DRUG ADDICTION PREDICTION USING MACHINE LEARNING

BY

MD. ARIFUL ISLAM ARIF
ID: 162-15-7871

SAIFUL ISLAM SANY
ID: 162-15-7868
AND

FAIZA ISLAM NAHIN
ID: 162-15-7722

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering.

Supervised By

MD. TAREK HABIB
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised By

FARAH SHARMIN
Senior Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY
DHAKA, BANGLADESH
JULY 202

APPROVAL

This research project titled "**Drug Addiction Prediction Using Machine Learning**", submitted by Md. Ariful Islam Arif, ID: 162-15-7871, Saiful Islam Sany, ID: 162-15-7868 and Faiza Islam Nahin, ID: 162-15-7722 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 08 July 2020.

BOARD OF EXAMINERS

Dr. Syed Akhter Hossain **Chairman**
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Dr. Md. Ismail Jabiullah **Internal Examiner**
Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Azmoon

Nazmun Nessa Moon **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Dr. Mohammad Shorif Uddin **External Examiner**
Professor
Department of Computer Science and Engineering
Jahangirnagar University

DECLARATION

We hereby declare that this project has been done by us under the supervision of **Md. Tarek Habib, Assistant Professor, Department of CSE,** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

SUPERVISED BY:

Habib

Md. Tarek Habib
Assistant Professor,
Department of CSE,
Daffodil International University

CO-SUPERVISED BY:

Farah Sharmin

Farah Sharmin
Senior Lecturer,
Department of CSE
Daffodil International University

Submitted by:

Arif

Md. Ariful Islam Arif
ID: 162-15-7871
Department of CSE
Daffodil International University

Saiful Islam Sany

Saiful Islam Sany

ID: 162-15-7868
Department of CSE
Daffodil International University

Faiza Islam Nahin

Faiza Islam Nahin
ID: 162-15-7722
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to Almighty **Allah** for His divine blessing makes us possible to complete the final year project/internship successfully.

We grateful and wish our profound indebtedness to **Md. Tarek Habib**, Assistant Professor, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "*Machine Learning*" to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stages have made it possible to complete this project.

We would like to express our heartiest gratitude to **Dr. Syed Akhter Hossain**, Head**,** Department of CSE, for his kind help to finish our project and **Dr. Md. Ismail Jabiullah**, Professor and **Nazmun Nessa Moon**, Assistant Professor and also to other faculty members and the staff of the CSE department of Daffodil International University.

We would like to thank our entire course mate at Daffodil International University, who took part in this discussion while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

Drugs and alcohol are dangerous to health and the body. Nowadays drug addiction has become a threat to Bangladeshi young people. Drugs and alcohol have a negative impact on our life. We have to keep an eye on the young people of our country not getting addicted to drugs quickly. We need to stay away from the drug before getting addicted to it. We will predict the risk of becoming addicted to drugs with machine learning. First, we study some related papers, journals, and online articles then we talk to doctors and drug addicts people; we find some common factors that related to becoming addicted to drugs. Then we collect data based on those factors, such as age, gender, profession, health ability, mental pressure, trauma, family and friend's history, incidents, etc. We collect data from both addicted and non-addicted people. We have two outcomes. One is 'Yes' means addicted and another is 'No' means not addicted. After data collection, we processed all the data and created a processed dataset. We applied machine-learning algorithms on our processed dataset. Since machine learning, artificial intelligence and deep learning used in various predictions and detection systems. We use k-nearest neighbor (*k*NN), logistic regression, support vector machine (SVM), naïve Bayes, random forest, adaptive boosting (ADA boosting), decision tree, multilayer perceptron (MLP) and gradient boosting classifier. In our work, out of nine algorithms, logistics regression gave the best performance based on accuracy and the accuracy of logistic regression was 97.91%.

TABLE OF CONTENTS

CONTENTS **PAGE**

CHAPTER **PAGE**

LIST OF TABLES

LIST OF FIGRES

CHAPTER 1

INTRODUCTION

1.1 Introduction

Drug addiction means the taking of various drugs illegally and being addicted to those drugs for their toxic and addictive effects. Drug addiction is one of the most malignant problems for a country. It can destroy a life and a nation easily. The toxic effects of drugs attacked and ragged a person mentally and physically. Drug addiction has taken a firm hold of our young generation of our country. A developing country like Bangladesh, addiction can bear a terrible effect on our society. According to the report of the daily star newspaper, narcotics drug production and miscellaneous use by the terrorist group have increased in Bangladesh [1]. Near about 25 lakh, people are drug-addicted. In Bangladesh, about 80 percent of the drug addicts are adolescents and young men of 15 to 40 years of age [2]. This social cancer has spread its poisonous claw all over the world. Frustration is the cause of this addiction. Unemployment problems, political cataclysm, lack of family ties, lack of love affection etc. give rise to frustration. Again, this addiction gives rise to social crimes. When addicted cannot afford to buy drugs, they commit many kinds of social crimes like hijacking, looting, plundering killing, robbery, etc. So in order to avoid drug addiction, we need to stay away from drugs. Stay away from drugs will only reduce the risk of getting addicted before you become addicted to it. We use machine learning for this prediction to becoming addicted to drugs.

1.2 Motivation

Drugs bring a heavy loss of health. They make the addicts' appetite less, abnormal, impatient, intolerant, unsocial, inhuman, cruel, heartless, and bankrupt. The addicts do not take care of their families but keep ties with other family members, respect their parents, superiors, and teachers. Nowadays drug addiction becomes a dangerous fact. Drug addiction affects the young generation from all lifestyles. In 2015, drug-addicted Oishee Rahman kill her parents [3]. It also difficult to moving alone for a woman in the city, because there are many drug-addicted people surrounded the city. When we go to a new place at that time we cannot find out those are addicted. An addicted friend can destroy the friend circle easily. According to the news of the Dhaka Tribune newspaper, there are around 7.5 million people addicted to drugs in Bangladesh. The dangerous thing is among them 80% are the youth and 50% of them are involved in different criminal activities [4]. We need to keep a special focus so that our youth do not become addicted to drugs. Therefore, we will try to anticipate in advance if we have a tendency to become addicted. We have not seen much research in this field. We are going to do this prediction using the machine learning technique.

1.3 Rationale of the Study

As we mentioned earlier, there is less significant work has done previously with drug addiction prediction in Bangladeshi perspective. That is why we are interested to work with drug addiction and machine learning techniques.

Machine learning is a branch of artificial intelligence that employs a variety of statistical, probabilistic, and optimization techniques that allows computers to “learn” from past examples and to detect hard-to-discern patterns from large, noisy or complex data sets. Machine learning methods used in a wide range of applications ranging from detecting and classifying. Machine learning is used for cancer prediction [5], a systemic review of software fault prediction [6], dermatological disease detection [7] and so on. Many types of detection and risk prediction are now conducting by using machine learning. Machine learning techniques may have a supplementary role in highly complex problems and provide a comparison to regression results. [8]. As machine learning has a vast field of work, we thought that we should apply machine learning for our work of prediction.

1.4 Research Questions

- How do we identify drug addicts?
- Why does machine learning seize out “interest” for carrying out the research?
- Is there any other way to identify drug addicts?
- What will our original data be like?
- Do we need to train our original data to the machine learning model?
- What amount of data do we collect and where do we collect it?
- Does our data and machine learning will be compatible?
- Should we use popular machine learning techniques or use a new machine learning technique?

1.5 Expected Outcome

We hope our research will help people to predict. By using this technique, people can easily and quickly know the risk of becoming addicted to drugs. People can also know more about prediction with machine learning. Successful deployment of existing or new machine learning algorithms for predicting drug addiction. An addicted person is harmful to our society we should keep away from him.

Sometimes we go to different places and different environments in need of work. We do not know if anyone is addicted there or not, if anyone is there, to what extent do I have a tendency to become addicted. This research will help us to predict the risk of becoming addicted when we enter a new environment. Parents can take care of their children. With our children, who can make friends with, who they are moving to, their daily lives, using this information, will let parents know if their child is prone to drug addiction. It will protect us and our society from the adverse effects of drug addiction. Besides, the department of drug control and law enforcement agency can make a collaboration with our model which will help them to identify addicted or not addicted people. Besides, the building of a large data set for drug addiction in the context of Bangladesh. Publication of one or more articles in international conference proceedings or journals.

1.6 Report Layout

This research paper contains the following contents as given below:

- Chapter **one** explains the introduction of the research with its motivation, rationale of the study, research questions, and expected outcome.
- Chapter **two** discusses related works, research summary, the scope of the problem, and challenges.
- Chapter **three** contains the workflow of this research, data collection procedure, and statistical analysis and feature implementation.
- Chapter **four** covers experimental evaluation and some relevant discussions, the outcome of research via numerically and graphically.
- Chapter **five** covers this research impact on society.
- Chapter **six** contains a summary of this research work along with the limitation and future work.

CHAPTER 2

BACKGROUND STUDY

2.1 Introduction

In this section, we will discuss related works, research summary, the scope of the problem, and challenges. In the part of the related work, we summarize some research papers, related works, underlying methods, classifiers and accuracies of which related to our work. In the research summary part, we prepare a summary of some related works and display them in a table for better and easy understanding. The scope of the problem part discusses how we can contribute to the problem with our work model. Finally, the Challenges part contains some words about the obstacles and dangers we encountered during the course of this research work.

2.2 Related Works

This literature review section of this research paper is going to present the near past related works done by some researchers on drugs and addiction prediction. We have followed and studied their work to understand the processes and methods expressed by them.

Dhiraj Dahiwade et al. [9] has proposed a general disease prediction system, which based on machine learning algorithms. They proposed disease prediction based on the syndromes of the patient. They used disease evidence in their dataset. They collected the living habits of a person and checkup information as to their data for the prediction system. They remove comma, punctuations and blank space in data preprocessing and used the dataset as a training dataset. The dataset of patient disease downloaded from UCI machine-learning website. They used *k*NN and CNN algorithms. In *k*NN, Euclidean distance, Hamming distance used as common distance metrics. In CNN, the dataset converted into vector from for implementation. Max pooling operation performed at the convolutional stage on CNN.

*k*NN took more time than CNN. They compare two algorithms based on the accuracy and time and found 84.5% accuracy in CNN, which was greater than *k*NN. They use Java for case implementation, MySQL in the backend.

Osman Hegazy et al. [10] has proposed a model for stock market prediction with machine learning technology. Their model can determine the future value of a company's stock on a financial exchange. In their proposed model, they integrated Particle Swarm Optimization (PSO) with LS-SVM. Their aim was to develop a machine learning hybrid model of PSO and LS-SVM. LS-SVM performs with three parameters cost penalty, insensitive loss function, kernel parameter. They used PSO for finding out the best parameter combination of these three parameters. Historical data and technical indicators were used in their model to predict daily stock prices. There proposed model contains six input vectors and one output that represents the future price. The proposed model tested on about 500 stock markets from 2009 to 2012. Information Technology, financial, health care, energy, communication, materials, and industrial sectors elected for testing. Relative strength index, money flow index, exponential moving average, stochastic oscillator, moving average convergence, these five technical indicators were calculated from the raw data set. They divided there are data set on 70% for training purposes and 30% for testing purposes. There proposed LS-SVM-PSO model compare with LS-SVM and ANN-BP algorithms. The compare based on error value, accuracy and mean square error. Where LS-SVM-PSO got the lowest error with the best accuracy and LS-SVM, ANN-BP got the highest error with the worst accuracy. Their proposed model performed better than other algorithms especially in the case with fluctuations in the time series function. Besides, their model was capable to solve the overfitting problem that did not solve in ANN. Therefore, their research proves that LS-SVM-PSO was the best algorithm among LS-SVM and ANN-BP for predicting stock market prices.

Lea Monica B. Alonzo et al. [11] has presented a detailed comparison between various machine learning algorithms that used to predict and assessment of coconut sugar quality. Their study had evaluated the accuracy and the average running time for each model. Color, order, test, purity were the physical characteristics and water activity, glucose, fructose, sucrose, and ash were the chemical properties of coconut. They used 350 images of coconut sugar collected from two coconut sugar production agencies; they are the Philippine Coconut Authority (PCA) and the United Coconut Association of the Philippines (UCAP). Collected images extracted into RGB values. RGB values used as input and classification of images as superior, good and reject used as output. The samples had 48 superior qualities, 110 good qualities and 192 rejected. They used python, scikit-learn and many classifiers for their prediction model. They divided total data set into 10 partitions among them nine partitions were used as input and rest one partition was used as output. They were used MLP artificial neural network (ANN), stochastic gradient descent (SGD), k-nearest neighbors (*k*NN), support vector machine (SVM), decision tree (DT) and random forest (RF) algorithm for coconut sugar quality prediction. They elected the best algorithm according to accuracy and average running time. SGD achieved the highest accuracy with 98.38% and *k*NN achieved the lowest running time with 19.35 sec. However, according to accuracy versus running time graph is SGD had the best performance for predicting coconut sugar quality. The aim of their study was to improve the quality assessment system that complaint with Philippine National standards. They would use border repository in the SGD algorithm for effectiveness in their further study and try to integrate computer vision with their approach for real-time assessment.

Amir Hamzeh Haghiabi et al. [12] has worked on predicting water quality in the machine learning approach. They applied artificial intelligence technology to predict the water quality of the Tireh River located in the southwest of Iran. In their research, they used temperature, pH, specific conductivity, bicarbonate, sulfates, chlorides, total dissolved solids, sodium, magnesium, calcium as their components. Regional water authority (RWA) measured the quality of the components in water. With the calculated value, they prepared their dataset. They divided the dataset with 80% train data and 20% test data. They applied artificial neural networks (ANN), support vector machine (SVM), and group method of data handling (GMDH) algorithms on their dataset. They did their selection process based on accuracy and DDR index. SVM had the best accuracy with the lowest DDR value. According to their research, water quality can predict using the SVM algorithm.

Yupu Zhang et al. [13] has proposed a method for predicting daily smoking behavior based on the machine learning algorithm. They did research on smoking behavior of every day when smokers smoked. They proposed a feature extraction module used on the characteristics of smokers. Smoker's information collected from the Chinese center for disease control and prevention. They collected data from the '2015 China adult tobacco survey report' which published by the Chinese center for disease control and prevention. The report made a survey on 15,095 people around the country. They also collected the data of daily smoking amount, smoking rate, age, gender, education level. In their classification model, they used nine factors as input and one factor as output. They used the decision tree algorithm because it can process a large number of data with short time training. A decision tree can integrate GBDT, random forest and XGBoost classification. In their prediction technique, they used the XGBoost decision tree algorithm. They evaluate with true positive, false positive, true negative and false negative. They calculated accuracy, precision, recall and F1-measure parameters in their model. They found the best accuracy of 84.11% with maximum depth is 5. The XGBoost model got 7.1% higher accuracy with a feature extraction module. In the future, they would use more data that are real and as their model related to time series, they would like to integrate natural language recognition to improve the ability of their model.

Ahmed M. Alaa et al. [14] has proposed a machine learning-based model for predicting disease risk of cardiovascular on Biobank participants. Their created ML-based model can predict CVD risk based on 473 variables. In their model, they used AutoPrognosis algorithmic tool. AutoPrognosis tool automatically selects and tunes ensembles of ML modeling pipelines. They compared their model with well-established risk prediction models like Framingham score, Cox proportional hazard. They used for 473 variables in their model. They considered patients usual walking, health rating, diabetes, and breathing rate for the analysis of the CVD disease risk. They collected data from 22 assessment centers across England, Wales and Scotland from 2006 to 2010. All their data also kept a store in UK Biobank. They used seven core risk factors; there are age, gender, systolic blood pressure, smoking status, hypertension, diabetes and BMI. They had 423604 participators from the UK. In their train model, AutoPrognosis conducted 200 iterations of the Bayesian Optimization Procedure. Generally, AutoPrognosis contains 5460 possible ML pipelines but in their model, they had used seven imputation algorithms, nine feature algorithms, twenty classification algorithms and three calibration methods. They used a random forest algorithm for post-hoc variable ranking. Their risk prediction AutoPrognosis model performance evaluated based on the area under the receiver operating characteristics curve (AUC-ROC). Their proposed model had AUC-ROC: 0.774 where the Framingham model had 0.724 and the Cox PH model had 0.734. Using their AutoPrognosis model for CVD risk prediction in UK Biobank increases accuracy. The limitation of their model was the absence of cholesterol biomarkers in a data repository.

