I. Introduction

To better support our youth, understanding what influences a young person to use alcohol will be useful in developing effective prevention strategies. This information is valuable not only to policy makers but could provide parents and guardians guidance into how to help their children avoid alcohol abuse and its related consequences.

The 2020 National Survey on Drug Use and Health is a cross-sectional survey, the primary purpose is to measure the prevalence of substance abuse. It is a wealth of data spanning back to 1971. The population surveyed were individuals aged 12 or older. This project utilized a subset of the larger dataset. This subset included demographic measures of race, grade at the time of the survey, sex, household income, poverty level, population density of residence and metro versus nonmetro status. Included are responses pertaining to the use and frequency of the use of alcohol, marijuana, and tobacco. There is a group of variables focused on attitudes regarding substance use; how the individual thinks about substance use, as well as what they perceive their peers, and their parents feel about substance use. Included are also reports of other behavior such as getting into fights, stealing, selling illicit drugs, as well as involvement in a religious community, support groups, and participation in youth activities.

II. Background

Decision Trees

Used in this project are classification trees, which predicts a qualitative response, and regression trees, which predicts a quantitative response. Growing a tree uses a top-down, greedy approach or recursive binary splitting. It begins at the top of the tree, where all observations belong to a single region. Then it successively makes the best split in the predictor space based on some splitting criterion. Example of splitting criterion can be measures of homogeneity or purity of the resulting split, it can be a measure of deviance, or

classification error rate. For regression trees it can also be measures of residual squared sum, or RSS. This process will continue until some stopping criteria is met. This stopping criterion can be specified. Examples of stopping criteria include tree depth, a specified number of observations in final region, or when no more improvement is possible. Cross validation as well as pruning, which reduces model complexity, can be utilized to improve a decision tree. Using a decision tree fits our objective, because while they may not have the same level of predictive accuracy, they are easily interpretable. Our dataset includes many categorial variables and decision trees can handle them without creating dummy variables. To improve upon predictiveness and consequently better variable importance aggregating methods will also be used.

Random Forest

This method builds a specified number of decision trees using bootstrapped training samples. At each split only a random sample of *m* predictors are considered. The rationale behind this is that if there exists one very strong predictor, that predictor will show up in every decision tree, all the trees will be similar. By limiting the number of predictors, other predictors will have more of a chance. The resulting average trees are more reliable. The random forest method will be suitable because it tends to be useful with correlated predictors which are present in the dataset. The random forest model will produce an out-of-bag, OOB, error estimation, a measure of prediction error. When training the random forest model, it uses bootstrap sample of the original data, the samples that are not used are the “out-of-bag” samples. They are used to calculate the OOB error estimation.

Boosting

In the context of decision trees, boosting uses an aggregation of decision trees. It grows the trees sequentially, where information from the previous tree is used to improve the predictability of the decision tree. It is an approach that learns slowly. The current tree

is trained to correct errors of the previous trees using residuals. There are three tuning parameters. For this project the shrinking parameter will be utilized. It is a small positive number used to control the rate at which boosting learns. It determines the contribution of the current tree to the ensemble. The number of trees and the number of splits in each tree are also tuning parameters that can be used.

III. Methodology

Data Preparation

This research specifically focuses on individuals that are still in high school at the time of the survey. This dataset includes individuals in college and beyond. Those observations were removed. Other fields required recategorization values or were replaced as missing values. For example, fields representing frequency of substance use recorded the values "993" to indicate that the individual did not use in the past year; those values were replaced with zero to reflect no use. In those same fields, "991" indicated that the individual had never used. Those values were changed to missing value since a binary variable exists in the dataset that recorded whether the individual had ever used.

Models

Our data set was randomly split, using 80% of the original dataset as the training set and 20% as the test set. For variables pertaining frequency of a substance use, there is a continuous variable and a discretized version of the variable. For each model one type was removed to reduce redundancy.

A Decision tree was used to predict whether the individual had every used alcohol. Pruning was used to improve the model, and the discretized variables were omitted. A testing error was calculated to measure the model performance.

Random forest was used to create a multi-class classification model. Continuous variables were utilized in this model. Multiple *m* values were tested and the OOB estimate

of error rate was used to find the best model. A testing error was then calculated as a measure of model performance.

Boosting was used for regression tree model, predicting the number of days in the past year the individual used alcohol. Discretized variables were removed from the dataset for this model. Testing was done to find a shrinking value that produced the lowest mean squared error and k-fold method was used for cross validation.

