FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

# Exploring the use of learning management systems data to early predict students' academic performance

**Pedro Afonso Paulino Ferreira de Castro**

DISSERTATION


Mestrado Integrado em Engenharia Informática e Computação

Supervisor: Vera Lucia Miguéis Oliveira e Silva

June 27th, 2018

# Exploring the use of learning management systems data to early predict students' academic performance

**Pedro Afonso Paulino Ferreira de Castro**

Mestrado Integrado em Engenharia Informática e Computação

June 27th, 2018

# Abstract

Education plays a huge rule in human development. Better education leads to being able to make wiser, more informed decisions. It drives humanity forward, since technology and innovation are strongly tied to education. It boosts a country's economy, since more specific jobs tend to generate more income. Hence, it is a matter of utmost importance to try to find methods that better suit the needs of each specific student, namely whether they are top-performing and need more challenging tasks or if they experience more difficulties and require more individual attention in order to achieve the expected results.

Universities keep enormous amounts of data about students that is often overlooked. If that information can be processed and analysed, conclusions about the students' overall performance can be drawn. Methods such as data mining can be used to achieve this goal. Data mining is the process of discovering patterns in really big data sets and turning them into understandable information that can then be used to develop models that are able to make more accurate predictions about future events.

With this in mind, this project's aim is to predict student success using their sociodemographic data and their academic performance during the first semester, as well as using data from their browsing history in the university's Information System during the first semester of their studies. Moreover, it also aims at understanding how much can the browsing history data improve the accuracy of a predictive model focused in academic performance. The model uses regression techniques to tackle this problem, namely Neural Networks, Support Vector Machines and Random Forests. The project uses the Faculty of Engineering of the University of Porto as case study, with information of 2023 students being used.

According to this investigation, the prediction of the academic success of the students in the end of the first semester seems to be viable and promising for the managers of academic institutions. However, the use of browsing history does not seem to make significant improvement in the predictive models' capacity to evaluate student performance.

# Acknowledgements

First, I'd like to thank my supervisor, Vera Lucia Miguéis Oliveira e Silva, PhD., for all the support given and always being available every time I needed.

Secondly, I'd like to thank my mother, who despite not understanding a single thing about programming, helped me with her constant support, specially when things were going wrong, much more than she could ever possibly imagine.

Lastly, I'd like to thank Flávio Couto, who also supported me a lot during the development of this project, both by giving opinions and suggestions everytime I faced some kind of difficulty and by being a good, supportive friend.

Afonso Castro

# Contents

# List of Figures

# List of Tables

LIST OF TABLES

# Abbreviations

| | |
|---|---|
| FEUP | Faculdade de Engenharia da Universidade do Porto |
| SIGARRA | Sistema de Informação para Gestão Agregada dos Recursos e dos Registos Académicos |
| URL | Uniform Resource Locator |
| LMS | Learning Management System |
| EDM | Educational Data Mining |
| ANN | Artificial Neural Network |
| GPA | Grade-Point Average |
| CGPA | Cumulative Grade Point Average |
| SVM | Support Vector Machine |
| SVR | Support Vector Regression |
| MIEC | Mestrado Integrado em Engenharia Civil |
| MIEEC | Mestrado Integrado em Engenharia Eletrotécnica e Computadores |
| MIEIC | Mestrado Integrado em Engenharia Informática e Computação |
| MIEM | Mestrado Integrado em Engenharia Mecânica |
| MIEIG | Mestrado Integrado em Engenharia Industrial e Gestão |
| MIEQ | Mestrado Integrado em Engenharia Química |
| MIB | Mestrado Integrado em Bioengenharia |
| MIEA | Mestrado Integrado em Engenharia do Ambiente |
| MIEMM | Mestrado Integrado em Engenharia Metalúrgica e de Materiais |

# Chapter 1

# Introduction

## 1.1 Context

Education plays a big part in our society's life. The American radio station National Public Radio described it as "the most important revolution of our time" [Poo14]. Education has several positive effects in our society. It provides economic growth for a country. For every year of education, a person's average earnings increase by 10 percent. By earning an income, people contribute to the country's economy as a whole. It also decreases the gender gap and poverty, while promoting health [CLE14].

We live in a world where the information and communication technologies are constantly growing, seeing more and more use in everything in our everyday lives. Education is no exception. It's becoming more and more common that schools and universities have computers with several types of software, ranging from complete office suites to programming tools. Apart from these, schools and universities are also investing in learning management systems (LMS), such as Moodle and CourseSites, to enhance teacher and student experience. These learning management systems aid in tasks such as project submissions and evaluation, allow teachers to make resources available to students and support forums for students to interact with each other and with their teachers.

The growth of information technologies has also promoted the emergence of tools supported by huge amounts of data. This is the case of data mining tools, which enable the possibility of discovering knowledge in big data repositories. Data mining techniques are used to discover patterns that might have gone unnoticed if they weren't deployed. They also provide capabilities to predict the outcome of future observations [TSK05]. One category of that set of techniques is supervised learning techniques. They receive a set of labeled examples as training data and make predictions for all unseen points [MRT12]. Some examples of data mining techniques that fit the category of supervised learning are some ensemble methods such as Random Forests and AdaBoost, Neural Networks and Support Vector Machines.

## 1.2 Motivation and Objectives

The current situation in what the birth rate of Portugal is concerned is not encouraging. This means that it is crucial that universities improve their overall image, in order to attract the fewer and fewer students that will apply each year to Portuguese universities.

Furthermore, as mentioned in section 1.1, education plays a big part in modern society, being directly linked to key factors such as economic growth, decreasing gender gap and poverty, while increasing health. Guaranteeing the best possible quality of education is, thus, a must. This can be achieved by finding teaching methods that cater to each student's specific needs, whether they are top-performing and need more challenging tasks for boosting their qualities or if they experience more difficulties and require more individual, specialized support to help them overcome their difficulties and achieve their goals.

The data gathered and generated by Learning Management Systems is oftentimes ignored by most institutions. However, these huge datasets hold enormous amounts of information that can (and should) be analyzed in order to improve the quality of the service provided by education institutions. If this information is studied and processed, conclusions from said data can be drawn, which would then lead to better overall quality in the education system. The generated knowledge could also be helpful when changes to the education system are being considered.

Thus, the aim of this study is to apply data mining techniques, more specifically, supervised learning techniques, in order to develop a predictive model capable of predicting a student's overall academic performance on the early stages of their academic career, using sociodemographic data, evaluation data from high school and their first academic semester. Moreover, figuring out eventual relations between academic performance and the use students make of Learning Management Systems is another goal. The results of this study constitute an important tool for institutions as they enable, in an early stage of the students' academic career, to anticipate the potential performance of the students and act accordingly. Academic institutions may design actions to mitigate potential failure and/or to design actions to provide a better experience to those students who may present a very promising academic career.

## 1.3 Dissertation Structure

Besides the introduction, this dissertation contains 5 more chapters. In chapter 2, a review on existing literature on the subject of Educational Data Mining, Learning Management Systems and Data Mining Techniques in Educational Mining is done. In chapter 3, the process followed for tackling the problem at hand will be described. In chapter 4 an analysis of the data is presented. In chapter 5, the results obtained from the generation of the models is described. Finally, in chapter 6, some final remarks are found, as well potential future work for the continuation of this project.

# Chapter 2

# Literature Review

In this chapter, a review on existing literature in the subject of Educational Data Mining (EDM), Learning Management Systems (LMS) and Data Mining Techniques in Educational Mining is done. We start with an overview of EDM, with some work done on each of its main categories. Secondly, we proceed to do an analysis on how Learning Management Systems have been used in EDM and more specifically, in understanding student performance. Lastly, we analyse some of the best performing data mining techniques that are used in the field, in light of the project's goal.

## 2.1 Educational Data Mining

The Educational Data Mining website[1] defines it as "an emerging discipline, concerned with developing methods for exploring the unique and increasingly large-scale data that come from educational settings and using those methods to better understand students, and the settings which they learn in". In other words, EDM tries to find patterns and make predictions that characterize learners' behaviours and achievements by using data mining techniques.

Peña-Ayala [PA14a] noticed that most of the work recently done in EDM could be divided into the following categories:

- **Student Modeling** - Student modeling is defined as attempt to characterize the learner through emotions, cognition, domain knowledge, learning strategies, achievements, features, learning preferences, skills, evaluation, and affects. By understanding the learner as an individual, it's easier to adapt the teaching experience to their needs. Some examples of Student Modeling include a study by Macfadyen and Dawson [MD10] to use LMS generated data to investigate which student online activities (such as total number of discussion messages posted) accurately predict academic achievement, by using regression modeling. Nacu et al. [NMSP16] gathered data from student and teacher activities, interactions between them and LMS to categorize their actions into learning skills and activities such as Self-directed learning, Creative Production and Social Learning.

[1] www.educationaldatamining.org

- **Student Behaviour Modeling** - In this category, the focus is on the behaviour that the learner is likely to have and how can the system be adapted to the user's tendencies. Baker et al. [BKA+] predict drop-outs and school failures using the students' social behaviour. He [He13] analyses online questions and chat messages recorded by a live video streaming. The study identifies discrepancies and similarities in the students' patterns and themes of participation between student–instructor interaction, as well as student–students interaction or peer interaction.

- **Student Performance Modeling** - In these studies, the focus is on estimating and anticipating performance of the students. Indicators of performance such as efficiency, achievement or competence are used. The goal is to estimate how well the learner is or will be able to accomplish a given task, reach a specific learning goal, or appropriately respond to a particular learning situation. Márquez, Romero and Ventura [MRV11] attempt to estimate final student performance and anticipate which students might fail using cost sensitive classification. Guruler et al. [GIK10a] attempt to predict if a student will fail, pass with an acceptable grade or have a good grade by using features such as socioeconomic background, language proficiency and if they receive a grant or not.

- **Assessment** - The supervision and evaluation of learners' domain knowledge acquisition, skills development and achieved outcomes. The purpose is to differentiate student proficiency as well as online and offline assessment. For example, Sohn and Ju [SJ10] perform conjoint analysis to assign weights to four components (an exam, high school grade, an essay and an interview) to help in recruiting high quality university candidates.

- **Student support and feedback** - These studies focus on the support that the system gives to the student and the feedback that the student provides back to the system. For example, Tsuruta et al. [TKD+13] try to find a way of matching a university's offer of courses to the students' needs.

- **Curriculum, domain knowledge, sequencing, and teachers support** - These studies focus on the customization of curriculum and teaching practices with the purpose of making the acquisition of domain knowledge easier to learners. Teachers support is the support teachers give to learners to make them achieve the aforementioned goal of acquisition of domain knowledge. In this category, for instance, Gaudioso et al. [GMHdO12] developed predictive models to assist students when they face problems, guiding them through the course materials in order to improve the effectiveness of the learning process.

Since this project's goal is to predict student performance using predictive models, it fits the **Student Performance Modeling** category the most. The prediction of a student's performance is a challenging problem, due to the myriad of characteristics and circumstances that might influence it. Socio-demographic information, such as age and gender has been used extensively in these studies, as well as information about prior studies, such as GPA in previous semesters, in high

school or marks in previous assignments [HS17, ABM13, AAK+16, LRSM15], if the student is studying in part or full-time [NZ14], economic factors such as the existence of a scholarship, if the student borrowed money or the financial situation of the family/parents [GIK10b, SCS+15], degree of development in certain soft skills, such as leadership and decision making [MKG14], behaviour (such as presence in classes, doing homework) [VMS07, WW14], level of peer support [WW14], a student's own perception of himself (e.g. probability of succeeding, confidence degree) [VMS07, MB12] and extra-curricular activities [NZ14].

### 2.1.1 Learning Management Systems and Academic Performance

Learning Management Systems and their use for predicting academic performance is a common field of study in EDM. Some studies that were already mentioned in section 2.1, like Macfadyen and Dawson's [MD10] use of LMS generated data to determine which online activities impacted student performance the most and are prime examples of how LMS's play a key role in EDM.

Information commonly extracted from LMS's includes number of LMS sessions, total session time [BSAD13, Pal13, MD09], date of first/last login to the LMS [BSAD13, Pal13], total number of individual LMS pages viewed [Pal13] and the number of actions taken [CAP+16]. Variables refering to LMS forum use have also been considered relevant. Examples of variables reltated to this issue include the total number of LMS discussion postings read and made [REZ+13, Pal13, MD09, JVMM12] and the number of words posted in said discussion postings [CAP+16]. The individual visualizations of each resource made available [LLM+14], the number of quizzes and assignments done [BSAD13, MB12, REZ+13, MD09], the grade obtained in graded activities [BSAD13] or if they passed or not [REZ+13, JVMM12], date and time taken to complete quizzes and exams [MB12, REZ+13, MD09, JVMM12, CAP+16], time in the discussion postings [JVMM12, CAP+16] and the number of days taken to turn in a task after it was assigned [CAP+16] have also been explored.

### 2.1.2 Data Mining Techniques in Educational Mining

Baker, in his review of the State of Educational Data Mining, in 2009 [BY09], proposed a classification of Data Mining methods in EDM that splits each method in prediction (which is then split into classification, regression and density estimation), clustering, relationship mining (which is then divided into association rule mining, correlation mining, sequential pattern mining and causal data mining), distillation of data for human judgment and discovery with models.

Considering the scope of the project developed, we will focus our analysis in prediction methods, more specifically, in classification and regression methods.

There have been attempts to understand which classification/regression methods (and data mining methods in general) provide best performance for educational data mining problems. For instance, Saranya et al [STU+13], in their survey of data mining techniques available for EDM, mention neural networks, decision trees and logistic regression as good data mining techniques for creating predictive models. Sen et al. [D12], in their attempt to identify the factors that lead

students to success or failure in tests, use decision trees, support vector machines (with non-linear kernel functions) and neural networks (with multi-layer perceptrons). Macfayden and Dawson [MD10] use logistic and multiple regression for their work on student online activities that help academic achievement mentioned in section 2.1. Bydžovská [Byd] built a classifier based on student data such as gender, year of birth, year of admission, number of credits gained from passed courses and average grades in order to evaluate the performance of classification and regression techniques and came to the conclusion that support vector machines, linear regression, additive regression and decision trees (specifically, RepTrees and Random Forests), were the regression algorithms that achieved the best results. Support Vector Machines were used by Strecht et al. [SCS+15] in both a regression and classification problem in their comparative study of algorithms for modelling student academic performance. They used data of 5779 students from 391 programmes. Artificial Neural Networks were used by Hoffait and Schyns [HS17] to determine if students have a good chance to succeed in their first academic year, if they are likely to fail, or if their outcome is uncertain. The dataset was composed of 6845 first year students. Random Forests were used by Vandamme et al. [VMS07] with a group of 533 first-year university students to determine if a student has a low, medium or high risk of failing their first academic year. Decision Trees were used by Mccuaig and Baldwin [MB12] to attempt to predict the grade (A, B, C or D, F) of 122 first year students using information about their LMS interactions and a survey to the student's confidence in their skills. Aluko et al. [AAK+16] used the K-nearest neighbour algorithm to determine the Cumulative Grade Points Average (CGPA) of 102 students using grades from previous exams. Marbouti et al. [MDDM16] used the Naive-Bayes algorithm to determine a student's grade in a course according to the learning objectives that they have shown to have met from one written exam, 10 quizzes and five homeworks of 3063 students.

This dissertation differs from the works previously mentioned since it makes use of logs from a university's information system from several years to determine the use students make from them: some studies make use of LMS data, but generally they skip aspects such as when in the semester (early or late, meaning they start working early in the semester or they leave things for the last moment) do they use them or information about number of accesses to course pages or course contents.

# Chapter 3

# Methodologies

The purpose of this project is to develop two prediction models that predict student academic performance. The first model will focus on using sociodemographic and academic information available by the end of the first semester, whereas the second model will use information pertaining to the usage habits of the students of SIGARRA. The dependent variable, in both predictive models, is a score calculated using the following formula:

$$\frac{GPA * CompletedCredits}{EnrolledCredits}$$

Where *GPA* represents the Grade Point Average of the student by the end of the degree, *Completed Credits* represents the total amount of ECTS the student completed and *Enrolled Credits* the total amount of ECTS the student enrolled in. ECTS (acronym for European Credit Transfer and Accumulation System) is an European standard for comparing the "volume of learning based on the defined learning outcomes and their associated workload" for higher education across the European Union and other collaborating European countries [Com15]. For successfully completed courses, ECTS credits are awarded. One academic year corresponds to 60 ECTS credits that are normally equivalent to around 1600 hours of total workload, irrespective of standard or qualification type.

Using this formula means that not only we consider the GPA of the student, but also how many credits (and consequently, courses) did they complete on the first try. Thus, we favour students who completed every course on the first try, penalizing students who took longer to complete their degree.

For the remainder of this chapter, the methodologies used in the project will be depicted. Firstly, an analysis to the data mining process will be done. After that, each step of said process will be analyzed with further depth, with special focus on the decisions taken in each one of them.

## 3.1 The data mining process

Before data is converted into knowledge, it needs to go through a process composed of several steps. Figure 3.1 illustrates the basic flow of this process. It should be noted that there could be

an existence of loops between several steps of the process, according to the needs of the person interested in the knowledge and the results obtained.

