Correlation between School/Personal life and Alcohol Consumption

Fengsheng Zhou(fengshen@usc.edu); Jinhao Huang(jinhaohu@usc.edu);
Lingyi Chen(lchen706@usc.edu); Tianjiao Hu(hutianji@usc.edu).

Background

As a legal, addictive substance, alcohol plays an important role in people's daily life and social interaction. Moderate drinking can improve blood circulation and promote the development of interpersonal relationships. However, over consumption of alcohol will have an important impact on individual physical and mental health. Nowadays, over consumption of alcohol among juveniles has become a significant issue. More and more young people are obsessed with alcohol. Alcohol is considered as an anesthetic to obtain short-term pleasure from their busy academic life. However, according to a new study published in *American Journal of Psychiatry*, over consumption of alcohol during adolescence would have a permanent, bad impact on people's brain development. The adolescents who drink more, less brain white matter would be produced and more gray matter would be reduced. This kind of decrease occurs in the lateral and anterior lobes of the brain, which are responsible for learning, emotional development, self-control and other abilities.[2]

Therefore, knowing what factors can affect alcohol over-consumption among juveniles would help us to solve the problem. From related works, we find that multiple factors influence college drinking, from an individual's genetic susceptibility to the positive and negative effects of alcohol, expectations regarding the benefits and detrimental effects of drinking, penalties for underage drinking, parental attitudes about drinking while at college, whether one is member of a Greek organization or involved in athletics, and conditions within the larger community that determine how accessible and affordable alcohol is.[3] [4] These factors can be generally categorized into personal factors and social factors. While, our research would like to discover and analyze some special factors of the consumption level. In our project, variables can be roughly categorized into school life and personal life. We plan to mainly focus on these two categories that can influence the level of alcohol consumption. And also, we want to find out whether family issues have something to do with alcohol over-consumption.

Problem Definition

Seeking factors of alcohol over-consumption is the best way to prevent the progression of addiction. Therefore, our group decides to do some statistical analysis to detect what factors might lead to students' alcohol over-consumption. Specifically, we want to analyze whether alcohol consumption will be impacted by different features we selected, like school, sex, age, locations, academic behaviors, health situations, and family relationships. Additionally, we would like to make a fit model to predict what kind of students are more likely to consume more alcohol.

Data Description

- **Data Source**

  https://www.kaggle.com/datasets/whenamancodes/alcohol-effects-on-study/versions/1?resource=download

  The dataset is downloaded from Kaggle, which is used for measuring students' achievements in secondary education of two Portuguese schools. After data cleaning, the total number of observations is 395.

- **Attribute Description**

The dataset uses both categorical variables and quantitative variables as factors. Since the dataset includes a total of 31 variables, we decide to pick the most relevant variables from the dataset. There are 12 variables we select including student's school, student's sex,student's age, student's home address type, weekly study time, number of past class failures, with a romantic relationship, quality of family relationships, workday alcohol consumption, weekend alcohol consumption, current health status, number of school absences, final grade. Among the above variables, age, grades and absences are quantitative variables. Most of the variables are numeric.

Categorical Variable	**Description**
student's school (school)	binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira
student's sex (sex)	binary: 'F' - female or 'M' - male

student's home address type (address)	binary: 'U' - urban or 'R' - rural
weekly study time (studytime)	numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours
number of past class failures (failures)	numeric: n if 1<=n<3, else 4
with a romantic relationship (romantic)	binary: yes or no
quality of family relationships (famerel)	numeric: from 1 - very bad to 5 - excellent
workday alcohol consumption (Dalc)	numeric: from 1 - very low to 5 - very high
weekend alcohol consumption (Walc)	numeric: from 1 - very low to 5 - very high
current health status (health)	numeric: from 1 - very bad to 5 - very good
Quantitative Variable	**Description**
student's age (age)	numeric: from 15 to 22
number of school absences (absences)	numeric: from 0 to 93
final grades (G3)	numeric: from 0 to 20, output target

