A Predictive Analysis of Academic Success at Universidade do Porto

Rafael Antonio Belokurows

Mestrado em Ciência de Dados

Departamento de Ciência de Computadores

2021

Orientador

Carlos Manuel Milheiro de Oliveira Pinto Soares, Professor Associado da Faculdade de Engenharia da Universidade do Porto

Coorientador

Fernando Manuel Augusto da Silva, Professor Catedrático da Faculdade de Ciências da Universidade do Porto

U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Todas as correções determinadas
pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, ______/______/________

MSc

Abstract

There has never been as much investment in higher education around the world as it has been done in recent years and, consequently, the pressure on students to achieve good performance and an academic degree it is on an all-time high. Regardless, failure and dropout rates in higher education still reach up to 30% at most universities, especially in the first academic year.

Through Data Mining and Machine Learning techniques, and using data from enrolments of 8 years and 50000 students, this study seeks to obtain an understanding of the factors that lead to academic failure at the University of Porto in a complete way and for the entire training offer of the 1st cycle - Degree and the institution's Integrated Master.

First, Feature Selection methods are used to identify a set with the most relevant attributes to predict academic success. Subsequently, this attribute subset is compared with various combinations of theoretical classes of variables - sociodemographic, related to admission, enrolment and the student's academic history. Then, Random Forest Classification algorithms are developed for each set of attributes to determine, based on available data, what is the best combination of attributes to predict academic success for the entire offering of Undergraduate and Integrated Master courses at the Universidade do Porto.

It was determined that the most efficient set of attributes to predict academic failure contains all 35 variables available in this study, which performed better than any other set or individual class of attributes. In addition to the general aggregated results, the predictive models are evaluated from two perspectives: a localized evaluation, looking at each Curricular Unit in itself and a more complete and general evaluation, comparing the results of different curricular units, programmes and scientific areas.

Results show that first-year curricular units and curricular units where there is a greater balance of classes of the target variable had the best possible predictive effectiveness. In addition, when studied separately, the most important attributes for predicting academic success are the ones that represent the academic history of the student. The top five attributes in predictive power considering the entirety of courses were: the Number of delayed Courses, Percentage of programme completion, Number of Courses already enrolled, Number of delayed years and No. of ECTS Credits to which the student has committed for the current semester.

Keywords: Data Mining, Educational Data Mining, Learning Analytics, Student Performance, Student Achievement, Classification, Random Forest

Resumo

Nunca se investiu tanto no ensino superior por todo o mundo como vem sendo feito nos últimos anos e, consequentemente, nunca foi tão grande a pressão sobre os alunos para alcançar bons desempenhos e um diploma acadêmico. Atualmente, as taxas de insucesso e abandono no ensino superior alcançam até 30% na maioria das universidades, especialmente no primeiro ano curricular. Através de técnicas de *Machine Learning*, e utilizando dados de inscrições de 8 anos e mais de 50000 estudantes, este estudo busca obter uma compreensão dos fatores que levam ao insucesso acadêmico na Universidade do Porto para toda a oferta formativa do $1^{\underline{o}}$ ciclo - Licenciatura e Mestrado Integrado - da instituição.

Primeiro, são utilizados métodos de *Feature Selection* para identificar um conjunto com os atributos mais relevantes para prever o sucesso acadêmico. Em sequência, é feita uma comparação deste conjunto com diversas combinações de classes teóricas de variáveis - sociodemográficas, relacionadas à admissão, à inscrição e ao histórico do estudante. São então desenvolvidos algoritmos de Classificação *Random Forest* para cada um deste conjunto de variáveis para determinar, qual o melhor conjunto de variáveis para prever o sucesso acadêmico para toda a oferta de cursos de Licenciatura e Mestrado Integrado da Universidade do Porto.

Determinou-se que o conjunto de atributos mais eficiente para prever o insucesso acadêmico contém todas as 35 variáveis disponíveis neste estudo, que obteve melhor desempenho do que qualquer outro conjunto ou classe individual de atributos. Além dos resultados agregados gerais, os modelos preditivos são avaliados sob duas óticas: avaliação local, olhando para cada Unidade Curricular e avaliação geral, comparando os resultados de diferentes Unidades curriculares, cursos e áreas científicas.

Os resultados mostram uma maior eficácia preditiva nas unidades curriculares do primeiro ano e naquelas onde há um maior equilíbrio de classes da variável objetivo. Quando estudados isoladamente, os atributos mais importantes para a previsão do sucesso acadêmico são os pertencentes ao grupo de atributos do histórico acadêmico, com destaque para o Número de UCs em atraso, Percentual de completude do curso, Número de Unidades Curriculares já cursadas, Número de anos em atraso e Créditos ECTS aos quais o aluno se comprometeu para o semestre atual, que tiveram o maior poder preditivo. Foram identificadas também diferenças bastante significativas no desempenho dos modelos entre várias unidades curriculares, cursos e áreas científicas.

Palavras-chave: Data Mining, Educational Data Mining, Learning Analytics, Sucesso Acadêmico, Desempenho Acadêmico, Classificação, Random Forest

Acknowledgements

The completion of this study would not be possible without the valuable feedback and inputs provided by Prof. Carlos Soares, my advisor and mentor for this project. I would also like to thank Prof. Fernando Silva, who advised in many ways, clearing a lot of obstacles and providing much needed push and guidance. Also, Pedro Strecht, who was a friend, a colleague and a technical advisor that supported me and the work done in every way he could. Thank you all for making this a better study, it was an honor and a pleasure working with each of you.

I would also like to thank the institutions that supported me with data, environment and the technical and business instructions necessary for carrying out this work, Universidade do Porto and INESC TEC. I learned a lot and hopefully this work gives something back to the institutions and the educational community.

While the academic and technical support is important, this study would not exist - and I would not be here - if not for the most important persons in my life. My partner Heloisa, the love of my life, who is always there for me every day, giving me the support, the love, and the push I need. Thanks for making me feel appreciated, loved, cared and for keeping me motivated. Lastly, my mom and my sister, who are far in distance, but always close in my mind and my heart, thanks for the support, this one is also for you and for dad, I think he would be proud. Thank you all for supporting me and for making me a better person.

Contents

List of Tables

List of Figures

Acronyms

CART	Classification and Regression Trees
CGPA	Cumulative Grade Point Average
CSV	Comma-separated values
ECTS	European Credit Transfer and Accumulation System
EDM	Educational Data Mining
ETL	Extraction, Transformation and Loading
FCUP	Faculdade de Ciências da Universidade do Porto
FDUP	Faculdade de Direito da Universidade do Porto
FFUP	Faculdade de Farmácia da Universidade do Porto
FMUP	Faculdade de Farmácia da Universidade do Porto
FN	False Negatives
FP	False Positives
FPCEUP	Faculdade de Psicologia e de Ciências da Educação
FSEL	Best Variables according to Feature Selection
GDP	Gross Domestic Product
HEI	Higher Education Institutions
ICBAS	Instituto de Ciências Biomédicas Abel Salazar
IQR	Inter Quartile Range
LA	Learning Analytics
LLM	Logit Leaf Model
LMS	Learning Management System
NLP	Natural Language Processing
OECD	Organisation for Economic Co-operation and Development
RFE	Recursive Feature Elimination
ROC	Receiver Operating Characteristic
STEM	Science, Technology, Engineering and Mathematics
TN	True Negatives
TP	True Positives
UNECE	United Nations Economic Commission for Europe

Chapter 1

Introduction

Access to education is at the forefront of most Federal Bills of Law as a basic right for every citizen. The current Portuguese Constitution states it abundantly clear: “Everyone should be granted access to education and culture”, in a free translation. These provisions are also clear-cut in the Brazilian and French Constitutions, and many others. It could be a reason for debate whether this should be explicitly recognized or not as a basic right by Law, but it’s perfectly clear by now all over the world that quality education is a necessity as big as one could have.

That being the case, the Education sector is one of the biggest beneficiaries of investments and public policies aimed at improving the quality of service and the quality of life for the whole population. On average, countries currently spend 4.5 percent of their Gross Domestic Product (GDP) and 14.5 percent of their entire public expenditure in maintaining and improving Education, and a good chunk of this investment goes in Higher Education policies.[97]

Of the total expenditure in education, according to reports by the European Commission with recent data, all but two member-countries of the European Union spend at least 20 percent of their education budget on tertiary education, with some spending as much as 37 percent.[76] In Portugal, even with the relatively low percentage of the public expenditure of 3.2 percent on education, it means that more than 1 billion euros are allocated each year for activities such as Research funding, financial aid for students and households in need, personnel and others, with the intention of furthering the level of education in the country.

When it comes to the student, the investment is perfectly justifiable. Several studies have shown a high association between educational attainment and obtaining better jobs, greater financial success and, in general terms, a better quality of life. In practically all Organisation for Economic Co-operation and Development (OECD) member-countries, the proportion of people who self-report as “being satisfied with their lives” is significantly higher among those with a higher education degree than among those who have completed only secondary education.[60] The same report shows an even more impressive variation for Portugal, with 90% of people who have attained higher education degrees reporting as “being satisfied with their lives”, in contrast with only 59% for people who had only completed high-school level education.

Unemployment rates also are almost always larger for people who have not completed tertiary education, 8 percent to 6 percent for the entire European Union. Completing college education also reflects higher salaries, on average. Aside from improving the financial condition of the student and that of those around them, higher education institutions also contribute to the preparation of citizens with a critical and free spirit, to live and work in a society that reflects criticism and freedom of opinion.[40]

Lowering completion rates and making sure more people obtain a college degree were among goals for the year 2020 by the European Commission[77]. In 2007, then US president Barack Obama started a program called American Graduation Initiative whose purpose was to improve graduation rates from 39% to 60% in the United States in a few years, investing 12 billion dollars in this endeavour[18]. And while those goals were not fully met[71], it shows how devoted governments and rule-makers are to this subject. Such programs and objectives are far-reaching and crucial to improve conditions for students all over the world.

The greatest adversities faced by Higher Education Institutions (HEI) on their way to achieve those goals are student dropout and student insuccess. Although fairly different from one another in definition, causes and consequences, both phenomena compromise efforts, time and resources spent by the government, private investors and members of the academia in general, and are cause of not only budgetary but humanitarian concern.

The dropout epidemic, as some researchers have been calling it[36], is all the more problematic when the institution is publicly funded, as it is currently the preponderant business model in most countries. The public funding model makes up for more than half of the students enrolled in tertiary education worldwide, despite the fact that, in some countries, more than half of universities are privately funded. In Portugal, this number is even more pronounced, with 65% of students currently in the tertiary level of education enrolled in publicly funded schools, according to recent data.

Abandoning a college programme before its completion, independently of its duration, is also dreadful for the students themselves. A student who leaves university loses their own time and resources, which they could have invested in something else that would be of most interest or closely related to their skills and profile. The same, but with a narrower range of consequences can be also said in the case of academic insuccess. Even though failing a class or a period of classes is definitely a less traumatic experience, it can also take a psychological or even physical toll on the student, especially when there are financial issues in the picture.

Also, the issue with respect to failing a class is twofold, as it can be considered independently or studied as a means of reaching a conclusion in a more comprehensive study, such as one for student abandonment. In the latter case, a student who fails in a handful of Course Units obviously feels frustrated and if this does not change, it can lead to early dropout.[4]

Most initiatives in place today to help students find their footing in higher education have one thing in common: they start with substantial work to identify relevant factors for student insuccess. There are a multitude of studies in the subjects of academic performance and dropout prediction, and being such multifaceted and broad subjects, those are issues that have already been thoroughly investigated by many viewpoints.

Studies in Brazil indicate that college students in a Brazilian HEI with lower economic statuses, history of poor academic performance and poor participation in academic activities were among the ones at risk of failing or dropping out.[45] With a different scope, another study, this time in the US observed that low motivation, economic constraints in the student's childhood and lack of mentorship in college all had significant statistical correlation with student dropout[56]. Such contrasting views show just how extensive the subject is and how the environment, student background and support by the university are just some of the determining factors to identify students at-risk.

Regardless of the approach taken and the use or not of data-oriented methods, a common goal of all studies in these areas is to better inform higher education decision-makers, to enable more effective policies to be created or better enforced to keep students enrolled and effectively receiving quality training.

One challenge ever present in this type of studies is the huge amount of data available. Most authors choose to limit their scope to only a small portion of the programmes offered by an educational institution. Namoun and Alshanqiti, in their 2021 survey of academic performance studies discovered that courses related to Natural Sciences, mainly those related to science, technology, engineering and mathematics were the target of more than 50% of the studies regarding this subject[93].

While this is a good way to present results and conclusions quickly and succinctly in projects that have a limited scope of time and budget, it can jeopardize the applicability of such studies in a real-world school environment. Worse yet, it can lead education managers to make institution-wise conclusions based on facts that were only observed in a small sample of the student population of an HEI. For that reason, it is paramount for a higher education institution offering courses in several domains to extend this type of research to its entire course offering. Only then can it develop effective programs that reach its entire roll of students.

The consequences of academic evasion and repeated academic insuccess can be devastating to all people involved in the matter, therefore, those issues call for solutions. Although it is not in the scope of this work to suggest actions that can be taken in order to prevent or deal with the issue of student insuccess, the scientific literature shows several precedents in which the prediction and correct identification of students at risk of failing or withdrawing from their courses enables and facilitates actions.

Ultimately, building upon some of the most influential and latest studies in the matter, the purpose of this work is to provide, to the best possible extent, a fresh analytical view of the factors that most strongly influence student insuccess. In addition, it is intended to explain, based on data provided by the University of Porto, the relationship between sociodemographic, admission-related and journey-related attributes and both types of events, as well as the interaction between each category of attributes.

Another important matter to look into is verifying the advance with which it is possible to determine with good statistical confidence that a student is at risk of failing a subject or dropping out of college. Only after obtaining a better understanding of the reasons for failure should it be possible to identify students in difficulty early and leverage timely positive and efficient policies in order to remedy the situation to the best extent possible.

1.1 Research Goals

One of the goals of this work is identifying patterns to help decision-making by educational managers in order to investigate academic insuccess of students in undergraduate courses using data mining techniques such as feature selection from available information in academic databases, for a better computational representation of the issue of student insuccess.

In big undertakings such as this, it is easy to lose focus and veer off course, so one key decision made together with University personnel was to limit the scope of the study to only First Degrees and Integrated Master's courses. Those courses are where the students are most prone to failing and dropping out, are the ones with the most young students - 18 to 24 years old - and are always at the forefront of the University's and country's current measures for counteracting insuccess in higher education environments. Another benefit of this structure is that those were also the most constant types of programmes across several factors, such as approval rate, degree attainment and duration.

The limitation of scope also allows for a clearer view of attributes and patterns, since combining courses and programs of various types, in addition to the other inherent complexity factors, could be another trend-smoothing element, causing the study not to present any clear conclusions.

It is intended first to carry out a stage of Feature Selection to identify the most relevant attributes according to scientific criterion. The second part will demonstrate how we use these attributes in the training and evaluation of a Machine Learning Classification system that allows predicting in advance and with reasonable accuracy which students are at risk of failure.

The Machine Learning task, performance prediction, consists of a binary classification task with a target variable indicating whether the student has been approved or not. In order to achieve simple and straightforward evaluation despite the complexity and size of the study, some goals were set in relation to the performance of the models, along with managers from the University of Porto. When it comes to anticipating an event that is important and possibly as harmful as failing or dropping out of higher education studies, it is important to pay special attention to the positive class of the binary variable, which represent the occurrence of these two events.

Accordingly, two of the most important measures used in the evaluation of models where efficiency in predicting instances of the positive class is the priority: Precision and Recall. On this study, with methodologies and knowledge available, we will seek to achieve the best Precision and Recall possible, meaning the prediction of students who will fail - the positive class is more important for the evaluation of the results obtained.

1.2 Research Questions

In addition to the more general objectives of identifying relevant factors and predicting academic insuccess with data made available by Universidade do Porto, to reach more informative and statistically significant conclusions, some more specific and quantifiable work objectives were defined. With this study, we intend to answer the following Research Questions:

1. What is the best combination of classes of attributes to predict academic insuccess?
2. How accurately can we identify what students are going to fail at a course at the beginning of the semester?
3. Are there any differences in the quality of predictions based on the Curricular Year?
4. Are there any differences in the quality of predictions based on scientific area, programme or school?
5. What are the most important individual attributes for predicting cases of academic insuccess in higher education institutions?

1.3 Innovation

Studies of student performance prediction usually have varied and broad scopes. While some seek to identify the most relevant factors, others focus on predicting students at risk of failure or to suggest measures to improve student performance and/or retention.

This study aims to determine what are the most important classes of attributes to predict student performance in an institution-wide study using information acquired from *Sigarra*, the university's Student Information System. Unlike other recent studies, ours intend to give a overview of the entire situation of academic success at the Universidade do Porto using data from all courses instead of a limited sample. Although we do not intend do dive deep into a pool of almost 2000 courses, this study could certainly bring some momentum and propel subsequent studies on the matter, especially with the framework and the data in place to do it.

1.4 Structure of the study

Chapter 2: Technical Background: Description and rationale behind Data Mining, Machine Learning techniques, statistic modelling and model evaluation measures to serve as a refresher for those already initiated in the subject and a quick overview for those who are not.

Chapter 3: Related Work: We present a review of the latest advancements available in current scientific literature in relation to the methods and objectives of the work. First, we discuss the importance of Learning Analytics and Educational Data Mining and what each of these generally studies. Later, we dive into some of the most influential recent studies of performance prediction, especially in relation to the attributes deemed as most important for predicting the occurrence of both events.

Chapter 4: Results: We present an overview of the Case Study. Data acquisition, data preprocessing and data preparation activities are briefly discussed and we elaborate on the findings and results obtained by the implementation of this work. We also evaluate the performance of models under several points of view, analysing the correctness of predictions and computation time of each combination of variables. Moreover, final remarks, limitations and possibilities for further studies are discussed.

Chapter 2

Background

2.1 Data Mining

There has been a enormous growth recently in data generated by everyone and collected by companies, providers, the government and other agents. An issue with this exponential growth in data generation is that without the appropriate tools and methods, the amount of data we generate is much greater than our processing and analysis capacity. Physical resources for storage and professional resources for analysing and extracting information from data are limited - and increasingly expensive. Recent studies show that 90% of data generated by the entire history of humanity was generated in the last two years2.1.

It is no wonder that professions such as Data Scientist, Data Analyst and other related ones are often listed as some of the sexiest professions on the market, with high salaries and benefits and that courses in these areas are some of the most sought after in the current industry. technology and STEM degrees. According to studies published by United Nations Economic Commission for Europe (UNECE), on behalf of the international statistical community, 90% of all the data in the world has been produced over the last 2 years.[61]

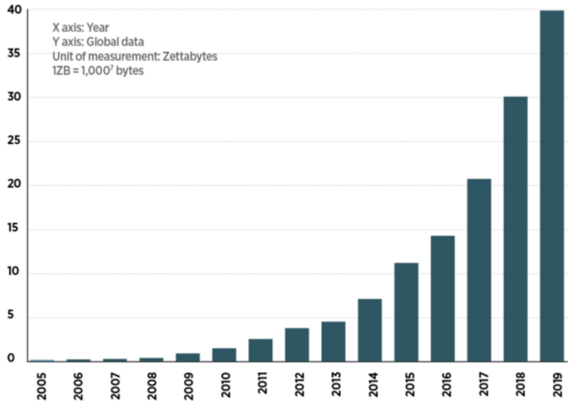


Figure 2.1: Data Growth in the last few years, according to UNECE, reproduced from Javed, 2018[61]

There is no clear-cut definition to explain the term "Data Mining", however, it is certain that it does not represent the whole picture when it comes to the importance of activities generally related to it. Tan et al., 2014[35], considers Data Mining an integral part of knowledge discovery in databases - KDD, which is the overall process of converting raw data into useful information, that certainly gives us more information and a better starting point to understand this area of science.

In the same manner, there is also no consensus on the scope of the discipline it covers. Although some people may refer to Data Mining as something more restricted, related only to the more concrete or physical activities of obtaining and preparing data effectively similar to a *mining* activity, perhaps the best definition was given by *Aggarwal* in his 2014 book says that "Data mining is the study of collecting, cleaning, processing, analysing, and gaining useful insights from data". a The concept of *mining* certainly plays a part when defining Data Mining and its activities, since some of the activities in data collection and analysis are somewhat analogous to a miner finding gold in the rivers or between the rocks. In the same way, the person responsible for analysing and extracting information from data sometimes has to pick his/her way through thousands or millions of records to discover patterns and to find the gold nugget, the piece of information to support decision-making, especially when it comes to the commercial sector.

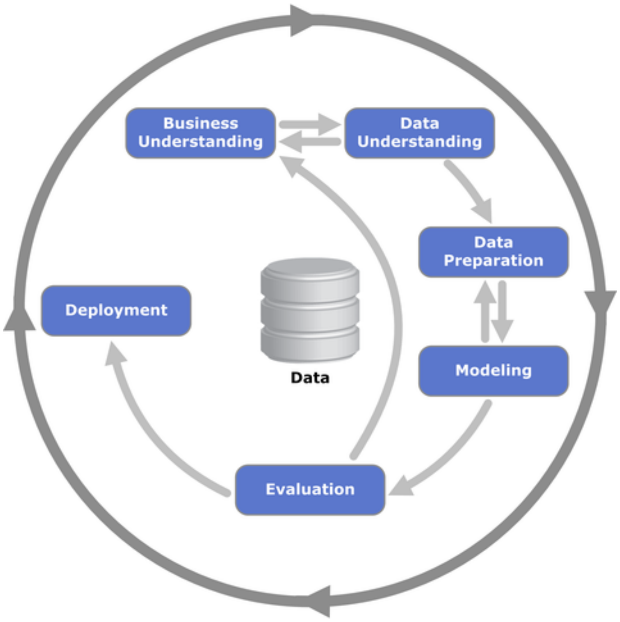


Figure 2.2: CRISP-DM framework, reproduced from Malik et al., 2018[62]

This is in line with what the CRISP-DM framework establishes. The CRISP-DM framework is a data mining process model that is widely accepted by the industry, and describes high-level approaches commonly used when carrying out a Data Mining project. Figure 2.2 shows an overview of this framework.

According to Han(2014)[22] "we don't live in the information age, but in the data age" and this is a major point to be raised. Any device, even those that appear not to be particularly smart by today's definitions, is logging and sending information about its usage to someone. However, generating this data is only the first step in making a decision or coming up with some relevant information.

