Machine Learning Techniques for Forecasting Drug Addiction

This paper is designed as examination project within the course "Machine Learning" at the University of Konstanz

Roberto Daniele Cadili
Matriculation No.: 906991
roberto-daniele.cadili@uni-konstanz.de

Burak Özturan
Matriculation No.: 944663
burak.oezturan@uni-konstanz.de

1. INTRODUCTION

Addiction is a psychological and physical inability to stop consuming a chemical, drug, or substance, even though it is causing psychological and physical harm. The term addiction does not only refer to dependence on substances such as heroin or cocaine. An individual who cannot stop taking a particular drug or chemical has a substance dependence. It is very often hard to determine which personality and social factors are more likely to induce individuals to consume drugs, due to drug addicted individuals' reluctancy to acknowledge their condition and the social stigma surrounding such topics in public debate. Having the opportunity of forecasting accurately which individuals are more exposed to such a risk may help both policy makers and physicians to design more effective prevention campaigns and to ensure the prompt administration of treatment to individuals in need.

This paper is based on [4] and analyses the performance of several standard machine learning techniques in forecasting drug addicted from non-drug addicted individuals. The paper is composed of three main parts. Firstly, the dataset is described and problems within the data and setting are discussed. The second part gives an overview on the used algorithms and their general mechanics. The third part summarizes the results and performances of the algorithms, and suggests possible ways of improving classifiers' accuracy.

2. DATA DESCRIPTION

Data were collected by Elaina Fehrman via using an online survey tool from Survey Gizmo providing high standard of anonymity, between March 2011 and March 2012 under the scope of psychological research at the University of Leicester School of Psychology [4].

In this section, we will analyze the main characteristics of the dataset. The study found 2051 respondent, who claim to be above 18 years old, over a 12-month period. 166 of these respondents did not give consistent answers according to validity check built into the survey, and, thus, they were excluded to abolish non-informant answers. Furthermore, 9 people are

found that they might overclaim in their drug usage because they said that they are also consuming a drug named 'semeron' which does not exist in reality. But those 9 people were not excluded and the final dataset contains 1885 respondents, 943 of which are males, and 942 females.

Most of the respondents (93.7%) are English speakers, and they are mostly from the UK (1044; 55.4%), the US (557; 29.5%), Canada (87; 4.6%), Australia (54; 2.9%), Ireland (20;1.1%) and New Zealand (5; 0.3%). The rest, 118 respondents (6.3%) come from other countries, none of which reached 1% of the sample individually.

Furthermore, to increase anonymity, instead of mentioning their age, respondents reported their age range; 18-24 years (643; 34.1%), 25-34 years (481; 25.5%), 35-44 years (356; 18.9%), 45-54 years (294; 15.6%), 55-64 (93; 4.9%), and over 65 (18; 1%). It means that, although the majority of the sample consist of young people (34.1%), adult respondents were also represented (40% of the sample is 35 or above).

When it comes to education, it seems that the sample has quite high standards; just under two thirds of the sample (59.5%) at least have degree or professional certificate level: 14.4% (271) of them claims that they hold a professional certificate or diploma, 25.5% (481) has an undergraduate degree, 15% (284) a master's degree, and 4.7% (89) a doctorate. Almost 26.8% (506) of the sample had received some college or university education, although they did not finish their studies and therefore do not hold any certificates. Lastly, 13.6% (257) had left school at the age of 18 or younger.

Participants are also asked to indicate their racial identity to which they feel that they belong to. The vast majority (91.2%; 1720) identifies as white, followed by smaller groups (1.8%; 33) who reported themselves as black, and (1.4%; 26) as Asian. The 7 remainder of the sample (5.6%; 106) reported themselves as 'Other' or 'Mixed' categories. Due to overwhelming majority of respondents consisting of white people, it is hard to conduct an analysis including racial categories.

Personality measurements

To assess personality traits of the sample, the Revised NEO Five-Factor Inventory (NEO-FFI-R) questionnaire was employed [3]. The scale is a 60-item inventory (survey) comprised of five personality domains or factors. The five factors are: N, E, O, A, and C with 12 items per each domain. The five traits can be described as [4]:

1. Neuroticism (N) is a long-term tendency to experience negative emotions such as 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) is a general appreciation for art, unusual ideas, and imaginative, creative, unconventional, and wide interests,

4. Agreeableness (A) is a dimension of interpersonal relations, characterized by altruism, trust, modesty, kindness, compassion and cooperativeness;

5. Conscientiousness (C) is a tendency to be organized and dependable, strong-willed, persistent, reliable, and efficient.

Respondents were asked to read the 60 NEO-FFI-R statements and indicate on a five-point Likert scale how much a given item applied to them (i.e. 0 = 'Strongly Disagree', 1 = 'Disagree', 2 = 'Neutral', 3 = 'Agree', to 4 = 'Strongly Agree'). In medical research of drug usage and personal traits, it is expected that drug usage is associated with high N and O, and low A and C. The influence of the E score is deemed to be drug specific [4].

The second measure is the Barratt Impulsiveness Scale (BIS-11) [9]. The BIS-11 is a 30-item self-report questionnaire, which measures the behavioral construct of impulsiveness, and comprises three subscales: motor impulsiveness, attentional impulsiveness, and non-planning [4]. The 'motor' aspect reflects acting without thinking, the 'attentional' component poor concentration and thought intrusions, and the 'non-planning' a lack of consideration for consequences [8]. The scale's items are scored on a four-point Likert scale and then items are aggregated, meaning that the higher BIS-11 scores, the higher the impulsivity level.

The third measurement is the Impulsiveness Sensation-Seeking (ImpSS), which consists of 19 TRUE/FALSE statements, measuring ImpSS accordingly.

Features such as Age, Education, Country, Ethnicity, NEOAC, BIS-11 and ImpSS are not numerical in their original form. They are either categorical or ordinal data. After collection, feature data were quantified by the data provider using specific techniques (e.g. polychoric correlation, non-linear CatPCA). However, these techniques are not in the scope of this project, therefore we will use the provided quantified version of features.

Thus far, we presented the original data set, but restricting the analysis only to these covariates would be limited because the consumption of some drugs might lead to usage of others. That's why, as literature suggests [6], we divided drugs into two groups, namely soft and hard drugs. Soft drugs (alcohol, cannabis, caffeine, chocolate, nicotine), which are legal or decriminalized in most countries, are used as covariates to improve the predictive ability of our classifiers. Impact of these soft drugs on human psychology and physiology are not as strong as the second group, namely hard drugs (amphetamine, amyl nitrite, benzos, cocaine, crack, ecstasy, heroin, ketamine, legal highs, LSD, methadone, mushroom, VSA). Both these drug groups are originally coded according to users' frequency of usage. But we re-coded them as users and non-users via method described in '*Drug Usage'* section for our analysis.

Predicting use of each drug separately with personality traits covariates might be misleading, in that some drugs have similar effects on human, i.e. they can be substitute for one another. In this scenario, if we look only at single drugs and do not consider their potential substitute our results would only capture one aspect of the big picture. To capture the whole picture, we decided to group some drugs into upper categorical groups based on high correlation between drugs. By doing so, we created 3 drug pleiades, namely Heroin Pleiade, Ecstasy Pleiade and Benzos Pleiade. We will analyze these characteristics in "3.2 Correlational Relationships" section.

Age	Gender	Education	Nscore	Escore	Oscore	Ascore	Cscore	Impulsive	SS
0.498	0.482	-0.059	0.313	-0.575	-0.583	-0.917	-0.007	-0.217	-1.181
-0.079	-0.482	1.984	-0.678	1.939	1.435	0.761	-0.143	-0.711	-0.216
0.498	-0.482	-0.059	-0.467	0.805	-0.847	-1.621	-1.015	-1.380	0.401
-0.952	0.482	1.164	-0.149	-0.806	-0.019	0.590	0.585	-1.380	-1.181
0.498	0.482	1.984	0.735	-1.633	-0.452	-0.302	1.306	-0.217	-0.216
2.592	0.482	-1.228	-0.678	-0.300	-1.555	2.040	1.631	-1.380	-1.549
1.094	-0.482	1.164	-0.467	-1.092	-0.452	-0.302	0.939	-0.217	0.080
0.498	-0.482	-1.738	-1.328	1.939	-0.847	-0.302	1.631	0.193	-0.526

Table 1: Dataset visualization sample.