Hongyan Zhu et al. [15] has worked on presymptomatic detection of tobacco disease with hyperspectral image and machine-learning classifiers. They worked on presymptomatic detection of tobacco using hyperspectral imaging which combined with variable selection method and machine learning classifier. They collected images of healthy and TMV-infected leaves with 2, 4, and 6 days post-infection. A push broom hyperspectral reflectance imaging system which covering the Spectral range of 380-1023 NM used in their model for preprocessing. Contrast, correlation, entropy and homogeneity used as a textural feature, which captured from hyperspectral images at selected EWs according to the gray level co-occurrence matrix (GLCM). Selected EWs had information about the detection and classification of TMV-infected tobacco leaves. They had 32 texture variables (4 texture features x 8 wavelengths) in their dataset. Tobacco plants (*Nicotiana tabacum*) collected from Zhejiang University. Total 180 tobacco plants collected in seven days period. Among 180 samples 120 samples used as calibration sets and rest 60 samples used as prediction sets. ENVI software used for image processing techniques in segmentation. By using a successive projection algorithm ineffective wavelength measurement, they got 459.58 nm as a result. Selected EWs used as input of machine learning algorithms. Partial least squares-discrimination analysis (PLS-DA), random forest (RF), support vector machine (SVM), backpropagation neural network (BPNN), extreme learning machine (ELM), least-square support vector machine (LS-SVM) algorithms applied on texture features, EWs and data fusion. ELM had the best accuracy with 98.3% on EWs, BPNN had the best accuracy with 93% on texture feature and SVM had the best accuracy with 96.7% with data fusion. Overall, ELM and BPNN could use on the successful detection of healthy and diseased tobacco leaves. They thought that healthy, 2DPI, 4DPI and 6 DPI leaves had a large scope to develop accurate and robust machine-learning models.

Xinyu Zhang et al. [16] has worked to predict HIV prognosis and mortality with smoking-associated DNA and machine learning classifier. They collected data of 1137 HIV positive people and then identified epigenome-wide significant CpGc for smoking. 698 CpGs selected for testing samples. They collected a DNA sample from WBCS of HIV patients from the Veterans Aging Cohort Study. Collected samples divided into the training set and testing set on an 8:2 ratio. They had 1137 samples; among them, high and low VACS index samples were not equal. The number of high VACS indexes was 237 and the low index was 900. For reducing potential bias they used four machine-learning algorithms, they were lasso and elastic-net regularized generalized linear model (GLMNET), support vector method (SVM), random forest (RF), and XGBoost. They used all these methods separately with each CpG group. They measured the sample set by the area under the curve (AUC) in receiver operating characteristic analysis. The efficiency of their prediction model was selected by receiver operator characteristic (ROC) curve analysis. Sensitivity, specificity, and AUC also considered in their prediction model's performance. Their selected 698 features got better performance with 0.78 AUC. To identify the best model for prediction they used tenfold cross-validation on the training set and they used GLMNET in their prediction model. Their model had a robust prediction of HIV prognosis and mortality by using DNA methylation-based machine learning.

Miguel Angel Fernandez-Granero et al. [17] has proposed a model for predicting exacerbations of obstructive pulmonary disease with machine learning features. They had proposed a new data-driven methodology for developing a prediction model. Their model had learned from past experiences and made a new pattern with clinical data. They tried to predict COPD with the patient's daily symptoms report with the help of a pattern recognition mechanism. They had prepared their dataset with the score of symptoms. They collected data and samples of 16 patients at home for six months each day. Pneumology and Allergy Department of the University Hospital Puerta del Mar of Cadiz (Spain) was monitoring the whole process. Patients were aged above 60 years and had cumulative tobacco consumption. They got a total of 789 records from the patients and applied these in their model. They used three classifiers; they were radial basis function neural network (RBF), k-means classifier (K-means) and probabilistic neural network (PNN). For graphical analysis, signal processing they used MATLAB. The classifiers had evaluated based on accuracy, specificity, sensitivity, confusion matrix, positive predictive value, and negative predictive value. PNN got the best accuracy with 89.3% accuracy, 84.1% sensitivity, and 92.5% specificity. In the future, they wanted to improve the consistency of the result with their proposed data-driven method.

Charles Frank et al. [18] has worked on smoking status prediction with machine learning and statistical analysis. They found the durability and effectiveness of machine learning algorithms on predicting smoking status based on patients' blood tests and vital readings. Firstly they established a static difference between smokers and non-smokers based on their blood test. After that, they had applied five machine-learning algorithms on their dataset to predict smoking status. They had collected patient blood test readings and vital readings as the input of their model. They collected all the information from a community hospital in the Greater Pittsburgh Area. They performed three blood tests on patients to find out the smokers and non-smokers; they were INR, HB, and HCT. They had a total of 534 samples of patients among them 311 were non-smokers and 87 were smokers. They divided the dataset into two parts, 66% data for training set and 34% data for the testing set. They used Weka for data preprocessing. They used the 'ReplaceMissingValue' filter for replacing missing values and the 'SMOTE' filter applied in the dataset for preprocessing. Then they applied five algorithms on the processed data. They used Naïve Bayes, MLP, Logistic regression classifier, J48, and Decision table algorithms to predict the smoking status of patients. Their evaluating process based on accuracy, precision, recall, and F-measure. Logistic regression had the best performance with 83.44% accuracy, 83% precision, 83.4% recall and 83.2% F-measure. They wanted to work on improving the outcome and accuracy in the future and also increase the dataset size and apply other methods for missing values that could replace the missing values with the average value.

Mary R. Lee et al. [19] has worked with a model that predicts alcohol use disorder by checking the treatment-seeking status with a machine learning classifier. They applied decision tree classifier to differentiate between treatment and non-treatment seeking group. They applied 'if-then' logic in the tree to find out the important logic and features. Treatment seeking status was count with yes or no. They prepared two separate decision trees for analysis of possible bias with gender. They collected data from the National Institute on Alcoholism and Alcohol Abuse (NIAAA) Intramural Research Program at the National Institute of Health Clinical Centre in Bethesda, MD with maintaining some protocols from 2008 to 2018. Their collected data domains were cognitive, mood, impulsivity, personality, aggression, and early life stress and childhood trauma. The data they collected from the clinical lab were alanine aminotransferase, aspartate aminotransferase, gamma-glutamyl transferase, hepatitis B antigen, and C antibody. They had a total of 778 numbers of data for their model. They used in Weka one attribute for treatment-seeking status and 178 attributes used for training the ADT. From there, they got 10 clinical variables; they were quantity and quality of drinking, depression, psychological problem, IQ, race, BMI, alcohol consumption, and a number of drinks in the last 30 days. They used 90% data for training and 10% data for testing purposes. They applied four classifier algorithms in ADT (alternating decision tree); they were random forest, random tree, logistic regression, and simple logistics model. They had evaluated their model based on accuracy. With a simple logistic model, ADT got the highest accuracy for Cross-validation was 86.1% and for independent validation was 78%. Their model will guide patient management and treatment process.

Sivan Kinreich et al. [20] has proposed a model on predicting the risk of Alcohol Use Disorder using machine-learning technology. They applied the SVM method as supervised machine learning to classify the symptoms of AUD people and normal people. To collect data, they recorded EEG for 4 minutes of all patients in dimly lit with a close eye and no movement. After preprocessing the data, they got features like, spectral power, coherence values, and correlation values. They prepared a prediction model that find out AUD using machine learning multidimensional features like gender, origin, and age. First, they collect healthy people, after a few years AUD affected, and unaffected people were separated from them. They collected data from 656 people in 12 to 30 years to prepare their models. Their age was divided into different age groups, like 12-15, 16-19, 20-30, etc. They also divided the origin into two parts; AA means African American and EA means European American. They used the regularization method for variable overfitting control, and they used the least absolute shrinkage and selection operator (LASSO) for feature selection. They applied a tenfold cross-validation procedure in regularization parameter detection and labeled response variable as control (not affected) and AUD (alcohol use disorder). They had considered true positive, true negative, accuracy, area under curve and F-score to evaluate their model. They found higher accuracy in AA than EA samples. Their model proves that using a wide range of multidimensional features increases the accuracy of the model. Their ML model proved that the combination of genetic data and EEG data get better accuracy than single using data. They wanted to make a valid result with large cohorts and use wide selection features in the future.

Divya Kumari et al. [21] has proposed a model of predicting alcohol abused using machine learning technology. An artificial neural network-based model was approached to predict alcohol users. The model had two ANN modules; they were ANN-D and ANN-C. They first applied ANN-D to the data they collected, and then applied ANN-C to those who were alcohol users. They considered age, gender, country, ethnicity, education, neuroticism, openness to experience, extraversion, agreeableness, conscientiousness, impulsive, sensation seeing as their models feature. These features considered in ANN-D and day, week, month, year, decade considered in ANN-C. ANN-D was used to determine whether a person was an alcohol user or not and ANN-C was used to determine when an alcoholic was drinking. They collected their data from the UCI machine learning database. They had a total of 1885 records with 12 attributes to use in their model. In ANN-D, they calculated target output 0 as not used and 1 as used. They applied a Levenberg Marquardt algorithm in both ANN-D and ANN-C. They used two hidden layers with 40 and 30 neurons in ANN-D and three hidden layers with 40, 20, and 30 hidden layers in ANN-C. Tan-sig was used as an output layer in ANN-D and pure linear was used as an output layer in ANN-C. Their model evaluated based on accuracy. ANN-D had 98.7% and ANN-C had 49.1% accuracy in their proposed model. They would try to increase the prediction accuracy of time to use in the future.

Md. Tarek Habib et al. [22] has done a study on Papaya disease recognition based on a machine learning classification technique. They used defective papayas color images. They converted all images into 300 x 300 pixels. Bicubic interpolation and histogram equalization were used for image processing. They used a total of 129 images of defective and defect-free in their model. They divided their dataset into two parts, two-third as a training dataset and one-third as a testing dataset. They have used several machine learning classification techniques. The techniques are SVMs, C4.5, Naïve Bayes, Logistic Regression, *k*NN, Random Forest, BPN, CPN, and RIPPER. They had worked with five common diseases in their work. Among these techniques, SVM has performed best. SVM has produced 95.2% accuracy among all classifiers.

2.3 Comparative Analysis and Summary

There are some work has already done about prediction and detection with the machine learning algorithm and data mining process. Nowadays, the use of machine learning technology has increased with the use of alcohol user prediction, tobacco user detection, and various disease detection. The comparison between these related works has shown in this part. Here, the comparison of different research works with their subject, methodology, and the outcome are given below in Table 2.1.

TABLE 2.1: SUMMARY OF RELATED RESEARCH WORK.

SL	Author name	Methodology	Description	Outcome
1.	Dhiraj Dahiwade,Prof. Gajanan Patle, Prof. Ektaa Meshram	*k*-nearest neighbors (*k*NN), CNN.	Machine learning-based general disease prediction system.	84.5% accuracy in CNN.
2.	Osman Hegazy , Omar S. Soliman , Mustafa Abdul Salam	Particle Swarm Optimization (PSO), LS-SVM	Stock market prediction model with machine learning.	LS-SVM-PSO got the highest accuracy and lowest error than LS-SVM and ANN-BP.
3.	Lea Monica B. Alonzo , Francheska B. Chioson, Homer S. Co, Nilo T. Bugtai, Renann G. Baldovino	MLP-ANN, stochastic gradient descent, *k*NN, SVM, decision tree, random forest.	Assessment and prediction of Coconut sugar quality with machine learning.	SGD had the best accuracy with 98.3%.
4.	Amir Hamzeh Haghiabi , Ali Heider Nasrolahi , Abbas Parsaie	ANN, SVM, and group method of data handling (GMDH).	Predicting water quality with machine learning approach.	SVM achieved the best accuracy with the lowest DDR error.
5.	Yupu Zhang , Jinhai Liu , Zhihang Zhang , Junnan Huang	XGBoost decision tree	predicting daily smoking behavior based on the machine learning algorithm	84.11% accuracy with depth 5 in XGBoost decision tree.
6.	Ahmed M. AlaaI , Thomas Bolton, Emanuele Di Angelantonio , James H. F. Rudd, Mihaela van der Schaar	AutoPrognosis	Predicting disease risk of cardiovascular on Biobank participants with machine learning.	AutoPrognosis had 0.774 AUC-ROC and increase accuracy.

7.	Hongyan Zhu, Bingquan Chu, Chu Zhang, Fei Liu, Linjun Jiang , Yong He	PLS-DA, Random forest, SVM, backpropagation neural network, extreme learning machine, LS-SVM.	Presymptomatic detection of tobacco disease with hyperspectral image and machine-learning classifiers	ELM had 98.3% accuracy.
8.	Xinyu Zhang, Ying Hu, Bradley E. Aouizerat, Gang Peng,Vincent C. Marconi, Michael J. Corley, Todd Hulgan, Kendall J. Bryant, Hongyu Zhao, John H. Krystal, Amy C. Justice, Ke Xu	GLMNET, SVM, random forest, XGBoost.	Predict HIV prognosis and mortality with smoking-associated DNA and machine learning classifier.	Area under curve had 0.78 with 698 features and GLMNET used for the best model.
9.	Miguel Angel Fernandez-Granero, Daniel Sanchez-Morillo, Miguel Angel Lopez-Gordo , Antonio Leon	Radial basis function neural network, K-means, probabilistic neural network	predicting exacerbations of obstructive pulmonary disease with machine learning features	PNN had 89.3% accuracy, 84.1% sensitivity and 92.5% specificity
10.	Charles Frank, Asmail Habach, Raed Seetan, Abdullah Wahbeh	Naïve Bayes, MLP, logistic regression, J48 and decision table.	smoking status prediction with machine learning and statistical analysis	Logistic regression had 83.44% accuracy, 83% precision, 83.4% recall.
11.	Mary R. Lee, Vignesh Sankar, Aaron Hammer, William G. Kennedy,	random forest, random tree, logistic regression and	Predicts alcohol use disorder by checking the treatment-seeking status	With the simple logistic model, ADT

	Jennifer J. Barb, Philip G. McQueen, Lorenzo Leggio	simple logistics model	with a machine learning classifier	had the highest accuracy.
12.	Sivan Kinreich ,Jacquelyn L. Meyers, Adi Maron-Katz, Chella Kamarajan, Ashwini K. Pandey, David B. Chorlian , Jian Zhang, Gayathri Pandey, Stacey Subbie-Saenz de Viteri, Dan Pitti, Andrey P. Anokhin, Lance Bauer, Victor Hesselbrock, Marc A. Schuckit, Howard J. Edenberg, Bernice Porjesz	Regularization method, LASSO.	Predicting the risk of Alcohol Use Disorder using machine-learning technology	Genetic data and EEG data had better accuracy.
13.	Divya Kumari, Sumran Kilam, Priyanka Nath, Aleena Swetapadma	ANN-D, ANN-C.	Predicting alcohol abused using machine learning technology	ANN-D had 98.7% and ANN-C had 49.1% accuracy.
14.	Md. Tarek Habib, Anup Majumder, Rabindra Nath Nandi, Farruk Ahmed, and Mohammad Shorif Uddin.	SVM, C4.5, naïve bayes, logistic regression, *k*NN, random forest, BPN, CPN and RIPPER.	Papaya disease recognition based on a machine learning classification technique	SVM got 95.2% accuracy.

Currently, a combination of machine learning, artificial intelligence and deep learning is being explored with new technologies that are used in any kind of prediction and detection model. Diagnosis and detection of material are being done recently using various machine-learning algorithms. ANN, *k*NN, CNN, SVM, logistic regression and many algorithms are popular for any detection model. From previous research, we can see that the *k*NN, SVM, random forest, ANN, naïve Bayes, and Decision tree algorithm's popularity and effectiveness for prediction or detection models are high. In our research, we have tried to implement *k*NN, SVM, logistic regression, naive Bayes, random forest and other algorithms to predict the risk of becoming addicted to drugs and alcohol in Bangladesh's perspective and we have 97.91% accuracy in logistic regression.

2.4 Scope of the Problem

The research work of ours is mainly building a model by analyzing data and applying machine-learning algorithms. Our proposed model can predict the risk of becoming addicted to drugs and alcohol. This prediction will have a significant impact on society. The young generation can stay away from drugs and alcohol. The Narcotics Control Directorate will be able to use this model for their various tasks. It is dangerous to become addicted to drugs and alcohol, and then it can be easy for anyone to become addicted to it. Therefore, this model would be useful for ordinary people and conscious people to stay away from drugs and drugs addicted. Parents worry too much about their children and worry about the future, which could hinder the development of a generation and reflect its adverse effects on society. Recently, as machine learning and artificial intelligence are being used for various object detection and disease predictions, the results are quite acceptable. Therefore, we decided that using machine learning, we would create a model of addiction risk prediction.

2.5 Challenges

While doing our research we are facing some problems. Data collection was very challenging for us. Drug addicts do not want to talk easily and do not want to admit. Besides, ordinary people and drug addicts cannot be easily distinguished. We read a lot of newspapers and talked to different people, talking to people in the neighborhood, but nobody was going to give any information about drugs or drugs addicted people. It was very difficult to collect information about drug addicts from bus stand, stations, and unknown places. After that we able to collect our data from the New Mukti Clinic, Brain and mind Hospital. We searched for some more Rehabilitation Centers but did not want to help her with any information. We needed to talk to Parsons to Parsons for the data collection but we could not do that because the hospital authorities had Privacy issues.

