Bachelor of Science in Computer Science & Engineering

Predicting Individual Substance Abuse Vulnerability and Key Risk Factors Identification.

by

Uwaise Ibna Islam

ID: 1504036

Department of Computer Science & Engineering

Chittagong University of Engineering & Technology (CUET)

Chattogram-4349, Bangladesh.

April, 2021

Predicting Individual Substance Abuse Vulnerability and Key Risk Factors Identification.

Submitted in partial fulfilment of the requirements for
Degree of Bachelor of Science
in Computer Science & Engineering

by

Uwaise Ibna Islam
ID: 1504036

Supervised by

Dr. Md. Iqbal H. Sarker
Assistant Professor
Department of Computer Science & Engineering

Chittagong University of Engineering & Technology (CUET)
Chattogram-4349, Bangladesh.

The thesis titled '**Predicting Individual Substance Abuse Vulnerability and Key Risk Factors Identification.**' submitted by ID: 1504036, Session 2019-2020 has been accepted as satisfactory in fulfilment of the requirement for the degree of Bachelor of Science in Computer Science & Engineering to be awarded by the Chittagong University of Engineering & Technology (CUET).

Board of Examiners

____________________________________ Chairman

Dr. Md. Iqbal H. Sarker
Assistant Professor
Department of Computer Science & Engineering
Chittagong University of Engineering & Technology (CUET)

____________________________________ Member (Ex-Officio)

Dr. Asaduzzaman
Professor & Head
Department of Computer Science & Engineering
Chittagong University of Engineering & Technology (CUET)

____________________________________ Member (External)

Muhammad Kamal Hossen
Associate Professor
Department of Computer Science & Engineering
Chittagong University of Engineering & Technology (CUET)

Declaration of Originality

This is to certify that I am the sole author of this thesis and that neither any part of this thesis nor the whole of the thesis has been submitted for a degree to any other institution.

I certify that, to the best of my knowledge, my thesis does not infringe upon anyone's copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my thesis, published or otherwise, are fully acknowledged in accordance with the standard referencing practices. I am also aware that if any infringement of anyone's copyright is found, whether intentional or otherwise, I may be subject to legal and disciplinary action determined by Dept. of CSE, CUET.

I hereby assign every rights in the copyright of this thesis work to Dept. of CSE, CUET, who shall be the owner of the copyright of this work and any reproduction or use in any form or by any means whatsoever is prohibited without the consent of Dept. of CSE, CUET.

Signature of the candidate

Date:

Acknowledgements

This thesis would not have been possible without motivation and support from some important individuals. Throughout the study there were many obstacles, and setbacks which I could only overcome because of the passionate support and motivation of these persons. First of all, I would like to express my gratitude to my honorable supervisor, Dr. Iqbal H. Sarker, PhD, Assistant professor, Department of Computer Science and Engineering, CUET. His passionate support, informative guidance and motivation at every stages of this study helped to maintain my confidence and patience much needed for the completion of this study. In every step, I informed my supervisor about current progress and the overall state of the study. He offered me valuable insights and provided advice in regular intervals which was extremely helpful for me.

I would also want to acknowledge the contribution and support of National Institute of Mental health and the rehabilitation centres and authorities for facilitating the Data collection process. I want to thank all the respondents who provided data for the purpose of this research and specially the addicted respondents who require careful attention.

It is my great pleasure to be able to thank all faculty members of the department who have taught me the major topics pertaining to the this thesis. Their teaching, motivation and support have guided me to fulfilling this degree across these four years.

Finally, I want to thank my father and mother for their unconditional love, support, encouragement, and contribution all throughout my life and academic career in every aspect along the years.

Abstract

Substance Abuse is the unrestrained and detrimental use of psychoactive chemical substances, unauthorized drugs, and alcohol that can ultimately lead a human to disastrous consequences. As patients with this behavior display a high value of relapse, the best intervention approach is to prevent at the very beginning. Thus in this study, we have developed a classifier to identify any individual's present vulnerability towards substance abuse by analyzing subjects' socio-economic environment. By carefully assessing the commonly involved causes, questionnaire was created for collecting data from healthy people and patients suffering from substance abuse to create the data-set. Using Pearson's chi-squared test of independence feature importance was measured to eliminate redundant features using backward elimination. Machine learning classification algorithms (Random Forest, Decision Tree, Logistic Regression, Support Vector Machine, K-Nearest neighbors, and Gaussian Naive Bayes) were trained with the remaining features(18 of 36) to build the classifier. Logistic regression algorithm trained these features yields the best accuracy. This model can successfully identify an individual's vulnerability with an accuracy of 96.72% and AUC value of 0.98. The 18 remaining features were decomposed into 53 factors and these factors were analysed using association rule mining to identify the key factors of substance abuse. This classifier and factors can help provide a pathway for vulnerable people to be enlisted for counseling to facilitate prevention at an early stage.

Keywords — Substance Abuse, Individual Vulnerability, Machine Learning, Predictive Classifier, Risk factors

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

Substance abuse: widely regarded as one of the most alarming problems in the world [1], refers to the harmful or hazardous use of psychoactive substances, including alcohol and illicit drugs. Psychoactive substance use can lead to dependence syndrome - a cluster of behavioral, cognitive, and physiological phenomena that develop after repeated substance use. It typically includes a strong desire to take the drug, difficulties in controlling its use, persisting in its use despite harmful consequences, a higher priority is given to drug use than to other activities and obligations, increased tolerance, and sometimes a physical withdrawal state [2].

About 4.9% adult population of the whole world roughly 240 million people have been identified with alcohol use disorder. Among them 7.8% of men and 1.5% of women, with alcohol causing an estimated 257 disability-adjusted life years lost per 100 000 population. 22.5% of adults all over the world, which is nearly a billion people smoke tobacco-related products. A significant percentage in the global male population 32.0% and a lesser yet large percentage of women 7.0% [3]. Substance abuse varies from country to country, based on demographic factors, demands, and other issues. Prolonged use of drugs can be fatal, from 1999 to 2017 more than 399,000 people died from an overdose in any form of opioid [4]. Alarmingly, opioid is just one of the most used drugs all around the world, other popularly used drugs are alcohol, amphetamine, cannabis, nicotine etc.

Substance abuse disorder is caused by the changes in the reward system in the brain than happens due to continuous abuse of substances [5]. As substance abuse disorder shows a tendency of relapse [6], it is ideal focusing on prevention to mitigate the abuse. So, in this study, we have tried to predict individual vulnerability by identifying key risk factors behind substance abuse which is an effective method for building a prevention model according to studies [7]. After reviewing relevant literature and discussion with local psychiatrists working on substance abuse, the following causes are found to be commonly involved in substance abuse disorder:

1. Peer influence [8][9]
2. Depression [10]
3. Family and relations [11][12]
4. Stress [13]
5. Occupational failure [14]
6. Curiosity [9] [15]
7. Religious affiliations [12]
8. Personality traits [16][17]

After identifying these causes, we have collected data from individuals through online survey and face to face interview with a questionnaire based on these aforementioned factors and therefore, a data-set has been prepared from the answers collected from individuals. After pre-processing the data-set and calculating the contribution of the feature variables on the target variable of the data-set using Pearson's chi-squared test of independence, redundant feature variables are eliminated and remaining ones are used to train and build the desired classifier using machine learning classification algorithms. The performances of these classifiers are evaluated and compared to figure out the optimal classification model. Finally, we used association rule mining algorithm on remaining features of the data-set containing respondents suffering from substance abuse disorder for finding key factors contributing to the the problem.

1.2 Framework/Design Overview

The study was implemented by breaking whole task into small segments and incremental updates. The initial constraint was availability of data and finding the right approach. First similar works performed on this area were examined rigorously. Study [18] was performed with similar kind of objectives. However, we changed the approach towards the implementation. For data collection we tried to extract necessary information regarding risk factors from news articles, scientific journals and other studies. After shortlisting a few factors mentioned in section 1.1, we needed backing from clinicians operating with patients of our importance for a long period. After factors finalization we had to figure out a way for collection of data to work on this factors.

Hence the questionnaire was identified as an effective way for initiation of data-collection. With the questionnaire we had the opportunity to get data from our desired demographic in a systematic approach. Questionnaire however had to go through clearance due to ethical concerns as expected for human data. Data collection can be done with questionnaire through several approaches. Although the SAQ method is easier and faster for obvious reasons, it was not a feasible approach for collection of data from addicted respondents. So, in-person interview method was adopted for data collection.

After data collection it required cleaning for getting it to workable stage. Once the dataset was functional, firstly we had to find the redundant features from the dataset as those features increases dimensions and consumes valuable training time. We worked on finding an approach to effectively eliminate features that failed to show necessary impact. Once the dataset was reduced to retracted size, it was more efficient to build the classifiers by training the algorithms. The ideal classifier had to be chosen out of all the trained ones based on certain measures. Finally, we had to decomposed the remaining features into unique factors to identify the most important factors for helping in the intervention and prevention approach.

1.3 Difficulties

Difficulties faced in this study can be categorized into two major sections. Ones that were identified prior to implementation in the planning stage and the ones which emerged in the later stages of implementation. The difficulties which were identified in the planning and implementation stages are:

- **Problem statement:** Substance abuse is a complex and multifarious factor. It might seem arcane for someone from computer science background. Thus understanding the problem and other fringe elements pertaining to it was the initial difficulty.
- **Psychometric analysis:** Psychometric analysis are difficult because of the wavering nature of human mind and therefore plethora of characteristics which varies from person to person. Moreover, finding better accuracy in this type of analysis is complicated and performance might be inconsistent.
- **Finding data source:** Although collection of data from healthy people is an easy task, addicted people are not willing to provide data voluntarily unless collected from a rehabilitation centre. Moreover, it's not possible to identify individuals suffering from substance abuse disorder externally. So, fixing sources from where sufficient data can be collected for model development was a difficulty.

1.4 Applications

Applying machine learning based method in systems working on substance abuse has been examined thoroughly in this study [19]. Machine learning and other advanced statistical approaches have yielded valuable insights in psychological analysis. This study can be applied in following ways:

1. This study can be used to identify individual vulnerability of substance abuse of any individual.
2. The important risk factors identified from the association rule mining can be used to address vulnerable individual.

3. An automated chat-bot can be developed to provide necessary recommendation to an user considering the vulnerability score and severity.

1.5 Motivation

Substance abuse otherwise known as drug addiction is an alarming problem in our country and worldwide. If a prevention approach can be build using the knowledge of computer science, then it would be a major achievement.

- The huge opportunity machine learning offers for analysis of the psychometric parameters and markers works as a key motivation.
- Population that are related to this frightening phenomenon constitutes a staggering amount of total population. Approximately 7,500,000 (7.5 million) people are suffering form substance abuse disorder and alarmingly more than 1/3rd of these number, about 2,500,000 (2.5 million) are children [1]
- The extent of substance abuse has been increasing at an alarming rate. In 2016, more than 29,000,000 (29 million) pills were seized by law enforcers, more than 35 times the amount confiscated in 2010[2] which is just one of the commonly used substance in the country.

1.6 Contribution of the thesis

The contributions of this study can be summarized as follows:

- Creating a real-time dataset by reviewing relevant literature and discussion with psychiatrists with clear specification.
- Eliminating redundant features by applying efficient approaches to reduce training time and drive better accuracy.

[1] Department of narcotics control, Bangladesh

[2] The Daily Star, Bangladesh

- Predicting the vulnerability of any individual towards substance abuse building a robust classifier by comparing performances of various classification algorithms.
- Calculating most impacting risk factors for substance abuse for facilitating intervention and providing better risk analysis.

