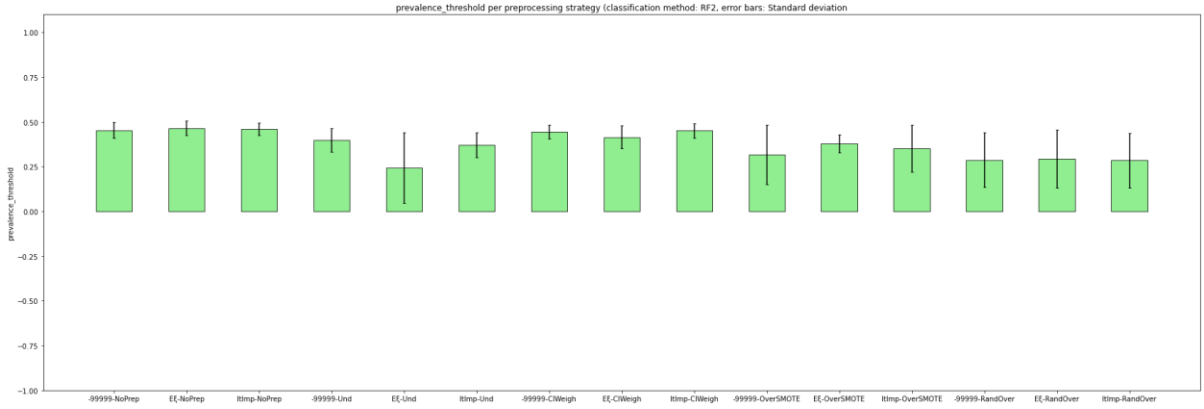

prevalence_threshold per preprocessing strategy (classification method: DT1, error bars: Standard deviation
prevalence_threshold
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

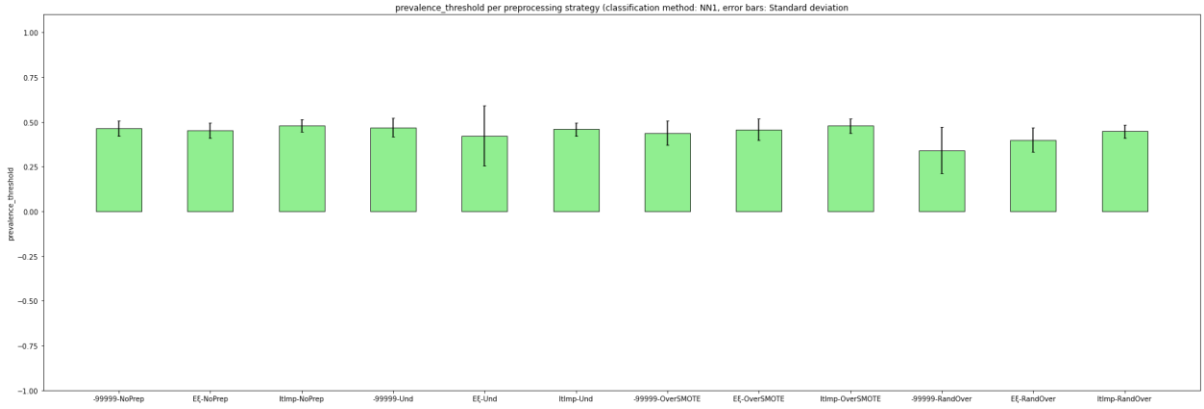

prevalence_threshold per preprocessing strategy (classification method: DT2, error bars: Standard deviation
prevalence_threshold
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

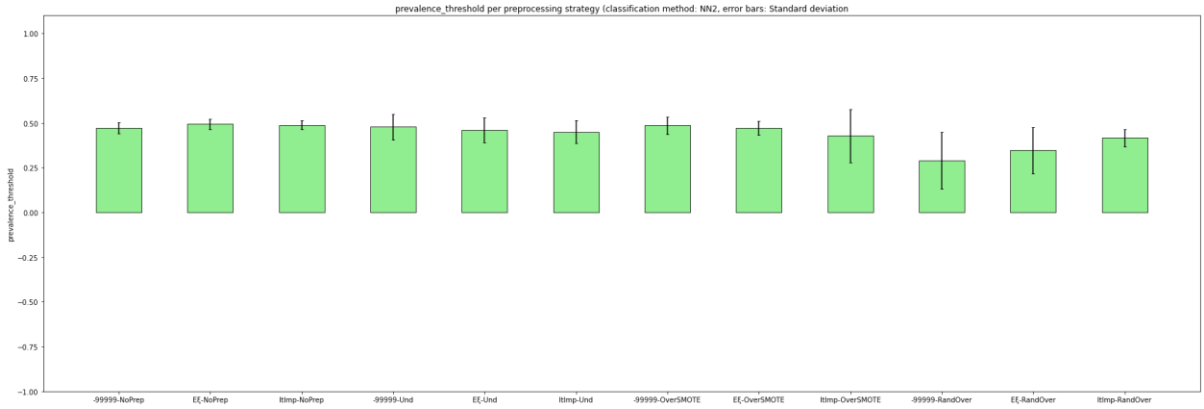

prevalence_threshold per preprocessing strategy (classification method: DT3, error bars: Standard deviation
prevalence_threshold
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

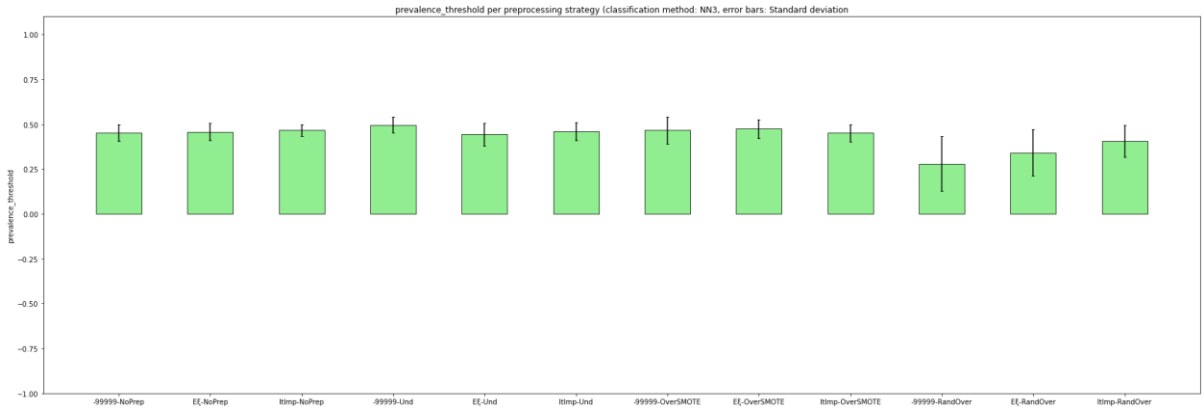

prevalence_threshold per preprocessing strategy (classification method: DT4, error bars: Standard deviation
prevalence_threshold
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

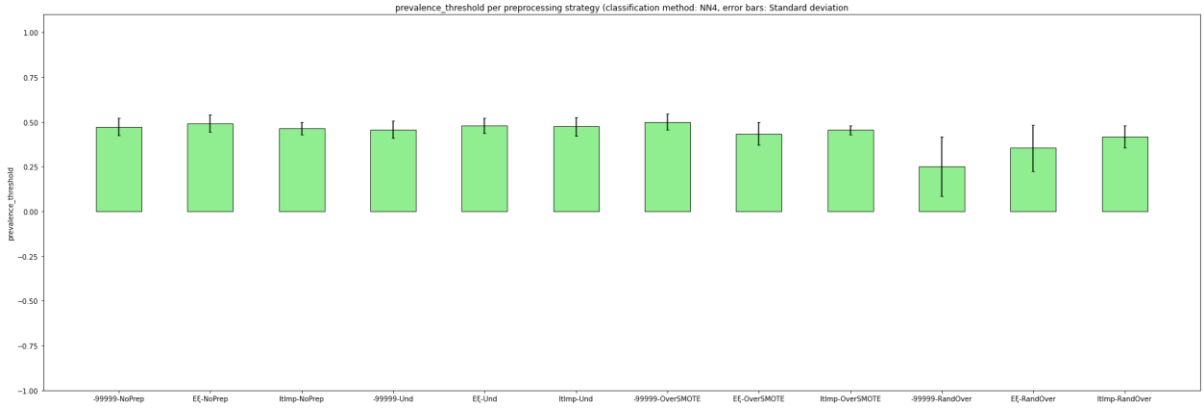

prevalence_threshold per preprocessing strategy (classification method: RF1, error bars: Standard deviation
prevalence_threshold
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

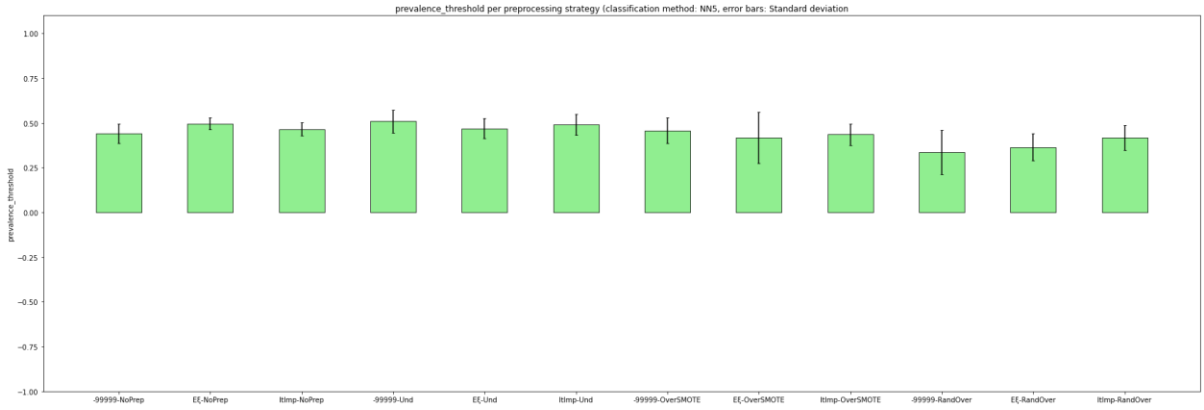

prevalence_threshold per preprocessing strategy (classification method: RF2, error bars: Standard deviation
prevalence_threshold
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

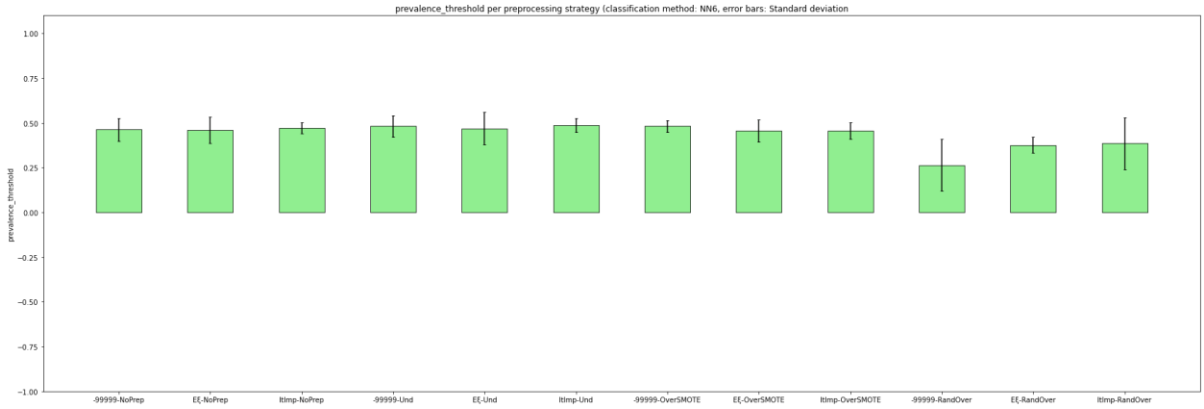

prevalence_threshold per preprocessing strategy (classification method: NN1, error bars: Standard deviation
prevalence_threshold
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

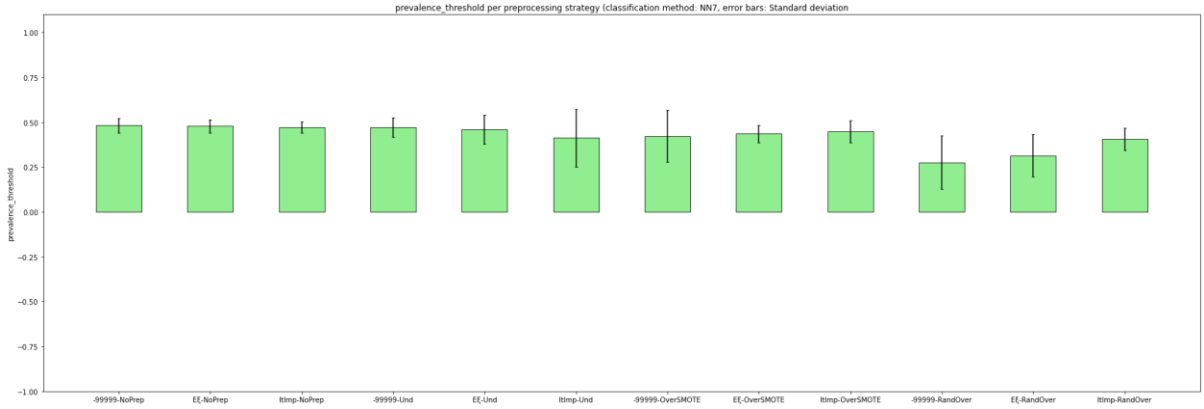

prevalence_threshold per preprocessing strategy (classification method: NN2, error bars: Standard deviation
prevalence_threshold
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

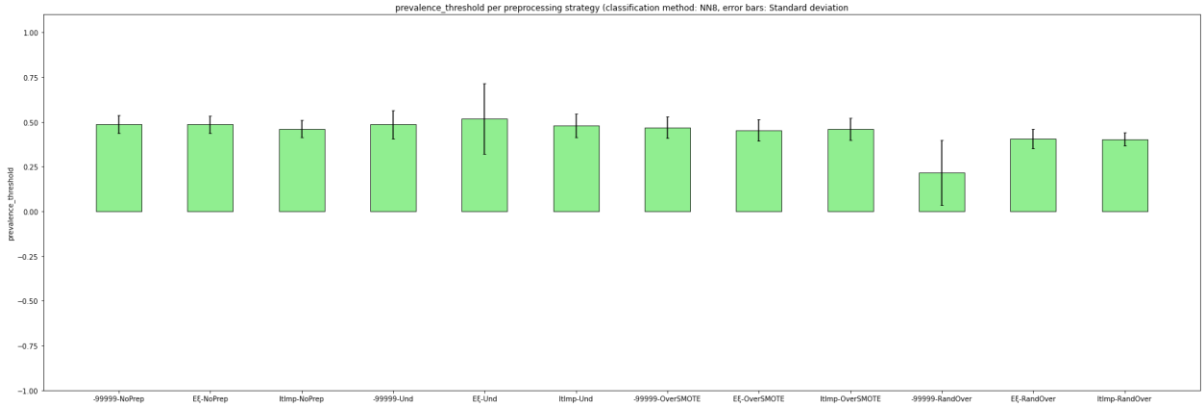

prevalence_threshold per preprocessing strategy (classification method: NN3, error bars: Standard deviation
prevalence_threshold
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

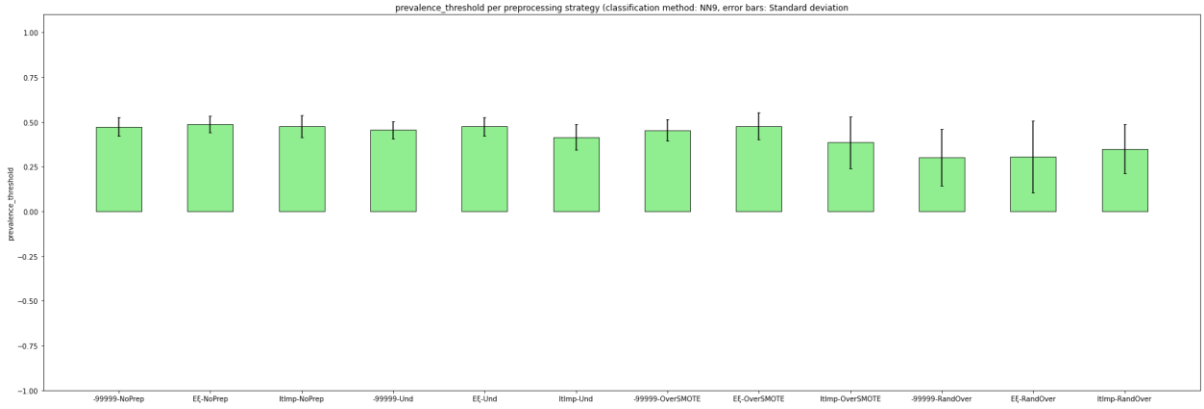

prevalence_threshold per preprocessing strategy (classification method: NN4, error bars: Standard deviation
prevalence_threshold
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ltImp-NoPrep
-99999-Und
EE-Und
ltImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ltImp-OverSMOTE
-99999-RandOver
EE-RandOver
ltImp-RandOver

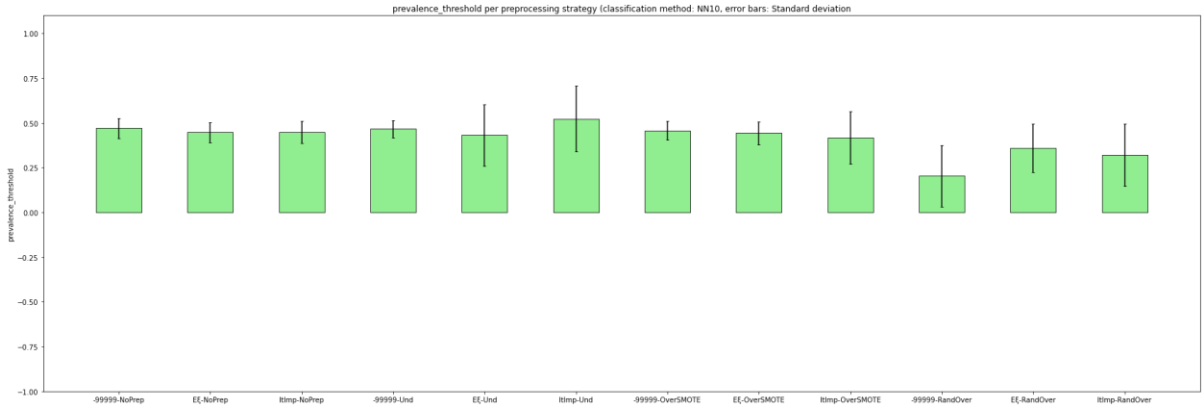

prevalence_threshold per preprocessing strategy (classification method: NN5, error bars: Standard deviation
prevalence_threshold
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ltImp-NoPrep
-99999-Und
EE-Und
ltImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ltImp-OverSMOTE
-99999-RandOver
EE-RandOver
ltImp-RandOver

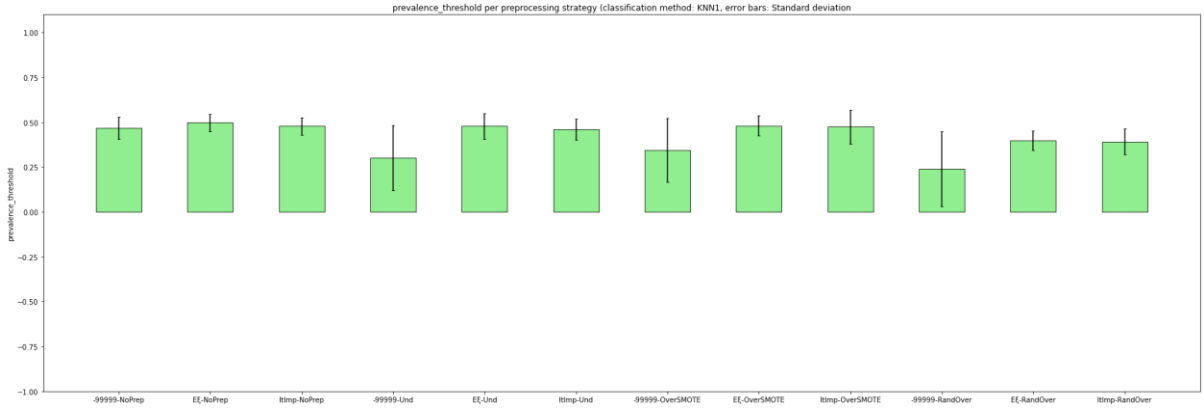

prevalence_threshold per preprocessing strategy (classification method: NN6, error bars: Standard deviation
prevalence_threshold
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ltImp-NoPrep
-99999-Und
EE-Und
ltImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ltImp-OverSMOTE
-99999-RandOver
EE-RandOver
ltImp-RandOver

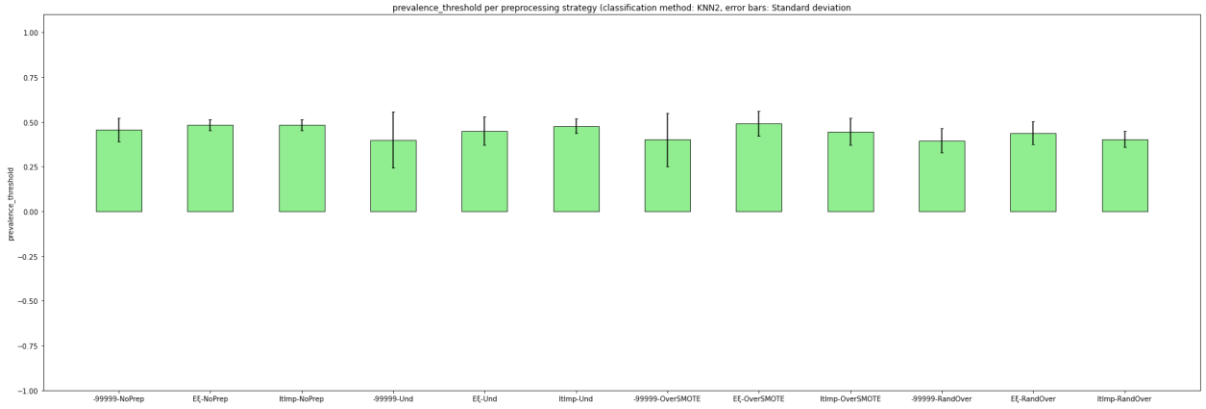

prevalence_threshold per preprocessing strategy (classification method: NN7, error bars: Standard deviation
prevalence_threshold
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ltImp-NoPrep
-99999-Und
EE-Und
ltImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ltImp-OverSMOTE
-99999-RandOver
EE-RandOver
ltImp-RandOver

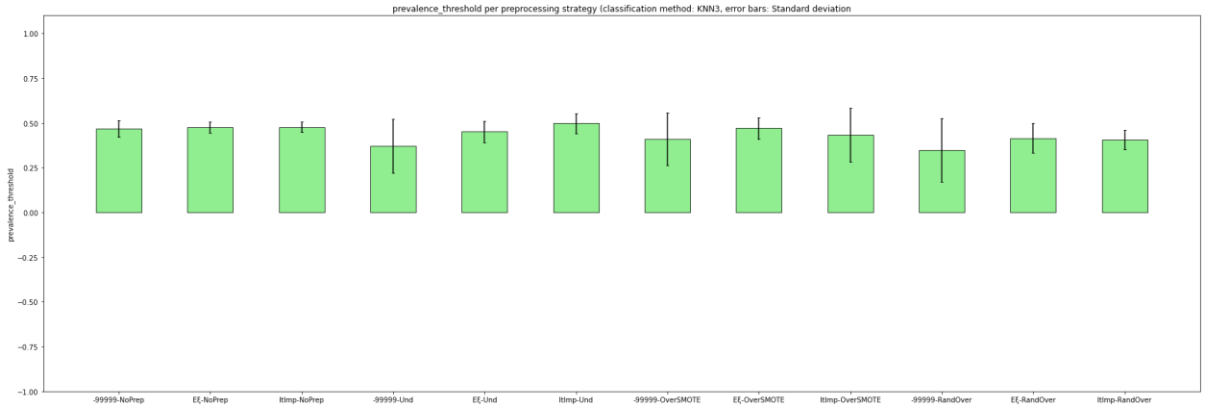

prevalence_threshold per preprocessing strategy (classification method: NN8, error bars: Standard deviation
prevalence_threshold
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

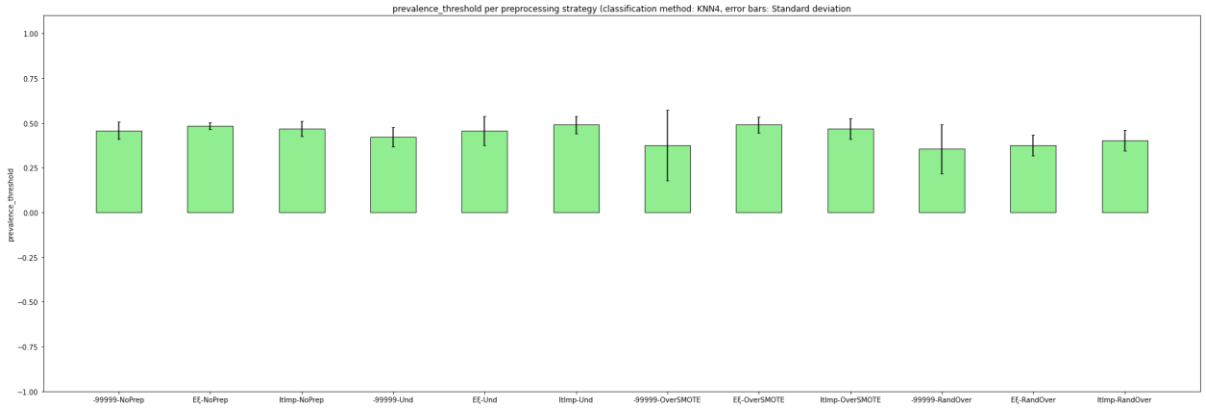

prevalence_threshold per preprocessing strategy (classification method: NN9, error bars: Standard deviation
prevalence_threshold
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

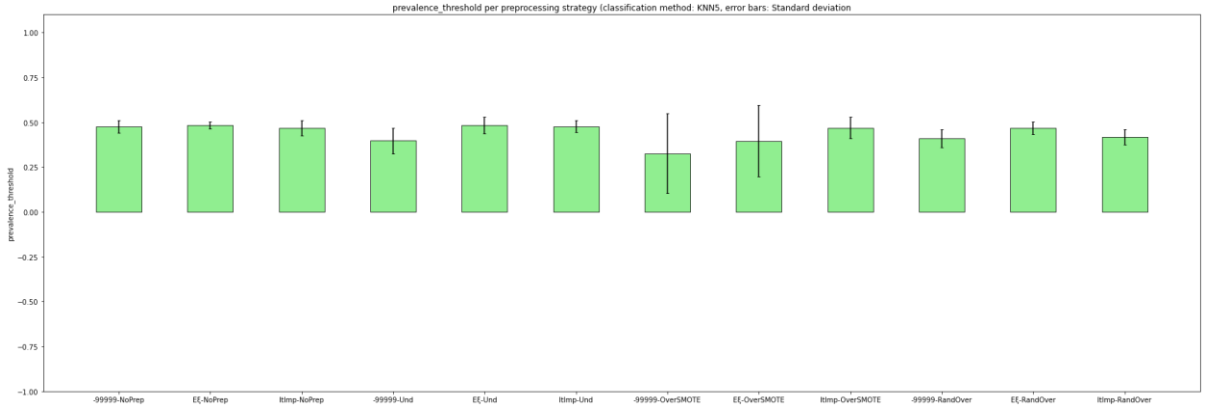

prevalence_threshold per preprocessing strategy (classification method: NN10, error bars: Standard deviation
prevalence_threshold
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

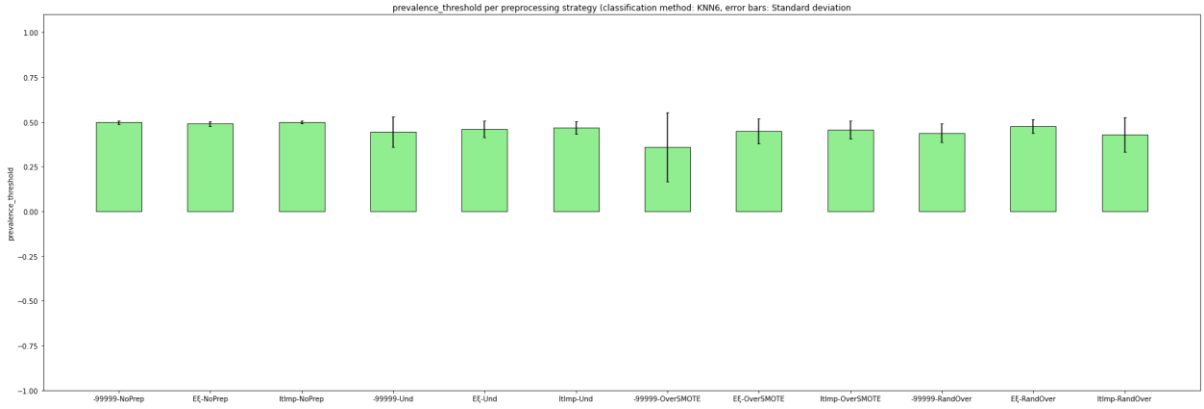

prevalence_threshold per preprocessing strategy (classification method: KNN1, error bars: Standard deviation
prevalence_threshold
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

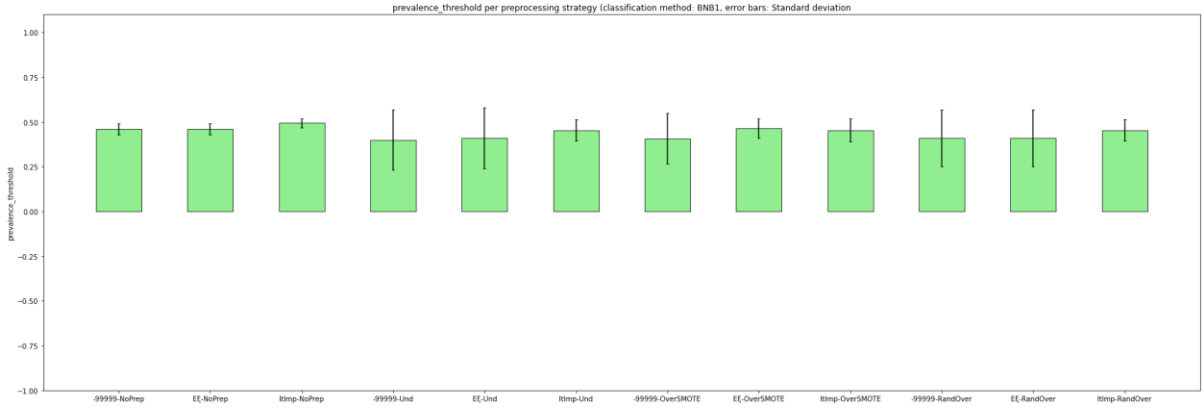

prevalence_threshold per preprocessing strategy (classification method: KNN2, error bars: Standard deviation
prevalence_threshold
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ltImp-NoPrep
-99999-Und
EE-Und
ltImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ltImp-OverSMOTE
-99999-RandOver
EE-RandOver
ltImp-RandOver

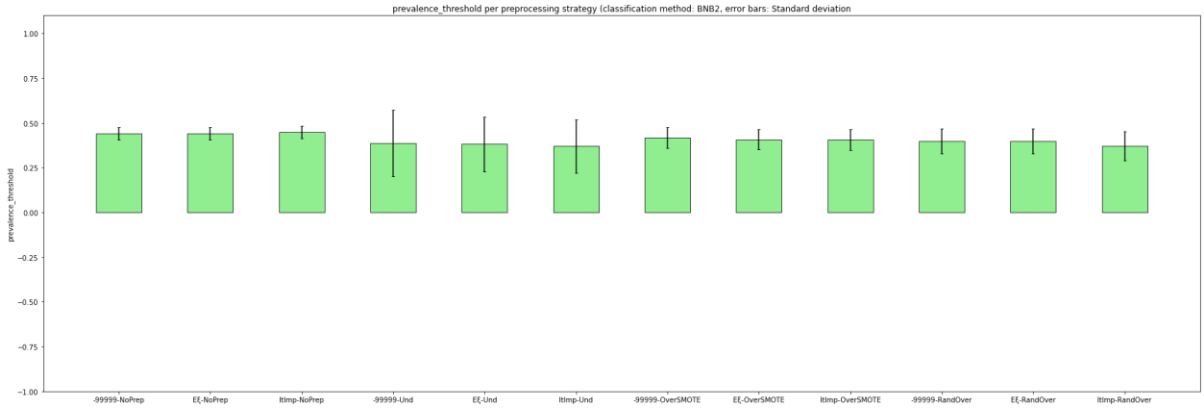

prevalence_threshold per preprocessing strategy (classification method: KNN3, error bars: Standard deviation
prevalence_threshold
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ltImp-NoPrep
-99999-Und
EE-Und
ltImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ltImp-OverSMOTE
-99999-RandOver
EE-RandOver
ltImp-RandOver

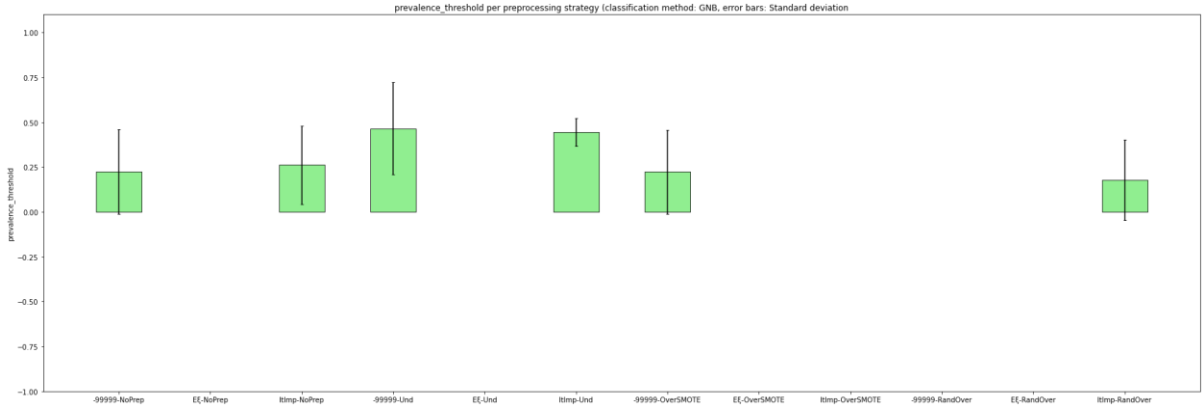

prevalence_threshold per preprocessing strategy (classification method: KNN4, error bars: Standard deviation
prevalence_threshold
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ltImp-NoPrep
-99999-Und
EE-Und
ltImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ltImp-OverSMOTE
-99999-RandOver
EE-RandOver
ltImp-RandOver

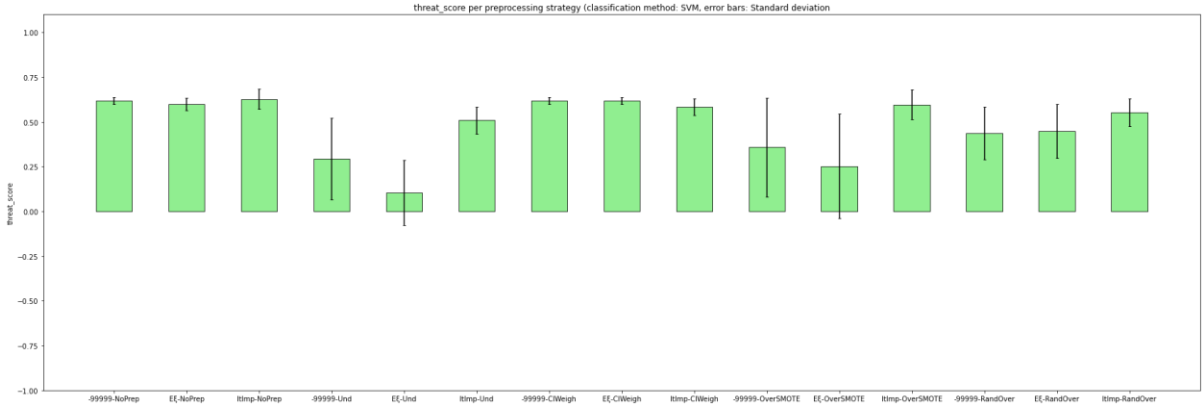

prevalence_threshold per preprocessing strategy (classification method: KNN5, error bars: Standard deviation
prevalence_threshold
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ltImp-NoPrep
-99999-Und
EE-Und
ltImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ltImp-OverSMOTE
-99999-RandOver
EE-RandOver
ltImp-RandOver

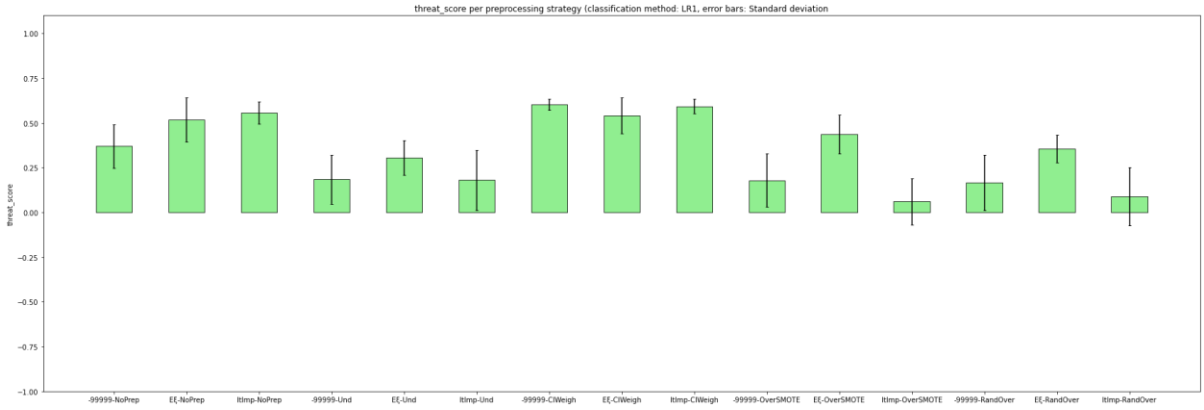

prevalence_threshold per preprocessing strategy (classification method: KNN6, error bars: Standard deviation
prevalence_threshold
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

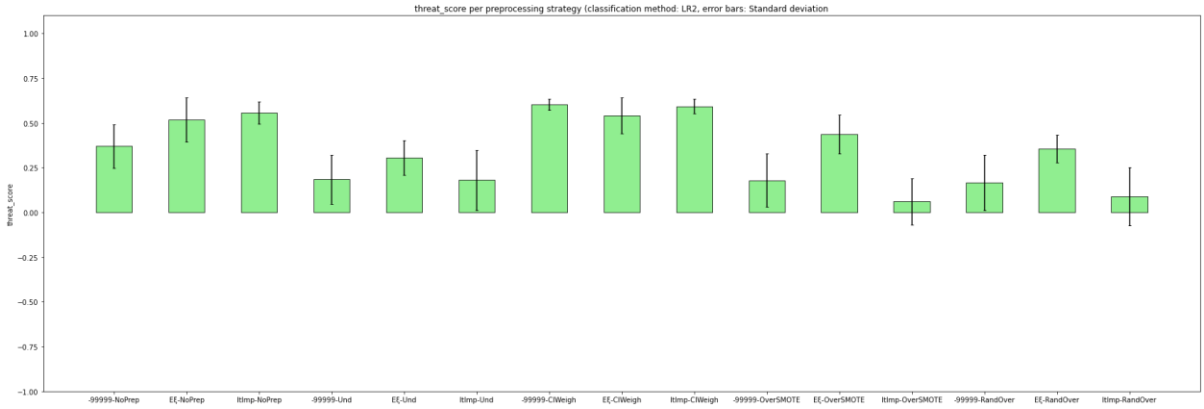

prevalence_threshold per preprocessing strategy (classification method: BNB1, error bars: Standard deviation
prevalence_threshold
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

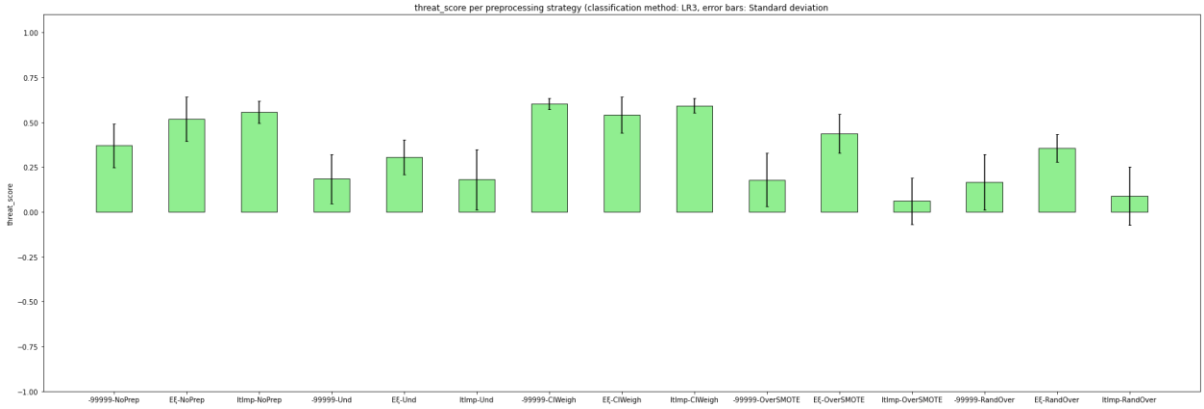

prevalence_threshold per preprocessing strategy (classification method: BNB2, error bars: Standard deviation
prevalence_threshold
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

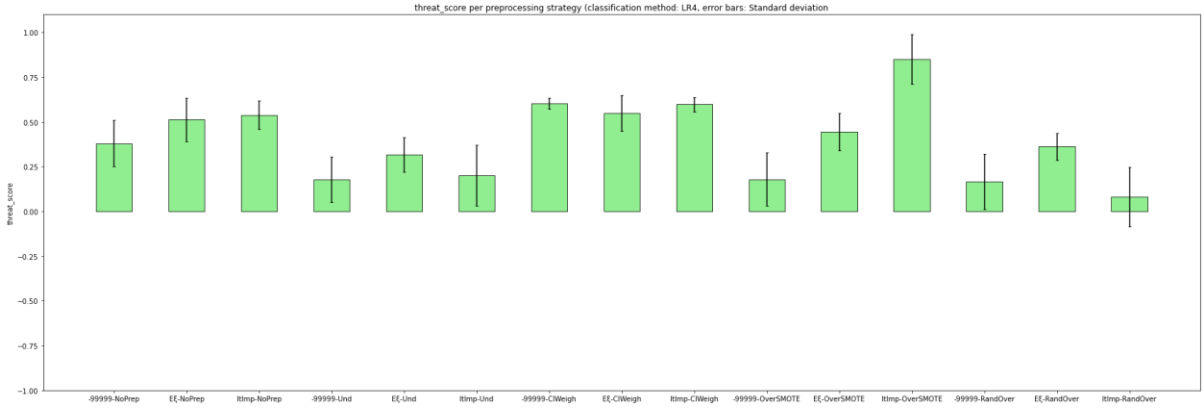

prevalence_threshold per preprocessing strategy (classification method: GNB, error bars: Standard deviation
prevalence_threshold
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

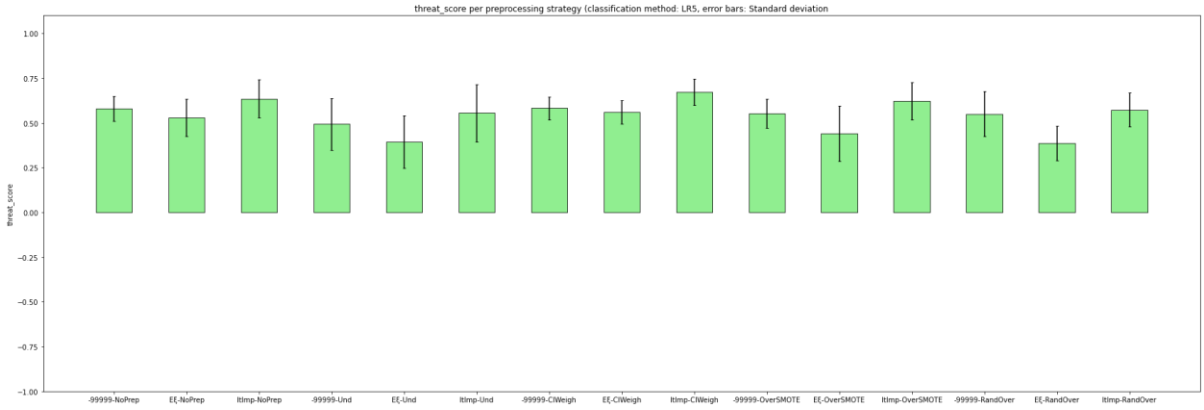

threat_score per preprocessing strategy (classification method: SVM, error bars: Standard deviation
threat_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

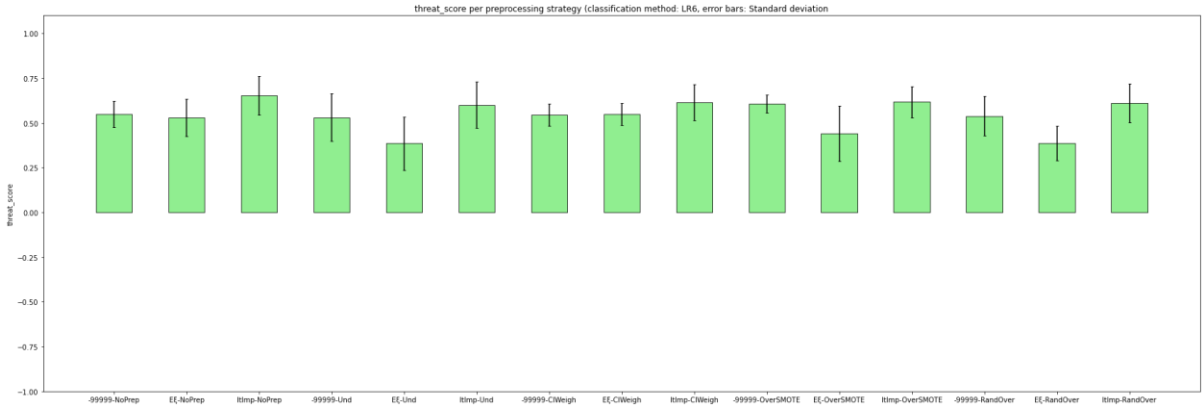

threat_score per preprocessing strategy (classification method: LR1, error bars: Standard deviation
threat_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

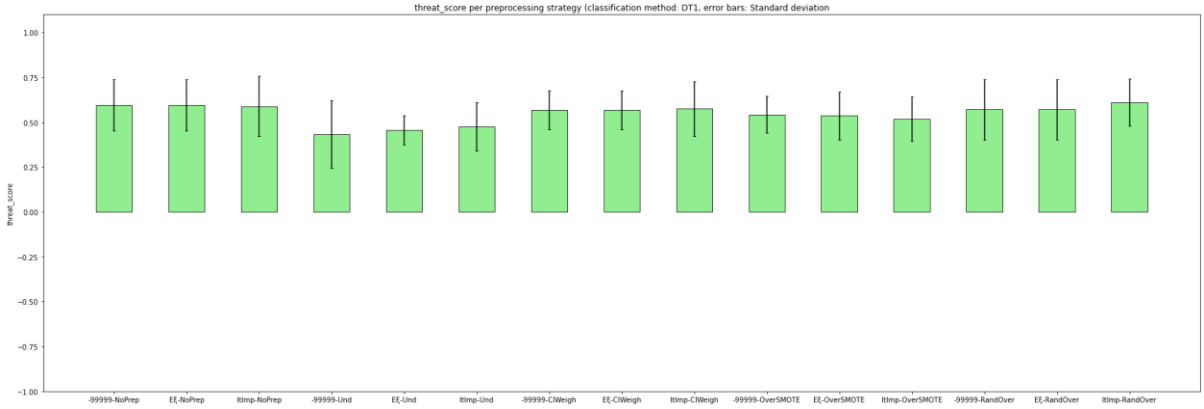

threat_score per preprocessing strategy (classification method: LR2, error bars: Standard deviation
threat_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

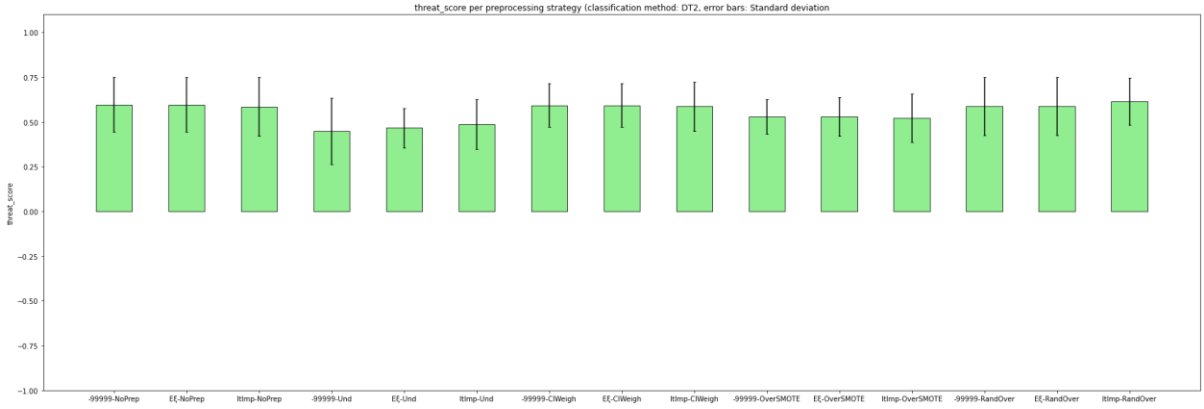

threat_score per preprocessing strategy (classification method: LR3, error bars: Standard deviation
threat_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

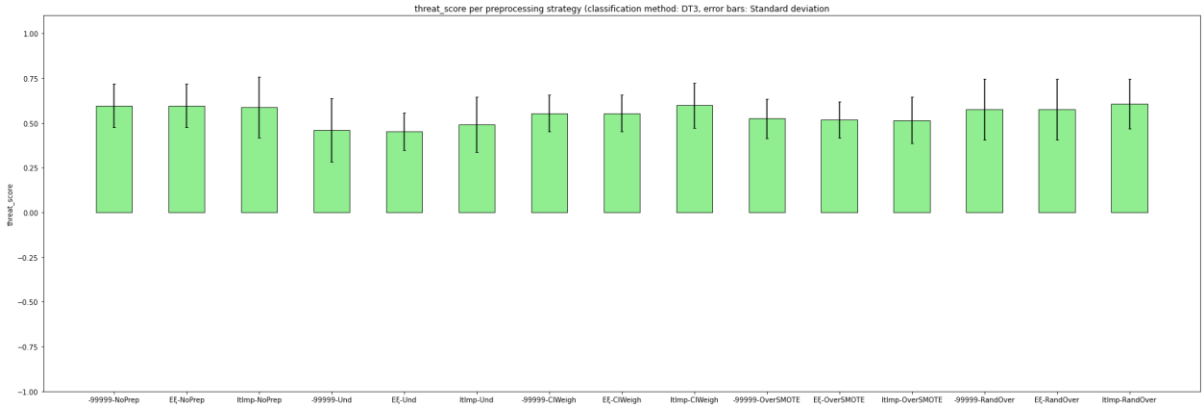

threat_score per preprocessing strategy (classification method: LR4, error bars: Standard deviation
threat_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

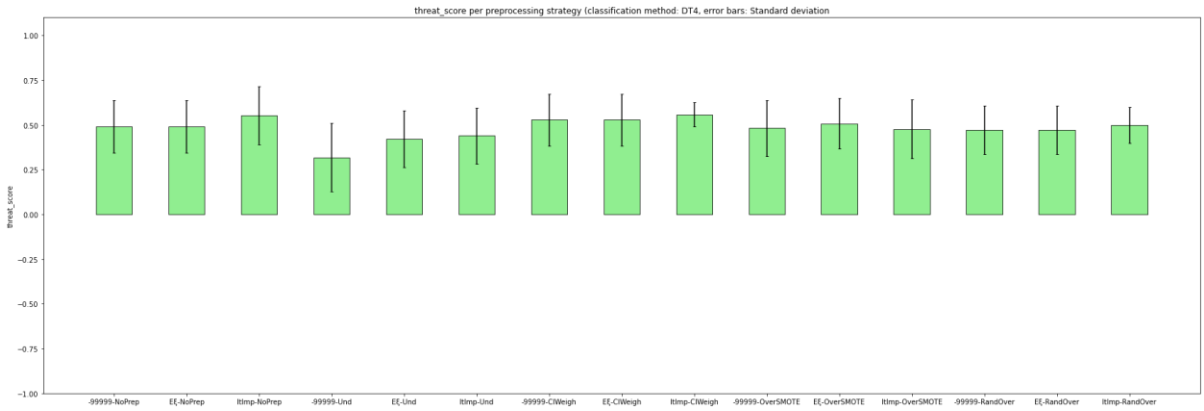

threat_score per preprocessing strategy (classification method: LR5, error bars: Standard deviation
threat_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

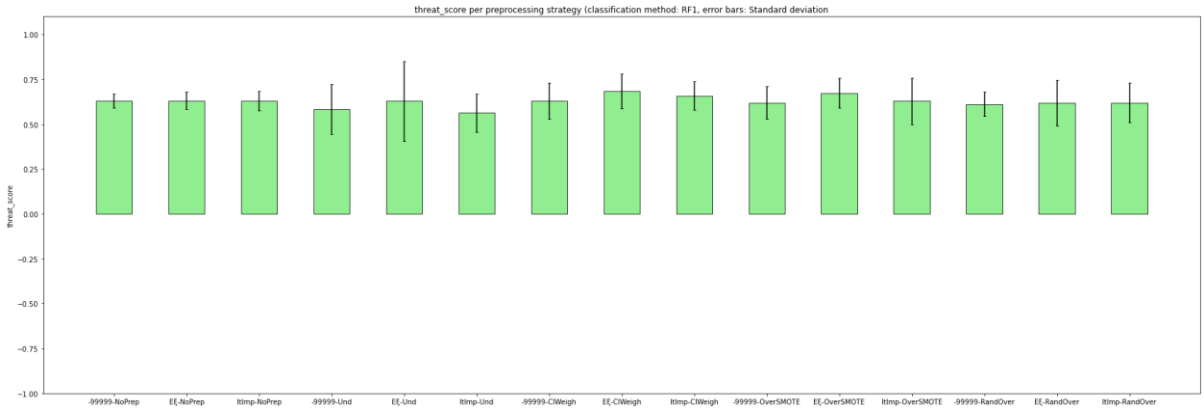

threat_score per preprocessing strategy (classification method: LR6, error bars: Standard deviation
threat_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

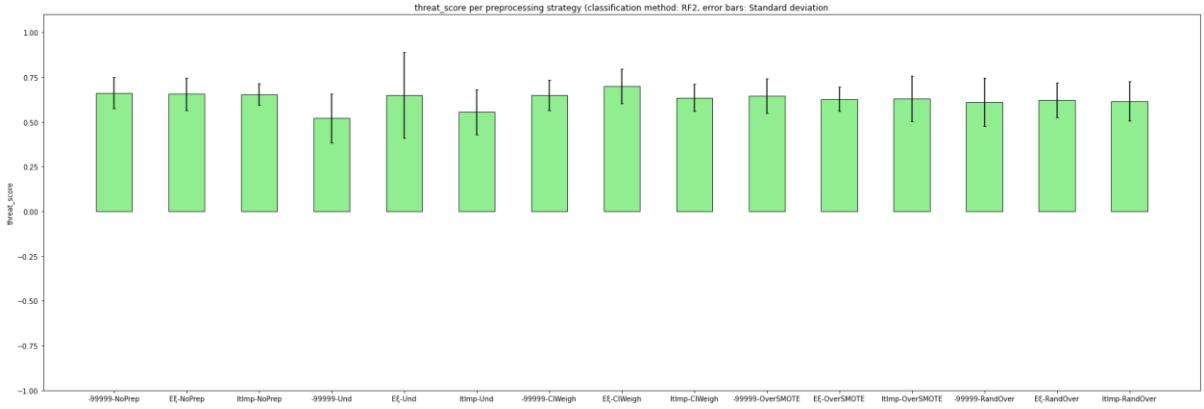

threat_score per preprocessing strategy (classification method: DT1, error bars: Standard deviation
threat_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

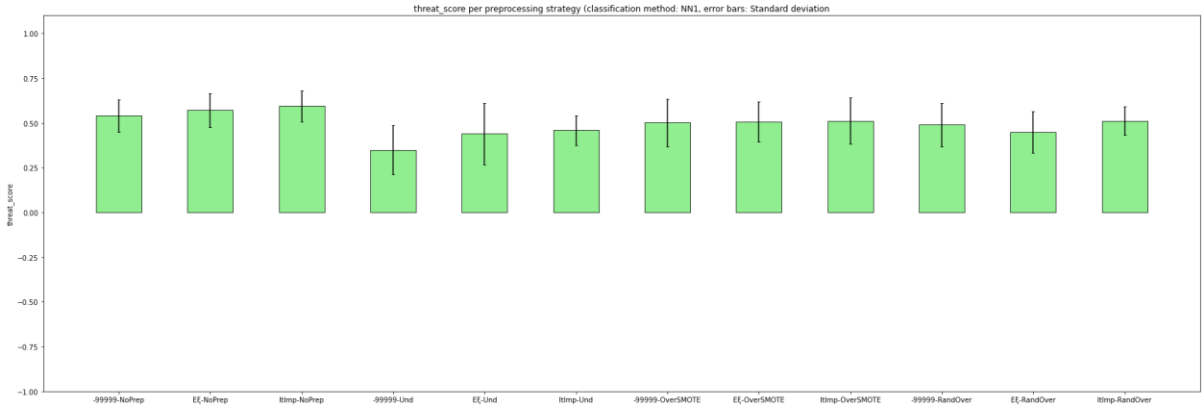

threat_score per preprocessing strategy (classification method: DT2, error bars: Standard deviation
threat_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

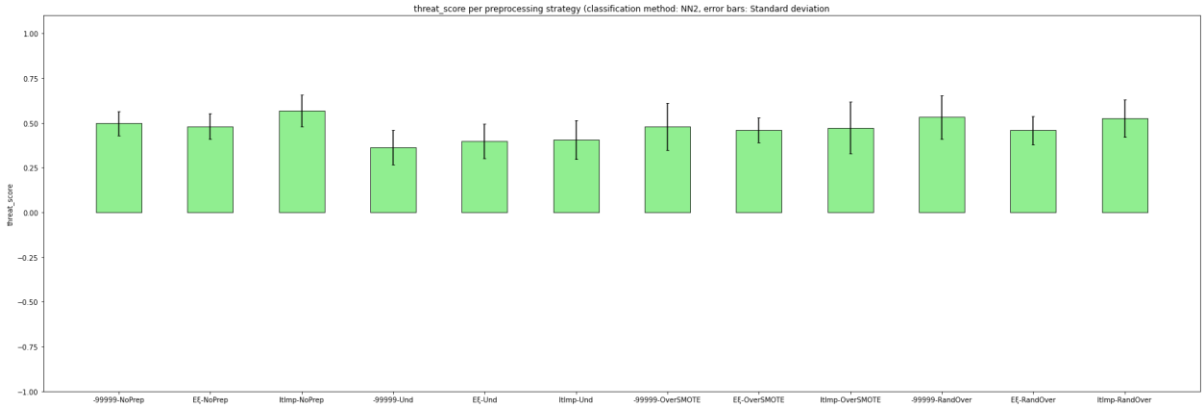

threat_score per preprocessing strategy (classification method: DT3, error bars: Standard deviation
threat_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

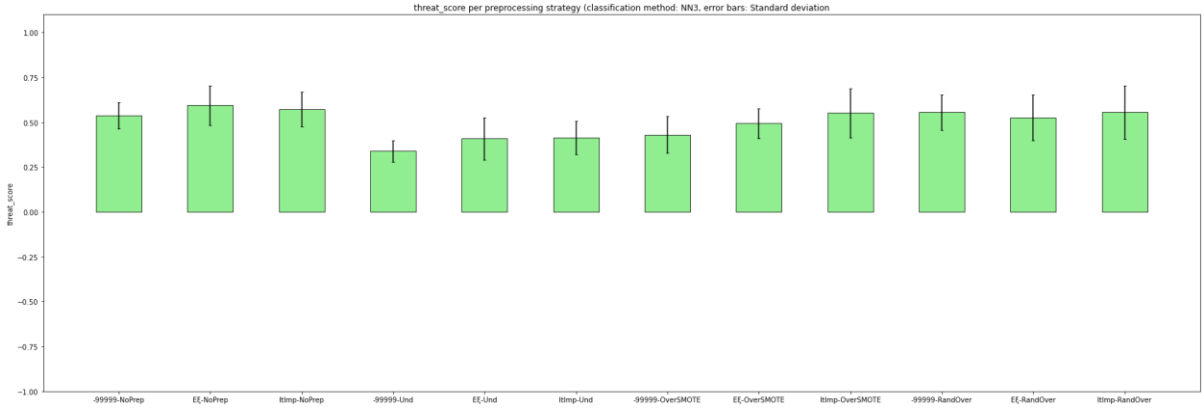

threat_score per preprocessing strategy (classification method: DT4, error bars: Standard deviation
threat_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

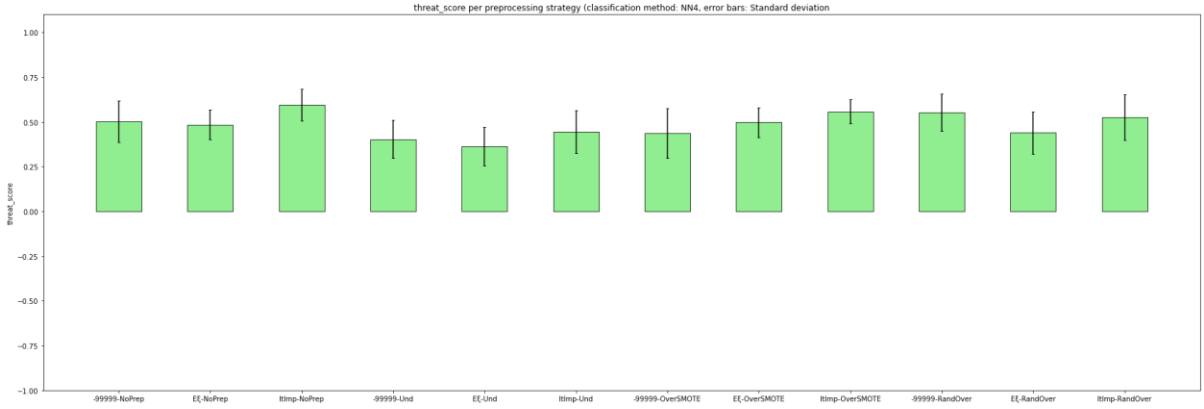

threat_score per preprocessing strategy (classification method: RF1, error bars: Standard deviation
threat_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

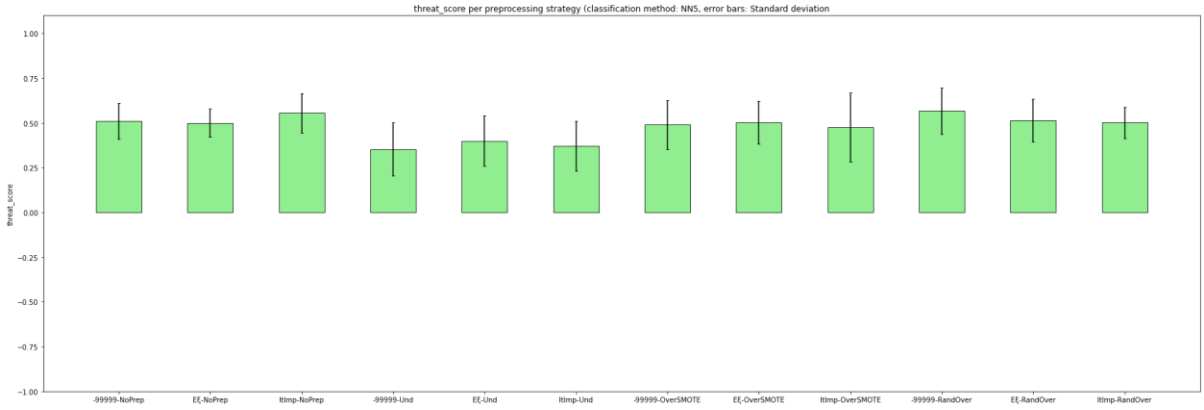

threat_score per preprocessing strategy (classification method: RF2, error bars: Standard deviation
threat_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

