Early Intervention: A Machine Learning Approach to Classify Drug Addiction

BY

Md. Asgar Ali Manik
ID: 211-15-3981

FINAL YEAR DESIGN PROJECT REPORT

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

Supervised By

Israt Jahan
Lecturer (Senior Scale)
Department of Computer Science and Engineering
Daffodil International University

Co-Supervised By

Ms. Sadia Jannat Mitu
Lecturer
Department of Computer Science and Engineering
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

Dhaka, Bangladesh
January 12, 2025

APPROVAL

This Project titled **Early Intervention: A Machine Learning Approach to Classify Drug Addiction,** submitted by **Md. Asgar Ali Manik,** ID No: **211-15-3981** 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 12-01-2025.

BOARD OF EXAMINERS

12/01/2025

Dr. Md. Taimur Ahad — **Chairman**
Associate Professor & Associate Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Mushfiqur Rahman — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Mr. Rahmatul Kabir Rasel Sarker — **Internal Examiner**
Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Sadat Hasan — **External Examiner**
Data Scientist (Senior Principal Officer)
Risk Management Division
BRAC Bank

©Daffodil International University

DECLARATION

We hereby declare that this project has been done by us under the supervision of **Israt Jahan, Senior Lecturer**, Department of Computer Science and Engineering, 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:

11.01.25

Israt Jahan
Lecturer (Senior Scale)
Department of Computer Science and
Engineering, Daffodil International University

Co-Supervised by:

Sadia 11.01.25

Ms. Sadia Jannat Mitu
Lecturer
Department of Computer Science and
Engineering, Daffodil International University

Submitted by:

Manik 12.01.25

Student Name: Md. Asgar Ali Manik
Student ID: 211-15-3981
Department of Computer Science and
Engineering, Daffodil International University

©Daffodil International University

ACKNOWLEDGEMENTS

This work would not have been possible without the support and contributions of many individuals over the past two semesters. We are deeply grateful to everyone who has assisted us in one way or another.

First, we express our heartfelt thanks and gratefulness to the almighty for His divine blessing making it possible for us to complete the **Final Year Design Project (FYDP)** successfully.

We are grateful and wish our profound indebtedness to **Israt Jahan, Senior Lecturer**, Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh. Deep knowledge and keen interest of our supervisor in the field of **Artificial Intelligence, Computer Vision, Digital Image Processing, Machine learning, Data mining, Natural Language Processing, Internet of Things, Cloud Computing** 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 themat all stages have made it possible to complete this project.

We would like to express our heartfelt gratitude to the Head of the Department of Computer Science and Engineering, for his kind help in finishing our project and also to other faculty members and the staff of the Department of Computer Science and Engineering, Daffodil International University.

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

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

ABSTRACT

Alcohol and drugs are harmful to the body and health. Drug addiction is becoming a menace to the youth of Bangladesh. We will use machine learning to forecast the likelihood of developing a drug addiction according to drugs symptoms. After reading relevant studies, journals, and internet publications and speaking with medical professionals and drug users, we were able to identify a few commonalities in the development of various types of drug addiction class. Then we collect my real data from Divisional Drug Addiction Treatment Centre, Department of Narcotics Control, Rajshahi. almost 21 on those features, such as Age, Gender, Living Situation, Motive of Drug Use, Time Spent Mostly, Failure in Life, Symptoms, Label etc. We collect our data from only addicted people from the agency. 8 classes of drug addicted people data have been collected such as 'Addicted-Heroin', 'Addicted-Alcohol', 'Addicted-Cannabis', 'Addicted-Meth', 'Addicted-Ecstasy', 'Addicted-Prescription Opioids', 'Addicted-Cocaine', 'Addicted-MDMA'. We collected the data, processed it all, and produced a processed dataset. We used machine learning methods on the dataset we had previously processed. Since different prediction and detection systems employ machine learning, artificial intelligence, and deep learning. We employ decision trees, random forests, XG Boost, naïve Bayes, Support Vector Classifier (SVC), and k-nearest neighbor (KNN). Among the six algorithms used in our experiment, decision tree models performed the best in terms of accuracy; the classifier's accuracy was 97.75%. Then create a web application according to the decision tree model based for predict various drug addiction using their symptoms.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
Introduction

1.1 Introduction

The worst social ill in contemporary society is drug addiction. Drug abuse has a direct negative influence on Bangladesh's social and economic landscape, and the drug's menace is only becoming worse. According to those familiar with the issue, the number of drug addicts in Bangladesh in 2018 was over seven million, with over five million of them dependent on "Yaba Pills". Every societal issue has a few characteristics. Addiction to drugs is a serious issue that is linked to societal and familial norms and behavior. Therefore, in order to prevent drug addiction, the government should include both the family and society, as these are the two most effective organizations. We chose this subject because, in our society, drug addiction is a complicated illness. Bangladesh is one of the most populous and least large countries in Asia. India encircles it on three strategically significant sides. Because of the massive mountains that encircle them, the eastern and northern sides are particularly significant. And the plain terrain on the western side is bountiful and fertile. The selling of illegal drugs and drug trafficking are mostly appropriate in the mountainous areas. These mountainous woodlands provide an easy place for traffickers to hide while securely transferring narcotics. However, the rise in drug abusers, the majority of whom are young, has made Bangladesh a very dangerous place for sustainable growth. Thus, one of the biggest obstacles to Bangladesh's progress is the anti-drug backlash.

Drug addiction is the use of drugs for recreational purposes without authorization and the subsequent development of a poisonous and compulsive drug addiction. Addiction to drugs is among the most serious issues facing a nation. It can easily wipe out a nation and a life. Drugs have poisonous consequences that assault and tear a person's body and mind. Drug addiction has firmly established itself among our nation's youth. In a developing nation such as Bangladesh, addiction may have a devastating impact on our culture. The daily star newspaper said that the terrorist group has boosted its manufacturing of drugs and consumption of other illicit substances in Bangladesh [1]. There are over 25 lakh drug addicts. Eighty percent of drug users in Bangladesh are young males and teenagers between the ages of fifteen and forty [2]. The deadly claw of this societal disease has extended to every

corner of the globe. This addiction stems from frustration. Frustration is fueled by issues with unemployment, political upheaval, a loss of family, a lack of love and affection, etc. Once again, social crimes are a result of this addiction. Addicts who are unable to pay for drugs commit a variety of social crimes, such as robbery, murdering, hijacking, looting, and plundering. We therefore need to abstain from drugs in order to prevent drug addiction. Avoiding drugs will only lessen your chance of developing an addiction before you do.

Drug addiction is the use of drugs for recreational purposes without authorization, which results in the dangerous and obsessive dependence on these substances. One of the worst problems a country may have is drug addiction. An estimated 7.5 million individuals in Bangladesh are thought to be drug addicts, according to news reported in the Dhaka Tribune. Of these, 50% are involved in different criminal activities, 80% are young people, and 43 percent of the state's unemployed population is a drug user [3]. Drug addiction is a severe problem that is linked to behavior standards in families, society, and the media. It may easily wipe out a nation and a person's life. Addiction to drugs has taken a solid hold on the younger population in our nation. Over seven million individuals in Bangladesh are drug addicts, with more than five million of these people being hooked to methamphetamine, according to a daily star newspaper article [4]. Drug addiction treatment is a serious problem that affects societal and family dynamics. To avoid and address this issue, it is essential that people, communities, and governments act swiftly and responsibly. So that this is the theme we chose. Most individuals are not aware of the processes that lead to drug addiction.

According to a theory, we may tell whether or not someone abuses drugs by observing their regular social contacts, familial dynamics, and health issues. Abuse of drugs results in social crimes. Aside from wreaking havoc on the family, drug users who are powerless to pay for their addictions also engage in a range of criminal activities, including robberies, killings, hijackings, and looting. Our country's workforce is made up of younger people. Unfortunately, because a large portion of them are extremely young, mostly teens, this group is not seen as a useful work force, but rather as a burden on society. Therefore, in order for the people to resume leading fulfilling lives, they needed our help. Drug abuse is a big problem these days, so we have to be cautious. Drug addiction causes havoc and serious problems for our families, community, nation, and economy. Therefore, parents and the government need to be conscious of this and take appropriate action. Everyone has to be aware of this and avoid being addicted to drugs. As a result, our goal was to develop a model through the use of supervised machine learning.

1.2 Motivation

The problem of drug addiction is intricate and multidimensional, impacting people, families, and civilizations worldwide. Its effects on the body, mind, and society are devastating. A crucial first step in early detection, mitigation, and improved treatment approaches is predicting drug addiction or the probability that a person will become addicted. The issues of managing drug addiction may be addressed, outcomes can be improved, and addiction tendencies can be predicted with the use of machine learning (ML) approaches. We also made the decision to do a study of the subject because of this. My boss also urges me to talk about this subject. As a result, I decided to write a paper with the title "Title." This type of research-based practice was motivated by these factors. Artificial intelligence is crucial since I find myself surrounded by intelligent devices.

1.3 Objectives

Machine learning (ML) is a very successful tool for problem solving. Supervised learning, an artificial intelligence technique, can be used to classify relevant data into drug addiction classes such as "Addicted-Heroin," "Addicted-Alcohol," "Addicted-Cannabis," "Addicted-Meth," "Addicted-Ecstasy," "Addicted-Prescription Opioids," "Addicted-Cocaine," and "Addicted-MDMA." Firstly, use effective models to train the analysis set (labeled data required) and see if any drug addiction category is similar to any of the eight groups. This is guided machine learning's primary objective. Artificial intelligence techniques can be used to identify drug addiction categories. As a result, we were able to identify the following goals:

- Using machine learning to classify each of the eight kinds of drug addiction subcategories.
- To compile information to foresee drug addiction.
- Use drug addiction classifications to categorize addiction.

1.4 Methodology

In the part, the study methodology is discussed, as with guidelines for gathering datasets, conducting each test, and utilizing each model to increase accuracy. This chapter also included recommendations for a methodological and all-data investigation. Therefore, this chapter aims to improve and simplify the facts offered. This study component offers a thorough overview of the complete process while concentrating on the techniques used to

assess drug addiction in statistics. This section will detail the study's whole methodology. Every given analysis has multiple solutions. The next step is to choose an ML approach. Building a repository of information is essential for developing the framework and executing the algorithm since, as we have already seen, we are utilizing six distinct machine learning techniques. The model is then trained using the gathered data. The data were then used to build training and testing sets. There is a common misunderstanding between "training dataset" and "testing dataset." After the input is taken out of the data set that must be created and fitted into various ML technique models, we have just access to a significant amount of the data required to evaluate our model. We then went on and explained everything using our simple process flow diagram.

1.5 Project Outcome

The primary issue with this study seems to be gathering and analyzing all of these little bits of data, given how difficult it was to dig through a single enormous data file. We cleaned and standardized the data collecting process using a number of tools and techniques. Due to the large number of papers and the variety of values they encompassed from different historical eras, it required some time to arrive at the appropriate conclusions. There are more data sets in this field. We never complete the study, therefore always have to put in a lot of effort on any connected tasks, which makes it difficult for me to come up with the best answers quickly. The development of natural language processing (NLP) algorithms for automatically categorizing drug addiction occurrences has garnered significant attention because to the potential advantages of identifying and mitigating such behavior.

1.6 Organization of the Report

The goals, concerns, research questions, and expected outcomes of the study were outlined in Chapter 1. This section also covers the general organization of the report.

All past research in this field is included in Chapter 2. They provide an example of the breadth that results from their narrowing of this research topic in the section that follows. The topic of the last discussion was the primary difficulties or barriers to this investigation. This chapter discusses the difficulties encountered throughout the project's development and contains sections on pertinent research summaries and studies.

A theoretical assessment of the research's conclusions is provided in Chapter 3. More details on the statistical methods particularly those used in the arithmetic section of the

investigation are provided in this chapter. Examples of real-world applications of machine learning techniques are also provided in this section. The next section covers the methods for collecting data and the framework used to assemble it. In the last stage, a single-family confusion evaluation matrix is used to evaluate the model and provide an acceptable tag for identifying the classifier. To ensure true accuracy while using machine learning algorithms, application assessment is required. This section covers the research topic and technique, operational efficiency, data collection plan, data processing, suggested methodology, teaching style, and requirements that must be completed for the project to proceed. This research offers a thorough justification for every machine learning technique and classifier used.

Chapter 4 presents the study's findings, an evaluation of the findings, and a conversation on the implications. A few test pictures are included in this chapter to aid in the project's execution. An review of the results and an application of AI techniques round up this chapter. Describe the web-based application that utilizes drug addiction symptoms to identify the many forms of drug addiction as well.

Chapters 5 and 6 had an outcome, an explanation of the planned course of action, and a research summary. A verified sample proving the report's structure complies with all standards is provided in the next section. Effects on the environment, society in general, and the Sustainable Development Goals The limitations on our job are highlighted in the chapter's conclusion, and these may have an effect on next generations of specialists in our field.

CHAPTER 2
Background

2.1 Introduction

This part usually includes a summary of the research, related functions, and research problems. In "Associated Work," I'll discuss research articles by other writers and compare our technique and accuracy with theirs. In the section on similar studies, I will address the methodologies, validity, and content of other research articles that are relevant to this investigation. I'll provide a synopsis of our related research in the study review section. I decide to go over each issue we encountered while conducting this research and how we improved the accuracy layer in the issues section. Everything was talked about in advance.

2.2 Literature review

This study paper's literature review portion will include recent related studies on drug use and addiction prediction by various scholars. We have observed and examined their work in order to comprehend the thought processes and strategies they have used.

The Decision Tree classification technique was demonstrated in one publication as a means of identifying the structural, physical, and chemical properties of chemicals that make them more likely to result in Adverse Drug Reactions (ADRs) [5]. A presentation of a structure–activity relationship study was made, which included allergic reactions and ADRs in the kidney, liver, and central nervous system (CNS) to categorize medications that may be suspected of causing adverse reactions. Utilizing a machine learning technique in conjunction with a decision tree classification algorithm, they have identified the chemical, physical, and structural components of substances that are predisposed to producing adverse drug reactions.

They learned about drug addiction therapy for both college and non-college students in a different publication. They make predictions in the form of classifications using the Logistic Regression technique.

This, depending on categorization, appropriately extracts features between the two crucial qualities [6]. In this article, they addressed treatment engagement suggestions for higher education. They concluded that the therapy providers seemed to be more successful in

keeping pupils for shorter amounts of time. Numerous studies have highlighted the significance of marijuana and cigarette usage as major risk factors for drug addiction [7]. Writers examined the connection between drugs, alcohol, and cigarettes in a study. They claimed that there is a strong correlation between drug misuse and cigarette smoking, that using cigarettes is widespread among drug users, and that a sizable portion of the questionnaire's questions had to do with smoking cigarettes. Cigarettes are a vital component.

Fahim Faisal and his colleagues at BRAC University conducted research on the topic of "Predicting the public and private life behaviors of a drug abuser in Bangladesh" in another study [8]. Between the ages of 15 and 40, they gathered data on men and women, and then they built a model using 498 examples. Additionally, they detailed how in this paper. We can ascertain the correlation among health problems, social, family, and substance abuse. To help highlight the findings, they employed a dataset that represents the culmination of responses to all 60 characteristics' queries. In actuality, each feature provides a response to a number of queries about a person's personal, social, familial, and health-related lives. Ultimately, the elements that influence people to use drugs have been identified with the use of a machine learning technique and classification algorithm.

Drug addiction is a serious problem in any nation, as I have stated. Alireza Amira badizadeh et al. (2017) employed a decision tree algorithm procedure on data gathered in an Iranian treatment center to comprehend and identify the risk variables for drug use. They discovered that the classification using decision trees approach is useful for these kinds of studies. [9]

Then, in order to estimate the probability of drug intoxication death, Young Jin Choi et al. (2019) [10] compared logistic regression models with other machine learning techniques. The study found that the decision-making tree model performed better. Additionally, logistic regression showed competitive when utilizing medical datasets, as these datasets demand rigorous precision, based on the tuning parameters chosen. [11]

Among the institutions in charge of studying and researching drug-related issues are the Department of Psychiatry and the National Drug and Alcohol Research Center. So, several of the people from these connected departments had done study on the adolescent alcohol usage. The effectiveness of seven machine learning algorithms was examined by Mohamad H Afzali et al. (2020) in order to forecast various amounts of alcohol use in mid-adolescence. According to the article, the best predictive performance is displayed by the elastic-net machine learning algorithm. The prediction performance was found to be

impacted by the baseline alcohol consumption, the sensation seeking personality profile, and the degree of alcohol use. According to the parameters taken into consideration for this research, the article claimed that it performed worse in logistic regression models. [12]

Machine learning has also been used to predict drug addiction in Bangladeshi population-wide data. In 2020, Md. Ariful Islam Arif et al. [13] conducted a comparison study of a number of machine-learning algorithms, including Random Forest, Multi-Layer Perception, K-Nearest Neighbor, Logistic Regression, Support Vector Machine, Naïve Bayes, Decision Tree, ADA Boost Classifier, and Gradient Boosting Classifier. The comparison was based on prediction performance. Based on the data and variables employed, including age, gender, occupation, stress, trauma, and so on, logistic regression outperformed the other nine algorithms in terms of accuracy. This study represents only one of several that Studies on drug addiction datasets in Bangladesh have employed machine learning techniques to forecast whether an individual is drug addicted or not. The results are cross-validated by psychological research that demonstrates the mechanism and severity of drug addiction in young people.

In order to categorize people at risk of addiction based on demographic, psychological, and genetic information, several research have used supervised learning algorithms, including logistic regression, support vector machines (SVM), and decision trees. For example, Duda et al. (2019) employed SVM models for estimating opioid addiction using healthcare records (EHRs), combining behavioral and patient history to achieve high accuracy rates [14]. Similar to this, Lin et al. (2020) created random forest models to find factors that predict drug use relapse, demonstrating how well machine learning handles complicated datasets [15].

Predicting addiction has also been done with neural networks. Recurrent neural networks (RNNs) and convolutional neural network networks (CNNs), two deep learning models, have proven to be particularly effective in spotting temporal patterns in longitudinal data. Using RNNs to examine data streams from wearable devices, Wu et al. (2021) predicted relapse with an accuracy of more than 85% [16]. This demonstrates how machine learning (ML) may incorporate a variety of data sources, such as environmental and physiological signals, into prediction models.

Unsupervised learning methods, such as dimensionality reduction and clustering, have been used to find hidden patterns in data pertaining to addiction. For instance, people have been divided into risk categories using clustering algorithms such as k-means, which have yielded useful information for focused treatments (Smith & Johnson, 2020) [17].

ML in addiction prediction has obstacles despite its potential. Given that biases in training datasets might result in distorted predictions, data availability and quality continue to be major concerns (Luxton et al., 2019) [18]. To guarantee the proper use of ML in this sector, ethical issues including privacy concerns and the possibility of stigmatization must also be taken into account.

2.2.1 Related research

Prediction and detection using machine learning algorithms and data mining techniques have previously been the subject of considerable research. These days, the detection of various diseases, the prediction of alcohol users, and the detection of drug users have all expanded the usage of machine learning technology. This section presents a comparison of these connected works. Here,

Table 2.1: Summary of Literature Reviewed.

