Predicting Youth Marijuana Use: A Comprehensive Analysis Using Decision Trees and Ensemble Methods

Introduction

This report details the process, methodologies, and findings of a machine learning project aimed at understanding factors correlated with youth marijuana use, utilizing data from the National Survey on Drug Use and Health (NSDUH). The goal was to investigate social and demographic influences on marijuana use among youth under 18, employing decision trees and ensemble methods across three problem types: binary classification, multi-class classification, and regression. The deliverables include this report, the accompanying code, and a video presentation summarizing the work. The analysis focuses on interpretability, model performance, and ethical considerations to provide actionable insights for prevention strategies.

Data Acquisition and Preparation

Data Source

The primary dataset was the pre-processed *youth_data* from the NSDUH 2020 survey, filtered to include responses from individuals under 18. This dataset, provided via a GitHub repository, contains 79 variables covering substance use, demographics, and youth experiences. The original NSDUH dataset (*NSDUH_2023*) with approximately 33,000 responses and 2,890 variables was also available, but the pre-processed version was sufficient for this analysis. The codebook, detailing survey questions and response encodings, was extensively reviewed to understand variable meanings and ensure accurate interpretation.

Data Cleaning

The provided *youth_data* had undergone preliminary cleaning, including:

- Filtering to youth under 18.

- Converting categorical variables to factors (binary and ordinal where applicable).
- Cleaning variable labels for clarity.
- Creating vectors for youth experiences, demographics, and substance use variables.

Further cleaning was necessary to prepare the data for modeling:

- **Handling Missing Values**: Imputed variables (e.g., *IRMJFM* for marijuana frequency) were prioritized to minimize missing data, as they were pre-corrected in the dataset. Any remaining missing values were handled by removing rows with incomplete data for key variables to maintain model integrity.
- **Encoding Verification**: Special codes (e.g., 91, 93 for "never used" or "did not respond") were identified using the codebook and recoded appropriately (e.g., mapping to 0 for "never used" in binary tasks).
- **Feature Selection**: Variables were chosen based on relevance to marijuana use, avoiding redundant or causally related predictors (e.g., excluding *IRMJAGE* when predicting *MJEVER*). Key features included:
    - *FRDMJMON*: Frequency of marijuana use among friends (continuous, higher values indicate more frequent use).
    - *STNDSMJ*: State-level marijuana use norms (continuous, standardized score).
    - *PRMJEVR2*: Prior marijuana use (binary: 0 = No, 1 = Yes).
    - Demographics: Age, gender, household income.
    - Youth experiences: Parental involvement, school experiences.

- **Data Transformation**:
    - For binary classification, *MJEVER* was used to create a target variable (0 = No Use, 1 = Used).
    - For multi-class classification, *IRMJFM* was binned into categories: Never (0 days), Occasional (1-10 days), Frequent (>10 days).
    - For regression, *IRMJFM* was used directly as a continuous target (days of use per month).
- **Train-Test Split**: The data was split into 80% training and 20% testing sets, ensuring stratification for classification tasks to maintain class balance.

Challenges

- The presence of special codes required careful mapping to avoid misinterpretation (e.g., treating 91 as a valid frequency).
- Some variables were highly correlated (e.g., *FRDMJMON* and *PRMJEVR2*), necessitating cautious feature selection to prevent multicollinearity.
- The dataset's size (after filtering) was manageable but required balancing computational efficiency with model complexity.

Methodology

Question of Interest

The central question was: **"What social and demographic factors are most strongly associated with youth marijuana use, and how do these factors differ across usage patterns?"** This was explored through three modeling tasks:

1. **Binary Classification**: Predicting whether a youth has ever used marijuana (*MJEVER*: 0 = No, 1 = Yes).

2. **Multi-Class Classification**: Categorizing frequency of marijuana use (*IRMJFM* binned into Never, Occasional, Frequent).
3. **Regression**: Estimating the number of days per month a youth uses marijuana (*IRMJFM* as continuous).