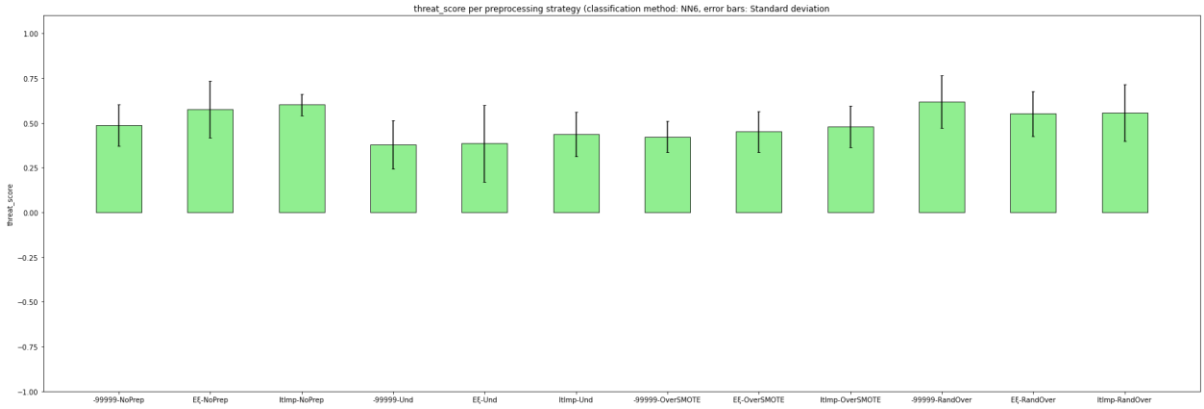

threat_score per preprocessing strategy (classification method: NN1, error bars: Standard deviation
threat_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

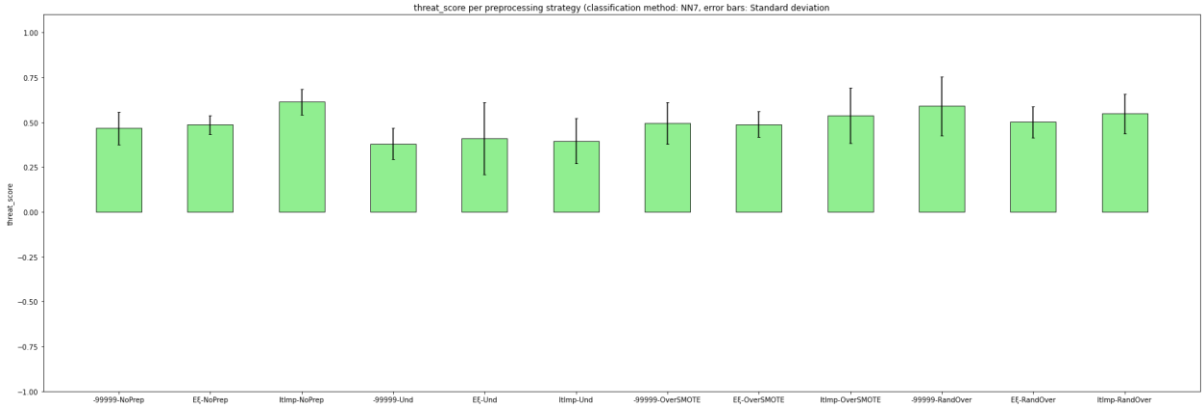

threat_score per preprocessing strategy (classification method: NN2, error bars: Standard deviation
threat_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

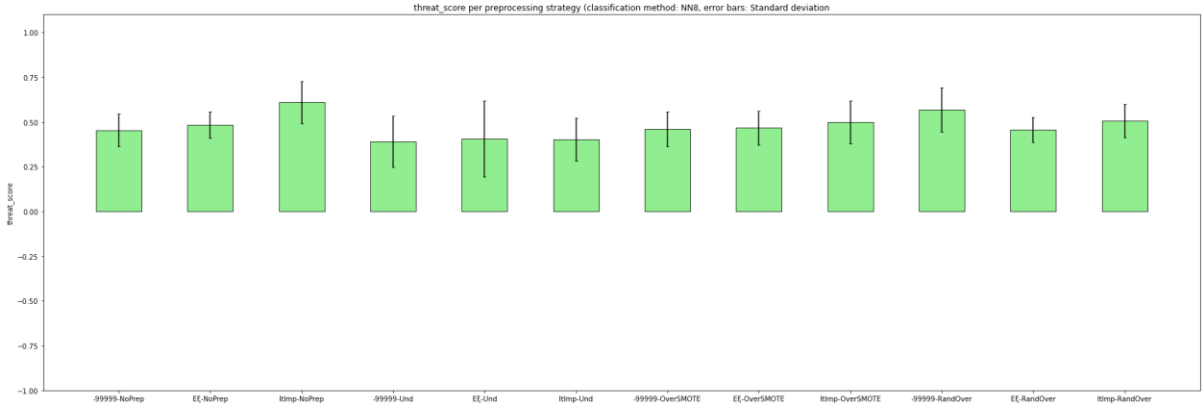

threat_score per preprocessing strategy (classification method: NN3, error bars: Standard deviation
threat_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

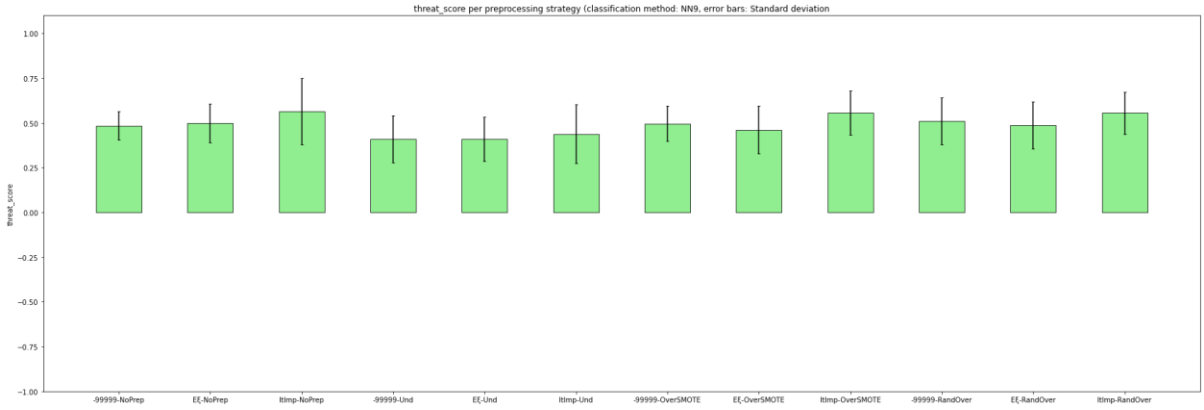

threat_score per preprocessing strategy (classification method: NN4, error bars: Standard deviation
threat_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

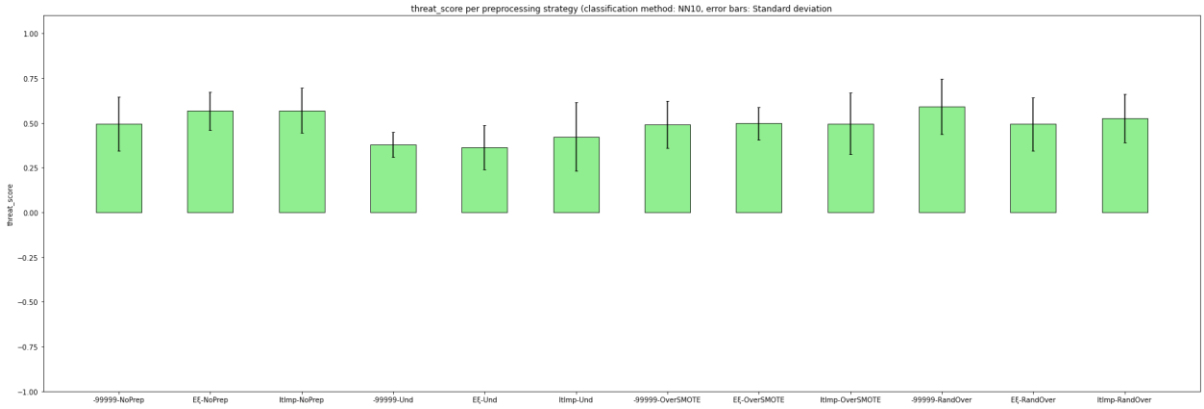

threat_score per preprocessing strategy (classification method: NN5, error bars: Standard deviation
threat_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

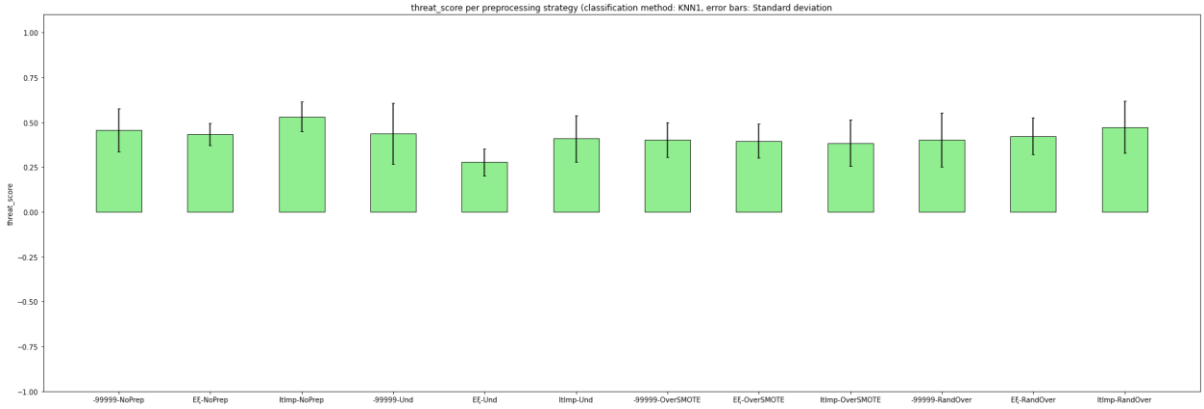

threat_score per preprocessing strategy (classification method: NN6, error bars: Standard deviation
threat_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

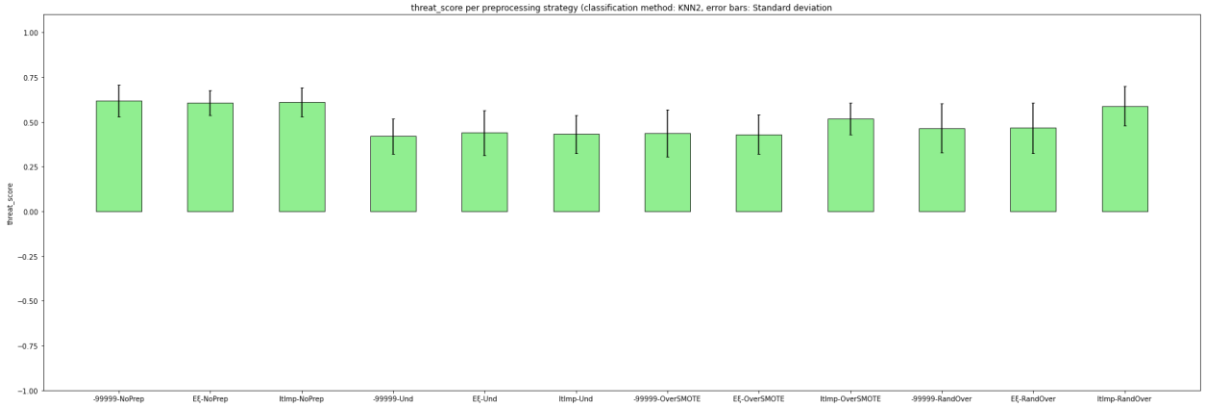

threat_score per preprocessing strategy (classification method: NN7, error bars: Standard deviation
threat_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

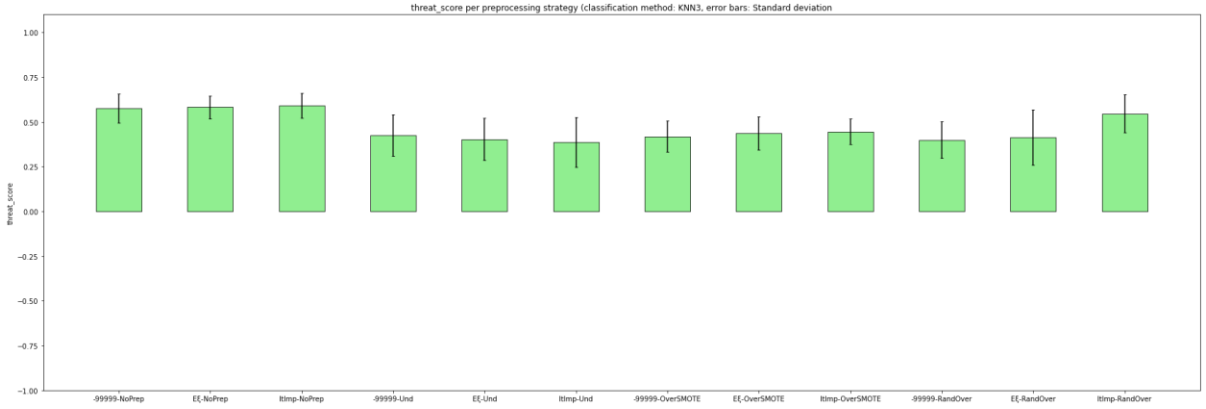

threat_score per preprocessing strategy (classification method: NN8, error bars: Standard deviation
threat_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

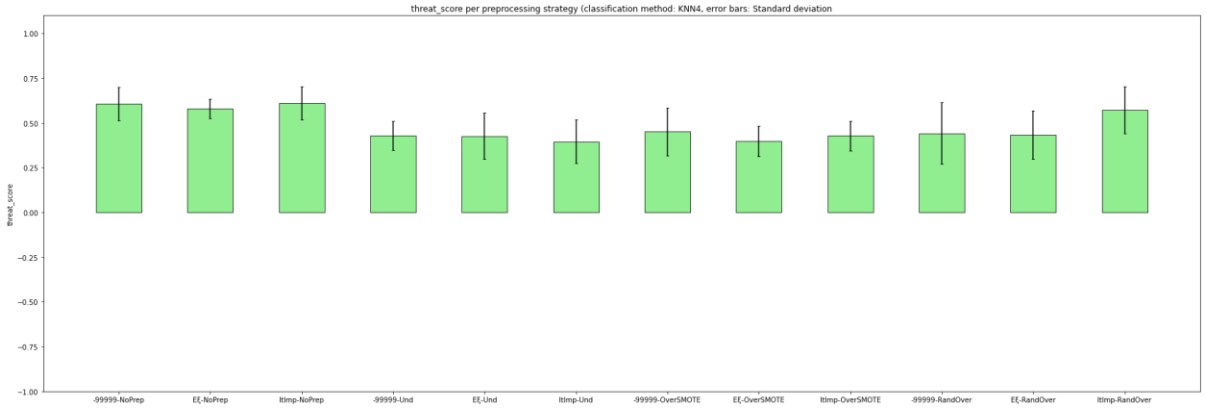

threat_score per preprocessing strategy (classification method: NN9, error bars: Standard deviation
threat_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

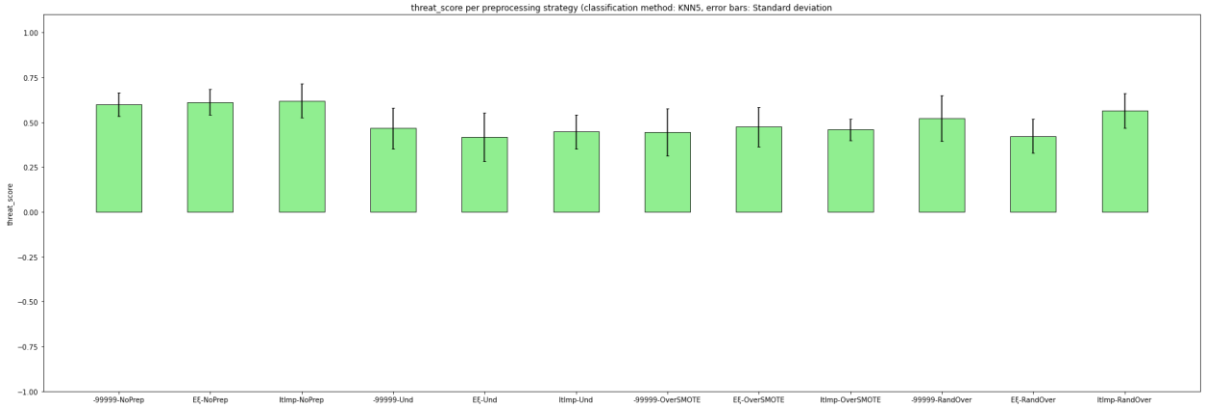

threat_score per preprocessing strategy (classification method: NN10, error bars: Standard deviation
threat_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

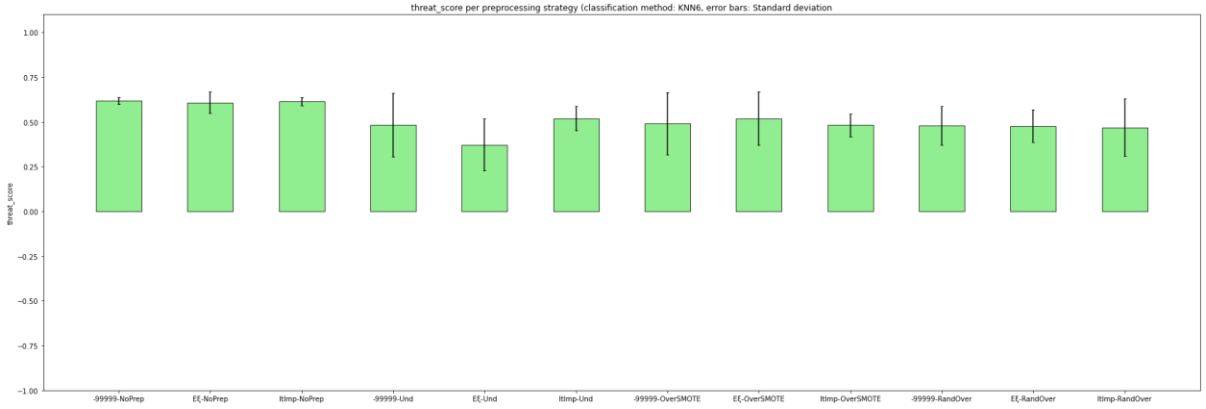

threat_score per preprocessing strategy (classification method: KNN1, error bars: Standard deviation
threat_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

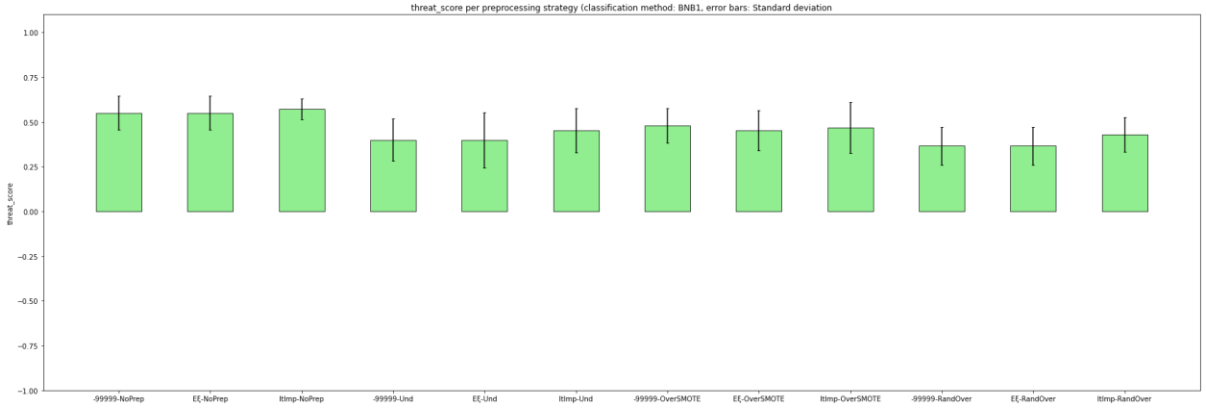

threat_score per preprocessing strategy (classification method: KNN2, error bars: Standard deviation
threat_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

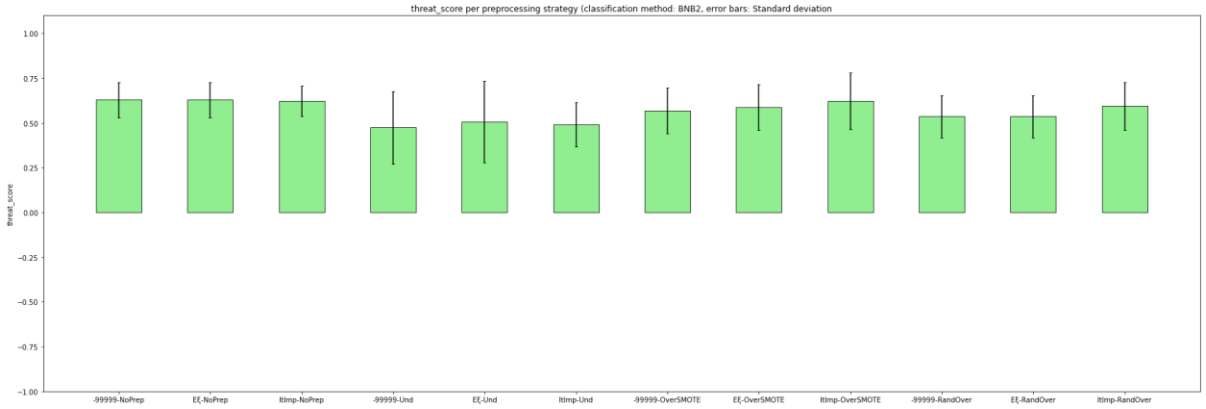

threat_score per preprocessing strategy (classification method: KNN3, error bars: Standard deviation
threat_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

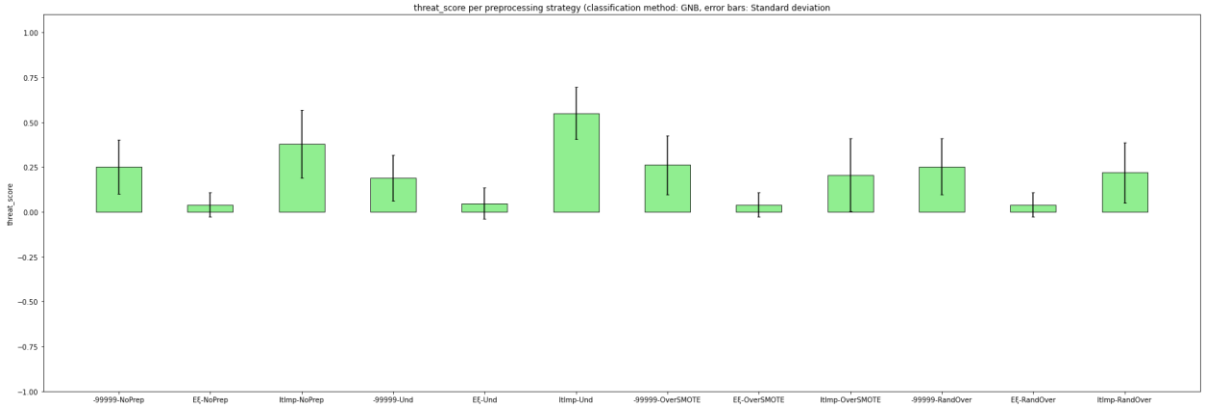

threat_score per preprocessing strategy (classification method: KNN4, error bars: Standard deviation
threat_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

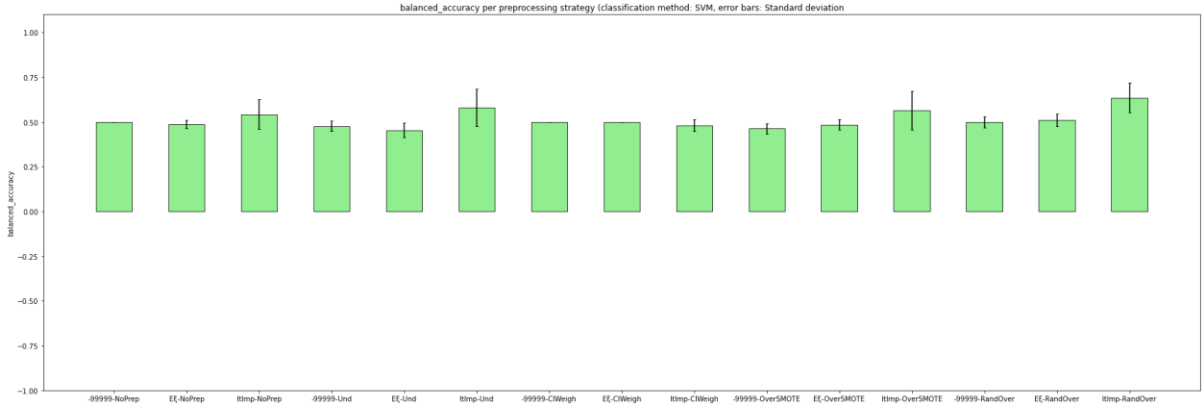

threat_score per preprocessing strategy (classification method: KNN5, error bars: Standard deviation
threat_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

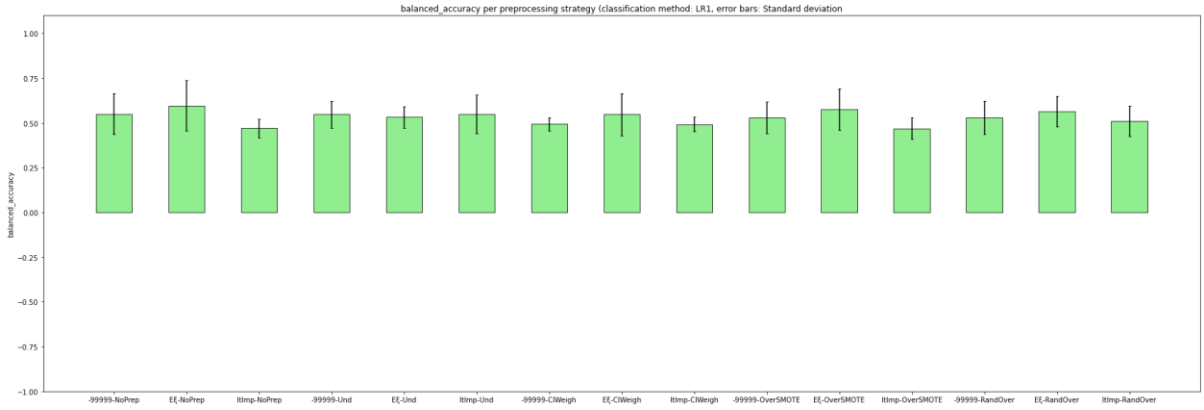

threat_score per preprocessing strategy (classification method: KNN6, error bars: Standard deviation
threat_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

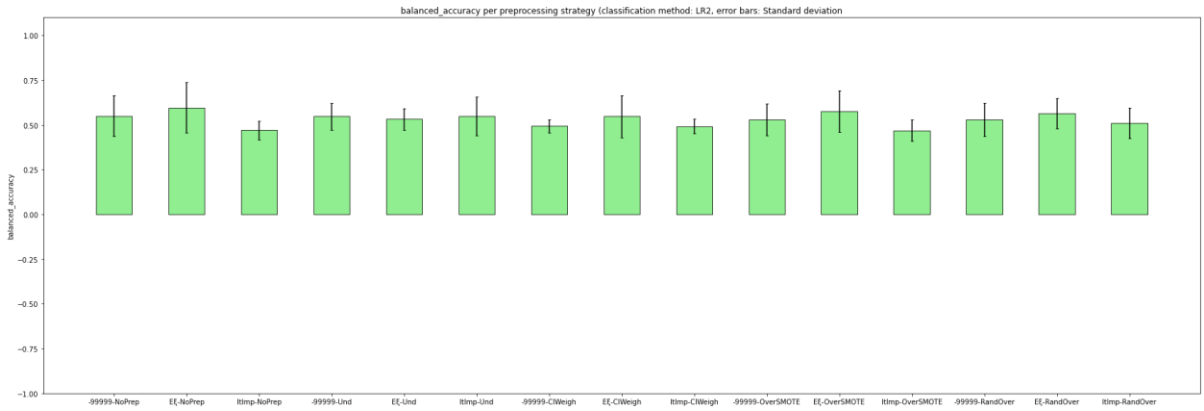

threat_score per preprocessing strategy (classification method: BNB1, error bars: Standard deviation
threat_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

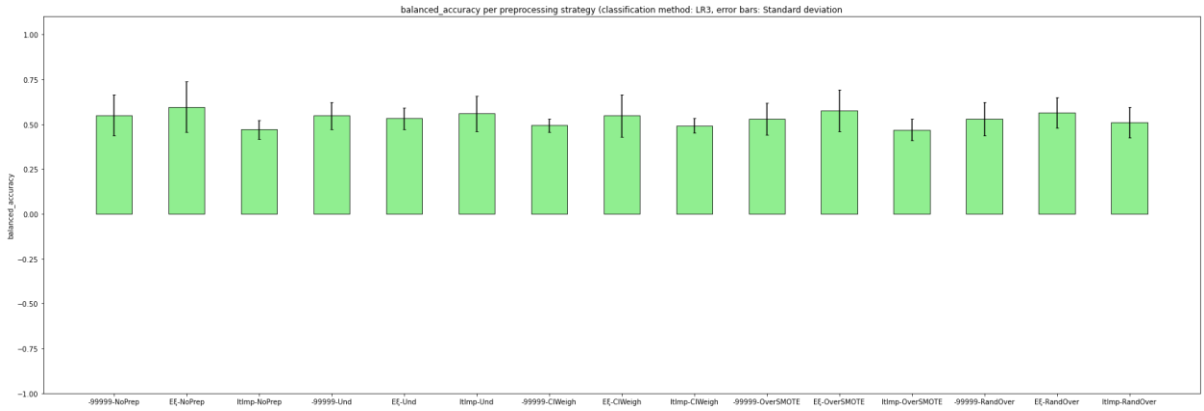

threat_score per preprocessing strategy (classification method: BNB2, error bars: Standard deviation
threat_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

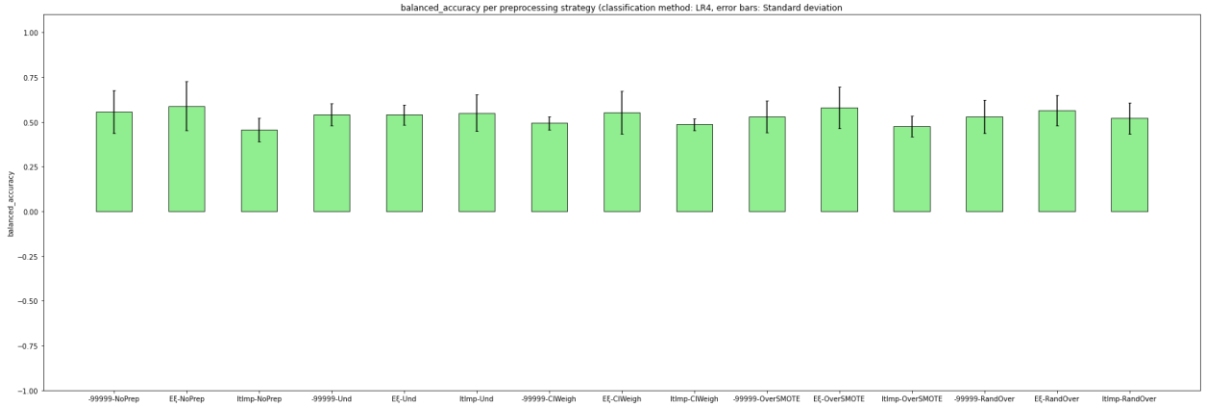

threat_score per preprocessing strategy (classification method: GNB, error bars: Standard deviation
threat_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

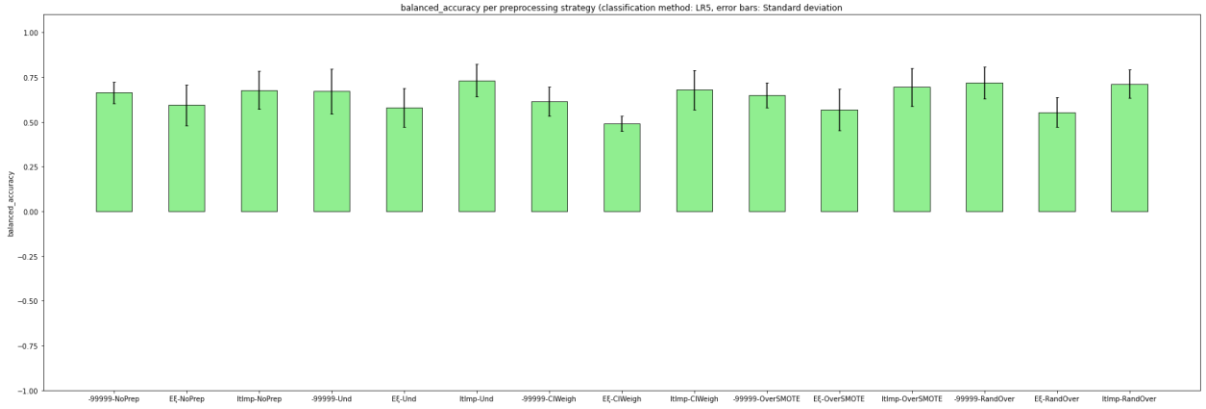

balanced_accuracy per preprocessing strategy (classification method: SVM, error bars: Standard deviation
balanced_accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

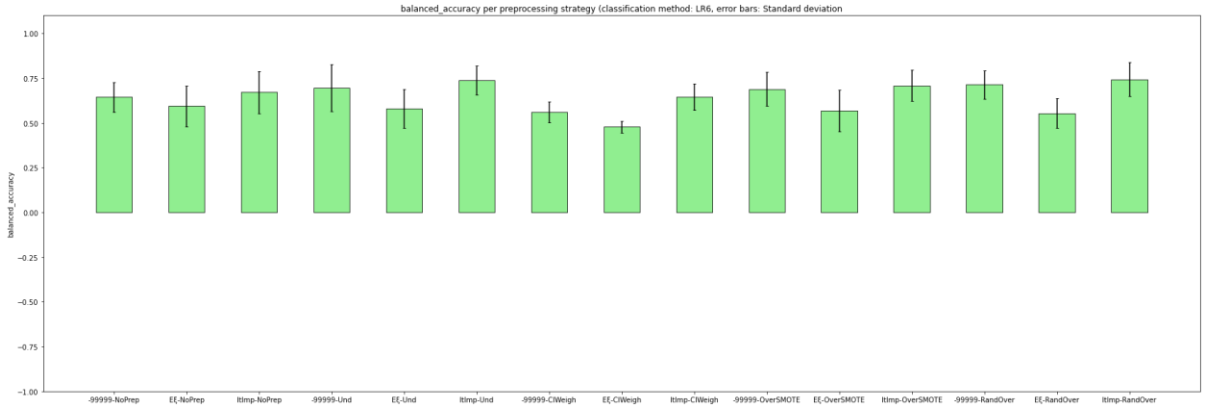

balanced_accuracy per preprocessing strategy (classification method: LR1, error bars: Standard deviation
balanced_accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

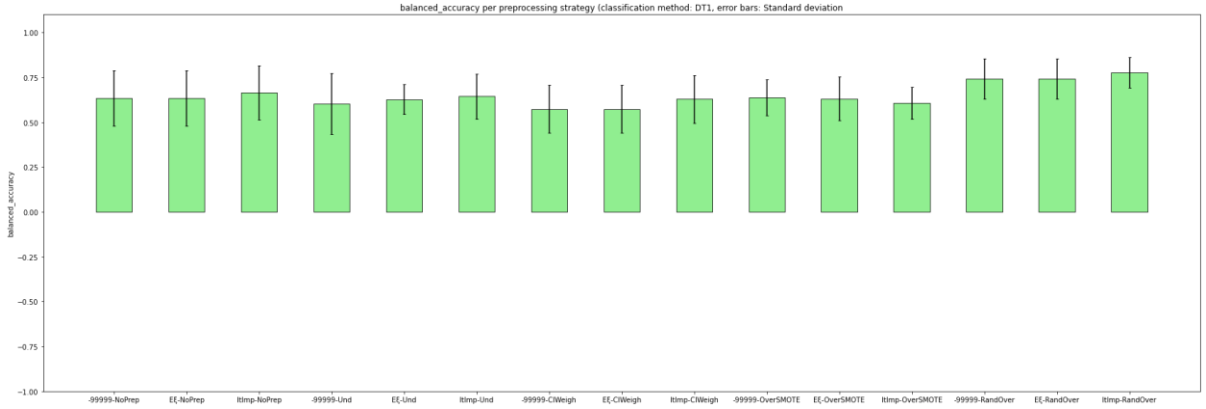

balanced_accuracy per preprocessing strategy (classification method: LR2, error bars: Standard deviation
balanced_accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

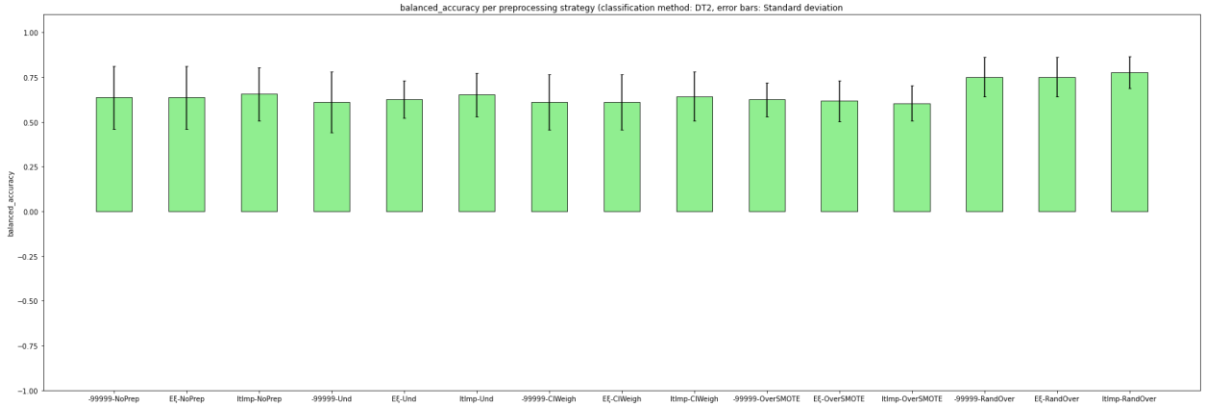

balanced_accuracy per preprocessing strategy (classification method: LR3, error bars: Standard deviation
balanced_accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

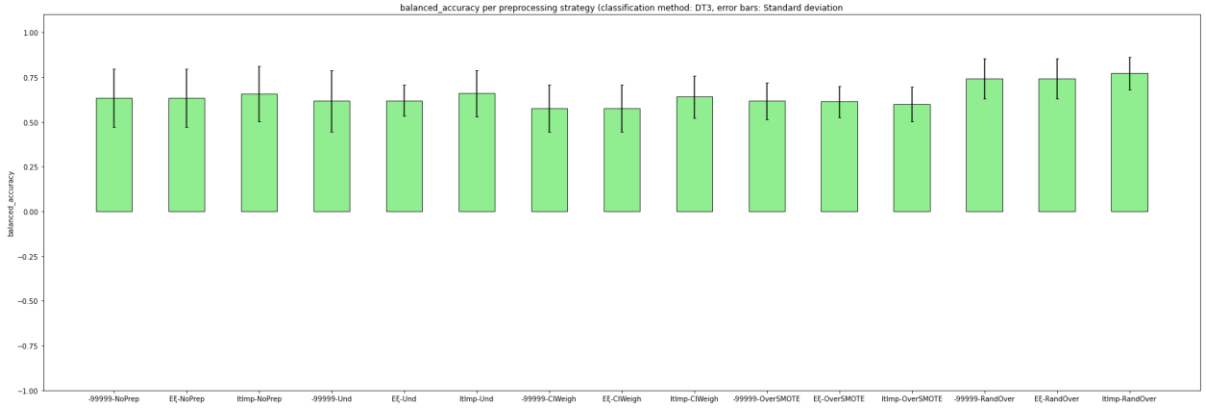

balanced_accuracy per preprocessing strategy (classification method: LR4, error bars: Standard deviation
balanced_accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

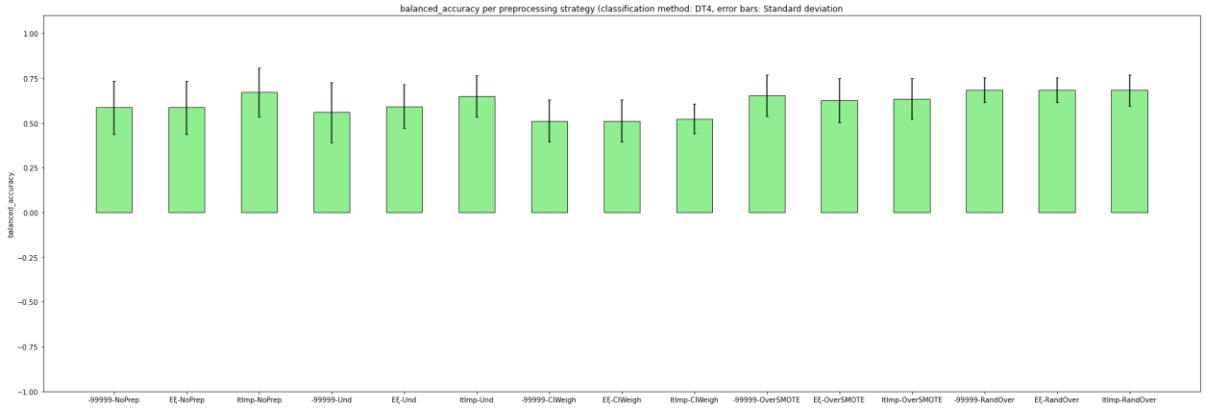

balanced_accuracy per preprocessing strategy (classification method: LR5, error bars: Standard deviation
balanced_accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

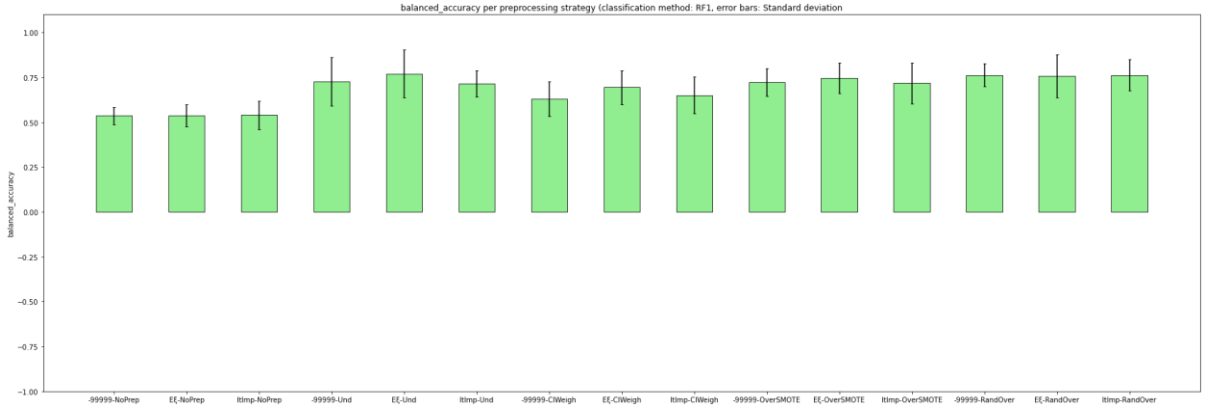

balanced_accuracy per preprocessing strategy (classification method: LR6, error bars: Standard deviation
balanced_accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

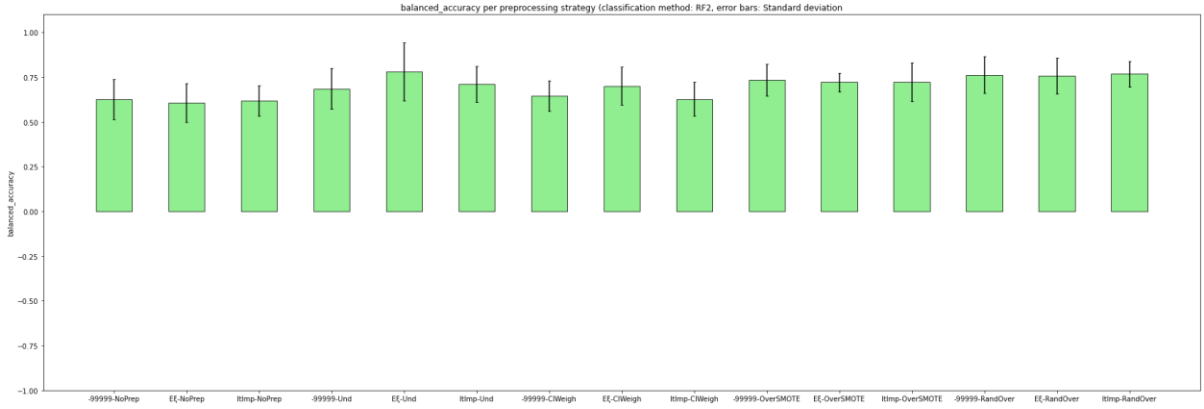

balanced_accuracy per preprocessing strategy (classification method: DT1, error bars: Standard deviation
balanced_accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

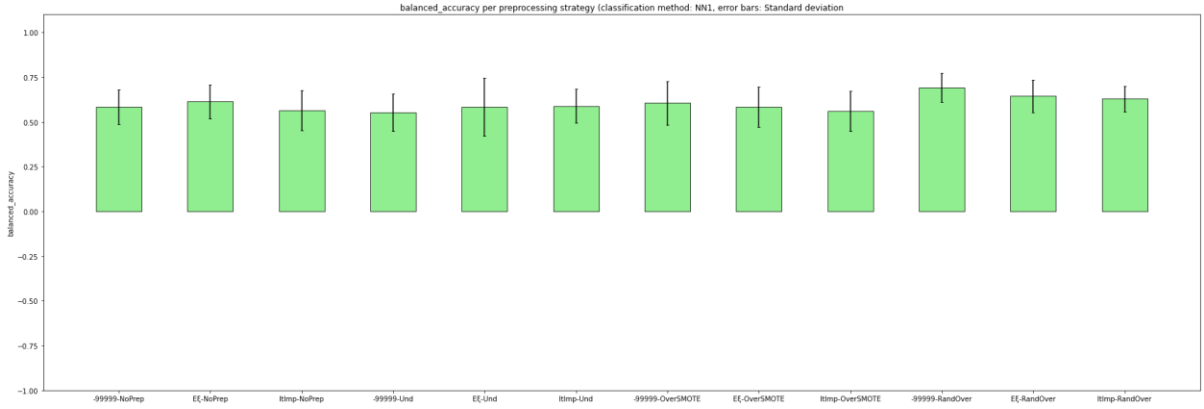

balanced_accuracy per preprocessing strategy (classification method: DT2, error bars: Standard deviation
balanced_accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

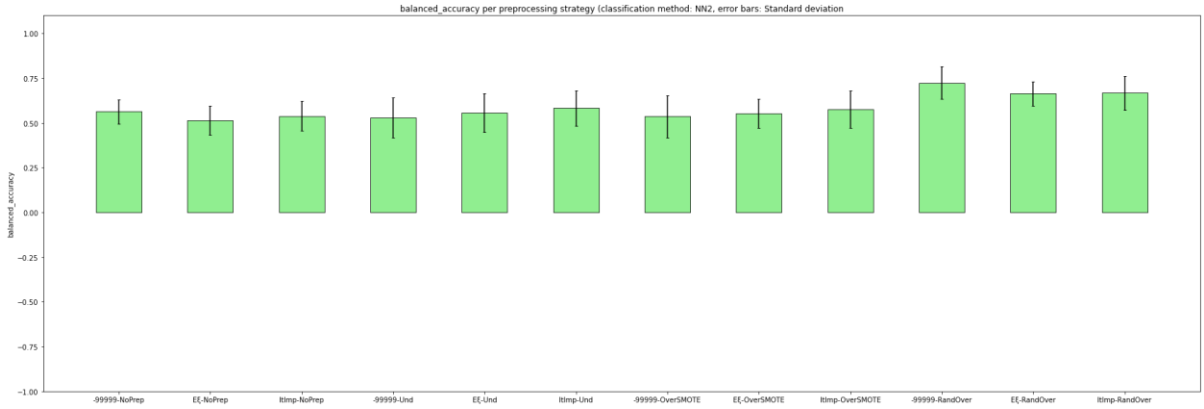

balanced_accuracy per preprocessing strategy (classification method: DT3, error bars: Standard deviation
balanced_accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

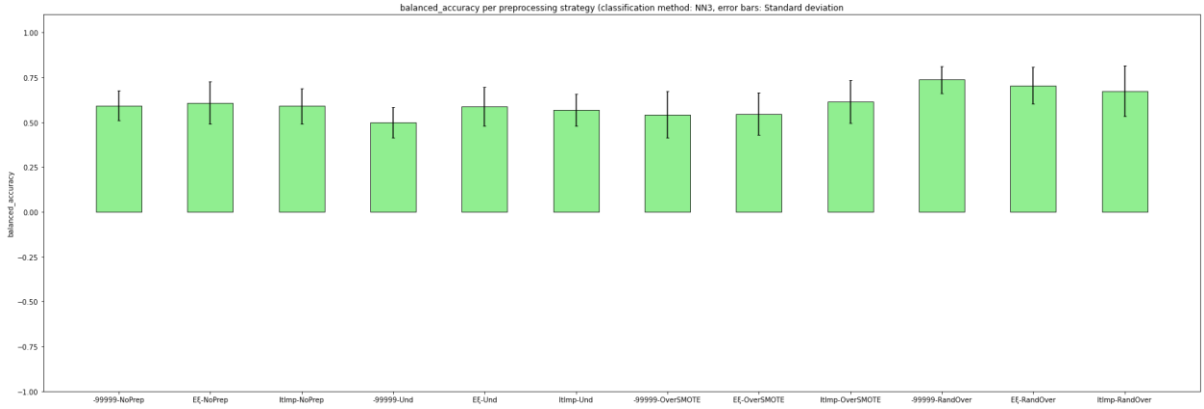

balanced_accuracy per preprocessing strategy (classification method: DT4, error bars: Standard deviation
balanced_accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

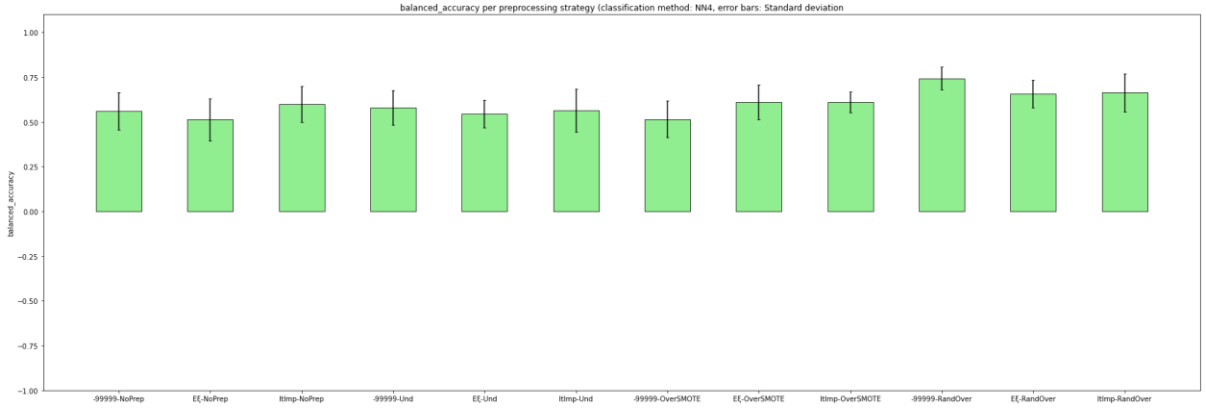

balanced_accuracy per preprocessing strategy (classification method: RF1, error bars: Standard deviation
balanced_accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

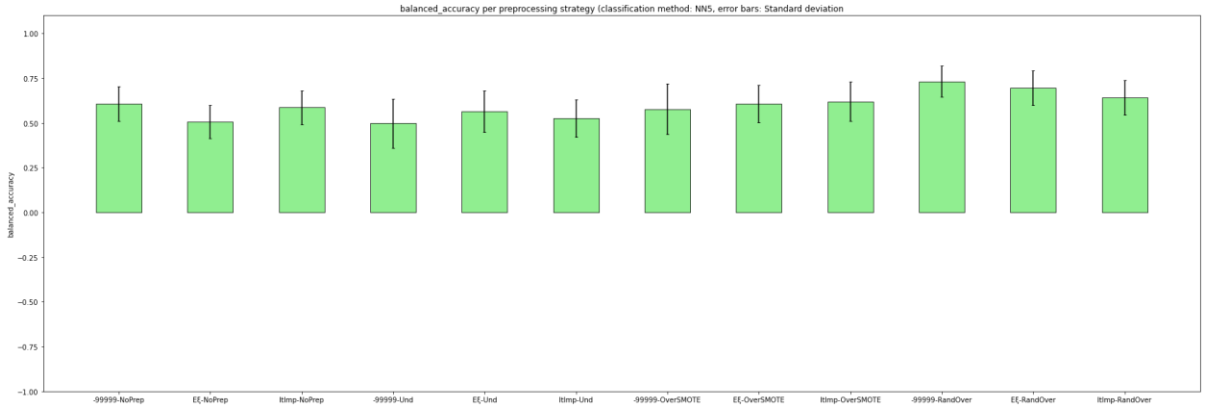

balanced_accuracy per preprocessing strategy (classification method: RF2, error bars: Standard deviation
balanced_accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

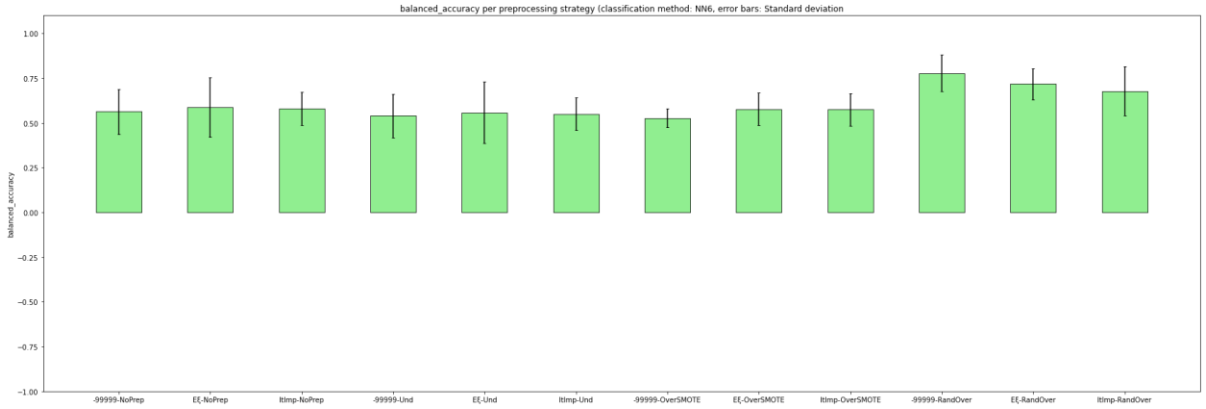

balanced_accuracy per preprocessing strategy (classification method: NN1, error bars: Standard deviation
balanced_accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

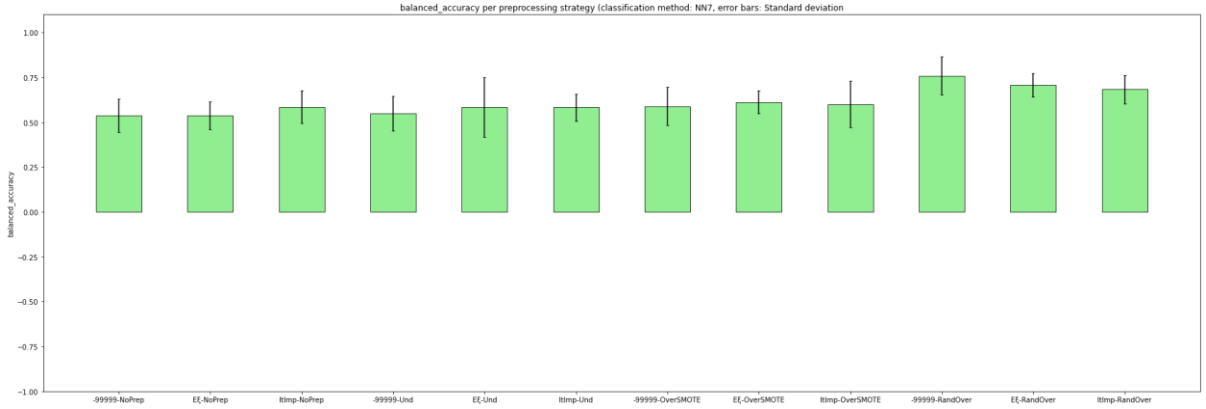

balanced_accuracy per preprocessing strategy (classification method: NN2, error bars: Standard deviation
balanced_accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

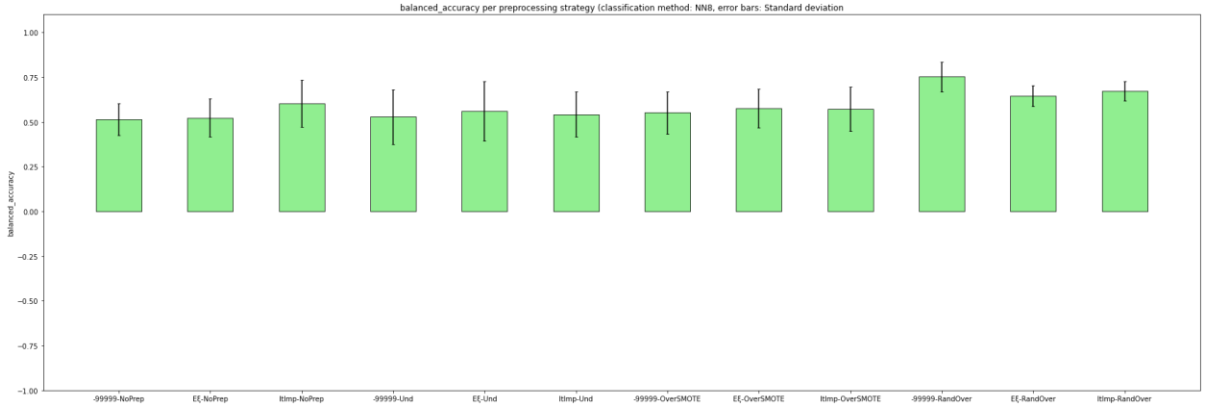

balanced_accuracy per preprocessing strategy (classification method: NN3, error bars: Standard deviation
balanced_accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

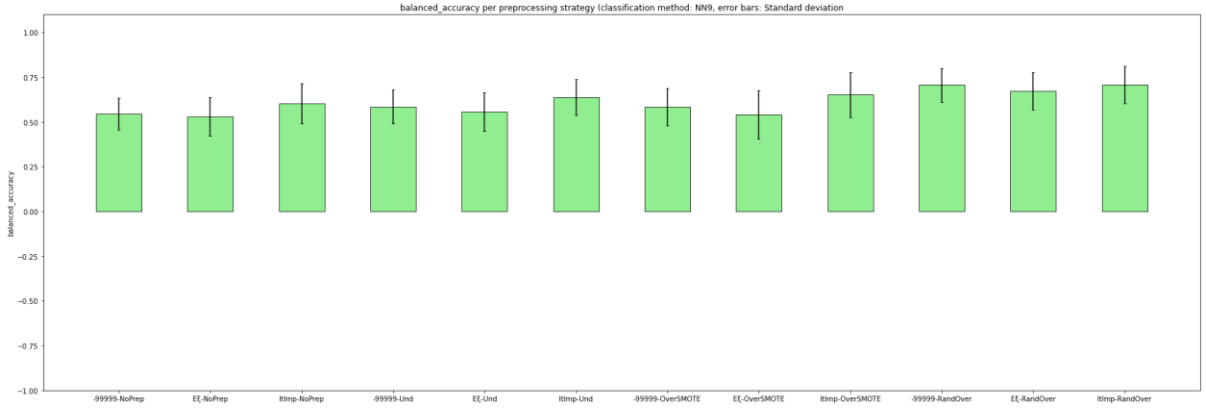

balanced_accuracy per preprocessing strategy (classification method: NN4, error bars: Standard deviation
balanced_accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

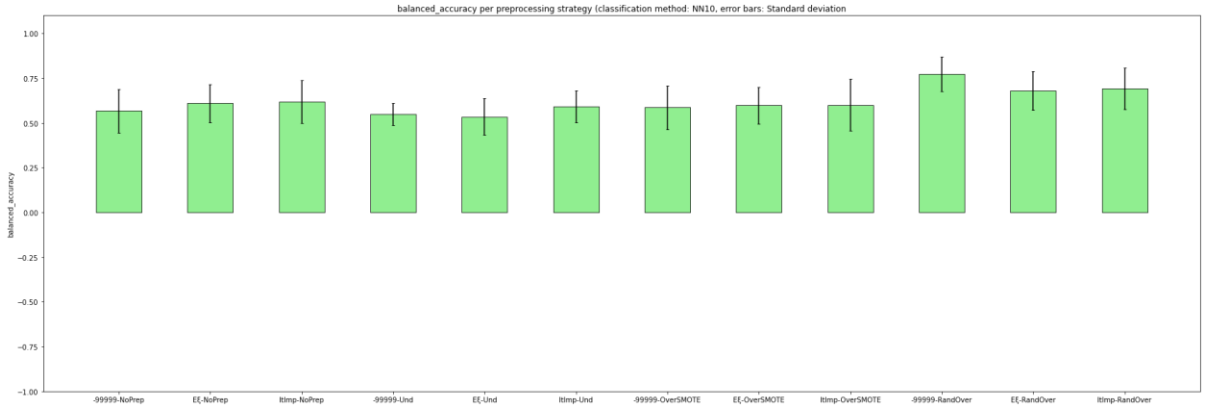

balanced_accuracy per preprocessing strategy (classification method: NN5, error bars: Standard deviation
balanced_accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

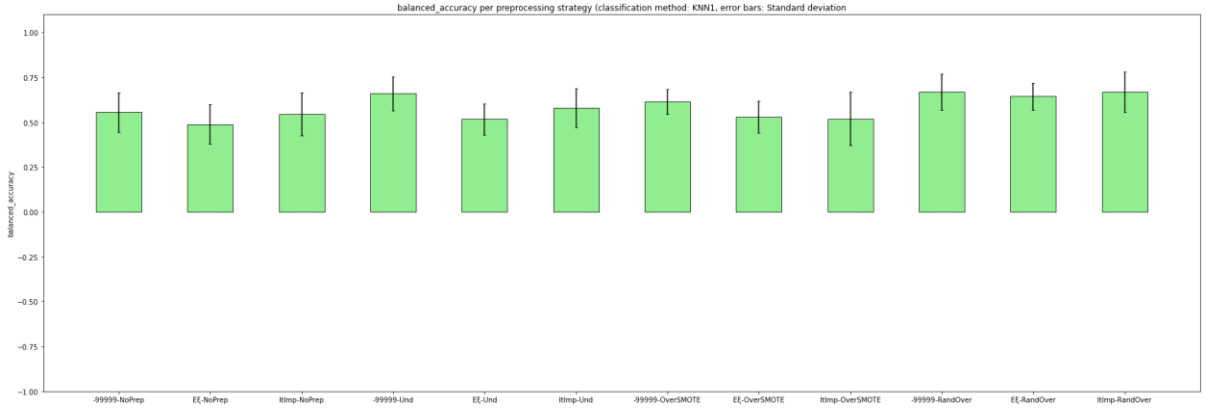

balanced_accuracy per preprocessing strategy (classification method: NN6, error bars: Standard deviation
balanced_accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

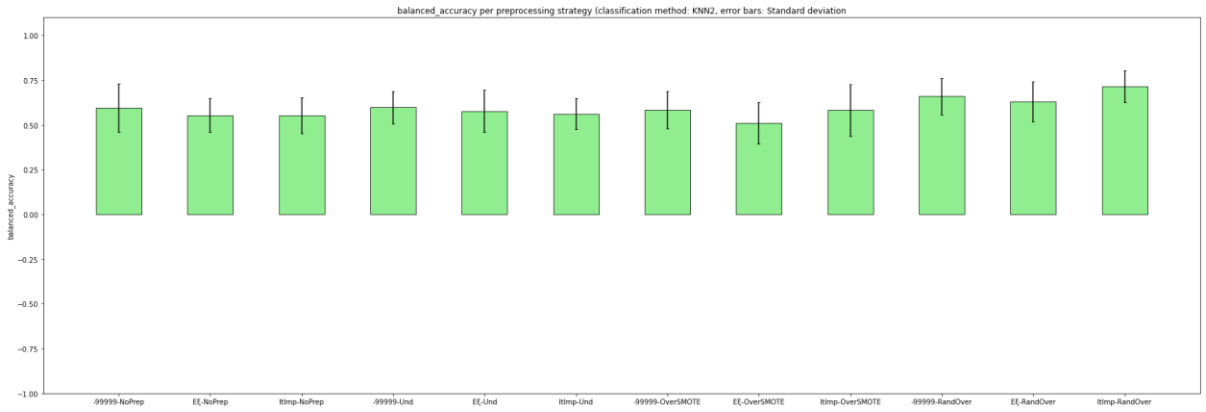

balanced_accuracy per preprocessing strategy (classification method: NN7, error bars: Standard deviation
balanced_accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

