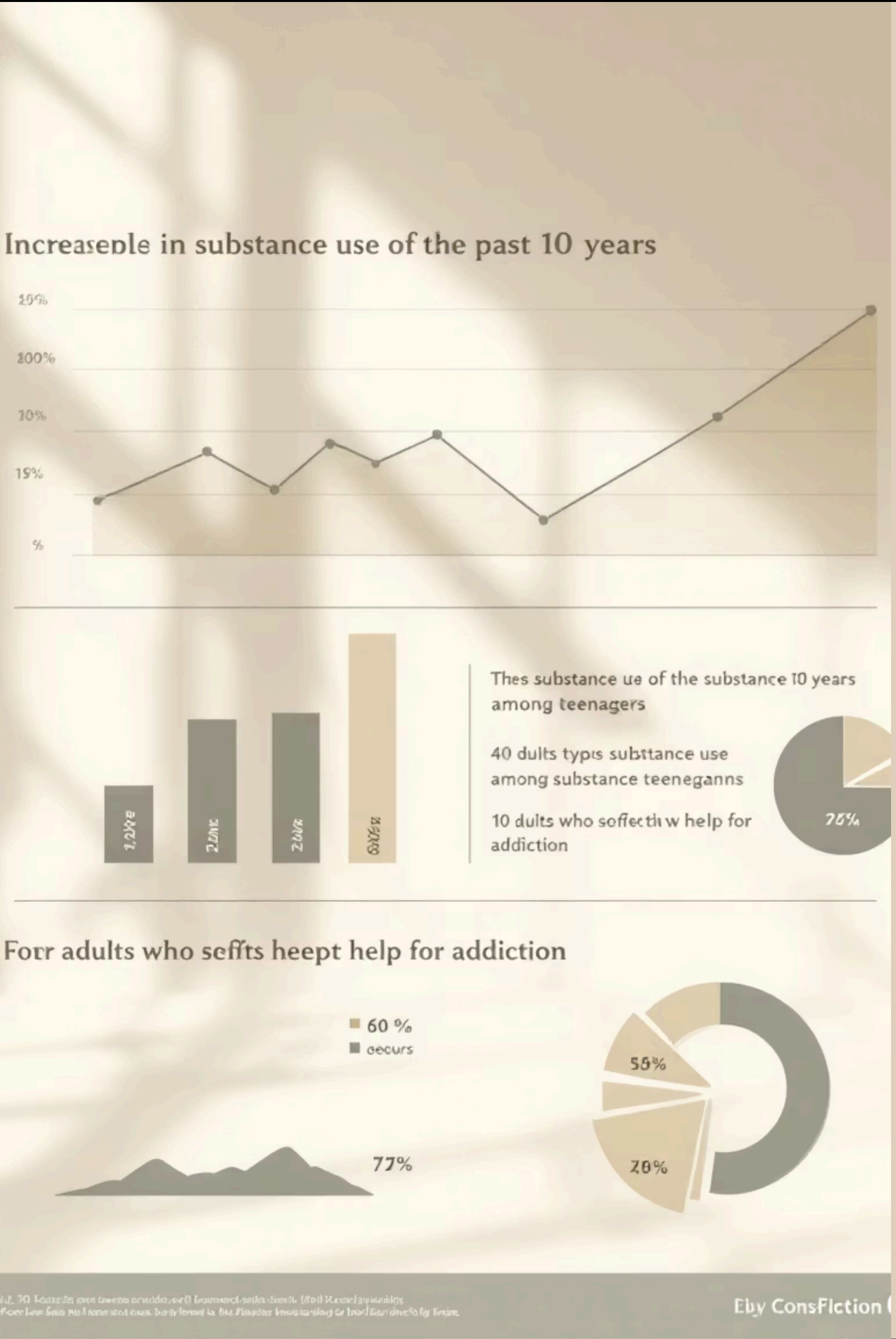
Understanding Youth Substance Use: A Data-Driven Investigation

Machine Learning Analysis of NSDUH Survey Data

 **by Khaja Moinuddin Mohammed**



Introduction to the Study

Study Goals

Predictions focus on binary cigarette use, marijuana usage frequency across multiple classes, and estimating the age of first cigarette use with regression.

