Introduction and Background

Despite the fact that many institutions deny that their students have a drinking issue, drinking is on the rise at most campuses. Pre-game warm-up parties, post-exam celebrations, weekend happy hours, and fraternity/sorority bonding events all contribute to American college students' vibrant campus life. But no celebration here is complete without an interesting taste - alcohol.

According to the National Institutes of Health, one-third of all individuals in the United States are alcoholics or binge drinkers ("Alcohol Facts and Statistics"). The National Institutes of Health also mentioned that for the past 20 years, the rate of alcohol misuse among college students in the United States has remained around 40%, with more students preferring hard liquor to less alcoholic beer ("Alcohol Facts and Statistics"). Participating in a binge may be a chance for many college students to experience freedom for the first time in their lives. But it's the type of binge that not only fosters an alcoholic culture among college students, but also has numerous severe effects, some of which overseas students and their families cannot afford.

Alcohol abuse has a frequent effect on students' physical and mental health, causing them to miss classes, flunk examinations, lose concentration, and thus interfere with their academics. Furthermore, most students who overuse alcohol must cope with the more serious repercussions of alcohol consumption, such as legal ramifications, personal damage ramifications, and psychological difficulties linked with alcohol addiction.

Family reasons, school-induced stress, absenteeism, parental relationships and views toward alcohol, and student grades are all variables that impact students' alcohol usage. Our project aims to investigate the link between students' family status, academic position, and alcohol consumption levels. Our objective is to first discover some key parameters that impact students' total alcohol intake and then utilize these key factors to develop a machine learning model that predicts students' overall alcohol consumption levels.

Literature Review

According to Collins, socioeconomic status (SES) influences students' alcohol consumption and related outcomes. Collins explains that students with higher SES are most likely to consume more alcohol than those with lower SES. The students with higher SES consume more alcohol compared to those with lower SES because they have a higher purchasing power. The students use the extra resources they have to buy alcohol for consumption. On the other hand, students with lower SES do not have adequate funds to buy alcohol. Collins also explains that students with lower SES are more likely to be burdened with negative alcohol-related consequences than those with higher SES. Lorant et al. describe the effect of community background on alcohol consumption among students. The authors explain that some communities accept alcohol consumption which makes students from such communities get used to it (Lorant et al). In such societies, alcohol consumption is encouraged through role models, peer attitudes, mass media, and the community's attitude in general.

Bergh describes the machine learning techniques based on artificial neural networks that can be used to predict students' alcohol consumption. The author explains how doc2vec, an unsupervised neural network, can be used to capture the semantic content of text messages sent by students (Bergh). The semantic content of the text messages is encoded as numeric vectors. A machine learning approach can be used to train a logistic regression model to predict students'

alcohol consumption during distinct life phases. Bi et al. describe two machine learning approaches that can be used to predict students' alcohol consumption. The first approach uses a temporally correlated support vector machine to construct a classifier that predicts alcohol consumption. The second approach combines feature selection and cluster analysis, identifying patterns based on averaged daily drinking behavior.

Dataset Description and Exploratory Aata Analysis of the dataset

There are two datasets: student-mat.csv from the Math course and student-por.csv from the Portuguese language course, each with 33 attributes. The information was gathered through a questionnaire of secondary school students enrolled in mathematics and Portuguese language classes. This dataset includes a wealth of intriguing social, gender, and academic information about students. It may be used for EDA or to forecast a student's final grade. However, instead of predicting the student's final grade, our project aims to find the impact of social backgrounds on student's alcohol consumption level. We use these intriguing features to predict the students' workday alcohol consumption level, which is represented by Dalc, and weekend alcohol consumption, which is Walc in this dataset.

In order to better integrate the data as well as more accurately predict alcohol consumption levels among students, we merged the two sets of data together in the first phase of data organization to gain more observations to train our model. The combined dataset had a total of 1044 samples with 33 distinct attributes. Since this dataset is collected from the survey results, many attributes belong to categorical value. As a result, we must first label-encode the categorical attributes into numerical data from 0 to the number of classes minus one in order to feed them into our desired model.

