Siva Sushmitha Meduri
Homework-1

Exploring Youth Drug Use Patterns: A Decision Tree Analysis

Abstract

This report investigates youth drug (cigarette, alcohol, marijuana)use patterns using decision tree analysis applied to survey data of 2020 year from the National Survey on Drug Use and Health(NSDUH). The study employs pre-processed data encompassing demographics, youth experiences, and drug use behaviors. The decision tree model which predicts the "Cigarette Frequency of past month" produced a Mean Squared Error around 2.67(random forest) and 3.44(decision trees) . The model accuracy when classifying the usage of "Alcohol" was 79.97% (using decision trees) and improved to 81.26% (using random forest). Similarly, the model accuracy when classifying the "Marijuana Frequency usage" of 86.4% (decision trees, random forest) is achieved.

Introduction

Drug use among youth remains a significant public health concern, with profound implications for individual well-being and societal welfare. The National Survey on Drug Use and Health (NSDUH) is a pivotal data source providing insights into substance use behaviors among diverse populations in the United States. The 2020 NSDUH survey captures a wide array of information, including demographics, youth experiences, and detailed drug use patterns.

The prevalence of drug use among youth is influenced by various factors, including social, economic, and environmental determinants. Understanding these complex dynamics is crucial for developing effective prevention and intervention strategies. Decision tree analysis offers a powerful approach to unravelling the underlying patterns of drug use among youth, leveraging the rich information contained within the NSDUH dataset.

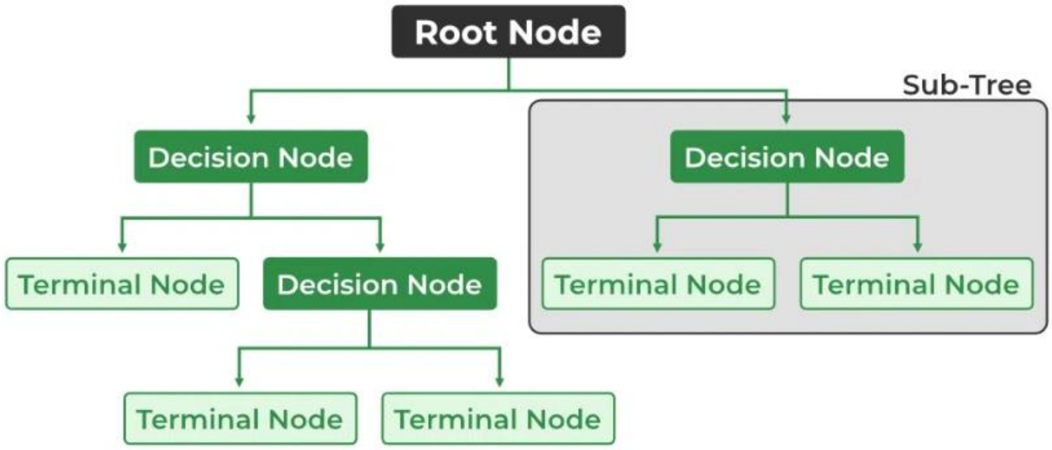
In this report, we focus on analyzing youth drug use patterns using decision tree analysis applied to the NSDUH 2020 dataset. By leveraging demographic characteristics, youth experiences, and other relevant variables, we aim to uncover the predictors of drug use among youth and understand the underlying mechanisms driving substance use behaviors. Through this analysis, we seek to contribute to the broader understanding of youth substance use and inform evidence-based interventions aimed at promoting healthy behaviors and reducing substance abuse among adolescents and young adults.

Background

Decision Trees:

Decision trees are a popular type of supervised machine learning algorithm used that is mostly used in classification problems. It is important to note, that it works for both categorical and continuous input and output variables. They work by recursively partitioning the feature space into regions, based on the values of input features, to predict the target variable. Each partition represents a decision node, where a decision is made based on a certain feature value. These decisions lead to the formation of branches that eventually terminate in leaf nodes, each corresponding to a predicted outcome.

The splitting process in decision trees is based on maximizing the purity of the resulting nodes, typically measured by metrics like Gini impurity or entropy. The algorithm iteratively selects the feature and split point that maximize purity, creating a binary tree structure.

