What’s Your Poison?

Feature Selection and Prediction of Student Alcohol Consumption

Mario Buczkowski  
 School of Computing and Information   
 University of Pittsburgh  
 Pittsburgh, PA, USA  
 mdb99@pitt.edu

Lucy Chen  
School of Computing and Information   
 University of Pittsburgh  
 Pittsburgh, PA, USA  
 yic76@pitt.edu

Akshay Madan  
School of Computing and Information   
 University of Pittsburgh  
 Pittsburgh, PA, USA  
akm88@pitt.edu

ABSTRACT

Underage drinking is a serious issue many countries face, including Portugal with a low starting drinking age of 14 years old. Not only can it impair brain development of adolescence, but it can also induce negative behaviors. There are many reasons and factors that contributes to underage drinking, from personal to environmental situations. In this project, several feature selection methods and machine learning classification models were applied to the “Student Alcohol Consumption” dataset to predict the alcohol consumption behavior and important demographic, social and school related attributes that influences the drinking behavior of secondary school students in Portugal. The results show that Logistic Regression performed the best amongst all other classification models with an accuracy of 88%, recall of 52%, and precision of 52%. In addition, the most important attributes that impact the alcohol consumption behavior of these students are gender, how often students go out with their friends, absence from school, and age.

1 INTRODUCTION

Alcohol, a popular social, celebratory, and even relaxing drink, can lead to serious health issues if consumed excessively either from heavy drinking or binge drinking. Some of the detrimental health issues include but not limited to liver disease, dementia, and cardiovascular failure [12]. Consuming alcohol as a teenager is a more serious issue because alcohol addiction at such a young age will impede brain development, and induce poor behaviors such as violence, criminal actions, and motor vehicle accidents. The study “Adolescent Alcohol Use: Risks and Consequences” by E. Jane Marshall concluded that heavy alcohol consumption in adolescences tends to persist into adulthood and lead to alcohol problems such as dependence, premature death, and diminished work capacity [4]. According to the National Institute on Alcohol Abuse and Alcoholism (NIAAA), every year approximately 5,000 young adults under the age of 21 die from motor vehicle crashes, homicides, suicides, and other form of injuries due to underage drinking [6]. Despite the consequences, more and more adolescents continue to drink. But why?

Research by students at Uppsala University on “Why Do Adolescents Drink? Motivational Patterns Related to Alcohol Consumption and Alcohol-Related Problems” identified three motives for underage drinking: social enhancement, coping with negative emotions and tension reduction, and dominance [2]. NIAAA details other reasons for increasing underage drinking including developmental transition to adolescence (puberty and increasing independence), expectation of alcohol and its consequences, personality and psychiatric comorbidity, hereditary factors, and environmental factors (parents and peers). There are many intervention programs sponsored by the government that are trying to reduce the underage drinking rate and these programs require continuing investigation of factors that contribute to underage drinking in order to provide the most effective intervention strategies.

This project aims to predict alcohol consumption behavior and to determine the factors that contribute to underage drinking of secondary students in Portugal using machine learning classification techniques. Data, gathered by Paulo Cortez and Alice Silva from the University of Minho, with school reports, several demographics, social and school related information, and alcohol consumption of students in a Math course and a Portuguese course from two public schools in Alentejo region of Portugal is used to conduct this classification project. Given alcohol consumption (weekend alcohol consumption and workday alcohol consumption) as the target variable and 31 demographics, social and school related attributes as features, we will preprocess the dataset, conduct feature engineering, and train many machine learning classification models to classify students as either drinkers or non-drinkers and determine the most significant features involved in classifying these students.

2 LITERATURE REVIEW

Fabio Pagnotta and Mohammad Amran Hossain from the University of Camerino used business intelligence and data mining techniques to classify drinking habits of students as either a drinker or not a drinker [5]. During the preprocessing stage, they discarded the family income feature due to lack of values and live with parent and personal computer features because almost 100% of the students responded yes to them. They merged the two different alcohol attribute, workday alcohol consumption and weekend alcohol consumption, into one attribute that represents the total weekly alcohol consumption, and then convert tit to binary value with lower than 3 drinks as 0, otherwise 1. They also converted absences into binary value (over 10 days as 0, otherwise 1). The 14 most important features were selected using linear correlation and loop with cross validation. As for the models, they used decision trees, including Random forest with an accuracy of 92%.