Alcohol	Caff	Cannabis	Choc	Nicotine
1	1	0	1	1
1	1	1	1	1
1	1	1	1	0
1	1	1	1	1
1	1	1	1	1
1	1	0	1	1
1	1	0	1	1

1	1	0	1	0

Table 1. *Continued.*

Drug Usage

In the original dataset, participants were asked about their usage of 18 legal and illegal drugs (alcohol, amphetamines, amyl nitrite, benzodiazepines, cannabis, chocolate, cocaine, caffeine, crack, ecstasy, heroin, ketamine, legal highs, LSD, methadone, magic mushrooms, nicotine and volatile substance abuse (VSA) and lastly fictitious drug (semeron) to detect over-claimers. As already mentioned in the previous section, for our project we included soft drugs as covariates, so that the actual number of predicted drug consumption refers to 14 drugs.

For our binary classification task, respondent consumption habits were coded with 0 or 1, referring to non-users and users, respectively. To discriminate users from non-users, it was necessary to take into account their drug consumption history, such as when they started or dropped consuming drugs. Therefore, in the survey respondents are asked to state whether they 'Never Used this drug', 'Used it Over a Decade Ago', or in the 'Last Decade, 'Year', 'Month', 'Week', or 'Day' for each drug separately. Basically, participants which belong to the category 'Used in the last day' also belong to the categories 'Used in the last week', 'Used in the last month', 'Used in the last year' and 'Used in the last decade'.

As mentioned above, the original dataset did not have users and non-user labels. Formally, only an individual in the class 'Never used' can be called a non-user, but it is not meaningful in the sense that an individual who used a drug more than decade ago cannot be considered a drug user either for most biological and psychological applications. There are several possible ways to discriminate participants into groups of users and non-users for binary classification:

1. If 'Never Used' and 'Used Over a Decade Ago' respondents are assigned as non-users, and others are users, so this separation is called 'decade-based'.

2. If 'Never Used', 'Used Over a Decade Ago' and 'Used in Last Decade' respondents are assigned as non-users, and others are users, so this separation is called 'year-based'.

3. If 'Never Used', 'Used Over a Decade Ago', 'Used in Last Decade', 'Used in Last Year' respondents are assigned as non-users, and others are users, so this separation is called 'year-based'.

4. Finally, if ‘Never Used’, ‘Used Over a Decade Ago’, ‘Used in Last Decade’, ‘Used in Last Year’ and ‘Used in Last Month’ respondents are assigned as non-users, and others are users, so this separation is called as ‘month-based’. Therefore, it can also be analyzed for weekly and daily consumption.

One can try some particular drug for any reason; therefore, we decided that using ‘decade-based’ separation was more appropriate in our case. By doing so, we assumed that if an individual did not consume a specific drug for more than 10 years, he/she can be classified as non-user, similarly to ‘Never Used’ individuals.

Consequently, our results from now on derived from decade-based classification. Other separation analyses could also be done to further research to look at other drugs usage tracks.

Drug	Number of Users (and %)
Amphetamines	679; 36.02%
Amyl nitrite	370; 19.63%
Benzodiazepines	769; 40.80%
Cocaine	687; 36.45%
Crack	191; 10.13%
Ecstasy	751; 39.84%
Heroine	212; 11.25&
Ketamine	350; 18.57%
Legal Highs	762; 40.42%
LSD	557; 29.55%
Methadone	417; 22.12%
Mushrooms	694; 36.82%
VSA	230; 12.20%

Table 2: Number of users and values in % for all drugs.

3. DATA INSPECTION

In the next following subsections, we will provide a detailed description of the data. To this end, a comprehensive descriptive statistic of the input features will be presented and it will be complemented by graphical representations of these. Moreover, correlational relationships among personality factors, single drugs and drugs pleiades will be investigated and presented.

3.1 Descriptive Statistics

The analysis of the input features relies on the fundamental concepts of descriptive statistics, in particular measures of central tendency and dispersion (see Table 3). It is important to remember that all original input features were either categorical or ordinal, therefore a quantified version of these was used. This implies that the values summarized in Table 3 are not directly interpretable, but they need to be associated to their encoded categories or ordinal intervals. From Table 3, we can see that the average age of an individual in the dataset approaches more closely the interval of 25-34 years old. An average individual has equal probability of being male or female, given that the gender input feature is almost perfectly balanced. In terms of education, the average schooling achieved by an individual approaches more closely the ordinal label "professional certificate or diploma", suggesting that on average our interviewees have successfully achieved higher levels of education. As far as the N-scores and the SS-scores are concerned, we can affirm that the average individual has middle levels of experiencing negative emotions (Neuroticism), and middle level of sensation seeking (SS). Slightly above middle levels are the E-scores and the O-scores, suggesting that the average individual is slightly more inclined to manifest outgoing, warm, cheerful characteristics (Extraversion), as well as a general appreciation for unusual ideas, art, and unconventional thinking (Openness to experience). Upper-middle are the levels depicted by the A-scores and the C-scores for the average individual in the dataset. This suggests that, on average, the interviewee is an altruist, modest and compassionate (Agreeableness) individual, who has a fairly good tendency to be organized, strong-willed and reliable (Conscientiousness). Conversely, the Impulsive score indicates that the average individual has lower-middle levels of impulsiveness.

As far as the average consumption of legal drugs and cannabis in the dataset is concerned, the average value approaches 1 for alcohol, caffeine and chocolate, suggesting that the consumption of such legal drugs is on average very widespread. Considerably less widespread, although still very popular, is the average consumption of cannabis and nicotine for individuals in the dataset. It is important to mention that a major drawback of using the mean value is that it may be sensible to outliers (i.e. a few outlying observations), for its computation is based on information contained in all the observations in the dataset.

Measures of dispersion are informative about how much variability there is in the observations. The standard deviation is one of the most commonly used measures of dispersion and shows the relation that the set of observations has to the mean of the sample. A large standard

deviation indicates that the observations in the dataset have a great deal of dissimilarity. Conversely, a small standard deviation indicates a great deal of similarity between the observations. All input features in the dataset, which are relative to personality traits, show large standard deviations, approaching the value of 1.

Column1	Age	Gender	Education	Nscore	Escore	Oscore	Ascore
count	1885	1885	1885	1885	1885	1885	1885
mean	0.035	0.000	-0.004	0.00005	-0.0002	-0.0005	-0.0002
std	0.878	0.483	0.950	0.998	0.997	0.996	0.997
min	-0.952	-0.482	-2.436	-3.464	-3.274	-3.274	-3.464
25%	-0.952	-0.482	-0.611	-0.678	-0.695	-0.717	-0.606
50%	-0.079	-0.482	-0.059	0.043	0.003	-0.019	-0.017
75%	0.498	0.482	0.455	0.630	0.638	0.723	0.761
max	2.592	0.482	1.984	3.274	3.274	2.902	3.464

Table 3: Descriptive statistics (count, mean, standard deviation, min, max and quantiles) for input features for raw data.

	Cscore	Impulsive	SS	Alcohol	Caff	Cannabis	Choc	Nicotine
count	1885	1885	1885	1885	1885	1885	1885	1885
mean	-0.0004	0.007	-0.003	0.964	0.980	0.671	0.981	0.671
std	0.998	0.954	0.964	0.187	0.139	0.470	0.135	0.470
min	-3.464	-2.555	-2.078	0	0	0	0	0
25%	-0.653	-0.711	-0.526	1	1	0	1	0
50%	-0.007	-0.217	0.080	1	1	1	1	1
75%	0.585	0.530	0.765	1	1	1	1	1
max	3.464	2.902	1.922	1	1	1	1	1

Table 3. *Continued.*

It is interesting to notice that standard deviations almost equal to 1, and mean values close to 0 suggest that our data may approximate a normal distribution. Fig. 1, 2, 3, 4, 5, 6 and 7 show the distribution of the observations in the dataset for each personality trait. Indeed, by looking at the "bell shape" approximation of the data distributions, the hypothesis of normality seems satisfied. This appears to be especially plausible for the five personality factors, with the exception of the C-score distribution, which is slightly skewed to the left. Similarly, Impulsive, and SS-score distributions are slightly skewed to the right and to the left, respectively.

