Title of the Thesis

A Comparative Analysis of Machine Learning Approaches for Cardiovascular Disease Prediction

Submitted By

Taniya Jannatul Hafsa Mim
ID: 201-16-500
Department of Computing & Information System
Daffodil International University

Supervised By

Md. Mehedi Hassan
Lecturer
Convener, Exam Committee
Department of Computing & Information System
Daffodil International University

Department of Computing and Information System
Daffodil International University.
Dhaka, Bangladesh

Submission Date: 23/01/2024

Approval

This Thesis titled **"A Comparative Analysis of Machine Learning Approaches for Cardiovascular Disease Prediction"**, Submitted by **Taniya Jannatul Hafsa Mim**, ID No: **201-16-500,** to the Department of Computing & Information Systems, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computing & Information Systems and approved as to its style and contents. The presentation has been held on 23-01-2024.

BOARD OF EXAMINERS

23/1/24

Md Sarwar Hossain Mollah — **Chairman**
Associate Professor and Head
Department of Computing & Information Systems
Faculty of Science & Information Technology
Daffodil International University

Dr. Shekh Abdullah-Al-Musa Ahmed — **Internal Examiner**
Assistant Professor
Department of Computing & Information Systems
Faculty of Science & Information Technology
Daffodil International University

Md. Nasimul Kader — **Internal Examiner**
Assistant Professor
Department of Computing & Information Systems
Faculty of Science & Information Technology
Daffodil International University

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

Declaration

I, Taniya Jannatul Hafsa Mim. hereby declare that the work in this dissertation tilled "**A Comparative Analysis of Machine Learning Approaches for Cardiovascular Disease Prediction**"; I have done this thesis under the supervision of Md. Mehedi Hassan. Lecturer, Convenor Exam Committee, Department of Computing and Information System (CIS) of Daffodil International University. I am also declaring that this thesis or any part of there has never been submitted anywhere else for the award of any educational degree like B,Sc., M.Sc., Diploma or other qualifications.

Supervised By

23.01.2024

Md. Mehedi Hassan
Lecturer, Convener, Exam Committee
Department of CIS
Daffodil International University

Submitted By

Mim 23-1-24

Taniya Jannatul Hafsa Mim
ID: 201-16-500
Department of CIS

Acknowledgements

It was my pleasure to research and complete the thesis paper on "**A Comparative Analysis of Machine Learning Approaches for Cardiovascular Disease Prediction**". I am grateful to Allah for his grace in my education.

I want to express my gratitude to my supervisor**, Md. Mehedi Hassan,** Lecturer, Convener, Exam Committee in the CIS Department at Daffodil International University, for his valuable time, advice, and inspiration to give this paper and my education my all. I want to thank my sir and send him my best wishes for his belief, steadfast guardianship, sound judgment, suitable outlook, and comprehension of various factors that helped make my thesis possible. He has done everything he can to ensure I know what is necessary to finish this thesis.

I also want to convey my heartfelt respect to **Mr. Sarwar Hossain Mollah**, the head of my department, for his helpful help at the beginning of the study and for making sure I understood the topic. His kind behaviors and helpful nature always encourage me to do better work.

Finally, I must respectfully thank to all of my lecturers of the same department for their unfailing help and patience in getting this thesis work done.

Dedication

I wholeheartedly dedicate this work to my cherished parents, **Jahangir Alam** and **Rabeya Begum**. Their boundless love, encouragement, and unwavering support have been my pillars of strength throughout this research journey. Their sacrifices and guidance have not only made this opportunity to study and delve into research possible but have also fueled my determination to make a meaningful and positive impact on society. This work stands as a testament to the values they instilled in me, and I am profoundly grateful for their enduring presence in my life.

