Drug Consumption Risk Evaluation

A PROJECT REPORT

for

DATA MINING TECHNIQUES (ITE2006)

in

B.Tech – Information Technology and Engineering

by

SAIDA REDDY.A	**(18BIT0299)**
LIKITH CHOWDARY.M	**(18BIT0039)**
SAI SUCHETAN REDDY.D	**(18BIT0094)**

Under the Guidance of

RANICHANDRA.C

Associate Professor, SITE

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

School of Information Technology and Engineering

June, 2021

DECLARATION BY THE CANDIDATE

We here by declare that the project report entitled **"DRUG CONSUMPTION RISK EVALUATION"** submitted by us to Vellore Institute of Technology University, Vellore in partial fulfillment of the requirement for the award of the course **Data Mining Techniques (ITE2006)** is a record of bonafide project work carried out by us under the guidance of **RANICHANDRA.C.** We further declare that the work reported in this project has not been submitted and will not be submitted, either in part or in full, for the award of any other course.

Place : Vellore Signature

Date : 04-06-2021

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

School of Information Technology & Engineering [SITE]

CERTIFICATE

This is to certify that the project report entitled **"DRUG CONSUMPTION RISK EVALUATION"** submitted by **LIKITH CHOWDARY.M (18BIT0039)** to Vellore Institute of Technology University, Vellore in partial fulfillment of the requirement for the award of the course **Data Mining Techniques (ITE2006)** is a record of bonafide work carried out by them under my guidance.

RANICHANDRA.C
GUIDE
Associate Professor, SITE

DRUG CONSUMPTION RISK EVALUATION

Abstract

This project presents different approaches to classify whether a person is consuming a specific drug or not. Various Machine Learning Classification techniques have been implemented involving the manipulation in the output labels and also various categorical input features. These techniques have been implemented assuring the major factors being age, gender, nationality, education and Ethnicity. The proposed methods have been tested on the test set and we achieved a good accuracy for the predication. The parameters used for the evaluation included Accuracy, Precision and Recall The experimental results suggest that using a particular method is subjective to its application.

Keywords – SVM, Decision Trees, Artificial Neural Networks, Drug Risk.

I. INTRODUCTION

In recent years, the use of addictive drugs and substances has turned out to be a challenging social problem worldwide. The illicit use of these types of drugs and substances appears to be increasing among elementary and high school students. After becoming addicted to drugs, life becomes unbearable and gets even worse for their users. Scientific studies show that it becomes extremely difficult for an individual to break this habit after being a user. Hence, preventing teenagers from addiction becomes an important issue. Thus we can do a classification of the people on the basis of their age, gender, education and other attributes and predict whether they have used drugs or not by using different supervised classification machine learning algorithms such as SVM, Decision Trees and Artificial Neural Networks.

II. BACKGROUND

In this section we will be the preliminary concepts used for our project work

- We have to get the dataset form trusted sources. We are working on the Drug consumption dataset prepared by University of California, Irvine.
- We are not only predicting whether a person consumes a specific drug or not, but also for how long he might have used the drug.
- We are predicting based on 13 attributes related to a person rather than 4-5 attributes.

III. Literature Survey

[1]. Pharmacovigilance is defined as the science and activities relating to the detection, assessment, understanding, and prevention of adverse drug events (WHO 2004). Post-approval adverse drug events are a major health concern The availability of various sources of healthcare data for analysis in recent years opens new opportunities for the data-driven pharmacovigilance researchMost studies in pharmacovigilance focus on structured and coded data, and therefore miss important textual data from patient social media and clinical documents in HER. f such systems is hampered by the bias in data and the pitfalls of the data mining algorithms adopted.

[2]. In this article we observed a proof of concept for a novel efficient ADR signal refinement method that filters instances of a DOI-HOI (Drug of interest-Health outcome of interest) signal and does not require knowledge of possible confounders. The recorded history of a patient experiencing the signal is used to filter instances where the medical event can be explained by alternative causes (other than the drug). The tentative results suggest that the method has the capability to efficiently refine ADR signals but each signal may require specific tuning to determine the optimal support and confidence values to be implemented

[3] .This paper describes four popular data mining algorithms Sequential minimal optimization (SMO), Bagging, REP Tree and decision table (DT) extracted from a decision tree or rule-based classifier to improve the efficiency of academic performance in the

educational institutions for students who consume alcohol. In this paper, we present a real-world experiment conducted at VBS Purvanchal University, Jaunpur, India. This method helps to identify the students who need special advising or counseling by the councilors/teachers to understand the danger of consuming alcohol

[4]. The problem of evaluating an individual's risk of drug consumption and misuse is highly important and novel. An online survey methodology was employed to collect data including personality traits (NEO-FFI-R), impulsivity (BIS-11), sensation seeking (ImpSS), and demographic information The best results with sensitivity and specificity being greater than 75% were achieved for cannabis, crack, ecstasy, legal highs, LSD, and volatile substance abuse. Sensitivity and specificity greater than 70% were achieved for amphetamines, amyl nitrite, benzodiazepines, chocolate, caffeine, heroin, ketamine, methadone, and nicotine

[5]. This study is aimed at analysing the predictive power of different psychosocial and personality variables on the consumption or nonconsumption of nicotine in a teenage population using different classification techniques from the field of Data Mining. To this end, we worked with a sample of 2666 teenagers, 1378 of whom do not consume nicotine while 1288 are nicotine consumers. The models analysed were able to discriminate correctly between both types of subjects within a range of 77.39% to 78.20%, achieving 91.29% sensitivity and 74.32% specificity .This study was carried out with the help of the National Plan on Drugs.

[6]. This review/survey paper based on the research carried out in the area of data mining depends for managing bulk amount of data with mining in social media on using composite applications for performing more sophisticated analysis using cloud platform. The objective of this paper is to introduce such type of tool which used in social network to characterised drug abuse This paper describes how to fetch important data for analysis from social network as Twitter, Facebook, and Instagram. Then big data techniques to extract useful content for analysis are discussed.

[7]. This paper describes four popular data mining algorithms Sequential minimal optimization (SMO), Bagging, REP Tree and decision table (DT) extracted from a decision tree or rule-based classifier to improve the efficiency of academic performance in the educational institutions for students who consume alcohol. Safety analysis is performed using multivariate analysis. Bayesian logistic regression and unsupervised machine learning approaches are used to find multiple adverse events. This method helps to identify the students who need special advising or counselling by the counsellors/teachers to understand the danger of consuming alcohol.

[8]. In this paper they evaluated what method works best at predicting successful treatment using a real life large database. Rather than a dataset tailored to showcase the methods properties, they used a dataset commonly used in the SUD field and hypothesized that SL will generate the best prediction model as measured by the area under the receiver operating characteristic curve (AUC) evaluated in a test sample not included in the training sample used to fit all prediction models.

[9]. In this paper ,Two ANN modules are designed, ANN-D to predict whether a person is using VSA or not and ANN-C to predict the time of use. Input features used are age, gender, country, ethnicity, education, neuroticism, openness to experience, extraversion, agreeableness, conscientiousness, impulsiveness, sensation seeking etc. Input features are given to the ANN-D module to predict if volatile substance abuse (VSA) has been done by the person or not. ANN-C module predicts the use of VSA in terms of time such as day, week, month, year, decade, before a decade, etc.

[10]. In this paper, a drug consumption data set from a survey, containing records for 1885 respondents, about their attributes, which include the five-factor model traits, impulsivity, sensation seeking, and other demographic characteristics, and CNS psychoactive drugs consumption history (18 drugs) was analyzed and used to develop a predictive model.Also, in this study, the Association Rule mining was used to discover and identify patterns in the drug consumption data to get more interpretable insight from the data set.

