**Changes made based on the feedback received.**

* For **categorical variables**, such as marital status and region, I applied **one-hot encoding** to convert them into binary indicator variables, enabling our model to process them numerically without introducing ordinal bias.
* For **ordinal variables** like education, I **mapped code ranges to meaningful class labels** — for example, grouping values into categories like *Less than High School*, *High School Graduate*, *Some College*, *Bachelor’s*, and *Graduate Degree*. This allowed us to preserve the ordering while improving interpretability.
* For Binary Classification and multi-class classification precision score, recall are calculated and model performance is evaluated based on these metrics.
* How the analysis and classification models are helpful in the real world is described as part of conclusions slide.
* Many slides are changed to include graphical representations and tables.

**Presentation**

My name is P.Devi Sowjanya and today I’ll be presenting my project on **Youth Drug Use Analysis & Prediction**, using data from the **National Survey on Drug Use and Health (NSDUH)**.

This project is on analysis and prediction of youth drug usage.

**Introduction :**

This project focuses on analyzing youth drug use using data from the National Survey on Drug Use and Health (NSDUH). The given survey data has various factors which play important role in predicting youth drug usage. I explored three predictive tasks:

First one is binary classification to determine marijuana use, the second one multiclass classification to estimate the frequency of marijuana use over the past year, and finally regression to predict the age at which individual first consumed marijuana.

Building prediction models using decision trees is a common machine learning technique, especially when the data contains both continuous and categorical variables. In this study, we will develop predictive models for youth drug usage using decision trees, including ensemble approaches.

The dataset includes demographic, behavioral, and peer/parental context indicators.

As part of this project Evaluated different models — starting from decision trees and extending to ensemble methods — and analyze which features are most predictive of usage patterns. The goal is not just to build accurate models, but also to draw meaningful insights that can help in public health interventions and early prevention efforts.

**Background:**

In this project I have used decision tree as the modelling approach which has an ability to handle both classification & Regression tasks.

Now I will talk about decision tree and approach used to overcome the disadvantages of decision trees.

* These decision trees work by recursively splitting the dataset into subsets based on feature values.
* The splits are made based on splitting criteria such as Gini index – a measure of purity of the resulting split, entropy, classification error for classification and squared error for regression.
* This process of splitting continues until stopping criteria is reached which can be specified in terms of depth of tree, optimal tree size which are usually found using cross -validation.
* Though decision trees are easy to visualize and interpret they tend to overfit the training data especially when the tree goes too deep. To address this, I applied **pruning**, which trims the tree by removing less important branches, resulting in better generalization to unseen data.

To further improve performance and reduce overfitting, I explored **ensemble methods** like **Bagging**, **Random Forest**, and **Boosting**.

Bagging builds multiple trees on different subsets of the data and combines them to reduce variance. For classification prediction is done by maximum vote and averaging for regression.

Number of trees, maximum depth of tree are some hyperparameters considered.

Random Forest improves on bagging by also introducing randomness in feature selection, which makes the model more robust.

Boosting, on the other hand, builds trees sequentially, with each tree trying to correct the errors of the previous ones — it's especially powerful for handling complex patterns in the data.

Hyperparameters: Number of trees, Learning rate, max\_depth

Learning rate controls how much each new model contributes to the overall prediction.

**Methodology**

Starting with exploratory data analysis I found missing and null values in the data these are removed and since the column names are in the form of codes with the help of codebook columns are renamed for better understanding.

For **categorical variables**, such as marital status and region, I applied **one-hot encoding** to convert them into binary indicator variables, enabling our model to process them numerically without introducing ordinal bias.

For **ordinal variables** like education, I **mapped code ranges to meaningful class labels** — for example, grouping values into categories like *Less than High School*, *High School Graduate*, *Some College*, *Bachelor’s*, and *Graduate Degree*. This allowed us to preserve the ordering while improving interpretability.

Moving on to variable selection I focused specifically on youth-related variables —which includes **peer influence**, **parental monitoring**, **school engagement**, and **perceptions of drug harm** — which were selected based on their **correlation with the target variables and also demographic details.**

Once data cleaning and variable selection is done, The dataset was split into **80% training and 20% testing** to evaluate model generalization.

For each of the three predictive tasks — **binary classification**, **multiclass classification**, and **regression** — we applied both **basic decision trees** and **ensemble methods**.

In the **classification tasks**, model performance was assessed using **accuracy** as the primary metric. We also used **cross-validation** to identify the optimal tree size or hyperparameters like the number of trees, max depth, and learning rate for ensemble methods. This allowed us to avoid overfitting and ensure the models generalized well.

For the **regression task**, where the goal was to predict the **age of first marijuana use**, I used **mean squared error (MSE)** as our evaluation metric. Cross-validation was again applied to select the best-performing models and parameters.

Lastly, I visualized **feature importance** to understand which factors had the strongest influence on each prediction task. This interpretability helped validate the relevance of our selected variables and provided actionable insights into youth behavior patterns.

**Binary classification:**

Our problem in binary classification is to find whether the youth use marijuana or not. For that the target variable is the column which indicates marijuana ever used.

The variables selected as discuss in the previous slide are used as predictors.

I trained a **basic decision tree** first, which achieved an accuracy of **84.08%** on the test data. However, the decision tree is very complex it overfitted to the training set, which led us to apply **pruning**, resulting in a much better accuracy of **87.87%**. This shows that simplifying the tree by removing unnecessary splits helped the model generalize better.