SL	Author name	Year	Methodology	Description	Outcome	Limitations
1.	D. Dahiwade, G. Patle et al. [14]	2019	k-nearest neighbors (KNN), CNN.	Method for predicting general diseases based on machine learning.	84.5% accuracy in CNN.	The study's applicability for wider healthcare applications is hampered by its narrow emphasis on a small number of disorders.
2.	Hegazy, Osman & Soliman, S. Omar & A. Salam et al. [15]	2013	Particle Swarm Optimization (PSO), LS-SVM.	Machine learning model for stock market forecasting.	LS-SVM-PSO got the highest accuracy and lowest error than LS-SVM and ANN-BP.	It is less appropriate for live stock market forecasts as it does not take real-time data

						processing into consideration.
3.	L. M. B. Alonzo, F. B. Chioson et al. [16]	2018	MLP-ANN, stochastic gradient descent, KNN, SVM, decision tree, random forest.	Machine learning is used to evaluate and forecast the quality of coconut sugar.	SGD had the best accuracy with 98.3%.	Reproducibility may be impacted by the lack of detail in the data pretreatment stages, such as normalization.
4.	A. H. Haghiabi et al [17]	2018	ANN, SVM, and group method of data handling (GMDH).	Water quality prediction using a machine learning method .	SVM had the lowest DDR error and the highest accuracy.	Seasonal and temporal fluctuations in water quality are not examined in the research, which may have an impact on forecasts.
5.	Y. Zhang et al. [18]	2019	XGBoost, Decision Tree.	utilizing the machine learning technique to forecast daily	84.11% accuracy in the XGBoost decision tree at layer 5.	Model optimization may be impacted by the decision tree algorithm's limited investigation of

				smoking habits		hyperparameter adjustment.
6.	A. M. Alaa et al. [19]	2019	Auto Prognosis.	using machine learning to forecast cardiovascular disease risk in Biobank members.	AutoPrognosis had 0.774 AUC-ROC and increase accuracy.	The application of automated machine learning technologies could restrict comprehension of the fundamental processes involved in decision-making.
7.	H. Zhu, B. Chu et al. [20]	2017	PLS-DA, Random forest, SVM, backpropagation neural network, extreme learning machine, LS-SVM.	Presymptomatic tobacco disease diagnosis using machine learning classifiers and hyperspectral images.	ELM had 98.3% accuracy.	Depends on expensive hyperspectral imaging equipment, which restricts its use for broad agricultural applications.

8.	X. Zhang, Y. Hu et al. [21]	2018	GLMNET, SVM, random forest, XGBoost.	Use a machine learning classifier with smoking-associated DNA to forecast HIV prognosis and death.	With 698 features, the area under the curve was 0.78, and the best model was GLMNET.	Potential biases in DNA methylation profiles brought on by environmental or demographic variables are not addressed in this work.
9.	M. A. F. Granero, D. S. Morillo et al. [22]	2015	Radial basis function neural network, Kmeans, probabilistic neural network.	using machine learning characteristics to forecast obstructive pulmonary disease exacerbations	PNN demonstrated 92.5% specificity, 89.3% accuracy, and 84.1% sensitivity.	Failure to take into account outside variables that might affect COPD exacerbations, such the environment.
10.	C. Frank, A. Habach et al [23]	2018	Naïve Bayes, MLP, logistic regression, J48 and decision table.	Predicting smoking status using statistical analysis and	Logistic regression had 83.44% accuracy, 83% precision, 83.4% recall.	The efficacy of various machine learning algorithms for predicting smoking status

				machine learning		is not compared in this study.
11.	Mary R. Lee, V. Sankar et al. [24]	2019	random forest, random tree, logistic regression and simple logistics model.	Determines whether someone is seeking treatment in order to predict alcohol use disorder with a machine learning classifier	With the simple logistic model, ADT had the highest accuracy.	Missing populations with alcohol use disorders who are not seeking treatment; restricted to those seeking treatment.
12.	S. Kinreich, J. L. Meyers et al. [25]	2019	Regularization method, LASSO.	Making predictions about the likelihood of alcohol use disorder with machine learning	Genetic data and EEG data had better accuracy.	The model's application to people lacking family history records may be limited or biassed due to its dependence on such data.

13.	D. Kumari, S. Kilam et al. [26]	2018	ANN-D, ANN-C.	Using machine learning to forecast alcohol misuse	ANN-D had 98.7% and ANN-C had 49.1% accuracy.	Inadequate information on the feature selection procedure, which may affect the accuracy of the model.
14.	M. T. Habib, A. Majumder et al. [27]	2018	SVM, C4.5, naïve bayes, logistic regression, KNN, random forest, BPN, CPN and RIPPER.	Recognition of papaya illness using a machine learning classification method	SVM got 95.2% accuracy.	In noisy or low-light picture situations, which are frequently seen in agricultural areas, the detecting system may not function properly.

2.3 Gap Analysis

A number of significant obstacles and possibilities are brought to light by the gap analysis of machine learning-based drug addiction diagnosis. Issues with current models include limited and skewed data, difficulties generalizing across varied populations, and a lack of integration across various data types (e.g., biological, psychological). Furthermore, models frequently have issues with transparency and real-time monitoring, which makes it challenging to incorporate them into clinical practice. Significant obstacles are also presented by privacy and ethical issues. To create more reliable and efficient addiction detection systems, there are still ways to improve, like diversifying datasets, integrating real-time and varied information, creating interpretable models, and making sure privacy-preserving methods are used.

Table 2.2: Gap Analysis.

Aspect	Current Issues/Challenges	Opportunities for Improvement
Data Availability	Limited and skewed datasets, often not representative of diverse populations.	Diversify datasets to include varied demographics and regions.
Generalization	Difficulty generalizing models across different populations and contexts.	Develop adaptive models capable of learning from diverse and heterogeneous data.
Data Integration	Lack of integration of biological, psychological, and other relevant data types.	Incorporate multi-modal data sources to improve accuracy and robustness.
Model Transparency	Current models lack transparency, making interpretation and trust in clinical contexts difficult.	Develop interpretable machine learning models to enhance trust and usability in clinical settings.
Real-time Monitoring	Insufficient real-time monitoring and prediction capabilities in current systems.	Implement real-time data collection and analysis mechanisms for timely intervention.
Privacy Concerns	Ethical issues regarding data privacy and patient confidentiality.	Employ privacy-preserving machine learning techniques, such as federated learning and differential privacy.
Clinical Integration	Models are not easily integrated into clinical workflows due to lack of usability and adaptability.	Design systems with user-friendly interfaces and compatibility with existing clinical tools and protocols.
Ethical Considerations	Insufficient focus on the ethical implications of automated predictions in sensitive areas like addiction diagnosis.	Establish guidelines and frameworks to ensure ethical use and decision-making.

2.4 Summary

This chapter contains all of the previous studies conducted in this area. In the next section, they give an illustration of the breadth that arises from their refining this study topic. The main challenges or obstacles to this inquiry were the subject of the most recent conversation. This chapter includes parts on relevant research summaries and studies and talks about the challenges faced during the project's development.

CHAPTER 3

Research Methodology

3.1 Methodology/Requirement Analysis & Design Specification

This chapter explains the requirements analysis and design specification for putting the research framework into practice, as well as the technique utilised to analyse and forecast drug addiction. It offers a methodical way to gather, prepare, and evaluate data as well as apply machine learning (ML) techniques to provide precise forecasts.

3.1.1 Overview

The study technique is covered in the section that follows, along with instructions for compiling datasets, running each test, and making use of each model to improve accuracy. Additionally, suggestions for a methodology and all-data study were presented in this chapter. Thus, in an attempt to make the data presented in this chapter better and simpler. This research component focuses on the methods utilized to evaluate drug addiction in statistics in addition to providing a detailed explanation of the entire procedure. The study's whole methodology will be covered in this part. There are several ways to solve every given analysis. The selection of an ML strategy is the next stage. Since we are using six different machine learning approaches, as we have already demonstrated, building a data store is a necessity for creating the framework and running the algorithm. The collected data is subsequently used to train the model. Following that, sets for training and testing were created using the data. People frequently confuse the terms "training dataset" and "testing dataset." We only have access to a sizable portion of the data needed to assess our model once the input is extracted from the data set that needs to be generated and fitted into different ML method models. We proceeded to use our basic process flow diagram to explain things.

3.1.2 Proposed Methodology

To accomplish its objectives, this research probably used a number of different approaches or techniques. The written content and all values had to be cleaned up, the workflow had to be chosen, the medicine labels had to be collected and manipulated, and the classifier's efficacy

had to be assessed using the results of a forest classification technique that was chosen at random.

Step 1: Gathering Data: Our entire dataset is sourced from **Divisional Drug Addiction Treatment Centre, Department of Narcotics Control, Rajshahi**. This whole real-time data set is used to forecast drug addiction. This company lacks a large, comprehensive dataset since it is difficult to collect data for the specific addicted substance of high-quality analyzers and classification type.

Step 2: Data processing: Each item of information was looked at separately once all practical means of obtaining data had been used. There are many instances of bad and ambiguous language all around us. It is advised that we go over the final piece of the dataset before utilizing it.

Step 3: Data prepare: The "Symptoms" and "Label" continue to direct the development and processing of the data for prediction even after the dataset has been put together. Training requires sorting the data, eliminating null values, and presenting it. The data has not been sufficiently prepared for separation.

Step 4: Model Selection: We decide on a prediction approach, train it using my data, and then assess it to increase reliability. In the field of machine learning, several filters are used. Even though several designs were used to enhance the component design and enable the machine learning model to detect the kind of drug addiction, only one instrument was ultimately selected to evaluate the data's dependability.

Step 5: Assessment of Performance: This phase's later sections address all the ramifications. These methodologies gave us a limited degree of consistency for the label groups of the eight distinct medication datasets after the training and evaluation phase. Accuracy data and f1 scores were generated to support the confusion matrices. This section provides a description of each result. These tactics did not provide us with sufficient dependability for the next two courses, even after testing and training. They produced a method for categorizing various therapeutic plants as well as visual aids for f1 measurement, recall, efficiency, and confusion matrix.

Step 6: Conclusion and Upcoming Projects: There will be a summary of this field's development plan given.

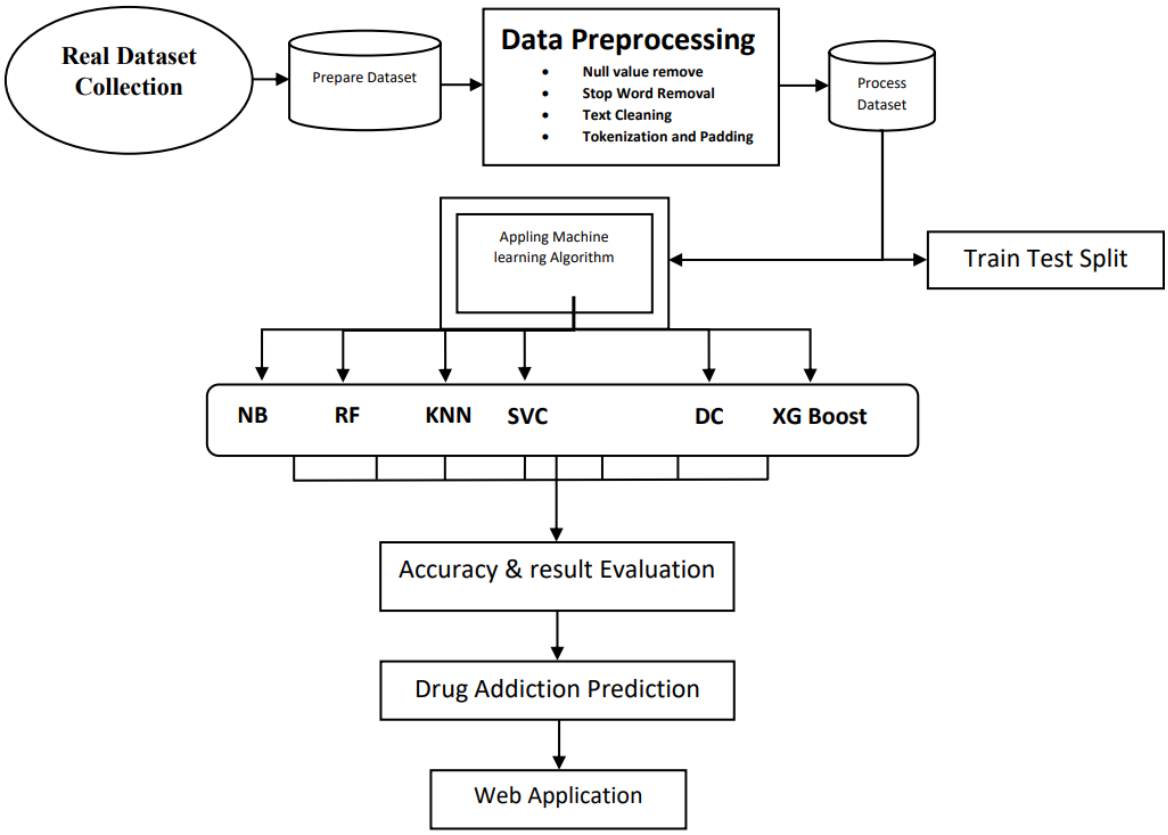


Fig 3.1: Entire Proposed model.

The stages of our study technique are depicted in Figure 3.1, which may be used to identify the specific drug addiction based on the symptoms of the patient. I use the "place name" to gather data in real time. The dataset was constructed utilizing all the necessary characteristics of drug addicts. We have reviewed the process used to gather the data, removed any unnecessary null values, and cleared up the language to ensure that the complete data collection only contains accurate and relevant information. We develop and improve models to investigate machine learning methods using previously collected data. We also utilize permutation approaches to address the issue of class heterogeneity, which will guarantee fair inclusion and increase the overall efficacy of the model. Our technique not only finds an overview, but also looks for ways to reduce the quantity of imprecise and wrong findings. By merging language comprehension approaches with machine learning technology, we may offer

an integrated solution that regulates the results in the drug addiction figures that are used to predict specific drug names based on their symptoms.

3.1.3 Implementation Requirements

An area of research is a field of study that is scrutinized and studied to elucidate concepts for overseeing modeling, data gathering, task fulfillment, and proving education, in addition to execution. As measurement experts, we talk about the instruments and methods we use. We used Python as our programming language, the company's Windows operating system, and a few other tools including NumPy, Scikit Learn, and OpenCV. Using GPU for run the program. The preferred platform for all teaching and evaluation was Google Colab. Python programmers may write code for data science and machine learning methods using Google Colab. In order to determine clusters, or the eight categories of drug addiction type categorization, these methods make use of machine learning-related statistical approaches.

Libraries used:

- **Matplotlib:** Plotting, scoring, organizing, and graphing are made easier with the Pyplot set of Matplotlib tools. This might be applied to form-building to draw attention to certain narrative points of view or the story's boundaries of plausibility.
- **NumPy:** The language's NumPy library has simplified vector processing. This topic includes detailed descriptions of matrix calculations, conversion indices, and the inverted transform of a Fourier transform. The NumPy Python packages include a number of tools and techniques for working with different kinds of matrices. The process of creating gadgets may be made more rational and useful with the aid of NumPy. NumPy is a Python module developed largely for numerical evaluation, to demonstrate it succinctly. You may also use "estimates. Py" to represent this.
- **Scikit-learn**: Scikit-learn is a powerful and intuitive tool for data modeling and analysis. Three Python utilities were used in the layout: NumPy, SciPy, and Matplotlib. These are open-source, publicly accessible tools that anybody may use.
- **Seaborn:** Matplotlib will be included in the upcoming edition of this popular Python data visualization application. This is a simple-to-use application for artistic data visualization.
- **H5py:** The Python h5py module allows users to access fragmented HDF5 code. After a significant amount of the data mostly integers is handled by NumPy, HDF5 is used for storage.

- **Pandas:** This freeware toolkit for examining and working with data that is particular to a language is provided by Pandas. Basic data kinds and statistical analysis methods can help ensure orderly data administration, particularly when working with summary data.
- **OS:** The Py System element offers a range of capabilities that enable staff members to communicate with one other using the same vernacular.

To perform in-depth research, the hardware, software, and data resources used in this setup for **Early Intervention: A Machine Learning Approach to Classify Drug Addiction.** have been carefully chosen. High-performance computers are computing systems with a computational core made up of powerful GPUs and a sturdy CPU. I am aware that no innovation can guarantee flawless outcomes. In a manner similar to that, we may modify the parameters of our model during training to improve accuracy.

3.2 Detailed Methodology & Design

The phases of our research methodology, which can be utilized to pinpoint a patient's particular drug addiction based on their symptoms. The "place name" is what I use to collect data instantly. All of the essential traits of drug addicts were used in the construction of the dataset. To make sure that the entire data collection only includes correct and pertinent information, we have examined the procedure used to collect the data, eliminated any superfluous null values, and clarified the terminology. Using previously gathered data, we create and refine models to study machine learning techniques. In order to ensure equitable inclusion and boost the model's overall effectiveness, we additionally employ permutation techniques to address the problem of class heterogeneity. In addition to obtaining an overview, our method searches for methods to minimize the number of inaccurate and imprecise results. We may provide an integrated solution that controls the outcomes in drug addiction statistics that are used to anticipate certain drug names based on their symptoms by combining machine learning technology with language comprehension techniques.

Data Collection

The data set is a massively scalable collection of relevant and easily accessible coordinates. Initially, we attempt to locate drug users in our area and other locations. However, we observed drug usage going on around us, albeit it was kept a secret, and drug users at the

bus and rail stations declined to provide assistance. Next, we made the decision to visit a drug addiction and recovery facility.

I collect my real time data from **Divisional Drug Addiction Treatment Centre, Department of Narcotics Control, Rajshahi**. Only addicted people drug related data have been collected. Also validate data help of **Md. Sheikh Robiul Akhter**. About 21 features have been collected for my resource purpose. Total 4000 data have been collected for various drug addicted people. 8 classes of drug addict people data collected in this dataset.

All of the data items found in each of these files are shown in Table 3.1, which organizes them into 21 main categories:

Table 3.1: Columns of Description in the Dataset.

Column's Name	**Description of the Column's**
Age	Age of drug addicted people
Gender	Male, Female
Living situation	Mention living alone, with family, friends etc.
Motivate of drug use	Motivation for the reason of drug use
Time spent mostly	Where to time spent drug addicted people like: At home, online gaming etc.
Failure in life	Yes, No.
Mental/Emotional Problems	Yes, No.
Suicidal Thoughts	Yes, No.
Family relationship	Good, Poor, Strained
Financial Status	High, Medium, Low
Addict in family	Yes, No.
Satisfied with workplace	Yes, No.

Case in Court	Yes, No.
Living with Drug user	Yes, No.
Smoking	Yes, No.
Drug Use	Yes
Drug Type	Heroin, Cocaine, Alcohol, MDMA, Cannabis, Prescription opioids, Meth, Ecstasy
Control Over use	Yes, No.
Level of drug use	Very high, high, moderate, low
Symptoms	Symptoms of taking drug
Label	'Addicted-Heroin','Addicted-Alcohol','Addicted-Cannabis','Addicted-Meth', 'Addicted-Ecstasy', 'Addicted-Prescription Opioids', 'Addicted-Cocaine', 'Addicted-MDMA'

Table 3.2: Raw data sample data.

