Study Factors for Student Performance Applying Data Mining Regression Model Approach

Shakir Khan

College of Computer and Information Sciences,
Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, Saudi Arabia

Summary

In this paper, we apply data mining techniques and machine learning algorithms using R software, which is used to predict, here we applied a regression model to test some factor on the dataset for which we assumed that it effects student performance. Model was built on an existing dataset which contains many factors and the final grades. The factors tested are the attention to higher education, absences, study time, parent's education level, parent's jobs, and the number of failures in the past. The result shows that only study time and absences can affect the students' performance. Prediction of student academic performance helps instructors develop a good understanding of how well or how poorly the students in their classes will perform, so instructors can take proactive measures to improve student learning. This paper also focuses on how the prediction algorithm can be used to identify the most important attributes in a student's data.

Key words: *data mining technique; big data , student performance; Linear Regression Model*

1. Introduction

Big data technologies are used to extract valuable and meaningful information from very big volumes of a wide variety, veracity and fast-growing data [1]. Using big data to develop different types of applications for education data mining, extracting knowledge from those data helps education sectors such as schools and universities to be smarter. The education data consider big data because the volume and the variety its daily produced a large amount of data about students, their activity and their interaction with learning systems or the platforms of learning, also the activities of learning, courses information which different from one another, also different information that help to improve the education processes quality. One of the applications on educational big data is predicting student performance; it is considered the oldest and most popular application of Data Mining in education. In this research, a model is developed to predict student performance using R-software to test factors' effect on student performance. First, I downloaded the dataset from UCI [2] and after that split the data set into training and testing datasets. In a dataset, a training dataset is used to build up a model, while a testing dataset is to validate the model. The main factors selected to test its effects on the students' performance are the attention to take higher education, absences, study time, mother education level, father education level, mother job, father job and the number of failures in the past. Here those factors are used to build a linear regression model on response variables for final grade G3. Table 1 shows the description of dataset variables model are the grades G3 then excluded from the testing dataset. On the basis of the model of prediction, results have been generated and it shows that only study time and absences can affect the students' performance. In educational data mining method, predictive modeling is usually used in predicting student performance. In order to build the predictive modeling, there are several tasks used, which are classification, regression and categorization. We will focus on regression, there are two main factors in predicting student's performances, which are attributes and prediction methods.

Manuscript received February 5, 2021
Manuscript revised February 20, 2021
https://doi.org/10.22937/IJCSNS.2021.21.2.21

2. Related Work

Acoording to [3], they developed a predictive model for students performance using linear regression, they examined their relationship between social media use and student performance on middle and high school students, they conclude to there are heavy associated with lower school connectedness and performance of student. [4] created classification rules and predicting students' performance on course program ,they analyze old enrolled students' data in other courses program in intervalbetween 2005-2010, they were able to predict the final grade , help thestudent's to improve their performance and reduce the failing ratio of the course. [5] produced qualitative predictive models that were able to predict the students' grades from the dataset they collected. Four decision tree algorithms were implemented, with the Naïve Bayes algorithm. It found that the student's performance is not totally dependent on their academic efforts. [6] used a linear regression model to predict students' performance that explain methods about how they could help students and teachers to improve their education. This study focuses on many factors to prove if they have an effect on

student performance based on a dataset in secondary education of two Portuguese schools. it includes academic and personal characteristics of the students in addition to the final grades. I prefer to use a linear regression model for data mining since I attend to determine the strength of those factors on the grades and predicting the effect. Data mining [7] can be generally defined as a technique to find patterns (extraction) or interesting information in large amounts of data [8]. This technique has been extensively applied in research in areas such as health, engineering [9], marketing [10], education, and others. Educational Data Mining (EDM) is an interdisciplinary research area that deals with the development of methods and techniques to explore data from educational contexts and which has been exploring different techniques to detect at-risk students [11]. Big data and education data mining process is discussed in [12] in which they applied two methods of data ming to validate the perfomance. [13] studied the students performance with Moodle learning management system, In the study, a correlation analysis is implemented to determine the impact of students' educational activity in the Moodle system on the final assessment. The results expose that gender affiliation compares with the overall performance but does not affect the selection of training materials. Moreover, it is shown that students who got the highest grades performed at least 210 logs during the course. His study shows that learning management systems enable generating new information about student behavior based on their digital profile.