R.L. Ackoff proposed in 1989 a model widely recognized and still used today in the field of Knowledge Management called DIKW, which organizes the pillars of knowledge management in a clear hierarchy.[9] As shown in figure 2.3, this hierarchy distinguishes notably the obtaining of raw data - lower and more comprehensive level, from more organized information, something actionable like knowledge or something more structured like Wisdom.

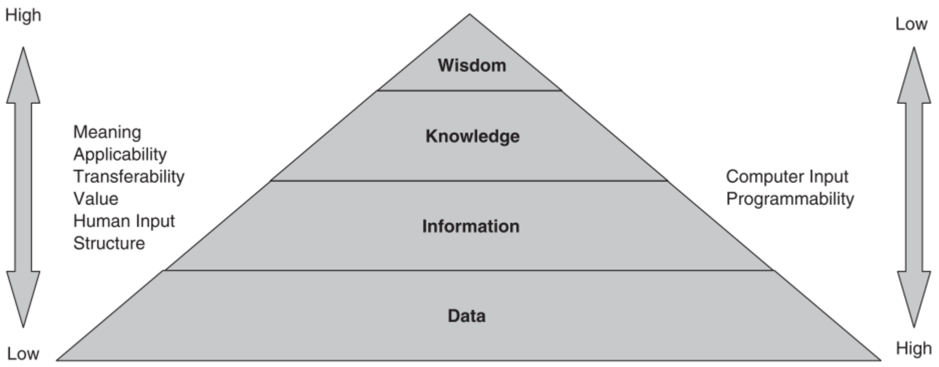


Figure 2.3: DIKW hierarchy of knowledge obtaining, reproduced from Rowley, 2006[9]

Although Data Mining is a comprehensive concept and one that has no fixed definition, there are some widely accepted frameworks in the industry that define its scope and that outline necessary and important activities to be carried out for most projects that could and should involve some type of Data Mining.

One of the most accepted frameworks in this regard is CRISP-DM, proposed in 2000 and which defines six major stages for a Data Mining endeavour, namely: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment. Each stage is subdivided into several smaller activities. Although not all stages are definitive, said framework is an excellent starting point for planning any project that has the goal of obtaining knowledge from data.

The stages and their main objectives:

Business Understanding: Understand business needs and set clear goals for the project.

Data Understanding: Collect, describe and carry out an initial exploration of the data.

Data Preparation: Filter, transform, and integrate data in order to prepare it for the next stage.

Modeling: Select one or more models from the various options available, build and test the model.

Evaluation: Select one or more models from the various options available, build and test the model.

Deployment: Put the model into production and review project errors and successes.

The process is cyclical and no phase is final or definitive. It is very beneficial and recommended to carry out periodic reviews of the process to feed back to the previous phases. With the feedback of the evaluation of a model or the data obtained in the final review of the process, it is possible to take measures to better treat the data and consequently have a better performance at the end of the project.

This framework is also very flexible to accommodate needs from every type of technology project. It is possible and advisable to adjust it to specific needs of each endeavour. One example is the Deployment stage, which is usually replaced by a guided knowledge extraction process in academic research settings.[89]

Some areas that take great benefit from Data Mining technologies are:

- Banks: identifying patterns to assist in managing relationship with customers.
- Credit Card: identifying market segments, identifying turnover patterns.
- Medicine: making more accurate diagnoses.
- Commerce and Marketing: Improving disposition of products on the shelves through analysis of consumer patterns.
- Education: Identifying patterns for common actions, clustering students and identifying factors influencing evasion or insuccess.

2.2 Machine Learning

Machine Learning is an area within the study of Artificial Intelligence that provides a set of algorithms to analyse large datasets to learn and discover patterns and make predictions about new datasets with the least human intervention possible.

Although the goals of Data Mining and Machine Learning may somewhat overlap and the two terms are sometimes used interchangeably, there are marked differences between the two areas. Both are analytical processes, related to the extraction of information for a specific purpose: learning from data and with the goal of improving decision-making processes. Moreover, both use algorithmic approaches to sift through data, different tools, and applications and require large amounts of data to be successful[35].

2.2.1 Common Machine Learning Tasks

Supervised Learning is a broader spectrum of Machine Learning techniques that uses labelled data to predict an outcome. As input data is fed into the model, it adjusts its weights through a reinforcement learning process, which ensures that the model has been fitted appropriately. This category is called Supervised because an example data set is used to learn the structure of the groups, just as a teacher supervises his or her students towards a specific goal.

Unsupervised Learning, on the other hand, is used to detect how data are organized and to produce what *descriptive models*. An Unsupervised Learning algorithm receives unlabelled examples and learns to detect hidden patterns based on the features of the data. Some of the Unsupervised Learning algorithms most commonly used in Data Mining are Clustering algorithms such as K-Means, Association Rule Mining algorithms such as Apriori ou Ecat and Principal Components Analysis.

Clustering is a type of Unsupervised Learning method, that is, a search is made for a structure in the data without necessarily comparing the results with pre-existing labels such as a Supervised Learning method would perform.

The most traditional and studied branch within the Machine Learning area is that of supervised learning methods, and the two most important tasks within this group are the tasks of Classification and Regression.[37] With a Regression task, the purpose of the algorithm is to develop a relationship between outputs and inputs using a continuous function to understand how the outputs are different for a given input. Some of the possible uses of a regression method are: (1) predicting the sales of a particular product (2) predicting the value of a property (3) calculating the life expectancy of a country or region or (4) estimating a patient's blood pressure, among thousands of possibilities. In addition to these traditional methods, there are several other types, as identified in several studies, such as Baker, Isotani and Carvalho, 2011[17].

Other Machine Learning methods present in this and other classifications are *Relationship Mining*, which can be classified in four main groups. In *Association Rule Mining*, the algorithm aims to observe frequently occurring patterns or associations between two or more variables in the data, looking for "if something then something" rules.

In *Correlation Mining*, the goal is to find correlation between positive or negative correlations between variables. Sequential Pattern Mining tries to find temporal sequences of events to determine what path of student behaviours leads to an eventual learning event of interest. *Causal Mining* aims to determine whether one event is causally related with another.

In the area of *Data distillation to facilitate human decisions*, the main objective is to present data in a more readable and visual way to facilitate human understanding and thus support important decisions based on the data. In addition, the data sometimes can be analysed, derived and transformed for an ulterior purpose such as identification, classification or human labelling. This way, data is depicted in a manner that enables a human to quickly place or classify its features.

Lastly, *Discovery with models* is a twofold process where a model of a construct is developed and validated using Machine Learning techniques and subsequently applied to data and used as a component in another analysis. One can even argue that most data-oriented studies internally perform *Discovery with models* to some extent through various activities commonly performed in studies of this type, such as: identification of outlier values through model predictions, selecting the variables that best explain the predicate of an analysis or analysing correlation between attributes in cross-sectional studies.[26]

2.2.1.1 Classification

A classification task is performed by building and training a model to learn the structure of a data sets partitioned into groups, also called labels or classes. The model is used to estimate the classes based on previously unseen data. A dataset with features or attributes is used to learn and discover patterns in the data.

Input
Output
Attribute set (x)
Classification
Class label (y)

Figure 2.4: Classification as the task of mapping an input attribute set x into its class label y, from Tan, 2014 [35]

Classification techniques are most suited for predicting or describing data sets with binary or nominal categories.[35] The most common type of Classification task is the Binary Classification with a True or False target variable. It is used, for instance, to predict if a test subject has a disease or not, or to predict if a e-mail message is spam or not. Multi-class Classification algorithms are present in several fields and they are used, for instance, to classify movies into genres or to determine the species of a plant based on its characteristics.

Most classification algorithms typically have two phases: the *learning step*, where a classification model is constructed learning from a training set of data. The goal in this phase is to learn the function $f(\bar{X})$ that best explain the relationship between the feature variables X and the binary class variable y, that can be described mathematically as:

$$y = f(\bar{X}) + \epsilon$$

where $f(\bar{X})$ is the function representing the true (but unknown) relationship between the feature variables and the class variable, and ϵ is the intrinsic error in the data that cannot be modelled.

The second phase of Classification algorithms is the *classification step*, where the model is utilized to predict class labels for unseen data with the same structure of attributes. Usually called the test dataset, this set is used to estimate the accuracy of the classification rules learned by the algorithm. If the accuracy is considered acceptable, the rules can be applied to the classification of new data points.

There are many type of classification models available and the most well-known include decision trees, rule-based classifiers, probabilistic models, instance-based classifiers, support vector machines, and neural networks.

2.2.1.2 Bias-Variance Trade-off

Most classification algorithms construct a model $g(\bar{X}, D)$ based on some kind of modelling assumption, e.g. a linear decision boundary, but these assumptions may not reflect the true model and that oversimplification brought on by the characteristics of the model, creates a bias in the modelling process. On the other hand, even if the assumptions made by the model are indeed correct, it is not possible to entirely capture the relationships in the data with any given training data set. Over different instances of a training data set, the predicted class labels will be different to an extent.

Controlling the flexibility of an algorithm is dependent on the balance between bias and variance. Bias occurs when a Machine Learning model is unable to capture the true relationship between the features and target of the data. Variance, on the other hand, is a result of the model making too complex assumptions, and refers to the sensitivity of predictions to the variability of training observations. In practical terms, a model with high variance may represent the data set accurately but could lead to *overfitting* with noisy or otherwise unrepresentative training data. *Overfitting* occurs when a prediction model has a good performance in the training stage, but a low generalization ability to new data. The prediction model learns not only the general patterns of the data, but also many specifics and noise found in the training set. It is considered that the model did not learn the main patterns or trends of the base data, but simply “memorized” all its cases.

In comparison, a model with high bias may *underfit* the training data due to a model that is too simplistic and that overlooks regularities in the data. To cope with this issue, it is important to build a learning algorithm flexible enough to correctly fit the data while also being sensitive to unseen data used for prediction and estimates.

The bias-variance trade-off is a particular property of Machine Learning models, that enforces a trade-off between how flexible the model is and how well it performs on unseen data. The flexibility of a model describes the ability to increase the degrees of freedom available to the model to fit to the training data.

In order to decrease the variance component of the prediction error, some methods may be used such as: (1) implementing Ensemble learning methods, that are able to leverage on both weak and strong learners in order to improve model prediction; (2) adding more data using a larger training set, reducing the data-to-noise ratio. . To reduce the bias, common strategies are: (1) changing the model to another who could better capture details in the data; (2) ensuring data quality and that training dataset is really representative of the whole; (3) tuning model hyperparameters like regularization and penalization for .

2.2.1.3 Decision Trees

Decision trees are a supervised learning methodology used for Classification and Regression that predict the value of the target variable by learning simple decision rules inferred from the data features. The data is continuously split according to a set of hierarchical decisions based on features present in the data, and it is arranged in a tree-like structure. Internal nodes represent the features, branches represent the decision rules and leaf nodes represent the outcome.

Decision Trees usually mimic human thinking ability while making a decision, so they are easy to understand and to explain, even to users not familiar with Machine Learning methods. In addition to that, trees work greatly when there is high non-linearity and complex relationships between dependent and independent variables, a setting that would not be ideal for a model that assumes linear relationships in the data. However, this flexibility can also be an issue, as this often causes an overfitted and possibly unstable model to even small changes to the data.

Algorithm 1: Generic decision tree training algorithm

```
Data: $D$
begin
    Create root node containing D;
    repeat
        Select an eligible node in the tree;
        Split the selected node into two or more nodes based on a pre-defined split
          criterion;
    until no more eligible nodes for split;
    Prune overfitting nodes from tree;
    Label each leaf node with its dominant class;
end
```

Popular Decision Tree algorithms are C4.5, ID3, C5.0 e CART[13]. For instance, Classification and Regression Trees (CART) constructs binary trees using the features and thresholds that yield the largest information gain at each node using the *Gini Index* as an attribute selection measure.[96] Unlike other Decision Tree algorithms, it supports Regression tasks as well. Another important and distinct characteristic of CART Decision Trees is that it only generates binary splits while other algorithms are able to do multiple splits at one particular node.

There are several measures of selecting the split criterion between nodes in a Decision Tree. Different algorithms have different choices of measure and that influences the way data runs through the algorithm and how the splits are selected. CART, for instance, uses the *Gini Index*, while C5.0 uses the Information Gain or Entropy[96].

The *Gini impurity Index* measures the impurity of D, of a data partition or a set of training tuples. The Gini impurity of a classification tree node is calculated using the count of each target category in all records corresponding to the specified node. The Gini impurities total is calculated as a sum of squares of count proportions across all target categories per node subtracted from one, and the result is multiplied by the number of records.[19]

To avoid overfitting and growing trees inefficiently with many unnecessary nodes, one of the most important processes in growing Decision Trees is pruning[13]. Pruning can be done in two ways: pre-pruning, during the creation of trees, or post-pruning after the trees are created.

Pre-pruning involves stopping the tree before it has completed classifying the training set based on termination criterion, thus preventing the generation of non-significant branches.In post-pruning, the tree is generated at its maximum size and then it is pruned using reliable evolution methods. Then the best performing subtree is chosen. Pre-pruning is faster but less efficient than post-pruning because of the risk of interrupting tree growth when selecting a sub-optimal tree[13].

2.2.1.4 Ensemble Methods

Different predictive models often arrive at different results due to inherent characteristics such as their sensitivity to noise and random artefacts in the data, their ability to capture the linearity of relationships and other specific characteristics. Also, in most cases, it is impossible for a single model, no matter how good it is, to fully capture the nuances that exist in the data.

Ensemble learning is a machine learning paradigm where multiple models called weak learners are trained, to solve the same problem and are then combined to get better results. The main hypothesis is that when weak models are correctly combined we can obtain a strong learner, that is more accurate and/or robust.

Ensemble Methods seek to increase the accuracy of predictions by reducing bias and variance, combining the results of multiple classifiers or multiple iterations of the same classifier. Although the functioning varies from one method to another, a weak (base) learner is commonly trained and either: (1) new models are trained to later add the forecasts to the base model; or (2) the same model is trained with new data, usually selected by sampling with substitution. Thus, and with the help of a correct parameter tuning, ensemble methods reduce classification errors by decreasing the combined model's bias and variance.

2.2.1.5 Bagging

Bagging, or bootstrapped aggregating is an approach designed to reduce the variance of the prediction generated by a Machine Learning algorithm.[37] The main idea behind this approach is that, given sufficiently independent predictors, the variance of a prediction can be reduced through a combination of bootstrapping, or sampling data with replacement, and aggregating the predictions of several classifiers, where the dominant vote between them is reported as the predicted class label.

Decision Trees, if grown deep enough, are great candidates for bagging, because they end up obtaining low bias and high variance, and the bagging works by reducing the variance. An issue with bagging, though, is that the assumption of independence of the predictors is usually not satisfied because of correlations between ensemble components, and there are a few methods that try to address that.

2.2.1.6 Random Forest

Random Forest is one of the most common algorithms for both Classification and Regression tasks, especially for its robustness and resistance to noise and outliers[37]. A Random Forest is an ensemble of Decision Trees in which randomness has explicitly been inserted into the model building process of each decision tree.

One common issue with Decision Trees is that trees ended up being correlated with one another, since the split choices at top levels of the tree are likely to remain approximately the same, even with bootstrap sampling. That leads to trees having the same attributes and splits,

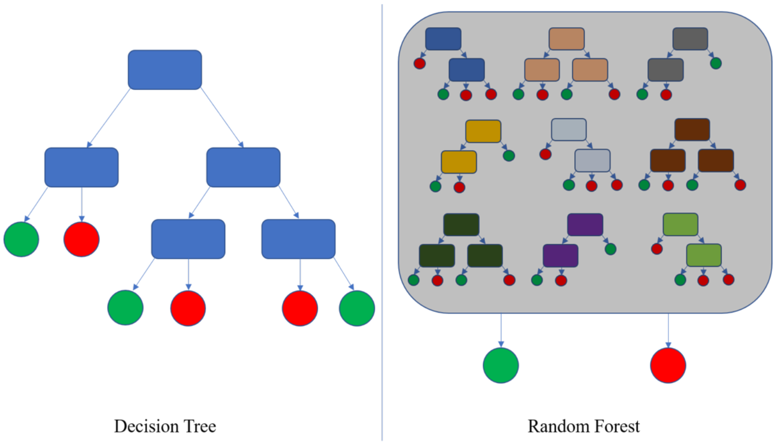


Figure 2.5: Visual comparison of a Decision Tree with a representation of Random Forest structure, reproduced from Beauchamp, 2020[72]

and hinders the possibility for error reduction that could be obtained from the bagging process described above.

A Random Forest algorithm addresses that by increasing the diversity of the component decision-tree models by introducing a second layer of randomness into the split criterion. Before growing the trees, it randomly selects a subset S of attributes q and the splits are executed only using this subset. Several trees are trained and the aggregation process is done similarly to the stand-alone Decision Trees method, taking a majority vote to determine the predicted label in the case of a Classification problem. This added randomness leads to less correlated component models and therefore, less variance and better results overall.

The overall result of this approach is very sensitive to the choice of the parameter q. Higher values of q do not generate a great improvement comparing to training a Decision-Tree on the entire feature subset and too small q makes necessary for a larger number of ensemble components. Additionally, dimensionality plays an important role, so much so that accuracy gains in less-dimensional settings are reduced. Nevertheless, the Random forests approach usually performs better than bagging and comparable to boosting, making it a great choice for Machine Learning tasks of Medium to High complexity.

2.3 Model Assessment

The next step after building any machine learning predictive model is to evaluate how well it did when compared to the observed values, also called “ground truth” or to any other models. The evaluation stage is also crucial to find hyper parameters that better explain the relations in the data without overtraining the model to a specific set of data.

Lastly, in most cases it is also useful and necessary to compare the performance of one model to those of others built with the same data. Some models adjust fit better to a certain type of data distribution.

2.3.1 Data Splitting

A well-established part of model assessment lies in splitting the data in a number of segments so the model is trained in part of the data first, and then evaluated against a new set of observations. This procedure ensures the model is free of biases, or at least has as few and smaller biases as possible. Furthermore, using the same portion of data to build and test a model would often lead to an overestimation of the accuracy of the model.

It is also paramount to perform all data preparation and preprocessing techniques separately on training and test data to prevent the occurrence of data leakage, as is also known in Machine Learning literature when data from the training part leaks into the test part, causing test observations to be impure. This provides an unfair advantage to the model and hinders the assessment of its predictive ability. Data leakage can also happen when working with time-related data, particularly when data from the future is leaked to the past.

The method of splitting the data makes for a truer test of the application of the models in a real-world setting, since the selected models will have to deal with new and previously unseen data. There are quite a few methods in use currently to split data for model assessment. Some of the most commonly used, as discovered by literature review, are discussed below.

2.3.1.1 Holdout

One of the more traditional ways of preparing data for validation is the holdout method, i.e. randomly splitting the data in two or more disjoint sets. The most usual setting is splitting the data in two parts, called train and test sets. Commonly used proportions are 70/30 and 80/20, with the largest number indicating the percentage of data put aside for training. The domain, the type of data and the size of data available are all factors that may influence the size of the split.

The first and typically largest portion of the data, often referred to as the training set, is used for building the models initially. Only after building and tuning the parameters of the models should the remaining data come into play, and that is the test set, a portion of the dataset that remains untouched for most of the process and is only used for final evaluation of the selected models.

Part of the scientific community goes further and advocates for splitting the data in three parts, with a second part for validating and fine tuning model parameters in between train and test sets, with relative sizes ranging from 15% to 30% of the full data.

While it is very useful to have additional data to fiddle around with parameters and build better models, in some situations, splitting the data in several parts can also cause more harm than good, especially when there are not many observations to start with. This exhaustive splitting and subsequent shortness of data can result in a weak base model and an uphill battle to achieve good results while fine tuning the model.

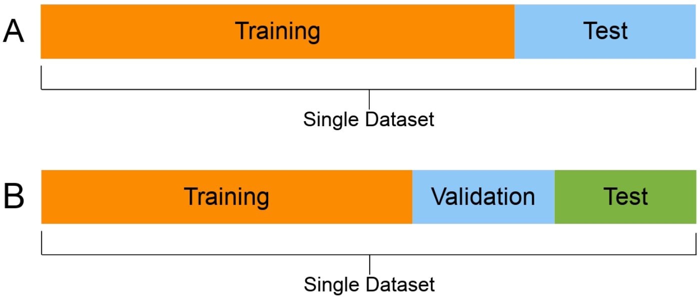


Figure 2.6: Holdout methods, with and without a validation set, reproduced from Buitinck, 2013[24]

2.3.1.2 Cross-Validation

Cross-validation is a technique used for assessing how a classifier will perform when classifying new instances of the task at hand. One iteration of cross-validation involves partitioning a sample of data into two complementary subsets: training the classifier on one subset (called the training set) and testing its performance on the other subset (test set).

In k-fold cross-validation, the original sample is randomly partitioned into k subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the classifier, and the remaining k 1 sub samples are used as training data. The cross-validation process is then repeated k times, with each of the k subsamples used exactly once as the test data. The k results from the folds are then averaged to produce a single performance estimation.

It is also possible to build the index for the k-number of splits based on a column or a set of columns to ensure each k fold represents the split of the entire column to the best extent possible. That ensures that the distribution of the dependent variable is the same throughout sets, for better validation of the machine learning models trained.

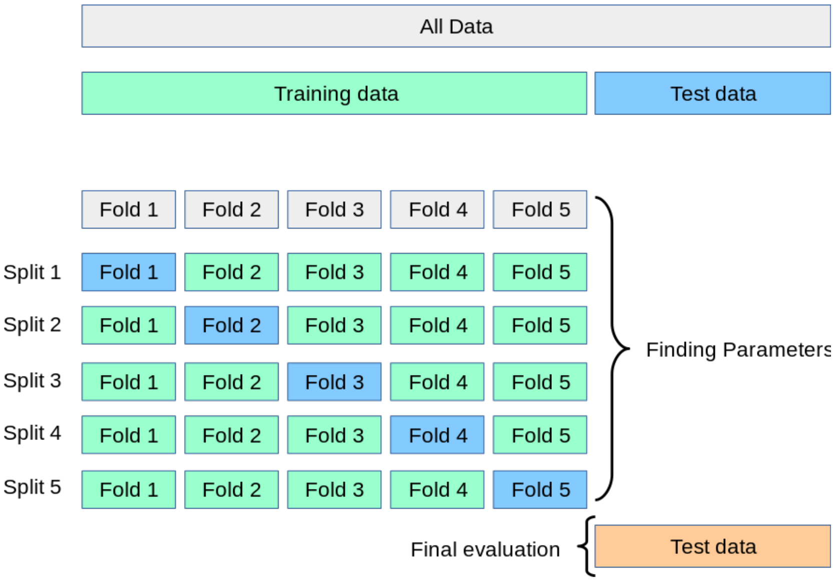

All Data
Training data
Test data
Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
Split 1 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
Split 2 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
Split 3 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
Split 4 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
Split 5 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
Finding Parameters
Final evaluation
Test data

Figure 2.7: Validation scheme for 5-fold Cross-Validation, reproduced from Buitinck, 2013[24]

Splitting the data in K folds combines benefits of both train/test split and train/validation/test split, since it feeds more data to the model right from the start than a 3-way split and provides more flexibility and accountability than a 2-way split, considering it will use K times the data actually available and the results of running the model in each split are then averaged for maximized statistical significance.