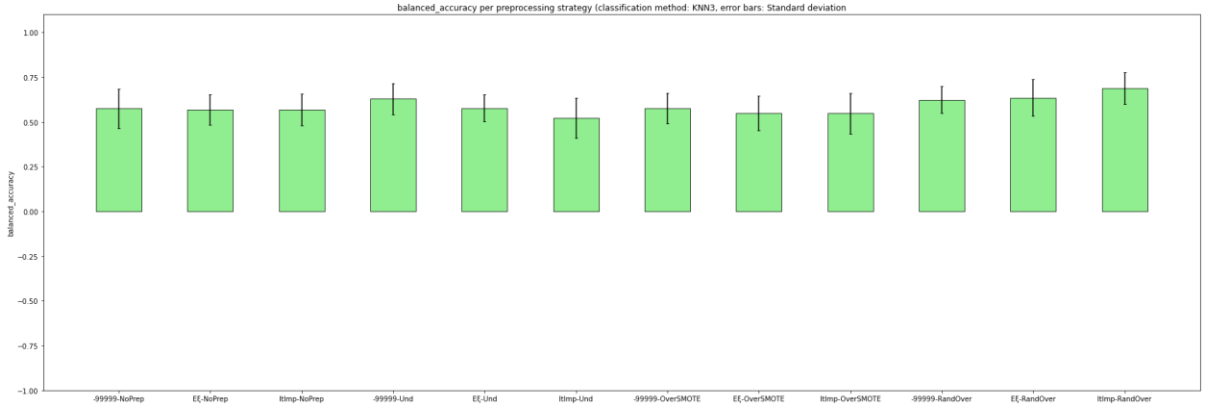

balanced_accuracy per preprocessing strategy (classification method: NN8, error bars: Standard deviation
balanced_accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

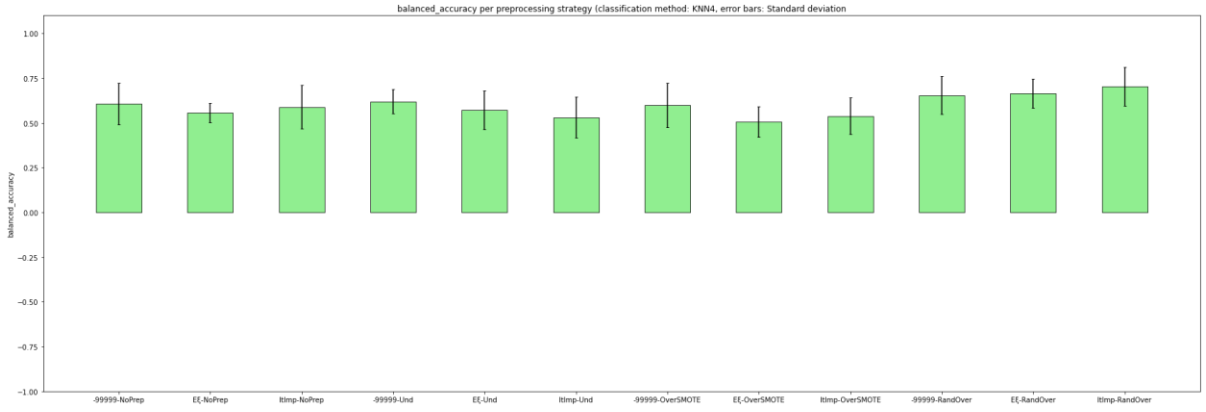

balanced_accuracy per preprocessing strategy (classification method: NN9, error bars: Standard deviation
balanced_accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

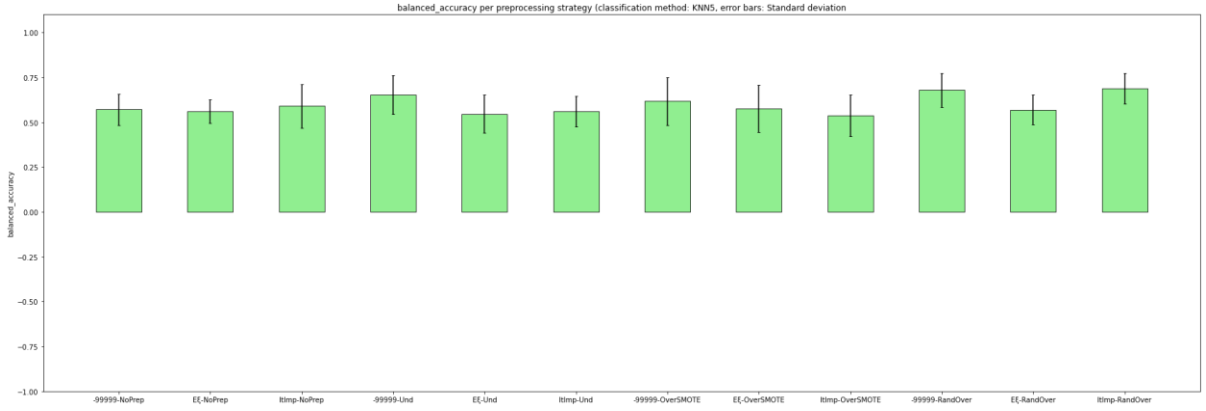

balanced_accuracy per preprocessing strategy (classification method: NN10, error bars: Standard deviation
balanced_accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

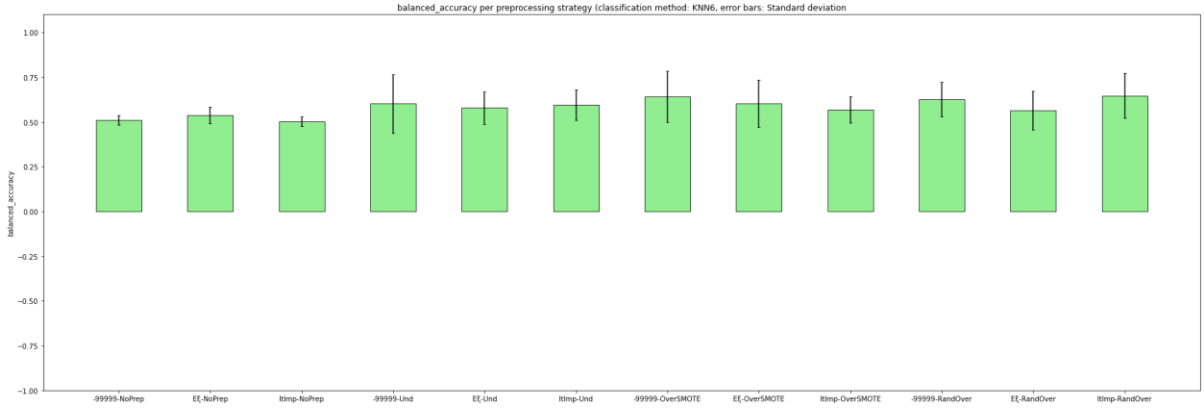

balanced_accuracy per preprocessing strategy (classification method: KNN1, error bars: Standard deviation
balanced_accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

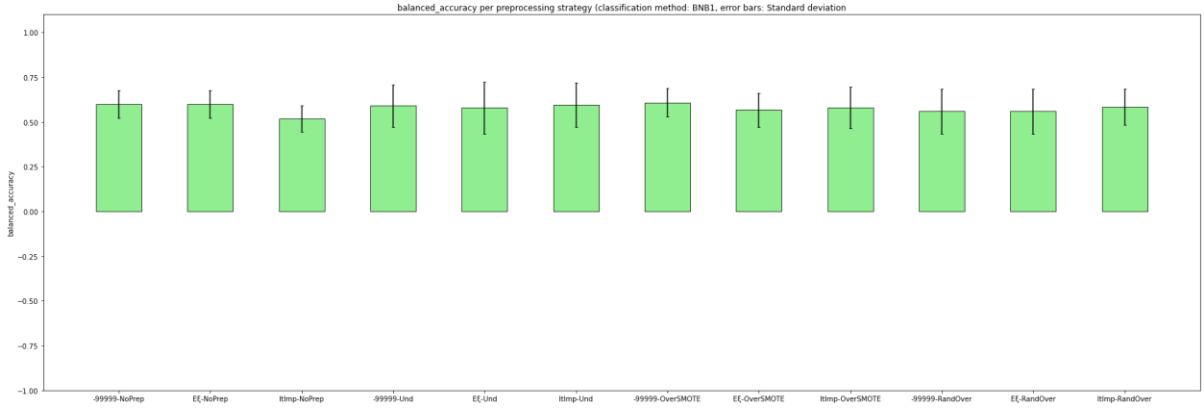

balanced_accuracy per preprocessing strategy (classification method: KNN2, error bars: Standard deviation
balanced_accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

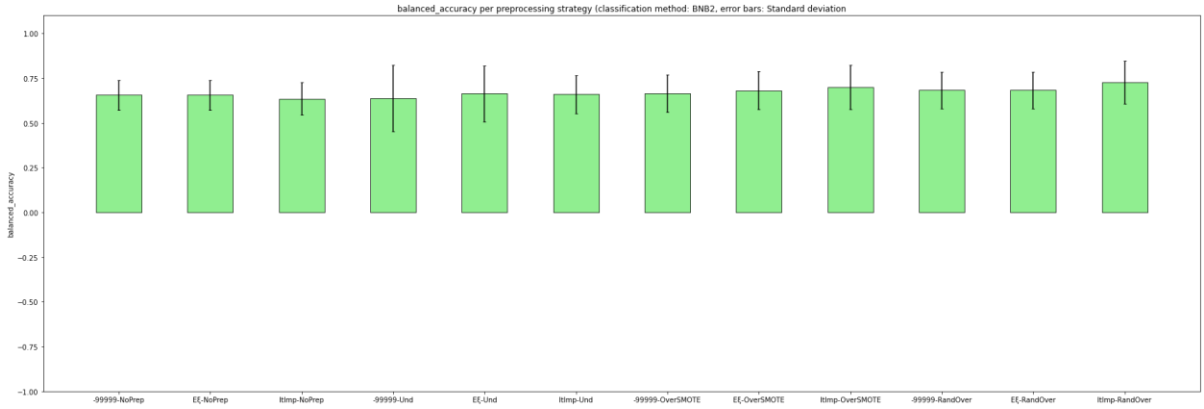

balanced_accuracy per preprocessing strategy (classification method: KNN3, error bars: Standard deviation
balanced_accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

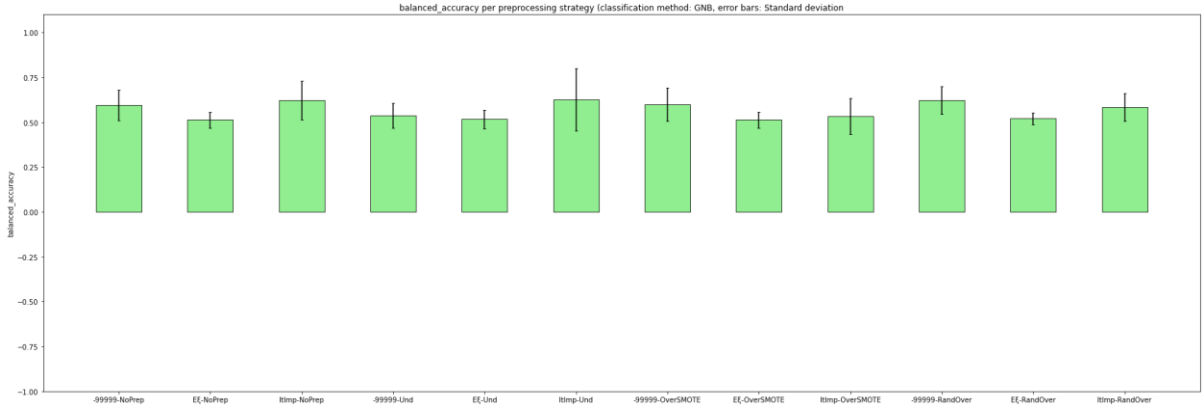

balanced_accuracy per preprocessing strategy (classification method: KNN4, error bars: Standard deviation
balanced_accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

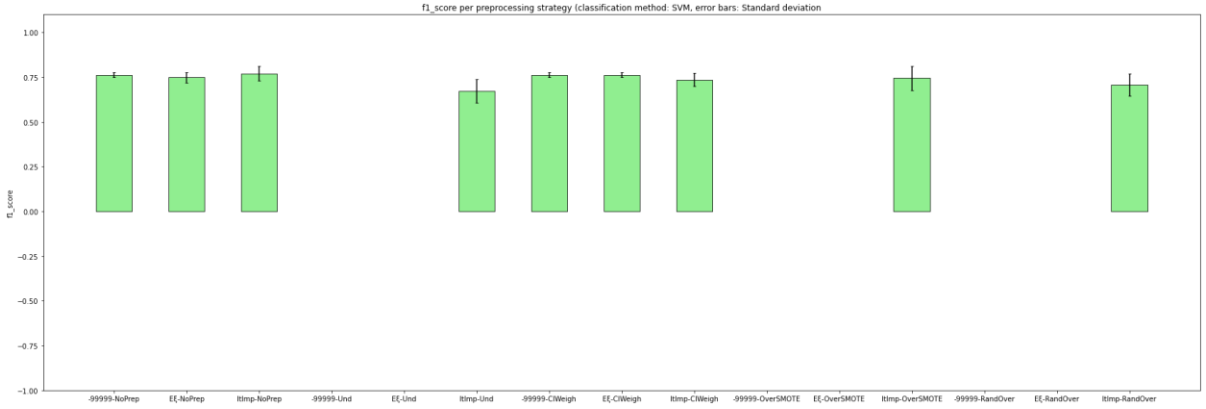

balanced_accuracy per preprocessing strategy (classification method: KNN5, error bars: Standard deviation
balanced_accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

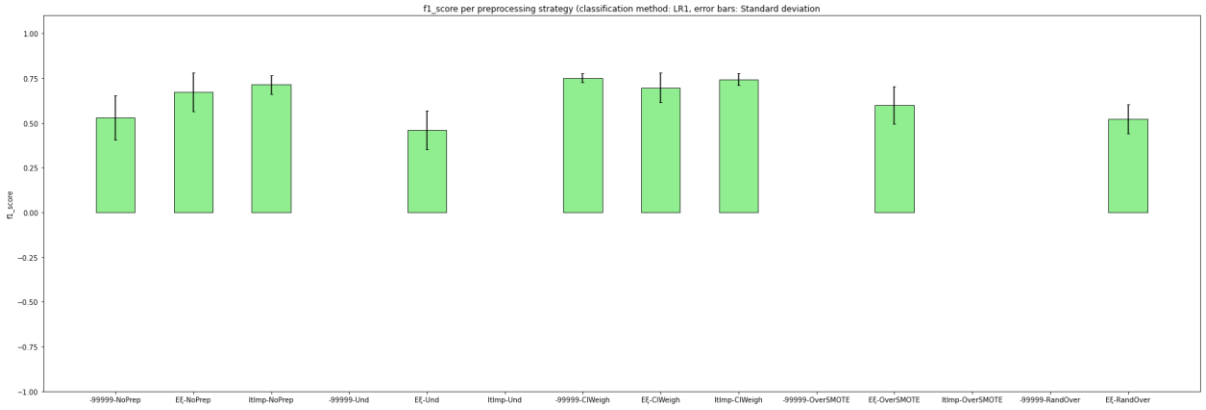

balanced_accuracy per preprocessing strategy (classification method: KNN6, error bars: Standard deviation
balanced_accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

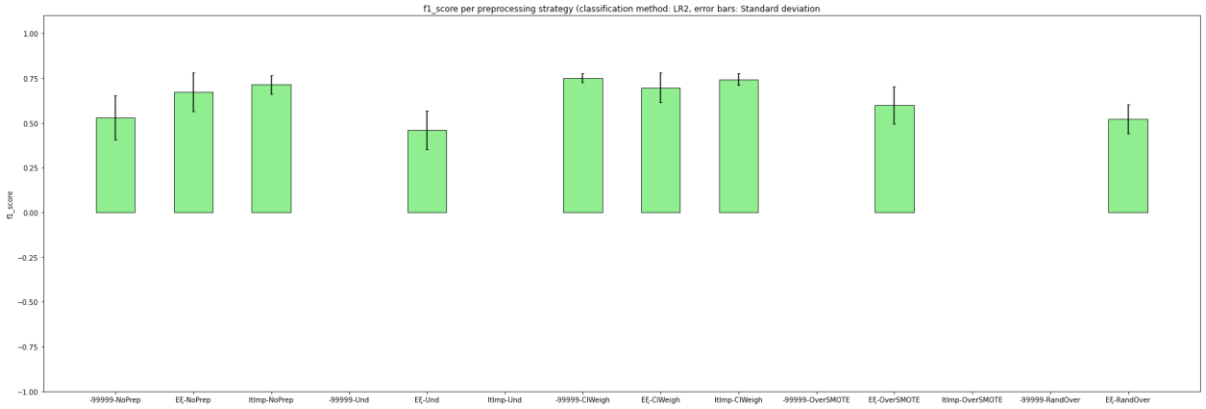

balanced_accuracy per preprocessing strategy (classification method: BNB1, error bars: Standard deviation
balanced_accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

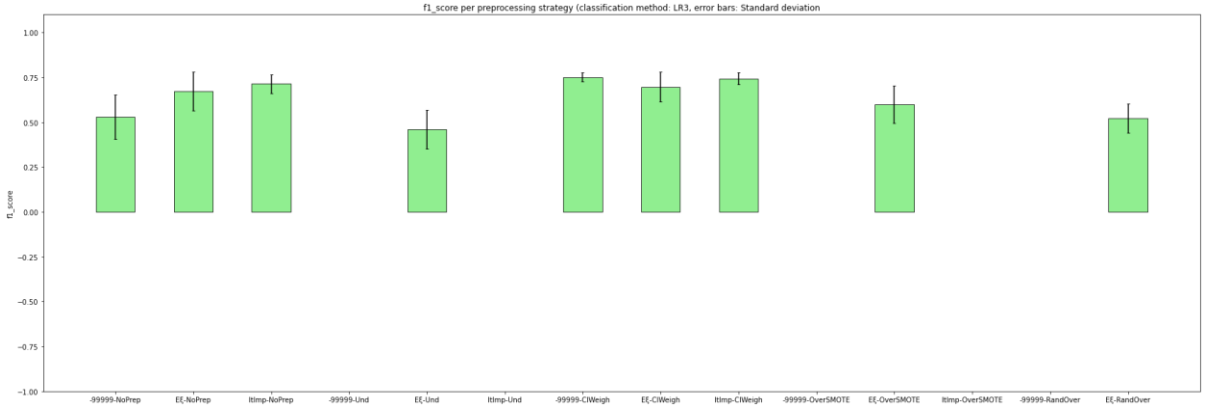

balanced_accuracy per preprocessing strategy (classification method: BNB2, error bars: Standard deviation
balanced_accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

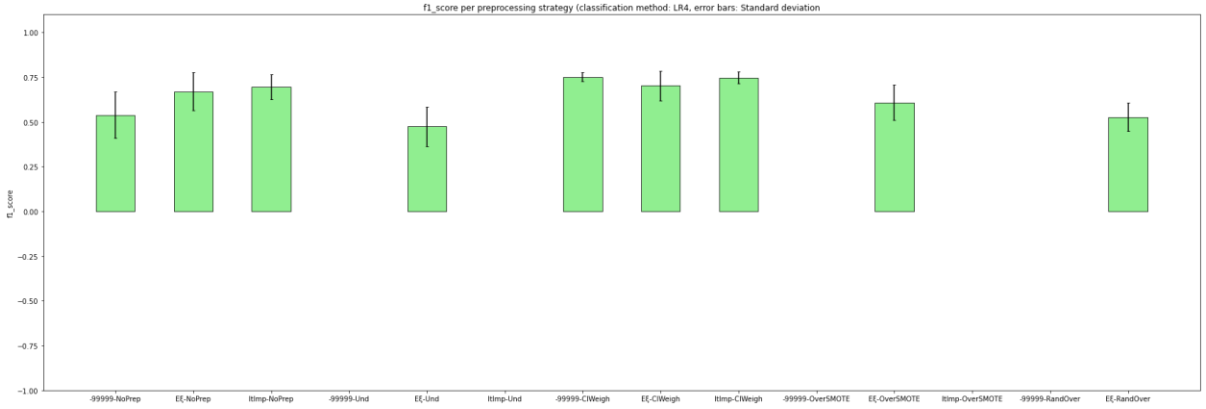

balanced_accuracy per preprocessing strategy (classification method: GNB, error bars: Standard deviation
balanced_accuracy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

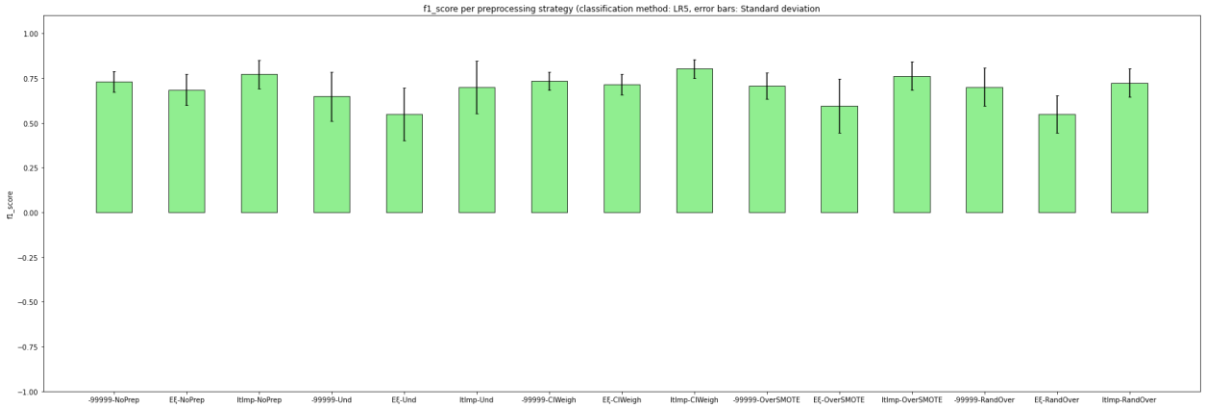

f1_score per preprocessing strategy (classification method: SVM, error bars: Standard deviation
f1_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

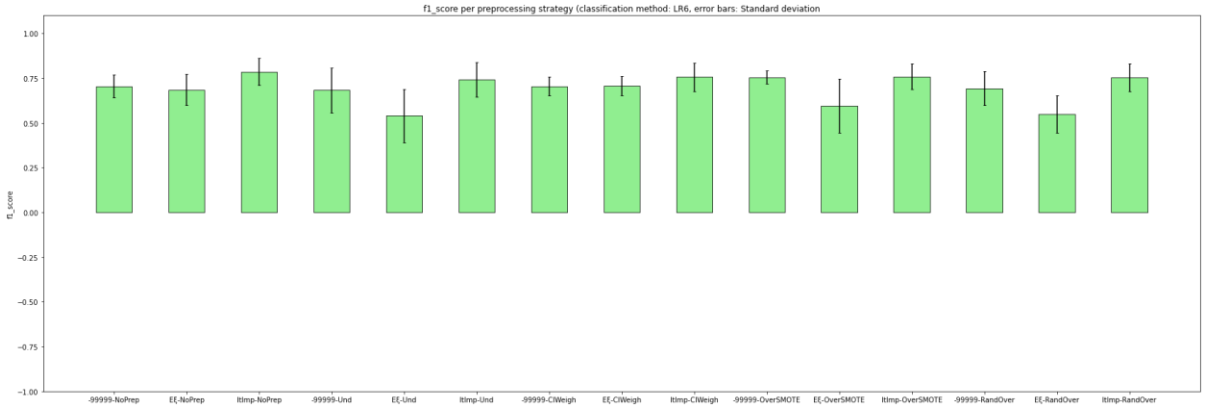

f1_score per preprocessing strategy (classification method: LR1, error bars: Standard deviation
f1_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver

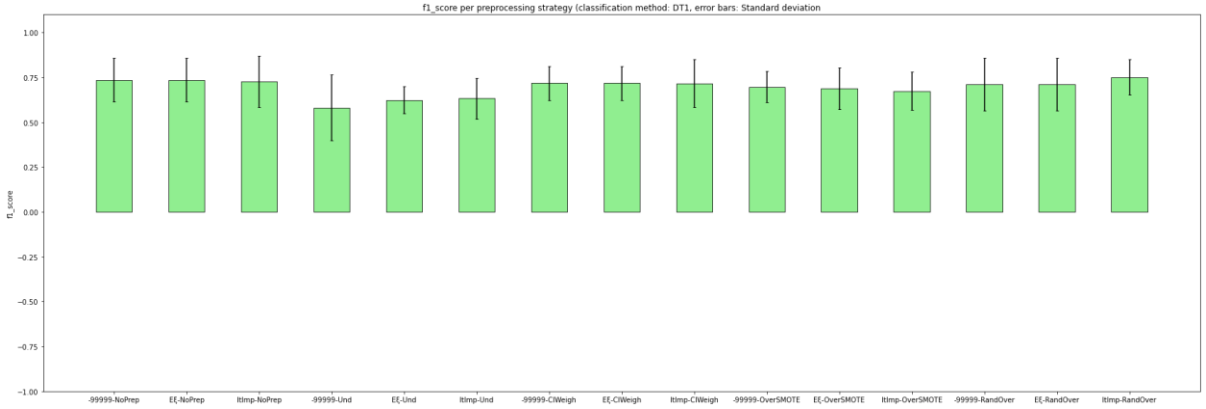

f1_score per preprocessing strategy (classification method: LR2, error bars: Standard deviation
f1_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver

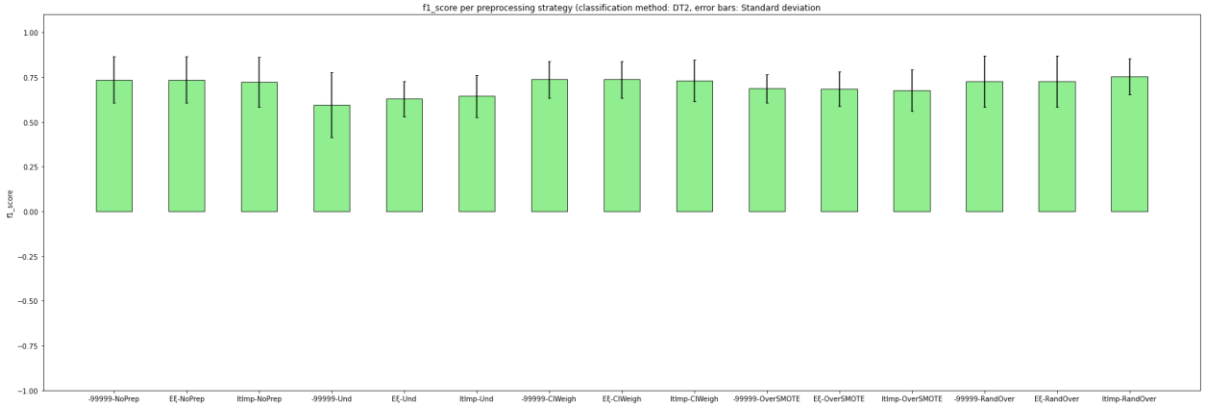

f1_score per preprocessing strategy (classification method: LR3, error bars: Standard deviation
f1_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver

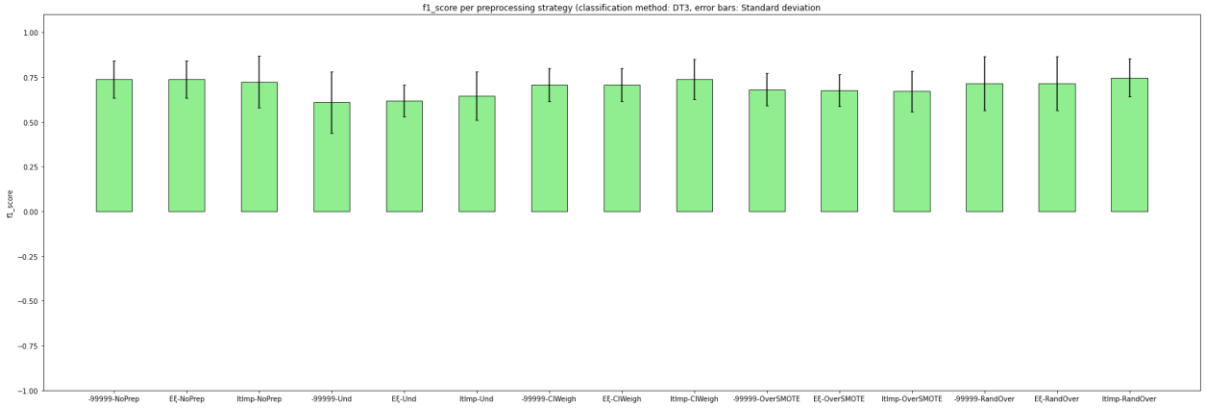

f1_score per preprocessing strategy (classification method: LR4, error bars: Standard deviation
f1_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver

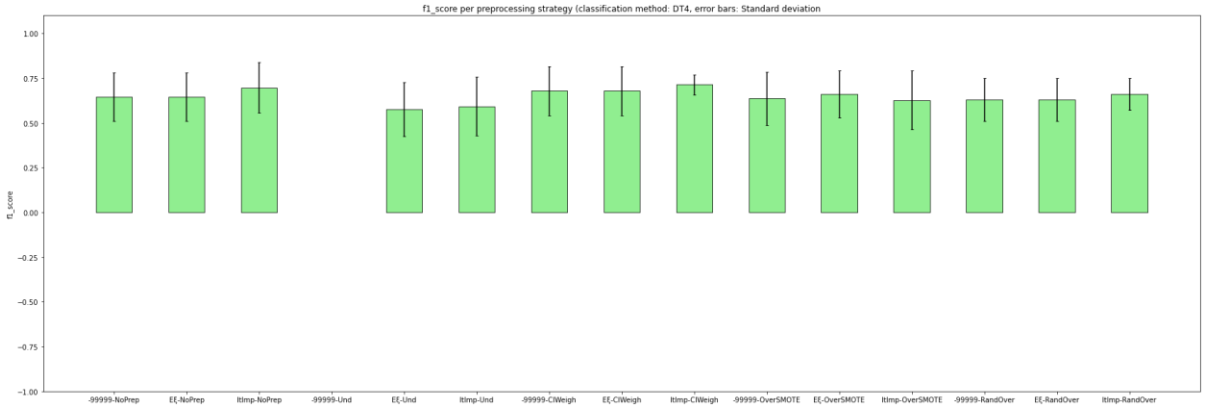

f1_score per preprocessing strategy (classification method: LR5, error bars: Standard deviation
f1_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

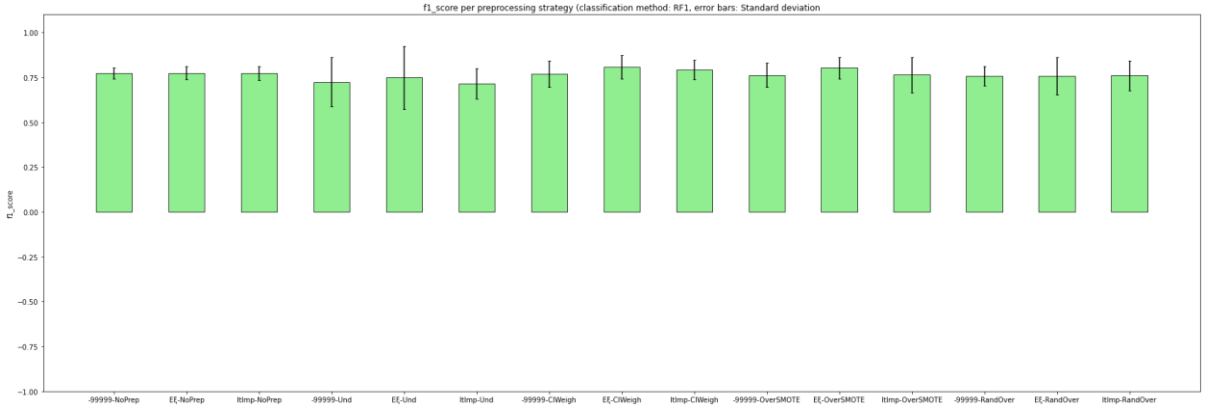

f1_score per preprocessing strategy (classification method: LR6, error bars: Standard deviation
f1_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

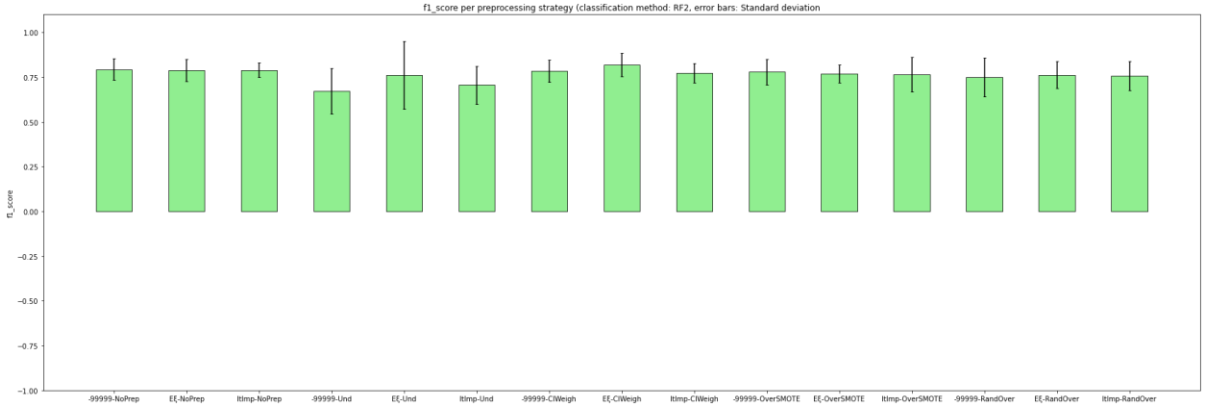

f1_score per preprocessing strategy (classification method: DT1, error bars: Standard deviation
f1_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

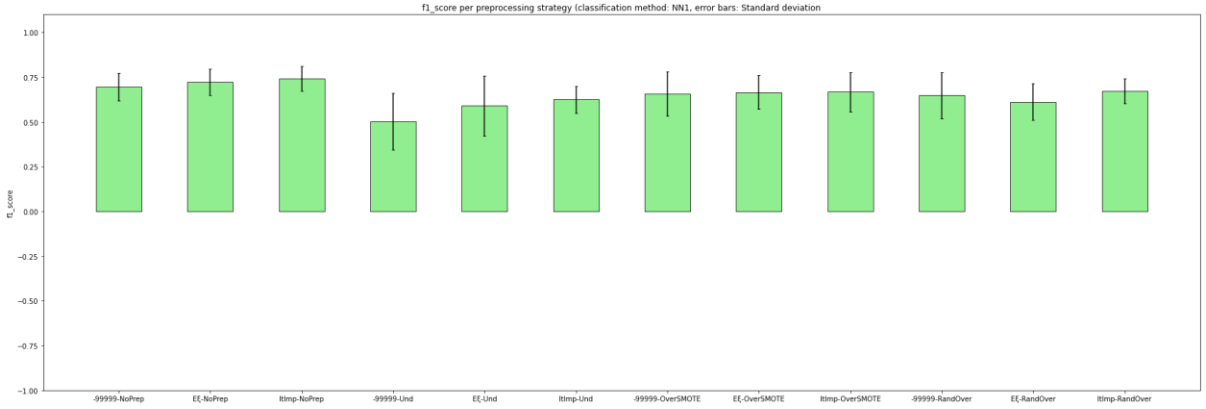

f1_score per preprocessing strategy (classification method: DT2, error bars: Standard deviation
f1_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

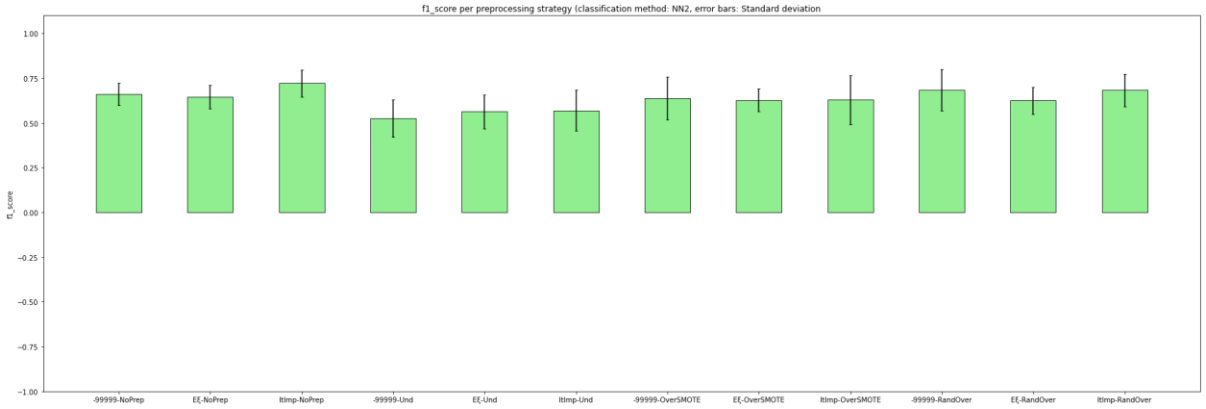

f1_score per preprocessing strategy (classification method: DT3, error bars: Standard deviation
f1_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

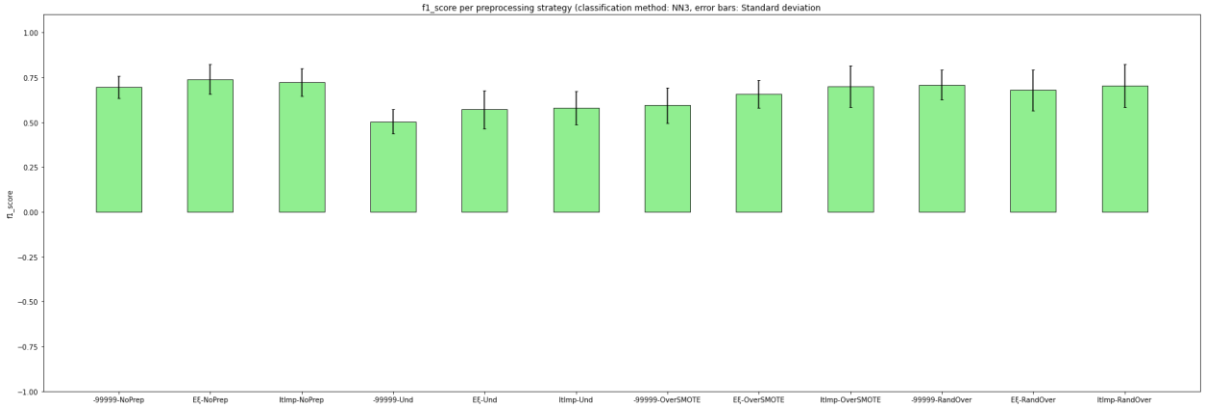

f1_score per preprocessing strategy (classification method: DT4, error bars: Standard deviation
f1_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

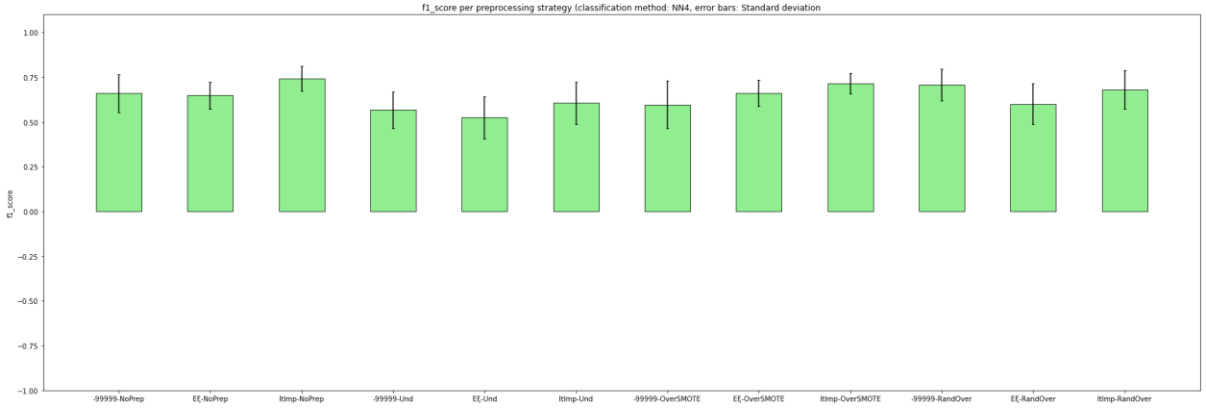

f1_score per preprocessing strategy (classification method: RF1, error bars: Standard deviation
f1_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

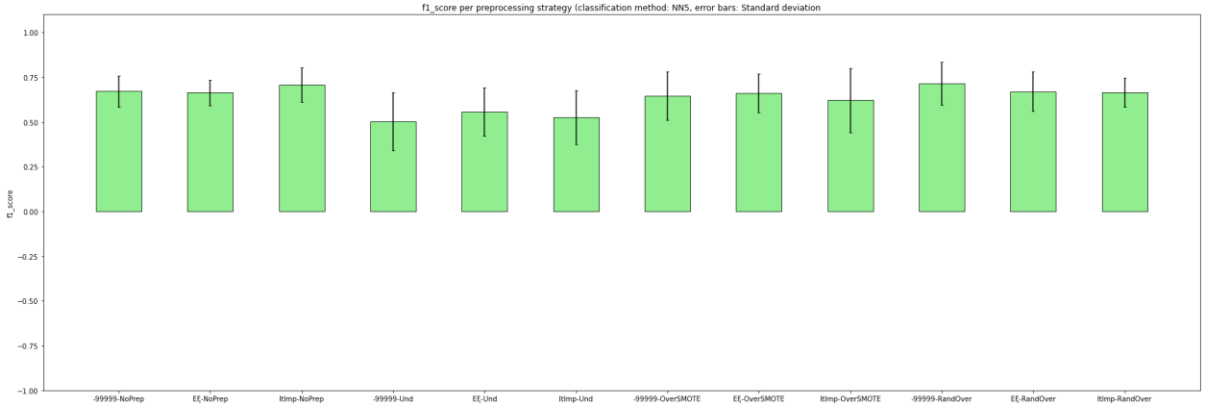

f1_score per preprocessing strategy (classification method: RF2, error bars: Standard deviation
f1_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

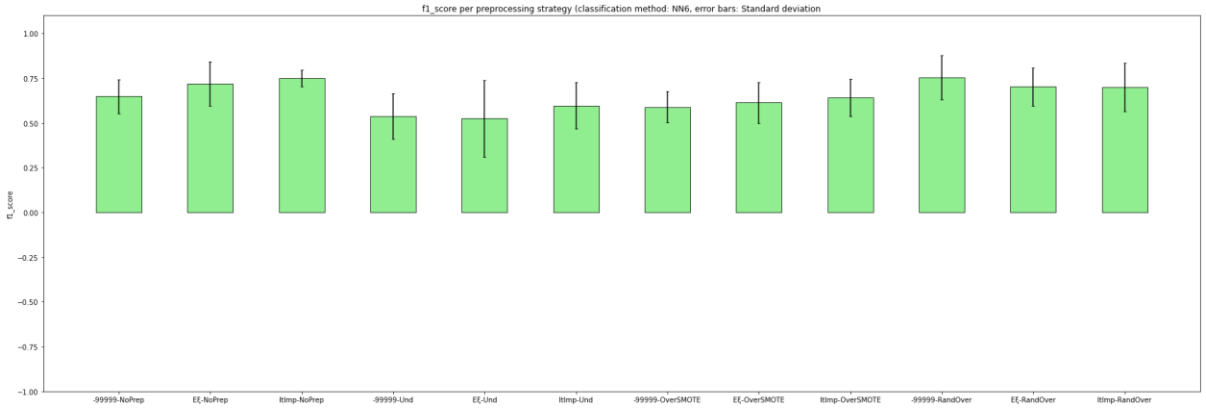

f1_score per preprocessing strategy (classification method: NN1, error bars: Standard deviation
f1_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

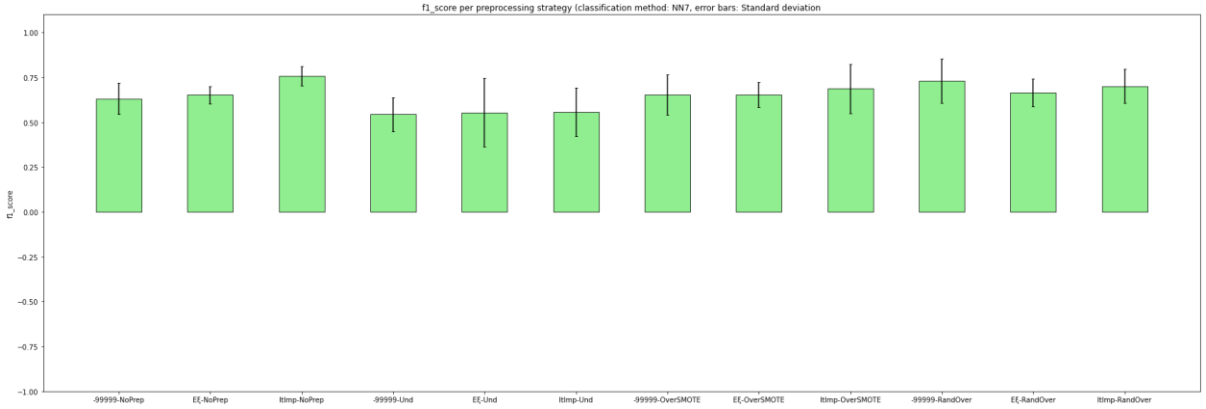

f1_score per preprocessing strategy (classification method: NN2, error bars: Standard deviation
f1_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

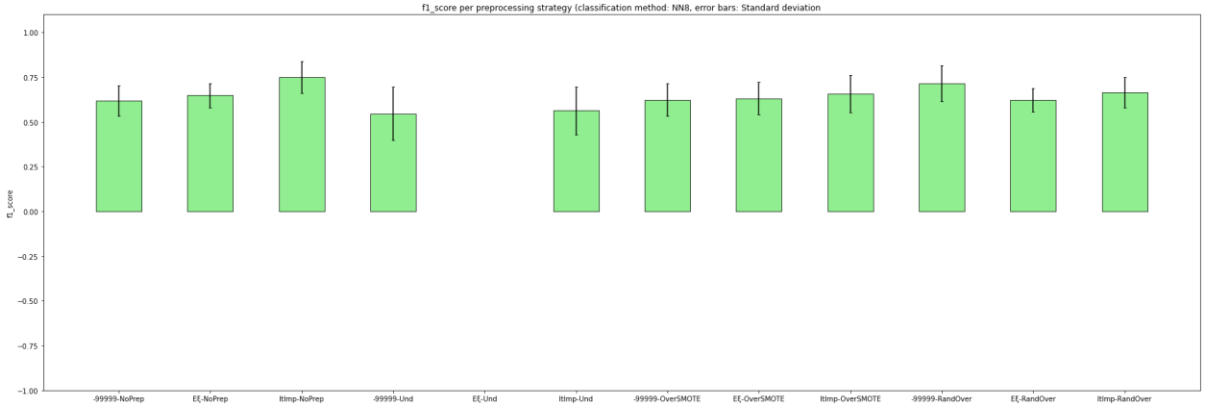

f1_score per preprocessing strategy (classification method: NN3, error bars: Standard deviation
f1_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

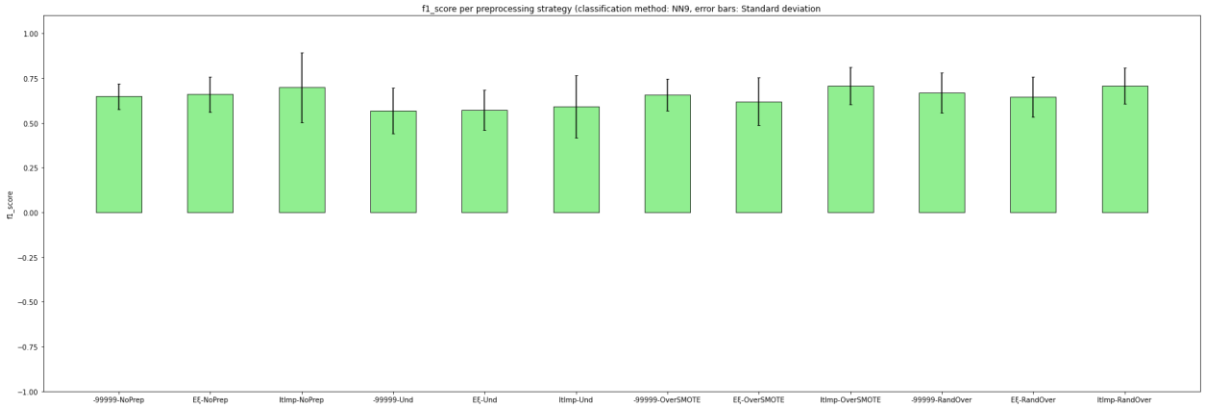

f1_score per preprocessing strategy (classification method: NN4, error bars: Standard deviation
f1_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

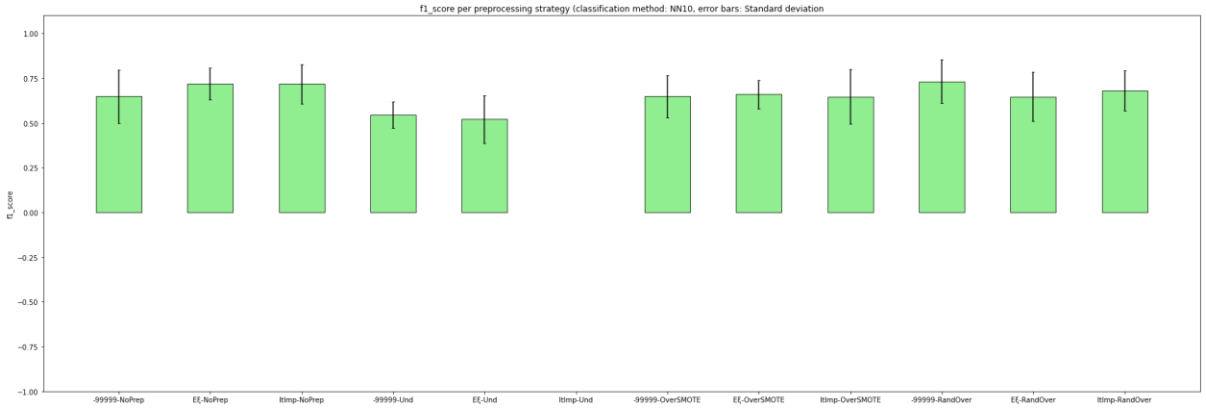

f1_score per preprocessing strategy (classification method: NN5, error bars: Standard deviation
f1_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

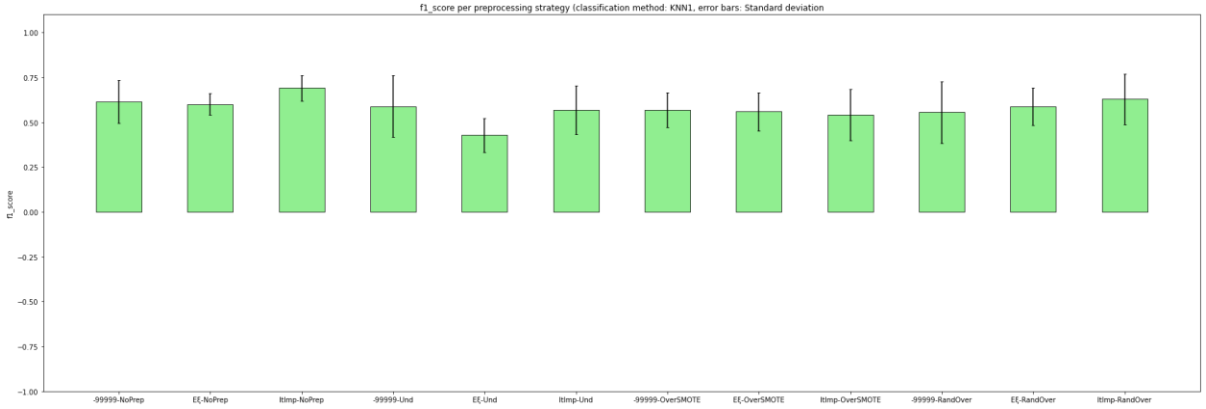

f1_score per preprocessing strategy (classification method: NN6, error bars: Standard deviation
f1_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

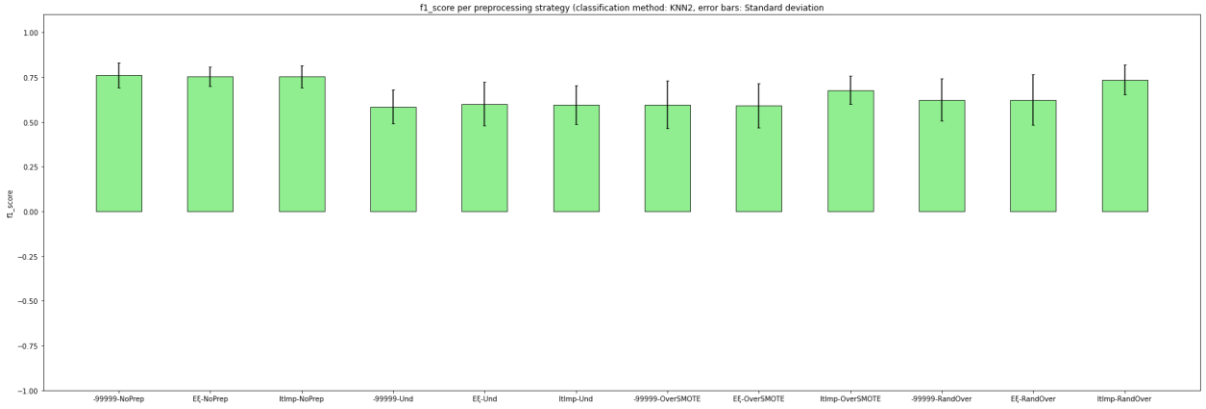

f1_score per preprocessing strategy (classification method: NN7, error bars: Standard deviation
f1_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

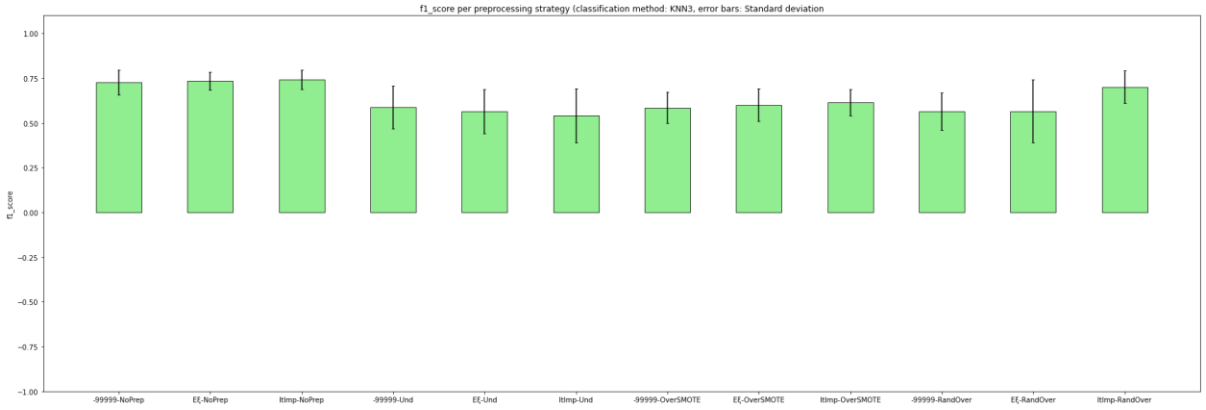

f1_score per preprocessing strategy (classification method: NN8, error bars: Standard deviation
f1_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ltImp-NoPrep
-99999-Und
EE-Und
ltImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ltImp-OverSMOTE
-99999-RandOver
EE-RandOver
ltImp-RandOver

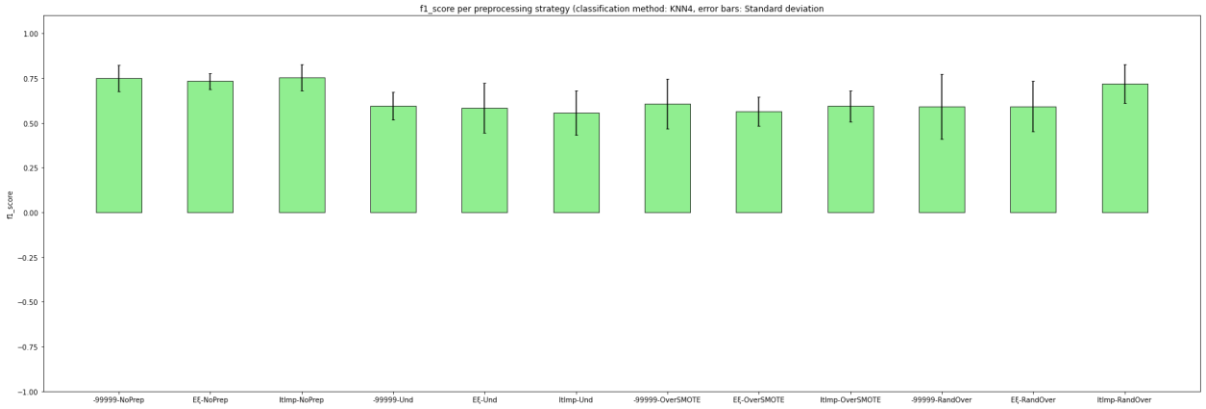

f1_score per preprocessing strategy (classification method: NN9, error bars: Standard deviation
f1_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ltImp-NoPrep
-99999-Und
EE-Und
ltImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ltImp-OverSMOTE
-99999-RandOver
EE-RandOver
ltImp-RandOver

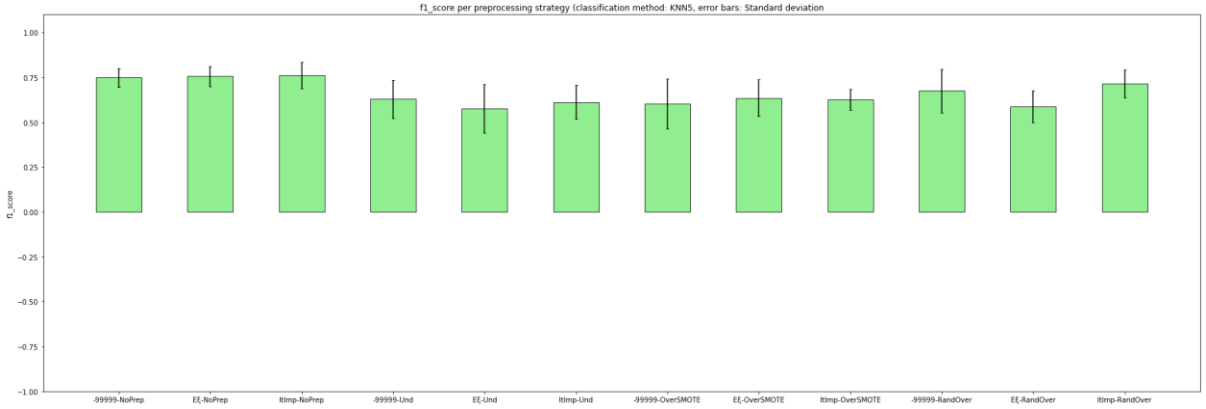

f1_score per preprocessing strategy (classification method: NN10, error bars: Standard deviation
f1_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ltImp-NoPrep
-99999-Und
EE-Und
ltImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ltImp-OverSMOTE
-99999-RandOver
EE-RandOver
ltImp-RandOver

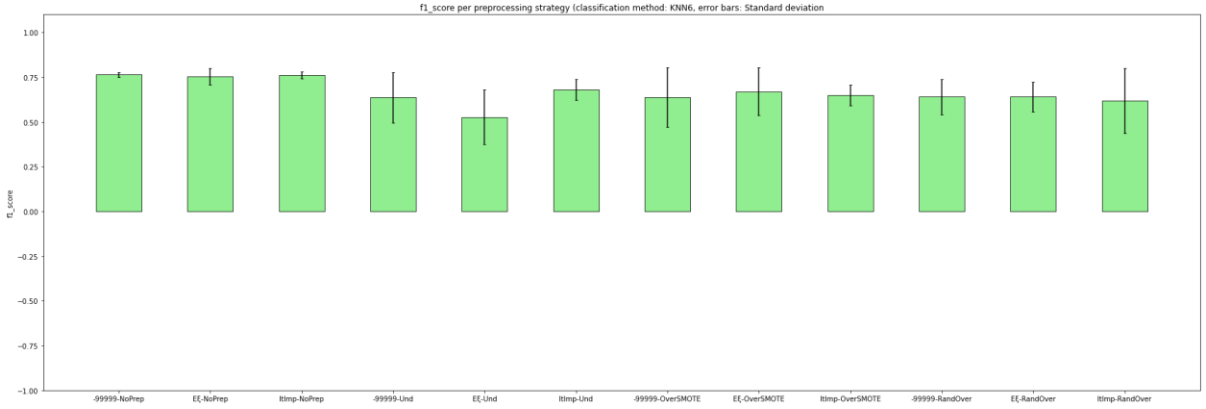

f1_score per preprocessing strategy (classification method: KNN1, error bars: Standard deviation
f1_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ltImp-NoPrep
-99999-Und
EE-Und
ltImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ltImp-OverSMOTE
-99999-RandOver
EE-RandOver
ltImp-RandOver

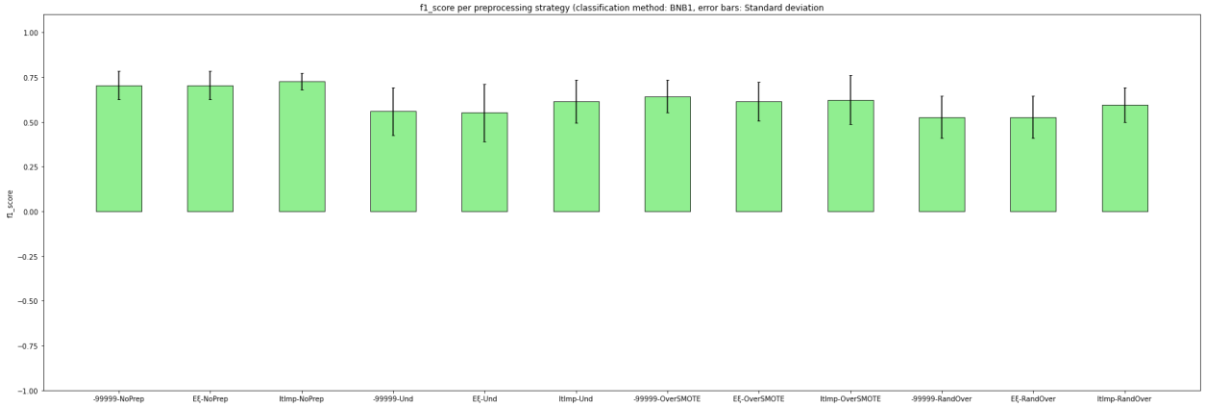

f1_score per preprocessing strategy (classification method: KNN2, error bars: Standard deviation
f1_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ltImp-NoPrep
-99999-Und
EE-Und
ltImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ltImp-OverSMOTE
-99999-RandOver
EE-RandOver
ltImp-RandOver

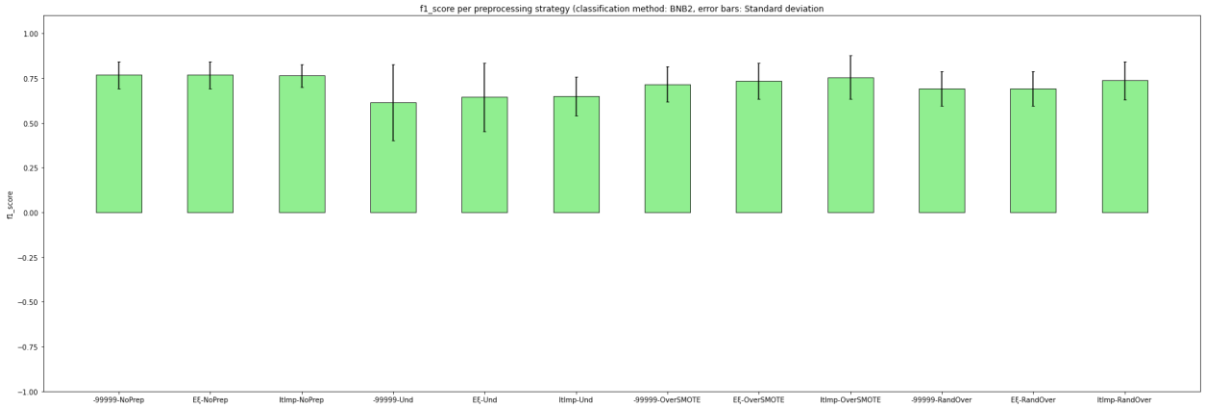

f1_score per preprocessing strategy (classification method: KNN3, error bars: Standard deviation
f1_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ltImp-NoPrep
-99999-Und
EE-Und
ltImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ltImp-OverSMOTE
-99999-RandOver
EE-RandOver
ltImp-RandOver