We also identified some redundant attributes in the dataset. G1, G2, and G3 all represent a student's academic performance. The only difference between them is that G1 is the first period grade, G2 is the second period grade, and G3 is the final grade. After we conducted the correlation heatmap on these three features(Fig. 3.1), we noticed an extremely high correlation between the three attributes. Using highly correlated attributes makes it hard to interpret the model since it will lower the importance of the correlated attributes on target value. Moreover, when the variables are strongly co-lineared, a change in one leads to alterations in the other, causing the model results to fluctuate significantly. Because we employ the Radial basis function (RBF) kernel, tiny changes in the model will cause the model outputs to become unstable. Since final grade should be a more unbiased view of a student's academic performance, we opted to utilize G3, the students' final grade, as one of the variables among the three grade-related qualities to predict a student's alcohol consumption level.

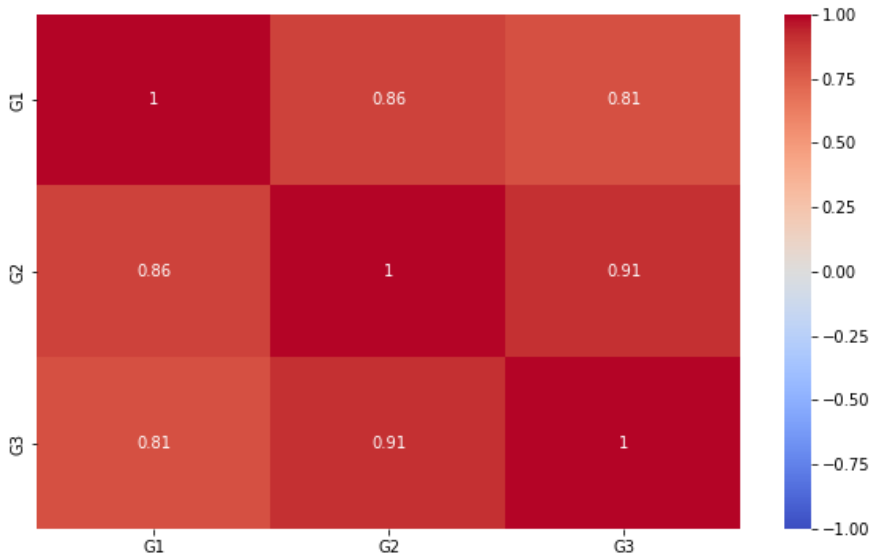


Fig. 3.1: High correlation between G1, G2 and G3

Since we are more interested in the overall alcohol consumption of the students. The second important decision we made is to combine Dalc, the weekday alcohol consumption level and Walc, the weekend alcohol consumption together into Alc, which is the alcohol consumption level. The new target value will now range from 2 to 10. We also noticed an imbalance in the different degrees of alcohol consumption levels (Fig. 3.2). In order to prevent our model from training with a biased dataset, we oversampled the dataset (Fig. 3.3) to have an even number of observations between all alcohol consumption levels. Oversampling enhances resolution and helps to eliminate aliasing and phase distortion. SMOTE (Synthetic Minority Oversampling Technique) is a popular oversampling approach for resolving imbalance issues. It seeks to balance the distribution of categories by recreating cases of randomly rising minorities.

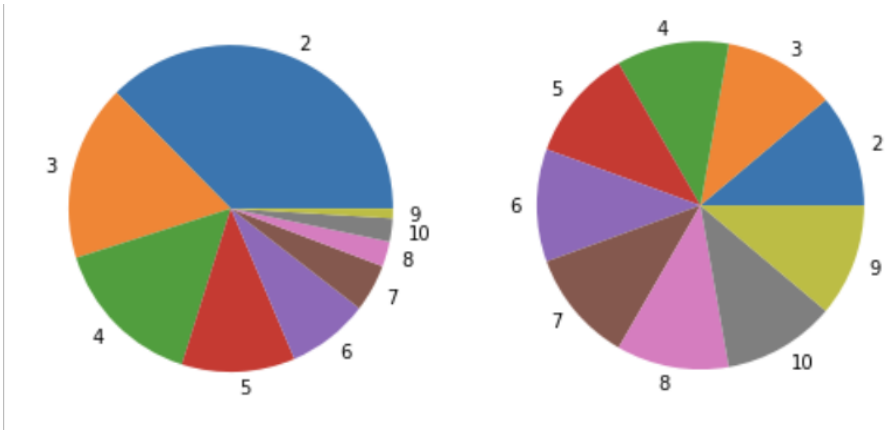


Fig. 3.2, 3.3: Alcohol consumption level pie chart before and after SMOTE oversampling

Proposed methodology

There are two main goals for our experiment. The first goal is to identify the important social element that impacts

Our initial proposal was to process the dataset first. We want to be able to process all nominal attributes into numerical attributes because we have two copies of the dataset and many of the 33 attributes were categorical.

