Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7759998

APPLICATION OF MACHINE LEARNING ALGORITHMS EFFECTS ON DRUG ABUSE CREATING DRUG ADDICTION BEHAVIOUR THROUGH ANALYSING PERSONALITY TRAITS

D. Kumaresan
Department of Computer and Information Science, Faculty of Science, Annamalai University, Annamalai Nagar, aucsedks@yahoo.co.in

Dr. Aranga. Arivarasan
Department of Computer and Information Science, Faculty of Science, Annamalai University, Annamalai Nagar, arivudatamining@gmail.com

Abstract

Mining good-looking understanding from accrued set of statistics stays constantly exciting. A number of researchers have furnished new offerings to this trouble via massive awesome perspectives. Several human beings don`t recognise how human beings get Addicted to Drug Consumption Risk (ADCR). They misleadingly recognize that individuals who consume drugs include deficiency in moral moralities and self-control. But the actual certainty is that the ADCR is a complicated disease. It additionally makes tough in quit from drug consumption and makes tougher in opposition to precise intents and will power. This is due to the fact the drug make modifications in brain in opposition to not consuming the drug. Modern study methodologies offer beneficial in understanding approximately how drug disturb the mind. It additionally classifies appropriate remedies to get over drug consumption to lead a peaceful life. In our approach the drawback in extraction of insightful understanding from statistics is estimated as crucial task. This study also lookout for competent datamining (DM) strategies to categorise the drug users and non-drug users based totally on drug consumption behaviour. Further targeted evaluation concerning eradication of dependency to drug consumption via analysing the features and appropriate measures are advised via this study. Three of the machine learning (ML) algorithms Decision Tree (DT), Support Vector Machine (SVM) and K-Nearest Neighbor (K-NN) are considered for implementation. From the data set information regarding drug consumption related to nervous function affective psychoactive five drugs (Alcohol, Caffeine, Cannabis, Cocaine, Heroin and Nicotine) was considered. The 70 % of dataset is divided to construct the training set and 30 % is for testing set. Finally, the Accuracy, AOC and the time occupied to construct the model are retrieved as classification results to achieve this research. The DT algorithm produces best result for the drug Alcohol (94.33). The SVM algorithm provide best results for the drugs Caffein (98.76), Cannabis (92.03), Cocaine (96.99), Heroin (97.52) and Nicotine (88.85). Among the five factor features the Nscore, Oscore and Cscore provide better contribution to identify the drug consumption risk. The ROC curve (AUC) for all the drugs [Alcohol (0.96), Caffein (0.93), Cannabis (0.98), Cocaine (0.92), Heroin (0.96) and Nicotine (0.97)] the SVM Classification Model provide the best accuracy. From our model it is observed that among the 12 features Age, Neuroticism,

Sensation Seeking, Oscore and Cscore contribute higher to predict the drug consumption risk. So, more personality trails are to be conducted by considering the contributing features.
Keywords: Drug, Decision Tree, SVM, K_NN, Drug Consumption Risk, AUC.

INTRODUCTION

The learning of drugs or chemicals and the side effects that they have on potential living beings is an essential task. This study provides use full knowledge about drugs. Through this knowledge we can understand the behaviour of body functions towards the drug. a drug is a chemical compound which reacts with the protein in human body [1]. This reaction leads to disrupt all physiological functions of the human. This knowledge leads to identify new medicine. The systemic circulation metabolises the identified chemicals and bind with dependable proteins. This process changes the working of the cell to certain level.

The working of the cell during the consumption of drugs certainly varies among people. This is because the drugs function in a different way among people. But at the same time, it will produce similar end result and side-effects. The side-effects quantitively vary among people. This variation is primarily because of variance in pharmacodynamics among ethnicity, age, genetic makeup and state of disease. Asians are potentially containing higher effectiveness while consuming drugs compared to Africans. The age plays a major role to Eliminate of the drug from our body. The babies and aged persons experience the effects of drugs for longer duration. It also consumes longer duration to be eliminated from the body.

The Drug addiction is considered as one of the chronic brain related sicknesses. It roots a person to take drugs frequently without worrying about the damage they cause. Frequent usage of drug will certainly transform the brain and cause addiction. The people who recovered from using drugs are undoubtedly at risk, if they consume the drugs again.

One important aspect here is that not all the people consume drugs [2] are get addicted. The reactions to the drugs will be varied among the people because of their body condition and brain activity. Few people never get addicted but few get addicted quickly as well some of them over a certain duration. Some people may become addicted quickly, or it may happen over time. The process of getting addicted to drugs involves many aspects. Few of them are genetic, geological, as well situational reasons.

The drug consumption and addiction will certainly be ruled out. This is by conducting awareness programs which include the family members, students from schools, publics, and through advertising in mass media. These actions will educate and extent people to realize danger of drug consumption [3]. In the recent times the machine learning which is a subset of artificial intelligence provide useful contribution in several health care modernizations. The modernization is achieved by advancement in new medical trials, the advancement in accessing the patient detail manipulation and providing medication for chronic diseases. The healthcare industry contains the potential to accommodate modern technological advancements. Machine learning is practical in solving a verity healthcare routine belonging. few of them include prediction and medication to various disease, medical image and diagnostics, Determining and developing new drugs and manipulating medication detail from medical records. Machine learning practice will guide the healthcare specialists to innovate specific medical exploration from accumulated high volumes of data. This achievement can also be personalized to individual physical characteristics.

