Universidade do Minho
Escola de Engenharia

Mestrado Integrado em Engenharia Biomédica
Mestardo Integrado em Engenharia e Gestão de Sistemas de Informação

UC: Data Mining para a Ciência de Dados

Data Mining

André Ferreira (A81350),
Bruno Fonseca (A83029),
Daniel Costa (A81434),
Luís Silva (A80981).

Teacher: Paulo Cortez

Braga, November of 2020

Abstract

The level of education in Portugal is constantly evolving, although there is still a significant rate of failure, namely in Portuguese and Mathematics subjects. Aiming to offer aid in education development, it is relevant to explore ways to come up with methods that use Data Mining (DM) techniques that analyse school information, along with personal aspects of the students and make predictions from there so that more efficient help can be given, according to each student's predicted performance.

The two referred subjects (Mathematics and Portuguese) were first modeled with binary/five-level classification and regression scenarios. Inside of each type of scenario, five DM models (Random Forest, Naive, Decision Trees, Nearest Neighbor and Support Vector Machine) and four input configurations were evaluated, without any grades, with only the first period or second period grades, and with both period grades. Additionally, it was applied an unsupervised learning technique, Clustering, to divide the data into two different classes of students (pass or fail at the respective subject). It was then constructed a set of association rules to find out relations between the attributes and the student's outcome, passing or failing the core class.

The results show that generally, a good predictive accuracy can be achieved, depending on the chosen type of algorithm. Despite this, the accuracy is greatly influenced by the presence, or not, of the first and/or second period grades made available. While having both period grades attained the best accuracy, even without one or any of those grades, it still was possible to predict student achievement with good reliability, using only personal socio-economical variables.

With this work, more efficient prediction methods can be built as a way to improve the quality of education, mainly focused on the students with higher risk of failure.

Index

List of figures

1 Introduction

Education is very important in any country, both in terms of humanistic and economic development. It is known that in the last decades, education in Portugal has grown immensely, as well as the number of students placed in higher education. Despite the large growth, there are still flaws and a considerable percentage of students who simply give up.

The main causes of these failures are the Portuguese and Mathematics subjects, which are very important since they are fundamental to a student's knowledge. Due to these reasons, it is relevant to conduct a study in order to analyze the main factors that lead to failure in these subjects, to improve the education standards [1][2].

The present work aims to implement several data mining techniques in order to discover knowledge and patterns from data collected from high school students in the subjects of Portuguese and Mathematics. With this in mind, a methodology is followed in order to find out if there are certain personal aspects that influence school success and are the most relevant, that is, the aspects that have the greatest impact on student performance [2]. The discovery of these patterns would be phenomenal, since these can lead to a better understanding of the mistakes that have been made, as a way to improve them and, in turn, increase the quality of teaching [1].

The following work is organized into four distinct phases. The first phase is the Introduction, where the motivation for the project mentioned. In the second phase, it shows what was discussed during the meetings, as well as the division of tasks performed. In the third phase, the performed procedure is shown, as well as its results. In the last phase, assessments on the work are made, as well as analysis of limitations and future prospects.

2 Project execution

2.1 Group Planning

In this initial phase of the work, a choice of databases to be used was made, as well as a division of bibliographies by the members of the group in order to explore data mining and the R programming language by themselves and then combine the learnt insights on the different subjects. It was discussed with all group members and to optimize the productivity, it was reached a consensus that it would be better if the majority of the work was done jointly, so that any doubts were promptly discussed and resolved. This applied to the coding, which was written by one person at a time, while the others would be researching for ways to complete the tasks at hand. This included some modeling, mostly all of the Classification and Regression tasks and extraction of the respective results. The whole group wrote some of the general sections of the report, i.e. Abstract, Deployment and Conclusion.

2.2 Division of the work

This work was done by:

- André Ferreira A81350
- Bruno Fonseca A83029
- Daniel Costa A81434
- Luis Silva A80981

2.2.1 André Ferreira

As I had not yet contacted the R language, I started watching some tutorials for beginners on YouTube. Then, we split some recommended literature, and I was left to study "R and Data Mining: Examples and Case Studies" [3]. This article is a very complete resume of the various capabilities of R in data mining, on how to deal with databases (CSV files, servers etc.), explore the imported data (charts and others), outlier detection and how to deal with deal, text mining (data mineralization of text) and some different data mining algorithms, like Decision Tree, Random Forest, Regression and

Clustering. Regarding the report, the Data Understanding and Induction Rules sections were produced mainly by me, although always supported by the other colleagues. As for Data Preparation, it had equal participation from all members, both in writing R code and in the report. The modeling section was discussed among all the group members synchronously but mostly written by me, except clustering.

2.2.2 Bruno Fonseca

As a way to introduce ourselves to the R programming language, reading material was split among us. I tackled the book "The Art of R Programming"[4] , which covered the fundamentals of how to work with the most important R Data Structures and Functions. To complement this with how to apply the fundamentals, introductory YouTube videos were also watched. After we all got our bases, the synchronous meetings allowed the group to develop skills together and develop the project together synchronously. In this report, I mostly worked in the "Business Understanding" section, coming from the Systems Management course, whilst also being present and contributing to the discussion for the rest of the report.

2.2.3 Daniel Costa

As a way to begin complementing the little knowledge I had of programming in R, I went through "Exploratory Data Analysis with R", by Roger Peng [5]. Throughout the book, the basics of EDA (Exploratory Data Analysis) were covered. EDA serves purposes such as identifying relationships between relevant variables or searching for missing data or errors. Along with providing links with examples, the book helped understanding the dplyr package and data frame manipulation, explained a workflow for EDA and how to perform each phase. It also explains some ways of data visualization plots. As for coding, I mainly did outlier detection and removal while also helping construct a majority of the rest of the code, making sure it would all work and be organized. In this report, I wrote about the outliers, while also being in charge of overseeing the entirety of the writing work, checking for grammar and formatting errors and sometimes making sure some points were explicit, by rewriting the ambiguous phrases.

2.2.4 Luís Silva

In order to increase my short repertoire of programming skills in R, I made a study which envolve certain YouTube tutorials simultaneously with the reading of the book "R Programming for Data science", written by Roger D. Peng [6]. This study, together with the learning during classes, increased my skills in R, namely in terms of the basic concepts of the programming language (loops, functions, etc), as well as in the knowledge of packages (rminer and dplyr) and topics related to Data Science, like data frame manipulation and extracting important knowledge from the data frame. The classes were really important in order to know more about the supervised and unsupervised algorithms applied in DM, both on a theoretical and a practical level. Regarding the report, I mainly did some of the sections or subsections of the report, like introduction, normalization, one-hot encoding and clustering, but always with the help of the other members of the group.

2.3 Auto Evaluation

Overall, the group worked well together, without having any significant issues towards each member. Both the meetings and the individual work always met the outlined objectives for each week. Taking all things into account, the group agrees with a grade of 18, since we applied both supervised learning and unsupervised learning. On the other hand, we also apply different pre-processing methods as well as different algorithms for classification and regression. This resulted in an in-depth study of the "Student Performance" database. Considering all of the previous points, where it is explained what each member contributed to the presented work and the good group dynamic that was developed, the group considers that everyone deserves the same grade, matching the report's overall grade.

3 Study CRISP-DM

The data presented was extracted from the "UC Irvine Machine Learning Repository" [1]. During the Data Mining Process, the CRISP-DM Methodology [7] (Cross Industry Standard Process for Data Mining) was followed, dividing the work in six parts, which will be explained later:

- Business Understanding;
- Data Understanding;
- Data Preparation;
- Modeling;
- Evaluation;
- Deployment.

3.1 Business Understanding

3.1.1 Business objectives

The Portuguese Ministry of Education currently has no sure way to know which of the various factors of life end up affecting students' performance the most, meaning that it cannot take proper measures to improve the current academic situation with accuracy.

The main focus of this work is to determine through data mining which socio-economic variables and/or school performance variables affect a student's success in the subjects of Portuguese and Mathematics the most. Another goal is to categorize similar types of students to identify target groups of students that have the highest risk of failure, so that more informed measures can be taken by professionals in education administration to improve the academic performance of the group of students that have the worst grades. This project is deemed successful if it can determine, with a good level of accuracy, the most impactful factors for the success of students and if it is possible to categorize types of students to facilitate the creation of target groups for improvement.

