A Comparative Machine Learning Study to Predict Drug Addiction in Bangladesh

Conference Paper · October 2021

DOI: 10.1109/AICT52784.2021.9620453

4 authors, including:

K M Rashedul Alam
Friedrich-Alexander-University of Erlangen-Nürnberg

Khair Ahammed
Noakhali Science & Technology University

A Comparative Machine Learning Study to Predict Drug Addiction in Bangladesh

K. M. Rashedul Alam[‡], Khair Ahammed[*], Mohammad Abu Tareq Rony[‡], Zannatul Ferdousi[§],
[*]Department of Data Science, FAU Erlangen–Nürnberg, Germany
[†]Department of Software Engineering, Noakhali Science and Technology University, Noakhali, Bangladesh
[‡]Department of Statistics, Noakhali Science and Technology University, Noakhali, Bangladesh
[§]Department of Computer Science Engineering, Southeast University, Dhaka, Bangladesh
Email: rashedulalam.km@gmail.com[*], khairahmad6@gmail.com[†], abutareqrony@gmail.com[‡] zannatulferdousi.cse@gmail.com[§]

***Abstract*—Drug Addiction is one of the growing threats all over the world. According to Dhaka Tribune, more than 7.5 million people are addicted to drugs in Bangladesh. There are a lot of differences between a drug-addicted and a non-addicted person on health condition, social life, personal life, and familial life behaviors. So, steps should be taken to prevent drug addiction with proper curative issues. In this paper, we dig for the influential factors behind drug addiction and possible solutions to reduce the drug addiction rate. The research is held on the people of Dhaka, Bangladesh. Most of the data of drug-addicted people are collected from 'Drug Rehab' and for non-addicted person data we have collected from different schools, colleges, and universities in Dhaka, Bangladesh. All are male and the age group of 17 to 45 years. Our primary data set is constructed including only 188 qualitative data. A total of 5 algorithms have been employed including Logistic Regression, Decision Tree, Random Forest, Naive Bayes, Support Vector Machine (SVM) and their results are compared. Among the algorithms Random Forest comes up with the highest accuracy of 97.3484%, XGBoost & Decision Tree Classifier delivers the accuracy of 96.2768% and 94.68%.**

***Index Terms*—Drug, Machine Learning, Statistical Analysis**

I. Introduction

Machine learning is a growing area of computational algorithms that seek to simulate human intelligence by learning from their surroundings [17]. Drug addiction is the worst social-economic problem in modern civilization [16]. Drug misconduct directly impacts the social and economic perspective in Bangladesh and is a very growing threat also. The statistics in Bangladesh for drug addiction in 2018, Bangladesh has more than seven million drug-addicted, and more than 5 million of them are hooked on "Methamphetamine", [4] according to people intimate with the situation. Every social problem has some attributes in society. Drug addiction is a major problem and it has a relation to social and familial behavior, standards. So, the government should comprise both the family and society which are the two most workable organizations to prevent drug addiction. So that we choose this topic because drug addiction is a complex disease in our society. But most people don't have enough knowledge about how people become addicted to drugs. [17]. In Asia, Bangladesh is a very small country by area and the population is very high [7]. It is surrounded by India from three sides. The Eastern and Northern sides are very important because they are surrounded by large mountains. And The western side is fertile and prolific plain land. The mountainous regions are mainly suitable for illegal drug sales and drug traffic. The drug smugglers can easily hide in these hilly forests and safely transport the narcotics. However, due to an increase in drug abusers, Bangladesh has become a high-risk country for long-term growth. The majority of the abusers are young. As a result, anti-drug results are critical challenges for Bangladesh's growth [18].To produce a well-educated young generation, the country's government should urgently pass legislation to address substance use and addiction. In the twenty-first century, a huge number of people are suffering from a drug addiction problem. Most of them are very young and teenage boys. These populations are becoming a burden for our society rather than being regarded as manpower. This population needs help from us to come back to a productive life. We wanted to establish a supervised machine learning-based model that will easily identify a person is drug addicted or not addicted. Because drug addiction is very harmful to our society. Again, Drug abusing is a social problem. Every parent, guardian of a drug abuser is concerned with this problem. Also, governments of every country of the world are fighting to eradicate it. However, addiction is not a crime, but it is internationally acknowledged as a disease by addiction professionals. So, like other diseases, such as dengue, cancer stages, heart disease can be measured it could be also measured using any social, physical, mental, family relationship-based indicators. We have decided to contribute to this sector, which might help the drug-addicted family as well as the government and other people who are concerned with the drug addiction problem in Bangladesh. Our main vision was to find out those relations between all the features based on learning data. We collected the data from drug rehab-related resources and work with the machine learning approach. By using machine learning approaches to Predict the public and personal life behaviors of a drug abuser is the aim of our research. In our thesis, we have discussed how we find the actual factors behind the drug addiction problem. We have made the questions based on the present situation, very common behavior, and real-life expression and social behavior of the drug-addicted people in our area. In our data set attributes are in fact answers mass questions related to people's personal life, health condition, social, and familial life. The data set is the outcome of expression, behavior, and health condition, which has consisted of a total of 18 features that help to find out the problems of drug addiction. We also find out the factors which are the most dominant people to take drugs.

