

**Faculty of Computing and Informatics**

**(FCI)**

TDS3301 DATA MINING

Assignment Part 3

Prepared by:

|  |  |
| --- | --- |
| Pay Eong Ian | 1122703150 |
| Tan Suk Mei | 1132700522 |
| Chew Yoke Teng | 1132701833 |
| Loh Yun Kuen | 1131122813 |

PART 3 CLASSIFICATION

For this assignment, we have chosen Student Performance dataset located at:

[https://archive.ics.uci.edu/ml/datasets/Student+Performance#](https://archive.ics.uci.edu/ml/datasets/Student+Performance)

There are 2 sets of dataset provided from the link, which are the students performance in Mathematics subject and Portuguese subject. Both dataset consist of 33 columns whereas the mathematics dataset contains 395 rows and portuguese dataset contains 649 rows. The rows represent the records of each students and columns represent the information of student and some attributes that are related to their performance in school such as study time, parents status, relationship status and so on.

We have decided to perform classification task to predict students ability to pass the mathematics subject based on certain variables and compare the accuracy performance of the classifiers.

1. Exploratory data analysis

Data exploration has been conducted to visualize and summarize the main characteristics of the dataset. The functions such as dim(), str(), names(), attributes(), summary() and others were been used in both dataset.

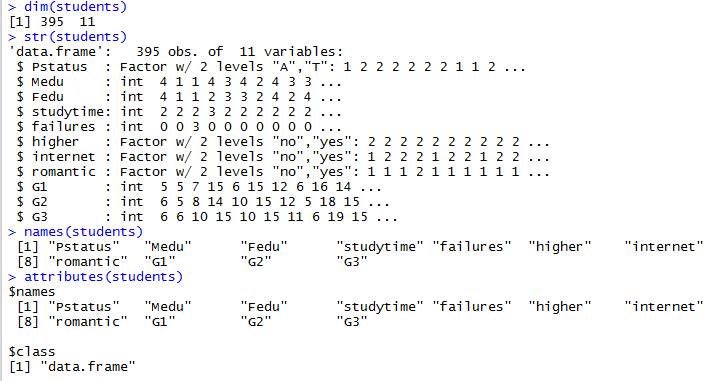


Figure above shows the structure of the Student Performance Mathematics dataset.

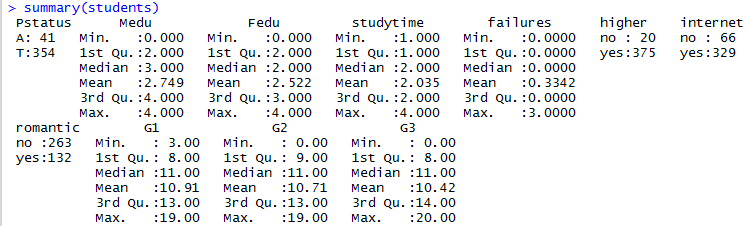


Figure above shows the summary of the Student Performance Mathematics dataset.

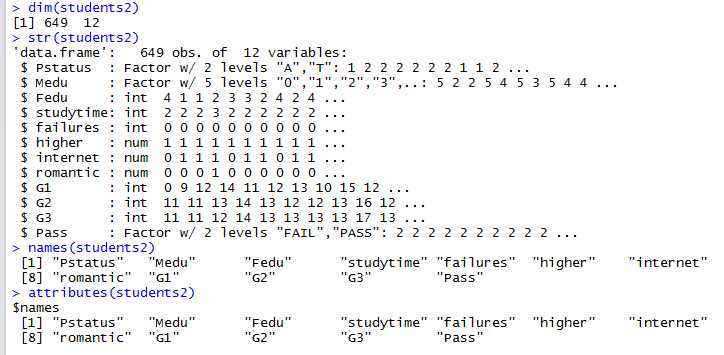


Figure above shows the structure of the Student Performance Portuguese dataset.