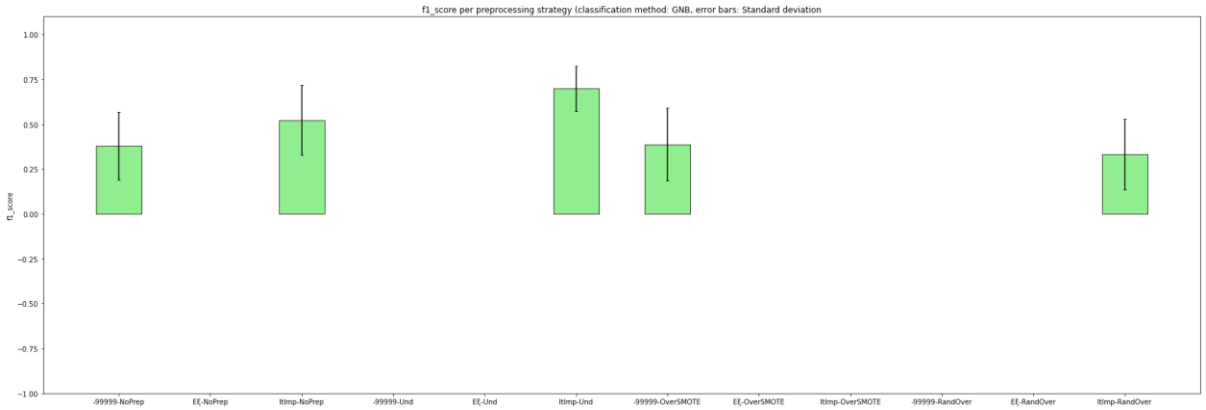

f1_score per preprocessing strategy (classification method: KNN4, error bars: Standard deviation
f1_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ltImp-NoPrep
-99999-Und
EE-Und
ltImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ltImp-OverSMOTE
-99999-RandOver
EE-RandOver
ltImp-RandOver

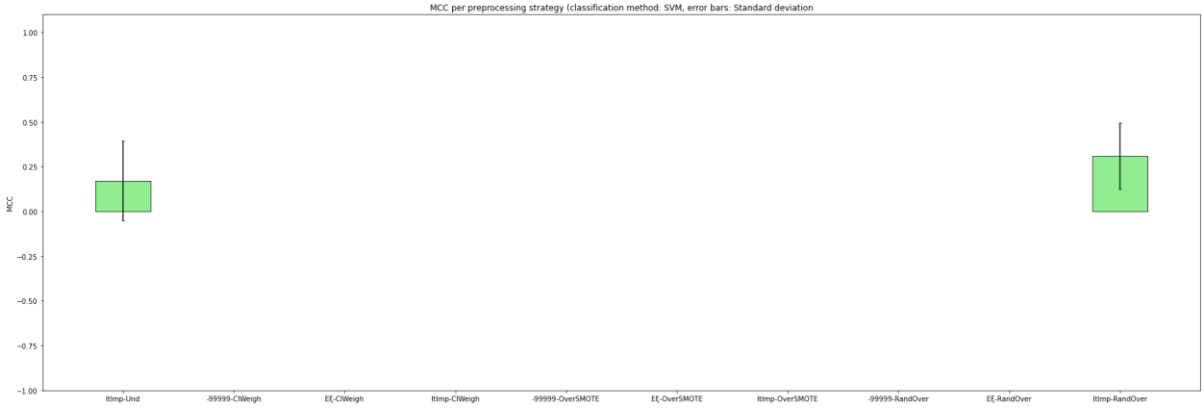

f1_score per preprocessing strategy (classification method: KNN5, error bars: Standard deviation
f1_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ltImp-NoPrep
-99999-Und
EE-Und
ltImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ltImp-OverSMOTE
-99999-RandOver
EE-RandOver
ltImp-RandOver

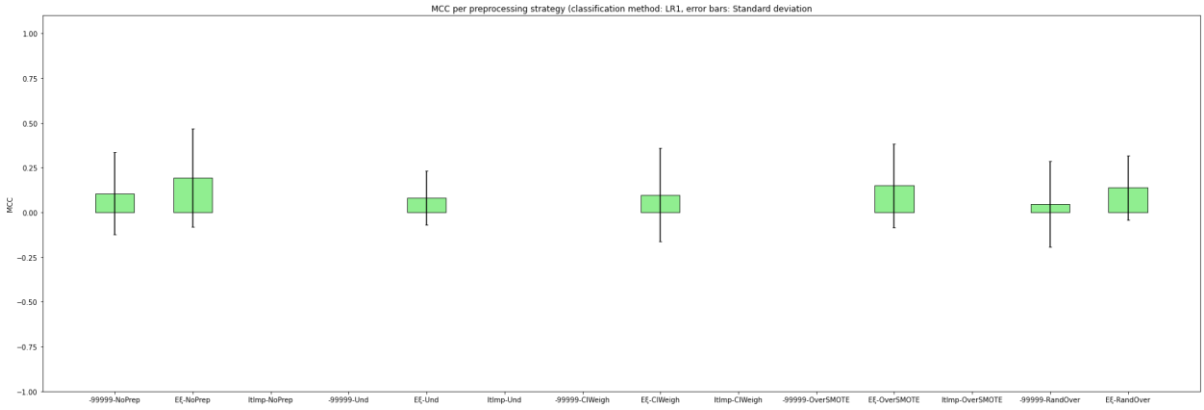

f1_score per preprocessing strategy (classification method: KNN6, error bars: Standard deviation
f1_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

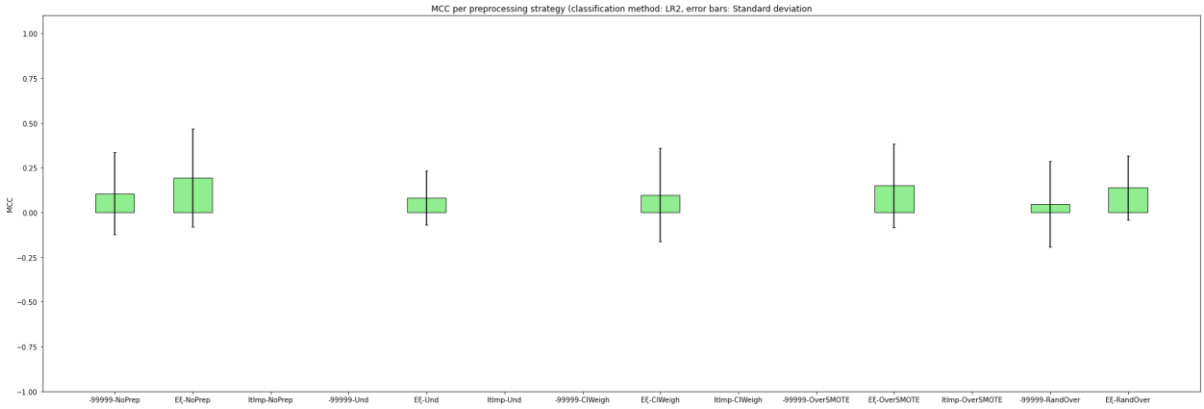

f1_score per preprocessing strategy (classification method: BNB1, error bars: Standard deviation
f1_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

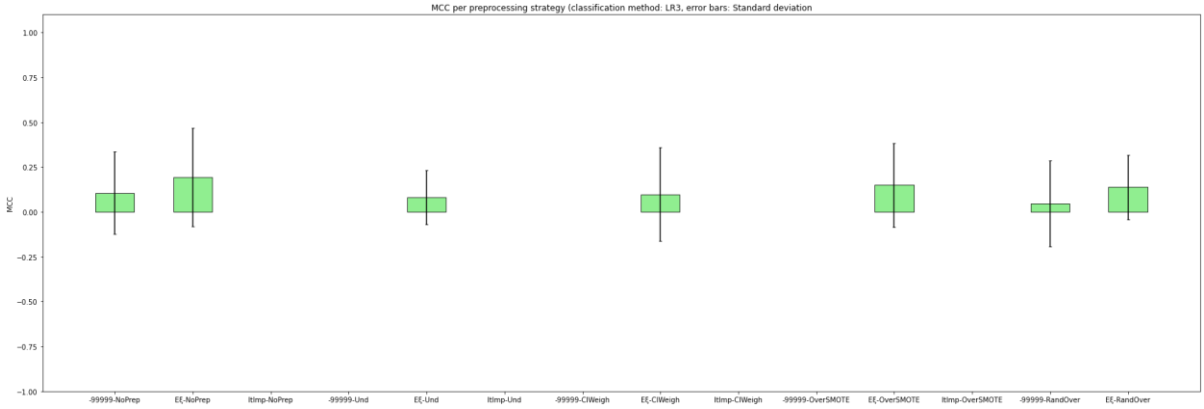

f1_score per preprocessing strategy (classification method: BNB2, error bars: Standard deviation
f1_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

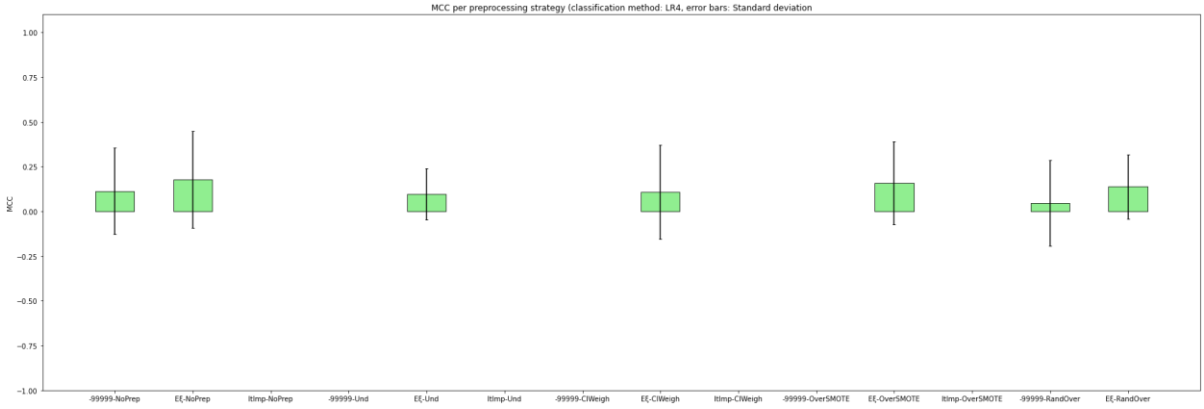

f1_score per preprocessing strategy (classification method: GNB, error bars: Standard deviation
f1_score
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

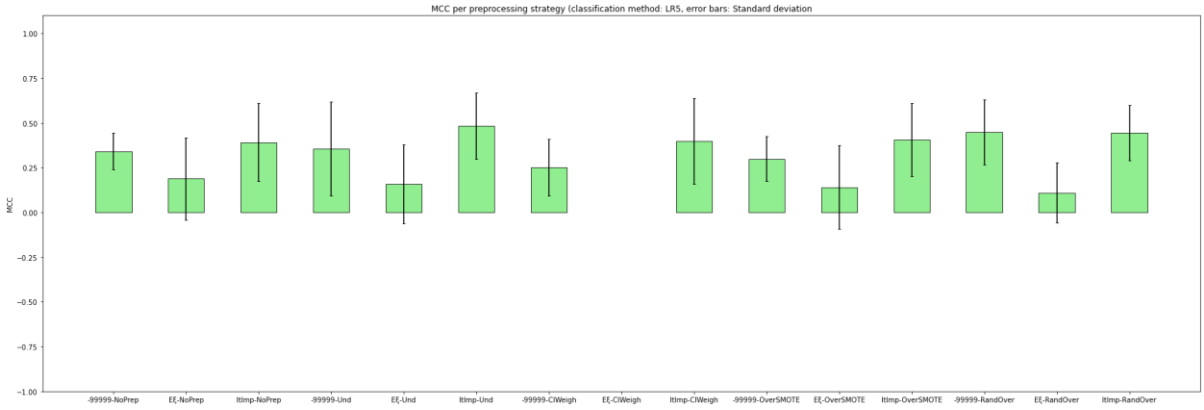

MCC per preprocessing strategy (classification method: SVM, error bars: Standard deviation
MCC
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
ItImp-Und
-99999-ClWeigh
EE-ClWeigh
ItImp-ClWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

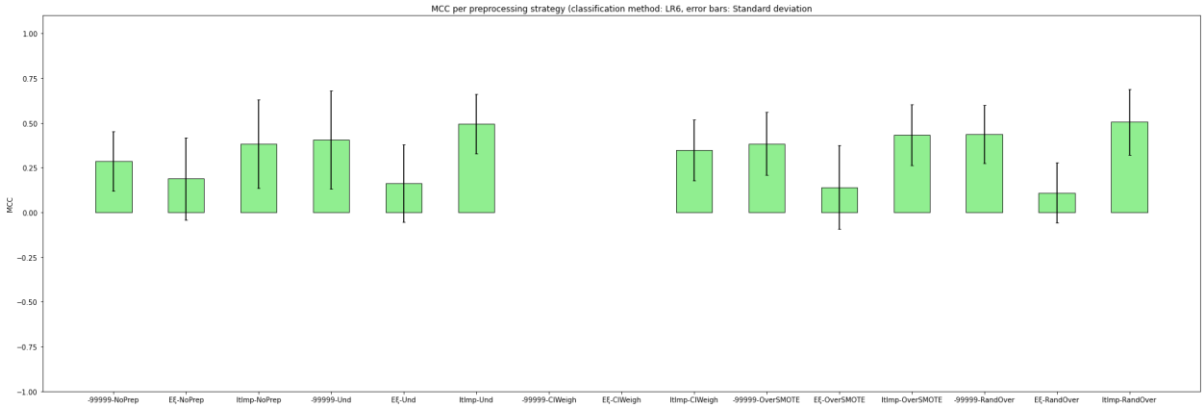

MCC per preprocessing strategy (classification method: LR1, error bars: Standard deviation
MCC
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-ClWeigh
EE-ClWeigh
ItImp-ClWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver

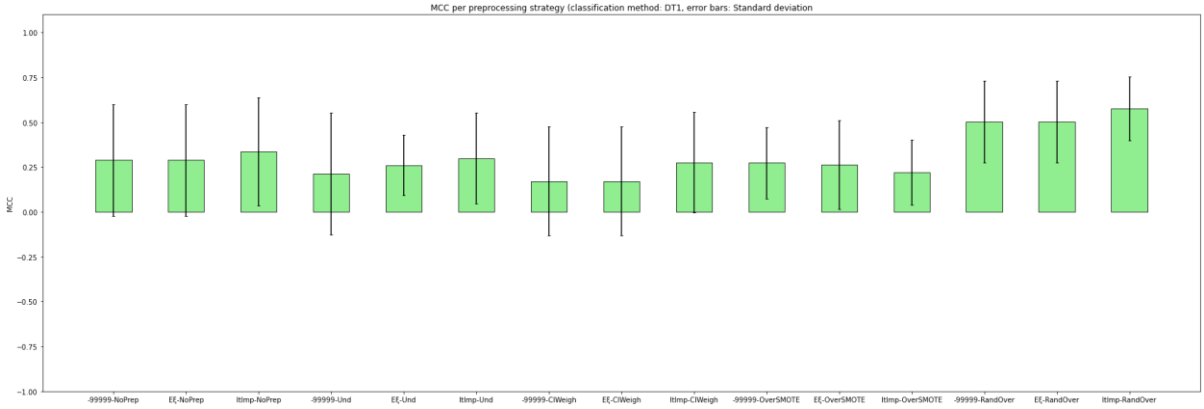

MCC per preprocessing strategy (classification method: LR2, error bars: Standard deviation
MCC
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-ClWeigh
EE-ClWeigh
ItImp-ClWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver

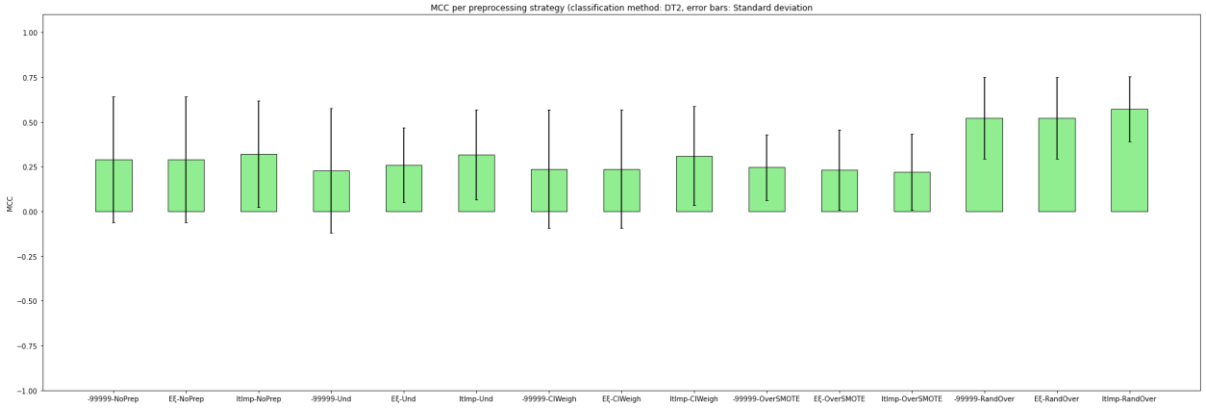

MCC per preprocessing strategy (classification method: LR3, error bars: Standard deviation
MCC
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-ClWeigh
EE-ClWeigh
ItImp-ClWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver

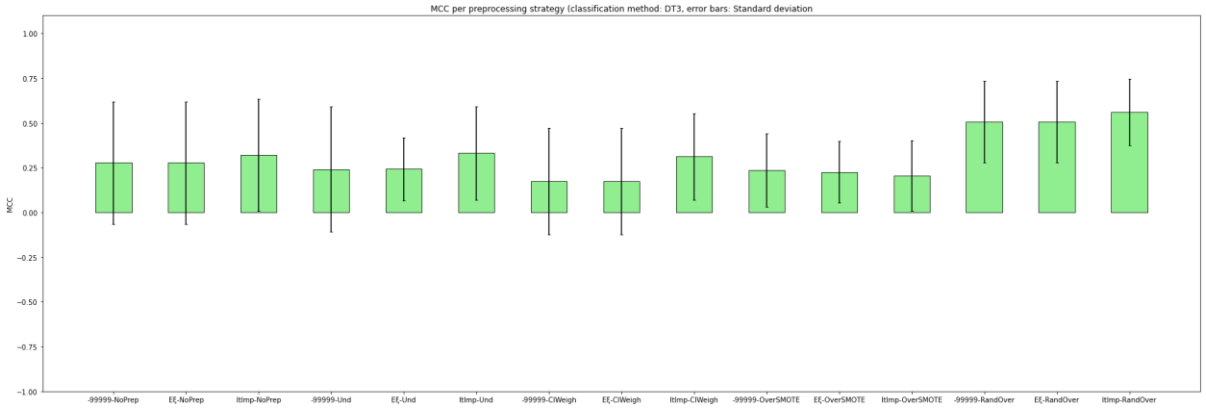

MCC per preprocessing strategy (classification method: LR4, error bars: Standard deviation
MCC
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver

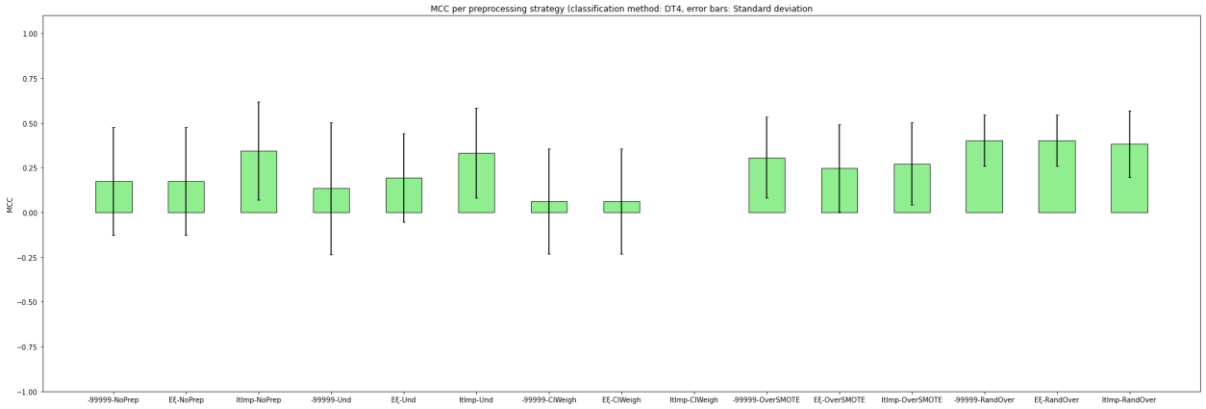

MCC per preprocessing strategy (classification method: LR5, error bars: Standard deviation
MCC
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

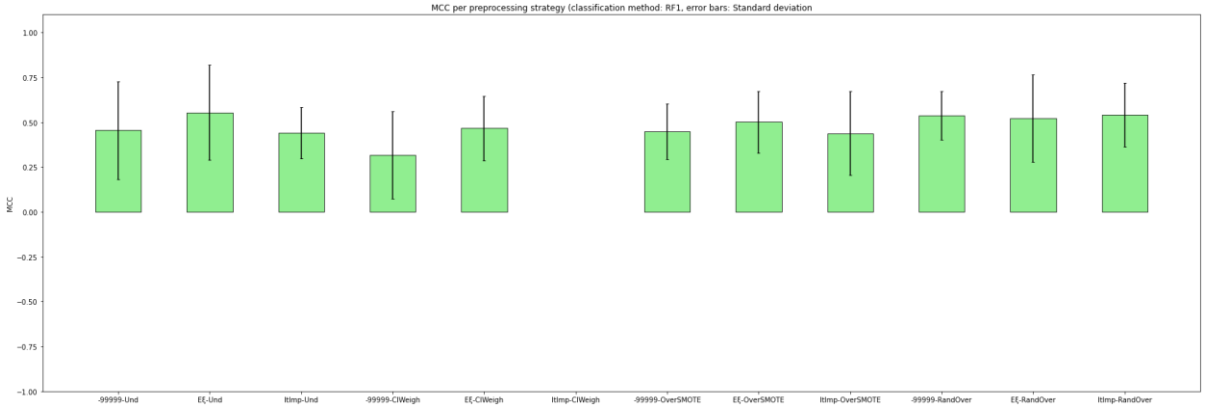

MCC per preprocessing strategy (classification method: LR6, error bars: Standard deviation
MCC
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

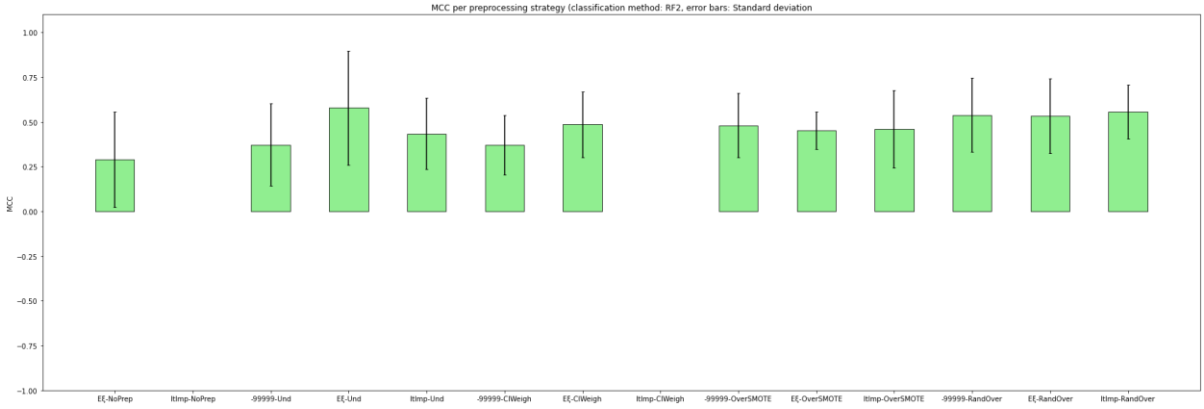

MCC per preprocessing strategy (classification method: DT1, error bars: Standard deviation
MCC
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

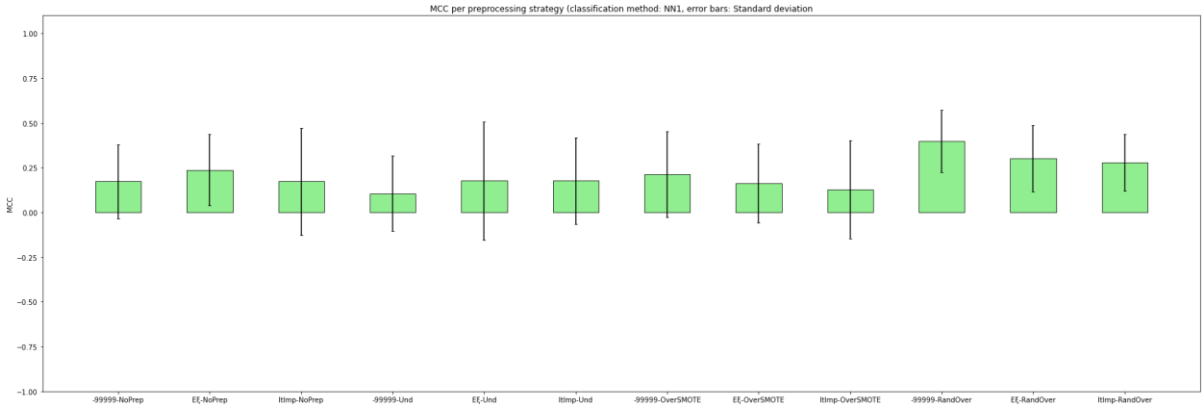

MCC per preprocessing strategy (classification method: DT2, error bars: Standard deviation
MCC
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

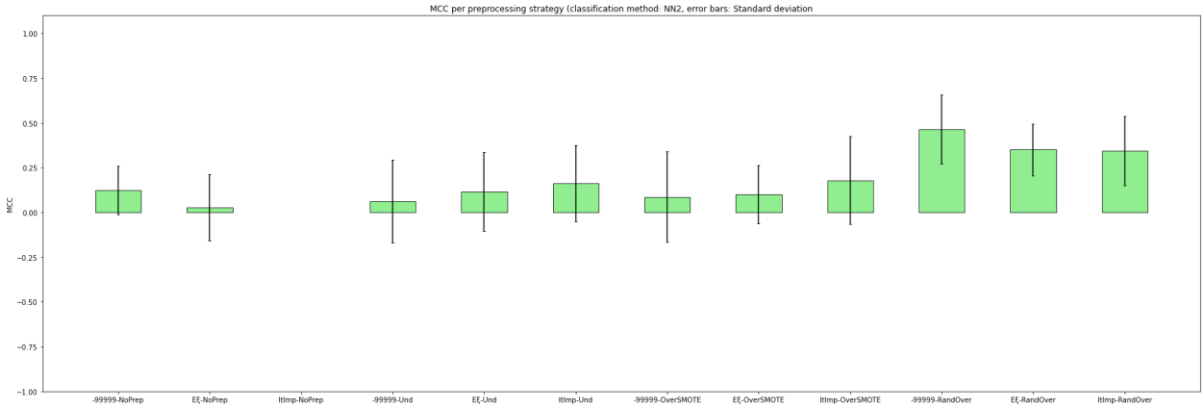

MCC per preprocessing strategy (classification method: DT3, error bars: Standard deviation
MCC
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

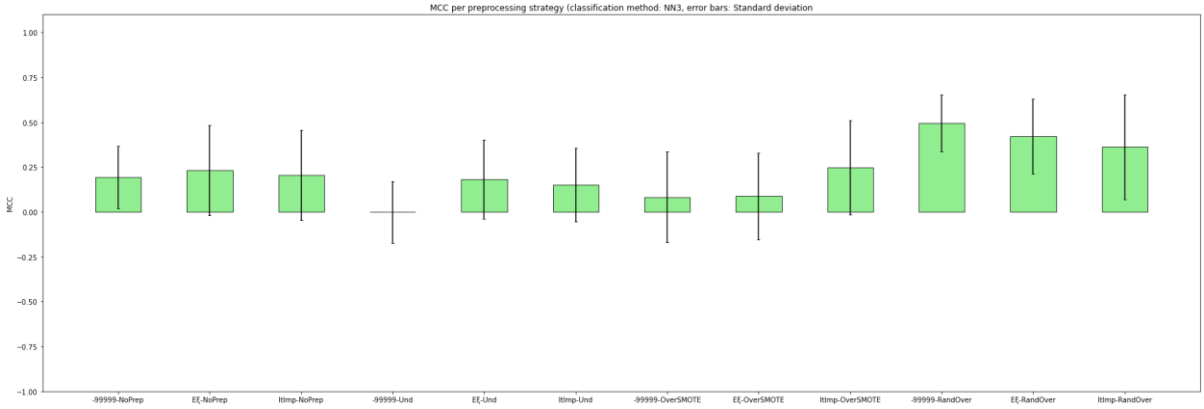

MCC per preprocessing strategy (classification method: DT4, error bars: Standard deviation
MCC
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

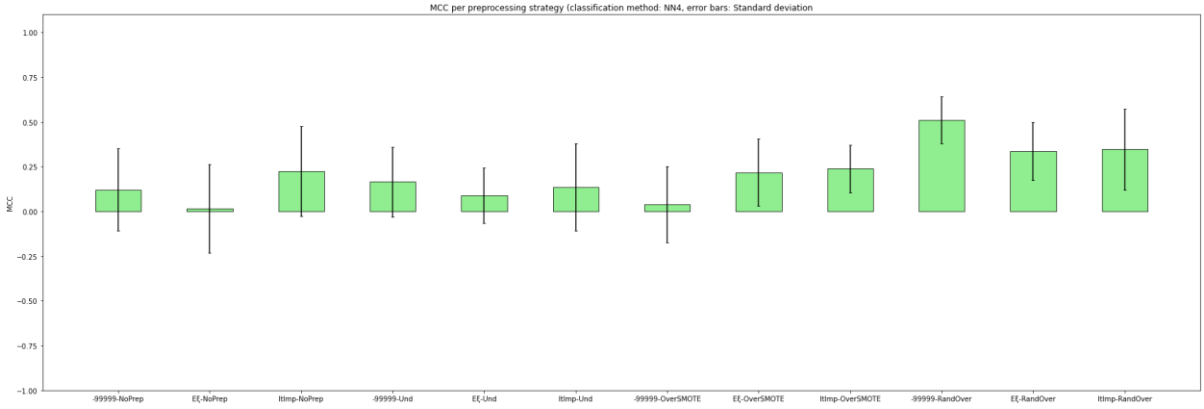

MCC per preprocessing strategy (classification method: RF1, error bars: Standard deviation
MCC
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

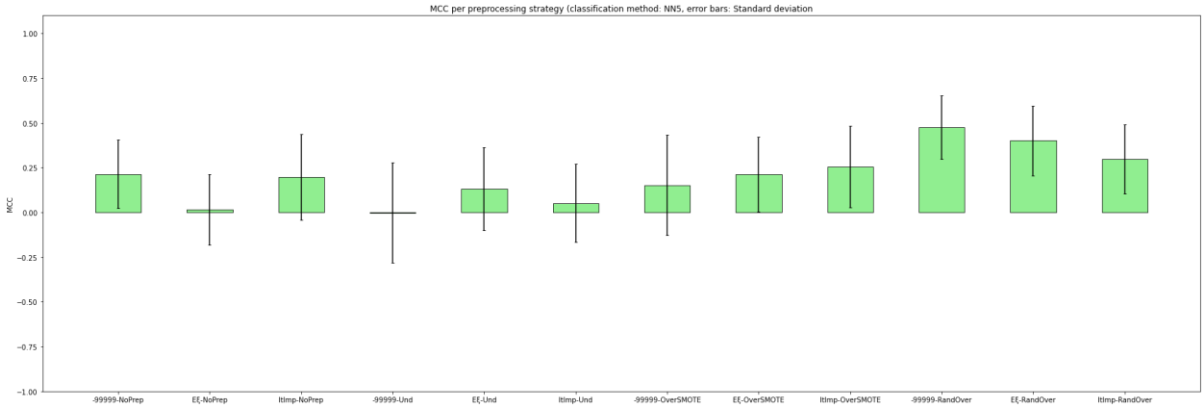

MCC per preprocessing strategy (classification method: RF2, error bars: Standard deviation
MCC
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CtWeigh
EE-CtWeigh
ItImp-CtWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

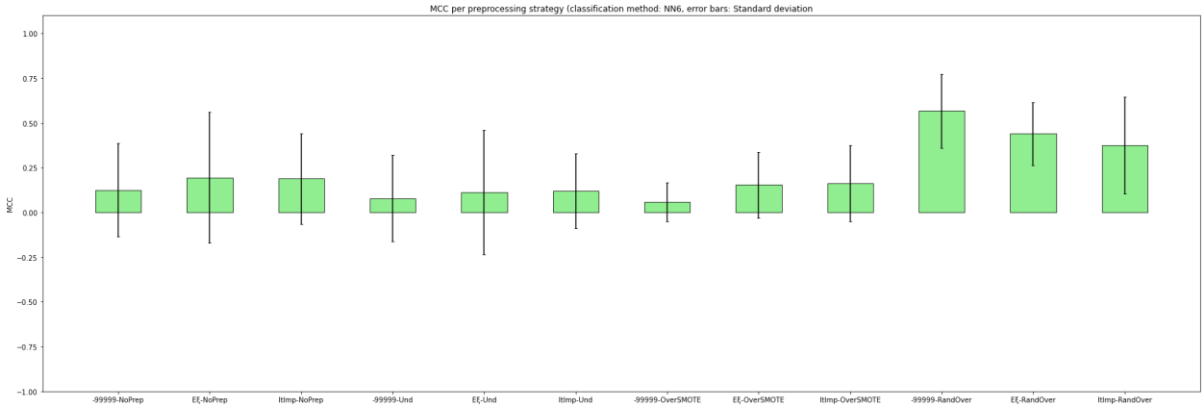

MCC per preprocessing strategy (classification method: NN1, error bars: Standard deviation
MCC
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

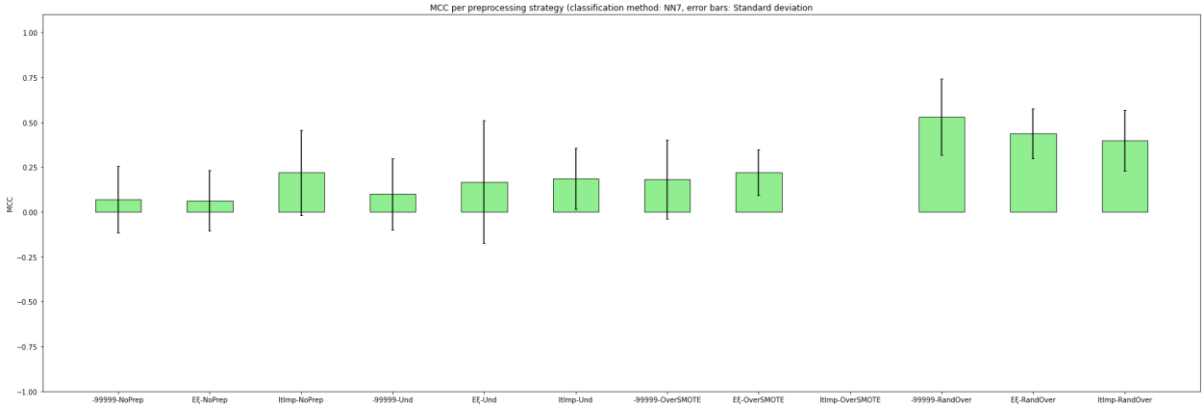

MCC per preprocessing strategy (classification method: NN2, error bars: Standard deviation
MCC
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

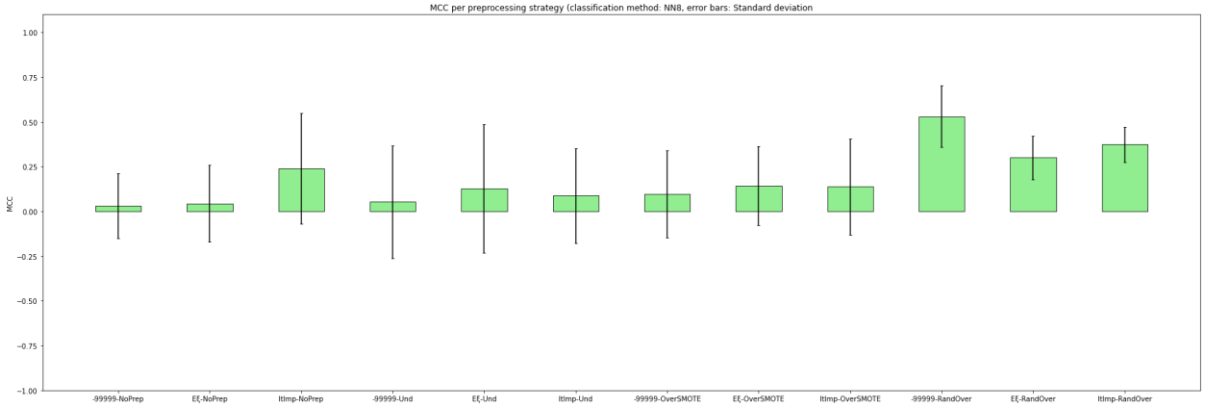

MCC per preprocessing strategy (classification method: NN3, error bars: Standard deviation
MCC
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

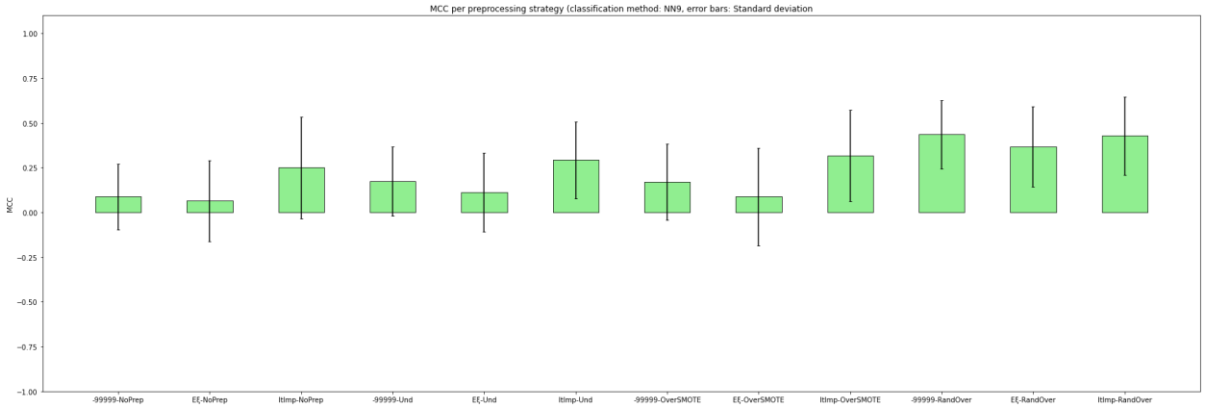

MCC per preprocessing strategy (classification method: NN4, error bars: Standard deviation
MCC
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

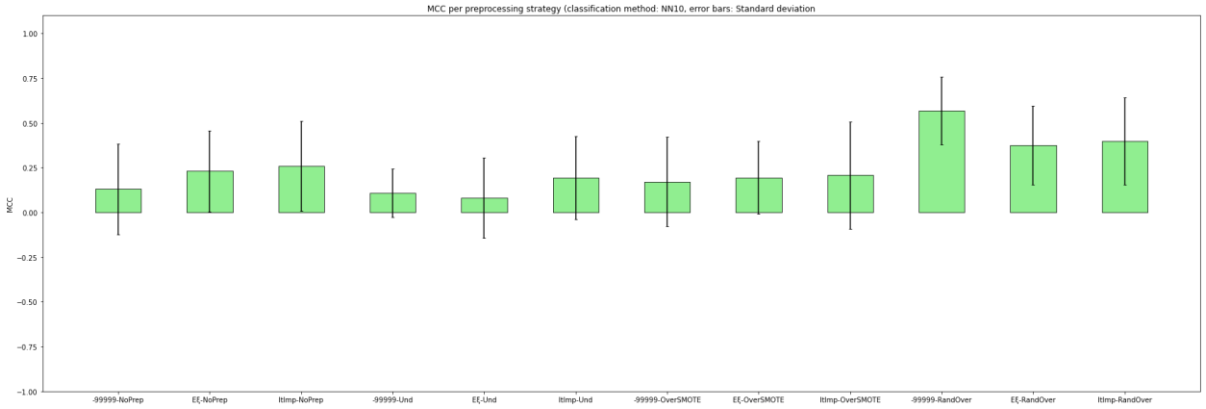

MCC per preprocessing strategy (classification method: NN5, error bars: Standard deviation
MCC
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

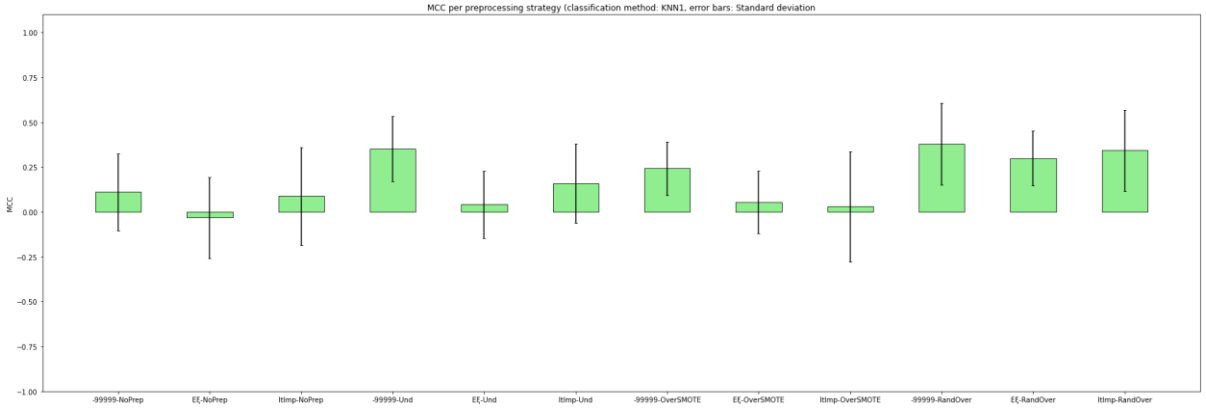

MCC per preprocessing strategy (classification method: NN6, error bars: Standard deviation
MCC
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

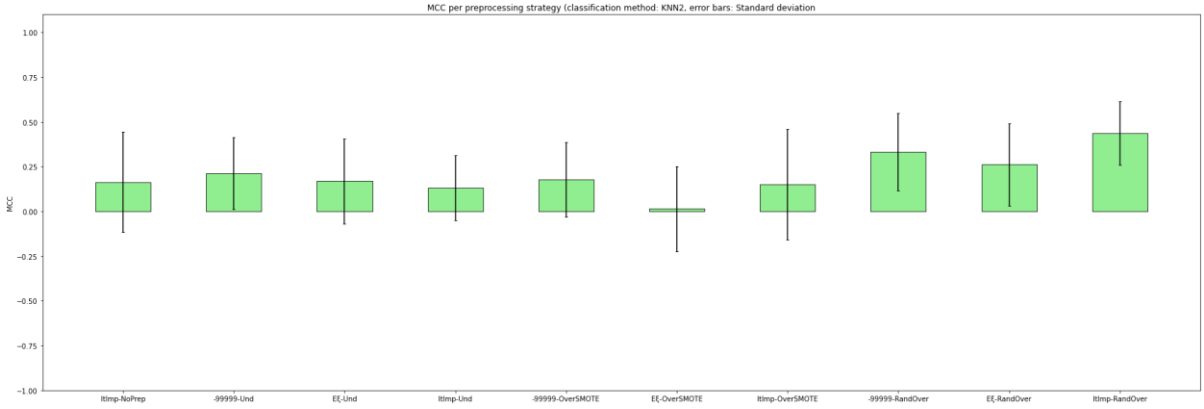

MCC per preprocessing strategy (classification method: NN7, error bars: Standard deviation
MCC
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

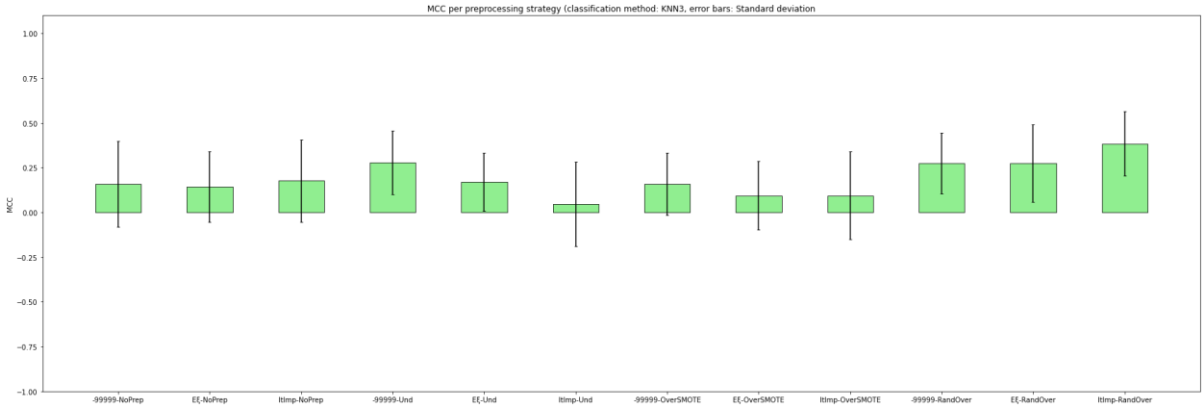

MCC per preprocessing strategy (classification method: NN8, error bars: Standard deviation
MCC
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

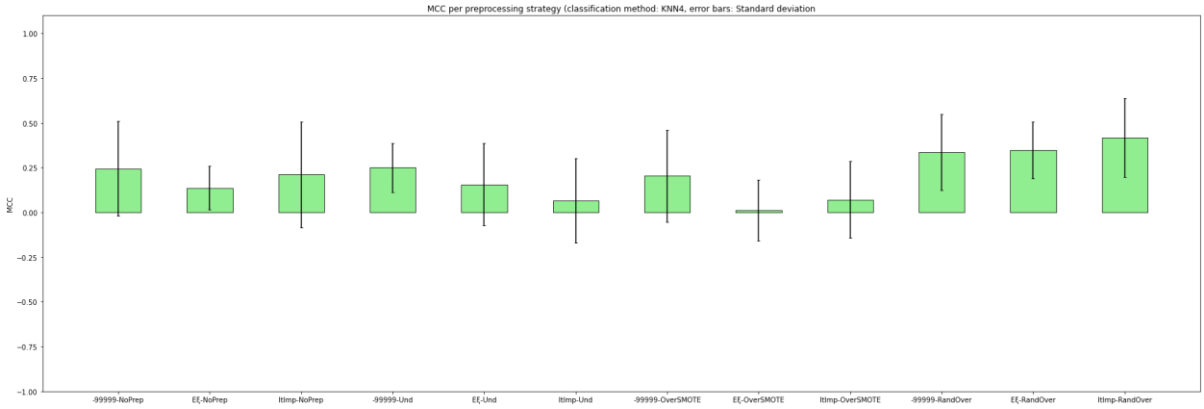

MCC per preprocessing strategy (classification method: NN9, error bars: Standard deviation
MCC
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

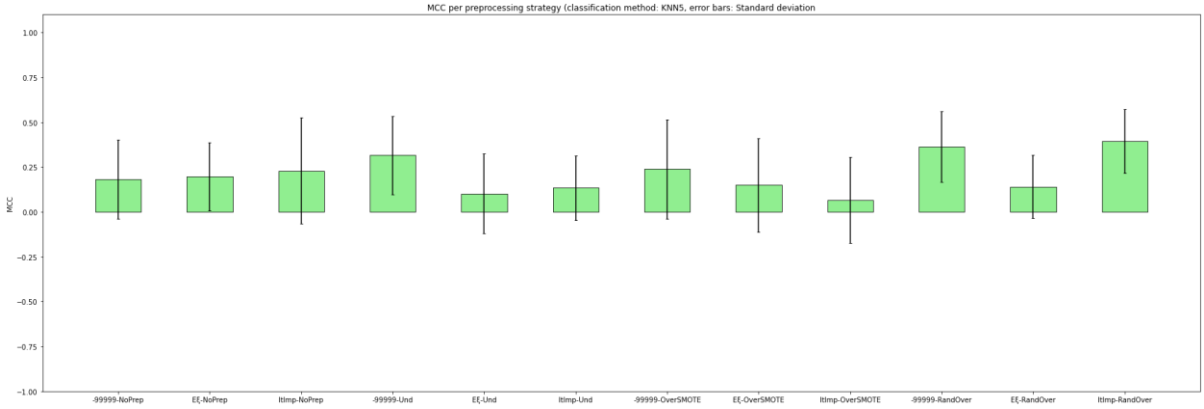

MCC per preprocessing strategy (classification method: NN10, error bars: Standard deviation
MCC
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

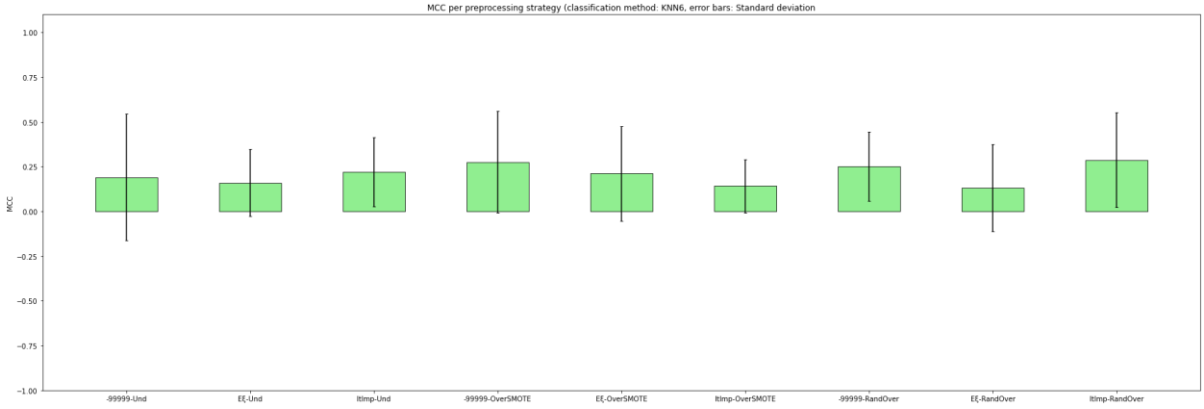

MCC per preprocessing strategy (classification method: KNN1, error bars: Standard deviation
MCC
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

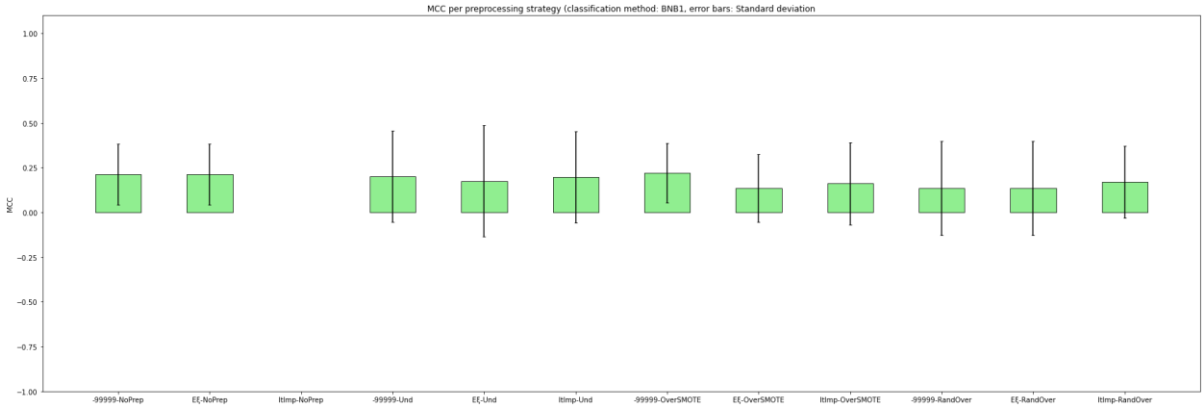

MCC per preprocessing strategy (classification method: KNN2, error bars: Standard deviation
MCC
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

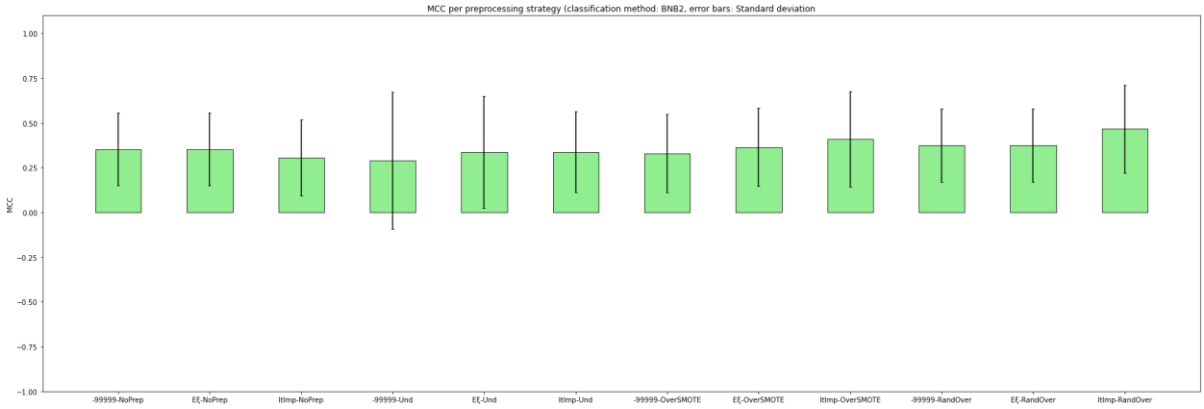

MCC per preprocessing strategy (classification method: KNN3, error bars: Standard deviation
MCC
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

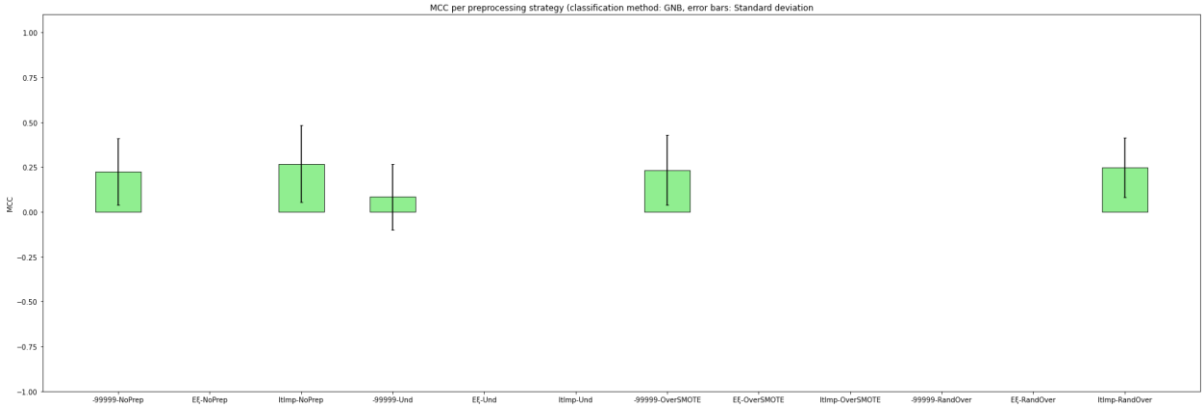

MCC per preprocessing strategy (classification method: KNN4, error bars: Standard deviation
MCC
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

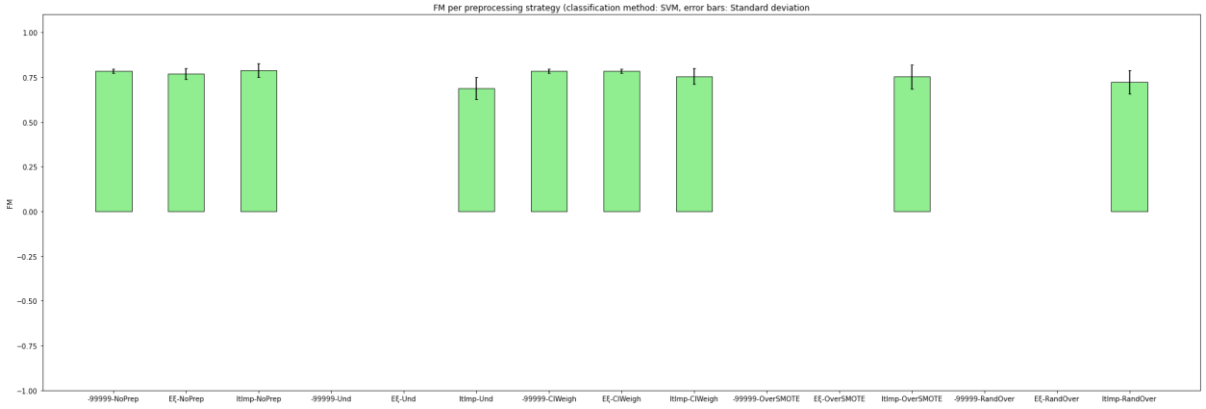

MCC per preprocessing strategy (classification method: KNN5, error bars: Standard deviation
MCC
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

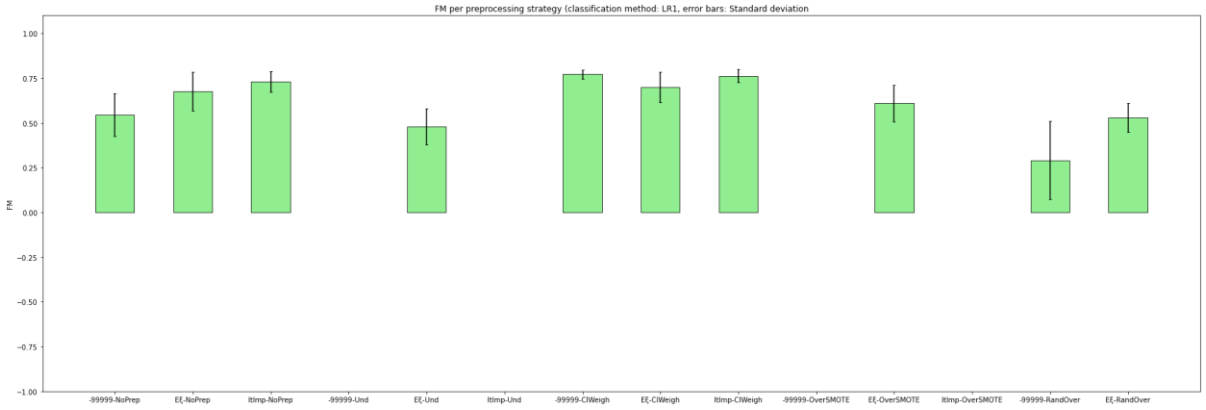

MCC per preprocessing strategy (classification method: KNN6, error bars: Standard deviation
MCC
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

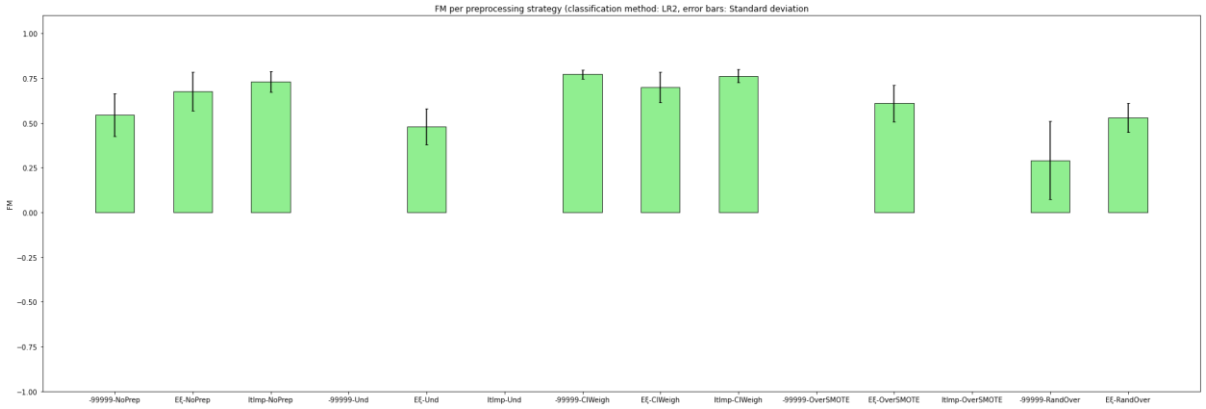

MCC per preprocessing strategy (classification method: BNB1, error bars: Standard deviation
MCC
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

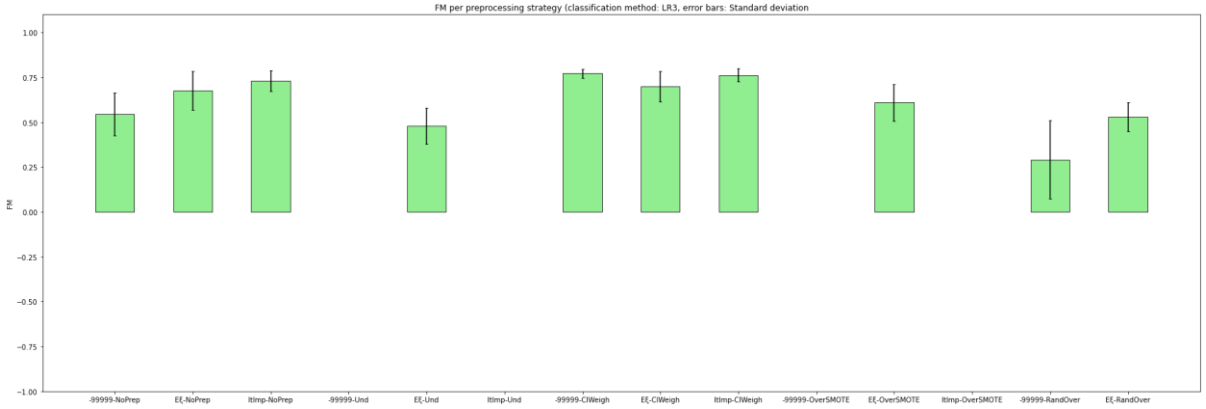

MCC per preprocessing strategy (classification method: BNB2, error bars: Standard deviation
MCC
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

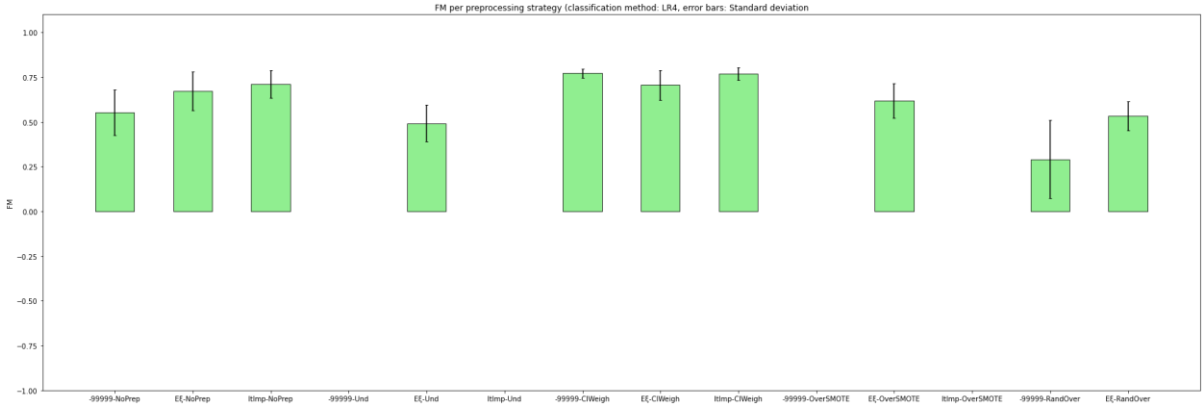

MCC per preprocessing strategy (classification method: GNB, error bars: Standard deviation
MCC
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

