STATISTICAL MACHINE LEARNING-2 PRACTICAL HOMEWORK-1 ON YOUTH DRUG ANALYSIS

By Alekhya Dabbiru

MS in DS

INTRODUCTION

- The “Youth Drug Analysis” project investigates the factors that are correlated with the drug use by using decision trees and ensemble methods.
- The youth_data file is a pre-cleaned subset of the NSDUH 2023 dataset, containing responses from approximately 2900 youth under 18 across 79 selected variables. It includes demographics, youth experiences, and use of various drugs, with full descriptions available in the NSDUH codebook.

MODELLING OVERVIEW

In the coding part, we've covered all the three required models using R programming in RStudio:

- BINARY CLASSIFICTION- Can we predict whether a youth has ever consumed alcohol?
- MULTI-CLASS CLASSIFICTION– How frequently has a youth used marijuana over the past year?
- REGRESSION - How many days did a youth consume alcohol in the past month?

THEORETICAL BACKGROUND

- **DECISION TREE:** A decision tree is a flowchart-like model used to make decisions based on features. For example, if you want to determine whether an animal is a dog, cat, crocodile, or other, you can use features such as "has hair," "sound it makes," and "domestic." [1]
- **PRUNING:** Pruning is a technique used to reduce the size of a decision tree by removing sections of the tree that provide little power in predicting target variables. The main goal is to enhance the model's generalization ability, thereby reducing overfitting. [1]

THEORTICAL BACKGROUND: ENSEMBLE METHODS

- **BAGGING:** Bootstrap Aggregating, also known as bagging, is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression. It decreases the variance and helps to avoid overfitting. It is usually applied to decision tree methods. Bagging is a special case of the model averaging approach. [2]

THEORTICAL BACKGROUND: ENSEMBLE METHODS

- **BOOSTING:** Boosting is an ensemble modeling technique designed to create a strong classifier by combining multiple weak classifiers. The process involves building models sequentially, where each new model aims to correct the errors made by the previous ones. [2]

- **RANDOM FOREST:** Random Forest is a type of classifier that uses many decision trees to make predictions. It takes different random parts of the dataset to train each tree and then it combines the results by averaging them. This approach helps improve the accuracy of predictions. Random Forest is based on ensemble learning.[3]

DATA CLEANING

- Variable meanings, coding schemes, and valid value ranges are documented in the NSDUH Codebook, which is essential for interpreting responses correctly (e.g., distinguishing between “no use,” “missing,” and “declined to answer”).
- Additional cleaning was performed during modeling, including removing extreme values (like 991/998 codes), converting categorical and binary variables , renaming variables for better readability (e.g., STNDALC to Friend_Drinks_Daily) and handling missing data with na.omit() to prepare for decision tree training and evaluation.

METHODOLOGY

BINARY CLASSIFICATION: In our binary classification model, we have used the “ALCFLAG” variable to predict whether a youth has ever consumed alcohol.

- A decision tree was built using demographic and youth experience predictors, and pruned to an optimal size of 3.
- We compared models using accuracy: boosting performed best (80.7%), followed by random forest (79.9%) and bagging (79.2%).
- The Key predictors included Friend_Drinks_Daily, Friends_Offers_Marijuana and Friend_Marijuana_Monthly.

METHODOLOGY

MULTI-CLASS CLASSIFICATION: In the multi-class classification model, we used the "MRJYDAYS" variable to capture how frequently marijuana was used over the past year, grouping responses into six usage levels: NEVER, RARE, OCCASIONAL, REGULAR, FREQUENT, DAILY.