Age	18	18	18	18	18
Gender	Male	Male	Male	Male	Male
Living Situation	Living with family	Alone	With friends	Living with parents	Living with family

Addict in Family	Financial Status	Family Relationship	Suicidal Thoughts	Mental/Emotional Problems	Failure in Life	Time Spent Mostly	Motive of Drug Use
Yes	Low	Strained	Yes	Yes	Yes	Friends' hangouts	Stress relief
No	Medium	Good	No	Yes	No	Clubs and parties	Social acceptance
Yes	Low	Poor	No	No	Yes	Street corners	Curiosity
Yes	High	Strained	Yes	Yes	Yes	Home	Escape from reality
No	Medium	Good	No	No	No	At home	Experimentation

Level of Drug Use	Control Over Use	Drug Type	Drug Use	Smoking	Living with Drug User	Case in Court	Satisfied with Workplace
High	No	Heroin	Yes	Yes	Yes	No	Yes
Moderate	Yes	Alcohol	Yes	Yes	Yes	No	No
Low	Yes	Cannabis	Yes	Yes	Yes	Yes	Yes
Very High	No	Meth	Yes	Yes	Yes	Yes	No
Moderate	Yes	Ecstasy	Yes	Yes	No	No	Yes

Symptoms	Euphoria, Nausea, Anxiety	Slurred speech, Coordination	Relaxation, Impaired memory	Insomnia, Paranoia	Euphoria, Increased energy
Label	Addicted-Heroin	Addicted-Alcohol	Addicted-Cannabis	Addicted-Meth	Addicted-Ecstasy

Age	Gender	Living Situ	Motive of	Time Sper	Failure in	Mental/Er	Suicidal Th	Family Re	Financial S	Addict in F	Satisfied v	Case in Cc	Living witl	Smoking	Drug Use	Drug Type	Control Ov	Level of D	Symptom	Label
18	Male	Living witl	Stress reli	Friends' h	Yes	Yes	Yes	Strained	Low	Yes	Yes	No	Yes	Yes	Yes	Heroin	No	High	Euphoria,	Addicted-
18	Male	Alone	Social acc	Clubs and	No	Yes	No	Good	Medium	No	No	No	No	Yes	Yes	Alcohol	Yes	Moderate	Slurred sp	Addicted-
18	Male	With frien	Curiosity	Street cor	Yes	No	No	Poor	Low	Yes	Yes	Yes	Yes	No	Yes	Cannabis	Yes	Low	Relaxatior	Addicted-
18	Male	Living witl	Escape fro	Home	Yes	Yes	Yes	Strained	High	Yes	No	Yes	Yes	Yes	Yes	Meth	No	Very High	Insomnia,	Addicted-
18	Male	Living witl	Experimer	At home	No	No	No	Good	Medium	No	Yes	No	No	No	Yes	Ecstasy	Yes	Moderate	Euphoria,	Addicted-
18	Male	With partr	Pain mana	At home	Yes	Yes	Yes	Strained	Low	Yes	No	No	Yes	No	Yes	Prescripti	No	High	Drowsine	Addicted-
18	Male	Living witl	Coping me	Social gatl	Yes	Yes	Yes	Poor	Medium	Yes	No	Yes	Yes	Yes	Yes	Cocaine	No	High	Increased	Addicted-
18	Male	Alone	Boredom	At home	No	No	No	Good	High	No	Yes	No	No	No	Yes	Alcohol	Yes	Moderate	Slurred sp	Addicted-
18	Male	Living witl	Peer pres	Clubs and	Yes	Yes	Yes	Poor	Low	Yes	Yes	Yes	Yes	Yes	Yes	MDMA	Yes	High	Euphoria,	Addicted-
18	Male	With fami	Escape fro	Online ga	Yes	Yes	Yes	Strained	Medium	Yes	No	Yes	Yes	No	Yes	Heroin	No	Very High	Intense cr	Addicted-
18	Male	Alone	Experimer	At home	No	No	No	Good	High	No	Yes	No	No	No	Yes	Cannabis	Yes	Moderate	Relaxatior	Addicted-
18	Male	Living witl	Stress reli	Friends' h	Yes	Yes	Yes	Strained	Low	Yes	No	Yes	Yes	No	Yes	Meth	No	Very High	Insomnia,	Addicted-
18	Male	Living witl	Social acc	Clubs and	No	Yes	No	Good	Medium	Yes	Yes	No	No	Yes	Yes	Cocaine	Yes	High	Increased	Addicted-
18	Male	With frien	Peer pres	Street cor	Yes	No	Yes	Poor	Low	Yes	Yes	Yes	Yes	Yes	Yes	Ecstasy	No	High	Euphoria,	Addicted-
18	Male	Living witl	Curiosity	At home	No	No	No	Good	Medium	No	No	No	No	No	Yes	Heroin	Yes	Moderate	Euphoria,	Addicted-
18	Male	Alone	Pain relie	At home	Yes	Yes	Yes	Poor	Low	Yes	No	Yes	Yes	Yes	Yes	Prescripti	No	High	Drowsine	Addicted-
18	Male	With partr	Escape fro	Social gatl	Yes	Yes	Yes	Strained	Medium	Yes	Yes	No	Yes	No	Yes	Meth	No	Very High	Insomnia,	Addicted-
18	Male	Living witl	Social acc	Clubs and	No	Yes	No	Good	High	No	No	No	No	Yes	Yes	Alcohol	Yes	Moderate	Slurred sp	Addicted-
18	Male	Living witl	Curiosity	Street cor	Yes	No	No	Good	Medium	Yes	Yes	Yes	Yes	Yes	Yes	Cocaine	Yes	High	Increased	Addicted-
18	Male	Alone	Boredom	Online ga	No	No	No	Poor	Low	Yes	No	No	No	No	Yes	Cannabis	Yes	Moderate	Relaxatior	Addicted-
18	Male	Living witl	Stress reli	Friends' h	Yes	Yes	Yes	Strained	Low	Yes	No	Yes	Yes	Yes	Yes	Heroin	No	High	Euphoria,	Addicted-
18	Male	Alone	Social acc	Clubs and	No	Yes	No	Good	Medium	No	Yes	No	No	Yes	Yes	Alcohol	Yes	Moderate	Slurred sp	Addicted-
18	Male	Living witl	Curiosity	At home	Yes	No	Yes	Poor	Low	Yes	Yes	Yes	Yes	Yes	Yes	MDMA	Yes	High	Euphoria,	Addicted-
18	Male	With partr	Pain mana	At home	Yes	Yes	Yes	Strained	Low	Yes	No	No	Yes	No	Yes	Prescripti	No	High	Drowsine	Addicted-

Fig 3.2: Raw data sample data.

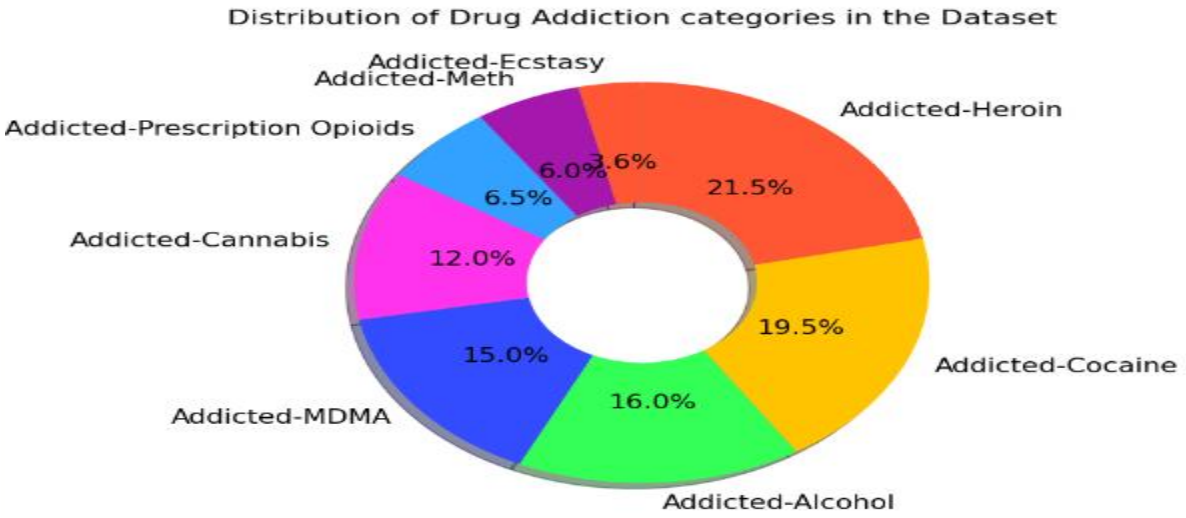


Fig 3.3: Drug Addiction categories in dataset.

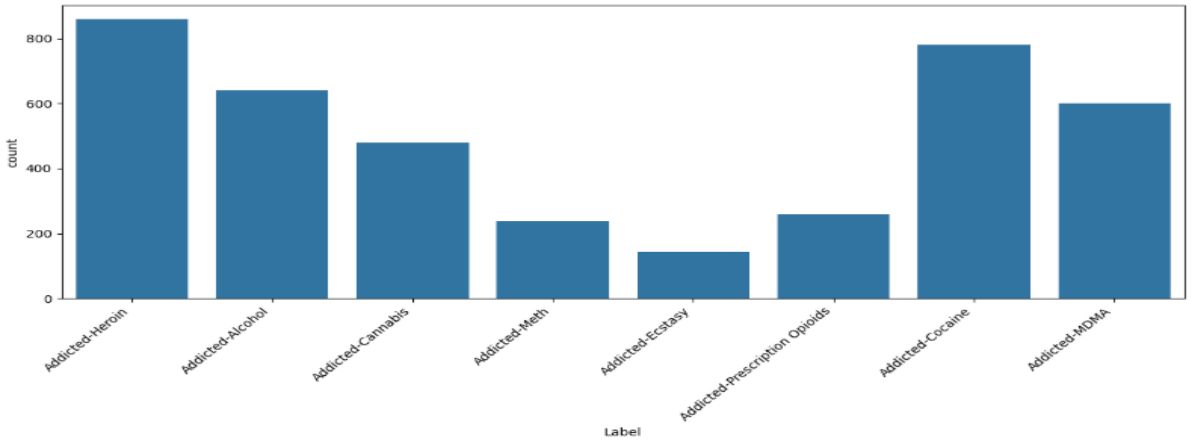


Fig 3.4: 8 classes of drug data contents.

Analyzing the Data

Obtaining data on the drug-related categories of addiction is a precondition for obtaining the required "Symptoms" and "Label" characteristics in the first stage of data processing. Data that may be utilized for testing and training can be obtained by combining many datasets. In order to begin fixing any mistakes, we began by removing any extraneous characters and symbols from the database's contents that had null values. Another method that transforms words into vectors of numbers that may be easily concatenated and utilized in machine learning models is tokenization.

Data Cleaning and Null value remove

Modifying numerical values is an essential stage in the creation of datasets. The two main strategies used in this strategy are meant to raise the standard and relevance of written content. News is edited before it is released, and it is only kept for a predetermined amount of words. Filtering procedures protect our commitment to offering instructive, legally-compliant, and high-quality content. To update the information gradually, extraneous symbols such as groups, stop categories, emoji removers, special signs, etc. are also removed using a text correction technique. There are no null value and 143 duplicates value found, both are cleaned. We use a meticulous preservation process to ensure that all of the data we gather is ready for a detailed examination. Each of the eight categories of drug addiction is represented by three rows, one for each statistic.

	0
Age	0
Gender	0
Living Situation	0
Motive of Drug Use	0
Time Spent Mostly	0
Failure in Life	0
Mental/Emotional Problems	0
Suicidal Thoughts	0
Family Relationship	0
Financial Status	0
Addict in Family	0
Satisfied with Workplace	0
Case in Court	0
Living with Drug User	0
Smoking	0
Drug Use	0
Drug Type	0
Control Over Use	0
Level of Drug Use	0
Symptoms	0
Label	0

Fig 3.5: After remove null & duplicate data.

Word Cloud

Words with great visual impact are important for our training and testing datasets as well as for our analysis, as Figs. 3.6 and 3.7 show. By giving a visual representation of the major phrases used in the report's content, this technique facilitates the discovery of themes and patterns in the language used. Significant insights on implicit and distinguishing characteristics can be gained. To help guide the process of producing our data, the visual is a helpful tool. The ensuing offers a genuine picture of explicit language growth in English as well as enabling it. With this simple technique, we can connect the dots between raw data and meaningful analysis, which aids in our decision-making process while creating a false-positive analysis strategy.

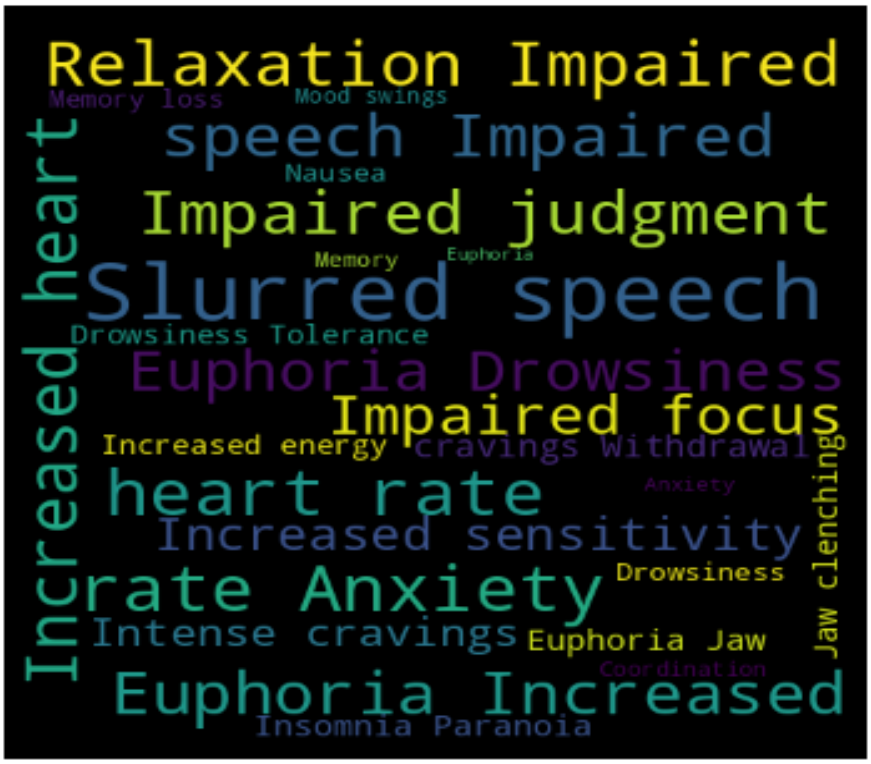


Fig 3.6: Train Data of Word Visualization.

Fig 3.7: Test Data of Word Visualization.

Stop Word Removal

When the content was enhanced, the language employed was acceptable. Even though they are commonly used, the phrases mentioned above usually have no meaning in the categorization system or in linguistic context. We utilize stop words in conjunction with a specially made custom "Stop words" module that we bought from GitHub to overcome this problem. The vast array of English numbers in the library has been painstakingly maintained to conform to the syntactic structure of our language. We might reduce the amount of unnecessary characters and speed up data processing by merging these libraries. We frequently eliminate punctuation to make our text categorization efforts more consistent, as the following examples demonstrate:

"if", "its", "Im", "no", "by", "at", "what" etc.

```
# Define a function to remove specified words (stop words) from a sentence
def remove_stop_words_and_single_alphabets(sentence):
    words = sentence.split()
    stop_words = ["it", "ve", "!","re", "me", "the", "oh", "we", "you","so", "he", "is", "." ,
    "she", "like", "they", "them", "my", "if", "to", "in", "im", "am", "on", "ok", "uh", "rt", "at",
    "us", "of",  "But", "over", "come", "real", "saying", "than", "off", "tell", "yours", "same", "any",
    "world", "back", "this", "if", "let", "mkr", "been", "thing", "should", "anything",
    "did", "its", "day", "still", "first", "too", "cant", "And", "had", "going", "make",
    "these", "only", "see", "has", "go", "why", "were", "there", "will", "because", "how",
    "the", "then", "an", "he", "if", "its", "Im", "no", "by", "at", "what", "u", "do",
    "amp", "i", "or", "so", "have", "be", "my", "who", "was", "are", "I", "to", "a", "the"]
    words = [word for word in words if len(word) > 1 and word.lower() not in stop_words]  # Remove single alphabets and stop words
    return ' '.join(words)
```

Fig 3.8: Custom English Stop Words.

Text Cleaning

One key aspect in the construction of the dataset is that we conduct text modification. This approach consists of two primary techniques meant to improve the caliber and pertinence of written material. Before being published, news is screened, and pieces are only retained for a predetermined 100 words. Filtering methods safeguard our promise to provide high-quality, legally-compliant, and educational content. A text correction approach is also applied to remove unnecessary symbols such as sections, stop categories, emoji removers, special markings, etc. in order to change the content in a methodical manner. The text has been edited to remove several standard symbols, line breaks, and English characters. We employ a comprehensive preservation procedure to guarantee that everything we collect is prepared for a close inspection.

Tokenization and Padding

Tokenization and padding are two fundamental components of contemporary data use. Words have to be tokenized, or translated into numerical sequences, in order for our system to understand English. Attributing a distinct number to every word establishes the link between the written word and symbolic numerals. Padding makes assuring that each sequence has a predetermined duration within an arranged training session, which raises the possibility of agreement. Our computers convert phrases into dollars and cents and verify that there is enough variation in the word lengths overall in order to assess the symptoms model. This stage is crucial in order for our computer systems to accurately assess the data and distinguish between the eight main categories that are represented in the drug addiction symptoms and label , as shown in Figure 3.9 [36].

Fig 3.9: Example of Tokenization.

Data Preparation

Even after removing duplicate data and null values, we did not randomly split the data to test the model and training during the data preparation step. We take the "Symptoms" and "Label" variables that are most important out of the 4,000 entries in the drug addiction dataset. The dataset is divided into two sections: Train and Test. There are 3200 data in the train and 800 data in the test.

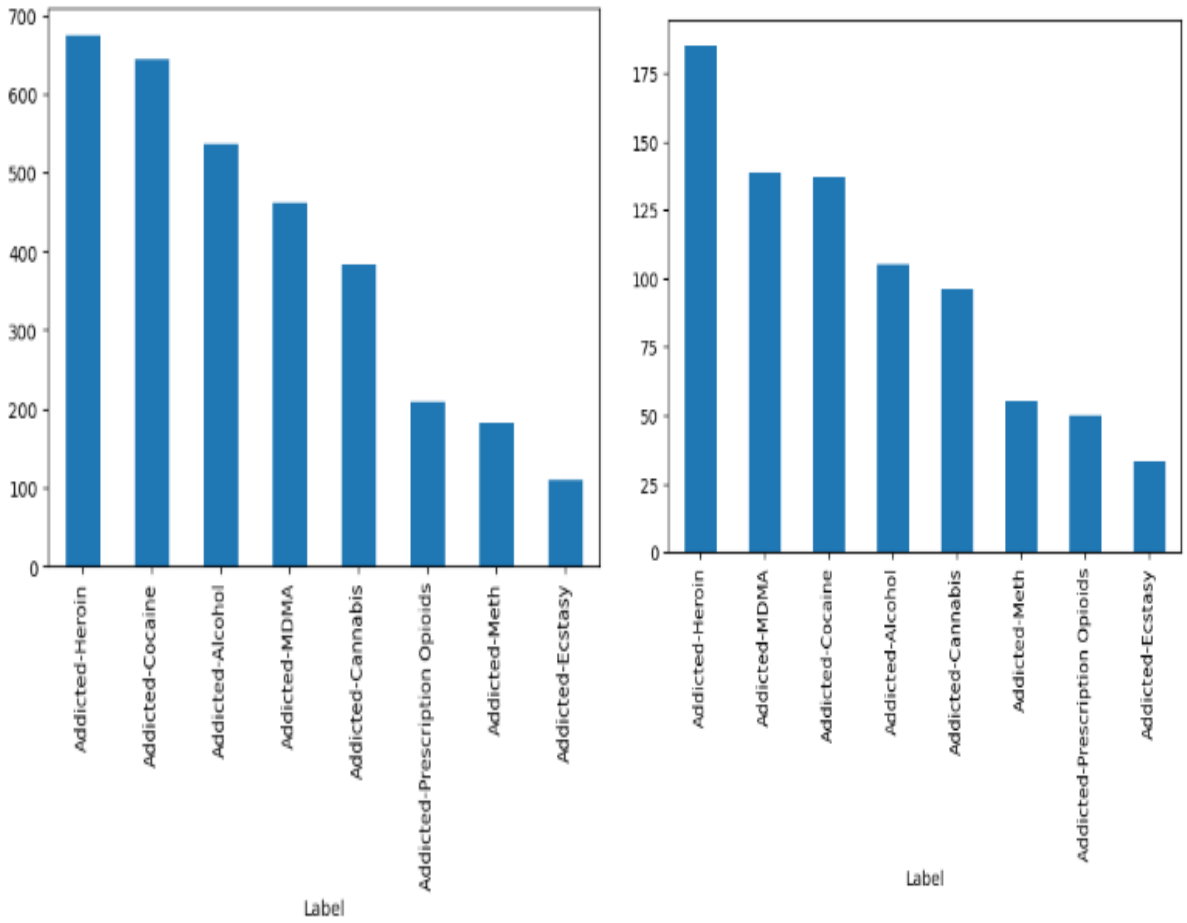


Fig 3.10: Drug addiction Data Count for Train and Test.

Statistical Analysis

There are 21 features or columns in this dataset. So here below fig 3.11. shows count plot of each feature.

Fig 3.11: All the features data count.

