International Journal of Electrical and Computer Engineering (IJECE)
Vol. 11, No. 5, October 2021, pp. 4471~4480
ISSN: 2088-8708, DOI: 10.11591/ijece.v11i5.pp4471-4480

Prediction of addiction to drugs and alcohol using machine learning: A case study on Bangladeshi population

Md. Ariful Islam Arif, Saiful Islam Sany, Farah Sharmin, Md. Sadekur Rahman, Md. Tarek Habib
Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh

Article Info

Article history:

Received Aug 11, 2020
Revised Dec 16, 2020
Accepted Mar 21, 2021

Keywords:

Addiction
Drugs and alcohol
Logistic regression
Machine learning
Prediction system

ABSTRACT

Nowadays addiction to drugs and alcohol has become a significant threat to the youth of the society as Bangladesh’s population. So, being a conscientious member of society, we must go ahead to prevent these young minds from life-threatening addiction. In this paper, we approach a machine-learning-based way to forecast the risk of becoming addicted to drugs using machine-learning algorithms. First, we find some significant factors for addiction by talking to doctors, drug-addicted people, and read relevant articles and write-ups. Then we collect data from both addicted and non-addicted people. After preprocessing the data set, we apply nine conspicuous machine learning algorithms, namely k-nearest neighbors, logistic regression, SVM, naïve bayes, classification, and regression trees, random forest, multilayer perception, adaptive boosting, and gradient boosting machine on our processed data set and measure the performances of each of these classifiers in terms of some prominent performance metrics. Logistic regression is found outperforming all other classifiers in terms of all metrics used by attaining an accuracy approaching 97.91%. On the contrary, CART shows poor results of an accuracy approaching 59.37% after applying principal component analysis.

Corresponding Author:

Md. Ariful Islam Arif
Department of Computer Science and Engineering
Daffodil International University
Dhanmondi, Dhaka-1207, Bangladesh
Email: ariful15-7871@diu.edu.bd

1. INTRODUCTION

Drug addiction, which means the taking of various drugs illegally and being addicted to those drugs for their toxic and addictive effects, is one of the most malignant problems for a country. It can destroy a life and a nation easily. In a developing country like Bangladesh, addiction can bear a terrible effect on our society. According to the report of the daily star newspaper, Bangladesh has become increasingly involved with terrorist groups involved in drug abuse and production, using Bangladesh's territory to smuggle drugs, which pose a threat to our country's youth society. [1]. Near about 25 lac people are drug-addicted. In Bangladesh, about 80 percent of the drug addicts are adolescents and young men of 15 to 40 years of age [2]. Dissatisfaction is the reason for this addiction. Joblessness issues, political upheaval, absence of family ties, absence of adoration friendship offer ascent to disappointment. In order to avoid drug addiction, we need to stay away from drugs. Stay away from drugs will only reduce the risk of getting addicted before one can be addicted to it. Nowadays drug addiction has become a dangerous fact for which the young generation from all lifestyles is affected silently. In 2015, drug-addicted Oishee Rahman killed her parents [3]. Even it is very difficult for a woman to roam around alone in the city because there are many drug-addicted people freely moving inside the city. When we go to a new place, we cannot find out those who are addicted. An addicted friend can destroy the friend circle easily. According to the news of the Dhaka tribune newspaper, there are around 7.5 million people addicted to drugs in Bangladesh. The dangerous thing about them is that 80% of them are the youth and 50% of them are involved in different criminal activities [4]. We need to keep a special focus so that our youth do not become addicted to drugs. Machine learning, a major branch of artificial intelligence (AI) can provide a solution to the problem just discussed above. The applications of machine learning vary on different application domains, e.g. cancer prediction [5], software fault prediction [6], dermatological disease detection [7], and risk prediction [8] and so on. Likewise, different conspicuous machine learning algorithms can be put into use for the work of prediction to drugs and alcohol.

This paper tries to anticipate in advance if someone has the risk of becoming addicted to drugs and alcohol. First, we read relevant articles from different national and international journals, conference proceedings, and magazines and write-ups from different websites and newspapers. Then we talk to doctors and drug-and-alcohol-addicted people and find some driving factors for addiction such as age, gender, profession, health ability, mental pressure, trauma, family-and-friends' history, life-changing incidents. Collecting raw data from both addicted and non-addicted people. We made an arduous endeavor for comparing our results with the results of similar research works even though no work has been observed, which addresses the problem of prediction of addiction to drugs and alcohol.

We have followed and studied related works in the near past done by some other researchers on drugs and addiction prediction and understand the processes and methods expressed by them. Here are some descriptions of recent notable research work on machine learning. Dahiwade *et al.* [9] proposed a general disease prediction system, which was based on machine learning algorithms. Hegazy et al. [10] proposed a model for stock market prediction with machine learning technology. Alonzo *et al.* [11] presented a detailed comparison between various machine learning algorithms used prediction and assessment of coconut sugar quality. Haghiabi *et al.* [12] worked on predicting water quality in the machine learning approach. Zhang *et al.* [13] proposed a method for predicting daily smoking behavior based on the machine-learning algorithm. They used the extreme gradient boosting (XGBoost) decision tree algorithm and found the best accuracy of 84.11% with maximum depth five. Alaa *et al.* [14] proposed a machine learning-based model for predicting disease risk of cardiovascular on Biobank participants. Zhu *et al.* [15] worked on pre-symptomatic detection of tobacco disease with hyperspectral image and machine-learning classifiers.