The next step is to use a heatmap to visualize the correlation between variables. Then, we want to utilize feature selection to identify the features that are most relevant to students' alcohol consumption levels. Random Forest (RF) might be an appropriate approach in this case. Random Forest is made up of a number of decision trees. By combining bootstrap aggregation with randomization of data node selection during decision tree building, it enhances the classification performance of a single tree classifier.

We evaluated SVM, logistic regression and random forest as models.

Experimental results

Feature Selection

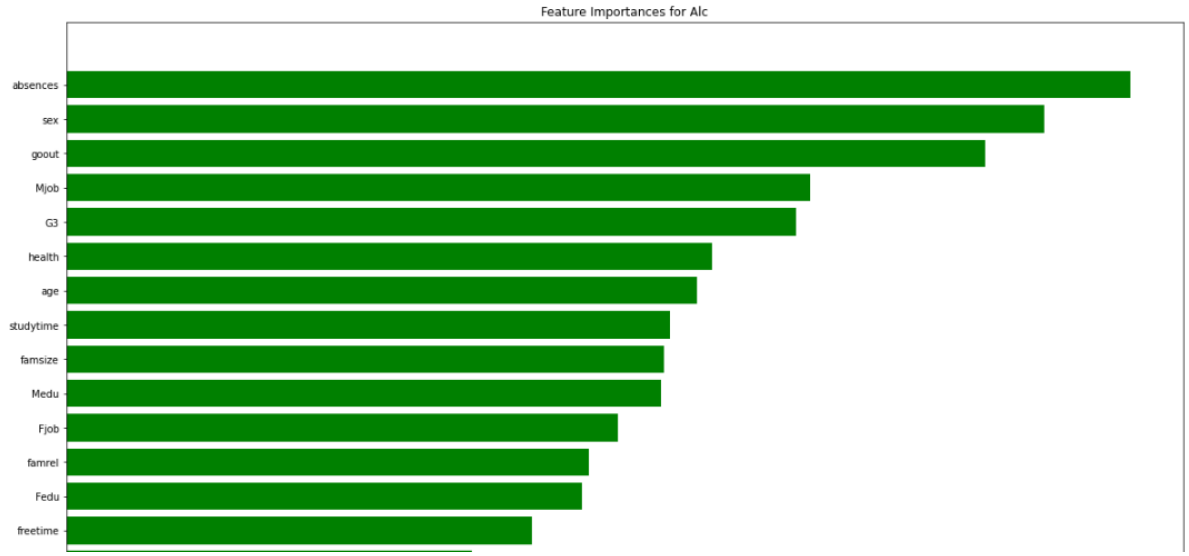


Fig. 2.6: Key Features Importances Plot

The selection of key features leads to a model with the lowest computing complexity while minimizing generalization mistakes caused by less relevant features. To determine the importance of features among the many different attributes in the dataset, we use Sklearn's RandomForestClassifier, which can collect the importance values of features so that after fitting the RandomForestClassifier model, we can access them through the feature_importances_ property. The following are some important steps:

- Splitting the data into training set and testing set, with a portion of 80:20
- Training the model RandomForestClassifier
- Obtaining the feature importance values
- Visualizing the importance of the features

The key features importance plot (Fig. 2.6) reveals some intriguing findings:

- The top three most important indicators of a student's degree of alcohol use are the number of school absences, the level of hanging out with friends, and the student's gender.
- A student's mother's job, final grade, health status, and age are all equally relevant in predicting alcohol intake.
- A student's weekly study hours, quality of family ties, father's job, family size, parental education level, and free time are somewhat more important predictors of alcohol intake.
- Factors such as the reason for selecting this school, parental educational support, whether or not they attended nursery, commute time from home to school, extracurricular activities, Internet access, romantic relationships, and parental cohabitation status showed limited predictive significance.
- By picking the most important features and discarding others, we observed a modest improvement in model performance.

