ISCTE IUL
University Institute of Lisbon

Department of Information Science and Technology

UNFOLDING THE DRIVERS FOR ACADEMIC SUCCESS: THE CASE OF ISCTE-IUL

A Dissertation presented in partial fulfilment of the Requirements for the Degree of Master in Computer Engineering

By

PAULO ALEXANDRE VIEIRA DINIZ FERREIRA GIL

Advisor:
PhD, Sérgio Moro, Assistant Professor,
ISCTE-IUL

Co-advisor:
PhD, Susana da Cruz Martins, Assistant Professor,
ISCTE-IUL

Setembro 2019

Abstract

Predicting the success of academic students is a major topic in the higher education research community. This study presents a data mining approach to predict academic success in a Portuguese University called ISCTE-IUL, unveiling the features that better explain failures. A dataset of 10 curricular years for bachelor's degrees has been analysed. Features' selection resulted in a characterising set of 68 features, encompassing socio-demographic, social origin, previous education, special statutes and educational path information. Understanding features' collection timings, distinct predicting was conducted. Based on entrance date, end of the first and the second curricular semesters, three distinct data models were proposed and tested. An additional model was designed for outlier degrees (i.e., a 4-year Bachelor). Six algorithms were tested for modelling. A support vector machines (SVM) model achieved the best overall performance and was selected to conduct a data-based sensitivity analysis. Relevance and impact review allowed extracting meaningful knowledge. This approach unfolded that previous evaluation performance, study gaps and age-related features play a major role in explaining failures at entrance stage. For subsequent stages, current evaluation performance features unveil their predicting power. Also, it should be noted that most of the features' groups are represented on each model's most relevant features, revealing that academic success is a combination of a wide range of distinct factors. These and many other findings, such as, age-related features increasing impact at the end first curricular semester, set a baseline for success improvement recommendations, and for easier data mining adoption by Higher Education institutions. Suggested guidelines include to provide study support groups to risk profiles and to create monitoring frameworks. From a practical standpoint, a data-driven decision-making framework based on these models can be used to promote academic success.

Keywords: Academic success; Data mining; Modelling; SVM; Features; Sensitivity analysis.

Resumo

O sucesso académico é um dos tópicos mais explorados nos estudos sobre o ensino superior. Este trabalho apresenta uma abordagem de *data mining* para a previsão do sucesso académico no ISCTE-IUL. Numa abordagem focada no insucesso, são estudados os fatores que explicam estes casos. Neste estudo foram utilizados dados de licenciatura de 10 anos curriculares. Foram analisadas 68 características sociodemográficas, origem social, percurso escolar anterior (ensino secundário), estatutos especiais e percurso académico. Foram adotados diferentes vetores de análise para o primeiro ano curricular (entrada e final dos primeiro e segundo semestres curriculares), dando origem a 3 modelos distintos. Um modelo suplementar foi projetado para cursos especiais. Entre os seis algoritmos de modelação testados, SVM obteve a melhor performance, sendo utilizado para a análise de sensibilidade. O processo de extração de conhecimento indicou que fatores como desempenho anterior, interrupções do percurso educacional e idade, demonstram grande impacto no (in)sucesso num estágio inicial. Nos estágios seguintes, fatores de performance atuais revelam um grande poder de previsão do (in)sucesso. A maior parte dos grupos de características faz-se representar, nas características mais relevantes de cada modelo. Estes e outros resultados, como o aumento do impacto dos fatores relacionadas com a idade no final do segundo semestre curricular, suportam a criação de recomendações institucionais. Por exemplo, criar grupos de apoio ao estudo para perfis de risco e criar ferramentas de monitorização são algumas das diretrizes sugeridas. Em suma, é possível criar uma ferramenta de apoio à decisão, baseada nos modelos apresentados, podendo ser utilizada pelo ISCTE-IUL para promover o sucesso académico.

Palavras-Chave: Sucesso académico; *Data mining*; Modelação; SVM; Características; Análise de sensibilidade.

Acknowledgments

I would like to acknowledge my advisor Professor Sérgio Moro for all the support and guidance since the beginning and through each stage of this work. Sérgio provided me with the tools that I needed to choose the right direction and successfully complete my dissertation. Thus, I had the chance to explore and enhance my understanding regarding each stage that comprises a structured data mining project.

I would like to thank my co-advisor, Professor Susana Martins for all dedication and shared knowledge, that provide me the opportunity to learn so much about higher education research field. Both provided me with brilliant insights that greatly improved the quality of this dissertation.

A special referral is due to António Casqueiro, António Luís Lopes, Pedro Ramos and the whole Information Systems Department of ISCTE-IUL for providing the academic dataset and their dedication on clarifying our questions about it.

I would like to express my very profound gratitude to my parents, close family and friends for all the support and continuous encouragement throughout my years of study and specially to my beloved wife through the process of researching and writing this dissertation. Finally, I would also like to dedicate this accomplishment to my firstborn Margarida that is about to be born. It would not have been possible without you all. Thank you.

Index

Table Index

Figure Index

Abbreviations

ABT	**A**nalytical **B**ase **T**able
ANN	**A**rtificial **N**eural **N**etworks
AUC	**A**rea **U**nder the ROC **C**urve
CC	**C**umulative **C**riteria
CRISP-DM	**Cr**oss-**I**ndustry **S**tandard **P**rocess for **D**ata **M**ining
DM	**D**ata **M**ining
DSA	**D**ata-based **S**ensitivity **A**nalysis
DT	**D**ecision **T**rees
ECTS	**E**uropean **C**redit **T**ransfer **S**cale
EDM	**E**ducational **D**ata **M**ining
ESCO	**E**uropean **S**kills, **C**ompetences, Qualifications and **O**ccupations
FPR	**F**alse **P**ositive **R**ate
GPA	**G**rade **P**oint **A**verage
HEI	**H**igher **E**ducation **I**nstitutes
LA	**L**earning **A**nalytics
MLPE	**M**ulti**l**ayer **Pe**rceptron
RF	**R**andom **F**orests
RQ	**R**esearch **Q**uestions
ROC	**R**eceiver **O**perating **C**haracteristic
SA	**S**ensitivity **A**nalysis
SQ	**S**earch **Q**ueries
SVM	**S**upport **V**ector **M**achines
TPR	**T**rue **P**ositive **R**ate

1. Introduction

Over the last decades, researchers on higher education are devoting more attention to academic success. While some studies establish theoretical frameworks to explain students' success, the vast majority analyse success through empirical research by considering the operational definition adopted by the studied institutions (York et al., 2015). Predicting students success has been a key topic for long in various scientific communities (Romero & Ventura, 2010). Researchers frequently seek insights regarding student characteristics and their impact on academic success.

Higher education institutes (HEI) are increasingly realizing the potential of their information systems and the data managed through it. Huge datasets originated by the most diverse activities and operations for the most diverse purposes are stored every day. As the volume of available data increases, the interest on exploring its potential and learn from it increases alongside (Canito et al., 2018). Data Mining (DM) is a computational method of processing large sets of existing data to obtain meaningful knowledge (Moro et al., 2018). Nowadays, data stored in educational databases is considerably large and increasing rapidly. These large datasets contain hidden patterns that can be explored through DM. Therefore, research areas such as higher education are expanding their interest in extracting meaningful and more complex knowledge from their data sources (Koedinger et al., 2008). Recently, a research area that combines DM and education has emerged and consolidated. Educational Data Mining (EDM) is a field that explores DM applied on different types of educational data (Howard et al., 2016). EDM uses data mainly obtained from educational information systems unfold knowledge and find answers to questions and problems concerning the education system.

This study aims to provide ISCTE-IUL - Instituto Universitário de Lisboa (ISCTE-IUL) meaningful information to increase academic success. It applies EDM techniques to an academic dataset provided by ISCTE-IUL information systems to unfold drivers for academic success. The resulting models' performance is evaluated and its suitability to predict potential success and failure cases are scrutinized. A knowledge extraction process is conducted, and the collected insights used to formulate guidelines and suggestions regarding institutional policies and pedagogical approaches to improve academic success. On an Institutional and management level, the suggested guidelines are expected to leverage decision-making, optimize allocation of educational resources and increase overall institutional productivity.

2. Literature review

Nowadays, HEI are challenged to provide the best curricula and programs as well as potentiate all means available into success. Pressure is intense in a context where very high value is attached to credentials. In fact, academic success concerns not only students and institutional but also governments, higher education policymakers and leaders (Kahu & Nelson, 2018). Great importance is being given by higher education' researchers to topics like student's dropout, persistence, learning processes and success. Conceiving the academic success as a final goal of a long run with diverse challenges and checkpoints, we could think of persistence as an endurance capability, enabler of success. Vincent Tinto (1997, 2006) explored five conditions that higher educational institutes might meet to enhance student persistence and success: institutional commitment, institutional expectations, support, feedback, and involvement or engagement.

Involvement and engagement are concepts strictly related with persistence that deserve especial attention within the higher education research community. Academic and social integration is also a condition for student success (Astin, 2012; Tinto, 1997). Kuh et al. (2006) describe and operationalize the concept of academic and social integration in ways that can be reasonably measured. The more students are academically and socially involved, the more likely they are to persist and graduate. This is especially true during the first year of study, when student membership is so tenuous yet so critical to subsequent learning and persistence (Tinto, 1999). Vincent Tinto (2006) points out and articulates a set of institutional initiatives for student success and additionally presents an important and well accepted theoretical model for institutional action that leads to success in higher education. As per this study, students enter an institution with a variety of abilities, skills, levels of high education preparation, attributes (such as social class, age and gender), attitudes, values and knowledge about higher education. At the same time students participate in external commitments like (family, work, community) (Figure 1). These set of features is being used as root to correlation and patterns studies regarding academic success.

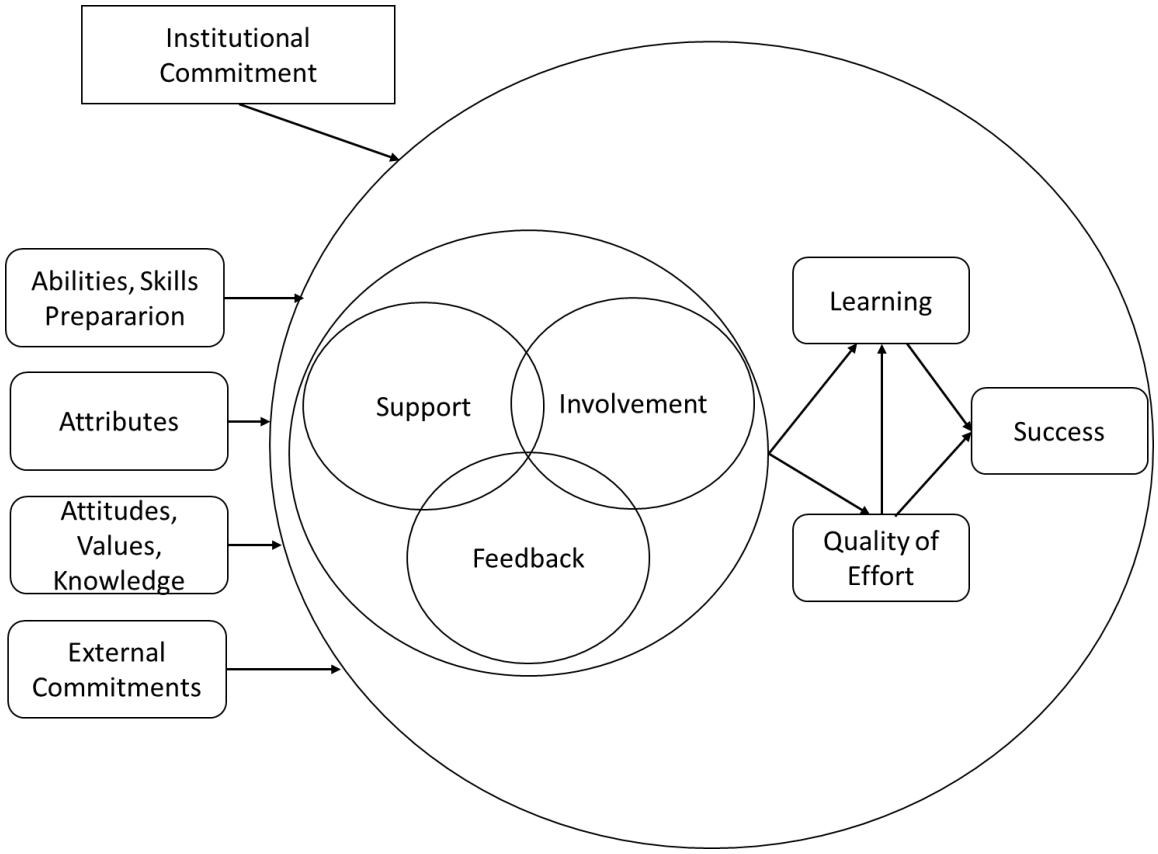


Figure 1 - Preliminary theoretical model for institutional action

(source: adapted from Tinto, 2006: p. 9).

Academic success concept is being applied as a wrap-up definition that aggregates a multiple number of student and institutional outcomes. Success, in conceptual terms, remains universal in its appeal and motivation for attainment or achievement of a goal (Hannon et al., 2017).

Research studies such as the ones by Parker et al. (2004) and Choi (2005) describe successful completion of course activities by students as ultimately improving students' academic achievement, defining assessment as Grade Point Average (GPA). There are a vast number of other studies that defines academic success as academic achievement referring to Grades or GAP (e.g., Gore Jr., 2006; Tracey et al., 2012). Kuh et al. (2006) and York et al. (2015) adopted the Astin's Inputs-Environments-Outcomes (IE-O) Model as the theoretical framework for their research, defining academic student success as the combination of the following factors: academic achievement, engagement in educationally purposeful activities, satisfaction, acquisition of desired knowledge, skills and competencies, persistence, attainment of educational outcomes, and post-college performance. The Astin model, first proposed in 1991 (Astin, 2012), clearly identifies academic success as an outcome of both input factors and the environment (Figure 2). The model also suggests that the environment functions as a mediator.

However, the relationship between environment and student outcomes cannot be understood without considering student inputs.

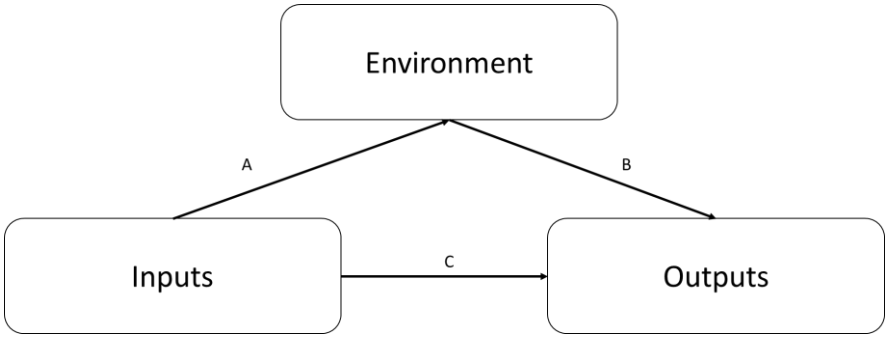


Figure 2 - Inputs-Environments-Outcomes (IE-O) Model

(source: Astin, 2012).

Pascarella & Terenzini (2005) extended Astin's framework by explaining higher education outcomes as functions of three sets of elements. The first set (inputs) is composed by demographic characteristics, family backgrounds, academic and social experiences that students bring to college. The second set (environment) encompasses people, programs, policies, cultures, and experiences that students encounter in HEI, whether on or off campus. The last set (outputs) includes students' characteristics, knowledge, skills, attitudes, values, beliefs, and behaviours as these exist after graduation. Academic success research track, based on how students perceive their own success, as well as their peers' success, has also gained some momentum recently. Students' perceptions of success depend on each individual's thoughts and beliefs. It is strongly influenced by the structures and cultures of the programs and demands of future occupation (Börjesson et al., 2016). Some relevant studies emphasize that successful students are described by their peers as those who achieve academic goals and at the same time are not apparently stressed by effort, still having time and energy for extra-academic activities and socializing. Nyström et al. (2013) coins this concept as 'stress-less achievement'. Wood and Breyer (2017) argues that achievement factors are relevant and relatively easy to measure, but also offer limited insight into the complexities of student success.

York et al. (2015) states that the improper usage of academic success definition by the research community when research aim is just narrowed to a portion of that concept may result in findings and conclusions that are not generalizable. The literature review conducted by York el al. (2015) highlights grades and GPA as the most used measures of academic success, accounting at making academic achievement the most commonly assessed aspect of academic success within higher education research community.

Assessments, as the main procedure for the measurement of studying outcomes, indicate the level of students' performance, which can be expressed qualitatively and quantitatively. According to Romero & Ventura (2007), the introduction of DM techniques in academic domains could improve decision-making processes in higher learning institutions. This improvement is expected to promote student's retention, transition rate and academic success (Figure 3).

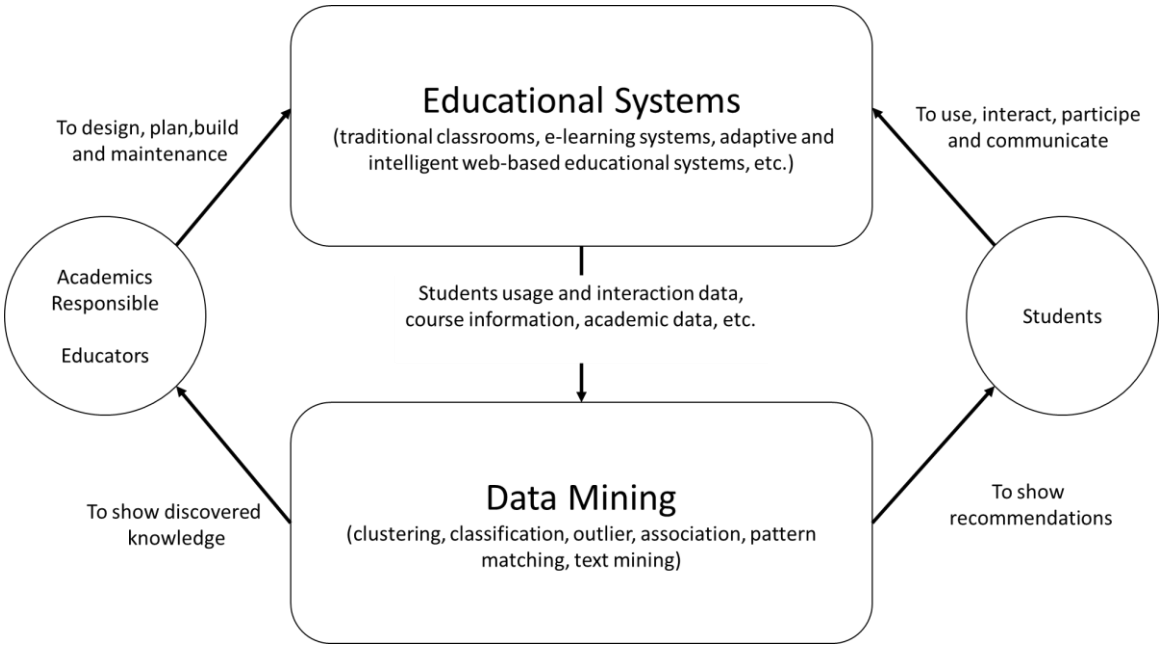


Figure 3 - Framework for data mining application in educational systems

(source: adapted from Romero & Ventura, 2007: p 2).

EDM is a field that explores data-mining approaches and techniques on different types of educational data, aiming at solving problems within the educational context (Baker & Yacef, 2009). It concerns to better understand students and the settings in which they learn (Baker, 2010). Over the years, students' enrolment and practicing in HEI has generated huge sets of student related data that may reflect the efficiency of the learning process (Koedinger et al., 2008). EDM seeks to use these data repositories to better understand learners and learning, and to understand patterns between academic features. Converting raw data originated by educational systems into useful information can potentially have a great impact on educational research and practice. Learning Analytics (LA) is an emerging technology-enhanced learning area based on online learning (Ferguson, 2012). Student's behaviour and usage data collected through online learning framework are being used recently as data source for DM analysis. Regardless of the origin, all DM techniques show one common characteristic: automated discovery of new relations and dependencies between attributes in the observed data.

The replicable approach for systematic literature review based on empirical studies presented by Kitchenham et al. (2009) has been adopted to collect relevant EDM studies that set academic success as the main DM goal. Thus, the following research questions have been addressed (RQ):

RQ1. How are DM researchers defining students' success?

RQ2. What are the main input features that influence the most the students' success?

A consolidated search strategy is crucial to assure the most relevant studies are retrieved during the search process (Santos et al., 2019). Therefore, 4 main search queries (SQ) have been designed by identifying the main academic success-related keywords in combination with the "data mining" term:

SQ1. "higher education success" AND "data mining"

SQ2. "academic success" AND "data mining"

SQ3. "higher education performance" AND "data mining"

SQ4. "academic performance" AND "data mining"

Each of these queries was executed on the following academic search engines since 2003: Google Scholar; ScienceDirect; Elsevier; ERIC; IEEEXplore. Searches were based on the title, abstract, and keywords (the sections that are indexed in most academic databases) and took place in July-August 2018. The collected knowledgebase was carefully reviewed and pruned, considering the following cumulative criteria (CC):

CC1. Clearly define academic success as the DM goal;

CC2. Clearly explore how DM model was defined and operationalized;

CC3. Present which input features were included in the model and resume their meanings;

CC4. Explicitly define context and data volume.

After applying the above inclusion/exclusion criteria, the body of knowledge was reduced to forty-one references, from which twenty-one are conference proceedings and the remaining twenty are journal articles. These references cover a fifteen-year timeframe from 2003 to 2018 and have been published under Computer Science/ Information Systems and/or Education focused' source titles (see Table A-1 in Appendix A for detailed stats). A few studies EDM literature review studies that identify current and future trends on the subject were also collected

during this research effort (Table A-2 in Appendix A details these studies on scope and contributions).

Table 1 characterizes the collect EDM studies on their main goals, the success metrics adopted, dataset volume, the number of features included in the model and the ones that influenced the most the success according to each study.

Academic success' modelling is significantly affected by diverse factors, such as, higher educational context, educational system and its specificities, available data, data granularity and data quality. Other aspects such as problem and modelling decisions lead to distinct operationalization of success and how it is measured. Regarding datasets there is a great diversity in terms of source, nature and volume. The data source used for empirical studies is mostly originated through surveys to students and/or form the HEI database. There is a wide variety of explanatory features to be found in literature, some studies made use of only few dozen features, while others counted with many dozens or even hundreds of features. Further analysis has been conducted through collecting the most observed features and grouping them in five distinct clustering groups. The following features' groups have been designed: socio-demographic features, social origin features, educational path features, previous education features and special statute features. Table 2 answers the following questions: which study includes each features' group in its DM model? Which are the three most used features within each group? Thus, it is possible to verify that approximately three quarters of the studies includes two or more distinct groups' representatives. By analysing each group share, it can be observed that more than three quarters of the studies include educational path features, more than two thirds include socio-demographic features and, approximately half include social origin features, previous education and special statute features. The importance of these feature's groups is discussed in academic success research literature and its usage incidence in data mining studies can be explained through great part of these being extracted from data that is generally collected as a requirement for admission or enrolment process. Individual analysis points gender as the most frequently used feature in overall literature. Student's behaviour and student's commitment are good examples of unconventional features that can be found in literature as well.

Summing up findings regarding RQ1, it is possible to cluster student's success operationalizations in the following main groups: passing grade in a specific module or course, passing grade in a specific exam, passing grade point average, student's graduation and student's

graduation with no failures. On RQ2, a wide spectrum of relevant explanatory features is observed, as there is a large number of distinct features pointed as the most relevant in the literature depending on each study's characteristics. There is no standard in the used datasets, as each study relies on distinct sources. Note that each study relies on data collected from potentially distinct academic systems and contexts with its own specificities. Even so, the following features' groups showed great impact on multiple studies: previous education features, educational path features and socio-demographic features.

Table 1 - Summary of the 41 reviewed studies.

Reference	Main Goal	Success Measure	Data Volume	Nr. Features	Relevant dependent variables
Minaei-Bidgoli et al. (2003)	Predict students' performance (Success). Identify students at risk.	Passing grade point average	Data set of 227 students who completed a homework set of 184 problems in LON-CAPA Web based system	10	Total_Correct _Answers Total_Number_of_Tries First_Got_Correct Time_Spent_to_Solve
Kotsiantis & Pintelas (2005)	Predict Students' marks and performance.	Passing the module (Introduction to Informatics)	Data set of 354 student records from Introduction to Informatics module	16	4th written assignment 3rd written assignment 2nd written assignment 4th face to face meeting Gender
Superby et al. (2006)	Classify students at early stages into three groups of risk	Move on to next academic year	Data set of 533 first-year university students' answers to 42 questions or question series.	375	High School last year' GPA Hours of mathematics in the last year of secondary education Hours students admit to attending class Student' confidence in his/her own abilities
Vandamme et al. (2007)	Predict academic performance of first-year students. Classify students into groups of risk.	Passing grade point average	Data set of 533 first-year university students' surveys	25	Attendance of courses by students Feeling of having made a good decision to enrol into the university
Romero et al. (2008a)	Compare several DM techniques for students' performance classification based on e-learning data	Passing grade point average	Data set of 438 Cordoba University students in 7 e-learning courses	11	Number of quizzes passed in e-learning framework Some others that are not described in the article.
Romero et al. (2008b)	Classify students based on e-learning data and the final marks obtained in their respective courses.	Passing grade point average	Data set of 438 Cordoba University students in 7 e-learning courses	11	Number of quizzes passed in e-learning framework Some others that are not described in the article.
Kabra & Bichkar (2011)	Predict the success (passing) of student in First Year of Engineering course.	Passing grade (final exam)	Data set of 346 first year of engineering students. University of Pune in Maharashtra, India	16	HSCCET - Marks obtained in common entrance test. SSCBoard - State, CBSE HSCPCM - Sum of Physics, Chemistry and Mathematics marks in HSC exam
Delen (2011)	Developed analytical models to predict	Passing the first year in University	Data set of 25,224 students' records - 1999 - 2006 timeframe - public	39	Earned/Registered = EarnedHours/RegisteredHours FallStudentLoan SpringStrudentLoan

Reference	Main Goal	Success Measure	Data Volume	Nr. Features	Relevant dependent variables
	freshmen student attrition.		university located in the mid-west region of the United States		ReceivedSpringAid AdmissionType
Barber & Sharkey, (2012)	Evaluate a likelihood of a given student failing/passing the current course.	Passing the current course	Data set of working adult students University of Phoenix, Arizona, USA.	24	Cumulative points earned (%) Points earned - prior courses (%) Ratio credits earned/attempted Point delta - prior courses >10%
Kovačić (2012)	Explore the profile of a student who successfully completes Information Systems course. Explore variables that influence persistence or dropout.	Students' persistence and graduation	Data set of 450 New Zealand students enrolled to Information Systems course	9	Ethnicity Course programme Course block
Osmanbegović & Suljić (2012)	Derive conclusions on students' academic success.	Passing grade at the exam (Business Informatics course)	Data set composed by Enrolment data and survey data collected from 257 students	12	GPA Entrance Exam Study Material Average Weekly hours devoted to studying
Watson et al. (2013)	Predict student's performance based upon various aspects of their ordinary programming behaviour	Passing grade	Data set of 45 students who studied Introduction to Programming (IP) course. University of Durham, UK 2012/2013	5 types	Programming behaviour features as a whole
Goker et al. (2013)	Apply data mining to improve an early warning system that may estimate the future academic success	Students' graduation	Data set of 200 students' records	25	Student's 9th level grades Absenteeism Knowledge Book reading habits Test Anxiety Primary school graduation certificate note Family income Parents educational qualification
Mishra et al. (2014)	Predict third semester performance	Third semester passing grade	Data set of 250 students' records collected from structured survey. Colleges affiliated to Guru Gobind Singh Indraprastha University, India	25	SECSEM - (% marks in 2nd Semester of MCA) GRAD - (% marks in Graduation) Leadership Drive - (Drive of the student)
Trstenjak & Donko (2014)	Predicting student success in Croatia using demographic features	Passing the first year in University - At least 42 ECTS at the end of the first year	Data set gathered from Croatian Information System of Higher Education Institutions (ISVU). Three years' timeframe 2010- 2012	15	Average grade Completed secondary school (the name of the secondary school that the student has completed) Student rights (designation of student rights and subventions)

