Wholly owned by UTAR Education Foundation
(Co. No. 578227-M)
DU012(A)

PREDICTIVE MODELLING
UECM3993

GROUP ASSIGNMENT

Session: May 2020

Group Leader: Chan Fus Keng

Dataset: Student Performance Data Set

Title: Application of Machine Learning Algorithms on Student Performance Analysis

Group Members:

Name	Student ID	Course
Chan Fus Keng	1703118	AS
Chang Dick Sheng	1703910	AS
Lee Jun Hau	1607030	AM
Ng Foo Shiong	1400284	FM
Ng Mun Fai	1601030	FM
Ting Bang Yew	1702769	AS

TABLE OF CONTENTS

TABLE OF CONTENTS ..... i - iv
CHAPTER 1 INTRODUCTION 1
1.1 General Introduction 1
1.1.1 Background of Study 1
1.1.2 Background of the data sets and Topic of Study 1
1.1.3 Objective of Study 1
1.1.4 Research Questions 2
1.1.5 Description of Data Set and Description of Data Attributes 2
1.1.6: The statistical model approaches in this assignment 4
1.2 Data Preparation 4
1.2.1 Load and prepare datasets 4
1.2.2 Cleaning datasets 5
1.2.3 Splitting training and testing data (numerical data) 5
1.2.4 Splitting training and testing data (categorical data) 6
CHAPTER 2 UNSUPERVISED LEARNING 6
2.1 Basic Descriptive Analysis (Numeric) 6
2.1.1 Fair Use Policy and Legal Disclaimer 6
2.1.2 Basic Descriptive Analysis on Numeric Data 7
2.1.2.1 For Math 7
2.1.2.1.1 Mean Test for Math 8
2.1.2.1.2 Skewness Test for Math 8
2.1.2.1.3 Kurtosis Test for Math 9
2.1.2.1.4 Normal Test (Jarque-Bera) for Math 9
2.1.2.2 For Portuguese 10
2.1.2.2.1 Mean Test for Portuguese 10
2.1.2.2.2 Skewness Test for Portuguese 11
2.1.2.2.3 Kurtosis Test for Portuguese 11
2.1.2.2.4 Normal Test (Jarque-Bera) for Portuguese 12
2.1.2.3 Analysis of correlation between final grade for Math and Portuguese 13
2.1.2.3.1 Scatterplots: analysis between final grades and 1st/2nd grade for Math 14
2.1.2.3.2 Scatterplots: analysis between final grades and 1st/2nd grade for Portuguese 15
2.1.2.3.3 Analysis of numeric social factor 16
2.1.2.3.4 Analysis of other numeric factors 17
2.1.3 Basic Descriptive Analysis (Categorical) 18
2.1.3.1 Distribution: Analysis by school 18
2.1.3.2 Distribution: Analysis by address 18
2.1.4 Chi-squared test of association 19
2.2 Correlation Analysis 20
2.2.1 Basic Introduction 20
2.2.2 Bartlett's Test of Sphericity 21
2.2.3 Correlation Analysis- Fact Checking 21

2.2.4 Correlation Analysis- Identify Most Significant Factors for Math.....23
2.2.5 Correlation Analysis- Identify Most Significant Factors for Portuguese 23
2.2.6 Weaknesses and Strengths of Correlation Analysis.....23
2.2.6.1 Strength.....23
2.2.6.2 Weakness.....23
2.3 Principal Component Analysis.....24
2.3.1 Math students.....24
2.3.1.1 Analysis of Biplot – Determining factors affecting result for Math.....24
2.3.1.2 Analysis of Principal Component for Math.....25
2.3.2 For Portuguese.....26
2.3.2.1 Analysis of Biplot – Determining factors affecting result for Portuguese..26
2.3.2.2: Analysis of Principal Component for Portuguese.....27
2.3.3: Weakness and Strength.....28
2.3.3.1 Strength.....28
2.3.3.2 Weakness.....28
CHAPTER 3 SUPERVISED LEARNING..... 29
3.1 Linear Regression Model.....29
3.1.1 Math students.....29
3.1.2 Portuguese students.....31
3.1.3 Weaknesses and Strengths of Linear Model approach.....32
3.1.3.1 Strength.....32
3.1.3.2 Weakness.....32
3.2 K-Nearest Neighbors.....33
3.2.1 Math students.....33
3.2.2 Portuguese student.....36
3.2.3 Weakness and Strength of KNN.....38
3.2.3.1 Strength.....38
3.2.3.2 Weakness.....38
3.3 Logistic Regression.....39
3.3.1 Linear regression vs logistic regression.....39
3.3.2 Multiple logistic regression.....39
3.3.2.1 For Math.....39
3.2.2.2 For Portuguese.....41
3.3.3 Strength and weakness of logistic regression.....43
3.3.3.1 Strength.....43
3.3.3.2: Weakness.....43
3.4 Decision Tree/ Tree Pruning.....44
3.4.1 Decision Tree.....44
3.4.1.1 Math.....44
3.4.1.2 Portuguese.....46
3.4.2 Weaknesses and Strengths.....48
3.4.2.1 Weakness.....48
3.4.2.2 Strengths.....48
3.4.3 Tree Pruning.....48
3.4.3.1 Math.....48
3.4.3.2 Portuguese.....50

3.4.4 Weaknesses and Strengths..... 51
3.4.4.1 Weakness..... 51
3.4.4.2 Strengths..... 51
3.5 Bootstrap Aggregating (Bagging)..... 52
3.5.1 For Math Students..... 52
3.5.2 For Portuguese Students..... 53
3.5.3 Weakness & Strength..... 54
3.5.3.1 Weakness..... 54
3.5.3.2 Strengths..... 54
CHAPTER 4 CONCLUSION..... 55
4.1 Review of all approaches we tried..... 55
4.2 Best fit model..... 55
4.3 Overall Strengths and Weaknesses of this report..... 56
4.4 Significant factors that affects the result generally..... 56
4.5 Future Work recommended to improvise the data sets..... 57
4.6 Recommendations for Education Officials..... 57

CHAPTER 1 INTRODUCTION

1.1 General Introduction

1.1.1 Background of Study

The cultural heritage and values are passed down by education from one generation to the next. Secondary education not only offers information and skills, but also inculcates values, training instincts, cultivating right attitudes and habits to encourage adolescents either step into higher education or to provide a workplace for students who have chosen to finish secondary education. Without secondary education to monitor young people's growth during their adolescence, they will be ill-prepared for tertiary or occupational education, and the risk of juvenile delinquency and underage pregnancy will also increase. Such negative impacts would increase the social and socio-economic tensions and expenditures. Psychologists have found that factors affecting student personality and achievement can be categorized into two categories-nature and nurturing. The two groups have been found to be playing complementary positions. As nature decides the child's level of intelligence and inherited ability, nurture helps enhance those inherent abilities. The nurture involves the home, the school and the peer groups to which the child belongs (Alokan et al., 2013).

In the last decades, the educational level of the Portugal has significantly improved. However, Portuguese secondary students' academic performances were still lower than the OECD countries' average. The present research aims to tackle student achievement in secondary education using Business Intelligence/ Data Mining techniques, with the goal of extracting high-level information from raw data, they deliver interesting automated tools that can support the field of education.

1.1.2 Background of the data sets and Topic of Study

The datasets we used were collected from two public schools, i.e. Gabriel Pereira and Mousinho da Silveira. The data samples are collected using questionnaire and school reports. In these datasets, we use the predictors such as school, sex, age, information about the students' study and lifestyle habits, family details, and three grades. The student's performance on Mathematics course and Portuguese language course in these two Portuguese schools will be the main target for our research. These two distinct subjects, Mathematics and Portuguese language course, were modelled under binary/five-level classification and regression tasks.

The main topic of our analysis is "Application of Machine Learning Algorithms on Student Performance Analysis". In this assignment, predictive models with and without G1 and G2 will be explored as the earlier grades (G1, G2) are known to be helpful in predicting the final grade (G3).

1.1.3 Objective of Study

This research project aims to:

1. Provide a greater understanding of the factors influencing Mathematics and Portuguese performance among secondary school students in Portugal.
2. Analyze and compare the variations between the Mathematics and Portuguese influences.
3. Evaluate the prediction of these variables in the forecasting of Mathematics and Portuguese performance using different predictive models.
4. Identify the most significant predictors and gain a better understanding of these factors.

1.1.4 Research Questions

In this assignment, the group members focused on solving these research questions:

1. Which factor can affect the overall performance of students taking either Mathematics or Portuguese language or both courses? [Solve using unsupervised and supervised learning]
2. Which supervised learning model can be best fitted to predict the outcome with given conditions.
3. What effective methods can be implemented to improve the student's grade based on the best fitted model?
4. What future work should be made to make the data set more analyzable and less random errors?

1.1.5 Description of Data Set and Description of Data Attributes

Total of 395 observations and 649 observations were collected for Mathematics and Portuguese courses respectively. There are 33 predictors of in both datasets respectively. Both include school, sex, age, information about the students' study and lifestyle habits, family details, and three grades.

This data focuses on the achievement of students in secondary education in two Portuguese schools. The data attributes include student ratings, features related to demographic, social, and education) and were compiled using school reports and questionnaires. Two output databases are presented in two distinct topics: Mathematics and Portuguese.

Attribution of data is shown on the next page.

Attribution of data is shown in the table below

No.	Attributes	Description and (Domain of the Variable)
1	school	student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
2	sex	student's gender (binary: 'F' - female or 'M' - male)
3	age	student's age (numeric: from 15 to 22)
4	address	student's home address type (binary: 'U' - urban or 'R' - rural)
5	famsize	family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
6	Pstatus	parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
7	Medu	mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)
8	Fedu	father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)
9	Mjob	mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
10	Fjob	father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
11	reason	reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
12	nursery	attended nursery school? (binary: yes or no)
13	Internet	Internet access at home? (binary: yes or no)
14	guardian	student's guardian (nominal: 'mother', 'father' or 'other')
15	traveltime	home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
16	studytime	weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
17	failures	number of past class failures (numeric: n if 1<=n<3, else 4)
18	schoolsup	extra educational support (binary: yes or no)
19	famsup	family educational support (binary: yes or no)
20	paid	extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
21	activities	extra-curricular activities (binary: yes or no)
22	higher	wants to take higher education (binary: yes or no)
23	romantic	with a romantic relationship (binary: yes or no)
24	famrel	quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
25	freetime	free time after school (numeric: from 1 - very low to 5 - very high)
26	goout	going out with friends (numeric: from 1 - very low to 5 - very high)
27	Dalc	workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
28	Walc	weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
29	health	current health status (numeric: from 1 - very bad to 5 - very good)
30	absences	number of school absences (numeric: from 0 to 93)
31	G1	first period grade (numeric: from 0 to 20)
32	G2	second period grade (numeric: from 0 to 20)
33	G3	final grade (numeric: from 0 to 20, output target)

1.1.6: The statistical model approaches in this assignment

We summarize all approaches we tried into the table below.

	Unsupervised Learning	Supervised Learning
Continuous Data	Basic Descriptive Analysis Correlation Analysis PCA	Linear Regression
Categorical Data	Basic Descriptive Analysis	k-Nearest Neighbor Logistic Regression Decision Tree Tree Pruning Bootstrap Aggregation

1.2 Data Preparation

1.2.1 Load and prepare datasets

Firstly, we prepare the two original data sets by downloading the csv files from the website named "UCI Machine Learning Repository". One data set is named "student-mat.csv" and another is named "student-por.csv". The two data set are used to generate third data set. The third set is the result of a list of students taking Math and Portuguese subjects together.

Once 3 data sets are prepared, we create a column of index number named" student_id". This student_id is needed for building supervising learning model.

Below shows how a single data of the third data set is produced:

Student A taking both Math and Portuguese subjects, hence data for "school", "address" and "age" are repeated.

Data Frame	School	Age	Address	G1	G2	G3
student-por (df1)	GP	18	U	1	2	3
student-mat (df2)	GP	18	U	4	5	6
Action: Merge "School", "Age" and "Address", then concatenate the data for "G1", "G2", "G3" for Portuguese and Math respectively in new data frame named "student-merge"						

student-merge (df3)

School	Age	Address	G1.x	G2.x	G3.x	G1.y	G2.y	G3.y
GP	18	U	1	2	3	4	5	6

where G1.x is the first period grade for Math, G3.y is the final grade for Portuguese

The variable "df1" [data frame 1], "df2" and "df3" are assigned to the respective data sets as according to the example above.

	student_id	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob
1	1	GP	F	15	R	GT3	T	1	1	at_home	other
2	2	GP	F	15	R	GT3	T	1	1	other	other
3	3	GP	F	15	R	GT3	T	2	2	at_home	other
4	4	GP	F	15	R	GT3	T	2	4	services	health
5	5	GP	F	15	R	GT3	T	3	3	services	services
6	6	GP	F	15	R	GT3	T	3	4	services	health
7	7	GP	F	15	R	GT3	T	3	4	services	teacher
8	8	GP	F	15	R	LE3	T	2	2	health	services
9	9	GP	F	15	R	LE3	T	3	1	other	other
10	10	GP	F	15	U	GT3	A	3	3	other	health

The diagram above shows the data frame for third data set.

1.2.2 Cleaning datasets

Then, for these three data sets we check for the missing values and remove that particular observation (row) if any missing value exists. We also check whether there are any duplicates of any single observations, and remove it. Once checking is done for these 3 data sets, we update the variables assigned (df1, df2 and df3) with the cleansed data sets.

1.2.3 Splitting training and testing data (numerical data)

It is very important to remember that the complete data set consists of noise. It is required to reduce such noise to prevent overfitting of the data, where we included too much noise and make the prediction of the model more inaccurate.

Therefore, to avoid the issue of overfitting, any complete data set (say df1) will now be separated into a training set and a testing set, with 75% data set randomly assigned into training set and the remaining 25% into testing set. The same goes to another 2 complete data set (say df2 and df3) where necessary.

By using command set.seed, we will assign the same 75% of that complete data set into the training set and same 25% into testing set every time we prepare the data from original data set. This allows us to generate the consistent outputs and results every time the command is executed. It eases our analysis. The default set.seed command is set.seed(123).

After executing command set.seed, we will assign 2 new variables for training set and testing data. Note that the original target output is numeric for training set and testing set.

Training data (set) is used to build predictive models and testing data (set) is used as a given condition for predicting the output in the model built by the training data.

After the new variables (say trainingdata and testingdata) have been assigned, we perform checking of the row number to ensure that the original data set has been separated properly.

1.2.4 Splitting training and testing data (categorical data)

In supervised learning, we use categorical target output (G3) instead of numeric output.

We first convert the target output of the 3 original data sets (df1, df2 and df3) using the following format.

$$G3 = \{ \begin{matrix} 1 \\ 0 \end{matrix} \quad \begin{matrix} if\ G3 \geq 10 \rightarrow passed \\ if\ G3 < 10 \rightarrow failed \end{matrix}$$

After that, we split the data set again into a training set and testing set, with 75% of the original data sets being randomly assigned to the training set. Lastly, perform checking on number of rows of the training and testing data respectively.

Note that cleansing of data for training and testing data is no longer required as we have done it prior to separation of the original data

CHAPTER 2 UNSUPERVISED LEARNING

2.1 Basic Descriptive Analysis (Numeric)

2.1.1 Fair Use Policy and Legal Disclaimer

Under the "Fair Use" Act, a copyrighted work can be used, cited or incorporated within another author's work legally without needing a license if it's being used explicitly for things like news reporting, researching purposes, teaching, commentary, criticism, and other such uses. The R software we use IS APPROPRIATE for this assignment and it is free of charge to download from the website https://cran.r-project.org/bin/windows/base/.

The document is a copyright (© 1998–2020) by Kurt Hornik, and it is without any warranty. Please refer to the following link below for verification below. (Hornik, 2020)

2.1.2 Basic Descriptive Analysis on Numeric Data

2.1.2.1 For Math

```
> basicStats(df1$G3,ci = 0.99)

                X..df1.G3
nobs        395.000000
NAs           0.000000
Minimum       0.000000
Maximum      20.000000
1. Quartile   8.000000
3. Quartile  14.000000
Mean         10.415190
Median       11.000000
Sum        4114.000000
SE Mean       0.230517
LCL Mean      9.818527
UCL Mean     11.011853
Variance     20.989616
Stdev         4.581443
Skewness     -0.727117
Kurtosis      0.366072
```

Kurtosis is a measure of peakedness for a distribution. Negative values indicate a flat (platykurtic) distribution, positive values indicate a peaked (leptokurtic) distribution, and a near-zero value indicates a normal (mesokurtic) distribution (boeh, n.d.). As a result, the normality test was used to check for continuous and ordinal variables for skewness and kurtosis. A symmetrical dataset will have a skewness equal to 0, which means a normal distribution will have a skewness of 0. Skewness essentially measures the relative size of the two tails.

Rule of Thumb:

The skewness can be used to test whether the distribution of the final grades of Math students is normal distribution. (Normality Testing, 2020). Only in the range of $(-\ 0.5,\ 0.5)$ of skewness, the distribution is considered normal. In the range of $(-1$ to $-0.5)$ or $(0.5$ to $1)$, the distribution is said to be skewed moderately, otherwise we consider a highly skewed distribution.

2.1.2.1.1 Mean Test for Math

Let μ denotes the mean of the final grades of students taking Math.

$H_0: \mu = 0 H_1: \mu \neq 0$

```
> t.test(df1$G3)

        One Sample t-test

data:  df1$G3
t = 45.182, df = 394, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 9.961992 10.868388
sample estimates:
mean of x
 10.41519
```

Test statistics = 45.182

α = 0.01, p-value ≈ 0

With α = 0.01, we **reject** H_0 and failed to conclude that the mean of the final grades of students taking Math is not equal to zero.

2.1.2.1.2 Skewness Test for Math

Let S(x) denotes the skewness of the final grades of students taking Math.

