Introduction to Big Data and Data Analytics
'20B12CS333'

Substance Abuse Analysis & Predictions

Members:

Shreya Grover - B3 - 18103113

Sarthak Sharma - B3 - 18103112

Akshay Mantri - B9 - 18103270

Shubh Anand - B12 - 18104069

Abstract:

It is highly important to evaluate the problem of a person's risk with substance abuse. Various studies have been done to evaluate the correlation between one's personality with their use of substances which may be legal or illegal. One of the studies, which is used by us, collected data including Big Five personality traits (NEO-FFI-R), impulsivity (BIS-11), sensation seeking (ImpSS), and demographic information. This data set contains 18 psychoactive substances. A correlation analysis has been done to find out a pattern between these drugs with various personality traits.An exhaustive search was performed to pick the foremost effective subset of input features and data processing methods to classify users and non-users for every substance. Different classifiers (Decision tree, Support Vector Machine, k- nearest neighbour, Adaboost) have been used to find out the most suitable classifier for each substance.

Introduction:

Normally, when the name "Drug/Substances" is heard, people generally relate it to narcotics. But, from the aspects of Pharmaceutical Sciences1, "Drugs" are a wide variety of chemical substances which when consumed in any manner – inhaled, injected, smoked, orally-consumed, absorbed (via skin) etc. affects the living body and brings chemical consumptions inside the body. Most of the substances are useful for the treatment of diseases found in Human body, and they are better known by the name "Medicine". Meanwhile, the other variant of Drugs are Recreational drugs. They also cause chemical changes in the state of the body, but rather their usage for treatment, these are used for recreation purposes by different age categories of humans. All of these drugs, more appropriately, the Narcotic drugs or recreational drugs leaves a devastating impact on the person who consumes it. These impacts remain in the body even after a few days of consumption in the human body. Since, they result in chemical changes, parameters and properties of blood, can be used to predict and determine when and what kind of narcotic one has consumed.

Studies have shown that the personality traits of the Five Factor Model (FFM) are the foremost comprehensive and adaptable system for understanding human individual differences. The study consists of various components which are Neuroticism(N), Extraversion (E), Openness to Experience(O), Agreeableness(A), and Conscientiousness(C). A variety of studies have illustrated that personality traits are associated with substance consumption. The importance of the connection between high N and therefore the presence of psychotic symptoms following cocaine-induced drug consumption. The personality traits of N,E, and C are highly correlated with hazardous health behaviors. a coffee score of C, and high score of E or high score of N

correlate strongly with multiple risky health behaviors. Alcohol used to be related to lower A and C, and better E. They found also that lower A and C, and better O are related to marijuana. the connection between low C and drug consumption is moderated by poverty; low C may be a stronger risk factor for illicit drug usage among those with relatively higher socioeconomic status. They found that prime N, and low A and C are related to higher risk of drug use (including cocaine, crack, morphine, codeine, and heroin). It should be mentioned that prime N is positively related to many other addictions like Internet addiction, exercise addiction, compulsive buying, and study addiction. An individual's personality profile plays a task in becoming a substance abuser . The personality profiles for the users and non-users of nicotine, cannabis, cocaine, and heroin are related to a FFM of personality samples from different communities. They also highlight the links between the consumption of those drugs and low C. A direct correlation between N and O, and drug use, while, increasing scores for C and A decreases risk of drug use. Previous studies demonstrated that participants who use drugs including alcohol and nicotine have a robust direct correlation between A and C and a robust indirect correlation for every of those factors with N. Three high-order personality traits are proposed as endophenotypes for substance use disorders: Positive Emotionality, Negative Emotionality, and Constraint. The statistical characteristics of groups of drug users and non-users are studied by many authors. They found that the personality profile for the users and non users of tobacco, marijuana, cocaine, and heroin are related to a better score on N and a really low score for C. Sensation seeking is additionally higher for users of recreational drugs.

Background Study:

- **Personality Measurements-**
  In order to assess personality traits of the sample, the Revised NEO Five-Factor Inventory (NEO-FFI-R) questionnaire was employed. The NEO-FFI-R may be a highly reliable measure of basic personality domains; internal consistencies are 0.84 (N); 0.78 (E); 0.78 (O); 0.77 (A), and 0.75 (C) Egan. the size may be a 60-item inventory comprising 5 personality domains or factors. The NEO-FFI-R may be a shortened version of the Revised NEO-Personality Inventory (NEO-PI-R). These factors are: N, E, O, A, and C with 12 items per subset. The five traits are often summarized as: 1. Neuroticism (N) may be a long-term tendency to experience negative emotions like nervousness, tension, anxiety and depression;
  2. Extraversion (E) is manifested in outgoing, warm, active, assertive, talkative, cheerful, and in search of stimulation characteristics;

3. Openness to experience (O) may be a general appreciation for art, unusual ideas, and imaginative, creative, unconventional, and wide interests,
4. Agreeableness (A) may be a dimension of interpersonal relations, characterized by altruism, trust, modesty, kindness, compassion and cooperativeness;
5. Conscientiousness (C) may be a tendency to be organized and dependable, strong-willed, persistent, reliable, and efficient.

