UNIVERZA V MARIBORU

FAKULTETA ZA ZDRAVSTVENE VEDE

GRADNJA NAPOVEDNIH MODELOV S POMOČJO STRUKTURIRANIH IN NESTRUKTURIRANIH PODATKOVNIH VIROV

(Magistrsko delo)

Maribor, 2017 Leon Kopitar

UNIVERZA V MARIBORU

FAKULTETA ZA ZDRAVSTVENE VEDE

GRADNJA NAPOVEDNIH MODELOV S POMOČJO STRUKTURIRANIH IN NESTRUKTURIRANIH PODATKOVNIH VIROV

(Magistrsko delo)

Maribor, 2017 Leon Kopitar

UNIVERZA V MARIBORU

FAKULTETA ZA ZDRAVSTVENE VEDE

Mentor: izr. prof. dr. Gregor Štiglic

Somentor: doc. dr. Andraž Stožer

ZAHVALA

Velika zahvala gre mentorju izr. prof. dr. Gregorju Štiglicu za potrpežljivo, odzivno in sprotno usmerjanje pri izdelavi zaključne naloge.

Prav tako se zahvaljujem somentorju doc. dr. Andražu Stožerju pri svetovanju in podajanju strokovnih popravkov.

Zahvaljujem se tudi družini, ki mi je nudila neizmerno podporo skozi vsa leta študija.

GRADNJA NAPOVEDNIH MODELOV S POMOČJO STRUKTURIRANIH IN NESTRUKTURIRANIH PODATKOVNIH VIROV

POVZETEK

Teoretična izhodišča: Sladkorna bolezen tipa 2 (SB2) je najpogostejša oblika sladkorne bolezni, predvsem v razvitih državah sveta. Za SB2 zboleva vedno več ljudi, in to zaradi neprimernega življenjskega stila, predvsem premalo fizične dejavnosti in nepravilnega prehranjevanja. Čeprav večina ljudi SB2 vidi kot samoumevno bolezen, ki se lahko pojavi v poznih letih, se mnogi ne zavedajo njene resnosti. SB2 predstavlja glavni vzrok za možgansko kap in bolezni srca. Poleg tega lahko privede do slepote, bolezni ledvic oziroma, v skrajnem primeru, tudi do smrti. S starostjo se tveganje za SB2 razumljivo povečuje, vendar pa lahko v veliki meri na povečanje tveganja vplivamo predvsem sami. Smrtnemu izidu so najbolj podvrženi bolniki s SB2, ki so bili hospitalizirani na enoti intenzivnega oddelka. Glavni namen magistrskega dela je bil preveriti vpliv najpogosteje ponavljajočih se korenov besed iz zapisov o zdravljenju bolnika na točnost napovednega modela za napoved preživetja bolnikov s SB2.

Metodologija raziskovanja: Analize smo opravili na filtrirani podatkovni zbirki MIMIC-III, ki hrani skupno 4236 zapisov o bolnikih s SB2. Analize so bile izvedene s programskim jezikom R s pomočjo naslednjih klasifikatorjev: Random Forest, Single C5.0 Ruleset, Glmnet (Lasso regresija), XGBoost ter GBM. Rezultate smo evalvirali z Bootstrap metodo, ponovljeno 100-krat.

Rezultati: Vsi napovedni modeli, zgrajeni na podatkih moškega vzorca, so bili v primerjavi z modeli, zgrajenimi na podatkih ženskega vzorca, statistično signifikantno uspešnejši pri napovedovanju umrljivosti bolnikov s SB2 ($\Delta AUC = +0{,}049$, $p < 0{,}001$). Z uporabo bigramov se rezultati napovedne uspešnosti statistično ne razlikujejo ($p > 0{,}001$). Ne glede na spol se rezultati pri napovedovanju z vključenim kriterijem SAPS izboljšajo v primerjavi z napovedovanjem, če kriterij SAPS ni prisoten ($\Delta AUC_{Ženske} = +0{,}0756$, $\Delta AUC_{Moški} = +0{,}082$).

Sklep: Napovedni model XGBoost je najprimernejši model za napovedovanje umrljivosti bolnikov s SB2. Prisotnost besed, ki se navezujejo na stimulacijo oziroma

spodbujanje, starost, gibanje, neodzivnost in diagnozo intracerebralne krvavitve, ima največji vpliv na uspešno napovedovanje umrljivosti bolnikov s SB2. Z vključitvijo bigramov se uspešnost napovednih modelov ne izboljša signifikantno. Uporaba pogosto uporabljenega kriterija SAPS, ki temelji na fizioloških podatkih, ostaja primarno vodilo pri napovedovanju umrljivosti bolnikov s SB2.

Ključne besede: sladkorna bolezen tipa 2, napovedni modeli, zapisi medicinskih sester

PREDICTIVE MODELING USING STRUCTURED AND UNSTRUCTURED DATA

ABSTRACT

Theoretical basis: Type 2 diabetes mellitus (T2DM) is the most common form of diabetes, especially in developed countries around the world. More and more people are getting T2DM due to an unadapted lifestyle characterized by physical inactivity and an excessive caloric intake. Although most people see T2DMM as a self-evident illness that can occur in older age, many are unaware of its severity. T2DM is the main cause of stroke and heart disease. In addition, it can lead to blindness, kidney disease or ultimately to death. With age, the risk for T2D is rising, but we can, to a large extent, influence the increase in risk through our own life choices. The main purpose of this paper was to examine the impact of the most commonly-repeated words from the nursing notes on the accuracy of the predictive model for predicting the survival of patients with T2DM.

Research methodology: The analyses were carried out on a filtered MIMIC-III database consisting a total of 4236 records of patients with T2D. The analyses were performed with the programming language R by using the following classifiers: Random Forest, Single C5.0 Ruleset, Glmnet (Lasso regression), XGBoost, and GBM. The results were evaluated with the Bootstrap method, repeated 100 times.

Results: All predictive models built on male sample data were statistically significantly more successful in predicting the mortality of patients with T2DMM in comparison with models built on female sample data. By using bigrams, the results of predictive performance were not statistically different ($p > 0,001$). Regardless of gender, results of predictions including the SAPS criterion were better than results of predictions without the SAPS criterion ($\Delta AUC_{Females} = +0,0756$, $\Delta AUC_{Males} = +0,082$).

Conclusion: Results show that XGBoost predictive model is the most appropriate model for predicting mortality of patients with T2DM. The presence of words related to stimulation, age, movement, unresponsive and diagnosis of intracerebral haemorrhage have the greatest impact in the successful predictability of mortality of

patients with T2D. By including bigrams, the performance of predictive models does not significantly improve.

Keywords: Type 2 Diabetes, predictive modelling, nursing notes

KAZALO VSEBINE

KAZALO TABEL

KAZALO SLIK

SEZNAM KRATIC

KRATICA	POLN POMEN	SLOVENSKI PREVOD
SB2/SB	Sladkorna bolezen tipa 2/Sladkorna bolezen	
LDL	low density lipoprotein	Lipoproteini z nizko gostoto
PHRP	Protecting Human Research Participants	Varovanje človekovih pravic v raziskavah
SAPS	Simplified Acute Physiology Score	Vrednost poenostavljene ocene akutne fiziologije
ID	Identification Device	Unikatni identifikator
MIMIC-III	Medical Information Mart for Intensive Care III	Zbirka medicinskih informacij za intenzivno oskrbo III
ZDA	Združene države Amerike	
AUC	The area under the curve	Površina pod krivuljo
SENS	Sensitivity	Senzitivnost
SPEC	Specificity	Specifičnost
ppv	Positive predictive value	Pozitivna napovedna vrednost
npv	Negative predictive value	Negativna napovedna vrednost
ROC	Receiver Operating Characteristic	Karakteristike delovanja sprejemnika
Glmnet	Lasso and Elastic-Net Regularized Generalized Linear Models	Lasso in Elastic-Net regularizirani splošni

		linearni modeli
GBM	Gradient boosting machine	
RF	Random forest	Naključni gozdovi
XGBoost	Extreme Gradient Boosting	

1 Uvod in opis problema

Sladkorna bolezen tipa 2 (SB2) je najpogostejša oblika sladkorne bolezni (90–95 %) (American Diabetes Association [ADA], 2014) in je prisotna povsod, predvsem v razvitih državah sveta (Rubino, 2008).

V letu 2015 je bilo diagnosticiranih okoli 415 milijonov pacientov, starih med 20 in 79 let, do leta 2040 pa bi se število obolelih lahko povzpelo na 642 milijonov obolelih za SB2 (Ogurtsova, et al., 2017). V Sloveniji naj bi bilo po ocenah skupno približno 168.200 sladkornih bolnikov (104.700 diagnosticiranih; 63.500 nediagnosticiranih) (International Diabetes Federation [IDF], 2015).

SB2 predstavlja glavni vzrok za možgansko kap in bolezni srca (Phillips, et al., 2014), bolniki s SB2 pa imajo kar trikrat večje tveganje za razvoj tuberkuloze. Poleg tega SB2 privede do zmanjšanja telesne odpornosti organizma proti okužbam, težav z vidom (slepota) in bolezni ledvic (Jeon & Murray, 2008). Incidenca unilateralne transmetatarzalne amputacije spodnjih okončin se pri bolnikih (starih 45 in več let) poveča za več kot osemkrat (Johannesson, et al., 2009).

V raziskavi, ki so jo izvedli McEwen, et al. (2012) so raziskovali faktorje, ki vplivajo na smrt posameznikov s SB2 v ZDA na populaciji oseb, starejših od 18 let, ki so imele status bolnika s SB2 vsaj eno leto. Izkazalo se je, da imajo na umrljivost največji vpliv višja starost, moški spol, bela rasa, nižji finančni prihodki, kajenje, zdravljenje z inzulinom, nefropatija, zgodovina dislipidemije, višja vrednost holesterola LDL, angina/miokardni infarkt in ostale koronarne bolezni, koronarna angioplastika, postopno srčno popuščanje, aspirin, zaviralci beta, uporaba diuretikov in visok Charlsonov indeks komorbidnosti (McEwen, et al., 2012). Poleg že naštetih komplikacij, ki so bile ugotovljene v okviru klasičnih kliničnih študij, je bilo med drugim smiselno raziskati, kateri dejavniki pripomorejo k boljši napovedni točnosti pri napovedovanju umrljivosti bolnikov s SB2 tudi na podlagi vse pogosteje uporabljenih elektronskih zapisov o pacientih.

Za merjenje življenjske ogroženosti pacienta po sprejetju na intenzivni oddelek je bil razvit točkovni sistem SAPS (Le Gall, et al., 1984). Določen je na podlagi dvanajstih fizioloških meritev v prvih 24 urah po sprejemu na oddelek intenzivne nege, informacij o preteklem zdravstvenem stanju in nekaterih drugih informacij,

pridobljenih ob sprejemu (Marafino, et al., 2015). V obstoječi študiji so na podlagi točkovnega sistema SAPS v prvih 24 urah dosegli le 79,1-odstotno uspešnost napovedi (Marafino, et al., 2015), kar smo poskušali v naši študiji izboljšati. Zato smo poizkusili ugotoviti, kolikšno napovedno uspešnost lahko dosežemo na podlagi podatkov, ki se zberejo po prvih šestih urah. Poleg tega smo si v magistrskem delu prizadevali prikazati, kakšna je dejanska učinkovitost takšnega točkovnega sistema v primerjavi z modernimi napovednimi modeli, ki jih lahko zgradimo na podlagi informacij iz zapisov o pacientu. Vse raziskave so bile opravljene le na bolnikih s SB2.

2 Namen in cilji raziskave

Namen magistrskega dela je bil preveriti vpliv najpogosteje ponavljajočih se korenov besed iz zapisov o zdravljenju bolnika na točnost napovednega modela za napoved preživetja bolnikov s SB2. Na ta način lahko gradimo uspešnejše napovedne modele in pridobimo dragocene informacije, ki lahko pripomorejo k nadaljnjim odločitvam pri zdravljenju bolnikov.