Here all columns has their own class shows as count plot. This fig 3.10. shows all count of each columns.

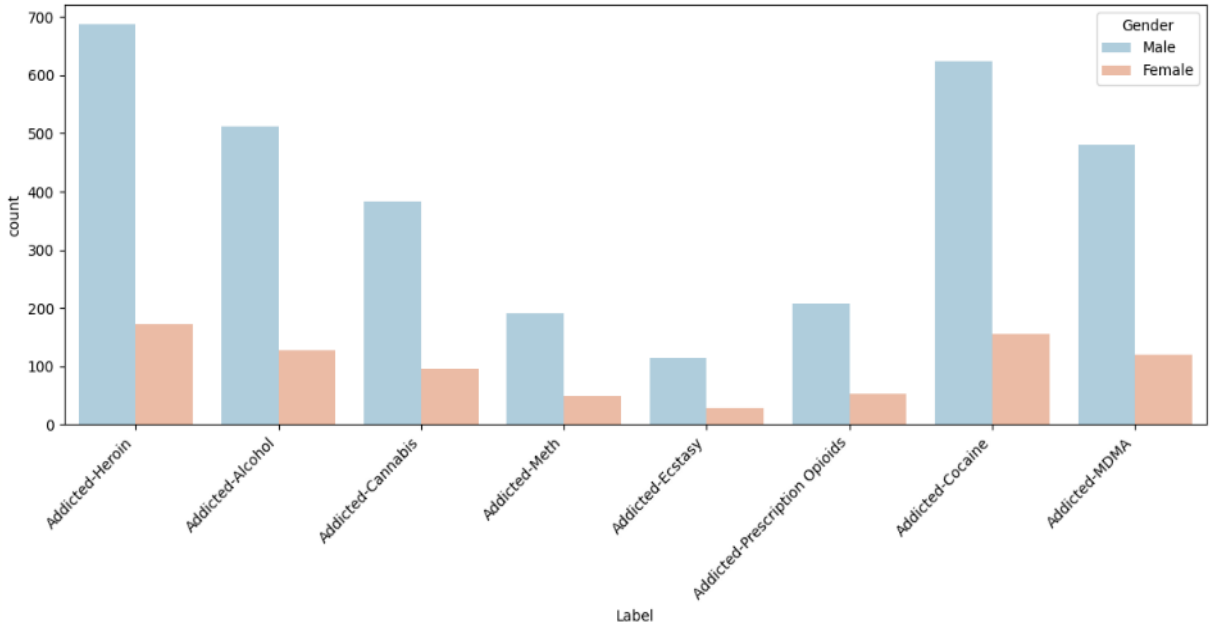


Fig 3.12: Drug addiction-based gender feature analysis.

The fig 3.12 shows the number of people addicted to many kinds of drugs. Two types of Gender data use in this work like: Male& Female. According to the graph blue shows the number of male and another shows the number of females. Most of the male are addicted to heroin and it's nearly 700. Female are addicted to heroin around 200.Lowest rate of addiction sector is Ecstasy and it's less than 100.

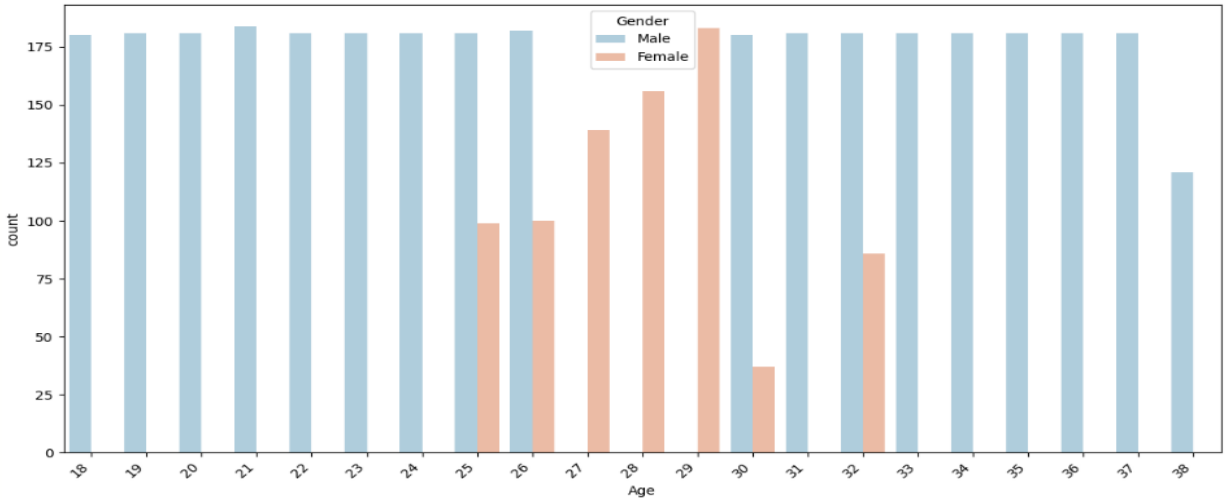


Fig 3.13: Age drug addiction-based gender feature analysis.

From this fig 3.13 we can see the age of male and female taking drugs. Two types of Gender data use in this work like: Male & Female and 18-38 years age people data used for analysis. All ages male are taking drugs but from 25 aged female's number is increasing. 29 aged females are taking drugs highly and it's 175.

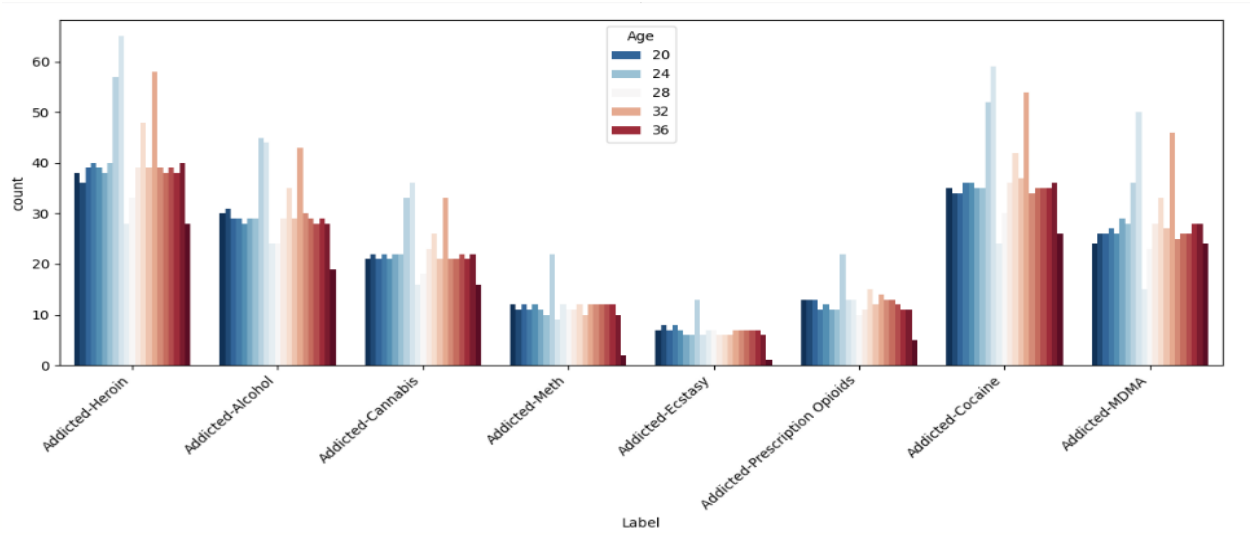


Fig 3.14: Drug addiction-based age feature analysis.

In this fig 3.14 shows that different age based specified drug addiction people taken drug different age. Here we saw 18-38 aged drug addicted people data have been collected. Here we analysis 8 classes of drugs addicted with different ages. Addicted-Heroin people who ahs age 24-28 people are the most taken of drug count almost 70 people. Addicted-Ecstasy people get less than of allover 18-38 age people.

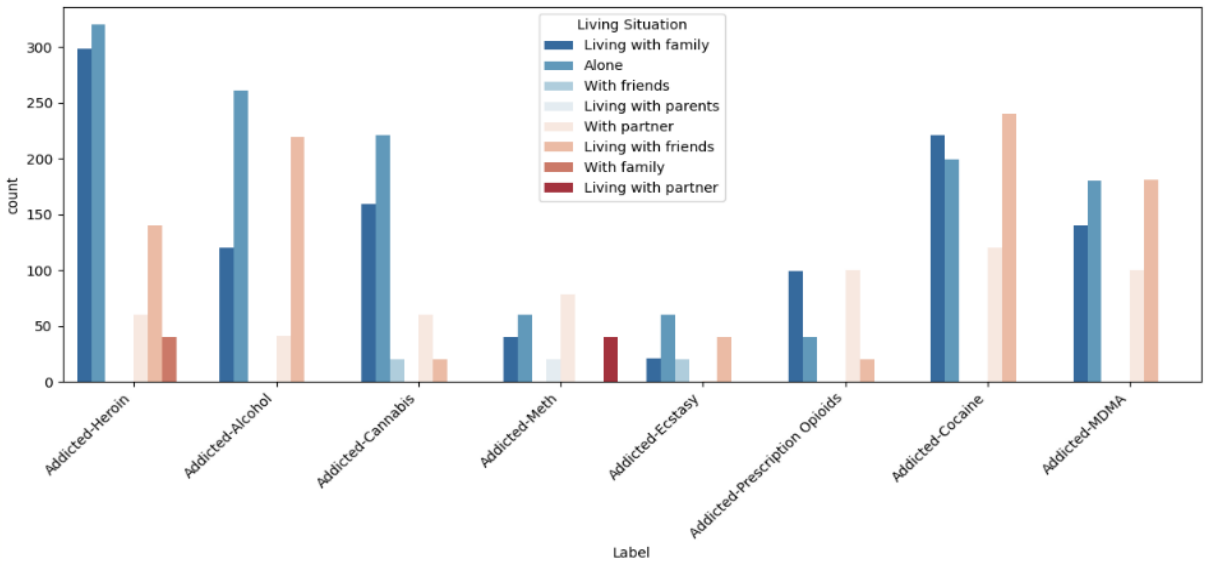


Fig 3.15: Drug addiction-based Living Situation feature analysis.

This fig 3.15 shows Drug addicted people living situation analysis. In Living situation features has 8 categories like: Living with family, Alone, with friends, living with parents, with partner, living with friends, with family, Living with partner. Addicted-Heroine people living alone most 350 counts of people. Then less 300 people who taken heroine living with family. After that Addicted-Ecstasy drug people have lowest rate of the category.

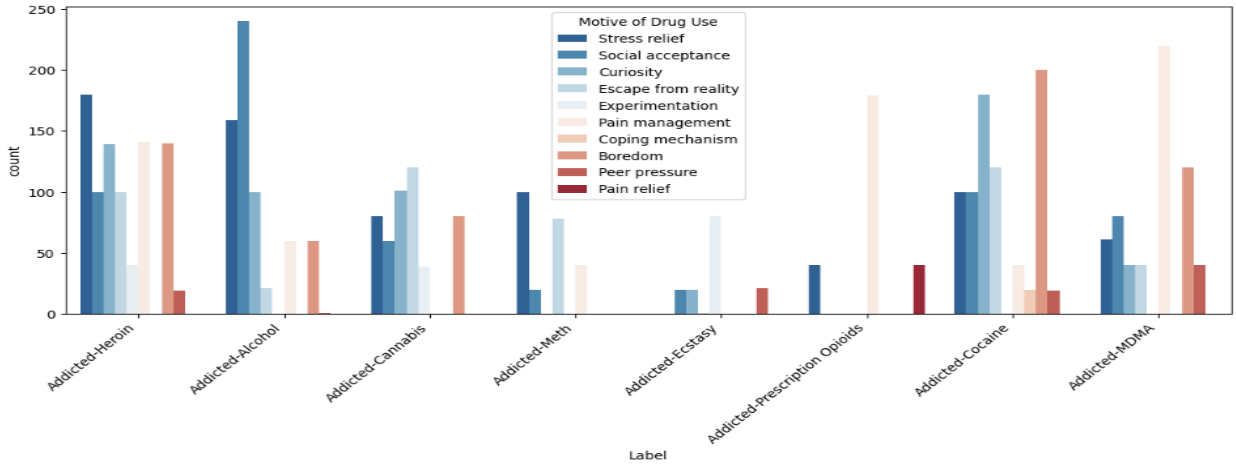


Fig 3.16: Drug addiction-based Motive Of Drug Use feature analysis.

This fig. 3.16 shows Motive of drug use based on drug addiction people. Some sub categories have been used in this analysis like: stress relief, Social acceptance, Curiosity, Escape from reality, Experimentation, Pain management, Coping Mechanism, Boredom, Peer pressure and Pain relief. Addicted-Alcohol people has the most rate of Motive of drug use name Stress relief get 160 and Social acceptance 240 people. Then Addicted-Ecstasy drug people get the lowest rate of motive to drug use like: social acceptance 20 people, Curiosity people have 20 etc.

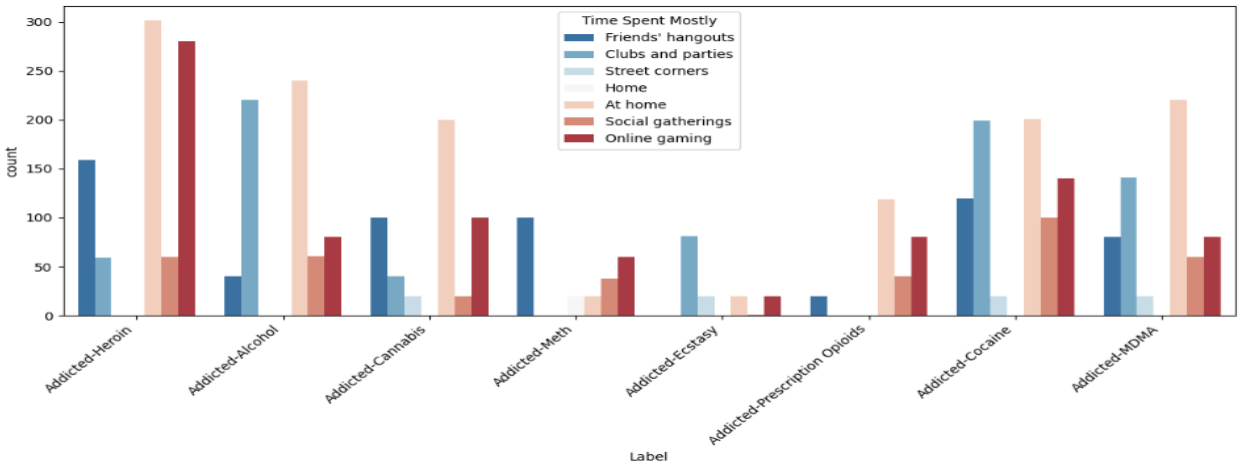


Fig 3.17: Drug addiction-based Time spent Mostly feature analysis.

The fig 3.17 shows the mostly spent time of an addicted person. Some features used like: Friends hangout, club and parties, street corners, Home, at home, social gatherings and Online gaming. From this graph we can see a heroin addicted spent mostly time by hanging out with friends. Alcohol addicted spent his time in clubs and parties. Heroin addicted person also stay at home around 300 and spent his/her time in online game.

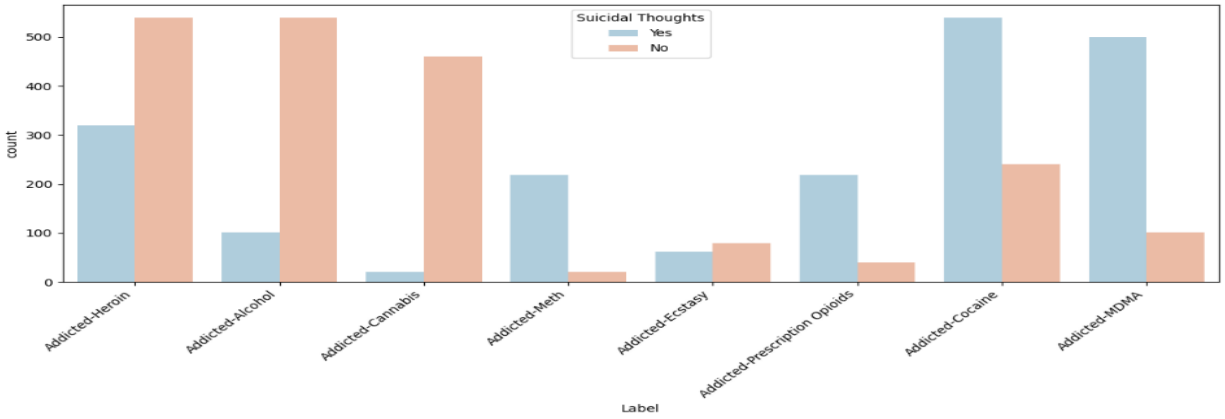


Fig 3.18: Drug addiction-based Motive Of Drug Use feature analysis.

From this fig 3.18 we can see that the bad effects of drugs and it can compel anyone to suicidal thoughts. Suicidal Thoughts has two feature like: Yes or No. Taking Cannabis the chances of suicidal thought are too low. But Cocaine, MDMA can oblige anyone for suicide.

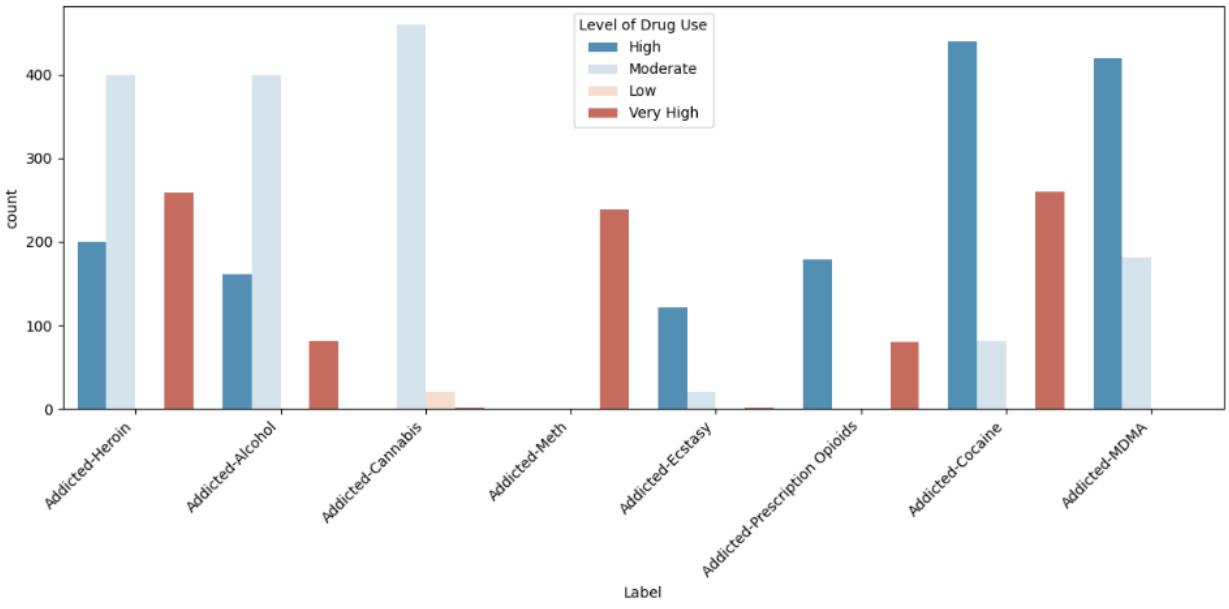


Fig 3.19: Drug addiction-based Level of Drug Use feature analysis.

The fig 3.19 describes the level of drug use. This feature has some categories like: High, Moderate, Low, Very High. Very high addicted sector is Heroin and MDMA. High sector is cocaine. Addicted Cannabis is the level of moderate.

