Drug Consumption by Personality Data

Štěpán Pešout
Universidade de Évora

January 19, 2022

Abstract

This paper deals with the construction of several classification models. First, it is a model that attempts to predict using of at least one drug of a predefined list. This model uses classifiers of several types (Rules, J48, SMO and NaiveBayes). Next, the use of 3 different substances is classified using the J48 algorithm, which generates a decision tree for the purpose of further analysis, which factors lead to the use of these drugs. After the necessary modifications in the dataset, the algorithms are applied and the results are summarized. Finally, possibilities for future work are presented.

1 Introduction

It is important to be able to tell individuals that they are at increased risk of using certain psychoactive substances, as this can increase the chances of minimising the impact of such use in the future. If the specific risk factors are known, it is possible to work with these people in a timely manner to effectively reduce potential harms. These can be both physical and psychological and in many cases even irreversible.

When the psychology of individual people needs to be quantified, psychologists now often use the Big Five Personality Model (sometimes also referred to as the CANOE or OCEAN model) [1]. Specifically, it is concerned with the following aspects of personality:

- **C**onscientiousness (impulsive, disorganized vs. disciplined, careful);
- **A**greeableness (suspicious, uncooperative vs. trusting, helpful);

- **N**euroticism (calm, confident vs. anxious, pessimistic);
- **O**penness to Experience (prefers routine, practical vs. imaginative, spontaneous);
- **E**xtraversion (reserved, thoughtful vs. sociable, fun-loving) [1].

From a medical perspective, the issue of people of different personality types taking psychoactive substances has already been addressed in an article on BMC Psychiatry written in 2008. This article shows how the values of the Five-Factor Model personality profiles vary for different drugs [2].

Later, this problem was also investigated using data mining and machine learning techniques. In 2015 an online survey methodology was employed to collect data including Big Five personality traits, impulsivity, sensation seeking alongside with demographic information. The optimised models sensitivities and specificities were around 70%, with some specific drugs even around 75% [3].

Even in recent years, machine learning models or neural networks have been developed to predict the use of psychoactive substances. Their accuracies has usually reached a maximum of 77%.

The goal of this project will be to build classification models that can best predict drug use based on several demographics and in particular personality profile. Furthermore, it will be briefly described how these attributes influence the use of few selected psychoactive substances.

2 Data

This paper will use the dataset that was compiled based on the survey from 2015, mentioned in the introduction. Given that several major studies have been written based on this dataset, it can be argued that it is relevant. Furthermore, it contains 1885 instances, which is a relatively sufficient number to build a plausible classification model.

Data are recorded on each research participant across 31 attributes (excluding ID). Of these, there are 5 demographics: level of education, age, gender, country of residence and ethnicity. The other 7 describe the personality of

the participant; neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness are attributes of the Big Five Personality Model. Impulsivity and sensation seeking are also recorded to better describe a person's personality. The remaining 19 attributes are data on the use of various legal or illegal drugs, however, one of them (Semeron) is a fictitious one in order to recognize over-claimers [3]. Value of the drug attributes are distributed in 7 classes, which are as follows:

- 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).

Among the most important papers that have been produced using this dataset, it is necessary to include in particular the aforementioned paper written directly by the authors of this dataset [3], as well as the book Personality Traits and Drug Consumption, which they wrote [4]. These data were also used in a study, which concerns predicting, who could be a potential drug consumer using machine learning techniques [5]. Also, the Extreme Learning Machine based on personality features has been trained for drug usage duration classification using this dataset [6].

3 Proposed solutions

3.1 Main objectives

This work will also deal with classification. The main goal will be to build a prediction model that can be used to determine with some accuracy whether a person is generally likely to use psychoactive substances or not. This will be considered as a binary classification problem.

A secondary aim will be to analyse what specific factors lead to the use of different substances. This will not be done for all substances, but only for

selected representatives of different categories of drugs. In this case, three separate classes will be distinguished.

Weka will serve as the main software to accomplish these data mining tasks. Given the number of available ready-made algorithms and their possibilities of additional modification, this is definitely a suitable choice. MySQL will be used to pre-edit the data before importing into Weka. Thanks to SQL it is possible to modify the data more efficiently according to the needs. Moreover, using SQL is also an fast and easy way to calculate some statistics that can determine further work.

