A Supervised Machine Learning Approach to Predict Vulnerability to Drug Addiction

Fahim Faisal
15201001
Arif Shahriar
15201002
Sohan Uddin Mahmud
15201006
Rakibul Alam Shuvo
15201025

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
August 2019

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Arif Shahriar 15201002	Fahim Faisal 15201001
Sohan Uddin Mahmud 15201006	Rakibul Alam Shuvo 15201025

Approval

The thesis/project titled "A Supervised Learning Approach to Predict Vulnerability to Drug Addiction" submitted by

1. Fahim Faisal (15201001)
2. Arif Shahriar (15201002)
3. Sohan Uddin Mahmud (15201006)
4. Rakibul Alam Shuvo (15201025)

Of Summer, 2019 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on August 7, 2019.

Examining Committee:

Supervisor:
(Member)

Amitabha Chakrabarty, PhD
Associate Professor
CSE Department
BRAC University

Program Coordinator:
(Member)

Jia Uddin, PhD
Associate Professor
CSE Department
Brac University

Head of Department:
(Chair)

Mahbubul Alam Majumdar, PhD
Professor and Chairperson
Department of Computer Science and Engineering
Brac University

Ethics Statement (Optional)

As we had to collect primary data from drug addicts from different institutes, drug addiction treatment centre, detoxification centre, university, college and job holders of different corporate offices, so we are committed that all the respondent's name will be kept anonymous.This data will be used only for research purpose.

Abstract

There are significant amount of differences between an addicted and non-addicted person on their social and familial behavior. In our thesis we tried to find out the characteristics of a person related to his social and familial life and also health issues that can prove his vulnerability to drug addiction. The research was held on the context of the people of Dhaka, Bangladesh and on an age group of 15 to 40 years. A primary data set was constructed which include 498 samples. For constructing the questionnaire Addiction Severity Index and WHO's Assist Scale were followed along with the help of psychologists and specialists on drug addiction. For addicted person's data we reached some rehabilitation center of Dhaka and for non-addicted person's data we communicated different aged group people of different colleges and universities. 498 samples where one sample consisted of 60 features were trained and tested by supervised machine learning approach. Reliability of the data set was validated by Cronbach's Alpha Nominal Test. 10 algorithms were incorporated including Neural Network, Deep Belief Network, Random Forest, XG-Booster etc. and their results were compared. Among the algorithms, XGB came up with the highest number of accuracy of 95.20% and KNN delivered the least which is 88.97%. In order to select important features mRMR, Chi-square, Principle Component Analysis techniques were used. From feature selection we got the key features of an addicted person's behavior that were influential for their drug abuse. This will help people to understand if a person is going to be vulnerable to addiction or not based on their health issues and social and familial behavior.

Keywords: Primary Data; mRMR; Deep Beilief Nerwork; Reliability; Vulnerability to Addiction; Neural Network; Random Forest

Dedication (Optional)

This research is dedicated to the individuals who are suffering from Substance Abusive Disorder and who also have a small spark of light that inspires them to come out and recover from this darkness.It is also a small tribute to the psychologist and addiction counselors,professionals who are trying to help people to stop from entering this severe addiction phase.

Acknowledgement

First and foremost, we want to thank Allah for His great support which helped us to continue our research without facing any kind of major difficulties. Furthermore,we wanted to thank all the helpful faculty members and specially our supervisor for tolerating our mistakes and giving continuous feedback to improve our work. And, We also want to thank our parents and peers who gave us enormous support throughout the semester.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

Drug addiction is a problem that every county of the world need to worry about. Every country has its own kind of extent and individualities to this problem. It has relation with the social and familial behavior, standards. It causes mental and physical damage of a person too. Our main vision was to find out those relations on the basis of learning data. For example, men and women who consumed alcohol are found to be more aggressive towards others [1]. Again, bonding with friends and family reportedly creates impact on a person to get into smoking [2]. These prove that peoples' social and behavioral concerns are somehow connected with and influenced by drug. Yes, drug does lot damage to the health but it also cause damage to person's private and public life. So we have though that we can predict if a person is connected to drug abuse or not by observing their daily social, familial actions and health issues as his or her social activities, different consequences of day-to-day life with people along with health issues can potentially indicate his or her openness to different types of drugs. We have gone through many studies that performed related work on drug abuse. For example, One study found out sources of referral for prescription opioid admission to substance use disorder treatment facilities and their relative completion success rates using secondary analysis of an existing data set (treatment episode datasets—discharge) [3]. Their data-set's variables are mainly focused on the frequency of taking drugs, specific group of drugs, specific drug's reporting time etc. However, their dataset played down the socio-cultural variables like bonding with family and friends, relation with peers, social behaviors etc. that we will mainly focus on our work. We had worked with the shortcomings of their works as well as many other works were investigated before performing the job which will be discussed in different section of the report.

1.2 Problem Statement

In this era of 2019, huge number of people are suffering from addiction problem. Most of them are young and teenagers. These populations are becoming a burden for our Society rather than being regarded as manpower. This population needs help from us to come back a productive life. We wanted to establish a machine learning based model with which we can easily identify a person's vulnerable stage of drug addiction. This stage may indicate that this person is in need of help. Again, Drug

abusing is social problem. Every parent, guardians of a drug abuser are concerned with this problem. Also, governments of every countries of the world are fighting this in order to eradicate it. However, addiction is not a crime but it is internationally acknowledged as a disease by addiction professionals. So, like other diseases, such as-diabetes, dengue, cancer stages can be measured it could be also measured using some social, physical, mental, family relationship-based indicators. We have decided to contribute in this sector which might help the government as well as parents and other people who are concerned with the problem.

1.3 Aim of Study

Predicting the public and private life behaviors of a drug abuser is the aim of our study. In our thesis we have discussed how we managed to catch out the relation between these social, familial, health issues and drug addiction. We have generated our questionnaire based on the common social, familial behavior of a drug user. Our dataset which is the outcome of our questionnaire is consisted of 60 features assisted to bring out the findings. All these features are in fact answer of numerous questions related to a person's social life, personal life, and familial life and health issues. Then, most important features were found which are close related with the behavioral pattern of drug addicted person. Thus, we figured out the features which told us that which kind of behavioral factors can tell us about a potential drug abuser.

1.4 Research Methodology

Our target is to locate the most potential individual who can turn into a drug addict in the near future. Having this objective, we have gathered information from different illicit drug users from the city to accumulate data about drug addicts. We make a questionnaire which concentrates on the important factors relating with the drug addiction. Information have been assembled from different sources including colleges, recovery centers and dependence treatment centers by utilizing both quantitative and subjective essential information accumulation techniques. In the wake of gathering information we do unwavering quality examination by utilizing Cronbach's Alpha nominal Test for approval. We utilize various number of features to guarantee that each factor is contacted. In addition to that, to choose the most significant highlights, we use feature selection algorithm like Minmum redundancy maximum relevance (mRMR) . Guideline Component Analysis (PCA) . Also we have used chi square test to decide characterizing features among the total dimensions. This is done to ensure that we get the most significant highlights for anticipating the helplessness to illicit drug use. We likewise use T-distributed SNE for envisioning the dataset. We used sequential based ensemble based model like Adaboost, sequential ensemble based method like Random Forrest, and tree based algorithms like Decision Tree. Deep belief Network has been used to probabilistically reconstruct its inputs to act as feature detectors and also supervised to perform classification. Lastly for plotting and visualizing data we used XGBoost algorithm. In the later part of this report, model have been compared to find the most suitable model for our prediction.

1.5 Thesis Outline

This report puts influence on constructing a prediction model which would be beneficial in detecting a potential addict in primary stage. The aim of the authors is to formulate a dataset from the context of our country which would be used for training existing supervised machine learning models to classify new observations. The overall report focuses on the steps that were followed by the researchers.

Firstly, introduction part (Chapter 1) states the motivation behind research which inspired authors to address this particular problem statement. The goals of our research and summary of the work is briefly discussed here.

In Literature review section (Chapter 2), we have discussed about papers from computer science background which have addressed similar issue. In addition to that, some statistical and psychological papers are mentioned which refers to the available secondary data. The purpose of background study was finding out the short comings of previous researches. Moreover, we have stated our contribution and reasons behind primary data collection.

In the data collection phase (Chapter 3), we have explained why we have used primary data instead of available secondary data. This portion also included a description of dataset. We also emphasized on the reliability and consistency of our generated dataset. Feature selection argued how the huge number of features can be reduced to decrease time complexity. Feature analysis focused on importance and significance of indicators with respect to outcome 'Flag'.

Furthermore, Model Selection (Chapter 4) includes our proposed models and comparative study of the prediction rate among respective models. Analysis of both traditional and advanced algorithms in case of our constructed data set are discussed in this section. Furthermore, outcomes are summarized along with visual representation to specify which model performs better for our data set.

Chapter 2

Related Work

Drug abuse is an occurrence that can be stimulated by many changes in life. We have found many works that had been done in the sector of drug abuse. In one paper, they have found out the relationship between addictive drug use and their short term and long term consequences in social behavior [1]. The use of addictive drugs can have profound short- and long-term consequences on social behaviors. For example they have found out that, Perturbations in the social environment, particularly during early development, can increase the vulnerability to drug abuse later in life. Similarly, they also said that social experiences and the presence or absence of social attachments during early development and throughout life can significantly effect drug consumption and the vulnerability to drug abuse. They focused on some of the most common addictive drugs such as psychostimulants, opiates, alcohol and nicotine. And they found relation with the following social behaviors: maternal, sexual, play, aggressive and bonding behaviors. The main purpose of our study is also to find out the correlation of an addicted person's social life, personal life and health issues.

In another paper, K. Kobus had vigorously reviewed different theoretical frameworks and empirical findings [2]. He specially put emphasis on peer influence on teenage smoking. For example, he said teenagers found an internal self-pressure to smoke if they see another friend is smoking in front of them in the paper. A lot of theoretical framework were discussed in the paper, specifically, social learning theory, social identity theory, primary socialization theory and social network theory. His empirical finding includes peer influence and selection as well as multiple reference points in different types of adolescent friendship. Best friendship, romantic friendship, peer group and social group were reviewed one by one. For instance, in his paper it was found that teenagers have been found to monitor or change their behavior or appearance, including 'trying on a cigarette' to expose the desired image. Such image exposes seem to be more important for teenage girls because they wanted to be attractive to boys and appealing to other girls. However, his study was based on theory and empirical knowledge where we did take the help of mathematical and learning based model.

We have seen many papers to put importance on cigarette smoking as a great impactor to drug abuse. In a paper, Writers studied the relation between alcohol, cigarette and drug [4]. They said, cigarette smoking is closely related with other

drug abuse and cigarette is very common who use to take drugs. For that, a significant number of questions of our questionnaire are related with cigarette smoking.

Proof showing the impacts of drug use and abuse on stress reactions and dopamine transmission is presented, along with adjusted rabid and persuasive reactions connected with wanting and slip to drug use [5]. In this report a progression of populace based and epidemiological examinations have recognized explicit stressors and individual-level factors that are prescient of substance use and misuse. Preclinical research additionally demonstrates that pressure introduction improves sedate self-organization and reestablishes medication looking for in medication experienced creatures. The pernicious impacts of early life stress, tyke abuse, and collected affliction on adjustments in the corticotrophin discharging factor and hypothalamic-pituitary-adrenal hub (CRF/HPA), the extra hypothalamic CRF, the autonomic excitement, and the focal noradrenergic frameworks are additionally exhibited. The authors focused only to the specific factors of drug addiction like stress, emotions and adaptive behavior. Moreover they found out that there is significant proof from populace based and clinical examinations supporting a positive relationship between psychosocial affliction, negative effect, and constant pain and fixation powerlessness.

In another paper, they test whether combined presentation to such stressors fundamentally predicts danger of DSM-IV liquor reliance issue in youthful grown-ups [6]. They additionally give engaging information that describes the examples of combined introduction to such occasions and paces of liquor reliance crosswise over sex, race/ethnic, and financial gatherings. They focused only on the social factor that is related to the drug addiction. In the previous paper, they made their sample in such a way so that every ethnically diverse people is been included in that particular area. On the contrary, we emphasize only drug addicted people and the factors that is strongly connected with the drug addiction. Moreover, we have used machine learning approach and algorithms to predict the person who will possibly be a drug addicted, whereas they signified the statistical analysis on alcohol dependence over a year.

Progresses in measurable techniques were depicted for compensative action inquire concerning with a selected spotlight on substance abuse prevention in this paper[7]. The foremost well known model for the estimation of program consequences for an identical ward live is that the restrictive linear regression model. They used multilevel analysis or random coefficient modeling to appropriately analyze clustered data . They also use LGM (Longitudinal Models) to quantify, clarify, and portray person contrasts in modification overtime. Also they use Survival Analysis in substance misuse anticipation to investigate starting of drug use [8]. The complexness of the statistical methods connected in substance misuse analysis is empowering. Multilevel models take under consideration the consolidation of impacts at varied levels, for instance, college and network even as individual impacts. These models take under consideration some fascinating trial of impacts crosswise over levels like the impact of network, school, and homeroom on individual substance use. The motivation behind this study was to administer a review of some advances in statistical methods for substance abuse prevention [9]. This thesis portrays progresses in factual techniques for counteractive action examine with a specific spotlight on

substance misuse avoidance. Standard investigation strategies are stretched out to the run of the mill research structures and qualities of the information gathered in counteractive action look into. Anticipation investigate frequently incorporates longitudinal estimation, bunching of information in units, for example, schools or centers, missing information, and clear cut just as persistent result factors. Factual strategies to deal with these highlights of aversion information are sketched out. Improvements in intervention, balance, and execution investigation take into consideration the extraction of increasingly point by point data from an aversion study.

In one paper the authors found out the substance abuse treatment in college student and non-college student. Where they use prediction model in form of Linear or Logistic Regression which extracts features properly between the two variables [10]. They concluded that treatment providers appeared to have superior results retaining students in smaller periods. Recommendations for higher education treatment engagement were discussed in their paper.

Another paper showed the statistical analysis of more than 10 algorithms [11]. They had demonstrated the statistical differences between many algorithms on the basis of accuracy measures. From their result, the top three algorithms with best accuracy are Super Learning, Random Forrest and Lasso All Predictors. They said, super learner is a method which run few learning algorithm inside it which might be the reason it provided the best result. However, random forest was also close to the super learner result though it was singly run on that study. In our thesis we have used random forest, deep super learner which shoed high accuracy in prediction.

Artificial Neural Network and Support Vector Machine are compared in a classification problem between drug and non-drug in early phase virtual compound filtering and screening [12]. In the paper, we have seen that SVM training provided a more robust result in terms of training with very smaller standard of error with the comparison to ANN. Support vector machine gave more accuracy than artificial neural network in every aspect of training data sets, molecule encoding and algorithm employed for neural network training. They have used 120 standard Ghose-Crippen fragment descriptors, a varied choice of 180 different things and physicochemical descriptors from the Molecular Operating Environment (MOE) package, and 225 topological pharmacophore (CATS) descriptors in order to make the comparison. However, we have compared 10 algorithm with each other for accuracy measures, features selection, confusion matrix result etc.

Besides, Neural Network has been used many times to predict drug's mechanism of action, drug content and hardness of intact tablets [13]. A total of 10 ANN correction models (5 each with 10 and 160 inputs at suitable wavelengths) and five isolated 4-factor incomplete least squares (PLS) correction models were spawned to predict drug substances of the test tablets from the shadowy data. Another paper also used ANN to predict drug's mechanism of action on the basis of its s pattern of activity against a panel of 60 malignant cell lines in the National Cancer Institute's drug screening program [14]. Their result has concluded many successful factors for example, (1) the cell line answer configurations are rich in info about mechanism. (2) Correctly designed neural networks can effectively use of that information. (3)

Trained networks can be used to categorize prospectively the more than 10,000 agents per year tested by the screening program.

In another paper they showed how they use Decision Tree algorithm to determine the chemical, physical, and structural properties of compounds that predispose them to causing ADRs [15]. A structure–activity relationship analysis was presented consisting of adverse drug reactions (ADRs) in the central nervous system (CNS), liver, and kidney, and also of allergic responses for classifying drugs that could be suspected of producing adverse reactions. With the help of a machine learning approach and decision tree algorithm they determined the chemical, physical, and structural stuffs of compounds that incline them to causing ADRs.

Finally, a machine learning based drug toxicity prediction research used the most commonly used machine learning algorithms like: Support Vector Machine, Random Forest, K-Nearest Neighbor, Naive Bayesian, Neural Network and Ensemble Learning [16]. Freely accessible data sets were used for drug toxicity prediction for building an machine learning model, each atomic descriptor and every piece of the fingerprints can fill in as an autonomous variable also known as a 'feature' within the extent of machine learning. Some valuable programming tools for instance, R, Weka, Python, and a few valuable QSAR modeling software for instance, KNIME, RDkit, provide executions of the machine learning algorithms that are widely use to model drug toxicity prediction.

Chapter 3

Data Collection and Feature Selection

3.1 Data Collection

In previous years, addiction problem have been addressed as a significant subject in diverse research fields. Our target is to incorporate the machine learning concepts in this research field. In the new era of artificial intelligence machine learning approaches are being used to target social problems. As they yields better accuracy and considered well suited to address specific problems we are using data science and machine learning approaches to produce a recommendation system for predictive analysis. In existing papers, the researchers have tried to find out the major attributes responsible for causing addiction problem but only few of them have tried to identify the problem in primary stage. However, our target in this research is to find out the persons who are prone to substance abuse and to design a system which will facilitate early detection of the problem. In order to answer the discussed research problem, we have used information (data) from all possible resources. We wanted to focus on the issue in context of our country. Since the crucial reasons and factors behind addiction problem may vary from society to society, we have tried to find out the features that result in vulnerability to substance dependence from the viewpoint of our country. There are two established methods of collecting data-secondary data collection and primary data collection. The purpose of this research work is the reconnaissance of drug addiction and to develop a prediction model of early detection of the problem. Data have been gathered from all possible sources including universities, rehabilitation centers and addiction treatment centers by using both quantitative and qualitative primary data collection methods. The collected data set is presented in [17]. However, there exists some secondary data-sets, which are not sufficient to develop our prediction model. In the dataset named table1.2, we get some important information. But this dataset focused mainly on 'substances'. It mainly focused on primary drugs, drug abuse along with alcohol consumption, Abusers without primary substances from a timeline representing 2004 to 2014 usage information. But we wanted to identify important social, mental, physical states and indicators that has significant contribution to increase vulnerability towards drug abuse.

Another data set found from HHS.gov official website [18] (Office of adolescent

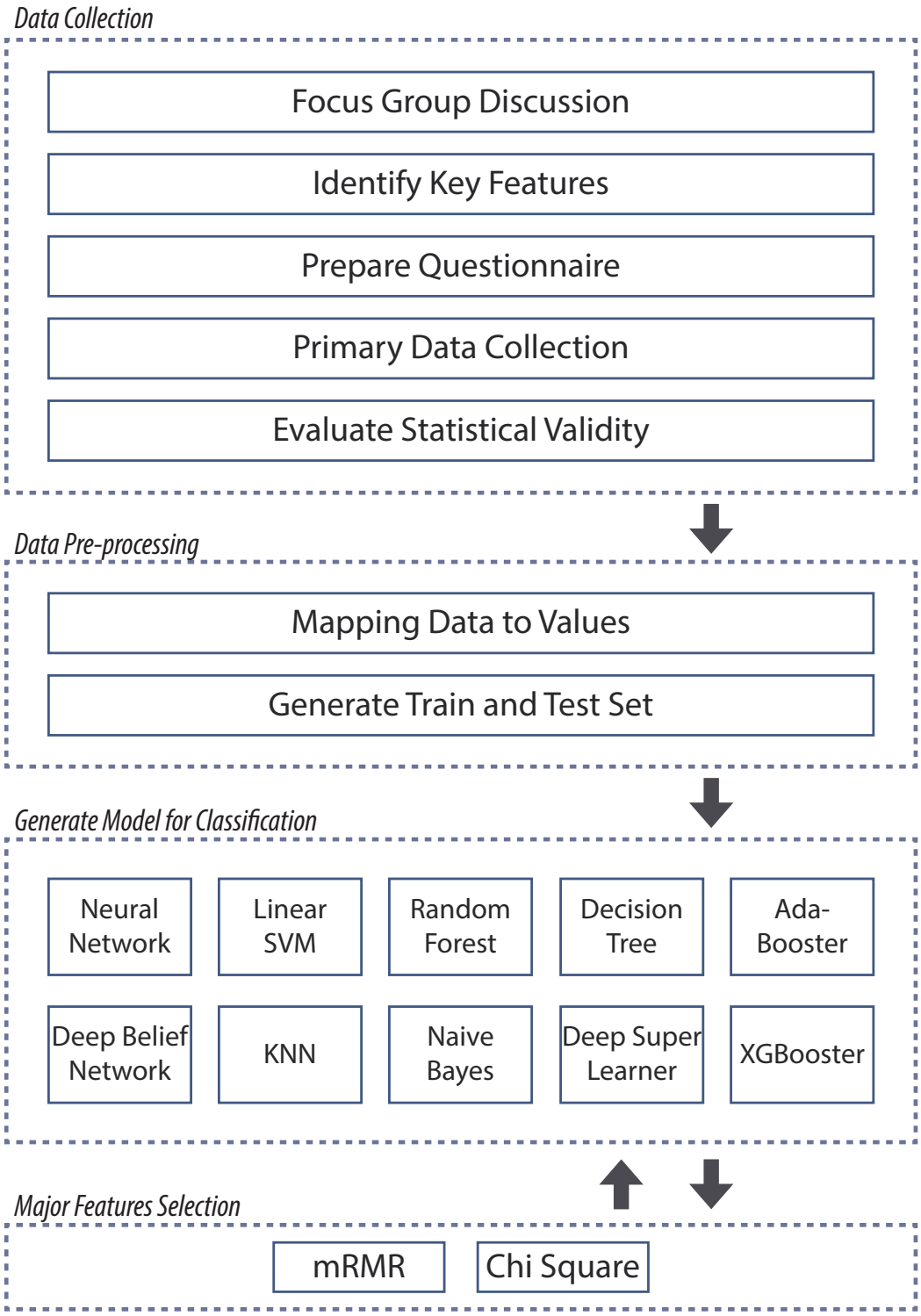


Figure 3.1: Overall system modules of the ML based proposed prediction model for vulnerability to drug addiction

Health) focused on e-cigarette usage , quit behavior , alcohol consumption related behavior, Cocaine usage ,Marijuana usage related behaviors among male and female high school students. In [19,20] database, author focused on 12 attributes which includes personality measurement, level of education, age, gender, country and ethnicity, use of 18 legal and illegal drugs. However, they did not cover familial, social factors and mental health issues rather they emphasized on personality traits. 'Percentage of 12th-graders who used any illicit drug in the past 30 days by sex and race/ethnicity, 2001–2014' data set focused on illicit drug use affected by sex, race, subculture among a time period. But from the perspective of our country, cultural factors have a great significance on drug abuse but subculture is not a crucial issue in our country. Our subcultures like Tribal has no crucial significance.

We have intended to use primary data collection method because it is very helpful in gathering data for a specific purpose and enables more flexibility of the researcher to customize it according to projected purpose. This type of studies consist of mathematical computations in several formats which will help us to analyze the data mathematically. There are few available methods for primary data collection-questionnaire, interviews and observations. We have used questionnaire as primary data collection method as it is a popular research instrument to get unbiased response from potential respondents and it have also helped us in getting responses from patients in rehab center by providing them anonymity and directly administering them to get better answers. In order to incorporate this method, we have used closed ended questions to obtain information and gather insight about the population. Since we have time constraint to complete our investigation we used this method which enables standardization of data and comparative study. Closed ended questions are preferable in scientific study as they can be easily interpreted into numerical data. Moreover, they are easier for programming and better for comparative analysis among multiple samples. This technique is also convenient for the respondents to answer and guarantees better understanding of questions by providing probable options. To collect our sample from population, we have chosen the treatment centers, rehabilitation centers and university students. Initially we have conducted a focus group discussion to get a better understanding about the key features and factors. By targeting the addicted patients in the treatment facilities we tried to obtain complete and precise information regarding their reasons behind addiction. The major factors that we came up with are socio-economic status, medium of education, frequency of substance abuse, money spent in buying substance, peer pressure, curiosity, family structure, lifestyle choices, social interaction and stress. According to their opinions, these are the factors that may result in addiction problem in long run. The impacts of substance dependence that we included in features are compulsive and violent behavior, emotions like guilt, anger, sadness; and feelings including suicidal and aggressive thoughts. The long and short term effects of substance abuse includes these particular aspects. Furthermore, we consulted with two counsellors of psychological unit of our university. Both of them shared their views and ideas with us regarding addiction problem and enlightened us with several perceptions to address the issue. They gave us necessary instructions concerning preparation of questionnaire. We went through many phases during preparation of questionnaire. The first phase included preparing a raw questionnaire with the help of available online resources. Moreover, the resources that we

used contains questions from EPSAD [21] (The European School Survey Project on Alcohol and Other Drugs) and other online sources. It was difficult to find out resources from renowned online sources as most of the contents regarding substance dependence require authorized access especially accessible for clinical psychologists and addiction professionals. However, we managed to gather around hundreds of closed ended questions associated with the major features that we came up with from FGD (Focus Group Discussion). Secondly, we categorized the questions according to their resemblance with the factors and ensured that each of the factors were covered thoroughly. Thirdly, we talked to our supervisor and he gave us feedbacks and recommended us to include issues associated with the context of our country. Since we have targeted this research to find the answers from perspective of our country, we added another section of questions to mainly focus this issue in context of our country. Furthermore, the counsellors from our university's psychological unit helped us to filter out the significant questions related to the derived factors. In addition to that, we consulted with clinical psychologists and addiction professionals to update and modify our questionnaire in such a way so that it would be able to extract all the all the required information precisely. As they have considerable experiences due to working in this field for prolonged period of time, they suggested us to follow ASSIST (WHO-ASSIST V3.0-BANGLA), ASI (Addiction Sensitivity Index), DMH scale and also advised us to take help from the case history which they uses to keep records of the admitted patients in rehab centers. Furthermore, they also helped us to sort the prepared questions in a logically coherent way so that the respondents feel comfortable while answering them.