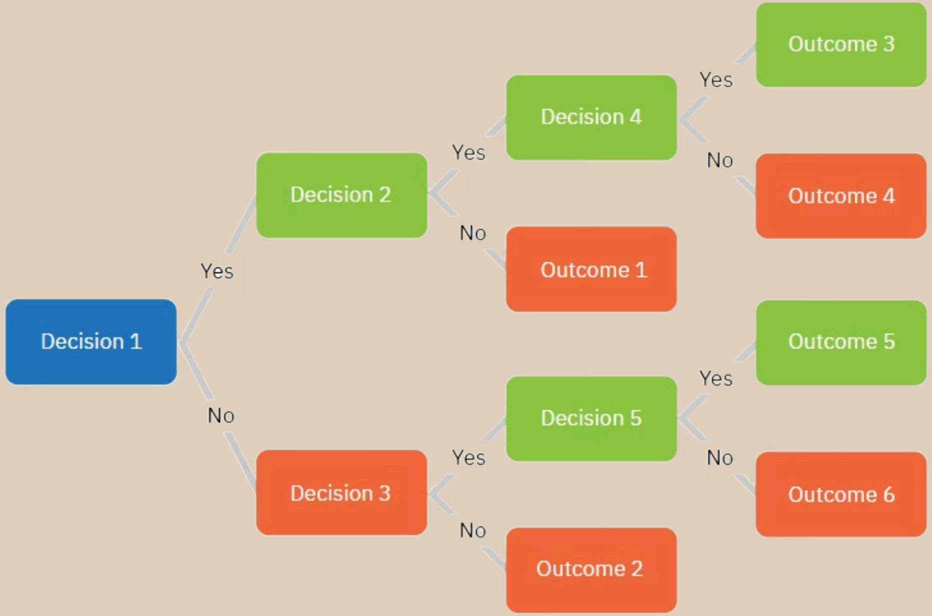
Dataset Overview

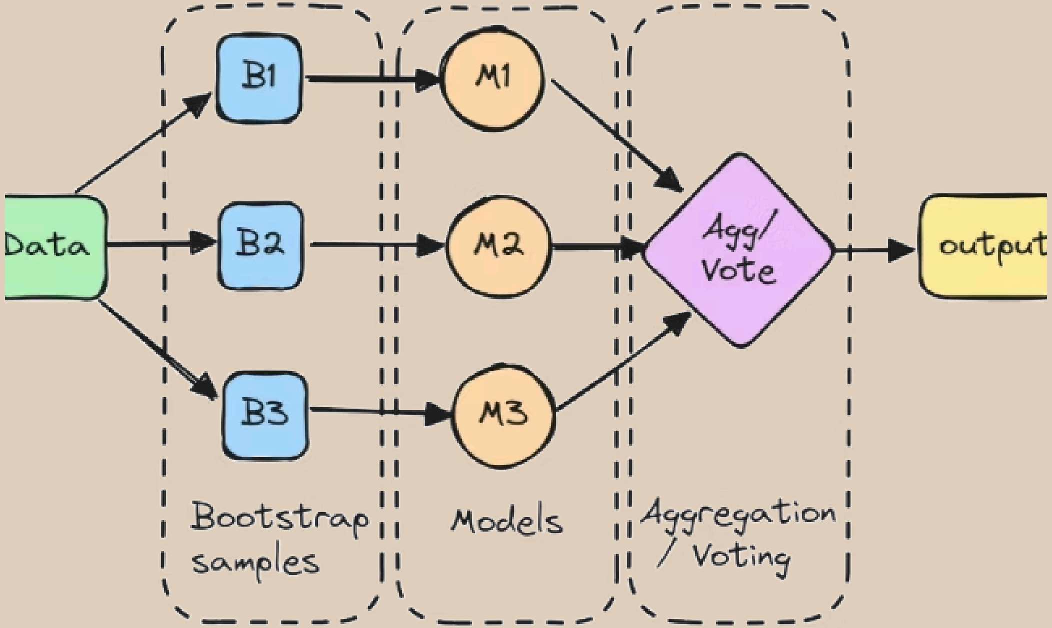
NSDUH provides over 32,900 youth records with more than 200 variables covering substance use, demographic information, and social environment contexts.

