Slide1:
Hello, and welcome to my presentation on predicting youth marijuana use using machine learning. My name is Navin, and today I'll be sharing how we can use decision trees and ensemble methods to identify key factors influencing marijuana use among youth under 18. This project leverages data from the National Survey on Drug Use and Health to provide insights that can inform prevention strategies. Let's get started!

Slide3:
Let's start with the foundation of this project: the data and methods. I used the National Survey on Drug Use and Health, focusing on youth under 18, which provides a rich dataset on drug use behaviors. Key variables include FRDMDN, which measures how often a youth's friends use marijuana, STDSMJ, reflecting state-level norms around marijuana, and PRMJEVR2, indicating prior use. I tackled three problem types: binary classification to predict if a youth has used marijuana, multi-class to assess frequency of use, and regression to estimate exact usage. For modeling, I started with a basic decision tree with a max depth of 3, then explored ensemble methods like Random Forest, Bagging, and Boosting to improve performance. This approach allowed me to capture both simple and complex patterns in the data.

Slide4:
Here's a visual of our basic decision tree, which predicts whether a youth has used marijuana. The tree has a max depth of 3, making it interpretable while capturing key patterns. Let's trace one path: at the root, we split on FRDMDN, the frequency of marijuana use among friends. If this is low, we move to STDSMJ, which reflects state-level norms. If that's also low, we check PRMJEVR2—prior use. When all these are low, the model predicts 'No Use,' with 721 samples in this leaf compared to 156 for 'Used.' This path highlights how peer influence, captured by FRDMDN, is a critical early predictor, suggesting that social circles play a big role in youth behavior.

Slide5:
Now, let's evaluate how well our models performed across the three tasks. For binary classification—predicting Used versus No Use—the basic decision tree achieved 85% accuracy, but Random Forest improved this to 90%. In multi-class classification, where we predicted frequency of use, the decision tree scored 75% accuracy, while Boosting reached 82%. For regression, predicting exact usage, the decision tree had an MSE of 0.20, but Bagging reduced this to 0.15, indicating better predictions. The takeaway? Ensemble methods like Random Forest consistently outperformed the basic tree, thanks to their ability to combine multiple trees and reduce overfitting. This suggests ensemble methods are more reliable for real-world applications.

Slide6:
Let's dive into what drives marijuana use among youth, using our feature importance analysis. This bar chart shows the most influential predictors in our model. At the top, with an importance of about 0.5, is FRDMDN—the frequency of marijuana use among friends. This suggests peer influence is the biggest factor. Next, with an importance of 0.2, is STDSMJ, reflecting statelevel norms around marijuana use, indicating that broader social context also matters. Other factors, like PRMJEVR2—prior use—play a role but are less dominant. The key takeaway here is that social factors, especially peer influence, are the strongest predictors, highlighting the importance of a youth's immediate environment in shaping their behavior.

Slide7:
An important aspect of this project was exploring how data representation affects our predictions. I modeled marijuana use in three ways. First, as a binary variable—yes or no use—which gave us 90% accuracy with Random Forest, making it great for straightforward predictions. Second, as an ordinal variable, capturing frequency levels like Never, Occasional, or Frequent, which achieved 85% accuracy and helped us understand usage intensity. Finally, as a numerical variable—exact usage frequency—where Bagging gave an MSE of 0.15, offering detailed trends but with more noise. The insight? Binary is best for simple yes/no questions, ordinal data helps with intervention planning by showing usage patterns, and numerical data reveals detailed trends but requires careful handling due to noise.

Slide8:
As with any project involving sensitive topics, we must consider the ethical implications. My approach is to focus on prevention, not punishment—for example, using these findings to design targeted education programs for at-risk youth. However, we need to be cautious. We should avoid implying causation; just because peer influence is a strong predictor doesn't mean it 'causes' marijuana use. We also need to address potential biases in the NSDUH data, which might underrepresent certain groups, like marginalized communities. My commitment is to ensure these findings are used to support youth, not stigmatize them, by promoting understanding and empathy in how we apply these insights.

Slide9:
To wrap up, this project revealed that social factors, particularly peer influence captured by FRDMDN and state norms via STDSMJ, are the strongest predictors of youth marijuana use. Random Forest proved to be the best model, achieving 90% accuracy in classification, highlighting the power of ensemble methods. For future work, I'd like to explore additional variables like mental health or family dynamics, and test deeper trees or even neural networks to improve predictions. Thank you for watching—I hope you found this insightful! If you have any questions or feedback, I'd love to hear them.

The link of my work: https://github.com/shanav-lo/naveen