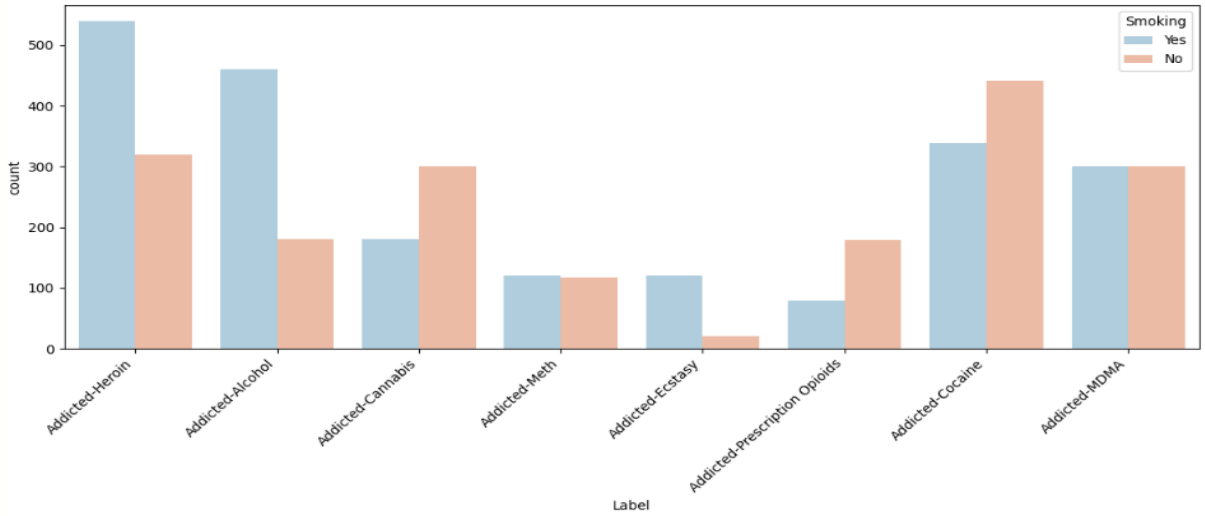


Fig 3.20: Drug addiction-based Smoking feature analysis.

The fig. 3.20 shows the number of smokers and nonsmokers who are addicted to drugs. Smoking features has 2 categories: Yes & No. Heroin addicted people are mostly smoked, it's more than 500 and nonsmokers around 300.Lowest number of smokers is prescription opioids.

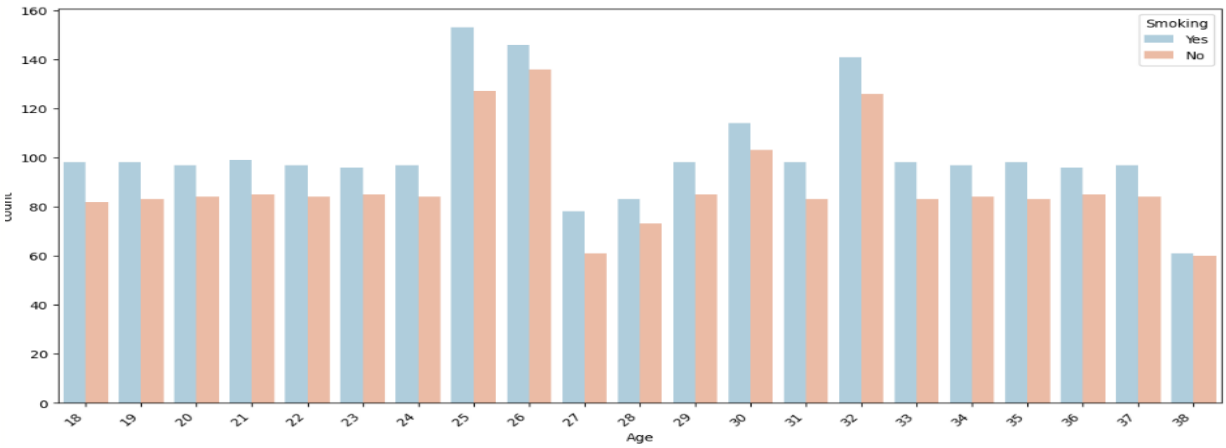


Fig 3.21: Drug addiction Age-based Smoking feature analysis.

The fig 3.21 shows the age of smokers and nonsmokers. Smoking features has 2 categories: Yes & No. 25 aged people are mostly smoked and it's around 160, who doesn't smoke it's count 120.Lowest count of smoker in this graph is 38 aged people. Smokers and nonsmokers count are mostly similar and it's around 60.

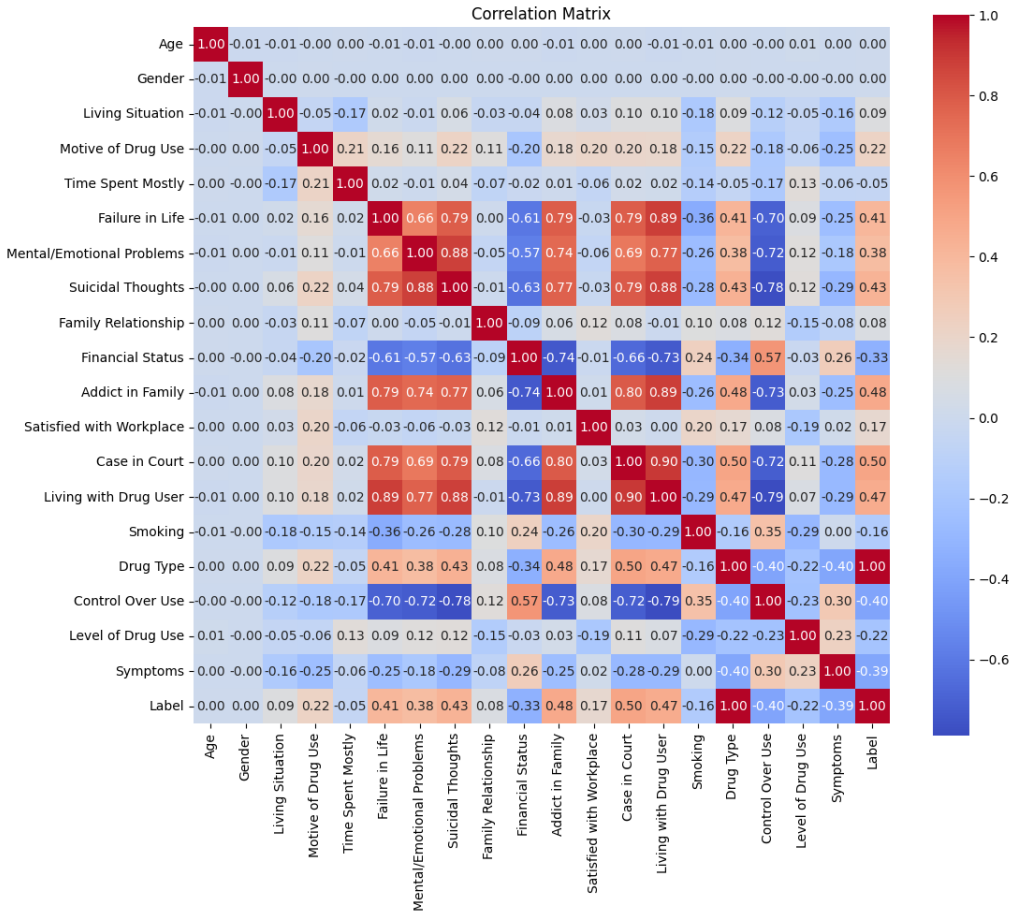


Fig 3.22: Correlation matrix analysis.

The Fig 3.22 correlation heatmap, which shows the dataset's pairwise relationships between variables, is shown here. Strong positive correlations are represented by colors of red, whereas strong negative correlations are represented by colors of blue. The heatmap shows important relationships, such as symptoms having a substantial correlation with level of drug use and other variables.

```
[109] df.describe()
```

	Age	Gender	Living Situation	Motive of Drug Use	Time Spent Mostly	Failure in Life	Mental/Emotional Problems	Suicidal Thoughts	Family Relationship	Financial Status	Addict in Family	Satisfied with Workplace	Case in Court	Living with Drug User	Smoking	Drug Type	Control Over Use	Leve Drug
count	3857.000000	3857.000000	2389.000000	3857.000000	3857.000000	3857.000000	3857.000000	3857.000000	3857.000000	3857.000000	3857.000000	3857.000000	3857.000000	3857.000000	3857.000000	3857.000000	3857.000000	3857.00
mean	27.851698	0.799585	0.470908	3.203785	3.288566	0.551724	0.525797	0.494685	1.026445	0.761732	0.572206	0.542131	0.484314	0.536168	0.541613	3.264195	0.532020	1.08
std	5.835646	0.400363	0.531751	2.568250	2.170496	0.497382	0.499399	0.500037	0.717538	0.781004	0.494823	0.498286	0.499819	0.498755	0.498330	2.644270	0.499038	1.14
min	18.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00
25%	23.000000	1.000000	0.000000	1.000000	1.000000	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000	0.00
50%	28.000000	1.000000	0.000000	3.000000	4.000000	1.000000	1.000000	0.000000	1.000000	1.000000	1.000000	1.000000	0.000000	1.000000	1.000000	3.000000	1.000000	1.00
75%	33.000000	1.000000	1.000000	5.000000	5.000000	1.000000	1.000000	1.000000	2.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	6.000000	1.000000	1.00
max	38.000000	1.000000	2.000000	9.000000	6.000000	1.000000	1.000000	1.000000	2.000000	2.000000	1.000000	1.000000	1.000000	1.000000	1.000000	7.000000	1.000000	3.00

Fig 3.23: Data Describe.

The following table summarizes the dataset's important statistics. For every characteristic, it provides metrics like count, mean, std (standard deviation), min, and quantiles (25%, 50%, 75%, max). It offers information on the distribution of the data, central tendency, variability, and missing values. For example, it shows that the Living Situation variable has a lower count than the others.

Model Evaluation

Numerous machine learning approaches were examined, such as KNN, Decision Tree, Random Forest classifier, SVM, Naive Bayes, XG Boost, and Random Forest. We leverage the ability of our models to capture language variations, context, and patterns in a variety of models. This in-depth examination helps us to accomplish our goal of creating a reliable classification system for drug addiction.

1. **Naive Bayes:** It is recommended to use the naïve Bayes method for machine learning even with a large number of records since it can handle large amounts of data. One of its strong points is NLP (natural language processing) tasks, such classifying drugs in texts. The filtering procedure is easy to use and fast. Comprehending the naive Bayes classifier completely requires an understanding of the Bayes theorem. We talk about the Bayes theory in our first exchange. The basis of this argument is the idea of conditional probability. Contingency probability is the chance that one event will occur given the possibility of another. We may use our prior knowledge and the conditional probability to determine the likelihood of an event. Naive Bayes methods are extensively employed in systems for guidance, sentiment analysis, and spam removal, to name a few uses. The main disadvantage of this approach is the requirement for distinct models, even with its speed and simplicity of use. Because the prediction components are interconnected, the classifier performs badly in the majority of real-world scenarios, as shown in Figure 3.24 [37]. The following formulation is widely used in various applications, including text classification and spam filtering [43].

   Bayes' Theorem

   $$P(y|x)=P(x)P(x|y)P(y) \quad \text{..... (i)}$$

   Naive Bayes Classifier Prediction

   Since P(x)P(x)P(x) is constant for all classes, the model predicts the class y^\hat{y}y^ that maximizes the posterior probability:

$$y^=argymaxP(y)i=1\prod nP(xi|y) \quad \text{..... (ii)}$$

Where:

- P(y) Prior probability of class y.
- P(xi|y)P(x_i \mid y)P(xi|y): Likelihood of feature xi given class y.
- ∏i=1nP(xi|y)\prod_{i=1}^n P(x_i \mid y)∏i=1nP(xi|y): Assumes conditional independence of features, simplifying computation.
- P(x)P(x)P(x): Marginal probability of the input xxx (constant across classes, so it can be ignored for classification).

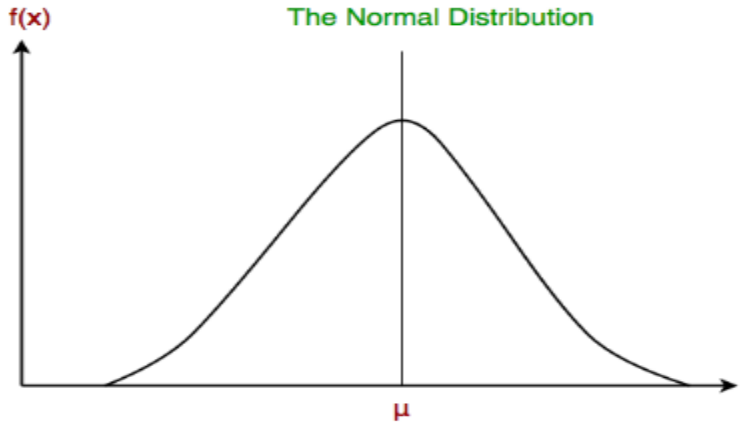


Fig 3.24: Naive Bayes models.

2. **Random Forest:** For the two classification categories, the RF Separator's tree-based method might be applied. An ML algorithm creates a hierarchical tree. This method is used by artificial intelligence to create hierarchical "decision trees". The randomized forest classification method builds several decision trees using a combination mechanism and then combines them all. This sheds lighter on the issues around overfitting. Machine learning is one of the hottest subjects in business right now because of its adaptability and ability to be utilized anywhere there is a lot of data. Owing to its many benefits, machine learning-based RF Encoder is usually chosen over other approaches. The approach was first developed in 1997 and was initially designed to work with very big datasets, as shown in Figure 3.25 [38]. The following formulation is commonly used in various applications, including decision tree ensembles and classification tasks [44].

$$y\ ^{\ } =arg\ c\ max\ \ t=1\ \sum T\ \ I(h\ t\ \ (x)=c) \quad \text{.....(iii)}$$

Where:

- y^: Final predicted class.

- c: Possible class labels.
- T: Total number of decision trees.
- ht(x): Prediction from the t-th tree for input x.
- I(·): Indicator function, 1 if ht(x)=c, otherwise 0.

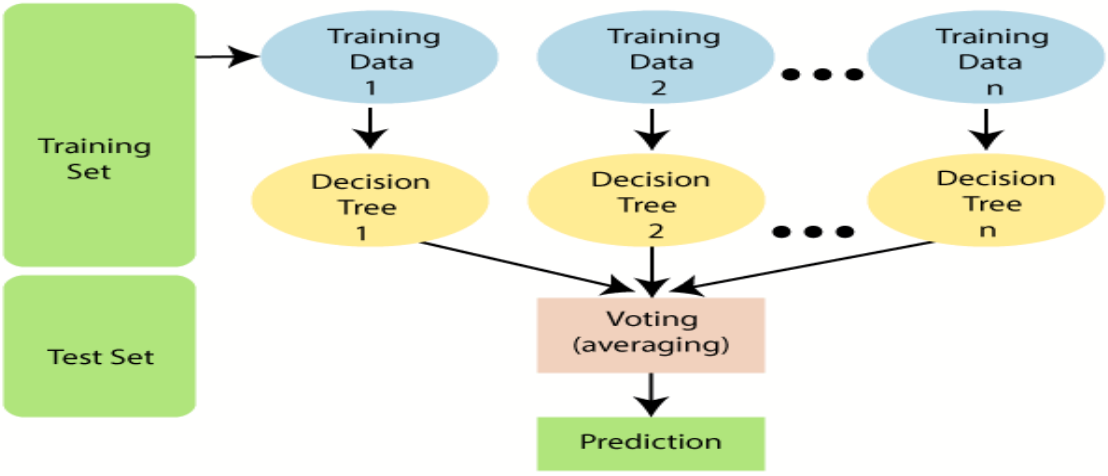


Fig 3.25: The process of the Random Forest classifier.

3. **XGBoost:** A well-liked and successful machine learning technique is called Extreme Gradients Boost, or "XG Boost," and it may be used to address clustering problems such as regression. Slope-enhanced decision-tree architectures may be produced more quickly and easily with the help of this technique. XG Boost is an ensemble learning strategy that combines the forecasts of many weak learners (typically decision trees) to enhance a model. One by one, each failed student makes up for previous mistakes by learning from their less fortunate peers. XG Boost has many normalizing penalty techniques built in to guard against overfitting. Due to the effectiveness of training using penalty-based regularizations, the algorithm is able to attain the appropriate degree of applicability. What is a non-linear system? Structures may be identified and trained on unexpected input using XG Boost. To be clear, right out of the box, everything is built and functional, as shown in Figure 3.26 [39]. The following formulation is commonly applied in various domains, including ensemble learning and regression tasks [45].

$$y\hat{}\ i\ =\phi(x\ i\ )=\ k=1\ \sum K\ \ f\ k\ \ (x\ i\ )\ \ .....(iv)$$

Where:

- y^i\hat{y}_iy^i: Predicted value for input xi.
- ϕ(xi) Overall model prediction as the sum of all tree outputs.

- fk(xi) Output of the k-th decision tree for input xix_ixi.
- K: Total number of trees.

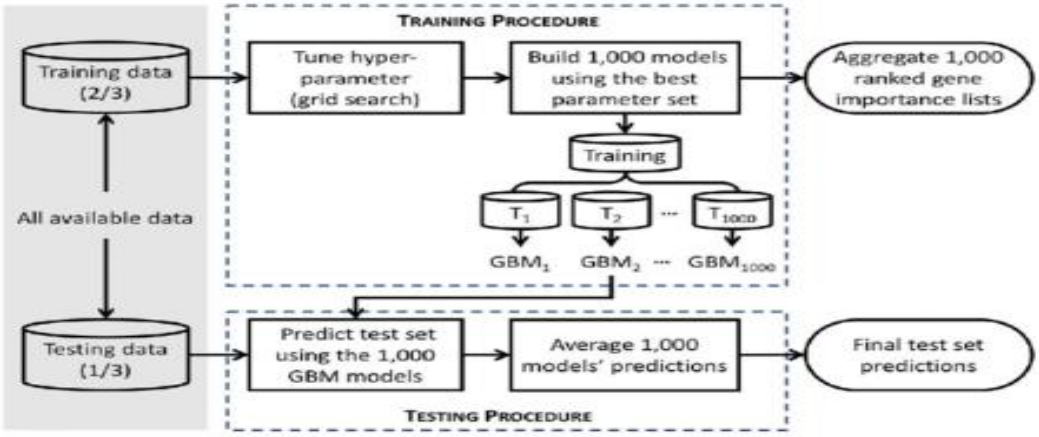


Fig 3.26: XG Boost model construction.

4. **K-Nearest Neighbors:** Machine learning professionals employ the robust and intuitive k-nearest-neighbors (KNN) technique to solve regression and classification problems. KNN employs the K near neighbors of the most recent data item in the training set to use the similarity principle when predicting the amount or title of a specific recent data point. This article will cover the supervised neural network (KNN) method, commonly referred to as the k-Nearest Neighbors approach, and its user-friendly features. Due to its simplicity and ease of use, the KNN technique is a widely used prediction technology. There is nothing to infer about the distribution of the underlying data. It is a flexible replacement for many other types of statistics in the areas of regression, classification, and other applications due to its capacity to handle both numerical and categorical data. This unapproved method creates predictions by comparing the degree of similarity between data points in a certain batch of data. Outliers affect K-NN less than they do other algorithms, as shown in Figure 3.27 [40]. The following formulation is commonly used in various applications, including k-nearest neighbors classification and pattern recognition [46].

$$y\hat{} = \arg c \max \; i \in N k \; \sum \; I(y\, i = c) \quad \text{.....(v)}$$

Where:

- Nk: Set of indices for the kkk-nearest neighbors.
- yi: Class label of neighbor iii.
- c: Possible class labels.
- I(·): Indicator function, 1 if yi=cy_i = cyi=c, otherwise 0.