II. Related Work

This section summarizes the related Machine Learning techniques employed for prediction. We have found many recent machine learning approaches, few recent research work in the domain of interest is summarized in Table. Besides, in Decision tree classification algorithm to find the structural, physical, and chemical characteristics of compounds that predispose them to cause Adverse Drug Reactions (ADRs) [1] A structure-activity relationship analysis was presented consisting of ADRs in the central nervous system (CNS), kidney, and liver also of allergic responses for classifying drugs that could be suspected of producing adverse reactions [1]. Decision trees are likely one of the most frequently popular methodologies for machine learning [14].

Logistic Regression algorithm for prediction in form of classification. Which extracts features properly between the two important attributes based on classification [2]. We have seen many papers to put the importance of mostly cigarette smoking and weed as having a great impact on drug abuse [3]. In another paper based on Predicting the public and private life behaviors of a drug abuser in Bangladesh [4]. In another paper, the authors emerged the statistical analysis of more than 8 algorithms [4]. They had demonstrated the statistical differences between many algorithms based on accuracy measures. As we see most people are used Random Forest, Decision Tree, Logistic Regression, KNN, Naïve Bayes, Neural Network, etc. [6]. To reach their final goals and some valuable programming tools, for instance, WEKA, Python provides executions of the machine learning algorithms that are widely used to model drug toxicity prediction [26] [27].

III. RESEARCH METHODOLOGY

In this section, we are going to discuss the procedures that we have followed to complete the thesis successfully. There have several steps that we followed.

A. Data Collection and Overview

In the current situation of Bangladesh, drug addiction is a very common and growing threat [28]. Our main objective is to organize the machine learning concepts and reduce the addiction rate. So, primarily we have collected data for 25 specific questions. Due to the missing data and privacy issue, we were not able to add all the features to the dataset. The research was held for the people of Dhaka, Bangladesh. Most of the samples for drug-addicted people, we have collected from 'Drug Rehab', and for non-addicted sample's data, we collected from various schools, colleges, and universities in Dhaka, Bangladesh. All the samples are male and the age group of 17 to 45 years. Our primary data set was constructed on 188 samples of 25 features. After cleaning and processing the initial dataset, the final dataset was of 188 samples and 19 features. For collecting the information, we followed a questionnaire methodology, asked for personal intervention, and looked for different aspects.