Cilji magistrskega dela:

- teoretični:
  - pregled literature s področja gradnje napovednih modelov za napoved umrljivosti bolnikov s SB2;

- empirični:
  - izdelava in primerjava napovednih modelov za napoved (klasifikacija) umrljivosti bolnikov s SB2, ki se zdravijo na oddelku intenzivne nege,
  - prikazati, kateri najpogosteje ponavljajoči se koreni besed najbolj vplivajo na preživetje bolnika s SB2 v 24 urah po sprejemu na intenzivni oddelek glede na spol in tip spremenljivke (unigram/bigram).

3 Teoretični del

3.1 Napovedno modeliranje

Napovedno modeliranje je definirano kot statistični model ali kot model strojnega učenja, ki se uporablja za predvidevanje vedenja v prihodnosti glede na preteklost. Napovedni modeli lahko vsebujejo enega ali več klasifikatorjev za določanje verjetnosti, da nabor podatkov pripada enemu ali drugemu razredu (Strickland, 2015).

Napovedno modeliranje obsega tri vrste napovednih modelov: modeli nagnjenosti (napovedi), modeli gručenja (segmenti), sodelovalno filtriranje (priporočila) (Strickland, 2015).

V nalogi so bili uporabljeni le modeli nagnjenosti, s pomočjo katerih so bile izračunane napovedi preživetja bolnikov s SB2.

3.2 Metrike uspešnosti klasifikatorja

Za primerjavo rezultatov napovedi preživetja bolnikov na oddelku intenzivne nege smo uporabili pet vrst klasifikatorjev: *Single C5.0 Ruleset, Random Forest, glmnet* (*Lasso*), *XGBoost* in *GBM*. Rezultate smo primerjali na podlagi metrike AUC (Območje pod krivuljo) (Praprott, et al., 2016). Krivulja ROC je mera, ki prikazuje uspešnost klasifikatorjev oziroma napovednih modelov. Prikazana je na dvodimenzionalnem grafu z začetno točko (0, 0) na skrajni levi spodnji strani grafa, poteka pa v diagonalni smeri vse do točke (1, 1). Območje pod krivuljo ROC se v skrajšani obliki običajno poimenuje kar AUC. Vrednost metrike AUC je omejena na interval [0, 1], pri čemer so klasifikatorji z vrednostjo, manjšo od 0,5, ki predstavlja vrednost naključne izbire, praktično neuporabni (Fawcett, 2006). Za evalvacijo modelov so v preteklosti pokazali, da se metrika AUC izkaže za boljšo mero kot mera točnosti (Huang & Ling, 2005).

4 Empirični del

4.1 Raziskovalna vprašanja in hipoteze

H1: Uspešnost napovednih modelov se bistveno razlikuje glede na spol.

H2: Uporaba bigramov skupaj z unigrami izboljša napovedno uspešnost modelov v primerjavi z uporabo izključno unigramov.

RV1: Ali je možno zgraditi napovedni model, ki bo na podlagi nestrukturiranih podatkov, zbranih v prvih šestih urah hospitalizacije na enoti intenzivnega oddelka, bolj uspešen od pogosto uporabljenega kriterija SAPS, ki je izračunan na podlagi točno določenih podatkov, zbranih v prvih 24 urah hospitalizacije?

4.2 Metodologija

4.2.1 Raziskovalne metode

Za napovedovanje uspešnosti preživetja bomo uporabili naslednjih pet pogosto uporabljenih vrst napovednih modelov: *Random Forest* (Casanova, et al., 2014), *Single C5.0 Ruleset* (Shao, et al., 2015), *Glmnet* (Lasso regresija) (Friedman, et al., 2010), *XGBoost* (Chen, et al., 2016), ter *GBM* (Fu, et al., 2015). Vsi napovedni modeli bodo zgrajeni in evalvirani s pomočjo *programskega jezika R* (R Development Core Team, 2017) in paketa *Caret* (Kuhn, et al., 2017).

V nadaljevanju podajamo krajše opise značilnosti petih vrst klasifikatorjev, ki so bili uporabljeni za gradnjo napovednih modelov v okviru magistrskega dela:

Random Forest ali naključni gozd je klasifikacijska ansambelska metoda, kjer se hkrati uporabi več regresijskih ali odločitvenih dreves. Vsako drevo napove vrednost odločitvenega atributa ali izbere klasifikacijo oziroma glasuje za določen razred. Na podlagi glasov algoritem izbere najbolj primerno rešitev (Casanova, et al., 2014).

Single C5.0 Ruleset je izboljšani drevesni klasifikator algoritma C4.5. Algoritem C5.0 na vsakem vozlišču odločitvenega drevesa izbere atribut, ki najbolj učinkovito razdeli skupino vzorcev podatkov v podskupine glede na izbrano evalvacijsko metriko. Posebnost klasifikatorjev s pravili (ruleset) je ta, da že vsebuje neurejene zbirke preprostih pravil "if-then" (Shao, et al., 2015).

Glmnet je posplošeni linearni model z vključenim kaznovanjem (*angl. penalized maximum likehood*). Velja za zelo hiter algoritem, ki poleg dobrih napovednih lastnosti omogoča tudi izbiro atributov že med samo gradnjo klasifikatorja (Friedman, et al., 2010).

XGBoost (Extreme Gradient Boosting) je metoda, ki omogoča vzporedno gradnjo večjega števila odločitvenih dreves na eni sami napravi, kar ji omogoča tudi do več kot 10-krat višjo časovno učinkovitost od obstoječih implementacij Gradient Boosting metode (Chen, et al., 2016).

GBM (Gradient Boosting Machine) je zelo prilagodljiva metoda, saj njen postopek učenja neprestano prilagaja nove modele z namenom, da zagotovi natančno oceno vrednosti odzivne spremenljivke (Natekin & Knoll, 2013). Končni klasifikator je tako sestavljen iz ansambla posameznih napovednih modelov, običajno v obliki odločitvenih dreves.

4.2.2 Raziskovalno okolje

V namen zaključne naloge smo uporabili podatkovno zbirko MIMIC-III (Johnson, et al., 2016). Ta hrani klinične podatke pacientov, ki so prosto dostopni raziskovalcem po vsem svetu. MIMIC-III je edina prosto dostopna zbirka s tako širokim naborom podatkov o pacientih, poleg tega pa se nabor podatkov razteza čez več kot desetletje (2001–2012). Podatki so bili pridobljeni v medicinskem centru Beth Israel Deaconess v Bostonu na različnih oddelkih intenzivne nege. Skupno vsebuje MIMIC-III 53423 različnih zapisov o hospitalizacijah pacientov, starih 16 in več let, ki so bili sprejeti na oddelek intenzivne nege. Povprečna starost pacientov je 65,8 let, pri čemer je 55,9 % pacientov moškega spola. Od vseh zapisov jih 11,5 % predstavlja hospitalizacijo s smrtnim izidom. Podatkovna zbirka je sestavljena iz 26 tabel, ki so med seboj povezane z unikatnim identifikatorjem (ID) (Johnson, et al., 2016). Podatkovna zbirka MIMIC-III je unikatna tudi v tem, da poleg podatkov v strukturirani obliki vsebuje tudi tekstovne oziroma t. i. nestrukturirane podatke. To v praksi pomeni, da imamo za vsako hospitalizacijo na voljo tudi zapise medicinskih sester, ki so bile zadolžene za nego posameznega pacienta. Iz takšnih zapisov lahko izločimo ključne besede oz. korene le-teh, ki predstavljajo pomembno informacijo

pri gradnji napovednih modelov. Poleg t. i. unigramov, ki predstavljajo koren posamezne besede, v besedilu poznamo tudi bigrame, kjer iz besedila izločimo različne besedne zveze oz. korena dveh sosednjih besed. Primer unigrama bi lahko bil unigram *monitor*, ki je lahko koren angleške besede *monitorized* (v slovenščini spremljan/a), poleg tega je lahko koren besede *monitoring* (v slovenščini spremljanje) in podobno. V primeru uporabe bigramov, ki ga sestavljata korena dveh povezanih besed, bi bigram *resp_care* lahko v angleškem jeziku predstavljal besedno zvezo *respond carefully* (v slovenskem jeziku *pazljivo odgovorite*) ali besedno zvezo *responsive care* (v slovenščini *odgovorna oskrba*).

4.2.3 Raziskovalni vzorec

Podatkovna zbirka, na kateri smo opravili analize, je bila pridobljena iz prosto dostopne podatkovne zbirke MIMIC-III. Zbirka je modificirana in filtrirana tako, da vsebuje le bolnike s SB2. Poleg tega so bile podatkovni zbirki dodane spremenljivke najpogostejših unigramov in bigramov v zapisih o bolnikih, ki smo jih izločili iz prostega besedila.

Podatkovna zbirka tako vsebuje skupno 4236 primerkov (t. j. hospitalizacij bolnikov s SB2) in 1058 spremenljivk, od teh sta dve osnovni (starost in spol), petdeset (50) spremenljivk, ki predstavljajo najbolj pogoste diagnoze po sprejetju na intenzivni oddelek, tisoč pet (1005) spremenljivk, ki predstavljajo najpogosteje ponavljajoče se korene besed, zapisanih v anamnezi, ter izhodna spremenljivka, ki predstavlja, ali je bolnik preživel hospitalizacijo. V podatkovni zbirki najvišji delež bolnikov predstavljajo moški (59 %). Povprečna starost moških, ki znaša 66,22 (95 % IZ: 65,75–66,69) let, je za približno tri leta nižja kot povprečna starost žensk, ki znaša 69,42 (95 % IZ: 68,86–69,98) let. Razmerje med razredoma izhodne spremenljivke, ki predstavlja umrljivost, je neuravnoteženo (približno 90 % preživelih proti 10 % umrlih). Enako neuravnoteženo razmerje je prisotno tudi znotraj posameznega spola.

Za evalvacijo rezultatov oz. napovedne uspešnosti modelov smo uporabili metodo *Bootstrap* s 100 ponovitvami (Efron & Tibshirani, 1994). Vse analize so bile izvedene s pomočjo *programskega jezika R*. Napovedni modeli pa so bili zgrajeni in

ovrednoteni s pomočjo knjižnice *Caret*. Pomembnost spremenljivk je bila določena s pomočjo metode *varImp*, ki je del knjižice *Caret*.

4.2.4 Postopki zbiranja podatkov

V raziskovalni nalogi so bili uporabljeni le podatki podatkovne zbirke MIMIC-III, ki je prosto dostopna za raziskovalne namene (Johnson, et al., 2016).

4.2.5 Etični vidik

Vsi uporabljeni podatki podatkovne zbirke MIMIC-III so prosto dostopni in anonimizirani, zato so vse raziskave etično sprejemljive, podatkovna zbirka pa je takšno mnenje pridobila tudi s strani pristojnih organov na Beth Israel Deaconess Medical Center, Boston, Massachusetts. Za dostop do podatkovne zbirke je predhodno potrebno opraviti usposabljanje PHRP (Protecting Human Research Participants) (National Institutes of Health, 2014). Podatkovna zbirka MIMIC-III je za raziskovalne namene na voljo brezplačno (Johnson, et al., 2016).

4.2.6 Predpostavki in omejitve raziskave

Omejitve raziskave:

- časovna zahtevnost evalvacije bolj kompleksnih klasifikatorjev, kjer je potrebno računati z nekoliko večjo varianco rezultatov,
- določeni zapisi ne vsebujejo tekstovnih oziroma nestrukturiranih podatkov, zato v študiji niso bili uporabljeni.

5 Rezultati

Vse primerjave napovednih modelov so bile izvedene s pomočjo metrike AUC, spremljali pa smo tudi senzitivnost in specifičnost pri klasifikaciji testnih primerkov. Poleg dejanskih vrednosti pomembnosti spremenljivk (ime stolpca *Vrednost*) so dejanske vrednosti prikazane tudi v intervalni obliki (od 0 do 100), kjer je spremenljivka z najnižjo vrednostjo ovrednotena z vrednostjo 0, spremenljivka z najvišjo pa s 100 (ime stolpca *Vred_scal*).

