Comparison of Decision Tree and Random Forest Algorithms in Student Performance in Exam

Davina Febryanthi Kurniawan
Information Systems, Faculty of Engineering and Informatics, Multimedia Nusantara University, Tangerang, Indonesia

davina.kurniawan@student.umn.ac.id

***Abstract*— Students' academic performance is a key feature in education. This study was carried out to assess factors affecting students' academic performance in one of Senior High School in USA. A hundred and students from this school were randomly selected for this study. All the Exams are a part of growth for the students; they are essential for knowledge & capability testing. Exams improve the student's overall personality, memory, and their revision skills. This study aims to examine the relationship between student performance and several factors which include gender race or ethnicity, parental level of education, lunch, test preparation course, math score, reading score, and writing score.**

***Index Terms*—Exam; Student; Performance; Factor, Affecting; Academics; Education.**

I. INTRODUCTION

Education is one of the most important aspects of human resource development. The students' performance plays an important role in producing best quality graduates who will become great leaders and manpower for the country thus responsible for the country's economic and social development [1]. Academic achievement is one of the major factors considered by employers in hiring workers especially for the fresh graduates. Thus, students have to put the greatest effort in their study to obtain good grades and to prepare themselves for future opportunities in their career at the same time to fulfil the employer's demand.

Students are the main asset for various school. Academic performance achievement is the level of achievement of the student's educational goal that can be measured and tested through examination, assessment and other form of measurements. Student performance is measured using grade average. A student's GPA is typically measured on a scale of zero to four with higher GPAs representing higher grades in the classroom. However, the academic performance achievement varies as different kind of student may have different level of performance achievement [2].

Every student has some responsibilities towards the institutions they are studying in. They are given duties to perform, which they are expected to discharge efficiently. Students play an active role in classrooms. Apart from the classroom or traditional learning, the world of education and learning is changing rapidly. Students' academic performance is affected by several factors which include students' gender race or ethnicity, parental level of education, lunch, test preparation course, math score, reading score, writing score.

The purpose of this research is to examine the relationship between several factor (gender race or ethnicity, parental level of education, lunch, test preparation course, math score, reading score, writing score). the standard examinations (typical True/False and multiple-choice questions) and students' performance for high school students in a USA with decision tree and random forest algorithm. However, the specific objectives are to:

1. Asses the general performance of high school students in USA.
2. Determine the students' gender race or ethnicity, parental level of education, lunch, test preparation course, math score, reading score, and writing score factors that related to their academic performance.

In addition to the above, the following hypothesis was tested:

Null Hypothesis

1. There is no significant relationship between level gender race or ethnicity and their score in academic performance.
2. There is no significant relationship between level gender race or ethnicity and their score in academic performance.
3. There is no significant relationship between level of their lunch and an improvement in academic performance.
4. There is no significant relationship between level of test preparation course and an improvement in academic performance.
5. There is no significant relationship between math score, reading score, and writing score

and an improvement in academic performance.

II. LITERATURE REVIEW

A. *Students' Academic Performance*

Academic achievement (i.e., GPA or grades) is one tool to measure students' academic performance. Based on the Center for Research and Development Academic Achievement [3], academic achievement is a construct to measure students' achievement, knowledge and skills. This measurement is holistically based on the students' age, the students' previous experience, and the students' capacity related to social and education skills. To measure academic achievement, educators use different types of assessment [4]. Assessment is a continuous process that brings some valuable information about the learning process (Linn and Gronlund, 1995). Hargis (2003) commented that the grading process is supposed to be motivating and provide goals. On the other hand, grades can provide incentives to the students to cheat. Grading has the additional benefit of provide records (data sets) of students' academic achievements.

B. *Factors Contributing to Academic Performance*

Several studies have been conducted in different countries to assess the factors which contribute to academic performance of students at different levels. In Pakistan, Farooq and Berhanu (2011) found that parents' education and socio-economic status have significant effect on a student's academic performance in Mathematics and English Language. Sibanda, Iwu and Olumide (2015) found that, regular study, punctuality in school and self-motivation are the key determining factors which influence students' academic performance in South Africa. Ali, Munir, Khan and Ahmed (2013) also found that daily study hours, parent's socio-economic status and age have a significant impact on academic performance.

According to Khan, Iqbal and Tasneem (2015) parents with higher level of education show much interest in the academic performance of their wards. They observed there is a positive significant relationship between the level of parents' education and students' academic performance. The same result was found by Muthoni (2013) in Kenya. She noticed that in Kenya Secondary schools, the level of education of a student parent is positively related to his/her performance. Similarly, Ogbugo-Ololube (2016), found that parents level of education has a positive relationship with academic performance. It was also observed by Ntitika (2014) that parents with higher level of education serve as a motivation for their children to work hard to achieve their academic goals. He added that such students have higher aspirations for their education. He found that parent's level of education has some level of impact on their wards academic performance. On the other hand, Amuda and Ali (2016) found that parent's level of education has no statistical impact on their wards academic performance.