Among ensemble methods, **Bagging** achieved **87.28%**, **Random Forest** and **Boosting** both reached **88.04%**. These ensemble models improved stability and slightly outperformed the unpruned baseline, though they took more time to train.

For all the 3 ensemble methods best parameters were found using grid search cross validation technique.

Feature importance were analysed for all the models and the above plot is for the boosting model One key insight from the plot is that **peer influence** and **perception of harm** were consistently top predictors.

**Multi – class classification**

For multiclass classification, I aimed to predict the **frequency of marijuana use in the past year**, based on the MRJYDAYS variable, grouped into 6 levels from "Rarely" to "Almost Daily" and "No Past Year Use."

Similar to the binary classification task, I used a decision tree classifier, pruned decision tree, random forest and boosting. All models were cross-validated to find the optimal parameters.

Contrary to the binary classification task, here the pruned decision tree proved to be better in terms of accuracy.

But if we look at the confusion matrix I noticed that the model was only good at prediction majority class ‘no pastyear use’. The model is very bad at predicting other classes.

This is due to the data was **extremely imbalanced**, with **"No Past Year Use" making up 87%** of the total responses. As a result, models tend to always predict that dominant class for boosting the accuracy.

From the variable importance we can say that the frequency of marijuana usage is hightly influence by peers.

**Regression**

Next moving to the regression problem the goal was to predict the age of the initial marijuana use. target feature is IRMJAGE variable which indiacate age of first marijuana use. We removed values like 991 (never used) to ensure the model focused only on users.

Decision tree regressor, and ensemble methods like random forest and boosting regressor were employed to predict the age of initial marijuana consumption and model performance is evaluated using mean squared error.

If we see the performance After pruning, MSE dropped to **2.23**, and ensemble methods further improved results. **Bagging achieved lowest MSE when compared to other ensemble methods.**

even though error metrics improved but if we see the plot of actual values vs predicted values  **showed that outputs didn’t align closely with actual values** — they didn’t fall along thedecision boundary, meaning the model didn’t predict precise ages well.

**Discussion**

Now I’m discuss some of the key finding I have observed while analysis and predicting using decision trees. the **basic decision tree model performed the worst** in every case. Their tendency to overfit was evident in both classification and regression tasks. However, applying **pruning** improved their accuracy substantially, especially in binary and multiclass classification.

We can also see from the variable importance plot features like **peer influence, parental monitoring, and school engagement** came up as the most important variables emphasizing their critical role in predicting youth drug use behavior.

One major challenge we encountered was **class imbalance in the multiclass problem**. The dominance of the “No Past Year Use” class skewed model predictions and artificially boosted accuracy.

Lastly, I observed **training time was longer** for ensemble models, but the gains in performance and robustness were well worth the cost especially in applications like public health, where predictive reliability is critical.

Now I’ll walk through how this **pruned decision tree** is used to classify whether a youth has used marijuana or not.

The tree starts at the top with the most informative feature:  
**friends\_use\_marijuana\_monthly**

**If the value is less than or equal to 1.5, meaning their friends don't regularly use, the model moves left; otherwise, it moves right.**

The next critical decision point is the youth’s **perception of harm** — captured by the feature **think\_marijuana\_is\_harmful**.  
If the youth perceives marijuana as harmful (≤ 1.5), the model is more likely to predict **non-use**;

Finally, we reach **leaf nodes**, which display:

* The **Gini impurity** (lower is better, ideally 0)
* The number of **samples**
* The **class distribution**: e.g., value = [105, 107] means 105 non-users, 107 users and the predicted class is No(non users).

**Conclusions:**

To summarize the outcome of my analysis

In **binary classification**, Gradient Boosting and Random Forest both achieved high accuracy (88.04%), showing strong performance in predicting whether youth used marijuana.

In the **multiclass setting**, the pruned decision tree slightly outperformed other models (87.47%), though the models struggled with imbalanced classes and underperformed on less frequent usage categories.

In **regression**, Random Forest again came out on top with an MSE of 1.91, but model predictions lacked precision, reinforcing the challenge of modeling subjective, personal variables like age of first use.

Through our analysis, we identified **key predictors** of health outcomes — notably, **peer influence**, **parental involvement**, and **socioeconomic status**.

These variables consistently emerged as significant across models, highlighting their role in shaping health behaviors and risk profiles.

Insights from this work can be used to **inform public health strategies**, particularly by guiding the design of **early intervention programs** that target high-risk groups based on these factors.

In terms of **future scope**, one of the most impactful improvements would be to address the **class imbalance** in the multiclass classification task. We plan to explore **resampling techniques like SMOTE** (Synthetic Minority Over-sampling Technique), which can help generate synthetic examples for minority classes and improve model sensitivity to rare usage patterns.

**References**

Here are the references for the resources I have used throughout the analysis. These include the primary dataset, machine learning libraries, and foundational literature that guided the modeling and evaluation process:

Through this analysis, we were able to identify key behavioral and social factors influencing drug use, and evaluate how different machine learning models perform across classification and regression tasks.

That concludes my presentation on youth marijuana use prediction using the NSDUH(**National Survey on Drug Use and Health (NSDUH), 2020**

) dataset.