Reference	Main Goal	Success Measure	Data Volume	Nr. Features	Relevant dependent variables
Slim et al. (2014)	Predict the performance of students early in their academic careers	Passing grade point average	Data set of 115746 students' records. University of New Mexico, Mexico.	1 per prior course	Prior Courses Grades
Olama et al. (2014)	Validate if the UMN LMS data elements are suitable to predict students success in early stages	Passing the course	Data set collected from Moodle. University of Minnesota (UMN). 2009 - 2013 timeframe	10	Homework
Martínez & Gómez (2014)	Determine patterns of academic success and failure for students	Students' graduation	Data set of TSAP course student's data. Institute of Curuzú, Cuatiá, Argentina. 2009-2013 timeframe	40	Type of residence (w/wo family) Student employment situation Weekly hours worked Relationship chosen career Parents educational qualification Priority assigned to the study
Simeunović & Preradović (2014)	Predict Students' Performance. Analyse factors which affect levels of success.	Passing grade point average	Data set of 354 students' records	17	Importance of mark Duration of studying Attendance at tests Intellectual capability
Natek & Zwilling (2014)	Predict the success rate of academic students	Passing final grade	Data set of 74 students' records in Informatics. 2010/11 - 2011/2012 timeframe + 34 students' records in Informatics 2012/2013.	8	Type of studies Age Employment Type of study
Mayilvaganan & Kalpanadevi (2014)	Predict academic students' performance. Improve classic Prediction/ classification techniques used in EDM to predict academic performance	Very Good/ Good/ Medium Learner	Data set of 197 student's records. PSG College of Arts and Science College, Coimbatore, India	18	Not clear
Hu et al. (2014)	Predict student learning performance based on activities in a fully online course	Passing the online course	Complete learning portfolio data of undergraduate students. Information Literacy and Information Ethics online course in a national university. 2009 - 2010 timeframe.	14	Course_LoginTime ReadTimeDOCCount Course_LoginAVGTime
Taruna & Pandey (2014)	Predict and classify students's marks in four distinct classes.	Student's graduation	Data set of 1000 undergraduate students' records. Engineering college in India.	11	Ag-g-7th - aggregate grades up to 7th semester Ag-g-3rd - aggregate grades up to 7th semester Backlogs

Reference	Main Goal	Success Measure	Data Volume	Nr. Features	Relevant dependent variables
Junco & Clem (2015)	Predicted final course grades using digital textbook usage metrics.	Passing course grade	Data set of 233 students. Iowa State University, USA	27	Number of days students spent reading
Cheewaprakobkit (2015)	Classifying a group of student academic achievement. Analyse factors affecting academic achievement.	Passing grade point average	Data set of 1600 students' records	22	Number of hours worked per semester
Stretch et al. (2015)	Predict individual performance of students in courses. Explore factors associated with success and failure.	Approval for classification. Passing grade average points for regression.	Data set of 5779 student's course	14	Not clear
Zimmermann et al. (2015)	Predictive value of undergraduate level indicators for subsequent graduate level success	Student's graduation	Data set of 171 student records in computer science	81	Third-year GPA
Zhou et al. (2015)	Predict students' performance of offline courses from their access records on general websites.	Passing Grade	Data set of 195 students' academic data + their internet access and navigation logs.	16	Visiting frequencies on 14 selected categories of websites Amount of time spent on online videos
Ahmad et al. (2015)	Predicting students' academic performance of first year bachelor students in Computer Science course	Passing the first year in University	Data set of 399 students' records. UniSZA, Malaysia. 8 years' timeframe 2006/2007 - 2013/2014	9	Not clear
Amrieh et al. (2015)	Predicting student's performance by including student's behaviour features	Passing total grade/mark	Data set of 150 students' records collected from e-learning system that called Kalboard 360	11	Raised hand on class Visited resources Announcements view Discussion Groups Relation
You (2016)	Identify indicators using LMS data to predict course achievement in online learning	Passing grade	Data set of 530 online course students in Gachon University, Republic of Korea	6	Regular Study Late submission Sessions Proof of reading the course information packets

Reference	Main Goal	Success Measure	Data Volume	Nr. Features	Relevant dependent variables
Badr et al. (2016)	Predict students's performance in a programming course based on their grades in other courses.	Passing grade in programming course	Data set of 203 graduate students. King Saud University (KSU). 2008–2014 Timeframe	57	English course performance
Vuttipittayamongkol (2016)	Predict students' success by using three main features: activity hours, English scores, and number of students admitted	Students' graduation	Data set of undergraduate students' records. Mae Fah Luang University, Thailand. 1999-2014 timeframe.	3 main features groups	English I course performance
Daud et al. (2017)	Explore the impact of proposed features on student performance prediction	Students' graduation	Data set of Around 700 students after data cleansing	23	Family expenditure Personal information features
Martins et al. (2017)	Explore Success Factors in Portuguese Higher Education	Students' graduation with no failures	Data set of 3000 students taken from academic surveys	32	Characteristics of the educational institution Type of education Field of education Age Student's commitment Schooling conditions at the starting point
Asif et al. (2017)	Predict performance and investigate how student's performance progresses during their studies	Students' graduation	Data set of 210 undergraduate students' records enrolled in a 4-year Information Technology bachelor's degree of a public engineering university in Pakistan	8	Pre-university marks 1st year courses' marks 2nd year courses' marks
Rahman & Islam (2017)	Study the impact of behavioural and student absent features on students' success	Passing performance	Data set called Student's Academic Performance data set' collected from e-learning system that called Kalboard 360. Comprises 480 instances	16	Absent days in class Behaviour features (such as raise hand on class, resource review, and group discussion etc.
Leppanen et al. (2017)	Study the impact of material usage on students' success	Passing grade in introduction to programming course	Data set of 271 students' records collected using a client-side data gathering component applied to e-learning content for Introduction to programming course. University of Helsinki	Not clear	Element-level usage
Kostopoulos et al. (2018)	Estimate students' academic success in a web-based university course combining	Passing the module (Introduction to Informatics)	Data set composed by a total of 3882 instances. Introduction to informatics	17	Type of employment New student OCS2, OCS4 and OCS5 (students' presence in the optional contact sessions

Reference	Main Goal	Success Measure	Data Volume	Nr. Features	Relevant dependent variables
	classification and regression rules		module of the computer science course. Three years' timeframe (2008–2010)		WR1, WR2, WR3 (students' grades in the written assignments)
Martins et al., (2018)	Predict the undergraduate academic performance of students from a Portuguese higher education institution	Students' graduation	Data set of 2159 students' records. Distinct bachelor's degree programmes in a Portuguese higher education institute. 2007 - 2015 timeframe	50	Ects_reprov_s Media_s Cod_escola Ects_aprov_s Ects_cred_tx Navalr_s Cod_curso Max_s Nível_esc_mae Ano_s Ano_mat
Fernandes et al. (2018)	Predict whether a student will pass/fail at the end of the school year.	Passing at the end of the school year	Data set of 238,575 records in 2015 and 247,297 records in 2016 - Public universities in Brasília district, Brazil	17	Grade Neighborhood School School_subjects Absence City Age

Table 2 – Features' groups used in EDM literature.

	Research Studies																																					
Features	Minaei-Bidgoli et al. (2003)	Kotsiantis & Pintelas (2005)	Superby et al. (2006)	Vandamme et al. (2007)	Romero et al. (2008a)	Romero et al. (2008b)	Kabra & Bichkar (2011)	Delen (2011)	Barber & Sharkey, (2012)	Kovačić (2012)	Osmanbegović & Suljić (2012)	Watson et al. (2013)	Goker et al. (2013)	Mishra et al. (2014)	Trstenjak & Donko (2014)	Slim et al. (2014)	Olama et al. (2014)	Martínez & Gómez (2014)	Simeunović & Preradović (2014)	Natek & Zwilling (2014)	Mayilvaganan & Kalpanadevi	Hu et al. (2014)	Taruna & Pandey (2014)	Junco & Clem (2015)	Cheewaprakobkit (2015)	Stretch et al. (2015)	Zimmermann et al. (2015)	Zhou et al. (2015)	Ahmad et al. (2015)	Amrieh et al. (2015)	Badr et al. (2016)	You (2016)	Vuttipittayamongkol (2016)	Daud et al. (2017)	Martins et al. (2017)	Asif et al. (2017)	Rahman & Islam (2017)	Leppanen et al. (2017)	Kostopoulos et al. (2018)	Martins et al. (2018)	Fernandes et al. (2018)	**Total**
SD		X	X	X			X	X	X	X	X		X	X	X			X	X	X	X		X	X	X	X	X		X	X				X	X		X		X	X	X	28
Gender		X	X	X			X	X	X	X	X		X	X	X			X	X	X	X		X	X	X	X	X		X	X				X	X		X		X	X	X	28
Age		X	X	X				X	X	X			X					X		X	X		X		X	X	X							X			X		X	X	X	19
Marital Status		X	X	X				X	X				X		X			X							X	X								X			X		X			13
Other		X	X	X			X	X	X	X	X			X	X			X	X				X	X	X	X			X	X				X			X		X	X	X	23
SO			X	X			X		X		X		X	X	X			X			X				X				X				X	X				X	X	X	17
Parents' Education			X	X									X	X	X			X							X				X					X					X		9
Parents'			X	X			X						X	X	X			X							X				X					X					X		9

Occupation																																										
Number of Household Members			X	X					X		X							X																								5
Other			X	X					X		X		X	X	X			X			X				X				X					X	X				X	X	X	16
EP	X		X	X	X	X	X	X	X	X		X		X		X	X				X	X	X	X	X		X	X	X	X	X	X	X		X	X	X	X	X	X	X	32
Course					X	X	X	X		X				X									X	X			X			X					X					X		12
Degree Type							X	X		X				X									X				X								X					X	X	9
HEI Type			X	X				X																											X					X	X	6
Other	X		X	X	X	X	X	X	X			X		X		X	X				X	X	X	X	X		X	X	X	X	X	X	X			X	X	X	X	X	X	30
PE			X	X			X	X	X	X	X		X	X	X			X			X		X	X					X					X	X	X				X		19
Entry Exams			X	X			X				X																								X					X		6
Entry Grade			X	X			X	X	X		X										X		X																	X		9
Entry Exams Grade			X	X			X																													X				X		5

Other			X	X				X	X	X	X		X	X	X			X			X		X	X				X																18
SS		X	X	X				X	X		X		X	X	X			X		X					X	X								X	X		X		X	X	X		19
Social Services			X	X				X	X		X			X												X																		7
Scholarship			X	X				X			X				X											X	X														X			8
Working Student		X	X	X					X				X					X		X						X								X	X		X		X	X			13
Other			X	X				X	X				X	X	X					X					X	X															X	X		12
Total	1	2	5	5	1	1	4	4	5	3	4	1	4	5	4	1	1	4	1	2	4	1	3	3	4	2	2	1	4	2	1	1	1	4	5	2	3	1	4	5	4		

Features: SD = Socio-demographic, SO = Social origin, EP = Educational Path, PE = Previous Education, SS = Special statutes

3. Materials and methods

Cross-Industry Standard Process for Data Mining (CRISP-DM) is the methodology implemented in this study. It is the most adopted methodology within DM' research domain and was designed to be applied in real-world business cases, helping to support business decisions and increasing DM projects' success. It defines a project as a cyclic process and applies a non-rigid sequence of six main stages (Chapman et al, 2000). As a flexible methodology, CRISP-DM allows and encourages iterative process, so it is a common practice to move back and forth between distinct stages promoting final DM results. CRISP-DM main stages are business understanding, data understanding, data preparation, modelling, evaluation and deployment (Shearer, 2000). As described in Chapman et al. (2000) the first three main stages are critical for the DM study's success and it is where the main implementation's effort and focus are required. R tool is an open source programming language and environment for statistical and data analysis. It is highly extensible and flexible, counting with a great variety of libraries and packages. There is a dynamic and enthusiastic DM research community, supporting and adopting R on their research studies. This study bases on R tool to apply CRISP-DM methodology. According to Roy & Garg (2017), R tool alongside with WEKA, is the most adopted tool in EDM literature. Furthermore, Rminer library, as an integrated R tool framework is used to implement complex DM analysis though an extensively documented set of functions (Cortez, 2010).

3.1. Domain understanding and data mining problem definition

ISCTE-IUL is a public HEI located in Lisbon, Portugal. It was established in 1972 and currently counts with approximately, 9000 students enrolled in undergraduate (46%) and postgraduate (54%) programmes, 450 teachers and 220 non-teaching staff. ISCTE-IUL provides 89 distinct degree programmes through the following school, i.e., study areas: Business, Sociology and Public Policy, Social Sciences, Technology and Architecture. Fénix@ISCTE-IUL (Fénix) is the information system adopted by ISCTE-IUL to manage educational processes. It allows candidates, students, teaching and non-teaching staff to have on-line access to services provided by ISCTE-IUL. Admission process management, applications and programme enrolments, academic path and evaluation monitoring are some critical features provided by Fénix. As part of the information process, Fénix manages and stores highly relevant academic-related data. It

sets the data source used in this study. This data was provided by the Information Systems Department of ISCTE-IUL and have been anonymised before being exposed for analysis. This procedure ensures that students is no longer identifiable through the provided dataset.

Business Understanding is CRISP-DM's kick-off stage and is based on understanding the domain and requirements from a business perspective and defining the main DM goal. Preliminary understanding efforts showed that is critical for HEI purposes, to maximize academic success rate. Thus, the ability to detect potential failures and dropouts, in early stages, assumes great importance to educational stakeholders. Understanding that ISCTE-IUL is interested to promote student's success, it is critical to focus on potential unsuccessful students. Therefore, a-priori detection of failures is set as a success enhancement feature for HEI. As important as its predicting capacity, the DM model, needs to be suitable for ISCTE-IUL to apply success enhancement actions in time to prevent failures. Thus, the earlier in the academic path, the DM model could predict failures, avoiding a high level of incorrectly predicted cases (false positives), the better.

As discussed in previous section, each EDM's approach depends on each specific higher educational system and the academic dataset it relies on, among other significant context characteristics. It leads studies to distinct dependent variable operationalization and how it is measured. This study adopts student's graduation with no failures as student's success operationalization. Thus, DM's goal and the main analysis' subject are devoted to predicting students that would not complete their degree's programme within the optimal number of curricular years. In other words, students that fails and/or repeats at least one curricular year. This study follows a classification DM approach, as it builds a predictive model that classifies a data record into one of two predefined classes. Predefined classes used for success are "Failure" and "Success".

3.2. Unfolding a higher education institution database

Data understanding stage is composed by data obtaining, data describing, data exploring and initial data quality review tasks. Academic dataset collected by Fénix set up the data source for this study's DM model. Fénix was implemented in 2008 and imported historical data from previous information systems. Its database uses a relational model composed by SQL tables. Among these tables there are a dozen main tables, covering candidacy, registration, enrolment,

evaluation, statutes and social services data mostly. A dozen lookup and matching tables are also provided, supporting data understanding and data quality tasks (see Figures B-1 and B-2 in Appendix B for data source model). There are three main tables within data source that defines student-HEI relationship: *Person*, *Candidacy* and *Registration* tables. *Person* table stores general student's data as socio-demographic and social origin data and new records are inserted during the admission process. *Candidacy* table stores data regarding student's candidacy to a degree provided by ISCTE-IUL. Note that the relationship between these two tables allows multiple candidacies to a single person. For instance, a student that interrupts and is transferred to a distinct degree, or a student that completes a bachelor's degree and apply for a master's degree, are some good examples. *Registration* table establishes the relationship between *Candidacy* table and relevant tables, such as, evaluations, courses and statutes related tables. It stores the current registration state that indicates whether the student completed registration's related degree.

After an initial data exploring process and stepping back to revisit DM goal's operationalization, it was detected that degree's standards have changed through time, such as the case of, Pre-Bologna and Post-Bologna process' degrees. Encompassing 1978/1979 to 2018/2019 curricular timeframe, the dataset includes both standards as well as the transition period. DM goal's operationalization may ensure that each subject of the prediction is comparable and defined through a well-identified characterization set for each feature. Pre-Bologna and Post-Bologna degrees are not comparable in terms of the adopted goal operationalization, as they differ in success premises and conditions, for instance, different curricular years' total. As per the same rational, bachelor's degree, master's degree and PhD's degree are not comparable as well. ISCTE-IUL implemented Bologna process in the curricular year of 2006/2007 and it is still currently in place. Priority criteria, such as, current higher education process and most recent and higher quality data, set up bases for selecting the subset of data to be used in this study. Accordingly, the original dataset was reduced to Bologna bachelor's degrees related records. General case, Bologna process sets a programme of 3 curricular years for bachelor's degrees, through which, passing all curricular courses, the student earns 180 ECTS (European Credit Transfer scale), completing it. Considering these premises and DM goal's operationalization, 2016/2017 to 2018/2019 entry curricular years' data were removed from the analysis, as it is not possible to calculate these cases' success. Another important issue is student's mobility within HEI, as students can be transferred from a HEI to another HEI. Most of the times transferred students can keep earned ECTS and

completed courses' evaluation. In these specific cases, it is not possible to guarantee success operationalization requirements as available data source doesn't provide information originated or collected by other HEI. This limitation forces transferred students' data to be excluded from the analysis.

Summing up decisions at this point, this analysis will be focused on bachelor's students who effectually enrolled a programme provided by ISCTE-IUL, between 2006/2007 and 2015/2016 (10 years' timeframe). This ensures success operationalization requirements to be met. Further data review revealed that, even being considered Bologna process's degrees, IGE and IGE-PL set a 4 curricular years' programme, so a similar analysis is conducted separately for these specific degrees. A further modelling and comparative analysis are performed to review features' relevance between regular bachelors and 4-years bachelors.

After redefining DM' requirements, an initial analytical base table (ABT) is created to aggregate features' candidates, originally spread among distinct tables of the relational database model. ABT is a flat table used for building analytical models, such as DM models. Each ABT's record represents the subject of the model and stores feature's data representing the subject. ABT records are set at registration level, so further data quality tasks and computed measures are designed ensuring registration's granularity. Table 3 summarizes the first ABT version composed by thirty-two features, by features' group, origin and main source table. Note that this initial gathering, are composed by multiple representatives for each of the five features' groups reviewed on literature (see Table C-1 in Appendix C for detailed feature's description and their classes).

Table 3 - First ABT version.

Feature	Features' Group	Origin	Source Table
Area	Socio-demographic	Extracted	
areaCode	Socio-demographic	Extracted	
Gender	Socio-demographic	Extracted	
yearOfBirth	Socio-demographic	Extracted	
fatherOccupation	Social Origin	Extracted	
motherOccupation	Social Origin	Extracted	
fatherOccupationConditionType	Social Origin	Extracted	Person
motherOccupationConditionType	Social Origin	Extracted	
Occupation	Socio-demographic	Extracted	
iscteFirstExecutionYear	Educational Path	Extracted	
maritalStatusType	Socio-demographic	Extracted	
Nationality	Socio-demographic	Extracted	
secondNationality	Socio-demographic	Extracted	

Feature	Features' Group	Origin	Source Table
entryYear	Educational Path	Extracted	
fatherLiteraryHabilitationType	Social Origin	Extracted	
motherLiteraryHabilitationType	Social Origin	Extracted	
workingStudentAtEntry	Special Statute	Extracted	Person Special Regime
partialTimeStudentAtEntry	Special Statute	Extracted	
specialEducationNeedsAtEntry	Special Statute	Extracted	
scholarshipAtEntry	Special Statute	Extracted	
dislocatedAtEntry	Special Statute	Extracted	
degreeCode	Educational Path	Extracted	Candidacy
degreeType	Educational Path	Extracted	
degreeSchool	Educational Path	Extracted	
entryGrade	Previous Education	Extracted	
precedentDegreeDesignation	Previous Education	Extracted	
precedentConclusionYear	Previous Education	Extracted	
secondarySchoolType	Previous Education	Extracted	
Ingression	Previous Education	Extracted	
highSchoolDegreeType	Previous Education	Extracted	
iscteWasFirstChoice	Previous Education	Extracted	
erasmusOutgoing	Educational Path	Extracted	Outgoing Mobility

Further data exploring efforts showed up a great number of new potential feature's candidates. There are student statutes and social services data to be found, stored by curricular semester and curricular year, respectively. This data relates to special statute features' group and the way it is stored allows special statute' time-based features to be created. Thus, fifty-four derived features total were created at this stage, a single feature per student statute by first curricular year's semester and a single feature per social service by first curricular year. Social services features were also split in two categories: accepted and requested. A requested social service doesn't mean that it has been accepted, so it adds extra detail to the analysis. Note that, as discussed in the beginning of this section, early-stage prediction is one of the main motivations for this study, so no data collected after the first curricular year are considered for feature's gathering purposes. In other words, features' gathering concentrates in freshmen characteristics' scrutiny. Table 4 reviews new derived features added to ABT (see Table C-1 in Appendix C for detailed feature's description and resultant classes).

Table 4 - Additional derived features included in ABT.

Feature	Features' Group	Origin	Source Table
workingStudentEntryYear1stSem	Special Statute	Derived	
InternationalStudentEntryYear1stSem	Special Statute	Derived	
partialTimeStudentEntryYear1stSem	Special Statute	Derived	
fctgrantOwnerEntryYear1stSem	Special Statute	Derived	
classSubRepresentativeEntryYear1stSem	Special Statute	Derived	
classRepresentativeEntryYear1stSem	Special Statute	Derived	
handicappedEntryYear1stSem	Special Statute	Derived	
pregnantOrChildrenUnder3EntryYear1stSem	Special Statute	Derived	
professionalAthleteEntryYear1stSem	Special Statute	Derived	
sasGrantOwnerEntryYear1stSem	Special Statute	Derived	
militaryEntryYear1stSem	Special Statute	Derived	
temporaryDisabilityEntryYear1stSem	Special Statute	Derived	
religiousEntryYear1stSem	Special Statute	Derived	
associativeLeaderEntryYear1stSem	Special Statute	Derived	
iscteAthleteEntryYear1stSem	Special Statute	Derived	
firefighterEntryYear1stSem	Special Statute	Derived	
erasmusGuestEntryYear1stSem	Special Statute	Derived	
deathOfSpouseOrFamilyEntryYear1stSem	Special Statute	Derived	
appearancePoliceOrMilitaryAuthorityEntryYear1stSem	Special Statute	Derived	
monitorEntryYear1stSem	Special Statute	Derived	Student Statutes
previousIBSStudentEntryYear1stSem	Special Statute	Derived	
top15IBSEntryYear1stSem	Special Statute	Derived	
workingStudentEntryYear2ndSem	Special Statute	Derived	
InternationalStudentEntryYear2ndSem	Special Statute	Derived	
partialTimeStudentEntryYear2ndSem	Special Statute	Derived	
fctgrantOwnerEntryYear2ndSem	Special Statute	Derived	
classSubRepresentativeEntryYear2ndSem	Special Statute	Derived	
classRepresentativeEntryYear2ndSem	Special Statute	Derived	
handicappedEntryYear2ndSem	Special Statute	Derived	
pregnantOrChildrenUnder3EntryYear2ndSem	Special Statute	Derived	
professionalAthleteEntryYear2ndSem	Special Statute	Derived	
sasGrantOwnerEntryYear2ndSem	Special Statute	Derived	
militaryEntryYear2ndSem	Special Statute	Derived	
temporaryDisabilityEntryYear2ndSem	Special Statute	Derived	
religiousEntryYear2ndSem	Special Statute	Derived	
associativeLeaderEntryYear2ndSem	Special Statute	Derived	
iscteAthleteEntryYear2ndSem	Special Statute	Derived	
firefighterEntryYear2ndSem	Special Statute	Derived	
erasmusGuestEntryYear2ndSem	Special Statute	Derived	
deathOfSpouseOrFamilyEntryYear2ndSem	Special Statute	Derived	

Feature	Features' Group	Origin	Source Table
appearancePoliceOrMilitaryAuthorityEntryYear2ndSem	Special Statute	Derived	
monitorEntryYear2ndSem	Special Statute	Derived	
previousIBSStudentEntryYear2ndSem	Special Statute	Derived	
top15IBSEntryYear2ndSem	Special Statute	Derived	
requestedSocialServiceEntryYear	Special Statute	Derived	Social Services
acceptedSocialServiceEntryYear	Special Statute	Derived	
requestedSStransportSupplementEntryYear	Special Statute	Derived	
requestedSSaccommodationSupplementEntryYear	Special Statute	Derived	
requestedSSresidenceRequestEntryYear	Special Statute	Derived	
requestedSSiscteFinantialSupportEntryYear	Special Statute	Derived	
acceptedSStransportSupplementEntryYear	Special Statute	Derived	
acceptedSSaccommodationSupplementEntryYear	Special Statute	Derived	
acceptedSSresidenceRequestEntryYear	Special Statute	Derived	
acceptedSSiscteFinantialSupportEntryYear	Special Statute	Derived	

Further data understanding effort exposed potential features based on pre-existing data. According to Barraza et al. (2019), feature engineering is key for data mining, Therefore, new features were designed applying non-straightforward logic, requiring distinct transformation, aggregation or/and calculation processes. These new set of features are labelled computed features due to the processes involved in their creation.

Candidacy preference related-table details higher education candidacy process. Each table's record specifies a candidacy preference by its order, HEI and degree, and stores the whole set of student's entry exams and respective grades. Portuguese higher education admission process allows a student to choose up to six candidacy preferences, that results in multiple records, per student. Therefore, five new computed features, encompassing previous education features' group, were designed for candidacy preference. Four of those, considering the relationship between student's preference, HEI and degree student ended up registering, and a fifth one for entry exams grades average. For instance, new computed feature, orderPreference, represents the order in which student has chosen HEI and degree he ended up registering.

Student's evaluation related data is stored by course and curricular semester. Evaluation records are updated each time there is an attempt to complete a specific course or improve its grade, reflecting its most recent state. Thus, new computed evaluation related features were designed, comprising overall evaluation by each semester of the first curricular year. Each course set up its ECTS, so each time a student achieves a passing grade in a specific course, its correspondent ECTS amount is earned. It is important to note that single course's analysis is not possible, as

each degree sets its specific programme, composed by distinct courses' set. Thus, six new computed educational path features are designed for student's evaluation. For instance, new computed feature, ectsCreditsEntryYear1stSem represents the total amount of ECTS earned in the first curricular semester. It is important to detail that weighted average features were calculated relying on the premise that 30 ECTS are the optimum amount of ECTS to be collected per semester. It is an approximate value, set for comparable analysis, as it is possible for students to earn more than 30 ECTS per semester. For instance, a student that passes all current semester courses and a pending previous semester course, will exceed 30 ECTS.

Additional computed features representing, student's age at entry, study gap time between precedent and current educational degree, and student's residence location are also developed. Residence location features are calculated by integrating Postal Code database[1], provided by the Portuguese Postal Office (CTT). This is public-domain data covered by Open Data Commons Public Domain Dedication and License (PDDL). This information set up lookups for location data to be represented in distinct aggregation levels, best suited for DM.

Table 5 represents computed features added to ABT (see Table C-1 in Appendix C for detailed feature's description and resultant classes)

Table 5 - Additional computed features included in ABT.

Feature	Features' Group	Origin	Source Table
firstChoice	Previous Education	Computed	Candidacy Preference
firstChoiceUniversity	Previous Education	Computed	
firstChoiceCourse	Previous Education	Computed	
orderPreference	Previous Education	Computed	
gapEntryExames	Previous Education	Computed	
entryAge	Educational Path	Computed	Person
entryAgeRange	Educational Path	Computed	
municipality	Socio-demographic	Computed	
district	Socio-demographic	Computed	
lisbonMetropolitanArea	Socio-demographic	Computed	
studyGap	Previous Education	Computed	Candidacy
studyGapYears	Previous Education	Computed	
ectsCreditsEntryYear1stSem	Educational Path	Computed	Enrolment
ectsCreditsEntryYear2ndSem	Educational Path	Computed	
averageEntryYear1stSem	Educational Path	Computed	
weightedAverageEntryYear1stSem	Educational Path	Computed	
averageGradeEntryYear2ndSem	Educational Path	Computed	

[1] Available for download from CTT website: https://www.ctt.pt/feapl_2/app/restricted/postalCodeSearch/postalCodeDownloadFiles.jspx

Feature	Features' Group	Origin	Source Table
weightedAverageEntryYear2ndSem	Educational Path	Computed	

3.3. Preparing the Analytical Base Table

Data preparation stage requires to take decisions on final features' set, establishing the foundation for modelling. Relevant tasks such as, data cleaning, data reduction and data discretization are performed at this point. Criteria such as, data quality, volume limitations and its relevance to DM goal are considered for decision taking. This final process, before moving towards modelling stage, was conducted thought 5 approaches.

The first approach consisted in data generalization, through replacing low level attributes with high level concepts. Following this guiding principle, low level residence location features have been removed. Replacing features make it possible for DM to apply location analysis, through higher-level concepts, allowing better modelling performance. The same reasoning has been applied to high school degree related features (Table 6).