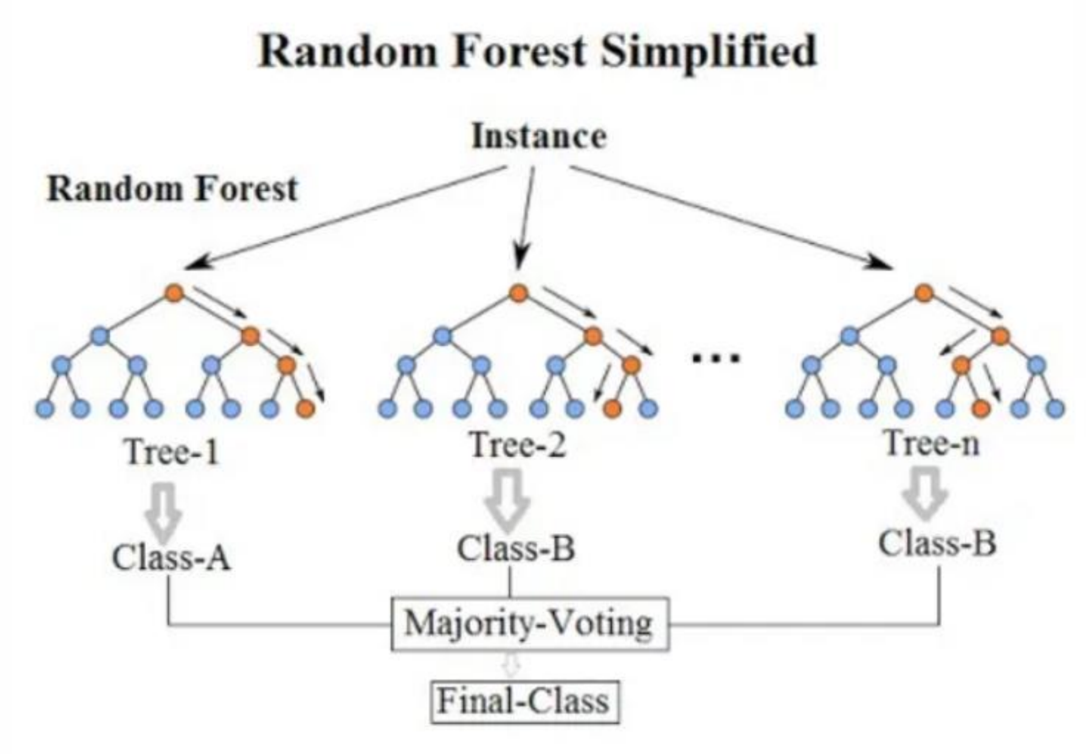


Decision trees have several advantages, including interpretability, ease of use, and the ability to handle both numerical and categorical data. However, they are prone to overfitting, especially with complex dataset.

Random Forest:

Random Forest is an ensemble learning method that builds multiple decision trees and combines their predictions to improve accuracy and reduce overfitting. It works by training each tree on a random subset of the training data and a random subset of features. During prediction, the ensemble of trees votes on the final outcome, with the majority vote determining the final prediction.

The algorithm introduces randomness through two key parameters: the number of trees (ntree) and the number of features considered at each split (mtry). By averaging predictions from multiple trees, Random Forest reduces variance and produces more robust models compared to individual decision trees.



Random Forest is effective for handling high-dimensional data, capturing complex interactions, and providing feature importance measures. However, it may be computationally intensive due to the training of multiple trees, especially with large datasets.

Methodology

Data Preparation:

The methodology employed in this analysis begins with data preparation, utilizing the National Survey on Drug Use and Health (NSDUH) 2020 dataset. I began by loading the 'youth_data' dataset, handling missing values through omission/case-wise deletion strategy to ensure data integrity. To understand the dataset, I conducted exploratory analysis using summary statistics, visualizations (including a correlation plot), and investigated the data structure. Target variables for alcohol use (binary), cigarette frequency (continuous), and marijuana use frequency (multi-class) are created based on the survey responses Following data preparation, the dataset is divided into training and testing sets using an 80-20 split ratio to facilitate model evaluation. The training set is utilized to train decision tree and random forest models, while the testing set is reserved for evaluating model performance. For the regression problem, I recoded the "ircigfm" variable to improve model interpretation.