3.2 Performance measurement

As for the binary classification problem, although the main goal is to classify as accurately as possible, it is also necessary to minimize the number of false negatives. This is because if someone, for instance, thinks that they are at increased risk of consuming these substances and this is not true, it is not such a problem, unlike the opposite situation. Since this is a relatively balanced dataset, it is possible to use accuracy as the primary performance evaluation method. However, precision and recall will be calculated as well in order to describe the status of false negatives and false positives.

For the three classes, performance evaluation will be a bit more complex. The performance of the model will also be primarily quantified using accuracy (1). Since the classes in this case are of ordinal type, it is also possible to measure an extended accuracy (2). It is defined as an accuracy that considers every instance classified correctly or with a difference equal to one as a positive (all the values in the main diagonal and all in the two adjacent diagonals to the main diagonal of the confusion matrix – Table 1). This method of measuring performance is not widely used, but a similar concept has been introduced in 2011 by Jaime Cardoso and Ricardo Sousa [7].

The model will be tested using 10-fold cross-validation, which will allow several different models to be built and then tested. As a result, the performance metrics can be expected to be slightly more accurate than using separate training and testing datasets.

		Predicted class		
		class = a	class = b	class = c
Actual class	class = a	a_a	a_b	a_c
	class = b	b_a	b_b	b_c
	class = c	c_a	c_b	c_c

Table 1: 3-class confusion matrix

$$accuracy\ (a) = \frac{a_a + b_b + c_c}{a_a + a_b + a_c + b_a + b_b + b_c + c_a + c_b + c_c} \tag{1}$$

$$extended\ accuracy\ (a_{ext}) = \frac{a_a + b_b + c_c + a_b + b_c + b_a + c_b}{a_a + a_b + a_c + b_a + b_b + b_c + c_a + c_b + c_c} \tag{2}$$

3.3 Algorithms

After making modifications to the dataset, such as splitting class attributes into only two or three classes and cleaning the dataset, it will be necessary for the main model to discard some unimportant attributes so that the classifiers perform better. This will be achieved using wrapper subset evaluation with J48 algorithm and cross-validation. A tree will be built for each fold and the frequency of representation of each attribute will be evaluated. This will allow to make an estimation, which attributes are more likely to be less important for decision making and should therefore be removed from the dataset.

For the model that is the main objective of this work (the binary classification models for multiple drugs), the ZeroRule and OneRule algorithms will be used to obtain a basic overview of the data. Furthermore, the J48 classifier will be used to generate a tree. This will be followed by a model built using Naive Bayes. Finally, Sequential minimal optimization (SMO) will be applied to the dataset.

For classification models, that predicts individual drug use, only the J48 algorithm will be used. Although the aim here is also to maximize accuracy,

an even more important task is to understand, what demographic or psychological factors lead to the use of particular substances. For this purpose, this algorithm is very suitable, as its output is a decision tree, which can be easily interpreted by humans. For the same reason, the minimum number of instances per leaf will be set to a high value in order to make the tree as simple as possible. Also, no attribute selection will be performed to see how individual attributes affect the decision tree.

3.4 Performance goals

The exactly same approaches (firstly binary classification models for multiple drugs and secondly single drug models with three classes) used in this paper have not been used anywhere in terms of available resources. It is known, that binary classification models for individual psychoactive substances in previous works have shown accuracies of around 70 to 75% for individual drugs. However, the work published by the original authors of the dataset claimed that when models are built to predict the use of multiple drugs, the accuracy is slightly higher [3].

Ideally, therefore, the binary classification models described would have an accuracy better than 75%. For single substance models, an accuracy of less than 70% is expected, as this is a multi-class prediction. Extended accuracy values (only for the multi-class models) are expected to be about 90%, because this method of performance evaluation identifies certain types of erroneous estimates as correct ones.

4 Dataset preparation

In order to achieve good and not-biased model performance, it is advisable to make certain adjustments to the dataset before a classifier is applied. It was uploaded to a MySQL database because SQL allows efficient searching and editing of the data.

