

ANALYSIS AND PREDICTION ON STUDENT LIFESTYLE EFFECT IN ACADEMIC PERFORMANCE

Magí Izquierdo Lechuga

Department of Computing and Mathematics

School of Informatics and Creative Arts

Dundalk Institute of Technology

Supervised By:

Dr Rajesh Jaiswal

This dissertation is submitted in part fulfillment of the Higher
Diploma in Science in Data Analytics

May 2020

Declaration of Authorship

I hereby certify that this material, which I now submit for assessment on the Higher Diploma in Science in Data Analytics programme of study is entirely my own work, and that I have exercised reasonable care to ensure that the work is original and does not, to the best of my knowledge, breach any law of copyright. In turn, related published work has been cited and acknowledged within the text of my work.

Signed: Magí Izquierdo Lechuga

Date: 25 May 2020

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Dr Rajesh Jaiswal, for all his guidance and support throughout the development of this project. He has helped me to give a novel approach to the presented chosen dataset.

Many thanks to Dr. Fiona Lawless for trusting me since the beginning and all the lecturers of this course, specially Dr. Kevin McDaid, Dr. Siobhan Connolly and Dr. Jack McDonnell for their patience. Although, I am not coming from a Mathematical background and nor a native English speaker, however, I am very satisfied with the work achieved so far.

Many thanks as well to DkIT environment and colleagues of the part-time course. Thanks for all their help and expertise provided during the in-person lectures. Looking forward to being able to put in practise all these new skills in my daily job and move forward in my professional career.

I would like to show my gratitude as well to the Springboard programme for providing the opportunity of a continuous education and economic support.

Cheers to my actual workmates in PayPal, family and friends who has always encouraged me to be a part of this experience and have never let me down over these two years of my new life in Ireland.

Abstract

Predicting individual academic and industrial performance have found use in numerous selection criteria and subsequent appraisals in various academic and commercial institutions. Over a past decade, researchers have successfully used machine learning based data mining algorithms to predict students' performance in academia. However, a lot of research still needs to be done. Through my research, I will attempt to study the existing techniques and aim to reduce redundancy as well as improve the statistical model predicting students' performance.

During 2005-2006, Cortez & Silva (2008) collected information of 788 Portuguese students from two secondary schools to understand the results of the Programme for International Student Assessment in the year 2000 as students in Mathematics and Portuguese Language were underperforming in comparison with the rest of Europe community. I used this dataset for my research. In order to reduce the redundancy within the dataset, the variables affecting most the students' academic performance were chosen to create a predictive model. Thereafter, statistical tools were developed to predict the final marks of the students. This selection of variables were made merely on the basis previous research done. Machine learning based two novel predictive models were implemented incorporating clustering and Principal Component Analysis techniques. The results confirmed that the preceding grades are the key to guarantee a reliable level of prediction's accuracy. In turn, to explore in more detail the importance of the other variables involved, a classification tree machine learning technique was carried out with a reduced dataset by removing the most significant variables (e.g. previous grades). Overall, presented results suggest that developed models allow predicting the final grade of the students with a reliability between 40% to 90% considering the inputted variables. Finally, Shiny student performance app is able to determine future performance based on current and historical data of the students and simplify this predictive process to academic professionals. Recommendations for a future work would be in the direction to perform a new and automated data collection by education centres, universities or companies and include additional crafting subjects that can strengthen predictive models and anticipate new lines of investigation and decision making tools.

Table of Contents

Acknowledgements	3
Abstract	4
List of Figures	6
List of Tables.....	9
1 Project Overview	10
1.1 Research Questions and Objectives	11
2 Literature Review	12
2.1 Education.....	12
2.2 History of Portuguese Education.....	13
2.3 Structure of Education.....	14
2.4 Secondary Education	14
2.5 Employment and Education Attainment	16
2.6 Funding.....	16
2.7 Previous Work.....	16
3 Research Methodology	17
3.1 Data Description and Presentation	18
3.2 Data Import and Cleaning	20
3.3 Statistical Analysis of Portuguese students' dataset.....	20
4 Research Analysis and Findings	23
4.1 Data Exploration and Cleaning	24
4.2 Preliminary Analysis	31
4.3 Student Performance Prediction	46
5 Discussion and Conclusions	60
6 References.....	63
7 Appendices.....	66
7.1 Ethics Documentation	66
7.2 Source Code	71

List of Figures

Figure 1. Overview of the education system. Source: GPS Education OECD.	15
Figure 2. Boxplot of Mathematics Age Outliers treatement	25
Figure 3. Mathematics Boxplot of Absence(A), Scatterplot of Age&Absence(B), Scatterplot of G1&Absence(D), Scatterplot of G2&Absence(C)	25
Figure 4. Boxplot of Mathematics G1(A), G2 (B), G3 (C)	26
Figure 5. Portuguese Language Boxplot of Absence(A), Scatterplot of Age&Absence(B), Scatterplot of G1&Absence(D), Scatterplot of G2&Absence(C)	27
Figure 6. Boxplot of Portuguese Language G1(A), G2 (B), G3 (C).....	27
Figure 7. A. Mathematics students receiving school support; B. Portuguese Lang. students receiving school support; C. Mathematics students receiving family support; D. Portuguese Lang. students receiving family support; E. Mathematics students receiving paid support.....	31
Figure 8. Correlation for Mathematics Continuous Variables	32
Figure 9. Correlation for Portuguese Language Continuous Variables	33
Figure 10. Mathematics Correlation Chart	34
Figure 11. Barplot of: A. Mathematics Studytime; B. Mathematics Past Failures; C. Mathematics Freetime; D.Mathematics Go Out time; E. Mathematics Weekend Alcohol Consumption; F. Mathematics Daily Alcohol Consumption.....	35
Figure 12. Boxplot of Mathematics Grade 3 and Failures	35
Figure 13. Boxplot of Mathematics Grade 3 and Higher Education.....	36
Figure 14. Boxplot of Mathematics Grade 3 and Free Time	36
Figure 15. Boxplot of Mathematics Grade 3 and Go Out Time.....	37
Figure 16. Boxplot of Mathematics Grade 3 and Types of Support	37

List of Figures

Figure 17. Scatterplot of Mathematics Grade 3 and Absences	38
Figure 18. Portuguese Language Correlation Chart.....	39
Figure 19. Barplot of: A. Portuguese Lang. Studytime; B. Portuguese Lang.Past Failures; C. Portuguese Lang.Freetime; D.Portuguese Lang.Go Out time; E. Portuguese Lang.Weekend Alcohol Consumption; F. Portuguese Lang.Daily Alcohol Consumption.	40
Figure 20. Boxplot of Portuguese Lang. Grade 3 and Study Time.....	40
Figure 21. Boxplot of Portuguese Lang. Grade 3 and Previous Failures.....	41
Figure 22. Boxplot of Portuguese Lang. Grade 3 and Higher Education	41
Figure 23. Boxplot of Portuguese Lang. Grade 3 and Daily Alcohol Consumption	42
Figure 24. Scatterplot of Portuguese Lang. Grade 3 and Absences.....	42
Figure 25. Boxplot of G3 and Gender for: A.Mathematics; B. Portuguese Language ...	43
Figure 26. Boxplot of G3 and Age for: A.Mathematics; B. Portuguese Language	44
Figure 27. Boxplot of G3 and Study Time for: A.Mathematics; B. Portuguese Language	44
Figure 28. Boxplot of G3 and Previous Failures for: A.Mathematics; B. Portuguese Language	45
Figure 29. Residuals vs Fitted and Normal QQ plot for Mathematics Linear Model.....	47
Figure 30. Regression Tree of G3.x in Mathematics Regression Analysis	47
Figure 31. Classification Tree of G3.x in Mathematics Binary Analysis	48
Figure 32. Classification Tree of G3.x in Mathematics 4 Level Analysis	49
Figure 33. Classification Tree of G3.x in Mathematics 4 Level Analysis without G2.x and G1.x	50
Figure 34. Dendrogram of Mathematics Students.....	50
Figure 35. Scree plot of Mathematics PCA	51

List of Figures

Figure 36. Mathematics PC1 and PC2 plot of weighted averages per student	52
Figure 37. Residuals vs Fitted and Normal QQ plot for Portuguese Language Linear Model	53
Figure 38. Regression Tree of G3.y in Portuguese Language Regression Analysis	53
Figure 39. Classification Tree of G3.y in Portuguese Language Binary Analysis	54
Figure 40. Classification Tree of G3.y in Portuguese Language 4 Level Analysis	55
Figure 41. Classification Tree of G3.y in Portuguese Language 4 Level Analysis without G2.y and G1.y	55
Figure 42. Dendrogram of Portuguese Language Students	56
Figure 43. Scree plot of Portuguese Language PCA.....	57
Figure 44. Portuguese Language PC1 and PC2 plot of weighted averages per student .	57
Figure 45. ShinyApp Introduction landing page.....	58
Figure 46. ShinyApp Portuguese Language input tab	59
Figure 47. ShinyApp Portuguese Language input tab with outcome	59

List of Tables

Table 1. Organisation of Studies , Source: Office of the Secretary-General of the European Schools (https://www.eursec.eu/en/European-Schools/studies/studies-organisation).....	13
Table 2. Portuguese Students Dataset, 33 variables and their attributes. Source: (Cortez & Silva, 2008)	19
Table 3.Summary statistics of Student background variables.	28
Table 4.Summary statistics of Student Mathematics Subject variables.....	29
Table 5. Summary statistics of Student Portuguese Language Subject variables.....	30
Table 6.Variables selected from the original Students dataset and used to create a working dataset	31
Table 7. Root Mean Squared Error Results for Caret package	60
Table 8. Accuracy Results for Machine Learning Techniques	61

1 Project Overview

Portugal 2000 PISA (Programme for International Student Assessment) results highlighted the education deficiencies of the country and identified a high rate of scholar abandonment without completing the compulsory education (Artlet, 2003). The report identified students with disadvantaged backgrounds frequently repeating grades on top of crowded schools in the major cities and small and isolated schools in rural areas due to the political instability of the country, heavily affected by the beginning of the wars of the century.

From the report mentioned above, researchers began an investigation about the causes of this educative deficit, mostly for Mathematics and Portuguese Language which were the most affected core subjects. In 2005, using questionnaires and school reports, Paulo Cortez and Alice Silva created a dataset and analysed it using Data Mining algorithms to predict the students grading using 32 explanatory variables from demographic, academic, family and lifestyle information of the students (Cortez & Silva, 2008). This research was done alongside with the development of the nowadays known as Educational Data Mining (EDM) research. Using data mining, machine learning and statistics techniques helped to resolve the research questions of that project.

Alternatively, to the approach explained above, reducing redundancy with only lifestyle and academic variables from Cortez & Silva (2008) dataset and using different predictive techniques, the intention of this new project is to develop a predictive algorithm for the grades of the Portuguese students and try to find a predictive pattern which, with the planned work, could probably be adjusted to other communities or used by public and private institutions.

This work is outlined in seven sections. Section 1 provides the study background and project research questions and objectives. A detailed literature review is presented in section 2, where Portuguese history is introduced with a special focus on education and its evolution until the time of Cortes and Silva (2008) findings. This is followed by the research methodology where data description and cleaning are provided. In addition, section 3 provides a detailed explanation of statistical analyses carried out to achieve the desired results. Section 4 introduces the research analyses and key findings resulting from the application of the techniques described in previous section 3. Finally, results

and future work are detailed in chapter 5. Ethics documents, and source code in R can be found in following sections.

The project Mahara page is available to follow up the details of the project, progress done over time and supervised meetings (Izquierdo, 2020). It is accessible in the following link: <https://mahara.dkit.ie/view/view.php?t=4oMvTpDeuHfWPnRqs6hJ>.

1.1 Research Questions and Objectives

There are several factors that can affect the students' academic performance including personal/psychological, geographical, socio-economic, environmental, and political factors. For this reason, there is the need to study and do a thorough research of these factors to be able to project students' performance. Therefore, the aim of this study is to be able to predict the final student marks from two Portuguese secondary schools and how these marks, as a direct measurement of the student performance, can be influenced by different factors of the student academic environment.

Specifically, this research will help to better understand the factors that affect the students' performance in Portuguese Language and Mathematics grades. Taking this into account, specific objectives of this study are:

- *Identify predictive academic and lifestyle variables that affect the performance of the students.*
- *Investigate which would be the best model or classification for prediction on student's final grades.*
- *Predict the student's grades based on the model/s developed.*

In addition, I aim to identifying the factors that influence the performance of students in final examinations and create a suitable data mining algorithm that can predict the grade of students. This tool will help targeting and be more aware of those students who are at risk to fail or with possibilities to improve their grades in a timely effective way.

2 Literature Review

First, and to understand the politic, socioeconomic and academic background of Portugal, a literature review of the main historical factors that have conditioned Portuguese students is presented. Next, the present status of the secondary education in Portugal is overviewed. Finally, an introduction to the previous work and research done in the field of academic prediction is reviewed.

2.1 Education

Education is the basis of the transmission of our heritage, generation after generation, which helps the global community to maintain a general knowledge and advance in its enrichment that benefits the existence of the human being. It helps to develop individual personality by making knowledgeable, competent, capable and skilful. It plays a vital role in the development of human capital and is linked with an individual's well-being and opportunities for better living (Eshetu, 2015).

In Europe this transmission of knowledge is being performed in schools which are official educational establishments controlled jointly by the governments of the Member States of the European Union (Eursec.eu., 2020), divided in different levels limited by groups of age (Table 1).

Secondary education is the final step before entering in the labour world or change to the tertiary education level. It is as well the rank of ages starting at 11 years old and finishing at 18 years old where the personality of the individual is growing, developing and shaping to become singular and be prepared for the future. From 11 to 18 years old, besides experiencing natural maturity, teenagers' degree of attention and learning skills can be affected in many ways. For example, daily lifestyle or region of residence can influence to the educational development of each respective student. For this reason, regular assessments with a criterion established by the Board of Governors are being performed to decide if the knowledge requirements are met to move up to the year above at the end of the school year (Eursec.eu., 2020).

2 Literature Review

Table 1. Organisation of Studies , Source: Office of the Secretary-General of the European Schools (<https://www.eursc.eu/en/European-Schools/studies/studies-organisation>)

Cycle	Classes	Age
'Early education' (Nursery)	1-2	4 and 5
Primary	1-5	6-10
Secondary		
Observation cycle	1-3	11-13
Pre-orientation cycle	4-5	14-15
Orientation cycle	6-7	16-18

2.2 History of Portuguese Education

Portuguese educational system has been negatively affected since 20th century. The French invasion, the dictatorial regime of 1926, the participation in the 1st World War and the neutralism in the 2nd, the independence of the Portuguese colonies, the antiparliamentary “Stado Nuovo” of 1933, and the continuous political and socio-economic instability. Portugal has suffered the collateral effects of all those events which culminated in 1974 by a military young movement that started the democratic regime and which developed into the first Constitutional Government in 1976, the same that in 1986 joined the European Economic Community. Up until that moment, the only educational improvement done was the extension of the compulsory education from 3 to 4 years and its extension on 1964, where the 4 years’ primary school were followed by 2 years of upper secondary preparatory cycle with an illiteracy rate of the 25% of Portuguese population (EURYDICE, 2020).

However, the educational reform was delayed due to a demographic increase and an unstructured political and economic system that were unable to adapt the education system until 1986 with the Education Act of the *Lei de Bases do Sistema Educativo*. This system was divided in basic and secondary education which, at the same time, offered 5 different paths and could be followed by a 2-year complimentary course with vocational training in the chosen area (EURYDICE, 2020).

During the 80’s decade, the 3 year technical courses appeared, additionally to the complimentary course and focused on students that wanted to enter the labour market and basic education was extended until the 15 years of age. From the 80’s until the 90’s, the number of students showed an increase from 130.000 to 500.000, respectively, and

followed with the extension of the secondary school until 18 years of age (EURYDICE, 2020).

Arriving to the XXI century, the change was visible from the 1% of residents aged 20 or older having a level of education equivalent to higher education to the 14.8% in 2011. From two thirds of residents aged over 15 years with no formal education to less than one third for men and from three quarters to almost one third for women. Nowadays, only the decrease in the birth rate is affecting the enrolment to education, where almost the 100% of younger people has access to it (EURYDICE, 2020).

2.3 Structure of Education

Since 1976 there is an International Standard Classification of Education which shows the actual education levels named ISCED and established by the UNESCO from pre-primary school to higher education levels in the Organisation for European Economic Co-operation (OEEC), nowadays known as the Convention on the Organisation for Economic Co-operation and Development (OECD). Marked in green as ISCED level 3 is the secondary school level before accessing any specialization, polytechnic course or Bachelor (Figure 1).

2.4 Secondary Education

Following Law no. 85/2009 of the 27th August of 2009, upper secondary education has become universal, free and compulsory for all students whose age is up to 18. It corresponds to three years of schooling (grades 10, 11 and 12) ISCED level 3 (science-humanities courses), or level 4 when completed via a vocational education course. The way to access to it is to successfully complete the basic education (República, 2009).

Although graduation from upper secondary education increased by 6% between 2005 and 2017, 15% of 25-34-year-olds did not attain upper secondary education in 2018, on average across OECD countries. Moreover, from those who did the average age on completion were 19.8 years old, 1.8 years later than expected. The proportion of 25-64-year-old men who have attained a vocational degree at the upper secondary or post-secondary level is one of the lowest among OECD and partner countries with 7.5 % men and 6.6% women in 2018 (OECD, 2019).

Education and training in upper secondary education aims to provide students with diverse training and learning that matches their interests, recognising that everyone has

2 Literature Review

capacity and can choose any educational and training provision available, with a view to continuing studies and/or working (EURYDICE, 2020).

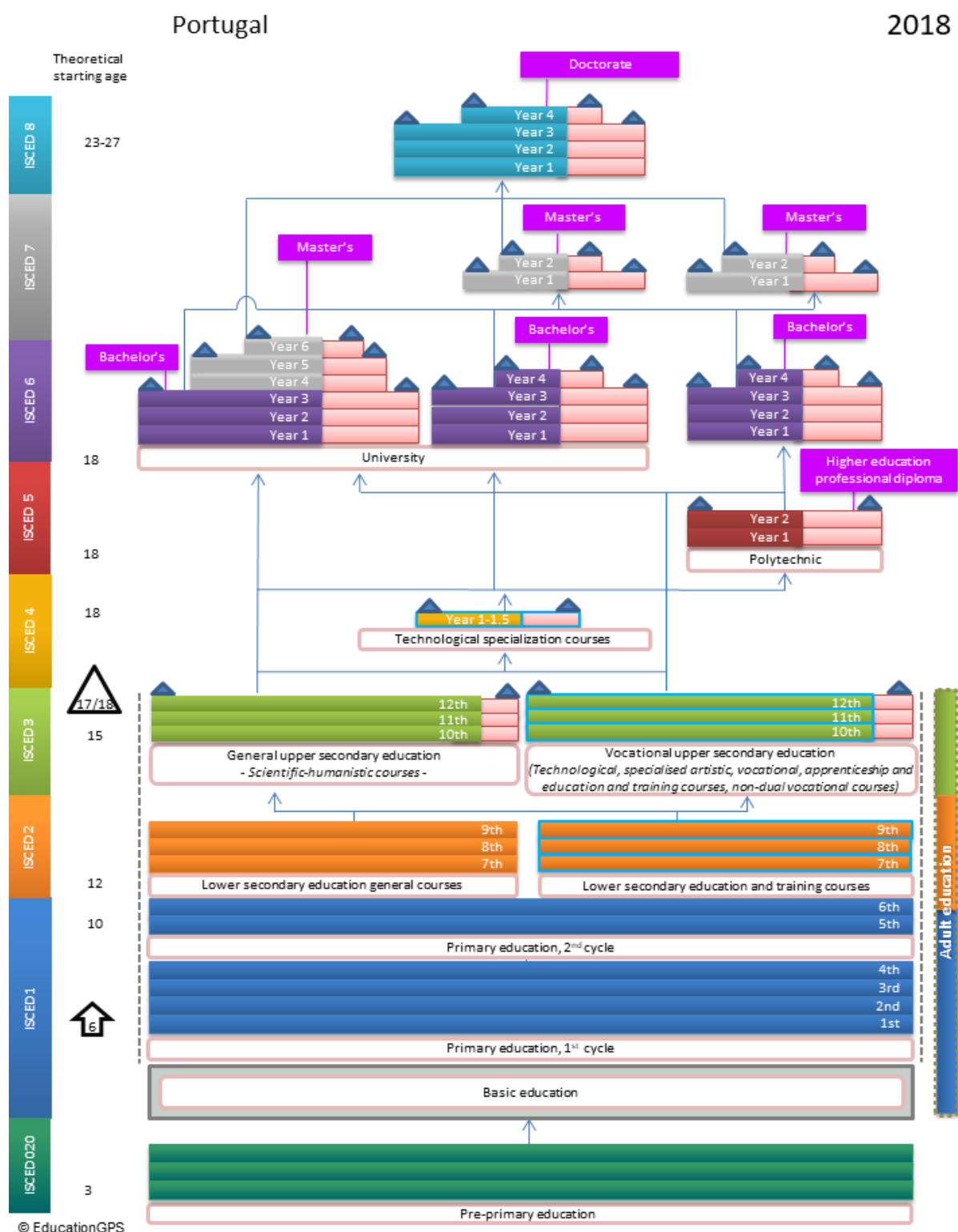


Figure 1. Overview of the education system. Source: GPS Education OECD.

2.5 Employment and Education Attainment

The employment rate among Portuguese citizens above 25 years old with a general upper secondary education is relatively high compared to other OECD and partner countries with the 82.9% in 2018 and with an inactivity rate decrease of 12.5% for adults with below upper secondary education, 9.1% for those with upper secondary education and 7.8% for adults with tertiary education (OECD, 2019).

2.6 Funding

Portuguese Education is funded by the Ministry of Education and the European commission with a running cost of 4,761,084,482€ and an investment of 126,717,381€ in the pre-school, basic and upper secondary education for 2019 (EURYDICE, 2020).

2.7 Previous Work

Each country has had a different history and variables that could affect the behaviour of the students. Researchers from different countries have been interested in studying the success of students over the last years. Most of the researchers like Eshetu and Amonge have done the analysis focused on one area of study such as parental socio-economic status. In turn, with the help of statistical inferences they were able to identify some of the factors that had large impacts on the students (Eshetu, 2015). This kind of descriptive analysis is good to understand the actual effects of the variables over the students using diverse techniques including cross tabulation, percentage, independent samples t-tests, Spearman's rho correlation and Analysis of Variance (ANOVAs).

Further work was done on the same direction but with a different focus by other researchers Paulo Cortez and Alice Silva with the intention of predicting the student's grades using Business Intelligence and Data Mining tools (Cortez & Silva, 2008). They were able to conclude that the background of the student is important, but the most relevant variables were the past marks, achieving an accuracy rate of 93% on a binary classification exercise and a 78.5% in a 5 level classification.

These techniques had been already used and named Educational Data Mining (EDM) when applied to educational purposes as it aids researchers gain a better insight in the mind of students, their learning processes, and also helps comprehend available datasets better (Ventura, 2010), (Filiz, 2019).

This literature review is unveiling an overview of factors that affect the Portuguese students since the beginning of the 20th century and that have an origin at daily lifestyle, family, sociocultural, socioeconomic and political backgrounds. It is as well presenting some of the most frequently used statistical models in the Educational Data Mining field and the techniques that have been developed during that period to better predict students' performance and that will be partly used in this project for the same purpose.

3 Research Methodology

Data generation is rapidly growing and being accumulated at a pace that cannot be followed (Fayyad, 1996). This accumulation of data is happening as well in the field of the education. For this purpose, data mining techniques are as well required in educational processes.

The structure for the analysis and prediction is started by the knowledge and understanding of the data with descriptive statistics which helps to better understand it and apply the most correct further analyses. In order to develop predictive analyses, models need to be created. For this specific case of study, a student model structure would represent the information about student's knowledge, motivation, meta-cognition and attitudes (Baker, 2009).

Predictive Machine Learning techniques (ML) are used to analyse, evaluate and predict the students' performance using classification techniques such as Naïve Bayes (Ramesh, 2013), decision trees (Pandey, 2013), K-means clustering (Varghese, 2011) and, for time series data, the Hidden Markov Model (Tadayon, 2019).

The data used in this project comes from two datasets composed by Paulo Cortez and Alice Silva from the University of Minho on Portuguese students and their performance in Mathematics (395 observations) and Portuguese Language (649 observations) courses during the 2005-06 academic year from the two public schools Gabriel Pereira and Mousinho da Silveira (Cortez & Silva, 2008).

The source of this data is the UCI Machine Learning public repository with a total of 382 students and 33 variables extracted from school reports and questionnaires focused in demographic, academic, family and lifestyle themes. The validity of the gathering of the data was reviewed by school professionals and tested on a small set of 15 students to get feedback. Further work exists from the authors of the dataset in the same field of

investigation of this project in order to be able to predict the students' performance using Data Mining (Cortez & Silva, 2008).

This dataset will be used to develop a predictive tool, being aware of the obsolescence of the chosen dataset and potential changes that Portuguese community could have experienced from 2006 to 2020. Any application of data prediction will be necessary to be reviewed and adapted for the best application and validity of the results.

3.1 Data Description and Presentation

Portuguese student's dataset is downloadable in a .csv format and in two different datasets that comprises the data of Mathematics and Portuguese Language subjects being two of the most unsuccessful core subjects as highlighted in the PISA report of year 2000. Just recently the grades have increased dramatically and fit in a middle classification with the rest of Europe results as the OECD confirmed in the latest publication (Schleicher, 2018).

The schools being analysed are Gabriel Pereira from the area of Évora and Mouzinho da Silveira from Portalegre. Both Mathematics and Portuguese Language are core subjects in the two schools and have the same relevance in the academic year for the two schools respectively. There is a total of 395 observations for Maths and 649 for Portuguese Language. The observations are divided by 33 variables (Table 2).

The structure of the dataset has been modelled to be binary (yes or no) and a maximum of five-level classification with five continuous numerical variables which are Age (from 15 to 21 years old), Absences (from 0 to 93), G1, G2 and G3. Those last ones referring to the final grades of each evaluation period and being G3 the final grade, with a range from 0 to 20 being 0 the minimum value and 20 the maximum. A minimum score of 10 points is required in order to pass the subject.

3 Research Methodology

Table 2. Portuguese Students Dataset, 33 variables and their attributes. Source: (Cortez & Silva, 2008)

<i>Attribute</i>	<i>Description (Domain)</i>
<i>sex</i>	student's sex (binary: female or male)
<i>age</i>	student's age (numeric: from 15 to 22)
<i>school</i>	student's school (binary: Gabriel Pereira or Mouzinho da Silveira)
<i>address</i>	student's home address type (binary: urban or rural)
<i>Pstatus</i>	parent's cohabitation status (binary: living together or apart)
<i>Medu</i>	mother's education (numeric: from 0 to 4), 0 – none, 1 – primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education.
<i>Mjob</i>	mother's job (nominal), teacher, health care related, civil services (e.g. administrative or police), at home or other.
<i>Fedu</i>	father's education (numeric: from 0 to 4), 0 – none, 1 – primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education.
<i>Fjob</i>	father's job (nominal), teacher, health care related, civil services (e.g. administrative or police), at home or other.
<i>guardian</i>	student's guardian (nominal: mother, father or other)
<i>famsize</i>	family size (binary: = 3 or > 3)
<i>famrel</i>	quality of family relationships (numeric: from 1 – very bad to 5 – excellent)
<i>reason</i>	reason to choose this school (nominal: close to home, school reputation, course preference or other)
<i>traveltime</i>	home to school travel time (numeric: 1 – < 15 min., 2 – 15 to 30 min., 3 – 30 min. to 1 hour or 4 – > 1 hour).
<i>studytime</i>	weekly study time (numeric: 1 – < 2 hours, 2 – 2 to 5 hours, 3 – 5 to 10 hours or 4 – > 10 hours)
<i>failures</i>	number of past class failures (numeric: n if 1 = n < 3, else 4)
<i>schoolsup</i>	extra educational school support (binary: yes or no)
<i>famsup</i>	family educational support (binary: yes or no)
<i>activities</i>	extra-curricular activities (binary: yes or no)
<i>paidclass</i>	extra paid classes (binary: yes or no)
<i>internet</i>	Internet access at home (binary: yes or no)
<i>nursery</i>	attended nursery school (binary: yes or no)
<i>higher</i>	wants to take higher education (binary: yes or no)
<i>romantic</i>	with a romantic relationship (binary: yes or no)
<i>freetime</i>	free time after school (numeric: from 1 – very low to 5 – very high)
<i>goout</i>	going out with friends (numeric: from 1 – very low to 5 – very high)
<i>Walc</i>	weekend alcohol consumption (numeric: from 1 – very low to 5 – very high)
<i>Dalc</i>	workday alcohol consumption (numeric: from 1 – very low to 5 – very high)
<i>health</i>	current health status (numeric: from 1 – very bad to 5 – very good)
<i>absences</i>	number of school absences (numeric: from 0 to 93)
<i>G1</i>	first period grade (numeric: from 0 to 20)
<i>G2</i>	second period grade (numeric: from 0 to 20)
<i>G3</i>	final grade (numeric: from 0 to 20)

3.2 Data Import and Cleaning

In order to acquire the dataset and load it into R, it has been downloaded from the UCI Machine Learning repository (<https://archive.ics.uci.edu/ml/datasets/student+performance>) in .rar, decompressed and saved as 2 files .csv (student-mat and student-por). Additional .txt file is available in the decompressed folder with information about the variables and a mention about 382 students belonging to the two datasets that can be identified by matching attributes (13 variables).