A group of faculty members from the Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Malaysia classified alcohol addiction with this dataset [8]. In the preprocessing step, they selected 14 features based on Fabio Pagnotta and Mohammad Amran Hossain’s research and cleaned the data by removing outliers, resolving data inconsistencies, and normalizing the dataset using Z-transformation. As for the classification algorithms, they used multilayer perceptron neural network (MLP) and self-tuning multilayer perceptron classifier (AutoMLP) with 10-fold cross validation. The MLP produced an accuracy of 61.78%, squared error of 0.360 and root mean squared error of 0.599, while the AutoMLP produced an accuracy of 64.54%, squared error of 0.322, and root mean squared error of 0.566.

In their paper [3], P. Cortez and A. Maria try to predict the students’ achievements using this database. They were the ones to collect this dataset. They used four datamining models (Decision Trees, Random Forest, Neural Networks and Support Vector Machines). They are able to develop a model that gives good predictive accuracy. This assumes that their grades of the first or second school are available. Their performance is found to be dependent on not only the past evaluations but also on other features like parent’s job and education, alcohol consumption, etc. The authors suggest this can be used to improve the quality of education by enhancing school resource management.

Another paper by Jessica Maikhanh Brown [1], uses KNN to predict the score based on the features. She categorizes the scores of the students. First, she adds them up to get the total score. Then, she creates bins on the bases of the average and how much above or below the average, a student’s net score is. Finally, she uses Weka to run the KNN algorithm on the dataset to predict the category of marks the student falls in. The accuracy is best when the K for the model is set to be 8, 48.87%.

Fabio Mendoza and other colleagues at the University of the Coast, Colombia, utilized this dataset to try and predict student alcohol consumption using Support Vector Machines, Decision Trees, Naïve-Bayes and KNN. This paper [5], focuses mainly on a comparison of metrics other than accuracy, namely True-positive rate, False-positive rate and recall. K-means Clustering was utilized to divide the data into the two categories trying to be predicted. The implementation of Support Vector Machines led to the highest overall percentages for precision, recall, and false-positive rate. The authors here stressed the importance of utilizing k-means clustering to partition the data prior to implementing any data mining algorithms.

Saurabh Pal and Vikas Chaurasia [9] utilized sequential minimal optimization, REP tree and decision table to look into the classifying students that indulge into excessive drinking. Their main focus was on finding a correlation between alcohol consumption and poor academic performance and in turn, find methods to rehabilitate students effected by alcohol by finding the underlying causes. Comparing all built models, the once built using bagging performed the best with 80.2% accuracy. Mean absolute error and root mean squared error were also used to further measure the model's performance. The Chi-squared test and Information Gain test were applied to measure the impact of variables on the model. Age and grades were among the more important features. Legal access to alcohol seemed to contribute to a large amount of how much alcohol a student consumes.

E. Jane Marshall conducted literature reviews on many studies on adolescent alcohol use [4]. Through her research, she categorized the risk factors into two groups: Environmental factors and Genetic factors. Environmental factor consists of increasing peer influences and decreasing family influences due to adolescents becoming more independent. Adolescents see peer acceptance as a high social reward and will drink and perform other risk-taking behaviors to feel more accepted. Genetic factors include family situation and lifestyle. Adolescents raised in a family with alcohol problem will most likely drink than those raised in a family without alcohol problem. Those suffering from depression or psychological dysregulation tend to be vulnerable to Alcohol Use Disorder.

3 METHOD

3.1 Exploratory Data Analysis

The datasets used for this project were composed by Paulo Cortez and Alice Silva from the University of Minho, Portugal [4]. They collected school reports and questionnaires about several demographic, social, and school related information from students in the Portuguese and Math courses at two public schools in the Alentejo region of Portugal due to the lack of success in these two courses. Originally, Cortez and Silva used this dataset to determine the factors that contribute to the academic success and failures of these students and to predict the students’ final grades; however, we think it would be more interesting to predict alcohol consumption and determine which attributes they collected impact the drinking behaviors of students.