Method Description

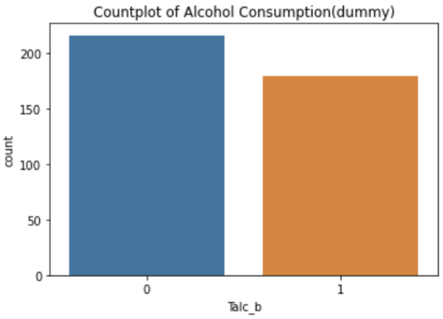
We used different kinds of methods to achieve our goal. The first one we used is the decision tree method. **Decision tree** is one of the methods in supervised learning classification. Leaves of the tree represent the class labels, while branches of the tree represent the features that lead to the classification of labels. And we want to find out the accuracy of our decision tree model. So we import metrics from the sklearn package and input the predicted y value and the test set in the metrics.accuracy_score function to get the accuracy. We can also get the recall and the precision of the method by importing their function similarly. We also import the **confusion**

matrix function to create a confusion matrix of this model to find out the performance of our model.

Besides the decision tree, we also used **logistic regression** to predict our model. Logistic regression is the model that can predict the probability of an event by having the log-odds of one or more variables. To check the performance of our regression model, we did the same thing as before. We get the accuracy, recall, precision and confusion matrix of our regression model.

In order to get a better result, we then use the third method — **Random Forest**. Random Forest is like the decision tree. It will construct multiple decision trees. We also perform the same thing like before to get the performance of our model.

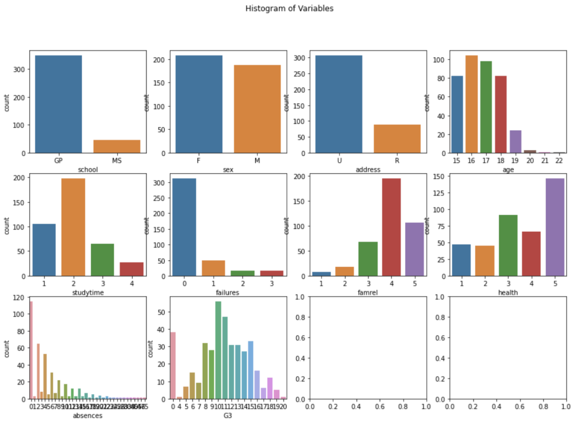
Experiment: experiment setup and analysis results



(figure 1)

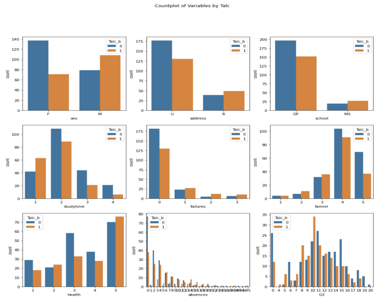
As our goal is to analyze how different factors influence the level of alcohol consumption, we first combined the variables Walc (weekend alcohol consumption) and Dalc (weekday alcohol consumption) into Talc (total alcohol consumption. Then, in order to differentiate the relationship between different factors and alcohol consumption levels, we created a variable called Talc_b (figure 1). We categorized rows on a range of alcohol consumption values. As shown in the figure, there are more than 200 students who have an alcohol consumption level between 0 to 4.

As shown in Figure 2, we did an overall analysis of all variables by plotting histograms. Most of the students are from Gabriel Pereira (GP) high school, under the age of 18 and living in urban areas. Almost 300 students have a good family relationship with the value more than 4.



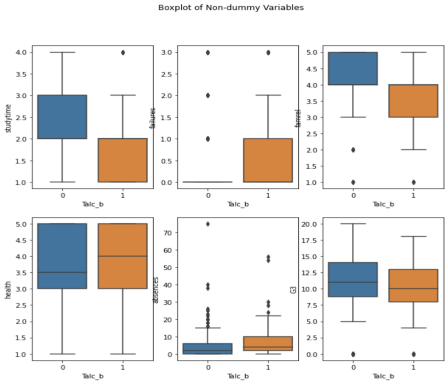
(figure 2)

From the count plot of all variables by Talc shown in figure 3, we can find that female students tend to drink less, while male students have a higher consumption value. Students with a longer study time, bad health conditions, higher grades, less times of failure and absences have a moderate alcohol consumption between 0 to 4.