Pri obravnavanju najuspešnejših spremenljivk posameznega napovednega modela so bile zavzete le spremenljivke oziroma koreni besed, katerih vrednosti intervalne oblike zavzamejo vrednost večjo ali enako 50.

5.1 Primerjava napovednih modelov glede na spol

V podpoglavjih (5.1.1 in 5.1.2) so predstavljeni rezultati posebej za napovedne modele, zgrajene tako na podatkih ženskega kot tudi moškega vzorca. V sledečem podpoglavju (5.1.3) je prikazana primerjava rezultatov obeh vzorcev ter izpis najpomembnejših spremenljivk za najboljše tri napovedne modele.

5.1.1 Rezultati – ženski spol

Zanimalo nas je, s katerim napovednim modelom dosežemo najboljše rezultate na podatkih ženskega vzorca (Tabela 1).

Tabela 1: Napovedne uspešnosti modelov, zgrajenih na podatkih ženskega vzorca

Model	AUC	PPV	NPV
C5.0Ruleset	0,583 0,576–0,591	0,907 0,907–0,908	0,287 0,269–0,305
GBM	**0,712** 0,707–0,716	0,909 0,908 - 0,909	0,399 0,380–0,418
Glmnet	**0,711** 0,706–0,716	0,908 0,907–0,908	0,333 0,316–0,350
XGBoost	**0,718** 0,714–0,721	0,907 0,907–0,908	0,452 0,425–0,480
RF	0,678 0,673–0,683	0,901 0,901–0,901	0,089 0,050–0,129

Napovedni model *XGBoost* je med obravnavanimi modeli dosegel najvišjo napovedno uspešnost (AUC = 0,718). Po uspešnosti sta mu najbližje *GBM* in *Glmnet* s 71,1-odstotno oziroma 71,2-odstotno napovedno uspešnostjo. Po kriteriju AUC se je najslabše odrezal napovedni model *C5.0Ruleset* (AUC = 0,583) (Tabela 1).

5.1.2 Rezultati – moški spol

Preverili smo, s katerim napovednim modelom dosežemo najboljše rezultate na podatkih moškega vzorca (Tabela 2).

Tabela 2: Napovedne uspešnosti modelov, zgrajenih na podatkih moškega vzorca

Model	AUC	PPV	NPV
C5.0Ruleset	0,605 0,597–0,613	0,907 0,907–0,908	0,318 0,303–0,333
GBM	**0,766** 0,762–0,769	0,911 0,911–0,912	0,470 0,455–0,485
Glmnet	**0,760** 0,756–0,763	0,904 0,903–0,905	0,336 0,316–0,356
XGBoost	**0,770** 0,767–0,774	0,908 0,908–0,909	0,480 0,464–0,496
RF	0,748 0,745–0,752	0,900 0,900–0,900	0,292 0,235–0,349

Pri modelih, zgrajenih na podatkih moškega vzorca, je najboljšo napovedno uspešnost dosegel napovedni model *XGBoost* (AUC = 0,770). Takoj za njim sta se po kriteriju AUC uvrstila *GBM* (AUC = 0,766) ter *Glmnet* (AUC = 0,760), najslabšo napovedno uspešnost pa je dosegel *C5.0Ruleset* (AUC = 0,605) (Tabela 2).

5.1.3 Primerjava med ženskim in moškim spolom

Iz tabel 1 in 2 je razvidno, da je bila pri obeh spolih dosežena najboljša napovedna uspešnost z modelom *XGBoost* ($AUC_{ženske} = 0,718$; $AUC_{moški} = 0,770$), sledita mu *GBM* ($AUC_{ženske} = 0,712$; $AUC_{moški} = 0,766$) in *Glmnet* ($AUC_{ženske} = 0,711$; $AUC_{moški} = 0,760$).

Najpomembnejše spremenljivke napovednega modela so tiste spremenljivke, ki imajo največji vpliv na napovedno uspešnost. V spodnji tabeli so prikazane najpomembnejše spremenljivke glede na spol za model *GBM* (Tabela 3).

Tabela 3: Prikaz najpomembnejših spremenljivk glede na spol (model GBM)

GBM						
Ženski vzorec				***Moški vzorec***		
Spremenljivke	***Vrednost***	***Vred_scal***		***Spremenljivke***	***Vrednost***	***Vred_scal***
u_stimuli	7,54	100,00		u_movement	7,59	100,00
AGE	6,41	84,95		u_stimuli	6,40	84,33
u_famili	5,55	73,61		u_ac	6,38	84,12
u_dr	5,47	72,57		AGE	6,36	83,80
u_unrespons	4,54	60,26		u_levo	5,22	68,80
PD_431	3,30	43,73		u_unrespons	4,80	63,22
b_ns_bolus	2,86	37,89		u_central	4,37	57,56
PD_0389	2,70	35,73		PD_0389	3,91	51,55
u_pt	2,65	35,06		PD_431	3,56	46,99
u_movement	2,38	31,59		u_made	3,40	44,87
u_pupil	2,38	31,56		u_liver	3,35	44,13
u_today	2,34	31,00		u_famili	3,34	44,03
u_follow	2,19	29,07		u_dopamin	3,26	43,03
u_hypotens	2,08	27,51		u_creat	3,22	42,49
u_sp	1,98	26,24		u_monitor	2,95	38,90
u_head	1,83	24,21		u_levoph	2,95	38,86
u_levoph	1,78	23,57		u_ffp	2,91	38,36
u_plan	1,68	22,29		u_line	2,79	36,76
u_sat	1,67	22,16		u_coccyx	2,78	36,69
u_lactat	1,56	20,65		u_intubated	2,44	32,23

Pri ženskem vzorcu so to spremenljivke *u_stimuli*, *AGE*, *u_famili*, *u_dr*, *u_unrespons*, pri moškem pa *u_movement*, *u_stimuli*, *u_ac*, *AGE*, *u_levo*, *u_unrespons*, *u_central*, *PD_0389* (Tabela 3).

Podobno raziskavo najpomembnejših spremenljivk glede na spol smo naredili tudi za model *Glmnet* (Tabela 4).

Tabela 4: Prikaz najpomembnejših spremenljivk glede na spol (model Glmnet)

Glmnet						
Ženski vzorec				*Moški vzorec*		
Spremenljivke	***Vrednost***	***Vred_scal***		***Spremenljivke***	***Vrednost***	***Vred_scal***
PD_03849	1,7506	100,00		PD_431	0,8703	100,00
PD_41001	1,1577	66,13		PD_430	0,8583	98,62
PD_430	0,9801	55,99		PD_0389	0,7783	89,43
PD_431	0,9425	53,84		PD_5770	0,6962	80,00
PD_44101	0,7763	44,34		PD_0380	0,6692	76,89
PD_4412	0,7727	44,14		PD_41401	0,5492	63,10
PD_85220	0,7475	42,70		PD_5849	0,4976	57,18
PD_5770	0,7049	40,27		PD_03849	0,4287	49,26
PD_0389	0,6391	36,51		PD_51884	0,3736	42,93
PD_03811	0,2356	13,46		b_chr_Hyperten	0,2586	29,71
u_stimuli	0,2223	12,70		PD_5712	0,1907	21,91
PD_99859	0,217	12,40		u_stimuli	0,1788	20,54
PD_5715	0,2113	12,07		u_heme	0,1736	19,95
PD_41401	0,1957	11,18		PD_43491	0,168	19,30
PD_85221	0,1863	10,64		u_movement	0,1163	13,36
PD_41011	0,163	9,31		b_resp_care	0,1151	13,23
u_unrespons	0,1418	8,10		u_intubated	0,1086	12,48
u_wife	0,134	7,65		u_ac	0,1067	12,26
u_dr	0,1076	6,15		u_coars	0,1048	12,04
b_pt_start	0,0995	5,68		b_respiratori_care	0,1003	11,52

Pri modelu *Glmnet* se med prvimi devetimi najpomembnejšimi spremenljivkami (za oba vzorca) uvrščajo le diagnoze oziroma spremenljivke strukturiranih podatkov, kot so na primer intracerebralna krvavitev (*PD_431*), subarahnoidna krvavitev (*PD_430*), nespecifična septikemija (*PD_0389*) pri analizi, izvedeni na moškem vzorcu, in septikemija zaradi ostalih gram-negativnih organizmov (razen Hemophilus influenzae, Escherichia coli, Pseudomonas, Serratia, Gram-negativni organizem (nespecifična Gram-negativna septikemija NOS)) (*PD_03849*), akutni miokardni infarkt anterolateralne stene (1 ST elevacija) (*PD_41001*) in subarahnoidna krvavitev (*PD_430*) pri analizi, izvedeni na ženskem vzorcu. Pri ženskem vzorcu bi izmed nestrukturiranih podatkov bil pri napovedovanju umrljivosti najbolj uporaben koren stimuli (spremenljivka *u_stimuli*), pri moškem vzorcu pa bigram *chr_Hyperten* (spremenljivka *b_chr_Hyperten*) (Tabela 4).

Poleg omenjenih napovednih modelov smo opravili tudi raziskavo najpomembnejših spremenljivk glede na spol tudi za model *XGBoost* (Tabela 5).

Tabela 5: Prikaz najpomembnejših spremenljivk glede na spol (model XGBoost)

XGBoost						
Ženski vzorec				***Moški vzorec***		
Spremenljivke	***Vrednost***	***Vred_scal***		***Spremenljivke***	***Vrednost***	***Vred_scal***
u_stimuli	0,0362	100,00		u_ac	0,0284	100,00
u_famili	0,0259	71,55		u_movement	0,0241	84,86
AGE	0,0246	67,96		AGE	0,0221	77,82
u_dr	0,0167	46,13		u_stimuli	0,0181	63,73
u_hypotens	0,0153	42,27		u_levo	0,0119	41,90
u_diet	0,0149	41,16		u_central	0,0117	41,20
u_unrespons	0,0146	40,33		u_unrespons	0,0106	37,32
u_extub	0,0133	36,74		u_creat	0,0104	36,62
u_head	0,0117	32,32		u_famili	0,0104	36,62
u_movement	0,0115	31,77		u_made	0,01	35,21
u_today	0,0112	30,94		PD_41401	0,0097	34,15
u_pupil	0,0109	30,11		u_liver	0,0093	32,75
u_co	0,0096	26,52		PD_0389	0,0088	30,99
b_ns_bolus	0,0089	24,59		u_stabl	0,0084	29,58
u_sp	0,0087	24,03		u_co	0,0074	26,06
PD_0389	0,0075	20,72		u_ffp	0,0074	26,06
u_good	0,0071	19,61		u_pain	0,0073	25,70
u_cool	0,0068	18,78		u_line	0,0073	25,70
u_status	0,0066	18,23		u_levoph	0,0067	23,59
u_ra	0,0065	17,96		u_renal	0,0066	23,24

Pri napovednem modelu z najvišjo povprečno napovedno uspešnostjo *XGBoost* se je po pomembnosti najvišje uvrstila spremenljivka *u_stimuli*, sledita ji spremenljivki *u_famili* ter *AGE*. Pri moškem vzorcu pa so se med štiri najbolj pomembne spremenljivke prav tako uvrstile *u_ac*, *u_stimuli* ter spremenljivka *AGE*. Poleg naštetih se je na tretje mesto uvrstila spremenljivka *u_movement* (Tabela 5).

5.1.4 H1: Uspešnost napovednih modelov se bistveno razlikuje glede na spol.

Za določitev razmerja uspešnosti napovednih modelov, zgrajenih na podatkih moškega in ženskega vzorca, je bil uporabljen statistični test dveh neodvisnih spremenljivk (angl. *Independent Samples T-Test*), v nadaljevanju *T-test*.

