Predicting Academic Performance Based on Alcohol Consumption Patterns

By: Wina Aaron, Yannick Tanyi, Emanda Bisrat

ABSTRACT

It is widely known that alcohol consumption during an individual's adolescent years can negatively impact brain development and one's ability to learn. This report presents a predictive modeling approach to understand whether alcohol has a greater influence on the parts of the brain involved with processing language versus mathematics. Various machine learning algorithms, including KMeans clustering and Random Forest classification, are utilized to create visualizations in order to analyze the trends between the alcohol consumption levels and the final grade of the students in the dataset. With overall accuracy scores of 80% (math scores) and 90% (Portuguese scores), the results of this research demonstrate the effectiveness of the proposed approach in accurately classifying the final grades of students given their alcohol consumption and other predictors. Additionally, findings reveal drinking levels have a stronger correlation with students' math scores than with their language results. Insights derived from the analysis provide valuable understanding of how varying magnitudes of alcohol consumption affect the students' ability to learn..

Key Words: Alcohol consumption, Brain development, K Means clustering, Random Forest, Predictive modeling, visualizations

EXPLORATION

The datasets utilized in this analysis are derived from Kaggle. The data for one of the datasets was gathered from a survey taken by students enrolled in a math class while the other dataset was taken by students enrolled in a Portuguese course. The data was collected in April of 2008 from secondary schools in Porto, Portugal, where the legal drinking age is 18, and has allowed for the exploration of interesting social, gender, and study information. It comprises anonymized student details, including demographics, levels of alcohol consumption, and performance in the classes.

There have been numerous projects done using the Kaggle: Student Alcohol Consumption dataset, and based on our research, individuals are utilizing varying variables as predictors for differing outcome variables. One particular report we found intriguing was completed by Andre Ye, a student attending the University of Washington. Mr. Ye had the objective of creating a model that accurately predicts the final score of students based on predictors he chose, which, as aforementioned, was a common theme in the projects created using this dataset. However, Mr. Ye tested a number of sophisticated models to find the model that provides the lowest mean absolute error. He then used the PermutationImportance function to find his most influential predictors, and successfully created visualizations that illustrate how dependent final grades are on each predictor.

Our initial exploratory data analysis included merging the two datasets, the first one including the final scores in Math and the second one including the final scores in Portuguese. We also dropped columns that were not needed for the context of our project. We initially compared weekday and weekend alcohol consumption, and found that weekday alcohol consumption affected final scores more. We then combined the weekday and weekend alcohol consumption predictors in a combined column in order to more closely analyze the Math and Portuguese Scores against weekly alcohol consumption. To address multicollinearity, we created a correlation matrix (Figure 1) to show if any of the predictors are highly correlated, and dropped the ones that had a correlation greater than 0.7. Finally, we wanted to see the distribution of the predictors not only to help understand the scale and magnitude of each variable, but to see the relationships between them which is shown in

