Predicting the Factors Influencing Teenagers Substance Abuse Problem Using Decision Tree Model

Karthika Selvaraj

April 13th, 2023

Abstract

This study utilizes Decision tree algorithms to predict the consumption of marijuana in teenagers using the National Survey on Drug Use and Health (NSDUH) data. This research involves building a classification tree using a training dataset to predict if a teenager consumes marijuana. The variables in the training dataset are selected using the information gain function. The performance of the model is refined by using optimal decision tree size from cross validation and pruning. The paper focuses on the ensemble methods such as bagging and random forest for the models that overfit the data, while boosting is used for the models that underfit the data.

1 Introduction and Overview

The NSDUH series is the primary source of statistical data on alcohol, tobacco, drug use, mental health, and other health-related issues in the general population in the United States. Every year, this survey is done. NSDUH applies to the general civilian population in the United States aged 12 and up. The data retrieved for substance use includes lifetime, recent, and recent month usage, as well as the age at first use, substance use treatment history and perceived need for treatment, and substance use disorders. According to data from the Nation Center for Drug Abuse Statistics, 83.88% of minors used marijuana within a month. By developing machine learning (ML) models and evaluating their efficacy in performing binary, multi-class classification, and regression, this dataset is analyzed to determine the etiology of teen substance misuse. To enhance the model's performance, we also run other ensemble models here.

2 Theoretical Background

2.1 Decision Tree Model

A supervised machine learning method used for both classification and regression is the decision tree model. By beginning at the top of the tree with a single strongly correlated node and separating the predictor space into two branches deeper down the tree, we employ a top-down approach called recursive binary splitting. The decision tree employs the mean or the mode response value for the training observations in the region to which the observation belongs to produce a forecast for a given observation. The training dataset is overfitted due to the enormous amount of features in the dataset, which causes a complicated tree structure with high variation. By choosing the ideal tree size, either through pruning or by altering the depth value for the ensembled models, the model can be improved.

2.2 Pruning

In machine learning and search algorithms, pruning is a data compression approach that decreases the size of decision trees by deleting parts of the tree that are unnecessary and redundant for classifying occurrences. Pruning lowers the final classifier's complexity, which increases predicted accuracy by reducing overfitting. For the test dataset, performance is improved since the smaller tree with fewer splits may result in lower variance and better interpretation at the expense of some bias. A learning tree should be pruned to get smaller while maintaining its predicted accuracy as determined by a K-fold cross-validation set. The mistake rate for different size decision tree models can also be calculated using the Gini Index, which is a measure of total variance across the classes.

2.3 Bagging

Models with high volatility include decision trees. It implies that a small modification to the training set of data will result in a completely different model. Typically, decision trees overfit. Bagging with an ensemble technique can be utilized to get around this problem. To create a low-variance model for bagging, the bootstrap technique is used to take a repeated sample from the training dataset. The average prediction of all the samples is calculated. The test and training dataset subsets typically have a 7:3 ratio. Therefore, the bootstrap samples only contain about two-thirds of the training set's original data points. Therefore, the error is determined from the unused samples for that specific bootstrap sample for each tree constructed using bootstrap samples. Calculate the cumulative error average. This is known as Out Of Bag Error (OOB).

2.4 Random Forest

The bagging principle and Random Forest's performance are extremely similar. When compared to bagging with the subset of predictors per tree, this method offers modest benefits and builds many trees using the boot-strapped training dataset. In other words, when building a random forest, the algorithm is not even permitted to take into account the majority of the predictors that are accessible at each branch of the tree. Assume the data set comprises numerous predictors that are only moderately strong and one very powerful predictor. This potent predictor will be used in the top split by the majority of the bagged tree collection. All of the trees that were bagged will then be correlated. Averaging a large number of uncorrelated models reduces variance more than averaging a large number of highly linked variables. The variance across a single tree will be significantly reduced by the random forest. They make such efficient models because they can control overfitting without significantly increasing bias-related error.