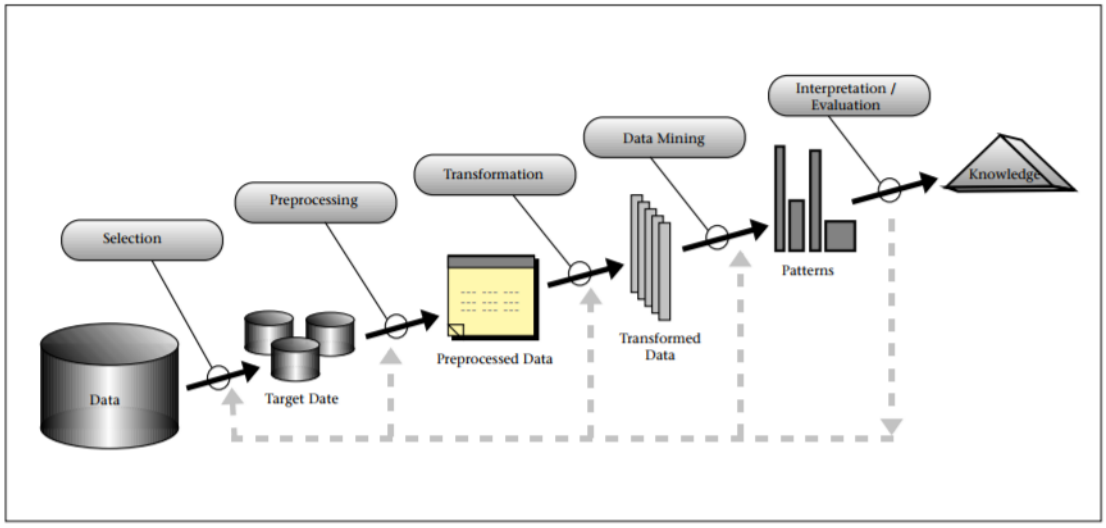



Figure 3.1: An Overview of the Steps That Compose the Data Mining Process [FPSS96a]

The data mining process is composed of the following steps [FPSS96a]:

- The first step consists of developing an understanding of the problem's domain and the prior knowledge required, while identifying the goals that we want to achieve with the data mining process. This step has already been described in previous sections.
- The second step is creating a target data set through **selection** of a data set or a subset of data on which knowledge is to be discovered.
- The third step involves data cleaning and **preprocessing**, through methods such as removing eventual noise in the data and deciding on strategies to deal with missing data fields.
- The fourth step involves data reduction: finding the useful features in the data and then applying the necessary **transformation** methods allows us to select and generate the variables that will be used for the model.
- The fifth step consists of matching the goals to a data mining method, such as summarization, classification, regression or clustering.
- The sixth step consists of choosing the data mining algorithms to be used for searching for patterns. It also includes deciding which models and parameters might be appropriate for finding said patterns in data sets.
- The seventh step is the **data mining** itself - searching for the patterns in data we're interested in.

- The eighth step involves the **interpretation** of the mined patterns, **evaluating** said patterns. This step will be discussed in section 6.

- The ninth step is acting on the discovered knowledge. Obviously this step is generally up to the person or organization that requested the process.

### 3.1.1 Data Selection

The data used for this project is the academic information of students who were in their first academic year in the academic years between 2006 and 2011. A class diagram of this data can be seen in figure 3.2.

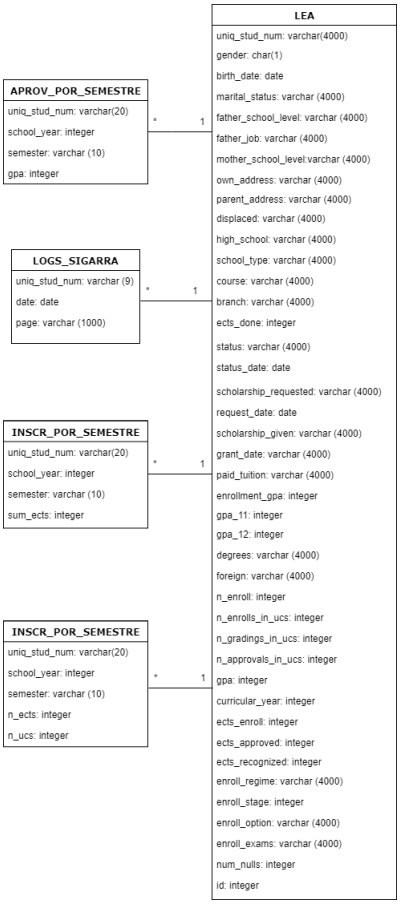



Figure 3.2: Class Diagram of the data

By analyzing the diagram, some attributes stand out as good potential candidates for being used for the model creation:

- **Sociodemographic information**, such as the student's date of birth, gender and marital status of the student, their parents' education level and job, the type of school they attended in high school, whether they applied for a scholarship or not and if they were granted the aforementioned scholarship or not.

- **Academic information**, such as the degree the student is enrolled in, the GPA of the student in the last year of high school, their enrollment GPA (the value used to determine who is chosen and who is not, obtained in the end of high school), their GPA at the end of their first academic year and the degree, the option they choose in the application form. We also have information on how many ECTS they were enrolled in, how many they were approved in and the GPA of each semester they studied at FEUP.
- **Browsing logs** of the student in SIGARRA by date/time.

Some variables contain information that was not available by their first academic semester and cannot be used for creating the model, such as diplomas they may have obtained, mobility, number of years they were enrolled in the course, total number of courses they were enrolled, evaluated and approved in, the final GPA in the degree and number of ECTS enrolled in, approved and acknowledged.

This leads us to the main part of the data selection step: **feature selection**. Feature selection consists in selecting a subset of attributes from a dataset for using in model construction. The following variables were selected:

- age: age of the student
- gender: gender of the student
- marital_status: marital status of the student
- father_school_level: Father's education level
- father_job: Father's job
- mother_school_level: Mother's education level
- mother_job: Mother's job
- school_type: Type of school (public/private)
- status: The status of the student (student-worker, ordinary, etc)
- degree: degree the student is enrolled in.
- scholarship_requested: whether the student applied for a scholarship or not
- scholarship_given: whether the student was granted a scholarship or not
- gpa_12: the student's GPA in the last year of of high school
- enrollment_gpa: the CGPA (Cumulative GPA) of the student used to apply to a degree
- stage: the application phase the student was accepted in his degree

- enroll_option: where the degree was in the student's application form

Other variables that also provide useful information were not directly included as independent variables for the model. Instead, they went through variable **transformation**. The transformations they were subjected to will be analysed with further detail in section 3.1.3. These attributes include the birth date of the student and all the information on how many ECTS they were enrolled in, how many they were approved in and the GPA per semester.

### 3.1.2 Data preprocessing

The dataset includes data from students that have already completed their degree and students that are still enrolled. Since we do not have information on the final grade of students that have yet to complete their degree, only the students that have already completed it were selected for the models.

Only students that are enrolled in their degree through the regular application phase (known as "Contingente Geral") were chosen. This means that no foreign students are included, as well as students that don't come from mainland Portugal (eg. Azores and Madeira).

The GPA of the student in the $12^{th}$ grade and the enrollment grade were also stored in an inconsistent way. Some records used a 0-20 scale whereas others used a 0-200 scale. The last ones were converted into a 0-20 scale.

Rows that hold polynomial values had nonexistent values filled with the value "Unavailable". This decision was made since most of the polynomial attributes already had a "Unavailable" possible value that has the same meaning.

**Outlier detection and removal**, a common part of data preprocessing that consists in removing samples that are too distant from other observations was not done for this project. The reason for this is that we are trying to predict student performance, which means that the concept of an outlier does not make much sense, since we want the model to consider distant values (these distant values represent a very important part of the dataset and should definitely be considered for the model).

Another common part of preprocessing is **data normalization**. Data normalization consists of reducing every feature to the same scale, in order to avoid that some features influence the model more than others due to having a wider range of possible values. Unlike outlier removal, this step is crucial for our dataset because we have several scales in our data (from integers that go from 1 to n, to probabilities that range between 0 and 1). Thus, every numerical feature was normalized to a value between 0 and 1, to guarantee that they all have the same scale. The algorithm chosen to perform this step was Min-Max Scaler. This scaler makes use of the following formula to normalize the numerical attributes:

$$y = \frac{x - min}{max - min}$$

Where $x$ is the value to be normalized, $max$ is the largest value in the pool of data to be normalized and $min$ is the smallest one.

### 3.1.3 Data transformation

Although the dataset contained the date of birth of the students, the big granularity that comes associated with it means that it is not very suitable for knowledge discovery, as is. This led into the decision of transforming this variable into an age variable. To obtain the age, the birth date of the student is subtracted to the school year of the grades record, which is obtained from the semester records (GPA, courses enrolled and approved and student status).

A score for the first academic semester is calculated in order to be used as an independent variable. This score is obtained using the following formula:

$$\frac{GPA * CompletedCredits}{EnrolledCredits}$$

Again, *GPA* represents the GPA of the student in the first semester, *Enrolled Credits* the credits the student enrolled in the first semester and *Completed Credits* the ones they completed.

Regarding SIGARRA's logs data, upon closer inspection, one can see that the only information that is stored is the URL that the student requested to the server. This information is not very useful because, like the birth date, has too much granularity (for example, if a student requests a course content, information about which specific resource was requested is included in the URL - we are not interested in the resource, but only in the fact that the student requested a resource). Thus, the URLs have been categorized. Three categories have been created: for every page that involves course content, for every student profile page and for every course page.

From this characterization, several variables have been generated:

- A variable for the number of sessions in each weekday and month of the first semester was added. A session is defined as "a sequence of page requests that start with a login into the system".
- A variable for the number of times they requested course content and another one for the percentage of those requests in comparison to the total amount of individual pages accessed. Four variables with similar information regarding access to the student's personal profile and to course pages have also been generated.
- A variable for each student's total number of sessions.
- A variable for the number of days that passed until the first access to SIGARRA by the student, in comparison to the first access in the dataset for the whole set of students in the first semester of the first academic year.

A table with the variables used in the models can be found in appendix A.

### 3.1.4 Data Mining classes

Data Mining methods can generally be split into several categories. The choice of one category over another is based on what the user wants to do and the kind of data we have. Said methods are as follows [FPSS96a]:

- **Clustering**- identification of a finite set of categories or clusters to describe the data. The categories can be mutually exclusive and exhaustive or consist of a richer representation, such as hierarchical or overlapping categories.
- **Summarization**- consists of finding a compact description for a subset of data.
- **Classification**- consists of generating a function (or model) that maps (classifies) a data item into one of several predefined classes. The output is, thus, composed of categorical variables.
- **Regression**- consists of learning a function that maps a data item to a real-valued prediction variable. The output being real variables is the main difference between regression and classification.
- **Dependency modeling**- consists of finding a model that describes significant dependencies between variables.

As said before, our objective is to create two predictive models that have two real values as dependent variables. This means that the most suitable category of methods to use is **regression**. Since we have labeled data that can serve as test data, and we want to generate a predictive model capable of predicting student performance using regression methods, we are interested in **supervised learning**. In supervised learning, labeled training data is used. A supervised learning algorithm analyzes the training data and produces an inferred function or model, which can be used for mapping new examples. Throughout the rest of this section, the supervised learning regression algorithms that better apply to the problem under analysis will be described.

### 3.1.5 Data Mining algorithms

#### 3.1.5.1 Random Forests

Random Forests are part of a class of algorithms called ensemble methods, that is, methods that employ multiple algorithms in order to produce better results than those obtained by using any of those algorithms alone. Random Forests can both be used for classification and regression. Random Forests employ the use of several decision trees where each tree is generated based on a subset of the original data set, each subset being independent from the others. This randomization helps reduce the correlation between the decision trees, which means the generalization error of the ensemble method can be improved.

Due to its ensemble nature, Random Forests are very accurate. They are also very robust to noise and tend to run faster compared to other ensemble methods) [PNSK06].

#### 3.1.5.2 Artificial Neural Networks

Artificial Neural Networks are a data mining method inspired by attempts to simulate animal neural systems. They are composed of several **input nodes** and an output node (which are the

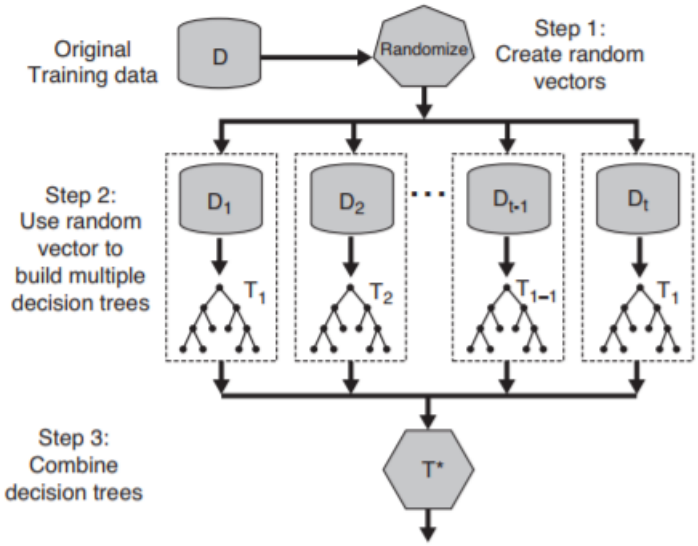



Figure 3.3: Random forests [PNSK06]

equivalent to neurons) connected via a **weighted link** (the equivalent to a synapse). The input and output nodes with the weighted links are called **perceptrons**. A schematic representation of a perceptron is shown in figure 3.4.

In this example, the inputs $x_1$, $x_2$ and $x_3$ take the value 0 or 1. The perceptron then computes its output value, which is either 1 or -1, by performing a weighted sum on each of the input nodes and then subtracting a bias value. So, assuming a bias of 0.4, the output generated by the perceptron would be the one listed in figure 3.5.

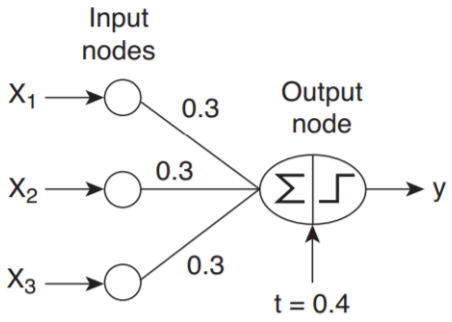



Figure 3.4: A perceptron.

$$\hat{y} = \begin{cases} 1, & \text{if } 0.3x_1 + 0.3x_2 + 0.3x_3 - 0.4 > 0; \\ -1, & \text{if } 0.3x_1 + 0.3x_2 + 0.3x_3 - 0.4 < 0. \end{cases}$$

Figure 3.5: Output of a perceptron.

The weight of each input value is then changed on each iteration, according to the following formula:

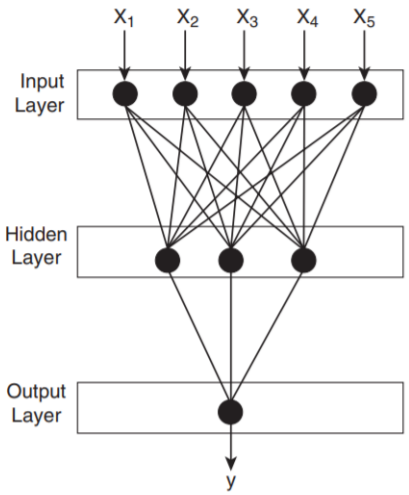
$$w_j^{(k+1)} = w_j^{(k)} + \lambda\left(y_i - \hat{y}_i^{(k)}\right)x_{ij}$$

Where $w^k$ is the parameter associated with the $i^{th}$ input link after the $k^{th}$ iteration. $\lambda$ is a parameter known as the **learning rate** that varies between 0 and 1 and $x_{ij}$ is the value of the $j^{th}$ attribute of the training example $x_i$. The new weight is a combination of the old weight and a term proportional to the prediction error (y - ŷ). If the prediction is correct, the prediction error equals 0 and the weight of the input node is not changed. However, if the prediction is wrong, it gets modified considering the following criteria:

- If y = +1 and ŷ = -1, then the prediction error is 2. To compensate for the error, the weights of links with positive inputs are increased and the ones with negative inputs are decreased.
- If y = -1 and ŷ = +1, then the prediction error is -2. To compensate for the error, the weights of links with positive inputs are decreased and the ones with negative inputs are increased.

The learning rate ($\lambda$) also has direct influence in the weight of the input nodes. If $\lambda$ is close to zero, then the new weight is mostly influenced by the old value for the weight, whereas if it is closer to 1, it is mostly influenced by the value generated for the new iteration.

Artificial Neural Networks with multiple layers can also be created, using the output nodes of a perceptron as input nodes of the next layer.




The aforementioned method's perceptrons only output a limited number of values (either -1 or 1). That makes them unsuitable for regression methods. However, it is possible to make the method account for a discrete output, thus making using ANNs in regression problems viable. That is achieved through using a different **activation function** other than a sign function. Examples of other activation functions are shown in figure 3.6.

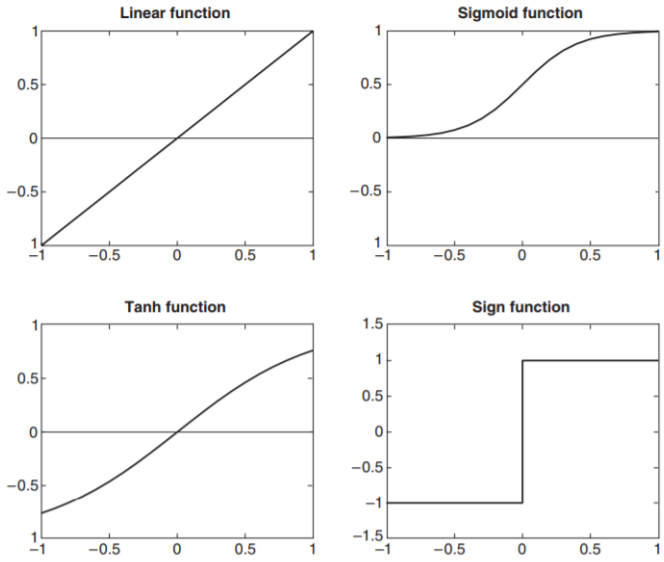



Figure 3.6: Types of activation functions in artificial neural networks

By using activation functions such as a Linear, a Sigmoid or a Tahn function, we can model a discrete variable, thus making it possible to use ANNs in regression methods.