There is also an R file with the import of the data and the identification of the 382 students. This final R file with 53 variables will be the one used to develop any future descriptive and predictive analysis where the first 13 variables are the matching attributes and the following 40 variables are 20 for each subject being Mathematics the ones ending in “.x” and Portuguese Language the ones ending in “.y”.

On a further step, missing values or NA’s and any possible outliers have been checked and have found that there are 0 NA. Respective outliers have been presented. Regarding outliers, for continuous data the age outliers for students older than 18 years old have been removed and outliers in grades and absences have been kept as plausible.

3.3 Statistical Analysis of Portuguese students’ dataset

3.3.1 Multiple Linear Regression

For the creation of a predictive model the `lm()` function in R has been used to build a multiple linear regression model using a backwards selection for the response variable G3 and the matrix of all explanatory variables used as an input (SJ, 2009). The adjusted R-squared obtained from the model has been used, alongside with the p-value, to compare different statistically significant models while removing variables. This metric explains the correspondence between the observed and modelled data, with a range from 0 to 1, 1 indicating the perfect model. In addition, the `anova()` function has been also used to test for differences between models, in combination with the search of the lowest AIC.

The `predict()` function requires of a predictive model and a new matrix of data from which the response variable will be predicted. For the validation of the results, a confusion matrix was created with the real values of the test and the predicted data in order to match the results and acquire a percentage of accuracy.

3.3.2 Caret Package

The Classification And REgression Training (caret) package in R has been used to be able to simplify different ML predictive techniques, divided into classification and regression types (Kuhn, 2008). For the implementation of the different models data needs a pre-treatment being divided by training and test datasets. Some modelling techniques require centralising and scaling the data such as k-NN (clustering) and SVM (Support Vector Machine). Additionally, to give a more robust estimate, model is cross validated on 10 cross folds 3 times where a subset of the data is used to train the model which is evaluated with the remaining 9 subsets (Brownlee, 2018).

For the classification techniques, the test metric used is the accuracy which is the correct predicted values divided by total observations of the model and it is showed as a mean result from 0 to 1, being 1 the greatest accuracy. On the other hand, for the regression techniques the Root Mean Squared Error (RMSE) is the metric to compare models, being the lowest the better. The regression approach will be complemented with the confusion matrix of the results and test dataset to achieve an accuracy outcome in a form of percentage (Kubat, 2017).

The modelling techniques used in the caret package with the train() function were: a) the linear methods Linear Discrimination Analysis (LDA) and Logistic Regression(GLM); b) non-linear methods Neural Networks (NN), Support Vector Machine (SVM), K-Nearest Neighbours (KNN) and Naïve Bayes (NB); and c) decision trees methods Classification and Regression Trees (CART), Bagging (TREEBAG), Random Forest (RF) and Boosting (GBM) (Kuhn, 2008).

The best accuracy results for the specific case of study were obtained with the detailed decision trees methods. For this reason, only these specific ones will be explained in more detail below:

- Classification and Regression Trees (CART): is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences. It is one way to display an algorithm that only contains conditional control statements. It is a supervised learning method that can be used to classify data. We have focused in the leaves created by the splits of the tree and the weight of the variables forcing that split, as well as, the final nodes which provide the final

classification. This method works for categorical and continuous data as well as non-linear approach (Soraya Sedkaoui, 2020).

- Bagging (TREEBAG): It is done to reduce the variance in the decision tree, and it uses bootstrapping to do so. The same algorithm is used on a number of subsets of the training set. Each model built from the different subsets is used to predict the response and the final classification is given by the outcome that was predicted by most models. With regression models the prediction is the mean of the predictions (James, 2013).
- Random Forest (RF): It uses bootstrapping as well, but it randomly selects a subset of the variables at each sample tree that usually is the size of the square root of the number of predictors (James, 2013).
- Boosting (GBM): Is the weighted sum of the predictions made by sequentially previous trees. Boosting does not involve bootstrap sampling; instead each tree is fit on a modified version of the original data set (James, 2013).

All these techniques were compared by the resulting mean of the accuracy and, additionally, the best performers were used in a confusion matrix with the predictive model and the test set to perform a second check on accuracy. Once the best performer was found, a posteriori analysis of variable importance was done and a new dataset with only the heaviest academic variables and a new predictive model was saved.

In general, the model that gives best results among the rest is the Random Forest, which is not a single decision tree, but an ensemble of a larger number of trees. Plotting the final model shows the error rate on the test and training datasets as the number of trees are increased. Therefore, for a better visualization of the behaviour of the classification algorithm, the plot provided for the analysis of the classification is a classification tree.

3.3.3 Clustering and PCA

On a further step and using classification techniques, clustering has been used with the K-means and Hierarchical clustering as a novel approach.

- Hierarchical Clustering: It has been applied in an agglomerative form where small clusters merge with each other repeatedly until they form a dendrogram. It has been used with the complete link which is the maximum distance between two clusters (Defays, 1977).

- K-means: A specific number of clusters is defined and are selected randomly by the algorithm as the initial cluster centres. It is based on the Euclidean distance, it assigns each observation a centroid and it precomputes the cluster centroid with all data points in the cluster. This action is repeated until the centroids do not change. The resulting information in R is the number of data points in each cluster and the mean for each variable of the cluster above or below 0, being 0 the less extreme value (Davidson, 2002).

Having a multivariate dataset has provided the opportunity to apply the Principal Component Analysis (PCA) to reduce the dimensionality of the data and identify some patterns. It summarizes the variance in a multivariate scatter of points, providing an overview of linear relationships between the variables (Kassambara, 2017). The function `prcomp()` collapses the variables into Principal Components (PC's) which are analysed as a single weighted mean of the student results (variables) and it can be plotted in a scatterplot (Kuhn, 2008). The importance of the PC's is usually required to be at the 80-90% of the variance explained which ideally will be in the first PC's to make the representation in a 2 dimensional plot reliable. The results of a PCA can as well be analysed by the variable mean in each PC which will determine the weighted mean in the scatterplot. In order to visualize the loadings of each variable, a directional arrow with the name of the variable has been added. Additionally, the variable G3 has been used to identify the final grades of the students and its distribution in the plot.

4 Research Analysis and Findings

In continuation with the aim of the project to predict the final grades of the students from academic and lifestyle background information, a deep scan of the data will be performed to seek any possible NA's and outliers. Once the dataset is ready for the preliminary analysis, this will be performed to find any details, characteristics and relationship between variables and subjects. Any finding in this chapter will be decisive on the decision taken for the further statistical and predictive analysis. Hopefully interesting results will be available in the conclusions and key findings at the end of this chapter.

4.1 Data Exploration and Cleaning

During the importation of the dataset, the two .csv files have been merged in order to obtain one single dataset with 382 unique students that share information in both datasets. After cleaning, the number of students has been reduced to 369 due the age outliers. This has generated a first part of the dataset which is the identification attributes (section 4.1.2), a second part with Mathematics subject information (section 4.1.3) and a third part with Portuguese Language information (section 4.1.4). The summary statistics of these three different groups of variables will be explored in the following sections.

4.1.1 Outliers and NA's

Variable “Age” has 3 outliers that fit in the range of the data from 15 to 22. Secondary education does not go further 18 years old. Therefore, these outliers will be deleted to normalize the data (Figure 2).

In a different way, for variable “Absence”, any data point higher than 20 are considered outliers. The reason of high absence is not disclosed but will treat data as plausible (Figure 3 A). When plotted together with age there are some students missing an entire semester starting on 16 years old and with the higher absence at the 18 years old (Figure 3 B). For G2, outliers have been identified, probably linked to students that didn't show up to the exam or failed with a 0. This can be visible when plotting Absence and G2 together where there are students with grade 0 and 0 absences but have higher marks in G1 (Figure 3 C & D). The most extreme value of absence is above 60 which it is almost a scholar trimester. It is relevant that G1 do not have any outlier when compared with G2 and G3 (Figure 4**Error! Reference source not found.**).

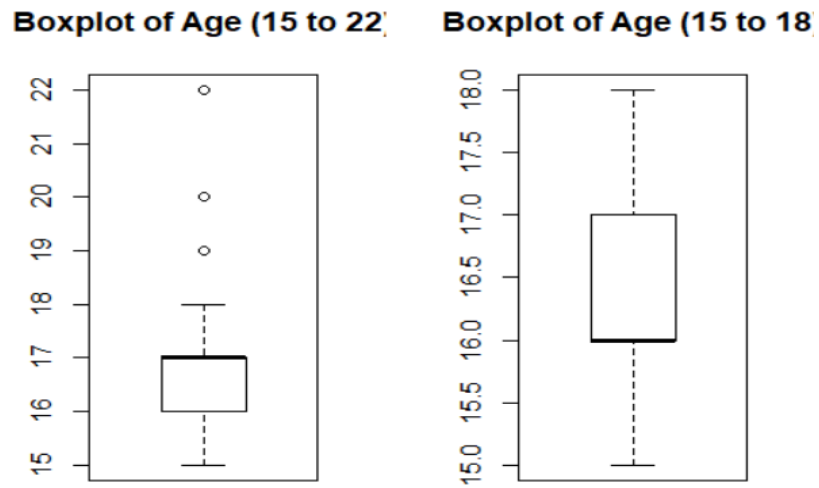


Figure 2. Boxplot of Mathematics Age Outliers treatment

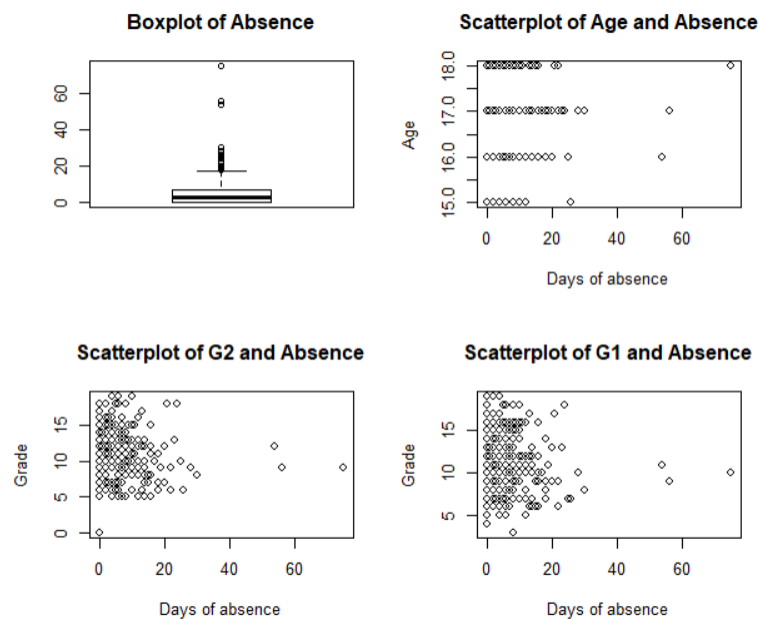


Figure 3. Mathematics Boxplot of Absence(A), Scatterplot of Age&Absence(B), Scatterplot of G1&Absence(D), Scatterplot of G2&Absence(C)

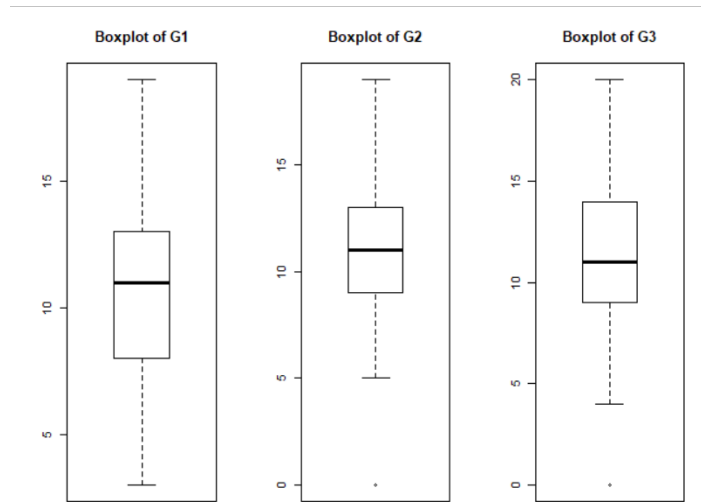


Figure 4. Boxplot of Mathematics G1(A), G2 (B), G3 (C)

In the Portuguese Language dataset, for “Absence” variable any data point higher than 15 is considered an outlier and will be treated as plausible (Figure 5 A). The highest outliers for absence are at the 17 years old but still are slightly higher than 30, much lower than in Mathematics (Figure 5 B). Portuguese Language seems to be a subject with a lower number of absence and a few more outliers in the final grades more specifically during the 3rd grading where there are three lower extreme values and one higher maintained from G2 (Figure 5 C&D, Figure 5 B&C). There are as well constant low values maintained from G1 to G3 (Figure 6Error! Reference source not found. A,B&C). After checking the data and acknowledging all the possible outliers will proceed to the exploration of it using summary statistics and visualization.

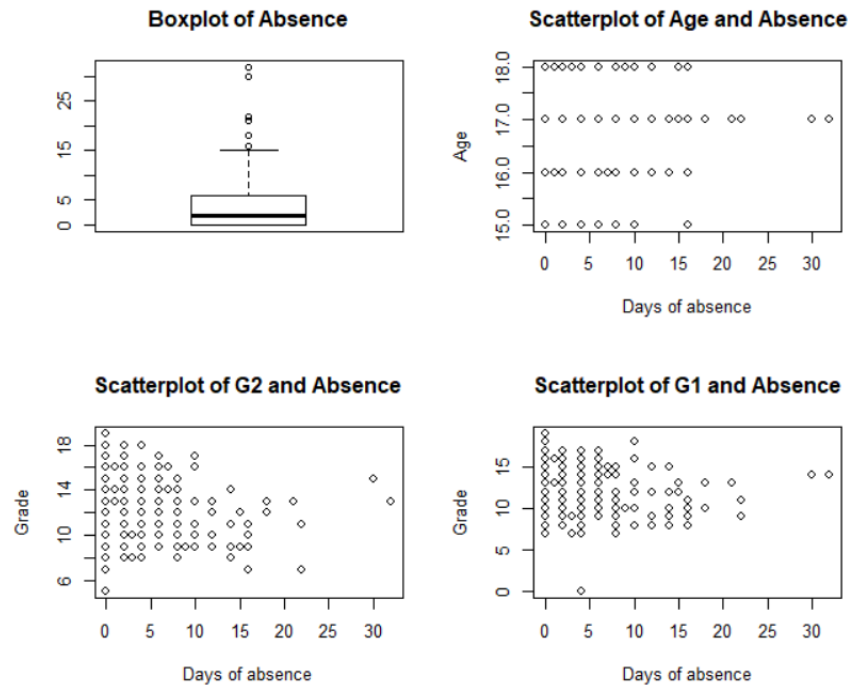


Figure 5. Portuguese Language Boxplot of Absence(A), Scatterplot of Age&Absence(B), Scatterplot of G1&Absence(D), Scatterplot of G2&Absence(C)

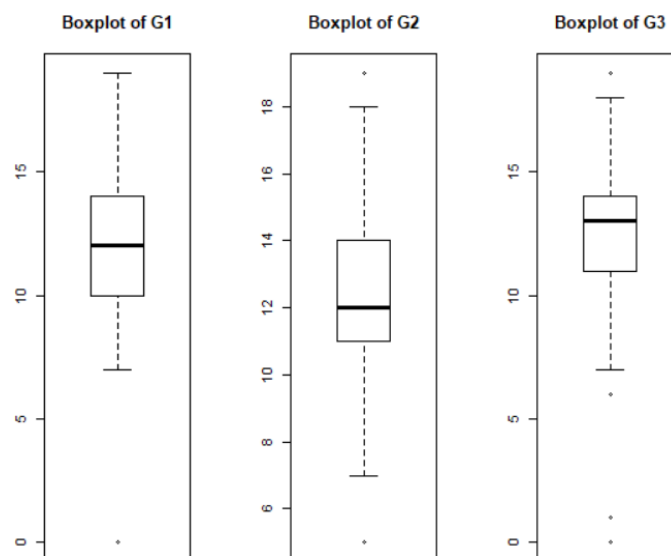


Figure 6. Boxplot of Portuguese Language G1(A), G2 (B), G3 (C)

4.1.2 Student Identification Attributes

At a first sight a basic profile of the most common students can be extracted from the data. The 90.2% of the students are from Gabriel Pereira School with a balanced distribution between girls and boys, a median age of 16 years' old and from it a 79.4% of them are from urban residence instead of rural, with an 85.4% of them with access to internet as detailed in Table 3.

When analysing the family status, 90% of families are married with 72.1% greater than 3 members. Parents are graduated from a secondary level with some tertiary education and with the majority of them working at services and other jobs. There is a substantial difference between women and men where women are 3 times more employed at home, 2 times more at health and 2 times more as teachers, in contrast to men, who are mainly employed in other jobs and services as demonstrated in Table 3.

Table 3. Summary statistics of Student background variables.

<i>school</i>		<i>sex</i>		<i>age</i>		<i>address</i>		<i>famsize</i>		<i>Pstatus</i>		<i>Medu</i>		<i>Fedu</i>	
<i>GP</i>	333	<i>F</i>	193	<i>Min.</i>	15	<i>R</i>	76	<i>GT3</i>	266	<i>A</i>	36	<i>Min.</i>	0	<i>Min.</i>	0
<i>MS</i>	36	<i>M</i>	176	<i>1st</i>	16	<i>U</i>	293	<i>LE3</i>	103	<i>T</i>	333	<i>1st</i>	2	<i>1st</i>	2
				<i>Median</i>	16							<i>Media</i>	3	<i>Media</i>	3
				<i>Mean</i>	16.5							<i>Mean</i>	2.82	<i>Mean</i>	2.58
				<i>3rd</i>	17							<i>3rd</i>	4	<i>3rd</i>	4
				<i>Max.</i>	18							<i>Max.</i>	4	<i>Max.</i>	4

<i>Mjob</i>		<i>Fjob</i>		<i>reason</i>		<i>nursery</i>		<i>internet</i>	
<i>at_home</i>	52	<i>at_home</i>	14	<i>course</i>	137	<i>no</i>	69	<i>no</i>	54
<i>health</i>	32	<i>health</i>	17	<i>home</i>	106	<i>ye</i>	300	<i>yes</i>	315
<i>other</i>	132	<i>other</i>	205	<i>other</i>	32				
<i>services</i>	91	<i>services</i>	102	<i>reputat</i>	94				
<i>teacher</i>	62	<i>teacher</i>	31						

4.1.3 Mathematics Summary Statistics

Mathematics is characteristic for being a subject with a mid-study time required and with the 8.4% of students being failed at least one previous class. The grades have a decrease tendency from 1st period to 3rd and the absence rate is between 0 and 7 days per student. From those students, 41.2% are receiving some kind of support, mostly from family or paid. There seem to be a relationship between paid classes and activities as it is exactly the opposite number of students. From them, 96% of them want to access the higher education. The 52.8% of them do some kind of activity and the 32% enjoy from a relationship while they enjoy from a mid-high amount of free time and time to go out. The consumption of alcohol during the weekend slightly increases but in general the health status of the students is good (Table 4).

4 Research Analysis and Findings

Table 4. Summary statistics of Student Mathematics Subject variables.

<i>guardian.x</i>		<i>traveltime.x</i>		<i>studytime.x</i>		<i>failures.x</i>		<i>schoolsup.x</i>		<i>famsup.x</i>		<i>paid.x</i>	
<i>father</i>	90	<i>Min.</i>	1	<i>Min.</i>	1	<i>Min.</i>	0	<i>no</i>	318	<i>no</i>	138	<i>no</i>	195
<i>mother</i>	270	<i>1st</i>	1	<i>1st</i>	1	<i>1st</i>	0	<i>yes</i>	51	<i>yes</i>	231	<i>yes</i>	174
<i>other</i>	9	<i>Median</i>	1	<i>Median</i>	2	<i>Median</i>	0						
		<i>Mean</i>	1.44	<i>Mean</i>	2	<i>Mean</i>	0.252						
		<i>3rd</i>	2	<i>3rd</i>	2	<i>3rd</i>	0						
		<i>Max.</i>	4	<i>Max.</i>	4	<i>Max.</i>	3						

<i>activities.x</i>		<i>higher.x</i>		<i>romantic.x</i>		<i>famrel.x</i>		<i>freetime.x</i>		<i>goout.x</i>		<i>Dalc.x</i>	
<i>no</i>	174	<i>no</i>	15	<i>no</i>	251	<i>Min.</i>	1	<i>Min.</i>	1	<i>Min.</i>	1	<i>Min.</i>	1
<i>yes</i>	195	<i>yes</i>	354	<i>yes</i>	118	<i>1st</i>	4	<i>1st</i>	3	<i>1st</i>	2	<i>1st</i>	1
						<i>Median</i>	4	<i>Median</i>	3	<i>Median</i>	3	<i>Med</i>	1
						<i>Mean</i>	3.935	<i>Mean</i>	3.2	<i>Mean</i>	3.108	<i>Mea</i>	1.5
						<i>3rd</i>	5	<i>3rd</i>	4	<i>3rd</i>	4	<i>3rd</i>	2
						<i>Max.</i>	5	<i>Max.</i>	5	<i>Max.</i>	5	<i>Max.</i>	5.000

<i>Walc.x</i>		<i>health.x</i>		<i>absences.x</i>		<i>G1.x</i>		<i>G2.x</i>		<i>G3.x</i>	
<i>Min.</i>	1	<i>Min.</i>	1	<i>Min.</i>	0	<i>Min.</i>	3	<i>Min.</i>	0	<i>Min.</i>	0
<i>1st</i>	1	<i>1st</i>	3	<i>1st</i>	0	<i>1st</i>	8	<i>1st</i>	9	<i>1st</i>	9
<i>Median</i>	2	<i>Median</i>	4	<i>Median</i>	3	<i>Median</i>	11	<i>Median</i>	11	<i>Median</i>	11
<i>Mean</i>	2.3	<i>Mean</i>	3.58	<i>Mean</i>	5.3	<i>Mean</i>	10.93	<i>Mean</i>	10.8	<i>Mean</i>	10.51
<i>3rd</i>	3	<i>3rd</i>	5	<i>3rd</i>	7	<i>3rd</i>	13	<i>3rd</i>	13	<i>3rd</i>	14
<i>Max.</i>	5	<i>Max.</i>	5	<i>Max.</i>	75	<i>Max.</i>	19	<i>Max.</i>	19	<i>Max.</i>	20

4.1.4 Portuguese Language Summary Statistics

In the other hand, Portuguese Language is characteristic for being a subject with a mid-study time required and with the 4% of students being failed at least one previous class. The grades have an increasing behaviour from 1st period to 3rd and the absence rate is between 0 and 6 days per student. From those students, 27.8% are receiving some kind of support, mostly from family and 96% of them want to access the higher education. The 52.5% of them do some kind of activity and the 32.5% enjoy from a relationship while they enjoy from a mid-high amount of free time and time to go out. The consumption of alcohol during the weekend slightly increases but in general the health status of the students is good (Table 5).

4 Research Analysis and Findings

Table 5. Summary statistics of Student Portuguese Language Subject variables

<i>guardian.y</i>	<i>traveltime.y</i>	<i>studytime.y</i>	<i>failures.y</i>	<i>schoolsup.y</i>	<i>famsup.y</i>	<i>paid.y</i>
<i>father</i> 90	<i>Min.</i> 1	<i>Min.</i> 1	<i>Min.</i> 0	<i>no</i> 319	<i>no</i> 137	<i>no</i> 343
<i>mother</i> 270	<i>1st</i> 1	<i>1st</i> 1	<i>1st</i> 0	<i>yes</i> 50	<i>yes</i> 232	<i>yes</i> 26
<i>other</i> 9	<i>Med</i> 1	<i>Med</i> 2	<i>Median</i> 0			
	<i>Mea</i> 1.44	<i>Mea</i> 2.05	<i>Mean</i> 0.117			
	<i>3rd</i> 2	<i>3rd</i> 2	<i>3rd</i> 0			
	<i>Max</i> 4	<i>Max.</i> 4	<i>Max.</i> 3			
<i>activities.y</i>	<i>higher.y</i>	<i>romantic.y</i>	<i>famrel.y</i>	<i>freetime.y</i>	<i>goout.y</i>	<i>Dalc.y</i>
<i>no</i> 175	<i>no</i> 15	<i>no</i> 249	<i>Min.</i> 1	<i>Min.</i> 1	<i>Min.</i> 1	<i>Min.</i> 1
<i>yes</i> 194	<i>yes</i> 354	<i>yes</i> 120	<i>1st</i> 4	<i>1st</i> 3	<i>1st</i> 2	<i>1st</i> 1
			<i>Median</i> 4	<i>Median</i> 3	<i>Median</i> 3	<i>Median</i> 1
			<i>Mean</i> 3.938	<i>Mean</i> 3.2	<i>Mean</i> 3.1	<i>Mean</i> 1.5
			<i>3rd</i> 5	<i>3rd</i> 4	<i>3rd</i> 4	<i>3rd</i> 2
			<i>Max.</i> 5	<i>Max.</i> 5	<i>Max.</i> 5	<i>Max.</i> 5
<i>Walc.y</i>	<i>health.y</i>	<i>absences.y</i>	<i>G1.y</i>	<i>G2.y</i>	<i>G3.y</i>	
<i>Min.</i> 1	<i>Min.</i> 1	<i>Min.</i> 0	<i>Min.</i> 0	<i>Min.</i> 5	<i>Min.</i> 0	
<i>1st</i> 1	<i>1st</i> 3	<i>1st</i> 0	<i>1st</i> 10	<i>1st</i> 11	<i>1st</i> 11	
<i>Medic</i> 2	<i>Med</i> 4	<i>Med</i> 2	<i>Median</i> 12	<i>Median</i> 12	<i>Median</i> 13	
<i>Mean</i> 2.3	<i>Mea</i> 3.57	<i>Mea</i> 3.64	<i>Mean</i> 12	<i>Mean</i> 12	<i>Mean</i> 13	
<i>3rd</i> 3	<i>3rd</i> 5	<i>3rd</i> 6	<i>3rd</i> 14	<i>3rd</i> 14	<i>3rd</i> 14	
<i>Max.</i> 5	<i>Max</i> 5	<i>Max.</i> 32	<i>Max.</i> 19	<i>Max.</i> 19	<i>Max.</i> 19	

4.1.5 Contrasting Mathematics and Portuguese Language

Therefore, from the information above, Mathematics and Portuguese Language look quite similar apart from a few characteristics. The most relevant is probably the difference in the support received were Mathematics receives more support from paid in comparison with Portuguese Language who receive less support and it is mainly from their families (Figure 7).

Another difference is the number of students who has repeated the subject which is the half for Portuguese Language and it decrease a rate of 2 as well the number of absences. In general, students seem to have better results in Portuguese Language.

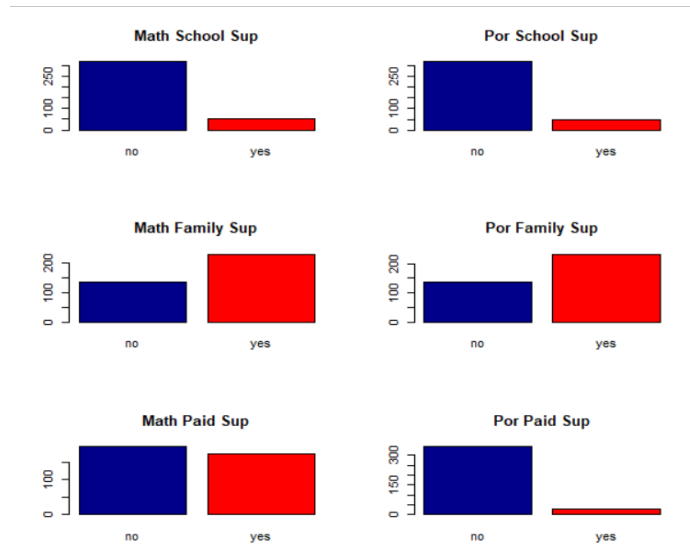


Figure 7. A. Mathematics students receiving school support; B. Portuguese Lang. students receiving school support; C. Mathematics students receiving family support; D. Portuguese Lang. students receiving family support; E. Mathematics students receiving paid support.