The relationship between gender and academic performance have been researched extensively for the past decade (Eitle, 2005 as cited in Farooq & Berhanu, 2011). According to Ghazvini and Khajehpour (2011) there is a difference between the cognitive levels of boys and girls. They noticed that the learning task of girls is more adaptive than boys. Omwirhiren and Anderson (2016) indicated that there is a statistical significant difference between the academic performance of males and females in Chemistry. They concluded that boys performed better than girls. Farooq and Berhanu (2011) on the other hand found that girls generally perform better than male students. Similarly, Nnamani and Oyibe (2016) and Jayanth et al. (2014) found that gender has a significant impact on academic performance.

C. *Decision Tree Algorithm*

Initially, the Decision Tree formation was used with ID3. Then it was modified because ID3 can only process discrete attribute values and can only be processed in small amounts. The C4.5 algorithm results from a further modification of ID3 and can process continuous and discrete attribute values. While the C5.0 algorithm is a modified version of C4.5, an extension of ID3 that can process large amounts of data, handle *missing values,* has better speed, and is efficient[4].

Decision Tree is one such modern solution to the decision making problems by learning the data from the problem domain and building a model which can be used for prediction supported by the systematic analytics [5], it consists of internal nodes and leaf nodes, used to put the values and check the values/attributes. Its classifier consists of two steps; one is preparing phase. Other is prune phase. The pruning phase helps to reduce the data and fit it in the decision tree.

The major advantage of using decision trees is that they are intuitively very easy to explain. They closely mirror human decision-making compared to other regression and classification approaches. They can be displayed graphically, and they can easily handle qualitative predictors without the need to create dummy variables. However, decision trees generally do not have the same level of predictive accuracy as other approaches, since they aren't quite robust. A small change in the data can cause a large change in the final estimated tree [6].

D. *Random Forest Algortihm*

A random forest is a machine learning technique that's used to solve regression and classification problems [7]. It utilizes ensemble learning, which is a technique that combines many classifiers to provide solutions to complex problems. The term "Random Forest Classifier" refers to the classification algorithm

made up of several decision trees. The algorithm uses randomness to build each individual tree to promote uncorrelated forests, which then uses the forest's predictive powers to make accurate decisions [8].

One of the biggest advantages of random forest is its versatility. It can be used for both regression and classification tasks, and it's also easy to view the relative importance it assigns to the input features [9]. Random forest is also a very handy algorithm because the default hyper parameters it uses often produce a good prediction result. Understanding the hyper parameters is pretty straightforward, and there's also not that many of them.

One of the biggest problems in machine learning is overfitting, but most of the time this won't happen thanks to the random forest classifier. If there are enough trees in the forest, the classifier won't overfit the model [10]. Another is the drawback of the Random Forest model is the instability of the existing accuracy results. This is because the input parameters and data are carried out more than once in a row, resulting in different accuracy values.

III. METHODOLOGY

A. *Object of Research*

The object of research is a variable or thing that is the target of a study. This study aims to to examine the relationship between several factor (gender race or ethnicity, parental level of education, lunch, test preparation course, math score, reading score, writing score) and student performance in senior high school in USA.

B. *Method of collecting data*

Data collection is the process of gathering and measuring information on variables of interest, in an established systematic fashion that enables one to answer stated research questions, test hypotheses, and evaluate outcomes. In this study, the data was obtained from the *Kaggle.com* website. The dataset that we used by can be accessed through the Kaggle website with the link address https://www.kaggle.com/spscientist/students-performance-in-exams.

Dataset Description

Data Collection: We got our dataset of Secondary schools. The dataset was collected using the school reported data in which student's social background data is present.

Data Preprocessing: Since the data was clean, with null values or attributes, we skipped the data cleaning phase and started preprocessing the data. In the Data Preprocessing phase, we added the pass or fail column and assigned values to it depending on the total scores, and we transformed it into numerical values. We concluded on Pass/Fail as our dependent variable and all other variables as the predictor variable.

- The dataset has 1000 rows and 8 columns
- There 5 categorical columns and 3 numerical columns
- The dataset has no null or duplicate values

Categorical columns are:

