Indian Institute of Technology Kanpur

"Influence of Personality Traits on Drug Use Behavior: A Comprehensive Analysis "

Submitted by:

Soumen konai (231080091)

Arnab Das (231080020)

Samrat Halder (231080074)

Souhardya Mitra (231080090)

Sujash Krishna Basak (231080093)

Supervised by:

Dr. Amit Mitra

Department of Mathematics and Statistics,IIT Kanpur

Abstract

This study dives into how different personality traits, like being anxious, outgoing, open-minded, kind, or organized, relate to the use of substances. By analyzing data gathered by Elaine Fehrman, we want to figure out which traits can predict substance use across various substances. This investigation can shed light on how our minds work when it comes to using substances, helping us create interventions that speak to specific traits.

To kick things off, we'll use logistic regression to unpack how each trait directly influences substance use, while keeping demographic factors in mind. Then, we'll bring in more advanced tools like Random Forest, XGBoost, and Artificial Neural Networks (ANNs) to sniff out any tricky, non-linear relationships or interactions. By pitting these models against each other, we can pinpoint the best method for predicting substance use based on someone's personality makeup.

Before we dig deep into the data, we will start by doing some Exploratory Data Analysis (EDA). This step will let us peek into how demographic factors like age and gender play into substance use, as well as uncover connections between different personality traits and types of substances used. This bird's-eye view will help us fine-tune our models and unearth key patterns in the data.

When all is said and done, uncovering the personality traits that link to risky substance use can pave the way for creating personalized prevention strategies and interventions. Our goal is to get a better grip on how personalities shape our health choices, adding new layers to substance use prevention efforts, and offering practical advice tailored to different groups of people.

Acknowledgment

Our journey in accomplishing this project has been greatly enriched by the support and guidance of many individuals, to whom we are deeply grateful. We extend our heartfelt appreciation to Dr. Amit Mitra, from the Department of Mathematics and Statistics at IIT Kanpur, for entrusting us with this project and providing invaluable mentorship throughout its course.

This experience has not only been an extraordinary learning opportunity but has also allowed us to apply theoretical knowledge to practical analysis, enhancing our understanding of nonparametric statistical methods.

We would also like to express our sincere gratitude to our friends for their unwavering support throughout this endeavor. Their encouragement and motivation were instrumental in enabling us to complete this project within the stipulated timeframe.

Date: 14.11.2024

Contents

Influence of Personality Traits on Drug Use Behavior: A Comprehensive Analysis

1 Introduction

Drugs come in many forms and can affect our bodies in different ways. When we hear "drugs," we often think of illegal ones, but there are also medicines that help us when we're sick and substances people use for fun. Medicines are there to make us better, but recreational drugs can lead to both mental and physical effects that stick around long after we've taken them. The fun side of drugs can be risky, not just for individuals but for our communities too, as it can influence behaviors that increase the chances of getting sick or hurt. How people use drugs around the world is shaped by different things like how they think and feel, who they spend time with, where they live, and how much money they have.

Our personalities, made up of traits like being anxious, outgoing, open-minded, kind, or organized, play a big role in how we interact with drugs. Some studies suggest that people who are more anxious or less organized might be more drawn to drugs. Anxious folks might turn to drugs to escape bad emotions, while less

organized people might have a hard time saying no to risky stuff. Being outgoing can also make someone more likely to try drugs for the excitement or connection it brings. People who are less kind and understanding might consume more alcohol or marijuana. When we look closely, we see that those who are less kind and organized tend to drink more and smoke more marijuana, while those who are very anxious might seek out harder drugs like cocaine or heroin. Factors like how much money someone has can also change how personality traits affect drug use, with wealthier folks who are less organized possibly getting involved in illegal drugs.

Understanding how personality traits and drug usage are connected is important for public health efforts. By knowing which traits raise the chances of using drugs, we can create better ways to help prevent drug problems and offer treatments that suit individuals better. By recognizing how our personalities influence drug use, public health initiatives can be more personal and effective, benefiting both individuals and the communities they live in.

2 Dataset Overview

The dataset was collected by Elaine Fehrman between March 2011 and March 2012. The research proposal was approved by the University of Leicester's Forensic Psychology Ethical Advisory Group in January 2011, and subsequently received favorable opinion from the University of Leicester School of Psychology's Research Ethics Committee (PREC).

The study recruited a total of 2051 participants over a 12-month period. Out of these, 166 participants did not respond correctly to a built-in validity check, and 9 participants were found to have endorsed using a fictitious recreational drug

(Semeron), which was included to identify over-claimers. After excluding these respondents, the final usable sample consisted of 1885 participants (943 males, 942 females).

The snowball sampling methodology recruited a primarily native English-speaking sample (93.7%), with participants from various countries:

- UK: 55.4% (1044 participants)
- USA: 29.5% (557 participants)
- Canada: 4.6% (87 participants)
- Australia: 2.9% (54 participants)
- New Zealand: 0.3% (5 participants)
- Ireland: 1.1% (20 participants)
- Other countries: 6.3% (118 participants), with no single country comprising more than 1% of the total sample.

3 Demographics

Participants reported their age in the form of an age band rather than their exact age:

- 18-24 years: 34.1% (643 participants)
- 25-34 years: 25.5% (481 participants)
- 35-44 years: 18.9% (356 participants)

- 45-54 years: 15.6% (294 participants)
- 55-64 years: 4.9% (93 participants)
- Over 65 years: 1% (18 participants)

The sample was highly educated, with nearly two-thirds of participants (59.5%) educated to at least a degree or professional certificate level:

- Professional certificate or diploma: 14.4% (271 participants)
- Undergraduate degree: 25.5% (481 participants)
- Master's degree: 15% (284 participants)
- Doctorate: 4.7% (89 participants)
- Some college/university (without certificate): 26.8% (506 participants)
- Left school at 18 or younger: 13.6% (257 participants)

In terms of ethnicity, the overwhelming majority (91.2%) of participants identified as white:

- White: 91.2% (1720 participants)
- Black: 1.8% (33 participants)
- Asian: 1.4% (26 participants)
- Other/Mixed: 5.6% (106 participants)

Due to the small number of participants from non-white ethnic groups, no analyses involving racial categories were performed.