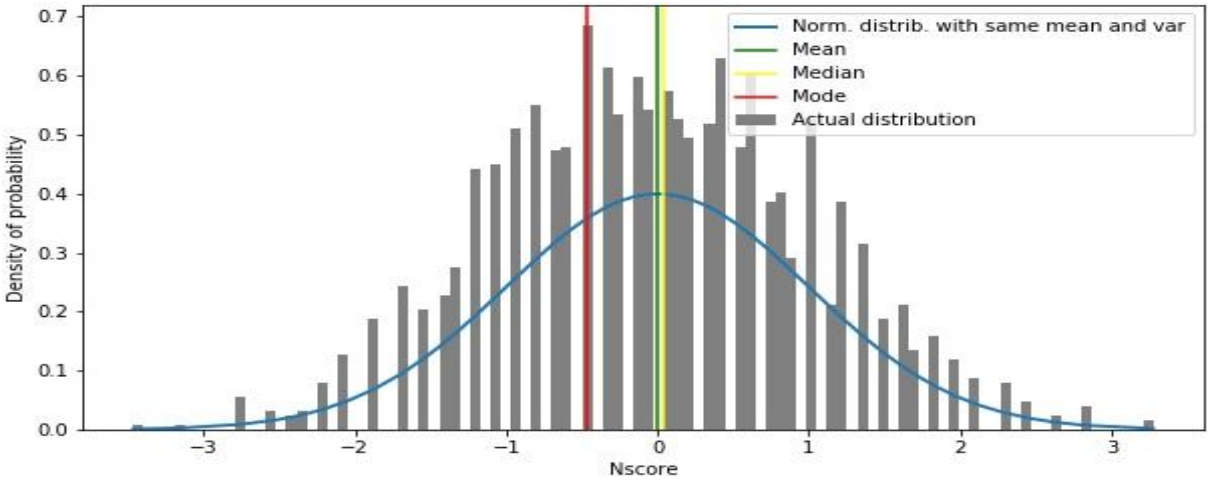

Fig. 1: Probability distribution of N-scores.

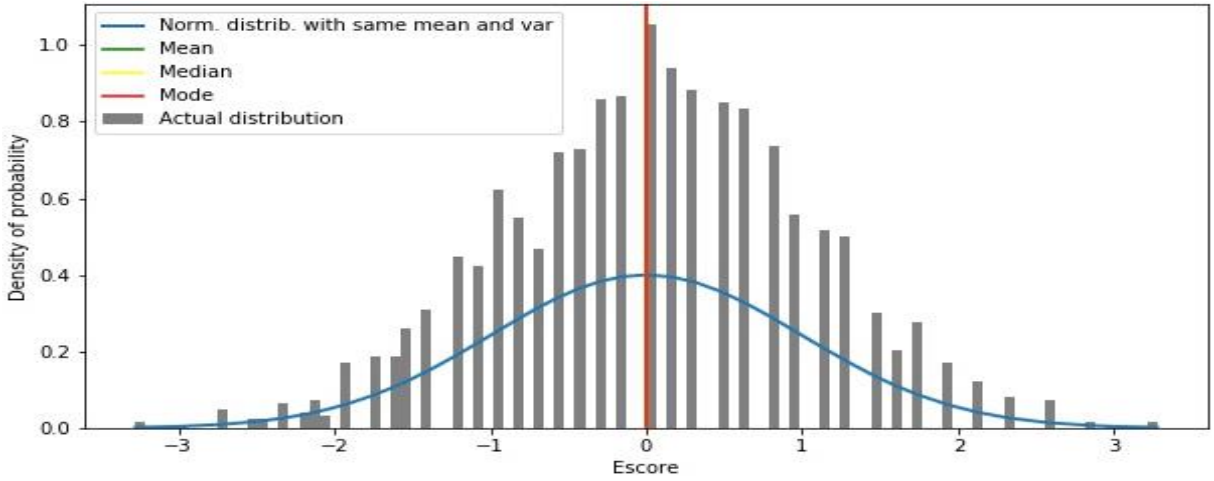

Fig. 2: Probability distribution of E-scores.

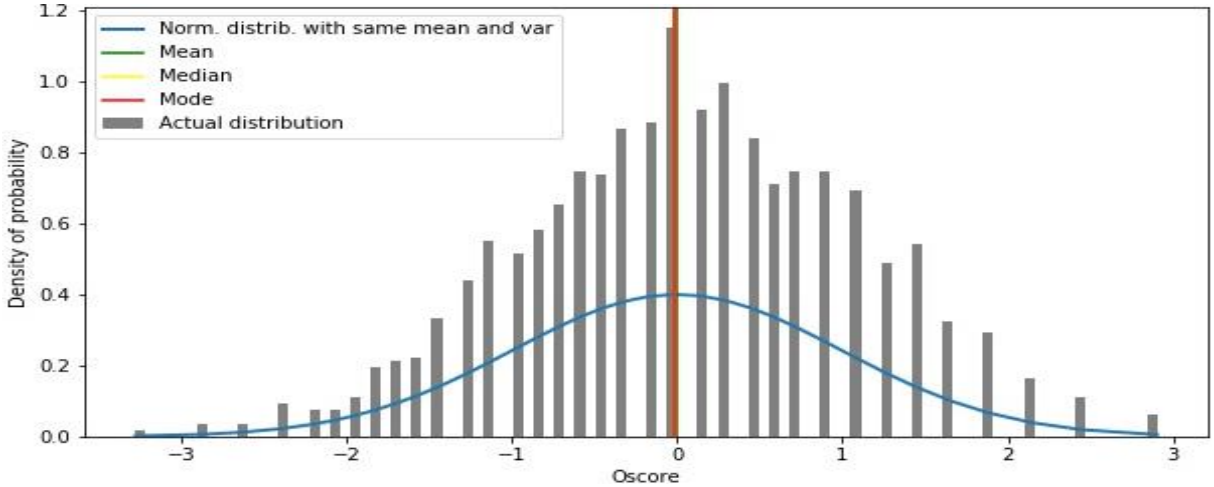

Fig. 3: Probability distribution of O-scores.

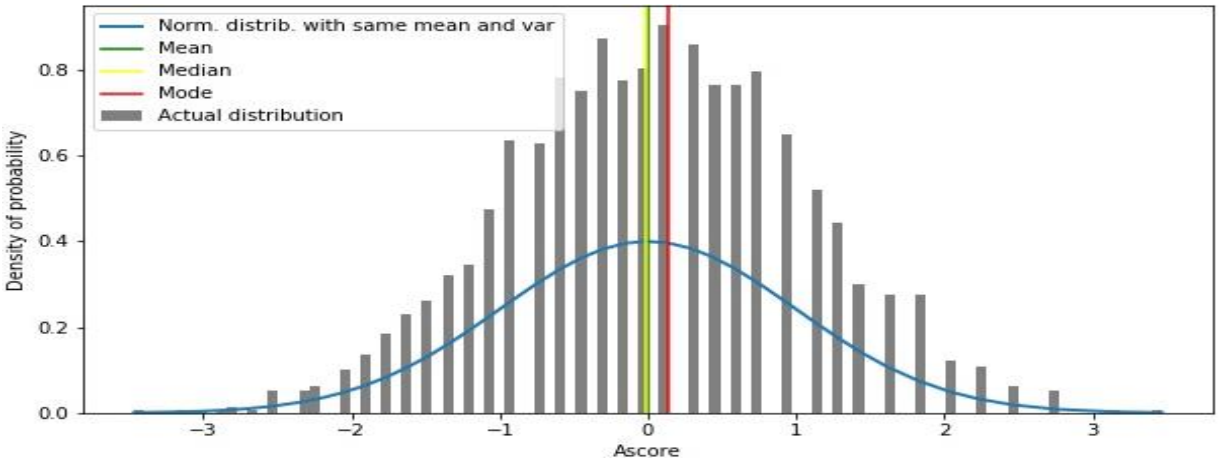

Fig. 4: Probability distribution of A-scores.