[11]. In this paper they partitioned the data in distinct training and test sets where not only pairs but also drugs/diseases were not overlapped. They tested several classifiers under

different cross validation schemes and compared our approach with existing methods. A binary feature matrix was generated using drug target, substructure and side effects and disease ontology terms. They observed that their model had better predictive performance than the existing models in disjoint cross-validation settings.

[12].In this paper they present a new data mining technique based on the bi-clustering paradigm, which is designed to identify drug groups that share a common set of adverse events (AEs) in the spontaneous reporting system (SRS) of the US Food and Drug Administration (FDA). Taxonomy of bi-clusters is developed, revealing that a significant number of bonafide adverse drug event (ADE) bi-clusters have been identified. Statistical tests indicate that it is extremely unlikely that the bi-cluster structures thus discovered, as well as their content, could have arisen by mere chance. The objective of this article is to describe a novel pharmacovigilance data mining technique designed to identify drug groups that share a common set of AEs, with which potential ADEs are analyzed and previously unrecognized ADEs may be identified.

[13]. The objective is to search potential therapeutic drugs for Parkinson's disease based on data mining and bioinformatics analysis and providing new ideas for research studies on "new application of conventional drugs." Results showed that metformin hydrochloride and other drugs had certain therapeutical effect on Parkinson's disease, and melbine (DMBG) can be used for treatment of Parkinson's disease and type 2 diabetes patient.

[14]. So, there work proposes to use an unsupervised data mining technique such as kmeans algorithm to group universities with similar characteristics. Then, the DEA is utilized for each cluster separately. The result shows a better improvements and a fair comparison of performance between universities.

[15]. In this paper they have conducted a survey in the island of mollarca over 93000 students have participated .Drug use motives are relevant to understand substance use amongst students. Data mining techniques present some advantages that can help to improve our understanding of drug use issue. The aim of this paper is to explore, through data mining techniques, the reasons why students use drugs.

[16]. In this paper they analysed drug use motives through a series of classification techniques included in Data Mining: two classical machine learning techniques, Decision Trees (DT) and Artificial Neural Networks (ANN); two modern statistical techniques, k-Nearest Neighbours (K-NN) and Naïve Bayes (NB); and a classical statistical technique, Logistic Regression (LogR). A random cluster sampling of schools was conducted in the island of Mallorca. Participants were 9,300 students.

[17]. Adverse drug reactions denote a major health problem all over the world. It describes any injury caused by taking a drug or overdose of drug or due to combination of two or more drugs. Detection of adverse drug reactions is compulsory because they affect large number of people and can help in raising early warning against adverse effects of drugs and help medical experts in making treatment effective and timely. In today's digital era a huge amount of data correlated to adverse effects of drugs is being collected at hospitals, drug retail stores and by drug producers.

[18]. In this paper, we explore the possibility of using big multimedia data, including both images and text, from social media in order to discover drug use patterns at fine granularity with respect to demographics. Instagram posts are searched and collected by drug related terms by analyzing the hashtags supplied with each post. A large and dynamic dictionary of frequent drug related slangs is used to find these posts. User demographics are extracted using robust face image analysis algorithms. These posts are then mined to find common trends with regard to the time and location they are posted, and further in terms of age and gender of the drug users.

[19]. As Computational methods capable of predicting these failures can reduce the waste of resources and time devoted to the investigation of compounds that ultimately fail.They proposed an original machine learning method that leverages identity of drug targets and off-targets, functional impact score computed from Gene Ontology annotations, and biological network data to predict drug toxicity. They demonstrate that their method (TargeTox) can distinguish potentially idiosyncratically toxic drugs from safe drugs and is also suitable for speculative evaluation of different target sets to support the design of optimal low-toxicity combinations.

[20]. There are many harmful effects of drug abuse, including changes in the user's brain, body, and spirit. This work does not deal with the results, but rather deals with the reasons for being dependent. There are some reasons leading adolescents to addiction. The first is about family- related issues, such as dissatisfaction with family relationships, antisocial family members, stress in the family, poverty or welfare usage in the family, illiterate parents, divorced parents, loss of one or both parents, and lack of people who could be a positive role model for the adolescent.

[21]. The objective of this work was to find out whether specific Type A Behavior Pattern (TABP) profiles exist which may constitute a risk factor for the consumption of legal and illegal substances (alcohol, tobacco, psychoactive drugs, cannabis, cocaine and hallucinogens) by young Spaniards.Specific personality profiles were identified which constitute either a risk factor or a protective factor for substance abuse. These results will prove useful to drug consumption prevention and treatment programs focused on the above-mentioned personality profiles.

[22]. The primary part is that the learning part, wherever the coaching knowledge is analyzed and classification rules square measure generated. Subsequent part is the classification, wherever check knowledge is classified into categories in step with the generated rules. Since classification algorithms need that categories be outlined primarily based on knowledge attribute values, we tend to had created associate degree attribute class for each knowledge that will have a worth of either level of addiction or level of extra tablets

[23]. In the present study, a UPLC-TOF-MS analytical strategy combined with online data acquiring and post-acquisition data-mining methods was established for heroin metabolite identification .Correlation analysis was used to investigate the internal relationship among the metabolites and seven respective metabolites were selected into one group as "Target-metabolites". Moreover, MRM screening method was employed via UPLC-MS/MS to validate the practicability and reliability of the identified metabolites.

[24] . In this paper they developed MEDICASCY, a multilabel-based boosted random forest machine learning method that only requires the small molecule's chemical structure for the drug side effect, indication, efficacy, and probable mode of action target predictions. however, it has comparable or even significantly better performance than existing approaches requiring far more information. Kmeans, KNN, Random forest, multi layer perceptron are used. MEDICASCY shows about 78% precision and recall for predicting at least one severe side effect and 72% precision drug efficacy.

[25].They aimed to validate the usefulness of artificial neural networks for the prediction of adverse drug reactions and focused on vancomycin -induced nephrotoxicity. For constructing an artificial neural network, a multilayer perceptron algorithm was employed. A 10-fold cross validation method was adopted for evaluating the resultant artificial neural network. Among these patients, 179 (15.7%) developed vancomycin -induced nephrotoxicity. The top three risk factors of vancomycin -induced nephrotoxicity which are relatively important in the artificial neural networks were average vancomycin trough concentration ≥ 13.0 mg/L and concomitant use of piperacillin–tazobactam and vasopressor drugs. The predictive accuracy of the artificial neural network was 86.3% and that of the multiple logistic regression model (conventional statistical method) was 85.1. The artificial neural network model predicting the vancomycin -induced nephrotoxicity showed good predictive performance.

[26]. As Computational methods capable of predicting these failures can reduce the waste of resources and time devoted to the investigation of compounds that ultimately fail.They proposed an original machine learning method that leverages identity of drug targets and off-targets, functional impact score computed from Gene Ontology annotations, and biological network data to predict drug toxicity. They demonstrate that their method (TargeTox) can distinguish potentially idiosyncratically toxic drugs from safe drugs and is also suitable for speculative evaluation of different target sets to support the design of optimal low-toxicity combinations.

