Slide 1-2:

Hi everyone, I am Chris and for this DATA5322 assignment, I decided to analyze factors that could potentially give us insight into youth marijuana use. Specifically, the question I want to answer is: “What factor(s) are most associated with youth marijuana use?”. Is it peer, family, demographic, health, education, something else, or some combination of factors? And so, the goal of this assignment is to investigate different factors to see which are the most predictive of marijuana use in youth under 18 years of age.

Slide 3:

The dataset I used for this assignment was from the National Survey on Drug Use and Health (NSDUH) 2023 Youth Subset where the respondents were between the ages of 12-17 years. In this narrowed down subset, there were 79 different variables and around 10,000 respondents.

Slide 4:

Our original question “What factor(s) are most associated with youth marijuana use?” can be split into 3 parts where each part covers a type of problem. (1) binary classification: has the respondent ever used marijuana? (2) multi-class classification: how many days has the respondent used marijuana in the past 30 days? And (3) regression: how many days has the respondent used marijuana in the past year?

For each of these parts, I used 2 different tree-based machine learning models to try and solve each problem.

Slide 5:

Each of the models mentioned in the previous slide have their own mechanics, strengths, and weaknesses.

Decision tree splits the data into subsets based on feature that provides the most information gain at each node. Then it recursively splits again and again until we have a tree like model where each leaf represents a class label or value. Some parameters we need to control are max depth, minimum sample split, and criterion. Although decision trees are easy to interpret, if we do not limit the max depth, it might overfit.

Bagging Classifier trains multiple decision trees on different bootstrap samples of the data and averages the predictions using mean or mode. We need to control the number of trees we create, the number of features each tree uses, and whether to bootstrap samples. This model improves stability and accuracy compared to one decision tree since we end up with less variance due to the averaging. Individual trees are interpretable but trying to interpret all is unnecessary and impossible since there will be so many. Although bagging reduces variance, it does not reduce bias. It is also less prone to overfitting, but might struggle with class imbalance unless sampling is adjusted. Training is also parallelizable and faster than something like boosting since we can train trees at the same time.

Random forest extends bagging by including a random subset of features at each split, which increases the diversity among trees. And this helps reduce correlation between trees and improves generalization. Again, we need to control the number of trees, the number of features considered at each split, and the max depth since we do not want too many trees or too many features. Random forests are a bit more robust than decision trees and bagging. Random forests also struggle with class imbalance unless we provide class weights to handle it.

Gradient Boosting builds sequential trees where each new tree tries to correct the errors of the previous tree using gradient descent to optimize a loss function. We want to control the number of boosting stages, learning rate, and max depth since we do not want to build endlessly or too fast or slow. This model is sensitive to overfitting if not tuned properly and is slower to train than bagging or random forests since we have to train in order. Tuning can be complex at times.

XGboost is an optimized version of gradient boosting that incudes regularization, tree pruning, and parallel computation, which makes this fast and efficient. Like gradient boosting, we need to control the learning rate, max depth, number of estimators, but also subsamples to control the fraction of data and features used per tree. It requires careful tuning and the regularization reduces the risk of overfitting.

Slide 6:

To clean my data, I imputed missing values with mean or mode depending on what that column’s specific values were. So if a column was just a flag variable, I used mode, and if it was something like grades, I averaged. I also renamed the variables to be more readable, and further mapped ordinal labels to midpoints for the regression target variable since we did not have any “continuous” variables.

I chose 16 of the 79 variables to be my features and chose another 3 for the target variables, which totals to 19 variables used. I chose the 16 based on intuition and interest. Since we are trying to find the predictive features of youth marijuana use, things like education, peer influence, or family variables we generally what I went with.

To tune the hyperparameters, I used gridsearchcv to find the optimal subset given a set of parameters. Since we had 16 variables, I provided the set of parameters reasonably like max depth of 3, 5, 7, 10 or learning rate to be .01, .05, .1 and so forth.

The metrics that I used were accuracy, precision, recall, and f1-score for classification models. I used r squared and root mean squared error for the regression ones. And because there was a bit of class imbalance, I leaned towards using the f1-scores, which essentially just indicates how well it balance precision and recall for all classes. To check model stability, I also did cross validation on each of the models

Then to come to my conclusion, I extracted feature importances from each of the models and compared them for commonalities.

Slide 7:

For binary classification, I used decision tree and bagging. Like I mentioned before, there was a noticeable class imbalance. So even with the proper adjustments, both model did not do well when classifying users of marijuana versus non-users. This is pretty intuitive since if you have more data to learn from, then obviously you should do better at classifying them. The overall accuracy was not bad for either model. Bagging did slightly better than the decision tree here.