The machine learning is applied in Medicinal companies to assist drug discovery and drug development. In future ML will possibly guide drugmakers to predict how the patients react to different drugs. During this course of work the drugmakers can categorize among patients who contain the potential to benefit from the identified drugs. Medical experts are facilitated with machine learning techniques contain the potential to simply understand patient histories without having much difficulty and chains of interconnection to various hospital. At a single point of action each patient's appropriate health information can be retrieved through updated medical systems.

Machine learning [4] aided AI apparatuses are working together with drug developers to identify new drug behaviours in quicker way compared with the past. Machine learning algorithms are used to process millions of data points and construct them to help the researches to interpret which components are successful and which are not successful. Machine learning applications has the potential to provide effective drug compounds in shorter duration without exercising much human efforts on each trial. The Machine learning algorithms generate more perfect and immediate disease identification and prediction.

The machine learning algorithms are applied in radiology [5] and pathology departments around the world. They analyse CT and X-RAY images to find solutions to the diseases. Machine learning systems are very much progressive when equipped with a combination of supervised, unsupervised or semi-supervised algorithms. Analysing a volume of diseased images, the ML can identify and distinguish several harmful diseases like cancer and viruses. Machine Learning Models provide assistance to medical experts in understanding and identification of spread of deadly viruses.

RELATED WORK

An exaugurated literature is carried out to perform the determined task. Several research works and survey are being carried out in the field of drug consumption risk prediction. One of the researchers carried out his research work to evaluate the efficiency of machine learning models in predicting daily marijuana use and identify factors associated with daily use among adults using the 2016–2019 Behavioural Risk Factor Surveillance System (BRFSS) Survey carried out in 2020. He utilized Logistic Regression, Decision Tree, and Random Forest with Gini function, and Naïve Bayes to perform his classification. The performance of all his classification models was compared by means of accuracy, AUC, precision, and recall. At the end he suggest that the most important factors for daily marijuana use were the current use of e-cigarette and combustible cigarette use, male gender, unmarried, poor mental health, depression, cognitive decline, abnormal sleep pattern, and high-risk behaviours [6]

The machine learning contains high potential towards effective analysis of different kind of data. There are several sources are available to accumulate the data. Few of them are biological trials, the survey results of longitudinal, consequences treatment, national and international surveys, medical history and information generated through social media. Considering the collected data management ability certainly the computers have the better influence. The ML has established to be exceptionally useful to filter the data efficiently to the healthcare services providers. The researchers utilize the sophistication of various ML algorithms like clustering, learning, classification, trees, fuzzy, neural networks, and time series analysis [31]. Never-the-less the healthcare industry also has consistently tried to come up with the latest approach into

their day-to-day practice. Nowadays in healthcare industries the ML has become extremely popular. The ML is of assistance in healthcare industry to examine the fraud identification and exploitation, customer relationship managing, booming patient care, and more widely in achieving reasonable solution accomplishing healthcare services. [8] One of the biggest setbacks is the larger quantity of data generated through healthcare connections are excessively complex and heterogeneous. Varieties of data source generate significant complication to the ML applications through missing data, inaccurate data, inconsistent data and information accumulated in various format. One of the important aspects considered in health care is that majority of diagnoses and treatments are inaccurate and subject to experimental investigations. Machine-learning (ML) procedures are currently applied to substance use disorder (SUD) data. several predictive applications are developed using this date to prevent the human lives. Identifying frequent drug abuse, measuring future risk by drug addiction and forecasting the available disease cure progression are few of the current requirements to aid the human lies. [7] The Super Learning (SL) practice accelerates the classifiers decision by merging identified prediction algorithms applicable for a particular prediction task. [8].

The researcher in his work provided an artificial neural network (ANN) based approach is proposed for prediction of alcohol user. Two ANN modules are designed, ANN-D to predict a person is an alcohol user or not and ANN-C to predict when it is used. The accuracy of the ANN-D module is found to be 98.7% and ANN-C module is 49.1%. [9]

The sensitive, emotional and public assistance characteristics of a person is exposed by his mental health influenced. The mental health decides how individual personality will think, feel or handle a situation. So, in a research work few implements classification algorithms Decision Tree, Random Forest and Naïve Bayes were applied to predict the mental health of the particular person. [10]. A researcher has developed a model to detect previously unknown Adverse Drug Reactions (ADR) using decision tree and fuzzy logic to generate a decision model. His model is equipped with a fuzzy inference engine, which find the causal relationship between a drug and a potential ADR. [11] Classifying the features of grownups with current drug use is limited by standard statistical methods. It also involves many exclusive methods. The authors [25] evaluate the prevalence and coordination among diagnoses using the ICD-11. The frequency of disaggregated ICD-10 and ICD-11 symptoms and variation in clinical features across diagnostic groups[32].

The educated elders are certainly a thoughtful person regarding the society. It is their responsibility to prevent young minds from life-threatening drug addiction. So, several researchers put together theirs useful hands in this regard. The authors in [26] applied the machine-learning-based forecast of risk of becoming addicted to drugs. Initially they collect data from both addicted and non-addicted people [33]. After pre-processing the data set, they apply nine well known 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 to predict the risk of becoming addicted to drugs [34]. The improvements in the field of data science and the development of many algorithms have extended the base for advancement in several scientific fields and industries. Among them the drug discovery and development require acceptable machine learning algorithms for the invention of several new drug candidates[35][36]. To improve the efficiency and quality of methods for designing useful drug targets the drug discovery and

development process overturned towards machine learning and deep learning techniques.The authors in [27] review, discussed the accompanying techniques in machine learning and deep learning algorithms employed in drug discovery process. They also investigate several applications which produce promising results and methods[37][38].