1.7 Thesis Organization

In the next parts of the paper, section 2.2 describes related works regarding this study along with their limitations and the improvements we hope to achieve with our proposed study. Section 4.2 provides, a short and precise description of the data-set, section 3.3 delineates the workflow by breaking it into separate modules, section 4.5 contains an exhaustive analysis about the performance and proves the efficacy of the study.In section 4.5.6, the pairs of factors which were observed to be most frequently present in addicted respondent are analyzed. Impact, scopes of future works, and limitations of the study are mentioned in 5.1 and finally section 5.2 concludes the study by describing scope of future work.

1.8 Conclusion

This study is focused on prevention and countering the problem of substance abuse. Knowledge about individual vulnerability helps to provide medical attention for those in precarious situation. Moreover, the risk factors identified can be taken into special consideration while dealing with patients with substance abuse to understand the extent of the problem. The objectives and motivation behind this study is mentioned above. A short description about the problem statement is provided in the beginning section. The challenges and difficulties are also stated.

Chapter 2

Literature Review

2.1 Introduction

Some remarkable works have been done in the field of substance abuse disorder: different causes of substance abuse disorder, the extent of the problem itself, relapse prevention, the impact of substance abuse disorder on daily activities etc. But studies about preventing newcomers from substance abuse are lesser than expected. Related works to this study are mentioned here.

2.2 Related Literature Review

Drug addiction among undergraduate students of private universities in Bangladesh [9]

Sani et al. analyses the factors responsible for drug addiction, identifies types of drugs being consumed among undergraduate students of private universities in Bangladesh from 160 samples in study [9]. Finally, the percentage of respondents receiving treatment and reported relapse into addiction was identified. Age range of respondents in the study is mentioned between 15 to 25 years. This represents the adolescent and early youth demographic within the urban educated population. Major causes for substance abuse were identified as influence from friends, (38.75%), exploring something new for fun (31.88%) and curiosity towards commonly abused substances and the phenomenon itself. A third of all the respondents were found to be involved in frequent substance abuse. Alarmingly, two out of every three respondents who participated in the study had been involved in different forms of substance abuse at least once. What's even more scary is that, even after receiving treatments around 42% have relapsed (resuming substance abuse after intervention through clinical approaches). This whole scenario exposed through this study points to a deepening crisis among youths of the population of this country. Generally, educated individuals are regarded as well-aware of the detrimental physical and psychological consequences caused by drug intake and actions of this sort. Multiple studies completed home and abroad buttress the fact that educated and financially solvent population are less present in the addicted demographic compared to the uneducated and financially weak counterparts. Numbers found in study [9] depicts a precarious situation regarding the problem for the privileged population and the impacts among the underprivileged demographic are even worse. All these evidences point towards the necessity for finding ways to counter the problem.

Through our proposed study we have built a predictive classifier with plans to identify vulnerable individuals at an early stage. The vulnerability predictor can classify vulnerable population with high degree of accuracy and also calculates the key risk-factors for substance abuse. Our study will therefore help to provide solutions to the problems unearthed by this aforementioned one. Moreover, this study focuses on data collected from private university students who doesn't represent the whole substance abuse scenario among students as a significant portion of students are admitted into public universities of Bangladesh. Interestingly, students living in dormitory display a greater prevalence of substance abuse and majority of these students are from the public universities [20]. Using machine learning and other statistical approaches like chi-squared test of independence, this study provides a comprehensive understanding of the problem itself and proposes a way forward for countering it.

A machine learning approach to predict volatile substance abuse for drug risk analysis[21]

Nath et. al develops a machine learning approach for analysing drug risk by predicting the volatile substance abuse. The specific machine learning tool used for the study is ANN (Artificial Neural Network). Two separate ANN modules name ANN-C and ANN-D for implementation of the solution. Two modules are functionally different: ANN-D was developed to predict the volatile substance abuse and ANN-C is designed to predict the duration of use [21]. Several input features are derived using the five-factor model of personality [22] and other research works. These derived input are then fed into the machine learning module ANN-D for binary prediction on individual data. ANN-C however, prognosticates about duration of usage using different units of time. This modules have accuracy scores of 81% and 71.9% respectively. This study uses only one method out of several machine learning classification algorithms that are available. Trying other classification algorithms for the same data-set would have provided an opportunity of comparison between the performances of the algorithms and therefore could have yielded better results. Moreover, it uses data-set from an available machine learning repository online which is not properly specified. For instance, the traits of substance abuse will shift with the change in regional demographic. Thus this model fails to specify the specific demographic it suits and incorporates. The only evaluation metrics used is accuracy score, other evaluation metrics like precision, recall, ROC curves analysis would provide better understanding of the accuracy and performance.

In our study, we have analyzed and compared the performances of classifiers by applying the various machine learning approaches which yield better accuracy and insights. During evaluation we considered important metrics like precision, recall, ROC curves to reach a conclusion regarding effectiveness of the classifier selected. The case and area that was considered while collecting data is also clearly delineated. This provides a better understanding of the usability of the predictive classifier. Furthermore, we have calculated the important risk-factors through backward elimination method. From the 18 risk factors found as the most important risk factor, we applied rule mining for even deeper insights and hence generated rules that contains the most important feature variables responsible for substance abuse. These risk factors should be addressed the most in order to tackle this major problem.

Utility of Machine-Learning Approaches to Identify Behavioral Markers for Substance Use Disorders: Impulsivity Dimensions as Predictors of Current Cocaine Dependence [23]

Ahn et. al aims at identifying the markers of cocaine (a form of substance abuse) dependence for prevention and intervention using machine learning approach. Primary target is to develop on the predicting markers after finding evidence from existing studies that cocaine dependence largely depends on impulsivity. As impulsivity is found to be multidimensional, specifying the exact dimension of it responsible for cocaine dependency becomes a problem. Therefore machine learning approach is adapted for identifying the multivariate predictive patterns for impulsivity phenotype in order to classify individual cocaine dependence. For the data-set they used data from 31 respondents. 23 of them were healthy respondents confirmed by Barratt Impulsiveness Scale - 11 [24] and five other neurocognitive tasks indexing different dimensions of impulsivity. Machine learning predictor can differentiate between individuals with an area under the curve value of 0.912.

This study has significant achievements and a good accuracy in terms of performance. But only drug it focuses on is cocaine and other than that the only factor being analyzed in this study is the impulsivity phenotypes which works as the basis for prediction. The data-set was collected from only 31 respondents which should be extended in order to provide the predictions better recognition. However, we not only have worked on cocaine or any single form of substance abuse but also on every form of substance abuse in Bangladesh for building a more comprehensive model. Furthermore, unlike this model which tries to find variations of a specific characteristic trait, our goal is to identify the most important risk factors for a better intervention approach.

Machine-Learning Prediction of Adolescent Alcohol Use:A Cross-Study, Cross-Cultural Validation [25]

This study was aimed at predicting adolescent alcohol abuse for Australia and Canada. Comparison between seven machine learning algorithms was performed to find the best predictor. Data was collected through 4 and 3 years follow up respectively in case of Canadian and Australian demographic. Canadian demographic yielded a data-set of (n=3,826) where the baseline age is around 13 and the participation of females are around 49%. Australian counterparts however, have a much smaller data-set with number of participants reduced to 2190. Baseline age of these respondents are around 13.5 years and the female fraction consists of 43.7% of the data-set. It predicts different levels of alcohol abuse for mid-adolescence by selecting one comparing these seven machine learning algorithms: Random forest, neural network, logistic regression, support vector machine, lasso regression, ridge regression, elastic-net. Precision, recall, and ROC curve analysis yields elastic-net classification algorithm as the most effective one. AUC values for Australian and Canadian demographic are (0.855±0.072) and (0.869±0.066) respectively.

This study has similar limitations like the previous study [23], in that it also focuses on alcohol abuse only and age demographic is also limited to adolescent only. Although the workflow of this study is similar to our approach, there is no effort for describing the impacting elements like psychopathology and personality contributing to alcohol abused disorder. Further delineation of these two multifarious terms and segmentation would be valuable for specifying the key prime elements. Intervention and preventive approaches will be much better equipped to handle the crisis while better informed regarding these compound terms. On the contrary, our study aims at finding key risk factors after building predictive classifier. Breaking these factors found from the initial method into further prime segments we try to specify every specific risk factors that might be having a greater influence on substance abuse compared to other risk factors. Notably, we started working with all the factors that are commonly involved by working on literature and knowing from psychiatrists and finally shortlisted the most important impacting elements.

New findings on biological factors predicting addiction relapse vulnerability [26]

This study tries to find out the biological factors that contribute to relapse vulnerability. For identifying the clinical, biological and other factors, Sinha et. al reviews the prospective studies which have previously explored risk factors of relapse. They find that several factors: clinical, factors related to patient or behavioral measures like symptoms of depression and stress or drug craving, all these can be used for predicting potential vulnerability of relapse. In case of biological measures, endocrine measures namely cortisol and corticotropin ratio helps find adrenal sensitivity. Neural measures like atrophy of brain in the medial frontal regions in case of withdrawal from substance abuse was important in relapse risk. They also discussed caveats related to specific type of drug being used and the duration of usage. Using these measures, they find the biological markers for relapse risk for identifying individuals most vulnerable in a clinical scenario. This process is aimed at prevention of relapse by facilitating early intervention approaches.

While relapse prevention is important such measures are rare in case of preventing substance abuse disorder at an early age which obviates the necessity of relapse and prevention altogether. It uses biological markers for identifying individuals who stands is particularly riskier position compared to other admitted in a clinical scenario. However, it doesn't provide a solution for relapse risk prevention in case of people who are receiving non-institutional treatment. Our study, on the contrary, works to prevent substance abuse at early stage so that, the amount of newcomers in this dangerous behaviour can be restricted. Such models for identifying individuals particularly at risk of substance abuse should be developed extensively. In this research we developed the predictive classifier with a view to identifying the particular individuals where intervention can help and work as an effective risk aversion formula. Moreover, we have tried to find out significant factors to look out for in case of treating a vulnerable patient so that the whole treatment and intervention method gets easier.

Machine-learning identifies substance-specific behavioral markers for opiate and stimulant dependence [27]

This is a study to with a goal to determining the several cognitive and neurobiological differences among the addicted demographic. However, the majority of population suffering from substance abuse disorder are known to have dependencies on multiple substances making it difficult to discern these differences.

This study attempted to identify the markers for substance abuse classifying two separate type of substance dependencies: heroin dependence and amphetamine dependence. In the data-set that was used, 44 respondents were solely dependant on heroin, 39 solely dependant on amphetamine and 58 respondents were poly-substance dependant, and rest of the respondents, 81 were healthy people in regards to substance abuse. Interestingly, majority of the dependant respondents were in abstinence for longer intervals. Measures that were used to differentiate between two separate classes were demographic, psychiatric, personality traits and impulsivity. All these factors were combined for predictors to be used for the machine learning algorithms to train. They found some promising results in revealing substance-specific profiles that were key for classification of heroin dependence and amphetamine dependence. Out of all the 54 predictors, only psychopathy was associated with both the types which marks good discriminating ability of the model.

The machine learning approach provides valuable insights in classifying Heroin dependence and amphetamine dependence as evidenced by the results. However, the various substance-dependant respondents were on protracted abstinence. Studies have found that neurological and behavioural changes can occur while abstinence from drugs [28]. So, this findings found from the study might conceal some facts that have gone repressed during abstinence. Moreover, the number of respondents used for finding the markers with machine learning was not enough considering the huge amount of data is needed to build an efficient predictive model. In the proposed study, we collected data from people who are currently undergoing treatment and interventions. They are clinically diagnosed as suffering from substance abuse disorder which eliminates the possibility of any changes in psychology of the addicted respondents. Therefore, helping to discern the differences between healthy and addicted respondents respectively. Moreover number of respondents that participated were significantly higher compared to this study resulting in a much comprehensive and inclusive predictive model. Furthermore, this study tries to find difference by using markers of two types of dependencies whereas, we aim to predict vulnerability for all forms of commonly known substances that are involved in the problem.

