

**Assignment Part 3 : Classification**

**DATA MINING**

**(TDS 3301)**

**LECTURER: Dr. Peter Ho**

|  |  |
| --- | --- |
| **Name** | **Student ID** |
| Benjamin Au Wen Wei | 1141327105 |
| Mah Siew Chin | 1132702455 |
| Tan Chong Raen | 1121115725 |
| Oh Soon Kit | 1132702874 |

**FACULTY OF COMPUTING AND INFORMATICS,**

**MULTIMEDIA UNIVERSITY**

**CYBERJAYA**

**3 February, 2017**

The final part of the assignment requires the group to complete a classification task. As usual, choose a dataset to perform the classification task on. Apply Decision Trees, Naïve Bayes and ANN on the classification task and compare the performance of the classifiers using measures such as accuracy, TPR, FPR etc. and so forth.

# **1. Exploratory data analysis**

The dataset that measures the Portuguese student’s performance in Mathematics was chosen from Paulo Cortez and Alice Silva’s work, “Using Data Mining To Predict Secondary School Student Performance” from its source: <http://archive.ics.uci.edu/ml/datasets/Student+Performance>. This dataset is composed of detailed observation for each student across multiple variables such as grades, demographic, social and school related features described as follow:

1 school – student’s school (binary: ‘GP’ – Gabriel Pereira or ‘MS’ – Mousinho da Silveira)

2 sex – student’s sex (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 â 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 guardian – student’s guardian (nominal: ‘mother’, ‘father’ or ‘other’)

13 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)

14 studytime – weekly study time (numeric: 1 – <2 hours, 2 – 2 to 5 hours, 3 – 5 to 10 hours, or 4 – >10 hours)

15 failures – number of past class failures (numeric: n if 1<=n<3, else 4)

16 schoolsup – extra educational support (binary: yes or no)

17 famsup – family educational support (binary: yes or no)

18 paid – extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)

19 activities – extra-curricular activities (binary: yes or no)

20 nursery – attended nursery school (binary: yes or no)

21 higher – wants to take higher education (binary: yes or no)

22 internet – Internet access at home (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)

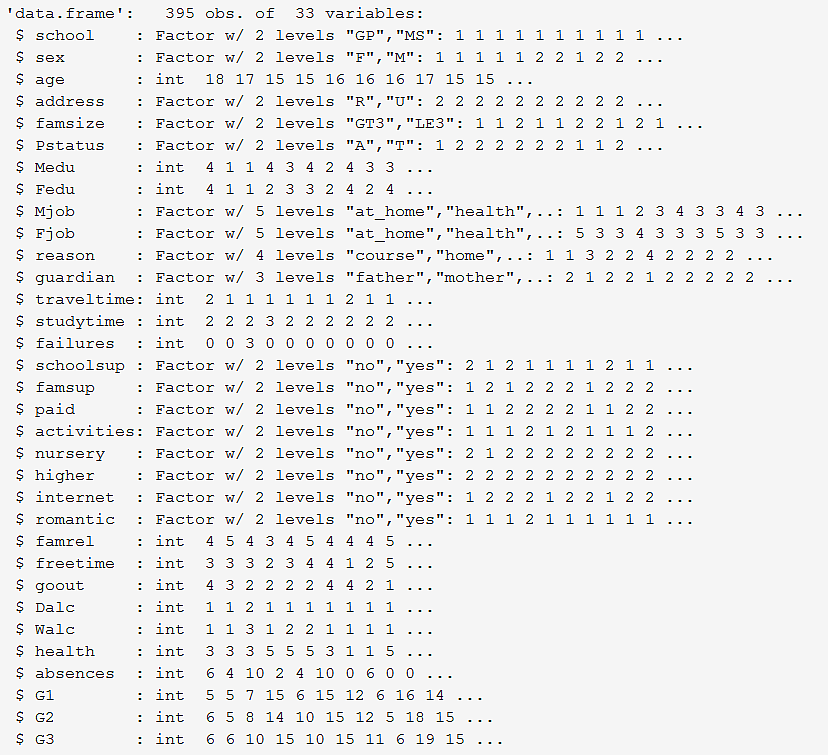
30 absences – number of school absences (numeric: from 0 to 93)

