STAT451 FINAL PROJECT

FRIDAY M., JOEL M. NAYDEEN M, EVE L.

INTRODUCTION- DATASET

- UC Irvine Machine Learning Repository
- Drug consumption patterns
  - Demographics
  - Physiological scoring/Personality-related scoring
  - Self-report drug use data for a range of substances
- Gave a decent foundation to start exploring relationships between individual traits and substance use behaviors
- https://archive.ics.uci.edu/dataset/373/drug+consumption+quantified

CLEANING PROCESS

- Data was quantified
- Dataset had one fictitious drug
  - 7 participants reported use
  - Removed the 7 participants from the dataset
- Chocolate was an included drug, removed from dataset
- Categorized drugs into hard drugs vs. non-hard drugs based on international classifications

QUESTIONS WE AIMED TO ANSWER

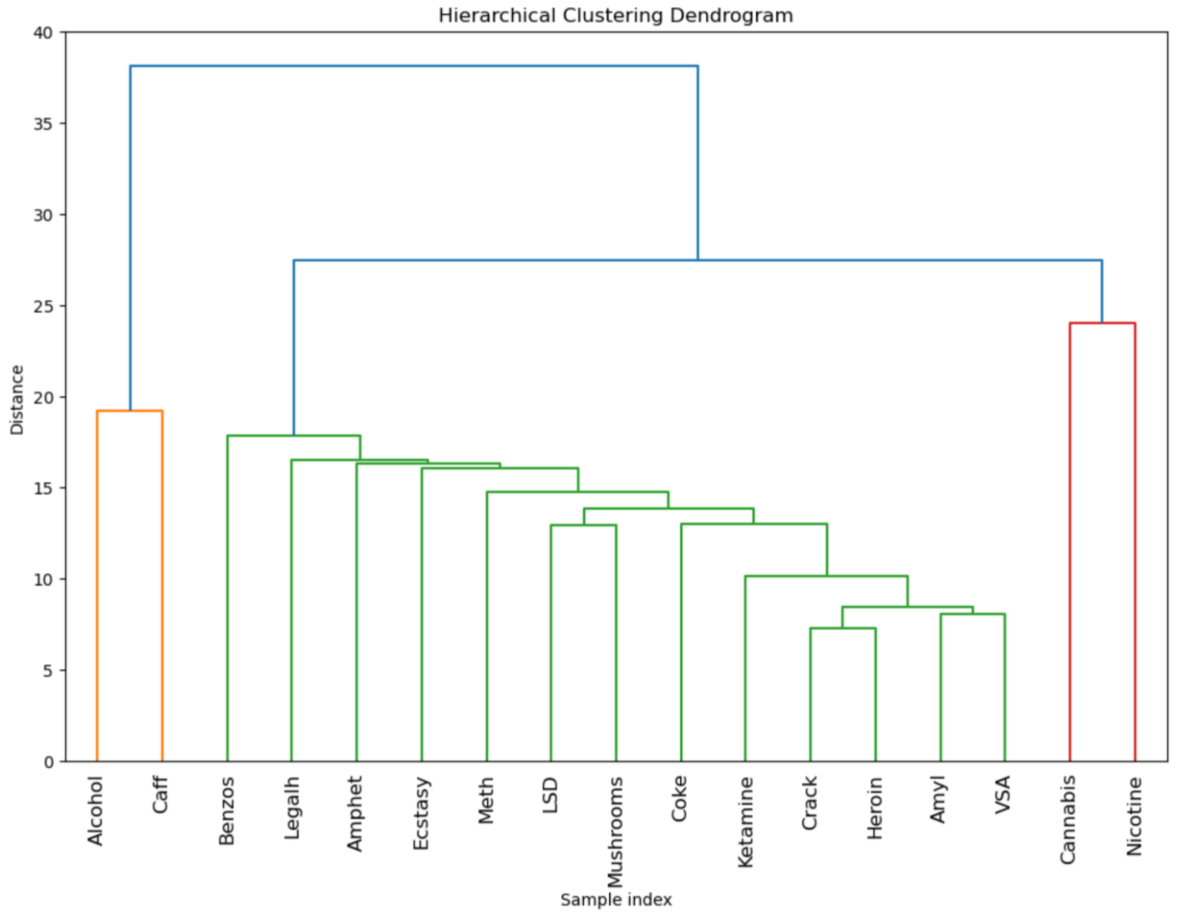
Question 1:What drugs are often used with other drugs?

Question 2:Which factors most strongly predict whether an individual uses cannabis/ consumes alcohol?

