**Investigating Factors Correlated with Youth Drug Use**

**Abstract**

This report presents a study that aimed to investigate the utility of decision tree models for regression and classification in predicting youth drug use using data from the National Survey on Drug Use and Health. The dataset focused on individuals below 18 years of age and included 15 carefully chosen variables. The analysis revealed that the Random Forest algorithm was not effective in predicting monthly cigarette smoking frequency. However, the classification models yielded promising results, with the binary classification model for marijuana use achieving high accuracy on the test data, and the multiclass classification model for alcohol use also achieving high accuracy. The study identified significant variables for the response variables and found that other drug-related variables, such as start age of alcohol consumption and monthly smoking frequency, were the best indicators for the models, while demographic and socioeconomic factors were not strongly correlated indicators. Overall, the findings suggest that the combination of variables and modeling approach used in this study was effective for classification but not suitable for regression in predicting youth smoking behavior.

**Introduction**

The National Survey on Drug Use and Health (NSDUH) is a comprehensive database that provides detailed information on substance use and mental health in the United States, making it a valuable resource for studying youth smoking behavior. The dataset used in this study was obtained from the NSDUH and included responses from individuals under 18 years old.

The NSDUH is widely recognized as the leading source of data on alcohol, tobacco, and drug use among the general population, covering individuals aged 12 and older. It includes information on various aspects of substance use, such as lifetime, past-year, and past-month use, treatment history, and perceived need for treatment. The dataset also captures demographic and socio-economic factors, as well as youth experiences that may be associated with smoking behavior among young people.

This report aims to investigate the correlation between youth drug use and various factors using decision tree models applied to survey data from the National Survey on Drug Use and Health (NSDUH). The survey covers a wide range of demographics and includes both household and non-institutional group quarters populations. The data for this report is sourced from the NSDUH 2020 dataset, which includes detailed information on respondents' demographics, youth experiences, and drug use behaviors.

The report employs various decision tree modeling techniques, including binary classification for identifying factors associated with cigarette use, multiclass classification for distinguishing between infrequent, occasional, and frequent marijuana use, and regression for predicting the number of days per year a person has used alcohol. Ensemble methods, such as random forests, are also used to build predictive models for youth drug use. The discussion will include a thorough analysis of the variables and their correlations with drug use, as well as the effectiveness of the decision tree models in predicting drug use behaviors among youth.

**Background**

**Decision Trees:**

Decision trees are a popular machine learning technique for performing both classification and regression tasks. However, decision trees can suffer from overfitting, which reduces their performance on new data. To improve decision tree performance, ensemble methods such as bagging, boosting, and random forests are often used.

A decision tree is a hierarchical tree-like structure that partitions data using binary decisions based on features. Internal nodes represent features, and branches represent the possible values of those features. Leaves represent class labels or regression values.

**Bagging**:

Bagging is an ensemble method that involves training multiple decision trees on random subsets of data with replacement and combining their predictions. The technique helps to reduce model variance and improve generalization performance. Bagging is particularly effective for reducing overfitting and enhancing decision tree performance in classification tasks.

**Boosting**:

Boosting is an iterative technique that adjusts weights of training samples to prioritize misclassified samples. Boosting creates a sequence of decision trees that are combined to make the final prediction. This approach focuses on misclassified samples and produces a strong ensemble of decision trees that improve model accuracy.

**Random Forests: An Ensemble Approach to Decision Trees**

Random forests are an ensemble method used in combination with decision trees to enhance their predictive capabilities. Multiple decision trees are constructed, and their predictions aggregated to improve accuracy and model robustness. This ensemble approach helps to reduce overfitting and enhance decision tree performance in classification tasks. Random forests provide an effective solution for building accurate and robust classification models using decision trees in machine learning. Figure 1 illustrates a decision tree with a classification problem.

