Slide 1

Good morning/afternoon everyone. Today I'll be presenting our research on "Understanding Youth Substance Use: A Data-Driven Investigation." I'm Khaja Moinuddin Mohammed, and I'll walk you through how we used machine learning to analyze substance use patterns among youth using national survey data.

Slide 2: Introduction

Our study had three main goals: predicting cigarette use as a binary outcome, classifying marijuana use frequency into multiple categories, and estimating the age of first cigarette use through regression. While we explored all three approaches, I'll focus primarily on our binary classification results today.

The National Survey on Drug Use and Health provided us with over 32,900 youth records containing more than 200 variables covering substance use behaviors, demographics, and social contexts. We employed four different machine learning models: Decision Tree, Bagging, Random Forest, and Gradient Boosting.

Slide 3: Theoretical Background I

Let's start with understanding our models. Decision trees split data using feature thresholds to create an interpretable structure. The advantage is their transparency - we can easily follow the decision path. However, they're prone to overfitting without proper pruning or depth limitations.

Bagging, or Bootstrap Aggregating, addresses this weakness by training multiple trees on bootstrapped samples and aggregating their predictions. This reduces variance and improves generalization. The diagram shows how multiple samples create different models whose outputs are combined for the final prediction.

Slide 4: Theoretical Background II

Building on bagging, Random Forest introduces feature randomness during the training process. By considering only a subset of features at each split, it further enhances model robustness and provides reliable feature importance metrics.

Gradient Boosting works differently - it sequentially builds trees to correct errors from previous ones. This typically yields higher accuracy but sacrifices some interpretability.

For evaluation, we used accuracy, precision, recall, F1 scores, and ROC AUC for classification, while regression models were assessed using RMSE and $R^2$ scores. It's important to note that complex ensembles can make individual decision pathways less transparent, and challenges like imbalanced classes affected our analysis.

Slide 5: Methodology I

For data preparation, we handled special survey codes by replacing them with NaN values and dropped rows with missing data to ensure integrity.

Our feature engineering strategy was crucial. We created interaction terms like School_Parental_Interaction to capture how these factors work together. We developed composite variables such as Peer_Influence to represent social dynamics, and for the regression task, we incorporated temporal features like Years_Since_First_Drink to capture sequence effects in substance initiation.

Slide 6: Methodology II

For model optimization, we tuned hyperparameters including tree depth, estimator count, and learning rates through cross-validation to balance bias and variance.

Class imbalance was a significant challenge - only about 14% of youth reported cigarette use. We addressed this using SMOTE and SMOTENC techniques to synthetically oversample minority classes.

We employed stratified k-fold cross-validation to ensure robust error estimates while maintaining class proportions across folds. Our project organization facilitated reproducibility through structured folders for data, code, and results.

Slide 7: Results

Our results show distinctive patterns across the three prediction tasks. For binary classification, we achieved accuracy up to 69.4% with the Decision Tree model. Multi-class prediction of marijuana frequency categories reached 74.4% accuracy using Gradient Boosting. Our regression models explained up to 46.4% of variance in age of first cigarette use.

The visualizations shown here include both classification boundaries and a sample decision tree structure. These help us understand how the models partition the feature space and make decisions based on the input variables.

Slide 8: Model Performance Comparison

Let's look more closely at our binary classification results. The Decision Tree achieved the highest accuracy at 69.4% and best ROC AUC at 0.645, indicating good overall discriminative ability. Bagging provided the best recall at 61.2%, capturing the highest proportion of actual smokers.

However, all models struggled with precision, ranging from 11.4% to 12.2%, reflecting the challenge of predicting a rare outcome. This means our models had many false positives, which is often preferable to false negatives in prevention contexts.

Slide 9: Discussion I

Comparing across tasks, we found that Random Forest and Bagging excelled in regression and multi-class predictions, while the Decision Tree offered surprisingly competitive binary classification accuracy despite lacking ensemble robustness.

All models faced precision challenges due to class imbalance. Gradient Boosting achieved the highest multi-class F1 scores but sacrificed interpretability.

For regression, our models accounted for up to 46% variance in age of first cigarette use, indicating moderate predictive success - a substantial improvement over the literature baseline of just 3.3%.

Slide 10: Discussion II

Now, I'd like to walk you through one critical path in our binary classification decision tree, shown in this visualization. Starting at the root node, we see the first split occurs at School_Parental_Interaction ≤ 1.044, separating youth with very low combined school and parental support from those with higher support.

Following the left branch, those with low support are then split by Income_Risk_Interaction ≤ 3.0. Going further left, we reach youth with low Income_Level ≤ 1.5. At this leaf node, we find 1044 samples with a 44.6% probability of having smoked - which is three times higher than the baseline rate in our dataset.

This path reveals that youth with weak school/parental support systems, lower income levels, and lower risk interactions have significantly elevated smoking probabilities, even without high-risk behaviors. This insight suggests that the absence of protective factors alone substantially increases vulnerability.

Variable encoding matters too - using binary, ordinal, or numerical representations affects where splits occur and how nuanced our predictions can be. Ethically, we focused on modifiable risk factors to avoid stigmatizing vulnerable youth.

Slide 11: Conclusions

Our findings have direct implications for prevention efforts. The models identify early alcohol use as a strong predictor of early cigarette initiation, suggesting that delaying alcohol use could have downstream benefits.

Peer and family environments emerged as crucial levers for intervention, aligning with existing public health literature but providing more precise interaction patterns.

Limitations include self-report bias, class imbalance challenges, and unmeasured confounders affecting prediction. Future work could incorporate longitudinal data, richer behavioral features, and explainable AI methods for more transparent decision support.

Slide 12: Thank You

Thank you for your attention.