Zhang *et al.* [16] worked to predict human immunodeficiency viruses (HIV) prognosis and mortality with smoking-associated deoxyribonucleic acid (DNA) and machine learning classifiers. Granero *et al.* [17] proposed a model for predicting exacerbations of obstructive pulmonary disease with machine learning features. Frank, Habach, and Seetan [18] worked on smoking status prediction with machine learning and statistical analysis. Logistic regression had the best performance with 83.44% accuracy, 83% precision, 83.4% recall and 83.2% *F*-measure in their work. Lee *et al.* [19] worked with a model that predicts alcohol use disorder by checking the treatment-seeking status with a machine learning classifier. Their collected data domains were cognitive, mood, impulsivity, personality, aggression, and early life stress and childhood trauma. Kinreich *et al.* [20] proposed a model on predicting the risk of alcohol use disorder (AUD) using machine-learning technology. Kumari *et al.* [21] proposed a model of predicting alcohol abused using machine learning technology. They considered age, gender, country, ethnicity, education, neuroticism, openness to experience, extraversion, agreeableness, conscientiousness, impulsive, sensation seeing as their models' feature. These features considered in ANN-D and day, week, month, year, decade considered in ANN-C. Habib *et al.* [22] had done a study on Papaya disease recognition based on a machine learning classification technique.

This paper is organized as follows: Section 1 describes the introduction. Section 2 gives a short review of the research method. Section 3 explains the result and discussion. Section 4 contains the conclusion.

2. RESEARCH METHOD

The system architecture of the prediction of addiction to drugs and alcohol is as demonstrated in Figure 1. Here a user has to answer the questions through a web application. The information collected from the user will go to the server and from there to the expert system. The outcome will be determined based on the input received by applying a logistic regression algorithm on the processed data. A definite result will be prepared in terms of the output obtained from the model. The results obtained through specific formatting can be viewed through the web application. We have collected 510 data of both addicted and non-addicted people among them 80% has been treated as training data and 20% as test data. Our data collection and data preprocessing techniques' layout will be shown in the next section. We have used nine machine-learning algorithms mentioned earlier. We have calculated the accuracy three times. The first time accuracy was calculated before using principal component analysis (PCA) on the processed data, and then the second time it was calculated after using PCA and finally, the accuracies were calculated using the algorithm on the unprocessed data. We have evaluated the classifiers based on accuracy and other metrics like sensitivity, specificity, precision, recall, and F_1-score. These working processes have been described in the following flow diagram in Figure 2.

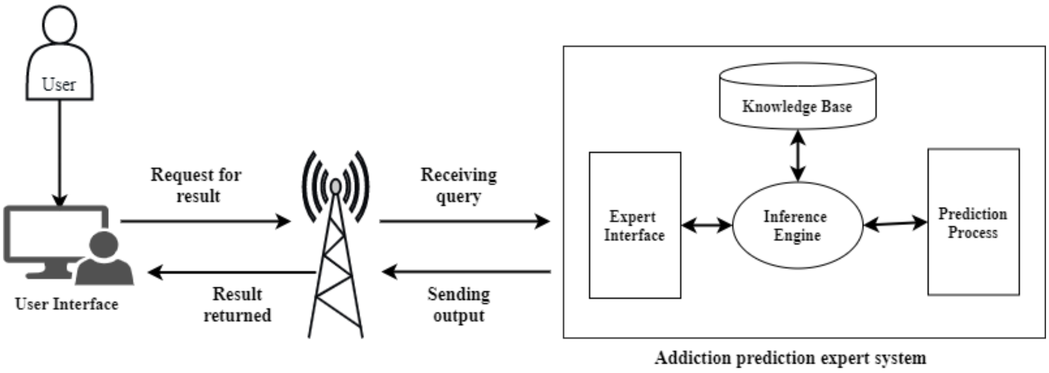

Figure 1. The system architecture of the prediction of addiction to drugs and alcohol

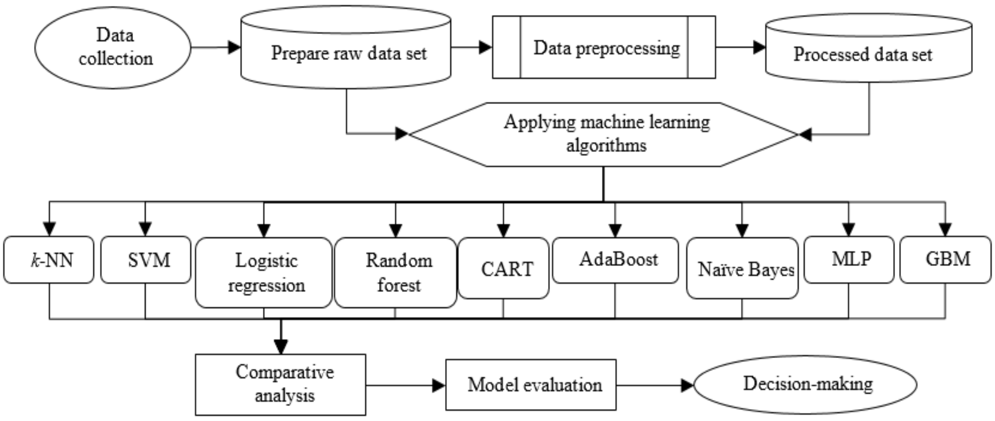

Figure 2. The methodology applied for predicting the addiction to drugs and alcohol

