MULTIMEDIA UNIVERSITY®

Bridging Boundaries, Connecting Minds

Faculty of Computing and Informatics

(FCI)

TDS3301

DATA MINING

Assignment

Part 3 Classification

Prepared by:

Name	ID
Yap Kit Boon	1141124580
Thong Yong Lin	1132702398
Ooi Yi Jie	1131122872
Chew Siu Min	1122703126

PART 3: ASSOCIATION RULE MINING

Dataset chosen: **Student Performance dataset**
From: https://archive.ics.uci.edu/ml/datasets/Student+Performance

Introduction

In this assignment, we have selected the 'student performance dataset' to do the classification task. There are two datasets, namely 'student-mat' and 'student-port', where 'student-mat' represent the students' performance in Mathematics subject and 'student-port' represent the students' performance in Portuguese language subject. Both datasets consist of 33 attribute columns and different number of data rows; the first consists of 395 data rows and the second consists of 649 data rows. Each data row represent one independent record and the attributes consist of information including student grades, demographic, social and school related features. The aim of classification tasks using this data is to separately classify the class variable which is the final grade (named G3) of students for each subject.

A. Exploratory data analysis

Exploratory data analysis (EDA) is an approach to analyzing datasets to summarize their main characteristics.

- View the structure of the dataset

```
'data.frame':   395 obs. of  33 variables:
 $ school    : Factor w/ 2 levels "GP","MS": 1 1 1 1 1 1 1 1 1 1 ...
 $ sex       : Factor w/ 2 levels "F","M": 1 1 1 1 1 2 2 1 2 2 ...
 $ age       : int  18 17 15 15 16 16 16 17 15 15 ...
 $ address   : Factor w/ 2 levels "R","U": 2 2 2 2 2 2 2 2 2 2 ...
 $ famsize   : Factor w/ 2 levels "GT3","LE3": 1 1 2 1 1 2 2 1 2 1 ...
 $ Pstatus   : Factor w/ 2 levels "A","T": 1 2 2 2 2 2 2 1 1 2 ...
 $ Medu      : int  4 1 1 4 3 4 2 4 3 3 ...
 $ Fedu      : int  4 1 1 2 3 3 2 4 2 4 ...
 $ Mjob      : Factor w/ 5 levels "at_home","health",..: 1 1 1 2 3 4 3 3 4 3 ...
 $ Fjob      : Factor w/ 5 levels "at_home","health",..: 5 3 3 4 3 3 3 5 3 3 ...
 $ reason    : Factor w/ 4 levels "course","home",..: 1 1 3 2 2 4 2 2 2 2 ...
 $ guardian  : Factor w/ 3 levels "father","mother",..: 2 1 2 2 1 2 2 2 2 2 ...
 $ traveltime: int  2 1 1 1 1 1 1 2 1 1 ...
 $ studytime : int  2 2 2 3 2 2 2 2 2 2 ...
 $ failures  : int  0 0 3 0 0 0 0 0 0 0 ...
 $ schoolsup : Factor w/ 2 levels "no","yes": 2 1 2 1 1 1 1 2 1 1 ...
 $ famsup    : Factor w/ 2 levels "no","yes": 1 2 1 2 2 2 1 2 2 2 ...
 $ paid      : Factor w/ 2 levels "no","yes": 1 1 2 2 2 2 1 1 2 2 ...
 $ activities: Factor w/ 2 levels "no","yes": 1 1 1 2 1 2 1 1 1 2 ...
 $ nursery   : Factor w/ 2 levels "no","yes": 2 1 2 2 2 2 2 2 2 2 ...
 $ higher    : Factor w/ 2 levels "no","yes": 2 2 2 2 2 2 2 2 2 2 ...
 $ internet  : Factor w/ 2 levels "no","yes": 1 2 2 2 1 2 2 1 2 2 ...
 $ romantic  : Factor w/ 2 levels "no","yes": 1 1 1 2 1 1 1 1 1 1 ...
 $ famrel    : int  4 5 4 3 4 5 4 4 4 5 ...
 $ freetime  : int  3 3 3 2 3 4 4 1 2 5 ...
 $ goout     : int  4 3 2 2 2 2 4 4 2 1 ...
 $ Dalc      : int  1 1 2 1 1 1 1 1 1 1 ...
 $ Walc      : int  1 1 3 1 2 2 1 1 1 1 ...
 $ health    : int  3 3 3 5 5 5 3 1 1 5 ...
 $ absences  : int  6 4 10 2 4 10 0 6 0 0 ...
 $ G1        : int  5 5 7 15 6 15 12 6 16 14 ...
 $ G2        : int  6 5 8 14 10 15 12 5 18 15 ...
 $ G3        : int  6 6 10 15 10 15 11 6 19 15 ...
```

G3 represents the final grade of students, and is represented in continuous manner (from 0 to 20). We can see the summary of the G3 for both subjects.