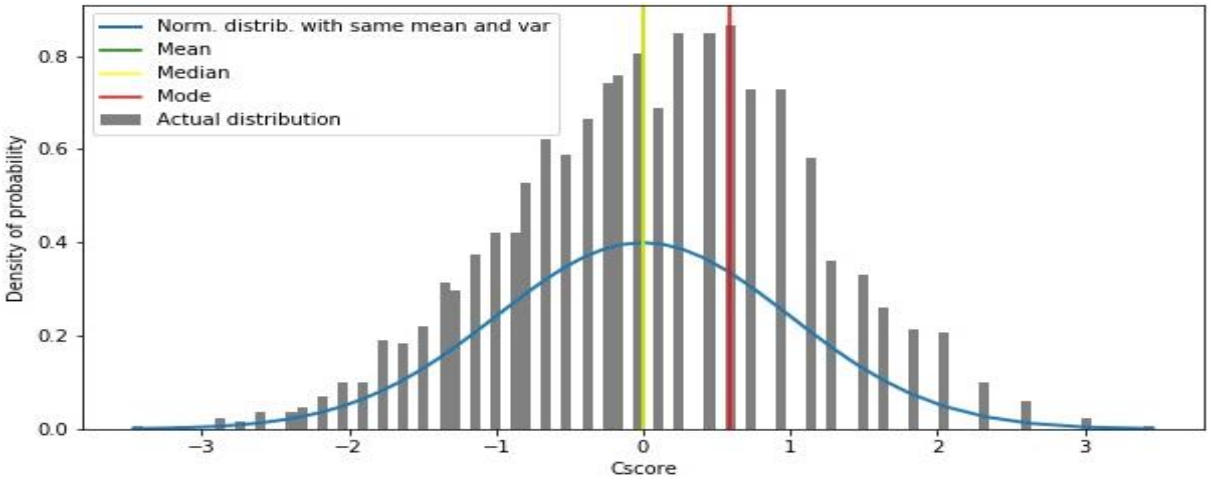

Fig. 5: Probability distribution of C-scores.

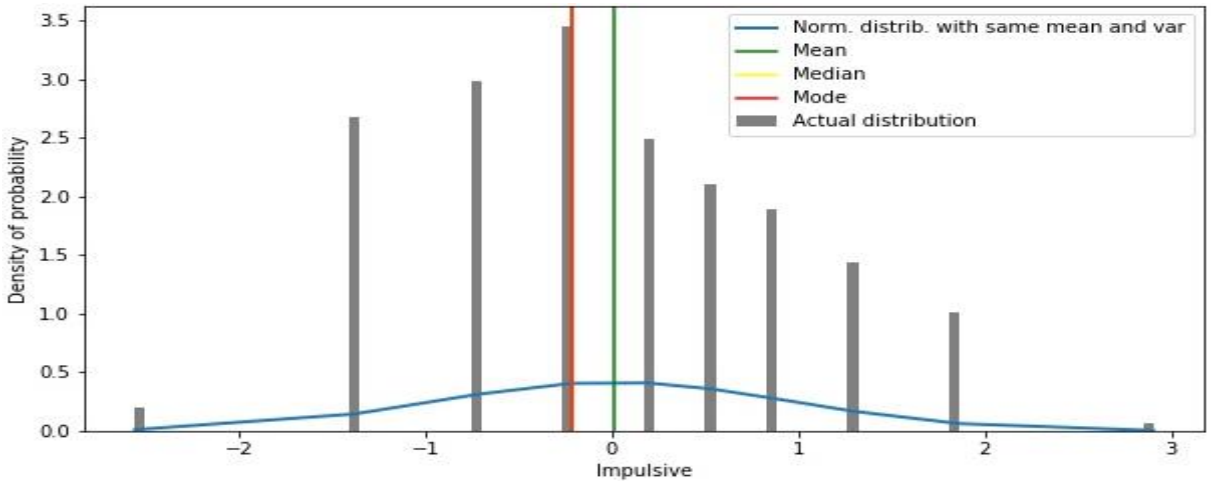

Fig. 6: Probability distribution of Impulsive scores.

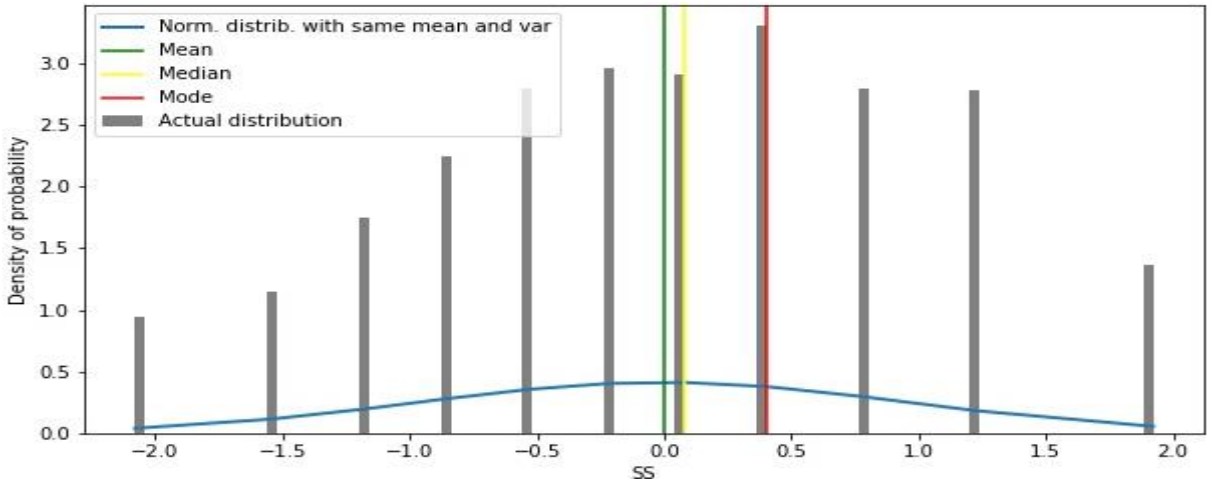

Fig. 7: Probability distribution of SS-scores.

3.2 Correlational Relationships

There are quite high correlational relationships between N-score, E-score and C-score. Whereas, two pairs of factors, N-score and O-score, and A-score and O-score, do not have significant correlation. The complete correlation matrix may be overviewed in Table 4:

	Nscore	Escore	Oscore	Ascore	Cscore
Nscore	1	-0.431	0.01	-0.217	-0.391
Escore	-0.431	1	-0.245	0.157	0.308
Oscore	0.01	0.245	1	0.039	-0.057
Ascore	-0.217	0.157	0.039	1	0.247
Cscore	-0.391	0.308	-0.057	0.247	1

Table 4: Correlation matrix for personality traits.

At the end of the “Data Description” section, we mentioned that grouping drugs into pleiades (groups) might enable us to capture the possibility that one drug might be substitute for another. By doing so, if an individual consumes one of the drugs in the pleiades, she/he can be classified as user of that pleiade and if she/he does not use any of them, she/he is non-user of that pleiade. Th first pleiade is the Heroin Pleiade where crack, cocaine, methadone and heroin are highly correlated with each other and heroin is the highest correlated drug for the rest of the drugs in the group.

Correlation coefficients are given in the Table 5 below.

	Crack	Coke	Meth	Heroin
Crack	1	0.396	0.367	0.509
Coke	0.396	1	0.351	0.414
Meth	0.367	0.351	1	0.494
Heroin	0.509	0.414	0.494	1

Table 5: Correlation matrix for drugs in Heroin pleiade.

The second pleiade is the Ecstasy Pleiade where ecstasy is the highest correlated drug for the group of amphetamine, cocaine, ketamine, LSD, legal highs, mushroom and ecstasy. Correlations between drugs in the group can be seen in the Table 6:

	Amphet	Coke	Ketamine	LSD	Legalh	Mushroom	Ecstasy
Amphet	1	0.58	0.412	0.49	0.481	0.481	0.597
Coke	0.58	1	0.449	0.442	0.445	0.48	0.633
Ketamine	0.412	0.449	1	0.462	0.393	0.436	0.512
LSD	0.49	0.442	0.462	1	0.518	0.68	0.599
Legalh	0.481	0.445	0.393	0.518	1	0.575	0.586
Mushroom	0.481	0.48	0.436	0.68	0.575	1	0.599
Ecstasy	0.597	0.633	0.512	0.599	0.586	0.599	1

Table 6: Correlation matrix for drugs in Ecstasy pleiade.

The third and last pleiade is the Benzos Pleiade representing highly correlated methadone, amphetamine, cocaine and benzodiazepines (see Table 7).

	Meth	Amphet	Coke	Benzos
Meth	1	0.415	0.351	0.468
Amphet	0.415	1	0.58	0.463
Coke	0.351	0.58	1	0.428
Benzos	0.468	0.463	0.428	1

Table 7: Correlation matrix for drugs in Benzos pleiade.