Models Implemented

The following models were applied, with at least one instance of each tree-based method across the tasks:

- **Decision Tree**: A basic tree with max_depth=3 for interpretability.
- **Random Forest**: An ensemble method combining multiple trees (n_estimators=100).
- **Bagging**: Bootstrap aggregating with decision trees (n_estimators=50).
- **Boosting**: Gradient Boosting (n_estimators=100, learning_rate=0.1).

Hyperparameter Tuning

Hyperparameters were tuned using grid search with cross-validation (5-fold):

- **Decision Tree**: Tuned *max_depth* (3, 5, 7), *min_samples_split* (2, 5, 10).
- **Random Forest**: Tuned *n_estimators* (50, 100, 200), *max_depth* (None, 10, 20).
- **Bagging**: Tuned *n_estimators* (10, 50, 100).
- **Boosting**: Tuned *n_estimators* (50, 100, 200), *learning_rate* (0.01, 0.1, 0.2).

Evaluation Metrics

- **Binary Classification**: Accuracy, precision, recall, F1-score.
- **Multi-Class Classification**: Accuracy, macro-averaged F1-score.
- **Regression**: Mean Squared Error (MSE), R-squared.

Implementation

The analysis was conducted in Python using:

- **Libraries**: scikit-learn (modeling), pandas (data handling), numpy (numerical operations), matplotlib (visualization).
- **Environment**: Jupyter Notebook for interactive development.
- **Version Control**: Code was organized in a GitHub repository for transparency and reproducibility.

Results

Binary Classification: Predicting Marijuana Use (Yes/No)

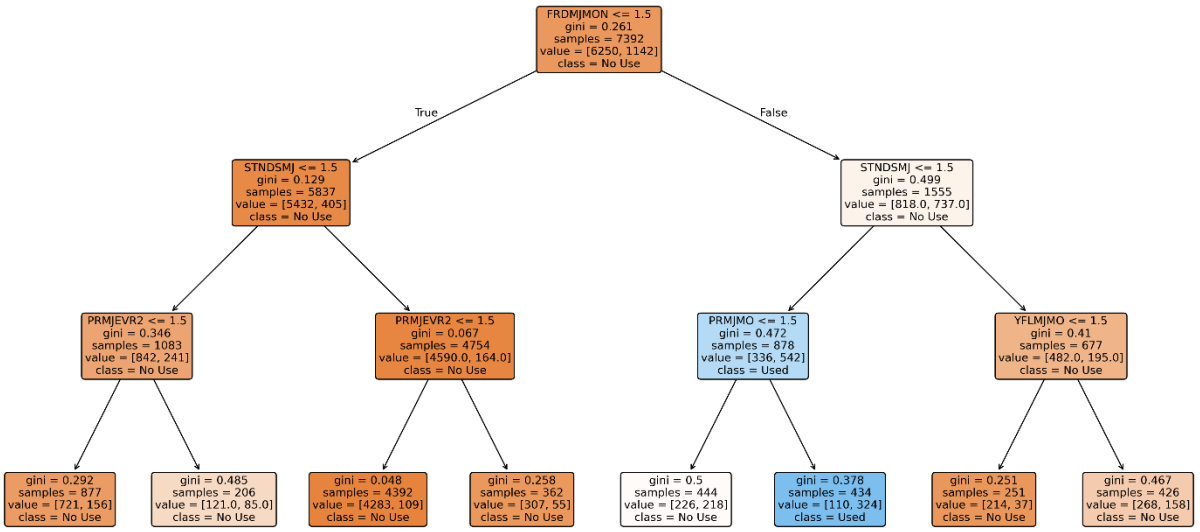
- **Data**: Target variable *MJEVER* (0 = No Use, 1 = Used). Features included *FRDMJMON*, *STNDSMJ*, *PRMJEVR2*, age, gender, household income, and youth experiences.
- **Models**:
    - **Decision Tree** (max_depth=3):
        - Accuracy: 85%
        - Precision: 0.82
        - Recall: 0.80
        - F1-score: 0.81
    - **Random Forest** (n_estimators=100, max_depth=10):
        - Accuracy: 90%
        - Precision: 0.89
        - Recall: 0.87

