Analysis of Student Performance using R

Abstract

Many students face challenges in securing good grades, primarily due to factors such as taking multiple courses in a single semester and various other reasons. Consequently, it becomes crucial to help students identify their areas of improvement and enhance their grades. Additionally, assessing the impact of past grades, as well as other factors such as the students' environment and habits, can provide valuable insights. By analyzing the interconnection among these factors, it is possible to develop a model that predicts expected future grades. The primary objective of this study is to facilitate students' improvement in their final grades by understanding the positive and negative influences on their academic performance. The findings indicate that, in addition to past grades, factors such as family influence, environment, and habits significantly affect students' final grades. Furthermore, a strong correlation between past and new grades was observed.

Introduction:

In today's educational landscape, institutions are increasingly concerned about their overall results and strive for students to achieve good grades. This emphasis on academic performance often prompts studies aimed at helping students identify areas of improvement and prioritize certain subjects or study areas. Additionally, these studies provide valuable insights for teachers, enabling them to provide targeted instruction and support to help students attain desired grades.

This paper presents an analysis of students' performance based on their past grades and various elements such as age, family size, study time, guardians, and more. The analysis employed two models: regression and random forest models. The dataset was divided into training and testing sets to evaluate the effectiveness of the models and ensure their accuracy. Furthermore, the results were visually represented using appropriate data visualization techniques, facilitating a detailed understanding of the numbers and models.

Each section of this paper describes different steps undertaken during the survey. Firstly, the data was cleaned and processed to ensure its suitability for analysis. Secondly, various data visualization tools were utilized to gain a preliminary understanding of the data, such as its skewness. Next, blank spaces in the dataset were removed, rendering it ready for investigation and model creation. Outliers were also identified and eliminated to prevent biased results. Lastly, two methods, regression and random forest, were employed to develop models after splitting the data into testing and training categories.

Research Problem:

The research aims to analyze the relationship between past grades (G1 and G2) as independent variables and final grades (G3) as the dependent variable, alongside other factors such as school and the marital status of parents.

Literature Review:

Several studies have investigated the correlation between past grades, independent variables, and final grades, employing various tools and methodologies. These studies offer valuable insights for conducting similar analyses. For instance, one study utilized five regression models, achieving an accuracy of approximately 82%. The data collection process involved school reports and questionnaires. Another research paper focused on employing machine learning and educational data mining to extract information about students' grades and their retention in different courses. Universities worldwide rely on such investigations to guide students in achieving good grades by identifying areas requiring improvement.

In a study titled "Predicting Student Performance: A Statistical and Data Mining Approach," the Weka tool and algorithms such as Naïve Bayes, Multi-Layer Perception, J48, SMO, and REPTree were used. The findings revealed that the type of occupation of parents had a significant impact on grades rather than the institution attended, enabling institutions to identify students at risk of academic decline. Another study emphasized the importance of past grades in predicting future grades and recommended incorporating online learning environments alongside traditional methods.

"Student Performance Prediction by Using Data Mining Classification Algorithms" employed different data mining algorithms, including OneR Rule Learner, Decision Tree, K-Nearest Neighbor, and Neural Network. The investigation identified two variables as highly influential on student grades: University Admission grades and the number of failures in the first-year

University examinations. Another study on educational data mining highlighted the concept of extracting patterns from educational data to improve overall student performance and the educational system. Clustering algorithms and Decision Tree algorithms were utilized in this study.

By combining the tools and methods presented in the aforementioned resources, this investigation aims to achieve a comprehensive analysis of the data. The primary focus of this study is the predictive accuracy of the models, supplemented by the conceptual and theoretical understandings derived from the aforementioned resources.

Methodology:

In this study, we utilized the Student Performance Data Set obtained from the UCI Machine Learning Repository to analyze student performance. The dataset consists of 1033 rows and 33 columns, with the target variable being G3, representing the grades earned in the third period.