RESEARCH METHOD

In this research article we have utilized three of the machine learning algorithms. They were decision tree, Support Vector Machine and the k-Nearest Neighbour. [12] The methodology applied in the determined research work is shown in Fig.3.1. among the drugs we choose six drugs for our implementation. They are Alcohol, Caffeine, Cannabis, Cocaine, Heroin and Nicotine [13]. The algorithm implemented in the proposed work as follows:

Step 1: Import the dataset.
Step 2: Explore the data to figure out what they look like.
Step 3: Pre-process the data.
Step 4: Split the data into attributes and labels.
Step 5: Divide the data into training and testing sets.
Step 6: Train the Machine learning algorithm.
Step 7: Make some predictions.
Step 8: Evaluate the results of the algorithm.
Step 9: Analyse ROC curve.

3.1 Decision tree

The decision tree is a supervised non-parametric learning algorithm. It is employed to execute classification as well as regression trial. It includes a root node, many branches, many inner nodes and leaf nodes. The structure of the decision tree [14] will be tree structure with a hierarchy. The decision tree begins with a root node. The root node does not include any inbound branches. The decision nodes are the outgoing branches from the root node.

analysing available features the branch nodes and inner nodes perform calculations construct homogenous subsets. These nodes are identified as leaf or terminal nodes. All feasible outcomes inside the dataset is determined through the leaf nodes. The Decision tree approach gets along with divide and conquer approach [15] associated with greedy search. This is to determine ideal split points inside the tree. The splitting process is performed by top-down approach till the end of all records or major set of records are classified in to a possible class. The algorithm as follows.

Step 1: Begin with training set associated with root node.
Step 2: Use the information gain to label each node.

Step 3: Recursively construct each subtree on the subset of training instances that would be classified down that path in the tree.

Step 4: For all remaining positive and all negative training instances label that node “1” or “0” respectively.

Step 5: For all remaining attributes label them with a majority vote of training instances left at that node.

Step 6: For all remaining instances label them with a majority vote of the parent’s training instances.

Step 7: Estimate the model’s acceptance through cross validation.

The primary task in DT is the determination of the attribute for the root node at each level. This is denoted as attribute selection. We have two popular attribute Information Gain and Gini Index. Entropy changes once every node in a decision tree is partitioned in to minor subsets as training instances. Information gain is level of change in entropy.

S - set of instances

A - attribute

Sv - subset of S with A = v,

and Values (A) is the set of all possible values of A, then

$$Gain\ (S, A) = Entropy(s) - \sum_{veValues(A)} \frac{\mathrm{I}s_v\mathrm{I}}{\mathrm{I}S\mathrm{I}}\ .\ Entropy(S_v)$$

SVM

In machine learning SVM is identified as a supervised learning algorithm. It is applied to analyse data. It can resolve classification and regression tasks. A support vector machine (SVM) executes classification by building an N-dimensional hyperplane which optimally splits the input data into two classes. SVM models are comparably associated with neural networks. In reality, the sigmoid kernel SVM model is comparable to a two-layer, feed-forward NN. The aim of SVM modelling [16] is to identify the ideal hyperplane which distinguish clusters of vectors. It separates one category of target variable on one side of the plane. The other category of target variable on the other size of the plane. The vectors which are very nearer towards the hyperplane are identified as support vectors.

Considering the pattern recolonization problems, the SVM learning algorithm is identified as one of the best applied statistic machine learning techniques [17]. In case of a linearly non-separable data while the same data is nonlinearly separable, the nonlinear SVM is applied. Then kernel function is used instead of an inner product, is to build a nonlinear support vector classification. The SVM classification algorithm is applied to build several types of learning activities through the use of various kernel functions.

One of the important tasks in applying SVM is the identification the optimal hyper-plane. The thumb rule here is that the identified hyper-plane must isolate the labels in a commanding way. Another way of selecting the hyper-plane which contain high margin. Because in case of a hyper-plane comprising low margin can lead towards miss-classification. While applying the SVM the margin between the data points and the hyperplane should be maximized. This is achieved through loss function. After the classification is performed if the obtained cost is 0 it means that the projected value and the true value are within same sign. If it is not the case then the loss value is to be determined through calculation. The regularization parameter is also summed with the cost function. Hear the principle of having the regularization parameter is to generate the similarity among maximization of margin and loss.

1. Input: set of samples of training set data;
2. x1, x2,x3…xn ----- input feature samples,
3. Here we may have lot of input features xn.
4. Output: set of weights w (or wi),
5. One [1] to all features in the dataset.

6. The linear compound project the y value.
7. Central difference: To minimize the number of weights which contain nonzero we apply optimization to maximize the margin. Because this margin is important in finalizing the optimal hyperplane.

$$\left[\left(\frac{1}{n}\right)\sum_{i=1}^{N}\max\left(0,1-\ y_i\ (w.x_i-b)\right)\right]+\ \lambda(w)^2$$

In our approach we concentrate on soft-margin SVM classifier by selecting a considerable minimum value for λ.