2.5 Boosting

Another ensemble method for building a set of predictors is boosting. This method involves teaching learners sequentially, starting with early learners fitting basic models to the data and moving on to later learners checking the data for flaws. In other words, we fit successive trees (random sample) to resolve the aim of resolving the net error from the previous tree at each stage. An input's weight is increased when a hypothesis incorrectly classifies it, increasing the likelihood that the subsequent hypothesis will classify it properly. Weak learners can become better performers by merging the entire set at the end. The shrinkage parameter regulates how quickly the model picks up new information. Cross validation is used to choose the appropriate shrinkage value. A large number of trees are needed for improved performance for modest shrinkage values. However, because of the slow learning rate, overfitting of the model might result from using several trees.

3 Methodology

The 32,893 records in the NSDUH dataset contain about 2890 variables that represent 17 distinct forms of drugs, including alcohol, marijuana, cocaine, and others. The age at which drugs were first used, the quantity consumed, the mental state, the experiences of youth, the presence of adult and teenage depression, and many other variables are the datasets. This investigation aims to identify the fundamental factors causing juvenile marijuana usage, a major public health concern. Preprocessing the data entails screening the data that reflects the kids and changing the variables to the proper data type. Utilizing the information gain approach, the pre-processed data is utilized to rate the features according to how important they are in getting teenagers to use marijuana. The element that has the most weight in influencing marijuana use is alcohol use. 12 variables are chosen, the effects of alcohol are therefore excluded, and a basic tree model is trained on a training set and used to forecast the test set. The training set's accuracy was 91%, while the test set's accuracy was 89%. The dataset is divided into training and test sets with a ratio of 7:3. The study seeks to provide answers to the following questions following data pre-processing and feature selection.

- **Substance Variables -** tobflag, irmjfy, iralcrc, cigdays, alcflag
- **Youth Experience -** FRDMEVR2, uadfwho, stndalc, EDUSCHGRD2, PRLMTTV2, eduschlgo
- **Health Condition -** HEALTH2
- **Economical Status -** irwrkstat

3.1 Consumption of Marijuna

A Binary Classification tree is constructed to predict whether teenagers consume marijuana. After careful selection of variables using the information gain, other marijuana related variables are also removed so that this method can consider other factors. The variables that is used to predict consumption of marijuana are: consumption of tobacco(tobflag), how friends and family feel about teens consuming marijuana(FRDMEVR2) who provides under age drinkers more alcohol (uadfwho), how many of the students consume marijuana(stndsmj).

3.2 Risk of Using Marijuana

A Multi-class classification tree is built to identify the factors that influence the teens to consume marijuana once or twice a week. The response variable “risk of smoking marijuana” is classified into four classes such as No, Low, Moderate and High Risk. The features that help in prediction of risks are as follows, how peers feel about trying marijuana (YFLTMRJ2), how parents feel about using marijuana (prmjmo), frequency of marijuana consumed in a year(irmjfy) and how recently teens consumed alcohol (iralcrc).

3.3 Frequency of Marijuana Consumption

A Regression is performed to predict the frequency of marijuana consumed in a year by teens. The response variable falls under the range of 1 to 365, while 0 is considered as no history of marijuana. The variables that contribute to the frequency of consumed tobacco(tobflag), who provides under age drinkers more alcohol (uadfwho), how friends and family feel about teens consuming marijuana monthly(YFLTMRJ2,prmjmo), how many of the students consume marijuana (stndsmj), number of cigars used in a month(cigdays), health condition (HEALTH2).

4 Computational Results

4.1 Variable Selection Using Information Gain Function

The subset of variables from the NSDUH dataset is used in this analysis. The Feature selection for the model is performed by using the information gain function to list the variable’s importance related to marijuana consumption. The other marijuana related variables are highly correlated. This directly influences the output which is removed from the model to identify other variables that influence the consumption.