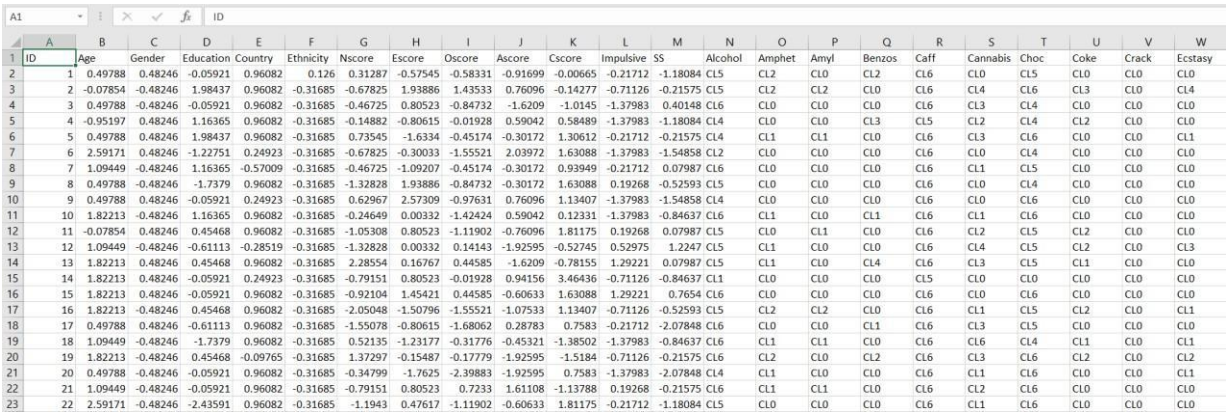
[27]. This paper describes four popular data mining algorithms Sequential minimal optimization (SMO), Bagging, REP Tree and decision table (DT) extracted from a decision tree or rule- based classifier to improve the efficiency of academic performance in the educational institutions for students who consume alcohol. In this paper, we present a real-world experiment conducted at VBS Purvanchal University, Jaunpur, India. This method

helps to identify the students who need special advising or counseling by the councilors/teachers to understand the danger of consuming alcohol

[28]. In this paper they developed MEDICASCY, a multilabel-based boosted random forest machine learning method that only requires the small molecule's chemical structure for the drug side effect, indication, efficacy, and probable mode of action target predictions. They demonstrate that their method (TargeTox) can distinguish potentially idiosyncratically toxic drugs from safe drugs and is also suitable for speculative evaluation of different target sets to support the design of optimal low-toxicity combinations.

IV. DATASET DESCRIPTION AND SAMPLE DATA

Database contains records for 1885 respondents. For each respondent 12 attributes are known: Personality measurements which include NEO-FFI-R (neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness), BIS-11 (impulsivity), and ImpSS (sensation seeking), level of education, age, gender, country of residence and ethnicity. All input attributes are originally categorical and are quantified. After quantification values of all input features can be considered as real-valued. In addition, participants were questioned concerning their use of 18 legal and illegal drugs (alcohol, amphetamines, amyl nitrite, benzodiazepine, cannabis, chocolate, cocaine, caffeine, crack, ecstasy, heroin, ketamine, legal highs, LSD, methadone, mushrooms, nicotine and volatile substance abuse and one fictitious drug (Semeron) which was introduced to identify over-claimers. For each drug they have to select one of the answers: never used the drug, used it over a decade ago, or in the last decade, year, month, week, or day.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	ID	Age	Gender	Education	Country	Ethnicity	Nscore	Escore	Oscore	Ascore	Cscore	Impulsive	SS	Alcohol	Amphet	Amyl	Benzos	Caff	Cannabis	Choc	Coke	Crack	Ecstasy
2	1	0.49788	0.48246	-0.05921	0.96082	0.126	0.31287	-0.57545	-0.58331	-0.91699	-0.00665	-0.21712	-1.18084	CL5	CL2	CL0	CL2	CL6	CL0	CL5	CL0	CL0	CL0
3	2	-0.07854	-0.48246	1.98437	0.96082	-0.31685	-0.67825	1.93886	1.43533	0.76096	-0.14277	-0.71126	-0.21575	CL5	CL2	CL2	CL0	CL6	CL4	CL6	CL3	CL0	CL4
4	3	0.49788	-0.48246	-0.05921	0.96082	-0.31685	-0.46725	0.80523	-0.84732	-1.6209	-1.0145	-1.37983	0.40148	CL6	CL0	CL0	CL0	CL6	CL3	CL4	CL0	CL0	CL0
5	4	-0.95197	0.48246	1.16365	0.96082	-0.31685	-0.14882	-0.80615	-0.01928	0.59042	0.58489	-1.37983	-1.18084	CL4	CL0	CL0	CL3	CL5	CL2	CL4	CL2	CL0	CL0
6	5	0.49788	0.48246	1.98437	0.96082	-0.31685	0.73545	-1.6334	-0.45174	-0.30172	1.30612	-0.21712	-0.21575	CL4	CL1	CL1	CL0	CL6	CL3	CL6	CL0	CL0	CL1
7	6	2.59171	0.48246	-1.22751	0.24923	-0.31685	-0.67825	-0.30033	-1.55521	2.03972	1.63088	-1.37983	-1.54858	CL2	CL0	CL0	CL0	CL6	CL0	CL4	CL0	CL0	CL0
8	7	1.09449	-0.48246	1.16365	-0.57009	-0.31685	-0.46725	-1.09207	-0.45174	-0.30172	0.93949	-0.21712	0.07987	CL6	CL0	CL0	CL0	CL6	CL1	CL5	CL0	CL0	CL0
9	8	0.49788	-0.48246	-1.7379	0.96082	-0.31685	-1.32828	1.93886	-0.84732	-0.30172	1.63088	0.19268	-0.52593	CL5	CL0	CL0	CL0	CL6	CL0	CL4	CL0	CL0	CL0
10	9	0.49788	0.48246	-0.05921	0.24923	-0.31685	0.62967	2.57309	-0.97631	0.76096	1.13407	-1.37983	-1.54858	CL4	CL0	CL0	CL0	CL6	CL0	CL6	CL0	CL0	CL0
11	10	1.82213	-0.48246	1.16365	0.96082	-0.31685	-0.24649	0.00332	-1.42424	0.59042	0.12331	-1.37983	-0.84637	CL6	CL1	CL0	CL1	CL6	CL1	CL6	CL0	CL0	CL0
12	11	-0.07854	0.48246	0.45468	0.96082	-0.31685	-1.05308	0.80523	-1.11902	-0.76096	1.81175	0.19268	0.07987	CL5	CL0	CL1	CL0	CL6	CL2	CL5	CL2	CL0	CL0
13	12	1.09449	-0.48246	-0.61113	-0.28519	-0.31685	-1.32828	0.00332	0.14143	-1.92595	-0.52745	0.52975	1.2247	CL5	CL1	CL0	CL0	CL6	CL4	CL5	CL2	CL0	CL3
14	13	1.82213	0.48246	0.45468	0.96082	-0.31685	2.28554	0.16767	0.44585	-1.6209	-0.78155	1.29221	0.07987	CL5	CL1	CL0	CL4	CL6	CL3	CL5	CL1	CL0	CL0
15	14	1.82213	0.48246	-0.05921	0.24923	-0.31685	-0.79151	0.80523	-0.01928	0.94156	3.46436	-0.71126	-0.84637	CL1	CL0	CL0	CL0	CL5	CL0	CL0	CL0	CL0	CL0
16	15	1.82213	0.48246	-0.05921	0.96082	-0.31685	-0.92104	1.45421	0.44585	-0.60633	1.63088	1.29221	0.7654	CL6	CL0	CL0	CL0	CL6	CL0	CL6	CL0	CL0	CL0
17	16	1.82213	-0.48246	0.45468	0.96082	-0.31685	-2.05048	-1.50796	-1.55521	-1.07533	1.13407	-0.71126	-0.52593	CL5	CL2	CL2	CL0	CL6	CL1	CL5	CL2	CL0	CL1
18	17	0.49788	0.48246	-0.61113	0.96082	-0.31685	-1.55078	-0.80615	-1.68062	0.28783	0.7583	-0.21712	-2.07848	CL6	CL0	CL0	CL1	CL6	CL3	CL5	CL0	CL0	CL0
19	18	1.09449	-0.48246	-1.7379	0.96082	-0.31685	0.52135	-1.23177	-0.31776	-0.45321	-1.38502	-1.37983	-0.84637	CL6	CL1	CL1	CL0	CL6	CL6	CL4	CL1	CL0	CL1
20	19	1.82213	-0.48246	0.45468	-0.09765	-0.31685	1.37297	-0.15487	-0.17779	-1.92595	-1.5184	-0.71126	-0.21575	CL6	CL2	CL0	CL2	CL6	CL3	CL6	CL2	CL0	CL2
21	20	0.49788	-0.48246	-0.05921	0.96082	-0.31685	-0.34799	-1.7625	-2.39883	-1.92595	0.7583	-1.37983	-2.07848	CL4	CL1	CL0	CL0	CL6	CL1	CL6	CL0	CL0	CL1
22	21	1.09449	-0.48246	-0.05921	0.96082	-0.31685	-0.79151	0.80523	0.7233	1.61108	-1.13788	0.19268	-0.21575	CL6	CL1	CL1	CL0	CL6	CL2	CL6	CL0	CL0	CL0
23	22	2.59171	-0.48246	-2.43591	0.96082	-0.31685	-1.1943	0.47617	-1.11902	-0.60633	1.81175	-0.21712	-1.18084	CL5	CL0	CL0	CL0	CL6	CL1	CL6	CL0	CL0	CL0

