Practical Homework-1 Script

Slide 1: Title Slide

"Hello, everyone. My name is *Pavan Sai*, and welcome to my presentation for *Statistical Machine Learning - 2*. In this project, I explored patterns in youth substance use using decision tree-based models. Our goal was to predict, classify, and estimate behaviors using real-world survey data. I worked under the guidance of *Dr. Mendible*, and this presentation walks through my findings using three key tasks—binary classification, multi-class classification, and regression."

Slide 2: Abstract

"This study leverages the 2023 NSDUH youth survey data to explore behavioral patterns in substance use. We performed three main tasks: identifying if a youth skipped school, classifying their marijuana usage frequency, and estimating how often they consumed alcohol annually.

We tested models including Decision Trees, Random Forests, Bagging, and Gradient Boosting. Key insights? Features like age of first use and peer influence were highly predictive. Gradient Boosting gave us the best regression results with an MSE of 2.56. This project shows how data can guide health and education interventions."

Slide 3: Questions Chosen

"For our analysis, we asked three well-structured questions:

- Binary Classification: Can we predict if a youth skipped school using substance use behavior?
- Multi-Class Classification: Can we classify marijuana use levels based on peer and home environment?
- Regression: Can we estimate alcohol consumption days based on parental involvement and demographics?

These questions are data-driven, realistic, and help us understand risk factors in youth behavior."

Slide 4: Dataset

"The data came from the *NSDUH 2023* survey, filtered to include only youths under 18. The dataset is very rich—it includes demographics, peer and parent interactions, and substance use indicators.

I used preprocessed data provided in youth_data.RData to avoid spending excessive time on cleaning, but I still applied:

- Dropping NAs and invalid codes.
- Re-bucketing marijuana use for classification.
- Downsampling where class imbalance was an issue.

This preprocessing was crucial to avoid bias and leakage in our models."

Slide 5: Binary Classification

"In the binary classification task, our goal was to predict whether a youth skipped school using substance use features. The target variable was SKIPPED_SCHOOL, with 1 meaning they skipped school and 0 otherwise.

We used four models:

- Decision Tree: 55.6% accuracy
- Random Forest: 62.2%
- Gradient Boosting: 57%
- Bagging: 66%

Bagging outperformed others, but during early trials we noticed unrealistically high accuracies due to data leakage, like using the same variable both as input and target in disguised forms. Once fixed, performance normalized and confusion matrices aligned better."

Slide 6: Multi-Class Classification

"This task aimed to classify marijuana usage into None, Occasional, and Frequent based on environment and peer factors. The target was derived from MRJMDAYS and rebucketed.

Random Forest again gave the best accuracy at 62.8%, but we faced major class imbalance initially. Occasional users were underrepresented, leading to misleading accuracy. We handled this using downsampling and selected features unrelated to the target variable's origin to avoid leakage."

Slide 7: Regression

"Our final task was to predict the number of days youth consumed alcohol in a year. The target was ALCYDAYS.

We tested:

- Decision Tree – MSE: 2.88

- Random Forest – MSE: 2.88, MAE: 1.17
- Gradient Boosting – MSE: 2.56, MAE: 1.14

GBM performed best overall. Interestingly, while linear regression was interpretable, its performance lagged. We also avoided overfitting by using fewer trees and adjusting depth in ensemble models. Feature importance helped interpret even the black-box models."

Slide 8: Challenges

"We faced several challenges:

- Data Leakage: Accidentally using target-derived variables. Fixed it by redesigning our feature selection process.
- Class Imbalance: Solved using downsampling.
- False Accuracy: Early models had 99% accuracy but failed on minority classes—confusion matrices were key in identifying this.
- Interpretability vs Performance: While ensemble models performed better, Decision Trees helped us explain the logic."

Slide 9: Conclusion

"This project taught me how critical proper preprocessing and feature selection is in modeling. Random Forest and GBM stood out in their respective tasks, but each model had its value.

Our modeling provides early insights into youth behavior, supporting interventions in education and health. Importantly, we balanced performance with interpretability—something real-world applications always demand."

Theoretical Background and Insights

Decision Tree

How It Works:
A decision tree splits data based on features that reduce impurity (Gini index for classification, variance for regression). It's easy to visualize and interpret.

Why We Used It:
We used decision trees as our starting point because they help uncover patterns and paths in a human-readable form—very useful when we want to say, 'If X happens, then Y is likely.'

Key Insight:
In our project, decision trees helped us confirm something important: youths who report low parental communication or support are more likely to skip school. For example, if a youth does not talk to parents about schoolwork or doesn't feel supported, the decision path often ended in the 'skipped school' class.
However, the tree model sometimes used only one or two variables—like PRLMTTV2 in the regression tree—indicating over-simplification, missing more complex behavioral patterns.

Random Forest

How It Works:
Random forest builds many decision trees using different data and feature subsets and averages their results. This reduces overfitting and improves accuracy.

