Assignment procedure Guide/ Script For video

Task: Use decision trees to investigate factors that are correlated with youth drug use
Data: youth drug use using survey data from the National Survey on Drug Use and Health
Problem Statement/ Question:

- Binary Classification: Predicting if a youth has or has not used Marijuana before.
- Multi-classification: : Investigate if having a father present is related to drug use history,
- Regression: Number of days a person has consumed alcohol in the past year.

Problem 1 Conclusion:

- Classification tree was built using relevant predictors (excluding features with potential data leakage like age of first use, frequency, etc.).
- Unpruned tree test error rate: 9.5%
- Pruned tree (4 nodes) test error rate: 6.84%
- Pruning reduced the test error slightly, indicating that a simpler tree generalizes slightly better.
- The pruned decision tree model performed slightly better, suggesting that a simplified model can make reliable predictions about marijuana use. This is useful for identifying at-risk youth with relatively high accuracy and interpretability.

Problem 2 conclusion:

- Random forest classifier was trained to predict IFATHER using substance_cols.
- Model Accuracy:
    - mtry = 5: 71.3%
    - mtry = 3: Best at 71.7%
    - mtry = 10: 70.08%
    - mtry = 15: 70.4%
    - Bagging model (mtry = all predictors): 70.5%
- Important predictors: IRALCFY, IRALCAGE, IRCIGAGE, IRMJFY —
- This indicates that teens who used substances or started early were more likely not to have a father figure.

Problem 3 conclusion:

- Gradient Boosting model tested multiple shrinkage (lambda) values: 0.001, 0.01, 0.05, 0.1, 0.3
- Training MSE was lowest at lambda = 0.3
- But Test MSE remained high across all lambda values, indicating possible overfitting