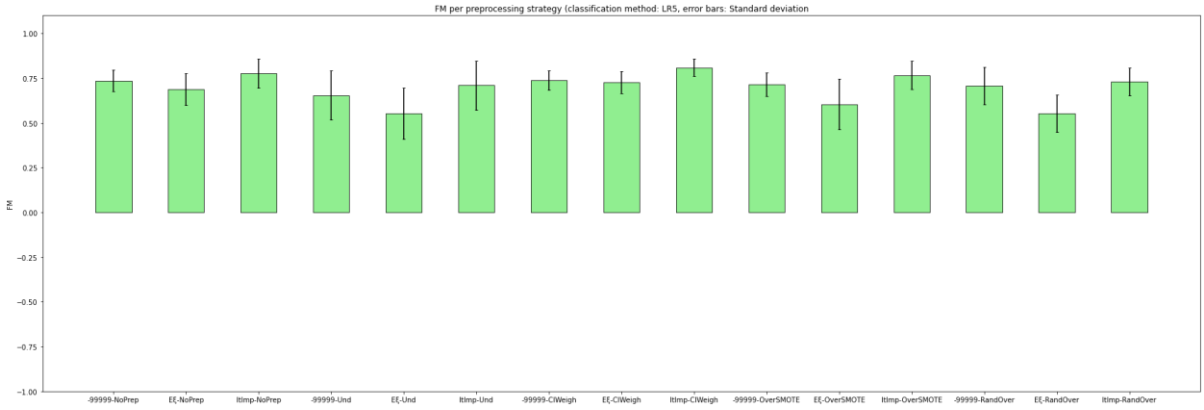

FM per preprocessing strategy (classification method: SVM, error bars: Standard deviation
FM
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

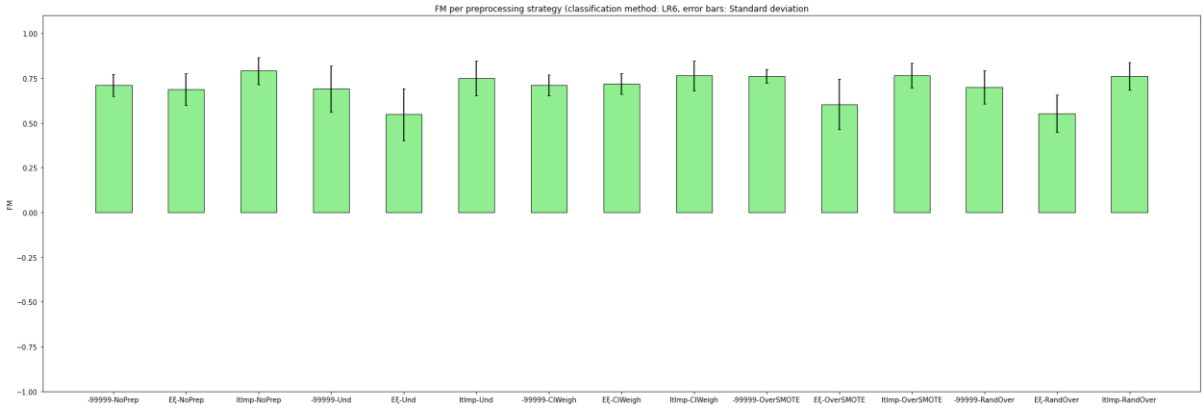

FM per preprocessing strategy (classification method: LR1, error bars: Standard deviation
FM
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver

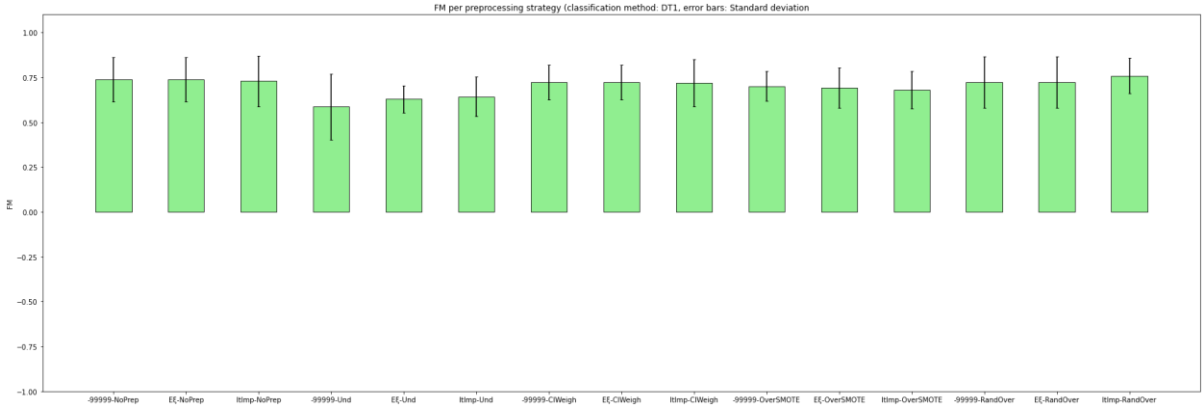

FM per preprocessing strategy (classification method: LR2, error bars: Standard deviation
FM
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver

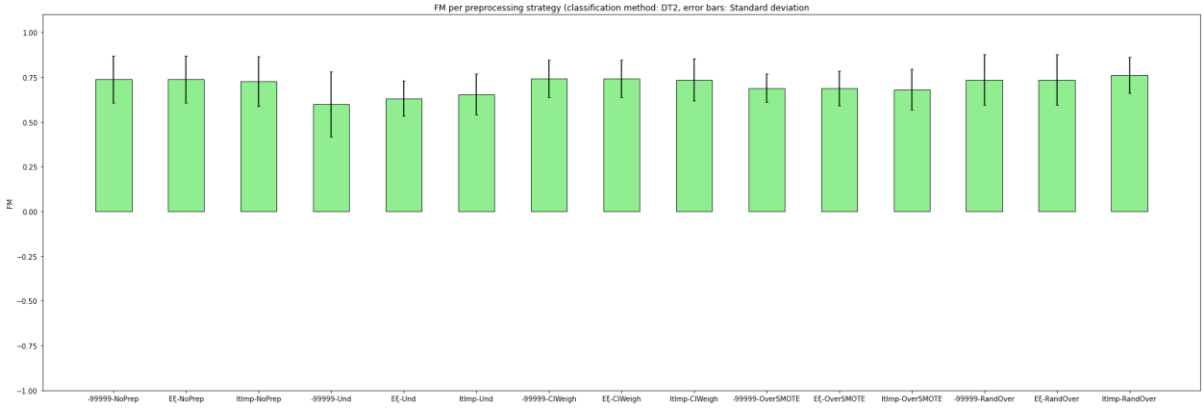

FM per preprocessing strategy (classification method: LR3, error bars: Standard deviation
FM
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver

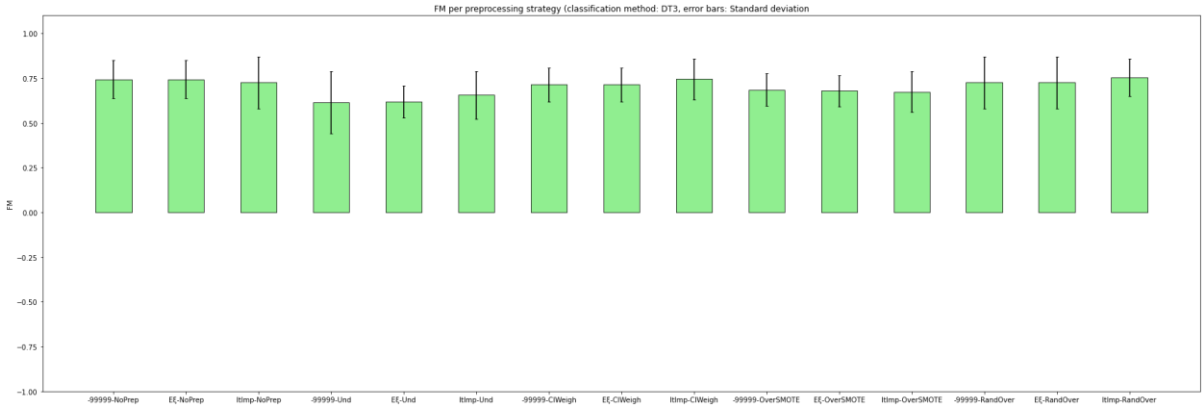

FM per preprocessing strategy (classification method: LR4, error bars: Standard deviation
FM
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver

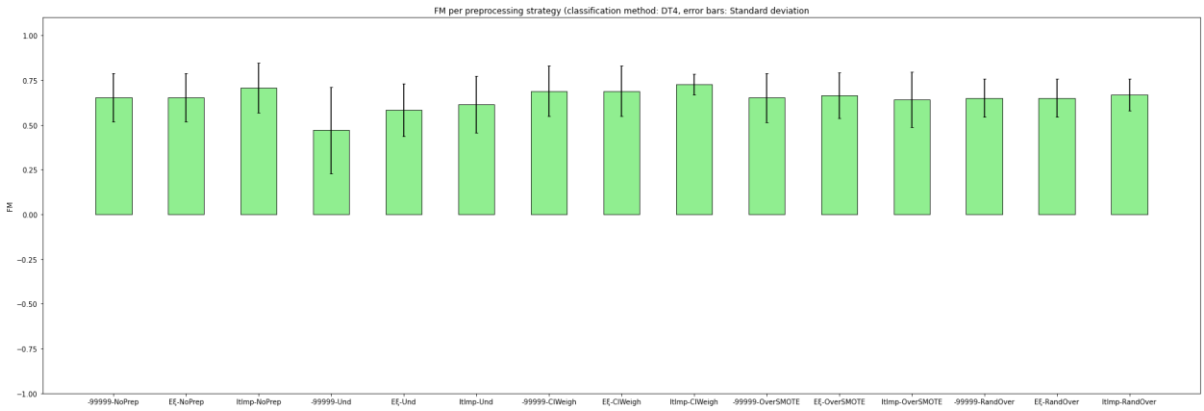

FM per preprocessing strategy (classification method: LR5, error bars: Standard deviation
FM
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

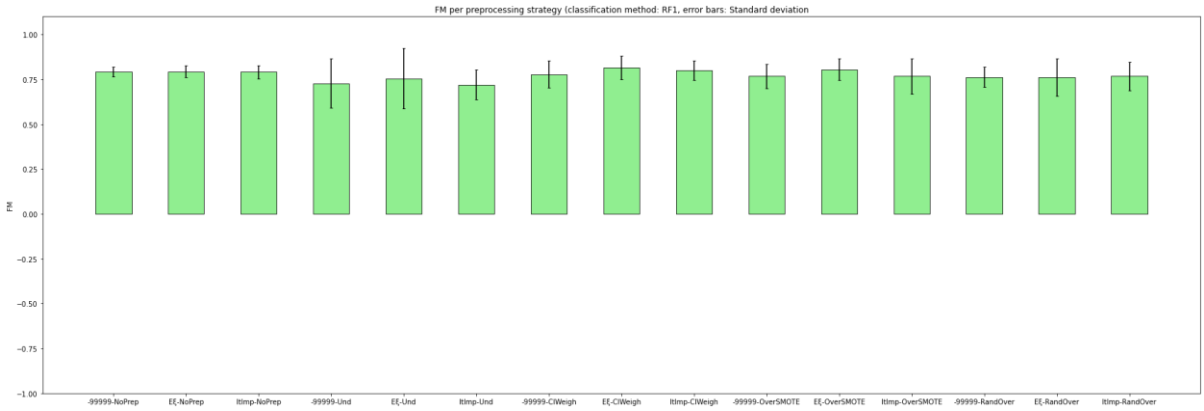

FM per preprocessing strategy (classification method: LR6, error bars: Standard deviation
FM
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

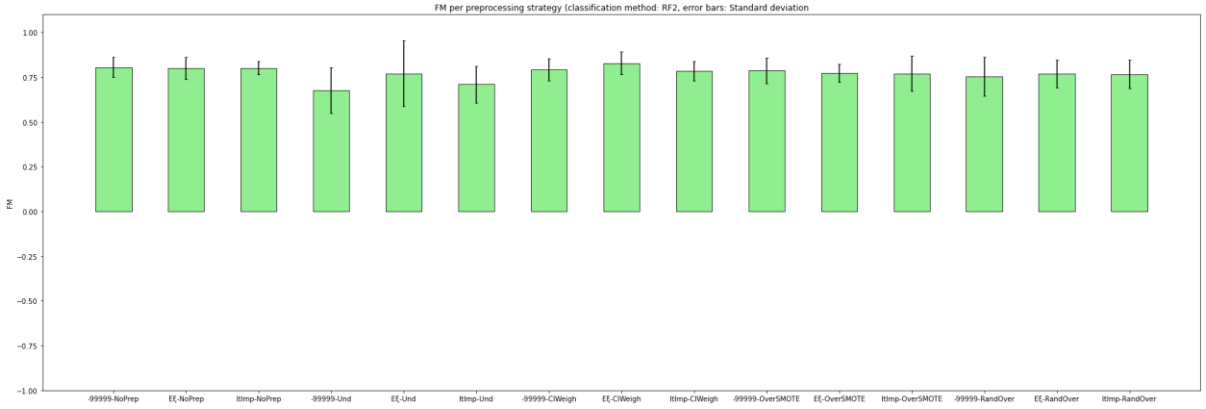

FM per preprocessing strategy (classification method: DT1, error bars: Standard deviation
FM
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

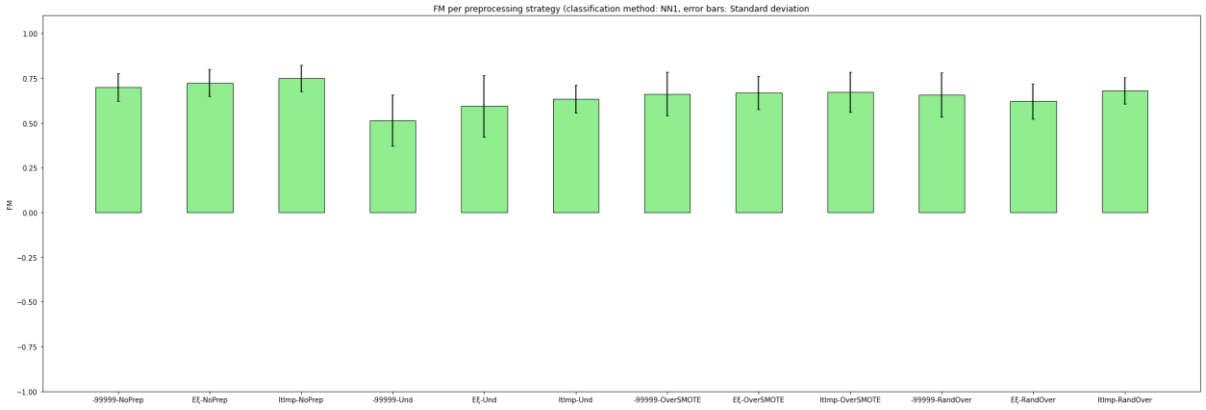

FM per preprocessing strategy (classification method: DT2, error bars: Standard deviation
FM
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

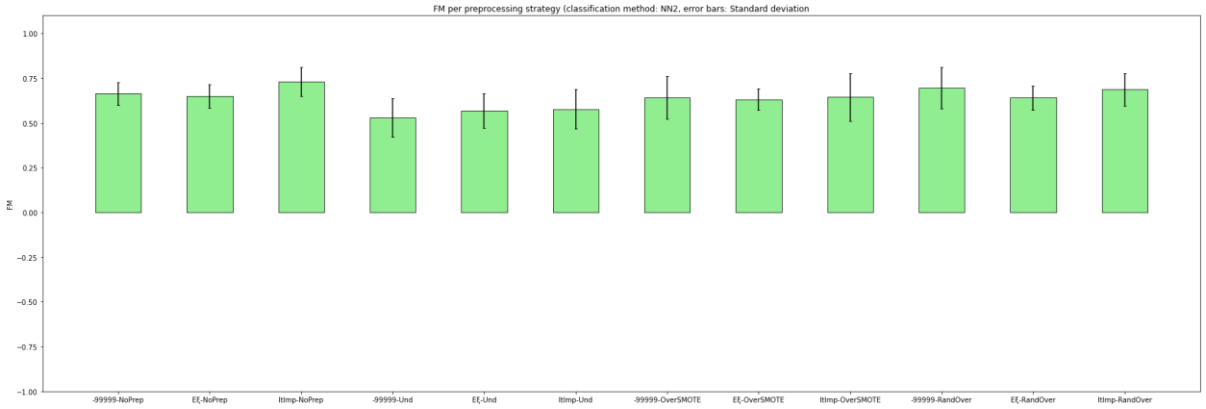

FM per preprocessing strategy (classification method: DT3, error bars: Standard deviation
FM
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

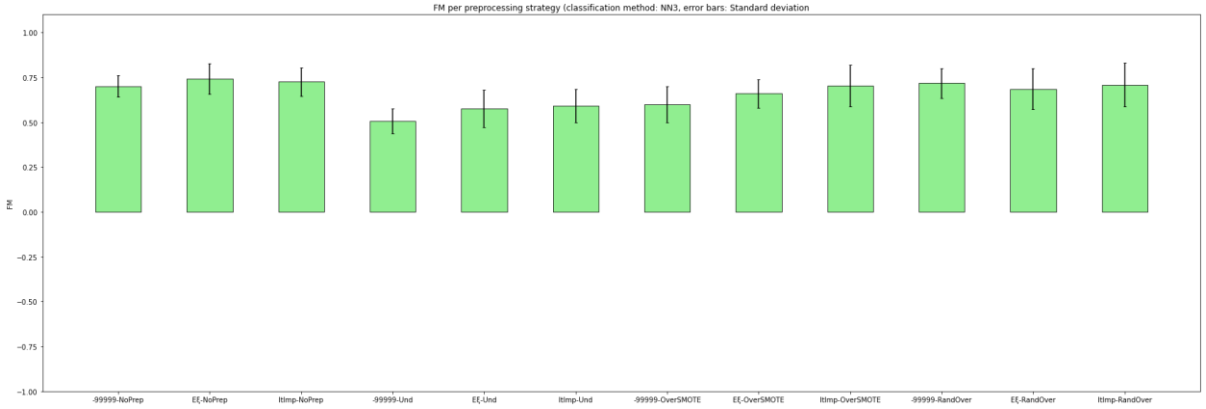

FM per preprocessing strategy (classification method: DT4, error bars: Standard deviation
FM
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

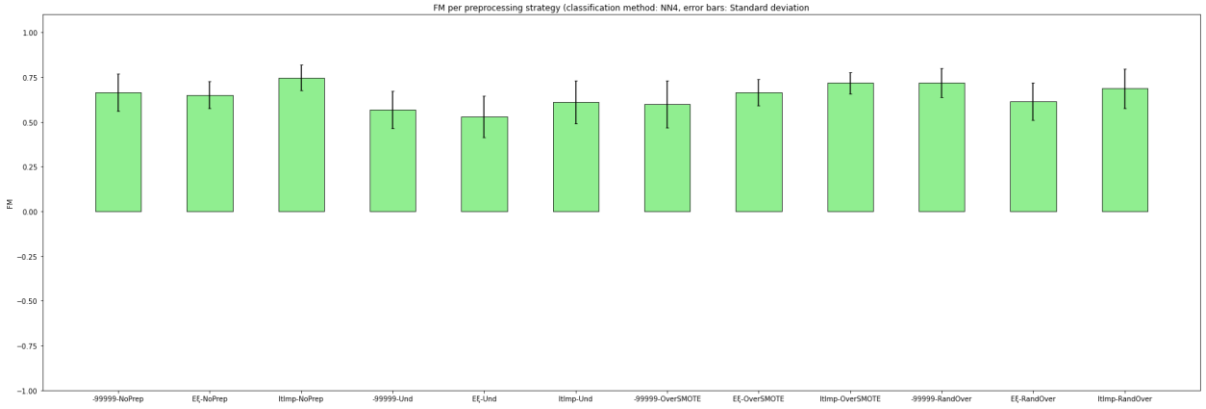

FM per preprocessing strategy (classification method: RF1, error bars: Standard deviation
FM
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

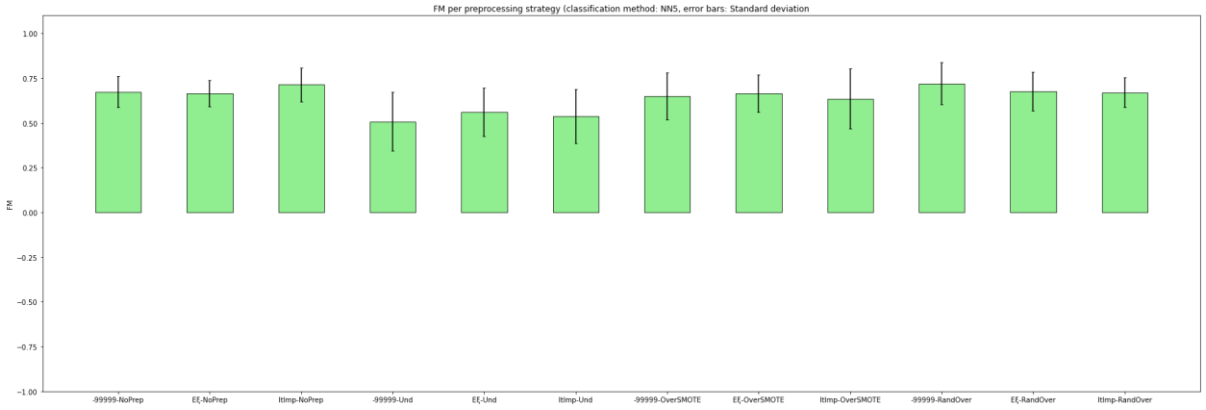

FM per preprocessing strategy (classification method: RF2, error bars: Standard deviation
FM
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

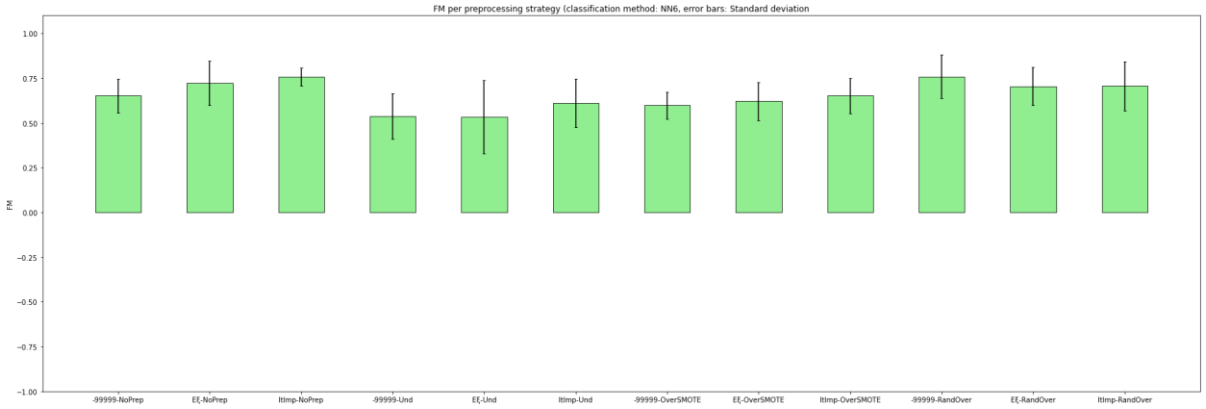

FM per preprocessing strategy (classification method: NN1, error bars: Standard deviation
FM
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

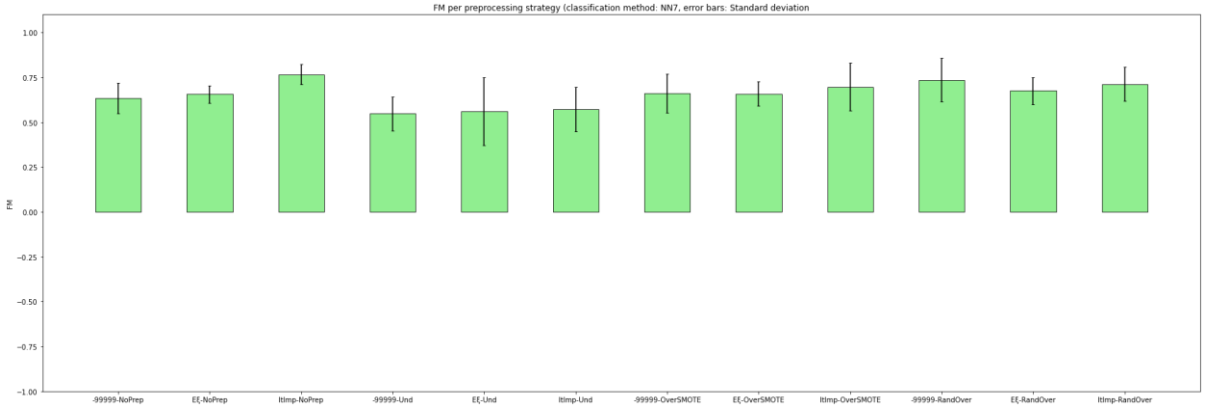

FM per preprocessing strategy (classification method: NN2, error bars: Standard deviation
FM
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

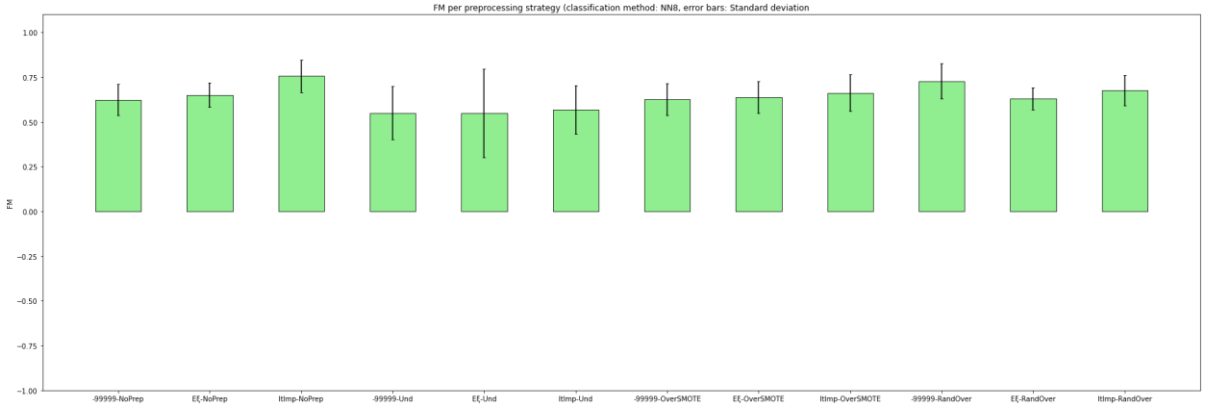

FM per preprocessing strategy (classification method: NN3, error bars: Standard deviation
FM
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

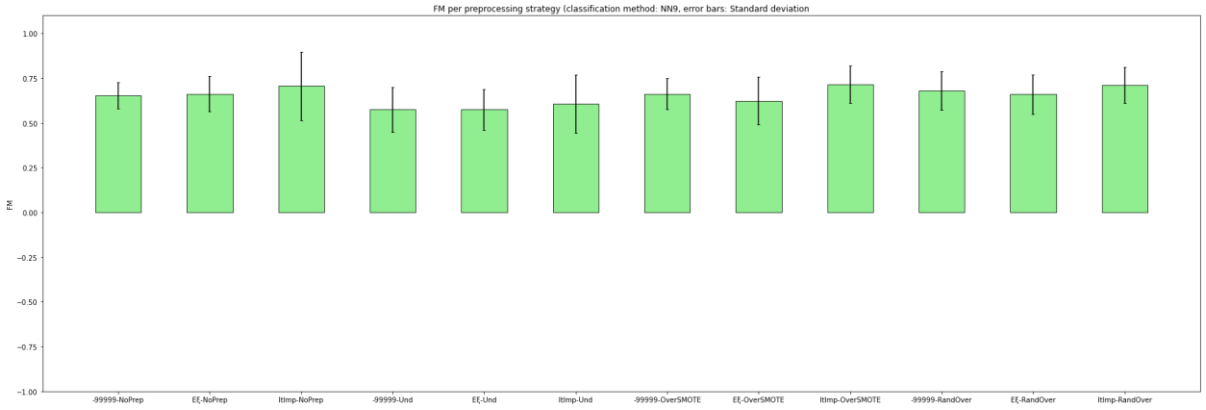

FM per preprocessing strategy (classification method: NN4, error bars: Standard deviation
FM
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

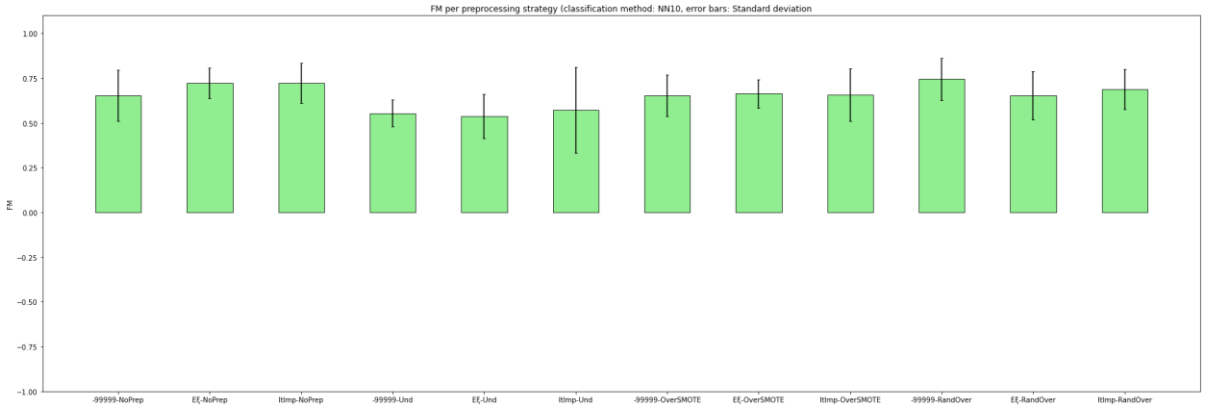

FM per preprocessing strategy (classification method: NN5, error bars: Standard deviation
FM
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

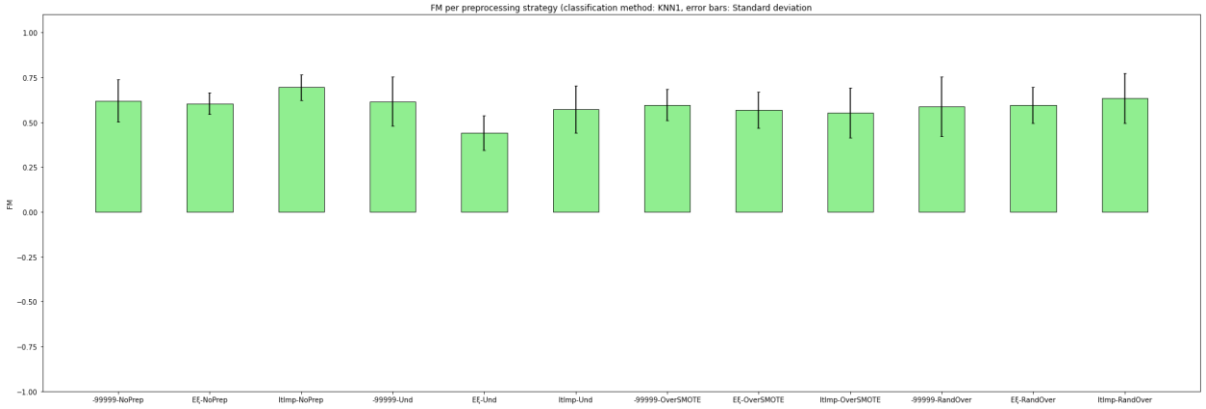

FM per preprocessing strategy (classification method: NN6, error bars: Standard deviation
FM
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

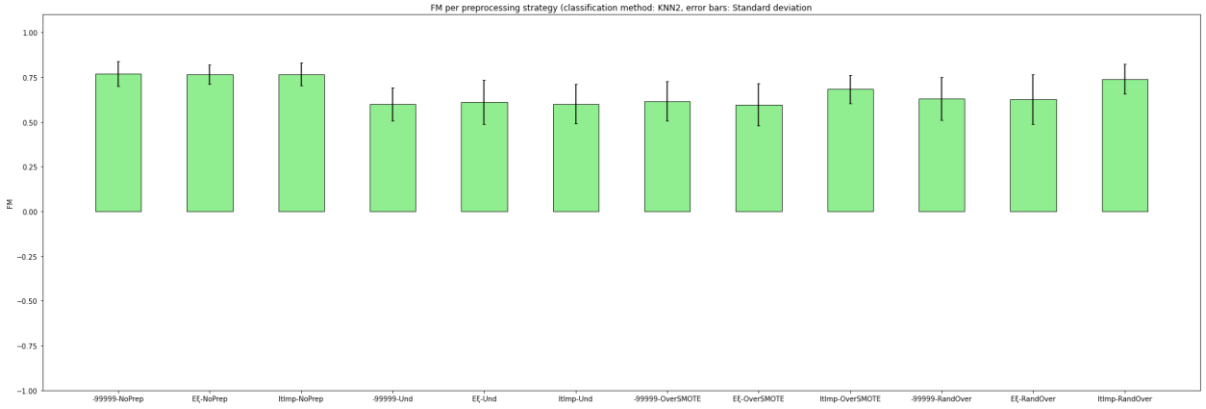

FM per preprocessing strategy (classification method: NN7, error bars: Standard deviation
FM
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

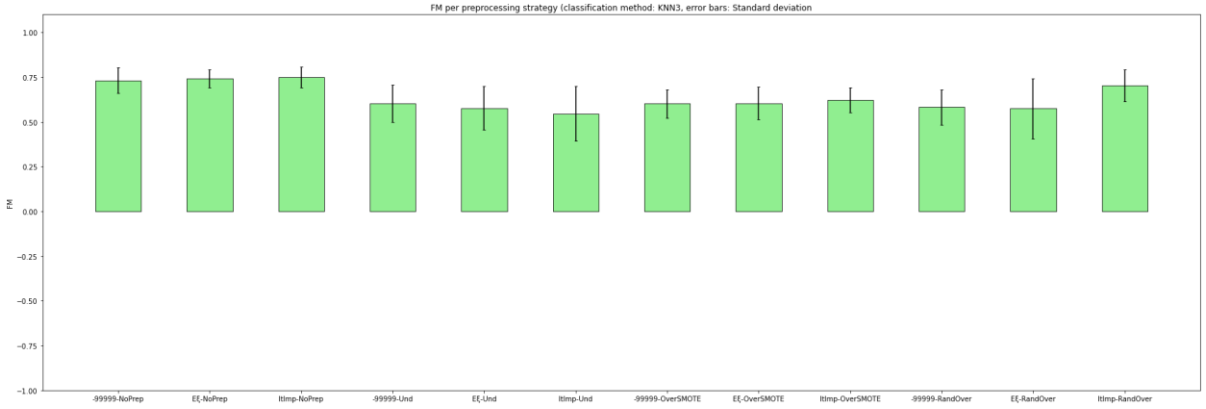

FM per preprocessing strategy (classification method: NN8, error bars: Standard deviation
FM
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

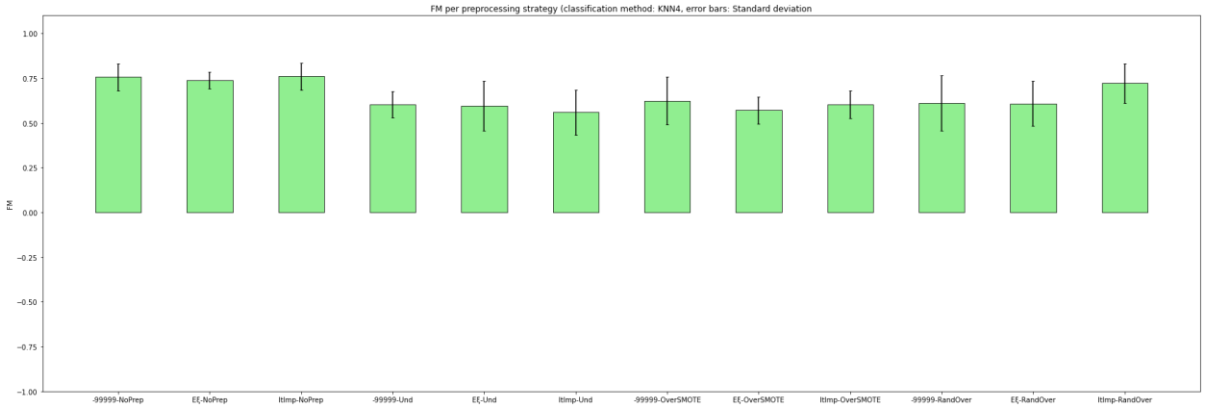

FM per preprocessing strategy (classification method: NN9, error bars: Standard deviation
FM
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

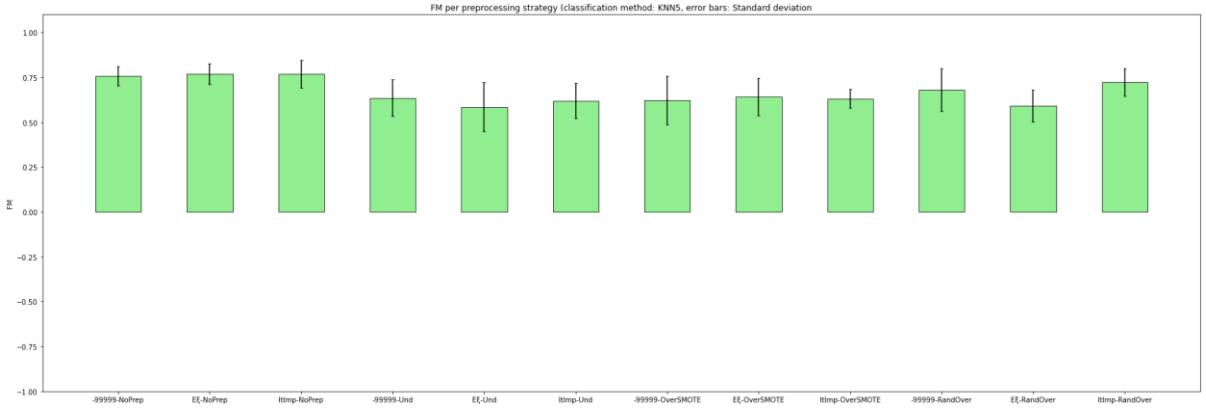

FM per preprocessing strategy (classification method: NN10, error bars: Standard deviation
FM
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

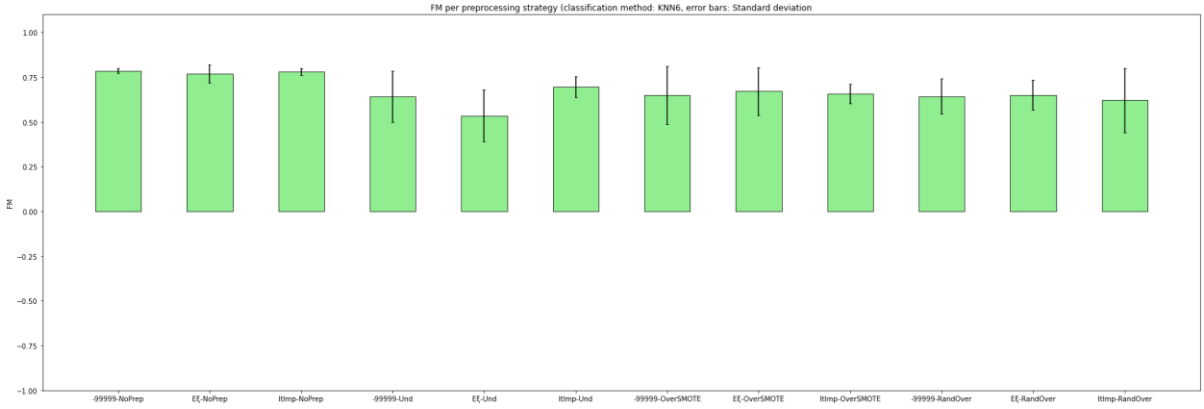

FM per preprocessing strategy (classification method: KNN1, error bars: Standard deviation
FM
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

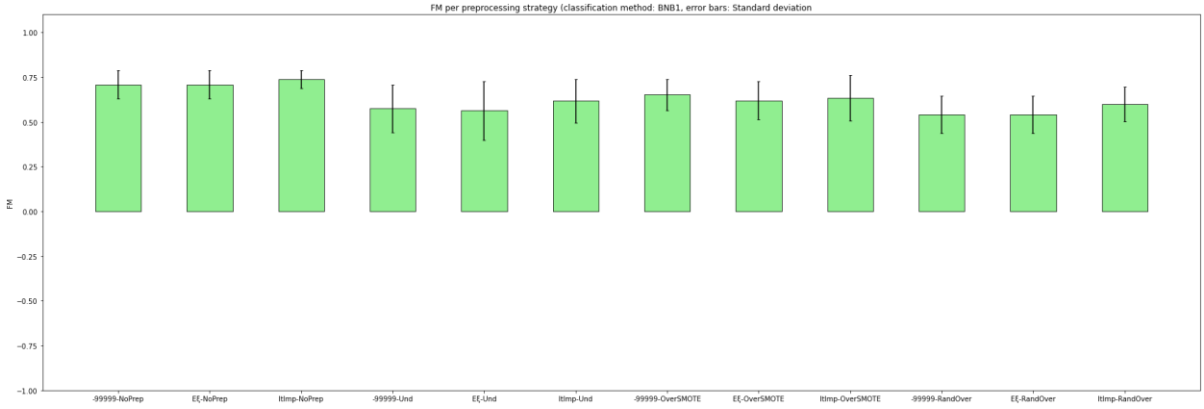

FM per preprocessing strategy (classification method: KNN2, error bars: Standard deviation
FM
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

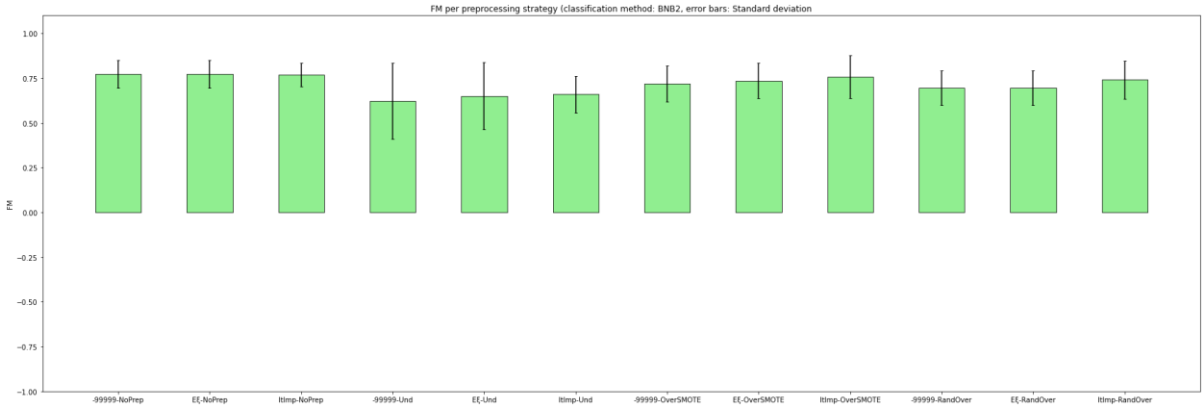

FM per preprocessing strategy (classification method: KNN3, error bars: Standard deviation
FM
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

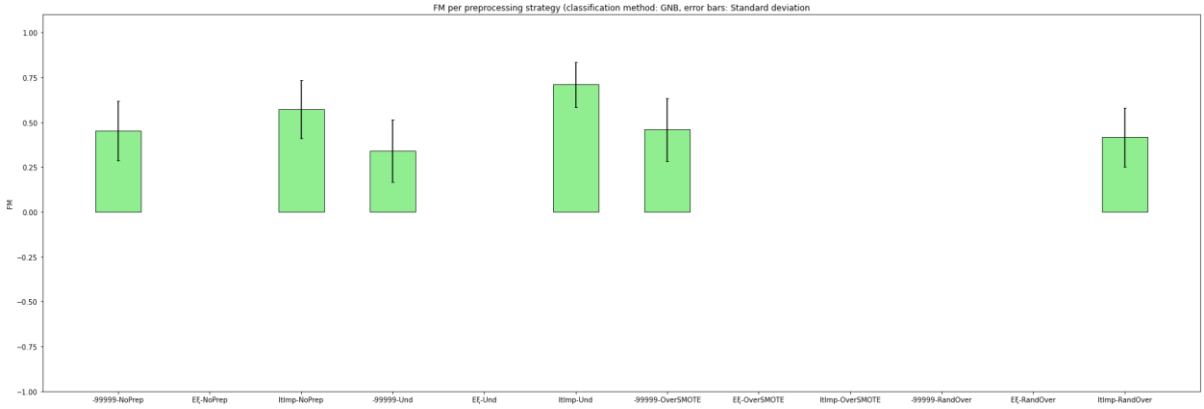

FM per preprocessing strategy (classification method: KNN4, error bars: Standard deviation
FM
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

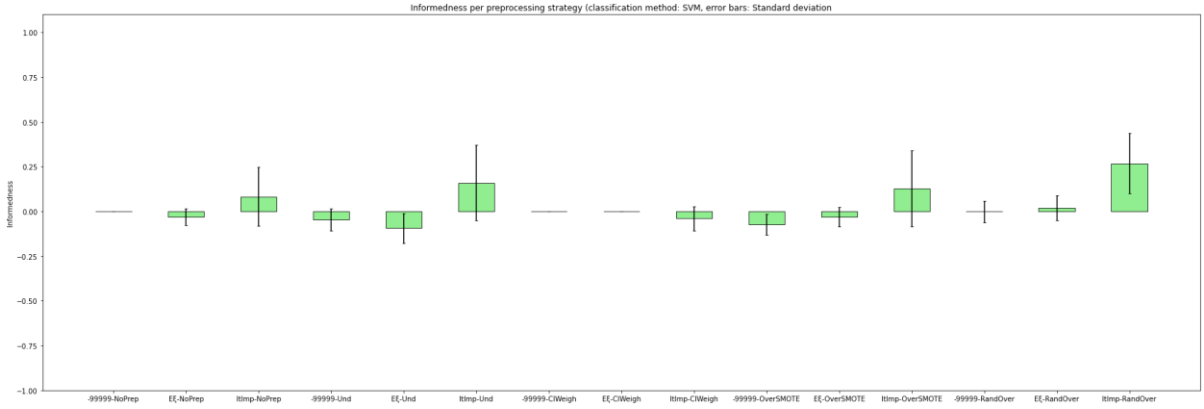

FM per preprocessing strategy (classification method: KNN5, error bars: Standard deviation
FM
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

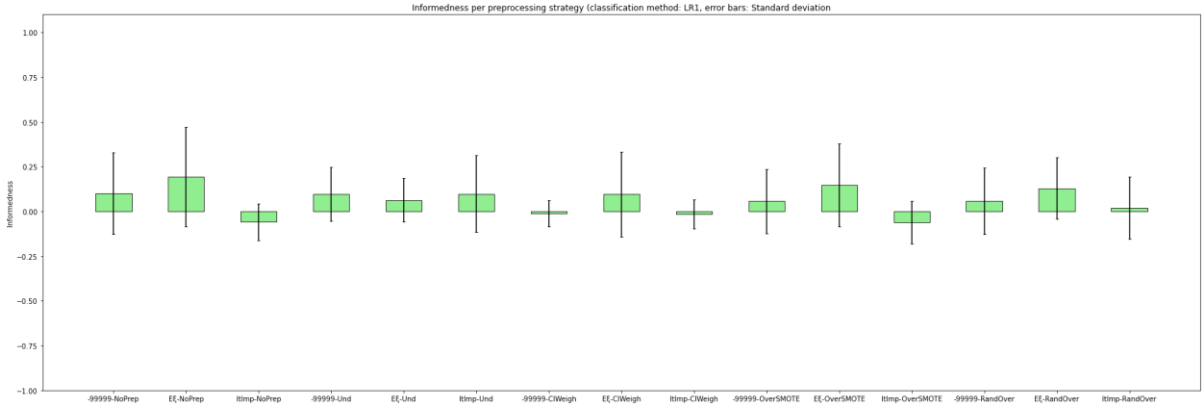

FM per preprocessing strategy (classification method: KNN6, error bars: Standard deviation
FM
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

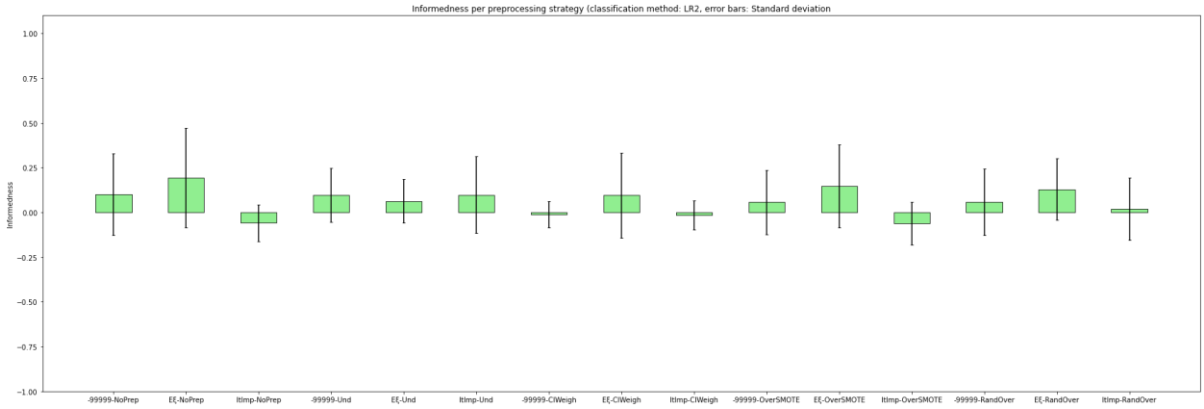

FM per preprocessing strategy (classification method: BNB1, error bars: Standard deviation
FM
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

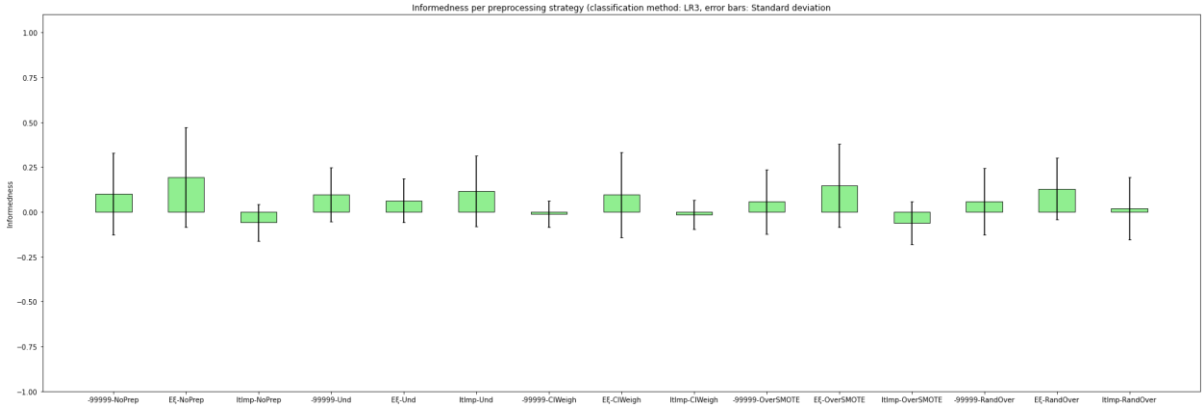

FM per preprocessing strategy (classification method: BNB2, error bars: Standard deviation
FM
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

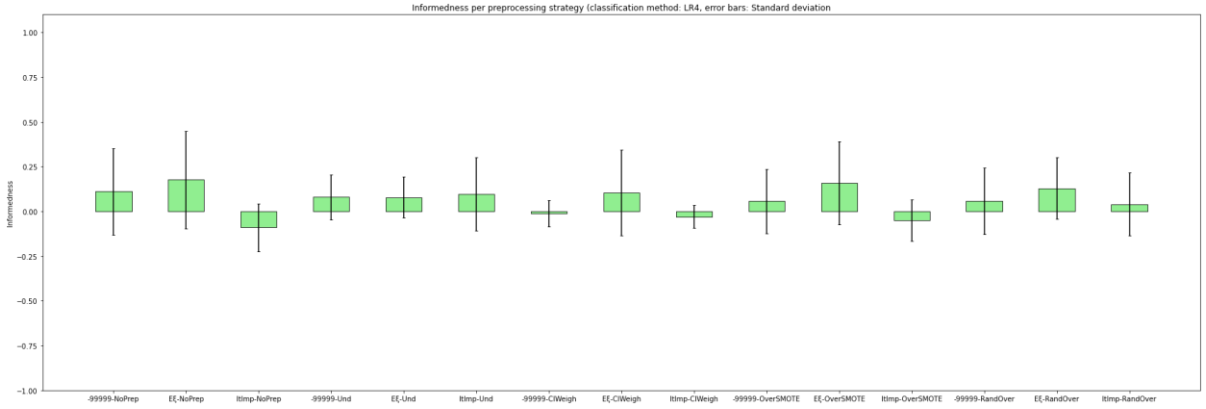

FM per preprocessing strategy (classification method: GNB, error bars: Standard deviation
FM
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

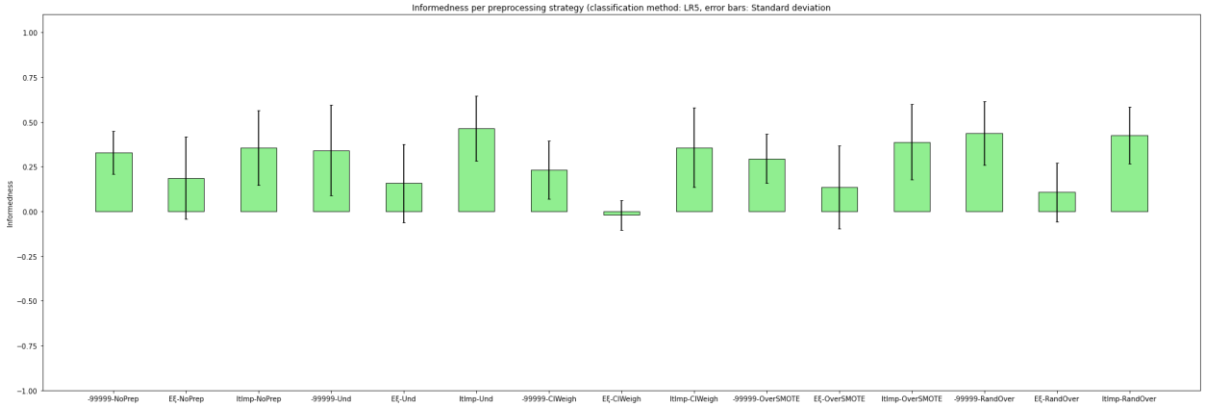

Informedness per preprocessing strategy (classification method: SVM, error bars: Standard deviation
Informedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

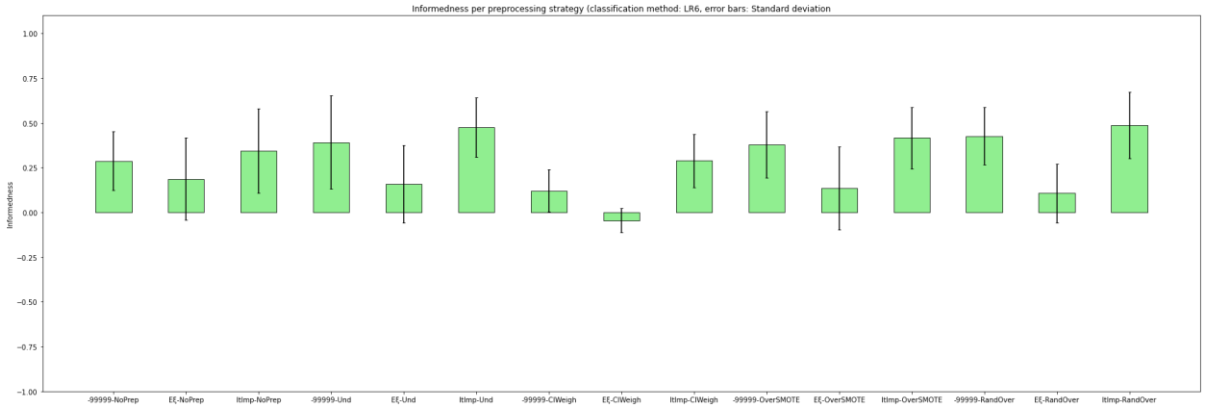

Informedness per preprocessing strategy (classification method: LR1, error bars: Standard deviation
Informedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

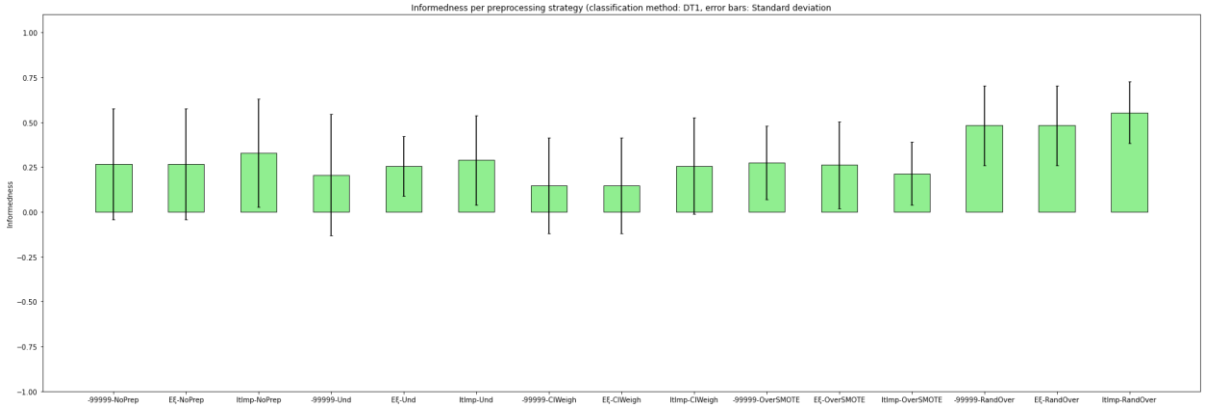

Informedness per preprocessing strategy (classification method: LR2, error bars: Standard deviation
Informedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

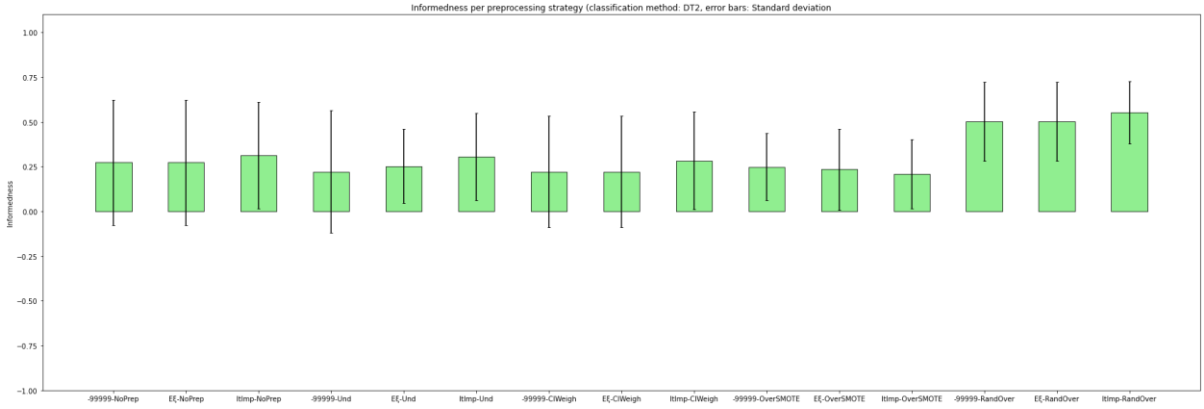

Informedness per preprocessing strategy (classification method: LR3, error bars: Standard deviation
Informedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

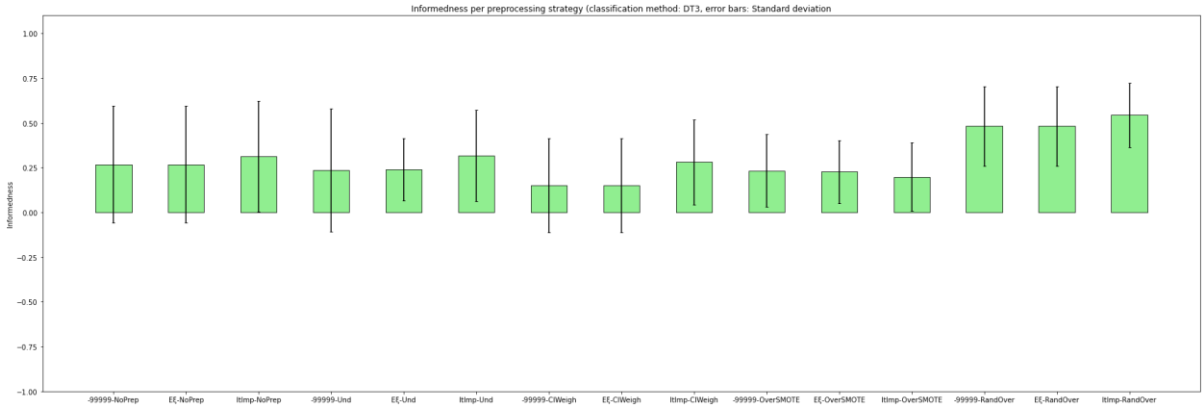

Informedness per preprocessing strategy (classification method: LR4, error bars: Standard deviation
Informedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

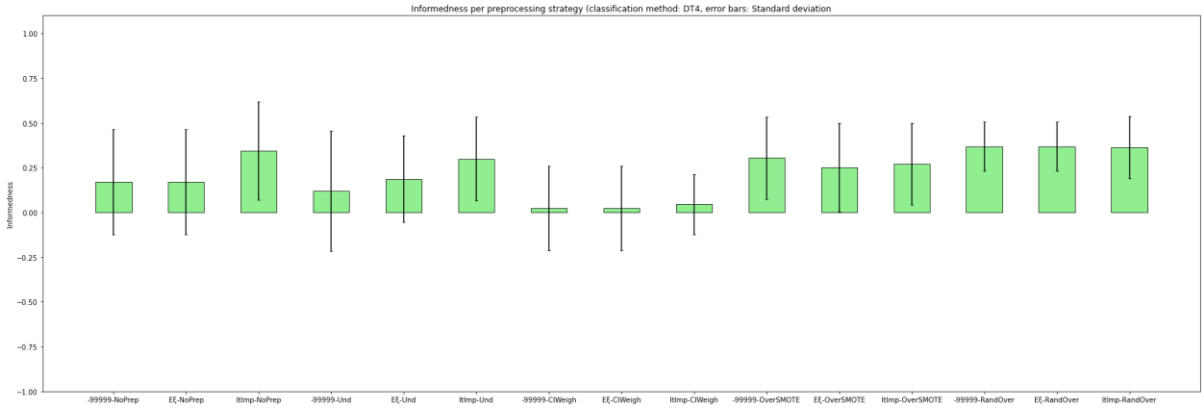

Informedness per preprocessing strategy (classification method: LR5, error bars: Standard deviation
Informedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

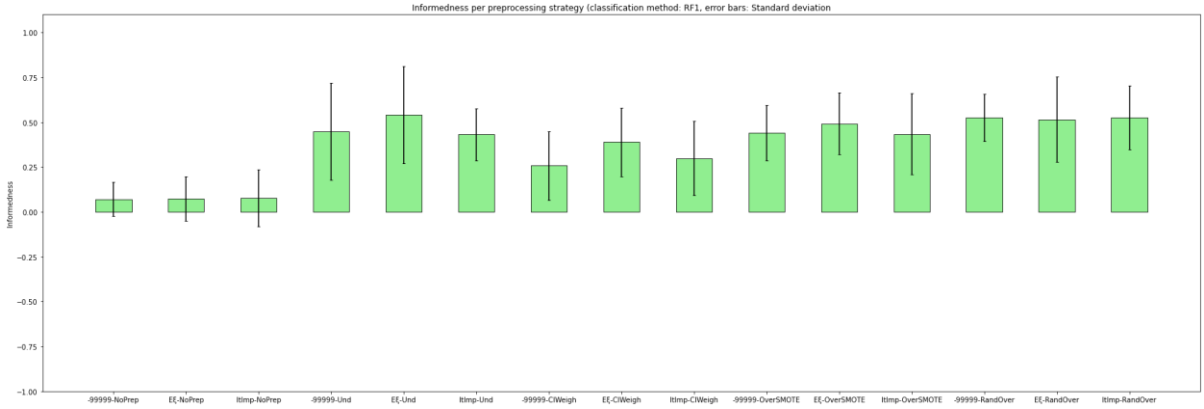

Informedness per preprocessing strategy (classification method: LR6, error bars: Standard deviation
Informedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