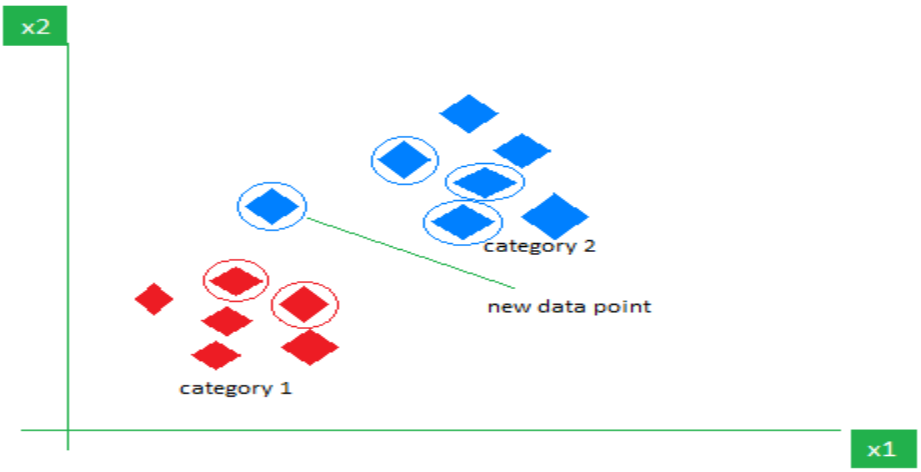


Fig 3.27: Operational Visualization of KNN Algorithm.

5. **SVM**: The process of creating a Support Vector Machine (SVM) classifier starts with gathering data and doing preprocessing, such as standardization. Next, the data is divided into testing and training sets. Using support vectors, SVM maximizes the margin between classes to find the ideal hyperplane that divides them during training. SVM uses kernel functions to move the data into a space with more dimensions in order to achieve better separation when the data cannot be separated linearly. After training, measures like accuracy and F1-score are used to assess the model's performance on the testing set. After the model has been verified, it may be used to forecast fresh data, and hyperparameter adjustment can be done to enhance the outcome, as shown in Figure 3.28 [41]. The following formulation is commonly used in various applications, including support vector machines and classification tasks [47].

$$f(x)=\text{sign}(\sum_{i=1}^{n} \alpha_i y_i K(x_i, x)+b) \quad \text{.....(vi)}$$

Where:

- xi: Support vectors (training points on the margin or violating it).
- yi: Class label of xix_ixi (+1+1+1 or −1-1−1).
- αi: Lagrange multiplier for xix_ixi, determined during training.
- K(xi,x): Kernel function measuring similarity between xix_ixi and xxx.
- b: Bias term.
- sign(·): Determines the class (+1+1+1 or −1-1−1) based on the sign of the decision function.

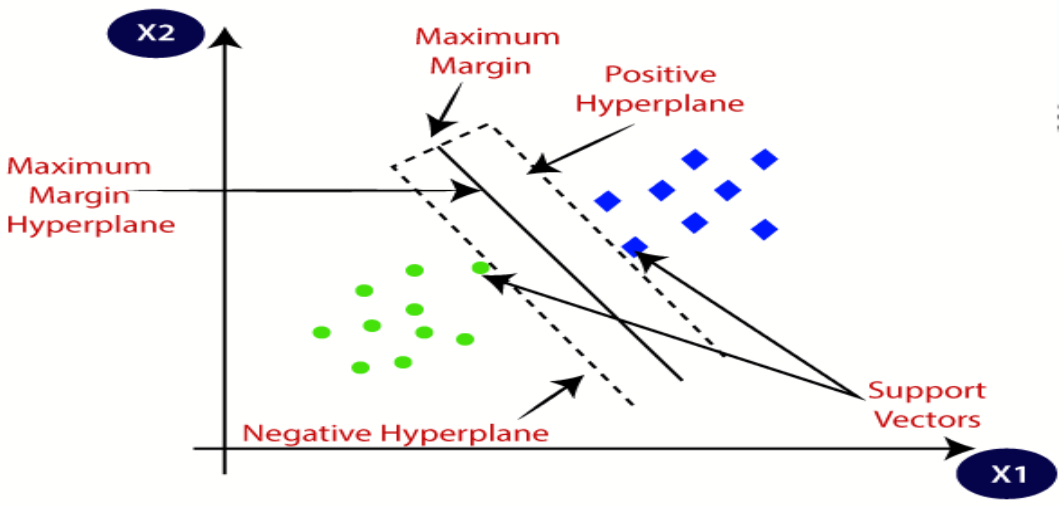


Fig 3.28: A demonstration of SVC.

6. **Decision Tree:** The first steps in the Decision Tree classifier workflow are data collection and preparation, which include encoding categorical variables and managing missing values. Subsets for testing and training are then created from the dataset. In the training phase, the algorithm splits the data recursively based on feature values to produce branches that lead to make a choice nodes and, eventually, leaf nodes, which represent class labels, creating a model that resembles a tree. At each split, the objective is to reduce impurity or increase information gain. Performance measures like accuracy, precision, and recall are computed to evaluate the tree's efficacy once it has been built using the testing data. Using pruning approaches, branches that contribute minimal predictive power can be removed in order to prevent overfitting. After undergoing validation, the Decision Tree model may be applied to generate predictions on fresh data, and its user-friendly layout facilitates comprehension of the decision-making procedure, as shown in Figure 3.29 [42]. The following formulations are widely used in various applications, including decision tree construction and classification tasks [48].

1. **Gini Impurity** (for classification):

$$Gini(S)=1-\sum_{i=1}^{C} p_i^2 \quad \text{.....(vii)}$$

2. **Entropy** (for classification):

$$Entropy(S)=-\sum_{i=1}^{C} p_i \log_2 p_I \quad \text{.....(viii)}$$

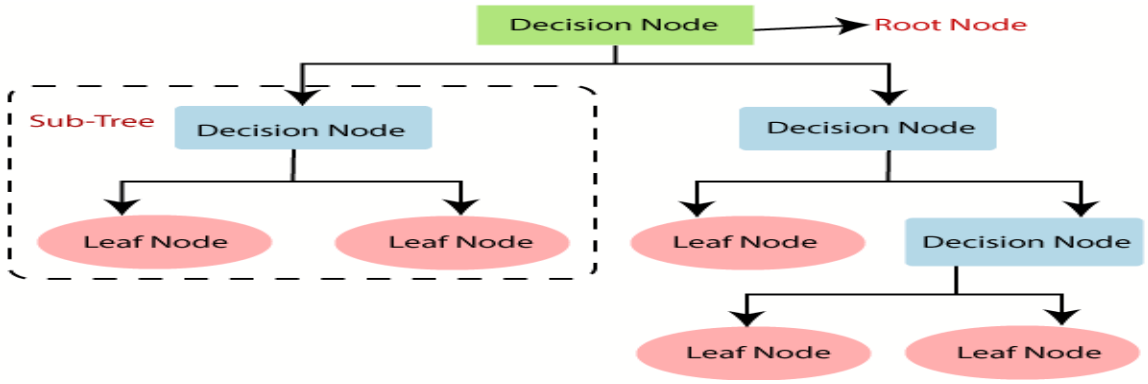


Fig 3.29: Workflow of Decision Tree.

3.3 Project Plan

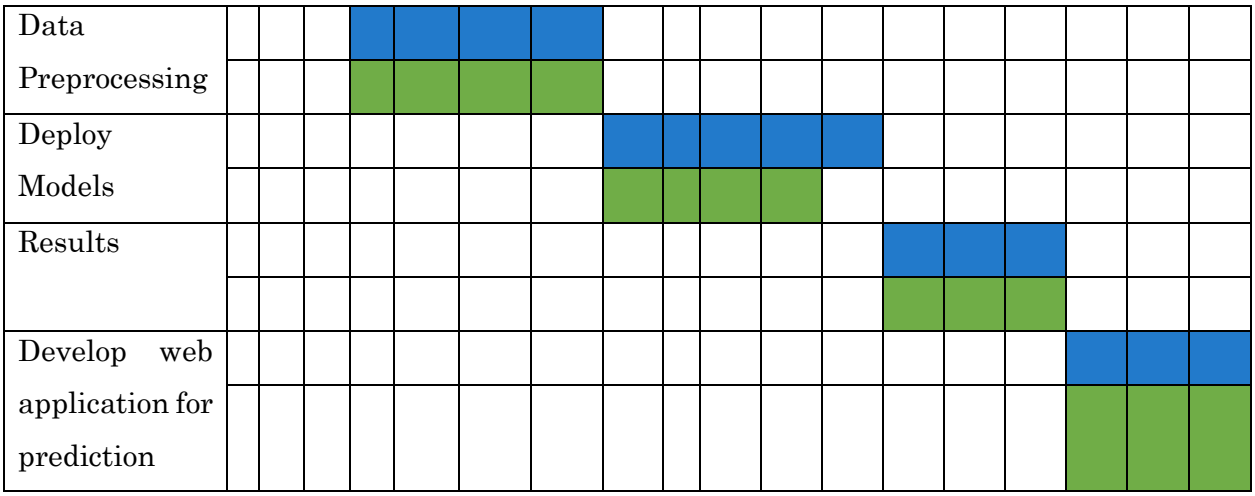
Table 3.3: Total project plan & time estimate.

S.No.	Next Task	Estimate completion time (MM-YY)
1	Real data collection	11-24
2	Complete very well of data preprocessing.	11-24
3	Choosing machine learning models deploying.	11-24
4	Reaching the greatest accuracy levels of more than 90%, web application design	11-24
5	Report writing	11-24

3.4 Task Allocation

Table 3.4: Task Allocation.

Tasks	Weeks																	
	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Data Collection																		



3.5 Summary

To ensure accuracy, details must be acquired after the earlier phases are finished. Our project's core idea required the completion of 10 components. If we are to reach our goal, all of these chores need to be completed.

- Data collection.
- Preprocessing of data.
- Eliminating duplicates and null values from the data.
- The tokenization process.
- Using models is advised for each of the six approaches.
- Building a working prototype of a label for drug addiction classification website.
- Examine and discuss the outcome.

The idea wouldn't function until we began coding it. The accuracy of six different techniques was examined. After it was finished, we assessed the method's accuracy. After examining the accuracy, we concluded that the design indicated above would be better appropriate for our requirements. A set of rules has been developed for each effort to categorize drug addiction following a thorough examination of relevant theoretical and numerical approaches and concepts. Here are a few results that could be noteworthy.

CHAPTER 4

Implementation and Results

4.1 Environment setup

To perform in-depth research, the hardware, software, and data resources used in this setup for **Early Intervention: A Machine Learning Approach to Classify Drug Addiction.** have been carefully chosen. High-performance computers are computing systems with a computational core made up of powerful GPUs and a sturdy CPU. I am aware that no innovation can guarantee flawless outcomes. In a manner similar to that, we may modify the parameters of our model during training to improve accuracy. The approach we took determined the different outcomes we obtained. We were able to accurately estimate the class label for drug addiction connected with all drug addicts and their symptoms thanks to six machine learning algorithms. Following the application of these techniques to assess the relative efficacy of every component of the entire structure, a number of verification procedures were conducted to ascertain the outcome of the forecast. After each statistic was chosen, each model used a single file including information from both publicly available web sources and our own study. We used Matlab and related pre-configured libraries to assess the algorithms' output when the data collecting was complete. The item is then assessed to determine if it satisfies the standards concerning the signs and symptoms of drug addiction using a comparable dataset. They can be categorized as "Addicted-Prescription Opioids," "Addicted-Cocaine," "Addicted-Heroin," "Addicted-Alcohol," "Addicted-Cannabis," "Addicted-Meth," "Addicted-Ecstasy," and so on. In this case, we extensively examined different models using relevant performance requirements. Metrics including recall, accuracy, precision, and total F1 score offer a thorough assessment of the strategies' effectiveness.

Accuracy: As a broad measure of the model's effectiveness, the accuracy evaluating classification was calculated as the ratio of the total number of response samples to all the samples.

Confusion matrix: By utilizing the confusion matrix to analyze the model's behavior, the occurrences of each class false positives, false negatives, true positives, and true negatives

were found and predicted. The research avenues and potential roadblocks in the quest for drug addiction were displayed in this matrix.

4.2 Testing and Evaluation/Performance/ Comparative Analysis

Six distinct algorithms were employed in this study. Before we could go to work, they had a lot of stuff to find. Yes, we did choose and begin working on the algorithm. Next, each strategy's accuracy was assessed. The decision tree classifier used ways to get an ultimate precision of 97.75%, which is used as a rough approximation for all six models. Once again, an online tool was created to identify the drug addiction group associated with their symptoms.

Precision: One popular criterion to evaluate the model's effectiveness is the accuracy with which the algorithm produces forecasts. The total number of true positives multiplied by the total number of accurate forecasts can be used to determine efficiency.

$$\text{precision} = \frac{TP}{TP+FP} \quad \text{.....(ix)}$$

Recall: Retrieval is the percentage of suitable cases that were ultimately found and retrieved, regardless of all relevant cases. A high recall rate indicates that the strategy yielded the most pertinent results.

$$\text{recall} = \frac{TP}{TP+FN} \quad \text{.....(x)}$$

F1-Score: The validity of a test is assessed based on its recall and accuracy. Recall and accuracy function best together.

$$\text{F1} = \frac{2 \times precision \times recall}{precision + recall} \quad \text{.....(xi)}$$

Accuracy: It is feasible to evaluate the efficacy of the model by combining recall and dependability, as the F1 score does. You may compute it using the following equation:

$$\text{accuracy} = \frac{TP+TN}{TP + FN + TN + FP} \quad \text{.....(xii)}$$

Bar Graph for Accuracy:

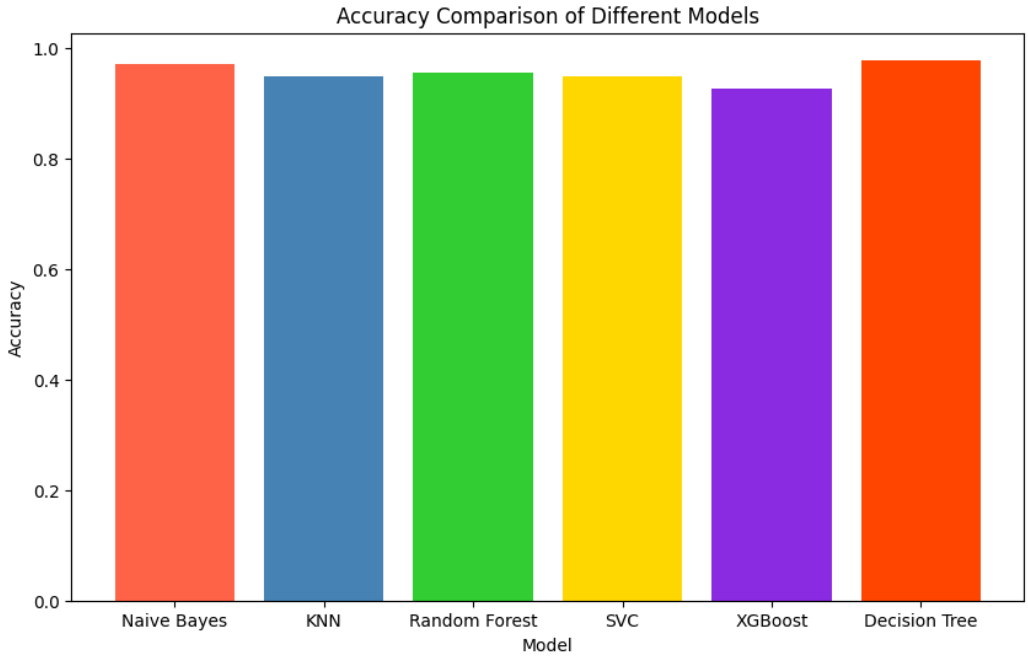


Fig 4.1: All model accuracy score bar chart comparison.

Bar Graph for F1:

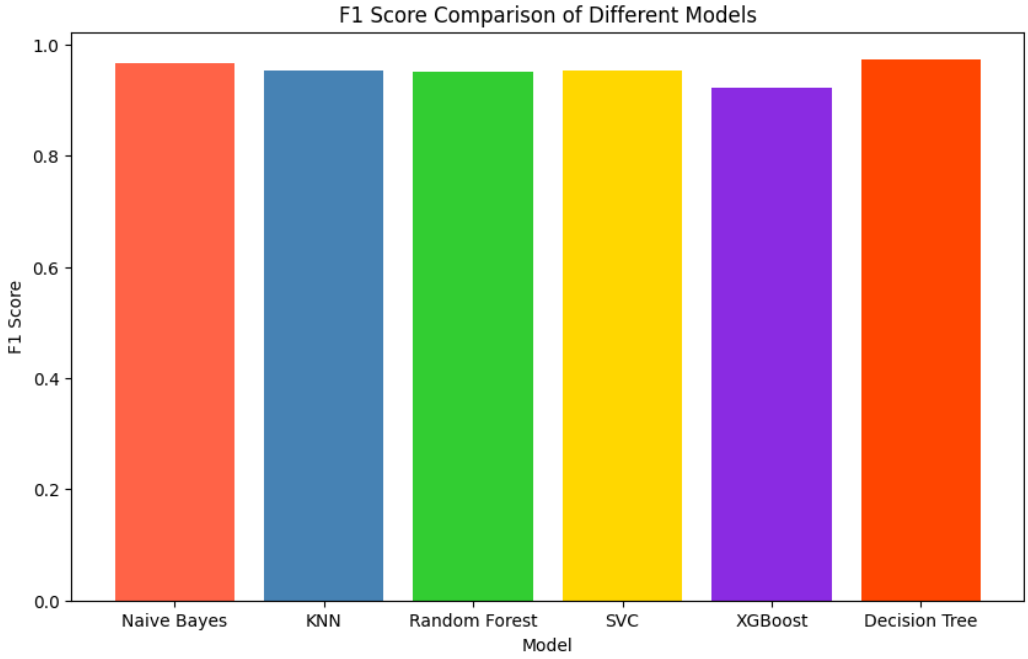


Fig 4.2: All model f1 score bar chart comparison.

Bar Graph for Recall:

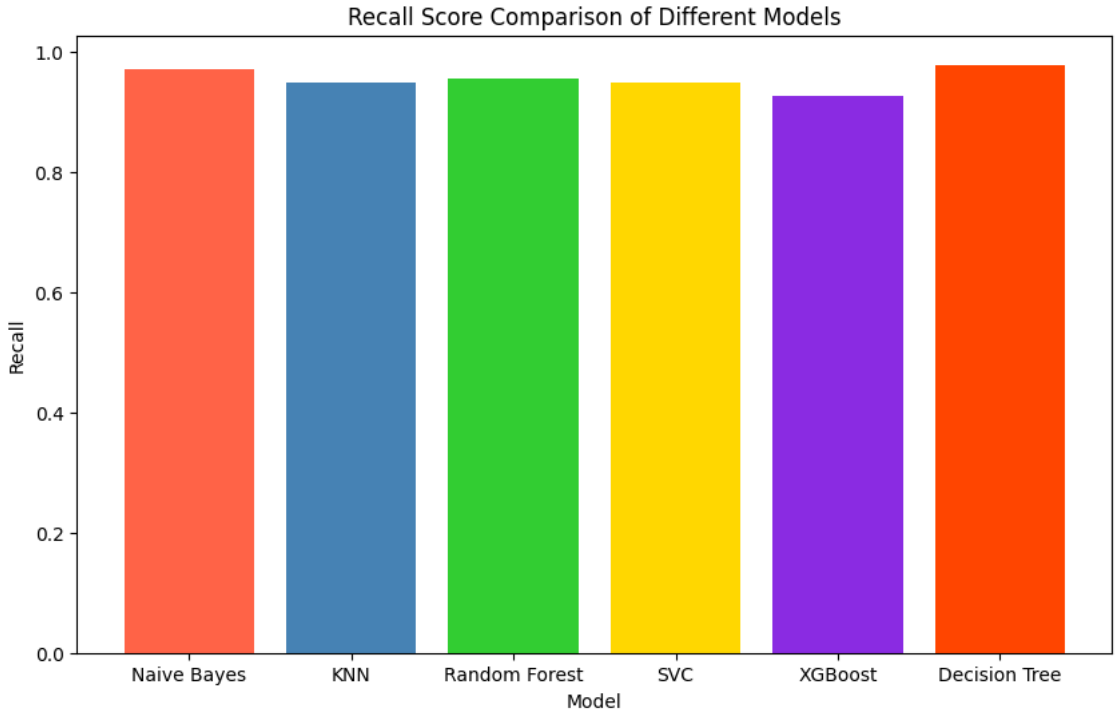


Fig 4.3: All model recall score bar chart comparison.