The SelectFromModel class in the Scikit-learn is used to extract the best features of a given dataset depending on the importance of the weights. We decided to pick 14 attributes to use as inputs to our machine learning model by generating an estimator using the SelectFromModel class, which accepts arguments such as an estimator (RandomForestClassifier instance) and a threshold. These 14 attributes are as follows:

Logistic regression

	precision	recall	f1-score	support
2	0.43	0.43	0.43	91
3	0.28	0.32	0.30	65
4	0.22	0.12	0.15	78
5	0.34	0.28	0.31	83
6	0.41	0.32	0.36	79
7	0.33	0.38	0.36	60
8	0.60	0.70	0.65	91
9	0.71	0.95	0.81	76
10	0.56	0.63	0.59	81
accuracy			0.46	704
macro avg	0.43	0.46	0.44	704
weighted avg	0.44	0.46	0.45	704

(Fig. 3.1: Testing result of Logistic Regression)

As shown in Fig 3.1, The overall accuracy of the logistic model is 0.46, which is not a great accuracy. The main reason for the logistic model performing badly is the non-linear relationship between the selected features and the target value, which is student alcohol consumption level. This classification report indicates that we need a model that is able to capture the non- linearity between the features and the Alc value.

Random Forest

	precision	recall	f1-score	support
2	0.68	0.70	0.69	70
3	0.82	0.71	0.76	77
4	0.78	0.81	0.79	69
5	0.92	0.88	0.90	80
6	0.96	0.85	0.90	87
7	0.89	0.96	0.92	74
8	0.89	0.97	0.93	78
9	0.99	1.00	0.99	84
10	0.94	1.00	0.97	85
accuracy			0.88	704
macro avg	0.88	0.88	0.87	704
weighted avg	0.88	0.88	0.88	704

Selected Feature

	precision	recall	f1-score	support
2	0.89	0.79	0.84	92
3	0.80	0.90	0.85	62
4	0.85	0.83	0.84	89
5	0.96	0.92	0.94	76
6	0.88	0.95	0.91	76
7	0.99	0.99	0.99	73
8	0.99	0.99	0.99	89
9	1.00	1.00	1.00	81
10	0.99	1.00	0.99	66
accuracy			0.93	704
macro avg	0.93	0.93	0.93	704
weighted avg	0.93	0.93	0.93	704

All Feature

Fig. 3.2: Testing result of Random Forest

Overall, both random forest models had a great performance. The accuracy for the selected features is 0.88 and the accuracy for all features is 0.93, and we can find that they are pretty close to each other. When we observe the classification report for each level, we noticed that the difference in overall accuracy mainly comes from classifying lower alcohol consumption levels. The performance of selected feature random forest models in the low alcohol consumption level is significantly lower than the all features ones. The main reason is that we eliminated some of the features. It will cause the random forest model to come up with a lower purity leaf node, which at last leads to a worse classification result.

SVM

	precision	recall	f1-score	support
2	0.76	0.76	0.76	95
3	0.70	0.78	0.74	72
4	0.86	0.70	0.77	87
5	0.90	0.91	0.90	78
6	0.87	0.86	0.87	87
7	0.92	1.00	0.96	66
8	0.96	0.96	0.96	75
9	1.00	1.00	1.00	69
10	0.96	0.99	0.97	75
accuracy			0.88	704
macro avg	0.88	0.88	0.88	704
weighted avg	0.88	0.88	0.87	704

Selected Feature

	precision	recall	f1-score	support
2	0.56	0.69	0.62	64
3	0.79	0.72	0.75	76
4	0.87	0.88	0.87	66
5	0.94	0.92	0.93	95
6	0.98	0.93	0.95	87
7	0.96	0.93	0.95	75
8	0.99	0.99	0.99	77
9	1.00	1.00	1.00	83
10	1.00	0.98	0.99	81
accuracy			0.90	704
macro avg	0.90	0.89	0.89	704
weighted avg	0.91	0.90	0.90	704

All Feature

Fig. 3.3: Testing result of SVM

SVM models also have great performance on both models. The gap between overall accuracy of the two models is even less than the random forest ones. Since SVM is a mathematical model, it won't impact its performance a lot when we eliminate the less important features. Another interesting discovery is that the accuracy for selected feature models is better than all feature models in classifying lower alcohol consumption levels. The reason might come from there being more noisy data in lower alcohol consumption levels. The all features model is

more vulnerable to the outliers and is trained to be more biased by noisy data. Thus, it ends up with a worse performance result.

