International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2020 IJSRCSEIT | Volume 6 | Issue 3 | ISSN : 2456-3307
DOI : https://doi.org/10.32628/CSEIT206399

Factors Impacting Students Academic Performance Using R Code Programming Language for Analysing the Data

Adesina Fatimat .O[1], Akande Ademola[1], Ajala Abiodun .L[2], Kolawole Tolulope [3], Ogundeji Tajudeen .O[1].
*[1]Department of Physics, The Polytechnic, Ibadan, Oyo State, Nigeria.
[2]Department of Mechanical Engineering, The Polytechnic, Ibadan, Oyo State, Nigeria.
[3]Department of Science Laboratory Technology, The Polytechnic, Ibadan, Oyo State, Nigeria.

ABSTRACT

This work intends to investigate and propose a model that can be used by students, parents, teachers and education policy makers to understand and predict high school student academic performance based on pre-defined factors identified as capable of impacting students' academic performance. This study uses ex post facto research design. An instrument measuring students' academic performance has been used to collect data from the management students and R-Code programming language was used to analyse the data collected and there was a positive and statistically significant impact of learning facilities, age, romance and proper guidance from parent on student performance.

Keywords : Academic Performance, Ex Post Facto, R-Code

I. INTRODUCTION

In the last two centuries, education has become a widely accepted catalyst for long-term economic growth because of the likely higher income earning of students who managed to progress based on academic performance. Academic performance remains a major determinant in the assessment of how a student has developed an understanding of subject-specific skills and transferable skills over a given period of time. In many countries, government and parents invest reasonable amount of resources to ensure that students not only acquire knowledge and transferable skills but perform satisfactorily to facilitate their usefulness to self and communities in general. However, despite the huge investment made by government and parents on education, student academic performance still does not reflect the expected outcome. Student academic performance is an important issue among educational policy makers, parents and government because of the possible economic loss due to the likely increase in the number of low educated workforce with less skills and possible less income tax revenue. In 2018, research conducted by Higher Education Statistic Agency in United Kingdom on dropout rate shows that on the average, 6.4 percent of students discontinued their studies within 2015 - 2016 academic year, representing a consistent rise for the third year in a row. However, the research could not specifically identify the factors influencing this dropout rate [1].

Therefore, the investigation of the factors impacting students' academic performance becomes necessary to help improve the retention rate in the higher institution, reduce the potential economic loss for government and parents. The scope of this work is primarily limited to the high school students;

therefore, basic school students' academic performance is outside the scope of this work.

II. METHODS AND MATERIAL

This work considered ex-post facto research design using mixed method (i.e qualitative and quantitative) appropriate because it provides a systemic scientific and analytical approach to examine dependent and independent variables. More specifically, ex-post facto research is appropriate in studies where independent variable or variables (multivariate) have occurred in the past but the researcher commenced a retrospective study of the dependent variable or variables to determine possible relationship and effect. This work examined absences, age, romantic, first term grade G1 and G2 as independent variables in retrospect against academic performance G3 as dependent variable with the aim of determining the relationship between these variables. This work used standard deviation, mean and regression analysis for the data interpretation and analysis. Student Performance G3 was measured with age, romantic, G1, G2 and absences using multiple regression analysis to facilitate the prediction of the impact of age, romantic, absences, G1 and G2 on student final academic performance G3. The dataset chosen is the 'Student Performance Data Set' - specifically related to performance in the field of Mathematics. This was accessed from UCI data on 18th February 2019 1900hrs. The original source for this dataset was: Paulo Cortez, University of Minho, Guimaraes, Portugal [18].

The chosen dataset, 'Student Performance', is made of 649 rows and 33 features. The dataset is multivariate and contains attributes related to student achievement in secondary education in two Portuguese schools. The dataset selected is concerned in the subject of Mathematics. The dataset was modelled under binary and 5 level classification and regression tasks. The target attribute is the final grade after the 3rd final year, G3. Some initial exploration follows, beginning with the distribution of grades and ages of the students as shown in figure 1:

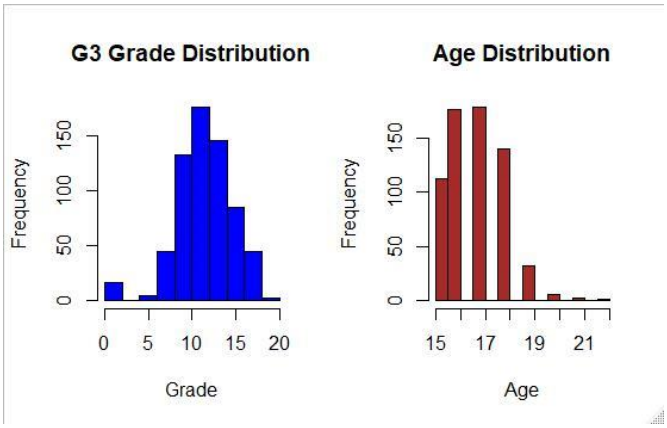


Figure 1.0: Distribution of Age and Grade G3

```
library(ISLR)
student.perf <- read.csv("C:/Users/User/Desktop/Ongoing Report/student-por.csv", sep=";")
par(mfrow = c(1,2))
# 1 rows and 2 columns hist(dfs$age,xlab="Age", main="Age Distribution", col ="grey")
hist(student.perf$G3,xlab="Grade", main="G3 Grade Distribution", col ="blue")
hist(student.perf$age,xlab="Age", main="Age Distribution", col ="brown")
```

Figure 1.1: R code that generates the above graph.

Box plots relating to time factors are shown in Figure 2. These are an important component that can be controlled by the student

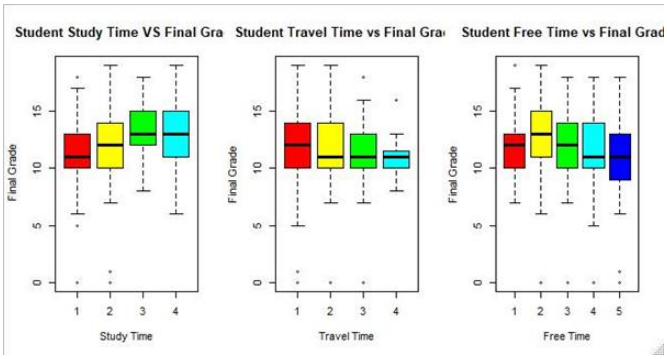


Figure 1.2: Boxplots of G3 Vs Time Factors

```
# Performance based on student study time
par(mfrow = c(1,3)) # 1 row and 3 columns
studyG3<- transform(student.perf,studytime=factor(studytime))
boxplot(G3~studytime, studyG3,col=rainbow(6),xlab="Study Time",
        ylab=" Final Grade", main="Student Study Time VS Final Grade")


# Performance based on student travel
travelG3<- transform(student.perf,traveltime=factor(traveltime))
boxplot(G3~traveltime, travelG3,col=rainbow(6),
xlab="Travel Time", ylab="Final Grade", main="Student Travel Time vs Final Grade"

# Performance based on student free time
freeG3<- transform(student.perf,freetime=factor(freetime))
boxplot(G3~freetime, freeG3,col=rainbow(6),
xlab="Free Time", ylab="Final Grade", main="Student Free Time vs Final Grade")
```

Figure 1.3: Rcode for above box graph

From the research carried out, it is shown that a few predictor variables should be chosen for experimentation. Variables relating to Failure, Parental education, Time and Social activities will be analysed initially. The output from the dataset has 3 grades; one grade per semester (3 semesters per term). This project focuses on the final and last grade, G3. Using a corrolation matrix, it is possible to identify predictor variables that are likley to influence the student grades from the given dataset. Before this can be run, the data must be processed to change all yes / no values into binary 1 / 0 values using ”lapply”. Variables will then be selected into a new dataset.