B. Data Pre-processing

We have collected the data of 188 people from the local rehab center and other sources and in the end, we have got so much messy data. So, for further processing, we needed to clean and process our data. There were several missing data and type mismatch features. So, we removed 6 features which lead to numerous percentages of missing values and impute other missing column values with their mean and frequency. Also, the data type of each feature has been checked and fixed. Finally, our data set contains 188 samples with 19 features. We have used a label encoder to encode the data to a numeric format for further processing. We have performed a different statistical analysis to find important features among those features. Chi-Square and PCA (Principal Component Analysis) are performed. The Chi-Square test has given us better accuracy in the final score. So, after finding out the score of each feature we've dropped the less important column that will help the model being simpler and more powerful. Finally, our data set contains 188 samples with 14 features.

C. Classification

Classification is a supervising technique that categorizes the data into the desired number of classes [15]. The goal of this work is to find out the factors behind drug addiction and predict a person's probability of being drug-addicted. In that manner, we can decrease the addiction rate and keep people especially teenagers away from this deadly addiction. So, we have employed 7 classifiers: Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), Bernoulli Naïve Bayes (BNB), Gaussian Naïve Bayes (GNB), Support Vector Machine (SVM) and XGBoost (XGB). Finally, we made a comparison of their performance based on different model evaluation metrics and find out the best fitting algorithm for this piece of the problem. We considered hyper-parameter tuning and 10-fold cross-validation for making our model more powerful and generic as well. The Table shows the parameter distributions of each classifier

IV. Evaluation Metrics

For executing the different machine learning models and finding the best one various evaluation metrics are there [20]. Different evaluation techniques are introduced based on the confusion metrics such as accuracy, precision, recall, and f-measure and our model evaluation is done based on these four evaluation criteria [8].

- **Accuracy:** It represents how many instances are correctly predicted with respect to total observations [21]. Hence, the accuracy is defined as follows:

TABLE I: SUMMARY OF RELATED WORK:

Source	Methodology	Objectives	Result
[8]	Clustering, classification and filtering method	Institute and family to identify student's academic performance and addiction	Pay greater attention to students and limit the effect of alcohol on their lives.
[9]	Feature extraction, classification algorithm, descriptive statistics	Selected significant models, classifying AUD and unaffected controls divided by ancestry, age and gender	The value of sampling uniformity, followed by stratified analysis and feature selection, to produce better prediction scores and allow for more accurate AUD production estimation.
[10]	Descriptive statistics, PCA, box plot, classification algorithm	Machine learning can be used to predict the probability of being addicted to drugs.	The risk prediction model with the best performance is logistic regression.
[4]	The Machine learning algorithm, feature selection, reliability test	Find out the characteristics of a person that can prove his vulnerability to drug addiction.	When a person's variables are linked in one dot, a possible drug abuser may be detected.
[11]	ANN	Artificial neural network (ANN) based approach for prediction of alcohol user	The proposed method can be used to reliably determine whether or not an individual is an alcoholic.
[12]	Decision tree and some other machine learning algorithms	Using a decision tree classifier, a subset of Psychological, behavioral, and laboratory tests that best predicted whether a person with (AUD) seeks care was defined.	The ADT classifier had ten measures that were related to psychological disorders as well as drug abuse.
[13]	Data Mining, decision tree and some other algorithms	Using a public dataset, educational data mining was used to investigate student alcohol consumption.	Males drink more alcohol than females, and a high degree of socializing with friends contributes to increased alcohol consumption.

