MULTIMEDIA UNIVERSITY®

Bridging Boundaries, Connecting Minds

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#

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.

A. 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.

```
> dim(students)
[1] 395  11
> str(students)
'data.frame':   395 obs. of  11 variables:
 $ 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 ...
 $ studytime: int  2 2 2 3 2 2 2 2 2 2 ...
 $ failures : int  0 0 3 0 0 0 0 0 0 0 ...
 $ 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 ...
 $ 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 ...
> names(students)
 [1] "Pstatus"   "Medu"      "Fedu"      "studytime" "failures"  "higher"    "internet"
 [8] "romantic"  "G1"        "G2"        "G3"
> attributes(students)
$names
 [1] "Pstatus"   "Medu"      "Fedu"      "studytime" "failures"  "higher"    "internet"
 [8] "romantic"  "G1"        "G2"        "G3"

$class
[1] "data.frame"
```

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

```
> summary(students)
 Pstatus      Medu            Fedu          studytime        failures      higher    internet
 A: 41   Min.   :0.000   Min.   :0.000   Min.   :1.000   Min.   :0.0000   no : 20   no : 66
 T:354   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:0.0000   yes:375   yes:329
         Median :3.000   Median :2.000   Median :2.000   Median :0.0000
         Mean   :2.749   Mean   :2.522   Mean   :2.035   Mean   :0.3342
         3rd Qu.:4.000   3rd Qu.:3.000   3rd Qu.:2.000   3rd Qu.:0.0000
         Max.   :4.000   Max.   :4.000   Max.   :4.000   Max.   :3.0000
 romantic        G1              G2              G3
 no :263   Min.   : 3.00   Min.   : 0.00   Min.   : 0.00
 yes:132   1st Qu.: 8.00   1st Qu.: 9.00   1st Qu.: 8.00
           Median :11.00   Median :11.00   Median :11.00
           Mean   :10.91   Mean   :10.71   Mean   :10.42
           3rd Qu.:13.00   3rd Qu.:13.00   3rd Qu.:14.00
           Max.   :19.00   Max.   :19.00   Max.   :20.00
```

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

```
> dim(students2)
[1] 649  12
> str(students2)
'data.frame':   649 obs. of  12 variables:
 $ Pstatus  : Factor w/ 2 levels "A","T": 1 2 2 2 2 2 2 1 1 2 ...
 $ Medu     : Factor w/ 5 levels "0","1","2","3",..: 5 2 2 5 4 5 3 5 4 4 ...
 $ Fedu     : int  4 1 1 2 3 3 2 4 2 4 ...
 $ studytime: int  2 2 2 3 2 2 2 2 2 2 ...
 $ failures : int  0 0 0 0 0 0 0 0 0 0 ...
 $ higher   : num  1 1 1 1 1 1 1 1 1 1 ...
 $ internet : num  0 1 1 1 0 1 1 0 1 1 ...
 $ romantic : num  0 0 0 1 0 0 0 0 0 0 ...
 $ G1       : int  0 9 12 14 11 12 13 10 15 12 ...
 $ G2       : int  11 11 13 14 13 12 12 13 16 12 ...
 $ G3       : int  11 11 12 14 13 13 13 13 17 13 ...
 $ Pass     : Factor w/ 2 levels "FAIL","PASS": 2 2 2 2 2 2 2 2 2 2 ...
> names(students2)
 [1] "Pstatus"   "Medu"      "Fedu"      "studytime" "failures"  "higher"    "internet"
 [8] "romantic"  "G1"        "G2"        "G3"        "Pass"
> attributes(students2)
$names
 [1] "Pstatus"   "Medu"      "Fedu"      "studytime" "failures"  "higher"    "internet"
 [8] "romantic"  "G1"        "G2"        "G3"        "Pass"
```