V. PROPOSED ALGORITHM WITH FLOWCHART

a.) Data Extraction:

We will import the dataset from University of California, Irvine Repository. We will download the zip file and extract the excel format of drugconsumption.csv dataset. Using libraries such as pandas we will import the dataset into the jupyter notebook environment.

b.) Data Visualization:

Based on all the attributes present in the dataset, to get better understanding about the dataset and to obtain various useful insights from the data, we will perform various types of visualization such as Box plot, Violin Plot, Sub plots, Spatial Analysis using matplotlib and seaborn libraries.

c.) Pre- Processing and Feature Engineering:

Before applying any algorithm or making any classifications, We need to clean the data, i.e, Pre-Processing Must be done. In the dataset mentioned above there are some attributes which stores categorical data, and these must be converted into numerical data. We will be using techniques such as One Hot Encoding to carry out the task.

Not all the Features are essential for our final goal of Classification. There will be some features which are more essential, i.e, have more impact on the output when compared to others. To know the correlation between attributes a heat map can be designed which some the amount of correlation between different attributes. Based on the results obtained in the above heat map we will be selecting the attributes.

d.) Model Selection:

The next important step is to design/ select an model to fit on our attributes. We will divide the attributes into attributes into two categories x & y. Where x contains all the attributes

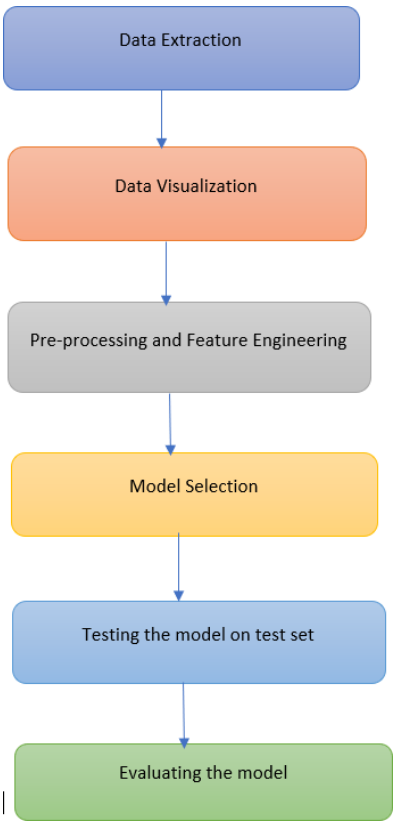
which are required for the classification, and y represent the predicted label attribute. We shoulf fit our model on x&y using a classification based machine algorithm such as SVM, Decision Trees and Neural Networks.

e.) Testing the model on the test set:

For any model designed it is very essential to test our model. So we should divide the data into train and test data. Firstly, we need to train the data on our training set and we need to test the model on the testing data to check out the performance of the model.

f.) Evaluating the Model:

Finally, we before we deploy our model we need to know the accuracy of the model. To find the accuracy of our model after it predicting the test data we can use various evaluation metrics such as Accuracy score. Based on all the evaluation metrics we will select the model which gives the highest accuracy and low error.



Data Mining Functionality:

Classification:

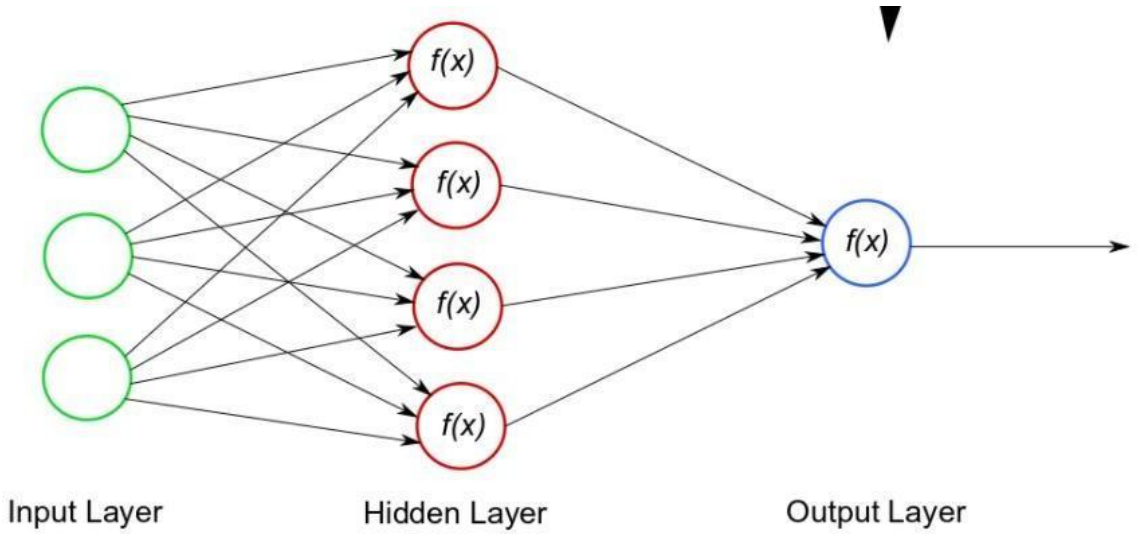
It is the organization of data in given classes. Classification uses given class labels to order the objects in the data collection. Classification approaches normally use a training set where all objects are already associated with known class labels. The classification algorithm learns from the training set and builds a model. The model is used to classify new objects.

In this project, we analyze the whether the person is addicted to drug or not based on the features such as age, gender, ethnicity etc. Based on the above features there are two possible labels: “safe” and “Not Safe”.