$H_0: S(x) = 0 H_1: S(x) \neq 0$

```
> #skewness test
> s <- -0.727117
> s.test <- s/sqrt(6/395)
> pnorm(s.test)
[1] 1.821221e-09

> pv1 <- 2*(1-pnorm(s.test))
> pv1
[1] 2

> s.test
[1] -5.899663
```

Test statistics, S* = $\frac{-0.727117}{\sqrt{6/395}} = -5.8997$

α = 0.01, p-value ≈ 0

Since p-value < α, ∴ reject H_0.

With α = 0.01, we **reject** H_0 and conclude that the final grades of students taking Math are significantly skewed to the left.

2.1.2.1.3 Kurtosis Test for Math

Let K(x) denotes the skewness of the final grades of students taking Math.

$H_0: K(x) - 3 = 0 H_1: K(x) - 3 \neq 0$

```
> k <- 0.366072
> k.test <- k/sqrt(24/649)
> pnorm(k.test)
[1] 0.971521
> pv2 <- 2*(1-pnorm(k.test))
> pv2
[1] 0.05695799
> k.test
[1] 1.903633
```

Test statistics, K* = $\frac{4.991680}{\sqrt{24/1258}} = 36.13945$

α = 0.01, p-value = 0.9715

Since p-value > α, ∴ do not reject H_0.

With α = 0.01, we do not **reject** H_0 and conclude that of the final grades of students taking Math have light tails.

2.1.2.1.4 Normal Test (Jarque-Bera) for Math

```
> normalTest(df1$G3, method="jb")

Title:
 Jarque - Bera Normalality Test

Test Results:
  STATISTIC:
    X-squared: 37.4883
  P VALUE:
    Asymptotic p Value: 7.236e-09
```

Test statistics, $JB = 37.4883$

α = 0.01, p-value ≈ 0

Since p-value < α, ∴ reject H_0.

With α = 0.05, we **reject** H_0 and conclude that the normality assumption of the final grades of students taking Math is rejected.

The test above shows the final grade of students taking Mathematics course. For the mean test, the p-value is approximately equal to 0, smaller than the significance level of 0.01, which brings the mean of the final grade of students taking

Mathematics course is not equals to zero. As we can see, the mean estimation of the sample is 10.41519, the grade given from 0 to 20, which the two Portuguese schools' students having an average on their final grade in Mathematics course.

For the skewness test, given the value of -5.8997, brings us negative skewness. Negative skewness indicates that the mean of the data values is less than the median, and the data distribution is left-skewed. Heavy-tailed test results brings us the kurtosis test to show the skewness of the data distribution, other than the skewness test, that we can conclude the results is a negative kurtosis since it indicates a "light tailed" distribution. The "minus 3" at the end of this formula is often described as a correction to make the normal distribution kurtosis equal to zero, because for a normal distribution the kurtosis is three. Normal test, also known as Jarque-Bera test, brings us value of 37.4883, which means the larger the Jarque-Bera values shown, the more the data deviates from the normal distribution of its dataset.

2.1.2.2 For Portuguese

```
> basicStats(df2$G3,ci=0.99)
```

	X..df2.G3
nobs	649.000000
NAs	0.000000
Minimum	0.000000
Maximum	19.000000
1. Quartile	10.000000
3. Quartile	14.000000
Mean	11.906009
Median	12.000000
Sum	7727.000000
SE Mean	0.126814
LCL Mean	11.578392
UCL Mean	12.233626
Variance	10.437140
Stdev	3.230656
Skewness	-0.908694
Kurtosis	2.664626

Table 2: For Portuguese

2.1.2.2.1 Mean Test for Portuguese

Let μ denotes the mean of the final grades of students taking Portuguese.

$H_0: \mu = 0 H_1: \mu \neq 0$

```
> t.test(df2$G3)
```

```
        One Sample t-test

data:  df2$G3
```

t = 93.885, df = 648, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
11.65699 12.15503
sample estimates:
mean of x
11.90601

Test statistics = 93.885

α = 0.01, p-value ≈ 0

With α = 0.01, we **reject** H_0 and failed to conclude that the mean of the final grades of students taking Portuguese is not equal to zero.

2.1.2.2.2 Skewness Test for Portuguese

Let S(x) denotes the skewness of the final grades of students taking Portuguese.

$H_0: S(x) = 0 H_1: S(x) \neq 0$

```
> s <- -0.908694
> s.test <- s/sqrt(6/649)
> pnorm(s.test)
[1] 1.682719e-21
> pv1 <- 2*(1-pnorm(s.test))
> pv1
[1] 2
> s.test
[1] -9.450709
```

Test statistics, S* = $\frac{-0.908694}{\sqrt{6/649}} = -9.450709$

α = 0.01, p-value ≈ 0

Since p-value < α, ∴ reject H_0.

With α = 0.01, we **reject** H_0 and conclude that the final grades of students taking Portuguese are significantly skewed to the left.

2.1.2.2.3 Kurtosis Test for Portuguese

Let K(x) denotes the skewness of the final grades of students taking Portuguese.

$H_0: K(x) - 3 = 0 H_1: K(x) - 3 \neq 0$

```
> k <- 2.664626
> k.test <- k/sqrt(24/649)
> pnorm(k.test)
[1] 1
> k.test
[1] 13.85648
```

Test statistics, K* = $\frac{4.991680}{\sqrt{24/649}} = 13.85648$

α = 0.01, p-value = 1. Since p-value > α, ∴ do not reject H_0.

With α = 0.01, we do not **reject** H_0 and conclude that of the final grades of students taking Portuguese have light tails.

2.1.2.2.4 Normal Test (Jarque-Bera) for Portuguese

```
> normalTest(df2$G3, method="jb")

Title:
 Jarque - Bera Normalality Test

Test Results:
  STATISTIC:
    X-squared: 284.2619
  P VALUE:
    Asymptotic p Value: < 2.2e-16
```

Test statistics, $JB = 284.2619$

α = 0.01, p-value ≈ 0

Since p-value < α, ∴ reject H_0.

With α = 0.05, we **reject** H_0 and conclude that the normality assumption of the final grades of students taking Portuguese is rejected.

The test above shows the final grade of students taking Portuguese course. For the mean test, the p-value is approximately equal to 0, smaller than the significance level of 0.01, which brings the mean of the final grade of students taking Portuguese course is not equals to zero. As we can see, the mean estimation of the sample is 11.90601, the grade given from 0 to 20, which the two Portuguese schools' students having an average on their final grade in Portuguese course.

For the skewness test, given the value of -9.450709, brings us negative skewness, which indicates that the mean of the data values in Portuguese course is less than the median, and the data distribution is skewed to the left. Heavy-tailed test results brings us the kurtosis test to show the skewness of the data distribution, other than the skewness test, that we can conclude the results is a negative kurtosis since it indicates a "light tailed" distribution. The "minus 3" at the end of this formula is often described as a correction to make the normal distribution kurtosis equal to zero, because for a normal distribution the kurtosis is three. Normal test, also known as Jarque-Bera test, brings us value of 284.2619, which means the larger the values of Jarque-Bera, the more the data deviates from the normal distribution of its dataset.

2.1.2.3 Analysis of correlation between final grade for Math and Portuguese

Pearson's Correlation test

```
Title:
 Pearson's Correlation Test

Test Results:
  PARAMETER:
    Degrees of Freedom: 380
  SAMPLE ESTIMATES:
    Correlation: 0.4803
  STATISTIC:
    t: 10.6761
  P VALUE:
    Alternative Two-Sided: < 2.2e-16
    Alternative      Less: 1
    Alternative   Greater: < 2.2e-16
  CONFIDENCE INTERVAL:
    Two-Sided: 0.3993, 0.554
         Less: -1, 0.5427
      Greater: 0.4128, 1

Description:
 Fri Mar 20 13:38:41 2020
```

Using correction analysis, the continuous and ordinal variables were analyzed. Parametric (Pearson Correlation) analyses were used to investigate the correlation between the students' performance on the final grade of both Mathematics and Portuguese courses. This is known as the best approach for calculating the correlation between interest variables, as it is based on the covariance method. This provides details about the association's significance, or correlation, and the path of the relationship. The degree of correlation is 0.4803, which leads to medium correlation. Medium degree of correlations shows the correlation of two series of data is neither large nor small. It could be the two variables somehow correlated to each other. So, the two-sided p-value is the key to this question. The Pearson correlation presents a p-value that is less than 0.01, it shows that the final grade for Mathematics course are corelated with the final grade for Portuguese course. Thus, the final grade for Mathematics course can represent the final grade for Portuguese course.

2.1.2.3.1 Scatterplots: analysis between final grades and 1st/2nd grade for Math

Scatterplots

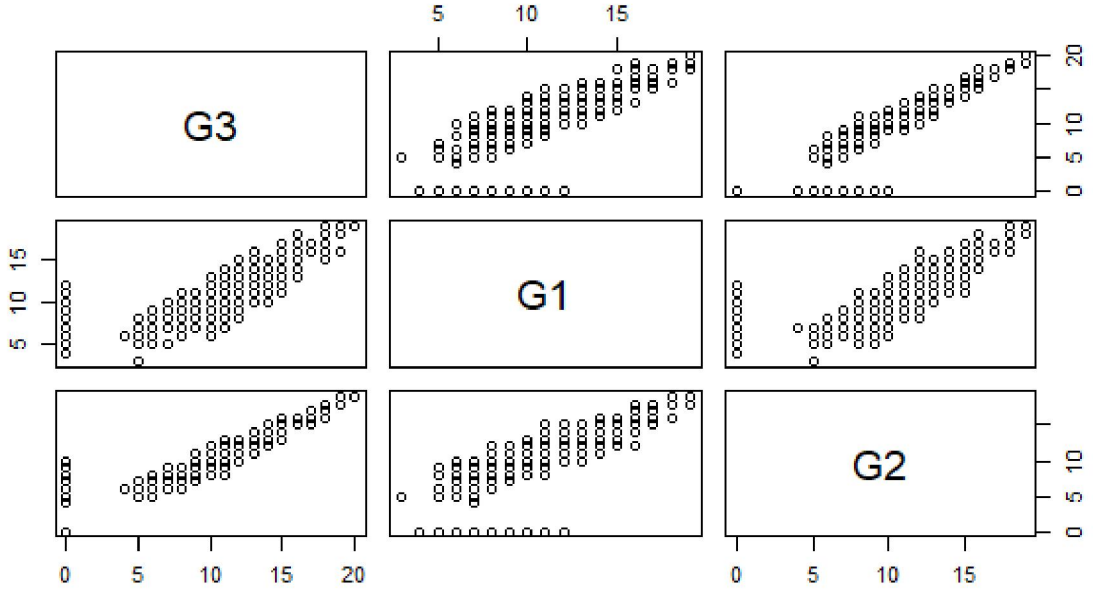


Table 3

A Scatter Analysis is used when you need to compare two data sets against each other to see if there is a relationship. Scatter plots are a way of visualizing the relationship; by plotting the data points you get a scattering of points on a graph (Hessing, n.d.). For Table 3, we compared the students' performance on final grade(G3) against first grade(G1) and second grade(G2) for Mathematics course using scatterplot. As comparison, the final grade of the students' performance on Mathematics course is increasing as first grade and second grade of the students' performance increases. The dataset shown in Table 3 is an uphill pattern from left to right, this indicates a positive relationship between first grade, second grade and final grade of students' performance. The diagram is in positive relationship, in a generally linear form and strong association between each other. Yet, the outliers of the student's final grade against first grade of the students' performance is more than then student's final grade against second grade of the students' performance. There's a total of 10 outliers in the plot of student's final grade against first grade that scores above 5 out of 20 for final grade, compared to the 8 outliers in the plot of student's final grade against second grade. Therefore, as the students' performance on first grade and second grade increase on their Mathematics course, their final grade will be increasing as well.

2.1.2.3.2 Scatterplots: analysis between final grades and 1st/2nd grade for Portuguese

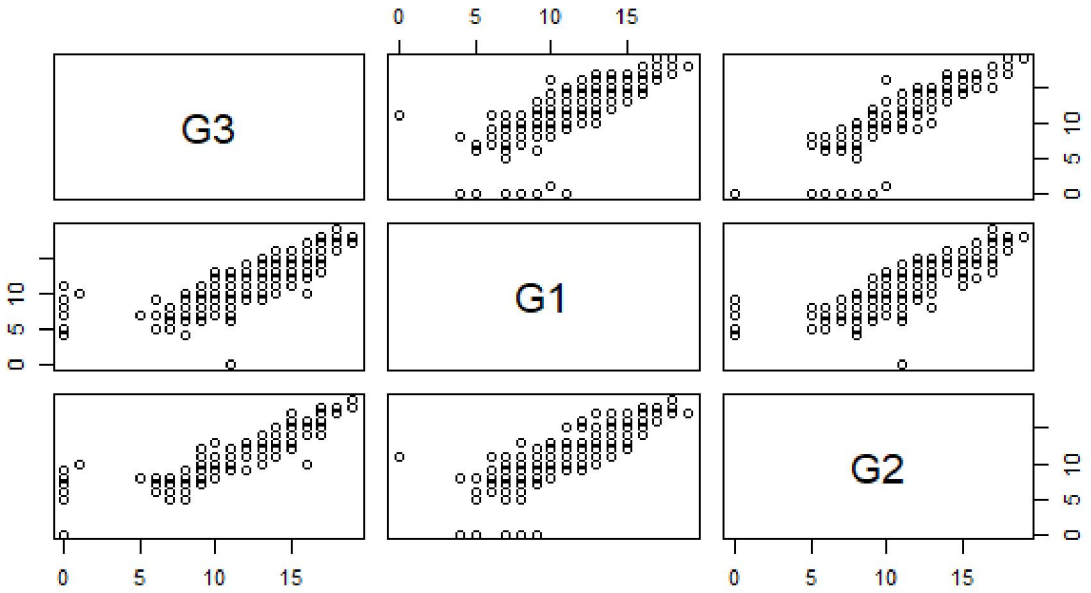


Table 4

We compared the students' performance on final grade(G3) against first grade(G1) and second grade(G2) for Portuguese course using scatterplot. The dataset shown in Table 3 is an uphill pattern as we move from left to right, this indicates a positive relationship between first grade, second grade and final grade of students' performance. The diagram is in positive relationship, in a generally linear form and strong association between each other. As comparison, the final grade of the students' performance on Portuguese course is increasing as first grade and second grade of the students' performance increases. The outliers of the student's final grade against first grade of the students' performance is more than then student's final grade against second grade of the students' performance. There's a total of 8 outliers in the plot of student's final grade against first grade compared to the 7 outliers in the plot of student's final grade against second grade. Therefore, as the students' performance on first grade and second grade increase on their Portuguese course, their final grade will be increasing as well.

2.1.2.3.3 Analysis of numeric social factor

Social factors are one of the possibilities that can affect the students' performance on their course in school. For our study, we determine whether the social factors will affect the students' performance in Mathematics course in two Portuguese schools.

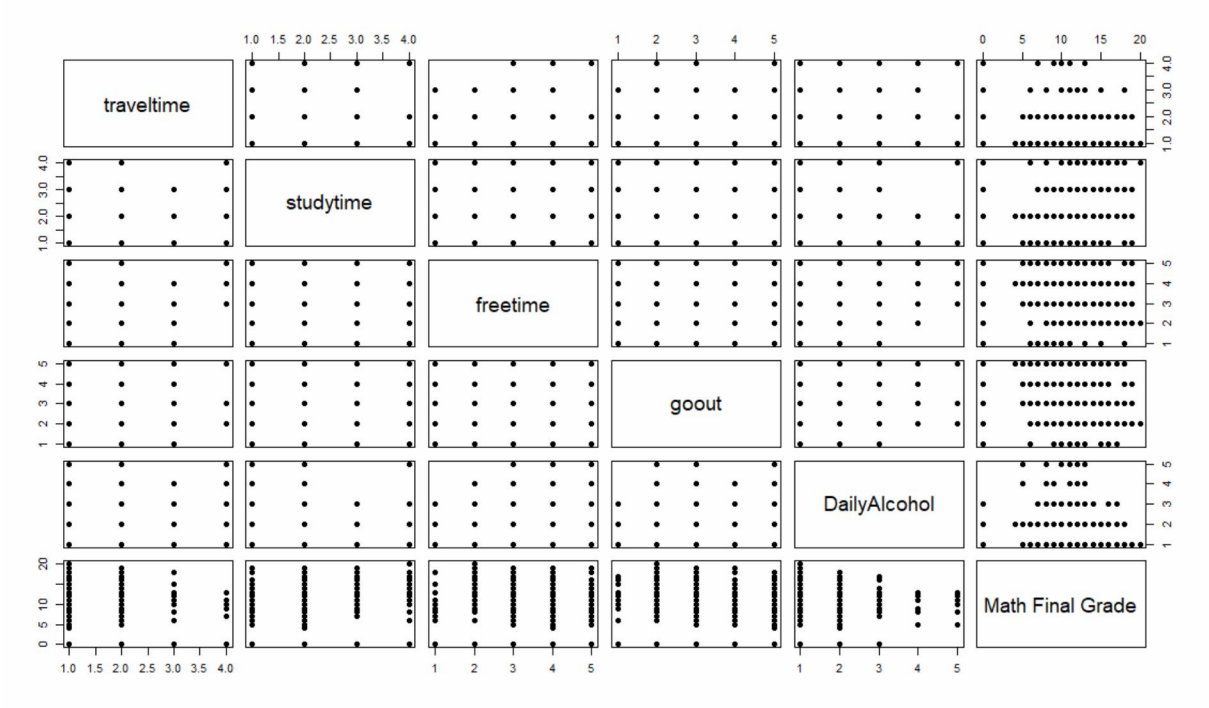


Table 5

Table 5 shows the relationship between all the social factors and the final grade of students' performance on Mathematics course. The travel time and study time of students is categorized as numeric number of 1 which is less than fifteen minutes, 2 which is in between fifteen and thirty minutes, 3 which is in between thirty minutes and one hour, lastly 4 is more than one hour. As the travel time increases, the number of students getting higher achievement in final grade on Mathematics course decreases. Thus, the longer time students travel, the lower final grades in Mathematics course the students had in average. Nevertheless, the study time for Mathematics course is positively associated with the achievement of students in their final grade for Mathematics course. The longer time the students take to do revision on Mathematics course, the better the result they can get. Still, the free time after school the students having more, the lower the final grade they got in Mathematics course. Furthermore, the longer time they spent with friends, the worse they got the result for the final grade in Mathematics course. Besides, daily alcohol assumption is worse as it will significantly be decreasing on the students' performance in Mathematics course. The higher amount they consume alcohol, the lower the students' performance in Mathematics course in Portuguese.