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

```
> summary(students2)
 Pstatus Medu        Fedu           studytime        failures          higher
 A: 80   0:  6   Min.   :0.000000   Min.   :1.000000   Min.   :0.0000000   Min.   :0.0000000
 T:569   1:143   1st Qu.:1.000000   1st Qu.:1.000000   1st Qu.:0.0000000   1st Qu.:1.0000000
         2:186   Median :2.000000   Median :2.000000   Median :0.0000000   Median :1.0000000
         3:139   Mean   :2.306626   Mean   :1.930663   Mean   :0.2218798   Mean   :0.8936826
         4:175   3rd Qu.:3.000000   3rd Qu.:2.000000   3rd Qu.:0.0000000   3rd Qu.:1.0000000
                 Max.   :4.000000   Max.   :4.000000   Max.   :3.0000000   Max.   :1.0000000
    internet          romantic            G1              G2
 Min.   :0.0000000   Min.   :0.0000000   Min.   : 0.00000   Min.   : 0.00000
 1st Qu.:1.0000000   1st Qu.:0.0000000   1st Qu.:10.00000   1st Qu.:10.00000
 Median :1.0000000   Median :0.0000000   Median :11.00000   Median :11.00000
 Mean   :0.7673344   Mean   :0.3682589   Mean   :11.39908   Mean   :11.57011
 3rd Qu.:1.0000000   3rd Qu.:1.0000000   3rd Qu.:13.00000   3rd Qu.:13.00000
 Max.   :1.0000000   Max.   :1.0000000   Max.   :19.00000   Max.   :19.00000
       G3              Pass
 Min.   : 0.00000   FAIL:100
 1st Qu.:10.00000   PASS:549
 Median :12.00000
 Mean   :11.90601
 3rd Qu.:14.00000
 Max.   :19.00000
```

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

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 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.

Pass
FAIL
FAIL
PASS
PASS
PASS
PASS
PASS
FAIL
PASS
PASS

C. Choice of performance measures

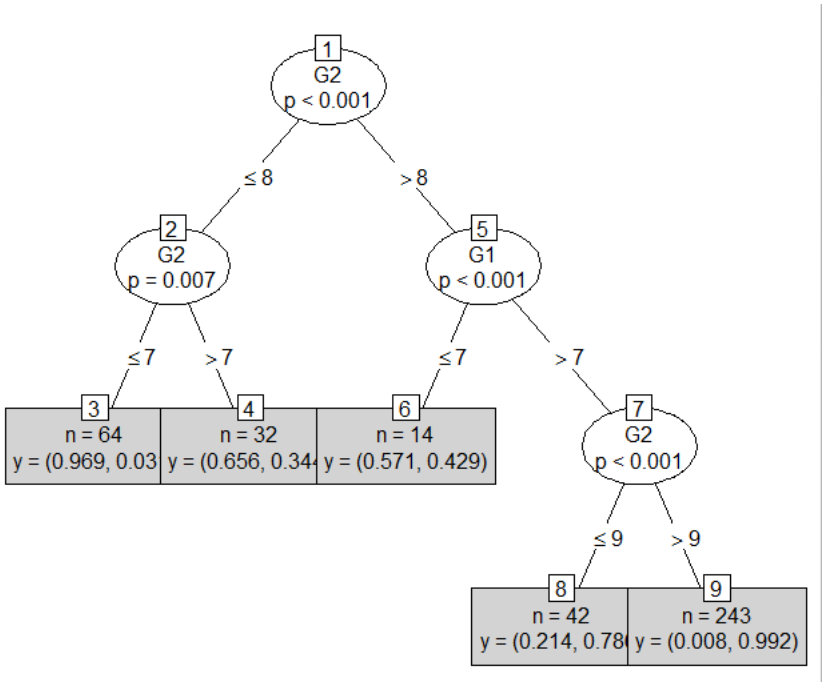
For compare the 3 classifiers, a confusion matrix needed to be create from simulated classification results. The chosen performance measures (based on confusion matrix) are accuracy, precision, recall and f1. Accuracy is defined as the fraction of instances that are correctly classified. Precision is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