KNN

	precision	recall	f1-score	support
2	0.73	0.58	0.65	79
3	0.69	0.79	0.73	70
4	0.87	0.80	0.83	83
5	0.91	0.91	0.91	68
6	0.89	1.00	0.94	65
7	0.94	0.99	0.97	85
8	1.00	0.98	0.99	87
9	0.95	1.00	0.98	82
10	1.00	0.99	0.99	85
accuracy			0.89	704
macro avg	0.89	0.89	0.89	704
weighted avg	0.89	0.89	0.89	704

Selected Feature

	precision	recall	f1-score	support
2	0.73	0.71	0.72	72
3	0.82	0.85	0.83	79
4	0.82	0.88	0.85	88
5	0.85	0.77	0.81	74
6	0.88	0.94	0.91	72
7	0.96	0.88	0.92	81
8	0.97	0.99	0.98	79
9	0.98	1.00	0.99	84
10	0.96	0.95	0.95	75
accuracy			0.89	704
macro avg	0.89	0.88	0.88	704
weighted avg	0.89	0.89	0.89	704

All Feature

Fig. 3.4: Testing result of KNN

Setting the K = 3 to construct the KNN models for selected Features and all features, we can see that removed features seem to have little impact on the overall performance in the KNN method as the overall accuracy has no difference. We can also see that the precision and f1 score at a lower level does decrease a lot, this may imply that some of the removed features may have some impact on the alcohol consumption at the lower level but no significant impact when the consumption level is high.

MLP

	precision	recall	f1-score	support
2	0.50	0.53	0.52	92
3	0.38	0.29	0.33	86
4	0.51	0.52	0.52	71
5	0.65	0.65	0.65	78
6	0.76	0.80	0.78	82
7	0.73	0.84	0.78	55
8	0.94	0.93	0.93	85
9	0.96	0.96	0.96	82
10	0.95	0.95	0.95	73
accuracy			0.71	704
macro avg	0.71	0.72	0.71	704
weighted avg	0.70	0.71	0.71	704

Selected Feature

	precision	recall	f1-score	support
2	0.50	0.32	0.39	79
3	0.48	0.55	0.52	76
4	0.55	0.49	0.52	85
5	0.61	0.59	0.60	94
6	0.71	0.78	0.74	87
7	0.68	0.92	0.78	60
8	0.90	0.89	0.89	71
9	1.00	0.95	0.97	79
10	0.86	0.92	0.89	73
accuracy			0.70	704
macro avg	0.70	0.71	0.70	704
weighted avg	0.69	0.70	0.69	704

All Feature

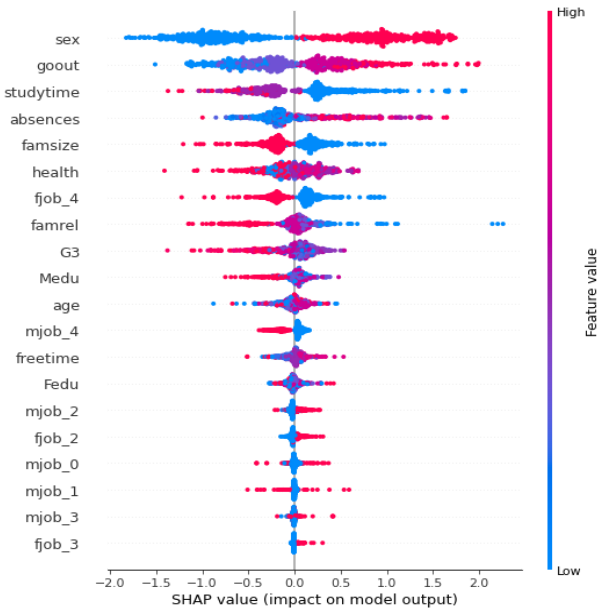
Fig. 3.5: Testing result of MLP

Setting the four hidden layers with (20, 40, 25, 10) nodes on the layers and setting the learning rate to 0.01, we construct the MLP models for selected Features and all features. The overall performance of both models seems to have lower accuracy, which may indicate that layer setting can still be optimized. We can see that the overall performance of the model with selected features is 0.01% higher than all features. This means that the selected feature we had seems to better fit the model layout we have settled.