#### 3.1.5.3 Support Vector Machine

**Support Vector Machine** is a method that when presented with a set of objects belonging to one of two possible values in an ambient space with n dimensions, builds a subspace with n-1 dimensions (a hyperplane) that separates the objects into one of the two categories. The hyperplane must be the **maximal margin hyperplane**.

Figure 3.7 represents a plot of a dataset of examples that belong to two different classes, represented as squares and circles. As we can see, there are infinite possible hyperplanes (represented by the lines) that classify the dataset perfectly.

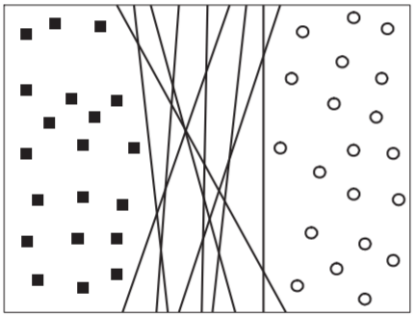


Figure 3.7: Possible decision boundaries for a linearly separable data set [PNSK06].

In figure 3.8, we can see two of those hyperplanes (represented by $B_1$ and $B_2$. Each of them is associated with a pair of hyperplanes $B_{i1}$ and $B_{i2}$. They are obtained by moving a parallel hyperplane until it touches the closest square and circle, respectively. From this, we can see that

the distance between $B_{11}$ and $B_{12}$ is considerably higher than the distance between $B_{21}$ and $B_{22}$. This distance is called the margin of the hyperplane and the **maximum margin hyperplane** is the largest possible margin hyperplane, which in this scenario turns out to be $B_1$. The reason why we want the largest margin hyperplane is because decision boundaries with large margins tend to have better generalization errors, which makes them less prone to overfitting and more suited to correctly classify previously unseen examples.

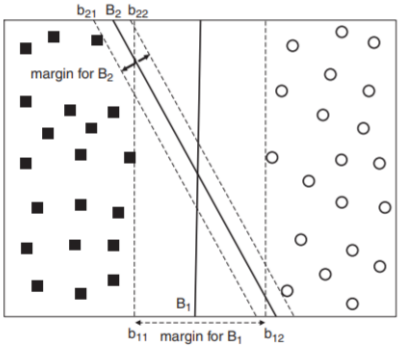



Figure 3.8: Possible decision boundaries for a linearly separable data set [PNSK06].

A linear SVM is a classifier that searches for the maximum margin hyperplane. Considering a classification problem consisting of N training examples where each example is denoted by a tuple ($x_i$, $y_i$), where $x_i$ represents the set of values for each attribute and $y_i$ the class label. The decision boundary of a linear classifier can be written in the following form:

$$\mathbf{w} \cdot \mathbf{x} + b = 0,$$

where $\mathbf{w}$ and $b$ are parameters of the model. If we have two points located on the decision boundary, then we have:

$$\mathbf{w} \cdot \mathbf{x}_a + b = 0,$$
$$\mathbf{w} \cdot \mathbf{x}_b + b = 0.$$

Subtracting the two equations will lead to:

$$\mathbf{w} \cdot (\mathbf{x}_b - \mathbf{x}_a) = 0,$$

where $x_b$ - $x_a$ is a vector parallel to the decision boundary. Since the dot product is zero, $\mathbf{w}$'s direction must be perpendicular to the decision boundary.

For any point located above the decision boundary, we have:

$$\mathbf{w} \cdot \mathbf{x}_s + b = k,$$

where $k > 0$. For any point located below the decision boundary, we have:

$$\mathbf{w} \cdot \mathbf{x}_c + b = k',$$

where k' < 0. If we label the formers as 1 and the latters as -1, then we have the following constraints:

$$y = \begin{cases} 1, & \text{if } \mathbf{w} \cdot \mathbf{z} + b > 0; \\ -1, & \text{if } \mathbf{w} \cdot \mathbf{z} + b < 0. \end{cases}$$

This technique works if constructing a linear decision boundary is possible. If that is not the case and a nonlinear SVM is required, a transformation of the data from its original coordinate space in **x** to a new space $\phi(x)$ is required. After the transformation is complete, the methodology previously described can be applied [PNSK06].

The technique described above is only suitable for classification problems. However, it is also possible to use SVMs with regression problems. If, instead of the previous constraints, we have the following:

$$y_i - wx_i - b \leq \varepsilon$$
$$wx_i + b - y_i \leq \varepsilon$$

we are saying that we want a hyperplane that has points on either side, but the distance between these points and the line must not be greater than $\varepsilon$. Thus, we create a hyperplane in the middle of the set of points making them as close as possible. This technique is commonly referred to as **Support Vector Regression** (SVR). Instead of focusing on minimizing the training error through finding the maximum margin hyperplane, SVR focuses on minimizing the generalization error to achieve better performance.

### 3.1.6 Validation methods

In this section methods to validate the predictive model generated will be described.

#### 3.1.6.1 Cross Validation

In cross validation, each record is used the same number of times for training and one time for testing. The data is split into N subsets. N-1 subsets are used as training data and 1 subset is used as test data. This procedure is then repeated N times, using a different subset for testing each time. This guarantees that each subset is used as test data exactly once.

This approach has the advantage of using as much data as possible for training, while the test data also covers the entire data set. This makes the use of cross validation very useful for parameter tweaking. However, the obvious drawback is that it's a computationally expensive task to perform, since the procedure is repeated N times [PNSK06]. However, our dataset is not large enough to make the extra computation power needed a problem, which means cross validation is an appropriate validation technique for this scenario.

#### 3.1.6.2 Performance Metrics

Performance metrics are used to understand how accurate a predictive model is. While it is obviously impossible to know for sure how will a model fare when classifying never seen data, it is possible to estimate its accuracy using the information currently possessed.

The methods that estimate the accuracy of a predictive model that will be described in this document are $R^2$, Mean Absolute Error and Mean Squared Error.

$R^2$, also called coefficient of determination, is a value that can be interpreted as the correlation between the predicted and the observed variables.

The Mean Absolute Error (MAE) measures the average of the difference between two variables, in this case, the value obtained by the predictive model for a test record that that test record's actual value for that feature. It is thus given by:

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

Derived from the Mean Absolute Error, the Mean Squared Error is, as the name implies, the Mean Absolute Error, squared. The key differences between them is that large errors have relatively greater influence on Mean Squared Error than they do the smaller error, which makes the MSE good for situations where big errors are very costly. The performance metrics used were the MSE and $R^2$.

### 3.1.7 Technologies

Currently, there are a lot of data mining technologies that offer the implementation of the appropriate methods for the development of a predictive model using supervised learning. In this section, the technologies used in this project will be listed. Since the project is being developed in an academic context, only free technologies were considered.

Table 3.1: Overview of Data Mining Tools

| Name | Description |
|---|---|
| *RapidMiner* [1] | RapidMiner is a data science software platform that provides an integrated environment for data preparation, machine learning, deep learning, text mining, and predictive analytics. It supports all steps of the machine learning process including data preparation, results visualization, model validation and optimization. |
| *R* [2] | R is a language and environment for statistical computing and graphics. R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering,...) and graphical techniques, and is highly extensible. |

*RapidMiner* was used for doing the preprocessing and the transformation steps, whereas *R* was used for the model creation and data mining.

# Chapter 4

# Data Exploration

In this section, an analysis to the dataset that will be fed to the models will be done, with a characterization of the students that compose the dataset as well as eventual correlations that might exist between the dependent and independent variables.

## 4.1 Sample Analysis

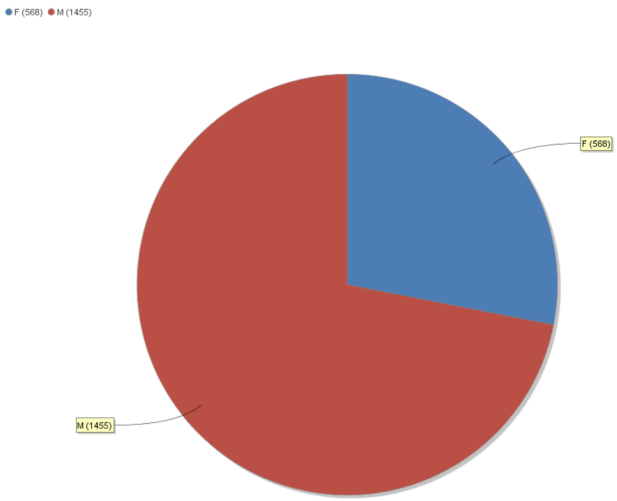



Figure 4.1: Gender distribution in the dataset.

In figure 4.1, we can see that the population of the dataset is predominately male, with 71.9% of the records being of said gender, in contrast with the 28.1% of female people.

In figure 4.2, we can see that the majority of the population is either 17 or 18 years old, with some 19 year old students. This is expectable, since students usually enroll in college in Portugal by the age of 17/18 (depending on whether their birthday is before or after September).

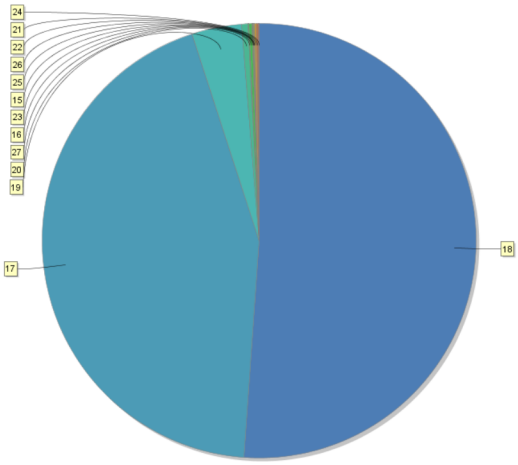



Figure 4.2: Age distribution in the dataset.

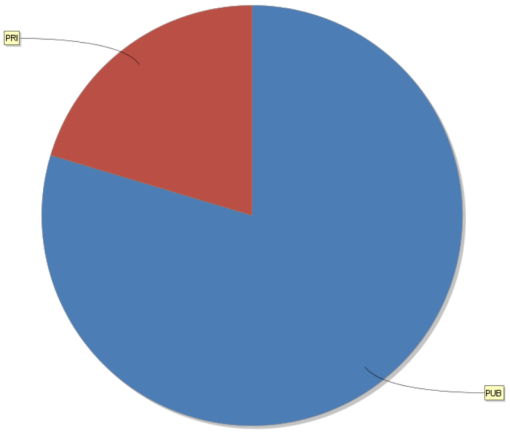



Figure 4.3: Distribution of the type of school the student came from in the dataset.

In figure 4.3, we can see that the majority of the students in the dataset come from public schools, with 20.4% coming from private schools and 79.6% coming from public schools.

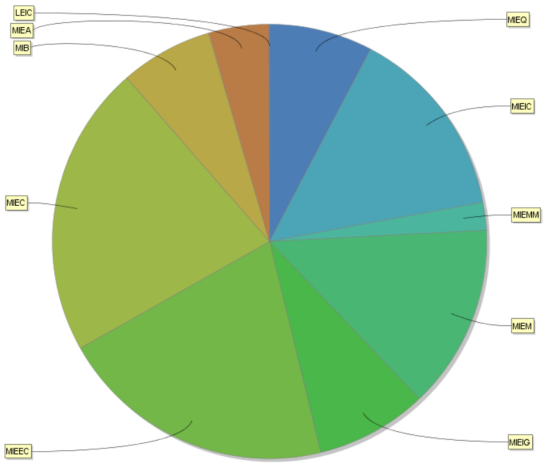



Figure 4.4: Distribution of the degree the student is enrolled in in the dataset.

In figure 4.4, we can see the distribution of the degrees the students are enrolled in, in the dataset. The dataset doesn't lean too much towards a specific degree, so we have a good representation of every degree.

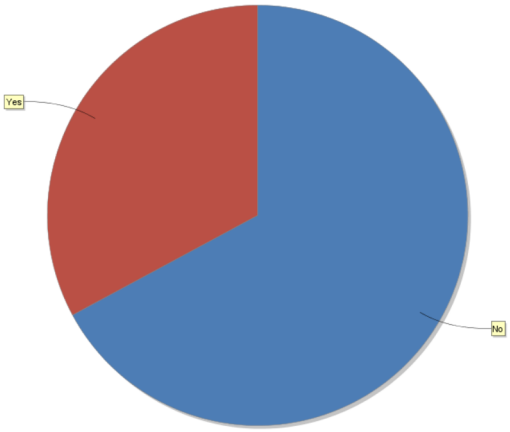



Figure 4.5: Did the student ask for a scholarship?

In figures 4.5 and 4.6, we can see that most of the students (67.1%) didn't ask for a scholarship. The percentage of students granted with a scholarship is 21.2%.

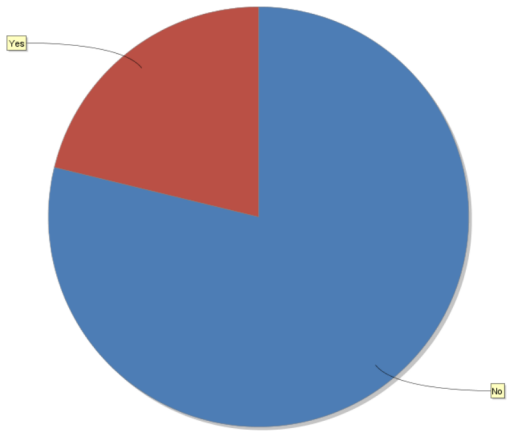



Figure 4.6: Was the student granted a scholarship?

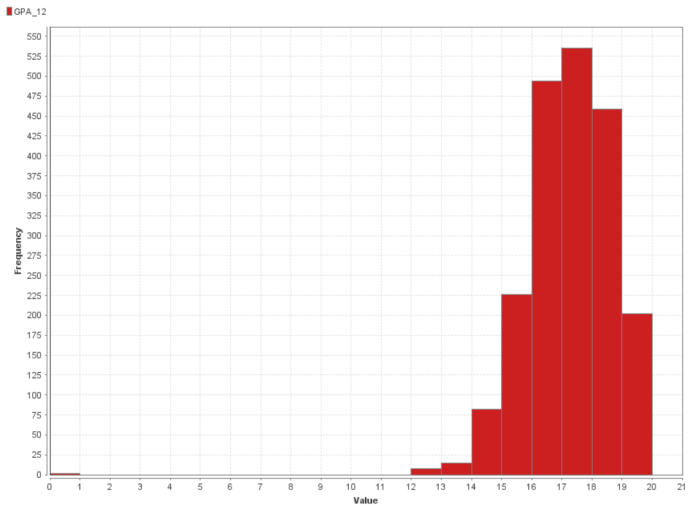



Figure 4.7: Distribution of the grades in the last year of high school

In figure 4.7, we can see the distribution of the grades in the last year of high school by the students. We can see that most students had a GPA between 15 and 16, with the higher frequencies being between 15 and 19. The average grade is 16.874. From this, we can conclude that the dataset is made of good students, since the grades are well above the Portuguese average (a fact that can be confirmed by checking the DGES - Direção Geral do Ensino Superior's website [1]).

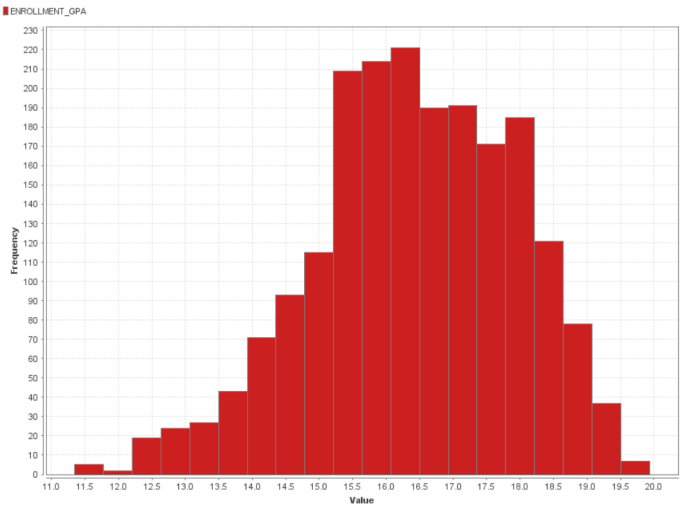



Figure 4.8: Distribution of the enrollment grades of students

In figure 4.8, we can see the distribution of enrollment grades of the students. The average is 16.415, slightly lower than the high school's last year's grades seen in figure 4.7. Most grades are in the 15 to 18 range.

