PREDICTING YOUTH DRUG USE

USING NSDUH DATA TO UNDERSTAND AND PREDICT YOUTH BEHAVIOR

By Ankit Bisleri

INTRODUCTION

Goal: Predict marijuana use and behavior patterns among youth.

- Tasks:
    - Binary Classification: Predict ever used marijuana (MRJFLAG).
    - Multiclass Classification: Predict usage frequency (MRJYDAYS).
    - Regression: Predict age of first marijuana use (IRMJAGE).

- Dataset: NSDUH 2020 Youth Data (30,000+ records).

- Challenges: Handling special missing codes (991, 993), severe class imbalance.

THEORETICAL BACKGROUND

- Decision Trees: Predict outcomes by splitting features into decision rules; easy to interpret but prone to overfitting.

- Random Forests: Ensemble of decision trees; reduces overfitting by averaging many trees; improves stability and accuracy.

- Gradient Boosting: Builds trees sequentially to correct prior mistakes; uses shrinkage (learning rate) to avoid overfitting; typically achieves higher accuracy but needs careful tuning.

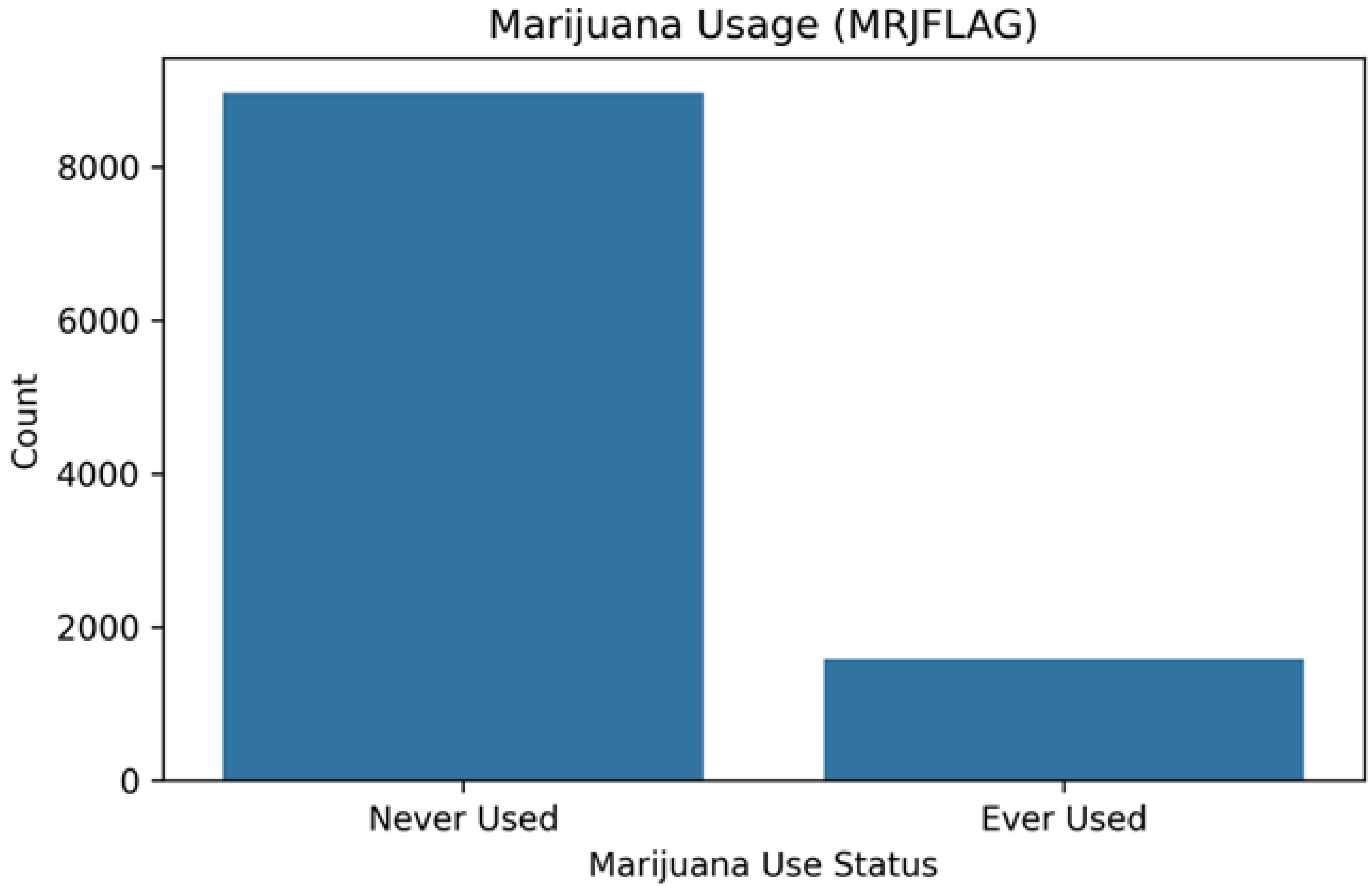
- SMOTE (for Multiclass Only):Synthetic Minority Oversampling Technique used to balance severely imbalanced classes.

