**MAY 19TH, 2021**

**THE EFFECTS**

**OF STUDENT ALCOHOL CONSUMPTION**

Rayanna Harduarsingh | Sandro O’Neill

IST 707 Final Project

Syracuse University

Rayanna Harduarsingh | Sandro O’Neill

IST 707 Final Project

Syracuse University

May 19th, 2021

**Shape

Description automatically generated with low confidence**

**ABSTRACT**

Student alcohol consumption is a prevalent public health issue, but little is known of its effects on academic achievement. Using real-world data obtained from a secondary school in Portugal, our objective was to predict how underage alcohol consumption affects a student’s performance. To model our predictions, we used four different analysis methods: Logistic Regression to discover correlations, Clustering to segment students based on several metrics, Classification (Naive Bayes, Support Vector Machine, Decision Tree, & Random Tree) to predict accurate relationships, and Association Rule Mining to find common variables with students who had high or low alcohol consumption. Although the classification models did not show a good predictive accuracy, the regression and association rule mining yield the most insights. Alcohol consumption did not heavily affect the grades of students, but it was more likely to be associated with higher absences, low study time, and failures. The results of this analysis could potentially identify what resources students need through student outreach and create policy to put an end to underage drinking.

Shape

Description automatically generated with medium confidence

**INTRODUCTION**

Underage drinking has been a serious long-term issue in the United States and all over the world. According to the 2019 National Survey on Drug Use and Health, about 7.05 million Americans between the ages of 12 and 20 reported current alcohol consumption. The influence of friends, siblings, media, and especially parents all factor into a child’s drinking habits. Some report having a range of 1-2 drinks to having more than 10 drinks in one night. Alcohol is a factor related to approximately 4,300 deaths among underage youths in the U.S. every year (Harding et al., 2016). This remains a huge concern for underage drinkers as underage drinking and its accompanying difficulties can have substantial negative implications. One of these consequences could be the decline in a student’s academic progress. One in two underage students reported that alcohol was “very easy” to obtain (Wechsler et al., 2002). However, little is known regarding the impact of underage drinking on academic achievement. Deficits on a variety of language and attentional activities, verbal and nonverbal memory tests, and specific working memory impairments have been discovered in youths aged 15–19 years with a history of alcohol misuse ([Schweinsburg et al., 2005](https://www-sciencedirect-com.libezproxy2.syr.edu/science/article/pii/S0376871609004554?via%3Dihub" \l "bib30)). An adolescent must acquire these cognitive skills as they are fundamental to succeed in school.

**LITERATURE REVIEW**

A data set from 2008, provided by the University of California Irvine (UCI), obtained student data from a survey of a Math and Portuguese language course taught in a secondary school located in Portugal. It includes various social, gender, and study data from mark reports and questionnaires such as the student’s age, alcohol consumption, mother’s education, and more. UCI analyzed this real-world data to predict student progress and, if possible, to pinpoint the fundamental factors that influence educational success or failure. Using multiple prediction methods such as Support Vector Machines, Decision Trees, and Random Forest, the analysts at UCI found that student achievement is highly affected by a number of school, demographic, and social variables such as number of absences, extra educational support, parent’s job and education, and as well as alcohol consumption (Cortez & Silva, 2008). In another study conducted by Mallie J. Paschall and Bridget Freisthler, they also examined the effects of heavy drinking on academic performance where they found alcohol-related academic problems. To continue this analysis on this focus area, our goal was to explore further into the factors that affect student performance and find more insights by focusing on one variable we deemed important, alcohol consumption, as it is a significant public health issue. By modeling significant variables based on alcohol consumption, we can acquire a deeper understanding on how much it influences educational success. The purpose of this study is to explore the relationship between a student’s alcohol consumption and their current grades to determine if alcohol has a negative effect on their academic performance or no effect at all.

**RESEARCH QUESTION**

The student data set is composed of answers from a survey as well as school reports to generate social, demographic, and school related variables. Key variables for our analysis include but are not limited to the student’s sex, age, study time, failures, extra-paid classes, wanting to take higher education, free time, weekday and weekend alcohol consumption, health, absences, and grades. The students are graded on a 19-point scale, with 0 being the lowest grade and 19 being the highest. Students are evaluated three times during the school year, with the last evaluation, G3, corresponding to the final grade. The social data collected such as the amount of alcohol consumption and going out helped us generate two research questions that will test our hypothesis:

**Hypothesis:** Higher drinking levels will significantly correlate with low final grades, study-time, and absences.

1. What variables associate with high and low alcohol consumption metrics?
2. Can other variables besides alcohol consumption affect student performance?