There are two datasets given, one from the Math course and one from the Portuguese course. The math course dataset contains 395 instances while the Portuguese course dataset contains 649 instances. Each instance corresponds to a student in the respective class. However, there are 382 students that take both courses, hence are in both datasets. There are 33 demographics, social, and school related attributes with two of them relating to alcohol consumption: Dalc (weekday alcohol consumption) and Walc (weekend alcohol consumption). The alcohol consumption value is based on a numerical rating on a scale from 1 to 5 with 1 indicating very low consumption and 5 indicating very high consumption. There are no missing values nor are there unique identification values for the students. There are 17 object datatype (qualitative) attributes and 16 integer datatype (quantitative) attributes. Some of the qualitative attributes are ratings from 1 to 5 and others are continuous over a wider range of values such as age and absences. The dataset is shown in the table below.

Table 1. All variables and descriptions of variables in the dataset collected by Cortez and Silva

|  |  |
| --- | --- |
| Variables | Details |
| school | student's school ('gp' - gabriel pereira or 'ms' - mousinho da silveira) |
| sex | student's sex ('f' - female or 'm' - male) |
| age | student's age (15 to 22) |
| address | student's home address type ('u' - urban or 'r' - rural) |
| famsize | family size ('le3' - less or equal to 3 or 'gt3' - greater than 3) |
| pstatus | parent's cohabitation status ('t' - living together or 'a' - apart) |
| medu | mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education) |
| fedu | father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education) |
| mjob | 1. mother's job ('teacher', 'health' care related, civil 'services', 'at\_home' or 'other') |
| fjob | father's job ('teacher', 'health' care related, civil 'services', 'at\_home' or 'other') |
| reason | reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other') |
| guardian | student's guardian ('mother', 'father' or 'other') |
| traveltime | home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour) |
| studytime | weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours) |
| failures | number of past class failures (numeric: n if 1<=n<3, else 4) |
| schoolsup | extra educational support (binary: yes or no) |
| famsup | family educational support (binary: yes or no) |
| paid | extra paid classes within the course subject (math or portuguese) (binary: yes or no) |
| activities | extra-curricular activities (binary: yes or no) |
| nursery | attended nursery school (binary: yes or no) |
| higher | wants to take higher education (binary: yes or no) |
| internet | internet access at home (binary: yes or no) |
| romantic | with a romantic relationship (binary: yes or no) |
| famrel | quality of family relationships (numeric: from 1 - very bad to 5 - excellent) |
| freetime | free time after school (numeric: from 1 - very low to 5 - very high) |
| goout | going out with friends (numeric: from 1 - very low to 5 - very high) |
| dalc | workday alcohol consumption (numeric: from 1 - very low to 5 - very high) |
| walc | weekend alcohol consumption (numeric: from 1 - very low to 5 - very high) |
| health | current health status (numeric: from 1 - very bad to 5 - very good) |
| absences | number of school absences (numeric: from 0 to 93) |
| g1 | first period grade (numeric: from 0 to 20) |
| g2 | second period grade (numeric: from 0 to 20) |
| g3 | final grade (numeric: from 0 to 20, output target) |

3.2 Preprocessing and Feature Engineering

*3.2.1 Target Variable*

The target variable is given to us as two variables: Dalc and Walc. Based on the study by Fabio Pagnotta and Mohammad Amran Hossain and other studies, we decided to combine the weekday and weekend alcohol consumption and find the combined average daily alcohol consumption (SumAlc) [1]. Since there are 5 weekdays and 2 weekend days, we calculated ‘SumAlc’ as:

The way ‘SumAlc’ was calculated let us maintain the same scale the alcohol consumption values were given to us. Then, we split the average daily consumption into two classes, drinkers and non-drinkers, based on ‘SumAlc’ and our selected threshold of 3. If the consumption is less than 3, then we consider that student to be a non-drinker; if the consumption is greater than or equal to 3, we consider that student to be a drinker.