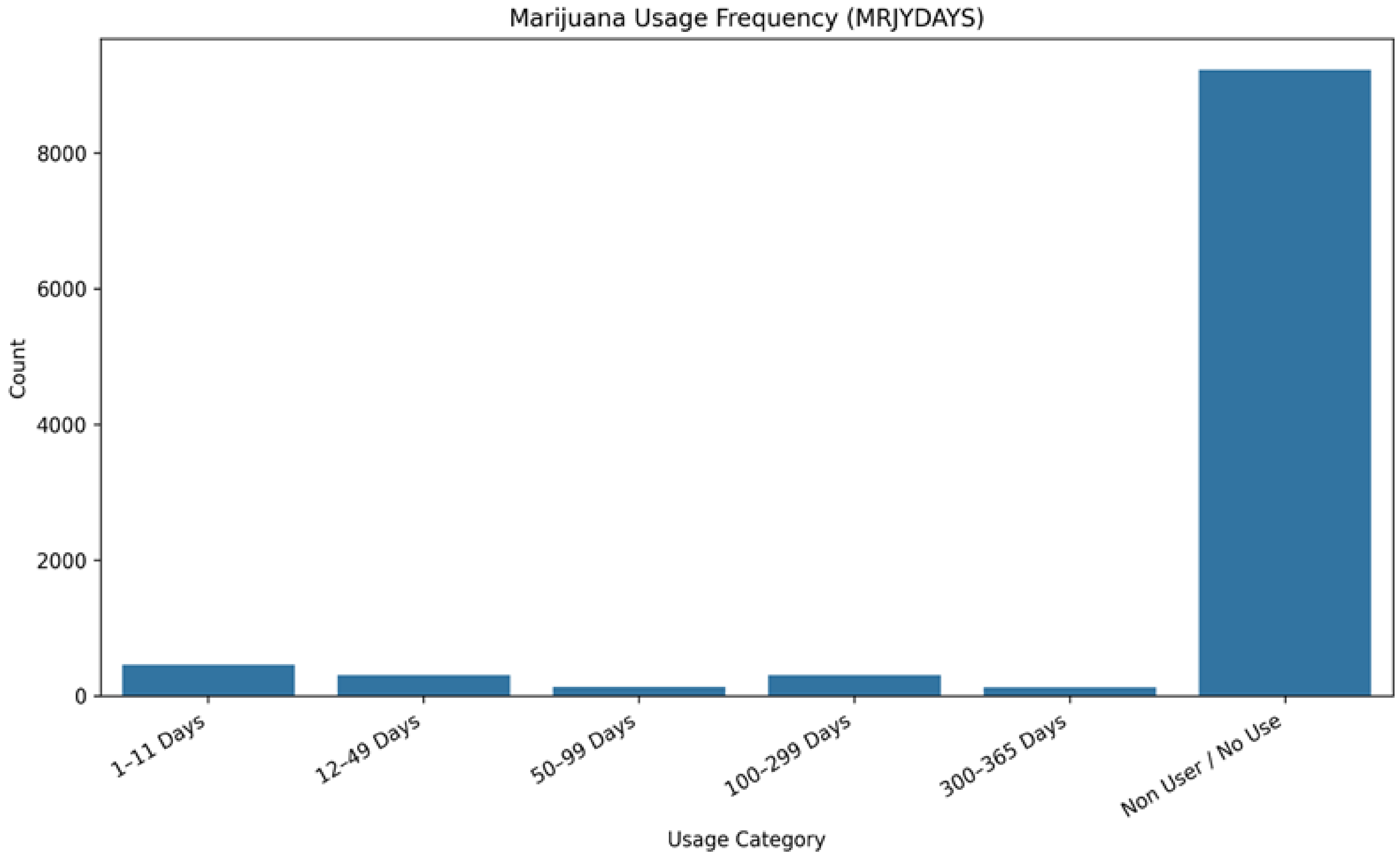
- Bagging (Bootstrap Aggregating):Trains multiple models on random subsets of data and averages their predictions to reduce variance and prevent overfitting.