Tabela 6: Rezultati statistično signifikantne razlike modelov glede na spol

Model	Podatkovna zbirka		P-vrednost	T	df
	Ženske	***Moški***			
C5.0Ruleset	0,583	**0,605**	< 0,001	-3.915	195,96
GBM	0,712	**0,766**	< 0,001	-19,07	188,64
Glmnet	0,711	**0,760**	< 0,001	-14,99	174,96
XGBoost	0,718	**0,770**	< 0,001	-20,41	197,89
RF	0,678	**0,748**	< 0,001	-23,07	177,13

Uspešnosti vseh napovednih modelov se na ravni rezultatov posameznega modela bistveno razlikujejo glede na spol ($p < 0,001$). Najvišja p-vrednost *T-testa* je bila določena pri napovednem modelu *C5.0Ruleset* ($p = 0,00012$), najnižja pa pri modelu *RF* ($p = 2,972e\text{-}55$) (Tabela 6).

5.2 Primerjava napovednih modelov glede na vključitev/izključitev bigramov

V podpoglavjih (5.2.1 in 5.2.2) so predstavljeni rezultati posebej za napovedne modele, zgrajene na podatkih, pri katerih so vključeni bigrami, kakor tudi rezultati napovednih modelov, zgrajenih na podatkih, kjer bigrami niso bili vključeni. V naslednjem podpoglavju (5.2.3) je prikazana primerjava rezultatov obeh zgoraj omenjenih analiz. Poleg tega so v podpoglavju 3 prikazane najpomembnejše spremenljivke za najboljše tri napovedne modele.

5.2.1 Rezultati – vključeni bigrami

Zanimalo nas je, s katerim napovednim modelom dosežemo najboljše rezultate na podatkih z vključenimi bigrami.

Tabela 7: Napovedne uspešnosti modelov, zgrajenih na podatkih z vključenimi bigrami

Model	AUC	PPV	NPV
C5.0Ruleset	0,527 0,523–0,531	0,904 0,903–0,904	0,428 0,398–0,45]
GBM	**0,771** 0,768–0,773	0,911 0,911–0,912	0,498 0,487–0,58]
Glmnet	**0,769** 0,766–0,771	0,905 0,904–0,905	0,428 0,411–0,446
XGBoost	**0,780** 0,777–0,782	0,909 0,909–0,910	0,547 0,532–0,563
RF	0,738 0,735–0,741	0,003 0,002–0,004	0,307 0,249–0,366

Med napovednimi modeli je najboljšo napovedno uspešnost dosegel napovedni model *XGBoost* (AUC = 0,780). Takoj za njim mu po napovedni uspešnosti sledita *GBM* (AUC = 0,771) ter *Glmnet* (AUC = 0,769). Najslabšo napovedno uspešnost smo dosegli z modelom *C5.0Ruleset* (AUC = 0,527) (Tabela 7).

5.2.2 Rezultati – izključeni bigrami

Preučili smo, s katerim napovednim modelom dosežemo najboljše rezultate na podatkih brez vključenih bigramov.

Tabela 8: Napovedne uspešnosti modelov, zgrajenih na podatkih brez vključenih bigramov

Model	AUC	PPV	NPV
C5.0Ruleset	0,527 0,524–0,531	0,904 0,903–0,904	0,423 0,396–0,451
GBM	**0,773** 0,771–0,775	0,912 0,911–0,912	0,514 0,502–0,526
Glmnet	**0,771** 0,768–0,773	0,904 0,904–0,950	0,430 0,411–0,448
XGBoost	**0,780** 0,777–0,783	0,910 0,909–0,910	0,553 0,538–0,569
RF	0,740 0,738–0,743	0,901 0,901–0,901	0,342 0,278–0,407

Med napovednimi modeli je najboljšo napovedno uspešnost dosegel napovedni model *XGBoost* (AUC = 0,780). Takoj za njim sta se po kriteriju AUC uvrstila *GBM*

(AUC = 0,773) ter *Glmnet* (AUC = 0,771), medtem ko se je napovedni model *C5.0Ruleset* (AUC = 0,527) uvrstil na zadnje mesto (Tabela 8).

5.2.3 Primerjava med zbirkama z vključenimi in izključenimi bigrami

Zanimale so nas razlike v napovedni uspešnosti glede na vključene oziroma izključene bigrame (Tabela 9).

Tabela 9: Primerjava napovednih uspešnosti glede na vključene oziroma izključene bigrame

Model	Podatkovna zbirka	AUC	SENS	SPEC
C5.0Ruleset	*bi_uni*	0,527 0,523–0,531	0,993 0,992–0,994	0,044 0,040–0,049
	uni	0,527 0,524–0,531	0,992 0,991–0,993	0,046 0,041–0,051
GBM	*bi_uni*	0,771 0,768–0,773	0,985 0,984–0,986	0,132 0,126–0,137
	uni	0,773 0,771–0,775	0,986 0,985–0,986	0,137 0,132–0,142
Glmnet	*bi_uni*	0,769 0,766–0,771	0,992 0,991–0,992	0,054 0,050–0,057
	uni	0,771 0,768–0,773	0,992 0,992–0,993	0,052 0,049–0,055
XGBoost	*bi_uni*	**0,780** 0,777–0,782	0,990 0,989–0,990	0,109 0,103–0,114
	uni	**0,780** 0,777–0,783	0,990 0,989–0,991	0,111 0,106–0,116
RF	*bi_uni*	0,738 0,735–0,741	0,999 0,999–0,999	0,003 0,002–0,004
	uni	0,740 0,738–0,743	0,999 0,999–0,999	0,004 0,003–0,005

Model *XGBoost* v obeh primerih napoveduje z enako napovedno uspešnostjo (AUC 0,780). Povprečja rezultatov treh napovednih modelov (*GBM*, *Glmnet*, *RF*) prikazujejo zanemarljivo prednost (ΔAUC = +0,002) pri napovedovanju umrljivosti v odsotnosti bigramov v primerjavi z rezultati, ko so bigrami vključeni v analizo (Tabela 9).

Prav tako smo preverili najpomembnejše spremenljivke glede na vključene oziroma izključene bigrame za model *GBM* (Tabela 10).

Tabela 10: Prikaz najpomembnejših spremenljivk glede na vključene in izključene bigrame (model GBM)

GBM						
bi_uni				***uni***		
Spremenljivke	***Vrednost***	***Vred_scal***		***Spremenljivke***	***Vrednost***	***Vred_scal***
u_stimuli	13,82	100,00		u_stimuli	14,07	100,00
AGE	11,19	80,98		AGE	11,41	81,11
u_famili	9,29	67,26		u_movement	9,41	66,92
u_movement	9,25	66,98		u_famili	9,27	65,88
u_unrespons	9,19	66,50		u_unrespons	9,23	65,64
PD_431	6,80	49,19		PD_431	7,12	50,59
u_ac	6,06	43,86		u_ac	6,15	43,74
PD_0389	5,88	42,59		u_levo	5,61	39,88
u_levo	5,62	40,68		PD_0389	5,51	39,18
u_made	4,81	34,82		u_made	4,83	34,36
u_liver	4,77	34,51		u_liver	4,80	34,14
u_levoph	4,37	31,59		u_levoph	4,16	29,60
u_pupil	3,60	26,07		u_lactat	3,82	27,18
u_dopamin	3,52	25,50		u_dopamin	3,64	25,90
u_lactat	3,48	25,19		u_multipl	3,54	25,19
u_multipl	3,40	24,64		u_poor	3,15	22,36
u_ffp	3,28	23,76		u_monitor	3,14	22,29
u_poor	3,28	23,74		u_pupil	3,11	22,08
u_monitor	3,09	22,36		u_ffp	2,98	21,15
PD_430	2,77	20,05		u_sp	2,72	19,30

Pri modelu *GBM* se v obeh primerih, z vključenimi bigrami in brez vključenih bigramov, med najbolj pomembne štiri spremenljivke uvrščajo *u_stimuli*, *AGE*, *u_famili* ter *u_movement* (Tabela 10).

Podobno raziskavo najpomembnejših spremenljivk smo izvedli tudi za model *Glmet* (Tabela 11).

Tabela 11: Prikaz najpomembnejših spremenljivk glede na vključene in izključene bigrame (model Glmnet)

GLMNET						
bi_uni				***Uni***		
Spremenljivke	***Vrednost***	***Vred_scal***		***Spremenljivke***	***Vrednost***	***Vred_scal***
PD_430	1,1162	100,00		PD_430	1,1083	100,00
PD_03849	0,9116	81,67		PD_03849	0,9101	82,12
PD_431	0,9107	81,59		PD_5770	0,9011	81,30
PD_5770	0,8814	78,96		PD_431	0,8973	80,96
PD_0389	0,7368	66,01		PD_0389	0,7349	66,31
PD_41401	0,6066	54,35		PD_41401	0,6175	55,72
PD_5849	0,2665	23,88		PD_5849	0,2648	23,89
PD_0380	0,2471	22,14		PD_0380	0,2553	23,04
PD_85220	0,2279	20,42		PD_85220	0,233	21,02
u_stimuli	0,2164	19,39		u_stimuli	0,2121	19,14
PD_41001	0,1217	10,90		PD_41001	0,1245	11,23
PD_5712	0,117	10,48		PD_5712	0,1205	10,87
b_bowel_sounds	0,113	10,12		PD_03811	0,0948	8,55
PD_03811	0,0877	7,86		u_heme	0,0855	7,71
u_heme	0,085	7,62		u_coars	0,0855	7,71
b_resp_care	0,0816	7,31		u_ac	0,0831	7,50
u_extub	0,0777	6,96		u_extub	0,077	6,95
u_coars	0,0758	6,79		u_movement	0,0767	6,92
u_movement	0,0755	6,76		PD_43491	0,0731	6,60
PD_43491	0,0747	6,69		u_unrespons	0,071	6,41

Pri modelu *Glmnet* se v obeh primerih (z vključenimi bigrami in brez vključenih bigramov) med prvih devet najpomembnejših spremenljivk uvrščajo le diagnoze oziroma spremenljivke strukturiranih podatkov, kot so intracerebralna krvavitev (*PD_431*), subarahnoidna krvavitev (*PD_430*), nespecifična septikemija (*PD_0389*) pri analizi, izvedeni na podatkovni zbirki z vključenimi bigrami, in prav tako subarahnoidna krvavitev (*PD_430*), septikemija zaradi ostalih gramnegativnih organizmov (razen *Hemophilus influenzae*, *Escherichia coli*, *Pseudomonas*, *Serratia*, Gramnegativni organizem (nespecifična Gramnegativna septikemija NOS)) (*PD_03849*) in akutni pankreatitis (*PD_5770*) pri analizi, izvedeni na podatkovni zbirki brez vključenih bigramov. Kot deseta najpomembnejša spremenljivka se v obeh primerih pojavi *u_stimuli*, vendar z le majhno pomembnostjo (~19) (Tabela 11).

Poleg omenjenih napovednih modelov smo opravili raziskavo najpomembnejših spremenljivk glede na vključene oziroma izključene bigrame tudi za model *XGBoost* (Tabela 12).

Tabela 12: Prikaz najpomembnejših spremenljivk glede na vključene in izključene bigrame (model XGBoost)

XGBoost						
bi_uni				***uni***		
Spremenljivke	***Vrednost***	***Vred_scal***		***Spremenljivke***	***Vrednost***	***Vred_scal***
u_stimuli	0,0483	100,00		u_stimuli	0,0452	100,00
u_movement	0,0323	66,87		u_movement	0,0302	66,81
u_famili	0,0271	56,11		u_famili	0,0277	61,28
u_ac	0,0266	55,07		AGE	0,0265	58,63
AGE	0,0264	54,66		u_ac	0,0252	55,75
u_unrespons	0,021	43,48		u_unrespons	0,0209	46,24
u_extub	0,0165	34,16		u_extub	0,0163	36,06
PD_41401	0,0164	33,95		PD_41401	0,0155	34,29
u_stabl	0,0137	28,36		u_stabl	0,014	30,97
PD_431	0,0134	27,74		u_liver	0,0135	29,87
u_diet	0,0134	27,74		PD_431	0,0133	29,42
u_liver	0,0131	27,12		u_diet	0,0126	27,88
u_levoph	0,0117	24,22		u_made	0,0126	27,88
u_made	0,0116	24,02		u_levo	0,0114	25,22
u_hypotens	0,0115	23,81		u_hypotens	0,0112	24,78
u_levo	0,0107	22,15		PD_0389	0,0109	24,12
PD_0389	0,0104	21,53		u_levoph	0,0102	22,57
u_poor	0,01	20,70		u_lactat	0,01	22,12
u_lactat	0,0094	19,46		u_poor	0,01	22,12
u_co	0,0092	19,05		u_co	0,0096	21,24

Pri napovednem modelu *XGBoost*, ki je dosegel najvišjo povprečno napovedno uspešnost, se med tri najbolj pomembne spremenljivke v obeh primerih uvrščajo spremenljivke *u_stimuli*, *u_movement* in *u_famili*. Sledita jim še spremenljivki *u_ac*, *AGE* v primeru vključenih bigramov oziroma v obratnem vrstnem redu v primeru brez vključenih bigramov (Tabela 12).