*3.2.2 Features*

For the qualitative features, we label encoded them to allow machine learning models to better process the data. For each of the features with only two categories (school, gender, address, family size, parent’s cohabitation status, extra educational support, family educational support, extra paid classes within course subject, extra-curricular activities, attended nursery school, wants to take higher education, internet access at home, and with a romantic relationship), we assigned one value as 0 and the other as 1. This set of features also includes variables with binary values with “no” set to 0 and “yes” set to 1. For the qualitative features with more than two categories (mother’s job, father’s job, reason to choose this school, and student’s guardian), we label encoded them with respect to the number of classes there are with 0 assigned to the “other” class since that class seemed broad. As for the quantitative features, there are no outliers or noisy data that needs to be removed since many of the features are on a scale.

After transforming the qualitative values into numerical values, we performed feature selection using p-values calculated from ordinary least square model. For the Portuguese course dataset, there are six features with p-values less than 0.05. These features include ‘sex’, ‘famsize’, ‘Pstatus’, ‘famrel’, ‘goout’, and ‘health’. However, for the math course dataset, there are 8 features with p-values less than 0.05: ‘sex’, ‘address’, ‘famsize’, ‘studytime’, ‘paid’, ‘famrel’, ‘goout’, and ‘absences.’ This difference might be due to the imbalanced number of instances since the Portuguese course has almost twice as many students as the math course. It might also be due to the course itself and type of students that are taking either or both. Most of the students surveyed form the math course are in both, but only half of the students surveyed in the Portuguese course are in both; hence the other half of the Portuguese course are taking Portuguese only. Due to the disparity between the two datasets, we merged the two datasets into one and worked with the merged dataset throughout the rest of the project.

*3.2.3 Merging Datasets*

To merge the two datasets, there needs to be an identifier that is unique to every student. Given that there is no name provided in the dataset for privacy reasons, we decided to use a hash of a few fields which are less likely to repeat for the students. Given that these values combined will be unique for different students, the hash of a string that is the fields concatenated them will also be unique for all students. So, the first job is to add a column that is the hash. Now, in the combined data frame we will have students that are present in both or in either. Two more columns are added that tell if a student is enrolled in Math course or in Portuguese course. The value is 1 if he/she is enrolled the course. Also, the grades for Math and Portuguese have to be separated since they are with respect to the course. For the students that are present in both courses, the grades are present in their respective columns. For others, the column of the course they are not enrolled in is empty. For the actual merging, we found out the intersection of the hashes of both data frames to find the common students. Then, a copy of the Math dataset is created which will later become the final dataset. Lastly, the hash values that are present in the intersection but do not appear in the Portuguese data frame are added to the copy of Math. This copy becomes our final merged data frame.

The newly merged dataset contains missing values for the grades attributes (G1\_por, G2\_por, G3\_por, G1\_mat, G2\_mat, G3\_mat) since they are with respect to the courses and not everyone was enrolled in both courses. Because the p-values for grades were not significant when the dataset was split, we decided to drop them.

Looking at the correlation matrix, the two features with the highest correlation with ‘SumAlc’ are ‘sex’ (0.3) and ‘goout’ (0.2).