We have run nine machine-learning algorithms on processed data set where the number of features was 23. Then we have used the PCA that is a kind of feature extraction method to grab the underlying variance of data in orthogonal linear projections. The independent used variable of a model is known as the dimensionality of that model. The number of variables can be reduced using a PCA; only the important variables were selected for the next task. Figure 3 has shown the scree plot where a variance is explained in the y-axis and number features showed in the x-axis. Using the scree plot and 90% variance explained as a threshold, we have calculated our principal component number and the number is 14. Normally it combines highly correlated variables to build up a short artificial set of variables [23].

k-NN is a simple supervised machine-learning algorithm. k-NN algorithm grabs similar things that exist in a close neighborhood [24]. Minkowski distance between the query points to other points is determined by using (1).

$$\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q} \tag{1}$$

Support vector machine is a supervised machine-learning algorithm. Data items are placed in n-dimensional space and the values of the features present the particular coordinate [24]. SVM builds a maximum margin separator, which is used for making decision boundaries with the largest possible distance. W is for weight vector and X is for is the set of points. By using (2), we can find out the separator.

$$W.X + b = 0 \tag{2}$$

Logistic regression uses logistic function and this logistic function serves as a sigmoid function. An *s*-shaped curve takes the real values and put them between 0 to 1 [24]. The logistic function is given as (3):

$$f(x) = \frac{1}{1+e^{-x}} \tag{3}$$

Naïve Bayes is one of the oldest algorithms of machine learning. This algorithm is based on Bayes theorem and basic statistics. It extends attributes using Gaussian distribution [23]. The Gaussian distribution with mean and standard deviation is described in (4).

$$p(x = v|C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(v-\mu_k)^2}{2\sigma_k^2}} \tag{4}$$

MLP means multilayer perception. MLP has a combination of multilayer neurons. The first layer is the input layer, the second layer calls the hidden layer, and the third layer is called the output layer. It takes input data through the input layer and gives the output from the output layer [24]. CART is a distribution-free decision tree learning technique. The decision tree is a tree-based model. The divide-and-conquer method is used for making the tree diagram. The Gini index is applied in CART where Gini index finds out the impurity of *D*, *D* represents the training tuples [23]. Gini index is defined as below:

$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^2 \tag{5}$$

Freund and Schapire proposed AdaBoost in 1996. It makes a classifier with a combination of multiple poorly performing classifier. It sets the weight of classifiers and trains the data in each iteration [23]. By using (6), we can compute the error rate of each tuple.

$$error(M_i) = \sum_{j=I}^{d} w_j \times err(\boldsymbol{X}_j) \tag{6}$$

Random forest makes a large collection of de-correlated trees for prediction purposes. It performs split-variable randomization. The random forest has a smaller feature search space at each tree split [23]. Gradient boosting machine builds an ensemble of shallow trees with tree learning and improving technique. GBM works with the principle of boosting weak learners iteratively by shifting focus towards problematic observation. It prepares a stage-wise fashion model like other boosting methods and normalizes them with arbitrary differentiable functions [25].

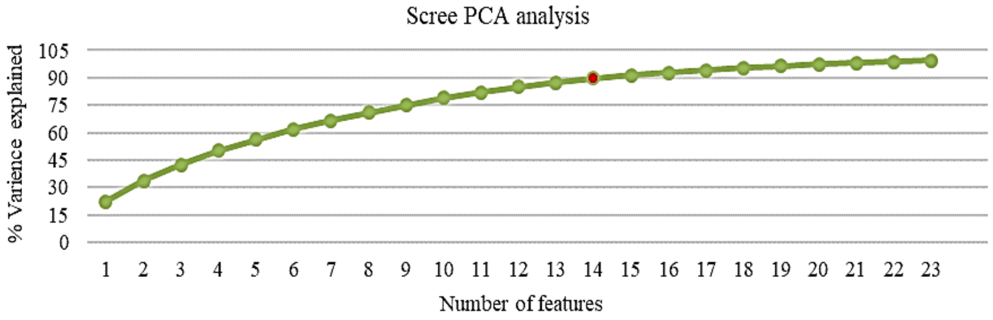

Figure 3. Scree plot where the number of principal components is shown in red color

We not only calculated the accuracy of several algorithms but also calculated sensitivity, specificity, precision, recall, F_1-score, and ROC curve and confusion matrix of each algorithm. In the case of model evolution, certain classifiers have been measured based on the test data set for better measurement. Sensitivity is the true positive rate. It is the ratio of how many positive tuples are correctly diagnosed.