31 G1 – first period grade related to the course subject (numeric: from 0 to 20)

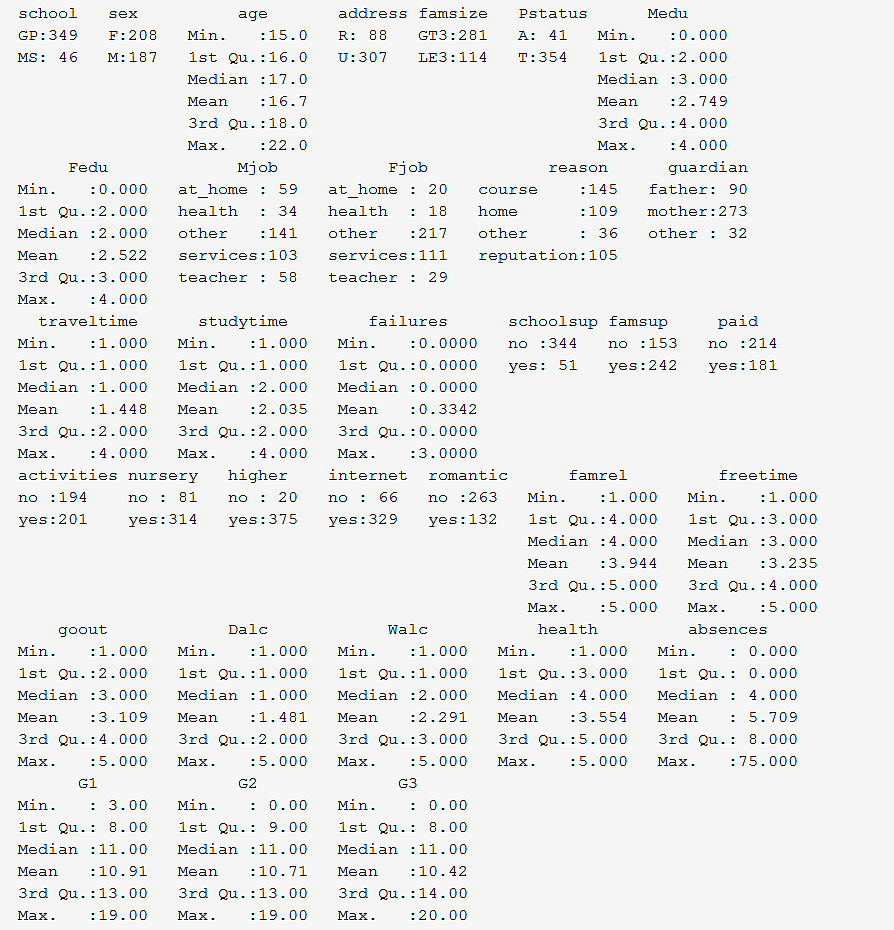
32 G2 – second period grade related to the course subject (numeric: from 0 to 20)

33 G3 – final grade related to the course subject (numeric: from 0 to 20, output target)

## **1.1) The Math dataset composed of 395 observations and 33 combinations of factors and numeric variables**



## **1.2) Summary of Math dataset**



## **1.3) Observed first 6 observations of dataset Math**

school sex age address famsize Pstatus Medu Fedu Mjob Fjob reason guardian

1 GP F 18 U GT3 A 4 4 at\_home teacher course mother

2 GP F 17 U GT3 T 1 1 at\_home other course father

3 GP F 15 U LE3 T 1 1 at\_home other other mother

4 GP F 15 U GT3 T 4 2 health services home mother

5 GP F 16 U GT3 T 3 3 other other home father

6 GP M 16 U LE3 T 4 3 services other reputation mother

traveltime studytime failures schoolsup famsup paid activities nursery higher internet

