Data Science Research on Education: The Analysis of Factors may Influence on Student Performance

Fenghuan Li
Xiamen Shuqiao Technology Co., Ltd., Xiamen 361115, China.

Abstract: In this Data science research on Education, it analyses the alcohol consumption, parent's education, study time and other factors may influence on student performance.

Keywords: Student Performance; Education; Alcohol Consumption; Parent's Education; Study Time

1. Introduction

Education is one of the most important factors influences the life. As Nelson Mandela says "Education is the most powerful weapon which you can use to change the world" (Ratcliffe, 2017). A lot of research and studies has conducted in education over the years. Education has been widely discussed in parents, teachers, schools, societies, and government authorities. The quality of the education is also attracted extensive attention worldwide. Academic performance normally be considered in reflect of successful level of scholastic or academic work, and the achievement combines the student's capacity and performance, it is multidimensional (Sharma & Jha, 2014). Therefore, this brings our interest on doing research and find more insights about student performance in education. Some of the factors from individuals will be used to find the potential relations on their performances, which could help us to find more insights of potential factors may influence the outcome of student performance. In this research, it will conduct analysis of the influence of alcohol consumption, family background, study time and other factors on student performance.

2. Literature and theory

In the modern society, measuring the academic performance is an important part in the learning progress, and gives the educator the idea on the progress of education goal, it is generally expecting the student to have higher achievement (Sharma & Jha, 2014). It is also being study that there are a lot of factors will impact on student performance (Harsha Aturupane, 2011).

The study on the effects of college drinking on study hours impact academic performance through the model analysis, found the higher frequency of drinking alcohol does not toward to the effect of grade point average, but study hours (Wolaver, 2002). However, research and study has reported that the underage consumption of alcohol will disrupts the learning and intellectual development, and this impact may continue to affect the individual into adulthood (Donald W. Zeigler, 2005). It has been found in study that student access to alcohol will result in reduce the average academic performance (Joung Yeob Ha, 2019). The parent's education is also considered as the factor effects on student's performance (Sharma & Jha, 2014). Based on the study of family background and student's performance, it has found that the parent's education background are the key determinants of student performance, and parent's education has positive impact on student's scores (Guimaraes & Sampaio, 2013). More time spend on study is considered as a factor influences on student's performance. It is being found that, there were significant effect of lecture attendance and study time on study performance, with better achievement on one additional hour of school attendance (Andrietti & Velasco, 2015).

3.Setup

In this research, it will focus on the alcohol consumption, family background, study time and other factors may influence on the student's performance. It will be further defining an experiment setup for the relationship between grades and all other variables, this

will further describe in approach section. Not only subject to these two questions, but more insights about student performance will also be discovered.

Hypothesis testing 1

H_0: The alcohol consumption on the weekdays does not impact the student grade

H_1: The alcohol consumption on weekdays impacts the student grade

Hypothesis testing 2

H_0: Mother's education level does not impact the final grade received by student.

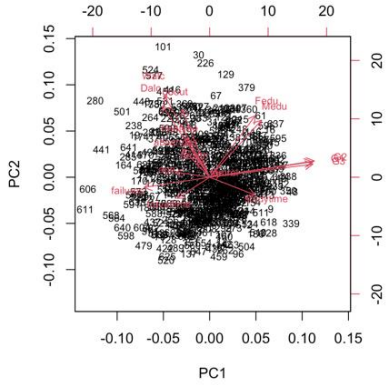
H_1: Mother's education level impacts the final grade received by student.

Hypothesis testing 3 (For classification)

H_0: All input factors have jointly no effect on students' grades

H_1: At least one of the independent variables have explanation power towards the variation in students' background status.

Quantify reliability will be conduct in python and R studio, significance testing, regression r square will be conducted, by implement the significance testing result, research question can be answered. More detail discussion about selected models training and testing results will be conducted and explored. Generally, the dataset is clean and complex, the initial checking the potential nulls checking have conducted, and there is no major transforming changed in the dataset.



Plot 1 - *Pre-processing scaled PCA*

Table 1 - Key Summary of dataset		
Factors	Category	Description
Dataset Dimension		649 x 45
G3 (Final grade)	Mean	11.91
G3Binary	Binary	1 Above mean
		0 Below mean
Dalc (Weekday alcohol consumption)	Numerical	1 to 5
Walc ((Weekend alcohol consumption)	Numerical	1 to 5
Medu (Mother's education)	Numerical	0 to 4
Fedu (Father's education)	Numerical	0 to 4
Studytime (weekly studytime)	Numerical	1 to 10

4. Approach

The dataset was download from Kaggle with size 108 KB, in the project we will focus on the larger data set student-por.csv since it covers more student within two schools. Initially, the student performance was measured by G3 as the numerical factors, in this research we will first perform numerical measurement by using the simple liner regression, multiple liner regression and random forest model to do the analysis.