Specificity is the true negative rate. It is the ratio of how many negative tuples are correctly diagnosed. Precision is the measurement of exactness. It is the ratio of true positive value and predicted positive value. The recall is the measurement of completeness. It is the ratio of true positive value and true negative value. F_1-score is the measurement of the harmonic mean of recall and precision. It considers both false positive and false negative values for calculation [23]. The confusion matrix is one of the most important performances of measurement techniques for machine learning classification. It can perform on the classification models with the set of test data and provide the true positive value and true negative value, false-positive value and false-negative value in a tabular format [23]. A feature set is developed by analyzing the main causes of drug addiction, through which it is possible to identify the person addicted to drugs. The feature lists of drug addiction are shown in Table 1.

$$Sensitivity = \frac{TP}{TP+FN} \times 100\% \tag{7}$$

$$Specificity = \frac{TN}{FP+TN} \times 100\% \tag{8}$$

$$Precision = \frac{TP}{TP+FP} \times 100\% \tag{9}$$

$$Recall = \frac{TP}{TP+FN} \times 100\% \tag{10}$$

$$F_1 \text{ score} = \frac{2\ x\ precision\ x\ recall}{precision+recall} \times 100\% \tag{11}$$

To identify the risk of becoming addicted to drugs we have considered each of these factors. We have found out about these factors by talking to various physicians, websites [26]-[30], and articles. The data set is a large collection of necessary and relatable coordinates that can easily be accessed and changed. We have seen someone around us taking drugs but it was a secret, and at the train station and bus station drug addicts refused to help. Then we have decided to go to the drug addiction center and rehabilitation center. We have also collected information from some private rehabilitation centers and clinics. New Mukti Clinic [31] and Brain and Mind Hospital [32] helped us with the information. In addition to providing information, we can learn from their consultants and doctors about many more important factors. Thus, we were able to collect data of 510 people based on 23 factors. There are 305-drug addicts' information and 205 healthy people's information. We have also collected our data from Daffodil International University, Sylhet Engineering College, Begum Rokeya University, New Mukti Clinic, Brain and Mind Hospital, and some other places. Data collection was the most challenging task for us. Nevertheless, we managed to collect some data where there were some missing data, categorical data, numerical and text data. Then we have decided that through data processing we would make this data suitable for different algorithms. Figure 4 has described our data preprocessing work. First, we started the work of data cleaning.

Table 1. Features for drugs and alcohol prediction

Feature Name	Evidence Based-on	Feature Name	Evidence Based-on	Feature Name	Evidence Based-on
Living with family	[27]	Stay alone	[28]	Stay outside at night	[30]
Interest in normal activities	[27]	How much you care about yourself	[27]	Think that drug addiction can be a solution	[28]
Age	[29]	Job losing	[30]	Losing weight	[27]
Residing address	[30]	Sexual harassment	[27]	Have addicted friends	[29]
Profession	[30]	Gender	[29]	Reason to become addicted	[29]
Distance with friends and family	[28]	Having odd sleep pattern	[30]	An addicted person at home	[26]
Working efficiency	[29]	Faced any trauma	[26]	Stress controlling skills	[30]
Relationship problem	[26]	Economic status	[29]		

We have checked if there is a null value in the data set. We have then encoded the level that converts the text data to numerical data. We have solved the missing value problem using the imputer and median. Then we have checked if there is a noisy value in the data set using a box plot. We have found some noisy values in our data set. Our data set had a noisy value in the 'age' feature and we solved the noisy value problem with an outlier quantile. Then we have dropped our outcome feature, that was, the addicted column. We finally have the processed data set in our hands.

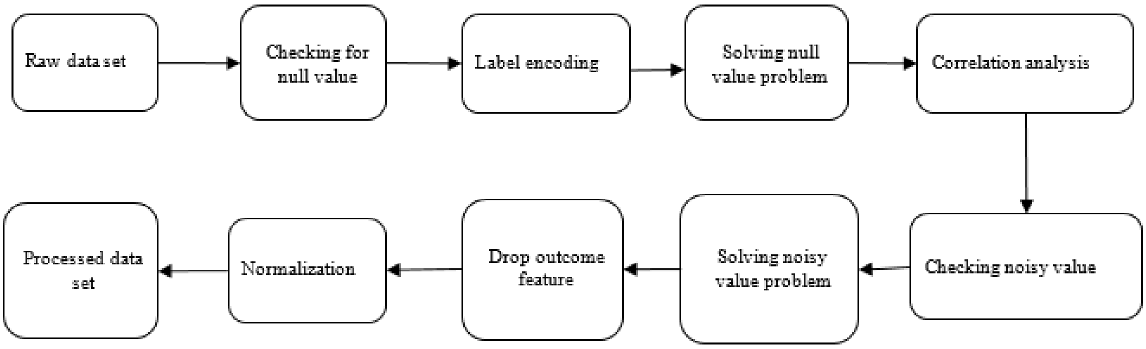

Figure 4. Steps of data preprocessing of gathered data

3. RESULTS AND DISCUSSION

This section will discuss the results of our research work in detail. For ease of understanding, we will present our work data with the help of some graphs and tables in two sections. Here we will provide a brief comparison with the work of others as well.