We were also not familiar with anaconda, jupyter notebook, and some new machine learning algorithms. It took us a while to know and learn about it at first, but with the help of our supervisor and doing more practice we can grab them easily. Then we continue to do our job very well and with enthusiasm.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

The purpose of this research is to establish a model for predicting the release of becoming drug-addicted. The Prediction Model is created based on the daily life information of people and some other related information. To create this model, we have applied various machine-learning algorithms. We used *k*NN, logistic regression, SVM, naïve Bayes, decision tree, ADA boost classifier, random forest, MLP, gradient boosting classifier in this research. Algorithms used in the model for classification purposes. We used twenty-four key factors that were very closely connected with addiction. We analyzed some of the features that were responsible for the outcome. We processed our dataset before implantation. We have calculated three types and compared them. We calculated and computed the accuracy, sensitivity, specificity, precision, recall, F_1 score, and roc-curve of each algorithm to select the appropriate algorithm for the model. We found logistic regression had the best accuracy and suitable for our proposed model.

3.2 Data Collection Process

The data set is a huge collection of necessary and relatable coordinates that can be easily accessed and changed. We first try to find out the whereabouts of drug addicts in our neighborhood and in different places. However, we saw someone around us taking drugs but it was a secret, and at the train station and bus station drug addicts refused to help. Then we decided to go to the drug addiction center and rehabilitation center. We also collect information from some private rehabilitation centers and clinics. Some of those hospitals were in their privacy issues and they refused to give us information. New Mukti Clinic and Brain & Mind Hospital helped us with the information. In addition to providing information, we can learn from their consultants and doctors about many more important factors. We were not able to collect the information as we went to the patients because the hospital authorities said it could damage patients' privacy. Therefore, we created the form and provided it to them and the hospital authorities helped us with information from their patients' and patients’ databases. We were able to collect data of 510 people based on 25 factors. There are 305-drug addicts’ information and 205 healthy people's information we have. We collected all our data from Daffodil International University, Sylhet Engineering College, Begum Rokeya University, Mukti Clinic, Brain & Mind Hospital and some other places. We collected our data based on the following factors:

- Feelings.
- Lives with family.
- Gender.
- Age.
- Living address.
- Profession.
- Distance with friends and family.
- Working efficiency.
- Stress controlling skills.
- Economic status.
- An addicted person at home.
- Faced any trauma.
- Relationship problem.
- Stay alone.
- How much you care about yourself.
- Lost job.
- Sexual harassment.
- Interest in normal activities.
- Odd sleep pattern.
- Stay outside at night.
- Loss of weight.
- Think that drug addiction can be a solution.

- Addicted friend.
- Reason to become addicted.
- Addicted or not addicted.

To identify the risk of becoming addicted to drugs we have to consider each of these factors. We find out about these factors by talking to various physicians, websites [23], [24], [25], [26], [27] and articles.

3.3 Research Subject and Instrumentation

At present, machine learning algorithms, data mining and deep learning are very popular for any prediction and detection. We will apply our collected data to various algorithms to see which algorithms will perform well for our model. We use various machine-learning algorithms; they are *k*NN, logistic regression, support vector machine (SVM), naïve Bayes, decision tree, random forest, multilayer perception (MLP), ADA boosting classifier and gradient boosting classifier. We used 'Python' as a programming language and 'Anaconda navigator', 'Jupyter notebook' as a data mining tool and 'Microsoft Excel' as our dataset in our research work.

3.3.1 Proposed Methodology

Our proposed methodology is shown below in Figure 3.1.

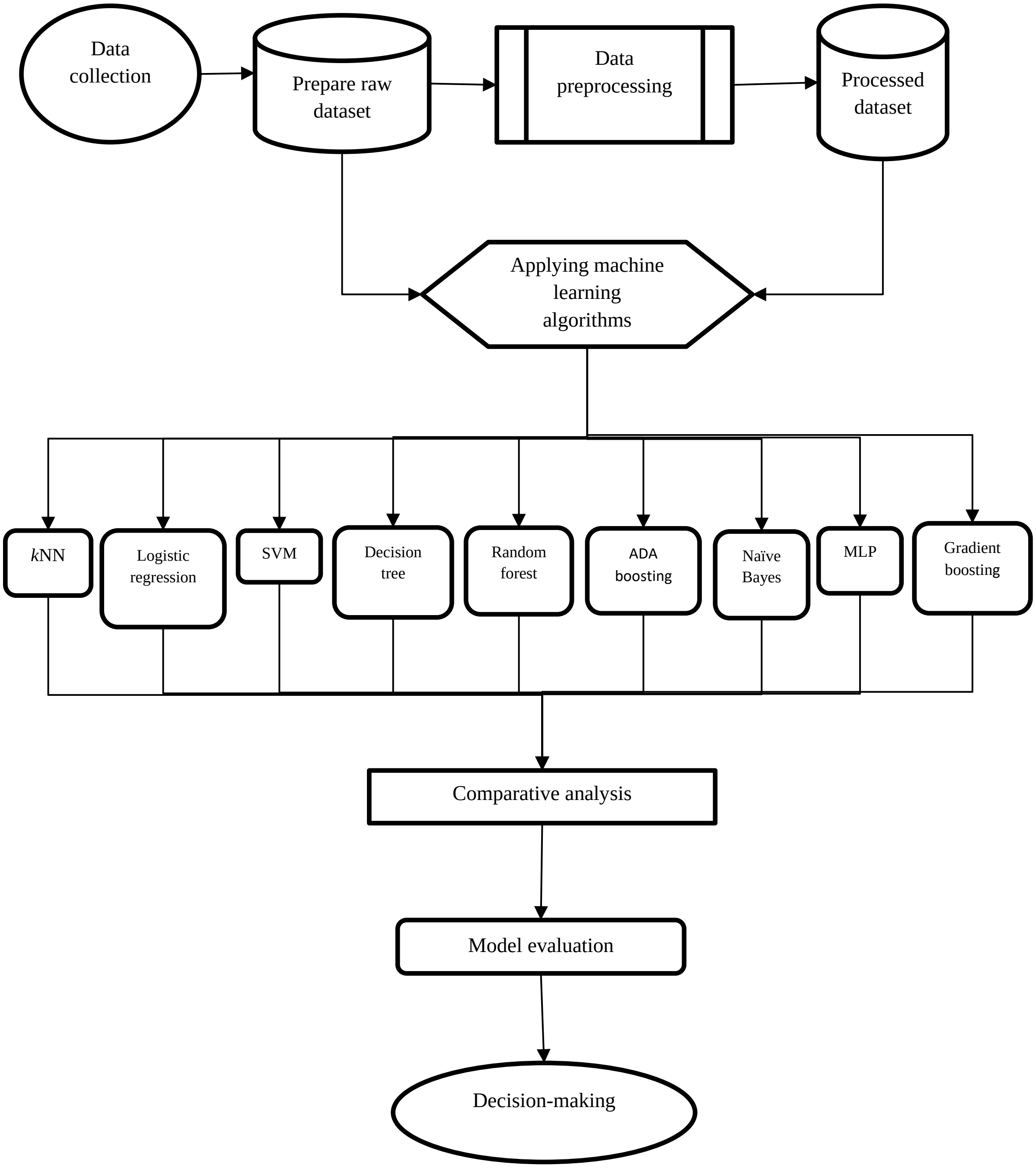


Figure 3.1: Steps of our proposed methodology

3.3.2 Data preprocessing

After collecting the data, we get some missing data, categorical data, numerical and text data. Then we decide that through data processing, we will make this data suitable for algorithms. Data processing is the ability to transform data into a suitable format after collecting data. Processing information or data in a specific format that helps to easily output.

Our data preprocessing method is shown below in Figure 3.2.

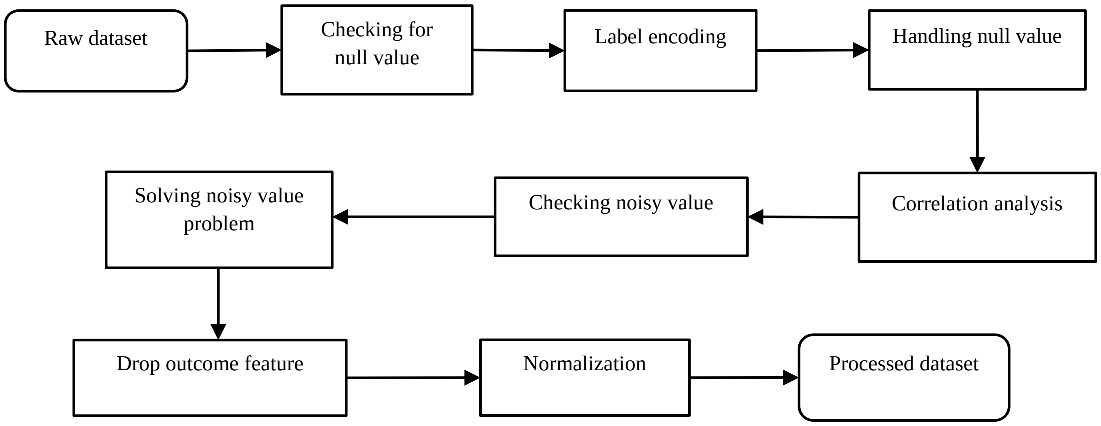


Figure 3.2: Steps of data preprocessing.

First, we started the work of data cleaning. We check if there is a null value in the data set. We then encode the level that converts the text data to numerical data. We solved the missing value problem using imputer and median. Then we check if there is a noisy value in the data set using a box plot. Here we can see that there was some noisy data in the numerical data. Then we analyze the correlation matrix as a data integration process. This matrix shows us the ratio of each data connected to each data. Data is highly connected by a positive value and the negative value means that the data is negatively connected and zero indicates that the data does not connect to itself. We remove noisy values by using outlier quantile detection. Then we drop our outcome feature, that was, the addicted column. We create a separate histogram of each feature that helps us with data reduction and data visualization in feature engineering. Through normalization, we completed the data transformation. Thus, we finally get the processed data set in our hands. This whole process of data processing was done using the “Jupyter Notebook” and “Anaconda navigator”.

3.4 Statistical Analysis

We were able to collect data of 510 people. We collected data on people from different occupations, different ages, and different districts. Figure 3.3 shows that in our data set how many addicted and non-addicted people were. We had prepared our model based on data from 305 addicted people and 205 addicted people.

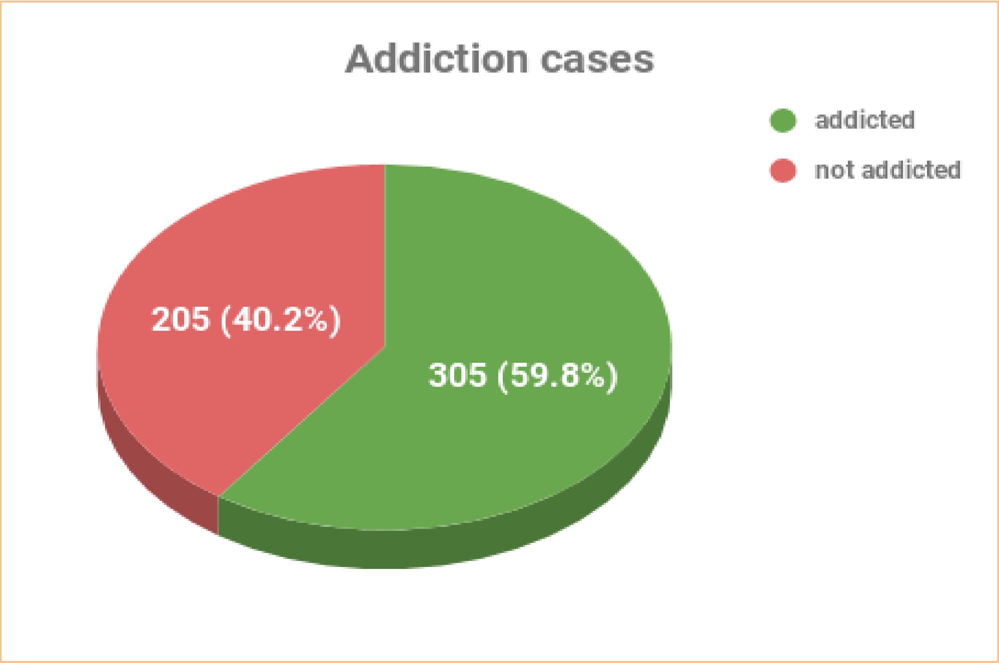


Figure 3.3: Addicted and not addicted cases.

Figure 3.4.2 shows that how many women and how many men were addicted. There were 8 females and 297 males were addicted. Again, there were 68 females and 136 males and one was not willing to reveal his/her gender. Figure 3.4 is shown below.

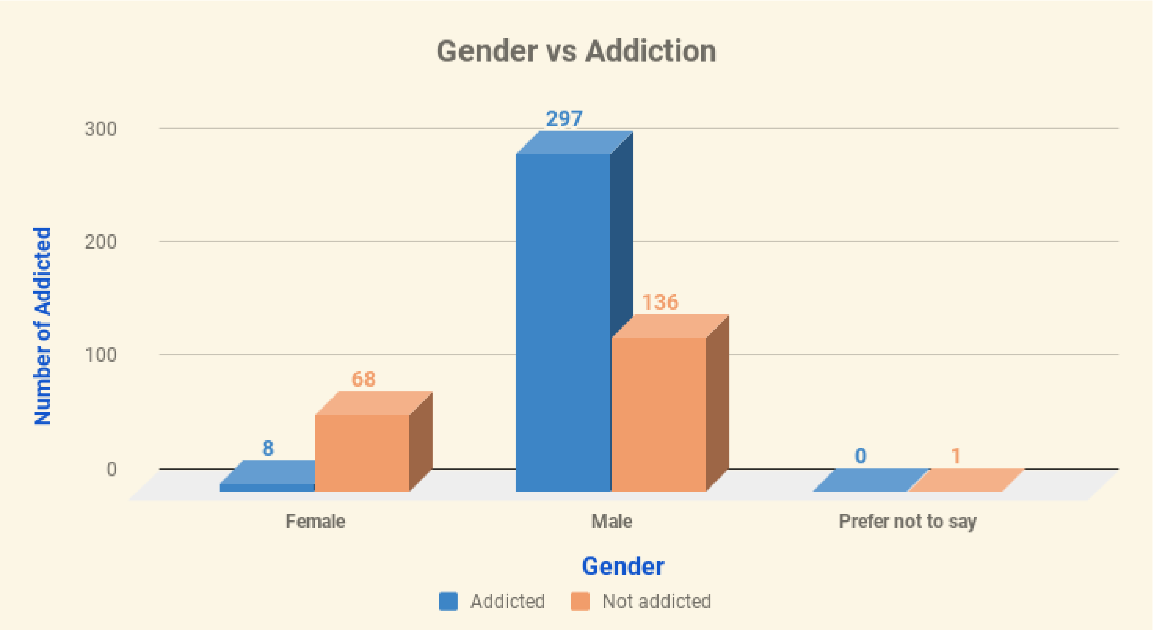


Figure 3.4: Addiction and gender case.

Figure 3.5 shows that information from people of some ages. This picture shows we have information about how many people of any age. Most of the data we collected was about young and middle-aged people.

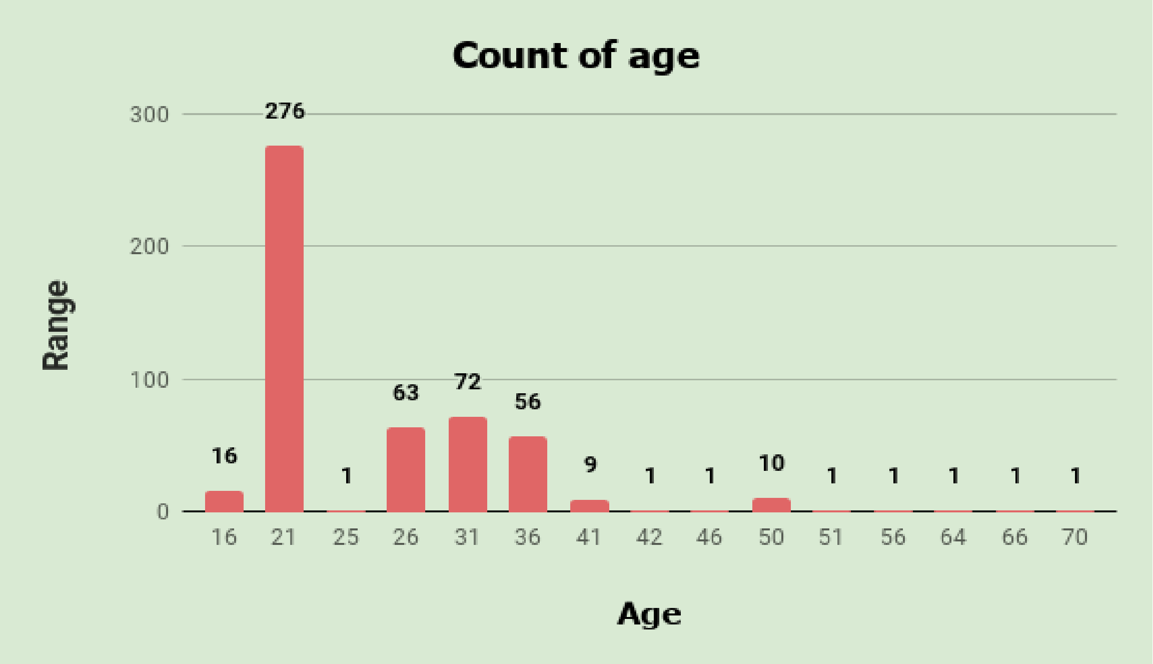


Figure 3.5: Addiction and age case.

