Proceedings of the 2023 International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/8/20230097

Predictive model of psychoactive drugs consumption using classification machine learning algorithms

Mothanna Almahmood[1], Hassan Najadat[1], Dalia Alzu'bi[1], Laith Abualigah[2,3,4,5,6,8] , Raed Abu Zitar[7], Sayel Abualigah[1], Faisal AL-Saqqar[2]

[1]Computer Information System Department, Jordan University of Science and Technology, Jordan.
[2]Computer Science Department, Prince Hussein Bin Abdullah Faculty for Information Technology, Al al-Bayt University, Mafraq 25113, Jordan.
[3]Hourani Center for Applied Scientific Research, Al-Ahliyya Amman University, Amman 19328, Jordan.
[4]Faculty of Information Technology, Middle East University, Amman 11831, Jordan.
[5]School of Engineering and Technology, Sunway University Malaysia, Petaling Jaya 27500, Malaysia
[6]School of Computer Sciences, Universiti Sains Malaysia, Pulau Pinang 11800, Malaysia.
[7]Sorbonne Center of Artificial Intelligence, Sorbonne University-Abu Dhabi, Abu Dhabi, United Arab Emirates

[8]aligah.2020@gmail.com

Abstract. It is difficult to predict the effect of drugs on the individuals, as its results are unpredictable and most often dangerous. For a police purpose that concerned with the protection of individuals, the problem of predicting drug abusing is highly important. A dataset was used from open-source website UCI, that includes specific attributes about using up of eighteen different psychoactive drugs. Our study aimed to use data mining classification techniques, in order to classify the individual into two categories: user or non-user. Eighteen classification models were built using different classification algorithms such as Gaussian Naive Bais, Logistic Regression, k-nearest neighbors, Random Forest, and Decision Tree. The accurate classifier was chosen by studying the accuracy, recall, precision, and f1-score measures for each one, and it was evaluated by the Holdout method. The results were obtained optimally, and we got 18 models, where each one had different high accurate outputs, that classify an individual to user and non-user. The final model is a combination of 18 models for 18 critical psychoactive drugs: Alcohol, Amphet, Amyl, Benzos, Caff, Cannabis, Choc, Coke, Crack, Ecstasy, Heroin, Ketamine, Legalh, LSD, Meth, Mushrooms, Nicotine and VSA. This study in turn may give a chance for the decision makers to reduce the risk of these drugs consumption, in order to avoid healthcare issues and keep the community in safe.

Keywords: predictive model, psychoactive drugs, classification, machine learning algorithms.

DOI: 10.54254/2755-2721/8/20230097

1. Introduction

The psychoactive drug is considered to have a strong effect on the nervous system, as the individual experiences temporary changes in perception, awareness and behavior [1]. These changes could probably make individuals doing abnormal behaviors, they may be illegal and completely affect the nature of their lives, and the society as a whole. There are many common types, and they differ from each other in terms of their effect, also they differ in terms of consumption, that can be taken via different ways. The main reason that makes the individuals to use these drugs, is because of its effect on changing awareness that is like other enjoyment activities, that serves satisfying the personality level, acceptance, and decreasing the level of distress [2]. In our research we study misdeed of psychoactive drug nevertheless if it's legal or illegal, just focusing on helping the decision makers to help the population to stop consumption these psychoactive drugs, that may affect their health level or leads them to death. Many people who abuse psychoactive drugs use the justification that they are only harming themselves in order to defend their dependence. However, addiction is not a one-person phenomenon; it impacts everyone around them, making it a societal issue that impacts people and their families [3]. There are many factors that correlated to the psychoactive drug abuse and associated with many personality traits. Psychologists concurred that the personality characteristic of the Five Factor Model (FFM) are complete criteria for understanding individuals' dissimilarities [4]. The FFM includes Neuroticism (N), Extraversion (E), Openness to Experience (O), Agreeableness (A), and Conscientiousness (C). In our study, we have used a dataset that is available from open source website "UCI repository" [5], that includes 12 attributes that are known for each individual: personality characteristics, impulsivity, sensation seeking, education, age, gender, residence, and ethnicity. The dataset includes details about the consumption of eighteen psychoactive drugs including alcohol, Amphet, amyl, benzos, cannabis, choc, coke, Caff, crack, ecstasy, heroin, ketamine, Legalh, LSD, meth, mushrooms, nicotine, and VSA. The participants who were selected in the data set are categorized to non-users, users over a decade ago, users last decade, users last year, users last month, users last week and users last day. We use one definition of 'drug using' based on the use, by identifying two labels; user and non-user, so that we have merged all periods of using from last month until last day in 'user' category and kept other periods as 'non-user' category.