DATA PREPARATION

- Replaced special missing codes (991, 993, 91, 93) with appropriate values or NA.
- Converted parental presence (IMOTHER, IFATHER) to binary indicators.
- Selected key features: Demographic, School, Parental Factors.
- Applied SMOTE oversampling only for Multiclass task.
- Train/Test Split: 70% training, 30% testing.

EDA CLASSIFICATION TARGETS

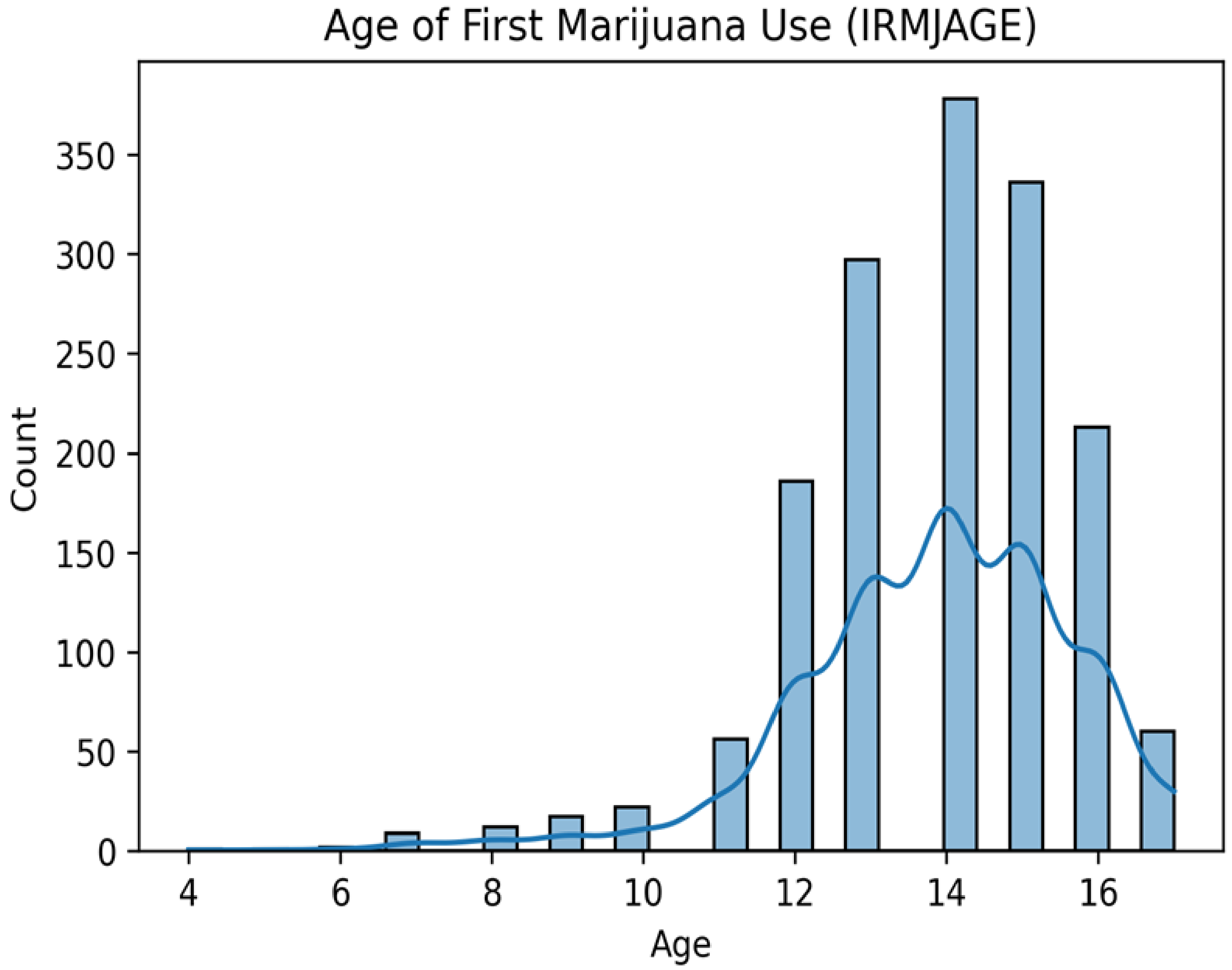




- MRJFLAG (Binary Target):
  - 85% never used; 15% used.

- MRJYDAYS(Usage Frequency in Past Year):
  - Majority report daily use; very few lower usage categories.