Bar Graph for Precision:

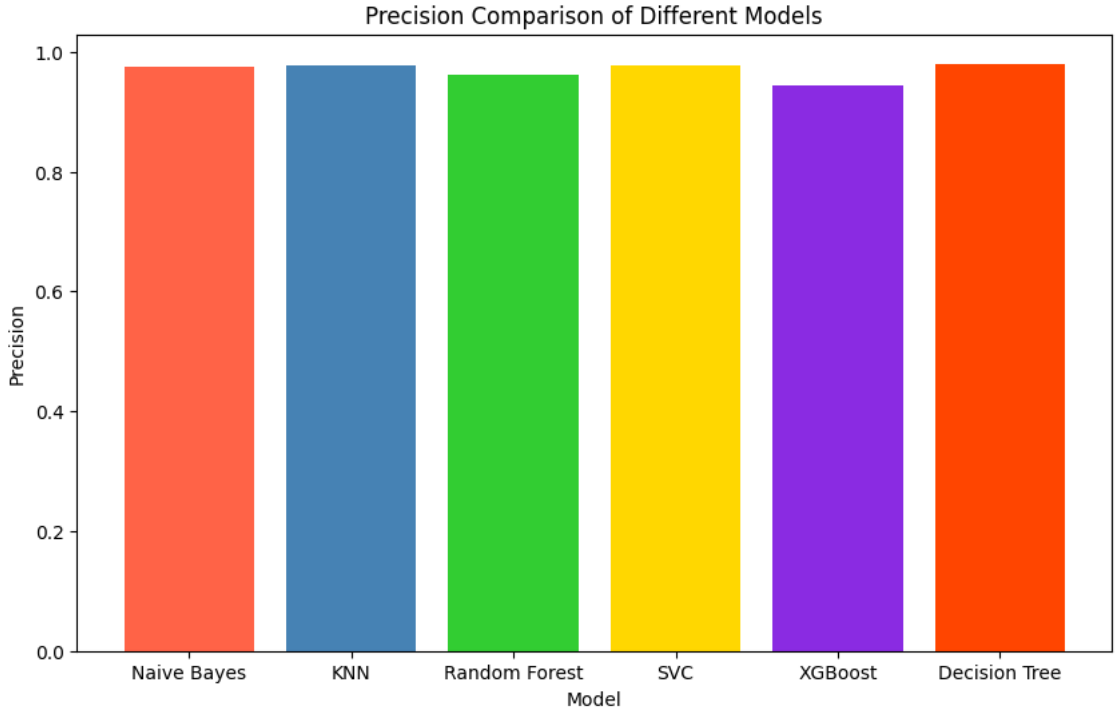


Fig 4.4: All model precision score bar chart comparison.

Comparative other work:

Table 3.4 provides a comparative analysis of several studies on substance addiction prediction, highlighting the datasets used, target classes, performance metrics, methodologies, and associated limitations. This comparison demonstrates both the advancements and challenges in the field of addiction prediction using machine learning. My work's AUC (1.00), F1 score (0.95), and accuracy (97.75%) outperform all other reviewed studies, showcasing a more effective prediction model.

Table 4.1: Comparative results of this work with recent research

Source	Dataset	Classes	AUC	F1 Score	Accuracy	Approach	Limitations
Gong et al. (2021) [28]	Public dataset	Prescription Opioids, Cocaine, Heroin, Alcohol, Cannabis, Meth, Ecstasy	Not available	Not available	Not available	Machine Learning (SVM, Random Forest)	Limited by the nature of public datasets, may not generalize to all populations
Lee et al. (2019) [29]	Clinical dataset	Alcohol Use Disorder (AUD) vs. Non-AUD	0.88	Not available	85	Support Vector Machine (SVM), Logistic Regression	Focused on a single addiction type (AUD), may not generalize to other drugs
Sahker et al. (2015) [30]	National Survey	Various (Alcohol, Marijuana, Tobacco)	Not available	0.81	77	Decision Trees, Random Forest, Logistic Regression	Limited by national data, might overlook behavioral aspects of addiction
Acion et al. (2017) [31]	Clinical dataset	Substance Use Disorder (SUD)	Not available	0.9	88	Ensemble Learning, Naive Bayes	Specific to substance use treatment, not broader addiction classification
Ash-Houchen & Lo	National Survey	Various (Illicit	Not available	Not available	Not available	Regression,	Limited to racial/ethnic differences,

(2020) [32]		Substances)				Statistical Analysis	doesn't account for other factors in addiction
Zhang et al. (2018) [33]	Drug Toxicity Dataset	Various Drugs	0.92	0.89	90	Machine Learning (Random Forest, SVM)	Focuses on drug toxicity rather than addiction classification, may lack addiction-related features
Weinstein et al. (1992) [34]	Cancer Drug Dataset	Various (Cancer Drugs)	0.94	Not available	91	Neural Networks, Decision Trees	Not directly related to addiction, focused on cancer drug prediction
Chen et al. (2001) [35]	Near-Infrared Spectroscopy Data	Drug Content and Tablet Hardness	N/A	N/A	N/A	Artificial Neural Network	Limited focus on pharmaceutical applications, not addiction or clinical drug behavior
My Work (2024)	Real Dataset (4000 individuals) collected from the Divisional Drug Addiction Treatment Centre, Rajshahi	Prescription Opioids, Cocaine, Heroin, Alcohol, Cannabis, Meth, Ecstasy, MDMA.	1.00	0.95	97.75	Decision Tree, SVM, Random Forest, XGBoost, KNN, Naive Bayes	Limited by dataset size and diversity, may require additional data for better generalization across different populations

4.3 Results & Discussion

The item is then assessed to determine if it satisfies the standards concerning the signs and symptoms of drug addiction using a comparable dataset. They may be categorized as "Addicted-Prescription Opioids," "Addicted-Cocaine," "Addicted-Heroin," "Addicted-Alcohol," "Addicted-Cannabis," "Addicted-Meth," "Addicted-Ecstasy". In this case, we extensively examined different models using relevant performance requirements. Metrics including recall, accuracy, precision, and total F1 score offer a thorough assessment of the strategies' effectiveness. In this section, many classifiers' output is displayed. PyCharm and Co Lab were two complementary tools that performed well in tandem throughout. Six classifiers in all were used. Decision Tree, SVM, Random Forest, KNN, XG Boost, and Naive Bayes.

Table 4.2: Accuracy, F1 Score, Precision, Recall Table.

Classifier	Accuracy Score	F1 Score	Precision	Recall
XG Boost	92.62%	92.26%	94.40%	92.62%
Random Forest	95.50%	95.15%	96.08%	95.50%
Naive Bayes	97.12%	96.75%	97.39%	97.12%
KNN	94.87%	95.41%	97.71%	94.87%
SVC	94.87%	95.41%	97.71%	94.87%
Decision Tree	97.75%	97.39%	98.00%	97.75%

Table 4.3: Accuracy of train and test model.

Classifier	Test accuracy	Train accuracy
XG Boost	92.62%	94.09%
Random Forest	95.50%	95.25%
Naive Bayes	97.12%	96.71%
KNN	94.87%	96.03%
SVC	94.87%	96.03%
Decision Tree	97.75%	97.81%

Classification reports:

	precision	recall	f1-score	support
Addicted-Alcohol	1.00	1.00	1.00	105
Addicted-Cannabis	1.00	1.00	1.00	96
Addicted-Cocaine	1.00	1.00	1.00	137
Addicted-Ecstasy	1.00	0.45	0.62	33
Addicted-Heroin	1.00	1.00	1.00	185
Addicted-MDMA	0.89	1.00	0.94	139
Addicted-Meth	1.00	1.00	1.00	55
Addicted-Prescription Opioids	1.00	1.00	1.00	50
accuracy			0.98	800
macro avg	0.99	0.93	0.95	800
weighted avg	0.98	0.98	0.97	800

Fig 4.5: Decision Tree classification reports.

The Fig 4.5 summaries the precision, recall, F1-score, and support for each class. The model's overall accuracy is 98%, with macro-averaged precision, recall, and F1-score values of 0.99, 0.93, and 0.95. Addicted-Ecstasy has a lower recall and F1-score than other classes, indicating difficulty in identifying this class.

Confusion matrix and Normalized Confusion matrix:

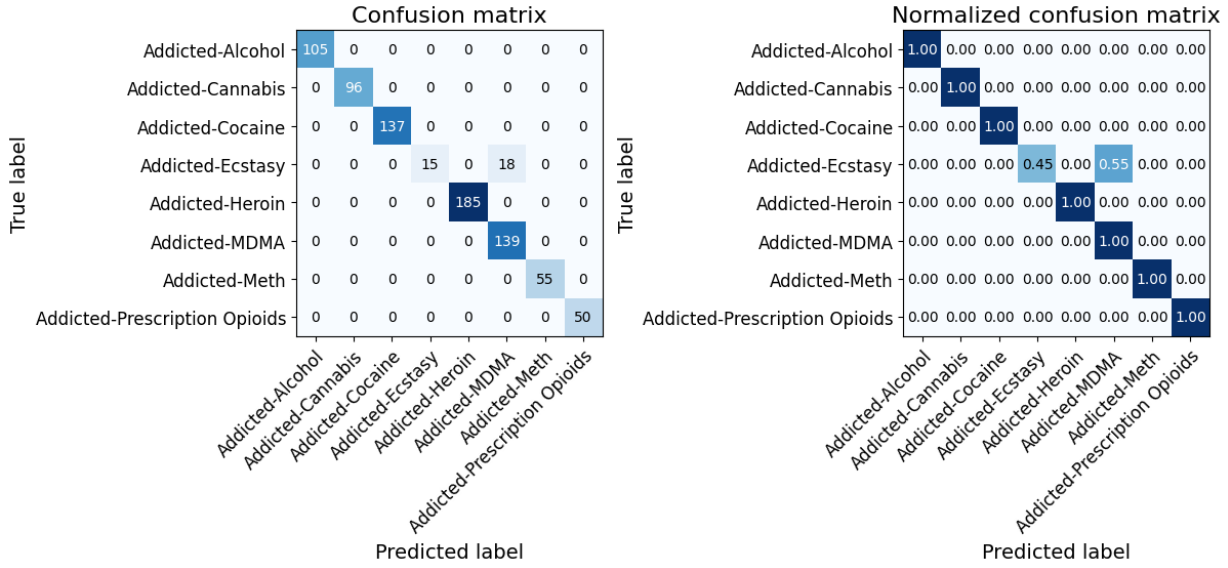


Fig 4.6: Confusion matrix and Normalized Confusion matrix of Decision Tree.

The Fig 4.6 left panel (Confusion Matrix) displays the raw counts of true positives, false positives, and false negatives for each addiction class. For example, the model accurately predicts most addiction classes, with the exception of Addicted-Ecstasy, which has misclassifications.

The Fig 4.6 right panel (Normalised Confusion Matrix) shows the normalised version of the confusion matrix, with values representing proportions rather than counts. For example, the classifier achieves 100% accuracy for most classes except Addicted-Ecstasy, which has a lower recall (0.45).

Accuracy of Training and Validation:

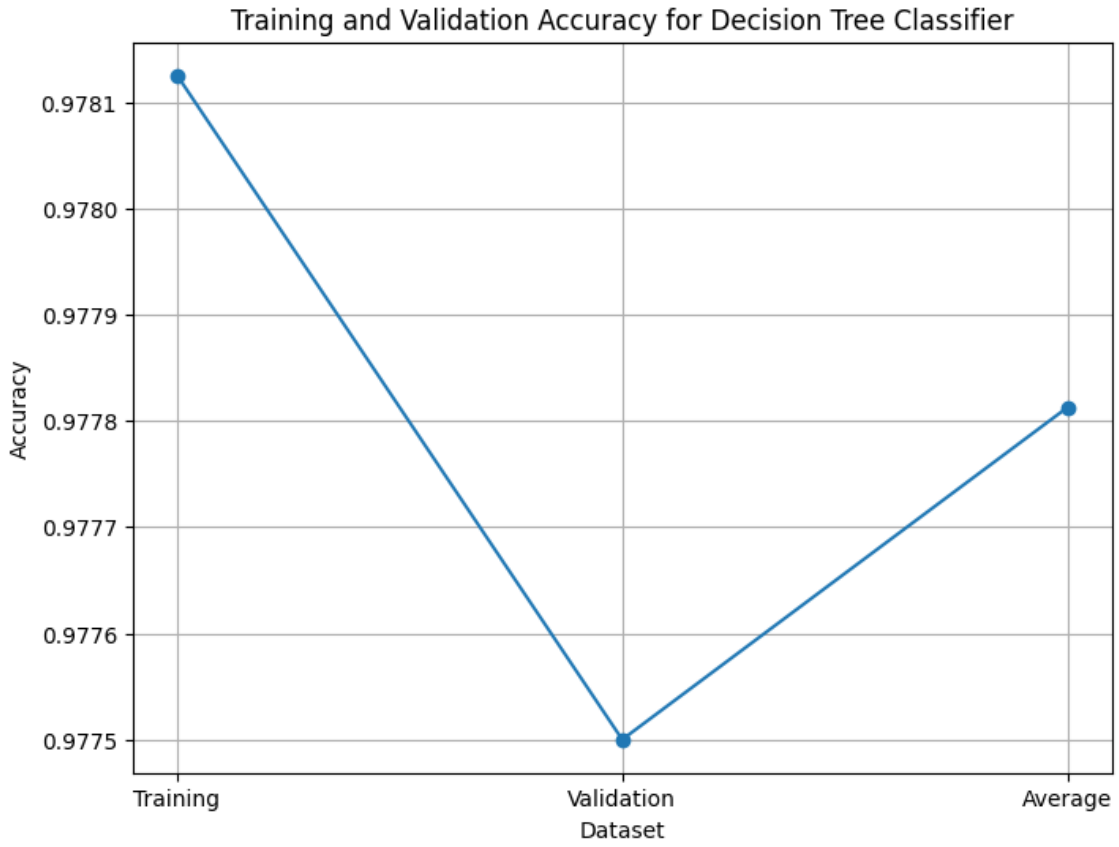


Fig 4.7: Decision Tree Accuracy of Training and Validation.

The Fig 4.8 shows the Decision Tree model's accuracy across the average, validation, and training datasets. The model is functioning effectively without overfitting, as seen by the training accuracy being marginally higher (~97.81%) than the validation accuracy (~97.76%). The model's strong generalization to unknown data is confirmed by the consistency of training and validation accuracy.

Loss Curve:

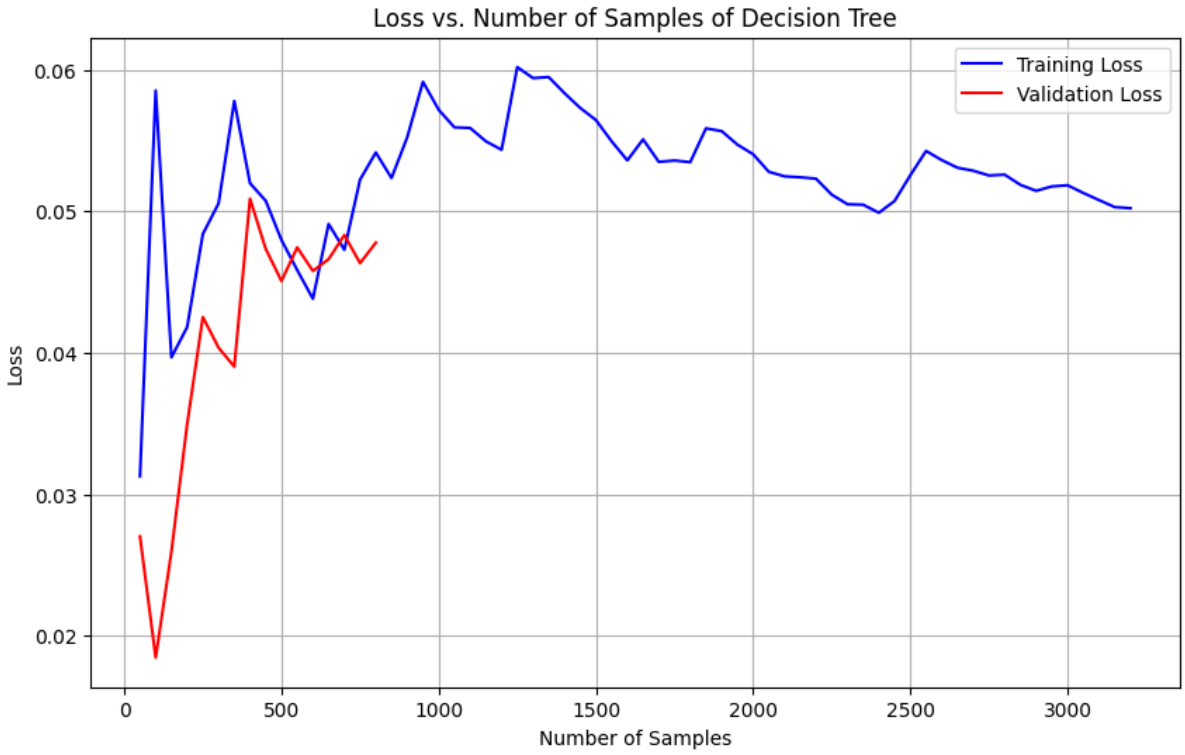


Fig 4.8: Loss curve for Decision Tree.

The Fig 4.9 plots the training and validation loss curves versus the amount of samples. Both curves show consistent and modest loss, indicating good model training without overfitting. However, initial fluctuations indicate that the model required a few data to stabilise during training.

Web prototype design:

Since all data is automatically classified using Decision Tree by the classifier to get optimum accuracy, just the Classification report is shown. Here below fig 4.10-4.12. shows web application predict design o for detection drug addiction using their symptoms.

Fig 4.9: Web prototype design of drug prediction.

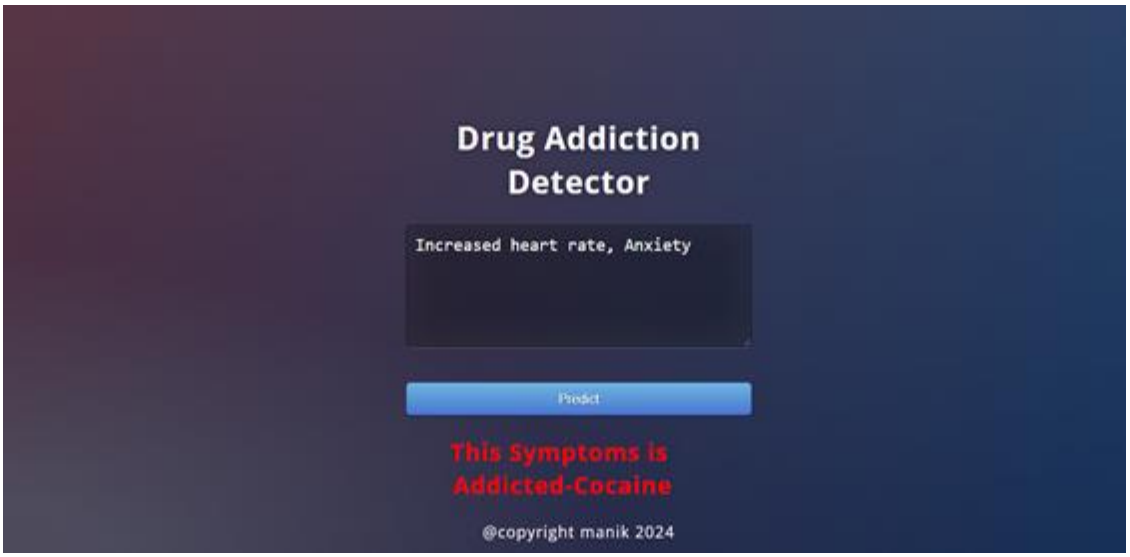


Fig 4.10: Addicted-Cocaine prediction using web application.

Fig 4.11: Addicted-Cannabis prediction using web application.

