The 3rd International Conference on Educational Innovation and Philosophical Inquiries (ICEIPI 2022)
DOI: 10.54254/2753-7048/2/2022384

Prediction of Final Academic Performance of Secondary School Students Based on Alcohol Consumptions and Diverse Backgrounds

Yu Xiong[1,a,*], Yuchen Ding[2], Jiayi Li[3], Wenting Zhang[4], Shutong Ni[5]

[1]School of Computer Science and Engineering, The Ohio State University, Columbus, Ohio 43210, U.S.
[2]Department of Physics, Applied Physics and Astronomy, Rensselaer Polytechnic Institute, Troy, NY 12180-3590, U.S.
[3]School of Banking and Finance, University of International Business and Economics, Beijing, 100029, China.
[4]Jericho Senior High School, Jericho, NY 11753-1202, U.S.
[5]Ulink college, Guangzhou, Guangdong 511458, China
a. xiongyu0124@gmail.com
**Corresponding author*

Abstract: The power of machine learning is prompting more educators to consider applying the technology to education. Also, alcoholism among students has become an important problem in society. Based on data from two US secondary schools, this paper investigates the impact of attributes including alcohol consumption, on students' grades, and aims to predict students' final grades in Portuguese. The normal distribution index was used as a reference to analyze and clean the data. The visualization of the data to explore which attributes were most correlated with final grade has been achieved. In this paper, ridge regressor, decision tree regressor, decision tree classifier, and KNN classifier were used as training models, with MSE, MAE, r2, and accuracy as testing indexes and cross validation to verify results. The final results show that alcohol consumption has a significant negative impact on students' academic performance.

Keywords: student education, alcohol consumption, machine learning, decision tree regressor, decision tree classifier

1. Introduction

Addiction to alcohol is harming students' health and academic performance. Neuroscientists at the University of California, San Diego, who've studied the brains of teens who binge drink, found brain damage that they believe reduces the attention span of boys and diminishes girls' ability to process and understand visual information [1].

This paper is mainly concerned with three directions. First, 10 attributes that have the greatest impact on student achievement according to correlation will be listed. This will serve as a useful reference for students who want to improve their grades in the future. Whether binge drinking really affected students' G3 (final grade) will become the second question. Dalc (workday alcohol consumption) and Walc (weekend alcohol consumption) will be considered important indicators. Third, we wanted to find out whether G1 (first period grade) and G2 (second period grade) had a significant impact on G3. All three directions will proceed simultaneously.

DOI: 10.54254/2753-7048/2/2022384

After data cleaning, this work used four different methods, namely ridge regressor, decision tree regressor, decision tree classifier and KNN classifier, to train the models. MSE, MAE, and r2 were used to measure the accuracy of the predictions that could be made by training models with ridge regressor and decision tree regressors. Similarly, accuracy was used to measure the results of the decision tree classifier and KNN classifier. Cross validation was used to verify the accuracy of the results.

Combined with the third analysis, which listed the 10 attributes that had the greatest impact on student achievement, the training results in this work showed that binge drinking did have a negative impact on student achievement. This paper found that workday alcohol consumption was the first of the 5 attributes that had the greatest negative impact on the model. Last with G1 and G2 included as attributes in the model, the training results obtained by the 4 methods are satisfactory. It can be concluded that alcohol consumption, G1, and G2 play important roles in predicting student achievement.

2. Background and Related Works

Many researchers have found that drinking alcohol has negative consequences on students' performance. Justina Ifeoma Ofuebe used the Z test with 0.05 level significance to find both female and male undergraduate students agree that the alcohol consumption has negative effects on their behaviors in the universities in South-East, Nigeria [2]. Chimwemwe Tembo used the multiple logistic regression analysis to state that the high level of alcohol consumption correlates with poor academic behavior and low mental health [3]. Mahmood R. Gohari used the first-order autoregressive multinomial logistic regression to conclude that quitting or decreasing on alcohol consumption may improve the students' academic behaviors in secondary schools [4].

