Slide 1: Title Slide
Hi everyone. I am Moinuddin. Today, I'll be presenting my analysis of youth substance use patterns using decision tree models. This project is titled Understanding Youth Substance Use with Decision Trees and leverages data from the National Survey on Drug Use and Health, or NSDUH.

The goal of this work is to identify key factors influencing youth substance use behaviors specifically cigarette use, marijuana use frequency, and the age of first cigarette use.

Slide 2: Introduction
Substance use among youth is a pressing public health issue with long-term consequences for individuals and society. Early intervention can prevent these consequences, but to intervene effectively, we need to understand the factors that influence these behaviors.

In this project, I focused on three predictive tasks:

1. For Binary Classification I am going to Predict whether a youth has ever smoked cigarettes.
2. For Multi-Class Classification I went with Categorize marijuana use frequency into "Never," "Seldom," "Sometimes," or "Frequent."
3. And lastly I Estimated the age at which youth first smoke cigarettes using regression.

The dataset used is the NSDUH survey, which includes over 32,000 responses from youth across the United States, covering variables like substance use behaviors, demographics, and social influences.

Slide 3: Theoretical Background
Let's briefly discuss the theoretical foundation behind our methods:

1. Decision Trees:
    - A decision tree is a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks.
    - They are interpretable but prone to overfitting without proper tuning.
    - Key hyperparameters include max_depth, min_samples_leaf, and class_weight.
2. SMOTE (Synthetic Minority Over-Sampling Technique):
    - SMOTE addresses class imbalance by synthesizing new samples for the minority class.
    - This ensures that our models don't favor majority classes at the expense of minority ones.
3. Classification vs. Regression:
    - Classification predicts categories (e.g., "Ever Used" vs. "Never Used").
    - Regression predicts continuous values (e.g., age of first cigarette use).

Slide 4: Methodology
Our methodology consisted of three main steps:

1. Data Cleaning:

- Special codes like 991 ("Never Used") were replaced with NaN values.
- Missing values were imputed using mean imputation for continuous variables.
- For regression tasks, non-users were filtered out using:

python

```
df = df[df['IRCIGAGE'] < 900]
```

2. Feature Engineering:
    - Composite features were created to capture meaningful relationships:

python (show this code as a picture when making the video for better understanding)

```
df['SCHOOL_SATISFACTION'] = df['SCHFELT'] * df['AVGGRADE']

df['PARENTAL_SUPPORT'] = df['PARHLPHW'] * df['PARCHKHW']
```

3. Model Implementation:
    - Decision trees were chosen for their interpretability.
    - Hyperparameters like max_depth and min_samples_leaf were tuned using cross-validation.
    - SMOTE was applied to balance classes in classification tasks.

Evaluation metrics included accuracy, precision, recall, F1-score for classification tasks, and MSE/R^2 for regression.

Slide 5: Binary Classification Results
For binary classification, our question was: "What factors predict whether a youth has ever smoked cigarettes?"

We trained a decision tree classifier with these parameters

(show this code as a picture when making the video for better understanding)

python

```
clf = DecisionTreeClassifier(
  max_depth=4,
  min_samples_leaf=5,
  class_weight='balanced',
  random_state=42
)
```

The resulting tree has:

- 25 nodes,
- 13 leaf nodes,
- and a maximum depth of 4.

Our metrics show:

- Accuracy: 65%
- Precision for "Ever Used": only 2%, due to class imbalance.
- Recall for "Ever Used": 52%.

Looking at the tree visualization, the most important split is on parental help (PARHLPHW). For example:

- If parental help is low (≤2.5) AND school safety is poor (≤3.5), we reach a leaf node predicting higher likelihood of cigarette use.

This highlights the importance of family support in preventing early cigarette use.

Slide 6: Multi-Class Classification Results
For multi-class classification, our question was: "How do demographic and social factors influence marijuana use frequency?"

This task involved predicting categories like "Never," "Seldom," "Sometimes," or "Frequent" marijuana use. The tree has:

- 63 nodes,
- 32 leaf nodes,
- and a maximum depth of 5.

Key predictors include:

1. Peer Influence (PEER_INFLUENCE)
2. Health Status (HEALTH2)
3. Risk Behaviors (YOGRPFT2)

One interesting path in this tree is:

- If peer influence is high AND risk-taking behavior is elevated, the model predicts "Frequent" marijuana use with high confidence.

Our metrics show:

- Accuracy: 75%
- Precision for "Seldom": 17%.
- Recall for "Seldom": 53%.

This suggests that interventions targeting peer resistance skills could be effective in reducing frequent marijuana use.

Slide 7: Regression Results
For regression, we asked: "Can we predict the age at which youth first smoke cigarettes based on family, peer, and school-related factors?"

The regression tree has:

- 53 nodes,
- 27 leaf nodes,
- and a maximum depth of 5.

Key predictors include:

1. Income Level (INCOME)
2. Peer Influence (PEER_INFLUENCE)
3. Standardized Alcohol Attitudes (STNDALC)
4. Risk-Taking Behavior (RISK_TAKING)

Our model performance metrics are:

- Mean Squared Error (MSE): ~5.8579
- Root Mean Squared Error (RMSE): ~2.4203
- R^2 Score: ~0.0328

The low R^2 indicates that predicting exact age is challenging with available features. However, income consistently predicts earlier initiation age when combined with negative peer influence.

Slide 8: Model Comparison
Let's compare the three decision trees:

Model	Nodes	Leaves	Depth	Key Predictors
Binary Classification	25	13	4	Parental Help, School Safety
Multi-Class	63	32	5	Peer Influence, Health Status
Regression	53	27	5	Income Level, Peer Influence

The binary classification tree is simpler but effective for identifying cigarette users based on family support and school safety. The multi-class tree is more complex to capture nuances in marijuana frequency categories. The regression tree balances complexity with interpretability but struggles with exact predictions due to limited features.

Slide 9: Discussion & Conclusion
In conclusion:

1. Parental help and school safety are key predictors of cigarette use.
2. Peer influence dominates marijuana frequency prediction.
3. Income level strongly predicts age of first cigarette use.

These findings suggest that interventions targeting family support systems and peer resistance skills could significantly reduce substance use among youth.

Slide 10: Thank You

Thank you for your time and attention.
I hope this analysis has provided valuable insights into the key factors influencing youth substance use, and how decision tree models can help us interpret and communicate these findings in a transparent way.

If you'd like to explore the code, data, or visualizations in more detail, please scan the QR code on this slide to access the full project repository on GitHub.
I welcome any questions, feedback, or discussion, you know where to find me.