Ultimately, a portion of the original data still remains to be seen and can be used for final evaluation of the model before deploying it to the real world.

2.3.1.3 Time Window

A similar method of validation used when there are time-based attributes, the time window guarantees benefits similar to the Cross-validation, making sure that the model is evaluated as it would be in a practical environment. The main application of this method is in the validation of Time Series data, but it can be applied to a series of other cases. It also makes it possible to use a weight or cost-matrix to give more value to more recent predictions instead of taking the simple average of each run of the model, especially when time-evolving data is concerned. In such applications, a simple average of the performance of each model would sometimes not be enough.

2.3.2 Evaluation measures

A Machine Learning predictive model does not stand alone and absolute in time, so when necessary to compare its performance with a certain baseline or against any other challengers, one good way to do it is through well-established evaluation metrics. These measures are adequate ways to compare models that go beyond model-specific or data-specific characteristics such as units, hyperparameters and size of the data.

In relation to the result generated by a predictive model, many models output numerical scores associated with each test instance and label value. This is particularly useful for the added flexibility in evaluating the predicted classes as positive or negative values, where the model outputs a numerical score usually between 0 and 1, and the user may set a threshold to determine where a prediction lies in one side or the other. The value of 0.5 is the default for all models, but there are ways of investigating further the effect of changing this number on the results evaluating a predictive model.

The selection of the threshold becomes critical and hugely affects the interpretation of the results generated by a predictive model. It results in a trade-off between the number of false positives and false negatives, as there seldom is one right threshold to be selected in a real-world setting. However, the trade-off curve can be quantified using a variety of measures, and two algorithms can be compared over the entire trade-off curve. Two examples of such curves are the precision–recall curve, and the Receiver Operating Characteristic (ROC) curve.

Some of the most important aspects regarding evaluation of predictive models:

		Predicted value		
		Positive	Negative	Total
Actual value	Positive	TP	FN	$a+b$
	Negative	FP	TN	$c+d$
	Total	$a+c$	$b+d$	N

Table 2.1: Confusion Matrix for a Binary Classification task

Confusion Matrix: Evaluating the performance of a classification model usually involves analysing the predictive ability or correctly separating the classes of the target variable. Using a threshold such as the default 0.5 enables the evaluation of predictions as a binary setting, even in cases when the model outputs a numerical value as a result of its computation.

A commonly used structure for evaluating binary Classification models is known as the *Confusion Matrix.*[22]. In a Confusion Matrix, the classification results are presented as a two-dimensional matrix, with a row and column for each class, as shown in table 2.1. There are four possible combinations for the prediction results of a binary classifier. True Positives (TP) and True Negatives (TN) values - the main diagonal of the matrix - correspond to the correctly predicted values from the classification model. False Positives (FP) and False Negatives (FN) values - the secondary diagonal - represent incorrect answers.

In addition to conveying conducive empirical results, the confusion matrix facilitates statistical analysis of predictions, allowing the calculation of various measures of evaluating the performance of a classification model. These measures are important mainly because they allow a quick view of the model's result in a single numerical component and standardized comparisons with reference lines or with the result of other models and/or studies.

Accuracy: The most traditional evaluation metric for Machine Learning models[31], it represents in a very straightforward way the number of correctly predicted instances by the model in relation to the total number of predictions. It is a useful metric when errors in predicting all classes are equally important, but that is seldom the case when trying to predict real-world behaviour.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Overall accuracy is an important measure for observing the quality of a model, but it must be analysed very carefully and never in isolation. In imbalanced datasets, where most instances belong to one class, the use of accuracy may lead to erroneous conclusions. A classifier would be able to achieve an accuracy of 90% in a dataset where there is a 90/10 distribution of the target variable simply by predicting all values as of the majority class, and such classifier would not be useful.[31]

Precision and Recall: Precision is the ratio of correct positive predictions to the overall number of positive predictions. It measures the reliability, or accuracy, of the information extracted.

$$Precision = \frac{TP}{TP + FP}$$

Recall is the ratio of correct positive predictions to the overall number of positive examples in the dataset. Recall is a measure of the amount of relevant information that the system extracts.

$$Recall = Sensitivity = \frac{TP}{TP + FN}$$

Specificity: Specificity is defined as the proportion of actual negatives which got predicted as negatives by a predictive model. This metric is a counterpart for Recall and is often used in cases where classification of true negatives is a priority. This metric can be also called True Negative Rate and it is defined by the following formula:

$$Specificity = \frac{TN}{TN + FP}$$

F1-measure: It summarizes both precision and recall in one single value, with a harmonic mean instead of the arithmetical one, intended to penalize extreme values of one measure of the other. The maximal value of F1 is 1.0 and is obtained when both precision and recall are 1. The minimum value of 0 is obtained whenever one of them is 0, even if the other one is 1. A higher F1-measure indicates greater performance, and while it does provide a better qualification measure than any of those by themselves, it is still dependent on the threshold, 0.5 is the default.

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN}$$

Kappa: A metric much more adaptable to different distributions of the dependent variable used to evaluate predictive models is the Cohen's Kappa coefficient. It is a statistic that represents the level of agreement between two different classifiers, comparing the accuracy obtained by a classifier in relation to the expected - and randomly calculated - accuracy.

$$Kappa = \frac{TotalAccuracy - RandomAccuracy}{1 - RandomAccuracy}$$

Total Accuracy is the notorious Accuracy used everywhere in Machine Learning and described in formula 2.3.2 above. Random Accuracy, according to Landis, 1977[3], is a hypothetical expected probability of agreement under an appropriate set of baseline constraints, defined as:

$$RandomAccuracy = \frac{(TN + FP) * (TN + FN) + (FN + TP) * (FP + TP)}{Total * Total}$$

Not only can the kappa statistic shed light into how a Machine Learning classifier performed, the kappa statistic obtained for one model is directly comparable to the kappa statistic for any other model used for the same classification task. Also according to [3], proposed levels of Kappa as a measure of strength of agreement can be defined as follows:

Kappa Statistic	Strength of Agreement
<0.00	Poor
0.00 - 0.20	Slight
0.21 - 0.40	Fair
0.41 - 0.60	Moderate
0.61 - 0.80	Substantial
0.81 - 1.00	Almost Perfect

Table 2.2: Kappa Statistic - according to Landis, 1977[3]

While this is a great measure to compare the performance of Machine Learning models, it is important to note that there are many scales and these coefficients should be compared with parsimony, taking into consideration the domain of application and the specific study.

2.4 Data Reduction

One of the biggest issues faced in the Machine Learning domain is to determine which attributes are the most important when undertaking any kind of predictive or analytic Data Mining task. However, it can be cumbersome and time-consuming to find relationships in the data and generate insights and predictions in highly-dimensional datasets, especially since more often than not, cost and time-efficiency are valid concerns.

Smaller computing times and reduced storage size are the most obvious and quantifiable benefits of applying dimensionality reduction techniques, but these methods may present a host of other advantages. Data quality is usually improved, since noisy, irrelevant and redundant data can be discovered and removed, and it also becomes clearer to visualize patterns in the data and to present the data itself.

2.4.1 Feature Selection

By feeding only relevant and insightful attributes into a machine learning model, it is possible to minimize computation times and make the best use of limited resources. Feature Selection is an important part of Machine Learning tasks and looks to find the best predictor variables out of all the variables available.

Reducing the dimensionality of the data by eliminating inappropriate attributes, improves the performance of learning algorithms and produces a more compact representation, which also enables an easier interpretation of the target concept, making the user focus on the most relevant variables. [52]

With filter methods, features are selected on the basis of their scores in various statistical tests for their correlation with the outcome variable. Filter methods are extremely diverse when it comes to their complexity and computational costs. Most take very little time and resources to run, such as calculating the correlation between predictors.

Univariate feature filters evaluate (and usually rank) a single feature, while multivariate filters evaluate an entire feature subset. Univariate filter methods consider especially two aspects:

Individual importance: The amount of information one variable by itself has in predicting the target variable. There are several methods for identifying variable importance, such as (1) Pearson correlation, a measure of the strength of a linear association between two variables, usually denoted by r[33]; (2) Chi-square statistic, used to show whether or not there is a relationship between two categorical variables, denoted by χ^2. (3) Student's-t statistic, indicates the difference between two groups of continuous variables, denoted by t.

Redundancy: One of the most used steps in reducing the number of attributes in the feature space and thus limiting the dimensionality of the models is removing variables that have very strong correlation with one another. The presence of highly correlated variables, i.e. those with an absolute Pearson's correlation coefficient of more than 0.90[85], make it difficult to evaluate their relative importance without running the risk of making inadequate inferences. Hence the need for identifying such variables and removing them from the models, to improve computation times and increase model effectiveness.

2.4.1.1 Forward and backward subset selection algorithms

Search algorithms such as the ones that use backward subset selection are multivariate computationally expensive methods, that require training and evaluating several models with different combinations of variables. Studies such as MacFayden and Dawson, 2010[15] and Oliveira Junior, 2015[43], leverage these type of algorithms with the end goal of predicting student performance.

This technique starts by building a Random Forest model across the entire set of predictors and it computes an importance score for each predictor. The least important predictor is then removed from the feature subset, another model is trained, and the importance scores are recalculated. And that process is repeated until predefined stopping criteria are met, or until the model's accuracy has reached its peak in comparison between subsets.

Although this process is extensive, it can be streamlined with recent algorithms, especially with techniques such as Recursive Feature Elimination (RFE), a Non-linear multivariate method implemented in several programming languages, which uses Machine Learning models to calculate variable importance and determine the subset of variables that gives the best result for the evaluation metric of choice.

RFE works as a backward elimination algorithm, which starts with the complete set of attributes and removes the attributes that degrade the accuracy of the underlying algorithm.[95] Just as all other backward feature selection methods, it is remarkably universal and simple in its application. These methods are called greedy-search methods because they make the best optimal choice at each step with the goal of this eventually leading to a globally optimum solution. This means that the algorithm picks the best immediate output, not consider the whole set of variables. While this does not guarantee that it will find a better solution, it is still a good choice for finding a good set of predictor variables within reasonable time.

Chapter 3

Related Works

In this chapter, we sought to make a systematic review of the works related to this study in terms of methods, data and objectives, drawing a relationship mainly in relation to the latest developments in the areas of Educational Data Mining and Learning Analytics. Several recent studies relevant to these two themes have been found and analysed, which shows that these are hot topics in the scientific community.

Educational Data Mining (EDM) and Learning Analytics (LA) are two distinct and continuously developing areas that work with educational data from all kinds of sources, but lately, there has been a huge influx of data from social media, Student Information Systems and others sources, so a lot of focus has been put in analysing Big Data for educational purposes.

3.1 Educational Data Mining and Learning Analytics

EDM and LA are two interdisciplinary communities of computer scientists, learning scientists, psychometricians, and researchers from other areas with the same objective of improve learning starting from data[84]. Educational Data Mining and Learning Analytics have a few common goals and similar interests, but also many distinguishing differences. First, LA places more emphasis on describing data and results using automated discovery as a tool to achieve its goal, namely, using findings to leverage humanized judgment. Meanwhile, EDM has strong origins in educational software and student modelling with a significant community in predicting outcomes. With the latter, automated discovery is more highly valued and harnessing human judgment is a tool used to achieve this goal.

Learning Analytics efforts are commonly implemented as a data-based method for detecting at-risk students, but the possibilities go way beyond this subject. Higher education institutions need to provide additional support, for their own benefit and that of their students, pursuing the ultimate goal of promoting academic achievement.[80]

Methods in EDM usually aim to better understand student behaviour in their learning process and analysing their interaction with the environment. There is also a significant and urgent need to provide appropriate computational environments for educational data mining, offering ease of use for each of the stakeholders involved.

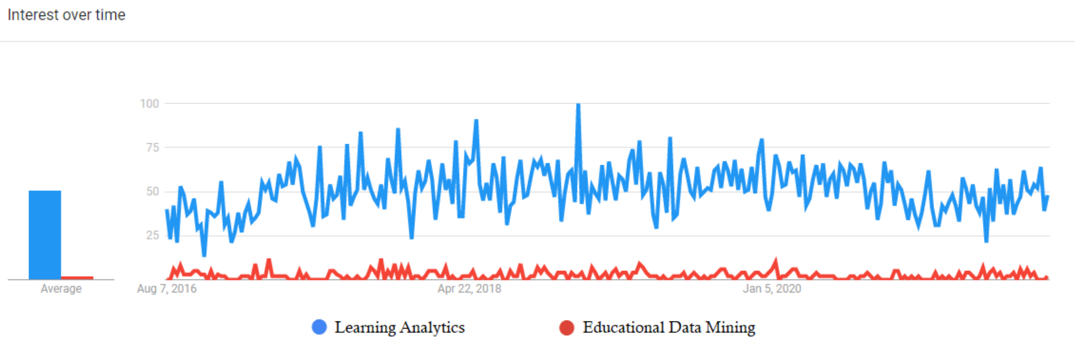


Figure 3.1: Google Searches for Educational Data Mining and Learning Analytics from 2015 to 2021

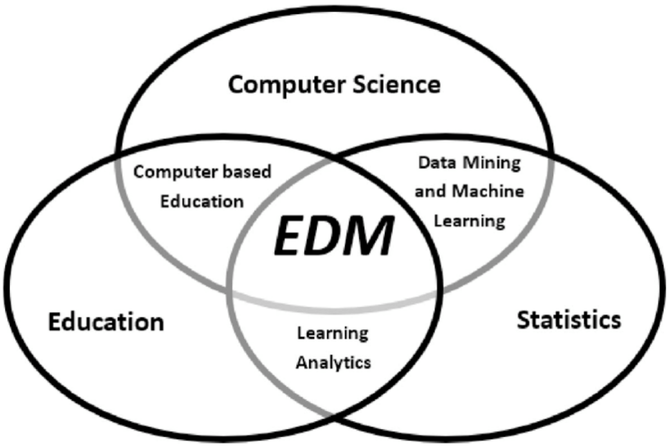


Figure 3.2: Educational Data Mining, Learning Analytics and closely related areas, reproduced from Romero, 2020[84]

3.1.1 Educational Data Mining

In recent years, extensive efforts have been made to build and extend the use of Student Information Systems within higher education institutions. The extended importance of such systems, for one, incurred in improvements in the entire information chain, with more efficient collection, preparation and evaluation of student-related information. This also allowed for the application of Data Mining algorithms and methods, generation of insights and, ultimately, the identification of relevant patterns of student and faculty behavior through state-of-the-art Data Science algorithms.

The field of Educational Data Mining leverages this growing availability of data and seeks to develop or adapt existing mining methods and algorithms in order to help better understand data in educational contexts.[17] Although it is difficult to precisely determine the date of emergence of such a comprehensive scientific area, it is noted that the first workshop in the area of EDM was held in 2005, still part of a conference on Artificial Intelligence. This goes to show how young this area is compared to other areas of technology.[84]

According to Romero & Ventura, 2020, Educational Data Mining can be defined as a combination of three major areas of science, computer science, education and statistics, as seen in figure 3.2. Still according to researchers in the area, the scientific area most connected with EDM is *Learning Analytics*.[84]

It has been historically difficult to study how much the differences between teachers and class groups influence specific aspects of the learning experience. This type of analysis is made much easier with educational data mining. On this subject, albeit in very different experiment settings, studies by Gašević[91] and Finnegan[11] observed that it is difficult to find attributes that explain extremely well relationships across different scientific areas, let alone smaller granularities, such as programmes and courses.

3.1.2 Learning Analytics

Learning Analytics is defined as the measurement, collection, analysis and reporting of data on the students and their learning contexts in order to understand and optimize learning and the environment where it takes place.[46] LA makes it possible to discover the learning difficulties that a student or group of students face in their day-to-day school activities, and suggest a customized learning process to the specific context of the student/user.

Also used in other contexts, studies on this field are not simply focused on student performance, but can also be used to assess Course Units, Programmes and entire Institutions. Another important use of LA methods is to monitor and predict student development, detecting possible issues and inadequacies in advance so that measures can be taken in order to improve results and use of resources for all involved.

More recent even than EDM, LA is a practice that emerged just over 10 years ago. Only in 2010 the first conference in the field was held, The International Conference on Learning Analytics and Knowledge. However, it is an area that has attracted many very prolific researchers, and the pool of studies in the area has been increasing every year.

Some other Learning Analytics methods are Social Network Analysis, which can be used to interpret and analyse the structure and relationships in collaborative tasks and interactions with communication tools [30] and Sentiment Analysis, a collection of Natural Language Processing (NLP) procedures used to extract information from texts - structured or otherwise - to determine the speaker's feelings, opinions and attitude towards a particular entity. It is a scientific area that has much to be explored[92], and through the methods mentioned, it can answer questions related to students engagement and student feedback toward a class, course, or institution.[92][6]

3.2 Common tasks in EDM and LA

There are several tasks generally developed in EDM, notably those that result directly from the analysis of data generated in students' interactions with learning environments. Some of the most used Educational Data Mining techniques are: decision trees, support vector machines, logistic regression, Bayesian networks, linear Regression, Neural Networks, K-means algorithm, Gaussian Mixture Model, Sequential Pattern Mining and Association Rules Mining.[30]

These methods have a wide range of applications, for example, from categorizing groups of students to recommending learning strategies based in specific patterns of student behavior to predicting performance in higher education.[1]

A few practical challenges that are present in various educational contexts are related, for example, to the lack of standardization of data, which ends up requiring a great preprocessing effort.[17] In addition, in some cases it is necessary to adapt classical data mining algorithms for the educational context, to deal with characteristics such as statistical non-independence and data hierarchy.

In an educational context, a fairly comprehensive classification of the most common tasks found in EDM settings was performed by Baker, Isotani and Carvalho in their 2011 work[17].

Prediction: these methods are mainly used to predict student educational outcomes, either academic performance in terms of grade obtained, using Regression methods, or whether the student completes the course successfully or not, through Classification methods. Prediction methods are also used in studies that aim to predict an attribute when labelled data is not available. For instance, that is sometimes the case when dealing with psychological factors, when the mere fact of labelling may alter the construct being studied.

For instance, a compelling use of Classification methods was in study by Vandamme et al.[10], where students were classified into three groups: 'low-risk' students, with a high probability of succeeding; 'medium-risk' students, who may succeed if the university takes appropriate measures; and 'high-risk' students, who have a high probability of failing or dropping out.

Clustering: a set of techniques generally used to find existing groups and patterns in the data, for example, groups of students that behave in a similar way or that have the same attributes or even the same marks in their exams.

Relationship Mining: aims to discover relationships between variables or sequence of actions. In EDM, common uses would be to find rules that explain student behaviors such as "if the student has many accumulated years then they are possibly at risk of dropping out", and to determine sequence of reoccurring events such as bad grades that lead to the student failing a course to them dropping out of college.

Distillation of Data for Human Judgment: information is generated to aid human identification and classification of student data.

Discovery with Models: many studies related to Educational Data Mining, deliberately or not, use this technique. Hershkovitz et al.[26], conducted a study in 2013 with young students about *carelessness*, generally defined as "giving the wrong answer despite having the needed skills for answering correctly". Such behaviour is harmful to a student and can indicate cognitive problems, influence of external factors such as pressure from parents, tutors and peers, or even the inadequacy of the exam model for the behaviour of students. This study was carried out using two models, comparing performance data of participants in scientific tests with psychological assessments that had previously divided students into groups.

[1] More on the applications of Data Mining methods on Educational Data Mining and Learning Analytics in 3.2

3.3 Scientific review on student performance prediction

Many studies using educational data have applied several statistical methods in order to determine which would achieve the most accurate prediction of the intended outcome variable. The main elements of focus in the scientific publications analysed were: student performance prediction, and student dropout prediction.

There are a multitude of studies in the subjects of academic performance and dropout prediction, and being such multifaceted and broad subjects, those are issues that have already been thoroughly investigated by many viewpoints.

Some studies analyse this problem from a more social and behavioural angle, seeking a greater understanding of the factors that lead to student dropout or academic success/failure. Others analyse the consequences these events bring to students' lives. A common characteristic of these studies is the use of data from surveys answered by the students themselves[59],[41], or even from the analysis of surveys used by the educational institutions applied to students at the time of student dropout.

Another entirely different class of studies in methods and goals are those conducted in Engineering and Information Sciences fields, especially with latest advances and popularization of tools and methods for performing data-oriented analysis. While structured tabulated data has been used for decades and statistical methods are certainly present in most scientific studies, latest computational advancements have made it possible to analyse thousands of records and reach conclusions in little time, something it was never thought possible only a few years ago.

3.3.1 Student performance

In educational contexts, the assessment of a student's learning development is an indispensable means of decision-making by the faculty regarding the continuity of the relationship and pedagogical routine with students, and today it is seen as a fundamental part of the education process.[53]

Regardless of the level and the age of the student, the evaluation process should focus on the student and on verifying the evolution of cognitive and non-cognitive aspects of the individual. Student assessment also provides a unified frame of reference for the entire educational system and serves as a basis for defining all the precepts of the modern education system.

Student performance, when objectively defined through a numerical grade or a letter on a scale, is an indirect measure of the student's knowledge or proficiency for what was assessed, and through this measure, it is inferred that there is a relationship between the student's performance in an assessment and the actual academic performance obtained by the student.[53]

There are several definitions of academic success in the literature, and accordingly, performance prediction studies have varied goals and contexts.[2] Broadly speaking, study success includes the successful completion of a degree in higher education, as seen in studies based on interactions

[2] A more complete list of recent influential studies on Performance Prediction can be found in attachment A.1

with activities on a Learning Management System[79] or based on a combination of work status, academic history and demographic attributes[42].

The results of a predictive model can be used to encourage those “potentially” low-performance students to develop a better learning strategy[27]. Faculty members and administrators can also direct affirmative action and support for students at risk of failing or dropping out of their Programmes, and when a student’s academic career is at stake, the sooner information is available and accessible, the better.