TABLE II: DESCRIPTIVE STATISTICS

	Minimum	Maximum	Sum	Mean	Std. Deviation	Variance	Skewness	Kurtosis
Gender	1	2	375	1.99	0.073	0.005	-13.711	188
Age	17	45	4589	24.41	3.625	13.142	1.885	7.587
Religion	1	2	361	1.92	0.272	0.074	-3.127	7.859
Study Medium	1	2	201	1.07	0.254	0.065	3.424	9.827
Education	1	4	464	2.47	0.727	0.528	1.041	-0.07
Family Status	1	3	486	2.59	0.752	0.565	-1.44	0.337
Relation with Family	1	3	453	2.41	0.668	0.446	-0.697	-0.589
Family Members	1	7	558	2.97	1.21	1.464	0.83	0.647
Marital Status	1	4	666	3.54	0.657	0.431	-1.246	0.839
Fathers Occupation	1	3	305	1.62	0.702	0.493	0.68	-0.729
Mothers Occupation	1	3	210	1.12	0.383	0.147	3.471	12.054
Smoke?	1	2	314	1.67	0.471	0.222	-0.73	-1.483
Total Friends	5	300	9290	49.41	48.759	2377.442	2.525	8.197
Percentage of Smoker Friend	2	100	12787	68.02	29.375	862.871	-0.825	-0.468
Take Weed?	1	2	287	1.53	0.501	0.251	-0.107	-2.01
Stay Out Night	1	2	276	1.47	0.5	0.25	0.129	-2.005
Parents Existence	1	2	357	1.9	0.302	0.091	-2.668	5.176
Spending Money	100	80000	1531500	8146.28	12767.331	163004745.4	3.235	10.937
Drug Addicted	1	2	272	1.45	0.498	0.248	0.216	-1.975

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

- **Precision and Recall:** Precision is the percentage of related instances found among the retrieved samples, whereas recall is the percentage of samples found [22].

$$Precision := \frac{(TP)}{(TP + FP)}$$

$$Recall = \frac{TP}{(TP + FN)}$$

- **F- measure:** F-Measure gives a way to combine precision and recall into a single measure [23].

$$F - measure = \frac{2 * Recall * Precision}{Recall + Precision}$$

Corresponding classifier's performance over Accuracy, Precision, Recall, and F-measure values are listed. TP defines True Positive, TN defines True Negative, FP defines False positive, FN defines False Negative [25].

A. *Parameter Settings Table*

It's natural for a decision tree or ensemble to over-fit the data as it grows [17]. Pruning is the process of selecting the most massive tree that is the most generalized and removing all of the branches below that level [24]. This dramatically improves performance on new data.

V. Experimental Result

We aimed at understanding the important factors behind drug addiction and also prevent an individual from being addicted to deadly drugs. So, the very first study was on finding the factors, and secondly to predict an individual that has higher chances to become drug-addicted.