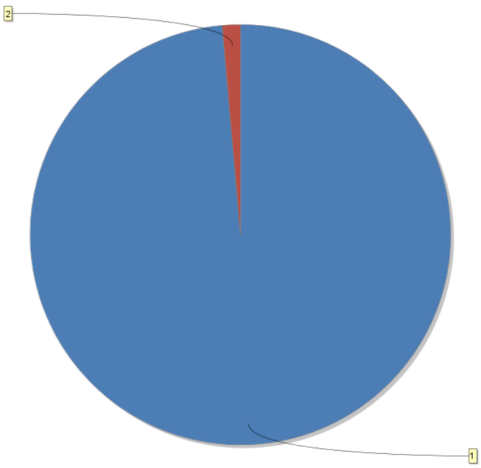



Figure 4.9: Distribution of the enrollment phase of students

[1] http://www.dges.gov.pt

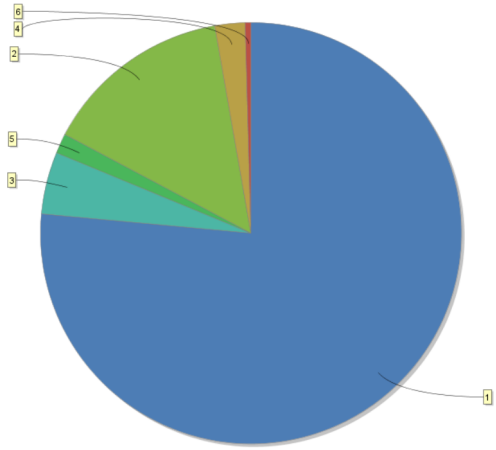



Figure 4.10: Distribution of the enrollment options of students

In figure 4.9, we can see that most students in the dataset enrolled in their degree in the first phase, which suggests that they enrolled in the degree they mostly wanted - a fact further corroborated by figure 4.10, where we can see that the vast majority of students enrolled in their first option.

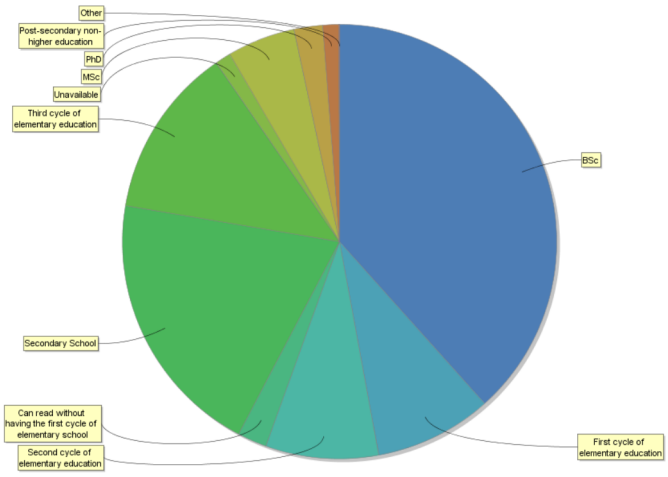



Figure 4.11: Distribution of the students' mother's education

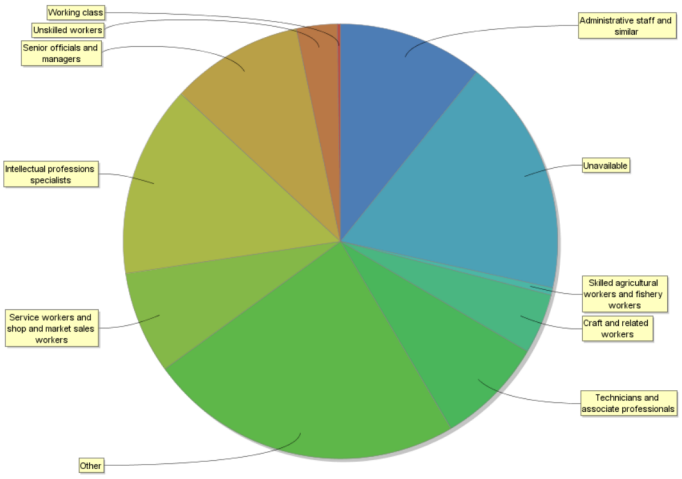



Figure 4.12: Distribution of the students' mother's jobs

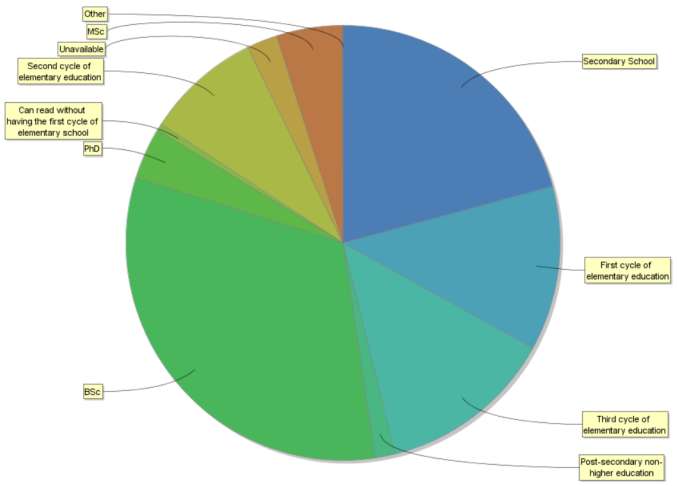



Figure 4.13: Distribution of the students' father's education

In figures 4.11, 4.12, 4.13, 4.14 one can see that most students have parents with tertiary education, with secondary education also having strong representation in the dataset. In what concerns their jobs, intellectual jobs are a common occurrence in both the father and mother's students. However, their mothers tend to work more in administration, whereas the fathers tend to lean more towards being technicians and senior officials/managers.

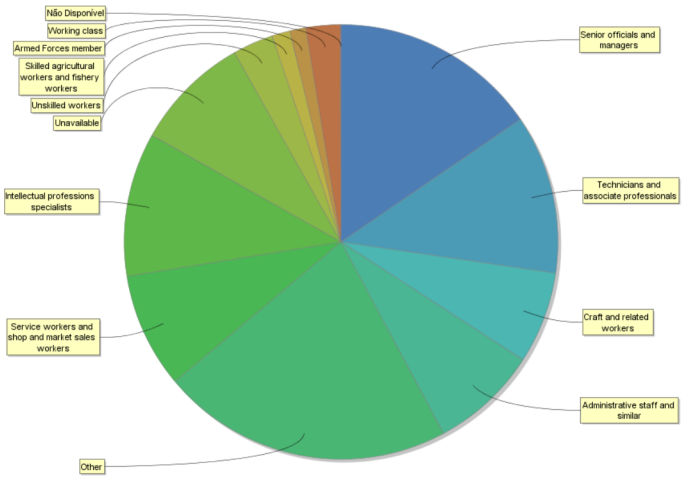



Figure 4.14: Distribution of the students' father's jobs

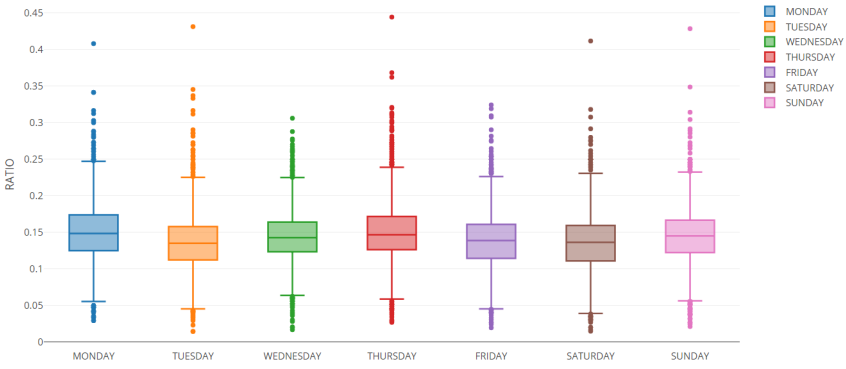



Figure 4.15: Distribution of the students' ratio of SIGARRA sessions per week day

Moving the focus to the browsing history of students, in figure 4.15, it can be seen that they seem to not display any relevant trends in when in the week they access SIGARRA. There is no clear tendency for having a bigger ratio of sessions on a specific week day, nor a clear tendency in what weekdays and weekend is concerned.

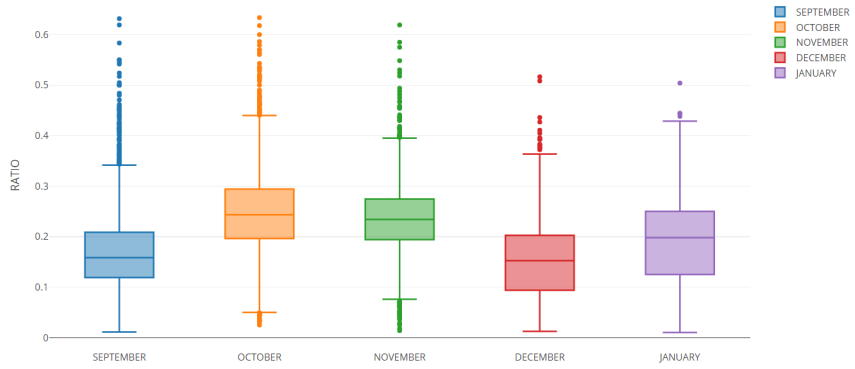



Figure 4.16: Distribution of the students' ratio of SIGARRA sessions per month

In figure 4.16, it can be seen that most of the accesses to SIGARRA happen in October, November and January. This can mostly be seen from an analysis to the quadrants and the median. The extremes, seem to also corroborate this tendency, since the months with the higher maxmimum were September, October and January. September and December having less representation can be explained by the fact that in September, classes have just started and thus there's not as much need to access SIGARRA, whereas in December it can be explained by the winter break.

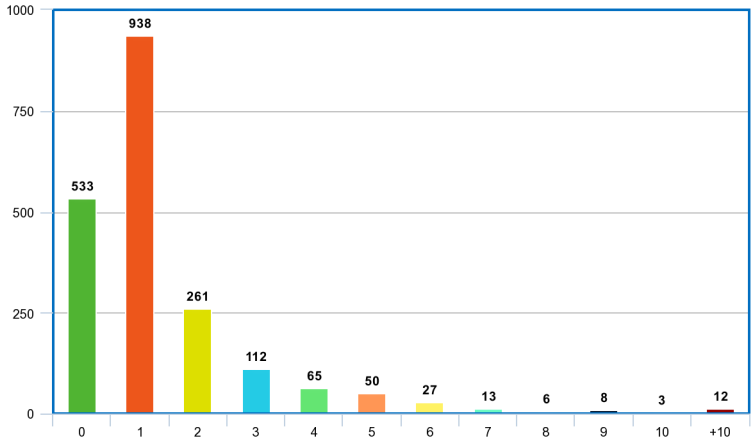



Figure 4.17: Distribution of the days that passed until the first access done by the student since the first access done by a student in the first semester of the first year.

As one can see from figure 4.17, most students accessed SIGARRA for the first time very quickly, with the big majortity of them accessing the Information System in 3 days or less.

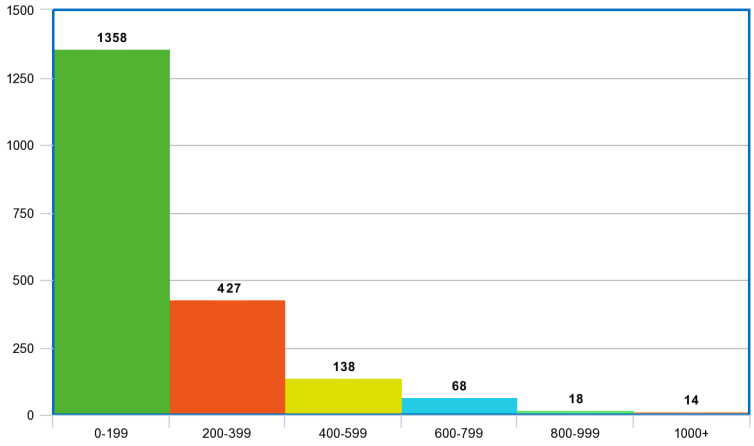



Figure 4.18: Number of accesses to course content in SIGARRA.

In figure 4.18, it can be seen that most students didn't access course content more than 100 times, with the vast majority having accessed between 0 and 200 times. It should also be noted that a significant amount of students that are in the 0 to 100 range have no records of accessing course content (959).

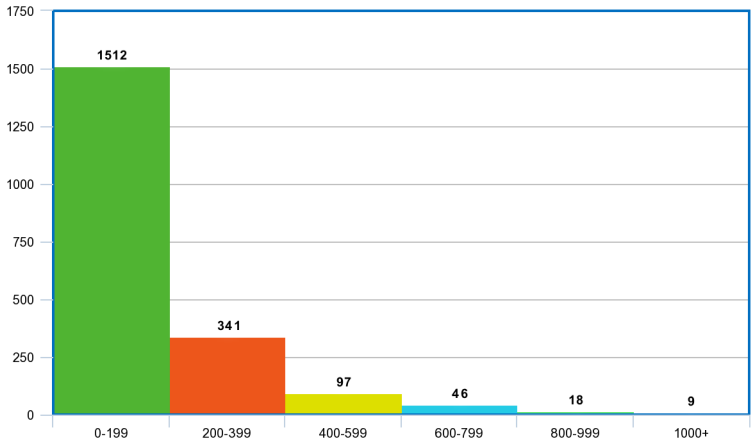



Figure 4.19: Number of accesses to the profile page in SIGARRA.

Similarly to the course content case, it is noticeable in figure 4.19 that most students didn't access their profile page more than 100 times, with the vast majority having accessed between 0 and 200 times. Once again, a significant amount of students that are in the 0 to 100 range have no records of accessing their profile (956).

Finally, one can see in figure 4.20 that like in the previous cases, the vast majority of students accessed a course page between 0 and 200 times. 957 of them have no records of ever accessing a course page.

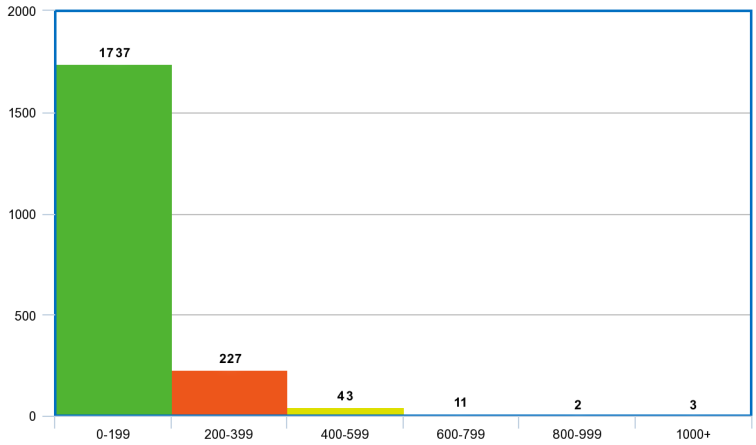



Figure 4.20: Number of accesses to a course's page in SIGARRA.

All in all, it can be concluded that the dataset is mostly composed of male students ranging between being 17 and 19 years old, coming from public school, with both good high school and enrollment grades (between 15 and 19 and between 15 and 18, respectively). Both parents tend to have tertiary or at least secondary education, while usually working in intellectual, technical or administrative jobs. They tend to access SIGARRA mostly in October, November and January.

## 4.2 Relations between academic results

In this section, some conclusions will be drawn through analysis of how the attributes correlate with the dependant variable, i.e. the academic performance score being used in the dataset.

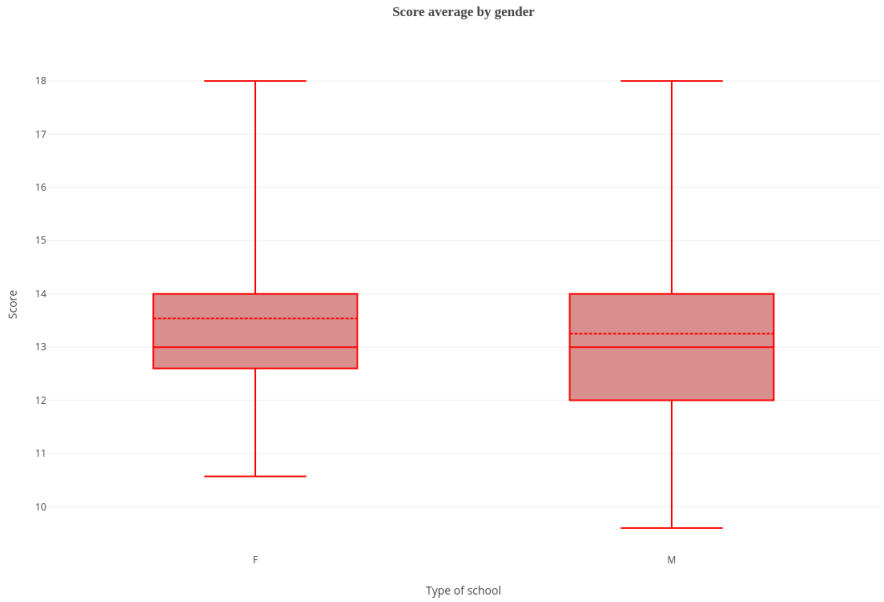



Figure 4.21: Relation between the score and gender.

In figure 4.21, we can see that females have on average marginally higher grades than males. This seems to be explained by the higher occurrence of male students with a score of 12 to 13 than females. There is no noticeable difference in the median for males and females. The lowest score values are also higher for females than for males.

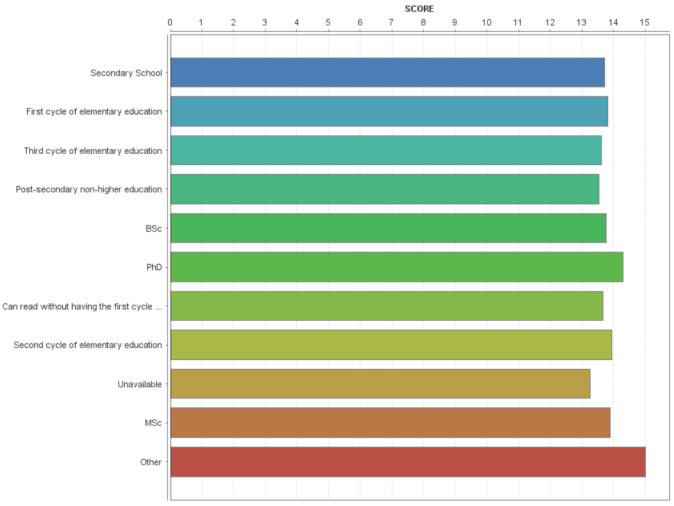



Figure 4.22: Relation between the score and the father's education level.