3.3 K – Nearest neighbor algorithm

The KNN is a supervised non-parametric learning classifier. To perform classifications or predictions this algorithm performs regionary among categorization from distinct data point [18] It makes an assumption that identical points can be identified next to one another. Because of this this algorithm is utilized as classification algorithm despite the fact that it can be applied to solve both regression and classification tasks. To solve the classification tasks this algorithm, allocate a class label by performing majority voting.

The KNN applies the average the k nearest neighbors to make predictions while applied to solve the regression tasks. The difference here is that the classification task is applied for discrete values. The regression task is applied for continuous values. Before performing any classification, the distance needs to be identified. Usually, the KNN applies Euclidean distance. At the training phase this algorithm stores the entire training set in the memory. So, it requires lot of storage to store all the available data for performing computation. Because of this fact this algorithm is identified as either instance-bases or memory bases algorithm.

To determine the specific class among all the collected data the KNN checks the k number of neighbors. The identification of k value is subject to input data. If the input data contain more outliers and noise higher values of k will certainly perform efficiently. The researcher suggests that if the k value is fixed as odd number, then we can come out of ties in classification. The idea of applying cross-validation will certainly lead to identify optimal k value of the dataset. The algorithm implemented is as follows:

Step 1: Load dataset.
Step 2: prepare the value for k
Step 3: Start the iteration from beginning to total number of training data set to identify the predicted class.
Step 4: Calculate the distance between test data and each row from the training set.
Step 5: with established distance value finalize the calculated distances in ascending order.
Step 6: Obtain the top k rows as of finalized array.
Step 7: Obtain the majority frequent class from existing rows.
Step 8: Retrieve the final predicted class.

K-nearest neighbor algorithm is applied to find the nearest neighbors of identified search point. Then only it is possible to assign the class label to that particular point. To perform this task, it is necessary to identify the distance among the search point and other collected data points. For this the distance metrics help us through decision boundaries. Through this we can

separate the search point into various regions. The researchers propose numerous distance measures. The most commonly use distance measure is that the Euclidean distance. This metric calculates a straight line among the search point and the other measured point. In our research work we applied this distance metric using the below given formula.

3.4 Data set

The dataset contains records for 1885 respondents. Each record has 12 attributes [19] These 12 attributes are listed in Table-1.

1. ID	2. Age
3. Gender	4. Education
5. Country	6. Ethnicity
7. Nscore	8. Escore
9. Oscore	10. Ascore
11. Cscore	12. Impulsive
13. Sensation Seeing (SS)	

Table-1. Feature Names

The five traits can be summarized as:

1. Neuroticism (N) is a long-term tendency to experience negative emotions such as nervousness, tension, anxiety and depression;
2. Extraversion (E) is manifested in outgoing, warm, active, assertive, talkative, cheerful, and in search of stimulation characteristics;
3. Openness to experience (O) is a general appreciation for art, unusual ideas, and imaginative, creative, unconventional, and wide interests,
4. Agreeableness (A) is a dimension of interpersonal relations, characterized by altruism, trust, modesty, kindness, compassion and cooperativeness;
5. Conscientiousness (C) is a tendency to be organized and dependable, trong-willed, persistent, reliable, and efficient.

In the dataset the features Age, Gender, Education, Country and Ethnicity are categorical data. The features Nscore, Escore, Oscore, Ascore, Cscore, Impulsive and Sensation Seeing are numerical data [20] The selected drug usage data is listed under seven categories. They are “Never Used”, “Used over a Decade Ago”, “Used in Last Decade”, “Used in Last Year”, “Used in Last Month”, “Used in Last Week”, and “Used in Last Day”. In the proposed work the data set is converted in to binary class classification. The seven classed were turned into two classes. They are “Drug used” and “Drug never used.

RESULTS

This research article consists of two important aspects. One is the detail about the dataset. The second one is the extracted results through the three machine learning algorithms. The algorithms are Decision Tree [DT], K Nearest Neighbor [KNN] and Support Vector Machine [SVM]. Here we aim at providing the detailed description about the examined research work.

4.1 Experimental evaluation

For our work we have taken all the 12 features for identification of the class. We converted the seven-class problem into binary class problem. So, there is only two classes. the class one is 'drugs never used' and the class two is 'drugs used'. Among the 18 drugs we have chosen only six drugs. They are namely, 1. Alcohol 2. Caffeine 3. Cannabis 4. Cocaine 5. Heroin and 6. Nicotine. The model is validated to count how efficiently it predict the arriving new data. The Table-2 shows the acquired classification results through Decision Tree algorithm evaluating the confusion matric.

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

$$Precision = \frac{TP}{TP + FP}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

	Accuracy	Precision	Sensitivity	AOC	Training time
Alcohol	94.3363	0.949731	0.992495	0.635172	2.18108
Caffeine	95.7522	0.957522	0.9912	0.5	0.122835
Cannabis	83.3628	0.84858	0.853968	0.885835	0.113361
Cocaine	93.4513	0.454545	0.138889	0.831075	0.0631207
Heroin	94.1593	0.625	0.27027	0.79461	0.034823
Nicotine	76.8142	0.796353	0.803681	0.796301	0.0329323