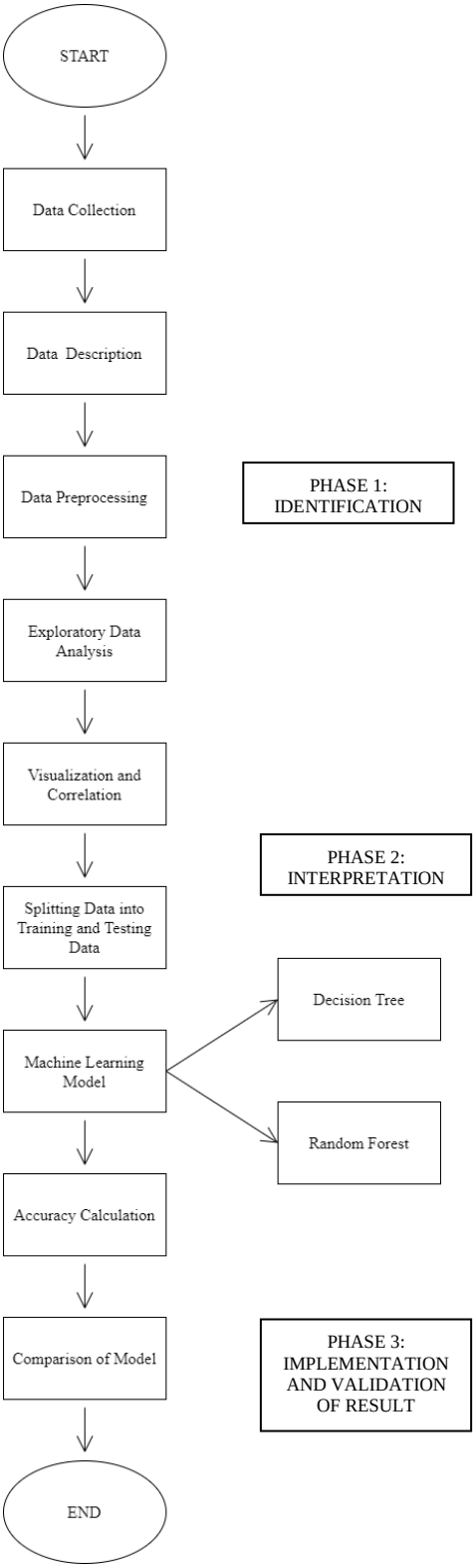
- Gender: Male or Female
- Race/ethnicity: 5 groups, from group A to group E
- Parental level of education: from high school to a master's degree
- Lunch: free/reduced or standard.

Numerical columns are:

- Math score: out of 100
- Reading score: out of 100
- Writing score: out of 100

The dataset contains the data of about 1000 students from the USA. This analysis aims to understand the influence of important factors such as parental level of education, the status of test preparation course etc. on the performance of the students in the exams.

C. Framework



D. Identification Phase

1. Load the library package that will be used and read the data file that has been prepared using the "read_csv ()" function, then save the read data file with the name "mydata".
2. Display the data structure with the "str()" function.
3. Create the data frame with "as.data.frame()" function.
4. Write a function to replace column names with "colnames()", create a variable names "namesofColumns" consists "Gender", "Race", "Parent_Education", "Lunch", "Test_Prep", "Math_Score", "Reading_Score", "Writing_Score", and save the "colnames()" into the variable that have been made previously.
5. For convenience, we create a table for 5 variable (Gender, Race, Parent_Education, Lunch, Test_Prep) by using "table()".
6. Create new column to determine student who pass the exam and assume the pass mark of each subject is 70/100.
7. Data Exploration

E. Interpretation and Testing Phase

1. Perform *splitting* data with "setseed()" and split 80% of data as *training set* and 20% as *testing set*.
2. Making predictions using Decision Tree with party and rpart models using a testing dataset from the training results that have been made previously.
3. Make predictions using the Random Forest algorithm to see the Out of Bag error rate estimation.

F. Implementation and Validation of Results Phase

After the interpretation and testing phase, we create a confusion matrix from the predict function made on both Decision Tree models with the rpart model and the Random Forest model. Then from the results of the confusion matrix accuracy that has been obtained, a comparison between the two models will be made to see which model is the best and can be used in real-life predictions.

IV. Result and Discussion

A. *Identification Phase*

```
> library(readr)
> library(ggplot2)
> library(tidyverse)
> library(dplyr)
> library(Amelia)
> library(knitr)
> library(psych)
> library(ROCR)
> library(caret)
> library(rpart)
> library(rpart.plot)
> library(randomForest)
```

Picture 4.1 Load *library package*

```
> str(mydata)
spec_tbl_df [1,000 x 8] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ gender                    : chr [1:1000] "female" "female" "female" "male" ...
 $ race/ethnicity            : chr [1:1000] "group B" "group C" "group B" "group A" ...
 $ parental level of education: chr [1:1000] "bachelor's degree" "some college" "master's degree"
"associate's degree" ...
 $ lunch                     : chr [1:1000] "standard" "standard" "standard" "free/reduced" ...
 $ test preparation course   : chr [1:1000] "none" "completed" "none" "none" ...
 $ math score                : num [1:1000] 72 69 90 47 76 71 88 40 64 38 ...
 $ reading score             : num [1:1000] 72 90 95 57 78 83 95 43 64 60 ...
 $ writing score             : num [1:1000] 74 88 93 44 75 78 92 39 67 50 ...
 - attr(*, "spec")=
  .. cols(
  ..   gender = col_character(),
  ..   `race/ethnicity` = col_character(),
  ..   `parental level of education` = col_character(),
  ..   lunch = col_character(),
  ..   `test preparation course` = col_character(),
  ..   `math score` = col_double(),
  ..   `reading score` = col_double(),
  ..   `writing score` = col_double()
  .. )
> #membuat data frame
> mydata <- as.data.frame(mydata)
> summary(mydata)
    gender          race/ethnicity     parental level of education
 Length:1000        Length:1000        Length:1000
 Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character


    lunch          test preparation course   math score
 Length:1000        Length:1000             Min.   :  0.00
 Class :character   Class :character        1st Qu.: 57.00
 Mode  :character   Mode  :character        Median : 66.00
                                            Mean   : 66.09
                                            3rd Qu.: 77.00
                                            Max.   :100.00
 reading score    writing score
 Min.   : 17.00   Min.   : 10.00
 1st Qu.: 59.00   1st Qu.: 57.75
 Median : 70.00   Median : 69.00
 Mean   : 69.17   Mean   : 68.05
 3rd Qu.: 79.00   3rd Qu.: 79.00
 Max.   :100.00   Max.   :100.00
```