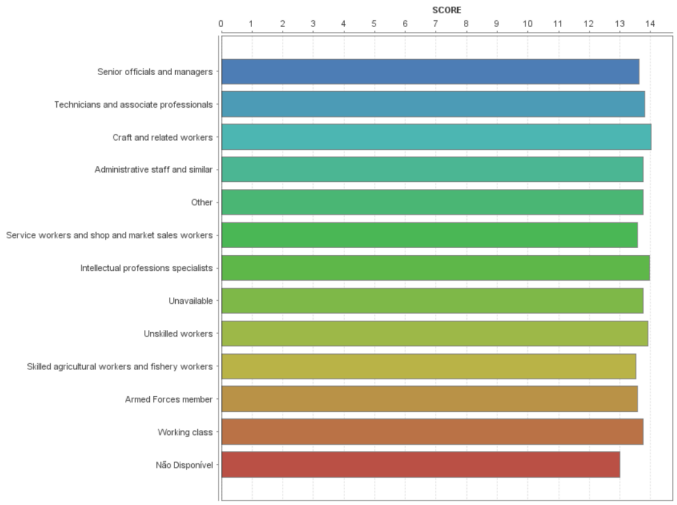



Figure 4.23: Relation between the score and the father's job.

In figure 4.22 we can see that there is a correlation between having a father that went far in education and having good grades. The highest grades come from students whose father has a BSc, MSc, PhD or concluded high school. It should be noted that despite "Other" having a high value of 15 for the score, there is only one occurrence, which means that it cannot be considered for statistical analysis. We can also see in figure 4.23 that having a father that works in intellectual jobs also tends to lead to higher grades.

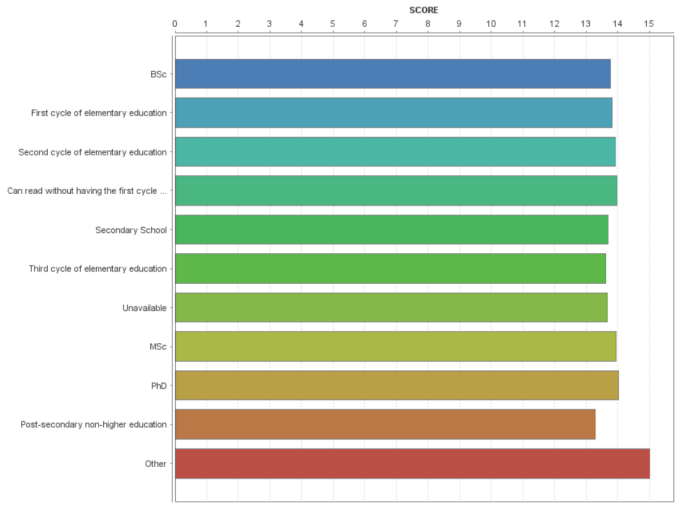



Figure 4.24: Relation between the score and the mother's education level.

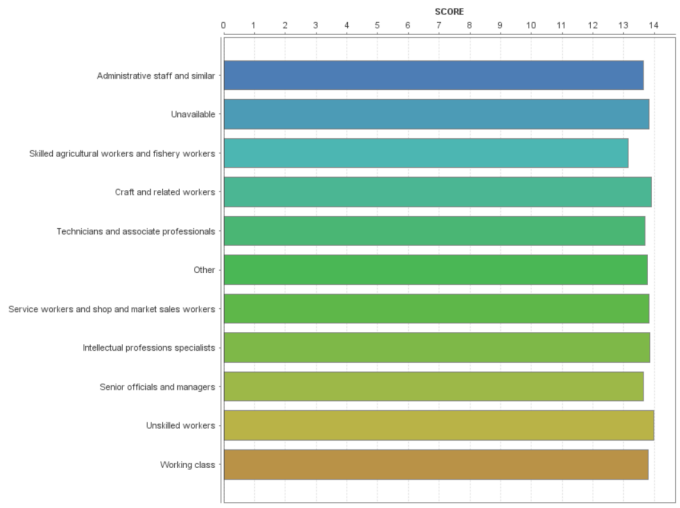



Figure 4.25: Relation between the score and the mother's job.

The correlation between the father's level and education and the student's score can also be seen for the mother, albeit much weaker. Although we can see that having a mother with a MSc or a PhD can lead to higher grades, the same doesn't apply to mothers with a BSc or high school, which weakens the correlation. No patterns can be seen between the mother's job and academic performance.

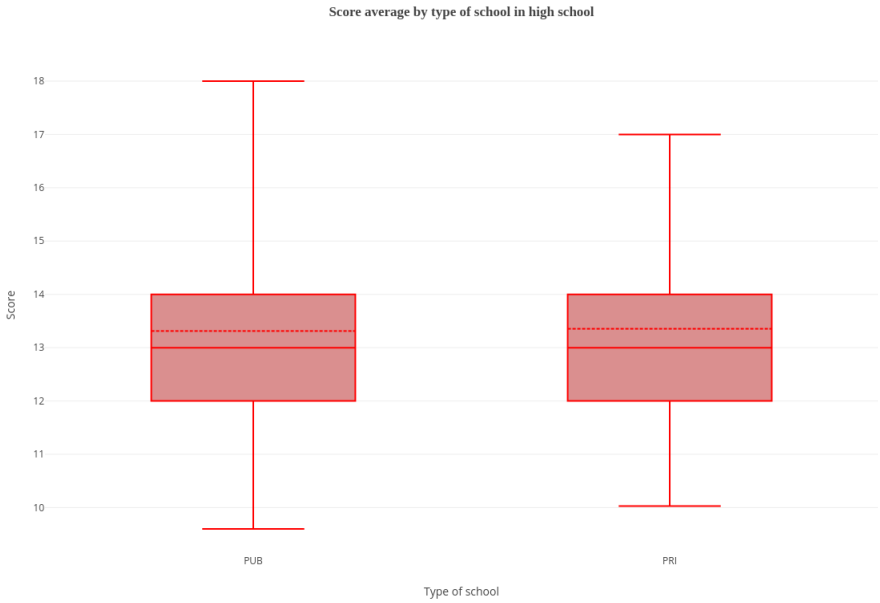



Figure 4.26: Relation between the score and the type of school the student came from.

According to figure 4.26, the results from students that came from public schools are very similar to the ones that came from private schools. However, the grades from students that come from public schools seem to have higher dispersion.

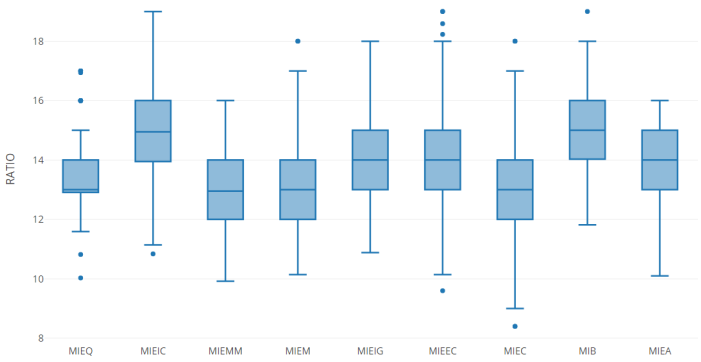



Figure 4.27: Score averages for each degree.

In figure 4.27, we can see the score averages from each degree. No real conclusions can be drawn from this information, though.

In figure 4.28, we can see that having a scholarship does not overly influence a student's performance. However, it should be noted that students that have been granted one tend to not have scores between 12 and 13, despite having very similar average median and average when compared to students that haven't been granted one. If we look at the dispersion of the data, we can see that the floor of the students that enrolled in the first phase is smaller than the one for the students that enrolled in the second one.

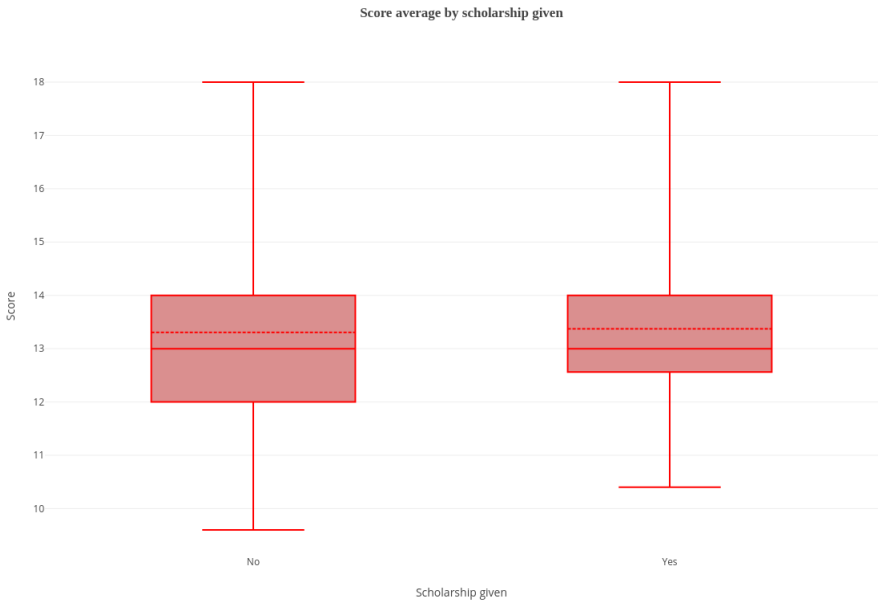



Figure 4.28: Relation between the score and whether the student was given a scholarship or not.

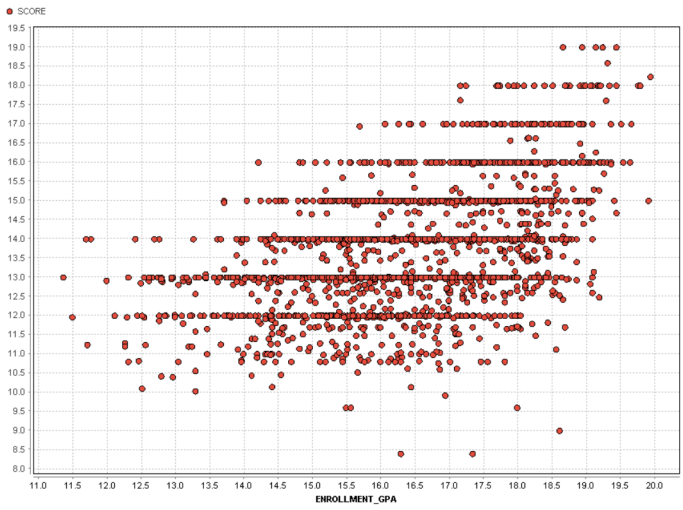



Figure 4.29: Relation between the score and the student's enrollment GPA.

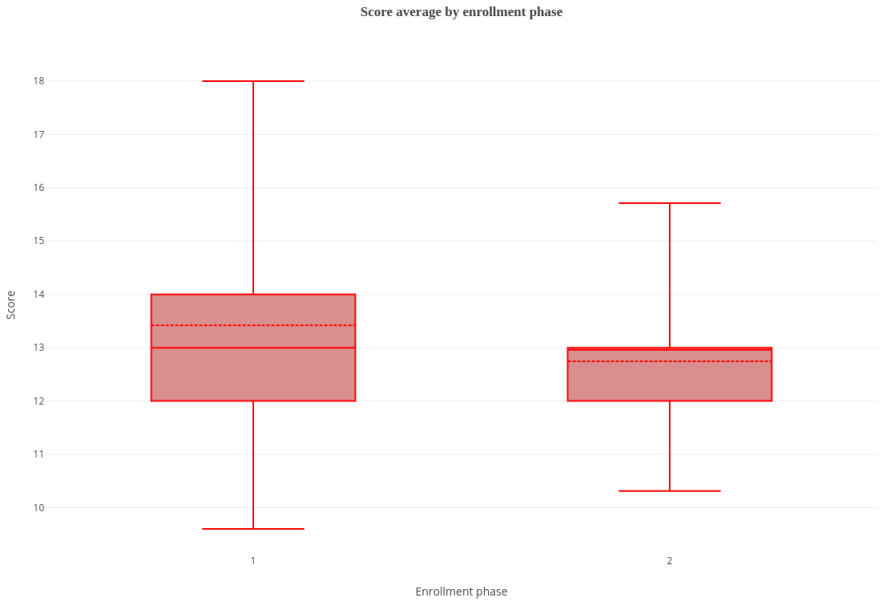



Figure 4.30: Relation between the score and the phase the student enrolled in.

In figure 4.29, we can see that there is a strong correlation between the GPA obtained to enroll in a degree and the performance of the student in an academic context. There also seems to be a clear difference of performance from students that enrolled in the degree in the first enrollment phase when compared to students that did so in the second phase, according to figure 4.30.

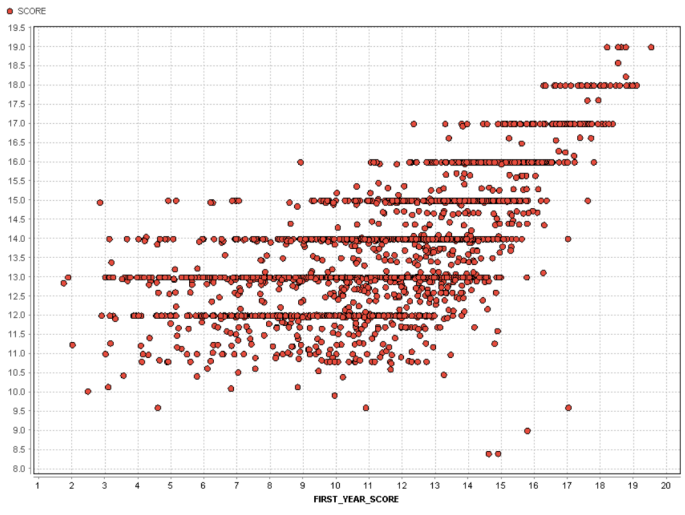



Figure 4.31: Relation between the score and the score of the student in the end of the first semester.

In figure 4.31, we can see that there is a very strong correlation between the performance of the student in the first year and their final performance.

Despite the clear difference in number of accesses to SIGARRA in the week compared to the weekend noted in section 4.1, there seems to be no correlation between the number of accesses per day of the week to SIGARRA, both in weekdays and in the weekend.

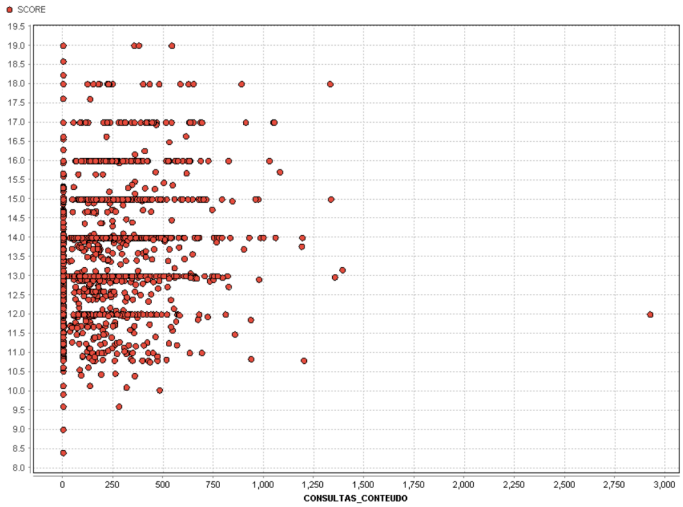



Figure 4.32: Comparison between accessing course content and score.

In figure 4.32, we can see that there does not seem to exist any correlation between accessing course content and the final score. The same seems to happen for course and profile pages, as can be seen in figures 4.33 and 4.34.

In figure 4.35 we can see that there does not seem to be any correlation between the number of sessions of a student when accessing SIGARRA.

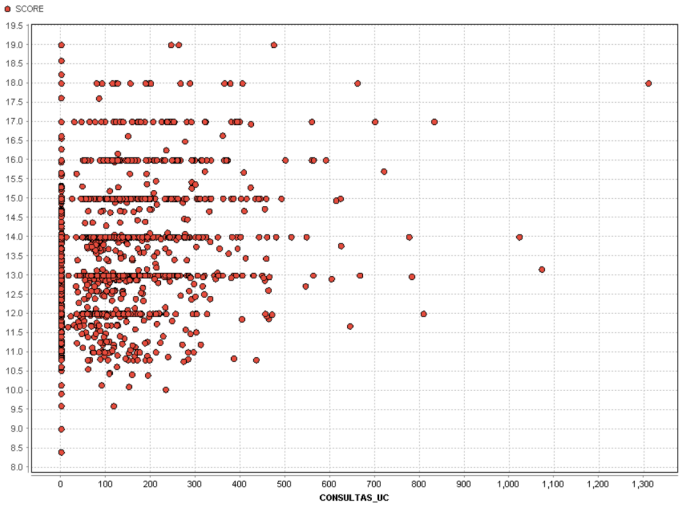



Figure 4.33: Comparison between accessing course pages and score.