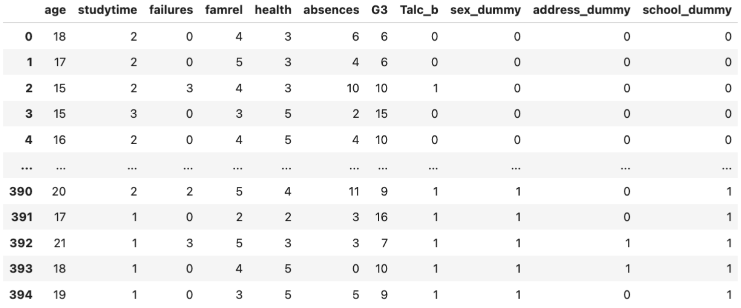
(figure 3)

We used a box plot shown in figure 4 to analyze the consumption level with numerical factors. We can see that students with longer study time, less failure, good family relationships, fewer school absences and better grades have a low alcohol consumption value. By analyzing Talc, we can conclude that school/personal life does have an effect on alcohol consumption.



(figure 4)

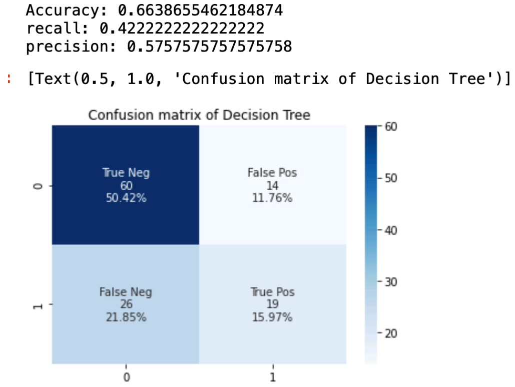
We also used dummy variable to represent several categorical variables including sex, address and school. And we completed our final dataset shown as figure 5, which will be used to fit models.



	age	studytime	failures	famrel	health	absences	G3	Talc_b	sex_dummy	address_dummy	school_dummy
0	18	2	0	4	3	6	6	0	0	0	0
1	17	2	0	5	3	4	6	0	0	0	0
2	15	2	3	4	3	10	10	1	0	0	0
3	15	3	0	3	5	2	15	0	0	0	0
4	16	2	0	4	5	4	10	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...
390	20	2	2	5	4	11	9	1	1	0	1
391	17	1	0	2	2	3	16	1	1	0	1
392	21	1	3	5	3	3	7	1	1	1	1
393	18	1	0	4	5	0	10	1	1	1	1
394	19	1	0	3	5	5	9	1	1	0	1

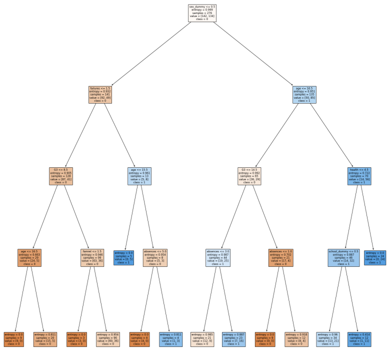
(figure 5. Dataset processing)

In order to find the best model fitting our data, we firstly tried the decision tree method and got an accuracy of 66% (figure 6). Below is one of the branches for our decision tree for the classification process (figure 7). To import this method, we first split the entire dataframe into the train set and the test set, using the train-test-split model from sklearn.model_selection package. We use test size = 0.3 to get 70% of the train set and 30% of the test set. Then we imported the DecisionTreeClassifier from the sklearn.tree package to fit the train and test set to get the result. We also get the accuracy, recall, precision of this method with the import of confusion matrix from sklearn.metrics.



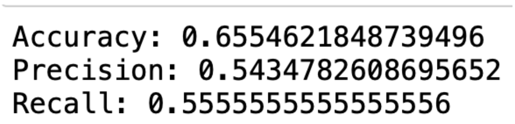
```
Accuracy: 0.6638655462184874
recall: 0.4222222222222222
precision: 0.5757575757575758

: [Text(0.5, 1.0, 'Confusion matrix of Decision Tree')]
```

(figure 6)



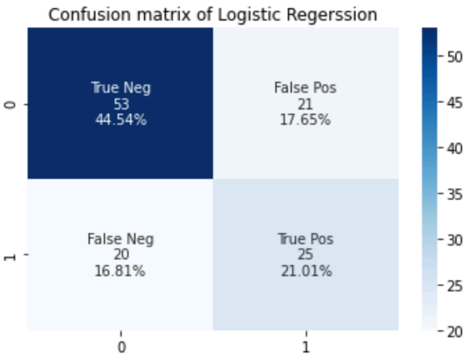
(figure 7)