Picture 4.2 Data Structure and Summary

```
> ## Replace column names
> colnames(mydata)
[1] "gender"                    "race/ethnicity"
[3] "parental level of education" "lunch"
[5] "test preparation course"     "math score"
[7] "reading score"               "writing score"
> namesOfColumns <- c("Gender","Race","Parent_Education","Lunch","Test_Prep","Math_Score","Reading_Score","Writing_Score")
> colnames(mydata) <- namesOfColumns
> table(mydata$Gender)
```

Picture 4.3 Replace column names

```
> table(mydata$Gender)

female   male
   518    482
> table(mydata$Race)

group A group B group C group D group E
     89     190     319     262     140
> table(mydata$Parent_Education)

associate's degree  bachelor's degree        high school    master's degree
               222                118                196                 59
      some college   some high school
               226                179
> table(mydata$Lunch)

free/reduced     standard
         355          645
> table(mydata$Test_Prep)

completed      none
      358       642
```

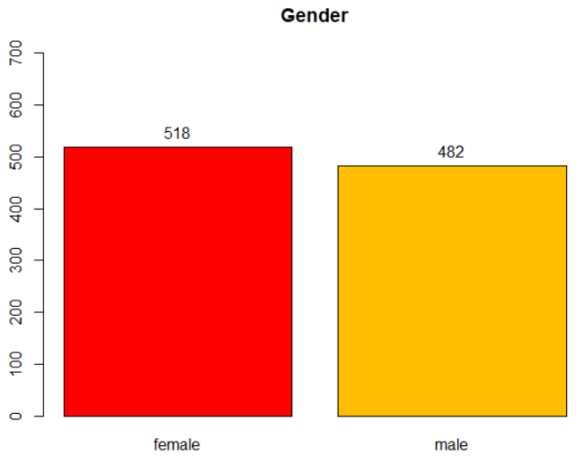
Picture 4.4 Tables

```
> new <- rep(1, nrow(mydata))
> mydata[ , ncol(mydata) + 1] <- new
> colnames(mydata)[which(names(mydata) == "...9")] <- "pass"
> for (i in 1:nrow(mydata)){
+   if((mydata$`Math_Score`[i]+ mydata$`Reading_Score`[i] + mydata$`Writing_Score`[i])/3 >= 70) {
+     mydata$pass[i] <-"pass"
+   }
+   else {
+     mydata$pass[i] <-"failed"
+   }
+ }
> View(mydata)
```

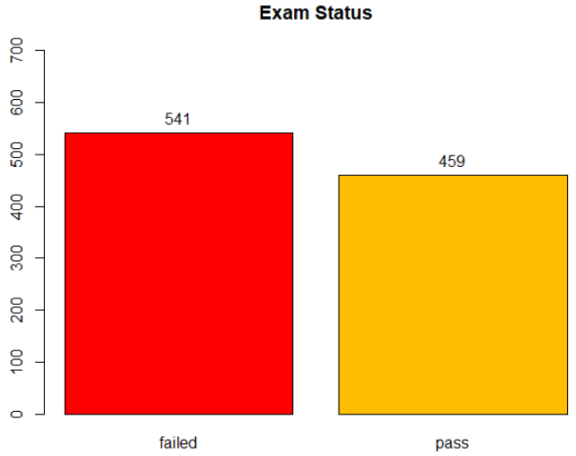
	Gender	Race	Parent_Education	Lunch	Test_Prep	Math_Score	Reading_Score	Writing_Score	V9	pass
1	female	group B	bachelor's degree	standard	none	72	72	74	1	pass
2	female	group C	some college	standard	completed	69	90	88	1	pass
3	female	group B	master's degree	standard	none	90	95	93	1	pass
4	male	group A	associate's degree	free/reduced	none	47	57	44	1	failed
5	male	group C	some college	standard	none	76	78	75	1	pass
6	female	group B	associate's degree	standard	none	71	83	78	1	pass
7	female	group B	some college	standard	completed	88	95	92	1	pass
8	male	group B	some college	free/reduced	none	40	43	39	1	failed
9	male	group D	high school	free/reduced	completed	64	64	67	1	failed
10	female	group B	high school	free/reduced	none	38	60	50	1	failed

