

**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

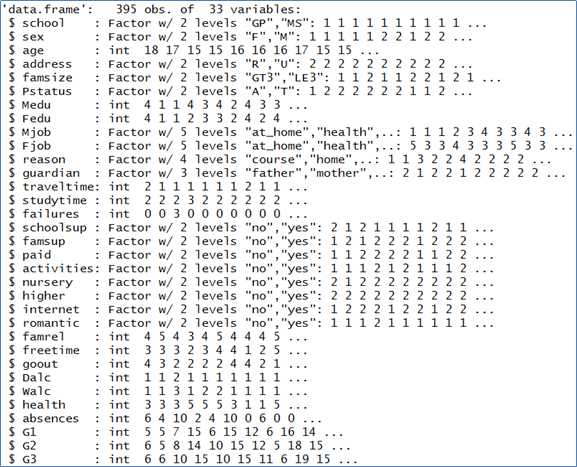
## Introduction

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



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.

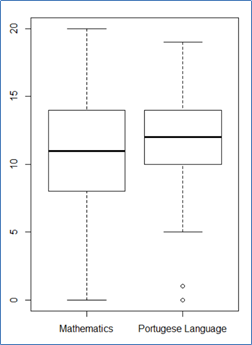
**Mathematics:**



**Portuguese Language:**

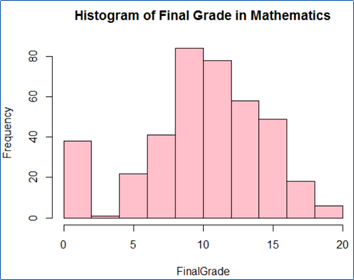


* We can produce boxplot to see the range of data distribution

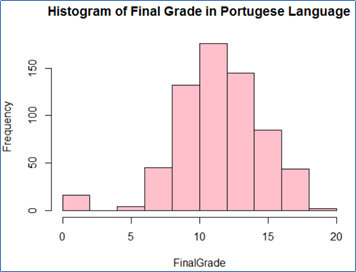


* We can also produce histograms to see the data and its frequency

**Mathematics:**

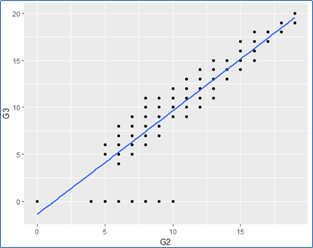


**Portuguese Language:**

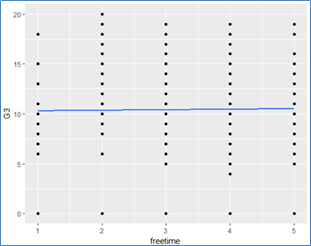


* We can find the correlation between variables by sketching it using ggplot

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

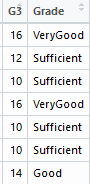
As the data set contains no missing data and data quality issues, there is no data cleaning task to be done. The preprocessing tasks we performed include thegeneral preprocessing task and specific task for the classifiers before proceeding to the classification.

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)

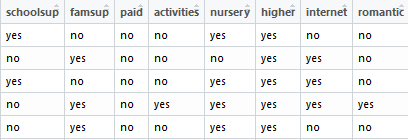


After categorizing,

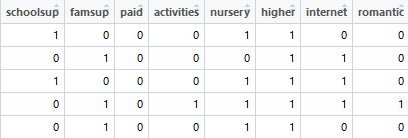


For **ANN**, column ‘schoolsup’, ‘famsup’, ‘paid’,’activities’, ‘nursery’, ‘higher’, ‘romantic’ and ‘internet’ data values (Yes/NO) have been **converted into binary number (1/0).**

Before preprocessing:

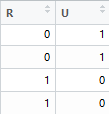


After preprocessing:



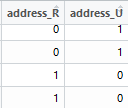
Next, all the **categorical data** from the table have been **normalized** and extracted to different data tables. The data contains the normalized data with binary number (1/0).

Before: After:

The columns have been renamed according to the corresponding column.

After discretization:



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.