The complete correlation matrix for all drugs can be accessed in Appendix C.

4. MACHINE LEARNING TECHNIQUES

In this section, we will summarize the algorithms, including Linear Regression, Logistic Regression, LDA, QDA, NBC, KNN, SVM, NN, Decision Tree, which we used to conduct our task. Information are taken from Andrew Ng's online course on Stanford [7], and Elements of Statistical Learning [5].

4.1 Linear Regression

In linear regression, the basic idea is that y (output) is linear function of x (features, attributes). So more formally:

$$h(x) = \sum_{i=1}^{d} (\theta_i X_i) = \theta^T X, \tag{1}$$

which allow us to calculate the cost function as:

$$J(\theta) = \frac{1}{2} \sum_{i=1}^{n} (h_\theta(x^{(i)}) - y^{(i)})^2. \tag{2}$$

then we can calculate the coefficients by taking partial derivatives for each parameter. There are a few other gradient descent methods, but we will use stochastic gradient descent update method which is:

$$\theta := \theta + a\left(y^{(i)} - h_\theta(x^{(i)})\right)x^{(i)} \tag{3}$$

4.1.1 Regularized Linear Regression

In some cases, results can be subject to overfitting, meaning that weights are only valid for the specific trainset and they lack generalization for the test set. To avoid this, some regularization methods can be applied to the original cost function:

$$\theta_j := \theta_j - a\left[\left(\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)}\right) + \frac{\lambda}{m}\theta_j\right] \tag{4}$$

4.2 Logistic Regression

In linear regression, dependent variable values might not be in the probability range of 0-1. To do so logistic regression can be used and hypotheses become:

$$h_{\theta}(X) = \frac{1}{1+ e^{-(\theta_0+\theta_1 X)}} \tag{5}$$

consequently, the loss function of LR is

$$(\theta) = -\frac{1}{m}\sum\left[y^{(i)} \log\left((h_{\theta}(x^{(i)}) \right) + \left(1 - y^{(i)}\right) \log\left(1 - (h_{\theta}(x^{(i)}) \right) \right] \tag{6}$$

while writing the loss function, we are obtaining a convex function, and so the above-mentioned Stochastic Gradient Method and regularization techniques can be used to find optimal coefficients as in the linear regression case.

4.3 Linear Discriminant Analysis (LDA)

LDA is one of the generative classifiers which uses Bayes rule to calculate conditional probabilities:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{7}$$

Then it follows that each feature has multivariate normal distribution with the same variance-covariance matrix

$$f_k(x) = \frac{1}{(2\pi)^{m/2}|\Sigma|} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma^{-1}(x-\mu_k)} \tag{8}$$

Features have different mean for each class so that, at an input x, we predict:

$$\delta_k(x) = log\pi_k - \frac{1}{2}\mu_k^T \Sigma^{-1}\mu_k + x^T \Sigma^{-1}\mu_k \tag{9}$$

for different class k=0 and k=1 and where:

$$\delta_0(x) = \delta_1(x) \tag{10}$$

are the decision boundaries.

4.4 Quadratic Discriminant Analysis (QDA)

QDA is a relaxed version of LDA in the sense that now we do not have to assume that for each class the variance-covariance matrix is the same. Hence, the objective function become:

$$\delta_k(x) = log\pi_k - \frac{1}{2}\mu_k^T \Sigma^{-1}{}_k \mu_k + x^T \Sigma^{-1}{}_k \mu_k - \frac{1}{2} x^T \Sigma_k^{-1} x - \frac{1}{2}\log|\Sigma_k| \quad (11)$$

4.5 Naïve Bayes Classifier (NBC)

NBC assumes features independence, and therefore uses the extended Bayesian Rule:

$$P(y|x_1, \dots, x_n) = \frac{P(x_1|y)P(x_2|y)\dots P(x_n|y)P(y)}{P(x_1)P(x_2)\dots P(x_n)} \quad (12)$$

so the class can be assigned, which maximizes the probability:

$$y = argmax_y P(y) \prod_{i=1}^{n} P(x_i|y) \quad (13)$$

4.6 K-Nearest Neighbours (KNN)

K-Nearest Neighbours is a non-parametric method for classification and regression. This method learns by searching the pattern space for the KNN that are closest to the unknown sample. To measure the closeness between the unknown sample and the surrounding data points common distance metrics, such as Euclidean, Mahalonobis, and Lp distances can be applied. Then each observation of the test sample is labelled using the most common class in the k-Neighbourhood.

$$f(x) = \begin{cases} 1, & \sum_{i=1}^{n} w_i(x) 1_{\{Y_i=1\}} \geq \sum_{i=1}^{n} w_i(x) 1_{\{Y_i=0\}} \\ 0, & otherwise \end{cases} \quad (14)$$

The optimal k parameter can be found by grid search or cross validation.

4.7 Support Vector Machine

This algorithm tries to find a hyperplane maximizing the distance of two different points (one for class 1 and one for class -1), i.e. support vectors. This can be formally written as:

- Hyperplane = ($\overline{w}.\overline{x}$ +b) = 0
- y = {-1,1}
- w = the orthogonal vector to the hyperplane

Minimize $||w||^2$ under the constraint $y_i\ (w.x_i + b) - 1 = 0$, for each i to n.

Afterwards, applying Lagrangian function and differentiation, we can find the solution for w where a_i is the Lagrangian-Multiplicator:

$$w = \sum a_i y_i x_i \tag{15}$$

4.8 Neural Network

This algorithm tries to emulate human neurons and synapses, as far as the learning process is concerned. A neural network is usually made of many different neurons, called perceptrons, which include input layers where input features are provided. The information contained in the input layers is multiplied with some initial random weights and biases, and these products are then fed into an activation functions such as Sigmoid, Softmax or ReLU. After this stage, the information can be transferred to the next layer, where new your information can be extracted from the features and the learning process can progress. This process is repeated several times through feed-forward and back propagation, where weights are constantly updated until the network converges to produce optimal weights:

a. calculate the actual output:

$$y_j(t) = f\left[w(t) \cdot x_j\right] \tag{16}$$

$$= f[w_0(t)x_{j,0} + w_1(t)x_{j,1} + w_2(t)x_{j,2} + \cdots + w_n(t)x_{j,n}$$

b. update the weights:

$$w_i(t+1) = w_i(t) + r \cdot (d_j - y_j(t))\, x_{j,i} \quad for\ all\ features\ 0 \le i \le n \tag{17}$$

r is the learning rate between 0 and 1, larger values make the weight change faster; $y = f(z)$ denotes the output from the perceptron for an input vector; and (x_1,d_1), (x_2,d_2), … (x_s,d_s) are samples of the trainset, where: x_j is the n-dimensional input vector, and d_j is the desired output value of the perceptron for that input.

4.9 Decision Tree

In this algorithm, we are creating tree starting from the initial node, and splitting it with the best decision criterion until reaching leaf node. The best decision criterion can be assessed using entropy, where H is a binary variable:

(18)

$$H(p_1) = -p_1 \log_2(p_1) - (1 - p_1) \log_2(1 - p_1),$$

while minimizing H, you are maximizing information gain at the current node.
Gini Index is another measure for decision criterion:

$$Gini = 1 - \sum_j p_j^2 = 2p_1(1 - p_1), \ where\ p_2 = 1 - p_1 \quad (19)$$

Again, where the Gini index is lowest, it gives us the point where we should split the decision tree. The Decision Tree algorithm iteratively uses those measures to reach the leaf node (either pure class or with predetermined method).

5. RESULTS

This chapter summarizes the performance of the classification techniques introduced in section 4. For each drug and drug pleiade in the dataset, we perform a binary classification task with the ultimate goal of classifying as accurately as possible individuals as drug users or non-drug users. In order to assess the performance of our classifiers and determine which classification algorithm is best suited to classify a specific drug or drug pleiade, accuracy was used as preferred performance metrics. Accuracy is defined as the ratio between the sum of true positive (TP) and true negative (TN), and the sum of TP, TN, false positive (FP) and false negative (FN):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (20)$$

Moreover, in order to quantify the diagnostic ability of our classifiers, two additional statistical measures were implemented: sensitivity and specificity. The latter are widely used in medicine and aim at measuring the proportion of actual positives that are correctly identified as such, and the proportion of actual negatives that are correctly identified as such, respectively [1].

$$Sensitivity = \frac{TP}{TP + FN} \quad (21)$$

$$Specificity = \frac{TN}{TN + FP} \quad (22)$$

All classification algorithms were implement using Scikit-learn, a free software machine learning library for the Python programming language. Before presenting the main results, subsection 5.1 to 5.7 provide a detailed insight into the algorithms and the adjustments made to them.