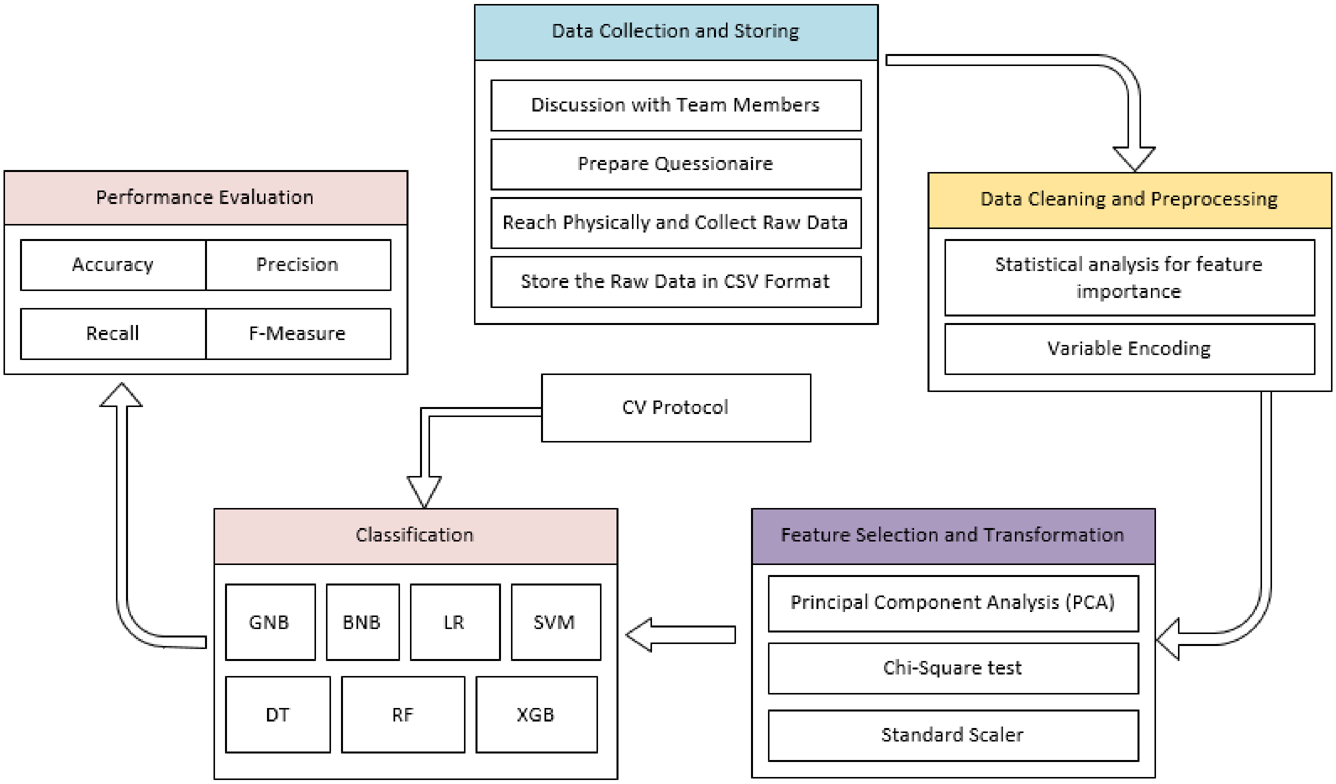

Fig. 1: Proposed Methodology

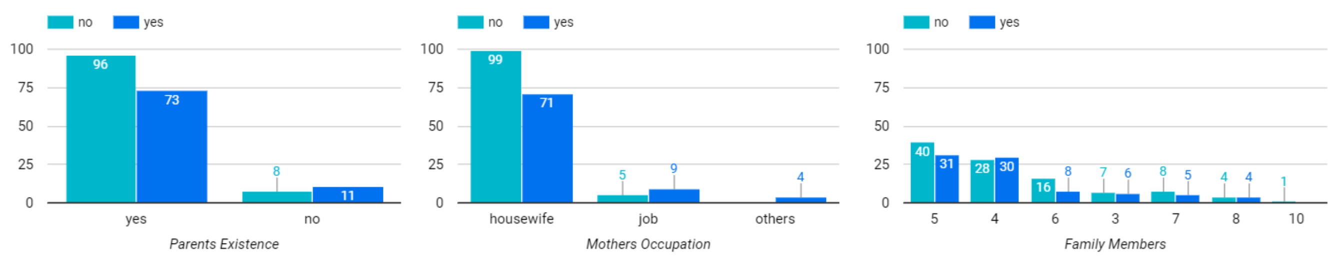

Fig. 2: Drug Addiction Distribution by Family conditions

TABLE III: ASSOCIATED PARAMETERS

Classifier	Parameters
DT	Criterion = gini, Splitter = best
RF	Max depth = none, random state = 0
LR	Liblinear solver, max iter = 2000
BNB	Alpha = 1.0, class prior = 0
GNB	Priors = none, var smoothing = 1e^-09
SVM	Kernel = rbf, max iter = -1
XGB	Learning rate = 0.1, max depth = 3

A. *Feature Discussion and Analysis*

We have done a uni-variate analysis to find the important factors. We have found some insights and important factors through analyzing the data set. We have considered binning the age values into 5 age groups. The data set consists of an age group of 17 years to 45 years. But for having limited data we won't make consumption about the age group above 30. Teenagers are mostly attracted to drugs and at the age of 17-21, there are higher chances to get addicted to drugs. Also, 22-30 age group there are risks get addicted to drugs. Smoke is considered the primary indicator of a drug-addicted person. Considering the age group and family status some shocking insight has been revealed. 'High Class' status people are more addicted to drugs of any age group. They are at high risk in terms of getting addicted to deadly drugs. On the other hand, middle and lower-class teenagers are mostly attracted by tobacco and later on after ages get addicted to deadly drugs.