Table 6 – High level concepts for data generalization

Replaced feature	Reason	Replaced by
area	Thousand distinct classes	district and lisbonMetropolitanArea
areaCode	Thousand distinct classes	
municipality	Hundred distinct classes	
precedentDegreeDesignation	Hundred distinct and bad quality classes	highSchoolDegreeType

On its turn, fatherOccupation and motherOccupation features were originally represented by hundreds of low-quality distinct values. A conceptual review process was carried out to design a meaningful higher aggregation level, setting bases for appropriate modelling. Six distinct classes were designed, from "Elementary occupations" to "Managers" taking ESCO[2] (European Skills, Competences, Qualifications and Occupations) multilingual classification of occupations, as reference (see Table C-2 in Appendix C for designed lookup table). Approximately half of these features' original data is missing or bad quality, those cases were replaced by "Others/Unknown". Due to this handling approach, a further performance

[2] ESCO is a Europe 2020 initiative, the current version is ESCO v1.0.3 (Last update 26/04/2018). DG Employment, Social Affairs and Inclusion of the European Commission developed ESCO in collaboration with stakeholders and with the European Centre for the Development of Vocational Training (Cedefop).

evaluation is conducted to validate both features impact in the model. Likewise, occupation feature is originally represented by hundreds of distinct values, most of them revealing poor quality. This feature's original data was replaced by 3 distinct classes, "Unknown" for missing values, "Student" for filled student-related values and "Filled Occupation" for filled occupation-related cases.

The second approach consisted in dealing with missing data' features. Hotdeck imputation algorithm was applied to entryGrade feature, which, approximately 18% data is missing. Hotdeck algorithm uses k-nearest neighbour method to identify the most similar case, based on a set of features, and replaces the missing data, by the value found in such example. A features' set comprised of 13 social origin and socio-demographic features were selected to feed hotdeck algorithm for similar case identification process (see Table C-3 in Appendix C). Further performance evaluation is performed to validate the impact of this feature in the model.

The following six previous education group' features: highSchoolDegreeType, firstChoice, firstChoiceUniversity, firstChoiceCourse, orderPreference and gapEntryExams, are particular cases of missing data features. Fénix stores candidacy preference data from 2013/2014 curricular year onwards. In contrast to regular data quality issues, these features' data are complete since the referred year. On the other hand, including these features in the model forces to discard 70 % of the dataset. Despite this fact, it was decided to keep these features and conduct a further performance evaluation so their impact on model's goal could be scrutinized. This features' set will hereafter be referred to as 30%FilledFeatures.

For the remaining missing data' features, a 1% threshold was set up for decision taking. "Unknown" dummy class was used to replace missing data for substantial missing data' features, greater or equal 1%. Missing data' records for features with residual missing values incidence, above 1%, were excluded (Table 7).

Table 7 - Missing data features handling

Feature	Nulls %	Action
fatherOccupationConditionType motherOccupationConditionType fatherLiteraryHabilitation motherLiteraryHatilitation secondarySchoolType secondNationality	>/= 1%	missing data replace by "Unknown" class
yearOfBirth nationality	< 1%	missing data records removed

precedentConclusionYear	
degreeCode	
ingression	

The third approach consisted in reviewing unnoticed or hidden dependencies between the DM goal and each feature. For instance, partial-time students are unable to meet operationalized success requirements. It is certain that a part-time student takes more than optimized number of curricular years to complete a degree. So, all records, which partialTimeStudentAtEntry is true were excluded. This action avoids further modelling to be influenced by this dependency. Consequently, partialTimeStudentAtEntry feature was removed from ABT. In contrast, records related to students that requested partial time statute during the 1st curricular year, represented by partialTimeStudentEntryYear1stSem and partialTimeStudentEntryYear2ndSem, were not excluded from the analysis. These two features were removed from ABT, as their impact to model's goal could introduce noise to overall features relevance.

The fourth approach consisted in removing single class features. A clear example is degreeType feature, that due to this proposed scope, is only represented by a single class: bachelor. As detailed in Table 8 a total of 22 features have been excluded from ABT as they comprise a single class.

Table 8 - Removed single class's features

Feature	Reason
degreeType	Single class for whole dataset
fctgrantOwnerEntryYear1stSem	
militaryEntryYear1stSem	
religiousEntryYear1stSem	
firefighterEntryYear1stSem	
erasmusGuestEntryYear1stSem	
deathOfSpouseOrFamilyEntryYear1stSem	
appearancePoliceOrMilitaryAuthorityEntryYear1stSem	
monitorEntryYear1stSem	
previousIBSStudentEntryYear1stSem	
top15IBSEntryYear1stSem	
fctgrantOwnerEntryYear2ndSem	
militaryEntryYear2ndSem	
religiousEntryYear2ndSem	
firefighterEntryYear2ndSem	
erasmusGuestEntryYear2ndSem	
deathOfSpouseOrFamilyEntryYear2ndSem	

appearancePoliceOrMilitaryAuthorityEntryYear2ndSem
monitorEntryYear2ndSem
previousIBSStudentEntryYear2ndSem
top15IBSEntryYear2ndSem
erasmusOutgoing

The fifth and final approach is based on outliers, impossible values and conflicting data' features. Table 9 details approaches followed for each of these features.

Table 9 - Outliers and conflicting data features approach

Feature	Issue	Fix
precedentConclusionYear	Outlier data: "12"; "9999"	Issued cases removed
studyGap	Outlier data: negatives values	Issued cases removed
iscteWasFirstChoice	Doesn't match firstChoiceUniversity values (residual cases)	Feature was removed
firstChoiceUniversity	Doesn't match iscteWasFirstChoice values (residual cases)	Kept as it shares baseline with other features

Table 10 summarizes the final ABT by features' group, data type and collection time, as the result of data preparation stage. Final dataset is composed by a total 9652 records for regular bachelor's degrees and 789 records for 4-year bachelor's degrees. A total of 74 features are provisionally represented, 36 special statute features, 12 education path features, 12 previous education features, 10 socio-demographic features and 4 social origin features. A further CRISP-DM iteration based on features' selection tuning will decide on motherOccupation, fatherOccupation, entryGradeHotdeck and 30%FilledFeatures inclusion in the final DM models.

Table 10 - Final ABT for DM modelling purposes.

Feature Name	Features' Group	Data Type	Collection time
gender	Socio-demographic	Cat.	Entrance
yearOfBirth	Socio-demographic	Num.	
fatherOccupationConditionType	Social Origin	Cat.	
motherOccupationConditionType	Social Origin	Cat.	
occupation	Socio-demographic	Cat.	
iscteFirstExecutionYear	Educational Path	Cat.	
maritalStatusType	Socio-demographic	Cat.	
nationality	Socio-demographic	Cat.	
secondNationality	Socio-demographic	Cat.	
entryYear	Educational Path	Cat.	
fatherLiteraryHabilitationType	Social Origin	Cat.	
motherLiteraryHabilitationType	Social Origin	Cat.	

entryAge	Educational Path	Num.	
entryAgeRange	Educational Path	Cat.	
district	Socio-demographic	Cat.	
lisbonMetropolitanArea	Socio-demographic	Cat.	
fatherOccupation	Socio-demographic	Cat.	
motherOccupation	Socio-demographic	Cat.	
degreeCode	Educational Path	Cat.	
degreeSchool	Educational Path	Cat.	
precedentConclusionYear	Previous Education	Cat.	
secondarySchoolType	Previous Education	Cat.	
ingression	Previous Education	Cat.	
highSchoolDegreeType	Previous Education	Cat.	
entryGradeHotDeck	Previous Education	Num.	
studyGap	Previous Education	Cat.	
studyGapYears	Previous Education	Num.	
firstChoice	Previous Education	Cat.	
firstChoiceUniversity	Previous Education	Cat.	
firstChoiceCourse	Previous Education	Cat.	
orderPreference	Previous Education	Num.	
gapEntryExames	Previous Education	Num.	
workingStudentAtEntry	Special Statute	Cat.	
specialEducationNeedsAtEntry	Special Statute	Cat.	
scholarshipAtEntry	Special Statute	Cat.	
dislocatedAtEntry	Special Statute	Cat.	
workingStudentEntryYear1stSem	Special Statute	Cat.	At the end of first curricular semester
InternationalStudentEntryYear1stSem	Special Statute	Cat.	
classSubRepresentativeEntryYear1stSem	Special Statute	Cat.	
classRepresentativeEntryYear1stSem	Special Statute	Cat.	
handicappedEntryYear1stSem	Special Statute	Cat.	
pregnantOrChildrenUnder3EntryYear1stSem	Special Statute	Cat.	
professionalAthleteEntryYear1stSem	Special Statute	Cat.	
sasGrantOwnerEntryYear1stSem	Special Statute	Cat.	
temporaryDisabilityEntryYear1stSem	Special Statute	Cat.	
associativeLeaderEntryYear1stSem	Special Statute	Cat.	
iscteAthleteEntryYear1stSem	Special Statute	Cat.	
ectsCreditsEntryYear1stSem	Educational Path	Num.	
averageEntryYear1stSem	Educational Path	Num.	
weightedAverageEntryYear1stSem	Educational Path	Num.	
workingStudentEntryYear2ndSem	Special Statute	Cat.	At the end of second curricular semester (first curricular year)
InternationalStudentEntryYear2ndSem	Special Statute	Cat.	
classSubRepresentativeEntryYear2ndSem	Special Statute	Cat.	
classRepresentativeEntryYear2ndSem	Special Statute	Cat.	
handicappedEntryYear2ndSem	Special Statute	Cat.	
pregnantOrChildrenUnder3EntryYear2ndSem	Special Statute	Cat.	
professionalAthleteEntryYear2ndSem	Special Statute	Cat.	
sasGrantOwnerEntryYear2ndSem	Special Statute	Cat.	

temporaryDisabilityEntryYear2ndSem	Special Statute	Cat.
associativeLeaderEntryYear2ndSem	Special Statute	Cat.
iscteAthleteEntryYear2ndSem	Special Statute	Cat.
requestedSocialServiceEntryYear	Special Statute	Cat.
acceptedSocialServiceEntryYear	Special Statute	Cat.
requestedSStransportSupplementEntryYear	Special Statute	Cat.
requestedSSaccommodationSupplementEntryYear	Special Statute	Cat.
requestedSSresidenceRequestEntryYear	Special Statute	Cat.
requestedSSiscteFinantialSupportEntryYear	Special Statute	Cat.
acceptedSStransportSupplementEntryYear	Special Statute	Cat.
acceptedSSaccommodationSupplementEntryYear	Special Statute	Cat.
acceptedSSresidenceRequestEntryYear	Special Statute	Cat.
acceptedSSiscteFinantialSupportEntryYear	Special Statute	Cat.
ectsCreditsEntryYear2ndSem	Educational Path	Num.
averageGradeEntryYear2ndSem	Educational Path	Num.
weightedAverageEntryYear2ndSem	Educational Path	Num.

Data type: Num. = Numerical, Cat. = Categorical.

4. Results and discussion

This section details the last three CRISP-DM stages, modelling, evaluation and deployment. Modelling techniques to be used are defined on the modelling stage as well as the model's training plan. Evaluation stage assesses DM model's performance and DM goal's predicting value. Finally, deployment stage details the extracted knowledge and presents guidelines on how to take advantage and apply it real-world context.

4.1. Modelling

There is a wide variety of modelling techniques for classification predictive modelling. The structure and nature of the dataset and DM goal are the main characteristics for modelling technique selection process. Considering it and the techniques that produced best results in the related works, supported by Shahiri et al. (2015) analysis, it was decided to implement the following four techniques: Decision Trees (DT) (Apté and Weiss, 1997), Random Forests (RF) (Breiman, 2001), Support Vector Machines (SVM) (Cortes and Vapnik, 1995) and Artificial Neural Networks (ANN) (Haykin, 1994). (See Table D-1 in Appendix D for DM techniques applied on related works). DT provides a tree-like representation with conditions associated to nodes that permit to classify a new instance in a predefined set of features, while random forests make use of multiple decision trees, merging them to get a stable prediction. Both resultant models represent classification rules on its path from root to leaf. SVM produces a hyperplane in a dimensional space represented by model's features. The resultant model distinctly classifies the data points in the dimensional space. ANN provide multilayer neural nets, consisting of input layer, multiple hidden layers, and an output layer. Each layer is composed by multiple nodes and every node in one layer is connected to every other node in the next layer. The resultant model represents the weights of all output layer's node.

Rminer provides *mining* function to implement DM models. This function requires an input algorithm for modelling, so the following six algorithms were selected to implement the described techniques: RPART, DT and CTREE (distinct DT algorithms), RF, SVM, and MLPE (multilayer perceptron), as ANN representative. Models' training plan is based on k-fold cross-validation method (Trevor et al., 2009). Through it, the training sample is partitioned into k-folds, holding the same number of instances, and each fold is left out of the learning process

and used as a testing set (Kim, 2009). As per Refaeilzadeh et al. (2009), cross-validation can be applied to estimate model's performance, model selection, and model's parameter tuning. The k parameter was set to 10 (k=10), as per the most recent related works' guidelines. Each DM model analysis is submitted to 20 runs in order to enhance results' robustness.

4.2. Evaluation

Evaluation and performance measurement are essential tasks for DM. Metrics such as the Receiver Operating Characteristic (ROC) curve and the corresponding Area Under the ROC Curve (AUC) (Bradley, 1997), based on the confusion matrix (Kohavi, 1998), are widely applied on classification models for measuring purposes. ROC is a probability curve that represents the relationship between true positive rate (TPR) and false positive rate (FPR), at each threshold value. TPR is also referred as sensitivity and FPR as the specificity counterpart, i.e., 1- specificity. Threshold is a cut-off probability value above which the target class is considered true. Each threshold originates a confusion matrix, which show the predictive performance of the model. The predictive performance on a specific threshold is calculated through crossing its predicted value with the real value. AUC is represented by a value between 0 and 1 and tells how much model is capable of distinguishing between classes. AUC=0.5 is the equivalent to perfectly random predictive model. Higher AUC values meant greater predicting performance. Rminer provides *mmetric* function for model's evaluation purposes. This function offers an array of options to be set, as *ROC*, *AUC* and *CONF* (for confusion matrices).

This stage is composed by three main tasks. The first task applies an extended fine feature selection approach. Thus, following CRISP-DM's cyclical approach, features' selection is tuned through models' evaluation, resulting in the ultimate features' set. The second task scrutinizes final models' performance. Three models are evaluated in this task, one for each data collection time, entrance, end of the first curricular semester and end of the second curricular semester. Each model relies on a distinct number of features depending on collection time it is based on. The last task is based on 4-year degrees' model.

4.2.1. Features' selection tuning

This tuning task aims to validate specific features' predictive value and decide upon their presence in final DM models. It starts with entryGradeHotdeck, motherOcuppation and fatherOccupation features. These features were previously submitted to extended data quality processes, though which, their original poor-quality values were filled or replaced. Four distinct test models are used for this evaluation test. Base test model relies on features collected at entrance, as shown in Table 10, except the three features being tested. The 2nd test model adds motherOccupation and FatherOccupation to base test model's features, 3rd test model adds EntryGradeHotdeck and 4th test model adds the whole three features being tested. Note that 30%FilledFeatures will be evaluated further, so they are discarded from this test.

Table 11 shows the predictive results for each model using the designated five modelling techniques, supporting feature selection. The best results are achieved by 4th test model through SVM technique (AUC = 0.7732). Comparing to the remaining SVM models, it is possible to state that, the three tested features, enhanced model's predictive performance. Second best results are for RF technique that demonstrates a similar performance increase. MLPE shows a mixed predictive impact, on the one hand, entryGradeHotdeck feature boosted performance results, on the other, motherOccupation and fatherOccupation features reduced it. Decision tree's tests show a divergent trend, achieving best results for base test models. This fact can be explained by the distinct approaches, each modelling techniques are based on. For instance, substantial number of classes for entryGradeHotdeck and high percentage of 'Unknown' classified records for motherOccupation and fatherOccupation, may impact decision trees performance negatively. As decision trees create a new branch for each feature's class, it may result in the predicting potential to be gradually diluted with the addition of features with these characteristics. Since it is possible to observe a considerable performance gain introduced by tested features on best performance models, it was decided to keep them (summing up 30 features).

Table 11 - AUC results for all preliminary test models.

RPART	DT	CTREE	SVM	RF	MLPE	Model Details
0.6772	0.6784	0.7295	0.7663	0.7491	0.7416	Base test model 27 features 9652 records
0.6768	0.6783	0.7281	0.7665	0.7514	0.7406	2nd test model 29 features (incl. parents' occupation) 9652 records

0.6769	0.6759	0.7281	0.7726	0.7585	0.7484	3rd test model 28 features (incl. entryGradeHotdeck) 9652 records
0.6764	0.6772	0.7273	0.7732	0.7611	0.7476	4th test model 30 features (incl.3 tested features) 9652 records

AUC mean values after 20 runs of 10-fold for each modelling technique

Complementary tuning task is focused on 30%FilledFeatures' predictive value. As discussed before, dataset is reduced from 9652 to 2713 records in order to accommodate these 6 features in the model. Table 12 summarizes the predictive performance results obtained by DM_30%FilledFeatures model. It combines the 6 features being tested to previously 30 selected features. The overall figures show a clear reduction in predictive capabilities, as the predictive potential introduced by these features does not pay off the dataset reduction in 70%. (See Figure E-1 and Table E-2 in Appendix E for ROC curve and confusion matrices) The results also demonstrate that the information gain introduced by these features is not enough for algorithms to leverage model's performance. In the light of this test results, it was decided to discard 30%FilledFeatures from further modelling.

Table 12 - AUC results for DM_30%FilledFeatures model.

RPART	DT	CTREE	SVM	RF	MLPE	Model Details
0.6630	0.6624	0.6991	0.7495	0.7375	0.7136	30%FilledFeatures model 36 features 2713 records

AUC mean values after 20 runs of 10-fold for each modelling technique

As a result of this tuning iteration it was decided to maintain entryGradeHotdeck, motherOcuppation and fatherOccupation, and remove 30%FilledFeatures from the final DM models. Previous education features' group representatives are reduced from 12 to 6, as all removed features belong to this group. Thus, the total number of features, considering all collection times was reduced to 68, which 30 are collected at entrance, additional 14 collected at the end of first curricular semester and finally, additional 14 collected at the end of the second curricular semester.

4.2.2. Final models' evaluation

Following the previous subsection considerations, all conditions are set to proceed with final DM models' evaluation. The first model being evaluated are composed by 30 features collected at entrance. This model is henceforth referred as DM_Entrance. Table 13 depicts SVM, as the best predictive model. It provides a significant AUC result, higher than 0.77. RF model also demonstrates a considerable predictive result surpassing 0.76, while MLPE model almost reaches 0.75. CTREE achieves the best result by far within the decision tree model, even performing considerably worse than the previous models.

Table 13 - AUC results for DM_Entrance model.

RPART	DT	CTREE	SVM	RF	MLPE	Model Details
0.6764	0.6772	0.7273	0.7732	0.7611	0.7476	DM_Entrance model 30 features 9652 records

AUC mean values after 20 runs of 10-fold for each modelling technique

Figure 4 shows the ROC curve for CTREE, as DT's representative, SVM, RF and MLPE. It plots FPR versus TPR performance for each technique at each threshold point. This analysis allows each model's discriminatory capacity to be compared. It is possible to observe that SVM curve achieves higher TPR values along the entire FPR-axis. SVM model proves its higher discriminatory capacity, outperforming remaining models for the whole cut-off probability's range. The points highlighted in the graphic represent a threshold value of 50%, for each model's curve. Finest threshold selection was determined by TPR versus FPR analysis through confusion matrices. Confusion matrices were designed for 9 distinct threshold values, from 0.1 to 0.9.

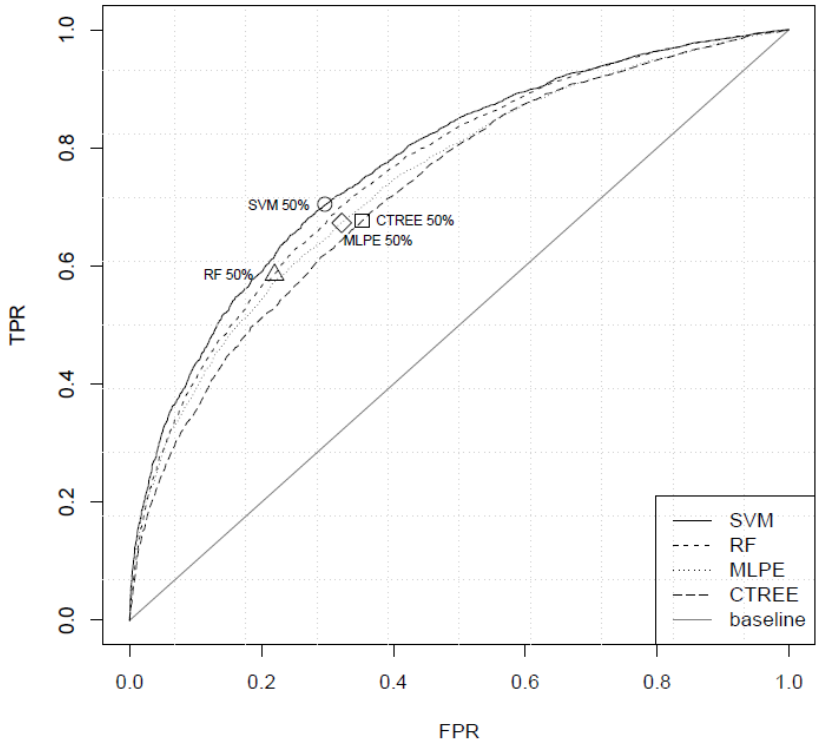


Figure 4 - ROC curves for DM_Entrance model.

Table 14 details 50% threshold analysis through confusion matrices and resulting sensitivity and 1-specificity values for DM_Entrance models. Good results correspond to high figures down the main diagonal, representing correct predictions, and low figures down the off diagonal, for incorrect predictions. This analysis is focused on "Failure" class predicting performance. So, starting with SVM, out of 4931 unsuccessful students, 3469 are classified correctly and 1462 are classified incorrectly. So, its sensitivity is found to be approximately 0.7, corresponding to a good TPR, especially considering the early stage predicting potential. SVM's sensitivity result is remarkably the highest, while 1-specificity result is approximately 0.3. Although, RF model's 1-specificify result is considerably lower, conferring a lower FPR, its sensitivity is not relevant for this threshold value (below 0.6).

Table 14 - Confusion matrices for DM_Entrance model.

Threshold = 50%					
SVM		Predicted		Sensitivity	1-specificity
		Failure	Success		
Target	Failure	3469	1462	0.7035	0.2959
	Success	1397	3324		
RF		Predicted		Sensitivity	1-specificity
		Failure	Success	0.5851	0.2197

Target	Failure	2885	2046		
	Success	1037	3684		
MLPE		Predicted		Sensitivity	1-specificity
		Failure	Success		
Target	Failure	3318	1613	0.6729	0.3215
	Success	1518	3203		
CTREE		Predicted		Sensitivity	1-specificity
		Failure	Success		
Target	Failure	3336	1595	0.6765	0.3527
	Success	1665	3056		

DM_EntryYear1Sem model establishes the basis for succeeding collection time model. It is composed by features collected at entrance and at the end of the first curricular semester, summing up to 44 features. Table 15 demonstrates a huge predictive performance boost compared to DM_Entrance model's results. In fact, all models registered great performance improvements, greater or equal than 13%. It can be observed that newly included features, resulted in a significant reduction of performance gap between models. Once more SVM and RF achieve the best AUC results, surpassing 0.90. On its turn, decision tree models show the highest predictive performance boost, increasing approximately 17%. It allows CTREE to overpass MPLE performance results.

Table 15 - AUC results for DM_EntryYear1Sem model

RPART	DT	CTREE	SVM	RF	MLPE	Model Details
0.8463	0.8466	0.8954	0.9097	0.9082	0.8936	DM_EntryYear1Sem model 44 features 9652 records

AUC mean values after 20 runs of 10-fold for each modelling technique

Figure 5 shows the ROC curves for DM_EntryYear1Sem models' performance analysis. RF curve clearly intersects SVM curve for an FPR close to 0.5. SVM slightly achieves better performance for lower values of FPR, while RF is slightly better above that value. On its turn, CTREE curve intersects MLPE curve for and FPR close to 0.2. MLPE achieves better performance than CTREE for lower values of FRP, being outperformed above that value. Threshold values of 50% and 30%, for each model's curve are highlighted in the figure. 30% threshold represents an optimized TPR/FPR trade-off, i.e., improved TPR values, for acceptable FPR values.

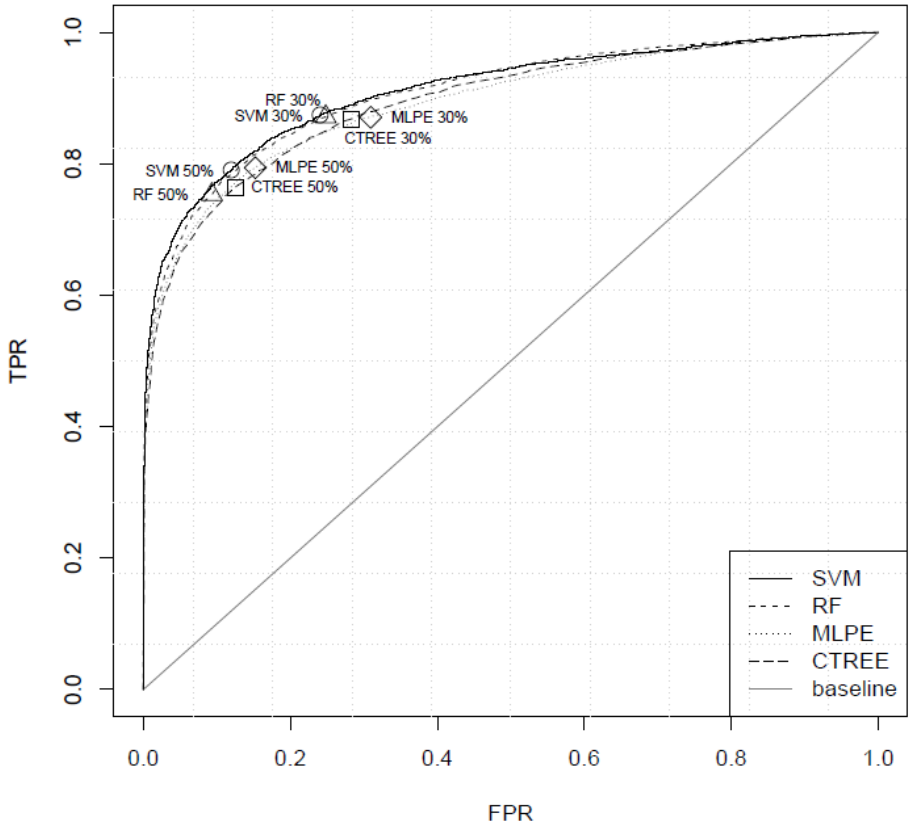


Figure 5 - ROC curves for DM_EntryYear1Sem model.

Table 16 details 50% and 30% threshold analysis through confusion matrices and resulting sensitivity and 1-specificity values for DM_EntryYear1Sem model.

Regarding 50% threshold, it is possible to observe that all models' sensitivity results increased significantly comparing to DM_Entrance results. MLPE shows the best sensitivity surpassing 0.79, while SVM almost reaches that value. For a similar sensitivity results, it is perceptible that SVM's 1-specitivity result is significantly lower, being a greater result. No less important, is the fact that overall 1-specificity results decreased significantly. It allows to perform additional analysis based on a lower threshold. Reducing the threshold to 30 %, it is noticeable that 1-sensitivity results are still above the ones verified for DM_Entrance applying 50% threshold. As for the sensitivity, there are clear increases, as all models show results around 0.87. SVM and RF achieve the best performance combination, as for such a great sensitivity result both show 1-specitivity around 0.24. SVM results are still slightly better. It is also interesting to verify that MLPE's 1-specitivity soared for this reduced threshold.

Table 16 - Confusion matrices for DM_EntryYear1Sem model.

Threshold = 50%			
SVM	Predicted	Sensitivity	1-specificity

		Failure	Success		
Target	Failure	3894	1037	0.7897	0.1193
	Success	563	4158		
RF		Predicted		Sensitivity	1-specificity
		Failure	Success		
Target	Failure	3714	1217	0.7532	0.0932
	Success	440	4281		
MLPE		Predicted		Sensitivity	1-specificity
		Failure	Success		
Target	Failure	3914	1017	0.7938	0.1523
	Success	719	4002		
CTREE		Predicted		Sensitivity	1-specificity
		Failure	Success		
Target	Failure	3768	1163	0.7641	0.1254
	Success	592	4129		
Threshold = 30%					
SVM		Predicted		Sensitivity	1-specificity
		Failure	Success		
Target	Failure	4303	628	0.8726	0.2402
	Success	1134	3587		
RF		Predicted		Sensitivity	1-specificity
		Failure	Success		
Target	Failure	4297	634	0.8714	0.2480
	Success	1171	3550		
MLPE		Predicted		Sensitivity	1-specificity
		Failure	Success		
Target	Failure	4293	638	0.8706	0.3093
	Success	1460	3261		
CTREE		Predicted		Sensitivity	1-specificity
		Failure	Success		
Target	Failure	4279	652	0.8678	0.2824
	Success	1333	3388		

DM_EntryYear2Sem model relies on the whole set of features collected by the end of the second curricular semester, i.e., the end of the first curricular year, being composed by 68 features. Table 17 shows DM_EntryYear2Sem model's increased predictive performance results. Newly included features allowed SVM and RF models to reach, approximately, 0.94. In contrast to DM_EntryYear1Sem model's evaluation results, RF obtains a slightly better overall performance then SVM and MLPE slightly overperforms CTREE. These results

indicate that the discriminatory capacity of the whole features' set, at this point, is so robust that distinct models' performance results tend to converge. Overall performance improvements, around 3%, aren't so prominent then on last evaluation loop, taking in consideration that the same number of additional analogous features were included. This fact can be explained by the great models' performance results achieved at this stage, quite close to 1.