4.1 Personality and demographic attributes

First, those research participants who reported taking the drug Semeron were identified. Since this drug does not exist, it is clear that those who claim to have taken it are over-claimers, who are not telling the truth. Thus, those instances that have some other value than never used for the Semeron attribute were removed. After that, the whole column was dropped.

The next step was to replace the intervals in the age attribute with the mean value. This turned the nominal attribute into a ratio attribute, which is more suitable in terms of subsequent processing by algorithms. A similar transformation was made even for the level of education. The lowest level (left school before 16 years) has a value of 1, while the highest, a doctorate degree, has a value of 9.

4.2 Class for the binary classification models for multiple drugs

The main classification problem, as already mentioned, will deal with the identification of users who regularly use a psychoactive substance. It was decided to include only those substances that less than two-thirds of users have ever tried, as the others can be described as widely used and are not dealt with in this paper.

Thus, in order to determine the class for this problem, it was necessary to calculate the percentages of users of each drug as the first step, as shown on Table 2. This was calculated as the ratio of the count of values that are neither equal to CL0 nor CL1 to all in a given column.

The resulting class is therefore a synthesis of the following 13 attributes: Amphetamine, Nitrite, Benzodiazepine, Cocaine, Crack, MDMA (Ecstasy), Heroin, Ketamine, Legal highs, LSD, Methadone, Mushrooms and Volatile substance. Thus, 5 attributes of the original 18 will not be part of this new class: Alcohol, Caffeine, Cannabis, Chocolate and Nicotine.

As already mentioned, this class is binary and can therefore only take the values "user" and "non-user" for individual instances. In order to exclude

Substance name	**Percentage of users**
Alcohol	96.4%
Amphetamines	35.9%
Nitrite	19.6%
Benzodiazepine	40.6%
Caffeine	98%
Cannabis	67%
Chocolate	98.2%
Cocaine	36.3%
Crack	10%
MDMA (Ecstasy)	39.7%
Heroin	11.1%
Ketamine	18.4%
Legal highs	40.3%
LSD	29.4%
Methadone	22%
Mushrooms	36.6%
Nicotine	66.9%
Volatile substance	12%

Table 2: Percentages of users of each drug

Original class	**New class**
CL0 (Never Used)	Never
CL1 (Used over a Decade Ago)	Never
CL2 (Used in Last Decade)	Rarely
CL3 (Used in Last Year)	Rarely
CL4 (Used in Last Month)	Regularly
CL5 (Used in Last Week)	Regularly
CL6 (Used in Last Day)	Regularly

Table 3: New classes for the single drug classification models

people who try new drugs but do not use them regularly, a participant is only marked as a user if they have a CL4, CL5 or CL6 value for any drug that is part of the shortlist. In the resulting dataset, there are 764 users and 1112 non-users.

Thus, only attributes representing a person's demographics and personality will be available when building a classification model. The reason for this is to avoid biasing the model due to other drugs. While the accuracy of the classification would be high, the result would not be of adequate value as it can be assumed that people who already use some drugs will also tend to use others.

4.3 Classes for the single drug classification models

In this case, the seven classes for each drug were simply divided into three in the way shown in the Table 3. Again, if a model is built for one substance, attributes reflecting the use of others will be eliminated. The reason for this is the same as for the previously described model.

5 Results

5.1 Binary classification models for multiple drugs

5.1.1 Attribute selection

At first, it was necessary to discard some unimportant attributes to avoid classifiers to process them. As already mentioned, this was done in order to improve the performance of the resulting models. For this purpose, the J48 algorithm was used to construct 3 types of decision trees with different minimum numbers of instances per leaf (10, 30 and 50) – as shown in the Table 4 – in a combination with 50-fold cross-validation.

After that, the percentages of representation of each attribute in the trees were averaged and all attributes with representation less than 15% were removed. Specifically, these were 4: ethnicity, extraversion score (escore), agreeableness score (ascore) and conscientiousness score (cscore), because they turned out not to be very important for prediction-making. Thus, the attributes had been removed from the dataset before the classification algorithms was applied.