Although there is a common general goal among works aimed at predicting student performance, which is to determine which students are at risk of failing, the approaches used to achieve this goal vary widely. The most straightforward approach is the one in which the data is in tabular format, and there are clearly defined attributes that allow to determine if the student will fail or succeed in a certain subject or course through training of Machine Learning Regression algorithms or classification, as it can be seen in works by Martins et al, 2019[69], and Tomasevic et al, 2020[86]. Some authors have opted to take a different approach, predicting student performance deriving predictors from different student behaviours, such as student procrastination patterns of behaviour[79] or social network interaction data[20].

In the same way, using rather uniquely derived attributes such as entries in a portfolio system[64], library book lending, cafeteria spending, and library entry frequency[87] certainly introduce a fresh view on the subject and show that the possibilities have not been exhausted to arrive at insightful conclusions in terms of student performance prediction.

Recent studies in student performance prediction have had varied scope, but a common element of focus have been class imbalance in the target variable, due to the distribution of the data[94]. Failure rates tend to vary between 15 and 30% in any given year in most countries, so that means data for the average higher education course will be mildly unbalanced by its own nature. Studies in the field tend to reflect this characteristic and one common step is performing some kind of data balancing technique with good results in some studies[94][65].

Other aspects also have been explored recently, with important works using data visualization concepts setting the stage for future studies in the field, such as Martins, 2018[63], Deng et al, 2019[66] and Kim et al. 2016[50].

3.3.1.1 Attributes

One of the common steps among all studies carried out on the topic of school dropout is the identification of common factors and positively or negatively associated with student performance or dropout prediction.

According to Argolo, 2017[53], determining characteristics that influence a student’s performance can be divided into two groups: (1) student characteristics, and (2) the school environment, which includes (a) the technical and professional specifics of the faculty and the (b) infrastructure and mode of operation of the education institution.

In terms of academic dropout and performance prediction, as in most computational tasks, it is often useful to split the work into smaller tasks, and thus, arranging the features available in the data into groups can come in handy. Grouping features makes them easier to treat, analyse and remove if there is the need for that. When one feature group is deemed to be irrelevant or not relevant enough, it can be eliminated altogether, helping reduce the dimensionality of the data at an early stage, saving resources and hours of manual treatment by the user.

From the research carried out in the scientific literature, however, this is not a path followed by many researchers, with few works performing this type of experiment. Some features are somewhat hard to define and categorize, especially the ones that deal with behaviour and psychological attributes. In addition to that, there is always some discretion on the part of the researchers on how to structure the work and the data available for each study. Because of that, studies in this field have arranged features in groups in several different ways.

The range of groups where attributes are split is varied, as there are not decisive definitions in this subject. Some studies, such as Tomasevic, 2020[86] and Vandamme, 2007[10] opted for a three-way split of their attributes with Demographic and Participation attributes common between the two.

Others opt for a more unequivocal division and arrange their predictors in more specific groups. While more groups does not equate to better information on student performance by itself, five seems to be one of the most common numbers of groups of attributes, as found in Martins et al, 2019[69] and Mitra and Goldstein, 2015[42].

There are many classifications of attributes currently in student performance prediction studies. Some authors use groups found in systematic review papers, but most studies choose to define their own classes of attributes. Through a research carried out in the scientific literature, the attributes most used in recent studies are listed, in order of number of appearances in table 3.1.

Academic History: it consists of all academically-registered attributes that represent the student's journey up to the time of study in their academic path. These variables are usually strongly related with student performance, as one student who thrives in some Course Units will most likely succeed in others as well. Likewise, a student who succeeds in some Course Units will probably not drop out of their programme.

A literature review conducted by Shahiri et al, 2015[44], identified that the most used attribute when trying to predict students' performance was the Cumulative Average, or Cumulative Grade Point Average (CGPA). According to the same author, there are two main reasons for that: the CGPA has a tangible value for future educational aspirations and career mobility, and it can be considered as an indication of realized potential.

The use of variables from this group in studies goes further than the cumulative average score. Some of the variables identified were the number of European Credit Transfer and Accumulation System (ECTS)[3] credits registered and approved, the grades in previous Course Units, the attendance mode - as full-time students have higher success rates in general[21].

Other attributes of this class deemed statistically relevant in recent studies include course marks, progress in the programme, and plagiarism count[78], student rank in their class,

[3] ECTS credits indicate the required workload to complete a study programme, or a module within a study programme. Usually 1 ECTS is equal to 25 to 30 study hours.

Year	Author(s)	Dem.	Adm.	Hist.	Int.	Partic.	Others
2012	Huang, Fang	-	-	x	x	-	-
2016	Marbouti et al.	-	-	-	x	-	-
2019	Martins et al.	x	x	x	-	-	-
2010	Macfadyen, Dawson	-	-	-	-	x	-
2015	Stretcht, Cruz	x	x	x	x	-	-
2019	Lau et al.	x	x	x	-	-	-
2017	Al-Shabandar	x	-	-	-	x	-
2016	Qiu et al.	x	-	-	-	x	-
2015	Mitra, Goldstein	x	x	x	x	x	Work related, Motivation
2018	Okubo	-	-	-	x	x	Library card
2020	Wang et al.	-	-	-	-	x	Library card
2020	Guerro-Higueras	-	-	-	-	x	-
2014	Baccaro	x	x	x	-	-	-
2014	Sharabiani	x	-	x	-	-	-
2013	Arora	x	-	x	x	-	Behavior, Motivation
2016	Gasevic	x	-	x	-	x	Work related
2021	Jovanović	-	-	-	-	x	-
2018	Miguéis	x	x	x	x	-	-
2017	Daud	x	-	-	x	-	Work related
2019	Hooshyar	-	-	x	x	x	-
2018	Darlington	-	x	x		x	Behavior, Motivation
	Total	**12**	**7**	**12**	**9**	**11**	**7**

Table 3.1: Variables used in recent studies of Student Performance

specialization field, and first year performance[90] and number of course units enrolled, completed and failed[81], showing that this group of attributes has wide

Social, Economic and Demographics: Student demographics are another group of attributes that are usually available. While every institution collects information on the student at the moment of enrolment, this type of data is complex to collect (in relation to its completeness, correctness, optionality) and is highly dependent on the socio-economic context of the students.[81] Nevertheless, those attributes usually are available to researchers and data analysis systems set up in the institution.

Some of the relevant attributes in this group include gender, age, family background, nationality, and whether the student has any kind of disability.[29] Other studies have found socio-economic attributes also to be highly associated with student performance[32]. Student demographic attributes, when combined with other groups of variables, even without participation and engagement data, can show very good results in estimating study success.[34] For instance, Lau, 2019[67] determined that female students had reasonably better grades in college than male students and the student mother's occupation was also significantly correlated with better performance.

Those attributes are found to be at least somewhat relevant in several studies, but there is some controversy regarding its use in research. According to Grebennikov, 2012[21],there is less agreement among researchers regarding the effects of student gender, language background and ethnicity on student retention and completion of their program.

As addressed by a growing number of authors such as Fynn[47],Simpson[8], and Riazy[83], using attributes such as gender, race, nationality and other demographic factors in student analysis could lead to some discrimination on the part of decision-makers, especially when the results of these studies can be determinant to generating new policies and driving financial or psychological support, and that could affect one demographic group more than another[83]. Inevitably, targeting support and policies, means that certain groups within the population receive disproportional levels of support.

While it is a relatively new topic in the domain of Machine Learning, it is one whose importance has grown in recent years. The simplest solution to deal with the unfairness of a model is the complete removal of attributes, blinding the algorithm to characteristics that may generate controversy and unfairness. However, this type of solution hinders the predictive ability of a model and may hide some otherwise interesting relationships in the data. Other proposed solutions include: the use of a differentiated measure to optimize the algorithms during their training, which allows a factor in some degree of equality between demographic groups, or the realization of post-processing optimization to ensure all groups are equally represented in predictive results.[8]

Participation and Attendance: Another group of variables that represent internal events, that is, contained within the scope of a discipline or course. Several studies use attributes related to student involvement in activities, time spent carrying out activities, discussions with peers and messages sent, among others[91],[15]. Accordingly, many studies have demonstrated a positive association with the students' learning outcomes and their level and regularity of interaction with the resources relevant to the given instructional conditions.[91],[15]

One of the most prolific sources of this group of attributes are interaction logs stored by Learning Management System (LMS) software like *Moodle*. Data stored by such tools can represent aspects of learning that are difficult to determine or discover in other ways, such as study patterns, participation in classes and activities, and extent of discussions with peers.[7]

Many aspects of learning in higher education can be incorporated into a Learning Management System, and as a result, these pieces of software can collect all kinds of interactions student and class content. However, the usefulness of many variables is dependent upon course site design and pedagogical goals, as not all courses use the same resources available in this kind of software. Some of the variables available used in performance prediction are number of discussion messages posted in the discussion forum, number of mail messages sent, and number of assessments completed[15].

As content available in Learning Management Systems evolves, other attributes may also become common such as the ones related to video content. Using Video Learning Analytics techniques allows access to attributes relating to student engagement with video lectures such as such as number of plays, number of times the student paused and rewinded the videos, among other features. In fully-online courses, these attributes have been found to be highly determinant of the student performance.[78]

Internal Assessment: Also as identified by Shahiri, 2015[44], a class of closely related attributes which shows great importance in studies of student failure. Internal Assessment variables are used in studies that try to predict student performance with information from within that same period, i.e. assignment marks, quiz results, lab works, frequency, etc.

This class of features is used mostly when the task at hand is predicting student success in a particular subject. Components of this group include all means of evaluation in this subject, such as: assignment marks, prior test results, lab work and attendance.[44] Grades from pre-requisite courses[27] and performance data from homework learning objectives have also been used to predict student performance in one specific course[51].

Admission, application and secondary level performance: Data from the college application process or secondary level performance of the student is often used to predict student performance. Although it is not present in as many studies as the attributes of other groups, this information can certainly serve as a good basis for comparison between students, as the entrance exams or national exams given to students are usually standardized and designed so that the results are comparable and applicable to a large number of students[53].

As university admission exams are quite intricate and specific processes, some with specific exams for several subjects, they can be a great source of information for predicting student insuccess. Studies have shown results in admission exams can be strongly correlated with performance in higher education, and the correlation is higher for some parts of the admission exam rather than others[67].

While grades in admission exams are the most common attribute of this class[67][69], other attributes found in recent students were the time it took for the student to complete basic level of education, the average grade in high school[90] and grades in standardized exams for high-school students[65].

There are several other less common groups of attributes used in studies of performance prediction in higher education. As Learning Management Systems and School Information Systems evolve and are able to collect and combine data from more sources, researchers are handling increasingly larger quantities of data and being given access to different types of variables that provide new capabilities.

Some studies leverage psychometric and behavioural attributes such Ketonen and Lenka, 2012[23], that identified the relationship between situational academic behaviours, self-study time and learning outcomes in a Finnish course. Among other findings, it was determined with reasonable statistical significance that self-study time is associated with academic outcomes through direct engagement and also through increased interest in the subject of the courses.

Student motivation[49] and professor-student rapport[28] were also among motivation-related factors that were investigated for its correlation with first-year performance. Social network interaction data[20], work-related data[42], information extracted from library control systems and even data related to dormitory access cards[87] are some of the types of data already used in recent studies, with various degrees of success.

Ensuring maximum student privacy is another challenge associated with predictive efforts of student performance. On the purely objective side, the addition of attributes related to the students' everyday and personal situations has produced valuable predictive results[87], but the collection and use of such information must be discussed and evaluated from an ethical and humanitarian point of view to ensure decision-makers attain perfect balance between protection of student privacy and predictive ability of machine learning models.

3.4 First year of college is critical

The first year of college is the one with the highest fail rate among all curricular years [27],[21]. Arias Ortiz and Dehon[14] find that first-year approval rates are a reliable measure of academic success since they observe that students failing their freshman year are more likely to drop out.

The same period is also when the issue of dropout is most present. Studies of dropout trends in colleges in the United States show that 30% of the students do not return for a second year[82] and studies in Organisation for Economic Co-operation and Development (OECD) member-countries from 2017 show that, on average, 23.6% and 18.5% of first-year students dropped out from their First Degree achieving courses in Private and Public schools, respectively[58].

To succeed academically, first-year students need to adapt to the new university life and learn how to study and efficiently manage their time and optimize their study practices.[57] Pursuing a college degree can be challenging and overwhelming even, at times, with internal and external pressure abound from all people close to a student. Therefore, it is paramount that there is a clear disclosure of the objectives of a course, the expected student profile and options for the labour market at the moment of graduation, since a poor fit between the course and the student is one of the main reasons for student failure, especially in the first year of college[21].

In scientific studies on the matter, some measures are mentioned for strengthening the student-university relationship. Before starting a course, the faculty must maintain good communication channels between prospective students and academic staff, so that students can correctly manage their expectations regarding their chosen field.[21] Thereafter, it is important that the faculty take timely and timely steps to renew interest in the course or field; thus, the student can maintain a good performance and achieve satisfactory academic results, both in relation to the grade and in relation to the knowledge and training acquired.[23]

Looking to identify gaps and to provide possible improvements for first-year students, several studies today leverage Educational Data Mining and Learning Analytics methods, whether making and evaluating predictions at different prediction timings[89] or ranking first-year students by risk percentage, allowing decision-makers to have a clear view of those students[55].

3.5 Summary

Thought-provoking findings and advancements have been made in the past few years in Educational Data Mining and Learning Analytics. Still, there is a lack of large-scale, rigorous scientific studies that show clear evidence of improvements in education outcomes and that, consequently, positively influence student success in higher education.[80]

Still, promising smaller-scale results are beginning to appear in scientific literature and indicate that EDM and LA are relevant themes that aggregate value to the higher education educational process as a whole. Cerro-Martinez[73] identified significant reduction in dropout rates after the implementation of a tool intended to promote and monitor better communication in asynchronous learning environments.

Study by Nizan et al., 2019[70], verified the effectiveness of a predictive model for supporting an effective teaching and learning process in a university in Malaysia. Lastly, Bao, 2021[88] evaluated the effectiveness of a Learning Analytics Dashboard and discovered that teachers using the dashboard used more advanced strategies to diagnose students' learning problems while also performing more timely interventions with those students.

While those are good results, there is still a need for greater acceptance and greater inclusion of Learning Analytics methods by decision-makers and government educational entities so that desirable improvements reach their intended beneficiaries, i.e. the students. Educational institutions have a few mountains to climb in applying technology and latest EDM and LA advancements to obtain the best possible results they can achieve.

Chapter 4

Results and discussion

This chapter presents an overview of the Case Study where we trained predictive models in data kindly made available by the Universidade do Porto. This chapter also showcases detailed results for the predictive task performed, and describes how each of its stages contributed to answering the Research Questions and meeting the goals set out for this study.

4.1 Case Study

This Case study aims to determine what are the most important attributes and classes of attributes to predict student performance using data from 8 years of enrollments at Undergraduate courses at the Universidade do Porto and to leverage Machine Learning methods to predict student failure with the least error possible.

Features used in this study were acquired from the university's Student Information System and seek to provide the most complete view of the student's path at the institution, representing four classes, namely sociodemographic, admission, enrollment and academic background attributes. To identify differences between courses and provide correct predictions, we opted to split the data into courses and build specific predictive models for each course.

Several Classification algorithms were tested and the predictive algorithm chosen for this study was *Random Forest* for its flexibility, robustness, and smaller error in predictions in preliminary tests, outperforming other algorithms such as Logistic Regression, Naive Bayes, Decision Trees and Support Vector Machines on a representative sample of the data. This study looks to leverage a binary Random Forest classification algorithm to generate timely and accurate predictions of academic insuccess. The target variable for this task is binary with positive class labelled as "fail" and negative class labelled as "pass".

The methodology employed, as shown in Figure 4.1, resembles the stages of the CRISP-DM framework.[1] Leveraging the cyclical characteristic of that framework, along with business knowledge and understanding of how the data is distributed, each stage of the work was applied and re-evaluated more than one time in a iterative manner, to ensure continuous improvement

[1] See attachment A.3 for a more complete description of the methodology employed in this study and a description of all stages

Observations	1.35 Million
Period	2012 to 2019
Scope	Years 1-3 of Undergraduate courses
Schools	14
Programmes	56
Courses	1702

Table 4.1: Overview of data for this study

and the achievement of the best results possible with the data available. A comprehensive description of all activities undertaken is made available in A.3, with detailed accounts of data acquisition, data preparation and Feature Selection activities.

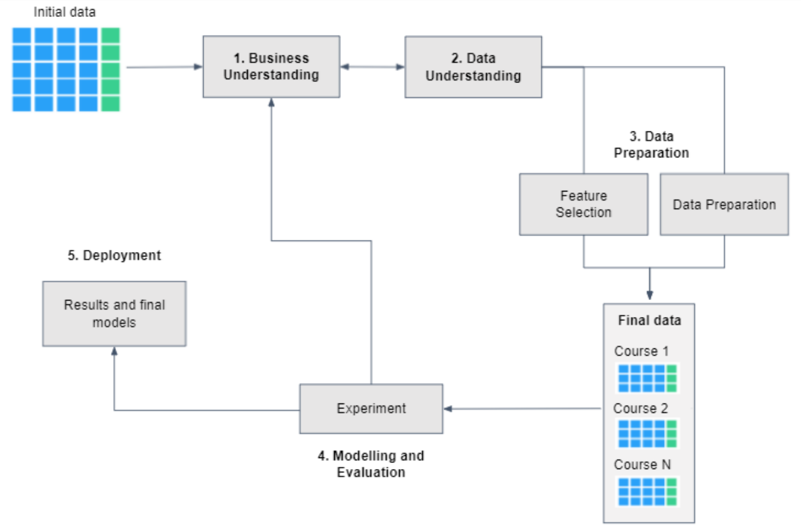


Figure 4.1: Overview of the methodology employed in this study

In the beginning of the study, we obtained a better understanding of the available data and maintained close collaboration with the university to ensure better data quality, treating missing values and performing necessary data transformations, making changes to the data acquisition process as we went along. To ensure the study would be reproducible, data preparation activities were stopped and the study used a snapshot of the data from May 15th, 2021. Unfortunately, data from 2020 enrollments was not available at that time and thus, was not used in this study.

As shown in table 4.1, this study was limited to years 1 through 3 of the first cycle of higher education, as defined earlier in 1.1. This filtering resulted in 1.35 Million observations, that spanned 1702 courses through 56 programmes.

Several data processing activities[2] were undertaken to prepare the data for training and evaluating predictive models: (1) data anonymization, to ensure students could not be identified through their attributes and results; (2) missing values treatment, based on the type and

[2] See A.3.2 for thorough description of data preparation activities

Class/Subset	**Attributes**
Attribute removal	Average Secondary, Average 11th grade
Set as category "Not Available"	Educational Level, Occupation, Country of Birth
Imputation of minimum value	Highest Grade, Lowest Grade

Table 4.2: Types of treatments of missing values used

Class/Subset	**Features available**
Grouping	School, Programme, Course, Curricular Year, Type of Degree
Sociodemographic	Age, gender, marital status, country of birth, student's and parents' occupation, student's and parent's level of education, whether the student had special needs
Admission	Student preference, Grade in admission process, Average in Secondary education
Enrolment	Student status, debt situation, type of enrollment
Academic History	Cumulative Average, First-year average, First-semester average, number of delayed courses, number of courses enrolled in semester, number of credits enrolled in semester, percentage of programme completion
Feature Selection	Best subset of variables obtained through Feature Selection methods

Table 4.3: Classes of attributes used in this study

completeness of variable; (3) grouping less representative factor levels, to reduce sparseness in some categorical attributes; (4) removal of attributes with zero variance, namely Type of Student and the identification of students with special needs that had distributions severely skewed;

Table 4.2 describes the three main types of data preparation activities performed in order to deal with missing data.

We chose to remove numerical attributes that had at least 50% of missing values, such as the averages in secondary education, to avoid imputing many values and causing disturbances in the distribution of data. For other numerical attributes with missing data, like the ones that represent the minimum and maximum grade obtained by the student in its academic path, we imputed the minimum value locally, looking only at one particular dataset (course), so as to not mix up information between different courses.[3]

For the categorical variables that had missing values, we imputed valid values as category "Not Available", allowing for maintaining the attributes while still showing that the information was not provided by the student, as it happens with most of these variables. While this choice of treatment comes with the risk of altering perceptions of patterns present in the data, it was the approach that showed the best results in preliminary tests.

Table 4.3 shows the classes attributes and some of the attributes belonging to each class. For a more exhaustive description of each attribute, refer to the appendix and tables A.3 and A.4.

[3] See attachment A.3 for a more complete description of treatment applied to each attribute

All predictive attributes were split in four classes, in similar way to studies such as Manhães, 2020[81], Moriconi, 2014[32], and Tomasevic, 2020[86]. This is done usually to compare the predictive power and informational value of each class to determine which is the most effective in predicting student failure, as it is also done in this research. The classes defined for this study are: (1) Sociodemographic; (2) Admission; (3) Enrolment (or Student status); and (4) Course-related (or Academic history).

The Sociodemographic class include a wide range of attributes, which represent students' physical attributes, such as age and gender, geographic attributes such as country of birth, nationality and residence, employment status of the student and their parents, educational attainment obtained by the student and their parents, as well as attributes indicating whether the student has applied and was awarded a scholarship grant by the university.

Admission-related attributes specify information related to the process of application and enrolment. Available attributes indicate the position of this programme in the student's order of preference at the time of application, the student's average grades in secondary education and the grade received by the student from the evaluation committee at the moment of the application for this programme.

Enrolment variables represent the status of the student in the current period. It is a class that represents information more used administratively such as the status of enrolment, and whether the student enjoys the prerogatives of a differentiated status for being employed while completing his studies.

The most comprehensive group of attributes is that of path-related variables, usually called Academic History in most studies[81],[32]. It contains 15 variables that represent the progress of the student at the moment of each of their enrolments. The cumulative numerical average, the number of credits the student has completed, the number of courses or years they are behind schedule, and the lowest and highest grade obtained by the student are among the attributes available in this set.

The next step in the study was splitting the main dataset into several specific datasets, one for each Course. This activity resulted initially in 5729 datasets, lately filtered based on some conditions related to the scope of the study. We filtered out: (1) Courses with names containing "dissertation", "thesis", and "internship"; (2) Courses with less than 100 students throughout all years; (3) Courses not offered in 2019, the last year of data available. This immediately reduced the number of courses from 5729 to 1702, so that only courses really suitable for the task at hand would be studied.