Slide 8:

Looking at feature importance for the two binary classification models, both placed friends use marijuana as the most important feature. Some other overlapping features were limited tv time, highest grade, and health status. It does seem like the decision tree had friends use marijuana as a significantly important feature compared to bagging classifier. Nothing super crazy. I think the results are pretty intuitive.

Slide 9:

Here is the decision tree that we got from tuning the hyperparameters. Interestingly enough, the first feature split here was limited tv time and not friends use marijuana. If we follow the path to the right most leaf node, it means the individual’s parents did not limit the amount of tv they watched, their friends used marijuana, their highest grade level was 8th grade or higher, then there is a high change that the individual has used marijuana before. This is pretty intuitive as well since if you are on the older side and have friends that also use marijuana plus have parents are not as engaged or strict (we can deduce this from the lack of limit of tv time), then you are clearly more likely to be exposed to marijuana.

Slide 10:

The results of the random forest and decision tree for multi-class results are better than the binary classification models in terms of accuracy. And although their accuracy was high because of the high accuracy of classifying the “never” class, the multi-class models did very poorly when classifying the “sometimes” and the “addict” classes. Both models surprisingly never predicted “sometimes” as the class. Again, this was due to the obvious class imbalance we had.

Slide 11:

Looking at the feature importance, both models again yielded that friends use marijuana was the most significant feature. We do see some new features pop up in contrast to the binary models’ feature importances. We see race, grades, and household income, and friends use drugs making an impact.

Slide 12:

Now let’s talk about regression. Like I mentioned earlier, I used the variable that represents the number of days the individual used marijuana in the past year. However, this variable was more an ordinal feature. It keys that paired with ranges, so 1 means the individual used any number of days between 1-11. But since we needed a more continuous feature, I decided to map each key with its range’s midpoint. So 1 would map to 6. I did this in hopes of getting an estimate to the actual number of days rather than a range of days and got these results. Both models did not so well, but they did do better than a model that simply predicts the average each time. The root mean squared error also tells us that the models were off by around 42 days, which is a lot.

Slide 13:

If we look at the feature importances, however, they say similar things to the other models we made, which means we are hopefully on the right track. Again, friends use marijuana was the most important feature in our models. We do see some new interesting features like school felt unsafe, group counseling, mother present, and private doctor visits, which indicate these models found more of a mental health pattern.

Slide 14:

So, summarizing what we found:

1. there is a clear class imbalance in our data that our models could not handle well. Despite having high accuracies for some of the models, the precision, recall, and f1 scores tell a slightly different story, that the models are accurate but only for classifying non-users. We also had roughly 10,000 respondents in the survey. This is not even remotely enough to use and start making conclusions.
2. The type of information does change the predictions and results of our models:
   1. Binary variables led the models to focus on splitting the data into two distinct groups which potentially simplified the model, which gave it more mediocre accuracy, but decent class balanced results.
   2. Ordinal variables led the models to potentially understand more nuanced patterns which led the results to be more focused on the “never” class than the others given our data
   3. Numerical variables (although used in a proxy sense) led the models to make more precise predictions given our data which led the results to be a little more varying.
3. I do think the survey could have handled the data different and include the actual number of days instead of a range since we can convert the actual days into ranges for use, but not the other way around. Further, the actual number of days might provide even more insight into trends for youth drug use.
4. Based on the results of the feature importances, friends use drugs was by far the most important. Not only that, friends offered drugs was also a contender, which means peer influence is by far the most associated with youth drug use. We do see other factors having solid impacts, such as household income, race, highest grade level, average grades, health, arguing with parents, and school felt unsafe.
5. What this tells us is that the most common aspects of our daily lives have the biggest impacts on whether someone uses marijuana, and targeted intervention should start there.
6. How we can communicate this to others is to be transparent. These predictive models are not absolute and like we saw, they can make mistakes. Models are only as good as the data they are trained on, so we still have a long way to go before we can use any model or data to accurately predict youth drug use. We do have a starting point, but we are very far from the finish line.

Also, correlation does not equal causation. Just because your friends do drugs or just because you meet some of these criterion does not mean you will do drugs. There might be more risk and exposure, but nothing is guaranteed.

It is also important to not blame the users the stigmatize them. These variables, again, are just a starting point. There is much more to consider and there could be reasons not captured by these variables as to the reason someone uses marijuana.