Models:

Next, computations are conducted using decision trees and random forests for both regression and classification tasks. For cigarette frequency (regression), I trained a decision tree, explored tree visualizations using both tree and rpart.plot packages, and calculated the evaluation metric mean squared error (MSE). To enhance the tree model's performance, I investigated pruning using cross-validation. Subsequently, I employed a Random Forest regressor, examining feature importances and comparing its MSE to the decision tree.

For the binary and multi-class classification problems (alcohol use and marijuana use frequency), I created appropriate target variables, split the data into training and testing sets, built decision tree classifiers, visualized the trees, and evaluated accuracy for classification tasks (and Mean Squared Error (MSE) for regression). These metrics provide insights into the effectiveness of the models in predicting alcohol use, cigarette frequency, and marijuana use frequency. Finally, I used Random Forest classifiers on both classification problems, again measuring accuracy and visualizing feature importances to compare performance improvements with ensemble methods.

Computational Results and Discussion

Plot:

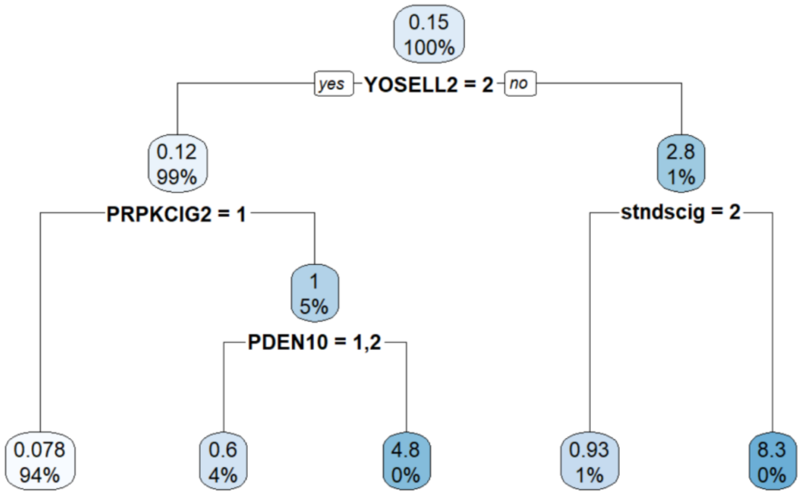
The decision tree generated for the regression variable, ircigfm, which represents cigarette frequency over a month.

Root Node:

The tree starts by splitting the data based on the variable YOSELL2 (RC-Youth Sold Illegal Drugs, (1 = yes, 2 = no)). This indicates that whether a youth sold illegal drugs is the most important factor affecting cigarette frequency.

Initial Split:

Branch on YOSELL2 = 1 (Sold Illegal Drugs): If a youth did sell illegal drugs (YOSELL2 = 1), the tree predicts their cigarette frequency to be 2.8. This suggests that youth who sold illegal drugs tend to have a higher cigarette frequency on average.



Branch on YOSELL2 = 2 (Did Not Sell Illegal Drugs): If a youth did not sell illegal drugs (YOSELL2 = 2), the tree predicts their cigarette frequency to be 0.12. The tree further splits the data based on another variable, PRPKCIG2.