2.1.2.3.4 Analysis of other numeric factors

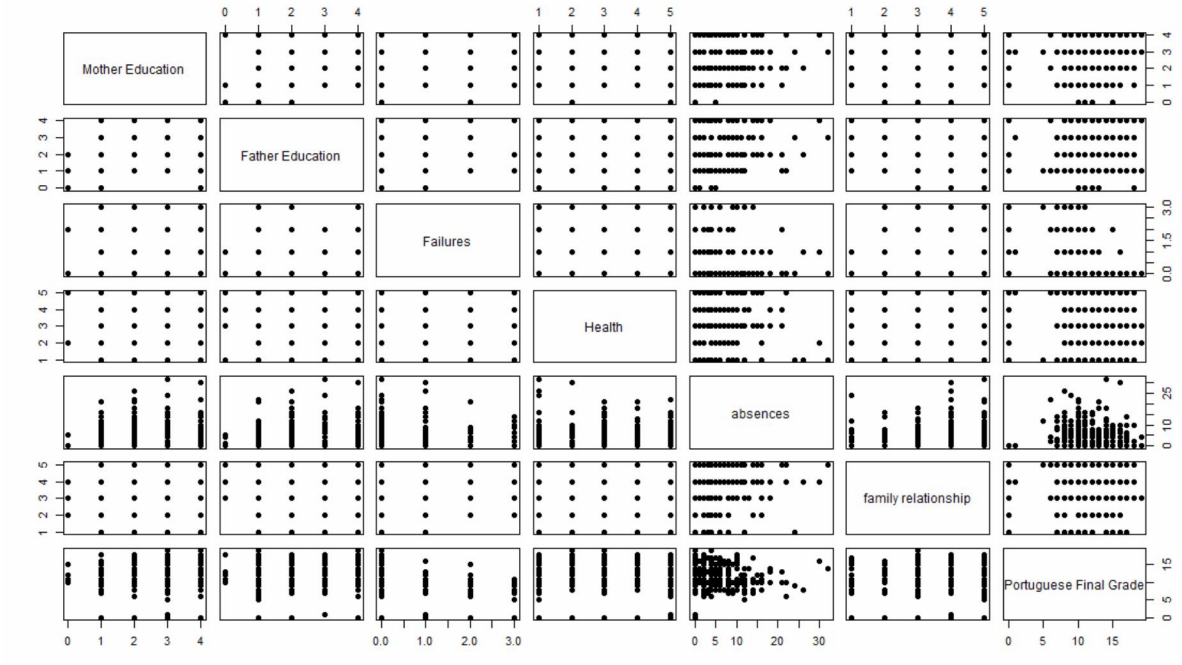
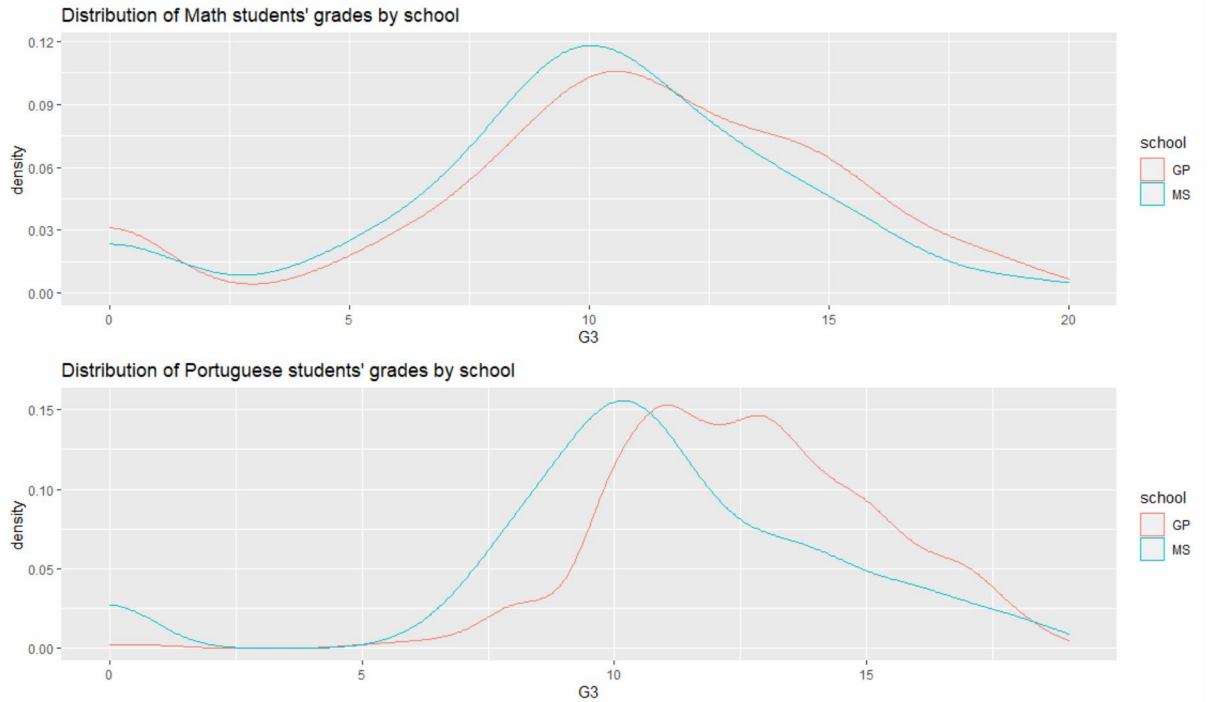


Table 6

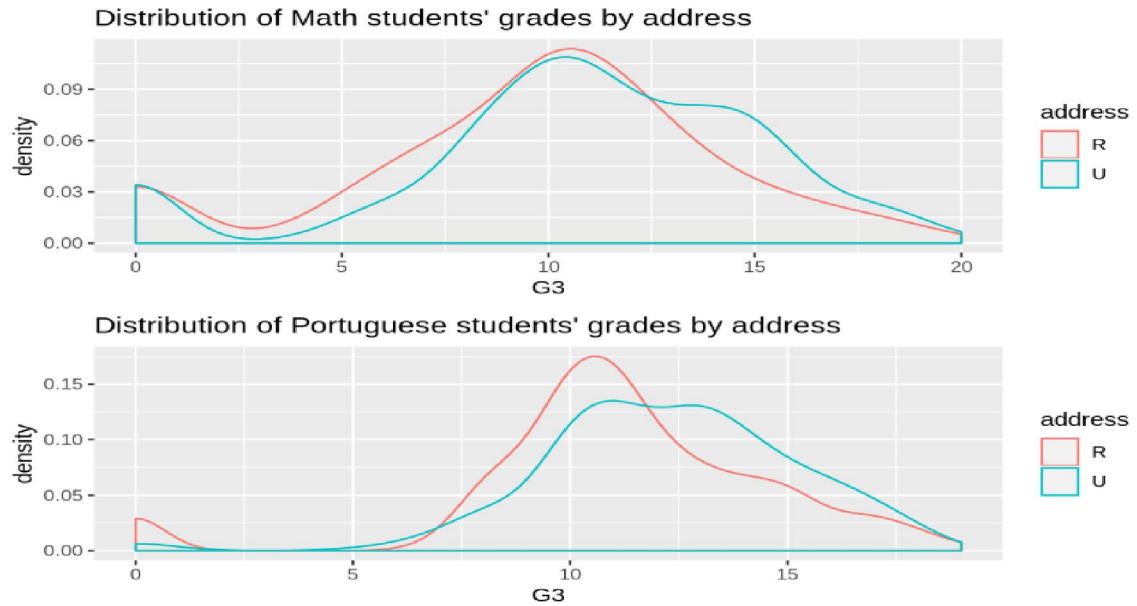
Table 6 shows other than the social factors that might affect the students' performance on final grade of Portuguese course. First, mother's education is significant as it can affect their child's performance in school. The higher mother's education is, the higher number of students having higher achievement in the final grade in Portuguese course. Same goes to father's education as both parents are the role model of their children, father and mother carries the same job and responsibilities in teaching their children, thus father and mothers' education can affect students' performance as well. The higher father's education is, the higher number of students having higher achievement in the final grade in Portuguese course. Notwithstanding, the past class failures is affecting much to students' performance in the final grade in Portuguese course. The higher number of past class failures students have, the lower the number of students getting higher achievement in the final grade in Portuguese course. Students with good school and good social connectedness are less likely to experience subsequent mental health issues and be involved in health risk behaviours, and are more likely to have good educational outcomes (Bond et al., 2007). The healthier the students, the better the result they have in Portuguese course. Furthermore, the number of school absences strongly affecting the students' performance in Portuguese course. The number of school absences the students significantly proved by non-linear form of scatterplot, showing that the higher the classes students absent, the lower number of students having higher achievement in the final grade in Portuguese course. Nevertheless, the family relationship can help on the students' performance a lot. Good communication at home influences a student's understanding of language a lot. The plot shows that the better relationship the students have in their family, the higher number of students having higher achievement in the final grade in Portuguese course.

2.1.3 Basic Descriptive Analysis (Categorical)

2.1.3.1 Distribution: Analysis by school



2.1.3.2 Distribution: Analysis by address



The distribution of math grades between the two schools is similar, but Gabriel Pereira's students outperform those of Mousinho da Silveira in Portugues Students living in more urban areas appear to outperform all Mathematics and Portuguese students from rural areas. Given that Gabriel Pereira is in a more urban location and that Mousinho da Silveira is more rural, this plot is in line with the plot before.

2.1.4 Chi-squared test of association

A chi-squared test is performed to test whether the following factors significantly affect the student results, at 1% significant level.

Ho: There is no association between the two variables.

H1: There is an association between the two variables.

```
        Pearson's Chi-squared test with Yates' continuity correction

data:  df1$activities and mathresult
X-squared = 0.019497, df = 1, p-value = 0.889
```

Table 1

For Table 1, the chi-squared test is performed using two variables which is the students' extra-curricular activities and final grade for Mathematics course. We need to test whether extra-curricular activities for students participating in school will affect the students' performance on their final grade of Mathematics course. Since p-value of 0.889 as shown in Table 1, is greater than the significance level of 0.01, we do not reject H_0. There is insignificant evidence to conclude that the extra-curricular activities and Mathematics course are independent, indicating extra-curricular activities are not affecting the students' performance on their final grade of Mathematics course.

```
        Pearson's Chi-squared test with Yates' continuity correction

data:  df1$romantic and mathresult
X-squared = 3.3451, df = 1, p-value = 0.0674
```

Table 2

For Table 2, the chi-squared test is performed using two variables which is the students' romantic relationship and final grade for Mathematics course. We need to test whether student's romantic relationship will affect the students' performance on their final grade of Mathematics course. Since p-value of 0.0674 as shown in Table 2, is greater than the significance level of 0.01, we do not reject H_0. There is insignificant evidence to conclude that the student's romantic relationship and Mathematics course are independent, indicating that whether student is in a romantic relationship or vice versa are not affecting the students' performance on their final grade of Mathematics course.

```
        Pearson's Chi-squared test with Yates' continuity correction

data:  df2$internet and portugueseresult
X-squared = 4.4887, df = 1, p-value = 0.03412
```

Table 3

For Table 3, the chi-squared test is performed using two variables which is the students' using Internet access at home and final grade for Portuguese course. We need to test whether student's using Internet access at home will affect the students' performance on their final grade of Portuguese course. Since p-value of 0.03412 as shown in Table 3, is greater than the significance level of 0.01, we do not reject H_0. There is insignificant evidence to conclude that the student's using Internet access at home and Portuguese course are independent, indicating that whether how long does

a student using Internet access at home of these two Portuguese schools are not affecting the students' performance on their final grade of Portuguese course.

```
        Pearson's Chi-squared test

data:  df2$reason and portugueseresult
X-squared = 17.988, df = 3, p-value = 0.0004423
```

Table 4

For Table 4, the chi-squared test is performed using two variables which is the reasons students chose their school and final grade for Portuguese course. We need to test whether the reasons student choose their school will affect the students' performance on their final grade of Portuguese course. Since p-value of 0.0004423 as shown in Table 4, is less than the significance level of 0.01, we reject H_0. There is insignificant evidence to conclude that the reasons student chooses their school and Portuguese course are independent. This indicates the reasons students chose either Gabriel Pereira or Mousinho da Silveira are affecting the students' performance on their final grade of Portuguese course. The mainly reasons could be the school they chose are close to their resident area, on school reputations, course they preferred on and others.

2.2 Correlation Analysis

2.2.1 Basic Introduction

Correlation Analysis comes in handy when the huge number of attributes are given. Not only it helps to analyze whether the target variables are correlated to each independent variable, it also helps to identify whether there is more than one interaction within the independent variables themselves only, i.e. without considering response variables. (STHDA, 2020)

In these 2 data sets, many attributes (independent variables) are given. Therefore, we are interested to use this unsupervised learning approach to perform correlation analysis. It helps to identify the factors affecting students, as mentioned in *1.1.4 Problem Statement*.

We first build a correlation matrix. Before analyzing the relationship of the variables, we perform Bartlett's Test of Sphericity (Admin, 2020). Bartlett's test of sphericity tests whether your correlation matrix is an identity matrix. If it is identity matrix, the variables in the Student Performance data sets are unrelated and therefore unsuitable for PCA analysis.

2.2.2 Bartlett's Test of Sphericity

Ho: the variables are orthogonal, i.e. not correlated. PCA analysis is not applicable.

H_1: the variables are not orthogonal, i.e. they are correlated enough to where the correlation matrix diverges significantly from the identity matrix.

```
> bart_spher(corMat, use = "everything")
Bartlett's Test of Sphericity
Call: bart_spher(x = corMat, use = "everything")
X2 = 17186.397
df = 2016
p-value < 2.22e-16
```

Based on the output above, and with 1% significant level, [alpha= 1% gives better accuracy], p- value is approximately 0, which is also less than 0.01.

Ho is rejected, therefore we have significant evidence to conclude that the correlation matrix is not an identity matrix. This correlation matrix can be used to perform correlation analysis and Principal Component Analysis.

2.2.3 Correlation Analysis- Fact Checking

By using 1% significant level, we performed a fact checking on the data set. 2 asterisks (**) means the two variables are related to each other at 1% significant level.

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Variable 1	Variable 2	Significant Code
nurseryyes	G3.x	(empty)
nurseryyes	G3.y	(empty)
paid.xyes	G3.x	*
paid.yyes	G3.y	**

Nursery will not affect the student's performance at all. Strictly speaking, if the student paid for extra Math class, it does not affect the final Math result. At 1% significant level, if the student paid for extra Portuguese class, (by common sense) he will get better result for Portuguese subject.

Variable 1	Variable 2	Significant Code
Fedu	Medu	***
Fedu	G3.y	***
Medu	G3.x	***
Fedu	famsup.xyes	*
Medu	famsup.yyes	*
G3.x	G3.y	***

The low/high education level of father and mother are correlated to the final result of the students, which implies that the final student's result for Portuguese and Math is dependent of their education level at 1% significant level. However, low/high

education level of the parents will not decide whether student receives family support or not. (It is about parent's love, by common sense.)

Variable 1	Variable 2	Significant Code
goout	Dalc	***
Dalc	Walc	***
Dalc	Health	(empty)
health.x	G3.x	.
health.y	G3.y	**

If student went out with friends, they would drink alcohol. Surprisingly, daily alcohol will not have direct effect on student's health as they are young. From the p-value of correlation matrix, if the student suffers from health problem, it would only affect the academic performance of Portuguese language at 1% significant level, but not for Math. More researches can be carried out to investigate this phenomenon.

Variable 1	Variable 2	Significant Code
studytime.x	G3.x	**
studytime.y	G3.y	***
absences.x	G3.x	.
absences.y	G3.y	***

Duration of study can reflect the student's attitude on studying. From the table above, it is observed that the attitude of the student must be good in order to get a good result in their Math and Portuguese finals. Interestingly at 1% significant level, there is no evidence to show that number of absences in Math class can affect their Math exam. This is logic as students can study Math at home. As long as students understand the concepts, they can still get a good result.

Variable 1	Variable 2	Highest Significant Code
guardian.x [mother and other]	G3.x	.
guardian.y [mother and other]	G3.y	.
famrel.x	G3.x	(empty)
famrel.y	G3.y	(empty)

From the table above, it shows that guardian and family relationship will not have effect on the final results for both Math and Portuguese subjects.

2.2.4 Correlation Analysis- Identify Most Significant Factors for Math

At the significant level of 1%, we can see that the most significant factors that affect student's Math final results are **age, addressU, Medu, Fedu, Mjob[other and health], Fjob[teacher], traveltime.x, studytime.x, failures.x, higher.xyes, romantic.x, gout.x, Dalc.x and Walc.x.**

A possible interpretation why these are the most significant factors is:

As the age increases, Math is more difficult to understand. If student lives in urban area, they could get enough resources, comfortable environment to do revision on Math. Based on the data set, when students put more time in socializing (include getting a boyfriend/girlfriend) instead of spending more time to do revision in Math, they could have difficulty in answering the exam questions as they didn't not practice much. If the student takes more time to travel far from home to school, it would indirectly affect the student's mood and students would have like time to study. If parents are well educated, they can assist students in academic performance, whereas it is unknown why parent's job can somehow affect the student's result.