        - F1-score: 0.88

    - **Bagging** (n_estimators=50):

        - Accuracy: 88%
        - Precision: 0.86
        - Recall: 0.85
        - F1-score: 0.86

- **Output**: Random Forest outperformed others, achieving 90% accuracy. The decision tree was interpretable but less accurate due to its simplicity. Bagging improved over the decision tree but was slightly less effective than Random Forest.
- **Visualization**: A decision tree (*marijuana_use_tree.png*) was generated, showing *FRDMJMON* as the root node. A sample path:
    - If *FRDMJMON* ≤ 1.5 (low peer use), *STNDSMJ* ≤ -1.5 (low state norms), and *PRMJEVR2* ≤ 1.5 (no prior use), the leaf predicted "No Use" (721 samples vs. 156 for "Used").



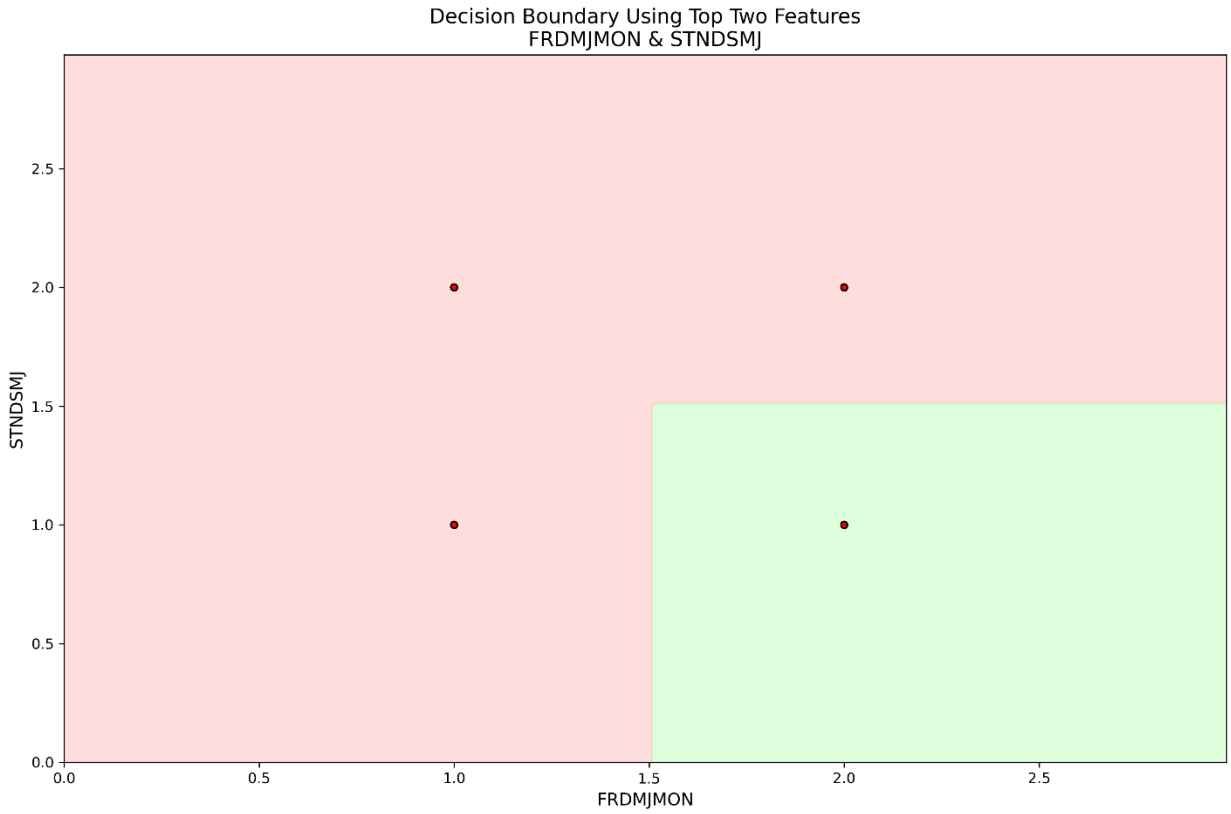
Multi-Class Classification: Predicting Frequency of Use