A computational model of the Cambridge gambling task with applications to substance use disorders [29]

Primary objective of this study is developing a computational model for cambridge gambling task[1] for applying in case of substance use disorder. The model displayed a greater rate of impulsivity for addicted people as it was prognosticated. The cambridge gambling task (CGT) is a widely used neurocognitive approach for assessment of impulsivity in case of both healthy and clinical demographic. This method tries to improve on the efforts of traditional analysis models which fail to fully capture the multiple cognitive mechanisms that engender impulsive actions. A hierarchical approach with bayesian modelling was adopted for the building the computational model. Data-set that was used has healthy respondents of (n= 124). Respondents with histories of substance abuse can be categorized into three categories:

- Heroin dependency (n = 79)
- Amphetamine dependency (n = 76)
- Dependency on multiple substances (n = 109)

So, the final data-set used has data from 382 respondents. Using bayesian comparison between models, **Romeu et. al** identified the best fitting model for examining differences between groups in cognitive model parameters.

This model has effectively used CGT for understanding impulsivity and application in the field of substance abuse disorder. However, there work is limited to only the specified types of substance mentioned in the study. It is well known that substances other than heroine and amphetamine are widely used globally. Moreover, impulsivity is just one of the reasons of substance abuse vulnerability. It also uses the bayesian model only while other methods usage and comparison could provide a more comprehensive model. In our proposed study, we have worked with other risk factors alongside impulsivity and also created an effective predictive classifier after analyzing performances between the several tested machine learning algorithms.

[1] https://www.cambridgecognition.com/cantab/cognitive-tests/executive-function/cambridge-gambling-task-cgt

Predicting Depression Levels Using Social Media Posts [30]

Aldarwish et al. try to find out depression among the youth population by analyzing social media posts taking advantage of the huge amount of people using social media these days. The goal was to try and classify an individual as either depressed or not depressed by analyzing social media posts. Machine learning classification algorithms, SVM (support vector machine) and NB (Naive Bayes) classifiers were used to build the binary classifier.

In study [30], **Aldarwish et al.** do not have a satisfactory *precision* & *recall* value. As people often post in social media with double meaning and all users are not equally expressive about personal life on social media, it is difficult for this classifier to identify actually depressed users

Identifying substance use risk based on deep neural networks and Instagram social media data [31]

Hassanpour et al. aim to build a method that can identify an individual's risk towards the use of alcohol, tobacco, and drugs using social media data. It applies Deep CNN (convolutional neural network) and LSTM (long short-term memory) on images and text consecutively extracted from the *Instagram* profiles of users and create a model to identify the risk of substance (Alcohol, tobacco, prescription drugs & illicit drugs) abuse [31]. This model showed a better success rate in identifying alcohol than other substances abused.

In study [31], Hassanpour et al. also use *Instagram* data to identify individual risk, as alcohol is more socially acceptable compared to drugs, posts containing drugs were self-censored providing less training data (regarding drugs compared to alcohol) to the classifier, inevitably resulting in better accuracy while predicting alcohol abuse. Moreover, this study additionally faces the same problems of study [30]. In our proposed study, we collect data from different patients diagnosed with substance abuse disorder from rehabilitation centers, providing accurate data to our classifier solving the problems regarding both these studies.

Predicting relapse to substance abuse as a function of personality dimensions [32]

Fisher et al. predict the certain personality traits which can be responsible for relapse after receiving treatment for substance abuse. This research examined 108 patients using the 5-factor model of personality used in the neo-personality inventory. After following up, they found out the same patient scored better in the same taste one year later on the personality domain of neuroticism and conscientiousness. The findings suggest that patients with both low conscientious and high neuroticism displayed a greater risk of relapse.

Fisher et al. works on preventing relapse and offers very specific personal traits responsible for the relapse whereas we focus on prevention at a very early stage which is always a better option. Moreover, we intend to incorporate people from random and diverse background to make in an inclusive predictive classifier. Our questionnaire have been prepared with a view to extracting valuable information of the various reasons of substance abuse to build a predictive classifier which can be implemented as a prevention model. We adopted an efficient strategy while selecting the questions by following several key steps to ensure our model encompasses every necessary aspect regarding substance abuse in Bangladesh.

2.3 Conclusion

The aforementioned studies shown in previous section, provides some key findings for execution of the idea in our research. They provide valuable information regarding commonly involved factors of substance abuse. Studies have used machine learning approaches before towards solving substance abuse, relapse or other important aspects of substance abuse. This approaches provide us with an overview and guidance about implementation of the overall framework. We have carefully delineated the limitations of the studies mentioned in the preceding section. Consecutively, we have identified and focused on the areas we intend to improve on and the problems we hope to solve in our study. Finally we described scrupulously how our findings and method overcome the various problems encountered. The goal and periphery elements were also described in fine detail.

2.3.1 Implementation Challenges

Studies have been done for examining the application and translation of computational tools into practice [33]. Computational tools and advanced statistical methods like machine learning have improved the accuracy and predictability performance of various psychometric analysis. However challenges remain about proper utilization and implementation. Creating predictions on large scale multimodal data using computational tools is not efficient. The predictions and analysis often fails to replicate same performance for slight change in demographic. Apart from these problems some important challenges were associated with this proposed study:

- **Dataset formation:** From confirming questions for data collection to the process of data collection itself was pretty arduous and time consuming. Finding key factors from substance abuse in Bangladesh and discussing with specialist for validation of the factors took a lot of effort. Furthermore, collecting data specially from rehabilitation centres was time expensive task. As there are around 20 to 30 patients in the private rehabilitation centres so we had to visit multiple centres in order to collect data.

- **Ethical Approval:** While conducting research on human data ethical review has to be performed and data privacy has to be protected. We collected data in a manner that conceals the identity of the respondents.

- **Optimum classifier selection:** Data-set was not properly balanced as we had more data from healthy respondents compared to addicted ones. So, we had to make sure the effectiveness of the classifier we select performs with equal precision while detecting both classes.

- **Key feature identification:** Apart from developing a model for prediction special impetus was given for finding the factors which contribute the most in case of substance abuse disorder. To find this features we adopted a gradual approach first we eliminated the features to chi-squared test of independence then used association rule mining to identify the key features out of the remaining ones.

Chapter 3

Methodology

3.1 Introduction

This section provides a complete workflow of the study. The major steps of the study is segmented and mentioned in the block diagram in figure 3.1. Each of the blocks are described in detail in section 3.3. These blocks are were implemented by breaking into small segments. While implementation a gradual incremental approach was adopted meaning: one step at a time method was applied during fulfillment of this study.

3.2 Diagram/Overview of Framework

Figure 3.1 provides a complete walk-through of the entire study. First data collection was necessary to obtain a usable data-set. Data collection was completed after into two major steps, 1. Preparing questionnaire and 2. Data collection. Questionnaire was prepared by reviewing literature and consultation with psychiatrists as shown in section 3.3.1.1. With this questionnaire data was collected using two different approaches as described in section 3.3.1.2. Collected data was pre-processed into usable dataset mentioned in section 3.3.2. Once the final dataset was available, redundant features were eliminated using the algorithm in table 3.1 as described in section 3.3.3. Finally, the machine learning classification algorithms mentioned in section 3.3.4 were trained and tested for finding the most effective classifier. Key factors were identified by decomposing the remaining features in section 3.3.3 into factors and association rule mining was performed between them as shown is section 3.3.5.

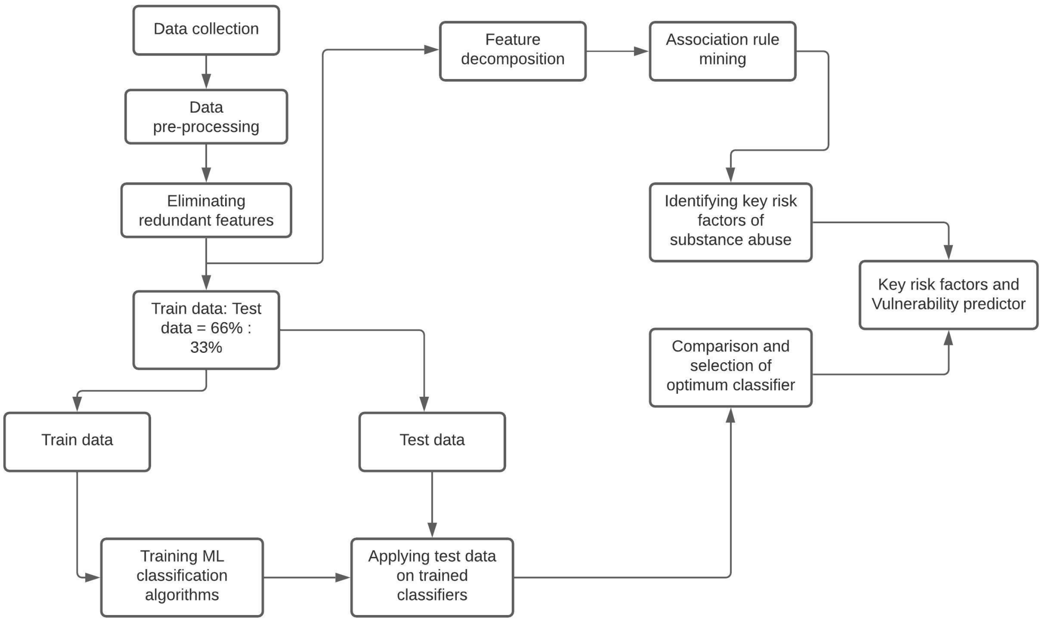

Data collection
Data pre-processing
Eliminating redundant features
Train data: Test data = 66% : 33%
Train data
Test data
Training ML classification algorithms
Applying test data on trained classifiers
Feature decomposition
Association rule mining
Identifying key risk factors of substance abuse
Key risk factors and Vulnerability predictor
Comparison and selection of optimum classifier

Figure 3.1: workflow of the study

3.3 Detailed Explanation

Different block mentioned in figure 3.1 are described in the subsections of this section. Data collection starts off the process, followed by redundant feature elimination and training algorithm and concludes with key factors identification.

3.3.1 Data collection

Data was collected using two separate steps by preparing questionnaire and collection of data from respondents. This two steps are described below:

3.3.1.1 Preparing Questionnaire

Before building the classifier, relevant literature was reviewed and there were discussions with local psychiatrists working on substance abuse disorder. The primary objective is to figure out the commonly involved causes responsible for the problem. Once the necessary literature was reviewed, some common causes are identified: peer influence, family and relationship, occupational failure, depression etc. These causes can vary geographically due to differences in social structure, to corroborate the findings from relevant literature, these causes have been reviewed and verified with the help of local psychiatrists working on this issue. Some special causes were taken into account which is endemic to Bangladesh only, e.g. religious mindset was identified as a key cause behind substance abuse in Bangladesh which is not always the case for many other countries. Considering these types of different aspects of substance abuse key reasons were finalized to be put into the questionnaire.

The questionnaire is prepared to find necessary information from an individual about the causes identified as key reasons behind substance abuse in Bangladesh. Thirty-Six Multiple choice questions (MCQ) formulated the questionnaire assessing reasons, local situations, and other aspects. Questions focus on family relationships, career, financial situation, social condition, peer culture, individual's mental health, and personal view on substance abuse. All of these questions are designed in a sophisticated way to extract the information needed to build a classification model. Once the draft questionnaire has been prepared, it is further examined by specialized psychiatrists. Corrections have been made by making few additions and exclusions to the draft. To make sure this questionnaire is actually capable of obtaining the appropriate information we need for our study, we have performed a pretesting among 25 participants. Once the pretesting has been completed, the questionnaire was ameliorated for the last trial to get the final questionnaire used in the study. Finally, the questionnaire was completed to be used for data collection. Inter-relationships between the features are too small to take into consideration.

