Drug Consumption Analysis

Group 4: Darien Young, Meishan Fan, Nghi Nguyen, Quang Tran, Yingying Liu
Department of Applied Data Science, San Jose State University
DATA 240 Data Mining/Analytic
December 12th, 2022

1.Introduction

Drug consumption is a common habit in the United States, with some being illegal and others being legal, such as nicotine and alcohol. With both types of drug consumption significantly affecting local communities, much research has been done on the topic. One popular branch of research is to determine if there are pre-determined characteristics that make an individual likely to consume a drug or characteristics that make them more susceptible to prolonged drug addiction. We chose this topic because of the plentiful amount of raw data available to us and the ability for us to personally see if our project results match with what is reported or observed in our local community.

Our data is Drug consumption (quantified) obtained from UC Irvine Machine Learning Repository. What interests us about this dataset is that the data uses personality measurements of the drug user. The personality measurements are EO-FFI-R (neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness), BIS-11 (impulsivity), and ImpSS (sensation seeking). Outside of the personality measurements, we have the drug users' age, ethnicity, gender, education, and country of residence; six features.

The dataset covers 18 drugs in total. This is a mix of legal and illegal addictive substances, but a majority of them are unlawful. Some of the drugs include cannabis, nicotine, and alcohol. Each individual is asked about their usage of each 18 drugs on a scale of 0-6 with a CL prefix. Each metric represents when they last took the drug: "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". The dataset can be switched to a binomial function by creating a threshold on the temporal scale with it becoming a simple yes or no on usage of the drug. In total, we have 32 features which is above the requirement of 20 features, and the dataset contains 1885 rows, aka respondents. The ratio between respondents and features falls within the general rule for machine learning of having the data be ten times the amount of features.

1.1 Data Source and Summary

Drug consumption (quantified) obtained from UC Irvine Machine Learning Repository contains a total 1885 records with 32 features on each. The 32 features can be broken down into three main categories: demographic characteristics, personality characteristics, and drugs.

Our demographic characteristics are made up of six features. These features are Age, Gender, Education, Country, and Ethnicity. Each feature on the dataset has its category bin based on the data owner’s own classification method. The values they used to represent each classification are not intuitive and require reviewing their documentation. To make it easier for our project, we also mapped out their values to values that are more intuitive.

Table 1
Feature Meanings

Age		Gender		Education		Country		Ethnicity	
Value	Meaning	Value	Meaning	Value	Meaning	Value	Meaning	Value	Meaning
-0.952	18-24	0.48246	Female	-2.4359	Left school before 16 years	-0.0977	Australia	-0.5021	Asian
-0.0785	25-34	-0.4825	Male	-1.7379	Left school at 16 years	0.24923	Canada	-1.107	Black
0.49788	35-44			-1.4372	Left school at 17 years	-0.4684	New Zealand	1.90725	Mixed-Black/Asian
1.09449	45-54			-1.2275	Left school at 18 years	-0.2852	Other	0.126	Mixed-White/Asian
1.82213	55-64			-0.6111	college or university, no certificate or degree	0.21128	Republic of Ireland	-0.2217	Mixed-White/Black
2.59171	65+			-0.0592	Professional certificate/diploma	0.96082	UK	0.1144	Other
				0.45468	University degree	-0.5701	USA	-0.3169	White
				1.16365	Masters degree				
				1.98437	Doctorate degree				

The personality characteristic is made up of 7 parameters: Nscore, Escore, Oscore, Ascore, Cscore, Impsulive, and SS. Table 2 explains what each of the definitions of these abbreviation and which personality measurement standard they are based off of:

Table 2
Dataset Label Definitions

Dataset Label	Definition	Personality measurement standard
Nscore	Neuroticism	NEO-FFI-R
Escore	Extraversion	NEO-FFI-R
Oscore	Openness	NEO-FFI-R
Ascore	Agreeableness	NEO-FFI-R
Cscore	Conscientiousness	NEO-FFI-R
Impulsiveness	Impulsiveness	BIS-11
SS	Sensation	ImpSS

The data set also contains 18 legal and illegal drug features: Alcohol, Amphetamine, Amyl_nitrite, Benzodiazepine, Caffeine, Cannabis, Chocolate, Cocaine, Crack, Ecstasy, Heroin, Ketamine, Legal_highs, LSD, Methadone, Mushrooms, Nicotine, Semeron, and VSA. Each of the 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".

Table 3

Drug Usage classification

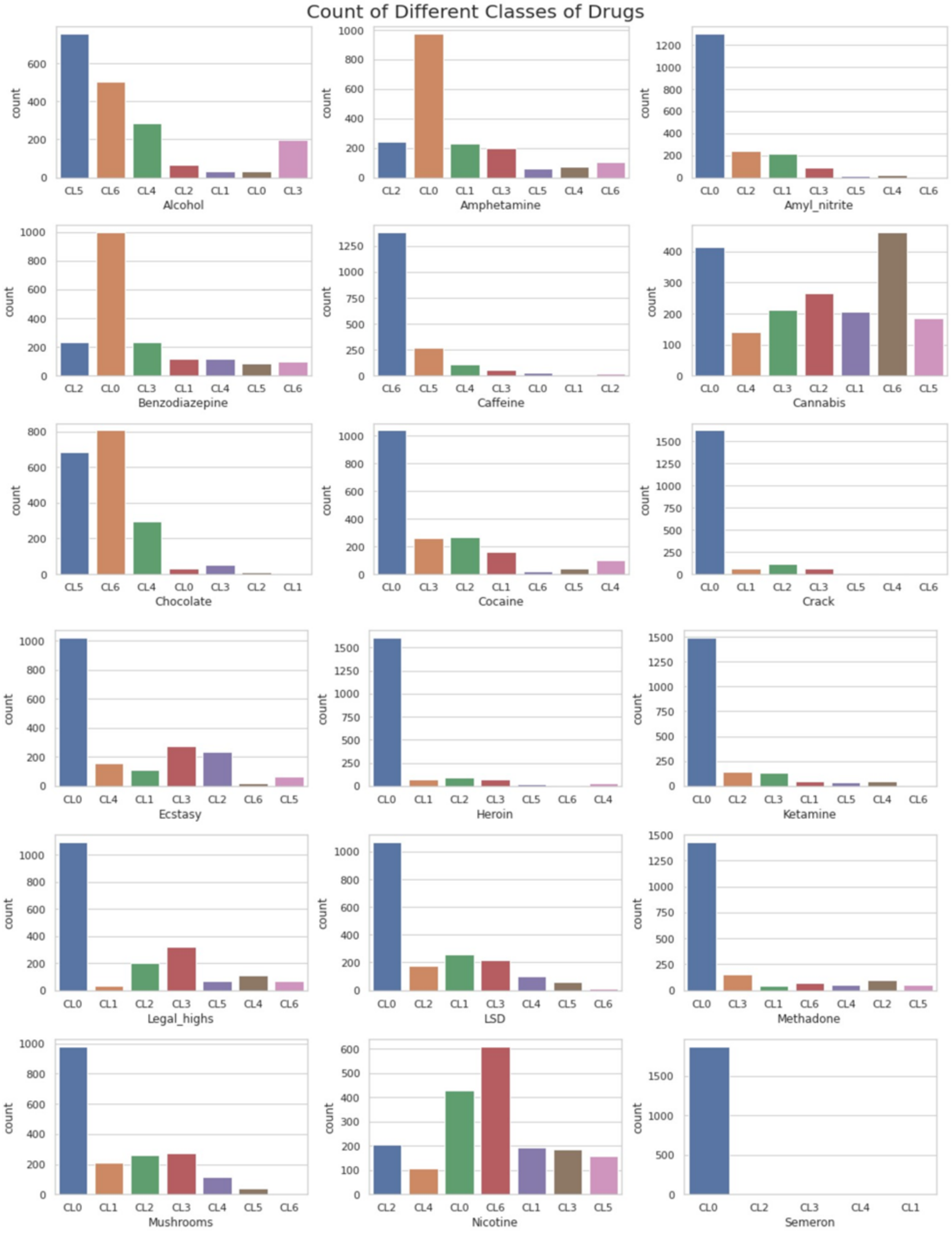
Data Classification	Definition
CL0	Never Used
CL1	Used over a Decade Ago
CL2	Used in Last Decade
CL3	Used in Last Year
CL4	Used in Last Month
CL5	Used in Last Week
CL6	Used in Last Day

The data set contains information on the consumption of 18 drugs, but the classes for most of the individual drugs are imbalanced (as shown in Figure 1). As we know, imbalanced datasets create challenges for machine learning. Since the machine learning models trained on imbalanced data usually fall victim to the accuracy paradox. To avoid this issue, we finally selected the feature of Cannabis, the most balanced feature among 18 drug consumptions, as the target attribute. Cannabis is a popular drug used by numerous demographics, and there is a growing decrease in stigma regarding cannabis consumption, especially with State's increasingly legalizing its usage.