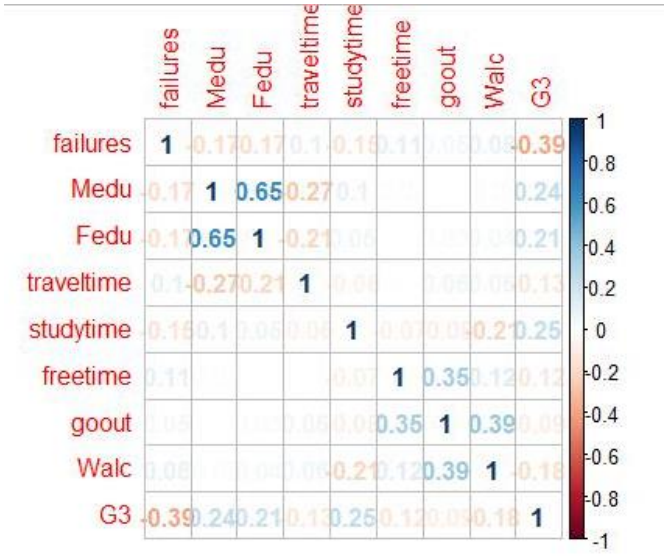


Figure 1.4: correlation graph for all variable predictor

```
library(corrplot)
student.pref <-  student.mat <- read.csv("C:/Users/User/Desktop/Ongoing Report/student-por.csv", sep=";",
                                         stringsAsFactors = TRUE)
student.perf[] <- lapply(student.mat,as.integer)

newdata<- student.mat[c("failures",
                        "Medu","Fedu",
                        "traveltime","studytime","freetime",
                        "goout","Walc",
                        "G3")]
corrplot(cor(newdata), method = "number")
```

Figure 1.5: Rcode for the above correlation graph

Correlation values greater than 0.9 are a concern; this indicates high-colinearity and therefore one predictor from the co-linear pair should be disguarded.

Another method to select predictors is to use stepwise forward regression

```
library(olsrr)
student.perfs <- read.csv("C:/Users/User/Desktop/Ongoing Report/student-por.csv", sep=";",
                          stringsAsFactors = TRUE)
student.perfs[] <- lapply(student.perfs,as.integer)
model <- lm(G3 ~ ., data = student.perfs)
k <- ols_step_forward_p(model)
k
```

The output from above code

```
                        Selection Summary
-------------------------------------------------------------------------------
        Variable                    Adj.
Step    Entered      R-Square    R-Square      C(p)        AIC         RMSE
-------------------------------------------------------------------------------
   1    G2             0.8437      0.8435    33.4932    2164.2708    1.2781
   2    G1             0.8478      0.8473    17.9909    2149.3096    1.2625
   3    failures       0.8490      0.8483    14.5164    2145.9120    1.2582
   4    Fjob           0.8500      0.8491    12.1840    2143.6085    1.2550
   5    school         0.8511      0.8499     9.6973    2141.1213    1.2517
   6    sex            0.8521      0.8507     7.2840    2138.6772    1.2484
   7    traveltime     0.8529      0.8513     5.6505    2136.9953    1.2458
   8    reason         0.8535      0.8517     4.9839    2136.2797    1.2442
   9    health         0.8543      0.8522     3.8164    2135.0393    1.2420
  10    absences       0.8547      0.8525     3.6985    2134.8636    1.2409
  11    Dalc           0.8552      0.8527     3.8468    2134.9554    1.2401
  12    famsup         0.8555      0.8528     4.4498    2135.5120    1.2397
  13    schoolsup      0.8558      0.8529     4.9731    2135.9829    1.2392
  14    address        0.8562      0.8530     5.5649    2136.5211    1.2388
  15    higher         0.8564      0.8530     6.3726    2137.2811    1.2386
-------------------------------------------------------------------------------
```

III. RESULT AND DISCUSSION

In the experiments, a multiple linear regression model will be produced using the predictors identified in the previous exploration steps. The code used is shown below:

```
student.per<- read.csv("C:/Users/User/Desktop/Ongoing Report/student-por.csv", sep=";",
                       stringsAsFactors = TRUE)
student.per[] <- lapply(student.per,as.integer)
student.perss <- student.per[c("G3","G2","famrel",
                               "absences", "G1","age","activities",
                               "Walc", "romantic","school","failures",
                               "schoolsup")]
Model=lm(G3~G2+G1+failures+absences+ age+activities+Walc+ romantic+school+schoolsup, data=student.perss)
student.perss$predictedG3<-predict(Model)
summary(Model)
confint(Model)
```

Results shown below:

```
Call:
lm(formula = G3 ~ G2 + G1 + failures + absences + age + activities +
    Walc + romantic + school + schoolsup, data = student.perss)

Residuals:
    Min      1Q  Median      3Q     Max
-9.0023 -0.4849 -0.0370  0.6221  6.0390

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.39174    0.89318   0.439 0.661103
G2           0.87742    0.03441  25.497  < 2e-16
G1           0.13656    0.03691   3.699 0.000235
failures    -0.24098    0.09485  -2.541 0.011299
absences     0.01907    0.01125   1.694 0.090716
age          0.03117    0.04475   0.696 0.486395
activities  -0.02200    0.09968  -0.221 0.825367
Walc        -0.07801    0.03961  -1.969 0.049363
romantic    -0.03640    0.10513  -0.346 0.729293
school      -0.20281    0.11299  -1.795 0.073139
schoolsup   -0.17976    0.16759  -1.073 0.283850

(Intercept)
G2          ***
G1          ***
failures    *
absences    .
age
activities
Walc        *
romantic
school      .
schoolsup
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.253 on 638 degrees of freedom
Multiple R-squared:  0.8518,    Adjusted R-squared:  0.8495
F-statistic: 366.8 on 10 and 638 DF,  p-value: < 2.2e-16
```