1 2 2 0 yes no no no yes yes no

2 1 2 0 no yes no no no yes yes

3 1 2 3 yes no yes no yes yes yes

4 1 3 0 no yes yes yes yes yes yes

5 1 2 0 no yes yes no yes yes no

6 1 2 0 no yes yes yes yes yes yes

romantic famrel freetime goout Dalc Walc health absences G1 G2 G3

1 no 4 3 4 1 1 3 6 5 6 6

2 no 5 3 3 1 1 3 4 5 5 6

3 no 4 3 2 2 3 3 10 7 8 10

4 yes 3 2 2 1 1 5 2 15 14 15

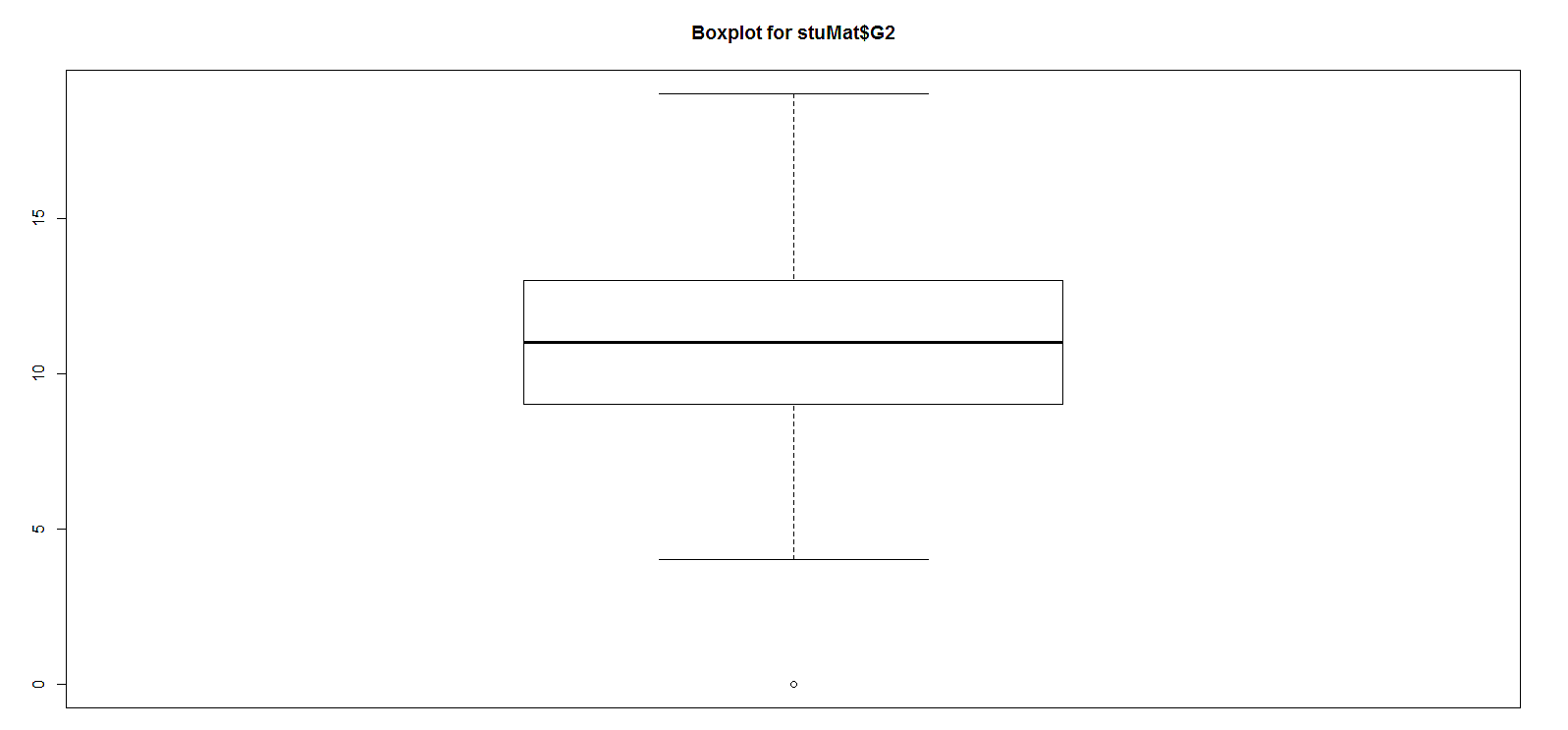
5 no 4 3 2 1 2 5 4 6 10 10

6 no 5 4 2 1 2 5 10 15 15 15