The feature of cannabis usage contains categorical and discrete data. The classes for the cannabis column in the original dataset include 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. In this project, we will change the classes from a multivariate of six into a binary output of "0" and "1".

Figure 1

The Numbers of Different Classes for Each Drug Consumption



After determining the target feature, we discriminated participants into groups of users and non-users for binary classification. In the raw data set, the people are grouped into six 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". Based on these categories, we want to discriminate the participant into two groups: 'Never used' and 'Used'. The research from Fehrman et al. (2015) suggests four ways to classify the participants into 'Never used' and 'Used' groups: decade-, year-, month-, and week-based user/non-user separation (Shown in Table 4). The decade-based separation placed 'Never used' and 'Used over a decade ago' into the class of 'non-users' and all other categories were grouped into the class 'users' of drugs; the year-based classification clustered the categories 'Used in last decade', 'Used over a decade ago', and 'Never used' into the group of non-users and all other categories are placed into the group of users; The month-base method combined the categories 'Never used', 'Used over a decade ago', 'Used in last decade', and 'Used in last year' to the class of non-users and all three other categories are clustered into the group of users; The week-based classification grouped the categories 'Used in last week' and 'Used in last month' into the category of users, and all others are placed into the category of non-users.

Table 4

Different Categories of Drug Users

	Week-based	**Month-based**	**Year-based**	**Decade-based**
User	Used in last day Used in last week	Used in last day Used in last week Used in last month	Used in last day Used in last week Used in last month Used in last year	Used in last day Used in last week Used in last month Used in last year Used in last decade
Nonuser	Used in last month Used in last year Used in last decade Never used over a decade ago	Used in last year Used in last decade Never used over a decade ago	Used in last decade Never used over a decade ago	Never used over a decade ago

The paper from Fehrman et al. (2015) also claimed that only a participant in the 'Never used' class can be formally called a non-user. Still, it is not a seminal definition, and in most applications, people who used a drug more than a decade ago cannot be considered drug users. Hence, in this project, we selected the year-based classification. After choosing the category method, we deployed data preparation to prepare the data for modeling. The flowchart shown in Figure 2 illustrates the main steps for data processing.

Our model will give a prediction if an individual is a consumer of cannabis based on a variety of attributes, such as demographics and other drug usage information, etc. In other words, we have a limit in prediction to seven categories. To classify the target into different classes, it is appropriate to utilize classification methods. The topic might be beneficial for cannabis sellers who want to know their customers and lawmakers to determine if their constituents are actively using cannabis.

Applying supervised, unsupervised, and hybrid (combined supervised and unsupervised) classification methods to the project is feasible. For unsupervised modeling, we can pretend that we don’t know the target values in advance and make the prediction models. Some possible algorithms are Multinomial Logistic Regression, Naïve Bayes, K-Nearest Neighbors, Decision Tree, and Support Vector Machines.

Regression is not an option for this topic because regression is for the assignment of values to continuous output. In other words, the target must be numerical and continuous. Unlike classification, regression does not have a limit in forecasting, which means the target value can be out of the original data range.

2. Literature Review

Abdullah et al. (2018) did research on predicting drug users using backpropagation.They used a few years of existing data as the input to predict drug users in the next upcoming year. The advantages of using back propagation methods have led to high accuracy of predictions by minimizing errors between the actual and the predicted values.

In research from Fehrman et al. (2015), the individual’s risk of drug abuse was evaluated based on the online-collected data set, consisting of demographic, personality traits, impulsivity, and sensation-seeking information.They deployed eight classification algorithms, including Random Forest, Decision Tree, Logistic Regression, Naïve Bayes, k-nearest Neighbors, Linear discriminant Analysis, Probability Density Function Estimation, and Gaussian Mixture, to classify users and non-users for 18 central nervous system psychoactive drugs. The experiment results showed that the performance of these models was surprisingly good based on sensitivity and specificity metrics which were mainly greater than 70%, evaluated by the cross-validation method. The best results for crack, ecstasy, cannabis, VSA, and LSD were greater than 75%. The K-nearest Neighbors model is the best classifier for caffeine and chocolate users. The Linear Discriminant Analysis is the best model to classify alcohol users. The Decision Tree is the best model to forecast the usage of most drugs.

In addition, Fehrman et al. (2015) also found that the consumption of some drugs is highly correlated. They employed correlation analysis using two correlation measures: the Relative Information Gain and the Pearson Correlation Coefficient. The research concluded that there were three correlation Pleiades of drugs: the heroin pleiad, the ecstasy pleiad, and the benzodiazepines pleiad, defined for the decade-, year-, month-, and week-based classifications. The decade-based correlation study presented that the usage of legal drugs, like caffeine, chocolate, and benzodiazepines, did not correlate with the usage of other drugs. However, the consumption of illegal drugs, like cocaine, LSD, and amphetamines, is correlated symmetrically.

Qiao et al. (2019) deployed four machine learning models, k-nearest Neighbors, Random Forest, Extreme Gradient Boosting, and Light Gradient Boosting Machine, to detect potential individuals of drug abuse and predict the last consumption time based on the demographic and personality traits information. Compared to KNN, RF, and XGBoost, the LightGBM had excellent performance, with the highest accuracy (0.8117, 0,7500) both on classifying drug users and estimating consumption time.

This paper also extracted important features based on RF, XGBoost, and LightGBM separately. They found that conscientiousness, neuroticism, and openness to experience are the most prominent features for detecting amyl nitrite users. Neuroticism, conscientiousness, and agreeableness are the key features for classifying methamphetamine users.

Hu et al. (2019) utilized an experiment consisting of 494 boys and 206 girls to predict the trajectory of substance use from an early age. Their starting ages ranged from 10-12 years of age and were periodically "followed up at 12-14, 16, 19, 22, 25, and 30 years of age" (p.1). After gathering data consisting of a wide variety of information on each person's health at each stage of life. They indexed the info using a substance use severity (SUS) index. This index targeted the harmfulness of any drugs consumed before the check up occurred. Hu et al. (2019) states that this index takes into account "fourteen substances that may have been used at least once during the past three years (heroin, cocaine, barbiturates, methadone, alcohol, benzodiazepines, amphetamine, tobacco, cannabis, solvents, LSD, anabolic steroids, ecstasy and nitrit"(p.2)

The SUS index score took the sum of usage multiplied by the harm that the drug caused on the body. Where the sum consists of each substance that they took. They evaluated the feature importance and determined the primary features that they would use in their machine learning models. After evaluating the features and determining importance they utilized six machine learning models to evaluate the trajectory of the drug usage per person. The six models consist of random forest, support vector machines, naive bayes, adaptive boost, nearest neighbors, and artificial neural network. After evaluating each model, they found that random forest and naive bayes outperformed the other four models but were almost equivalent in terms of performance.

Islam et al. (2020) target is to classify a person's "vulnerability towards substance abuse by analyzing subjects' socio-economic environment" (p.1). They collected data by utilizing a questionnaire that specifically targeted common factors that would lead a person to abuse drug usage. They used Pearson's chi-squared test to identify important features for machine learning and index the results they received from the questionnaire. This was done by assessing the accuracy score of each feature and comparing those scores to each other. This reduced the number of features to 18, which ranged from problems with family to religious affiliations. After feature selection, they decided to utilize Random Forest, KNearest Neighbors, Decision Tree, Linear SVC, Gaussian Naive Bayes, and Logistic Regression. They evaluated each machine learning model using ROC, AUC, and accuracy. This resulted in logistic regression being the most accurate out of all the models that they tested.