The objective of our study was to predict if an individual is using a psychoactive drug or not, according to personality profiles. Our study shows the strong association relationship between the personality characteristics and belonging to classes of users and nonusers for the eighteen psychoactive drugs. We have applied different data mining techniques, such as; Gaussian Naive Bais, Logistic Regression, k-nearest neighbors, Random Forest, and Decision Tree. The results of classification task were high and accurate. We evaluated each classifier by Holdout method, and the accuracy of prediction was very high. Final model was a collection of eighteen models for eighteen critical psychoactive drugs. Our study would help the decision makers to reduce the risk of the psychoactive drugs consumption to avoid its issues and keep the community in safe. In Section 2 we summarize the most important related works to our research, and in section 3 we describe our methodology, whereas in section 4 we show the experiments and results analysis, and finally in section 5 we finish by concluding our study.

2. The proposed method

In order to create a model, we must consider the issue we are attempting to resolve. In this case, we must determine whether or not a person is taking drugs, and what is the type and period of the drug that he is using, so the problem therefore becomes: How well can we predict that an individual using drugs or not, and what is the period of using if he is a user? In this work we have used data mining algorithms to predict if an individual is using drugs or not, and what the period of his using.

2.1. Dataset

The dataset of interest concerns drug consumptions information retrieved from open-source website "UCI repository" [5], and contains (1886) instances with (31) attributes without aby missing values.

DOI: 10.54254/2755-2721/8/20230097

The dataset comprises nominal and numerical attributes, where ID, Age, Gender, Education, Country, Ethnicity, Nscore, Escore, Oscore, Ascore, Cscore, Impulsive and SS makes up the numerical attributes, and their description in the Table 1.

Table 1. Dataset Description.

Name	Description
ID	number of records in original database
Age	age of participant
Gender	gender of participant
Education	level of education of participant
Country	country of current residence of participant
Ethnicity	ethnicity of participant
Nscore	Neuroticism
Escore	Extraversion
Oscore	Openness to experience
Ascore	Agreeableness
Cscore	Conscientiousness
Impulsive	Impulsiveness
SS	sensation seeing

2.2. Classification

The goal of our study is to identify drug consumption trends in order to create a multi-label classification-based prediction model. Each character in the multi-label classification issue may be related to one or more extracted features. Python programming language was used to apply several machine learning methods. The basic principle of each used algorithm is defined in the following:

- Gaussian Naïve Bais: A Bayes' theorem-applying supervised learning technique that makes the "naive" assumption that each pair of features is conditionally independent given the result of the classification model [6,7,8].
- Logistics Regression: Using coordinate descent and least-squares regression, a generalized linear model was developed. Additionally, it uses related methods for stochastic gradient descent [9].
- KNN: An approach for supervised learning in which data samples are categorized based on the consensus of the k closest training points. Due to the lack of a training process, it is a powerful method. The parameter k in our mode was set to 5 by us [9,10].
- Random Forest: An estimator that uses categorizing decision trees on various dataset sub-samples in order to enhance prediction accuracy and prevent overfitting [9].
- Decision Tree: A supervised learning technique that creates a tree-like structure in order to generate decision rules from the dataset's characteristics and produce a model that forecasts the value of the final feature.

Based on the preprocessed dataset, we have used all the above machine learning algorithms through python to establish classification models respectively on each one of the 18 datasets [11].