From the results, the significance of the predictors is displayed based on the ’p-value’ - a p-value below 0.05 indicates this predictor is significant. For any p-value below 0.05 the t-value should be above 1.96 so therefore, a new model is created using just the most significant predictors (p-val less than 0.05) and (t-val

more than 1.96): G1 Grade, G2 Grade, Absences, Age, and Romantic

```
student.perf <-  read.csv("C:/Users/User/Desktop/Ongoing Report/student-por.csv", sep=";",
                          stringsAsFactors = TRUE)
student.perf[] <- lapply(student.perf,as.integer)
student.perfs <-student.perf[c("G3","G2","G1","absences","age", "romantic")]

Models=lm(G3~G2+G1+absences+age+romantic, data=student.perf)
student.perfs$predictedG3 <- predict(Models)
summary(Models) # Summary
confint(Models) #confidence interval
paste("AIC:",AIC(Models))# Akaike's information criterion
paste("BIC:",BIC(Models)) # Bayesian information criterion
```

Results shown below (Summary followed by Confidence Interval and AIC/BIC

```
Call:
lm(formula = G3 ~ G2 + G1 + absences + age + romantic, data = student.perf)

Residuals:
    Min      1Q  Median      3Q     Max
-9.4478 -0.4494 -0.0911  0.6745  5.8058

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.227813   0.760183  -0.300   0.7645
G2           0.896415   0.034157  26.244  < 2e-16 ***
G1           0.154120   0.036588   4.212 2.89e-05 ***
absences     0.020298   0.010902   1.862   0.0631 .
age         -0.002055   0.042443  -0.048   0.9614
romantic    -0.025207   0.104974  -0.240   0.8103
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.262 on 643 degrees of freedom
Multiple R-squared:  0.8486,    Adjusted R-squared:  0.8474
F-statistic: 720.7 on 5 and 643 DF,  p-value: < 2.2e-16

> confint(Models) #confidence interval
                 2.5 %     97.5 %
(Intercept) -1.720553644 1.26492737
G2           0.829341554 0.96348828
G1           0.082273863 0.22596684
absences    -0.001110133 0.04170690
age         -0.085398482 0.08128775
romantic    -0.231341194 0.18092627
> paste("AIC:",AIC(Models))# Akaike's information criterion
[1] "AIC: 2151.77584124241"
> paste("BIC:",BIC(Models)) # Bayesian information criterion
[1] "BIC: 2183.10387025934"
>
```

With this refined model the F-statistic has increased from 193.4 to 380.9 which indicates a better fit. As it can be seen, the Std. Error for G1, G2, Absences, Age and Romantic has also slightly decreased, which is another indicator of a better fit. However, the t-value should be greater than 1.96 for any predictor with a p-value less than 0.05. This is true for G1,G2 and Absences but not Age or Romantic. Therefore age and romantic is of less significance despite its low p-value. A minor reduction in AIC and BIC also indicates this model has improved fit over the first. The regression equation for this work has the following form:

= + 1 2 + 2 + 3 1 + 4 + 5

Where:

SAP = Student Academic Performance

Abs = Absences

Rom = Romantic

Age = Age

G1 = First semester/term grade

G2 = Second semester/term grade

b = strength of the extent of impact of the independent variable on the predictor

a = Dependent variable; G3

From the above regression equation:

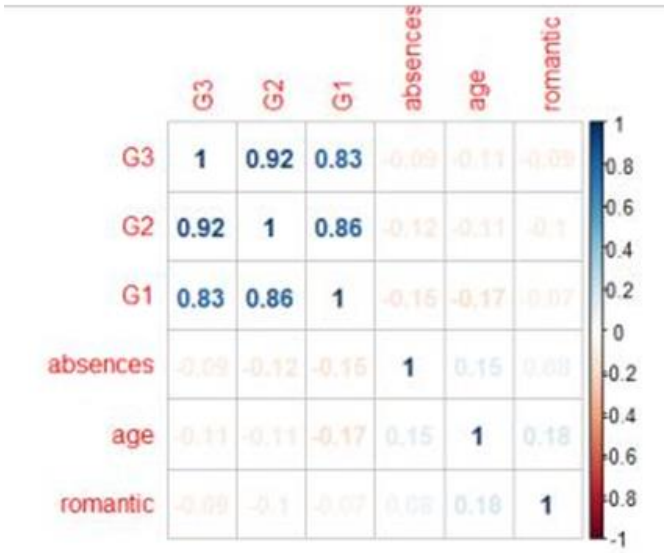
SAP = 1.33776 + 0.95515G2 + 0.04461abs + 0.18089G1 + (-0.17046)Age + (-0.40330)Rom ..(1)