Table 17 - AUC results for DM_EntryYear2Sem model.

RPART	DT	CTREE	SVM	RF	MLPE	Model Details
0.8882	0.8886	0.9257	0.9378	0.9380	0.9263	DM_EntryYear2Sem model 68 features 9652 records

AUC mean values after 20 runs of 10-fold for each modelling technique

Figure 6 shows ROC analysis for DM_EntryYear2Sem models. SVM achieves the best performance for an FPR below 0.3. RF intersects SVM around that value, outperforming it for above values. CTREE and MLPE show a close performance for all FPR axis. Their curves intersect each other for several times. 20% threshold value was scrutinized, following same threshold selection reasoning applied previously.

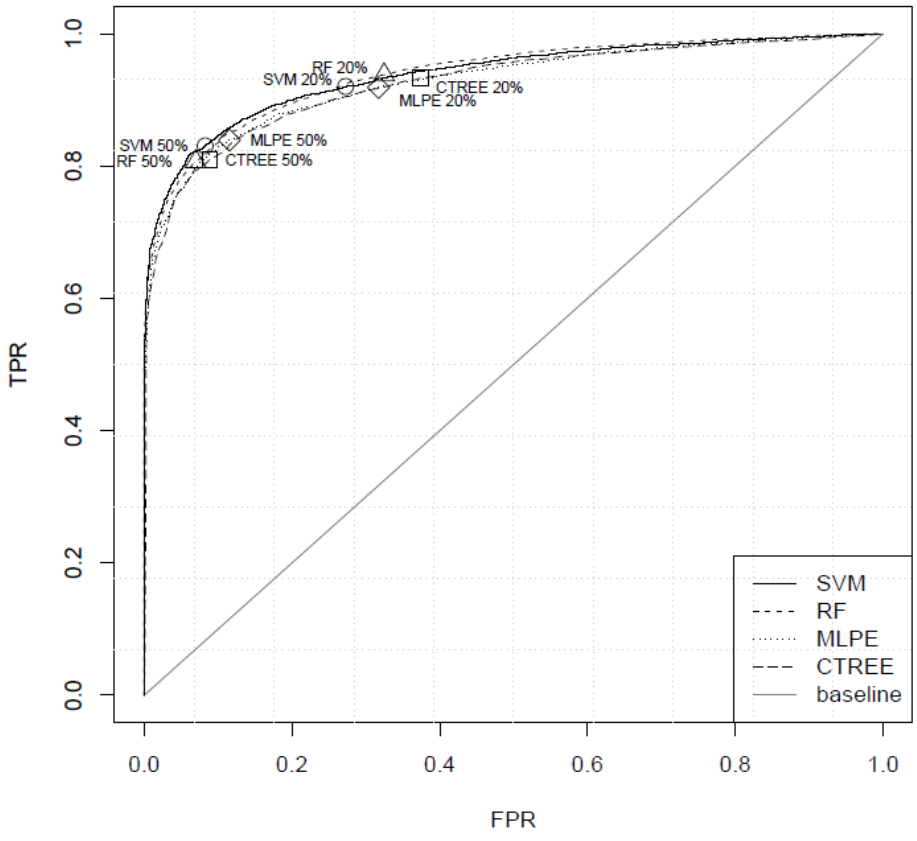


Figure 6 - ROC curves for DM_EntryYear2Sem model.

Table 18 details 50% and 20% threshold analysis through confusion matrices and resulting sensitivity and 1-specificity values for DM_EntryYear2Sem model.

Regarding 50% threshold, it is possible to observe that all models' sensitivity results increased comparing to DM_EntryYear1Sem results. Following DM_EntryYear1sem confusion matrices trend, MLPE shows the best sensitivity almost reaching 0.84, while SVM slightly surpasses 0.83. Although, SVM's 1-specitivity result is significantly lower, resulting in greater TPR/FPR ratio. Overall 1-specificity results decreased to values below 0.1. It allows to perform additional analysis based on an even lower threshold. Reducing the threshold to 20 %, 1-sensitivity results are around 0.3 for the most part of the models while CTREE increases to 0.37. Sensitivity results increase, as RF and CTREE surpasses 0.93 and, SVM and MLPE almost reach 0.92.

At this point the models' sensitivity is so high and close that special attention is given to 1-specificity review. So, CTREE even achieving the second-best sensitivity, it is under the spotlight due to its poor 1-specificity results. MLPE loses some track to SVM, as for a similar sensitivity results, MLPE performs worst in terms of 1-specificity. Comparing RF and SVM, RF achieves a slightly better sensitivity while SVM achieves a reduced and considerably better 1-specificity results.

Table 18 - Confusion matrices for DM_EntryYear2Sem model.

Threshold = 50%					
SVM		Predicted		Sensitivity	1-specificity
		Failure	Success		
Target	Failure	4094	837	0.8303	0.0826
	Success	390	4331		
RF		Predicted		Sensitivity	1-specificity
		Failure	Success		
Target	Failure	3977	954	0.8065	0.0699
	Success	330	4391		
MLPE		Predicted		Sensitivity	1-specificity
		Failure	Success		
Target	Failure	4137	794	0.8390	0.1154
	Success	545	4176		
CTREE		Predicted		Sensitivity	1-specificity
		Failure	Success		
Target	Failure	3993	938	0.8098	0.0888
	Success	419	4302		

Threshold = 20%					
SVM		Predicted		Sensitivity	1-specificity
		Failure	Success		
Target	Failure	4536	395	0.9199	0.2724
	Success	1286	3435		
RF		Predicted		Sensitivity	1-specificity
		Failure	Success		
Target	Failure	4624	307	0.9377	0.3247
	Success	1533	3188		
MLPE		Predicted		Sensitivity	1-specificity
		Failure	Success		
Target	Failure	4293	638	0.9187	0.3173
	Success	1460	3261		
CTREE		Predicted		Sensitivity	1-specificity
		Failure	Success		
Target	Failure	4600	331	0.9329	0.3739
	Success	1765	2956		

Figure 7 shows a wrapped-up analysis for the reviewed models performance, considering each DM model per features' collection time.

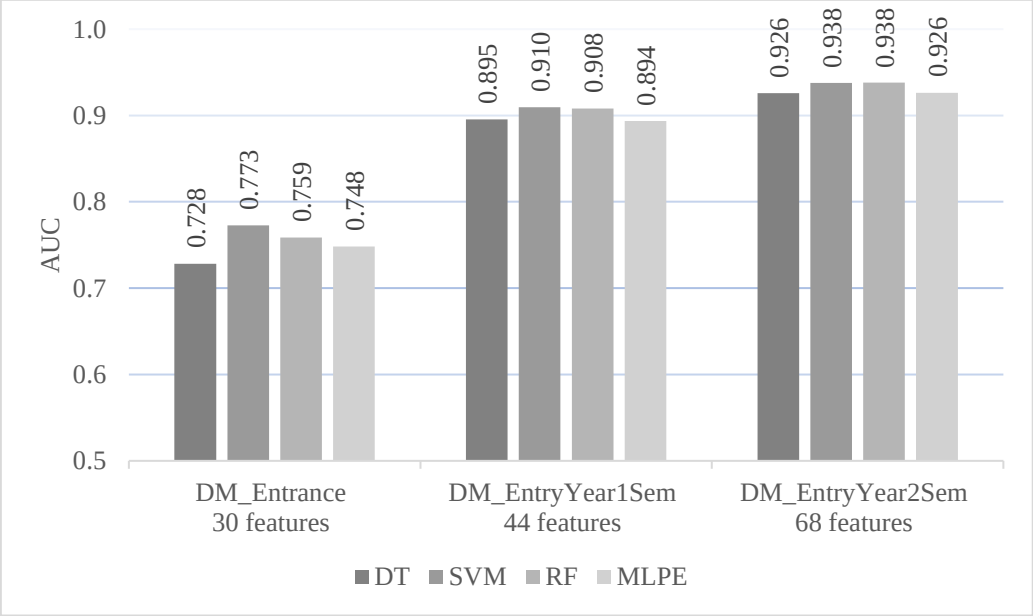


Figure 7 - Shows a wrapped-up analysis for reviewed models performance.

Concluding this evaluation analysis, it is important to highlight the following findings: SVM is clearly the best model for DM_Entrance, as it outperforms other models for all threshold range.

DM_Entrance model can be developed to predict student's performance before the beginning of the first curricular semester. This a-priori predictive model shows good evaluation results (AUC =0.77 for SVM). DM_EntryYear1Sem model is able to predict student's performance by the end of the first curricular semester. It provides an enhanced predictive potential in early stages of the academic path, achieving improved evaluation results (AUC around 0.91 for SVM). On its turn DM_EntryYear2Sem model can be set up by the end of the first curricular year. As expected, the predictive potential is even improved achieving near perfect performance (AUC around 0.94 for SVM and RF models). Following these findings, SVM models are selected to be submitted to feature's relevance analysis for knowledge extraction.

These results demonstrate that it is reasonable to predict academic failure at early stages, as the models show good to great performance levels. Even exclusively relying on information collected through admission process, it is possible to achieve good predictive results. As the first curricular year proceeds it is possible to enrich the model with new information, enhancing its predictive potential to near perfect results. Thus, it is possible setup a framework to act in three different times, supporting success policies and permitting it to be re-adjusted along the course.

4.2.3. Extended evaluation for 4-years Bologna bachelor's degrees

Based on features collected at entrance, DM_Entrance_IGE model encompasses 4-year bachelor's data, being composed by only 789 records. It is represented by a much smaller dataset, as there are just two degree represented: IGE and IGE-PL. Both degrees belong to the same school (ISTA). There is just a single class to be found for degreeSchool feature, as it represents school's information. Thus, degreeSchool feature is removed from DM_Entrance_IGE model, which is then composed by 29 features, minus one than DM_Entrance model. Table 19 depicts SVM technique as the best predictive performance (AUC = 0.7434) and RF as the second. Same trend verified on DM_Entrance model's AUC results, is verified here.

Table 19 - AUC results for DM_Entrance_IGE model and all modelling techniques.

RPART	DT	CTREE	SVM	RF	MLPE	Model Details
0.6912	0.6972	0.7270	0.7434	0.7433	0.7236	DM_Entrance_IGE model 29 features 789 records

AUC mean values after 20 runs of 10-fold for each modelling technique

Figure 8 shows the ROC curves for DM_Entrance_IGE models' performance analysis. It is noticeable that curves are much sharper compared to DM_Entrance models' representation. This behaviour relates to a much smaller dataset being represented in this model. SVM curve shows poor performance for FPR values below 0.3, increasing its performance and clearly surpassing other models above that value. On its turn CTREE demonstrates the best performance for FPR values below 0.3, losing its momentum as FPR increases. In a lower extent, MLPE's curve mimics SVM trend, as well as RF's curve mimics CTREE trend. Highlighted threshold values of 50% and 80% show a manifest difference in terms of performance between each model.

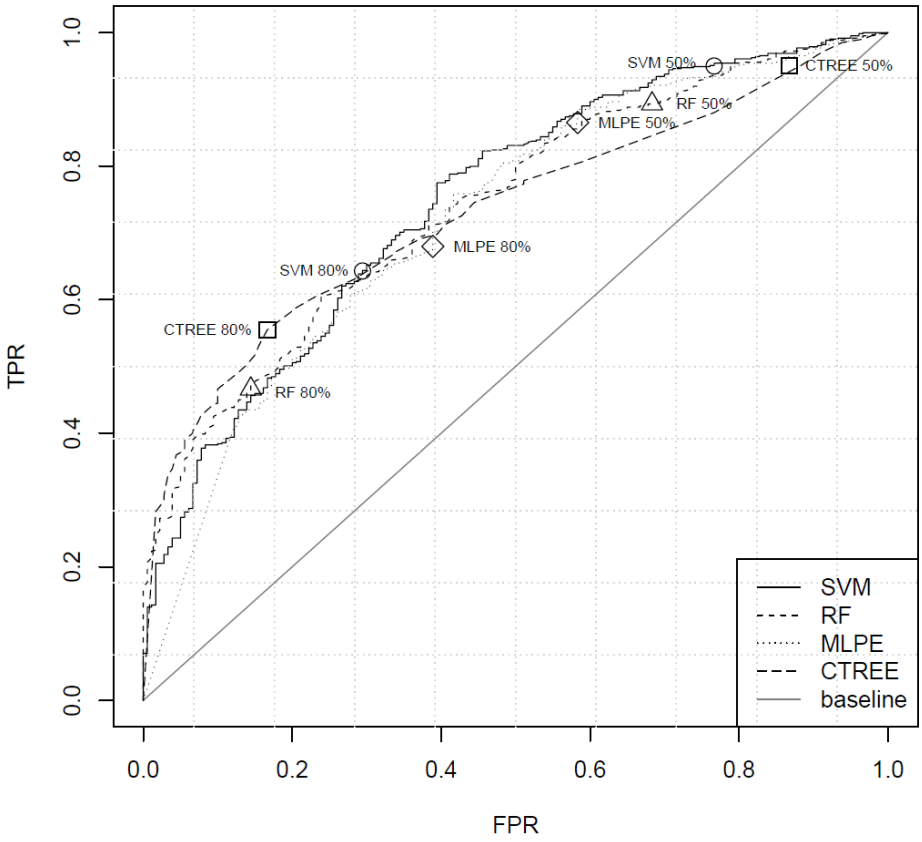


Figure 8 - ROC curves for DM_Entrance_IGE model.

80% threshold value is considered, once it is the approximate value, which SVM achieves a considerably low FPR value (0.3), registering a significant TPR (0.65). Confusion Matrices detailed in Table 20, show that distinct models are not efficient for 50% threshold values. Even extremely accurate predicting failures, a huge percentage of success cases are incorrectly predicted as failure. This can be confirmed through high 1-specificity figures. Nevertheless, 80% threshold analysis show some interesting results especially for SVM and MLPE models.

Success conditions for this model are significantly different from DM_Entrance model. So, it is relevant to ISCTE-IUL to understand whether the same features affect both models the same way, i.e., demonstrate similar predictive relevance. This additional evaluation's results show a reasonable predictive performance, in particular, for such a small dataset. These findings encourage an extended features' relevance analysis for DM_Entrance_IGE, allowing a subsequent comparative review.

Table 20 - Confusion matrices for DM_Entrance_IGE model.

Threshold = 50%					
SVM		Predicted		Sensitivity	1-specificity
		Failure	Success		
Target	Failure	579	30	0.9507	0.7667
	Success	138	42		
RF		Predicted		Sensitivity	1-specificity
		Failure	Success		
Target	Failure	544	65	0.8933	0.6833
	Success	123	57		
MLPE		Predicted		Sensitivity	1-specificity
		Failure	Success		
Target	Failure	527	82	0.8654	0.5833
	Success	105	75		
CTREE		Predicted		Sensitivity	1-specificity
		Failure	Success		
Target	Failure	579	30	0.9507	0.8679
	Success	138	21		
Threshold = 80%					
SVM		Predicted		Sensitivity	1-specificity
		Failure	Success		
Target	Failure	392	217	0.6437	0.2944
	Success	53	127		
RF		Predicted		Sensitivity	1-specificity
		Failure	Success		
Target	Failure	284	325	0.4663	0.1444
	Success	26	154		
MLPE		Predicted		Sensitivity	1-specificity
		Failure	Success		
Target	Failure	414	195	0.6798	0.3889
	Success	70	110		
CTREE		Predicted		Sensitivity	1-specificity
		Failure	Success	0.5550	0.1667

Target	Failure	338	271
	Success	30	150

4.3. Knowledge extraction and guidelines for implementation

Sensitivity analysis (SA) method described by Cortez and Embrecht (2011) was adopted to perform feature's relevance analysis. SA allows to assess the importance of the input features to a given model (Saltelli et al., 2000). SA' characteristics potentiate meaningful knowledge to be extracted from DM models. As discussed previously, final SVM models were selected to be submitted to feature's relevance analysis, due to their finest evaluation results. Rminer provides *importance* function to implement SA. Data-based sensitivity analysis algorithm (DSA) is selected among others, as it induces several features values to be changed simultaneously, allowing interactions between input features to be detected. Additionally, it uses values taken from the dataset used for training in order to avoid testing all possible combinations, increasing computational efficiency.

4.3.1. DM_Entrance model's DSA

Figure 9 shows the relevance for the most impacting features in DM_Entrance model (see Table F-1 in Appendix F for complete DSA). Each bar depicts the relevance of a single feature in the model. Feature's relevance is measured through its contribution percentage to the output. Each of the illustrated features, 8 out of 30, demonstrates great relevance, above 5%. Their combined contribution to the model surpasses 63%. It is also interesting to note that approximately half of the features' set, show a relevance above 2%, summing up 80%. So, the remaining half, showed a much-reduced predictive influence, summing up 20%.

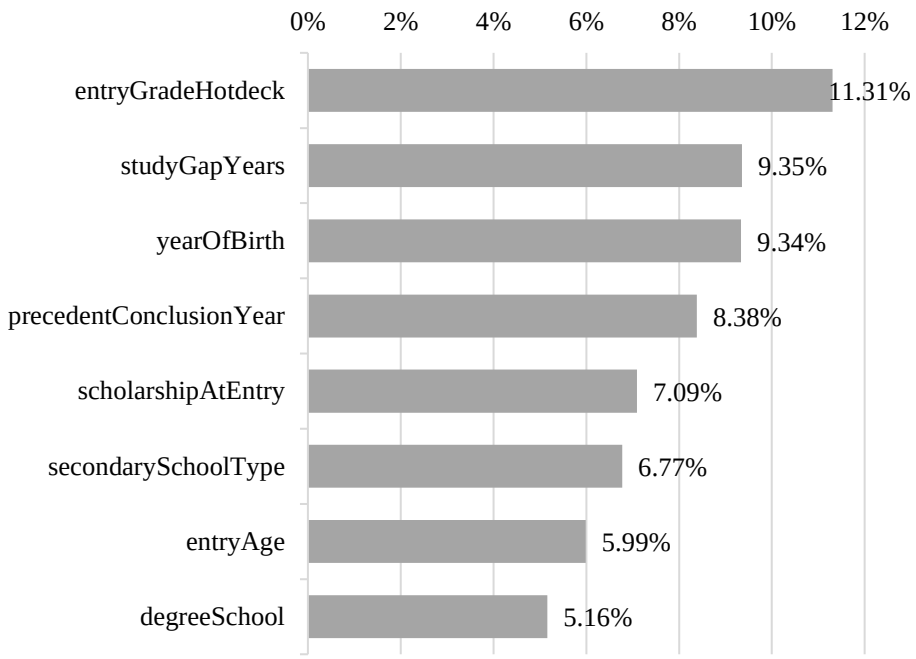


Figure 9 - Features' relevance for DM_Entrance model.

Reviewing high impact features on its characteristics, it is noticeable that all features' groups are represented, except social origin. Most represented features' group is previous education, placing entryGradeHotDeck as the most impacting feature, studyGapYears, precedentConclusionYear and secondarySchoolType, in second, fourth and sixth, respectively. Although, half of these features belong to previous education features' group, the most remarkable aspect is, the diversity in terms of characteristics and nature.

Even submitted to an imputation process, during data preparation stage, entryGradeHotdeck feature keep its prominent importance and shows the highest relevance, around 11.3%. The detailed influence of entryGradeHotDeck feature is depicted in Figure 10.

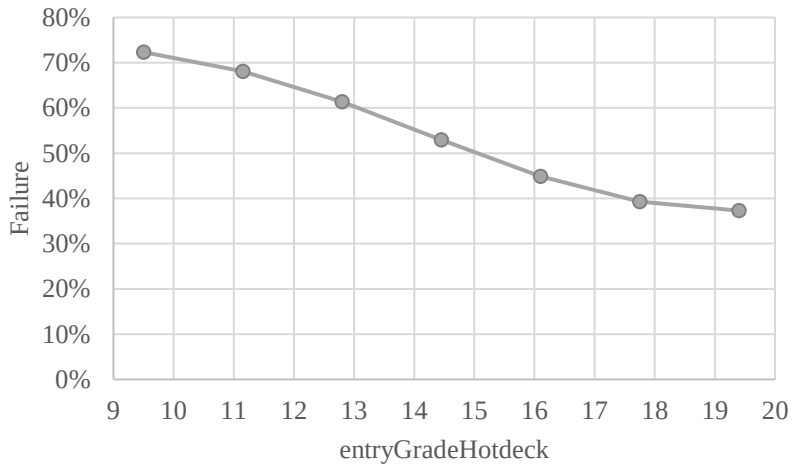


Figure 10 - Impact of entryGradeHotdeck on DM_Entrance model.

This feature quantifies high school evaluation performance, so it is expected that students that achieved higher previous evaluation performances are more likely to succeed in the higher education. As highlighted in Tinto (1999), high school evaluation performance provides insight into potential academic performance of the freshmen and shows strong positive effect on persistence. Previous education features are commonly pointed out as relevant predictors of academic success. Related studies, such as, Osmanbegović & Suljić (2012), Goker et al. (2013), Trstenjak & Donko (2014) and Asif et al. (2017), present previous evaluation related features as the most impacting features on their models. As per Trstenjak & Donko (2014), great part of socio-demographic and social origin features doesn't change over time, having previously influenced high school evaluation performance. This helps explaining the leveraged relevance of entryGradeHotDeck feature in the model. The initial perception regarding previous student's performance is confirmed, as lower entryGradeHotdeck values, presents a much stronger contribution to failure, especially for entry grade values below 13.

The second most impacting feature is studyGapYears, contributing with around 9.3%. It is quite an interesting finding, since no similar feature has included in related works' models. Figure 11 shows no significant impact for studyGapYears values below 10.

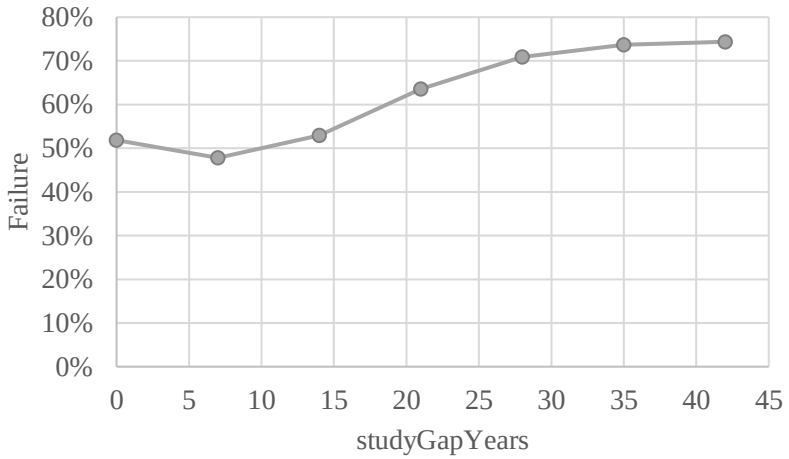


Figure 11 - Impact of studyGapYears on DM_Entrance model.

Even so several years' gap shows slightly inferior impact than gap's absence. For gaps above 10 years, a prominent influence is verified and it is possible to infer that big gaps between the precedent study year and fresh enrolment plays a great role in unsuccessful cases.

The third most impacting feature is yearOfBirth, registering a similar contribution percentage. In order to review its impact illustrated in Figure 12, it is important to remind that original dataset was trimmed to enrolments between 2006/2007 and 2015/2016.

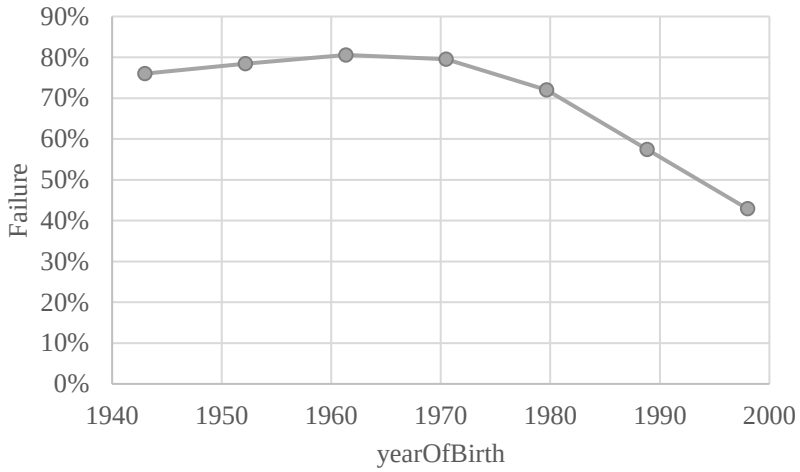


Figure 12- Impact of yearOfBirth on DM_Entrance model.

In general terms, yearOfBirth show considerable to high contribution to failure for values below 1990. This impact trend demonstrates that failure is higher among older students, as most of these cases represent students that enrolled in later life stages. These findings follow indications presented in Natek & Zwilling (2014), Martins et al. (2017) and Fernandes et al. (2018).

The remaining most impacting features presents contributions between around 8.4% and 5.2%. Ranked in fourth place, precedentConclusionYear is related with studyGapYears. Considering each case's entryYear, the older the precedentConclusionYear, the higher studyGapYears. So, as expected, Figure 13 shows that older precedentConclusionYear are more likely to explain failure.

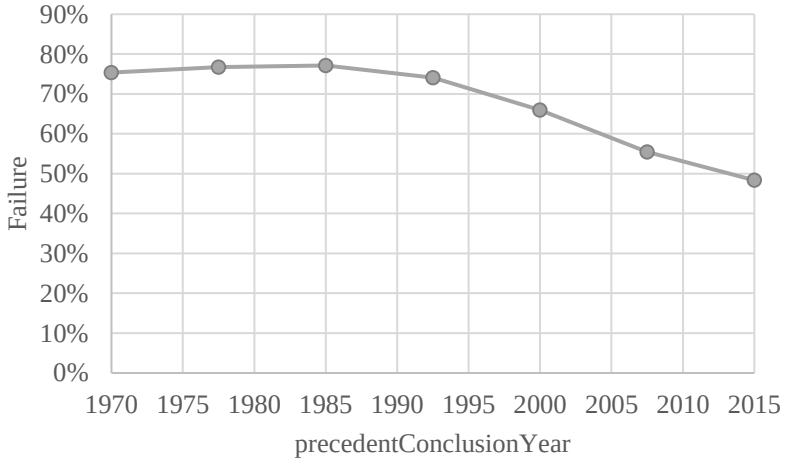


Figure 13 - Impact of precedentConclusionYear on DM_Entrance model.

This feature's weight in the model reaffirms the importance of gaps between the previous educational stage and the current degree's enrolment.

As assessed in Delen (2011), scholarship feature, represented by scholarshipAtEntry demonstrate its predictive potential in the model. According to Figure 14, students, which are granted a scholarship at entry, are more likely to avoid failure. This feature's impact validates Herzog (2005) claims, that scholarships create basis for successful academic paths.

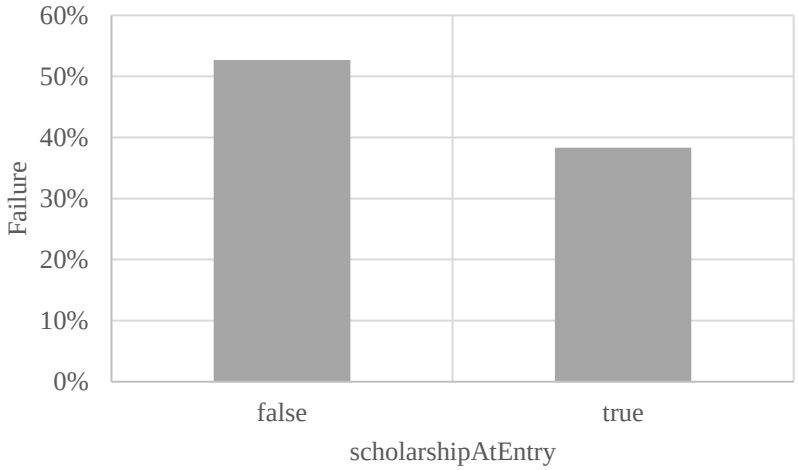


Figure 14 - Impact of scholarshipAtEntry on DM_Entrance model.

The impact of high school's sector on academic failure is depicted in Figure 15. It is observable that private sector high schools demonstrate greater contribution to unsuccessful paths than public or public/private high schools.