4 Variables in the Dataset

Each record in the dataset corresponds to a participant and contains the following 12 attributes:

1. ID
2. Age (Age band)
3. Gender (Male/Female)
4. Education (Level of education)
5. Country (Country of residence)
6. Ethnicity (Cultural background)
7. Nscore (Neuroticism score)
8. Escore (Extraversion score)
9. Oscore (Openness to Experience score)
10. Ascore (Agreeableness score)
11. Cscore (Conscientiousness score)
12. Impulsivity (Measure of impulsiveness)
13. Sensation Seeking (SS) (Measure of sensation seeking behavior)

5 Drug Use Variables

Participants were also questioned regarding their use of 18 legal and illegal drugs, including:

- Alcohol
- Amphetamines
- Amyl nitrite
- Benzodiazepine
- Cannabis
- Chocolate
- Cocaine
- Caffeine
- Crack
- Ecstasy
- Heroin
- Ketamine
- Legal highs
- LSD
- Methadone

- Mushrooms
- Nicotine
- Volatile substance abuse
- Semeron (a fictitious drug to identify over-claimers)

For each drug, participants had to choose from one of the following categories:

- Never used
- Used over a decade ago
- Used in the last decade
- Used in the last year
- Used in the last month
- Used in the last week
- Used in the last day

This results in 18 classification problems, where each label variable corresponds to one drug and contains seven classes.

6 Motive of the Project

Understanding the Link Between Personality Traits and Drug Use:

- The project aims to explore how personality traits, based on the Five Factor Model (Neuroticism, Extraversion, Openness, Agreeableness, and Conscientiousness), influence an individual's propensity to use legal or illegal drugs.

Prediction of Drug Use Based on Behavioral Traits:

- By analyzing personality scores (Nscore, Escore, Oscore, Ascore, and Cscore), the project seeks to develop predictive models that can classify individuals' likelihood of drug use based on their personality profiles.

Identifying Risk Factors for Substance Abuse:

- The project investigates which personality traits or demographic factors (e.g., age, gender, education) are associated with higher risks of substance abuse, including both legal substances (like alcohol or nicotine) and illegal drugs (like cocaine or heroin).

Exploring the Role of Impulsivity and Sensation Seeking in Drug Use:

- This project aims to analyze how impulsivity and sensation-seeking behaviors contribute to drug consumption, providing insights into the behavioral mechanisms behind substance abuse.

Comparing Drug Use Patterns Across Demographics:

- By leveraging the diversity of the dataset, the project will examine how drug use patterns vary across different countries, age groups, and educational levels, helping to identify socio-cultural factors that influence substance use.

Contribution to Public Health and Drug Prevention:

- The findings from this study could provide valuable information for public health initiatives, helping to create targeted intervention programs that reduce drug consumption based on psychological and demographic risk profiles.

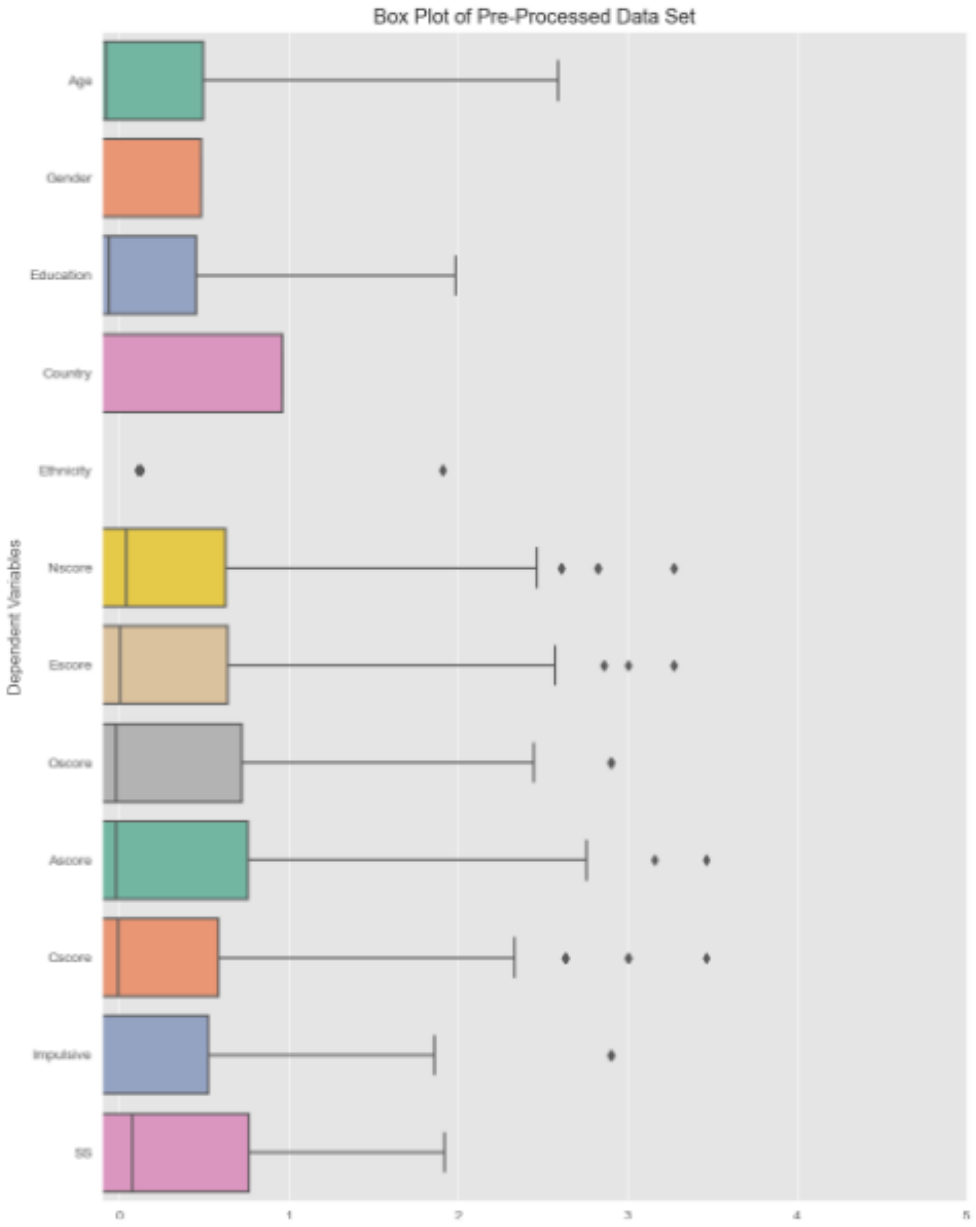
Behavioral Insights for Addiction Counseling:

- The results may assist in the development of more effective counseling strategies by understanding the underlying personality traits associated with different types of substance use, contributing to addiction treatment approaches.