3.3.1.2 Collection of respondents' data

The key objective of data collection is to find answers to identical questions from both healthy people and patients, the questionnaire was prepared in that manner. Once we have recorded the responses for the identical questions from both groups, we could analyze which answers were most contrasting between the two groups. The responses which contrast the most are considered as necessary features in the data-set and the ones that don't are redundant. Responses have been collected through an online survey and in-person interviews for healthy participants and patients respectively. Different approaches are adopted for obvious reasons which are explained below.

As for the patients suffering from substance abuse disorder, data have been collected through in-person interviews because of the social stigma and the difficulty in verifying or identifying an individual as a patient. It is inevitably difficult to get a voluntary response from people suffering from this problem because of the negative attitude of society towards an individual known as a patient. Consequently, recognizing and collecting enough individual data required for this study voluntarily by SAQ (self-assessment questionnaire) or other means e.g. online survey from patients is implausible. So, data from these patients are collected by in-person interviews from the patients admitted into rehabilitation centers across Dhaka city with consent. Patients admitted to these institutions solve the problems regarding social stigma and identification. Each interview takes approximately 20 minutes to be conducted.

Data have been collected from healthy people through a google form delivered personally to the recipients by email or via social media from voluntary participants confirmed to have not been involved in substance abuse. Link to the form have been The form contains prefatory statements describing the purpose of the study and asking for permission to let their information be used for the research. Once answers have been collected from both healthy people and patients, these answers are put into two separate spreadsheets for organizing the data-set.

3.3.2 Pre-processing of Dataset

In the spreadsheet, each question is shortened into a variable name which constitutes each column in the data-set and the answers of every single question are replaced with categorical values between '1', '2' or '3' depending on the question being binary or a 3 point-Likert scale question. Finally, a target variable titled "goal" is added to both the data-sets to distinguish between healthy people and patients. Values within the column "goal" is binary: '0' and '1' are assigned to the variable 'goal' in the data-sets of healthy people and patient respectively. These two data-sets are conjoined to produce functional data-set. Column titled ‘age’ was dropped from the final data-set as all the participants belonged to the same age group (18 to 30 years.) and therefore had no impact on the target variable. Finally, we have 36 variables in our data-set with 35 feature variables and one target variable. In case of missing data, if a row contains more than 5 missing data it is purged from the data-set and if a column has missing value it was replaced with the average values within that column. The features of the dataset are mentioned in detail in table 4.1.

3.3.3 Eliminating redundant features

Column variables/features in this data-set are all of the categorical types as it is shown in Table 4.2. To calculate the impact of variables, we had to use a method to find out the relationship between categorical variables. Pearson’s chi-squared test of independence is a non-parametric method that finds the relationship between two categorical variables. We use the chi-squared test of independence to calculate the dependency of feature variables on the target variable. As our data-set is free of the limitations chi-squared test has (sample size is greater than 20, expected frequency of each category in a column is greater than 5, values in each column of the data-set are numerical and randomly distributed), it is a suitable metric for our data-set [34]. All these operations have been performed in *jupyter notebook* which is free and python-based software. . In chi-squared test of independence there is a null hypothesis that all the variables are independent of one another. The chi-squared test calculates the dependency between two variables by creating a contingency table of the two variables. A contingency table (also known as a cross-tabulation, cross-tab, or two-way table) is an arrangement

Table 3.1: Algorithm used for classifier buildup

Step	Action
1	Pre-processing of data-set for feature selection and classifier buildup
2	Applying chi-squared test of independence for feature importance
3	Sorting the features in order of chi-squared statistic, χ^2 value
	For sorted features:
4	.... Training mentioned classification algorithm with available features
5	.... Recording performance of trained classifiers using discussed metrics
6	.... Removing the feature with lowest χ^2 value
7	.... Repeat n times (for n number of features in data-set)
8	Evaluating and comparing the recorded performances between classifiers for selection

in which data is classified according to two categorical variables. The categories for one variable appear in the rows, and the categories for the other variable appear in columns. Each variable must have two or more categories. Each cell reflects the total count of cases for a specific pair of categories. If the table has R number of rows and C number of columns the Chi-squared test of independence statistic, χ^2 is given by the equation:

$$\chi^2 = \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

here, o_{ij} is the observed cell count in the ith row and jth column of the table and e_{ij} is the expected cell count in the ith row and jth column of the table. Expected cell count e_{ij} is calculated by the equation:

$$e_{ij} = \frac{\text{row } i \text{ total} \times \text{col } j \text{ total}}{\text{grand total}}$$

The calculated χ^2 value is then compared to the critical value from the χ^2 distribution table with degrees of freedom, df = (R - 1)(C - 1), and chosen confidence level. If the calculated χ^2 value > critical χ^2 value, then we reject the null hypothesis.

In the case of our data-set, we used the sci-kit-learn library function: $sklearn.feature\_selection.chi2(X, y)$ where X is each of all the feature variables and y is the target variable. This function returns Chi-squared statistic, χ^2 values for each of the feature variables with the target variable by creating a contingency table for each of the feature variables. We have sorted all the values in descending order to list the most impacting features on the target variable.

According to the chi-squared test of independence between two variables the higher the chi-squared statistic value the stronger the relationship is between two variables. There are several statistical thresholds for this value to understand the strength of the relationship between two categorical variables. But, in our case, we have employed backward elimination of features so that we can identify and eliminate the redundant features specifically for this data-set as described in algorithm 3.1. Backward elimination is a technique that starts by using all the features to train the algorithms and then in each step drops the feature with the least chi-square statistic value until the best performing classifier is identified. Features used in the best performing classifier are identified as the necessary features of the data-set. We have started the process by training all 35 feature variables to build the classifier and in each step, we drop the variable with the least chi-squared statistic value χ^2, while recording the performance in each step. Finally, the classifier with the best performance is identified which is trained with 18 features as mentioned in subsection 4.5.1. These 18 feature variables are regarded as necessary features on the target variable in our data-set and the remaining 17 are deemed as redundant hence eliminated from the dataset.

3.3.4 Training machine learning algorithms

The first 35 variables in the data-set are feature variables and the 36$^{\text{th}}$ variable is the target variable. Feature variables and target variables are stored in two separate data-frames. Data-set was split into train data-set and test data-set with a ratio of 66.66% to 33.33% using the sci-kit-learn function *train_test_split*(). Train data-set was fed into machine learning classification algorithms (Random Forest, Decision Tree, Logistic regression, Linear support vector machine, KNearest Neighbors, Gaussian Naive Bayes) to build the predictive binary classifier. The output data of the classifiers are compared with the target variable of the test data-set to observe the accuracy of classifiers.

As we chose the backward elimination technique we started to train the algorithms with all 35 features and in each step, we took out the variable with least Chi-squared statistic value, χ^2 from the data-set and thus the models were being trained with one less feature in every step and the process was repeated until the classifier with the best performance was found. We used the sci-kit-learn library functions $RandomForestClassifier()$, $KNeighborsClassifier()$, *GaussianNB*(), *LogisticRegression*(), $DecisionTreeClassifier()$, $LinearSVC()$ to train and build the binary classifiers. Once training was done this classifiers became capable of predicting the outcome for a completely new entry in the data-set.

Binary Classifier is an intelligent classifier that takes feature variables as input and predicts the possibility of the outcome variable's probability of appertaining to either of the two target variables. In our case, it will take information about an individual(feature variables) as input and predict in percentage whether that individual belongs to the *healthy* category or the *addicted* category.

Decision tree classifier builds classification models in the form of a tree structure. It breaks down a data set into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. Each node represents the decision and the leaf represents the outcome of the decision. For our dataset, the decision tree takes the features as input and the leaf nodes predict whether the individual belongs to the *healthy* or *addicted* category.

Random forest classifier is as its name forest suggests a collection of trees. It creates many decision trees on randomly selected data samples and tries to calculate the outcome. The tree has the best accuracy is selected by voting as the solution. It is very accurate for randomly distributed and big data-set. In our case, the algorithms are trained with the features of the dataset. After that, it predicts the class based on the trained data whether the instance fits for the *healthy* or *addicted* characteristics through voting.

Gaussian naive Bayes is an algorithm having a Probabilistic Approach. It involves prior and posterior probability calculation of the classes in the data-set and the test data given a class respectively. Prior probabilities of all the classes are calculated using the same formula. If we consider our case, it takes the features of individuals and calculates the probability associated with the input features. Finally, it predicts the output for those features based on higher probability.

K-nearest neighbours algorithm uses similarity in features to predict the values of new datapoints which further means that the new data point will be assigned a value based on how closely it matches the points in the training set. The value of K, which determines how much neighboring value of entry has to be compared to classify, should be carefully chosen to get optimal classification performance. It is worth mentioning that the KNN algorithm hasn't any training phase. Considering our case, the algorithm selects a coordinate for input features and predict the class *healthy* or *addicted* based on the majority of the nearest labelled data.

Linear support vector is an algorithm that creates a line or a hyperplane which separates the data into classes. It tries to fit the hyperplane in an optimal position based on the training data and therefore uses it to classify between different categories. Based on the input features, the algorithm selects whether the features belong to the *healthy* or *addicted* part of the trained hyperplane and predicts the class accordingly.

Logistic regression uses a function known as the logistic function or Sigmoid function to build a model. Input values are combined linearly using weights or coefficient values to predict an output value. Output values are binary values that can otherwise be considered as the classes of classification. According to our input features, the algorithm selects a regression line in the training phase with our training data. When we input the data of any individuals, the model selects the position of the feature according to the trained regression line. Finally, the activation function decides whether the instance belongs to the *healthy* or *addicted* category.

3.3.5 Key factors analysis

After performing feature elimination as identifying the best classifier mentioned in section 4.5, List of features in figure 4.5.1, are sorted in order of their respective χ^2 value. This list used for training the classifier trained with apropos classification algorithms eliminated 17 unnecessary features leaves us with 18 remaining features of the dataset. Now, to find out the key factors contributing to substance abuse we analyse 18 remaining features further. First, as each of these features are categorical, so each of these may point to multiple factors.

features	factor no.	factors
belief_in_notion	1	believes that drugs relieve mental stress
	2	have mixed feelings on this notion
	3	doesn't believe drugs relieve mental stress
relation_w_parents	4	very good relation/shares everything in-between
	5	ordinary relations/not too bad
	6	poor relationship/miscommunication
curiosity_in_drugs	7	strong interest in testing new drugs
	8	general curiosity like everything but fearful
	9	no curiosity and think it's dangerous
access_to_drugs	10	have personal access to drugs
	11	have access via friends or other mediums
	12	no access to drugs whatsoever
life_satisfaction	13	fully satisfied/happy with life
	14	not satisfied not disgusted either
	15	unsatisfied/disgusted with life
party_join	16	joins party where drugs are used frequently
	17	doesn't join parties where drugs are used
friends_w_drugs	18	2 or more friends are addicted to drugs
	19	1 or 2 friends are addicted to drugs
	20	none of the friends are addicted to drugs
friends_no.	21	have 10+ friends
	21	have 2-10 friends
	22	have fewer or equal to 2 friends
fam_in_drugs	23	someone in the family takes drugs
	24	someone in the family used to take drugs
	25	no one in the family ever taken drugs
happiness_in_relationship	26	very happy in marriage/relationship
	27	content with current situation
	28	unhappy and looking for a change
curse_for_exp.	29	family always curses for failing to meet expectations
	30	curses sometimes for failing to meet expectations
	31	family never curses and is happy
sex	33	male
	34	female
religious_mindset	35	Has strong religious belief
	36	moderate belief, attends the rituals/weekly prayers
	37	Doesn't have a religious identity, don't care about it
failed_love	38	bitter experiences from past relationships
	39	failed in previous relations but moved on
	40	did not fail in previous relations
	41	have never been in a relationship
risk_tendency	42	Impulsive and often does risky things
	43	calm and not into risky activities
occupational_success	44	successful in career
	45	average career success
	46	failed in career/desired position
occupation_v_friends	47	More successful than most of the friends
	48	successful like most of the friends
	49	less successful than most of the friends
	50	least successful among all the friends
success_in_family	51	most successful among family members
	52	average success among family members
	53	least successful among the family members

Table 3.2: Decomposition of necessary features into factors

For instance, *relation_w_parents* feature can point to any one of these three factors:

1. Very good relation/healthy communication
2. Ordinary not too good, not too bad
3. Very bad relationship/ miscommunication.

