TDS 3301 DATA MINING

ASSIGNMENT

PART 3

Student Performance Dataset

GROUP MEMBERS	STUDENT ID
LIYA SAFFURA	1132702377
AMIR RIDHWAN	1132701767
SURAYA IBRAHIM SHAH	1151303737
ILI FADHILAH AHMAD HIZZAD	1151303720

Contents

Introduction

In this assignment, we have used the Student Performance data. There are two sets of data, ‘student-mat.csv’ and ‘student-por.csv’, and we have chosen to perform classification tasks using ‘student-mat.csv’ and have read it into an R dataframe called ‘math’.

After performing exploratory data analysis, we have found that the ‘math’ dataset contains 395 rows and 33 variables. The rows indicate each student’s record and the columns are "school", "sex", "age", "address", "famsize", "Pstatus", "Medu", "Fedu", "Mjob", "Fjob", "reason", "guardian", "traveltime", "studytime", "failures", "schoolsup", "famsup", "paid", "activities", "nursery", "higher", "internet", "romantic", "famrel", "freetime", "goout", "Dalc", "Walc", "health", "absences", "G1", "G2" and "G3". The table below shows the data dictionary for the ‘math’ dataset.

Data Dictionary

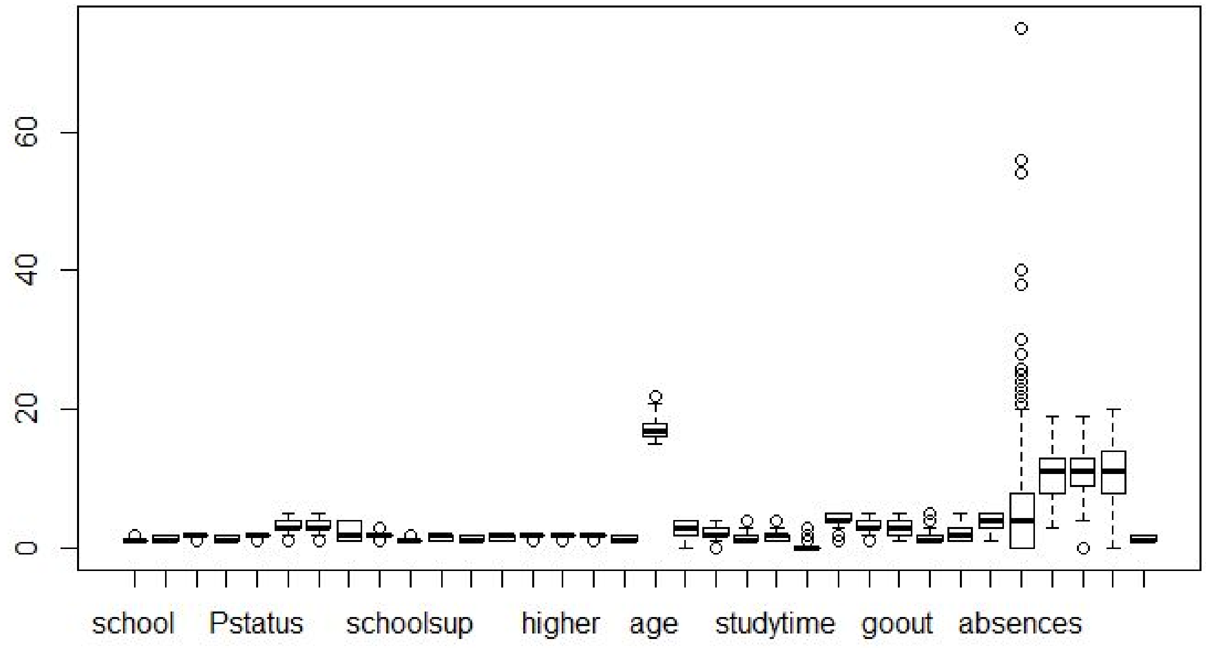
Column Name	**Description**	**Data Type**
school	Name of school	Binary: "GP" - Gabriel Pereira or "MS" - Mousinho da Silveira
sex	Gender of the student	Binary: "F" - female or "M" - male
age	Student’s age	Numeric
address	Student's home address type	Binary: "U" - urban or "R" - rural
famsize	Family size	Binary: "LE3" - less or equal to 3 or "GT3" - greater than 3
Pstatus	Parent's relationship status	Binary: "T" - living together or "A" - apart
Medu	Mother's education	Numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education
Fedu	Father’s education	Numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education
Fjob	Father’s occupation	Nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other"
Mjob	Mother’s occupation	Nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other"