3. Dataset Description

The dataset requirement for this research is fulfilled through UCI repository [2]. The dataset includes student achievement in secondary education of two Portuguese schools. The data attributes include student grades, demographic, social and school-related features. The target attribute final grade G3. It is has 33 attributes and 649 instances. I choose variables which should include in this model higher, absences, study time, Medu, Fedu, Mjob, Fjob, Failure, G1, G2 and G. Table 1 show the description of dataset variables

4. Methodology

The methodology deals with different stages of the model it consists of data collection, data preprocessing, generating training and testing dataset, building the model, prediction the details about stages describe below:

4.1 Data Collection

The dataset is collected by UCI repository their source is Paulo Cortez from University of Minho from Portugal [2], the data include student grades, demographic, social and school related features, and it was collected by using school reports and questionnaires.

4.2 Data Preprocessing

In this stage dataset is prepared for data mining technique, preprocessing methods include cleaning, variable transformation and so on. The data was needed in preprocessing for this model to choose the factor which I assume to effect on student performance. Here I focus my attention to choose the response variable because I try to build model and prediction, before selecting my respond variables the hypothesis is built to test the relation between grades and those factor as follow:

H1: The attention to take higher education effect on student performance.
H2: The absences effect on student performance.
H3: The study time effect on student performance.
H4: Parent education effect on student performance.
H5: Parents job effect on student performance.
H6: Number of past class failures effect on student performance.

I attend to keep only those factors and test their correlations with G3, the requirement of the model is numerical only.

Table 1. List of the variables that measures students' performance

Variables	type	Description
Higher	Char	wants to take higher education(yes/no)
absences	Integer	number of school absences
study time	Integer	weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
Medu	Integer	mother's education (0 - none, 1 - primary education (4th grade), 2 (5th to 9th grade), 3 (secondary) education or 4(higher education))
Fedu	Integer	father's education (0 - none, 1 - primary education (4th grade), 2 (5th to 9th grade), 3 (secondary) education or 4(higher education))
Mjob	Integer	mother's job ('teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at home' or 'other')
Fjob	Char	father's job ('teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at home' or 'other')
Failure	Integer	number of past class failures (numeric: n if 1<=n<3, else 4)
G1	Integer	first period grade
G2	Integer	second period grade
G3	Integer	final grade

The variables of dataset shown in figure1, figure 2 is the summary information of dataset and figure 3 shows a description of attributes of dataset

```
> names(studentData)
 [1] "Medu"      "Fedu"      "Mjob"      "Fjob"      "studytime" "failures"
 [7] "higher"    "absences"  "G1"        "G2"        "G3"
>
```

Fig. 1. Variables' Names

```
> summary(studentData)
      Medu            Fedu            Mjob           Fjob
 Min.   :0.000   Min.   :0.000   at_home : 59   at_home : 20
 1st Qu.:2.000   1st Qu.:2.000   health  : 34   health  : 18
 Median :3.000   Median :2.000   other   :141   other   :217
 Mean   :2.749   Mean   :2.522   services:103   services:111
 3rd Qu.:4.000   3rd Qu.:3.000   teacher : 58   teacher : 29
 Max.   :4.000   Max.   :4.000
   studytime       failures      higher       absences
 Min.   :1.000   Min.   :0.0000   no : 20   Min.   : 0.000
 1st Qu.:1.000   1st Qu.:0.0000   yes:375   1st Qu.: 0.000
 Median :2.000   Median :0.0000             Median : 4.000
 Mean   :2.035   Mean   :0.3342             Mean   : 5.709
 3rd Qu.:2.000   3rd Qu.:0.0000             3rd Qu.: 8.000
 Max.   :4.000   Max.   :3.0000             Max.   :75.000
       G1              G2              G3
 Min.   : 3.00   Min.   : 0.00   Min.   : 0.00
 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
>
```

Fig. 2 .Dataset Summary

```
> str(studentData)
'data.frame':   395 obs. of  11 variables:
 $ Medu     : int  4 1 1 4 3 4 2 4 3 3 ...
 $ Fedu     : int  4 1 1 2 3 3 2 4 2 4 ...
 $ Mjob     : Factor w/ 5 levels "at_home","health",..: 1 1 1 2 3 4 3
 3 4 3 ...
 $ Fjob     : Factor w/ 5 levels "at_home","health",..: 5 3 3 4 3 3 3
 5 3 3 ...
 $ 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 ...
 $ absences : int  6 4 10 2 4 10 0 6 0 0 ...
 $ 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 ...
>
```

Fig. 3. Dataset's Structure

4.3 Data Analysis

The dataset is now ready to use in prediction model but it needs to analize and describe more about their relations with the student grades to understand the grades distribution and I visualize its distribution on the number of students with grades and figure 4 shows that result for visualization.