**DATA DESCRIPTION & PRE-PROCESSING**

The .csv data file we obtained from Kaggle contained 649 observations across 32 variables. For the purpose of this analysis, we combined two variables: Dalc (weekday alcohol consumption) and Walc (weekend alcohol consumption) into a single variable “alcohol”. Dalc and Walc ranged 1-5, so the combined “alcohol” variable had values ranging from 2-10. After combining them, the Dalc and Walc variables were removed as they would interfere with our classification algorithms. The right-skewed histogram below gives the number of students by their self-reported alcohol weekly alcohol consumption.

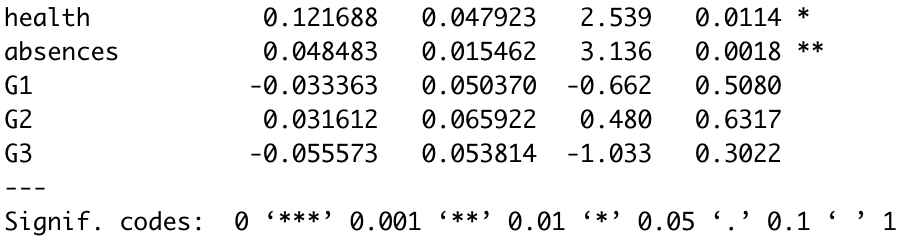
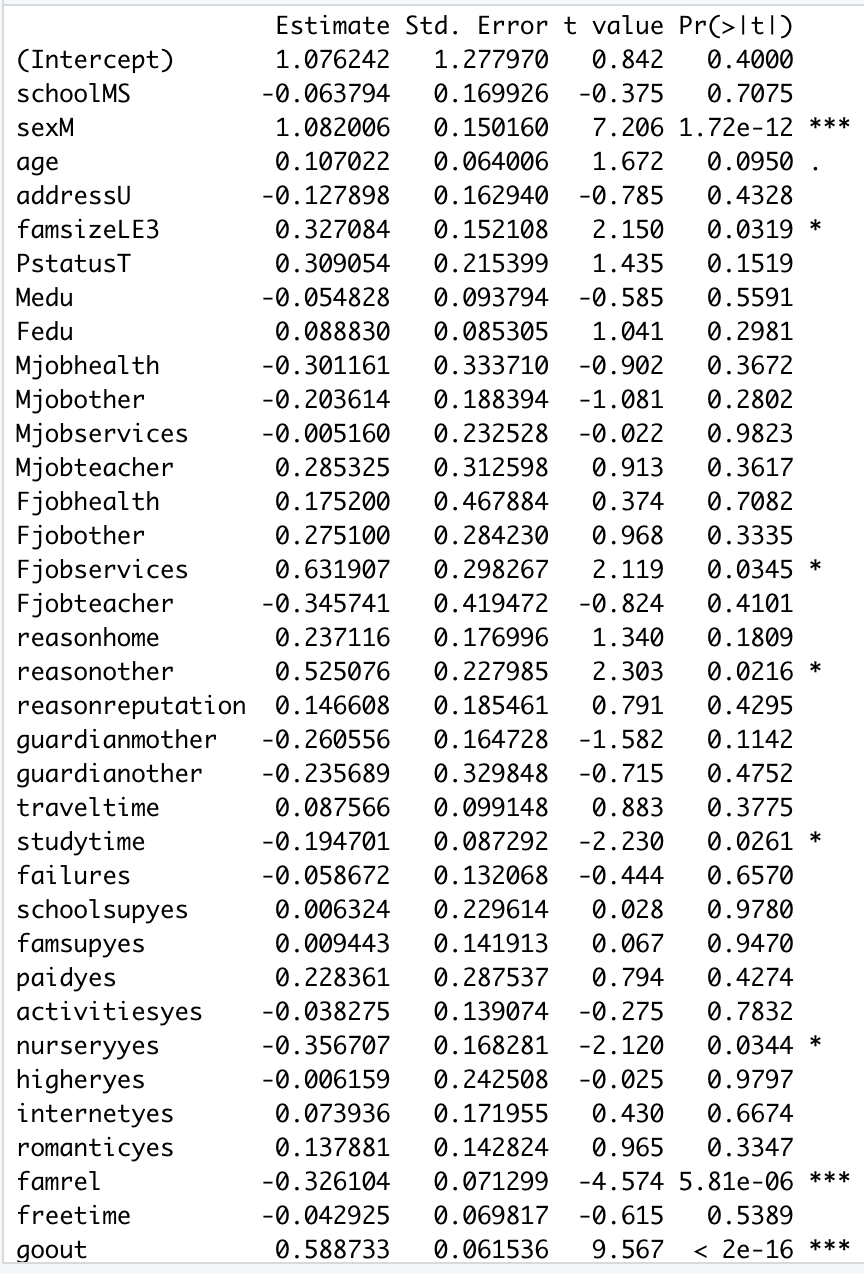
Chart, histogram