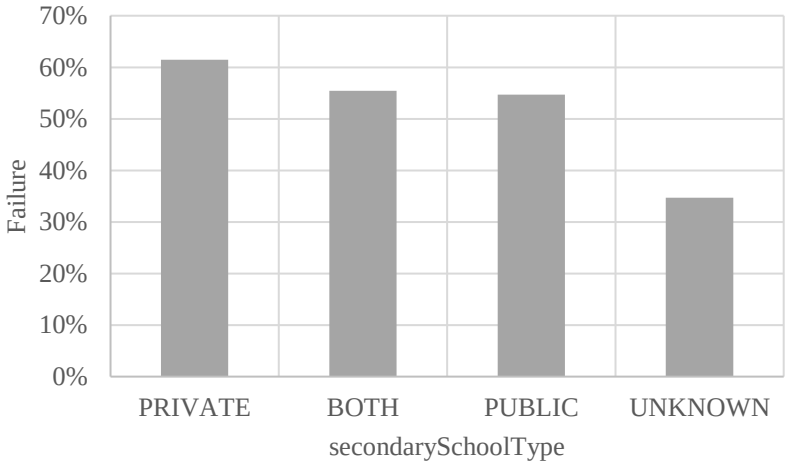


Figure 15 - Impact of secondarySchoolType on DM_Entrance model.

Although high school related features are frequently found in related works, its sector is a little explored aspect. Even so, Martins et al. (2017) included a corresponding feature, not achieving similar expressive impact. It is important to refer that "Unknown" class represents more cases

(1153 cases), than "Private" and "Both" classes together, 761 and 333 cases respectively. This fact may affect this feature's relevance and result in misleading findings.

According to entryAge contribution shown in Figure 16, it is interesting to review it alongside with YearOfBirth (Figure 12). Both analyses are naturally related, so following the same rational, higher figures, especially above 25 years old, present a higher contribution to failure.

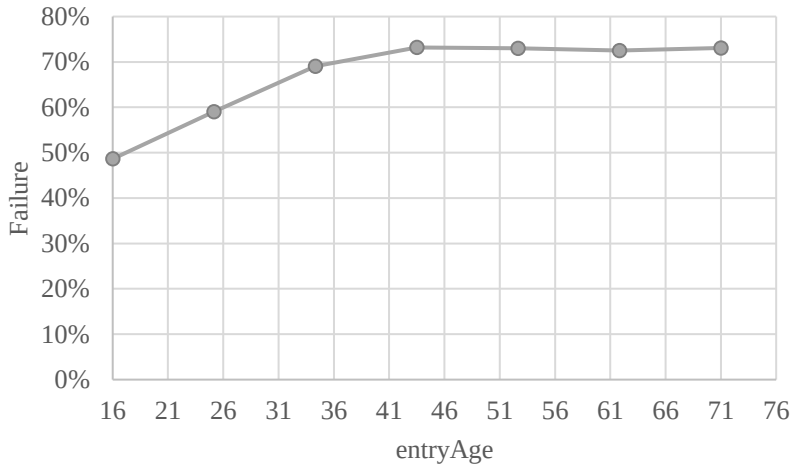


Figure 16 - Impact of entryAge on DM_Entrance model.

The last relevance analysis for DM_Entrance model is focused on degreeSchool (Figure 17). It shows that unsuccessful cases are more likely to occur in ISTA school's degrees than the remaining schools' degrees. The high relevance of this feature is also been identified in Martins et al. (2017) and Martins et al. (2018), both studies relying on Portuguese educational system. ISTA school are composed by architecture and mostly engineering degrees. Higher impact on failure for engineering degrees were expected results, that corroborate Martins et al. (2017) indications. In a reduced scale ECSH+EG joint degrees and ECSH degrees show higher contribution to failure than EG and ESPP.

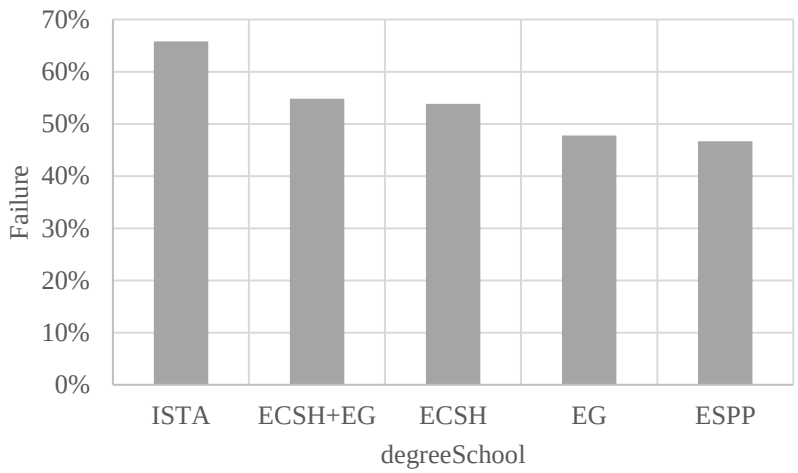


Figure 17 - Impact of degreeSchool on DM_Entrance model.

4.3.2. DM_EntryYear1Sem model's DSA

Following DSA is based on DM_EntryYear1Sem model. Figure 18 shows the relevance of the 8 most impacting features in the model (see Table F-2 on the Appendix F for extended DM_EntryYear1Sem DSA). Several features, collected at the end of first curricular semester, showed great impact, placing 4 features among the 8 most relevant. This impact confirms the directions discussed in model's evaluation, that pointed a performance enhancement, provided by these features. The combined contribution of these 8 most impacting features is close to 65%. There is a clear tendency for residual contributions, as only 12 features out of 44 show a relevance above 2%, summing up approximately 74%, while the other 30 features contribute with the remaining 26%.

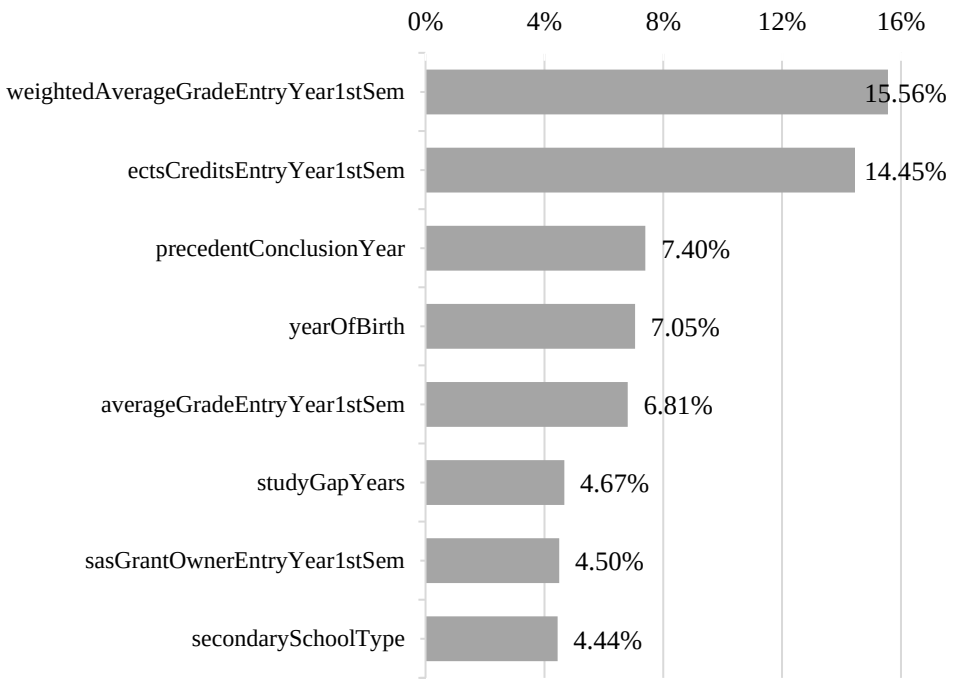


Figure 18 - Features' relevance for DM_EntryYear1Sem model.

The two higher relevance features, weightedAverageGradeEntryYear1stSem and ectsCreditsEntryYear1stSem are educational path features' group representatives. Particularly, they represent first curricular semester evaluation' information, achieving a combined relevance greater than 30%, around 15.5% and 14.5% respectively. Both features approximately double the relevance of the third most relevant feature. This enhanced impact assess how relevant are these two features to explain failures at this point of the curricular path. These feature's relevance supports the findings presented in Martins et al. (2018), that relying on the same educational system, observed similar results for these features. Other studies, such as, Mishra et al. (2014), Slim et al. (2014), Zimmermann et al. (2015) and Asif et al. (2017) demonstrate similar level of impact for equivalent features in their models. Asif et al., (2017) points two groups of academic students according to their performance, high-performing students and low-achieving students and claim that many students tend to stay in the same kind of groups for all academic path. This standpoint may provide some insight regarding these features' great impact in the model. Figures 19 and 20 demonstrate that the lower their values (worst evaluation performance), the stronger their contribution to academic failure.

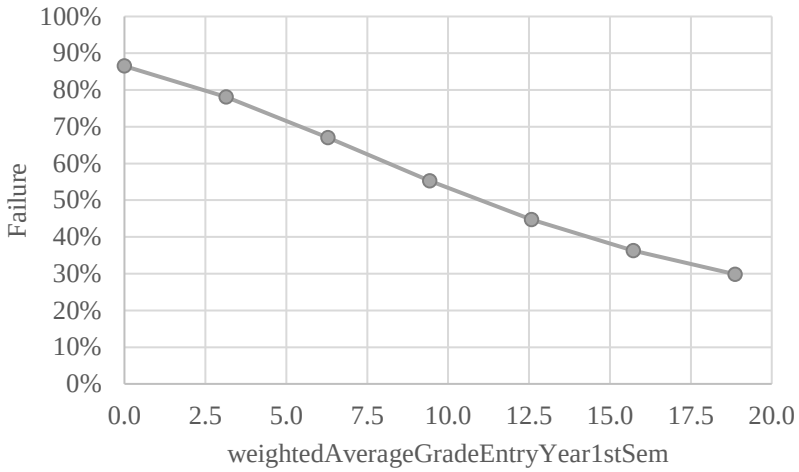


Figure 19 - Impact of weightedAverageGradeEntryYear1stSem on DM_EntryYear1Sem model.

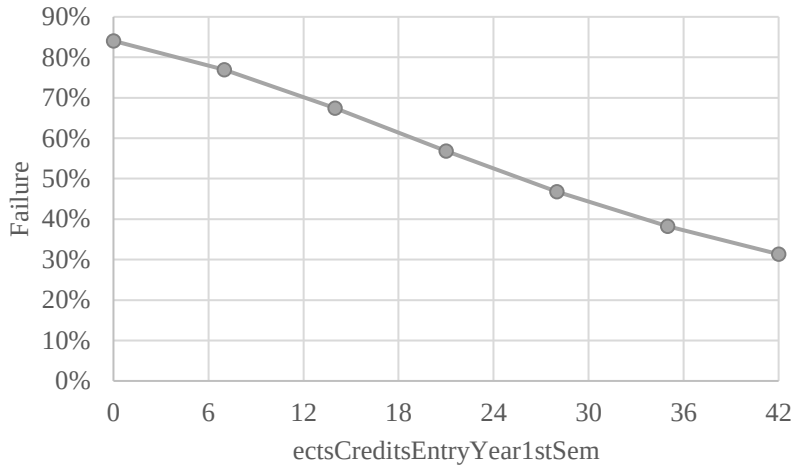


Figure 20 - Impact of ectsCreditsEntryYear1stSem on DM_EntryYear1Sem model.

Focusing on ectsCreditsEntryYear1stSem, although there is no significant impact for values above 18 ECTS, there is a prominent increasing impact for higher values. This increasing impact correspond to a lower number of ECTS earned. A conforming impact, mostly influenced by its weighted calculation, is registered for weightedAverageGradeEntryYear1stSem feature. This explains the negligible impact for values slightly below 10. Substantially worse grades (below 7), show an outstanding higher impact. A third evaluation-related feature named averageGradeEntryYear1stSem is the fifth most relevance feature. Figure 21 reveals its truncated importance when compared to weightedAverageGradeEntryYear1stSem. This behaviour can be explained by the fact that it doesn't reflect the number of passed courses or earned ECTS, just a simple average grade of the passed courses.

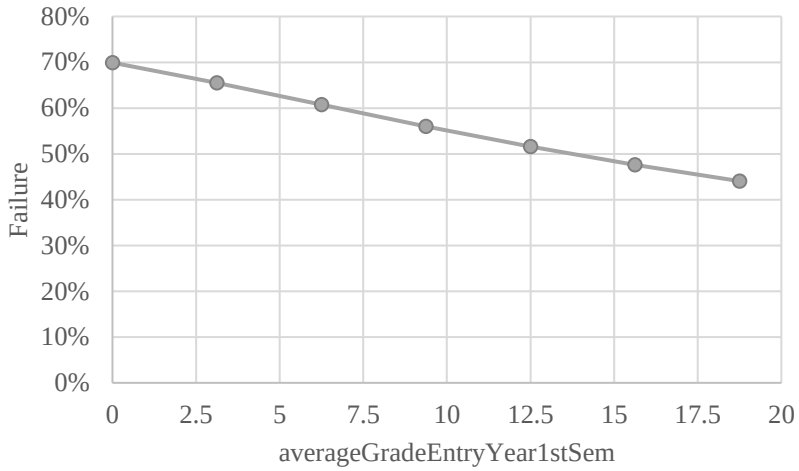


Figure 21 - Impact of averageGradeEntryYear1stSem on DM_ EntryYear1Sem model.

It is interesting to verify the high relevance of sasGrantOwnerEntryYear1stSem feature, as its impact depicted on Figure 22 show that first curricular semester students granted with social support are more likely to fail. This feature's impact can be addressed as a cause/effect event, once its negative impact to success seems to be more related with the required conditions for its acquirement. In other words, lower socio-economical level might show its impact through this feature. These results are aligned with Herzog (2005) insights that point social support/aid as a retention enabler that is not able to potentiate student's success.

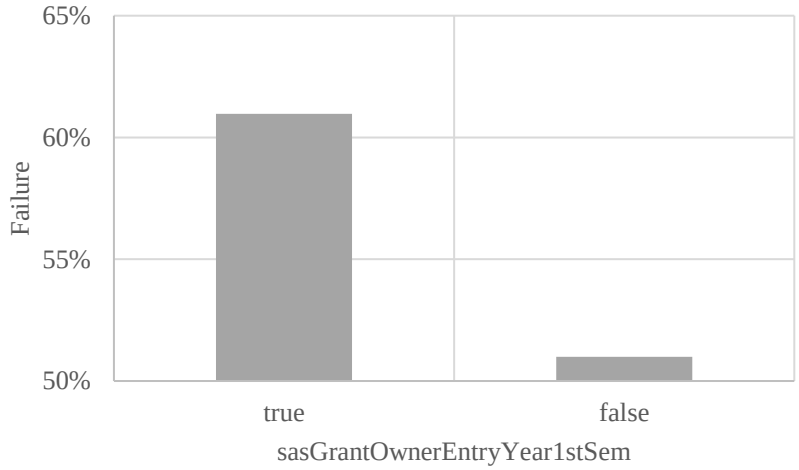


Figure 22 - Impact of sasGrantOwnerEntryYear1stSem on DM_ EntryYear1Sem model.

This freshly included feature belongs to special statute features' group. Besides sasGrantOwnerEntryYear1stSem, only workingStudentEntryYear1stSem presents a substantial relevance, around 1.5%, while all other freshly included special statute features, show negligible explanatory importance. This detail can be explained by special statutes' specificity

and granting requirements, that leads to a residual statute acquirement, which is reflected in DSA results.

As DM_Entrance model, all features' groups except social origin are represented on the higher relevance features. In contrast, educational path features are now represented in greater number. This model's DSA illustrates some curious aspects as the reduced relevance of entryGradeHotdeck feature. It is just the 13th higher impact feature on this model registering around 2% value. Similar behaviour is observed in the related works. According to Slim et al. (2014) replacing impact of fresh evaluation features illustrate the influence of a student's present performance on predicting future performance. On its turn entryGradeHotdeck relates to prior higher education performance, explaining how good student's evaluation performance was before enrolling a higher education degree. A possible explanation for this reduced importance relates to the explanatory context covered by freshly collected evaluation-related features. It is possible that entryGradeHotdeck and new evaluation-related features' explanatory contexts overlap, resulting in the explanatory impact to be transferred to most recent and consequently most impacting evaluation-related features.

The remaining high relevance features maintain a similar importance and impact shown in previous DSA (see Appendix G for remaining features).

4.3.3. DM_EntryYear2Sem model's DSA

Figure 23 shows the relevance of the 8 most importance features in DM_EntryYear2Sem model (see Table F-3 on the Appendix F for extended DM_EntryYear2Sem DSA). The combined contribution of the 8 most impacting features in the model is approximately 63%, denoting an importance spreading tendency, that can be credited to the increasing number of features. Only 12 out of 68 features show an impact above 2%, registering a combined impact around 71%, while the other 56 features explain the remaining 29%.

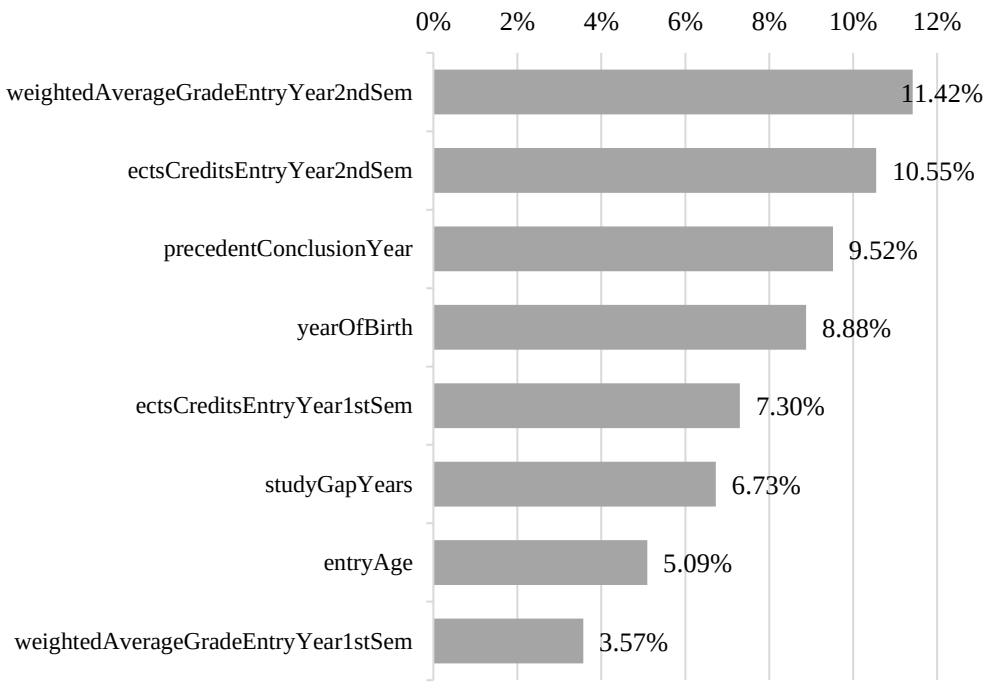


Figure 23 - Features' relevance for DM_EntryYear2Sem model.

Following DM_EntryYear1Sem model's trend, most recent evaluation-related features are the most important features. So, weightedAverageGradeEntryYear2ndSem and ectsCreditsEntryYear2ndSem, being collected at the end of second curricular semester, show the highest impact in the model, around 11%. These features' relevance is aligned with Zimmermann et al. (2015) insights regarding the higher impact of most recent evaluation performances over the academic path. Figures 24 and 25 show the influence of these two features in DM_EntryYear2Sem model.

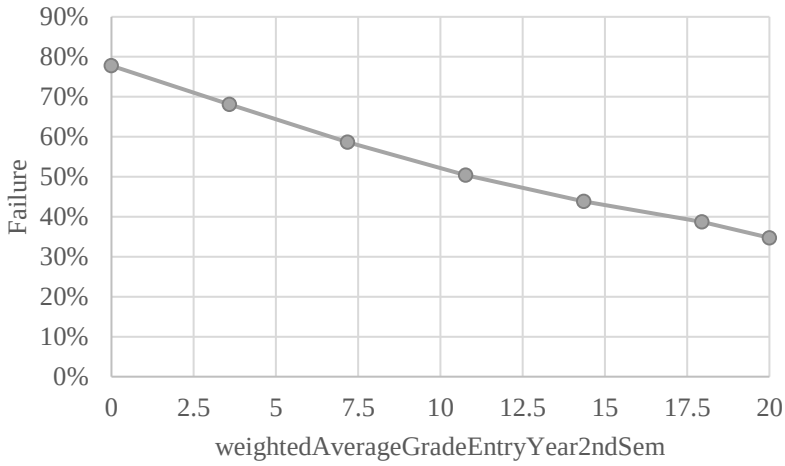


Figure 24 - Impact of weightedAverageGradeEntryYear2ndSem on DM_ EntryYear2Sem model.

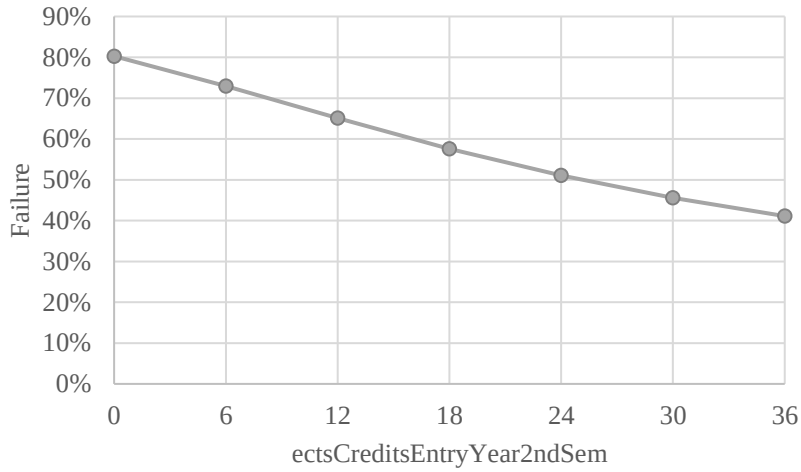


Figure 25 - Impact of ectsCreditsEntryYear2ndSem on DM_ EntryYear2Sem model.

Despite showing equivalent trend, compared to most recent evaluation-related features in DM_EntryYear1Sem model, a slightly lower contribution is verified. This can be explained by the fact that second semester evaluation-related features share their importance with first semester evaluation-related features in DM_EntryYear2Sem model, and by the greater number of features in this model. The inclusion of new high relevance features supports the model's performance improvement.

In contrast to DM_EntryYear1Sem DSA, there is no big percentage gaps between the 8 most relevant features. Although the first semester evaluation-related features still show high importance, they are exceeded by several previously collected features. For instance, precedentConclusionYear, yearOfBirth, studyGapYears and entryAge features that increased their impact in the model compared to DM_EntryYear1Sem DSA results. This set of features are based on similar explanatory vector, as all rely on time-domain analysis. Time-domain features show leveraged impact on failure at this point (end of first curricular year). Student's decision on retention, transition or dropout are potentially more influenced by individual life cycles at the end of first curricular year, directly impacting success (see Appendix H for these and the remaining feature's impact analysis). No special statute feature is represented in the most relevant features, as their importance fade in benefit of freshly added higher importance features. Even that most part of the special statute's features consistently show negligible importance, social support/aid, scholarship and working statute related features still show residual explanatory relevance.

4.3.4. DM_Entrance_IGE model's DSA

The last DSA are based on the supplementary DM_Entrance_IGE model. Figure 26 illustrates the 8 most relevant features in this model (see Table F-4 on the Appendix F for extended DM_EntryYear1Sem DSA).

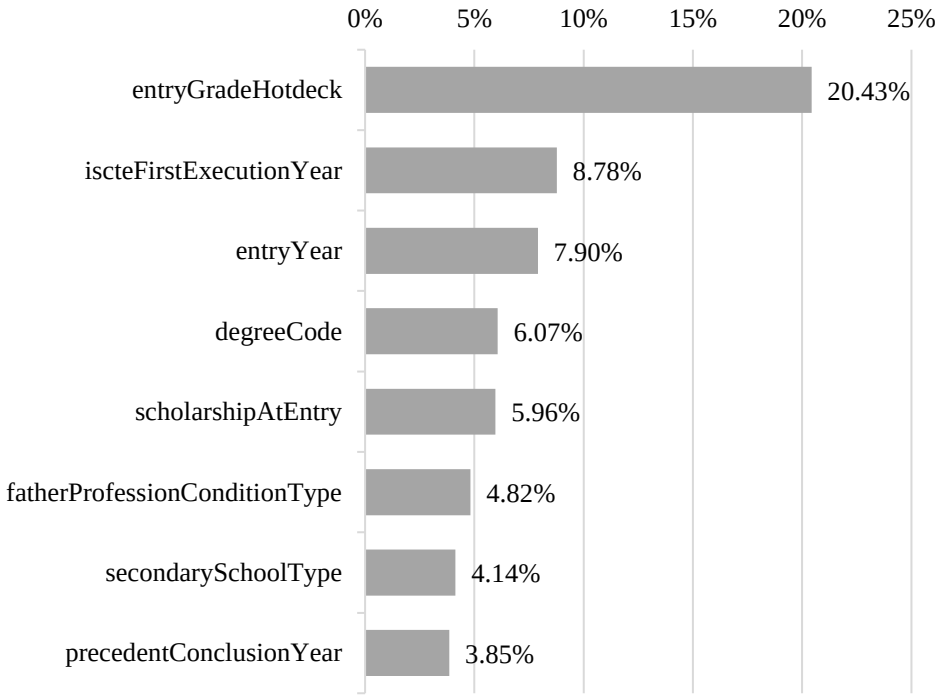


Figure 26 - Features' relevance for DM_Entrance_IGE model.

Compared with DM_Entrance model's DSA, it is observable that entryGradeHotdeck are equally the most impacting feature. Nevertheless, it shows a much higher importance on this model, above 20%, emphasizing its importance in such a smaller dataset model. Subsequent most important features are iscteFirstExecutionYear and entryYear. It is interesting to stress that comparing both models, these two features swap their explanatory relevance with other two time-domain features (studyGapYears and yearOfBirth). These features relate to a current time information while studyGapYears and yearOfBirth are based on previous events. This characteristic and the fact that entryGradeHotdeck shows a much higher importance in this model can relate to a significant fluctuation in required entry grade for IGE and IGE-PL degrees over the years. This hypothesis is corroborated by the fluctuating contribution of iscteFirstExecutionYear and entryYear to failure (see Appendix I for all feature's impact analysis).

On its turn, degreeCode is the 4th most impacting features. Its importance relates to the reduced number classes in this model (only 2 classes). DM dataset used in Kovačić (2012) is also composed by two distinct bachelor's degrees. The author emphasizes the importance of degree when few distinct degrees are submitted to DM analysis. Even based on similar programmes, post-labour nature shows its influence, as IGE-PL demonstrate a much higher contribution to academic failure.

The high importance of fatherOccupationConditionType, around 5%, it is an interesting and unanticipated results. Reviewing the original dataset, this feature is represented by a smaller percentage of "Unknown" class in this model compared to DM_Entrance model, 15% and 23%, respectively. This feature's improved quality in this model's dataset may potentiate its explanatory contribution to be revealed. It is possible to notice through impact analysis that unemployed and retired conditions are more likely to contribute to failure than the other classes. The remaining high relevance features demonstrate similar impact to that shown in DM_Entrance model.

4.3.5. Practical implications

The overall success of an EDM project is very much accounted for providing educational stakeholder, such as, coordinators, teachers and managers, with meaningful information when making decisions concerning educational policies, courses offered, etc (Fernandes et al., 2018). The DSA raised important insights regarding academic success, stressing out which features are more likely to explain unsuccessful paths in different stages of the first curricular year. The analysis is enriched thought identifying specific classes that have higher influence on student's failure within high impact features. As mentioned previously, the earlier in student's academic path, ISCTE-IUL could predict potential failures, the earlier educational support and performance improving policies can be applied to risk groups. Considering it and the impact results obtained for distinct models, we suggest the following institutional guidelines for improving ISCTE-IUL bachelor's freshmen success:

- Providing specific study supporting groups for lower entry grade's students, since the beginning of first curricular semester. Some literature suggests that low performing high school students tend to maintain their low performance level on further higher education. This study's findings corroborate this vision. Considering entry grade values' influence on

failure, this action should be focused on students with entry grade below 13 values. Resources' limitation could force this action's scope to be restricted to lower entry grades' students or degrees that demonstrated higher failure tendency, such as, ISTA school's degrees. The balance between these two groups would optimize this action's effectiveness, considering efforts and available resources.