7 Exploratory Data Analysis (EDA)

7.1 Pre-Processed Data Set

- Below is the box plot of the pre-processed data set:

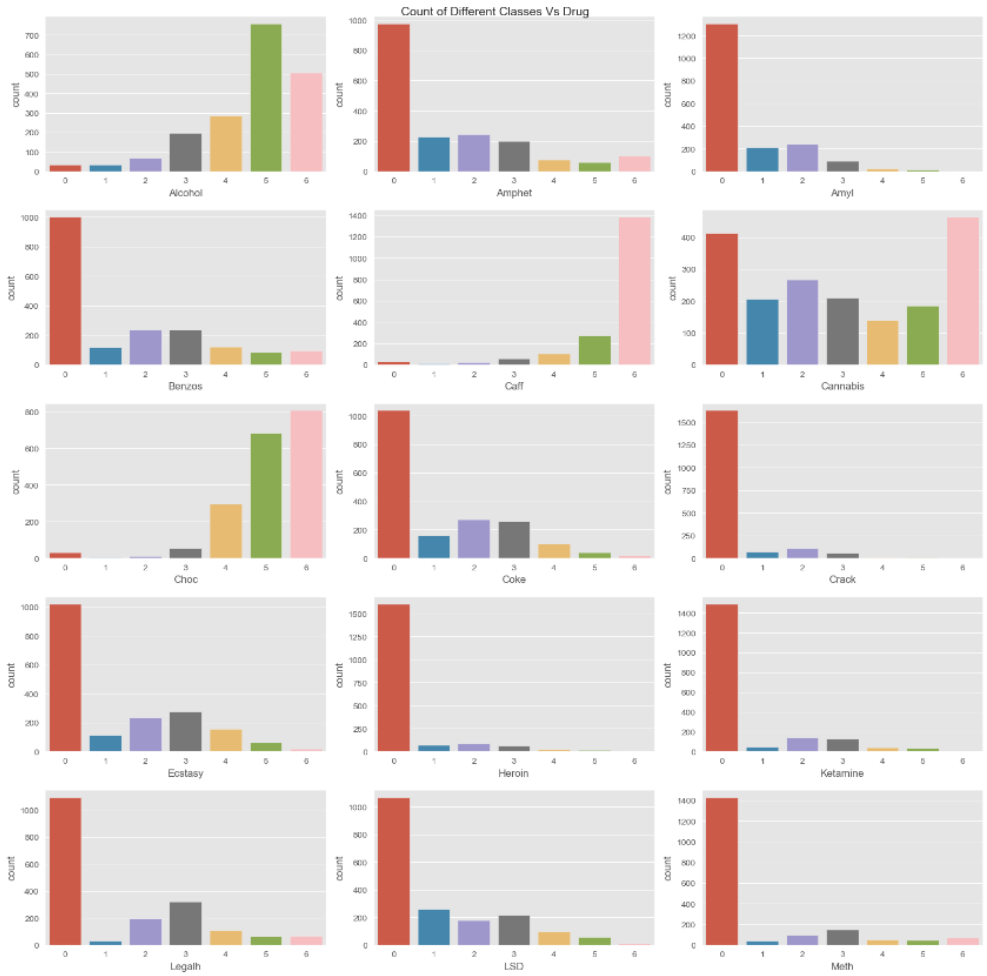


The box plot provides a visual summary of various variables in a dataset. Each box captures the middle 50 % of the data, with the line inside showing the median. The "whiskers" of each box extend to the smallest and largest values within 1.5 times the box's range, while any values beyond the whiskers are shown individually as outliers.

For example, the **Age** variable has a roughly normal distribution, with a median around 0.5 to 1, suggesting a balanced spread of ages. Meanwhile, the **Nscore** variable is right-skewed, with a noticeable cluster of outliers, indicating that some people score significantly higher than most. Observing these patterns helps in selecting suitable statistical methods for further analysis, as different distributions may call for different approaches.

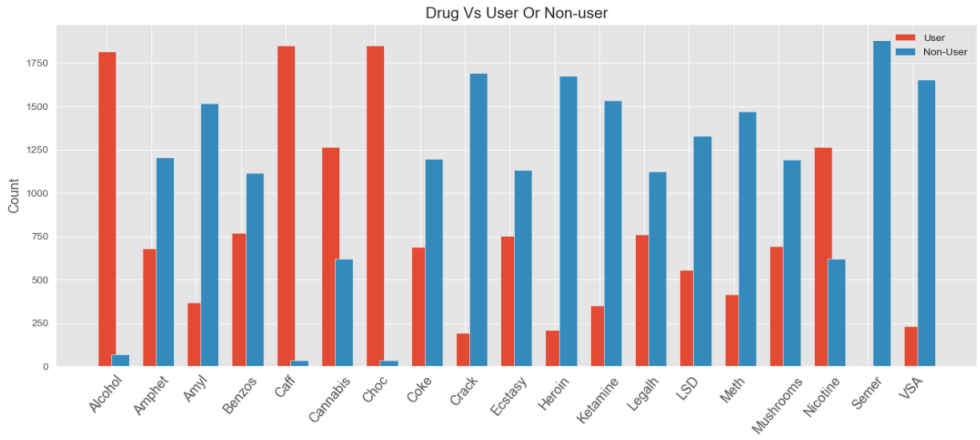
7.2 Count of Different Classes Vs Drugs

- Below is the box plot of Count of Different Classes Vs Drugs:



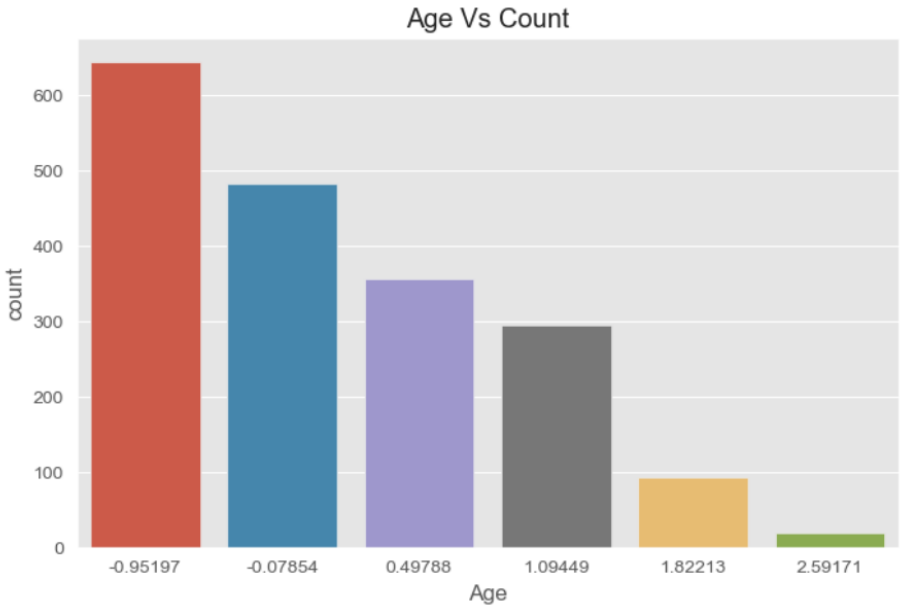
7.3 Count of Drug Vs User Or Non-user

- Below is the Count plot of Drug Vs User Or Non-user:



7.4 Age vs Count

- Below is the plot of Age vs Count:

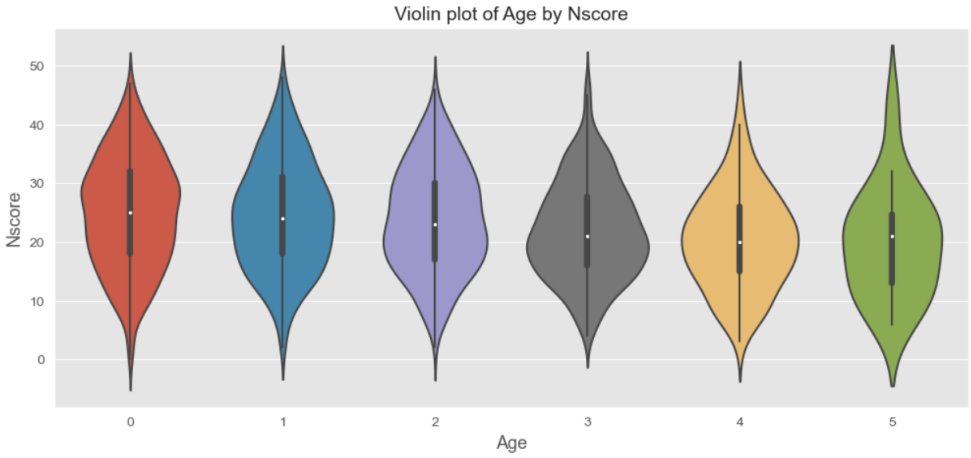


The plot shows the distribution of age in a dataset. The x-axis represents the age range, and the y-axis represents the count of individuals within each age range. The plot reveals that the majority of the individuals fall within the age range of

-0.95197. The count gradually decreases as the age range increases. This suggests that the dataset contains individuals who are mostly clustered around the lower end of the age spectrum.

7.5 Age by Nscore

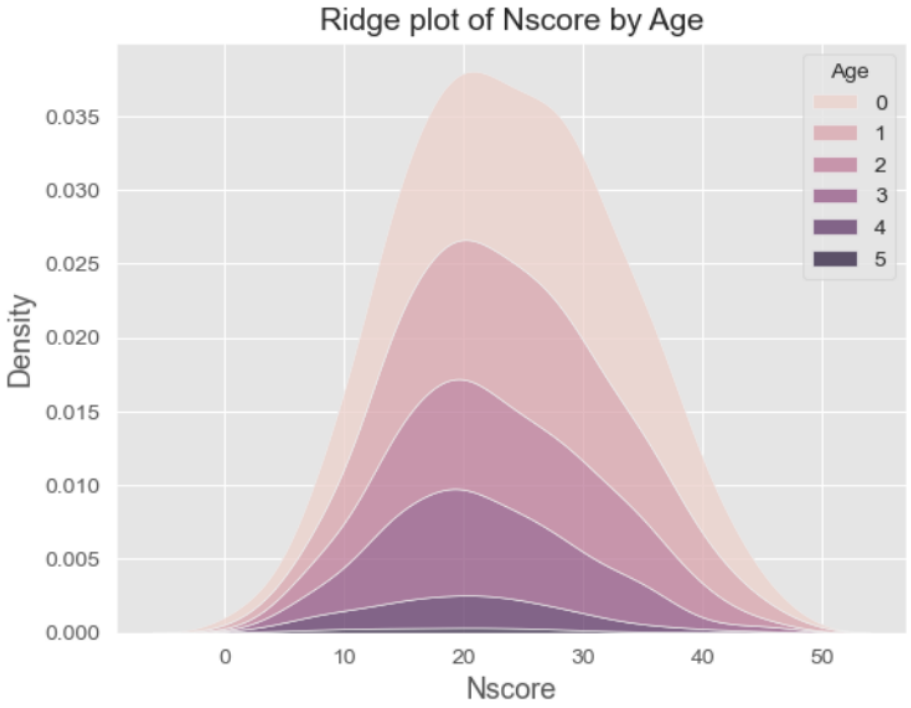
- Below is the Violin plot of Age by Nscore:



The violin plot shows the distribution of Nscore for different Age groups. We can see that the distribution of Nscore is quite different for each Age group. For example, the Nscore for Age group 0 is generally lower than the Nscore for Age group 1. The plot also shows that the spread of Nscore is wider for Age group 5 than for other groups. This indicates that there is more variability in Nscore for Age group 5. Overall, the violin plot provides a useful visualization of the relationship between Age and Nscore.

7.6 Ridge plot of Nscore by Age

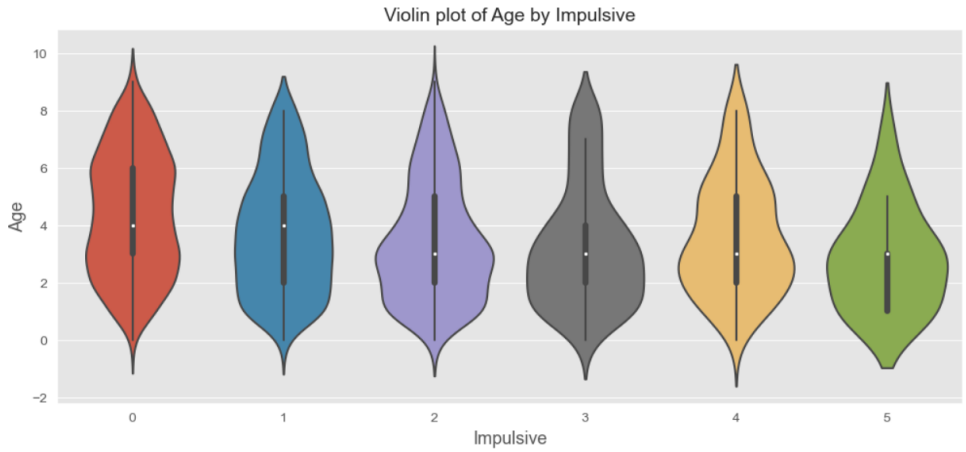
- Below is the Ridge plot of Nscore by Age:



The ridge plot displays the distribution of Nscore values across different age groups. Each colored line represents a different age group (0 to 5 in this case),The Nscore is highest for the youngest age group (age 0) and decreases as age increases. The distribution of Nscore is bimodal for the youngest age group, with a peak around 20 and another peak around 30. The distribution of Nscore becomes unimodal for older age groups, with the peak shifting towards lower Nscore values. This suggests that Nscore is negatively associated with age.