Figure 3.6 shows the profession and addiction cases. This picture shows the occupations of people in which we collect information and how many of them were drug addicts.

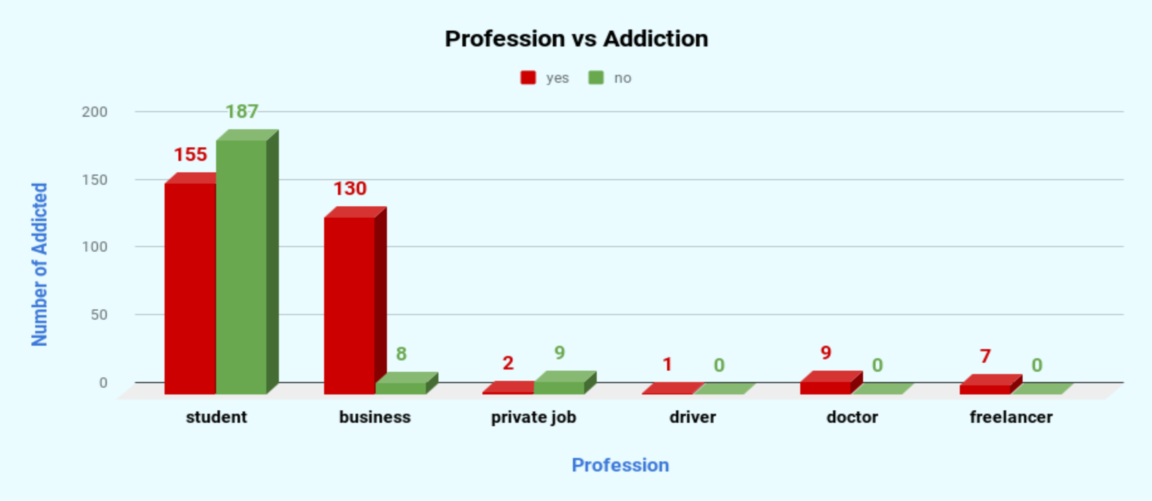


Figure 3.6: Addiction and profession case.

Figure 3.7 shows, that the information we collect is in which district they lived and the number of them. Maximum people were from Dhaka.

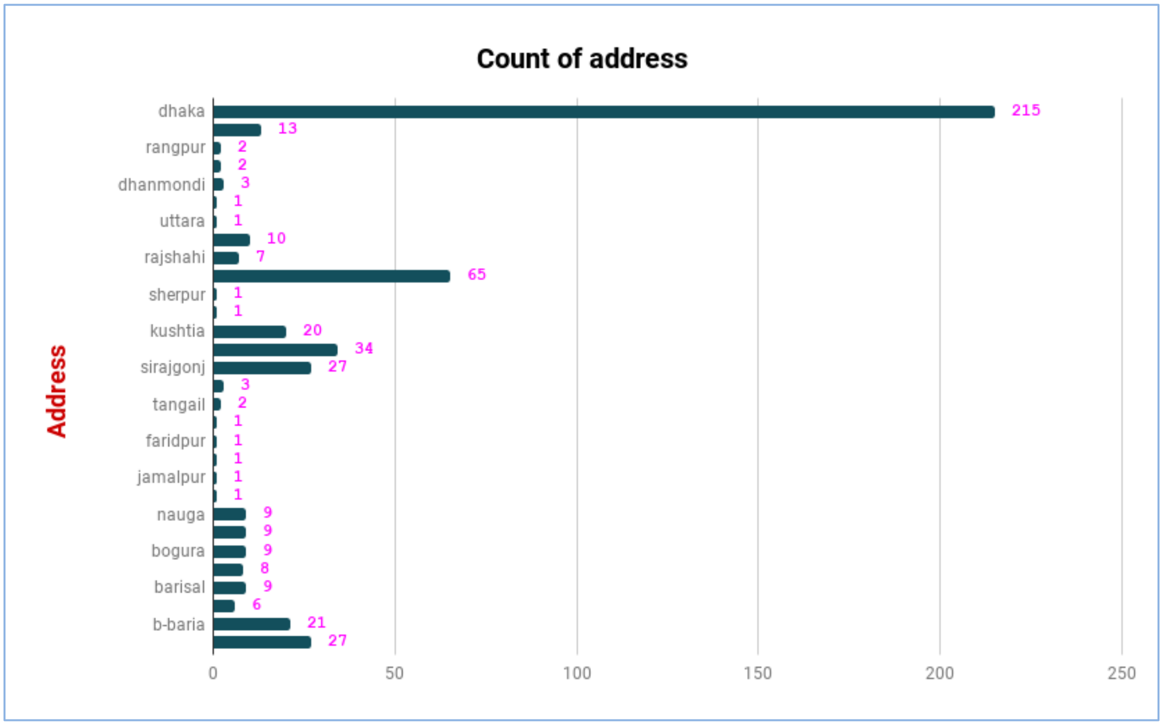


Figure 3.7: Addiction and address case.

Figure 3.8 shows, that the socioeconomic status of the people we have collected our data. Collecting data from lower economic class people was difficult, so their value was zero.

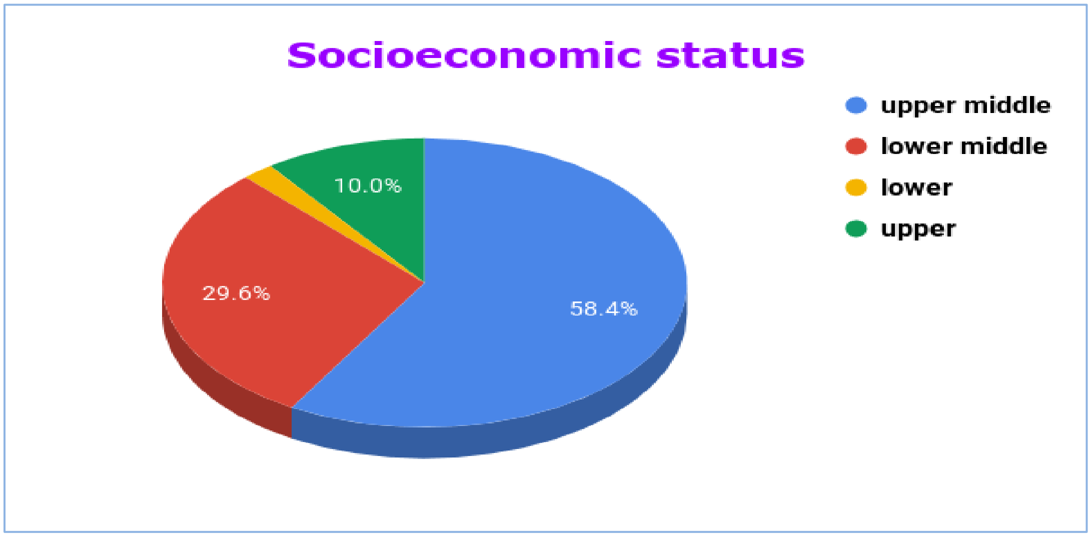


Figure 3.8: Social-economic condition.

Figure 3.9 shows here are the results of how much people care about themselves. It seems drug addicts usually do not care about themselves.

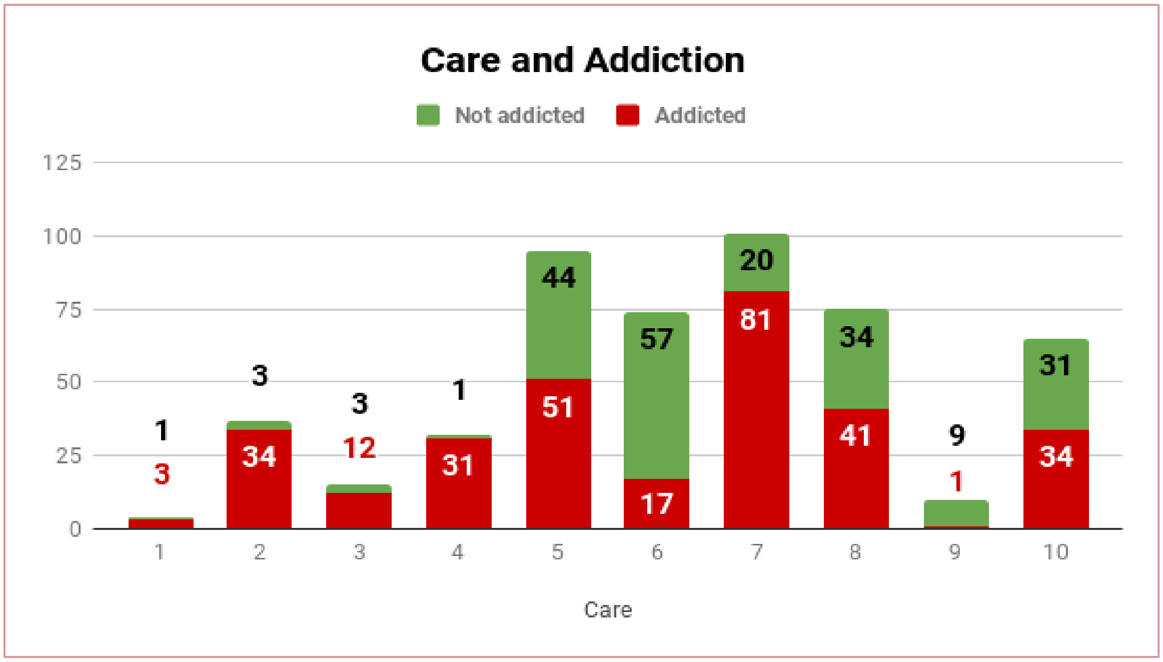


Figure 3.9: Care rate.

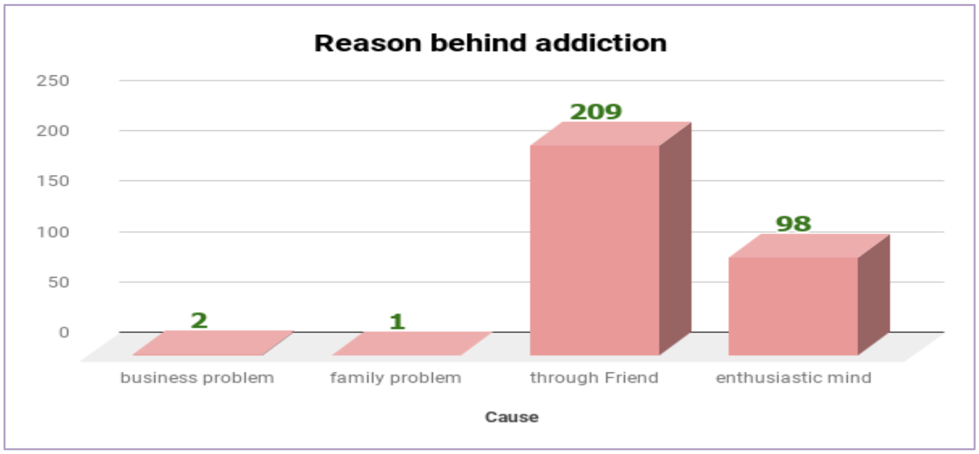


Figure 3.10: Reason behind addiction.

Figure 3.10 shows the reasons for their addiction to drugs. Many people become addicted to drugs because of having a close relation with addictive friends. Curiosity often leads to drug addiction. What kind of data we had on our data set and how they relate to our outcome feature. Figure 3.11 shows the correlation between the features. In addition, we find some noisy values in our dataset. We solve the noisy value problem and using a box plot, we demonstrate the result.

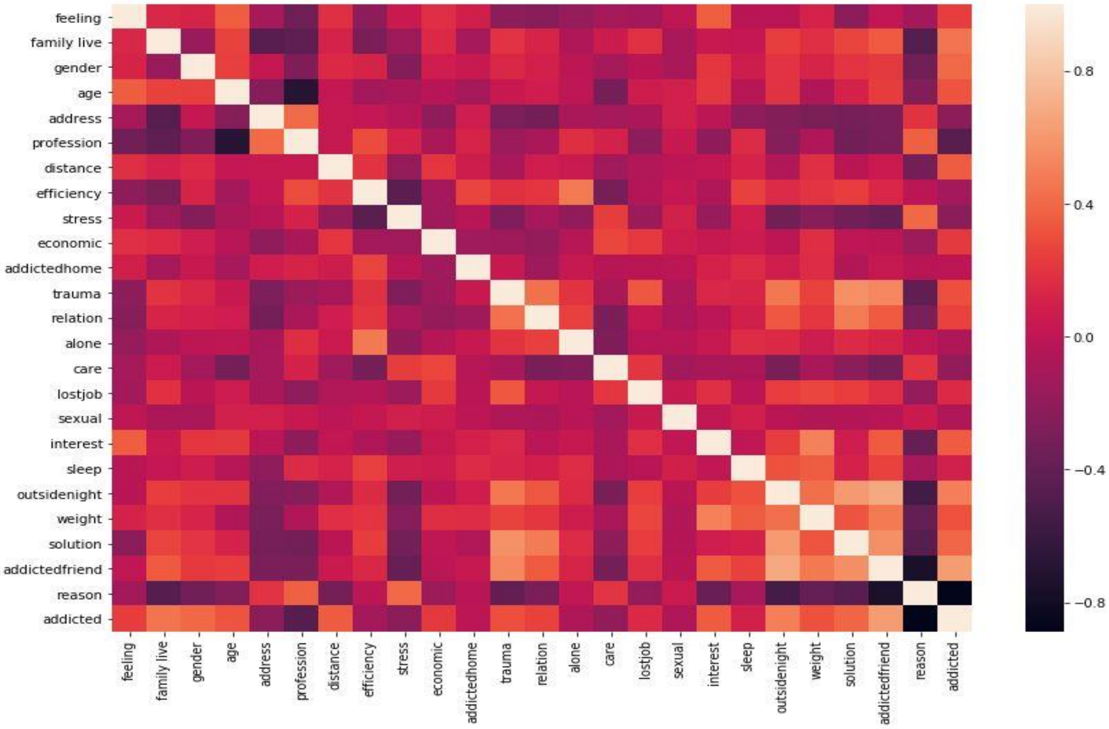


Figure 3.11: Correlation Matrix.

Now a correlation matrix describes the features connectivity to others feature. The statistics have shown that 209 people addicted because of their friends and 98 people addicted to drugs for curiosity mind.

Figure 3.12 shows which feature has noisy value. We had noisy value in the ‘age’ feature and solved it with outlier quantile

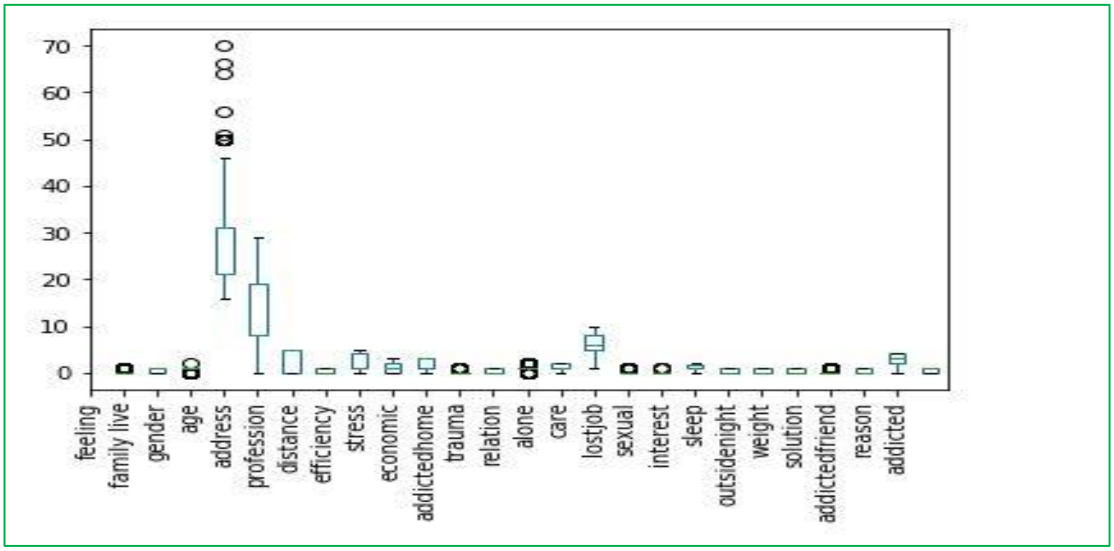


Figure 3.12: Box plot with noisy value.