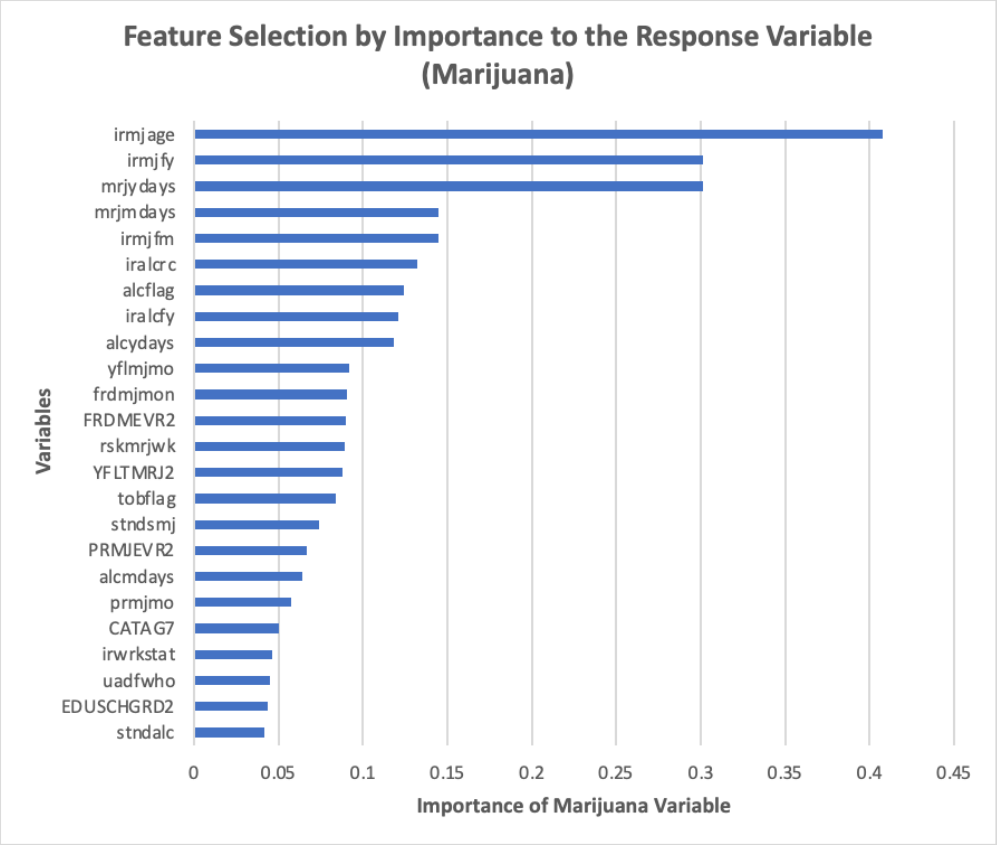


Fig 1: Variables listed by their importance with marijuana using Feature Selection

4.2 Classification of Marijuana Usage

The tree model classifies if a teenager has consumed marijuana or not. This is performed by identifying the highly correlated variables to check if marijuana has been consumed, which resulted in finding out if tobacco consumption is the primary factor apart from alcohol being a deciding factor. It is very evident that drugs such as alcohol, tobacco and marijuana are highly interdependent factors. The secondary factors are how the peers or the parents feel about using marijuana and how many students in their grade are using it. This shows that the teenagers cultivate these habits from the people they are socializing with.

- If a person is not consuming tobacco and their friends are against using marijuana, they are not prone to marijuana consumption.
- If a person consumes tobacco and their friends are against using marijuana, but their parents somewhat disapprove, they still consume marijuana.

