Project Report

- **Project Name: Drug Usage Predictions & Clustering based on human metrics and other drug-related factors**

- **Team name**
  - The Tenderloins

- **Team members**
  - Varun Singh (singhva@bc.edu)
  - Lahari Somasi (lahari@stanford.edu)
  - Teo Ivancevic (tivancev@stanford.edu)
  - Muhammad Talha Salani (salani01@stanford.edu)

- **Discussion Section 2**

Introduction:

Our project aims to explore patterns and correlations in drug usage among various demographic groups using the Drug Consumption dataset from the UCI Machine Learning Repository. This dataset provides detailed information on the drug consumption habits of 1885 respondents, including their use of both legal and illegal substances. Additionally, the dataset includes personality metrics, demographic information, and quantified categorical attributes, making it a rich source for predictive modeling and clustering analyses.

The purpose of our project is to investigate several key questions about drug usage. We are interested in predicting the likelihood of drug usage based on individual characteristics, identifying clusters of drugs that tend to be used together, and analyzing how drug usage varies across different demographic groups. These questions are significant because they can provide insights into the factors that influence drug use, identify potential "drug buddies," and help understand the demographic distribution of drug usage. By applying predictive models and clustering techniques, we aim to uncover patterns that could inform public health strategies and interventions.

Dataset Description:

a. Data Provenance:

The dataset we are using for our analysis is the "Drug Consumption (quantified)" dataset, which is publicly available from the UCI Machine Learning Repository. The dataset can be accessed at:(https://archive.ics.uci.edu/dataset/373/drug+consumption+quantified).

Provenance:

The dataset was contributed by the UCI Machine Learning Repository, which hosts a wide range of datasets used for research and educational purposes. This particular dataset contains information on drug consumption patterns among 1885 respondents. The data was collected through self-reported surveys where respondents provided information about their drug use, personality traits, and demographic background. The dataset was originally used in research to explore the relationships between personality traits and drug usage, and it is frequently cited in studies focusing on drug consumption patterns.

Structure:

The dataset is structured as a tabular file, where each row represents an individual respondent, and each column represents a specific attribute or feature. The dataset includes the following key components:

- Observational Units (Rows): Each row corresponds to a single respondent, providing a unique record of that individual's drug usage, personality traits, and demographic information.

- Features (Columns): The dataset contains a total of 31 columns, which can be broadly categorized into three groups:

    1. **Demographic Information:** Includes attributes such as age, gender, level of education, country of residence, and ethnicity.

    2. **Personality Traits:** Comprises NEO-FFI-R measurements (neuroticism, extraversion, openness to experience, agreeableness, conscientiousness), BIS-11 (impulsivity), and ImpSS (sensation seeking).

    3. **Drug Usage:** Contains information on the usage of 18 drugs, including both legal and illegal substances (e.g., alcohol, cannabis, cocaine, heroin), and a fictitious drug (Semeron) to identify over-claimers. Each drug usage is classified into seven categories: "Never Used," "Used Over a Decade Ago," "Used in Last Decade," "Used in Last Year," "Used in Last Month," "Used in Last Week," and "Used in Last Day."

    4. **Size and Shape:** The dataset comprises 1885 records (rows) and 31 attributes (columns). Each attribute has been quantified, and

the values are numerical, allowing for statistical analysis and machine learning applications.

Data Quality:

The dataset is generally well-structured and suitable for analysis. However, there are some aspects to consider:

- **Missing Values:** The dataset does not contain explicit missing values, but the classification of drug usage into categories means that some responses may be less granular, potentially affecting the precision of the analysis.

- **Data Quality Issues:** Given that the data is self-reported, there is a potential for bias or inaccuracies, particularly with sensitive topics like drug usage. The inclusion of a fictitious drug (Semeron) helps identify respondents who may not have answered truthfully.

b. Data Preparation:

To prepare the dataset for analysis, several steps were taken to ensure the data was in a suitable format:

1. Conversion of Categorical Data: The drug usage data was initially categorical, representing different levels of consumption. These categories were converted into numerical values to facilitate correlation analysis and predictive modeling. Each category was assigned a numerical value from 0 ("Never Used") to 6 ("Used in Last Day").

2. Feature Selection: For our specific analyses, we focused on a subset of features, particularly the drug usage columns and the demographic and personality attributes. This selection allowed us to build models that address the research questions related to drug usage patterns, correlations, and clustering.

3. Data Normalization: The personality trait scores and other continuous variables were normalized to ensure that they were on a comparable scale. This step was essential for clustering and regression models.

4. Handling of Outliers: The dataset includes a fictitious drug to identify over-claimers. Responses indicating usage of this drug were flagged as potential outliers. Depending on the analysis, these records were either excluded or analyzed separately.