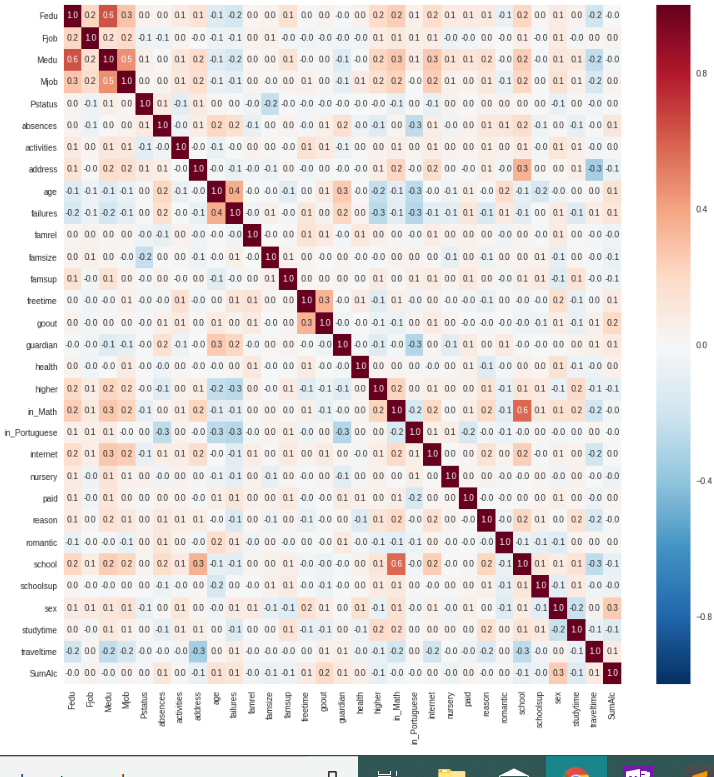


Figure 1. Correlation matrix heatmap of all the preprocessed variables in the merged dataset.

A deeper look into the gender separation between drinkers and non-drinkers shows that males are more likely to be drinkers than females.

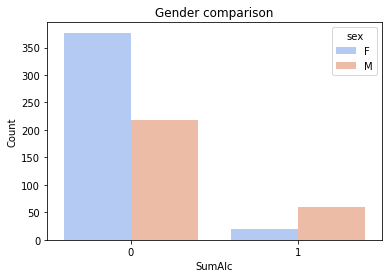


Figure 2. Count of female and male drinkers in the dataset

As for going out comparison, the number of drinkers increase as the going out rate increases.

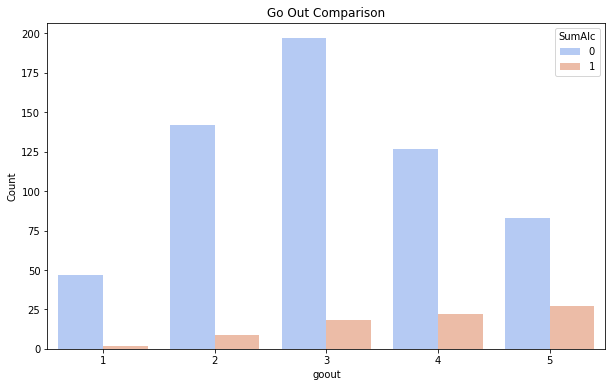


Figure 3. Count of drinkers and non-drinkers dependent based on going out rate

*3.2.4 Feature Selection*

Three methods were used to select the most important features and reduce the dimensionality of the data. The first method selected important features using p-values calculated from ordinary least square model Only features with p-value less than 0.05 were considered to be significant. There are five features that satisfy this condition: ‘sex’, ‘absences’, ‘age’, ‘goout’, and ‘in\_Portuguese’. However, ‘in\_Portuguese’ was dropped after looking into the built-in feature importance function of the Random Forest and Decision Tree Model and that improved many of the models. The remaining four features become our p-value selected feature subset. It is no surprise that the ‘sex’ and ‘goout’ features are part of this subset since the two have the highest correlation with ‘SumAlc’ as seen in the correlation matrix.

The second method selected most important feature using RFE (recursive feature elimination) on a model, which we chose to be Support Vector Machine [10]. RFE recursively constructs a model and removes the most important feature, based on the feature importance estimator, in that set of features, and repeats the process with the remaining features until the features are exhausted. The features removed are ranked according to when that feature was eliminated. The rankings were printed, and we utilized the top 10 ranked features as our RFE selected feature subset. The 10 ranked features are 'goout', 'absences', 'age', 'sex', 'traveltime', 'studytime', 'health', 'famrel', 'in\_Portuguese', 'school', 'guardian'.