4.2 Preliminary Analysis

After the exploration of the data and before starting the preliminary analysis, as the main purpose of this project is to be able to predict the final grade from the Academic and Lifestyle variables of the dataset, it will be necessary to create a new dataset with only those variables (Table 6).

Table 6. Variables selected from the original Students dataset and used to create a working dataset

Attribute	Description (Domain)
<i>sex</i>	student's sex (binary: female or male)
<i>age</i>	student's age (numeric: from 15 to 22)
<i>studytime</i>	weekly study time (numeric: 1 – < 2 hours, 2 – 2 to 5 hours, 3 – 5 to 10 hours or 4 – > 10 hours)
<i>failures</i>	number of past class failures (numeric: n if 1 = n < 3, else 4)
<i>schoolsup</i>	extra educational school support (binary: yes or no)
<i>famsup</i>	family educational support (binary: yes or no)
<i>activities</i>	extra-curricular activities (binary: yes or no)
<i>paidclass</i>	extra paid classes (binary: yes or no)
<i>nursery</i>	attended nursery school (binary: yes or no)
<i>higher</i>	wants to take higher education (binary: yes or no)
<i>romantic</i>	with a romantic relationship (binary: yes or no)
<i>freetime</i>	free time after school (numeric: from 1 – very low to 5 – very high)
<i>goout</i>	going out with friends (numeric: from 1 – very low to 5 – very high)
<i>Walc</i>	weekend alcohol consumption (numeric: from 1 – very low to 5 – very high)
<i>Dalc</i>	workday alcohol consumption (numeric: from 1 – very low to 5 – very high)
<i>absences</i>	number of school absences (numeric: from 0 to 93)
<i>G1</i>	first period grade (numeric: from 0 to 20)
<i>G2</i>	second period grade (numeric: from 0 to 20)
<i>G3</i>	final grade (numeric: from 0 to 20)

Once presented the variables, the dataset is divided in the student's variables plus the subject's variables creating two subsets of 19 variables each. Therefore, this preliminary analysis is divided in two sections, a first one for Mathematics and a second one for Portuguese Language. The variables for each of the datasets have been transformed to Factors with their respective levels and the Integers into Numerical data. This transformation has been done for the analysis purposes and requirements of R coding.

The first step for the analysis is the plotting and visualization of each one of the variables, followed by the correlation and pair plots of the continuous numerical. For the categorical data will create single and combined tables.

From the correlation chart (Figure 8, Figure 9) of the continuous data, G3 is the variable selected as responsive. This decision has been done based on the fact that the 3 grades are highly correlated between each other and G3 is the most representative of the grades and the one intended to be predicted as a main focus of the project.

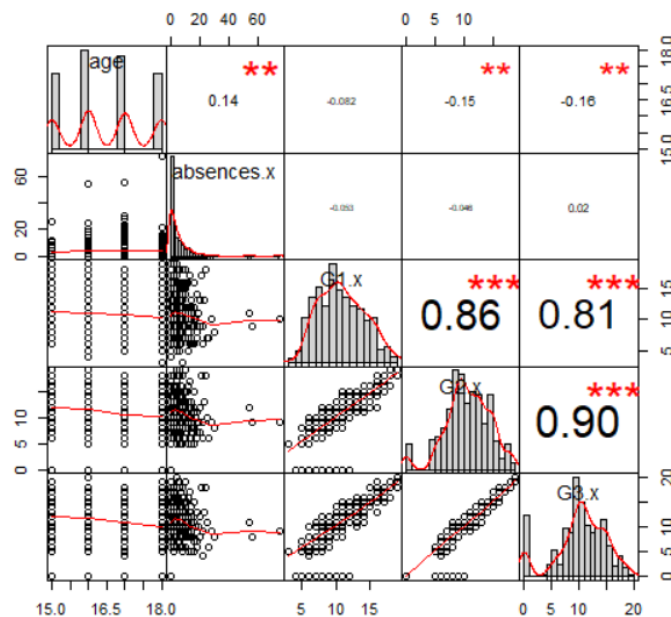


Figure 8. Correlation for Mathematics Continuous Variables

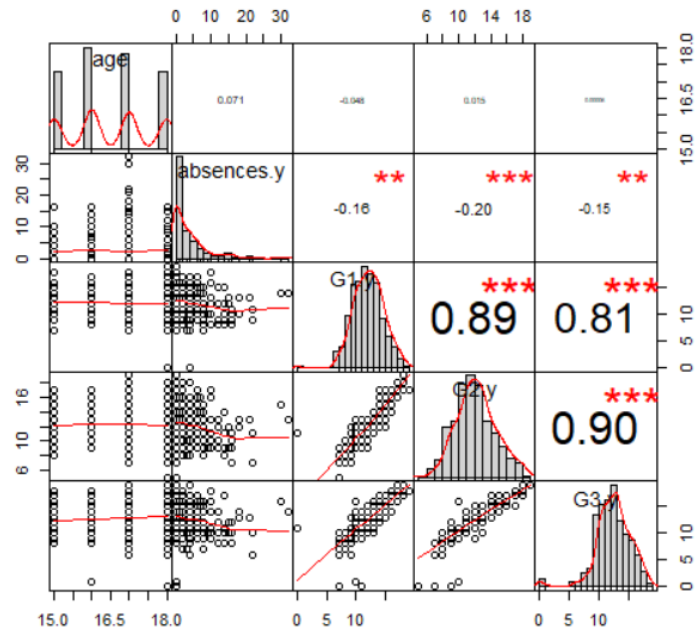


Figure 9. Correlation for Portuguese Language Continuous Variables

4.2.1 Mathematics Data Exploration

From the remaining 19 variables after the first cleaning of the data, and deciding which will be the responsive variable, a correlation table is requested to understand the relationship between them. In Figure 10, it has been marked with colour gradient the relationship between variables. Therefore, it can be observed that “failures” has the most predominant negative correlation with “higher education” and the three grades. Additionally, “higher education” has the highest positive relationship with “grades”. “Study time” as well seems to have a positive relationship with “grades” but the support given by either “paid classes”, “school support” and “family support” does not seem to help improving the grades. Finally, the time spent “going out” is related with a negative impact to the “grades” but it is positive correlated with “daily and weekly alcohol consumption” which at the same time, drinking during the week is as well positive related with drinking during the weekend.

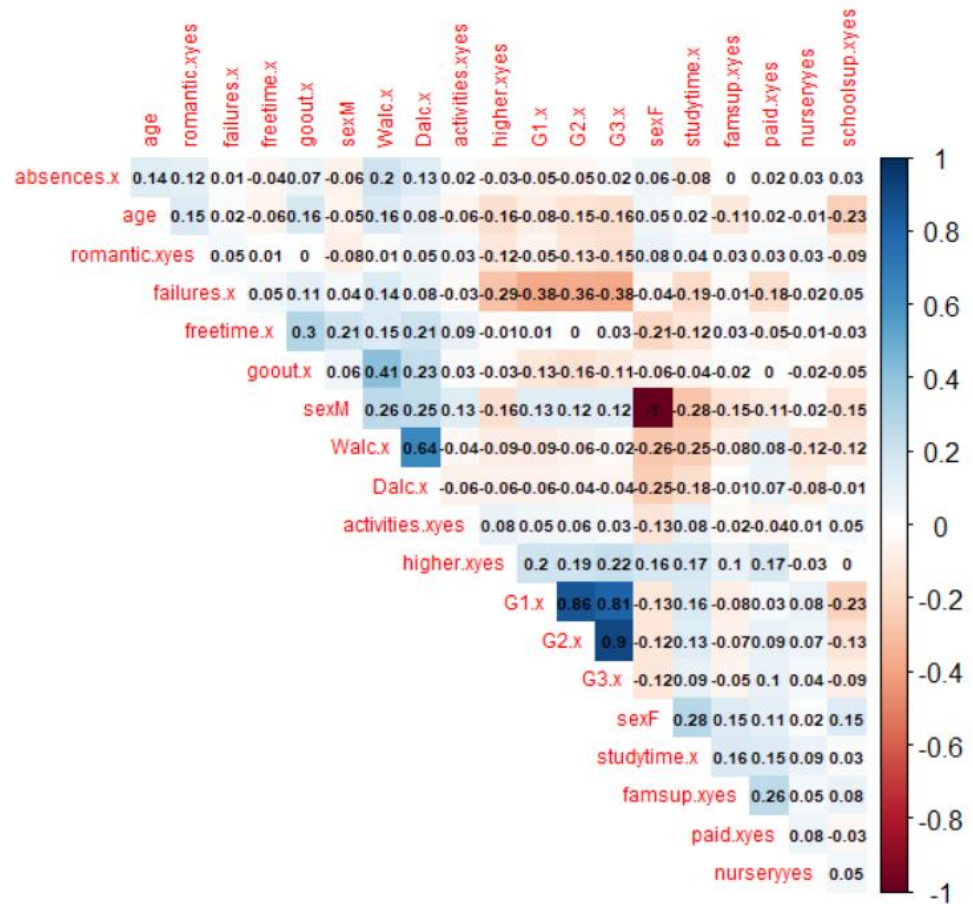


Figure 10. Mathematics Correlation Chart

When analysing mathematics categorical variables, the previous description comes back in the form of graphics. As seen before in the summary statistics, there is a low/mid-level of study time, with not much past failures, a centred distribution of the free time and going out time and a low consumption of alcohol that slightly increases during the weekend (Figure 11).

4 Research Analysis and Findings

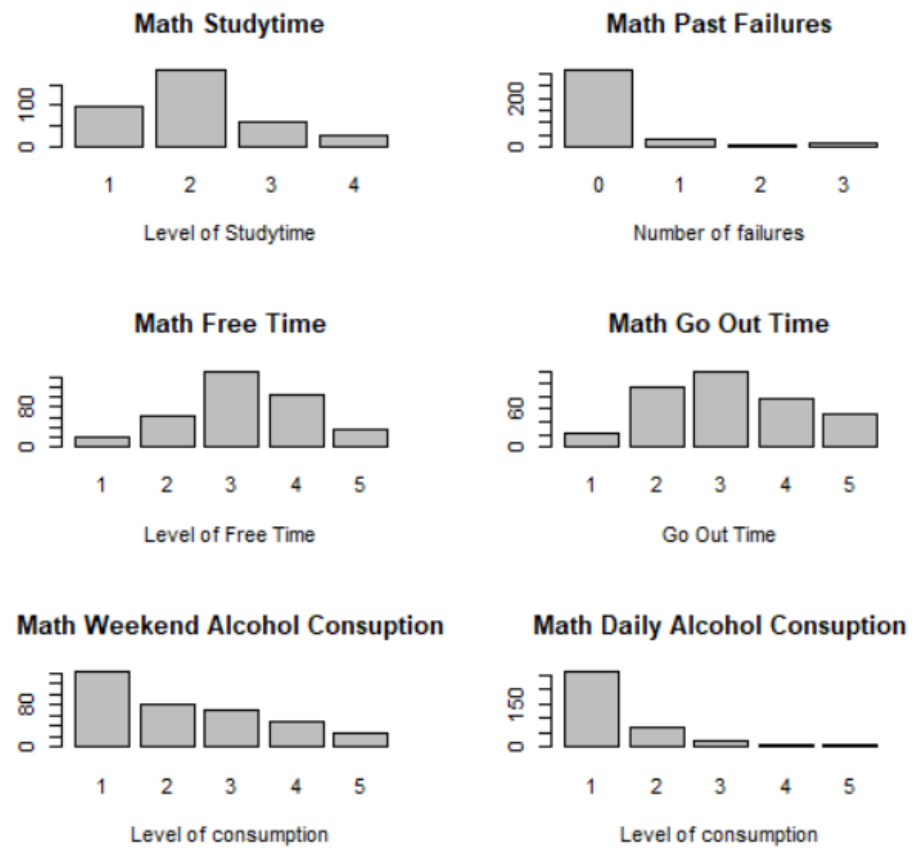


Figure 11. Barplot of: A. Mathematics Studytime; B. Mathematics Past Failures; C. Mathematics Freetime; D. Mathematics Go Out time; E. Mathematics Weekend Alcohol Consumption; F. Mathematics Daily Alcohol Consumption.

Going into detail, being for the first time at class has an apparent positive effect respect those who have already failed Mathematics. There are still some pass grades for those who repeat class for first time but not for the rest (Figure 12).

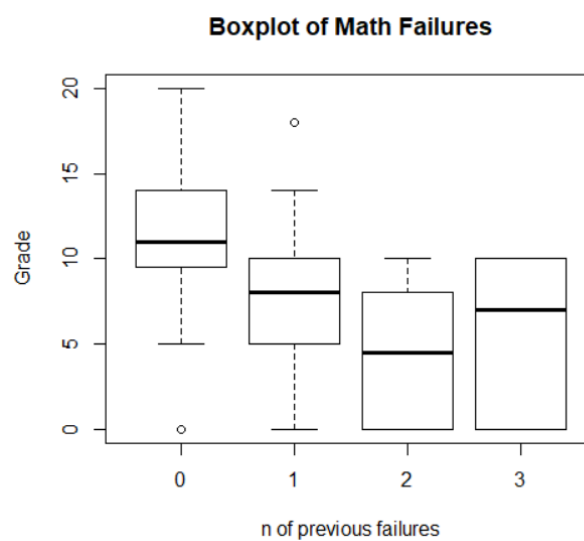


Figure 12. Boxplot of Mathematics Grade 3 and Failures

4 Research Analysis and Findings

For those who are interested in going to the Higher Education programmes after the secondary school Mathematics seems to be a hard subject as median is close to 10 as well as the mean but in general students perform well, whereas the students who does not want to do a Higher Education programme the general grade results are below 10 with not many pass (Figure 13).

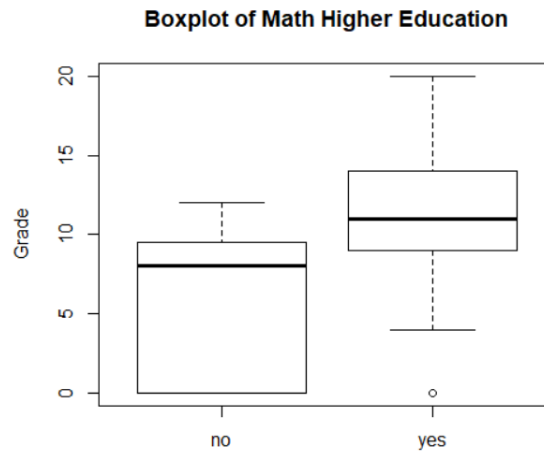


Figure 13. Boxplot of Mathematics Grade 3 and Higher Education

Having free time does not seem to have a great variance on the grade as when observing the boxplot, the best performers seems to be the ones with a bit and a lot of free time having slightly better marks respect the rest (Figure 14).

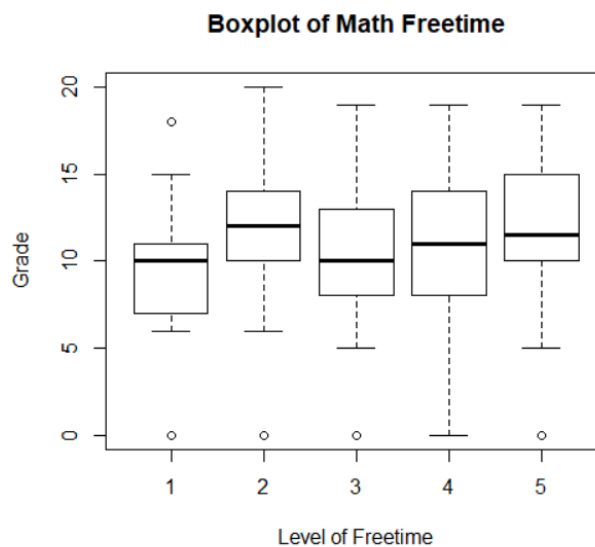


Figure 14. Boxplot of Mathematics Grade 3 and Free Time

4 Research Analysis and Findings

The fact of the students going out has a slightly increase from very low to low but after that, it seems to be a very light decreasing tendency that equals to the very low going out students (Figure 15).

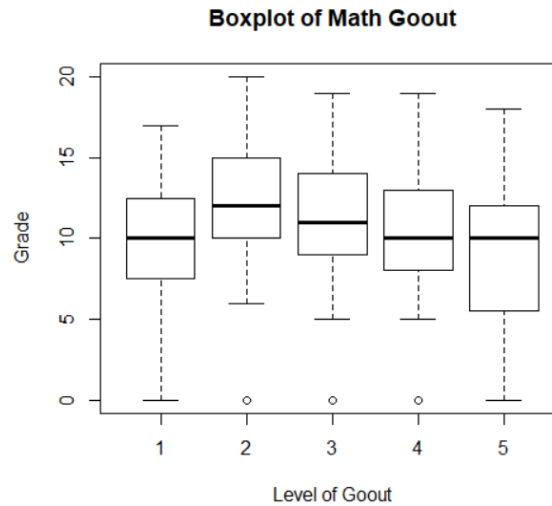


Figure 15. Boxplot of Mathematics Grade 3 and Go Out Time

There are 3 different support types for the students: school, family and paid supports. From those, school support and paid support, are the ones reducing the low marks on G3. However, paid support is the only showing a negative impact when not provided as G3 goes all the range from 0 to 20 while in the other supports starts at 4 to 20 (Figure 16).

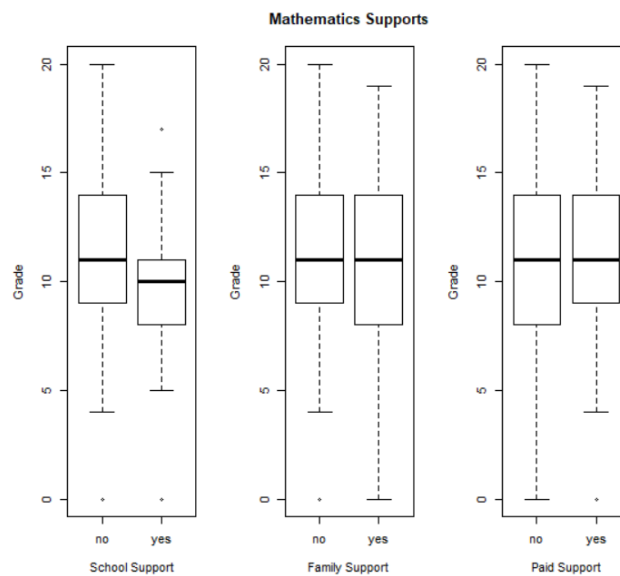


Figure 16. Boxplot of Mathematics Grade 3 and Types of Support

4 Research Analysis and Findings

In general, the third grade has a negative trend when plotted with the absences as, as higher the number of absences the lower is the mark (Figure 17). However, not the lowest number of absences means the highest mark and the highest number of absences the lowest mark.

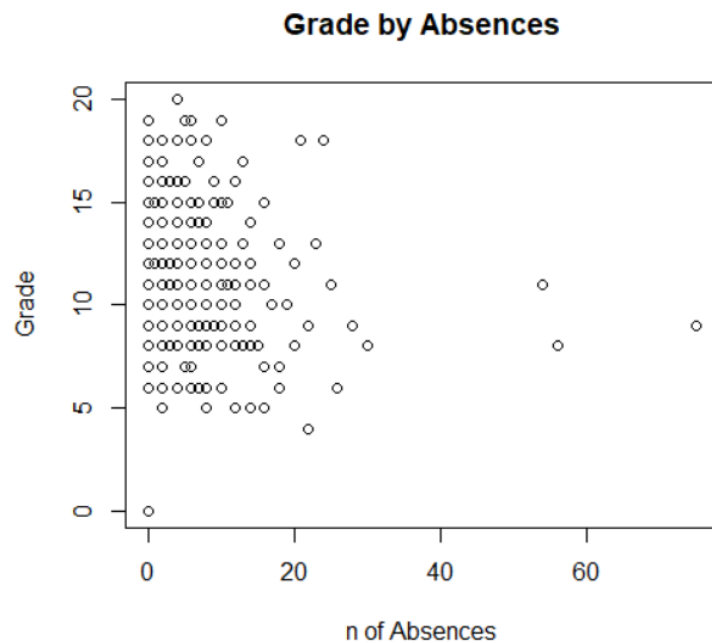


Figure 17. Scatterplot of Mathematics Grade 3 and Absences

4.2.2 Portuguese Language Data Exploration

Focusing on Portuguese Language, the correlation chart at Figure 18 presents again the highest negative correlation of “previous failures” with “grades”, but this time the highest positive correlation is shared between “higher education” and “study time”. However, the “daily/weekly consumption of alcohol” and the “absences”, still have a negative correlation with “grades”.

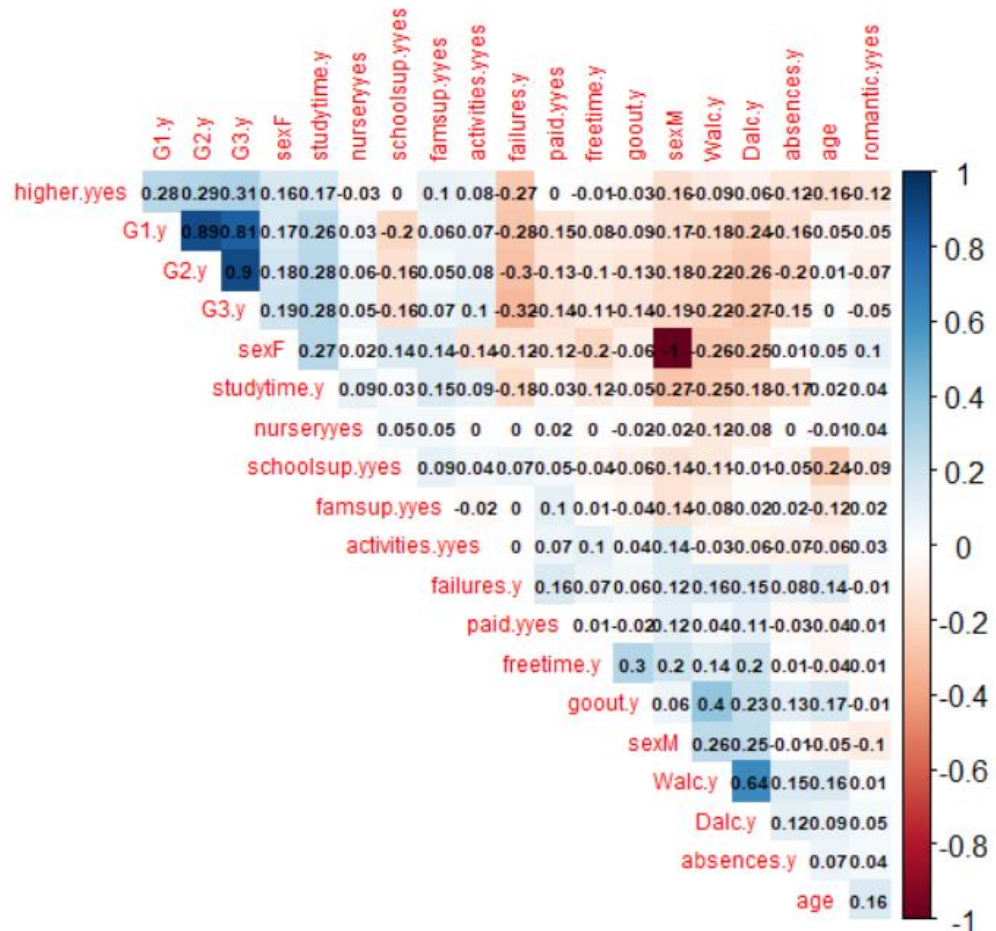


Figure 18. Portuguese Language Correlation Chart

When analysing Portuguese categorical variables, there is no apparent difference than the Mathematics subject results. There is a low/mid-level of study time, a centred distribution of the free time and going out time and a low consumption of alcohol that slightly increases during the weekend. Failures can be the only difference were Portuguese Language has no increase of failures after the decrease of the 1st failure. (Figure 19).

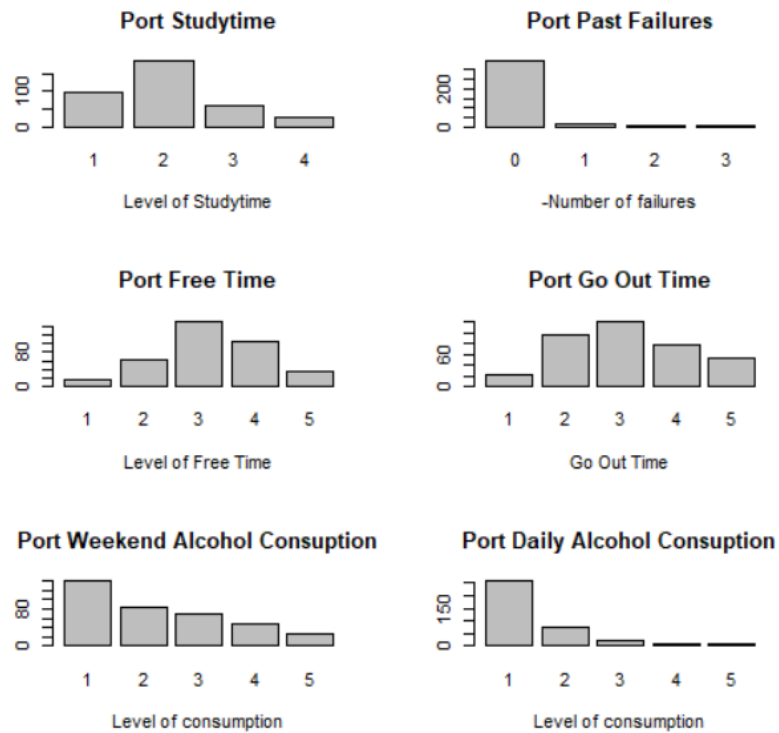


Figure 19. Barplot of: A. Portuguese Lang. Studytime; B. Portuguese Lang.Past Failures; C. Portuguese Lang.Freetime; D.Portuguese Lang.Go Out time; E. Portuguese Lang.Weekend Alcohol Consumption; F. Portuguese Lang.Daily Alcohol Consumption.

In general, there is a positive trend to improve marks with a longer time of study (Figure 20).

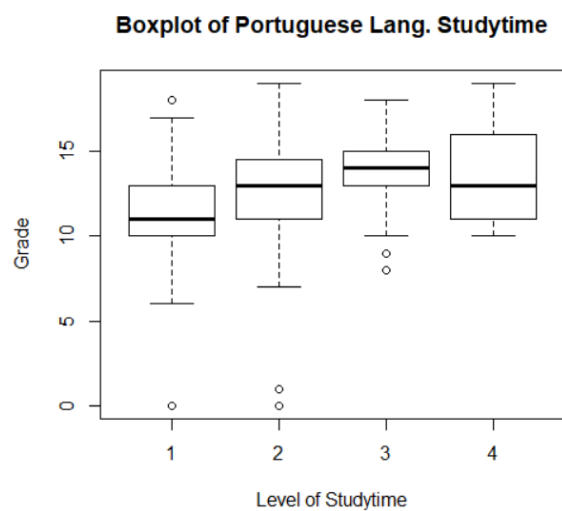


Figure 20. Boxplot of Portuguese Lang. Grade 3 and Study Time

4 Research Analysis and Findings

The students with previous failures have a tendency to reduce their grades where having failed 1 or 3 classes are quite close to the pass mark of 10, as well as 2 classes failed where the range of marks is larger but still quite close to 10 (Figure 21).

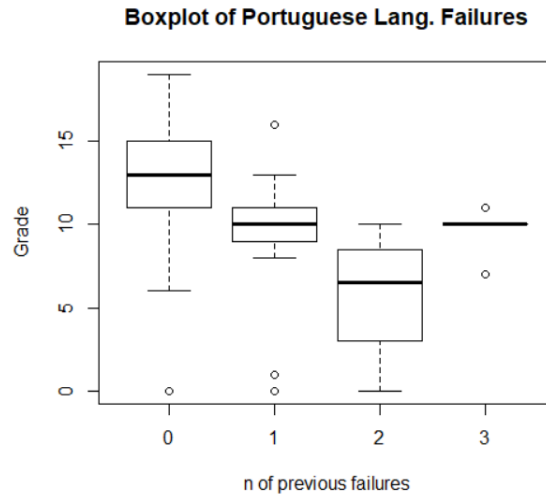


Figure 21. Boxplot of Portuguese Lang. Grade 3 and Previous Failures

The variable G3 is being affected by the students who intent to go to a Higher School programme performing better in results than the ones who does not want to access the Higher Education programme. This second group is sitting close below the 10 marks with only a few students passing the subject. Still, there are some students that want to continue studying that perform below the 10 marks as well (Figure 22).