The questionnaire answers are primarily stored as string values in a csv file which are later converted to numerical values using classification and scaling technique. For instance, the questions that includes binary responses can be interpreted into numeric values easily. The questions which indicates whether the sample respondent faces withdrawal symptom or not includes 'yes' or 'no' response by labelling with 0 and 1. However, the questions which consists of various categories can be classified into some explicit classes that corresponds to the specific response of the question's attribute. For example – the question that specifies the family income of the respondent includes few ranges of earnings as answers like more than 100000, 50000-100000, 20000-50000 and less than 20000 which are classified using interval data ranging from 0 to 3. The same technique is also applied for the similar type of questions like occupation of respondent. Some questions included multiple answers which were broken down into several features where each of them was treated as binary response. For example, the question which identifies whether the respondent is suffering from any mental illness or not includes responses 'Depression, guilt', 'Tension and anxiety', 'Insomnia and anger' where each of them represented individual features. We mapped them as binary response 1 or 0 depending on whether he/she has encountered the problem or not. Since we wanted to predict whether the person is vulnerable to substance abuse or not, we have mapped binary classes 'Addicted' and 'Sober' consecutively to 1 and 0. In our thesis, our main focus is to classify the instant spaces into two categories. Approximately, we have collected data about more than 60 attributes from 500 people.

In this report, our main focus is to classify instant spaces into two categories/flag.

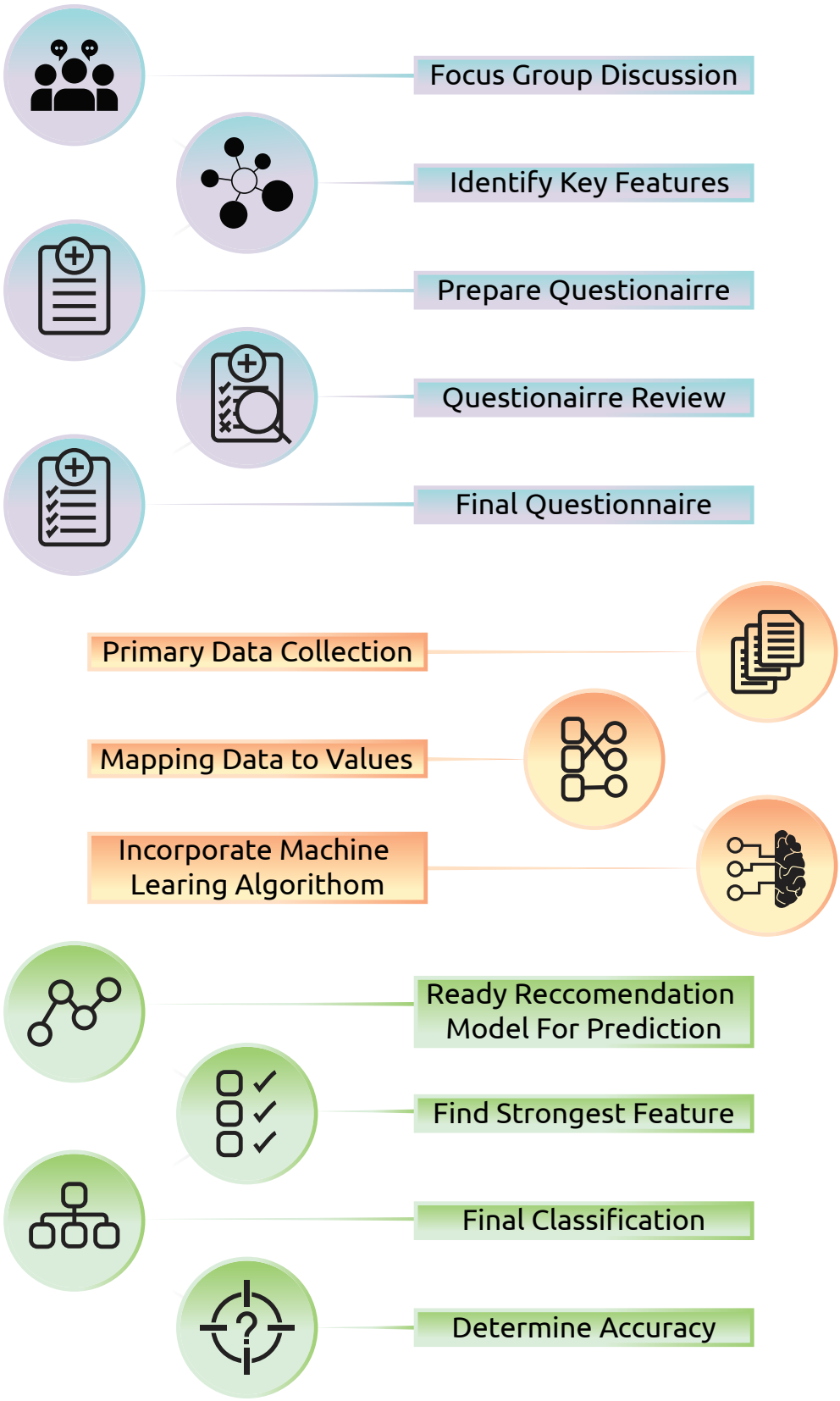


Figure 3.2: Work flow of methodology

Approximately, we have collected data about more than 50 attributes from more than approximately 498 people. As we are following questionnaire methodology, it required personal intervention and it needed direct administering the correspondents. Therefore, then number of samples are limited. After rectifying and linearizing dataset depending on attribute characteristics, we can analyze the dataset. We need to eliminate ambiguity also. In terms of scaling, we have followed WHO's Assist scale, DMH scale and ASI scale.The Addiction Severity Index was first familiarized in S in 1989 as an instrument to assess the deficiencies, changes that common between individuals who are prone to substance abuse disorder. It is the most popularly used instrument with high success rate to differentiate the addicts suffering from SUD(Substance Use Disorder).It is used in a broad range to assess severity in multiple criteria-Addiction, mental health, prison involvement, urge for treatment, homeless condition and their psychological problems pattern. This is a valid scale which can detect an addict with confirmation. We have taken the pattern of Logical order like general information section, Occupation section, Family and Social section, Physical and Mental Health related section and legal activities section which helped to keep the questionnaire in a coherent manner. Again, it helped as to identify the key indicators to be considered as attribute for dataset which will be used for learning and testing. The WHO-Assist V3 was also used .It is concerned with all the experience of lifetime, especially drug abuse history regarding past 3 months. They include smoking, drinking, inhaling and injective drugs. Also the use of sleeping pills, morphine, pethidine (painkiller) without prescription is regarded as addictive behavior. There are 8 questions. If someone is not a social/addict he/she can skip question 2- 6 if all the answers regarding substance use is 'NO'. We have taken the type of possible drug abuse from this assist. But as we are considering a majority number of non addicts for training also so we cannot use this scaling on other sectors. MH scale focuses on some Mental Health regarding questions which can assess the current mental health of a subject. And it can also differentiate addict's mental condition from a psycho patient or a mentally disordered patient. It helped us to distinguish some mental health criteria to detect specific addictive behaviors. From the DMH scale we have included some mental health related issues to reflect and identify some mental health states. Those attributes can be considered as vulnerability parameter for drug abusers which can help as to differentiate between addicts and non addicts through empirical learning and analysis approach.

Firstly, we have some questions to get information about family. It's about the subject's relationship with family members. We wanted to understand his degree of relationship, Family income to identify his relationship pattern of family. We also used relationship with spouse and tried to understand if it is more prone to addiction. We wanted to follow uniform scaling. But due to the variation in our questions' pattern we had to follow different type of questions with quantitative, informative, analytic answers. Secondly, we wanted to analyze peer pressure, relationship with peers and how much subject can be influenced by friendship. Thirdly, we have focused on physical and health issues .By removing ambiguous data such as mentally ill, psycho patients, depression patients from addicts we can ensure that it helps our machine learning process to identify individuals that are prone to addiction use disorder only. (Not any other mental health issues).

Column Serial Number	Feature Name
1	gender
2	medium of study
3	educational qualification
4	nationality
5	religion
6	family members
7	relationship with family
8	Family Income
9	Daily money
10	addiction in family origin
11	friends number
12	friend's social class
13	stay at friend's house
14	borrow money rate
15	marital status
16	ever broken up
17	stay out at night
18	live with substance abuser
19	With whom spend time
20	occupation
21	monthly avg income
22	illegal income
23	problem in workplace
23(A)	If yes, following problems
24	feel sick
25	diseases as disrupting life
26	failed in life
27	A) Depression, sorrow and hopelessness
27	B) Anxiety, irritated
27	C) Hallucinations
27	D) Lack of attention/ memorization

Table 3.1: Feature Name List Part-A(1-27)

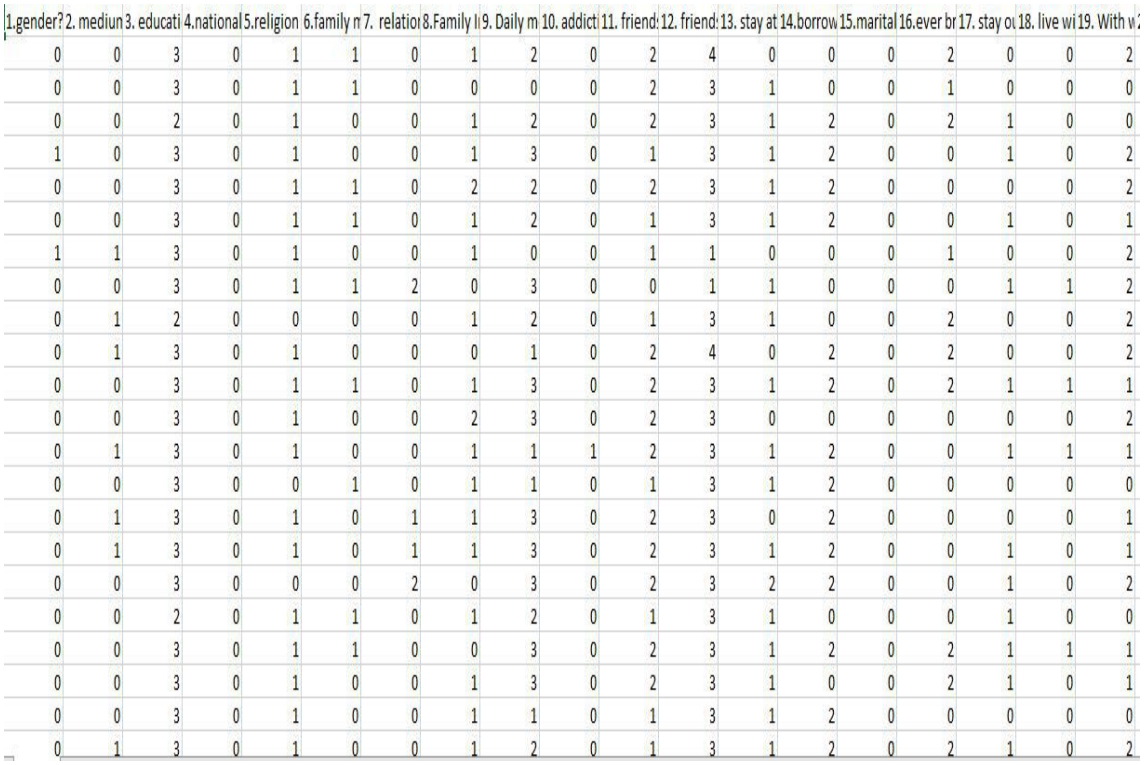
28	A) Depression, sorrow and hopelessness
28	B) Anxiety, irritated
28	C) Hallucinations, see imaginary things, hear voices
28	D) Lack of attention/ memorization problem
29	family has mental health issues
30	suicidal thoughts
31	attempt to suicide
32	difficult to maintain routine
33	ever used any substance without the doc's permission
34	A) Stimulant (Methamphetamine, Cocaine)
34	B) Sedative (Sleeping Pills, Alcohol)
34	C) Depressant (Heroin, Phencidil)
34	D) Hallucinogen (LSD, Piot)
34	D) Others (Cannabis etc.)
35	family members use substance
36	age of first substance use
37	hurt anyone or to anger
38	consulted doctor for using substance
39	smoke?
40	percentage of smoker friends
41	peer pressure to engage in drug
42	withdrawal symptoms
43	fail to fulfill social duties
44	stole money from parents
45	ever been arrested
46	case/lawsuit going on
47	arrested for keeping substance
48	arrested for selling or dealing substance
Flag	

Table 3.2: Feature Name List Part-B(28-49)

We included two images of partial view of scaled 3.3 and unscaled dataset 3.3in the above figure.We have used class based labelling with numerical numbers for scaling.We have used the name of different features with Feature Serial number. So, for the benefit of the readers we have included a table3.13.2 of all the features

1. What is	2. In whic	3. What is	4. What is	5. Mentio	6. How mu	7. How is	8. What i	9. Amoun	10. Is the	11. How n	12. Most	13. Do yo	14. Do yo	15. What	16. Have	17. Do yo	18. Do yo	19. With v	20. What	21. If you	22. Do yo	23. Do yo	23(A). If ye	24. Do yo	25. Do yo	26.
Male	Bengali	H.S.C/ A l	Banglade	Muslim	More thai	Good rela	More thai	300-1000	Yes	Few	Middle cl	No, I do n	Only in en	Unmarrie	No	No	Yes	Family	Study		No	Yes	Others	Yes		No
Male	Bengali	Undergra	Banglade	Muslim	Less than	Good rela	More thai	300-1000	No	Many	Middle cl	Sometime	No	Divorced	Yes, Some	No	No	Friends	Business	Between	No	No		No		No
Male	English	Undergra	Banglade	Muslim	More thai	Good rela	More thai	300-1000	No	Many	Higher cla	Sometime	No	Married	Yes, I did	Yes	Yes	Friends	Nothing	Between	No	No	Others	No		Yes
Male	English	Undergra	Banglade	Muslim	More thai	Good rela	More thai	300-1000	No	Many	Higher cla	Sometime	No	Married	Yes, I did	Yes	Yes	Friends	Nothing	Between	No	No	Others	No		Yes
Male	Bengali	S.S.C/ O l	Banglade	Muslim	Less than	Not good	50,000 to	Less than	No	Few	Higher cla	No, I do n	Only in en	Unmarrie	Yes, I did	No	No	Alone	Nothing		Yes	Yes	Others	Yes	Others	Yes
Male	English	Undergra	Banglade	Muslim	More thai	Good rela	More thai	300-1000	No	Few	Higher cla	Sometime	No	Married	Yes, I did	Yes	Yes	Family	Nothing	Between	No	No		No		No
Male	Bengali	Undergra	Banglade	Muslim	More thai	Good rela	Less than	Less than	No	Many	Middle cl	No, I do n	Only in en	Unmarrie	Yes, I did	Yes	Yes	Family	Study		No	No	Mental ill	Yes		Yes
Male	Bengali	Undergra	Banglade	Christian	More thai	Good rela	More thai	Nothing /	Yes	Many	Middle cl	No, I do n	No	Married	No	No	Yes	Family	Job	More thai	No	No		No	Others	No
Male	Bengali	Undergra	Banglade	Muslim	More thai	Good rela	20,000-5(	Nothing /	No	Many	Middle cl	Sometime	Only in en	Unmarrie	No	Yes	No	Alone	Business	Under 10,	No	No		No	Others	No
Male	English	Undergra	Banglade	Muslim	More thai	Satisfactc	50,000 to	Nothing /	No	Many	Middle cl	Sometime	Only in en	Unmarrie	Yes, I did	Yes	No	Alone	Job	Under 10,	No	No	Others	No	Others	No
Male	Bengali	Undergra	Banglade	Muslim	Less than	Good rela	50,000 to	More thai	No	Many	Higher cla	No, I do n	Only in en	Unmarrie	Yes, I did	No	Yes	Family	Study	Between	No	No	Others	No	Others	No
Male	Bengali	Undergra	Banglade	Muslim	More thai	Good rela	More thai	Nothing /	No	Many	Middle cl	Sometime	Only in en	Unmarrie	Yes, I did	No	No	Family	Job	Between	No	No		Yes	Diabetes	Yes
Male	Bengali	Undergra	Banglade	Muslim	Less than	Good rela	Less than	Nothing /	No	Many	Middle cl	Sometime	Only in en	Unmarrie	Yes, Some	Yes	Yes	Friends	Study		No	No		Yes		Yes
Male	Bengali	Undergra	Banglade	Muslim	More thai	Good rela	50,000 to	Nothing /	No	Many	Middle cl	Sometime	Only in en	Unmarrie	Yes, Some	Yes	Yes	Family	Nothing	Between	No	No		No	Others	Yes
Male	Bengali	Undergra	Banglade	Muslim	More thai	Good rela	50,000 to	Less than	No	Many	Middle cl	Sometime	I borrow r	Unmarrie	Yes, I did	Yes	Yes	Family	Study		No	No		Yes		Yes
Male	English	Undergra	Banglade	Hindu	Less than	Good rela	50,000 to	Nothing /	No	Many	Middle cl	Sometime	Only in en	Unmarrie	Yes, I did	Yes	Yes	Friends	Job	Between	No	No		No		No
Male	English	Undergra	Banglade	Muslim	More thai	Good rela	50,000 to	Nothing /	No	Many	Middle cl	Sometime	Only in en	Unmarrie	Yes, I did	No	Yes	Friends	Job	Between	No	Yes	Others	Yes	Others	No
Male	Bengali	Undergra	Banglade	Muslim	Less than	Satisfactc	More thai	More thai	No	Many	Higher cla	Sometime	No	Married	Yes, Some	Yes	Yes	Friends	Business	More thai	No	No	Others	No	Hepatitis	No
Male	Bengali	Undergra	Banglade	Hindu	More thai	Good rela	Less than	Nothing /	Yes	Many	Middle cl	Sometime	Only in en	Unmarrie	Yes, I did	Yes	Yes	Friends	Study		No	Yes	Others	Yes		No
Male	Bengali	Undergra	Banglade	Muslim	Less than	Good rela	50,000 to	Less than	No	Many	Higher cla	Sometime	No	Unmarrie	Yes, I did	Yes	Yes	Friends	Study	Under 10,	No	No		No	Others	No
Male	Bengali	Undergra	Banglade	Muslim	Less than	Satisfactc	More thai	Less than	No	Many	Middle cl	Sometime	Only in en	Unmarrie	Yes, Some	No	No	Family	Study		No	No		Yes		No
Male	English	Undergra	Banglade	Muslim	More thai	Good rela	50,000 to	300-1000	No	Many	Middle cl	Sometime	No	Unmarrie	No	Yes	No	Friends	Study	Under 10,	No	No		No		Yes
Male	Bengali	Undergra	Banglade	Muslim	Less than	Satisfactc	50,000 to	Nothing /	No	Many	Middle cl	Sometime	Only in en	Unmarrie	Yes, Some	No	No	Family	Study	Between	No	Yes	Mental ill	No	Others	Yes
Male	English	S.S.C/ O l	Banglade	Muslim	More thai	Satisfactc	Less than	Less than	Yes	Few	Middle cl	No, I do n	No	Unmarrie	Yes, Some	Yes	Yes	Friends	Business	Between	No	No		No	Others	Yes
Male	Bengali	Undergra	Banglade	Muslim	Less than	Good rela	50,000 to	More thai	No	Many	Middle cl	Sometime	Only in en	Unmarrie	Yes, Some	Yes	No	Alone	Study	Under 10,	No	No		No		Yes
Female	Bengali	Undergra	Banglade	Muslim	More thai	Good rela	More thai	300-1000	Yes	Many	Middle cl	Yes, too o	No	Unmarrie	Yes, I did	Yes	No	Friends	Study	Under 10,	No	Yes	Mental ill	No	Others	No
Male	Bengali	Undergra	Banglade	Muslim	More thai	Good rela	20,000-5(	Nothing /	No	Many	Higher cla	No, I do n	No	Married	No	No	No	Friends	Study	Between	No	No	Others	No	Others	No
Male	Bengali	Undergra	Banglade	Muslim	Less than	Satisfactc	20,000-5(	Less than	No	Many	Middle cl	Yes, too o	Only in en	Unmarrie	Yes, I did	Yes	Yes	Friends	Job	Between	No	No	Others	No	Others	No
Male	Bengali	H.S.C/ A l	Banglade	Muslim	Less than	Good rela	Less than	Nothing /	No	Many	Higher cla	Yes, too o	I borrow r	Unmarrie	Yes, I did	No	No	Alone	Business	Between	Yes	Yes	Mental ill	Yes		Yes
Male	Bengali	H.S.C/ A l	Banglade	Muslim	Less than	Satisfactc	More thai	Less than	No	Many	Middle cl	Sometime	Only in en	Unmarrie	Yes, I did	Yes	Yes	Friends	Study	Under 10,	No	No	Others	Yes	Others	Yes

Figure 3.3: Non-Scaled Dataset



1.gender?	2. mediun	3. educati	4.national	5.religion	6.family n	7. relatio	8.Family l	9. Daily m	10. addict	11. friend	12. friend	13. stay at	14.borrow	15.marital	16.ever br	17. stay ou	18. live wi	19. With w	2
0	0	3	0	1	1	0	1	2	0	2	4	0	0	0	2	0	0	2	
0	0	3	0	1	1	0	0	0	0	2	3	1	0	0	1	0	0	0	
0	0	2	0	1	0	0	1	2	0	2	3	1	2	0	2	1	0	0	
1	0	3	0	1	0	0	1	3	0	1	3	1	2	0	0	1	0	2	
0	0	3	0	1	1	0	2	2	0	2	3	1	2	0	0	0	0	2	
0	0	3	0	1	1	0	1	2	0	1	3	1	2	0	0	1	0	1	
1	1	3	0	1	0	0	1	0	0	1	1	0	0	0	1	0	0	2	
0	0	3	0	1	1	2	0	3	0	0	1	1	0	0	0	1	1	2	
0	1	2	0	0	0	0	1	2	0	1	3	1	0	0	2	0	0	2	
0	1	3	0	1	0	0	0	1	0	2	4	0	2	0	2	0	0	2	
0	0	3	0	1	1	0	1	3	0	2	3	1	2	0	2	1	1	1	
0	0	3	0	1	0	0	2	3	0	2	3	0	0	0	0	0	0	2	
0	1	3	0	1	0	0	1	1	1	2	3	1	2	0	0	1	1	1	
0	0	3	0	0	1	0	1	1	0	1	3	1	2	0	0	0	0	0	
0	1	3	0	1	0	1	1	3	0	2	3	0	2	0	0	0	0	1	
0	1	3	0	1	0	1	1	3	0	2	3	1	2	0	0	1	0	1	
0	0	3	0	0	0	2	0	3	0	2	3	2	2	0	0	1	0	2	
0	0	2	0	1	1	0	1	2	0	1	3	1	0	0	0	1	0	0	
0	0	3	0	1	1	0	0	3	0	2	3	1	2	0	2	1	1	1	
0	0	3	0	1	0	0	1	3	0	2	3	1	0	0	2	1	0	1	
0	0	3	0	1	0	0	1	1	0	1	3	1	2	0	0	0	0	0	
0	1	3	0	1	0	0	1	2	0	1	3	1	2	0	2	1	0	2	

Figure 3.4: Scaled Dataset

Feature Description	
Number of features	60
Class of target variables	2
Scales Followed	ASI, DSM, WHO ASSIST
Mapping values	Discrte values of 0-4 Dichotomous values of 0-1
Questionnaire Information	
Number of Participants	500
Number of Questions	49
Rehab-centers covered	5
Major factors	social behavior & status, likeness, psychological behavior, psychiatric history, rational condition

Table 3.3: Summery of data set

sequentially.

3.2 Relaiability Analysis

3.2.1 Cronbach's Alpha Nominal Test for Validation

Initially, we have used Cronbach's Alpha nominal test for ensuring the validity and consistency of our formed questionnaire. Validity indicates the meaningfulness of the measurement. It is used to calculate the degree and level of internal consistency-which means how well related the data are. It ensures that all features are positively and strongly sufficient co-variate with own selves. The coefficient depends on the type of data and relationship with target outcome. Scores more than 0.7 are considered proper. It is also affected by the number of items. Alpha is computed using eqn (3.1) where k = number of indicators and r = average correlation among all indicators. Cronbach's alpha is better for Likert scales, it is also applicable for interval level variables, dichotomous and continuously scored variables. The derived value of Standardized Cronbach's alpha score is 0.799. where, 79.9% of the variance in the scores is reliable variance. It is also called a ratio of true scored variance to total variance. Unstandardized alpha is considered from a covariance perspective where standardized Cronbach's alpha considers correlation among indicators according to the specified formula and it assumes that all of the items have equal covariances.

$$\alpha = \frac{k\bar{r}}{1 + (k-1)^{\bar{r}}} \tag{3.1}$$

Then according to Our statistical analysis Marital Status is a less important feature and it resulted in a lower value of inter item correlation which is closer to 0.1 with respect to outcome.Again, 'Lack of Attention' has a correlation value of -.074 with respect to the outcome and an inter item correlation closer to 0.01.Another confusing element of our questionnaire was 'others' option in which 'List of Substances used' by which we meant cannabis which is also used due to medical reasons and many

sobers grew up a habit of using this element.If this item was deleted It gained a scale variance - closer to approximately 114.So, after excluding these two items the value of standardized Alpha was 0.806 (Unstadnardized Cronbach's alpha up to 0.736) which is close to a satisfactory validity level. Coefficient alpha helps us to identify if a group of categorical questions/indicators can successfully summarize the total value. For an example, all the questions in the group of 'Substance use' related questions and peer pressure' from question 33-43 has a Cronbach's alpha standardized value of 0.792 for these $\mathcal{N}$=12 items. Another group containing 'legal issues' from questions 44-48 has a standardized Cronbach's alpha is 0.711 which also indicates a satisfactory level. And the group consisting of 'family and friends/Social influence' containing questions some of '6-19' has an standardized alpha value of 0.602 for $\mathcal{N}$=10 items.