DOI: 10.54254/2755-2721/8/20230097

3. Experimental Results and Analysis

We utilized the train test split function from the sk-learn package and the hold-out approach to estimate classification performance when training the model [12]. The technique we utilized divides the dataset into two halves, randomly selecting 30% of it as the test set and 70% of it as the train set, where the evaluation of algorithms for each dataset are displayed. The comparison of accuracy results shows us the best classifier in each model for predicting drug consumption. We have selected 18 models: Alcohol, Amphet, Amyl, Benzos, Caff, Cannabis, Choc, Coke, Crack, Ecstasy, Heroin, Ketamine, Legalh, LSD, Meth, Mushrooms, Nicotine and VSA. Because they obtained high results of accuracy level that lies between 70-98%. Table 2 indicates that the most accurate method for Alcohol drug is Random Forest, with 96% accuracy level. Table 2 indicates that the most accurate method for Amphet drug is Logistic Regression, with 72% accuracy level. Table 2 indicates that the most accurate method for Amyl drug is Random Forest, with 80% accuracy level. Table 2 indicates that the most accurate method for Benzo's drug is Logistic Regression, with 71% accuracy level.

Table 2. Accuracy for classifiers.

Model	Decision Tree	Random Forest	Gaussian Naive Bais	K Neighbors Classifier	Logistic Regression
Alcohol	0.918728	0.961131	0.636042	0.765018	0.680212
Amphet	0.618375	0.710247	0.674912	0.671378	0.719081
Amyl	0.731449	0.795053	0.530035	0.655477	0.676678
Benzos	0.590106	0.701413	0.681979	0.648410	0.712014
Caff	0.946996	0.977032	0.681979	0.830389	0.710247
Cannabis	0.749117	0.814488	0.782686	0.768551	0.810954
Choc	0.941696	0.982332	0.752650	0.832155	0.696113
Coke	0.628975	0.696113	0.662544	0.613074	0.628975
Crack	0.805654	0.858657	0.680212	0.712014	0.699647
Ecstasy	0.655477	0.710247	0.701413	0.681979	0.713781
Heroin	0.800353	0.855124	0.729682	0.694346	0.699647
Ketamine	0.710247	0.745583	0.660777	0.628975	0.657244
Legalh	0.683746	0.787986	0.763251	0.740283	0.683746
LSD	0.689046	0.759717	0.743816	0.720848	0.750883
Meth	0.720848	0.766784	0.726148	0.689046	0.713781
Mushrooms	0.671378	0.719081	0.747350	0.706714	0.756184
Nicotine	0.660777	0.743816	0.685512	0.628975	0.712014
VSA	0.809187	0.846290	0.736749	0.742049	0.727915

Table 2 indicates that the most accurate method for Caff drug is Decision Tree, with 95% accuracy level. Table 2 indicates that the most accurate method for Cannabis drug is Random Forest, with 81% accuracy level. Table 2 indicates that the most accurate method for Choc drug is Decision Tree, with 94% accuracy level. Table 2 indicates that the most accurate method for Coke drug is Random Forest, with 70% accuracy level. Table 2 indicates that the most accurate method for Crack drug is Random Forest, with 85% accuracy level. Table 2 indicates that the most accurate method for Ecstasy drug is Logistic Regression, with 71% accuracy level. Table 16 indicates that the most accurate method for Heroin drug is Random Forest, with 85% accuracy level. Table 2 indicates that the most accurate method for Ketamine drug is Random Forest, with 74% accuracy level. The above Table 2 indicates that the most accurate method for Legalh drug is Random Forest, with 79% accuracy level. Table 2 indicates that the most accurate method for LSD drug is Random Forest, with 76% accuracy level. Table 2 indicates that the most accurate method for Meth drug is Random Forest, with 77% accuracy level. Table 2 indicates that the most accurate method for Mushrooms drug is Logistic Regression,

DOI: 10.54254/2755-2721/8/20230097

with 76% accuracy level. Table 2 indicates that the most accurate method for Nicotine drug is Random Forest, with 71% accuracy level. Table 2 indicates that the most accurate method for VSA drug is Random Forest, with 85% accuracy level. Thus, from all the above results of the eighteen model, we consider Random Forest, Decision Tree and Logistic Regression as the most suitable algorithms for predicting the psychoactive drug consumption of individuals. The accuracy rate was different, since that all model had accuracy level lie between 70-96%. With these high results of prediction, we have combined the 18 models into one model, that accept inputs and give prediction outputs about individual psychoactive drug consumption (user and non-user). Finally, we can say that our research target is claimed.

