Drug Addiction Prediction System by Machine Learning Techniques: A Case Study

[1]Dr.M.Deepa,[2]Sri Ranjani.S,[3]Sowmiya.V, [4]Tamizhan.E, [5]Venkat Vijay.M.P

[1]Associate Professor, [2,3,4,5]Student
[1,2,3,4,5]Department of Computer Science & Engineering, Paavai College of Engineering, Namakkal, India

Abstract: Today's youth in society, as well as the population of Tamil Nadu, face a serious threat from drug and alcohol addiction. Therefore, as responsible members of society, we must act to shield these impressionable minds from potentially fatal addiction. In this article, we take a machine learning-based approach to predicting the likelihood of developing a drug addiction. First, by speaking with doctors, drug addicts, and reading pertinent publications and write-ups, we identify several key causes of addiction.Next, we gather information from both addicted and non-addicted individuals. We apply nine notable machine learning algorithms—k-nearest neighbors, logistic regression, SVM, nave bayes, classification and regression trees, random forest, multilayer perception, adaptive boosting, and gradient boosting machine—on the preprocessed data set. We then assess how well each of these classifiers performs in terms of some key performance metrics. By achieving an accuracy close to 95.01%, logistic regression is determined to surpass all other classifiers in terms of all measures. The findings of CART, on the other hand, are subpar, with an accuracy of about 50.37% after using principal component analysis.

Index Terms: Addiction Drugs and alcohol, Logistic regression , Machine learning, Prediction system.

I. Introduction

Every nation in the globe needs to be concerned about the issue of drug addiction. Every nation deals with this issue differently and to varying degrees. It is connected to social and familial norms and conduct. It harms a person's physical and emotional health as well. Alcohol consumption has been linked to increased aggression toward others in both men and women, for instance [1]. A person's decision to start smoking is apparently influenced by their relationships with friends and family [2]. These demonstrate how social and behavioral issues are linked to and influenced by drugs in some way. Yes, drugs are very bad for your health, but they may also ruin your personal and professional life. These demonstrate how social and behavioral issues are linked to and influenced by drugs in some way. Drugs are very bad for your health, but they may also ruin your personal and professional life. We can determine whether a person is connected to drug abuse or not by observing their daily social, familial, and health issues. As his or her social activities, various consequences of day-to-day life with people, as well as health issues, may potentially indicate his or her openness to different types of drugs. The root cause of this addiction is dissatisfaction. Political unrest, family disconnection, lack of adoration friendship, and joblessness issues all contribute to disappointment. We must abstain from using drugs if we want to prevent addiction. Avoiding drugs simply lowers the chance of developing a drug addiction before it develops. Today's youthful generation, from all walks of life, is discreetly impacted by the dreadful reality that drug addiction has become. Drug-dependent Oishee Rahman killed her parents in 2015 [3].

The friendship circle might easily be destroyed by an addictive companion. Around 7.5 million individuals in Tamilnadu are drug addicts, according to the Dhaka Tribune newspaper. They are dangerous since 50% of them are involved in various criminal activities and 80% of them are young people [4]. In order to prevent our young from developing a drug addiction, we must maintain a specific emphasis. The issue raised above may have a solution thanks to machine learning, a key component of artificial intelligence (AI). Machine learning is used in a variety of application domains, such as risk prediction [8], cancer prediction [5], software error prediction [6], dermatological disease detection [7], and others.

This essay aims to foresee the possibility of someone developing a drug or alcohol addiction. We started by reading pertinent publications from various national and international journals, conference proceedings, magazines, and newspaper and internet pieces. Then, after speaking with medical professionals and drug and alcohol addicts, we identify certain risk factors for addiction, including age, gender, career, health status, mental stress, trauma, family and friend history, and life-altering events. obtaining unprocessed data from both addicted and non-dependent individuals. Despite the fact that there hasn't been any work that specifically addresses the issue of predicting drug and alcohol addiction, we made a valiant effort to compare our findings with those of other research projects that looked in a similar direction.