Cuttler et al. (2016) used a large sample set from an anonymous online survey that assessed cannabis usage practices and their effects on short-term and withdrawal. Cuttler (2016) received a dataset containing 57.7% male and 43.3% female, with a majority being college educated in someway; at least an associate degree. The study aimed was to examine the different reactions to cannabis based on sex. They leveraged the p-score on their data to determine what was helpful information. The exhaustive survey covered many known adverse effects of cannabis and the reaction expected from the use of cannabis, both positive and negative. They found that women overall suffered more negative effects than their male counterparts.

To find the relationship between nicotine usage and cannabis usage, Steinberg (2022) conducted a survey to gather data and analyze it. Multivariate logistic regression was used to determine the relationship between nicotine and cannabis as Steinberg (2022) accepted all mediums of usage on both drugs. They utilized the null hypothesis and p-score value to determine the significance of the data to conclude the results. They found that there was a relationship between nicotine usage and cannabis usage, but more research needs to be done.

3. Feature selection methods

3.1 Data Preparation

Before performing feature selection, we should deploy data preparation first. The main steps of data preparation are shown in Figure 2. We first downloaded the raw data set (Table 5) from the UCI data repository. Then we performed data pre-processing, like checking missing and outlier values, encoding categorical features, and selecting the most informative attributes. After that, we standardized the data set since there are different scales with different features. Finally, the data was split into 80% training and 20% testing datasets. The final data set is shown in Table 6. Figure 3 shows that Cannabis is still the most balanced feature after data preparation.

Figure 2

Flowchart of Data Preparation

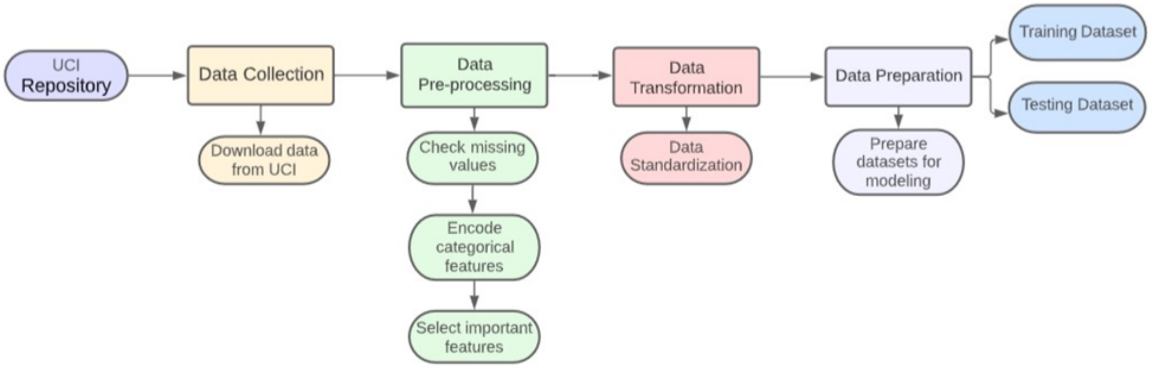
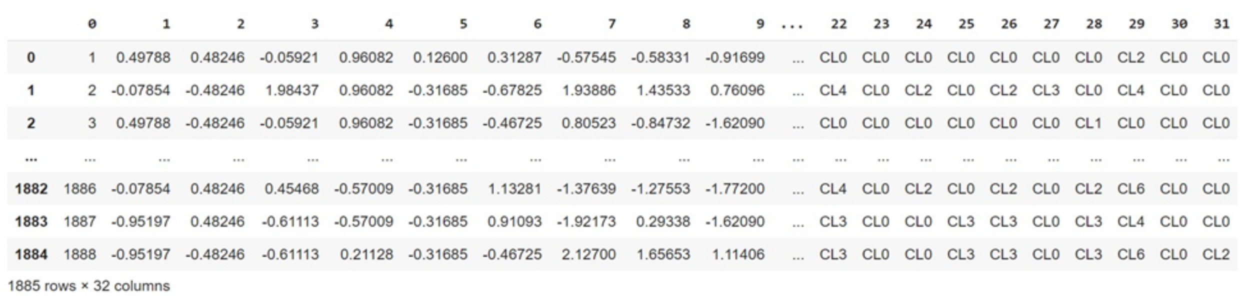


Table 5

The Sample of Raw Data Set

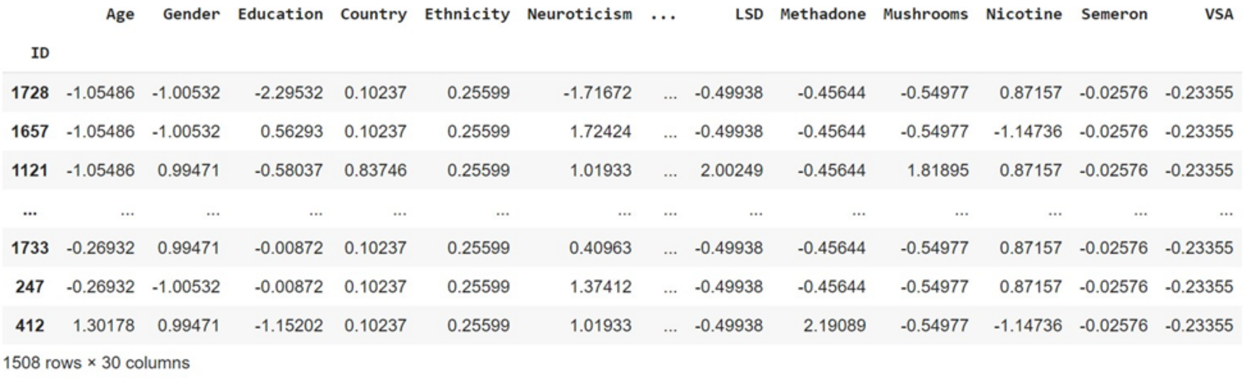


	0	1	2	3	4	5	6	7	8	9	...	22	23	24	25	26	27	28	29	30	31
0	1	0.49788	0.48246	-0.05921	0.96082	0.12600	0.31287	-0.57545	-0.58331	-0.91699	...	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL2	CL0	CL0
1	2	-0.07854	-0.48246	1.98437	0.96082	-0.31685	-0.67825	1.93886	1.43533	0.76096	...	CL4	CL0	CL2	CL0	CL2	CL3	CL0	CL4	CL0	CL0
2	3	0.49788	-0.48246	-0.05921	0.96082	-0.31685	-0.46725	0.80523	-0.84732	-1.62090	...	CL0	CL0	CL0	CL0	CL0	CL0	CL1	CL0	CL0	CL0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1882	1886	-0.07854	0.48246	0.45468	-0.57009	-0.31685	1.13281	-1.37639	-1.27553	-1.77200	...	CL4	CL0	CL2	CL0	CL2	CL0	CL2	CL6	CL0	CL0
1883	1887	-0.95197	0.48246	-0.61113	-0.57009	-0.31685	0.91093	-1.92173	0.29338	-1.62090	...	CL3	CL0	CL0	CL3	CL3	CL0	CL3	CL4	CL0	CL0
1884	1888	-0.95197	-0.48246	-0.61113	0.21128	-0.31685	-0.46725	2.12700	1.65653	1.11406	...	CL3	CL0	CL0	CL3	CL3	CL0	CL3	CL6	CL0	CL2