As someone's relationship with parents can be categorized into any of these three categories we call these factors. Each of these 18 feature variables were decomposed into the number of categorical column values they represent. If there are three categories of a features like the preceding one, then that feature will be decomposed into three factors. We performed this method for all the 18 remaining features and it yielded 53 factors from these 18 features. Breakdown of these features are described in table 3.2. This factors are analysed through association rule mining and key factors are identified among the 53 factors described in section 4.5.6.

3.4 Conclusion

The sections above delineate the complete list of activities that was performed during this study. The list begin with questionnaire preparation and ends with identification of the key factors. The steps were planned in accordance with the introduction and literature review sections. The performance and efficacy of these methods adopted are analysed in detail in the following section. This process was adopted with a goal to creating a real-time dataset, developing an efficient classifier and finally to identify key risk factors of substance abuse.

Chapter 4

Results and Discussions

4.1 Introduction

In this study, we calculated two important aspects of substance abuse: key factors responsible for substance abuse and individual vulnerability. To identify the redundant features in data-set, a backward elimination method is applied. To measure the accuracy we have used scikit-learn function $accuracy\_score()$ which takes the predicted output from the classifiers of test data-set and compares it with the actual results to compute the accuracy and prints the accuracy in percentage score. In our case, test data-set is the size of one-third of the original data-set so accuracy on the test data set can be considered as a good metric to understand the performance of the classifier on a new entry in other words on a new individual's data. We have compared the accuracy score of all the six different algorithms with features using performance metrics described in section 4.5, to find the classifier with the best output. Starting training with all 35 feature variables and coming down all the way to a single feature variable we discerned that classifier trained with 18 best chi-squared statistic value, χ^2 has the best accuracy. Then, we applied association rule mining on these 18 features to find the key factors of substance abuse mentioned below.

4.2 Dataset Description

After reviewing literature and discussions with local psychiatrists to know about the local aspects of substance abuse, the questionnaire is formed based on the risk factor identified. Then we apply pretesting to finalize a total of 36 questions in the questionnaire. These questions are MCQ (Multiple choice questions) with two or more answers to choose from:

Q1. What is your age?

Q2. What is your Gender?

Q3. What is your Marital status?

Q4. What is your Relationship status right now? (if unmarried)

Q5. Do you belong to a broken family? (parents live separately)

Q6. Are you happy about your marriage or relationship?

Q7. what is the financial situation of your family?

Q8. Is or was any of your family members ever involved in drugs?

Q9. How is your relationship with your parents?

Q10. Do your family often curse you for failing to fulfil their expectations?

Q11. How successful are you among your family members?

Q12. Who are you living with right now?

Q13. What is your present occupational status?

Q14. How do you rate your occupational success?

Q15. How do you rate your workplace?

Q16. What's your occupational status compared to friends?

Q17. What's the number of your friends or accomplices?

Q18. Do you get influenced by your friends' activity?

Q19. How many of your friends have taken drugs so far?

Q20. Do you like to hang out with friends?

Q21. Do you join parties or hang-outs where substance is abused frequently?

Q22. How fast can you get into a society or community?

Q23. Do you have curiosity or fantasy about drugs?

Q24. Do you have access to drugs(could you manage if you wanted)?

Q25. Do you hate a drug addict as a person?

Q26. Are you satisfied with your life until now?

Q27. Are you happy with your physical outlook?

Q28. Are you suffering from any serious illness?

Q29. Have you lost any of your closed ones recently?

Q30. How do you rate your living place?

Q31. What is your religious mindset?

Q32. Are you involved in sports or any form of physical exercise?

Q33. Have you ever failed in love? (break-up, refusal of love)

Q34. Do you enjoy taking risks or having new experiences?

Q35. Do you think that people don't find interest in conversation with you?

Q36. Do you believe in the notion that drugs can relieve you from mental stress?

Table 4.1: feature variables description

key factors	Question no.	feature name	feature type
Family and relations	Q3	*marital_status*	discrete
	Q4	*reln_status*	binary
	Q5	*broken_family*	binary
	Q6	*happiness_in_relationship*	ordinal
	Q7	*financial_condition*	ordinal
	Q8	*fam_in_drugs*	discrete
	Q9	*relation_w_parents*	ordinal
	Q10	*curse_for_exp*	ordinal
	Q11	*success_in_family*	ordinal
	Q12	*living_with*	discrete
	Q33	*failed_love*	discrete
Occupational failure	Q13	*occupation_stat*	discrete
	Q14	*occupation_succ*	ordinal
	Q15	*work_rating*	ordinal
	Q16	*occupation_v_friends*	ordinal
Peer influence	Q17	*friends_no.*	ordinal
	Q18	*friends_influence*	ordinal
	Q19	*friends_w_drugs*	ordinal
	Q20	*hangout_type*	discrete
	Q21	*party_join*	binary
	Q22	*socializing_capability*	ordinal
Curiosity	Q23	*curiosity_in_drugs*	ordinal
Depression & stress	Q26	*life_satis.*	ordinal
	Q27	*outlook_concept*	ordinal
	Q28	*major_illness*	binary
	Q29	*lost_closed_ones*	binary
	Q35	*lack_interest*	binary
	Q36	*Belief_in_notion*	discrete
Social structure	Q24	*access_to_drugs*	discrete
	Q30	*home_rating*	discrete
Personality	Q25	*hate_addict*	binary
	Q32	*physical_exercise*	binary
	Q34	*risk_tendency*	binary
Religious affiliations	Q31	*religious_mindset*	discrete
Demographic	Q1	*age*	discrete
	Q2	*gender*	discrete

Detailed analysis of the questionnaire: 8 questions are binary type questions involving two answers to choose from (Questions no. : 2, 4, 5, 21, 25, 32, 34, 35). The rest of the questions are 3 point Likert scale questions (28). Each of these questions constitutes a column in the data-set. Answers to these questions are categorical which have been later transformed into numerical values '1', '2' and '3' depending on the number of answers. Numerical transformation helped training the machine learning algorithms afterwards. Data is collected separately into two different data-sets for healthy people and the ones suffering from substance abuse separately. A new column titled ‘goal’ is added to both data-sets with values ‘0’ and ‘1’ for healthy people and patients respectively. Lastly, these two separate data-set are conjoined to create the final data-set. This data-set contains 37 columns with 36 aforementioned questions as feature variables and the ‘goal’ column added later on as the target variable which is used for future works in this study.

In table 4.1, a description of all the 36 feature variables is illustrated. The key factor behind each question, question number, the shortened names used in the data-set of the features, and feature type for each of the feature variables. Most of the key factors were identified from the literature review, however, social structure and personality are added subsequently after discussions with local psychiatrists, and questions Q1 and Q2 were background-related questions. 2^{nd} column denotes the question number in the original questionnaire. Feature name shows all the names of the features as used in the data-set and finally, the feature type column provides the answer type of each of the question or feature variable.

A snapshot of the data-set is given in table 4.2. As there are 37 columns in the data-set, it is not feasible to accommodate all the columns in the table, but this microcosm represents the data-set succinctly. In table 4.2, these five columns represent questions no: 7, 8, 9, 10, & 11 respectively. As for the column names, questions are shortened accordingly for convenience during exploratory data analysis. For example, question no. 8: “Is/was any of your family members ever involved in drugs?” is shortened to fam_in_drugs which makes it both

Table 4.2: Snapshot of the final data-set

financial_condition	fam_in_drugs	relation_w_parents	curse_for_exp	occupation_succ
medium	Never	ordinary	sometimes	most
solvent	Never	friendly	sometimes	most
weak	Never	friendly	always	least
solvent	someone was	friendly	never	most
solvent	Never	miscommunication	sometimes	average
medium	Never	friendly	sometimes	average

understandable and easier to use in our data-set. Values within the columns are visibly categorical in nature.

4.3 Impact Analysis

As specified earlier this study was performed help prevention and to facilitate treatment of substance abuse. The study was designed to achieve a high degree of social and environmental impact. The impacts are described in the following sections.

4.3.1 Social and Environmental Impact

This research was performed to address a social issue. Substance abuse has deep underlying social consequences. Left untreated this problem can cause severe trouble to the members within a society. Social impacts of this study can be summarized as:

1. If the more vulnerable demographic can be identified early by applying this research, it will be really helpful for the vulnerable individual as well as the society.

2. the risk factors identified in this study can help to assess the condition of substance abuse within the society members and hence actions can be taken to fix underlying problems.

Although this study doesn't have any direct environmental impact, reducing substance abuse can help to significantly reduce some of the biodegradable elements used for substance abuse. For example bottles used for alcohol abuse, other plastics and synthetic elements that are used in various kind of substance abuse can be reduced which in turn will benefit the environment. However, this can depend on a lot of other peripheral factors.

4.3.2 Ethical Impact

Substance abuse is a problem that is interrelated to several ethical concepts. While in an imbalanced mental state as a consequence of substance abuse, ethical qualities of a human can be compromised. The ethical impacts of this study can be summarized as:

1. This study itself is performed as part of an ethical responsibility: looking after the vulnerable people of the community. Thus is has greater ethical consequences.

2. While in an intoxicated state, studies have found people lose their ethical consciousness. These study aims at tackling the issue which in turn can help solve this dangerous problem by acting upon the vulnerable individuals.

4.4 Evaluation of Framework

As it is a binary classification method, to validate the efficacy of the model, other metrics like $precision$, $recall$ or TPR (True Positive Rate), TNR (True Negative Rate) and $f1-$ $measure$ and their implications are also important. To understand these terminologies a $confusion\ matrix$ has to be explained. A confusion matrix in this study is a 2x2 Table, where each column and row is a specific instance of the predicted class and actual class respectively. Table 4.3 shows a confusion matrix and the terms from the table 4.3 are described below:

Table 4.3: Confusion Matrix

		Predicted	
		Healthy	Addicted
Actual	Healthy	TP	FN
	Addicted	FP	TN

TP (True positive): How many of the outcomes are correctly predicted by the classifier as *healthy* in the data-set.

FP (False positive): How many of the outcomes are incorrectly predicted by the classifier as *healthy* whereas they actually are *addicted* in the data-set.

FN (False negative): How many of the outcomes are wrongly predicted by the classifier as *addicted* whereas they are actually *healthy* in the data-set.

TN (True negative): How many outcomes are correctly predicted as *addicted* in the data-set by the classifier.

$$Precision = \frac{\text{TP}}{\text{TP} + \text{FP}}, Recall = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$TNR = \frac{\text{TN}}{\text{TN} + \text{FP}}, F1 = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

Calculating *Precision*, *Recall*, TNR, and $f1$-*score* we can observe how accurately the classifier is performing in predicting both healthy and addicted classes which is important because sometimes in the case of binary classification, there can be great discrepancies between the performances of two different classes which is not desirable.

We calculated these values and also plotted the *ROC* (receiver operator characteristic) curve and observed the *AUC* (area under the curve) value. *ROC* curves are frequently used to graphically show the connection or trade-off between clinical sensitivity and specificity for every possible cut-off for a test or a combination of tests. It is a plot where False Positive Rate (1-TNR) is placed along the X-axis and True Positive Rate (*Recall*) is placed along the Y-axis. *AUC* is the two-dimensional area under the *ROC* curve which is used for comparing performance between predictive models. *AUC* of a test can be used as a criterion to measure the test's discriminating ability, i.e. how good is the test in a given clinical situation. A higher *AUC* value generally signifies higher predictive performance. Comparing all these measures the best performing classifier is chosen.

4.5 Evaluation of Performance

The first objective of this study is to eliminate the redundant features of substance abuse from the dataset. Using Pearson's chi-squared statistic of independence and backward elimination method 17 feature variables were identified as redundant and hence eliminated.