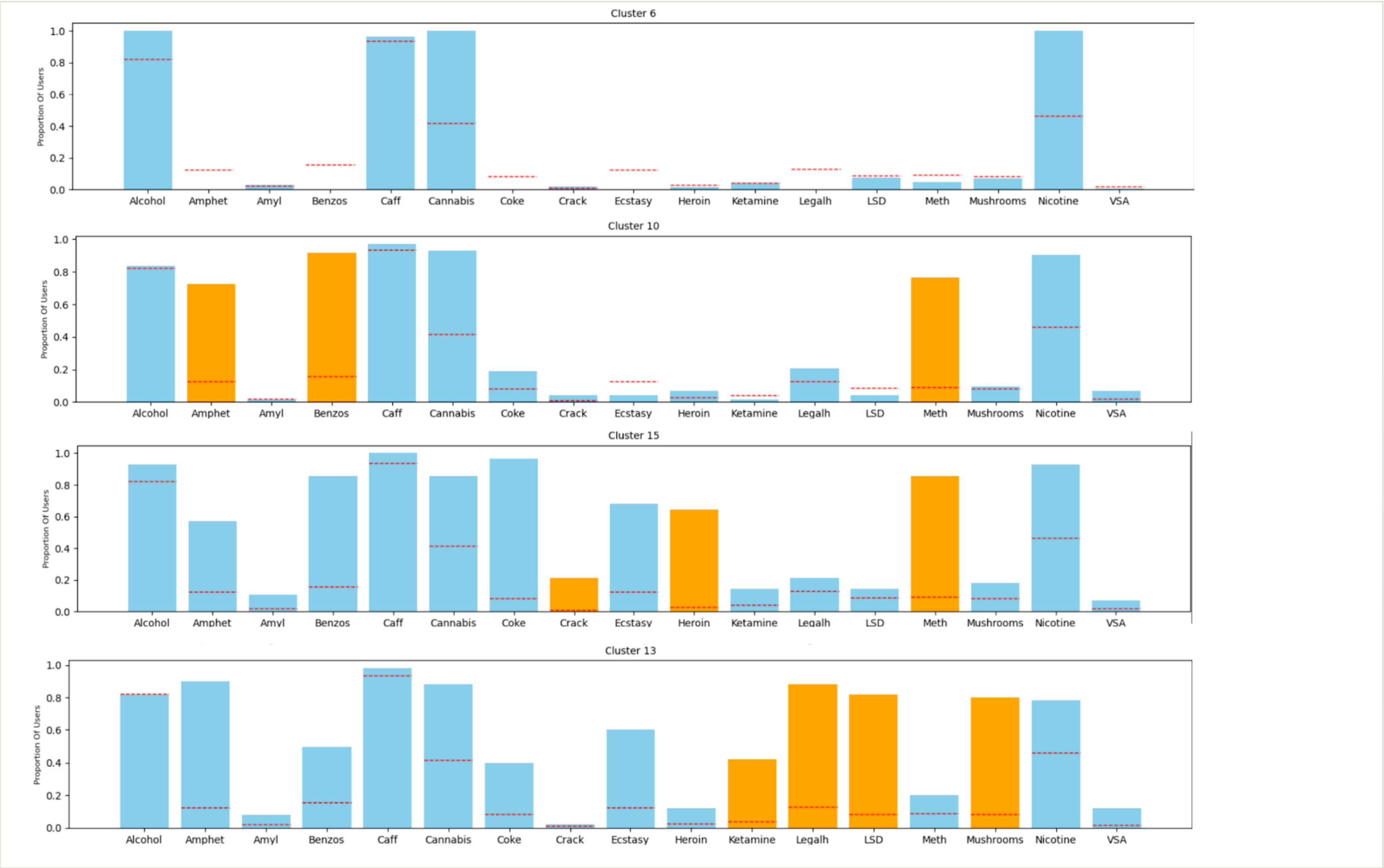
Question 3: Based on demographic and personality traits, can we classify individuals into low, moderate, or high-risk groups for substance abuse?

QUESTION 1: WHAT DRUGS ARE COMMONLY USED TOGETHER?

- Dendrogram
  - Groups drugs based on similarities in usage patterns, using hierarchical clustering
- Goal: identify clusters of drugs that are often used together or have similar usage behaviors
- Key findings
  - Alcohol and Caffeine are closely related, suggesting they are commonly used together
  - Cannabis and Nicotine form another distinct pair
  - All other substances fall into broader, less tightly related groups