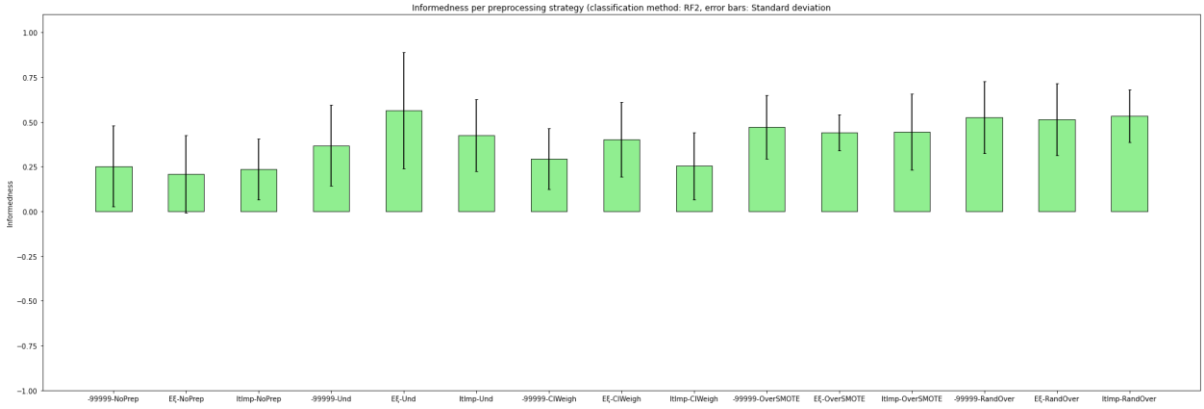

Informedness per preprocessing strategy (classification method: DT1, error bars: Standard deviation
Informedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

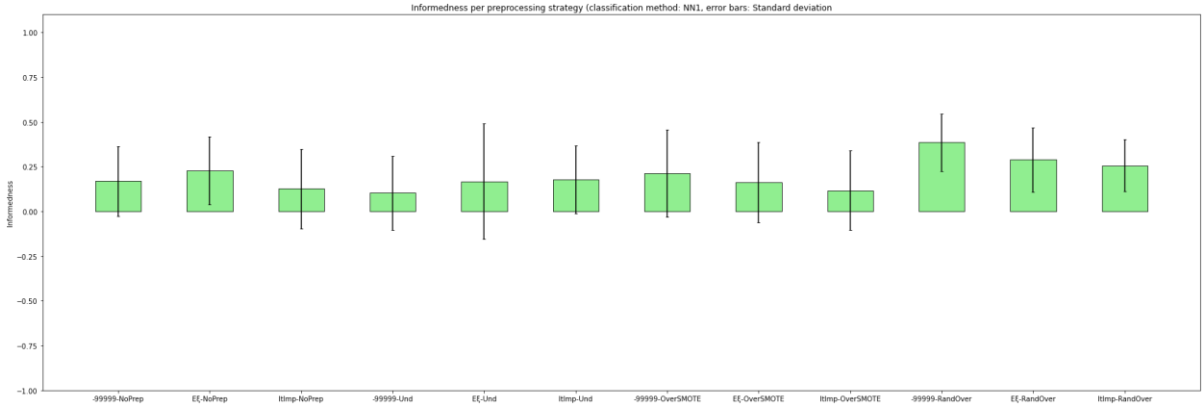

Informedness per preprocessing strategy (classification method: DT2, error bars: Standard deviation
Informedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

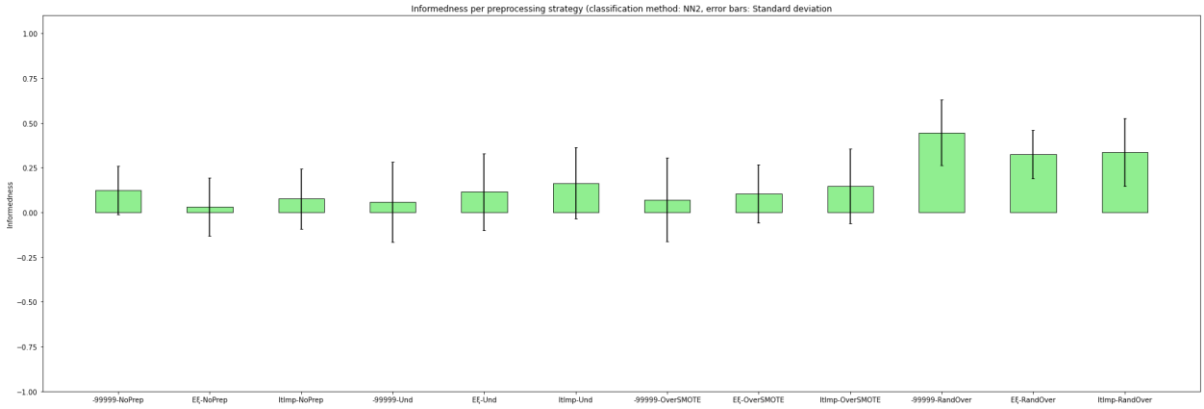

Informedness per preprocessing strategy (classification method: DT3, error bars: Standard deviation
Informedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

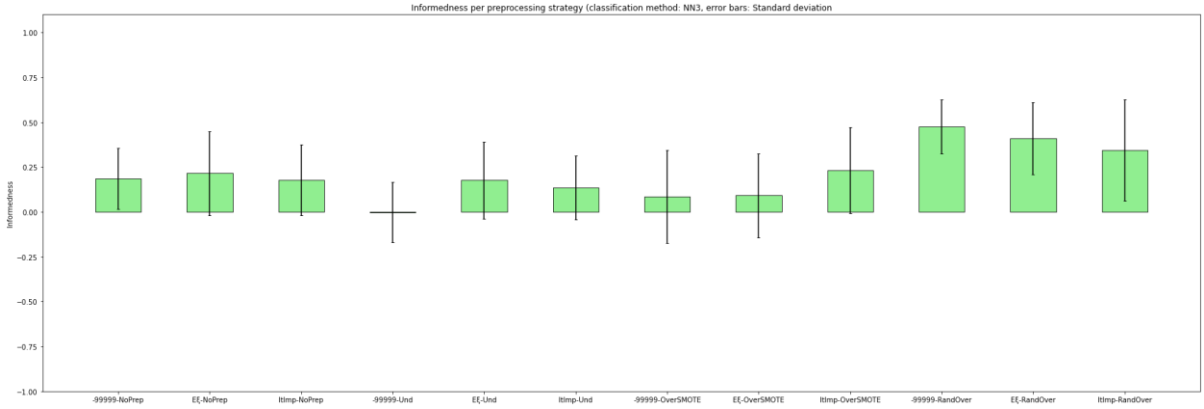

Informedness per preprocessing strategy (classification method: DT4, error bars: Standard deviation
Informedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

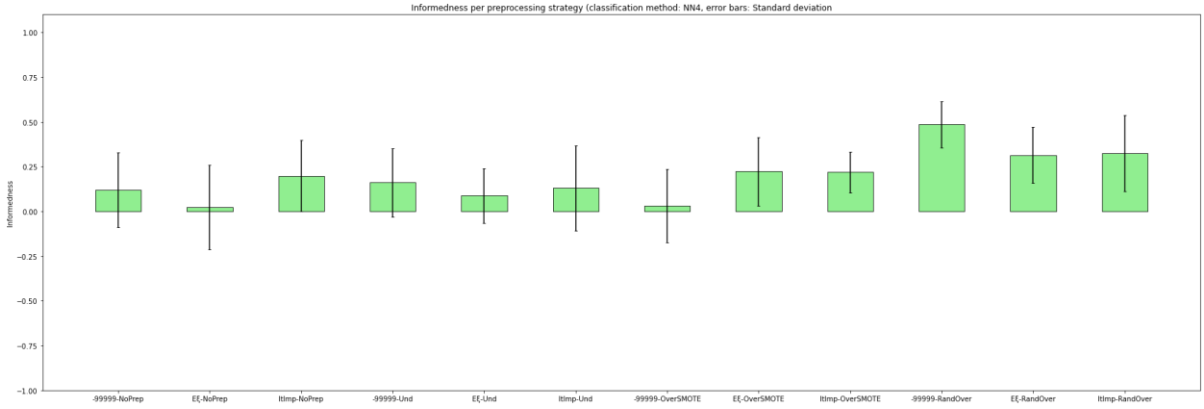

Informedness per preprocessing strategy (classification method: RF1, error bars: Standard deviation
Informedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

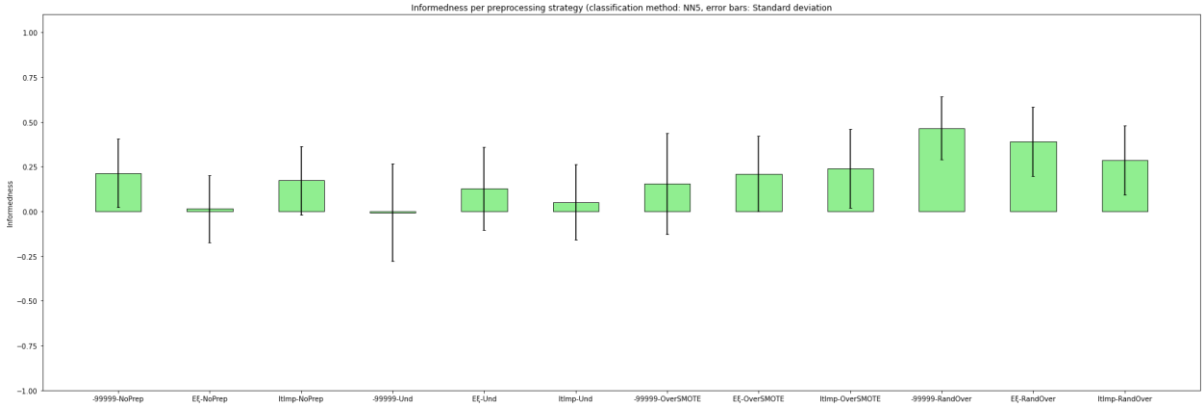

Informedness per preprocessing strategy (classification method: RF2, error bars: Standard deviation
Informedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

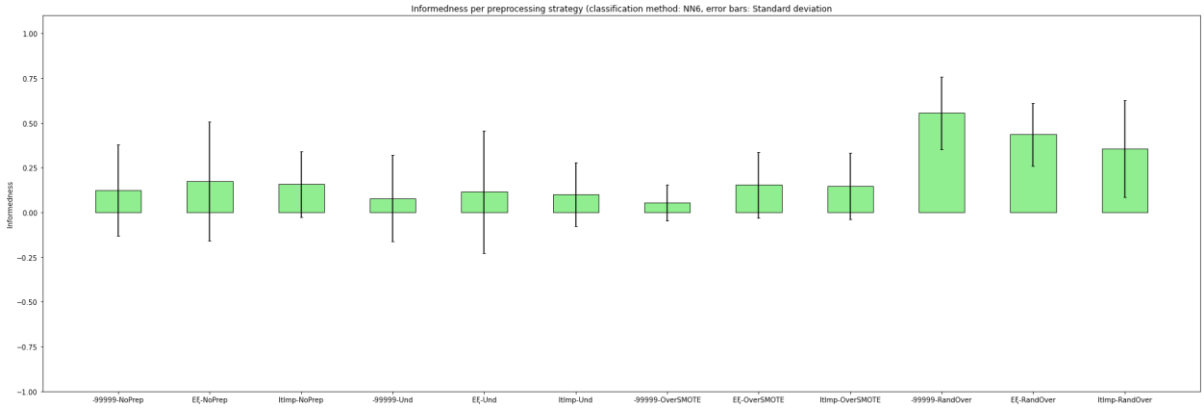

Informedness per preprocessing strategy (classification method: NN1, error bars: Standard deviation
Informedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

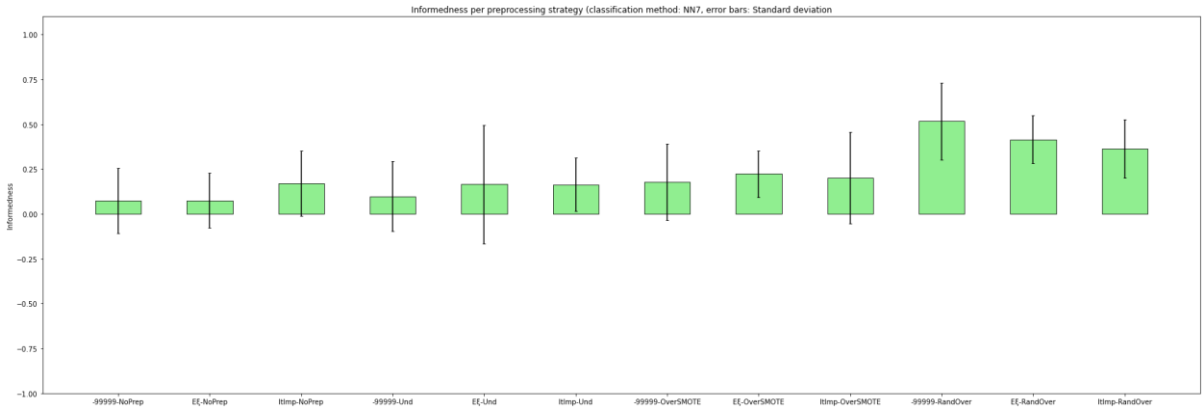

Informedness per preprocessing strategy (classification method: NN2, error bars: Standard deviation
Informedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

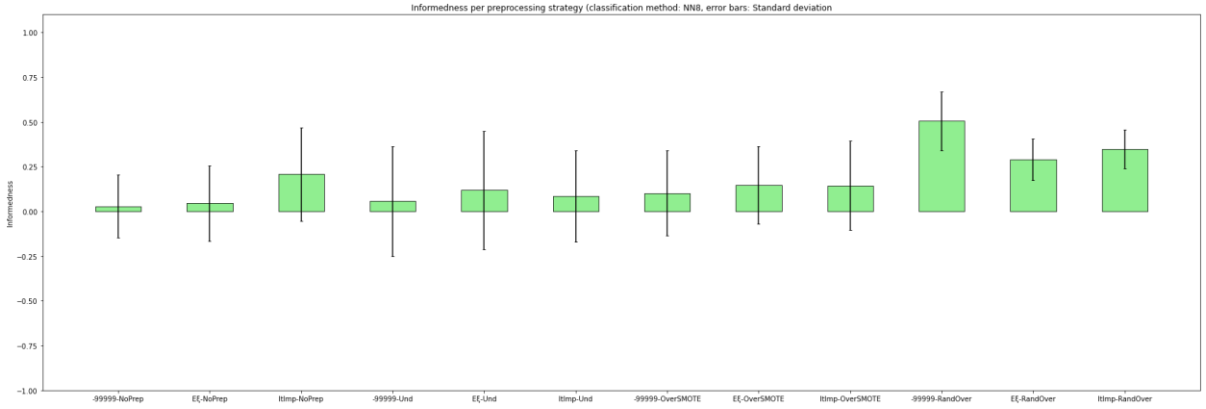

Informedness per preprocessing strategy (classification method: NN3, error bars: Standard deviation
Informedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

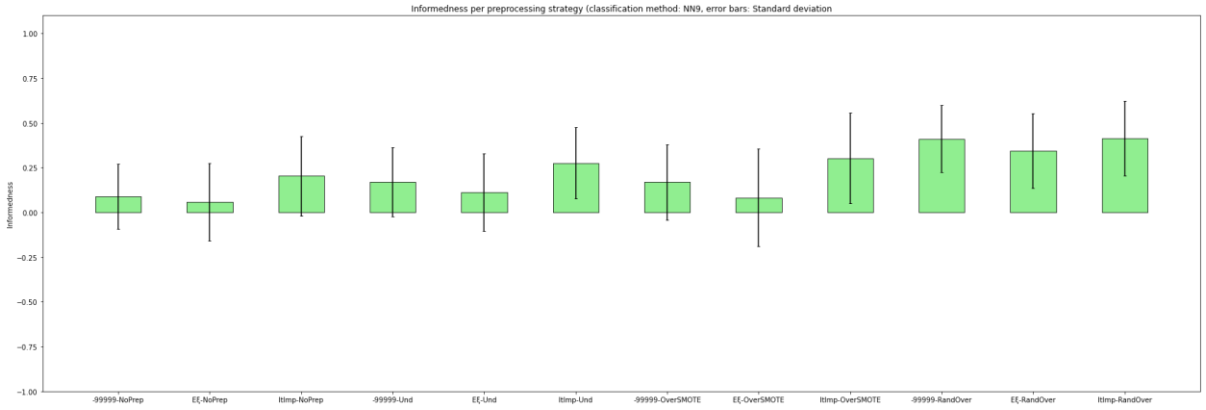

Informedness per preprocessing strategy (classification method: NN4, error bars: Standard deviation
Informedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

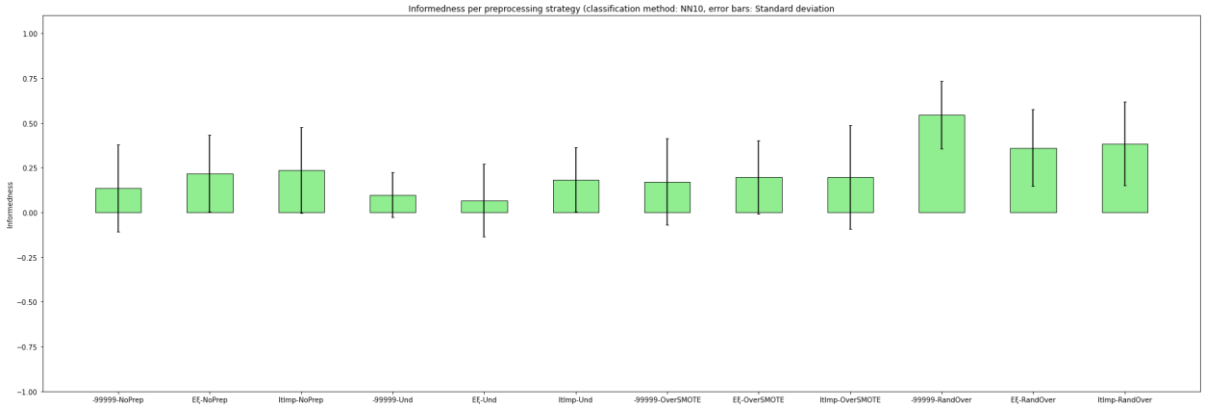

Informedness per preprocessing strategy (classification method: NN5, error bars: Standard deviation
Informedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

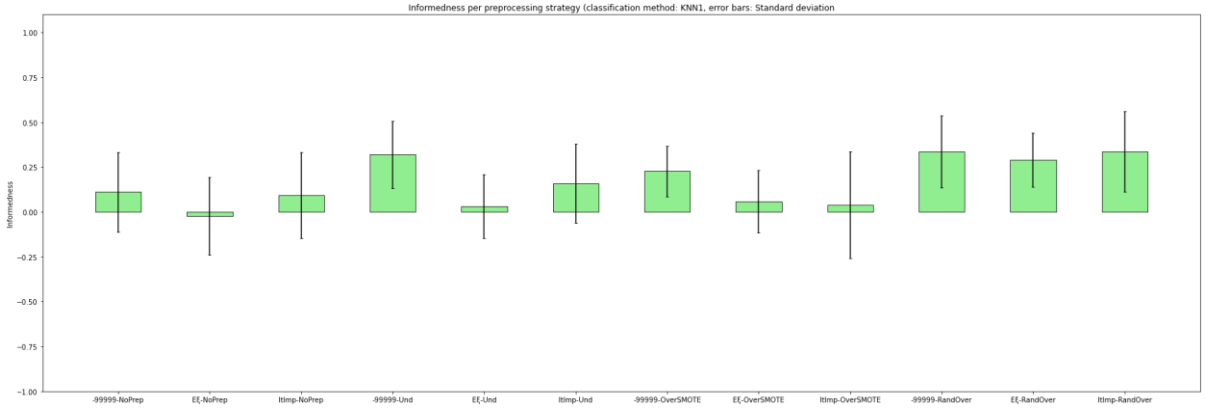

Informedness per preprocessing strategy (classification method: NN6, error bars: Standard deviation
Informedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

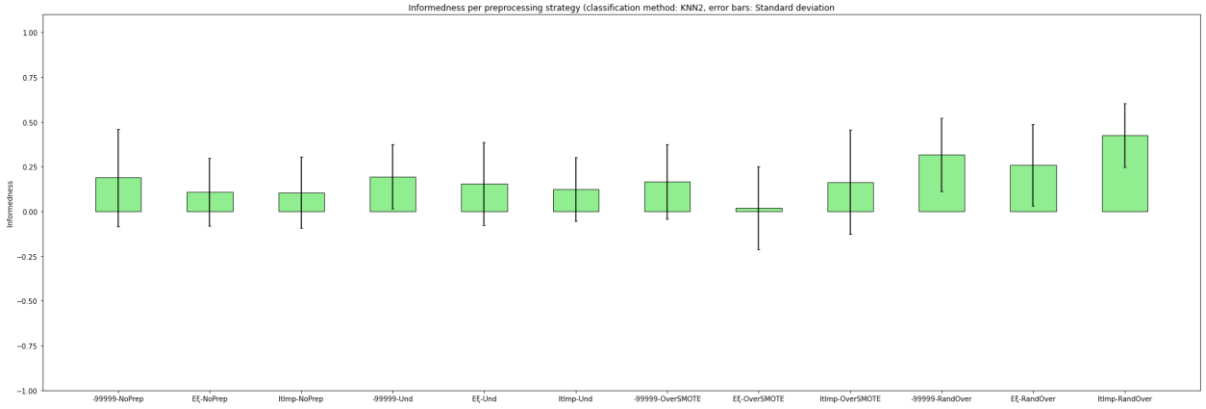

Informedness per preprocessing strategy (classification method: NN7, error bars: Standard deviation
Informedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

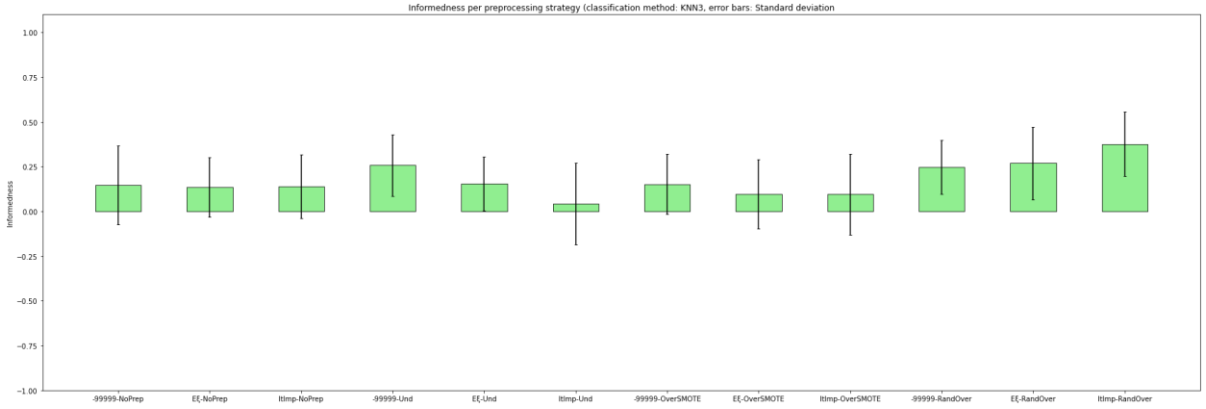

Informedness per preprocessing strategy (classification method: NN8, error bars: Standard deviation
Informedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

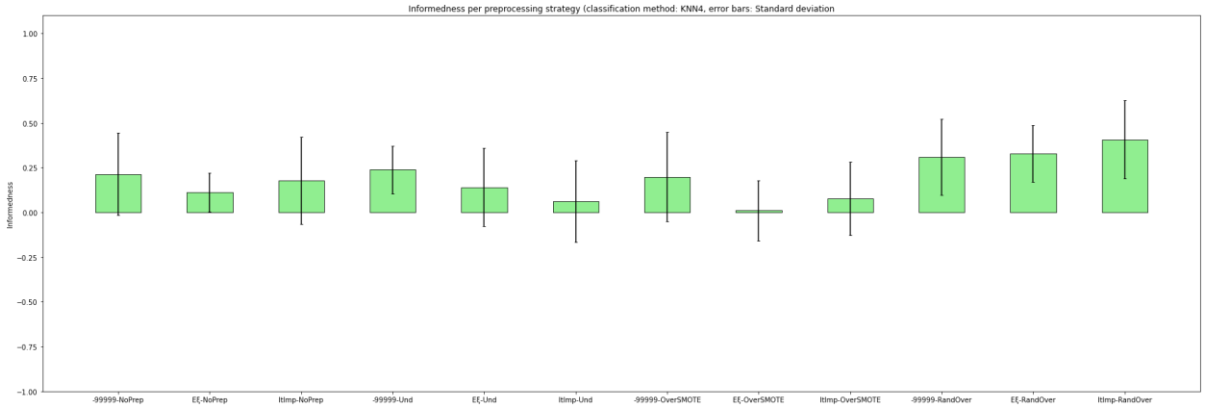

Informedness per preprocessing strategy (classification method: NN9, error bars: Standard deviation
Informedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

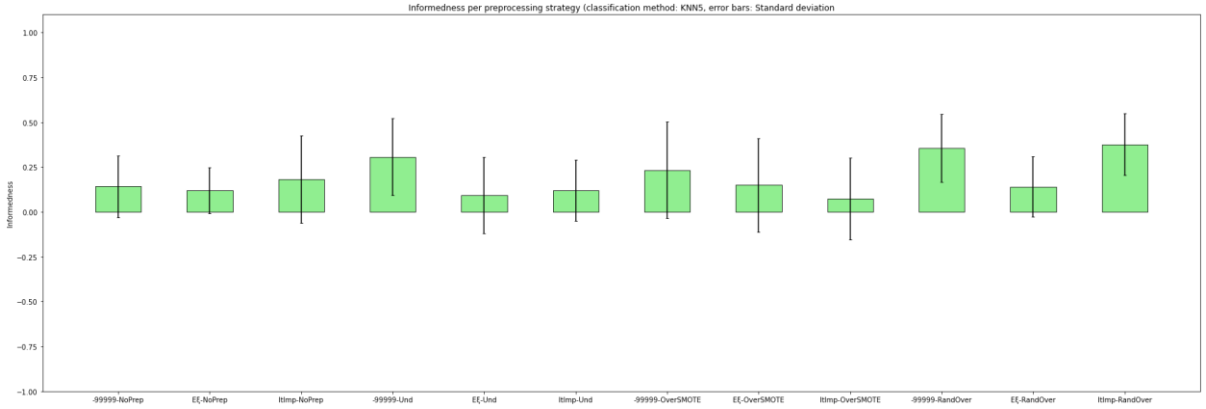

Informedness per preprocessing strategy (classification method: NN10, error bars: Standard deviation
Informedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

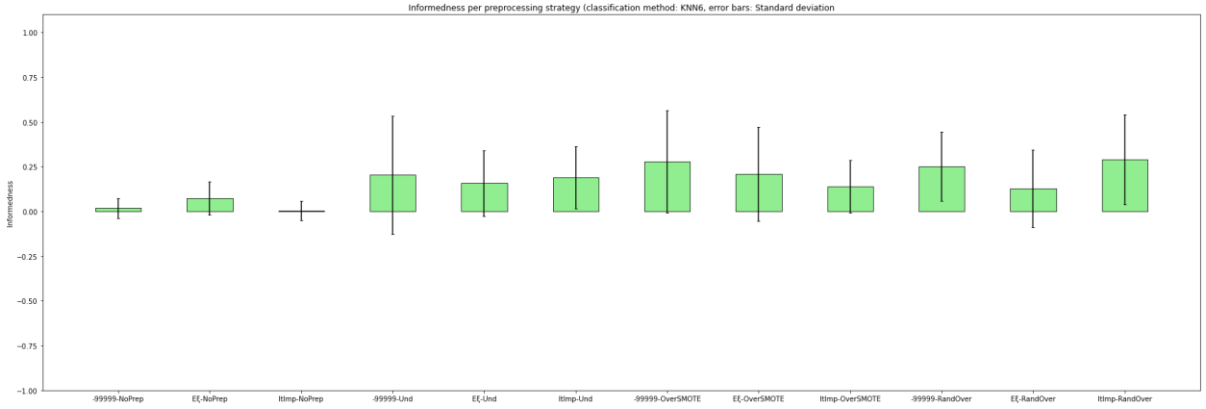

Informedness per preprocessing strategy (classification method: KNN1, error bars: Standard deviation
Informedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

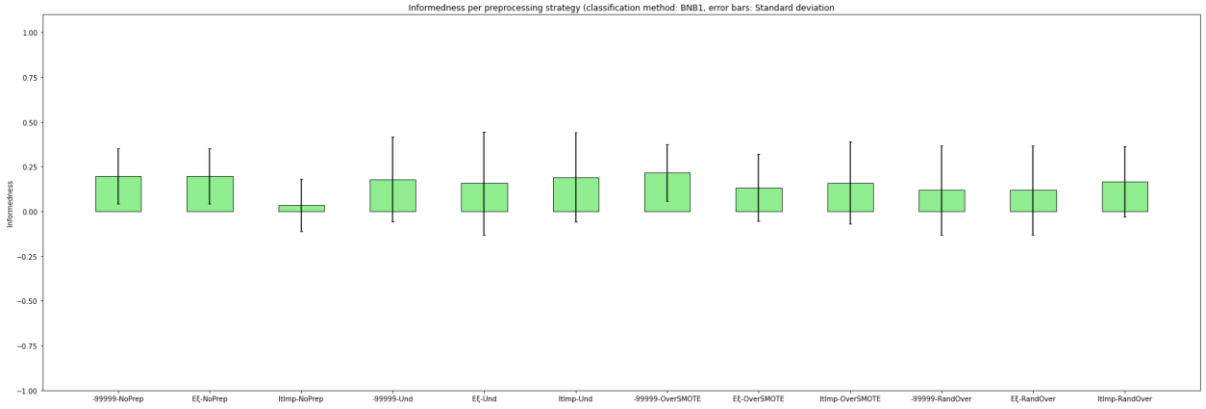

Informedness per preprocessing strategy (classification method: KNN2, error bars: Standard deviation
Informedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

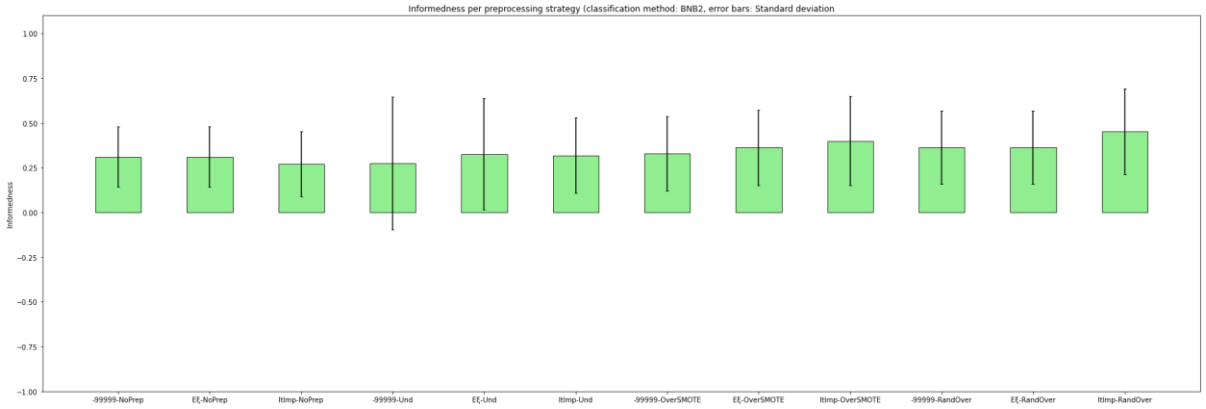

Informedness per preprocessing strategy (classification method: KNN3, error bars: Standard deviation
Informedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

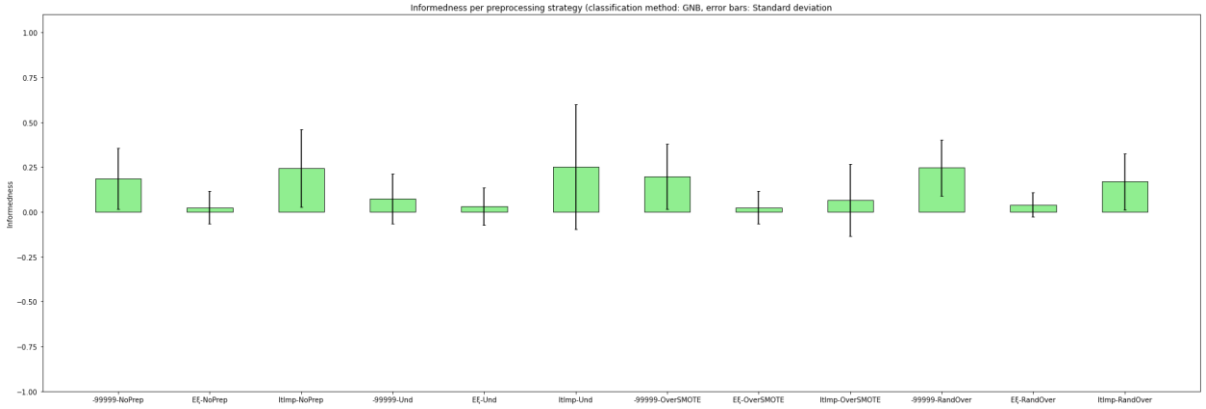

Informedness per preprocessing strategy (classification method: KNN4, error bars: Standard deviation
Informedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

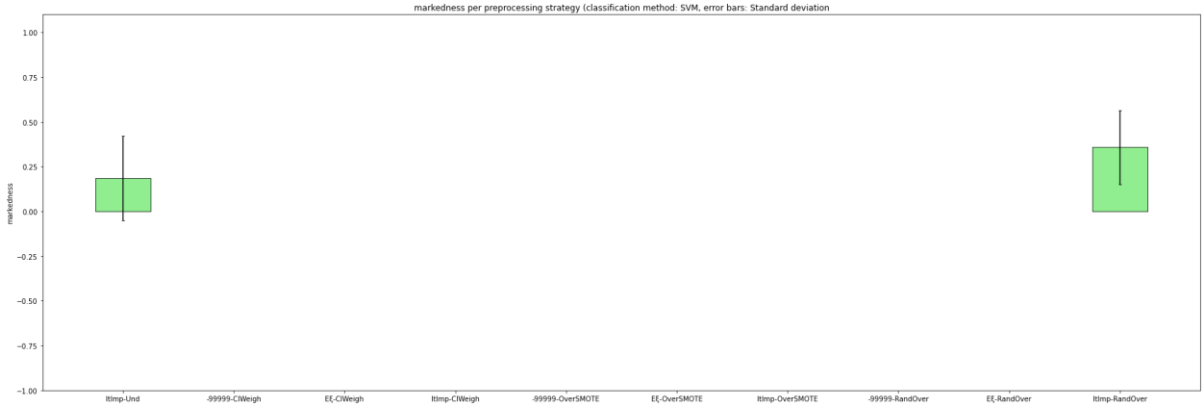

Informedness per preprocessing strategy (classification method: KNN5, error bars: Standard deviation
Informedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

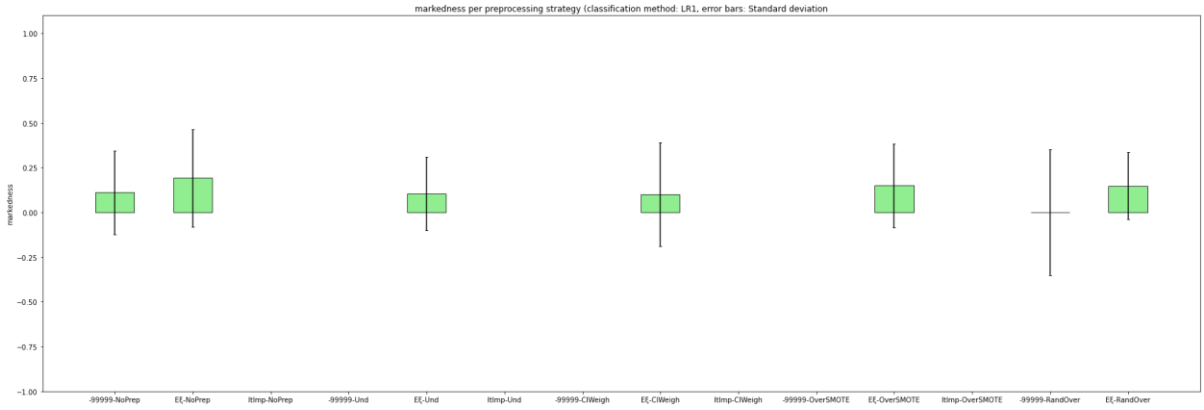

Informedness per preprocessing strategy (classification method: KNN6, error bars: Standard deviation
Informedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

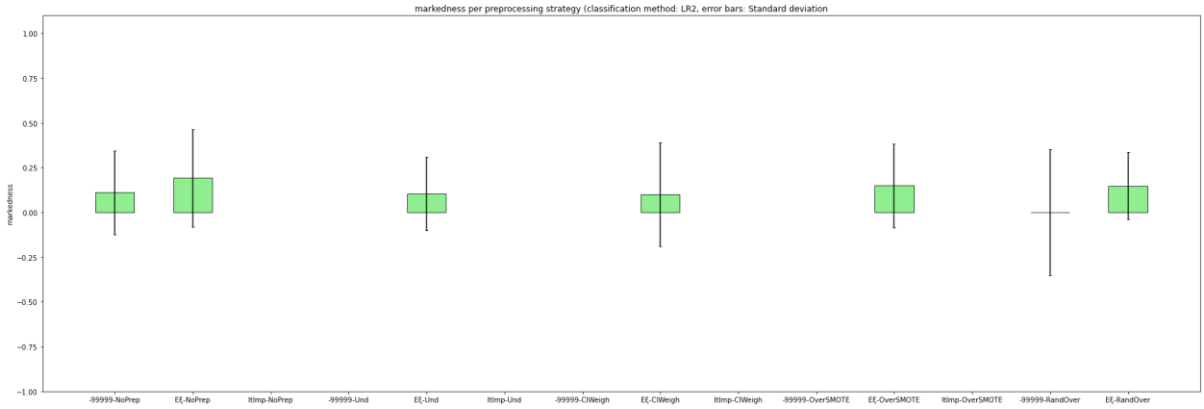

Informedness per preprocessing strategy (classification method: BNB1, error bars: Standard deviation
Informedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

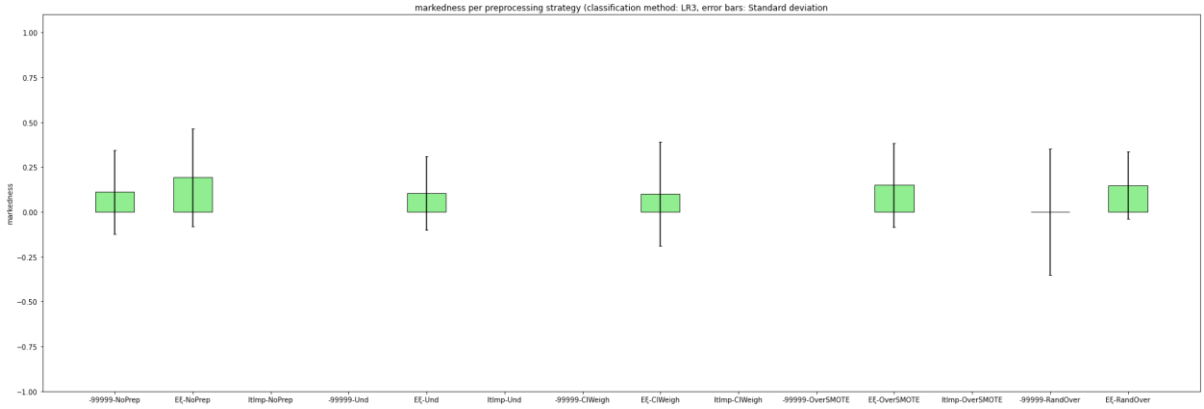

Informedness per preprocessing strategy (classification method: BNB2, error bars: Standard deviation
Informedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

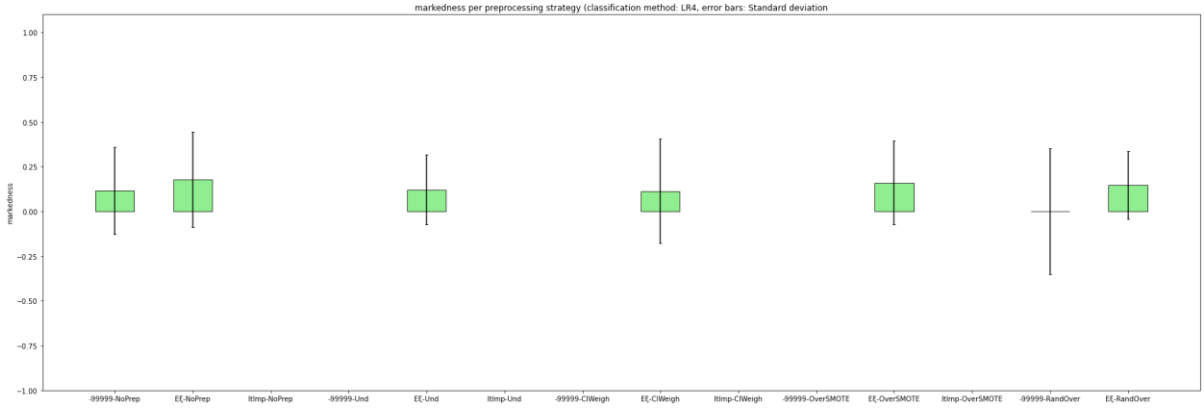

Informedness per preprocessing strategy (classification method: GNB, error bars: Standard deviation
Informedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

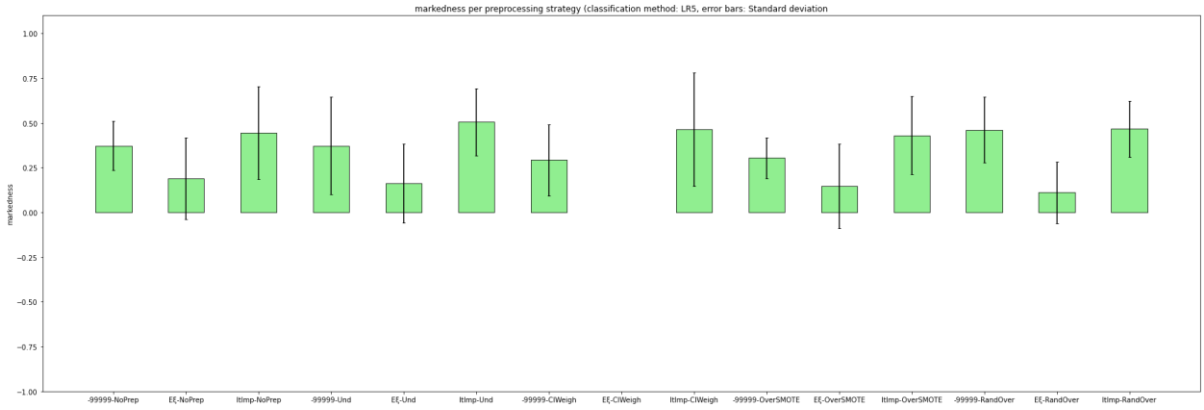

markedness per preprocessing strategy (classification method: SVM, error bars: Standard deviation
markedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
ItImp-Und
-99999-ClWeigh
EE-ClWeigh
ItImp-ClWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

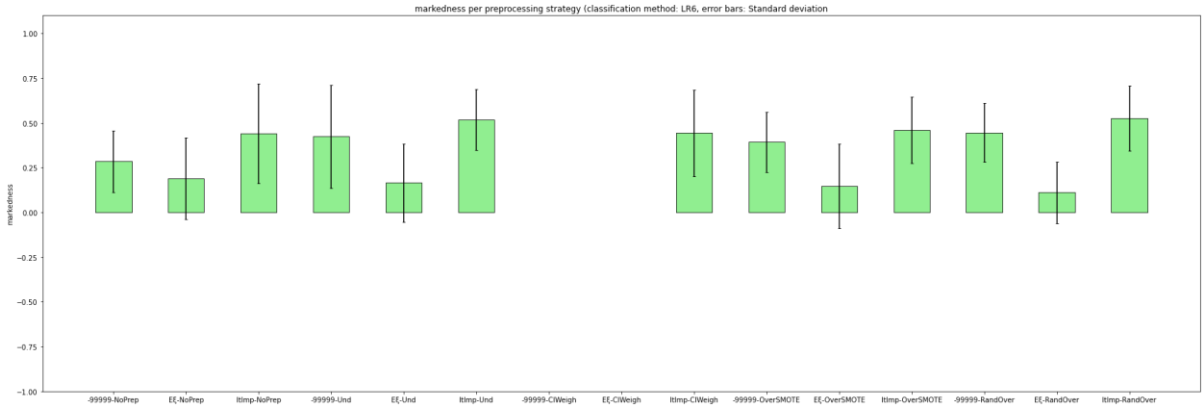

markedness per preprocessing strategy (classification method: LR1, error bars: Standard deviation
markedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-ClWeigh
EE-ClWeigh
ItImp-ClWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver

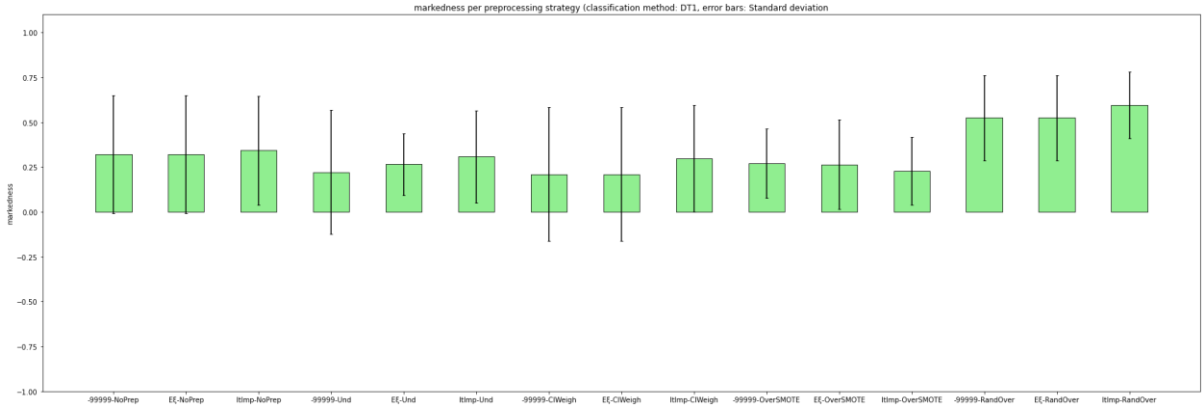

markedness per preprocessing strategy (classification method: LR2, error bars: Standard deviation
markedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-ClWeigh
EE-ClWeigh
ItImp-ClWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver

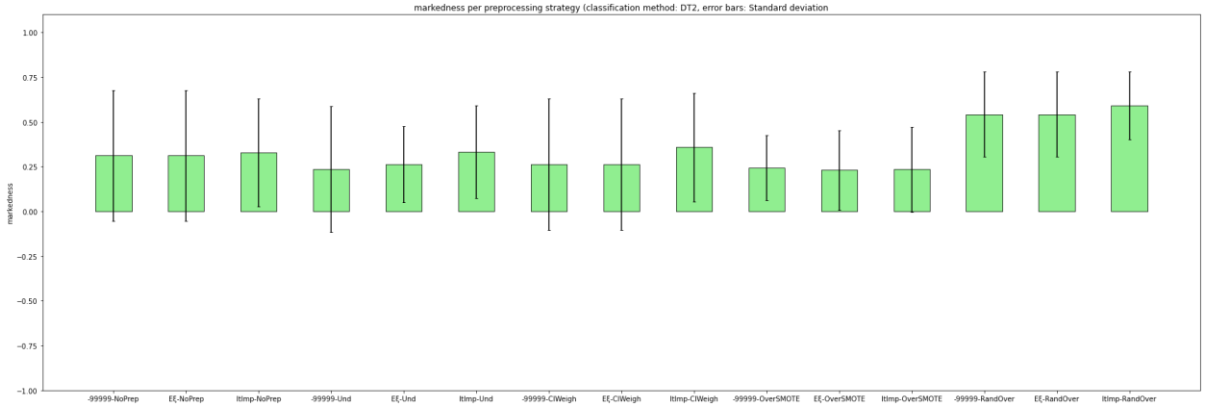

markedness per preprocessing strategy (classification method: LR3, error bars: Standard deviation
markedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-ClWeigh
EE-ClWeigh
ItImp-ClWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver

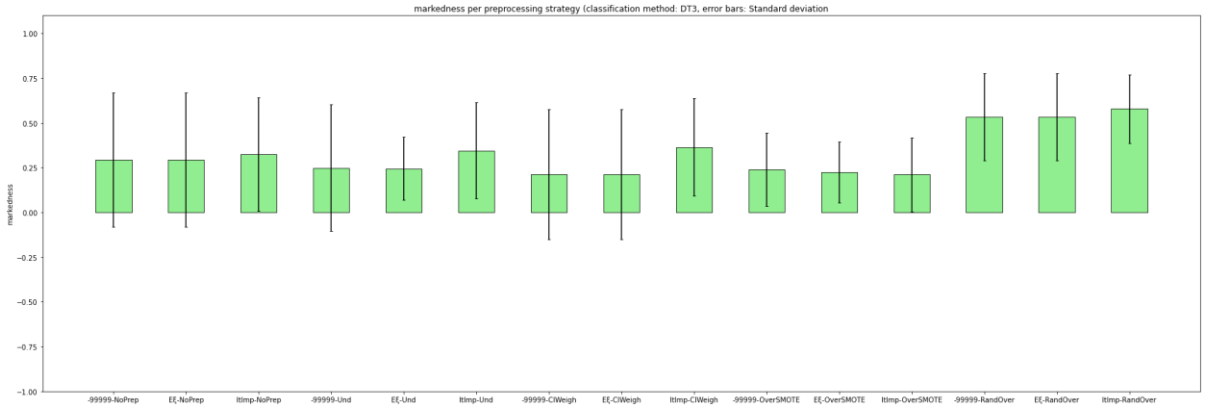

markedness per preprocessing strategy (classification method: LR4, error bars: Standard deviation
markedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver

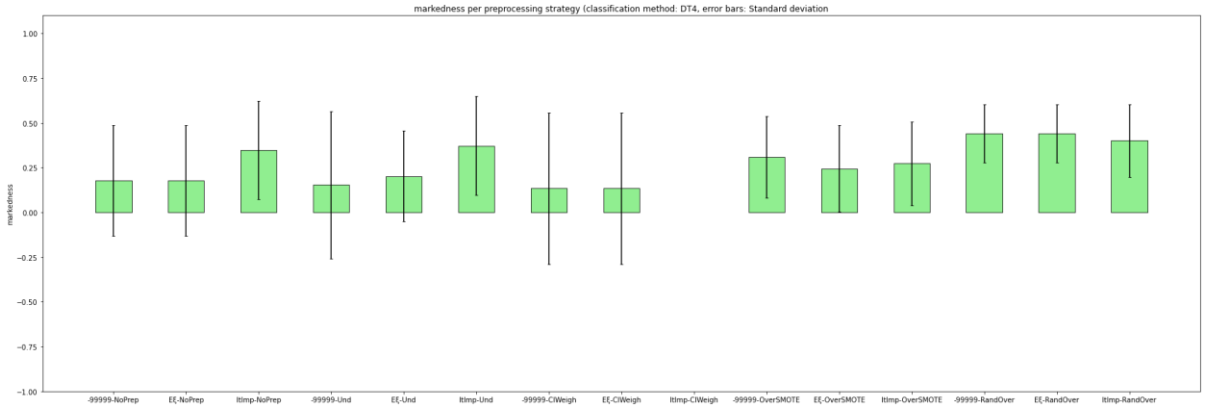

markedness per preprocessing strategy (classification method: LR5, error bars: Standard deviation
markedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

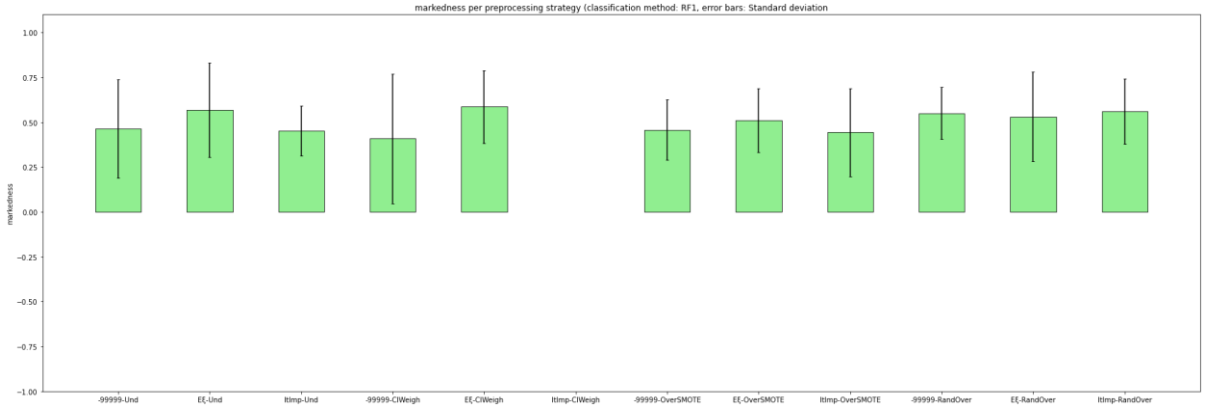

markedness per preprocessing strategy (classification method: LR6, error bars: Standard deviation
markedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

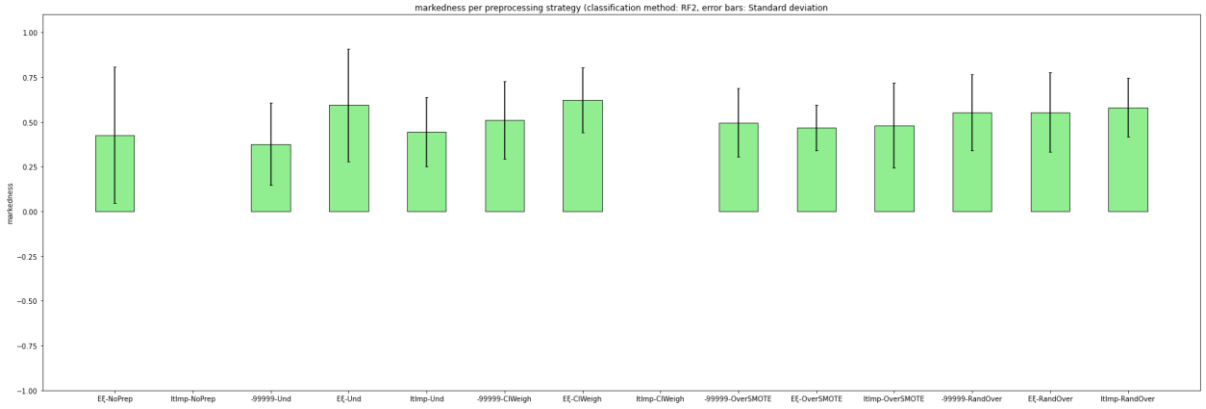

markedness per preprocessing strategy (classification method: DT1, error bars: Standard deviation
markedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

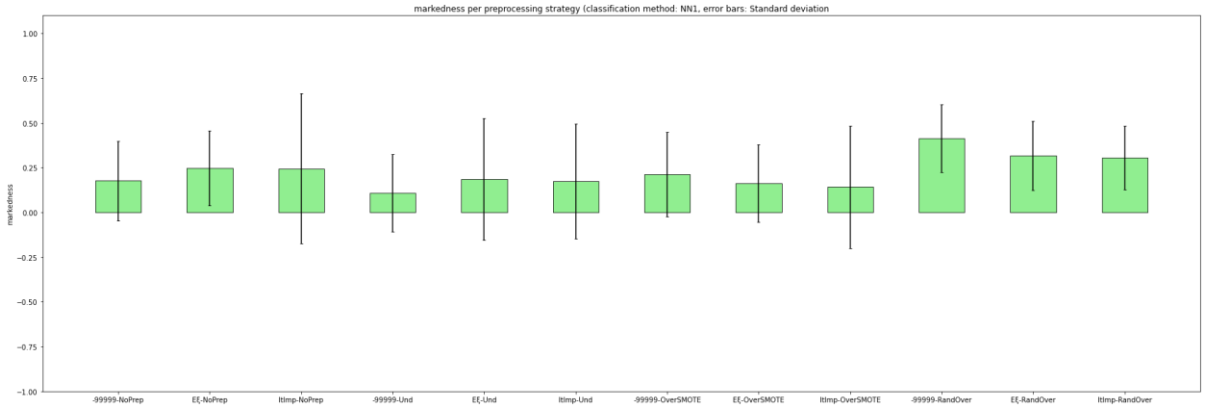

markedness per preprocessing strategy (classification method: DT2, error bars: Standard deviation
markedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

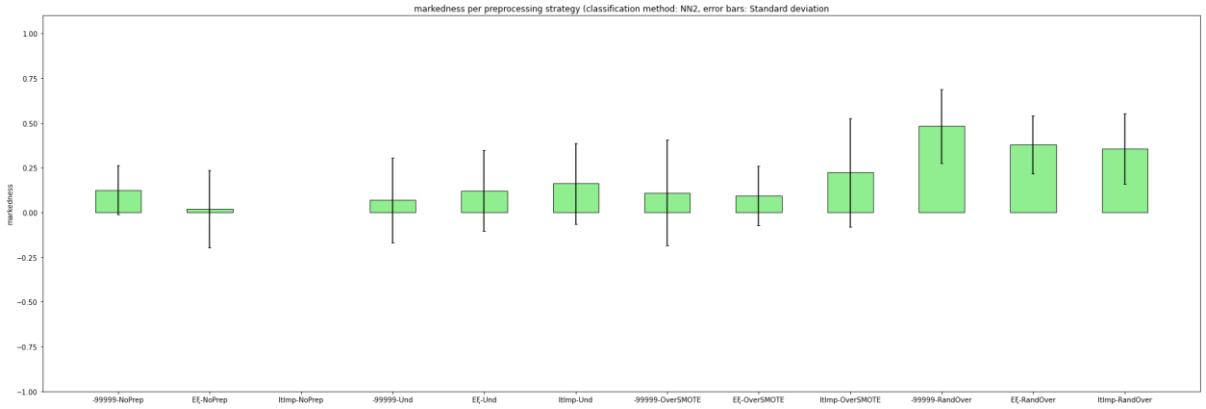

markedness per preprocessing strategy (classification method: DT3, error bars: Standard deviation
markedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

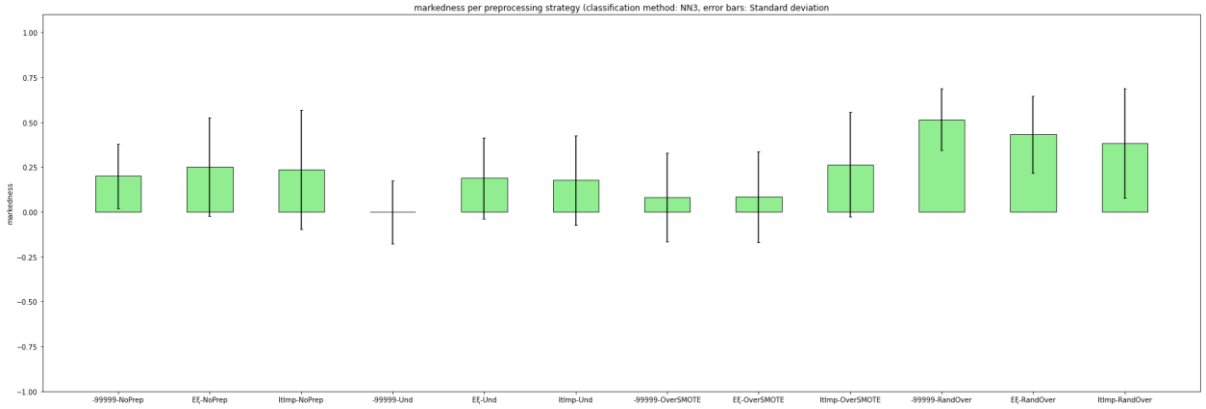

markedness per preprocessing strategy (classification method: DT4, error bars: Standard deviation
markedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

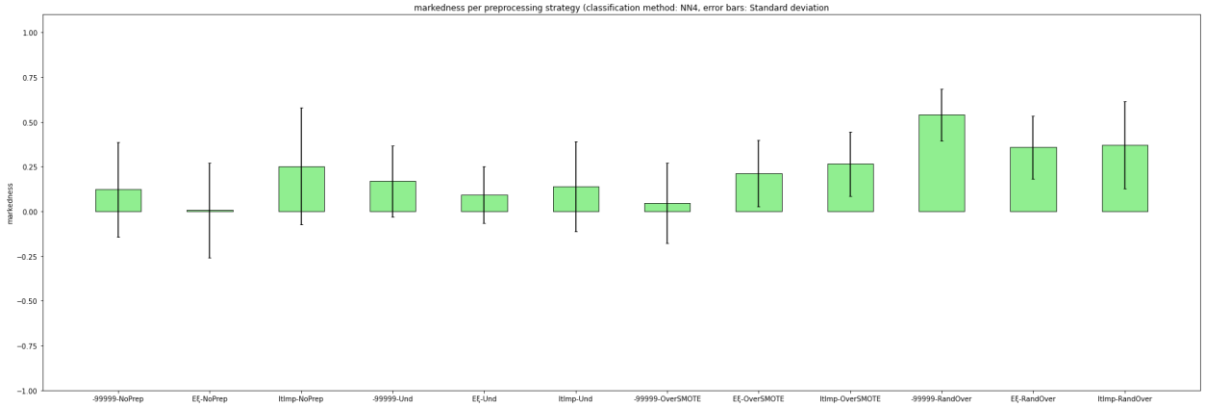

markedness per preprocessing strategy (classification method: RF1, error bars: Standard deviation
markedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

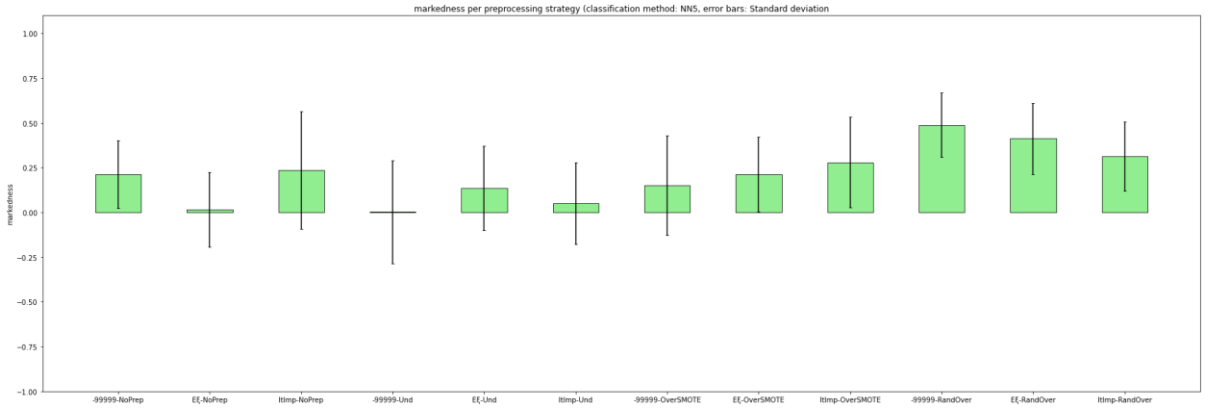

markedness per preprocessing strategy (classification method: RF2, error bars: Standard deviation
markedness
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CtWeigh
EE-CtWeigh
ItImp-CtWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

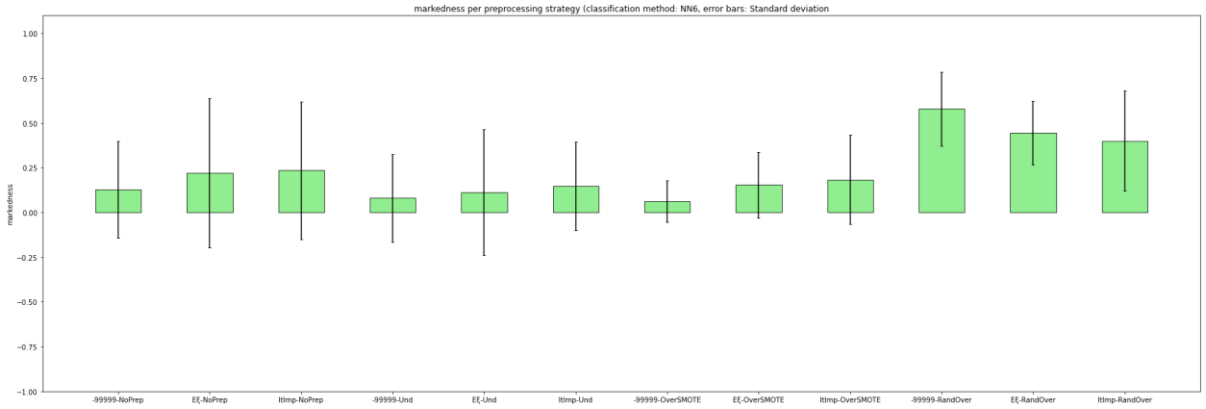

markedness per preprocessing strategy (classification method: NN1, error bars: Standard deviation
markedness
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