3.3 Feature Selection

3.3.1 MRMR feature extraction

Since our data had large number of attributes, it was necessary to reduce dimension of data space. Our data initially had dimension $[\mathcal{N} * \mathcal{F}]$ where $\mathcal{N} = x_1, x_2 \ldots, x_n$ the number of samples in our observation and $\mathcal{F}$ denotes to $f_1, f_2 \ldots, f_n$ the number of features considered for classification of target variable. In our empirical analysis, numerical values of $\mathcal{N}$ and $\mathcal{F}$ were consecutively 498 and 60. Though there were many methods available for feature selection, we used Minimum Redundancy Maximum Relevance algorithm [22] for selecting major features based on mutual information score to reduce misclassification errors. In our dataset, all of the features do not necessarily have equal impact on target variable. So the purpose of using this algorithm was to find out the features f_i which had satisfactory mutual information with target variable c_i such that the resulting features were selected based on mutual information score $\mathrm{M}(f_i, c_i)$. The equation which was used to obtain mutual information between feature f_i and class variable c_i is stated below.

$$M(f, c) = P(f_i, c_i) * log(\frac{P(f_i, c_i)}{P(f_i) * P(c_i)}) \tag{3.2}$$

Here P indicated computed probability estimate of corresponding features. The algorithm resulted in a descending list $\mathcal{T}$ of top k features where target variable c_i was highly dependent on the features $f_1, f_2 \ldots\ldots, f_k$ belonging to $\mathcal{T}$. Initially, we obtained a list of 40 major characterizing features having significant impact on target variable. Though the set of features which had maximum relevance could have increased accuracy of classification, they tend to have redundancy between them. The set of features was obtained by calculating the average of the mutual information gained from the given equation for N samples. However, the collective probability of the features influencing the target variable would not reduce significantly if we could minimize redundancy by selecting mutually exclusive set of features . Therefore, it has tried to find out a tradeoff between maximum relevance and minimum redundancy to find out an optimal set of features which can be expressed as the maximum value of the following function.

$$mRMR(F) = max(D) - min(R) \tag{3.3}$$

The function has taken F features as parameter and returned a set of ordered features $\mathcal{T}$ where the first feature had most significant impact on label variable. It could be said that for each feature f_i belonging to $\mathcal{T}$, they have high correlation with the class variable. Moreover, they exhibit lower correlation between themselves at the same time according to the resulting order of features. After applying mRMR algorithm, we implemented classification algorithms to see whether the major features were able to predict decision variable more accurately. Support vector machine exhibited 92% accuracy, 94% sensitivity and 90.67% specificity. Furthermore, Random Forrest classifier also demonstrated improved accuracy of 94.00% with 97.01% sensitivity. From result, it can be stated that major characterizing features were able to reduce misclassification errors considerably. In addition to that, we have also implemented mRMR feature selection algorithm based on information theoretic feature selection to maximize conditional likelihood of the features. The function implemented minimum redundancy maximum relevance feature selection algorithm by returning three lists of features which included index of selected features in a hierarchical order, corresponding objective function value and corresponding mutual information between features and target variables. Corresponding objective function value was determined by calculating the correlation of the feature variables to the decision variable. The mutual information values are plotted against the feature variables in the following graph. Though we considered a subset of features including 40 major characterizing attributes, we have only plotted few of them for better visualization.

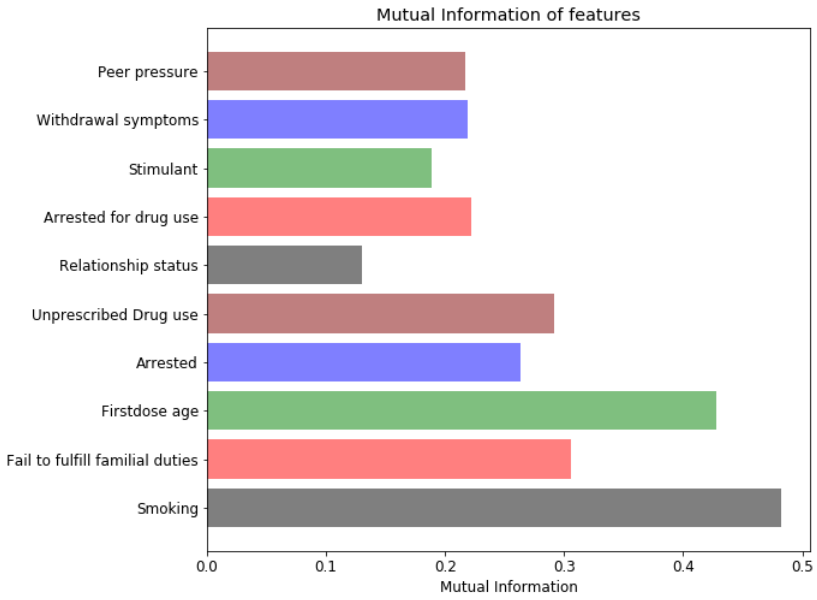


Figure 3.5: Mutual Information plot of features (different colors used for better visualization)

3.3.2 Principle Component Analysis

It was quite difficult to visualize our data set due to large number of features. Principal component analysis method was used to reduce dimension of feature space. A set of 59 features was combined to form only 2 orthogonal principle components so that we can visualize scatter plot of two independent components. The target classes demonstrated nice separation in figure 3.6 which was even linearly separable.

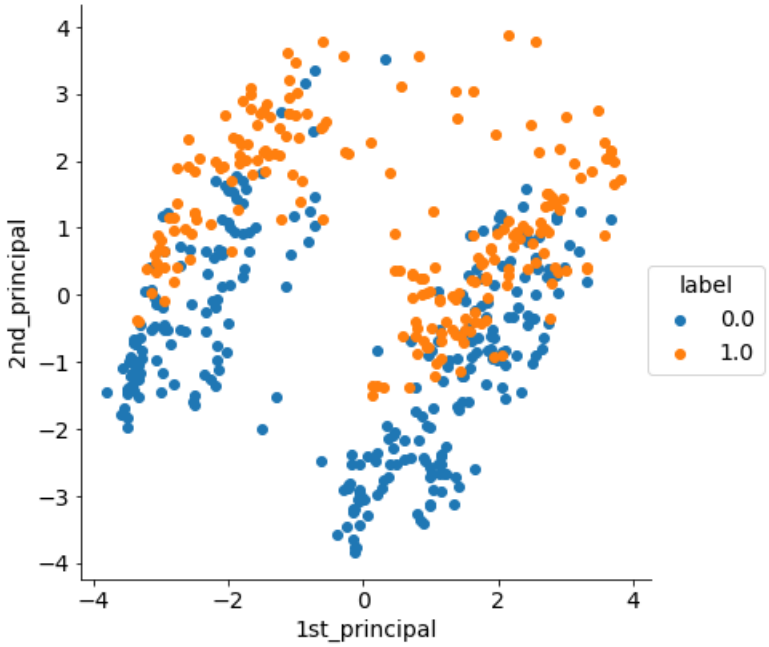


Figure 3.6: Scatter plot after principal component analysis of features

3.3.3 Chi Square test to determine dependecies

$$\Delta f = f_{observed} - f_{expected} \tag{3.4}$$

$$Chi\ square, x^2 = \sum \frac{\Delta f}{f_{expected}} \tag{3.5}$$

If data don't come in numerical value but comes in format of frequency. As different indicators has different labels we have used two way Chi square / ANOVA test. This test requires that individual observation are independent of each other and Exp frequencies should not be very small. It is used for understanding significance with pearson Chi square test. We have used it to prove that our chosen indicators of the developed questionnaire has significance on 'vulnerability to addiction'. Then main formula uses the equation given above, where in eqn (3.4) $f = frequency$. Exp cell C is calculated using another formula which is, $e = \frac{x_i + x_j}{total}$ where x_i, x_j stands for row and column numbers and *total* stands for grand total. We have used Chi square test to understand if my independent variables has significance on my dependent classification output variable .We have formed a cross table from analyzing the Exp value of the Sober and addicted in other feature's class according to null hypothesis. According to null hypothesis, we wanted to mean that this feature has no significance on being a SUD vulnerable person or not. According to Chi square feature importance, my top 10 features include peer pressure, relationship problems, urge for getting a sober life, sedative and drug abuse, drug abuse without doctor's permission and also the age of first drug abuse. Where age of first substance abuse has a great significance on affecting the outcome 'vulnerability' which is more than 0.65. For an example, for 'relationship with family, Expected Count was a way much different than Actual Count. Significance of this indicator is less than threshold probability value α =0.05 which is ignorable. So, it assures that null hypothesis

of independence of 'family relationship' and 'vulnerability flag' is false and it indicates that they are dependent and related. Its impact depends on the Pearson Chi square value50.552 and Cramer's $\mathcal{V}$ value 0.327. Dependency and significance of all the features with Flag is discussed in Table 3.4.

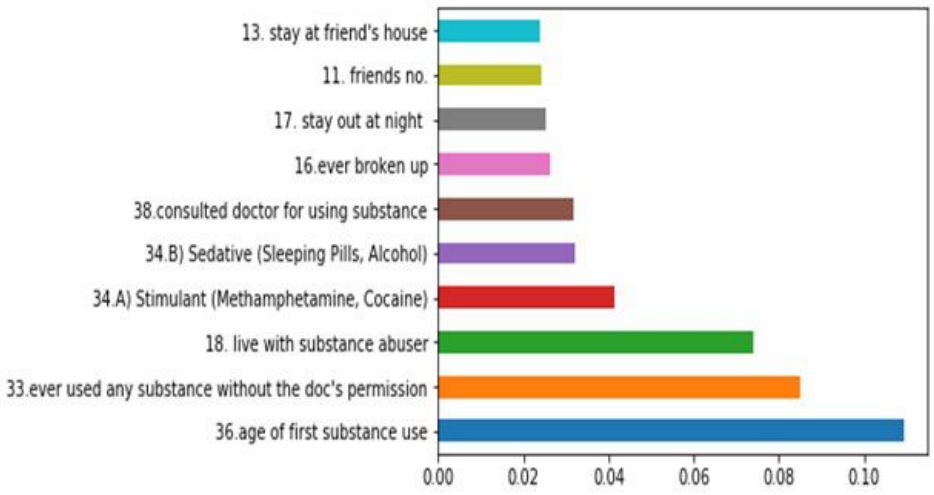


Figure 3.7: Chi Square feature importance graph (different colors used for better visualization)

3.4 T-distributed SNE Implementation for Visualization

In our data set there are k dimensions such that $\mathcal{F}$=$\{f_1$,f_2,$\ldots f_k\}$ where each of the dimension represents independent features which have an considerable impact on target variable. However, visualizing data with such a large dimensional space is difficult so we have used dimension reduction tool in order to visualize our constructed data set. T-distributed stochastic neighbor embedding [24] is a popular tool for high dimensional data visualization. After applying the algorithm, the data set was mapped to a two dimensional data points $\mathcal{D}$=$\{d_1$,$d_2\ldots\ldots$,$d_n\}$ where each of them d_i was clearly visible. This technique facilitated preservation of both local and global structure of data set. Firstly, the distances between high dimensional data points are converted to joint probabilities. For an instance, a data point d_i would select its neighbor d_k according to their probability density where conditional probability $P_{j|i}$ should be higher for closer data points. The equation for computing conditional probability is given below.

$$P_{j|i} = \frac{exp\left(\frac{-||d_i - d_j||^2}{2\sigma^2}\right)}{\sum_{n\neq i} exp\left(\frac{-||d_i - d_k||^2}{2\sigma^2}\right)} \tag{3.6}$$

		Addicted Flag	**Sober Flag**	**Chi Sqr value**	**Asymptotic Significance**	**Phi Cramer's V**	**Ques**
Good	Count	113	216	50.55	1.05E-11	0.327	Ques 7
	Exp C	147.1	181.9				
Not good	C	31.0	7.0				
	Exp C	17.0	21.0				
Satisfactory	Count	68	39				
	Exp C	47.9	59.1				
No	Count	60	136	40.79	1.38E-9	0.293	Ques 16
	Exp C	87.7	108.3				
Yes,I did	Count	86	98				
	Exp C	82.3	101.7				
Yes,someone did	Count	66	28				
	Exp C	42.0	52.0				
None	Count	2	174	227.56	4.65E-49	0.693	Ques 36
	Exp C	78.7	97.3				
15-21	Count	92	34				
	Exp C	56.4	69.6				
<15	Count	77	19				
	Exp C	42.9	53.1				
>21	Count	41	35				
	Exp C	34.0	42.0				
No	Count	5	184	260.18	3.18E-57	0.741	Ques 39
	Exp C	84.5	104.5				
Yes,everyday	Count	189	45				
	Exp C	104.7	129.3				
Yes,sometimes	Count	18	33				
	Exp C	22.8	28.2				
-	-	-		6.11	0.047	0.114	Ques 13
-	-	-	-	4.32	0.038	0.095	Ques 26

Table 3.4: Chi Square cross table

Furthermore, pairwise resemblances were computed for both low dimensional and high dimensional data points. The equation below shows how similarity for low dimensional data points c_i and c_j can be measured.

$$Q_{j|i} = \frac{\exp(-||c_i - c_j||^2)}{\sum_{n \neq i} exp(-||c_i - c_k||^2)} \tag{3.7}$$

For a datapoint d_i, when the distance value increased between d_i and d_j the resulting conditional probability became considerably smaller. As a result, the corresponding data points could not be determined correctly. So symmetrized conditional probability was considered for high dimensional data points. It was calculated using following equation so that each and every data points were considered properly during visual representation.

$$P_{ij} = \frac{P_{i|j} + P_{j|i}}{2n} \tag{3.8}$$

After calculating conditional probabilities in low dimensional and high dimensional data points, the deviation of the similarities is minimized using a gradient descent method in order to visualize data points in low dimensional space. The data points are plotted using tSNE in figure 3.8 and comparison between PCA vs tSNE is shown in figure 3.9.

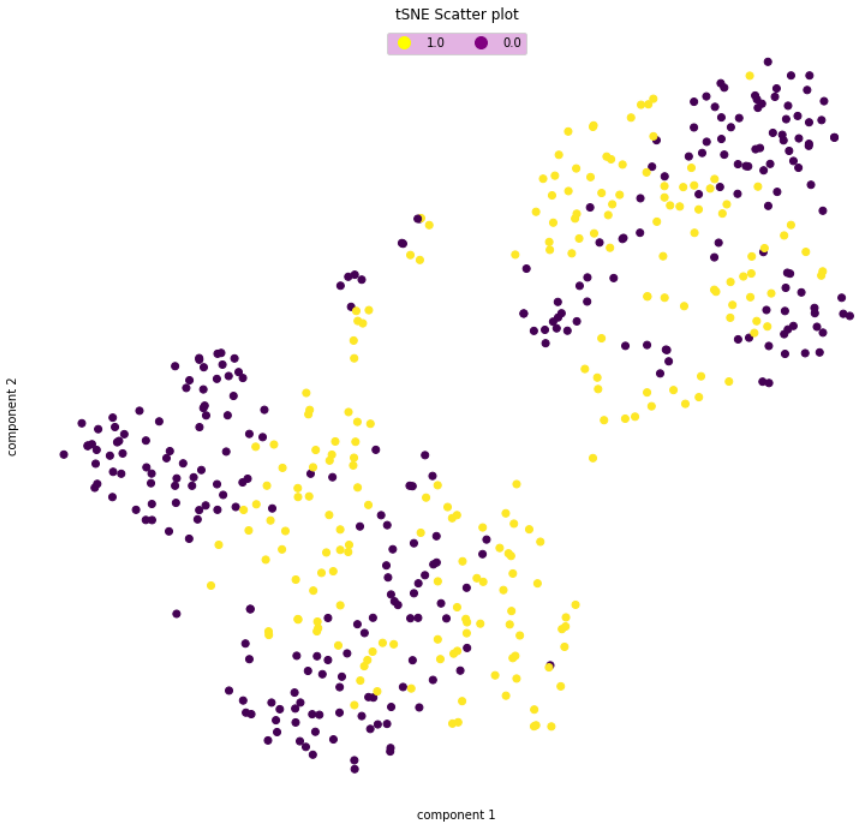


Figure 3.8: Scatter plot after T-distributed stochastic neighbor embedding

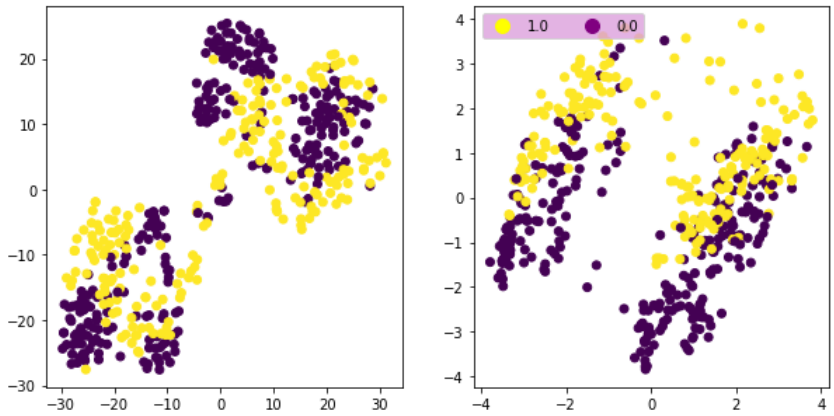


Figure 3.9: Comparison between tSNE and PCA

3.5 Feature Analysis

In our research, the purpose is to identify the person who is vulnerable to drug addiction to the near future. Also, to build a model which will enable the existing systems to identify drug addicted people. To do so we make a questionnaire of over 60 attributes where the participants have to answer in a binary form. Some important features have been identified later by running some algorithms like SVM, Random Forrest, Linear Regression and Artificial neural network (ANN) which put really high value in the field of identifying the prospective drug addicted person. We give both addicted and non-addicted people the same questionnaire and later matched their answer. By doing this we found that some questions (attributes) have high factor for finding the vulnerable person towards drug addiction. Some important questions are:

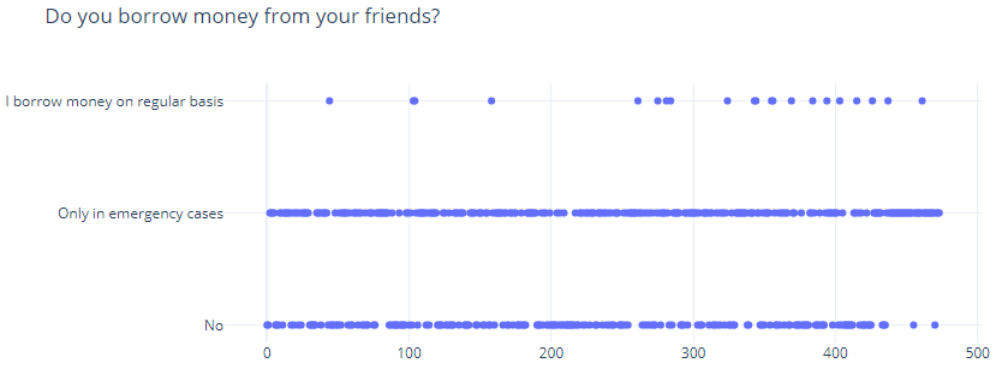


Figure 3.10: Scatter diagram of the answer given by the non-addicted people

In this question, we tend to focus on the psychological state of the participant. Here the larger cluster is "only in emergency case". We have visited several universities and schools, additionally rehabilitation center across the town to require answers from drug dependant folks. After we use algorithms to check whether or not there's any co relation between the drug addicts answer and non-addict's answer, we tend to saw a positive co-relation in figure3.11. Therefore it's clear that those non-addicted who mark purpose ("only in emergency cases"), is on risk of potential Drug addicted. In another vital feature wherever we tend to target the participant's social rank.

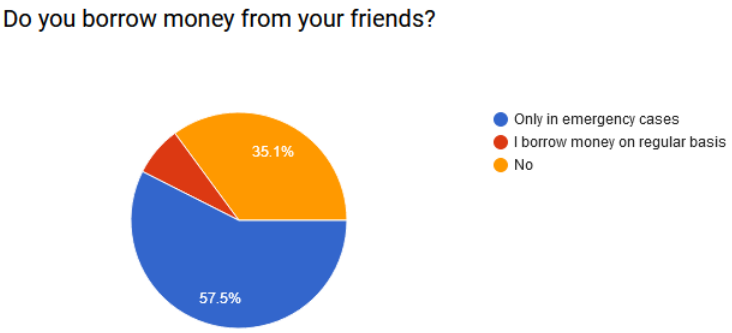


Figure 3.11: pie chart of the answer given by the addicted people

During this pie diagram in figure3.13 .Most of the participants are unmarried. With the assistance of algorithmic program like Linear regression, we tend to found that there is a positive relation in figure3.12 towards the vulnerability of substance abuse.

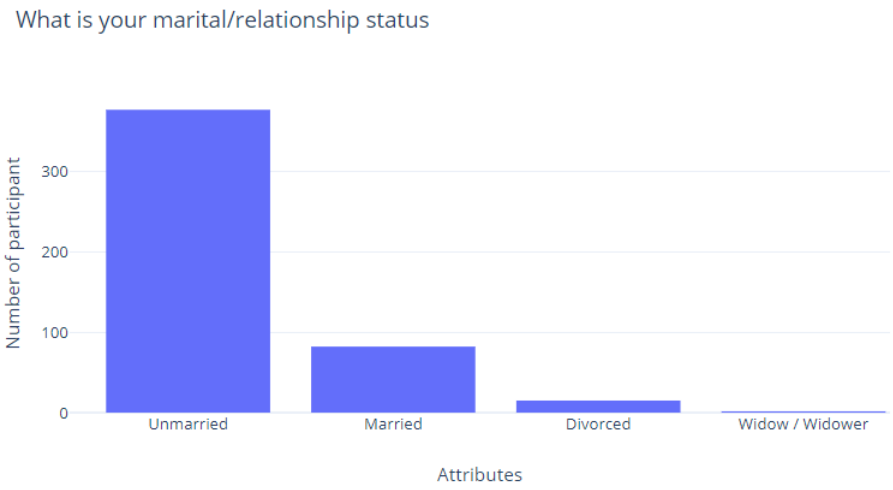


Figure 3.12: Histogram for marital/relationship status(non-addicted)

Another vital feature wherever we tend to asked a few criteria of personality. Here it says that the majority of the folks that are flagged as addicted (figure: 3.15), they broke up with somebody. Our algorithms tag this question as vital criteria for distinctive potential drug abuse as a result of once gathered information from alcoholic individuals (figure: 3.15), most of their responses aforesaid that they broke up with somebody that leads them to the present toxic path.

Our major feature selector like mRMR and Chi square finds out that this can be one amongst the necessary feature to flag the potential addicted person. This chart shows that variety of individual's (figure: 3.16) response that they live with somebody who is drug addict or alcohol user. By examination with the knowledge collected from the drug addicts (figure: 3.17), it's clear that there's positive co-relation; that means those folks who live with alcohol/ substance abuser are prone to suffer from substance abuse disorder.

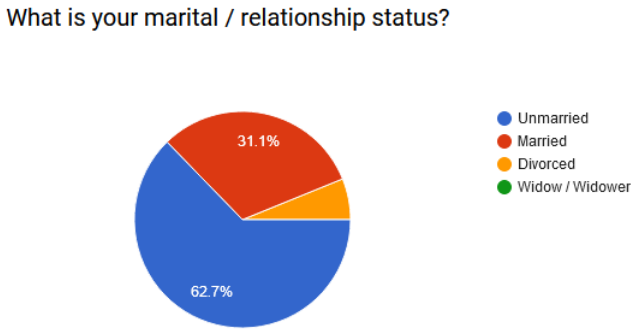


Figure 3.13: Pie chart for marital/relationship status(Addicted)

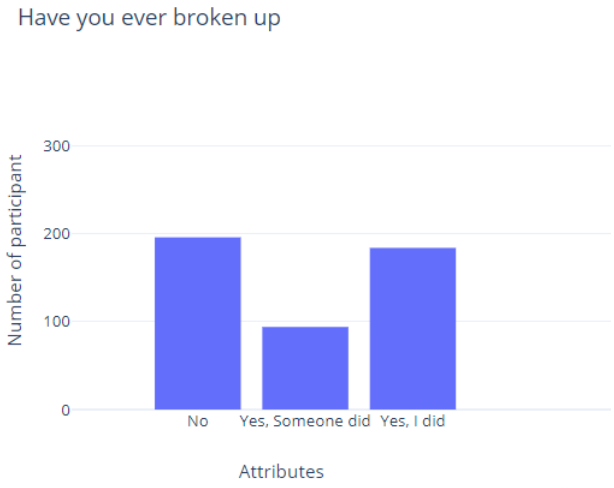


Figure 3.14: histogram of the answers given by the non-addicted people

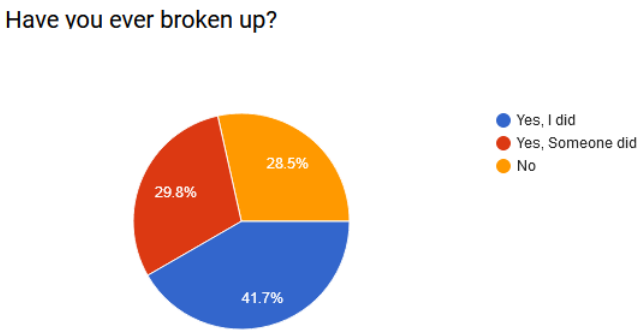


Figure 3.15: pie chart of responses collected from addicts

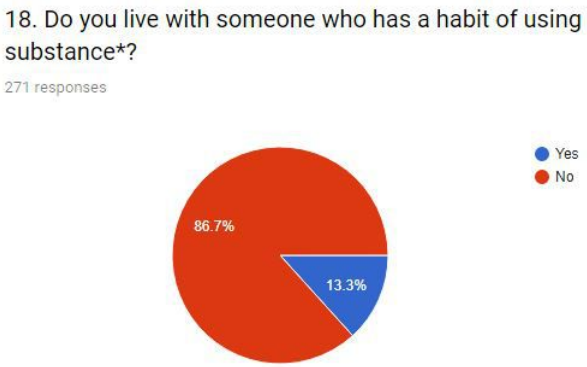


Figure 3.16: histogram of the answers given by the non-addicted people

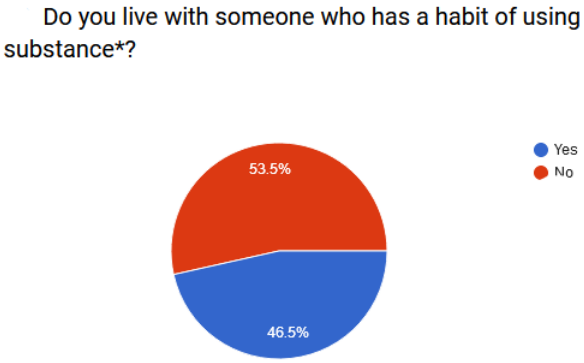


Figure 3.17: pie chart of responses collected from addicts

Most of the participants responded about facing problem in their workplace which can be related to the drug abuse (figure.3.18, 3.19). Which is clearly visible from figures, that they face problem (51.8%) in their workplace or educational institution. Our aim was to find out their personal lifestyle and consequences they face because of drug abuse as well as other problems, it hamper their life. When we run chi square to test whether it has co-relation, this gives us a strong co-relation with the drug addicted people.