However, other researchers might have controversial conclusions on the relation between alcohol consumption and academic performance. A. I. Balsa used the Add Health's longitudinal design to estimate how alcohol consumption affects learning in high school. They came out with the conclusion that alcohol consumption has very small negative effects on male's grades, but it has negligible effects on female's grades [5]. M. Lopez-Frias used the Multiple logistic regression analysis to find that although the low grade came with high alcohol intake, they couldn't draw the conclusion of cause-and-effect relationship between alcohol intake and academic behavior. There are many other factors that influence a student's academic achievement [6].

3. Dataset

The dataset, called Student Alcohol Consumption, has 33 attributes with 649 samples. The attributes in this dataset include the student's school, gender, family information. Attributes such as Dalc and Walc are ordinal types. Some attributes, such as school and sex, are of the nominal type. Other attributes, such as age and G3, are of interval type. Among these attributes, G3 is the output and this work want to study the influence of these attributes on students' final grade on Portuguese.

3.1. Data Cleaning Processes

3.1.1. Samples with a 0 Score

Samples with a 0 score in G1 and G3, possibly caused by dropping out or transferring to another school, may interfere with our training. To avoid this, this research removed all these samples.

DOI: 10.54254/2753-7048/2/2022384

3.1.2. Sparse Samples

Because the volume of samples aged 20 or over is limited, it is difficult to draw meaningful conclusions. Thus, this work decided to group samples aged 20 or over together. This work also groups samples whose absences value is 15 or over together for the same reason.

3.1.3. Attributes with Nominal Type

Convert nominal type (String/Boolean) to an exploitable type (Int). This will allow the gaps between data to be calculated and compared by machine. Take Higher (the willingness of students to pursue higher education) as an example. The value of Higher can only be yes or no, and this work divided higher into two attributes higher_yes and higher_no. If higher is yes, higher_yes is 1 and higher_no is 0. If higher is no, higher_no is 1 and higher_yes is 0.

3.1.4. Attributes with Less Reference Value

Four attributes named famsize, Pstatus, famsup and romantic are discarded, because the mean of G3 is almost unchanged when the four attributes change. This will help the machine focus on more meaningful data and attributes. The box plot of them is shown in Figure 1.

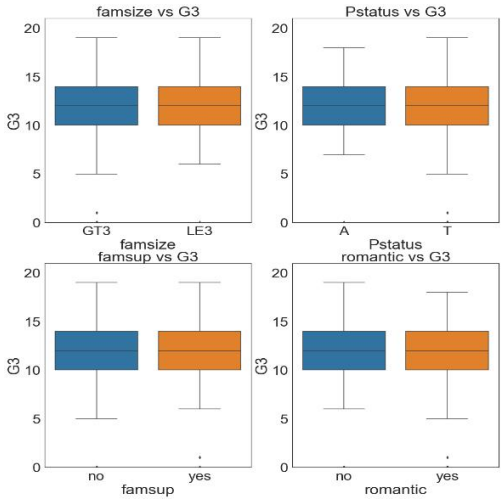

Figure 1: Box plot of G3 in terms of famsize, Pstatus, famsup and romantic.

3.2. Attributes Analysis

3.2.1. Correlation Ranking

In Table 1 and Table 2, there lists the five attributes with the greatest positive correlation with G3 and the five attributes with the greatest negative correlation with G3, respectively. This work will focus on attributes related to students' scores and drinking behavior in the following in the following parts.

Table 1: The five attributes with the greatest positive correlation with G3.

correlation	G2	G1	higher_yes	Medu	Studytime
G3	0.933784	0.876187	0.337096	0.273355	0.246042

DOI: 10.54254/2753-7048/2/2022384

Table 2: The five attributes with the greatest negative correlation with G3.

correlation	failures	Higher_no	School_MS	absences	Dalc
G3	-0.388053	-0.337096	-0.220368	-0.212730	-0.211458

3.2.2. Attributes Dalc & Walc

Walc is not in the Table 1 and 2, but it is the sixth most correlated attribute among the attributes that have a negative correlation with G3. This part discusses the relationship between Dalc, Walc, and G3 together. Dalc indicates workday alcohol consumption (numeric: from 1 - very low to 5 - very high), and Walc indicates weekend alcohol consumption (numeric: from 1 - very low to 5 - very high). As can be seen from Figure 2, 3, 4, and 5, the less alcohol students consume on weekdays and weekends, the better their grades are likely to be, but weekend alcohol consumption has a smaller impact than Workday alcohol consumption.