These data preparation steps ensured that the dataset was ready for the specific analytical questions we aimed to address, allowing for a more accurate and insightful analysis.

Exploratory Data Analysis (EDA):

Relevant EDA that was conducted included the plotting and visualization between subsets of our feature variables (ie. age, gender, education) and our target variables. Since all variables in these visualizations were categorical, the color and the intensity of each color for each observation in the scatter plot represented the degree of classification for that specific combination of feature variables.

Doing this for different subsets of our feature variables gave us a better grasp of the data and allowed us to see what variables actually affected our targets and if certain groupings of two feature variables (ie. ethnicity and education) affected our targets differently than other groupings of two feature variables (ie. gender and country of origin)

5.1. Question 1 (Varun)

Question: Given the non-drug related metrics (ie. Age Range, Gender, Education level, country of origin, ethnicity, Nscore, Escore, Oscore, AScore, CScore, Impulsiveness) of an individual, predict their frequency of use for cannabis. Our target variable, “frequency of use” for cannabis will consist of two levels: “user” or “nonuser”. From the original dataset, levels CL0, CL1, and CL2 will map to “non-user” and levels CL4, CL5, and CL6 will map to “user”. Essentially this is a binary classification problem.

Design

1. **Data Used:**

   Dataset: Drug Consumption (quantified) dataset from the UCI Machine Learning Repository.

   Features: age, gender, education, country, ethnicity, n-score, e-score, o-score, a-score, c-score, impulsive, ss.

   Target Variable for Predictive Modeling: cannabis_binary

2. **Data Preparation Steps:**

   1. **Conversion of Categorical Features:** Transformed “age”, “gender”, “education”, “country”, and “ethnicity” from numerical variables to categorical (since these feature variables are categorical by nature). This allowed us to one-hot encode these variables. categorical drug usage responses into numerical

values, where each response category (e.g., "Never Used," "Used in Last Month") was mapped to a numerical scale (0 to 6).

2. Normalization: As for our numerical variables (n-score, e-score, o-score, a-score, c-score, impulsive, and ss) we standardized them using the standard scaler method.

3. Target Variable creation: Created binary column for our target drug (cannabis) which contained the label “user” or “non-user” for each observation in our dataset. This target column will be the dependent variable for our classification model aimed at shedding light on this overall question.

Technical Approach:

1. Predictive Modeling:

-Model Used:

- KNN Classifier:
- Specific Metrics Used:
  number of neighbors: 19
  distance metric: “euclidean”

-Objective: Predict cannabis usage based on basic and psychological metrics of an individual.

-Evaluation Metrics: Accuracy and 10-fold cross-validation score to determine the model's performance in predicting cannabis usage.

Implementation

1. **Predictive Modeling:**
   - Conducted a grid-search cross-validation to find the best hyperparameters for our KNN classifier.
     - The range for the number of neighbors for our grid search was between 1 and 20 inclusive.
     - The two distance metrics we compared for our use case were “euclidean” and “manhattan”.
   - Split our dataset into training and test portions using 10-fold cross validation.

- Constructed a confusion matrix based on the predictions made for frequency of cannabis use by our model using the best hyperparameters and the ground truths for frequency of cannabis use.

- Evaluated the performance of our model based on the accuracy score (True Positives and True Negatives) derived from this confusion matrix.

Results:

- **Accuracy:**
- The classification model used for this question was roughly 80% correct in its classification for frequency of cannabis use, exhibiting decent performance.

5.2. Question 2 (Lahari)

Formulation of Question 2:

Cluster individuals based on their personality and types of drugs consumed. Determine the optimal number of clusters and interpret the clusters formed.

Design and Implementation:

We have a dataset containing information about individuals' personality scores (Nscore, Escore, Oscore, Ascore, Cscore), impulsivity (Impulsive), sensation seeking (SS), and their consumption of various drugs. Our plan is to use this data to perform a clustering analysis that groups individuals into distinct clusters based on these features. We will be using the Elbow method to find the optimal number of clusters.