The attributes of the dataset include:

- School: Categorical variable indicating the school attended (1 for GP - Gabriel Pereira, 0 for MS - Mousinho da Silveira).
- Sex: Categorical variable indicating the student's gender (1 for female, 0 for male).
- Age: Numeric variable representing the age of the student (ranging from 15 to 22 years old).
- Address: Categorical variable indicating the type of address (0 for rural, 1 for urban).
- Famsize: Categorical variable indicating the size of the student's family (0 for LE3 - less than 3, 1 for GT3 - greater than 3).
- Pstatus: Categorical variable indicating the status of the student's parents' relationship (0 for A - Apart, 1 for T - living together).
- Medu: Numeric variable indicating the level of the mother's education.
- Fedu: Numeric variable indicating the level of the father's education.
- Reason: Categorical variable indicating the reason for choosing the school (0 for other, 1 for reputation, 2 for home, 3 for course).
- Guardian: Categorical variable indicating the guardian of the student (0 for other, 1 for mother, 2 for father).
- Traveltime: Numeric variable indicating the time taken to travel to school (1 for <15 min, 2 for 15 min to 30 min, 3 for 30 min to 1 hour, 4 for >1 hour).
- Studytime: Numeric variable indicating the weekly study time (1 for <2 hours, 2 for 2 to 5 hours, 3 for 5 to 10 hours, 4 for >10 hours).

- Failures: Numeric variable indicating the number of past class failures (ranging from 0 to 3).
- Schoolsup: Categorical variable indicating educational support provided by the school (0 for no, 1 for yes).
- Famsup: Categorical variable indicating educational support provided by the family (0 for no, 1 for yes).
- Paid: Categorical variable indicating whether the student paid for extra classes (0 for no, 1 for yes).
- Activities: Categorical variable indicating participation in extra activities (0 for no, 1 for yes).
- Nursery: Categorical variable indicating attendance at a nursery school (0 for no, 1 for yes).
- Higher: Categorical variable indicating the student's aspiration for higher education (0 for no, 1 for yes).
- Internet: Categorical variable indicating internet access (0 for no, 1 for yes).
- Romantic: Categorical variable indicating whether the student was in a romantic relationship (0 for no, 1 for yes).
- Famrel: Numeric variable indicating the student's relationship with their family (ranging from 1 to 5).
- Freetime: Numeric variable indicating the student's free time after school (ranging from 1 to 5).
- Goout: Numeric variable indicating the student's frequency of going out with friends (ranging from 1 to 5).
- Dalc: Numeric variable indicating the student's alcohol consumption during weekdays (ranging from 1 to 5).
- Walc: Numeric variable indicating the student's alcohol consumption during weekends (ranging from 1 to 5).
- Health: Numeric variable indicating the student's medical condition (ranging from 1 to 5).

- Absences: Numeric variable indicating the number of classes the student was absent from (ranging from 1 to 5).
- G1: Numeric variable indicating the first period grades (ranging from 0 to 20).
- G2: Numeric variable indicating the second period grades (ranging from 0 to 20).

In our methodology, we first cleaned and processed the data to ensure its suitability for analysis. We then employed various data visualization techniques to gain insights into the data, such as assessing skewness and distributions. Blank spaces in the dataset were eliminated to prepare it for investigation and model creation. Additionally, we identified and removed outliers to avoid biased results.

For model development, we utilized two approaches: regression and random forest. The dataset was divided into testing and training sets to assess the models' accuracy and effectiveness. By implementing these steps, we aimed to analyze the relationship between past grades, independent variables, and final grades, considering other factors such as the school attended and the marital status of parents.