## **1.4) Check Outlier**

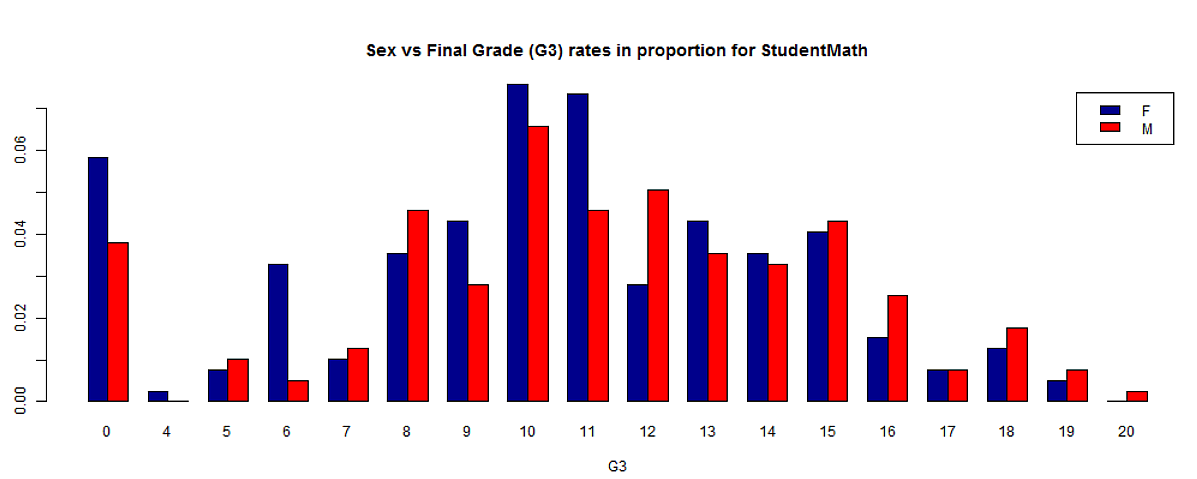


We found that the is outlier for absences variable but it should be acceptable since the range for number of school absences (numeric: from 0 to 93)



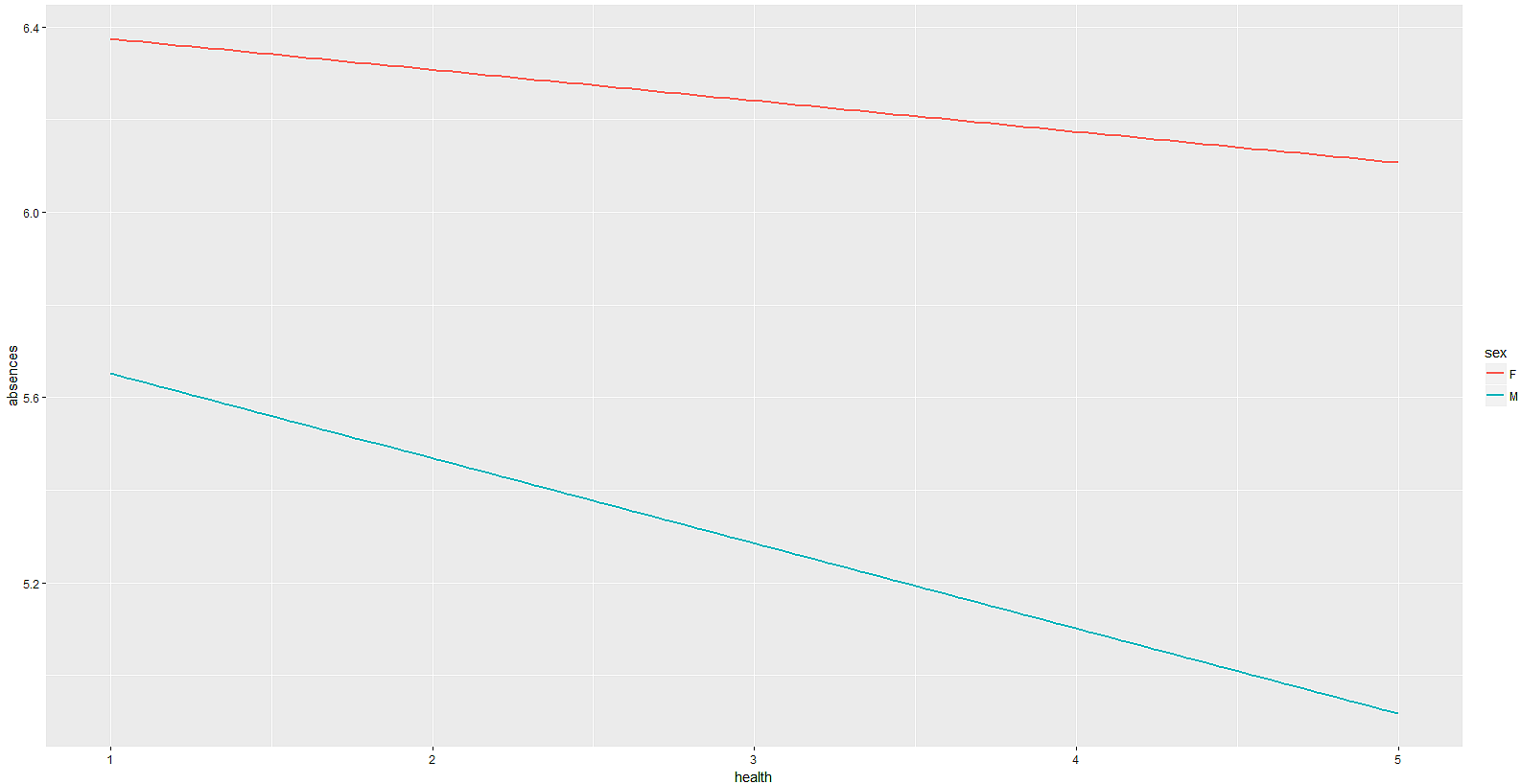
We found that the outlier for G2 variable is acceptable. Since, the target range for Grade 2 (G2) is from 0 to 20, so we won't correct it because we wanted to observe the relationship for grade G1, G2 and G3