Figure 3.13 shows that noisy value on age is removed. Normally noisy value remains in numerical data and there are few noisy values finds in texture data.

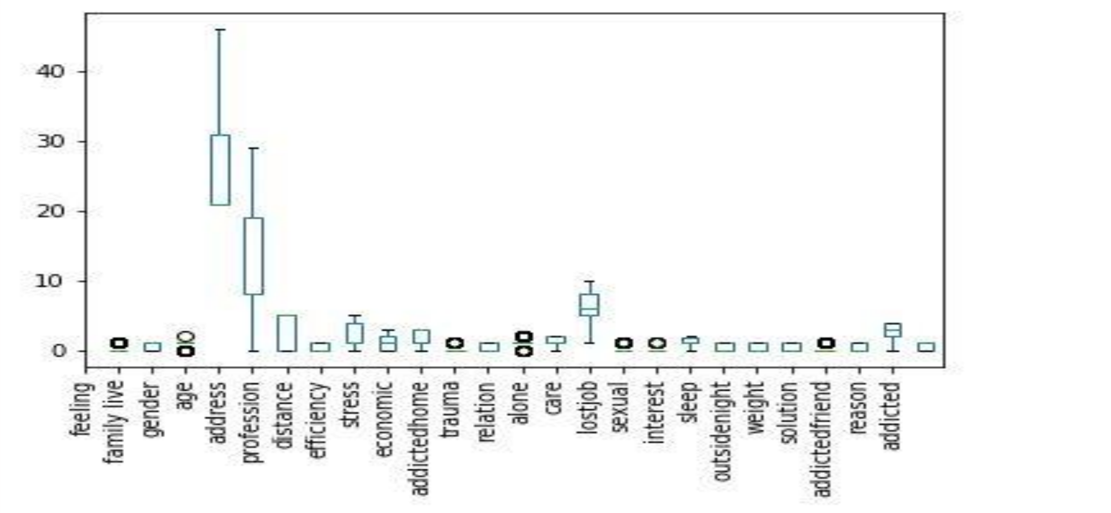


Figure 3.13: Box plot without noisy value.

3.5 Implementation Requirements

We need data mining tools, data processing tools, data storing tools to implement our work. We collect data through Google forms and using handwritten forms. We created data sets with Microsoft Excel. For data preprocessing and algorithms implementation, we used “Anaconda-navigator” and “Jupyter notebook”.

Anaconda Navigator is one kind of graphical user interface for the desktop. It allows users to launch application and conda packages, environment and channel without any command-line command. Anaconda has complete and open-source data science packages [28].

CHAPTER 4

EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Introduction

In the previous section, we discussed the dataset and dataset processing processes. The processed data is used in some algorithms and the results of the algorithm will be discussed in this section. *k*NN, logistic regression, support vector machine (SVM), naïve Bayes, decision tree, random forest, multilayer perception (MLP), ADA boosting classifier and gradient boosting classifier all of these algorithms are used and the results are analyzed to see which algorithm provides the best accuracy. There are basically three steps to calculate accuracy. The accuracies first diagnosed before using PCA on the processed data, then calculate the accuracies after using PCA and finally the accuracies are calculated using the algorithm on the unprocessed data. We collect 510 data of both addicted and non-addicted persons among them 80 percent is used as training data and 20 percent is used as test data. The name of our dataset is 'Drug Dataset v1'.

4.2 Experimental Results & Analysis

We used nine machine-learning algorithms and compared them with each algorithm by calculating their accuracy, confusion matrix, precision, recall, F_1 score, sensitivity, and specificity.

4.2.1 Experimental Evaluation

We run nine machine-learning algorithms on processed datasets where the number of the feature was 24. Then we use the PCA. PCA means principal component analysis. It is one kind of feature extraction method, which uses to grab the underlying variance of data in orthogonal linear projections. In the case of dimensionality reduction, PCA is used. The independent used variable of a model is known as the dimensionality of that model. The number of variables can be reduced using a PCA, only the important variables are selected for the next task. Normally it combines highly correlated variables together to build up a short artificial set of variables [29].

Figure 4.1 shows the accuracy of nine algorithms. It appears that before using PCA, *k*NN has achieved 96.8% accuracy, SVM has achieved 93.75% accuracy, logistic regression has achieved 84.37% accuracy, Naïve Bayes has achieved 87.5% accuracy, random forest has achieved 66.67% accuracy, decision tree has achieved 50% accuracy, ADA boosting classifier has achieved 69.79% accuracy, MLP has achieved 78.13% accuracy, gradient boosting classifier has achieved 73.96% accuracy.

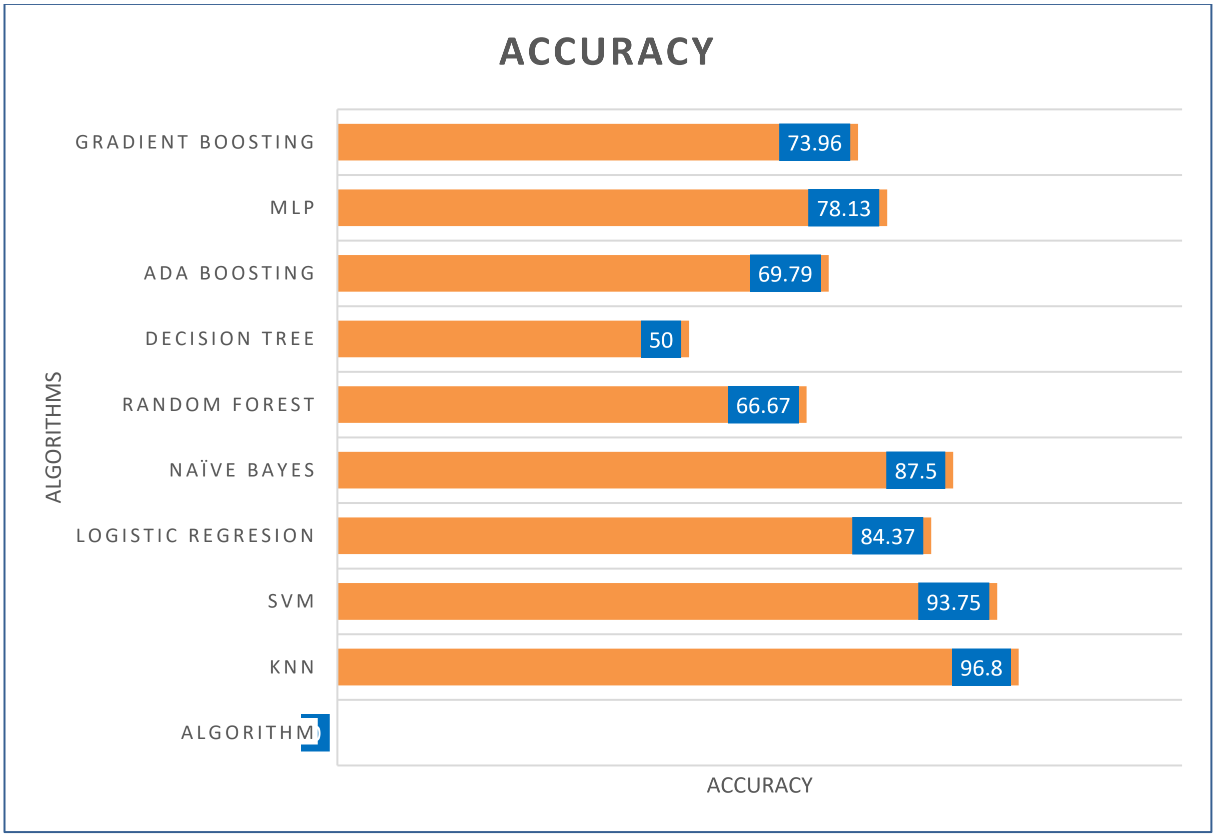


Figure 4.1: Accuracy before applying PCA.

We calculate the accuracy again using PCA and in PCA, we used 14 features instead of 24 features. Since the number of features has changed due to the use of PCA, the change in accuracy is noticed.

Figure 4.2 shows that the accuracy of nine algorithms after performing PCA. After using PCA, we can see that the accuracy of some algorithms has increased and some algorithms have decreased and some algorithms have remained unchanged. , *k*NN has achieved 82.29% accuracy, SVM has achieved 95.83% accuracy, logistic regression has achieved 97.91% accuracy, naïve Bayes has achieved 92.7% accuracy, random forest has achieved 73.95% accuracy, decision tree has achieved 59.37% accuracy, ADA boosting classifier has achieved 71.87% accuracy, MLP has achieved 72.91% accuracy, gradient boosting classifier has achieved 59.38% accuracy.

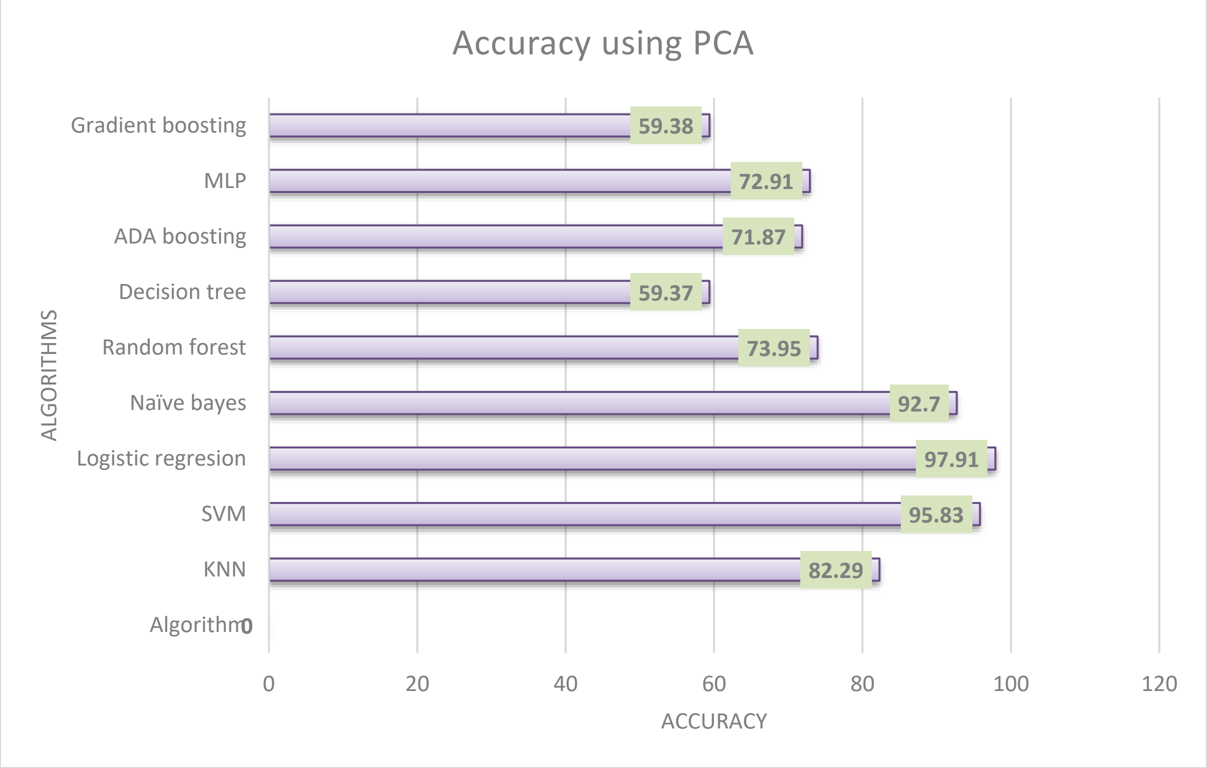


Figure 4.2: Accuracy after applying PCA.

Figure 4.3 shows the accuracy of nine algorithms with unprocessed data. *k*NN has achieved 81.37% accuracy, SVM has achieved 59.01% accuracy, logistic regression has achieved 58.82% accuracy, naïve Bayes has achieved 57.84% accuracy, random forest has achieved 73.52% accuracy, decision tree has achieved 57.84% accuracy, ADA boosting classifier has achieved 71.56% accuracy, MLP has achieved 58.82% accuracy, gradient boosting classifier has achieved 73.52% accuracy.

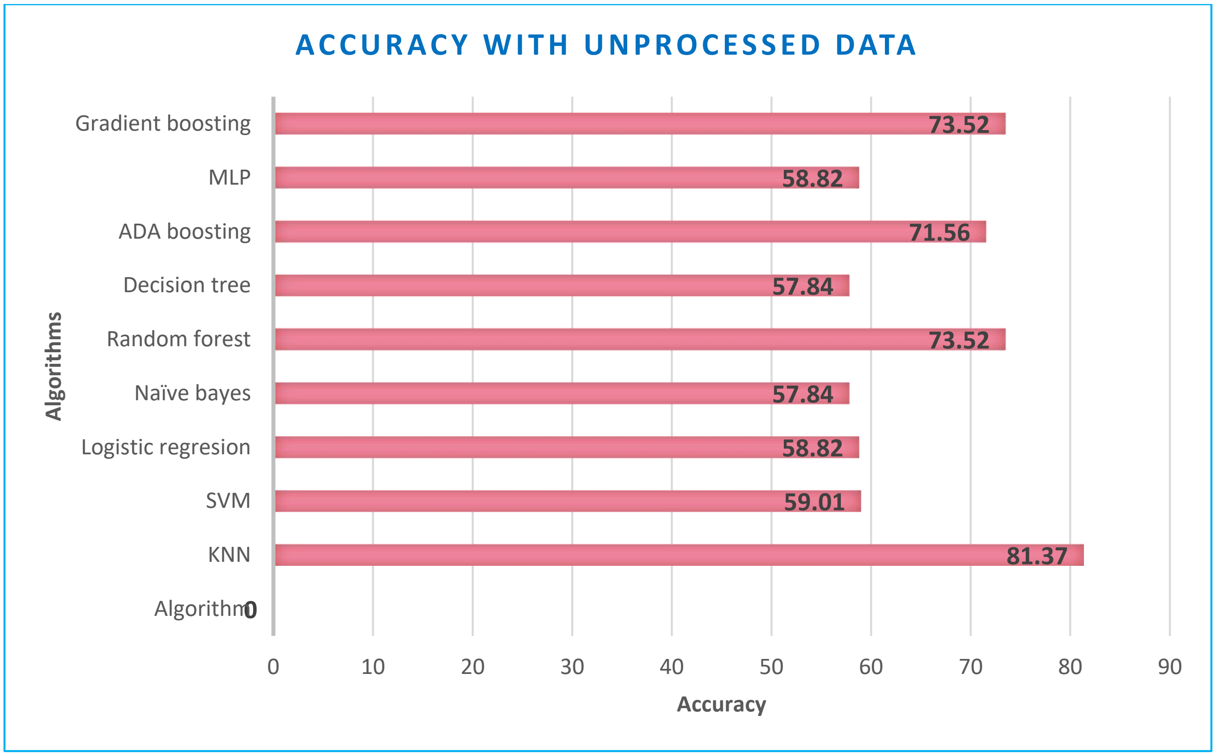


Figure 4.3: Accuracy with unprocessed data.

K-nearest neighbors (*k*NN) is a simple supervised machine-learning algorithm. Classification and regression problems can be solved with the *k*NN algorithm. *k*NN algorithm memorizes the training observation for classifying the unseen test data. *k*NN algorithm grabs similar things that exist in a close neighborhood [30].

Support vector machine is a supervised machine-learning algorithm. This also used for both classification and regression problems. Data items are placed in n-dimensional space and the values of the features are presented the particular coordinate. It creates the most homogeneous points in each subsection which is why it is called hyperplane [30].

Logistic regression used logistic function and this Logistic function is called a sigmoid function. An S-shaped curve takes the real values and put them between 0 to 1 [30].

Naïve Bayes is one of the oldest algorithms of machine learning. This algorithm is based on Bayes theorem and basic statistics. Class Probabilities and conditional Probabilities are used in the Naive bias model. It extends attributes using Gaussian distribution [29].

Decision tree is a tree-based model. It distributing the features into the smaller section with similar response value using splitting rules. The divide-and-conquer method uses for making the tree diagram. Decision tree needs a small pre-processing and it can easily control the categorical features without preprocessing [29].

Yoav Freund and Robert Schapire propose ADA boosting or Adaptive boosting in 1996. It makes a classifier with a combination of multiple poorly performing classifier. It set the weight of classifiers and train the data in each iteration [29].