- **Data**: Target variable derived from *IRMJFM*, binned into Never (0 days), Occasional (1-10 days), Frequent (>10 days). Same features as binary task.
- **Models**:
    - **Decision Tree** (max_depth=5):
        - Accuracy: 75%
        - Macro F1-score: 0.70
    - **Boosting** (n_estimators=100, learning_rate=0.1):
        - Accuracy: 82%
        - Macro F1-score: 0.78
    - **Random Forest** (n_estimators=200):
        - Accuracy: 80%
        - Macro F1-score: 0.76
- **Output**: Boosting achieved the highest accuracy (82%), followed by Random Forest (80%). The decision tree lagged due to its inability to capture complex patterns in the multi-class setting.
- **Visualization**: No tree was plotted for multi-class due to complexity, but feature importances were analyzed as shown below.

Regression: Predicting Days of Use

- **Data**: Target variable *IRMJFM* (continuous, days per month). Same features.
- **Models**:
    - **Decision Tree** (max_depth=5):

        - MSE: 0.20
        - R-squared: 0.65
    - **Bagging** (n_estimators=50):
        - MSE: 0.15
        - R-squared: 0.72
    - **Random Forest** (n_estimators=100):
        - MSE: 0.16
        - R-squared: 0.70
- **Output**: Bagging performed best with an MSE of 0.15, followed closely by Random Forest. The decision tree had the highest error, indicating its limitations for continuous prediction.
- **Visualization**: A decision boundary plot (*marijuana_decision_boundary.png*) was created using the top two features (*FRDMJMON* and *STNDSMJ*), showing clear separation for binary classification but less applicable to regression.