Machine Learning Models

We utilized Decision Tree, Bagging, Random Forest, and Gradient Boosting algorithms to analyze complex patterns in the data.

Theoretical Background I: Tree-Based Models



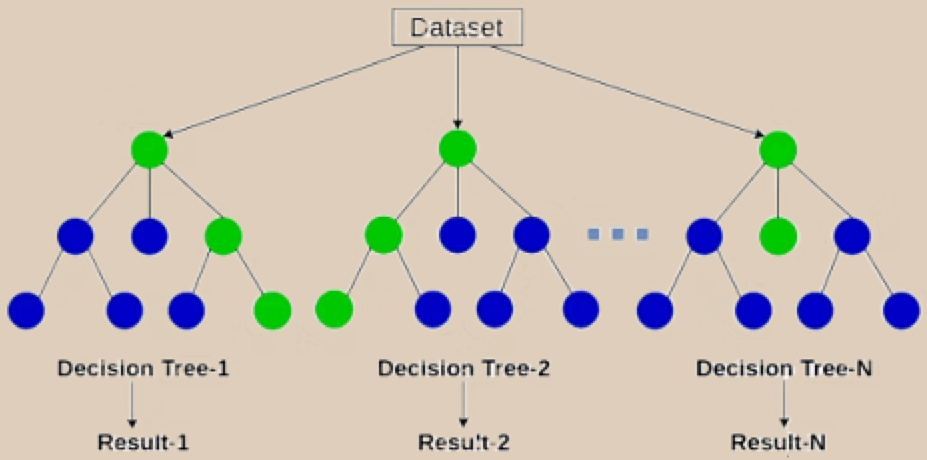


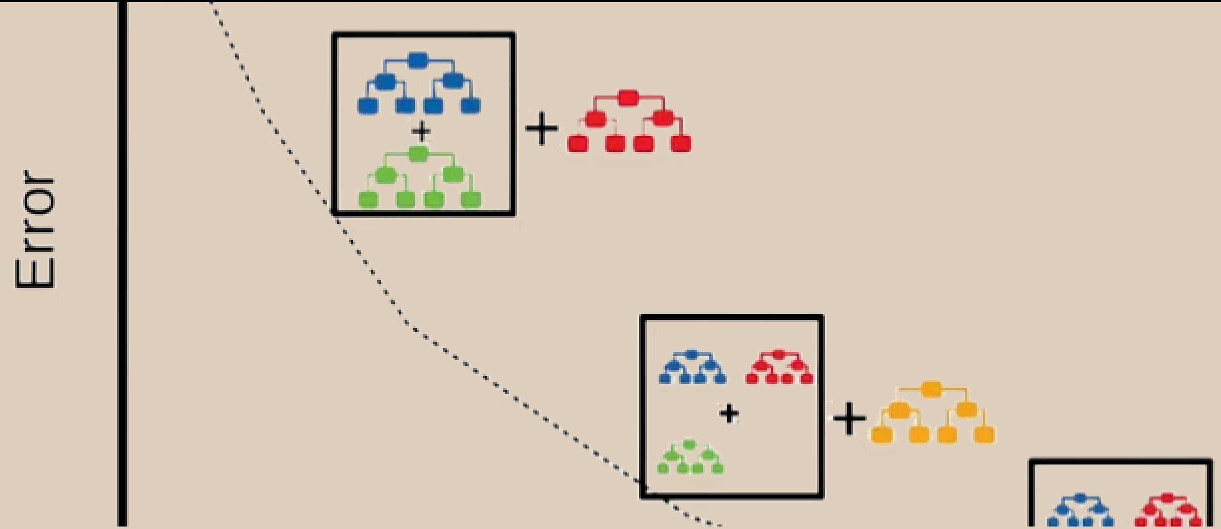
Decision Trees

Models that split datasets by feature thresholds to build an interpretable tree structure but prone to overfitting without pruning or depth limits.