5.2.4 H2: Uporaba bigramov skupaj z unigrami izboljša napovedno uspešnost modelov v primerjavi z uporabo izključno unigramov.

Z uporabo statističnega testa T-test smo preverili razlike v uspešnosti napovednih modelov, zgrajenih na podatkovnih zbirkah z vključenimi in izključenimi bigrami.

Tabela 13: Rezultati statistično nesignifikantne razlike modelov glede na vključene oziroma izključene bigrame

Modeli	bi_uni	uni	p-vrednost	t	df
C5.0Ruleset	0,527	0,527	0,890	-0,139	197,05
GBM	0,771	0,773	0,211	-1,254	194,95
Glmnet	0,769	0,771	0,347	-0,943	197,98
XGBoost	0,778	0,780	0,862	-0,175	196,71
RF	0,738	0,740	0,248	-1,158	197,91

Uspešnosti vseh napovednih modelov se na ravni rezultatov posameznega modela statistično ne razlikujejo glede na spol ($p > 0,05$) (Tabela 13).

5.3 Primerjava napovednih modelov glede na uporabo kriterija SAPS

Pri primerjavi napovednih modelov glede na uporabo kriterija SAPS smo pri gradnji modelov uporabili le osnovne spremenljivke starost, SAPS ter nestrukturirani podatki (bigrami in unigrami).

5.3.1 Rezultati – vključen kriterij SAPS

Zanimalo nas je, s katerim napovednim modelom dosežemo najboljše rezultate na podatkih ženskega vzorca z vključenim kriterijem SAPS (Tabela 14).

Tabela 14: Napovedne uspešnosti modelov, zgrajenih na podatkih z vključenim kriterijem SAPS (Ženski vzorec)

Ženski vzorec			
Model	**AUC**	**PPV**	**NPV**
C5.0Ruleset	0,611 0,598–0,624	0,898 0,897–0,899	0,304 0,285–0,323
GBM	0,743 0,739–0,747	0,897 0,896–0,898	0,423 0,400–0,445
Glmnet	0,765 0,762–0,769	0,893 0,893–0,894	0,567 0,531–0,603
XGBoost	0,740 0,735–0,745	0,897 0,896–0,898	0,435 0,405–0,465
RF	0,714 0,709–0,720	0,888 0,888–0,888	0,374 0,319–0,429

Pri modelih, zgrajenih na podatkih ženskega vzorca, pri katerih je vključen kriterij SAPS, je najboljšo napovedno uspešnost dosegel napovedni model *Glmnet* (AUC = 0,765). Takoj za njim sta se po kriteriju AUC uvrstila *GBM* (AUC = 0,743) ter *XGBoost* (AUC = 0,740), medtem ko se je kot najslabši ponovno izkazal napovedni model *C5.0Ruleset* (AUC = 0,611) (Tabela 14).

Podobno smo preverili, s katerim napovednim modelom dosežemo najboljše rezultate na podatkih moškega vzorca z vključenim kriterijem SAPS (Tabela 15).

Tabela 15: Napovedne uspešnosti modelov, zgrajenih na podatkih z vključenim kriterijem SAPS (Moški vzorec)

Moški vzorec			
Model	**AUC**	**PPV**	**NPV**
C5.0Ruleset	0,647 0,633–0,661	0,900 0,899–0,902	0,411 0,394–0,427
GBM	0,764 0,760–0,768	0,901 0,900–0,902	0,583 0,563–0,603
Glmnet	0,777 0,773–0,781	0,886 0,886–0,887	0,615 0,579–0,650
XGBoost	0,773 0,769–0,777	0,898 0,897–0,899	0,604 0,582–0,626
RF	0,751 0,746–0,756	0,883 0,883–0,883	0,819 0,770–0,865

Najboljšo napovedno uspešnost smo dosegli z napovednim modelom *Glmnet* (AUC = 0,777). Takoj za njim sta se po kriteriju AUC uvrstila *XGBoost* (AUC = 0,773) ter *GBM* (AUC = 0,764), najslabšo napovedno uspešnost pa smo ponovno izmerili pri *C5.0Ruleset* (AUC = 0,647) (Tabela 15).

5.3.2 Rezultati – izključen kriterij SAPS

Zanimalo nas je, s katerim napovednim modelom dosežemo najboljše rezultate na podatkih ženskega vzorca z izključenim kriterijem SAPS (Tabela 16).

Tabela 16: Napovedne uspešnosti modelov, zgrajenih na podatkih brez vključenega kriterija SAPS (ženski vzorec)

Ženski vzorec			
Model	**AUC**	**PPV**	**NPV**
C5.0Ruleset	0,547 0,540–0,554	0,891 0,891–0,892	0,259 0,224–0,294
GBM	0,666 0,661–0,671	0,892 0,891–0,892	0,256 0,228–0,285
Glmnet	0,674 0,669–0,679	0,891 0,891–0,892	0,405 0,366–0,443
XGBoost	0,670 0,664–0,676	0,892 0,891–0,892	0,289 0,251–0,326
RF	0,638 0,633–0,644	0,888 0,888–0,890	0,028 0,018–0,038

Z izključitvijo kriterija SAPS se je med napovednimi modeli, zgrajenimi na ženskem vzorcu podatkov, najbolje obnesel model *Glmnet* (AUC = 0,674), takoj za njim pa modela *XGBoost* (AUC = 0,670) ter *GBM* (AUC = 0,666) (Tabela 16).

Podobno nas je zanimalo, s katerim napovednim modelom dosežemo najboljše rezultate na podatkih moškega vzorca (Tabela 17).

Tabela 17: Napovedne uspešnosti modelov, zgrajenih na podatkih brez vključenega kriterija SAPS (moški vzorec)

Moški vzorec			
Model	**AUC**	**PPV**	**NPV**
C5.0Ruleset	0,558 0,549–0,568	0,886 0,885–0,887	0,246 0,215–0,276
GBM	0,684 0,680–0,689	0,887 0,887–0,888	0,360 0,330–0,391
Glmnet	0,689 0,684–0,694	0,884 0,883–0,884	0,277 0,236–0,318
XGBoost	0,693 0,688–0,698	0,887 0,887–0,888	0,361 0,333–0,389
RF	0,678 0,672–0,683	0,882 0,882–0,880	0,375 0,343–0,407

Za razliko od rezultatov uspešnosti napovednih modelov na podatkih ženskega vzorca ima med modeli, zgrajenimi na podatkih moškega vzorca, najboljšo napovedno uspešnost model *XGBoost* (AUC = 0,693), sledita mu *Glmnet* (AUC = 0,689) in *GBM* (AUC = 0,684) (Tabela 17).

5.3.3 Primerjava med zbirkama z vključenim in izključenim kriterijem SAPS za posamezni spol.

Zanimale so nas razlike v napovedni uspešnosti napovednega modela, zgrajenega na podatkih ženskega vzorca, glede na vključen oziroma izključen kriterij SAPS (Tabela 18).

Tabela 18: Primerjava napovednih uspešnosti glede na vključen/izključen kriterij SAPS (ženski vzorec)

Ženski vzorec				
Model	**Podatkovna zbirka**	**AUC**	**SENS**	**SPEC**
C5.0Ruleset	*SAPS*	0,611 0,598–0,624	0,958 0,954–0,962	0,137 0,125–0,149
	brez SAPS	0,547 0,540–0,554	0,975 0,970–0,979	0,057 0,049–0,065
GBM	*SAPS*	0,743 0,739–0,747	0,981 0,979–0,983	0,105 0,095–0,116
	brez SAPS	0,666 0,661–0,671	0,981 0,979–0,983	0,054 0,047–0,060
Glmnet	*SAPS*	0,765 0,762–0,769	0,994 0,993–0,995	0,060 0,054–0,066
	brez SAPS	0,674 0,669–0,679	0,991 0,989–0,993	0,041 0,036–0,046
XGBoost	*SAPS*	0,740 0,735–0,745	0,980 0,977–0,982	0,107 0,099–0,116
	brez SAPS	0,670 0,664–0,676	0,981 0,978–0,984	0,054 0,046–0,062
RF	*SAPS*	0,714 0,709–0,720	0,999 0,999–1	0,003 0,002–0,004
	brez SAPS	0,638 0,633–0,644	1 0,999–1	,001 -0,001–0,001

V obeh primerih (s in brez kriterija SAPS) smo z napovednim modelom *Glmnet* dosegli najvišjo napovedno uspešnost (AUC_{SAPS} = 0,765; $AUC_{Brez\ SAPS}$ = 0,674). Sledita mu *GBM* (AUC_{SAPS} = 0,743; $AUC_{Brez\ SAPS}$ = 0,666) in *XGBoost* (AUC_{SAPS} = 0,740; $AUC_{Brez\ SAPS}$ = 0,670) (Tabela 18).

Podobno smo preverili tudi pri napovednih modelih, zgrajenih na podatkih moškega vzorca (Tabela 19).

Tabela 19: Primerjava napovednih uspešnosti glede na vključen/izključen kriterij SAPS (moški vzorec)

Moški vzorec				
Model	**Podatkovna zbirka**	**AUC**	**SENS**	**SPEC**
C5.0Ruleset	*SAPS*	0,647 0,633–0,661	0,958 0,955–0,962	0,207 0,195–0,220
	brez SAPS	0,558 0,549–0,568	0,968 0,964–0,971	0,072 0,063–0,081
GBM	*SAPS*	0,764 0,760–0,768	0,980 0,978–0,981	0,197 0,187–0,208
	brez SAPS	0,684 0,680–0,689	0,983 0,981–0,984	0,066 0,060–0,072
Glmnet	*SAPS*	0,777 0,773–0,781	0,996 0,995, 0,997	0,044 0,039, 0,048
	brez SAPS	0,689 0,684–0,694	0,990 0,986–0,993	0,026 0,019–0,034
XGBoost	*SAPS*	0,773 0,769–0,777	0,984 0,983–0,986	0,163 0,153–0,172
	brez SAPS	0,693 0,688–0,698	0,983 0,981–0,985	0,069 0,060–0,077
RF	*SAPS*	0,751 0,746–0,756	1 1–1	0,011 0,009–0,013
	brez SAPS	0,678 0,672–0,683	1 1–1	0,001 0,001–0,001

Pri moškem vzorcu pa so doseženi rezultati najvišje napovedne uspešnosti z vključenim kriterijem SAPS zaznane pri modelu *Glmnet* (AUC $_{SAPS}$ = 0,777). Sledita mu *XGBoost* (AUC $_{SAPS}$ = 0,773) ter *GBM* (AUC $_{SAPS}$ = 0,764). Pri napovednih modelih, zgrajenih na podatkih brez vključenega kriterija SAPS, ima najvišjo napovedno uspešnost *XGBoost* (AUC $_{Brez\ SAPS}$ = 0,693), sledita mu *Glmnet* (AUC $_{Brez\ SAPS}$ = 0,689) in *GBM* (AUC $_{Brez\ SAPS}$ = 0,684) (Tabela 19).