Abstract

Cardiovascular disease (CVD) is a major global health issue with a particularly noticeable impact in Bangladesh. Cardiovascular disease (CVD) is a major global cause of death and disease that affects the heart and blood vessels, with serious consequences such as heart attacks, strokes, and heart failure. Understanding the seriousness of the condition and taking early identification and preventative action are essential to reducing the negative impacts of CVD. In an effort to lower mortality due to CVD, many researchers are attempting to predict CVD using machine learning models. However, very few researchers have actually employed real world data. This study compared several machine learning methods that can reliably predict the occurrence of cardiovascular disease based on actual patient records from Bangladesh's healthcare system. I have collected data for my study from two well-known hospitals in Dhaka, Bangladesh. There are 1019 instances and 9 attributes in my dataset. Six machine learning models have been used: K-Nearest Neighbor, XGBoost, Decision Tree, Random Forest, Support Vector Machine and Logistic Regression. After using stratified 5-fold cross validation, my model XGBoost demonstrated improved performance in both training (84.22%) and testing (86.05%). Several performance measurement techniques, including the ROC curve, precision, recall, and F1-score have been used to assess the performance of my models.

Preface

This bachelor thesis is centered around predicting cardiovascular disease using data collected from two hospitals in Dhaka, Bangladesh. The primary objective is to develop a predictive model for cardiovascular disease, utilizing insights gained from the analysis of datasets obtained from the hospitals. This work aims to contribute to the advancement of cardiovascular health research and diagnostics in the context of the local healthcare landscape.

This thesis includes 6 chapters which are briefed as follows:

Chapter-1

This chapter provides a brief overview of cardiovascular disease, outlining its global and local impact, specifically in Bangladesh. It discusses the motivations behind the research, highlights its contributions, and outlines the objectives and anticipated outcomes.

Chapter-2

Chapter 2 delves into a comprehensive examination of related work on cardiovascular disease, exploring both global and Bangladesh-specific perspectives. The literature review synthesizes existing research, identifying key findings, scopes, and challenges in the field.

Chapter-3

This chapter provides an overview of the study's methodology, including preprocessing procedures, a thorough explanation of the data properties, and data collecting from two hospitals in Dhaka.

Chapter-4

Chapter 4 shows the outcomes of applying various machine learning models to the dataset. It provides a detailed examination of the results obtained from each model and employs diverse evaluation metrics for performance assessment. The chapter includes a comprehensive comparison table, offering insights into the effectiveness of different algorithms in predicting cardiovascular disease.

Chapter-5

Conclusion of the research.

Chapter-6

Future work and References.

Contents

List of Figures

List of Tables

List of Abbreviations

CVD – Cardiovascular disease

ML – Machine learning

NCD – Non communicable disease

SVM – Support Vector Machine

RF – Random Forest

KNN- K-Nearest Neighbor

XGB- Extreme Gradient Boosting

LR- Logistic Regression

DT- Decision

Chapter 1

Introduction

1.1 Introduction

Cardiovascular disease (CVD) poses a significant global health threat, emerging as a leading cause of both mortality and morbidity on a worldwide scale. The World Health Organization (WHO) reports that heart diseases claimed the lives of approximately 17.9 million individuals in 2019, constituting 32 percent of all global deaths [1]. Bangladesh, which is dealing with changes in lifestyle, increased stress, and changing food habits, is not immune from this worrying situation. A 2018 non-communicable disease risk factor survey indicates that 15.5 percent of Bangladeshis aged 40-69 face a risk of cardiovascular diseases, with a 2021 Bangladesh Bureau of Statistics (BBS) report revealing a rise in cardiac arrest-related deaths in 2020 compared to 2019 [2]. The report, titled "Monitoring the Situation of Vital Statistics of Bangladesh," discloses that heart attacks caused 21.1% of total deaths in 2020, up from 17.9% in 2019 [2]. While chronic diseases like CVD, cancer, chronic lung disease, and diabetes are on the upswing, concerns about the escalating healthcare burden are growing.

The increase in non-communicable diseases (NCDs) over the world is a serious public health concern that claims a large number of lives each year. Cardiovascular diseases (NCDs)—which include strokes and coronary heart disease—are the leading causes of death worldwide, particularly in low- and middle-income nations like Bangladesh in South Asia. Because of the early development of these diseases and the country's enormous population, the nation has a significant burden of CVD. The main illness pattern in Bangladesh has shifted over the past few decades from predominantly communicable to predominantly non-communicable diseases due to epidemiological transition; cardiovascular disease (CVD) plays a major role in this one [3]. In Bangladesh, NCDs account for a substantial portion of total deaths, with CVDs alone contributing to approximately 30% of the mortality [4].