Bagging

An ensemble method training many trees on bootstrapped samples, aggregating outputs to reduce variance and enhance generalization with out-of-bag validation.





Theoretical Background II: Ensembles & Evaluation

Random Forest

Introduces feature randomness during bagging, enhancing robustness and enabling feature importance assessment.

Gradient Boosting

Sequentially fits trees to residuals, achieving high accuracy but sacrificing interpretability.

Assessment Metrics

- Classification: Accuracy, Precision, Recall, F1 Score, ROC AUC
- Regression: Root Mean Squared Error (RMSE), R^2 Score

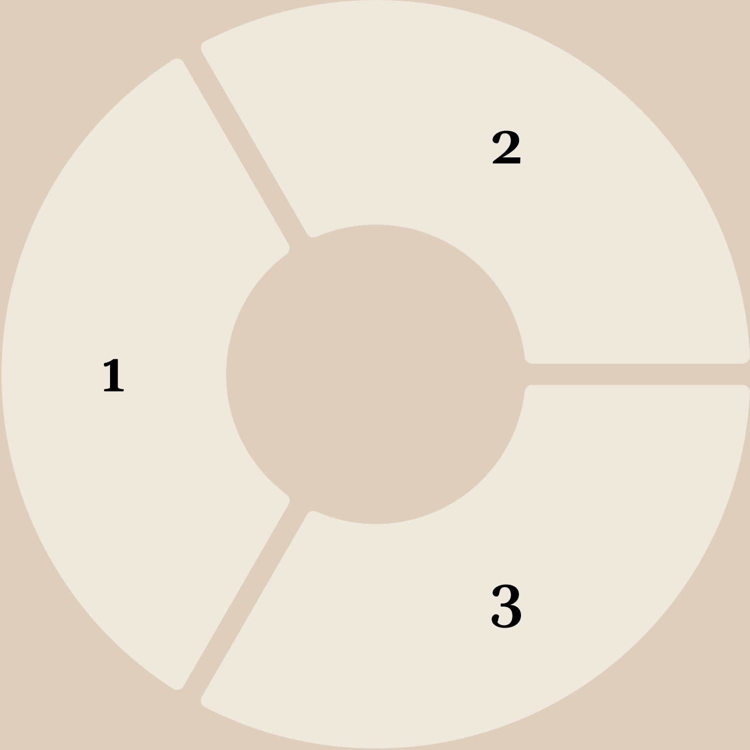
Limitations

Complex ensembles obscure individual decision pathways; challenges include imbalanced classes and survey sample biases that affect prediction validity.

Methodology I: Data & Feature Engineering

Data Cleaning

Special survey codes indicating missing or non-applicable values were replaced with NaN and rows with missing data were dropped for data integrity.



Feature Engineering

Constructed interaction terms (e.g., School_Parental_Interaction), composite variables (e.g., Peer_Influence), and polynomial features (e.g., Income_Squared) to capture complex relationships.

Temporal Variables

Created features like Years_Since_First_Drink to incorporate timing effects relevant to substance use initiation.

Methodology II: Modeling & Validation

Model Tuning

Hyperparameters like max_depth, number of estimators, and learning rate were optimized through cross-validation to balance bias and variance effectively.

Handling Imbalanced Classes

Implemented SMOTE and SMOTENC synthetic over-sampling techniques to address class imbalance, improving model recall and precision.

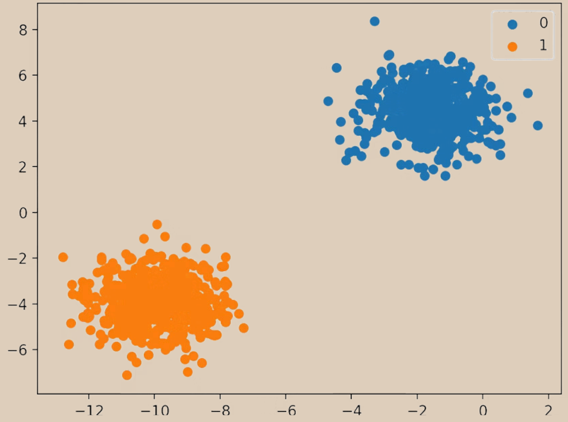
Validation Strategy

Stratified k-fold cross-validation ensured robust error estimates, maintaining class proportions across folds.