Data Processing:

At first, we installed the required packages using “install package” command and used them by “library” command.

```
#install.packages
#install.packages("caTools")
#install.packages("ggplot2")
library(tidyverse) # utility functions
library(rpart) # for regression trees
library(randomForest) # for random forests
library(caTools)
library(ggplot2)
```

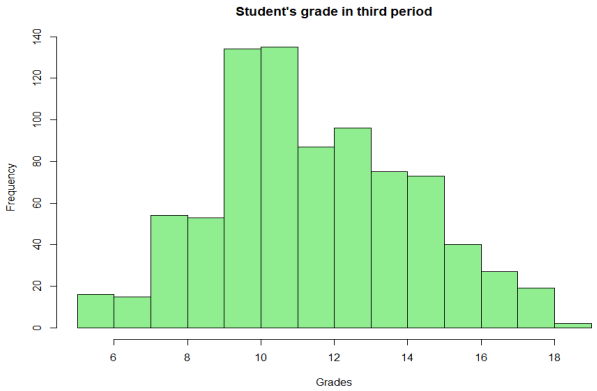
Then we merged the datasets “student” and “student2” into studentData using merge command.

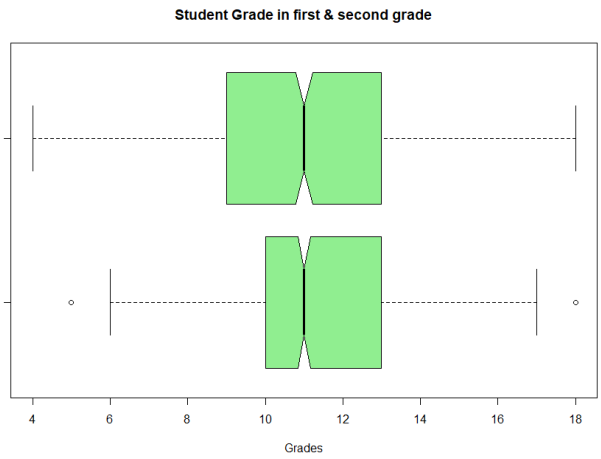
```
student <- read.table('student-mat.csv' , sep = ";" , header = TRUE)
student2 <-read.table('student-por.csv' , sep =  ";" , header = TRUE)

studentData <- merge(x =student2, y = student, all = TRUE , no.dups = TRUE)
```

Then we found out various mathematical parameters such as the structure, variance, range, and standard deviation of the target variable G3 using the following commands.

```
> var(studentData$G3) #variance
[1] 7.482093
> sd(studentData$G3)  #standard deviation
[1] 2.735341
> range(studentData$G3) #range
[1]  5 19
```

We used various data visualization methods such as boxplot, ggplot and scatter plot to visualize our data and get more insights of the dataset.





As, data cleaning plays the most important part of the model creation process. We started deleting the missing values of the dataset using “omit” command.

```
studentData <- na.omit(studentData) # deleting the missing values
sum(is.na(studentData))
```

After that we converted the categorical data with the yes and no values into binary data with values 1,0 using “mutate” command In columns such as schoolsup, famsup, paid ,activities , higher, nursery, internet, romantic, sex, address, school, famsize, Pstatus.

```
studentData <- studentData %>%
  mutate(schoolsup = ifelse(schoolsup == "no",0,1)) # #changing categorical data to binary

studentData <- studentData %>%
  mutate(famsup = ifelse(famsup == "no",0,1)) #changing categorical data to binary

studentData <- studentData %>%  #changing categorical data to binary
  mutate(paid = ifelse(paid == "no",0,1))
```

Then we changed the categorical data into numerical data .

```
studentData$guardian[studentData$guardian %in% ("mother")] <- 1  #changing categ
studentData$guardian[studentData$guardian %in% ("father")] <- 2
studentData$guardian[studentData$guardian %in% ("other")] <- 0


studentData$reason[studentData$reason %in% ("other")] <- 0
studentData$reason[studentData$reason %in% ("course")] <- 3
studentData$reason[studentData$reason %in% ("home")] <- 2
studentData$reason[studentData$reason %in% ("reputation")] <- 1
```

We performed the outlier detection method on selected columns to make sure they don't impact the model accuracy. We used IQR method on G1, G2, G3, age, Fjob , Mjob columns to perform the outlier analysis.