Random forest makes a large collection of de-correlated trees for prediction purposes. It reduces tree correlation by injecting randomness into the tree growing process. It performs split-variable randomization. Random forest has a smaller feature search space at each tree split [29].

Gradient boosting classifier build an ensemble of shallow trees with tree learning and improving technique. Gradient boosting classifier works with the principle of boosting weak learners iteratively by shifting focus towards problematic observation. It prepares a stage-wise fashion model like others boosting methods and normalizes them with arbitrary differentiable functions [29].

MLP means multilayer perception. MLP has a combination of multilayer neurons. The first layer is the input layer, the second layer is called the hidden layer and the third layer is called the output layer. It takes input data through the input layer and gives the output from the output layer [29].

Table 4.1 shows that logistic regression has achieved the highest accuracy among all of them with 97.91% accuracy. *k*NN has achieved the highest accuracy before applying PCA. Again, logistic regression has achieved the highest accuracy after applying PCA with 97.91%. *k*NN has achieved the highest accuracy with 81.37% on unprocessed data.

TABLE 4.1: SUMMARY OF ACCURACY

Algorithms	**Accuracy before applying PCA (%)**	**Accuracy after applying PCA (%)**	**Accuracy with unprocessed data (%)**
*k*NN	96.8	82.29	81.37
SVM	93.75	95.83	59.01
Logistic regression	84.37	97.91	58.82
Naïve Bayes	87.5	92.7	57.84
Random forest	66.67	73.95	73.52
Decision tree	50	59.37	57.84
ADA boosting classifier	69.79	71.87	71.56
MLP	78.13	72.91	58.82
Gradient boosting classifier	73.96	59.38	73.52

4.2.2 Descriptive Analysis

We not only calculated the accuracy of several algorithms but also calculated sensitivity, specificity, precision, recall, f-score, and roc-curve and confusion matrix of each algorithm. Evaluation of that model is required for any model selection. In the case of model evolution, certain classifiers have to be measured. Classifications are measured based on the test data set for better Measurement.

Sensitivity is the true positive rate. That is, sensitivity is the ratio of how many positive tuples correctly diagnosed.

$$Sensitivity = \frac{TP}{TP + FN} \times 100\% \quad (1)$$

Specificity is the true negative rate. That is, specificity is the ratio of how many negative tuples are correctly diagnosed.

$$Specificity = \frac{TN}{FP + TN} \times 100\% \quad (2)$$

Precision is the measurement of exactness. It is the ratio of true positive value and predicted positive value.

$$Precision = \frac{TP}{TP + FP} \times 100\% \quad (3)$$

A recall is the measurement of completeness. It is the ratio of true positive value and true positive value.

$$Recall = \frac{TP}{TP + FN} \times 100\% \quad (4)$$

F1 score is the measurement of the harmonic mean of recall and precision. It considers both false positive and false negative values for calculation.

$$F_1 \text{ score } = \frac{2 \ x \ precision \, x \ recall}{precision + recall} \times 100\% \quad (5)$$

Receiver operating characteristics (roc) curves is very useful for visual comparison of classification models. ROC curve is made with a true positive rate and false-positive rate. The diagonal line is representing the random guessing. The curve of a model is close to random guessing, which is a less accurate model. Therefore, for an accurate model, the curve will be far away from the random guessing line. The ROC curves of our using algorithms are given below.

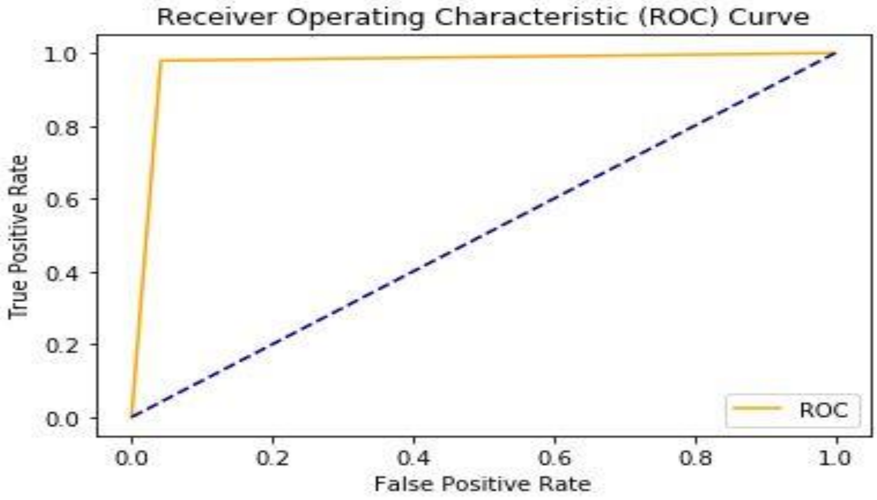


Figure 4.4: ROC curve of the *k*NN algorithm.

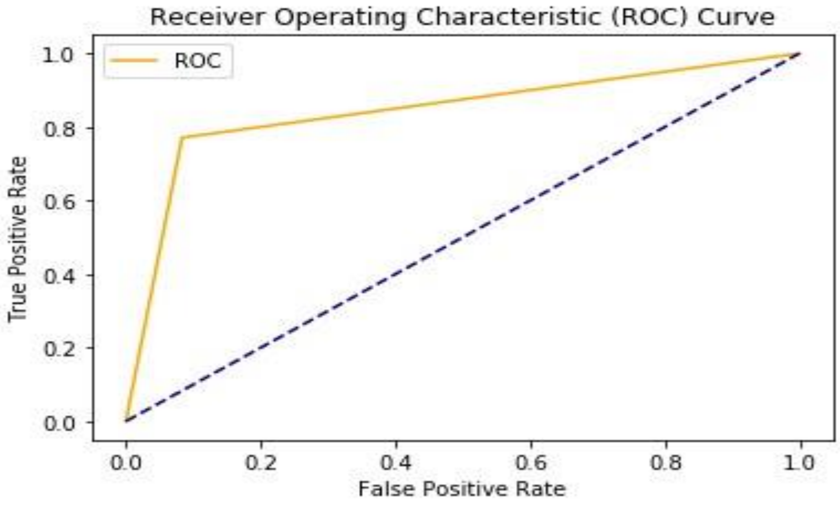


Figure 4.5: ROC curve of the Logistic regression algorithm.

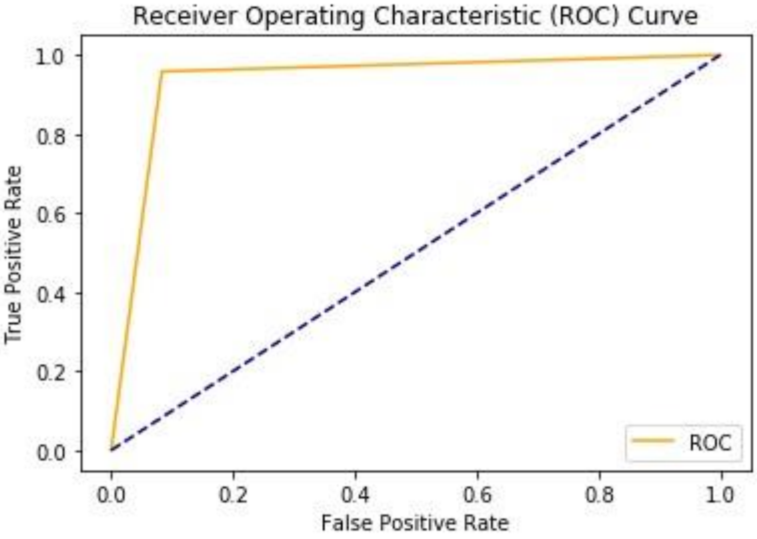


Figure 4.6: ROC curve of the SVM algorithm.

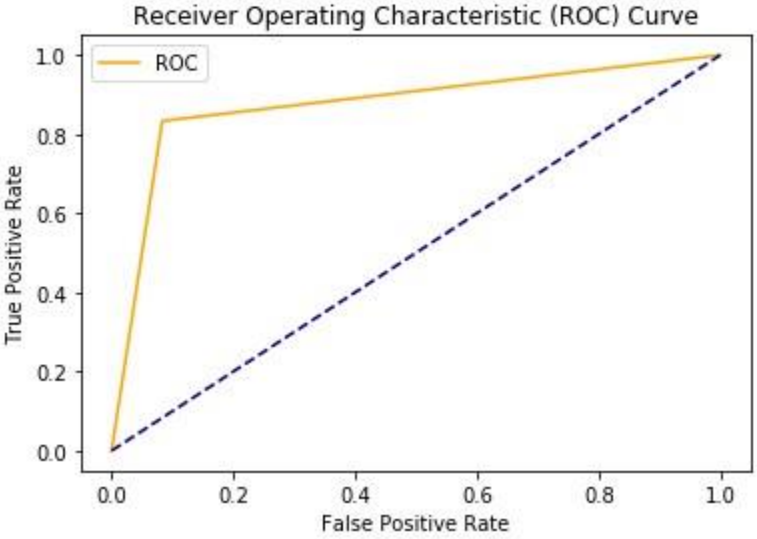


Figure 4.7: ROC curve of Naïve Bayes algorithm.

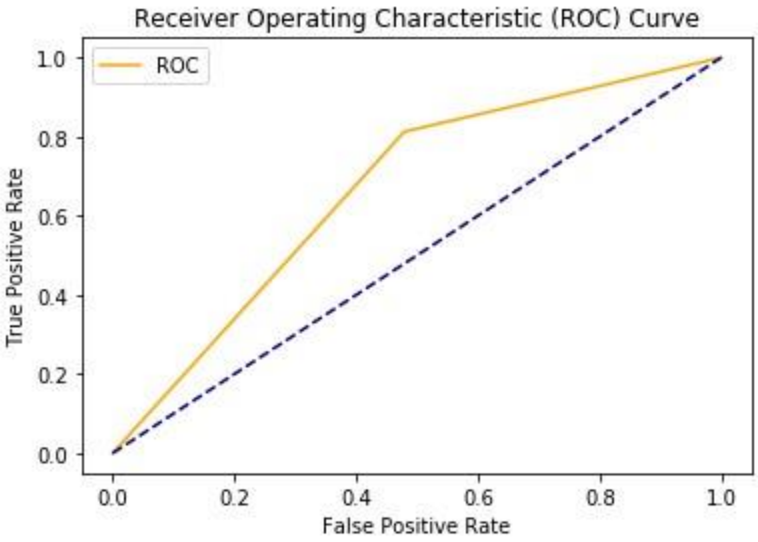


Figure 4.8: ROC curve of the Random forest algorithm.

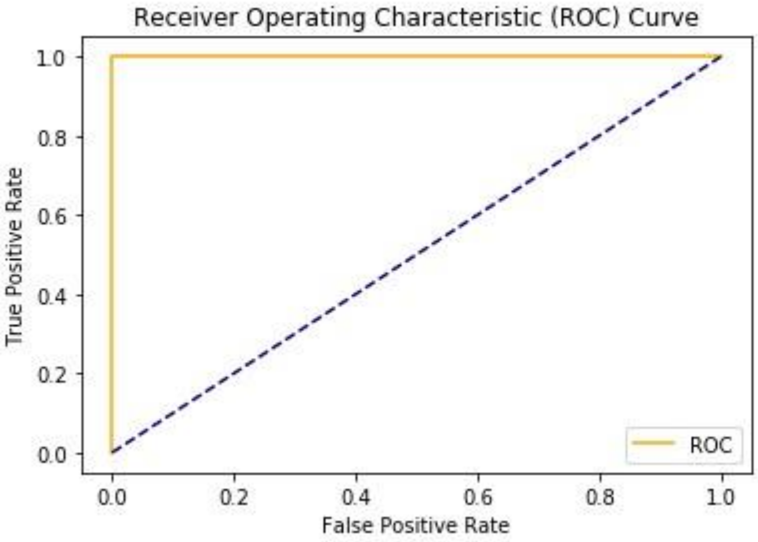


Figure 4.9: ROC curve of the Decision tree algorithm.

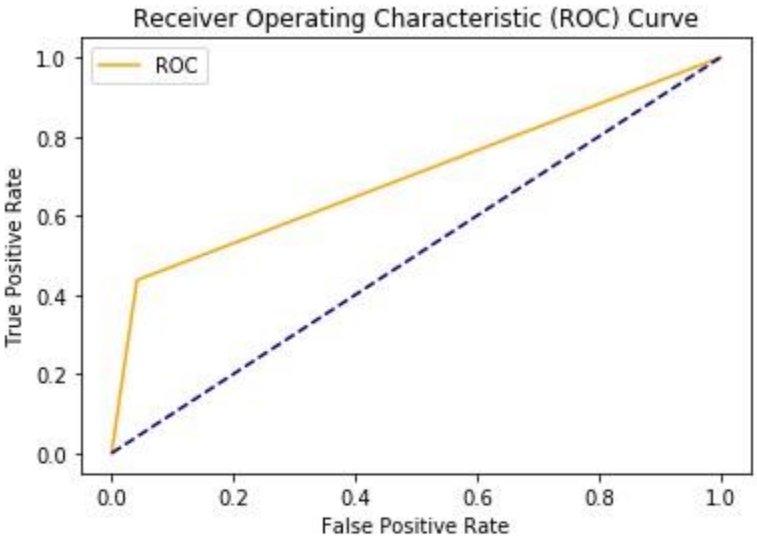


Figure 4.10: ROC curve of ADA boosting classifier.

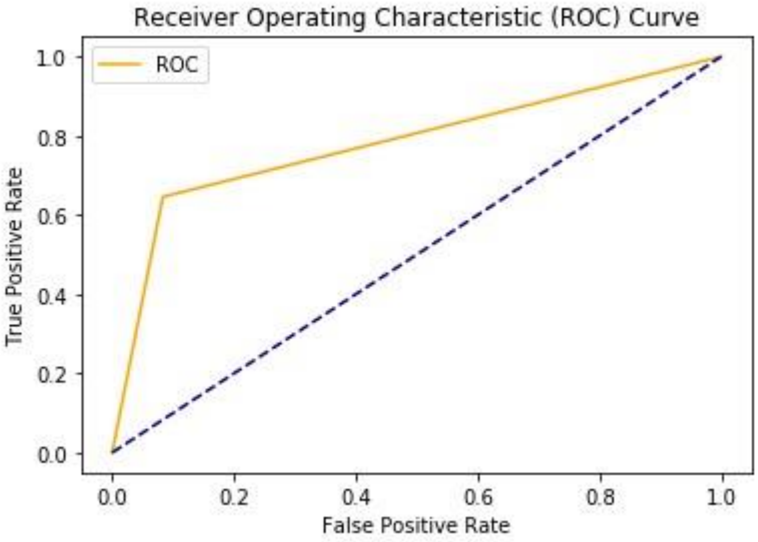


Figure 4.11: ROC curve of MLP algorithm.

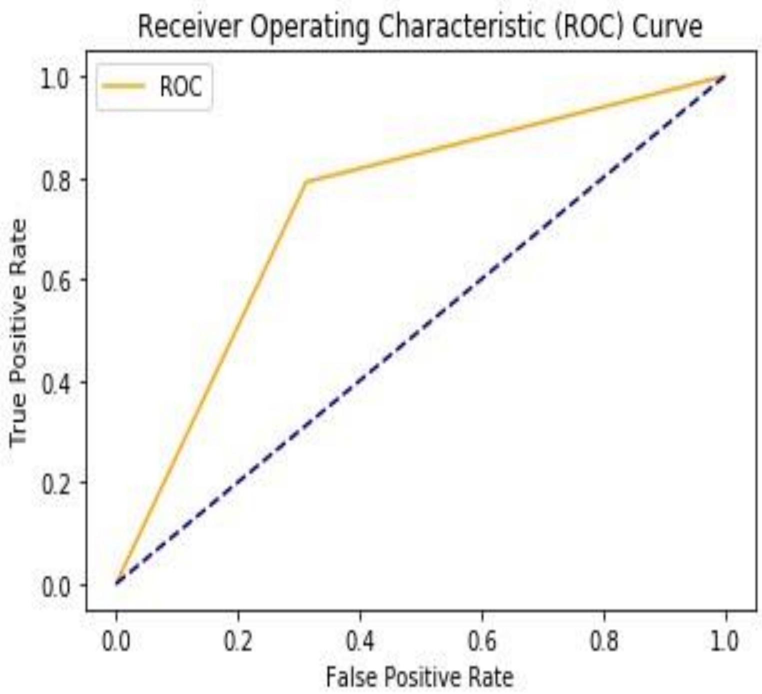


Figure 4.12: ROC curve of Gradient boosting classifier.

Confusion Matrix is one of the most important performance measurement techniques for machine learning classification. It will perform on the classification models with the set of test data and provide the true positive value, true negative value, false-positive value and false-negative value in a tabular format. The Confusion Matrix is very important for measuring the performance of any classifier.