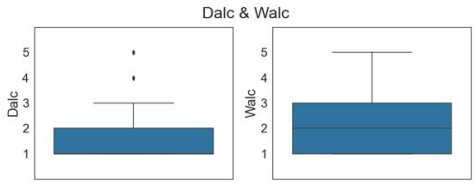

Figure 2: Box plot of attributes Dalc and Walc.

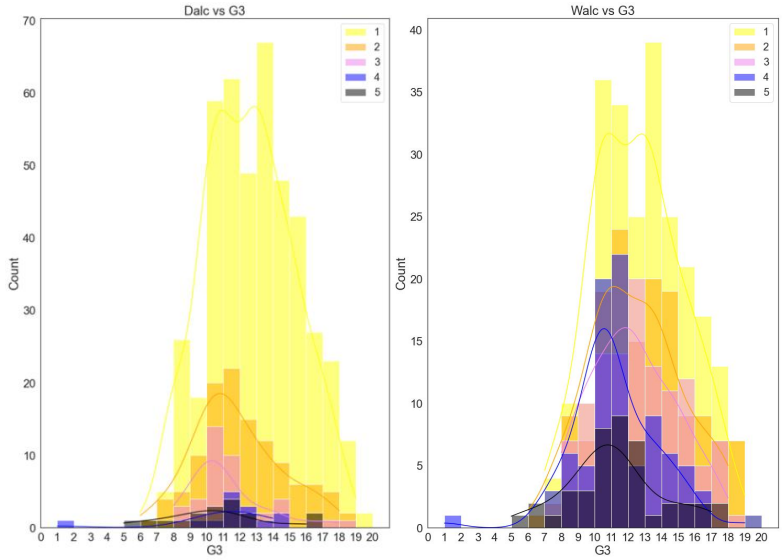

Figure 3 & 4: Distribution of G3 divided in terms of Dalc & Walc.

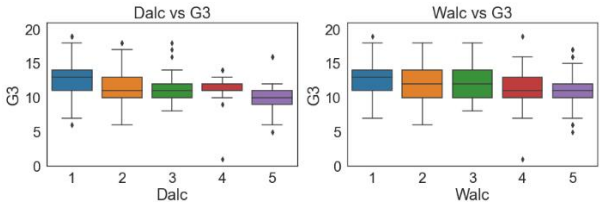

Figure 5: Boxplot of G3 divided in terms of Dalc and Walc.

DOI: 10.54254/2753-7048/2/2022384

3.2.3. Attributes G1 & G2

There exist very strong correlations between G1 and G3, and between G2 and G3. If one creates the boxplots for G1 versus G3 and G2 versus G3 for each value. They both show the direct positive relationship with G3 in the y-axis. However, the degree of contribution for each score in G1 and G2 compared with G3 is not very strong. Instead, this work groups the scores into 3 categories, which include low, medium, and high scores. As shown in Figure 6 and 7, the scores from 3 to 8 are the low score group, because we drop the 0 score and none of the student have 1 or 2 score on G1 and G2; the scores from 9 to 14 are the medium score group; and the scores from 15 to 20 are the high score group, because the highest score is 20. The boxplots diagrams show that the distributions of three groups are approximately normal and as the scores in G1 and G2 increase, the scores in G3 increase as well. Based on Figure 6 and 7, one observes that the low-grade group in both G1 and G2 has mean and median of approximately 8 to 9, the medium grade group has that of approximately 12, and the high-grade group has that of approximately 16. One can assume if a student does well on G1 and G2, he has a high possibility of getting higher grades on G3.

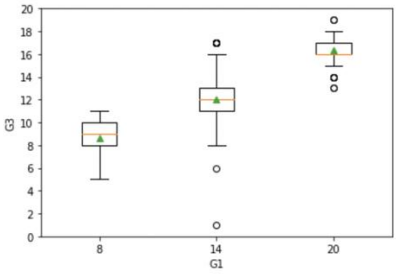

Figure 6: Box plot of G3 of samples in each grade group of G1.