3.1. Experimental evaluation

A data set is prepared by gathering 510 peoples' data. The statistics have shown that 209 people are addicted because of their friends and 98 people are addicted to drugs for curiosity. Table 2 shows the correlation between the features. Data are highly connected by a positive value and the negative value means that the data is negatively connected and zero indicates that the data does not connect to itself. Besides this, it also shows us how the features correlated with the outcome.

Table 3 describes the performance of each algorithm. It reviews the performance of algorithms sensitivity, specificity, recall, precision, and F_1-score. Based on this performance of algorithms, we would determine which algorithm will fit best for our problem domain. It has been shown that logistic regression performs the best based on accuracy. Again, based on sensitivity, specificity, recall, precision, the CART performs better. However, after performing unprocessed data and PCA, the CART's performance was not good. So, considering everything, the best performance of the model was found using a logistic regression algorithm. We have used nine algorithms here. Each algorithm uses certain parameters and the value of these parameters varies. The parameter values of all the algorithms for training the model are discussed in Table 4. The values given here we selected their optimal value by experiment.

Table 2. Correlation between other features with outcome feature

Features	Correlation Values	Features	Correlation Values	Features	Correlation Values
Have an addicted friend	0.620413	Job losing	0.148141	Economic status	0.219804
Stay outside at night	0.494180	Lives with family	0.449630	Gender	0.409784
Amount of caring about oneself	-0.178059	Having odd sleep pattern	0.094546	Faced any trauma	0.301965
Having a relationship problem	0.257227	Reason to become addicted	-0.882967	Working efficiency	-0.126149
Drug addiction could be a solution	0.392257	Stress controlling skills	-0.217813	An addicted person at home	-0.013045
Distance with friends and family	0.356072	Interest in normal activities	0.352754	Sexual harassment	-0.063114
Age	0.322807	Stay alone	-0.074514	Losing weight	0.321901
Profession	-0.456458	Living address	-0.217732		

It appears that before using PCA, *k*-NN has achieved 96.8% accuracy, SVM has achieved 93.75% accuracy, logistic regression has achieved 84.37% accuracy, naïve Bayes has achieved 87.5% accuracy, the random forest has achieved 66.67% accuracy, CART has achieved 50% accuracy, AdaBoost has achieved 69.79% accuracy, MLP has achieved 78.13% accuracy, GBM has achieved 73.96% accuracy. After using PCA, we can see that the accuracy of some algorithm has increased, some has decreased, and some algorithm has remained unchanged. *k*-NN has achieved 82.29% accuracy, SVM has achieved 95.83% accuracy, logistic regression has achieved 97.91% accuracy, naïve Bayes has achieved 92.7% accuracy, the random forest has achieved 73.95% accuracy, CART has achieved 59.37% accuracy, AdaBoost has achieved 71.87% accuracy, MLP has achieved 72.91% accuracy and GBM has achieved 59.38% accuracy. The difference in the accuracy of the algorithm has obtained before and after the use of PCA is shown in Figure 5.

Table 3. Classifier performance evaluation

Algorithms	Accuracy	Sensitivity	Specificity	Precision	Recall	F_1-score
k-NN	82.29%	95.80%	97.90%	95.91%	97.91%	96.90%
SVM	95.83%	91.66%	95.83%	92.0%	95.83%	93.87%
Logistic regression	97.91%	91.66%	77.08%	90.24%	77.08%	83.14%
Naïve Bayes	92.70%	91.66%	83.33%	90.90%	83.33%	86.95%
Random forest	73.95%	52.08%	81.25%	62.90%	81.25%	70.90%
CART	59.37%	100.0%	100.0%	100.0%	100.0%	100.0%
AdaBoost	71.87%	95.83%	43.75%	91.30%	43.75%	59.15%
MLP	72.91%	91.66%	64.58%	88.57%	64.58%	74.69%
GBM	59.38%	68.75%	79.17%	71.69%	79.16%	75.24%

Table 4. Detailed specifications of the algorithms used

Algorithm	Specifications
k-NN	Number of neighbors = 1 Weight function used for prediction, weights = *c*, where *c* is a constant Power parameter, *p* = 2 Distance metric: Minkowski distance = $(\sum_{i=1}^{k}(\lvert x_i - y_i\rvert)^q)^{\frac{1}{q}}$
SVM	C = 1.0 Kernel: radial basis function = $exp(-\gamma \lVert x - x_n \rVert^2)$ Gamma: scale = $\frac{1}{number\ of\ features \times X.var()}$
Logistic regression	Penalty = *l*2 C = 1.0 Number of random states = 0 Maximum number of iterations = 100
Naïve Bayes	Distribution: Gaussian distribution = $f(x,\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}e^{\frac{-(x-\mu)^2}{2\sigma^2}}$ Mean, $\mu_y = \frac{1}{N}\sum_i x^{(i)}$ Variance, $\sigma_y = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}$
CART	Distribution measure: Gini index, $Gini(D) = 1 - \sum_{i=1}^{m} p_i^2$ Maximum depth = 0 Minimum samples split = 2
AdaBoost	Number of estimators = 100 Learning rate = 1.0 Number of random states = 102
Random forest	Number of estimators = 100 Maximum depth = 2 Number of random states = 0
MLP	Alpha = 0.0001 Network architecture: 23-5-2-1 Number of random states = 94
GBM	Number of estimators = 2 Learning rate = 0.15 Maximum depth = 5

We have also calculated the accuracy with an unprocessed data set. *k*-NN has achieved 81.37% accuracy, SVM has achieved 59.01% accuracy, logistic regression has achieved 58.82% accuracy, naïve Bayes has achieved 57.84% accuracy, the random forest has achieved 73.52% accuracy, CART has achieved 57.84% accuracy, AdaBoost has achieved 71.56% accuracy, MLP has achieved 58.82% accuracy and GBM has achieved 73.52% accuracy with the unprocessed data set.