## **1.5) two-way comparison on the number of males and females vs final grade rates in proportion using the dataset Math**



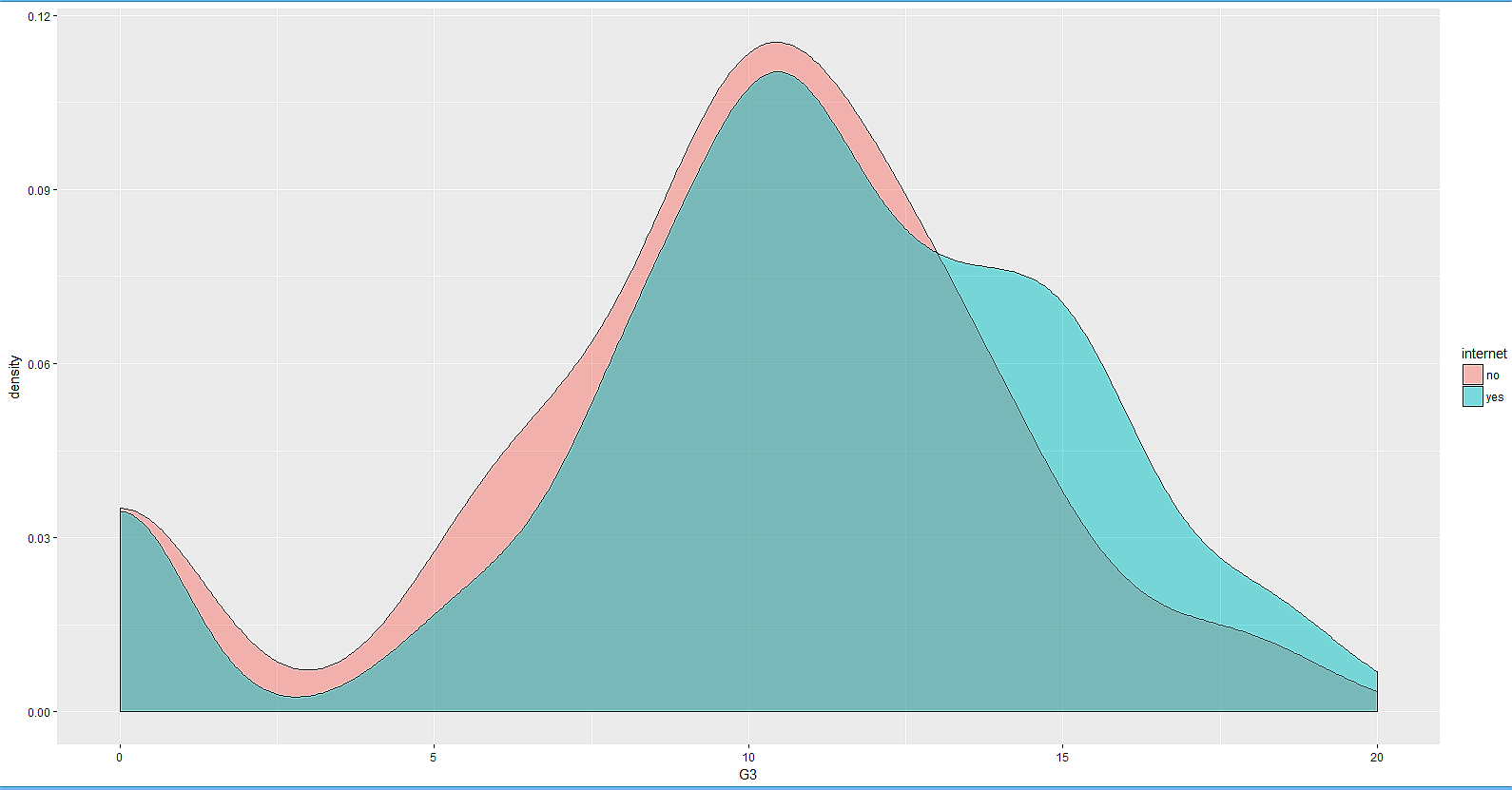
We found that the number of male students have higher final grade (G3) than the number of female students observed from grade 15-20.