TABLE IV: EXPERIMENTAL RESULTS

Age Group	Smoke	Take Weed	Drug Addicted
High Class (Avg. Spending Money = 26,266.67Tk)			
(17-21]	100%	83.33%	100%
(21-26]	100%	100%	85%
(26-31]	100%	100%	100%
Class (Avg. Spending money = 4880Tk)			
(17-21]	61.11%	44.44%	44.44%
(21-26]	58.65%	41.34%	30.76%
(26-31]	56.25%	43.75%	43.75%
Lower Class (Avg. Spending money = 3350Tk)			
(17-21]	100.00%	75%	0.00%
(21-26]	63.63%	45.45%	45.45%
(26-31]	66.67%	66.67%	66.67%

Table 2 illustrates the percentages of the different factor's outcomes.

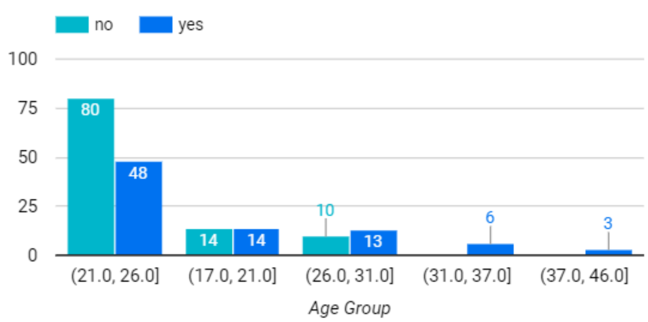

Fig. 3: Drug Addiction Distribution by Age Group

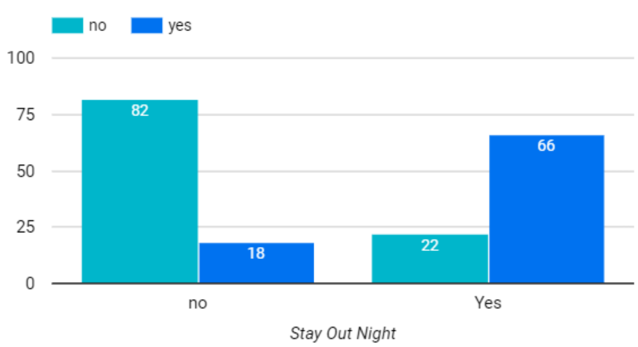

Fig. 4: Staying out at night

The family can play a major role here in preventing their son/daughter from being drug-addicted. Those people who don't have parents are more likely to be addicted to drugs. Also, larger family reduces the drug addiction rate than smaller families. Figure 2 illustrates the relations of drug addiction with family. Another most important factor is staying out at night. The people who stay out at night are more likely to be addicted than others. Figure 3 illustrates the probabilities of being drug-addicted who are staying out at night.

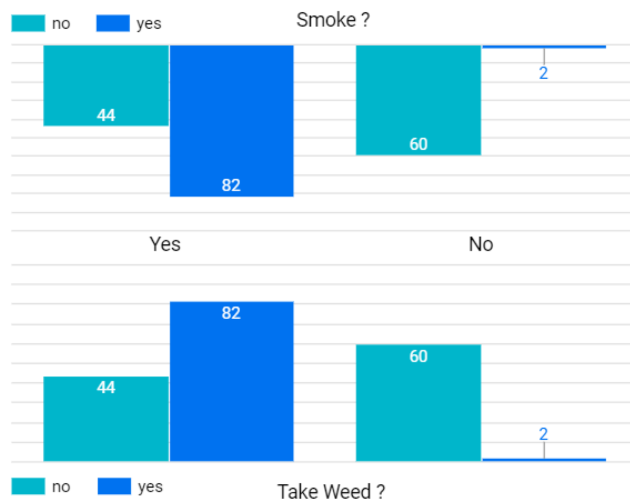

Fig. 5: Take Weed

TABLE V: THE PERFORMANCES OF DIFFERENT MODELS

Algorithms	Accuracy	Precision	Recall	F1-score
Naive Bayes (Bernoulli)	0.8936	0.8404	0.94	0.8876
Random Forest	0.9734	0.9647	0.976	0.9704
Logistic Regression	0.8882	0.8387	0.929	0.8813
Decision Tree	0.9468	0.9404	0.94	0.9404
Support Vector Machine	0.9148	0.8617	0.964	0.9101
Naive Bayes (Gaussian)	0.867	0.8241	0.893	0.8571
XGBoost	0.9627	0.9425	0.976	0.959