1. **Data Used:** The dataset consists of 1,885 rows representing individual profiles with the following features:
    - Personality scores: Nscore, Escore, Oscore, Ascore, Cscore
    - Impulsivity: Impulsive
    - Sensation Seeking: SS
    - Drug consumption: Alcohol, Amphet, Amyl, Benzos, Caff, Cannabis, Choc, Coke, Crack, Ecstasy, Heroin, Ketamine, Legalh, LSD, Meth, Mushrooms, Nicotine, Semer, VSA
2. **Proposed Solution:** We will follow these steps:
    a. Load & Preprocess the dataset: Download and read the data into a suitable format. Clean the data and prepare it for clustering.
    b. Feature Selection: Choose the features mentioned in #3.
    c. Standardize the Data: Scale the features to have similar ranges.
    d. Apply Clustering: Perform K-means clustering to identify distinct groups.
    e. Determine the Number of Clusters
    f. Interpret the Clusters: Analyze the characteristics of each cluster.
3. **Proposed Evaluation:** We will evaluate the clustering by analyzing the characteristics of each cluster. Specifically, we will look at the mean values of the personality scores, impulsivity, sensation seeking, and drug consumption for each cluster. This will help us

understand the distinct groups formed by the clustering process and their defining features.

For each cluster, we will calculate:

- Mean values of personality scores, impulsivity, and sensation seeking.
- Mean values of drug consumption for each type of drug.

We will use these metrics to interpret the clusters and understand the relationships between personality traits, behaviors, and drug consumption habits.

Results:

Interpretation of Clusters: Given the features `['Nscore', 'Escore', 'Oscore', 'Ascore', 'Cscore', 'Impulsive', 'SS', 'Alcohol', 'Amphet', 'Amyl', 'Benzos', 'Caff', 'Cannabis', 'Choc', 'Coke', 'Crack', 'Ecstasy', 'Heroin', 'Ketamine', 'Legalh', 'LSD', 'Meth', 'Mushrooms', 'Nicotine', 'Semer', 'VSA']`, clustering the data into four groups would result in clusters that capture distinct personality profiles and drug consumption behaviors.

1. **Cluster 0: The High Sensation Seekers and Experimenters**
    a. Personality Traits:
        i. High Sensation-Seeking and Impulsivity (SS and Impulsive): This cluster is characterized by high scores in sensation-seeking and impulsivity. These individuals are more likely to seek out new and thrilling experiences, sometimes without considering the risks.
        ii. Moderate to High Neuroticism (Nscore): These individuals may also exhibit moderate levels of neuroticism, indicating some degree of emotional instability or anxiety.
    b. Drug Consumption Patterns:
        i. High Use of Recreational Drugs: This cluster tends to have higher-than-average use of recreational drugs like Cannabis, Ecstasy, LSD, and Mushrooms. The experimentation with various substances aligns with their high sensation-seeking trait.
        ii. Moderate Alcohol and Nicotine Use: Alcohol and nicotine usage is moderate, suggesting that while they indulge in recreational substances, they may not heavily rely on alcohol or nicotine.
    c. Summary: This cluster represents individuals who are likely to be thrill-seekers and are more open to experimenting with different substances. They may be emotionally volatile and engage in drug use as a form of exploration or self-medication.
2. **Cluster 1: The Disciplined and Conservative Users**
    a. Personality Traits:
        i. High Conscientiousness (Cscore) and Low Neuroticism (Nscore): These individuals are disciplined, responsible, and emotionally stable. They likely plan their actions carefully and avoid risky behaviors.