Many diseases that affect the heart and blood arteries and cause arterial damage to important organs are commonly referred to as cardiovascular diseases. Even though many CVDs can be avoided by making lifestyle changes including giving up smoking, eating a healthy diet, exercising, and controlling obesity, early detection and prevention of CVD is still essential.

Over 165 million people live in Bangladesh, a developing nation in South Asia with a highly dense population [5]. According to reports, the country has a prevalence of 4.5–5.0% of CVDs [6, 7], accounting for almost 31% of all deaths globally [8]. This poses a substantial challenge to the country's healthcare system, necessitating efficient methods for early detection and prediction. Traditional testing systems often prove to be inefficient, costly and time-consuming. In response to these challenges, the integration of computer-based methods, particularly machine learning, emerges as a crucial asset in enhancing predictive capabilities for cardiovascular diseases.

Machine learning (ML) is a technology that can learn from data and make predictions or decisions. ML can help in healthcare by providing accurate and reliable diagnosis, prognosis, treatment, and management of diseases.

A number of noteworthy research papers in the field of medicine have investigated the use of machine learning algorithms to forecast common diseases. For example, in 2018 Ramalingam et al. conducted a performance analysis and offered a survey of several models based on these methods and techniques [9]. The model that demonstrated superior accuracy of 90.78% is K-nearest neighbor. These studies highlight the potential of machine learning techniques in early detection and prevention of CVD.

In this research, I have used ML to predict CVD based on different types of data, such as age, gender, blood pressure, cholesterol, glucose level, physical activity, smoking, etc. I found the best model based on my data to predict whether the patient have CVD or not. I have tested and evaluated my ML technique. I have also addressed the specific challenges posed by cardiovascular diseases in Bangladesh by adopting a machine learning-based approach.

1.2 Research Objective

The objectives of this study are-

- Investigate the impact of various data types, including age, gender, blood pressure, cholesterol, glucose level, physical activity, and smoking, on CVD prediction.
- Using a variety of patient data, evaluate and compare various machine learning models to determine which one is the best at predicting CVD.
- Address the specific challenges associated with cardiovascular diseases in the context of Bangladesh's healthcare environment.
- Provide valuable insights to healthcare professionals and policymakers for early detection and prevention strategies.

1.3 Motivation

Nowadays heart problem is increasing rapidly in Bangladesh. This research is focused on addressing this growing health concern in Bangladesh. Unhealthy lifestyle and eating habits, like smoking, consuming oily foods, are leading to issues like high cholesterol and diabetes, contributing to a rise in heart-related problems in the country. I have conducted this study to find a better way to predict and detect heart problems early. By using machine learning, I aim to identify potential heart issues sooner, ultimately saving lives and promoting better health in Bangladesh.

1.4 Rationale of the Study

After extensive review of existing literature, it has come to my attention that there is a significant gap in cardiovascular disease (CVD) prediction research using machine learning in Bangladesh. While some researchers have explored this domain, a limited number have utilized real patient data. Notably, my research stands out by collecting data from two hospitals, making it unique in the context of Bangladesh. Conducting research on real patient data is paramount, considering it has not been extensively explored in the country. This uniqueness and the utilization of authentic patient data can contribute valuable insights to the field of CVD prediction.

Chapter 2

Literature Review

2.1 Introduction

In this chapter, I'll explore existing research on cardiovascular disease (CVD) prediction, comparing methodologies, findings, and accuracy across different studies. I'll highlight my approach and discuss the challenges I faced, along with the valuable lessons learned during my study.

2.2 Related Work

In an effort to accurately assess overall risk, a great deal of research has been done to predict cardiovascular diseases and identify the critical risk factors for heart disease. In this field, integrating machine learning techniques is essential. Using a variety of global datasets, researchers have investigated a wide range of machine learning techniques for the early identification and prevention of cardiovascular illnesses. Significant progress in the early prevention of cardiovascular disease has resulted from their efforts. Nonetheless, there is a chance for improvement if real data obtained directly from different healthcare facilities in diverse areas is used to train the model. It is important to acknowledge the important contributions made by earlier researchers while also being aware of the potential limits associated with relying on datasets, such those seen on Kaggle. These datasets, which are frequently used because of their accessibility and scope, might not accurately represent the complex patterns found in actual patient data that is gathered from hospitals. By using real, location-specific patient data, predictive models may become more accurate and useful by providing a more thorough understanding of the various ways that cardiovascular illnesses present.