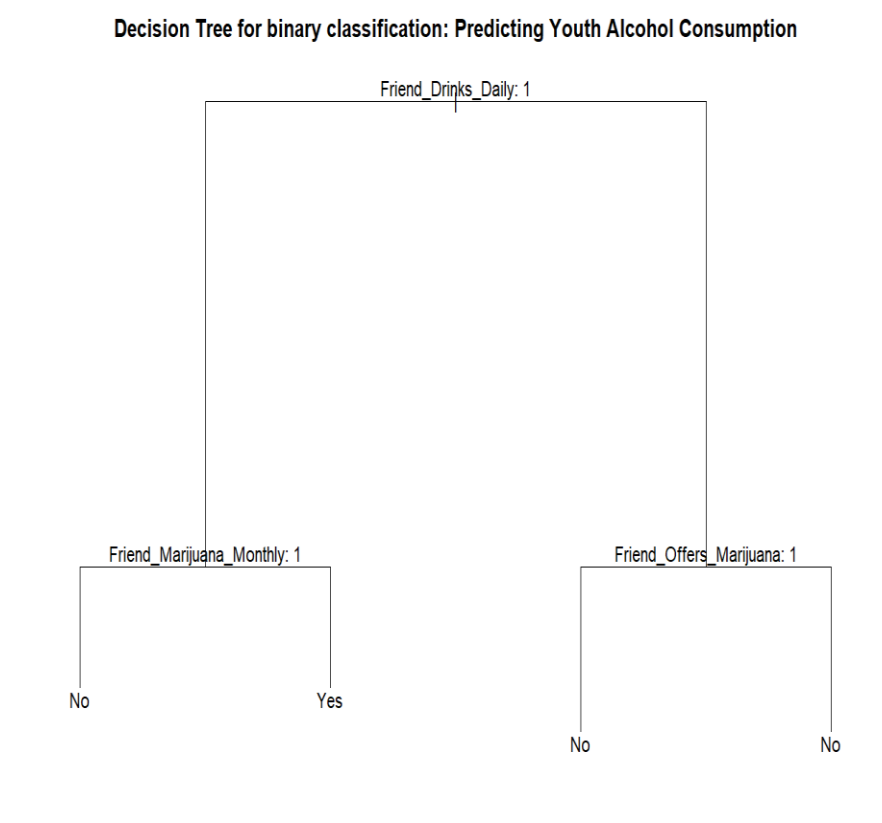
- A decision tree was built using demographic and youth experience predictors and pruned to an optimal size of 3.
- Model accuracy was evaluated on the test set: random forest performed (88.8%), and decision tree (86.6%).
- The key predictors included Friend_Uses_Marijuana_Monthly and Friend_Smokes_Marijuana.