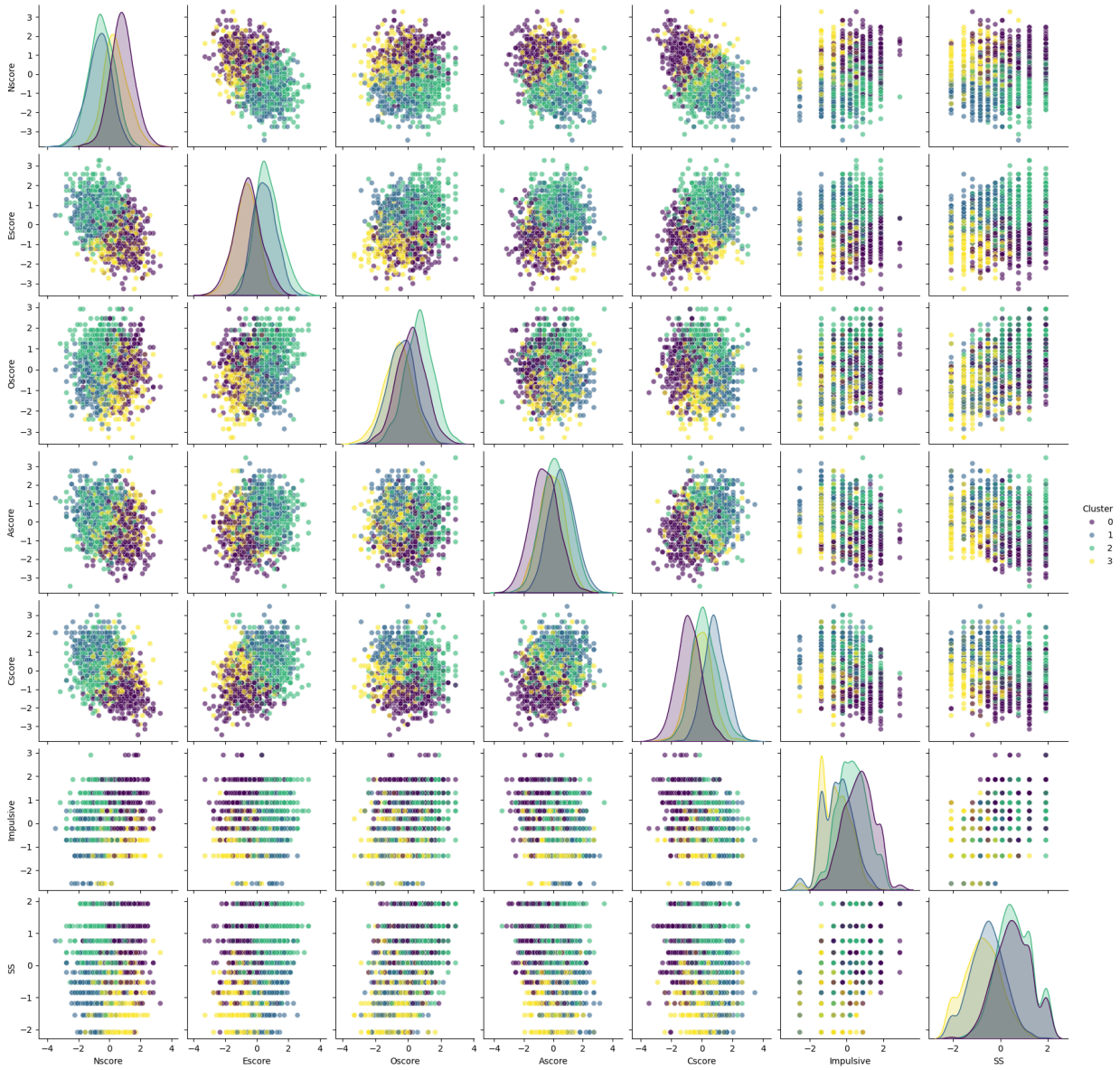
    ii. Low Sensation-Seeking and Impulsivity (SS and Impulsive): This group scores low on sensation-seeking and impulsivity, indicating a preference for stability and routine.
  b. Drug Consumption Patterns:
    i. Low to Minimal Drug Use: This cluster exhibits very low usage of most drugs, including recreational substances like Cannabis, Ecstasy, and LSD. Their conservative nature likely leads them to avoid drug use altogether.
    ii. Moderate to Low Alcohol Use: Alcohol consumption might be moderate but controlled, aligning with their disciplined nature.
  c. Summary: Cluster 1 consists of individuals who are conservative, careful, and avoid engaging in drug use. They prioritize stability and self-control, which is reflected in their minimal substance use.
3. **Cluster 2: The Socially Engaged and Balanced Users**
  a. Personality Traits:
    i. High Extraversion (Escore) and Agreeableness (Ascore): This cluster is sociable, outgoing, and cooperative. These individuals are likely to engage in social activities and enjoy being around others.
    ii. Moderate Sensation-Seeking and Conscientiousness (SS and Cscore): They have a balanced approach to life, with moderate levels of sensation-seeking and conscientiousness.
  b. Drug Consumption Patterns:
    i. Moderate Alcohol and Cannabis Use: Alcohol and Cannabis are used moderately, often in social settings or as a means to relax.
    ii. Low to Moderate Use of Other Substances: Use of other drugs like Amphetamines, Cocaine, and Ecstasy may be occasional and likely influenced by social circles rather than personal inclination.
  c. Summary: Cluster 2 represents socially active individuals who use substances like alcohol and cannabis moderately. They strike a balance between enjoying life and maintaining self-discipline, engaging in drug use primarily in social contexts.
4. **Cluster 3: The Emotional and Dependent Users**
  a. Personality Traits:
    i. High Neuroticism (Nscore) and Low Extraversion (Escore): This cluster consists of individuals who are more emotionally unstable, anxious, and introverted. They may struggle with social interactions and rely on substances as a coping mechanism.
    ii. Low Agreeableness and Conscientiousness (Ascore and Cscore): These individuals may have difficulties in social relationships and might lack the discipline to avoid risky behaviors.
  b. Drug Consumption Patterns:
    i. High Use of Hard Drugs: This cluster has a higher likelihood of consuming more dangerous substances like Heroin, Crack, and Methamphetamines. Their drug use may be driven by emotional struggles and a lack of social support.

ii. High Nicotine and Alcohol Use: They may also heavily rely on nicotine and alcohol, indicating potential substance dependence.

c. Summary: Cluster 3 includes individuals who are emotionally vulnerable and may use substances as a form of escape. Their drug consumption patterns suggest a potential for addiction or dependence, with a focus on more harmful substances.