Description automatically generated

**METHODS & RESULTS**

**LOGISTIC REGRESSION**

We converted all of our character variables into factors and ran a logistic regression, with alcohol as our response variable and all other variables as the explanatory. We found that 11 of the 31 variables were significant at alpha = 0.05. The results of the regression can be seen below, surprisingly grades were not a significant predictor of alcohol consumption which was one of our initial predictions. The most notable predictors with the lowest p-values were: goout (1-5 range of how often a student went out), sexM (male student) and famrel (1-5 range of family relationship quality). Some results were expected, such as increased study time decreasing alcohol consumption, but there were some surprises such as health (1-5 range of health quality) increasing alcohol consumption which seems contradictory.



Chart, line chart

Description automatically generated**CLUSTERING**

Chart, scatter chart

Description automatically generatedWe subsetted our initial dataframe into a new object “numdata” which only included numerical variables in order to be compatible with k-means clustering. After scaling the data, we determined our optimal number of clusters to be 2, as can be seen from the “elbow” in the wss plot below. After generating our 2 clusters, we can see from the cluster plot that there is some overlap between the two. When the number of clusters were increased to 3 and 4 we saw significantly more overlap and decided to continue with 2 clusters. Using the colMeans() command on our “numdata” object with our clusters nested in brackets, we were able to view the average values of the numeric variables in the two clusters. In the first cluster (size 277), the average age, alcohol consumption, absences and failures were above the data set average, while the second (size 372) had lower than average values for the same variables. The grades for the first three quarters were lower on average for the first cluster, and higher for the second.

A picture containing text, newspaper, screenshot, receipt

Description automatically generated

**CLASSSIFICATION**

To test the predictive performance of alcohol consumption, we used the algorithms Support Vector Machine, Naïve Bayes, Decision Tree, and Random Forest to build our models. We split 70% of the data to use to train our prediction models, and the remaining 30% was left for us to test our models on. We also ran a 3-fold cross-validation procedure across each prediction model. We experimented with multiple different tuning parameters as well as tried the advanced 10-fold cross validation procedure, however, both procedures either yielded the same results or did not make much of a difference. Initially, we trained our models with alcohol as our factor variable in response to all the other explanatory variables. We were getting pretty low accuracies, even after predicting on our test set. In addition, we paired alcohol with other explanatory variables one at a time such as absences, final grades, failures, and more to predict a relationship, but we were still deriving low accuracies in the 9%-20% range. However, in order to improve the accuracy, we tried a different approach which would ultimately create our final prediction models. Based on our regression model from earlier, we identified quite a few statistically significant predictors that had a low p-value. We chose to use the top 4 predictors as our explanatory variables in our training models which were free time, going out, absences, and study-time in response to alcohol. For our SVM model, we tuned the parameter for our kernel to use the “polynomial” method instead of the classic linear as it produced a higher accuracy. For naïve bayes, we trained it using the default naïve bayes method. For decision tree, we used the “rpart” method and for random forest we used the “rf” method. Each model was tuned with a search grid as well multiple different metrics to try and derive the highest accuracy as possible. It was then used to predict the testing set in the output of a confusion matrix. After using the statistically significant predictors from our regression, our accuracies definitely increased, however, it was still not a result we were looking for, hence trying other methods.

|  |  |
| --- | --- |
| MODEL | PREDICTION ACCURACY |
| SVM | **39.5%** |
| NAÏVE BAYES | **37.9%** |
| DECISION TREE | **34%** |
| RANDOM FOREST | **32%** |