IV. Results

Decision Tree

A decision tree was the first model constructed predicted a binary outcome, whether an individual will use alcohol. Figure 1 is the resulting decision tree. The structure of the tree before and after pruning remained unchanged.

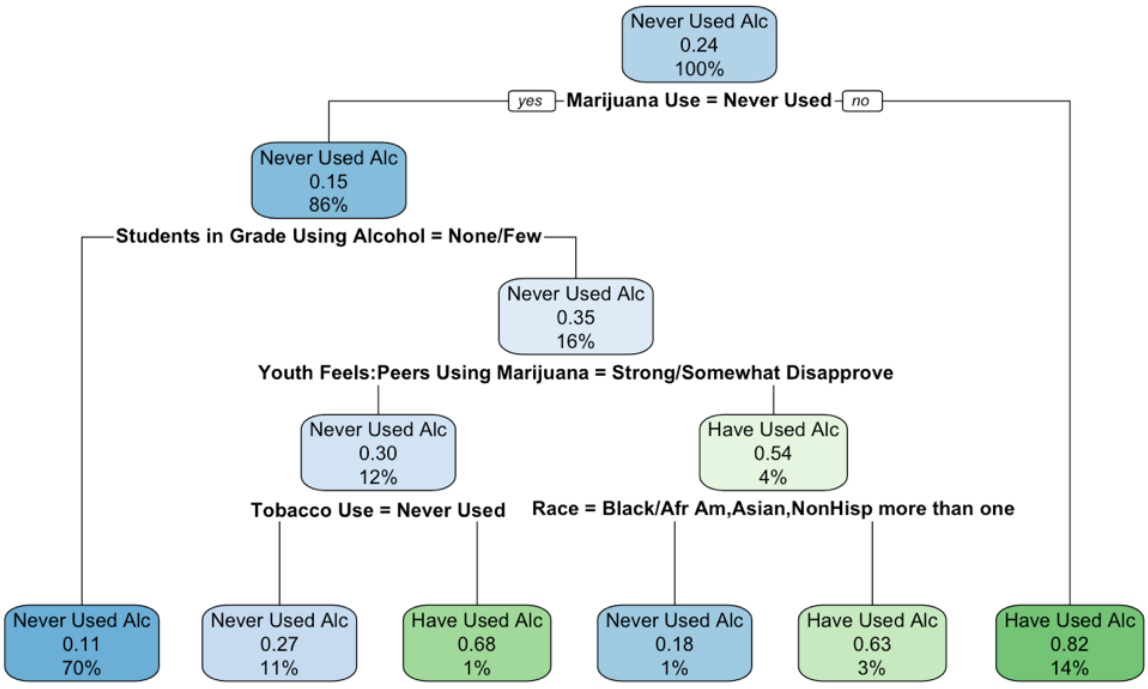

Figure 1 Decision Tree Plot: Use of Alcohol

To understand this tree, we will evaluate the path leading the second to the last terminal on the right resulting in having used alcohol. In the root node at the very top, the first number under text “Never Used Alc” represents the proportion of the observation belong in that category. Below that figure is the percentage of all observations in that region. Below the box is the first split variable which indicates whether the individual has ever used marijuana. By convention, the “yes” branch is on the left and the “no” branch is on the right. If the individual responded “yes”, the next node predicts that the proportion, .15, of those that responded “yes” to marijuana use will not have used alcohol, and these individuals will account for 86% of the entire population. The next split uses the variable how many students in the individual’s grade that were using alcohol. Following the right branch the next node predicts that the proportion, .35, of those that did not report that “None/Few” of their peers were using alcohol, will not have used alcohol, this group is 16% of the entire population. Continuing, if the individual did not respond that they strongly or somewhat disagreed with their peers’ using marijuana to the following node, if the individual is more than one race, then the proportion, .it is predicted that the individual will have used alcohol. This group of individuals represent 3% of the entire population. Note that the variable in the last split is not binary, the left branch represents the responses: Black/African American, Asian, or Non-Hispanic and the right branch represents the response more than one race. The terminal nodes combined is the entire population, percentages will add up to 100%.

Figure 2 shows the relative importance of the variables contributing to the model.