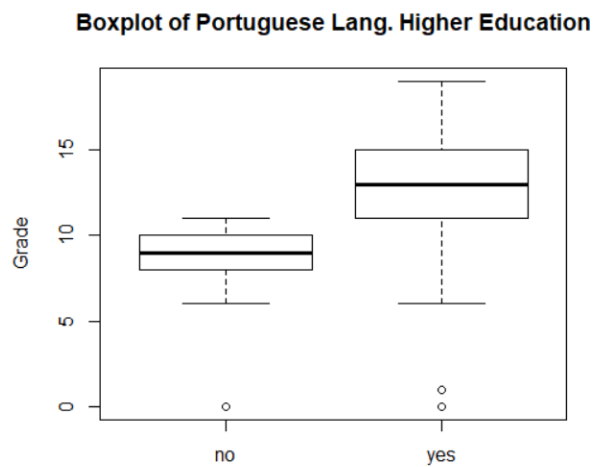


Figure 22. Boxplot of Portuguese Lang. Grade 3 and Higher Education

4 Research Analysis and Findings

There is a sensible decrease of the higher grades linked with the increase of alcohol consumption during the week and where a level 4 and 5 of consumption has the lowest marks (Figure 23).

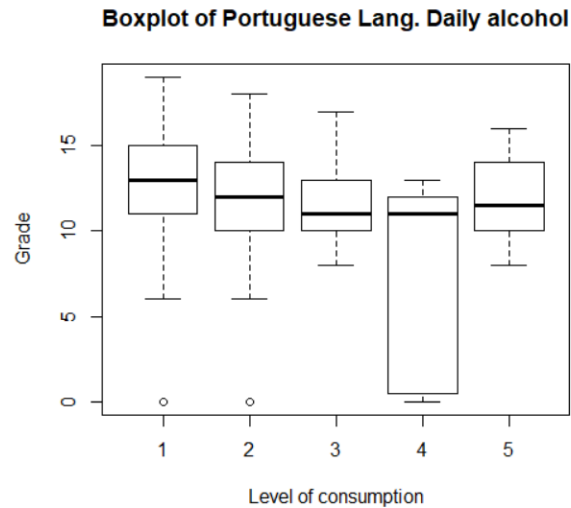


Figure 23. Boxplot of Portuguese Lang. Grade 3 and Daily Alcohol Consumption

In the following scatterplot of the grade depending on the absences, there is a negative trend showing a decrease on the grades with the increase in the number of absences. However, grades do not fall below 5 except for some outliers at 0 or close to 0. (Figure 24).

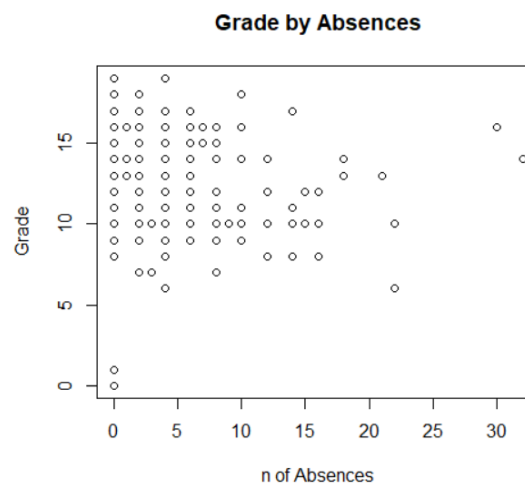


Figure 24. Scatterplot of Portuguese Lang. Grade 3 and Absences

4.2.3 Subject Comparisons

Some of the variables are worth to be presented and compared between subjects as for example the gender of the students and how this affects the grading. In that case, there is no much difference between both groups but it is remarkable the fact that girls seem to perform slightly better than boys at Portuguese Language (Figure 25 **B**). In the other hand, boys seem to perform slightly better than girls at Mathematics subject (Figure 25 **A**).

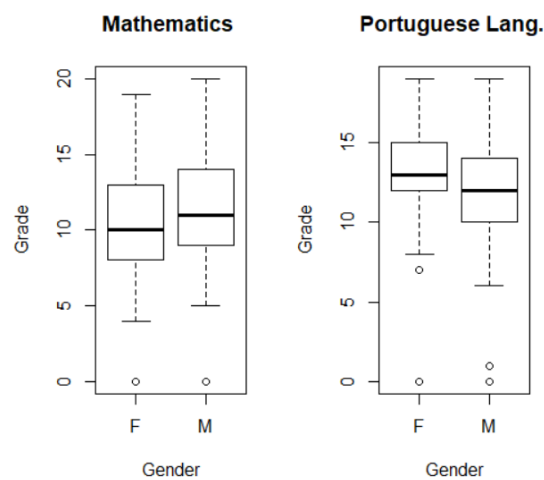


Figure 25. Boxplot of G3 and Gender for: A.Mathematics; B. Portuguese Language

In the comparison for the age variable, the results are as well some kind of opposite where the Mathematics subject seems to have a negative affectation with the older students which slightly decrease their marks (Figure 26 **A**). On the other hand, Portuguese Language has a positive tendency with the older students, even though is quite stabilized (Figure 26 **B**).

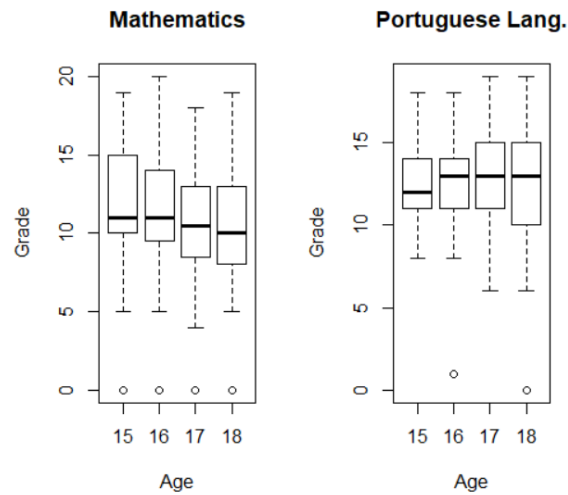


Figure 26. Boxplot of G3 and Age for: A.Mathematics; B. Portuguese Language

The variable studytime when represented with the Mathematics subject does not seem to have a big effect on the grading except for the small increase after the 2 value (Figure 27 A). In comparison with Portuguese Language, this one is increasing level after level until level 3 where level 4 does not seem to continue with the positive trend of the relationship (Figure 27 B).

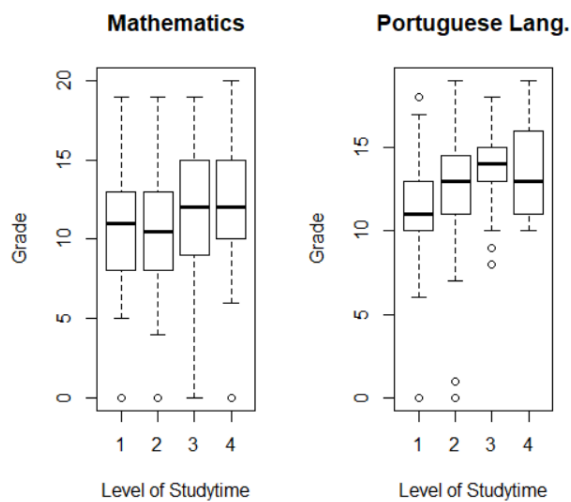


Figure 27. Boxplot of G3 and Study Time for: A.Mathematics; B. Portuguese Language

Having a failed class in the past has a negative effect in comparison with not having failed any class and in this case has a similar affectation in both boxplots, as there is a decrease in the grading. The fact of failing 2 times is having the worst results and failing 3 times seems to increase slightly. Portuguese Language has grading closer to the pass mark where Mathematics is roughly reaching that mark (Figure 28).

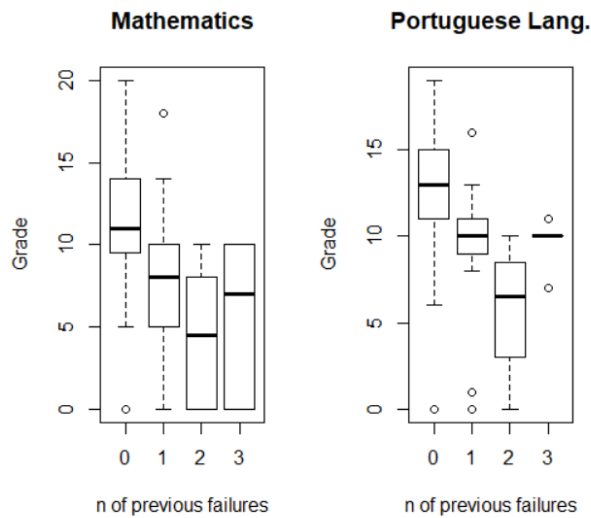


Figure 28. Boxplot of G3 and Previous Failures for: A.Mathematics; B. Portuguese Language

In a first sight to the data, there are variables that present a direct affectation to the 3rd Grading variable. The rest of variables not presented here are not showing as much as difference between groups except G1 and G2 which are directly correlated to G3.

Almost all the previous plots from Mathematics and Portuguese Language have some outliers which means that a few students failed dramatically the subject with grading equal or close to 0 and some others are outperformers even though the number of absences, alcohol consumption, having a previous failed class, etc... These outliers need to be treated as plausible data given the fact that it is a real value and it is acceptable for this dataset.

The continuation of this preliminary analysis will be based on the implementation of Machine Learning techniques to find the best predictive model for each subject. Other techniques that are in the scope of the project are Clustering and PCA's.

4.3 Student Performance Prediction

The intention of the Project is predicting to have a final grade outcome and decide if it can be done with the available data. Following Cortez & Silva (2008) approach, the layout will be to analyse G3 variable as Regression (0-20), Binary (Pass-Fail) and 4 Level (HighPass, LowPass, LowFail and HighFail). Differently to previous published work, there will be a division between training and test datasets, 4 Levels instead of 5 and the techniques used in R.

The presented statistical analyses will be structured in two main sections, Mathematics and Portuguese, and within each section four subcategories to detail the results of each data analysis and prediction. Further, a third and fourth sections will be added to present the discussion on the results between each subject and the presentation of a useful tool that can be a strong resource for academic purposes.

4.3.1 Mathematics Models

As done along the dissertation, will start with Mathematics Statistical Analysis. After the preliminary analysis it can be stated that G3 is normally distributed with some outliers at grade 0 which will be treated as plausible. The rest of variables are mainly factors of 2 and up to 5 levels. During all the analysis, data has been divided into training and test datasets with a division of 80/20 which it is equal to 296 students for training and 73 students for test.

4.3.1.1 Regression

During the analysis of G3 as a numerical variable representing the grades from 0 to 20, six linear models has been fitted in the research of the ideal combination of variables with a significant p-value of 95% (<0.05). The final model achieved is made of 4 variables which are “G1.x”, “G2.x”, “absences.x” and “romantic.x” form which only romantic is affecting the response variable G3 negatively. All detailed explanatory variables effect G3 significantly (p-values < 0.05). In turn, the final model has a R2 adjusted of 0.83 (p-value < 0.0001). However, the prediction of the model has a low accuracy of 34.72% and, therefore, it has been considered not valid. The problem can be observed when plotting the Normal Q-Q plot and the residuals vs fitted where the assumptions are not met as the plot is heavy tailed and skewed left (Figure 29).

4 Research Analysis and Findings

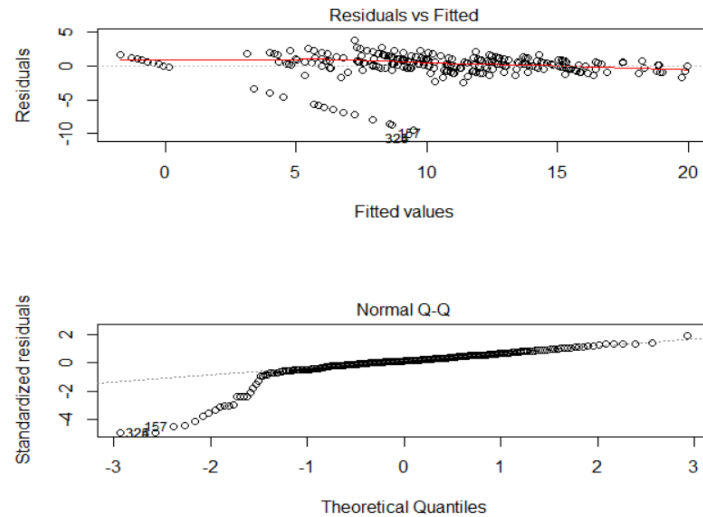


Figure 29. Residuals vs Fitted and Normal QQ plot for Mathematics Linear Model

As not having a success with this approach, the package “caret” has been used with different Machine Learning techniques to have different results to compare. From these results, the best performer on prediction is the Random Forest with the lowest Root Mean Squared Error of 1.48 and a prediction rate of 44.44%, almost a 10% higher than the normal linear model. Taking into consideration the prediction rate, the Classification Tree is divided mainly by “G2.x” with a 48% of variable importance and “G1.x” with a 34% variable importance. It shows a first split of 53% at grades greater than 10.5 going to a final leave of grade 14 and the 47% of the remaining data being divided in 2 final leaves of 3.2 and 8.7 grades (Figure 30).

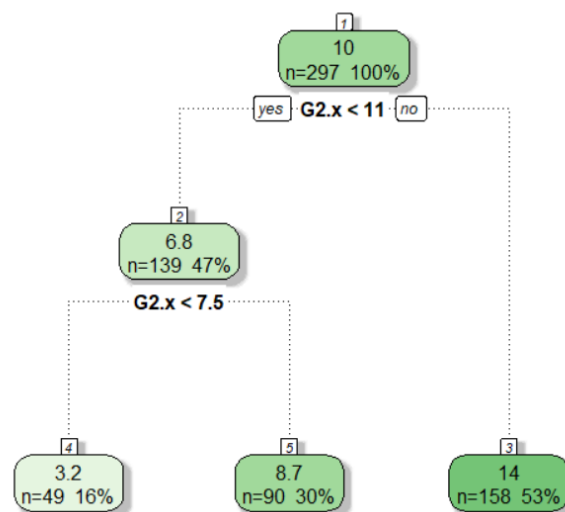


Figure 30. Regression Tree of G3.x in Mathematics Regression Analysis

4.3.1.2 Binary Classification

In order to see more depth after the application of regression, the “caret” package is applied again once the data has been transformed into binary Pass/Fail. The results for this approach, show that the higher performer in that occasion is Boosting with a 92.3% accuracy, followed by Random Forest with 91.5%. The variable importance is as well focused on “G2.x” with a 52% and “G1.x” with a 31%. The primary splits have been done on “G2.x” being greater than 9.5 and “G1.x” being greater than 10.5. The tree has only 2 leaves as data is binary and the 64% of students pass the subject (Figure 31).

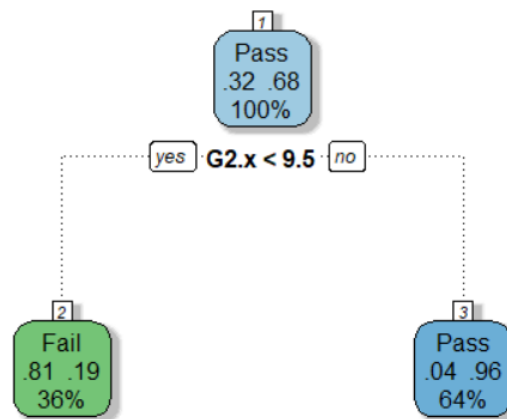


Figure 31. Classification Tree of G3.x in Mathematics Binary Analysis

4.3.1.3 Four Levels Classification

Binary approach has a good predicting accuracy but it does not provide much information of the data. When applying the 4 levels, the outcome still has a high accuracy rate of 83% for Random Forest and 82.5% for Boosting. A confusion matrix of the predicted values and the test dataset shows that Boosting performs better out of the “caret” package with an accuracy of 84.93% vs 82.82%. The tree presents a variable importance of 59% for “G2.x” and 39% for “G1.x” with a first split at “G2.x” being greater than 13.5 and “G1.x” greater than 11.5. Additionally, a second split is performed on the lower marks differenced on “G2.x” and “G1.x” being greater than 10.5 and 11.5 respectively (Figure 32).

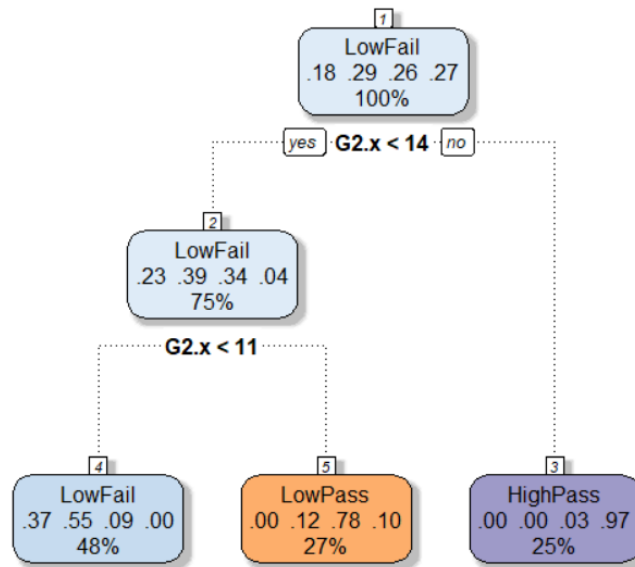


Figure 32. Classification Tree of G3.x in Mathematics 4 Level Analysis

The 4 Level approach allows us to have more information about the data but still is hidden by the heaviness of “G2.x” and “G1.x”. Therefore, the same process has been applied removing “G2.x” and “G1.x” to understand the behaviour of the data without these variables. The findings are that accuracy has heavily decreased to a 39.08% for Boosting and when creating the confusion matrix and compared with the test dataset, the best performer is Random Forest with a 35.61% over Boosting with a 31.50%. The tree shows now more divisions affected by “absences.x”(45%), “paid.x”(32%), “famsup.x”(6%), “freetime.x, level 2”(3%), “goout.x, level 3”(3%), “activities.x”(3%), “failures.x, level 2”(2%). The first split is done by the students with no absences which, after that, is divided between “High Fail” (17%) and “High Pass” (46%) depending on the paid classes. The last split, identified as well by the absences being higher than 6.5, generates the last two groups of “Low Fail” (14%) and “Low Pass” (23%), (Figure 33).

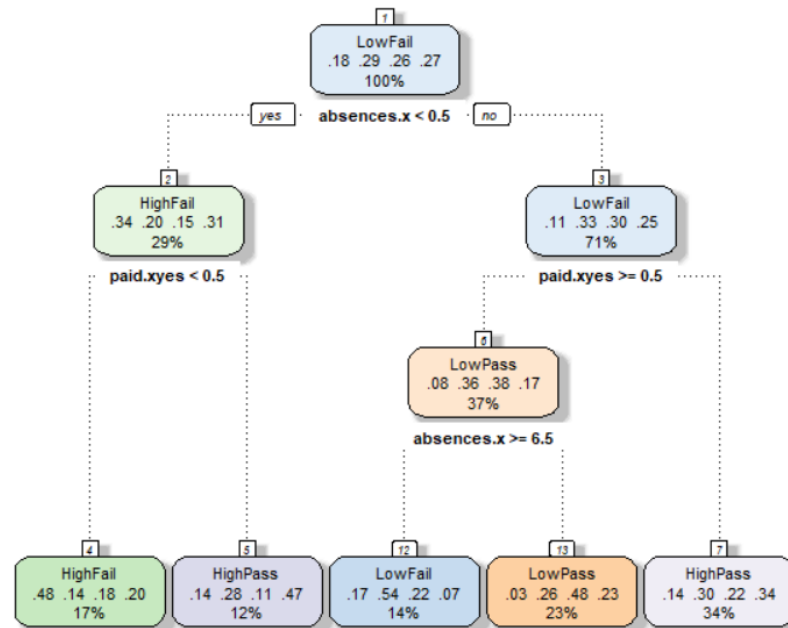


Figure 33. Classification Tree of G3.x in Mathematics 4 Level Analysis without G2.x and G1.x

4.3.1.4 Clustering / PCA

A hierarchical clustering using the bottom up approach represents all the students structure showing as well an appropriate division on 4 clusters in the dendrogram (Figure 34).

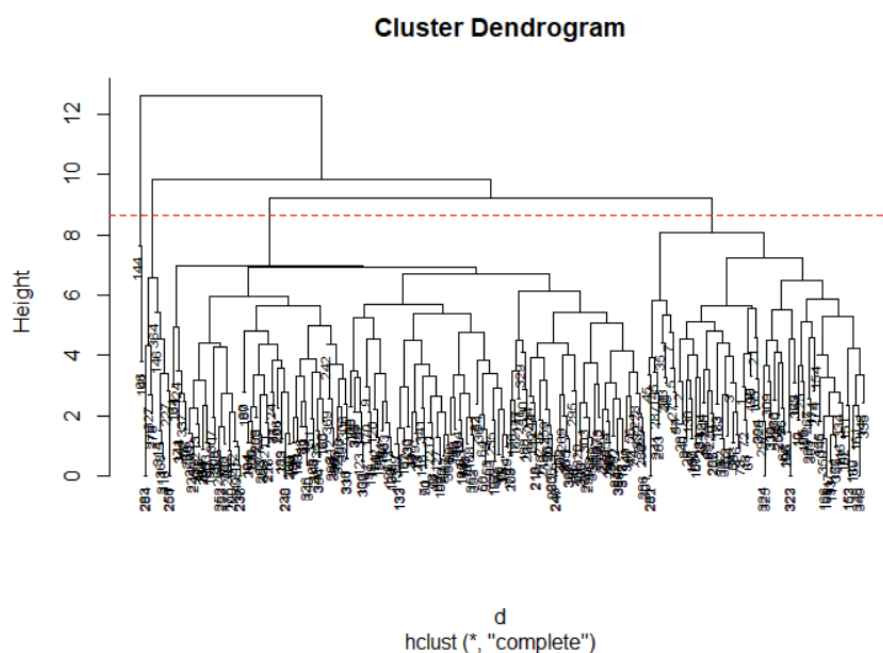


Figure 34. Dendrogram of Mathematics Students

Furthermore, the 4-cluster division of the data presents four groups of 88, 134, 23 and 124 students where first cluster means being above average on absences and activities and falling below average on family support. The second cluster is high above average on G1 and G 2 and falling below average on failures. Regarding the third cluster, G1 and G 2 are far below average along with study time and higher education while failures and paid classes are highly above average. Finally, the fourth cluster is similar to the first, being moderate, but with means above average for study time, school, family and paid classes support and below average on activities.

Continuing with the analysis, PCA shows the need of 8 Principal Components (PC) in order to explain at least 80% of the total variance with a double elbow on 4th PC and the 6th PC in the scree plot (Figure 35).

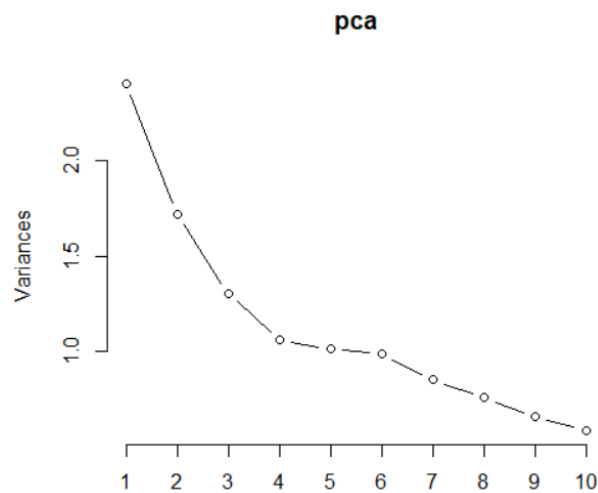


Figure 35. Scree plot of Mathematics PCA

Focusing in PC1 and PC2, only the 34.34% of the variance is explained. The first component being a weighted average of the twelve variables with lower G1 and G2 and high in failures explaining the 20% of the variance. The second component is high on school and family support, paid classes and study time and mid high on higher education. It is low on sex, G1 and G2 but not the lowest and it explains the 14.3% of the variance. Overall, the distribution of PCs clearly show differences in G3 distribution where High Pass, Low Pass, Low Fail and High Fail are clearly grouped in colours and linearly distributed along the plot (Figure 36).

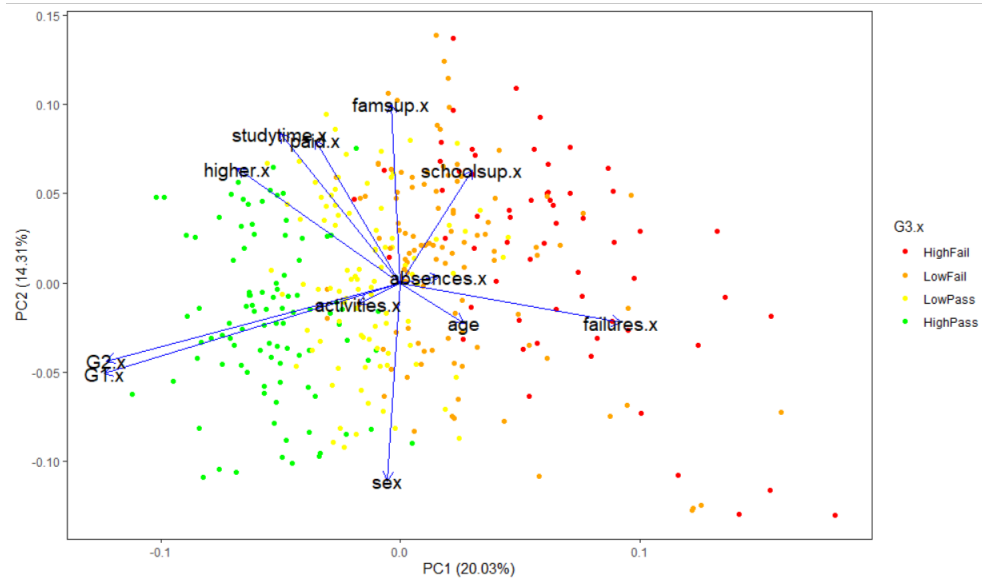


Figure 36. Mathematics PC1 and PC2 plot of weighted averages per student

4.3.2 Portuguese Language Models

Portuguese Language has less failed students with some differences in the variables when compared with Mathematics. For prediction purposes the same Machine Learning algorithms will be applied to the Portuguese Language and will try to find the best model for the subject.

4.3.2.1 Regression

As previously done with Mathematic subject, a manual approach to the regression has been done with unsuccessful results as the best model with only “G2.y” with positive slope and “failures.y” affecting “G3.y” negatively to the slope. Both variables with a significant effect on G3 (p-values < 0.05) and with a R^2 adjusted of 0.81(p-value < 0.0001), indicating strong correspondence between the observed and modelled data. The prediction of the model has a mid-accuracy of 52.02% and assumptions are met except for the outliers that are preserved as valid. Normal Q-Q plot is bimodal and skewed left (Figure 37). For the purposes of the project, the accuracy is considered insufficient and the model has been rejected.

4 Research Analysis and Findings

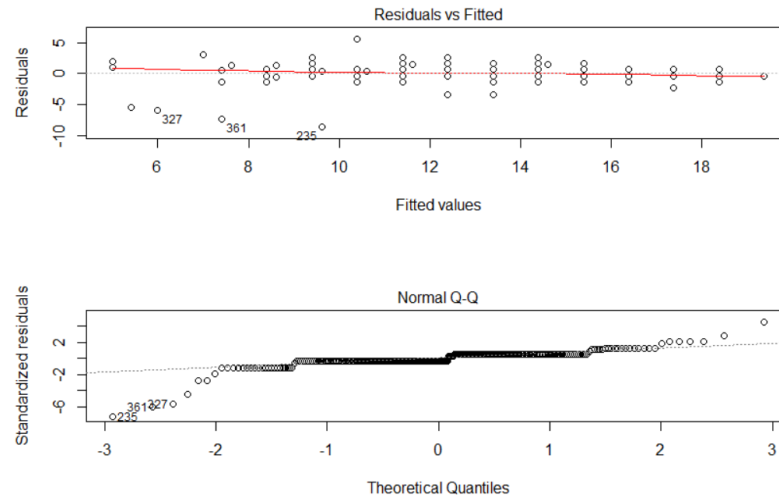


Figure 37. Residuals vs Fitted and Normal QQ plot for Portuguese Language Linear Model

From the caret package, the best performer on prediction is the Random Forest with the lowest RMSE of 1.15 and a prediction rate of 54.79%, very close to the normal linear model. The Classification Tree is divided by “G2.y” with a 56% of variable importance and “G1.y” with a 36%. It shows a first split of 28% at grades greater than 13.5 going to a final leaf of grade 16 and the 72% of the remaining data being divided in 2 final leaves of 9.2 and 12 grades, 23% and 49% of the students respectively (Figure 38).