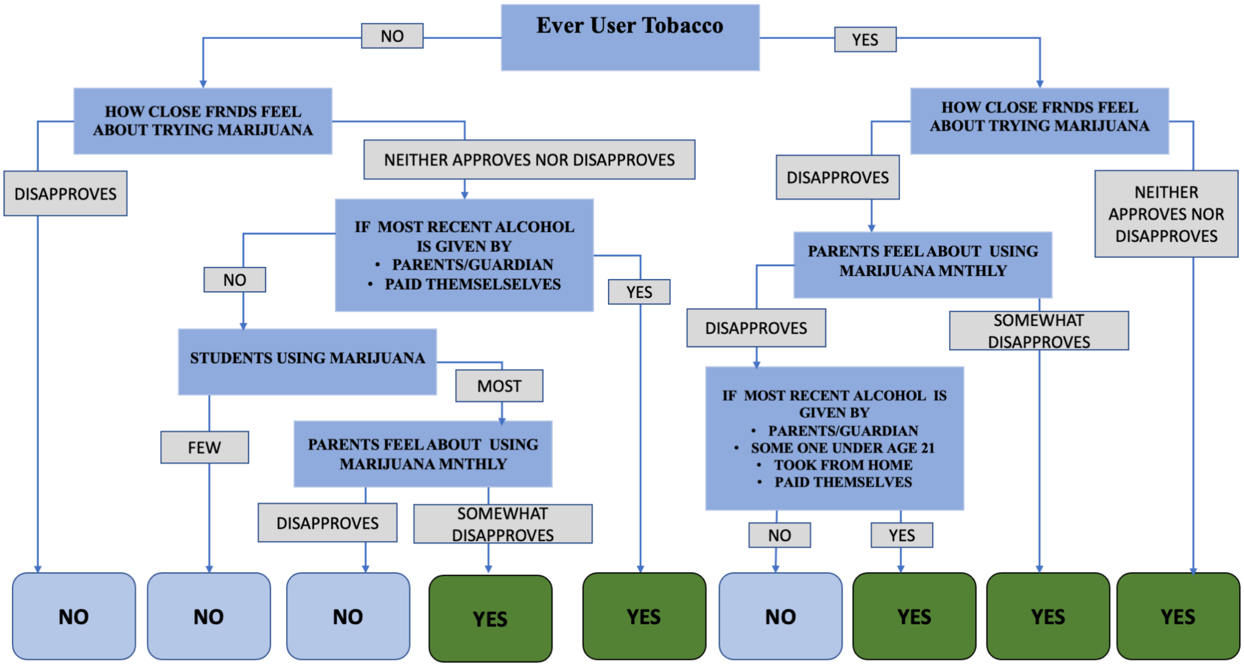

Ever User Tobacco
NO
YES
HOW CLOSE FRNDS FEEL ABOUT TRYING MARIJUANA
HOW CLOSE FRNDS FEEL ABOUT TRYING MARIJUANA
DISAPPROVES
NEITHER APPROVES NOR DISAPPROVES
DISAPPROVES
NEITHER APPROVES NOR DISAPPROVES
IF MOST RECENT ALCOHOL IS GIVEN BY
• PARENTS/GUARDIAN
• PAID THEMSELSELVES
NO
YES
PARENTS FEEL ABOUT USING MARIJUANA MNTHLY
DISAPPROVES
SOMEWHAT DISAPPROVES
STUDENTS USING MARIJUANA
MOST
FEW
PARENTS FEEL ABOUT USING MARIJUANA MNTHLY
DISAPPROVES
SOMEWHAT DISAPPROVES
IF MOST RECENT ALCOHOL IS GIVEN BY
• PARENTS/GUARDIAN
• SOME ONE UNDER AGE 21
• TOOK FROM HOME
• PAID THEMSELVES
NO
YES
NO
NO
NO
YES
YES
NO
YES
YES
YES

Fig 1: Decision Tree - Consumption of marijuana in teenagers

This model was initially trained with the training dataset and had a performance rate of 91% for the training and 90% for the test set. The difference occurring in the accuracy rate of training and test dataset is due to the overfitting of the model using training data. Since the model has already produced a good performance rate using the ensemble method, a test is performed to check if this method can improve the model further. The bagging and random forest method has not shown any significant improvement in the model. The boosting method has increased the accuracy rate to 91%. Overall the accuracy of this model is good even without using the ensemble methods. From the graph below the boosting method performs better with the use of shrinkage parameter(λ) as 0.05 rather than using bagging and random forest method.