5.3.4 RV1: Ali je možno zgraditi napovedni model, ki bo na podlagi nestrukturiranih podatkov, zbranih v prvih šestih urah hospitalizacije na enoti intenzivnega oddelka, bolj uspešen od pogosto uporabljenega kriterija SAPS, ki je izračunan na podlagi točno določenih podatkov, zbranih v prvih 24 urah hospitalizacije?

Z uporabo statističnega testa T-test smo preverili razlike v uspešnosti napovednih modelov, zgrajenih na podatkih ženskega vzorca z vključenim in izključenim kriterijem SAPS (Tabela 20).

Tabela 20: Rezultati statistično signifikantne razlike modelov glede na vključen/izključen kriterij SAPS (ženski vzorec)

Ženski vzorec					
Modeli	**Brez SAPS**	**SAPS**	**P-vrednost**	**t**	**df**
C50Ruleset	0,547	0,611	< 0,001	-8,51	153,09
GBM	0,666	0,743	< 0,001	-22,52	193,89
Glmnet	0,674	0,765	< 0,001	-2833	186,53
XGBoost	0,670	0,740	< 0,001	-18,35	193,33
RF	0,638	0,714	< 0,001	-19,10	197,96

S pomočjo rezultatov T-testa je razvidno, da imajo vsi modeli, zgrajeni na podatkih ženskega vzorca ter uporabljenim kriterijem SAPS, signifikantno višjo napovedno uspešnost od modelov brez uporabljenega kriterija SAPS ($p < 0,001$) (Tabela 20).

Podobno smo preverili tudi za napovedne modele, zgrajene na podatkih moškega vzorca (Tabela 21).

Tabela 21: Rezultati statistično signifikantne razlike modelov glede na vključen/izključen kriterij SAPS (moški vzorec)

Moški vzorec					
Modeli	**Brez SAPS**	**SAPS**	**P-vrednost**	**t**	**df**
C5.0Ruleset	0,558	0,647	< 0,001	-10,67	175,24
GBM	0,684	0,764	< 0,001	-25,68	195,13
Glmnet	0,689	0,777	< 0,001	-27,31	197,10
XGBoost	0,693	0,773	< 0,001	-25,41	193,28
RF	0,678	0,751	< 0,001	-20,77	194,84

Z rezultatov T-testa je razvidno, da smo z vsemi modeli, zgrajenimi na podatkih moškega vzorca ter uporabljenim kriterijem SAPS, dosegli signifikantno višjo napovedno uspešnost od modelov brez uporabljenega kriterija SAPS ($p < 0,001$) (Tabela 21).

6 Interpretacija in razprava

SB2 je v porastu. Za SB2 zboleva vedno več ljudi zaradi neprilagojenega življenjskega stila, bodisi zaradi pomankanja časa za udejstvovanje pri fizičnih dejavnostih bodisi zaradi nepravilnega prehranjevanja (Bang, 2009; Paulweber, 2010). Čeprav večina ljudi SB2 vidi kot samoumevno bolezen, ki se lahko pojavi v poznih letih, se mnogi ne zavedajo njene resnosti. Ta lahko povzroči velike težave: od izgube vida (Jeon & Murray, 2008), do amputacij okončin (Johannesson, et al., 2009) oziroma v skrajnem primeru tudi do smrti (Brown, et al., 2015). S starostjo se tveganje za SB2 povečuje, vendar pa lahko v veliki meri na povečanje tveganja vplivamo predvsem sami (Bang, 2009).

Smrtnemu izidu so najbolj podvrženi tisti bolniki SB2, ki so bili hospitalizirani na enoti intenzivnega oddelka. Prav zaradi tega nas je v nalogi zanimalo, ali je mogoče uspešno napovedati tveganje za umrljivost le na podlagi zapisov medicinskih sester, kar bi lahko služilo zdravnikom kot pomoč pri sprejemanju nadaljnjih odločitev. Predvsem nas je zanimalo, kateri koreni besed najbolj vplivajo na umrljivost posameznika. Poleg tega smo v nalogi preverili, kateri izmed klasifikatorjev lahko s pomočjo strukturiranih in nestrukturiranih podatkov najbolje napove tveganje, da bo pacient umrl. Med drugim smo tudi preverili, ali prisotnost bigramov vpliva na končni izid ter ali je smiselno uporabiti različna modela za posamezni spol.

V preteklosti so v podobni raziskavi (Marafino, et al., 2015) razvili klasifikatorje, ki temeljijo samo na zapisih medicinskih sester v prvih 24. urah po sprejetju bolnika na intenzivni oddelek, in hkrati preverili, katere spremenljivke najbolj vplivajo na umrljivost pacientov. Upoštevati je treba, da zapisi niso bili osredotočeni le na določene paciente, ampak so bili v vzorec zavzeti vsi pacienti, ki so bili sprejeti na oddelek intenzivne nege. Za spremenljivke so v analizi uporabili le tekstovne oziroma nestrukturirane podatke, razbrane iz zapisov medicinskih sester. V primerjavi z našo analizo so posledično zaradi obsega različnih pacientov imeli na voljo bistveno več spremenljivk (91.317 unigramov, 1.751.205 bigramov, skupaj 1.842.522), ki so bile določene iz takrat aktualne podatkovne zbirke MIMIC-II. Uporabljena sta bila dva različna klasifikatorja, logistična regresija in linearna

metoda podpornih vektorjev, optimizirana s stohastičnim gradientnim spustom (stochastic gradient descent (SGD)) in z vključeno Elastic net regularizacijo, in sicer regularizacijo L1 (lasso regresija) in regularizacijo L2 (ridge regresija). Najboljši rezultat (AUC = 0,897) so dosegli z uporabo kombinacije logistične regresije in L2-regularizacije (Marafino, et al., 2015).

V naši analizi smo za nadaljnjo primerjavo zavzeli le rezultate dveh najboljših napovednih modelov (*GBM* in *XGBoost*), ki v primerjavi z napovednim modelom *Glmnet* vključujeta v napoved predvsem nestrukturirane, tekstovne podatke. Prav tako so v razpravi upoštevani le rezultati najpomembnejših spremenljivk napovednih modelov, zgrajenih nad celotnim vzorcem, ne glede na spol.

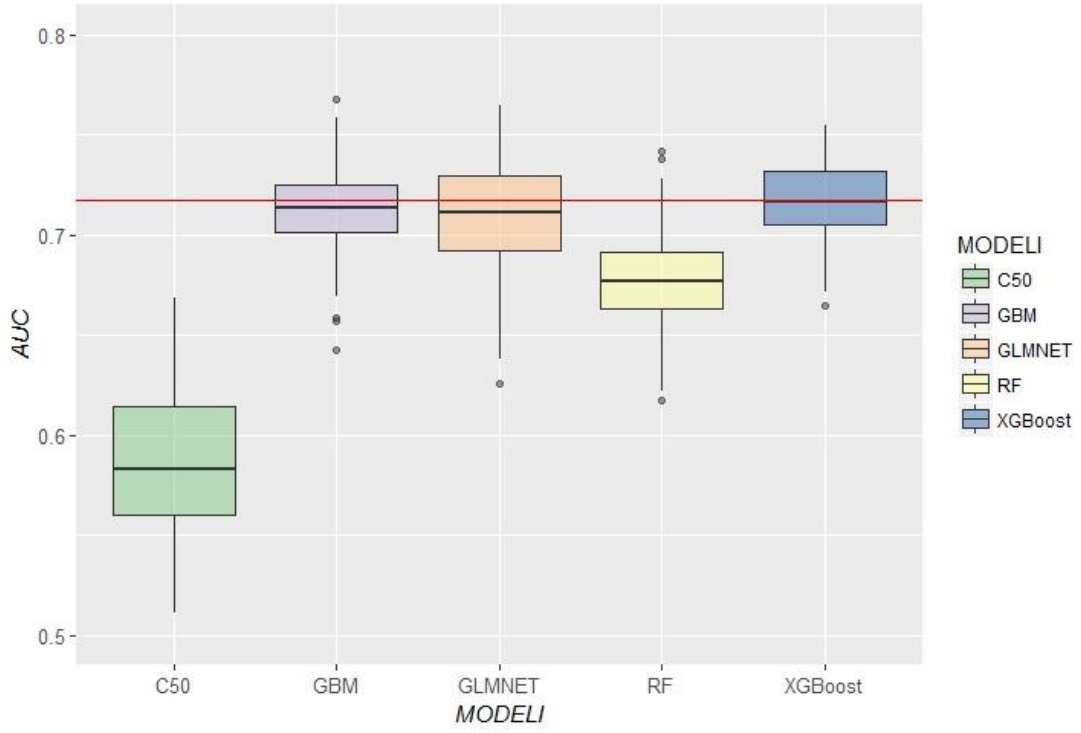
Spremenljivka *u_stimuli*, ki predstavlja stimulacijo oziroma spodbujanje, je v naši raziskavi prisotna pri samem vrhu najpomembnejših spremenljivk za napovedovanje umrljivosti. Besedo »stimulacija« bi lahko povezali s spremenljivkama *u_levo* in *u_levoph*, ki predstavljata levophed z aktivno učinkovino neropinefrinom oziroma noradrenalinom. Ena izmed lastnosti hormona noradrenalina je povišanje krvnega tlaka (Chistriakov, et al., 2015). Znano pa je, da je prisotnost SB2 povezana s hipertenzijo oziroma povišanjem krvnega tlaka (ADA, 2014). Obe spremenljivki se sicer v primerjavi s študijo (Marafino, et al., 2015) ne nahajata med najvplivnejšimi spremenljivkami za napovedovanje umrljivosti (*u_levo* (31,98 % ±9.67), *u_levoph* (27,00 % ±4,29)). V naši študiji je imela spremenljivka *AGE* (»starost«) prav tako veliko pomembnost pri napovedovanju umrljivosti, kar so tudi ugotovili v študiji, kjer so staranje povezali z višjim tveganjem, da bo pacient umrl (ADA, 2014; Marafino, et al., 2015). Poleg omenjenih se je na vrhu pojavila tudi spremenljivka *u_movement*, ki pooseblja gibanje, kjer je bila besedna zveza v zapisih najverjetneje omenjena v povezavi s pomanjkanjem vsakodnevnega gibanja, kar le poveča tveganje za pojav SB2 (ADA, 2014). Sledi spremenljivka *u_unrespons* (*unresponsive*), ki se navezuje na neodzivnost. Prav slednja je bila v študiji (Marafino, et al., 2015) navedena kot tretja najpomembnejša spremenljivka, ki ima pozitiven vpliv na umrljivost. Spremenljivka *PD_431* je diagnoza, ki nakazuje, da je prišlo pri bolniku do intracerebralne krvavitve. Intracerebralna krvavitev lahko povzroči hemiparezo oziroma omrtvelost določenega dela telesa (Jorgensen, et al.,

1995). Prav zaradi tega lahko prisotnost diagnoze (intracerebralna krvavitev) povežemo s pojavnostjo spremenljivke *u_unrespons* v zapisih medicinskih sester. Sledita spremenljivka *u_famili* (*familiar*), ki pomeni »znan«, ter spremenljivka *u_ac*, ki se lahko navezuje na vrsto različnih besed, kot so na primer »izpolniti« (*accomplish*), »sprejeti« (*accept*), »dejanje«, »aktivno« (*act-ion*/*-ive*) in podobno.

Napovedni modeli glede na spol

Spodnji grafični prikaz prikazuje primerjavo napovedne uspešnosti modelov, zgrajenih na podatkih ženskega vzorca (Slika 1).