Lastly, the implementation of the Random Forest Classification algorithm for R programming language available in package *caret* was trained for all 1702 Course datasets, to ensure each course had its own self-contained predictive effort. Only information for that particular course was used to train and to test the models, especially to avoid data leakage.

4.1.1 Feature Selection

As identifying relevant factors for student success was one of the most important undertakings of this study, we cross-referenced different sets of predictive attributes, in order to guarantee we

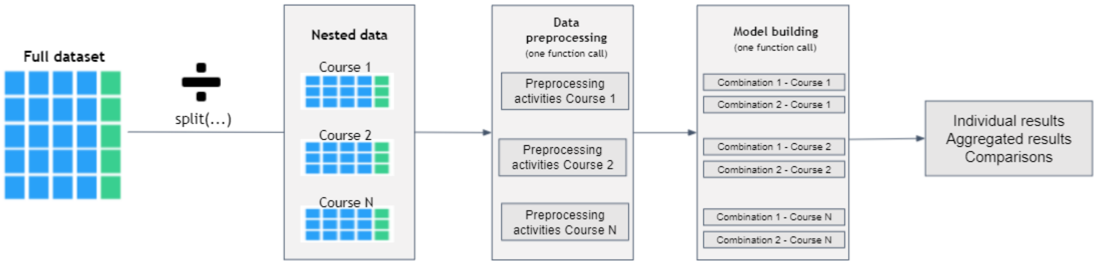


Figure 4.2: Overview of workflow for the predictive stage of the experiment

would always achieve the best predictive modelling performance given the data available. For this reason, we performed Feature Selection, to identify a reduced subset of variables that would be the best for predicting academic success, and then, we tested this subset against the classes of attributes previously defined to verify which was the best set of attributes overall.

The first part of the Feature Selection stage consisted of identifying and eliminating highly correlated variables. The Pearson coefficient of correlation was measured between each numerical predictor, and a few pairs of variables presented extremely a high correlation coefficient with each other, as listed on table 4.4.

variable1	variable2	cor
CREDITS_COMPLETED	COURSES_COMPLETED	0.958
PROGRAMME_COMPLETION	CREDITS_COMPLETED	0.920
CREDITS_COMPLETED	CREDITS_APPROVED	0.906
COURSES_COMPLETED	CREDITS_APPROVED	0.877
PROGRAMME_COMPLETION	COURSES_COMPLETED	0.859
PROGRAMME_COMPLETION	CREDITS_APPROVED	0.856

Table 4.4: Highly correlated variables

Attributes CREDITS_COMPLETED, COURSES_COMPLETED, CREDITS_APPROVED, and PROGRAMME_COMPLETION all provide information about the student's path in similar terms, so after a few tries of removing some of those variables and repeating the correlation analysis on the resultant subset, the best option was found to be the removal of the variables representing the number of ECTS credits completed, the number of Course Units completed and the number of ECTS credits the student has been approved on. After this process, no pair of attributes presented a correlation coefficient of more than 0.7. Subsequently, we selected a stratified sample of 200 courses representing all faculties, curricular years, and programmes and ran Recursive Feature Elimination (RFE), a backward Feature Selection method with an underlying Random Forest algorithm for each of those courses. The output of this process was a list of features ranked by importance for each course. All scores were then aggregated and the 20 features that showed the highest mean importance score would be selected as the top 20

Attribute	**Class**	**Mean Rank**
Delayed Courses	Course	2.75
Average 1st Year	Course	5.63
Average Approved Courses	Course	6
Programme Completion	Course	6.31
Enrolled Courses	Course	7.82
Average 1st year 1st semester	Course	9.06
Delayed Years	Course	10.21
Age	Sociodemographic	11.42
Credits enrolled in semester	Course	11.98
Application Ranking	Admission	12.13
Number Enrollments in Course	Course	12.47
Courses enrolled in semester	Course	13.05
Highest grade	Course	13.75
Admission Regime	Admission	14.36
Debt Situation	Enrollment	16.23
Application Preference	Admission	17.27
Educational Level Parent 1	Sociodemographic	19.12
Occupation Parent 2	Sociodemographic	19.16
Educational Level Parent 2	Sociodemographic	19.44
Occupation Student	Sociodemographic	19.67

Table 4.5: Variables chosen by Feature Selection method

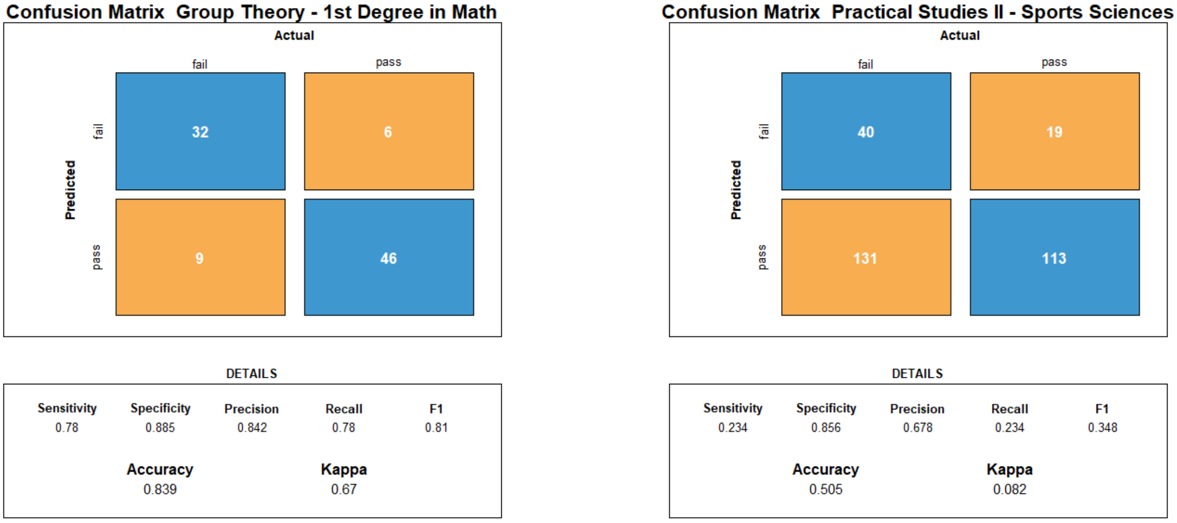
variables for the purpose of this task.

Table 4.5 shows the 20 most important variables, along with the average ranking of importance. Features representing the number of Delayed Courses, average grade on the 1st year, cumulative average grade on all years, rate of program completion and number of courses enrolled were found to be the most influential according to the data used.

4.2 Results

After training and evaluating predictive models, we aggregated results obtained for each of the major Machine Learning evaluation measures and show the average results for those metrics in table 4.7. The best performance overall was achieved by combination SAEC, i.e. when training the models with all 35 variables available. Although it is a model with a larger set of variables than the other combinations, it did not require much additional time to train it.

In addition to an overview of aggregated results, it is also important to investigate results of individual models trained for some courses in particular. It is clear that there is great diversity in performance between courses. For some courses, the models performed particularly well, as they did, for instance, in a course named Group Theory, part of the curriculum of the First Degree in



(a) Good predictive effectiveness

(b) Bad predictive effectiveness

Figure 4.3: Confusion matrices obtained for a model with good and bad results

Result	Mean Grade
True Positives	5.54
False Negatives	6.84
False Positives	12.30
True Negatives	14.23

Table 4.6: Mean grades per result obtained by the predictions

Mathematics, as seen in figure 4.3a. In this particular case, most instances of the test set were correctly predicted resulting in a good looking Confusion Matrix, Precision at 0.942 and Recall at 0.780.

In contrast, models were not accurate for all courses, such as the model obtained for "Practical Studies II", part of the curriculum of the First Degree in Sports Sciences, as show in figure 4.3b. Most negative instances of the target variable were correctly predicted, but the model did not do a good job of predicting failure. Most of the cases predicted as failure by the model were correct, hence the Precision at 0.678, but it only identified 40 of 171 positive instances on the test set, resulting in a 0.234 Recall.

Looking to summarize and evaluate predictions and having evaluated all models through Confusion Matrices and common Machine Learning evaluation metrics, we looked to compare the result of each prediction with the actual grade obtained by the student. While the actual grade obtained by the student was available in the data, it was removed from all other facets of this study so we would not incur in data leakage. Table 4.6 shows mean grades for each type of predictive result: True Positives, False Negatives, False Positives and True Negatives.

It is clear that, on average, the models do a good job of identifying mostly students with the lowest and the highest grades, as shown by the average grades for True Positives and True

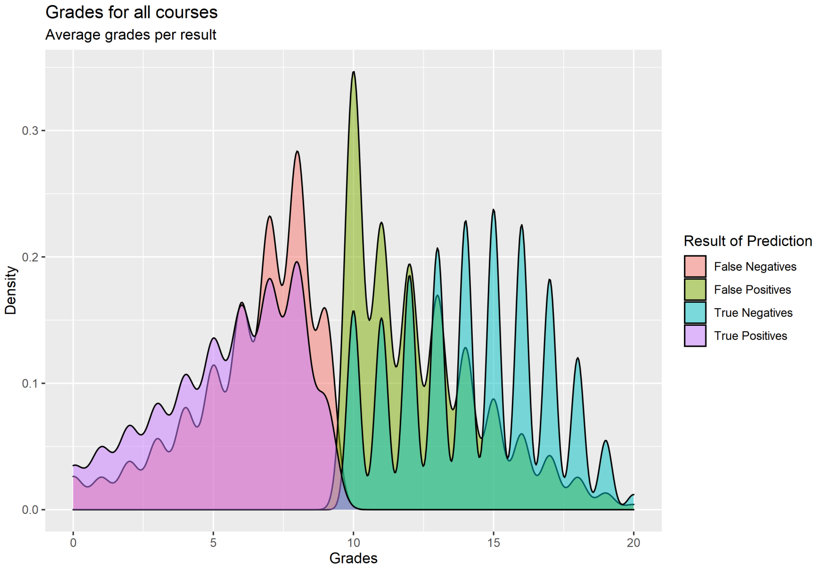


Figure 4.4: Distribution of grades per type of result

Negatives. Somewhere in the middle lies the difficulty, i.e. the students who obtain borderline grades around the necessary 10 for approval. Average grades of 6.84 and 12.30 for False Negatives and False Positives respectively, show that the models have a lot of room for improvement in quality of predictions.

Looking beyond the mean values, we built four distributions of occurrences of grades for each type of result, shown in figure 4.4 that clearly confirms the fact that the models performed better on students who ended up obtaining lower and higher grades. Most students who achieved grades 8 through 12 were not correctly identified, as also shown in the same figure. While those are inherently the most difficult cases to predict, those are also key for the management of higher education institutions, as a little aid, encouragement, or counseling could help a student obtain an approval when they otherwise would not.

Additionally, we looked to identify if the models had better quality of predictions based on the size of the courses, i.e. the number of students enrolled over all years of data. As shown in figure 4.5, the distributions for each type of result of predictions for the 25% of courses with the smallest number of students look very different than the overall results, with many more incorrectly identified students in all grades, especially in the borderline grades, as discussed.

The results indicate some degree of disparity in quality of predictions, with some models doing a great job at identifying students at risk of failure and others not being able to do the same. This goes to show that this type of predictive effort should be well validated and if implemented, great consideration should be taken for the specificity of each Scientific area, programme or course.

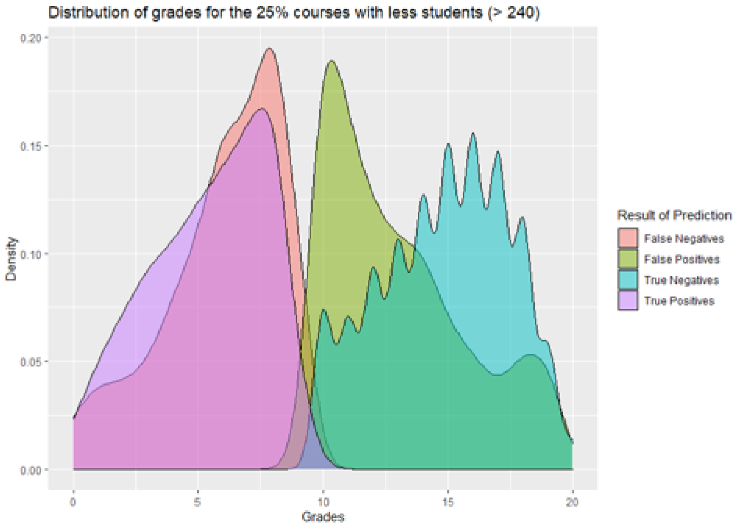


Figure 4.5: Distribution of grades for courses with the fewer students

4.3 Research Questions

4.3.1 Research Question 1: What is the best combination of classes of attributes to predict academic insuccess?

Predictive attributes were divided in four classes: Sociodemographic, Admission, Enrollment and Academic History and one of the goals of this study was to determine how informative each of those classes is for predicting academic in success with available data. Classification models were trained with various combinations of classes to determine the best one, and additionally, whether this classes of attributes would outperform the results obtained by the top 20 variables as determined by Feature Selection methods.

Combination	Accuracy	kappa	f1	Recall	spec	Precision	Time to run
C	0.834	0.291	0.453	0.528	0.864	0.393	1h44m
SC	0.833	0.341	0.491	0.524	0.876	0.467	1h57m
AC	0.831	0.312	0.469	0.526	0.871	0.433	1h48m
EC	**0.839**	0.326	0.481	**0.536**	0.867	0.442	1h37m
SAC	0.833	0.345	0.493	0.527	0.878	0.478	2h14m
SEC	0.835	0.353	0.504	0.533	0.879	0.486	2h09m
AEC	0.833	0.327	0.487	0.535	0.874	0.449	1h53m
SAEC	0.836	**0.359**	**0.508**	**0.536**	**0.881**	**0.496**	2h28m
FSEL	0.831	0.334	0.492	0.524	0.878	0.472	2h01m

Table 4.7: Evaluation Metrics for models on each combination

Deciding which model performed best in a data-oriented process is rarely a unanimous decision, as there are several common trade-offs when training predictive models. Often a model that

Group	Variable
Sociodemographic	age, sex, marital_status, country_nationality, country_birth, country_official_residence, country_classes_residence, foreign_student, displaced_student, educational_level_student, educational_level_parent1, educational_level_parent2, occupation_student,occupation_parent1, occupation_parent2, special_needs_student, applied_scholarship, has_scholarship
Admission	admission_regime, application_preference, application_ranking
Enrolment	dedication_regime, debt_situation, status_student, working_student
Course-Related	average_1st_year_1st_semester, average_1st_year, average_approved_courses, highest_grade, lowest_grade, number_enrollments_in_course, courses_enrolled_in_semester, credits_enrolled_in_semester, credits_approved, delayed_courses, delayed_years, enrolled_courses, courses_completed, credits_completed, programme_completion

Table 4.8: Best subset of variables

predicts positive cases well will not do as well with negative cases and therefore a model that has good results in Specificity, a metric that is used to evaluate how well the model predicts instances of the negative class, may not show the same predictive power when it comes to Precision or Recall, measures that give more weight to the performance in relation to the positive cases.

The combination of attributes related to the student's academic history, identified with the abbreviation C, provided the best performance in an individual evaluation of the classes of variables, showing a better performance than each of the other classes alone. Using this class as basis, the combinations of 2, 3 and 4 classes were evaluated, for a total of 8 combinations.

As shown in table 4.7, the combination of classes of variables that resulted in the best performance was SAEC[4], with all its 35 attributes shown in table 4.8. Objectively, while the difference in the average of each metric between different combinations is not particularly large, SAEC models still outperformed all others, obtaining better results in Kappa, Precision, Recall, and Specificity with margins of 3 to 5 percent on each evaluation metric.

4.3.2 Research Question 2: How accurately can we identify what students are going to fail at a course at the beginning of the semester?

From the entire set of variables made available for this study, all attributes with the exception of two are available at the start of the school period, and those attributes are: the number of the courses the student is enrolled in the current semester and the number of times the student already enrolled in that particular course. That allows for prematurely identify students at risk of failure, since the university has most of the data early in the semester.

[4] Sociodemographic, Admission, Enrolment and Course-related attributes

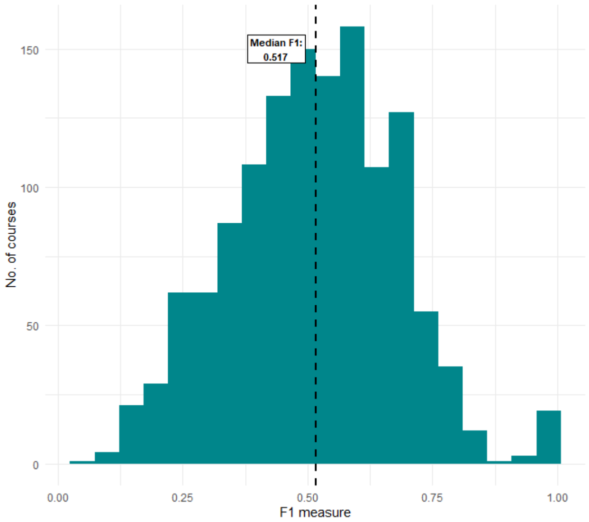


Figure 4.6: Histogram of F1-measure throughout all courses

As seen in table 4.9, a median Recall value of 0.545 indicates that for 50% of the courses, we can correctly identify 5 students out of each 10 that will fail in their courses. In addition to that, a closer look at the distribution of each evaluation measure and to see how the models performed better for some courses than for others.

Table 4.9 and figure 4.6 show those distributions and some elements stand out. (1) For 20% of the courses, the Kappa coefficient obtained by the best model was negative, meaning poor performance in relation to the expectation. (2) With such imbalanced underlying distributions, Specificity is high with 75% of the courses achieving at least 0.821. (3) 50% of the courses obtained Precision above 0.444 and Recall above 0.545. (4) The distribution of F1 measures throughout all courses, as seen in figure 4.6 approximately follows a Normal distribution, and shows a healthy number of courses as having F1-measures of 0.6 and up.

For additional context, this time in relation to the class imbalance of the target variable, it can be seen in table 4.10 that the results change according to different distribution of the classes of the target variable. Courses with more uniform distributions of the target variable usually had better balance in the predictions. On the other hand, it proved more difficult to obtain balanced positive and negative predictions in courses with extremely unbalanced approval rates.

Metric	Min.	1st Qu.	**Median**	Mean	3rd Qu.	Max.
Kappa	-0.286	0.130	**0.322**	0.305	0.461	1
Precision	0.000	0.250	**0.444**	0.429	0.625	1
Recall	0.000	0.400	**0.545**	0.530	0.680	1
F1	0.067	0.400	**0.517**	0.511	0.621	1
Specificity	0.000	0.821	**0.900**	0.869	0.952	1

Table 4.9: Distribution of evaluation measures for the best model on all courses

According to table 4.10 above, the courses that achieved the best results in Accuracy and

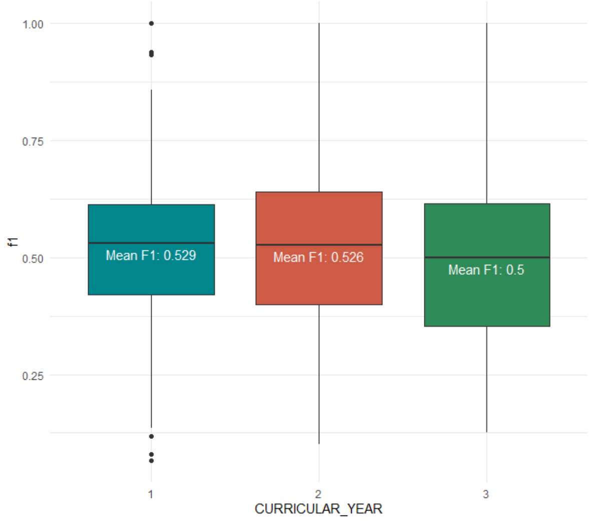


Figure 4.7: Boxplot of F1-measure throughout all courses per Curricular Year

Specificity were the ones that had almost exclusively cases of the negative class - represented by the last group. On the other hand, the ones that had the best Recall, Precision and F1, were the ones with mostly positive cases. The Kappa coefficient, as a more balanced measure of effectiveness between both classes, had the best results in the middle, where there are more balance in the distribution of the target variable.

Approval Rate	No. of Courses	Accuracy	Kappa	Recall	Specificity	Precision	F1
0.0 - 0.5	84	0.632	0.262	**0.594**	0.694	**0.745**	**0.641**
0.5 - 0.7	407	0.717	0.345	0.575	0.788	0.575	0.548
0.7 - 0.8	386	0.803	**0.361**	0.546	0.861	0.474	0.504
0.8 - 0.9	489	0.872	0.305	0.500	0.908	0.348	0.463
0.9 - 1	381	**0.943**	0.204	0.451	**0.956**	0.216	0.472

Table 4.10: Comparison of measures obtained by best model for all approval rate groups

4.3.3 Research Question 3: Are there any differences in the quality of predictions based on the curricular year?

Another meaningful dimension with which we can analyze the results of the predictive models is that of the Curricular Year, since a 1st-year student will behave differently than a student from the 2nd or 3rd year, and thus, require different measures to improve their decision ability in terms of their educational future. The measure of choice for this review was the F1-measure, since it is the harmonic mean of Precision and the Recall in one metric, two of the most important individual Machine Learning metrics as defined in this study.

Figure 4.7 shows mainly two things. Primarily, F1 results for 1st-year and 2nd-year courses were almost identical, on average, with a difference of only 0.003. Furthermore, there seems to be a clear difference between the performance obtained in the first two years in comparison with the 3rd year, perhaps indicating there is more uncertainty that could not be captured by the features available for this study.

To investigate the possibility of difference between F1 measures obtained for courses of each curricular year, we conducted two runs of the Wilcoxon rank sum test[5] between (1) the first and the second curricular years, where the 0.003 difference is very subtle, and; (2) between the 1st and the 3rd curricular years, where there is a larger difference of 0.3. For the first test, a W statistic of 120328 and a p-value of 0.843, well above the 0.05 significance level, mean we cannot reject the null hypothesis that there is not a difference in medians between year 1 and year 2.

In the second case, with a W statistic of 103743 and a p-value of less than 0.01, we can confidently reject the null hypothesis in favor of the alternative hypothesis, which is, there is really a difference in median F1 values obtained for 2nd-year and 3-year courses.