Algorithm used:

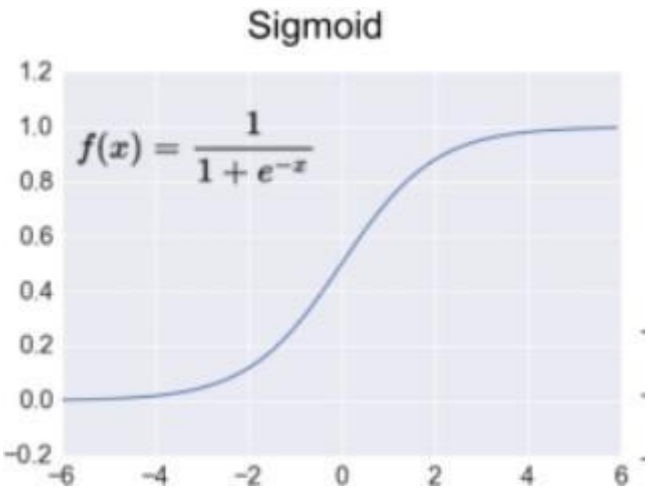
a.) Neural Network.

In our project we are going to use a Neural Network for the purpose of classification whether the person consumes a specific drug or not. For this classification, we are going to use a Neural Network with an Input layer, an Output layer, and one Hidden layer. This kind of model is called Single layer Neural Network or Perceptron.

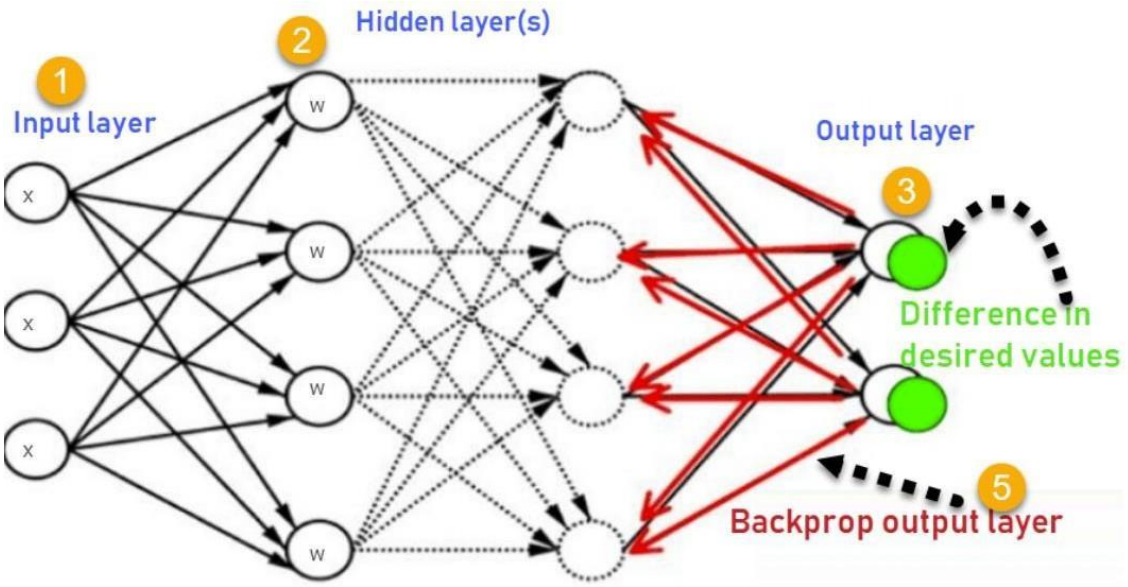


A single-layer neural network represents the most simple form of neural network, in which there is only one layer of input nodes that send weighted inputs to a subsequent layer of receiving nodes, or in some cases, one receiving node. each neuron is connected to numerous other neurons, allowing signals to pass in one direction through the network from input to output layers, including through any number of hidden layers in between.

The activation function keeps values forward to subsequent layers within an acceptable and useful range, and forwards the output. In our project we are using to use sigmoid activation function. The sigmoid function "squashes" values to the range 0 and 1.

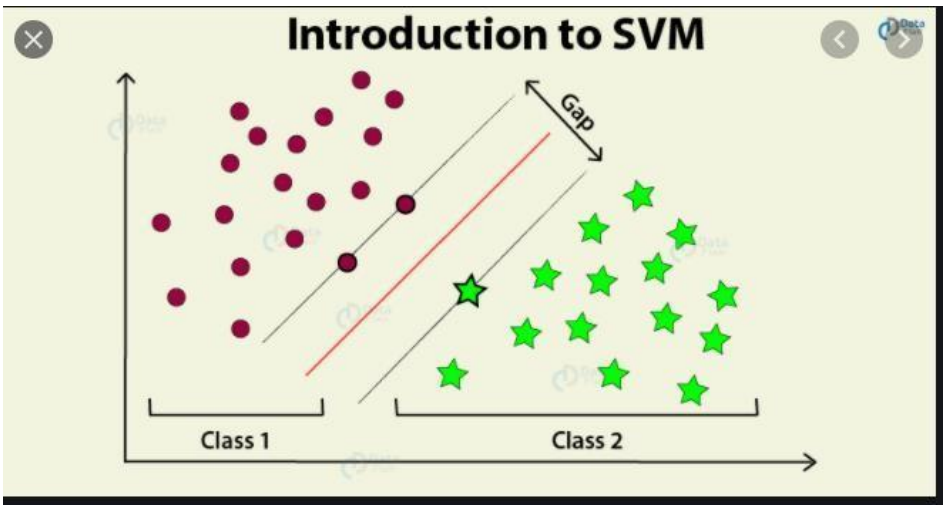


Backpropagation is a short form for "backward propagation of errors." It is a standard method of training artificial neural networks. This method helps to calculate the gradient of a loss function with respects to all the weights in the network. It is the method of fine-tuning the weights of a neural net based on the error rate obtained in the previous epoch (i.e., iteration). Proper tuning of the weights allows you to reduce error rates and to make the model reliable by increasing its generalization.



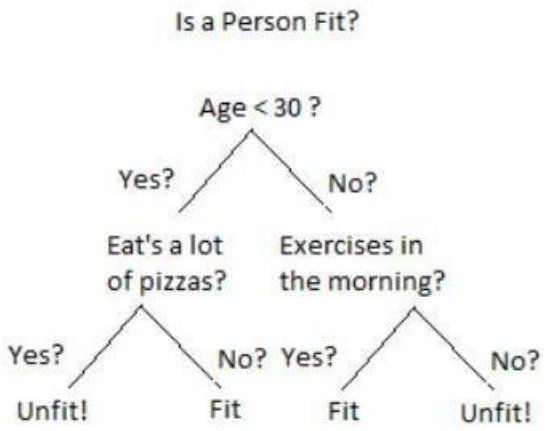
b.) SVM:

A support vector machine (SVM) is a supervised machine learning model that uses classification algorithms for two-group classification problems. We carry out plotting in the n-dimensional space. Value of each feature is also the value of the specific coordinate. Then, we find the ideal hyperplane that differentiates between the two classes. After giving an SVM model sets of labelled training data for each category, they're able to categorize new data. In this project an SVM model is trained using the dataset and is used to classify whether a person is addicted to following drug or not.



c.) Decision Trees:

Decision tree learning is a supervised machine learning technique for inducing a decision tree from training data. A decision tree (also referred to as a classification tree or a reduction tree) is a predictive model which is a mapping from observations about an item to conclusions about its target value. In the tree structures, leaves represent classifications (also referred to as labels), nonleafy nodes are features, and branches represent conjunctions of features that lead to the classifications.



VI. EXPERIMENTS RESULTS

In this project we have built a model on 3 Machine learning algorithms and predicted the test data on these model. We tried not only to find whether a person consumes drugs or not, but also for how he has been consuming drugs. For the binary classification we used SVM for binary classification and decision trees and to find out the time we used models such as Artificial Neural Networks and Support vector Machines, and compared these models on the basis of their accuracy score, which is the percent of correct outcomes to the total number of outcomes.