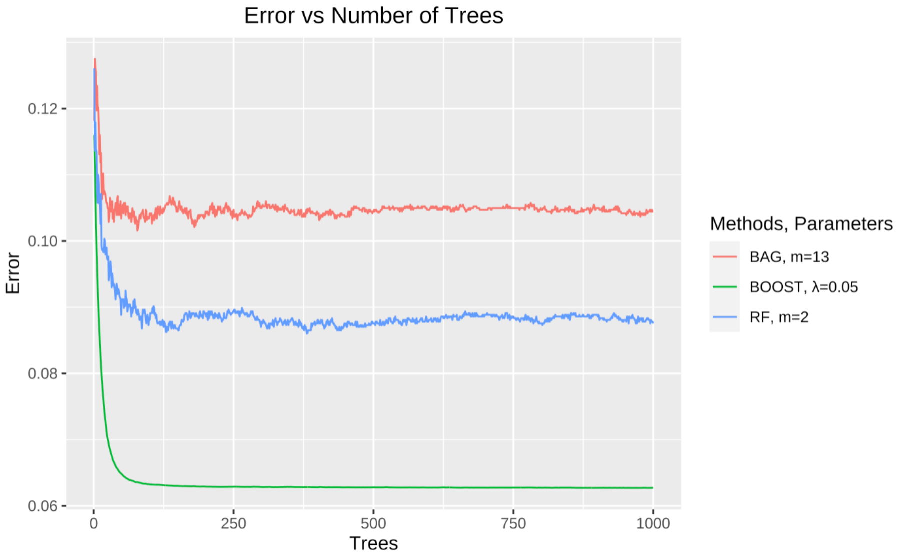
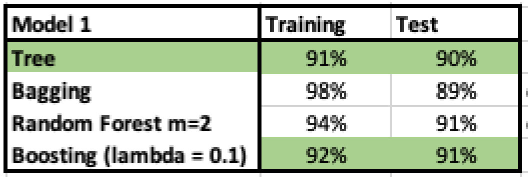


Fig 2: Performance of the model by different methods on training data



Model 1	Training	Test
Tree	91%	90%
Bagging	98%	89%
Random Forest m=2	94%	91%
Boosting (lambda = 0.1)	92%	91%

Fig 3: Table - model accuracy on test and training data

4.3 Classifying the Risk of teenagers consuming marijuana

A multi classification is performed to analyze the risk of the teenagers consuming marijuana once or twice in a week. The risk rate is classified into no, low, moderate and high variables. The variables that determine the risk are primarily by how the peers and parents feel about using marijuana, the frequency of marijuana and alcohol consumption.

- If the teenagers, their parents and their peers neither approve nor disapprove about using marijuana and the teenagers have the habit of alcohol consumption, then they are at high risk of consuming marijuana at least once a week.

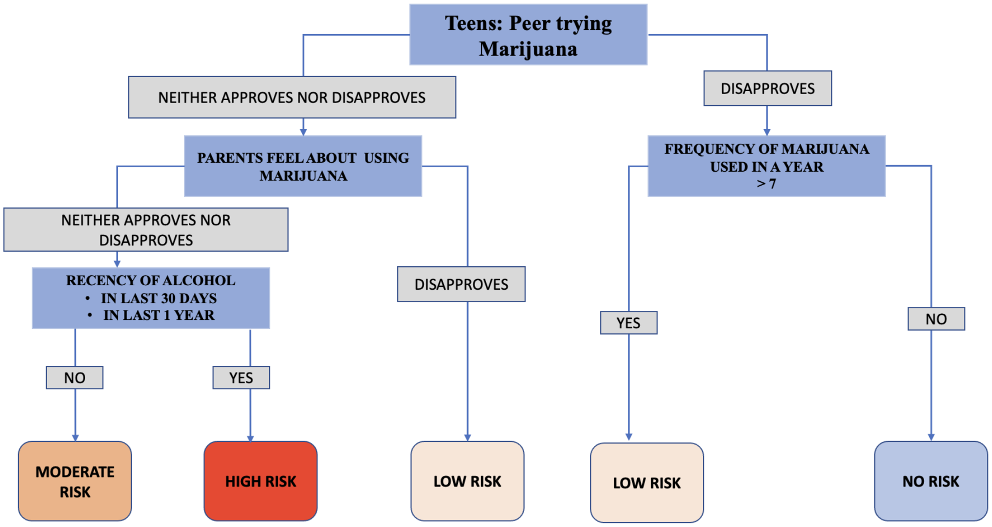

Teens: Peer trying Marijuana
NEITHER APPROVES NOR DISAPPROVES
DISAPPROVES
PARENTS FEEL ABOUT USING MARIJUANA
FREQUENCY OF MARIJUANA USED IN A YEAR > 7
NEITHER APPROVES NOR DISAPPROVES
RECENCY OF ALCOHOL
• IN LAST 30 DAYS
• IN LAST 1 YEAR
DISAPPROVES
YES
NO
NO
YES
MODERATE RISK
HIGH RISK
LOW RISK
LOW RISK
NO RISK