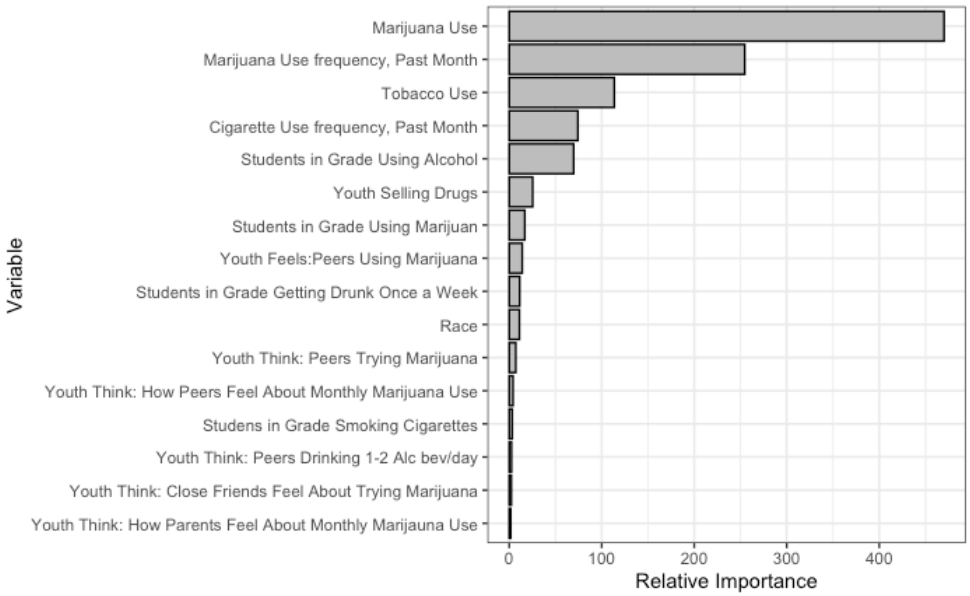

Figure 2 Decision Tree: Variable Importance

This figure does not show all 69 predictor variables, only the ones with highest relative importance. In this plot the relative importance with not scaled to 100 which is often standard for variable importance ratings. A confusion matrix was constructed and a test error of 14.9% was calculated.

	Never Have Used Alcohol	Have Used Alcohol
Never Have Used Alcohol	769	115
Have Used Alcohol	49	165

Figure 3 Decision Tree: Confusion Matrix

Random Forest

Random forest was used for a multiclass classification model. For this model, discretize variables produced greater prediction accuracy. Testing *m* values in the 1-66 range, 7 predictors were found to be the best *m* value. Using m = 7 predictors at each split, the model produced an OOB estimate of error rate of 18.27% and a test error of 21.77%. In

comparison, using continuous variables and the same *m* value, the training error of 61.43% was achieved and a testing error of 69.81va%. Figure 4 is the relative variable importance plots for the model with the discretized variables. The variables are ranked by average decrease in mean accuracy and mean decrease in Gini index. It is an average because each class produced different variable importance scores. All scores can be seen in the full variable importance table.