Second Split

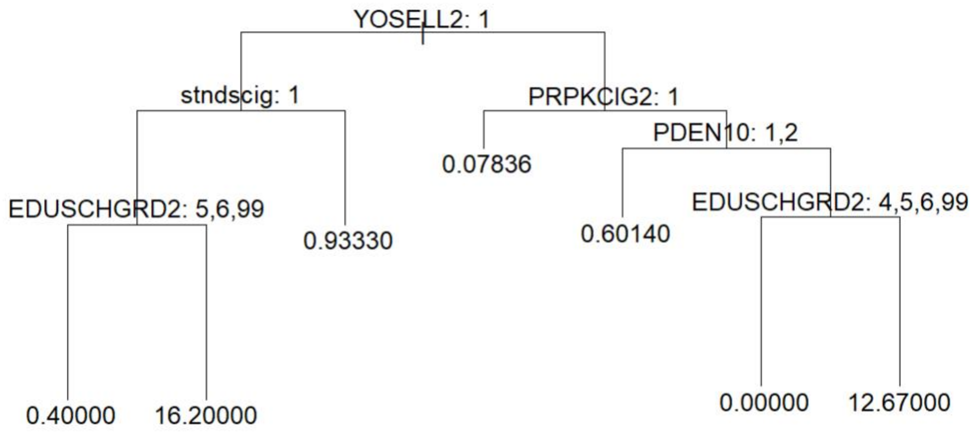
For youths who sold illegal drugs (YOSELL2=Yes), the tree splits the data based on the variable PRPKCIG2, which inquires about How parents feel about youth smoke pack of Cigarettes per day (Yes or No).

- If PRPKCIG2 is "Strongly Disapprove" (meaning the youth believes their parents would strongly disapprove), then the predicted cigarette frequency is lower (around 0.78 cigarettes per day). This suggests that parental disapproval of smoking might be associated with lower cigarette use among youth.
- If PRPKCIG2 is "Somewhat Disapprove or Neither" (meaning the youth is unsure or believes their parents somewhat disapprove), then the predicted cigarette frequency is the highest in the tree (around 5 cigarettes per day). This is interesting because it suggests that uncertainty or a more neutral parental stance regarding smoking might be linked to the highest levels of cigarette use in this group.

If parents somewhat disapprove or have no opinion of youth smoking, the final split considers whether students in the Yth grade smoke cigarettes (STNDSCIG, 1 = Most/All, 2 =Few/None).

- If most or all students in the youth's grade smoke (STNDSCIG = 1), the tree predicts the highest cigarette frequency (8.3).

- If none or few students in the youth's grade smoke (STNDSCIG = 2), the tree predicts a medium cigarette frequency (0.93).

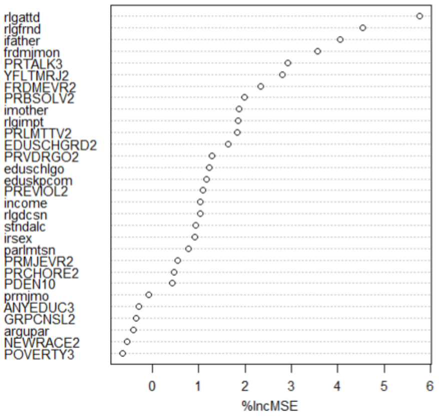


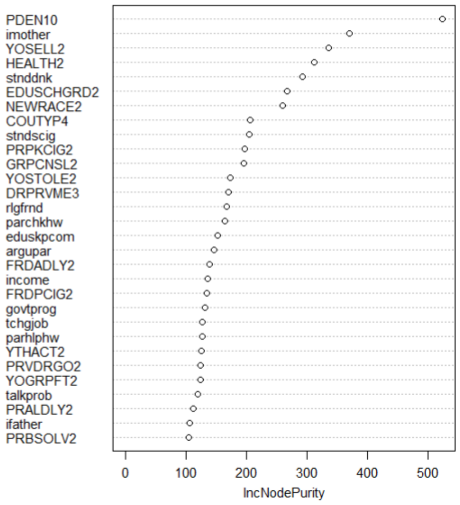
Third Split

This split suggests that population density might be an essential factor influencing cigarette frequency PDEN10 (Population Density). Potentially, youth in denser areas (PDEN10=1) have different cigarette use patterns compared to those in less dense areas (PDEN10=2) or non-CBSA locations (PDEN10=3).

- If the youth live in a segment in a CBSA with 1 million or more people (PDEN10 = 1), then they are predicted to have a lower cigarette frequency (0.6 on the probability scale).
- If the youth live in a segment in a CBSA with fewer than 1 million people (PDEN10 = 2) or in a segment not in a CBSA (PDEN10 = 3), then they are predicted to have a higher cigarette frequency (4.8 on the probability scale).

Examining variable importance through Random Forest:





The variable importance plot you generated from your random forest regression model ranks the features (variables) by their contribution to predicting drug use (the target variable). The features at the top of the plot are the most important for predicting drug use.