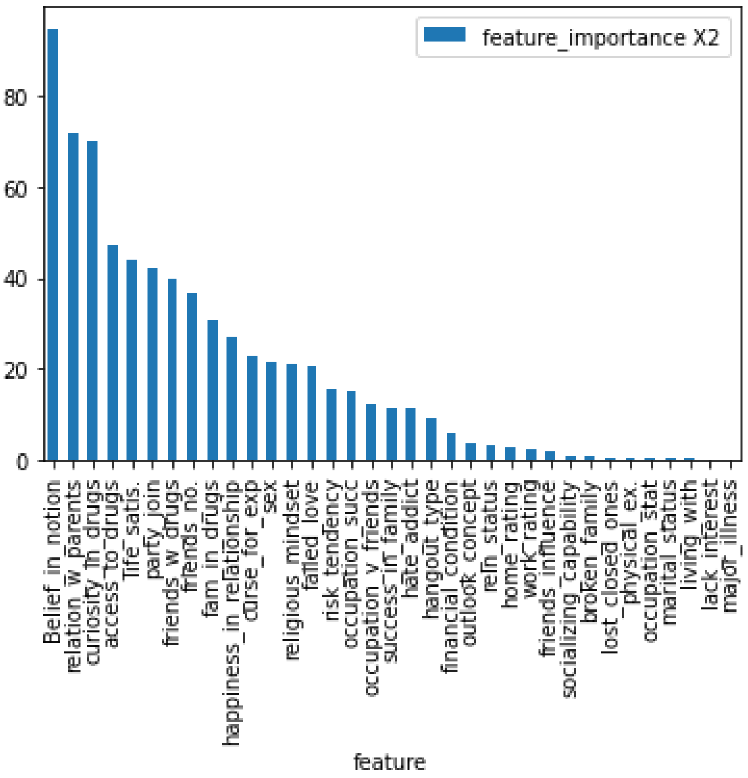


Figure 4.1: Chi-squared, χ^2 feature importance of all feature variables

4.5.1 Chi-squared feature importance

In Figure 4.5.1, the feature importance of all variables is plotted along the Y-axis in descending order of their chi-squared statistic value. If we look at the importance of different features, we can see that $Belief\_in\_notion$, $relation\_w\_parents$, $curiosity\_in\_drugs$, $friends\_no$ & $life\_satis.$ are the 5 features having the best chi-square statistic value, χ^2 and features like $curse\_for\_exp$, $occupation\_succ$, $hate\_addict$, sex & $hangout\_type$ have the least. Taking a closer look at these features, it can be discerned that some of the features have a forward relationship with substance abuse while other features have an inverse relationship. e.g. $relation\_w\_parents$ has an inverse relationship with substance abuse because people having a better relationship with parents were observed to be less vulnerable towards substance abuse disorder. Conversely, $friends\_w\_drugs$ has a direct relation with substance abuse because individuals having more friends suffering from substance abuse disorder were observed to be more vulnerable towards substance abuse.

Now if we take a closer look at 18 remaining features from initial 35, used to train the classifier with best performance as showed in subsection 4.3, questions on *belief_in_notion* and *life_satis*.were based on depression & stress; *relation_w_parents*, *failed_love*, *fam_in_drugs* and *curse_for_exp*, *happiness_in_relat* were based on family & relations; *friends_w_drugs*, *friends_no*, *party_join* were based on peer influence; *occupation_v_friends* & *occupation_succ* were based on career failure and unemployment; *curiosity_in_drugs* was based on curiosity; *access_to_drugs* was based on social structure; *risk_tendency*, *hate_addict* was based on personality; *religious_mindset* was based on religious affiliations and finally *sex* was an introductory question. From the insights we found from the literature review and follow up with local psychiatrists as mentioned in the Introduction section. Presence of all these features among the remaining ones proves that our findings are in compliance with the supporting literature and the Pearson's chi-squared statistic is good choice.

4.5.2 Accuracy scores

As feature selection was performed using the backward elimination method, performance in each step was recorded to compare and analyze between the classifiers. Classifier having the best performance in each iteration was identified and then compared with other best performing models in various iterations. Here, in Table 4.4, the accuracy scores of classifiers are displayed for 4 steps:

Table 4.4: % Accuracy of ML Classifiers trained with features selected by χ^2 values

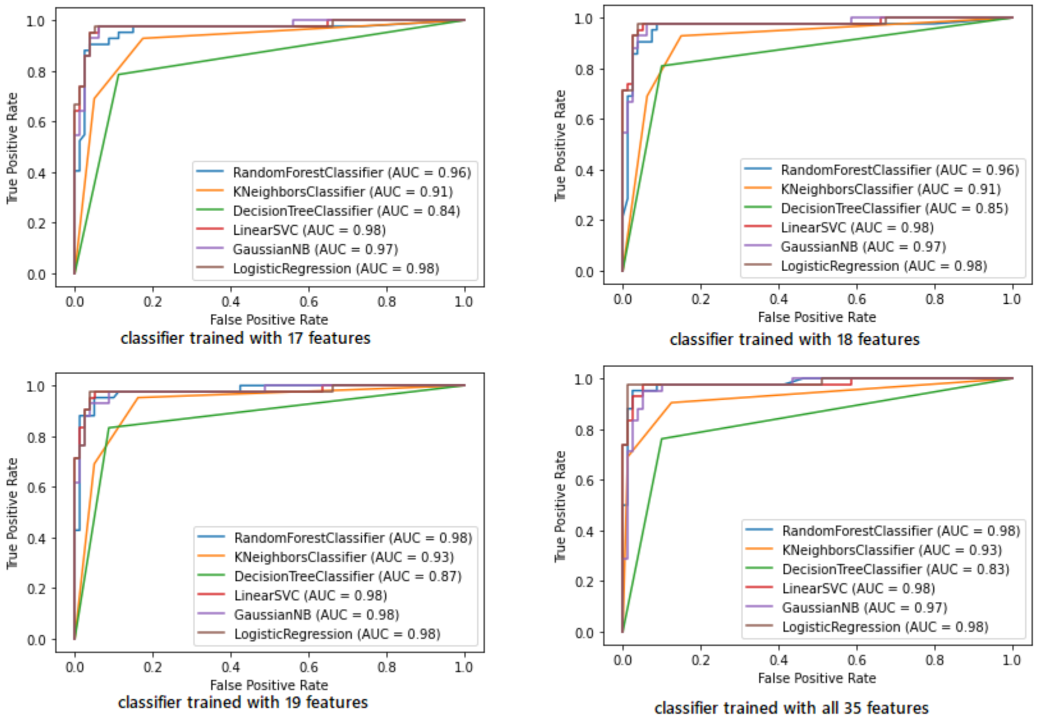
Classifier trained with	17 features	18 features	19 features	all 35 features
Random forest	90.98	92.62	92.62	95.08
KNearest Neighbors	86.06	85.24	86.06	88.52
Decision Tree	85.24	86.88	88.52	85.24
Linear SVC	95.08	95.08	95.08	95.90
Gaussian Naive Bayes	91.80	91.80	90.16	92.62
Logistic Regression	95.08	96.72	95.90	94.26

As displaying the results for all 35 steps is impossible to accommodate and extravagant as well. We, therefore, have displayed the accuracy scores of the models trained with 17, 18, 19 & finally all 35 features respectively to show how the performance of the classifier trained with 18 feature variables performs better than any other classifiers. The best performance in each column was compared between all the columns resulting in the logistic regression classifier trained with 18 feature variables selected by chi-squared statistic value, χ^2 having the highest accuracy score of 96.72%. As mentioned earlier, this accuracy score shows how accurately a model can predict the class of a random individual from the data-set. As the size of the data set is (n= 486) and data is split into a 2:1 ratio for train and test data-set, this accuracy score shows that how accurately the model can predict the class of 162 participants within the test data-set.

4.5.3 ROC curve and AUC value analysis

Analyzing the data in table 4.4, we can see logistic regression classifier has the best accuracy in classifiers trained with 18 and 19 feature variables, whereas linear support vector classifier has the best accuracy in model trained with all 35 features. Apart from classification accuracy we also need to compare the *ROC* curves for each of this classifier. *ROC* curves provide a good description of the discriminating ability and *AUC* values were used to compare between the classifiers. *AUC* values are required because in our study, we needed to understand the performance of the classifiers in case of both *healthy* and *addicted* classes where accuracy score may not explain the full scenario as our data-set contains more *healthy* people compared to ones who are *addicted* or suffering from substance abuse disorder.

To determine exactly how the classifiers are performing, the *ROC* curve for each of the cases mentioned in figure 4.2 which represents classifiers trained with 17, 18, 19 & all 35 features respectively. *AUC* values for each classifier is displayed at the bottom right corner of the images. Calculating *AUC* values together with accuracy scores we can reach a convincing decision towards the selection of our desired model. If we analyze the *AUC* values of classifiers that had the best accuracy score in each case highlighted in table 4.4 and figure 4.2, we see logistic regression has the best combination with an accuracy score of 95.08% with an *AUC* value of .98 in case of classifiers trained with 17 features, for 18 features logistic regression again has the best combination with an accuracy score of 96.72% with an *AUC* value of .98, with 19 features logistic regression has the best combination again with an accuracy of 95.90% and the same *AUC* value



classifier trained with 17 features

classifier trained with 18 features

classifier trained with 19 features

classifier trained with all 35 features

Figure 4.2: ROC curves for classifier trained with 17, 18 and 19 best χ^2, and all 35 feature variables

of .98, finally when trained with all 35 features, linear support vector has the best combination with an accuracy score matching previous logistic regression classifier of 95.90% with again a same AUC value of .98.

4.5.4 Precision and Recall value analysis

Among all the classifiers, logistic regression classifier trained with 18 feature variables has the best combination of accuracy and the best AUC value. Now to consider the other metrics such as *precision* and *recall*, table 4.5 provides a description for each of the classifiers mentioned above for each of the binary classes *healthy & addicted*. *Precision* tells us about the success probability of making a correct positive class classification whereas *recall* explains how sensitive the model is towards identifying the positive class. $f1 - Score$ is the weighted average of *precision* and *recall*. Therefore, this score takes both false positives and false negatives into account. These scores are required in case of imbalanced data-set where entries within the data-set are not equally distributed among the classes, in our case: binary classes.

In Table 5, RF denotes random forest, KNN denotes K-nearest neighbors, DT stands for the decision tree, LSVC means linear support vector, GNB means gaussian naive Bayes, and LGR stands for logistic regression classifiers respectively.

Table 4.5: Precision and Recall scores for classifiers discussed in Table 4.4

Classifier		17 features		18 features		19 features		all 35 features	
	Class	Prec.	Recall	Prec.	recall	Prec.	Recall	Prec.	Recall
RF	*healthy*	0.96	0.90	0.97	0.91	0.97	0.91	0.97	0.95
	Addicted	0.83	0.93	0.85	0.95	0.85	0.95	0.91	0.95
KNN	*healthy*	0.85	1.00	0.85	0.94	0.85	0.95	0.86	0.99
	Addicted	0.88	0.69	0.85	0.69	0.88	0.69	0.97	0.69
DT	*healthy*	0.89	0.89	0.90	0.90	0.91	0.91	0.88	0.90
	Addicted	0.79	0.79	0.81	0.81	0.83	0.83	0.80	0.76
LSVC	*healthy*	0.97	0.95	0.97	0.95	0.99	0.94	0.99	0.95
	Addicted	0.91	0.95	0.91	0.95	0.89	0.98	0.91	0.98
GNB	*healthy*	0.99	0.89	0.99	0.89	0.99	0.86	0.97	0.91
	Addicted	0.82	0.98	0.82	0.98	0.79	0.98	0.85	0.95
LGR	*healthy*	0.97	0.95	0.99	0.96	0.99	0.95	0.99	0.93
	Addicted	0.91	0.95	0.93	0.98	0.91	0.98	0.87	0.98

The weighted average is necessary as we are working with an imbalanced dataset. Among classifiers trained with 17 feature variables, LGR and $LSVC$ has the same highest weighted average $f1 - score$ of 0.95; in case of 18 feature variables, LGR has the best score with an upgrade on the classifier trained with 17 features yielding the highest weighted f1-score of 0.97; with 19 features LGR again has the equal-weighted $f1 - score$ of 0.96, and while trained with all the features, $LSVC$ matches LGR with a weighted $f1 - score$ of 0.96.