3.1.2 Assess situation

To accomplish this task, a group of 4 data mining students will each use their own personal computer (totalling 4), each with a work hour pool of 140 hours (totalling 460

work hours)[8].The amount of students in the work group is fixed.

The raw data provided is a fixed extract in the "student-mat.csv" and the "student-por.csv" files available in the "Student Performance Data Set"[9]. The R tool must be used for the development of this data mining project. Other data and text mining tools may be used for support. To develop the data mining model, the "R Project" software environment for statistical computing and graphics (https://www.r-project.org/) and the "RStudio Desktop - Open Source Edition" IDE for R development (https://rstudio.com/) will be used. The data provided is assumed to be factual. The data mining model must not be plagiarized as well as the obtained results.

This project must be completed within and presented at the deadline at 12th of January, 2021. Every student in the group must abide by the contract "Contrato Data Mining para Ciência de Dados" available in annex, detailing the contract object, contracted commitments and evaluation criteria. Every citation must be correctly identified.

Terminology

- **Binary -** A property type where the value can only assume one of two possibilities. In a datatype, it can either be "true"(1) or "false"(0). In this case, during the modeling of student classification, it assumes "Passes" or "Fails" for a student's final grade;
- **Five-level -** A classification system used in the data mining models based on Erasmus grade conversion;
- **Regression -** A set of statistical processes for estimating the relationships between a dependent variable (often called the 'outcome variable') and one or more independent variables (often called 'predictors'). The most common form of regression analysis is linear regression, in which a researcher finds the line that most closely fits the data according to a specific mathematical criterion;
- **Classification -** Classification is a process related to categorization, the process in which ideas and objects are recognized, differentiated and understood. The student's final grade will undergo this process in a binary classification and a five-level classification;

- **Clustering -** The task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups;
- **Accuracy -** closeness of the measurements to a specific value. Not to confuse with precision, which is the closeness of the measurements to each other;
- **Sensitivity -** How the uncertainty in the output of a mathematical model or system (numerical or otherwise) can be divided and allocated to different sources of uncertainty in its inputs;
- **Root Mean Squared Error (RMSE) -** The standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are. In other words, it tells you how concentrated the data is around the line of best fit;
- **Mean Absolute Error (MAE) -** Measure of errors between paired observations expressing the same phenomenon.

Risks and Contingencies

Table 1: Risks and Contingency Plans

Risk	Consequence	Contingency Plan
Group Member's Personal Computer Failure	Inability to continue the execution of tasks	Inform through the group chat the other group members of the current situation. The other group members must distribute the inactive member's workload amongst themselves in the group chat while the issue is not fixed.
Group Member Illness/Injury	Inability to continue the execution of tasks	Inform the other group members of the current situation. The other group members must distribute the inactive member's workload amongst themselves in the group chat.
Group Member's impossibility to attend Data Mining Class	Group Member will not receive the Professor's feedback to improve upon the project	The Group Members that did attend the Data Mining class must accurately relay the information unto the Member that couldn't attend.

Costs and Benefits

This project is not funded by any organization and is entirely developed through the effort of the student work group so there is no assumed cost for it's execution besides the work hour input by each group project member. Hardware wear is very minimal and will not be accounted for. There are no transportation or food costs since every member is working from home. The benefit of this project, assuming it is successful, is subjective to whether or not the proper authorities to which the project results are destined to will find use for and benefit from the results. The main benefit is intangible since its ultimately derived from the outcome of the use of the project results and if it helped the students with the highest risk of failure.

3.1.3 Data Mining Goals

The project must produce a model that allows us to identify the academic and socioeconomic factors that determine the success or failure of a student in the subject of Portuguese and Mathematics. With that in mind, several data mining techniques, namely *Random Forest*, *Naive*, *Decision Tree*, *Nearest Neighbor* and *Support Vector Machine* are to be implemented in order to see which one is better at predicting the outcome of a student's final grade. Clustering models, namely *kmeans* and *DBSCAN* will also be tested for the purpose of grouping the students that have the highest risk of failure.

The success criteria for the classification and regression models is to at least be more accurate than a purely random decision, meaning that binary classification and regression models must have results with an accuracy better than the equivalent of 50% and five-level classification and regression models must have an accuracy better than the equivalent of 20%. The goal is to be as accurate as possible, hence the use of multiple models.

3.1.4 Project Plan

The execution of this project is of a flexible nature. The group decided that the most efficient use of time dedicated to work would be to hold scheduled weekly meetings outside of class periods, where all members could synchronously develop the project.

At the end of each meeting, tasks are assigned to each member to be completed until the next meeting, which will also be scheduled unanimously according to each group member's availability. The amount of work and the nature of each task is assigned according to the next deadline (Figure 1).

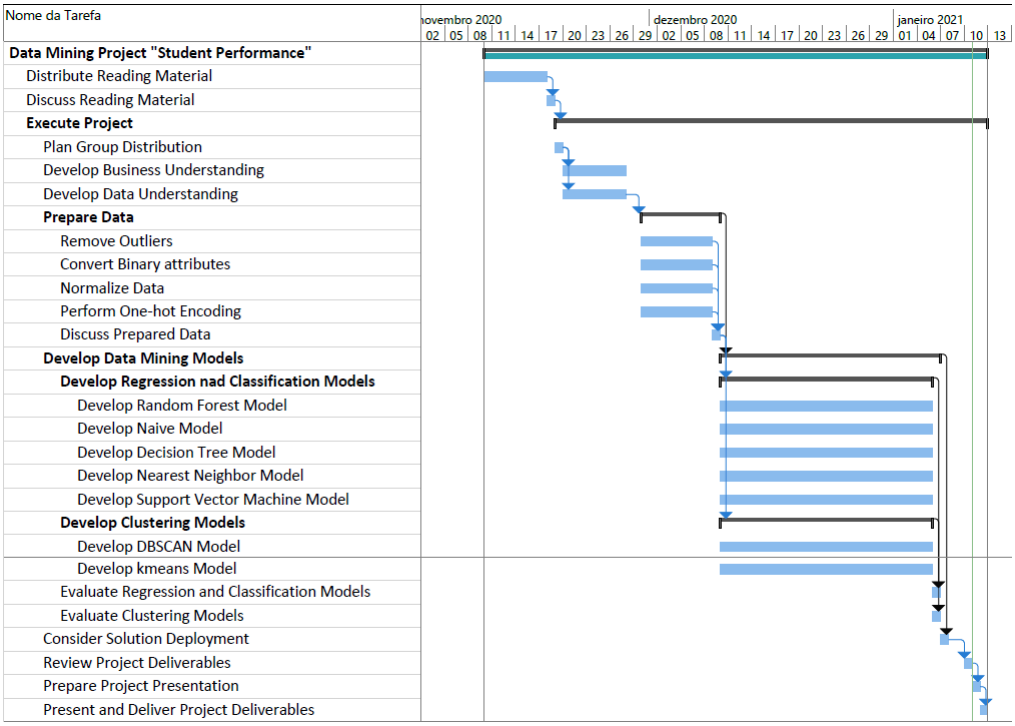

Figure 1: Project tasks and timeline illustrated by a Gantt Chart.

3.2 Data Understanding

Considering that all data in the data sets is factual, there are some analysis that need to be done before choosing or applying a machine learning algorithm.

These data sets were constructed by extracting data from two Portuguese schools (region of Alentejo) during the school year of 2005-2006. One of the two data sets refers to the grades obtained in the Mathematics with 395 examples, while the other refers to the Portuguese containing 649 records [1]. There are 32 attributes for each instance and a target attribute (G3) as shown in the Table 2 (obtained by surveys) and Table 3 (obtained by school reports).