Discussion:

This analysis delves into the predictors of youth substance use, revealing several significant factors across various domains. Social influences play a pivotal role, with variables such as exposure to substance use by others, including peer marijuana use (`yflmjmo`) and parental attitudes towards smoking (`PRPKCIG2`), influencing youth behaviors. Additionally, school-related factors emerge as influential predictors, including grade level (`EDUSCHGRD2`) and the prevalence of substance use among peers (`STNDSCIG` and `stndalc`). Moreover, individual attitudes, particularly regarding trying marijuana (`YFLTMRJ2`), demonstrate a strong association with alcohol use.

Interpretation and Implications:

Social Influences: These findings strongly underscore how powerful social norms, both from peers and parents, can be in shaping youth substance use behaviors. School Environment: The school context appears to be a critical factor regarding exposure to substances and social influences. Individual Attitudes: Even though this data cannot determine causation, the strong association between a youth's own feelings about marijuana and their alcohol use warrants further investigation.

Multifaceted Prevention: Effective prevention programs should address both individual attitudes towards substances and broader social influences from peers and family. School-Based Interventions: The significance of school-related factors suggests targeted interventions within the school environment could have significant impact. Early Identification: Understanding these key predictors can help identify youth potentially at higher risk, allowing for early interventions.

Results

For predicting cigarette frequency, the decision tree model had a mean squared error (MSE) of 3.4427. When compared to a random forest model, the MSE improved to 2.677, indicating better accuracy. Important predictors are YOSELL2 (RC-Youth Sold Illegal Drugs), PRPKCIG2(parents feel about youth smoke pack of Cigarettes per day), STNDSCIG(students in the Yth grade smoke cigarettes).

In the case of alcohol use, the decision tree model achieved an accuracy of 79.97%. After employing a random forest model, the accuracy increased to 81.26%, demonstrating improved performance. Important Predictors are YFLTMRJ2(How youth feel when they try Marijuana), stndalc (students in Yth grade drink Alcohol Beverages), EDUSCHGRD2(WHAT GRADE IN NOW/WILL BE IN).

Regarding marijuana use frequency prediction, the decision tree model achieved an accuracy of 86.4%. Upon implementing a random forest model, the accuracy remained relatively stable at 86.3%, indicating similar performance to the decision tree model. Important predictors are yflmjmo(How Yth Feels: Peers Using Marijuana Monthly), stndsmj(Students in Yth Grade Use Marijuana), prmjmo (Yth Think: Parents Feel About Yth Use Marijuana Monthly).

Conclusion

In this study, we employed decision tree analysis to investigate youth drug use patterns using data from the 2020 National Survey on Drug Use and Health (NSDUH). By examining predictors such as social influences, school-related factors, and individual attitudes, we aimed to understand the underlying mechanisms driving drug use among youth. Our findings underscore the significant role of social norms and environmental influences, with variables such as exposure to substance use by others and parental attitudes towards smoking emerging as key predictors.

Moreover, the school context appears pivotal in shaping youth behaviors, highlighting the importance of targeted interventions within educational settings. Early identification of high-risk youth can facilitate timely interventions, while targeted programs within schools hold promise for easing usage among adolescents. By understanding and addressing the complex dynamics of youth drug use, we can work towards promoting healthier behaviors and reducing the prevalence of drug misuse in this vulnerable population.

References

1. National Survey on Drug Use and Health (NSDUH):
   Substance Abuse and Mental Health Services Administration (SAMHSA). (Year). National Survey on Drug Use and Health (NSDUH). Retrieved from NSDUH Website: https://www.datafiles.samhsa.gov/dataset/national-survey-drug-use-and-health-2020-nsduh-2020-ds0001
   Codebook : NSDUH-2020-DS0001-info-codebook.pdf (samhsa.gov)
2. Hastie, T., Tibshirani, R., & Friedman, J. (2009):
   Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer. ISBN: 978-0-387-84857-0

Appendix

This is the link of .html file of my code in Rstudio:

code_file.html

file:///D:/1_MASTERS/3rd%20quarter-Spring%202024/Machine%20learning%202/Youth_Parse_data_assignment_1/code_file.html