3.2. Comparative analysis of result

To evaluate the goodness of our proposed addiction prediction system, we need to compare our work with some recent and relevant research works. We should take it into account that the presumption adopted by the researchers in collecting samples and reporting results of their research activities in processing those samples will have an intense indication of our endeavor for comparative performance evaluation. We have strived to compare our work with the other's based on some of the parameters like sample size, size of feature set, algorithm, and accuracy. Table 5 shows a comparative overview of other works and our work.

Zhang *et al*. [13] performed a prediction on daily smoking behavior with five features after collecting data from 15,095 people. Zhu *et al*. [15] worked on tobacco disease detection with 180 hyperspectral images with 32 features. In paper [16], prediction of HIV prognosis and mortality with smoking-associated DNA was done with roughly 0.78 AUC. Prediction on the smoking status by collecting patients' blood tests and health associated vital readings was done in [18]. Lee *et al.* [19] predicted alcohol use disorder by checking the treatment-seeking status of patients and they did not mention the accuracy of their work. In the paper [20] also, prediction on the risk of alcohol use disorder with different types of data were done yet they did not mention the classifier and accuracy. Prediction on alcohol abuse with ANN was seen in the work performed by Kumari *et al*. [21], and it showed an accuracy of 98.7%. Concerning the overall picture depicted in this section, our attained accuracy of more than 97.91%. has turned out to be good as well as promising enough. The reason behind our proposed solution to achieve a very high accuracy is that the features deployed are computationally simple to calculate and have very high discriminatory information to predict the risk of becoming addicted to drugs. As we have mentioned before, most of the other works are not very close to ours. So it would not be wise to explicitly compare the worthiness of our approach with other works.

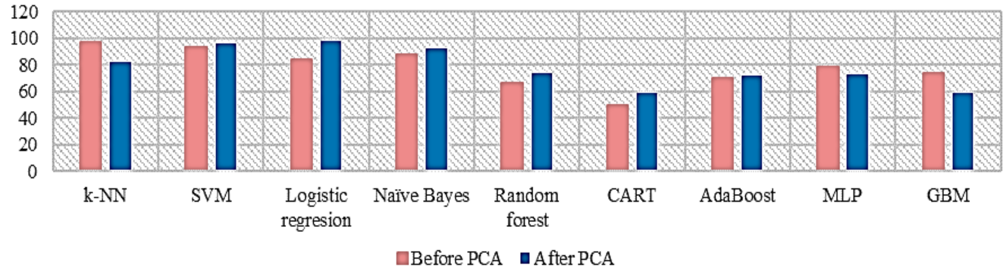

Figure 5. Comparison of accuracy between before and after PCA

Table 5. Results of the comparison of our work and other works

Method/Work Done	Addiction Dealt with	Problem Domain	Sample Size	Size of Feature Set	Algorithm	Accuracy
This work	Drugs and alcohol (risk)	Prediction	510	23	Logistic regression	97.91%
Zhang *et al.* [13]	Smoking behavior	Prediction	15095	5	XGboost	84.11%
Zhu *et al.* [15]	Tobacco diseases	Detection	180	32	ELM	98.3%
Zhang *et al.* [16]	HIV prognosis with smoking-associated DNA	Prediction	1137	698	GLMNET	0.78 AUC
Frank *et al.* [18]	Smoking status	Prediction	534	3	Logistic regression	83.44%
Lee *et al.* [19]	Alcohol use disorder (treatment seeking)	Prediction	778	10	Logistic regression	*NM*[1]
Kinreich *et al.* [20]	Alcohol use disorder (risk)	Prediction	656	3	*NM*[1]	*NM*[1]
Kumari *et al.* [21]	Alcohol abuse	Prediction	1885	12	ANN	98.7%

[1] *NM*: Not Mentioned.

4. CONCLUSION

In this paper, we have performed an in-depth exploratory work for predicting the risk of becoming addicted to drugs and alcohol using different machine learning techniques. First, we have formed the basis, i.e. feature set for this predictive work after talking to doctors and drugs-and-alcohol-addicted people and reading different articles and write-ups. Data have been collected and thoroughly preprocessed. The prediction of risk for addiction to drugs and alcohol has been accomplished with nine conspicuous classifiers. The merits of those classifiers have been measured in terms of six conspicuous performance metrics. The relative merits of the results achieved have been assessed by analyzing the results of similar works thereafter. We have achieved an accuracy of 97.91% with logistic regression classifier, which is good as well as promising. There remains a potential future work with a very large set of addicted and non-addicted people's data to cover an as much wider range of addicted and non-addicted people as required for Bangladesh.