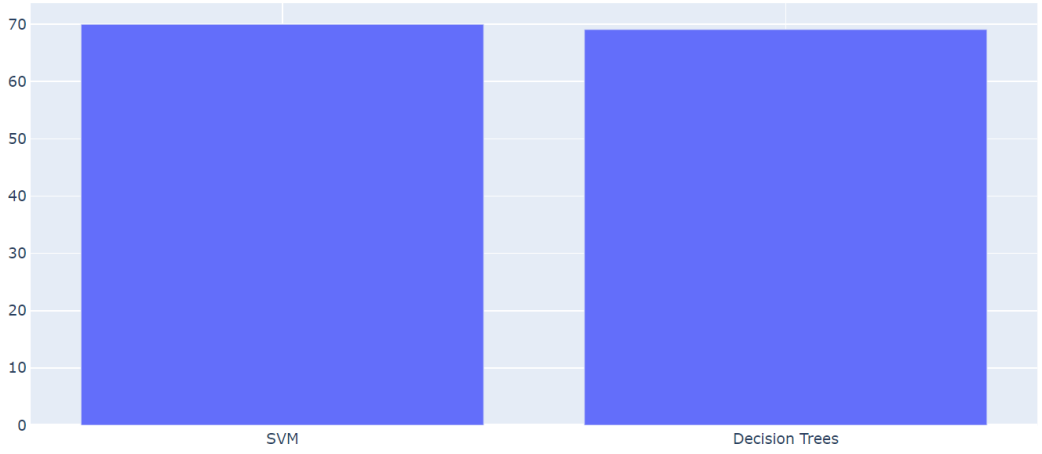
VII. COMPARATIVE STUDY / RESULTS AND DISCUSSION

In this project, we have made an attempt to classify Drug Consumption Risk using machine learning techniques. Two algorithms namely SVM and Decision Trees are implemented along with a single hidden layer Artificial Neural Network. The Evaluation metrics which we used to compare the models here are accuracy score.

The Accuracy Score which we got are as follows:

- SVM for Binary Classification: 69.96466431095406.
- Decision Trees: 69.08127208480566.

From the above results, Support Vector Machine is slightly better method compared to Decision Trees Technique with 69.9% accuracy which means the model built for the prediction Drug Risk Evaluation gives 66.9% right prediction.



VIII. CONCLUSION AND FUTURE WORK

Drug use is a risk behaviour that does not happen in isolation. It includes numerous risk factors, which are defined as any attribute, characteristic, or event in the life of an individual that increases the probability of drug consumption. Psychological, social, environmental, economic, and individual factors are correlated with initial drug use. As our dataset also contain 13 attributes as we are giving many attributes to model can confuses the model that's why we have achieved an accuracy around 70 percent.

We can improve the accuracy by selecting only a few attributes which are highly corelated to the drug rather than all the 13 features. Moreover, we can also use boosting techniques to improve the accuracy of the model.

IX. REFERENCES

1. Xiao Liu (2016) "HEALTH DATA ANALTYICS: DATA AND TEXT MINING APPROACHES FOR PHARMACOVIGILANCE"
2. Elaine Fehrman (2017) "The Five Factor Model of Personality and Evaluation of Drug Consumption Risk"
3. Saurabh Pal, Vikas Chaurasia (2017) "PERFORMANCE ANALYSIS OF STUDENTS CONSUMING ALCOHOL USING DATA MINING TECHNIQUES"
4. Pal, S., & Chaurasia, V. (2017). Performance analysis of students consuming alcohol using data mining techniques. *International Journal of Advance Research in Science and Engineering*, *6*(2), 238-250.
5. Nath, Priyanka, Sumran Kilam, and Aleena Swetapadma. "A machine learning approach to predict volatile substance abuse for drug risk analysis." In *2017 Third*

International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN), pp. 255-258. IEEE, 2017.

6. Acion, Laura, Diana Kelmansky, Mark van der Laan, Ethan Sahker, DeShauna Jones, and Stephan Arndt. "Use of a machine learning framework to predict substance use disorder treatment success." *PloS one* 12, no. 4 (2017): e0175383.
7. Celebi, Remzi, Özgün Erten, and Michel Dumontier. "Machine Learning Based Drug Indication Prediction Using Linked Open Data." In *SWAT4LS*. 2017.
8. Juan J. Montaño-Moreno*, Elena Gervilla-García, Berta Cajal-Blasco y Alfonso Palmer (2017) "Data mining classification techniques: an application to tobacco consumption in teenagers"
9. Shadma Qureshi,Sonal Rai,Shiv Kumar (2017) "Mining Social Media Data for Understanding Drugs Usage"
10. E Fehrman, A K Muhammad, E M Mirkes, V Egan, A N Gorban (2017) Analysis of Drug Consumption Data Using Data Mining Techniques and a Predictive Model using Multi-label Classification
11. Susanne Stoelben Jutta Krappweis & Wilhelm Kirch (2017) Adolescents' drug use and drug knowledge
12. Jiménez, Rafael, Joella Anupol, Berta Cajal, and Elena Gervilla. "Data mining techniques for drug use research." *Addictive behaviors reports* 8 (2018): 128-135.
13. Chuan Xu, Jiajun Chen, Xia Xu, Yingyu Zhang and Jia Li (2018) "Potential Therapeutic Drugs for Parkinson's Disease Based on Data Mining and Bioinformatics Analysis"
14. Anusha N, Rajashree, Srikanth Bhat K (2018) "A Survey on Medical Data by using Data Mining Techniques"
15. Rafael Jiménez, Joella Anupol, Berta Cajal, Elena Gervilla (2018) "Data mining techniques for drug use research"
16. Rosario Ruiz-Olivares, Valentina Lucena , Antonio F. Raya, Javier Herruzo(2019) Personality profiles and how they relate to drug consumption among young people in Spain
17. Mahyoub, Mohammed A., Laith Abu Lekham, Emad Alenany, Lubna Tarawneh, and Daehan Won. "Analysis of Drug Consumption Data Using Data Mining Techniques and a Predictive Model using Multi-label Classification." In *Proceedings of the 2019*

IISE Annual Conference, Orlando, pp. 864-869. 2019.Faruk BULU(2019) An urgent precaution system to detect students at risk of substance abuse through classification algorithms

18. G. Thailambal , R. Subramani , S. Saradha (2019) DRUGS USAGE PREDICTION IN WEKA TOOL USING C4.5 CLASSIFICATION ALGORITHM
19. Anusha N, Rajashree, Srikanth Bhat K (2019) A Survey on Medical Data by using Data Mining Techniques
20. Zhou, Hongyi, Hongnan Cao, Lilya Matyunina, Madelyn Shelby, Lauren Cassels, John F. McDonald, and Jeffrey Skolnick. "MEDICASCY: A Machine Learning Approach for Predicting Small-Molecule Drug Side Effects, Indications, Efficacy, and Modes of Action." *Molecular Pharmaceutics* 17, no. 5 (2020): 1558-1574.
21. Zhang, Dan, Jintao Lv, Bing Zhang, Xiaomeng Zhang, Hao Jiang, and Zhijian Lin. "The characteristics and regularities of cardiac adverse drug reactions induced by Chinese materia medica: A bibliometric research and association rules analysis." *Journal of Ethnopharmacology* 252 (2020): 112582.
22. Imai, Shungo, Yoh Takekuma, Hitoshi Kashiwagi, Takayuki Miyai, Masaki Kobayashi, Ken Iseki, and Mitsuru Sugawara. "Validation of the usefulness of artificial neural networks for risk prediction of adverse drug reactions used for individual patients in clinical practice." *Plos one* 15, no. 7 (2020): e0236789.
23. Rafael Jiménez, Joella Anupol, Berta Cajal, Elena Gervilla (2020) "Data mining techniques for drug use research"
24. Rosario Ruiz-Olivares, Valentina Lucena , Antonio F. Raya, Javier Herruzo(2019) Personality profiles and how they relate to drug consumption among young people in Spain
25. Saurabh Pal, Vikas Chaurasia (2020) performance analysis of students consuming alcohol using data mining techniques
26. L.Thomas Robinson ,Soniya.G, , D.Brintha Sweetly(2020) Mining Social Media Data for Understanding Drugs Usage
27. Faruk BULU(2020) An urgent precaution system to detect students at risk of substance abuse through classification algorithms