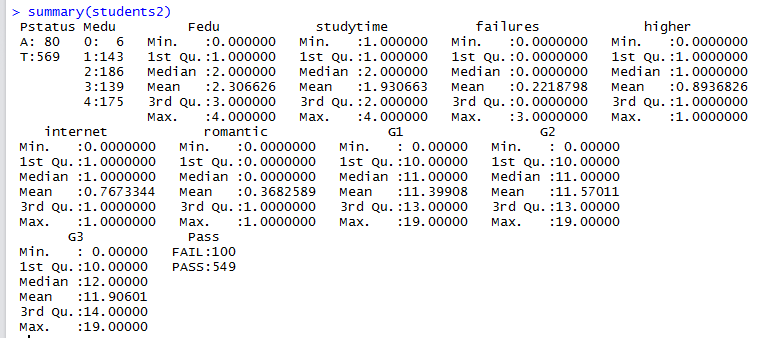


Figure above shows the summary of the Student Performance Portuguese dataset.

1. 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 pre-processing task we performed include removing unnecessary columns such as school, age, sex, family size, address, travel time to school and reason to choose this school. A new column ‘PASS’ is created by categorizing the data to pass or fail from class variable ‘G3’ as shown below.

https://lh6.googleusercontent.com/-vLbdsiK5LyE8IW2TJ8GtRyjXy-7rPJ8QpKZHNwzpHHoBK3tnNnusAn_B5ChHnplP_vRtWtMaWvxPRl-uHyOu1F0RoLfhiFL1K-RSWF_NRj1HFybt0z3uBTrkd7xeA4yH_5x-AqM

1. Choice of performance measures

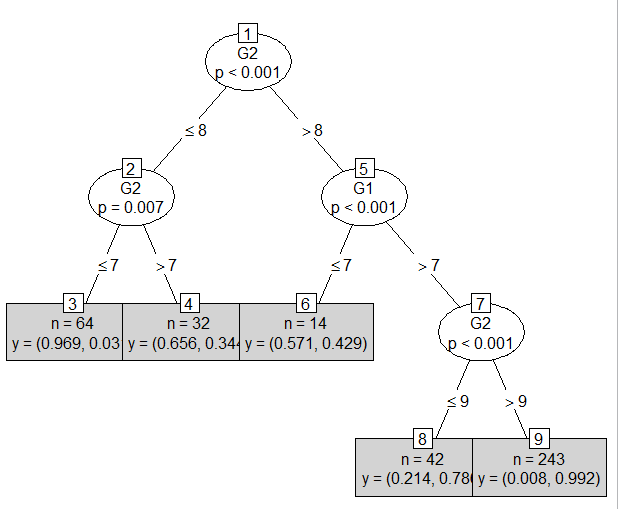
For compare the 3 classifiers, a confusion matrix needed to be create from simulated classification results. The chosen performance measures (based on confusion matrix) are accuracy, precision, recall and f1. Accuracy is defined as the fraction of instances that are correctly  classified. Precision is the fraction of correct predictions for a certain class, recall is the fraction of instances of a class that were correctly predicted. F1 is the weighted average of precision and recall.

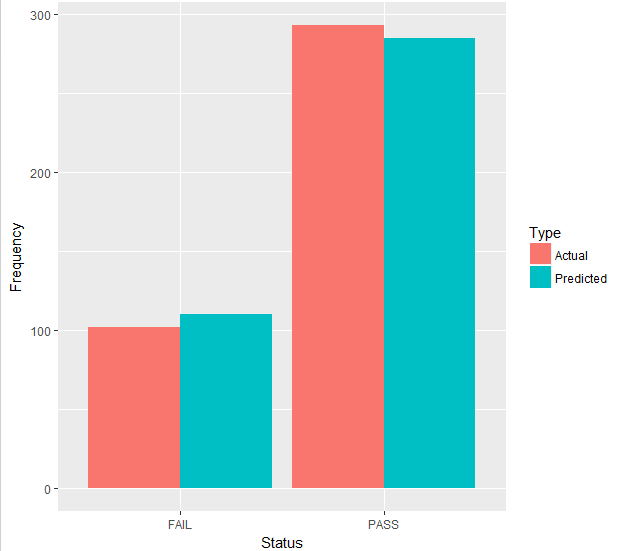
Please refer to .R for detail steps.