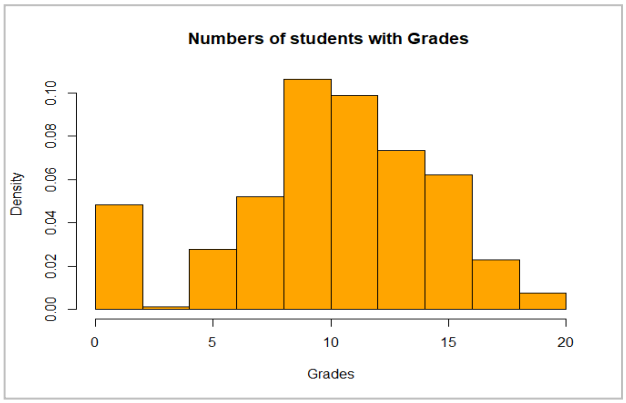

Fig. 4. Distribution of Final Grades

In this section, I tried to understand the relations between grades and the variables and measure the correlations of the factors to show its affect on the grade prediction. I create dummy variables for categorical variables to test its correlation to G3.

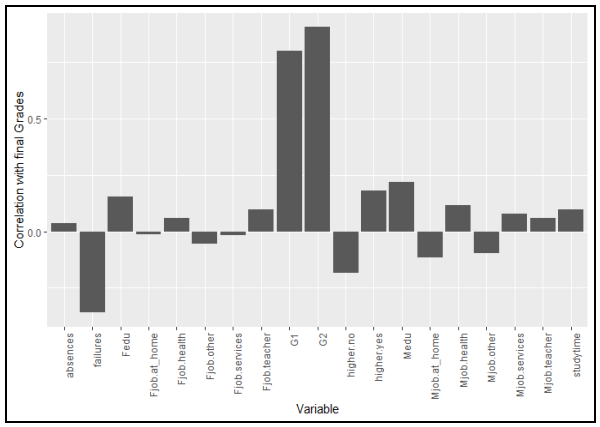

Fig. 5. Visualize the Correlation between Final Grade and Variables

The correlation figure 5 shows that there is no high correlations between the final Grade and variables unless the G1 and G2. The heatmaps show higher seven variabels to each other well I focuses only in their correlation to final Greads G3.

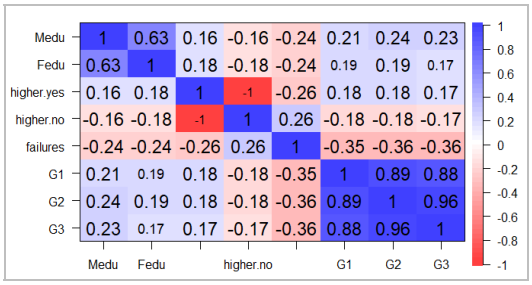

Fig. 6. The Heatmaps Show Higher Seven Variables to each other

A linear regression model was built with those variable to add value to the final grade, it will make difference with the highly correlated variabels first and second period grades 'G1' and G2 since the final grads is correlated without second grade G2 to show the value of adding value to the model not just prediction depends on the previous grades.