4. Conclusion

As we mentioned before, the problem of predicting drug abusing is highly important, for its benefit on the whole community. From this standpoint our study came out, to use different machine learning classification techniques for classification problem, such as Gaussian Naive Bais, Logistic Regression, k-nearest neighbors, Random Forest, and Decision Tree. In order to classify the individuals into two different labels (user and non-user), which belong to the class of each drug of the known eighteen drugs. The dataset was fully cleaned and helpful, so that we did not have to do any process of pre-processing, except dropping out the unneeded attribute "ID". The dataset helped our goal, because of its information about personality traits and others. The most accurate classifier was picked up for each model by studying the accuracy measure of each one, and it was evaluated by Holdout method. The dataset gave the best results with using Random Forest and Logistic Regression, with a high level of accuracy that is good enough for prediction. The accuracy rate was different from model to another. And combing these eighteen models together in one model, was efficient and useful to do one single task of prediction. This work could be a new step to go over reducing the drugs abusing, and increasing the healthcare level, by studying the reasons that may make an individual abuse on a drug, and provide suitable solutions and fix all issues, in order to stop doing this behavior, also this work may help the police for their investigations, which in turn may help the whole community. In the future work, we hope to use larger dataset with more information that help us to predict add more information to our model, so that we can predict the period of consumption for the user (in days) for all types of drugs, and that may help our research target, to find the appropriate period for drug cessation for all cases.

References

[1] Psychoactive drug. (n.d.). Retrieved December 26, 2019, from https://www.sciencedaily.com/ter-ms/psychoactive_drug.htm.

[2] Department of Health | 3.1 Reasons why people use drugs. (n.d.). Retrieved December 26, 2019, from https://www1.health.gov.au/internet/publications/publishing.nsf/Content/drugtreat-pubs-front5-wk-toc~drugtreat-pubs-front5-wk-secb~drugtreat-pubs-front5-wk-secb-3~drugtreat-pubs-front5-wk-secb-3-1.

[3] Why Drug Addiction Is a Social Problem -. (n.d.). Retrieved December 26, 2019, from https://www.hotelcaliforniabythesea.com/2019/02/26/why-drug-addiction-is-a-social-problem.

[4] Costa, P. (2018). Neo PI-R professional manual. (January 1992).

[5] UCI Machine Learning Repository: Drug consumption (quantified) Data Set. (n.d.). Retrieved De-cember 26, 2019, from https://archive.ics.uci.edu/ml/datasets/Drug+consumption+%28quantified%29#.

[6] Welcome to Python.org. (n.d.). Retrieved December 27, 2019, from https://www.python.org.

[7] Perceptron Definition | DeepAI. (n.d.). Retrieved December 13, 2019, from https://deepai.org/machine-learning-glossary-and-terms/perceptron.

[8] Sathishkumar, V. E., & Cho, Y. (2019, December). Cardiovascular disease analysis and risk assessment using correlation based intelligent system. In Basic & clinical pharmacology &

DOI: 10.54254/2755-2721/8/20230097

toxicology (Vol. 125, pp. 61-61). 111 RIVER ST, HOBOKEN 07030-5774, NJ USA: WILEY.

[9] Sathishkumar, V. E., Park, J., & Cho, Y. (2020). Using data mining techniques for bike sharing demand prediction in metropolitan city. Computer Communications, 153, 353-366.

[10] VE, S., Shin, C., & Cho, Y. (2021). Efficient energy consumption prediction model for a data analytic-enabled industry building in a smart city. Building Research & Information, 49(1), 127-143.

[11] VE, S., Park, J., & Cho, Y. (2020). Seoul bike trip duration prediction using data mining techniques. IET Intelligent Transport Systems, 14(11), 1465-1474.

[12] Chen, J., Shi, W., Wang, X., Pandian, S., & Sathishkumar, V. E. (2021). Workforce optimisation for improving customer experience in urban transportation using heuristic mathematical model. International Journal of Shipping and Transport Logistics, 13(5), 538-553.