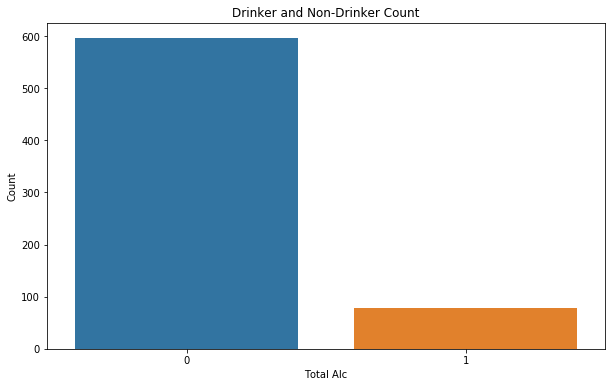
The third method reduced the dimensionality by using Principal Component Analysis (PCA). PCA reduces the feature space to one that contains only the subset of features (known as principal components) that accounts for as much of the variability in the data as possible [11]. The number of principal components was specified to 4 components and 10 components in attempt to match the number of features in the subset form the previous two methods

*3.2.4 Scaling and Splitting*

The dataset is split into training and testing sets so that parameter tuning can be performed with training set and model performance can be fairly evaluated with testing set. Because not all the features are on the same scale, scaling was used. Both Standard Scaler and Min-Max Scaler were tested.

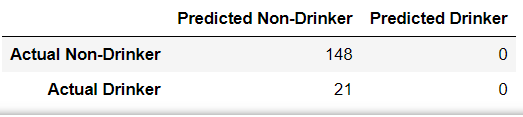
3.3 Balancing the Dataset

One of the main issues with using this dataset is the imbalance of data. There are 78/674 (12%) instances labeled as drinkers and the rest are non-drinkers as shown below.



**Figure 4. Count of drinkers and non-drinkers in the merged dataset. This shows the imbalanced of the data**

Due to this imbalance of data, the models all initially underfitted and predicted all test set to be 0 (non-drinker), as shown in the confusion matrix below, which led to accuracies around 88% for each, but precision and recall values no greater than 0.



**Figure 5. Confusion matrix of one of the many models that underfitted and predicted all testing set data to be 0**

The solution to this problem is using oversampling methods to balance the dataset. Three oversampling methods were tested to balance the training set. The first is using a resampling technique and repeatedly randomly chose drinker instances from the training set to append to the train dataset until there is an equal number of drinker and nondrinker instances in the training set. The second method is using SMOTE, which generates synthetic data based on the training set using n-nearest neighbors. The last method is using ADASYN which is similar to SMOTE in that it adds synthetic data but there is also an additional small value added to make the synthetic data more like a real value. Due to the imbalance of the dataset, the performance of all the models were measured by not only the accuracy, but also the recall and precision. We were careful to use these methods only on the training data. Initially, we applied oversampling methods on the entire dataset before splitting it into training and testing set, but it became clear that training set data was bleeding into the testing set. The non-parametric methods were predicting at a very high percentage for accuracy, precision, and recall, and it was clear that something was wrong. We re-evaluated our results and applied over-sampling only to the training data which resulted in more realistic results and tweaked our models to best obtain new results.

3.4 Machine Learning Models

*3.4.1 Overview*

Various classifiers were chosen and applied to the dataset after all the preprocessing, feature engineering, and balancing of the dataset were applied. Each model was applied with one or two tuning parameters and the rest default. The usage for loops and10-fold Cross validation with f1-score as the scorer were applied to all models for finding the optimal tuning parameters values. The performance of the models is evaluated by recall and precision score in addition to accuracy. The models were trained with all combinations of resampling technique and subsets of features from earlier preprocessing steps. All the classification models used, and the tuning parameters altered are listed below. Bagging and boosting assembling technique were applied to Decision Tree and Random Forest.