2.2.5 Correlation Analysis- Identify Most Significant Factors for Portuguese

At the significant level of 1%, we can see that the most affected student's Portuguese final result are **school[GP and MS], sex[M], age, addressU, Medu, Fedu, Mjob[other and health], reason[reputation], traveltime.y, studytime.y, failures.y, paid.yyes, higher.yyes, freetime.y, gout.y, Dalc.y and Walc.y, health.y and absence.y**

A possible interpretation why these are the most significant factors is:

Indeed, when taking language subjects, it would be good to choose a well-known school as language itself is hardly to master without proper guidance from teachers. It looks like male students prefer language subject, and the reason is yet to be found. Similar with Math, students should balance their academic lifestyle and social lifestyle as these would affect student's final result. Students who travel far from home can somehow affect the result for languages subject. If the students can pay for extra classes and have motivation to pursue higher education, these would help them to pass the Portuguese exam. In addition, if students have more free time, they could make use of the time to polish the Portuguese language skills or release stress, as it could help students to pass Portuguese exam as well.

2.2.6 Weaknesses and Strengths of Correlation Analysis

2.2.6.1 Strength

Correlation analysis is useful as a point for further research. It can assist researchers to perform very detailed analysis.

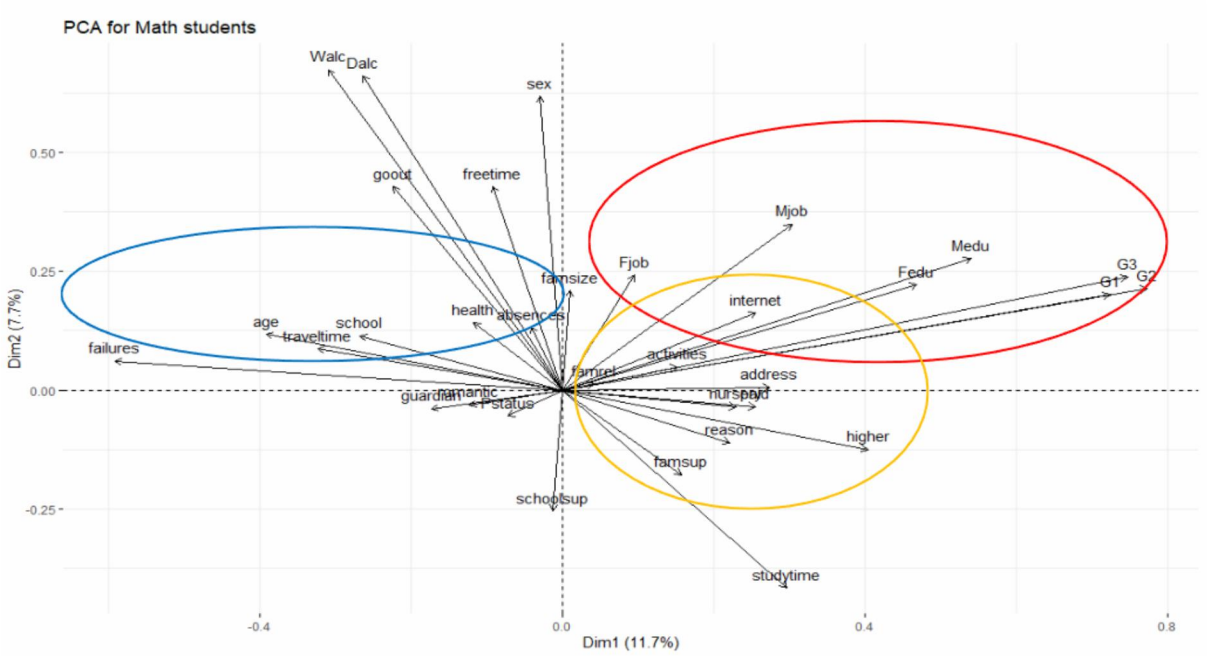
2.2.6.2 Weakness

The correlation analysis tests for the linear relationship between two variables, therefore the significance codes in the matrix we used may be misleading. No (linear) correlation doesn't mean it also has no quadratic correlation and so on.

2.3 Principal Component Analysis

2.3.1 Math students

2.3.1.1 Analysis of Biplot – Determining factors affecting result for Math



Based on the biplot above, the predictor variables circled in red have stronger positive association with the final grade (G3) of the students taking Math. The variables in blue circle would have stronger negative association with G3 for Math. The elements in orange circle seem to have positive association with G3 as they are going in approximately same direction but the strength of association is insignificant.

{Mjob, Fjob, Medu, Fedu, activities, internet, address, G1, G2} ⊆ Red Circle

{guardian, romantic, Pstatus, failures, age, traveltime, school} ⊆ Blue Circle

{nursery, reason, higher, famsup, studytime} ⊆ Orange Circle

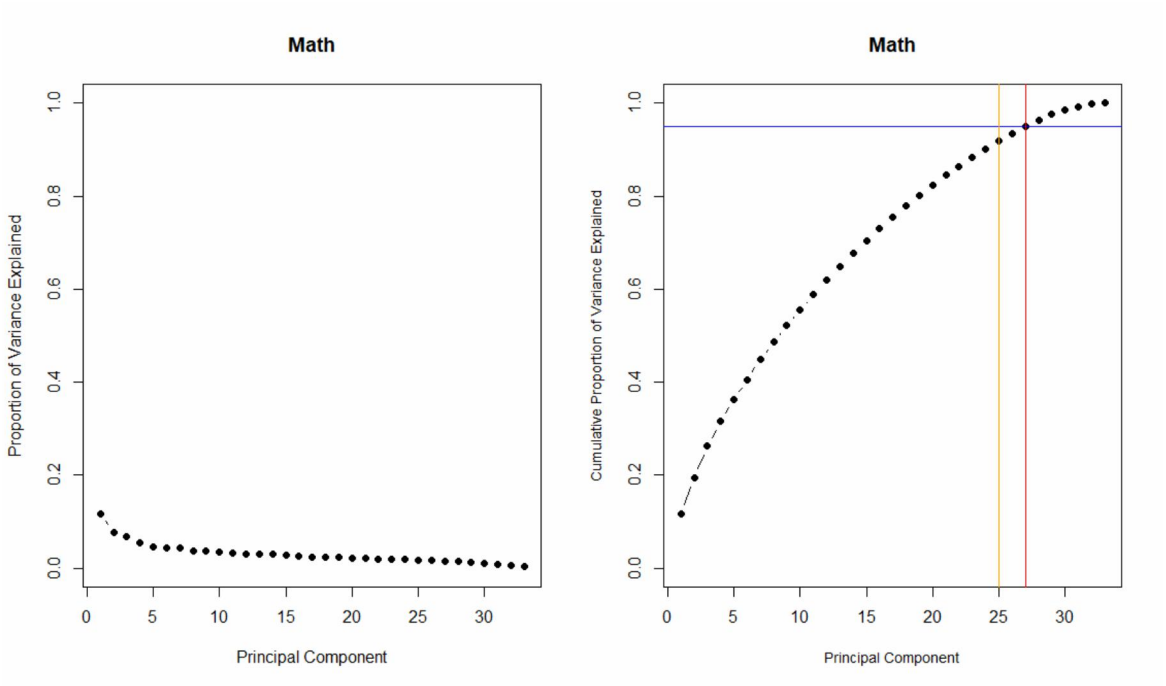
The elements not circled are not affecting student's overall result for Math significantly.

We can interpret in many ways based on the elements in red and blue circle. For example, if the student has internet access at home, the student can use google to find resources to study, therefore the final result is going to be better. If the student is having a romantic relationship, it might be difficult for the student to focus on study and hence fail the Math exam.

2.3.1.2 Analysis of Principal Component for Math

```
> summary(mathMod.pc)
Importance of components:
                          PC1     PC2     PC3     PC4
Standard deviation     1.9680 1.59348 1.49302 1.33405
Proportion of Variance 0.1174 0.07694 0.06755 0.05393
Cumulative Proportion  0.1174 0.19431 0.26186 0.31579
```

PC1(The first principal component) has a proportion of variance (PVE) of 0.1174. It means that one principal component can only explain 11.74% of variation of the data. This is not good as we oversimplified the problem, so we need more principal components (PC) so that the prediction can be more accurate. PC2 and PC1 helps to explain only 19.43% of variation (which is not good enough), therefore more PCs until we can explain at least 95% of variation.




```
> abline(h=0.95,col="blue") #h= horizontal, v=vertical
```

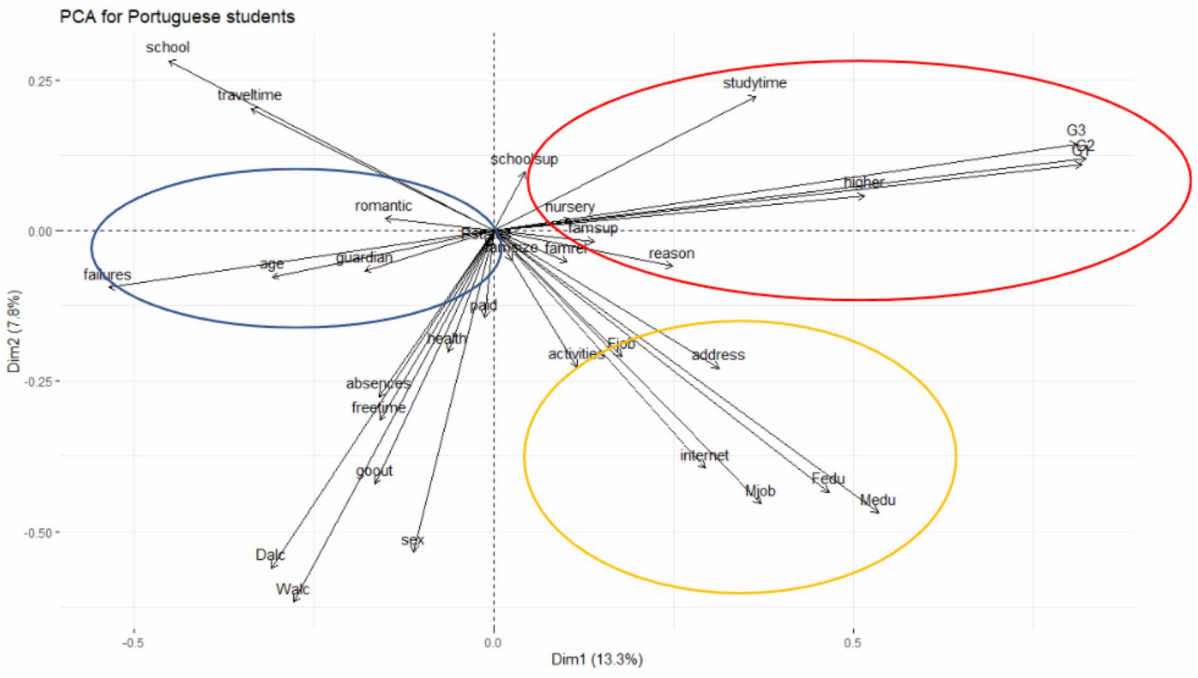
Blue line is $y = 0.95$

Based on the Scree plots above, we need at least 27 out of 33 Principal Components to explain at least 95% of variation. Only 6 dimensions are allowed to be reduced, where 1 dimension equals one principal component.

Therefore, the 395 observations of students taking Math may not be a good data set as the dimension is still large, i.e. 27-Dimension.

2.3.2 For Portuguese

2.3.2.1 Analysis of Biplot – Determining factors affecting result for Portuguese



Based on the biplot above, the predictor variables circled in red have stronger positive association with the final grade (G3) of the students taking Portuguese. The variables in blue circle would have stronger negative association with G3 for Portuguese. The elements in orange circle seem to have positive association with G3 as they are going in approximately same direction but the strength of association is insignificant.

{schoolsup, studytime, nursery, famsup, higher, G1, G2} ⊆ Red Circle

{romantic, failures, age, guardian} ⊆ Blue Circle

{activities, internet, Fjob, Mjob, Fedu, Medu, address} ⊆ Orange Circle

The elements not circled are not affecting student's overall result for Portuguese significantly.

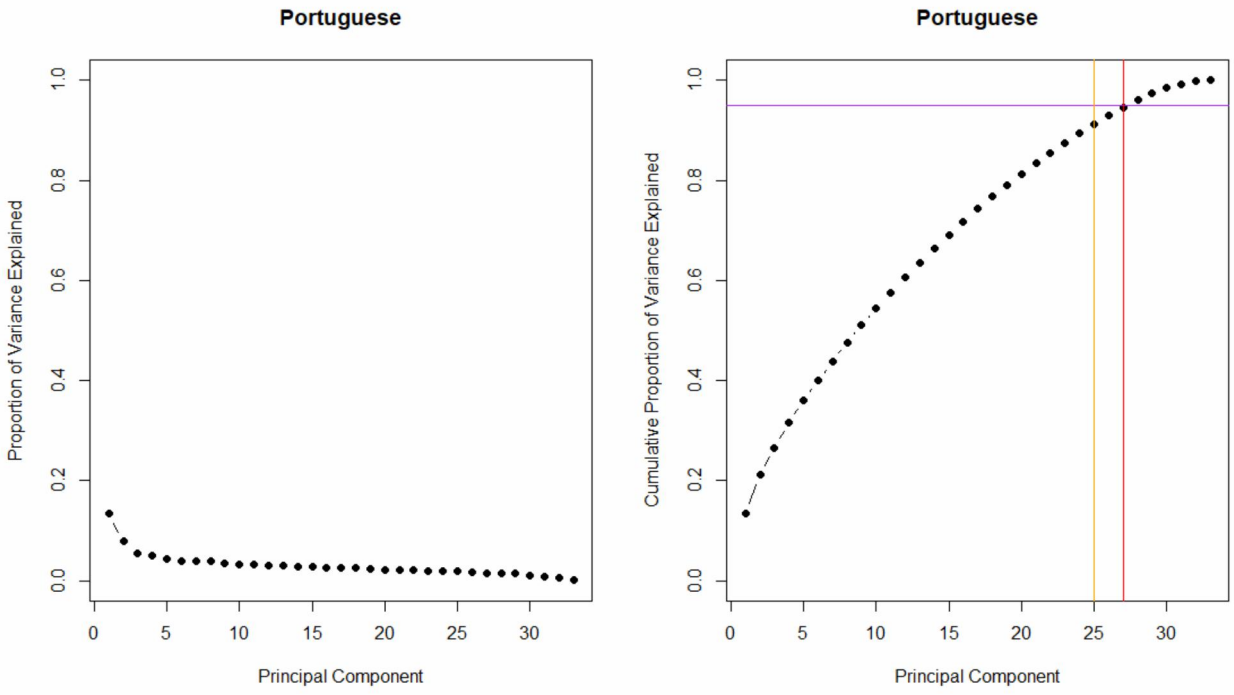
We can interpret in many ways based on the elements in red and blue circle. For example,

If the student receives school's educational support, the final result is going to be better. If the student is having romantic relationship, it might be difficult for the student to focus on study and hence fail the Portuguese exam.

2.3.2.2: Analysis of Principal Component for Portuguese

```
> summary(portMod.pc)
Importance of components:
                          PC1     PC2     PC3     PC4
Standard deviation     2.0986 1.60465 1.33318 1.29705
Proportion of Variance 0.1335 0.07803 0.05386 0.05098
Cumulative Proportion  0.1335 0.21148 0.26534 0.31632
```

PC1(The first principal component) has a proportion of variance (PVE) of 0.1335. It means that one principal component can only explain 13.35% of variation of the data. This is not good as we oversimplified the problem, so we need more principal components (PC) so that the prediction can be more accurate. PC2 and PC1 helps to explain only 21.15% of variation (which is not good enough), therefore more PCs until we can explain at least 95% of variation.



> abline(h=0.95,col="purple") #h= horizontal, v=vertical

Purple line is $y = 0.95$

Based on the Scree plots above, we need at least 27 out of 33 Principal Components to explain at least 95% of variation. Only 6 dimensions are allowed to be reduced, where 1 dimension equals one principal component.

Therefore, the 649 observations of students taking Portuguese may not be a good data set as the dimension is still large, i.e. 27-Dimension.

2.3.3: Weakness and Strength

2.3.3.1 Strength

PCA could be a versatile technique that works well in practice. It is fast and straightforward to implement. It is easy to test the algorithms with and without PCA to check performance. In additional, PCA offers several variations and extensions to tackle specific roadblocks.

2.3.3.2 Weakness

The new principal components aren't interpretable, which can be a deal-breaker in some settings. Additionally, it is required to manually set or tune a threshold for cumulative explained variance.

CHAPTER 3 SUPERVISED LEARNING

3.1 Linear Regression Model

In this section, the original data sets, "student-mat" and "student-por" are used to predict the final grade of a student taking mathematics and Portuguese respectively. Firstly, data cleansing, reshuffling, resampling is performed using R software to avoid overfitting problem.

There are certain conditions that must be fulfilled in order to verify that the data set is adequate to build a linear regression model, i.e. linearity must be accepted, variation of the observation around regression line is constant (homoscedastic) , normality of the target output (G3) must be accepted (MarinStatsLectures, 2013).

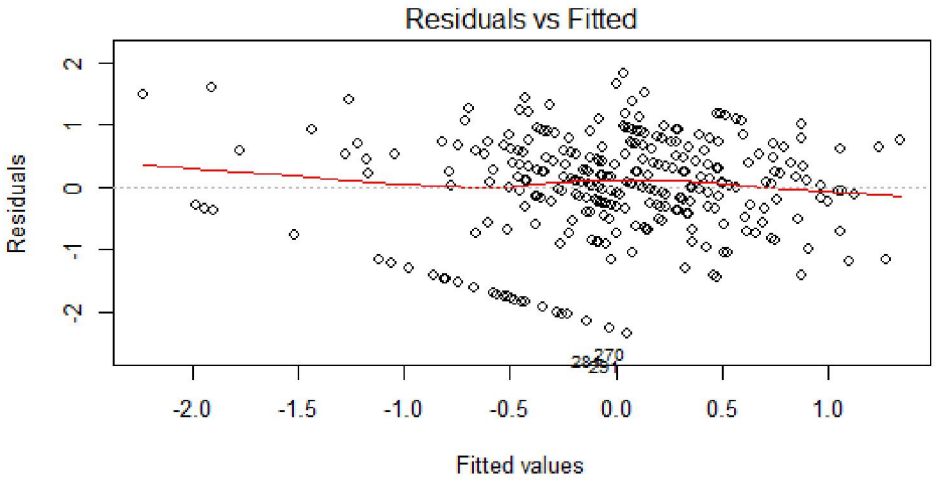
After several attempts using different set.seed in R command, the plots give the same conclusion.