- Monitoring performance evolution of a specific students' group. This group would be gathered using the following criteria: low entry grade (below 13); older students (above 26 years old); large study gap (above 20 years); and students that came from private high schools. Focusing on admission time collected data these are the most impacting characteristics for student's failure. This information can be used to gather potential risk students. So, we suggest gathering this students' group and monitor their evaluation performance since the beginning of the first curricular semester. Each time that a pre-established performance threshold is breached an institutional action should be triggered. Performance threshold can be defined in distinct ways, such as, failing a crucial course or getting insufficient grade on the first written assessment. This information could also feed an alerting system for at risk students.
- Identifying students that collect less than 18 ECTS or achieve weighted average grade below 7, at the end of the first curricular semester. Extended institutional support can be provided to these students, such as, helping them defining individual study plan for second curricular semester, clearly identifying effort requirements and work balance for better performance achievement. It can be also considered to provide additional support for students that failed courses that take precedence over second semester courses. This pedagogical support is expected to benefit students that is showing difficulties at this early stage of the first curricular year. The more this students' group fell academically integrated, the higher their chances of achieving better performances on subsequent semesters, promoting their success.
- Again, at the end of the second curricular semester, poor performance students should be identified. Proceeding with pedagogical support is important at this stage. Specially for older students or students in life cycles that differ from the great part of their peers, as they may be on the verge of dropping out that inevitably leads to failure. Although, considering the adopted success operationalization, some students will not be able to succeed at this stage of the curricular path, the suggested actions could at least improve their retention rate and potentially mark a turning point into further better performance.

5. Conclusions

EDM has been introduced as an upcoming game-changer for HEI and general educational stakeholders. Academic success is one of the most explored topics, as the knowledge extraction regarding unsuccessful path, in an early stage, has been drawing the attention of HEI. Relying on Fénix dataset, we have gathered a final dataset of 9652 records for regular Bologna bachelor's degrees and 789 records for 4-year Bologna bachelor's degrees. A total of 68 features were used to compose final models after a meticulous features' selection tuning process. This set is represented by 36 special statute features, 12 education path features, 10 socio-demographic features, 6 previous education features, and 4 social origin features. Three distinct models were designed for regular Bologna bachelor's degrees based on distinct stages of first curricular year (entrance, end of the first curricular semester and end of the second curricular semester) and a supplementary model for 4-year bachelors at entrance stage. DT, RF, SVM, and MLPE algorithms were applied on these models for performance evaluation, while modelling robustness was ensured through 10-fold cross-validation. Confusion matrices, ROC curve and AUC metrics were used for models' performance evaluation. DM_Entrance model achieved great performances, particularly for SVM, RF and MLPE algorithms. SVM algorithm clearly achieves the best performance (AUC= 0.77) for this model. This is a robust result, as this model rely exclusively on data collected at admission stage, encompassing socio-demographic, social origin and previous education path features. These algorithms' performance trend is also verified on the remaining models. SVM results can be explaining in part by the improved performance of the algorithm in small size datasets. SVM versions of DM_entryYear1Sem and DM_entryYear2Sem models registered performance results of around, 0.91 and 0.94, respectively. RF's results on DM_entryYear2Sem model is worth mentioning, as it achieved similar performance to SVM, surpassing it on diverse threshold values. This RF performance boost can be explained by algorithm's improved ability to deal with a mixture of numerical and categorical features, bearing in mind that relevant numerical features amount has increased significantly, with the inclusion of first and second semesters' students evaluation features. Although relying on slightly later stages, reducing timings for decision-making and actions to be taken, these models provide an enhanced predictive potential, achieving great performances. These results demonstrate that collecting fresh features during the first curricular year, such as, student's evaluation performance features, it is possible to enrich model's ability to predict unsuccessful cases, while reducing false positive detections. SVM version of DM_Entrance_IGE model achieved a surprisingly good performance (AUC =

0.74). Even relying on different success conditions, it is very interesting to observe that it demonstrates an approximate performance level when compared to DM_Entrance. This result supports the understanding that a well-designed and tuned model, achieves good performance even for reduced datasets.

The DSA was performed, taking SVM models, to unfold drivers of academic failure. For entrance stage (DM_Entrance model) it was possible to observe that the 8 most relevant features, in a total of around 63%, belong to socio-demographic, previous education and special statute features. As widely pointed out on related works, previous student's performance, represented by entryGradeHotDeck, gets the highest rank in relevance analysis, around 11.3%. The second most relevant feature, studyGapYears, around 9.3%, is of great interest, as no similar feature has been described on related works. This finding highlights the importance of business understanding, data understanding, data preparation to final model's performance, as studyGapYears is a computed feature. The remaining relevant features are yearOfBirth, around 9.34%, precedentConclusionYear, around 8.38%, scholarshipAtEntry, around 7.09%, secondarySchoolType, around 6.77%, entryAge, around 5.99%, and degreeSchool, around 5.16%. Although the predominance of previous education features on the most relevant features' set, the diversity in terms of features' groups are remarkable. Time-domain characteristics represented through studyGapYears, yearOfBirth, precedentConclusionYear and entryAge, emphasize the predicting impact of educational gaps and distinct life cycles in the model. The positive impact of scholarshipAtEntry reveals the effective implications of a supporting mechanism provided by the HEI based on students' performance. High school's sector represented by secondarySchoolType, showed an unexpected high relevance. High school's sector is an unusually explored characteristic in EDM literature. Some prudence is recommended for this feature's impact interpretation, as there are a considerable number of "Unknown" class cases to be found in the dataset. In Fénix context, high school institutes, such as, private social welfare entities, private social solidarity institutes, and private colleges are all represented as private high schools. Following these findings, we recommend investigating this aspect on future works. The impact of degreeSchool' classes corroborate the findings described by related works based on same educational system, as it points ISTA (architecture and mostly engineering degrees) as the highest impact class on unsuccessful cases.

For the end of first curricular semester stage (DM_EntryYear1Sem model), the 8 most relevant features, registered a total relevance close to 65%. The two higher relevance features, weightedAverageGradeEntryYear1stSem, around 15.5%, and ectsCreditsEntryYear1stSem,

around 14.5%, are educational path features. This high relevance demonstrates how strong is the impact of current students' performance evaluation features as predictors of academic success. The impact of educational path features is so prominent that a clear reduction on remaining features are verified, such as previous performance evaluation features. The remaining high relevance features are precedentConclusionYear, around 7.40%, yearOfBirth, around 7.05%, averageGradeEntryYear1stSem, around 6.81%, studyGapYears, around 4.67%, sasGrantOwnerEntryYear1stSem, around 4.50% and secondarySchoolType, around 4.44%. Socio-demographic, previous education and special statute features are still represented in high relevance features' set, as well as, new educational path features, that is now the most represented features' group. Impact results for sasGrantOwnerEntryYear1stSem raise some concern, as it shows clearly that students, which are granted a social support during the first curricular semester are more likely to not succeed. Despite the fact that this impact can be addressed to the root cause, as it relates, in first instance, to lower socio-economical level, it raises some warnings regarding how social support efforts are being capitalized into promoting success.

With regards to the end of second curricular semester stage (DM_EntryYear2Sem model), the combined contribution of the 8 most relevant features was around 63%. As reviewed on DM_EntryYear1Sem model, the most recent evaluation performance features, weightedAverageGradeEntryYear2ndSem, around 11.4%, and ectsCreditsEntryYear2ndSem, around 10.5%, compose the two most relevant features. Although these features' robust relevance, they share importance with first semester evaluation performance features, ectsCreditsEntryYear1stSem, around 7.30%, and weightedAverageGradeEntryYear1stSem, around 3.57%. The remaining four relevant features are precedentConclusionYear, around 9.52%, yearOfBirth, around 8.88%, studyGapYears, around 6.73%, and entryAge, around 5.09%. There are no special statute features to be found on high relevant feature's set, as their importance fade in benefit of freshly added higher importance features. An inspiring aspect of this analysis is the improved impact of time-domain features on this model compared to previous stage model. Some literature suggests that students' decision on retention, transition or dropout are influenced by individual's life cycles and social integration, showing its higher impact at the end of first curricular year. These premises, directly impacting academic success, partially explain these features leveraged impact at this stage.

Supplementary DSA analysis based on DM_Entrance_IGE model for entrance stage, presented similar relevance and impact for most part of the features when compared to DM_Entrance.

The exceptions are degreeCode and fatherOccupationConditionType that show relevant impact in this model. The impact of degreeCode strictly relates to post-labour degree's characteristics, as post-labour IGE degree shows a much higher contribution to failure. The high relevance of fatherOccupationConditionType sets up a somehow surprising result, as no other social origin feature achieved high impact in this study. It is possible that higher data quality percentage in this specific model, could potentiate its explanatory contribution. This feature impact analysis showed unemployed and retired classes as the most likely to explain academic failure. These interesting finding potentially foresees that social origin features' predictive potential might have been obfuscated by the bad quality of source data. Remaining relevance results demonstrate that great part of reviewed features impact regular and 4-year bachelors in a similar fashion.

Based on the extracted knowledge, we suggest the following set of institutional guidelines to promote academic success in ISCTE-IUL: provide study support groups for lower entry grade's students since the beginning of the first curricular semester, especially for ISTA degrees' students; create an alerting and monitoring framework for students that present impacting characteristics, such as, low entry grades (below 13), older students (over 26 years old), big study gap (above 20 years); provide additional educational mentoring and guidance to students that collected less than 18 ECTS or achieved weighted average grade below 7 at the end of first curricular semester; and provide pedagogical support and guidance to poor performance students at the end of second curricular semester (with special attention to older students). It is also recommended to maintain the performance and achievement related policies, such as scholarships, and potentially invest in smaller complementary benefits. On the other hand, additional mentoring and follow-ups are recommended to students that receive social support, as this type of support did not show positive impact on success.

A great part of this study's effort consisted in data quality tasks. Nevertheless, predictive potential has been lost due to some bad quality data, this is a limitation on this study. For instance, precedentDegreeDesignation and highSchoolDegreeType ended up excluded, reducing high school's characterization spectrum. This potentially important explanatory vector ended up being characterized by a single feature (secondarySchoolType). Even being included, poor data quality features could find their true explanatory potential put at risk, as it could lead their relevance in the model to be enhanced, reduced or flattened. For instance, after an extensive data processing effort, we were just able to fill around half of the cases for fatherOccupation and motherOccupation, while the remaining half had to be filled with

"Unknown/Others" class. All these features have in common that their original data have been collected through open text fields. So, we suggest automating a standard data collection process in order to collect data, while ensuring its quality at the same time. Consistent and coherent academic data is easier to analyse and include in further DM models and frameworks. Simple processes, as empty/incomplete fields validation could be applied to academic forms in order to reduce inadequate data. Creating a segmented list of answers for each field would enhance the quality of collected data. For instance, it would be interesting to use a static list based on ESCO's multilingual classification of occupations for collecting parents' occupation data. The above suggestions would facilitate and promote DM applications as it would potentially reduce the data preparation, cleansing and quality stages' effort as well as increasing the number of data and specially the number of candidates' features to be included in the model.

For future work we propose the following avenues: designing individual school's DM models based on presented models, in order to capture specific school's characteristics; considering additional data sources, such as, end of semester's student satisfaction surveys; scrutinizing the effect of post-labour feature on academic failure; and extending data quality approaches on social origin, candidacy preference and high school related features and revisiting their impact on predicting academic failure. Ultimately, an information system encompassing these models can be used as a data-driven decision-making framework for ISCTE-IUL to support and optimize institutional policies and actions for academic success.

6. References

Ahmad, F., Ismail, N., & Aziz, A. A. (2015). The prediction of students' academic performance using classification data mining techniques. *Applied Mathematical Sciences, 9*(129), 6415-6426.

Amrieh, E. A., Hamtini, T., & Aljarah, I. (2015). Preprocessing and analyzing educational data set using X-API for improving student's performance. In *Applied Electrical Engineering and Computing Technologies* (AEECT), 2015 *IEEE Jordan Conference on* (pp. 1-5). IEEE.

Apté, C., & Weiss, S. (1997). Data mining with decision trees and decision rules. *Future generation computer systems*, 13(2-3), 197-210.

Asif, R., Merceron, A., Ali, S. A., & Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers & Education, 113*, 177-194.

Astin, A. W. (2012). *Assessment for excellence: The philosophy and practice of assessment and evaluation in higher education*. Rowman & Littlefield Publishers.

Badr, G., Algobail, A., Almutairi, H., & Almutery, M. (2016). Predicting students' performance in university courses: a case study and tool in KSU mathematics department. *Procedia Computer Science, 82*, 80-89.

Baker, R. S. (2010). Data mining for education. *International Encyclopedia of Education, 7*(3), 112-118.

Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *JEDM Journal of Educational Data Mining, 1*(1), 3-17.

Barber, R., & Sharkey, M. (2012). Course correction: Using analytics to predict course success. In Proceedings of the 2nd international conference on learning analytics and knowledge (pp. 259-262). ACM.

Barraza, N., Moro, S., Ferreyra, M., & de la Peña, A. (2019). Mutual information and sensitivity analysis for feature selection in customer targeting: A comparative study. *Journal of Information Science, 45*(1), 53-67.

bin Mat, U., Buniyamin, N., Arsad, P. M., & Kassim, R. (2013). An overview of using academic analytics to predict and improve students' achievement: A proposed proactive intelligent intervention. In *Engineering Education (ICEED), 2013 IEEE 5th Conference on* (pp. 126-130). IEEE.

Börjesson, M., Broady, D., Le Roux, B., Lidegran, I., & Palme, M. (2016). Cultural capital in the elite subfield of Swedish higher education. *Poetics, 56*, 15-34.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7), 1145-1159.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.

Canito, J., Ramos, P., Moro, S., & Rita, P. (2018). Unfolding the relations between companies and technologies under the Big Data umbrella. *Computers in Industry, 99*, 1-8.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and Wirth, R. (2000). *CRISP-DM 1.0 -Step-by-step data mining guide*, CRISP-DM Consortium.

Cheewaprakobkit, P. (2015). Predicting student academic achievement by using the decision tree and neural network techniques. *Catalyst, 12*(2), 34-43.

Choi, N. (2005). Self-efficacy and self-concept as predictors of college students' academic performance. *Psychology in the Schools, 42*(2), 197-205.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3), 273-297.

Cortez, P. (2010). Data mining with neural networks and support vector machines using the R/rminer tool. In *Industrial Conference on Data Mining* (pp. 572-583). Springer, Berlin, Heidelberg.

Cortez, P., & Embrechts, M. J. (2011, April). Opening black box data mining models using sensitivity analysis. In *2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)* (pp. 341-348). IEEE.

Daud, A., Aljohani, N. R., Abbasi, R. A., Lytras, M. D., Abbas, F., & Alowibdi, J. S. (2017). Predicting student performance using advanced learning analytics. In *Proceedings of the 26th International Conference on World Wide Web Companion* (pp. 415-421). International World Wide Web Conferences Steering Committee.

Delavari, N., Phon-Amnuaisuk, S., & Beikzadeh, M. R. (2008). Data mining application in higher learning institutions. *Informatics in Education-International Journal*, 7, 31-54.

Delen, D. (2011). Predicting student attrition with data mining methods. *Journal of College Student Retention: Research, Theory & Practice*, *13*(1), 17-35.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, *55*(10), 78-87.

Dutt, A., Ismail, M. A., & Herawan, T. (2017). A systematic review on educational data mining. *IEEE Access*, *5*, 15991-16005.

Eggens, L., Van der Werf, M. P. C., & Bosker, R. J. (2008). The influence of personal networks and social support on study attainment of students in university education. *Higher education*, 55(5), 553-573.

Ferguson, R. (2012). Learning analytics: drivers, developments and challenges. International Journal of Technology Enhanced Learning, 4(5/6), 304-317.

Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R., & Van Erven, G. (2019). Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. *Journal of Business Research, 94*, 335-343.

Goker, H., Bulbul, H. I., & Irmak, E. (2013). The estimation of students' academic success by data mining methods. In Machine Learning and Applications (ICMLA), 2013 *12th International Conference on* (Vol. 2, pp. 535-539). IEEE.

Gore Jr, P. A. (2006). Academic self-efficacy as a predictor of college outcomes: Two incremental validity studies. *Journal of Career Assessment*, 14(1), 92-115.

Hannon, O., Smith, L. R., & Lã, G. (2017). Success at University: The Student Perspective. In *Success in Higher Education* (pp. 257-268). Springer, Singapore.

Haykin, S. (1994). *Neural networks: a comprehensive foundation*. Prentice Hall PTR.

Herzog, S. (2005). Measuring determinants of student return vs. dropout/stopout vs. transfer: A first-to-second year analysis of new freshmen. *Research in higher education*, 46(8), 883-928.

Howard, S. K., Ma, J., & Yang, J. (2016). Student rules: Exploring patterns of students' computer-efficacy and engagement with digital technologies in learning. *Computers & Education*, *101*, 29-42.

Hsia, L. H., Huang, I., & Hwang, G. J. (2016). Effects of different online peer-feedback approaches on students' performance skills, motivation and self-efficacy in a dance course. *Computers & Education*, *96*, 55-71.

Hu, Y. H., Lo, C. L., & Shih, S. P. (2014). Developing early warning systems to predict students' online learning performance. *Computers in Human Behavior*, *36*, 469-478.

Junco, R., & Clem, C. (2015). Predicting course outcomes with digital textbook usage data. The Internet and Higher Education, 27, 54-63.

Kabra, R. R., & Bichkar, R. S. (2011). Performance prediction of engineering students using decision trees. *International Journal of Computer Applications*, *36*(11), 8-12.

Kahu, E. R., & Nelson, K. (2018). Student engagement in the educational interface: understanding the mechanisms of student success. *Higher Education Research & Development*, *37*(1), 58-71.

Kim, J. H. (2009). Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational statistics & data analysis*, 53(11), 3735-3745.

Kitchenham, B., Brereton, O. P., Budgen, D., Turner, M., Bailey, J., & Linkman, S. (2009). Systematic literature reviews in software engineering–a systematic literature review. *Information and Software Technology*, 51(1), 7-15.

Koedinger, K., Cunningham, K., Skogsholm, A., & Leber, B. (2008). An open repository and analysis tools for fine-grained, longitudinal learner data. In *Educational Data Mining*.

Kohavi, R. (1998). Glossary of terms. *Special issue on applications of machine learning and the knowledge discovery process*, 30(271), 127-132

Kotsiantis, S. B., & Pintelas, P. E. (2005). Predicting students marks in hellenic open university. In *Advanced Learning Technologies*, 2005. ICALT 2005. *Fifth IEEE International Conference on* (pp. 664-668). IEEE.

Kovačić, Z. (2012). Predicting student success by mining enrolment data. Research in Higher Education Journal, 1.

Kostopoulos, G., Kotsiantis, S., Pierrakeas, C., Koutsonikos, G., & Gravvanis, G. A. (2018). Forecasting students' success in an open university. *International Journal of Learning Technology*, *13*(1), 26-43.

Kuh, G. D., Kinzie, J. L., Buckley, J. A., Bridges, B. K., & Hayek, J. C. (2006). *What matters to student success: A review of the literature* (Vol. 8). Washington, DC: National Postsecondary Education Cooperative.

Leppänen, L., Leinonen, J., Ihantola, P., & Hellas, A. (2017). Predicting Academic Success Based on Learning Material Usage. In *Proceedings of the 18th Annual Conference on Information Technology Education* (pp. 13-18). ACM.

Lundberg, C. A., & Schreiner, L. A. (2004). Quality and frequency of faculty-student interaction as predictors of learning: An analysis by student race/ethnicity. *Journal of College Student Development*, 45(5), 549-565.

Martínez, D. L. L. R., & Gómez, C. E. P. (2014). Contributions from Data Mining to Study Academic Performance of Students of a Tertiary Institute. *American Journal of Educational Research, 2*(9), 713-726.

Martins, M. P., Migueis, V. L., & Fonseca, D. S. B. (2018). A data mining approach to predict undergraduate students' performance. In 2018 *13th Iberian Conference on Information Systems and Technologies* (CISTI) (pp. 1-7). IEEE.

Martins, S., Carvalho, H., Ávila, P., & da Costa, A. F. (2017). Policies for Widening Participation and Success Factors in Portuguese Higher Education. *Creative Education, 8*, 210-230.

Mayilvaganan, M., & Kalpanadevi, D. (2014). Comparison of classification techniques for predicting the performance of students' academic environment. In *Communication and Network Technologies* (ICCNT), 2014 *International Conference on* (pp. 113-118). IEEE.

Minaei-Bidgoli, B., & Punch, W. F. (2003). Using genetic algorithms for data mining optimization in an educational web-based system. In *Genetic and Evolutionary Computation Conference* (pp. 2252-2263). Springer, Berlin, Heidelberg.

Mishra, T., Kumar, D., & Gupta, S. (2014). Mining students' data for prediction performance. In *Advanced Computing & Communication Technologies (ACCT)*, 2014 *Fourth International Conference on* (pp. 255-262). IEEE.

Moro, S., Cortez, P., & Rita, P. (2018). A divide-and-conquer strategy using feature relevance and expert knowledge for enhancing a data mining approach to bank telemarketing. *Expert Systems*, *35*(3), e12253.

Moro, S., Laureano, R., & Cortez, P. (2011). Using data mining for bank direct marketing: An application of the crisp-dm methodology. In *Proceedings of European Simulation and Modelling Conference-ESM'2011* (pp. 117-121). EUROSIS-ETI.

Natek, S., & Zwilling, M. (2014). Student data mining solution–knowledge management system related to higher education institutions. *Expert Systems with Applications, 41*(14), 6400-6407.

Olama, M. M., Thakur, G., McNair, A. W., & Sukumar, S. R. (2014). Predicting student success using analytics in course learning management systems. In *Next-Generation Analyst II* (Vol. 9122, p. 91220M). International Society for Optics and Photonics.

Osmanbegović, E., & Suljić, M. (2012). Data mining approach for predicting student performance. *Economic Review, 10*(1), 3-12.

Papamitsiou, Z., & Economides, A. A. (2014). Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. *Journal of Educational Technology & Society, 17*(4), 49-64.

Parker, J. D., Summerfeldt, L. J., Hogan, M. J., & Majeski, S. A. (2004). Emotional intelligence and academic success: Examining the transition from high school to university. *Personality and individual differences, 36*(1), 163-172.

Pascarella, E. T., & Terenzini, P. T. (2005). *How college affects students: A third decade of research* (Vol. 2). San Francisco: Jossey-Bass.

Pena, A., Domínguez, R., & de Jesus Medel, J. (2009). Educational data mining: a sample of review and study case. *World Journal On Educational Technology, 1*(2), 118-139.

Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Systems with Applications*, 41(4), 1432-1462.

Rahman, M. H., & Islam, M. R. (2017). Predict Student's Academic Performance and Evaluate the Impact of Different Attributes on the Performance Using Data Mining Techniques. In 2017 *2nd International Conference on Electrical & Electronic Engineering (ICEEE)* (pp. 1-4). IEEE.

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-validation. *Encyclopedia of database systems*, 532-538.

Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40*(6), 601-618.

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications, 33*(1), 135-146.

Romero, C., Ventura, S., Espejo, P. G., & Hervás, C. (2008). Data mining algorithms to classify students. In *Educational Data Mining* 2008.

Romero, C., Ventura, S., & García, E. (2008). Data mining in course management systems: Moodle case study and tutorial. *Computers & Education, 51*(1), 368-384.

Roy, S., & Garg, A. (2017). Analyzing performance of students by using data mining techniques a literature survey. In *Electrical, Computer and Electronics (UPCON), 2017 4th IEEE Uttar Pradesh Section International Conference on* (pp. 130-133). IEEE.

Sachin, R. B., & Vijay, M. S. (2012). A survey and future vision of data mining in educational field. In *Advanced Computing & Communication Technologies (ACCT), 2012 Second International Conference on* (pp. 96-100). IEEE.

Saltelli, A., Tarantola, S., & Campolongo, F. (2000). Sensitivity anaysis as an ingredient of modeling. *Statistical Science*, 15(4), 377-395.

Santos, M. R., Laureano, R. M., & Moro, S. (2019). Unveiling Research Trends for Organizational Reputation in the Nonprofit Sector. *VOLUNTAS: International Journal of Voluntary and Nonprofit Organizations*, DOI: 10.1007/s11266-018-00055-7.

Shahiri, A. M., & Husain, W. (2015). A review on predicting student's performance using data mining techniques. *Procedia Computer Science, 72*, 414-422.

Shearer, C. (2000). The CRISP-DM model: the new blueprint for data mining. *Journal of data warehousing*, 5(4), 13-22.

Simeunović, V., & Preradović, L. (2014). Using data mining to predict success in studying. *Croatian Journal of Education, 16*(2), 491-523.

Slim, A., Heileman, G. L., Kozlick, J., & Abdallah, C. T. (2014). Predicting student success based on prior performance. In *Computational Intelligence and Data Mining (CIDM),* 2014 *IEEE Symposium on* (pp. 410-415). IEEE.

Strecht, P., Cruz, L., Soares, C., Mendes-Moreira, J., & Abreu, R. (2015). A Comparative Study of Classification and Regression Algorithms for Modelling Students' Academic Performance. *International Educational Data Mining Society*, ERIC.

Sukhija, K., Jindal, M., & Aggarwal, N. (2015). The recent state of educational data mining: A survey and future visions. In *MOOCs, Innovation and Technology in Education (MITE), 2015 IEEE 3rd International Conference on* (pp. 354-359). IEEE.

Superby, J. F., Vandamme, J. P., & Meskens, N. (2006). Determination of factors influencing the achievement of the first-year university students using data mining methods. In *Workshop on Educational Data Mining* (Vol. 32, p. 234).

Taruna, S., & Pandey, M. (2014). An empirical analysis of classification techniques for predicting academic performance. In *Advance Computing Conference* (IACC), 2014 *IEEE Internationa*l (pp. 523-528). IEEE.

Tinto, V. (1997). Classrooms as communities: Exploring the educational character of student persistence. *The Journal of Higher Education, 68*(6), 599-623.

Tinto, V. (1999). Taking retention seriously: Rethinking the first year of college. *NACADA journal, 19*(2), 5-9.

Tinto, V. (2006). Research and practice of student retention: What next? *Journal of College Student Retention: Research, Theory & Practice, 8*(1), 1-19.

Tracey, T. J., Allen, J., & Robbins, S. B. (2012). Moderation of the relation between person–environment congruence and academic success: Environmental constraint, personal flexibility and method. *Journal of Vocational Behavior, 80*(1), 38-49.

Trevor, H., Robert, T., & JH, F. (2009). The elements of statistical learning: data mining, inference, and prediction.

Trstenjak, B., & Donko, D. (2014). Determining the impact of demographic features in predicting student success in Croatia. In *Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 2014 *37th International Convention on* (pp. 1222-1227). IEEE.

Vandamme, J. P., Meskens, N., & Superby, J. F. (2007). Predicting academic performance by data mining methods. *Education Economics, 15*(4), 405-419.

Vora, D. R., & Iyer, K. (2018). EDM – survey of performance factors and algorithms applied. *International Journal of Engineering & Technology, 7*(2-6), 93-97.

Vuttipittayamongkol, P. (2016). Predicting factors of academic performance. In *Defence Technology* (ACDT), 2016 *Second Asian Conference on* (pp. 161-166). IEEE.

Watson, C., Li, F. W., & Godwin, J. L. (2013). Predicting performance in an introductory programming course by logging and analyzing student programming behavior. In 2013 IEEE 13th International Conference on Advanced Learning Technologies (pp. 319-323). IEEE

Wood, L. N., & Breyer, Y. A. (2017). Success in higher education. In *Success in higher education* (pp. 1-19). Springer, Singapore.

York, T. T., Gibson, C., & Rankin, S. (2015). Defining and Measuring Academic Success. *Practical Assessment, Research & Evaluation, 20*(5), 1-20.

You, J. W. (2016). Identifying significant indicators using LMS data to predict course achievement in online learning. The Internet and Higher Education, 29, 23-30.

Zhou, Q., Zheng, Y., & Mou, C. (2015). Predicting students' performance of an offline course from their online behaviors. In *DICTAP* (pp. 70-73).

Zimmermann, J., Brodersen, K. H., Heinimann, H. R., & Buhmann, J. M. (2015). A model-based approach to predicting graduate-level performance using indicators of undergraduate-level performance. *Journal of Educational Data Mining, 7*(3), 151-176.

