Hrishabh Kulkarni

Script

Slide 1: Title Slide

Good morning, everyone. Today, I'll walk you through my analysis of youth drug use patterns using decision trees and ensemble methods. This work, supervised by Dr. Mendible, leverages NSDUH survey data to answer three critical questions about marijuana, alcohol, and cigarette use among youths."

Slide 2: Mission & Dataset

My goal was to predict drug use behaviors using three modeling approaches: binary classification for marijuana use, multi-class for alcohol frequency, and regression for cigarette days. The dataset covers demographics, parental behaviors, and social influences- key factors known to shape youth decisions.

Slide 3: Theoretical Background

Let's briefly recap the methods:

Decision Trees split data recursively using parameters like max_depth to avoid overfitting.

Bagging & Random Forest build multiple trees; RF adds randomness by selecting subsets of features (mtry).

Boosting iteratively corrects errors, controlled by shrinkage and n.trees.

Why It Matters:

"Pruned trees simplify interpretation, while boosting often delivers the highest accuracy—as we'll see in my results."

Slide 4: Key Questions & Variables

We focused on three questions:

Binary: Predict whether a youth has ever used marijuana (binary: Yes/No) based on demographics, youth experiences, and peer/parental influences?

Multi-Class: How frequently do youths use alcohol (categories: None, 1-2 days/month, 3+ days/month) based on their social environment and demographics?

Regression: What is the best factors to predict the number of days a youth has used cigarettes in the past 30 days (continuous count)?

Predictors like ‘ParentsSmoke’ and ‘PovertyLevel’ were chosen for their documented impact on substance use."

Slide 5: Data Cleaning

We cleaned the data by removing NAs, balancing splits (70/30), and converting variables. For instance, marijuana use MRJ became a factor (Yes/No), while cigarette days IRCTGFM stayed numeric.

Slide 6: Hyperparameter Tuning

"We tuned parameters rigorously:

Decision Trees: cp=0.01 for pruning, minsplit=10 to control splits.

Boosting: shrinkage=0.001 (regression) and 0.01 (binary) to optimize learning rates."

Talk about RF and other in ppt.

Slide 7: Decision Tree Path Analysis

Discuss on one high-risk path in my pruned marijuana tree:

Friends Use MJ=Yes -> 100% usage likelihood.

School Punishment=Moderate -> Drops to 77% avoidance.

Parental Disapproval=Low + Friends Offered MJ=Yes -> 46% usage.

(Visualize the tree path)

Takeaway:

Peer influence dominates, but school policies and parental attitudes can mitigate risks—especially when combined with behavioral markers like group fights.

Slide 8: Binary Results

Boosting outperformed here with an AUC >0.9. While sensitivity was strong (95%), specificity lagged at 47%- meaning we’re better at identifying non-users than users. This hints at needed improvements, like cost-sensitive learning.

Discuss on plots and table.

Slide 9: Multi-Class Challenges

A stark reality: my models achieved 77% accuracy by always predicting the majority class ('None'). Rare categories were ignored—a classic imbalance problem.

Show class distribution plot

Slide 10: Regression Insights

"Random Forest won here with the lowest MSE (123.84). The partial dependence plot reveals a linear trend: as parental smoking increases, so does youth use. This underscores the need for family-focused interventions."

Show RF vs. Boosting MSE

Slide 11: Future Scope

Three key next steps:

Add geographic data to capture environmental effects.

Fix imbalance with SMOTE or class weights in XG Boost.

Model interactions ('Parents Smoke + Friends Smoke').

These could unlock deeper insights- like whether peer effects amplify in high-poverty areas.

Slide 12: References & GitHub

Talk about Citation and where and why have used in the code.

All methods and citations are listed here. The full code is on GitHub- feel free to have a look,

Conclude: peer influence and parental behaviors are pivotal, but data quality and model choice dictate what we can learn. Thank you, and I'd love your thoughts on where to take this next! Or improvements or changes if any.