The proposed Simple liner regression model:

$$G3 = \beta_0 + \beta_1 * Dalc + \varepsilon$$

$$G3 = \beta_0 + \beta_1 * Medu + \varepsilon$$

Proposed multiple liner regression model:

$$G3 = \beta_0 + \beta_1 * Dalc + \beta_2 * Medu + \varepsilon$$

$$G3 = \beta_0 + \beta_1 * Dalc + \beta_2 * Walc + \beta_3 * Medu + \beta_4 * Fedu + \beta_5 * famrel + \beta_6 * freetime + \beta_7 * goout + \beta_8 * health + \beta_9 * absences + \beta_{10} * traveltime + \beta_{11} * studytime + \varepsilon$$

The simple linear regression model has conducted, the variable for the null hypotheses can be evaluated by using the p-value. All of the features will be input into multiple linear regression model at the beginning, after the evaluation on each parameter. By looking at the importance of the features, ten of the relatively important and interested features will input into the multiple linear regression model. These factors are Walc (weekday alcohol consumption), Medu (Weekend alcohol consumption), Medu (mother's education), Fedu (father's education), famrel (quality of family relationship), freetime (free time after school), goout (going out with friends), health (current health status), absences (number of school absences), traveltime (home to school travel time), studytime (weekly study time). The Random Forest is also conducted based on G3 for this numerical analysis.

In order to find more insights in this research, we decided to classify the student's final grade into two groups, named 1 and 0, 1 representing above the average and 0 representing below the average. Therefore, this dataset can be used to do the classification analysis with more models can be explored. G3, the numerical final grade will be then converted into the G3Binary (above or below average) which can be used in the classification analysis.

The logistic regression, decision tree and random forest can be further conduction on this classification research. In order to widely explored the dataset on what are the factors depend the student's final grade, the numerical features used in the multiple linear numerical analysis stage will then be used again in the classification models. The logistic regression will be used and based on the binomial distribution. The dataset will be split into 80% for training and 20% for testing for the model, decision tree and random forest model will be tested and measure by F1 and the accuracy will be evaluated. These methods could be further used in the comparison.

Results and analysis

Part 1 – For the regression results

Proposed simple liner regression model	
$G3 = \beta_0 + \beta_1 * Dalc + \varepsilon$	H_0: The coefficient in the model is equal to 0. H_1: The coefficient in the model is not equal to 0. P value(β_0) is 2e-16 less than 0.05, statistically significant, reject the null. P value(β_1) is 1.43e-07 less than 0.05, statistically significant, reject the null. P value of the model is 1.432e-07, model is statistically significant. R square of the model is 0.04, which means the variable in the regression model does not explain much of the variance in dependent variable.
$G3 = \beta_0 + \beta_1 * Medu + \varepsilon$	H_0: The coefficient in the model is equal to 0. H_1: The coefficient in the model is not equal to 0. P value(β_0) is 2e-16 less than 0.05, statistically significant, reject the null. P value(β_1) is 5.75e-10 less than 0.05, statistically significant, reject the null. P value of the model is 5.752e-10, model is statistically significant. R square of the model is 0.06, which means the variable in the regression model does not explain much of the variance in dependent variable.
Proposed multiple liner regression model	

G3 = β_0 + β_1 *Dalc + β_2 *Medu + ε	H_0: The coefficient in the model is equal to 0. H_1: The coefficient in the model is not equal to 0. P value(β_1) is 7.72e-08 less than 0.05, statistically significant, reject the null. P value(β_2) is 3.14e-10 less than 0.05, statistically significant, reject the null. P value of the model is 2.467e-15, model is statistically significant. R square of the model is 0.10, which means only 10% the variance in dependent variable is explained by the model.
G3 = β_0 + β_1 *Dalc + β_2 *Walc + β_3 *Medu + β_4 *Fedu + β_5 *famrel + β_6 *freetime + β_7 *goout +β_8 *health+β_9 *absences+ β_{10} *traveltime + β_{11} *studytime + ε	H_0: The coefficient in the model is equal to 0. H_1: The coefficient in the model is not equal to 0. P value(β_1) for Dalc is 0.00386 less than 0.05, statistically significant, reject the null. P value(β_2) is 0.00659 less than 0.05, statistically significant, reject the null. P value of the model is 2.2e-16, model is statistically significant. R square of the model is 0.16, which means only 16% the variance in dependent variable is explained by the model.
Since the R square for all the model seems very low, we have conducted a multiple liner regression model with all the available features, the result shown at appendix that the all features model R-square is just about the 33%, however, considered the R square will eventually increase by add so many features, 33% is considered very low. The above proposed model is the model tunned and partly selected relevant the importance features.	
Random Forest in Multiple liner model	
By Using the same input of the last multiple liner regression model which we have been selected some of the features from all available features model, by inputing Dalc, Walc, Medu, Fedu, famrel, freetime, gout, health, absences, traveltime, studytime, and the number of tree is 500, the mean of squared residuals is 8.957, 14.04% variance explained. Although, 14% is not high but it improved from the other methods. among these 11 features. The studytime is the top important variable that impact on student final grade the percentage increase in MSE is about 14%. Follow by Medu, Fedu, Dalc and Walc, which means mother and father's education background and weekdays and weekend's alcohol consumption are important to student grade. Family background took over 20% and alcohol consumption took nearly 20% which highly relevant to our project research.	