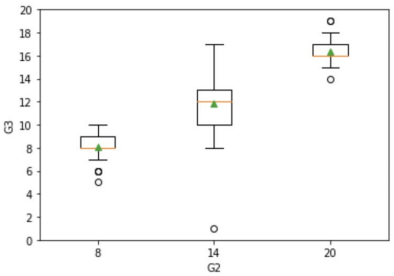

Figure 7: Box plot of G3 of samples in each grade group of G2.

4. Methods

Based on the previous pre-processing steps, the processed dataset is used in the prediction. Because most of the attributes, including age, are discrete nominal values, one hot encoding is used to digitize them. In this way, each class under an attribute is used to create a new column with true or false value associated with each sample. In our prediction, linear regression[7], decision tree regressor and classifier[8], and KNN classifier[9] will be considered as the classifiers. We also utilize cross validation[10] as the validator.

By looking at the coefficients of linear regression, we find that more than half of the coefficients sometimes diverge, which may be caused by that the input matrix is too sparse, so a similar model, Ridge, with a regularization term is used instead. Ridge regression is a model tuning method that is used to analyze data suffer from multicollinearity[11]. With the regularization, one can control the coefficients in the normal range, although this does not contribute too much to those metrics score.

DOI: 10.54254/2753-7048/2/2022384

For classifiers, with the knowledge that we can only get a low accuracy, the parameters of classifiers are modified in order to attain a training accuracy 0.5 or below.

One already knows from the EDA that students' final grade has a high positive correlation with those two previous exam scores, so this work splits the goal into two parts, try to use students' general background to improve the prediction of final score with the data set that include their previous test scores, and try to predict students' final grade based only on their background without using previous scores. All four models are performed on both cases.

Because using the 20 levels of final grade as the classes in the classification is too much, this work considered group the students' grade into three or four classes, for example, less than 10, 10 to 14, and more than 14, as three classes. However, the accuracy results we get from the classifiers are not much higher than the probability to guess by assuming all the students are in the class that has the most students, which is, in the three classes cases, class 10 to 14. Thus, this operation is not meaningful. In addition, because the predictions we got from regression model are floats, but the grade only takes integers, this work tried to round the predictions to the nearest integer and then calculate the metrics. Nevertheless, the improvement was not significant, so this work decided to stick to the cross validation as metrics scores without rounding.

Moreover, if we do not regroup the classes of final grade, a prediction can still be valid if it is only off by 1 point, as there is some intrinsic fluctuation about anyone's performance in each exam, so another accuracy score for classifiers is calculated when we consider the minus 1 to plus 1 range of the score to be correct.

5. Experiments and Results

Based on our methods, we consider two parts in this section, include previous exam scores or not include them in the prediction. This is because the previous two exams, G1 and G2, showed a high positive correlation with G3 in the correlation analysis. We wanted to explore whether the machine could accurately predict students' final score in Portuguese based on other background factors, with or without G1 and G2. For the first part, the results of all four classifiers used with previous exam scores in the input included is shown in Table 3. The metrics used for regressor are MSE, MAE, and r2. For classifiers, we use accuracy. After considering the prediction to be accurate if it is of by one point, the modified accuracy is included as Acc(Modified) in the table. For the second part, the results of these models fit on data without previous exam scores are included in Table 4.

Table 3: Predict model result on data including previous exams.

Method	MSE	MAE	r2	Accuracy	Acc(Modified)
Ridge Regressor	0.87	0.67	0.88		
Decision Tree Regressor	1.05	0.73	0.85		
Decision Tree Classifier				0.43	0.89
KNN Classifier				0.28	0.75

Table 4: Predict model result on data without previous exams.