Fig 4: Decision Tree - Risk of teenagers consuming marijuana

Initial accuracy rate for this model using the test dataset was 46 %. The performance of the bagging model including all the predictors and the random forest techniques with the subset of the predictors are plotted by error rate. The number of trees to figure out how the model has improved with the increase in the number of trees are also included in the performance of the bagging model. The graph shows that the random subset of two predictors used in the model has developed well with the accuracy rate of 47% for the test data, and performs better when compared to the bagging model (15 predictors) with the test data accuracy of 30%. The boosting method is only for the under-fitting model. From the graph it is clear that the boosting only enhances the error rate for this model even though the optimal learning rate also known as shrinkage parameter (0.4) is identified through cross-validation for the larger number of trees. The boosting model performance rate was 39%

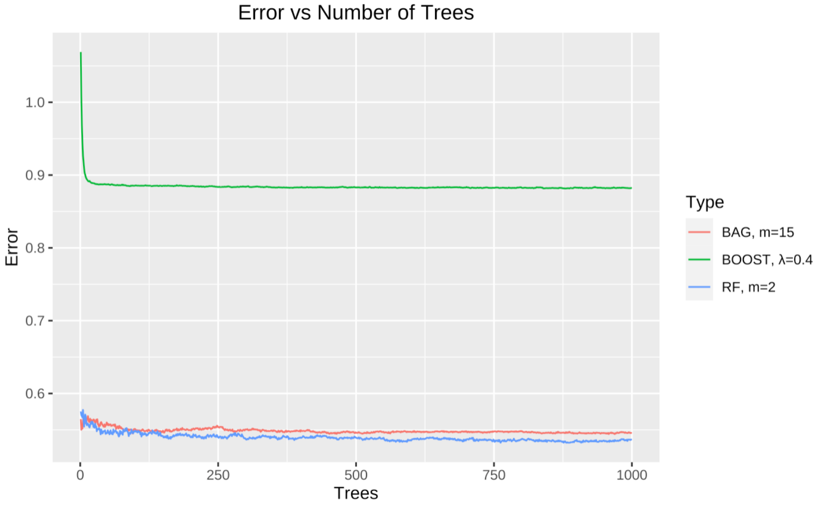
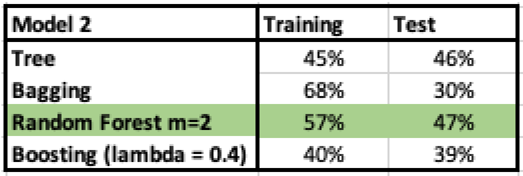


Fig 5: Performance of the model by different methods on training data



Model 2	Training	Test
Tree	45%	46%
Bagging	68%	30%
Random Forest m=2	57%	47%
Boosting (lambda = 0.4)	40%	39%

Fig 6: Table - model accuracy on test and training data

4.4 Prediction of Frequency in Marijuana Consumption

A Regression model is used to find how frequently teenagers consume marijuana in a year. The frequency ranges from 1 to 365. The important variables used to identify are the combination of variables from the model 1 and 2. The primary factors considered here are, whether they consume tobacco, how parents feel about them using marijuana monthly, how many students of their grade use marijuana, number of cigars used in a month, who recently gave them alcohol.

- If a teenager used tobacco and their parents somewhat disapprove of using it, most of the students in their grade are prone to using marijuana, and they result in using more than six cigarettes in a month. They tend to use marijuana 223 times in a year.
- If a teenager does not use tobacco and also does not get access to alcohol at home, this results in teenagers using marijuana twice a year.