- **Decision Tree-**

  The decision tree approach is a classifier that constructs a tree like structure, which can be used to choose between several courses of action. Binary decision trees are used in this study. A decision tree is comprising of nodes and leaves. Each node can have a child node. If a node has no child node, it's called a leaf or a terminal node. Any decision tree contains one root node, which has no parent node. Each non terminal node calculates its own Boolean with the value ‘true’ or ‘false’). According to the result of this calculation, the decision for a given sample would be delegated to the left child node (’true’) or to the right child node (’false’). Each leaf (terminal node) features a label which shows what percentage samples of the training set belong to every class. The probability of every class is estimated as a ratio of the amount of samples during this class to the entire number of samples within the leaf. There are many methods for developing a choice tree. We use the methods supported by Gini gain.

- **Support Vector Machine(SVM)-**

  “Support Vector Machine” (SVM) may be a supervised machine learning algorithm which may be used for both classification or regression challenges. However, it is mostly used in classification problems. The objective of the support vector machine algorithm is to seek out a hyperplane in an N-dimensional space (N — the amount of features) that distinctly classifies the info points. To separate the 2 classes of knowledge points, there are many possible hyperplanes that would be chosen. Our objective is to seek out a plane that has the utmost margin, i.e. the maximum distance between data points of both classes. Maximizing the margin distance provides some reinforcement in order that future data points are often classified with more confidence.

- **K- Nearest Neighbours (kNN) -**

  The basic concept of kNN is the class of an object is the class of the majority of its k nearest neighbours . This algorithm is very sensitive to distance definition.

There are several commonly used variants of distance for kNN: Euclidean distance; Minkovsky distance; and distances calculated after some transformation of the input space.In this study, we used three distances: the Euclidean distance, the Fisher's transformed distance and the adaptive distance. Moreover, we used a weighted voting procedure with weighting of neighbours by one of the standard kernel functions. The kNN algorithm is well known. The adaptive distance transformation algorithm is described in. kNN with Fisher's transformed distance is less known.

- **Adaboost Classifier-**

  Adaboost is used in conjunction with other learning algorithms in an effort to improve accuracy. The results of other algorithms are combined to create a weighted sum which hopefully improves performance.

System Architecture:

- **Dataset Description**

  Database contains records for 1885 respondents.Our dataset had 12 different features, 5 of them were general information on a person and the other 7 were personality scores that described that person. The general information features were age, gender, level of education, ethnicity, and country of residence; while the 7 personality scores were, neuroticism, extraversion, openness to experience, agreeableness, conscientiousness, impulsivity, and sensation seeking.

Neuroticism	Extraversion	Openness to experience	Agreeableness	Conscientiousness	Impulsivity	Sensation seeking

Figure 1. Seven personality scores included in the dataset

Age	Gender	Level of Education	Ethnicity	Country of Residence

Figure 2. Five general information features

In addition, participants were questioned concerning their use of 18 legal and illegal drugs (alcohol, amphetamines, amyl nitrite, benzodiazepine, cannabis, chocolate, cocaine, caffeine, crack, ecstasy, heroin, ketamine, legal highs, LSD, methadone, mushrooms, nicotine and volatile substance abuse and one fictitious drug (Semeron)

which was introduced to identify over-claimers. For each drug they have to select one of the answers: never used the drug, used it over a decade ago, or in the last decade, year, month, week, or day. Database contains 18 classification problems. Each of independent label variables contains seven classes: "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".

- **Methodology-**

  Data Preprocessing - The data was first cleaned to find any missing values which we got none.

  Transformation of Class Labels- We transformed the seven class labels into binary classification by union of a part of classes into one new class, i.e, “Never Used” and “Used over a Decade ago” into “Non- User” and the rest were being transformed into "User".

  Algorithms Supported- The classifiers used in our project are Decision trees, k-Nearest Neighbours, Support Vector machines. We also used AdaBoost to maximize our classification accuracy.