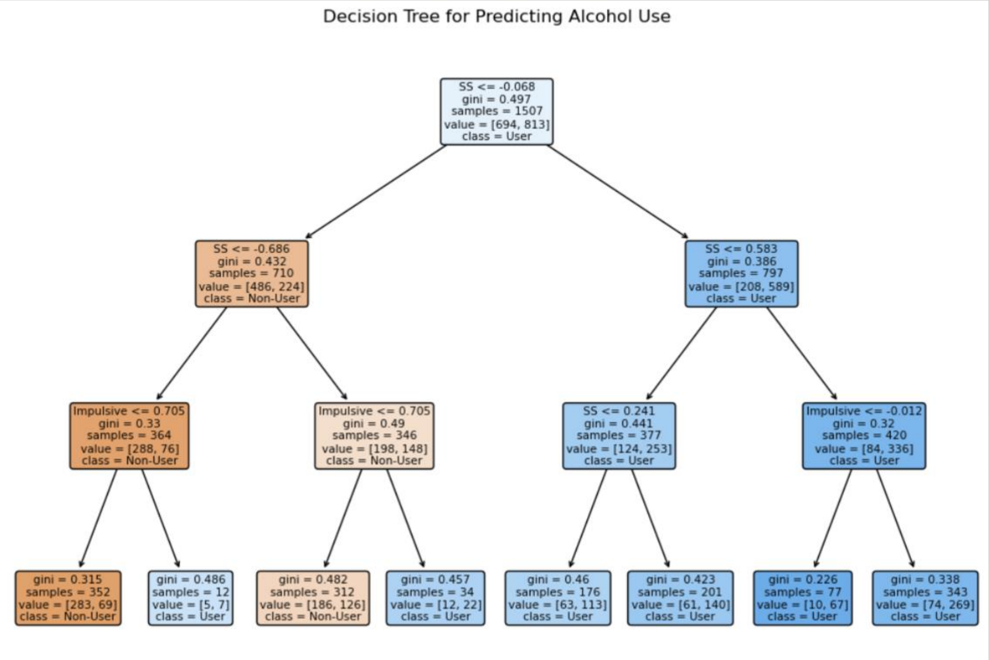


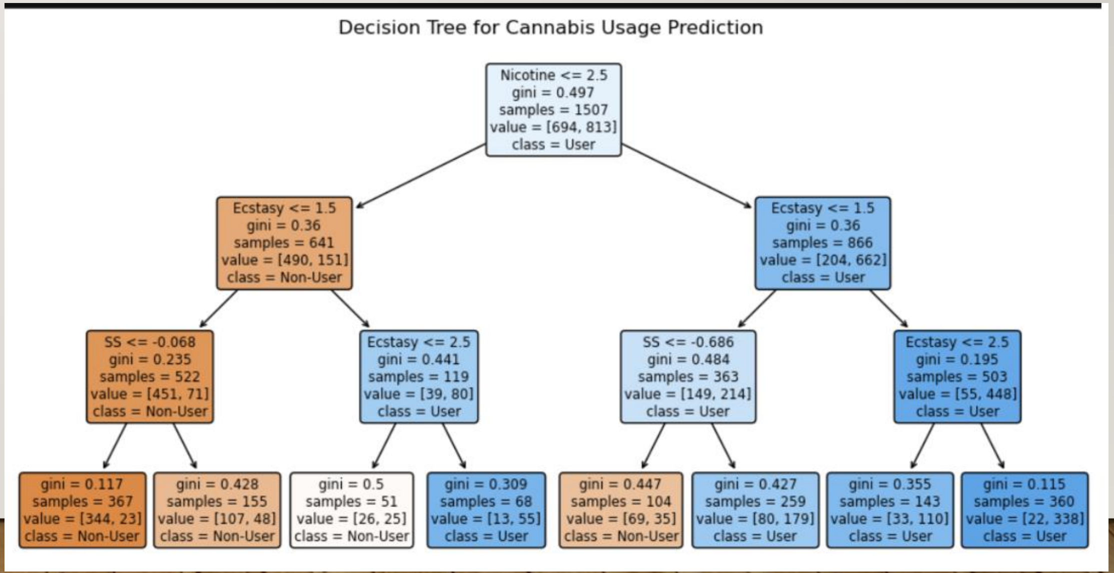
K-MEANS CLUSTERING (QUESTION 1)



QUESTION 2: WHICH FACTORS MOST STRONGLY PREDICT WHETHER AN INDIVIDUAL USES CANNABIS/ CONSUMES ALCOHOL?

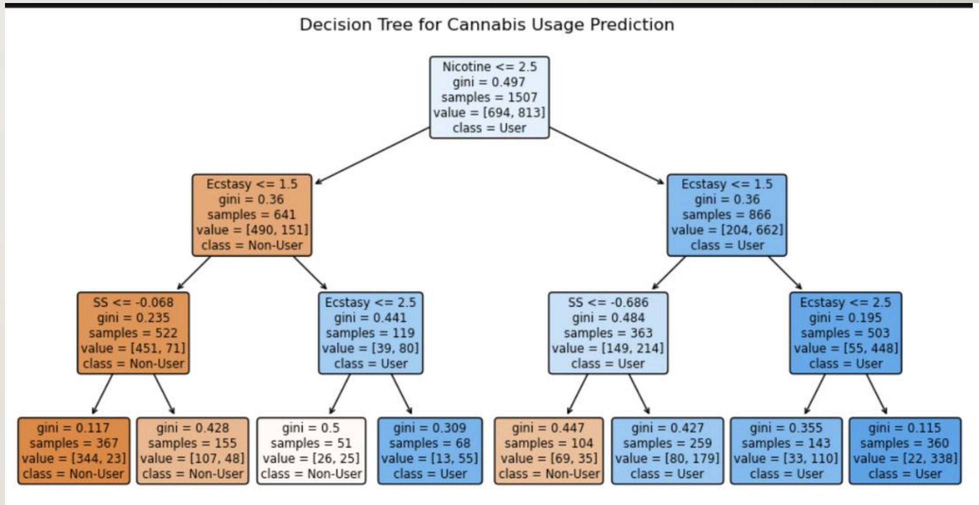
- Decision Tree for Alcohol
- Decision Tree for Cannabis


Decision Tree for Predicting Alcohol Use
SS <= -0.068 gini = 0.497 samples = 1507 value = [694, 813] class = User
SS <= -0.686 gini = 0.432 samples = 710 value = [486, 224] class = Non-User
SS <= 0.583 gini = 0.386 samples = 797 value = [208, 589] class = User
Impulsive <= 0.705 gini = 0.33 samples = 364 value = [288, 76] class = Non-User
Impulsive <= 0.705 gini = 0.49 samples = 346 value = [198, 148] class = Non-User
SS <= 0.241 gini = 0.441 samples = 377 value = [124, 253] class = User
Impulsive <= -0.012 gini = 0.32 samples = 420 value = [84, 336] class = User
gini = 0.315 samples = 352 value = [283, 69] class = Non-User
gini = 0.486 samples = 12 value = [5, 7] class = User
gini = 0.482 samples = 312 value = [186, 126] class = Non-User
gini = 0.457 samples = 34 value = [12, 22] class = User
gini = 0.46 samples = 176 value = [63, 113] class = User
gini = 0.423 samples = 201 value = [61, 140] class = User
gini = 0.226 samples = 77 value = [10, 67] class = User
gini = 0.338 samples = 343 value = [74, 269] class = User