reason	Reason to choose this school	Nominal: close to "home", school "reputation", "course" preference or "other"
guardian	Student's guardian	Nominal: "mother", "father" or "other"
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
studytime	Student’s weekly study time	Numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours
failures	Number of past class failures	Numeric: n if 1<=n<3, else 4
schoolsup	Extra educational support	Binary: yes or no
famsup	Family educational support	Binary: yes or no
paid	Extra paid classes	Binary: yes or no
activities	Extra-curricular activities	Binary: yes or no
nursery	Attended nursery school	Binary: yes or no
higher	Wants to take higher education	Binary: yes or no
internet	Internet access at home	Binary: yes or no
romantic	In a romantic relationship	Binary: yes or no
famrel	Quality of family relationships	Numeric: from 1 - very bad to 5 - excellent
freetime	Free time after school	Numeric: from 1 - very low to 5 - very high
goout	Going out with friends	Numeric: from 1 - very low to 5 - very high
Dalc	Workday alcohol consumption	Numeric: from 1 - very low to 5 - very high
Walc	Weekend alcohol consumption	Numeric: from 1 - very low to 5 - very high

health	Current health status	Numeric: from 1 - very bad to 5 - very good
absences	Number of school absences	Numeric: from 0 to 93
G1	First period grade	Numeric: from 0-20
G2	Second period grade	Numeric: from 0-20
G3	Final period grade	Numeric: from 0-20

Preprocessing Tasks

The dataset has no missing values and no duplicate records. The only outlier in the dataset is in "absences" column which has the value 75 while the median is 4.



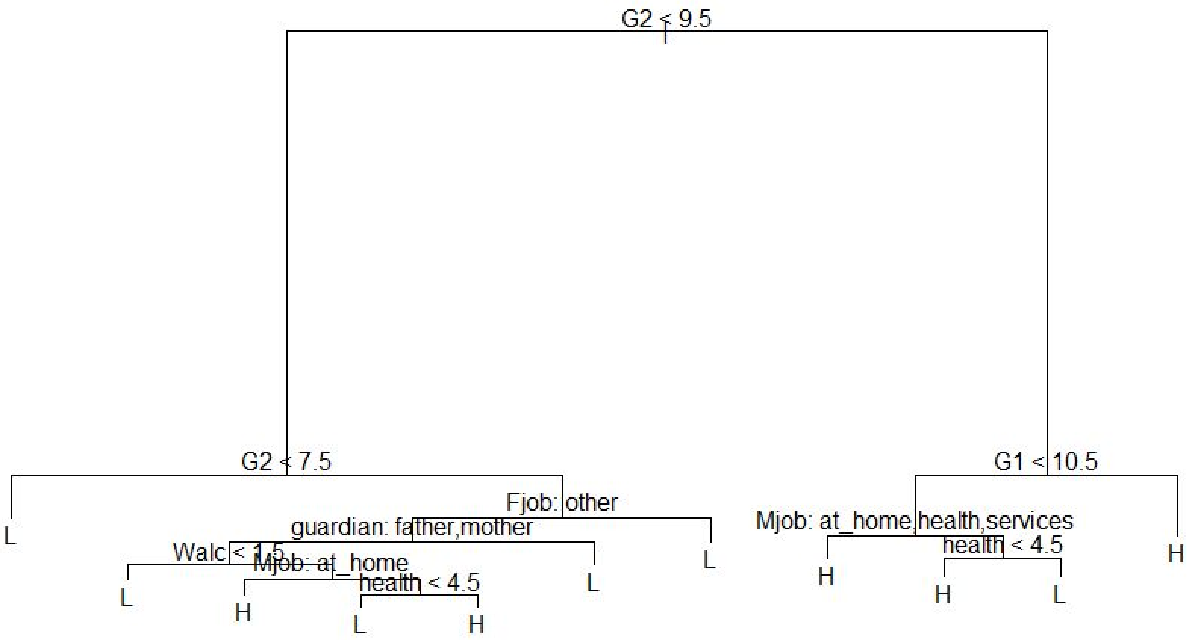
Boxplot of dataframe 'math'

Data cleaning was not done on the dataset as it has no missing values and duplicate records, but the dataset's column is rearranged for easy processing. A new column has been created as a class variable, named 'achievement'. The value of achievement is labeled H (High) if the value of G3 is above 10, and L (Low) if the value of G3 is less than 10. New vectors have been created from the columns in the 'math' dataframe that contains factors of two levels. They are "schoolsup", "famsup", "paid", "activities", "nursery", "higher", "internet", "romantic", and "achievement". We have normalized the values by converting them to 1 and 0. We have also created scaled.math, which contains scaled data from columns "age", "Medu", "Fedu", "traveltime", "studytime", "failures", "famrel", "freetime", "goout", "Dalc", "Walc", "health", "absences", "G1", "G2", "G3". After scaling these data, scaled.math was combined with the other variables that were converted to 1 and 0. This data is called ‘data’.