Appendix

Appendix A – Related works details

Table A-1 - Related works publication details

Reference	Source Type	Source Title	Publisher	Scope Focus	Indexed DB			Citations			Quartile	
					Sco	ISI	GS	Sco	ISI	GS	Sco	ISI
Minaei-Bidgoli et al. (2003)	CP	Genetic and evolutionary computation conference	Springer	Edu	Y	Y	Y	64	33	203		
Kotsiantis & Pintelas (2005)	CP	International Conference on Advanced Learning Technologies	IEEE	CS/IS	Y	Y	Y	36	10	77		
Superby et al. (2006)	CP	Workshop on Educational Data Mining	International Educational Data Mining Society	CS/IS	N	N	Y	-	-	137		
Vandamme et al. (2007)	JA	Education Economics	Taylor & Francis	Edu	N	N	Y	-	-	142		
Romero et al. (2008a)	CP	International Conference on Educational Data Mining	International Educational Data Mining Society	CS/IS	Y	N	Y	165	-	340		
Romero et al. (2008b)	JA	Computers & Education	Elsevier	CS/IS Edu	Y	Y	Y	474	351	1107	Q1	Q1
Kabra & Bichkar (2011)	JA	International Journal of Computer Applications	Foundation of Computer Science (FCS)	CS/IS	N	N	Y	-	-	87		
Delen (2011)	JA	Journal of College Student Retention: Research, Theory & Practice	SAGE	Edu	Y	N	Y	13	-	58	Q2	
Barber & Sharkey (2012)	CP	International conference on learning analytics and knowledge	ACM	CS/IS	Y	N	Y	43	-	110		
Osmanbegović & Suljić (2012)	JA	Journal of Economics & Business/Economic Review	University of Tuzla, Faculty of Economics, Bosnia and Herzegovina	Other	N	N	Y	-	-	150		
Watson et al. (2013)	CP	International Conference on Advanced Learning Technologies	IEEE	CS/IS	Y	Y	Y	53	13	85		
Goker et al. (2013)	CP	Machine Learning and Applications (ICMLA)	IEEE	CS/IS	Y	Y	Y	3	0	6		
Mishra et al. (2014)	CP	International Conference on Advanced Computing & Communication Technologies	IEEE	CS/IS	Y	Y	Y	33	10	43		
Trstenjak & Donko (2014)	CP	International Convention on Information and Communication Technology, Electronics and Microelectronics ({MIPRO})	IEEE	CS/IS	Y	Y	Y	5	2	11		

Reference	Source Type	Source Title	Publisher	Scope Focus	Indexed DB			Citations			Quartile	
					Sco	ISI	GS	Sco	ISI	GS	Sco	ISI
Slim et al. (2014)	CP	Computational Intelligence and Data Mining (CIDM)	IEEE	CS/IS	Y	Y	Y	8	2	13		
Olama et al. (2014)	CP	Next-Generation Analyst {II}	SPIE	CS/IS	Y	Y	Y	3	0	3		
Martínez & Gómez (2014)	JA	American Journal of Educational Research	Science and Education Publishing	Edu	N	N	Y	-	-	8		
Simeunović & Preradović (2014)	JA	Croatian Journal of Education	Faculty of teacher education, Zagreb, Croatia	Edu	N	Y	Y	-	4	7		Q4
Natek & Zwilling (2014)	JA	Expert Systems with Applications	Elsevier	Edu	Y	Y	Y	58	48	97	Q1	Q1
Mayilvaganan & Kalpanadevi (2014)	CP	Communication and Network Technologies (ICCNT	IEEE	CS/IS	Y	Y	Y	17	8	42		
Hu et al. (2014)	JA	Computers in Human Behavior	Elsevier	CS/IS	Y	Y	Y	37	27	72	Q1	Q1
Taruna & Pandey (2014)	CP	Advance Computing Conference (IACC)	IEEE	CS/IS	Y	Y	Y	19	11	28		
Junco & Clem (2015)	JA	Internet and Higher Education	Elsevier	CS/IS Edu	Y	Y	Y	28	26	55	Q1	Q1
Cheewaprakobkit (2015)	JA	Catalyst	Asia-Pacific International University, Institute Press	CS/IS	N	N	Y	-	-	2		
Stretch et al. (2015)	CP	International Conference on Educational Data Mining (EDM)	International Educational Data Mining Society	CS/IS	N	N	Y	-	-	27		
Zimmermann et al. (2015)	JA	Journal of Educational Data Mining	International Working Group on Educational Data Mining	CS/IS Edu	N	N	Y	-	-	22		
Zhou et al. (2015)	CP	International Conference on Digital Information and Communication Technology and its Applications (DICTAP)	IEEE	CS/IS	Y	Y	Y	4	2	9		
Ahmad et al. (2015)	JA	Applied Mathematical Sciences	Hikari	CS/IS	Y	N	Y	16		35	Q3	
Amrieh et al. (2015)	CP	Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)	IEEE	CS/IS	Y	Y	Y	9	0	18		
You (2016)	JA	Internet and Higher Education	Elsevier	CS/IS Edu	Y	Y	Y	43	26	78	Q1	

Reference	Source Type	Source Title	Publisher	Scope Focus	Indexed DB			Citations			Quartile	
					Sco	ISI	GS	Sco	ISI	GS	Sco	ISI
Badr et al. (2016)	JA	Procedia Computer Science	Elsevier	CS/IS	Y	Y	Y	2	0	11	Q3	
Vuttipittayamongkol (2016)	CP	Asian Conference on Defence Technology ({ACDT})	IEEE	CS/IS	Y	N	Y	2	-	4		
Daud et al. (2017)	CP	26th International Conference on World Wide Web Companion	ACM	CS/IS	N	N	Y	-	-	25		
Martins el al. (2017)	JA	Creative Education	Scientific Research Publishing	Edu	N	N	Y	-	-	0		
Asif et al. (2017)	JA	Computers & Education	Elsevier	CS/IS Edu	Y	Y	Y	33	12	64	Q1	Q1
Rahman & Islam (2017)	CP	International Conference on Electrical & Electronic Engineering (ICEEE)	IEEE	CS/IS	Y	N	Y	1	-	0		
Leppanen et al. (2017)	CP	Conference on Information Technology Education - {SIGITE}	ACM	CS/IS	Y	N	Y	0	-	2		
Kostopoulos et al. (2018)	JA	International Journal of Learning Technology	Inderscience	CS/IS	Y	Y	Y	1	0	1	Q4	
Martins et al. (2018)	CP	Iberian Conference on Information Systems and Technologies (CISTI)	IEEE	CS/IS	Y	Y	Y	0	0	0		
Fernandes et al. (2018)	JA	Journal of Business Research	Elsevier	Other	Y	Y	Y	2	0	4	Q1	Q1

Table A-2 – EDM literature reviews

Reference	Nr. studies	Timeframe	Scope	Main Contributions
Romero & Ventura (2007)	81	1995 - 2005	Present EDM applications.	Presents EDM specific applications grouping it by tasks and point some future research guidelines.
Delavari et al. (2008)	8	2002 - 2004	Explore existing EDM areas.	Presents the capabilities of DM in the higher educational context.
Baker & Yacef (2009)	45	1995 - 2009	Review the history and current trends in EDM.	Identifies EDM research problems.
Pena et al. (2009)	91	1995 - 2009	Explore three main topics: DM, EDM and web-based Education Systems.	Presents conclusion regarding Web-based Education Systems.

Reference	Nr. studies	Timeframe	Scope	Main Contributions
Romero & Ventura (2010)	235	1995 - 2009	Explore the type of data and DM techniques used in EDM Categorize type of educational task that they resolve.	Presents most common educational environment tasks resolved thought DM and present some future research guidelines.
Sachin & Vijay (2012)	26	1997 - 2011	Survey the applications of data mining techniques to traditional educational systems.	Discusses and summarize key applications of EDM.
bin Mat et al. (2013)	22	2001 - 2013	Explore academic analytic tools in educational institutions and how institution can predict student performance and achievement.	Evaluates DM applications. Presents guidelines to improve students' achievement prediction
Papamitsiou & Economides (2014)	40	2008 - 2013	Categorize the research questions, methodology and findings within EDM literature.	Evaluates findings of the collected studies and highlighted four distinct major directions of the EDM empirical research.
Peña-Ayala (2014)	240	2010 - first quarter of 2013	Summarize, organize, analyse, and discuss EDM approach outcomes.	Motivates and points opportunities and guidelines within distinct EDM subjects.
Sukhija et al. (2015)	19	2001 - 2015	Explore EDM objectives, components, tools and techniques applied.	Suggests four main gaps in EDM research field.
Shahiri et al. (2015)	25	2002 - 2015	Review explanatory attributes analysed by the various techniques of classification.	Compares method of prediction for analysing the performance of students'.
Dutt et al. (2017)	> 100	1983 - 2016	Cluster algorithm and its applicability in the context of EDM.	Presents insights on educational data clustering and avenues for further research.
Roy & Garg (2017)	20	2006 - 2016	Review EDM studies that analyse student's performance.	Summarizes DM techniques and tools and present their Pros and Cons.
Vora & Iyer (2018)	33	2006 - 2017	Present current EDM state and identify the algorithms applied, goals and methods.	Identifies lacunas and challenges in Algorithms applied, performance factors considered, and data used in EDM.

Appendix B – Data source model

Figure B-1 – Data source main tables' model

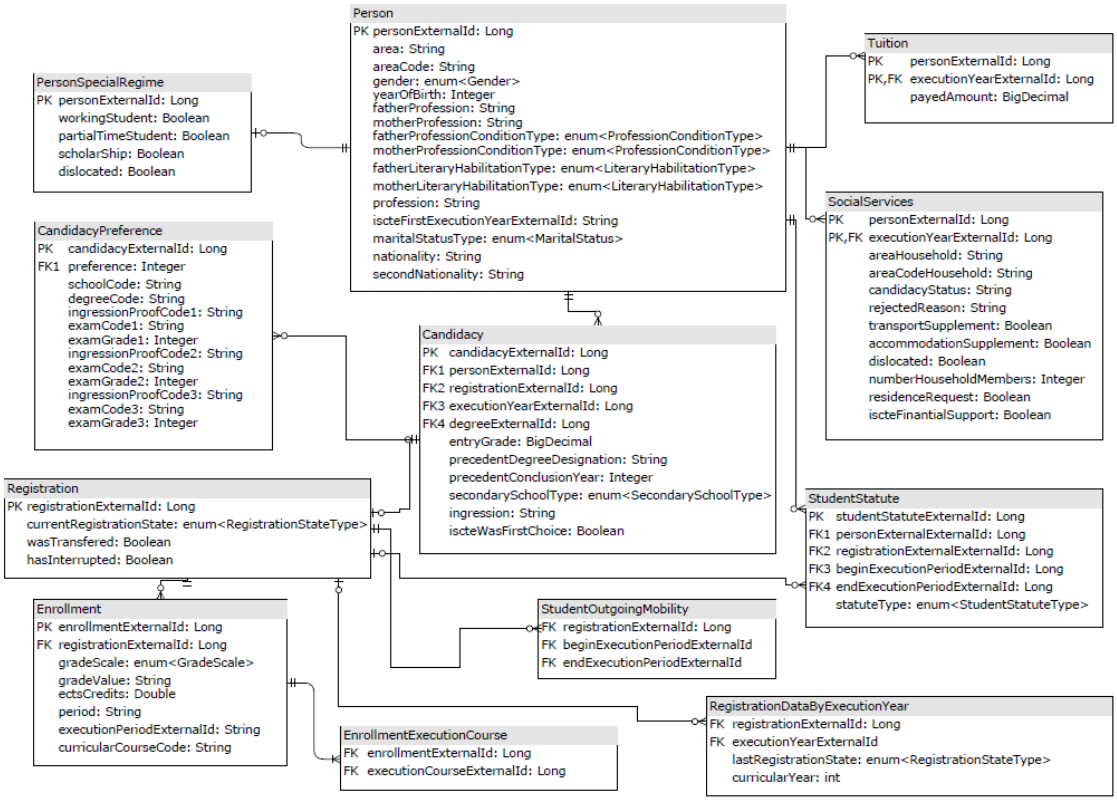
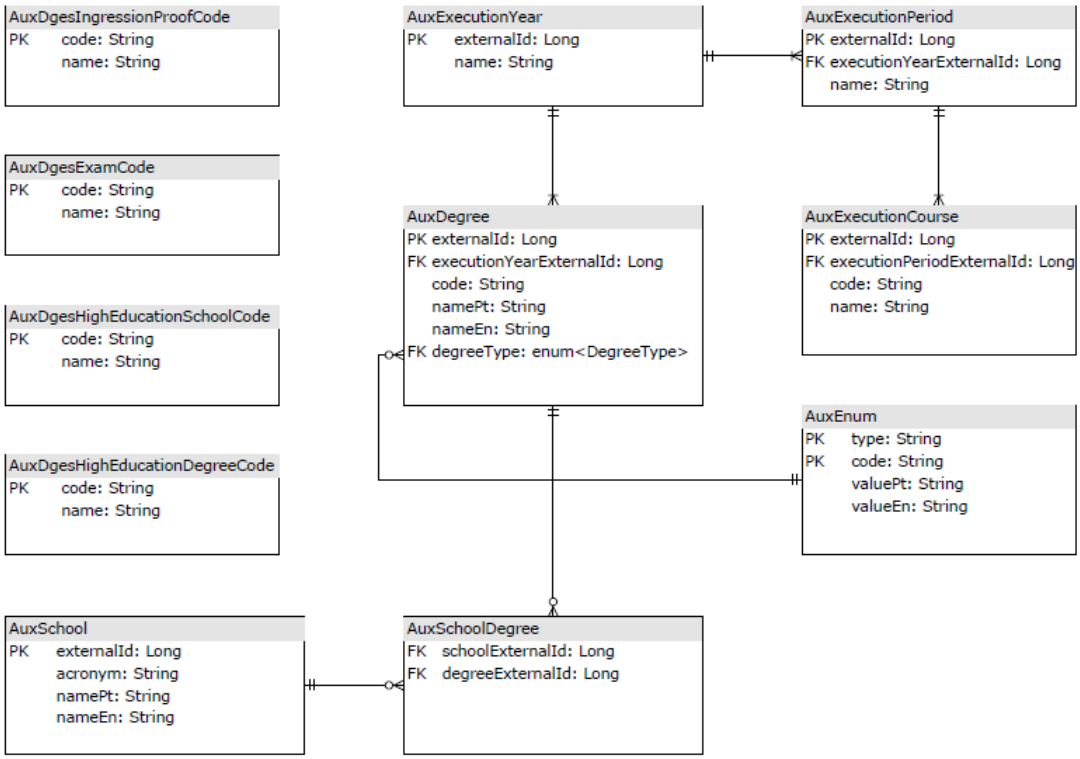


Figure B-2 - Figure B-1 – Data source lookup tables' model



Appendix C – Features detail

Table C-1 –Features' description and classes

Feature	Description	Classes
area	Student's residence area	"1495-041";"2795-050"; etc.
areaCode	Student's residence postal code	"Linda-a-Velha";"Algés"; etc.
gender	Gender	"M";"F"
yearOfBirth	Year of birth	"1986";"1967"; etc.
fatherOccupation	Father's occupation	"Managers";"Elementary occupations";etc.
motherOccupation	MotherOccupation	"Managers";"Elementary occupations";etc.
fatherOccupationConditionType	Father's condition in labour force	"Unemployed"; "Dep.worker"; "Imp.worker";etc.
motherOccupationConditionType	Mother's condition in labour force	"Unemployed"; "Dep.worker"; "Imp.worker";etc.
occupation	Student's occupation	"Student";"Filled Occupation", "Unknown"
iscteFirstExecutionYear	Curricular year of first admission in ISCTE-IUL	"2006/2007";"2007/2008"; etc
maritalStatusType	Marital status	"Married";"Single";etc.
nationality	Nationality	"Portuguese";"Brazilian";etc.
secondNationality	Second Nationality	"Angolan";Portuguese";etc.
entryYear	Year of registration in current degree	"2006/2007";"2007/2008"; etc
fatherLiteraryHabilitationType	Father's literary education	"Illiterate", "Higher education";etc.
motherLiteraryHabilitationType	Mother's literary education	"Illiterate", "Higher education";etc.
workingStudentAtEntry	Working statute (required at admission process)	"True;"False".
partialTimeStudentAtEntry	Partial time statute (required at admission process)	"True;"False".
specialEducationNeedsAtEntry	Special Education statute (required at admission process)	"True;"False".
scholarShipAtEntry	Scholarship (granted at admission process)	"True;"False".
dislocatedAtEntry	Dislocated statute (required at admission process)	"True;"False".
degreeCode	Code that represents each degree	"IGE";"LEI";etc.
degreeType	Degree Type	"Bachelor";"Master";etc.
degreeSchool	Degree school	"ISTA";"EG";etc
entryGrade	Entry Grade for HEI admission	"9.5" to "20"
precedentDegreeDesignation	High school degree description	"Ciências ";" Desporto"; etc.
precedentConclusionYear	High school conclusion year	"1984";"2005"; etc.
secondarySchoolType	High school's sector	"Public";"Private";etc

ingression	Ingression type	"CNA";"CM23"; etc
highSchoolDegreeType	High school degree type	"Scientific_Humanistic";"Other";etc.
iscteWasFirstChoice	Was ISCTE-IUL the university chosen in 1st place?	"True;"False".
erasmusOutgoing	Student accepted for Erasmus outgoing	"True;"False".
workingStudentEntryYear1stSem	Working student statute granted during 1st semester	"True;"False".
InternationalStudentEntryYear1stSem	International student statute granted during 1st semester	"True;"False".
partialTimeStudentEntryYear1stSem	Partial time statute granted during 1st semester	"True;"False".
fctgrantOwnerEntryYear1stSem	Fct grant granted during 1st semester	"True;"False".
classSubRepresentativeEntryYear1stSem	Class sub-representative statute granted during 1st semester	"True;"False".
classRepresentativeEntryYear1stSem	Class representative statute granted during 1st semester	"True;"False".
handicappedEntryYear1stSem	Handicapped statute granted during 1st semester	"True;"False".
pregnantOrChildrenUnder3EntryYear1stSem	Pregnant or children under 3 years old statute granted during 1st semester	"True;"False".
professionalAthleteEntryYear1stSem	Professional athlete statute granted during 1st semester	"True;"False".
sasGrantOwnerEntryYear1stSem	Social support statute (SAS) granted during 1st semester	"True;"False".
militaryEntryYear1stSem	Military statute granted during 1st semester	"True;"False".
temporaryDisabilityEntryYear1stSem	Temporary disability statute granted during 1st semester	"True;"False".
religiousEntryYear1stSem	Religious statute granted during 1st semester	"True;"False".
associativeLeaderEntryYear1stSem	Associative leader statute granted during 1st semester	"True;"False".
iscteAthleteEntryYear1stSem	ISCTE-IUL athlete statute granted during 1st semester	"True;"False".
firefighterEntryYear1stSem	Firefighter statute granted during 1st semester	"True;"False".
erasmusGuestEntryYear1stSem	Erasmus guest statute granted during 1st semester	"True;"False".
deathOfSpouseOrFamilyEntryYear1stSem	Death of spouse or family statute granted during 1st semester	"True;"False".
appearancePoliceOrMilitaryAuthorityEntryYear1stSem	Appearance in police or military authority statute granted during 1st semester	"True;"False".
monitorEntryYear1stSem	Monitor statute granted during 1st semester	"True;"False".
previousIBSStudentEntryYear1stSem	Previous IBS student statute granted during 1st semester	"True;"False".
top15IBSEntryYear1stSem	Top 15 IBS statute granted during 1st semester	"True;"False".
workingStudentEntryYear2ndSem	Working student statute granted during 1st semester	"True;"False".
InternationalStudentEntryYear2ndSem	International student statute granted during 2nd semester	"True;"False".
partialTimeStudentEntryYear2ndSem	Partial time statute granted during 2nd semester	"True;"False".
fctgrantOwnerEntryYear2ndSem	FCT grant granted during 2nd semester	"True;"False".
classSubRepresentativeEntryYear2ndSem	Class sub-representative statute granted during 2nd semester	"True;"False".
classRepresentativeEntryYear2ndSem	Class representative statute granted during 2nd semester	"True;"False".

handicappedEntryYear2ndSem	Handicapped statute granted during 2nd semester	"True;"False".
pregnantOrChildrenUnder3EntryYear2ndSem	Pregnant or children under 3 years old statutes granted during 2nd semester	"True;"False".
professionalAthleteEntryYear2ndSem	Professional athlete statute granted during 2nd semester	"True;"False".
sasGrantOwnerEntryYear2ndSem	Social support statute (SAS) granted during 2nd semester	"True;"False".
militaryEntryYear2ndSem	Military statute granted during 2nd semester	"True;"False".
temporaryDisabilityEntryYear2ndSem	Temporary disability statute granted during 2nd semester	"True;"False".
religiousEntryYear2ndSem	Religious statute granted during 2nd semester	"True;"False".
associativeLeaderEntryYear2ndSem	Associative leader statute granted during 2nd semester	"True;"False".
iscteAthleteEntryYear2ndSem	ISCTE-IUL athlete statute granted during 2nd semester	"True;"False".
firefighterEntryYear2ndSem	Firefighter statute granted during 2nd semester	"True;"False".
erasmusGuestEntryYear2ndSem	Erasmus guest statute granted during 2nd semester	"True;"False".
deathOfSpouseOrFamilyEntryYear2ndSem	Death of spouse or family statute granted during 2nd semester	"True;"False".
appearancePoliceOrMilitaryAuthorityEntryYear2ndSem	Appearance in police or military authority statute granted during 2nd semester	"True;"False".
monitorEntryYear2ndSem	Monitor statute granted during 2nd semester	"True;"False".
previousIBSStudentEntryYear2ndSem	Previous IBS student statute granted during 2nd semester	"True;"False".
top15IBSEntryYear2ndSem	Top 15 IBS statute granted during 2nd semester	"True;"False".
requestedSocialServiceEntryYear	Requested any social service during 1st year	"True;"False".
acceptedSocialServiceEntryYear	Granted any social service during 1st year	"True;"False".
requestedSStransportSupplementEntryYear	Requested transport supplement during 1st year	"True;"False".
requestedSSaccommodationSupplementEntryYear	Requested accommodation Supplement during 1st year	"True;"False".
requestedSSresidenceRequestEntryYear	Requested residence during 1st year	"True;"False".
requestedSSiscteFinantialSupportEntryYear	Requested ISCTE-IUL financial support during 1st year	"True;"False".
acceptedSStransportSupplementEntryYear	Granted transport supplement during 1st year	"True;"False".
acceptedSSaccommodationSupplementEntryYear	Granted accommodation supplement during 1st year	"True;"False".
acceptedSSresidenceRequestEntryYear	Granted residence during 1st year	"True;"False".
acceptedSSiscteFinantialSupportEntryYear	Granted ISCTE-IUL financial support during 1st year	"True;"False".
firstChoice	Was It the first choice (University+degree)?	"True";"False".
firstChoiceUniversity	Was ISCTE-IUL the first choice?	"True";"False".
firstChoiceCourse	Was the enrolled degree the first choice?	"True";"False".
orderPreference	which order of preference did the student registered?	"1";"2";"3";"4";"5";"6".
gapEntryExames	Grade average points for entry exams	"9.5" to "20"
entryAge	Student's age at entry	"16" to "74"

entryAgeRange	Student's age at entry	"[16-18]";"[19-23]"; etc.
municipality	Student’s residence municipality	"Lisboa";"Oerias"; etc.
district	Student's residence district	"Lisboa";"Setúbal”; etc.
lisbonMetropolitanArea	Does student live within Lisbon metropolitan area?	"True";"False".
studyGap	Any time gap since previous educational programme?	"True;"False".
studyGapYears	Time Gap since previous educational programme	"0";"1";"2"; etc.
ectsCreditsEntryYear1stSem	number of course passed in the entry year 1st semester	"0";"6";"12"; etc.
ectsCreditsEntryYear2ndSem	number of course passed in the entry year 2nd semester	"0";"6";"12"; etc.
averageEntryYear1stSem	average grade of the passed courses in entry year 1st semester	"0" to "20"
weightedAverageEntryYear1stSem	weighted average grade of the passed courses in entry Year 1st semester	"0" to "20"
averageGradeEntryYear2ndSem	average grade of the passed courses in entry year 2nd semester	"0" to "20"
weightedAverageEntryYear2ndSem	weighted average grade of the passed courses in entry year 2nd semester	"0" to "20"

Table C-2 – ESCO Occupations' lookup table

Portuguese occupations' classification	Aggregated final class
Managers	Managers
Professionals Technicians and associate professionals	Heads of, specialists and technicians
Clerical support workers Service and sales workers	Office, services and commerce Workers
Skilled agricultural, forestry and fishery workers Craft and related trades workers Plant and machine operators and assemblers	Industry, transports and agriculture workers
Elementary occupations	Elementary occupations

Table C-3 – Features' set selected to feed hotdeck algorithm for entryGradeHotDeck

Selected Feature
gender
yearOfBirth
maritalStatusType
nationality
areaCode
degreeCode
entryYear
secondarySchoolType
ingression
fatherLiteraryHabilitationType
motherLiteraryHabilitationType
fatherOccupationConditionType
motherOccupationConditionType

Appendix D – DM Algorithms on literature

Table D-1 – DM algorithms used on related works

Research Work	**DM algorithms**
Minaei-Bidgoli et al. (2003)	DT Bayes Network 1NN KNN Kernel Density Estimation MLP Genetic algorithm Combination of multiple classifiers
Kotsiantis & Pintelas (2005)	DT ANN Linear Regression Locally weighted linear Regression SVM
Superby et al. (2006)	DT RF ANN
Vandamme et al. (2007)	Discriminant analysis ANN DT
Romero et al. (2008a) Romero et al. (2008b)	ADLinear PolQuadraticLMS Kernel Density Estimation KNN C4.5 CART AprioriC CN2 Corcoran XCS GGP SAI MaxLogitBoost SAP AdaBoost LogitBoost GAP GP Chi NNEP RBFN RBFN Incremental RBFN Decremental GANN MLP
Kabra & Bichkar (2011)	DT
Delen (2011)	ANN DT Logistic Regression
Barber & Sharkey (2012)	Logistic Regression Naïve Bayes
Kovačić (2012)	CHAID Exhaustive CHAID QUEST

Research Work	**DM algorithms**
	CART
	Logistic Regression
Osmanbegović & Suljić (2012)	ANN
	DT
	Bayes network
Watson et al. (2013)	Linear Regression
Goker et al. (2013)	Naïve Bayes
	J48
	Bayes Network
	RBF
Mishra et al. (2014)	J48
	DT
Trstenjak & Donko (2014)	Naïve Bayes
	SVM
Slim et al. (2014)	Bayesian Belief Network (BBN)
Olama et al. (2014)	Logistic Regression
	ANN
Martínez & Gómez (2014)	Clustering Techniques
	Association Generators
	DT
Simeunović & Preradović (2014)	DT
	Logistic Regression
	ANN
Natek & Zwilling (2014)	DT
Mayilvaganan & Kalpanadevi (2014)	C4.5
	AODE
	KNN
	Naïve Bayes
Hu et al. (2014)	C4.5
	LGT
	CART
Taruna & Pandey (2014)	DT
	IBK
	NBT
	Bayes network
	Naïve Bayes
Junco & Clem (2015)	Linear Regression
Cheewaprakobkit (2015)	DT
	ANN
Stretch et al. (2015)	KNN
	RF
	AdaBoost
	CART
	SVM
	Naïve Bayes
	OLS
	SVM
	CART
	KNN
	RF
	AdaBoost
Zimmermann et al. (2015)	Linear Regression
Zhou et al. (2015)	Naïve Bayes
Ahmad et al. (2015)	Naïve Bayes
	DT
	Rule Based Methods

Research Work	**DM algorithms**
Amrieh et al. (2015)	ANN Naïve Bayes DT
You (2016)	Linear Regression
Badr et al. (2016)	CBA rule-generation algorithm
Vuttipittayamongkol (2016)	Linear Regression
Daud et al. (2017)	Bayes Network Naïve Bayes SVM C4.5 CART
Martins el al. (2017)	Logistic Regression
Asif et al. (2017)	DT Rule Based Methods 1NN Naïve Bayes ANN RF
Rahman & Islam (2017)	Naïve Bayes KNN DT ANN
Leppanen et al. (2017)	SVM
Kostopoulos et al. (2018)	ANN Naïve Bayes SVM DT Rule Based Methods
Martins et al. (2018)	RF
Fernandes et al. (2018)	Gradient Boost Machine (GBM)