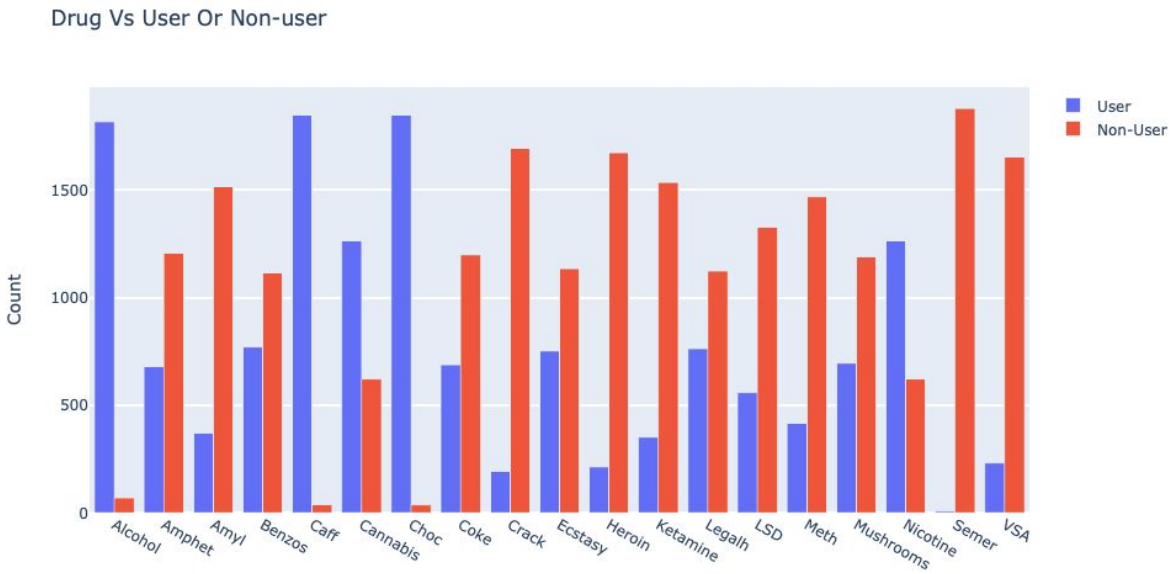
Novelty:

The paper references for our project did not classify the class labels as we did.They predicted per drug typically using a decision tree. Their averages came in and around 70-85 percent, however they were classifying much easier and much harder datasets than us. It is hard to say how we compare.

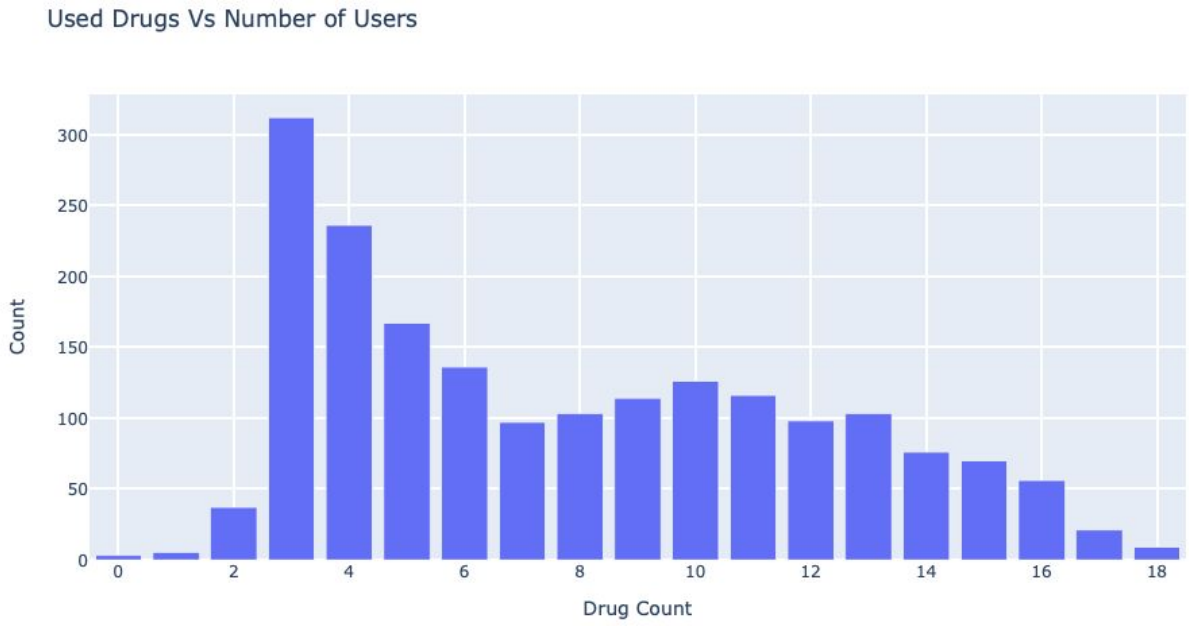
Results:

The data set contains seven categories of drug uses: ‘Never used’, ‘Used over a decade ago’, ‘Used in last decade’, ‘Used in last year’, ‘Used in last month’, and ‘Used in last week’. We form four problems based on four dichotomies of these classes (see ‘Drug use’ Section): the decade-, year-, month-, and week-based user/non-user separations. We identified the relationship between personality profiles (NEO-FFI-R) and drug consumption for the decade-, year-, month-, and week-based classification problems. We evaluated the risk of drug consumption for each individual according to their personality profiles. This evaluation was performed separately for each drug for the

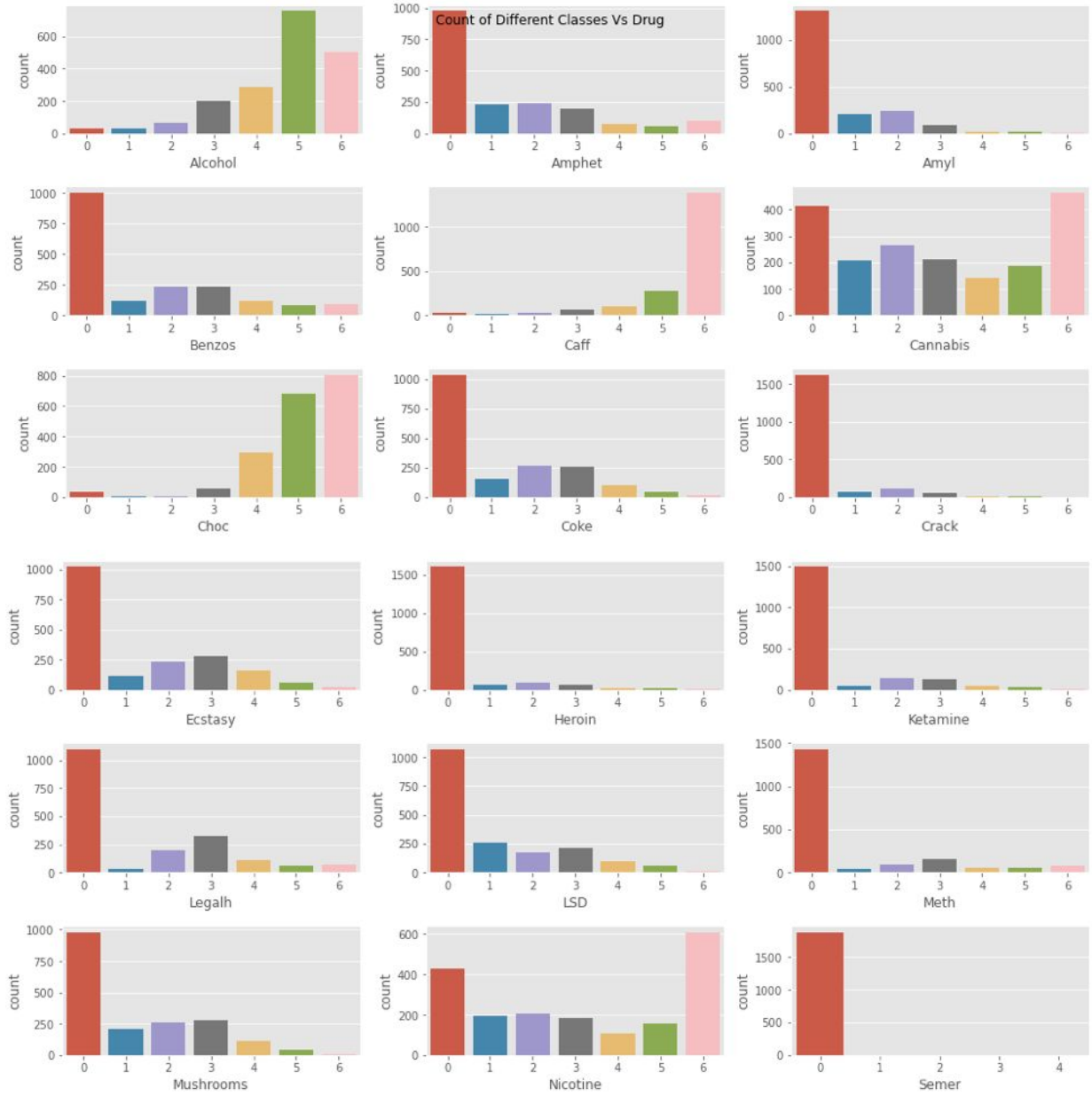
decade-based user/non-user separation. We also analyzed interrelations between the individual drug consumption risks for different drugs.