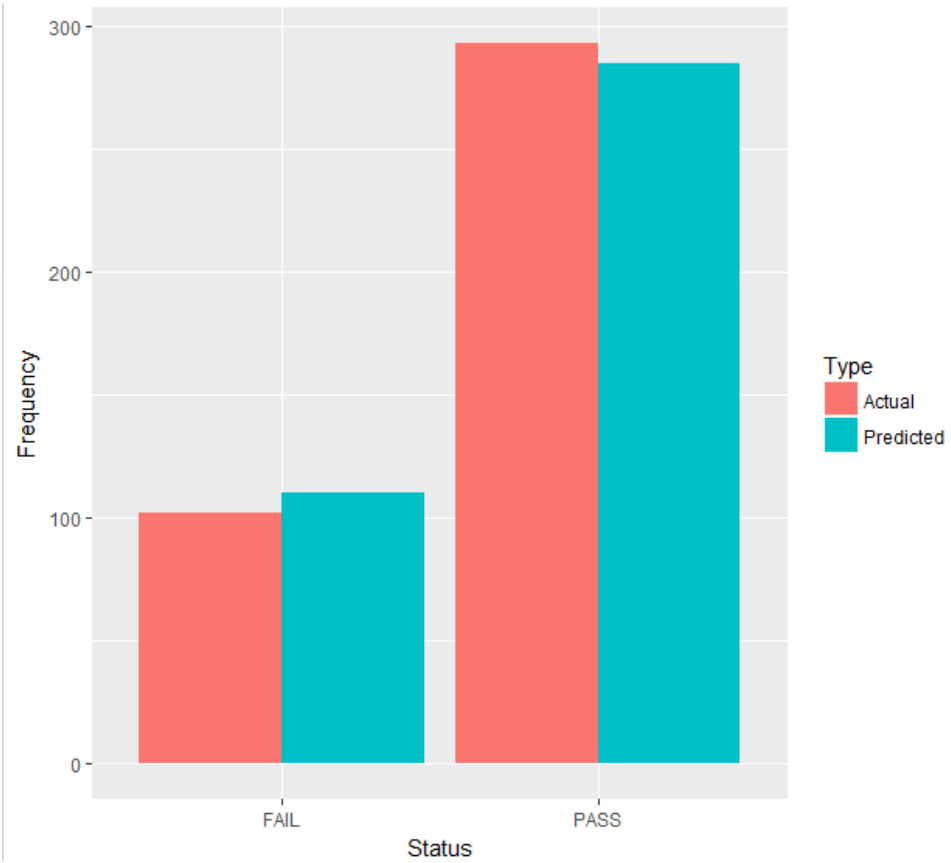
Please refer to .R for detail steps.

D. Performance of the 3 classifiers

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

Decision Tree






```
> table(predict(tree, newdata=students), students$Pass,dnn=c('Predicted','Actual'))
         Actual
Predicted FAIL PASS
     FAIL   91   19
     PASS   11  274
> df.confmatrix <- data.frame(table(predict(tree, newdata=students), students$Pass,dnn=c('Predicted','Actua
l')))
> data_long <- gather(df.confmatrix, Type, Status, Predicted:Actual)
> data_long <- data_long %>% group_by(Status,Type) %>% summarise(Frequency=sum(Freq))
> ggplot(data_long, aes(x=Status,y=Frequency,fill=Type)) + geom_bar(stat='identity', position='dodge')
```