4.5.5 Comparative analysis between the classifiers

As we have calculated accuracy scores, $precision$, $recall$, $f1 - scores$ and ROC curve with AUC values, we can compare these evaluation scores among all the classifiers to understand the effectiveness and therefore compare performances among all the classifiers. Figure 4.3 shows a comparative analysis, between the best performing classifiers mentioned in table 4.4, table 4.5 and figure 4.2 added with all the classifiers trained with feature variables ranging from 10 to 20. AUC values and weighted average $f1 - scores$ is multiplied by 100 as the accuracy is being used in percentage. As we can see, the logistic regression classifier, trained with 18 features, has the best accuracy score of 96.72%, best average weighted $f1 - score$ of 0.97, and a joint best AUC value of 0.98. The only classifier that has a similar AUC value is the linear support vector classifier trained with all

35 features but has both a lower accuracy score of 95.90% and weighted average f1-score of 0.96. So, comparing all the classifiers it can be seen that logistic regression classifier trained with 18 feature variables selected by the chi-squared statistic value, chi^2 has the best performance among all the classifiers.

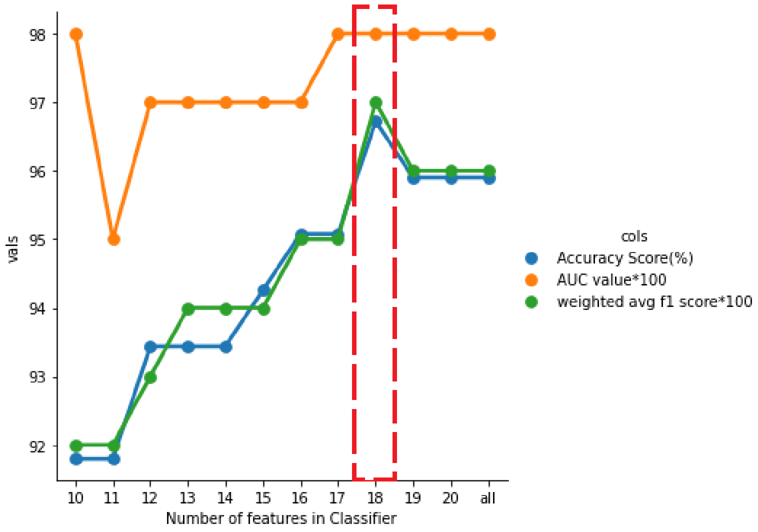


Figure 4.3: Comparison between best performing classifiers

The reason that performances of only four cases were shown in the table 4.4, table 4.5 and figure 4.2, figure 4.3 because it will be imprudent, redundant, and nearly impossible to accommodate all necessary aforementioned evaluation scores for each case with 35 feature variables meaning 35 possible cases. The performance of classifiers trained with less than 10 variables gradually declined with the exclusion of each feature variable in every step. On the contrary, the classifiers trained with more than 20 variables had accuracy scores equal to or less than the ones trained with 20 feature variables which are evident in the performance of the classifier trained with all 35 feature variables. Moreover, training classifiers with as few feature variables as possibles not only reduces the time required for training but also eliminates the possibilities of over-fitting. Considering all these factors, logistic regression classifier trained with 18 feature variables selected by highest chi-squared statistic value, χ^2 is the most effective model to differentiate accurately between *healthy* and *addicted* classes with an accuracy score of 96.72%, weighted f1 average of 0.97 and AUC value of 0.98. This classifier can provide the affinity of an individual towards either of these two classes in percentage scores. Individuals having a greater percentage of affinity towards *addicted* class will therefore be requested to enlist themselves in prevention counseling programs as necessary. A person having a higher percentage of affinity towards the class *healthy* can therefore be excused from further proceedings.

Analyzing the result, an accuracy score of 96.72% can be considered as a good score which is the percentage of data correctly identified by the classifier. *Precision* and *recall* scores were well above 0.90 for both classes *healthy* and *addicted* which shows that the classifier is predicting in both cases. As for the *AUC* value, in the case of researches regarding psychological diagnosis, scores above 0.70 are considered as a good *AUC* score and here the value is significantly greater than that standing at 0.97.

4.5.6 Key risk factors of substance abuse

The 53 factors mentioned in table 3.2 can be examined to monitor which of these 53 contribute the most in substance abuse. To find that out, we have to find an appropriate method to calculate the key factors from the 53 factors available.

Apriori is an association rule mining algorithm that identifies concurrent items within a dataset calculating frequency of the items. We used this algorithm to identify which factors were concurrently present in case of addicted respondents. Rules were mined containing various factors and their perfomance was measured using the *support*, $lift$ and $confidence$ between the factors of any rule. The calculated relationship between features of left hand side and right hand side are labelled as rule. In a rule : A => B, confidence = 80%; means the possibility of B being present when A is present already is 80%. *Apriori* identifies the relationship features in terms of these three metrics mentioned in the preceding sentence which given us important understanding of the psychological behaviour from the data-set. These metrics are calculated by:

Support: identifies the frequency of any factor within the data-set.

Lift: measure of the percentage of two factors being present simultaneously instead of being present alone in the data-set.

Confidence: denotes the possibility of a factor being present when another factor is also present.

Calculating these metrics for the 53 factors present within the data-set of "addicted" people consisting mostly of urban youth in Bangladesh, we can find which factors have higher values and hence are more influential behind the problem. Threshold values were set to eliminate the redundant rules. From the ones that reaches the threshold, we picked top 10 rules based on the set threshold values of aforementioned metrics. Table 4.6 provides detailed analysis of these rules.

Table 4.6: rules for identifying key risk factors

Data set	Discovered Rules	confidence
addicted	If (Unsatisfied in life, cursed by family members for failing to meet expectations, male), then addicted	86.7%
	If (Bad relation with parents, have access to drugs, bitter experience from past relationship), then addicted	81.8%
	If (Unsatisfied in life, have access to drugs, unsuccessful career), then addicted	90%
	If (Bad relation with parents, joins party with drugs, number of friends 5-10, male) then addicted	90%
	If (Unsatisfied in life, have access to drugs, more than 2 addicted friends), then addicted	90.9%
	If (Curious about various drugs, unsuccessful among family members, unsuccessful Career believes drug addiction relieves from stress) then addicted	90%
	If (believes drug addiction relieves from stress, joins party with drugs, Moderately religious, male, least successful in family), then addicted	88.9%
	If (have own access to drugs, more than 2 addicted friends, bitter experience from previous relationships, joins party with drugs), then addicted	82.6%
	If (bad relationship with parents, no feelings regarding life satisfaction, joins party with drugs, bitter experience of past relations, least successful among friends, have access to drugs via friends) then addicted	83.3%
	If (believes drug addiction relieves from stress, joins party with drugs, Male, moderately religious, least successful among friends, least successful among family members), then addicted	83.3%

These 10 rules were handpicked from all the generated rules based on the performance. Closely examining these rules we can see each of these rules is a combination of 53 factors. This rules will give us a comprehensive understanding about the key risk factors. The factors on the lest hand side are known as antecedent and the factor on the right hand side is know as consequent. In table 4.6, first rule specifies that, for male respondents in our dataset, if a person is unsatisfied in life, continuously cursed by family members for failing to fulfill their expectations then he has a 86.7% probability of being addicted. If a person is male, has bad relation with parents, has access to drugs and bitter experience from past relationship then he has a 81.8% probability of being addicted. When unsatisfied with life, has access to drugs, unsuccessful in career, , and male then 90% probability of being addicted. If a respondent in male, unsatisfied in life, have access to drugs and more than 2 addicted friends, then 90.9% probability of being addicted. If curious about various drugs, has unsuccessful in terms of career among family members, unsuccessful career and believes that drug relieves one from mental stress then 90% probability of being addicted. If a person believes drug relieves one from mental stress, joins party where drugs are consumed, is moderately religious and least successful among family members, and male then he has 88.9% probability of addicted. In case of a respondent having own access to drugs, joining parties where drugs are taken, having more than 2 addicted friends and having bitter experience from past relationship then there is a 82.6% probability of that person being addicted. If a respondent has bad relationship with parents, no feelings regarding life satisfaction, joins party where drugs are used, bitter experience about previous relations, have access to drugs via friends, and least successful among family members then that person has 83.3% probability of being addicted. Finally, if the respondent believes that drugs relieve mental stress, joins parties where drugs are used in a male, is moderately religious, is least successful among friends , and least successful among family members then he has a 83.3% probability of being addicted .

If we calculate the frequency of risk factors we can find the key risk factors of substance abuse. They are:

1. **Unsatisfied in life**

2. **Bad relations with parents**

3. **Bitter experience from past relations**

4. **Access to drugs**

5. **Joins party where drugs are consumed**
6. **Unsuccessful among family and friends**
7. **Believes drugs relieve from stress**
8. **More than 2 addicted friends**

4.6 Conclusion

In this section, the evaluation method and results are clearly explained. The efficacy and reasoning behind the selection of the effective classifier and the key risk factors are described as well. We eliminated the redundant features using backward elimination and Pearson's chi-squared test combined. Then we selected the classifier with most efficiency based on accuracy, weighted f1-measure and ROC/AUC values. Finally, the key factors were identified by factor decomposition and association rule mining.

Chapter 5

Conclusion

5.1 Conclusion

The basis of this study was the abundant information available on substance abuse in previous studies and projects. Necessary information culled from the related researches paved the way to figure out common causes of substance abuse. As aspects of substance abuse change from place to place, fact-checking of causes by local professionals operating on this issue was essential. Subsequently, to gain data from participants on these causes, the questionnaire was carefully designed safeguarding individual privacy and finalized after pre-testing. Responses from participants were converted into data-set after some pre-processing. Using the Pearson chi-squared statistic, χ^2 between feature variables and target variable, feature importance was calculated. Using the backward elimination method, performance of the classifiers was analyzed in each case and compared to select the efficient classifier and eliminate redundant features. Logistic regression classifier trained with 18 features, selected by having the best chi-squared statistic value was found the be the yielding the best performance. Decomposition of these 18 features and performing factor analysis finds out the key factors of substance abuse.

Using only the 18 necessary features, out of the initial 35 helped acquire better accuracy by excising dubious and ill-performing features. Participants were respondents aged between 20-35yrs. (largely youths) which provided a better opportunity to understand behavioral traits as similar age groups usually display similar patterns of behavior. Local impacts that occur uniquely to Bangladesh were taken into consideration to measure risk factors in a feasible way to improve the effectiveness of the model. A great feature of Pearson's chi-squared test of independence is it provides insights about two categorical features using the frequency of individual categories which is useful for binary classification. Although the performance of the classifier is slightly down in case of the class *addicted*, this shortcoming can be overcome by collecting more data from patients suffering from substance abuse disorder. As far as the overall performance is concerned, a good accuracy score was achieved because we prepared the questionnaire according to the information we needed and pretested it before finalizing leaving no scope for unnecessary questions to stay in the questionnaire. Data-set was created from the information gathered from the questionnaire resulting in an effective data-set for our binary classification.