Showing 1 to 11 of 1,000 entries, 10 total columns

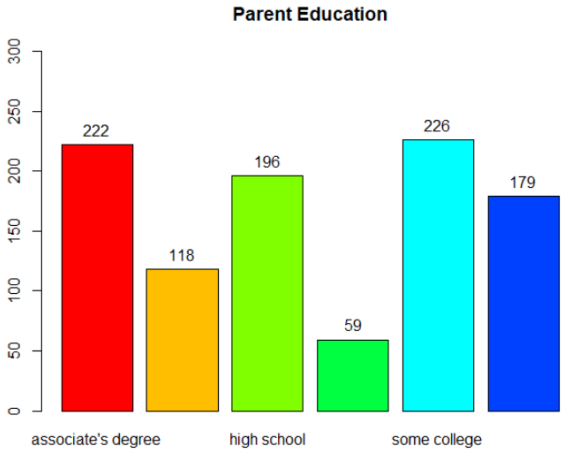
Picture 4.5 Create new column



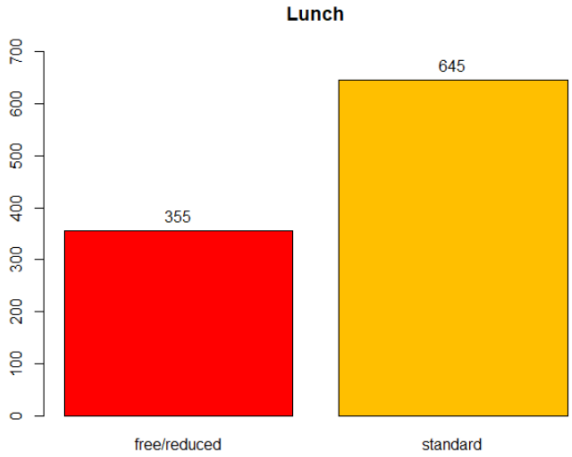
Picture 4.6 Visualization of Gender



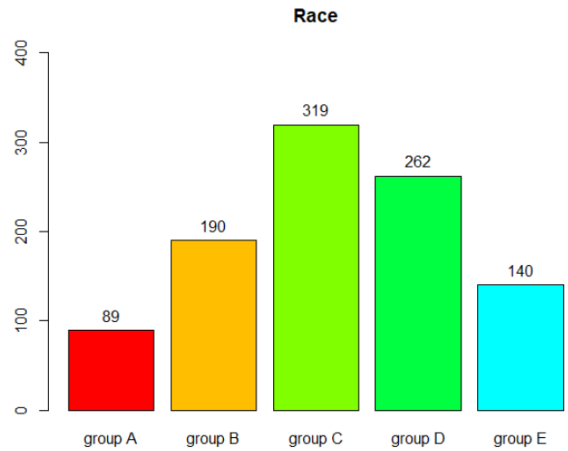
Picture 4.7 Visualization of Exam Status



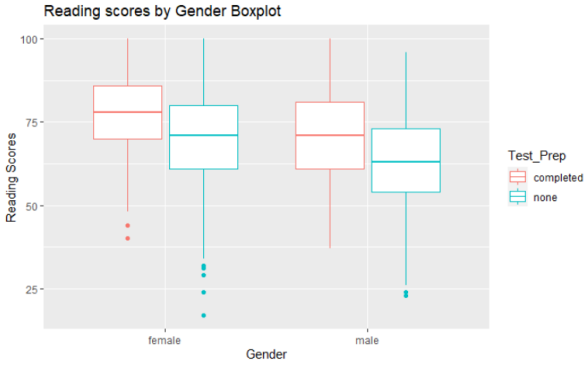
Picture 4.8 Visualization of Parent Education



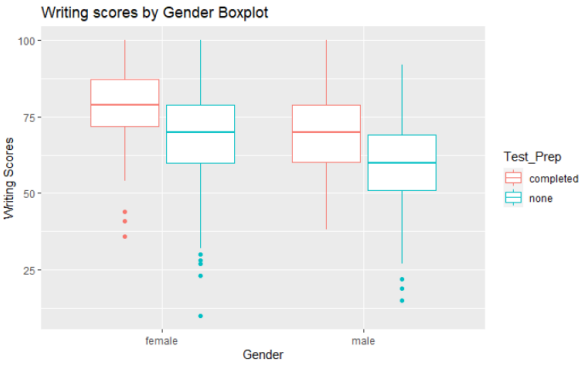
Picture 4.9 Visualization of Students' Lunch



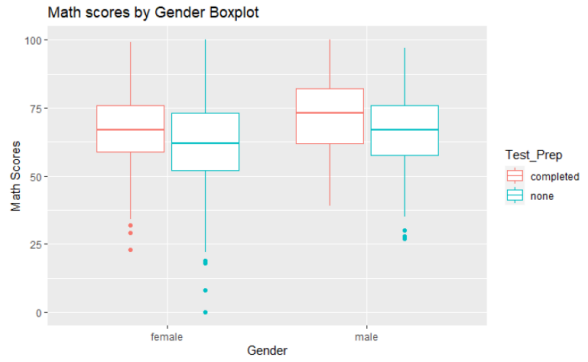
Picture 4.10 Visualization of Students' Race