## **1.6) The relationship between health and attendance in class and the gender of the student**



We found that female students have lower attendance on average and as the health scale increases, the absence decreases as expected for both male and female students.

## **1.7) Relationship between access to internet and the performance of the students**

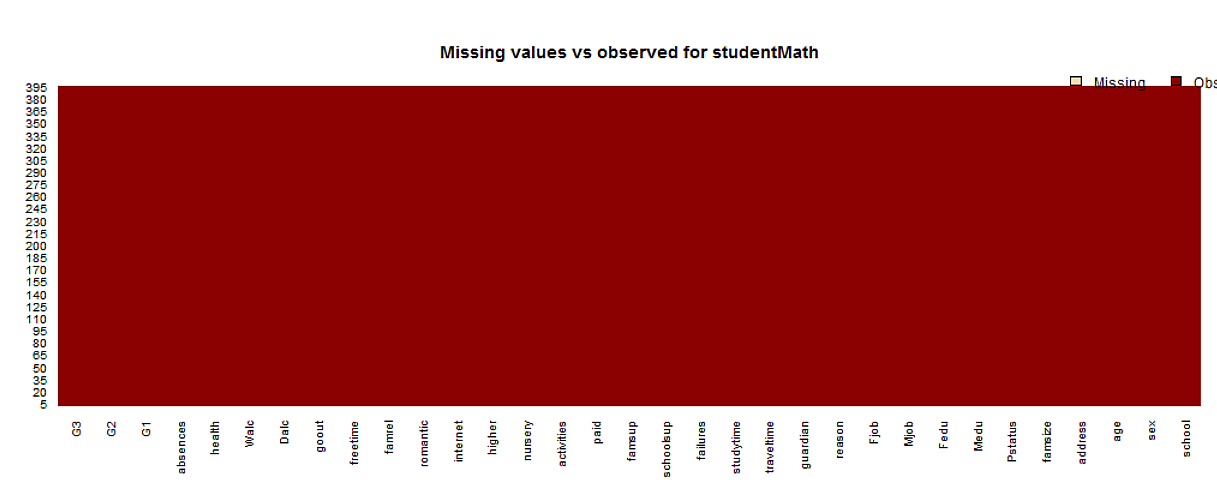


We found that the students who have access to internet obtained better final grade (G3) compare to students without internet access.