7.7 Age By Impulsive

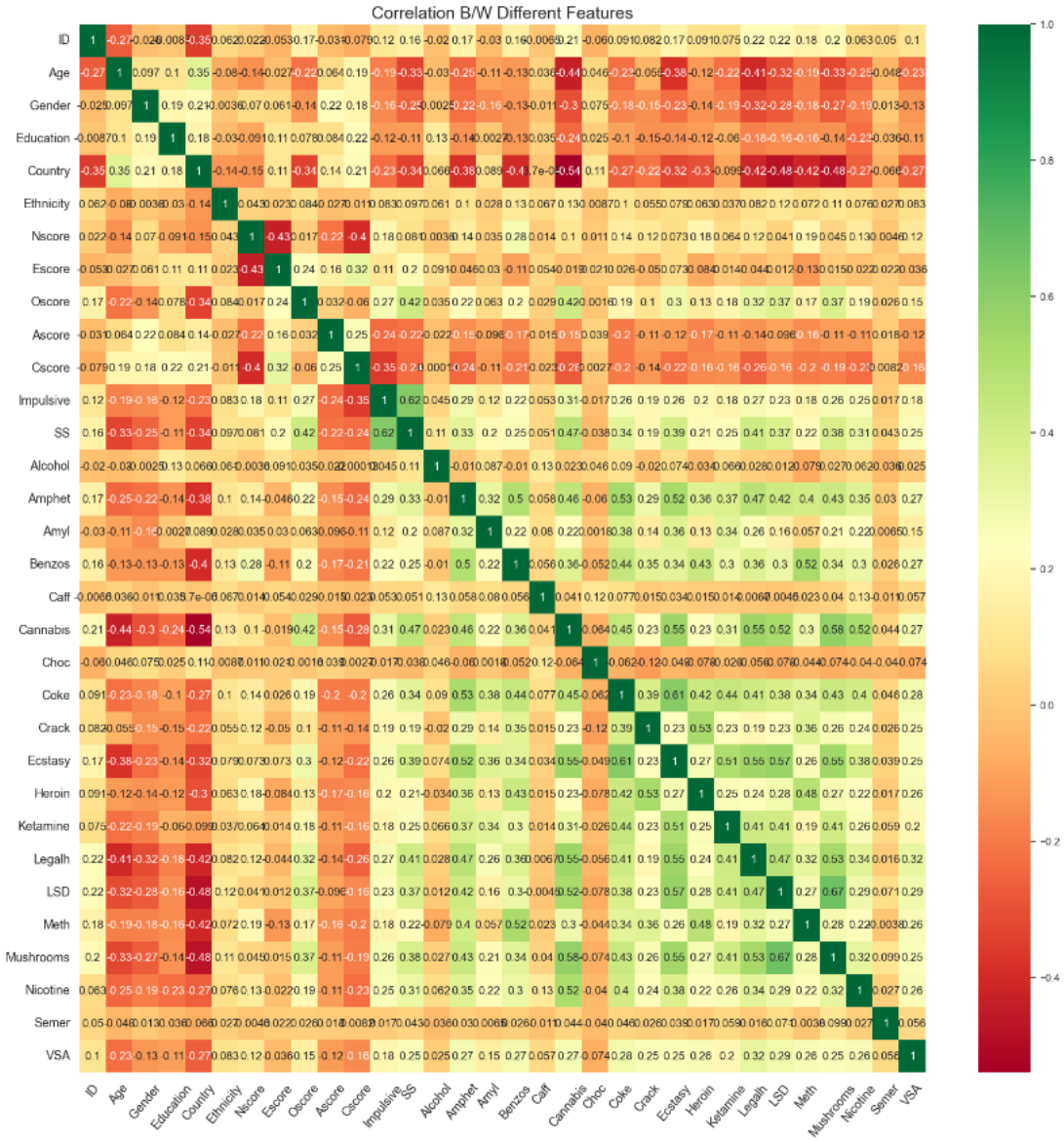
- Below is the Violin Plot of Age By Impulsive:



The violin plot shows the distribution of age across different levels of impulsivity. We can see that there is a general trend of increasing age with increasing impulsivity. The distribution of age is also wider for higher levels of impulsivity. However, we should be careful to make definitive conclusions as the spread of the data is very wide, and more evidence is needed to confirm whether or not this trend is statistically significant.

7.8 Correlation B/W Different Features

- Below is the Heat map of Different Features:



The heatmap displays the correlation matrix between different features (variables) in the dataset. The color intensity represents the strength of the correlation, with red indicating positive correlation (variables move in the same direction), blue indicating negative correlation (variables move in opposite directions), and white indicating no correlation.

"Unveiling Drug Use Patterns through Personality Analysis: Leveraging AI and Machine Learning Techniques"

8 Assessing AI models involves understanding their explainability and interpretability.

8.1 Classification Performance Matrix

In evaluating classification models, a confusion matrix is frequently used to analyze performance, especially in binary classification tasks. The confusion matrix for a binary classifier is structured as follows:

	Predicted Positive	Predicted Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

where:

- TP (True Positive): Instances that are truly positive and correctly predicted as positive by the model.
- TN (True Negative): Instances that are truly negative and correctly predicted as negative.
- FP (False Positive): Instances that are actually negative but are incorrectly predicted as positive (Type I error).
- FN (False Negative): Instances that are actually positive but are incorrectly predicted as negative (Type II error).

8.1.1 Precision

Precision indicates how many of the model's positive predictions are correct. It is calculated as the ratio of true positive predictions to the total positive predictions made by the model:

$$\text{Precision} = \frac{TP}{TP + FP}$$

8.1.2 Recall

Recall (also called sensitivity or true positive rate) measures the model's effectiveness in identifying all actual positives. It is defined as the ratio of true positive predictions to the total actual positive instances:

$$\text{Recall} = \frac{TP}{TP + FN}$$

8.1.3 F1-Score

The F1-score is the harmonic mean of precision and recall, providing a single measure that balances both. It is given by:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

8.1.4 Accuracy

Accuracy measures the proportion of correct predictions out of all predictions made:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

This metric indicates the overall performance of the model in terms of correctly classified instances.

8.1.5 Macro Average

Macro averaging computes the metric independently for each class and then averages the results. For multi-class problems, precision, recall, and F1-score can be calculated as:

$$\text{Macro Average} = \frac{1}{N} \sum_{i=1}^{N} \text{Metric}_i$$

where N is the number of classes.

8.1.6 Weighted Average

Weighted averaging calculates the metric for each class and then takes a weighted average, where each class's weight is proportional to its support (i.e., the number of actual instances for that label). For precision, recall, and F1-score, this is computed as:

$$\text{Weighted Average} = \frac{\sum_{i=1}^{N} \text{Metric}_i \times \text{Support}_i}{\sum_{i=1}^{N} \text{Support}_i}$$

where N is the number of classes.

9 Multinomial Logistic Regression(When the response variable has more than two categories present)

9.1 Introduction

Multinomial logistic regression is an extension of binary logistic regression that allows for the prediction of categorical outcomes with more than two categories. It is commonly used when the dependent variable has multiple unordered categories. Unlike binary logistic regression, which is used for binary classification tasks, multinomial logistic regression can handle multiple classes simultaneously.

9.2 Overview

In multinomial logistic regression, the dependent variable Y can take on K different categories. The goal is to model the probabilities of each category given the values of the independent variables $X_1, X_2, \ldots, X_k$. The model estimates the probability of each category relative to a reference category, which is typically chosen arbitrarily. The probabilities for all categories sum up to 1 for each observation.