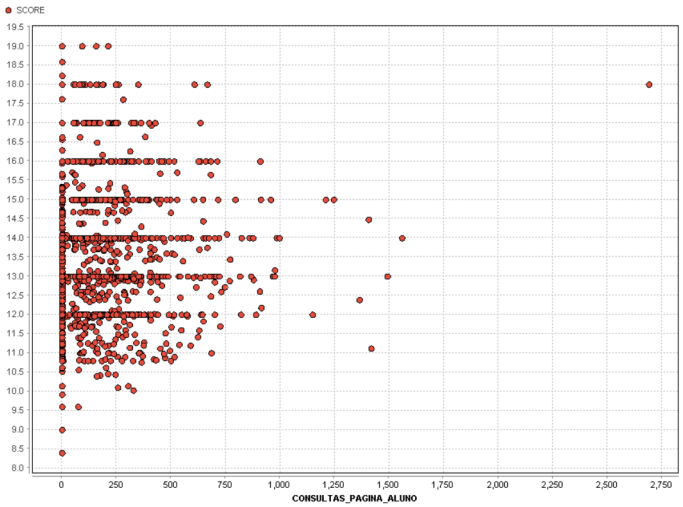



Figure 4.34: Comparison between accessing profile pages and score.

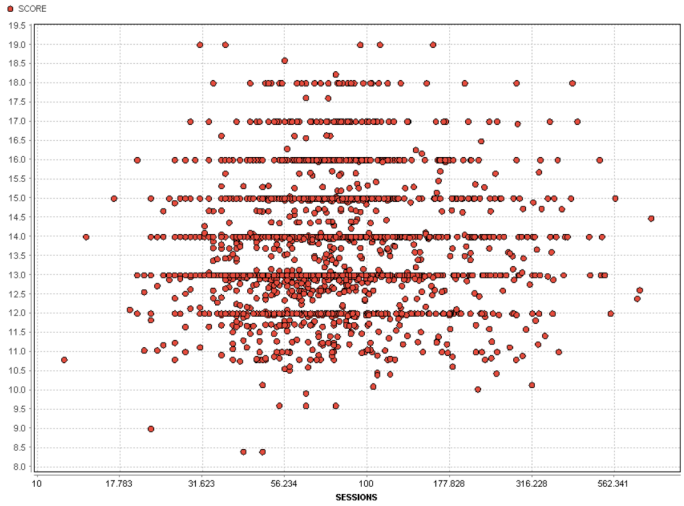



Figure 4.35: Comparison between number of sessions in SIGARRA and score.

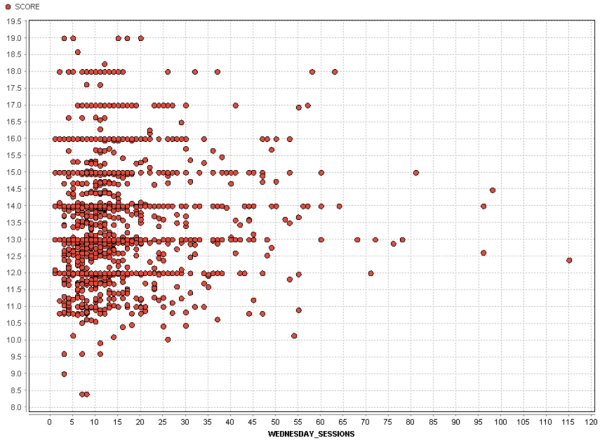



Figure 4.36: Comparison between number of sessions in SIGARRA on Wednesday and score.

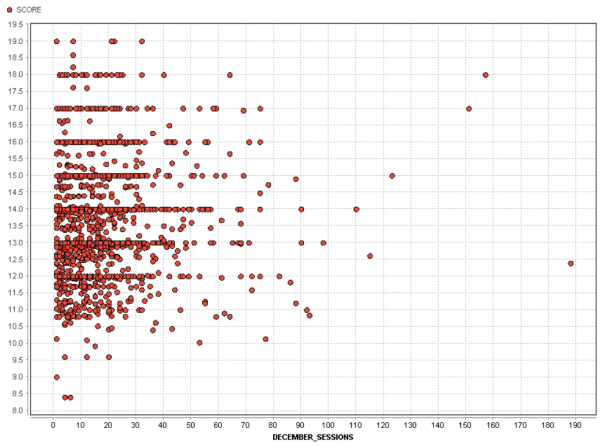



Figure 4.37: Comparison between number of sessions in SIGARRA in December and score.

The same can also be said about the number sessions per weekday and month. Figures 4.36 and 4.37 show the number of sessions of each user on Wednesday and in December as example.

From this analysis we can conclude that the enrollment grade and the first year grade are the attributes most likely to have a key role in the model. However, other attributes such as the parent's education are also likely to have some preponderance.

# Chapter 5

# Results

Before delving into the results obtained, it is important to outline the parameters used for the data mining methods used to obtain them. The process of finding a set of optimal parameters for a learning algorithm is called **hyperparameter optimization**. The method chosen to accomplish this process was **grid search**. Grid search consists of an exhaustive search through a specified array of values for each parameter being optimized. Each possible combination between the parameters is given as input to the algorithm. Once every possible combination was tested, the one that led to better results is chosen.

In what concerns the Random Forests, for both models, 1000 trees were used. The only parameter that was subject to grid search was the amount of features to consider when building the trees, being set to 7 for the first model and 17 for the second one.

For SVM, a linear kernel was used. Two parameters were optimized: the C (cost) parameter, which determines the smoothness of a decision surface, was set to 1000. A high C value aims at classifying a high amount of training examples correctly, at the cost of being more likely to fall into overfitting, whereas a smaller C is more loose, letting some training examples not be classified correctly but, in return, being less prone to overfitting [PVG+11]. The cost was set to 0.125 for the first model and 512 for the second model. The other parameter optimized with grid search was $\gamma$ (gamma). This parameter is a value that determines the influence that a single parameter has, with a small output value meaning a high influence. The values $3.05e^{(}-5)$ and 8 were set for the first and second model, respectively.

In what concerns Neural Networks, 1 hidden layer was used. Two parameters were optimized: the number of nodes in the hidden layer, which were set as 50 nodes for the first model and 60 nodes for the second one, and the weight decay, a parameter that determines how much the weight of new iterations matters for the weight of the nodes as more iterations are completed. 1 was set as the weight decay for both models. The maximum number of iterations for the algorithm was also defined as 100 for both models.

As mentioned in section 3, the first model will only make use of sociodemographic and academic information. The attributes that were considered by the model were:

- Gender

- Age
- Father's education level
- Father's job
- Mother's education level
- Mother's job
- School type
- Degree
- Scholarship requested
- Scholarship given
- GPA in the $12^{th}$ grade
- Enrollment GPA
- The application stage the student was accepted into the degree
- The position of the degree student's application form.
- The score of the student in the first semester.

After the preprocessing and transformation steps described in section 3, 2023 instances remained. Random Forests and Support Vector Machines make use of the full list for both the first and second model. However, for Neural Networks, further feature selection beyond the one described in chapter 3 was done. The procedure used to determine which features were included and which features were not included involved extracting the importance value determined by R of each feature in each of the 10 iterations of the k-fold cross validation algorithm for the Random Forests and then averaging them. If the average was less than 20, the feature was excluded. The results are shown in the following table:

Table 5.1: Feature selection results for the first model

| **Attribute** | **Average** | **Selected for the model** |
|---|---|---|
| The score of the student in the first semester | 969.22 | Yes |
| Enrollment GPA | 230.86 | Yes |
| Degree | 228.92 | Yes |
| GPA in the $12^{th}$ grade | 181.50 | Yes |
| Father's job | 65.52 | Yes |
| Mother's job | 52.21 | Yes |
| Mother's education level | 48.45 | Yes |

Continued on next page

**Table 5.1 – continued from previous page**

| Attribute | Average | Selected for the model |
|---|---|---|
| Father's education level | 37.84 | Yes |
| The position of the degree student's application form | 9.887 | No |
| Age | 4.12 | No |
| Scholarship requested | 2.31 | No |
| Gender | 2.21 | No |
| Type of school | 2.06 | No |
| Scholarship given | 2.03 | No |
| The application stage the student was accepted into the degree | 0.38 | No |

From the table, one can see that the variables that more strongly correlate with the dependent variable are the score of the student in the first semester and the enrollment GPA. This seems to be in line with the conclusions drawn in section 4, where it was concluded that both these variables correlated strongly with the score.

The performance of the three algorithms for this model is shown in the following table:

Table 5.2: Performance of the first model

| Algorithm | $R^2$ | MSE |
|---|---|---|
| Random Forest | 0.826 | 0.84 |
| Neural Network | 0.791 | 0.98 |
| Support Vector Machine | 0.768 | 1.07 |

From table 5.2 we can see that the model proposed seems to have potential in determining the potential success of the students. Furthermore, the Random Forest algorithm produces the best results, getting a better $R^2$ and MSE than Neural Networks and Support Vector Machines, who come off as second and third, respectively.

For the second model the variables related to SIGARRA's log files were added. These include the following:

- Number of sessions on Monday
- Number of sessions on Tuesday
- Number of sessions on Wednesday
- Number of sessions on Thursday
- Number of sessions on Friday
- Number of sessions on Saturday
- Number of sessions on Sunday

- Number of sessions in September
- Number of sessions in October
- Number of sessions in November
- Number of sessions in December
- Number of course content requests
- Number of student profile page requests
- Number of course page requests
- Days until the student's first access to SIGARRA since the semester started

In addition to these, each week day and month has a variable associated with the percentage of sessions had in said week day or month.

Much like for the first model, the variables were selected according to their importance. The values used were the same ones for the first model, as well as the values that correspond to the new variables. The results for the features that are exclusive to the second model are shown in the following table:

Table 5.3: Feature selection results for the second model

| **Attribute** | **Average** | **Selected for the model** |
|---|---|---|
| Percentage of sessions in October | 27.83 | Yes |
| Percentage of sessions on Tuesday | 27.29 | Yes |
| Percentage of sessions on Sunday | 24.05 | Yes |
| Percentage of sessions on Wednesday | 23.03 | Yes |
| Percentage of sessions in November | 22.09 | Yes |
| Percentage of sessions in December | 21.67 | Yes |
| Percentage of sessions on Thursday | 21.65 | Yes |
| Percentage of sessions on Friday | 21.56 | Yes |
| Percentage of sessions on Monday | 21.17 | Yes |
| Percentage of sessions on Saturday | 21.16 | Yes |
| Percentage of sessions in September | 20.65 | Yes |
| Percentage of sessions in January | 20.44 | Yes |
| Course content requests | 15.42 | No |
| Number of sessions in October | 15.20 | No |
| Number of sessions in September | 14.50 | No |

Continued on next page

**Table 5.3 – continued from previous page**

| Attribute | Average | Selected for the model |
|---|---|---|
| Number of sessions in November | 13.33 | No |
| Number of course content requests | 12.93 | No |
| Number of sessions in December | 12.87 | No |
| Number of sessions in January | 12.43 | No |
| Number of sessions on Thursday | 12.35 | No |
| Number of sessions on Sunday | 12.05 | No |
| Number of sessions | 11.80 | No |
| Number of sessions on Wednesday | 11.36 | No |
| Number of sessions on Friday | 11.22 | No |
| Number of sessions on Saturday | 11.09 | No |
| Number of sessions on Monday | 9.88 | No |
| Number of sessions on Tuesday | 9.69 | No |
| Number of student profile page requests | 9.08 | No |
| Days until first SIGARRA access | 8.89 | No |

In general, these variables are less important than the variables already included in the first model. It can be seen that the conclusions drawn in chapter 4 about the non-existance of a correlation between variables such as the number of sessions or the course content requests are corroborated by the results displayed on the table. However, the small correlation that existed between the number of sessions in the months - particularly in December - does not seem to hold true. The information about the percentage of sessions seems to be more valuable for the model instead.

The dataset for this model went through the same preprocessing and transformation steps used for the first model. The performance is shown in the following table:

Table 5.4: Performance of the second model

| Algorithm | $R^2$ | MSE |
|---|---|---|
| Random Forest | 0.828 | 0.84 |
| Neural Network | 0.778 | 1.04 |
| Support Vector Machine | 0.755 | 1.14 |

From table 5.4 we can see that the Random Forest algorithm produces the best results, getting a better $R^2$ and MSE than Neural Networks and Support Vector Machines, who come off as second and third, respectively.

Comparing the results between table 5.2 and table 5.4, one can see that only for Random Forests has the browsing history information been useful, and only marginally - in fact, the results for the SVMs and Neural Networks, the results are slightly inferior. This means that the inclusion of the logs from SIGARRA does not seem to benefit the predictive models, which means that there is not a usage pattern that influences the academic performance of students.

# Chapter 6

# Conclusions and Future Work

## 6.1 Conclusions

In order to better understand as early as possible which students are likely to be top-performing and which ones are more likely to face difficulties throughout their academic career, as well as determine the impact that Information System usage might have in their performance, two predictive models were developed - one that does not make use of the aforementioned information about Information System used and one that does. These regression models try to predict the score of the student in the end of their degree - defined as the multiplication between their GPA and the completed/enrolled ECTS credits ratio.

The second model seems to perform marginally better than the first model when using Random Forests and slightly worse when using Neural Networks and Support Vector Machines, unlike what was expected from the literature review. Therefore, there are not evidences that the variables referring to the access to SIGARRA enable to have insights on the academic performance of students. This seems to reveal that the interaction of students with this LMS is not different according to the academic performance of the users. However, we can hypothesize that this may be due to the small variety in the SIGARRA logs provided. With more information about the categories of pages students access and when in the day they access them, more variables could be created, which could increase the potential of the information extracted from the use of SIGARRA.

It should also be noted that the Random Forest algorithm has consistently proven to be the best one performance-wise. This algorithm has been getting very good results in several studies similar to this one, and these results further prove that Random Forests are among the best algorithms available when looking for a solution in EDM and should never be disregarded by anyone who wants to tackle an academic performance problem.

The project will culminate in the submission of a paper which is being concluded. The paper is expected to be submitted in July 2018. The current state of the paper can be found in appendix B.

## 6.2 Future work

The future work that can be made for this project lies mostly in two key aspects: one of them is the improvement of the models developed. More data in what concerns the students' use of SIGARRA would definitely be the best way of improving the performance of the models, since that means that move variables can be generated, thus increasing the pool of variables the model can work with. Moreover, having data from more years to work with may also improve model performance, since that means that more data can be used for training and testing the model. Another enhancement that could be done is resorting to feature selection algorithms to do the aforementioned preprocessing step. The selection of features with higher predictive potential could potentially lead to better results. The model also takes in consideration solely students that enrolled through the regular contingent, which means that students from other countries or from Madeira and Azores are not considered.

The other aspect where the project can be improved on is how the information is presented. The creation of a application with an easy to use interface and with results displayed in an easy way to understand could be very useful for the end-user of the knowledge extracted from the dataset. Although this last aspect is out of scope of this dissertation, it is undeniably true that presenting information in a graphical, easy to use and understand manner can improve information absorption and make interpreting the results much more feasible for less tech-savvy people.

# References


[AAK+16] Ralph Olusola Aluko, Olumide Afolarin Adenuga, Patricia Omega Kukoyi, Aliu Adebayo Soyingbe, and Joseph Oyewale Oyedeji. Predicting the academic success of architecture students by pre-enrolment requirement: using machine-learning techniques. *Construction Economics and Building*, 16(4):86, dec 2016.

[ABM13] Pauziah Mohd Arsad, Norlida Buniyamin, and Jamalul-lail Ab Manan. A neural network students' performance prediction model (NNSPPM). In *2013 IEEE International Conference on Smart Instrumentation, Measurement and Applications (ICSIMA)*, pages 1–5. IEEE, nov 2013.

[AEH12] Mohammed M Abu Tair and Alaa M El-Halees. International Journal of Information and Communication Technology Research Mining Educational Data to Improve Students' Performance: A Case Study. 2(2), 2012.

[BKA+] Ryan S J D Baker, Jessica Kalka, Vincent Aleven, Lisa Rossi, Sujith M Gowda, Angela Z Wagner, Gail W Kusbit, Michael Wixon, Aatish Salvi, and Jaclyn Ocumpaugh. Towards Sensor-Free Affect Detection in Cognitive Tutor Algebra.

[BSAD13] Angela Bovo, Stéphane Sanchez, Olivier Héguy Andil, and Yves Duthen. Analysis of students clustering results based on Moodle log data. 2013.

[BY09] Ryan S.J.d. Baker and Kalina Yacef. *Journal of Educational Data Mining JEDM.*, volume 1. International Educational Data Mining Soc, oct 2009.

[Byd] Hana Bydžovská. A Comparative Analysis of Techniques for Predicting Student Performance.

[CAP+16] Rebeca Cerezo, Miguel S Anchez-Santill An, M Puerto, Paule Ruiz, and J Carlos Nú. Students' LMS interaction patterns and their relationship with achievement: A case study in higher education. 2016.

[CLE14] HANNAH CLEVELAND. The Positive Effects of Education, 2014.

[Com15] European Comission. Ects users' guide, 2015.

[FPSS96a] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From Data Mining to Knowledge Discovery in Databases. 1996.

[FPSS96b] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From Data Mining to Knowledge Discovery in Databases. 1996.

[GIK10a] Huseyin Guruler, Ayhan Istanbullu, and Mehmet Karahasan. A new student performance analysing system using knowledge discovery in higher educational databases. *Computers & Education*, 55(1):247–254, aug 2010.

[GIK10b] Huseyin Guruler, Ayhan Istanbullu, and Mehmet Karahasan. A new student performance analysing system using knowledge discovery in higher educational databases. *Computers & Education*, 55:247–254, 2010.

[GMHdO12] Elena Gaudioso, Miguel Montero, and Felix Hernandez-del Olmo. Supporting teachers in adaptive educational systems through predictive models: A proof of concept. *Expert Systems with Applications*, 39(1):621–625, jan 2012.

[He13] Wu He. Examining students' online interaction in a live video streaming environment using data mining and text mining. *Computers in Human Behavior*, 29(1):90–102, jan 2013.