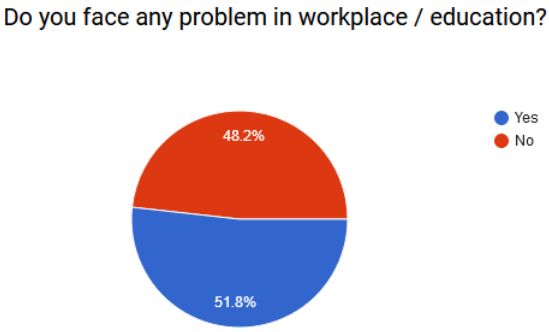


Figure 3.18: pie chart for problems in workplace/ education

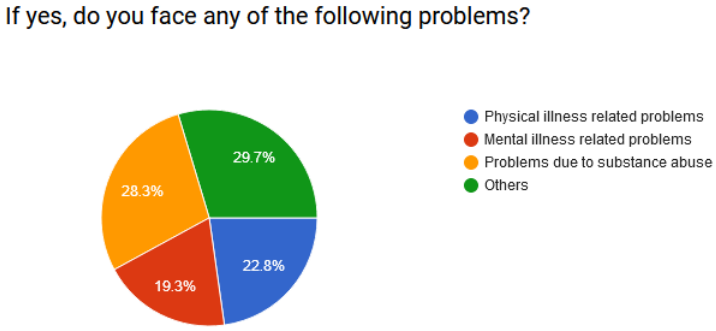


Figure 3.19: pie chart for various difficulties faced in workplace/ education

Furthermore, we wanted to focus on participant's physical condition. This is one of the important feature because sedative type medicine is the first step toward potential drug abuse. When we compare with drug addicted people (figure 3.21), they mostly use sedative medicine that leads them to become a drug addict.
Another most important feature which is related to the physical and mental health is whether the participant's smoke or not. The pie chart (figure: 3.22) represent the non-addicted whereas another pie plot (figure: 3.23) represents the addicted people. Most of the non-addicted response is positive. By matching the response with the addicted people it is clearly visible that friends influence in a person's life can play a significant role. So a friend can influence someone to use drug addiction too. When we take responses from drug addicted people, they agreed with one point, by smoking frequently they started to use drug too. As both addicted and non-addicted data indicates a positive relation, this is an important factor to predict whether an

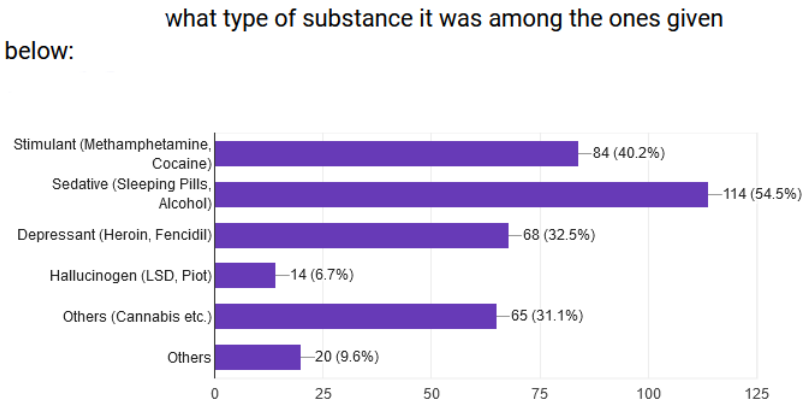


Figure 3.20: Bar diagram for various types of substance (Addict)

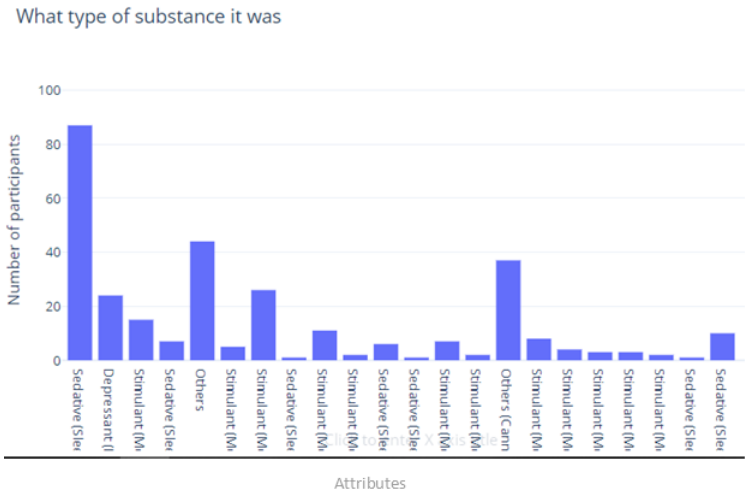


Figure 3.21: Bar diagram for various types of substance (addict)

individual is vulnerable to addiction or not.

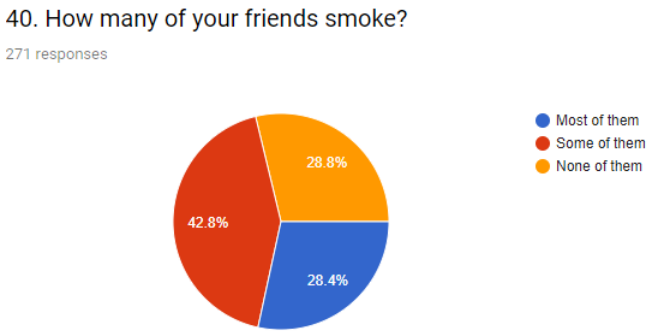


Figure 3.22: Influence of smoking (non-addict)

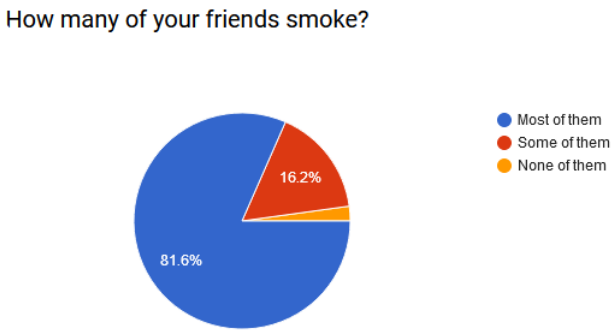


Figure 3.23: Influence of smoking (addict)

In this feature, it would be determined that overall the bulk of non-addicted person do smoke (figure 3.24). On the opposite hand, Drug addicted person (figure 3.25) showed nice interest in regular smoking. However, the foremost of smokers are people who have productive ages as against unproductive ages. So, there's a high risk that these non dependant of us are danger at risk of addiction in future.
Half of the addicted people reportedly mentioned that they started smoking due to peer pressure followed by avoiding tension (figure: 3.27), There is a stereotype that smoking with friends is a symbol of maturity . The smoking status among the young population is the great threat for a healthy country. So this should be addressed properly to create awareness among them and as a proper measures should be taken to prevent smoking among the valuable youth population like non addict folks. If we look at the pie plot (figure: 3.27) which is showing us a visualization of drug addicted people, most of them are showing the same result, they are enjoying staying with friends. So it is a clear that peer pressure can be one of the most important factor to become a drug addict.

It was also found that about 54.2% respondents have faced problem to maintain daily routine but 45.8% respondents have faced no difficulties to do so (figure: 3.28a). Most of the addicted respondents (59.2%) gave their opinion that they were not able to complete their tasks for addiction (figure: 3.28b) but few of them said that they

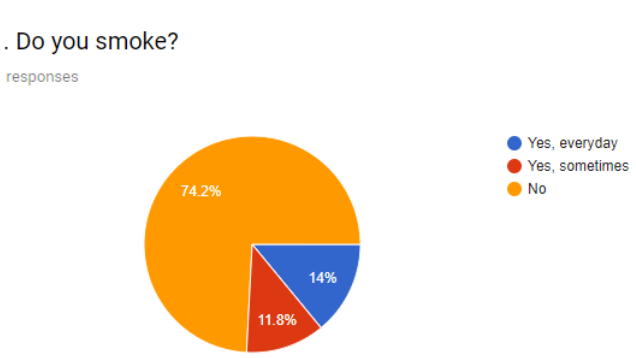


Figure 3.24: pie chart for smoking (non-addict)

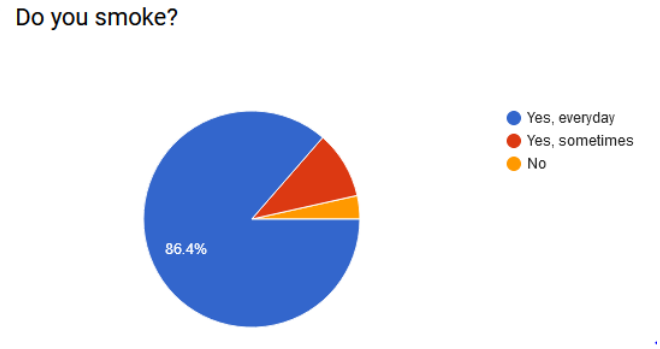


Figure 3.25: pie chart for smoking (addict)

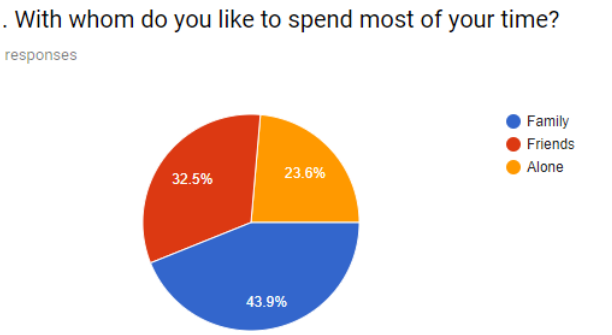


Figure 3.26: pie chart for personal interest (non-addict)

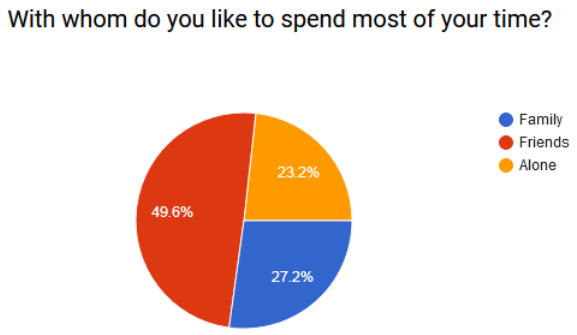


Figure 3.27: pie chart for personal interest (addict)

are not affected by addiction. So there is positive co relation between the addicted and non-addicted data which lead this attribute as an important factor to predict the vulnerable drug addict.

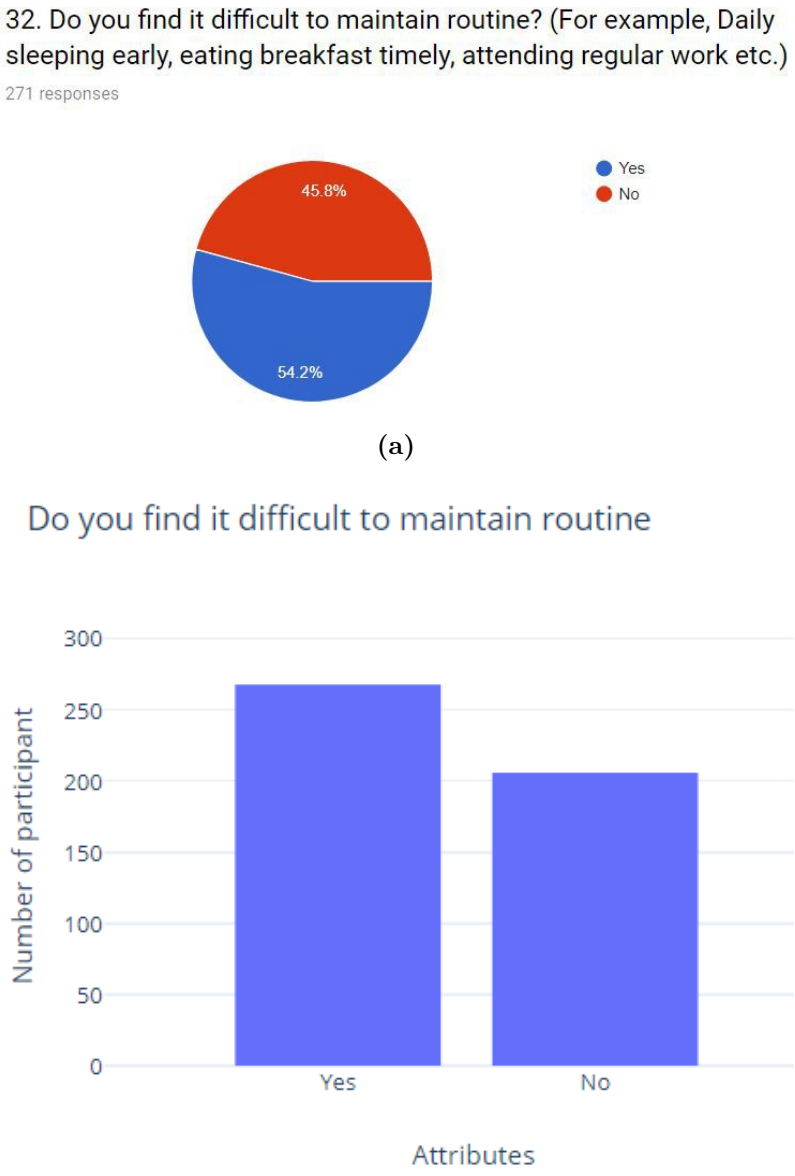
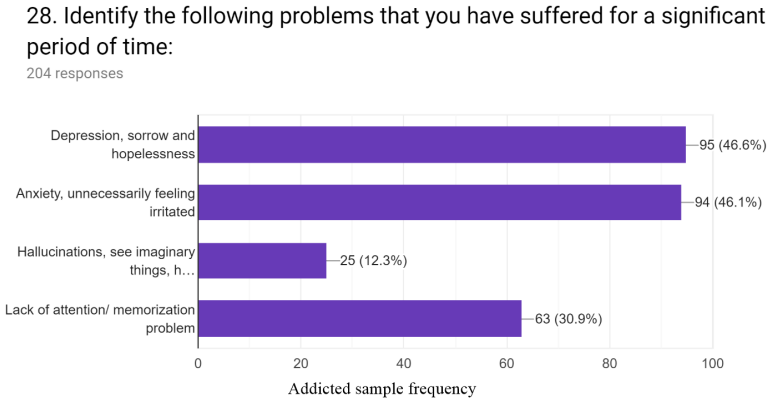


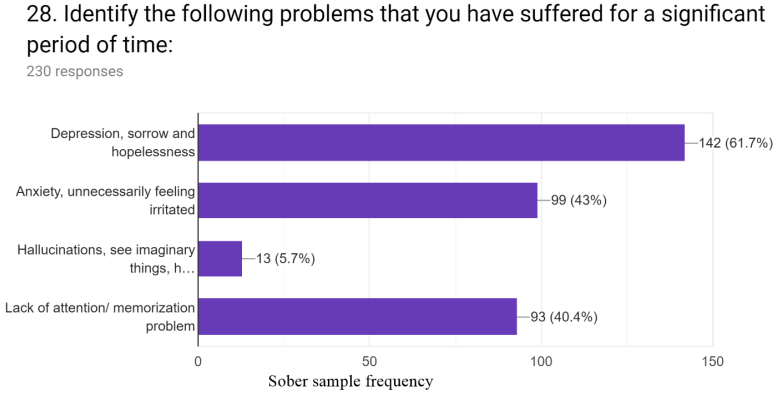
Figure 3.28: Graphs of difficulties to maintain daily routine life
(a)pie chart for sober peoples' samples,(b)Histogram of addicted peoples' samples

From the figures (figure 3.29a,3.29b) it can be inferred that 'Anxiety', 'Hallucination' can be considered as significant impact of addiction problem. According to medical study, it has been proved that drugs like hallucinogen and marijuana directly affect abuser's central nervous system which may lead to hallucination and schizophrenia. In addition to that, participants of focused group discussion also shared their concerns regarding severe anxiety. According to the opinion of narcotic drug abusers from FGD, they tend to feel anxious as their regular required dose of drug increases rapidly. Furthermore, unnecessarily feeling irritated is a symptom of methamphetamine withdrawal. In our empirical analysis, 12.3% addicts reported

about hallucination and 46.1% patients suffered from anxiety (figure 3.29a).



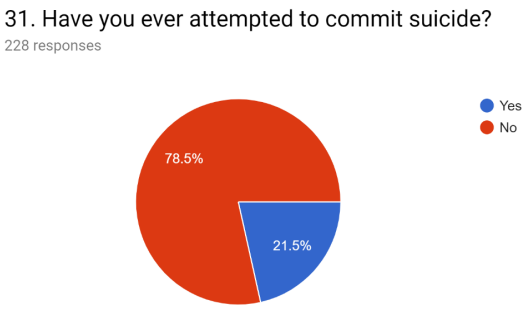
(a)



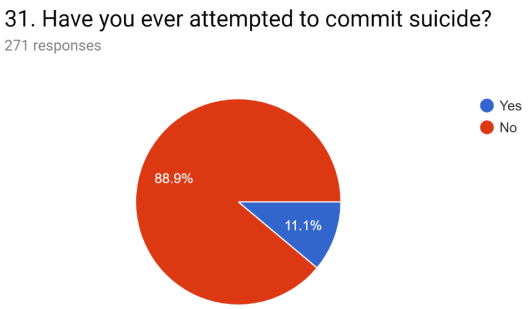
(b)

Figure 3.29: Bar plot of Mental health issues
(a)Plot of addicts,(b)Plot of sober people

In our survey, we intended to find out whether there is a relation between drug addiction and suicide. From our constructed dataset, we have found out that persons who have made suicide attempts are more likely to be victim of substance abuse disorder. Drug addiction significantly contributes to suicidal behavior due to abusers' financial problems and impulsive attitude. From figure 3.30a, 3.30b we have seen that 21.5% addicted person have attempted to commit suicide; whereas only 11.1% of sober people have exhibited suicidal tendency. As a result, this feature can be considered as an indicator to identify substance abusers.
Early and middle childhood is a crucial time when our brain is shaped and formed through learning process which can be affected by some external factors. A considerable percentage of our sample have claimed that they were exposed early to drugs due to their family members substance abuse disorder. From figure 3.31a, we can observe that 23.2% of our addicted samples admitted that they had an early exposure to addictive substance because of family members addiction problem.
Our main purpose was to find out indicators that will enable early detection of the problem. Age of first unprescribed substance use can be a strong indicator for

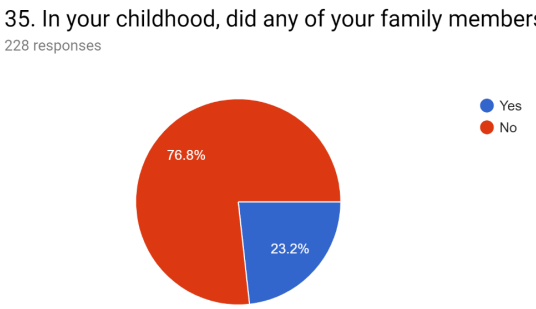


(a)

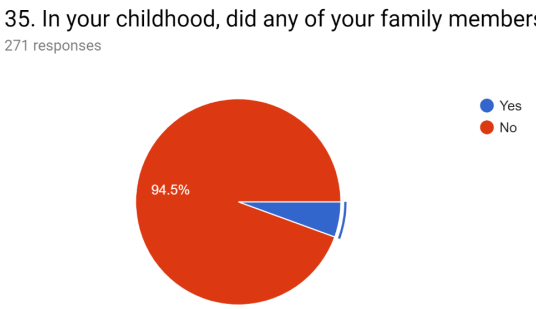


(b)

Figure 3.30: Respondents attempt to commit suicide
(a)Plot of addict's suicidal attempts,(b)Plot of sobers people's suicidal attempts



(a)



(b)

Figure 3.31: Family members' substance abuse disorder
(a)Plot of addict's responses,(b)Plot of sober people's responses

identifying this problem. From our research, we found out that persons who have taken their first dose at an early age are more prone to addiction problem. In our data set, 43.5% of addicts has reported that they have taken their first dose of substance at a range of 15-21 years (figure 3.32).

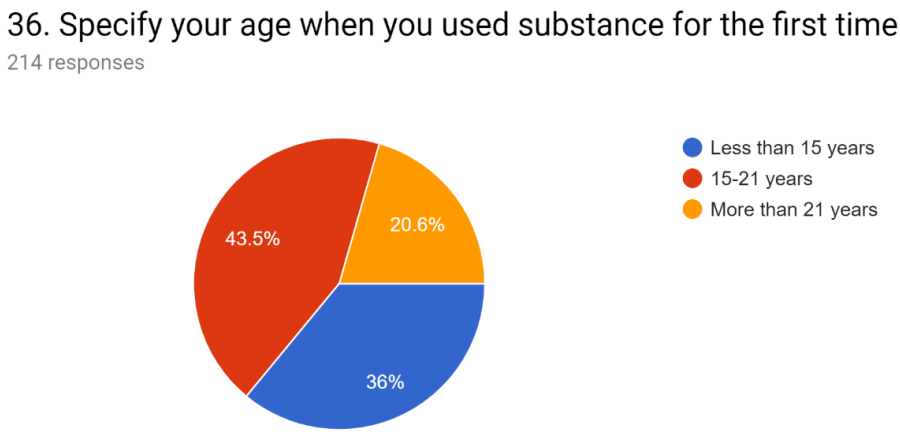
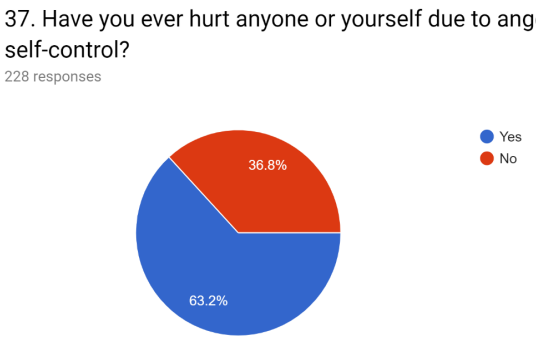


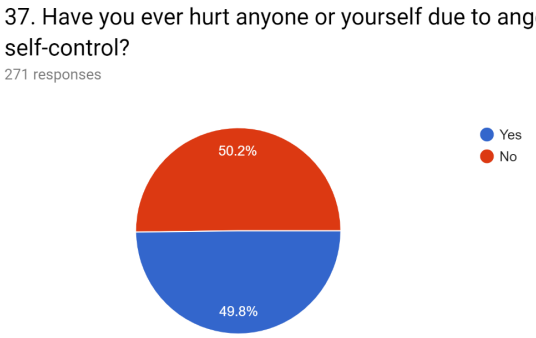
Figure 3.32: Age of first unprescribed substance use

As we have discussed earlier, addicted people exhibit impulsive attitude and tend to suffer from withdrawal symptoms such as severe anger, unnecessarily feeling irritated. So, they have a tendency to hurt themselves or other people specially during the withdrawal period. Some of our respondents have claimed that they used to harm themselves through self-mutilation due to depression, guilt feelings and anger. In figure 3.33a, it is clearly visible that 63.2% of addicts have admitted that they have caused harm to others or themselves. On the other hand, 49.8% of sober people have hurt others or themselves due to anger or loosing self-control (figure 3.33b).

Whenever an addicted person leaves drug he tends to suffer from withdrawal symptoms. There are two phases of withdrawal symptom - acute withdrawal and post-acute withdrawal. Substance abuser may suffer up to few weeks from acute withdrawal which includes physical withdrawal signs. The next phase includes more mental withdrawal signs rather than physical withdrawal symptoms. Post-acute withdrawal may last up to 72-96 hours. When SUD patient stops taking drug, his brain requires the flow of chemical in order to restore equilibrium. His brain cannot normally restore to stability without presence of dependent chemicals. Therefore, an addict's brain chemistry starts to become normal which will be restored to new equilibrium condition. Only persons who are dependent on substances may suffer from it as they have abused drugs for a prolonged period of time. This indicator contributes significantly to detection of a substance abuser. So we have included a question regarding withdrawal symptoms to identify potential drug addicts. In figure 3.34a it can be observed that 38.3% addicts face withdrawal symptoms regularly and 33.8% addicts face withdrawal signs less frequently. Therefore, total 72.1% addict faces withdrawal symptoms when they try to stop taking drugs or try to give up smoking. However, only 23% of sober people faces withdrawal symptoms (figure 3.34b).