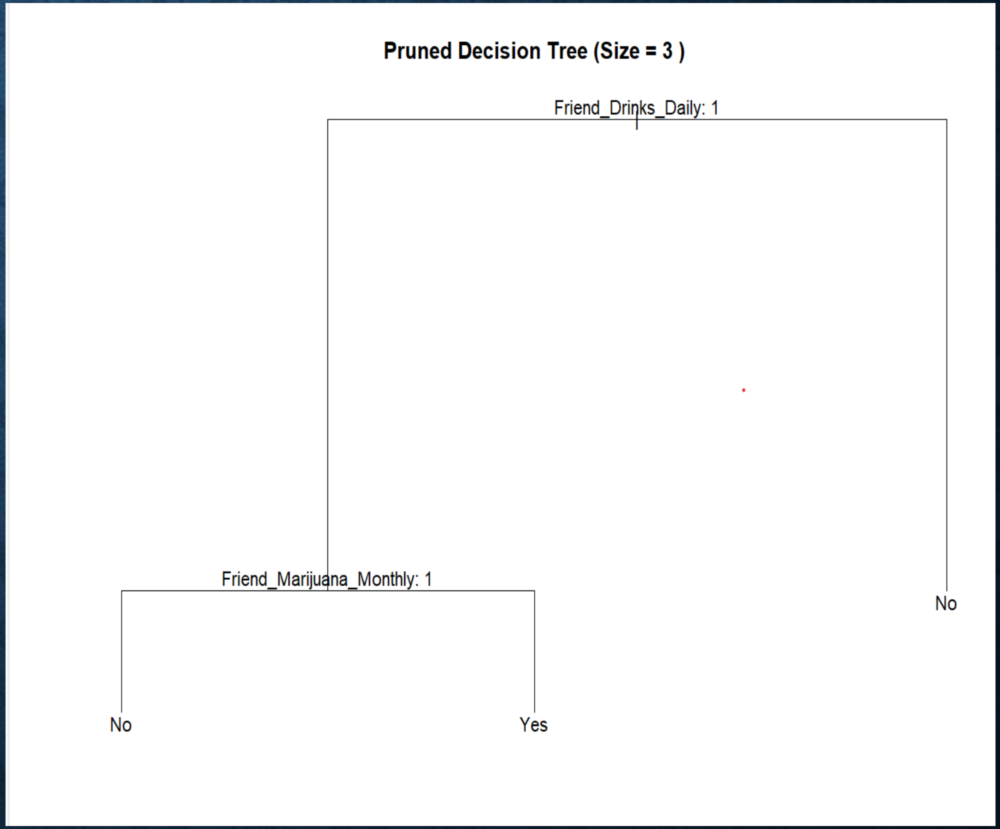
METHODOLOGY

REGRESSION: In regression model, we used the "IRALCFM" variable to predict how many days a youth consumed alcohol in the past month.

- A decision tree was built using demographic and youth experience predictors, and pruned to an optimal size of 2.
- Model performance was evaluated using Mean Squared Error (MSE): bagging performed best (MSE = 1.60), followed by the pruned decision tree (MSE = 1.69) and the full decision tree (MSE = 1.77).
- The key predictors included Friend_Marijuana_Monthly and Friend_Ever_Used_Marijuana.

BINARY CLASSIFICATION: DECISION TREE BEFORE VS AFTER PRUNING


Decision Tree for binary classification: Predicting Youth Alcohol Consumption
Friend_Drinks_Daily: 1
Friend_Marijuana_Monthly: 1
Friend_Offers_Marijuana: 1
No
Yes
No
No



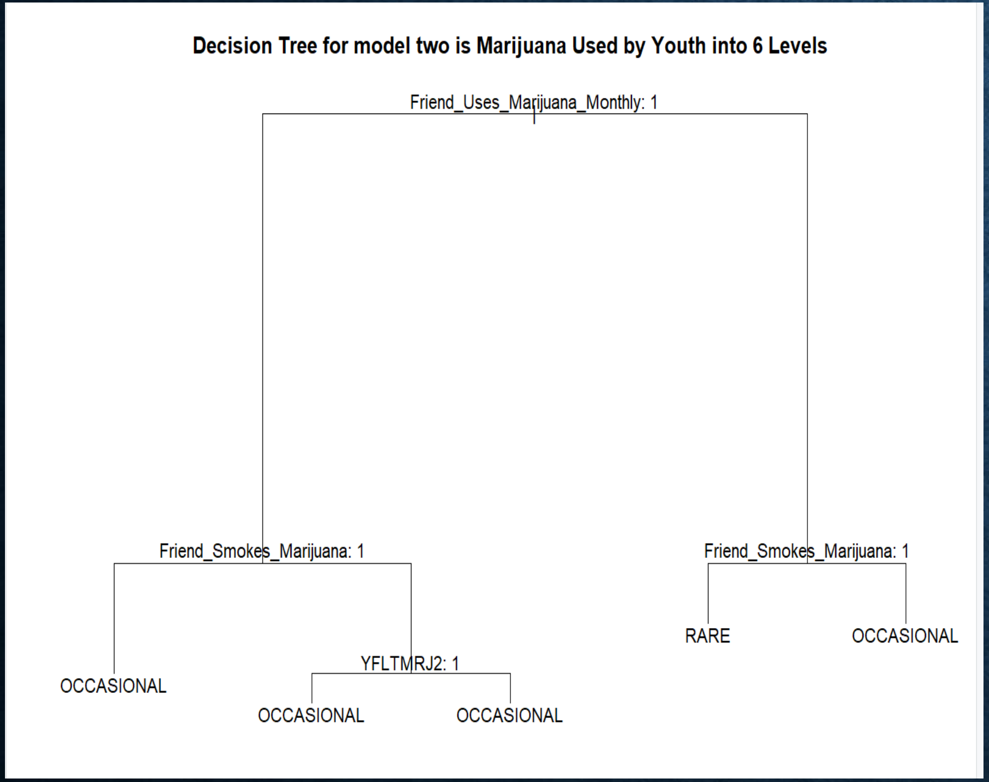
FLOW OF THE TREE FOR BINARY CLASSIFICATION MODEL