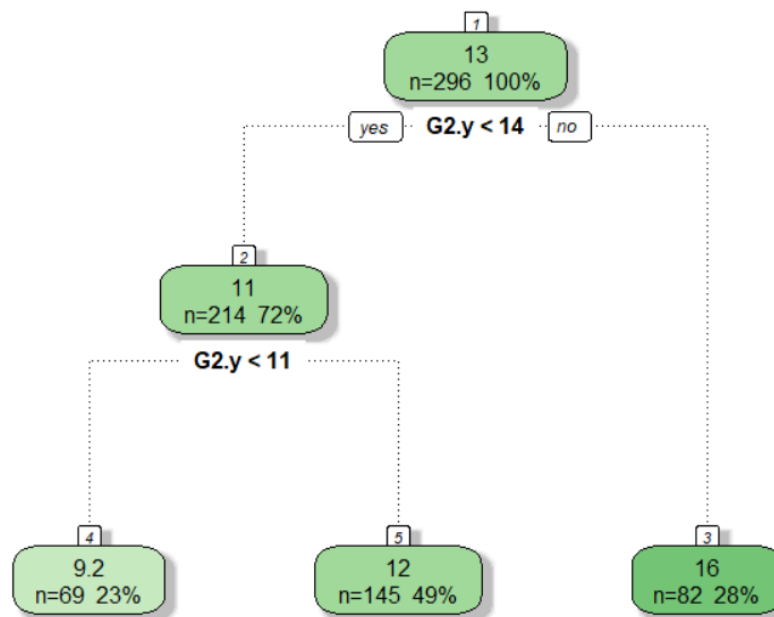


Figure 38. Regression Tree of G3.y in Portuguese Language Regression Analysis

4.3.2.2 Binary Classification

The results for the Pass / Fail approach, show that the higher performer in that occasion is the normal Classification Tree without improving with an accuracy of 95.4% and Random Forest is very close with a 95.15%. The variable importance is exclusive to “G2.y” with a 71% and “G1.y” with a 29%. The primary splits have been done on “G2.y” and “G1.y” being greater than 8.5. The tree has only 2 leaves as data is binary and the 96% of students pass the subject (Figure 39). When comparing results with the test dataset, Boosting has the best performance, increasing the accuracy to a 98.6% and Random Forest keeps exactly the same accuracy of the Classification Tree.

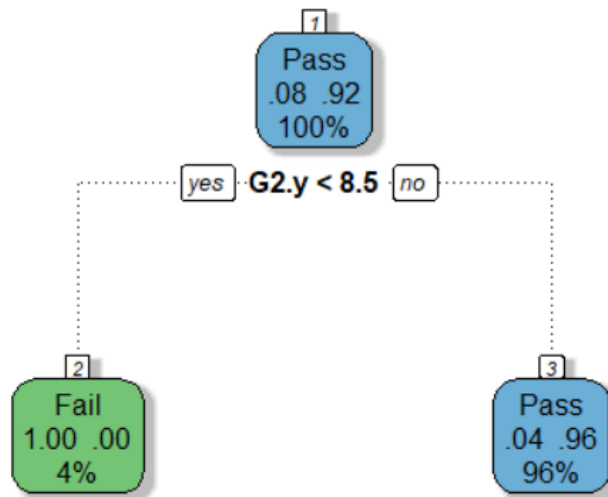


Figure 39. Classification Tree of G3.y in Portuguese Language Binary Analysis

4.3.2.3 Four Levels Classification

When applying the 4 levels on the Portuguese Language subject, the outcome keeps a high accuracy rate of 86.2% for Random Forest, but still slightly below the binary results. A confusion matrix of the predicted values and the test dataset shows that Classification Tree performs better with an accuracy of 83.56% and Random Forest falls to 76.71%.

The tree presents a variable importance of 55% for “G2.y” and 34% for “G1.y” with a first split at grade above 13.5 for both, grouping 28% of the students in “High Pass”. The remaining 72% of students are divided on “G2.y” smaller than 10.5 and “G1.y” smaller than 9.5, being 49% of it a “Low Pass” and 24% a “Low Fail” (Figure 40).

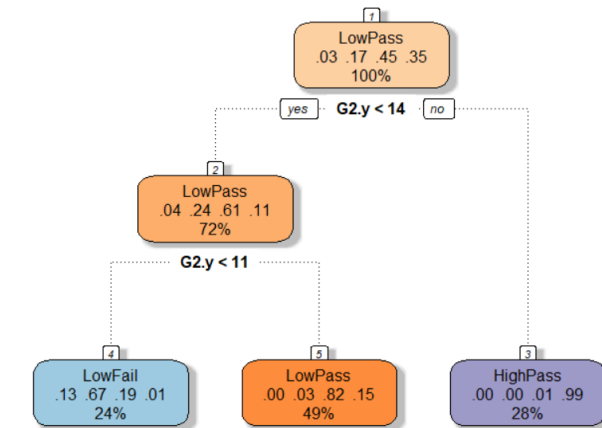


Figure 40. Classification Tree of G3.y in Portuguese Language 4 Level Analysis

In a second step, when removing “G2.y” and “G1.y” from the dataset and repeat the models, accuracy decreases to a 48.2% for Classification Tree outperformed by Bagging with a 49.6%. Now the Classification Tree, present further divisions with a variable importance of 40% for higher education, 27% for school support, 22% on age, 3% on failures and 8% others. The first split is done on a 4% to “Low Fail” for those who want to attend higher education, do school support, are older than 16.5 years old, haven’t fail the subject before and are below 8.5 absences. For the remaining 96% there is a 13% “Low Pass” for those attending school support and 40% for those who are not at school support but are older than 16.5 years old. The last leave is for the 43% of the students on a “High Pass” and are not going to attend higher education, do not receive school support and are younger than 16.5 years old (Figure 41).

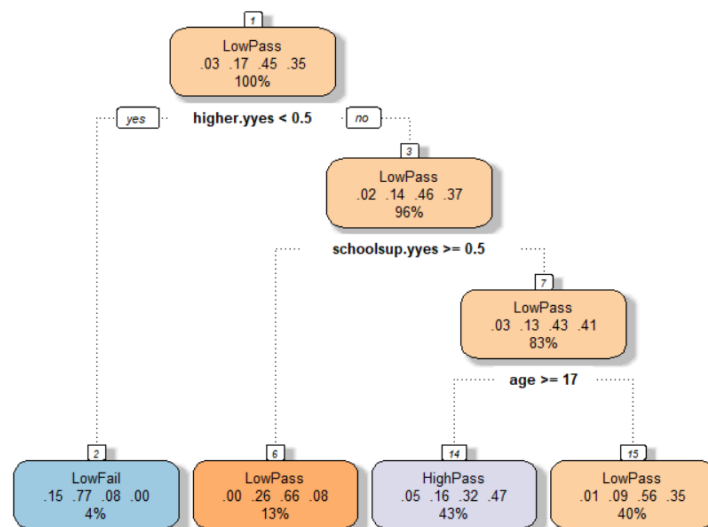


Figure 41. Classification Tree of G3.y in Portuguese Language 4 Level Analysis without G2.y and G1.y

4.3.2.4 Clustering/PCA

Portuguese Language Dendrogram has a hard to decide structure on the clustering. However, once again will decide on the 4 clusters as it is the longest Dendrogram distance that can be cut with a horizontal line (Figure 42).

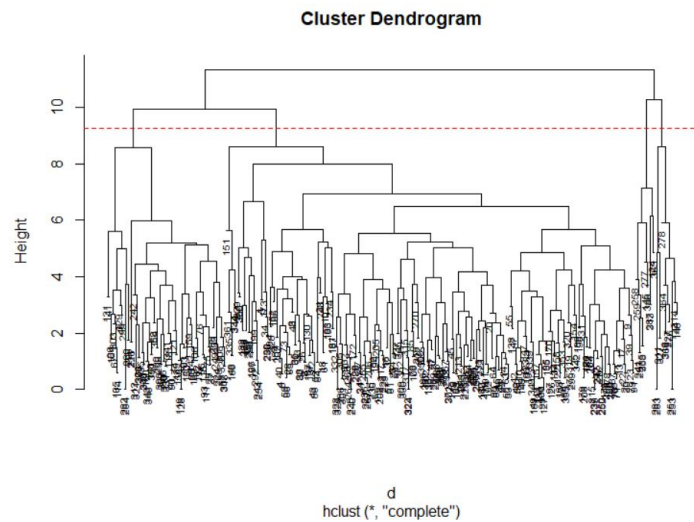


Figure 42. Dendrogram of Portuguese Language Students

The K-means cluster analysis, has been kept to the 4 cluster division of the data into four groups of 15, 39, 191 and 124 students with cluster 1 means being extreme below average on G1 and G2, family support, study time, activities and higher education and above average on age, sex, absences and failures. The second cluster is as well below average on G1 and G2, activities and study time but above average on age, failures, family support and high above average on absences. Regarding the third cluster, is as well below average on G1 and G2, age, absences and study time but not as extreme as the other clusters. Above average can be found school support and paid classes. Finally, the fourth cluster is the only one with G1 and G2 above average along with study time and family support. It has as well the most extreme means below average for absences, failures, school support and paid classes.

Following the analysis, PCA shows again the need of 8 Principal Components in order to explain at least 80% of the total variance with no obvious elbow after PC2 in the scree plot (Figure 43).

4 Research Analysis and Findings

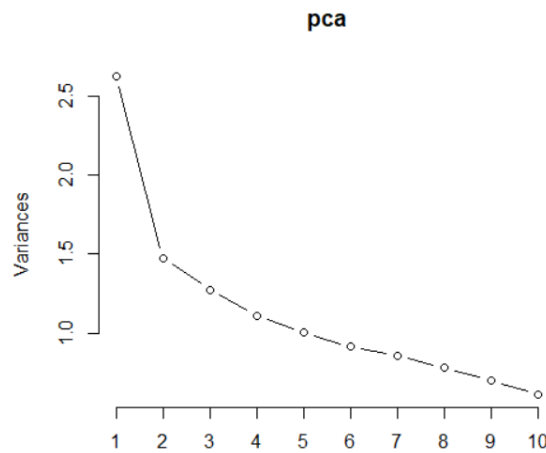


Figure 43. Scree plot of Portuguese Language PCA

Most of the variance is being explained by PC1 with a 21.85% but together with PC2 only 34.13% of the variance is being explained. Analysing PC1 and PC2 in deep, the first component is a weighted average of the twelve variables with lower G1 and G2 and high in failures. The second component explains the 12.3% of the variance and is high on school and family support, paid classes and study time and mid high on higher education. It is mid low on G1 and G2 but not the lowest. Overall, the distribution of PCs clearly show differences in G3 distribution where High Pass, Low Pass, Low Fail and High Fail are clearly grouped in colours and linearly distributed along the plot (Figure 44).

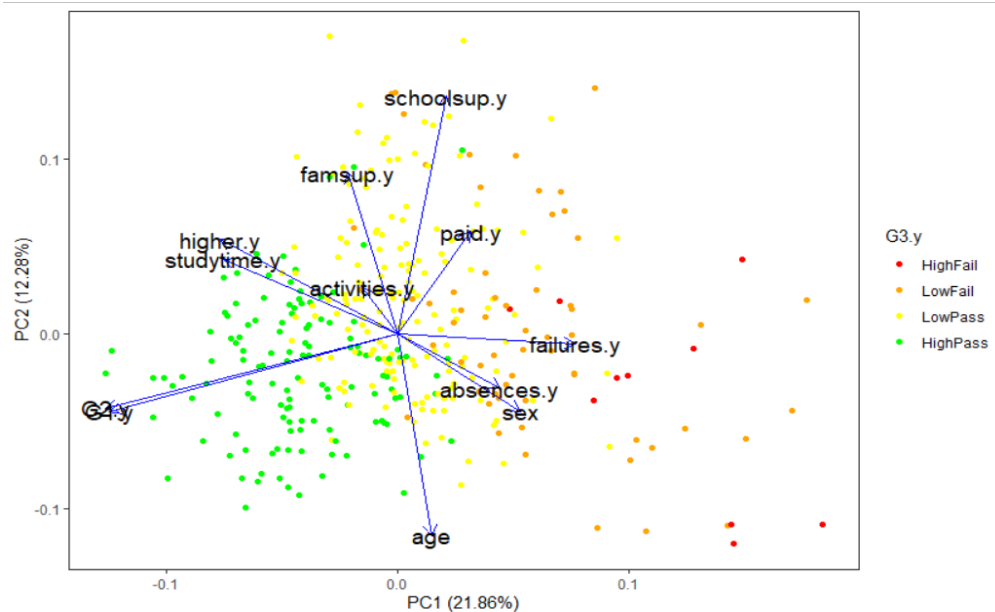


Figure 44. Portuguese Language PC1 and PC2 plot of weighted averages per student

4.3.3 Predicting Tool, ShinyApp

The strongest academic variables determined by the predictive models have been saved in a new dataset. A tool inspired for the benefit and improvement of the students' performance with a fast, reliable and user friendly interface has been created. It allows teachers to do fast decision during the academic course. The App emulates the random forest predicting model that has been created comparing the results of the different models and selecting the best performer. The landing page has a dashboard interface with information and structure of the App (Figure 45). Second and third tabs include the variables of the model to input student information (Figure 46) and when the selection is run by the "Enter" button the outcome of the model appears (Figure 47). Now, teachers and professionals in the education field can use this tool to predict final grades in a minute with an accuracy from 35% to 90% depending on the completion of the input selection. This tool can help to make fast and reliable decisions on the structure of the classes, reinforcement subjects and homework demand.

Student Performance App has been deployed online and can be visited at:

<https://kawkawbala.shinyapps.io/StudentPerformanceApp/>

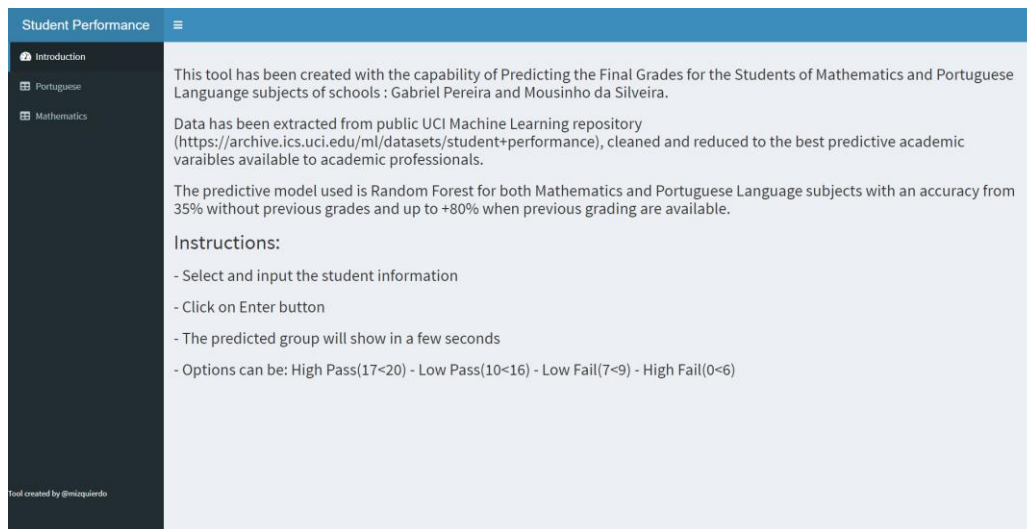
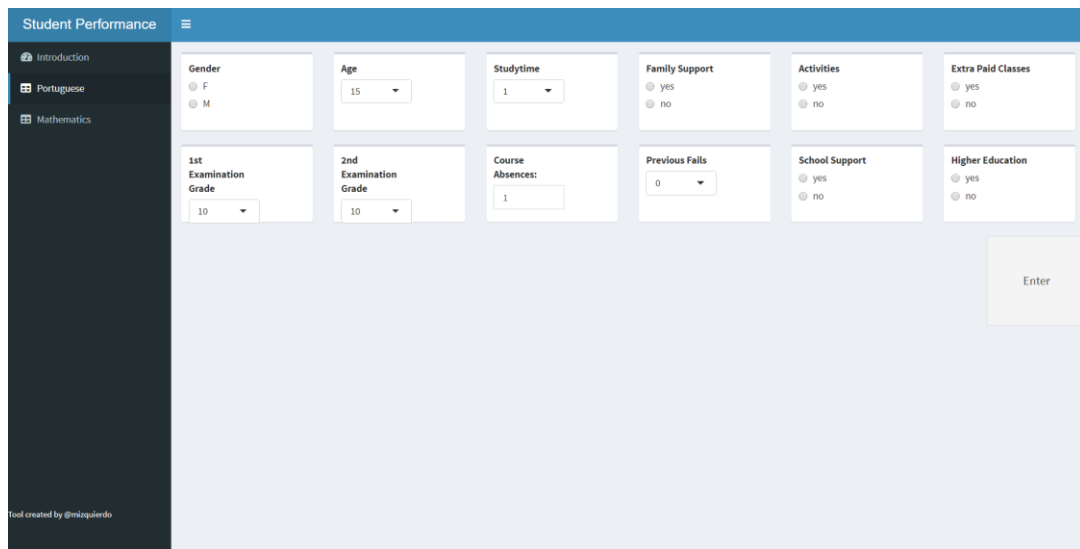


Figure 45. ShinyApp Introduction landing page

4 Research Analysis and Findings

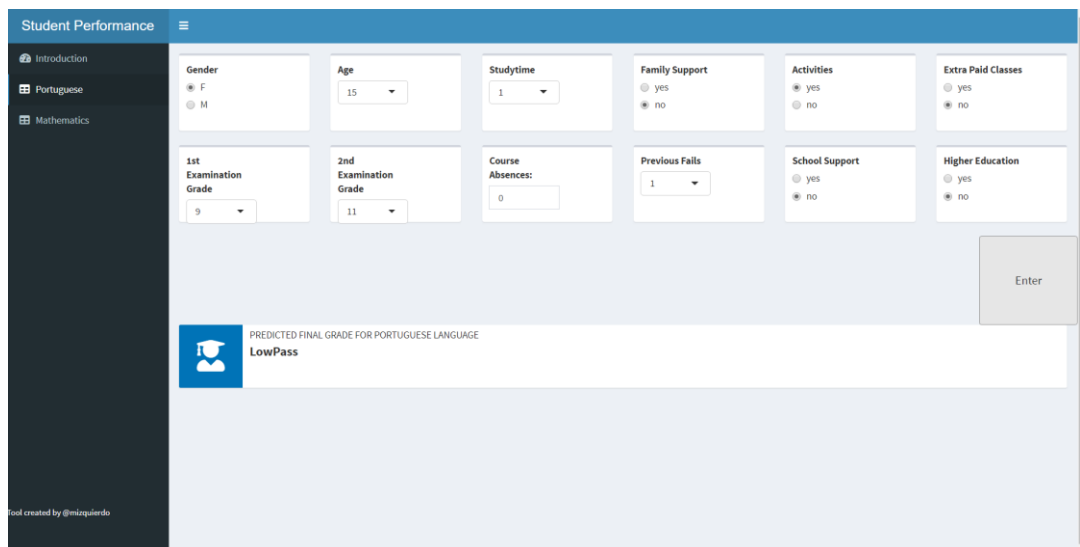


The image shows a ShinyApp interface titled "Student Performance". On the left is a dark sidebar with navigation links: "Introduction", "Portuguese" (selected), and "Mathematics". The main area contains a grid of input fields for student data:

- Gender:** Radio buttons for "F" and "M".
- Age:** A dropdown menu showing "15".
- Studytime:** A dropdown menu showing "1".
- Family Support:** Radio buttons for "yes" and "no".
- Activities:** Radio buttons for "yes" and "no".
- Extra Paid Classes:** Radio buttons for "yes" and "no".
- 1st Examination Grade:** A dropdown menu showing "10".
- 2nd Examination Grade:** A dropdown menu showing "10".
- Course Absences:** A text input field containing "1".
- Previous Fails:** A dropdown menu showing "0".
- School Support:** Radio buttons for "yes" and "no".
- Higher Education:** Radio buttons for "yes" and "no".

An "Enter" button is located at the bottom right of the input grid. At the bottom left of the main area, it says "Tool created by @mizquiedo".

Figure 46. ShinyApp Portuguese Language input tab



This image shows the same ShinyApp interface as Figure 46, but with the predicted outcome displayed. The input values are: Gender (F), Age (15), Studytime (1), Family Support (no), Activities (yes), Extra Paid Classes (no), 1st Examination Grade (9), 2nd Examination Grade (11), Course Absences (0), Previous Fails (1), School Support (no), and Higher Education (no). The "Enter" button has been clicked, and the result is shown in a white box at the bottom:

PREDICTED FINAL GRADE FOR PORTUGUESE LANGUAGE
LowPass

The sidebar and footer text remain the same.

Figure 47. ShinyApp Portuguese Language input tab with outcome

5 Discussion and Conclusions

The work presented here highlights the importance of having a better understanding of the academic-related backgrounds of students and how this influence their marks, directly impacting their educational progress.

When observing the manual regression both subjects didn't meet the assumptions but the best models present a better performance on Portuguese Language with a much simpler model based on "G2.y" and "failures.y" and having a prediction accuracy of 52.02% versus 34.72% for Mathematics with "G1.x", "G2.x", "absences.x" and "romantic.x".

The caret package, with no doubt, has simplified the Machine Learning Techniques. For both subjects we can quickly identify the underperformer techniques such as the Linear Discriminant Analysis, Logistic Regression, Neural Network, Support Vector Machine, k-Nearest Neighbours and Naive Bayes. Therefore, the analysis is focused on the Classification and Regression Tree, Bagging, Random Forest and Boosting.

During the regression analysis for the prediction of G3, Random Forest is the best performer (Table 7) with better results on Portuguese Language which has higher grades than Mathematics having the same tree distribution and splits.

Table 7. Root Mean Squared Error Results for Caret package

RMSE Table	logistic	svm	knn	bagging	rf	gbm	Predict
Mathematics	1.950405	2.08871	3.007368	1.602039	1.477337	1.590868	0.4444
Portuguese	1.243961	1.472411	2.021377	1.223969	1.145132	1.250465	0.5479

The Binary analysis shows a high accuracy as well on Random Forest with more than 90% accuracy which it is slightly improved by the Classification Tree in Portuguese Language (Table 8). The variables with more weight are again G2 and G1 and the results show a higher chance to pass the subject on Portuguese Language with the 96% of students having pass the subject against 64% on Mathematics. However, the tree splits in a lower grade in Portuguese Language, 8.5 instead of 9.5.

The 4 level results are as well the same for Mathematics and Portuguese Language with Random Forest as the best performer with an accuracy above 80% (Table 8) heavily relying on G2 and G1 but the Classification Tree for Mathematics starts at "Low Fail"

5 Discussion and Conclusions

instead of “Low Pass” for Portuguese Language. After the first split, Mathematics has almost the 50% of the students in “Low Fail” and 27% in “Low Pass” which is just the opposite in Portuguese Language which has almost the 50% at “Low Pass” and 24% at “Low Fail”.

When removing G2 and G1 the best performer model is still Random Forest except for Portuguese Language which has a roughly 1% improvement in the Bagging model (Table 8). The results show that Mathematics is the only with “High Fail” results on a 17% of the students and Portuguese Language has more percentage of “High Pass”. The first split is in this case lower for Portuguese with a 9.5 against 11.5 in G1. From it, it can be concluded that the 12 variables age, sex, G1, G2, absences, failures, schoolsup, paid, famsup, activities, studytime and higher, will be the ones used in the Shiny predictive tool for a minimal performance if no grades for G2 and G1 are available.

Table 8. Accuracy Results for Machine Learning Techniques

Accuracy Table		lda	logistic	svm	knn	nb	cart	bagging	rf	gbm
Mathematics	Binary	0.849226	0.8863169	0.844624	0.78282	0.813005	0.888616	0.905323	0.91563	0.923446
	4 Level	0.713921		0.659151	0.488877	0.757435	0.747861	0.816988	0.830327	0.824925
	w/o G2&G1						0.333402	0.3626	0.374368	0.390845
Portuguese Language	Binary	0.910984	0.9009121	0.919142	0.919142	0.919142	0.95398	0.950417	0.951528	0.947198
	4 Level	0.74514		0.686876	0.58857	0.667711	0.808855	0.820829	0.861721	0.818571
	w/o G2&G1						0.482328	0.496316	0.492053	0.460235

Clustering has the four clusters divisions with two clusters above 100 students and the other two below 100 students. In both subject's clusters seems to be divided by three clusters with extreme values and one with closer to mean values.

The Principal Component Analysis has a minimum requirement of 8 PC's for an explanation of at least the 80% of the variance with a difference between scree plots where Mathematics has a two elbows plot and Portuguese Language has no obvious elbow. The results of the two first PC's is as well similar in both subjects with slightly over 20% of the variance and the main difference is in school support and age being more representative in Portuguese Language and sex and failures for Mathematics.

Finally, observing the previous work by (Cortez & Silva, 2008) and given the changes done on the dataset such as redundancy reduction, student academic background focus, increased training dataset, 4 level grouping and implementation of new Machine Learning techniques, it can be concluded that the previous grades are essential on the prediction of the final grades. Additionally, the same variables keep having importance after removing grade 2 and 1 except for the ones removed to reduce the number of

variables and redundancy. Furthermore, it has been confirmed that reducing the size of the dataset does not imply substantial differences in the predicting results, however, the caret package has presented better predicting results increasing from 2 to 11% the predicting accuracy depending on the approach selected. Once again the Random Forest and Decision Tree have given the best results outperforming over the linear and non linear approaches.

Overall, presented results suggest that developed models allow predicting the final grade of the students with a reliability between 40 to 90 % considering the inputted variables. Finally, Shiny student performance app is able to determine future performance based on current and historical data of the students and simplify this predictive process to academic professionals.

In addition, the presented analyses could be very useful in a situation such as the one lived nowadays with a covid-19 pandemic that left, a lot of educational centres, without final real grades either the opportunity of concluding the academic year. Novel approaches such as the developed Shiny App are key as easy-friendly tools to be used and support academic professionals on the decision of evaluating final assessments.

Future work would need to improve and automatize the data collection with better retention and automatic techniques implementation for a fast and reliable decisions as well as adding additional subjects and not only the core ones. This could help to introduce craft subjects that are not considered in this project and the additional variables left out for this research that can add robustness to the prediction model. It would, therefore, help teachers, educational centres and companies apply better decisions on the different groups of students, adapt the structure of the subjects or guide students in their future careers or jobs.