**Table 2. All the classification models applied, and the tuning parameters altered**

|  |  |
| --- | --- |
| Model | Tuning Parameters |
| K-nearest neighbors | n- neighbors |
| Decision Tree  (plus Bagging and AdaBoost) | Max depth |
| Random Forest (plus Bagging) | Max depth |
| Logistic Regression | Regularization penalty (Ridge vs Lasso), regularization parameter C |
| Linear Discriminant Analysis | Default |
| Multilayer Perceptron | Activation function, sovler |
| Support Vector Machine | Kernel, regularization parameter C |
| Extreme Gradient Boosting | Max\_depth, min child weight, scale positive weight |

*3.4.2 Threshold Change*

Due to the imbalanced dataset, the probability threshold was also adjusted after fitting each model. It was desirable to improve the precision and the recall of the models as far as possible with the new threshold. So, the Precision-Recall Curve was plotted to find which point on the curve, i.e. which threshold value gave the best recall and the best precision. Once the threshold was detected, it was used for the classification.

4 RESULTS

After applying all the possible combinations of oversampling techniques and feature selection methods, results show that Min-Max Scaler performed better than Standard scaler and oversampling using the resampling method performed better than the SMOTE and ADASYS methods. In addition, using PCA feature dimensionality reduction did not improve the performance of the models like RFE features and P-value selected features did. The following three charts showcases the performance results of all the classification models with the resampling method applied on the main three sets of features.

Figure 6. Accuracy, precision, and recall scores of all the models with and without change in threshold using all the features



Figure 7. Accuracy, precision, and recall scores of all the models with and without change in threshold using RFE selected features subset

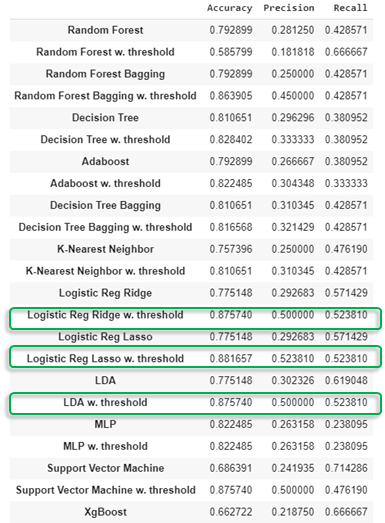


Figure 8. Accuracy, precision, and recall scores of all the models with and without change in threshold using all p-value selected features subset

As shown in the results above, Logistic Regression with Ridge and Lasso regularization performed the best across all subset of features with precision and accuracy above 0.50 when the threshold was changed to 0.68. Without the change in threshold, majority of the models performed poorly. However, Logistic Regression (with Lasso Regularization) performed better with all the features than with RFE feature subset or p-value selected feature subset. This makes sense because the more features there are, the more flexible and more fitted the model becomes. Even with all the features, the model did not overfit to the training data, as proven by the highest accuracy, recall, and precision. Although Logistic Regression performed the best with all the features, its performance with the RFE feature subset and the p-value selected feature subset did not stray far behind; the difference in performance is only due to difference in one or two predictions. This indicates that the features that explains majority of the variability in the alcohol consumption behavior of students in this dataset are the four p-value selected features. This can be further supported with the increase in performance of LDA with a subset of the important features. LDA did not perform as well as Logistic Regression with all the features, however it did perform just as well as Logistic Regression with a smaller subset of the features. This suggests that the LDA model was overfitting with all the features but fitted well with reduced number of features. The four important features that had a major effect on the alcohol consumption behavior are ‘gender’, ‘goout’, ‘age’, and ‘absences’. It is not surprising that ‘gender’ and ‘goout’ are significant factors in this prediction since the two have the highest correlation coefficient with ‘SumAlc’ as shown in the correlation matrix earlier. We tried Extreme Gradient Boosting due to its success on many Kaggle competitions and also because it was expected to work well with an imbalanced dataset such as ours. This is because it has a special tuning parameter called scale positive weight that precisely helps remove the bias in the model due to large number of negative or positive values. Although, when applied to the dataset, this model did not seem to work as well as it was expected to. Overall, the best model for this dataset is Logistic Regression with the parameters and finals results shown below.