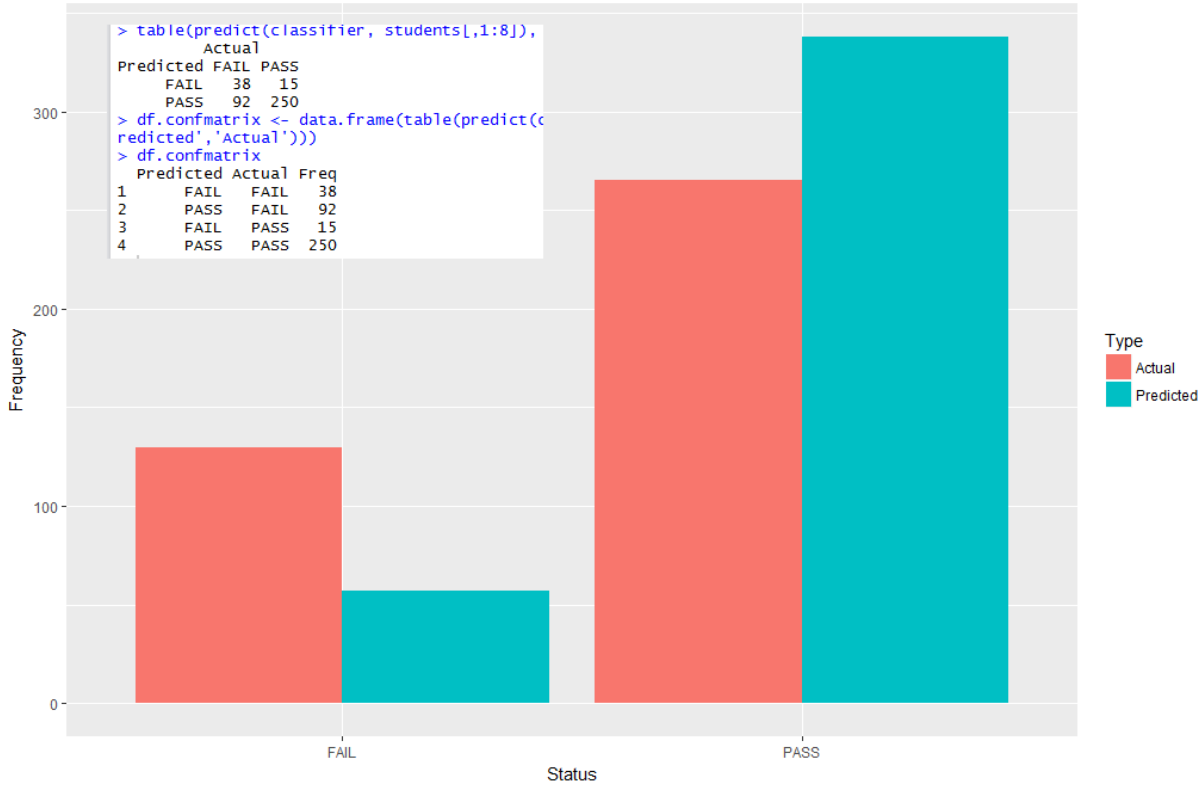
Decision tree is used on variables G1 and G2 together to predict students' pass-ability. The above figure had shown about the result of prediction of pass and fail of the students from the chosen dataset. It is clearly shown that the predicted fail rate is lesser than the actual fail rate, whereas the predicted pass rate is higher than actual pass rate.

```
> accuracy
[1] 0.9240506329
> data.frame(precision,recall,f1)
       precision       recall           f1
no  0.5721649485 0.5904255319 0.5811518325
yes 0.6169154229 0.5990338164 0.6078431373
```

Accuracy : 92%
Precision : 57% Fail, 62% Pass
Recall : 59% Fail, 60% Pass
F1 : 58% Fail, 61% Pass

Naive Bayes

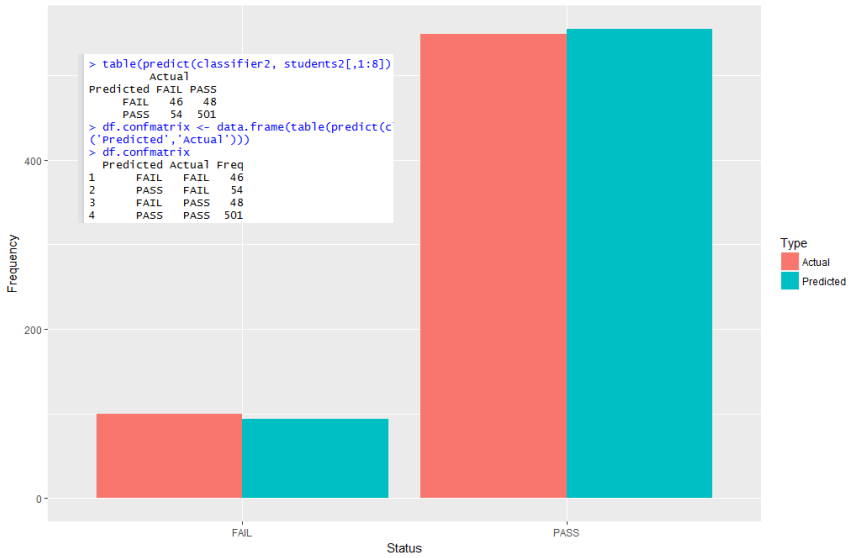
Naive Bayes is used on Parent Status, Education of mother and father, romantic status, study time, number of failures, higher chance to further study, and internet access in living place to predict the pass-ability, .The library (e1071) had been used for naive bayes function. Column no 1 to 8 and Pass 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
[1] 0.7291139
> data.frame(precision,recall,f1)
     precision    recall        f1
FAIL 0.3076923 0.7017544 0.4278075
PASS 0.9358491 0.7337278 0.8225539
```

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
[1] 0.8428351
> data.frame(precision,recall,f1)
     precision    recall        f1
FAIL 0.4600000 0.4893617 0.4742268
PASS 0.9125683 0.9027027 0.9076087
```

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.