- For predicting whether a youth has ever consumed alcohol, the tree has been split into 2 branches:
- Split-1: Checks if the youth's friend drinks daily, if Friend_Drinks_Daily = 1, then the tree moves to another split, else predicts "No".
- Split-2: This occurs when Friend_Drinks_Daily = 1, then Checks if the friend uses marijuana monthly, if Friend_Marijuana_Monthly = 1 then tree predicts "No", else the tree predicts "Yes".

FLOW OF THE NODE FOR BINARY CLASSIFICATION MODEL

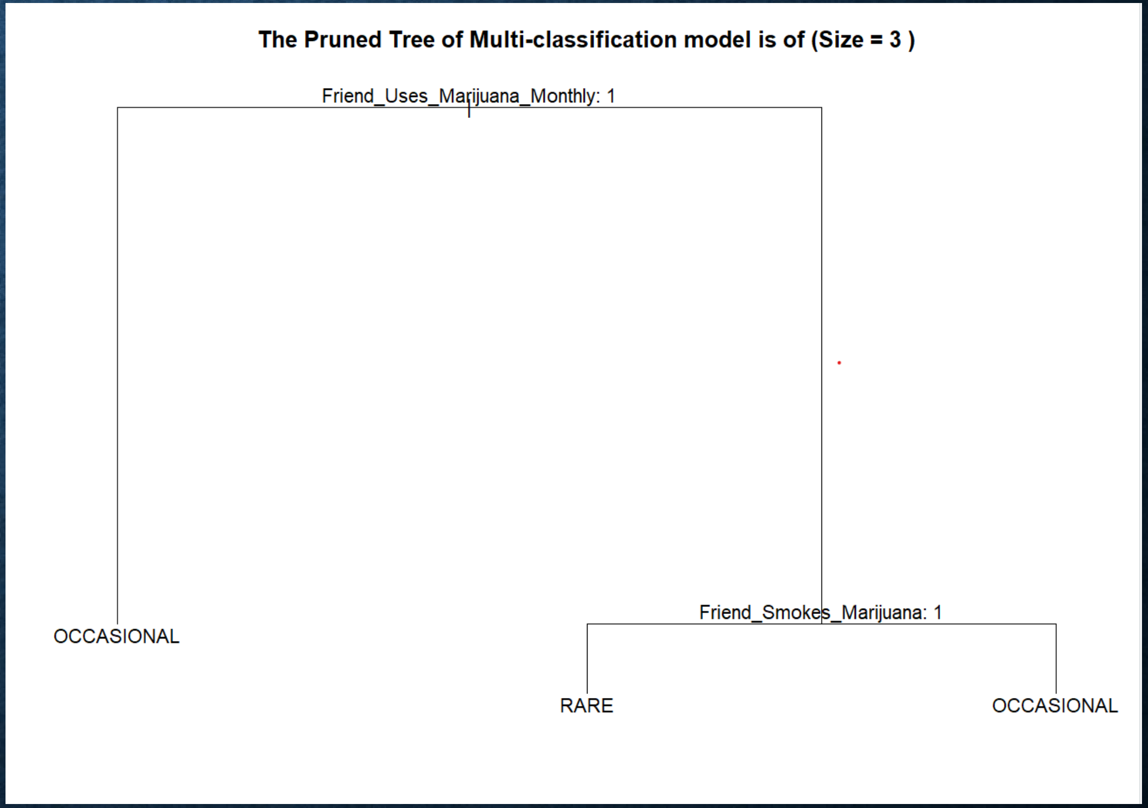
- At one noteworthy end node, the model predicts that the youth has consumed alcohol(YES). This node is reached by first checking if Friend_Drinks_Daily equals 1 , meaning the youth's friend drinks daily. Then the model looks at Friend_Marijuana_Monthly, when this variable equals 2, indicating that the friend uses marijuana frequently, the prediction is "Yes."
- The model suggests that when a youth has friends who both drink daily and frequently consume marijuana, their likelihood of consuming alcohol is high. This does not prove that one behavior causes the other, it just shows a relation between them that helps us understand the pattern better.