1885 rows × 32 columns

Table 6

The Sample of Training Data Set Before Feature Selection

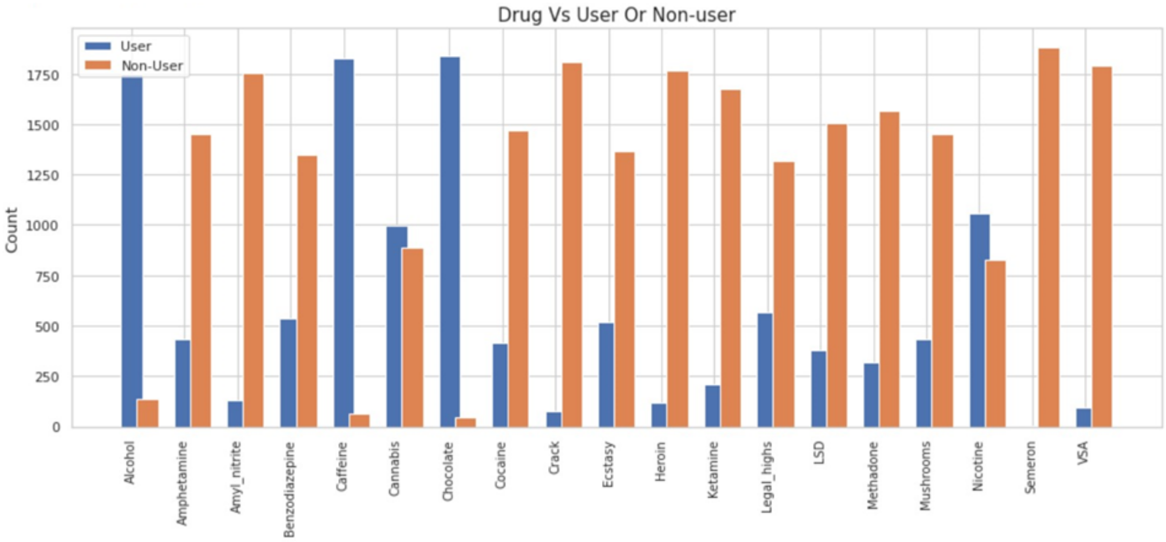


	Age	Gender	Education	Country	Ethnicity	Neuroticism	...	LSD	Methadone	Mushrooms	Nicotine	Semeron	VSA
ID													
1728	-1.05486	-1.00532	-2.29532	0.10237	0.25599	-1.71672	...	-0.49938	-0.45644	-0.54977	0.87157	-0.02576	-0.23355
1657	-1.05486	-1.00532	0.56293	0.10237	0.25599	1.72424	...	-0.49938	-0.45644	-0.54977	-1.14736	-0.02576	-0.23355
1121	-1.05486	0.99471	-0.58037	0.83746	0.25599	1.01933	...	2.00249	-0.45644	1.81895	0.87157	-0.02576	-0.23355
...	...	...	...	...	...	...	...	...	...	...	...	...	...
1733	-0.26932	0.99471	-0.00872	0.10237	0.25599	0.40963	...	-0.49938	-0.45644	-0.54977	0.87157	-0.02576	-0.23355
247	-0.26932	-1.00532	-0.00872	0.10237	0.25599	1.37412	...	-0.49938	-0.45644	-0.54977	0.87157	-0.02576	-0.23355
412	1.30178	0.99471	-1.15202	0.10237	0.25599	1.01933	...	-0.49938	2.19089	-0.54977	-1.14736	-0.02576	-0.23355

1508 rows × 30 columns

Figure 3

The Count of User and Non-user for Different Drugs



3.2 Feature selection

3.2.1 Logistic Regression Combined Hypothesis Test Method

The null hypothesis (H0) is that the value of the coefficient is zero, which means the corresponding feature is statistically insignificant and has no effect on the target feature. The alternative hypothesis (H1) is that the value of the coefficient is not zero, which means the corresponding feature is statistically significant and influences the target feature.

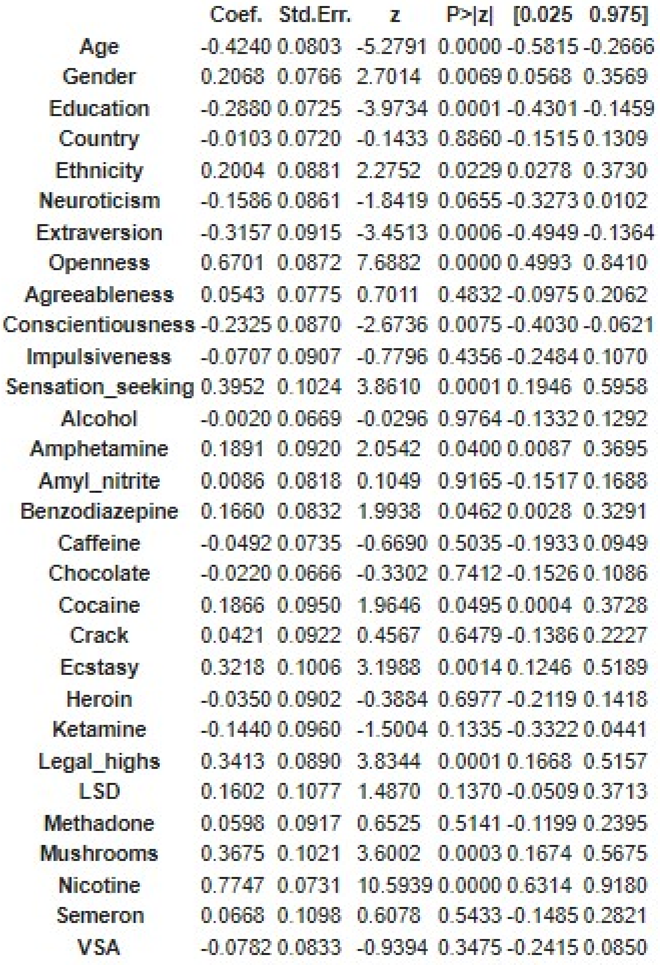
- Null hypothesis (H0): ai = 0 (no effect),
- Alternative hypothesis (H1): ai ≠0 (important).

The p-values based on logistic regression are shown in Table 7. Based on a 99% confidence level, if the p-value < 0.01, we cannot accept H0, which means the corresponding feature is statistically significant. The important features selected are:

- Personal Characteristics: 'Age', 'Gender', and 'Education'.
- Personality: 'Extraversion', 'Openness', 'Conscientiousness', and 'Sensation_seeking'.
- Other Drugs: 'Ecstasy', 'Legal_highs', 'Mushrooms', and 'Nicotine'.