3.1.1 Math students

Consider set.seed(123) and generate the plots,

The plots below were generated by R software for inspections.

- Checking linearity



Rule of Thumb: If the linearity is accepted, the red line in this plot should be fairly straight. (MarinStatsLectures, 2013).

By inspecting this plot and strictly speaking, the red line appears to be little curvilinear instead of flat. Linearity is rejected.

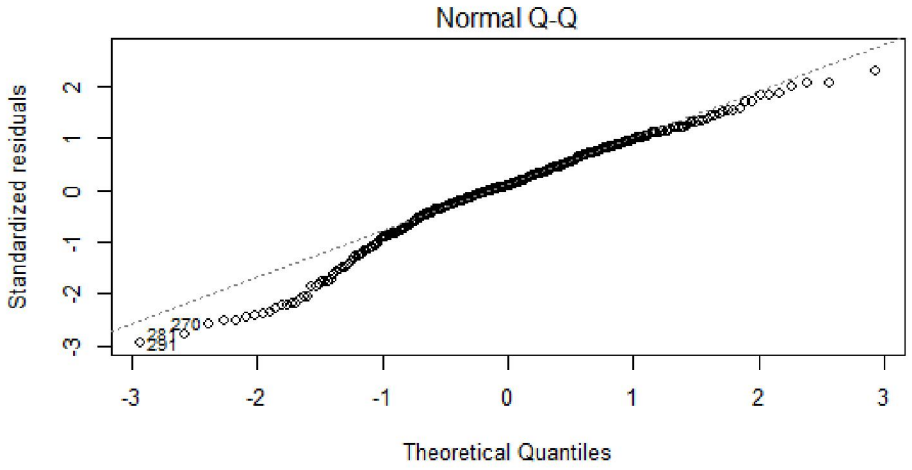
- Checking constancy of the variance (homoscedasticity)

Rule of thumb: all observations should form a "cloud" pattern, i.e. distributed evenly around the regression line. (MarinStatsLectures, 2013).

Using the same "Residual vs Fitted" plot, we observed a straight line with negative gradient below the regression line. Strictly speaking, the observations are not distributed evenly and therefore it is concluded that the variance of the observation is not constant.

- Checking normality of the error term

Rule of thumb: In Quantile-Quantile plot, all observations should lie around the dotted line (MarinStatsLectures, 2013), and the observations at both ends should converge back to the dotted line (UECM2263 Applied Statistical Model).



Based on this QQ plot, the observations tend to converge back to the dotted line at both ends, however most of the observations did not fall onto the dotted line. It should be concluded that the normality is rejected.

To further verify the normality of the error terms, Shapiro test is performed. (STHDA, 2020)

Shapiro Test provides the following hypotheses:

H_0: The sample distribution of the data of the target output (G3) is normal

H_1: The sample distribution of the data of the target output (G3) is not normal

We say "sample distribution" as the data set consists of only around 600 data whereas the population should be more than 600. And also, our data is split into training data and testing data. It should be valid to say "sample" too as we used training data to build the model instead.

From the output below,

```
Shapiro-Wilk normality test

data:  train.mat$G3
W = 0.93018, p-value = 1.518e-10
```

At 1% significant level, p-value is approximately 0, which is less than 0.01.

H_0 is rejected. We have significant evidence to say that the sample distribution is not normal.

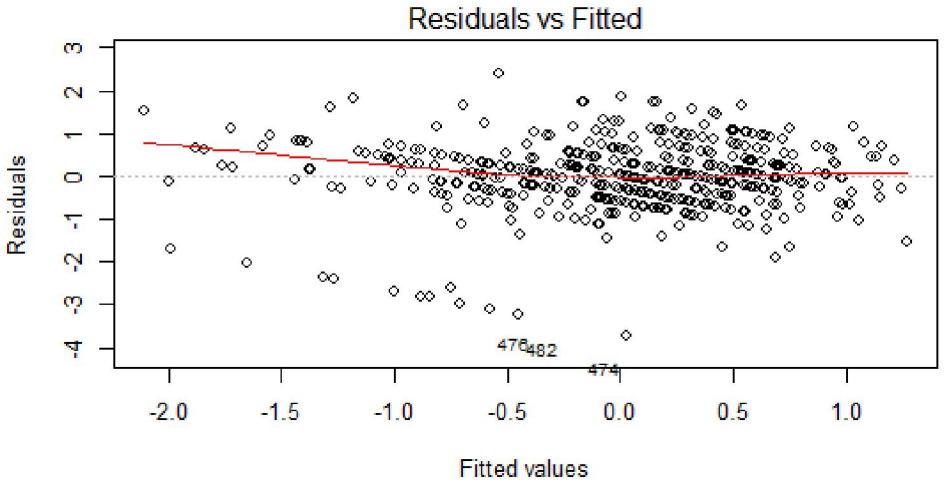
In short, the data set "student-mat" is not suitable for building linear regression model as all assumptions required are violated. We can use alternative approaches to build a predictive model instead.

3.1.2 Portuguese students

Consider set.seed(123) and generate the plots,

The plots below were generated by R software for inspections.

- Checking linearity



Rule of Thumb: If the linearity is accepted, the red line in this plot should be fairly straight.

By inspecting this plot and strictly speaking, the red line does not appear to be flat enough. Linearity is rejected.

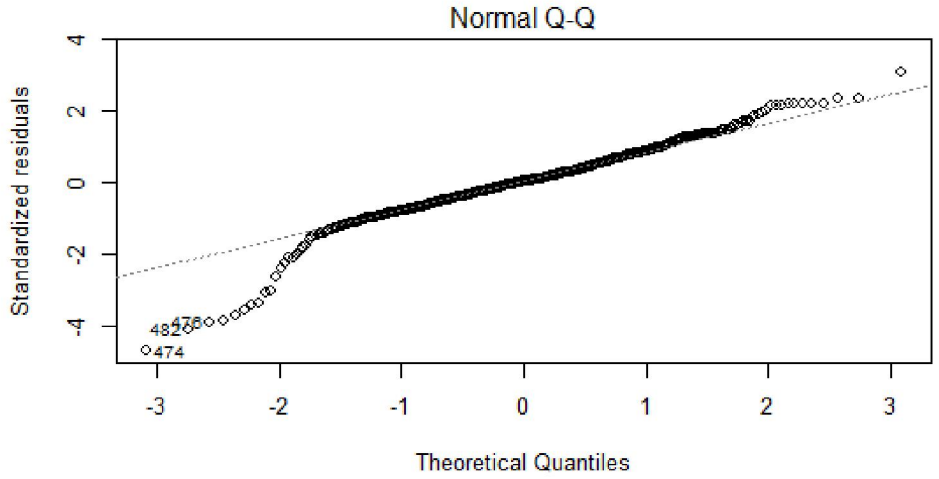
- Checking constancy of the variance (homoscedasticity)

Rule of thumb: all observations should form a “cloud” pattern, and there is no pattern in the plot.

Using the same “Residual vs Fitted” plot, we observed many straight lines with negative gradient. Strictly speaking, as there exists at least one pattern (straight line) instead of scattering around, therefore it is concluded that the variance of the observation is not constant.

- Checking normality of the error term

Rule of thumb: In Quantile-Quantile plot, all observations should lie around the dotted line, and the observations at both ends should converge back to the dotted line.



Based on this QQ plot, the observations tend to diverge from the dotted line at one end. In addition, most of the observations did not fall onto the dotted line. It should be concluded that the normality is rejected.

To further verify the normality of the error terms, Shapiro test is performed.

From the output below,

```
Shapiro-Wilk normality test

data:  train.por$G3
W = 0.92742, p-value = 1.27e-14
```

At 1% significant level, p-value is also close to 0, which is less than 0.01. H_0 is rejected. We have significant evidence to say that the sample distribution is not normal.

In short, the data set "student-por" is also not suitable for building linear regression model as all assumptions required are violated. We can use alternative approaches to build a predictive model instead.

3.1.3 Weaknesses and Strengths of Linear Model approach

3.1.3.1 Strength

Linear regression model comes in handy when the predictor variables and response variable has a linear relationship in between The linear model would be great to estimate the outputs without taking much consideration of the complicated details of the training model. It is great for supervised learning process. (Halthor, A. 2017).

3.1.3.2 Weakness

Linear regression model would over simplifies many real-world application problems and it is not recommended to use this approach

when dealing with actual scenario. In actual world, these independent variables/ predictor variables usually have either no relationship or more than a hidden relationship and interaction in between that would significantly affect the accuracy of the predicted outputs. Sometimes, these independent variables do not exist in a linear relationship, but quadratic relationship and so on. (Halthor, A. 2017). It is not recommended for noisy data like these 2 data sets according to Venkat Reddy (2020), even though we performed data resampling.

3.2 K-Nearest Neighbors

3.2.1 Math students

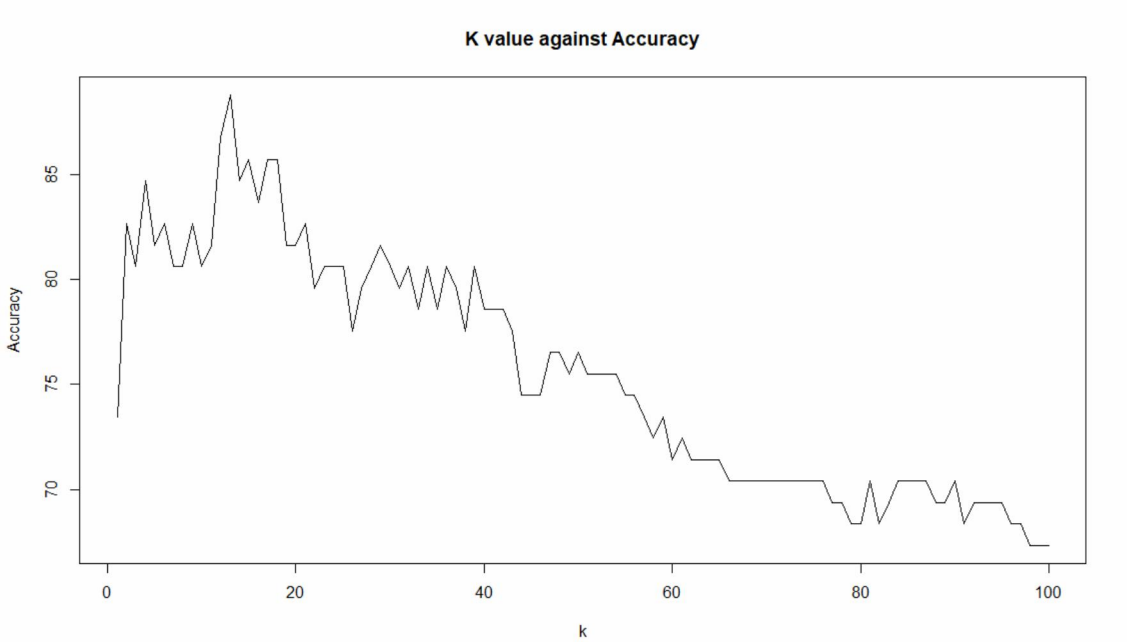
Consider set.seed(300) in R command,

```
> confusionMatrix(cm)
Confusion Matrix and Statistics
knn_model  0  1
        0 16  3
        1 16 63
               Accuracy : 0.8061
                 95% CI : (0.7139, 0.879)
    No Information Rate : 0.6735
    P-Value [Acc > NIR] : 0.002576

                  Kappa : 0.5077

 Mcnemar's Test P-Value : 0.005905

            Sensitivity : 0.5000
            Specificity : 0.9545
         Pos Pred Value : 0.8421
         Neg Pred Value : 0.7975
             Prevalence : 0.3265
         Detection Rate : 0.1633
   Detection Prevalence : 0.1939
      Balanced Accuracy : 0.7273

       'Positive' Class : 0
```

The output above shows the confusion matrix of our predictions and some evaluations of the confusion matrix. From the output above, it is observed that the accuracy of the model is 80.61%. when the value of k is 7. This indicates that the model has 80.61% chance to predict the output correctly when one set of input data is given.

The sensitivity of this model is 0.50 and this means that whenever a student is classified as failed by the model, there is a 50% probability that it is correct. The specificity of the model is 0.9545. This shows that 95.45% of the students that passed the Math subject is identified correctly. From these two measures, we can see that

this k-NN model is suitable in classifying the students that passed the Math accurately.



In order to increase the accuracy of our model, we need to find the best k value. The output above is generated by looping through all the k values from 1 to 100. We can observe that the optimum value of k is 13 since it has the highest accuracy which is approximately 88.78%. From the graph, we can see that the line reached a peak when k equals to 13 and started to decrease gradually when the value of k increases. The R command below verifies it.

```
> max(k_accuracy)
[1] 88.77551

> which.max(k_accuracy)
[1] 13
```

When the k value is small, overfitting might happen and increase the variation of the predictions. Diagram below shows the confusion matrix for the KNN model using k = 13. We can see that the correct number of predictions is higher.

```
> confusionMatrix(cm)

Confusion Matrix and Statistics

knn_model  0  1
        0 21  0
        1 11 66

          Accuracy : 0.8878
            95% CI : (0.808, 0.9426)
   No Information Rate : 0.6735
   P-Value [Acc > NIR] : 7.778e-07

            Kappa : 0.72
```

```
Mcnemar's Test P-Value : 0.002569

            Sensitivity : 0.6562
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 0.8571
             Prevalence : 0.3265
         Detection Rate : 0.2143
   Detection Prevalence : 0.2143
      Balanced Accuracy : 0.8281

       'Positive' Class : 0
```

Note that the sensitivity greatly improved to 0.6562 compared to accuracy measures corresponding to k=7. The 13-NN model has a 65.62% probability to classify a student who
failed in Math correctly. In addition, the 13-NN can exactly predict which student will pass the Math subject.

In short, we select 13-NN model as our candidate model for Math.

3.2.2 Portuguese student

Consider set.seed(300) in R command,

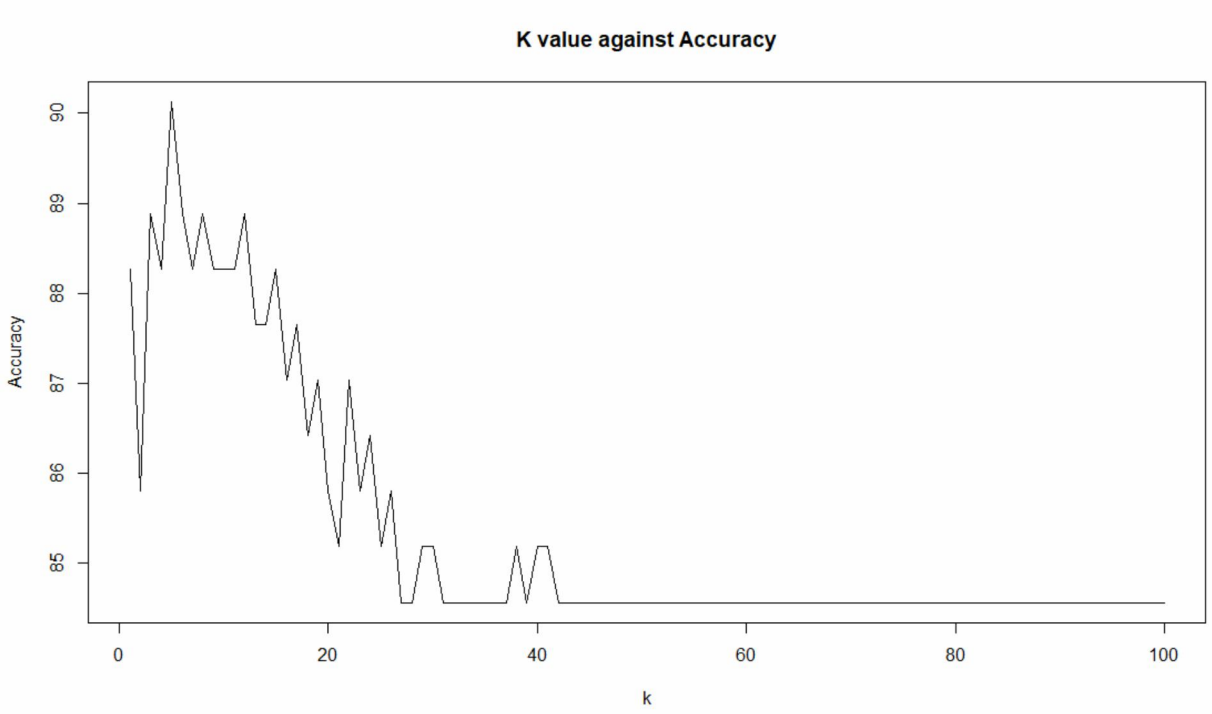
```
> confusionMatrix(cm)

Confusion Matrix and Statistics


knn_model  0  1
        0  7  1
        1 18 136

               Accuracy : 0.8827
                 95% CI : (0.8229, 0.9279)
    No Information Rate : 0.8457
    P-Value [Acc > NIR] : 0.1132671

                  Kappa : 0.3777

 Mcnemar's Test P-Value : 0.0002419

            Sensitivity : 0.28000
```

Specificity : 0.99270
Pos Pred Value : 0.87500
Neg Pred Value : 0.88312
Prevalence : 0.15432
Detection Rate : 0.04321
Detection Prevalence : 0.04938
Balanced Accuracy : 0.63635

'Positive' Class : 0

The output above shows the confusion matrix of our predictions and some evaluations of the confusion matrix when the value of k is 7. From the output above, we can observe that out of the 162 students, the model classified 88.27% of them correctly. The sensitivity of this model is 0.28 and this means that the model has a 28% probability to classify a student who failed in Portuguese correctly. The specificity of the model is 0.9927. This shows that 99.27% of the students that passed are identified correctly. In short, this model is also suitable in classifying the students that passed the Portuguese subject accurately.