Table 4.2 shows the confusion matrix of all algorithms used in our model. Now, model evaluation of each classifier is described with value in the following table.

TABLE 4.2: CONFUSION MATRIX OF ALL CLASSIFIER.

<table>
<tr><th>Algorithms</th><th colspan="4">Confusion Matrix</th><th>Algorithms</th><th colspan="4">Confusion matrix</th></tr>
<tr><td rowspan="4">*k*NN</td><td rowspan="4">True Class</td><td></td><td>No</td><td>Yes</td><td rowspan="4">Logistic
Regression</td><td rowspan="4">True Class</td><td></td><td>No</td><td>Yes</td></tr>
<tr><td>No</td><td>46</td><td>2</td><td>No</td><td>44</td><td>4</td></tr>
<tr><td>Yes</td><td>1</td><td>47</td><td>Yes</td><td>11</td><td>37</td></tr>
<tr><td colspan="3">Predicted Class</td><td colspan="3">Predicted Class</td></tr>
</table>

<table>
<tr><td>SVM</td><td><table><tr><td rowspan="4">True Class</td><td></td><td>No</td><td>Yes</td></tr><tr><td>No</td><td>44</td><td>4</td></tr><tr><td>Yes</td><td>2</td><td>46</td></tr><tr><td colspan="3">Predicted Class</td></tr></table></td><td>Gaussian
Naïve
Bayes</td><td><table><tr><td rowspan="4">True Class</td><td></td><td>No</td><td>Yes</td></tr><tr><td>No</td><td>44</td><td>4</td></tr><tr><td>Yes</td><td>8</td><td>40</td></tr><tr><td colspan="3">Predicted Class</td></tr></table></td></tr>
<tr><td>Random
Forest</td><td><table><tr><td rowspan="4">True Class</td><td></td><td>No</td><td>Yes</td></tr><tr><td>No</td><td>25</td><td>23</td></tr><tr><td>Yes</td><td>9</td><td>39</td></tr><tr><td colspan="3">Predicted Class</td></tr></table></td><td>Decision
Tree</td><td><table><tr><td rowspan="4">True Class</td><td></td><td>No</td><td>Yes</td></tr><tr><td>No</td><td>48</td><td>0</td></tr><tr><td>Yes</td><td>0</td><td>48</td></tr><tr><td colspan="3">Predicted Class</td></tr></table></td></tr>
<tr><td>ADA
Boosting
classifier</td><td><table><tr><td rowspan="4">True Class</td><td></td><td>No</td><td>Yes</td></tr><tr><td>No</td><td>46</td><td>2</td></tr><tr><td>Yes</td><td>27</td><td>21</td></tr><tr><td colspan="3">Predicted Class</td></tr></table></td><td>MLP</td><td><table><tr><td rowspan="4">True Class</td><td></td><td>No</td><td>Yes</td></tr><tr><td>No</td><td>44</td><td>4</td></tr><tr><td>Yes</td><td>17</td><td>31</td></tr><tr><td colspan="3">Predicted Class</td></tr></table></td></tr>
<tr><td>Gradient
Boosting
classifier</td><td colspan="3"><table><tr><td rowspan="4">True Class</td><td></td><td>No</td><td>Yes</td></tr><tr><td>No</td><td>33</td><td>15</td></tr><tr><td>Yes</td><td>10</td><td>38</td></tr><tr><td colspan="3">Predicted Class</td></tr></table></td></tr>
</table>

Table 4.3 describes the performance of each algorithm. Based on these performances of algorithms and their accuracy performance, which algorithm will fit for our model that will be decided. Based on this accuracy it can be seen that logistic regression performs the best. Again based on sensitivity, specificity, recall, precision, the decision tree performs better. However, after performing unprocessed data and PCA, decision Tree's performance was

not good. So considering everything, it is possible to get the best performance in the model using algorithms.

TABLE 4.3: CLASSIFIER PERFORMANCE EVALUATION.

Algorithms	**Accuracy (%)**	**Sensitivity (%)**	**Specificity (%)**	**Precision (%)**	**Recall (%)**	***F₁* score (%)**
*k*NN	82.29	95.8	97.9	95.91837	97.917	96.90722
SVM	95.83	91.66	95.83	92	95.833	93.87755
Logistic regression	97.91	91.66	77.08	90.2439	77.083	83.14607
Naïve Bayes	92.7	91.66	83.33	90.90909	83.333	86.95652
Random forest	73.95	52.08	81.25	62.90323	81.25	70.90909
Decision tree	59.37	100	0	100	100	100
ADA boosting	71.87	95.83	43.75	91.30435	43.75	59.15493
MLP	72.91	91.66	64.58	88.57143	64.583	74.6988
Gradient boosting	59.38	68.75	79.17	71.69811	79.167	75.24752

4.3 Comparative Analysis

The purpose of our work is to predict the addiction to drugs and alcohol. In paper [13], predicting daily smoking behavior with five features and collected data from 15095 people. In paper [15], tobacco disease detection with 180 hyperspectral images with 32 features. In paper [16], predicting HIV prognosis and mortality with smoking-associated DNA with 0.78 AUC. In paper [18], predicting smoking status by collecting patients' blood tests and health associated vital readings. In paper [19], predicting alcohol use disorder by checking the treatment-seeking status of patients and they did not mention the accuracy of their work. In paper [20], predicting the risk of alcohol use disorder with a different types of data and they did not mention the classifier and accuracy. In paper [21], predicting alcohol abuse with ANN and achieved 98.7% accuracy. Table 4.4 shows a general overview of other works including our work.

TABLE 4.4 RESULTS OF THE COMPARISON OF OUR WORK AND OTHERS' WORKS

Method/ Work Done	Object(s) Deal with	Problem Domain	Sample size	Size of Feature set	Algorithm	Accuracy
This work	Drugs and alcohol (risk)	Prediction	510	23	Logistic regression	97.91%
Zhang et al. [13]	Smoking behavior	Prediction	15095	5	XGboost	84.11%
Zhu et al. [15]	Tobacco diseases	Detection	180	32	ELM	98.3%
Zhang et al. [16]	HIV prognosis with smoking-associated DNA	Prediction	1137	698	GLMNET	0.78 AUC
Frank et al. [18]	Smoking status	Prediction	534	3	Logistic regression	83.44%
Lee et al. [19]	Alcohol use disorder (treatment seeking)	Prediction	778	10	Logistic regression	*NM*
Kinreich et al. [20]	Alcohol use disorder (risk)	Prediction	656	3	*NM*	*NM*
Kumari et al. [21]	Alcohol abuse	Prediction	1885	12	ANN	98.7%

[1] *NM*: Not Mentioned

4.4 Discussion

This section reviews the performance of algorithms, accuracy, sensitivity, specificity, recall, precision, F_1 score, and ROC curve. Also discussed here are the equations of evolution models and their function. We can see that the logistic regression algorithm yields the highest accuracy with 97.91%. As well as the logistic regression algorithm achieved 91.66% sensitivity, 77.08% specificity, 90.24% precision, 77.08% recall and 83.14% F_1 score. Finally, we find out that using the logistic regression algorithm we can get the best performance on our drug addiction risk prediction model.

CHAPTER 5

IMPACT ON SOCIETY, ENVIRONMENT, AND SUSTAINABILITY

5.1 Impact on Society

Addiction to drugs and alcohol risk prediction with the machine learning model will have a positive impact on society. Humans are social beings so people of all religions and castes have to move together to live in a society. On the way together, a drug addict becomes an obstacle. People can become addicted to drugs at any time. In the discussion above, we have seen that people are addicted to drugs from curiosity, from drug addicts' friends, and many use drugs to stay away from various traumas. This is how the younger generation of our society is slowly turning to drugs. Parents and guardians should always be aware of and care for their children. It is the parents' responsibility to give their children time, use them in a friendly manner, and observe their activities. If ever a parent is in doubt, he/she will use this model to provide the information and data needed here so that he/she will know the possibility of their child being addicted to the drug. In this way, we can protect our young society before becoming addicted to drugs. We think our drug addiction prediction model will be used for the development of all in society.

5.2 Impact on Environment

Our model is certainly not detrimental to the environment. No chemicals, combustibles, and organic acids are needed to operate this model. Therefore, this model will not have any adverse effects on the environment and biodiversity. Drugs and alcohol addicts use different types of plastics to serve drugs. All these plastic and waste materials are a threat to our environment. The use of this model will keep people away from drugs and reduce the number of drug addicts. Even if the number of drug addicts decreases, the number of plastics used for serving drugs will decrease. If plastic is used less, it will certainly be beneficial for our environment.

5.3 Ethical Aspects

This addiction prediction model is not anti-moral and does not violate human rights in such a way The model does not collect any personal information, name, identity etc. so there will be no privacy problem. This model does not undermine a person's right to enjoy or use, but rather plays an important role in making a person aware. The risk of addiction to drugs and alcohol prediction model was created with respect to all types of rules and with respect to privacy and confidentiality issues. So using machine-learning technology, the model of prediction of addiction to drugs can be managed without any problems.

5.4 Sustainability Plan

The sustainability plan has three parts they are community financial and organizational. The Sustainability Plan gives us a realistic idea of any project run and future plans for the project. Our model cohort mission is to find the tendency to be addicted to drugs. This model has to be targeted to make it easy for people to adjust and it is important to keep in mind that people do not suffer from inferiority to use this model. Police, law enforcement, and narcotics control departments can use this model to speed up their work.

CHAPTER 6

SUMMARY, CONCLUSION, RECOMMENDATION, AND IMPLICATION FOR FUTURE RESEARCH

6.1 Summary of the Study

Our work is divided into several parts like Data collection, Data preprocessing, Methodology implementation, and Experimental evaluation. We collected the necessary data from Mukti Clinic and Brain & Mind Hospital. We collected data on both drug addicts and healthy individuals. After data collection, we do data processing and work on data processing and implementation using Anaconda Navigator and Jupyter Notebook. After preprocessing, we run nine machine-learning algorithms and they are *k*NN, Logistic regression, Gaussian naïve Bayes, Random forest, ADA boosting, Decision tree, multilayer perceptron and Gradient boosting classifier and their performance on their accuracy, sensitivity, precision, etc. is considered. It is noticeable that the logistic regression algorithm gives the best performance. Therefore, the tendency to be addicted to drugs has to be modeled using the logistic regression algorithm for prediction.

6.2 Limitations and Conclusions

Our study is about addiction to the drug prediction systems which machine learning algorithms. We have some limitations and deficiencies in our work and model. The data set we used was not comparatively large, it would have been better to use a larger and richer data set. Due to some limitations, people from different professions, people from different districts and different classes could not collect data. Many advanced methods could also be used for data processing, and the model could be presented beautifully using different variations in the application of algorithms.

With our proposed model, it is possible to determine the tendency of drug addiction. We hope that this model will use very easily by the common people once it is fully formed and will be able to realize the importance of this model in raising awareness. It is important to always be vigilant in order to avoid the horrors of the drug and not get addicted to it. People gradually come into contact with the drug and become addicted to the drug so it is difficult to stop the person from being addicted if they do not take effective measures at the beginning. We are hopeful that this model will keep people away from drug exposure and that people will be aware of their situation and control himself or herself.

6.3 Implication for Further Study

Nowadays technology and modern science make our life fast and easier. We want to use our model in the future in a software or web application or an Android application, in the continuation that information technology and the internet are used in our country. In the future, we will be able to increase the accuracy of our model using a larger database. In addition, by creating user-friendly GUIs, the software created by the model can be reached to the people. In the future, implementing new algorithms, adding different parameters, and adding some more features can be made more effective from the model. In the future, a robust database can be created by collecting data from different categories of people according to the district. In addition, with the help of the Department of Drug Control, the model can be made larger and taken forward.

REFERENCES

[1]. Dhiraj Kumar Nath, “Control of Drug Abuse Is A Must", *The Daily Star*, 2019. [Online]. Available: https://www.thedailystar.net/health/health-alert/control-drug-abuse-must-1515874. [Accessed: 18- June-2019].

[2]. M. N. Shazzad, S. Abdal, M. S. Majumder, J. ul Sohel, S. M. Ali, and S. Ahmed, “Drug Addiction in Bangladesh and its Effect”, *MEDTODAY*, vol. 25, no. 2, pp. 84-89, Feb. 2014.

[3]. Shaheen Mollah, “Restricted, she killed parents", *The Daily Star*, 2013. [Online]. Available: https://www.thedailystar.net/news/restricted-she-killed-parents. [Accessed: 25- June- 2019].

[4]. Arifur Rahman Rabbi, “43% of unemployed population addicted to drugs", *Dhaka Tribune*, 2019. [Online]. Available: https://www.dhakatribune.com/bangladesh/dhaka/2019/02/27/43-of-unemployed-population-addicted-to-drugs. [Accessed: 25- June- 2019].

[5]. Cruz, A. Joseph, and D. S. Wishart. “Applications of Machine Learning in Cancer Prediction and Prognosis.” *Cancer Informatics*, Jan. 2006.

[6]. C. Catal, B. Diri, “A systematic review of software fault prediction studies” , *Expert Systems with Applications*, Volume 36, Issue 4, 2009, Pages 7346-7354, ISSN 0957-4174.

[7]. V. B. Kumar, S. S. Kumar and V. Saboo, "Dermatological disease detection using image processing and machine learning," *2016 Third International Conference on Artificial Intelligence and Pattern Recognition (AIPR)*, Lodz, pp. 1-6, 2016.

[8]. E. W. Steyerberg, T. van der Ploeg, and B. V. Calster (2014), “Risk prediction with machine learning and regression methods. Biom. J.”, 56: 601-606, 2014.

[9]. D. Dahiwade, G. Patle and E. Meshram, "Designing Disease Prediction Model Using Machine Learning Approach," *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, Erode, India, 2019, pp. 1211-1215.

[10]. Hegazy, Osman & Soliman, S. Omar & A. Salam, Mustafa. (2013). “A Machine Learning Model for Stock Market Prediction”, *International Journal of Computer Science and Telecommunications*. 4. 17-23.

[11]. L. M. B. Alonzo, F. B. Chioson, H. S. Co, N. T. Bugtai and R. G. Baldovino, "A Machine Learning Approach for Coconut Sugar Quality Assessment and Prediction," *2018 IEEE 10th International Conference on Humanoid, Nanotechnology, Information Technology,Communication and Control, Environment and Management (HNICEM)*, Baguio City, Philippines, 2018, pp. 1-4.

[12]. A. H. Haghiabi, A. H. Nasrolahi, A. Parsaie; “Water quality prediction using machine learning methods”, *Water Quality Research Journal 1 February 2018*; 53 (1): 3–13.

[13]. Y. Zhang, J. Liu, Z. Zhang and J. Huang, "Prediction of Daily Smoking Behavior Based on Decision Tree Machine Learning Algorithm," *2019 IEEE 9th International Conference on Electronics Information and Emergency Communication (ICEIEC)*, Beijing, China, 2019, pp. 330-333.

[14]. A. M. Alaa, T. Bolton, E. D. Angelantonio, J. H. F. Rudd, M. van der Schaar. “Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants”, *PLoS One*. 2019;14(5) e0213653. doi:10.1371/journal.pone.0213653. PMID: 31091238; PMCID: PMC6519796.

[15]. H. Zhu, B. Chu, C. Zhang. et al. “Hyperspectral Imaging for Presymptomatic Detection of Tobacco Disease with Successive Projections Algorithm and Machine-learning Classifiers”, *Sci Rep 7*, 4125 (2017).

[16]. X. Zhang, Y. Hu, B. E. Aouizerat et al. “Machine learning selected smoking-associated DNA methylation signatures that predict HIV prognosis and mortality”, *Clin Epigenet 10,* 155 (2018).

[17]. M. A. F. Granero, D. S. Morillo, M. A. L. Gordo, A. Leon (2015) "A Machine Learning Approach to Prediction of Exacerbations of Chronic Obstructive Pulmonary Disease”, *Artificial Computation in Biology and Medicine. IWINAC 2015*. Lecture Notes in Computer Science, vol 9107. Springer, Cham, 2015.

[18]. C. Frank, A. Habach, R. Seetan, A. Wahbeh "Predicting Smoking Status Using Machine Learning Algorithms and Statistical Analysis", *Advances in Science, Technology and Engineering Systems Journal*, vol. 3, no. 2, pp. 184-189 (2018).

[19]. Mary R. Lee, V. Sankar, A. Hammer, W. G. Kennedy, J.J. Barb, P. G. McQueen, L. Leggio, “Using Machine Learning to Classify Individuals With Alcohol Use Disorder Based on Treatment Seeking Status”, *EClinicalMedicine*, Volume 12,2019,Pages 70-78,ISSN 2589-5370.

[20]. S. Kinreich, J. L. Meyers, A. Maron-Katz. et al. “Predicting risk for Alcohol Use Disorder using longitudinal data with multimodal biomarkers and family history: a machine learning study*”. Mol Psychiatry* (2019).