VI. Conclusions and Future Work

Bangladesh has millions of drug-addicted people and most of them are young. Drug addiction is one of the growing threats. Every day a lot of people are being addicted to drugs. Drug use remains at an unacceptable level and continues to bring affliction to human beings. It associated finances convict and, to some extent, terrorist activities. We found some important factors which are the most dominant people to take drugs and it will help people to learn about drug addiction. We wish our research will make a good impression on our society.

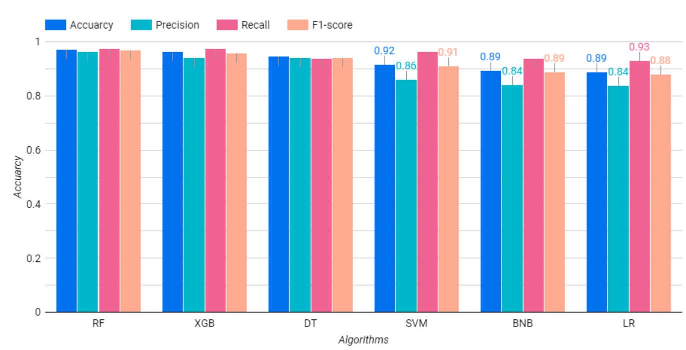

Fig. 6: Experimental Result for Different Algorithms

As we have collected complete data set for only males and age groups of 17 to 45 years old. All drug abusers from rehab and non-addicted only school, college, and university in Dhaka, Bangladesh. So, we expect that if future research is meditated on this topic, the area of data collection is expected to be much broader all over the country. We have some specific features. We believe we could achieve a better classification result by using some of the selected algorithms which yielded better accuracy with modified hyper-parameter tuning. However, the prediction model has been developed from the 188-sample data which is not large enough to predict the class. In future research, we will work with both males, females people. We will include more features and a more efficient prediction tool can be developed to get more satisfactory results.

REFERENCES

[1] Hammann, Felix, et al. ”Prediction of adverse drug reactions using decision tree modeling.” Clinical Pharmacology & Therapeutics 88.1 (2010): 52-59.

[2] Sahker, Ethan, Laura Acion, and Stephan Arndt. ”National analysis of differences among substance abuse treatment outcomes: College student and nonstudent emerging adults.” Journal of American College Health 63.2 (2015): 118-124.

[3] Myers, Mark G., and John F. Kelly. ”Cigarette smoking among adolescents with alcohol and other drug use problems.” Alcohol Research & Health 29.3 (2006): 221.

[4] Shahriar, Arif, et al. ”A Machine Learning Approach to Predict Vulnerability to Drug Addiction.” 2019 22nd International Conference on Computer and Information Technology (ICCIT). IEEE, 2019.

[5] Acion, Laura, et al. ”Use of a machine learning framework to predict substance use disorder treatment success.” PloS one 12.4 (2017): e0175383.

[6] Zhang, Li, et al. ”Applications of machine learning methods in drug toxicity prediction.” Current topics in medicinal chemistry 18.12 (2018): 987-997.

[7] Peng, Hanchuan, Fuhui Long, and Chris Ding. ”Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy.” IEEE Transactions on pattern analysis and machine intelligence 27.8 (2005): 1226-1238.

[8] Trivedi, Tanvi, and Devangi Kotak. ”Exploring Prediction Modeling of Students Alcohol and Drug Addiction Affecting Performance using Data Mining Approach.”

