Slide 1-2:

Hi, 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:

How I approached our investigation was by creating 3 different types of models. Binary classification using a variable that indicates whether the respondent has ever used marijuana, multiclass classification using a variable that indicates how many days the respondent has used marijuana in the past 30 days, and regression using a variable that indicates how many days the respondent has used marijuana in the past year.

Slide 5:

These are the models that I used. Each of the models 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.

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.

Slide 6:

For data cleaning, I just dropped rows with missing and invalid values, mapped binary variables to 0/1, one hot encoded categorical variables, renamed variables, and mapped ordinal labels to midpoints as a proxy for regression since we did not have any continuous variables.

The variables I chose are listed here. I chose these variables because I thought they capture all the sentiment that I am sure we have all heard at like middle school or high school assemblies. So things like don’t do drugs because your friends do drugs. Or if you need help with addiction, reach out to someone kind of things.

I tuned hyperparameters using grid search to easily find some optimal hyperparameters.

Then used metrics like recall and f1-score to measure performance of classification models and r squared and root mean squared error for regression models.

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

Slide 7:

Here are the results for binary classification. Some notable results were the accuracy, recall, and samples. There was a significant class imbalance, which indicate that the accuracy is a bit misleading. Although 87% is decent, we want to see how the model performed in regards to true positives. So for decision tree, of the 1149 non users, our model correctly identified 96% of them, while for the 184 users, our model only correctly identified 28% which is terrible. This is pretty intuitive since, if our model has more non users, it will do better in classifying non users.

Slide 8:

Both binary models placed how close friends feel about using marijuana as the most important feature followed by highest grade completed. This just means that peer influence is the most significant factor in predicting youth marijuana use. This is also pretty intuitive since youth spend most of their time in school and around friends, they also spend a lot of time outside of school with friends. So, if friend do it, it must be cool so I want to do it and vice versa is the consensus. With highest grade completed, when you are in middle school, you are exposed to some marijuana and have access to some, in high school you have more access and exposure, and in college you have much more exposure and access, so intuitive as well. Interestingly enough, parental factors were not very important features.

Slide 9:

Similarly both of the multiclass models performed well on non users, but very poorly on users.

Some notable results was that neither model here predicted sometimes users. And another interesting result was that the random forest model predicted addict class one time and got it correct so it had precision of 100% which I thought was funny.

Slide 10:

We also see similar feature importance. Again, we see that peer influence/environment is the most influential factor. We do see health status appear, and this just indicates how healthy the individual is. But nothing too crazy or surprising.

Slide 11:

the results of our regression models did sort of poorly. Since the dataset version I used did not have a continuous variable for regression, I used a proxy variable that mapped the midpoint of an ordinal variable’s range to represent the number of days the youth used marijuana in the past year. Since the ranges were wide, the midpoints we far apart as well, so our results are not that great. Off by about 41 days, but did do better than a model that only predicts average at each point.

Slide 12:

Again, we see that peer factors are the most influential. Friends have much more of an influence on youth drug use than something like parents or even demographics. We do see some new interesting factors like mother present, which just indicates whether a mother figure is present in the household. For xgboost we also see sex, which I am kind of surprised we did not see higher up in other models. I figured males would be more prone to drug use than females but I guess not.

Slide 13:

Here is the first tree in our gradient boost regressor which tries to predict the number of days the youth used marijuana in the past year. If we follow the path to the bottom left leaf node, we can see this model indicates that how close friends feel about using marijuana was the most important factor followed by mother presence then gender were key splits. Interpret-wise, the bottom left leaf node means that if your close friends neither approve nor disapprove of marijuana use, your mother is not present, and you are female, then the expected or predicted number of days of marijuana use is 102 days, which is a lot. Considering our data did not have many addicts, this first tree might have some corrections in the next tree.

Slide 14:

To summarize our results, all models had class imbalance, so we need to look at weighted f1 score which tells us how well the model handled class imbalance. From our results it seems the multi class models did the best, but improvements should be made to account for the minority class. Also noting runtime, random forest took the longest by far and the decision trees took the least amount of time which is expected.

For regression models xgboost did the best, which is also expected since it is a better model in general. It also took far less time to train than gradient boosting since like we said, gradient boosting trains sequentially, while xgboost trained in parallel.

Slide 15:

In conclusion, peer influence/pressure/environment was the most important factor in predicting youth marijuana use. So surround yourself with responsible friends, don’t always do what your friends do.

We also saw some class imbalance and our data size was pretty small, so we still have a long way to go before we can use the data or even the models to accurately predict youth marijuana use and the factor(s) associated with it. Our result should be taken as only a starting point.

And correlation does not equal causation. Just because you tick a couple of the boxes, does not mean you will use marijuana. You might be more at risk or more exposed, but nothing is guaranteed. But like we saw, the common consensus was peer influence, so surround yourself with people you want to be around and be responsible and you will be fine.

~~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.~~