In order to improve the model, we plot a graph of Accuracy vs k value again and perform looping for k ∈ {1: 100}



From the output above, we can deduce that the optimum value of k is 5 since it has the highest accuracy which is approximately 90.12%. The R commands and outputs below verify this.

```
> max(k_accuracy)
[1] 90.12346

> which.max(k_accuracy)
[1] 5
```

From the graph, we can see that the line reached a peak when k equals to 4 and started to decrease gradually when the value of k increases. Therefore, we will choose 5 as our optimum k value since it has the highest accuracy. Diagram below shows the confusion matrix for the 5-NN model.

```
> confusionMatrix(cm)

knn_model  0  1
        0 11  2
        1 14 135

               Accuracy : 0.9012
                 95% CI : (0.8446, 0.9425)
    No Information Rate : 0.8457
    P-Value [Acc > NIR] : 0.02725

                  Kappa : 0.5292

 Mcnemar's Test P-Value : 0.00596

            Sensitivity : 0.44000
            Specificity : 0.98540
         Pos Pred Value : 0.84615
         Neg Pred Value : 0.90604
             Prevalence : 0.15432
         Detection Rate : 0.06790
   Detection Prevalence : 0.08025
      Balanced Accuracy : 0.71270

       'Positive' Class : 0
```

Note that sensitivity increased significantly to 44% compared to accuracy measures corresponding to k=7. The 5-NN model has 44% probability to classify a student who failed in Portuguese correctly.

In short, we use 5-NN model as our candidate model for Portuguese.

3.2.3 Weakness and Strength of KNN

3.2.3.1 Strength

KNN algorithm is very easy to be implemented as we only need to determine the k value and calculate the distance between two objects. Furthermore, the KNN algorithm is easy to understand.

3.2.3.2 Weakness

Firstly, k-nearest neighbor algorithm does not allow categorical features, because we are not able to calculate the arbitrary distance between two

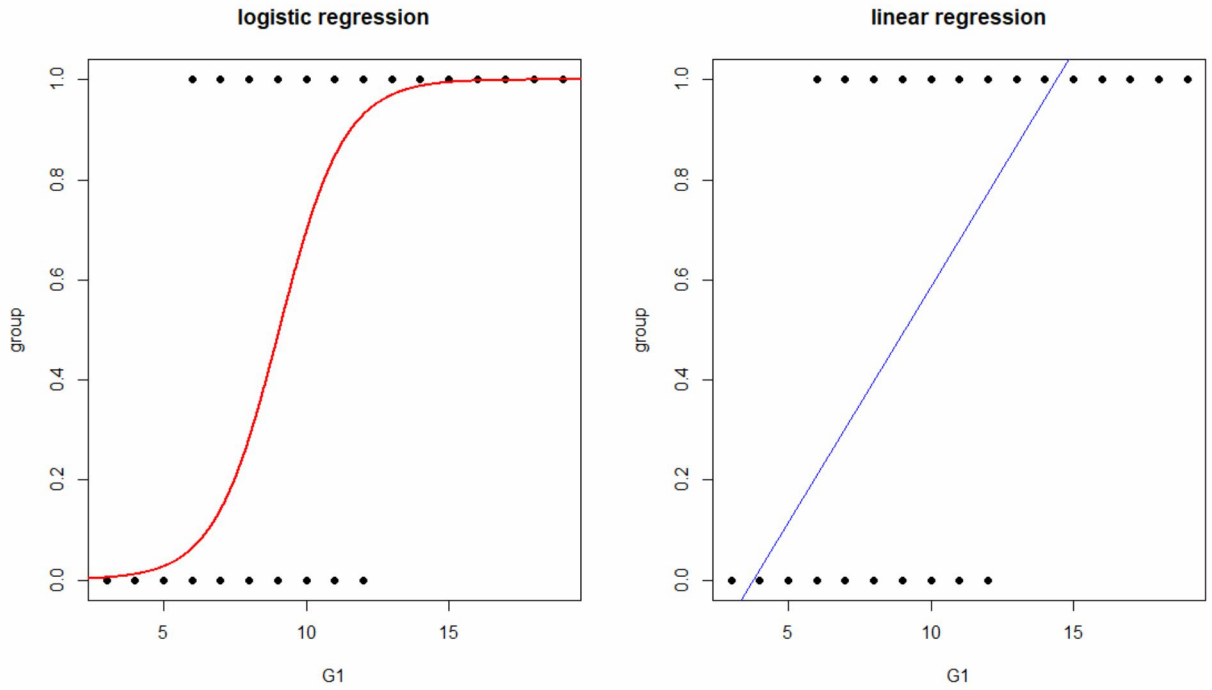
categories. For example, we cannot calculate the distance between colors. Therefore, before applying the KNN algorithm, we have to convert the categorical columns into multiple binary vectors.

Besides, we need to find the best k value before using KNN algorithm. This can increase the computational cost as we need to perform the classification multiple times.

3.3 Logistic Regression

3.3.1 Linear regression vs logistic regression

Why Logistic Regression?



First, we consider the linear regression line. We can extrapolate the regression line to predict the value of group related to G3 which can give us either a negative value or a value more than 1. This does not make sense because we classified the group of a student as either failed or passed.

The logistic regression line gives us a categorical prediction instead of continuous prediction. Therefore, in this case we will choose logistic regression since we want to classify the students as either passed or failed.

3.3.2 Multiple logistic regression

3.3.2.1 For Math

```
> summary(logistic)

Call:
glm(formula = group ~ . - G3 - G2 - G1, family = binomial, data = d1)


Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     4.90464    2.68263   1.828 0.067505 .
schoolMS        0.24985    0.46348   0.539 0.589841
sexM            0.47319    0.31331   1.510 0.130967
age            -0.26491    0.13449  -1.970 0.048880 *
addressU        0.19695    0.34019   0.579 0.562623
famsizeLE3      0.18573    0.29795   0.623 0.533039
PstatusT       -0.43621    0.44541  -0.979 0.327404
Medu            0.09321    0.19693   0.473 0.636003
Fedu            0.08193    0.16847   0.486 0.626734
Mjobhealth      0.36782    0.69504   0.529 0.596662
Mjobother      -0.43924    0.41006  -1.071 0.284096
Mjobservices    0.29079    0.46467   0.626 0.531446
Mjobteacher    -0.92403    0.60371  -1.531 0.125875
Fjobhealth     -0.17389    0.82047  -0.212 0.832159
Fjobother       0.19057    0.57762   0.330 0.741461
Fjobservices   -0.11946    0.59490  -0.201 0.840850
Fjobteacher     0.67140    0.77139   0.870 0.384098
reasonhome      0.34278    0.32805   1.045 0.296068
reasonother     0.54395    0.50197   1.084 0.278530
reasonreputation 0.49858   0.34888   1.429 0.152973
guardianmother -0.09562    0.33631  -0.284 0.776156
guardianother   0.06715    0.60618   0.111 0.911793
traveltime      0.03598    0.19731   0.182 0.855309
studytime       0.24113    0.17697   1.363 0.173031
failures       -0.89409    0.20856  -4.287 1.81e-05 ***
schoolsupyes   -0.89924    0.37943  -2.370 0.017790 *
famsupyes      -0.60713    0.29322  -2.071 0.038396 *
paidyes         0.22566    0.29066   0.776 0.437514
activitiesyes  -0.14465    0.26952  -0.537 0.591495
nurseryyes     -0.38310    0.34186  -1.121 0.262437
higheryes       0.67864    0.62895   1.079 0.280587
internetyes     0.36476    0.35384   1.031 0.302604
romanticyes    -0.30925    0.28004  -1.104 0.269462
famrel          0.18149    0.14967   1.213 0.225290
```

```
freetime      0.14395   0.14364   1.002 0.316272
goout        -0.53424   0.14304  -3.735 0.000188 ***
Dalc         -0.10266   0.20114  -0.510 0.609779
Walc          0.24785   0.15270   1.623 0.104555
health       -0.12755   0.09880  -1.291 0.196712
absences     -0.01388   0.01613  -0.860 0.389551
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As shown in the table above, we have taken all variables in data set "student-mat" into consideration, except G1 and G2 (as they are strongly correlated). At significance level of 5%, the following variables have significant relationship with the group (pass/fail) of a student.

Age, failures, family support, school support and free time.

This means that only these variables have direct effect on student final grade for Math.

The logistic model at 5% significant level, in this case is defined as

$$P(Y=1|\mathbf{X}) = \frac{1}{1+\exp exp\,(\alpha)}$$

Where $\alpha = -\,[(4.9046 - 0.2649 * age - 0.8941 * failures - 0.8992 * schoolsup[yes] - 0.6071 * [famsupyes] - 0.5342 * goout)]$

This logistic model is also another candidate model for Math.

The confusion matrix of this logistic model is shown below.

```
> cm2

            Predicted_value
Actual_value FALSE TRUE
           0    13   19
           1     3   63

> (13+63)/(13+19+3+63)

[1] 0.7755102
```

The accuracy of the model is 0.7755, which implies that the model has 77.55% of chance to predict whether student will pass or fail the Math subject correctly.

3.2.2.2 For Portuguese

```
> summary(logistic)

Call:
glm(formula = group ~ . - G3 - G2 - G1, family = binomial, data = d2)


Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -0.59198    2.53384  -0.234 0.815273
schoolMS        -2.00545    0.34865  -5.752 8.81e-09 ***
sexM            -0.40170    0.33435  -1.201 0.229578
age              0.22207    0.13684   1.623 0.104616
addressU         0.20457    0.31342   0.653 0.513945
famsizeLE3       0.36658    0.33199   1.104 0.269521
PstatusT         0.22030    0.43796   0.503 0.614954
Medu            -0.02666    0.18291  -0.146 0.884112
Fedu             0.22985    0.18379   1.251 0.211074
Mjobhealth      -0.20383    0.67968  -0.300 0.764256
Mjobother       -0.07702    0.34996  -0.220 0.825819
Mjobservices     0.16293    0.46251   0.352 0.724636
Mjobteacher      0.58988    0.74583   0.791 0.429000
Fjobhealth      -1.77394    0.99081  -1.790 0.073392 .
Fjobother       -0.63136    0.57605  -1.096 0.273074
Fjobservices    -1.13159    0.60728  -1.863 0.062410 .
Fjobteacher     -1.55474    0.97127  -1.601 0.109437
reasonhome       0.31000    0.38640   0.802 0.422395
reasonother      0.11391    0.40562   0.281 0.778830
reasonreputation 0.42150    0.45092   0.935 0.349920
guardianmother  -0.77525    0.36379  -2.131 0.033087 *
guardianother   -0.09165    0.69762  -0.131 0.895474
traveltime       0.27169    0.20379   1.333 0.182476
studytime        0.16928    0.19293   0.877 0.380272
failures        -1.05535    0.20616  -5.119 3.07e-07 ***
schoolsupyes    -0.55017    0.49980  -1.101 0.270994
famsupyes        0.11748    0.29027   0.405 0.685678
paidyes         -0.80342    0.50304  -1.597 0.110241
activitiesyes    0.40074    0.29296   1.368 0.171346
nurseryyes      -0.48082    0.34900  -1.378 0.168293
higheryes        1.43259    0.36835   3.889 0.000101 ***
internetyes     -0.14855    0.33141  -0.448 0.653985
romanticyes     -0.28217    0.28472  -0.991 0.321670
famrel           0.07614    0.13132   0.580 0.562072
freetime        -0.09230    0.13916  -0.663 0.507169
```

```
goout        -0.02382    0.13823  -0.172 0.863200
Dalc         -0.02839    0.17541  -0.162 0.871420
Walc         -0.13031    0.14941  -0.872 0.383124
health       -0.04633    0.10281  -0.451 0.652226
absences      -0.08476    0.02965  -2.859 0.004254 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As shown in the table above, we have taken all variables in this model into consideration, except G1 and G2 (as they are strongly correlated). At significance level of 5%, the following variables have significant relationship with the group (pass/fail) of a student.

School, guardian, failures and thoughts of wanting to pursue for higher education

This means that only these variables have direct effect on student final grade for Portuguese.

The logistic model at 5% significant level, in this case is defined as

$$P(Y=1|\mathbf{X}) = \frac{1}{1+\exp exp\,(\beta)}$$

where $\beta = -\,[(-\,0.5920 - 2.0055 * school[MS] - 0.7753 * guardian[Mother] - 1.0554 * failures + 1.4326 * higher[Yes] - 0.08476 * absences)]$

This logistic model is also another candidate model for Portuguese.

The confusion matrix of this logistic model is shown below.

```
> cmm2
       failed passed
failed     12     13
passed      6    131

> (12+131)/(12+13+6+131)
[1] 0.882716
```

The accuracy of the model is 0.8827, which implies that the model has 88.37% of chance to predict whether student will pass or fail the Portuguese subject correctly.

3.3.3 Strength and weakness of logistic regression

Logistic regression calculates the probability of a given outcome and predicts the outcome based on the probability.

3.3.3.1 Strength

By using logistic regression, we can get the probability of an outcome along with the final outcome. Logistic regression is useful in predicting categorical outcomes but not continuous outcomes since the output lies between 0 and 1. For example, we cannot use logistic regression to predict the salary of a person based on his age. Furthermore, logistic regression is easy to implement and very efficient to train.

3.3.3.2: Weakness

We cannot use logistic regression to solve nonlinear problems.

3.4 Decision Tree/ Tree Pruning

3.4.1 Decision Tree

3.4.1.1 Math

Firstly, training data is used to build a decision tree model.

The top 5 rules that significantly affected the student performance for Math will be determined based on Figure 1.

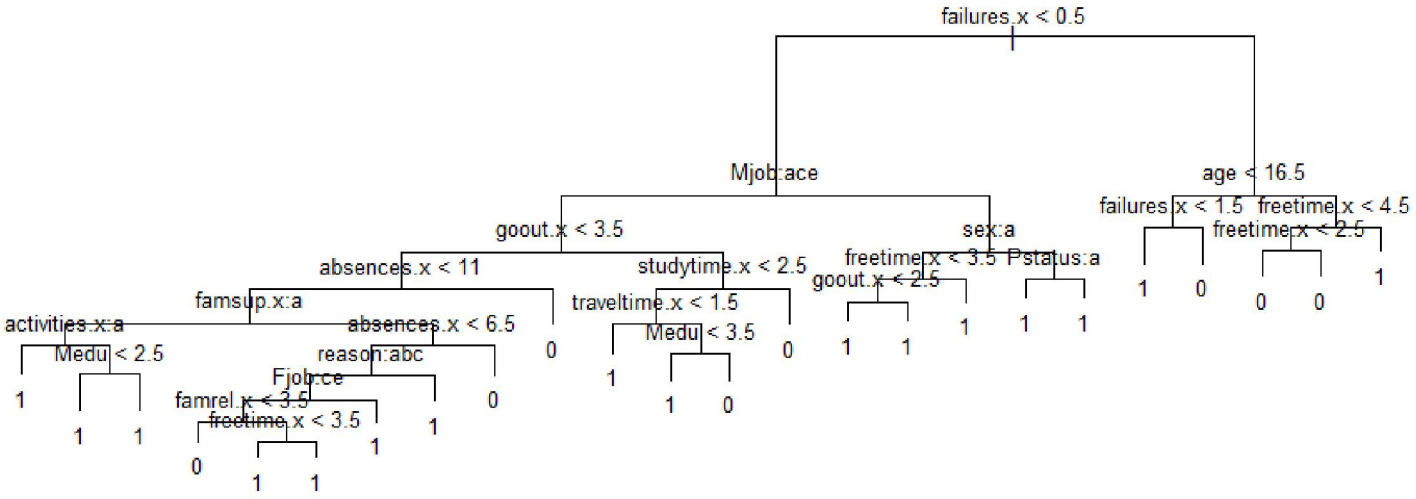


Figure 1: Decision Tree for Math Course

The extract of the output below helps to explain all unknowns found in Figure 1.

```
> tree.d1
 4) Mjob: at_home,other,teacher 142 178.900 1 ( 0.32394 0.67606 )
32) famsup.x: no 38  25.570 1 ( 0.10526 0.89474 )
64) activities.x: no 20   0.000 1 ( 0.00000 1.00000 ) *
65) activities.x: yes 18  19.070 1 ( 0.22222 0.77778 )
33) famsup.x: yes 53  63.150 1 ( 0.28302 0.71698 )
```

132) reason: course,home,other 36 42.540 1 (0.27778 0.72222)
264) Fjob: other,teacher 23 30.790 1 (0.39130 0.60870)
265) Fjob: at_home,services 13 7.051 1 (0.07692 0.92308) *
133) reason: reputation 10 0.000 1 (0.00000 1.00000) *

5) Mjob: health,services 77 55.540 1 (0.11688 0.88312)
10) sex: F 41 40.470 1 (0.19512 0.80488)
11) sex: M 36 9.139 1 (0.02778 0.97222)
22) Pstatus: A 5 5.004 1 (0.20000 0.80000) *
23) Pstatus: T 31 0.000 1 (0.00000 1.00000) *

Output 1

Example of interpretation:

In Output:

4) Mjob: at_home,other,teacher 142 178.900 1 (0.32394 0.67606)

In Figure 1:

Mjob: ace means Mother's job ∈ {a = at home, c = other, e = teacher}

Based on the decision tree, it can be deduced that "failure" is the first important factor that will decide whether the student is going to pass or fail the Math subject. After "failure", there are 2 significant variables that will decide whether student can pass or fail Math, i.e. "Mother's job" or "age of less than 17 (in plain language: age of 17, in term of Math: age = 16.5)".