Method	MSE	MAE	r2	Accuracy	Acc(Modified)
Ridge Regressor	4.93	1.76	0.30		
Decision Tree Regressor	5.71	1.91	0.19		
Decision Tree				0.15	0.42

DOI: 10.54254/2753-7048/2/2022384

Classifier					
KNN Classifier				0.11	0.45

One can obviously find that the results in Table 3 are superior to those in Table 4. One can conclude that G1 and G2 are extremely important considerations for the training process of the model. Student performance on each test may be more volatile than one thinks. It can be concluded that G1 and G2 are very important factors to be considered during model training. In addition to the student's background information, the first two test scores are also very important indicators in the prediction process. If one doesn't include any previous score information, it's almost impossible to accurately predict a person's score. The forecast is not quite in line with expectations. Ideally, even without considering G1 and G2, one should be able to get better training results based on other considerations. And two indicators related to alcoholism should be strong. Although this analysis in the database found that binge drinking on weekdays and weekends did have a significant negative impact on student performance. The effects of G1 and G2 are still the most important. This is also reflected in the section of correlation analysis.

6. Conclusion and Future Work

In conclusion, the three directions proposed have been reasonably answered. The highest 5 factors that have a positive influence on our model are G2, G1, willingness of higher education, mother's education, and study time. The highest 5 factors that have a negative influence are number of past class failures, unwillingness of higher education, school in Mousinho da Silveira, number of school absences, workday alcohol consumption. Regarding alcohol consumption, the focus of this work, it is found that the level of alcohol consumption had a significant negative impact on students' academic performance. The students who drank more alcohol were more likely to have a poor final grade in Portuguese. This is consistent with previous analysis. Drinking on weekdays had a greater effect than drinking on weekends. This suggests that alcohol can interfere with students' ability to concentrate on homework and in class during the week, leading to poor grades in final exams. In order to achieve better final grades, students should reduce the frequency of alcohol abuse and avoid alcohol addiction. Also, it is easy to understand the high correlation between G1, G2 and G3. If students get good performance on G1 and G2, then the chance they get good performance on G3 is large.

The result G3 of training process including G1 and G2 is better than the one without G1 and G2. This result is possibly caused by the limit of the dataset. To improve the performance of exam score prediction, we propose that further study may consider collecting data about students' behavior only during the time between this and the next exams, and this time dependent data may be used to predict whether he will improve on his next exam.

References

[1] *Amy Keller. (2020). Drug Use in Middle School. https://www.drugrehab.com/teens/middle-school/*

[2] *Ezurike, C, Nweke, O, Ofuebe, I, Isiaku, B, Ncheke, C, Obetta, C, & Okeke, P. (2021). Effects of Alcohol Consumption on Undergraduate Students' Behaviour in Universities in South-East, Nigeria. Journal of Critical Reviews., 8(2): 206-218.*

[3] *Tembo, C., Burns, S., & Kalembo, F. (2017). The association between levels of alcohol consumption and mental health problems and academic performance among young university students. PLoS ONE, 12(6): 1–13.*

[4] *Gohari, M. R., Zuckermann, A. M. E., & Leatherdale, S. T. (2021). A longitudinal examination of alcohol cessation and academic outcomes among a sample of Canadian secondary school students. Addictive Behaviors, Volume 118.*

[5] *Balsa, A. I., Giuliano, L. M., & French, M. T. (2011). The effects of alcohol use on academic achievement in high school. Economics of education review, 30(1): 1–15.*

DOI: 10.54254/2753-7048/2/2022384

[6] *Lopez-Frias, M., Fernandez, M. D. L. F., Planells, E., Miranda, M. T., Mataix, J., & Llopis, J. (2001). Alcohol consumption and academic performance in a population of Spanish high school students. Journal of Studies on Alcohol, 62(6): 741+.*

[7] *Peter Grant. (2021). What Is Linear Regression?. https://builtin.com/data-science/linear-regression*

[8] *Saedsayad. (2021). Decision Tree – Regression. https://www.saedsayad.com/decision_tree_reg.htm*

[9] *Marina Chatterjee. (February 3, 2020). A Quick Introduction to KNN Algorithm. https://www.mygreatlearning.com/blog/knn-algorithm-introduction/*

[10] *Amitrajit Bose. (2019). Cross Validation — Why & How. https://towardsdatascience.com/cross-validation-430d9a5fee22*

[11] *Great Learning Team. (2020). What is Ridge Regression?. https://www.mygreatlearning.com/blog/what-is-ridge-regression/*