ACKNOWLEDGEMENTS

Authors would like to thank all the individuals and organizations who have helped the authors by providing data during data collection. As well as many thanks to the reviewers for providing valuable and constructive reviews.

REFERENCES

[1] Control of Drug Abuse is a Must, [Online]. Avaible: https://www.thedailystar.net/health/health-alert/control-drug-abuse-must-1515874.

[2] M. N. Shazzad, S. J Abdal, M. S. M. Majumder, J. Ul Alam Sohel, S. M. M. Ali, and S. Ahmed, “Drug Addiction in Bangladesh and its Effect,” in *Medicine Today*, vol. 25, no. 2, pp. 84-89, 2014, doi: 10.3329/medtoday.v25i2.17927.

[3] Restricted, she killed parents, [Online]. Avaible: https://www.thedailystar.net/news/restricted-she-killed-parents parents.

[4] 43% of the unemployed population addicted to drugs, [Online]. Avaible: https:// www.dhakatribune.com/bangladesh/dhaka/2019/02/27/43-of-unemployed-population-addicted-to-drugs.

[5] J. A. Cruz and D. S. Wishart, “Applications of Machine Learning in Cancer Prediction and Prognosis,” in *Cancer Informatics*, vol. pp. 59-77, 2006, doi: 10.1177/117693510600200030.

[6] C. Catal and B. Diri, “A systematic review of software fault prediction studies,” *Expert Systems with Applications*, vol. 36, no. 4, pp. 7346-7354, 2009, doi: 10.1016/j.eswa.2008.10.027.

[7] V. B. Kumar, S. S. Kumar and V. Saboo, "Dermatological disease detection using image processing and machine learning," *2016 Third International Conference on Artificial Intelligence and Pattern Recognition (AIPR)*, Lodz, Poland, 2016, pp. 1-6, doi: 10.1109/ICAIPR.2016.7585217.

[8] E. W. Steyerberg, T. V. D. Ploeg, and B. V. Calster, “Risk prediction with machine learning and regression methods,” in *Biometrical Journal,* vol. 56, no. 4, pp. 601-606, 2014, doi: 10.1002/bimj.201300297.

[9] D. Dahiwade, G. Patle and E. Meshram, "Designing Disease Prediction Model Using Machine Learning Approach," *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC),* Erode, India, 2019, pp. 1211-1215, doi: 10.1109/ICCMC.2019.8819782.

[10] O. Hegazy, O. S. Soliman, and M. A. Salam, “A Machine Learning Model for Stock Market Prediction,” *International Journal of Computer Science and Telecommunications*, vol. 4, no. 12, pp. 17-23, 2013.

[11] L. M. B. Alonzo, F. B. Chioson, H. S. Co, N. T. Bugtai and R. G. Baldovino, "A Machine Learning Approach for Coconut Sugar Quality Assessment and Prediction," *2018 IEEE 10th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM)*, Baguio City, Philippines, 2018, pp. 1-4, doi: 10.1109/HNICEM.2018.8666315..

[12] A. H. Haghiabi, A. H. Nasrolahi, and A. Parsaie, “Water quality prediction using machine learning methods,” *Water Quality Research Journal,* vol. 53, no. 1, pp. 3-13, 2018, doi: 10.2166/wqrj.2018.025.

[13] Y. Zhang, J. Liu, Z. Zhang and J. Huang, "Prediction of Daily Smoking Behavior Based on Decision Tree Machine Learning Algorithm," *2019 IEEE 9th International Conference on Electronics Information and Emergency Communication (ICEIEC)*, Beijing, China, 2019, pp. 330-333, doi: 10.1109/ICEIEC.2019.8784698.

[14] A. M. Alaa, T. Bolton, E. D. Angelantonio, J. H. F. Rudd, and M. V. D. Schaar, “Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants,” in *PLOS ONE*, vol. 14, no. 5, 2019, Art. No. e0213653, doi: 10.1371/journal.pone.0213653.

[15] H. Zhu, B. Chu, C. Zhang, F. Liu, L. Jiang, and Y. He, “Hyperspectral Imaging for Presymptomatic Detection of Tobacco Disease with Successive Projections Algorithm and Machine-learning Classifiers,” *Scientific Reports,* vol. 7, no. 1, pp. 1-12, 2017, doi: 10.1038/s41598-017-04501-2.

[16] X. Zhang *et al.*, “Machine learning selected smoking-associated DNA methylation signatures that predict HIV prognosis and mortality,” *Clinical Epigenetics,* vol. 10, no. 1, pp. 1-15, 2018, doi: 10.1186/s13148-018-0591-z.

[17] M. A. F. Granero, D. S. Morillo, M A. L. gordo, and A. Leon, "A Machine Learning Approach to Prediction of Exacerbations of Chronic Obstructive Pulmonary Disease,” in *Artificial Computation in Biology and Medicine. IWINAC 2015, Springer,* pp. 305-311, 2015, doi: 10.1007/978-3-319-18914-7_32.