This clustering provides a nuanced understanding of the relationships between personality traits and drug consumption behaviors. Each cluster represents a different combination of personality profiles and substance use, offering insights that can be used for targeted interventions, further research, or understanding the diverse behaviors within the population.

Visualization - Pair Plots for Personality scores, Impulsivity, and Sensation seeking



5.3. Question 3 (Teo)

Question: How can information about the demographic group of a person be used to predict the likelihood and recency for use of each drug.

Explanation: The goal of this question is to explore how certain demographic parameters, like age, education and gender could be used to predict drug use over time.

Design

1. **Data Used:**

   Dataset: Drug Consumption (quantified) dataset from the UCI Machine Learning Repository.

   Features: Usage data of 18 different drugs and demographic attributes (specifically age, education and gender).

   Target Variable for Predictive Modeling: Multi-class (for all drugs)

2. **Data Preparation Steps:**

   Loading and Cleaning Data: Import the dataset, handle any missing values, and ensure data integrity.

   Encoding and Scaling: Convert categorical variables into numerical form and apply feature scaling to standardize the data.

   Train-Test Split: Split the dataset into training and testing sets for model evaluation.

Technical Approach

1. **Model Selection:**

   K-Nearest Neighbors (KNN): I chose KNN because it's straightforward and works well with smaller datasets. I used cross-validation to find the best number of neighbors (K).

   Random Forest: I selected Random Forest for its ability to handle complex data and provide probability estimates for each drug use category.

2. **Model Training and Evaluation:**

   Cross-Validation: I applied 5-fold cross-validation to ensure the model's results were reliable and not just due to a lucky split.

Evaluation Metrics: I evaluated the models using accuracy, F1 score, and a confusion matrix to see how well they classified different drug use categories.

3. **Probability Prediction with Random Forest:**

   Estimating Probabilities: The Random Forest model was used to predict the probability of each drug use category, which helps understand the confidence of the predictions.

Implementation

1. **Class Imbalance:**

   Some drug use categories were underrepresented. I adjusted class weights in the Random Forest model to improve F1 scores and ensure these categories were not ignored.

2. **Hyperparameter Tuning:**

   I used cross-validation to find the best K value for KNN and set n_estimators=100 for Random Forest, balancing model performance with computational efficiency.

3. **Feature Importance:**

   Random Forest's feature importance helped identify key demographic and psychological factors, providing insights into the model's decision-making process.

4. **Probability Calibration:**

   I checked the accuracy of probability predictions to avoid overconfidence in incorrect predictions. Calibration methods like isotonic regression were considered if needed.

5. **Test Set Performance:**

   The models were tested on unseen data to evaluate real-world performance, using metrics like accuracy, F1 score, and confusion matrix analysis.

Results

K-Nearest Neighbors (KNN)

- **Accuracy Trend:**

    - Visualized in a line plot (Figure 1), the accuracy of the KNN model improved as the number of neighbors (K) increased, although the overall accuracy remained relatively low.
- **Confusion Matrix Analysis:**
    - The KNN confusion matrix (Figure 2) revealed significant misclassifications across multiple drug use categories, indicating that the model struggled to distinguish between certain classes.
- **Summary:**
    - The results suggest that while tuning K improved KNN's performance slightly, the model's ability to correctly classify drug use categories is limited, likely due to the complexity and overlap in the dataset.

Random Forest

- **Confusion Matrix Analysis:**
    - The Random Forest confusion matrix (Figure 3) demonstrated better performance in identifying certain categories compared to KNN but still faced challenges, particularly in differentiating among less frequent classes.
- **Model Strengths:**
    - The Random Forest model captured more nuanced patterns in the data, reflecting its robustness and ability to handle feature complexity better than KNN.
- **Summary:**
    - Despite the improved classification over KNN, the Random Forest model still struggled with clear class separations, suggesting that additional data preprocessing or more advanced techniques might be needed to enhance performance further.