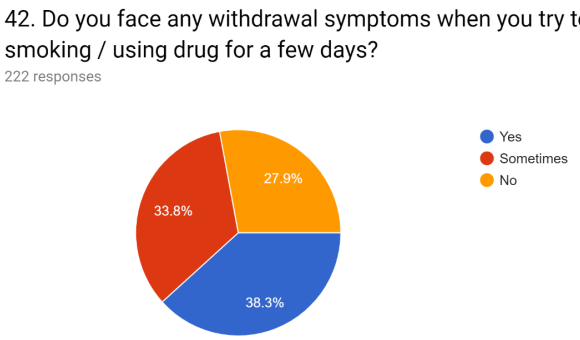


(a)

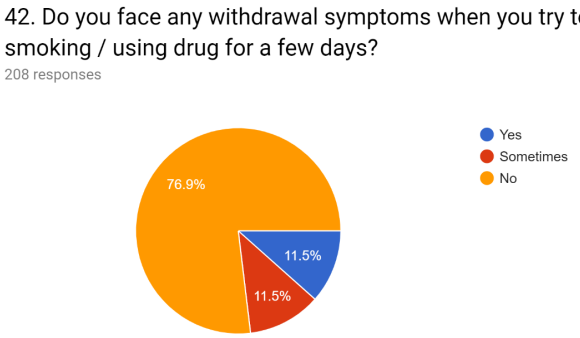


(b)

Figure 3.33: Tendency to cause self-harm or hurt anyone due to anger (a)Plot of addict's responses,(b)Plot of sober people's responses



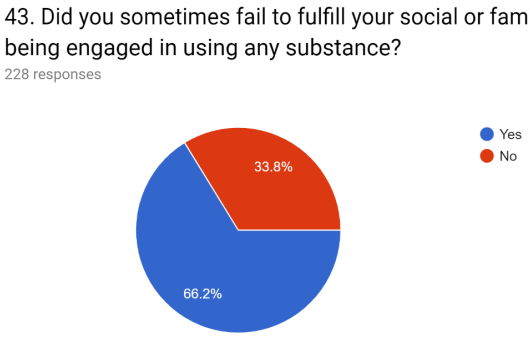
(a)



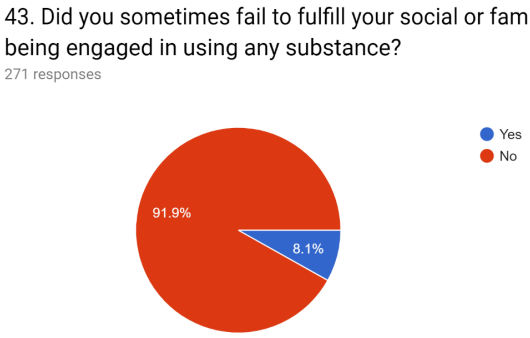
(b)

Figure 3.34: Pie chart of faced withdrawal symptoms
(a)Pie chart of addict's responses,(b)Pie chart of sober people's responses

From focused group discussion, we came to know about the consequences of addiction. In primary stage of addiction, addict can somehow manage to fulfil his familial duties. However, he will gradually become more desperate, impatient and irresponsible. It becomes more difficult for such an individual to keep his promises, commitments and meet expectations of his family members. The short term consequences may lead to trust issue, avoiding family members and violent behavior at home. Failure to fulfil familial duties can play an important role as distinguishing feature between addicts and sober people. From pie chart visualization of our addicted sample (figure 3.35a), we can see that 66.2% of drug abusers have failed to fulfil familial or social duties whereas only 8.1% of sober population (figure 3.35b) failed to fulfil familial duties.



(a)

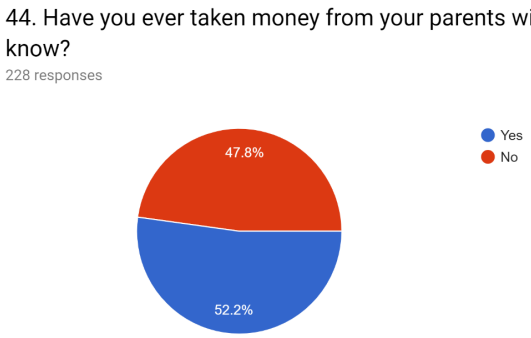


(b)

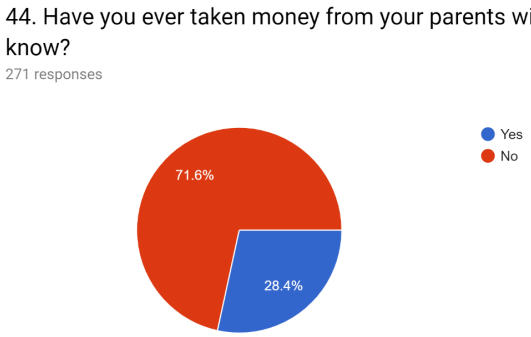
Figure 3.35: Pie chart of failure to fulfil familial or social duties
(a)Pie chart of addict's responses,(b)Pie chart of sober people's responses

In this question (figure 18a, 18b), we intended to focus on moral degradation which is caused by drug addiction. Drug abuser will eventually face financial difficulties due to his rapidly increasing expenses of buying substances. Therefore, he or she will be engaged in illegal activities to cover the increased cost of buying drugs. However, their exposure to illegal activities begins at home. As they require more drugs than their affordability and they intend to maintain a regular supply, they get engaged in immoral activities like stealing. In our study, a significant association (52.2%) have been found between illegally taking money from parents and drug dependency (figure 3.36a). On the other hand, pie chart (figure 3.36b) demonstrated majority

of (71.6%) sober people have never taken money illegally from their parents.



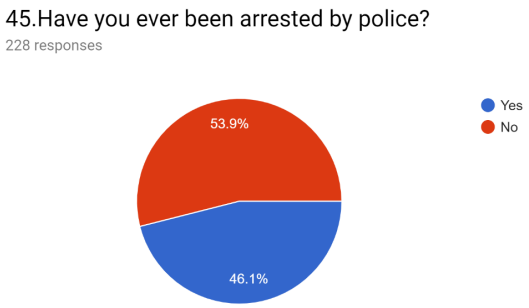
(a)



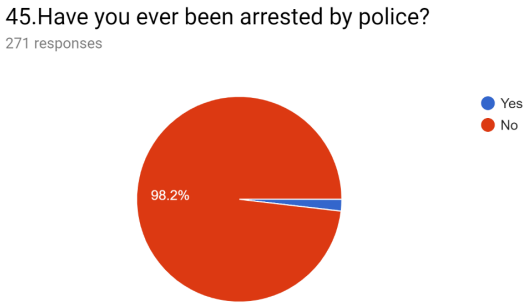
(b)

Figure 3.36: Pie chart of illegally taking money from parents
(a)Pie chart of addict's responses,(b)Pie chart of sober people's responses

Clear evidences from our empirical analysis (figure 3.37a) indicates that drug consumers have a high probability of getting involved in illegal activities. A large number of subjects (46.1%) reported that they were arrested during their addiction period. Their criminal record indicated that a considerable amount of crimes was committed to obtain money for drugs. Although initially they have taken their first dose for recreational purpose, later on it has become difficult for them to live properly without drugs. On the contrary, only 1.8% of sober population were arrested by police (figure 3.37b). We discussed about the crimes where the offender was under influence of addictive substances; but carrying and using drug is also considered as a heinous crime. Bar chart in figure 3.38 demonstrated that 34.4% of drug consumers were arrested due to illegal possession of drug.



(a)



(b)

Figure 3.37: Pie chart of illegal criminal records
(a)Pi-chart of addicted people arrested by police,(b)Pi-chart of sober people arrested by police,

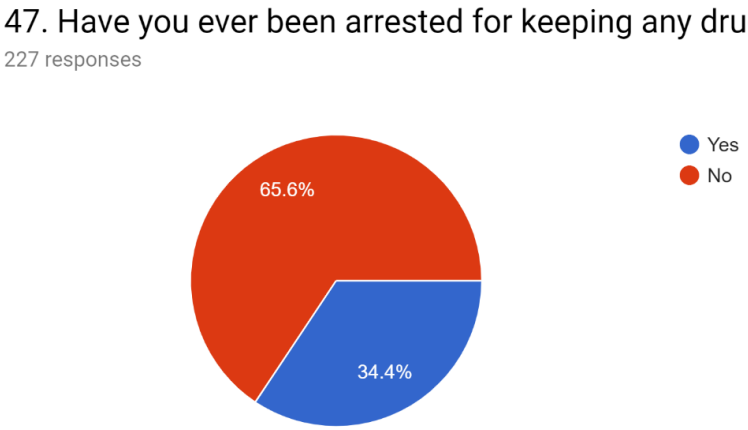


Figure 3.38: Consumers arrested for illegal possession of drug

3.5.1 Heatmap of Data

Secondly, we have generated heat map of our data. Heat map is a better visual representation for a large data set with more than 60 features and approximately 500

samples. Because color shade consumes less space to represent data than numbers. The diagonal represents the variable/feature being correlated with itself. Of course, it should be 1. The right bar indicates green shade for 0 to strongly positively correlated values. In figure 3.39, the darker Red shaded part indicates that they are negatively correlated. we can see that smoke, age of first substance dose age, unprescribed drug use tendency are strongly correlated with Addicted flag. Again , borrow money is negatively correlated with Flag outcome. Secondly, in figure 3.40 the darker Red part indicates negative correlation. we can see that smoke, arrested for keeping drugs, arrested for selling drugs, age of substance abuse, abuse medicine without doctor's permission, peer influence, withdrawal symptom severity, failure to fulfill social/family duties, ever been arrested are strongly correlated with Addicted flag. Again , addiction in family origin and marital status are negatively correlated. It allows us to visualize the correlation between quantitative variable of importance.

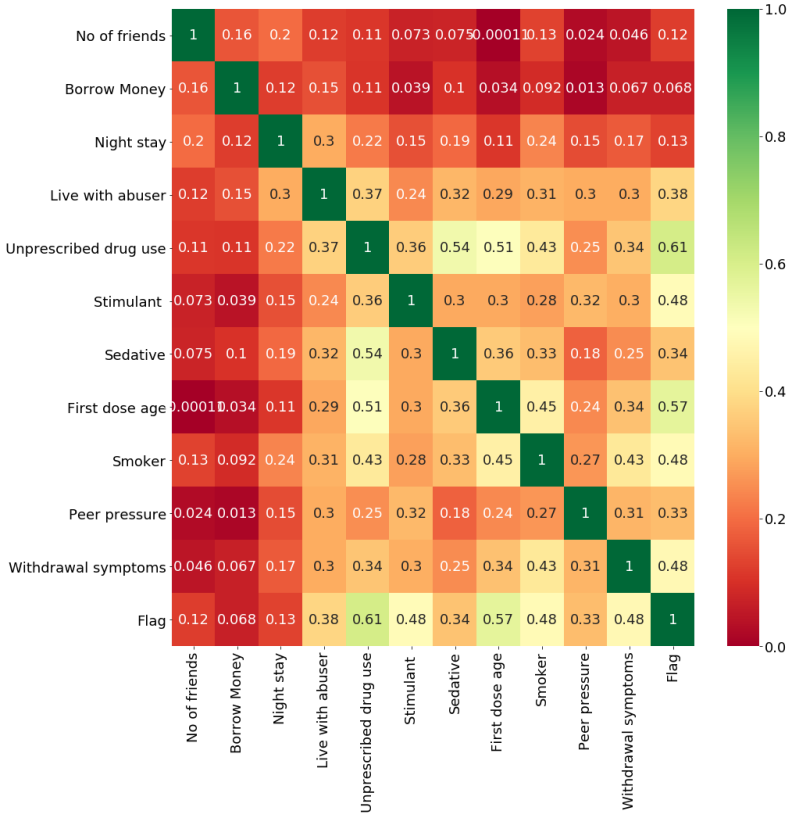


Figure 3.39: Partial Heat map based on feature correlations

3.5.2 Ensemble approach for feature voting

As we have a vast dataset in terms of feature number which increases the execution time complexity and which become more complex to analyze to identify important factors that should be identified as relapse triggers or vulnerability measurement parameters. After applying Minimum Redundancy and Maximum Relevance we narrowed our feature number to 40. Again, we have used Chi square feature importance to identify a list of most important features with highest pearson correlation estimates. Then we used them to determine the dependencies and Asymptotic Sig-

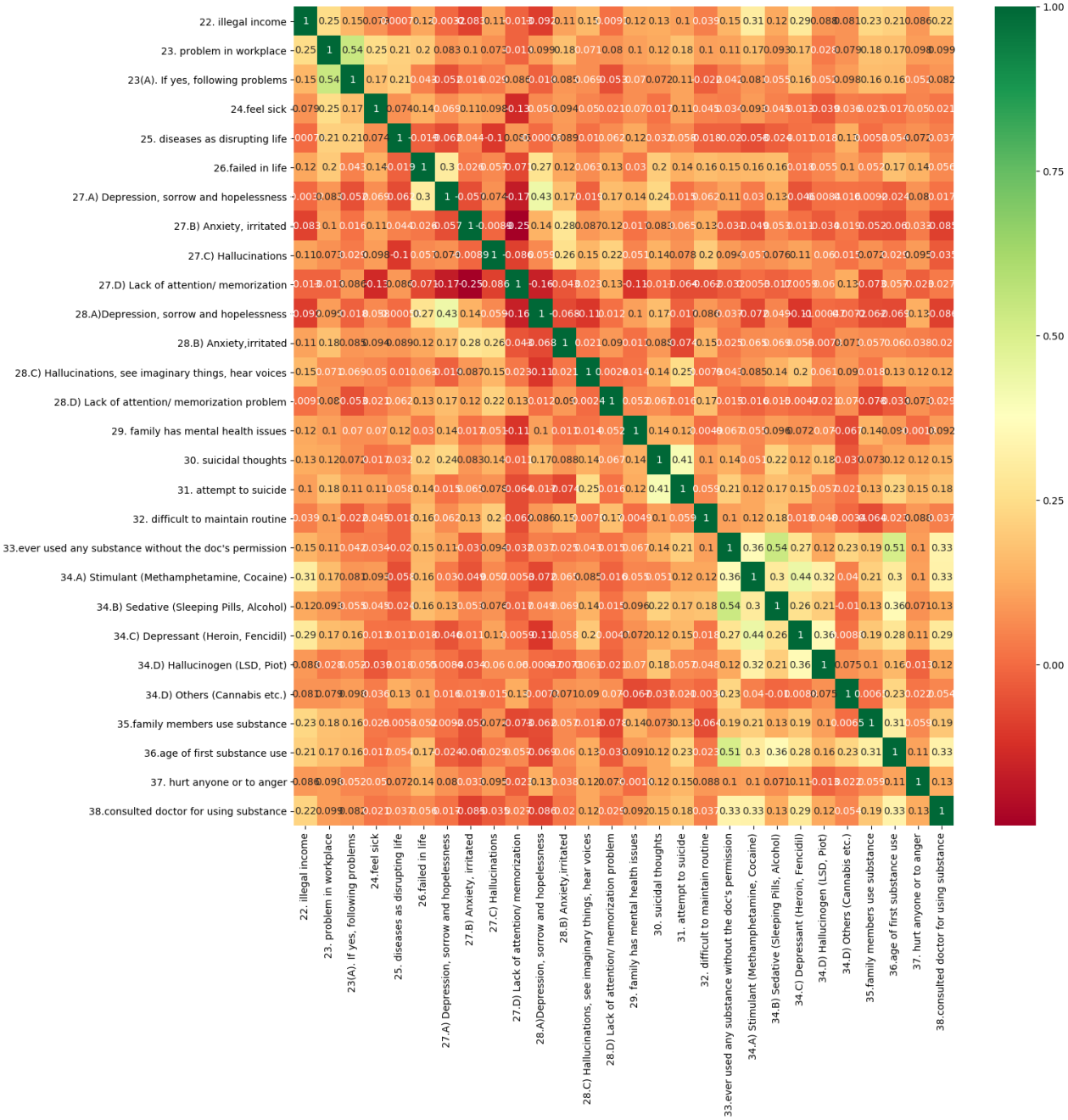


Figure 3.40: Heat map based on feature correlations

nificance with the main 'Addiction Flag' outcome. But to make a list of most important features we also have used an ensemble feature selection approach. In the first phase, we went through some sieving. For filtering, we have used sklearn's k-best algorithm using pearson's correlation to reduce the feature number to half(which is 30 in our case).It takes two arrays of indicator and outcome and returns array of Score and Pvalue. After using min-max Scaler to scale the scattered dataset we obtained a list of 30 variables. Then, we have used chi2 selector for identifying best 30 variables. It is used to identify non negative statistical and empirical values between features and outcome.

It also returned Chi2 array along with shape of 30 value and also the pvalues. These two Algorithms aided to maximize the correlation among feature and final outcome. They also helped to identify features with highest dependency on addiction vulnerability. Their contribution impact factor that highly influences degree of vulnerability could be reduced down to half of the original features. This phase is called as filtration because it helps us to reduce down feature numbers before applying any classification algorithm. But this step is not enough for us because this stage doesn't give any feedback to the learning algorithm that why we should consider which subset and which is better for Successful prediction where sensitivity and specificity both are upgraded along with precision and accuracy. On the other hand, another popular methodology-wrapper could help us through giving and taking feedback directly from learners where the subgroup of indicators is mainly picked up on the basis of classification report through learning.

In the second phase, we have used Wrapping with help of Recursive Feature Elimination. RFE helps to rank features on the basis of recursive feature elimination. These deduction policy helps to pick different subsets of indicators to check whether they can outperform the normal learning success rate. It recursively tries to consider smaller set of features. Importance of features are obtained through co-efficient and feature importance attributes. By ranking it tries to eliminate the least important features through pruning. Initially, we have used logistic regression as estimator parameter. In this stage, Recursive Feature Elimination with logistic regression estimator was iterated up to 3 counts with different subset of feature until the classifier's report touched a better threshold range.Sequentially, it was 59,49,39 in 3 steps. We also assigned number of reduced features to be 30. We also modified the parameter step to be 10 which indicates number of features to be reduced at each step. Again, we have also used Support Vector Regression for wrapping purpose. We have taken the vote of both RFE and SVR. SVR's function is quite similar to SVM in case of classification technique. In this case, a threshold limit is maintained to the approximation of SVM. We have used linear SVR for wrapping purpose. For an example, after observing 3.41 'ever broken up', 'arrested for keeping drugs' was voted by RFE but not supported by SVR. But on the other hand, 'Failed to fullifil social Responsibility' was both supported by RFE and SVR. In the third stage, we have used 3 supervised machine learning classification algorithms. It can be considered as an embedded method [23][24] where learning and feature selection are interconnected and integrated in such a manner that it is more fruitful to identify most important features in terms of some algorithms. Here, we have incorporated 3 algorithms-Logistic Regression, Gradient boosting and Random Forest classifier. We

featrues	Chi-2	LightGBM	Logistics	Pearson	RFE	Random Forest	SVR	Total
45.ever been arrested by	True	True	True	True	True	True	True	7
43.fail to fulfill your	True	True	True	True	True	True	True	7
34.D) Others (Cannabis e	True	True	True	True	True	True	True	7
34.A) Stimulant (Methamp	True	True	True	True	True	True	True	7
33. ever used substance	True	True	True	True	True	True	True	7
47.arrested for keeping	True	True	True	True	True	True	False	6
42.withdrawal symptoms	True	True	True	True	True	True	False	6
39.smoke?	True	True	True	True	True	True	False	6
38.consulted any doctor	True	True	True	True	True	True	False	6
36.age of first dose	True	True	True	True	True	True	False	6
1. gender	True	True	True	True	True	True	False	6
7. relationship with fam	True	False	True	True	True	True	False	5
20.occupation	False	True	True	True	True	True	False	5
2. Medium of study	True	True	True	True	True	False	False	5
15.relationship status	False	True	True	True	True	True	False	5
46.case/lawsuit going on	True	True	False	True	True	True	False	5
40.No. of yoursmoker fri	False	True	True	True	True	True	False	5
34.B) Sedative (Sleeping	True	False	True	True	True	False	False	4
27.A) Depression, sorrow	False	False	True	True	True	True	False	4
23.problem in workplace	False	True	True	True	True	False	False	4
21.range of monthly avg	True	False	True	True	True	False	False	4
17.stay out at night?	False	True	True	True	False	True	False	4
16.ever broken up	True	True	True	True	False	False	False	4
14.borrow money	False	True	True	True	True	False	False	4
13.stay at friend's hous	False	True	False	True	False	False	False	4
8.range of Family Income	False	True	False	True	False	True	False	3
37.hurt anyone due to an	True	False	False	True	True	False	False	3
34.D) Hallucinogen (LSD,	True	False	False	True	True	False	False	3
3.educational qualificat	True	False	False	True	True	False	False	3
29.mental health related	False	False	True	True	True	True	False	3
28.D) Lack of attention/	False	False	True	True	True	False	False	3
28.C) Hallucinations, se	False	False	False	True	True	False	False	3
26.failed in your life	True	False	True	True	True	False	False	3
12.friends class :	False	False	False	True	False	False	False	3
9.Amount of daily hand m	True	True	False	True	False	True	False	3
6. No. of family members	False	True	False	True	False	True	False	3
5. religion:	False	True	False	True	False	False	False	2
48.arrested for selling	False	True	False	True	False	False	False	2
44.ever taken money from	False	False	False	True	False	False	False	2
	True	False	False	True	False	False	False	2
	True	False	False	True	False	False	False	2

Figure 3.41: partial output image for feature selection using Ensemble voting based feature selection

have considered the vote for each algorithm using sklearn's select from Model where these 3 algorithms are used as estimators. The estimator algorithms must have coefficient or feature_selection attribute to accommodate this ensemble approach. Firstly, in case of we have used L1 penalty.L1 is usually least absolute deviation which represent error percentage/loss. L1 penalty is used for comparison of sparsity on the basis of C value. This indicates C=1.Sparsity means proportion of zero coefficient. Sparsity with L1 penalty is usually 6.25%.Logistic Regression returned 26 indicators shown in 3.41.