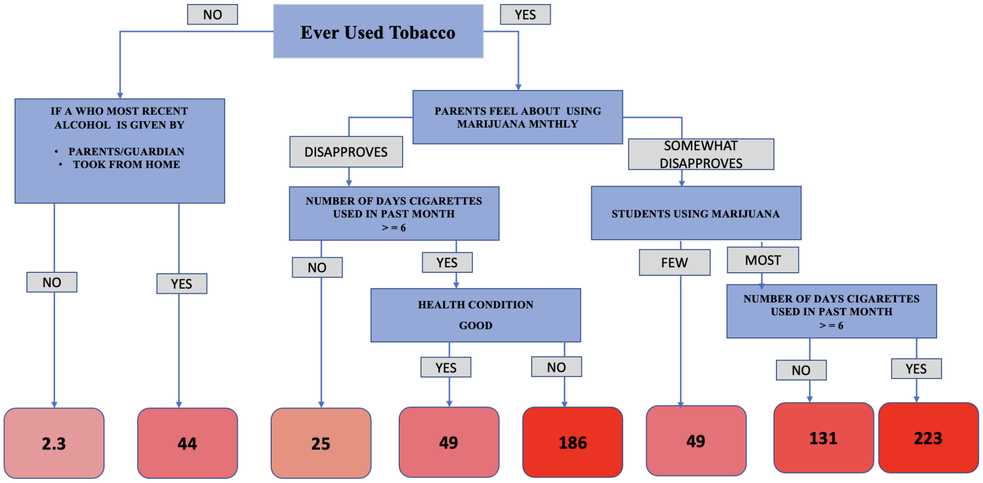

NO
Ever Used Tobacco
YES
IF A WHO MOST RECENT ALCOHOL IS GIVEN BY
• PARENTS/GUARDIAN
• TOOK FROM HOME
PARENTS FEEL ABOUT USING MARIJUANA MNTHLY
DISAPPROVES
SOMEWHAT DISAPPROVES
NUMBER OF DAYS CIGARETTES USED IN PAST MONTH >= 6
STUDENTS USING MARIJUANA
NO
YES
NO
YES
FEW
MOST
HEALTH CONDITION GOOD
NUMBER OF DAYS CIGARETTES USED IN PAST MONTH >= 6
YES
NO
NO
YES
2.3
44
25
49
186
49
131
223

Fig 7: Decision Tree - Frequency in Marijuana Consumption

The initial training MSE of the model was 1263 and the test MSE was 1430. The difference in the MSE clearly shows that the model overfits the data. This overfitting can be avoided using the random forest method by selecting the optimal subset value of the variables as two which was obtained from the cross-validation technique. The performance rate of the test data has improved to 1282 and the training data has 822 Mean square error rate. Despite all these methods overfitting still occurs. The boosting method is used to overcome the model overfitting by using the best learning rate which is 0.05 obtained from the cross-validation technique. Thus the training MSE for the bagging method is 1363 and for the test MSE is 1376. The error rate for the test and training set is very close thus the model's performance result was good without any overfitting or underfitting of the training data.