EDA REGRESSION TARGET

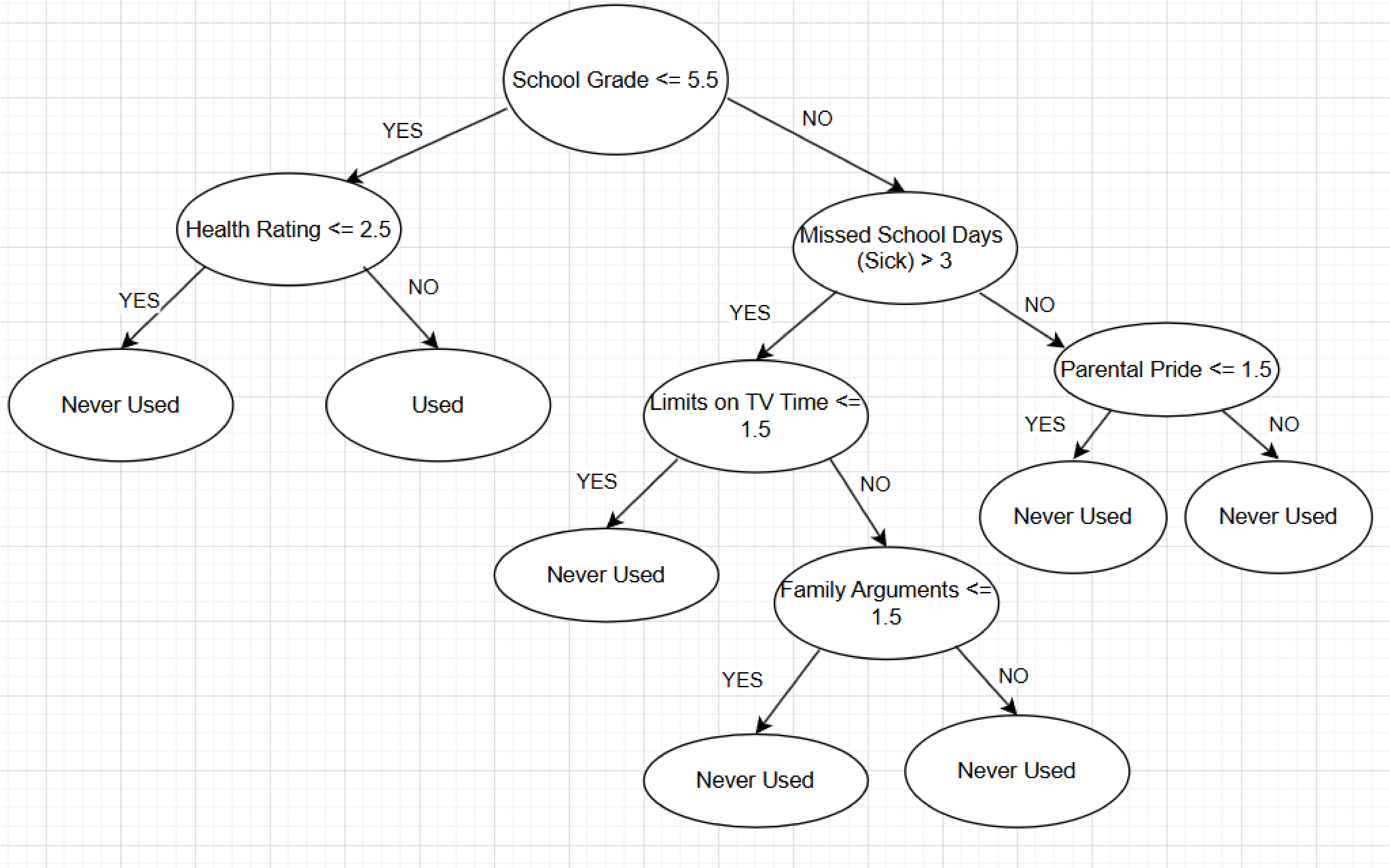


- IRMJAGE (Age at First Use):
  - Most youth first try marijuana between ages 15–17.
  - Few users start after age 25.

BINARY CLASSIFICATION: PREDICTING MARIJUANA USE

- Target: MRJFLAG (0 = Never used, 1 = Ever used).
- Features Used:
  - Demographic: Gender, Race, Family Income, Health.
  - School Factors: School Attendance, Grade.
  - Parental Factors: Homework Help, Chores, Parental Supervision.
- Models Trained:
  - Decision Tree Classifier
  - Random Forest Classifier
- Handling Class Imbalance:
  - Class weighting used (class_weight=balanced) in models.

PRUNED TREE



Jordan:

- School Grade = 6
- Missed School Days = 4
- TV Limits = 2
- Family Arguments = 1

Prediction:

Never Used Marijuana

BINARY CLASSIFICATION RESULTS & COMPARISON

Model	Accuracy	Precision (Class 1)	Recall (Class 1)	F1-Score (Class 1)	Comments
Unpruned Decision Tree	76.7%	0.23	0.26	0.24	Slightly Overfitting
Pruned Decision Tree	85.4%	0.00	0.00	0.00	Improved accuracy but fails for minority class
Random Forest	83.6%	0.28	0.08	0.12	Handles majority class very well
Random Forest (Balanced)	83.6%	0.28	0.08	0.12	Same as RF, reweighted classes but minimal change

- Pruning improved overall accuracy compared to Unpruned Tree.
- Random Forest provided more stable performance but struggled with minority class (MRJFLAG=1).
- Class imbalance remains a major challenge even after balancing weights.
- Important Predictors: School Grade, Missed School Days, Race/Ethnicity.