As for the computational load for running Wrapper Subset Evaluation, on a regular laptop this operation took 8 minutes and 20 seconds for 10 minimum instances per leaf, for 30 minimum instances it took 5 minutes and 51 seconds and for 50 minimum instances it was 4 minutes and 28 seconds. It suggests that this is a relatively computationally demanding operation, but, however, it only has to be done once. This is because when more data is added in the future, unlike classification models that would have to be rebuilt, the decision on the importance of the attributes remains unchanged.

5.1.2 Classification

As already mentioned, the ZeroRule and OneRule algorithms was used to obtain a basic overview of the data. The accuracy of ZeroRule was 0.59 (every predicted class was non-user), which in a binary classification model proves that it is a balanced dataset. OneRule had an accuracy of 0.69, with classification based on age. If age is excluded, education comes next, followed

Attribute	**inst. > 10**	**inst. > 30**	**inst. > 50**	**Average**
age	100%	100%	100%	100%
gender	100%	92%	82%	91.3%
education	96%	68%	56%	73.3%
ethnicity	44%	0%	0%	14.7%
nscore	12%	36%	10%	19.3%
escore	12%	20%	10%	14%
oscore	30%	24%	36%	30%
ascore	14%	6%	14%	11.3%
cscore	0%	6%	6%	4%
impulsive	10%	50%	4%	21.3%
ss	100%	92%	62%	84.7%

Table 4: Percentages of representation of each attribute in the J48 trees with different minimum numbers of instances per leaf

by sensation seeking. For all of these models, the accuracy is higher than 0.68, but the recall decreases with each excluded attribute, with values between 0.45 and 0.55. This means that these models are poor, because they have very high numbers of false negatives.

Furthermore, the J48 classifier with 10 minimum instances per leaf and confidence factor for pruning with the value of 0.2 was used to generate a decision tree. This model showed an accuracy of 0.76. Although this value alone would meet the goal of a value higher than 0.75 (which has been achieved within other similar models in the past), the result is not so satisfactory. The problem still lies in the precision and recall values of 0.71 and 0.67 respectively. This means, among other things, that the model successfully detects users of psychoactive substances in approximately two-thirds of cases only.

Another classifier used was NaiveBayes. The accuracy value was similar to J48 (0.77), as was the precision (0.7). However, recall has been slightly improved, reaching a value of 0.72. This means that there was some reduction in the number of false negatives. Thus, the prediction model built using

Classifier	Accuracy	Precision	Recall
ZeroRule	0.59	–	–
OneRule	0.69	0.65	0.54
J48	0.76	0.71	0.67
NaiveBayes	0.77	0.7	0.72
SMO	0.76	0.71	0.69

Table 5: Performance of all classifiers used

NaiveBayes can be considered to be a slightly better compared to J48.

The values of these performance metrics were not significantly different even when using Sequential Minimal Optimization (SMO). Accuracy here has reached the value of 0.76, while precision and recall were 0.71 and 0.69, respectively. This model can be therefore rated as slightly better than J48 and worse than NaiveBayes, but the differences are very small.

The performance of all classifiers used is recorded in Table 5 from the perspective of accuracy, precision and recall. In terms of the computing power required, these models take fractions of a second to build even with the 10-fold cross-validation. This is very advantageous because in some practical use they could be rebuilt frequently based on new recorded data.

5.2 Single drug classification models

As already mentioned, this is a classification problem with 3 classes. Only the J48 algorithm will be used, because the most important task here is to understand, what demographic or psychological factors lead to the use of particular substances and the generated tree can be easily interpreted by humans. For the same reason, the minimum number of instances per leaf will be set to 40 in order to make the tree as simple as possible.

The models will deal with each of three drugs: cannabis, cocaine and LSD. This decision was made because they are representatives of different categories of substances and their use is therefore likely to be determined by different combinations of factors.

5.2.1 Cannabis use classification

The resulting tree constructed using the J48 algorithm had an accuracy of 0.6 while the extended accuracy was 0.89. Such a low result was expected because there was an attempt to generate a simple tree that is human readable.

The main factors leading to cannabis consumption according to this model are age, gender and sensation seeking. In the youngest age group (under 21), the model even rates all males as regular users and all females as at least occasional users.

