**Slide 1**

Hello everyone, my name is Ankit Bisleri, and today I’ll be presenting my project: Predicting Youth Drug Use.

This analysis uses the NSDUH youth dataset to explore how factors like demographics, school behavior, and parental support influence the likelihood and patterns of marijuana use among adolescents.

**Slide 2**

The data comes from the National Survey on Drug Use and Health, focusing on U.S. youth between the ages of 12 and 17.  
It includes over 10,000 responses, covering substance use, family environment, school attendance, and more.

In terms of cleaning, special codes like 991 and 993, which represent “never used” or “not used in the past year,” were replaced with NaN and dropped. I also recoded the outcome variables like MRJFLAG and MRJYDAYS - to make them interpretable for modeling.

I built three kinds of models:

* Binary classification to predict if someone has ever used marijuana,
* Multi-class classification for usage frequency,
* And regression to estimate the age of first use.

My features were grouped into demographic, school-related, and parental categories.

**Slide 3**

The research questions I am addressing are related to marijuana

1. Has the youth ever used marijuana?
2. How frequently do they use it in a year?
3. At what age did they first try it?

**Slide 4**

First, I used a decision tree to predict whether a youth had ever used marijuana (MRJFLAG).  
This model achieved 99% training accuracy, with the most important features being:

* School Grade
* Missed School Days
* Race/Ethnicity
* Family Income, and
* Health Rating

These features reflect both environmental and behavioral signals tied to risk.

Now let’s look at one of the decision tree models I built for binary classification — predicting whether a youth has ever used marijuana or not.

This tree uses features like school performance, parental support, and family environment. Each node splits the population based on one question or threshold.

To explain how it works, I’ll walk you through one specific path in the tree — starting from the top and ending at a terminal leaf node.

We begin at the root:

The first split asks if the school grade is 5.5 or lower. If it’s higher — which likely means the student is older — we move to the right.

Next: The model looks at Parental Pride — basically, whether the youth feels their parent is proud of them. If the value is low, we go right again — indicating weaker parental support.

Then: It checks how often family arguments happen. If arguments are frequent, we move further right.

The next split is based on how often the student misses school due to illness. If they miss school frequently, we keep going right.

Finally:

We check if their school grade is below or above 7.5. If it's lower — so around middle school level — we land at our terminal node.

At this end node:We have 82 youth who never used marijuana, and 33 who have.

That means around 29% in this group are users — which is noticeably higher than average.

**Slide 5**

However, the original tree was overfit. So I used cost complexity pruning to simplify it.  
After pruning, the model had just 2 leaf nodes and used only 2 predictors. Accuracy dropped to 84.34%, but the tree became more interpretable.  
I experimented with deeper trees, but increasing leaf nodes beyond 12 didn’t improve performance.

In this tree, we have one split School Grade<=5.5, and we land at the terminal node

**Slide 6**

I then trained a Random Forest using the same features.  
This model also achieved around 84% accuracy, but its top predictors were different:

* Missed School Days,
* Race/Ethnicity,
* and School Grade

Overall, demographic features were most important, followed by parental and school-related factors.

**Slide 7**

For multi-class classification, I binned past-year marijuana use (MRJYDAYS) into six frequency levels.

The decision tree initially achieved 75.89% accuracy.

But after pruning to 14 leaf nodes, the model improved to 86.87% — a nearly 10% increase in performance.

Again, the top predictors were Missed School Days, School Grade, and Health Rating.

**Slide 8**

One issue I faced was class imbalance.

The model had 93% accuracy for non-users, but nearly 0% for all other classes.

I used SMOTE, a synthetic oversampling method, to improve balance. This helped the model perform better across usage levels.

**Slide 9**

With the balanced dataset, I trained a Random Forest, which achieved 85% accuracy.

The top features remained consistent:

Health Rating, Missed School Days and School Grade

**Slide 10**

Next, I used regression to predict the age of first marijuana use (IRMJAGE).  
I started with a Bagging Regressor (Random Forest ensemble) with 100 trees.  
The model had a Mean Squared Error of 2.66, and the scatter plot shows predicted vs. actual ages.  
The model captures general trends, but it struggles with precision due to the binned nature of age data.

**Slide 11**

I also tested a Gradient Boost Regressor.  
According to Partial Dependence Plots, the most influential variables were:

* Grade Level — youth in lower grades had higher predicted usage,
* and Marijuana Use Days in the Past Month, which showed only a small marginal effect.

This model had an MSE of 3.67, slightly worse than the bagging model.

**Slide 12**

Doubt

**Slide 13**

To conclude, let’s talk about the most important takeaways from the models I built.

Across all tasks — binary, multiclass, and regression — one thing stood out clearly: school performance, especially school grades, was the most consistent and impactful predictor of youth drug use. Students with lower grades were far more likely to use substances, and this pattern held across all models.

Parental support also played a key role. For example, when youth reported that their parents were proud of them or helped with homework, the likelihood of drug use was noticeably lower. On the flip side, families with frequent arguments or low involvement were associated with higher use.

We also saw some influence from health-related factors — students with poorer health or who missed school due to illness showed higher usage rates. Similarly, models picked up on the presence of a father in the household and having rules around screen time as protective features.

From the modeling perspective, decision trees helped us clearly trace paths from background characteristics to the likelihood of drug use. Multiclass models helped differentiate users based on severity or frequency, while regression models were useful for estimating age of first use.

Finally, the partial dependence plots visually confirmed these relationships, showing that as school performance improved or parental support increased, the predicted likelihood of use decreased.

So overall, what we’ve learned is that school, family, and support systems really matter when it comes to youth drug use. And interpretable models like decision trees make it easier to uncover these insights in a way that feels practical and actionable.