After getting the 66% accuracy of the decision tree, we want to try other methods to see if the accuracy can be increased. Then, we tried the logistic regression method. For the Logistic Regression method, we import LogisticRegression from sklearn.linear_model. We use the LogisticRegression function to fit our already split train and test data and use the logreg.predict function to predict the y values based on the test value of x. Then we also create the confusion matrix and accuracy, recall, precision like we did for the decision tree method. We tuned the depth of the tree and got a max_depth of 2. The accuracy will be 65% (figure 8 & 9).



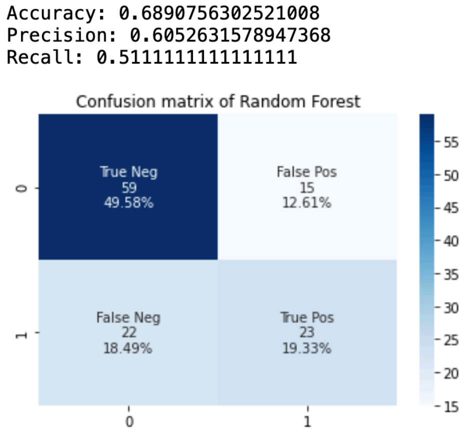
```
Accuracy: 0.6554621848739496
Precision: 0.5434782608695652
Recall: 0.5555555555555556
```

(figure 8. Logistic regression measurements)



(figure 9)

Lastly, since the decision tree and logistic regression resulted in poor accuracy, we tried random forest. To use this method, we import RandomForestClassifier from sklearn.ensemble. We then used the RandomForestClassifier to fit our training data and get the result. The next step is basically the same as the two methods before. We create the confusion matrix and get accuracy, recall, precision of our model. The best max_depth for the random forest after tuning is 4. Luckily, for the random forest method, we got an accuracy that is slightly higher than the other two methods with a value of 69% (figure 10).



```
Accuracy: 0.6890756302521008
Precision: 0.6052631578947368
Recall: 0.5111111111111111
```

(figure 10)

Observation and Conclusion

By analyzing the relation between alcohol consumption and various chosen variables, our study established that variables including study time, health condition, failure, absences and family relationship does indeed have a significant effect on student's alcohol consumption. Some variables show a positive relationship: An increasing number of failures and school absences increase the level of alcohol consumption. An increasing level of study time and family relationships decrease the level of alcohol consumption.

In the model section, we used different methods like Decision Tree, Logistic Regression, and Random Forest to find our goal which is whether school/personal life will have an effect on alcohol consumption. Eventually, we got an accuracy of 66% for Decision Tree, 65% for Logistic Regression, and 69% for Random Forest. We used three methods because we wanted to find out whether using different methods can improve our model performance. Unfortunately, Logistic Regression didn't show much of the improvement. But Luckily, the Random Forest did improve the accuracy to almost 70%. Although the performance is not very perfect, it still shows that all three methods favor the positive result. Therefore, our final result of this project is that school/personal life does have an effect on alcohol consumption.

References

1. Camilla Molin.(June,2020). *A statistical analysis of the performance in mathematics of secondary students in Portugal.* https://uu.diva-portal.org/smash/get/diva2:1464106/FULLTEXT01.pdf

2. Lindsay M. Squeglia., Susan F. Tapert. (2015). "Brain Development in Heavy-Drinking Adolescents". *American Journal of Psychiatry.172*(6),531-542 https://ajp.psychiatryonline.org/doi/full/10.1176/appi.ajp.2015.14101249

3. El Ansari, W., Stock, C., &amp; Mills, C. (2013). "Is alcohol consumption associated with poor academic achievement in university students?" *International journal of preventive medicine*. *113*(4),1175-1188 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3843305/

4. Aaron White, Ralph Hingson. (2013). The burden of alcohol use: excessive alcohol consumption and related consequences among college students. "*Alcohol Research: Current Reviews.*" *35*(2),201-218 https://link.gale.com/apps/doc/A382084212/GPS?u=usocal_main&sid=bookmark-GPS&xid=cdd17037