MULTICLASS CLASSIFICATION: PREDICTING USAGE FREQUENCY

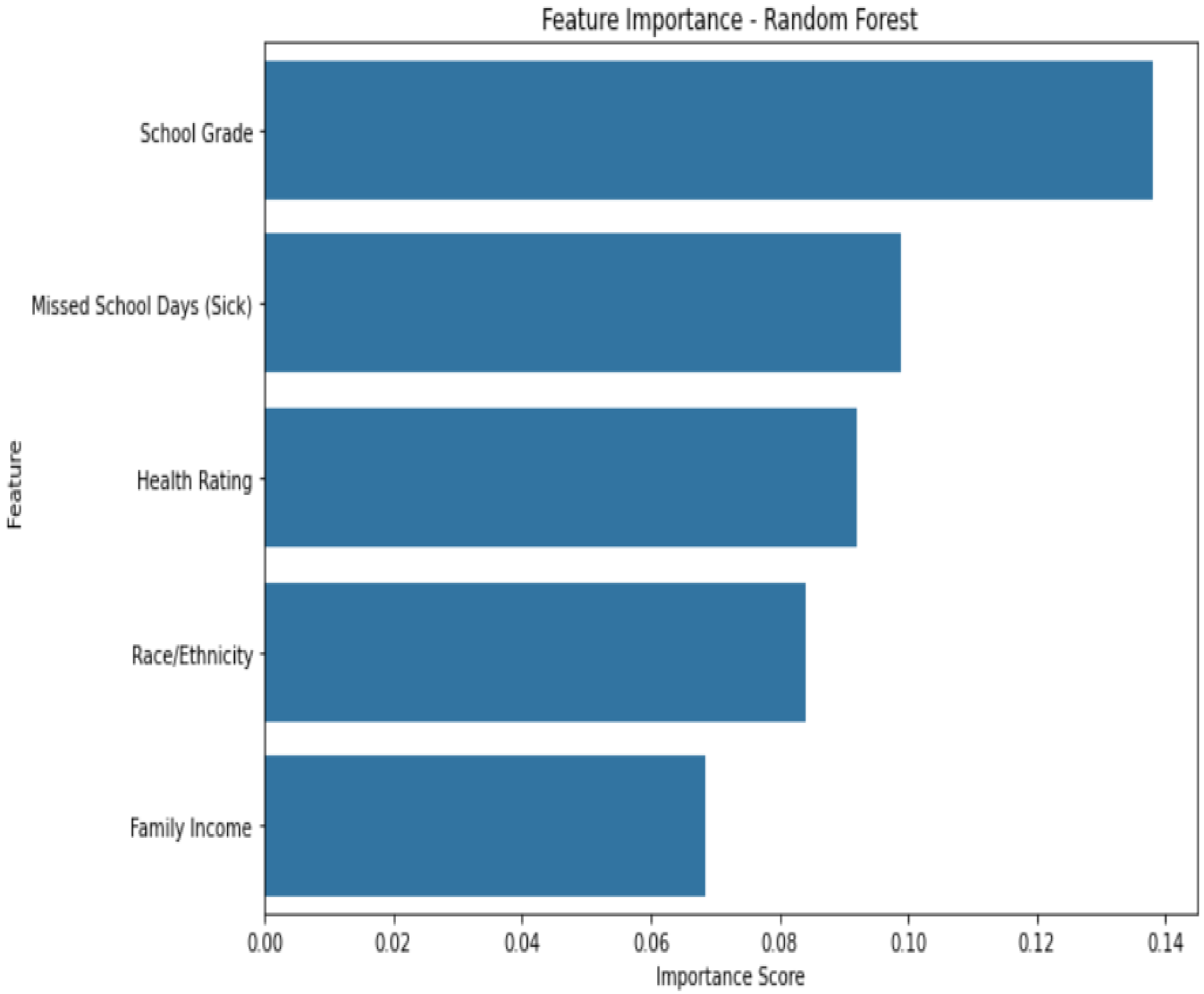
- Target: MRJYDAYS (Usage Frequency in Past Year, 6 classes).
- Classes:
  - 1 = 1–2 days
  - 2 = 3–5 days
  - 3 = 6–19 days
  - 4 = 20–39 days
  - 5 = 40–99 days
  - 6 = 100–365 days (Daily Use)
- Challenges:
  - Severe class imbalance (83% in Class 6).
- Models Trained:
  - Decision Tree and Random Forest Classifier
- SMOTE:
  - Applied to training data to balance classes.