[18] C. Frank, A. Habach, and R. Seetan, "Predicting Smoking Status Using Machine Learning Algorithms and Statistical Analysis," *Advances in Science, Technology and Engineering Systems Journal*, vol. 33, no. 3, pp. 184-189, 2018, doi: 10.5555/3144687.3144703.

[19] M. R. Lee, V. Sankar, A. Hammer, W. G. Kennedy, J. J. Barb, McQueen *et al.*, “Using Machine Learning to Classify Individuals with Alcohol Use Disorder Based on Treatment Seeking Status,” *EClinicalMedicine,* vol. 12, pp. 70-78, 2019, doi: 10.1016/j.eclinm.2019.05.008.

[20] S. Kinreich, J. L. Meyers, A. Maron-Katz, C. Kamarajan, A. K. Pandey, D. B. Chorlian *et al.*, “Predicting risk for Alcohol Use Disorder using longitudinal data with multimodal biomarkers and family history: a machine learning study,” *Molecular Psychiatry,* vol. 26, pp. 1133-1141, 2021, doi: 10.1038/s41380-019-0534-x.

[21] D. Kumari, S. Kilam, P. Nath, and A. Swerapadma, “Prediction of alcohol abused individuals using artificial neural network,” *International Journal of Information Technology,* vol. 10, no. 2, pp. 233-237, 2018, doi: 10.1007/s41870-018-0094-3.

[22] M. T. Habib, A. Majumber, R. N. Nandi, F. Ahmed, and M. S. Uddin, “A Comparative Study of Classifiers in the Context of Papaya Disease Recognition,” in *Proceedings of International Joint Conference on Computational Intelligence. Algorithms for Intelligent Systems,* Springer, 2020, pp. 417-429, doi: 10.1007/978-981-13-7564-4_36.

[23] J. Han, M. Kember, and J. Pei, “Data Mining Concept and Technique,” *3rd Edition, Morgan Kaufmann*, pp. 332-398, 2012.

[24] S. J. Russell and P. Norvig, “Artificial Intelligence a Modern Approach,” *3rd Edition, Upper Saddle River, NJ: Prentice Hall*, pp. 725-744, 2001.

[25] H. Jerome Friedman, "Greedy function approximation: a gradient boosting machine," *The Annals of Statistics*, vol. 29, no.5, pp. 1189-1232, 2001.

[26] What Is Drug Addiction, [Online]. Avaible: https://www.webmd.com/mental-health/addiction/drug-abuse-addiction#2.

[27] Teen Drug Abuse and Recovery, [Online]. Avaible: https://www.nextgenerationvillage.com/drugs/.

[28] 10 Reasons Why People Abuse Drugs, [Online]. Avaible: https://www.recoveryconnection.com/10-reasons-people-abuse-drugs/.

[29] What is Drug Addiction, [Online]. Avaible: https://www.healthyplace.com/addictions/drug-addiction/what-is-drug-addiction-drug-addiction-information.

[30] The Causes and Effects of Drug Addiction, [Online]. Avaible: https://www.altamirarecovery.com/drug-addiction/causes-effects-drug-addiction/.

[31] New Mukti Clinic, [Online]. Avaible: https://www.newmukti.com/.

[32] Brain and Mind Hospital, [Online]. Avaible: https://brainandmindhospital.com/.

BIOGRAPHIES OF AUTHORS

Md. Ariful Islam Arif obtained his B.Sc. degree in Computer Science and Engineering from Daffodil International University, Dhaka, Bangladesh in 2020. His research interest includes artificial intelligence, computer vision, computer networking, and machine learning.

Saiful Islam Sany completed B.Sc. in Computer Science and Engineering from Daffodil International University, Dhaka, Bangladesh in 2020. His research interest is in Computer Networks, Network Security, and Web Development.

Md. Sadekur Rahman obtained his B.Sc. and M.Sc. degree in Applied Mathematics & Informatics from the Peoples' Friendship University of Russia. Now he is working as a Senior Lecturer at the Department of Computer Science and Engineering in Daffodil International University. He has a number of publications in international and national journals and conference proceedings. His research interest includes Data Mining, Artificial Intelligence, Pattern Recognition, and Natural Language Processing.

Farah Sharmin obtained her B.Sc. and M.Sc. degree in Computer Science and Engineering from University of Dhaka, Bangladesh. Now she is working as a Senior Lecturer at the Department of Computer Science and Engineering in Daffodil International University. She has a number of publications in international and national journals and conference proceedings. Her research interest includes Artificial Intelligence, Reversible logic, VLSI circuit design, Quantum computing.

Md. Tarek Habib is continuing his Ph.D. degree at the Department of Computer Science and Engineering in Jahangirnagar University. He obtained his M.S. degree in Computer Science and Engineering (Major in Intelligent Systems Engineering) and B.Sc. degree in Computer Science from North South University in 2009 and BRAC University in 2006, respectively. Now he is an Assistant Professor at the Department of Computer Science and Engineering in Daffodil International University. He is much fond of research. He has had a number of publications in international and national journals and conference proceedings. His research interest is in Artificial Intelligence, especially Artificial Neural Networks, Pattern Recognition, Computer Vision and Natural Language Processing.