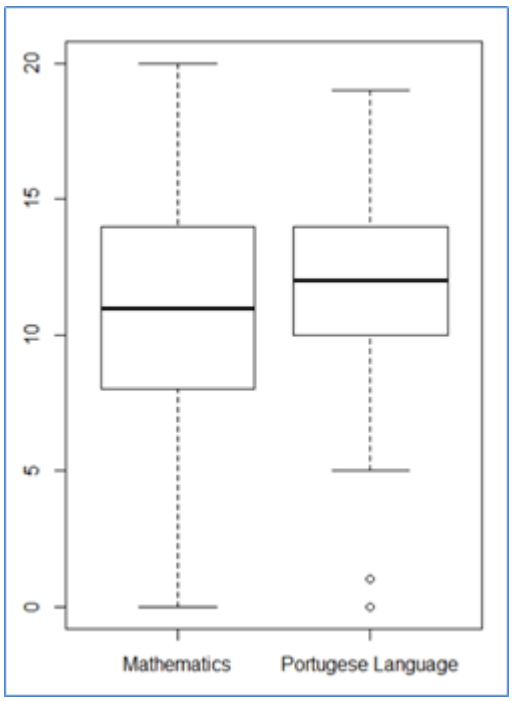
Mathematics:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.00	8.00	11.00	10.42	14.00	20.00

Portuguese Language:

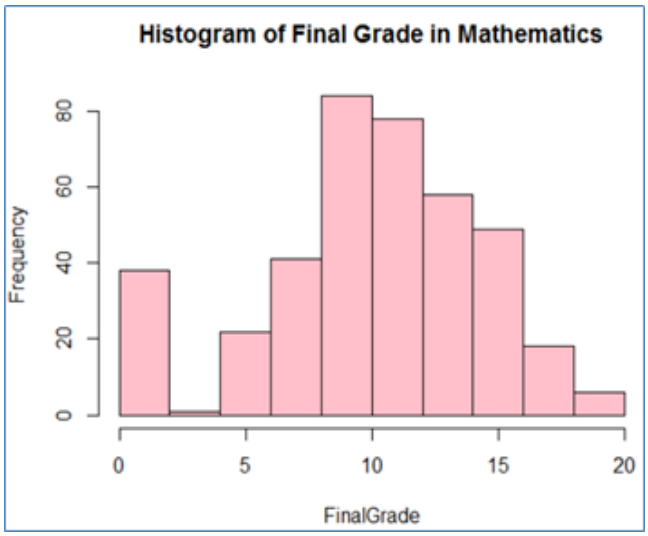
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.00	10.00	12.00	11.91	14.00	19.00

- We can produce boxplot to see the range of data distribution

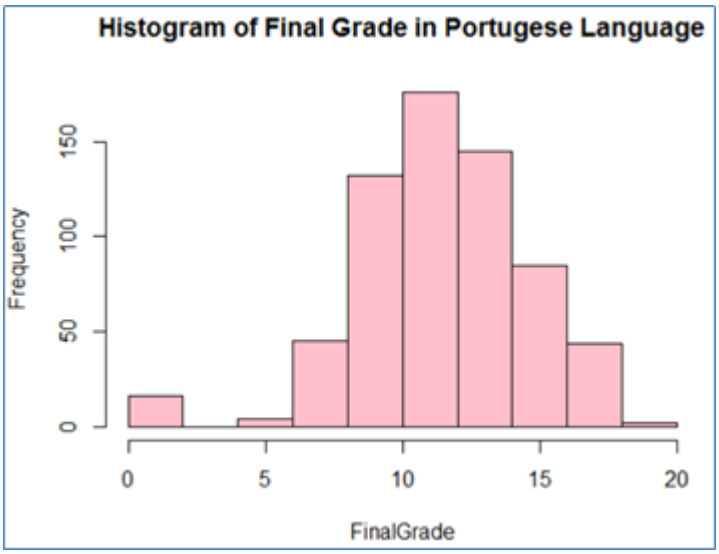


- We can also produce histograms to see the data and its frequency

Mathematics:

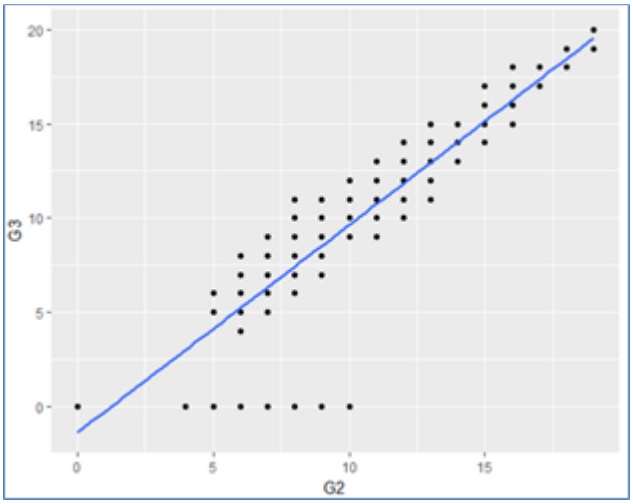


Portuguese Language:

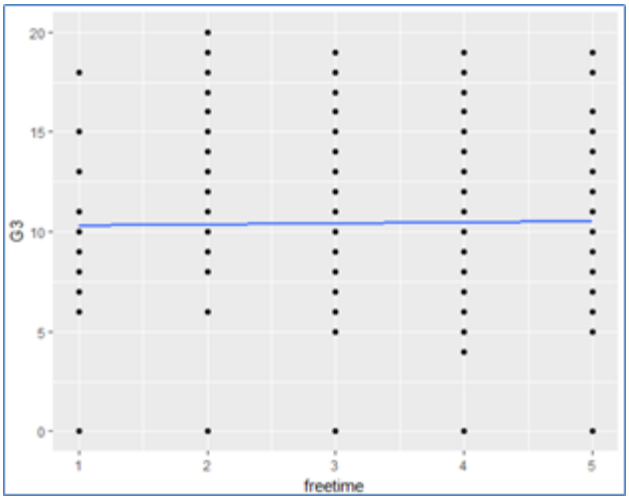


- We can find the correlation between variables by sketching it using ggplot

The graph below shows that there is positive correlation between G2 (second period grade) and G3 (ie: the student will most likely get high final grade, if he gets high second period grade beforehand)



- The 'freetime' (free time after classes), on the other hand, does not show obvious correlation with student's final grade



B. Pre-processing tasks

As the data set contains no missing data and data quality issues, there is no data cleaning task to be done. The preprocessing tasks we performed include the general preprocessing task and specific task for the classifiers before proceeding to the classification.

First of all, the **unnecessary columns** not needed in the analysis are **removed**, including column 'school', 'age', 'sex', 'famsize'.

A new column 'Grade' is created by **categorizing the data** from class variable **'G3'** according to the academic grading in Portugal retrieved from Academic Grading in Portugal in Wikipedia (https://en.wikipedia.org/wiki/Academic_grading_in_Portugal)

Grade	Qualification
20 ⋮ 17.5	Excellent
17.4 ⋮ 15.5	Very good
15.4 ⋮ 13.5	Good
13.4 ⋮ 9.5	Sufficient
9.4 ⋮ 3.5	Weak
3.4 ⋮ 0	Poor

After categorizing,

G3	Grade
16	VeryGood
12	Sufficient
10	Sufficient
16	VeryGood
10	Sufficient
10	Sufficient
14	Good

For **ANN**, column 'schoolsup', 'famsup', 'paid','activities', 'nursery', 'higher', 'romantic' and 'internet' data values (Yes/NO) have been **converted into binary number (1/0).**

Before preprocessing:

schoolsup	famsup	paid	activities	nursery	higher	internet	romantic
yes	no	no	no	yes	yes	no	no
no	yes	no	no	no	yes	yes	no
yes	no	no	no	yes	yes	yes	no
no	yes	no	yes	yes	yes	yes	yes
no	yes	no	no	yes	yes	no	no

After preprocessing:

schoolsup	famsup	paid	activities	nursery	higher	internet	romantic
1	0	0	0	1	1	0	0
0	1	0	0	0	1	1	0
1	0	0	0	1	1	1	0
0	1	0	1	1	1	1	1
0	1	0	0	1	1	0	0

Next, all the **categorical data** from the table have been **normalized** and extracted to different data tables. The data contains the normalized data with binary number (1/0).

Before:

address
U
U
R
R

After:

R	U
0	1
0	1
1	0
1	0

The columns have been renamed according to the corresponding column.

After discretization:

address_R	address_U
0	1
0	1
1	0
1	0

Besides, the **numeric data** is scaled using mins and maxs to **convert** the data into value between **0 and 1**.

For **Naïve Bayer**, the **ordinal data** is **converted into factors** with several levels. **Categorizing data** in **'G1' and 'G2'** also performed according to academic grading in Portugal.