6 References

- Amirah M., W. H. (2015). A Review on Predicting Student's Performance Using Data Mining Techniques. *Procedia Computer Science*, Vol.72, 414-422.
- Artlet, C. B.-M. (2003). *Learners for Life. Student Approaches to Learning. Results from PISA 2000*. Paris: Organisation for Economic Cooperation and Development.
- Baker, R. S. (2009). The State of Educational Data Mining in 2009 : A Review and Future Visions. *Journal of Educational Data Mining*. 1. , 3–16.
- Brownlee, J. (2018). *Statistical Methods for Machine Learning: Discover how to Transform Data into Knowledge with Python*.
- Cortez, P., & Silva, A. (2008). Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., *Proceedings of 5th FUBUTEC 2008*, 5-12.
- Davidson, I. (2002). *Understanding K-means non-hierarchical clustering*. New York: SUNY Albany.
- Defays, D. (1977). An efficient algorithm for a complete link method. *The Computer Journal*.
- Eshetu, A. A. (2015). Parental Socio-economic Status as a Determinant Factor on Academic Performance of Students in Regional Examination : Case of Dessie town; Ethiopia. *International Journal of Academic Research in Education and Review*, 3(9), 247-256. Obtenido de <http://www.academicresearchjournals.org/IJARER/Index.htm>
- Eursec.eu. (09 de January de 2020). *Office of the Secretary-General of the European Schools*. Obtenido de <https://www.eursec.eu/en/European-Schools/studies/studies-organisation>
- EURYDICE, E. C. (09 de January de 2020). *European Commission*. Obtenido de https://eacea.ec.europa.eu/national-policies/eurydice/content/historical-development-60_en

6 References

- Fayyad, U. P.-s. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine Volume 17 Number 3*, 37–54.
- Filiz, E. &. (2019). Finding the best algorithms and effective factors in classification of Turkish science student success. *Journal of Baltic Science Education*. 18, 239-253.
- Izquierdo, M. (25 de 05 de 2020). *Dkit Mahara*. Obtenido de <https://mahara.dkit.ie/view/view.php?t=4oMvTpDeuHfWPnRqs6hJ>.
- James, G. W. (2013). *An introduction to statistical learning*. New York: Springer.
- Kassambara, A. (2017). *Practical guide to principal component methods in R: PCA, M (CA), FAMD, MFA, HCPC*. STHDA.
- Kubat, M. (2017). *An introduction to machine learning*. Cham: Springer International .
- Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, vol. 28, 5, 1.
- OECD. (2019). *Education at a Glance 2019: OECD Indicators*. Paris: OECD Publishing.
- Pandey, M. &. (2013). A Decision Tree Algorithm Pertaining to the Student Performance Analysis and Prediction. *International Journal of Computer Applications*. 61., 1-5.
- Ramesh, V. (2013). Predicting Student Performance: A Statistical and Data Mining Approach. *INTERNATIONAL JOURNAL OF COMPUTER APPLICATIONS*. 63., 35–39.
- República, A. d. (27 de August de 2009). “Diário da República” n.º 166. *Law 85/2009*, págs. 5635–5636.
- Schleicher, A. (2018). *Insights and Interpretations, PISA 2018*. OECD.
- SJ, S. (2009). *A Modern Approach to Regression with R*. New York: Springer.
- Soraya Sedkaoui, M. K. (2020). En *Sharing Economy and Big Data Analytics* (pág. 201). John Wiley & Sons.

6 References

- Tadayon, M. &. (2019). *Predicting Student Performance in an Educational Game Using a Hidden Markov Model*.
- Varghese, B. &. (2011). Clustering Student Data to Characterize Performance Patterns. *IJACSA. Special Issue. 10.14569*.
- Ventura, C. R. (2010). Educational Data Mining: A Review of the State of the Art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 40, no. 6, 601-618.

7 Appendices

7.1 Ethics Documentation

ADMINISTRATION DETAILS

Researcher: Magi Izquierdo Lechuga

School/Research Centre/Programme (as applicable) School of Informatics and Creative Media

Title of Project: Analysis and prediction on student lifestyle effect in academic performance

Supervisor/Research Centre Director/Head of Department: Dr. Rajesh Jaiswal

Date: 18/11/19

Type of research			
Undergraduate	Postgraduate	Staff member	External to DKIT
<u>X</u>			

There is an obligation on the lead researcher to bring to the attention of the School Ethics Committee any issues with ethical implications not clearly covered by this application form.

APPLICATION FORM CHECKLIST

Please complete the ethics application form below and provide additional information as attachments.

My application includes the following documentation:	INCLUDED (mark as YES)	NOT APPLICABLE (mark as N/A)
Recruitment advertisement		X
Participant Information Leaflet		X
Participant Informed Consent form		X
Questionnaire/Survey		X
Interview/Focus Group Questions		X
Debriefing material		X
Evidence of approval to gain access to off-site location		X
Ethical Approval from external organisations. If ethical approval from external organisations is pending give details below		X
Details		

PROJECT DETAILS

a) Lay description (Maximum 200 words)

Please outline, in terms that any non-expert would understand, what your research project is about, including what participants will be required to do. Please explain any technical terms or discipline-specific phrases.

As a part-time student in the High Diploma in Science and Data Analytics offered in Dundalk Institute of Technology. The final project of this course consists in developing and analyzing a dataset and study potential interested organizations which could be interested in the idea in terms of providing information and/or taking advantage of the output of the project.

The actual dataset comprises information of 395 Portuguese teenagers about their school marks in Math and Portuguese language, family lifestyle and extra daily activities (32 different variables in total for each case of study). Therefore, I aim to create a model to relate the degree of success (observed with final grades) of teenager students with information about their lifestyles.

It is important to note that the data obtained is anonymized and publically available.

b) Research objectives (Maximum 150 words)

Please summarise briefly the objectives of the research,

Objectives:

- Identification and analysis of response and explanatory academic variables of the dataset.
- Create a predictive model to relate the degree of success of teenager students with information about their lifestyles.
- App development and intended to better understand student performance based on key information about their lifestyles and potentially help to improve student educational success.

c) Research location and duration

Location(s)/Population*	DKIT facilities & Researcher's home
Research start date	26/10/2019
Research end date	30/04/2020
Approximate duration	6 months

* If location/Population other than DKIT campus/population, provide details of the approval to gain access to that location/population as an appendix.

PARTICIPANTS

		YES	NO	N/A
Do participants fall into any of the following special groups?	Minors (under 18 years of age)			X
	People with learning or communication difficulties			X
	Patients			X
	People in custody			X
	People engaged in illegal activities (e.g. drug-taking)			X
Have you given due consideration to the need for satisfactory Garda clearance?				X

SAMPLE DETAILS

Approximate number	NA
Where will participants be recruited from?	NA
Inclusion Criteria	NA
Exclusion Criteria	NA
Will participants be remunerated, and if so in what form? NA	

Justification for proposed sample size and for selecting a specific gender, age, or any other group if this is done in your research. NA

RISKS TO PARTICIPANTS

- a) Please describe any risks to participants that may arise due to the research. Such risks could include physical stress, emotional distress, perceived coercion e.g. lecturer interviewing own students. Detail the measures and considerations you have put in place to minimize these risks
- b) What will you communicate to participants about any identified risks? Will any information be withheld from them about the research purpose or procedure? If so, please justify this decision.

N/A

INFORMED CONSENT

	YES	NO	N/A
Will you obtain active consent for participation?			X
Will you describe the main experimental procedures to participants in advance?			X
Will you inform the participants that their participation is voluntary and may be withdrawn at any point?			X
If the research is observational, will you ask for their consent to being observed?			X
With questionnaires, will you give participants the option of omitting questions they do not want to answer?			X
Will you tell participants that their data will be treated with full confidentiality and that, if published, it will not be identifiable as theirs?			X
Will the data be anonymous?			X
Will you debrief participants at the end of their participation?			X
Will your project involve deliberately misleading participants in any way or will information be withheld? If you answer yes, give details and justification for doing this below.			X

- a) Please outline your approach to ensuring the confidentiality of data (that is, that the data will only be accessible to agreed upon parties and the

safeguarding mechanisms you will put in place to achieve this.) You should include details on how and where the data will be stored, and who will have access to it.

Original dataset is retrieved from “UCI Machine Learning Repository” (<https://archive.ics.uci.edu/ml/datasets/student+performance>) and it has been donated by Paulo Cortez, University of Minho, on 27/11/2014 and it will serve as an old academy data used for analysis.

It is formed by 395 Portuguese teenager students from two different schools from Mathematics and Portuguese language subjects.

Storage of the data will be researcher’s local drive with no file sharing or cloud storage involved.

It is important to note that the data obtained is anonymized and publically available.

b) Please outline how long the data will be retained for, if it will be destroyed and how it will be destroyed.

The retention of the data will be limited to the duration of the project and its related processes.

DECLARATION

I have read and understand the DkIT guidelines for ethical practices in research and have read and understand the data protection guidelines.

Signed:



Name: MAGI IZQUIERDO LECHUGA

Date: 18/11/2019
(Researcher)

Signed:



Name: Rajesh Jaiswal

Date: 26/02/2020
(Supervisor/Research Centre Director/Head of Department)

Ethical Approval Application - Feedback Form	
Application No.	53 Resubmission
Date.	11 March 2020
Applicants Name.	Magi Lzquierdo
Supervisor.	Rajesh Jaiswal
Project Title.	Analysis and prediction on student lifestyle effect in academic performance
Decision. (Approved/ Not Approved/ Approved Subject to indicated Requirements)	Approved
Comments	

7.2 Source Code

7.2.1 Project Code

```
# set directory where .csv files are located
#getwd()
#setwd("C:/Users/jik_6/Documents/R/win-library/3.5")
#read tables in.
d1=read.table("student-mat.csv",sep=";",header=TRUE)
d2=read.table("student-por.csv",sep=";",header=TRUE)

# Merge Tables
d3=merge(d1,d2,by=c("school","sex","age","address","famsize","Pstatus","Medu","Fed
u","Mjob","Fjob","reason","nursery","internet"))
print(nrow(d3)) # 382 students

# Look for NA's or Outliers and remove
#install.packages("tidyverse")
library(tidyverse)

d3 %>%
  summarise(count = sum(is.na(.)))

par(mfrow=c(1,2))

boxplot(d3$age, main= "Boxplot of Age (15 to 22)")
#There are 3 values with outliers. Will remove them as will focus on under/to 18 years
old.

boxplot(d3$absences.x, main="Boxplot of Absence")
plot(d3$absences.x,d3$age, main="Scatterplot of Age and Absence", ylab="Age", xlab
= "Days of absence")
# Any datapoint higher than 20 is considered an outlier. In the scatterplot the most
extreme
#values seems to be above 50. The reason of high absence is not disclosed but will
#treat data as plausible.

boxplot(d3$G1.x, main="Boxplot of G1")

boxplot(d3$G2.x, main="Boxplot of G2")
# has outliers on 0 grades, probably linked to the most extreme value of Absence as it is
close to 60, which
# is almost an schoolar trimester. It is relevant that G1 does not have any outlier.

boxplot(d3$G3.x, main="Boxplot of G3")
# Outliers are grades 0.

plot(d3$absences.x,d3$G2.x, main="Scatterplot of G2 and Absence", ylab="Grade",
xlab = "Days of absence")
```

7 Appendices

```
plot(d3$absences.x,d3$G1.x, main="Scatterplot of G1 and Absence", ylab="Grade",  
xlab = "Days of absence")
```

```
#Remove age outliers
```

```
d3<-filter(d3,d3$age<19)
```

```
str(d3)
```

```
# from 382 to 369 students
```

```
par(mfrow=c(1,3))
```

```
boxplot(d3$age, main= "Boxplot of Age (15 to 22)")
```

```
boxplot(d3$age, main= "Boxplot of Age (15 to 18)")
```

```
# Remove students with 0 Absences and 0 Grading.
```

```
d3<-d3[!(d3$absences.x==0 & d3$G1.x==0 & d3$G2.x==0 & d3$G3.x==0),]
```

```
str(d3)
```

```
# There seem to be some students that have attended the whole semester and graded 0  
but not in all gradings.
```

```
# Probably have not attended the exam in some of the grading periods.
```

```
# Portuguese Language
```

```
dev.off()
```

```
boxplot(d3$absences.y, main="Boxplot of Absence")
```

```
# Any datapoint higher than 15 is considered an outlier. In the scatterplot the most  
extreme
```

```
#values seems to be around 30. The reason of high absence is not disclosed but will
```

```
#treat data as plausible.
```

```
plot(d3$absences.y,d3$age, main="Scatterplot of Age and Absence", ylab="Age", xlab  
= "Days of absence")
```

```
par(mfrow=c(1, 3))
```

```
boxplot(d3$G1.y, main="Boxplot of G1")
```

```
# Has 1 outlier at the 0 value.
```

```
boxplot(d3$G2.y, main="Boxplot of G2")
```

```
# has 1 outlier at each side, one at 5 and the higher one at 19.
```

```
boxplot(d3$G3.y, main="Boxplot of G3")
```

```
#Several outliers below 7 and one higher around 19
```

```
par(mfrow=c(2, 2))
```

```
plot(d3$absences.y,d3$G2.y, main="Scatterplot of G2 and Absence", ylab="Grade",  
xlab = "Days of absence")
```


7 Appendices

```
plot(d3$absences.y,d3$G1.y, main="Scatterplot of G1 and Absence", ylab="Grade",
xlab = "Days of absence")

dev.off()

# Summary Statistics
#Student
summary(d3[1:13])
#Math
summary(d3[14:33])
#Por
summary(d3[34:53])

# Summary Stat MATH

Math_sup<-c(school=d3$schoolsup.x,Family=d3$famsup.x,Paid=d3$paid.x)
addmargins(prop.table(table(Math_sup)))

barplot(table(Math_sup), main = " General Math support",names.arg=c("NO", "YES"))

addmargins(prop.table(table(d3$failures.x)))
addmargins(prop.table(table(d3$absences.x)))
addmargins(prop.table(table(d3$higher.x)))
addmargins(prop.table(table(d3$activities.x)))
addmargins(prop.table(table(d3$romantic.x)))

# Summary Stat Por

Por_sup<-c(school=d3$schoolsup.y,Family=d3$famsup.y,Paid=d3$paid.y)
addmargins(prop.table(table(Por_sup)))

barplot(table(Por_sup),main = " General Por Lang. support",names.arg=c("NO",
"YES"))

addmargins(prop.table(table(d3$failures.y)))
addmargins(prop.table(table(d3$absences.y)))
addmargins(prop.table(table(d3$higher.y)))
addmargins(prop.table(table(d3$activities.y)))
addmargins(prop.table(table(d3$romantic.y)))

# Summary Stat Both

par(mfrow=c(1, 2))
barplot(table(Math_sup), main = " General Math support",names.arg=c("NO", "YES"))
barplot(table(Por_sup),main = " General Por Lang. support",names.arg=c("NO",
"YES"))

counts1 <- table(d3$schoolsup.x)
counts2 <- table(d3$famsup.x)
counts3 <- table(d3$paid.x)
counts4 <- table(d3$schoolsup.y)
```

7 Appendices

```
counts5 <- table(d3$famsup.y)
counts6 <- table(d3$paid.y)
```

```
par(mfrow=c(3, 2))
```

```
barplot(counts1, main="Math School Sup",col=c("darkblue","red"))
barplot(counts4, main="Por School Sup",col=c("darkblue","red"))
barplot(counts2, main="Math Family Sup",col=c("darkblue","red"))
barplot(counts5, main="Por Family Sup",col=c("darkblue","red"))
barplot(counts3, main="Math Paid Sup",col=c("darkblue","red"))
barplot(counts6, main="Por Paid Sup",col=c("darkblue","red"))
```

```
# create the new dataset with Academia and Lifestyle variables and remove outliers.
```

```
d4<-d3 %>%
  select(sex, age, nursery, studytime.x, failures.x, schoolsup.x,
         famsup.x,activities.x,paid.x,higher.x,romantic.x,freetime.x,
         goout.x,Walc.x,Dalc.x,absences.x, G1.x,G2.x,G3.x,studytime.y,
         failures.y, schoolsup.y,famsup.y,activities.y,paid.y,higher.y,
         romantic.y,freetime.y,goout.y,Walc.y,Dalc.y,absences.y, G1.y,G2.y,G3.y)
```

```
summary(d4)
str(d4)
```

```
# Divide Math - Por
```

```
d4M<-d4 %>%
  select(sex, age, nursery, studytime.x, failures.x, schoolsup.x,
         famsup.x,activities.x,paid.x,higher.x,romantic.x,freetime.x,
         goout.x,Walc.x,Dalc.x,absences.x, G1.x,G2.x,G3.x)
```

```
d4P<-d4 %>%
  select(sex, age, nursery, studytime.y,failures.y, schoolsup.y,famsup.y,
         activities.y,paid.y,higher.y,romantic.y,freetime.y,goout.y,Walc.y,
         Dalc.y,absences.y, G1.y,G2.y,G3.y)
```

```
str(d4M)
str(d4P)
```

```
##Mathematics
```

```
# Plot variables (integers to be plotted as follows) dont forget pairs plot for continuous.
str(d4M)
```

```
### transform integers into Factors
```

```
d4M[,"studytime.x"]<-as.numeric(factor(d4M[,"studytime.x"]))
class(d4M$studytime.x)
```

7 Appendices

```
plot(d4M$studytime.x, main="Math Studytime", xlab="Level of Studytime")

d4M[, "failures.x"] <- as.numeric(factor(d4M[, "failures.x"]))
class(d4M$failures.x)
plot(d4M$failures.x, main="Math Past Failures", xlab="Number of failures")

d4M[, "freetime.x"] <- as.numeric(factor(d4M[, "freetime.x"]))
class(d4M$freetime.x)
plot(d4M$freetime.x, main="Math Free Time", xlab="Level of Free Time")

d4M[, "goout.x"] <- as.numeric(factor(d4M[, "goout.x"]))
class(d4M$goout.x)
plot(d4M$goout.x, main="Math Go Out Time", xlab="Go Out Time")

d4M[, "Walc.x"] <- as.numeric(factor(d4M[, "Walc.x"]))
class(d4M$Walc.x)
plot(d4M$Walc.x, main="Math Weekend Alcohol Consumption", xlab="Level of
consumption")

d4M[, "Dalc.x"] <- as.numeric(factor(d4M[, "Dalc.x"]))
class(d4M$Dalc.x)
plot(d4M$Dalc.x, main="Math Daily Alcohol Consumption", xlab="Level of
consumption")

str(d4M)

d4M$age <- as.numeric(d4M$age)
d4M$absences.x <- as.numeric(d4M$absences.x)
d4M$G1.x <- as.numeric(d4M$G1.x)
d4M$G2.x <- as.numeric(d4M$G2.x)
d4M$G3.x <- as.numeric(d4M$G3.x)

### Plotting Maths ###
dev.off()

d4Mc <- d4M %>%
  select(age, absences.x, G1.x, G2.x, G3.x)
pairs(d4Mc)
cor(d4Mc)

#install.packages("PerformanceAnalytics")
library(PerformanceAnalytics)
chart.Correlation(d4Mc)

boxplot(d4M$G3.x ~ d4M$sex, main= "Boxplot of Math Sex", xlab= "Sex",
ylab="Grade")
# Light difference with slightly higher grade for boys.
boxplot(d4M$G3.x ~ d4M$age, main= "Boxplot of Math Age", xlab= "Age",
ylab="Grade")
#Age seems to affect negatively to the grade as median decreases gradually until 19 yo
were almost
```

```

# nobody pass the subject.
boxplot(d4M$G3.x~d4M$nursery, main= "Boxplot of Math Nursery", xlab= "Attended
Nursery", ylab="Grade")
# Does not seem to be significant difference
boxplot(d4M$G3.x~d4M$studytime.x, main= "Boxplot of Math Studytime", xlab=
"Level of Studytime", ylab="Grade")
# As more time spendend studing seems to perform better but the difference does not
seem relevant.
boxplot(d4M$G3.x~d4M$failures.x, main= "Boxplot of Math Failures", xlab= "n of
previous failures", ylab="Grade")
# Having previous failed classes decreases the grade.
boxplot(d4M$G3.x~d4M$schoolsup.x, main= "Boxplot of Math School Support",
ylab="Grade")
#School support has a less variance of grade for students who has it but it does not seem
to improve the grade.
boxplot(d4M$G3.x~d4M$famsup.x, main= "Boxplot of Math Family Support",
ylab="Grade")
# Yes and NO are almost the same.
boxplot(d4M$G3.x~d4M$paid.x, main= "Boxplot of Math Paid Support",
ylab="Grade")
#Does not seem to be a big difference between yes and no.
boxplot(d4M$G3.x~d4M$activities.x, main= "Boxplot of Math Activities",
ylab="Grade")
# grades for "Yes" are slightly higher.
boxplot(d4M$G3.x~d4M$higher.x, main= "Boxplot of Math Higher Education",
ylab="Grade", xlab = "")
# There is a considerable difference between students who want to go to Higher
Education or not.
boxplot(d4M$G3.x~d4M$romantic.x, main= "Boxplot of Math Romantic",
ylab="Grade")
# NOt having a relationship helps slightly to have better marks.
boxplot(d4M$G3.x~d4M$freetime.x, main= "Boxplot of Math Freetime", xlab= "Level
of Freetime", ylab="Grade")
# Level 2 on Freetime seems to be a bit higher but not much.
boxplot(d4M$G3.x~d4M$goout.x, main= "Boxplot of Math Goout", xlab= "Level of
Goout", ylab="Grade")
# Having a mid level of going out has a positive effect on the G3 variable respect not
going out at all
# or going out very frequently.
boxplot(d4M$G3.x~d4M$Walc.x, main= "Boxplot of Math Weekend alcohol", xlab=
"Level of consumption", ylab="Grade")
# Does not seem to have a large effect but the range of grades is decreased by the
highest grades.
boxplot(d4M$G3.x~d4M$Dalc.x, main= "Boxplot of Math Daily alcohol", xlab=
"Level of consumption", ylab="Grade")
# IT affect on the highest grades but medians are almost the same.
plot(d4M$G3.x~d4M$absences.x, main= "Grade by Absences", ylab="Grade", xlab="n
of Absences")
# Grade has a negative trend
plot(d4M$G3.x~d4M$G1.x, main= "Grade 1 vs Grade 3", ylab="Grade3", xlab="Grade
1")
# Positive Trend with some values=0

```

```
plot(d4M$G3.x~d4M$G2.x, main= "Grade 2 vs Grade 3", ylab="Grade3", xlab="Grade
2")
# Positive Trend with some values=0
```

```
par(mfrow=c(1,3))
boxplot(d4M$G3.x~d4M$schoolsup.x, ylab="Grade", xlab = "School Support")
boxplot(d4M$G3.x~d4M$famsup.x, main= "Mathematics Supports", ylab="Grade",
xlab = "Family Support")
boxplot(d4M$G3.x~d4M$paid.x, ylab="Grade", xlab = "Paid Support")
dev.off()
```

```
#### Portuguese ####
```

```
# transform integers into Factors
```

```
d4P[, "studytime.y"]<-as.numeric(factor(d4P[, "studytime.y"]))
class(d4P$studytime.y)
plot(d4P$studytime.y, main="Port Studytime", xlab="Level of Studytime")
```

```
d4P[, "failures.y"]<-as.numeric(factor(d4P[, "failures.y"]))
class(d4P$failures.y)
plot(d4P$failures.y, main="Port Past Failures", xlab="-Number of failures")
```

```
d4P[, "freetime.y"]<-as.numeric(factor(d4P[, "freetime.y"]))
class(d4P$freetime.y)
plot(d4P$freetime.y, main="Port Free Time", xlab="Level of Free Time")
```

```
d4P[, "goout.y"]<-as.numeric(factor(d4P[, "goout.y"]))
class(d4P$goout.y)
plot(d4P$goout.y, main="Port Go Out Time", xlab="Go Out Time")
```

```
d4P[, "Walc.y"]<-as.numeric(factor(d4P[, "Walc.y"]))
class(d4P$Walc.y)
plot(d4P$Walc.y, main="Port Weekend Alcohol Consupction", xlab="Level of
consumption")
```

```
d4P[, "Dalc.y"]<-as.numeric(factor(d4P[, "Dalc.y"]))
class(d4P$Dalc.y)
plot(d4P$Dalc.y, main="Port Daily Alcohol Consupction", xlab="Level of
consumption")
```

```
str(d4P)
```

```
d4P$age<-as.numeric(d4P$age)
d4P$absences.y<-as.numeric(d4P$absences.y)
d4P$G1.y<-as.numeric(d4P$G1.y)
d4P$G2.y<-as.numeric(d4P$G2.y)
d4P$G3.y<-as.numeric(d4P$G3.y)
```

Plotting Portuguese Language

```
d4Pc<-d4P %>%
```

```
  select(age,absences.y,G1.y,G2.y,G3.y)
```

```
  pairs(d4Pc)
```

```
  cor(d4Pc)
```

```
  chart.Correlation(d4Pc)
```

```
boxplot(d4P$G3.y~d4P$sex, main= "Boxplot of Portuguese Lang. Sex", xlab= "Sex",
  ylab="Grade")
```

```
# Light difference with slightly higher grade for girls.
```

```
boxplot(d4P$G3.y~d4P$age, main= "Boxplot of Portuguese Lang. Age", xlab= "Age",
  ylab="Grade")
```

```
#Age seems to affect positively to the grade as median increase gradually. Exception is
for 19 and 22
```

```
# where both are lower than the rest.
```

```
boxplot(d4P$G3.y~d4P$nursery, main= "Boxplot of Portuguese Lang. Nursery", xlab=
  "Attended Nursery", ylab="Grade")
```

```
# Does not seem to be significant difference
```

```
boxplot(d4P$G3.y~d4P$studytime.y, main= "Boxplot of Portuguese Lang. Studytime",
  xlab= "Level of Studytime", ylab="Grade")
```

```
# As more time spendened studing seems to perform better.
```

```
boxplot(d4P$G3.y~d4P$failures.y, main= "Boxplot of Portuguese Lang. Failures",
  xlab= "n of previous failures", ylab="Grade")
```

```
# Having previous failed classes decreases the grade.
```

```
boxplot(d4P$G3.y~d4P$schoolsup.y, main= "Boxplot of Portuguese Lang. School
  Support", ylab="Grade")
```

```
#School support has a less variance of grade for students who has it but it does not seem
to improve the grade.
```

```
boxplot(d4P$G3.y~d4P$famsup.y, main= "Boxplot of Portuguese Lang. Family
  Support", ylab="Grade")
```

```
# Yes and NO are almost the same.
```

```
boxplot(d4P$G3.y~d4P$paid.y, main= "Boxplot of Portuguese Lang. Paid Support",
  ylab="Grade")
```

```
#Does not seem to be a big difference between yes and no.
```

```
boxplot(d4P$G3.y~d4P$activities.y, main= "Boxplot of Portuguese Lang. Activities",
  ylab="Grade")
```

```
# grades for "Yes" are slightly higher.
```

```
boxplot(d4P$G3.y~d4P$higher.y, main= "Boxplot of Portuguese Lang. Higher
  Education", ylab="Grade", xlab = "")
```

```
# There is a considerable difference between students who want to go to Higher
Education or not.
```

```
boxplot(d4P$G3.y~d4P$romantic.y, main= "Boxplot of Portuguese Lang. Romantic",
  ylab="Grade")
```

```
# It does not affect much the fact of having or not a relationship.
```

```
boxplot(d4P$G3.y~d4P$freetime.y, main= "Boxplot of Portuguese Lang. Freetime",
  xlab= "Level of Freetime", ylab="Grade")
```

```
# Level 2 on Freetime seems to be a bit higher but not much.
```

```

boxplot(d4P$G3.y~d4P$goout.y, main= "Boxplot of Portuguese Lang. Goout", xlab=
"Level of Goout", ylab="Grade")
# Having a low/mid level of going out has a positive effect on the G3 variable respect
not going out at all
# or going out very frequently.
boxplot(d4P$G3.y~d4P$Walc.y, main= "Boxplot of Portuguese Lang. Weekend
alcohol", xlab= "Level of consumption", ylab="Grade")
# Does not seem to have a large effect but the tendency is to decrease with the level of
consumption.
boxplot(d4P$G3.y~d4P$Dalc.y, main= "Boxplot of Portuguese Lang. Daily alcohol",
xlab= "Level of consumption", ylab="Grade")
# IT affect on the highest grades but medians are almost the same.
plot(d4P$G3.y~d4P$absences.y, main= "Grade by Absences", ylab="Grade", xlab="n
of Absences")
# Grade has a negative trend
plot(d4P$G3.y~d4P$G1.y, main= "Grade 1 vs Grade 3", ylab="Grade3", xlab="Grade
1")
# Positive Trend with some values=0
plot(d4P$G3.y~d4P$G2.y, main= "Grade 2 vs Grade 3", ylab="Grade3", xlab="Grade
2")
# Positive Trend with some values=0 for G3

### Compraison ###
par(mfrow=c(1,2))

#CORRELATION
#install.packages("corrplot")
library(corrplot)

chart.Correlation(d4Mc)
chart.Correlation(d4Pc)

Cor_Math<-cor(data.frame(model.matrix(~.-1, data=(d4M))))
Cor_Port<-cor(data.frame(model.matrix(~.-1, data=(d4P))))
dev.off()
corrplot(Cor_Math, method="color", type="upper",order="hclust",addCoef.col =
"black",
         tl.cex = 0.75,number.cex = 0.6,cl.cex = 1,diag=FALSE)
corrplot(Cor_Port,method="color", type="upper",order="hclust",addCoef.col = "black",
         tl.cex = 0.75,number.cex = 0.6,cl.cex = 1,diag=FALSE)

# SEX
boxplot(d4M$G3.x~d4M$sex, main= "Mathematics ", xlab= "Gender", ylab="Grade")
boxplot(d4P$G3.y~d4P$sex, main= "Portuguese Lang. ", xlab=
"Gender",ylab="Grade")

# AGE
boxplot(d4M$G3.x~d4M$age, main= "Mathematics", xlab= "Age", ylab="Grade")
boxplot(d4P$G3.y~d4P$age, main= "Portuguese Lang.", xlab= "Age", ylab="Grade")