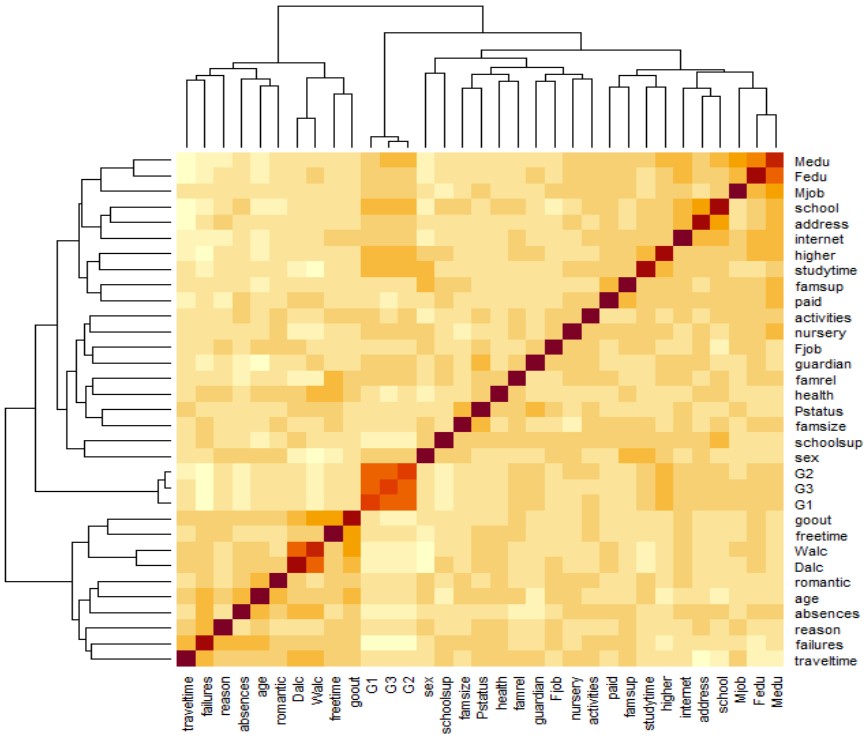
```
##outliers detection in G1
boxplot(studentData$G1)
#find Q1, Q3, and interquartile range for values in column A
Q1 <- quantile(studentData$G1, .25)
Q3 <- quantile(studentData$G1, .75)
IQR <- IQR(studentData$G1)

#only keep rows in dataframe that have values within 1.5*IQR of Q1 and Q3
studentData <- subset(studentData, studentData$G1> (Q1 - 1.5*IQR)
                      &  studentData$G1< (Q3 + 1.5*IQR))
```

Feature Selection:

To make the model more accurate we selected the most important features using two different feature selection methods such as heatmap and forward feature selection method. And we found out very high correlation between G1, G2 and G3 columns.

```
model3 <- lm(G3 ~.,data = studentData)
step <- stepAIC(model3 , direction = "forward")
step
summary(step)
```

```
Call:
lm(formula = G3 ~ 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, data = studentData)

Coefficients:
(Intercept)       school          sex          age      address      famsize      Pstatus
  -0.579462     0.144348     0.075548     0.077566     0.051229     0.065435    -0.155702
       Medu         Fedu         Mjob         Fjob       reason     guardian   traveltime
   0.049720    -0.059605     0.016154     0.013793     0.011369     0.027354     0.094141
  studytime     failures    schoolsup       famsup         paid   activities      nursery
  -0.003004    -0.047612    -0.144365     0.072187    -0.319661    -0.046490    -0.080440
     higher     internet     romantic       famrel     freetime        goout         Dalc
   0.124197     0.145757     0.068455     0.059816    -0.012198    -0.043493     0.023837
       Walc       health     absences           G1           G2
  -0.021853    -0.049720    -0.013739     0.181000     0.772491
```



Splitting The Dataset:

It is important to split the data into training and testing sets because we want our model to learn first to predict the outcome. Training set helps to fit the model and testing set helps to validate the model by getting the predictions on the remaining observations called testing set.