![Diagram

Description automatically generated]()

Fig 1: Classification Tree Example ([xoriant.com](https://www.xoriant.com/blog/decision-trees-for-classification-a-machine-learning-algorithm))

**PRUNING: Simplifying Decision Trees**

Pruning is a technique that simplifies decision trees and prevents overfitting. This method involves removing unnecessary nodes or branches from the tree to create a more interpretable and compact structure. Two common pruning techniques are pre-pruning and post-pruning. Pre-pruning limits the tree's growth by stopping it from growing beyond a certain depth or number of samples, while post-pruning involves growing the tree to its fullest potential and then removing unnecessary branches based on a pruning criterion. By reducing the complexity of decision trees, pruning is an effective way to improve their generalization performance. [xoriant.com](file:///Users/hanishak/Desktop/Hanisha/MSDS/spring/ML2/git/5.%09https:/www.xoriant.com/blog/decision-trees-for-classification-a-machine-learning-algorithm)

**Binary and Multiclass Regression in Regression Analysis**

Regression analysis is a statistical method that is widely used for modeling the relationship between a dependent variable and independent variables that may affect the outcome. In binary regression, the dependent variable takes on binary values, and the objective is to predict the probability of an event occurring. On the other hand, multiclass regression is used when the dependent variable has multiple categories. Linear regression is a commonly used technique in both binary and multiclass regression. This method fits a linear relationship between the dependent and independent variables.

**Methodology**

**Data Cleaning and Variables Selection:**

Before conducting any analysis on a dataset, it is important to ensure that the data is clean and free of errors. The dataset used in this report was carefully prepared for analysis by addressing various issues such as redundancy, inconsistency, and missing values. The first step in this process is to replace out-of-range values with appropriate values for several variables. This step ensured that the data was accurate and that any potential outliers that could distort the results were eliminated.

Next, categorical variables were recoded using the recode factor () function to create more readable labels. For instance, the father\_Presence and Mother\_Presence variables were recoded into "Yes," "No," and "NA" categories to facilitate the interpretation of the data. Similarly, other variables such as Alcohol\_Monthly\_Class, ever\_Used\_Marijuana, and Income were recoded into more meaningful categories.

To further refine the dataset, rows containing NAs in father\_Presence and Mother\_Presence variables were filtered out since these variables were crucial for the analysis. Additionally, any remaining rows containing NAs were removed to ensure that the final dataset was free of any missing data.

Finally, all unused columns were dropped from the data frame, resulting in a new data frame with 15 variables of interest that were then renamed into more readable labels. The selected variables included Alcohol\_Monthly\_Class, ever\_Used\_Marijuana, Monthly\_Smoking\_Frequency, Smokeless\_Tobacco\_Start\_Age, Alcohol\_Start\_Age, Marijuana\_Start\_Age, Monthly\_Marijuana\_Frequency, Monthly\_Alcohol\_Frequency, Race, Father\_Presence, Mother\_Presence, Monthly\_Alcohol\_Frequency, Race, Sex, General\_Health\_Rank, Income\_Group, and Metro\_Size.

To focus the analysis on youth drug use, responses from respondents under 18 years old were filtered, and out-of-range responses were replaced with appropriate values. Additionally, responses indicating uncertainty or older age for variables capturing father and mother presence were replaced with NAs to ensure data integrity.

**PREDICTIVE MODELS**

In this study, predictive models were developed for substance use behaviors using regression and classification methods. The data was split into training and testing datasets to evaluate the performance of the models.

To predict whether an individual has ever used marijuana, a binary classification tree model was constructed. The resulting model consisted of only three nodes and was not pruned during the determination of both the training and test mean squared errors (MSEs). Afterward, the important variables were identified for further analysis. The model was evaluated using a confusion matrix, and a plot was created to compare the predicted versus actual values of ever used marijuana.

To predict the MonthlySmokingFrequency for multiclass classification, a Random Forest model was trained using both the out-of-bag technique and normal random forest models. The model was evaluated using its summary, and relative influence was extracted to assess the importance of each predictor variable. Importance variables were determined, with only the age when first smoked being related to the analysis. No variables were dropped during the analysis.

For regression, a bagging model was built using the random Forest () function to predict monthly smoking frequency. The model was evaluated using mean squared error (MSE) and a plot was created to compare the predicted versus actual values of monthly smoking frequency. Variable importance was assessed using the importance () function and further analyzed using the "%IncMSE" and "IncNodePurity" metrics.