MULTICLASS CLASSIFICATION RESULTS & COMPARISON

Model	Accuracy	Macro Avg. Precision	Macro Avg. Recall	Comments
Unpruned Decision Tree	79.99%	0.21	0.23	Bias toward majority class; poor minority class recall.
Pruned Decision Tree	44.10%	0.20	0.25	Simplified model; slight improvement for rare classes.
Random Forest	87.30%	0.16	0.17	High overall accuracy but ignores minority classes.
Random Forest (Balanced)	86.80%	0.19	0.17	Minor recall improvement for rare classes; slight accuracy drop.

- Without SMOTE, Random Forest achieved very high accuracy (~87%) but was dominated by Class 6.
- After applying SMOTE, minority class recall improved slightly, though overall accuracy dropped very slightly (~0.5%).
- Decision Tree models showed very poor performance for minority classes without resampling.

MULTICLASS CLASSIFICATION: PREDICTING USAGE FREQUENCY



- School Grade was the most important predictor of marijuana use.
- Missed School Days (Sick) and Health Rating also strongly influenced the prediction.
- Race/Ethnicity and Family Income had moderate impact.
- Academic performance and health-related behavior were more predictive than family income or demographics alone.

REGRESSION TASK: PREDICTING AGE OF FIRST MARIJUANA USE

- Target :IRMJAGE (Age of First Marijuana Use)

- Models Trained:
  - Decision Tree Regressor
  - Random Forest Regressor
  - Gradient Boosting Regressor

- Data Preparation:
  - Removed non-users (IRMJAGE=991)

- Imputed missing demographic, school, parental features

- Goal: Predict age of first use accurately with minimum RMSE.

REGRESSION RESULTS & COMPARISON

Model	RMSE	MAE	Comments
Decision Tree Regressor	2.04	1.50	High error, overfitting likely, poor generalization.
Random Forest Regressor (Default)	1.53	1.14	Better than Decision Tree, but not tuned yet.
Random Forest Regressor (Tuned)	1.53	1.13	Slight improvement with tuning (max_depth=10, max_features='sqrt').
Gradient Boosting Regressor (Tuned)	1.47	1.10	Best performance; benefited from learning rate 0.01 (shrinkage)

- Gradient Boosting achieved the lowest RMSE (~1.47 years), meaning most accurate age prediction.
- Tuned Random Forest performed better than default Random Forest.
- Decision Tree Regressor had the highest error, confirming instability without ensemble methods.
- Hyperparameter tuning and shrinkage were critical for improving model performance.

ETHICAL CONSIDERATIONS

- Bias Awareness:
  - Predictive models may reflect and reinforce existing societal biases present in survey data.
- Youth Sensitivity:
  - Predicting substance use among youth requires extra caution to avoid stigmatization or misuse of predictions.
- Fairness Across Groups:
  - Models showed imbalance issues — heavier users were easier to predict than rare users, suggesting fairness challenges.
- Data Privacy and Integrity:
  - NSDUH data was anonymized; no personally identifiable information was used or exposed.

DISCUSSIONS

- Binary Classification:
  - Pruned Decision Tree and Random Forest achieved ~84–85% accuracy.
- Multiclass Classification:
  - Random Forest (with SMOTE) achieved ~86–87% accuracy; slight improvement in minority class recall.
- Regression Prediction:
  - Gradient Boosting Regressor performed best with RMSE ~1.47 years.
- Impact of Tuning:
  - Hyperparameter tuning (max_depth, max_features, shrinkage) significantly improved Random Forest and Gradient Boosting models.
- Key Learnings:
  - Handling class imbalance, avoiding overfitting, and ethical modeling practices are critical when predicting sensitive youth behaviors.

CONCLUSIONS

- Machine learning can identify at-risk youth by analyzing school, family, and demographic factors.
- Models like Random Forest and Gradient Boosting support early intervention and targeted prevention.
- These insights can guide schools, counselors, and public health programs in resource allocation.
- Predictive tools act as early warning systems — aiding professionals, not replacing them.

THANK YOU