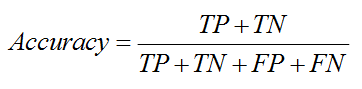
Before: After:

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,



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

## D. Performance of the 3 classifiers

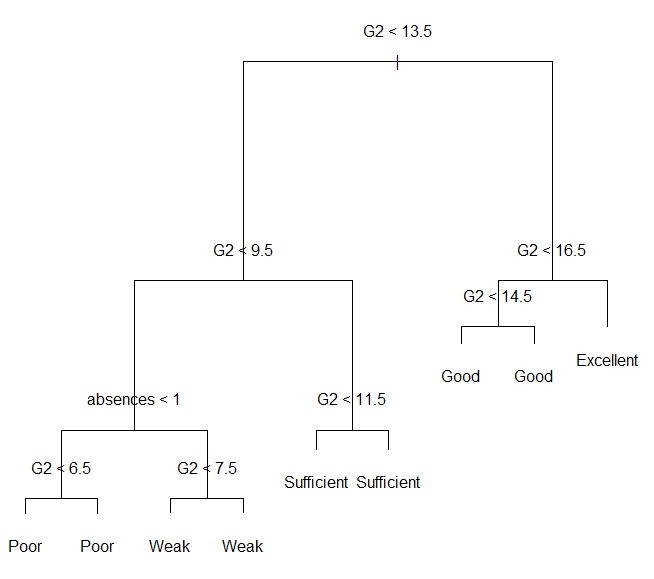
Three classifiers have been used, which is decision tree (DT), Artificial Neutral Network (ANN) and Naive Bayes (NB). All the classifiers are used to predict “Grade” attributes by all the attributes except G3. The following diagrams are showing the result of each classifiers for Mathematics and Portuguese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

### Mathematics Student:

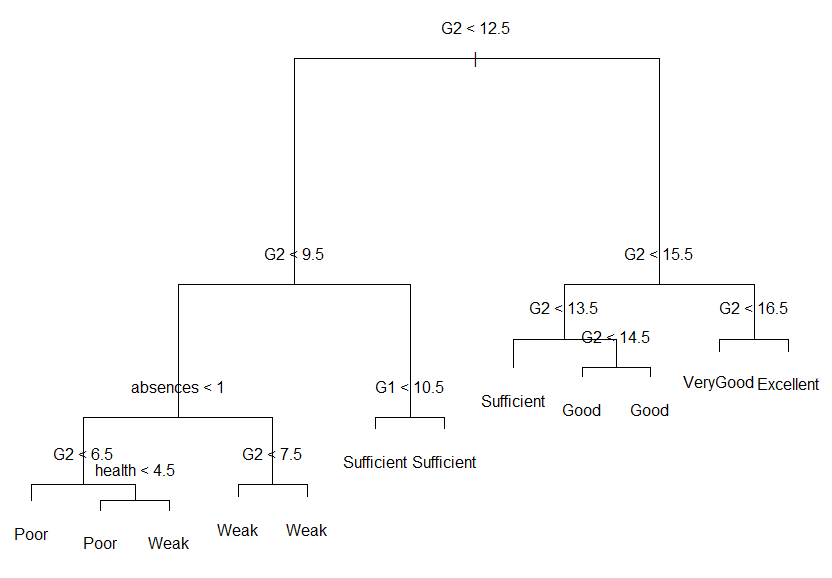
### Decision Tree

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

Before pruned:

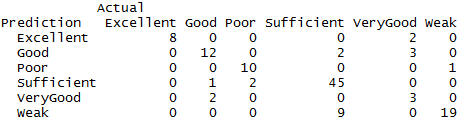


After pruned:



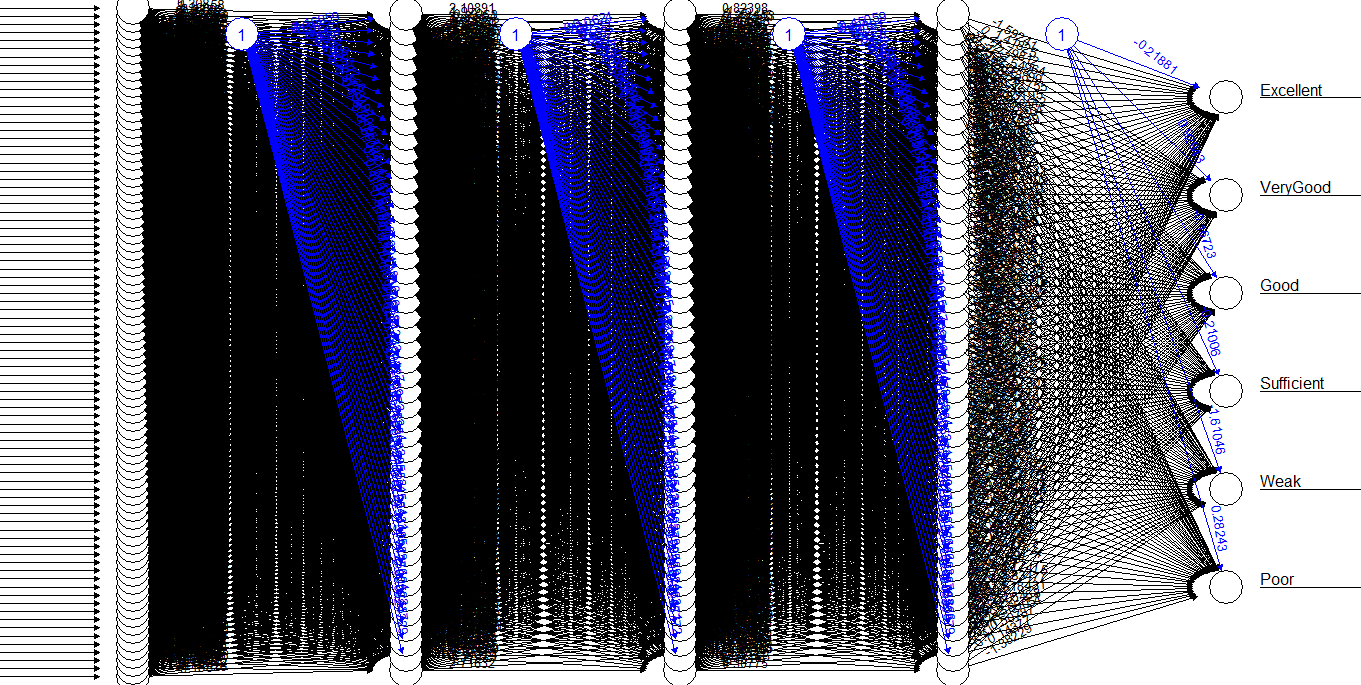
As we can observed, it will be much better after decision tree is pruned, because the decision tree before pruned doesn’t include ”VeryGood” Grade in the output level. Therefore, we choose to use the pruned decision tree.

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



Accuracy: **81.51%**.

### 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**



Accuracy: **96.64%**

**2. Very Good**



Accuracy: **91.60%**

**3. Good**



Accuracy: **79.83%**

**4. Sufficient**



Accuracy: **68.91%**

**5. Weak**



Accuracy: **79.83%**

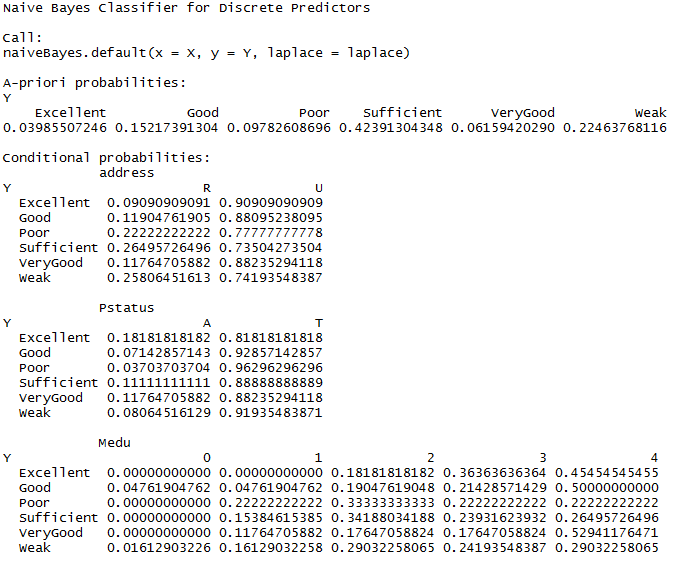
**6. Poor**



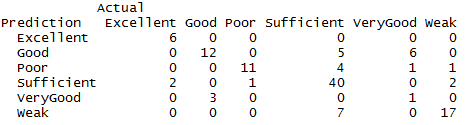
Accuracy: **89.08%**

### Naïve Bayes

The following is showing the result of Naive Bayes Classifier:

****

This is Confusion Matrix for using Naive Bayes method:



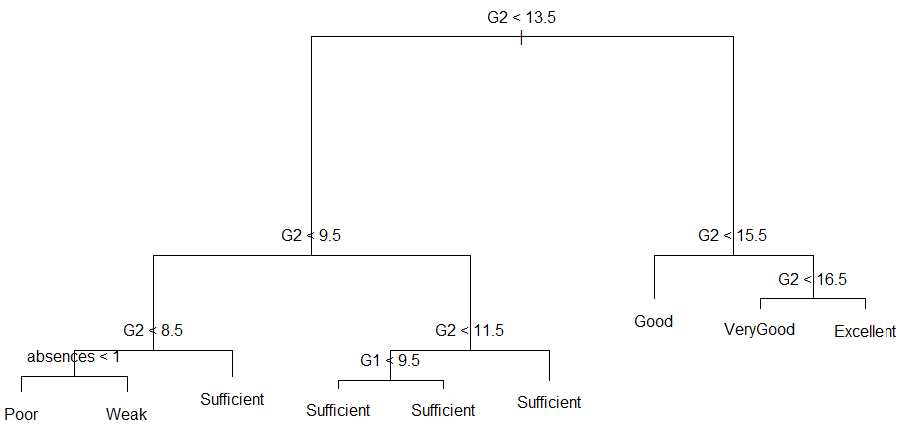
Accuracy: **73.11%**

### Portuguese Language:

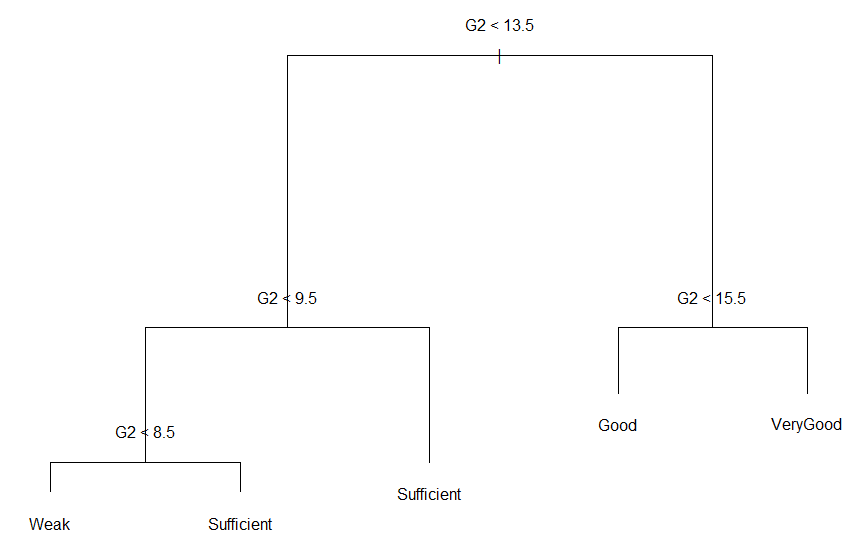
### 1. Decision Tree

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

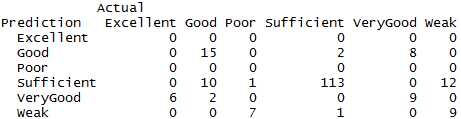
Before pruned:



After pruned:



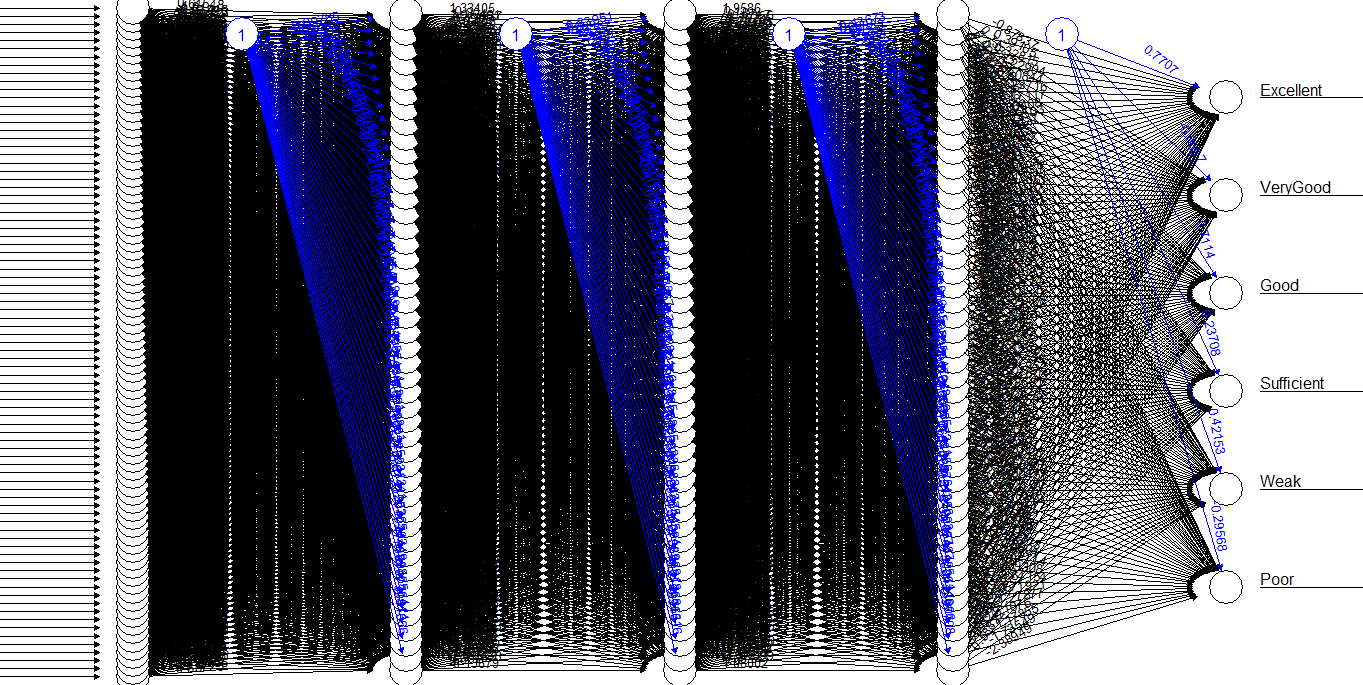
The Confusion Matrix for using Decision tree method:



Accuracy: **74.87%**

### Artificial Neural Network

Visualization of ANN for Grade:



1. **Excellent**

https://lh3.googleusercontent.com/8fBPGXmBVCSPkenf06UIjn-mf0__VCqTGf_PyKokl3yuKnGlqolptK8x3K7EhWXIzFt1_ilSB4gVkw7cVbDpNttcAO10UBBBO_Hce7myuvIEY3_Z1PD38JrfkJbRXEGd5HT8YK3G

**Accuracy: 94.87%**

1. **Very Good**

**https://lh3.googleusercontent.com/JWouu3m7Kj30DFSho-21m4EYpAFv7f9OUoZ0C0hNxJI0DMTu6s2rJ_BgCLKnsWqqvcoVm_XLFjVAApJ3bDGkrMLdmqB56tAO8Wq6tpdrVZZ2Zlchji39nkRN4UwCVk2hu7xQkemR**

**Accuracy: 85.64%**

1. **Good**

**https://lh3.googleusercontent.com/wYES89h484Kppv_LYLYtDtZFO86XXTq8cNux_0hCU1ddbjGbBkCjsgUg_44Ncqyz2Pxmm6LD0IOjsftEU1_vt0u6hqQ245vLSOjNXlAoQ7B-Bf1W0xnXbq90STzhEJ1UpFWAunWO**

**Accuracy: 75.90%**

1. **Sufficient**

**https://lh6.googleusercontent.com/yq3G8eZIz6zztLrgzIHwNEsVtC_p1T9sxfmgjmviN2bGnt1R9a2zGoWK-EnC3wBJr90yj0HPfqOSvaTUqQAUrlAoSRCcaXEyg0td5zka2Q9MAwt0uWElRbf3O6monEzHLZjGWCY2**

**Accuracy: 71.28%**

1. **Weak**

**https://lh5.googleusercontent.com/l95C56hd2awJX6WBcbcUSi6b4lLQaaF7QR5QAdF1Uh_HkBxOrTjLsat7Tg3JC2XQ0Ces5Ct5-IxTwxbBeaVKfEjYOMSOgyTl_N3vZStrsBOzimsdi-Pm40uYtV_E-JXduM36_Be9**

**Accuracy: 87.18%**

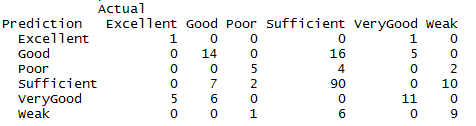
1. **Poor**

**https://lh5.googleusercontent.com/FYVFWdDBWZFuSN0-68NtuFfK_zOM53jxYrvbD39CAY5w0zor8fVVagvnhhCiJwVp6oJvPTJkZPT54aQxJ5Ds5-61QP3M3es02ymRW-tbyBWvuoi3O1TCYsu1N6XmhdXAv22mwmUm**

**Accuracy: 97.95%**

### Naïve Bayes

The following is showing the result of Naive Bayes Classifier.

****

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.