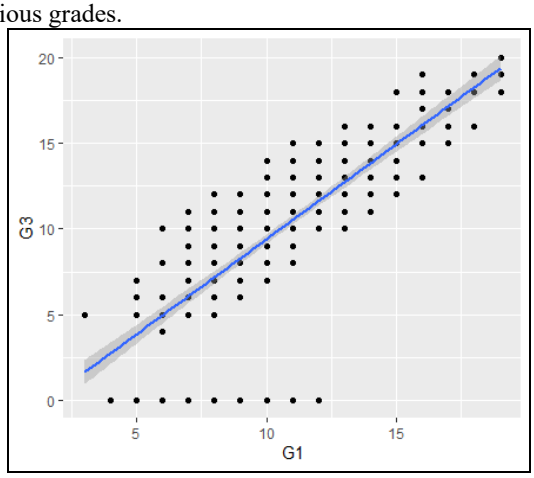

Fig. 7. Correlation of G1,G3

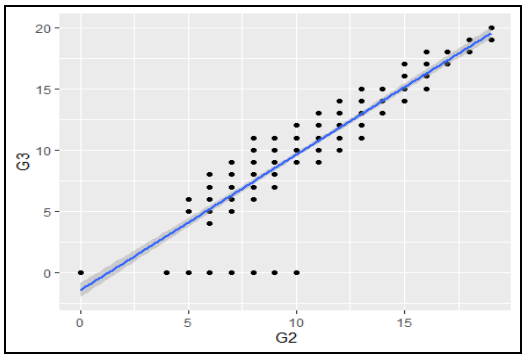

Fig. 8. Correlation of G2,G3

Final grades or G3 will modeled by what I assumed in the hypothesis, which are absences, study time, Medu, Fedu, Mjob, Fjob and Failure attrebutes, also becouse it is hard to predict the final grades without first period grade and the second period G2 so I will add them to the model.

4.4 Generating Training And Test Dataset

This stage is important when we build model in machine learning, it is helpful to train the model on the part of dataset. Well, other part is to test or validate the model here; we split our data where 75% of the dataset will uesd in training the model and 25% to validate the model. Training dataset is used to build model in which the variabels of the model may be adjusted and then the resulted model is applied on the test dataset to provide an unbiased evaluation of a model that build on the training dataset.

4.5 Model Generation

The model used for preduction is linear regression model which is the common model used for predictions. Linear regression model is mathematical equation to approximate reality then make predictions from this approximation.
First I build model using all variabels thet I asummed in the hypotesis it is effect on the student performance

```
> summary(model1)

Call:
lm(formula = G3 ~ G1 + G2 + failures + Medu + Fedu + higher +
    studytime + absences + Mjob + Fjob, data = training)

Residuals:
    Min      1Q  Median      3Q     Max
-9.1009 -0.4725  0.2812  0.9556  3.5175

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.66023    0.90228  -1.840   0.0668 .
G1            0.16532    0.06755   2.447   0.0150 *
G2            0.96637    0.05958  16.219   <2e-16 ***
failures     -0.24384    0.17896  -1.363   0.1741
Medu          0.10927    0.16858   0.648   0.5174
Fedu         -0.11358    0.14433  -0.787   0.4320
higheryes    -0.08494    0.56367  -0.151   0.8803
studytime    -0.24417    0.14418  -1.694   0.0915 .
absences      0.02957    0.01326   2.230   0.0265 *
Mjobhealth    0.43667    0.60807   0.718   0.4733
Mjobother     0.18308    0.38006   0.482   0.6304
Mjobservices  0.30048    0.42872   0.701   0.4840
Mjobteacher   0.42623    0.53079   0.803   0.4227
Fjobhealth    0.69358    0.71777   0.966   0.3347
Fjobother     0.26794    0.48693   0.550   0.5826
Fjobservices -0.14758    0.51023  -0.289   0.7726
Fjobteacher   0.19206    0.66225   0.290   0.7720
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.923 on 281 degrees of freedom
Multiple R-squared:  0.8261,    Adjusted R-squared:  0.8162
F-statistic: 83.45 on 16 and 281 DF,  p-value: < 2.2e-16
```

Fig. 9. Summary of the Model

The p-value of the model is less than the significant level which is 0.05 and the adjusted R-squered is higher than 0.7. The model works but when we read the pr(>|t|) there are some variables not less the significant level becouse of that the model need refinement.

Second model was built on the variables that have p-value at significant level. The model become:

```
> summary(model2)

Call:
lm(formula = G3 ~ G1 + G2 + studytime + absences, data = training

Residuals:
    Min      1Q  Median      3Q     Max
-9.3953 -0.3716  0.3068  1.0236  3.1748

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.71581    0.46006  -3.730  0.00023 ***
G1           0.16856    0.06423   2.624  0.00914 **
G2           0.98867    0.05665  17.453  < 2e-16 ***
studytime   -0.23057    0.13749  -1.677  0.09460 .
absences     0.02944    0.01291   2.280  0.02335 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.915 on 293 degrees of freedom
Multiple R-squared:  0.8201,    Adjusted R-squared:  0.8177
F-statistic:   334 on 4 and 293 DF,  p-value: < 2.2e-16
```

Fig. 10. Summary of the Final Model

Depend on this dataset and the model's factors can effect on the students performence are the study time and the absences. So we can reject H1, H4 and H5.