Parameter tuning was performed for each model to optimize its performance. The goal of this study was to develop accurate predictive models for substance use behaviors using a variety of techniques, and the results were evaluated using error metrics and variable importance analysis.

**RESULTS**

**REGRESSION TREE MODEL (Monthly Smoking Frequency)**

The Regression Tree model was utilized to predict monthly cigarette smoking frequency, employing bagging and random forest approaches. The importance of predictor variables in the model was determined using %IncMSE (Percentage Increase in Mean Squared Error) and IncNodePurity (Increase in Node Purity) values. Notably, variables such as "alcohol\_monthly\_class", "smokeless\_tobacco\_start\_age", "alcohol\_start\_age", "monthly\_marijuana\_frequency", "general\_health\_rank", and "income\_group" were identified as influential due to their higher %IncMSE and IncNodePurity values.

A picture containing text, screenshot, number

Description automatically generated

Fig 2: Importance Variables Fig 3: Variable Importance plot

The accuracy of the model in predicting monthly cigarette smoking frequency was assessed using Mean Squared Error (MSE), which yielded a value of 2.287. This indicates relatively good accuracy of the Regression Tree model with the selected predictor variables.

Chart

Description automatically generated

Fig 4: Out-Of-Bag Error Plot

Upon examining the Regression Tree model with around 100 trees, it becomes apparent that the out-of-bag (OOB) error rate stabilizes. This implies that augmenting the model with additional trees does not yield substantial enhancements in predicting monthly cigarette smoking frequency. Hence, utilizing 100 or more trees can strike a favorable balance between model complexity and accuracy in this context. The results of the Regression Tree model highlight the significant impact of variables such as "alcohol\_monthly\_class", "smokeless\_tobacco\_start\_age", "alcohol\_start\_age", "monthly\_marijuana\_frequency", "general\_health\_rank", and "income\_group" on predicting monthly cigarette smoking frequency. Further exploration of these variables can provide insights into the factors influencing smoking behavior. However, it is important to interpret these results in the context of the specific data and study population, and consider other factors such as model assumptions, sample size, and potential confounding variables when interpreting the findings.

**BINARY CLASSIFICATION (Marijuana drug use)**

The binary classification model for predicting marijuana drug use was trained on the train dataset, using all variables except for "marijuana\_start\_age" and "monthly\_marijuana\_frequency" to predict the target variable "ever\_used\_marijuana". The resulting model was then applied to the test dataset, and the predicted values of "ever\_used\_marijuana" were stored in the drug use prediction object. A contingency table was generated using the "table" function, which shows the counts of predicted values compared to the actual values of "ever\_used\_marijuana". The table indicates that 1029 instances were correctly predicted as "NeverUsed”, and 79 instances were correctly predicted as "EverUsed". However, there were 119 instances predicted as "NeverUsed" but labeled as "EverUsed", and 44 instances predicted as "EverUsed" but labeled as "NeverUsed".

A picture containing text

Description automatically generated

Fig 5: Binary Classification for Marijuana Drug Use

Based on the size of the resulting tree, it was determined that model pruning was not needed as the tree was not overly complex and did not require further simplification.

The results suggest that in predicting marijuana drug use using the selected predictor variables, such as "AlcoholStartAge" and "monthly alcohol frequency". However, it is important to conduct further analysis and interpretation of the results, considering other factors such as model assumptions, sample size, and potential confounding variables, to ensure a robust and reliable assessment of the model's performance in predicting marijuana drug use.

**MULTI-CLASS CLASSIFICATION (Alcohol drinking case)**

In the context of multinomial classification for alcohol drinking behavior, the results of the Generalized Boosted Regression Model (GBM) revealed several important predictors. The most influential predictor was "alcohol\_start\_age," accounting for 33.80% of the relative importance, indicating that the age at which an individual starts consuming alcohol significantly impacts their drinking behavior. Another significant predictor was "monthly\_marijuana\_frequency" with a relative importance of 17.31%, suggesting that the frequency of marijuana use on a monthly basis also plays a role in predicting alcohol drinking behavior.