Table 2: Attributes - Surveys

Attribute	Type	Description
school	Binary	School 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira
sex	Binary	Gender 'F'-female or 'M'-male
age	Numeric	Age from 15 to 22
address	Binary	Home address type 'U' - urban or 'R' - rural
famsize	Binary	Family size 'LE3'-less or equal to 3 or 'GT3'-greater than 3
Pstatus	Binary	Parent's cohabitation status 'T' - living together or 'A' - apart
Medu	Numeric	Mother's education 0-none, 1-primary education (4th grade), 2-5th to 9th grade, 3-secondary education or 4-higher education
Fedu	Numeric	Father's education 0-none, 1-primary education (4th grade), 2-5th to 9th grade, 3-secondary education or 4-higher education
Mjob	Nominal	Mother's job 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other'
Fjob	Nominal	Father's job 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other'
reason	Nominal	Reason to choose this school close to 'home', school 'reputation', 'course' preference or 'other'
guardian	Nominal	Student's guardian 'mother', 'father' or 'other'
traveltime	Numeric	Home to school travel time 1-<15 min., 2-15 to 30 min., 3-30 min. to 1 hour, or 4->1 hour
studytime	Numeric	Weekly study time 1-<2 hours, 2-2 to 5 hours, 3-5 to 10 hours, or 4 - >10 hours
failures	Numeric	Number of past class failures n if 1<=n<3, else 4
schoolsup	Binary	Extra educational support yes or no
famsup	Binary	Family educational support yes or no
paid	Binary	Extra paid classes within the course subject (Math or Portuguese) yes or no
activities	Binary	Extra-curricular activities yes or no
nursery	Binary	Attended nursery school yes or no
higher	Binary	Wants to take higher education yes or no
internet	Binary	Internet access at home yes or no
romantic	Binary	with a romantic relationship yes or no
famrel	Numeric	Quality of family relationships from 1 - very bad to 5 - excellent
freetime	Numeric	Free time after school from 1 - very low to 5 - very high
goout	Numeric	Going out with friends from 1 - very low to 5 - very high
Dalc	Numeric	Workday alcohol consumption from 1 - very low to 5 - very high
Walc	Numeric	Weekend alcohol consumption from 1 - very low to 5 - very high
health	Numeric	Current health status from 1 - very bad to 5 - very good

Table 3: Attributes - School reports

Attribute	Type	Description
absences	Numeric	Number of school absences from 0 to 93
G1	Numeric	First period grade from 0 to 20 of Math or Portuguese
G2	Numeric	Second period grade from 0 to 20 of Math or Portuguese
G3	Numeric	Final grade from 0 to 20 of Math or Portuguese

For the Mathematics data set, using the R language, it was possible to verify that there are no missing values (Figure 2) and, with the function "summary", it's possible notice that each attribute is made of only valid values (Figure 3).

```
> sapply(base, function(x) sum(is.na(x)))
    school        sex        age    address    famsize    Pstatus       Medu       Fedu       Mjob       Fjob     reason   guardian
         0          0          0          0          0          0          0          0          0          0          0          0
traveltime  studytime   failures  schoolsup    famsup       paid activities    nursery     higher   internet   romantic    famrel
         0          0          0          0          0          0          0          0          0          0          0          0
  freetime      goout       Dalc       Walc     health   absences         G1         G2         G3
         0          0          0          0          0          0          0          0          0
```

Figure 2: Number of missing values of each attribute in Mathematics data set.

```
> summary(base)
    school               sex                 age             address            famsize            Pstatus               Medu
 Length:395         Length:395         Min.   :15.0   Length:395         Length:395         Length:395         Min.   :0.000
 Class :character   Class :character   1st Qu.:16.0   Class :character   Class :character   Class :character   1st Qu.:2.000
 Mode  :character   Mode  :character   Median :17.0   Mode  :character   Mode  :character   Mode  :character   Median :3.000
                                       Mean   :16.7                                                         Mean   :2.749
                                       3rd Qu.:18.0                                                         3rd Qu.:4.000
                                       Max.   :22.0                                                         Max.   :4.000
      Fedu            Mjob               Fjob              reason            guardian          traveltime      studytime
 Min.   :0.000   Length:395         Length:395         Length:395         Length:395         Min.   :1.000   Min.   :1.000
 1st Qu.:2.000   Class :character   Class :character   Class :character   Class :character   1st Qu.:1.000   1st Qu.:1.000
 Median :2.000   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :1.000   Median :2.000
 Mean   :2.522                                                                               Mean   :1.448   Mean   :2.035
 3rd Qu.:3.000                                                                               3rd Qu.:2.000   3rd Qu.:2.000
 Max.   :4.000                                                                               Max.   :4.000   Max.   :4.000
    failures        schoolsup           famsup             paid           activities          nursery             higher
 Min.   :0.0000   Length:395         Length:395         Length:395         Length:395         Length:395         Length:395
 1st Qu.:0.0000   Class :character   Class :character   Class :character   Class :character   Class :character   Class :character
 Median :0.0000   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character
 Mean   :0.3342
 3rd Qu.:0.0000
 Max.   :3.0000
   internet           romantic             famrel          freetime         goout            Dalc            Walc
 Length:395         Length:395         Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
 Class :character   Class :character   1st Qu.:4.000   1st Qu.:3.000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.000
 Mode  :character   Mode  :character   Median :4.000   Median :3.000   Median :3.000   Median :1.000   Median :2.000
                                       Mean   :3.944   Mean   :3.235   Mean   :3.109   Mean   :1.481   Mean   :2.291
                                       3rd Qu.:5.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:2.000   3rd Qu.:3.000
                                       Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000
     health         absences            G1              G2              G3
 Min.   :1.000   Min.   : 0.000   Min.   : 3.00   Min.   : 0.00   Min.   : 0.00
 1st Qu.:3.000   1st Qu.: 0.000   1st Qu.: 8.00   1st Qu.: 9.00   1st Qu.: 8.00
 Median :4.000   Median : 4.000   Median :11.00   Median :11.00   Median :11.00
 Mean   :3.554   Mean   : 5.709   Mean   :10.91   Mean   :10.71   Mean   :10.42
 3rd Qu.:5.000   3rd Qu.: 8.000   3rd Qu.:13.00   3rd Qu.:13.00   3rd Qu.:14.00
 Max.   :5.000   Max.   :75.000   Max.   :19.00   Max.   :19.00   Max.   :20.00
```

Figure 3: Summary of each attribute in Mathematics data set.

The "sex" attribute is roughly balanced (208 instances of "F" and 187 of "M"), against the attributes "address", "famsize", "Pstatus" that have a discrepant number of instances for each value (R-88/U-307, LE3-114/GT3-281 and A-41/T-354, respectively). The attributes "Medu" and "Fedu" are significantly balanced but, on the other hand, "traveltime", "studytime" and "failures" are poorly distributed with a number of entries for a value discrepant from the others. The following binary attributes "higher",

"school", "schoolsup", "internet", "nursery", "romantic", "famsup", "paid" and "activities" are ordered from the least balanced to the most balanced ("higher" has a difference of 355 and "activities" just a difference of 7). The remaining attributes are considerably balanced, and it is important to highlight the attribute "absences" and "Dalc" which have few entries for high values. Finally, for the label "G3" there is a high percentage of values next to the grade "10".

For the Portuguese data set there are no missing values either (Figure 4) and, using the function "summary" again, it is possible to notice that each attribute only has valid values (Figure 5).

```
> sapply(base, function(x) sum(is.na(x)))
    school        sex        age    address    famsize    Pstatus       Medu       Fedu       Mjob       Fjob     reason   guardian 
         0          0          0          0          0          0          0          0          0          0          0          0 
traveltime  studytime   failures  schoolsup     famsup       paid activities    nursery     higher   internet   romantic     famrel 
         0          0          0          0          0          0          0          0          0          0          0          0 
  freetime      goout       Dalc       Walc     health   absences         G1         G2         G3 
         0          0          0          0          0          0          0          0          0 
```