# **2. Pre-processing tasks**

## **2.1) Data Cleaning**

### **2.1.1) Missing Value**



No missing data was found in the dataset.

### **2.1.2) Does the data contain other special values?**

any(sapply(stuMat, is.nan))

[1] FALSE

### **2.1.3) No duplicate record**

anyDuplicated(stuMat)

[1] 0

## **2.2) Preparation steps**

a) Since the dataset is already cleaned and in a tidy format, the majority of the cleaning is done is selecting the necessary variables for this project.

b) We split the dataset in a 70% training set and a 30% test set and apply with 3 different classifiers Decision Trees, Naïve Bayes and ANN on the classification task and compare the performance of the classifiers.

c) The classifier predicts if a student will pass (G3 >= 10) or fail (G3 <10) Mathematics on the final grade (G3) variable modeled as a binary classification. We naming the G3 to a binary pass or fail grade called final and use this as our class that we wish to predict for future students.

d) To assess the predictive performances, 20 runs of a 10- fold cross-validation were applied to each configuration parameters. The selected parameters to determine the final grade include:Previous grade result : G1 and G2

Attendance

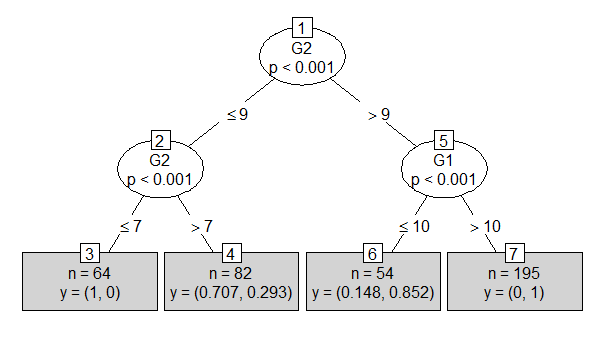
Father’s Job : Fjob

Internet : access to Internet

e) There are different steps needed to process in prior for the 3 classifier model generated described as follow:

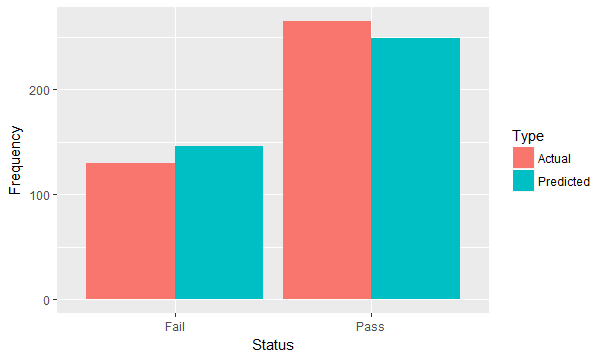
***1) Decision Trees***

From our model, G3 represents the final grade of student performances in Mathematics, a new column named ‘Pass\_Or\_Fail’ is created based on G3 where is must be at least 10 to pass the subject else is considered failed. There are only certain columns to be plotted on tree namely Pstatus, Medu, traveltime, romantic, G1, G2, G3, Pass\_Or\_Fail. Pass\_Or\_Fail is taken as the root node of the tree, from there it will branch out 2 paths to decision making.



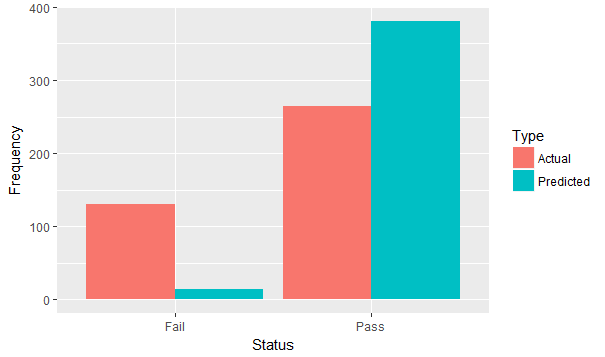
Based on the plotted tree above, if G2 is less than or equals to 9, is will branch to another G2 where again to determine how many scored less than or equals to 7 or more than 7. Nevertheless, scoring below 10 will result in failure. On the other side, if is more than 9 it goes to G1, then branches out again for one side scoring less than or equals to 10 and more than 10 on the other side.

n is the number of students, y has 2 values where the first value is probability of students achieving the results, followed by probability of students not achieving the results. The total number of n is 395.