Slika 1: Prikaz napovednih uspešnosti modelov (AUC), zgrajenih na podatkih ženskega vzorca

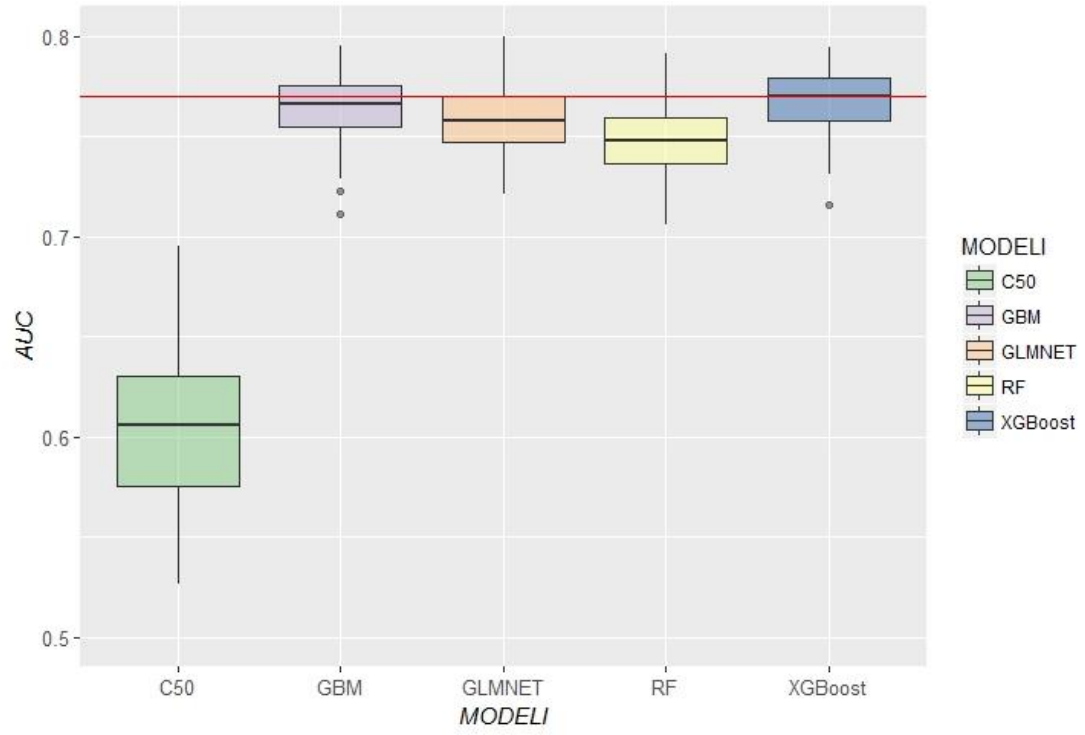


Pri analizi na podatkih, kjer je vključen le ženski vzorec, ima najvišje povprečje napovedne uspešnosti model *XGBoost* (AUC = 0,718 (95 % IZ: 0,714–0,721)), kar je ponazorjeno tudi s horizontalno rdečo premico (Slika 1). Prvi kvartili (Q1) napovednih modelov, zgrajenih s klasifikatorji *GBM* (AUC_{Q1} = 0,701), *Glmnet* (AUC_{Q1} = 0,692) in *XGBoost* (AUC_{Q1} = 0,706), imajo višjo vrednost kot tretji kvartil napovednega modela RF (AUC_{Q3} = 0,691), kar pomeni, da večinski del rezultatov napovedne uspešnosti prvih treh modelov vsebuje višjo vrednost AUC kot v primeru

rezultatov modela *RF*. Absolutna razlika v vrednosti prvega kvartila (Q1) in tretjega kvartila (Q3) je pri modelu *XGBoost* ($AUC_{|Q3 - Q1|}$ = 0,026) manjša kot pri modelu *Glmnet* ($AUC_{|Q3 - Q1|}$ = 0,038), kar je moč opaziti tudi pri vrednostih intervala zaupanja (Tabela 1) in na prikazu primerjave uspešnosti med napovednimi modeli (Slika 1).

Prav tako smo želeli določiti najuspešnejši napovedni model, zgrajen na podatkih moškega vzorca. Primerjava napovednih modelov je razvidna iz naslednjega grafičnega prikaza (Slika 2).

Slika 2: Prikaz napovednih uspešnosti modelov (AUC), zgrajenih na podatkih moškega vzorca



Rezultati analize, opravljene na podatkih moškega vzorca, prikazujejo podobne rezultate. Najvišja povprečna vrednost AUC je prisotna pri modelu *XGBoost* (AUC = 0,770 (95 % IZ: 0,767–0,774)) (Tabela 2), kar je ponazorjeno tudi s horizontalno rdečo premico (Slika 1). Rezultati povprečja kažejo na to, da je vrednost Q1 pri *XGBoost* (AUC_{Q1} = 0,759) višja kot povprečna vrednost pri modelu *RF* (AUC = 0,748). Spodnja meja (Q1) večinskega dela rezultatov napovedne uspešnosti modela

XGBoost je približno enaka ($|AUC_{XGBoostQ1} - AUC_{Glmnet}| = 0{,}001$) povprečni vrednosti rezultatov modela *Glmnet* (AUC = 0,760) (Slika 2).

Vsi modeli, zgrajeni na podatkih moškega vzorca, so v povprečju prikazali boljše rezultate napovedi umrljivosti bolnikov v primerjavi z modeli, zgrajenimi na podatkih ženskega vzorca (ΔAUC = +0,049) (Tabela 6).

H1: Uspešnost napovednih modelov se bistveno razlikuje glede na spol.

Razlike med rezultati napovedne uspešnosti modelov na podatkih ženskega in moškega vzorca se signifikantno statistično razlikujejo ($p < 0{,}001$) na ravni posameznega modela (Tabela 6). Na podlagi omenjenih rezultatov lahko hipotezo (H1) dokončno potrdimo. Povprečna uspešnost napovednih modelov, zgrajenih in testiranih na podatkih moškega vzorca, je bila v vseh primerih višja od uspešnosti napovednih modelov, namenjenih za žensko populacijo.

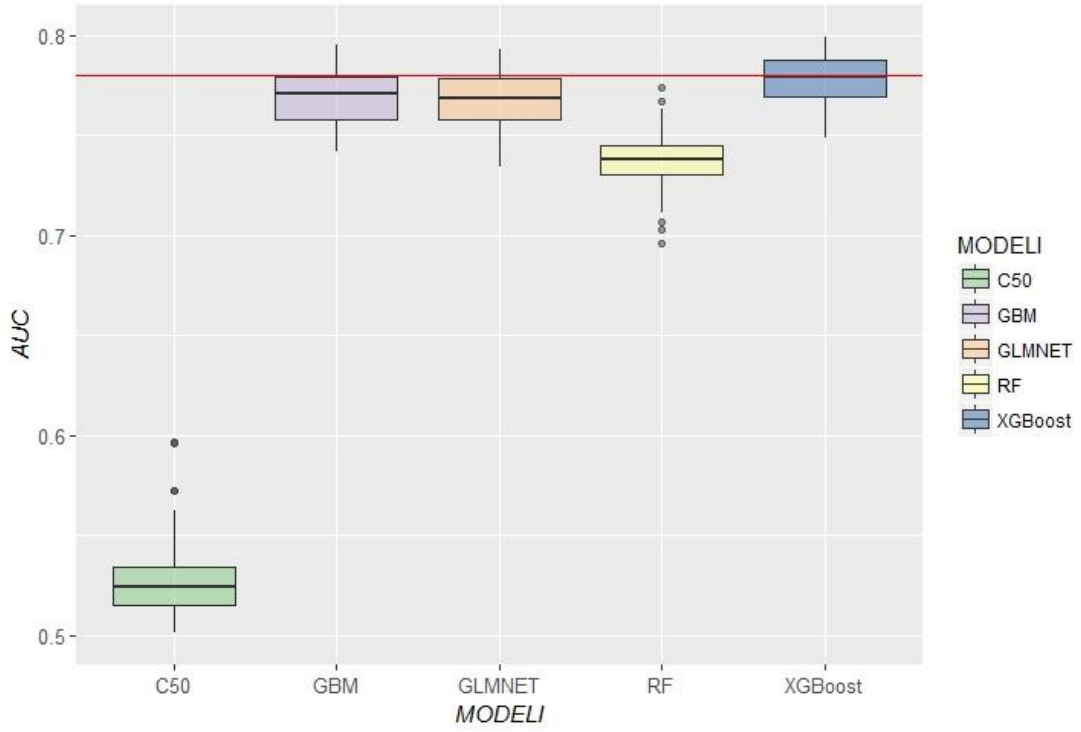
Gradnja skupnega napovednega modela je primernejša izbira

Zanimivo je bilo tudi raziskati, ali je v tem primeru smiselno graditi ločene modele glede na spol. Vse primerjave se znotraj modela statistično razlikujejo ($p < 0{,}05$). Napovedna modela *C5.0Ruleset* glede na spol v povprečju napovedujeta bolje (ΔAUC = +6 %) od združenega modela, vendar še zmeraj premalo, da bi bila v praksi uporabna (< 61 % napovedna uspešnost). Vse modele, *XGBoost* (ΔAUC = +3,6 %), *Glmnet* (ΔAUC = +3,4 %) in *GBM* (ΔAUC = +3,2 %), je smiselno graditi za skupno populacijo. Model *RF*, zgrajen na podatkih moškega vzorca, zanemarljivo izboljša napovedno uspešnost glede na združeni model (ΔAUC = +1 %), medtem ko napovedni model za žensko populacijo poslabša napovedno uspešnost (ΔAUC = -6 %) v primerjavi z združenim modelom.

Napovedni modeli glede na uporabo/neuporabo bigramov

Primerjave napovednih modelov glede na uporabo oziroma neuporabo bigramov so grafično ponazorjene na naslednjem grafičnem prikazu (Slika 3).

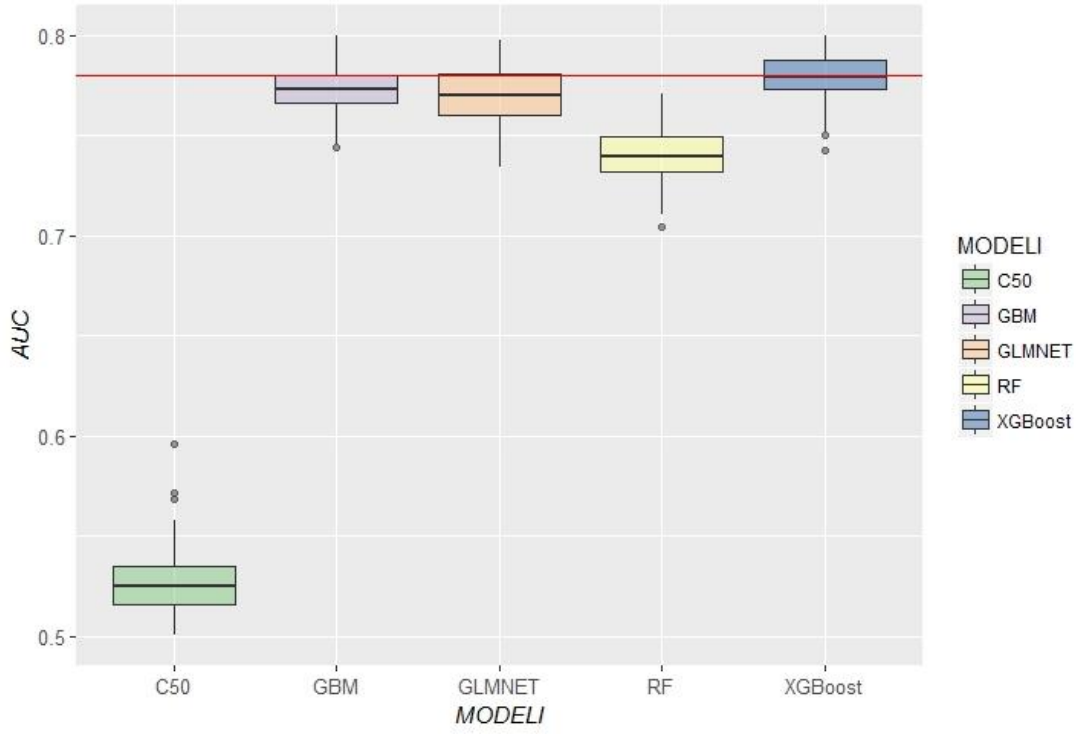
Slika 3: Prikaz napovednih uspešnosti modelov (AUC), zgrajenih na podatkih z vključenimi bigrami



Pri analizi na podatkih, kjer so bili uporabljeni vsi podatki, je najvišjo povprečno vrednost AUC dosegel model *XGBoost* (AUC = 0,7797) (Tabela 7), na sliki ponazorjen s horizontalno rdečo premico (Slika 1). Povprečna vrednost *XGBoost* presega vrednosti tretjega kvartila (Q3) skoraj vseh napovednih modelov (*Glmnet* AUC_{Q3} = 0,779; RF AUC_{Q3} = 0,745; *C5.0Ruleset* AUC_{Q3} = 0,534), izjema je *GBM* (AUC_{Q3} = 0,7798), čigar Q3 je za malenkost višji od povprečne vrednosti AUC modela *XGBoost* (ΔAUC = +0,0001). Poleg tega je prvi kvartil *XGBoost* (AUC_{Q1} = 0,769) na približno enaki vrednosti kot povprečni vrednosti AUC najbližjih modelov *GBM* (AUC = 0,771) in *Glmnet* (AUC = 0,769) (Slika 3).