Decision Tree for Cannabis Usage Prediction
Nicotine <= 2.5 gini = 0.497 samples = 1507 value = [694, 813] class = User
Ecstasy <= 1.5 gini = 0.36 samples = 641 value = [490, 151] class = Non-User
Ecstasy <= 1.5 gini = 0.36 samples = 866 value = [204, 662] class = User
SS <= -0.068 gini = 0.235 samples = 522 value = [451, 71] class = Non-User
Ecstasy <= 2.5 gini = 0.441 samples = 119 value = [39, 80] class = User
SS <= -0.686 gini = 0.484 samples = 363 value = [149, 214] class = User
Ecstasy <= 2.5 gini = 0.195 samples = 503 value = [55, 448] class = User
gini = 0.117 samples = 367 value = [344, 23] class = Non-User
gini = 0.428 samples = 155 value = [107, 48] class = Non-User
gini = 0.5 samples = 51 value = [26, 25] class = Non-User
gini = 0.309 samples = 68 value = [13, 55] class = User
gini = 0.447 samples = 104 value = [69, 35] class = Non-User
gini = 0.427 samples = 259 value = [80, 179] class = User
gini = 0.355 samples = 143 value = [33, 110] class = User
gini = 0.115 samples = 360 value = [22, 338] class = User

CANNABIS PREDICTION

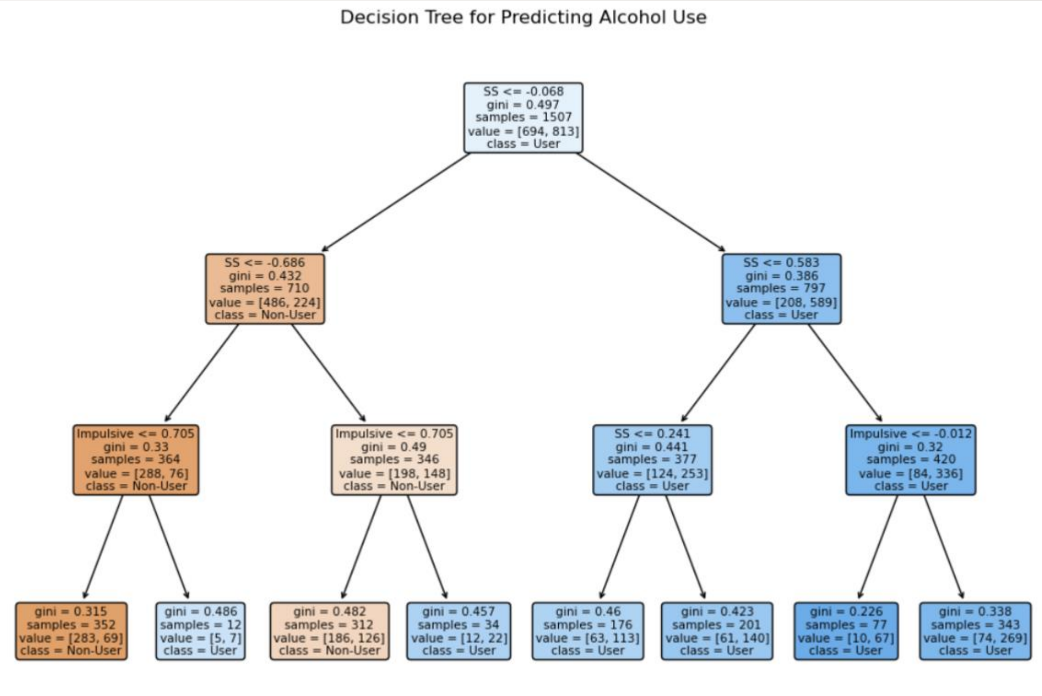
- Starting point: root node starts with Nicotine, split at <= 0.25.
- Other important predictors: Ecstasy usage and the Sensation Seeking (SS) variable.
- Outcomes
  - Final classifications: User or Non-User
  - If nicotine >2.5 and Ecstasy <= 1.5, individual predicted to be Cannabis User
  - Inverse, likely to not be a User