Appendix

CODE:-

[1]
```
import warnings
```

[2]
```
#import the necessary libraries
    import pandas as pd
    import seaborn as sns
    import numpy as np
    import matplotlib.pyplot as plt
    #set the backgroung style sheet
    sns.set_style("whitegrid")
    %matplotlib inline
```

[3]
```
#importing the libaries required for data visualization
   from plotly.offline import download_plotlyjs, init_notebook_mode, plot,
iplot
   import plotly.offline as py
   from plotly.graph_objs import Scatter, Layout
   py.init_notebook_mode(connected=True)
   import plotly.graph_objs as go
   import plotly.figure_factory as ff
```

[4]
```
from google.colab import drive
   drive.mount('/content/drive')
```

[5]
```
#load the csv file in a dataframe using read_csv function
   df = pd.read_csv("/content/drive/MyDrive/datamining/drug_consumption
(1).csv",encoding="latin-1")
```

[6]
```
 #make a copy of the dataframe
    copy_df = df.copy()
    #information about the dataset
    df.info()
```

[7]
```
#print first 5 rows of the dataset
   df.head()
```

[8]
```
#more information about the data
   df.describe()
```

[9]
```
#finding whether there are any null values in the dataset
   df.isna().sum()
```

[10]
```
#data visualization using box plot on dependent variables
    feature_col_names = ['Age', 'Gender', 'Education', 'Country',
```

```
'Ethnicity', 'Nscore',
        'Escore', 'Oscore', 'Ascore', 'Cscore', 'Impulsive', 'SS']
  plt.style.use('ggplot')
  f, ax = plt.subplots(figsize=(11, 15))
  ax.set(xlim=(-.100, 5))
  plt.ylabel('Dependent Variables')
  plt.title("Box Plot of Pre-Processed Data Set")
  ax = sns.boxplot(data = df[feature_col_names], orient = 'h', palette =
'Set2')
```

[11]
```
#creating 2 more dataframes for each drug
    columns = ['Alcohol','Amphet', 'Amyl', 'Benzos', 'Caff', 'Cannabis',
'Choc', 'Coke', 'Crack',
           'Ecstasy', 'Heroin', 'Ketamine', 'Legalh', 'LSD', 'Meth',
'Mushrooms','Nicotine', 'Semer', 'VSA']
   cp = ['User_Alcohol','User_Amphet', 'User_Amyl', 'User_Benzos',
'User_Caff', 'User_Cannabis', 'User_Choc', 'User_Coke', 'User_Crack',
           'User_Ecstasy', 'User_Heroin', 'User_Ketamine', 'User_Legalh',
'User_LSD', 'User_Meth', 'User_Mushrooms','User_Nicotine', 'User_Semer',
'User_VSA']
```

[12]
```
# encoding into numeric data for the original data set
    from sklearn.preprocessing import LabelEncoder
    for column in columns:
    le = LabelEncoder()
    df[column] = le.fit_transform(df[column])
```

[13]
```
# encoding into numeric data for the copied dataset
    for column in columns:
    le = LabelEncoder()
    copy_df[column] = le.fit_transform(copy_df[column])
```

[14]
```
#dataset after encoding
    df[columns]
```

[15]
```
#make a new column for each drug which contain the information that a
user is using that drug or not
    for i in range(len(columns)):
    copy_df.loc[((copy_df[columns[i]]==0) |
(copy_df[columns[i]]==1)),cp[i]] = 'Non-user'
    copy_df.loc[((copy_df[columns[i]]==2) | (copy_df[columns[i]]==3) |
(copy_df[columns[i]]==4) | (copy_df[columns[i]]==5) |
(copy_df[columns[i]]==6)),cp[i]] = 'User'
```

[16]
```
#Visualization of the total amount of consumption for each drug
    fig, axes = plt.subplots(5,3,figsize = (16,16))
    fig.suptitle("Count of Different Classes Vs Drug",fontsize=14)
```

```
k=0
for i in range(5):
for j in range(3):
     sns.countplot(x=columns[k], data=copy_df,ax=axes[i][j])
     k+=1

plt.tight_layout()
plt.show()
```

[17]
```
#counting the total number of user or non user.
count_of_users = []
count_of_non_users = []
```

[18]
```
for i in range(len(columns)):
s = copy_df.groupby([cp[i]])[columns[i]].count()
count_of_users.append(s[1])
count_of_non_users.append(s[0])
```

[19]
```
#visualization of total user and non user of drug for every specific drug
trace1 = go.Bar(
x=columns,
y=count_of_users,
name='User',
marker = dict(color="rgb(117, 127, 221)")
)
trace2 = go.Bar(
x=columns,
y=count_of_non_users,
name='Non-User',
marker = dict(color="rgb(191, 221, 229)")
)
data = [trace1, trace2]
layout = go.Layout(
title= 'Drug Vs User Or Non-user',
yaxis=dict(title='Count', ticklen=5, gridwidth=2),
barmode='group'
)
fig = go.Figure(data=data, layout=layout)
py.iplot(fig, filename='grouped-bar')
```

[20]
```
#visualization of popluation of drug addicted for specific countries on world map
df['Country'].value_counts()
con = ['UK','USA','Canada','Australia','Ireland','New Zealand']
```

[21]
```
data = [dict(
        type='choropleth',
        locations = con,
        locationmode='country names',
        z=(df['Country'].value_counts().values),
        text=con,
        colorscale='portland',
        reversescale=True,
   )]
   layout = dict(
    title = 'A Map About Population of Drug Addicted in Each Country',
    geo = dict(showframe=False, showcoastlines=True,
projection=dict(type='Mercator'))
   )
   fig = dict(data=data, layout=layout)
   py.iplot(fig, validate=False, filename='world-map')
```

[22]
```
#Violin plot of Age vs Nscore
    plt.figure(figsize=(12,5))
    sns.violinplot(x='Age', y='Nscore', data=df)
    plt.title('Violin plot of Age by Nscore',fontsize=14)
    plt.xlabel('Nscore',fontsize=13)
    plt.ylabel('Age',fontsize=13)
    plt.show()
```

[23]
```
#violin plot of Age vs Impulsive
    plt.figure(figsize=(12,5))
    sns.violinplot(x='Age', y='Impulsive', data=df)
    plt.title('Violin plot of Age by Impulsive',fontsize=14)
    plt.xlabel('Impulsive',fontsize=13)
    plt.ylabel('Age',fontsize=13)
    plt.show()
```

[24]
```
#heatmap to find the correlation between different Features.
    corrmat = df.corr()
    plt.figure(figsize=(20,20))
    sns.set(font_scale=1)
    hm = sns.heatmap(corrmat,cmap = 'RdYlGn',annot=True,
              yticklabels = df.columns, xticklabels = df.columns)
    plt.xticks(fontsize=13,rotation=50)
    plt.yticks(fontsize=13)
    plt.title("Correlation B/W Different Features",fontsize=18)
    plt.show()
```