Furthermore, other predictors such as "marijuana\_start\_age" (8.06%), "general\_health\_rank" (8.03%), "race" (7.54%), "income\_group" (5.99%), "metro\_size" (5.64%), "smokeless\_tobacco\_start\_age" (4.76%), "monthly\_smoking\_frequency" (3.76%), "sex" (3.03%), "father\_presence" (1.39%), "ever\_used\_marijuana" (0.35%), and "mother\_presence" (0.33%) also showed significant contributions to the model's predictive power.

Chart

Description automatically generated

Fig 6: Multi class Classification result for Alcohol drinking Case

It is important to note that these relative importance values reflect the contribution of each predictor towards explaining the variance in the outcome variable of alcohol drinking behavior in the context of multinomial classification. These findings highlight the significance of various demographic and behavioral factors, such as age of alcohol initiation, frequency of marijuana use, general health rank, race, income group, metro size, and other related variables, in predicting alcohol consumption patterns.

**DISCUSSION**

The study analyzed different predictive models to identify influential predictors of smoking behavior, marijuana drug use, and alcohol drinking behavior. The Regression Tree model for monthly cigarette smoking frequency achieved good accuracy with an MSE value of 2.287 and identified six influential predictors, while the binary classification model for marijuana drug use showed a lower accuracy, with 119 instances predicted as "Never Used" and 44 instances predicted as "Ever Used”. The multi-class classification model for alcohol drinking behavior revealed several important predictors, with "alcohol\_start\_age" and "monthly\_marijuana\_frequency" being the most influential predictors.

Comparing the results of the three models, the predictors of smoking behavior and marijuana drug use are different from those of alcohol drinking behavior. "Monthly\_marijuana\_frequency" was a significant predictor in both the Regression Tree model for smoking behavior and the binary classification model for marijuana drug use, but not as important in the multi-class model for alcohol drinking behavior. However, "alcohol\_start\_age" was a crucial predictor in all three models, indicating its significance in predicting drinking behavior. The study underscores the value of predictive models in identifying important predictors of smoking, drug use, and drinking behavior, which can inform interventions and prevention strategies.

**CONCLUSION**

In conclusion, various methods and models were utilized to analyze smoking behavior, marijuana drug use, and alcohol consumption. Regression analysis was used to identify predictors of smoking behavior, binary classification was employed to predict marijuana drug use, and multi-class classification was utilized to understand different levels of alcohol consumption. The regression analysis for smoking behavior identified significant demographic and behavioral factors that influence smoking behavior. The binary classification for marijuana drug use revealed important predictors that can be utilized for predicting marijuana use. The multi-class classification for alcohol consumption highlighted the significance of various factors, including age of alcohol initiation, frequency of marijuana use, general health rank, race, income group, metro size, and related variables, in predicting different levels of alcohol consumption.

Overall, the use of diverse methods and models in analyzing smoking behavior, marijuana drug use, and alcohol consumption has provided valuable insights that can guide efforts to address these behaviors and promote healthier behaviors in the population.

**CITATION**

1. NSDUH (2020). National Survey on Drug Use and Health, 2020. Retrieved from:

<https://www.datafiles.samhsa.gov/sites/default/files/field-uploads-protected/studies/NSDUH>-

2020/NSDUH-2020-datasets/NSDUH-2020-DS0001/NSDUH-2020-DS0001-info/NSDUH-2020-

DS0001-info-codebook.pdf

1. SAMHSA (2020). National Survey on Drug Use and Health, 2020. Retrieved from:

<https://www.datafiles.samhsa.gov/dataset/national-survey-drug-use-and-health-2020-nsduh-2020>-

ds0001

1. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data

Mining, Inference, and Prediction (2nd Edition). Retrieved from:

<https://hastie.su.domains/ISLR2/ISLRv2_website.pdf>

1. AWS (Amazon Web Services). (n.d.). What is boosting? Retrieved from

<https://aws.amazon.com/what-is/boosting/>

1. <https://www.xoriant.com/blog/decision-trees-for-classification-a-machine-learning-algorithm>

1. <https://scikit-learn.org/stable/modules/ensemble.html>