[HF13] Shaobo Huang and Ning Fang. Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models. *Computers & Education*, 61:133–145, 2013.

[HS17] Anne-Sophie Hoffait and Michaël Schyns. Early detection of university students with potential difficulties. *Decision Support Systems*, 101:1–11, sep 2017.

[JVMM12] Milos Jovanovic, Milan Vukicevic, Milos Milovanovic, and Miroslav Minovic. Using data mining on student behavior and cognitive style data for improving e-learning systems: a case study. *International Journal of Computational Intelligence Systems*, 5(3):597–610, jun 2012.

[LLM+14] Juan A Lara, David Lizcano, María A Martínez, Juan Pazos, and Teresa Riera. A system for knowledge discovery in e-learning environments within the European Higher Education Area - Application to student data from Open University of Madrid, UDIMA. 2014.

[LRSM15] Marcia Laugerman, Diane T Rover, Mack C Shelley, and Steven K Mickelson. Determining Graduation Rates in Engineering for Community College Transfer Students Using Data Mining. 2015.

[MB12] Judi Mccuaig and Julia Baldwin. Identifying Successful Learners from Interaction Behaviour. 2012.

[MD09] Leah P Macfadyen and Shane Dawson. Mining LMS data to develop an "early warning system " for educators: A proof of concept. *Computers & Education*, 54:588–599, 2009.

[MD10] Leah P. Macfadyen and Shane Dawson. Mining LMS data to develop an "early warning system" for educators: A proof of concept. *Computers & Education*, 54(2):588–599, feb 2010.

[MDDM16] Farshid Marbouti, Heidi A Diefes-Dux, and Krishna Madhavan. Models for early prediction of at-risk students in a course using standards-based grading. 2016.

[MKG14] Tripti Mishra, Dharminder Kumar, and Sangeeta Gupta. Mining Students' Data for Prediction Performance. In *2014 Fourth International Conference on Advanced Computing & Communication Technologies*, pages 255–262. IEEE, feb 2014.

[MRT12] Mehryar. Mohri, Afshin. Rostamizadeh, and Ameet. Talwalkar. *Foundations of machine learning*. MIT Press, 2012.

[MRV11] Carlos Márquez, Cristóbal Romero, and Sebastian Ventura. Predicting school failure using data mining. pages 271–276, 01 2011.

[NMSP16] Denise Nacu, Caitlin K. Martin, Michael Schutzenhofer, and Nicole Pinkard. Beyond Traditional Metrics. In *Proceedings of the Third (2016) ACM Conference on Learning @ Scale - L@S '16*, pages 197–200, New York, New York, USA, apr 2016. ACM Press.

[NZ14] Srečko Natek and Moti Zwilling. Student data mining solution–knowledge management system related to higher education institutions. 2014.

[PA14a] Alejandro Peña-Ayala. Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Systems With Applications*, 41:1432–1462, 2014.

[PA14b] Alejandro Peña-Ayala. Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Systems With Applications*, 41:1432–1462, 2014.

[Pal13] Stuart Palmer. Modelling Engineering Student Academic Performance Using Academic Analytics*. *International journal of engineering education*, 29(1):132–138, 2013.

[PNSK06] Tan Pang-Ning, Michael Steinbach, and Vipin Kumar. *Introduction to data mining*. 2006.

[Poo14] John Poole. Why Education Is The Most Important Revolution Of Our Time, 2014.

[PVG+11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[REZ+13] Cristobal Romero, Pedro G. Espejo, Amelia Zafra, Jose Raul Romero, and Sebastian Ventura. Web usage mining for predicting final marks of students that use Moodle courses. *Computer Applications in Engineering Education*, 21(1):135–146, mar 2013.

[RV07] C. Romero and S. Ventura. Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, 33(1):135–146, jul 2007.

[RV10] Cristóbal Romero and Sebastián Ventura. Educational Data Mining: A Review of the State of the Art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(6):601–618, nov 2010.

[Saa16] Amjad Abu Saa. Educational Data Mining & Students' Performance Prediction. *IJACSA) International Journal of Advanced Computer Science and Applications*, 7(5), 2016.

[SCS+15] Pedro Strecht, Luís Cruz, Carlos Soares, João Mendes-Moreira, and Rui Abreu. A Comparative Study of Classification and Regression Algorithms for Modelling Students' Academic Performance. 2015.

[SJ10] So Young Sohn and Yong Han Ju. Conjoint analysis for recruiting high quality students for college education. *Expert Systems with Applications*, 37(5):3777–3783, may 2010.

REFERENCES


[STU+13] S Saranya, N Tamilselvi, P Usha, M Yasodha, and V Padmapriya. Data Mining Techniques in EDM for Predicting the Pupil " s Outcome. *International Journal of Advanced Research in Computer and Communication Engineering*, 2(6), 2013.

[TKD+13] Setsuo Tsuruta, Rainer Knauf, Shinichi Dohi, Takashi Kawabe, and Yoshitaka Sakurai. An Intelligent System for Modeling and Supporting Academic Educational Processes. pages 469–496. Springer, Berlin, Heidelberg, 2013.

[TSK05] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to data mining*. Pearson, 1st edition, 2005.

[VMS07] J.-P. Vandamme, N. Meskens, and J.-F. Superby. Predicting Academic Performance by Data Mining Methods. *Education Economics*, 15(4):405–419, dec 2007.

[WW14] Robert Whannell and Patricia Whannell. Identifying tertiary bridging students at risk of failure in the first semester of undergraduate study. *Australian Journal of Adult Learning*, 54(2), 2014.

[D12] Baha Şen, Emine Uçar, and Dursun Delen. Predicting and analyzing secondary education placement-test scores: A data mining approach. *Expert Systems with Applications*, 39(10):9468–9476, aug 2012.

# Appendix A

# Dataset Variables

Table A.1: Dataset variables.

| Attribute | Type | Description | Values(frequency)/Mean(std. deviation |
|---|---|---|---|
| 1 | Numerical | Age of the student | 17.634 (0.759) |
| 2 | Categorical | Gender of the student | Male:71.9%<br>Female:28.1% |
| 3 | Categorical | Father's education level | Can read but doesn't have the First Cycle of Elementary Education: 0.5%<br>First Cycle of Elementary Education: 12.2%<br>Second Cycle of Elementary Education: 8.6%<br>Third Cycle of Elementary Education: 13.1%<br>Postsecondary Non-higher Education:1.5%<br>Secondary Education: 20.8%<br>Technological Specialization (higher education): 3.7%<br>BSc: 32.2%<br>MSc: 4.9%<br>PhD: 3.9%<br>Not available: 2.3% |
| 4 | Categorical | Father's job | Senior officials and managers: 15.5%<br>Unavailable: 8.7%<br>Technicians and associate professionals : 11.8%<br>Intellectual professions specialists: 10.7%<br>Service workers and shop and market sales workers: 8.6%<br>Craft and related workers: 6.9%<br>Administrative staff and similar : 8.0%<br>Unskilled workers: 3.0%<br>Armed Forces member: 1.2%<br>Working class: 2.6%<br>Skilled agricultural workers and fishery workers: 1.3% |
| 5 | Categorical | Mother's education level | Undergraduate level: 26.6%<br>Unknown:18.8%<br>3rd cycle of elementary school:12.2%<br>Can read but doesn't have the First Cycle of Elementary Education: 2.2%<br>First Cycle of Elementary Education: 8.7%<br>Second Cycle of Elementary Education: 8.4%<br>Third Cycle of Elementary Education: 12.8%<br>Postsecondary Non-higher Education:1.2%<br>Secondary Education: 20.0% |

Continued on next page

**Table A.1 – continued from previous page**

| Attr. | Type | Description | Values:frequency/Mean(std. deviation) |
|---|---|---|---|
| | | | BSc: 38.4% |
| | | | MSc: 5.1% |
| | | | PhD: 2.1% |
| | | | Not available: 1.2% |
| 6 | Categorical | Mother's job | Other: 23.5%<br>Unavailable: 17.6%<br>Intellectual professions specialists: 14.3%<br>Administrative staff and similar : 10.8%<br>Senior officials and managers: 9.8%<br>Technicians and associate professionals : 8.1%<br>Service workers and shop and market sales workers: 7.6%<br>Craft and related workers: 4.5%<br>Unskilled workers: 3.0%<br>Skilled agricultural workers and fishery workers: 0.5%<br>Working class: 0.2% |
| 7 | Categorical | Type of school (public/private school) | Public: 79.6%<br>Private: 20.4% |
| 8 | Categorical | Degree | MIEC: 21.7%<br>MIEEC: 20.6%<br>MIEIC: 14.4%<br>MIEM: 13.7%<br>MIEIG: 8.4%<br>MIEQ: 7.7%<br>MIB: 6.9%<br>MIEA: 4.4%<br>MIEMM: 2.1% |
| 9 | Categorical | The student asked for a scholarship | Yes:32.1%<br>No:67.1% |
| 10 | Categorical | The student was granted a scholarship. | Yes:21.2%<br>No:78.8% |
| 11 | Numerical | The student's GPA in the $12^th$ grade. | 16.514 (1.477) |
| 12 | Numerical | The student's grade of application to the degree (GPA + National Exams) | 16.415 (1.519) |
| 13 | Categorical | The application phase the student was accepted into the degree. | 1: 98.6%<br>2: 1.4% |
| 14 | Categorical | Where the degree was in the student's preference in the application list. | 1: 76.5%<br>2: 14.5%<br>3: 4.7%<br>4: 2.3%<br>5: 1.5%<br>6: 0.4% |
| 15 | Numerical | The score of the student in the end of the first semester | 11.707 (3.229) |
| 16* | Numerical | The number of SIGARRA sessions of the student on Monday. | 13.965 (11.590) |
| 17* | Numerical | The percentage of SIGARRA sessions of the student on Monday. | 0.150 (0.041) |
| 18* | Numerical | The number of SIGARRA sessions of the student on Tuesday. | 12.980 (11.641) |
| 19* | Numerical | The percentage of SIGARRA sessions of the student on Tuesday. | 0.137 (0.039) |

Continued on next page

**Table A.1 – continued from previous page**

| Attr. | Type | Description | Values:frequency/Mean(std. deviation) |
|---|---|---|---|
| 20* | Numerical | The number of SIGARRA sessions of the student on Wednesday. | 13.265 (10.705) |
| 21* | Numerical | The percentage of SIGARRA sessions of the student on Wednesday. | 0.144 (0.036) |
| 22* | Numerical | The number of SIGARRA sessions of the student on Thursday. | 14.003 (11.838) |
| 23* | Numerical | The percentage of SIGARRA sessions of the student on Thursday. | 0.150 (0.042) |
| 24* | Numerical | The number of SIGARRA sessions of the student on Friday. | 12.910 (11.125) |
| 25* | Numerical | The percentage of SIGARRA sessions of the student on Friday. | 0.138 (0.039) |
| 26* | Numerical | The number of SIGARRA sessions of the student on Saturday. | 12.084 (8.996) |
| 27* | Numerical | The percentage of SIGARRA sessions of the student on Saturday. | 0.136 (0.041) |
| 28* | Numerical | The number of SIGARRA sessions of the student on Sunday. | 13.034 (10.273) |
| 29* | Numerical | The percentage of SIGARRA sessions of the student on Sunday. | 0.145 (0.040) |
| 30* | Numerical | The number of SIGARRA sessions of the student in September. | 14.499 (12.408) |
| 31* | Numerical | The percentage of SIGARRA sessions of the student in September. | 0.174 (0.085) |
| 32* | Numerical | The number of SIGARRA sessions of the student in October. | 23.406 (21.193) |
| 33* | Numerical | The percentage of SIGARRA sessions of the student in October. | 0.249 (0.084) |
| 34* | Numerical | The number of SIGARRA sessions of the student in November. | 21.625 (17.220) |
| 35* | Numerical | The percentage of SIGARRA sessions of the student in November. | 0.236 (0.071) |
| 36* | Numerical | The number of SIGARRA sessions of the student in December. | 15.125 (15.411) |
| 37* | Numerical | The percentage of SIGARRA sessions of the student in December. | 0.151 (0.074) |
| 38* | Numerical | The number of SIGARRA sessions of the student in January. | 17.585 (16.521) |
| 39* | Numerical | The percentage of SIGARRA sessions of the student in January. | 0.151 (0.074) |
| 40* | Numerical | The number of times a student requested a resource from a course. | 158.013 (220.822) |
| 41* | Numerical | The number of times a student checked the student's page. | 130.837 (199.604) |
| 42* | Numerical | The number of times a student checked a course's page. | 90.625 (126.164) |
| 43* | Numerical | The number of sessions in SIGARRA the student had. | 92.241 (70.846) |

Continued on next page

**Table A.1 – continued from previous page**

| Attr. | Type | Description | Values:frequency/Mean(std. deviation) |
|---|---|---|---|
| 44* | Numerical | Days that passed until the first access to SIGARRA by the student, in comparison to the first access in the dataset for the whole set of students in the first semester of the first academic year | 1.770 (10.936) |
| 45 | Numerical | Dependent variable. Student's final degree GPA * Ratio between completed and enrolled courses. | 13.321 (1.650) |

# Appendix B

# Paper to be submitted

## B.1 Abstract

Education is one of the key aspects of human and economic growth. One having better education means that they are able to more competently tackle the challenges that arise in their lives. Thus, finding methods that cater to the needs of each student, whether they show signs of excellence in their academic results or have been struggling to achieve performance levels that meet the ones of their peers, is an endeavour of any university that wishes to provide their students top quality education. Universities generally hold really big databases with information about their students. This information is oftentimes ignored. However, it can be processed and analyzed, which means that conclusions about the student's performance can be drawn from said information. This task can be achieved with the use of data mining. With this in mind, this project's aim is to predict student success using this data, in their first academic year. This data includes sociodemographic information about the students, as well as information about their academic performance. The database also holds information on their browsing history in the university's Information System. In addition to this, the project also aims to understand how mcuh can this browsing history information be used to predict the accuracy of a predictive performance that does not include it. Regression will be used to approach this problem, more specifically, Neural Networks, Support Vector Machines and Random Forests. In order to validate the models proposed, the project will use the Faculty of Engineering of the University of Porto as case study.

## B.2 Introduction

Education plays a big part in our society's life and has several positive effects in our society. The current situation in what birth rate of Portugal and other European contries is concerned is not encouraging. This means that it's crucial that universities improve their overall image, in order to attract the fewer and fewer students that apply each year to Portuguese universities. Guaranteeing the best possible quality of education is, thus, a must. This can be achieved by finding teaching methods that cater to each student's specific needs, whether they are top-performing and need more

challenging tasks for boosting their qualities or if they experience more difficulties and require more individual, specialized support to help them overcome their difficulties and achieve their goals.

Furthermore, it's becoming more and more common that schools and universities have computers with several types of software, ranging from complete office suites to programming tools. Apart from this softwares, schools and universities are also investing into learning management systems (LMS), such as Moodle and CourseSites, to enhance the teacher and students' experience. These learning management systems aid in tasks such as project submissions and evaluation, allow teachers to make resources available to students and support forums for students to interact with each other and with their teachers.

The data gathered and generated by Learning Management Systems is oftentimes ignored by most institutions. However, these huge datasets hold enormous amounts of information that can (and should) be analyzed in order to improve the quality of education. If this information could be studied and processed, conclusions from said data can be drawn, which would then lead to better overall quality in the education system. The generated knowledge could also be helpful when changes to the education system are being considered.

Thus, the aim of this study is to apply data mining techniques, more specifically, supervised learning techniques, in order to develop a predictive model capable of predicting a student's overall academic performance on the early stages of their academic career, using sociodemographic data, evaluation data from high school and first academic year. Moreover, figuring out eventual relations between academic performance and the use students make of Learning Management Systems is another goal of this project.

The paper is structured as follows. The following section presents the related studies, in order to emphasize the contributions of the current study. Section 3 introduces the methods and data used, the variables included in the proposed model, and the performance evaluation criteria. Section 4 addresses the results and the discussion. Section 5 highlights the conclusions and section 6 the limitations and ideas for future research.

## B.3 Related Studies

The use of data mining in the context of education is not new. There has been an increasing number of attempts at using data available to universities for trying to better understand what influences student performance.

The progress done in this field has been surveyed several times already [RV07, RV10, PA14b]. These surveys are a thorough review to Educational Data Mining as a whole, in its multiple branches, hence providing several examples of studies that focus on predicting student performace and in the interaction between Educational Data Mining and Learning Management Systems.

The prediction of a student's performance is a challenging problem, due to the myriad of characteristics and circumstances that might influence it. Socio-demographic information, such as age and gender has been used extensively in these studies, as well as information about prior

studies, such as GPA in previous semesters, in high school or marks in previous assignments [HS17, ABM13, AAK+16, LRSM15], if the student is studying in part or full-time [NZ14], economic factors such as the existence of a scholarship, if the student borrowed money or the financial situation of the family/parents [GIK10b, SCS+15], degree of development in certain soft skills, such as leadership and decision making [MKG14], behaviour (such as presence in classes, doing homework) [VMS07, WW14], level of peer support [WW14], a student's own perception of himself (e.g. probability of succeeding, confidence degree) [VMS07, MB12] and extra-curricular activities [NZ14].