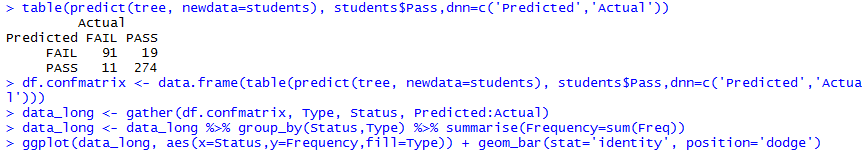
1. Performance of the 3 classifiers

Three classifiers have been used are decision tree, naive bayes and artificial neural network.

Decision Tree

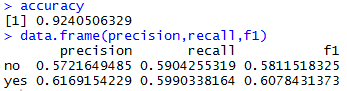






Decision tree is used on variables G1 and G2 together to predict students’ pass-ability.

The above figure had shown about the result of prediction of pass and fail of the students from the chosen dataset. It is clearly shown that the predicted fail rate is lesser than the actual fail rate, whereas the predicted pass rate is higher than actual pass rate.



Accuracy : 92%

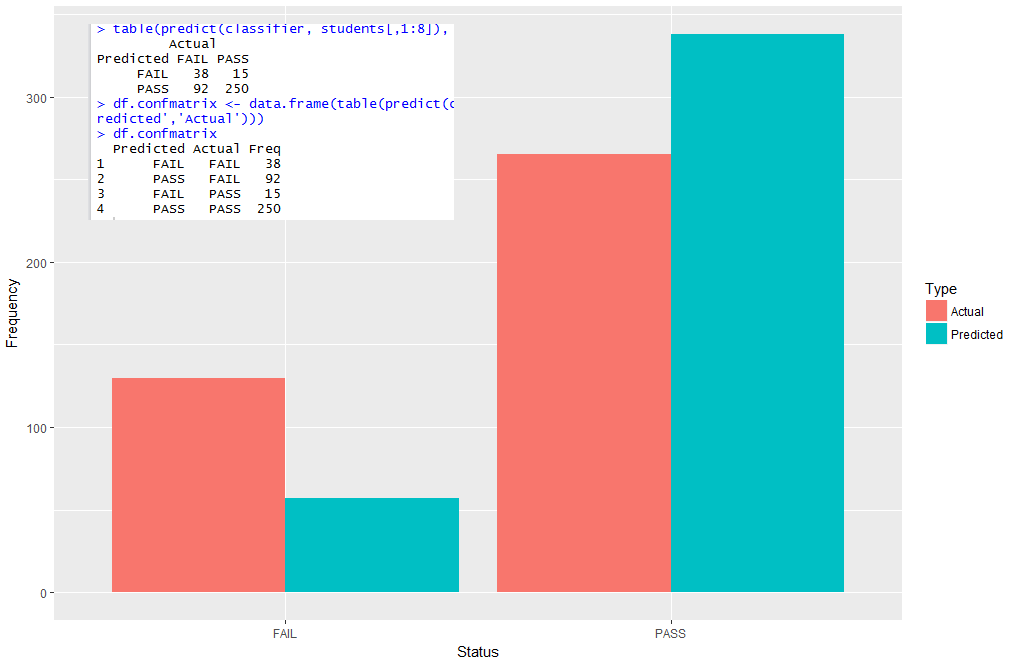
Precision : 57% Fail, 62% Pass

Recall : 59% Fail, 60% Pass

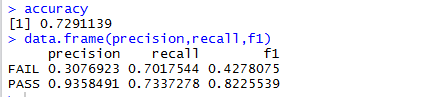
F1 : 58% Fail, 61% Pass

Naive Bayes

Naive Bayes is used on Parent Status, Education of mother and father, romantic status, study time, number of failures, higher chance to further study, and internet access in living place to predict the pass-ability, .The library (e1071) had been used for naive bayes function. Column no 1 to 8 and Pass were the classifier of the naive bayes function. After the function created, the next process is the prediction of pass and fail in the table and frame the data frame. It’s will create the confusion matrix.