MULTI-CLASS CLASSIFICATION: DECISION TREE BEFORE VS AFTER PRUNING



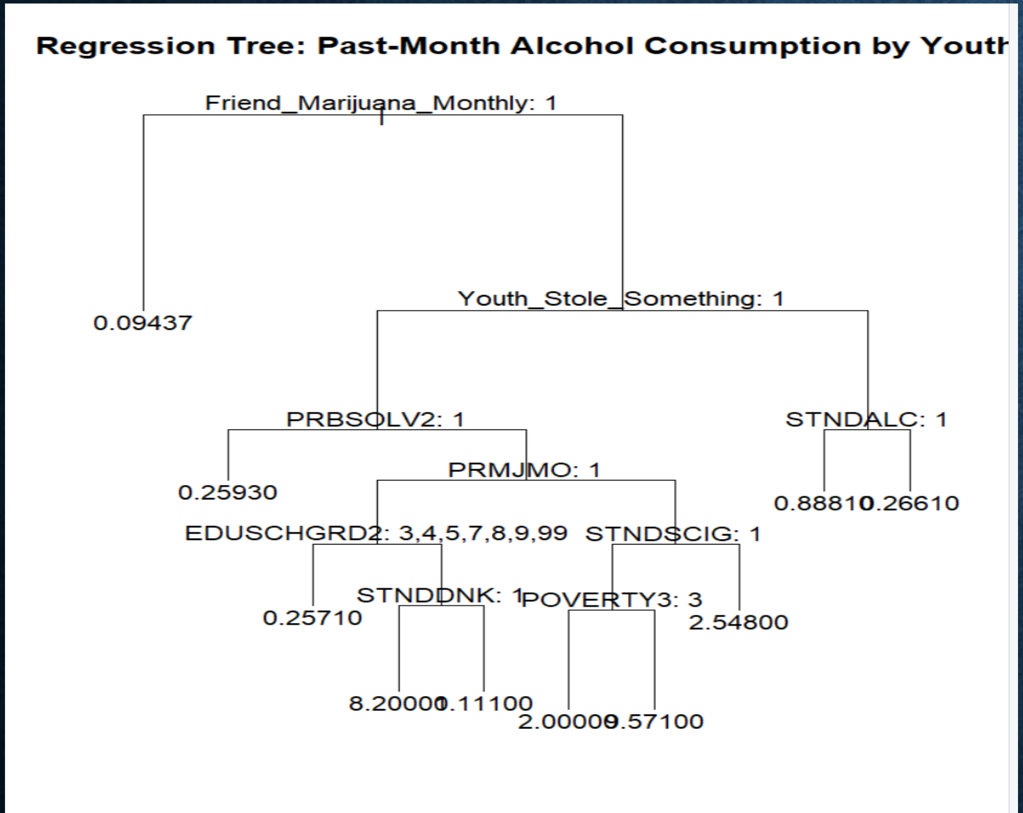
Decision Tree for model two is Marijuana Used by Youth into 6 Levels