Figure 4: Number of missing values of each attribute in Portuguese data set.

```
> summary(base)
    school              sex                 age           address             famsize             Pstatus              Medu      
 Length:649         Length:649         Min.   :15.00   Length:649         Length:649         Length:649         Min.   :0.000  
 Class :character   Class :character   1st Qu.:16.00   Class :character   Class :character   Class :character   1st Qu.:2.000  
 Mode  :character   Mode  :character   Median :17.00   Mode  :character   Mode  :character   Mode  :character   Median :2.000  
                                       Mean   :16.74                                                         Mean   :2.515  
                                       3rd Qu.:18.00                                                         3rd Qu.:4.000  
                                       Max.   :22.00                                                         Max.   :4.000  
      Fedu            Mjob               Fjob              reason            guardian          traveltime      studytime     
 Min.   :0.000   Length:649         Length:649         Length:649         Length:649         Min.   :1.000   Min.   :1.000  
 1st Qu.:1.000   Class :character   Class :character   Class :character   Class :character   1st Qu.:1.000   1st Qu.:1.000  
 Median :2.000   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :1.000   Median :2.000  
 Mean   :2.307                                                                               Mean   :1.569   Mean   :1.931  
 3rd Qu.:3.000                                                                               3rd Qu.:2.000   3rd Qu.:2.000  
 Max.   :4.000                                                                               Max.   :4.000   Max.   :4.000  
    failures        schoolsup           famsup             paid           activities          nursery             higher         
 Min.   :0.0000   Length:649         Length:649         Length:649         Length:649         Length:649         Length:649        
 1st Qu.:0.0000   Class :character   Class :character   Class :character   Class :character   Class :character   Class :character  
 Median :0.0000   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character  
 Mean   :0.2219                                                                                                                    
 3rd Qu.:0.0000                                                                                                                     
 Max.   :3.0000                                                                                                                     
   internet           romantic             famrel         freetime        goout            Dalc            Walc           health     
 Length:649         Length:649         Min.   :1.000   Min.   :1.00   Min.   :1.000   Min.   :1.000   Min.   :1.00   Min.   :1.000  
 Class :character   Class :character   1st Qu.:4.000   1st Qu.:3.00   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.00   1st Qu.:2.000  
 Mode  :character   Mode  :character   Median :4.000   Median :3.00   Median :3.000   Median :1.000   Median :2.00   Median :4.000  
                                       Mean   :3.931   Mean   :3.18   Mean   :3.185   Mean   :1.502   Mean   :2.28   Mean   :3.536  
                                       3rd Qu.:5.000   3rd Qu.:4.00   3rd Qu.:4.000   3rd Qu.:2.000   3rd Qu.:3.00   3rd Qu.:5.000  
                                       Max.   :5.000   Max.   :5.00   Max.   :5.000   Max.   :5.000   Max.   :5.00   Max.   :5.000  
    absences            G1              G2              G3       
 Min.   : 0.000   Min.   : 0.0   Min.   : 0.00   Min.   : 0.00  
 1st Qu.: 0.000   1st Qu.:10.0   1st Qu.:10.00   1st Qu.:10.00  
 Median : 2.000   Median :11.0   Median :11.00   Median :12.00  
 Mean   : 3.659   Mean   :11.4   Mean   :11.57   Mean   :11.91  
 3rd Qu.: 6.000   3rd Qu.:13.0   3rd Qu.:13.00   3rd Qu.:14.00  
 Max.   :32.000   Max.   :19.0   Max.   :19.00   Max.   :19.00  
```

Figure 5: Summary of each attribute in Portuguese data set.

Approximately, all the attributes follow a similar distribution as in the Mathematics data set, being important to note that the "sex" attribute is approximately balanced (383 instances of "F" and 266 of "M"), "school" has less discrepancy between values "MS" and "GP" and for the label "G3", such as in the other data set, there is a high percentage of values next to the grade "10".

3.3 Data Preparation

Data preparation is a fundamental stage of data analysis. This step is highly important, since missing values, incomplete data, noise data and inconsistent data can have a high impact on the performance of the algorithms.

In our work, since there are no missing values, our stage of data preparation involved the analysis of outliers, data binarization, data normalization and one-hot encoding.

3.3.1 Outliers

In order to evaluate any outliers existing in the data, as well as possibly remove them, for each of the relevant numerical attributes (such as the age, G1/G2 grades, number of absences). Anything beyond the whiskers is marked as an "outlier" and is plotted separately as an individual point.

As a matter of example, as it is observed in Figure 6, there are a lot of outliers when checking the "absences" attribute, all of them being of a high number of absences. Although they might affect the classification and regression results that are going to be studied later on, they are not removed because they were obtained with mark reports, as well as having high importance, along with G1 and G2 grades, for classification and regression, as it is mentioned in Section 3.4.

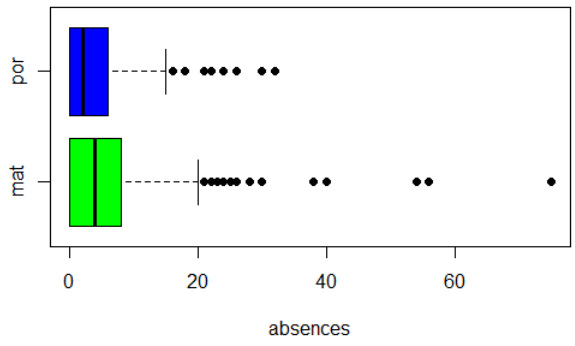

Figure 6: Boxplot of the absences attribute for both Mathematics (mat) and Portuguese (por) data sets.

Also, according to the boxplot in Figure 7, in both Mathematics and Portuguese data sets, there are several data points (students) which their G2 grade is being considered

as an outlier. A similar situation happens when referring to G1 grade. Despite this, due to being inside the grading scale and also for the same reason in the "absences" case, both G1 and G2 grades were left untouched.

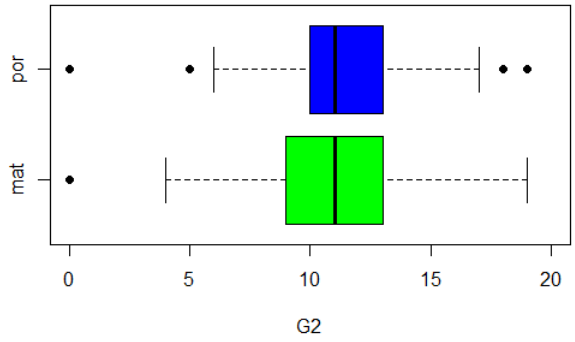

Figure 7: Boxplot of the G2 attribute for both Mathematics (mat) and Portuguese (por) data sets.

The only attribute that was considered to be better to remove outliers was the students' age. In the boxplot of Figure 8, it is seen that students with 22 years old are considered to be outliers. According to the data sets, in both Mathematics and Portuguese data sets, there was only 1 student of that age, thus being removed from the data sets.

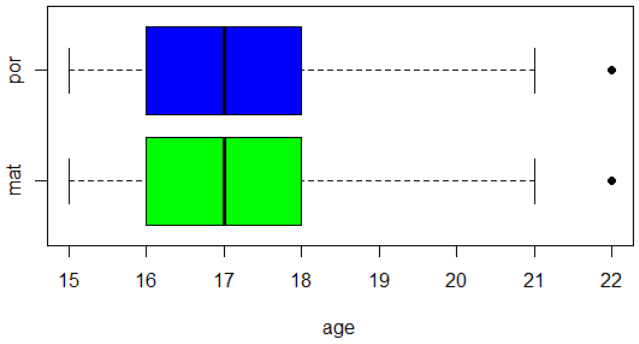

Figure 8: Boxplot of the age attribute for both Mathematics (mat) and Portuguese (por) data sets.

3.3.2 Binary attributes

The binary attributes are changed from "yes" and "no" to 1 and 0 before applying the machine learning mechanisms, as these algorithms handle numerical values better than characters.

3.3.3 Normalization

Some Machine Learning algorithms are sensitive to the scale presented in the data. To overcome this problem, the data is normalized so that it presents a similar scale [0-1]. This normalization can be done in several ways, such as standardization or min-max scale.

In our work, it was chosen the standardization algorithm to perform data normalization.

3.3.4 One-hot encoding

There are two types of data in the database used, categorical and numeric attributes. Some algorithms used in Machine Learning have problems with the use of categorical attributes, therefore being necessary to transform them into numeric attributes. Label encoding or One-hot encoding algorithms can be used for this transformation.

For this purpose, the One-hot encoding algorithm was chosen, since it is much more efficient. Label encoding assigns a number to each column, which is not relevant because certain algorithms will give more importance to higher numbers. On the other hand, the One-Hot encoding algorithm creates several columns with the names of the attributes of a column that has categorical data. These columns only display the values of 1 and 0, depending on whether the column has the categorical value or not, respectively.