Table 7

P-values Based on Logistic Regression



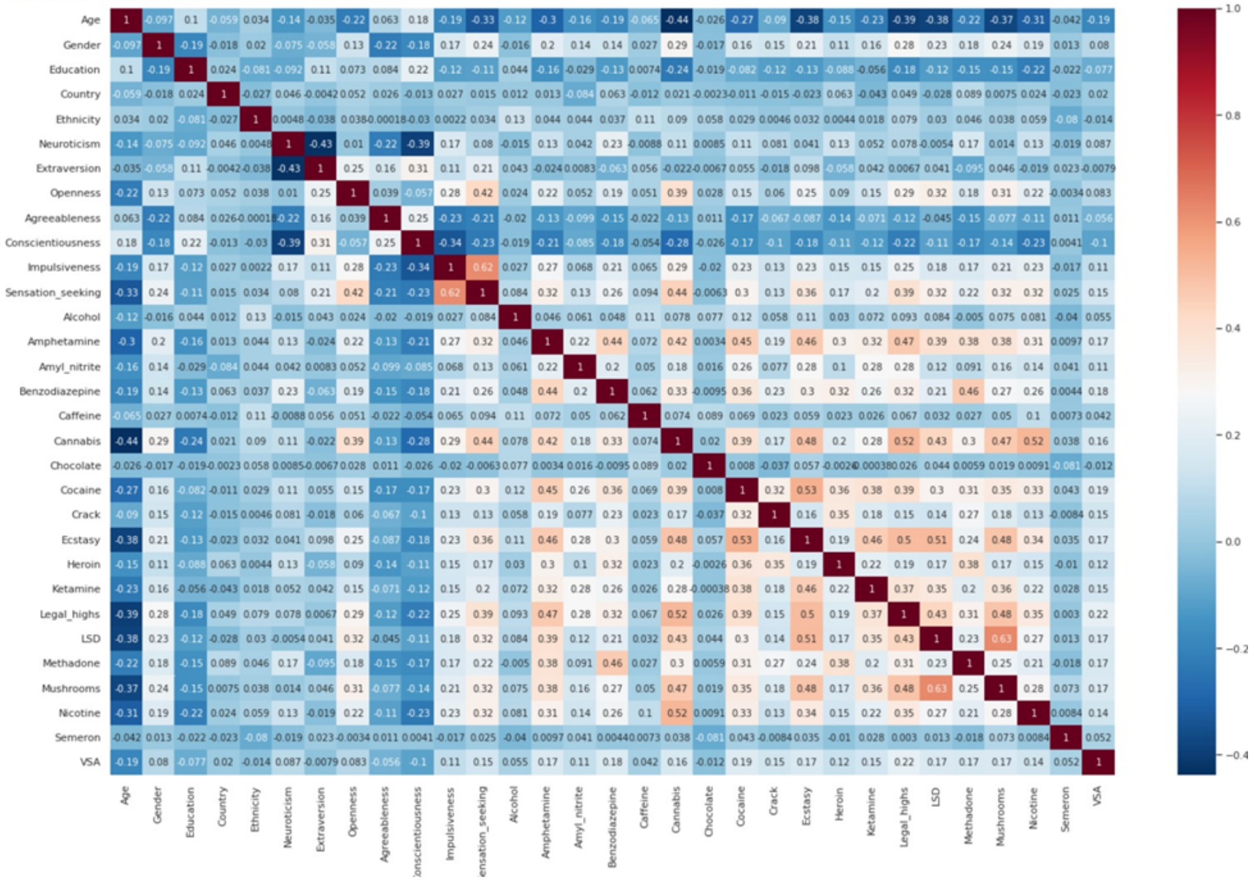
	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Age	-0.4240	0.0803	-5.2791	0.0000	-0.5815	-0.2666
Gender	0.2068	0.0766	2.7014	0.0069	0.0568	0.3569
Education	-0.2880	0.0725	-3.9734	0.0001	-0.4301	-0.1459
Country	-0.0103	0.0720	-0.1433	0.8860	-0.1515	0.1309
Ethnicity	0.2004	0.0881	2.2752	0.0229	0.0278	0.3730
Neuroticism	-0.1586	0.0861	-1.8419	0.0655	-0.3273	0.0102
Extraversion	-0.3157	0.0915	-3.4513	0.0006	-0.4949	-0.1364
Openness	0.6701	0.0872	7.6882	0.0000	0.4993	0.8410
Agreeableness	0.0543	0.0775	0.7011	0.4832	-0.0975	0.2062
Conscientiousness	-0.2325	0.0870	-2.6736	0.0075	-0.4030	-0.0621
Impulsiveness	-0.0707	0.0907	-0.7796	0.4356	-0.2484	0.1070
Sensation_seeking	0.3952	0.1024	3.8610	0.0001	0.1946	0.5958
Alcohol	-0.0020	0.0669	-0.0296	0.9764	-0.1332	0.1292
Amphetamine	0.1891	0.0920	2.0542	0.0400	0.0087	0.3695
Amyl_nitrite	0.0086	0.0818	0.1049	0.9165	-0.1517	0.1688
Benzodiazepine	0.1660	0.0832	1.9938	0.0462	0.0028	0.3291
Caffeine	-0.0492	0.0735	-0.6690	0.5035	-0.1933	0.0949
Chocolate	-0.0220	0.0666	-0.3302	0.7412	-0.1526	0.1086
Cocaine	0.1866	0.0950	1.9646	0.0495	0.0004	0.3728
Crack	0.0421	0.0922	0.4567	0.6479	-0.1386	0.2227
Ecstasy	0.3218	0.1006	3.1988	0.0014	0.1246	0.5189
Heroin	-0.0350	0.0902	-0.3884	0.6977	-0.2119	0.1418
Ketamine	-0.1440	0.0960	-1.5004	0.1335	-0.3322	0.0441
Legal_highs	0.3413	0.0890	3.8344	0.0001	0.1668	0.5157
LSD	0.1602	0.1077	1.4870	0.1370	-0.0509	0.3713
Methadone	0.0598	0.0917	0.6525	0.5141	-0.1199	0.2395
Mushrooms	0.3675	0.1021	3.6002	0.0003	0.1674	0.5675
Nicotine	0.7747	0.0731	10.5939	0.0000	0.6314	0.9180
Semeron	0.0668	0.1098	0.6078	0.5433	-0.1485	0.2821
VSA	-0.0782	0.0833	-0.9394	0.3475	-0.2415	0.0850

3.2.2 Correlation-based Method

The high correlation between two input features results in multicollinearity, which may affect the model’s performance. After deploying the logistic regression method, we will leverage the correlation-based method to select features further. There are various correlation metrics such as Pearson, Kendall, Spearman, etc. In this project, we choose the Pearson correlation coefficient. The correlations between features are shown in the following heat maps (Tables 8, 9).

Table 8

Correlations between Features Before Feature Selection Based on Logistic Regression



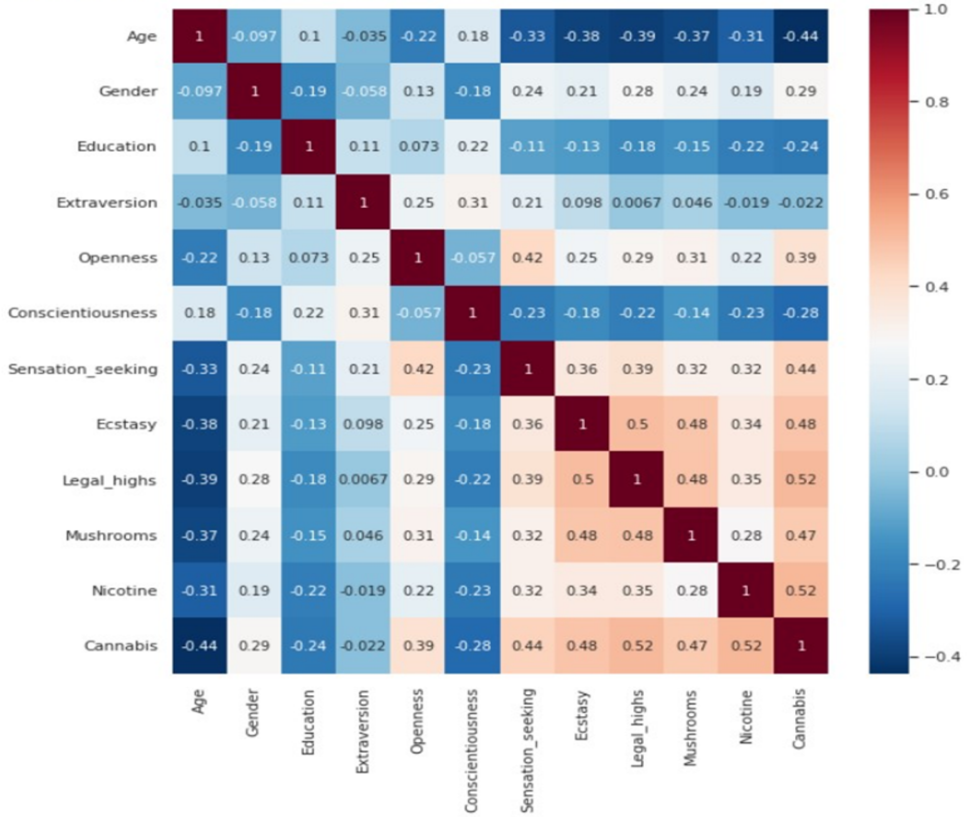
From the correlation heatmaps, the highest Pearson correlation coefficient between ‘Mushrooms’ and ‘LSD’ is 0.63. The results seem to be not bad. Further, after feature selection using logistic regression, the highest correlation between input features is reduced from 0.63 to 0.5. In research from Mendis (2019), there is no universal threshold, but a good heuristic is that multicollinearity is present with correlations above 0.5. So, there is no need to eliminate any features based on the correlation method. The final important features still are:

- Personal Characteristics: 'Age', 'Gender', and 'Education'.
- Personality: 'Extraversion', 'Openness', 'Conscientiousness', and 'Sensation_seeking'.
- Other Drugs: 'Ecstasy', 'Legal_highs', 'Mushrooms', and 'Nicotine'.

The sample of the final data set after feature selection is shown in Table 10.