Graph 1:- Visualisation of the count of 'User' vs 'Non User'



Graph 2:- Number of user for each drug



Graph 3:- Count of different classes(7 classes in the original dataset) vs each Substance

Support Vector Machines(SVM)

SVMs can perform both unsupervised and supervised learning. It is a binary linear classifier method. It works by being given a training set, with the data being marked as belonging to one of two categories. The SVM then creates a model which assigns new samples to one category or the other.

Classification Report:

	precision	recall	f1-score	support
0	0.71	0.74	0.73	173
1	0.88	0.87	0.88	393
accuracy			0.83	566
macro avg	0.80	0.80	0.80	566
weighted avg	0.83	0.83	0.83	566

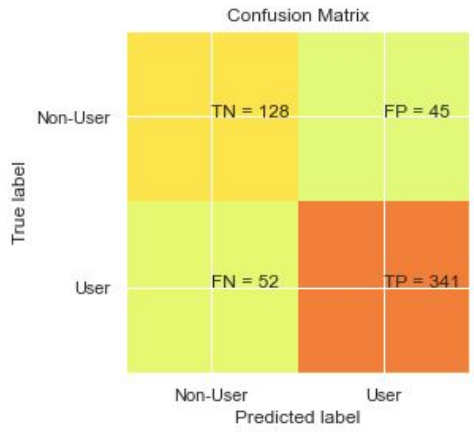


Figure1- Classification Matrix for SVM

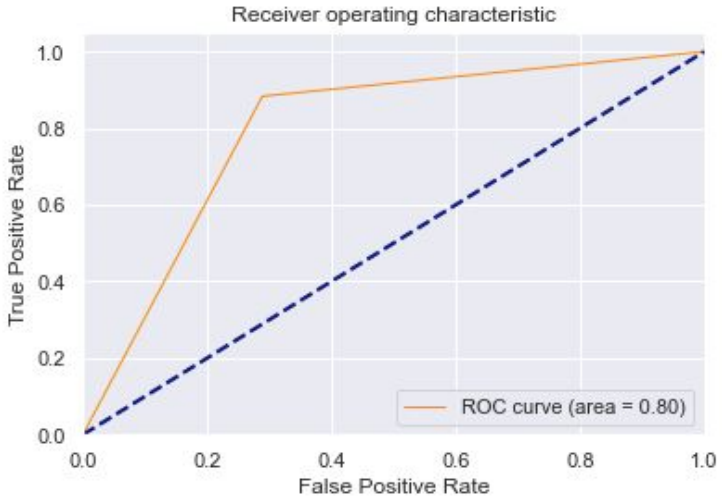


Figure2- SVM Classifier ROC Curve

Decision Tree-

The decision tree is a supervised algorithm that uses observations about an item to reach conclusions about the item's value. The interior nodes of the tree corresponds to the input variable, where the edges to the children nodes represent the possible values of the input.

Classification Report:

	precision	recall	f1-score	support
0	0.63	0.76	0.69	173
1	0.88	0.80	0.84	393
accuracy			0.79	566
macro avg	0.76	0.78	0.76	566
weighted avg	0.81	0.79	0.79	566

