Slide 1: Title Slide

Hello everyone, my name is Pavan Sai, and this is my project for Statistical Machine Learning 2.

In this presentation, I'll walk you through how I explored patterns in youth substance use using tree-based models and ensemble methods on the NSDUH 2023 dataset.

We'll cover three tasks: binary classification, multiclass classification, and regression. Along the way, I'll show you how each model behaved, how we evaluated them, and what insights we gained.

Slide 2: Abstract

This study aimed to analyze how behavioral, demographic, and environmental factors influence youth outcomes.

We built decision trees and ensemble models like Bagging, Random Forests, and Gradient Boosting to predict school attendance, marijuana usage frequency, and alcohol consumption days.

Some of the most predictive variables included the age of first marijuana use and frequency of alcohol consumption.

Gradient Boosting performed especially well in regression, and peer/parent influence shaped multiclass outcomes. The overall goal was to not just predict, but understand what drives these behaviors.

Slide 3: Questions Chosen

We explored three core research questions:

First: Can we predict if a youth skipped school recently based on their substance use behavior?
Second: Can we classify their frequency of marijuana use into three groups using home and peer environment data?
And third: Can we predict how many days they drank alcohol using demographics and parental involvement?

Slide 4: Dataset

We used the NSDUH 2023 dataset, focusing on youth under 18.

Key features fell into four categories:
Substance use behavior, such as how frequently they used alcohol, marijuana, or cigarettes.
Parental involvement — things like homework check-ins or emotional support.
Peer influence — for example, whether their friends used substances.
And demographics like gender, income, or education.

We handled missing values, cleaned special codes like 991, and performed downsampling where needed to fix class imbalance.

Slides 5–6: Theoretical Background

Let me briefly explain the models we used:

Decision Trees split the data based on feature values to form pure groups. They're easy to interpret but can overfit, so we used pruning via cross-validation.

Bagging trains multiple trees on bootstrapped samples and averages their predictions. It's a great way to reduce variance.

Random Forests improve on bagging by adding feature-level randomness at each split, which helps decorrelate trees and reduce overfitting.

And **Gradient Boosting** builds trees sequentially, each one correcting the errors of the previous. It's powerful, especially when tuned, but more sensitive to imbalance.

Slide 7: Binary Classification – Methodology

For binary classification, our target was SKIPPED_SCHOOL, created from the variable EDUSKPCOM.

We used features like substance frequency (IRALCFM, IRMJFM), age of first use, and behavioral flags like MRJFLAG.

We converted codes like 991 to NA, ensured proper factor types, and didn't downsample since class distribution was fair.

We tested Decision Tree, Bagging, Random Forest, and GBM, and used a tuned threshold of 0.4 for probabilistic models to improve recall.

Slide 10–12: Binary Results – Heatmap & F1 Score

This heatmap shows prediction patterns — true positives were strong, but some recall gaps existed initially.

Our best performer was Bagging with an F1 score of 0.766. It benefited from reduced variance across multiple trees.

Random Forest followed with 0.693. GBM trailed slightly at 0.610 due to sensitivity to threshold tuning.

The pruned Decision Tree had the lowest score — it was interpretable but underfit the data.

Slide 8: Multiclass Methodology

For multiclass, we re-bucketed MRJMDAYS into a new variable SKIP_LEVEL with values: None, Occasional, and Frequent.

Features came from peer and home environment — variables like PARCHKHW, FRDMJMON, and PRMJEVR2.

Because Frequent users were underrepresented, we manually downsampled each class to equal sizes using slice_sample() or sample_n().

We applied Decision Tree, Random Forest, and GBM with multinomial loss.

Slide 13–15: Multiclass Results – Heatmap & F1 Score

Here's how the models performed:

The 'None' class had the highest F1 scores (~0.78) since it had the clearest patterns and support.

'Occasional' was harder to predict due to overlaps with both extremes. Interestingly, Decision Tree overfitted to it slightly.

All models struggled with 'Frequent', which was rare in the data. GBM performed best with an F1 of 0.429 but still lacked strong recall.

This showed us that **class balance and feature clarity** were key in multiclass prediction.

Slide 9: Regression Methodology

In the regression task, we predicted ALCYDAYS — the number of days youth drank alcohol in the past year.

Predictors included demographics (IRSEX, NEWRACE2), parenting features (PARCHKHW, PRTALK3), and behaviors like YTHACT2.

Models used were Decision Tree, Random Forest, and GBM, and we tuned them using cross-validation with different controls per model.

Slide 16–18: Regression Results & Comparison

Random Forest had the lowest error across all metrics: MAE of 1.17, MSE of 2.76, and RMSE of 1.66.

GBM came close with MAE of 1.19, showing that boosting works well on regression if tuned right.

The Decision Tree was the most interpretable but underfit — its MAE was 1.23.

Scatter plots confirmed these differences, and the comparison bar chart makes it easy to see RF's overall edge.

Slide 19–20: Discussion & Challenges

In early models, we mistakenly used EDUSKPCOM and MRJMDAYS as predictors — this leaked label information and caused unrealistic accuracy (up to 95%).

After removing them, accuracy dropped to 51% — more realistic. This highlighted the importance of careful feature selection.

Confusion matrices showed precision was fine, but recall only improved after threshold tuning.

For multiclass, downsampling boosted GBM's macro-F1 from ~0.52 to ~0.586.

We learned that simple accuracy is misleading — heatmaps and F1 scores helped uncover hidden flaws.

Slide 21: Conclusion

To summarize:

Decision Trees were easy to explain but didn't scale well to complex data.
Bagging and Random Forest were stable and performed well, especially in binary classification.
GBM shined in regression and multiclass tasks after proper tuning.

More importantly, we learned the value of avoiding leakage, balancing classes, and choosing models that fit both the data and the interpretability needs.

These findings could support early intervention policies in youth substance monitoring.

Slide 22: Thank You

Thank you for listening! I'm happy to take any questions.