Friend_Uses_Marijuana_Monthly: 1
Friend_Smokes_Marijuana: 1
Friend_Smokes_Marijuana: 1
RARE
OCCASIONAL
YFLTMRJ2: 1
OCCASIONAL
OCCASIONAL
OCCASIONAL

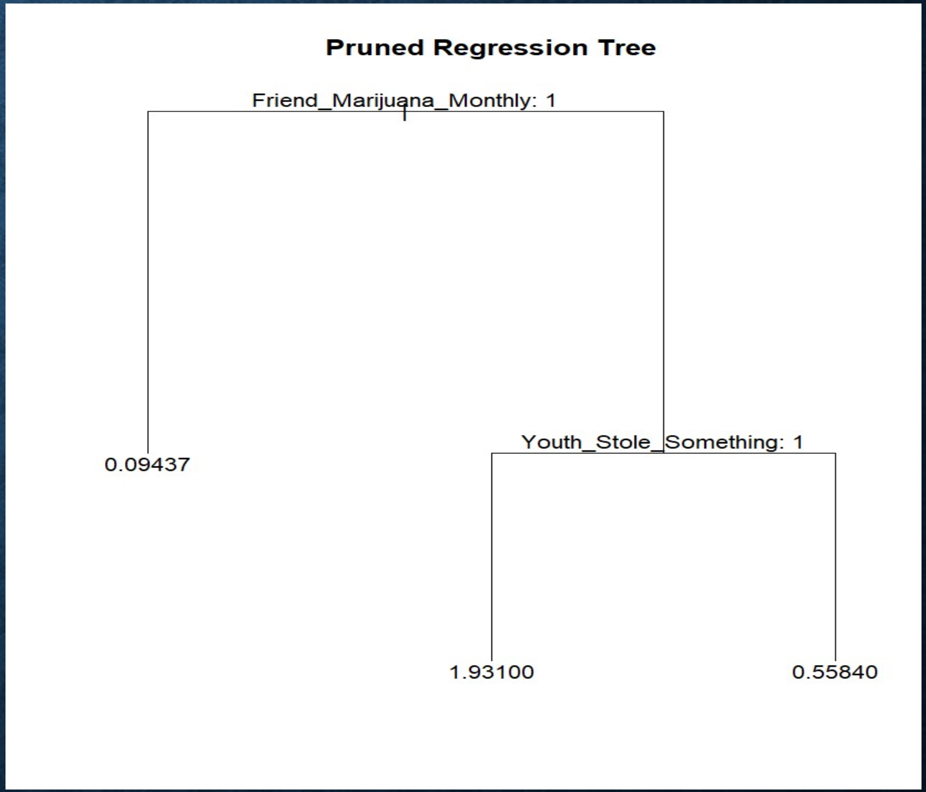


The Pruned Tree of Multi-classification model is of (Size = 3)