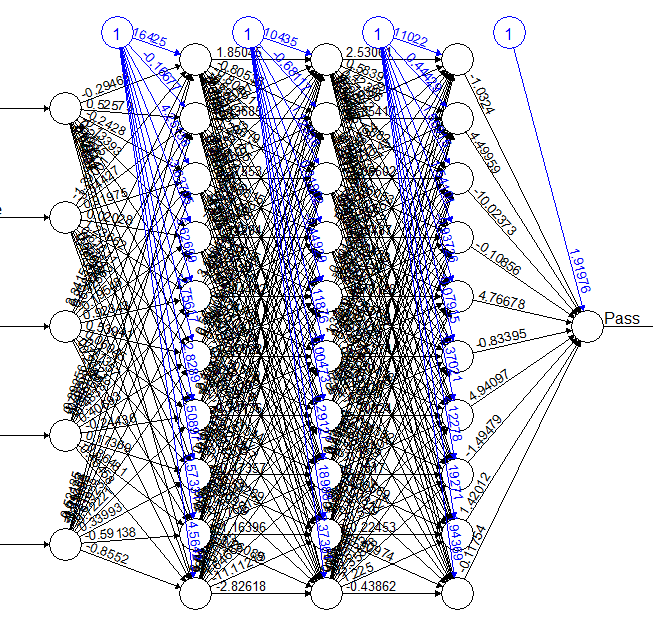
Based on the bar chart above, it is shown that the most of the students in fact did better than predicted, however quite a number of students did worst than prediction. Difference between prediction and actual results is acceptable, therefore the prediction made in this model may be used for future reference in predictive work.

***2) Naïve Bayes***



Based on the bar chart above, we can see that there is a huge difference between the actual result and predicted result where predicted fail result shows a noticeable low frequency as compared to the actual fail result while predicted pass result shows a noticeable high frequency as compared to the actual pass result. This unstable prediction result of this model unlikely to be used for future reference in predictive work.

## ***3) ANN***



The ANN required the dataset be normalizes using one-of-n normalizations.

The nominal 3 variables (e.g Mjob) were transformed into a 1-of-C encoding and all attributes were standardized to a zero mean and one standard deviation.

**3. Choice of performance measures**

The performance of 3 classifier generated are compared by using the following evaluation measures: accuracy, precision and recall.

# **4. Performance of the 3 classifiers**

The Naïve Bayes and Decision Tree classifiers gave very similar results on term of accuracy, but when it comes to precision, the Decision Tree was the most precise. Decision tree prediction on the same dataset showed very less errors in predicting pass or fail result of students while Naive Bayes method showed huge different and unstable on predictions and hence it is not a good predicting model for this dataset.

**5. Suggestion as to why the classifiers behave differently.**

Artificial neural network was superior when the number of categorical variables was two or more; the artificial neural network performance improved faster than that of the other methods as the number of classes of categorical variable increased.

Neuron networks are more affected by irrelevant inputs than the DT/RF algorithms, since the latter explicitly perform an internal feature selection.

It is found that the performance of classification techniques varies with different data sets. Factors that affect the classifier’s performance are data set, number of tuples and attributes, type of attributes, system configuration. Multilayer perceptron outperformed with two datasets and Naives Bayes updatable has given good results with a data set.

It can be seen that low error rates are obtained while maintaining a reasonably low false positive rate and that the decision tree classifier has the highest precision among the three classifiers.

Decision tree has the ability to use both numeric and categorical variables as predictors:Our data is a mix of categorical and numeric variables and other commonly used methods such as linear or logistic regressions can use either of these categories but not both in a single algorithm.

The obtained results reveal that it is possible to achieve a high predictive accuracy, provided that the first and/or second school period grades are known

# 

# 

# **6. Extra Credit: Completing a Kaggle competition – proven by team score**

Our team named R Pro have completed [Titanic: Machine Learning from Disaster competition in Kaggle.](https://www.kaggle.com/c/titanic)