Table-2. Performance Evaluation of Decision Tree Algorithm

4.2 Identifying the motivation for predictive models

Certainly, it is a necessity to find the motivation before training the ML algorithms. The motivation involves identifying the important requirements involving determination of the predictor variables, dataset, and most importantly the contributing feature importance and selection. A range of predictor variables contain direct relationship in finding output categories. In our experiment the features Nscore, Escore, Oscore, Ascore, Cscore, Impulsive and Sensation Seeing contain the direct relationship. This is tabulated in Table-5. The remaining features contain indirect relationship. The relationship among the variables is calculated and joint to find the relationship to the output variable The Table-3 shows the results by SVM algorithm. The Table-4 shows classification results by K-NN algorithm. From the classification model the feature importance for all the features involving in the training are shown in the Figure-1. The Receiver Operating Characteristics (ROC) diagram is very much valuable in finding the accuracy. It provides strong visualization and measures involved in organizing the classifiers performance. The area under the ROC curve (AUC) is usually applied to find

the equivalence among various tests using the indicator variables. It is a measure identify the test accuracy. It determines a two-dimensional visualization of set of classifiers prediction. The value achieved through ROC curve fulfil the equation $0 \leq$Accuracy ≤ 1. The value very closer to one indicates a best investigative result. The Table-5 display the results obtained for the five-factor model [21]. The Figure-2 shows the accuracy achieved by the classification algorithm for our chosen drugs.

	Accuracy	Precision	Sensitivity	AOC	Training time
Alcohol	93.4513	0.934164	0.965961	0.963238	1260.77
Caffeine	98.7611	0.987203	0.99356	0.930296	1387.99
Cannabis	92.0354	0.909968	0.943333	0.983245	1447.63
Cocaine	96.9912	0.78481	0.645833	0.923154	1495.4
Heroin	97.5221	0.666667	0.8	0.962214	1547.88
Nicotine	88.8496	0.873596	0.945289	0.967132	1616.61

Table-3. Performance Evaluation of KNN Algorithm

	Accuracy	Precision	Sensitivity	AOC	Training time
Alcohol	90.2655	0.97314	0.918129	0.836482	0.35563
Caffeine	95.3982	0.996205	0.956284	0.914845	0.0526398
Cannabis	80.885	0.97235	0.674121	0.91815	0.0758401
Cocaine	93.2743	0.136364	0.731707	0.865295	0.0407798
Heroin	95.3982	0.666667	0.857	0.809291	0.0943116
Nicotine	75.7522	0.909574	0.713987	0.840959	0.02871

Table-4. Performance Evaluation of KNN Algorithm

By comparing the results achieved through the three-classification algorithm we are able to conclude that the Decision tree algorithm produce best result for the drug Alcohol (94.33).

4.3 Evaluation of the prevailing predicted ML algorithms

One of the important tasks in constructing the predictive models with several machine learning algorithms is that the best-trained ML algorithm. The validation results determined through the confusion matrix is applied here to identify the best trained machine learning algorithm. From the Figure-2 the accuracy is applied to determine the best-trained ML algorithm. The SVM algorithm provide best results for the drugs Caffein (98.76), Cannabis

(92.03), Cocaine (96.99), Heroin (97.52) and Nicotine (88.85). The K-NN achieves best results for none of the drugs. But at the same time, it also provides a good classification result. Over all it is observed that the SVM performs well for a majority of drugs. For Alcohol also the difference between DT and SVM is also simply 0.883.

	Nscore,	Escore,	Oscore,	Ascore,	Cscore
Alcohol	0.000363	0	0.000232	0.000206	0.000487
Caffeine	0	0	0.000135	0	0
Cannabis	0.000218	0.000166	0.000558	0	0.000282
Cocaine	0.00023	0.000169	0	0	0.000448
Heroin	0.000413	0.000162	0.000369	0.000219	9.97E-05
Nicotine	0.000289	9.03E-05	0.000195	0.00027	0.000562

Table-5. NEO-Five Factor Feature importance in classification

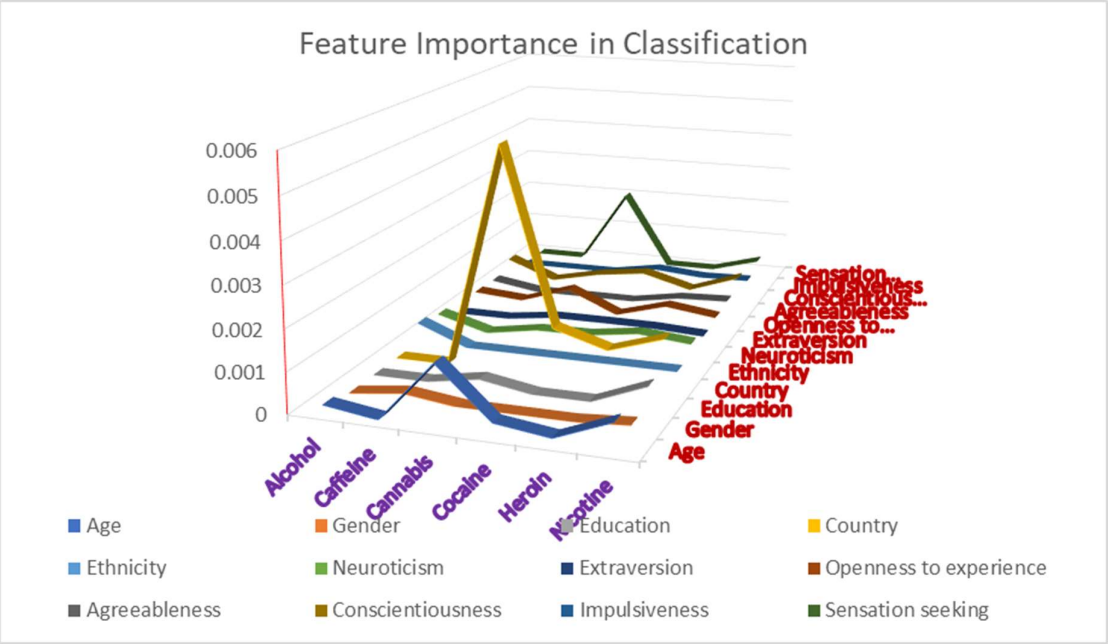

Figure 1 Feature Importance in Classification Process.