5. Conclusion And Future Work

Predicting students performance is mostly useful to help the educators and learners improving their learning and teaching process. In this research, I test the factors effect on the student performance depend on existing dataset which contain many factors can effect on the final grade. The selected factors were the attention to take higher education, absences, study time, mother education level, father education level, mother job, father job and the number of failures on the past. I test their effects using linear regression model on respond variable final grade G3. The result obtained was only absences and study time which can effect on students' performance. The dataset is old and need more detailes, and more data can be collected through surveys with local school, college or universit and will be more effective and give rbost result and can applied on saudi students situations.

6. References

[1] Al-Kabi, M. N., & Jirjees, J. M. (2019). Survey of Big Data applications: health, education, business & finance, and security & privacy. Journal of Information Studies & Technology (JIS&T), 2018(2), 12.

[2] UCI, (2014) Student Performance Data Set https://archive.ics.uci.edu/ml/datasets/student+performance

[3] Sampasa-Kanyinga, H., Chaput, J. P., & Hamilton, H. A. (2019). Social media use, school connectedness, and academic performance among adolescents. The journal of primary prevention, 40(2), 189-211.

[4] Ahmed, A.B.E.D. and Elaraby, I.S., 2014. Data Mining: A prediction for Student's Performance Using Classification Method. World Journal of Computer Application and Technology, 2(2), pp.43-47.

[5] Saa, A. A. (2016). Educational data mining & students' performance prediction. International Journal of Advanced Computer Science and Applications, 7(5), 212-220

[6] Saxena, K., Jaloree, S., Thakur, R.S., & Kamley, S. (2018). Linear Regression Technique for Student Academic Performance Prediction.

[7] Harwati, Ardita Permata Alfiani, and Febriana Ayu Wulandari. "Mapping Student's Performance Based on Data Mining Approach (A Case Study)." Agriculture and Agricultural Science Procedia 3 (January 1, 2015): 173–77. doi:10.1016/j.aaspro.2015.01.034

[8] Ngai E.W.T. A,, Li Xiu B, Chau D.C.K. (2009) Application Of Data Mining Techniques In Customer Relationship Management: A Literature Review And Classification, Journal Of Expert Systems With Applications 36 (2009) 2592–2602

[9] Raval M Kalyani (2012) Data Mining Techniques, International Journal Of Advanced Research In Computer Science And Software Engineering Volume 2, Issue 10.

[10] Ridwan Mujib, Suyono Hadi, M. Sarosa 2013 Penerapan Data Mining Untuk Evaluasi KinerjaAkademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier, Jurnal EECCIS Vol.7, No. 1

[11] Romero, C.; Ventura, S. Data mining in education. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 2013, 3, 12–27. https://doi.org/10.1002/widm.1075

[12] Khan, S., & Alqahtani, S. (2020). Big Data Application and its Impact on Education. International Journal Of Emerging Technologies In Learning (IJET), 15(17), pp. 36-46. http://dx.doi.org/10.3991/ijet.v15i17.14459

[13] Zhang, Y., Ghandour, A., & Shestak, V. (2020). Using Learning Analytics to Predict Students Performance in Moodle LMS. *International Journal Of Emerging Technologies In Learning (IJET), 15*(20), pp. 102-115. doi: http://dx.doi.org/10.3991/ijet.v15i20.15915

Dr. Shakir Khan received his BSc, MSc and PhD in computer science in 1999, 2005 and 2011 respectively. He is member of the International Association of Online Engineering (IAOE) and IEEE, He is currently working as Associate Professor at College of Computer and Information Sciences in Imam Mohammad Ibn Saud Islamic University, Riyadh (Saudi Arabia). His research interest is Big Data, Data Science, Data Mining, Machine Learning, Internet of Things (IoT), and eLearning, Artificial Intelligence, Emerging Technology, Open-Source Software, Library Automation and Mobile / Web Application. He published many research papers in international journals and conferences in his research domain. He has around 15 years of teaching and research experience in India and Saudi Arabia. Dr. Khan is teaching bachelor and master's degree courses in the college of computer at Imam University. He is reviewer for many international journals.