3.4 Modeling

3.4.1 Classification and Regression

For each data set, three different scenarios were tested, more concretely:

- SC1 - The label will be binary, if G3>=10 then pass (1), else fail (0) (Figure 9);

- SC2 - The label will be divided in 5 levels (from insufficient (5) to very good (1)) (Figure 10);
- SC3 - The real grade (from 1 to 20, where 1 is the worst and 20 the best grade) (Figure 11).

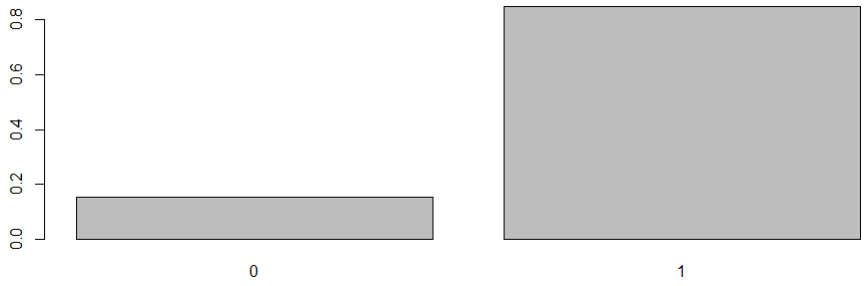

Figure 9: Percentage of each value in G3 label (SC1).

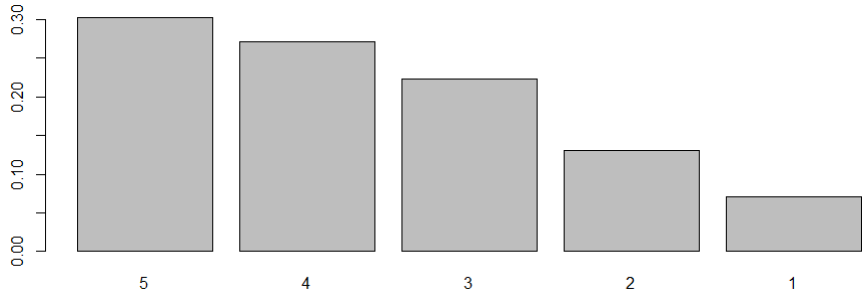

Figure 10: Percentage of each value in G3 label (SC2).

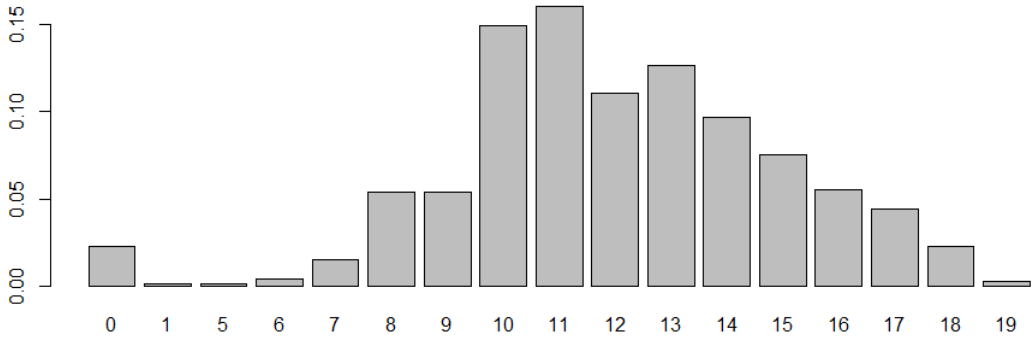

Figure 11: Percentage of each value in G3 label (SC3).

In order to test the predictability of the social and compartmental attributes, four input configurations were used:

- InA - With all the attributes;

- InB - Without the G1;
- InC - Without the G2;
- InD - Without the G1 and G2.

These two attributes (G1 and G2 grades) are the most predictive for the G3 label when random forest algorithm is used (considering Node Purity, Figure 12), probably because they are influenced by all the other attributes and/or the G3 score is also calculated based on the G1 and G2 scores. Also, InB configuration is not realistic as we can't know the second period grade without first knowing the first period grade. This configuration only served to compare with InC, to see which of the grades had more impact on the prediction's performance.

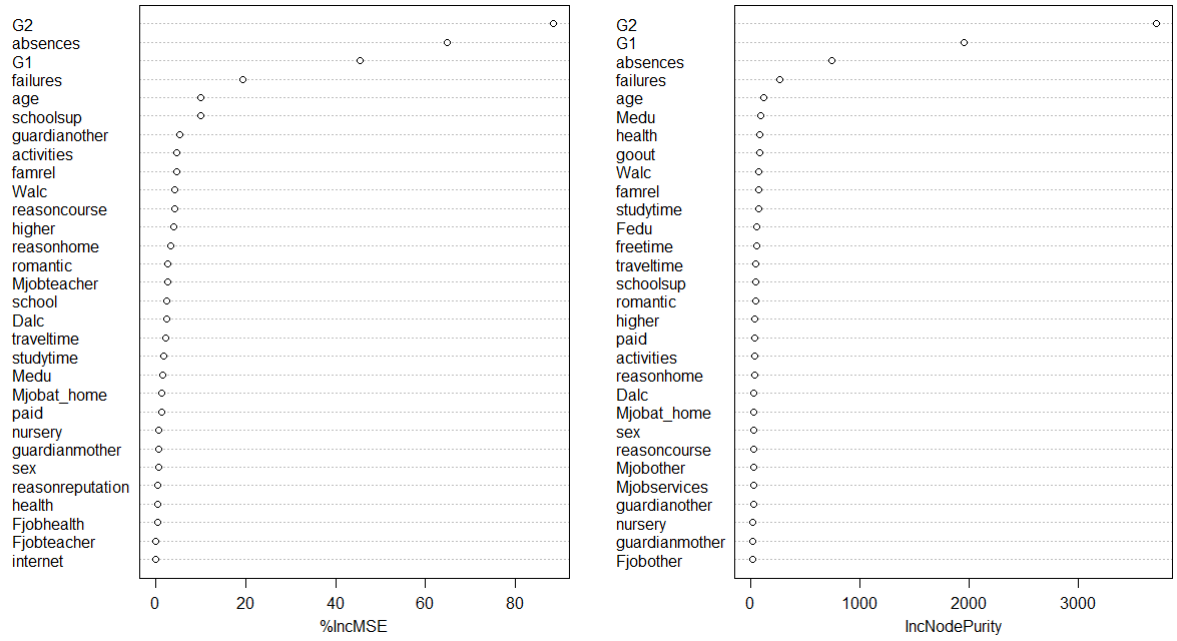

Figure 12: Attribute importance for prediction of label G3. Mean Decrease Accuracy (%IncMSE) and Mean Decrease Gini (IncNodePurity) sorted by importance.

The SC2 scenario is based on the Erasmus grade conversion system (Table 4) [1].

Table 4: Five-level classification system based on Erasmus grade conversion

Country	**1** (excellent/very good)	**2** (good)	**3** (satisfactory)	**4** (sufficient)	**5** (fail)
Portugal/France	16-20	14-15	12-13	10-11	0-9
Ireland	A	B	C	D	F

3.4.2 Association Rules

Association Rule learning are rule-based machine learning methods to discover relationships between variables in databases [10][11]. Most algorithms can only handle categorical attributes, so it is necessary to change all values from numeric to categorical. For this, the following code was used:

```
sapply(colnames(baseAR),function(name){ paste(name,baseAR[,name],sep="_")})
```

changing all the attributes of each line to their column name and value, separated by "_". Only will be considered one scenario, SIR1 - G1, G2 and G3 are changed to binary (if G1/G2/G3>=10 then pass, else fail).

Since the main difference between algorithms is the computational efficiency, only the "Apriori" algorithm will be used [12], due its high efficiency. The association between the various attributes and the success or failure of the students will be the objective of this application, so that there is more attention in these situations, if possible.

3.4.3 Clustering

Clustering is an unsupervised learning method that divides the population into groups that are more similar to each other.

In order to evaluate the existing clustering algorithms, k-Means and DBSCAN (Density-Based Spatial Clustering on Applications with noise) were implemented.

There are differences between these algorithms in terms of implementation and the main rule on which it is based. While K-Means is based on Euclidean distance, it needs an initial definition of the desired number of clusters and is strongly influenced by initializations, DBSCAN is based on density (similar points in the same space), does not need to explain the number of clusters required and it is not strongly influenced by initializations. However, for an accurate DBSCAN a good definition of the threshold distance parameters is required.