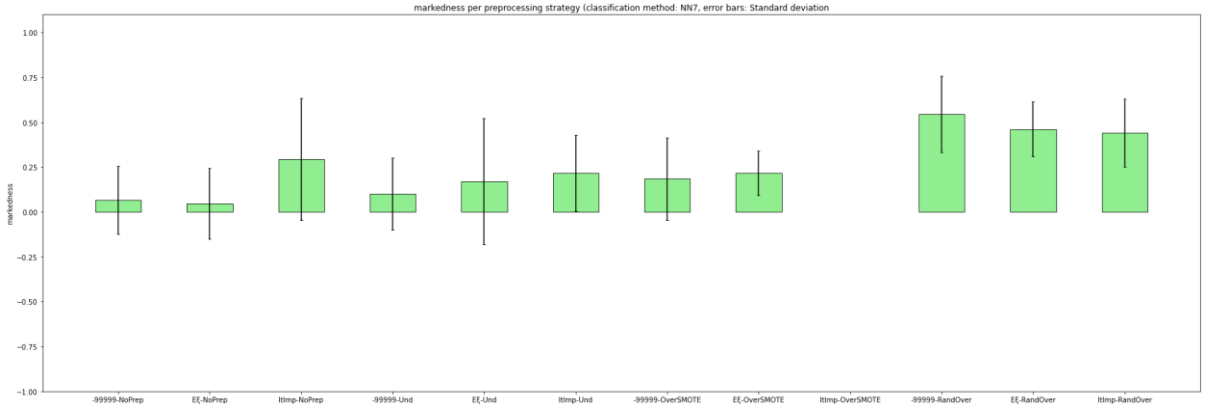

markedness per preprocessing strategy (classification method: NN2, error bars: Standard deviation
markedness
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

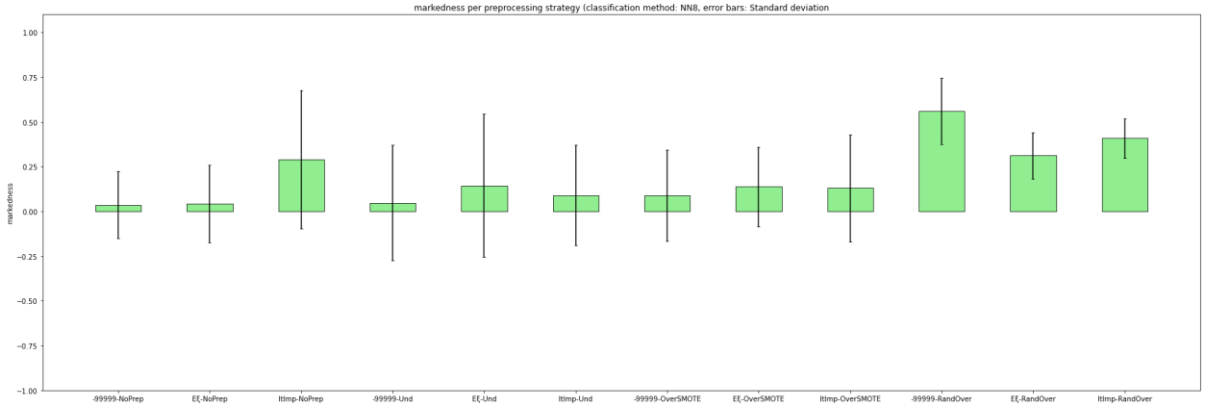

markedness per preprocessing strategy (classification method: NN3, error bars: Standard deviation
markedness
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

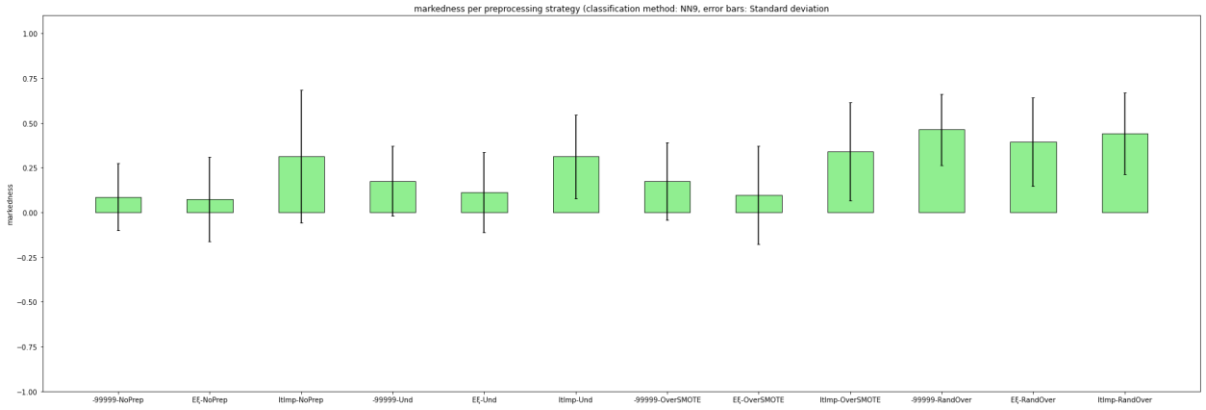

markedness per preprocessing strategy (classification method: NN4, error bars: Standard deviation
markedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

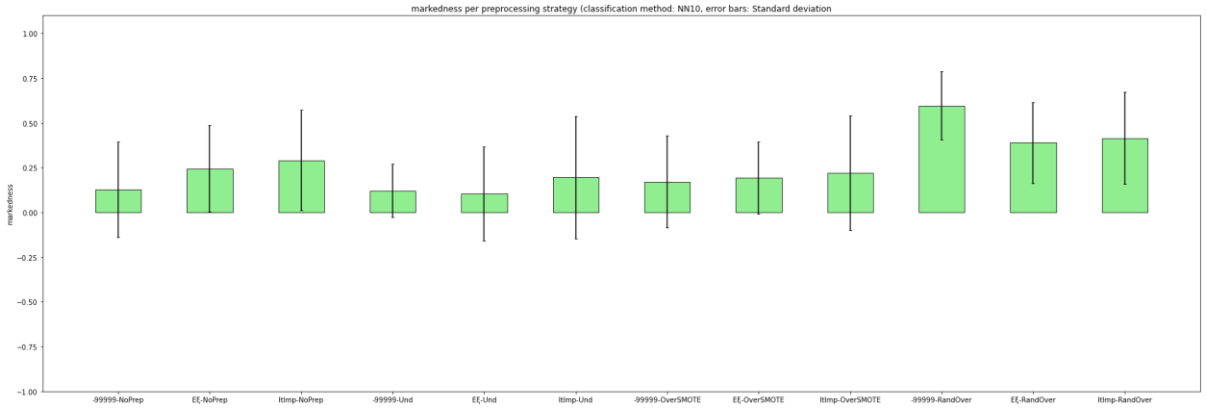

markedness per preprocessing strategy (classification method: NN5, error bars: Standard deviation
markedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

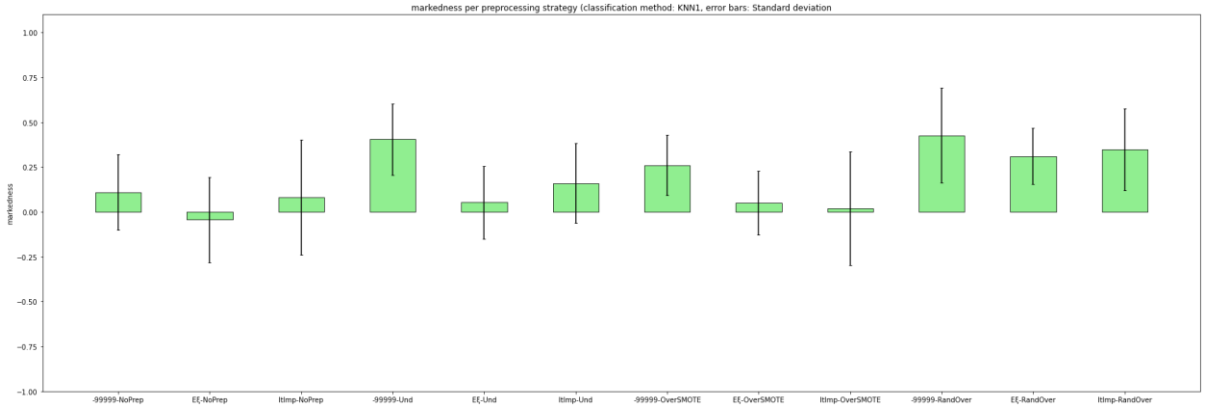

markedness per preprocessing strategy (classification method: NN6, error bars: Standard deviation
markedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

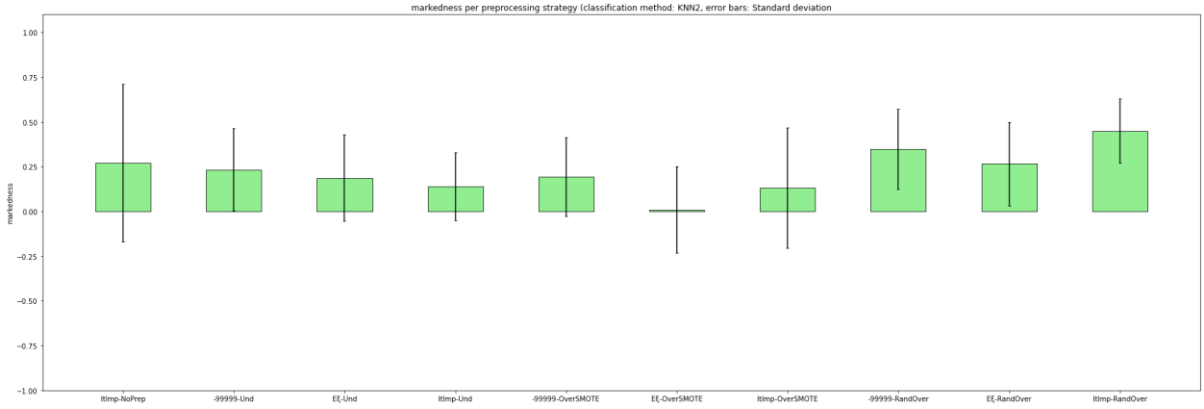

markedness per preprocessing strategy (classification method: NN7, error bars: Standard deviation
markedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

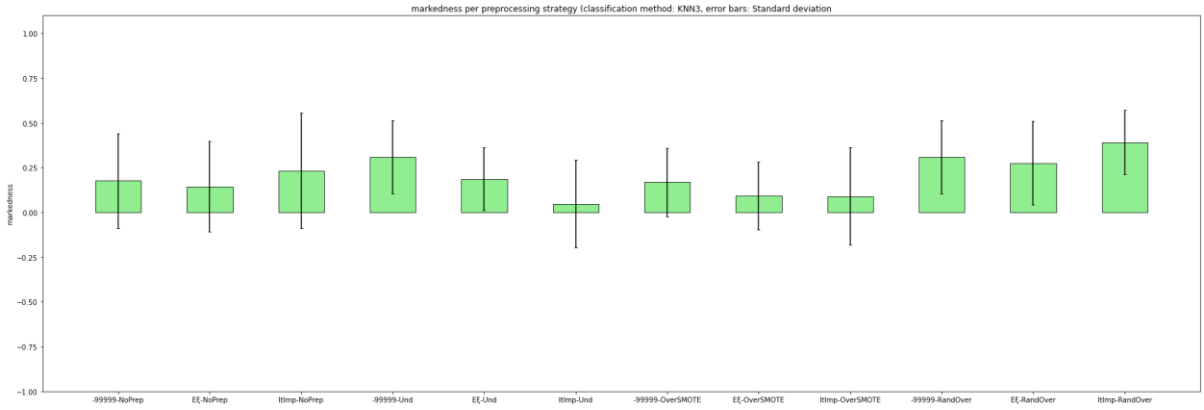

markedness per preprocessing strategy (classification method: NN8, error bars: Standard deviation
markedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

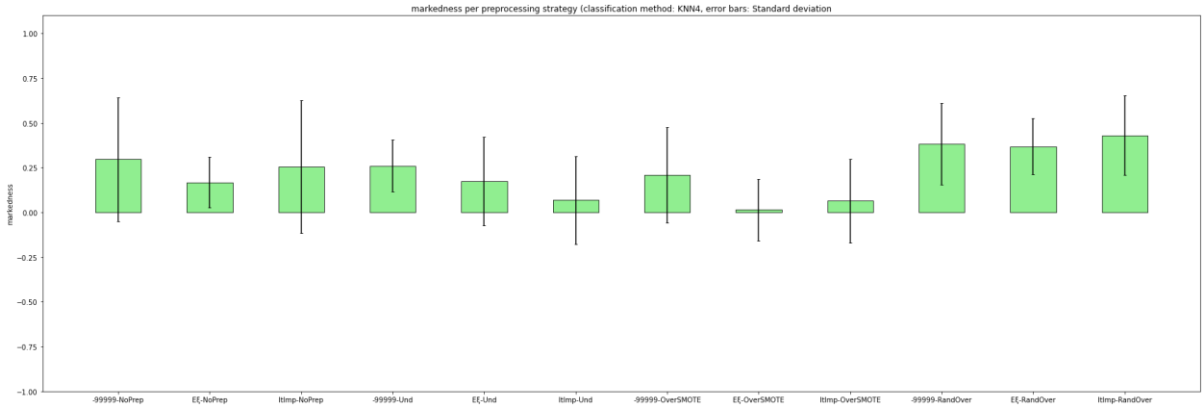

markedness per preprocessing strategy (classification method: NN9, error bars: Standard deviation
markedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

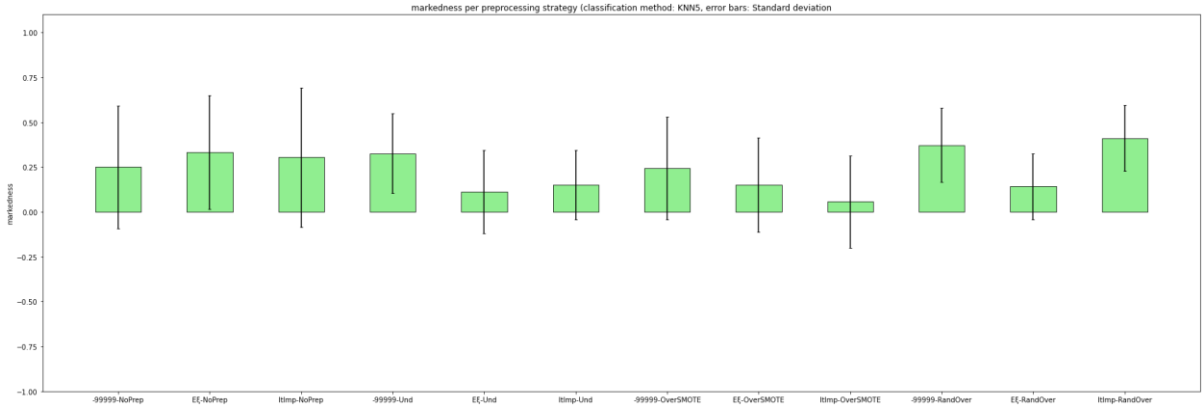

markedness per preprocessing strategy (classification method: NN10, error bars: Standard deviation
markedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

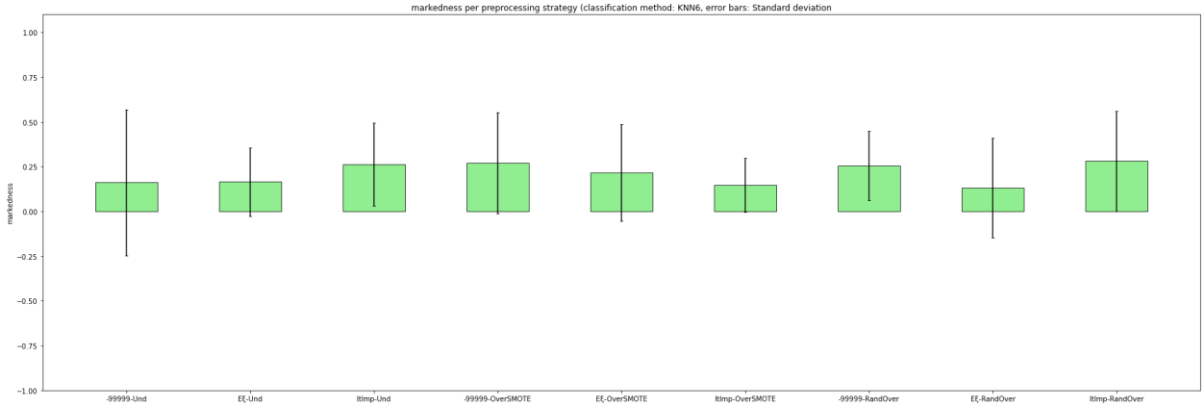

markedness per preprocessing strategy (classification method: KNN1, error bars: Standard deviation
markedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

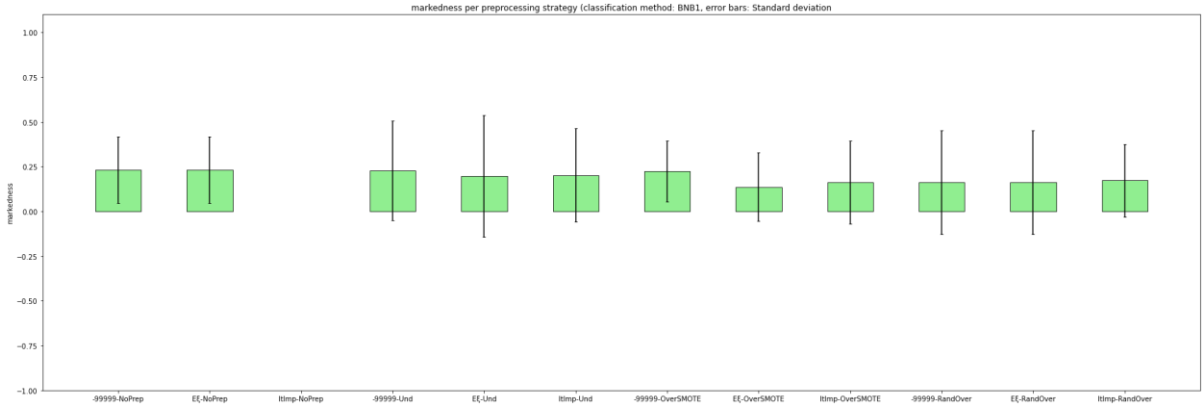

markedness per preprocessing strategy (classification method: KNN2, error bars: Standard deviation
markedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

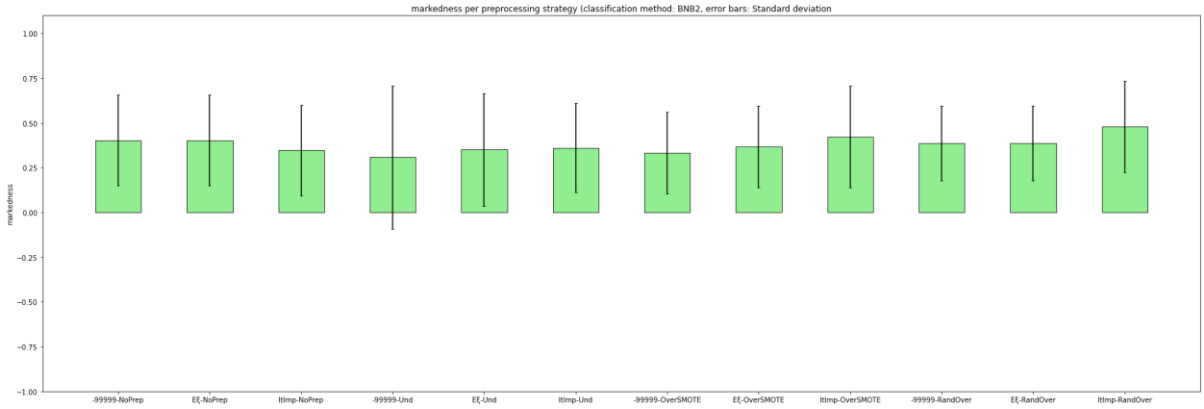

markedness per preprocessing strategy (classification method: KNN3, error bars: Standard deviation
markedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

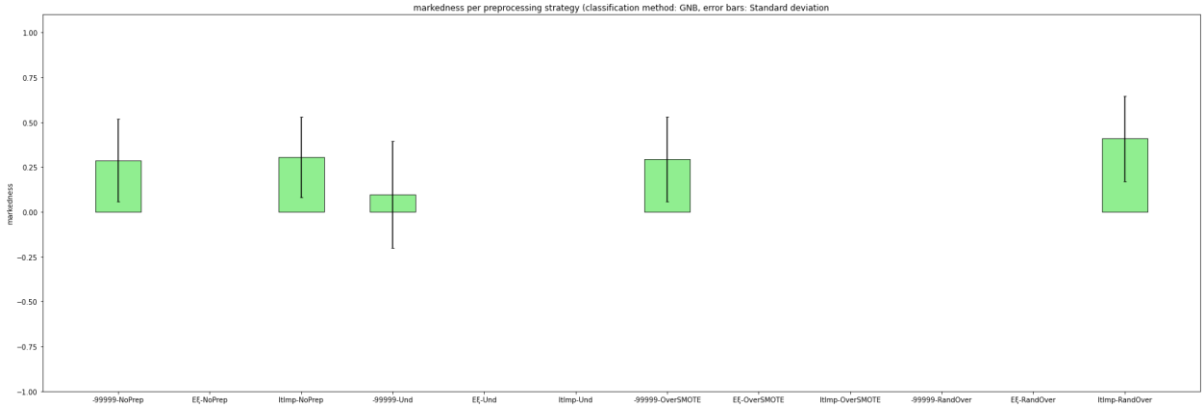

markedness per preprocessing strategy (classification method: KNN4, error bars: Standard deviation
markedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

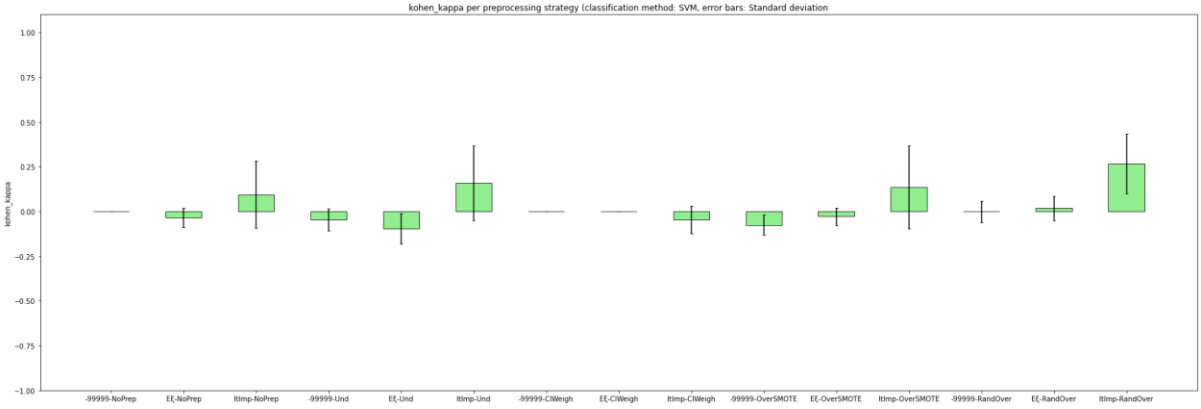

markedness per preprocessing strategy (classification method: KNN5, error bars: Standard deviation
markedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

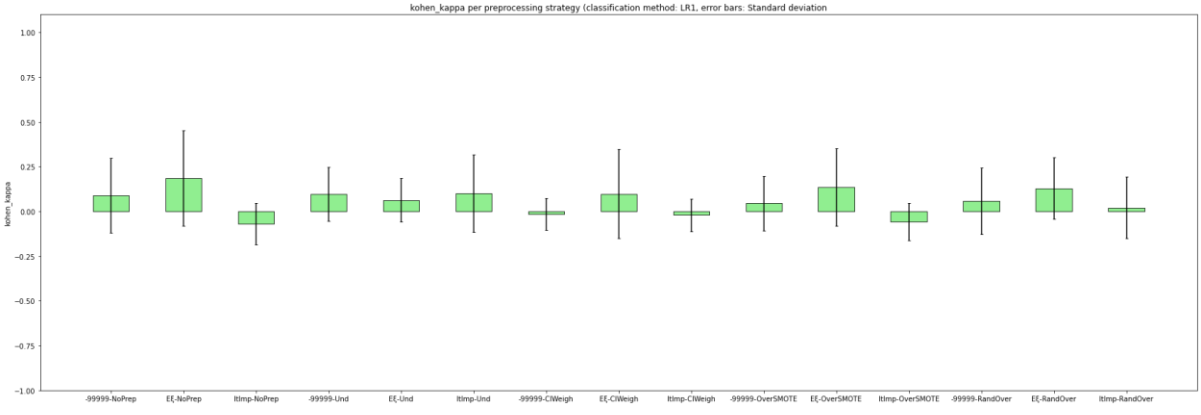

markedness per preprocessing strategy (classification method: KNN6, error bars: Standard deviation
markedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

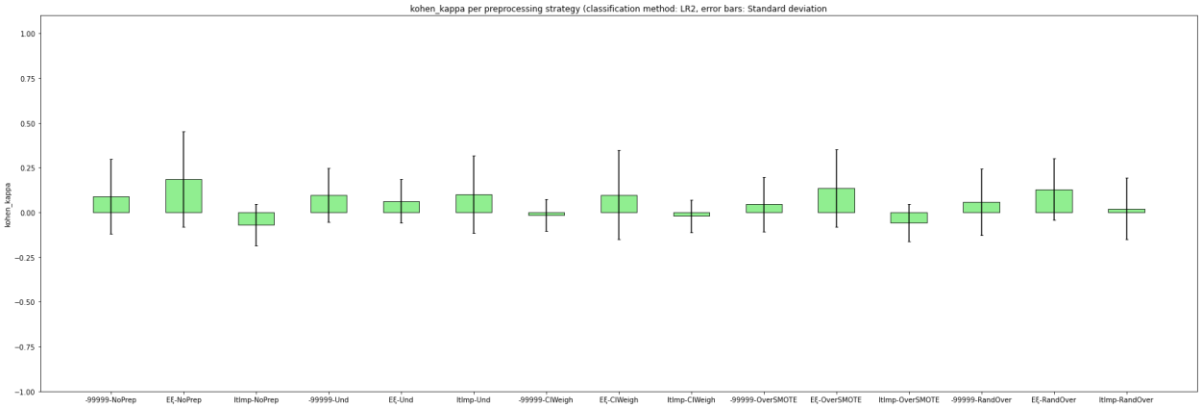

markedness per preprocessing strategy (classification method: BNB1, error bars: Standard deviation
markedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

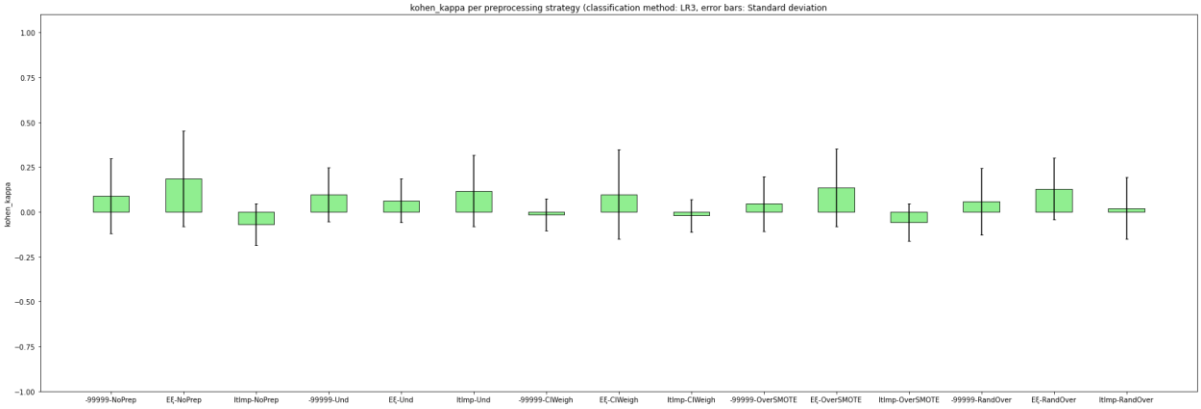

markedness per preprocessing strategy (classification method: BNB2, error bars: Standard deviation
markedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

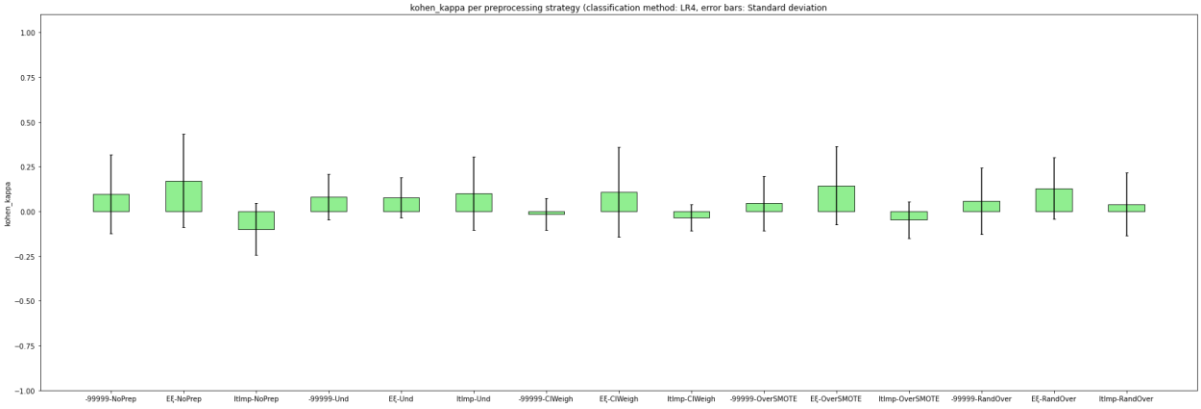

markedness per preprocessing strategy (classification method: GNB, error bars: Standard deviation
markedness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

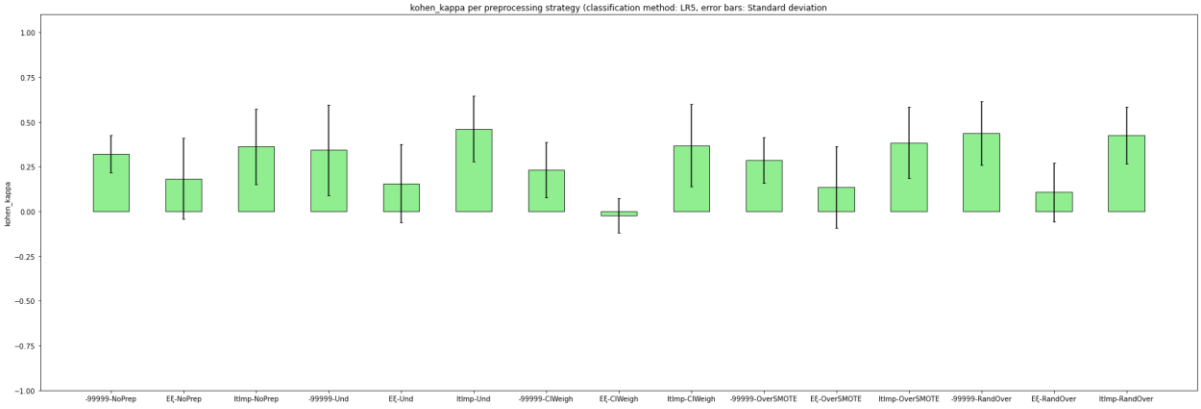

kohen_kappa per preprocessing strategy (classification method: SVM, error bars: Standard deviation
kohen_kappa
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

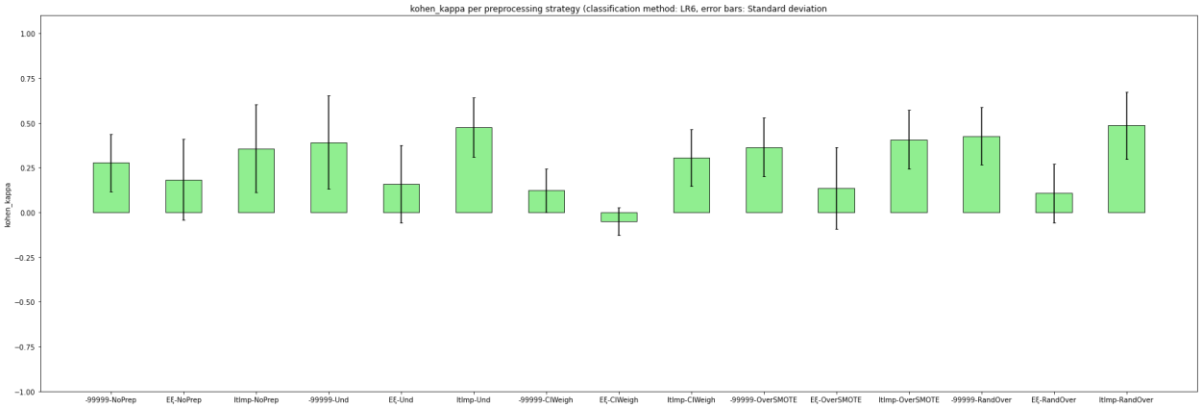

kohen_kappa per preprocessing strategy (classification method: LR1, error bars: Standard deviation
kohen_kappa
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

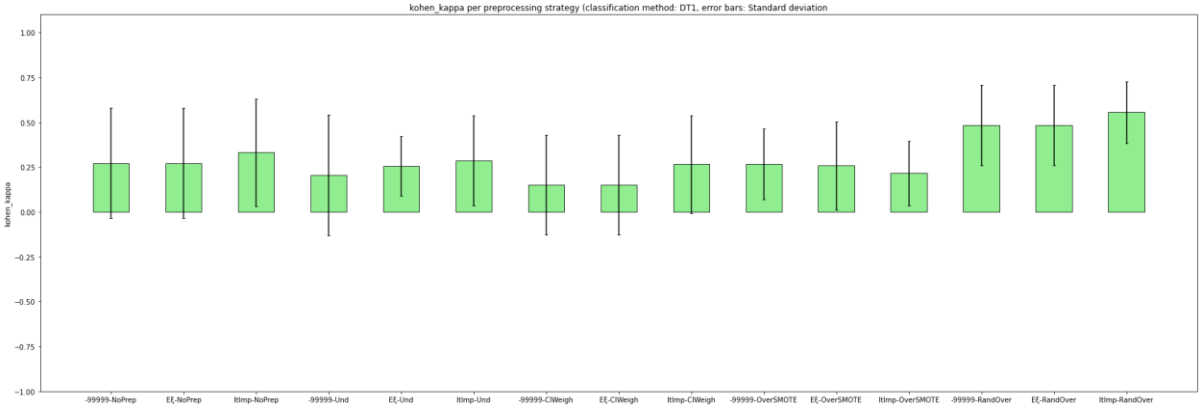

kohen_kappa per preprocessing strategy (classification method: LR2, error bars: Standard deviation
kohen_kappa
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

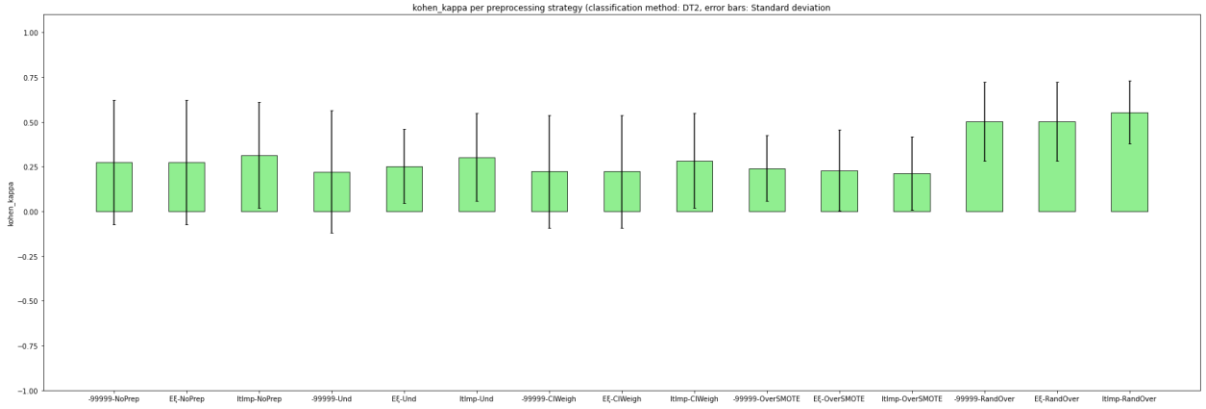

kohen_kappa per preprocessing strategy (classification method: LR3, error bars: Standard deviation
kohen_kappa
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

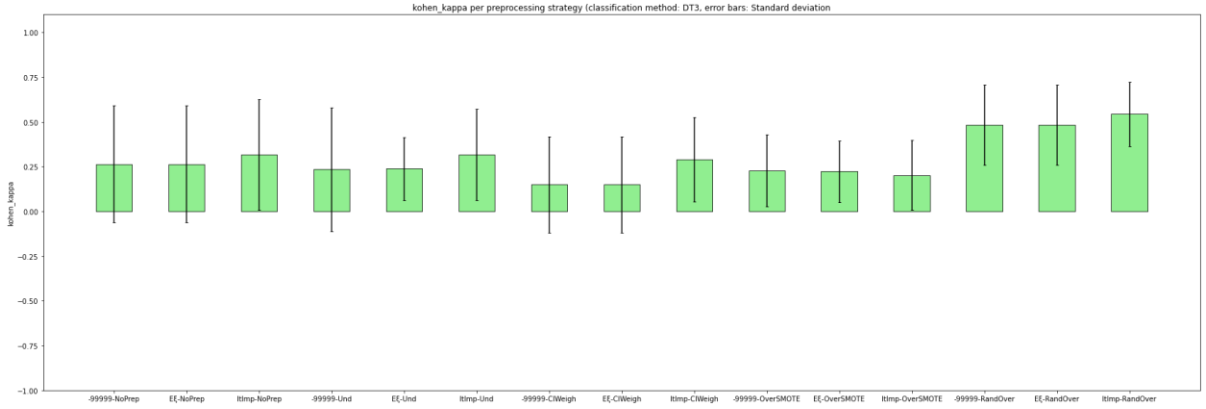

kohen_kappa per preprocessing strategy (classification method: LR4, error bars: Standard deviation
kohen_kappa
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

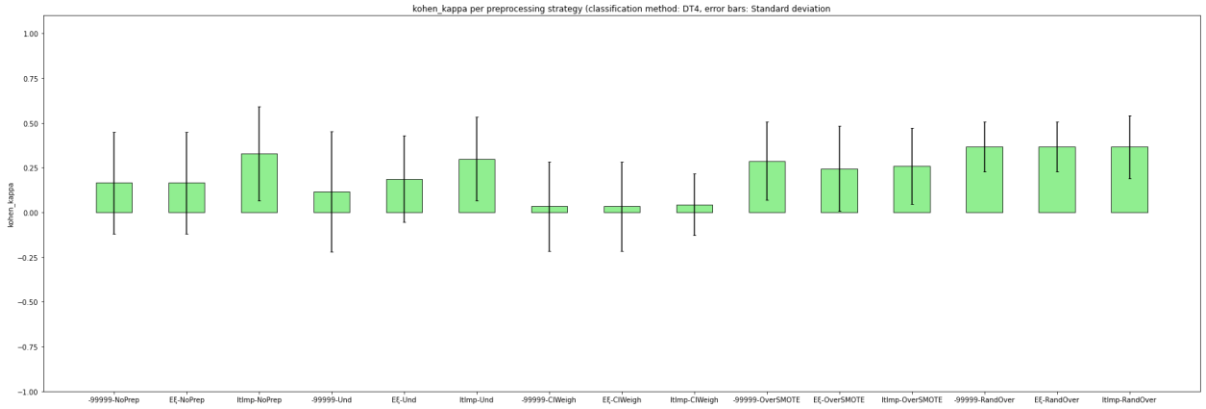

kohen_kappa per preprocessing strategy (classification method: LR5, error bars: Standard deviation
kohen_kappa
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

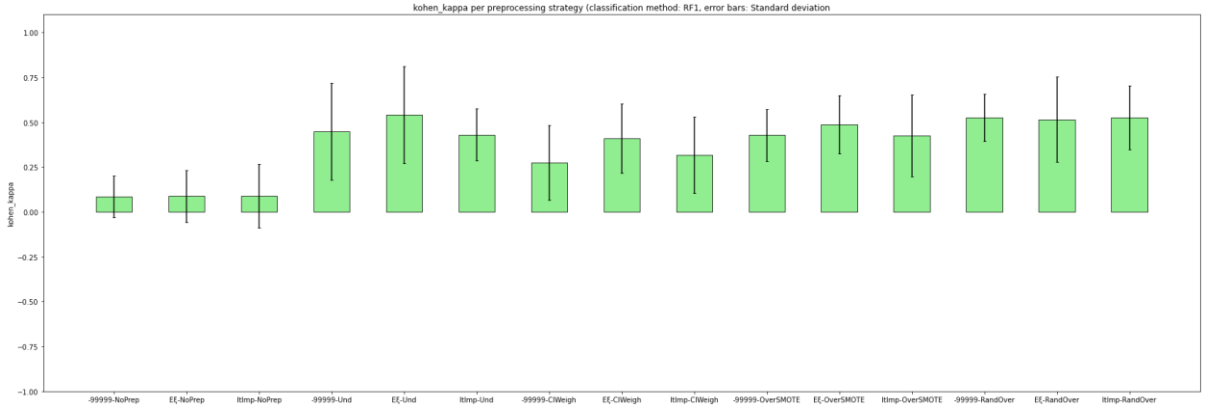

kohen_kappa per preprocessing strategy (classification method: LR6, error bars: Standard deviation
kohen_kappa
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

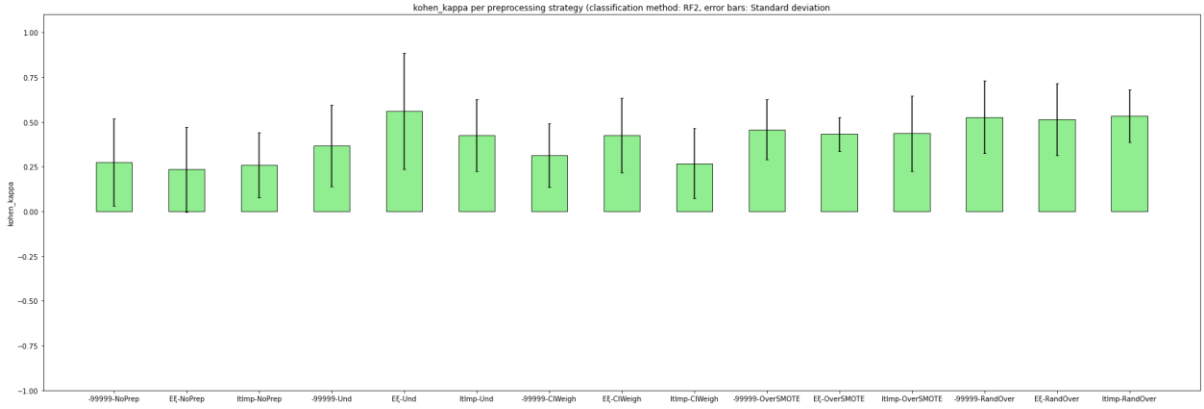

kohen_kappa per preprocessing strategy (classification method: DT1, error bars: Standard deviation
kohen_kappa
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

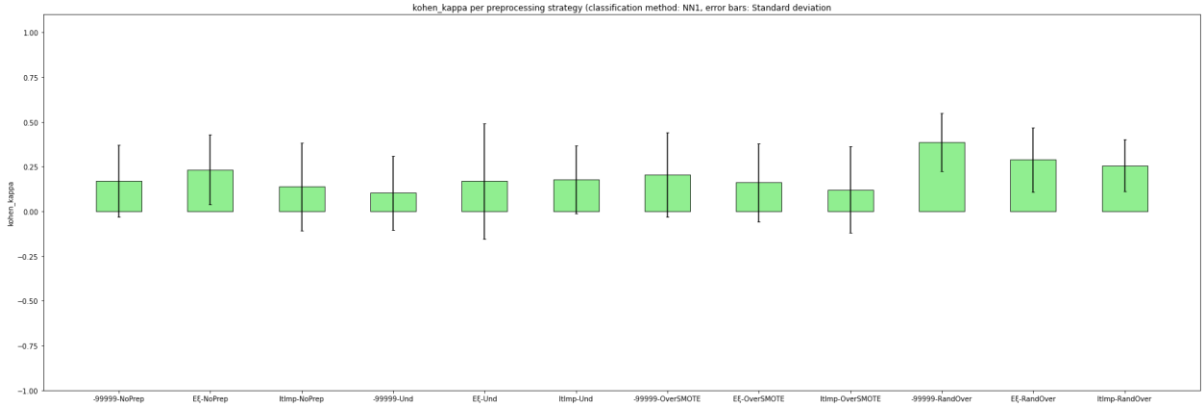

kohen_kappa per preprocessing strategy (classification method: DT2, error bars: Standard deviation
kohen_kappa
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

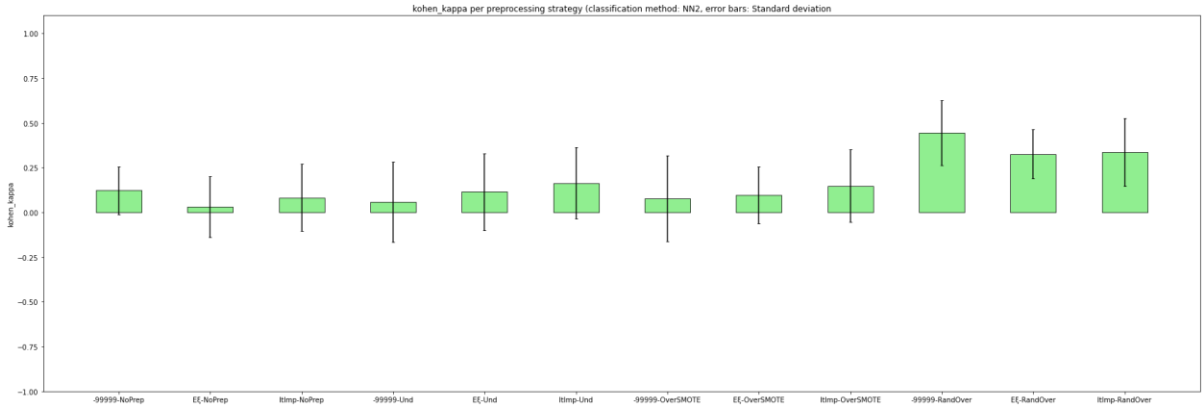

kohen_kappa per preprocessing strategy (classification method: DT3, error bars: Standard deviation
kohen_kappa
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

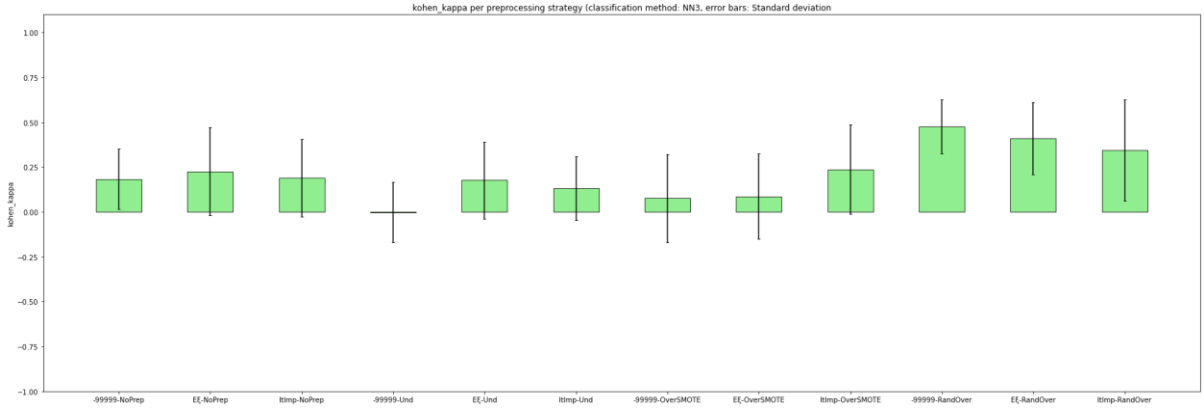

kohen_kappa per preprocessing strategy (classification method: DT4, error bars: Standard deviation
kohen_kappa
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

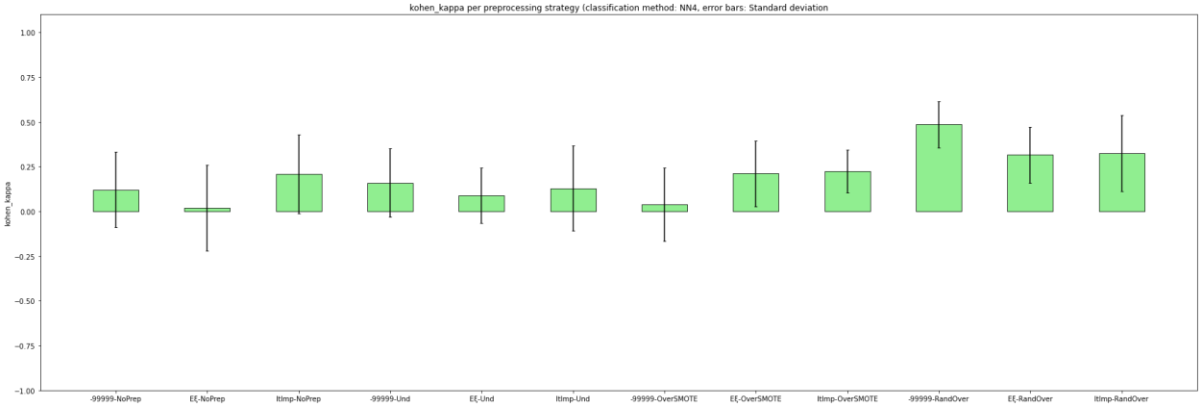

kohen_kappa per preprocessing strategy (classification method: RF1, error bars: Standard deviation
kohen_kappa
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

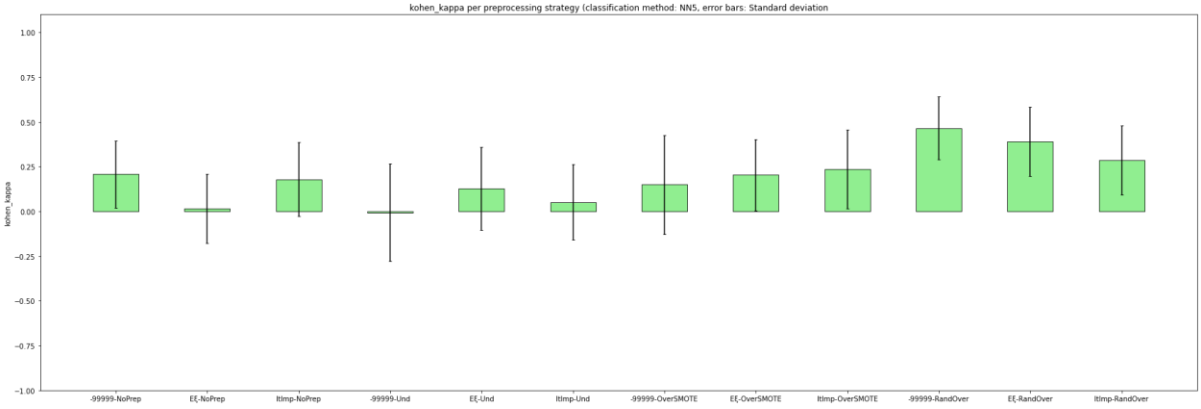

kohen_kappa per preprocessing strategy (classification method: RF2, error bars: Standard deviation
kohen_kappa
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-ClWeigh
EE-ClWeigh
ItImp-ClWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

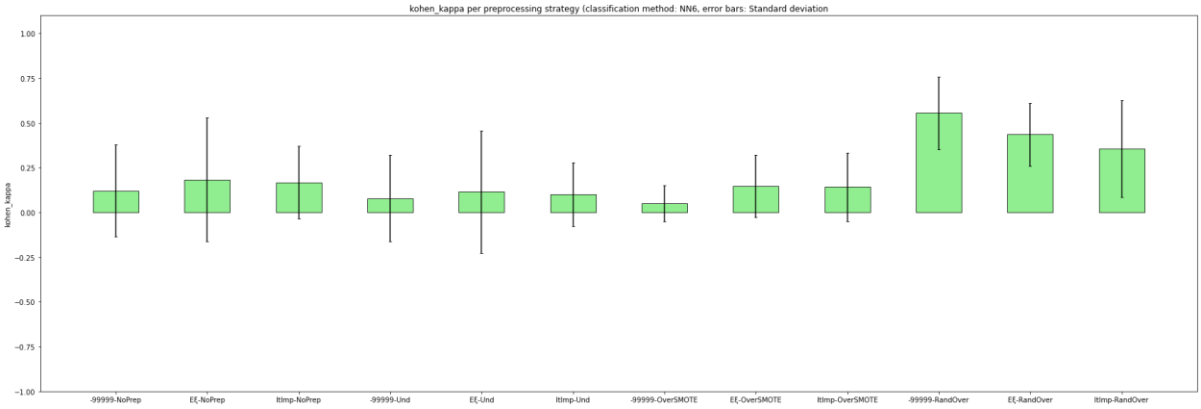

kohen_kappa per preprocessing strategy (classification method: NN1, error bars: Standard deviation
kohen_kappa
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

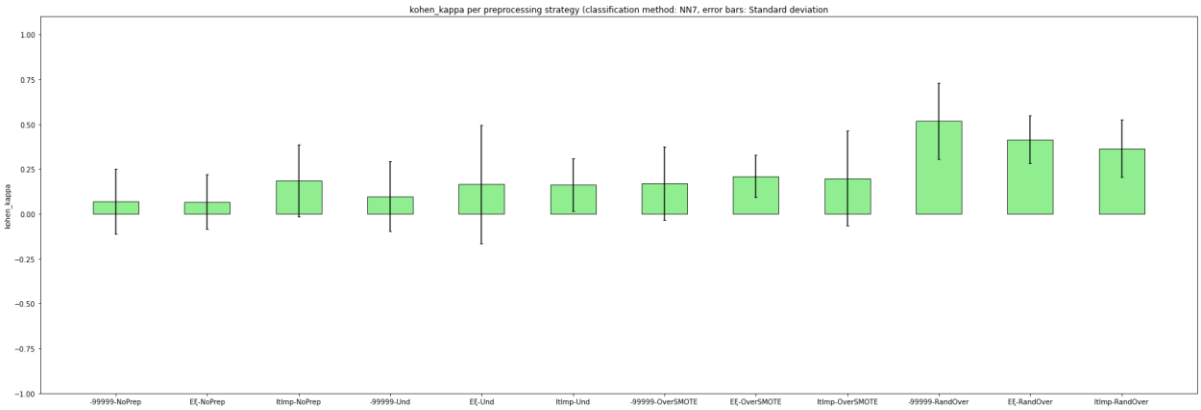

kohen_kappa per preprocessing strategy (classification method: NN2, error bars: Standard deviation
kohen_kappa
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

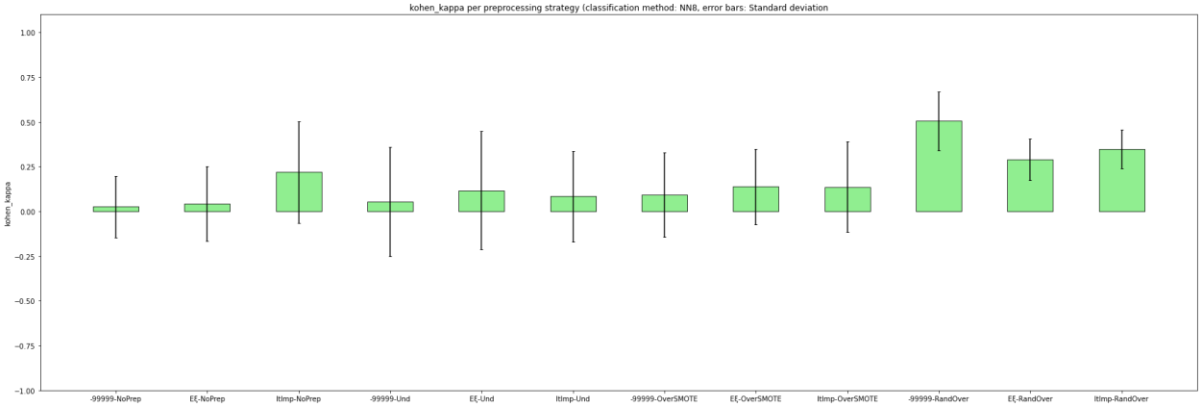

kohen_kappa per preprocessing strategy (classification method: NN3, error bars: Standard deviation
kohen_kappa
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

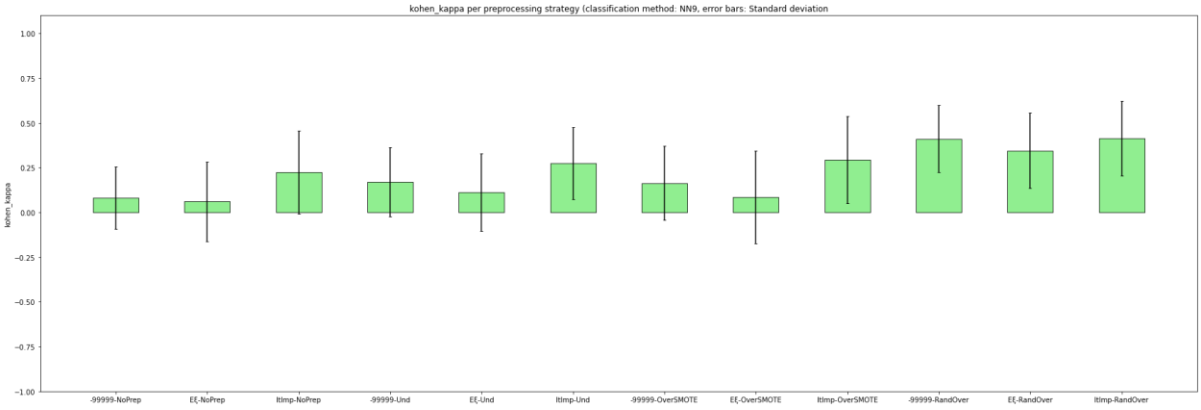

kohen_kappa per preprocessing strategy (classification method: NN4, error bars: Standard deviation
kohen_kappa
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

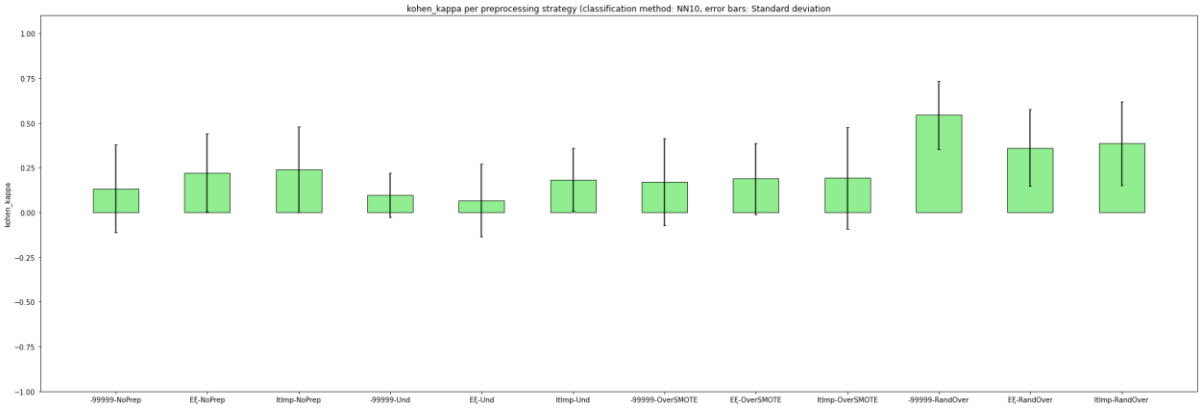

kohen_kappa per preprocessing strategy (classification method: NN5, error bars: Standard deviation
kohen_kappa
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

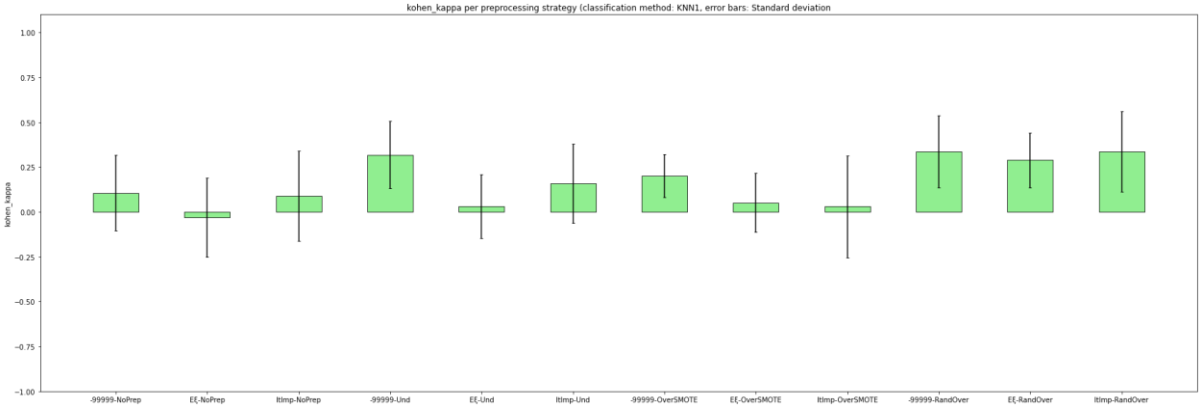

kohen_kappa per preprocessing strategy (classification method: NN6, error bars: Standard deviation
kohen_kappa
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

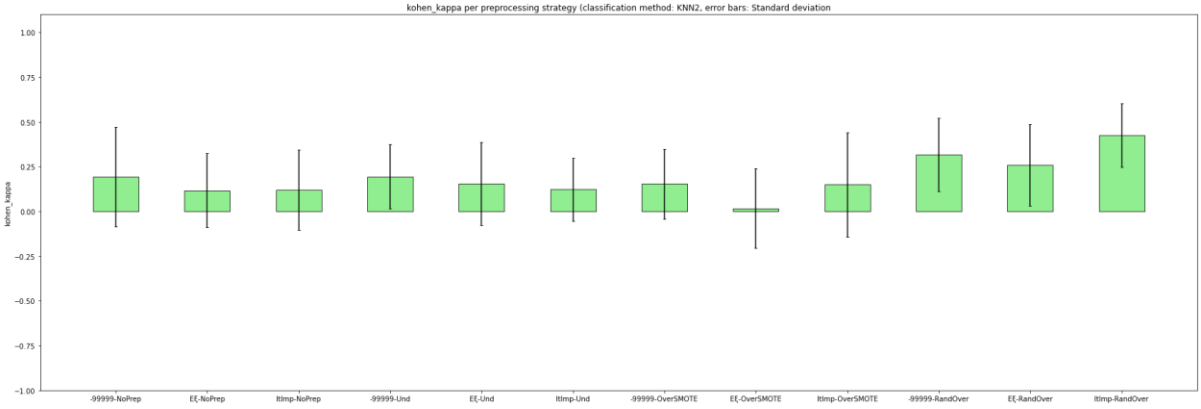

kohen_kappa per preprocessing strategy (classification method: NN7, error bars: Standard deviation
kohen_kappa
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

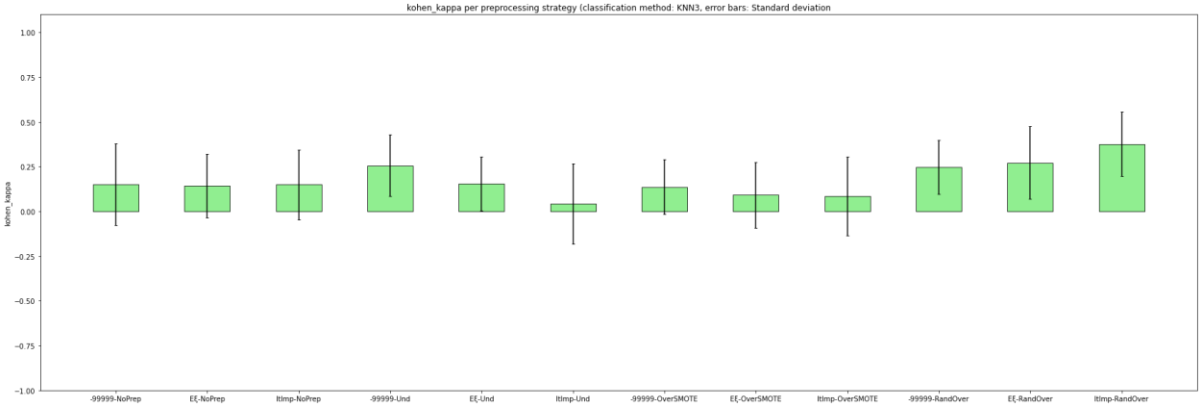

kohen_kappa per preprocessing strategy (classification method: NN8, error bars: Standard deviation
kohen_kappa
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