Conclusion and discussion

One of the limitations is the values in the original data, for example, the consumption level in the original data was collected by survey. As people seem to have different standards when describing their personal consumption levels. Therefore, we decided to treat the Alcohol consumption level in the model as a categorical value. By doing this the model we trained has relatively better accuracy in predicting the consumption level as the difference between each consumption level is not clear and unscaled. This is clearly shown in the data when the consumption level is low, the data seems to have a lot of noise and precision in each of the models have lower accuracy of prediction in lower level.

By comparing all the models with all features and selected features, the features we selected are good indicators to predict the alcohol consumption value. In order to know how related the alcohol consumption level is, we output the feature importance graph with the SHAP value calculated from the model we had. Based on the graph, we can find that some features can be clearly observed to have a positive or negative effect on alcohol consumption levels like sex, go out times, studying time, absences, family size, and family relationships. For example, in the feature value of study time, people who drink more alcohol usually spend less time studying. We can also see that people who have large family sizes seem more likely to have higher alcohol consumption. And people who spend more time going out with friends and less free time after school seem more likely to consume more alcohol. And being a male seems more likely to consume more alcohol which is probably due to some physiological effect. The father's job and the mother's job are special features here. There are four types of jobs that have positive impacts: teacher, healthcare-related, civil services, and at-home jobs. However, we contribute other jobs to 'others'(fjob_4 and mjob_4), and it has a negative impact on alcohol consumption level.



(Fig. 4.1: Feature Importance graph)

From the information we find above, we can conclude that people who have more social activities and more social interaction in regular life seem to have more alcohol consumption. We

can also see that people who spend less time in study, for example, more absence, or less time studying have a greater chance of having high-level alcohol consumption. And what seems to be irrational is that student health status seems to have no clear impact on alcohol consumption and students seem to drink more when their health condition is great. This is probably caused by students are usually young adults and their physical condition is all relatively the healthiest in their whole life. This will make students become more careless about their physical condition when taking alcohol. And consume alcohol as much as they can. This shows why students will the greatest health condition consumes the most alcohol.

In our study, Random Forest, SVM and KNN all performed well, boasting an accuracy rate of about 90 percent while the Logistic model did not perform well as it cannot capture nonlinearity. The accuracy obtained by training selected features with three models is not very different from that obtained by training all features. We finally selected Random Forest as our training model because of its high accuracy and the fact that this model predicts all drinking levels well compared to other models.

The purpose of our study is not to ban alcohol consumption by students altogether, but to reduce the negative effects of alcohol consumption on students by identifying the factors that lead them to drink. Schools and families can use our model to predict a child's level of drinking and thus reduce alcohol abuse.

References

"Alcohol Facts and Statistics." *National Institute on Alcohol Abuse and Alcoholism*, U.S. Department of Health and Human Services, https://www.niaaa.nih.gov/publications/brochures-and-fact-sheets/alcohol-facts-and-statistics.

Bergh, Adrienne. "A Machine Learning Approach to Predicting Alcohol Consumption in Adolescents from Historical Text Messaging Data." *Chapman University Digital Commons*, https://digitalcommons.chapman.edu/cads_theses/2/.

Bi, J., Sun, J., Wu, Y., Tennen, H., & Armeli, S. "A machine learning approach to college drinking prediction and risk factor identification." ACM Transactions on Intelligent Systems and Technology (TIST), 4(4), 1-24. 2013. https://dl.acm.org/doi/10.1145/2508037.2508053

Collins, Susan E. "Associations Between Socioeconomic Factors and Alcohol Outcomes." Alcohol research : current reviews vol. 38,1 (2016). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4872618/

Lorant, V., Nicaise, P., Soto, V.E. et al. Alcohol drinking among college students: college responsibility for personal troubles. BMC Public Health 13, 615 (2013). https://doi.org/10.1186/1471-2458-13-615