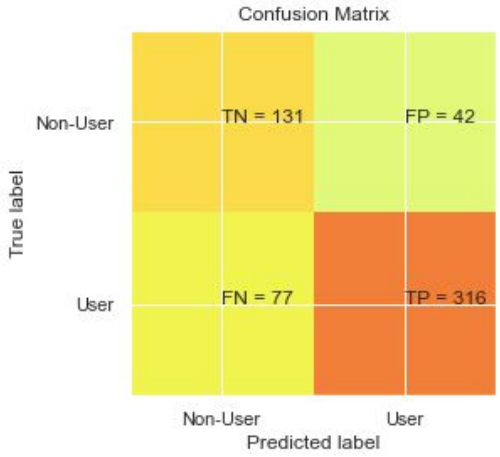


Figure3- Classification Matrix for Decision Tree

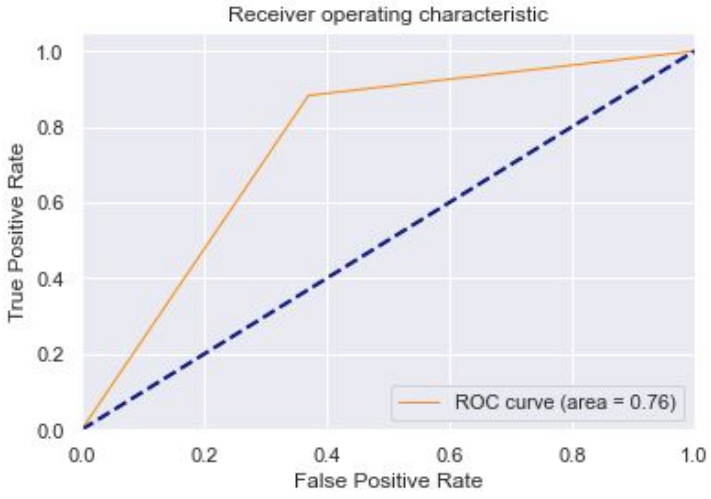
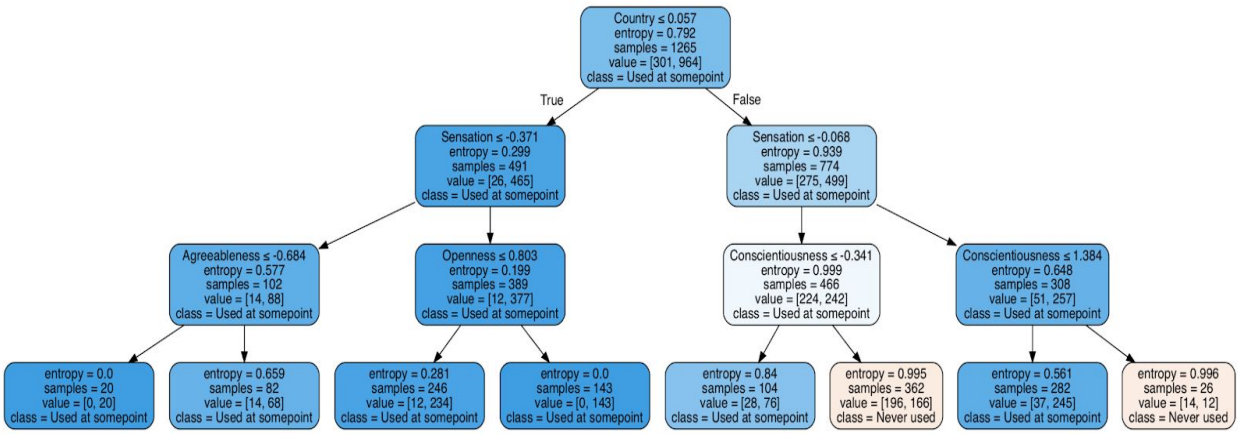


Figure4- Decision Tree ROC Curve



Example of Decision Tree for the cannabis dataset.

K- Nearest Neighbour -

KNN is a non-parametric supervised classification method. It works by selecting a data point to be the center cell, then growing the cluster until it contains a determined amount of data points. Once the cluster, typically referred to as a neighborhood, is fully created, we then search the neighborhood and count the number of data points that belong to each class. The class with the highest amount of data point in the neighborhood is determined to be the class of the original data point.

Classification Report:

	precision	recall	f1-score	support
0	0.60	0.65	0.63	173
1	0.84	0.81	0.82	393
accuracy			0.76	566
macro avg	0.72	0.73	0.73	566
weighted avg	0.77	0.76	0.76	566