Before:

G1	G2
16	17
15	17
11	9
7	7

After:

G1Grade	G2Grade
Excellent	Excellent
VeryGood	VeryGood
Sufficient	Sufficient
Weak	Weak

Discretization of numeric variables are also performed to convert them into different categories.

C. Choice of performance measures

For classification problems, the primary source for accuracy estimation is the confusion matrix. To do this, we separate out 70% of the data to be train data, and the rest of the data will be the test data. After the model is constructed using the train data, the test data is fed into the model and the predicted result is compared with the real result. We are to calculate the classification accuracy using the formula,

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

While TP represent True Positive, TN represent True Negative, FP represent False Positive, and FN represent False Negative.

D. Performance of the 3 classifiers

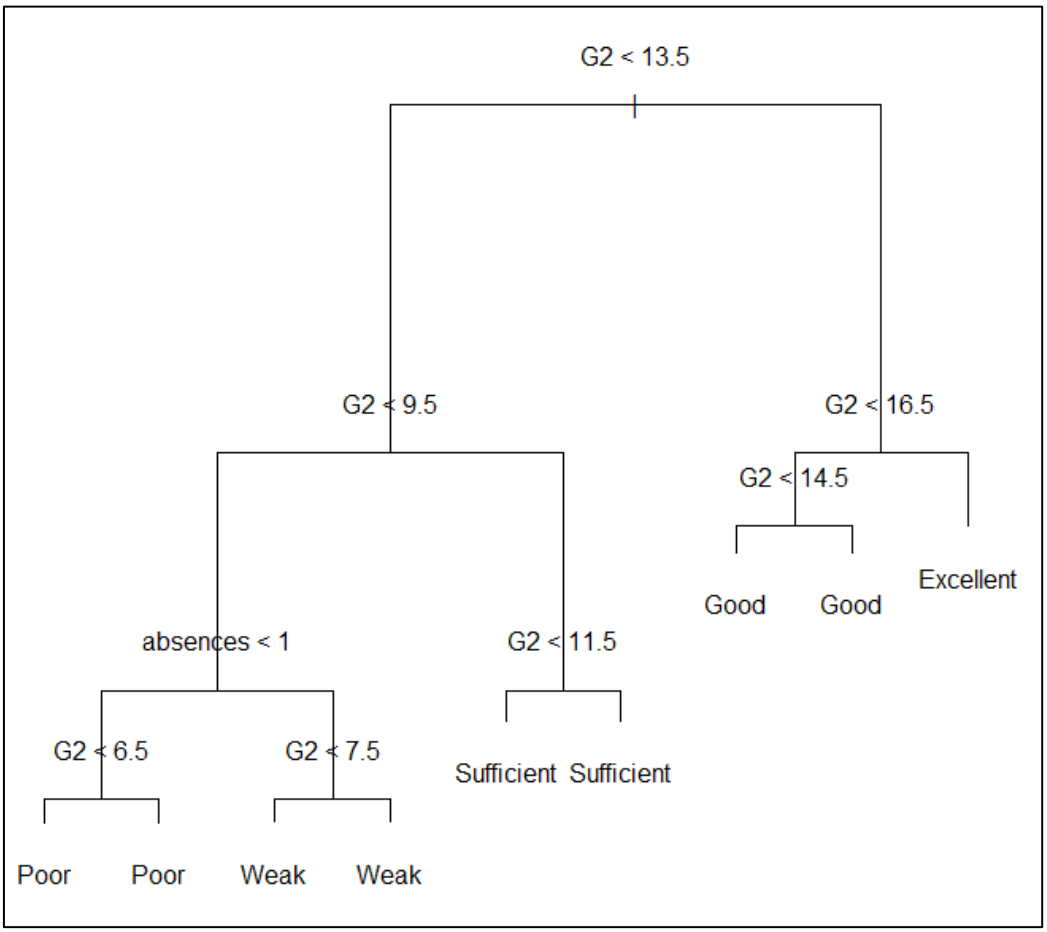
Three classifiers have been used, which is decision tree (DT), Artificial Neutral Network (ANN) and Naive Bayes (NB). All the classifiers are used to predict "Grade" attributes by all the attributes except G3. The following diagrams are showing the result of each classifiers for Mathematics and Portuguese subjects.