Two possible reasons for the better results obtained predicting outcomes for courses from the first and second years were: (1) a higher average number of observations available for 1st-year and 2nd-year courses with 680 and 546 observations respectively, while the third year had only 462 on average, a drop of 30% from the first year; (2) approval rate is lower for first and second-year courses at 69.6% and 76.2% when comparing with the 83.9% of the third year, meaning that, on average, the distribution of the target variable is more balanced in the first two years.

While those are not decisive characteristics by themselves, a larger number of training observations and better balanced distributions of the target variable are aspects that usually help obtain good results when training predictive models.

4.3.4 Research Question 4: Are there any differences in the quality of predictions based on scientific area, programme or school?

When aggregating results obtained by the best combination of variables for each of the available dimensions, it is clear that there is a great heterogeneity between levels such as scientific area, programme, school, and curricular year. Thorough follow-up analysis of each of these groups is imperative for making any decisions based on the generated predictions.

Scientific Area	Mean F1	Courses	Rank	Approval %	Students
Applied Mechanics	0.677	6	1	56.3	1536
Engineering Sciences	0.673	6	2	64.5	514
Specialty Sciences	0.625	9	3	82.8	846
...	...	...	...		
Landscape Architecture	0.364	15	73	79.7	175
Fine Arts - Sculpture	0.333	6	74	85.4	195
Biomedical Engineering	0.286	9	75	90.8	230

Table 4.11: Scientific Areas with best and worst F1 measures

[5] Non-parametric test to determine difference in median between two independent samples

Using the F1 metric once more, the results obtained by the best combination of variables were aggregated by scientific area, and table 4.11 presents the highest and lowest values for this measure[6]. This indicates in which areas the predictive models achieved the best and worst aggregate quality of predictions in terms of F1.

Table 4.12 shows the results aggregated per School. Considerable variance is apparent when comparing different faculties: Faculdade de Direito da Universidade do Porto (FDUP) was the school with the best average F1 score at 0.559 and Faculdade de Farmácia da Universidade do Porto (FFUP) had the worst score, at 0.446.

School	Mean F1	Courses	Rank	Approval %	Students
FDUP	0.559	52	1	75.8	809
FEP	0.527	80	2	76.1	1590
FCUP	0.527	338	3	77.4	341
...	...	...	...		
FAUP	0.453	27	12	84.5	871
FMDUP	0.449	48	13	83.1	645
FFUP	0.446	34	14	80.8	1573

Table 4.12: Schools with best and worst F1 measures

Table 4.13 shows average F1 values aggregated per each of the 56 programmes evaluated. The programme with the best results in terms of F1 was the First Degree in Geology, part of Faculdade de Ciências da Universidade do Porto (FCUP) with 0.631 F1 on average, and the worst was the Master's Degree in Medicine offered by Instituto de Ciências Biomédicas Abel Salazar (ICBAS), at 0.373 average F1.

Programme	F1	Courses	Rank	Approval %	Students
First Degree in Geology	0.631	26	1	69.5	247
First Degree in Engineering Sciences	0.613	35	2	64.2	315
First Degree in Mathematics	0.596	27	3	58.8	173
...	...	...	...	...	
Bachelor in History of Art	0.441	12	54	77.8	247
First Degree Landscape Architecture	0.424	26	55	89.3	789
Master's in Medicine	0.373	27	56	73.8	189

Table 4.13: Programmes with best and worst F1 measures

Three instance of the Kruskal-Wallis test[7] were performed to identify if the differences in F1 values for each group are indeed statistically relevant. Table 4.14 shows results of the three

[6] Only Scientific areas with more than 5 courses were considered for this analysis to ensure proper representativeness

[7] Non-parametric test used to determine if the difference in medians between three or more groups of the same attribute are statistically significant

Group	Chi-squared Statistic	Degrees of freedom	p-value
Scientific Area	263.7	153	<0.01
Programme	132.4	50	<0.01
School	47.1	13	<0.01

Table 4.14: Result of Kruskal-Wallis tests between groups

Kruskal-Wallis tests performed for each attribute and p-values smaller than 0.01 for every test indicate with statistical confidence that there are significant differences in quality of predictions between Scientific Areas, Programmes and Schools.

4.3.5 Research Question 5: What are the most important individual attributes for predicting cases of academic insuccess in higher education institutions?

To answer this Research Question, results obtained by training the models using combination of variables named *SAEC*, the one that contains all predictor variables, since it was the combination that generated the most effective predictions. Then, Feature Importance scores obtained by each Random Forest model were aggregated for all courses, to obtain the average importance of all variables, shown in table 4.15 below. Values are normalized and scaled on a range from 0 to 100 for convenience and better presentation.

The number of courses that the student is delayed on, the percentage of programme completion, the student's cumulative average, the average grade obtained by the student in the 1st year and the number of courses to which the student is enrolled in at the present school term were considered the 5 most important attributes when averaging local importance scores for all courses. This outcome reaffirms results found on the Feature Selection task, where several of the same attributes had also been considered relevant, especially attributes related to the academic history of the student.

Rank	Variable	Mean Importance
1	DELAYED_COURSES	82.21
2	PROGRAMME_COMPLETION	57.71
3	ENROLLED_COURSES	43.30
4	DELAYED_YEARS	36.35
5	CREDITS_ENROLLED_IN_SEMESTER	35.92
6	HIGHEST_GRADE	33.27
7	COURSES_ENROLLED_IN_SEMESTER	30.66
8	NUMBER_ENROLLMENTS_IN_COURSE	27.39
9	APPLICATION_PREFERENCE	24.49
10	SEX	24.36
11	LOWEST_GRADE	23.22
12	HAS_SCHOLARSHIP	21.98
13	OCCUPATION_STUDENT	21.40
14	COUNTRY_BIRTH	20.57
15	APPLIED_SCHOLARSHIP	20.50

Table 4.15: Mean aggregated variable importance - top variables

Looking to identify if the same attributes remained important throughout courses from all curricular years, follow-up analysis shows that the five most important variables for all courses are also highly influential on a year-by-year basis.

As it can be seen in the table 4.16, the variable Delayed Courses was the most important on courses from all curricular years. Some features are also important for courses of all years such as the percentage of programme completion, and the number of courses the student has enrolled in the current semesters. Other variables seem to have more predictive power for first-year courses, especially sociodemographic variables such as Age and Gender, which are the second and seventh most important in the first year, but considered less important in the following years.

Feature	Year1	Rank	Year2	Rank	Year3	Rank	Mean
DELAYED_COURSES	77.98	1	85.38	1	83.11	1	82.16
PROGRAMME_COMPLETION	40.02	5	63.86	2	69.60	2	57.83
AGE	60.83	3	35.91	4	34.90	5	43.88
ENROLLED_COURSES	44.95	4	44.02	3	40.72	3	43.23
APPLICATION_RANKING	62.60	2	27.89	8	25.74	8	38.74
CREDITS_ENROLLED	37.09	6	34.87	5	35.85	4	35.94
HIGHEST_GRADE	32.73	9	32.65	6	34.53	6	33.30
COURSES_ENROLLED	32.46	10	29.06	7	30.55	7	30.69
NUMBER_OF_ENROLLMENTS	29.17	13	27.21	9	25.70	9	27.36
APPLICATION_PREFERENCE	34.10	8	18.50	13	21.02	11	24.54
SEX	34.34	7	20.52	11	18.06	20	24.31
LOWEST_GRADE	25.72	18	20.83	10	23.26	10	23.27
SCHOLARSHIP	31.13	11	16.80	18	18.10	19	22.01
STUDENT_OCCUPATION	25.35	21	20.32	12	18.41	14	21.36
COUNTRY_BIRTH	28.32	14	15.81	22	17.69	24	20.61

Table 4.16: Aggregated feature importance per year

Chapter 5

Conclusion

In this work, it was investigated how data stored in academic record systems can be transformed into potentially useful information to support decision-making in order to reduce school failure in higher education, using data mining techniques. The main objectives included identifying: (1) students at risk of insuccess failing in the curricular units to which they are enrolled; and (2) the most influential attributes for correct and timely identification of these students.

Unlike most recent works on student performance prediction, for this study, predictive models were trained for the entire curriculum structure of the Universidade do Porto. Training course-specific models allowed for an individual analysis of factors related to academic failure in each Course, which can also be used by education managers to monitor performance and the occurrence of failure within various dimensions: course, programme, scientific area and faculty, among others.

Preparing, training and evaluating this amount of data and this number of predictive models was certainly a daunting task, but we can draw interesting conclusions from the results, that certainly provide food for thought for next endeavours, be it at this institution or others that seek to implement an institution-wide predictive effort.

Regarding the five Research Questions proposed in this study, we were able to find the best possible answers according to the data available and methods used. In Research Question 14.3.1, we identified that among all combinations of classes of variables, the best individual for predicting student performance was the one with variables with the student's academic history. In addition to that, the combination identified as *SAEC* containing all 35 predictor variables was the one that generated the best quality of predictions, although the difference between combinations was never decisive.

Predictions of student performance can be made early, at the start of the semester. In Research Question 24.3.2, we can correctly identify 5 students out of each 10 that will fail in their courses for 50% of the courses and this makes it possible to set up alerts and interventions to support these students. Although 5 out of 10 students identified as failing would not need this support for obtaining approval, the additional support provided by the education institution could consolidate learning and over time, shorten the student's path to obtaining their academic degree, and improve motivation and knowledge retention.

Greatest predictive ability was achieved for courses (1) of the 1st and 2nd curricular years; (2) with more students over the years; (3) with more balance in the target variable. Those three findings are supported by data, as seen in Research Question 34.3.3, and while their mere discovery has informational value on its own, this could certainly act as guidelines for implementation of such models in this or another higher education institution.

In Research Question 3 it was identified that courses from the first and second curricular years achieved better quality of predictions than the 3rd-year courses. Subsequently, in Research Question 44.3.4, statistical tests showed that there is definitively difference in quality of predictions between Scientific Areas, Faculties and Programmes.

As shown in Research Question 54.3.5, predictive variables such as the number of delayed courses, the percentage of progress within the course and the number of Course Units enrolled in the semester are constantly important, regardless of the area, course or curricular year. In addition to a strong predictive power, such variables are easily trackable by education managers, thus allowing a subsequent implementation of the information generated by the predictive models for the monitoring of at-risk students and the development of motivational or pedagogical measures.

Although it is out of the scope of this work to evaluate subjective dimensions and teaching and learning processes that lead to failure in higher education, we hope that the numbers and results presented provide positive momentum and that can foster pedagogical and educational changes in higher education institutions to better accommodate and meet the needs of students in difficulty. Moreover, it is our hope that the results of this study and many other studies of academic performance are not confined to a single institution, so it is difficult to prove the applicability of a study with such implicitly divergent databases and attributes.

5.1 Study limitations

One of the strengths of the study, but one that also proved to be one of the biggest difficulties in its development, was the amount of data. In technical terms, the large number of datasets and observations in general made the execution of the data preparation and model building algorithms somewhat time-consuming, making iterations and iterations of the algorithms take a few hours to execute. However, this is something reasonably expected when it comes to such a comprehensive study of the entire training offer of a large higher-level educational institution.

Also regarding the data, although we had the collaboration of knowledgeable professionals from the University of Porto for verifying the availability and quality of data available, something was left to be desired in terms of data completeness with a few variables having to be removed because of the high percentage of missing values. We were able to observe that the data quality improved significantly after the year 2014, and after investigations carried out with university personnel, we found that improvements were implemented at that time in the institution's student management software.

Even so, some variables that could be influential and beneficial in predicting academic performance, such as the average grade of students in secondary education, had to be removed because of its number of missing values. Others, such as the variables that represented the occupation and educational level of the students' parents, had a fair amount of missing values, and some data imputation had to be carried out to leverage some kind of information out of these variables in the model building stage.

Conclusively, we believe that the results achieved were quite satisfactory by and large, and the limitations found were not detrimental to achieving the desired goals. As for the possible lack of depth in the study, more specific information can certainly be extracted *as-is* from the models built, without the need for much adaptation or additional work.

5.2 Future work

The study of attributes that lead to academic insuccess traditionally requires extensive and complex work due to its multidimensionality and the number of factors that are usually out of the control of the researchers. Although it was possible to draw interesting and hopefully actionable conclusions, the Universidade do Porto and many other Portuguese and European universities would benefit from follow-up studies in this field, especially with the complexity of these issues.

With the data made available by the University for this study, it was only possible to explore relationships and dive into the issue of academic insuccess. As explored by Kaplan, 1997[4] among other authors, continued academic insuccess and the occurrence of negative academic experiences can lead to insatisfaction, lack of motivation and feelings of inadequacy, which can explain student dropout to a large extent. There is a lack of studies in the literature considering both subjects, insuccess and dropout, so a follow-up study looking to link both events would certainly be very beneficial for the educational community.

As noted by Huang and Fang[27], learning is an extremely complex process involving many psychological factors such as learning styles, self-efficacy, achievement goals, motivation, interest, and teaching and learning environment. Fundamentally, obtaining and combining data from several sources regarding many aspects of the student life and journey would also make for a great addition for follow-up studies, as many done recently. Data on student participation in online activities and assessments in a Learning Management System (LMS) system[91][48][79], self-reports of interest, motivation and level of preparedness[64][87][42], and more data on financial background and current economic situation[42][54] are examples of supplementary data that would add much information - and possibly predictive ability - to the findings already made in this study.

Interpretability and transparency in Machine Learning are also critical aspects of building predictive models. There is a growing desire in the intelligence and decision-making sectors of many fields today[33][83] to have clear and easy-to-interpret insights so that all stakeholders can use this information, even those without deep technology knowledge. The continuation of studies in the scope of interpretability would certainly be of great value to the Universidade do Porto and the academic community as a whole.

Lastly, follow-up studies on the performance of the predictive models when applied in a real-world setting would also desirable. Ever-changing patterns and relationships and new dynamics in the educational environment due to the Covid pandemic should merit continuing studies and model evaluation in the following years.

5.3 Conflicts of Interest

On behalf of all authors, the corresponding author states that there is no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Bibliography

[1] William G. Spady. "Dropouts from higher education: An interdisciplinary review and synthesis". In: *Interchange* 1.1 (Apr. 1970), pp. 64–85. ISSN: 0826-4805. DOI: 10.1007/BF02214313. URL: http://link.springer.com/10.1007/BF02214313.

[2] Vincent Tinto. "Dropout from Higher Education: A Theoretical Synthesis of Recent Research". In: *Rev. Educ. Res.* 45.1 (Mar. 1975), pp. 89–125. ISSN: 0034-6543. DOI: 10.3102/00346543045001089. URL: http://journals.sagepub.com/doi/10.3102/00346543045001089.

[3] J. Richard Landis and Gary G. Koch. "The Measurement of Observer Agreement for Categorical Data". In: *Biometrics* 33.1 (Mar. 1977), p. 159. ISSN: 0006341X. DOI: 10.2307/2529310. URL: https://www.jstor.org/stable/2529310?origin=crossref.

[4] Diane S. Kaplan, B. Mitchell Peck, and Howard B. Kaplan. "Decomposing the academic failure–dropout relationship: A longitudinal analysis". en. In: *J. Educ. Res.* 90.6 (July 1997), pp. 331–343. ISSN: 19400675. DOI: 10.1080/00220671.1997.10544591.

[5] Isabelle Iguyon and André Elisseeff. "An introduction to variable and feature selection". en. In: *J. Mach. Learn. Res.* 3 (2003), pp. 1157–1182. ISSN: 15324435. DOI: 10.1162/153244303322753616. URL: http://www.crossref.org/deleted%7B%5C_%7DDOI.html.

[6] Laurie P. Dringus and Timothy Ellis. "Using data mining as a strategy for assessing asynchronous discussion forums". In: *Comput. Educ.* 45.1 (Aug. 2005), pp. 141–160. ISSN: 03601315. DOI: 10.1016/j.compedu.2004.05.003. URL: https://linkinghub.elsevier.com/retrieve/pii/S0360131504000788.

[7] Shane Dawson, Bruce Burnett, and Mark O'Donohue. "Learning communities: an untapped sustainable competitive advantage for higher education". In: *Int. J. Educ. Manag.* 20.2 (Feb. 2006), pp. 127–139. ISSN: 0951-354X. DOI: 10.1108/09513540610646118. URL: https://www.emerald.com/insight/content/doi/10.1108/09513540610646118/full/html.

[8] Ormond Simpson. "Preprint PREDICTING STUDENT SUCCESS 2006". In: 21 (2006), pp. 125–138.

[9] Jennifer Rowley. "The wisdom hierarchy: Representations of the DIKW hierarchy". In: *J. Inf. Sci.* 33.2 (2007), pp. 163–180. ISSN: 01655515. DOI: 10.1177/0165551506070706.

[10] J.-P. Vandamme, N. Meskens, and J.-F. Superby. "Predicting Academic Performance by Data Mining Methods". en. In: *Educ. Econ.* 15.4 (Dec. 2007), pp. 405–419. ISSN: 0964-5292. DOI: 10.1080/09645290701409939. URL: http://www.tandfonline.com/doi/abs/10.1080/09645290701409939.

[11] Catherine Finnegan, Libby V. Morris, and Kangjoo Lee. “Differences by course discipline on student behavior, persistence, and achievement in online courses of undergraduate general education”. In: *J. Coll. Student Retent. Res. Theory Pract.* 10.1 (2008), pp. 39–54. DOI: 10.2190/CS.10.1.d. URL: http://journals.sagepub.com/doi/10.2190/CS.10.1.d.

[12] Max Kuhn. “Building predictive models in R using the caret package”. In: *J. Stat. Softw.* 28.5 (2008), pp. 1–26. ISSN: 15487660. DOI: 10.18637/jss.v028.i05.

[13] G. G. Moisen. “Classification and Regression Trees”. eng. In: *Encycl. Ecol. Five-Volume Set.* 1. CRC Pre. Boca Raton, Fla.: Chapman & Hall/CRC, 2008, pp. 582–588. ISBN: 9780080914565. DOI: 10.1016/B978-008045405-4.00149-X.

[14] Elena Arias Ortiz and Catherine Dehon. “What are the factors of success at University? A case study in Belgium”. In: *CESifo Econ. Stud.* 54.2 (May 2008), pp. 121–148. ISSN: 1610241X. DOI: 10.1093/cesifo/ifn012. URL: https://academic.oup.com/cesifo/article-lookup/doi/10.1093/cesifo/ifn012.

[15] Leah P. Macfadyen and Shane Dawson. “Mining LMS data to develop an "early warning system" for educators: A proof of concept”. en. In: *Comput. Educ.* 54.2 (2010), pp. 588–599. ISSN: 03601315. DOI: 10.1016/j.compedu.2009.09.008.

[16] Ying Zhang et al. “Use data mining to improve student retention in higher education: A case study”. en. In: *ICEIS 2010 - Proc. 12th Int. Conf. Enterp. Inf. Syst.* Vol. 1 DISI. Funchal, Madeira, Portugal: SciTePress - Science, 2010, pp. 190–197. ISBN: 9789898425041. DOI: 10.5220/0002894101900197. URL: http://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0002894101900197.

[17] Ryan Baker, Seiji Isotani, and Adriana Carvalho. “Mineração de Dados Educacionais: Oportunidades para o Brasil”. In: *Rev. Bras. Informática na Educ.* 19.02 (Aug. 2011). ISSN: 1414-5685. DOI: 10.5753/rbie.2011.19.02.03. URL: http://www.br-ie.org/pub/index.php/rbie/article/view/1301.

[18] Aaron M. Kuntz, Ryan Evely Gildersleeve, and Penny A. Pasque. “Obama’s American Graduation Initiative”. In: *Peabody J. Educ.* 86.5 (Nov. 2011), pp. 488–505. DOI: 10.1080/0161956X.2011.616130.

[19] Oded Rokach, Lior; Maimon. *Data mining with decision trees.* Vol. 81. Series in Machine Perception and Artificial Intelligence. 2011. DOI: 10.1142/9097. URL: https://www.worldscientific.com/worldscibooks/10.1142/9097.

[20] Michael Fire et al. “Predicting Student Exam’s Scores by Analyzing Social Network Data”. In: 2012, pp. 584–595. DOI: 10.1007/978-3-642-35236-2_59. URL: http://link.springer.com/10.1007/978-3-642-35236-2%7B%5C_%7D59.

[21] Leonid Grebennikov and Mahsood Shah. “Investigating attrition trendsin order to improvestudent retention”. In: *Qual. Assur. Educ.* 20.3 (2012), pp. 223–236. ISSN: 09684883. DOI: 10.1108/09684881211240295.

[22] Jiawei Han, Micheline Kamber, and Jian Pei. *Data Mining: Concepts and Techniques.* 2012. DOI: 10.1016/C2009-0-61819-5.

[23] Elina Ketonen and Kirsti Lonka. “Do Situational Academic Emotions Predict Academic Outcomes in a Lecture Course?” In: *Procedia - Soc. Behav. Sci.* 69.February 2015 (2012), pp. 1901–1910. ISSN: 18770428. DOI: 10.1016/j.sbspro.2012.12.144.

[24] Lars Buitinck et al. “API design for machine learning software: experiences from the scikit-learn project”. In: (2013), pp. 1–15. arXiv: 1309.0238. URL: http://arxiv.org/abs/1309.0238.

[25] Ana Maria Eyng et al. “Diversidade e padronização nas políticas educacionais: Configurações da convivência escolar”. In: *Ensaio* 21.81 (Dec. 2013), pp. 773–800. ISSN: 01044036. DOI: 10.1590/S0104-40362013000400007. URL: http://www.scielo.br/scielo.php?script=sci%7B%5C_%7Darttext%7B%5C&%7Dpid=S0104-40362013000400007%7B%5C&%7Dlng=pt%7B%5C&%7Dtlng=pt.

[26] Arnon Hershkovitz et al. “Discovery With Models”. In: *Am. Behav. Sci.* 57.10 (Oct. 2013), pp. 1480–1499. ISSN: 0002-7642. DOI: 10.1177/0002764213479365. URL: http://journals.sagepub.com/doi/10.1177/0002764213479365.

[27] Shaobo Huang and Ning Fang. “Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models”. In: *Comput. Educ.* 61.1 (2013), pp. 133–145. DOI: 10.1016/j.compedu.2012.08.015. URL: https://linkinghub.elsevier.com/retrieve/pii/S0360131512002102.

[28] Janie H. Wilson and Rebecca G. Ryan. “Professorâ€“Student Rapport Scale: Six Items Predict Student Outcomes”. In: *Teach. Psychol.* 40.2 (2013), pp. 130–133. ISSN: 15322802. DOI: 10.1177/0098628312475033.