Why We Used It:
We relied on Random Forests to capture more subtle and complex patterns in youth behavior that single trees couldn't detect.

Key Findings:
- In binary classification, youths who first tried alcohol or marijuana at an early age were more likely to skip school.
- In multi-class classification, peer marijuana use and perception of parental disapproval were strong signals of frequent use.
- For regression, features like income level, parental involvement, and race were top contributors.
These insights suggest that early prevention efforts should focus on peer education and strengthening home environments, especially in lower-income households.

Bagging

How It Works:
Bagging builds multiple full-featured trees on bootstrapped datasets and averages their predictions. It reduces variance without increasing bias.

Why We Used It:
Bagging offered robustness in binary classification where we had slight class imbalance and potential outliers.

What We Learned:
Bagging improved recall slightly for school skipping, especially when we adjusted the classification threshold (e.g., 0.4 instead of 0.5). This meant we were better at catching students who were skipping, even if it meant more false positives.
The model showed that even when substance use isn't extreme, the combination of multiple mild risk factors—like peer exposure and lack of parental monitoring—can predict school disengagement.

Gradient Boosting Machine (GBM)

How It Works:
Boosting adds models sequentially to reduce the error made by the previous model. GBM does this with shallow trees and weighted errors.

Why We Used It:
GBM was essential in regression where small errors mattered. It helped us fine-tune the estimation of alcohol use days.

Insights From GBM:
- Parental monitoring, especially rules around TV or friends (PRLMTTV2, PARCHKHW), were significant in predicting alcohol days.
- Race and income remained strong predictors even after controlling for other variables.
- We also discovered that peer influence mattered less than assumed, suggesting that internal household structure plays a bigger role in frequency of alcohol use.
In short, even when friends use substances, a structured home can buffer against high use.

Methodology

Data Understanding & Preprocessing

We began by understanding the data provided in youth_data.Rdata. This included:

- **Familiarizing ourselves** with variable labels and encoded responses using the codebook.
- Identifying **missing values and outliers** such as 991, 993, or 998 codes and converting them to NA.
- **Dropping irrelevant or redundant features**, especially those directly revealing the target variable, to avoid data leakage.
- For binary classification, we created a new variable SKIPPED_SCHOOL from the EDUSKPCOM column.
- For multi-class classification, we recoded the MRJMDAYS variable into three buckets: None, Occasional, and Frequent marijuana use.
- For regression, we focused on predicting ALCYDAYS — the number of days alcohol was consumed in the past year — using both demographic and parental influence variables.

Model Setup and Implementation

We trained the following models for each task:

Binary Classification – "Did the student skip school?"

Models Used:

- Decision Tree
- Random Forest
- Bagging (with threshold tuning)
- Gradient Boosting Machine (GBM)

Hyperparameters:

- Random Forest: ntree = 500, mtry = 4
- GBM: n.trees = 5000, shrinkage = 0.01, interaction.depth = 3
- Bagging: mtry = number of predictors, tested thresholds: 0.3, 0.4, 0.5

Multi-Class Classification – "How frequently does a student use marijuana?"

Target: Recoded MRJMDAYS into 3 classes.

Features: Chosen from home environment and peer influence domains only.

Models Used:

- Decision Tree
- Random Forest
- GBM (multinomial distribution)

Class Imbalance Handling:

- We applied **downsampling** to ensure equal representation across the three marijuana usage categories.
- Also tried **restricted model complexity** to reduce overfitting.
- Explored class weight adjustments but finally stuck with sampling.

Regression – "How many days per year does a student consume alcohol?"

Target: ALCYDAYS

Features: Parental behavior, gender, race, income, education.

Models Used:

- Decision Tree
- Random Forest
- GBM (Gaussian loss function)

Hyperparameter Tuning:

- For RF: ntree = 350, mtry = 3
- For GBM: n.trees = 3000, shrinkage = 0.01, cv.folds = 5

Model Evaluation & Validation

Each model was trained using a **70-30 train-test split**. For evaluation:

Classification Models:

- Confusion matrix
- Accuracy, Precision, Recall, F1-Score
- Class-specific sensitivity and specificity

Regression Models:

- Mean Squared Error (MSE)
- Mean Absolute Error (MAE)
- Root Mean Squared Error (RMSE)

We also generated **variable importance plots** to interpret which features most influenced each model.

Key Refinements and Challenges

- To prevent **data leakage**, we carefully excluded any predictors directly related to the target.
- **Downsampling** was preferred over weighting due to skewed class distributions in multi-class classification.
- **Threshold tuning** in bagging and GBM significantly improved recall in the binary classification task.
- We encountered **overfitting** in early stages, which was handled by limiting model complexity and adjusting sampling.
- We monitored the **OOB error vs. ntree plots** to choose the optimal number of trees for ensemble methods.