Prav tako smo želeli določiti najuspešnejši napovedni model, zgrajen na podatkih brez vključenih bigramov. Primerjava napovednih modelov je razvidna iz naslednjega grafičnega prikaza (Slika 4).

Slika 4: Prikaz napovednih uspešnosti modelov (AUC), zgrajenih na podatkih brez vključenih bigramov



Rezultati analize, opravljene na podatkih brez vključenih bigramov, prikazujejo, da povprečna vrednost AUC modela *XGBoost* (AUC = 0,780) (Tabela 8) presega vrednosti Q3 dveh napovednih modelov (*RF* AUC_{Q3} = 0,750; *C5.0Ruleset* AUC_{Q3} = 0,535), izjemi sta *Glmnet* (AUC_{Q3} = 0,781) in *GBM* (AUC_{Q3} = 0,780) (Slika 4).

Na podlagi rezultatov analiz z oziroma brez vključenih bigramov pridemo do ugotovitve, da so napovedni modeli *GBM*, *Glmnet* in predvsem *XGBoost* najbolj učinkoviti glede napovedne uspešnosti (Tabela 9).

H2: Uporaba bigramov skupaj z unigrami izboljša napovedno uspešnost modelov v primerjavi z uporabo izključno unigramov.

Razlike med rezultati napovedne uspešnosti modelov na podatkih z in brez bigramov se v splošnem signifikantno statistično ne razlikujejo ($p > 0{,}001$) na ravni posameznega modela (Tabela 13). Na podlagi omenjenih rezultatov lahko hipotezo (H2) dokončno zavržemo. Uspešnost napovednih modelov, zgrajenih in testiranih na podatkovnih zbirkah z oziroma brez bigramov, se v splošnem ne razlikuje. Podobno so ugotovili tudi v študiji, ki sicer ni obravnavala le bolnikov SB2, kjer prisotnost

raznolikih bigramov ni povečala uspešnosti napovednega modela (Marafino, et al., 2015).

RV1: Ali je možno zgraditi napovedni model, ki bo na podlagi nestrukturiranih podatkov, zbranih v prvih šestih urah hospitalizacije na enoti intenzivnega oddelka, bolj uspešen od pogosto uporabljenega kriterija SAPS, ki je izračunan na podlagi točno določenih podatkov, zbranih v prvih 24 urah hospitalizacije?

Vsi modeli, zgrajeni na podatkih ženskega vzorca z vključenim kriterijem SAPS, so v povprečju prikazali boljše rezultate napovedi umrljivosti bolnikov v primerjavi z modeli, zgrajenimi na podatkih ženskega vzorca brez vključenega kriterija SAPS (ΔAUC = +0,0756) (Tabela 18). Podobno velja za modele, zgrajene na podatkih moškega vzorca. Modeli, pri katerih je bil vključen kriterij SAPS, so v povprečju prikazali boljše rezultate napovedi umrljivosti bolnikov v primerjavi z modeli brez vključenega kriterija SAPS (ΔAUC = +0,082) (Tabela 19). Razlike med rezultati napovedne uspešnosti modelov na podatkih brez in z uporabljenim kriterijem SAPS so statistično signifikantne na nivoju posameznega modela ($p < 0,001$) glede na ženski (Tabela 20) kakor tudi moški vzorec (Tabela 21). Na podlagi rezultatov lahko zatrdimo, da na podatkih (ne glede na spol) ni možno zgraditi napovednega modela, ki bi bil na podlagi nestrukturiranih podatkov, zbranih v prvih šestih urah hospitalizacije na enoti intenzivnega oddelka, bolj uspešen od pogosto uporabljenega kriterija SAPS.

Do podobne ugotovitve so prišli tudi v raziskavi, v kateri sicer niso imeli tarčnih pacientov, vendar so prav tako napovedovali umrljivost z uporabo zapisov medicinskih sester nekaj manj kot 15.000 odraslih pacientov intenzivne nege. Z uporabo tematskih modelov (Hierarchical Dirichlet Processes (HDP)) so prikazali, da je z dodatkom zapisov medicinskih sester možno signifikantno izboljšati uspešnost algoritma, ki sprva napoveduje samo na podlagi fizioloških podatkov (SAPS-I algoritem) (Lehman, et al., 2012).

7 Sklep

V raziskavi smo ugotovili, da je za napovedovanje umrljivosti bolnikov s SB2 najprimernejša izbira napovednega modela, ki temelji na algoritmu *XGBoost,* kot alternativni izbiri pa lahko uporabimo tudi modela *GBM* in *Glmnet*. Prisotnost besed, ki se navezujejo na stimulacijo oziroma spodbujanje, starost, gibanje, neodzivnost in diagnozo intracerebralne krvavitve, ima največjo težo pri uspešnem napovedovanju umrljivosti bolnikov s SB2.

Uporaba dodatnih spremenljivk v obliki bigramov ni potrebna, saj se z vključitvijo uspešnost napovedovanja napovednih modelov ne izboljša signifikantno. Uporaba pogosto uporabljenega kriterija SAPS, ki temelji na fizioloških podatkih, ostaja primarno vodilo pri napovedovanju umrljivosti bolnikov s SB2.

Uporaba enotnega napovednega modela za oba spola je najprimernejša izbira, s tem pa posledično tudi zmanjšamo zapletenost interpretacije pri napovedih.

V nadaljevanju bi bilo analizo smiselno opraviti še z ostalimi algoritmi in naprednimi tehnikami podatkovnega rudarjenja, med katere spadajo ansambli, globoko učenje z uporabo večslojnih nevronskih mrež in ostale tehnike.

Literatura

American Diabetes Association, 2014. Diagnosis and classification of diabetes mellitus. *Diabetes Care,* 37(Suppl. 1), pp. S81–S90.

Bang, H. et al., 2009. A patient self-assessment diabetes screening score. *Annals of Internal Medicine,* 151(11), pp. 775–783.

Brown, E., Natoli, N., McLaughlin, R. & Mehta, K., 2015. Pathways and barriers to diabetes pathways and barriers to diabetes. *Procedia Engineering,* 107, pp. 387–394.

Casanova, R. et al., 2014. Application of random forests methods to diabetic retinopathy classification analyses. *PLoS ONE,* 9(6), p. e98587.

Chen, T., He, T. & Benesty, M., 2016. *Extreme gradient boosting.* [pdf] Available at: https://cran.r-project.org/web/packages/xgboost/xgboost.pdf [Accessed 14. 07. 2016].

Chistriakov, D. A., Ashwell, K. W., Orekhov, A. N. & Bobryshev, Y. V., 2015. Innervation of the arterial wall and its modification in atherosclerosis. *Autonomic Neuroscience,* 193, pp. 7–11.

Efron, B. & Tibshirani, R., 1994. *An introduction to the bootstrap.* London: Chappman Hall/CRC.

Fawcett, T., 2006. An introduction to ROC analysis. *Pattern Recognition Letters,* 27(8), pp. 861–874.

Friedman, J., Hastie, T. & Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software,* 33(1), pp. 1–22.

Fu, H. et al., 2015. Early glycemic response predicts achievement of subsequent treatment targets in the treatment of type 2 diabetes: a post hoc analysis. *Diabetes Therapy,* 6(3), pp. 317–328.

Huang, J. & Ling, C. X., 2005. Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering,* 17(3), pp. 299–310.

International Diabetes Federation, 2015. *IDF Diabetes Atlas.* 7th ed. Bruselj: International Diabetes Federations.

Jeon, C. Y. & Murray, M. B., 2008. Diabetes mellitus increases the risk of active tuberculosis: a systematic review of 13 observational studies. *PLoS Med,* 5(7), p. e152.

Johannesson, A. et al., 2009. Incidence of lower-imb amputation in the diabetic and nondiabetic general population: a 10-year population based cohort study of initial unilateral and contralateral amputations and reamputations. *Diabetes Care,* 32(2), pp. 275–280.

Johnson, A. E. et al., 2016. Data descriptor: MIMIC-III, a freely accessible critical care database. *Scientific Data,* 3(160035), p. 9.

Jorgensen, H. S., Nakayama, H., Raaschou, H. O. & Olsen, T. S., 1995. Intracerebral hemorrhage versus infarction: stroke severity, risk factors and prognosis. *Annals of neurology,* 38(1), pp. 45–50.

Kuhn, M. et al., 2017. *caret: classification and regression training.* [Online] Available at: https://CRAN.R-project.org/package=caret [Accessed 14. 06. 2017].

Le Gall, J. et al, 1984. A simplified acute physiology score for ICU patients. *Critical Care Medicine,* 12(11), pp. 975–977.

Lehman, L.-W. et al, 2012. Risk stratification of ICU patients using topic models inferred from unstructured progress notes. *AMIA Annual Symposium Proceedings,* 2012, pp. 505–511.

Marafino, B., Boscardin, W. & Dudley, R., 2015. Efficient and sparse feature selection for biomedical text classification via the elastic net: application to ICU risk stratification from nursing notes. *Journal of Biomedical Informatics,* 54, pp. 114–120.

McEwen, L. N. et al., 2012. Predictors of mortality over 8 years in type 2 diabetic patients: translating research into action for diabetes (TRIAD). *Diabetes Care,* 35(6), pp. 1301–1309.

Natekin, A. & Knoll, A., 2013. Gradient boosting machines, a tutorial. *Front Neurorobot,* 7(21), p. 21.

National Institutes of Health, 2014. *Protecting human research participants.* [Online] Available at: http://phrp.nihtraining.com/index.php [Accessed 3. 4. 2016].

Ogurtsova, K. et al., 2017. IDF Diabetes atlas: global estimates for the prevalence of diabetes for 2015 and 2040. *Diabetes Research and Clinical Practice,* 128, pp. 40–50.

Paulweber, B. et al., 2010. A european evidence-based guideline for the prevention of type 2 diabetes. *Hormone and Metabolic Research,* 42(Suppl. 1), pp. S3–S36.

Phillips, L. S., Ratner, R. E., Buse, J. B. & Kahn, S. E., 2014. We can change the natural history of type 2 diabetes. *Diabetes Care,* 37(10), pp. 2668–2676.

Praprott, R. et al., 2016. Validation of the German diabetes risk score among the general adult population: findings from the German health interview and examination surveys. *BMJ Open Diabetes Research & Care,* 4(1), p. 10.

R Development Core Team, 2017. *R: a language and environment for statistical computing.* [Online] Available at: http://www.R-project.org [Accessed 15. 06. 2017].

Rubino, F., 2008. Is type 2 diabetes an operable intestinal disease? A provocative yet reasonable hypothesis. *Diabetes Care,* 31(Suppl. 2), pp. S290–S296.

Shao, C.-Y. et al., 2015. CypRules: a rule-based P450 inhibition prediction server. *Bioinformatics,* 31(11), pp. 1869–1871.

Strickland, J., 2015. *Predictive Analytics using R.* Colorado Springs: Lulu.com