Appendix E – DM_30%FilledFeatures model evaluation

Figure E-1 - ROC curve for DM_30%FilledFeatures model

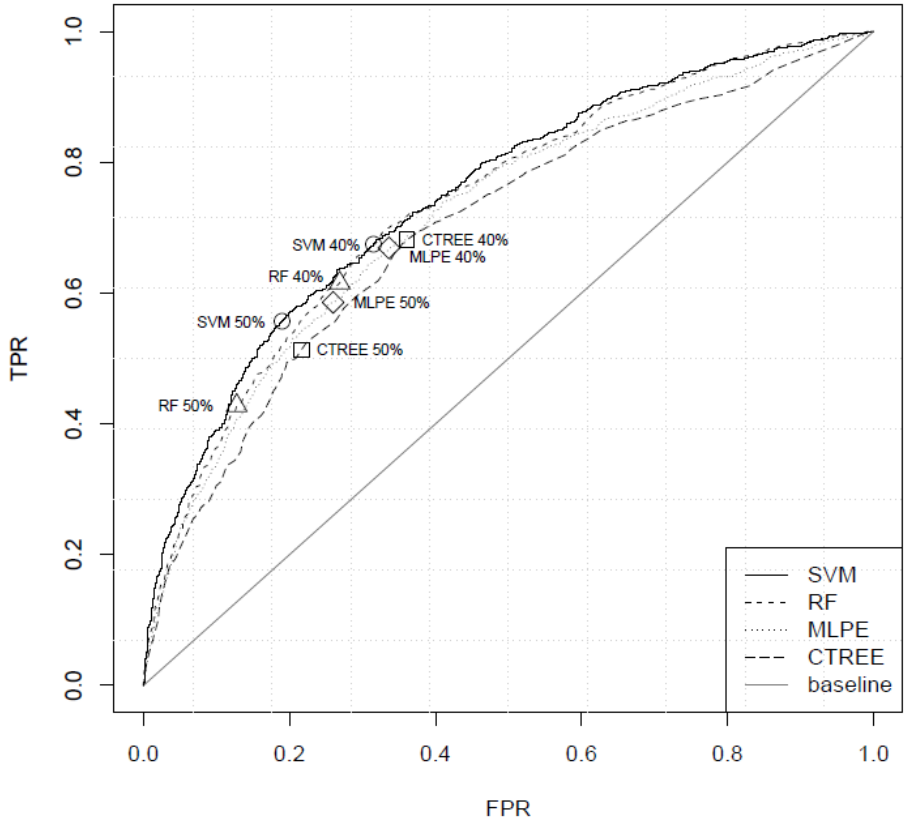


Table E-2 - Confusion Matrices for DM_30%FilledFeatures model

Threshold = 50%					
SVM		Predicted		Sensitivity	1-specificity
		Failure	Success		
Target	Failure	641	513	0.5555	0.1899
	Success	296	1263		
RF		Predicted		Sensitivity	1-specificity
		Failure	Success		
Target	Failure	494	660	0.4281	0.1276
	Success	199	1360		
MLPE		Predicted		Sensitivity	1-specificity
		Failure	Success		
Target	Failure	676	478	0.5858	0.2598
	Success	405	1154		
CTREE		Predicted		Sensitivity	1-specificity
		Failure	Success		
Target	Failure	592	562	0.5130	0.2162
	Success	337	1222		

Threshold = 40%					
SVM		Predicted		Sensitivity	1-specificity
		Failure	Success		
Target	Failure	777	377	0.6733	0.3149
	Success	491	1068		
RF		Predicted		Sensitivity	1-specificity
		Failure	Success		
Target	Failure	710	444	0.6153	0.2688
	Success	419	1140		
MLPE		Predicted		Sensitivity	1-specificity
		Failure	Success		
Target	Failure	772	382	0.6690	0.3361
	Success	524	1035		
CTREE		Predicted		Sensitivity	1-specificity
		Failure	Success		
Target	Failure	786	368	0.6811	0.3605
	Success	562	997		

Appendix F – Complete features' importance

Table E-1 - Features' importance on DM_Entrance Model

Feature	Importance
entryGradeHotdeck	11.31%
studyGapYears	9.35%
yearOfBirth	9.34%
precedentConclusionYear	8.38%
scholarshipAtEntry	7.09%
secondarySchoolType	6.77%
entryAge	5.99%
degreeSchool	5.16%
degreeCode	3.65%
gender	3.52%
fatherLiteraryHabilitationType	2.64%
ingression	2.43%
nationality	2.28%
district	2.09%
secondNationality	1.97%
motherLiteraryHabilitationType	1.89%
occupation	1.75%
entryYear	1.67%
motherOccupationConditionType	1.59%
fatherOccupationConditionType	1.55%
entryAgeRange	1.28%
dislocatedAtEntry	1.15%
lisbonMetropolitanArea	1.14%
maritalStatusType	1.10%
workingStudentAtEntry	1.02%
fatherOccupation	0.97%
iscteFirstExecutionYear	0.89%
motherOccupation	0.87%
studyGap	0.77%
specialEducationNeedsAtEntry	0.40%

Table V-2 - Features' importance on DM_EntryYear1Sem Model

Feature	Importance
weightedAverageGradeEntryYear1stSem	15.56%
ectsCreditsEntryYear1stSem	14.45%
precedentConclusionYear	7.40%
yearOfBirth	7.05%
averageGradeEntryYear1stSem	6.81%
studyGapYears	4.67%
sasGrantOwnerEntryYear1stSem	4.50%
secondarySchoolType	4.44%
entryAge	3.28%
degreeShool	3.12%
scholarShipAtEntry	2.51%

Feature	Importance
entryGradeHotdeck	1.99%
gender	1.71%
degreeCode	1.57%
iscteFirstExecutionYear	1.55%
workingStudentEntryYear1stSem	1.48%
entryYear	1.45%
secondNationality	1.31%
occupation	1.17%
ingression	1.11%
workingStudentAtEntry	1.11%
studyGap	1.10%
nationality	0.99%
fatherLiteraryHabilitationType	0.83%
fatherOccupation	0.73%
maritalStatusType	0.73%
motherOccupationConditionType	0.69%
fatherOccupationConditionType	0.67%
motherLiteraryHabilitationType	0.67%
temporaryDisabilityEntryYear1stSem	0.63%
district	0.60%
pregnantOrChildrenUnder3EntryYear1stSem	0.49%
classSubRepresentativeEntryYear1stSem	0.46%
professionalAthleteEntryYear1stSem	0.43%
lisbonMetropolitanArea	0.42%
specialEducationNeedsAtEntry	0.40%
handicappedEntryYear1stSem	0.29%
entryAgeRange	0.28%
iscteAthleteEntryYear1stSem	0.27%
classRepresentativeEntryYear1stSem	0.26%
motherOccupation	0.25%
associativeLeaderEntryYear1stSem	0.21%
internationalStudentEntryYear1stSem	0.21%
dislocatedAtEntry	0.16%

Table E-3 - Features' importance on DM_EntryYear2Sem Model

Feature	Importance
weightedAverageGradeEntryYear2ndSem	11.42%
ectsCreditsEntryYear2ndSem	10.55%
precedentConclusionYear	9.52%
yearOfBirth	8.88%
ectsCreditsEntryYear1stSem	7.30%
studyGapYears	6.73%
entryAge	5.09%
weightedAverageGradeEntryYear1stSem	3.57%
secondarySchoolType	3.42%
averageGradeEntryYear2ndSem	2.50%
degreeSchool	2.27%

Feature	Importance
scholarshipAtEntry	1.86%
acceptedSocialServicesEntryYear	1.49%
averageGradeEntryYear1stSem	1.31%
fatherLiteraryHabilitationType	1.13%
degreeCode	1.11%
ingression	1.11%
iscteFirstExecutionYear	1.10%
entryYear	1.06%
entryGradeHotdeck	0.99%
nationality	0.99%
secondNationality	0.92%
studyGap	0.90%
sasGrantOwnerEntryYear1stSem	0.90%
sasGrantOwnerEntryYear2ndSem	0.90%
workingStudentAtEntry	0.81%
workingStudentEntryYear1stSem	0.81%
workingStudentEntryYear2ndSem	0.81%
occupation	0.73%
gender	0.55%
fatherOccupation	0.54%
district	0.47%
fatherOccupationConditionType	0.43%
entryAgeRange	0.41%
motherOccupationConditionType	0.39%
motherLiteraryHabilitationType	0.37%
maritalStatusType	0.35%
temporaryDisabilityEntryYear1stSem	0.35%
motherOccupation	0.34%
requestedSSaccommodationSupplementEntryYear	0.30%
acceptedSSaccommodationSupplementEntryYear	0.30%
professionalAthleteEntryYear1stSem	0.28%
professionalAthleteEntryYear2ndSem	0.28%
acceptedSSiscteFinantialSupportEntryYear	0.27%
requestedSSresidenceRequestEntryYear	0.26%
requestedSocialServicesEntryYear	0.26%
iscteAthleteEntryYear1stSem	0.25%
iscteAthleteEntryYear2ndSem	0.25%
pregnantOrChildrenUnder3EntryYear1stSem	0.23%
associativeLeaderEntryYear1stSem	0.23%
associativeLeaderEntryYear2ndSem	0.23%
internationalStudentEntryYear1stSem	0.23%
internationalStudentEntryYear2ndSem	0.23%
classSubRepresentativeEntryYear1stSem	0.21%
classSubRepresentativeEntryYearYear2ndSem	0.21%
specialEducationNeedsAtEntry	0.21%
acceptedSStransportSupplementEntryYear	0.19%
requestedSStransportSupplementEntryYear	0.19%
classRepresentativeEntryYear1stSem	0.16%

Feature	Importance
classRepresentativeEntryYearYear2ndSem	0.16%
lisbonMetropolitanArea	0.15%
dislocatedAtEntry	0.14%
acceptedSSresidenceRequestEntryYear	0.12%
temporaryDisabilityEntryYear2ndSem	0.10%
handicappedEntryYear2ndSem	0.06%
handicappedEntryYear1stSem	0.06%
pregnantOrChildrenUnder3EntryYear2ndSem	0.04%
requestedSSiscteFinantialSupportEntryYear	0.01%

Table E-4 - Features' importance on DM_Entrance_IGE Model

Feature	Importance
entryGradeHotdeck	20.43%
iscteFirstExecutionYear	8.78%
entryYear	7.90%
degreeCode	6.07%
scholarshipAtEntry	5.96%
fatherOccupationConditionType	4.82%
secondarySchoolType	4.14%
precedentConclusionYear	3.85%
entryAgeRange	3.69%
district	3.44%
fatherLiteraryHabilitationType	2.94%
motherLiteraryHabilitationType	2.93%
occupation	2.58%
yearOfBirth	2.51%
motherOccupationConditionType	2.39%
lisbonMetropolitanArea	2.35%
ingression	2.26%
secondNationality	2.02%
fatherOccupation	1.98%
motherOccupation	1.67%
studyGapYears	1.64%
gender	1.55%
nationality	1.20%
maritalStatusType	0.78%
workingStudentAtEntry	0.59%
studyGap	0.54%
entryAge	0.51%
dislocatedAtEntry	0.32%
specialEducationNeedsAtEntry	0.17%

Appendix G – High relevance features' impact on DM_EntryYear1Sem model

Figure VI-1 – precedentConclusionYear impact on DM_EntryYear1Sem model.

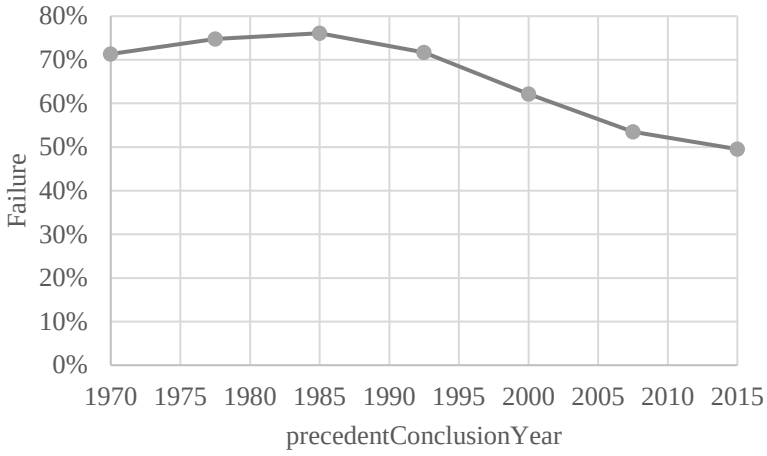


Figure F-2 – yearOfBirth impact on DM_EntryYear1Sem model.

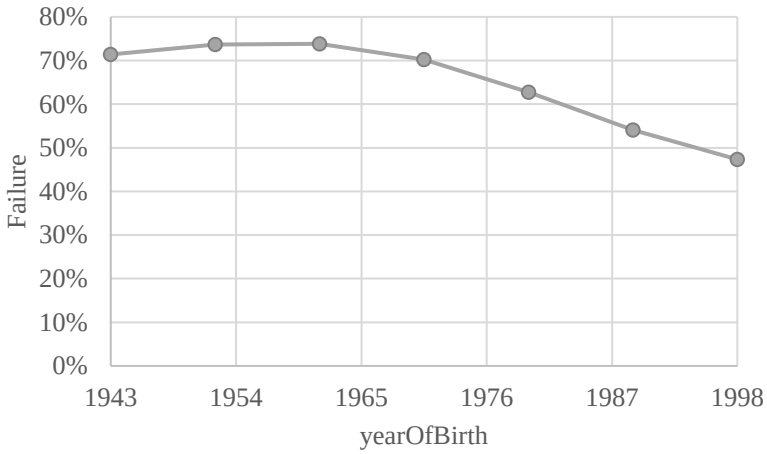


Figure F-3 – studyGapYears impact on DM_EntryYear1Sem model.

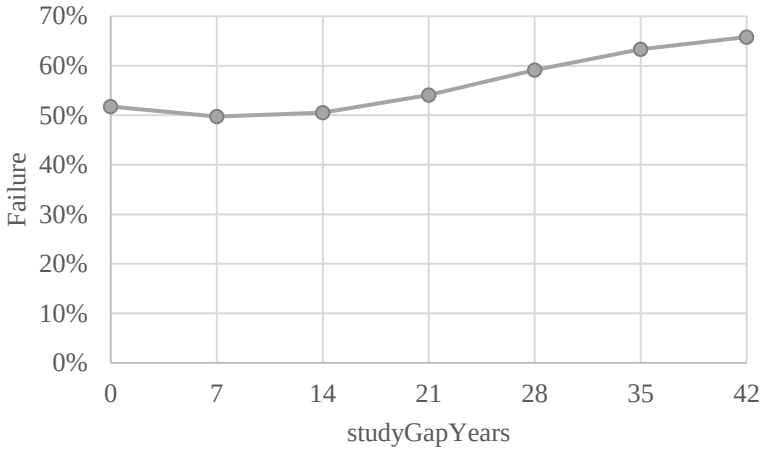
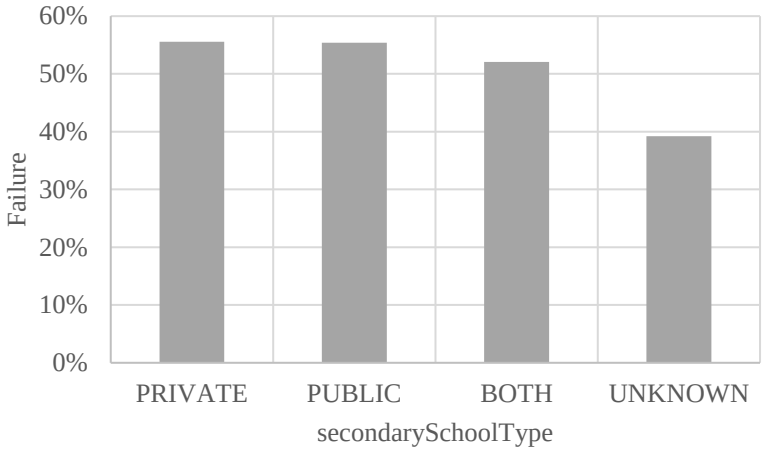


Figure F-4 – secondarySchoolType impact on DM_EntryYear1Sem model.



Appendix H – High relevance features' impact on DM_EntryYear2Sem model

Figure VII-1 – precedentConclusionYear impact on DM_EntryYear2Sem model.

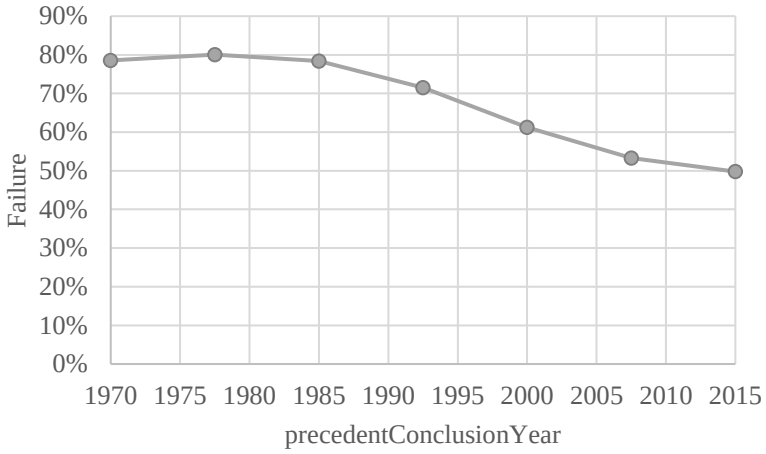


Figure G-2 – yearOfBirth impact on DM_EntryYear2Sem model.

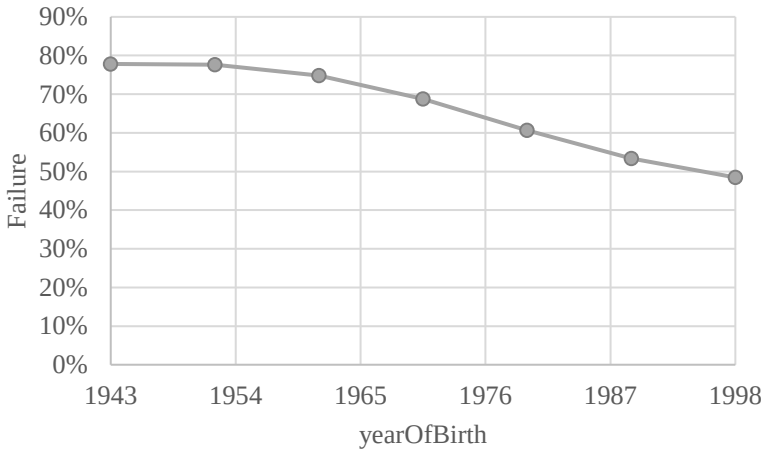


Figure G-3 – ectsCreditsEntryYear1stSem impact on DM_EntryYear2Sem model.

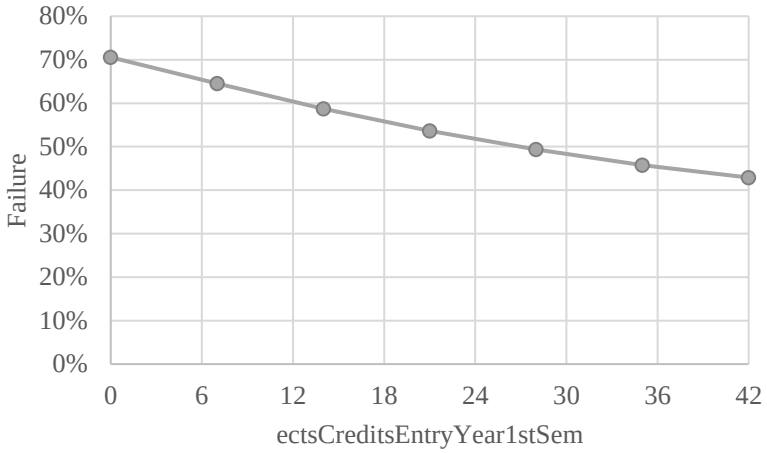


Figure G-4 – studyGapYears impact on DM_EntryYear2Sem model.

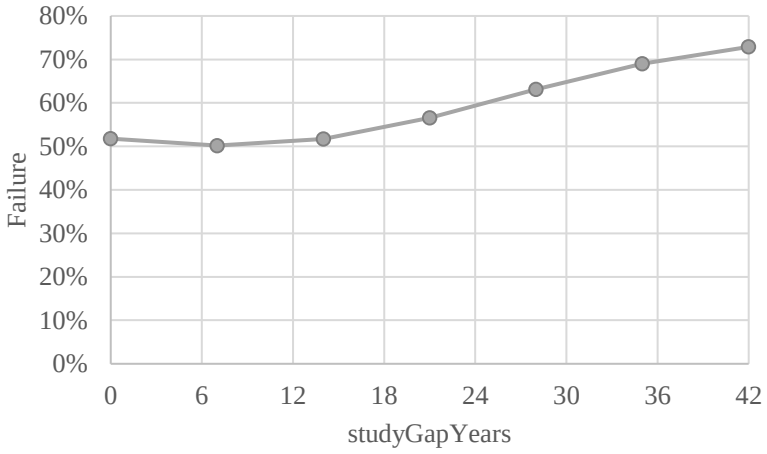


Figure G-5 – entryAge impact on DM_EntryYear2Sem model.

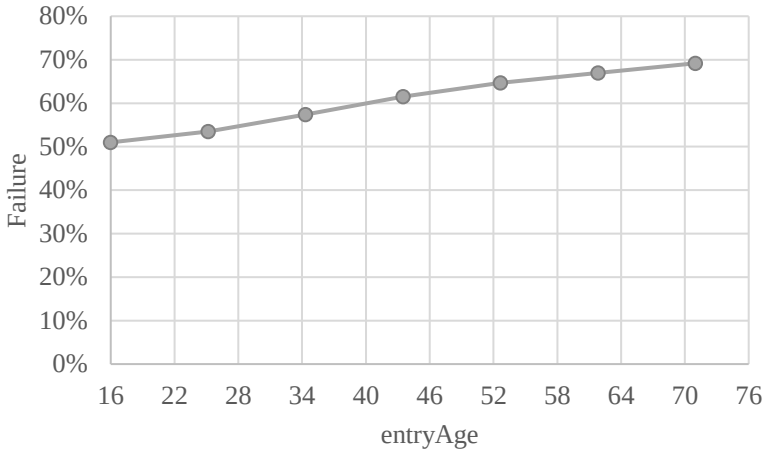
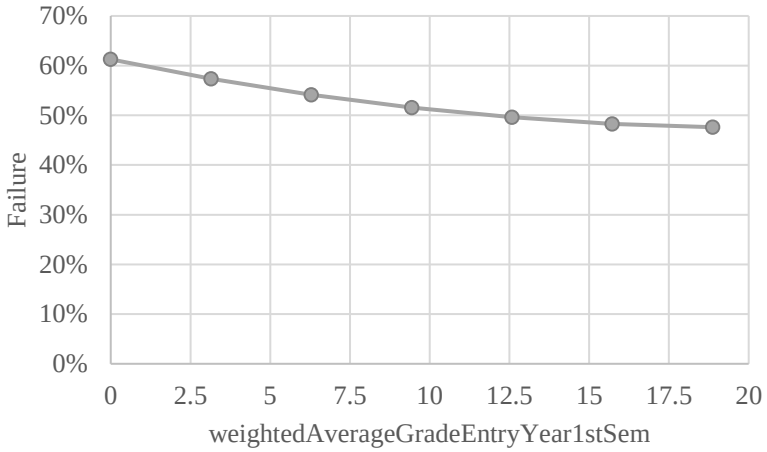


Figure G-5 – weightedAverageGradeEntryYear1stSem impact on DM_EntryYear2Sem model.



Appendix I – High relevance features' impact on DM_Entrance_IGE model

Figure VIII-1 – precedentConclusionYear impact on DM_Entrance_IGE model.

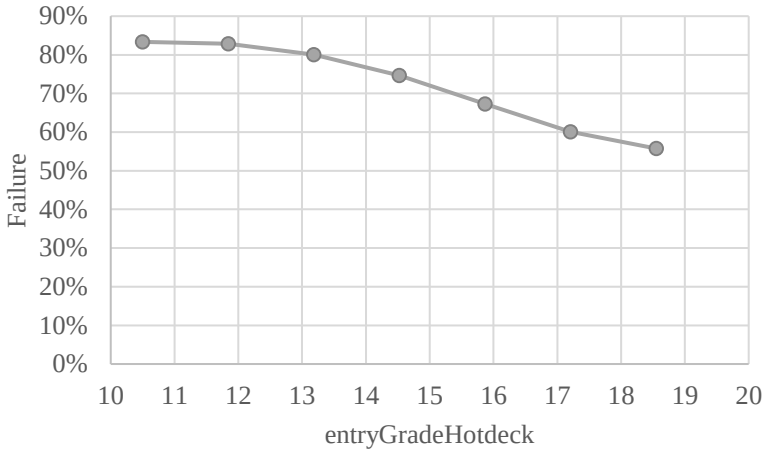


Figure H-2 – iscteFirstExecutionYearimpact on DM_Entrance_IGE model.

Figure H-3 – entryYear impact on DM_Entrance_IGE model.

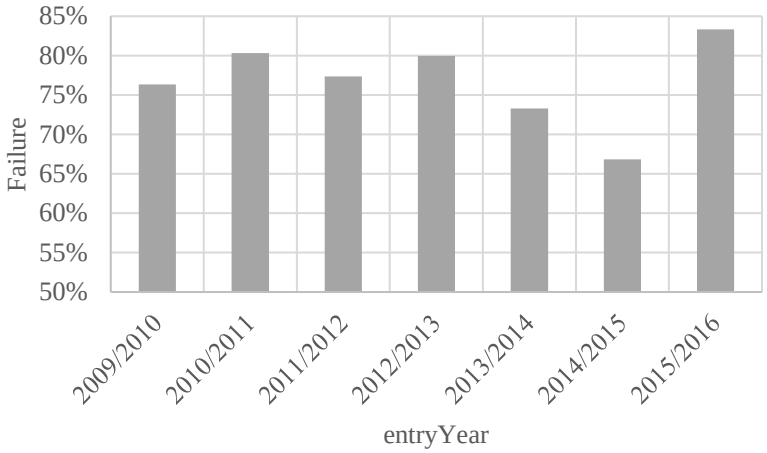


Figure H-4 – degreeCode impact on DM_Entrance_IGE model.

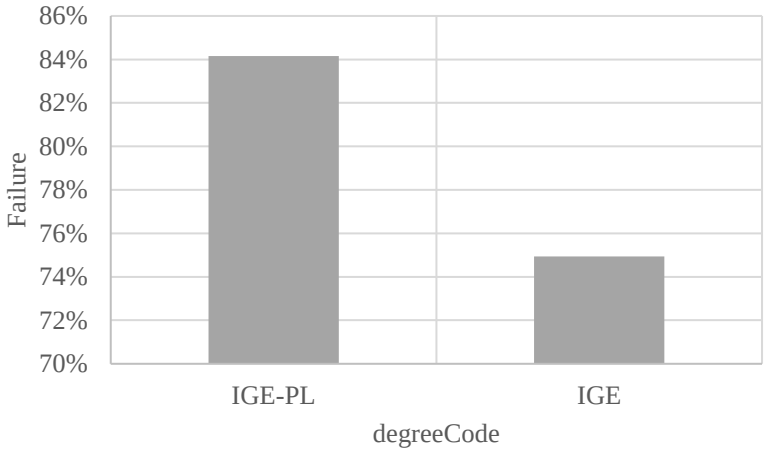


Figure H-5 – scholarshipAtEntry impact on DM_Entrance_IGE model.

Figure H-6 – fatherProfessionConditionType impact on DM_Entrance_IGE model.

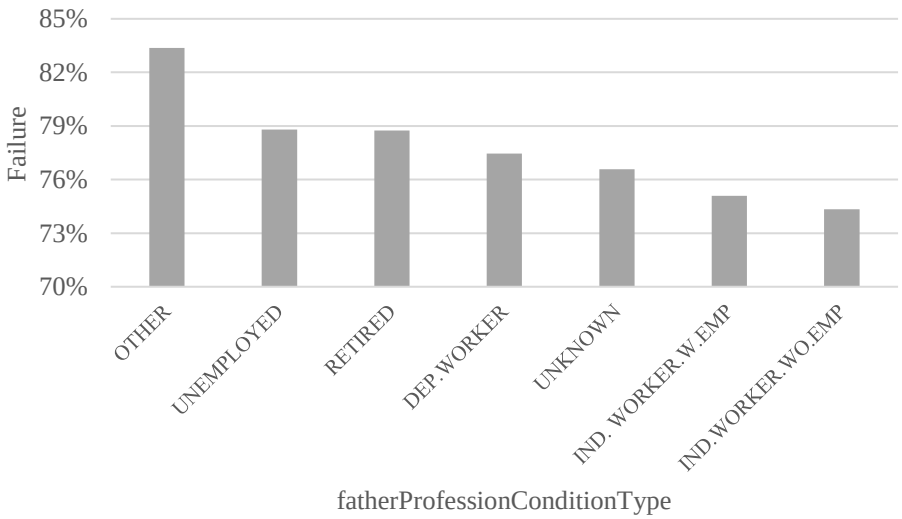


Figure H-7 – secondarySchoolType impact on DM_Entrance_IGE model.

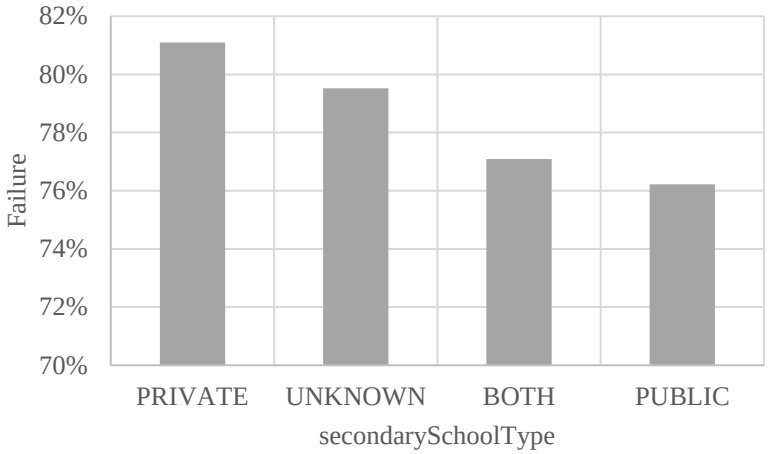


Figure H-8 – precedentConclusionYear impact on DM_Entrance_IGE model.