Equation (1) indicates that if absences could be improved by one unit, there would be a 0.95515 unit change on student academic performance when other independent variables remain constant. As shown in the equation, the coefficients are positive for 3 out the 5 variables. This is an indication that G2, absences and G1 have direct relationship with student academic performance more than the age and romantic variables. This work measures the t-value for individual variables in the model to ascertain the significance of each variable with student academic performance. The result shows that four out of the five variables are more significant predictors of the student academic performance: Second semester/term grade G2 (t=19.193, p =.000 less than 0.05); Absence(t = 3.641, p = 0.000 less than 0.05); First semester/term grade G1 (t = 3.246, p = 0.001); Age (t = -2.186, p = 0.029) and Romatic (t = -1.930, p = 0.054). Although, the result of the standard error (??) suggests that romantic variable has the highest impact (?? = 0.20898) on student performance however, the significance of this variable is quite less compared to other four variables when P-value is considered. Similarly, the next variable with higher impact is age (0.07796), however, this variable has a negative t-value. Therefore, G1, G2 and absences has higher impact with respective ?? - value of 0.05573, 0.04977, and 0.01225. The confidence interval of this student performance model is 95%. It can be said that G2, G1, and absences have the highest significant and direct relationship with student performance. In terms of impact, age and romantic variables provide the largest impact but their negative t-value and estimate

value will continue to make them a minus from the predictive model. Corrolation of the predictors is shown:

```
##corrolation of predicitve model
student.perf1 <- read.csv("C:/Users/User/Desktop/Ongoing Report/student-por.csv", sep=";",
                          stringsAsFactors = TRUE)
student.perf1[] <- lapply(student.perf1,as.integer)
student.per1<- student.perf1[c("G3","G2","G1","absences","age", "romantic")]
corrplot(cor(student.per1), method = "number")
```

Results shown below:




```
# Create Training and Test datasets set.seed(100)
# Random sampling seed
set.seed (200)
trainingRowIndex <- sample(1:nrow(student.perf), 0.8*nrow(student.perf))
trainingData <- student.perf[trainingRowIndex, ] # Model training data
testData <- student.perf[-trainingRowIndex, ] # Test data
# Use newModel to predict G3 on test data
newModel <- lm(G3~G2+G1+absences+age+romantic, data=trainingData)
# Default to 95% confidence and prediciton interval
g3Predict <- predict(newModel, testData, interval="prediction")
# Calculate prediction accuracy and error rate
Act_vs_Pred <- data.frame(cbind(actuals=testData$G3, predicteds=g3Predict))
correlation_accuracy<- cor(Act_vs_Pred) # 91.2%
head(Act_vs_Pred, 10)
```

The corrolation values are 0.9 maximum, greater than 0.9 is a concern, therefore these are borderline acceptable. The next step is to test the model on new data. Therefore the dataset will be split into 80 percent training and 20 percent testing. The model will be built on the training data and then used to predict the test data

Results shown below:

	actuals	fit	lwr	upr
1	11	9.47406	6.867169	12.08095
12	13	11.90581	9.456749	14.35487
18	14	14.27644	11.836573	16.71631
26	12	11.23768	8.796559	13.67879
27	12	12.32850	9.881910	14.77509
30	12	11.47040	9.026747	13.91405
35	12	12.40519	9.966757	14.84363
40	12	13.73310	11.285188	16.18101
42	11	11.21849	8.766197	13.67078
50	12	12.49333	10.050363	14.93629

The relative importance of each predictor can be measured with the following code :

Results shown in figure 4

```
# Calculate Relative Importance for Each Predictor
library(relaimpo)
# Bootstrap Measures of Relative Importance (1000 samples)
boot <- boot.relimp(newModel, b = 1000, type = c("lmg", "last", "first", "pratt"),
rank = TRUE, diff = TRUE, rela = TRUE)
plot(booteval.relimp(boot,sort=TRUE)) # plot result
```