REGRESSION: DECISION TREE BEFORE VS AFTER PRUNING



Regression Tree: Past-Month Alcohol Consumption by Youth



Pruned Regression Tree

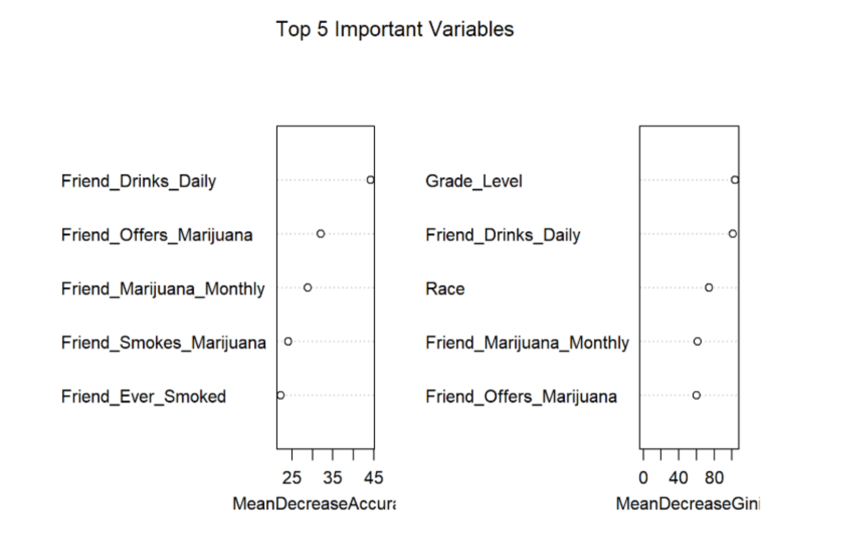
DISCUSSION 1: SOME OF THESE VARIABLES ARE THE SAME INFORMATION CODED INTO BINARY, ORDINAL (CATEGORICAL AND ORDERED), AND NUMERICAL VARIABLES. HOW DO THE PREDICTIONS CHANGE USING EACH DATA TYPE? WHAT IS EACH TELLING YOU, AND WHEN IS IT APPROPRIATE TO USE EACH?

- In this project case, several key predictors such as Friend_Drinks_Daily, Friend_Marijuana_Monthly, and Friend_Offers_Marijuana are seen in the code with multiple data types including binary, ordinal, and numerical, each conveying different nuances.
- For example, a binary variable like Friend_Drinks_Daily with responses ‘Yes’ or ‘No’ yields straightforward decision splits that are easy to understand. An ordinal variable, such as the binned version of marijuana usage frequency with levels NEVER, RARE, OCCASIONAL, REGULAR, FREQUENT, and DAILY, provides a sense of progression and captures gradations in behavior.
- A numerical variable offers fine granularity and enables the model to detect subtle thresholds, however, these values must be used cautiously as they assume a linear relationship that may not always hold. In our case, binary coding provided clear and direct influences, ordinal coding captured the intensity of usage patterns, and numerical data offered additional detail when the variation was significant. Ultimately, the choice of data type should match the level of detail required and the inherent ordering present in the variable.

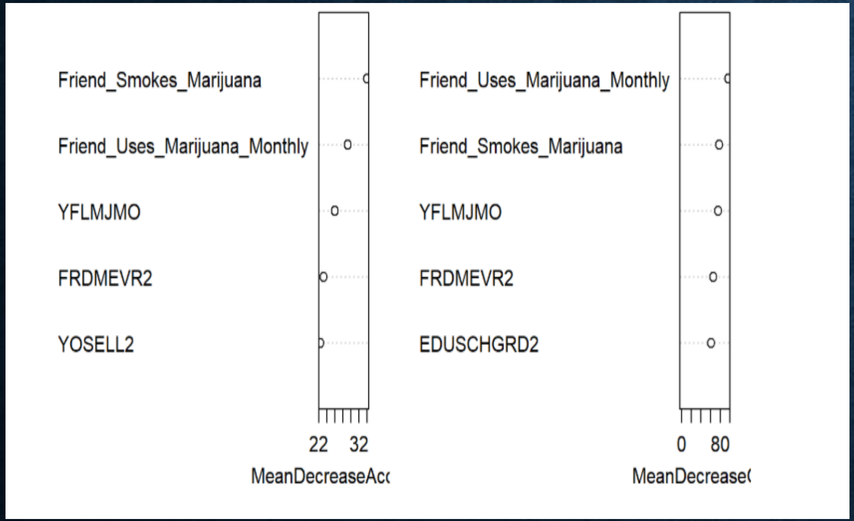
DISCUSSION 2: WHICH VARIABLES TEND TO BE IMPORTANT FOR PREDICTING DRUG USE? HOW CAN THESE BE INTERPRETED? WHAT ARE THE IMPLICATIONS OF THIS OUTCOME? HOW CAN YOU, AS THE DATA SCIENCE COMMUNICATOR, DISCUSS THESE FINDINGS IN AN ETHICAL WAY?