```

7 Appendices

```
# STDYTIME
boxplot(d4M$G3.x~d4M$studytime.x, main= "Mathematics", xlab= "Level of
Studytime", ylab="Grade")
boxplot(d4P$G3.y~d4P$studytime.y, main= "Portuguese Lang.", xlab= "Level of
Studytime", ylab="Grade")

# FAILURES
boxplot(d4M$G3.x~d4M$failures.x, main= "Mathematics", xlab= "n of previous
failures", ylab="Grade")
boxplot(d4P$G3.y~d4P$failures.y, main= "Portuguese Lang.", xlab= "n of previous
failures", ylab="Grade")

# ROMANTIC
boxplot(d4M$G3.x~d4M$romantic.x, main= "Boxplot of Math Romantic",
ylab="Grade")
boxplot(d4P$G3.y~d4P$romantic.y, main= "Boxplot of Portuguese Lang. Romantic",
ylab="Grade")

##### Analysis Mathematics Regression
#####
#install.packages("car")
#install.packages("mlbench")
#install.packages("caret")

# load libraries
library(mlbench)
library(caret)
library(car)

d4M<-d4M %>%
  mutate(studytime.x=as.factor(studytime.x),
         failures.x=as.factor(failures.x),
         freetime.x =as.factor(freetime.x ),
         goout.x=as.factor(goout.x),
         Walc.x =as.factor(Walc.x),
         Dalc.x =as.factor(Dalc.x ))
str(d4M)

inTrain = createDataPartition(y = d4M$G3.x, p = .80, list = FALSE)
Mathtrain = d4M[inTrain,]
Mathtest = d4M[-inTrain,]

str(Mathtrain)
str(Mathtest)

MathReg<-lm(G3.x~.,data=Mathtrain)
summary(MathReg)
plot(MathReg)
```



```
MathReg1<-lm(G3.x~.-famsup.x,data=Mathtrain)
summary(MathReg1)
plot(MathReg1)
```

```
MathReg2<-lm(G3.x~.-famsup.x-paid.x-freetime.x-goout.x,data=Mathtrain)
summary(MathReg2)
plot(MathReg2)
```

```
MathReg3<-lm(G3.x~.-famsup.x-paid.x-freetime.x-goout.x-sex-age-
nursery,data=Mathtrain)
summary(MathReg3)
plot(MathReg3)
```

```
MathReg4<-lm(G3.x~.-famsup.x-paid.x-freetime.x-goout.x-sex-age-
nursery,data=Mathtrain)
summary(MathReg4)
plot(MathReg4)
```

```
MathReg5<-lm(G3.x~.- freetime.x - nursery - sex - paid.x -
studytime.x - famsup.x - goout.x - age - schoolsup.x - Walc.x -
Dalc.x - factor(failures.x, exclude = c(3,4)), data = Mathtrain)
summary(MathReg5)
plot(MathReg5)
```

```
MathReg6<-lm(G3.x~ G2.x + G1.x + absences.x+romantic.x,data=Mathtrain)
summary(MathReg6)
plot(MathReg6)
```

```
par(mfrow=c(2,1))
```

```
anova(MathReg,MathReg1,MathReg2,MathReg3,MathReg4,MathReg6)
AIC(MathReg,MathReg1,MathReg2,MathReg3,MathReg4,MathReg5,MathReg6)
BIC(MathReg,MathReg1,MathReg2,MathReg3,MathReg4,MathReg5,MathReg6)
vif(MathReg6)
```

```
# Assumptions of normality are not met in regression. The best results are accomplished
when removing all
# non academic variables except romantic situation.
library(stats)
```

```
pred<-predict(MathReg6,Mathtest)
round_pred<-round(pred)
# Predict Accuracy
sum(round_pred==Mathtest$G3.x)/72*100
#34.72%
```

```
# Confusion Matrix
table(round_pred,Mathtest$G3.x)
```

```
##### Analysis Portuguese Regression
#####
```

7 Appendices

```
d4P<-d4P %>%
  mutate(studytime.y=as.factor(studytime.y),
         failures.y=as.factor(failures.y),
         freetime.y =as.factor(freetime.y),
         goout.y=as.factor(goout.y),
         Walc.y =as.factor(Walc.y),
         Dalc.y =as.factor(Dalc.y ))
str(d4P)

inTrain = createDataPartition(y = d4P$G3.y, p = .80, list = FALSE)
Portrain = d4P[inTrain,]
Porttest = d4P[-inTrain,]

str(Portrain)
str(Porttest)

PortReg<-lm(G3.y~.,data=Portrain)
summary(PortReg)
plot(PortReg)

PortReg1<-lm(G3.y~.-freetime.y-age,data=Portrain)
summary(PortReg1)
plot(PortReg1)

PortReg2<-lm(G3.y~.-freetime.y-nursery-sex-paid.y-studytime.y-famsup.y-age-
schoolsupt.y-Walc.y-Dalc.y
+activities.y*higher.y+absences.y*G2.y,data=Portrain)
summary(PortReg2)
plot(PortReg2)

PortReg3<-lm(G3.y~ G2.y +higher.y*activities.y+failures.y,data=Portrain)
summary(PortReg3)
plot(PortReg3)

PortReg4<-lm(G3.y~ G2.y+factor(failures.y, exclude = c(4)),data=Portrain)
summary(PortReg4)
plot(PortReg4)

par(mfrow=c(2,1))

anova(PortReg,PortReg1,PortReg2,PortReg3,PortReg4)
AIC(PortReg,PortReg1,PortReg2,PortReg3,PortReg4)
BIC(PortReg,PortReg1,PortReg2,PortReg3,PortReg4)
vif(PortReg4)

# Assumptions of normality are not met in regression.
#Surprisingly, G1.y is not statistically significant
```

7 Appendices

```
pred<-predict(PortReg4,Porttest)
pred
predround<-round(pred)

predround[is.na(predround)] <- 0

plot(predround,Porttest$G3.y)
abline(a=0,b=1)

summary(pred)
summary(Porttest$G3.y)

lmtest<-lm(pred~Porttest$G3.y)
summary(lmtest)

sum((predround==Porttest$G3.y)==TRUE)
sum((predround==Porttest$G3.y)==FALSE)
table()

table1<-data.frame(predround,Porttest$G3.y)
table1

sum(predround==Porttest$G3.y)/73*100
#predict accuracy of 52.05%

# check predicting error
with(Porttest, table(pred, G3.y))
```

```
##### CARET REGRESSION
MATHEMATICS #####
```

```
# Create test and training data
```

```
inTrain = createDataPartition(y = d4M$G3.x, p = .80, list = FALSE)
training = d4M[inTrain,]
```

7 Appendices

```
testing = d4M[-inTrain,]

# rename dataset to keep code below generic
dataset <- training
str(dataset)

## control and seed
control <- trainControl(method="repeatedcv", number=10, repeats=3)
seed <- 7

# Accuracy
metric <- "Accuracy"

# Preproces
preProcess=c("center", "scale")

#### models

# Logistic Regression
set.seed(seed)
fit.glm <- train(G3.x~., data=dataset, method="glm", trControl=control)
# SVM Radial
set.seed(seed)
fit.svmRadial <- train(G3.x~., data=dataset, method="svmRadial", preProc=c("center",
"scale"), trControl=control, fit=FALSE)
# kNN
set.seed(seed)
fit.knn <- train(G3.x~., data=dataset, method="knn", preProc=c("center", "scale"),
trControl=control)
# CART
set.seed(seed)
fit.cart <- train(G3.x~., data=dataset, method="rpart", trControl=control)
# Bagged CART
set.seed(seed)
fit.treebag <- train(G3.x~., data=dataset, method="treebag", trControl=control)
# Random Forest
set.seed(seed)
fit.rf <- train(G3.x~., data=dataset, method="rf", trControl=control)
# Stochastic Gradient Boosting (Generalized Boosted Modeling)
set.seed(seed)
fit.gbm <- train(G3.x~., data=dataset, method="gbm", trControl=control,
verbose=FALSE)
?train

#BayesNaive cannot be run for regression

####

results <-
resamples(list(logistic=fit.glm,SVM=fit.svmRadial,KNN=fit.knn,BAGGED=fit.treebag,
rf=fit.rf, BOOSTED=fit.gbm))
# Table comparison
```

7 Appendices

```
summary(results)
```

```
###
```

```
# boxplot comparison  
bwplot(results)  
# Dot-plot comparison  
dotplot(results)
```

```
## nice plot #  
#install.packages("rattle")  
#install.packages("rpart")  
library(rpart)  
library(rattle)  
library(rpart.plot)  
library(RColorBrewer)  
dev.off()
```

```
fancyRpartPlot(fit.cart$finalModel)  
summary(fit.cart)
```

```
# PREDICT
```

```
Pred<-predict(fit.rf, newdata = testing)  
PredRound<-round(Pred)  
sum(PredRound==testing$G3.x)/72*100
```

```
# Prediction Accuracy of 44.44%
```

```
##### CARET REGRESSION  
PORTUGUESE #####
```

```
inTrain = createDataPartition(y = d4P$G3.y, p = .80, list = FALSE)  
training = d4P[inTrain,]  
testing = d4P[-inTrain,]
```

```
# rename dataset to keep code below generic  
dataset <- training
```

```
str(dataset)
```

```
### models
```

```
# Logistic Regression
```

```
set.seed(seed)
```

```
fit.glm <- train(G3.y~., data=dataset, method="glm", trControl=control)
```

```
# SVM Radial
```

```
set.seed(seed)
```

```
fit.svmRadial <- train(G3.y~., data=dataset, method="svmRadial", preProc=c("center",  
"scale"), trControl=control, fit=FALSE)
```

```
# kNN
```

```
set.seed(seed)
```

```
fit.knn <- train(G3.y~., data=dataset, method="knn", preProc=c("center", "scale"),  
trControl=control)
```

```
# CART
```

```
set.seed(seed)
```

```
fit.cart <- train(G3.y~., data=dataset, method="rpart", trControl=control)
```

```
# Bagged CART
```

```
set.seed(seed)
```

```
fit.treebag <- train(G3.y~., data=dataset, method="treebag", trControl=control)
```

```
# Random Forest
```

```
set.seed(seed)
```

```
fit.rf <- train(G3.y~., data=dataset, method="rf", trControl=control)
```

```
# Stochastic Gradient Boosting (Generalized Boosted Modeling)
```

```
set.seed(seed)
```

```
fit.gbm <- train(G3.y~., data=dataset, method="gbm", trControl=control,  
verbose=FALSE)
```

```
#BayesNaive cannot be run for regression
```

```
###
```

```
results
```

```
<-
```

```
resamples(list(logistic=fit.glm,SVM=fit.svmRadial,KNN=fit.knn,BAGGED=fit.treebag,  
rf=fit.rf, BOOSTED=fit.gbm))
```

```
# Table comparison
```

```
summary(results)
```

```
###
```

```
# boxplot comparison
```

```
bwplot(results)
```

```
# Dot-plot comparison
```

```
dotplot(results)
```

```
fancyRpartPlot(fit.cart$finalModel)
```

```
summary(fit.cart)
```

```
#PREDICT
```

```
Pred<-predict(fit.rf, newdata = testing)
PredRound<-round(Pred)
sum(PredRound==testing$G3.y)/73*100
```

```
# Prediction Accuracy of 54.79%
```

```
##### CARET BINARY MATHEMATICS
#####
```

```
# Create Binary G3.x
Binaryd4M<-d4M
Binaryd4M$G3.x<-factor(ifelse(Binaryd4M$G3.x >= 10, "Pass", "Fail"))
```

```
str(Binaryd4M)
# 2= Pass , 1= Fail
```

```
#Create Train and Test Datasets
```

```
inTrain = createDataPartition(y = Binaryd4M$G3.x, p = .80, list = FALSE)
training = Binaryd4M[inTrain,]
testing = Binaryd4M[-inTrain,]
```

```
# rename dataset to keep code below generic
dataset <- training
str(dataset)
```

```
# MODELS
```

```
# Linear Discriminant Analysis
set.seed(seed)
fit.lda <- train(G3.x~., data=dataset, method="lda", metric=metric, preProc=c("center",
"scale"), trControl=control)
# Logistic Regression
```

7 Appendices

```
set.seed(seed)
fit.glm <- train(G3.x~., data=dataset, method="glm", metric=metric, trControl=control)
# SVM Radial
set.seed(seed)
fit.svmRadial <- train(G3.x~., data=dataset, method="svmRadial", metric=metric,
preProc=c("center", "scale"), trControl=control, fit=FALSE)
# kNN
set.seed(seed)
fit.knn <- train(G3.x~., data=dataset, method="knn", metric=metric,
preProc=c("center", "scale"), trControl=control)
# Naive Bayes
set.seed(seed)
fit.nb <- train(G3.x~., data=dataset, method="nb", metric=metric, trControl=control)
# CART
set.seed(seed)
fit.cart <- train(G3.x~., data=dataset, method="rpart", metric=metric, trControl=control)
# Bagged CART
set.seed(seed)
fit.treebag <- train(G3.x~., data=dataset, method="treebag", metric=metric,
trControl=control)
# Random Forest
set.seed(seed)
fit.rf <- train(G3.x~., data=dataset, method="rf", metric=metric, trControl=control)
# Stochastic Gradient Boosting (Generalized Boosted Modeling)
set.seed(seed)
fit.gbm <- train(G3.x~., data=dataset, method="gbm", metric=metric, trControl=control,
verbose=FALSE)

# Check models
results <- resamples(list(lda=fit.lda, logistic=fit.glm, svm=fit.svmRadial, knn=fit.knn,
nb=fit.nb, cart=fit.cart,
                        bagging=fit.treebag, rf=fit.rf, gbm=fit.gbm))
# Table comparison
summary(results)

# boxplot comparison
bwplot(results)
# Dot-plot comparison
dotplot(results)

# Boosting best results at the price of long procesing time.

# LOGISTIC, CART, NB, KNN,SVM and LDA lower performance. RF, CART and
LDA have outliers.

fancyRpartPlot(fit.cart$finalModel)
summary(fit.cart)

#PREDICT

Pred<-predict(fit.gbm, newdata = testing)
sum(Pred==testing$G3.x)/73*100
```


7 Appendices

Prediction Accuracy of 89.04%

```
Pred<-predict(fit.rf, newdata = testing)
```

```
sum(Pred==testing$G3.x)/73*100
```

Prediction Accuracy of 90.41%

#• Random forest has a better accuracy when compared to the test data.

```
##### CARET BINARY PORTUGUESE  
#####
```

```
# Create Binary G3.y
```

```
Binaryd4P<-d4P
```

```
Binaryd4P$G3.y<-factor(ifelse(Binaryd4P$G3.y >= 10, "Pass", "Fail"))
```

```
str(Binaryd4P)
```

```
# 2= Pass , 1= Fail
```

```
inTrain = createDataPartition(y = Binaryd4P$G3.y, p = .80, list = FALSE)
```

```
training = Binaryd4P[inTrain,]
```

```
testing = Binaryd4P[-inTrain,]
```

```
# rename dataset to keep code below generic
```

```
dataset <- training
```

```
str(dataset)
```

```
# MODELS
```

```
# Linear Discriminant Analysis
```

```
set.seed(seed)
```

```
fit.lda <- train(G3.y~., data=dataset, method="lda", metric=metric, preProc=c("center",  
"scale"), trControl=control)
```

```
# Logistic Regression
```

```
set.seed(seed)
```

```
fit.glm <- train(G3.y~., data=dataset, method="glm", metric=metric, trControl=control)
```

```
# SVM Radial
```

```
set.seed(seed)
```

```
fit.svmRadial <- train(G3.y~., data=dataset, method="svmRadial", metric=metric,  
preProc=c("center", "scale"), trControl=control, fit=FALSE)
```

```
# kNN
```

```
set.seed(seed)
```

7 Appendices

```
fit.knn <- train(G3.y~., data=dataset, method="knn", metric=metric,
preProc=c("center", "scale"), trControl=control)
# Naive Bayes
set.seed(seed)
fit.nb <- train(G3.y~., data=dataset, method="nb", metric=metric, trControl=control)
# CART
set.seed(seed)
fit.cart <- train(G3.y~., data=dataset, method="rpart", metric=metric, trControl=control)
# Bagged CART
set.seed(seed)
fit.treebag <- train(G3.y~., data=dataset, method="treebag", metric=metric,
trControl=control)
# Random Forest
set.seed(seed)
fit.rf <- train(G3.y~., data=dataset, method="rf", metric=metric, trControl=control)
# Stochastic Gradient Boosting (Generalized Boosted Modeling)
set.seed(seed)
fit.gbm <- train(G3.y~., data=dataset, method="gbm", metric=metric, trControl=control,
verbose=FALSE)

# Check models
results <- resamples(list(lda=fit.lda, logistic=fit.glm, svm=fit.svmRadial, knn=fit.knn,
nb=fit.nb, cart=fit.cart,
                        bagging=fit.treebag, rf=fit.rf, gbm=fit.gbm))
# Table comparison
summary(results)

# boxplot comparison
bwplot(results)
# Dot-plot comparison
dotplot(results)

# Classification Tree best results.

# NB, KNN,SVM, GLM and LDA poor performance.

fancyRpartPlot(fit.cart$finalModel)
summary(fit.cart)

#PREDICT

Pred<-predict(fit.gbm, newdata = testing)
sum(Pred==testing$G3.y)/73*100
# Prediction Accuracy of 98.63%

Pred<-predict(fit.rf, newdata = testing)
sum(Pred==testing$G3.y)/73*100
# Prediction Accuracy of 95.89%

Pred<-predict(fit.cart, newdata = testing)
sum(Pred==testing$G3.y)/73*100
# Prediction Accuracy of 95.89%
```

##!!When comparthg predictions with test dataset, Boosting is the most accurate technique. Random Forest does not
improve the accuracy for the Decision Tree.

Portuguese Language has a Higher Rate of Pass in comparison with Mathematics. The Binary division does not
gives much information as almost all the students pass both subjects. Will create a new 4 Level division
to better clasify the students and create 2 groups of performance below and above 10.
Groups will be:
High Fail $0 < 7$
Low Fail $7 < 10$
Low Pass $10 < 13$
High Pass $13 < 20$

High Fail and High Pass are bigger groups as it is expected to have less students as more extreme is the value.

CARET 4LEVEL MATHEMATICS
#####

Create 4Levels G3.x
Levels4M<-d4M
Levels4M\$G3.x<-cut(Levelsd4M\$G3.x, br=c(-1,7,10,13,20), labels =
c("HighFail","LowFail","LowPass","HighPass"))

Levels4M\$G3.x
str(Levelsd4M)
1=HighFail, 2=LowFail, 3=LowPass, 4=HighPass

#Create Training and Test datasets
inTrain = createDataPartition(y = Levels4M\$G3.x, p = .80, list = FALSE)
training = Levels4M[inTrain,]
testing = Levels4M[-inTrain,]

rename dataset to keep code below generic
dataset <- training
str(dataset)

MODELS

Linear Discriminant Analysis

7 Appendices

```
set.seed(seed)
fit.lda <- train(G3.x~., data=dataset, method="lda", metric=metric, preProc=c("center",
"scale"), trControl=control)
# SVM Radial
set.seed(seed)
fit.svmRadial <- train(G3.x~., data=dataset, method="svmRadial", metric=metric,
preProc=c("center", "scale"), trControl=control, fit=FALSE)
# kNN
set.seed(seed)
fit.knn <- train(G3.x~., data=dataset, method="knn", metric=metric,
preProc=c("center", "scale"), trControl=control)
# Naive Bayes
set.seed(seed)
fit.nb <- train(G3.x~., data=dataset, method="nb", metric=metric, trControl=control)
# CART
set.seed(seed)
fit.cart <- train(G3.x~., data=dataset, method="rpart", metric=metric, trControl=control)
# Bagged CART
set.seed(seed)
fit.treebag <- train(G3.x~., data=dataset, method="treebag", metric=metric,
trControl=control)
# Random Forest
set.seed(seed)
fit.rf <- train(G3.x~., data=dataset, method="rf", metric=metric, trControl=control)
# Stochastic Gradient Boosting (Generalized Boosted Modeling)
set.seed(seed)
fit.gbm <- train(G3.x~., data=dataset, method="gbm", metric=metric, trControl=control,
verbose=FALSE)

# Check models
results <- resamples(list(lda=fit.lda, svm=fit.svmRadial, knn=fit.knn, nb=fit.nb,
cart=fit.cart,
                        bagging=fit.treebag, rf=fit.rf, gbm=fit.gbm))
# Table comparison
summary(results)

# boxplot comparison
bwplot(results)
# Dot-plot comparison
dotplot(results)

# Logistic has been deleted as predictive variables are not continuous, therefore cannot
run the model.
# Boosting and RF takes a long processing time +10 sec
# Best Performer is RF along with GBM.

fancyRpartPlot(fit.cart$finalModel)
summary(fit.cart)

Pred<-predict(fit.gbm, newdata = testing)
sum(Pred==testing$G3.x)/73*100
```

7 Appendices

```
# Prediction Accuracy of 84.93%
```

```
Pred<-predict(fit.rf, newdata = testing)
sum(Pred==testing$G3.x)/73*100
# Prediction Accuracy of 80.82%
```

```
Pred<-predict(fit.cart, newdata = testing)
sum(Pred==testing$G3.x)/73*100
# Prediction Accuracy of 71.23%
```

```
# Summary shows better accuracy on RF but GBM has the best Accuracy when
compared with the test data.
```

```
summary(fit.gbm)
```

```
# GBM main influence variables are: G2,G1,Absences, Age, Walc, Activities and
Failures.
```

```
# Save predictive model for Shiny purposes
```

```
# saveRDS(fit.gbm,file="Math_Pred.rda")
```

```
##### CARET 4LEVEL PORTUGUESE
#####
```

```
# Create 4Levels G3.y
```

```
Levels4P<-d4P
```

```
Levels4P$G3.y<-cut(Levelsd4P$G3.y, br=c(-1,7,10,13,20), labels =
c("HighFail","LowFail","LowPass","HighPass"))
```

```
Levels4P$G3.y
```

```
str(Levelsd4P)
```

```
# 1=HighFail, 2=LowFail, 3=LowPass, 4=HighPass
```

```
#Create Training and Test datasets
```

```
inTrain = createDataPartition(y = Levels4P$G3.y, p = .80, list = FALSE)
```

```
training = Levels4P[inTrain,]
```

```
testing = Levels4P[-inTrain,]
```

```
# rename dataset to keep code below generic
```

```
dataset <- training
```

```
str(dataset)
```

```
# MODELS
```

```
# Linear Discriminant Analysis
```

7 Appendices

```
set.seed(seed)
fit.lda <- train(G3.y~., data=dataset, method="lda", metric=metric, preProc=c("center",
"scale"), trControl=control)
# SVM Radial
set.seed(seed)
fit.svmRadial <- train(G3.y~., data=dataset, method="svmRadial", metric=metric,
preProc=c("center", "scale"), trControl=control, fit=FALSE)
# kNN
set.seed(seed)
fit.knn <- train(G3.y~., data=dataset, method="knn", metric=metric,
preProc=c("center", "scale"), trControl=control)
# Naive Bayes
set.seed(seed)
fit.nb <- train(G3.y~., data=dataset, method="nb", metric=metric, trControl=control)
# CART
set.seed(seed)
fit.cart <- train(G3.y~., data=dataset, method="rpart", metric=metric, trControl=control)
# Bagged CART
set.seed(seed)
fit.treebag <- train(G3.y~., data=dataset, method="treebag", metric=metric,
trControl=control)
# Random Forest
set.seed(seed)
fit.rf <- train(G3.y~., data=dataset, method="rf", metric=metric, trControl=control)
# Stochastic Gradient Boosting (Generalized Boosted Modeling)
set.seed(seed)
fit.gbm <- train(G3.y~., data=dataset, method="gbm", metric=metric, trControl=control,
verbose=FALSE)

# Check models
results <- resamples(list(lda=fit.lda, svm=fit.svmRadial, knn=fit.knn, nb=fit.nb,
cart=fit.cart,
                        bagging=fit.treebag, rf=fit.rf, gbm=fit.gbm))
# Table comparison
summary(results)

# boxplot comparison
bwplot(results)
# Dot-plot comparison
dotplot(results)

# Logistic has been deleted as predictive variables are not continuous, therefore cannot
run the model.
# Boosting and RF takes a long processing time +10 sec
# Best Performer is RF

fancyRpartPlot(fit.cart$finalModel)
summary(fit.cart)

Pred<-predict(fit.gbm, newdata = testing)
```

7 Appendices

```
sum(Pred==testing$G3.y)/73*100
# Prediction Accuracy of 79.45%
```

```
Pred<-predict(fit.rf, newdata = testing)
sum(Pred==testing$G3.y)/73*100
# Prediction Accuracy of 76.71%
```

```
Pred<-predict(fit.cart, newdata = testing)
sum(Pred==testing$G3.y)/73*100
# Prediction Accuracy of 83.56%
```

```
# Summary shows better accuracy on RF but Classification Tree has the best Accuracy
when compared with the test data.
# Classification Tree relies on G2, G1, Goout, Higher, Walc(5), Freetime (5)
```

```
summary(fit.gbm)
# GBM main influence variables are: G2,G1,Absences, Age, Goout, Schoolsup and
Failures.
```

```
# Save predictive model for Shiny purposes
# saveRDS(fit.rf,file="Port_Pred.rda")
```

```
##### Given the results, only a few variables are relevant on the ML techniques.
Therefore, will reduce the data
##### to be more specific when requesting information from the APP, reduce to MAX
10 variables between Mathematics
##### and Portuguese.
##### Binary is the most accurate model but it lacks of information for Decision
Making.
```

```
##### MATHEMATICS (4 LEVEL)
WITHOUT G1 and G2 #####
```

```
# Create 4Levels G3.x
RedMath<-d4M %>%
  select(-G1.x,-G2.x)
```

```
LevelsRedMath<-RedMath
LevelsRedMath$G3.x<-cut(L LevelsRedMath$G3.x, br=c(-1,7,10,13,20), labels =
c("HighFail","LowFail","LowPass","HighPass"))
```

```
LevelsRedMath$G3.x
str(L LevelsRedMath)
```

7 Appendices

```
# 1=HighFail, 2=LowFail, 3=LowPass, 4=HighPass
```

```
#Create Training and Test datasets
```

```
inTrain = createDataPartition(y = LevelsRedMath$G3.x, p = .80, list = FALSE)
```

```
training = LevelsRedMath[inTrain,]
```

```
testing = LevelsRedMath[-inTrain,]
```

```
# rename dataset to keep code below generic
```

```
dataset <- training
```

```
str(dataset)
```

```
# MODELS
```

```
# CART
```

```
set.seed(seed)
```

```
fit.cart <- train(G3.x~., data=dataset, method="rpart", metric=metric, trControl=control)
```

```
# Bagged CART
```

```
set.seed(seed)
```

```
fit.treebag <- train(G3.x~., data=dataset, method="treebag", metric=metric,  
trControl=control)
```

```
# Random Forest
```

```
set.seed(seed)
```

```
fit.rf <- train(G3.x~., data=dataset, method="rf", metric=metric, trControl=control)
```

```
# Stochastic Gradient Boosting (Generalized Boosted Modeling)
```

```
set.seed(seed)
```

```
fit.gbm <- train(G3.x~., data=dataset, method="gbm", metric=metric, trControl=control,  
verbose=FALSE)
```

```
# Check models
```

```
results <- resamples(list(cart=fit.cart,bagging=fit.treebag, rf=fit.rf, gbm=fit.gbm))
```

```
# Table comparison
```

```
summary(results)
```

```
# boxplot comparison
```

```
bwplot(results)
```

```
# Dot-plot comparison
```

```
dotplot(results)
```

```
str(dataset)
```

```
Pred<-predict(fit.gbm, newdata = testing)
```

```
sum(Pred==testing$G3.x)/73*100
```

```
# Prediction Accuracy of 31.50%
```

```
Pred<-predict(fit.rf, newdata = testing)
```

```
sum(Pred==testing$G3.x)/73*100
```

```
# Prediction Accuracy of 35.61%
```

```
Pred<-predict(fit.cart, newdata = testing)
```

```
sum(Pred==testing$G3.x)/73*100
```


7 Appendices

Prediction Accuracy of 30.13%

Predictive Model with better accuracy is RF

```
fancyRpartPlot(fit.cart$finalModel)
summary(fit.cart)
```