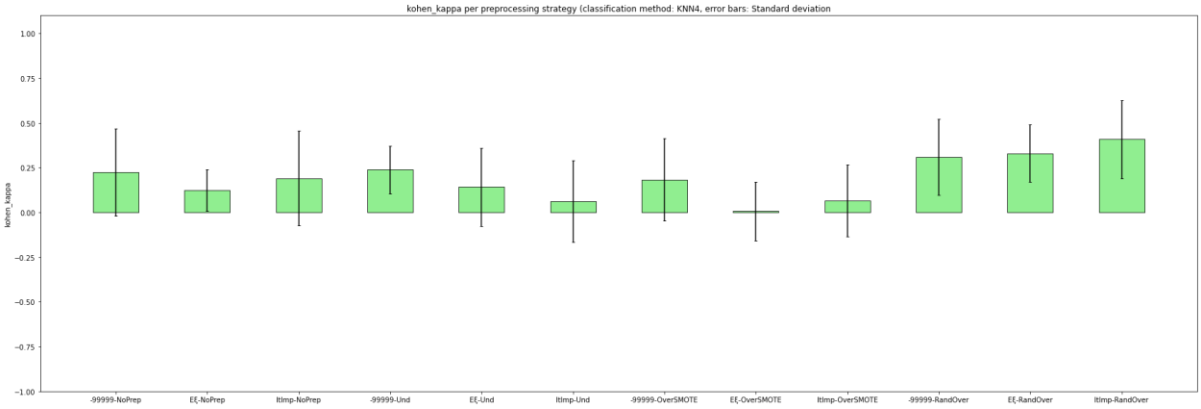

kohen_kappa per preprocessing strategy (classification method: NN9, error bars: Standard deviation
kohen_kappa
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

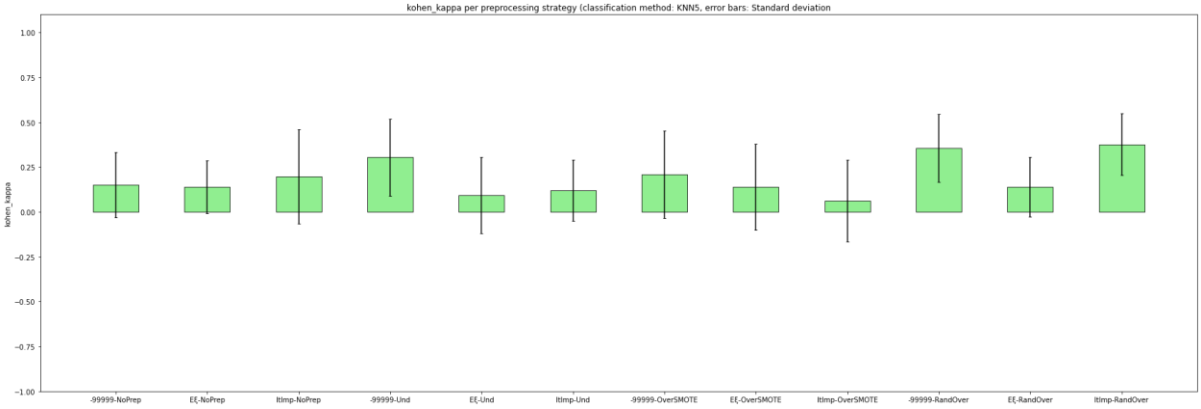

kohen_kappa per preprocessing strategy (classification method: NN10, error bars: Standard deviation
kohen_kappa
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

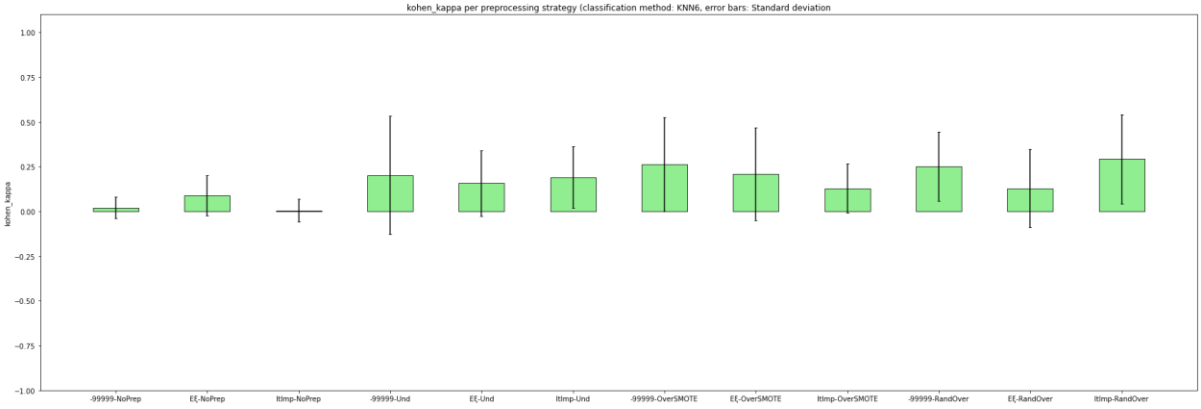

kohen_kappa per preprocessing strategy (classification method: KNN1, error bars: Standard deviation
kohen_kappa
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

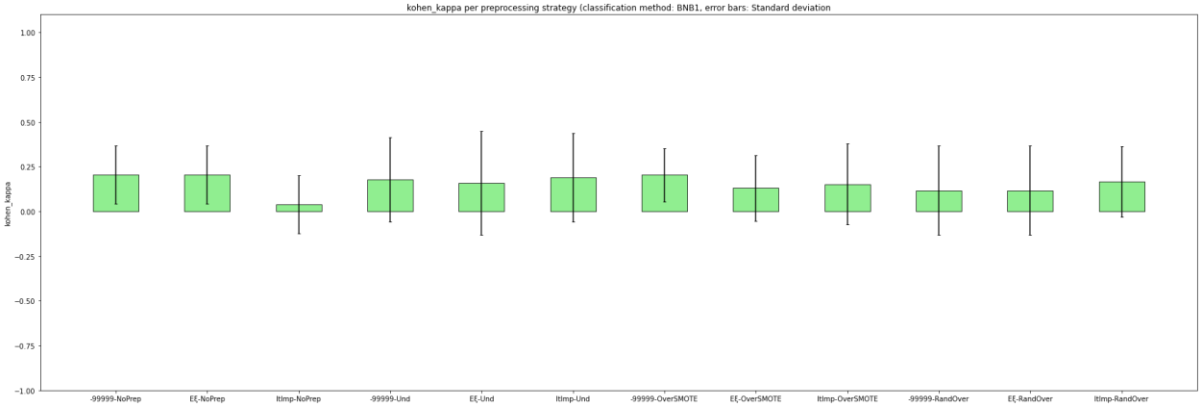

kohen_kappa per preprocessing strategy (classification method: KNN2, error bars: Standard deviation
kohen_kappa
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

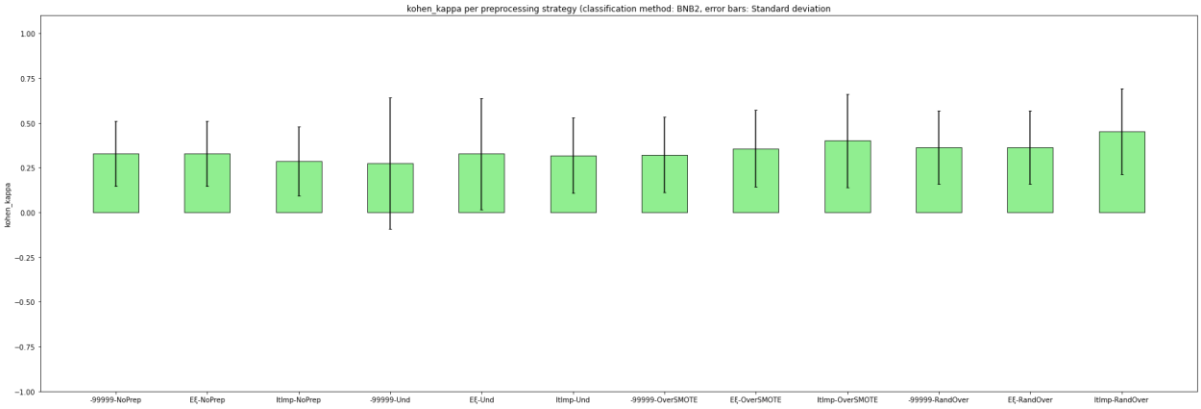

kohen_kappa per preprocessing strategy (classification method: KNN3, error bars: Standard deviation
kohen_kappa
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

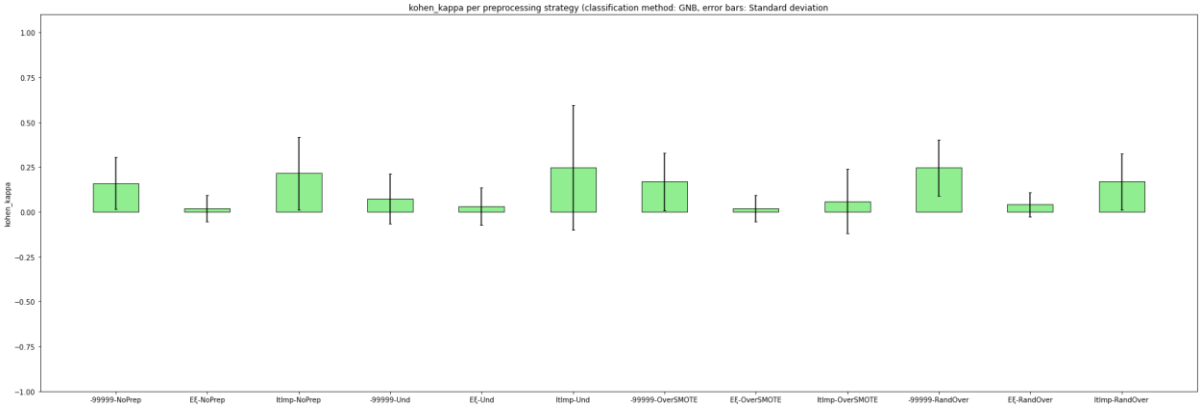

kohen_kappa per preprocessing strategy (classification method: KNN4, error bars: Standard deviation
kohen_kappa
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

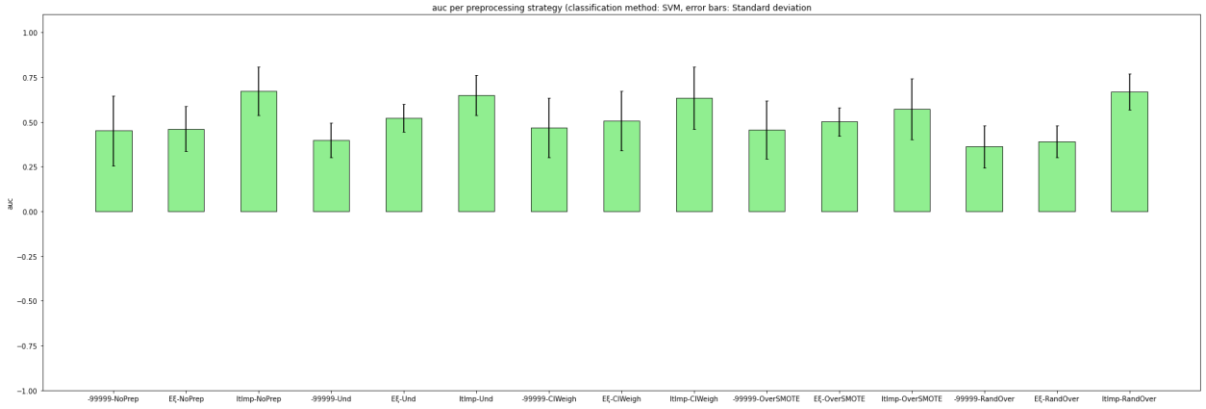

kohen_kappa per preprocessing strategy (classification method: KNN5, error bars: Standard deviation
kohen_kappa
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

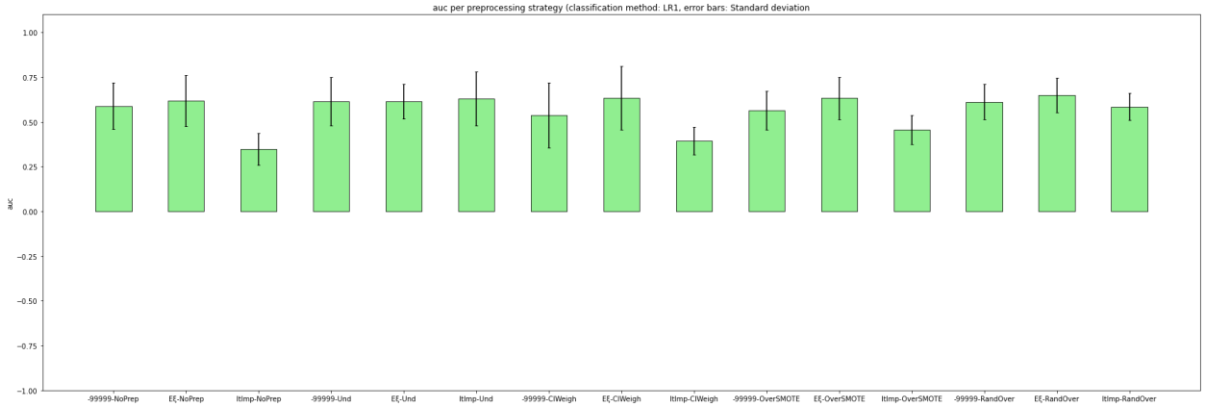

kohen_kappa per preprocessing strategy (classification method: KNN6, error bars: Standard deviation
kohen_kappa
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

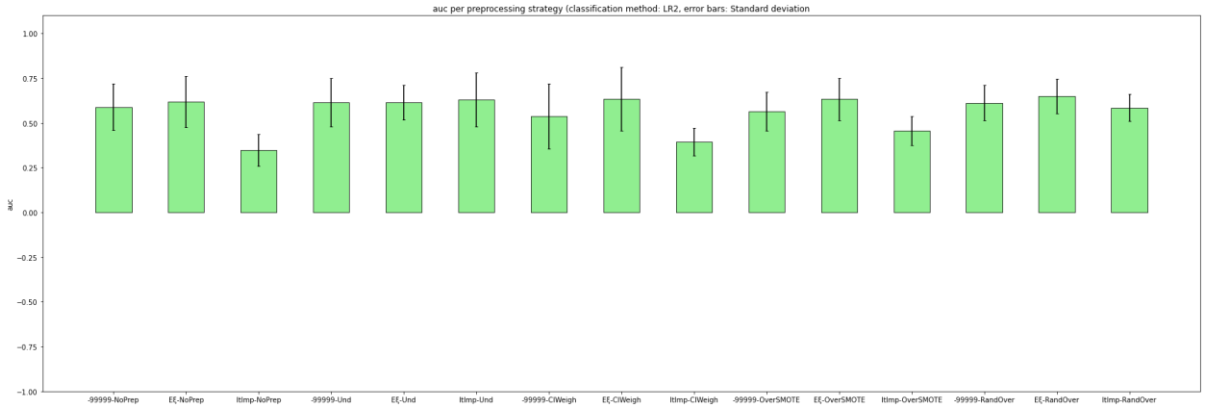

kohen_kappa per preprocessing strategy (classification method: BNB1, error bars: Standard deviation
kohen_kappa
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

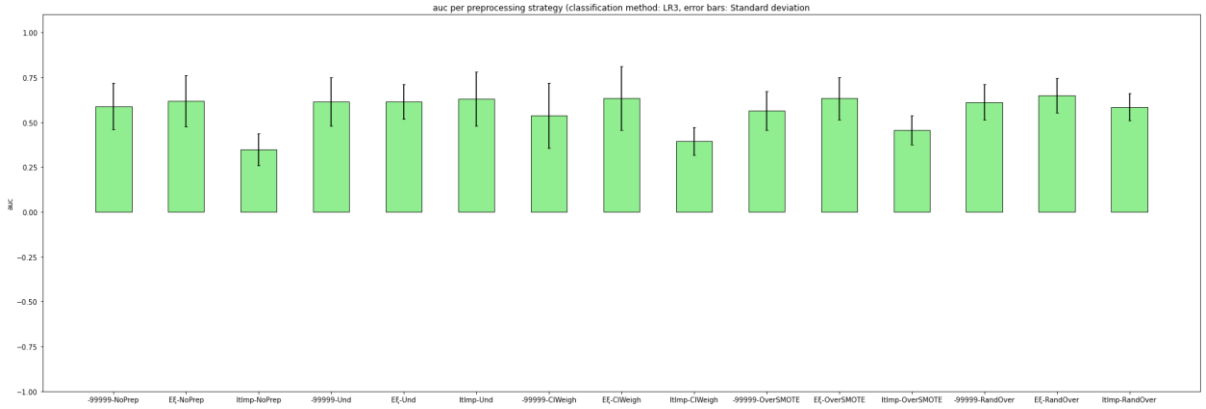

kohen_kappa per preprocessing strategy (classification method: BNB2, error bars: Standard deviation
kohen_kappa
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

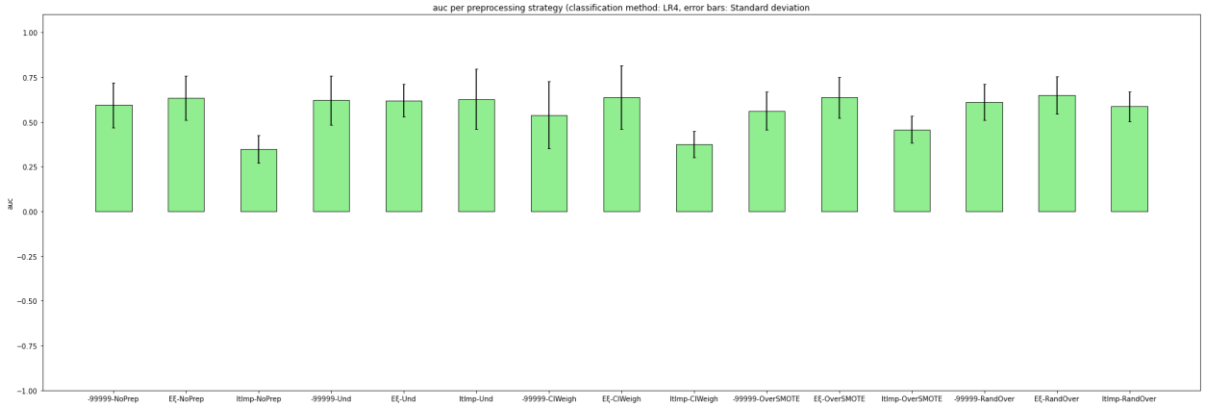

kohen_kappa per preprocessing strategy (classification method: GNB, error bars: Standard deviation
kohen_kappa
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

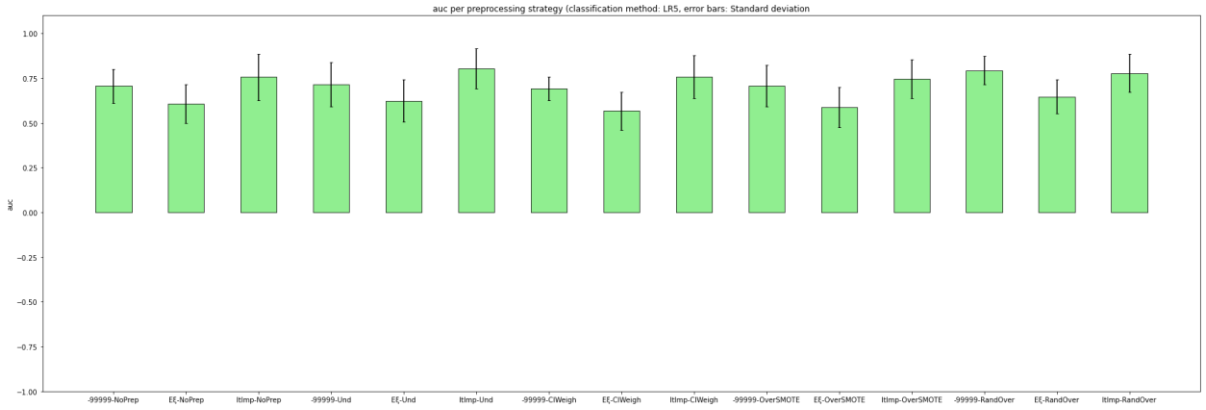

auc per preprocessing strategy (classification method: SVM, error bars: Standard deviation
auc
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

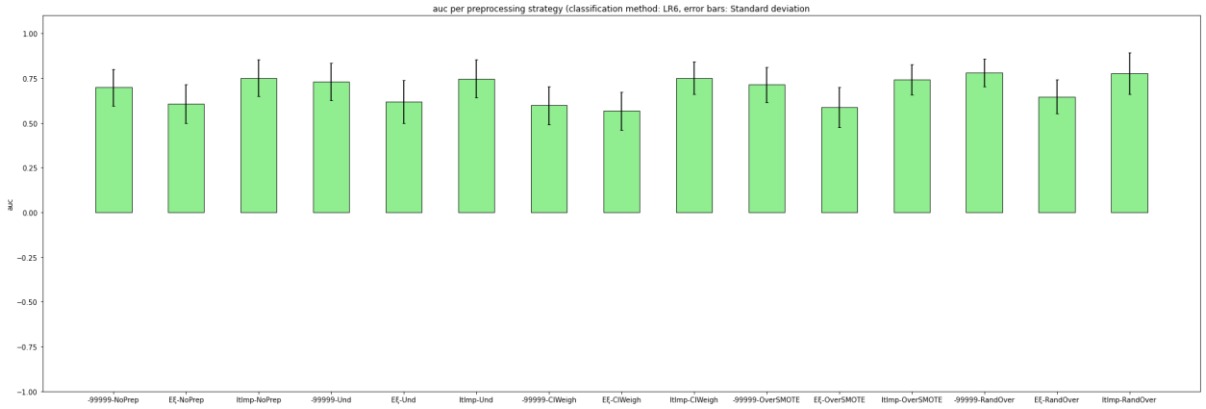

auc per preprocessing strategy (classification method: LR1, error bars: Standard deviation
auc
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

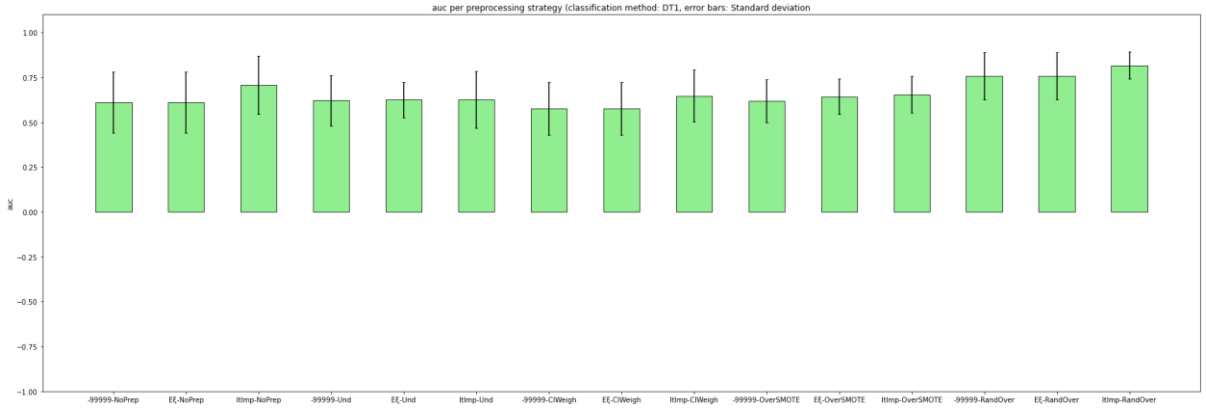

auc per preprocessing strategy (classification method: LR2, error bars: Standard deviation
auc
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

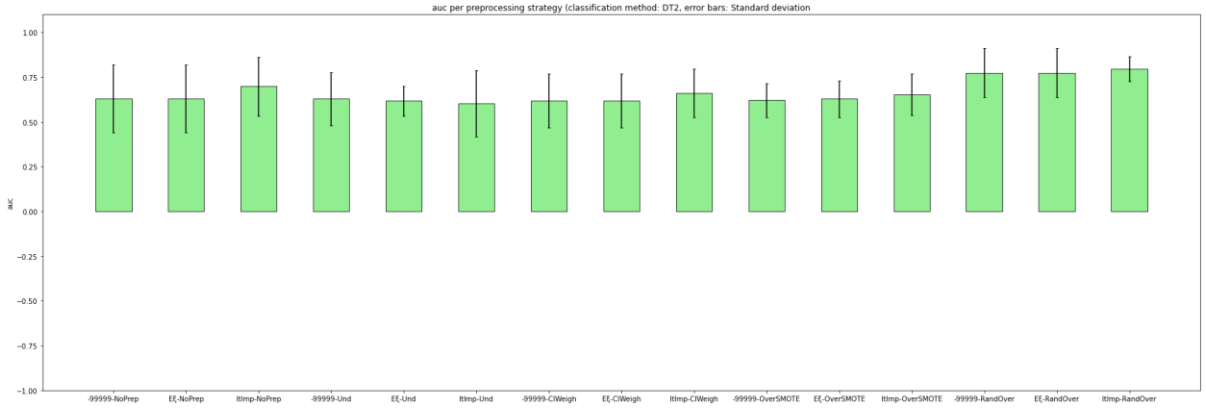

auc per preprocessing strategy (classification method: LR3, error bars: Standard deviation
auc
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

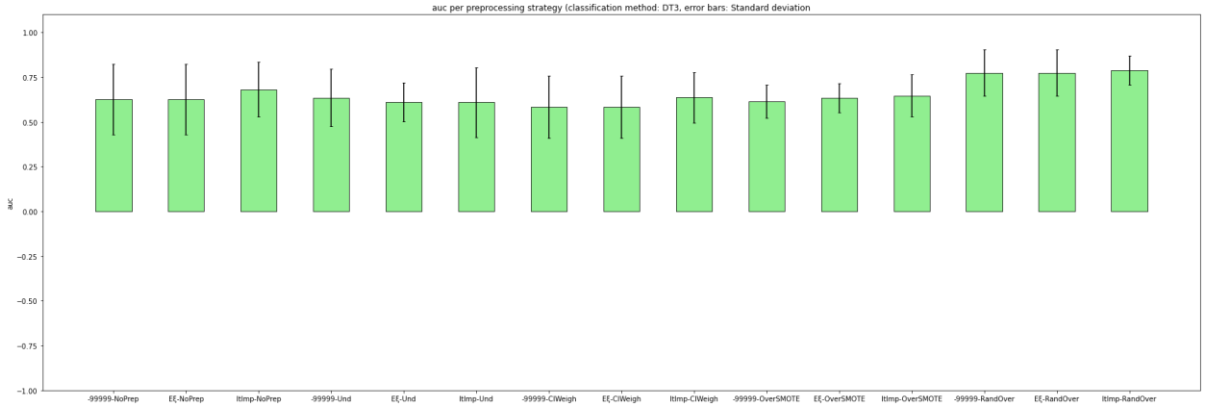

auc per preprocessing strategy (classification method: LR4, error bars: Standard deviation
auc
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

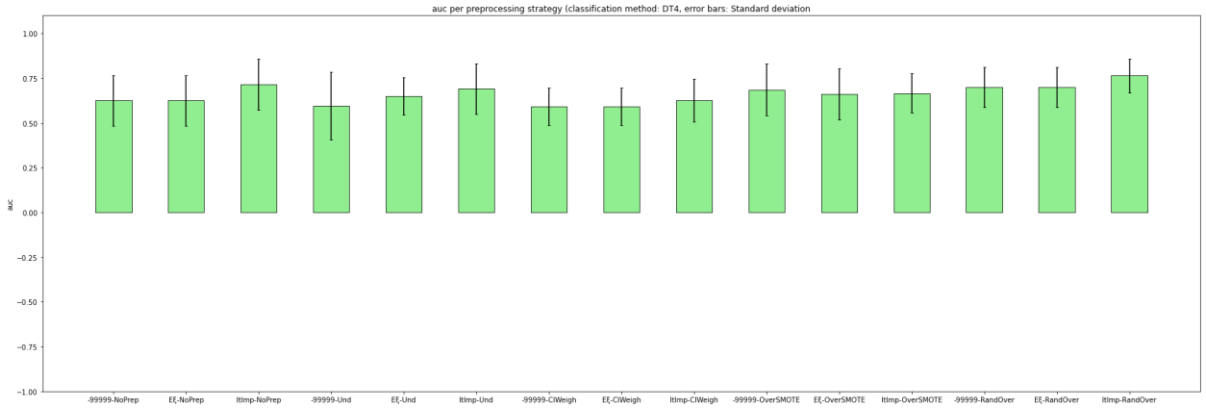

auc per preprocessing strategy (classification method: LR5, error bars: Standard deviation
auc
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

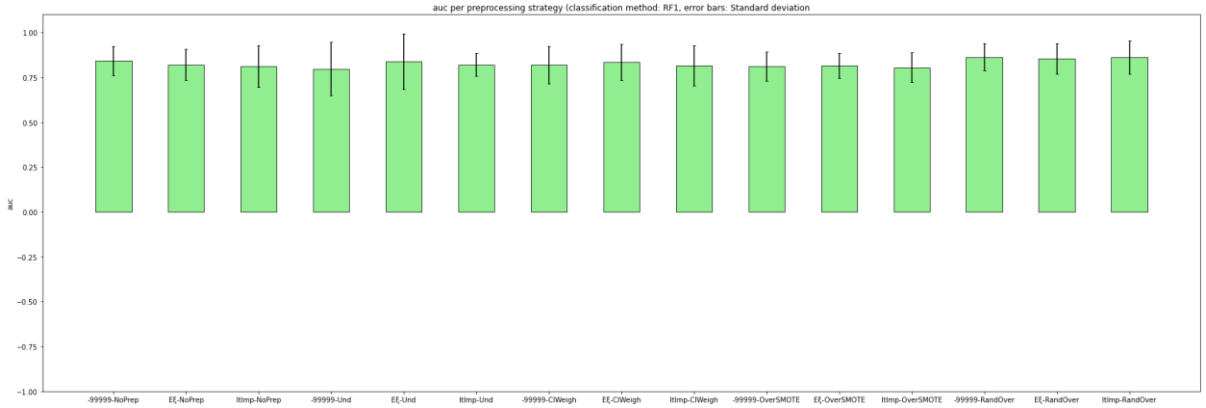

auc per preprocessing strategy (classification method: LR6, error bars: Standard deviation
auc
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

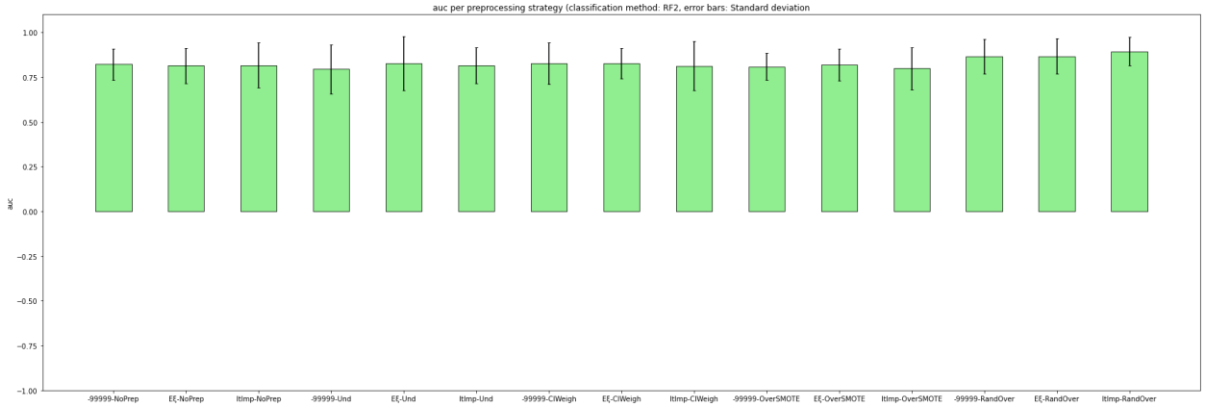

auc per preprocessing strategy (classification method: DT1, error bars: Standard deviation
auc
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

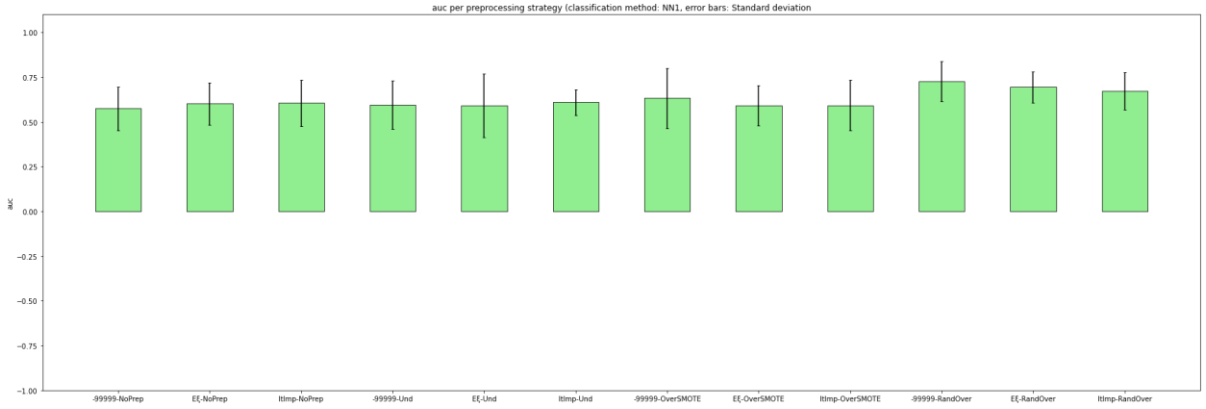

auc per preprocessing strategy (classification method: DT2, error bars: Standard deviation
auc
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

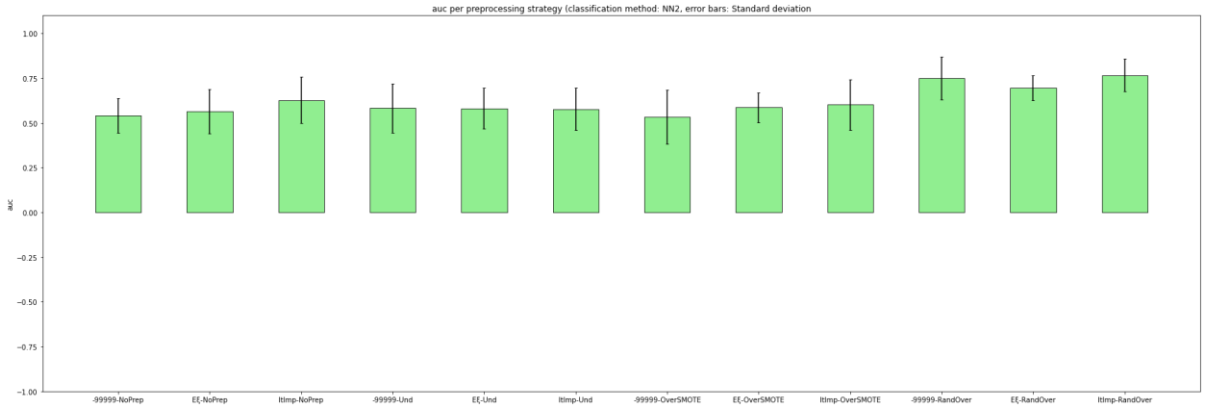

auc per preprocessing strategy (classification method: DT3, error bars: Standard deviation
auc
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

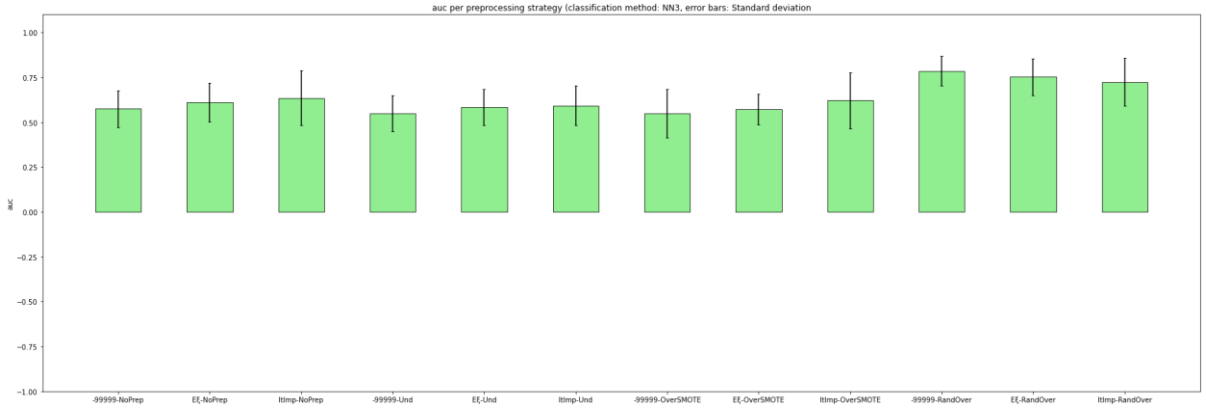

auc per preprocessing strategy (classification method: DT4, error bars: Standard deviation
auc
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

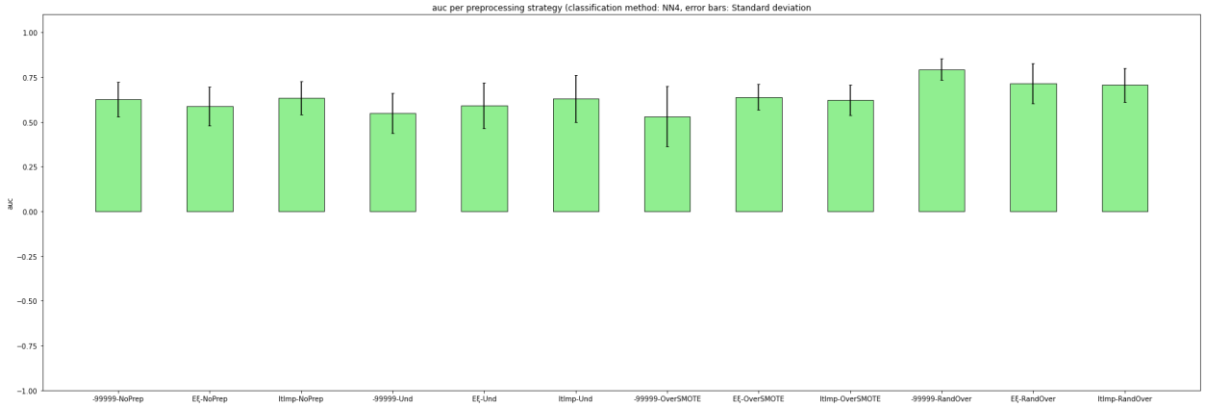

auc per preprocessing strategy (classification method: RF1, error bars: Standard deviation
auc
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

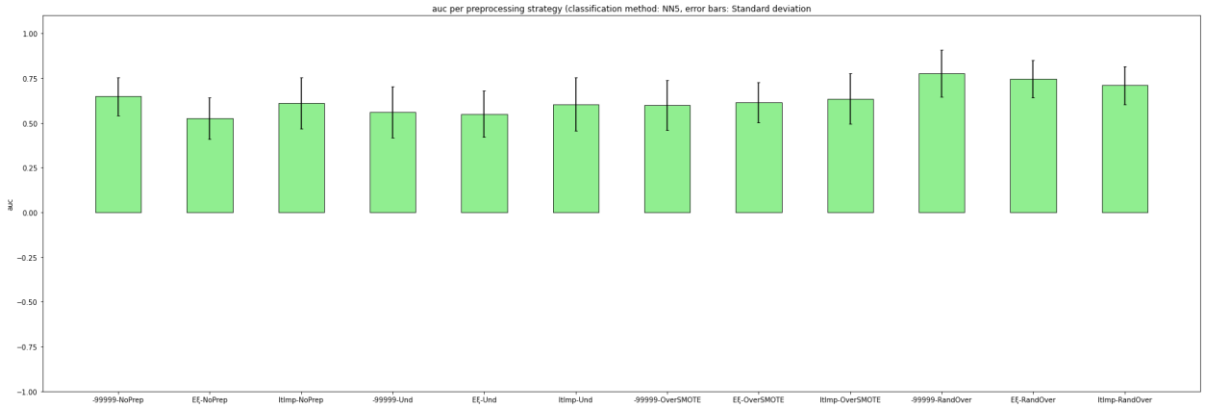

auc per preprocessing strategy (classification method: RF2, error bars: Standard deviation
auc
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-CIWeigh
EE-CIWeigh
ItImp-CIWeigh
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

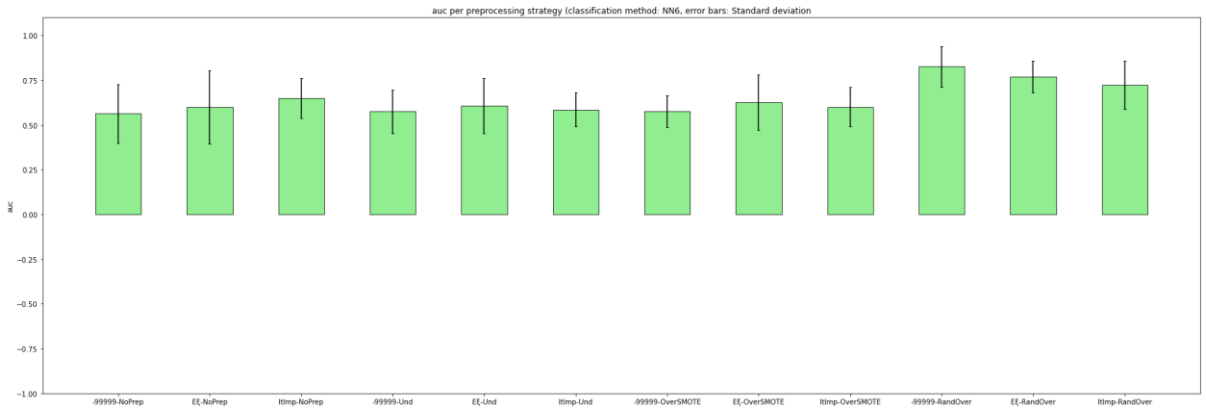

auc per preprocessing strategy (classification method: NN1, error bars: Standard deviation
auc
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

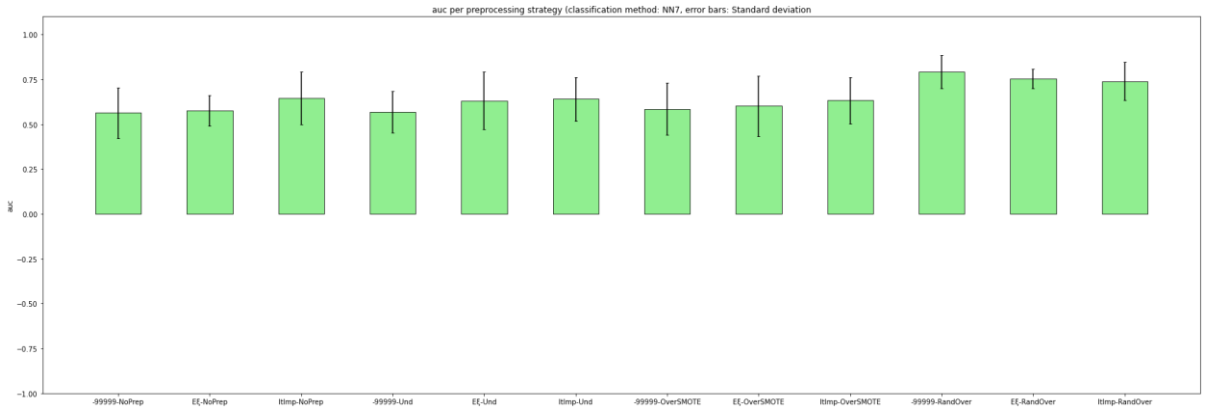

auc per preprocessing strategy (classification method: NN2, error bars: Standard deviation
auc
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

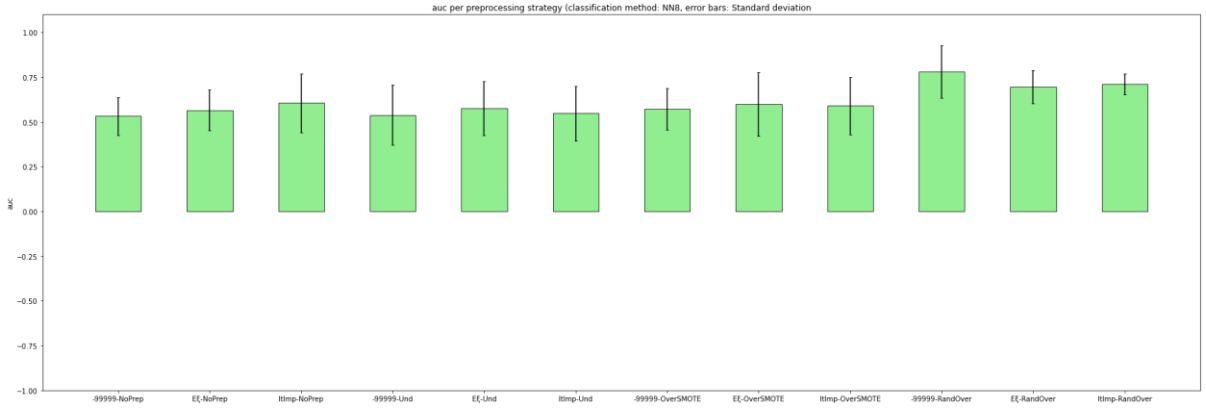

auc per preprocessing strategy (classification method: NN3, error bars: Standard deviation
auc
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

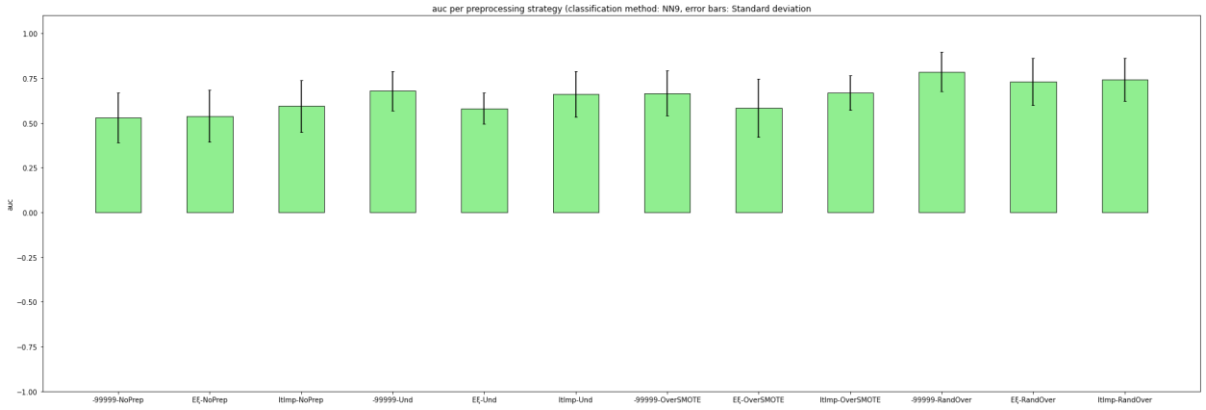

auc per preprocessing strategy (classification method: NN4, error bars: Standard deviation
auc
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

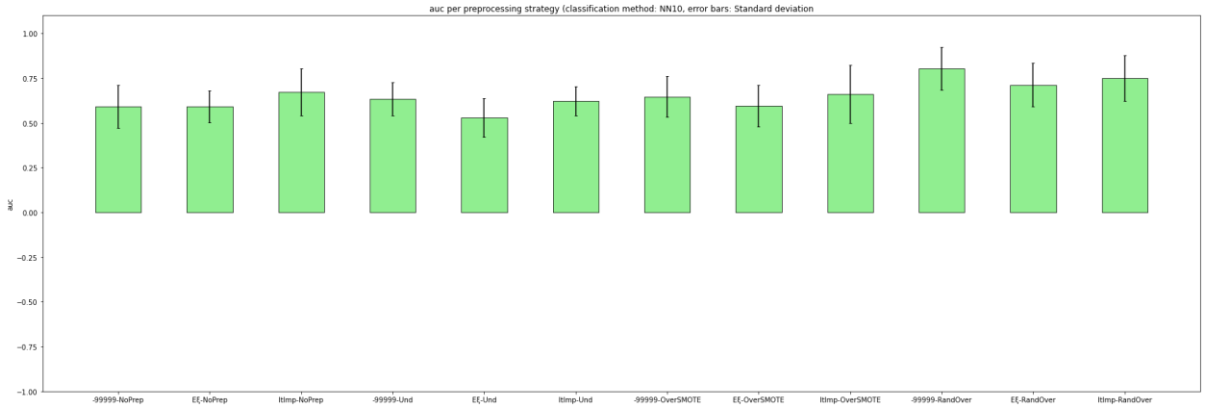

auc per preprocessing strategy (classification method: NN5, error bars: Standard deviation
auc
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

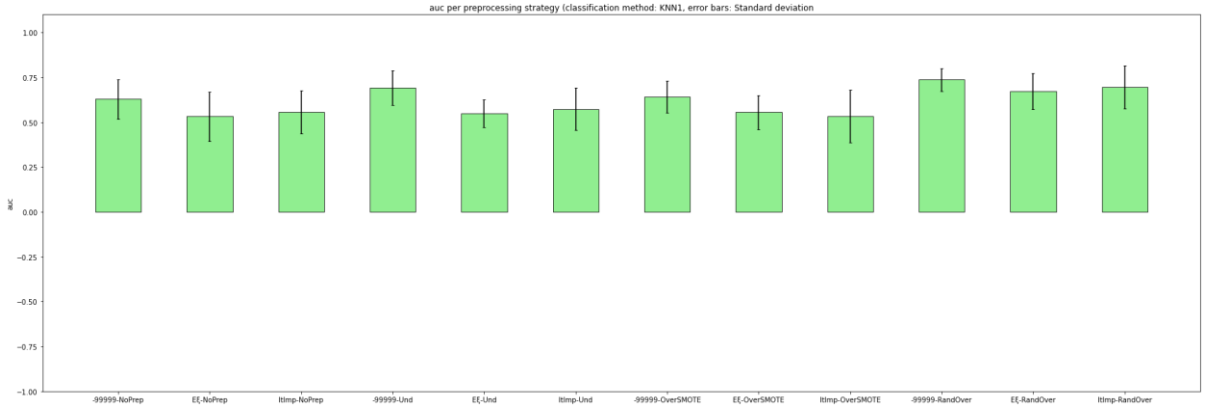

auc per preprocessing strategy (classification method: NN6, error bars: Standard deviation
auc
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

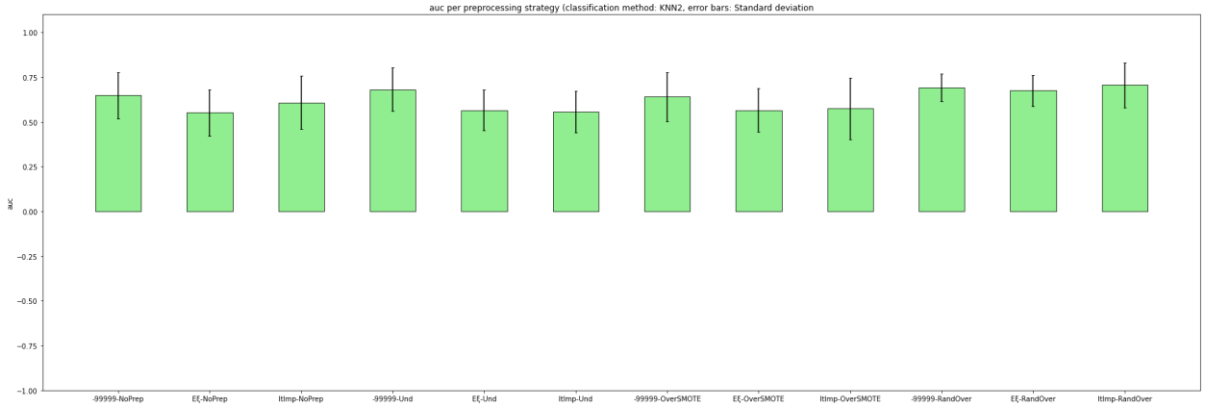

auc per preprocessing strategy (classification method: NN7, error bars: Standard deviation
auc
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

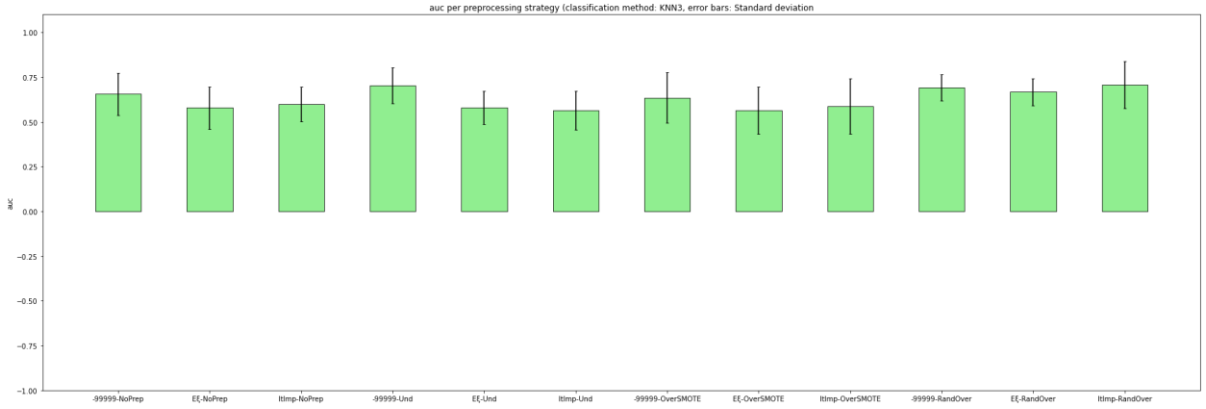

auc per preprocessing strategy (classification method: NN8, error bars: Standard deviation
auc
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

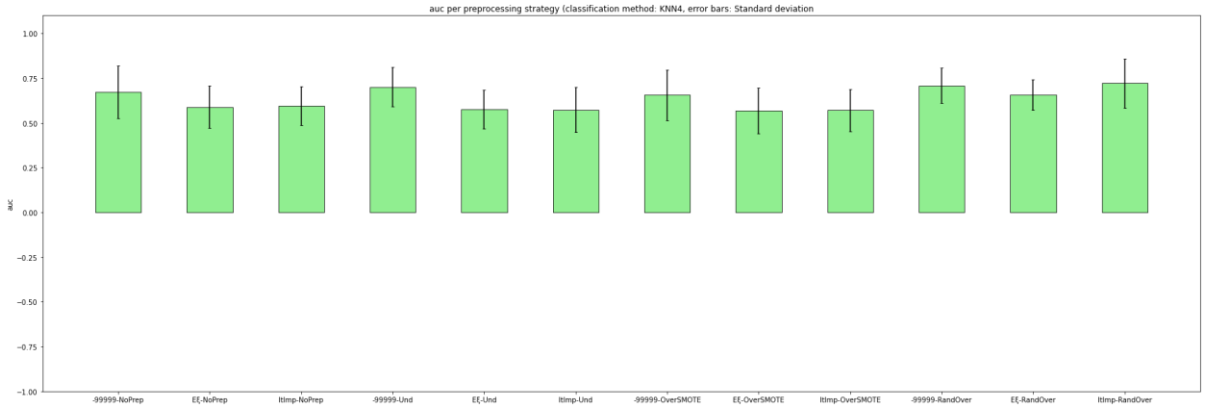

auc per preprocessing strategy (classification method: NN9, error bars: Standard deviation
auc
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

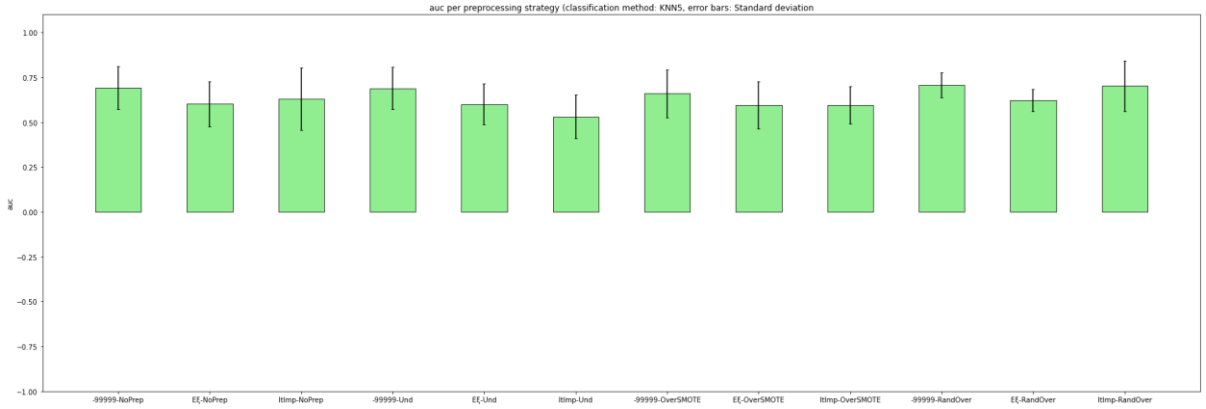

auc per preprocessing strategy (classification method: NN10, error bars: Standard deviation
auc
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

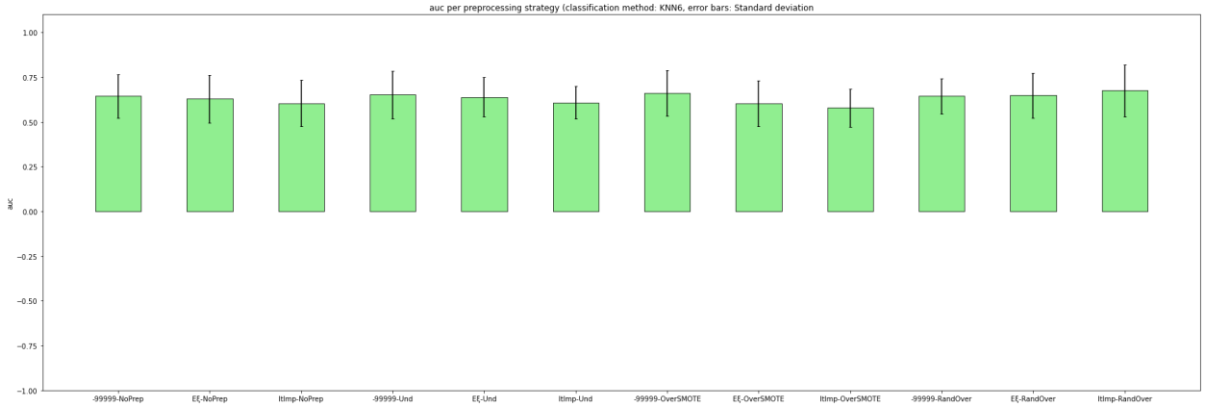

auc per preprocessing strategy (classification method: KNN1, error bars: Standard deviation
auc
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

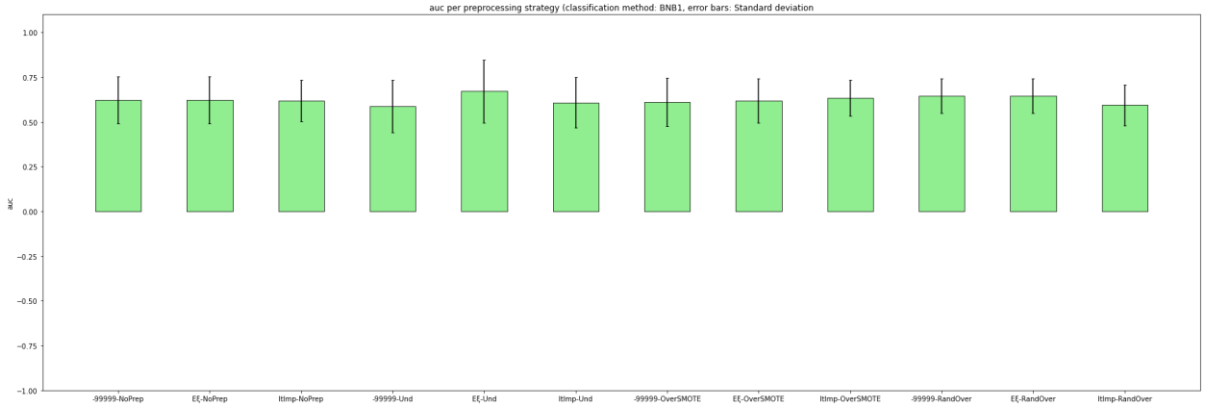

auc per preprocessing strategy (classification method: KNN2, error bars: Standard deviation
auc
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

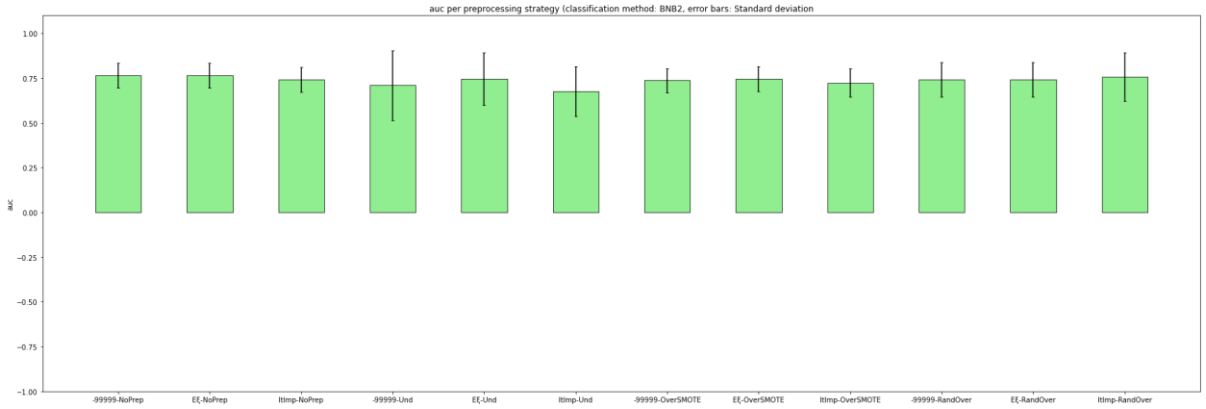

auc per preprocessing strategy (classification method: KNN3, error bars: Standard deviation
auc
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver

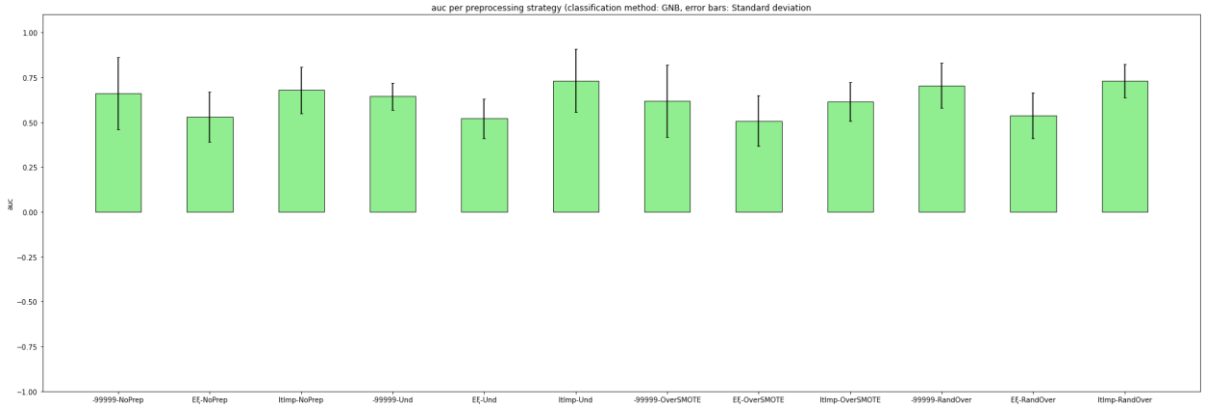

auc per preprocessing strategy (classification method: KNN4, error bars: Standard deviation
auc
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver


auc per preprocessing strategy (classification method: KNN5, error bars: Standard deviation
auc
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver


auc per preprocessing strategy (classification method: KNN6, error bars: Standard deviation
auc
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver


auc per preprocessing strategy (classification method: BNB1, error bars: Standard deviation
auc
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver


auc per preprocessing strategy (classification method: BNB2, error bars: Standard deviation
auc
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver


auc per preprocessing strategy (classification method: GNB, error bars: Standard deviation
auc
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
-99999-NoPrep
EE-NoPrep
ItImp-NoPrep
-99999-Und
EE-Und
ItImp-Und
-99999-OverSMOTE
EE-OverSMOTE
ItImp-OverSMOTE
-99999-RandOver
EE-RandOver
ItImp-RandOver