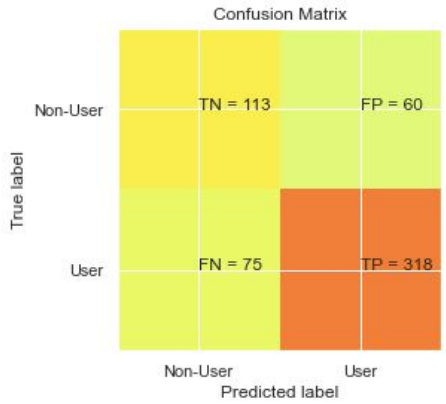


Figure5- Classification Matrix for kNN Classifier

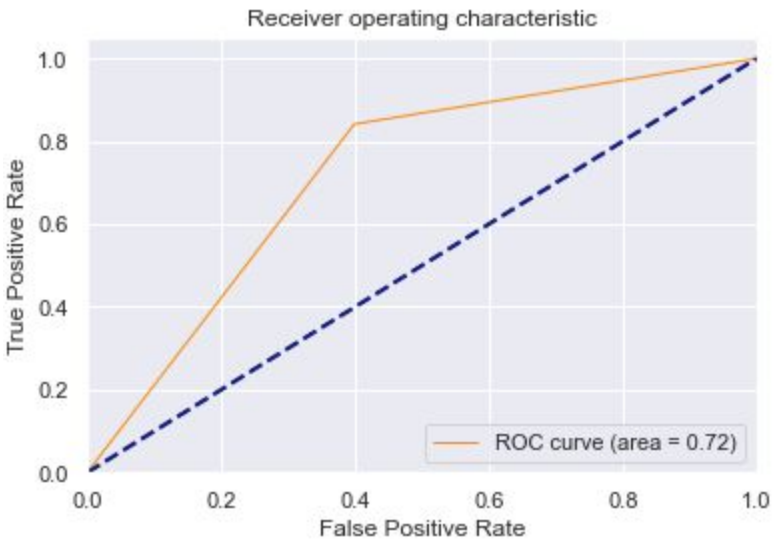


Figure6- kNN ROC Curve

Adaboost Classifier-

Adaboost is used in conjunction with other learning algorithms in an effort to improve accuracy. The results of other algorithms are combined to create a weighted sum which hopefully improves performance.

Classification Report:

	precision	recall	f1-score	support
0	0.69	0.72	0.70	173
1	0.87	0.86	0.87	393
accuracy			0.82	566
macro avg	0.78	0.79	0.79	566
weighted avg	0.82	0.82	0.82	566

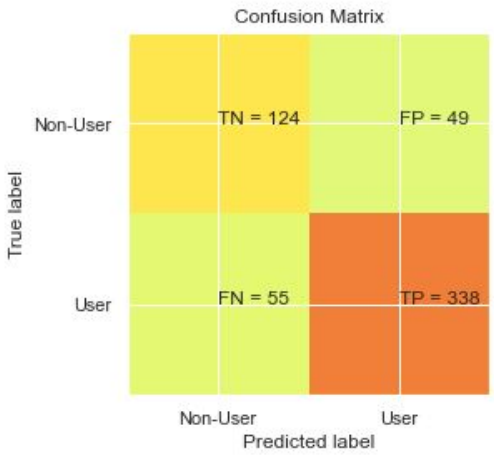


Figure7- Classification Matrix for Adaboost

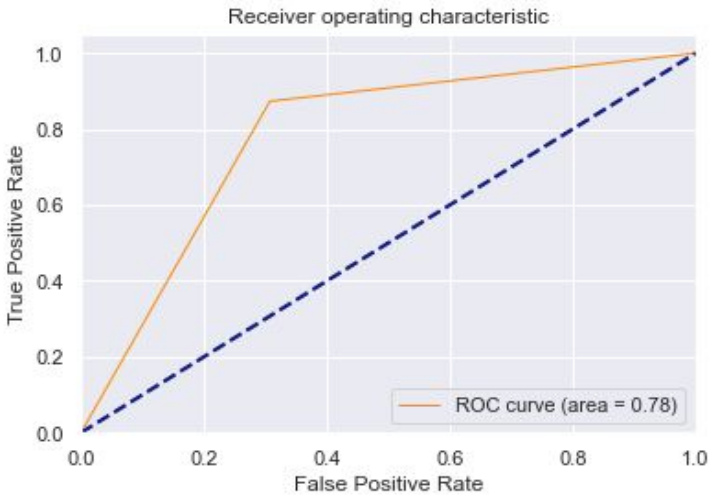
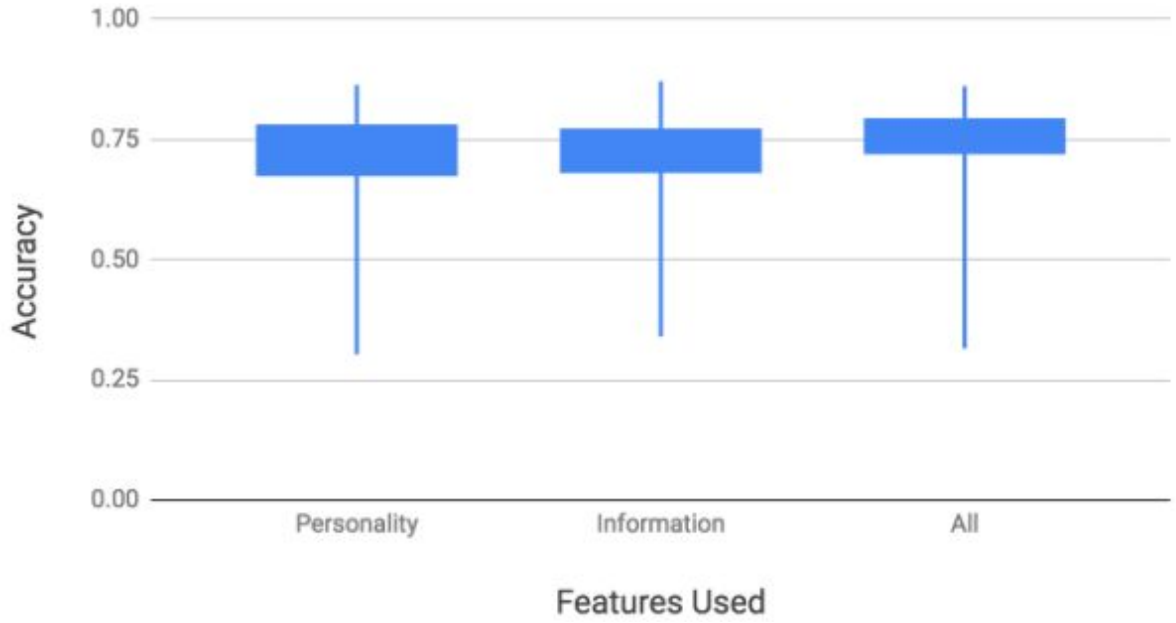