Without G1 and G2 the most influence variables that can be known by teachers are:
Absences, Age, Activities, Paid, Studytime, Famsup.

```
##### PORTUGUESE (4 LEVEL)
WITHOUT G1 and G2 #####
# Create 4Levels G3.y
RedPort<-d4P %>%
  select(-G1.y,-G2.y)

LevelsRedPort<-RedPort
LevelsRedPort$G3.y<-cut(L LevelsRedPort$G3.y, br=c(-1,7,10,13,20), labels =
c("HighFail","LowFail","LowPass","HighPass"))

LevelsRedPort$G3.y
str(L LevelsRedPort)
# 1=HighFail, 2=LowFail, 3=LowPass, 4=HighPass

#Create Training and Test datasets
inTrain = createDataPartition(y = LevelsRedPort$G3.y, p = .80, list = FALSE)
training = LevelsRedPort[inTrain,]
testing = LevelsRedPort[-inTrain,]

# rename dataset to keep code below generic
dataset <- training
str(dataset)

# MODELS

# CART
set.seed(seed)
fit.cart <- train(G3.y~., data=dataset, method="rpart", metric=metric, trControl=control)
# Bagged CART
set.seed(seed)
fit.treebag <- train(G3.y~., data=dataset, method="treebag", metric=metric,
trControl=control)
# Random Forest
```

7 Appendices

```
set.seed(seed)
fit.rf <- train(G3.y~., data=dataset, method="rf", metric=metric, trControl=control)
# Stochastic Gradient Boosting (Generalized Boosted Modeling)
set.seed(seed)
fit.gbm <- train(G3.y~., data=dataset, method="gbm", metric=metric, trControl=control,
verbose=FALSE)

# Check models
results <- resamples(list(cart=fit.cart,bagging=fit.treebag, rf=fit.rf, gbm=fit.gbm))
# Table comparison
summary(results)

# boxplot comparison
bwplot(results)
# Dot-plot comparison
dotplot(results)

Pred<-predict(fit.gbm, newdata = testing)
sum(Pred==testing$G3.y)/73*100
# Prediction Accuracy of 50.68%

Pred<-predict(fit.rf, newdata = testing)
sum(Pred==testing$G3.y)/73*100
# Prediction Accuracy of 47.94%

Pred<-predict(fit.cart, newdata = testing)
sum(Pred==testing$G3.y)/73*100
# Prediction Accuracy of 42.46%

fancyRpartPlot(fit.cart$finalModel)
summary(fit.cart)
summary(fit.gbm)
summary(fit.rf)

## Without G1 and G2 the most influence variables are:
# Absences, Age,Sex, Schoolsup, Romantic, Activities, Famsup, Higher, Failures
```

```
##### Create the best predictive models fro both subjects
#####
##### and save them for Shiny purposes
#####
```

```
Export_Port<-Levels4P %>%
  select(age,sex,G1.y,G2.y,G3.y,absences.y,failures.y,schoolsup.y,paid.y,      famsup.y,
activities.y, studytime.y, higher.y)
```

```
#New_Port<-LevelsPort_Model %>%
# mutate(age=as.factor(age),
#       G1.y=as.factor(G1.y),
#       G2.y=as.factor(G2.y),
#       absences.y=as.factor(absences.y))
```

```
str(Export_Port)
write.csv(Export_Port,"Export_Port.csv")
```

```
Export_Math<-Levels4M %>%
  select(age,sex,G1.x,G2.x,G3.x,absences.x,failures.x,schoolsup.x,paid.x,      famsup.x,
activities.x, studytime.x, higher.x)
```

```
str(Export_Math)

write.csv(Export_Math,"Export_Math.csv")
```

```
# Predict models
```

```
# RF for Math
set.seed(seed)
fit.rf.Math <- train(G3.x~., data=New_Math, method="rf", metric=metric,
trControl=control, verbose=FALSE)
```

```
saveRDS(fit.rf.Math, file="Math_Pred.rda")
```

```
#RF for Portuguese
fit.rf.Port <- train(G3.y~., data=New_Port, method="rf", metric=metric,
trControl=control)
summary(fit.rf.Port)
```

```
saveRDS(fit.rf.Port, file="Port_Pred.rda")
```

```
##### PCA ANALYSIS
MATHEMATICS #####
#install.packages("ggfortify")
```

7 Appendices

```
library(ggfortify)
```

```
New_Math<-Levels4M %>%  
  select(age,sex,G1.x,G2.x,G3.x,absences.x,failures.x,schoolsup.x,paid.x,      famsup.x,  
  activities.x, studytime.x, higher.x)
```

```
str(New_Math)
```

```
New_Math<-New_Math %>%  
  mutate(schoolsup.x=as.numeric(schoolsup.x),  
    activities.x=as.numeric(activities.x),  
    famsup.x=as.numeric(famsup.x),  
    sex=as.numeric(sex),  
    failures.x=as.numeric(failures.x),  
    paid.x=as.numeric(paid.x),  
    studytime.x=as.numeric(studytime.x),  
    higher.x=as.numeric(higher.x))
```

```
str(New_Math)
```

```
New_Math1<-New_Math %>%  
  select(-G3.x)
```

```
cov(New_Math1)  
cor(New_Math1)
```

```
New_Math1_std<-scale(New_Math1)  
cov(New_Math1_std)
```

```
d <- dist(New_Math1_std, method = "euclidean")
```

```
# Hierarchical clustering using Complete Linkage  
hc1 <- hclust(d, method = "complete" )  
?hclust  
# Plot the obtained dendrogram  
plot(hc1, cex = 0.6, hang = 0)
```

```
k = 4  
n = nrow(New_Math1_std)  
MidPoint = (hc1$height[n-k] + hc1$height[n-k+1]) / 2  
abline(h = MidPoint, lty=2, col="red")
```

```
Math_clusters <- kmeans(New_Math1_std, centers = 4)
```

```
Math_clusters  
barplot(Math_clusters$centers[,1])
```

7 Appendices

```
barplot(Math_clusters$centers[1,])
barplot(Math_clusters$centers[2,])
barplot(Math_clusters$centers[3,])
barplot(Math_clusters$centers[4,])

pairs(New_Math1, col = Math_clusters$cluster)

pca<-prcomp(New_Math1_std)
summary(pca)
barplot(pca$rotation[,2])
barplot(pca$rotation[,1])

plot(pca,type="l")

plot(pca$x[,1],pca$x[,2],xlab="PC 1",ylab="PC 2",col = New_Math$G3.x)
plot(pca$x[,1],pca$x[,2],xlab="PC 1",ylab="PC 2",col = Math_clusters$cluster)

PCA_Math<-autoplot(pca, data = New_Math, colour = "G3.x",
  loadings = TRUE, loadings.colour = 'blue',
  loadings.label = TRUE, loadings.label.size = 5, loadings.label.colour="black")

PCA_Math+
  scale_colour_manual(values=c("HighPass" = "green", "LowPass" = "yellow",
"LowFail"="orange", "HighFail"="red"))+
  theme_bw()+
  theme(panel.background = element_blank(),
    panel.grid.major = element_blank(), #remove major-grid labels
    panel.grid.minor = element_blank(), #remove minor-grid labels
    plot.background = element_blank())

##### PCA ANALYSIS
PORTUGUESE #####

New_Port<-Levels4P %>%
  select(age,sex,G1.y,G2.y,G3.y,absences.y,failures.y,schoolsup.y,paid.y, famsup.y,
activities.y, studytime.y, higher.y)

str(New_Port)

New_Port<-New_Port %>%
  mutate(schoolsup.y=as.numeric(schoolsup.y),
    activities.y=as.numeric(activities.y),
    famsup.y=as.numeric(famsup.y),
    sex=as.numeric(sex),
    failures.y=as.numeric(failures.y),
    paid.y=as.numeric(paid.y),
    studytime.y=as.numeric(studytime.y),
```

7 Appendices

```
higher.y=as.numeric(higher.y))

str(New_Port)

New_Port1<-New_Port %>%
  select(-G3.y)

cov(New_Port1)
cor(New_Port1)

New_Port1_std<-scale(New_Port1)
cov(New_Port_std)

d <- dist(New_Port1_std, method = "euclidean")

# Hierarchical clustering using Complete Linkage
hc1 <- hclust(d, method = "complete" )
?hclust
# Plot the obtained dendrogram
plot(hc1, cex = 0.6, hang = 0)

k = 4
n = nrow(New_Port1_std)
MidPoint = (hc1$height[n-k] + hc1$height[n-k+1]) / 2
abline(h = MidPoint, lty=2, col="red")

Port_clusters <- kmeans(New_Port1_std, centers = 4)

Port_clusters

barplot(Port_clusters$centers[,1])

barplot(Port_clusters$centers[,2])
barplot(Port_clusters$centers[,3])
barplot(Port_clusters$centers[,4])

pairs(New_Port1, col = Port_clusters$cluster)

pca<-prcomp(New_Port1_std)
summary(pca)
barplot(pca$rotation[,2])
barplot(pca$rotation[,1])

plot(pca,type="l")
```

```

plot(pca$x[,1],pca$x[,2],xlab="PC 1",ylab="PC 2",col = New_Port$G3.y)
plot(pca$x[,1],pca$x[,2],xlab="PC 1",ylab="PC 2",col = Port_clusters$cluster)

PCA_Port<-autoplot(pca, data = New_Port, colour = "G3.y",
                   loadings = TRUE, loadings.colour = 'blue',
                   loadings.label = TRUE, loadings.label.size = 5,
                   loadings.label.colour="black")

PCA_Port+
  scale_colour_manual(values=c("HighPass" = "green", "LowPass" = "yellow",
                                "LowFail"="orange", "HighFail"="red"))+
  theme_bw()+
  theme(panel.background = element_blank(),
        panel.grid.major = element_blank(), #remove major-grid labels
        panel.grid.minor = element_blank(), #remove minor-grid labels
        plot.background = element_blank())

#Export Final Data for Shiny Project
write.csv(New_Port,'Portuguese_Language.csv')

write.csv(New_Math,'Mathematics.csv')

##### SAND BOX
#####

##### NMDS TRYOUT (https://mb3is.megx.net/gustame) #####

New_Port<-LevelsPort_Model %>%
  select(G3.y,G2.y,absences.y,age,schoolsup.y,romantic.y,activities.y,famsup.y)

New_Port<-LevelsPort_Model %>%
  mutate(schoolsup.y=as.numeric(schoolsup.y),
         activities.y=as.numeric(activities.y),
         famsup.y=as.numeric(famsup.y),
         sex=as.numeric(sex),
         failures.y=as.numeric(failures.y),
         paid.y=as.numeric(paid.y),
         studytime.y=as.numeric(studytime.y))

New_Port<-New_Port %>%

```

7 Appendices

```
mutate(absences.y=absences.y^1/8)

str(New_Port)
summary(New_Port)


#install.packages("vegan")
library(vegan)

New_Port1<-New_Port %>%
  select(-G3.y)

fm<-metaMDS(New_Port1,autotransform = FALSE)

fm
stressplot(fm)

plot(fm)

data.scores <- as.data.frame(scores(fm)) #Using the scores function from vegan to
extract the site scores and convert to a data.frame
data.scores$G3.y <- New_Port$G3.y # add the grp variable created earlier
head(data.scores) #look at the data

species.scores <- as.data.frame(scores(fm, "species")) #Using the scores function from
vegan to extract the species scores and convert to a data.frame
species.scores$species <- rownames(species.scores) # create a column of species, from
the rownames of species.scores
head(species.scores) #look at the data

#to group Grading groups as a polygon shape in ggplot
HighPass <- data.scores[data.scores$G3.y == "HighPass",
][chull(data.scores[data.scores$G3.y ==
"HighPass", c("NMDS1",
"NMDS2"))], ] # hull values for High Pass
LowPass <- data.scores[data.scores$G3.y == "LowPass",
][chull(data.scores[data.scores$G3.y ==
"LowPass", c("NMDS1", "NMDS2"))],
] # hull values for Low Pass
LowFail <- data.scores[data.scores$G3.y == "LowFail",
][chull(data.scores[data.scores$G3.y ==
"LowFail", c("NMDS1", "NMDS2"))], ]
# hull values for Low Fail
HighFail <- data.scores[data.scores$G3.y == "HighFail",
][chull(data.scores[data.scores$G3.y ==
"HighFail", c("NMDS1", "NMDS2"))],
] # hull values for High Fail
hull.data <- rbind(HighPass, LowPass, LowFail, HighFail) #combine 4 groups
hull.data

ggplot() +
```



```

geom_point(data=data.scores,aes(x=NMDS1,y=NMDS2,colour=G3.y),size=3, alpha =
0.5) + # add the point markers
scale_colour_manual(values=c("HighPass" = "green", "LowPass" = "yellow",
"LowFail"="orange", "HighFail"="red")) + #manual colours for date points
scale_fill_manual(values=c("HighPass" = "green", "LowPass" = "yellow",
"LowFail"="orange", "HighFail"="red")) + #manual colours for fill polygons

geom_text(data=species.scores,aes(x=NMDS1,y=NMDS2,label=species),fontface="bol
d", size=4) + # add the species labels
coord_equal() + #important for the dimension of the NMDS
theme_bw()+
theme(panel.background = element_blank(),
      panel.grid.major = element_blank(), #remove major-grid labels
      panel.grid.minor = element_blank(), #remove minor-grid labels
      plot.background = element_blank(),
      legend.position="top")

```

Classification Tree: 4L Mathematics

```

#install.packages("ISLR")
#install.packages("tree")

```

```

library(ISLR)
library(tree)

```

```
# Select Main decision variables
```

```
Math_Model<-d4M %>%
```

```
  select(age,sex,G1.x,G2.x,G3.x,absences.x,failures.x,schoolsup.x,paid.x,      famsup.x,
activities.x, studytime.x, higher.x)
```

```
# Create G3 Levels
```

```
LevelsMath_Model<-Math_Model
```

```
LevelsMath_Model$G3.x<-cut(L LevelsMath_Model$G3.x, br=c(-1,7,10,13,20), labels =
c("HighFail","LowFail","LowPass","HighPass"))
```

```
str(L LevelsMath_Model)
```

```
## Classification Tree ##
```

```
set.seed(seed)
```

```
train <- sample(1:nrow(L LevelsMath_Model), 297)
```

```
# 369 observations, train = 297 observ.
```

```
tree.Math <- tree(G3.x~., LevelsMath_Model, subset=train)
```

```
plot(tree.Math)
```

```
text(tree.Math, pretty=0)
```

```
tree.Math
```

7 Appendices

```
#Predict
tree.pred <- predict(tree.Math, LevelsMath_Model[-train,], type="class")
# check predicting error
with(LevelsMath_Model[-train,], table(tree.pred, G3.x))
(13+12+16+18)/72
# Predict shows a result of 86.11%

## Bagging ###
library(ipred)
bagging.Math <- bagging(G3.x~., LevelsMath_Model, subset=train, coob = T)
#Predict Bagging
bagging.pred <- predict(bagging.Math, LevelsMath_Model[-train,], type="class")
# check predicting error
with(LevelsMath_Model[-train,], table(bagging.pred, G3.x))
(11+1+1)/72
# Confusion Matrix error, not doing proper diagonal. Find alternative (Jack)
sum(bagging.pred==LevelsMath_Model[-train,]$G3.x)/72*100

# Predict shows a result of 84.72%

## Random Forest ##
library("randomForest")

RF.Math <- randomForest(G3.x~., LevelsMath_Model, subset=train)
#Plot RAndom FOrEst

plot(RF.Math)
print(RF.Math)

#Predict Random Forest
RF.pred <- predict(RF.Math, LevelsMath_Model[-train,], type="response")
# check predicting error
table_Math_RF<-with(LevelsMath_Model[-train,], table(RF.pred, G3.x))
accuracy_Test_RF <- sum(diag(table_Math_RF)) / sum(table_Math_RF)
print(paste('Accuracy for Math test', accuracy_Test_RF*100))

# Predict shows a result of 81.94%

## Boosting ##
library("gbm")
?gbm
Boosting.Math <- gbm(G3.x~., data = LevelsMath_Model[train,], distribution =
"gaussian", n.trees =10000, shrinkage = 0.01, interaction.depth = 4)
#Predict Boosting
Boosting.pred <- predict(Boosting.Math, LevelsMath_Model[-train,], n.trees
=10000,type="response")
# check predicting error
Math_Roun<-round(Boosting.pred)
with(LevelsMath_Model[-train,], table(Math_Roun, G3.x))
(11+15+16+19)/72*100
```

7 Appendices

Predict shows a result of 84.72%

Boosting has the same accuracy than the Classification Tree, other improvement techniques have a lower performance

Classification Tree: 4L Portuguese

Select Main decision variables

```
Port_Model<-d4P %>%
```

```
  select(age,sex,G1.y,G2.y,G3.y,absences.y,failures.y,schoolsup.y,paid.y,      famsup.y,
activities.y, studytime.y, higher.y)
```

Create G3 Levels

```
LevelsPort_Model<-Port_Model
```

```
LevelsPort_Model$G3.y<-cut(LevelsPort_Model$G3.y, br=c(-1,7,10,13,20), labels =
c("HighFail","LowFail","LowPass","HighPass"))
```

Classification Tree

```
set.seed(seed)
```

```
train <- sample(1:nrow(LevelsPort_Model), 297)
```

369 observations, train = 297 observ.

```
tree.Port <- tree(G3.y~., LevelsPort_Model, subset=train)
```

```
plot(tree.Port)
```

```
text(tree.Port, pretty=0)
```

```
tree.Port
```

#Predict

```
tree.pred <- predict(tree.Port, LevelsPort_Model[-train,], type="class")
```

check predicting error

```
with(LevelsPort_Model[-train,], table(tree.pred, G3.y))
```

```
(1+9+28+17)/72*100
```

Predict shows a result of 76.38%

Bagging

```
library(ipred)
```

```
?bagging
```

```
bagging.Port <- bagging(G3.y~., LevelsPort_Model, subset=train, coob=T)
```

#Predict Bagging

```
bagging.pred <- predict(bagging.Port, LevelsPort_Model[-train,], type="class")
```

check predicting error

```
with(LevelsPort_Model[-train,], table(bagging.pred, G3.y))
```

```
(1+1+4)/72*100
```

Confusion Matrix error, not doing proper diagonal. Find alternative (Jack)

```
sum(bagging.pred==LevelsPort_Model[-train,]$G3.y)/72*100
```

Predict shows a result of 84.72%

Random Forest

```

library("randomForest")

RF.Port <- randomForest(G3.y~., LevelsPort_Model, subset=train)
#Plot RAndom FOrEst

plot(RF.Port)
print(RF.Port)

#Predict Random Forest
RF.pred <- predict(RF.Port, LevelsPort_Model[-train,], type="response")
# check predicting error
table_Port_RF<-with(LevelsPort_Model[-train,], table(RF.pred, G3.y))
accuracy_Test_RF <- sum(diag(table_Port_RF)) / sum(table_Port_RF)
print(paste('Accuracy for Math test', accuracy_Test_RF*100))

# Predict shows a result of 80.55%

## Boosting ##
library("gbm")
?gbm
Boosting.Port <- gbm(G3.y~., data = LevelsPort_Model[train,], distribution =
"gaussian", n.trees =10000, shrinkage = 0.01, interaction.depth = 4)
#Predict Boosting
Boosting.pred <- predict(Boosting.Port, LevelsPort_Model[-train,], n.trees
=10000,type="response")
# check predicting error
Port_Roun<-round(Boosting.pred)
with(LevelsPort_Model[-train,], table(Port_Roun, G3.y))
(1+11+22+18)/72*100
# Predict shows a result of 72.22%

# Bagging has the highest accuracy rate and the worst performer is Boosting.

```

7.2.2 Shiny Code

```

#install.packages("shiny","dplyr","reactable","shinydashboard","tidyverse")
#install.packages("flexdashboard")
#install.packages("randomForest")
#install.packages("ipred")
#install.packages("ISLR")
#install.packages("tree")
#install.packages("shinyWidgets")
#install.packages('rsconnect')

library(ISLR)

```

7 Appendices

```
library(tree)

library(shiny)
library(tidyverse)
library(dplyr)
library(reactable)
library(shinydashboard)
library(shinyWidgets)
library(flexdashboard)
library(reactable)
library(stringr)
library(caret)
library(mlbench)
library(rpart)
library(ipred)
library(randomForest)
library(rsconnect)

# Load Data Portuguese Language
Import_Math<-read.csv(file="Export_Math.csv")
Import_Port<-read.csv(file="Export_Port.csv")
#Math_Pred<-readRDS("Math_Pred.rda")
#Port_Pred<-readRDS("Port_Pred.rda")

Import_Math<- Import_Math %>% select(-c(X))
Import_Port<- Import_Port %>% select(-c(X))

ui <- dashboardPage(

  # Application title
  dashboardHeader(title="Student Performance Software"),

  # Dashboard Sidebar
  dashboardSidebar(sidebarMenu(
    menuItem("Introduction", tabName = "Intro", icon = icon("dashboard")),
    menuItem("Portuguese", tabName = "Port", icon = icon("table")),
    menuItem("Mathematics", tabName = "Math", icon = icon("table")),
    br(),
    br(),
    br(),
    br(),
    br(),
    br(),
    br(),
    br(),
    br(),
    br(),
    br()
  ))
```

[illegible]

```
dashboardBody(  
  tabItems(  
    tabItemLabel("Dashboard"),  
    tabContentPanel(
```

$$),$$

```

#Tab 2
tabItem(tabName = "Port",
  fluidRow(
    box( radioButtons("Gender","Gender",choices = c("F", "M")), width = 2, height
= 110),

    box( selectInput("Age","Age",choices = 15:18, width = 100),width = 2, height
= 110),

    box( selectInput("Studytime","Studytime",choices = 1:4, width=100),width = 2,
height = 110),

    box(radioButtons("Familiysupport", "Family Support",choices = c("yes",
"no")),width = 2, height = 110),

    box(radioButtons("Activities", "Activities",choices = c("yes", "no")),width = 2,
height = 110),

    box(radioButtons("Paidclasses", "Extra Paid Classes", choices = c("yes",
"no")),width = 2, height = 110)
  ),
  fluidRow(

    box( selectInput("Grade1", "1st Examination Grade", choices=1:20, selected =
10,width = 100),width = 2, height = 110),

    box( selectInput("Grade2", "2nd Examination Grade", choices=1:20,selected =
10, width = 100),width = 2, height = 110),

    box(numericInput("Absences", "Course Absences:", value = 1,min = 0,max =
150, width = 100),width = 2, height = 110),

    box( selectInput("Failures","Previous Fails",choices = 0:3, width = 100),width
= 2, height = 110),

    box( radioButtons("Schoolsupport", "School Support",choices = c("yes",
"no")),width = 2, height = 110),

    box( radioButtons("HigherEducation", "Higher Education",choices = c("yes",
"no")),width = 2, height = 110),

    #radioButtons("Nursery","Nursery",choices = c("Yes", "No")),

    #radioButtons("Romantic", "In a Relationship", choices = c("Yes", "No")),

    #selectInput("Free Time", "Free Time", choices=1:5),

    #selectInput("Go Out", "Going Out with Friends", choices=1:5),

```

7 Appendices

```
#selectInput("Weekend Alcohol", "Weekend Alcohol", choices=1:5),

#selectInput("Week Alcohol", "During Week Alcohol", choices=1:5),

actionButton("Enter", "Enter",style="padding:50px; font-size:120%;float:right
")
),

# Show results

#tableOutput("Student2"),
#textOutput("Student"),

infoBoxOutput("modelSummary", width = NULL)

),

#Tab 3
tabItem(tabName = "Math",
  fluidRow(
    box( radioButtons("Gender","Gender",choices = c("F", "M")), width = 2,
height = 110),

    box( selectInput("Age","Age",choices = 15:18, width = 100),width = 2, height
= 110),

    box( selectInput("Studytime","Studytime",choices = 1:4, width=100),width =
2, height = 110),

    box(radioButtons("Familiysupport", "Family Support",choices = c("yes",
"no")),width = 2, height = 110),

    box(radioButtons("Activities", "Activities",choices = c("yes", "no")),width =
2, height = 110),

    box(radioButtons("Paidclasses", "Extra Paid Classes", choices = c("yes",
"no")),width = 2, height = 110)
  ),
  fluidRow(

    box( selectInput("Grade1", "1st Examination Grade", choices=1:20, selected
= 10,width = 100),width = 2, height = 110),
```



```
box( selectInput("Grade2", "2nd Examination Grade", choices=1:20,selected
= 10, width = 100),width = 2, height = 110),
```

```
box(numericInput("Absences", "Course Absences:", value = 1,min = 0,max =
150, width = 100),width = 2, height = 110),
```

```
box( selectInput("Failures","Previous Fails",choices = 0:3, width = 100),width
= 2, height = 110),
```

```
box( radioButtons("Schoolsupport", "School Support",choices = c("yes",
"no")),width = 2, height = 110),
```

```
box( radioButtons("HigherEducation", "Higher Education",choices = c("yes",
"no")),width = 2, height = 110),
```

```
#radioButtons("Nursery","Nursery",choices = c("Yes", "No")),
```

```
#radioButtons("Romantic", "In a Relationship", choices = c("Yes", "No")),
```

```
#selectInput("Free Time", "Free Time", choices=1:5),
```

```
#selectInput("Go Out", "Going Out with Friends", choices=1:5),
```

```
#selectInput("Weekend Alcohol", "Weekend Alcohol", choices=1:5),
```

```
#selectInput("Week Alcohol", "During Week Alcohol", choices=1:5),
```

```
actionButton("Enter2",          "Enter",style="padding:50px;          font-
size:120%;float:right ")
),
```

```
# Show results
```

```
infoBoxOutput("modelSummary2", width = NULL)
```

```
)
)
)
)
```

```
# Define server logic
```

```
server <- function(input, output, session) {
```

```
#Portuguese
```

```
observeEvent(input$Enter, {
  sex=as.character(input$Gender)
  age=as.numeric(input$Age)
  studytime.y=as.numeric(input$Studytime)
  famsup.y=as.character(input$Familiysupport)
  activities.y=as.character(input$Activities)
  paid.y=as.character(input$Paidclasses)
```

```

G1.y=as.numeric(input$Grade1)
G2.y=as.numeric(input$Grade2)
absences.y=as.numeric(input$Absences)
failures.y=as.numeric(input$Failures)
schoolsup.y=as.character(input$Schoolsupport)
higher.y=as.character(input$HigherEducation)

#t <- data.frame(as.factor(input$Gender),input$Age, as.factor(input$Familiysupport),
as.factor(input$Studytime),
#           as.factor(input$Activities), as.factor(input$Paidclasses),input$Grade1,
input$Grade2,input$Absences,
#           as.factor(input$Failures), as.factor(input$Schoolsupport),
as.factor(input$HigherEducation))

z<-data.frame(age,sex,G1.y,G2.y,absences.y,
failures.y,schoolsup.y,paid.y,famsup.y,activities.y,studytime.y,
              higher.y)

#train_Port <- sample(1:nrow(Import_Port), 296)
#tree.port <- tree(G3.y~., data=Import_Port, subset = train_Port)
#summary(tree.port)

#RF.Port <- randomForest(G3.y~., Import_Port,subset=train_Port)

inTrain = createDataPartition(y = Import_Port$G3.y, p = .80, list = FALSE)
training = Import_Port[inTrain,]
testing = Import_Port[-inTrain,]

fit.rf <- train(G3.y~., data=training, method="rf")

output$modelSummary <- renderInfoBox({
  infoBox("Predicted Final Grade for Portuguese Language", predict(fit.rf,
newdata= z), icon = icon("user-graduate"),color = "blue")
})
})

# Mathematics

observeEvent(input$Enter2, {
  sex=as.character(input$Gender)
  age=as.numeric(input$Age)
  studytime.x=as.numeric(input$Studytime)
  famsup.x=as.character(input$Familiysupport)
  activities.x=as.character(input$Activities)
  paid.x=as.character(input$Paidclasses)
  G1.x=as.numeric(input$Grade1)
  G2.x=as.numeric(input$Grade2)
  absences.x=as.numeric(input$Absences)
  failures.x=as.numeric(input$Failures)
  schoolsup.x=as.character(input$Schoolsupport)
  higher.x=as.character(input$HigherEducation)

```

```

z_Math<-data.frame(age,sex,G1.x,G2.x,absences.x,
failures.x,schoolsup.x,paid.x,famsup.x,activities.x,studytime.x,
higher.x)

inTrain_Math = createDataPartition(y = Import_Math$G3.x, p = .80, list = FALSE)
training_Math = Import_Math[inTrain,]
testing_Math = Import_Math[-inTrain,]

fit.rf.Math <- train(G3.x~., data=training_Math, method="rf")

output$modelSummary2 <- renderInfoBox({
  infoBox("Predicted Final Grade for Mathematics", predict(fit.rf.Math, newdata=
z_Math), icon = icon("user-graduate"),color = "blue")
})
})

}

# Run the application
shinyApp(ui = ui, server = server)

```