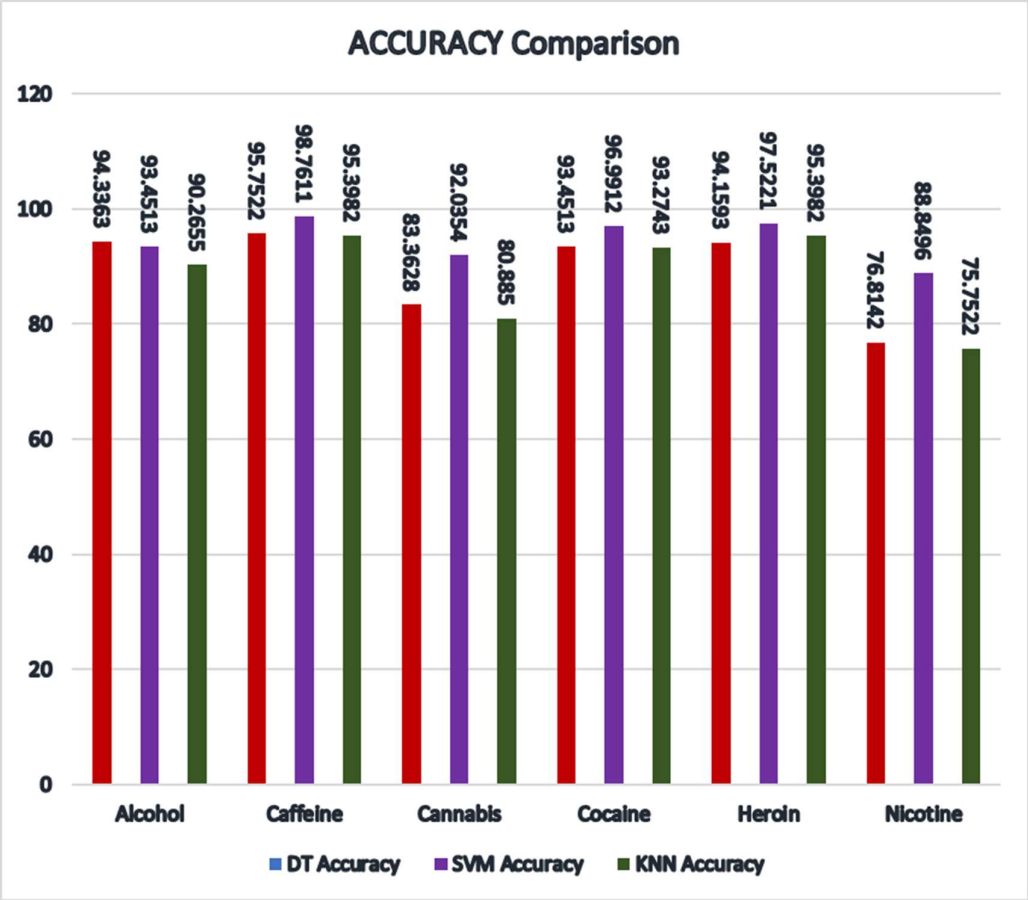

Figure-2. Accuracy chart of DT, SVM And K-NN Algorithm.

DISCUSSION

5.1 Identification of the requirements for prediction performance

The medical field require employment of several predictive models to identify people at risk of coming down with various diseases (Nobleetal.,2011). In this work we proposed a valuable predictive model for identifying people who are in risk of drug consumption by using the Decision Tree, SVM and K-NN. This work consists of 18 classification models to identify the drug consumption risk against six dugs [1. Alcohol 2. Caffeine 3. Cannabis 4. Cocaine 5. Heroin and 6. Nicotine.] and three ML algorithms [DT, SVM, and K-NN]. Both traditional and newer analytic strategies were connected with the dataset provided by WHO. The data set consist of six categorical and seven numerical data. The comparison of the proposed three machine learning algorithms performance is also being carried out. A decision tree contains the capacity to retrieve a non-linear association and create possible and sensible rules (Kammerer et al., 2005).

5.2 Training and validation of machine learning algorithms

To build prediction models for marijuana use, an iterative process of datamining was followed. The extracted dataset was stored in a .MAT format as database. Our final dataset contained 13 features and 1885 records that were fed in the machine learning algorithms. The drug use feature had two classes with users who consumed drug as 1, and never-consumed drug as 0. We applied MATLAB software to construct the proposed predictive models. In terms of

algorithm selection, we chose three commonly applied machine learning algorithms, DT, SVM, and K-NN for this research we split our dataset in 70/30 with 70% to train the models and 30% to internally validate them.

5.3 Evaluation of trained machine learning algorithms

We performed ten-fold cross-validation and reported the accuracy, precision, sensitivity and training time for each of the constructed predictive models. These results were retrieved through constructing confusion matrix. To identify and compare the important factors influencing addiction to drug consumption risk (ADCR), we testified the features with predictor weight. The determined predictor weight must be greater than 15% from the top performing model on validation data as shown in Figure-1. To extract additional validation for the acquired predictive models we also performed AUC statistic.