Classification Tasks

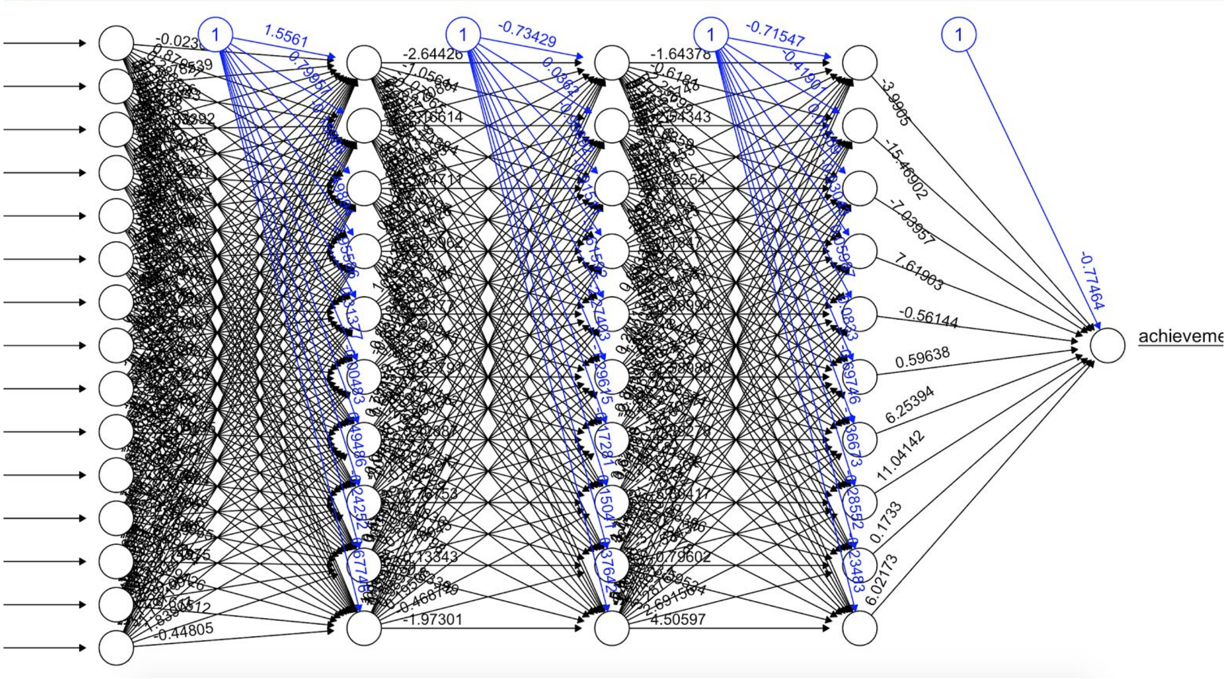
A. Decision Tree

A training is created using the math dataset which will be used to run with the test set created. The test set will produced the variable achievement's result. The decision tree is the made by classifying the results using the training set data. From the tree, achievement on math.test is predicted.



B. Neural Network

To perform neural network task, we partitioned the dataset into 70% training and 30% testing, based on split boolean vector. Then we created a formula to be used in the neural network function. The neural network function from neuralnet library then perform the neural network algorithm function.

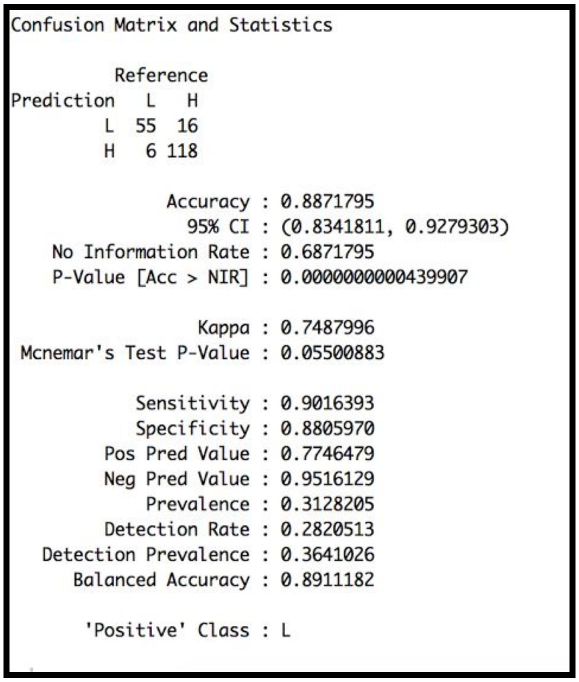


C. Naive Bayes

A random subset from ‘data’ is created to be used as the training set. Then the test set is obtained by taking the rows which are not in training set. The naiveBayes function was used on the training set, data.train to obtain the Naive Bayes model. The model is then used on data.test to obtain the conditional probabilities.