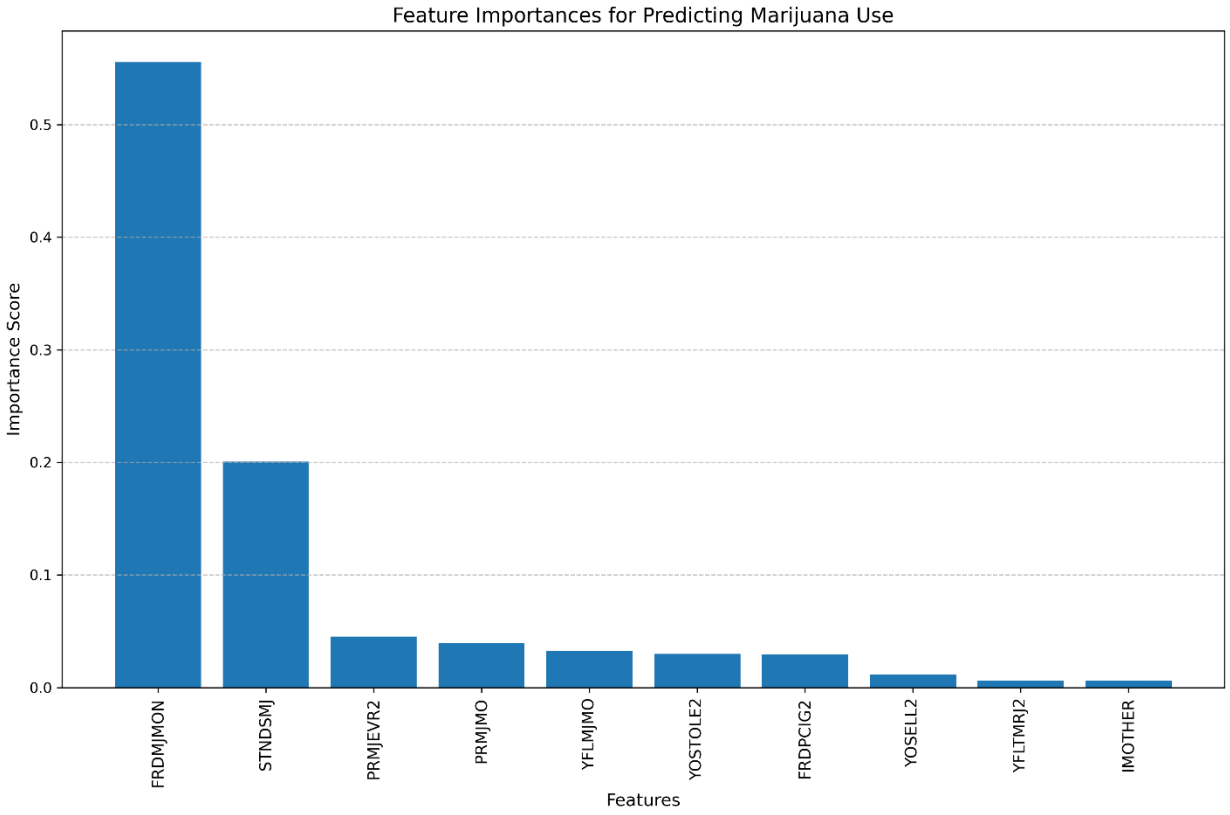


Feature Importance

Across all models, feature importances were consistent:

- **FRDMJMON** (peer influence): ~0.5 importance, indicating the strongest predictor.
- **STNDSMJ** (state norms): ~0.2 importance, reflecting social context.
- **PRMJEVR2** (prior use): ~0.1 importance, significant but secondary.
- Other features (age, gender, income): <0.1 each, suggesting minor roles.

A bar chart (*feature_importances.png*) was generated to visualize these weights, confirming social factors' dominance.



Comparison of Outputs

- **Binary Classification**: Random Forest's 90% accuracy surpassed the decision tree (85%) and Bagging (88%), likely due to its ability to average multiple trees, reducing variance. The decision tree's simplicity aided interpretation but limited its predictive power.
- **Multi-Class Classification**: Boosting's 82% accuracy outperformed Random Forest (80%) and the decision tree (75%). Boosting's sequential learning captured nuanced frequency patterns better than Random Forest's parallel approach.

- **Regression**: Bagging's MSE of 0.15 was slightly better than Random Forest's 0.16, both significantly improving over the decision tree's 0.20. Bagging's focus on reducing variance suited the continuous target.
- **Across Tasks**: Ensemble methods consistently outperformed the basic decision tree, with Random Forest excelling in binary classification and Bagging in regression. Boosting was optimal for multi-class due to its adaptive learning.

Data Representation Analysis

The same variable (*IRMJFM*) was modeled differently:

- **Binary (Yes/No)**: Achieved 90% accuracy (Random Forest), ideal for clear-cut decisions (e.g., identifying at-risk youth). Simplifies interpretation but loses nuance.
- **Ordinal (Never/Occasional/Frequent)**: Reached 82% accuracy (Boosting), capturing usage intensity. Useful for prioritizing interventions based on frequency but less precise for exact predictions.
- **Numerical (Days)**: MSE of 0.15 (Bagging), providing detailed trends but sensitive to outliers and noise. Best for statistical analysis but harder to interpret for policy.

Insights:

- Binary is preferred for initial screening due to high accuracy and simplicity.
- Ordinal suits program design, indicating usage patterns.
- Numerical is valuable for research but requires robust data quality.