5.1 Linear classifier with SGD training

The implementation of the linear classifier with SGD training was performed twice both including and excluding regularization. Such a dual implementation was made in order to detect a possible risk of overfitting, which would then affect accuracy in the test set. To both implementations was applied a squared loss function, which allows for a more stable and closed form solution (by setting its derivative to 0). Using a squared loss function entails the risk of huge deviations in some of the samples. This may result in reduced accuracy in the presence of outliers in the dataset. However, such a circumstance seems to be excluded because the dataset was made available online after the removal of invalid responses. As for the learning rate, the default implementation "optimal" was used, which corresponds to eta = 1.0 / (alpha * (t + t0)) where t0 is chosen by a heuristic proposed by Leon Bottou [2], and alpha has a default value of 0.0001. For the regularised model, a L2 regularization method was used, which adds squared magnitude of coefficient as penalty term to the loss function and, unlike L1 regularization, avoids the risk of shrinking less important feature's coefficient to zero. Lambda, the constant that multiplies the regularization term, in the Scikit-learn library is coded as alpha, and it is set to 0.001. Alpha is also used to compute the learning rate, when this is set to "optimal". Both the regularized and the non-regularised models were set to 10000 iterations, and a stopping criterion with a default value of 0.001, which measures how small the residual of the ultimate solution is supposed to be.

5.2 Logistic Regression

Analogously to the linear classifier with SGD training, the logistic regression classifier was implemented both in a regularised and non-regularized version. Parameters and adjustments mimic those already implemented in the linear classifier with the exception of the loss function, which uses a logarithmic loss in order to measure the performance of a classification model where the prediction input is a probability value between 0 and 1.

5.3 LDA, QDA, Gaussian Naïve Bayes

The implementation of LDA, QDA and Gaussian Naïve Bayes follows the basic defaults settings provided by the Scikit-learn library.

5.4 Support Vector Machine

The implementation for the Support Vector Machine specifies the kernel type to be used in the algorithm to be the default Radial Basis Function (RBF) with a gamma value of 1/no. of features. Avoiding the use of kernel function “linear” allows SVM to handle non-linearly separable data. Tuning parameter C, which weights in-sample classification errors and thus controls the generalisation ability of the SVM classifier, was set to a value of 1. By choosing 1, the expectation is that the SVM classifier will choose a large margin decision boundary, which may lead to a larger number of misclassifications on the one hand, but it also ensures a higher generalization power by reducing the risk of overfitting on the training sample.

5.5 K-Nearest Neighbours

The implementation of K-Nearest Neighbours follows closely the defaults parameters offered by the Scikit-learn library, assigning uniform weights to all points in each neighbourhood, and choosing Minkowski with p=2 – equivalent to Euclidean metric – as distance metrics. The choice of the k parameter is made possible by performing grid search for each drug in a range of k from 1 to 10. Subsequently, we looked at accuracy values obtained for each drug and for each values of k, and selected the best accuracy value for each drug to be used for classifier performance comparison.

5.6 Neural Network

The implementation of the Neural Network relies on one hidden layer containing 100 neurons, whose weights optimization is done by the ‘adam’ solver, a stochastic gradient-based optimizer. The model is trained on a batch size of 25 observations and implements an adaptive learning rate, which keeps the initial learning rate constant to 0.001 as long as training loss keeps decreasing. Each time two consecutive epochs fail to decrease training loss by at least 0.0001 – the value of the stopping criterion, or fail to increase validation score by at least the value of the stopping criteria, provided that early stopping is on, the current learning rate is divided by 5. The Neural Network aims at reducing overfitting on the training set by including an L2 regularization parameter equal to 0.0001. As for the activation function, ReLU was the preferred choice. The proportion of training data to set aside as validation set for early stopping is 10%. The model is set to a maximum of 200 epochs.

5.7 Decision Tree

The implementation of the decision tree supports the "gini" criterion to determine the Gini impurity, as the latter proved to yield better performances than the Information Gain. The strategy used to choose the split at each node is set to "best" in order to choose the best split. As for the maximum depth of the tree, this is set to the default "None", which implies that nodes are expanded until all leaves are pure or until all leaves contain less than the minimum number of samples – set to 2 – required to split an internal node. Finally, in order to yield better performances, the maximum number of features to consider when looking at the best split is set to "log2", and the minimum impurity decrease, which determines that a node will split if this split induces a decrease of the impurity greater than or equal to a specific value, is set to 0.005.

5.8 General results

The best accuracy scores corresponding to the best classifier/s for each drug in the test set are summarized in Table 8. Overall, the general level of accuracy seems satisfactory, given that all scores are largely above 50%, and the lowest best accuracy score reached 69%. It is not very surprising that once legal drugs (chocolate, caffeine, nicotine, alcohol) and cannabis are used as covariates, it is less likely to observe cases of very high accuracy (above 90%). Previous to the use of soft drugs as covariates, high frequency of above-90%-accuracy-scores was largely due to the fact that most individuals do in fact consume alcohol, caffeine and chocolate or have done so in the last ten years. This translated into a very homogeneous labelling, where the vast majority of individuals were indeed drug users. In that sense, for some legal drugs, it was possible to assume almost a one-class classification task. Similarly, but for opposite reasons, among all illegal drugs, VSA, crack and heroin were and are still classified with very high accuracy because they are the least consumed illegal drugs. Hence, this implies once again a class-unbalanced classification task in favour of non-drug users. Semeron, the fictious drug used to detect over-claimers, is classified again with extremely high accuracy, and this is not surprising because the number of over-claimers in the entire dataset is 9, and in our randomly split test set there is no over-claimer.

It is interesting to notice that introducing all legal drugs and cannabis as covariates resulted in an improvement of the general level of accuracy. Including consumption information about these soft drugs helped increase the accuracy levels (2-3% increase) reached by classifiers, and predict more precisely the risk that an individual may consume other (illegal) drugs. With the

exception of amphetamine, legal highs and methadone, which are classified less accurately (fall of 2-3% approx.), the remaining drugs are classified with accuracy levels that range between 72% and 79%, if we disregard VSA, semeron, crack and heroin. It is especially noteworthy to mention how the introduction of soft drug covariates increased sensibly the accuracy scores for cocaine (from 65% to 73%) and ecstasy (from 73% to 79%), two of the most abused and harmful hard drugs.

With respect to the classifiers, often more than one classifier reached the same high level of accuracy. In relative terms, KNN was able to classify with high accuracy 7 out of 14 drugs, followed by Logistic Regression regularized with 6 out of 14, and Logistic regression with 4 out of 14 drugs. In absolute terms, KNN and Logistic regression regularised performed better than any other classifier, both of which classified 4 out of 14 drugs with the highest accuracy. It is not surprising that KNN turned out to be the best classifier because of the k-neighbour selection process that optimised KNN performance before comparing it with other classifiers. Moreover, it is also not surprising that Logistic Regression (with and without regularization) performed well, given that the Sigmoid Function used in Logistic Regression returns outputs ranging from 0 to 1, which are often interpreted as probabilities. Therefore, Logistic Regression in its basic form is especially suited to classify or predict binary classes. Surprisingly, despite data approximation of normality, LDA and QDA performed poorly, classifying successfully only 1 and 3 drugs out of 14 drugs respectively, relative to other equally accurate classifiers. Disappointing performances were also those of SVM, and the-lately-introduced Decision Tree technique, both of which classified accurately only 3 drugs out 14, relative to other equally accurate classifiers. Finally, also the Neural Network failed to meet our expectations of the best classifier both in absolute and relative terms. With the hope of improving the performance of our Neural Network, we implemented the Principal Component Analysis (PCA), a dimensionality reduction technique that re-expresses the available dataset by extracting the most relevant information. Moreover, PCA is used to reduce redundancy, and minimizes noise. Our expectation was, thus, that PCA might increase the chances of the Neural Network to outperform other machine learning techniques. However, not only did implementing PCA yield lower accuracy scores for all classifiers, but it also did not improve the predictive ability of the Neural Network. Therefore, we resolved to dismiss the use of dimensionally reduced data. Two possible reasons that may explain such a poor result for the Neural Network might be related to the general architectural complexity of the network, and the large number of parameters to optimize. The latter might have turned out to be counterproductive for the task at hand, and might have affected the model performance in favour of other classifiers.