[29] Thais Accioly Baccaro and Gilberto Tadeu Shinyashiki. “Relação entre desempenho no vestibular e rendimento acadêmico no ensino superior”. In: *Rev. Bras. Orientac. Prof.* 15.2 (2014), pp. 165–176. ISSN: 19847270.

[30] Susana Maria Sousa Martins Leite de Faria. “Educational Data Mining e Learning Analytics na melhoria do ensino online”. In: (2014), p. 138. URL: http://repositorioaberto.uab.pt/handle/10400.2/3511.

[31] Victoria López, Alberto Fernández, and Francisco Herrera. “On the importance of the validation technique for classification with imbalanced datasets: Addressing covariate shift when data is skewed”. en. In: *Inf. Sci. (Ny).* 257 (Feb. 2014), pp. 1–13. ISSN: 00200255. DOI: 10.1016/j.ins.2013.09.038. URL: https://www.sciencedirect.com/science/article/pii/S0020025513006804.

[32] Gabriela Miranda Moriconi and Paulo Augusto Meyer Mattos Nascimento. “Fatores associados ao desempenho dos concluintes de Engenharia no Enade 2011”. In: *Estud. em Avaliação Educ.* 25.57 (Apr. 2014), p. 248. DOI: 10.18222/eae255720142831.

[33] Alejandro Peña-Ayala. “Educational data mining: A survey and a data mining-based analysis of recent works”. In: *Expert Syst. Appl.* 41 (2014), pp. 1432–1462. DOI: 10.1016/j.eswa.2013.08.042.

[34] Ashkan Sharabiani et al. "An enhanced bayesian network model for prediction of students' academic performance in engineering programs". In: *IEEE Glob. Eng. Educ. Conf. EDUCON* April (2014), pp. 832–837. DOI: 10.1109/EDUCON.2014.6826192.

[35] Pan-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining*. 2014.

[36] Erin Dunlop Velez. *Drop-Out Epidemic: Understanding the College Drop-Out Population*. Tech. rep. 2014, p. 28.

[37] Charu C. Aggarwal. *Data Mining*. Cham: Springer International Publishing, 2015. ISBN: 978-3-319-14141-1. DOI: 10.1007/978-3-319-14142-8. URL: http://link.springer.com/10.1007/978-3-319-14142-8.

[38] Nidhi Arora and Jatinderkumar R Saini. "A Fuzzy Probabilistic Neural Network for Student?s Academic Performance Prediction". In: *Int. J. Innov. Res. Sci. Eng. Technol.* 2013.9 (2015), pp. 4425–4432. ISSN: 2319-8753.

[39] Michael Mogessie Ashenafi, Giuseppe Riccardi, and Marco Ronchetti. "Predicting students' final exam scores from their course activities". In: *2015 IEEE Front. Educ. Conf.* IEEE, Oct. 2015, pp. 1–9. ISBN: 978-1-4799-8454-1. DOI: 10.1109/FIE.2015.7344081. URL: http://ieeexplore.ieee.org/document/7344081/.

[40] Pedro Belo and Catarina Oliveira. "The Relation between Experiences and Expectations with University Dropout". In: *Procedia - Soc. Behav. Sci.* 187 (May 2015), pp. 98–101.

[41] Rajeev Bukralia, Amit V. Deokar, and Surendra Sarnikar. "Using Academic Analytics to Predict Dropout Risk in E-Learning Courses". In: 2015, pp. 67–93. DOI: 10.1007/978-3-319-11575-7_6. URL: http://link.springer.com/10.1007/978-3-319-11575-7%7B%5C_%7D6.

[42] Sinjini Mitra and Zvi Goldstein. "Designing early detection and intervention techniques via predictive statistical models—A case study on improving student performance in a business statistics course". In: *Commun. Stat. Case Stud. Data Anal. Appl.* 1.1 (Jan. 2015), pp. 9–21. ISSN: 2373-7484. DOI: 10.1080/23737484.2015.1063409. URL: http://www.tandfonline.com/doi/full/10.1080/23737484.2015.1063409.

[43] José gonçalves de Oliveira Junior. "Identificação de padrões para a Análise de Evasão em Cursos de Graduação usando Mineração de Dados Educacionais". In: (2015).

[44] Amirah Mohamed Shahiri, Wahidah Husain, and Nur'Aini Abdul Rashid. "A Review on Predicting Student's Performance Using Data Mining Techniques". en. In: *Procedia Comput. Sci.* Vol. 72. 2015, pp. 414–422. DOI: 10.1016/j.procs.2015.12.157. URL: https://linkinghub.elsevier.com/retrieve/pii/S1877050915036182.

[45] Tufi Machado Soares et al. "Fatores associados ao abandono escolar no ensino médio público". In: *Educ. e Pesqui.* 41.3 (Sept. 2015), pp. 757–772. DOI: 10.1590/S1517-9702201507138589.

[46] Rebecca Ferguson et al. *Research Evidence on the Use of Learning Analytics - Implications for Education Policy*. 2016. 2016, pp. 1–152. ISBN: 9789279644412. DOI: 10.2791/955210.

[47] Angelo Fynn. "Ethical considerations in the practical application of the unisa socio-critical model of student success". In: *Int. Rev. Res. Open Distance Learn.* 17.6 (2016), pp. 206–220. ISSN: 14923831. DOI: 10.19173/irrodl.v17i6.2812.

[48] Dragan Gašević et al. “Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success”. In: *Internet High. Educ.* 28 (Jan. 2016), pp. 68–84. ISSN: 10967516. DOI: 10.1016/j.iheduc.2015.10.002. URL: https://linkinghub.elsevier.com/retrieve/pii/S1096751615300038.

[49] Stephanie M. Gratiano and William John Palm. “Can a five-minute, three-question survey foretell first-year engineering student performance and retention?” In: *ASEE Annu. Conf. Expo. Conf. Proc.* 2016-June.June (2016). ISSN: 21535965. DOI: 10.18260/p.26427.

[50] Jeonghyun Kim, Il-Hyun Jo, and Yeonjeong Park. “Effects of learning analytics dashboard: analyzing the relations among dashboard utilization, satisfaction, and learning achievement”. In: *Asia Pacific Educ. Rev.* 17.1 (Mar. 2016), pp. 13–24. ISSN: 1598-1037. DOI: 10.1007/s12564-015-9403-8. URL: http://link.springer.com/10.1007/s12564-015-9403-8.

[51] Farshid Marbouti, Heidi A. Diefes-Dux, and Krishna Madhavan. “Models for early prediction of at-risk students in a course using standards-based grading”. en. In: *Comput. Educ.* 103 (Dec. 2016), pp. 1–15. ISSN: 03601315. DOI: 10.1016/j.compedu.2016.09.005. URL: https://linkinghub.elsevier.com/retrieve/pii/S0360131516301634.

[52] Ian H. Witten et al. *Data Mining: Practical Machine Learning Tools and Techniques.* Elsevier, 2016, pp. 1–621. ISBN: 9780128042915. DOI: 10.1016/c2009-0-19715-5. URL: https://linkinghub.elsevier.com/retrieve/pii/C20090197155.

[53] Rodrigo Ferrer de Argôlo. “Determinantes de Desempenho dos Estudantes do Ensino Superior : O Caso do Curso de Pasicologia da UFBA”. In: *Programa Pós-Graduação em Educ. da Fac. Educ. da Univ. Fed. da Bahia* (2017).

[54] Ali Daud et al. “Predicting student performance using advanced learning analytics”. In: *26th Int. World Wide Web Conf. 2017, WWW 2017 Companion* October (2017), pp. 415–421. DOI: 10.1145/3041021.3054164.

[55] Anne Sophie Hoffait and Michaël Schyns. “Early detection of university students with potential difficulties”. en. In: *Decis. Support Syst.* 101 (Sept. 2017), pp. 1–11. ISSN: 01679236. DOI: 10.1016/j.dss.2017.05.003. URL: https://linkinghub.elsevier.com/retrieve/pii/S0167923617300817.

[56] Salam Khan. “Analysis of Social, Psychological and Other Factors on College Dropout Rates among African American Students”. In: *Adv. Appl. Sociol.* 07.08 (2017), pp. 319–326. ISSN: 2165-4328. DOI: 10.4236/aasoci.2017.78020. URL: http://www.scirp.org/journal/doi.aspx?DOI=10.4236/aasoci.2017.78020.

[57] Amanda J. Sebesta and Elena Bray Speth. “How should i study for the exam? Self-regulated learning strategies and achievement in introductory biology”. In: *CBE Life Sci. Educ.* 16.2 (June 2017). Ed. by Janet Batzli, ar30. ISSN: 19317913. DOI: 10.1187/cbe.16-09-0269. URL: https://www.lifescied.org/doi/10.1187/cbe.16-09-0269.

[58] Jonathan Williams. “Portuguese Higher Education ADDRESSING THE COMPLETION CHALLENGE IN PORTUGUESE HIGHER”. In: 81 (2017).

[59] Whitney Alicia Zimmerman and Glenn Johnson. "Exploring factors related to completion of an online introductory statistics course". In: *Online Learn. J.* 21.3 (2017), pp. 191–205. DOI: 10.24059/olj.v21i3.1017. URL: http://olj.onlinelearningconsortium.org/index.php/olj/article/view/1017.

[60] Dominique Guellec et al. "OECD Review of the Tertiary Education, Research and Innovation System in Portugal". en. In: (2018), pp. 1–56.

[61] Mohammed Javed, P. Nagabhushan, and Bidyut B. Chaudhuri. "A review on document image analysis techniques directly in the compressed domain". In: *Artif. Intell. Rev.* 50.4 (2018), pp. 539–568. DOI: 10.1007/s10462-017-9551-9.

[62] M. M. Malik, S. Abdallah, and M. Ala'raj. "Data mining and predictive analytics applications for the delivery of healthcare services: a systematic literature review". In: *Ann. Oper. Res.* 270.1-2 (2018), pp. 287–312. ISSN: 15729338. DOI: 10.1007/s10479-016-2393-z.

[63] Tiago Martins, Daniel Gonçalves, and Sandra Gama. "Visualizing Historical Patterns in Large Educational Datasets". In: *Int. J. Creat. Interfaces Comput. Graph.* 9.1 (2018), pp. 32–48. DOI: 10.4018/ijcicg.2018010103.

[64] Fumiya Okubo et al. "On the prediction of students' quiz score by recurrent neural network". In: *CEUR Workshop Proc.* 2163 (2018), pp. 1–6. ISSN: 16130073.

[65] Tatiana A. Cardona and Elizabeth A. Cudney. "Predicting student retention using support vector machines". en. In: *Procedia Manuf.* Vol. 39. 2019, pp. 1827–1833. DOI: 10.1016/j.promfg.2020.01.256. URL: https://linkinghub.elsevier.com/retrieve/pii/S2351978920303206.

[66] Haozhang Deng et al. "PerformanceVis: Visual analytics of student performance data from an introductory chemistry course". In: *Vis. Informatics* 3.4 (Dec. 2019), pp. 166–176. ISSN: 2468502X. DOI: 10.1016/j.visinf.2019.10.004. URL: https://linkinghub.elsevier.com/retrieve/pii/S2468502X1930049X.

[67] E. T. Lau, L. Sun, and Q. Yang. "Modelling, prediction and classification of student academic performance using artificial neural networks". en. In: *SN Appl. Sci.* 1.9 (Sept. 2019), p. 982. ISSN: 25233971. DOI: 10.1007/s42452-019-0884-7. URL: http://link.springer.com/10.1007/s42452-019-0884-7.

[68] Sunbok Lee and Jae Young Chung. "The machine learning-based dropout early warning system for improving the performance of dropout prediction". In: *Appl. Sci.* 9.15 (2019). DOI: 10.3390/app9153093.

[69] Maria P.G. Martins et al. "A Data Mining Approach for Predicting Academic Success – A Case Study". In: *Adv. Intell. Syst. Comput.* Vol. 918. Cham: Springer International Publishing, 2019, pp. 45–56. DOI: 10.1007/978-3-030-11890-7_5. URL: http://link.springer.com/10.1007/978-3-030-11890-7%7B%5C_%7D5.

[70] Shahrul Nizam Ismail, Suraya Hamid, and Haruna Chiroma. "The utilization of learning analytics to develop student engagement model in learning management system". In: *J. Phys. Conf. Ser.* 1339 (Dec. 2019), p. 012096. ISSN: 1742-6588. DOI: 10.1088/1742-6596/1339/1/012096. URL: https://iopscience.iop.org/article/10.1088/1742-6596/1339/1/012096.

[71] The Hechinger Report. *10 years later, goal of getting more Americans through college is way behind schedule.* 2019. URL: https://hechingerreport.org/10-years-later-goal-of-getting-more-americans-through-college-is-way-behind-schedule/ (visited on 08/17/2021).

[72] Beauchamp. *Decision Tree x Random Forest.* 2020. URL: https://commons.wikimedia.org/wiki/File:Decision%7B%5C_%7DTree%7B%5C_%7Dvs.%7B%5C_%7DRandom%7B%5C_%7DForest.png (visited on 09/20/2021).

[73] Juan Pedro Cerro Martínez, Montse Guitert Catasús, and Teresa Romeu Fontanillas. "Impact of using learning analytics in asynchronous online discussions in higher education". In: *Int. J. Educ. Technol. High. Educ.* 17.1 (2020). DOI: 10.1186/s41239-020-00217-y.

[74] Janneth Chicaiza et al. "Application of data anonymization in Learning Analytics". In: *ACM Int. Conf. Proceeding Ser.* January (2020). DOI: 10.1145/3378184.3378229.

[75] Kristof Coussement et al. "Predicting student dropout in subscription-based online learning environments: The beneficial impact of the logit leaf model". In: *Decis. Support Syst.* 135.December 2019 (2020), p. 113325. DOI: 10.1016/j.dss.2020.113325. URL: https://doi.org/10.1016/j.dss.2020.113325.

[76] European Comission. *Report on Educational expenditure statistics.* 2020. URL: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Educational%7B%5C_%7Dexpenditure%7B%5C_%7Dstatistics (visited on 08/17/2021).

[77] European Comission. *Targets for the European Union in 2020.* 2020. URL: https://ec.europa.eu/info/topics/education-and-training%7B%5C_%7Den (visited on 08/17/2021).

[78] Raza Hasan et al. "Predicting student performance in higher educational institutions using video learning analytics and data mining techniques". In: *Appl. Sci.* 10.11 (2020). ISSN: 20763417. DOI: 10.3390/app10113894.

[79] Danial Hooshyar, Margus Pedaste, and Yeongwook Yang. "Mining educational data to predict students' performance through procrastination behavior". In: *Entropy* 22.1 (Dec. 2020), p. 12. ISSN: 10994300. DOI: 10.3390/e22010012. URL: https://www.mdpi.com/1099-4300/22/1/12.

[80] Dirk Ifenthaler and Jane Yin-Kim Yau. "Utilising learning analytics to support study success in higher education: a systematic review". In: *Educ. Technol. Res. Dev.* 68.4 (Aug. 2020), pp. 1961–1990. DOI: 10.1007/s11423-020-09788-z. URL: https://link.springer.com/10.1007/s11423-020-09788-z.

[81] Laci Mary Barbosa Manhães and Sérgio Manuel Serra da Cruz. "PREDIÇÃO DO DESEMPENHO ACADÊMICO DE ALUNOS DA GRADUAÇÃO UTILIZANDO MINERAÇÃO DE DADOS". pt. In: *Simpósio Pesqui. Operacional e Logística da Mar. - Publicação Online.* São Paulo: Editora Blucher, May 2020, pp. 2050–2064. DOI: 10.5151/spolm2019-148. URL: http://www.proceedings.blucher.com.br/article-details/34563.

[82] First-year Persistence. *Persistence and retention 2020.* Tech. rep. 2020, pp. 1–17. URL: https://nscresearchcenter.org/persistence-retention/.

[83] Shirin Riazy, Katharina Simbeck, and Vanessa Schreck. “Fairness in learning analytics: Student at-risk prediction in virtual learning environments”. In: *CSEDU 2020 - Proc. 12th Int. Conf. Comput. Support. Educ.* 1.Csedu (2020), pp. 15–25. DOI: 10.5220/0009324100150025.

[84] Cristobal Romero and Sebastian Ventura. “Educational data mining and learning analytics: An updated survey”. In: *WIREs Data Min. Knowl. Discov.* 10.3 (May 2020). ISSN: 1942-4787. DOI: 10.1002/widm.1355. URL: https://onlinelibrary.wiley.com/doi/10.1002/widm.1355.

[85] Barbara G. Tabachnick. “Multivariate Statistics”. In: *Essentials Polit. Res.* (2020), pp. 173–208. DOI: 10.4324/9780429500749-17.

[86] Nikola Tomasevic, Nikola Gvozdenovic, and Sanja Vranes. “An overview and comparison of supervised data mining techniques for student exam performance prediction”. In: *Comput. Educ.* 143 (Jan. 2020), p. 103676. ISSN: 03601315. DOI: 10.1016/j.compedu.2019.103676. URL: https://linkinghub.elsevier.com/retrieve/pii/S0360131519302295.

[87] Xinhua Wang et al. “Student performance prediction with short-term sequential campus behaviors”. In: *Inf.* 11.4 (2020), pp. 1–20. ISSN: 20782489. DOI: 10.3390/INFO11040201.

[88] Haogang Bao et al. “The effects of a learning analytics dashboard on teachers’ diagnosis and intervention in computer-supported collaborative learning”. In: *Technol. Pedagog. Educ.* 30.2 (2021), pp. 287–303. DOI: 10.1080/1475939X.2021.1902383. URL: https://www.tandfonline.com/doi/full/10.1080/1475939X.2021.1902383.

[89] Paulo Diniz Gil et al. “A data-driven approach to predict first-year students’ academic success in higher education institutions”. In: *Educ. Inf. Technol.* 26.2 (2021), pp. 2165–2190. ISSN: 15737608. DOI: 10.1007/s10639-020-10346-6. URL: http://link.springer.com/10.1007/s10639-020-10346-6.

[90] Omiros Iatrellis et al. “A two-phase machine learning approach for predicting student outcomes”. In: *Educ. Inf. Technol.* 26.1 (2021), pp. 69–88. ISSN: 15737608. DOI: 10.1007/s10639-020-10260-x.

[91] Jelena Jovanović et al. “Students matter the most in learning analytics: The effects of internal and instructional conditions in predicting academic success”. en. In: *Comput. Educ.* 172 (Oct. 2021), p. 104251. ISSN: 03601315. DOI: 10.1016/j.compedu.2021.104251. URL: https://linkinghub.elsevier.com/retrieve/pii/S0360131521001287.

[92] Michelangelo Misuraca, Germana Scepi, and Maria Spano. “Using Opinion Mining as an educational analytic: An integrated strategy for the analysis of students’ feedback”. In: *Stud. Educ. Eval.* 68 (Mar. 2021), p. 100979. ISSN: 0191491X. DOI: 10.1016/j.stueduc.2021.100979. URL: https://linkinghub.elsevier.com/retrieve/pii/S0191491X21000055.

[93] Abdallah Namoun and Abdullah Alshanqiti. *Predicting student performance using data mining and learning analytics techniques: A systematic literature review.* en. Dec. 2021. DOI: 10.3390/app11010237. URL: https://www.mdpi.com/2076-3417/11/1/237.

[94] Michael Scholz and Tristan Wimmer. "A comparison of classification methods across different data complexity scenarios and datasets". In: *Expert Syst. Appl.* 168 (Apr. 2021), p. 114217. ISSN: 09574174. DOI: 10.1016/j.eswa.2020.114217. URL: https://linkinghub.elsevier.com/retrieve/pii/S0957417420309428.

[95] Max Kuhn. *Recursive Feature Elimination algorithm| Caret Package for R programming language.* URL: https://topepo.github.io/caret/recursive-feature-elimination.html (visited on 08/09/2021).

[96] Alvin Nguyen. *Comparative Study of C5.0 and CART algorithms.* Tech. rep.

[97] UNESCO Institute for Statistics. *Government expenditure on education.* URL: https://data.worldbank.org/indicator/SE.XPD.TOTL.GB.ZS (visited on 08/01/2021).

[98] Hadley Wickham. *R for Data Science - Many Models.* URL: https://r4ds.had.co.nz/many-models.html%7B%5C#%7Dlist-columns-1.

Appendix A

Attachments

This section contains additional information for the Scientific Literature review - A.1, description of the main Case Study - A.3.

A.1 Scientific Literature review

The following table displays studies deemed relevant in the scientific literature in relation to the subject of student performance prediction. We focused on more recent work, as the methods have advanced at great pace in recent years. There is a wide variety of countries where the works were applied. Some of the most influential countries in all areas, but also in terms of studies in student performance, as analysed by this study, were the United States and China.

Year	Author	Country	Name of the paper
2010	Macfadyen and Dawson	Canada	Mining LMS data to develop an “early warning system” for educators
2012	Huang, Fang	USA	Predicting student academic performance in an engineering dynamics course
2014	Baccaro	Brazil	Relação entre Desempenho no Vestibular e Rendimento Acadêmico
2014	Sharabiani	USA	Bayesian Network Model for Prediction of Students’ Performance in Engineering
2015	Stretcht and Cruz	Portugal	A Comparative Study of Classification and Regression Algorithms for Modelling Students’ Academic Performance
2015	Mitra and Goldstein	USA	Designing early detection and intervention techniques - improving student performance
2016	Marbouti et al.	USA	Models for early prediction of at-risk students in a course using standards-based grading
2016	Qiu et al.	China	Modeling and Predicting Learning Behavior in MOOCs
2016	Gasevic	UK	Learning Analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success
2017	Al-Shabandar et al.	USA	Machine Learning Approaches to Predict Learning Outcomes in MOOC
2018	Okubo	Japan	On the Prediction of Students’ Quiz Score by Recurrent Neural Network
2018	Miguéis	Portugal	Early segmentation of students acc. to academic performance: A predictive modelling approach
2019	Martins et al.	Portugal	Data Mining Approach for Predicting Academic Success
2019	Lau	China	Modeling, prediction and classification of student performance using neural networks
2019	Hooshyar	Estonia	Mining educational data to predict students’ performance through procrastination behavior
2020	Wang et al.	China	Student Performance Prediction with Short-Term Sequential Campus Behaviors
2020	Guerro-Higueras et al.	Spain	Academic Success Assessment through Version Control Systems
2020	Wang	China	Student performance prediction with short-term sequential campus behaviors
2021	Jovanović	Serbia	Students matter most: effects of internal and instructional conditions in predicting success

Table A.1: Latest influential works on Student performance prediction

Table 3.1 shows in summary the classes of variables most used to predict student performance, according to study carried out on recent influential works in the field. Sociodemographic, Academic History and Participation in educational artefacts are the three most used classes of variables. Additionally, such students feed on average 24 variables to their predictive models.