[21]. D. Kumari, S. Kilam, P. Nath. et al. “Prediction of alcohol abused individuals using artificial neural network”. *Int. j. inf. tecnol*. 10, 233–237 (2018).

[22]. M. T. Habib, A. Majumder, R. N. Nandi, F. Ahmed and M. S. Uddin, “Machine Vision Based Papaya Disease Detection,” *Journal of King Saud University – Computer and Information Sciences*, June 2018.

[23]. “What Is Drug Addiction?” 2019. [Online]. Available: https://www.webmd.com/mental-health/addiction/drug-abuse-addiction#2. [Accessed: 07- Aug- 2019].

[24]. "Teen Drug Abuse and Recovery." 2019. [Online]. Available: https://www.nextgenerationvillage.com/drugs/. [Accessed: 07- Aug- 2019].

[25]. "10 Reasons Why People Abuse Drugs." 2019. [Online]. Available: https://www.recoveryconnection.com/10-reasons-people-abuse-drugs/ [Accessed: 07- Aug- 2019].

[26]. "Causes of Drug Addiction - What Causes Drug Addiction?" 2019. [Online]. Available: https://www.healthyplace.com/addictions/drug-addiction/causes-of-drug-addiction-what-causes-drug-addiction [Accessed: 29- Nov- 2019].

[27]. "The Causes and Effects of Drug Addiction." 2019. [Online]. Available: https://www.altamirarecovery.com/drug-addiction/causes-effects-drug-addiction/ [Accessed: 07- Aug- 2019].

[28]. "Anaconda Navigator." 2020. [Online]. Available: https://docs.anaconda.com/anaconda/navigator/ [Accessed: 15- Mar- 2020].

[29]. Jiawei Han, Micheline Kamber, Jian Pei, Data Mining Concept and Technique, 3rd Edition, Morgan Kaufmann, 2012,pp. 332-398.

[30]. Stuart J.Russell, Peter Norvig, Artificial Intelligence a Modern Approach, 3rd Edition, Upper Saddle River, NJ : Prentice Hall, 2010,pp. 725-744.

[31]. Friedman, Jerome H. "Greedy Function Approximation: A Gradient Boosting Machine." The Annals of Statistics, vol. 29, no. 5, 2001, pp. 1189–1232. JSTOR.

APPENDICES

Abbreviation

k-NN = k-nearest neighbors.

SVM = Support Vector Machine.

MLP = Multilayer Perception.

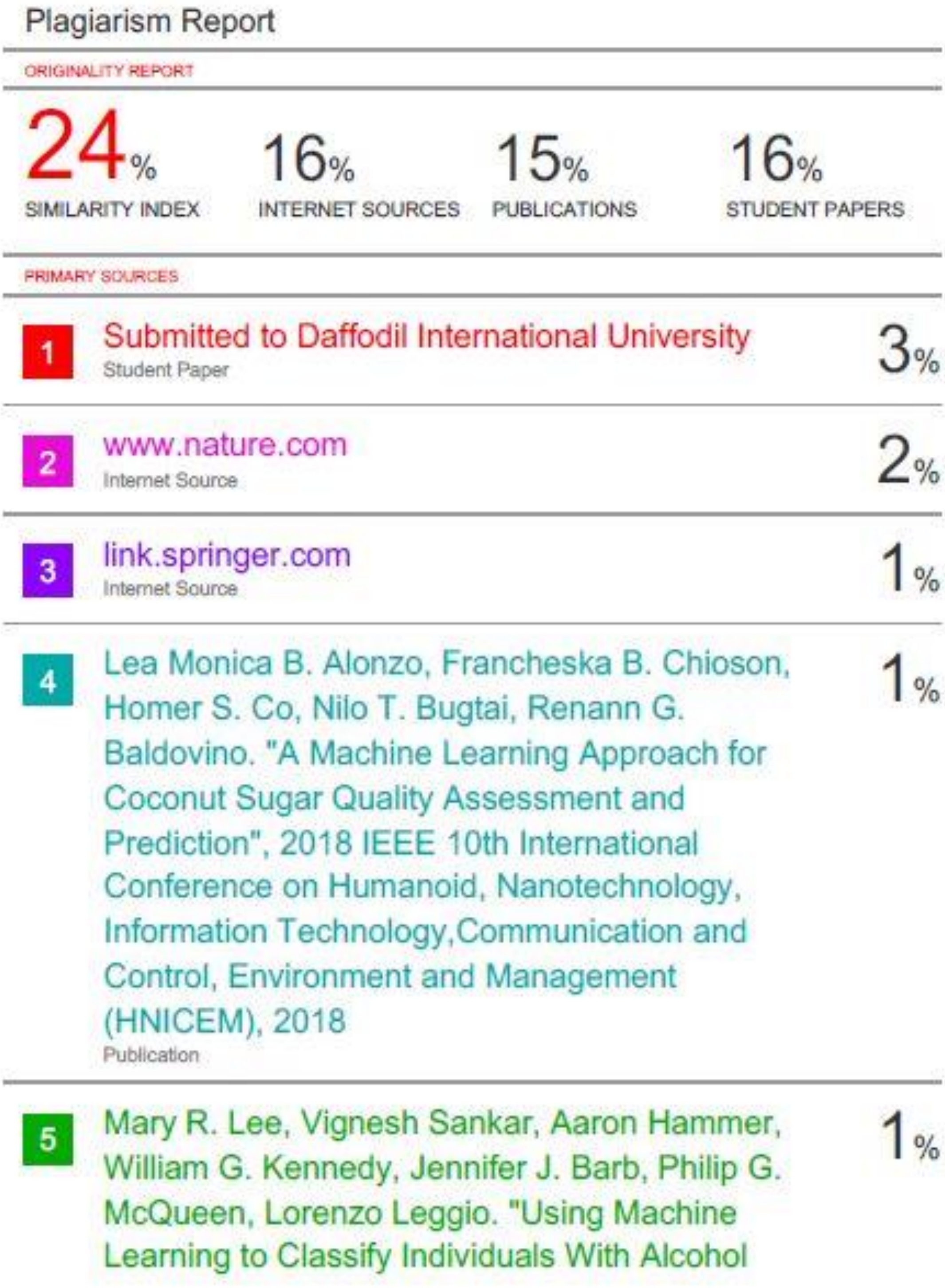
ANN = Artificial Neural Network.

Appendix: Research Reflections

At the beginning of this research work, we had very little idea about machine learning and artificial intelligence detection and prediction. Our supervisor was very kind and sincere. He gave us valuable guidance and helped us a lot. In this whole time of research, we learned many new techniques, learned new information, learned how to use algorithms, how to work with different methods. I also learned about the Anaconda-navigator and Jupyter notebook and Python programming language. Initially, there were problems working with these, but gradually we became more and more familiar with Anaconda-navigator and Jupyter notebook and Python.

Finally, by doing the research we have gained courage and been inspired to do more in the future.

PLAGIARISM REPORT



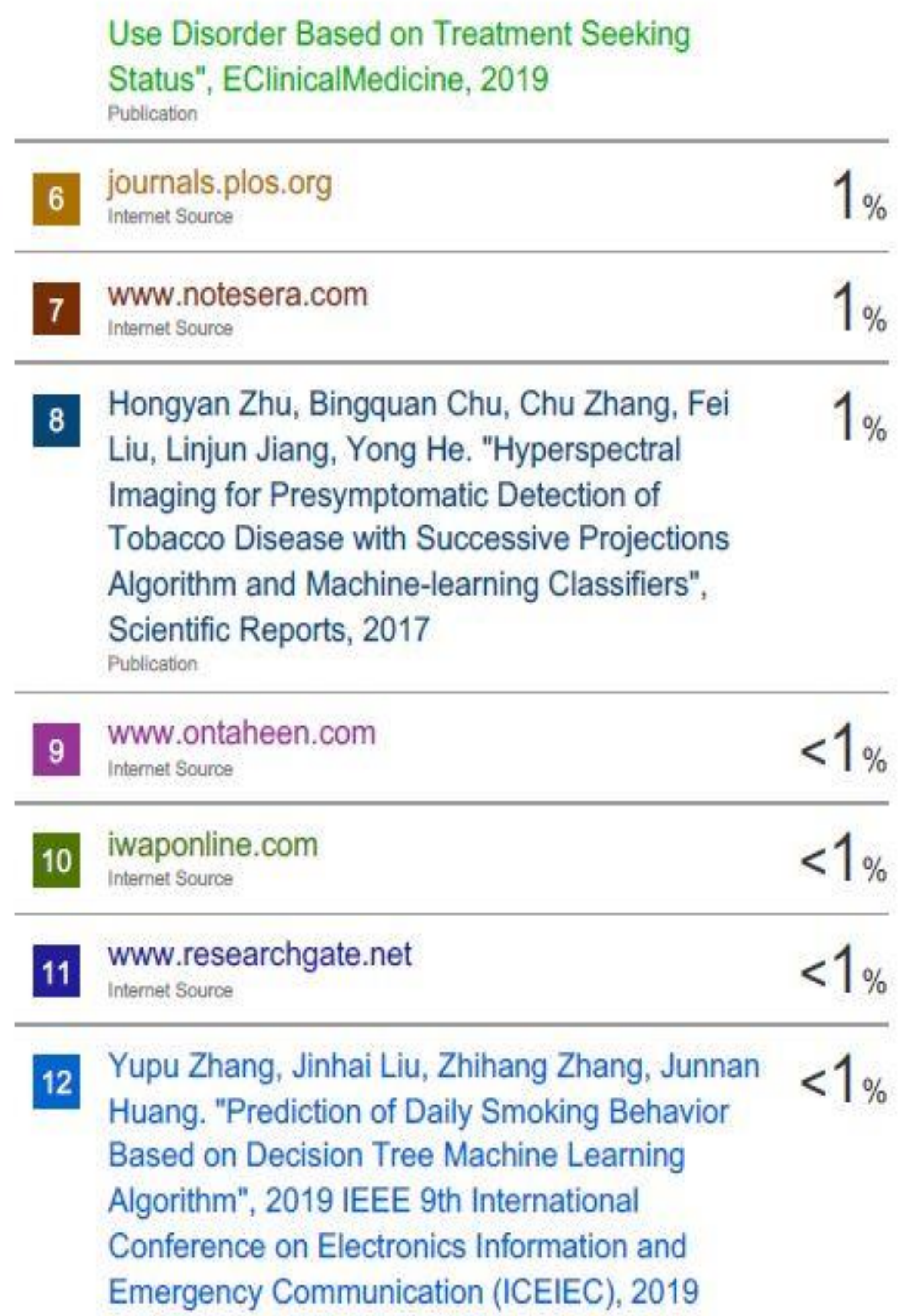
Plagiarism Report

ORIGINALITY REPORT

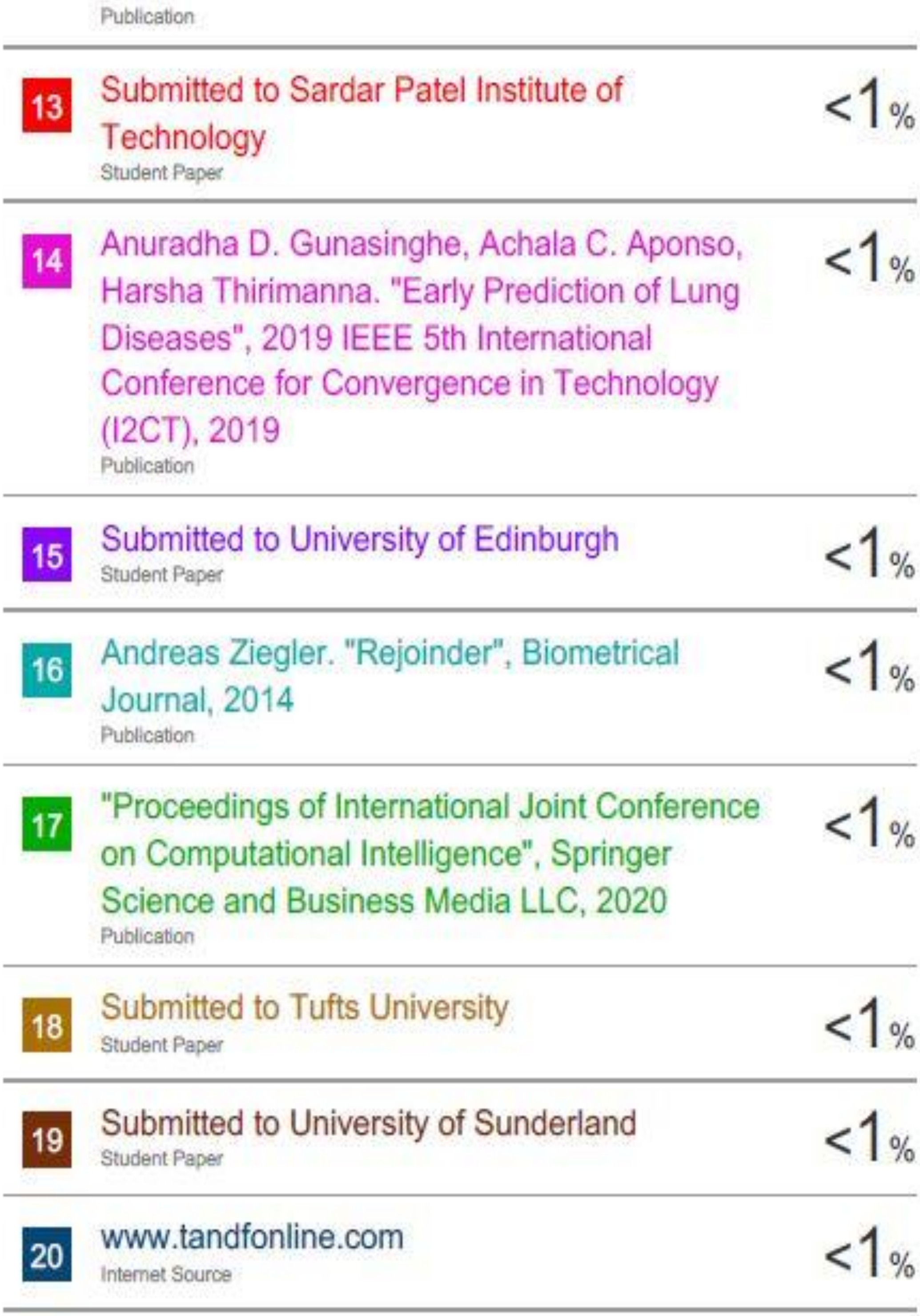
24%	16%	15%	16%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	3%
2	www.nature.com Internet Source	2%
3	link.springer.com Internet Source	1%
4	Lea Monica B. Alonzo, Francheska B. Chioson, Homer S. Co, Nilo T. Bugtai, Renann G. Baldovino. "A Machine Learning Approach for Coconut Sugar Quality Assessment and Prediction", 2018 IEEE 10th International Conference on Humanoid, Nanotechnology, Information Technology,Communication and Control, Environment and Management (HNICEM), 2018 Publication	1%
5	Mary R. Lee, Vignesh Sankar, Aaron Hammer, William G. Kennedy, Jennifer J. Barb, Philip G. McQueen, Lorenzo Leggio. "Using Machine Learning to Classify Individuals With Alcohol	1%



	Use Disorder Based on Treatment Seeking Status", EClinicalMedicine, 2019 Publication	
6	journals.plos.org Internet Source	1%
7	www.notesera.com Internet Source	1%
8	Hongyan Zhu, Bingquan Chu, Chu Zhang, Fei Liu, Linjun Jiang, Yong He. "Hyperspectral Imaging for Presymptomatic Detection of Tobacco Disease with Successive Projections Algorithm and Machine-learning Classifiers", Scientific Reports, 2017 Publication	1%
9	www.ontaheen.com Internet Source	<1%
10	iwaponline.com Internet Source	<1%
11	www.researchgate.net Internet Source	<1%
12	Yupu Zhang, Jinhai Liu, Zhihang Zhang, Junnan Huang. "Prediction of Daily Smoking Behavior Based on Decision Tree Machine Learning Algorithm", 2019 IEEE 9th International Conference on Electronics Information and Emergency Communication (ICEIEC), 2019	<1%



Publication

13	Submitted to Sardar Patel Institute of Technology Student Paper	<1%
14	Anuradha D. Gunasinghe, Achala C. Aponso, Harsha Thirimanna. "Early Prediction of Lung Diseases", 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), 2019 Publication	<1%
15	Submitted to University of Edinburgh Student Paper	<1%
16	Andreas Ziegler. "Rejoinder", Biometrical Journal, 2014 Publication	<1%
17	"Proceedings of International Joint Conference on Computational Intelligence", Springer Science and Business Media LLC, 2020 Publication	<1%
18	Submitted to Tufts University Student Paper	<1%
19	Submitted to University of Sunderland Student Paper	<1%
20	www.tandfonline.com Internet Source	<1%