The figure and confusion matrix of the Student Performance Mathematics dataset above had shown  the predicted pass and fail were not even close to the actual pass and fail. More actual fail and less actual pass than predicted pass and fail in mathematics subject.

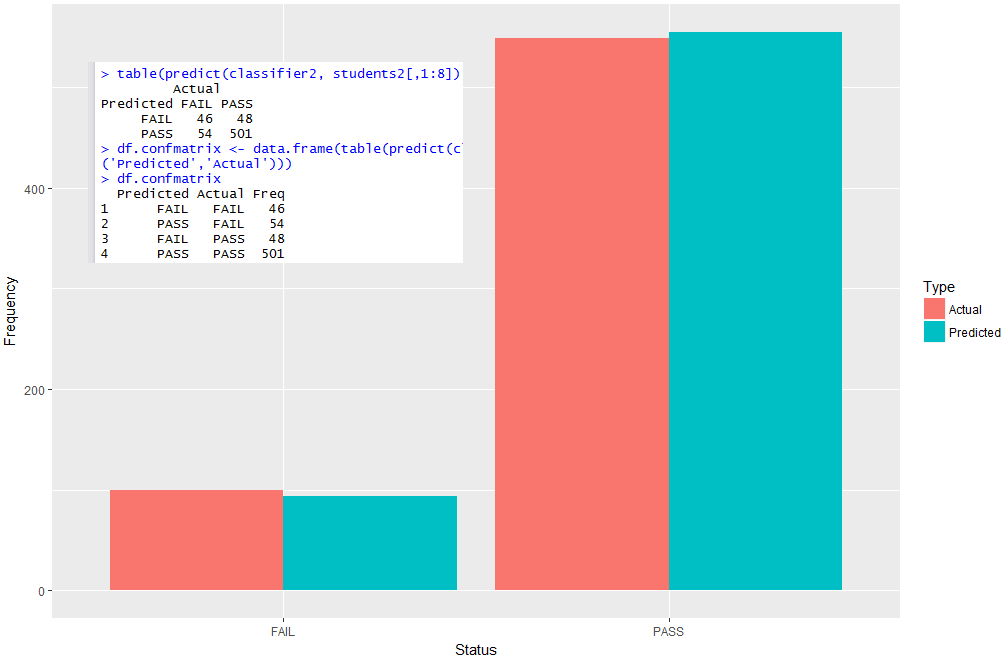


Accuracy: 73%

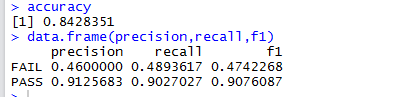
Precision: 31% Fail, 94% Pass

Recall: 70% Fail, 73% Pass

F1: 43% Fail, 82% Pass



The figure and confusion matrix of the Student Performance Portuguese dataset above had shown  the predicted pass and fail were close to the actual pass and fail. The actual fail is high a little than predicted fail, the actual pass is low a little than predicted pass.



Accuracy: 84%

Precision: 46% Fail, 91% Pass

Recall: 49% Fail, 90% Pass

F1: 47% Fail, 91% Pass

E. Suggestion as to why the classifiers behave differently.

Overall, decision tree prediction on the same dataset showed lesser errors on predictions making G1 and G2 suitable for predicting Grades and Pass-ability of students

Whereas, Naive Bayes method showed many errors on predictions. Naive Bayes method is not a good predicting model for this dataset.