The top 5 decision rules that can be extracted from the decision tree in Figure 1 and all outputs in Output 1 are as follows:

1. ***If failures > 0.5 ∧ age < 16.5 ∧ failures.x < 1.5, then fail.***
2. ***If failures > 0.5 ∧ age > 16.5 ∧ freetime > 4.5, then pass***
3. ***If failures < 0.5 ∧ Mjob ∈ {health, services} ∧ sex = M ∧ Pstatus = T, then pass***
4. ***If failures < 0.5 ∧ Mjob ∈ {at_home, other, teacher} ∧ goout > 3.4 ∧ studytime > 2.5, then fail***
5. ***If failures < 0.5 ∧ Mjob ∈ {at_home, other, teacher} ∧ goout < 3.4 ∧ absence > 11, then fail***

The R software helps to generate the confusion matrix based on the predicted output and actual result of the testing data set.

The output of confusion matrix and accuracy are as follows:

```
> table(tree.d1pred,ndatatest$score.x)

tree.d1pred  0  1
         0 18 14
         1 20 63

> (18+63)/(18+14+20+63)
[1] 0.7043478
```

		Actual	
		0 (Failed)	**1 (Passed)**
Predicted	**0 (Failed)**	18	14
	1 (Passed)	20	63

$$Accuracy_{Math} = \frac{(18+63)}{(18+14+20+63)} = 0.7043$$

This accuracy tells us that out of 115 testing data, there is a 70.43% chance that this decision tree will correctly predict whether student will pass/fail the Math subject given a set of x's.

We include this decision tree model into the candidate model for Math

3.4.1.2 Portuguese

Using the same algorithm in 3.4.1,

The top 5 rules that significantly affected the student performance for Math will be determined first based on Figure 2.

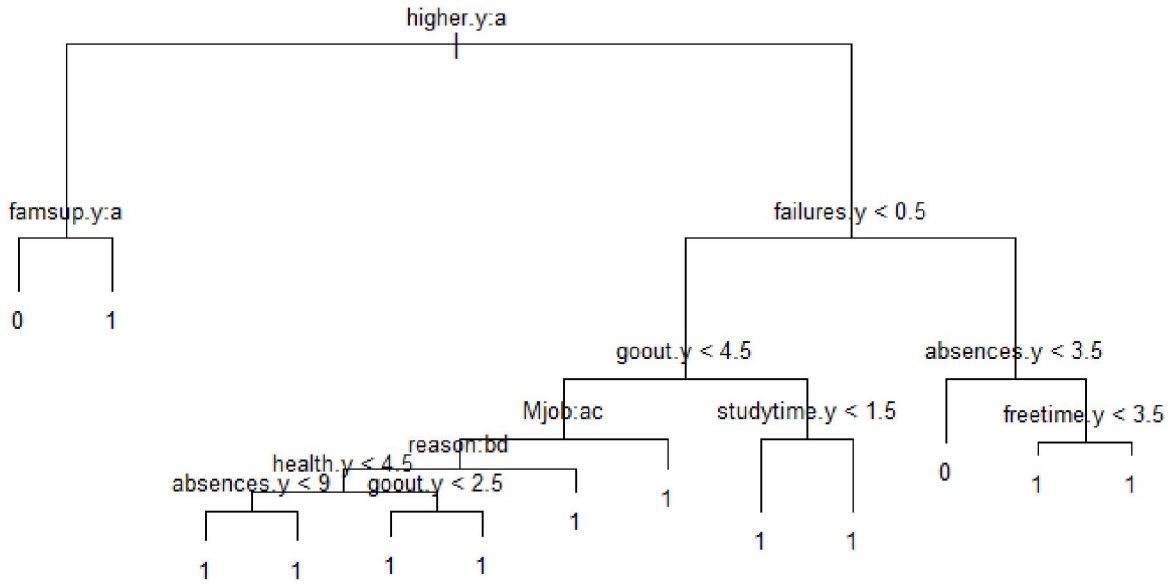


Figure 2: Decision Tree for Portuguese Language Course

The extract of the output below helps to explain all unknowns found in Figure 2.

```
> tree.d2
 2) higher.y: no 12  15.280 0 ( 0.66667 0.33333 )
 4) famsup.y: no 6   0.000 0 ( 1.00000 0.00000 ) *
 5) famsup.y: yes 6   7.638 1 ( 0.33333 0.66667 ) *
 3) higher.y: yes 255 108.500 1 ( 0.05490 0.94510 )
24) Mjob: at_home,other 100  26.950 1 ( 0.03000 0.97000 )
48) reason: home,reputation 58  23.510 1 ( 0.05172 0.94828 )
49) reason: course,other 42   0.000 1 ( 0.00000 1.00000 ) *
25) Mjob: health,services,teacher 108   0.000 1 ( 0.00000 1.00000 ) *
```

Output 2

Example of interpretation:

In Output:

2) higher.x: no 12 15.280 0 (0.66667 0.33333)

In Figure 2:

higher.x: a means higher.x == “no”

Based on the decision tree, it can be deduced that “higher.y” (students want to pursue for higher education) is the first important factor that will decide whether the student is going to pass or fail the Portuguese subject, then followed by “famsup.y” and “failures.y”. This implies that family supports and failures played an important role in determining the final result of the students taking Portuguese language.

The top 5 decision rules that can be extracted from the decision tree in Figure 2 and all outputs in Output 2 are as follows:

1. ***If higher = no ∧ famsup = no, then fail***
2. ***If higher = yes ∧ failures > 0.5 ∧ absence < 3.4, then fail***
3. ***If higher = yes ∧ failures < 0.5 ∧ goout > 4.5, then pass***
4. ***If higher = yes ∧ failures < 0.5 ∧ goout < 4.5 ∧ Mjob ∈ {health, service, teacher}, then pass***
5. ***If higher = yes ∧ failures < 0.5 ∧ goout < 4.5 ∧ Mjob ∈ {at_home, other}, then pass***

The output of confusion matrix and accuracy are as follows:

```
> table(tree.d2pred,ndatatest$score.y)
tree.d2pred  0  1
          0  2  5
          1  8 100
```

```
> (2+100)/(2+5+8+100)
[1] 0.8869565
```

		Actual	
		0 (Failed)	**1 (Passed)**
Predicted	**0 (Failed)**	2	5
	1 (Passed)	8	100

$$Accuracy_{Portuguese\ Language} = \frac{(2+100)}{(2+5+100+8)} = 0.8870$$

This accuracy tells us that out of 115 testing data, there is a 88.70% chance that this decision tree will correctly predict whether student will pass/fail the Portuguese language given a set of *x*'s.

We include this decision tree model into the candidate model for Portuguese.

3.4.2 Weaknesses and Strengths

3.4.2.1 Weakness

Decision trees are less appropriate for estimation tasks where the goal is to predict the value of a continuous attribute.

Decision trees are prone to errors in classifications problems with many class and relatively small number of training examples.

3.4.2.2 Strengths

Decision trees are able to generate understandable rules.

Decision trees perform classification without requiring much computation.

Decision trees provide a clear indication of which fields are most important for prediction or classification.

3.4.3 Tree Pruning

3.4.3.1 Math

Tree Pruning model is built using same training data set from 3.4. It is the improvised version of decision tree.

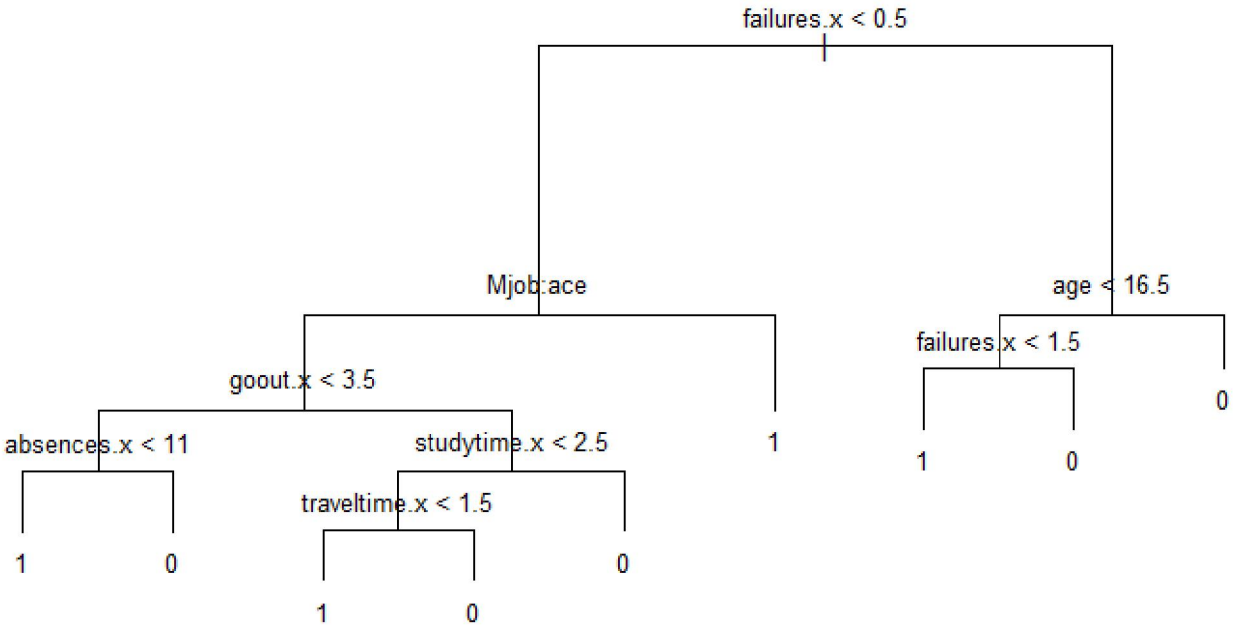


Figure 3: Tree Pruning Plot for Math Course

Use the same extract of output for tree.d1 to explain the unknowns in the Figure 3.

> tree.d1

4) Mjob: at_home,other,teacher 142 178.900 1 (0.32394 0.67606)

5) Mjob: health,services 77 55.540 1 (0.11688 0.88312)

Example of interpretation:

In Output:

4) Mjob: at_home,other,teacher 142 178.900 1 (0.32394 0.67606)

In Figure 1:

Mjob: ace means Mother's job ∈ {a = at home, c = other, e = teacher}

Based on Figure 3, it can be deduced that "failure" is still the first important factor that will decide whether the student is going to pass or fail the Math subject, whereas "Mother's job" or "age of less than 17 (in plain language: age of 17, in term of Math: age = 16.5)" are still the second most important factors as stated in *3.4.1.1: Math*.

The top 5 decision rules that can be extracted from Figure 3 are as follows:

1. ***If failures > 0.5 ∧ age > 16.5, then fail.***
2. ***If 0.5 < failures < 1.5 ∧ age < 16.5, then pass***
3. ***If failures > 1.5 ∧ age < 16.5, then fail***
4. ***If failures < 0.5 ∧ Mjob ∈ {at home, other, teacher} ∧ goout < 3.4 ∧ absences < 11, then pass***
5. ***If failures < 0.5 ∧ Mjob ∈ {at home, other, teacher} ∧ goout < 3.4 ∧ absences > 11, then fail***

The output of confusion matrix and accuracy are as follows:

```
> table(tree.d1pred,ndatatest$score.x)
tree.d1pred  0  1
          0 17 11
          1 21 66
> (17+66)/(17+11+21+66)
[1] 0.7217391
```

		Actual	
		0 (Failed)	**1 (Passed)**
Predicted	**0 (Failed)**	17	11
	1 (Passed)	21	66

$$Accuracy_{Math} = \frac{(17+11)}{(17+11+21+66)} = 0.7217$$

This accuracy tells us that out of 115 testing data, there is now a 72.17% chance that this decision tree will correctly predict whether student will pass/fail the Math subject given a set of *x*'s.

The accuracy of tree pruning model (72.17%) is slightly higher than the accuracy of decision tree model. (70.43%)

We include this tree pruning model into the candidate model for Math.

3.4.3.2 Portuguese

With the similar situation in 3.4.1.1, we determine the top 5 rules based on Figure 4.

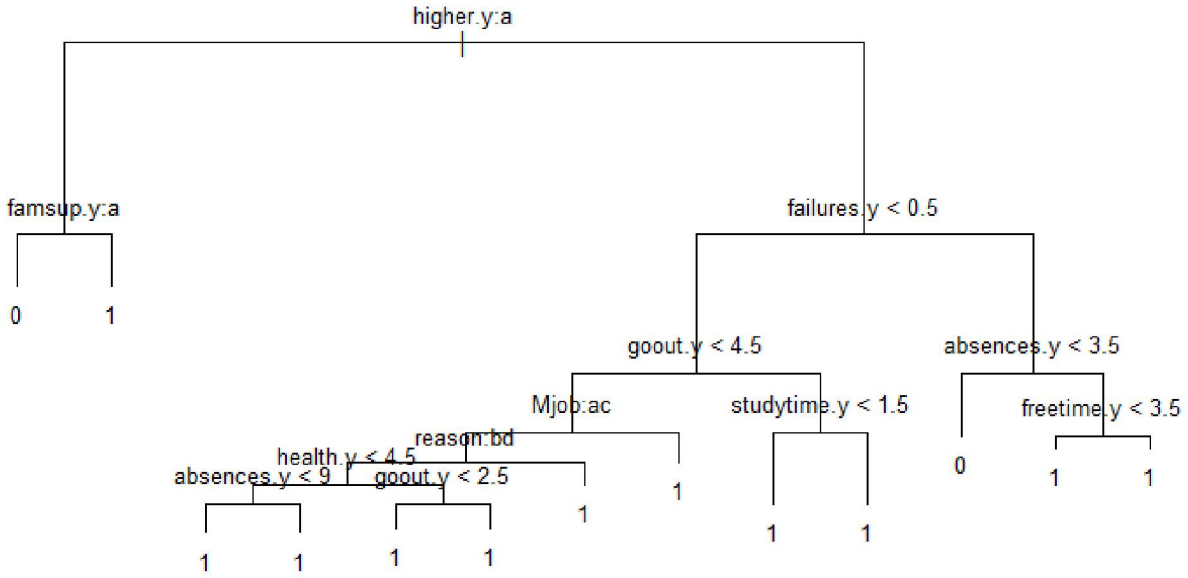


Figure 4: Tree Pruning Plot for Portuguese Language

Use the same extract of output for tree.d2 to explain the unknowns in the Figure 4.

```
> tree.d2
 2) higher.y: no 12  15.280 0 ( 0.66667 0.33333 )
 4) famsup.y: no 6   0.000 0 ( 1.00000 0.00000 ) *
 5) famsup.y: yes 6   7.638 1 ( 0.33333 0.66667 ) *
 3) higher.y: yes 255 108.500 1 ( 0.05490 0.94510 )
24) Mjob: at_home,other 100  26.950 1 ( 0.03000 0.97000 )
48) reason: home,reputation 58  23.510 1 ( 0.05172 0.94828 )
49) reason: course,other 42   0.000 1 ( 0.00000 1.00000 ) *
25) Mjob: health,services,teacher 108   0.000 1 ( 0.00000 1.00000 ) *
```

Example of interpretation:

In Output:

2) higher.x: no 12 15.280 0 (0.66667 0.33333)

In Figure 2:

higher.x: a means higher.x == "no"

Based on Figure 4, it can be deduced that "higher.y" (students want to pursue for higher education) is still the first important factor that will decide whether the student is going to pass or fail the Math subject, whereas "famsup.y" and "failures.y" are still the second most important factors as stated in *3.4.1.2: Portuguese*.

The top 5 decision rules that can be extracted from Figure 4 are as follows:

1. ***If higher = no ∧ famsup = no, then fail***
2. ***If higher = no ∧ famsup = yes, then pass***
3. ***If higher = yes ∧ failures < 0.5 ∧ absences < 3.4 ∧ freetime < 3.4, then pass***

4. ***If higher = yes ∧ failures < 0.5 ∧ absences < 3.4 ∧ freetime > 3.4, then pass***
5. ***If higher = yes ∧ failures < 0.5 ∧ absences > 3.4, then fail***

The output of confusion matrix and accuracy are as follows:

```
> table(tree.d2pred,ndatatest$score.y)

tree.d2pred  0  1
           0  2  5
           1  8 100
```

		Actual	
		0 (Failed)	**1 (Passed)**
Predicted	**0 (Failed)**	2	5
	1 (Passed)	8	100

$$Accuracy_{Portuguese\ Language} = \frac{(2+100)}{(2+5+100+8)} = 0.8870$$

It seems that the accuracy of this model and its interpretation is still the same as in *3.4.1.2 Portuguese,* i.e. 88.70% of predictions will be correct.

We include this tree pruning model into the candidate model for Portuguese.

3.4.4 Weaknesses and Strengths

3.4.4.1 Weakness

As an improvised approach for Decision Tree, the weakness for this approach are similar to Decision Tree

3.4.4.2 Strengths

As an improvised approach for Decision Tree, the strengths for this approach are similar to Decision Tree.