Visualizations

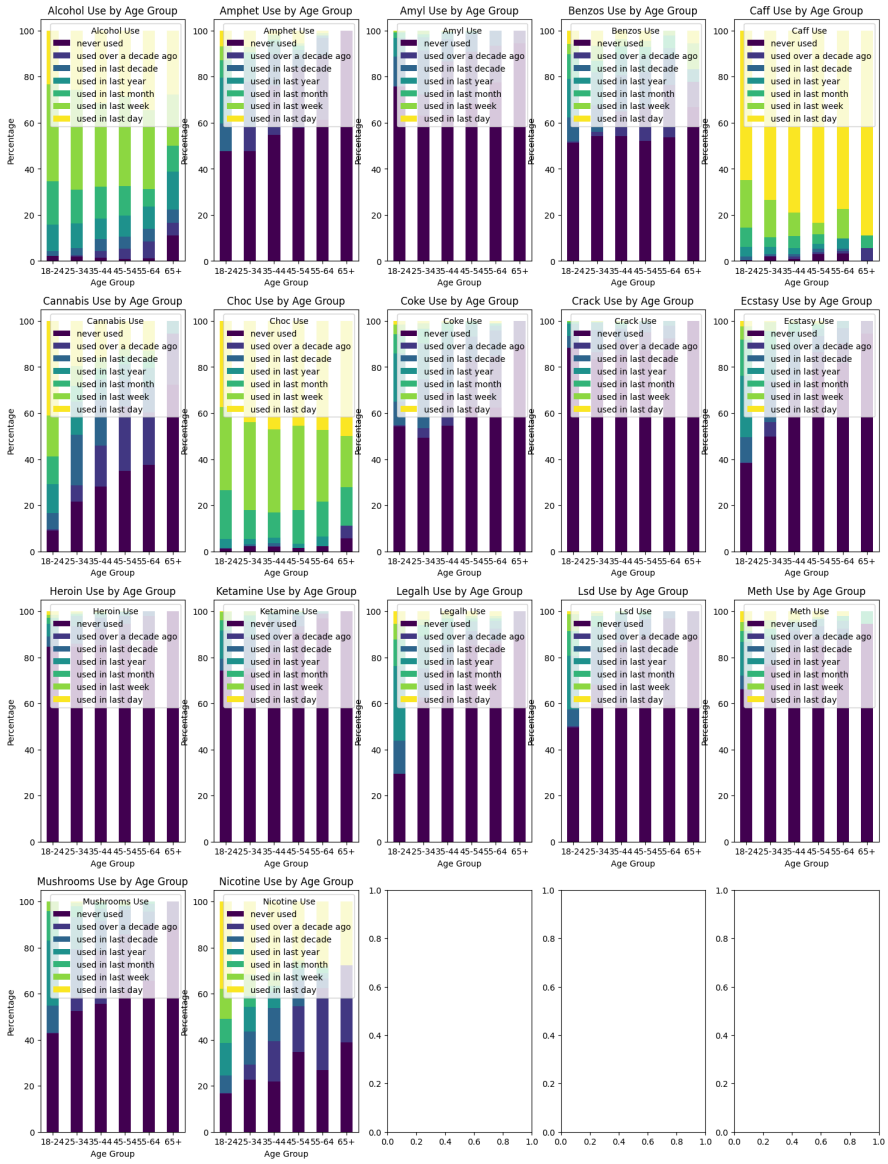


Figure 0

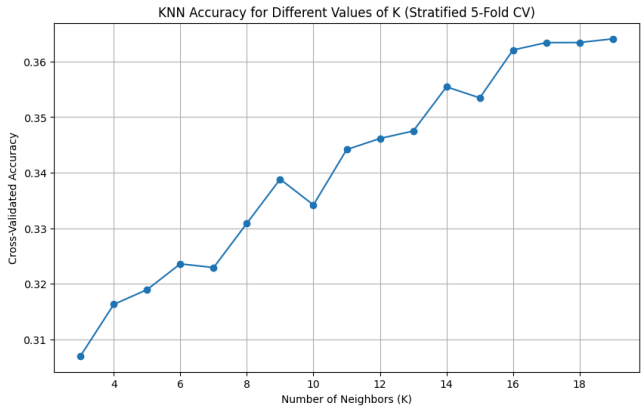


Figure 1

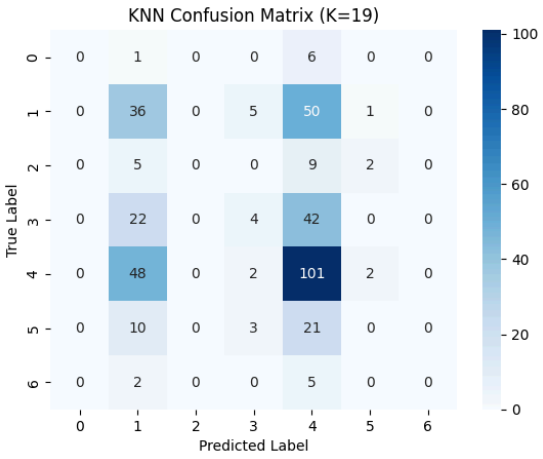


Figure 2

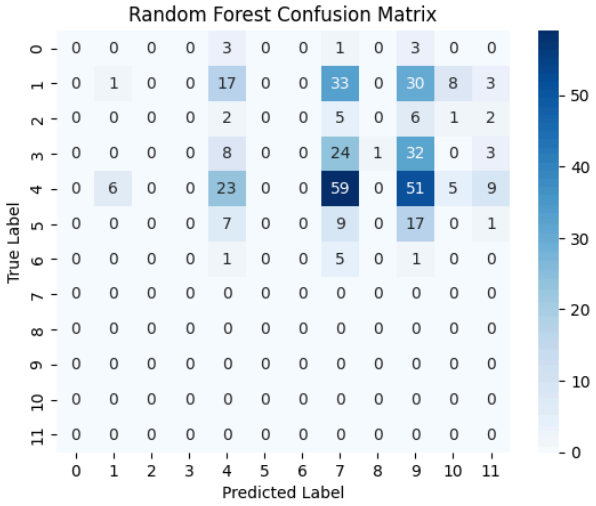


Figure 3

5.4. Question 4 (Talha)

Question: How does the usage of one drug correlate with the usage of other drugs?

Explanation: This question aims to explore the relationships between the usage patterns of various drugs. Understanding these correlations can reveal how the usage of one drug might be associated with the usage of others, providing insights into common substance use patterns. This is relevant for identifying potential clusters of substance use and for informing targeted interventions. It also helps to understand whether some drugs are commonly used together, which might be influenced by social, cultural, or behavioral factors.

Design

1. **Data Used:**

   Dataset: Drug Consumption (quantified) dataset from the UCI Machine Learning Repository.

Features: Usage data of 18 different drugs, personality traits, and demographic attributes.

Target Variable for Predictive Modeling: Usage of cannabis.

2. **Data Preparation Steps:**

    1. **Conversion of Categorical Data:** Transformed categorical drug usage responses into numerical values, where each response category (e.g., "Never Used," "Used in Last Month") was mapped to a numerical scale (0 to 6).

    2. **Normalization:** Normalized continuous variables to bring them onto a comparable scale, which is necessary for effective model training and evaluation.