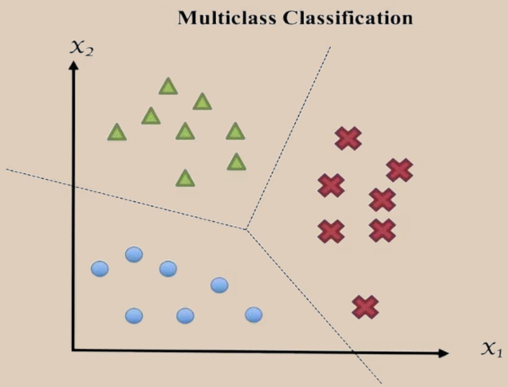
Project Organization

Structured folders for data, notebooks, scripts, and results facilitated reproducibility and workflow management.

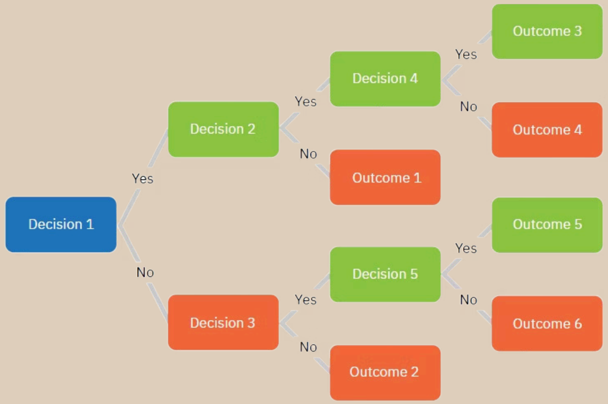
Results: Model Evaluations and Visualizations



Binary Classification



Multi-Class Classification



Decision Tree Visualization

Performance tables demonstrate superior accuracy of ensemble methods. Feature importance plots show peer influence and alcohol use timing as key predictors. Model diagnostics support robust predictive capacity with room for improvement in class imbalance contexts.

Discussion I: Model Insights and Comparative Analysis

Model Performance Comparison

Model	Accuracy	Precision	Recall	F1	ROC AUC
Decision Tree	0.694	0.122	0.520	0.198	0.645
Bagging	0.627	0.114	0.612	0.192	0.642
Random Forest	0.634	0.115	0.605	0.193	0.641
Gradient Boosting	0.642	0.117	0.599	0.195	0.644

Performance Highlights

Random Forest and Bagging excel in regression and multi-class predictions. Decision Tree offers competitive binary classification accuracy but lacks ensemble robustness.

Error Trends

All models face challenges with precision due to data imbalance. Gradient Boosting achieves highest multi-class F1 scores but sacrifices interpretability.

Explained Variance

Regression models account for up to 46% variance in age of first cigarette use, indicating moderate predictive success.

Discussion II: Decision Paths, Variable Types, and Ethical Considerations

1

Notable Tree Path

Low School_Parental_Interaction → High Income_Risk_Interaction → High Peer_Influence → Elevated probability of youth smoking (7x increase).

2

Variable Encoding Effects

Binary, ordinal, and numerical encodings influence split granularity and thresholding, enabling nuanced predictive pathways.

3

Ethical Implications

Focusing on modifiable risks such as family and peer environments promotes targeted prevention without stigmatizing vulnerable youth.

Conclusions and Future Directions

Broader Impacts

Results can guide school and community interventions focused on early alcohol use and social environment modifications to reduce youth smoking initiation.

Study Limitations

Self-report bias, class imbalance, and unmeasured confounders limit predictive performance and generalizability.

Future Work

Enhance models with longitudinal data, richer behavioral/contextual features, and integrate explainable AI methods for transparent decision support.

Thank You