5.4 Limitation of the study

The Drug-testing is carried out for two important understandings. One is to identify the behaviour of organization's employees. The second one is to identify safety, efficiency, and healthiness [30]. Biological test reports and self-appraisals are the two most commonly used direct methods. Both methods have their own limitations. To avoid these limitations indirect methods are applied to identify the drug consumption. These methods deal-with measuring and observing the behaviours or reactions among drug users. They too have considerable limitations, but they can combine with the information generated through biological test reports and self-appraisals.

CONCLUSION AND FUTURE WORK

The machine learning models are very much essential to produce use full results. In the field of medicine, the perfection of the prediction is very essential because it deals with human lives. The models are also very much essential to recommend the required treatment measures to the affected patients. From our model the persons very much suspect to be in risk by drug consumption (ADCR) are to be recommended for a better treatment. The patients with minimal suspect still need to be advised for a required suitable treatment.

One of our goals of this research is to determine the feature importance which serve to conclude the performance of Decision tree, SVM and K-NN. Among the five factor features the Nscore, Oscore and Cscore provide [28,29] better contribution to identify the drug consumption risk (ADCR). While considering the ROC curve (AUC) [23] for all the drugs [Alcohol (0.96), Caffein (0.93), Cannabis (0.98), Cocaine (0.92), Heroin (0.96) and Nicotine (0.97)] the SVM Classification Model provides the best accuracy. It out performs the other two algorithms Decision tree and K-NN algorithm. In further the drug use history [24] of individuals could also be added as input to predict drug abuse risk. This problem may also be carried out as a seven-class problem also.

References

[1] Mattson, C. L., Tanz, L. J., Quinn, K., Kariisa, M., Patel, P., & Davis, N. L. Trends and geographic patterns in drug and synthetic opioid overdose deaths—United States, 2013–2019. Morbidity and Mortality Weekly Report. 2021;70(6): 202.

[2] Mooney-Leber, S. M., & Gould, T. J. The long-term cognitive consequences of adolescent exposure to recreational drugs of abuse. Learning & memory. 2018;25(9): 481-491.

[3] Fernández, D., Zabala, M. C., Ros, L., Martinez, M., Martínez, A., Latorre, J. M., & Ricarte, J. J. Testing the properties of the triarchic model of psychopathy in a community sample: Self-reported trait aggression and drug consumption associations. Scandinavian journal of psychology. 2019; 60(4): 377-385.
[4] Stephenson, N., Shane, E., Chase, J., Rowland, J., Ries, D., Justice, N., ... & Cao, R. Survey of machine learning techniques in drug discovery. Current drug metabolism. 2019; 20(3): 185-193.
[5] Kulkarni, S., Seneviratne, N., Baig, M. S., & Khan, A. H. A. Artificial intelligence in medicine: where are we now? Academic radiology. 2020;27(1): 62-70.
[6] Parekh, T., & Fahim, F. Building risk prediction models for daily use of marijuana using machine learning techniques. Drug and alcohol dependence. 2021;225: 108789.
[7] Barenholtz, E., Fitzgerald, N. D., & Hahn, W. E. Machine-learning approaches to substance-abuse research: emerging trends and their implications. Current opinion in psychiatry. 2020; 33(4): 334-342.
[8] Acion, L., Kelmansky, D., van der Laan, M., Sahker, E., Jones, D., & Arndt, S. Use of a machine learning framework to predict substance use disorder treatment success. PloS one. (2017);12(4): e0175383
[9] Kumari, D., Kilam, S., Nath, P., & Swetapadma, A. Prediction of alcohol abused individuals using artificial neural network. International Journal of Information Technology. 2018; 10(2): 233-237.
[10] Laijawala, V., Aachaliya, A., Jatta, H., & Pinjarkar, V. Classification algorithms based mental health prediction using data mining. In 2020 5th International Conference on Communication and Electronics Systems (ICCES), IEEE. 2020, June; 1174-1178).
[11] Mansour, A. M. Decision tree-based expert system for adverse drug reaction detection using fuzzy logic and genetic algorithm. International Journal of Advanced Computer Research. (2018); 8(36): 110-128.
[12] Wiyono, S., & Abidin, T. Comparative study of machine learning KNN, SVM, and decision tree algorithm to predict student's performance. International Journal of Research-Granthaalayah. (2019); 7(1): 190-196.
[13] Asicioglu, F., Genc, M. K., Bulbul, T. T., Yayla, M., Simsek, S. Z., Adioren, C., & Mercan, S. E. L. D. A. 2021.
[14] Asicioglu, F., Genc, M. K., Bulbul, T. T., Yayla, M., Simsek, S. Z., Adioren, C., & Mercan, S. E. L. D. A. Investigation of temporal illicit drugs, alcohol and tobacco trends in Istanbul city: Wastewater analysis of 14 treatment plants. Water Research. (2021); 190: 116729.
[15] Mrva, J., Neupauer, Š., Hudec, L., Ševcech, J., & Kapec, P. Decision support in medical data using 3D decision tree visualisation. In 2019 E-Health and Bioengineering Conference (EHB). 2019, November; 1-4. IEEE.
[16] Trabelsi, A., Elouedi, Z., & Lefevre, E. Decision tree classifiers for evidential attribute values and class labels. Fuzzy Sets and Systems. 2019; 366: 46-62.
[17] Tao, Z., Huiling, L., Wenwen, W., & Xia, Y. GA-SVM based feature selection and parameter optimization in hospitalization expense modeling. Applied soft computing. 2019; 75: 323-332.