We used the proportion of 0.7:0.3 to split the dataset into training and testing sets.

```
# training and testing data
split <- sample.split(studentData, SplitRatio = 0.7)
split

train <- subset(studentData, split == "TRUE")
test <- subset(studentData, split == "FALSE")
dim(train)
dim(test)
```

MODEL CREATION

1. Linear Model

We used a linear regression model with the most important features selected by the feature selection techniques and evaluated Mean Absolute Error for the model.

```
# LINEAR MODEL
lin_mod=lm(G3~G1+G2+age+Fedu+Medu+Mjob+sex+Pstatus+absences
           +famsize+Dalc+famrel+traveltime, data=train)
pred_test <- predict(lin_mod, test)
cat("Test MAE:", round(mean(abs(pred_test-test$G3)),4))
prediction = (data.frame((pred_test), (test$G3)))
```

```
> cat("Test MAE:", round(mean(abs(pred_test-test$G3)),4))
Test MAE: 0.6689
```

We tested the model by fetching first few values and compared the predicted values with the original ones.

```
> head(prediction,10)
   Predicted G3 Real G3
1     13.346106      13
2      9.808645      10
3      9.314285       9
5      6.789918       5
10    13.322415      13
19    16.370543      16
24    15.331873      15
30    10.796553      11
38    10.419350      10
39     8.230484       8
```

Regression is a way to check relationship between two variables in which one is independent on another variable. This model is sometimes used to predict the dependent variable value based on value of independent variable. (GeeksforGeeks, 2021)

2. **Random Forest Model**

We created one more model with random forest method as below and evaluated the same by calculating the Mean Absolute Error.

```
# Fitting Random Forest to the train dataset
set.seed(120)  # Setting seed
rf = randomForest(formula = G3~G1+G2+age+Fedu+Medu+Mjob+sex+Pstatus+absences
                  +famsize+Dalc+famrel+traveltime ,
                                data = train ,
                                ntree = 500)
pred_test_rf <- predict(rf, test)
cat("Test MAE:", round(mean(abs(pred_test_rf-test$G3)),4))
prediction_rf = (data.frame((pred_test_rf), (test$G3)))
```

```
> cat("Test MAE:", round(mean(abs(pred_test_rf-test$G3)),4))
Test MAE: 0.696
```

Random forest model is combination of various decision tree outputs to get a single outcome. It is very easy to use, and it is used in both classification and regression problems. (Education, 2021)

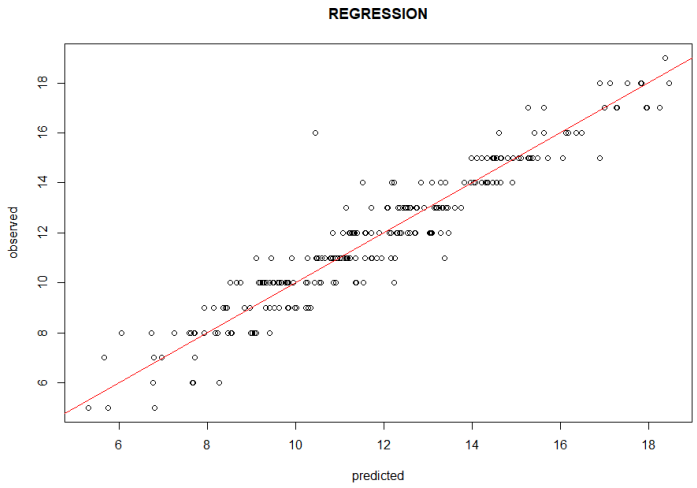
And we tested the model by fetching first few values and compared the predicted values with the original ones.

```
> head(prediction_rf,10)
   Predicted G3 Real G3
1     12.945432      13
2      9.571464      10
3      9.616300       9
5      7.983800       5
10    13.007467      13
19    15.711867      16
24    15.361079      15
30    10.200967      11
38    10.506011      10
39     8.453700       8
```