- In the binary classification model, the top predictors identified were Friend_Drinks_Daily, Friend_Marijuana_Monthly, and Friend_Offers_Marijuana; in the multi-class classification model, Friend_Uses_Marijuana_Monthly and Friend_Smokes_Marijuana emerged as key predictors; and in the regression model, Friend_Marijuana_Monthly and Youth_Stole_Something were found to be critical.
- These findings suggest that friends-related and risk-behavior variables strongly influence our drug- use predictions, but they represent associations rather than proof of influence and therefore require cautious interpretation.
- As data science communicators, we should avoid labelling people based on these factors. Instead, it's best to point out the limits of this study and highlight the need for further research so that we present a balanced view of what our data really tells us.

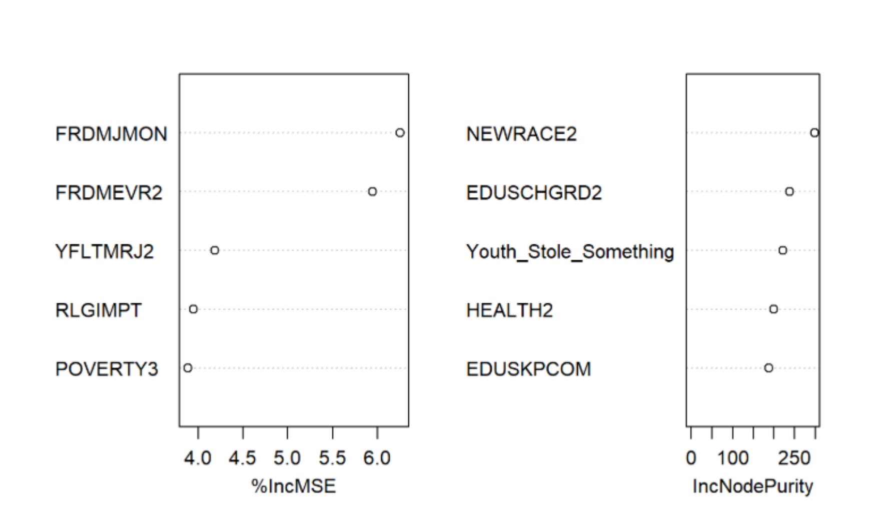
THE TOP FIVE MOST IMPORTANT VARIABLES



BINARY CLASSIFICATION MODEL



MULTI-CLASS CLASSIFICATION MODEL



REGRESSION MODEL

CONCLUSION

- The study shows that decision trees and ensemble methods do a good job predicting youth drug use.
- In the binary model, boosting led with 80.7% accuracy (influenced by factors like Friend_Drinks_Daily and Friend_Marijuana_Monthly). For marijuana use frequency(multi-class classification model), random forest performed at 88.8% accuracy(with Friend_Uses_Marijuana_Monthly and Friend_Smokes_Marijuana as top predictors). In the regression model, predicting alcohol consumption days, bagging performed well with mean squared error of 1.60 (key predictors like Friend_Marijuana_Monthly and Youth_Stole_Something).
- Overall, these results suggest that friend's behavior and certain risk factors are strongly related to the drug use, though they show associations rather than proving direct links.

BIBLIOGRAPHY

[1] Yepes, A. F. T. Understanding Tree Algorithms: Decision Trees, Pruning, Bagging, and Boosting. Medium. Retrieved from https://medium.com/@aftellez/understanding-tree-algorithms-decision-trees-pruning-bagging-and-boosting-5c50ef27d7f1

[2] GeeksforGeeks. Bagging vs Boosting in Machine Learning. Retrieved from https://www.geeksforgeeks.org/bagging-vs-boosting-in-machine-learning/

[3] GeeksforGeeks. Random Forest Algorithm in Machine Learning. Retrieved from https://www.geeksforgeeks.org/random-forest-algorithm/

[4] Class presentation slides and worksheets. Data 5322 Coursework, Seattle University, 2025.

THANK YOU!!!