DECISION TREES:

IMPACT OF FAMILY BACKGROUND AND SOCIOECONOMIC STATUS ON MARIJUANA USE

INTRODUCTION

- Objective: Predict marijuana use among youth based on socioeconomic, familial characteristicts.
- Research Question: Does financial and family background propel individuals towards marijuana use?
- Dataset: Data has 10,561 rows and 79 variables. It has numeric and categorical data. It also has ordered and unordered variables as well.
- Models Used: Decision trees, Random Forests, Bagging, Boosting, Pruning.
- Metrics: Accuracy and Mean Squared Error.

THEORETICAL BACKGROUND

- Decision Trees:

1. A model that splits the data with respect to certain features until the final decision (terminal mode). It is one of the easiest models to interpret, especially visually.
2. Overfitting is very common. To avoid this, we prune the trees.
3. Pruning is just cutting of unneccesary branches, based on certain error rates and feature weightage.

Ensemble Methods: A set of models aimed at improving performance through collaboration.

- Bagging:

1. This method makes use of the bootstapping methods. It trains multiple models. on random samples of data.
2. All the trained models get to vote to make predictions. The votes are aggregated.
3. Affective in reducing variance, can be tuned by minimising the OOB error.

- Boosting:

1. In bagging, models are trained in parallel, whereas in boosting, models are trained sequentially.
2. In bagging, every model gets equal weight to their vote, but in boosting, models' vote have weights to them.
3. Affective in reducing bias, an be tuned by using the depth and shrinkage parameters.

- Random Forests:

1. Random Forest is very similar to bagging where in it makes a collection of decision trees.
2. Bagging makes use of all predictors, whereas you can tune the mtry parameter in Random Forest.

METHODOLOGY

- Data Cleaning and Filtering:

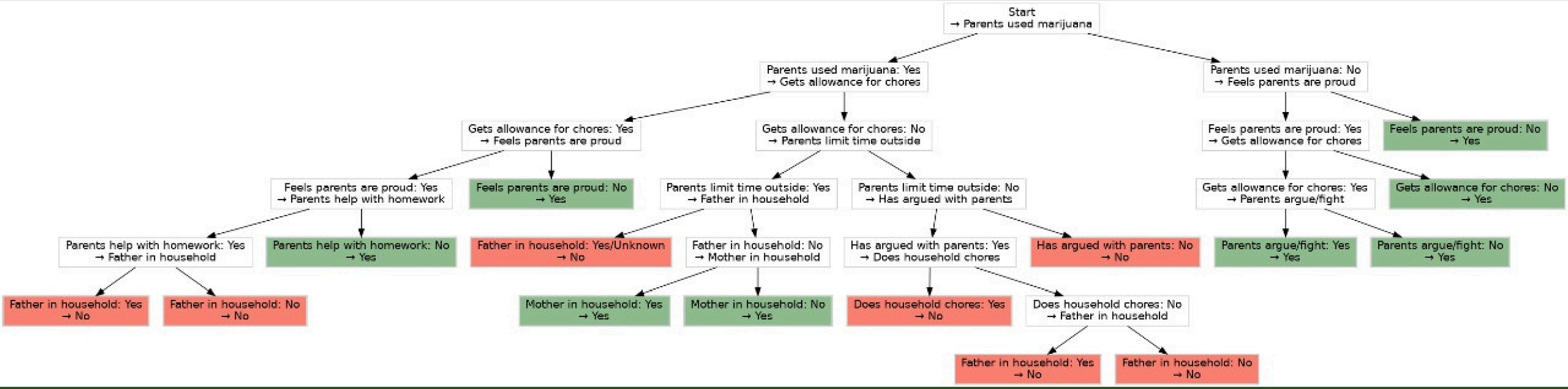
1. Converted numeric factors to labels for better interpretation.
2. Removed NA values for boosting and bagging.
3. Cleaned and filtered selected variables, prevalent to the research question.

- Model Implementation and Training

1. Split the data into training and testing in the ratio of 30:70.
2. Binary Classification - used decision trees and hyperparameter tuned it by pruning.
3. Multi-Class Classification - used decision trees and optimised it by bagging.
4. Regresion - used random Forests and optimised by boosting.

Let us look at the Binary Tree Classification in detail.

DISCUSSION



EXAMPLE (SELF-CREATED):
RICHA. (2025). VISUALIZATION OF MARIJUANA USE TRENDS [AI-GENERATED IMAGE]. CHATGPT/DALL·E.

This plot is of a pruned decision tree. The response variable is MRJFLAG. The predictors used are PARCHKHW, PARHLPHW, PRCHORE2, PRLMTTV2, PARLMTSN, PRGDJOB2, PRPROUD2, ARGUPAR, PRMJEVR2, PRMJMO, IMOTHER, IFATHER.