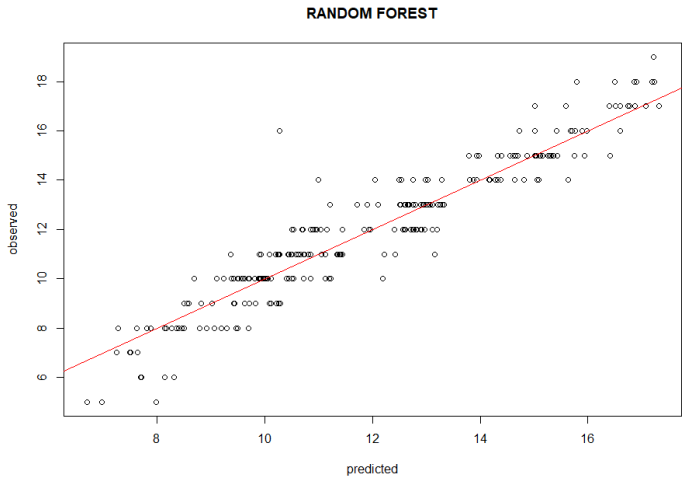
Conclusion and Results

Mean absolute error is defined as the average difference between observed value and the predicted value.

MAE for linear regression turns out to be 0.6689



Random forest – MAE for random forest turns out to be 0.6960



Results have been found that highest correlation is between G1 and G2 with respect to G3. That means past grades impact the final grades the most. Apart from that we have also come across the outcomes that other factors which are also impactful on the final grades are school, status of parent's marriage and school's educational support.

From all the analysis done above according to the results obtained we can just see that the school's management support system allows to collect additional features and along with that it enables to obtain valuable engine has been used.

Overall, it can be concluded that past grades are very impactful in along with some other factors on the final grades of the students so that they can improve their performance.

References

1. *Iqbal, Zafar & Qadir, Junaid & Mian, Adnan & Kamiran, Faisal. (2017). Machine Learning Based Student Grade Prediction: A Case Study. https://www.researchgate.net/publication/319350236_Machine_Learning_Based_Student_Grade_Prediction_A_Case_Study*

2. *RAMESH, VAMANAN & P.PARKAVI, & Ramar, K.. (2013). Predicting Student Performance: A Statistical and Data Mining Approach. INTERNATIONAL JOURNAL OF COMPUTER APPLICATIONS. 63. 975-8887. https://www.researchgate.net/publication/254558610_Predicting_Student_Performance_A_Statistical_and_Data_Mining_Approach*

3. *Audu, A. E. (2021, December 15). Predicting High School Students Grades with Machine Learning (Regression). Medium.* https://medium.com/geekculture/predicting-high-school-students-grades-with-machine-learning-regression-3479781c185c

4. *Cortez, P., & Silva, A. (n.d.). USING DATA MINING TO PREDICT SECONDARY SCHOOL STUDENT PERFORMANCE.* http://www3.dsi.uminho.pt/pcortez/student.pdf

5. Kabakchieva, Dorina. (2012). Student Performance Prediction by Using Data Mining Classification Algorithms. International Journal of Computer Science and Management Research. 1. 686-690. Microsoft Word - paper101.docx (researchgate.net)

6. UCI Machine Learning Repository: Student Performance Data Set. (2014). Uci.edu. https://archive.ics.uci.edu/ml/datasets/Student+Performance

7. *S, S. T. (2018). Prediction of Students Academic Performance using Data Mining: Analysis. International Journal of Engineering Research & Technology, 3(30). https://doi.org/10.17577/IJERTCONV3IS30011*

8. GeeksforGeeks. (2021, November 29). Regression and its Types in R Programming. https://www.geeksforgeeks.org/regression-and-its-types-in-r-programming/

9. Education, I. C. (2021, January 26). Random Forest. https://www.ibm.com/cloud/learn/random-forest