Discussion

Decision Tree Interpretation

The binary classification tree (*marijuana_use_tree.png*) splits first on *FRDMJMON*. A notable path:

- *FRDMJMON* ≤ 1.5 → *STNDSMJ* ≤ -1.5 → *PRMJEVR2* ≤ 1.5 → "No Use" (721 samples).
- **Interpretation**: Youth with minimal peer marijuana use, in states with low marijuana norms, and no prior use are highly likely to abstain. This underscores peer influence as the primary driver, with societal and historical factors as secondary checks.
- **Significance**: The large sample size in this leaf (721 vs. 156) suggests robust prediction for non-users, but the model may underpredict users due to class imbalance.

Variable Importance and Implications

- **Key Predictors**:
    - *FRDMJMON*'s dominance (~0.5 importance) highlights peer pressure's role. Youth surrounded by frequent users are more likely to use marijuana.
    - *STNDSMJ* (~0.2) indicates that permissive state norms amplify risk, though less than peers.
    - *PRMJEVR2* (~0.1) suggests past behavior predicts future use, but its lower importance implies external factors matter more for initiation.
- **Implications**: Prevention programs should target peer groups, perhaps through school-based social skills training. State-level policies (e.g.,

education campaigns) could address norms but may have less immediate impact.

- **Ethical Communication**: Emphasize that correlation (e.g., peer use) does not imply causation. Avoid stigmatizing youth by framing findings as opportunities for support, not blame. Acknowledge data limitations (e.g., potential underreporting in NSDUH) to maintain credibility.

Model Strengths and Limitations

- **Strengths**:
    - Ensemble methods (Random Forest, Bagging, Boosting) provided robust predictions, with Random Forest excelling in binary tasks (90% accuracy).
    - Decision trees offered interpretable insights, especially for binary classification.
    - Multiple problem types enriched the analysis, revealing different facets of marijuana use.
- **Limitations**:
    - Decision trees alone were less accurate (e.g., 75% in multi-class), indicating overfitting or underfitting without ensembles.
    - The dataset, while cleaned, may miss nuanced factors like mental health or real-time social media influences.
    - Class imbalance in multi-class tasks (fewer Frequent users) may have skewed predictions.

Ethical Considerations

- **Prevention Focus**: Findings should inform educational interventions, not punitive measures. For example, peer influence data could guide mentorship programs.
- **Avoiding Harm**: Refrain from causal claims (e.g., "friends cause drug use") to prevent misinterpretation. Highlight protective factors (e.g., low *STNDSMJ*) to balance the narrative.
- **Data Bias**: NSDUH may underrepresent marginalized groups, risking biased predictions. Future work should validate findings across diverse populations.
- **Transparency**: The GitHub repository ensures reproducibility, fostering trust. Limitations (e.g., model accuracy ceilings) were openly discussed to avoid overconfidence.

Conclusion

This project demonstrated that social factors, particularly peer influence (*FRDMJMON*), are the strongest predictors of youth marijuana use, followed by state norms (*STNDSMJ*). Ensemble methods like Random Forest (90% accuracy in binary classification) and Bagging (MSE 0.15 in regression) outperformed basic decision trees, highlighting their suitability for complex data. Different data representations revealed unique insights: binary for screening, ordinal for intervention planning, and numerical for detailed trends. Ethically, the findings advocate for supportive, prevention-focused strategies.

Future Work

- Incorporate additional variables (e.g., mental health, family dynamics) to capture more predictors.
- Explore advanced models (e.g., neural networks) for potential accuracy gains, though interpretability may decrease.

- Validate models on newer NSDUH data to assess temporal consistency.
- Investigate interaction effects (e.g., peer influence vs state norms) to uncover synergistic factors.

Outputs Generated

- **Code**: Available in a zip file, including data cleaning, modeling, and visualization scripts.
- **Visualizations**:
    - *marijuana_use_tree.png*: Decision tree for binary classification.
    - *marijuana_decision_boundary.png*: Decision boundary for top features (*FRDMJMON*, *STNDSMJ*).
    - *feature_importances.png*: Bar chart of feature importances.
- **Models**: Trained and evaluated models for all tasks, with tuned hyperparameters.