Mathematics Student:

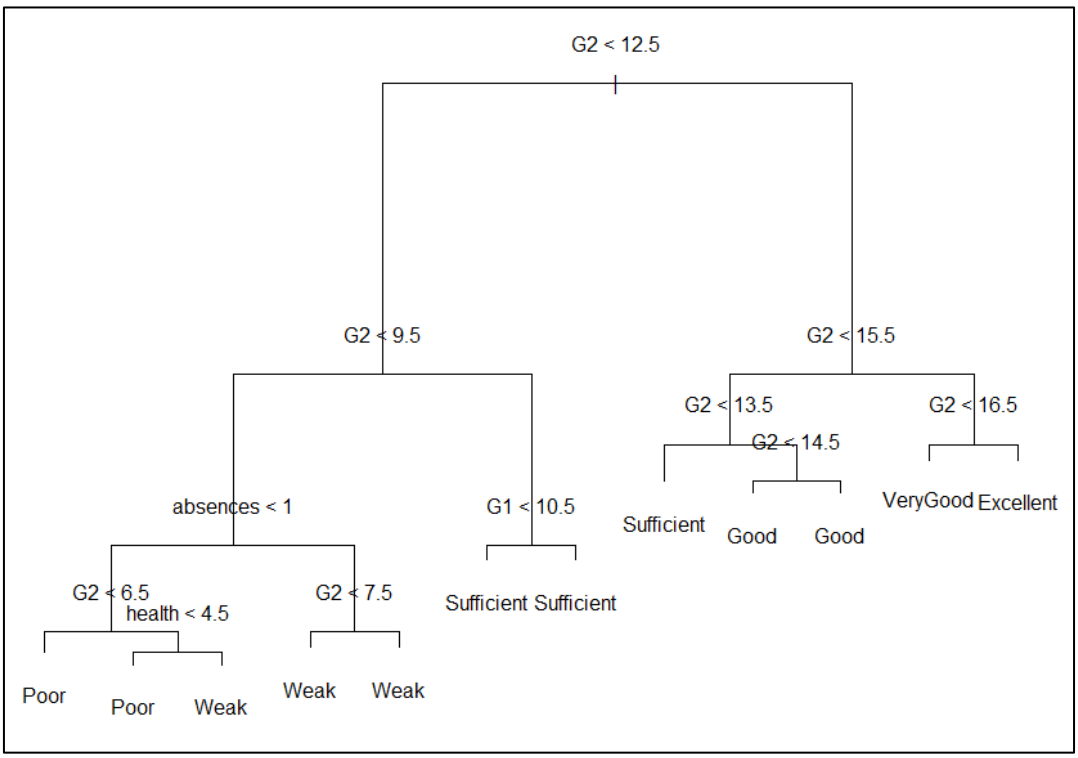
1. Decision Tree

Diagrams below are the visualization of decision tree by all the attributes except G3, to predict "Grade" of a student in Mathematics:

Before pruned:



After pruned:



As we can observed, it will be much better after decision tree is pruned, because the decision tree before pruned doesn't include "VeryGood" Grade in the output level. Therefore, we choose to use the pruned decision tree.

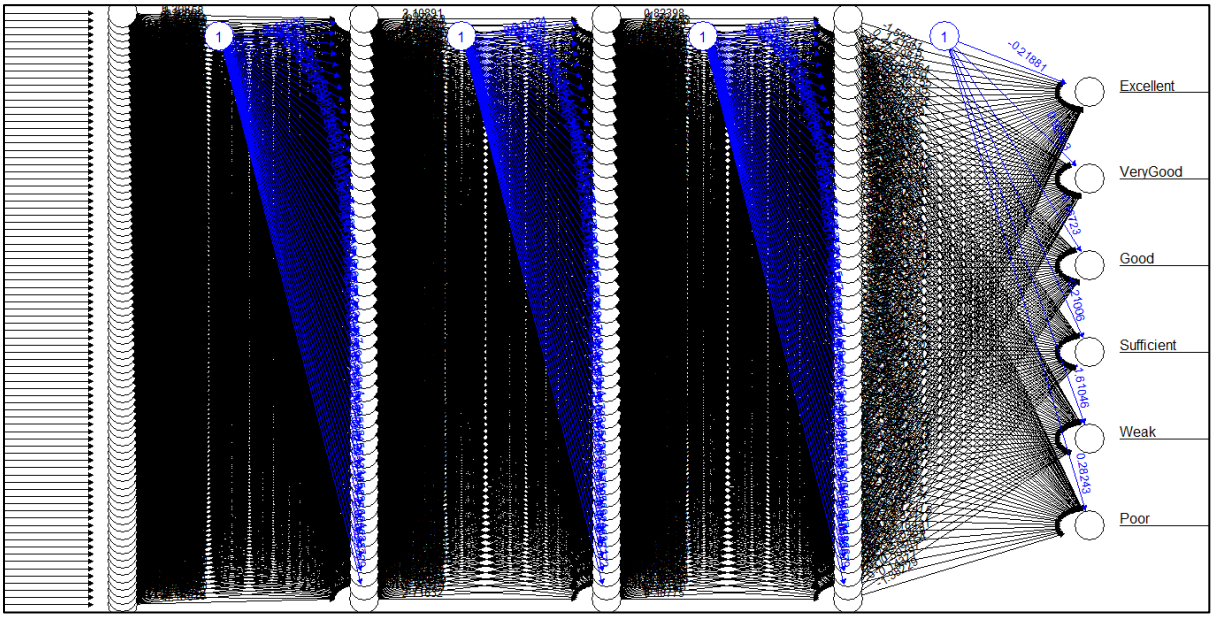
This is the result of confusion matrix for the pruned decision tree:

	Actual					
Prediction	Excellent	Good	Poor	Sufficient	VeryGood	Weak
Excellent	8	0	0	0	2	0
Good	0	12	0	2	3	0
Poor	0	0	10	0	0	1
Sufficient	0	1	2	45	0	0
VeryGood	0	2	0	0	3	0
Weak	0	0	0	9	0	19

Accuracy: **81.51%**.

2. Artificial Neural Network

As we have categorized Grade into six classes, so we also have six output for this method. The number of input is 83 attributes, we use the mean of number of input and output to determine the number of hidden layer elements, which is around 45 neurons in hidden layers. The diagram below is showing the visualization of its neural network method.



The following are showing the confusion matrix tables below are separated by the grade using ANN method.

1. Excellent

	Actual	
Prediction	0	1
0	114	3
1	1	1

Accuracy: **96.64%**

2. Very Good

	Actual	
Prediction	0	1
0	108	7
1	3	1

Accuracy: **91.60%**

3. Good

	Actual	
Prediction	0	1
0	108	7
1	3	1

Accuracy: **79.83%**

4. Sufficient

	Actual	
Prediction	0	1
0	46	16
1	21	36

Accuracy: **68.91%**

5. Weak

```
          Actual
Prediction  0  1
         0 83  9
         1 15 12
```

Accuracy: **79.83%**

6. Poor

```
          Actual
Prediction   0   1
         0 100   8
         1   5   6
```

Accuracy: **89.08%**

3. Naïve Bayes

The following is showing the result of Naive Bayes Classifier:

```
Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
    Excellent          Good          Poor     Sufficient      VeryGood          Weak
0.03985507246 0.15217391304 0.09782608696 0.42391304348 0.06159420290 0.22463768116

Conditional probabilities:
            address
Y                         R             U
  Excellent  0.09090909091 0.90909090909
  Good       0.11904761905 0.88095238095
  Poor       0.22222222222 0.77777777778
  Sufficient 0.26495726496 0.73504273504
  VeryGood   0.11764705882 0.88235294118
  Weak       0.25806451613 0.74193548387

            Pstatus
Y                         A             T
  Excellent  0.18181818182 0.81818181818
  Good       0.07142857143 0.92857142857
  Poor       0.03703703704 0.96296296296
  Sufficient 0.11111111111 0.88888888889
  VeryGood   0.11764705882 0.88235294118
  Weak       0.08064516129 0.91935483871

            Medu
Y                         0             1             2             3             4
  Excellent  0.00000000000 0.00000000000 0.18181818182 0.36363636364 0.45454545455
  Good       0.04761904762 0.04761904762 0.19047619048 0.21428571429 0.50000000000
  Poor       0.00000000000 0.22222222222 0.33333333333 0.22222222222 0.22222222222
  Sufficient 0.00000000000 0.15384615385 0.34188034188 0.23931623932 0.26495726496
  VeryGood   0.00000000000 0.11764705882 0.17647058824 0.17647058824 0.52941176471
  Weak       0.01612903226 0.16129032258 0.29032258065 0.24193548387 0.29032258065
```

This is Confusion Matrix for using Naive Bayes method:

	Actual					
Prediction	Excellent	Good	Poor	Sufficient	VeryGood	Weak
Excellent	6	0	0	0	0	0
Good	0	12	0	5	6	0
Poor	0	0	11	4	1	1
Sufficient	2	0	1	40	0	2
VeryGood	0	3	0	0	1	0
Weak	0	0	0	7	0	17