9.3 Mathematical Formulation

The multinomial logistic regression model is formulated using the softmax function, which generalizes the logistic function for multiple categories. The softmax function is defined as follows:

$$P(Y = k|X) = \frac{e^{\beta_{0k}+\beta_{1k}X_1+...+\beta_{pk}X_p}}{\sum_{j=1}^{K} e^{\beta_{0j}+\beta_{1j}X_1+...+\beta_{pj}X_p}}$$

where:

- $P(Y = k|X)$ is the probability of the dependent variable being in category k given the values of the independent variables $X_1, X_2, \ldots, X_p$.
- $\beta_{0k}, \beta_{1k}, \ldots, \beta_{pk}$ are the coefficients associated with category k.
- K is the total number of categories.

The softmax function ensures that the predicted probabilities sum up to 1 across all categories for each observation.

9.4 Parameter Estimation

Similar to binary logistic regression, the parameters of the multinomial logistic regression model are estimated using Maximum Likelihood Estimation (MLE). The likelihood function for multinomial logistic regression is the product of the probabilities of observing the given outcomes given the predictor variables and model parameters.

The log-likelihood function is then obtained by taking the natural logarithm of the likelihood function. The goal is to find the values of $\beta_{0k}, \beta_{1k}, \ldots, \beta_{pk}$ that maximize the log-likelihood function. This optimization problem is typically solved using numerical optimization algorithms such as gradient descent or Newton-Raphson method.

9.5 Advantages of Multinomial Logistic Regression

- Allows for the prediction of categorical outcomes with more than two categories.
- Provides interpretable coefficients representing the effect of each independent variable on the probability of each category relative to the reference category.
- Can handle multicollinearity among independent variables.

9.6 Why Multinomial Logistic Regression

Multinomial logistic regression was chosen for our project because:

- Our dependent variable has more than two unordered categories.
- We are interested in understanding the influence of multiple independent variables on the probabilities of each category.
- Multinomial logistic regression provides a flexible and interpretable framework for modeling and predicting categorical outcomes with multiple categories.

9.7 Evaluation of the outcomes

Based on the accuracy results from the logistic regression model applied to the drug dataset, we can draw several insights. The accuracy levels vary significantly across different substances, highlighting differences in model performance. The highest accuracy is achieved for "Semer" at 99.73 %, indicating that the model can predict this drug usage pattern with high reliability. Other drugs with relatively strong model performance include "Coke" at 88.59 %, "Heroin" at 85.94 %, "Ketamine"

at 81.70 %, and "VSA" at 76.39 %. These accuracies suggest the model is effective in capturing usage trends for these substances.

In contrast, certain drugs, such as "Nicotine" (39.26%), "Alcohol" (39.52%), and "Cannabis" (41.64%), yield lower accuracy scores, indicating challenges in accurately classifying usage patterns for these substances. This could be due to overlapping usage categories or greater diversity in consumption patterns among users.

Overall, the model demonstrates reasonable accuracy across multiple drug categories but may benefit from further refinement or additional features to improve predictive accuracy for substances with lower scores.

10 Decision Trees

Decision Trees are flexible, interpretable models used in supervised learning for both classification and regression. They work by dividing the feature space into distinct regions through a series of binary splits based on feature values. Each node in a decision tree corresponds to a feature and split, guiding the data down to leaf nodes, which provide the final prediction. Decision trees are widely used because of their simplicity, interpretability, and their ability to handle both continuous and categorical data.

10.1 Components of Decision Trees

10.1.1 Root Node

The root node represents the entire dataset and initiates the partitioning process. At this step, the algorithm selects a feature that maximizes information gain for classification or minimizes variance for regression.

10.1.2 Internal Nodes

Internal nodes represent features and their thresholds, dividing the dataset into subsets. The feature and threshold that best reduce impurity are chosen at each node. Common impurity measures include Gini impurity and entropy for classification and variance for regression.

10.1.3 Leaf Nodes

Leaf nodes signify the output prediction (either class or value). Once a sample reaches a leaf, the decision tree provides the associated prediction.

10.1.4 Splitting Criteria

The splitting criteria aim to reduce impurity or increase information gain. Impurity measures like Gini and entropy quantify uncertainty, helping the algorithm find optimal splits.

10.2 Mathematical Expressions

10.2.1 Gini Impurity

Gini impurity evaluates the purity of a dataset D with C classes:

$$\text{Gini}(D) = 1 - \sum_{i=1}^{C} p_i^2$$

where p_i is the proportion of instances in class i. Gini values range from 0 (pure) to 1 (maximally impure).

10.2.2 Information Gain

Information gain calculates the reduction in impurity by splitting based on a feature A with possible values $\{v_1, v_2, \ldots, v_k\}$:

$$\text{IG}(D, A) = \text{Impurity}(D) - \sum_{j=1}^{k} \frac{|D_{v_j}|}{|D|} \cdot \text{Impurity}(D_{v_j})$$

where $|D_{v_j}|$ is the number of samples in D for which feature A has value v_j.

10.2.3 Decision Rule

A decision tree's decision rule is a series of conditional splits based on feature values. For example:

$$\text{if feature1} \leq \text{threshold1}:$$

$$\text{if feature2} \leq \text{threshold2}:$$

$$\text{return class1}$$

$$\text{else: return class2}$$

$$\text{else: return class3}$$

This rule guides the tree in making predictions.

10.3 Analysis of Classification Performance

```
--------------------------------------------------
Evaluating Decision Tree for LSD...
Performance for LSD:
Accuracy: 0.4399
Confusion Matrix:
[[196  47  30  29   9   9   1]
 [ 37  21   8   8   2   0   1]
 [ 20  11   8   7   4   4   1]
 [ 22   4   7  18   7   8   2]
 [ 11   1   3   7   5   2   0]
 [  3   1   1   6   1   1   0]
 [  1   1   0   1   0   0   0]]
Classification Report:
              precision    recall  f1-score   support

         CL0       0.68      0.61      0.64       321
         CL1       0.24      0.27      0.26        77
         CL2       0.14      0.15      0.14        55
         CL3       0.24      0.26      0.25        68
         CL4       0.18      0.17      0.18        29
         CL5       0.04      0.08      0.05        13
         CL6       0.00      0.00      0.00         3

    accuracy                           0.44       566
   macro avg       0.22      0.22      0.22       566
weighted avg       0.47      0.44      0.45       566

ROC AUC: 0.5889
```

Figure 1: Classification Report

After applying the Decision Tree Classifier on our classification problem, it was observed that the model performed poorly with certain drugs, such as Alcohol, Amyl, Amphet, Benzos, LSD, Caffeine, and Heroin. The model showed low accuracy across these drugs. We suspect that the imbalance in the number of users for various drugs and the complex interaction patterns among drug users could be contributing to the high variance in performance metrics. The AUC-ROC values for these drugs were also low, lying in the range of 50-60%, indicating that the

model's performance was close to random guessing. Consequently, we decided to explore ensemble methods, specifically a Random Forest, in our study.