Picture 4.11 Visualization of Reading Scores by Gender



Picture 4.12 Visualization of Writing Scores by Gender

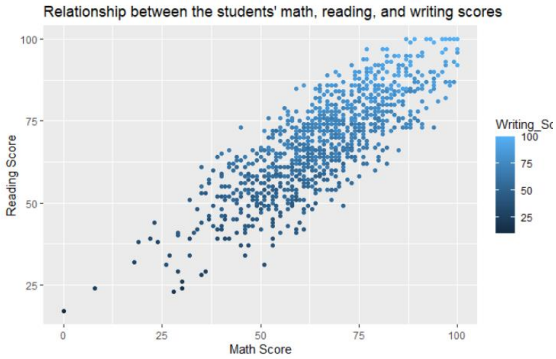


Picture 4.13 Visualization of Math Scores by Gender

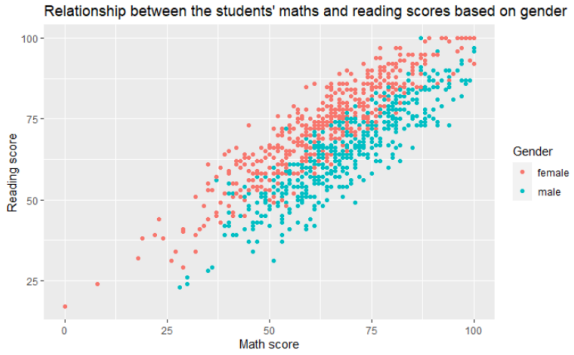
Boxplot is a graph that provides us with measures of central tendency, spread and visual of outliers:

median: the middle value of the dataset.
first quartile: the middle number between the smallest number and the median of the dataset.
third quartile: the middle value between the median and the highest value (not the "maximum") of the dataset.
interquartile range: 25th to the 75th percentile.
outliers, maximum, minimum.
Summary for boxplot visualizations:

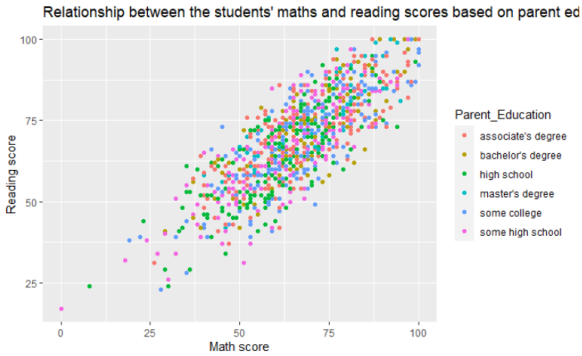
1. students who completed the prep class had better scores in all three tests.
2. male students have received better scores in Math while female students in reading and writing.
3. there is a presence of outliers in all three tests.



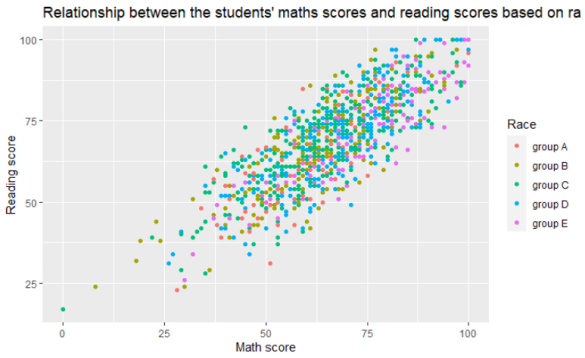
Picture 4.14 Relationship between the Students' and Each Subject



Picture 4.15 Relationship between the Students' Math and Reading Score Based on Gender



Picture 4.16 Relationship between the Students' Math and Reading Score Based on Parent Education

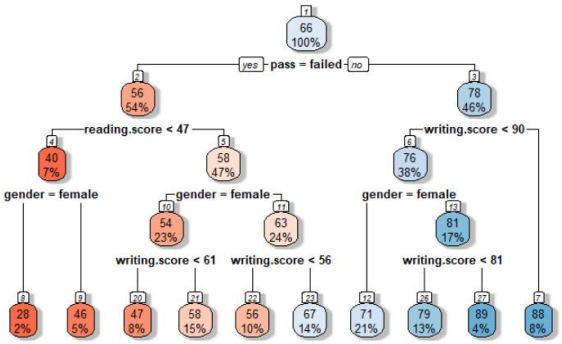


Picture 4.17 Relationship between the Students' Math and Reading Score Based on Race

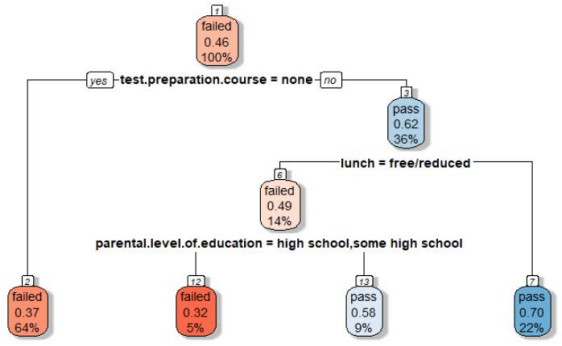
B. Interpretation and Testing Phase

The data will be split in the interpretation and testing section, which previously used the researcher's NIM setseed. Then the data will be divided into two parts, 80% for training set, and 20% for testing set. The training dataset will be stored in an object named "train," and the testing dataset will be stored in an object called "test".