Figure8- Adaboost ROC Curve

Accuracy Statistics For the Substances				
Classifiers	**Min**	**Mean**	**Median**	**Max**
Decision Tree	36.53%	66.69%	72.85%	86.11%
KNN	39.90%	66.71%	71.71%	84.94%
Support Vector	38.45%	67.57%	72.98%	86.34%
AdaBoost	39.40%	67.10%	73.08%	85.21%

Table 1:- Accuracy Statistics for the Substances

Overall, most of the algorithms reached a respectable level of accuracy. SVM gave us the highest level of accuracy, though the other algorithms were not far behind it.Since Adaboost was added on as a bonus, we were pleased that it netted such accurate results.



Graph 4:- Accuracy vs Features used

This graph shows our prediction accuracies only using certain features. Surprisingly, using personality scores or general information about someone makes almost no difference. This tells us that income, country, ethnicity, gender, education, and age are equally likely to determine if someone is likely to do drugs as personality scores.

The best classifier we found came from a decision tree. Although SVM had the maximum accuracy, we want a classifier that also has a high f1 score as well as a high accuracy.

Drug	Best Accuracy
Alcohol	99.22%
Amphetamines	70.65%
Amyl Nitrite	74.04%
Benzodiazepine	71.34%
Cannabis	83.82%
Chocolate	99.20%
Cocaine	70.47%
Caffeine	99.39%
Crack	89.67%
Ecstasy	75.94%
Heroin	88.19%
Ketamine	84.84%
Legal Highs	80.75%
LSD	79.18%
Methadone	81.99%
Mushrooms	75.65%
Nicotine	80.79%
Volatile Substance Abuse	80.78%

Table2 :- Best accuracy for each Substance

Conclusions:

In conclusion, the best classification methods we found when running against the dataset were Decision Trees, kNN, Support vector machine, and AdaBoost. Our very

best classifier was Decision trees, even though it did not have the highest accuracy, its f1 score was high. When running against individual drugs we saw much higher accuracies for a handful of drugs, regardless of the classification method used. This is because the drugs were either very commonly used or not very common at all. So if the usage for a particular drug was on either extreme we found that the algorithms usually had a much higher accuracy. This helped for more accurate results.
These results are important as they examine the question of the relationship between drug use and personality comprehensively and engage the challenge of untangling 44 correlated personality traits (the FFM, impulsivity, and sensation seeking), and clusters of substance misuse (the correlation pleiades). The work acknowledged the breadth of a common behaviour which may be transient and leave no impact, or may significantly harm an individual. We examined drug use behaviour comprehensively in terms of the many kinds of substances that may be used (from the legal and anodyne, to the deeply harmful), as well as the possibility of behavioural over-claiming.

References:

- E. Fehrman, A. K. Muhammad, E. M. Mirkes, V. Egan, A. N. Gorban (2017). The Five Factor Model of personality and evaluation of drug consumption risk.
- Elaine Fehrman. Vincent Egan. Evgeny M. Mirkes. (2016). Drug consumption Classification.
- Kshitiz Sirohi (2019). Support Vector Machines
- Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Published by Addison Wesley(??).Introduction to Data Mining
- Tavish Srivastava (2014). Introduction to k-Nearest Neighbors: A powerful Machine Learning Algorithm