Decision Tree for Cannabis Usage Prediction

ALCOHOL PREDICTION

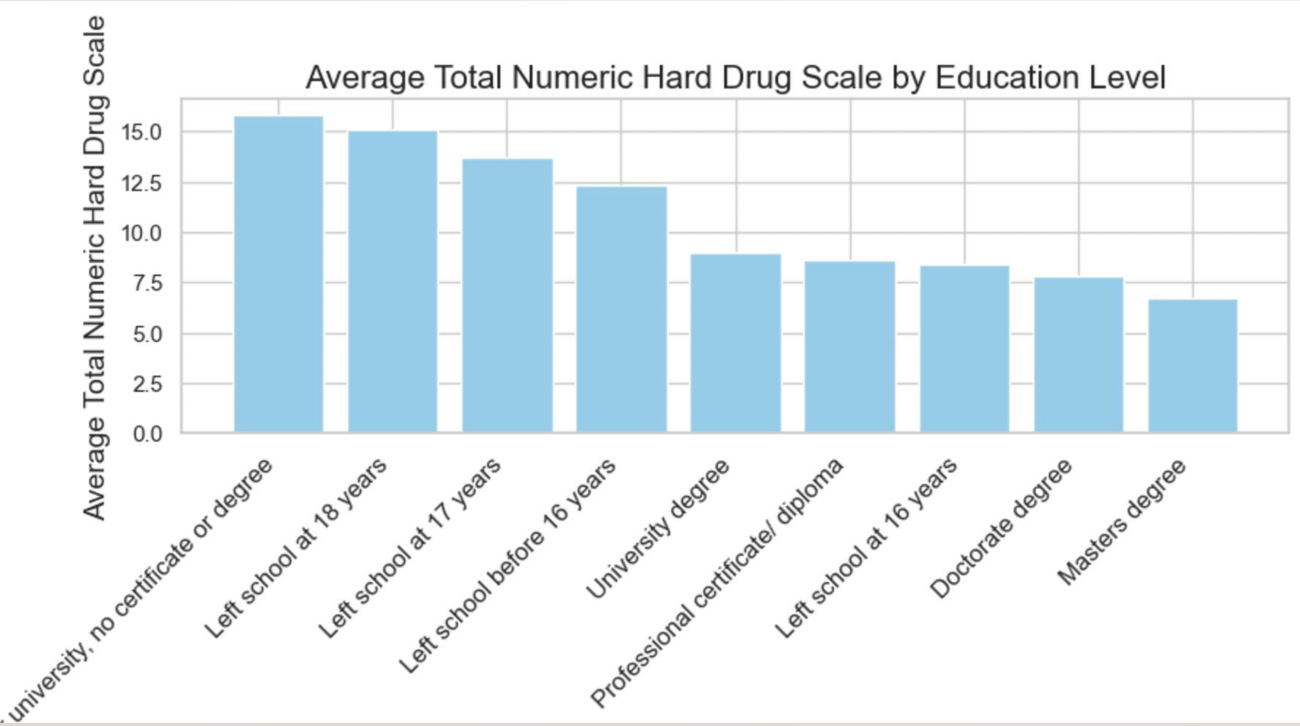
- Starting point: root node starts with the Sensation Seeking score, split at 0.068.
    - Indicates individuals with a low SS are more likely to not use alcohol
    - Other important factors was the Impulsiveness personality score
- Outcomes
    - SS > 0.068 and Impulsive <= 0.705, individual is likely to be a user
    - Inverse outcome, likely to not use alcohol