A.2 Review of studies on student performance prediction

This section looks to summarize influential studies found in the literature on Student Performance prediction.

The objective of the study performed by Martins et al, 2019[69], was to train a Random Forest-based Regression algorithm to determine at an early stage the global academic performance of students of a Portuguese polytechnic higher education institution. The dataset contained socio-demographic, curricular and access-related attributes, which were split into groups. Most valuable groups of variables were determined and this allowed to decrease the dimensionality early, lowering the complexity of the prediction task.

Hooshyar et al, 2019[79], used student participation data collected from Moodle to predict the performance of students with learning difficulties through procrastination behaviour. First, they developed a feature vector based on engagement logs and, after having setting definitions for what is considered procrastination behaviour, students were clustered in 3 groups. Classification algorithms were applied to the student data and results shown that it was possible to predict students' performance through their procrastination behaviour with an accuracy of 96%.

Tomasevic et al, 2020[86], used a dataset of approximately 30.000 students acquired from an open university to compare the performance of several machine learning algorithms on predicting student performance. Three different approaches were used: (1) Similarity-based algorithms: such as K-Nearest Neighbors - KNN - and distance-weighted KNN. (2) Machine learning modelling: Support Vector Machines, Neural Networks, Decision Trees, among others. (3) Probabilistic: Naive Bayes and Bayesian Linear Regression. Data used in this study included student performance, student engagement and student demographic attributes. Several combinations of variables were tested to determine which algorithm and combination of variables generated the predictions with least error. Of all the algorithms used, Neural Networks and SVM performed better with all variables although not the best on computing times.

In a smaller extent, many authors focus on studying the successful completion of individual learning tasks, based, for instance, on participation data[39] or social network interaction[20].

Okubo et al., 2018[64] used a very broad set of attributes, obtaining data from 3 different systems regarding student participation, entries in a portfolio system and quiz results to predict grades for 15 short-term courses with high performance and low error using Neural Networks.

In their turn, Wang et al.'s 2020 work[87], used highly unique attributes in his study of predicting the performance of students from a technical course at a Chinese college. This author encoded student behaviours related to library book lending, cafeteria spending, and library entry frequency, among others, and developed a semi-automatic implementation of a Sequence-based Classifier model.

Martins, 2018[63], designed line plots, scatterplots and matrices visualizations that allow the evaluation of aspects related to student approval and achievement statistics from courses offered on this institution, also building a framework for using these charts in other case studies.

Work by Deng et al, 2019[66] presented a visual analytics tool for analysing student admission and course performance data and investigating homework and exam question design named *PerformanceVis*.

With an investigation more focused on user satisfaction and perception, Kim et al. 2016[50] found a slight but statistically significant correlation between satisfaction with using a Learning Analytics dashboard and student achievement.

Tomasevic's 2020 study[86], for example, divided the variables into three groups, namely Demographics (D), Engagement (E) and Performance (P). Although a more limited approach from the point of view of the range of variables, with only 17 attributes split into three groups, the experiments performed were quite comprehensive, since each ML algorithm was trained with each of the variable combinations.

The combination D+E+P, that is, when all variables were used, had the best performance in almost all algorithms, with the exception of the overall winner, a Neural Networks Algorithm with the Engagement and Performance variables, which obtained the best performance across all tested setups. From this result it is possible to abstract that not always adding more variables to a model will result in better results for a job.

Study by Martins et al (2019)[69], arranged predictors in five groups: Curricular, Matriculation, Demographic, Socioeconomic and Access-related and trained Machine Learning models in combinations of those groups. Vandamme, 2007[10] also split all attributes in three groups, sociodemographic, student engagement and student perceptions, i.e. their views on their academic context, professors, courses, etc.

One of the most complete recent studies in terms of broadness of scope of variables is Mitra and Goldstein, 2015[42], that used attributes from five classes, as the own authors categorized: Demographic factors, Academic History, Work-related factors, Course-related factors, Academic self-concept factors. While the study only covered one course, it combined several features not often utilized together making for a very comprehensive and thorough investigation.

A.2.1 Attributes

Hasan, 2020[78] used student academic information, along with interaction data to predict end semester examination grades in online courses. While in a different context and limited scope, some of the features available were course marks, progress in the course and plagiarism count, i.e. if the student had been accused of any plagiarism.

Iatrellis et al., 2021[90], determined several variables related to the academic history were indeed relevant, with information on: student rank in their class, specialization field, first year performance, and whether the student participated of a mobility program, obtained any internships or awards.

Manhães and Cruz, 2019[81] used several variables representing past performance of the students to predict dropout in a Brazilian university: the number of Course units enrolled, completed and failed, and the student's performance coefficients updated every school period.

A study by Moriconi and Nascimento, 2014[32] with students graduating from Engineering courses found a high association of socio-economic background measures with student performance, where courses with more economically disadvantaged students achieved performance in terms of grades. Likewise, a study by Baccaro and Shinyashiki, 2014, with 4000 students from various courses at a Brazilian university, found a statistically significant relationship between student performance and demographic attributes such as age, gender, economic status and type of establishment attended in high school.[29]

Macfayden and Dawson, 2010[15], extracted and used as predictors several variables from a Learning Management System (LMS) and studied student grades with Regression algorithms. Some of the variables deemed most important for predicting performance were (1) Total number of discussion messages posted in the discussion forum; (2) Total number of mail messages sent; (3) total number of assessments completed.

Study by Jovanovic, 2021[91], used high and low-level attributes - less and more granular, respectively - to predict the grades obtained in online courses. It was determined that engagement-based predictors had more explanatory power than demographic or other groups of attributes. The overall time spent online, regularity in the daily counts of discussion forum postings and participation in weekly activities were the most determinant attributes for this particular set of data.

Lau, 2019[67] predicted grades obtained from students in a Chinese undergraduate course using grades obtained in 5 different parts of admission exams: Chinese, Maths, English, Science and Proficiency along with a few demographic variables. The study showed that some parts of the admission exams had better correlations to the student's cumulative average in college, with the English exam being the most strongly correlated.

A.3 Case Study

A.3.1 Data Overview

The dataset used in the study was extracted from Universidade do Porto's main Student Information System, *Sigarra*. It contains comprehensive information on the student life cycle in the university. It uses an Oracle relational database with a more classical and well defined data model but its size and lack of optimization were two issues faced while querying and extracting its data. Although there is a Data Warehouse under development to be finished in this current year, the university still did not have an established process for extracting data from this database nor did it have the extensive documentation one may think should be available for this kind of data.

One of the challenges when extracting and preparing this dataset was understanding the attributes and their definitions. Concepts such as the one of Occurrence, the admission regimes and student statuses, for instance, may look straightforward at first glance, but a quick dive into the data shows those are not always simple as they look. Even with a steep learning curve, thanks to extensive documentation, clarifications from helpful colleagues and an introductory course on these definitions provided by the university, understanding of the definitions and distributions became easier with familiarization with the data.

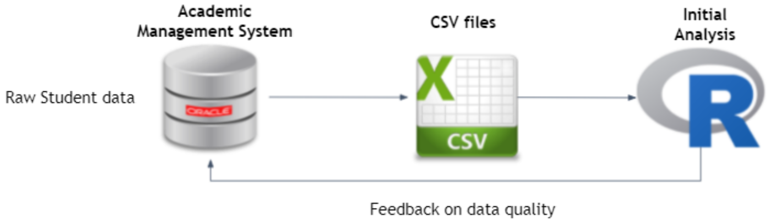


Figure A.1: Data Extraction and Data quality verification process

Data extracted from academic management systems were made available iteratively in Comma-separated values (CSV) files that were imported and processed using R programming language. To ensure better data quality, the university carried out a monthly review of files, delivering updated CSV files, result of improvements in data extraction queries. At the same time, the development of an updated Extraction, Transformation and Loading (ETL) system was started in Python language, but its conclusion was not possible until the end of this work.

The dataset consisted of 1.35 Million observations and 54 attributes with data available from 8 academic years, 2012 through 2019. The smallest granularity of the data is that of one enrolment, that is, each observation represents the enrolment of a student in a Course in a given academic period. Only data from Undergraduate courses was available and table A.2 shows the distribution of enrolments per curricular year and type of programme.

Curricular Year	Integrated Master	First Degree	Total
1	243328	271405	514733
2	244724	221063	465787
3	204624	173170	377794
Total	692676	665638	1358314

Table A.2: Number of students per Curricular Year and Type of Programme enrolled

A.3.1.1 Attributes

Regarding the analysis of attributes, the main target of this study, they comprise information of varied types, ranging from distinct and continuous numerical attributes to categorical attributes with diverse levels of cardinality.

Feature	%Missing	Most common value	Mean
AGE	0.00	19	20.48
APPLICATION_RANKING	18.09	170	164.50
AVERAGE_11_GRADE	94.16	17	17.05
AVERAGE_12_GRADE	38.52	17	151.54
AVERAGE_SECONDARY	94.18	17	21.69
AVERAGE_1ST_YEAR_1ST_SEMESTER	0.00	13	12.65
AVERAGE_1ST_YEAR	0.00	12	13.00
AVERAGE_APPROVED_COURSES	0.00	0	8.71
HIGHEST_GRADE	32.21	17	16.36
LOWEST_GRADE	32.21	10	10.65
NUMBER_ENROLLMENTS_IN_COURSE	0.00	1	1.30
COURSES_ENROLLED_IN_SEMESTER	0.00	5	6.17
CREDITS_ENROLLED_IN_SEMESTER	0.00	30	31.24
CREDITS_APPROVED	0.11	0	54.70
DELAYED_COURSES	0.00	0	1.98
DELAYED_YEARS	0.15	0	0.32
ENROLLED_COURSES	0.00	0	16.32
COURSES_COMPLETED	0.00	0	11.21
CREDITS_COMPLETED	0.11	0	61.82
PROGRAMME_COMPLETION	0.11	0	0.26

Table A.3: Numerical predictor variables

A complete and detailed univariate analysis of each attribute available in the dataset was then performed. As for numerical variables, density plots, scatter plots and histograms were studied mainly in search for patterns and outliers. In categorical variables, frequency tables were created to identify variables with prevalent levels or variables in which the change from one category to another meant a large change in the probability of obtaining a positive or negative result for the target variable.

Tables A.4 and A.3 contains information on all attributes, categorical and numerical.

A.3.1.2 Target variable

The target variable for this task is APPROVAL_STATUS, a binary categorical variable, where "pass" indicates that the student passed this course and "fail" indicates failure to achieve the minimum goals for approval. Overall approval rate considering all Courses was 71.59%, determined by the number of observations with approval status equal to “pass” in relation to the number of observations in the dataset. Conversely, 28.41% of the ratings indicated disapproval.

By grouping approval rates by Course, it can be seen in figure A.2 that the average approval rate among all Courses is approximately 81% and that most courses have an approval rate of

Feature	%Missing	Most frequent value
SEX	0.00	F
MARITAL_STATUS	0.02	SINGLE
COUNTRY_NATIONALITY	0.00	PT
COUNTRY_BIRTH	21.85	PT
COUNTRY_OFICIAL_RESIDENCE	0.42	PT
COUNTRY_CLASSES_RESIDENCE	2.77	PT
FOREIGN_STUDENT	0.00	N
DISPLACED_STUDENT	0.00	N
EDUCATIONAL_LEVEL_STUDENT	78.82	Secondary - Year 12
EDUCATIONAL_LEVEL_PARENT1	21.36	Secondary - Year 12
EDUCATIONAL_LEVEL_PARENT2	21.66	Bachelor's
OCCUPATION_STUDENT	21.84	Student
OCCUPATION_PARENT1	21.92	Employee
OCCUPATION_PARENT2	21.87	Employee
SPECIAL_NEEDS_STUDENT	0.00	N
APPLIED_SCHOLARSHIP	0.00	N
HAS_SCHOLARSHIP	0.00	N
ADMISSION_REGIME	8.82	Regular admission
APPLICATION_PREFERENCE	19.92	1
DEDICATION_REGIME	0.00	Full-time
DEBT_SITUATION	0.00	N
STATUS_STUDENT	1.99	Ordinary

Table A.4: Categorical predictor variables

around 90%, making for consistently unbalanced classes for a Classification task. Additionally, 150 courses have a 100% approval rate, and in 12 courses, less than 1% of the students were approved.

A.3.2 Data preparation

To carry out the descriptive analysis and understanding of the data, it was decided to keep the dataset in its original state to present any kind of insights in their pure form, without tampering. However, we identified the need to perform several data processing operations that would be necessary for the continuation of the work.

A.3.2.1 Missing values

A wide-ranging analysis of missing values was performed during most of the work, with the objective of obtaining a dataset as complete as possible to generate insights and predictions at

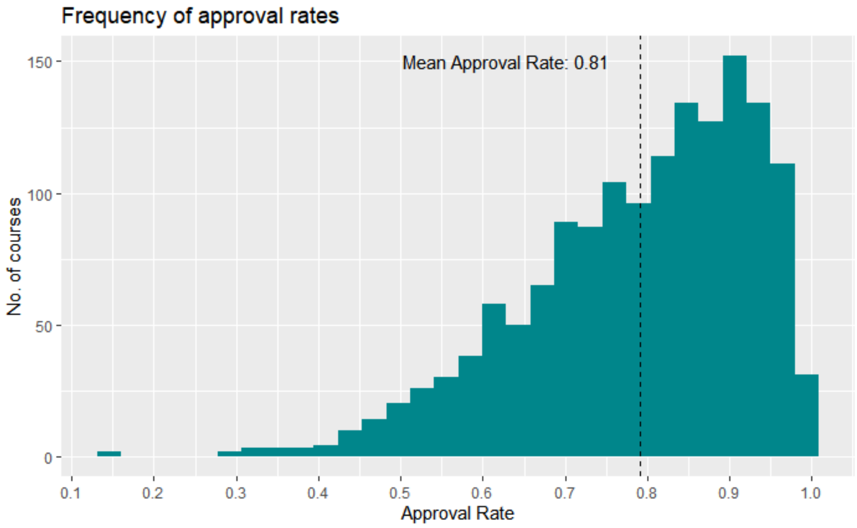


Figure A.2: Distribution of approval rates among all courses

the end of the project. From the findings made in this activity, some adjustments were made directly to the data extraction queries by the competent team of the Universidade do Porto.

Even with the time devoted to this task, data quality was still somewhat problematic, especially with regard to variables that depended on user input. Many of the sociodemographic features available are informed by the students themselves through forms filled in person or online. Few attributes are for student enrolment, and many of these are unavailable or with their quality compromised throughout the entire student period. This can be verified in variables that represent the educational level of the student or those of their parents, for instance, where at least 20% of values are missing. Equivalent behaviour is observed in several other variables, as can be seen in table A.5.

Table A.5 shows all variables with missing values and the percentage in relation to the total number of observations in the enrolment dataset. The first two variables, which represent the grades of students in the 11th grade and the mean grade for the entire secondary education, were the ones with the highest percentage of missing values. With access to the calculations made in the extraction queries, we observed the latter is calculated as a mean of the average of grades in the 11th and 12th grades, so the high number of missing values in one unquestionably explains the high number of the other. With more than 94% of missing values, the only plausible solution was to remove both variables from the dataset.

The other numerical variable that provides information on performance in secondary education, AVERAGE_12_GRADE, was also excluded. Although it has a smaller percentage of missing values, it has a bimodal distribution, with approximately half of its values below 20 and the other half between 150 and 200. This can be either a result of miscalculation or an indication of two conflicting measure systems being used at the same time in the country for students coming out of High School. Either way, the safe choice was removing the variable to avoid misrepresentation of patterns.

There is also a high number of missing values in variables APPLICATION_PREFERENCE,

Attribute	**% missing**	**Transformation**
AVERAGE_SECONDARY	94,19%	Removal, too many missing
AVERAGE_11_GRADE	94,18%	Removal, too many missing
EDUCATIONAL_LEVEL_STUDENT	78,77%	Set as "Not Available"
AVERAGE_12_GRADE	38,68%	Removal, odd distribution
HIGHEST_GRADE	32,22%	Local conservative imputation
LOWEST_GRADE	32,22%	Local conservative imputation
OCCUPATION_PARENT1	21,84%	Set as "Not Available"
COUNTRY_BIRTH	21,79%	Set as "Not Available"
OCCUPATION_PARENT2	21,76%	Set as "Not Available"
OCCUPATION_STUDENT	21,74%	Set as "Not Available"
EDUCATIONAL_LEVEL_PARENT2	21,57%	Set as "Not Available"
EDUCATIONAL_LEVEL_PARENT1	21,31%	Set as "Not Available"
APPLICATION_PREFERENCE	19,88%	Set as "Not Available"
APPLICATION_RANKING	18,03%	Local conservative imputation
ADMISSION_REGIME	8,77%	Set as "Not Available"
COUNTRY_CLASSES_RESIDENCE	2,79%	Set as "Not Available"
STATUS_STUDENT	1,97%	Set as "Not Available"
COUNTRY_OFICIAL_RESIDENCE	0,38%	Set as "Not Available"
DELAYED_YEARS	0,15%	Local conservative imputation
CREDITS_APPROVED	0,11%	Local conservative imputation
CREDITS_COMPLETED	0,11%	Local conservative imputation
PROGRAMME_COMPLETION	0,11%	Local conservative imputation
MARITAL_STATUS	0,02%	Set as "Not Available"

Table A.5: Missing values and transformation done on each variable

ADMISSION_REGIME and APPLICATION_RANKING, showing that there is certainly a data gap at the University in relation to student admission process. Lastly, HIGHEST_GRADE and LOWEST_GRADE, had a high number of missing values because they are only calculated from the end of the 1st term, since it would be impossible to determine the highest or lowest grade if the student had not yet obtained any assessment at this level. Such values were labelled as missing for enrolments in the 1st period and these variables were removed from the statistical models of 1st-year Course Units.

A.3.2.2 Substitution of Missing Values by category

Some categorical variables, especially those that depend on user-entered data, had very erratic data distributions. On the one hand, they had a very significant presence of missing values. Furthermore, some had a wide spread of information across various categories.

The variables that represent the educational level of the student's parents, for example, have 13 categories, which allows very specific analyses from the point of view of an Exploratory

Analysis, but such distributions can do more harm than good when it comes to train statistical models. Too stratified data can be statistically insufficient to discover relationships, and a large number of categories increases the complexity and processing time of a predictive model.

Educational Level Parent	# of obs.	%
higher education - bachelor's degree	277019	0.20
secondary education - year 12 or equivalent	229889	0.17
third stage of basic education - year 9	140794	0.10
second stage of basic education - year 6	104327	0.08
not available	86373	0.06
first stage of basic education - year 4	74018	0.05
master's degree	59117	0.04
higher education - graduate	37303	0.03
doctorate	27372	0.02
intermediate secondary education	10250	0.01
technical education	9508	0.01
able to read without having attended year 4	8465	0.01
not able to read or write	919	0.00
NA	292960	0.22

Table A.6: Variable before lumping factors and adding category

As mentioned above and shown in A.6, this variable has a high rate of missing values, 78% of the dataset. Many statistical algorithms do not handle missing values well and that is the case with the implementation of the Random Forest model used in this work, which outputs an error when trying to run it when there is as much as 1 missing value in the data.

To try to take advantage of the informational value of this variable and to avoid removing it from the dataset, its missing values were transformed into a category named "Not available", since the data not being not informed by the student may represent information already in itself. This same operation was carried out for other categorical variables with a high rate of missing values, such as the educational level of parents and students, occupation of parents and students, and geographic attributes, representing the country of birth, nationality, etc.

A.3.2.3 "Grouping" less representative factor levels

To address the sparsity of some attributes that have many factor levels, they went through a process called *lumping*. All categories that did not reach 5% of representation in an attribute were grouped as a category called "Others". The algorithm iterates through every factor variable binning together categories that do not represent at least our predetermined percentage of the data, in this case, set to 5%. Although at some risk of losing information, that transformation allows for less sparse and more representative models.

As seen in table A.6, there are several levels of the parent educational level variable that have less than 5% of representation, and after the transformation, the of the data was severely reduced. Like many other operations when it comes to data processing, there is a trade-off between a possible loss of information and an operation aimed at more efficiency and reproducibility of the model. While the “other” category may not represent much information on its own, the most representative factors that hopefully will be relevant when constructing the statistical models were kept.

Educational Level Parent	# of obs.	%
not available	379333	0.28
higher education - bachelor’s degree	277019	0.20
secondary education - year 12 or equivalent	229889	0.17
third stage of basic education - year 9	140794	0.10
second stage of basic education - year 6	104327	0.08
first stage of basic education - year 4	74018	0.05
Other	152934	0.11

Table A.7: Variable after lumping factors and adding category

Table A.7 shows the result of this operation, which was performed on all categorical variables with more than 2 categories.

A.3.2.4 Conservative local imputation

The first of the local transformations is intended to correct the problem of missing values in some numeric variables of the dataset, namely those representing: (1) Highest and Lowest grade of the student throughout their entire academic path (2) Number of delayed years behind schedule (3) Rate of programme completion (4) Number of ECTS credits enrolled in that school period (5) Classification given in the student application process All of those attributes had up to 30% of the missing values in the overall dataset and while this number raises some concern, imputation was the preferred solution because it allows to keep the variables instead of removing them.

Simpler strategies for value imputation in Machine Learning use, for instance, the mean, median or minimum value of a column for the entire set of observations. There are more robust algorithms such as K-Nearest Neighbors that calculate what should be the imputation value based on proximity measures between observations, but for sake of simplicity and conservatism, we chose to impute the minimum non-missing value of each of these variables within the specific dataset of the Course Unit.

This method was the most conservative path, so it would affect as little as possible the algorithms that would be executed next, both for feature selection and statistical forecasting models, thus avoiding producing extreme values or values that would magnify non-existent trends in the data.

A.3.2.5 Removal of variables from first-year Course datasets

Two attributes that could pose problems regarding data quality and correctness are those that represent student average grades in the 1st semester and in the 1st curricular year. These attributes are available for 1st year Course Units, so if statistical models were to be trained including these variables, which were calculated later and based on the student's performance in this period, would result in a *data leakage* situation, where we would unduly have access to information that shouldn't be available for training models.

For the above reasons, one of the most important local treatments performed was the removal of the variables AVERAGE_1ST_YEAR and AVERAGE_1ST_YEAR_1ST_SEMESTER from all datasets of courses of the 1st curricular year.