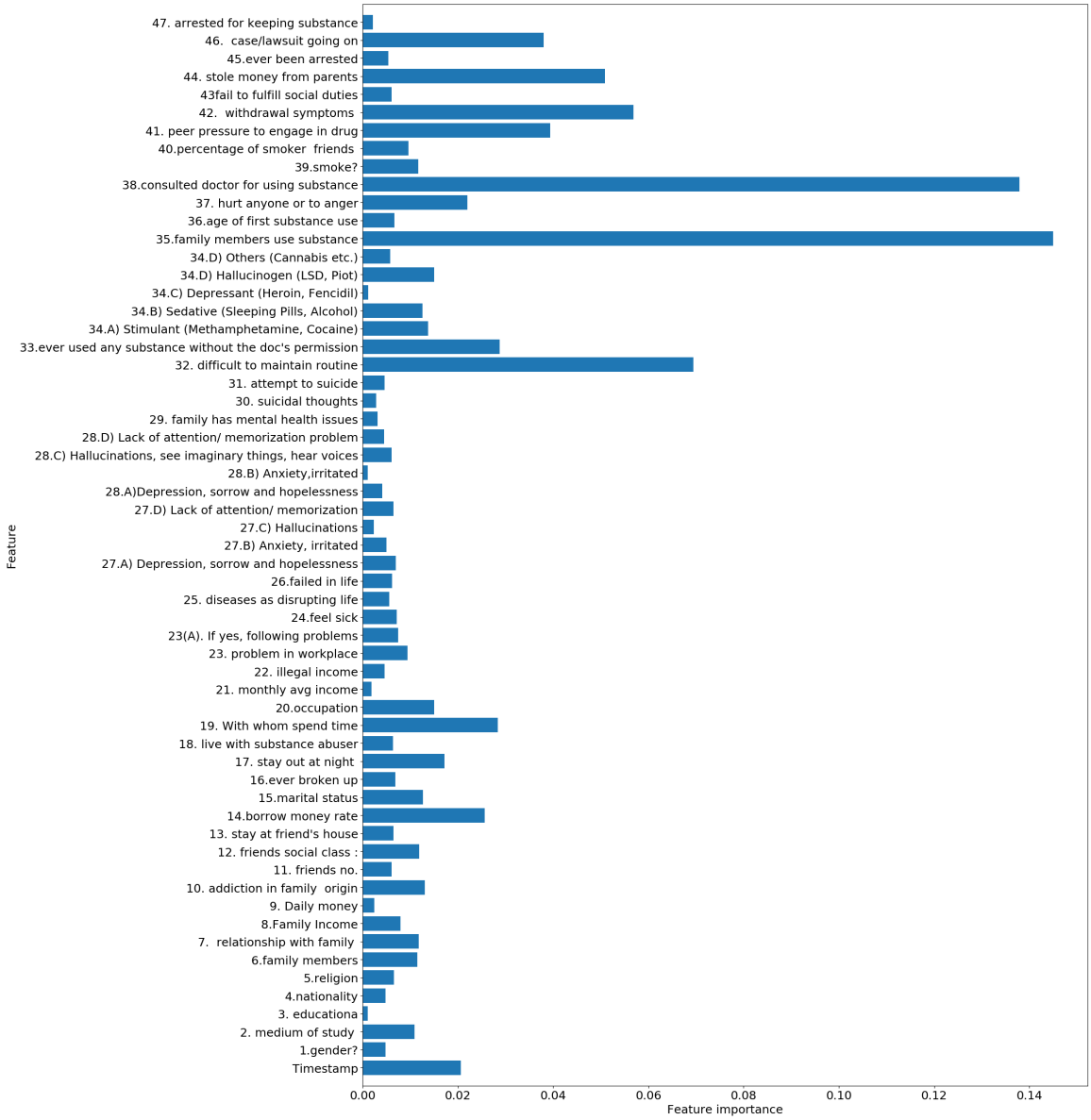


Figure 3.42: Feature importance using Random Forest Classifier

3.42 represents feature importance of Random Forest Classifier, Gradient Booster algorithm. When we have used Random Forest Classifier as estimator for Feature selection, we have used n_estimator as 30 .It returned a shape of 25 features. Furthermore, we have used Light Gradient Boosting Model classifier. A large number of tree-based structure helps to identify purity of selected nodes. As for each tree, each attribute is used as root it can gain more information about each indicator's

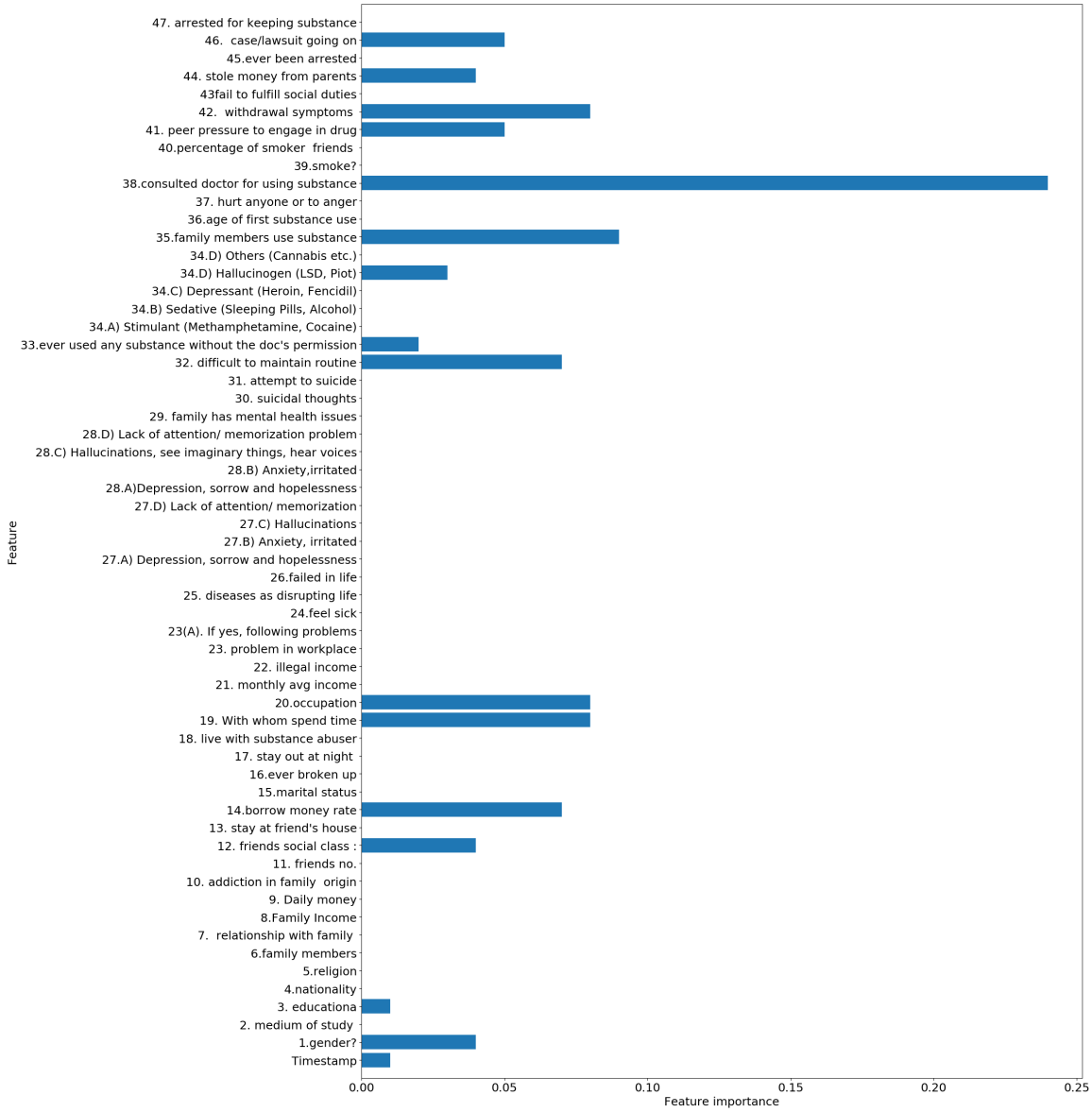


Figure 3.43: Feature importance using Light Gradient Booster Classifier

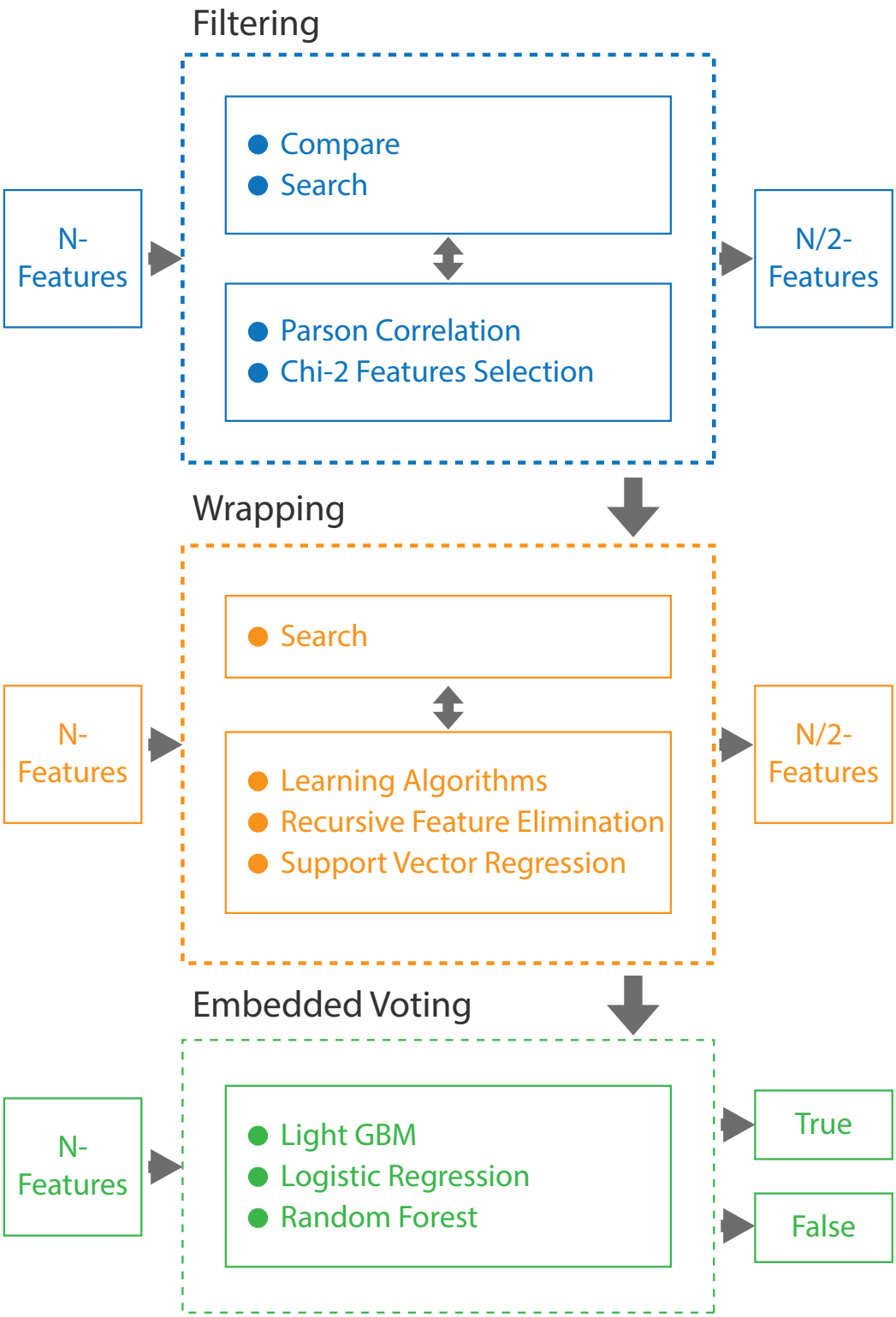


Figure 3.44: Processes of Filtering,Wrapping and Ensemble voting based feauture selection

contribution to the outcome. Each tree is different than another tree which ensures that this trees within the forest are nor highly correlated. It covers all important attribute's perspective as root so it helps to avoid overfitting error. They attempt to reduce gini impurity. They try to identify which subset of trees can reduce the impurity drastically. According to Random Forest estimator, 'Consulted doctor previously for using substance' is the most important attribute. Again, when using light gradient booster because it uses tree based structure .It expands the tree in respect of width in a horizontal manner. This different approach helps to grow the leaf according to max delta loss. 'Light' is used due to faster execution process. It mainly focuses on accuracy. I have used 'gdbt' as boosting parameter. It stands for gradient boosting Decision tree. Number of leaves is 31 by default. We have settled the threshold value parameter of Select from Model as 1.25*Median.One method Get_support returns the index of selected features. Threshold parameter is used for feature selection. It has returned 27 attributes after selection. It detected 'Family member's substance use', 'consulted doctor for substance abuse' as most important features.

Chapter 4

Model Selection and Result Analysis

4.1 Machine Learning

The term Machine Learning was first used by Arthur Samuel in 1959 [25]. Machine Learning is the systematic learning of algorithms and statistical models. A computer uses Machine Learning in order to accomplish a task successfully without using explicit commands, rather take advantage of patterns and interpretation. It is considered as a part of Artificial Intelligence. According to writer Tom Mitchell machine learning is defined as, "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E" [26]. Machine learning algorithms are being used in a wide range of applications. It is used in every sector of Artificial Intelligent such as Image Processing, Computer Vision. Mostly, Machine Learning is preferable where it is infeasible to construct an algorithm of explicit instructions for performing the job. Also, Machine Learning is related to computational statistics in many ways, which basically used for making predictions using computers. In our study we have incorporated a machine learning approach to predict the vulnerability to drug addiction.

4.2 Supervised Learning

Machine Learning is categorized as different types. For example, supervised learning, unsupervised learning, reinforcement learning etc. For our study we have used supervised learning as our problem is associated with both input and output. A mathematical model is built consisting both input and wanted output in supervised learning [27]. The data is recognized as training data which has multiple training samples. One training sample contains one or more input, in our case 60 input and a desired output. Each training data is a matrix and each training example is an array or vector in the mathematical model of the supervised learning. An iterative optimization of a function helps to find out the way to learn the task in order to predict the output [28]. Finally, an input from outside which was not part of the training data is used to be predicted by the system. This is how supervised learning find out the prediction of a problem. However, unsupervised learning algorithms

works with data which only consider inputs, and discover different types of structures in the data, for example, grouping or clustering of data. Again, the task of reinforcement learning is to how software agents are going to take measures in a situation so as to take full advantage of some concept of cumulative return. Our problem was to find out from a test data to predict if the falls under addicted or non-addicted class. For that, we have used training data consist of 60 input features and one output. Our supervised learning algorithms found out the learning method from a data set of about 500 data where 70% were used as training data and rest are for testing data.

4.3 Neural Network Implementation

We have implemented neural networks to generate a model for binary classification. Multilayer perceptions were used as the model was expected to make complex decisions depending on 60 features which requires more than linear separation between classes. In paper [29], it has been discussed that hidden units can facilitate classification task. For our dataset, there were two classes '0' and '1' where the transformation function was $y = C(x)$. Either an individual sample would belong to C_0 or C_1; for every sample x_i belonging to C_i a hidden unit should be preserved so that the weight would be equal to the pattern of class C_i. The output layer should also be chosen using such an activation function so that the output would classify result $x_i \in C_1$ when the neuron in predecessor layer would result in value '1'. As the algorithm is well suited to work with numerical data, we mapped the responses to numeric values. We have selected 30 percent of the data frame as test set which includes 150 samples from total population. Since some of the attribute values are scattered so we applied standard scaler module to scale dataset. Keras sequential model facilitated to build fully connected layers of neural network. In the input layer we have specified input shape as 59 because we had 59 feature attributes except the class label in our dataset. The feature values were added after being multiplied with the weights which acted as input for the activation functions. The equation of calculating weighted sum is given below. Here w_i refers to weights, x_i and $\mathcal{S}$ indicates feature values and weighted sum consecutively.

$$\mathcal{S} = f\left(\sum_{i=1}^{n} w_i * x_i\right) \tag{4.1}$$

Furthermore, we have used dense layer to implement fully connected layers where the first layer had 60 as dense value. Firstly, 'relu' activation function was used in the first layer and 60 hidden units were used to enable the network to learn more complex relationships. Because Rectified linear units have faster convergence property and they solved the vanishing gradient problem. Secondly, two hidden layers were implemented having dense arguments of 20 and 12 which also uses 'relu' activation function. Moreover, we have added advanced activation function Leaky relu as layer to solve the dying relu problem of avoiding negative values by mapping them to 0. Finally, the output layer using 'sigmoid' function had a dense layer of 1 which resulted in binary class values representing the potential person's vulnerability to addiction. 'Sigmoid' function was used for predicting class values as they fall between the range of 0 and 1. We compiled the model with 'binary_crossentropy'

loss function and passed 'accuracy' to metrics argument in order to observe accuracy throughout training. The model was trained for 50 epochs to optimize the model and to make sure that the error on the training data was reduced to satisfactory level. After building the model we have evaluated its performance on test data. Therefore, we have calculated accuracy score 92.67% from our derived confusion matrix. In addition to that, F1 Score was 93% which was calculated from weighted average of precision and recall. The ROC curve is shown in the figure below which was plotted using tpr, npr values obtained from confusion matrix.

$$\begin{bmatrix} 87 & 6 \\ 4 & 53 \end{bmatrix}$$

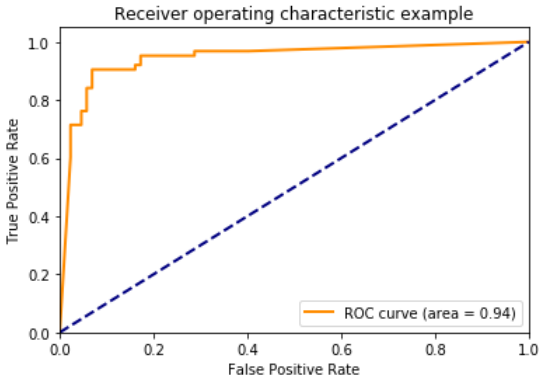


Figure 4.1: Neural Network ROC curve

4.4 Support Vector Machine Implementation

We have used support vector machine algorithm to reduce misclassification errors because SVMs separate classes using an optimal decision boundary. Data preprocessing was performed to divide the data into train and test sets and separating the attributes and class labels. 125 samples representing 25 percent of the total population was used for testing purpose. As our aim was to classify data, we used built-in SVM classifier with linear kernel. Since logistical regression yielded better accuracy in our dataset, it indicated that our data was linearly separable. So, linear kernel achieved accuracy of 90.4% which was better comparing to prediction results of polynomial, gaussian and sigmoid kernels. Linear kernels use the following equation where dot product of x and x_i are performed; x and x_i consecutively denotes the input for prediction and each of the support vectors. The coefficients $\mathcal{B}_0$ and a_i are calculated during the learning process of algorithm.

$$f(x) = \mathcal{B}_0 + \sum_{n=1}^{n} \Big((a_i * (x.x_i)) \Big) \tag{4.2}$$

After evaluating performance from confusion matrix, the obtained result was satisfactory. Because the algorithm has achieved 90.4% accuracy, 94% sensitivity and 88% specificity. The achieved confusion matrix is shown below.

$$\begin{bmatrix} 66 & 9 \\ 3 & 47 \end{bmatrix}$$

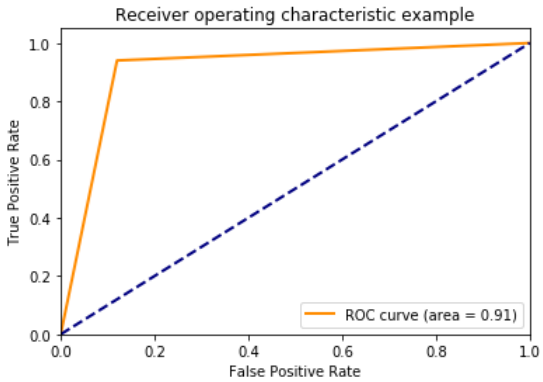


Figure 4.2: Support Vector Machine ROC curve

4.5 Decision Tree Implementation

Primarily, we have used a popular supervised learning approach-Decision Tree to predict outcome. The reason behind using this algorithm is that it helps to predict result because after scaling our data contains mostly conditional control statement/classes.

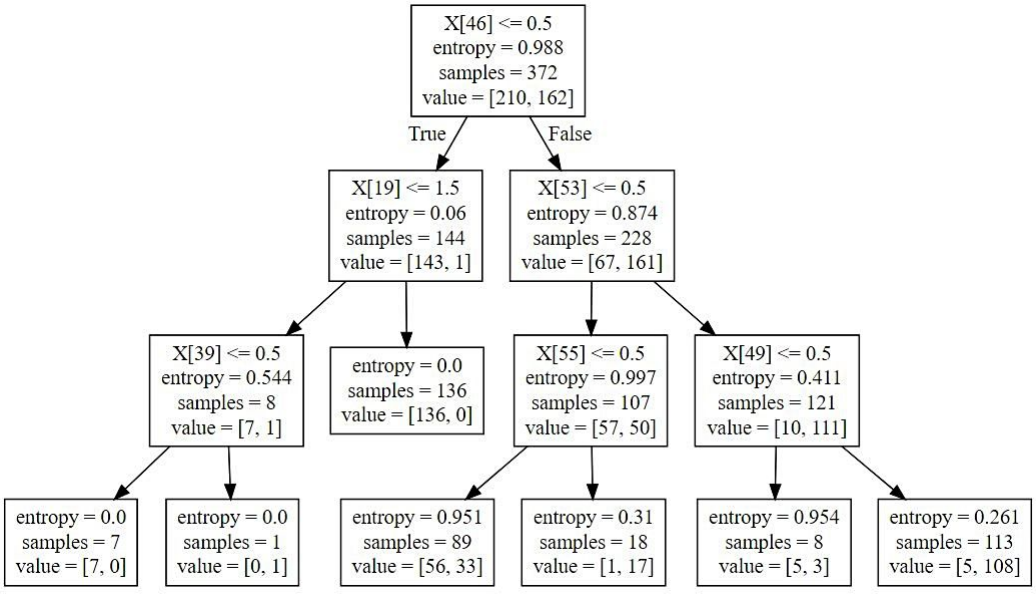


Figure 4.3: Partial view of Decision Tree

This algorithm helped us to detect features best for dataset taken as root – the attribute 'smoking'. It is regarded as a gateway to severe drug addiction. A built-in decision tree classifier method from sklearn was used with passing modified arguments. We have selected criterion parameter to be entropy, random_state parameter as an int so that random seeds could be generated using np. random. The maximum depth of the tree is chosen according to a graph for parameter tuning. In this graph,

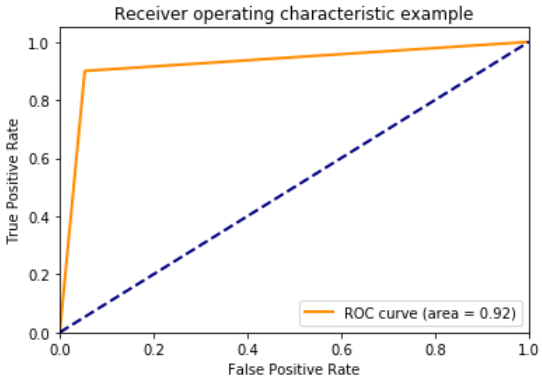


Figure 4.4: Decision Tree ROC curve

we have identified which max_depth value can yield best accuracy of train set. Here, we found the value of max_depth is 57 which gives highest accuracy 88%. Again, from the confusion matrix we have calculated the TPR, 0.8113 and FPR,0.0694. We have used these values to form the ROC curve with AUC=0.96. Sensitivity is 0.9306 and Specificity is 0. 8113.Here the main strategy is to construct a decision tree. This model estimates Exp value for each alternative and select better alternative based on gain value where gain is generated by deducting multiplication of weight average, entropy of child from Entropy of Parent Node. Entropy $= \sum p(x) * \log P(x)$. Entropy also helped us to detect strongest features of questionnaire/dataset. One of the drawbacks of this implementation is that it could be sensitive to small modification on dataset as it may bring major change in final resultant. The generated confusion matrix from model is given below.

$$\begin{bmatrix} 70 & 5 \\ 4 & 46 \end{bmatrix}$$

4.6 Random Forest Implementation

Secondly, we have used Random Forest for classification which uses an ensemble approach consisting of decision trees as building blocks. The strength of this algorithm is that, during tree construction- it chooses training samples randomly with replacement which helps in case of our data as most of the addict's behavior follows a similar pattern. Random features are selected for splitting nodes. In our case, max features for splitting are by default square root of total features which is roughly 7 Increasing number of trees for voting or bootstrap aggregating helps to ensure less bias. Mean Area Under the Curve Score is 0.981 where Accuracy calculated from the generated Confusion matrix is .9333 or approximately 93.33%, f1 score is 0.93. We have assigned the percentage of train test split to be 33% . Total 165 samples have been used for training purpose. To break down, 92 Sober and 73 addict samples has been considered for forming train set. We have used Standard Scaler for scaling our train and test set. It helped to normalize the data within a range. We have plotted a graph for tuning estimator parameter for this algorithm. As the total number of Independent features is less than 60 I formed a graph of training set accuracy versus 'n_estimators' where estimator=67 yields better accuracy. The performance measurements were derived from the following confusion matrix.

$$\begin{bmatrix} 70 & 5 \\ 4 & 46 \end{bmatrix}$$

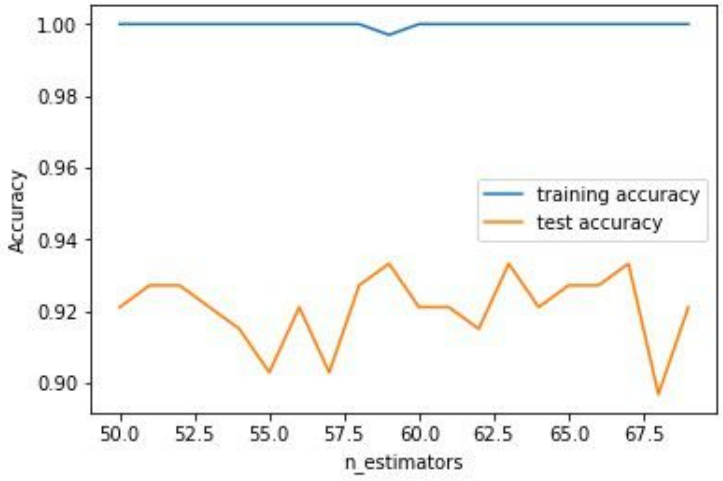


Figure 4.5: Accuracy vs n_estimators graph

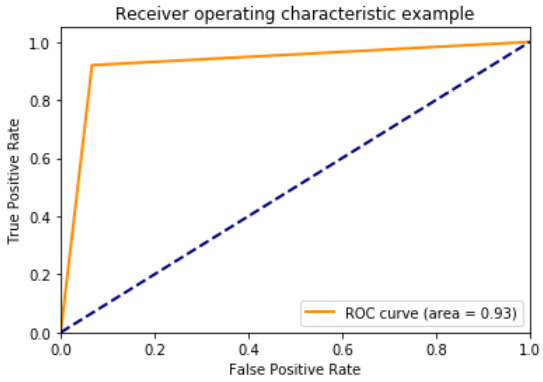


Figure 4.6: Random Forest ROC curve

4.7 Ada-booster Algorithm Implementation

Ensemble machine learning approach was implemented to avoid overfitting problem. A set of classifiers were integrated to facilitate creation of an improved classifier. The series of classifiers $\mathcal{C}=\{c_1,c_2,c_3.\ldots\ldots,c_n\}$ comprised of n low performing classifier. This ensemble method has gained popularity due to it's voting approach. All of the individual classifiers voted and final prediction was achieved based on majority voting. The algorithm utilized several base learners where each of them contributed to generate the final outcome. In addition to that, whole process was executed parallelly by combining different machine learning methods into a single model. In boosting approach, various low accuracy classifiers were merged to gain better accuracy. The low performing model was identified which resulted in misclassification. In each ensemble a model was introduced to correctly classify the instances which were wrongly predicted by previous models. Adaptive boosting classifier [30] set weight of classifiers during training process to ensure classification of unusual instances.