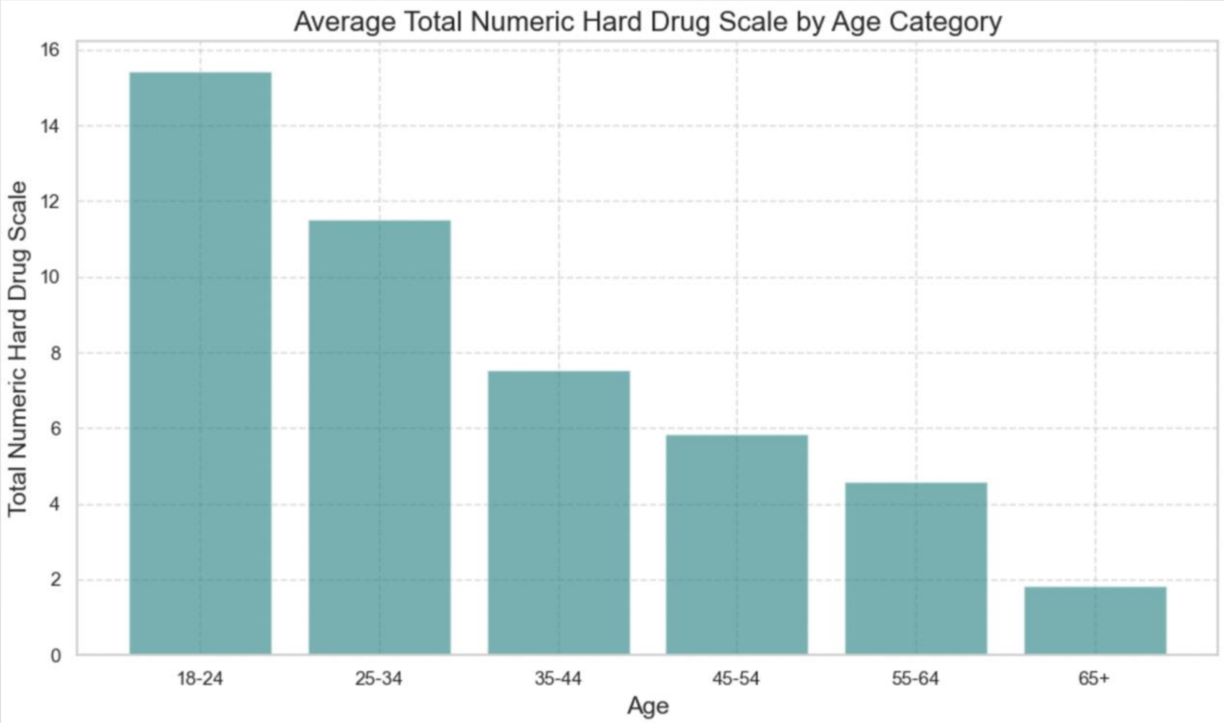

Decision Tree for Predicting Alcohol Use

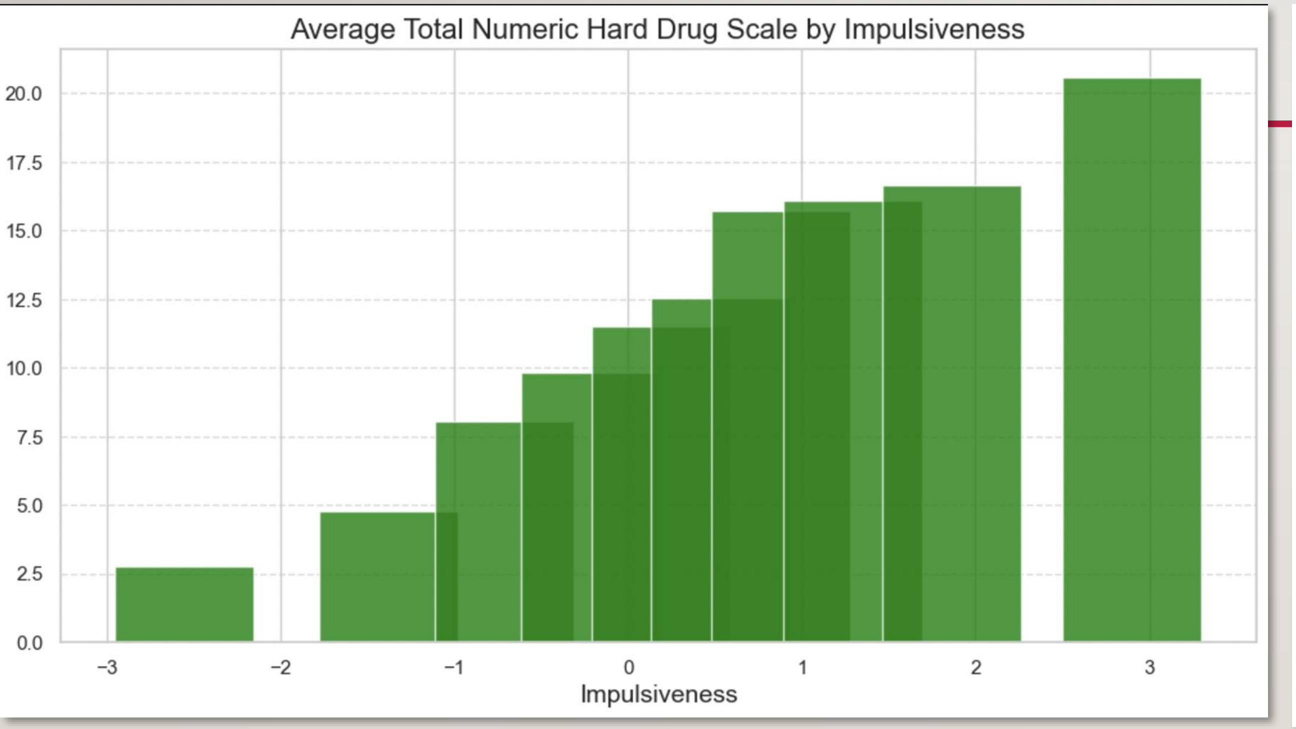
COMPARING THE TREES

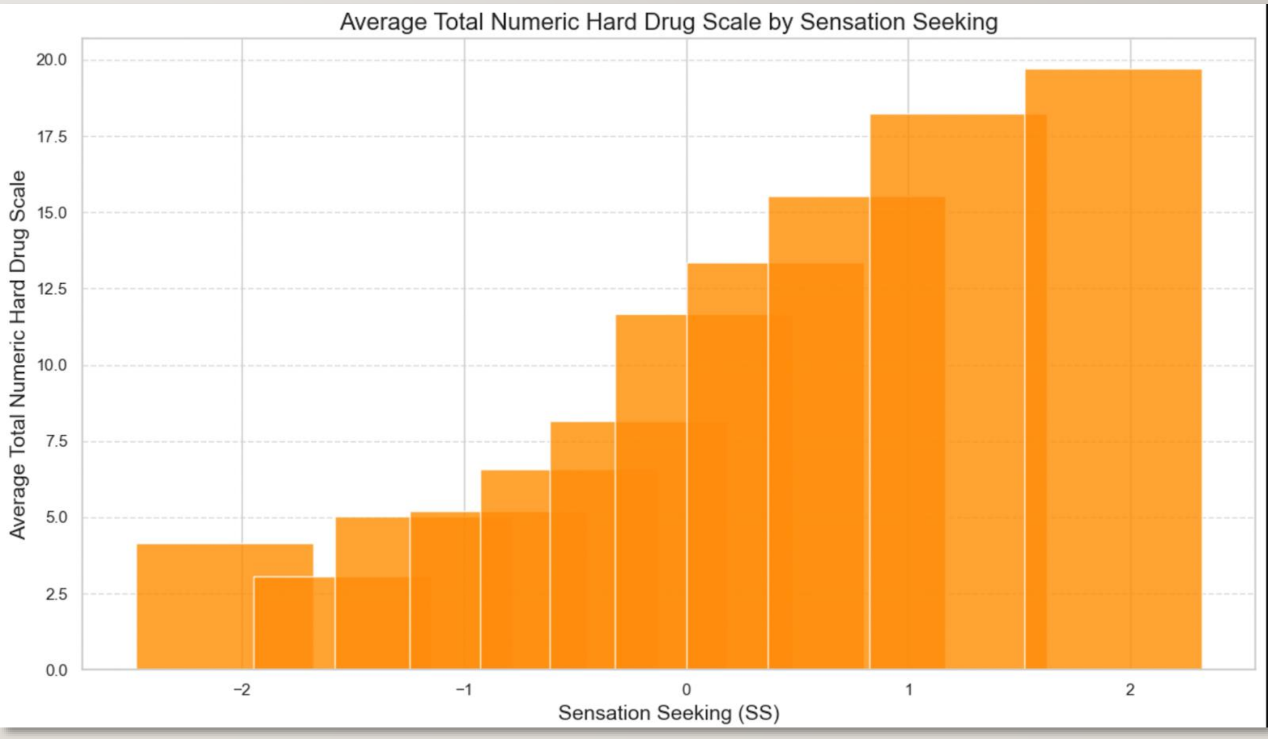
- Cannabis:
  - Nicotine and Ecstasy usage are key predictors, highlighting the role of co-usage with other substances
- Alcohol:
  - Behavioral Traits like Sensation Seeking and Impulsiveness are more significant, reflecting a stronger psychological component.