**ASSOCIATION RULE MINING**

A picture containing table

Description automatically generatedWe added an identity variable “ID” with the respective observation numbers to our data-frame, then used the mutate\_if() function to convert all character variables into factors and discretize all numeric variables. By changing the data types, we were able to use the apriori() function on the data-frame in order to generate association rules. With roughly 3.78 being the average alcohol consumption across the entire dataset, we set the left-hand side apriori parameter to a combined list of alcohol values 5-10 for above average alcohol consumption. With our initial parameters set to 90% confidence, support equal to 0.001 and maximum length of 5, close to 3 thousand rules were generated. After inspecting the counts, we noticed that a lot of the rules had counts equal to 1. By filtering the rules by count numbers more than 3, we were left with 13 good rules which can be seen in the screenshot below. This process was repeated for the left hand side set to alcohol = 2 for below average/no alcohol consumption (26 good rules), along with higher (23 rules) and lower absence (18 rules) numbers. For absences, higher count and support filters had to be used in order to obtain the final results.

Table

Description automatically generated with medium confidence

Table

Description automatically generated with medium confidence

A picture containing table

Description automatically generated

**DISCUSSION**

Out of all the methods used for our analysis, the regression model and association rules yielded the most insightful analysis. We were able to draw appropriate conclusions about the effects of student alcohol consumption. Underage drinking does affects students, but not dramatically based off of this data set. Our hypothesis could not be proven as there was not much of a correlation or relationship between alcohol and grades. However, we were able to answer our second research question, “Can other variables besides alcohol consumption affect student performance?”. Alcohol consumption was found to be more associated with higher absences, low study time, and failures, which are all key components to have a good academic standing.

In the study conducted by Mallie J. Paschall and Bridget Freisthler, they too examined the effects of heavy drinking on academic performance (focused on Grade Point Average) from students based on Freshmen attending the University of California at Berkeley. They concluded that heavy alcohol use did not have a significant effect on students’ academic performance as well. In the original analysis conducted by the University of California Irvine (UCI) using the Portuguese class data set, although they did not focus particularly on alcohol consumption as the response variable, they focused on all the variables to determine what factors affect student performance. They also found that student achievement was affected by number of absences, extra educational support, parent’s job and education, as well as going out, which were variables we also found in our association rules. Although some of our models did not achieve high accuracies on testing sets, we were still able to identify the significant variables that contributed to higher rates of alcohol consumption.

After performing our analysis and coming to our conclusions, we reflected on how our analysis could have been more promising. We noticed some data observations were repetitive having several of the same values which is why we believe our accuracies were low in our prediction models from our classification method. However, being that it derived low accuracies, it could have that alcohol consumption was simply not a strong predictor for student performance. Having more observations in the dataset could potentially increase the accuracies of our models while also yielding more interesting association rules.

**REFERENCES**

Frances M. Harding, Ralph W. Hingson, Michael Klitzner, James F. Mosher, Jorielle Brown,

Robert M. Vincent, Elizabeth Dahl, Carol L. Cannon, Underage Drinking: A Review of Trends and Prevention Strategies, American Journal of Preventive Medicine, Volume 51, Issue 4, Supplement 2, 2016, Pages S148-S157, ISSN 0749-3797,

https://doi.org/10.1016/j.amepre.2016.05.020.

Paschall, M. J., & Freisthler, B. (2003). Does heavy drinking affect academic performance in college? Findings from a prospective study of high achievers. *Journal of Studies on Alcohol,* *64*(4), 515-519. doi:10.15288/jsa.2003.64.515

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 (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.

Schweinsburg, A. D., Schweinsburg, B. C., Cheung, E. H., Brown, G. G., Brown, S. A., & Tapert, S. F. (2005). Fmri response to spatial working memory in adolescents with comorbid marijuana and alcohol use disorders. *Drug and Alcohol Dependence,* *79*(2), 201-210. doi:10.1016/j.drugalcdep.2005.01.009

University of California Irvine. (2016). Student Alcohol Consumption [Data file]. Retrieved

from https://www.kaggle.com/uciml/student-alcohol-consumption

Wechsler, H., Lee, J. E., Nelson, T. F., & Kuo, M. (2002). Underage College Students' Drinking Behavior, Access to Alcohol, and the Influence of Deterrence Policies: Findings from the Harvard School of Public Health College Alcohol Study. *Journal of American College Health,* *50*(5), 223-236. doi:10.1080/07448480209595714