In 2021, Weicheng et al. investigated the use of machine learning techniques for cardiovascular disease prediction [10]. Their dataset, which consists of 70,000 data points across 12 variables, was obtained from the Svetlana Ulianova research group. They used machine learning techniques including Support Vector Machines (SVM), Random Forest (RF), and Logistic Regression (LR) to perform 5-fold cross-validation in order to predict CVD. SVM performs better, with an accuracy of 71.48% and an AUC of 78.84%. They applied various feature integration techniques, such as summing, averaging, geometric mean, and selecting the maximum value, which are used for combining feature vectors. The study mainly incorporates correlation coefficient methods for feature extraction.
In 2020, Syed et al. used deep learning methods to forecast cardiovascular disease [11]. They used machine learning methods and deep learning techniques, including Tensor Flow Keras, K-Nearest Neighbor (KNN), Decision Trees (DT), Support Vector Machines (SVM), and Artificial Neural Networks (ANN). The ANN generated a greater accuracy of 85.24% out of these. They obtained the information they needed from Kaggle.
In 2020, Atharv et al., used machine learning methods to predict CVD, with a particular focus on BMI [12]. They also talked about how to deal with outliers and missing values, among other things. The dataset has 12 attributes: Alcohol Intake, Systolic Blood Pressure, Smoking, Diastolic Blood Pressure, Age, Blood Pressure, Height, Weight, Physical Activity, Gender, and Cholesterol. The BMI features were implemented by computing Height and Weight. Several algorithms were used. XGB, Decision Tree, K-Nearest Neighbor, Naive Bayes, Neural Network, Logistic Regression, and Light GBM classifiers are a few of these. The decision tree classifier performed more accurately in training and testing (72.68% and 73.13%).
In their model for predicting CVD, Ibashisha et al., used a set of procedures [13]. The steps include pre-processing the data, exploring the dataset to demonstrate the effects of various attributes (such as gender, age, cholesterol, glucose, smoking, and alcohol consumption) on disease, selecting and extracting features for better accuracy and disease Prediction using Machine Learning model. They also

added risk factors (such as blood pressure and BMI) to increase accuracy. KNN outperformed the other algorithms they used, including Naive Bayes (NB), Decision Trees (DT), Random Forests (RF), Support Vector Machines (SVM), and Linear Discriminant Analysis (LDA).They employed K-fold cross-validation with various measures, including accuracy, precision, recall, and F1-score, for evaluation.

Using medical data, Vansh et al., created a machine learning model to forecast incidences of cardiovascular disease [14]. They included this prediction model into an iOS application for mobile devices. Users can enter their information, and the system will automatically run it through the model and provide a result, enabling them to emphasize efficiency on the platform and take the necessary preventive action. XGBoost performed better from their model, with a score of 72.70%.

Abdul et al., tested nine different classifiers on heart disease datasets [15]. While Support Vector Machine (SVM) performed the best, they found common issues in previous systems. One problem was that these systems became less efficient when dealing with larger datasets, leading to slower processing. Additionally, they struggled to classify datasets effectively, especially when faced with complex or diverse data. The models heavily relied on extracting attributes from the data, and if this process wasn't optimized, it negatively impacted the overall performance. The researchers also noted the sensitivity of the classifiers to hyperparameter tuning, meaning their performance depended heavily on finding the right settings. This could make the models less reliable and adaptable to different datasets. Moreover, the models might not generalize well to new, unseen data and could potentially introduce bias or variability in their predictions.

In contrast to other researchers, Arsalan et al., [16] gathered their data from the Lady Reading Hospital (LRM) and the Khyber Teaching Hospital (KTH), which are Pakistan's two biggest teaching hospitals. For classification and prediction, machine learning techniques such as support vector machines (SVM), logistic regression (LR), decision trees (DT), random forests (RF), and Naïve Bayes (NB) were used. The study used metrics like the recursive operating characteristic curve and confusion matrix in its extensive exploratory and experimental studies. The RF method was found to be the most accurate in the performance evaluation, with the following values for CVD: 85.01%, 92.11%, and 87.73% for prediction accuracy, sensitivity, and recursive operative characteristic curve, respectively.