At older ages, the value for sensation seeking is particularly important, with people who have a higher value being more likely to use. Furthermore, low conscientiousness scores, high openness to experience scores and a combination of neuroticism and low agreeableness often lead to cannabis use.

5.2.2 Cocaine use classification

The model built using J48 was slightly more accurate for cocaine than for cannabis. Accuracy was 0.64 and extended accuracy was 0.98.

Sensation seeking plays a major role in cocaine use. It is also low agreeableness score, high neuroticism and higher openness to experience. Males and extroverted people are more frequent users.

5.2.3 LSD use classification

The model built to describe LSD use has an accuracy of 0.7 and an extended accuracy of 0.96. This shows that the group of users has a slightly more specific personality profile, which is easier to predict from the personality data.

The highest number of users are those under the age of 30 who are open to new experiences. High scores in conscientiousness and agreeableness are also important factors. People without a university degree and more impulsive people are more likely to use LSD.

Factor	Cannabis	Cocaine	LSD
Age	Lower	–	Lower
Gender	Males	Males	–
Education	–	–	Lower
Ethnicity	–	–	–
Conscientiousness	Lower	–	Higher
Agreeableness	Lower	Lower	Higher
Neuroticism	Higher	Higher	–
Openness to experience	Higher	Higher	Higher
Extraversion	–	Higher	–
Sensation seeking	Higher	Higher	–
Impulsivity	–	–	Higher

Table 6: Influence of demographic and personality factors to drug use

5.2.4 Summary

Users of psychoactive substances differ from each other (LSD users from others in particular). The only factor found common to users of all substances studied is openness to new experiences. For other personality and demographic factors, differences can be found for each drug. Table 6 records this information, obtained by extraction from J48 trees. It shows whether a given factor influences use and, if so, how.

6 Conclusions and future work

This work has shown that the most at-risk group of people are young adult males. In contrast, ethnicity plays a negligible role. These demographic factors are similar for all drugs.

From a psychological perspective, high scores on sensation seeking and openness to new experiences are particularly conducive to drug use, but other factors vary by substance. This limits the accuracy of models classifying the risk of using multiple drugs simultaneously. Another problem with the multi-substance models is the higher number of false negatives.

Although drug use is to some extent determined by previously mentioned demographic and personality factors, these are by no means the only conditions. It is therefore clear, that some other factors can play a significant role as well.

Given that psychologists more or less agree that the CANOE model of personality is a reasonably good description of reality [1], and given that sensation seeking and impulsivity were available in the dataset in addition to this model, it is likely that adding another demographic attributes would have helped predict more accurately rather than additional psychological ones.

This dataset could be analysed from many perspectives in the future. For example, it would be possible to rank drugs by their harmfulness and analyse, if the consumption of less harmful substances lead to the use of more problematic ones. Also, it would be possible to examine which drugs people of different personalities are more likely to try once and which they use regularly. Given that some substances are dangerous when combined with others at the same time, it would be useful in the future to find out which types of people tend to use these combinations.

7 References

[1] Lim, A. The Big Five personality traits. SimplyPsychology (2020). https://www.simplypsychology.org/big-five-personality.html

[2] Terracciano, A., Löckenhoff, C.E., Crum, R.M. et al. Five-Factor Model personality profiles of drug users. BMC Psychiatry (2008). https://doi.org/10.1186/1471-244X-8-22

[3] 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. arXiv (2017). https://arxiv.org/abs/1506.06297

[4] Fehrman E., Muhammad A. K., Mirkes E. M., Egan V., Gorban A. N., Levesley J. Personality traits and drug consumption: A story told by Data (2019). Springer.

[5] Qiao Z., Chai T., Zhang Q., Zhou X., Chu Z. Predicting potential drug abusers using machine learning techniques. International Conference on Intelligent Informatics and Biomedical Sciences (2019).
https://ieeexplore.ieee.org/document/8991550

[6] Adinugroho S., Sari Y. A., Hidayat N. Drug usage duration classification using Extreme Learning Machine based on personality features. International Conference on Sustainable Information Engineering and Technology (2019).
https://ieeexplore.ieee.org/document/8986131

[7] Cardoso J., Sousa R. Measuring the Performance of Ordinal Classification (2011).
https://doi.org/10.1142/S0218001411009093