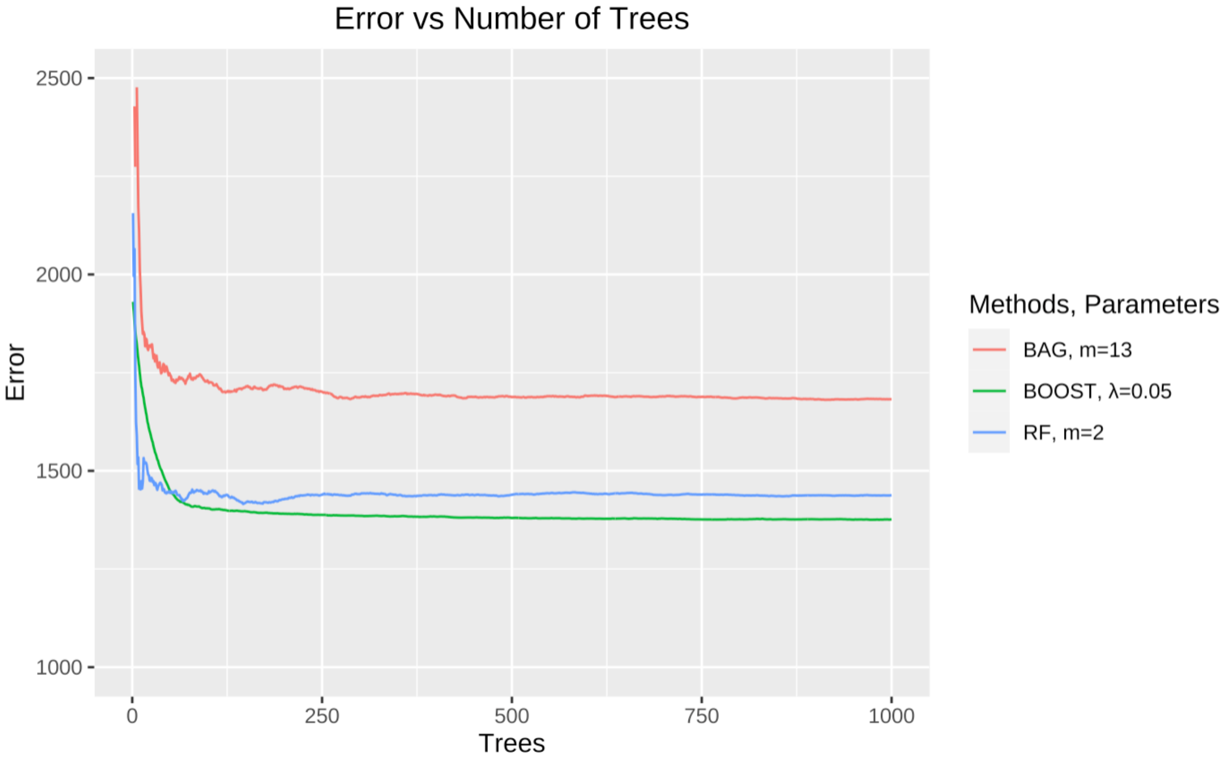
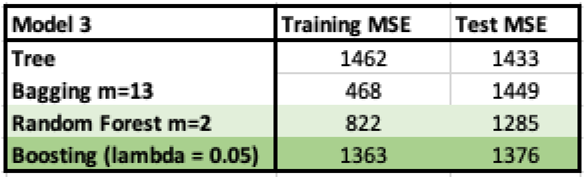


Fig 8: Performance of the model by different methods on test data



Model 3	Training MSE	Test MSE
Tree	1462	1433
Bagging m=13	468	1449
Random Forest m=2	822	1285
Boosting (lambda = 0.05)	1363	1376

Fig 6: Table - model accuracy on test and training data

Conclusion

This study used a Decision tree algorithm to significantly classify marijuana consumption by teenagers in the United States using the National Survey of Drugs Use and Health dataset. Among many variables available, the predictor variables were handpicked using feature selection. The primary factors that were responsible for marijuana consumption were the other substance usage such as alcohol and tobacco. This is because the use of one substance leads to the other. Here the study used a binary classification model to classify if a teenager used marijuana with a performance rate of 91%. The multi-class classification model using random forest was used to predict the risk of consuming marijuana more than once a week with the accuracy rate of 47%. Finally a regression model combined with the use of a boosting method was used to predict the frequency of marijuana usage in a year. This model’s training and test error were similar thus making the model a better fit. A thorough investigation of the NSDUH dataset could definitely aid in efficient use of the data in order to get better outcomes. Future research based on various machine learning techniques can be used to analyze if it could perform better when using same predictors.

References

(n.d.). *National Survey on Drug Use and Health (NSDUH)*. SAMHDA. https://www.datafiles.samhsa.gov/dataset/national-survey-drug-use-and-health-2020-nsduh-2020-ds0001

(n.d.). *Drug Use Among Youth: Facts & Statistics*. NCDAS. https://drugabusestatistics.org/teen-drug-use/

Alam, M. (2022, April 25). *Feature selection: A comprehensive list of strategies*. Towards Data Science. https://towardsdatascience.com/feature-selection-a-comprehensive-list-of-strategies-3fecdf802b79

A. T. (2020, February 17). *How is information gain calculated?* R Bloggers. https://www.r-bloggers.com/2020/02/how-is-information-gain-calculated/

James, G., Witten, D., Hastie, T., & Tibshirani, R. (n.d.).An Introduction to Statistical Learning. Retrieved April 10, 2023, from https://www.statlearning.com/