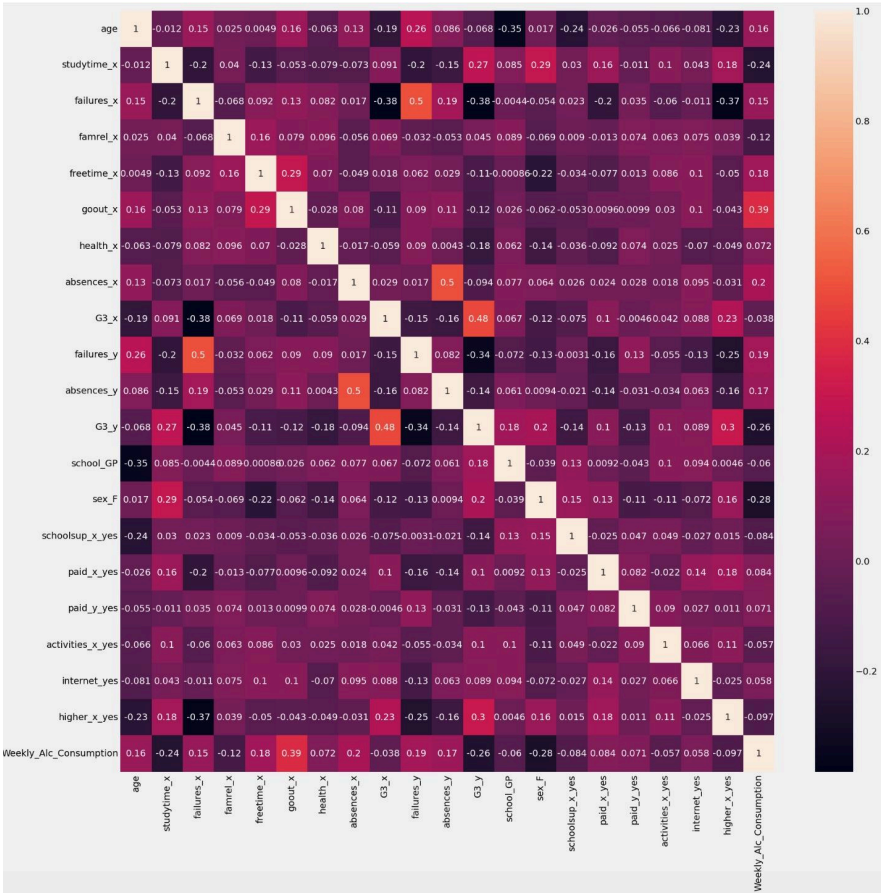

Figure 1

Figure 2 below. We decided to include the distributions of the final math grade and final Portuguese grade, as these will be our focus for the report.

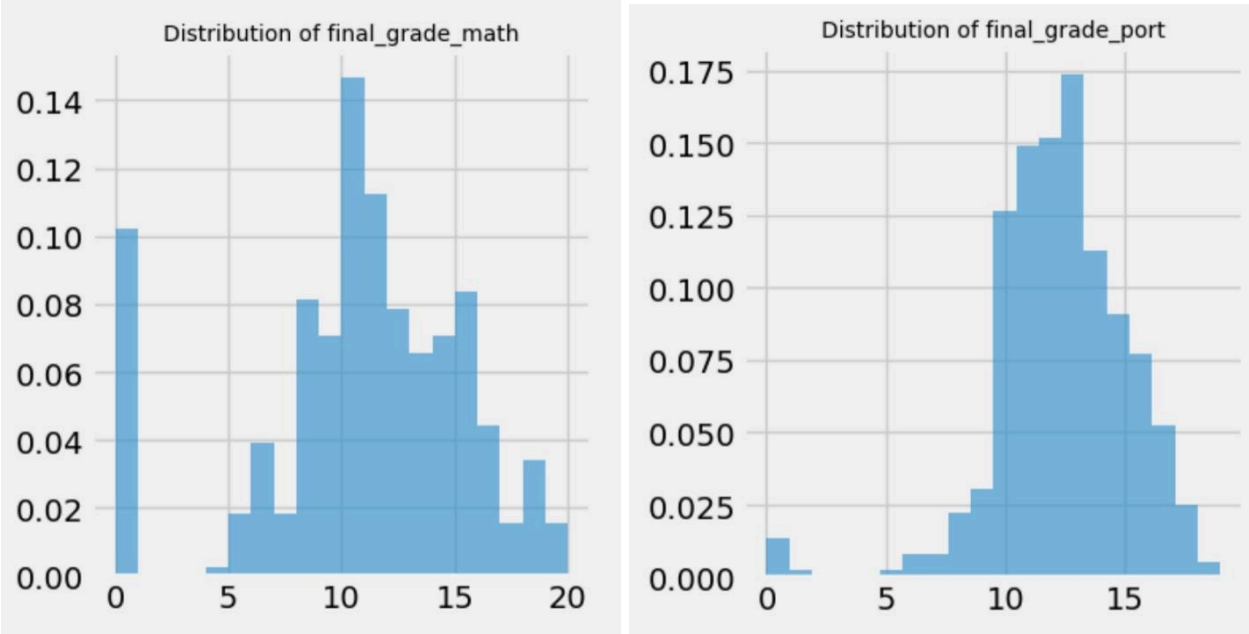

Figure 2

METHOD 1 - K-MEANS CLUSTERING:

For our first method, we used K-means clustering in order to compare Math and Portuguese Final Grades with weekly student alcohol consumption, and how they differ for each subject. For math scores, we first used the elbow method and determined that 3 clusters would be the optimal k-means value. The x-axis measures the weekly alcohol consumption from a range of 2 to 10, 2 being very low consumption and 10 being very high consumption. The y-axis measures the Final Math Grade, ranging from 0 to 20. In the visualization, the 3 centroids lie from 3-5 on the x-axis, and along with the increase in variance as the consumption increases, this indicates that most students lie in the low-medium range of consumption. In addition, this shows that the lower the weekly consumption, the better the grade.