DRUG	BEST ACCURACY	BEST CLASSIFIER/S
Amphet	0.698	Logistic Regr reg
Amyl	0.772	KNN
Benzos	0.720	KNN
Coke	0.735	Logistic Regr reg
Crack	0.905	KNN
Ecstasy	0.794	QDA
Heroin	0.899	Logistic Regr
Ketamine	0.794	LinearSGD, Logistic Regr, Logistic Regr reg, LDA, SVM, NN, Decision Tree
Legalh	0.783	QDA, KNN
LSD	0.772	Logistic Regr reg
Meth	0.778	Logistic Regr, SVM, KNN, Decision Tree
Mushroom	0.772	KNN
Semer	1	LinearSGD, Logistic Regr, Logistic Regr reg, QDA, SVM, KNN, NN, Decision Tree
VSA	0.910	Logistic Regr reg

Table 8: Best accuracy score and best classifier/s per drug.

After determining which classifier/s achieved the highest level of accuracy for each drug, sensitivity and specificity scores were computed in order to quantify the diagnostic ability of the classifier/s. Indeed, if a treatment is to be administered, it is necessary to be certain that the patient is presenting a specific medical condition that actually requires that treatment in order to avoid undesired side effects. Analogously, patients who are drug addicted and need treatment should be promptly classified as such. Table 9 shows sensitivity and specificity scores for each drug. Already at first glance, we can easily notice that the results show overall high specificity scores. More specifically, for crack, ketamine, meth, semeron and VSA the specificity score is 100%, while for amyl nitrite, heroin and LSD the specificity score is above 90%. High specificity scores indicate that individuals who do not consume such drugs are correctly identified as such, and thus treatment should not be administered. As one can notice specificity and sensitivity scores are inversely proportional, meaning that as sensitivity increases, specificity decreases and vice versa. In that sense, it is not surprising that to low sensitivity scores obtained for crack, heroine, ketamine and VSA correspond high specificity scores. Moreover, it is often the case that for those drugs with high specificity but low sensitivity scores, the actual number of drug-addicted individuals is very low and not correctly classified, while non-addicted individuals are correctly classified and represent the vast majority of cases (see Table 12 in Appendix). High sensitivity scores are found for benzos, cocaine, ecstasy,

legal highs and mushroom, which suggest that actual drug-addicted individuals are correctly identified as such. These scores are very encouraging, in that they reached very high percentages, ranging from over 60% to 94%, and thus treatment may be promptly administered to individuals in need. Overall, these results seem to be encouraging, in that they would prevent physicians from administering a treatment to patients that do not suffer from a given condition. Furthermore, for some of the most harmful drugs, using classifiers as a diagnostic tool turned out to be successful, as these enable physicians to forecast more accurately which patients are exposed to a higher risk of drug addiction, and administer treatment promptly. On the other hand, for some drugs such as amphetamine, amyl nitrite and LSD, the implemented algorithms seem to fail to correctly identify drug-addicted individuals as having such a condition. In that sense, physicians would need to conduct supplementary tests to determine the actual total amount of individuals who suffer from drug-addition, and who may otherwise miss the chance of being treated.

	SENSITIVITY	SPECIFICITY
Amphet	0.525	0.781
Amyl	0.348	0.909
Benzos	0.611	0.786
Coke	0.859	0.649
Crack	0.053	1
Ecstasy	0.940	0.670
Heroin	0.1	0.994
Ketamine	0	1
Legalh	0.886	0.709
LSD	0.306	0.936
Meth	0	1
Mushroom	0.708	0.812
Semer	inf	1
VSA	0.056	1

Table 9: Sensitivity and Specificity scores per drug.

In section 3.2, correlations matrices summarizing the correlational relationships among drugs were showed and discussed. After identifying those drugs that show strong positive correlation, drug pleiades were built and classification algorithms were used to label individuals as drug users or non-drug users. Table 10 shows the best accuracy scores for each pleiade and its best classifier/s. Overall, results suggest very satisfactory accuracy levels, with scores ranging between 78% and 89%. Indeed, using pleiades to conduct a classification task turned out to be more effective in terms of accuracy performances. However, this was obtained at the expenses

of not being able to identify straightforwardly which single drug an individual is more likely to consume. As far as classifiers are concerned, once again Logistic Regression with and without regularization seemed to ensure better performances relative to other classifiers. Surprisingly, in the benzos pleiade classification task, SVM performed in absolute terms better than any other classifier, although in the single drug classification task SVM classified correctly only 3 out 14 drugs.

DRUG PLEIADE	BEST ACCURACY	BEST CLASSIFIER/S
HeroinPl	0.778	Logistic Regr, QDA
EcstasyPl	0.889	LinearSGD, Logistic Regr reg, KNN
BenzoPl	0.810	SVM

Table 10: Best accuracy score and best classifier/s per drug.

What is really interesting, however, are the findings about sensitivity and specificity scores. The employed classification algorithms turned out to be very satisfactory diagnostic tools. Indeed, extremely high sensitivity scores are reported for all pleiades, ranging between 92% to over 93%. The latter indicate that actual drug-addicted individuals were correctly identified as such. As for specificity scores, these show much more variation, ranging from 62% to 82%. Overall, findings are very encouraging because they suggest that drug-addicted individuals have very good chances of receiving treatment. Moreover, including classification for drug pleiades may increase the effectiveness of prevention and treatment administration, as there is a strong positive correlation among drugs and, thus drug users, within each pleiade. On the other hand, however, specificity scores are less encouraging, with the exception of the ecstasy pleiade. Low specificity scores suggest that non-drug-addicted individuals are not correctly identified as such, exposing them to the risk of receiving treatment (and its side effects), when in fact they wound not need it. It is, however, preferable to administer treatment to healthy individuals, then to omit to treat drug-addicted patients.

DRUG PLEIADE	SENSITIVITY	SPECIFICITY
Heroin Pleiade	0.927	0.624
Ecstasy Pleiade	0.930	0.824
Benzos Pleiade	0.931	0.667

Table 11: Sensitivity and Specificity scores per drug pleiade.

6. CONCLUSION

Adopting machine learning techniques to classify drug addicted from non-drug addicted individuals has yield in most cases fairly high levels of accuracy, indicating that such techniques represent a valuable resource in the detection of individuals who may require social and health assistance. Physicians and policy makers may also take advantage of classification algorithms to improve the timeliness of treatment administration or to better target individuals in drug prevention campaigns.

Further investigations may attempt to reach even higher accuracy by including additional personality trait information, increase the sample size or investing further efforts in carrying out additional fine-tuning and optimization adjustment to the parameters of many of the machine learning algorithms employed in this project.

7. REFERENCES

[1] Altman, D.G., Bland, J.M. Diagnostic tests. 1: Sensitivity and specificity. *The British Medical Journal*, 308 (6943): 1552, 1994.

[2] Bottou L. Stochastic Gradient Descent Tricks. In: Montavon, G., Orr, G.B., Müller, K.R. (eds) *Neural Networks: Tricks of the Trade. Lecture Notes in Computer Science*, vol. 7700. Springer, Berlin, Heidelberg, 2012.

[3] Costa, P.T., MacCrae, R.R. *Revised NEO-Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO FFI): Professional manual*, Psychological Assessment Resources, Odessa (FL), 1992.

[4] Fehrman, E., Muhammad, A.K., Mirkes, E.M., Egan, V., Gorban, A.N. The Five Factor Model of Personality and Evaluation of Drug Consumption Risk. In: Palumbo F., Montanari A., Vichi M. (eds) *Data Science. Studies in Classification, Data Analysis, and Knowledge Organization*, Springer, Cham, 2017.

[5] Hastie, T., Tibshirani, R., Friedman, J. *The Elements of Statistical Learning*. New York, NY, USA: Springer New York Inc, 2001.

[6] Janik, P., Kosticova, M., Pecenak Prof, J., Turcek, M. Categorization of psychoactive substances into “hard drugs” and “soft drugs”: a critical review of terminology used in current scientific literature. *The American Journal of Drug and Alcohol Abuse*, 43:6, 636-646, 2017.