II. LITERATURE REVIEW

We have followed and examined related works on drugs and addiction done in the recent past by some other researchers, and we comprehend the procedures and approaches they stated. Here are some summaries of current significant machine learning research. A generic disease prediction system based on machine learning methods was proposed by Dahiwade et al. [9]. A model for stock market forecasting using machine learning technologies was put forth by Hegazy et al. [10]. Alonzo et al. [11] provided a thorough comparison of various machine learning methods used for coconut sugar quality prediction and assessment. Using a machine learning method, Haghiabi et al. [12] worked on forecasting water quality. Based on a machine-learning algorithm, Zhang et almethod .'s [13] for forecasting daily smoking behaviour was put forth. The best accuracy was 84.11% with a maximum depth of five using the extreme gradient boosting (XGBoost) decision tree technique. A machine attempting to learn strategy for predicting cardiovascular disease risk in Bioscience participants was proposed by Alaa et al. [14]. The pre-symptomatic detection of tobacco disease using hyperspectral images and machine-learning classifiers was the focus of Zhu et al[15] .'s research.

By using smoking-associated deoxyribonucleic acid (DNA) and machine learning classifiers, Zhang et al. [16] attempted to predict the prognosis and mortality of human immunodeficiency viruses (HIV). Using machine learning characteristics, Granero et al. [17] created a model for anticipating obstructive lung disease exacerbations. With the use of statistical analysis and machine learning, Frank, Habach, and Seetan [18] worked on predicting smoking status. In their investigation, logistic regression performed the best, with 83.44% accuracy, 83% precision, 83.4% recall, and 83.2% F-measure. By examining the treatment-seeking status with a machine learning classifier, Lee et almodel .'s [19] predicts alcohol consumption disorder. Cognitive, mood, impulsivity, personality, aggressiveness, early-life stress, and childhood trauma were the data collection domains.

A model for estimating the risk of alcohol use disorder (AUD) utilising machine learning technologies was put out by Kinreich et al. [20]. A model for predicting alcohol misuse using machine learning technologies was put out by Kumari et al. [21]. They evaluated the characteristics of their models to include age, gender, nationality, ethnicity, education, neuroticism, openness to experience, extraversion, agreeableness, conscientiousness, impulsive, and sensation seeing. These characteristics were taken into account in ANN-D, as well as day, week, month, year, and decade. Based on a machine learning classification approach, C. Habib et al. [22] studied the identification of papaya illness.

The structure of this essay is as follows: The introduction is described in Section 1. An overview of the past literature is provided in Section 2. The outcome and discussion are explained in Section 3. The conclusion is found in Section 4.

III. RESULTS AND DISCUSSION

We will go into great depth about the findings of our research in this part. We will divide our work data into two areas for convenience of comprehension and show it with the aid of some graphs and tables. Here, we'll also give a quick comparison to some other authors' work.

Data from 510 individuals are collected to create a data set. According to studies, 98 individuals become hooked to drugs because of curiosity, while 209 people become addicted because of their friends. The relationships between the characteristics are shown in Table 1. Positive values show that the data are strongly related, while negative values denote that the data are weakly connected and zero denotes that the data are not connected to one another. In addition, it demonstrates the correlation between the attributes and the result. Each algorithm's performance is described in Table 2.

Table 1. 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		

The effectiveness of the algorithms' sensitivity, specificity, recall, accuracy, and F1-score are examined. We would choose the method that best fits our issue area based on how well each performed. Again, the CART performs better based on sensitivity, specificity, recall, and accuracy. However, the CART's performance was poor when unprocessed data and PCA were applied. So, taking everything into account, a logistic regression approach was used to find the model's optimum performance. Here, nine algorithms have been applied. Each algorithm makes use of specific parameters, each with a different value. Table 3 discusses the parameter values for each approach used to train the model. The values shown here were determined through experimentation to be at their best.

Table 2. Classifier performance evaluation

Algorithms	Accuracy	Sensitivity	Specificity	Precision	Recall	*F1-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%

Prior to using PCA, it appears that k-NN has achieved 96.8% accuracy, SVM has achieved 93.75% accuracy, logistic regression has achieved 84.37% accuracy, naive Bayes has achieved 87.5% accuracy, random forest has achieved 66.67% accuracy, CART has achieved 50% accuracy, AdaBoost has achieved 69.79% accuracy, MLP has achieved 78.13% accuracy, and GBM has achieved 73.96% accuracy. We can observe that some algorithms' accuracy has grown, some algorithms' accuracy has dropped, and some algorithms' accuracy has stayed constant after employing PCA. In addition to k-accuracy NN's of 82.29%, SVM's accuracy of 95.83%, logistic regression's accuracy of 95.01%, and naive Bayes accuracy of 92.7%, The random forest obtained accuracy of 73.95%, CART achieved accuracy of 50.37%, AdaBoost achieved accuracy of 71.87%, MLP earned accuracy of 72.91%, and GBM achieved accuracy of 59.38%. Figure 1 depicts the variation in algorithmic accuracy before and after the use of PCA.