Table 3. Best Logistic Regression parameters and finals results with p-value selected features subset

|  |  |
| --- | --- |
| Logistic Regression | |
| Regularization parameter C | 1 |
| Solver (default) | ‘liblinear’ |
| Probability Threshold | 0.68 |
| Accuracy | 0.88 |
| Precision | 0.52 |
| Recall | 0.52 |

5 CONCLUSION

Underage drinking is an issue that many countries are trying to curb especially since according to the Center for Disease Control and Prevention, general teenage alcohol consumption is on the decline but binge drinking among admitted drinkers is increasing [12]. There are many intervention programs aimed at limiting underage drinking. In order to successfully do so, one must discover and focus on the factors that contribute to underage drinking. There are several studies that have concluded certain elements that impact underage drinking and can help with intervention programs.

With the dataset provided by Cortez and Silva, the goal of this project was to predict alcohol consumption behaviors of students in Portugal and determine the factors that impact their drinking habits. Several feature selection and dataset balancing techniques were tested on various machine learning classification models. Overall, the performance of the models varied but the logistic regression (both ridge and the lasso regularization) performed the best with the adjusted probability threshold of 0.68. The most important features that affect prediction of alcohol consumption behaviors were determined to be age, how often students go out with their friends, absences from school, and age. The attributes age, gender, and how often students go out with their friends, align with the ones determined by NIAAA. As for absences, one can assume that students who miss school a lot probably spend more time going out with friends and drinking, and care less about school and as a result make the poor decision to not attend. Although this dataset was initially used for predicting student's grades, and not alcohol consumption, it did allow us to predict alcohol consumption to an extent with fair results. Further research with a larger dataset that is intended for studying adolescent alcohol consumption is required to better predict alcohol consumption behaviors and understand more about underage drinking.

REFERENCES

[1] Brown, Jessica Maikhanh. Predicting math test scores using K-Nearest Neighbor. In 2017 IEEE Integrated STEM Education Conference (ISEC), pp. 104-106. IEEE, 2017.

[2] Comasco, Erika & Berglund, Kenneth & Oreland, Lars & Nilsson, Kent. (2010). Why Do Adolescents Drink? Motivational Patterns Related to Alcohol Consumption and Alcohol-Related Problems. Substance use & misuse. 45. 1589-604. 10.3109/10826081003690159.

[3] Cortez, Paulo, and Alice Maria Gonçalves Silva. "Using data mining to predict secondary school student performance." (2008): 5-12.

[4] E. Jane Marshall. Adolescent Alcohol Use: Risks and Consequences, Alcohol and Alcoholism. Volume 49, Issue 2, March/April 2014, Pages 160–164, https://doi.org/10.1093/alcalc/agt180

[5] Fabio, Mendoza-Palechor, De la Hoz-Manotas Alexis, Morales-Ortega Roberto, Martínez-Palacio Ubaldo, Díaz-Martínez Jorge, and Combita-Niño Harold. "Research Article Designing A Method for Alcohol Consumption Prediction Based on Clustering and Support Vector Machines." Research Journal of Applied Sciences, Engineering and Technology 14, no. 4 (2017): 146-154.

[6] National Institute on Alcohol Abuse and Alcoholism (NIAAA). U.S. Department of Health and Human Services, www.niaaa.nih.gov/.

[7] Pagnotta, F, Amran Hossain H. Using Data Mining to Predit Secondary School Student Alcohol Consumption. http://www3.dsi.uminho.pt/pcortez/student.pdf

[8] Shamala Palaniappan, Norhamreeza A Hameed, Aida Mustapha, Noor Azah Samsudin. (2017). “Classification of Alcohol Consumption among Secondary School Students”.

[9] Pal, Saurabh & Chaurasia, Vikas. (2017). Performance Analysis of Students Consuming Alcohol Using Data Mining Techniques. IJARSE. 6. 13. Conference

[10] Brownlee, Jason. (2014). Feature Selection for Machine Learning in Python. <https://machinelearningmastery.com/feature-selection-machine-learning-python/>

[11] Brems, Matt. A One-Stop Shop for Principal Component Analysis. <https://towardsdatascience.com/a-one-stop-shop-for-principal-component-analysis-5582fb7e0a9c>

[12] Excessive Alcohol Use. Center for Disease Control and Prevention. https://www.cdc.gov/chronicdisease/resources/publications/aag/alcohol.htm