11 Random Forest

Random Forest is an ensemble method that aggregates multiple decision trees to enhance performance and mitigate overfitting. It trains multiple trees on different data samples and averages their predictions (regression) or uses majority voting (classification).

11.1 Decision Trees

11.1.1 Construction

At each tree node:

- A subset of features is randomly chosen.
- The feature and threshold with optimal information gain (classification) or minimum impurity (regression) are selected.
- The data splits recursively until a stopping condition is met (e.g., maximum depth).

11.1.2 Mathematical Expressions

For classification:

$$\text{Gini impurity:} \quad G(t) = 1 - \sum_{i=1}^{C} p_i^2$$

$$\text{Entropy:} \quad H(t) = -\sum_{i=1}^{C} p_i \log_2(p_i)$$

$$\text{Information gain:} \quad IG(t) = H(\text{parent}) - \sum_{j \in \text{children}} \frac{N_j}{N} H(j)$$

For regression:

$$\text{MSE}(t) = \frac{1}{N} \sum_{i=1}^{N} (y_i - \bar{y}_t)^2$$

11.2 Bootstrap Sampling

Bootstrap sampling creates multiple datasets by randomly sampling from the original data with replacement. A sample's probability of exclusion in a bootstrap dataset approaches $\frac{1}{e}$ as the sample size increases.

11.3 Random Feature Selection

A subset of features is randomly selected at each node, introducing randomness and helping prevent overfitting. Typically, this subset size, m, is set to $\sqrt{M}$ or $\log_2(M)$.

11.4 Voting/Averaging

Final predictions are obtained by aggregating the predictions of individual trees. For classification, the most frequent class is chosen, and for regression, the average

prediction is used.

11.5 Implementation of Random Forest Classifier

```
Performance for LSD:
Accuracy: 0.5936
Confusion Matrix:
[[287   5   2  24   2   1   0]
 [ 51  19   3   4   0   0   0]
 [ 36   6   4   9   0   0   0]
 [ 37   2   1  24   3   1   0]
 [ 16   1   0  10   2   0   0]
 [  8   0   1   4   0   0   0]
 [  1   0   0   2   0   0   0]]
Classification Report:
              precision    recall  f1-score   support

         CL0       0.66      0.89      0.76       321
         CL1       0.58      0.25      0.35        77
         CL2       0.36      0.07      0.12        55
         CL3       0.31      0.35      0.33        68
         CL4       0.29      0.07      0.11        29
         CL5       0.00      0.00      0.00        13
         CL6       0.00      0.00      0.00         3

    accuracy                           0.59       566
   macro avg       0.31      0.23      0.24       566
weighted avg       0.54      0.59      0.53       566

ROC AUC: 0.7818
--------------------------------------------------
```

Figure 2: Classification Report

After implementing a Random Forest Classifier, we noticed an improvement in the classification accuracy for each drug. However, the issue of unequal support across classes remained in the dataset. As a result, significant increases in accuracy were observed only for Heroin, Amyl, Caffeine, and LSD, with each achieving approximately 75% accuracy. Additionally, the AUC-ROC for all drugs increased substantially, suggesting that the bagging technique used in Random Forests was able to capture the variability in response at various classification thresholds.

12 XGBoost

XGBoost (eXtreme Gradient Boosting) is a powerful boosting algorithm that builds a model by sequentially adding trees that correct previous errors.

12.1 Decision Trees in XGBoost

XGBoost uses decision trees as weak learners in a boosting framework.

12.1.1 Tree Construction

At each iteration:

- A new tree minimizes residual errors from previous predictions.
- The split minimizes the loss function (e.g., mean squared error or log loss).
- Regularization is applied to control complexity and reduce overfitting.

The process continues until a stopping criterion is met (e.g., a maximum number of iterations or minimum reduction in loss).

12.1.2 Mathematical Expressions

The objective function of XGBoost combines a loss function and a regularization term to penalize model complexity. For a model with K trees, the objective function can be expressed as:

$$\text{Objective:} \quad \mathcal{L} = \sum_{i=1}^{N} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

$$\text{Regularization term:} \quad \Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$$

Where:

- $l(y_i, \hat{y}_i)$ is the loss function (e.g., squared error for regression or log loss for classification).
- $\Omega(f_k)$ is the regularization term for tree k to penalize complex trees.
- γ controls the penalty for the number of leaves T in the tree.
- λ controls the penalty for the $L2$ norm of leaf weights w_j.

12.2 Boosting Procedure

XGBoost employs a boosting procedure in which each new tree corrects the residual errors of the previous trees. Given the residuals r_i for each sample i, XGBoost optimizes the objective function by using a second-order Taylor expansion for faster and more accurate updates:

$$\text{Approximate objective:} \quad \mathcal{L}^{(t)} = \sum_{i=1}^{N} \left(g_i f_t(x_i) + \frac{1}{2} h_i f_t(x_i)^2 \right) + \Omega(f_t)$$

Where:

- g_i and h_i are the first and second-order gradients of the loss function with respect to predictions.
- $f_t(x_i)$ is the prediction of tree t for sample i.

12.3 Regularization and Pruning

XGBoost includes regularization parameters to control overfitting:

- γ: Minimum loss reduction required to make a further partition on a leaf node.
- λ and α: $L2$ and $L1$ regularization terms on leaf weights to penalize overly complex trees.

XGBoost also includes early stopping to halt training if the validation score does not improve after a set number of rounds.

12.4 Voting/Averaging

In XGBoost, the final prediction is obtained by summing the weighted predictions of all trees. For classification tasks, a softmax function is applied if probabilities are needed. For regression tasks, the sum of all tree predictions is taken as the final output.

12.5 Exploration of XGBoost Classifier

To further leverage ensemble techniques, we applied the XGBoost Classifier. However, we found that the accuracy for drugs other than Heroin and LSD decreased,

and the AUC-ROC values similarly reflected this trend. This outcome confirmed our hypothesis regarding the imbalanced dataset and insufficient support for certain classes.

```
Evaluating XGBoost for LSD...
Performance for LSD:
Accuracy: 0.5777
Confusion Matrix:
[[266  16   6  22   5   5   1]
 [ 41  24   9   3   0   0   0]
 [ 29   8   7   9   1   1   0]
 [ 26   2   6  24   5   3   2]
 [ 17   0   2   5   5   0   0]
 [  8   1   0   3   0   1   0]
 [  1   0   0   2   0   0   0]]
Classification Report:
              precision    recall  f1-score   support

           0       0.69      0.83      0.75       321
           1       0.47      0.31      0.38        77
           2       0.23      0.13      0.16        55
           3       0.35      0.35      0.35        68
           4       0.31      0.17      0.22        29
           5       0.10      0.08      0.09        13
           6       0.00      0.00      0.00         3

    accuracy                           0.58       566
   macro avg       0.31      0.27      0.28       566
weighted avg       0.54      0.58      0.55       566

ROC AUC: 0.7718
------------------------------------------------
```

Figure 3: Classification Report

13 Artificial Neural Networks (ANN)

Artificial Neural Networks (ANNs) are a subset of deep learning models inspired by the structure and function of the human brain. They consist of interconnected nodes (neurons) organized in layers that enable complex pattern recognition and feature extraction. ANNs are commonly used in both classification and regression

tasks due to their ability to capture non-linear relationships in data.