This study was performed among the urban young population of Bangladesh which does not represent the whole scenario of substance abuse in the country. Causes might vary for changes in the target population in both home and abroad. For example, in the rural youth population, the number of uneducated people is higher with less vigilant surveillance from law enforcement agencies compared to metropolitan cities, which might result in a significant change on the list of causes. Internationally, the scope and type of substance abuse vary due to different reasons which are to be considered to implement the model overseas. Data collected from the people suffering from substance abuse disorder was sufficient for our study but not a huge amount due to various limitations. An increase in the size of the data-set will reveal some more valuable insights which will help to predict the vulnerability with a higher degree of precision. Due to the COVID-19 outbreak in midst of the study, collecting data from rehabilitation centers became increasingly difficult. A larger data-set can be collected in the future which will produce an even better accuracy in the *addicted* class to create a classifier with even better performance and also incorporate an even larger population.

The purpose of this study is to investigate an individual's risk of substance abuse to prevent substance abuse disorder at an early age. The trend of substance abuse in almost every country is that it follows a hot spot or a region with more abuse than other places [20]. The population of these regions is at a greater risk of falling victim, using this model vulnerable individuals can be identified to be enrolled in counseling programs. This can be a way forward to reduce the effect of this deleterious behavior on these affected communities. In the case of an individual, it has already been shown, how this study can identify risk to prevent them from falling into a hazardous situation. It has already been addressed again and again, how serious and damaging this problem is to Bangladesh and all over the world. So many important factors within a society are directly influencing by substance abuse disorder, one of them is the development of the youth population. Sadly, the number of people suffering from substance abuse disorder is increasing globally which is a portent for larger problems ahead. This study focuses on prevention by creating a model to assess vulnerability which can be a little step forward towards addressing this global problem. Finally, from this study, it can be said that data science has an important role to play in understanding these intricate behavioral traits of human beings to address important psychological issues. Health officials can use the huge potential of data to build a safer and better future.

5.2 Future Work

This study leaves a wide scope for future works to be done in this area. Data in this project was only collected from urban youths of Bangladesh, with better resources and more time, the focused range of the population in this study can be increased which will enable this model to predict within a larger community. A similar process can be adopted in different countries where risk factors may vary a little. An automated counseling chat-bot online can be developed with the help of psychiatrists to suggest participants further actions according to the individual vulnerability predicted from the classifier. Different people can suffer from substance abuse due to different reasons like depression, family problems, economic crisis, etc. Once the specific reasons can be identified it can be treated with better precision.

Chapter 6

Publications

Publications

Part of this thesis work was presented in 20th International Conference on Hybrid Intelligent Systems, 2020 available as:

1. Predicting Individual Substance Abuse Vulnerability Using Machine Learning Techniques [35]

References

[1] L. Degenhardt and W. Hall, 'Extent of illicit drug use and dependence, and their contribution to the global burden of disease,' *The Lancet*, vol. 379, no. 9810, pp. 55–70, 2012 (cit. on p. 1).

[2] *Mental health and substance abuse*. [Online]. Available: `https://www.who.int/westernpacific/about/how-we-work/programmes/mental-health-and-substance-abuse` (cit. on p. 1).

[3] L. R. Gowing, R. L. Ali, S. Allsop, J. Marsden, E. E. Turf, R. West and J. Witton, 'Global statistics on addictive behaviours: 2014 status report,' *Addiction*, vol. 110, no. 6, pp. 904–919, 2015 (cit. on p. 1).

[4] L. Scholl, P. Seth, M. Kariisa, N. Wilson and G. Baldwin, 'Drug and opioid-involved overdose deaths—united states, 2013–2017,' *Morbidity and Mortality Weekly Report*, vol. 67, no. 51-52, p. 1419, 2019 (cit. on p. 1).

[5] A. E. Kelley and K. C. Berridge, 'The neuroscience of natural rewards: Relevance to addictive drugs,' *Journal of neuroscience*, vol. 22, no. 9, pp. 3306–3311, 2002 (cit. on p. 1).

[6] T. H. Brandon, J. I. Vidrine and E. B. Litvin, 'Relapse and relapse prevention,' *Annu. Rev. Clin. Psychol.*, vol. 3, pp. 257–284, 2007 (cit. on p. 1).

[7] J. D. Hawkins, R. F. Catalano and J. Y. Miller, 'Risk and protective factors for alcohol and other drug problems in adolescence and early adulthood: Implications for substance abuse prevention.,' *Psychological bulletin*, vol. 112, no. 1, p. 64, 1992 (cit. on p. 2).

[8] E. R. Oetting and F. Beauvais, 'Peer cluster theory, socialization characteristics, and adolescent drug use: A path analysis.,' *Journal of counseling psychology*, vol. 34, no. 2, p. 205, 1987 (cit. on p. 2).

[9] M. N. Sani, 'Drug addiction among undergraduate students of private universities in bangladesh,' *Procedia-social and behavioral sciences*, vol. 5, pp. 498–501, 2010 (cit. on pp. 2, 7, 8).

[10] K. R. Conner, M. Pinquart and S. A. Gamble, 'Meta-analysis of depression and substance use among individuals with alcohol use disorders,' *Journal of substance abuse treatment*, vol. 37, no. 2, pp. 127–137, 2009 (cit. on p. 2).

[11] A. E. Barrett and R. J. Turner, 'Family structure and substance use problems in adolescence and early adulthood: Examining explanations for the relationship,' *Addiction*, vol. 101, no. 1, pp. 109–120, 2006 (cit. on p. 2).

[12] S. J. Bahr, S. L. Maughan, A. C. Marcos and B. Li, 'Family, religiosity, and the risk of adolescent drug use,' *Journal of Marriage and the Family*, pp. 979–992, 1998 (cit. on p. 2).

[13] R. Sinha, 'Chronic stress, drug use, and vulnerability to addiction,' *Annals of the new York Academy of Sciences*, vol. 1141, p. 105, 2008 (cit. on p. 2).

[14] G. E. Nagelhout, K. Hummel, M. C. de Goeij, H. de Vries, E. Kaner and P. Lemmens, 'How economic recessions and unemployment affect illegal drug use: A systematic realist literature review,' *International Journal of Drug Policy*, vol. 44, pp. 69–83, 2017 (cit. on p. 2).

[15] J. P. Pierce, J. M. Distefan, R. M. Kaplan and E. A. Gilpin, 'The role of curiosity in smoking initiation,' *Addictive behaviors*, vol. 30, no. 4, pp. 685–696, 2005 (cit. on p. 2).

[16] A. M. Belcher, N. D. Volkow, F. G. Moeller and S. Ferré, 'Personality traits and vulnerability or resilience to substance use disorders,' *Trends in cognitive sciences*, vol. 18, no. 4, pp. 211–217, 2014 (cit. on p. 2).

[17] N. Nair, N. Newton, E. Barrett, T. Slade, P. Conrod, A. Baillie and M. Teesson, 'Personality and early adolescent alcohol use: Assessing the four factor model of vulnerability,' *Journal of Addiction & Prevention*, vol. 4, no. 2, pp. 1–6, 2016 (cit. on p. 2).

[18] A. Shahriar, F. Faisal, S. U. Mahmud, A. Chakrabarti and M. G. Rabiul Alam, 'A machine learning approach to predict vulnerability to drug addiction,' in *2019 22nd International Conference on Computer and Information Technology (ICCIT)*, 2019, pp. 1–7. DOI: 10.1109/ICCIT48885.2019.9038605 (cit. on p. 3).

[19] K. K. Mak, K. Lee and C. Park, 'Applications of machine learning in addiction studies: A systematic review,' *Psychiatry research*, vol. 275, pp. 53–60, 2019 (cit. on p. 4).

[20] A. B. Heydarabadi, A. Ramezankhani, H. Barekati, M. Vejdani, K. Shariatinejad, R. Panahi, S. H. Kashfi and M. Imanzad, 'Prevalence of substance abuse among dormitory students of shahid beheshti university of medical sciences, tehran, iran,' *International journal of high risk behaviors & addiction*, vol. 4, no. 2, 2015 (cit. on pp. 8, 47).

[21] P. Nath, S. Kilam and A. Swetapadma, 'A machine learning approach to predict volatile substance abuse for drug risk analysis,' in *2017 Third International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)*, IEEE, 2017, pp. 255–258 (cit. on pp. 8, 9).

[22] J. S. Wiggins, *The five-factor model of personality: Theoretical perspectives.* Guilford Press, 1996 (cit. on p. 9).

[23] W.-Y. Ahn, D. Ramesh, F. G. Moeller and J. Vassileva, 'Utility of machine-learning approaches to identify behavioral markers for substance use disorders: Impulsivity dimensions as predictors of current cocaine dependence,' *Frontiers in Psychiatry*, vol. 7, p. 34, 2016 (cit. on pp. 10, 11).

[24] J. H. Patton, M. S. Stanford and E. S. Barratt, 'Factor structure of the barratt impulsiveness scale,' *Journal of clinical psychology*, vol. 51, no. 6, pp. 768–774, 1995 (cit. on p. 10).

[25] M. H. Afzali, M. Sunderland, S. Stewart, B. Masse, J. Seguin, N. Newton, M. Teesson and P. Conrod, 'Machine-learning prediction of adolescent alcohol use: A cross-study, cross-cultural validation,' *Addiction*, vol. 114, no. 4, pp. 662–671, 2019 (cit. on p. 10).

[26] R. Sinha, 'New findings on biological factors predicting addiction relapse vulnerability,' *Current psychiatry reports*, vol. 13, no. 5, p. 398, 2011 (cit. on p. 11).

[27] W.-Y. Ahn and J. Vassileva, 'Machine-learning identifies substance-specific behavioral markers for opiate and stimulant dependence,' *Drug and alcohol dependence*, vol. 161, pp. 247–257, 2016 (cit. on p. 12).

[28] H. Garavan, K. Brennan, R. Hester and R. Whelan, 'The neurobiology of successful abstinence,' *Current opinion in neurobiology*, vol. 23, no. 4, pp. 668–674, 2013 (cit. on p. 13).

[29] R. J. Romeu, N. Haines, W.-Y. Ahn, J. R. Busemeyer and J. Vassileva, 'A computational model of the cambridge gambling task with applications to substance use disorders,' *Drug and Alcohol Dependence*, vol. 206, p. 107 711, 2020, ISSN: 0376-8716. DOI: `https://doi.org/10.1016/j.drugalcdep.2019.107711`. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/S0376871619304880` (cit. on p. 14).

[30] M. M. Aldarwish and H. F. Ahmad, 'Predicting depression levels using social media posts,' in *2017 IEEE 13th international Symposium on Autonomous decentralized system (ISADS)*, IEEE, 2017, pp. 277–280 (cit. on p. 15).

[31] S. Hassanpour, N. Tomita, T. DeLise, B. Crosier and L. A. Marsch, 'Identifying substance use risk based on deep neural networks and instagram social media data,' *Neuropsychopharmacology*, vol. 44, no. 3, pp. 487–494, 2019 (cit. on p. 15).

[32] L. A. Fisher, J. W. Elias and K. Ritz, 'Predicting relapse to substance abuse as a function of personality dimensions,' *Alcoholism: Clinical and Experimental Research*, vol. 22, no. 5, pp. 1041–1047, 1998 (cit. on p. 16).

[33] W.-Y. Ahn and J. R. Busemeyer, 'Challenges and promises for translating computational tools into clinical practice,' *Current Opinion in Behavioral Sciences*, vol. 11, pp. 1–7, 2016 (cit. on p. 17).

[34] M. L. McHugh, 'The chi-square test of independence,' *Biochemia medica: Biochemia medica*, vol. 23, no. 2, pp. 143–149, 2013 (cit. on p. 22).

[35] U. I. Islam, I. H. Sarker, E. Haque and M. M. Hoque, ‘Predicting individual substance abuse vulnerability using machine learning techniques,’ in *Hybrid Intelligent Systems*, A. Abraham, T. Hanne, O. Castillo, N. Gandhi, T. Nogueira Rios and T.-P. Hong, Eds., Cham: Springer International Publishing, 2021, pp. 412–421, ISBN: 978-3-030-73050-5 (cit. on p. 49).