3.5 Evaluation

3.5.1 Classification and Regression - with outliers

For a better analysis of the results it was used a confusion matrix, which can give the data mining model performance (to ensure that the results are statistically relevant), namely the Accuracy (1), Sensitivity (2), Specificity, Precision, Classification error, Cohen's kappa and Recall. Only the first two metrics will be used to evaluate the SC1 and SC2 scenarios, with their respective results being in Table 5 and 7 and the Root Mean Squared Error (RSME) (3) and Mean Absolute Error (MAE) (4) to the SC3 scenario, where Table 9 shows RSME results and and Table 10 shows MAE results. The higher the precision and sensitivity the better the model, unlike RSME and MAE where the smaller the value is, the better.

$$Accuracy = TP + TN/(TP + TN + FP + FN) \tag{1}$$

$$Sensitivity = TP/(TP + FN) \tag{2}$$

$$RMSE = \sqrt{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2 / N} \tag{3}$$

$$MAE = \sqrt{\sum_{i=1}^{N} |y_i - \hat{y}_j|} \tag{4}$$

It is worth noting that, with the Naive algorithm, the results for each type of classification is the same in all scenarios because this algorithm assumes the most common class of G3 and applies it to the predictions and on the regression model it assumes the mean value of all G3 grades.

Table 5: Binary classification results using accuracy (bold value is best for each input configuration)

Input	Math					Portuguese				
Model	RF	NV	DT	NN	SVM	RF	NV	DT	NN	SVM
InA	91.0	68.3	**91.6**	77.9	88.5	**91.3**	86.2	90.9	86.7	89.3
InB	90.9	68.3	89.5	68.9	**91.8**	89.0	86.2	**92.1**	86.1	85.6
InC	80.3	68.3	**87.9**	72.6	79.8	85.9	86.2	**88.2**	86.6	85.1
InD	65.6	68.3	65.1	57.6	**69.0**	**86.6**	86.2	85.6	85.5	86.4

In order to visualise the significance of the predictions, the confusion matrix of the Decision Tree algorithm using the Maths database was elaborated. As most of the predictions are on the main diagonal (in bold), the model presented is an excellent choice since it fits well with the data. The confusion matrix is presented in the Table 6, with a sensibility of 90.63% (considering fail as true positive) which is a good value, proving that this is a great model. The decision tree algorithm has a high accuracy for most situations although, for Portuguese, the best model has been obtained using random forest.

Table 6: Confusion matrix of the mathematical binary classification

		References	
		0	1
Predictions	0	**29**	7
	1	3	**80**

Table 7: Five-level classification results using accuracy (bold value is best for each input configuration)

Input	Math					Portuguese				
Model	RF	NV	DT	NN	SVM	RF	NV	DT	NN	SVM
InA	75.0	47.1	**86.9**	42.6	56.6	67.3	28.9	**74.1**	44.4	55.1
InB	65.6	47.1	**85.2**	38.5	43.1	67.0	28.9	**74.4**	35.8	51.0
InC	60.3	47.1	**64.7**	40.3	47.6	54.5	28.9	**60.4**	36.8	43.6
InD	45.6	47.1	**47.2**	28.0	41.3	42.2	28.9	38.8	30.9	**43.1**

The confusion matrix was also designed for the classification between 5 levels. The confusion matrix of the Decision Tree algorithm using the Maths database is represented in the Table 8, with a sensibility of 93.85% for the class 5 (negative grade) and 76,19% for the class 4 (from 10 to 11). The sensibility value of class 5 is very good because almost all students with a real negative final grade will be classified correctly, unlike students with grades between 10 and 11 (class 4), where there is a high percentage that is poorly classified, but in this particular situation, since some are classified as fail (3 of 21) there will not be so much loss. For this scenario, the decision tree algorithm obtained the best accuracy for all configurations except for the last one, where the best accuracy was obtained using the SVM algorithm, in the Portuguese data set.

Table 8: Confusion matrix of the mathematical 5 levels classification

		References				
		5	4	3	2	1
Predictions	5	**61**	3	0	0	0
	4	4	**16**	0	0	0
	3	0	2	**12**	2	0
	2	0	0	0	**15**	1
	1	0	0	0	0	**2**

Table 9: The real grade results using RMSE (bold value is best for each input configuration)

Input	Math					Portuguese				
Model	RF	NV	DT	NN	SVM	RF	NV	DT	NN	SVM
InA	**1.735**	4.590	1.996	3.635	2.214	**1.313**	3.233	1.476	2.301	1.468
InB	**1.906**	4.590	1.996	4.015	2.279	**1.429**	3.233	1.476	2.549	1.481
InC	**2.451**	4.590	2.664	4.189	2.979	1.785	3.233	**1.730**	2.713	1.891
InD	**3.930**	4.590	4.361	4.828	4.237	**2.665**	3.233	2.934	2.668	2.713

The Figure 13 presents a residual plot to show how far away the projections are compared to the actual grades (values in the red line).

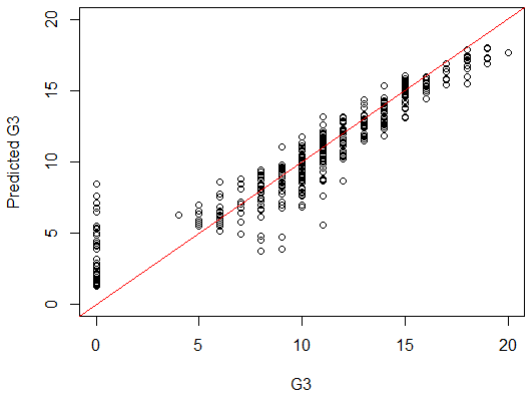

Figure 13: Correlation between the model's predictions and its actual results.

Table 10: The real grade results using MAE (bold value is best for each input configuration)

Input	Math					Portuguese				
Model	RF	NV	DT	NN	SVM	RF	NV	DT	NN	SVM
InA	**1.134**	3.438	1.218	2.671	1.386	**0.8144**	2.409	0.8551	1.639	0.8866
InB	1.282	3.438	**1.218**	2.992	1.400	0.8915	2.409	**0.8551**	1.835	0.9109
InC	**1.776**	3.438	1.886	3.097	1.973	1.2190	2.409	**1.199**	1.933	1.210
InD	**2.966**	3.438	3.264	3.644	3.119	**1.9270**	2.409	2.163	1.915	1.933

3.5.2 Classification and Regression - without outliers

In this scenario, considering both metrics, the random forest algorithm obtained the best results for most configurations, unlike the previous situations (binary and five-level).

Since only students over the age of 21 were removed from the data sets, the following tables (Table 11 and 12) summarize the results if all outliers were removed, using boxplots.

Table 11: Binary and 5-levels classification results using accuracy, without outliers (bold value is best for each input configuration)

	Binary			**5-Levels**		
Model	RF	DT	SVM	RF	DT	SVM
InA	90.29	89.32	**92.23**	74.77	**86.92**	41.12
InB	**88.07**	**88.07**	**88.07**	73.87	**84.68**	55.86
InC	77.27	**82.73**	73.64	57.80	**60.55**	46.79
InD	**76.79**	69.64	75.00	**41.44**	34.23	39.64

Table 12: The real grade results using RMSE and MAE, without outliers (bold value is best for each input configuration)

	RMSE			**MAE**		
Model	RF	DT	SVM	RF	DT	SVM
InA	**1.78**	2.02	2.34	**1.15**	1.30	1.41
InB	**1.95**	2.02	2.39	**1.29**	1.43	1.30
InC	**2.30**	2.45	2.78	**1.67**	1.75	1.84
InD	**3.64**	4.04	3.94	**2.76**	3.08	2.91

For the first configuration in the binary scenario, the SVM algorithm has the best accuracy (higher than the result obtained with the decision tree in the original model). Considering the 5-level model, the result was quite identical to the previous one, with no relevant difference. In the last scenario, the results were noticeably worse, in both metrics, when the outliers were removed, with the random forest algorithm again showing the best performance.