In our empirical analysis, we started with importing the Ada-boost classifier from sklearn library. The train and test were formulated in a 7:3 ratio so that 30% of dataset was preserved for testing purpose. The parameters were chosen carefully; where we have specified 56 'n_estimators' and default value of 1 as 'learning_rate'. After implementing random forest algorithm, we have plotted an accuracy vs 'n_estimators' graph which yielded better prediction rate at 56 'n_estimators'. As a result, 56 'n_estimators' were preferred which guaranteed that boosting process would be terminated whenever we reach maximum number of 'n_estimators'. In each iteration, the misclassified observations were assigned higher weight so that they would get more probability for classification in next iteration. Furthermore, the trained classifiers were also allocated weights according to their respective accuracy. During final prediction process a vote was performed among the classifiers. After incorporating Ada-boost with linear support vector machine as base learner we achieved comparatively less accuracy of 88.0%. On the other hand, using decision tree as base learner we achieved 94.0% accuracy which could be considered as better accuracy. So decision tree was primarily used as base learner or weak learner.The performance measurements were derived from the following confusion matrix.

$$\begin{bmatrix} 86 & 7 \\ 2 & 55 \end{bmatrix}$$

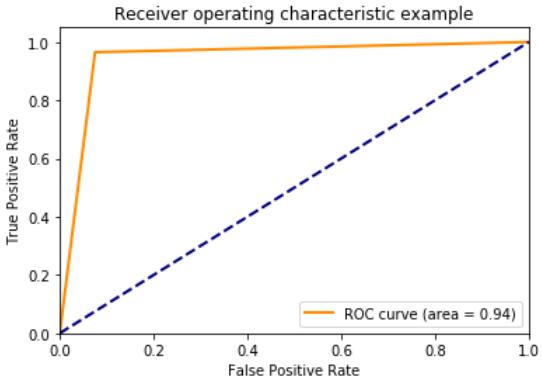


Figure 4.7: Adaptive Booster ROC curve

4.8 Deep belief Network Implementation

We went through trial of different learning techniques to identify which techniques can successfully used for predicting probable Substance abuser with better success rate. One of our trial included a different portion of learning machine known as Deep belief Network with the help of Restricted Boltzmann Machine applied accoridng to the article [31]. As, it has a similarity with Multilayer Perceptron in case of network hierarchy -which yielded better prediction rate in the context of our generated dataset, We used Stacked restricted Boltzmann Machine which has a significance due to a different learning practice. This practice embraces unseen layer of one RBM to be exploited as a visible layer of another RBM in pre-training phase. The Successor RBM uses previous RBM's output as its training input until full training completes. It finishes after all of the hidden layers are completely trained in a

different manner with more depth than MLP. It tries to dynamically advance the model slowly in an iterative way through multiple layer's feedback. Each layer able the opportunity to be well adopted and trained with the input through successive steps. Each layer's feature activation function can be used as an input for further layers. Another advantage of this algorithm is that it is useful for future real world implementation of our problem where the percentage of labelled data set could be very few in number. It utilizes Backward propagation to execute resourceful updates of weight so that learning can become more consistent and effective.

$$\mathcal{B} = -\sum_{i}^{n} A_i - \sum_{j}^{n} B_i - \sum_{i}^{n}\sum_{j}^{n} C_{i,j} \tag{4.3}$$

A = hidden layer units*Bias weights
B = visible layer units*Bias weights
C = visible layers*hidden layer units*weight matrix

It includes undirected graphical connection where parameters are modified to ensure that the generated probability distribution guarantees a fitted model with a sound training procedure. It focuses on gradient based maximum likelihood. Here, there are relationship between different layer but no relationship between variables within same layer. In the first phase, we chose one of the algorithms that yielded standard accuracy of our dataset. We have stacked 3 RBMs where first RBMs hidden layer with adjusted weight which consecutively functioned as an input layer for the second RBM's input training set to ensure a more stochastic and strong classification. After 3 RBMs we also cascaded an AdaBooster Classifier model where RBMs output served as a well fitted input. This resulted with 89.6% percent accuracy where Sensitivity was 92.65 percent and Specificity approximately 86% Precision was 90%.The AUC curve for this Deep Belied Network is shown in 4.8

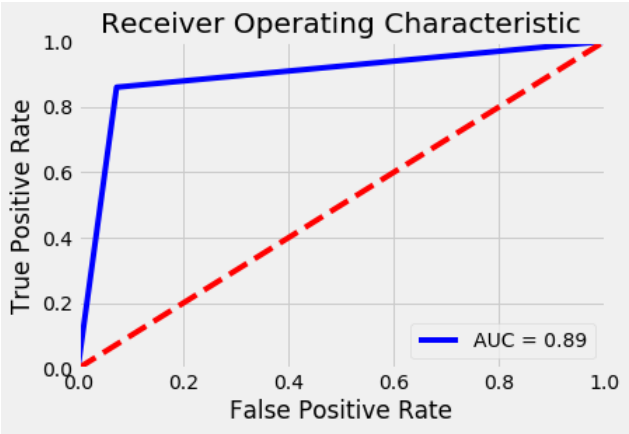


Figure 4.8: ROC curve for RBM and Adaptive booster

In the second phase, we have stacked 2 RBMs with Logistic Regression Classifier. These Deep Belief Network yielded a much better result among the randomly picked up 125 training data, 68 non addicts and 57 addicts were used for initial training purpose. Accuracy of the training was 91.2, Specificity was 0.96, Sensitivity 0.86, Precision 0.91, f1_score was 0.91.the AUC curve under the ROC is shown in 4.9

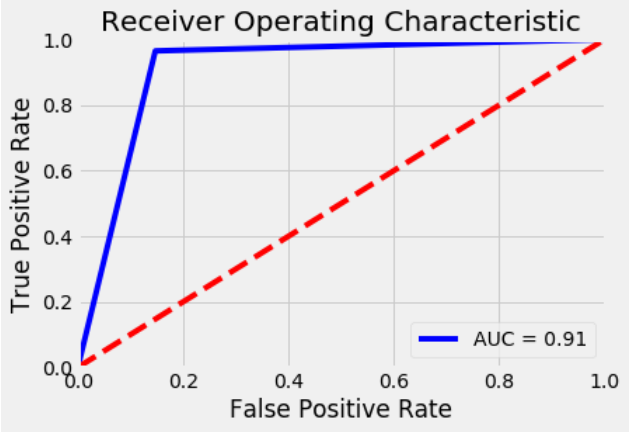


Figure 4.9: ROC curve for RBM and Logistic Regression

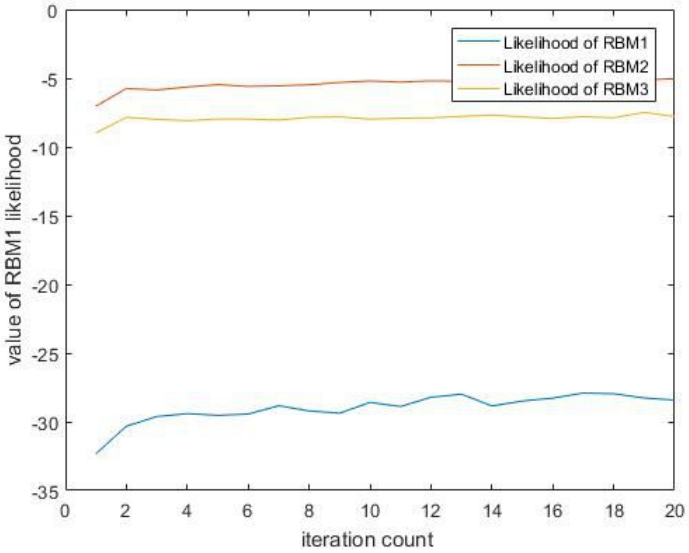


Figure 4.10: Pseudo likelihehood with 20 iterations of RBM1,RBM2,RBM3

In the graph 4.10, we have showed the likelihood of RBM1, RBM 2 and RBM 3's pseudo-likelihood adjustment up to 20 iterations. Pseudo-Likelihood usually indicates estimate of the joint probability distribution of a collection observed dataset. This Figure is a partial graphical representation of the likelihoods. Initially, RBM 1's started from -32.33 and continued to stabilize within the range of -26.28 after 99th iteration. Gradually, it improved and it started from -7.04 while RBM2 used the previous initial unit's hidden layer as it's visible layer it improved up to –4.45 within 20 iterations. Finally, in case of RBM2, modified wright helped to alleviate the likelihood within a steady range of -8.99 to -7.08 till 100 iterations. The graph visualization helps us to interpret mathematically that after the 2nd RBM's 100 iterations, pseudo likelihood started to decrease. So to utilize the maximized likelihood, we have used a deep belief network built with 2 RBMs stacked with Logistic Regression classifier which helped to reach a more higher accuracy of 91.2%, which was satisfactory.

4.9 KNN Algorithm Implementation

K-nearest neighbor shortly known as KNN is an algorithm widely used for prediction in classification problem. For instance, K-nearest Neighbor along with Artificial Neural Network and Decision Tree has been used in the experiment of stock market prediction where they examined the predictability of Dow Jones Industrial Average index and showed that all periods are not equally random [32]. It is non parametric and lazy learning algorithm which is why it needs lesser time in training and a bit longer time in testing. We also used it in our work to predict the accuracy of the model. The training and testing set is divided as 30% is for testing set and the rest are for training set. However, for default value of K which is 5 we got an accuracy of 84%. Then we had looked up for an optimum value of K which would let us know the best number of neighbors for the model. For that, we test every value from 1 to 50 as a K value and saw which one gave us the best accuracy. From Fig. 4.11 we can see that the value of 6, 7 and 29 are giving the highest number of accuracy which is 86%. Study has shown that a lesser amount of neighbors are most flexible fit and will have small amount of bias but high amount of variance and also in this case noise will influence the result. Again, a larger number of neighbors will have a smoother choice area that means small amount of variance but higher amount of bias and in this case the noise will be less but it will be computationally expensive. We don't want noise to influence our result and which is why we took 29 as the K number to go with.

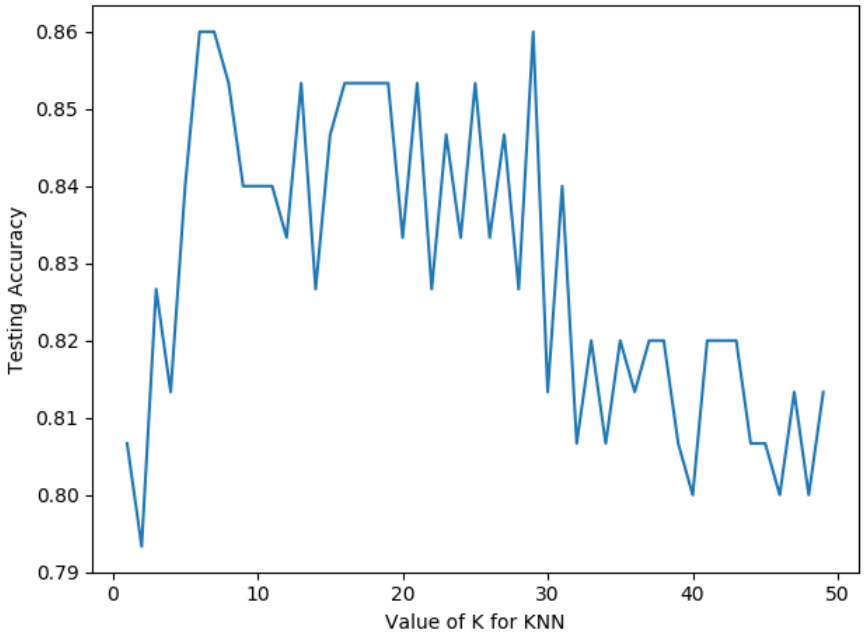


Figure 4.11: KNN accuracy vs k numbers

For this value of K, the accuracy is 86% as mentioned before. From confusion matrix we got the Precision of 0.935, Specificity of 0.736 and Recall of 0.852. We got AUC score 0.863 from ROC curve (Fig. 4.13).

$$\begin{bmatrix} 52 & 1 \\ 13 & 84 \end{bmatrix}$$

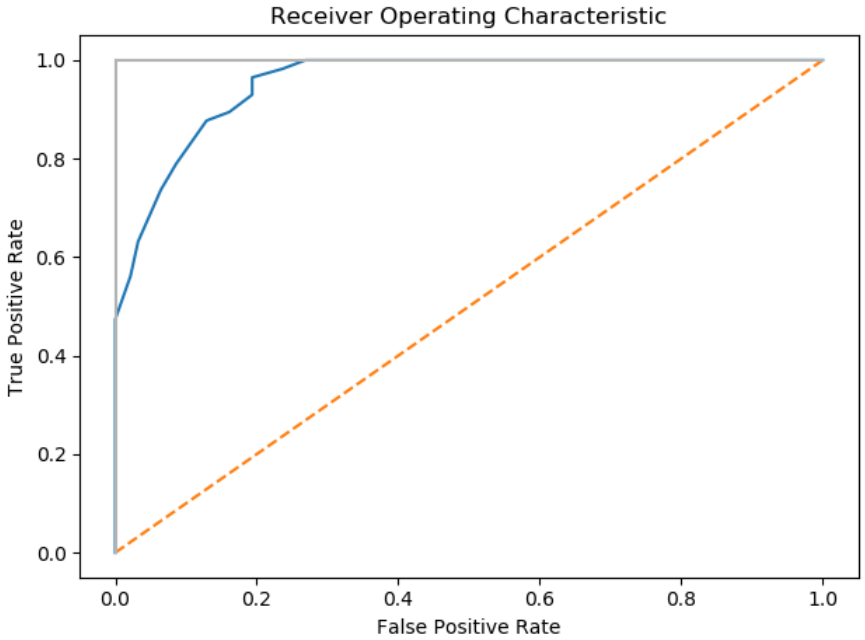


Figure 4.12: ROC curve for KNN algorithm

4.10 Naïve Bayes Algorithm Implementation

Naïve Bayes is one of the fastest algorithm according to the computational cost and a widely used one as well in prediction problems. In one research, scholars have studied the modeling of battery degradation under different usage conditions and ambient temperatures and finally predicted the online state-of-health (SoH) estimation and remaining useful life of lithium-ion batteries [33]. They have used Naïve Bayes algorithm for calculation and also compare the result with the result of Support Vector Machine result. However, in our study we have decided to use Naïve Bayes as it is a fast method to work with. We have particularly used Gaussian Naïve Bayes method in our work and discuss the result of it. We all know the Bayes Theorem as follow:

$$P(A|B) = \frac{[P(B|A) * P(A)]}{P(B)} \tag{4.4}$$

Here, A is class, B is Data. In our example, we have one observation to predict and two possible classes which are 1. Addicted, 2. Non-Addicted or sober. As a result, we will determine two posteriors: one for addicted and one for non-addicted.

P(person is addicted | person's data)=[P(person's data | person is addicted)*P(person is addicted)]/ P(person's data)
P(person is sober | person's data)=[P(person's data | person is sober)*P(person is non-addicted)]/ P(person's data)
From our observation, we had an accuracy of 90% for Gaussian Naïve Bayes method. Besides, the confusion matrix show us the precision of 0.988, specificity of 0.8 and recall of 0.865. The training and testing data are split just as before with a 30% of the data for testing. From receiver operation characteristics we can see the AUC score is 0.97.

$$\begin{bmatrix} 42 & 6 \\ 15 & 87 \end{bmatrix}$$

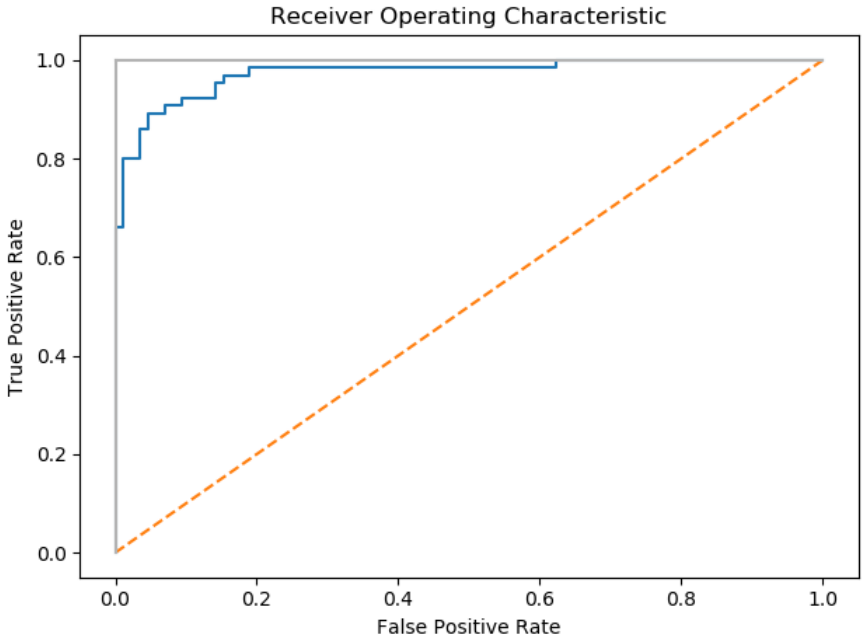


Figure 4.13: ROC curve for Naïve Bayes Algorithm

4.11 Deep Super Learner Implementation

After gathering the experience of classification report, we came to know that neural network worked well with our dataset. But it requires a more complex structure with higher complexity. The number of Epochs for neural network was 200 which also made the process slower but helped to yield better accuracy. The efficient structure required three layers of 60,20 and 12 inputs. So, overall performance's time complexity was higher and the number of passed hyper -parameter were more than traditional algorithm. So, to overcome this time requirement we wanted to use another recent algorithm-Deep Super Learner. In the mentioned paper [34], author approached with a new ensemble algorithm called Deep Super Learner which maximizes the performance of some base learners while minimizing the loss percentage. So, after using the algorithm from github we have achieved a better performance in respect to our 458 samples dataset which contains more than 60 features. For base learners, we have used ExtraTrees Classifier, KNeighbors Classifier, Random Forest Classifier, Gradient Boosting Classifier, Logistic Regression. Exta tree classifier learns using a collection of randomized decision trees exploiting random subsets of data to achieve a better average accuracy. KNN is a parametric classification which works better with input comprising K contiguous data. As, our human samples have a greater range of similarity in many attributes we assumed that it might also help the Super Learner's collaborative approach. Random Forest and Gradient Boosting works better for binary classification with dichotomous data, so we also have included them as base learners. Logistic Regression is also rich in statistical analysis and also works resourcefully for Binary classification.

in the phase of implementation,We have settled proportion of training set as 20 percent. This algorithm uses weight gained from combination of these traditional base learner algorithms to find an ideal combination which yields better and stable accuracy. It uses different learners in a ordered structure based on the report of logistic-loss. It is the non-positive log likelihood which is usually generated for two labels. It also practiced usage of stratified kfold to train and fit the dataset with base learners. All the base learner's precision and recall are shown in the figure 4.14.This deep super learner takes a list of base learners as parameters along with kfolds as attribute where k is assigned to be 5.It also uses weighted prediction to measure performance enhancement in each iteration. Our dataset received highest performance after only 2 iterations where each iteration focuses on adjusting the weights to lessen logistic loss.The above figure 4.14 displayed the precision and recall report per label of each algorithms along with deep super learner. As this algorithm is similar with neural network but the usage of traditional base learners like- RF ,GB ,Logistic regression , KNN ,extra tree classifier etc. helped to improve accuracy with log loss per iteration within a reduced amount of time following a less complex process. From the figure, we can see we can see extreme randomized tree and Random forest produced within range of .92 and .94 . But, both of their precision had a decreasing slope whether Deep Super Learner maintained a steady and stable recall rate.Sequentially, iteration count 0 Loss percentage was 0.24318 and after iteration 1 the loss was .24333.But ,we can also see the Precision rate of Deep Super Learner ranged from .90 to .92 which is satisfactory because this algorithm availed the combination of different base learner and successfully computed the

desired precision and recall rate with lesser time than Neural Network.

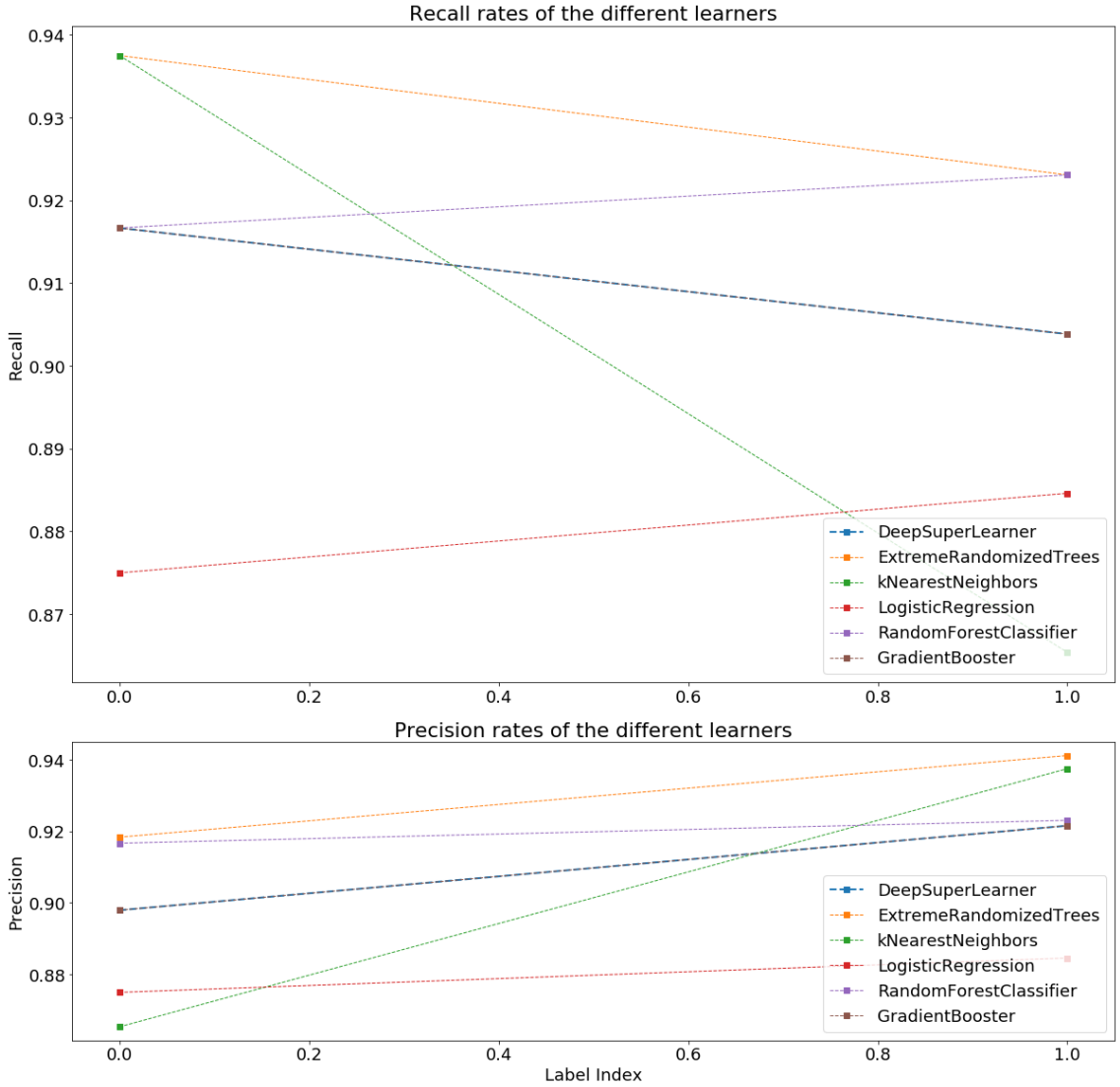


Figure 4.14: Precision and Recall rate for different learners and Deep Super Learner

4.12 XGBoost Algorithm Implementation

Extreme gradient boosting algorithm is a state-of-the-art machine learning algorithm which is used both in classification and regression task. We have implemented this algorithm to yield better result due to its higher performance in benchmark assessment. Because the algorithm is scalable and specially designed to process computations in parallel and a distributed fashion [35]. It is an optimized boosting algorithm which combines weak learners to provide a better prediction rate. Weak learners refer to the models which can produce better result than random predicting. Moreover, for k set of models $\mathcal{M}$=m_1,m_2,... m_k every model m_i assigns weight to the instances classified by the previous model. For an instance, a model m_1 has correctly classified n number of observations, the following model m_2 will allocate lower weights to the correctly predicted observations (n) and comparatively higher weights to the poorly classified observations. The models are collection of classifier

trees which contributes to the prediction task. In addition to that, it is an ensemble approach where explicit scores are allocated to specific trees and it takes advantage of misclassification error to produce a better model in each iteration. Initially, we have used trees as base learner which is default parameter of XGboost algorithm. The data set was converted to Dmatrix which is a special data structure used by this algorithm to gain efficiency. Furthermore, the parameters included learning rate which was set to 0.1 to avoid overfitting problem; 'max_depth' was assigned 30 which specified that the tree could grow up to 30 depth in each iteration. Another parameter 'colsample_bytree' was set to 0.3 which ensured that the model would not be able to overfit. Since we had greater number of features (59), the set of trees was specified to 56 as 'n_estimator'. Here 'n_estimator' referred to the number of trees that would be constructed. The value of alpha was set to 10 which facilitated regularization; so the model was disciplined when it became more compound. Secondly, we formulated the train and test set (25%) and random state was implemented to ensure reproducibility of cross-validation outcomes. Finally, the model was trained and it yielded accuracy of 95.2% and sensitivity of 98%. The true positive rate was plotted against false positive rate in figure 4.18. The achieved confusion matrix is shown below.

$$\begin{bmatrix} 70 & 5 \\ 1 & 49 \end{bmatrix}$$

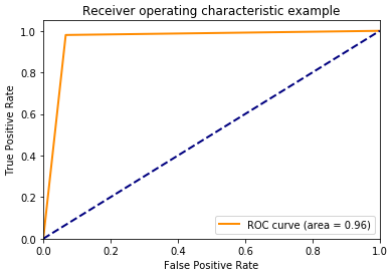


Figure 4.15: Reciever Operating Curve of Gradient Boosting Algorithm

4.13 Results and Analysis

After building the model, the performance was evaluated to determine how correctly the model would be able to predict vulnerability to addiction. The measures that we have used for evaluating performance were based on four parameters of the confusion matrix. The parameters were TP, FP, TN and FN where True positive and True negative indicated how many observations were predicted correctly. Accuracy determined the ratio of the correctly predicted samples to the total samples. The correct prediction rate of classifier was obtained from following equation:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.5}$$

Furthermore, Precision indicated the ratio of correctly predicted positive observations to total positive observations of test set.