Table 9

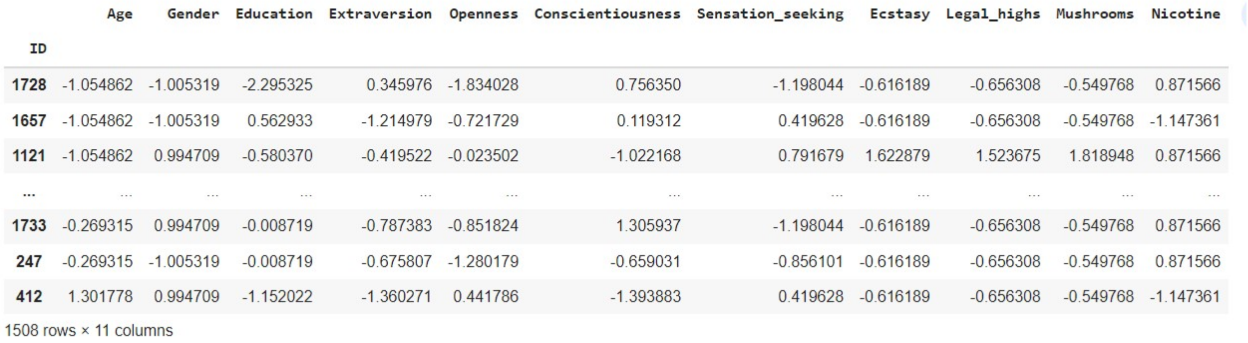
Correlations between Features After Feature Selection Based on Logistic Regression



	Age	Gender	Education	Extraversion	Openness	Conscientiousness	Sensation_seeking	Ecstasy	Legal_highs	Mushrooms	Nicotine	Cannabis
Age	1	-0.097	0.1	-0.035	-0.22	0.18	-0.33	-0.38	-0.39	-0.37	-0.31	-0.44
Gender	-0.097	1	-0.19	-0.058	0.13	-0.18	0.24	0.21	0.28	0.24	0.19	0.29
Education	0.1	-0.19	1	0.11	0.073	0.22	-0.11	-0.13	-0.18	-0.15	-0.22	-0.24
Extraversion	-0.035	-0.058	0.11	1	0.25	0.31	0.21	0.098	0.0067	0.046	-0.019	-0.022
Openness	-0.22	0.13	0.073	0.25	1	-0.057	0.42	0.25	0.29	0.31	0.22	0.39
Conscientiousness	0.18	-0.18	0.22	0.31	-0.057	1	-0.23	-0.18	-0.22	-0.14	-0.23	-0.28
Sensation_seeking	-0.33	0.24	-0.11	0.21	0.42	-0.23	1	0.36	0.39	0.32	0.32	0.44
Ecstasy	-0.38	0.21	-0.13	0.098	0.25	-0.18	0.36	1	0.5	0.48	0.34	0.48
Legal_highs	-0.39	0.28	-0.18	0.0067	0.29	-0.22	0.39	0.5	1	0.48	0.35	0.52
Mushrooms	-0.37	0.24	-0.15	0.046	0.31	-0.14	0.32	0.48	0.48	1	0.28	0.47
Nicotine	-0.31	0.19	-0.22	-0.019	0.22	-0.23	0.32	0.34	0.35	0.28	1	0.52
Cannabis	-0.44	0.29	-0.24	-0.022	0.39	-0.28	0.44	0.48	0.52	0.47	0.52	1

Table 10

The Sample of Training Data Set After Feature Selection



	Age	Gender	Education	Extraversion	Openness	Conscientiousness	Sensation_seeking	Ecstasy	Legal_highs	Mushrooms	Nicotine
ID											
1728	-1.054862	-1.005319	-2.295325	0.345976	-1.834028	0.756350	-1.198044	-0.616189	-0.656308	-0.549768	0.871566
1657	-1.054862	-1.005319	0.562933	-1.214979	-0.721729	0.119312	0.419628	-0.616189	-0.656308	-0.549768	-1.147361
1121	-1.054862	0.994709	-0.580370	-0.419522	-0.023502	-1.022168	0.791679	1.622879	1.523675	1.818948	0.871566
...	...	...	...	...	...	...	...	...	...	...	...
1733	-0.269315	0.994709	-0.008719	-0.787383	-0.851824	1.305937	-1.198044	-0.616189	-0.656308	-0.549768	0.871566
247	-0.269315	-1.005319	-0.008719	-0.675807	-1.280179	-0.659031	-0.856101	-0.616189	-0.656308	-0.549768	0.871566
412	1.301778	0.994709	-1.152022	-1.360271	0.441786	-1.393883	0.419628	-0.616189	-0.656308	-0.549768	-1.147361

1508 rows × 11 columns

4. Results

The four machine learning algorithms we apply are Naive Bayes, Logistic Regression, Decision Tree, and Random Forest. More details on the models will be discussed in section 5.1, Model Explanation.

Table 11 below shows the comparisons of the result before and after feature selections of our prediction among the four models. LR stands for Logistic Regression, NB means Naive Bayes, DT is Decision Tree, and RF represents Random Forest. One thing that really stands out in this comparison is that, in the Naive Bayes model, before feature selection, its precision and recall for non-drug users appear to be zero, which means there are no zero true positives.

However, after feature selection, the result improved a lot. To compare the F1 score among the models, the random forest model appears to have the highest score.

Table 11

Prediction Result Comparisons Among Models

	Precision								**Recall**							
	LR		**NB**		**DT**		**RF**		**LR**		**NB**		**DT**		**RF**	
Feature Selection	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After
Not Drug User	0.79	0.8	0	0.76	0.84	0.86	0.81	0.78	0.84	0.86	0	0.93	0.78	0.71	0.86	0.92
Drug User	0.85	0.87	0.53	0.92	0.82	0.77	0.87	0.91	0.81	0.81	1	0.73	0.87	0.9	0.82	0.78
	F1 Score								**Support**							
	LR		**NB**		**DT**		**RF**		**LR**		**NB**		**DT**		**RF**	
Feature Selection	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After
Not Drug User	0.82	0.83	0	0.83	0.81	0.77	0.83	0.84	177	177	177	177	177	177	177	177
Drug User	0.83	0.84	0.69	0.82	0.84	0.83	0.84	0.84	200	200	200	200	200	200	200	200
Accuracy	0.82	0.83	0.53	0.82	0.83	0.81	0.84	0.84	377	377	377	377	377	377	377	377

5. Discussion

5.1 Model Explanation

5.1.1 Without feature selection

As table 8 shows, there are some moderate connections before feature selection. Impulsiveness has a correlation of 0.623223 with Sensation_seeking, and that index for LSD and Mushrooms is 0.62661.

Naive Bayes assumes all features are conditionally independent, which is not true for our dataset because the data before feature selection has some correlated variables (as mentioned above). However, there is a chance that those variables do not strongly affect the performance of the model since the correlation of about 0.6 is not considered remarkable. Additionally, Naive Bayes expects all attributes to be relevant and contribute equally to the outcome. Logistic Regression Significance Test shows several irrelevant predictors before feature selection (initially we have 32 features, and it reduces to 11 features after feature selection). As a result, the performance of Naive Bayes can be significantly low before feature selection.

Logistic Regression also requires no multicollinearity among explanatory variables, but it is somewhat less sensitive than Naive Bayes. The threshold for high multicollinearity is usually 0.6, and we just see a few in table 8; thus, it is possible that the raw features do not lower the performance of Logistic Regression so much.

On the other hand, other approaches like Decision Tree and Random Forest have their own feature selection techniques. For example, during the Decision Tree induction, the algorithm chooses the most informative variable to split at each node. If there are unimportant features, they are simply not picked. Random Forest has similar characteristics to Decision Tree as it is an ensemble of various decision trees. In other words, Decision Tree and Random Forest may perform well even before feature selection.

5.1.2 With feature selection

Table 9 shows the correlations between all independent variables after feature selection. Observe that the correlations between the chosen features are weak (no magnitude values are greater than 0.5).

- Naive Bayes: the weak correlations between all predictors in table 9 improve the performance of Naive Bayes. For this project, we apply the Gaussian Naive Bayes Classifier. The classifier works best if the predictors' values are continuous and follow Gaussian distributions. Observe that many of our selected features have Gaussian-like distributions, but the shapes are not perfect bell curves, and not all features have that same distribution. In other words, the effectiveness of the model can not be determined in advance.

- Logistic Regression assumes that the response variable has only two outcomes, and we have only two classes, 0 and 1, for the target. The algorithm also works well where the dataset is linearly separable, meaning there is a linear relationship between the explanatory variables and the logit of the response variable. Since this second condition is not tested in our project, it is difficult to claim if Logistic Regression will perform well before the model development phase. Last but not least, Logistic Regression expects no multicollinearity among explanatory variables, which fits our case after feature selection.

- Decision Tree has various advantages: the algorithm is easy to interpret, does not require normalization, does not require data scaling, missing could be handled during the modeling process, and so forth. Yet these characteristics do not necessarily benefit this project because the data have to go through the full data pre-processing procedure for comparison purposes. A disadvantage is Decision Tree has a higher risk of overfitting than Random Forest.