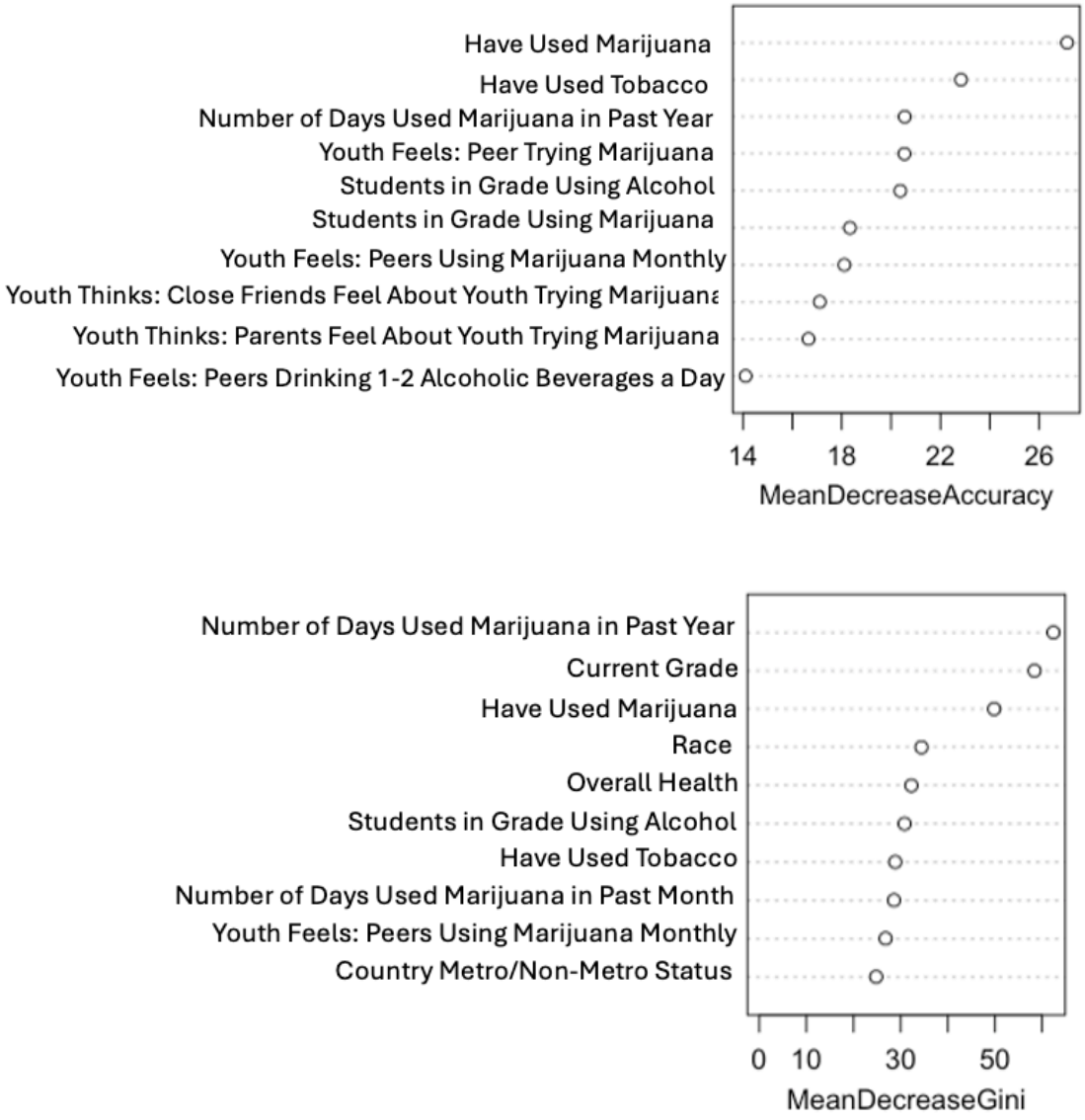

Figure 4 Random Forest: Variance Importance

Figure 5 is the confusion matrix with the individual class error rate.

	1-11 Days	12-49 Days	50-99 Days	100-299 Days	300-365 Days	No Use
1-11 Days	104	8	1	0	272	0.73
12-49 Days	56	18	0	0	69	.87
50-99 Days	20	7	0	0	20	1.0
100-299 Days	15	5	0	0	12	1.0
300-365 Days	50	2	0	0	2281	0.02

Figure 5 Random Forest: Confusion Table

Boosting

Using boosting on the dataset containing continuous variables, a series of values were tested to choose the shrinking value that produced the lowest means squared error, MSE. Figure 6 shows the shrinkage values tested and the correlated MSE. The shrinkage value used for this model is 0.4, 1,000 trees were made, and the interaction depth was. The result was a model where 66 predictors were considered and 65 had a non-zero influence with a mean squared error of 1545.145. Figure 7 is a plot of the top 10 most important variables.