Some work has also been done with Learning Management Systems in Educational Data Mining. Information commonly extracted from LMS's includes number of LMS sessions, total session time [BSAD13, Pal13, MD09], date of first/last login to the LMS [BSAD13, Pal13], total number of individual LMS pages viewed [Pal13] and the number of actions taken [CAP+16]. Variables refering to LMS forum use have also been considered relevant. Examples of variables reltated to this issue include the total number of LMS discussion postings read and made [REZ+13, Pal13, MD09, JVMM12] and the number of words posted in said discussion postings [CAP+16]. The individual visualizations of each resource made available [LLM+14], the number of quizzes and assignments done [BSAD13, MB12, REZ+13, MD09], the grade obtained in graded activities [BSAD13] or if they passed or not [REZ+13, JVMM12], date and time taken to complete quizzes and exams [MB12, REZ+13, MD09, JVMM12, CAP+16], time in the discussion postings [JVMM12, CAP+16] and the number of days taken to turn in a task after it was assigned [CAP+16] have also been explored.

In order to analyze the data and extract knowledge from it, a plethora of data mining techniques can be used. Student performance problems are usually tackled with the use of regression and/or classification techniques. Classification consists of generating a function (or model) that maps (classifies) a data item into one of several predefined classes. The output is, thus, composed of categorical variables. Regression consists of learning a function that maps a data item to a real-valued prediction variable. The output being real variables is the main difference between regression and classification [FPSS96b]. Among the various regression and classification techniques, Support Vector Machines, Artificial Neural Networks, Random Forests, Decision Trees, K-Nearest Neighbour and Naive-Bayes seem to be the most widely used ones. Support Vector Machines are used by [SCS+15] in both a regression and classification problem in their comparative study of algorithms for modelling student academic performance. They used data of 5779 students from 391 programmes. Artificial Neural Networks are used by [HS17] to determine if students have a good chance to succeed in their first academic year, if they are likely to fail, or if their outcome is uncertain. The dataset is composed of 6845 first year students. Random Forests are used by [VMS07] with a group of 533 first-year university students to determine if a student has a low, medium or high risk of failing their first academic year. Decision Trees are used by [MB12] to attempt to predict the grade (A, B, C or D, F) of 122 first year students using information about their LMS interactions and a survey to the student's confidence in their skills. [AAK+16] use the K-nearest neighbour algorithm to determine the Cumulative Grade Points Average (CGPA) of

102 students using grades from previous exams. [MDDM16] uses the Naive-Bayes algorithm to determine a student's grade in a course according to the learning objectives that they have shown to have met from one written exam, 10 quizzes and five homeworks of 3063 students.

This paper differs from the works previously mentioned since it makes use of logs from a university's information system from several years to determine the use students make from them: some studies make us of LMS data, but generally they skip aspects such as in which time periods (morning/afternoon/evening/night) or when in the semester (early or late, meaning they start working early in the semester or they leave things for the last moment) do they use them. Although some studies make use of the student's part/full time situation, they don't make use of their student status (student/worker, athlete, etc).

Table A1 summarizes the studies mentioned above that focus on academic performance but don't directly use Learning Management system, whereas table A2 summarizes the ones that use LMS in their attempts at predicting student results.

## B.4 Methods and data

### B.4.1 Proposed method

This study aims to predict a student's academic performance in the early stages of their academic career, namely using anonymous browsing history information. For this purpose we will use as case study a portuguese institution, i.e. Faculty of Engineering of University of Porto. Therefore, the academic database of SIGARRA, FEUP's Information System, will be explored in order to understand how much can the browsing history data improve the accuracy of a predictive model focused in academic performance.

In light of this, we propose to construct two predictive regression models: one that only uses sociodemographic data and academic performance from the student's first year and another one that adds browsing data from SIGARRA as independent variables. In order to create the models, we are going to use three data mining algorithms that are capable of creating a regression model: Random Forests, Support Vector Machines and Artificial Neural Networks. The dependent variable, in both predictive models, is a score calculated using the following formula:

$$\frac{GPA * CompletedCredits}{EnrolledCredits}$$

Where *GPA* represents the Grade Point Average of the student by the end of the degree, *Completed Credits* represents the total amount of ECTS the student completed and Enrolled Credits the total amount of ECTS the student enrolled in. When a student does not get approval to a course, they have to enroll again in the same course until is able to conclude it with success. ECTS (acronym for European Credit Transfer and Accumulation System) is an European standard for comparing the "volume of learning based on the defined learning outcomes and their associated workload" for higher education across the European Union and other collaborating European

countries [Com15]. For successfully completed studies, ECTS credits are awarded. In Portugal, one academic year corresponds to 60 ECTS credits that are normally equivalent to around 1600 hours of total workload, irrespective of standard or qualification type.

Using this formula means that not only we consider the GPA of the student, but also how many credits (and consequently, courses) did they complete on the first try. Thus, we favour students who completed every course on the first try, penalizing students who took longer to complete their degree.

Following the literature, the set of independent variables includes factors such as high school background, family background and personal information. We also include information about the student's results in their first year, such as how many credits did they complete, how many credits did they enroll in and their GPA.

### B.4.2 Data

This study uses data from FEUP's information system, SIGARRA, from 2006 to 2011. Information from 8 programmes has been used, totaling 2023 students.

Sociodemographic information about the students includes information about their age, gender and marital status. In terms of socio-cultural status, we use information about the parents' education level and field of work. High school background is represented by the type of school they were in (public of private) and their GPA. Regarding the enrolment process, the application phase the students were accepted into their degree and the order of preference the degree was in, in the enrolment form. Information about their results during the first year in the degree consists of their GPA and the ratio between credits completed and enrolled.

This is the data used for the model that doesn't account for the browsing history data. To the other one, we make use of the number of times students requested course content, as well as page requests for the student's personal profile and course pages. To add to this, we also use information about each student's session. We define a session as "a sequence of page requests that start with a login into the system". In light of this, we also use the number of sessions a user was involved in, the number of sessions in each day of the week, the number of sessions in each month that is part of the first semester (from September to January). We also include the percentage of sessions that were had in each day of the week and each month in comparison with the total number of sessions. Lastly, we also include the number of days that passed until the first access to SIGARRA by the student in each academic year, in comparison to the first access in the dataset (by the whole set of students).

In the "Dataset Variables" subsection in the Appendices, a table with information about the several fields for the dataset for the second model can be found. Fields marked with an asterisk (*) are only used in the second model.

### B.4.3 Regression Methods

Three data mining techniques are used to create the predictive models: Random Forests, Support Vector machines and Artificial Neural Networks. In the next sections each of the algorithms used for this study is briefly described.

#### B.4.3.1 Random Forests

Random Forests are part of the class of algorithms called ensemble methods. Ensemble methods produce several predictive models and combine them into one final predictive model. This makes the final model less prone to overfitting and more stable (in the sense that we are much less prone to randomness that can lead into underperforming models). Random Forests employ the use of several decision trees where each tree is generated based on a subset of the original data set, each subset being independent from the others. This randomization helps reduce the correlation between the decision trees, which means the generalization error of the ensemble method can be improved.

#### B.4.3.2 Artificial Neural Networks

Artificial Neural Network is a data mining method inspired by attempts to simulate animal neural systems. They are composed of several input nodes and an output node (which are the equivalent to neurons) connected via weighted link (the equivalent to a synapse). The value output by the output node is determined by an activation function. In classification problems, a sign function is usually used. However, since we are dealing with a regression problem, an activation function such as a Linear, a Sigmoid or a Tahn function are used.

#### B.4.3.3 Support Vector Machines

Support Vector Machine is a method that when presented with a set of objects belonging to one of two possible values in an ambient space with n dimensions, builds a subspace with n-1 dimensions (a hyperplane) that separates the objects into one of the two categories. The hyperplane built is the one that has the largest possible margin between the area of the two categories, since this means that the model is less prone to generalization errors. Although this method is mostly adequate for classification problems with n attributes, it can be used in regression problems. Several methods have been proposed to tackle this, being one of the most popular reducing the problem into several classification subproblems.

### B.4.4 Evaluation criteria

Evaluation criteria is used to understand how accurate a predictive model is. This allows us to infer the ability of the predictive model to correctly respond to unseen data.

For this study, cross-fold validation with 10 folds is used. This means that the data is split into 9 training folds and 1 testing fold. The process is repeated 10 times, with each of the folds being

used once as the testing fold. This guarantees that each subset is used as test data exactly once. This approach has the advantage of using as much data as possible for training, while the test data also covers the entire data set.

We use as performance metrics the $R^2$ (also called coefficient of determination) and the Mean Squared Error. The coefficient of determination is a value that can be interpreted as how inferrable the dependent variable is from the independent variables. It ranges from 0 to 1. The Mean Absolute Error (MAE) measures the average of the difference between two variables, in this case, the value obtained by the predictive model for a test record that that test record's actual value for that feature. Derived from the Mean Absolute Error, the Mean Squared Error is, as the name implies, the Mean Absolute Error, squared. The key differences between them is that large errors have relatively greater influence on Mean Squared Error than they do the smaller error, which makes the MSE good for situations where big errors are very costly. The Mean Squared Error is, thus, given by:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Where $\hat{Y}$ is a vector of $n$ predictions, and $Y$ is the vector of observed values of the variable being predicted.

## B.5 Results and discussion

Before the final results were generated, hyperparameter optimization was done. This process consists of choosing a set of parameters from a range of possible values. The method used to accomplish this was grid search, which is an exhaustive search through a specified array of values for each parameter being optimized. Each possible combination between the parameters is given as input to the algorithm. Once every possible combination was tested, the one that led to better results is chosen.

In what concerns Random Forests, 1000 trees were used for both models. The number of features to consider when building the trees was subject to grid search. The best results were achieved with 7 variables for the model without the browsing history and 19 variables for the model with the browsing history.

For Neural Networks, the maximum number of iterations was defined as 100 for both models. Both models will use one hidden layer. Two parameters were subject to grid search: the weight decay and the size. The weight decay is a parameter that determines how much the weight of new iterations matters for the weight of the nodes as more iterations are completed, which was set to 1 for both models. The size, which is the number of nodes in the hidden layer, was set as 50 for the first model and 60 for the second one. The data was also subject to feature selection, with less performing variables being removed from the dataset given as input to the Neural Networks.

For Support Vector Machines, a linear kernel was used. Two parameters were optimized: the cost (C) parameter and the gamma ($\gamma$). The cost parameter determines the smoothness of a decision surface and was set to 0.125 for both models. A high C value aims at classifying a high amount

of training examples correctly, at the cost of being more likely to fall into overfitting, whereas a smaller C is more loose, letting some training examples not be classified correctly but, in return, being less prone to overfitting. The gamma is a value that determines the influence that a single parameter has, with a small output value meaning a high influence, and was set to 0.00012207 for both models.

The results obtained for the two models are shown in the following tables:

Table B.1: Performance of the first model

| **Algorithm** | **$R^2$** | **MSE** |
|---|---|---|
| Random Forest | 0.826 | 0.84 |
| Neural Network | 0.791 | 0.98 |
| Support Vector Machine | 0.768 | 1.07 |

Table B.2: Performance of the second model

| **Algorithm** | **$R^2$** | **MSE** |
|---|---|---|
| Random Forest | 0.828 | 0.84 |
| Neural Network | 0.778 | 1.04 |
| Support Vector Machine | 0.755 | 1.14 |

We can see that, in both models, the Random Forest algorithm produces the best results, getting a better $R^2$ and MSE than Neural Networks and Support Vector Machines, who come off as second and third, respectively.

## B.6 Conclusions

This study proposes the creation of two regression models, supported by data mining techiques, with the intent of predicting student overall performance. The goal is to understand the impact that browsing history from the university's Information Systems may have in their performance. This, one of the models uses only sociodemographic information, whereas the other uses both sociodemographic information and browsing history data. These regression models try to predict the score of the student in the end of their degree - defined as the multiplication between their GPA and the completed/enrolled ECTS credits ratio.

Random Forests seem to perform marginally better with the browsing history, whereas Neural Networks and Support Vector Machines seem to perform slightly worse, unlike what was expected from the literature review. This may be because of the small variance in the logs provided. With more information about the categories of pages students access and when in the day they access them, more variables could be created, which could potentially increase the difference of performance between the two models.

The Random Forest algorithm was the one that performed the best. This algorithm has been getting very good results in several studies related to Educational Data Mining, specially in assessing student performance and should not be ignored by anyone who wants to tackle an problem of this nature.

## B.7 Future work

Potential future work for this project can be divided into two categories: one being the improvement of model developed that makes use of browsing history and the other being improving the way information is displayed.

For improving the second model's performance, having more data related to the students browsing history would definitely be the best initial course of action. As previously mentioned, more information about the browsing history means that more variables can be created, which will most likely boost the performance of the model. Another enhancement that could be considered is using more sophisticated feature selection methods. The model also takes in consideration solely students that enrolled through the regular contingent, which means that students from other countries or from Madeira and Azores are not considered.

About information displaying, the creation of an application with an easy to use interface and with results displayed in a human friendly way could prove to be paramount for a less tech-savvy stakeholder to make use of the knowledge generated by the models.

## B.8 APPENDICES

### B.8.1 Related literature

Table A1: Studies addressing students' academic performance.

| Study | Main Objective | # Instances | Techniques | Dependent variable |
|---|---|---|---|---|
| [AEH12] | Predict student's final Grade Point Average in their degree. | 3360 students | Association Rules, Naive-Bayes | Final Grade split into five categorical values: Poor, Average, Good, Very Good, Excellent |
| [AAK+16] | Predict academic success of architecture students based on information provided in prior academic performance. | 101 students | Discriminant analysis, K-nearest Neighbour | Pass or fail the programme |
| [ABM13] | To predict the academic performance of Electrical Degree students. | 886 students | Neural Networks | Cumulative Grade Point Average at semester 8. |
| [GIK10b] | Determine profiles of students whose GPA is at least 2.0 (graduate) and students whose GPA is at least 3.0 (graduate with distinction) | At least 2699 students | Decision Trees | GPA is at least 3.0 + GPA is at least 2.0. |
| [HS17] | Determine which students may be struggling to complete their first academic year. | 6845 students | Logistic Regression, Neural Networks, Random Forest | Complete first year or not |

Continued on next page

**Table A1 – continued from previous page**

| Study | Main Objective | # Instances | Techniques | Dependent variable |
|---|---|---|---|---|
| [HF13] | Predict student academic performance in engineering. | 323 students | Multiple Linear Regression, Multilayer Perception Network, Radial Basis Function and Support Vector Machines | Students' scores on the dynamics final comprehensive exam |
| [LRSM15] | Determining graduation rates in Engineering for community college transfer students. | 472 students | Logistic Regression | Graduate or not |
| [MDDM16] | Determining students that are at risk of failing using data from previous assessments. | 2907 students | Naive-Bayes, Support Vector Machines, K-nearest Neighbour | Pass or fail the course |
| [MKG14] | Predict performance in the third semester of MCA students. | 250 students | Decision Trees and Random Forest | BAVG (Smaller than 60%), AVG (60%-70%) , ABVG (70%-79%) and EXCL (Larger than 80%) |
| [NZ14] | Predict the students' final grade in a course exam | 106 students | Decision Trees | Low (0-5), Medium (6-7) and Hight (8-10) |
| [Saa16] | Discover relations between students' personal and social factors, and their educational performance in the previous semester to then predict performance in the upcoming semesters. | 270 students | C4.5 decision tree, ID3 decision tree, CART decision tree, CHAID | Excellent (more than 3.60), Very Good (3.00-3.59), Good (2.50-2.99), Pass (less than 2.5) |
| [SCS+15] | Predict both the approval or failure and the grade of a student in a course or degree. | 5779 students | k-Nearest Neighbours, Random Forest, AdaBoost, CART decision trees, Support Vector Machines, Naive-Bayes, Ordinary Least Squares | Prediction of approval/-failure and prediction of grade |
| [VMS07] | Predict if the student has a low, medium or high risk of failing a programme | 533 students | Random Forests, Neural Networks, Decision Trees | High, medium or low risk of failure. |
| [WW14] | Identify students at risk of failure in the first semester. | 92 students | Multiple linear regression | Mean result in the first semester |

Table A2: Studies addressing students' academic performance using Learning Management Systems.

| Study | Main Objective | # Instances | Techniques | Dependent variable |
|---|---|---|---|---|
| [BSAD13] | Using Moodle data to keep students from falling behind their peers and giving up | 101 students | Expectation Maximisation, Hierarchical Clustering, Simple K-Means, X-Means | Clusters to classify students |

Continued on next page

**Table A2 – continued from previous page**

| Study | Main Objective | # Instances | Techniques | Dependent variable |
|---|---|---|---|---|
| [CAP+16] | Examine students' asynchronous learning processes via an Educational Data Mining approach using data extracted from Moodle logs | 140 students | Expectation Maximization, K-means | Final marks of students in a specific course. |
| [JVMM12] | Evaluate student performance by using information from Moodle's features | 260 students | AdaBoost, Bagging, C4.5, Linear Discriminant Analysis, Logistic Regression, Naive-Bayes, Neural Networks, Random Forest | E (Excellent), G (Good) or P (Poor) results |
| [LLM+14] | Early identification of students that might drop out a course | 400 students | ? | The student will drop out of the course or not |
| [MD09] | Identify at-risk students and allow for more timely pedagogical interventions using LMS tracking data. | 118 students | Multiple Regression, Logistic Regression | Student is at risk of failure (final grade lower than 60%), otherwise 'performing adequately or better' |
| [MB12] | Predict the success or failure of a student using data gathered by LMS, student interaciton with course material and self-reports. | 122 students | Decision Trees | Final grade |
| [Pal13] | Use student data stored in institutional systems to predict student performance. | 132 students | Binary Logistic Regression | Fail or not a course |
| [REZ+13] | Predict the marks that university students will obtain in the final exam of a course. | 1011 students | Decision trees, Neural Networks, Rule Induction | FAIL (if value is lower than 5), PASS (if value is between 5 and 7;), GOOD (if value is between 7 and 9), EXCELLENT (if value is above 9) |