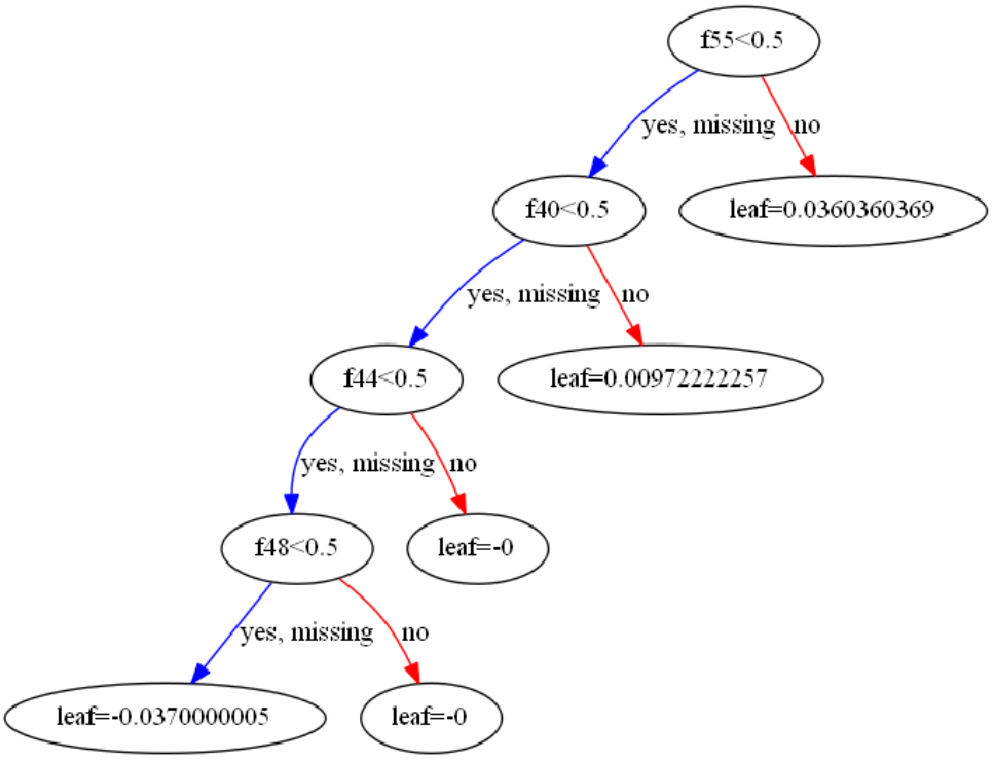


Figure 4.16: Generated tree from Gradient Boosting Algorithm

$$Precision = \frac{TP}{TP + FP} \tag{4.6}$$

Sensitivity specified how accurately the model could predict outcomes of samples comparing to all the actual outcomes in test set. In our experiment, when the model classified an observation as 'addicted' representing the person was vulnerable to addiction; sensitivity recognized the pattern of anticipating correct flag. The equation to compute sensitivity or recall is

$$Sens = \frac{TP}{TP + FN} \tag{4.7}$$

Specificity checked how often the model could predict negative values among all actual negative values. When the model generated outcome 'sober' indicating the individual was not prone to addiction, it actually projected how often the model could predict correct negative outcome.

$$Spec = \frac{TN}{TN + FN} \tag{4.8}$$

Finally, F1-score was the weighted average of precision and sensitivity. It was considered as a better measure of performance because it even worked when the model had uneven class distribution. Equation to determined F1-score is:

$$F1 - score = \frac{2(Sensitivity * Precision)}{Sensitivity + Precision} \tag{4.9}$$

Algorithm	**TN**	**FP**	**FN**	**TP**
Neural Network	87	6	4	53
Random Forest	70	5	4	46
SVM	66	9	3	47
Decision tree	70	5	4	46
Logistical regression	64	11	3	47
KNN	52	1	13	84
Naive Bayes	42	6	15	87
XGB	70	5	1	49
DBN	63	5	5	52
Ada Booster	86	7	2	55

Table 4.1: Generated confusion matrix from models

Performance evaluation of algorithms on our constructed dataset are shown in Table 4.2. Initially the algorithms were implemented on all the features and later on they were implemented on the major characterizing features derived from mRMR feature selection algorithm.

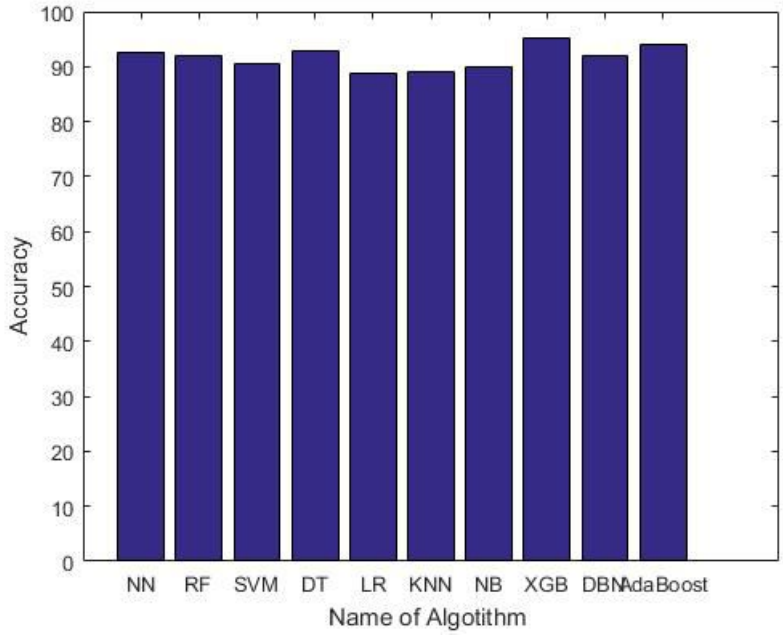


Figure 4.17: Accuracy of all Algorithm

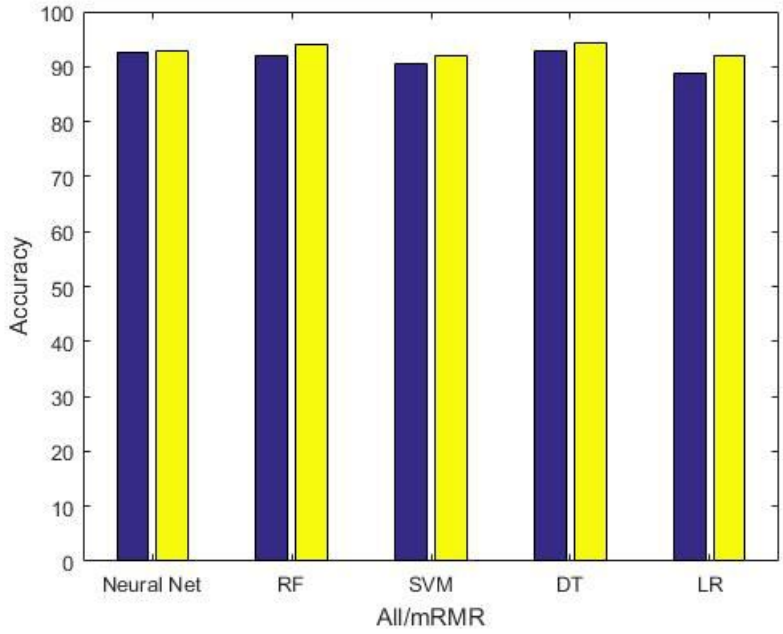


Figure 4.18: Visual representation of Accuracy indicating difference between before(All) after mRMR

Algorithm	**Acc**	**Sens**	**Spec**	**F1-sc**	**Prec**	**Feature**
Neural	92.67	92.98	92.47	93.00	93.00	All
Network	92.72	94.12	91.25	93.00	93.00	mRMR
Random	92.00	92.98	91.17	92.00	92.00	All
Forest	94.00	97.02	91.38	94.00	92.00	mRMR
SVM	90.40	94.00	88.00	90.00	91.00	All
	92.00	94.00	90.67	92.00	92.00	mRMR
Decision	92.80	92.00	93.33	93.00	93.00	All
Tree	94.40	92.00	96.00	94.00	94.00	mRMR
Logistic	88.80	94.00	85.33	89.00	90.00	All
Regression	92.00	96.00	89.00	92.99	93.00	mRMR
KNN	88.97	93.55	73.68	86.50	84.50	All
Nayive Bayes	90.00	86.50	91.00	86.50	98.00	All
XGB	95.20	98.00	93.33	95.00	95.00	All
DBN	92.00	92.65	91.23	92.00	92.00	All
Ada Booster	94.00	96.49	92.47	94.00	94.00	All

Table 4.2: Performance evaluation of algorithms on dataset

Chapter 5

Conclusion and Future Work

The main purpose of our study is to find out the important social, familial and health factors related to drug abuse of the age group of 15-40.It can be said that, if peer influence on smoking, un-prescribed drug use at home, legal subjects, and family responsibilities etc. factors of a person can be connected in one dot, then a potential substance abuser can be found. Family, friends and closely connected people can find out this important factors in a person. Therefore, we can claim that presence of this major indicators in an individual can facilitate identifying his state of addiction severity.However, though significant progress has been recorded, there's no space for satisfaction. Drug use remains at an unacceptable level and continues to bring misery to human beings. It conjointly finances criminal and, to some extent, terrorist activities. A considerable percentage of youngsters across the world still die each year owing to medication, either as an instantaneous results of drug abuse, or indirectly from exposure to infectious diseases, primarily HIV, transmitted by contaminated injection paraphernalia. As we generated a complete data set of drug abuser and non- addicted only in the Dhaka city. So we hope that if further research is intended on this topic, the location or area of the data collection are expected to be much broader all over Bangladesh. Additionally, we have got inadequate time for our in depth analysis as time limitation was set. As a result, we only focused on specific factors like peer influence, un-prescribed drugs use at home, family responsibilities etc. We specified some attributes (60) to identify vulnerability of drug addiction. Moreover, to determine the significant difference between the means of two groups, which may be related in certain features, an implementation of t-test (a type of inferential statistic) is required for more satisfactory result. So this hypothesis testing tool in needed in further extension to allow testing of an assumption applicable to a population. Besides, to get a far better response from the people the scale of the form ought to be reduced. In our questionnaire we have forty eight questions which we found monotonous to answer. For this reason the questionnaire should be more selective and precise in further extension.The major features identified from this research could help us to use a more extended dataset with less features and more sample size.We believe,we could achieve a better classification result by using some of the selected algorithm which yielded better accuracy with modified hyper-parameter tuning. We Hope that, better accuracy can be assembled by using more attributes as we only worked on some certain limited attributes. In terms of accuracy, additional attributes based on these research's outcome should be included in further research.

Bibliography

[1] K. Young, K. Gobrogge, Z. Wang, "The role of mesocorticolimbic dopamine in regulating interactions between drugs of abuse and social behavior", Neuroscience and biobehavioral reviews, vol. 35, no. 3, pp. 498-515, 2011. Available: 10.1016/j.neubiorev.2010.06.004.

[2] K. Kobus, "Peers and adolescent smoking", Addiction, vol. 98, pp. 37-55, 2003. Available: 10.1046/j.1360-0443.98.s1.4.x.

[3] E. Sahker, B. S. Marie, S. Arndt, "Referrals and Treatment Completion for Prescription Opioid Admissions: Five Years of National Data", Journal of Substance Abuse Treatment, vol. 59, pp. 109-114, 2015. Available: 10.1016/j.jsat.2015.07.010.

[4] M. Myers, J. Kelly, "Cigarette Smoking Among Adolescents With Alcohol and Other Drug Use Problems", Alcohol research & health: the journal of the National Institute on Alcohol Abuse and Alcoholism, vol. 29, no. 3, pp. 221-7,2006.

[5] R. Sinha, "Chronic Stress, Drug Use, and Vulnerability to Addiction", Ann N Y Acad Sci, vol. 1141, pp. 105–130, 2008. Available: 10.1196/annals.1441.030.

[6] D. Lloyd and R. Turner, "Cumulative lifetime adversities and alcohol dependence in adolescence and young adulthood", Drug and Alcohol Dependence, vol. 93, no. 3, pp. 217-226, 2008. Available: 10.1016/j.drugalcdep.2007.09.012.

[7] T. Little, K. Schnabel and J. Baumert, "Modeling longitudinal and multilevel data: Practical issues, applied approaches, and specific examples", 2000.

[8] R. Kosterman, J. Hawkins, J. Guo, R. Catalano and R. Abbott, "The dynamics of alcohol and marijuana initiation: patterns and predictors of first use in adolescence", American Journal of Public Health, vol. 90, no. 3, pp. 360-366, 2000. Available: 10.2105/ajph.90.3.360.

[9] D. MacKinnon and C. Lockwood, "Advances in statistical methods for substance abuse prevention research", Prev Sci, vol. 4, no. 3, pp. 155–171, 2003. Available:10.1023/a:1024649822872

[10] E. Sahker, L. Acion and S. Arndt, "National Analysis of Differences Among Substance Abuse Treatment Outcomes: College Student and Nonstudent Emerging Adults", Journal of American College Health, vol. 63, no. 2, pp. 118-124, 2015. Available: 10.1080/07448481.2014.990970.

[11] L. Acion, D. Kelmansky, M. van der Laan, E. Sahker, D. Jones and S. Arndt, "Use of a machine learning framework to predict substance use disorder treatment success", PLOS ONE, vol. 12, no. 4, p. e0175383, 2017. Available: 10.1371/journal.pone.0175383.

[12] E. Byvatov, U. Fechner, J. Sadowski and G. Schneider, "Comparison of Support Vector Machine and Artificial Neural Network Systems for Drug/Nondrug Classification", ChemInform, vol. 35, no. 5, 2004. Available: 10.1002/chin.200405237.

[13] Y. Chen, S. Thosar, R. Forbess, M. Kemper, R. Rubinovitz and A. Shukla, "Prediction of Drug Content and Hardness of Intact Tablets Using Artificial Neural Network and Near-Infrared Spectroscopy", Drug Development and Industrial Pharmacy, vol. 27, no. 7, pp. 623-631, 2001. Available: 10.1081/ddc-100107318.

[14] J. Weinstein et al., "Neural computing in cancer drug development: predicting mechanism of action", Science, vol. 258, no. 5081, pp. 447-451, 1992. Available: 10.1126/science.1411538.

[15] F. Hammann, H. Gutmann, N. Vogt, C. Helma and J. Drewe, "Prediction of Adverse Drug Reactions Using Decision Tree Modeling", Clinical Pharmacology & Therapeutics, vol. 88, no. 1, pp. 52-59, 2010. Available: 10.1038/clpt.2009.248.

[16]]L. Zhang, H. Zhang, H. Ai, H. Hu, "Applications of Machine Learning Methods in Drug Toxicity Prediction", Current Topics in Medicinal Chemistry, vol. 18, no. 12, pp. 987-997, 2018. Available: 10.2174/1568026618666180727152557.

[17] Vulnerability to Drug Addiction Indicators Dataset. [Online] available: https://drive.google.com/drive/u/1/folders/1gv4H50WzEwwj2YQkNGHMzVA5hq4-EzK5.

[18] Drug abusers dataset [Online]. available: https://www.hhs.gov/ash/oah/facts-and-stats/national-and-state-data-sheets/adolescents-and-substance-abuse/united-states/index.html.

[19] E.Fehrman, A.K. Muhammad, E. M. Mirkes, V. Egan and A. N. Gorban, "The Five Factor Model of personality and evaluation of drug consumption risk.", arXiv, 2015.

[20] Drug consumption (quantified) Data Set. [Online]. available: https://archive.ics.uci.edu/ml/datasets/Drug+consumption.

[21] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne",Journal of machine learning research, vol. 9, no. Nov, pp. 2579–2605, 2008.

[22] H. Peng, F Long and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy",IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 8, pp.1226-1238, 2005.

[23] N.Thomas Lal , C. Olivier, W. Jason, E. Andre, "Embedded Methods".Feature Extraction. Studies in Fuzziness and Soft Computing, vol 207. Springer, Berlin, Heidelberg, 2006.

[24] I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh,Feature extraction: foundations and applications. Springer, 2008, vol. 207.

[25] A. Samuel, "Some Studies in Machine Learning Using the Game of Checkers", IBM Journal of Research and Development, vol. 3, no. 3, pp. 210–229, 1959. Available: 10.1147/rd.33.0210.

[26] T. Mitchell, Machine Learning. New York: McGraw-Hill, 1997

[27] S. Russell and P. Norvig, Artificial Intelligence: A Modern Approach, 3rd ed. Prentice Hall, 2010.

[28] M. Mohri, A. Rostamizadeh and A. Talwalkar, Foundations of machine learning. The MIT Press, 2012.

[29] B. Krose, B. Krose, P. van der Smagt, and P. Smagt, "An introduction to neural networks", 1993.

[30] J.Zhu, H. Zou, S. Rosset, T. Hastie, "Multi-class AdaBoost", 2009.

[31] A. Fischer and C. Igel, "Training restricted boltzmann machines: An introduction",Pattern Recognition, vol. 47, pp. 25–39, 2014.

[32] B. Qian and K. Rasheed, "Stock market prediction with multiple classifiers", Applied Intelligence, vol. 26, no. 1, pp. 25-33, 2006. Available: 10.1007/s10489-006-0001-7.

[33] S. Ng, Y. Xing and K. Tsui, "A naive Bayes model for robust remaining useful life prediction of lithium-ion battery", Applied Energy, vol. 118, pp. 114-123, 2014. Available: 10.1016/j.apenergy.2013.12.020.

[34] S. Young, T. Abdou, and A. Bener, "Deep super learner: A deep ensemble for classification problems", pp. 84–95, 2018.

[35] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system", in Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, ACM, 2016, pp. 785–794.

Chapter 6

Appendix A: Prepared Questionnaire

The questionnaire that we have used for data collection purpose is attached below.

This questionnaire is approved and reviewed

by

Mr. Shami Suhrid
Psycho social counselor & lecturer,couseling unit,BRAC University
Member of Bangladesh Counselors' Association
Tasnuva Haque
Phsycho-social counselor & lecturer,BRAC University
National membership in BECPS
Miraz Uddin Ahmed
Chairman & Founder,Challenges rehabilitation centre for addicts
Member of NARCOB Assosciation
Kamrun Nahar Sumi
Psychologist & Addiction Professional,Challanges Bridge to recovery
Azad Siddique
Chairman,Sheba Foundation of Recovery
Member of NARCOB Assosciation
Farhan Rahman Khan
Director & Addiction counselor,Sheba Foundation of Recovery
Rezaul karim
Addiction Counselor & Program cordinator,Promises Centre for de-addiction

Vulnerability to drug addiction prediction (Flagged)

A Questionnaire for collecting primary data regarding research purpose

*Required

The following questions will be used only for research purposes. This will keep the participant name, identity and personal information hidden. We sincerely request you to help and participate in it honestly and openly.

1. **1. What is your gender?** *
   Mark only one oval.
   - ◯ Male
   - ◯ Female

2. **2. In which medium did you study in School/College?** *
   Mark only one oval.
   - ◯ Bengali
   - ◯ English

3. **3. What is your educational qualification?** *
   Mark only one oval.
   - ◯ Primary Education
   - ◯ S.S.C / O levels
   - ◯ H.S.C / A levels
   - ◯ Undergraduate / more

4. **4. What is your nationality?** *
   Mark only one oval.
   - ◯ Bangladeshi
   - ◯ Others

5. **5. Mention your religion:** *
   Mark only one oval.
   - ◯ Hindu
   - ◯ Muslim
   - ◯ Buddhists
   - ◯ Christians

6. **6. How many members do you have in your family?** *
Mark only one oval.

- ◯ Less than or equal 4
- ◯ More than 4

7. **7. How is your relationship with your family members?** *
Mark only one oval.

- ◯ Good relationship
- ◯ Satisfactory relationship
- ◯ Not good

8. **8. What is the range of you Family Income?** *
Mark only one oval.

- ◯ More than 1,00,000
- ◯ 50,000 to 1,00,000
- ◯ 20,000-50,000
- ◯ Less than 20000

9. **9. Amount of money your parent(s) give you daily?** *
Mark only one oval.

- ◯ More than 1000
- ◯ 300-1000
- ◯ Less than 300
- ◯ Nothing / I don't take

10. **10. Is there any addiction in your family of origin?** *
Mark only one oval.

- ◯ Yes
- ◯ No

11. **11. How many friends do you have?** *
Mark only one oval.

- ◯ Few
- ◯ Many
- ◯ None

12. **12. Most of your friends belong to :** *
Tick all that apply.

- ☐ Higher class
- ☐ Middle class
- ☐ Lower class

13. **13. Do you often stay at friend's house?** *
Mark only one oval.

- Yes, too often
- Sometimes
- No, I do not stay

14. **14. Do you borrow money from your friends?** *
Mark only one oval.

- Only in emergency cases
- I borrow money on regular basis
- No

15. **15. What is your marital / relationship status?** *
Mark only one oval.

- Unmarried
- Married
- Divorced
- Widow / Widower

16. **16. Have you ever broken up?** *
Mark only one oval.

- Yes, I did
- Yes, Someone did
- No

17. **17. Do you like to stay out at night?** *
Mark only one oval.

- Yes
- No

18. **18. Do you live with someone who has a habit of using alcohol / substance*?** *
Mark only one oval.

- Yes
- No

19. **19. With whom do you like to spend most of your time?** *
Mark only one oval.

- Family
- Friends
- Alone

20. **20. What is your occupation?** *
Mark only one oval.

- Job
- Business
- Study
- Nothing

21. **21. If you have a job / business, what is the range of your monthly average income?**
Mark only one oval.

- Under 10,000
- Between 10,000-30,000
- Between 30,000-50,000
- More than 50,000 thousand

22. **22. Do you have any type of illegal income?** *
Mark only one oval.

- Yes
- No

23. **23. Do you face any problem in workplace / education?** *
Mark only one oval.

- Yes
- No

24. **23(A). If yes, do you face any of the following problems?**
Mark only one oval.

- Physical illness related problems
- Mental illness related problems
- Problems due to substance abuse
- Others

25. **24. Do you occasionally feel sick or weak?** *
Mark only one oval.

- Yes
- No

26. **25. Do you have any of the following diseases that are causing disruption in your life?**
Mark only one oval.

- Diabetes
- AIDS / STD (Sexually Transmitted Diseases)
- Hepatitis B/C
- Others

27. **26. Do you think you have failed in your life? (Think yourself worthless)** *
Mark only one oval.

- ◯ Yes
- ◯ No

28. **27. Have you ever suffered from the following mental/emotional problem?** *
Tick all that apply.

- ☐ Depression, guilt and inferiority complex
- ☐ Tension and anxiety
- ☐ Insomnia and severe anger
- ☐ Others

29. **28. Identify the following problems that you have suffered for a significant period of time:**
Tick all that apply.

- ☐ Depression, sorrow and hopelessness
- ☐ Anxiety, unnecessarily feeling irritated
- ☐ Hallucinations, see imaginary things, hear voices
- ☐ Lack of attention/ memorization problem

30. **29. Does anyone of your family have mental health related issues?** *
Mark only one oval.

- ◯ Yes
- ◯ No

31. **30. Do you ever have suicidal thoughts?** *
Mark only one oval.

- ◯ Yes
- ◯ No

32. **31. Have you ever attempted to commit suicide?** *
Mark only one oval.

- ◯ Yes
- ◯ No

33. **32. Do you find it difficult to maintain routine? (For example, Daily sleeping early, eating breakfast timely, attending regular work etc.)** *
Mark only one oval.

- ◯ Yes
- ◯ No

34. **33. Have you ever used any substance without the permission of the doctor? (For example, sleeping pills, alcohol etc.)** *
Mark only one oval.

- ◯ Yes
- ◯ No

35. **34. If yes, then what type of substance it was among the ones given below:**
Tick all that apply.

- ☐ Stimulant (Methamphetamine, Cocaine)
- ☐ Sedative (Sleeping Pills, Alcohol)
- ☐ Depressant (Heroin, Fencidil)
- ☐ Hallucinogen (LSD, Piot)
- ☐ Others (Cannabis etc.)

36. **35. In your childhood, did any of your family members use substance?** *
Mark only one oval.

- ◯ Yes
- ◯ No

37. **36. Specify your age when you used substance for the first time (If you did)**
Mark only one oval.

- ◯ Less than 15 years
- ◯ 15-21 years
- ◯ More than 21 years

38. **37. Have you ever hurt anyone or yourself due to anger or losing self-control?** *
Mark only one oval.

- ◯ Yes
- ◯ No

39. **38. Have you ever consulted any doctor for using any substance / alcohol?** *
Mark only one oval.

- ◯ Yes
- ◯ No

40. **39. Do you smoke?** *
Mark only one oval.

- ◯ Yes, everyday
- ◯ Yes, sometimes
- ◯ No

41. **40. How many of your friends smoke?** *
Mark only one oval.
- Most of them
- Some of them
- None of them

42. **41. Do your friends influence you to engage in drug / smoke?** *
Mark only one oval.
- Yes
- No

43. **42. Do you face any withdrawal symptoms when you try to control or stop smoking / using drug for a few days?**
Mark only one oval.
- Yes
- Sometimes
- No

44. **43. Did you sometimes fail to fulfill your social or familial duties because of being engaged in using any substance?** *
Mark only one oval.
- Yes
- No

45. **44. Have you ever taken money from your parents without letting them know?** *
Mark only one oval.
- Yes
- No

46. **45.Have you ever been arrested by police?** *
Mark only one oval.
- Yes
- No

47. **46. Is there any case/lawsuit going on court against you?** *
Mark only one oval.
- Yes
- No

48. **47. Have you ever been arrested for keeping any drug/substance?**
Mark only one oval.
- Yes
- No

49. **48. Have you ever been arrested for selling or dealing drug/substance?**
Mark only one oval.

- Yes
- No

Thank you for participation and your patience