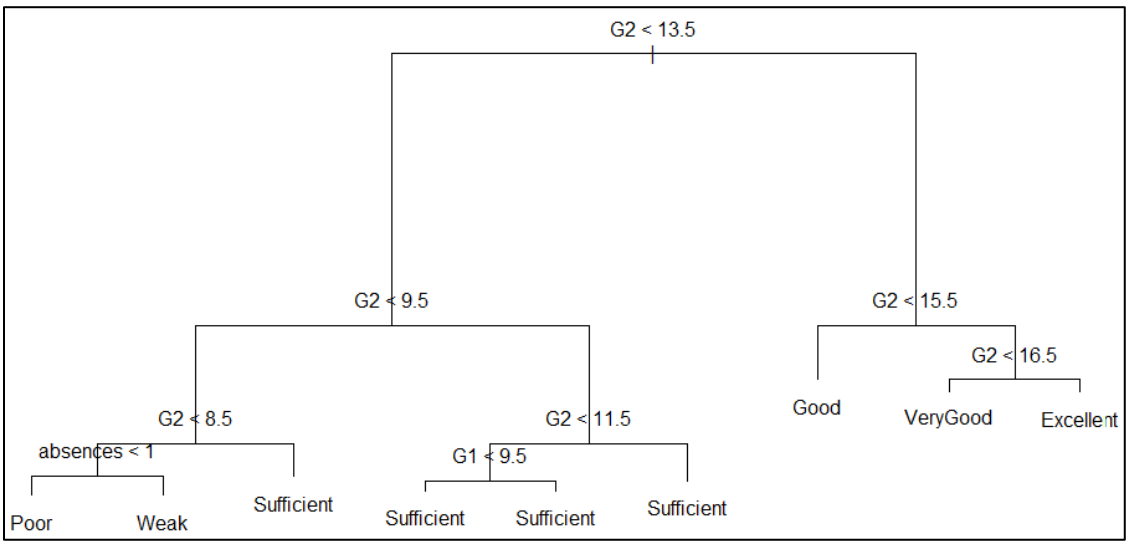
Accuracy: **73.11%**

Portuguese Language:

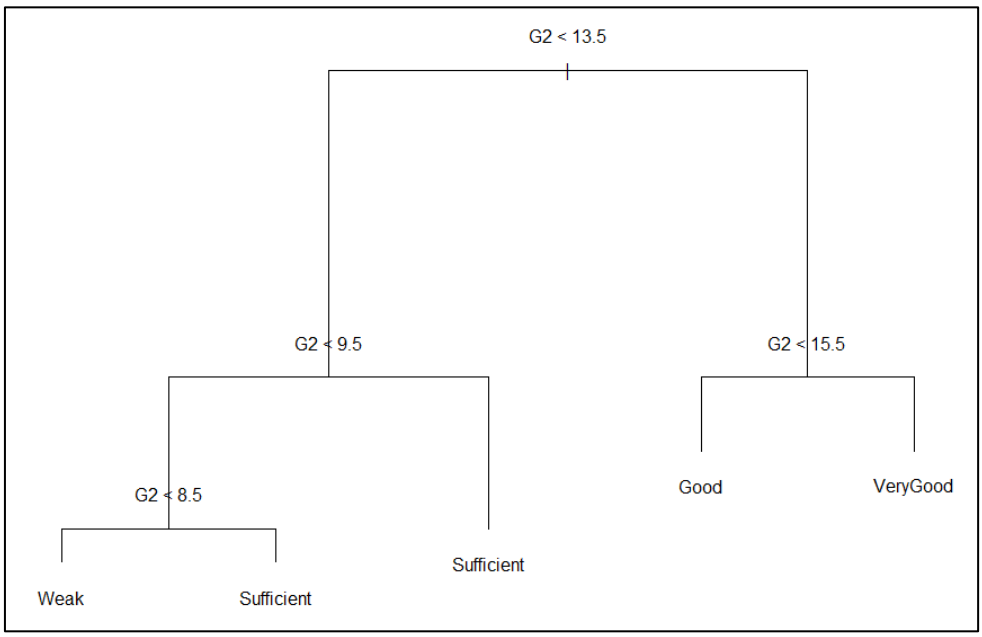
1. Decision Tree

Diagrams below are the visualization of decision tree by all the attributes except G3, to predict "Grade" of a student in Mathematics:

Before pruned:



After pruned:



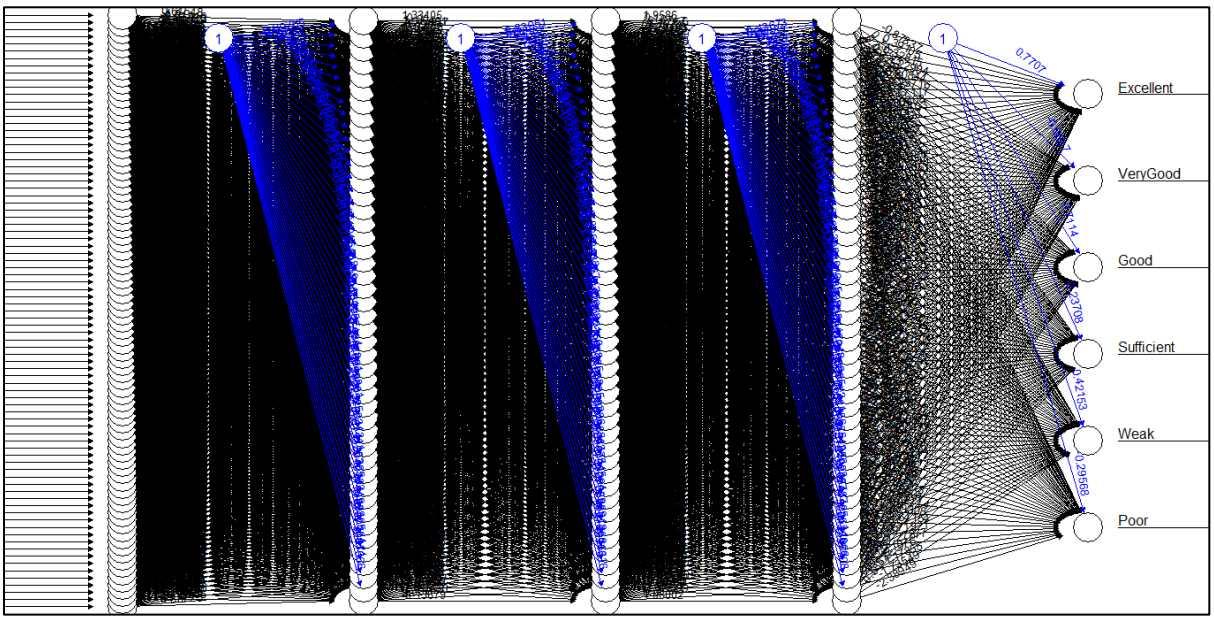
The Confusion Matrix for using Decision tree method:

Prediction	Actual Excellent	Good	Poor	Sufficient	VeryGood	Weak
Excellent	0	0	0	0	0	0
Good	0	15	0	2	8	0
Poor	0	0	0	0	0	0
Sufficient	0	10	1	113	0	12
VeryGood	6	2	0	0	9	0
Weak	0	0	7	1	0	9

Accuracy: **74.87%**

2. Artificial Neural Network

Visualization of ANN for Grade:



1. Excellent

Prediction	Actual 0	1
0	185	5
1	5	0

Accuracy: 94.87%

2. Very Good

Prediction	Actual 0	1
0	163	15
1	13	4

Accuracy: 85.64%

3. Good

	Actual	
Prediction	0	1
0	139	26
1	21	9

Accuracy: 75.90%

4. Sufficient

	Actual	
Prediction	0	1
0	61	28
1	28	78

Accuracy: 71.28%

5. Weak

	Actual	
Prediction	0	1
0	159	15
1	10	11

Accuracy: 87.18%

6. Poor

	Actual	
Prediction	0	1
0	189	2
1	2	2

Accuracy: 97.95%

3. Naïve Bayes

The following is showing the result of Naive Bayes Classifier.

	Actual					
Prediction	Excellent	Good	Poor	Sufficient	VeryGood	Weak
Excellent	1	0	0	0	1	0
Good	0	14	0	16	5	0
Poor	0	0	5	4	0	2
Sufficient	0	7	2	90	0	10
VeryGood	5	6	0	0	11	0
Weak	0	0	1	6	0	9

Accuracy: **66.67%**

After we done all the three classifiers, the overall result show that the best performance is doing Decision Tree, as it have achieved 81.51% in student_math dataset and 74.87% in student_por dataset. This is because it is the fastest algorithm among the others two methods, but if the time consuming is not a concern, we would like to say that the ANN is also a good choice among them, because it has the most accuracy in predicting some of the category in Grade from the result showing.

E. Suggestion as to why the classifiers behave differently.

Decision tree is the best choice when there are nominal and numeric attributes in a dataset, like the 'student_mat' and 'student_por'. Unlike ANN and Naive Bayes classification process, we do not need to discretize numeric variables or dealing with dummy variables when doing decision tree. As the result, its accuracy is the highest among the three classification methods.

ANN, results second highest accuracy followed by decision tree. It produces lower accuracy than decision tree because of the large amount of nominal attributes and excessive normalization into extra columns (up to 83 columns in our case).

Naive Bayes, on the other hand, results the lowest accuracy compared to decision tree and ANN methods. To get high classification accuracy in Naive Bayes, we need to ensure that the attributes are independent of each others. But it is not the case in our dataset as we still use all attributes, except 'G3' in our classification process. As a result, the accuracy is lower than the other two classifiers. However, Naive Bayes is the best choice to check independency between attributes in the dataset.