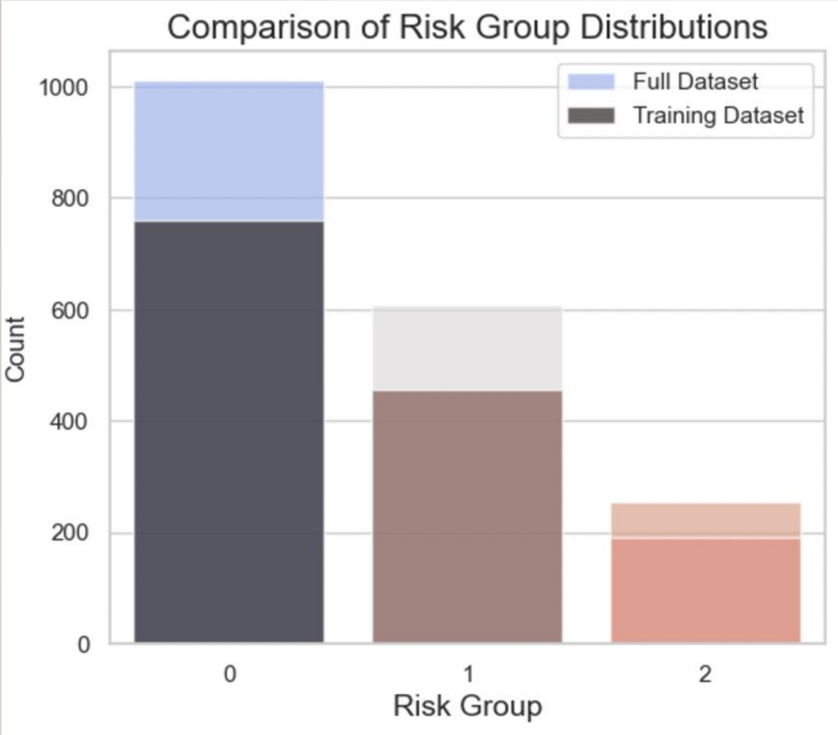
QUESTION 3: BASED ON DEMOGRAPHIC AND PERSONALITY TRAITS, CAN WE CLASSIFY INDIVIDUALS INTO LOW, MODERATE, OR HIGH-RISK GROUPS FOR SUBSTANCE ABUSE?

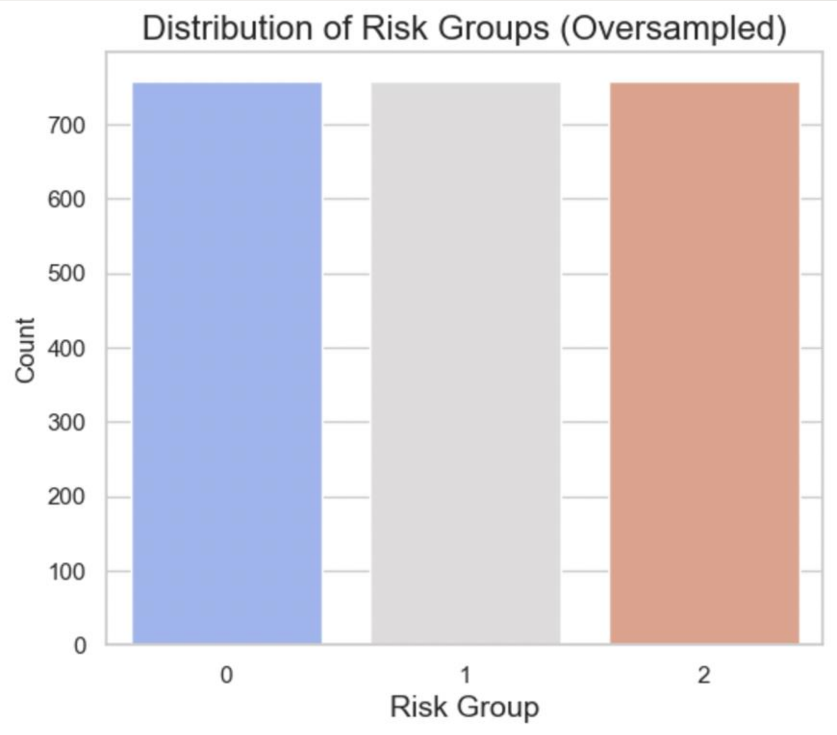





Average Total Numeric Hard Drug Scale by Impulsiveness
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
−3
−2
−1
0
1
2
3
Impulsiveness


Average Total Numeric Hard Drug Scale by Sensation Seeking
Average Total Numeric Hard Drug Scale
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
−2
−1
0
1
2
Sensation Seeking (SS)