13.1 Network Structure

An ANN is composed of an input layer, one or more hidden layers, and an output layer:

- The **input layer** receives the input data, with each neuron representing a feature.
- The **hidden layers** consist of neurons connected with weights and biases that process input data through nonlinear transformations.
- The **output layer** provides the final prediction, with the number of neurons corresponding to the target output dimensions.

Each neuron in a hidden layer applies a transformation defined by an activation function to introduce non-linearity and enable learning complex patterns.

13.1.1 Mathematical Expressions

The output of a neuron j in layer l is calculated as:

$$\text{Neuron output:} \quad a_j^{(l)} = \sigma\left(\sum_{i=1}^{n} w_{ij}^{(l)} a_i^{(l-1)} + b_j^{(l)}\right)$$

Where:

- $w_{ij}^{(l)}$ is the weight connecting neuron i in layer $l-1$ to neuron j in layer l.
- $a_i^{(l-1)}$ is the activation of neuron i in the previous layer.

- $b_j^{(l)}$ is the bias term for neuron j in layer l.
- σ is the activation function, such as ReLU, sigmoid, or tanh.

13.2 Activation Functions

Activation functions introduce non-linearity to the model, enabling it to learn complex patterns. Common activation functions include:

$$\text{Sigmoid:} \quad \sigma(x) = \frac{1}{1 + e^{-x}}$$

$$\text{ReLU:} \quad f(x) = \max(0, x)$$

$$\text{Tanh:} \quad f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

13.3 Forward and Backward Propagation

During training, an Artificial Neural Network (ANN) adjusts its weights and biases through the processes of forward and backward propagation:

- **Forward Propagation**: The input data flows through each layer, with each neuron computing its output until the network reaches the final prediction at the output layer.
- **Loss Function**: The difference between the predicted and actual values is calculated using a loss function, such as Mean Squared Error (MSE) for regression or Cross-Entropy Loss for classification tasks.
- **Backward Propagation**: Using the chain rule, the gradients of the loss with respect to the weights and biases are computed. These gradients are then used to adjust the weights to minimize the loss.

13.3.1 Mathematical Expressions for Backpropagation

Let $\mathcal{L}$ represent the loss function. The gradients for updating weights and biases are determined as follows:

$$\text{Weight Gradient:} \quad \frac{\partial \mathcal{L}}{\partial w_{ij}} = \frac{\partial \mathcal{L}}{\partial a_j} \cdot \frac{\partial a_j}{\partial w_{ij}}$$

$$\text{Bias Gradient:} \quad \frac{\partial \mathcal{L}}{\partial b_j} = \frac{\partial \mathcal{L}}{\partial a_j} \cdot \frac{\partial a_j}{\partial b_j}$$

13.4 Optimization

To optimize the network and minimize loss, ANNs use algorithms such as Stochastic Gradient Descent (SGD) and its variants (e.g., Adam, RMSprop), which iteratively adjust weights and biases until convergence.

13.4.1 Gradient Descent Update Rule

The gradient descent rule is applied to update the weights:

$$w_{ij} \leftarrow w_{ij} - \eta \cdot \frac{\partial \mathcal{L}}{\partial w_{ij}}$$

Where:

- η is the learning rate, controlling the step size for each weight update.

13.5 Evaluation of the outcomes

Once training is complete, new input data can be passed through the network to obtain predictions. For classification tasks, a softmax activation function may

be used in the output layer to produce class probabilities, while regression tasks output raw predictions.

Classification Report:

	precision	recall	f1-score	support
CL0	0.66	0.55	0.60	222
CL1	0.72	0.88	0.79	189
CL2	0.69	0.58	0.63	200
CL3	0.71	0.65	0.68	201
CL4	0.72	0.80	0.76	193
CL5	0.79	0.82	0.81	206
CL6	0.84	0.92	0.88	189
accuracy			0.74	1400
macro avg	0.73	0.74	0.73	1400
weighted avg	0.73	0.74	0.73	1400

Figure 4: Classification Report

The classification report offers an in-depth evaluation of the model's performance in this 7-class classification task, with each class representing a distinct category of drug consumption. This analysis has been conducted for each drug, with a detailed breakdown here provided for the drug "Benzos." The report examines essential metrics such as precision, recall, F1-score, and support for each class, alongside aggregate scores for accuracy, macro average, and weighted average.

Precision

Class CL6 shows the highest precision (0.84), while CL0 has the lowest (0.66), suggesting that the model is more accurate in identifying CL6 than CL0.

Recall

The model has the best recall for CL6 (0.92), indicating it correctly captures most instances of this class. In contrast, recall for CL0 is lower (0.55), with other classes ranging from 0.58 to 0.88.

F1-Score

CL6 has the highest F1-score (0.88), reflecting a balance between precision and recall, whereas CL0 has the lowest F1-score (0.60), indicating less reliable predictions for this class.

Overall Accuracy

Overall, the model displays moderate accuracy in predicting each class. The macro and weighted averages for precision, recall, and F1 scores are approximately 0.73 to 0.74, indicating balanced performance across classes without significant bias, and showing that the model can generalize its performance effectively across different classes.

14 Final Conclusion of our Project

Summary of the Study

Our study looked at how different personality traits influence drug use by analyzing data from 1,885 people. We wanted to create a model that could predict how individuals use drugs based on traits like moodiness, sociability, openness to new experiences, friendliness, reliability, impulsiveness, and seeking thrills. We

focused on 18 drugs, from those never tried to those used recently.

We found that logistic regression was good at predicting use for certain substances like Semeron and cocaine, with high accuracy. However, it struggled with popular drugs like alcohol and cannabis, where accuracy was lower. This might be because these commonly used drugs are enjoyed in various ways by different people.

When we switched to an artificial neural network (ANN), we saw better results in terms of precision and recall for specific drug categories, reaching an impressive F1 score of 0.88. This score tells us the model was well-rounded in predicting usage levels for certain substances. We also tested decision trees, random forests, and XGBoost, which showed moderate accuracy, with overall scores around 0.73-0.74. Our models generally performed well, especially for drugs with clear usage patterns. However, improving accuracy for widely used drugs may require tweaking our methods or gathering more detailed data to capture the subtleties of how people consume them.