Part 2 - For the classification results:

In the second part of analysis, we separated the students into two group into binary with 1 presenting who above average, and 0 representing below the average. In such way, as it become binary outcomes, the classification analysis can be further conducted. We can further explore whether the jointed factors such as alcohol consumption, family education background, time effect student study all of these features that used in the first stage multiple regression analysis, will also remain the same and fit into the classification model. For the dataset, we will randomly split the dataset into 80% training, and 20% for testing. The logistic regression, decision tree classifier, random forest classifier will be conducted, the accuracy, precision, recall, F1 score of each model will be calculated for evaluation of difference. Consider to the grade prediction task, precision and accuracy measurement will be focused on model evaluation.

5. Logistic regression

To begin with, we use the logistic regression to fit on the model. First to apply the grid search to find the best parameters for optimizing the logistic regression model. The best parameter C has set to 1 and penalty sets to I2. For the result, the weighted seems health as the figures are seems quite balance, as there is no single attribute is really high or low. In this model, the accuracy is 0.63 and the F1 score is 0.49 of the student who is below the average. However, the recall, correct gold label is only 0.39 for the student below average prediction is considering very low. Recall is a significant figure on the evaluation, low recall rate suggest the model is not very suitable.

6. Decision tree

Further, the decision tree model is performed. The max_depth = 2 will be set in the performance of the decision tree. The decision

tree model has improved the accuracy to 0.6, the F1 score for the student below to average prediction is 0.63 and the precision is only 0.54. Compared to the Logistic regression model it has both improved in the accuracy and F1, this suggesting that decision tree classifier has better performed than Logistic regress model.

7. Random forest

The third model we look at the random forest model, by setting the number of trees by 200. By computing the outcome, the model has the highest accuracy 0.64 among these models we have been evaluated. The result, the weighted seems still quite balance, as there is no single attribute is very high or low. Compared the decision tree model, the accuracy of the random forest model has slightly improved. However, for the F1 score of perdition of the student below the average is lower to 0.52, and the precision is 0.66. But the prediction for the student above the average has increased in F1 score which reached 0.71 with the precision is 0.63. Overall, the random model seems the best among three models.

8. Conclusion

In the classification part, we used the three different models for analysis, we also tested and evaluated the classification result. However, the accuracy for these models was not high, this may not be able to expect the input features could impact on student performance, that might be caused by the limitation on selected features, many other factors may need to bring into the consideration, more data input will help us to improve the accuracy. Also, not all of factors impacting of student performance in life can be measure and record in the databased, and these could be the limitation of the dataset.

To sum up, answer to our hypothesis which set up at the beginning of the research, according to the p value in the liner regression both of the null hypothesis is rejected, which confirmed to us that the alcohol consumption in weekdays will impact on student's final grade. The mother's education background is also significantly influence on student's final grade. The coefficient significant testing in multiple regression model that also confirm that weekday's alcohol consumption and mother's education level will impact on student's performance. However, there are some limitations on model as the liner regressions do not fit the student performance quite well, this may be due to there are many other factors may influence on students' performance, the dataset maybe is relatively not large. More of the data input may improve the accuracy, and model could be further tunned.

References

[1] Andrietti, V., & Velasco, C. (2015). Lecture Attendance, Study Time, and Academic Performance: A Panel Data Study. The Journal of Economic Education(46:3), 239-259.

[2] Donald W. Zeigler, P. C. (2005). The neurocognitive effects of alcohol on adolescents and college students. Preventive Medicine, 40(1), 23–32.

[3] Guimaraes, J., & Sampaio, B. (2013). Family background and students' achievement on a university entrance exam in Brazil. Education Economics, 21:1, 38-59.

[4] Harsha Aturupane, P. G. (2011, Apr 26). The impact of school quality, socioeconomic factors, and child health on students' academic performance: evidence from Sri Lankan primary schools. Education Economics, 21(1), 2–37.

[5] Joung Yeob Ha, A. C. (2019). Legal access to alcohol and academic performance: Who is affected? Economics of Education Review, 72, 19–22.

[6] Nelson Mandela. In Ratcliffe, S. (Ed.), Oxford Essential Quotations. : Oxford University Press. Retrieved 27 Sep. 2023, from https://www.oxfordreference.com/view/10.1093/acref/ 9780191843730.001.0001/q-oro-ed5-00007046.

[7] Sharma, G., & Jha, M. (2014). Academic Performance in Relation to Parents' Education, Institution and Sex. Journal of Psychosocial Research, 9(1), 171-178.

[8] UCI Machine Learning. (2020, 11 21). Student Alcohol Consumption. Retrieved from Kaggle: https://www.kaggle.com/uciml/student-alcohol-consumption/download.

[9] Wolaver, A. M. (2002). Effects Of Heavy Drinking In College On Study Effort, Grade Point Average, And Major Choice. Contemporary Economic Policy, 20(4), 415–428.