The drug class recognition process is explained by the ML DC classifiers approach and the predicted Drug Addition features in the data labeling, as can be shown in Figs 4.13 further down. In addition to the eight kinds of pollutants found in the natural world category, the two types of misleading information that students can recognize are included in the following table. Machine learning-based algorithms are a wonderful alternative to the more popular detection or forecasting approaches since they can learn unsupervised. It was quite successful, especially when compared to more conventional methods.

```
[106] # Preprocess new text data (replace 'new_text_data' with your actual new text data)
      new_text_data = ["Slurred speech, Memory loss",'Relaxation, Impaired memory','Increased heart rate, Anxiety','Euphoria, Drowsiness','Drowsiness, Tolerance']

      new_text_data_processed = [clean_text(text, CONTRACTION_MAPPING) for text in new_text_data]
      new_text_data_sequences = tokenizer.texts_to_sequences(new_text_data_processed)
      new_text_data_padded = pad_sequences(new_text_data_sequences, maxlen=X_SEQ_LEN)

      # Use the trained model for predictions
      new_text_predictions = text_clf_decision.predict(tokenizer.sequences_to_texts_generator(new_text_data_padded))

      # Decode the predicted labels to class names
      predicted_class_names = encoder.inverse_transform(new_text_predictions)

      # Check the predicted class for each text
      for text, predicted_class in zip(new_text_data, predicted_class_names):
          print(f"Symptoms: '{text}' Means '{predicted_class}'.")

Symptoms: 'Slurred speech, Memory loss' Means 'Addicted-Alcohol'.
Symptoms: 'Relaxation, Impaired memory' Means 'Addicted-Cannabis'.
Symptoms: 'Increased heart rate, Anxiety' Means 'Addicted-Cocaine'.
Symptoms: 'Euphoria, Drowsiness' Means 'Addicted-Heroin'.
Symptoms: 'Drowsiness, Tolerance' Means 'Addicted-Prescription Opioids'.
```

Fig 4.12: Drug addiction predictor.

4.4 Summary

An assigned symptom is used to gauge the drug addiction label. This section discusses this process. The steps involved in developing a model included choosing a model, gathering and evaluating data, eliminating null values, enhancing the text with additional columns, and evaluating the model's performance in relation to the findings about drug addiction level recognition. The findings of the experiment are analyzed and presented in this portion of the article. Nevertheless, we find that the dependability level using many approaches is rather high. The work that each of us have been doing is visually summarized in this way. The memory, precision, F1 score, supports, heat maps, and other data are displayed in these graphics. Here, we use our data to determine, depending on each drug addiction category's symptom, which predicted drug class is most frequently used.

CHAPTER 5
Engineering standards and Design Challenges

5.1 Compliance with the standards

Adherence to pertinent criteria is crucial while creating a machine learning-based drug addiction detection system in order to guarantee precision, security, and moral rectitude. Generally speaking, important standards cover topics like machine learning procedures, data privacy, and medical device regulation.

5.2 Impact on Society, Environment and Sustainability

Machine learning predicts drug and alcohol addiction, enabling early intervention and healthy community development. It empowers the next generation by treating addiction's underlying causes, improving quality of life, and reducing crime and healthcare pressure, promoting a sustainable and peaceful society.

5.2.1 Impact on Life

The use of machine learning models to forecast the danger of alcohol and drug addiction will benefit society. Because humans are social creatures, members of various classes and faiths must coexist in a community. Along the journey, a drug addict poses a challenge. Drug addiction may strike anyone at any time. As we saw in the previous session, many people take drugs to cope with trauma or as a method to quell curiosity about drugs or to avoid the company of other drug users. This is the method by which our society's younger generation is gradually becoming drug addicts. Children need to be taken care of and their parents should constantly be aware of them. It is the duty of parents to spend time with their children, treat them with kindness, and keep an eye on their actions. When a parent is unsure, they may utilize this approach to gather the necessary facts and data to determine whether their child is at risk of developing a drug addiction. We can prevent drug addiction in our young culture by doing this. We believe that everyone in society will benefit from the use of our drug addiction prediction model.

5.2.2 Impact on Society & Environment

Without a doubt, our model does not harm the environment. This model does not require any chemicals, combustibles, or organic acids to function. As a result, neither biodiversity nor the environment will suffer from this paradigm. Different kinds of plastic are used by alcohol and drug abusers to serve narcotics. Our environment is in danger because of all these plastics and waste products. By using this strategy, drug usage will be prevented and the number of drug addicts will decline. The quantity of plastics required to provide drugs will decline even if the number of drug addicts declines. Our environment will undoubtedly benefit from decreased plastic use.

5.2.3 Ethical Aspects

This addiction prediction algorithm does not infringe upon human rights or act in an anti-moral manner. There won't be a privacy issue because the model doesn't gather any name, identification, or other personal information. Rather of infringing upon someone's freedom to use or enjoy something, this model helps raise awareness. The risk of drug and alcohol addiction prediction model was developed with consideration for all kinds of regulations as well as concerns about confidentiality and privacy. Therefore, the model for predicting drug addiction can be controlled without any issues utilizing machine-learning technology.

5.2.4 Sustainability Plan

The three components of the sustainability strategy are organizational, financial, and community. The Sustainability Plan provides us with a practical understanding of how the project will operate in the future. The goal of our model cohort is to identify drug addiction tendencies. The goal of this model must be to make it simple for individuals to adapt, and it's critical to remember that using it does not imply inferiority complex. This strategy can help law enforcement, police, and drug control agencies function more quickly.

5.3 Project Management and Financial Analysis

Thesis Management and Financial Analysis focusses on planning, monitoring, and effectively conducting the research process while guaranteeing optimal resource allocation. Successfully achieving the project's financial and academic goals requires budgeting, expense tracking, and cost analysis. Our project is related to one that makes use of machine learning. The financial analysis for our project has to be provided in a part of this section. Table 5.1 below

gives an overview of the approximate costs for the different components of the machine learning project.

Table 5.1: Estimated Cost for drug addiction prediction.

SN.	Components	Estimated Cost (BDT)
01	Visiting Stakeholders	600-1000
02	Software's and Tools	1500-2000
03	Data Collection and Processing	500-1000
04	Report Writing	600-1000
05	Contingency (10% of total)	1500-2000
	Total Estimated Cost	4,700-7,500

5.4 Complex Engineering Problem

Complex engineering problems contain extensive and multidimensional obstacles that necessitate significant technical knowledge, critical thinking, and novel approaches. These issues frequently entail competing demands, multidisciplinary factors, and substantial societal, environmental, or financial ramifications.

5.4.1 Complex Problem Solving

Establish a mapping with categories for problem solution in this area. To provide justification, provide subsections for every mapping (see Table 5.1).

Table 5.2: Mapping with complex problem solving.

EP1 Dept of Knowledge	EP2 Range of Con-flicting Requirements	EP3 Depth of Analysis	EP4 Familiarity of Issues	EP5 Extent of Applicable Codes	EP6 Extent of Stake-holder Involvement	EP7 Interdependence
√	√	√	√	√		√

Mapping with Knowledge Profile for EP1

This table 5.2) is designed to map the EP1 to the Knowledge Profile.

Table 5.3: Mapping with knowledge Profile.

K3 Engineering Fundamentals	K4 Specialist Knowledge	K5 Engineering Design	K6 Engineering Practice	K8 Research Literature
				√

5.4.2 Engineering Activities

Provide an engineering activity mapping in this section. To provide justification, provide subsections for every mapping (see Table 5.4).

Table 5.4: Mapping add subsections to put rationale.

Engineering Activity	Rationale
1. Data Collection	Our entire dataset is sourced from **Divisional Drug Addiction Treatment Centre, Department of Narcotics Control, Rajshahi**. This whole real-time data set is used to forecast drug addiction. This company lacks a large, comprehensive dataset since it is difficult to collect data for the specific addicted substance of high-quality analyzers and classification type.
2. Data Preprocessing	Each item of information was looked at separately once all practical means of obtaining data had been used. There are many instances of bad and ambiguous language all around us. It is advised that we go over the final piece of the dataset before utilizing it.
3. Model Development	We decide on a prediction approach, train it using my data, and then assess it to increase reliability. In the field of machine learning, several filters are used. Even though several designs were used to enhance the component design and enable the machine learning model to detect the kind of drug addiction, only one instrument was ultimately selected to evaluate the data's dependability.

4. Model Evaluation and Testing	This phase's later sections address all the ramifications. These methodologies gave us a limited degree of consistency for the label groups of the eight distinct medication datasets after the training and evaluation phase. Accuracy data and f1 scores were generated to support the confusion matrices. This section provides a description of each result. These tactics did not provide us with sufficient dependability for the next two courses, even after testing and training. They produced a method for categorizing various therapeutic plants as well as visual aids for f1 measurement, recall, efficiency, and confusion matrix.
5. Deployment and Integration	Deploying the model into a user-friendly platform (e.g., web or mobile app) enables real-time predictions. Security measures like encryption and access control are critical for protecting sensitive health data.
6. Continuous Monitoring and Maintenance	Ongoing monitoring ensures the model remains accurate over time. Regular updates based on new data and performance feedback help maintain the model's relevance and compliance with evolving standards.

Table 5.5: Mapping with complex engineering activities.

EA1 Range of re-sources	EA2 Level of Interaction	EA3 Innovation	EA4 Consequences for society and environment	EA5 Familiarity
√	√	√	√	√

5.5 Summary

The report discusses the development of a machine learning-based drug addiction prediction system, focusing on societal, environmental, and ethical aspects. The system aims to mitigate addiction's adverse effects, improve quality of life, and reduce societal pressures like crime and healthcare costs. It also emphasizes environmental benefits by reducing harmful materials associated with drug abuse. Ethical safeguards ensure user privacy and compliance

with human rights. The project, estimated to cost BDT 4,500-7,500, involves meticulous data collection, preprocessing, model development, evaluation, and deployment. The goal is to promote a healthier, more sustainable society using technology for social good.

CHAPTER 6
Conclusion

6.1 Summary

Drug addiction is a complicated, multidimensional problem that has an impact on people, families, and communities. In order to treat the underlying causes of addiction and promote recovery, prevention techniques, early intervention, and availability to high-quality treatment choices are essential. Furthermore, lowering stigma and raising awareness can help create an atmosphere that is more encouraging for individuals who are impacted by addiction. Prioritizing community support, education, and resource accessibility can help us lessen the effects of drug addiction and encourage healthy lifestyles for all. It is feasible to ascertain the likelihood of drug addiction using our suggested model. Once this model is complete, we expect that the general public will be able to utilize it with ease and recognize its significance in spreading awareness. Being watchful at all times is crucial to avoiding the dangers of drugs and preventing drug addiction. We expect that by using this approach, individuals would avoid drug exposure and become conscious of their circumstances and take self-control. We exclusively get our data from agency-affiliated addicts. We collect 21 features and 4000 real data. Data on 8 kinds of drug addicts has been gathered. We used machine learning techniques to the previously processed dataset. Given that several methods for detection and prediction make use of deep learning, artificial intelligence, and machine learning. We use XG Boost, naïve Bayes, decision trees, random forests, support vector machines (SVM), and k-nearest neighbor (KNN). With a classifier accuracy of 97.75%, decision tree models outperformed the other six methods we tested in our experiment. Next, develop a web application that uses a decision tree model to forecast different drug addictions based on their symptoms.

6.2 Limitations

There are limitations to our research. We are unable to gather a large amount of data. Although we gathered 4000 pieces of data, it is insufficient. It will be more advantageous if we get more data. Another drawback is that we are unable to get a significant amount of female data. We can predict if a female is drug addicted or not more accurately if we gather data from more females. Our research focuses on drug prediction systems that use machine

learning algorithms and addiction. There are certain restrictions and shortcomings in our model and approach. A larger and richer data collection would have been preferable to the very small one we chose. Certain restrictions prevented individuals from different classes, districts, and occupations from gathering data. Data processing might also be done using a variety of sophisticated techniques, and the model could be elegantly shown by utilizing many algorithmic variants.

6.3 Future Work

Our lives are now quicker and easier thanks to technology and contemporary science. We wish to continue using information technology and the internet in our nation by incorporating our model into software, a web application, or an Android application in the future. With a larger database, we will be able to improve our model's accuracy in the future. Furthermore, the model's software may be made accessible to the public by developing graphical user interfaces. The model may be improved in the future by introducing new features, altering its settings, and adding new algorithms. By gathering information from various groups of individuals based on the district, a strong database may be produced in the future.

Reference

[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: Dec. 11, 2024].

[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] A. A. Choudhury, Md. R. H. Khan, N. Z. Nahim, S. R. Tulon, S. Islam, and A. Chakrabarty, “Predicting Depression in Bangladeshi Undergraduates using Machine Learning,” 2019 IEEE Region 10 Symposium (TENSYMP). IEEE, Jun. 2019.

[4] Faisal, Fahim, et al. “A supervised machine learning approach to predict vulnerability to drug addiction”. Diss. Brac University, 2019.

[5] F. Hammann, H. Gutmann, N. Vogt, C. Helma, and J. Drewe, “Prediction of Adverse Drug Reactions Using Decision Tree Modeling,” Clinical Pharmacology & Therapeutics.

[6] E. Sahker, L. Acion and S. Arndt,” National Analysis of Differences Among Substance Abuse Treatment Outcomes: College Student and Nonstudent Emerging Adults”, Journal of American College Health.

[7] M. Myers, J. Kelly, “Cigarette Smoking Among Adolescents With Alcohol and Other Drug Use Problems”, Alcohol research & health: the journal of the National Institute on Alcohol Abuse and Alcoholism, vol. 29, no. 3, pp. 2217,2006.

[8] A. Shahriar, F. Faisal, S. U. Mahmud, A. Chakrabarti, and M. G. Rabiul Alam, “A Machine Learning Approach to Predict Vulnerability to Drug Addiction,” 2019 22nd International Conference on Computer and Information Technology (ICCIT). IEEE.

[9] A. Amirabadizadeh, H. Nezami, M. G. Vaughn, S. Nakhaee, and O. Mehrpour, “Identifying Risk Factors for Drug Use in an Iranian Treatment Sample: A Prediction Approach Using Decision Trees,” Substance Use & Misuse. Informa UK Limited.

[10] Yupu Zhang, Jinhai Liu, Zhihang Zhang, Junnan Huang(2021) “Prediction of Daily Smoking Behavior Based on Decision Tree Machine Learning Algorithm”.

[11] M. H. Afzali, M. Sunderland, S. Stewart, B. Masse, Jean Seguin, Nicola Newton, Maree Teesson, Patricia Conrod(2019). Machine-Learning Prediction of Adolescent Alcohol Use: A Cross-Study, Cross-Cultural Validation

[12] Ahnaf Atef Choudhury, Md Rezwan Hassan Khan, Nabuat Zaman Nahim, Sadid Rafsun Tulon, Samiul Islam, Amitabha Chakrabarty(2020). “Predicting Depression in Bangladeshi Undergraduates Using Machine Learning”.

[13] Arif, Md. Ariful Islam; Sany, Saiful Islam; Nahin, Faiza Islam, ”Drug Addiction Prediction Using Machine Learning”.

[14] 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.

[15] 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.

[16] 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,".

[17] 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.

[18] 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.

[19] 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.

[20] 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).

[21] 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).

[22] 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.

[23] 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).

[24] 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,

[25] 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).

[26] 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).

[27] 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.

[28] H. Gong, C. Xie, C. Yu, N. Sun, H. Lu, and Y. Xie, "Psychosocial Factors Predict the Level of Substance Craving of People with Drug Addiction: A Machine Learning Approach," International Journal of Environmental Research and Public Health.

[29] M. R. Lee et al., "Using Machine Learning to Classify Individuals With Alcohol Use Disorder Based on Treatment Seeking Status," EClinicalMedicine.

[30] E. Sahker, L. Acion, and S. Arndt, "National Analysis of Differences Among Substance Abuse Treatment Outcomes: College Student and Nonstudent Emerging Adults," Journal of American College Health.

[31] L. Acion, D. Kelmansky, M. van der Laan, E. Sahker, D. Jones, and S. Arndt, "Use of a machine learning framework to predict substance use disorder treatment success," PLOS ONE.

[32] W. Ash-Houchen and C. C. Lo, "Racial/Ethnic Differences in Illicit Substance Use: A Temporal Ordered Test of General Strain Theory," Journal of Drug Issues. SAGE Publications.

[33] L. Zhang et al., "Applications of Machine Learning Methods in Drug Toxicity Prediction," Current Topics in Medicinal Chemistry.

[34] J. N. Weinstein et al., "Neural Computing in Cancer Drug Development: Predicting Mechanism of Action," Science.

[45] Y. Chen, S. S. Thosar, R. A. Forbess, M. S. Kemper, R. L. Rubinovitz, and A. J. Shukla, “Prediction of Drug Content and Hardness of Intact Tablets Using Artificial Neural Network and Near-Infrared Spectroscopy,” Drug Development and Industrial Pharmacy.
[36] Unknown Author, "Illustration for Tokenization in NLP," Image, Medium, [Online]. Available: https://miro.medium.com/v2/resize:fit:640/format:webp/0*PQsctgROiuiJRs0S.png. [Accessed: Dec. 11, 2024].
[37] Unknown Author, " The working of a naive bayes," Image, Medium, [Online]. Accessed: Dec. 11, 2024].
[38] Unknown Author, "The working of a random forest," Image, ResearchGate, [Online]. Available: https://www.researchgate.net/publication/384513281/figure/fig1/AS:11431281281361826@1727807125706/The-working-of-a-random-forest.png. [Accessed: Dec. 11, 2024].
[39] Unknown Author, "A schematic of the XGBoost workflow," Image, ResearchGate, [Online]. Available: https://www.researchgate.net/publication/338195121/figure/fig6/AS:840744516452353@1577460521513/A-schematic-of-the-XGBoost-workflow-The-shaded-area-indicates-the-data-and-its.png. [Accessed: Dec. 11, 2024].
[40] Unknown Author, " The working of a knn," Image, GeeksforGeeks, [Online]. Available: https://media.geeksforgeeks.org/wp-content/uploads/20200616145419/Untitled2781.png. [Accessed: Dec. 11, 2024].
[41] Q. Suhair, "The principal operation of SVM classifier," Image, ResearchGate, [Online]. Available: https://www.researchgate.net/profile/Qudsieh-Suhair/publication/373389772/figure/fig4/AS:11431281183900032@1693109152244/The-principal-operation-of-SVM-classifier-40.png. [Accessed: Dec. 11, 2024].
[42] O. Takawira, "Example of a structure of decision tree," ResearchGate, [Online]. Available: https://www.researchgate.net/profile/Oliver-Takawira/publication/360009739/figure/fig1/AS:1145931231694848@1650222704166/Example-of-a-structure-of-decision-tree-Source-Charbuty-et-al-2021-A-decision-tree-is.png. [Accessed: Dec. 11, 2024].
[43] A. McCallum and K. Nigam, "A comparison of event models for naive Bayes text classification," in Proceedings of the AAAI-98 Workshop on Learning for Text Categorization, 1998, pp. 41–48.

[44] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, Oct. 2001.
[45] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, Oct. 2001.
[46] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21–27, Jan. 1967.
[47] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, 20, no. 3, pp. 273–297, 1995.
[48] J. R. Quinlan, "Induction of decision trees," Machine Learning, vol. 1, no. 1, pp. 81–106, 1986.

Asgar Report Checking V7

ORIGINALITY REPORT

13%	7%	5%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	3%
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	2%
3	Submitted to BRAC University Student Paper	2%
4	Dinesh Goyal, Bhanu Pratap, Sandeep Gupta, Saurabh Raj, Rekha Rani Agrawal, Indra Kishor. "Recent Advances in Sciences, Engineering, Information Technology & Management - Proceedings of the 6th International Conference “Convergence2024” Recent Advances in Sciences, Engineering, Information Technology & Management, April 24–25, 2024, Jaipur, India", CRC Press, 2025 Publication	<1%
5	H.L. Gururaj, Francesco Flammini, S. Srividhya, M.L. Chayadevi, Sheba Selvam. "Computer Science Engineering", CRC Press, 2024 Publication	<1%