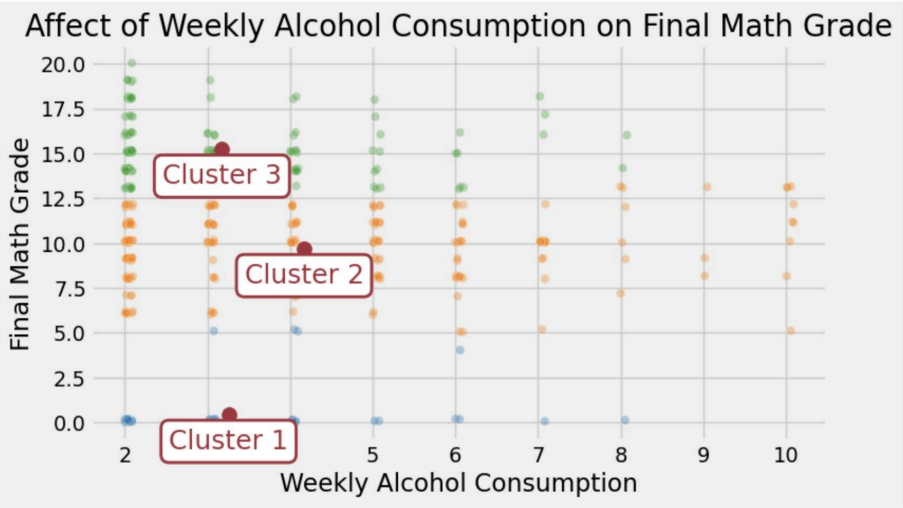

Figure 3

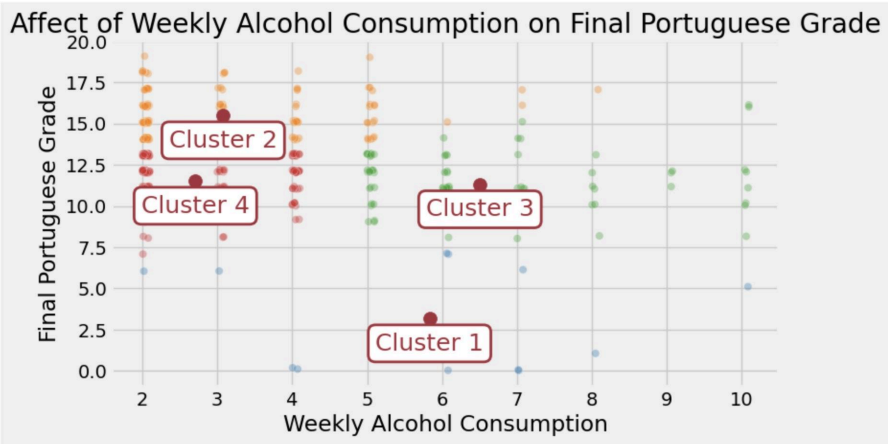

Figure 4

For final Portuguese scores, we determined that the optimal k-means value is 4 by using the elbow method. Using the same axes, this visualization shows that as the weekly alcohol consumption increases, the final grade gets worse. It is also apparent that there is less variability in the final scores because they tend to be higher.

When comparing both visualizations, it is clear that the final scores for math are much more varied than portuguese. In math, the data points on the y axis generally ranged from 5.0 to 17.5 score while for Portuguese, the final scores generally ranged from 10.0 to 18.0. In the context of our report, this indicates that weekly alcohol consumption affects final math scores much more in comparison to final Portuguese scores. This shows that increased alcohol consumption affects the parietal/occipital/frontal parts of the brain more than the temporal lobe, which is involved in processing language.

In addition, we then calculated the lowest sum of squared errors, which measures the distance between each data point and its assigned centroid. For final math scores, the lowest sum of squared errors found was 2589.591 and for final Portuguese scores, the lowest sum of squared errors was 1435.715. A lower SSE value indicates that the data points within each cluster are closer to their respective centroids, suggesting a better clustering solution. As described above, the Portuguese clustering has a significantly lower SSE, indicating its a better solution.

METHOD 2 - Random Forest Classification:

For our second method, we used Random Forest Classification to predict final Math and Portuguese scores based on student alcohol consumption. The Random Forest classifier was utilized due to its ability to handle non-linear relationships and complex interactions among predictors. By increasing the number of estimators to 100 and adjusting parameters such as maximum depth and cost-complexity pruning alpha, we aimed to optimize the model's performance.

After training the Random Forest classifiers for both Math and Portuguese scores, we evaluated their accuracy using test data. We changed our final score variables to binary by making it Pass/Fail, pass being a score greater than or equal to 10 and fail being a score less than 10. For predicting Math scores, the accuracy achieved by the model with increased estimators was approximately 80.52%, while for predicting Portuguese scores, the accuracy reached approximately 90.91%. These accuracy scores provide insight into the model's ability to correctly classify students' final grades based on their alcohol consumption levels and other predictors.