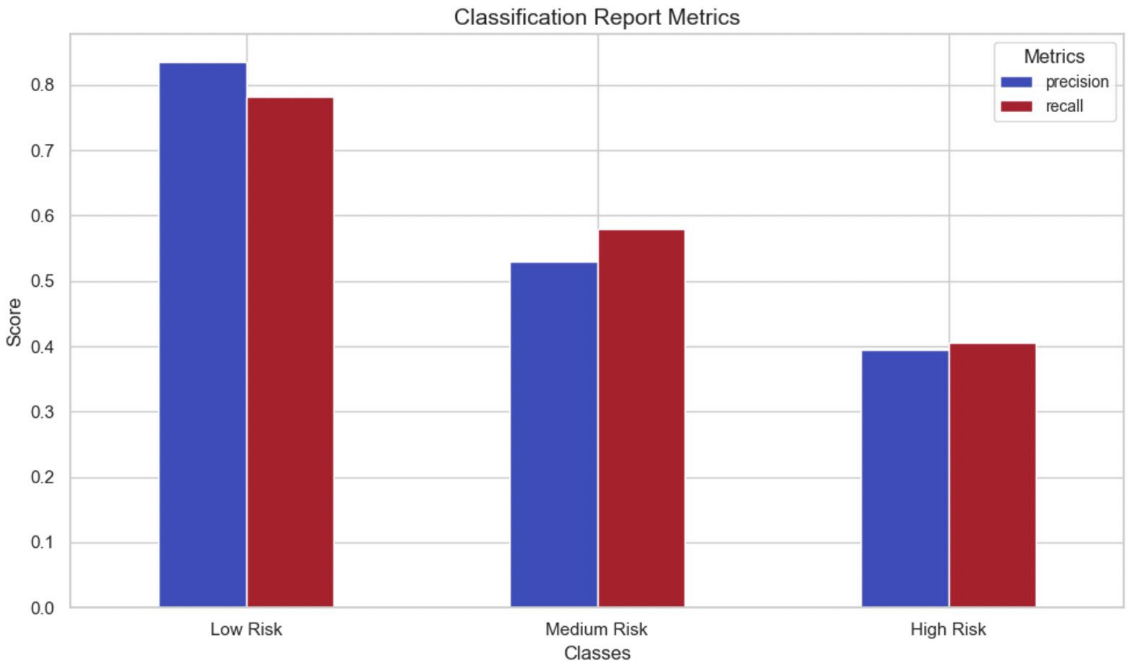


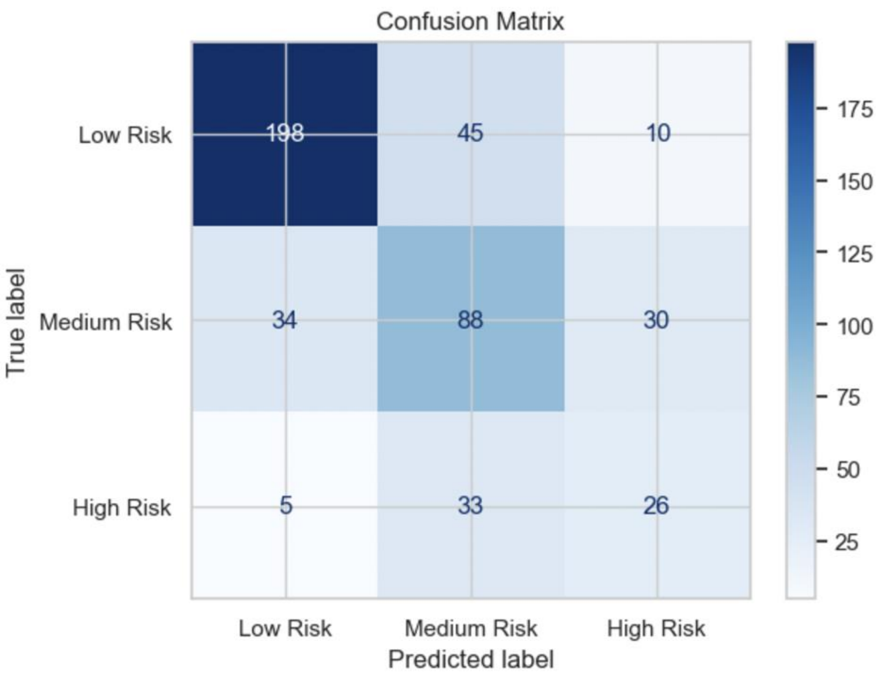


OVERSAMPLING

RANDOMFORESTCLASSIFIER

- Mean CV Score: 0.8564
- Test Accuracy: 0.6652





CONSIDERATIONS - BUILDING A MORE ACCURATE MODEL

- Two classes - Low Risk, High Risk
- More features
  - Childhood maltreatment
  - Familial substance abuse
- Individual Risk factors
  - Neurodevelopmental disorders
  - Mental illness
- Social risk factors
  - Bullying
  - Gang affiliation

CONCLUSION

- **Patterns of Co-Usage:**
Alcohol and caffeine, as well as cannabis and nicotine, show strong co-usage tendencies, reflecting shared social or behavioral contexts.
- **Substance Use Predictors:**
- **Cannabis:** Nicotine and Ecstasy usage, along with Sensation Seeking scores, were strong predictors.
- **Alcohol:** Psychological traits such as Sensation Seeking and Impulsiveness dominated prediction accuracy.
- **Risk Classification:**
While the Random Forest model achieved promising results (Mean CV: 85.6%), further improvements could involve deeper demographic and psychosocial variables.