[25]
```
#Model Building
```

```
   yp = []
   for i in df['Benzos']:
    if(i==0):
        yp.append([1,0,0,0,0,0,0])
    elif(i==1):
        yp.append([0,1,0,0,0,0,0])
    elif(i==2):
        yp.append([0,0,1,0,0,0,0])
    elif(i==3):
        yp.append([0,0,0,1,0,0,0])
    elif(i==4):
        yp.append([0,0,0,0,1,0,0])
    elif(i==5):
        yp.append([0,0,0,0,0,1,0])
    elif(i==6):
        yp.append([0,0,0,0,0,0,1])
   yp = np.array(yp)
```

[26]
```
#yp data frame
    yp
```

[27]
```
#feature selection
    feature_col_names = ['Age', 'Gender', 'Education', 'Country',
'Ethnicity', 'Nscore',
       'Escore', 'Oscore', 'Ascore', 'Cscore', 'Impulsive', 'SS']
    predicted_class_names = ['Benzos']

   X = df[feature_col_names].values
   y = df[predicted_class_names].values
```

[28]x

[29]y

[30]yp

[31]
```
#splitting the data in test data and train data
    from sklearn.model_selection import train_test_split
    X_train, X_test, y_train, y_test = train_test_split(X, yp,
test_size=0.30, random_state=42)
    X_train1, X_test1, y_train1, y_test1 = train_test_split(X, y,
test_size=0.30, random_state=42)
```

Artificial Neural Network

[32]
```
#building a neural network
    num_inputs = len(X_train[0])
    hidden_layer_neurons = 13
    np.random.seed(4)
    #setting weights and bias for hidden layer
   b1 = 2*np.random.random(num_inputs) -1
   w1 = 2*np.random.random((num_inputs, hidden_layer_neurons)) - 1
   w1
```

[33]
```
NameError                                 Traceback (most recent call last)
<ipython-input-5-b12013d34c52> in <module>()
      1 #building a neural network
----> 2 num_inputs = len(X_train[0])
      3 hidden_layer_neurons = 13
      4 np.random.seed(4)
      5 #setting weights and bias for hidden layer
NameError: name 'X_train' is not defined
```

[34]
```
#setting the weights for output layer
    num_outputs = 7
    b2 = 2*np.random.random(num_inputs) -1
    w2 = 2*np.random.random((hidden_layer_neurons, num_outputs)) - 1
   w2
```

[35]
```
def draw_neural_net(ax, left, right, bottom, top, layer_sizes):
    n_layers = len(layer_sizes)
    v_spacing = (top - bottom)/float(max(layer_sizes))
    h_spacing = (right - left)/float(len(layer_sizes) - 1)
    # Nodes
    for n, layer_size in enumerate(layer_sizes):
        layer_top = v_spacing*(layer_size - 1)/2. + (top + bottom)/2.
        for m in range(layer_size):
            circle = plt.Circle((n*h_spacing + left, layer_top -
m*v_spacing), v_spacing/4.,
                                color='w', ec='k', zorder=4)
            ax.add_artist(circle)
    # Edges
    for n, (layer_size_a, layer_size_b) in enumerate(zip(layer_sizes[:-1],
layer_sizes[1:])):
        layer_top_a = v_spacing*(layer_size_a - 1)/2. + (top + bottom)/2.
        layer_top_b = v_spacing*(layer_size_b - 1)/2. + (top + bottom)/2.
        for m in range(layer_size_a):
```

```
            for o in range(layer_size_b):
                line = plt.Line2D([n*h_spacing + left, (n + 1)*h_spacing +
left],
                                  [layer_top_a - m*v_spacing, layer_top_b
- o*v_spacing], c='k')
                ax.add_artist(line)
```

[36]
```
#plotting the neural network
    fig = plt.figure(figsize=(12, 12))
    ax = fig.gca()
    ax.axis('off')
    draw_neural_net(ax, .1, .9, .1, .9, [12, 13, 7])
```

[37]
```
# sigmoid function representation
    xp = np.linspace( -5, 5, 50 )
    yp = 1 / ( 1 + np.exp( -xp ) )
    plt.plot( xp, yp )
```

[38]
```
#back propagation for reducing error
    error = []
    b1=0
    b2=0
    learning_rate = 0.2 # slowly update the network
    for epoch in range(1000):
    l1 = 1/(1 + np.exp(-(np.dot(X_train, w1) + b1))) # sigmoid function
    l2 = 1/(1 + np.exp(-(np.dot(l1, w2) +b2 )))
    er = (abs(y_train - l2)).mean()
    l2_delta = (y_train - l2)*(l2 * (1-l2))
    l1_delta = l2_delta.dot(w2.T) * (l1 * (1-l1))
    w2 += l1.T.dot(l2_delta) * learning_rate
    w1 += X_train.T.dot(l1_delta) * learning_rate
    error.append(er/(epoch*0.1))
    print('Error:', er)
    #we updated the weights and biases of different layers
```

[39]
```
er = set(error)
```

[40]
```
#epoch vs error graph after back propagation
    sp = pd.Series(error)
    sp.plot()
    plt.title("Epoch Vs Error Rate",fontsize=13)
    plt.xlabel("Epoch")
```

[41]
```
#normalizing the data between 0 & 1.
    from sklearn.preprocessing import StandardScaler
    sc = StandardScaler()
    X_train = sc.fit_transform(X_train)
```

```
    X_test = sc.transform(X_test)
```

[42]
```
x_train
```

[43]
```
#imporing the libraries required for model building
    import keras
    from keras.models import Sequential
   from keras.layers import Dense, Activation, Dropout
   from keras.preprocessing.text import Tokenizer
```

SVM classification

[44]
```
from sklearn.svm import SVC
svm = SVC(kernel="rbf", C=2,random_state=0)
svm.fit(X_train1, y_train1.ravel())
```

[45]
```
x_train1
```

[46]
```
x_test1
```

[47]
```
y_train1
```

[48]
```
y_test1
```

[49]
```
import sklearn.metrics as metrics
```

[50]
```
pred = svm.predict(X_test1)
    accu = metrics.accuracy_score(y_test1,pred)
    accu
```

SVM for binary Classification

[51]
```
feature_col_names = ['Age', 'Gender', 'Education', 'Country',
'Ethnicity', 'Nscore',
       'Escore', 'Oscore', 'Ascore', 'Cscore', 'Impulsive', 'SS']
  predicted_class_names = ['User_Benzos']

  X = copy_df[feature_col_names].values
  y = copy_df[predicted_class_names].values
  X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.30, random_state=42)
```

[52] x

[53] y

[54]
```
svm = SVC(kernel="linear", C=1,random_state=0)
 svm.fit(X_train, y_train.ravel())
```

[55]
```
pred = svm.predict(X_test1)
accu = metrics.accuracy_score(y_test,pred)
print("\nAcuuracy Of the Model: ",accu,"\n\n")
```

Decision Tree for Binary Classification

[56]
```
from sklearn.tree import DecisionTreeClassifier
clf_dtc =
DecisionTreeClassifier(criterion='gini',max_depth=4,random_state=0)
clf_dtc.fit(X_train, y_train.ravel())
```

[57]
```
pred = clf_dtc.predict(X_test1)
accu = metrics.accuracy_score(y_test,pred)
accu
```

[58]
```
from plotly.offline import iplot
```

[59]
```
import plotly.graph_objs as go
labels = ["SVM","Decision Trees"]
usage = [69.99,69.08]
data = [go.Bar(x = labels,y = usage)]
fig = go.Figure(data=data)
iplot(fig)
```

[60]
```
from google.colab import drive
drive.mount('/content/drive')
```