- Random Forest:

  - Works well with an unbalanced dataset, which fits our case. Although we try to choose the most balanced target and make it more smooth in the data processing phase, our classes are not extremely balanced (one of the reasons is that we do not apply any sampling techniques)
  - Leverages the power of multiple decision trees, which reduces overfitting issues of Decision Tree. Generally, it can be said that Random Forest is better than Decision Tree in terms of performance, especially with large data. However, as our dataset is not too large, the difference might be light.
  - A drawback of Random Forest is it has high training time, which should not be a concern for this project in general.

All in all, Random Forest seems to be the best candidate for our project.

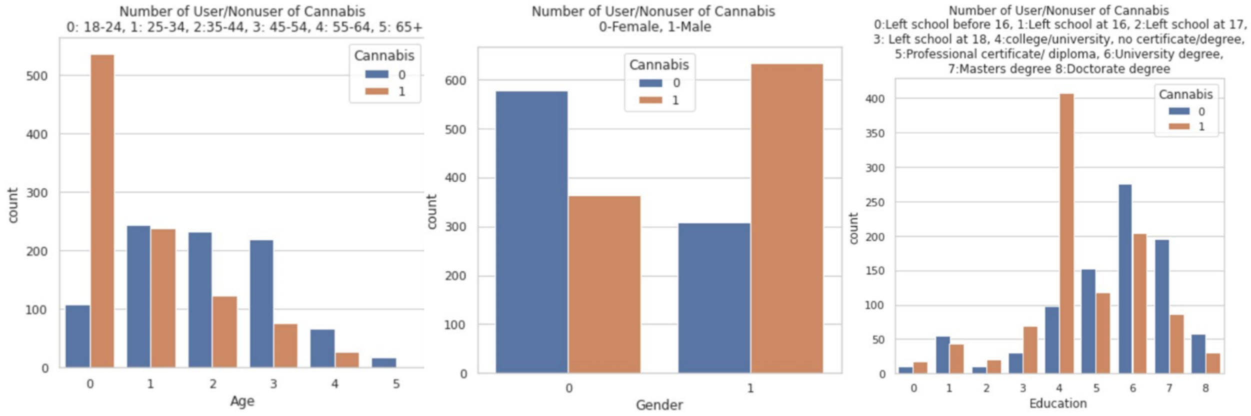
5.2 Feature Explanation

In the demographic characteristics, Figure 4, we see that marijuana is used in a specific range of groups. With age having high usage in young adults, usage leaning more towards men and college without a degree. NIDA (2019) has found that marijuana is the most commonly used drug after tobacco and alcohol. In addition, they've found that it's often the first illegal drug experiment for those coming of age. In many parts of the world, marijuana is increasingly being decriminalized or legalized with the usual target age similar to cigarettes or alcohol which is between 18 and 21. The fact that marijuana is considered a coming-of-age type of drug and the legal age around the drug, it explains why there is such high usage of marijuana at the younger spectrum than the older spectrum. Because marijuana is coming-of-age, the drug is linked to education.

Many individuals going to college view it as a coming-of-age or their first time being truly independent of their households. This is further emphasized through popular culture of movies, music, and tv shows. A common TV trope is using marijuana in college. This also creates an atmosphere and pressure for individuals to replicate what they see to enjoy the college experience. This feeling often subsides over time, or the reality of the employment forces them to re-evaluate their marijuana usage. Many employers still drug test, and marijuana is notoriously known to stay in one's system for long periods of time. This creates a significant incentive, combined with being satisfied with experimenting, for many to stop or reduce their usage after finishing college. This is why group 4, college/university; no degree has such a high positive response for marijuana usage. While this explains the importance of age and education, it doesn't explain the gender imbalance.

Figure 4

Number of User/Nonuser for Cannabis Based on Personal Characteristics

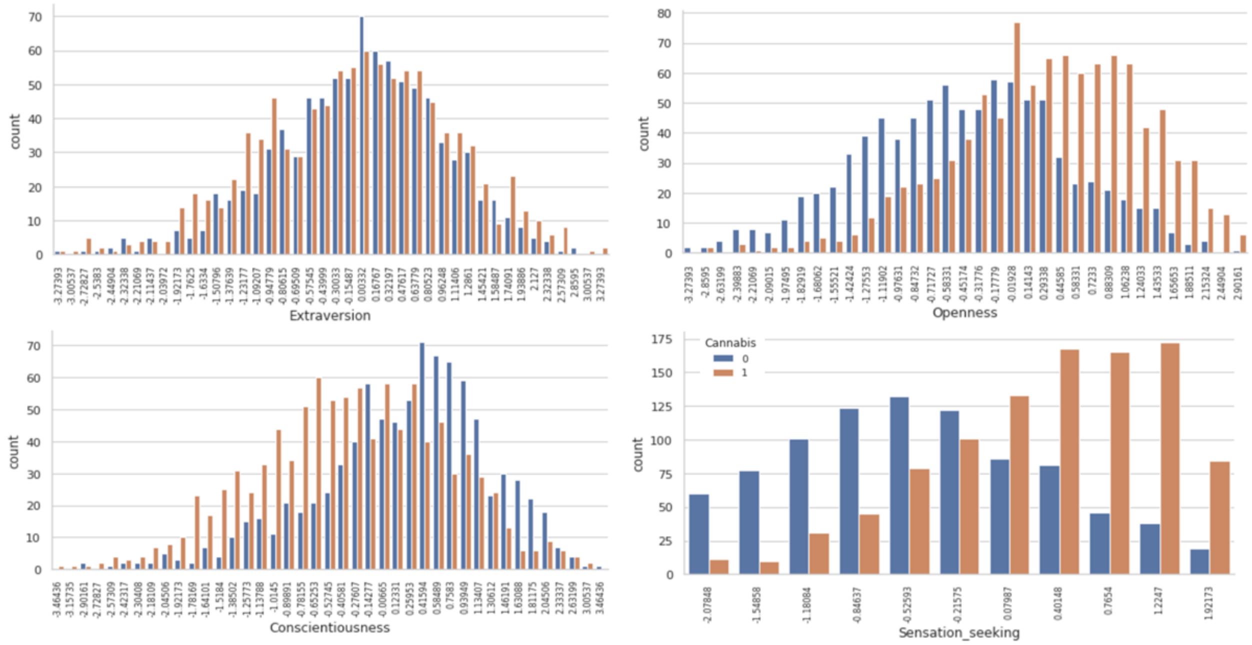


Cuttler et al. (2016) looks into why there is a disparity between cannabis usage among men and women. They found that men reported using cannabis more frequently and in higher quantities than women. Their results found that women were apprehensive before even considering using cannabis, and women were more prone to have negative effects. Women were apprehensive about using cannabis for many reasons but one primary is that they were concerned of the effects it would have on their reproductive health, and what would happen if they used it while unaware they were pregnant. For women who still decided to use it, many reported higher negative effects of the drug than men, especially during withdrawal. Women reported more instances of nausea and anxiety. The conclusion of this specific research correlates with our feature data. Cannabis is considered a stimulant and depressant with one or the other playing a bigger role depending on the cannabis strain used and the individual. This fact places a greater importance on one's personality.

We found that openness and sensation-seeking personality traits are more likely to have used cannabis, shown in Figure 5. Marijuana is an experimental drug and also a drug that brings excitement through the societal expectation created. Those rated high on openness are more likely to experiment or be less held back by negative concerns. Those who are sensation-seeking are also not restrained by negative considerations and are excited about experimentation. Those who are likely Conscientiousness are less likely to use cannabis. Conscientious individuals have a desire to do a task well and to take obligations to others seriously. Marijuana is known to impair judgment, so it is logical to not use it as it is detrimental. Extraversion personalities are equally distributed for both cannabis and non-cannabis users. Unlike the other three personalities, extraversion has no simple linear relationship with cannabis. Figure 5 shows people with a higher or lower level of extraversion are prone to using cannabis. From Table 9, the correlation coefficient between extraversion and cannabis is only -0.022. But it is still selected as the critical feature because logistic regression can detect linear and nonlinear relationships. Extraversion measures how energetic, sociable, and friendly a person is. Since marijuana is both a stimulant and a depressant, it can aid in social events while also impairing one during a social event. This highly depends on the individual and/or the cannabis strain they use. Though because of societal expectations, marijuana does lean towards being a party drug, so one who is higher in Extraversion is more likely to use cannabis so they can fit in and not have a negative effect on the party. On the flip side, we see cannabis use also being higher among those who have a low score on extraversion, and that may be because they are using marijuana as a depressant, such as a sleep aid or a relaxer. Cannabis has also been said to be a gateway drug meaning that those who use cannabis are likely to use other drugs afterward. Some use cannabis after trying other drugs.