    3. Feature Selection: Focused on drug usage columns and excluded the target drug (cannabis) from the features set in the predictive model.

Technical Approach:

1. Correlation Analysis:

- Constructed a correlation matrix to evaluate how the usage of one drug is related to the usage of others.

- Used a heatmap for visualization to identify strong and weak correlations between drug usages.

2. Predictive Modeling:

-Model Used: Logistic Regression.

-Objective: Predict cannabis usage based on the usage of other drugs.

-Evaluation Metrics: Accuracy, cross-validation scores, and a classification report to assess the model's performance in predicting cannabis usage.

Implementation

1. **Correlation Analysis:**

-Generated a correlation matrix using numerical drug usage data.

- Visualized the matrix with a heatmap to highlight the relationships between different drugs.

2. **Predictive Modeling:**
   - Split the data into training and testing sets using a 70/30 split.

   - Trained a logistic regression model on the training set to predict cannabis usage.

   - Evaluated the model using cross-validation and tested its performance on the test set.

Results

1- Correlation Analysis:

- The heatmap of the correlation matrix (Figure 1) revealed varying degrees of correlation between drug usages.

- Drugs like cannabis, alcohol, and cocaine showed relatively higher correlations with each other, indicating that users of these substances often use them concurrently.

- Conversely, some drugs, such as heroin and cannabis, showed lower correlations, suggesting more independent usage patterns.

2- Predictive Modeling:

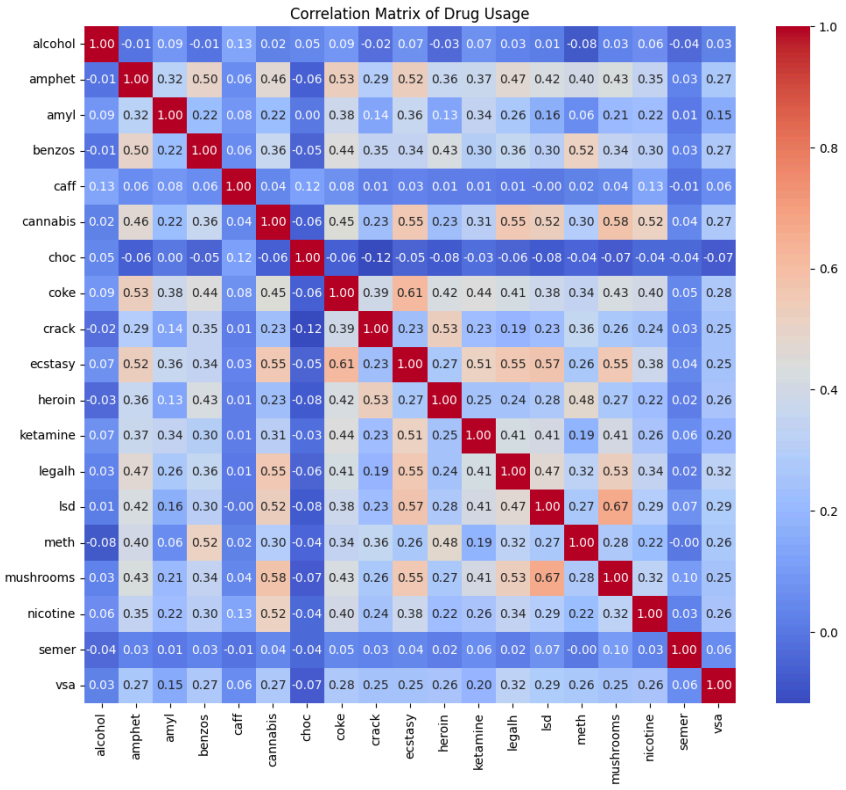
- Cross-Validation Scores: The model achieved an average cross-validation score of approximately 43%, reflecting modest performance.

- Accuracy: The logistic regression model had an accuracy of about 46% on the test set.

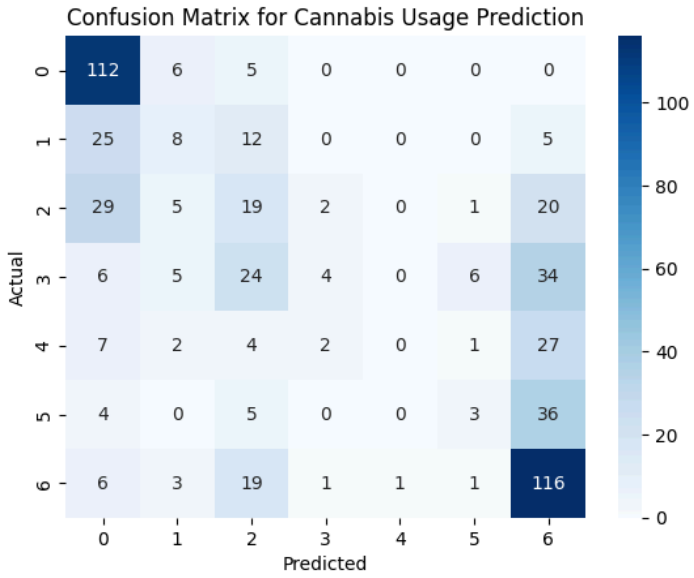
- Classification Report: Detailed metrics (Table 1) showed high precision and recall for non-users but lower performance in predicting other levels of cannabis usage.

3- Visualizations:

- Figure 1: Correlation Heatmap of Drug Usage. This heatmap visualizes the correlations between the usage of different drugs. Strong correlations are indicated by darker colors, while weaker correlations are shown in lighter colors.



- Table 1: Classification Report for Cannabis Usage Prediction. This table summarizes the model's precision, recall, and F1-score for each class of cannabis usage.



Discussion of Findings:

Correlation Analysis:

The correlation analysis revealed varying degrees of association between drug usage levels. For example, cannabis, alcohol, and cocaine showed high correlations with each other, indicating that users of one are likely to use the others. In contrast, heroin and cannabis had little to no correlation, suggesting independent usage patterns. These results reflect common patterns in substance use, where social and cultural factors influence concurrent usage. However, correlation does not imply causation, so results should be interpreted with caution.

Predictive Modeling:

The logistic regression model aimed to predict drug usage (e.g., cannabis) based on other drugs but achieved only 46% accuracy, with cross-validation scores around 43%. The model performed well in predicting non-users (class 0) but struggled with other classes, particularly moderate to heavy users (classes 1-6). Low F1-scores for these classes indicate challenges in distinguishing between different levels of usage. This difficulty may stem from overlapping usage patterns or the model's inability to capture complex relationships between drug use behaviors.

Answer to the Original Question:

The correlation analysis effectively identified relationships between drug usages, with some drugs showing strong associations. Predictive modeling, however, faced challenges, suggesting that predicting drug usage based on others is complex and requires more advanced models or additional features to improve accuracy. Overall, while certain drugs are commonly used together, predicting individual drug usage remains difficult due to the multifaceted nature of substance use behavior. These insights provide context for understanding drug usage patterns and could guide further research or interventions.