Considering the last configuration, in all scenarios, except SC2, the removal of outliers significantly improved the results. In view of this, for SC1, the removal of outliers improved the overall results. In the case of SC2, there was no relevant change when G1 and G2 grades are available (when they are not removed, the results become worse). For SC3, if the first and second period grades are available, the existence of all instances favors the results, unlike when they are not used (InD) where removing outliers improves the model.

3.5.3 Association Rules

It is possible to identify situations with high relation in databases using some measures such as support (5), confidence (6) and lift (7) [11].

$$Support = \frac{frq(X,Y))}{N} \tag{5}$$

$$Confidence = \frac{frq(X,Y))}{frq(N)} \tag{6}$$

$$Lift = \frac{Support}{Supp(X) * Supp(Y)} \tag{7}$$

For each data set, the rules that result in "fail" or "pass" in the G3 attribute will be evaluated, using SIR1. In the Tables 13, 14 and 15 are some rules that result in "pass" using the "Apriori" algorithm, sorted by lift, support and confidence, respectively. The minimum support, confidence and maxlen (maximum number of attributes per rule) were set to 0.2, 0.8 and 4.

Table 13: Results of association rules for G3 pass using Math data set, ordered by lift

lhs	support	conf	coverage	lift	count
famsup_no, G2_pass, schoolsup_no	0.2278	1.000	0.2278	1.4907	90
activities_no, G2_pass,romantic_no	0.2076	1.000	0.2076	1.4907	82
activities_no, Dalc_1,G2_pass	0.2025	1.000	0.2025	1.4907	80
G2_pass, guardian_mother, paid_no	0.2051	1.000	0.2051	1.4907	81
Dalc_1, G1_pass, G2_pass	0.4152	1.000	0.4177	1.4815	164

Table 14: Results of association rules for G3 pass using Math data set, ordered by support

lhs	support	conf	coverage	lift	count
G2_pass	0.6101	0.9679	0.6304	1.4427	241
G2_pass, higher_yes	0.5949	0.9671	0.6152	1.4415	235
G1_pass	0.5747	0.8972	0.6405	1.3374	227
G1_pass, G2_pass	0.5595	0.9822	0.5696	1.4641	221
G1_pass, higher_yes	0.5595	0.8984	0.6228	1.3391	221

Table 15: Results of association rules for G3 pass using Math data set, ordered by confidence

lhs	support	conf	coverage	lift	count
famsup_no, G2_pass, schoolsup_no	0.2278	1.0000	0.2278	1.4906	90
activities_no, G2_pass, romantic_no	0.2076	1.0000	0.2076	1.4906	82
activities_no, Dalc_1, G2_pass	0.2025	1.0000	0.2025	1.4906	80
G2_pass, guardian_mother, paid_no	0.2051	1.0000	0.2052	1.4906	81
Dalc_1, G1_pass, G2_pass	0.4152	0.9939	0.4177	1.4815	164

Analysing the Table 13, it is possible to verify that there are 4 rules with the highest lift value, having in common G2_pass, so it is possible to affirm that this attribute has a big relation to the final success. In Table 14, the rules with the attributes G2_pass and G1_pass has the biggest support values. Finally, Table 15 has, once again, G2_pass and G1_pass as the most important attributes. Given this, for success in Mathematics, the rule with the highest lift and confidence is not to have family educational support, to pass G2 and not to have extra educational support. The rule with highest support is just pass G2.

For the unsuccessful situation in G3 the minimum support value and confidence were changed to 0.1 and 0.6 when G3_fail, respectively, because there are far fewer instances of students who failed. The rule that presents the greatest lift and confidence is to fail in G1 and G2 and have access to the internet (5.0220 and 0.7738, respectively). The greatest support belongs to the rule in which there is failure in G2 (0.1371).

For the Portuguese data set in G3 fail situation, the rules with highest support value

is fail in G2 (0.1371). The biggest confidence and lift are reached when fail in G1 and G2 and attended at the nursery school (0.7738 and 5.0220, respectively).

Changing the minimum amount of support and confidence back to 0.2 and 0.8, the rule with the highest support is not to have extra paid Portuguese lessons (0.7997). Pass in G1 and G2 and having a mother with higher education is the rule with the highest support and lift values (1 and 1.1821, respectively).

3.5.4 Clustering

In order to differentiate groups of students (students who fail and who pass) a study of clustering was carried out with the Mathematics database. However, this study was not performed directly, i.e. using all the variables in the database.

With this in mind, the most relevant variables were chosen to carry out the division of students (age, G2 and absences). After the choice of the variables, they were normalized for better clustering operation.On the other hand, as it was wished to carry out an evaluation of different clustering algorithms, K-Means and DBSCAN algorithms were applied.

In the case of DBSCAN, before the algorithm implementation, two parameters had to be calculated (minimum points and threshold distance). For the minimum parameter of points, the value of 4 points was chosen. Based on this value, the optimal threshold distance computing value was calculated with the help of k-nearest neighbor distances. The result of this calculation is shown in Figure 14.

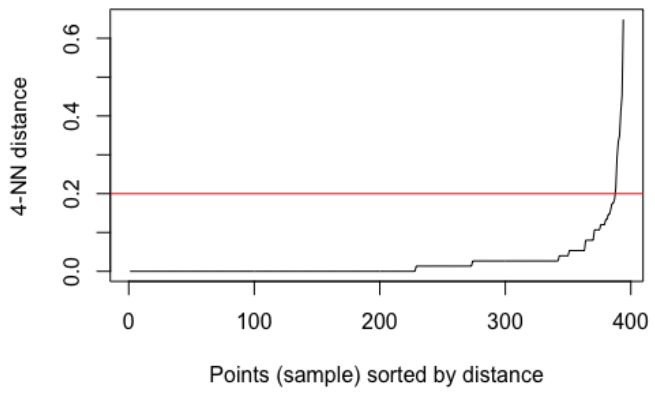

Figure 14: K-Distance Plot in order to find the optimal distance value.

It can be seen in the Figure 14 that the optimal threshold distance value is around of 0.2, which represents the neck of the plot. Therefore, the calculated parameters were used as inputs to the DBSCAN algorithm.

The results of the different tests (K-Means and DBSCAN) can be seen in the Figure 15.

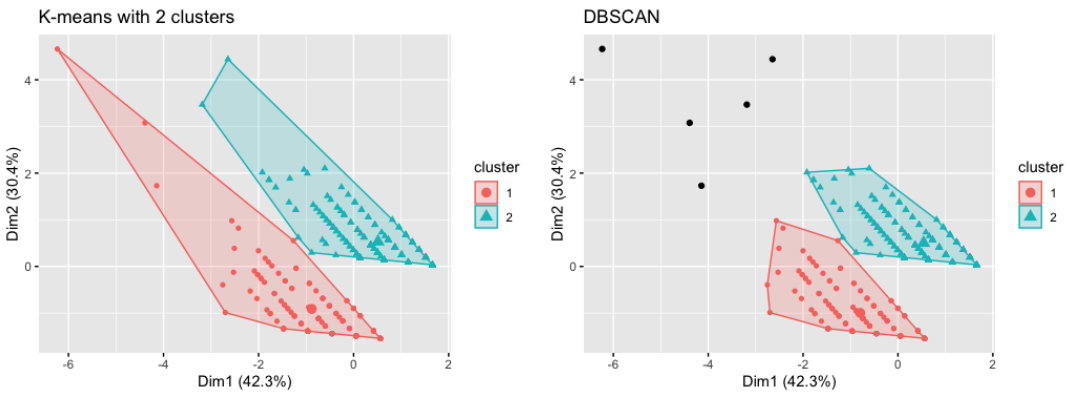

Figure 15: Clustering using K-Means and DBSCAN.

In the right plot of the Figure 15 , some black dots can be seen. These dots represent outliers, proving that the DBSCAN algorithm is more robust against outliers.

When the DBSCAN algorithm is performed with the number of minimum points equal to 4 and the threshold distance value of 0.4, the algorithm presents the same result as K-Means. This result can be seen in the Figure 16.