Table 3. 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	95.01%
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	*NM1*
Kinreich *et al.* [20]	Alcohol use disorder (risk)	Prediction	656	3	*NM1*	*NM1*
Kumari *et al.* [21]	Alcohol abuse	Prediction	1885	12	ANN	98.7%

We must compare our study with other current and pertinent studies to see how well our suggested addiction prediction system works. We must keep in mind that the presumption used by the researchers when gathering samples and disclosing the outcomes of their research activities when handling those data will have a significant impact on our effort for comparing performance evaluation. We have made an effort to contrast our work with that of others based on criteria including sample size, selected feature size, algorithm complexity, and accuracy. After gathering data from 15,095 individuals, Zhang et al. [13] conducted a prediction on daily smoking behaviour using five parameters. In order to identify tobacco illness, Zhu et al. [15] used 180 hyper spectral pictures with 32 characteristics. In research [16], HIV prognosis and death were predicted using smoking-associated DNA with an AUC of around 0.78. In [18], smoking status was predicted using patient blood tests and vital signs related to their health. By assessing patients' treatment-seeking behaviour, Lee et al. [19] made a prediction about alcohol use disorder without mentioning the precision of their findings. Additionally, in the study [20], risk of alcohol use disorder was predicted using several types of data, but the classifier was not mentioned.

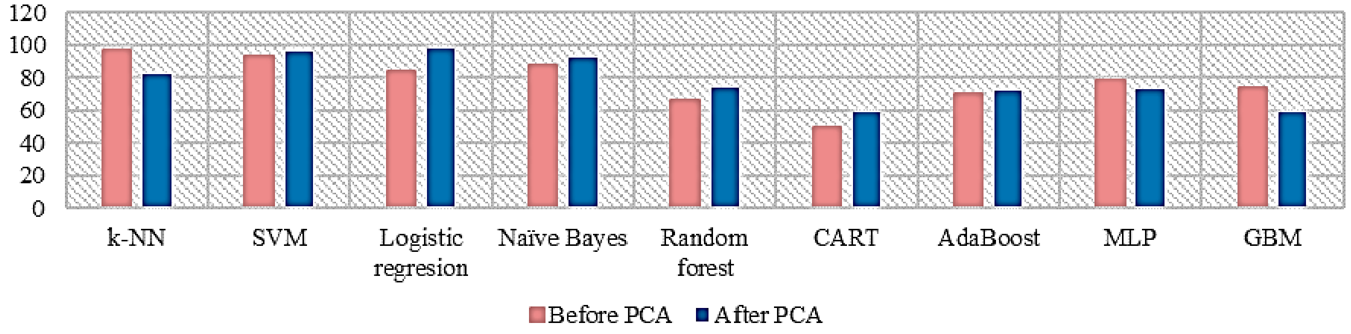

Figure 1: Comparison of accuracy between before and after PCA

IV. CONCLUSION

In this study, we used a variety of machine learning approaches to do a thorough exploratory work for predicting the likelihood of developing a drug or alcohol addiction. First, after speaking with physicians, drug and alcohol addicts, and reading various publications and write-ups, we developed the foundation, or feature set, for this prediction work. Data have been gathered and meticulously prepared. Nine obvious classifiers have been used to predict the likelihood of drug and alcohol addiction. Six glaring performance indicators have been used to evaluate the merits of those classifiers. By examining the outcomes of subsequent identical studies, the relative qualities of the results obtained have been evaluated. With the logistic regression classifier, we were able to attain an accuracy of 95.01%, which is both encouraging and positive. To cover as wide a variety of addicted and non-addicted persons as is necessary for Tamil nadu, there is still a potential future work with a very large set of data on hooked and semi people.

REFERENCES

1. K. Young, K. Gobrogge, Z. Wang, “The role of mesocorticolimbic dopamine in regulating interactions between drugs of abuse and social behavior”, Neuroscience and biobehavioral reviews, vol. 35, no. 3, pp. 498-515, 2011.
2. K. Kobus, ”Peers and adolescent smoking”, Addiction, vol. 98, pp. 37-55, 2003. Available: 10.1046/j.1360-0443.98.s1.4.x.
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:// ww.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: 0.1109/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,
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: 0.1016/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,