The confusion matrices below further illuminate the performance of the models. In the case of Math scores, the confusion matrix reveals that out of the total predictions made, 11 students were correctly classified as failing and were indeed failing, 3 students were incorrectly classified as failing when they actually passed, 51 students were correctly classified as passing when they were indeed passing, and 12 students were incorrectly classified as passing when they actually failed. Similarly, for Portuguese scores, the confusion matrix indicated that out of the total predictions made, 3 students were correctly classified as failing and were indeed failing, 1 student was incorrectly classified as failing when they actually passed, 67 students were correctly classified as passing and were indeed passing, and 6 students were incorrectly classified as passing when they actually failed. These results highlight the model's effectiveness in predicting student outcomes based on alcohol consumption, with a notable emphasis on correctly identifying students who may still pass even with a higher alcohol consumption.

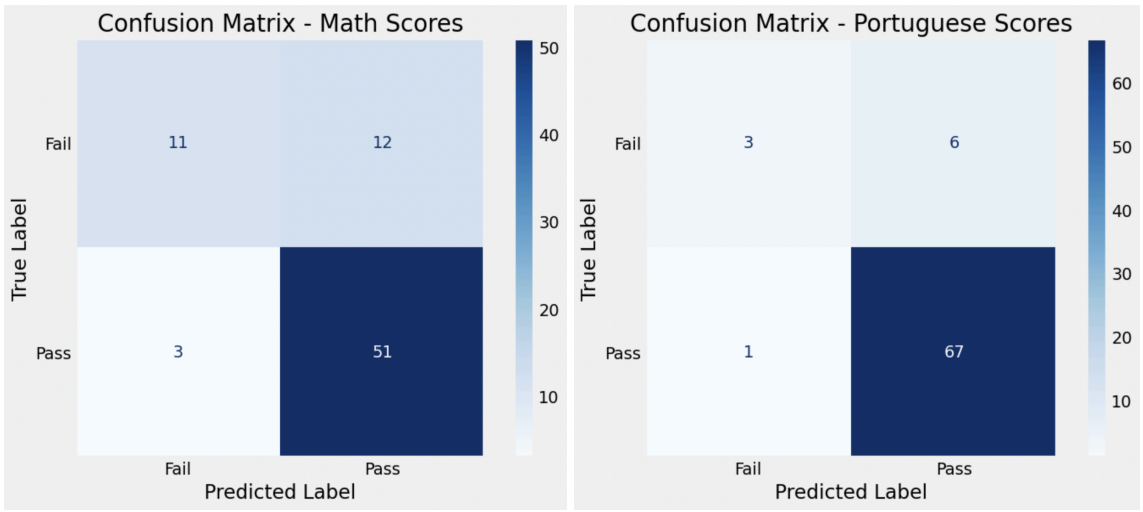

Figure 5

We can also see several trends that emerge from looking at both figures regarding the predictive modeling of student outcomes based on alcohol consumption– in the case of Math scores, the model tends to underestimate the negative impact of alcohol on the parietal and frontal lobes, as evidenced by a higher rate of false positives. This suggests that students who may be at risk of failing due to alcohol consumption are not being accurately identified

by the model. Similarly, for Portuguese scores, while the model shows strong specificity in predicting passing outcomes, it still overlooks the detrimental effects of alcohol on the temporal lobe, as indicated by a higher rate of false negatives. This indicates that students who may struggle academically due to alcohol consumption are not being effectively identified. In short, the confusion matrices portray the models ability to correctly classify the performance of students, and as shown in figure 5, the model is better at accurately identifying students who are passing or failing. We can contribute this to the fact that the variance in distributions for final grades in math is greater as the histogram is more normally distributed while the distribution for Portuguese is left skewed.

CONCLUSION:

From our two methods, we concluded that student alcohol consumption affects math and Portuguese scores affect final math grades in comparison to final Portuguese grades. K-means clustering was used to group students based on their alcohol consumption patterns while Random Forest Classifier was used to assess the effect of features on academic performance and evaluate the accuracy of the model. Both methods have limitations as well. For instance, Random Forest Classifier relies heavily on tuning the parameters for model performance. For K-means clustering, some limitations include sensitivity to outliers and needing to specify the number of clusters in advance. Because random forest is a predictive model, it fits the context of our project better and is more appropriate with respect to the academic outcomes. Nevertheless, both methods were valuable in the context of our report as K-means clustering was helpful in identifying patterns and Random Forest was helpful in producing the highest accuracy.

REFERENCES

P. Cortez and A. Silva. *Using Data Mining to Predict Secondary School Student Performance* In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBU TEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.**https://www.kaggle.com/datasets/uciml/student-alcohol-consumption/data**

Kulasingham, Joshua P., et al. "Cortical Processing of Arithmetic and Simple Sentences in an Auditory Attention Task." *Journal of Neuroscience*, Society for Neuroscience, 22 Sept. 2021, www.jneurosci.org/content/41/38/8023.

This research paper was used in an article we read in order to fully understand the cortical difference in processing language and mathematics.