[7] Ng, A. Stanford Lecture Notes. Available at: http://cs229.stanford.edu/notes/cs229-notes1.pdf [Last accessed: 15th August 2019].

[8] Snowden, R.J., Gray, N.S. Impulsivity and psychopathy: Associations between the Barrett Impulsivity Scale and the Psychopathy Checklist revised. *Psychiatry Research*, 187(3): 414–417, 2011.

[9] Stanford, M.S., Mathias, C.W., Dougherty, D.M., Lake, S.L., Anderson, N.E., Patton, J.H. Fifty years of the Barratt Impulsiveness Scale: An update and review. *Personality and Individual Differences*, 47(5): 385–395, 2009.

Dataset:

Fehrman, E., Egan, V. Drug consumption, collected online March 2011 to March 2012, English-speaking countries. Drug consumption (quantified) Data Set, donated to UCI Machine Learning Repository on 17th October 2016. Available at: http://archive.ics.uci.edu/ml/datasets/Drug+consumption+%28quantified%29 [Last accessed: 15th August 2019].

APPENDIX

A. CONFUSION MATRICES

AMPHET	Predicted Negative	Predicted Positive
Actual Negative	100	28
Actual Positive	29	32
AMYL	Predicted Negative	Predicted Positive
Actual Negative	130	13
Actual Positive	30	16
BENZOS	Predicted Negative	Predicted Positive
Actual Negative	92	25
Actual Positive	28	44
COKE	Predicted Negative	Predicted Positive
Actual Negative	72	39
Actual Positive	11	67
CRACK	Predicted Negative	Predicted Positive
Actual Negative	170	0
Actual Positive	18	1
ECSTASY	Predicted Negative	Predicted Positive
Actual Negative	71	34
Actual Positive	5	79
HEROIN	Predicted Negative	Predicted Positive
Actual Negative	168	1
Actual Positive	18	2
KETAMINE	Predicted Negative	Predicted Positive
Actual Negative	150	0
Actual Positive	39	0
LEGALH	Predicted Negative	Predicted Positive
Actual Negative	78	32
Actual Positive	9	70
LSD	Predicted Negative	Predicted Positive
Actual Negative	131	9
Actual Positive	34	15
METH	Predicted Negative	Predicted Positive
Actual Negative	147	0
Actual Positive	42	0

MUSHROOM	Predicted Negative	Predicted Positive
Actual Negative	95	22
Actual Positive	21	51
SEMER	Predicted Negative	Predicted Positive
Actual Negative	189	0
Actual Positive	0	0
VSA	Predicted Negative	Predicted Positive
Actual Negative	171	0
Actual Positive	17	1

Table 12: Confusion matrix per drug.

HEROIN PLEIADE	Predicted Negative	Predicted Positive
Actual Negative	58	35
Actual Positive	7	89
ECSTASY PLEIADE	Predicted Negative	Predicted Positive
Actual Negative	61	13
Actual Positive	8	107
BENZOS PLEIADE	Predicted Negative	Predicted Positive
Actual Negative	58	29
Actual Positive	7	95

Table 13: Confusion matrix per drug pleiade.

B. ACCURACY MATRICES

	Amphet	Amyl	Benzos	Coke	Crack	Ecstasy	Heroin	Ketamine	Legalh	LSD	Meth	Mushroom	Semer	VSA
LinearSGD	0.677	0.746	0.661	0.587	0.899	0.757	0.894	0.794	0.772	0.534	0.476	0.619	1	0.905
LinearSGD_reg	0.635	0.709	0.646	0.471	0.799	0.608	0.831	0.688	0.730	0.614	0.725	0.698	0.952	0.767
Logistic Regr	0.672	0.751	0.688	0.704	0.894	0.624	0.899	0.794	0.751	0.614	0.778	0.730	1	0.905
Logistic Regr reg	0.698	0.757	0.608	0.735	0.899	0.783	0.889	0.794	0.772	0.772	0.772	0.714	1	0.910
LDA	0.667	0.751	0.661	0.730	0.894	0.751	0.894	0.794	0.757	0.741	0.772	0.725	0.979	0.899
QDA	0.519	0.651	0.677	0.550	0.730	0.794	0.693	0.646	0.783	0.646	0.704	0.720	1	0.894
NB	0.651	0.524	0.667	0.683	0.471	0.788	0.677	0.635	0.757	0.635	0.646	0.735	0.450	0.497
SVM	0.677	0.757	0.661	0.730	0.899	0.746	0.894	0.794	0.762	0.746	0.778	0.730	1	0.905
KNN	0.683	0.772	0.720	0.730	0.905	0.767	0.894	0.788	0.783	0.767	0.778	0.772	1	0.905
NN	0.661	0.757	0.672	0.624	0.899	0.704	0.894	0.794	0.772	0.714	0.751	0.714	1	0.905
Decision Tree	0.667	0.757	0.619	0.603	0.899	0.677	0.894	0.794	0.735	0.741	0.778	0.619	1	0.905

Table 14: Accuracy matrix for all drugs and classifiers.

	HeroinPl	EcstasyPl	BenzosPl
LinearSGD	0.725	0.889	0.661
LinearSGD_reg	0.709	0.730	0.614
Logistic Regr	0.778	0.725	0.799
Logistic Regr reg	0.772	0.889	0.794
LDA	0.746	0.884	0.799
QDA	0.778	0.868	0.799
NB	0.735	0.878	0.788
SVM	0.746	0.878	0.810
KNN	0.772	0.889	0.794
NN	0.698	0.884	0.799
Decision Tree	0.608	0.820	0.757

Table 15: Accuracy matrix for all drug pleiades and classifiers.

C. CORRELATION MATRICES FOR ALL DRUGS

	Amphet	Amyl	Benzos	Coke	Crack	Ecstasy	Heroin	Ketamine	Legalh	LSD	Meth	Mushroom	Semer	VSA
Amphet	1,000	0,372	0,463	0,580	0,323	0,597	0,359	0,412	0,481	0,490	0,415	0,481	0,016	0,304
Amyl	0,372	1,000	0,226	0,381	0,144	0,392	0,137	0,345	0,268	0,213	0,084	0,271	0,019	0,130
Benzos	0,463	0,226	1,000	0,428	0,326	0,383	0,395	0,303	0,348	0,352	0,468	0,367	0,049	0,294
Coke	0,580	0,381	0,428	1,000	0,396	0,633	0,414	0,449	0,445	0,442	0,351	0,480	0,055	0,277
Crack	0,323	0,144	0,326	0,396	1,000	0,280	0,509	0,256	0,203	0,268	0,367	0,276	0,043	0,278
Ecstasy	0,597	0,392	0,383	0,633	0,280	1,000	0,300	0,512	0,586	0,599	0,316	0,599	0,031	0,289
Heroin	0,359	0,137	0,395	0,414	0,509	0,300	1,000	0,275	0,237	0,347	0,494	0,306	0,040	0,293
Ketamine	0,412	0,345	0,303	0,449	0,256	0,512	0,275	1,000	0,393	0,462	0,235	0,436	0,046	0,193
Legalh	0,481	0,268	0,348	0,445	0,203	0,586	0,237	0,393	1,000	0,518	0,334	0,575	0,030	0,314
LSD	0,490	0,213	0,352	0,442	0,268	0,599	0,347	0,462	0,518	1,000	0,344	0,680	0,067	0,299
Meth	0,415	0,084	0,468	0,351	0,367	0,316	0,494	0,235	0,334	0,344	1,000	0,343	0,015	0,278
Mushroom	0,481	0,271	0,367	0,480	0,276	0,599	0,306	0,436	0,575	0,680	0,343	1,000	0,074	0,253
Semer	0,016	0,019	0,049	0,055	0,043	0,031	0,040	0,046	0,030	0,067	0,015	0,074	1,000	0,036
VSA	0,304	0,130	0,294	0,277	0,278	0,289	0,293	0,193	0,314	0,299	0,278	0,253	0,036	1,000

Table 16: Correlation matrix for all drugs.

DECLARATION OF AUTHORSHIP

We hereby declare that the paper submitted is our own unaided work. All direct or indirect sources used are acknowledged as references.

This paper was not previously presented to another examination board and has not been published. We both spend comparable time and effort to this project concerning research, coding, creating the report and the presentation.

Konstanz, 18.08.2019

_____

Place and Date

_____ _____

Roberto Daniele Cadili Burak Özturan