Performance Measures of Decision Tree, Naive Bayes and Neural Network

A. Decision tree

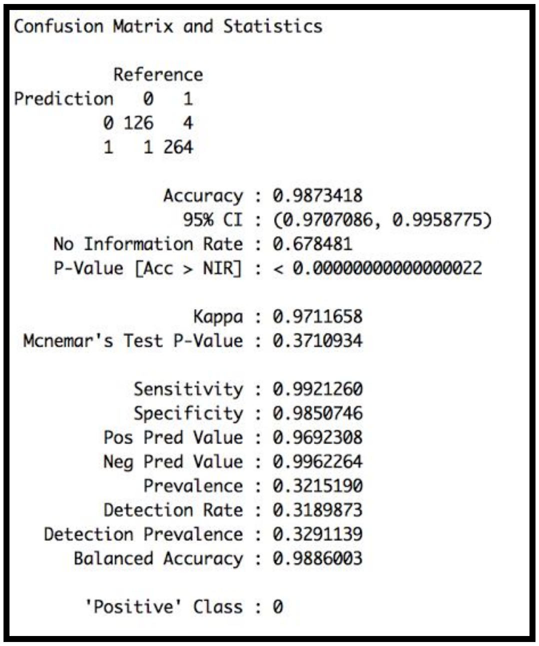


```
Confusion Matrix and Statistics

          Reference
Prediction   L   H
         L  55  16
         H   6 118

               Accuracy : 0.8871795
                 95% CI : (0.8341811, 0.9279303)
    No Information Rate : 0.6871795
    P-Value [Acc > NIR] : 0.0000000000439907

                  Kappa : 0.7487996
 Mcnemar's Test P-Value : 0.05500883

            Sensitivity : 0.9016393
            Specificity : 0.8805970
         Pos Pred Value : 0.7746479
         Neg Pred Value : 0.9516129
             Prevalence : 0.3128205
         Detection Rate : 0.2820513
   Detection Prevalence : 0.3641026
      Balanced Accuracy : 0.8911182

       'Positive' Class : L
```

The confusion matrix shows that there are 55 True Positives and 118 True Negatives.

Performance measures:
Accuracy: 0.8871795
TPR (True Positive Rate) / Sensitivity: 0.9016393
FPR (False Positive Rate) / Specificity: 0.8805970
Misclassification Rate: 0.035

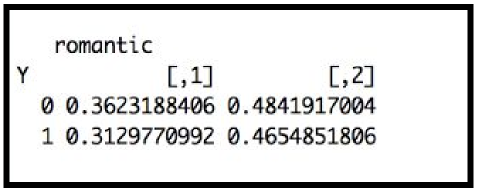
B. Neural Network



```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 126   4
         1   1 264

               Accuracy : 0.9873418
                 95% CI : (0.9707086, 0.9958775)
    No Information Rate : 0.678481
    P-Value [Acc > NIR] : < 0.0000000000000022

                  Kappa : 0.9711658
 Mcnemar's Test P-Value : 0.3710934

            Sensitivity : 0.9921260
            Specificity : 0.9850746
         Pos Pred Value : 0.9692308
         Neg Pred Value : 0.9962264
             Prevalence : 0.3215190
         Detection Rate : 0.3189873
   Detection Prevalence : 0.3291139
      Balanced Accuracy : 0.9886003

       'Positive' Class : 0
```

The confusion matrix shows that there are 126 True Positives and 264 True Negatives.
Performance measures:
Accuracy: 0.9873418
TPR (True Positive Rate) / Sensitivity: 0.9921260
FPR (False Positive Rate) / Specificity: 0.9850746

C. Naive Bayes



```
   romantic
Y           [,1]         [,2]
  0 0.3623188406 0.4841917004
  1 0.3129770992 0.4654851806
```

Naive Bayes model gives us the conditional probability, where the condition is either ‘yes’ or ‘no’. In the image above, it shows that if the value is 0, then the probability for column 1 is 0.3623188406 and for column 2 is 0.4841917004.

Suggestions on Why the Classifiers Behave Differently

Decision tree, Naive Bayes and Artificial Neural Network are behaving differently because they are structurally, functionally, or philosophically different. Decision tree provides us with a flowchart describing how we should classify an observation. We start at the root of the tree, and the leaf where we end up determines the classification we predict. For Naive Bayes the graph represents the conditional dependencies of different variables in the model. Each node represents a variable, and each directed edge represents a conditional relationship. Essentially, the graphical model is a visualization of the chain rule. Lastly in Artificial Neural Network, each node is a simulated "neuron". The neuron is essentially on or off, and its activation is determined by a linear combination of the values of each output in the preceding "layer" of the network.

References

1. ‘Student-mat.csv’ dataset
    - https://archive.ics.uci.edu/ml/datasets/Student+Performance