[9] Kinreich, Sivan, et al. ”Predicting risk for Alcohol Use Disorder using longitudinal data with multimodal biomarkers and family history: a machine learning study.” Molecular psychiatry (2019): 1-9.

[10] Arif, Md, et al. Drug Addiction Prediction Using Machine Learning. Diss. Daffodil International University, 2020.

[11] Kumari, Divya, et al. ”Prediction of alcohol abused individuals using artificial neural network.” International Journal of Information Technology 10.2 (2018): 233-237.

[12] Safavian, S. Rasoul, and David Landgrebe. ”A survey of decision tree classifier methodology.” IEEE transactions on systems, man, and cybernetics 21.3 (1991): 660-674.

[13] Pisutaporn, Auth, Burit Chonvirachkul, and Daricha Sutivong. ”Relevant factors and classification of student alcohol consumption.” 2018 IEEE international conference on innovative research and development (ICIRD). IEEE, 2018.

[14] Sharma H. and Kumar S. “A survey on decision tree algorithms of classification in data mining”; International journal of science and research (IJSR);2016

[15] Kotsiantis, Sotiris B., I. Zaharakis, and P. Pintelas. ”Supervised machine learning: A review of classification techniques.” Emerging artificial intelligence applications in computer engineering 160.1 (2007): 3-24.

[16] Heather, Nick. ”A conceptual framework for explaining drug addiction.” Journal of Psychopharmacology 12.1 (1998): 3-7.

[17] El Naqa, Issam, and Martin J. Murphy. ”What is machine learning?.” machine learning in radiation oncology. Springer, Cham, 2015. 3-11.

[18] Islam, Azizul, and Md Faruque Hossain. ”Drug abuse and its impact on Bangladesh.” International Journal of Sociology and Anthropology 9.11 (2017): 143-156.

[19] Ahammed, M. S. Satu, M. I. Khan and M. Whaiduzzaman, ”Predicting Infectious State of Hepatitis C Virus Affected Patient’s Applying Machine Learning Methods,” 2020 IEEE Region 10 Symposium (TENSYMP), Dhaka, Bangladesh, 2020, pp. 1371-1374, doi: 10.1109/TENSYMP50017.2020.9230464.

[20] Handelman, Guy S., et al. ”Peering into the black box of artificial intelligence: evaluation metrics of machine learning methods.” American Journal of Roentgenology 212.1 (2019): 38-43.

[21] Picek, Stjepan, et al. ”The curse of class imbalance and conflicting metrics with machine learning for side-channel evaluations.” IACR Transactions on Cryptographic Hardware and Embedded Systems 2019.1 (2019): 1-29.

[22] Tatbul, Nesime, et al. ”Precision and recall for time series.” arXiv preprint arXiv:1803.03639 (2018).

[23] Palomba, Fabio, and Damian Andrew Tamburri. ”Predicting the emergence of community smells using socio-technical metrics: a machine-learning approach.” Journal of Systems and Software 171 (2021): 110847.

[24] Maxwell, Aaron E., Timothy A. Warner, and Fang Fang. ”Implementation of machine-learning classification in remote sensing: An applied review.” International Journal of Remote Sensing 39.9 (2018): 2784-2817.

[25] Koizumi, Yuma, et al. ”SNIPER: Few-shot learning for anomaly detection to minimize false-negative rate with ensured true-positive rate.” ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019.

[26] Mitrpanont, Jarernsri, et al. ”A study on using Python vs Weka on dialysis data analysis.” 2017 2nd International Conference on Information Technology (INCIT). IEEE, 2017.

[27] Viloria, Amelec, et al. ”Comparative Analysis Between Different Automatic Learning Environments for Sentiment Analysis.” International Symposium on Distributed Computing and Artificial Intelligence. Springer, Cham, 2020.

[28] Rahman, Mohammad Mizanur, et al. ”Psycho-social factors associated with relapse to drug addiction in Bangladesh.” Journal of Substance Use 21.6 (2016): 627-630.