Figure 5

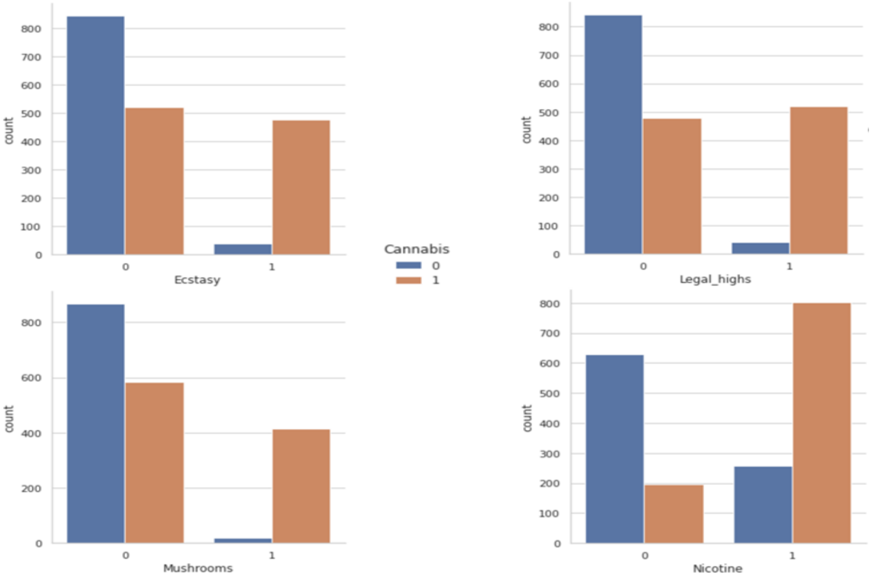
Number of User/Nonuser for Cannabis Based on Personality



In Figure 6, we found the drugs most correlated to Cannabis usage are ecstasy, legal highs, mushrooms, and nicotine. With those taking the drug, 1 on the x-axis, also having a high correlation of taking cannabis also. Legal highs are mood-altering or stimulant substances whose sale is not banned by current legislation. Legal highs, such as stimulants and sedatives, often provide similar effects as cannabis. This similar effect equally provides an incentive to try a more potent drug, cannabis, or the usage of cannabis makes them more willing to take the legal high. Mushrooms are a drug that causes hallucinations. Certain variants of marijuana cause hallucination but are often mild, and tolerance can be built.Mushrooms strongly provide this hallucination effect and make it tougher to build a tolerance. An interesting aspect of our data is that those who do not take ecstasy, legal high, or mushrooms still have a high usage of cannabis. This can be perceived that cannabis is a drug that leads one to utilize other drugs, but the relationship does not work the other way around. Nicotine, though, is unique in that when someone does nicotine, they are likely to use cannabis while its vice versa where no usage of nicotine results in a high likelihood of not using cannabis. Steinberg et al. (2022) found that 40% of medical marijuana users also use nicotine. As there are more non-medical marijuana users, it can be assumed that the rate of nicotine usage and marijuana usage overlap is much higher. The study found that cannabis users are likelier to use nicotine products than the general population. Also, cannabis is known to be easier to consume than nicotine and often in a form that is already familiar to nicotine users. This creates a strong relationship between cannabis usage and nicotine usage. The only real connection between marijuana and other drugs is nicotine, where those who don't take nicotine also don't take marijuana.

Figure 6

Number of User/Nonuser for Cannabis Based on Other Drug Consumptions



5.3 Result Explanation

The performances of the selected models are mostly improved after the feature selection process. The most significant effect of the feature selection process is on Naive Bayes model (accuracy score increases from 53% to 82%, precision, and F1 scores also elevate remarkably), while the effect is less on Logistic Regression (enhance indexes including precision, recall, F1 score, and accuracy score), and Random Forest's metrics slightly vary. A particular case is Decision Tree where the accuracy score decreases from 83% to 81%.

As explained in section 5.1, Logistic Significance Test helps reduce the number of correlated variables and irrelevant predictors, thus matching the assumption of Naive Bayes and boosting the results.

All Logistic Regression indexes benefit from the attribute selection process as they all improve. It is understandable because the Logistic Test derives from the Logistic Regression logic.

After feature selection, the Decision Tree tends to have slightly lower performance in terms of F1 score and accuracy score. A plausible explanation is that Decision Tree, and Logistic Regression have different feature selection criteria. For example, the Logistic Test in this project takes a confidence level of 99%, which might eliminate one or more variables that Decision Tree believes to be important.

Compared with Decision Tree, Random Forest is less sensitive to the impact of different feature selection methods because its process is different. While Decision Tree gives high importance to a particular set of explanatory variables, Random Forest chooses features randomly for each of the decision trees in its ensemble, meaning that not every tree sees all the features and is thus less prone to overfitting.

In conclusion, Random Forest has the best performance concerning accuracy and F1 scores, while other algorithms work reasonably well. For precision and recall, some models have the best results for the Not Drug User class, but worse results for the Drug User class; consequently, it is unreasonable to claim that an algorithm is the best.

References

Abdullah, D., Pardede, A. M., Umami, L., Manurung, R. T., Suryani, R., Surya, S., Saddhono, K., Mulyaningsih, I., Sudarsana, I. K., Brata, D. P. N., Mahatmaharti, R. A., Novziransyah, N., Amalia, A., Effendi, S. U., Samidah, I., & Murwati, M. (2019). Drug users prediction using back propagation EducationalMethod. *Journal of Physics: Conference Series*, *1361*(1), 012055. https://doi.org/10.1088/1742-6596/1361/1/012055

Cuttler, C., Mischley, L. K., & Sexton, M. (2016). Sex Differences in Cannabis Use and Effects: A Cross-Sectional Survey of Cannabis Users. Cannabis and cannabinoid research, 1(1), 166–175. https://doi.org/10.1089/can.2016.0010

Fehrman, E., Mirkes, E., Muhammad, A., Egan, V., & Gorban, A. (2015, February 20). The five factor model of personality and evaluation of drug consumption risk. arXiv.org. https://arxiv.org/abs/1506.06297

Hu, Z., Jing, Y., Xue, Y., Fan, P., Wang, L., Vanyukov, M., Kirisci, L., Wang, J., Tarter, R. E., & Xie, X.-Q. (2020). Analysis of substance use and its outcomes by Machine Learning: II. derivation and prediction of the trajectory of substance use severity. *Drug and Alcohol Dependence*, *206*, 107604. https://doi.org/10.1016/j.drugalcdep.2019.107604

Islam, U. I., Sarker, I. H., Haque, E., & Hoque, M. M. (2021). Predicting individual substance abuse vulnerability using machine learning techniques. *Hybrid Intelligent Systems*, 412–421. https://doi.org/10.1007/978-3-030-73050-5_42

Mendis, A. (2019b, September 17). *How Bad is Multicollinearity?* KDnuggets. Retrieved December 11, 2022, from https://www.kdnuggets.com/2019/09/multicollinearity-regression.html

NIDA. 2019, December 24. Cannabis (Marijuana) DrugFacts. Retrieved from https://nida.nih.gov/publications/drugfacts/cannabis-marijuana on 2022, December 12

Qiao, Z., Chai, T., Zhang, Q., Zhou, X., & Chu, Z. (2019, November). Predicting potential drug abusers using machine learning techniques. 2019 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS). https://doi.org/10.1109/iciibms46890.2019.8991550

Steinberg, M. L., Rosen, R. L., Billingsley, B., Shah, D., Bender, M., Shargo, K., Aamir, A., & Bridgeman, M. B. (2022). Tobacco/nicotine use among individuals using cannabis for therapeutic purposes. *The American Journal on Addictions*, *31*(6), 486–493. https://doi.org/10.1111/ajad.13323