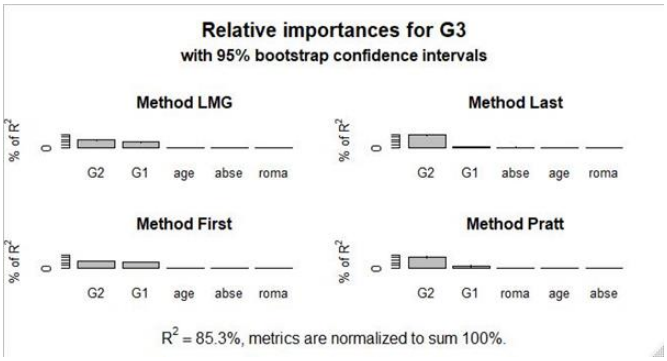


Figure 1.8: relative importance of predicitor variables chosen to predict G£

IV.CONCLUSION

This work as shown that G2, absences and G1 are the most significant factors for predicting student academic performance. The two other factor factors that could impact student academic performance as shown from the exploration and experimentation are age and romance. This work has been able to develop an analytic predictive model using five factors which are justified statistically, to determine student performance. However, this work has not examined language, learning facilities and tuition as factor that could impact student academic performance. Therefore, future research may be conducted to analyse these factors to develop an enhanced student performance model.

V. REFERENCES

[1]. BBC News. Scottish university dropout rate lowest for 19 years. Scotland UK. 2018. url: https : //www.bbc.co.uk/news/uk-scotland-43333615.

[2]. Jordan Navarro et al. “The Relative Age Effect and Its Influence on Academic Performance”. In:PloS one.2015.doi:10.1371/journal.pone.0141895.eCollection2015.

[3]. J. Too J. M. Momanyi and C. Simiyu. “Effect of Students’ Age on Academic Motivation and Academic Performance among High School Students in Kenya”. In: Asian Journal of Education and E-Learning. Vol. 3. 5. 2015. url: https://ajouronline.com/index.php/AJEEL/article/view/3130.

[4]. S. Malik S. P. Singh and P. Singh. “Factors Affecting Academic Performance of Students”. In: Peripex Indian Journal of Research. Vol. 5. 4. 2016. url: doi:10.15373/22501991.

[5]. Y. Gooding. “The relationship between parental educational level and academic success of college freshmen ”.Retrospective Theses and Dissertations.” In: (2001). url: https://lib.dr.iastate.edu/ rtd/429.

[6]. P. Cortez. “Student Performance Data Set”. In: Paulo Cortez, University of Minho, Guimar˜aes, Portugal. url: https://archive.ics.uci.edu/ml/datasets/student+performance

[7]. M. S. Farooq et al. “Factors Affecting Students ’ Quality of Academic Performance : A Case of Secondary”. In: 2012. url: https : / / www .semanticscholar . org / paper / FACTORS - AFFECTING - STUDENTS-%E2%80%99-QUALITY-OF-ACADEMIC-%3A-Farooq Chaudhry/028a5f97888d3bb42e4e17864aaf8eb3f0f74695.

[8]. J. Smith W. Arulampalam R. A. Naylor. “Am I missing something? The effects of absence from class on student performance, Economics of Education Review”. In: 31.4 (2012), pp. 363–375. issn: 0272-7757.doi:https://doi.org/10.1016/j.econedurev.2011.12.002..url:http://www.sciencedirect.com/science/article/pii/S0272775711001786.

[9]. G-R. Gonzalo. “Follow the Leader: Student Strikes, School absenteeism and Long Term Implication for Education Actors.”In: (Nov.2017). url:https://warwick.ac.uk/fac/soc/economics/staff/ ggaete/strikes_ggr_jmp.pdf.

[10]. L. Kann et al. C. N. Rasberry G. F. Tiu. “Health-Related Behaviours and Academic Achievement Among High School Students - United States, 2015” .pp 921-927. Doi: http //dx.doi.org10.15585/mmwr.mm6635al.

Cite this article as :

Adesina Fatimat. O, Akande Ademola, Ajala Abiodun. L, Kolawole Tolulope, Ogundeji Tajudeen. O, "Factors Impacting Students Academic Performance Using R Code Programming Language for Analysing the Data", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 6, Issue 3, pp.404-409, May-June-2020. Available at doi : https://doi.org/10.32628/CSEIT206399
Journal URL : http://ijsrcseit.com/CSEIT206399