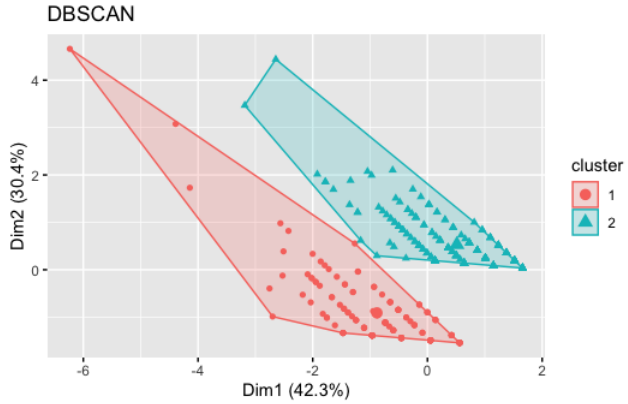

Figure 16: Clustering using DBSCAN with threshold distance of 0.4 units.

In terms of internal evaluation, both methods had a Dunn index value of 0.894, which means that the clusters are compacted and well separated[13].

On the other hand, a external evaluation was made using the Rand Index. In this test, both methods had a value of 0.850, which is an excellent value, since it means that there is a great similarity between the realized partition and the real one.

3.6 Deployment

Although the models won't properly deployed, it is possible to outline a plan to make them useful in some way. Being successful, this work should be able to be implemented in a real life situation, where students' characteristics and previous school performance would be analysed and, based on that, predict each student's success, whether it is pass or fail in Mathematics and Portuguese, or having grades within a certain interval.

The primary objective is to be able to preemptively detect a student's possible fail on a subject (binary scenario), so that measures can be taken by the responsible educators in order to help the student to improve their grades. Depending on when the work is intended to be deployed, different input configurations can be used, explained in Section 3.4.1.

If the school wants to interact and help the students in these matters at the beginning of the academic year, the InD configuration can be explored in such way, given that it predicts grades according only to personal aspects, not accounting for grades. On the other hand, if it is at the end of the first period, when G1 is already know, then InC configuration can give better results than the previous one. Finally, if both G1 and G2 are available to be used, that is, at the end of the second period, InA will certainly predict better than the other configurations.

More than only identifying a pass or fail situation, the five level scenario can also be applied, to split the students in grade intervals. This could be used to, for example, get different amounts of help, depending on their predicted grade interval. While the fail students are to be given more attention in general when compared to others, students that pass but still have low grades should also be helped, as they have a higher risk of failing too.

At the end of the year, the new data can be added to the models so that these give even better prediction results, further enhancing the predictive algorithms.

4 Conclusion

This work consisted on applying data mining techniques to predict the students' performance based on socio-economical variables as well as school reports for the Mathematics and Portuguese subjects. These personal characteristics and school grades can be applied to predict the student's final grade and be used to evaluate future methodologies to better support each student according to their education needs.

Based on this premise, several Data Mining techniques were tested and analyzed, to achieve better results. Some main algorithms were chosen and tested using different scenarios (binary and five-levels classification; and regression), to evaluate the effect of previous grades (G1 and G2) on the final grade (G3). In the classification task, the Decision Tree algorithm was better in in the majority of the scenarios and independently of the classification being binary or five-level and Math or Portuguese. For the regression task, the Random Forest algorithm performed better than the other parts, with the exception of some scenarios. Both in the classification and regression tasks the original goals were surpassed, being viable to be implemented in a real life scenario.

In addition, Association Rules were studied to find out relations between the data, namely identifying interesting rules between variables and the class G3 grade. As expected, the rules with the greatest support and confidence in relation to success in G3 are the pass or fail in G1 and G2. As seen in Section 3.4.1, the attributes with a higher importance with G3 were the school reports' attributes (G1, G2 and absences).

Clustering algorithms were used to distinguish between groups of people more likely to fail and groups more likely to pass Mathematics. Two of the main clustering algorithms (K-Means and DBSCAN) were implemented, which despite the different ways of acting, presented similar results with specific input parameters. For internal evaluation, as both algorithms presented a significant Dunn index (89.4%), it was found that these are able to distinguish data in two different clusters. With respect to an external evaluation, both methods presented a 85.0% in Rand Index, which means that the predicted clusters are very similar to the real ones.

In view of this, the generated models can be applied to improve the efficiency of teaching in Portuguese schools, helping the students who need it most.

References

[1] Paulo Cortez and Alice Maria Gonçalves Silva. Using data mining to predict secondary school student performance. 2008.

[2] Dorina Kabakchieva. Predicting student performance by using data mining methods for classification. *Cybernetics and information technologies*, 13(1):61–72, 2013.

[3] Yanchang Zhao. R and data mining: Examples and case studies. *R and Data Mining*, pages 1–4, 2011.

[4] Norman Matloff. The art of r programming. *http://heather. cs. ucdavis. edu/˜ matloff/132/NSPpart. pdf¿. Acesso em*, 1(03):2018, 2009.

[5] Roger Peng. *Exploratory data analysis with R*. Lulu.com, 2012.

[6] Roger D Peng. *R programming for data science*. Leanpub, 2016.

[7] Peter Chapman, Janet Clinton, Randy Kerber, Tom Khabaza, Thomas Reinartz, C. Russell H. Shearer, and Robert Wirth. Crisp-dm 1.0: Step-by-step data mining guide. 2000.

[8] Universidade do Minho. Creditosects na uminho. `(https://alunos.uminho.pt/EN/students/mobilityprograms/Pages/CreditosECTSnaUMinho.aspx`. accessed 12/13/20.

[9] UCI. Student performance data set. `(http://archive.ics.uci.edu/ml/datasets/Student+Performance`. accessed 12/13/20.

[10] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, pages 207–216, 1993.

[11] Kurt Hornik, Bettina Grün, and Michael Hahsler. arules-a computational environment for mining association rules and frequent item sets. *Journal of Statistical Software*, 14(15):1–25, 2005.

[12] Michael Hahsler. Mining associations with apriori. `(https://www.rdocumentation.org/packages/arules/versions/1.6-6/topics/apriori`. accessed 12/21/20.

[13] Zahid Ansari, MF Azeem, Waseem Ahmed, and A Vinaya Babu. Quantitative evaluation of performance and validity indices for clustering the web navigational sessions. *arXiv preprint arXiv:1507.03340*, 2015.

Appendices

Universidade do Minho

Professor: Paulo Alexandre Ribeiro Cortez

Contrato Data Mining para a Ciência de Dados

1. OBJETO DO CONTRATO

É objeto do presente contrato o seguimento pelos contratados de todas as regras elaboradas no Ponto 2, de forma a obter como resultado final o relatório do projeto "Data Mining para a Ciência de Dados 2020/21" (Projeto DMCD 2020/21).

A falta de seguimento das regras impostas neste contrato, levará a uma reunião de emergência com o intuito de se proceder a uma expulsão do arguido e, como tal, a posterior não realização da unidade curricular.

2. OBRIGAÇÕES DOS CONTRATADOS

(1) Garante-se que não se irão falsificar resultados, nem copiar/plagiar projetos (de outros grupos) ou de conteúdos derivados da Internet, sem que estes sejam devidamente identificados e referenciados (quem é o autor, onde foi publicado) e esta utilização não seja exagerada (face ao restante trabalho desenvolvido);

(2) Garantir presença nas reuniões semanais, acordadas atempadamente nas reuniões prévias;

  (a) O incumprimento de alguma destas regras terá de ser devidamente justificado e, se possível, a reunião será reagendada com a máxima rapidez possível;

(3) Realizar as tarefas / cumprir os objetivos dentro do tempo estipulado;

(4) Participação ativa de cada interveniente em cada reunião e nas respetivas tarefas.

3. VALORES DADO AOS CONTRATADOS

Com o cumprimento das obrigações previstas no Ponto 2, os contratados terão uma avaliação dos restantes intervenientes da forma mais clara e objetiva possível, sendo também importante que cada um reconheça o seu desempenho e o compare o mesmo com o dos restantes membros.

CONTRATADO: André Filipe de Sousa Ferreira

André Filipe de Sousa Ferreira _____ Nº de aluno: A81350

CONTRATADO: Bruno Garcia Carneiro Meira da Fonseca

Bruno Garcia Carneiro Meira da Fonseca _____ Nº de aluno: A83029

CONTRATADO: Daniel Vilaça Costa

Daniel Vilaça da Costa _____ Nº de aluno: A81434

CONTRATADO: Luís Pedro Magalhães da Silva

Luís Pedro Magalhães da Silva _____ Nº de aluno: A80981

Ano letivo: 2020/2021