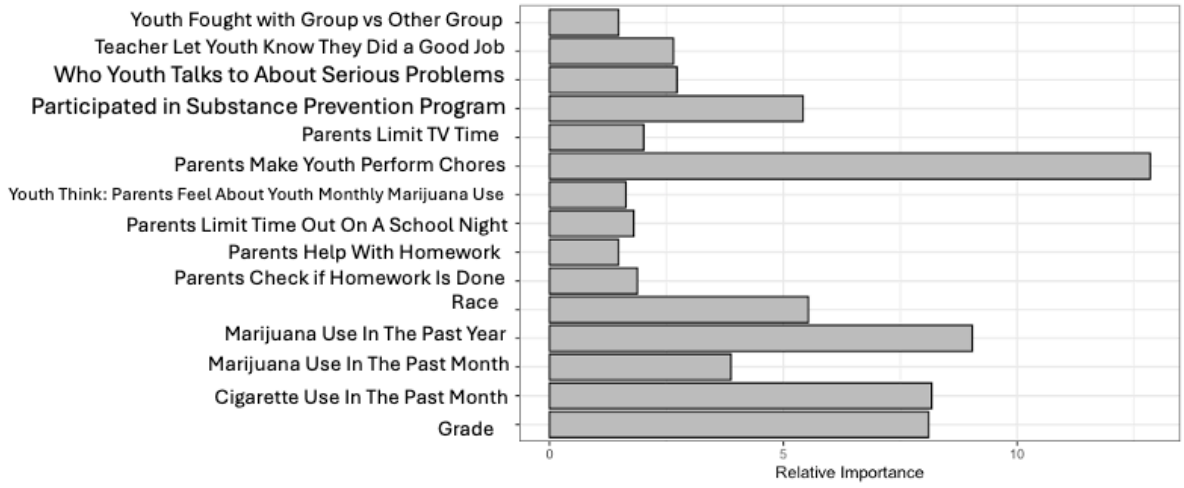

Figure 6 Boosted: Variable Importance

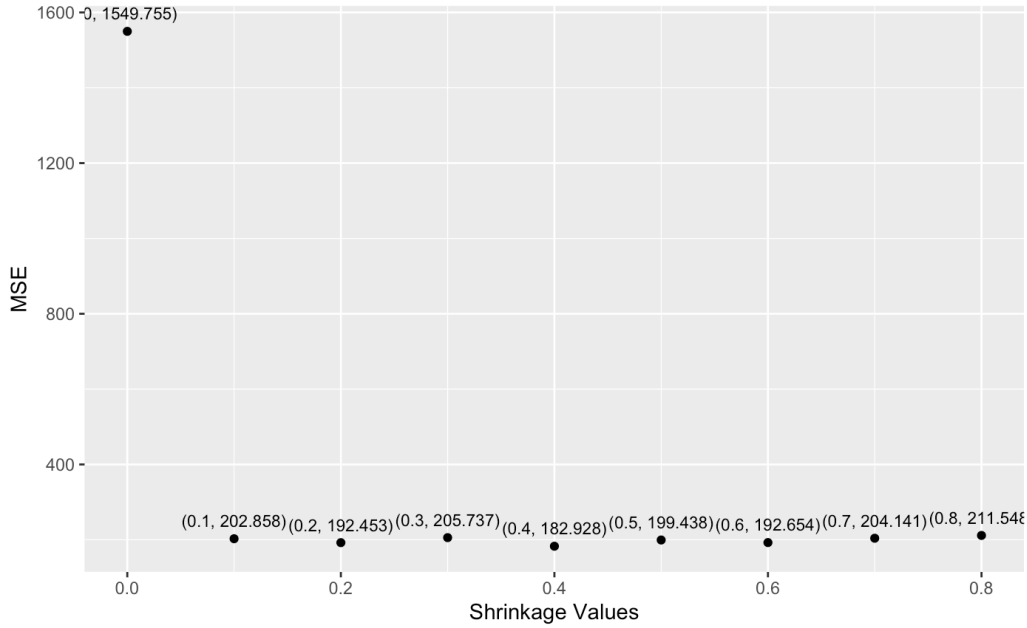

Figure 7 Boosting: Shrinkage Values

V. Discussion

This research shows that it is possible to utilize this survey to predict alcohol use among the youth with decent accuracy. Using these models' relative variable importance scores, we can see common influential predictors among the different models. Predictors associated with use of other substances appeared as important variables with high relative importance scores in all models. In particularly, the number of days in the past year an individual has used marijuana. Different variables associated with youth perception and feelings about substance use appeared in all models. In the random forest and boosted model, how the individual feels about their peers trying marijuana, and how they think their friends and parents feel about trying marijuana are high scoring variables. Less frequent among the most important variables were those associated with demographic information. Grade and race were an important determinant in both the random forest and boosted model. Parent involvement and youth activity only appeared as top predictors in the boosted model.

It is not completely clear why there is discrepancy in the testing error between the models using continuous and discretize variables. Seen in the two different random forest models is the largest change in model performance. While discretizing variables can lead to information loss, it improved predictive accuracy of the models. A boosted model was also tested using discretized variables, with 1,000 trees and a shrinkage value of .6. The result was an MSE of 2333.594, which makes the model using discretized variables significantly better. Some possible explanations may be discretized variables reduces overfitting, it makes the models more robust, bins also automatically handle outliers.

IIV. Conclusion

This research found some insight into the determinants of alcohol use among youth, further exploration would be useful in finding more conclusive evidence. Based on these results, to develop alcohol use prevention strategies, next steps may be to further explore using a response variable that includes the use of other substance, given how influential marijuana use is on alcohol use. The boosted model suggests that exploration into the youth participation in illegal activities may also be useful. It would seem from these results parent involvement as well as that individual perception about how peers and parents view substance are influential on an individual's decision to use alcohol. Strategies that support parents and guardians may be effective in mitigating underage drinking.

References

Underage Drinking, *Center for Disease Control and Prevention*. *https://www.cdc.gov/alcohol/fact-sheets/underage-drinking.htm*