3.5 Bootstrap Aggregating (Bagging)

3.5.1 For Math Students

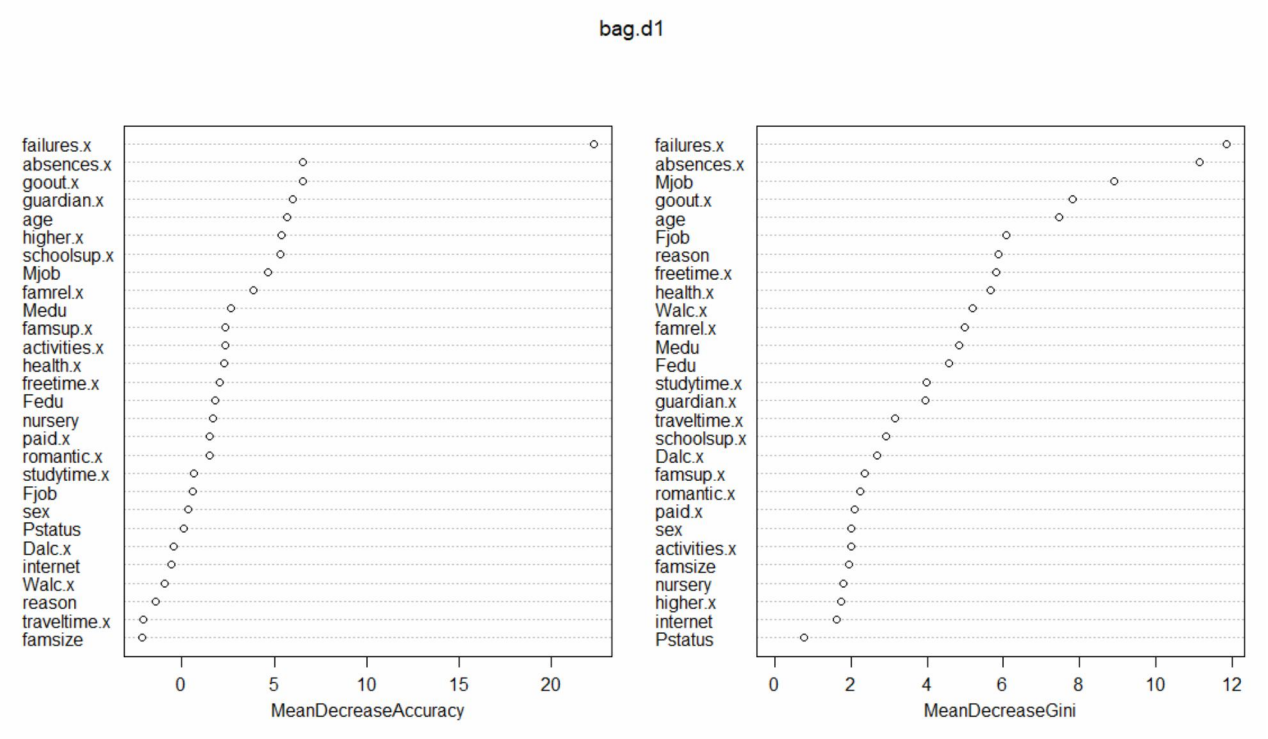


Figure 9.1

According to the output above, it is obvious that students' performance depends most on the past failure rate of Mathematics. Secondly, frequency of absences in Mathematics class decides the students' performance in this subject, followed by the frequency of going out with friends. From the result shown, we can also see that the guardian plays an important role in affecting student's academic performance in this subject. The result also indicates that the academic performance depends on the age of student.

Confusion Matrix for Math:

```
> confusionMatrix(cm1)
Confusion Matrix and Statistics
prediction1  0  1
         0 14  7
         1 18 57

          Accuracy : 0.7396
            95% CI : (0.64, 0.8238)
   No Information Rate : 0.6667
   P-Value [Acc > NIR] : 0.07762

            Kappa : 0.359

 Mcnemar's Test P-Value : 0.04550
```

Sensitivity : 0.4375
Specificity : 0.8906
Pos Pred Value : 0.6667
Neg Pred Value : 0.7600
Prevalence : 0.3333
Detection Rate : 0.1458
Detection Prevalence : 0.2188
Balanced Accuracy : 0.6641

'Positive' Class : 0

Based the output above, the accuracy of the model bootstraps aggregating is 73.96%. This shows that the model has 73.96% chance to predict the output correctly with the given input data. With a sensitivity of 0.4375, it means that there is a 43.75% chance of predicting a student that will fail in Math correctly by the model, t. The specificity of the model is 0.8906 means that 89.06% of the students that passed Math are predicted correctly. The model is better in predicting the students passed precisely.

We include this bagging model into the candidate model for Math.

3.5.2 For Portuguese Students

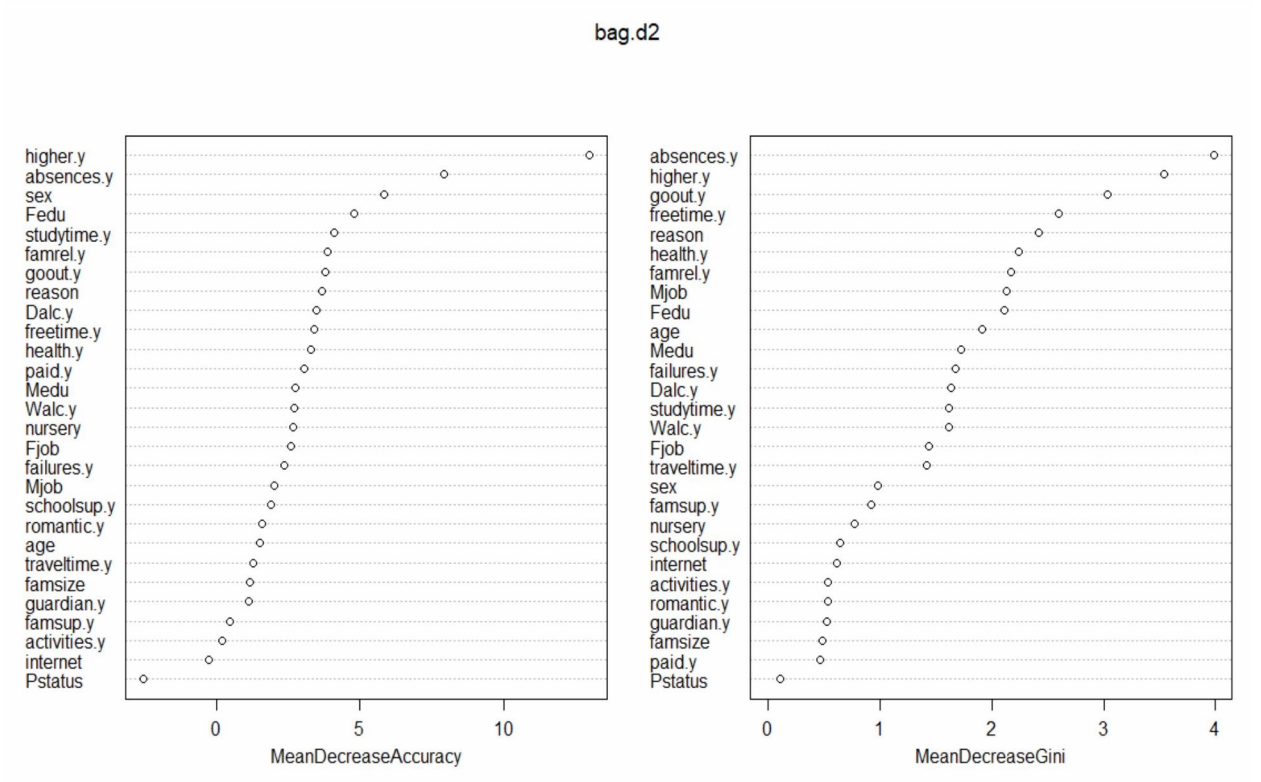


Figure 9.2

Based on the output above, the desires of pursuing higher education in Portuguese affects the students' performance in this subject the most. Whether students attend the class for this subject is the next important factor that decides their academic performance. From the result given, sex plays an important role in the performance in this subject. It seems like the level of education of the students' fathers exerts

certain influence on their performance in Portuguese, followed by the time they spent in studying for this subject weekly.

Confusion Matrix for Portuguese:

```
> confusionMatrix(cm2)
Confusion Matrix and Statistics
prediction2  0  1
         0  2  0
         1  6 88

               Accuracy : 0.9375
                 95% CI : (0.8689, 0.9767)
    No Information Rate : 0.9167
    P-Value [Acc > NIR] : 0.30271

                  Kappa : 0.3793

 Mcnemar's Test P-Value : 0.04123

            Sensitivity : 0.25000
            Specificity : 1.00000
         Pos Pred Value : 1.00000
         Neg Pred Value : 0.93617
             Prevalence : 0.08333
         Detection Rate : 0.02083
   Detection Prevalence : 0.02083
      Balanced Accuracy : 0.62500

       'Positive' Class : 0
```

Based the output above, the accuracy of the model bootstraps aggregating is 93.75%. This shows that the model has almost a perfect chance to predict all the output correctly with the given input data. With a sensitivity of 0.25, it means that for student that is classified as failed in Math by the model, there is only 25% chance of getting it correct. The specificity of the model with 1 means that as long as there is a prediction about the students will pass this subject, the students will surely pass as a result. The model is best in predicting the students passed precisely.

We include this bagging model into the candidate model for Portuguese.

3.5.3 Weakness & Strength

3.5.3.1 Weakness

Random Forest weaknesses are that when used for regression they cannot predict beyond the range in the training data, and that they may over-fit data sets that are particularly noisy. Random Forest is more suitable in classification rather than regression. On the other hand, they also lack some interpretability, which sometimes it is hard to interpret the factors that are given for the predicted output.

3.5.3.2 Strengths

The advantage of the random forest is that if there are a lot of data and lots of predictor variables, random forest is the best solution. They can deal with messy, real data. If there are lots of extraneous predictors, it has no problem. It automatically does a good job of finding interactions as well. There are no assumptions that the response has a linear (or even smooth) relationship with the predictors. Random forest runtimes are quite fast, and they are able to deal with unbalanced and missing data.

CHAPTER 4 CONCLUSION

4.1 Review of all approaches we tried

Education is a crucial element in our society. In this study, we have addressed the prediction of secondary student grades of two core classes, Mathematics and Portuguese courses by using demographic, family-related data. Dataset was collected from two public secondary schools, Galriel Pereora and Mousiho da Silveira. In layman terms, a total of 33 predictors were collected including gender, age, health status, absence, home address, student guardian, parents' marital status, mother's educational background, father's educational background, mother's job, father's job, home internet, family size, reason for choosing the school, home to school travel time, study time, free time after school, past failures, school and family education support, curricular activities, extra paid class, nursery school, higher education intention, romantic relationship, going out with friends, weekday and weekend alcohol consumptions, quality of family relationship, student Mathematics grade and Portuguese grade.

The K-Nearest Neighbors model is able to predict approximately 88.78% of students' Mathematics grades. This means this 13-NN model has 65.62% chances to classify a student who failed in Mathematics course correctly. The best k value which we used is 13 as when k equals to 13 it reached the highest peak in the graph. Age, failures, absences, go out with friends, guardian, school educational support, family educational support, mother's occupation and free time were found to be significant predictors of student's Mathematics performance. The Bootstrap Aggregating (Bagging) model able to predict approximately 93.75% of students' Portuguese grades and there is only 25% chance of classified as failed in Portuguese course. Sex, father's education, absences, failures, school, family education support and higher education intention, were found to be significant predictors of student's Portuguese performance.

4.2 Best fit model

Table A: Summary of the Accuracy of all candidate models for Mathematics course

Approaches	Accuracy of Predicted value
13-Nearest Neighbor	88.78%
Logistic Regression	77.55%
Decision Tree	70.43%
Tree Pruning	72.17%
Bagging	73.96%

Table B: Summary of the Accuracy of all candidate models for Portuguese course

Approaches	Accuracy of Predicted value
5-Nearest Neighbor	88.78%
Logistic Regression	88.37%
Decision Tree	88.7%
Tree Pruning	88.7%
Bagging	93.75%

Based on Table A, we choose 13-Nearest Neighbor as our model to estimate the predicted final result for students taking Math. The Accuracy of 13-NN is the highest, i.e. 88.78% among other models (in Table A).

Based on Table B, we choose Bootstrap Aggregating (Bagging) as our model to estimate the predicted final result for students taking Portuguese. The Accuracy of Bagging the highest, i.e. 93.75% among other models (in Table B).

4.3 Overall Strengths and Weaknesses of this report

A few models we used for this study including the basic descriptive analysis, correlation analysis, Principal Component Analysis (PCA), linear regression model, K-Nearest Neighbors (KNN) model, logistic regression model, decision tree and Bootstrap Aggregating (Bagging) model. Correlation analysis can perform a very detailed analysis for further research but they online test for linear relationship between two predictors. PCA testing algorithms is easy but they required to set by manual sometimes and some cannot be interpretable. Linear regression model is suggested for supervised learning process, yet over simplifies many real-world application problems. KNN is easy to implement by calculating the k value, but its algorithm does not allow categorical features. Logistic regression model can get the probability of an outcome along with the final outcome easily, yet it does not solve for nonlinear problems. Decision tree are prone to errors in classifications problems but this model allows to generate understandable rules. Bagging model usually cannot predict beyond the range in the training data. This model can deal with messy, real data by automatically finding interactions.

4.4 Significant factors that affects the result generally

This research identified the significant factors influencing Mathematics and Portuguese performance in secondary school students. Suggestions on how to improve the impact of these factors have been provided in this study. Student academic performance can be improved in combination with the efforts of the school, the family and the students themselves. Instead of finding direct correlations between predictors and grades, we found that the clustering of predictors produced a more interesting and productive analysis. While different models provided different sets of variables, the groupings made sense in the context of the data set. During our analysis we found that it was hard to predict final grades accurately when we omitted previous grades regardless of the type of model we were building. In particular,

while we originally thought that urban students appeared to outperform rural students, we found no evidence that the student's address is a major predictor in their grade. Likewise, we found no evidence that travel time, study time and intention to pursue higher education were significant in predicting a grade for a student. Nevertheless, we have been able to confirm that the role of parents is meaningful in determining student performance.

4.5 Future Work recommended to improvise the data sets

In future research, the dataset can be collected from different types of secondary schools, for examples, private schools on multiple locations mixing together such as urban and suburban area. The questionnaire can be configured for a more detailed and subject-oriented approach.

4.6 Recommendations for Education Officials

In addition to the effort to be made by school and family to enhance the academic performance of students, governments should be active in enhancing the ambition of students to pursue higher education. The strategy should concentrate on reducing regional inequalities and creating more job opportunities for rural and small-town students, or even suburban area secondary school students. Throughout small towns and rural areas, further colleges may be opened. This will provide more incentives for local students to take part in higher education, as most graduates would be working in banking, legal and other professional services or businesses.

REFERENCES

[1] P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.

[2] Hornik, K. (2020, February 20). Frequently Asked Questions on R. Retrieved March 26, 2020, from https://CRAN.R-project.org/doc/FAQ/R-FAQ.html

[3] Normality Testing - Skewness and Kurtosis. (n.d.). Retrieved March 26, 2020, from https://help.gooddata.com/doc/en/reporting-and-dashboards/maql-analytical-query-language/maql-expression-reference/aggregation-functions/statistical-functions/predictive-statistical-use-cases/normality-testing-skewness-and-kurtosis

[4] Correlation Test Between Two Variables in R. (n.d.). Retrieved March 18, 2020, from http://www.sthda.com/english/wiki/correlation-test-between-two-variables-in-r Admin. (2020, March 11). A Guide to Bartletts Test of Sphericity. Retrieved March 18, 2020, from https://www.statology.org/a-guide-to-bartletts-test-of-sphericity/

[5] Holmes, E., Koehler, B., & Valent, H. (2019, May 19). Final Project: Student Performance Analysis. Retrieved March 16, 2020, from https://hugovalent.com/mlearn.html

[6] Normality Test in R. (n.d.). Retrieved March 23, 2020, from http://www.sthda.com/english/wiki/normality-test-in-r

[7] MarinStatsLectures. (2013, November 14). Checking Linear Regression Assumptions in R | R Tutorial 5.2 [Video file]. Retrieved from https://www.youtube.com/watch?v=eTZ4VUZHzxw&t=326s

[8] Halthor, A. (2017, November 26). What are the advantages and disadvantages of linear regression? Retrieved March 24, 2020, from https://www.quora.com/What-are-the-advantages-and-disadvantages-of-linear-regression

[9] Basic evaluation measures from the confusion matrix. Retrieved March 23, 2020, from https://classeval.wordpress.com/introduction/basic-evaluation-measures/

[10] Santiago, P. et al. (2012), OECD Reviews of Evaluation and Assessment in Education: Portugal 2012, OECD Publishing. http://dx.doi.org/10.1787/9789264117020-en

[11] Cheng, L. (1970, January 1). [PDF] Exploring the Factors that Affect Secondary Student's Mathematics and Portuguese Performance in Portugal: Semantic Scholar. Retrieved from https://www.semanticscholar.org/paper/Exploring-the-Factors-that-Affect-Secondary-and-in-Cheng/7b3b4522747e1f295ed5abb3b449913d285732ac?tab=abstract&citingPapersSort=is-influential&citingPapersLimit=10&citingPapersOffset=0&year[0]=&year[1]=&citedPapersSort=year&citedPapersLimit=10&citedPapersOffset=0

[12] Numerical Data Descriptive Statistics. (n.d.). Retrieved from https://ucr.github.io/descriptives_numeric

[13] 1.3.5.11. Measures of Skewness and Kurtosis. (n.d.). Retrieved from https://itl.nist.gov/div898/handbook/eda/section3/eda35b.htm

[14] Are the Skewness and Kurtosis Useful Statistics? (2019, June 21). Retrieved from https://www.spcforexcel.com/knowledge/basic-statistics/are-skewness-and-kurtosis-useful-statistics

[15] Pearsons Correlation Coefficient. (n.d.). Retrieved from https://www.statisticssolutions.com/pearsons-correlation-coefficient/

[16] Disclaimer Examples. (n.d.). Retrieved from https://www.termsfeed.com/blog/disclaimer-examples/

[17] degree of correlation lrh63944 -Statistics for Economics. (n.d.). Retrieved from https://www.topperlearning.com/doubts-solutions/degree-of-correlation-lrh63944

[18] Evergreen, S. (n.d.). Scatterplot. Retrieved from https://www.betterevaluation.org/en/evaluation-options/scatterplot

* Special thanks to the lecturers who conducted lectures on UECM3993 Predictive Modelling (Dr. Liew How Hui), UECM3493 Introduction to Time Series and Forecasting (Ms Lee Yap Jia), and UECM2263 Applied Statistical Model (Dr. Mahboobeh Zangeneh Sirdari) so that we are guided to write most of the R codes for this assignment.