[18] Richter, A. N., & Khoshgoftaar, T. M. A review of statistical and machine learning methods for modeling cancer risk using structured clinical data. Artificial intelligence in medicine. 2018; 90: 1-14.
[19] Xing, W., & Bei, Y. Medical health big data classification based on KNN classification algorithm. IEEE Access. 2019; 8: 28808-28819.
[20] Qiao, Z., Chai, T., Zhang, Q., Zhou, X., & Chu, Z. Predicting potential drug abusers using machine learning techniques. In 2019 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS). 2019, November; 283-286. IEEE.
[21] Han, P. The Application of Machine Learning Methods in Drug Consumption Prediction. In Advances in Computer, Communication and Computational Sciences. 2021; 497-507: Springer, Singapore.
[22] Widiger, T. A. (Ed.). The Oxford handbook of the five-factor model. 2017; Oxford University Press.
[23] Muschelli, J. ROC and AUC with a binary predictor: a potentially misleading metric. Journal of classification. 2020; 37(3): 696-708.
[24] Shing, J. Z., Ly, K. N., Xing, J., Teshale, E. H., & Jiles, R. B. Prevalence of hepatitis B virus infection among US adults aged 20–59 years with a history of injection drug use: National Health and Nutrition Examination Survey, 2001–2016. Clinical Infectious Diseases. 2020; 70(12): 2619-2627.
[25] Degenhardt, L., Bharat, C., Bruno, R., Glantz, M. D., Sampson, N. A., Lago, L., ... & WHO World Mental Health Survey Collaborators. Concordance between the diagnostic guidelines for alcohol and cannabis use disorders in the draft ICD-11 and other classification systems: analysis of data from the WHO's World Mental Health Surveys. Addiction. 2019;114(3): 534-552.
[26] Arif, M. A. I., Sany, S. I., Sharmin, F., Rahman, M. S., & Habib, M. T.Prediction of addiction to drugs and alcohol using machine learning: A case study on Bangladeshi population. International Journal of Electrical and Computer Engineering. 2021: 11(5), 4471.
[27] Patel, L., Shukla, T., Huang, X., Ussery, D. W., & Wang, S. Machine learning methods in drug discovery. Molecules. 2020; 25(22): 5277.
[28] Rawat, B. P. S., Jagannatha, A., Liu, F., & Yu, H. Inferring ADR causality by predicting the Naranjo Score from Clinical Notes. In AMIA Annual Symposium Proceedings. 2020; 1041. American Medical Informatics Association.
[29] Ito, N., Funasaka, K., Furukawa, K., Kakushima, N., Hirose, T., Muroi, K., ... & Fujishiro, M. A novel scoring system to predict therapeutic intervention for non-variceal upper gastrointestinal bleeding. Internal and Emergency Medicine. 2022; 17(2): 423-430.
[30] Natarajan, V. Anantha, M. Sunil Kumar, Rizwan Patan, Suresh Kallam, and Mohamed Yasin Noor Mohamed. "Segmentation of nuclei in histopathology images using fully convolutional deep neural architecture." In 2020 International Conference on computing and information technology (ICCIT-1441), pp. 1-7. IEEE, 2020.
[31] Sreedhar, B., BE, M. S., & Kumar, M. S. (2020, October). A comparative study of melanoma skin cancer detection in traditional and current image processing techniques. In 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC) (pp. 654-658). IEEE.

[32] Davanam, G., Pavan Kumar, T., & Sunil Kumar, M. (2021). Novel Defense Framework for Cross-layer Attacks in Cognitive Radio Networks. In International Conference on Intelligent and Smart Computing in Data Analytics (pp. 23-33). Springer, Singapore.
[33] Peneti, S., Sunil Kumar, M., Kallam, S., Patan, R., Bhaskar, V., & Ramachandran, M. (2021). BDN-GWMNN: internet of things (IoT) enabled secure smart city applications. Wireless Personal Communications, 119(3), 2469-2485.
[34] Sangamithra, B., Neelima, P., & Kumar, M. S. (2017, April). A memetic algorithm for multi objective vehicle routing problem with time windows. In 2017 IEEE International Conference on Electrical, Instrumentation and Communication Engineering (ICEICE) (pp. 1-8). IEEE.
[35] Balaji, K., P. Sai Kiran, and M. Sunil Kumar. "Resource aware virtual machine placement in IaaS cloud using bio-inspired firefly algorithm." Journal of Green Engineering 10 (2020): 9315-9327.
[36] Ganesh, Davanam, Thummala Pavan Kumar, and Malchi Sunil Kumar. "Optimised Levenshtein centroid cross-layer defence for multi-hop cognitive radio networks." IET Communications 15.2 (2021): 245-256.
[37] Normand, J., Lempert, R. O., O'Brien, C. P., & National Research Council., Etiology of Alcohol and Other Drug Use: An Overview of Potential Causes. Under the Influence? Drugs and the American Work Force. 1994.
[38] Harrer, S., Shah, P., Antony, B., & Hu, J. Artificial intelligence for clinical trial design. Trends in pharmacological sciences. 2019; 40(8): 577-591.