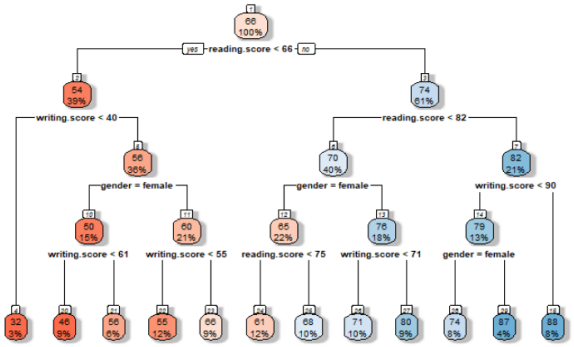
- *Decision Tree*



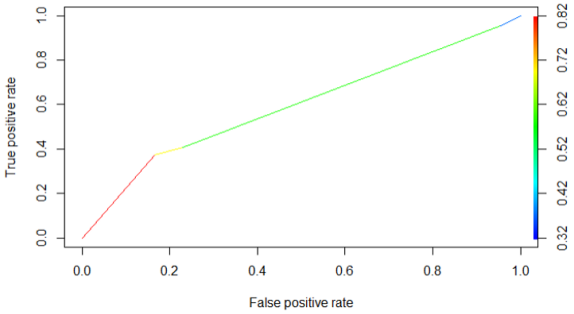
Picture 4.18 Decision Tree party model (exam status, subjects, and gender)



Picture 4.19 Decision Tree rpart model (course, lunch, and parental education)



Picture 4.19 Decision Tree party model (subjects and gender)



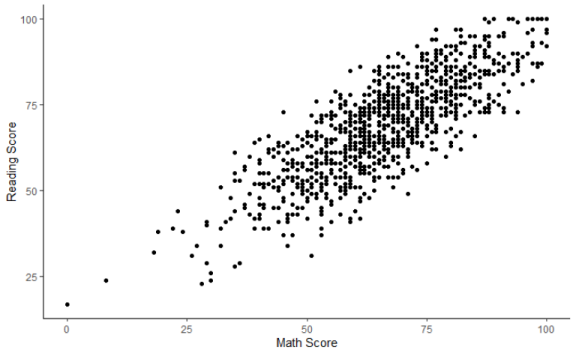
Picture 4.20 True False Positive Rate in Decision Tree

```
> library(caret)
> table_tree1 <- table(pred_math,test$pass)
> confusionMatrix(table_tree1)
Confusion Matrix and Statistics

          
pred_math failed pass
   failed     84   54
   pass       25   37

               Accuracy : 0.605
                 95% CI : (0.5336, 0.6732)
    No Information Rate : 0.545
    P-Value [Acc > NIR] : 0.050739

                  Kappa : 0.182

 Mcnemar's Test P-Value : 0.001631

            Sensitivity : 0.7706
            Specificity : 0.4066
         Pos Pred Value : 0.6087
         Neg Pred Value : 0.5968
             Prevalence : 0.5450
         Detection Rate : 0.4200
   Detection Prevalence : 0.6900
      Balanced Accuracy : 0.5886

       'Positive' Class : failed
```

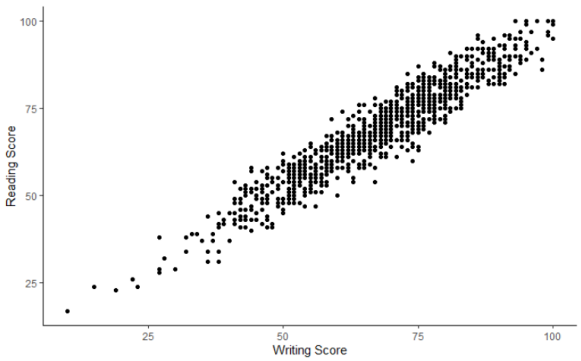
Picture 4.21 Confusion matrix in the Decision Tree model rpart

In the Decision Tree model rpart, the accuracy is 60.5% with a sensitivity value of 77.06%, a specificity value of 40.66%, a positive pred value of 60.87 %, and a negative pred value of 59.68%.

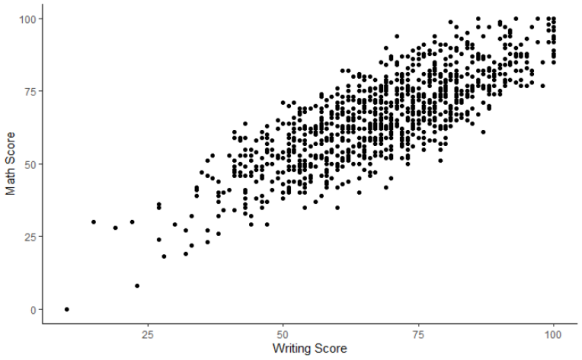
- Random Forest



Picture 4.22 Random Forest Model (Reading and Math Score)