FLOW OF THE TREE:

- 50.9% of individuals in the dataset have used marijuana.

- First split is on parental marijuana use (PRMJEVR2)

- Factors that lower likelihood of marijuana use:

1. Parents have used marijuana (PRMJEVR2 = Yes)
2. Father present in household (IFATHER = Yes)
3. Parents help with homework (PARHLPHW = Yes)
4. Limits on TV or outdoor time (PRLMTTV2, PARLMTSN = Yes)
5. Feels parents are proud (PRPROUD2 = Yes)

- Factor that increases likelihood of use:

1. Does not feel parents are proud (PRPROUD2 = No).

FLOW OF NODE 42:

- Node 42 predicted 56.5% youth to have used marijuana, has deviance 84.92, P(No) = 43.5%, P(Yes) = 56.4%

- Conditions to reach this group:
  - Parents have used marijuana (PRMJEVR2 = Yes).
  - No allowance for chores (PRLMTTV2 = No).
  - Parents limit time outside (PARLMTSN = Yes).
  - Father not present (IFATHER = No).
  - Mother present (IMOTHER = Yes).

- Despite of present mothers, a lot of teens use marijuana.

- Likely contributing factors:
  - Exposure to parent drug use.
  - Lack of father figure in the household.
  - No allowance for chores.

- Possible intervention focus:
  - Support for single-parent homes
  - Parental substance use prevention
  - Encouraging positive behavior reinforcement

PATH TO NODE 42:

NODE 1 (ROOT)

- → PRMJEVR2 = YES → NODE 2

NODE 2

- → PRLMTTV2 = NO → NODE 5

NODE 5

- → PARLMTSN = YES → NODE 10

NODE 10

- → IFATHER = NO → NODE 21

NODE 21

- → IMOTHER = YES → NODE 42

Q. Some of these variables are the same information coded into binary, ordinal (categorical and ordered), and numerical variables. How do the predictions change using each data type? What is each telling you, and when is it appropriate to use each?

- There are two types of variables: IMOTHER (binary), INCOME(multi-class) and IRMJFY (numerical).
- Binary variables are easy to split, as opposed to categorical data with multiple classes.
- Trees, I have noticed through this assignment have a preference for binary varibales as they fit the memo of a decision tree perfectly.
- We have to be careful with multi-class ordered variables, otherwise they are wrongly interpreted. Can be used when a certain class matters.
- Numerical variables are better to find out thresholds.

Q. Which variables tend to be important for predicting drug use? How can these be interpreted? What are the implications of this outcome? How can you, as the data science communicator, discuss these findings in an ethical way?

- The importance of variables depends on moel to model. But for the tree we saw in this presentation, the most important variables are those that reflect family structure and emotional support from parents (eg: IFATHER, PRPROUD2 etc), which makes a lot of sense.

- As a data science communicator, always scale the data so that privacy is maintained.

- Always remember that these are just machine made predicitons and not every child who is not given allowance for chores uses marijuana.

- Since this dataset included very sensitive data like income of families and government allowance, handle that data carefully and never make judgement.

RESULTS

Binary Trees:

- Achieved a validation accuracy of 70.3% before pruning.
- Post pruning, achieved an accuracy of 72.2%.

Multi-Class Trees:

- Base multi-class decision tree achieved an accuracy of 25.25%, reflecting class imbalance and model variance.
- Tuned bagging model (mtry = 5, ntree = 25) improved performance to 29.46% accuracy, highlighting ensemble robustness.

Regression Trees:

- Implemented gradient boosting with 5 shrinkage values (0.001 to 0.3).
- Best MSE observed at shrinkage = 0.1 .

CONCLUSION

The study maps the relationship between a variety of factors and the use of marijuana. The insights from this study can be analysed further and awareness can be raised about them, with the end goal being limiting the number of teens that use marijuana.

We can look explicitly as to which predictor has maximum weight in deicidng if a teen uses marijuana and then try our best as a society to work on those predictors. Drug use ruins lives, especially when exposed so young.

BIBLIOGRAPHY

- Ko, M. (2022, July). Lecture 8: Bagging and Random Forests. University of Washington, CSE 416: Introduction to Machine Learning. https://courses.cs.washington.edu/courses/cse416/22su/lectures/8/lecture_8.pdf

- Gopal, M. (2023, March 7). What is Bagging in Machine Learning? A Guide with Examples. DataCamp. https://www.datacamp.com/tutorial/what-bagging-in-machine-learning-a-guide-with-examples

- Plot of decision tree made by AI.

- Harrell, D. (2023, August 10). Understanding the classification tree from tree package [Online forum post]. Posit Community. https://forum.posit.co/t/understanding-the-classification-tree-from-tree-package/168582

- References from class notes and worksheet codes were made.

THANK YOU!