Picture 4.23 Random Forest Model (Reading and Writing Score)



Picture 4.24 Random Forest Model (Math and Writing Score)

```
> pred
Confusion Matrix and Statistics


rf   1  2
  1 98 47
  2 23 32

               Accuracy : 0.65
                 95% CI : (0.5795, 0.7159)
    No Information Rate : 0.605
    P-Value [Acc > NIR] : 0.108891

                  Kappa : 0.2269

 Mcnemar's Test P-Value : 0.005977

            Sensitivity : 0.8099
            Specificity : 0.4051
         Pos Pred Value : 0.6759
         Neg Pred Value : 0.5818
             Prevalence : 0.6050
         Detection Rate : 0.4900
   Detection Prevalence : 0.7250
      Balanced Accuracy : 0.6075

       'Positive' Class : 1
```

Picture 4.25 Confussion Matrix in Random Forest)

In the Random Forest model, the accuracy is 65% with a sensitivity value of 80.99%, a specificity value of 40.51%, a positive pred value of 65.79%, and a negative pred value of 58.18%.

- Validation of Results

Based on the results above using Decision Tree model rpart and the Random Forest model show a level of accuracy that is not much different. The Decision Tree prediction model using rpart shows an accuracy rate of 60.5%. In comparison, the Random Forest prediction model shows an accuracy rate of 65%.

IV. CONCLUSION

A. *Conclusion*

This study has taught us that, factors such as parental level of education, socioeconomic disadvantage (schools' lunch), test preparation courses affect the students' performances in the exams. But there are many exceptions as well. There are students with a low parental level of education scoring full marks. Also, some students have not completed the test preparation course getting full marks. These students may have their own strategies for test preparations. Socioeconomic disadvantage also has many exceptions. These students did not allow economic obstacles to affect their efforts. So, many factors are affecting the students' performances. Some have great effects while some not. Also, there are other factors to be considered as well which are not mentioned in the dataset.

Decision Tree and Random Forest algorithm used for the relationship between each variable and see the accuracy to predicting student performance. By dividing the data into two models, 80% training set and 20% testing set. This study obtained the highest accuracy value of 65% using the Random Forest algorithm. The Random Forest algorithm provides a sensitivity value of 80.99%, a specificity value of 40.51%, a positive pred value of 65.79%, and a negative pred value of 58.18%. Based on these values, the researcher concludes that the Random Forest algorithm performs better than the Decision Tree algorithm when making predictions on the student performance dataset.

B. *Suggestion*

Our plan for the upcoming is to get more data and train the model on that data to get more accuracy. By giving more data we can enhance the accuracy as well this research will be helpful for the faculties as predicting the student those are at risk at the early stage ,As predicting the students status will aids them to get the mandatory action and help for the faculties to determine the student those need more attention and help.

REFERENCES

[1] Oladebinu Tokunbo Olufemi1 , Amos Adekunle Adediran (Ph.D) and Dr. W.O. Oyediran, FACTORS AFFECTING STUDENTS' ACADEMIC PERFORMANCE IN COLLEGES OF EDUCATION IN SOUTHWEST, NIGERIA. Vol.6, No.10, pp.43-56, October 2018

[2] Chew Li Sa, Dayang Hanani bt. Abang Ibrahim, Emmy Dahliana Hossain, Mohammad bin Hossin, Student Performance Analysis System (SPAS). Universiti Malaysia Sarawak (UNIMAS), 2015.

[3] Christopher B. Davison and Gandzhina Dustova, A quantitative assessment of student performance and examination format. Vol 18, Ball State University, 2017.

[4] Hussain, S., Dahan, N. A., Ba-Alwib, F. M., & Ribata, N. (2018). Educational data mining and analysis of students' academic performance using WEKA. Indonesian Journal of Electrical Engineering and Computer Science, 9(2), 447-459

[5] Patil, S., & Kulkarni, U. (2019, April). Accuracy prediction for distributed decision tree using machine learning approach. In 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI) (pp. 1365-1371). IEEE.

[6] Budiman, E., Dengan, N., Kridalaksana, A. H., & Wati, M. (2017, November). Performance of decision tree C4. 5 algorithm in student academic evaluation. In International Conference on Computational Science and Technology (pp. 380-389). Springer, Singapore.

[7] Hoque, M. I., Kalam Azad, A., Tuhin, M. A. H., & Salehin, Z. U. (2020). University students result analysis and

prediction system by decision tree algorithm. Adv Sci Technol Eng Syst J, 5(3), 115-122.

[8] Speiser, J. L., Miller, M. E., Tooze, J., & Ip, E. (2019). A comparison of random forest variable selection methods for classification prediction modeling. Expert systems with applications, 134, 93-101.

[9] Salal, Y. K., Abdullaev, S. M., & Kumar, M. (2019). Educational data mining: Student performance prediction in academic. IJ of Engineering and Advanced Tech, 8(4C), 54-59.

[10] Sullivan, W. (2017). Machine Learning For Beginners: Algorithms, Decision Tree & Random Forest Introduction. Journal of Chemical Information and Modeling M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.