Nonetheless, due to legal restrictions and restricted access to extensive datasets, there is very little field research being done in south Asian nations like Bangladesh. In spite of these challenges, Bangladeshi researchers are making a significant contribution and working to remove barriers in order to improve cardiovascular health outcomes.

Tamanna et al., in 2021 [17] focused on the early prediction of cardiovascular disease using machine learning techniques. The XGBoost algorithm was identified as the best model, achieving an accuracy of 75.10%. For feature selection, the researchers utilized the Random Forest algorithm. The evaluation metrics included accuracy, precision, sensitivity, specificity, and F1-score, with the model demonstrating notable performance across all metrics.

Using the Kaggle dataset, Khairul et al., in 2022 [18] tested four machine learning algorithms: Random Forest, KNN, XGBoost, and Decision Tree. They discovered that the model worked better when the dataset was reduced in size. In order to assess the impact of lifestyle factors on the risk of cardiovascular disease, they extensively narrowed down the dataset by selecting patients who smoke, drink alcohol, and engage in physical activity. They also used a 5-fold cross-validation method on their training set of data. The accuracy was 73.72% before the dataset was reduced. Following the dataset reduction, it reached 81.14%.

Instead of gathering data from internet repositories, Mohammad et al., 2021 [19] did it by interviewing residents of the Sylhet district of Bangladesh, six hospitals, and educational institutions. They employed a cross-validation approach to test the model after removing noisy input. They demonstrated that their model works better with eight qualities than with eighteen. SVM achieved the best overall performance (91%).

Md et al., 2022 [20] used the mean imputation method to address null values. The info-gain strategy of feature selection helps to enhance the performance of the classification model. The research compares the performance of three classifiers—Random Forest, Naive Bayes, and K-Nearest Neighbors (KNN)—comparing results with a reduced dataset of 10 features. Using 10 features, Random Forest notably obtains the highest classification accuracy of 95.63%. Additionally, they used a 10-fold cross-validation method to assess how well the model worked.

Voting ensemble classifiers were utilized by M. Shahiduzzaman et al., 2022 [21] and K-nearest neighbor outperformed them with an accuracy of 75% which outperformed Bernoulli Naive Bayes, Random Forest, Gradient Boosting, and Logistic Regression.

2.3 Bangladesh Perspective

Cardiovascular disease is the main cause of illness and mortality in Bangladesh, and the number of patients has been noticeably rising. The population's general health and well-being are seriously threatened by the escalating tendency. The rising incidence of CVD is a result of unhealthy lifestyle choices like increased tobacco use, inactivity, and poor eating habits. It is imperative to address these factors in order to avoid and manage cardiovascular diseases. Bangladesh's healthcare system might have trouble keeping up with the rising cost of CVD. Early detection and effective preventative interventions can help control the demand on healthcare resources. Significant economic consequences may result from CVD, as it can reduce productivity and place a financial strain on both the healthcare system and individual patients. Therefore, I can state that my will be a great innovation for Bangladesh. By raising awareness and enabling early prevention and detection of cardiovascular disease (CVD), it holds the potential to save numerous lives.

2.4 Significance of the study

In a nation where CVD is becoming more common, a predictive model created with machine learning methods is an innovative and progressive approach. It is important because the predictive model has the ability to enable early detection and thereby changing healthcare outcomes. My research attempts to produce a tool that can predict an individual's risk of acquiring CVD by utilizing machine learning. The potential to anticipate the CVD has significant implications since it facilitates prompt intervention and preventative actions. It has also a huge societal influence. It may lessen the overall burden of CVD on the healthcare system in addition to improving individual health by facilitating early awareness and intervention.

2.5 Challenge of this work

While conducting this research, I ran into a lot of obstacles. The first difficulty I ran across was determining which data attribute to use to predict CVD. Not all features are readily available in Bangladesh. The second difficulty I encountered was obtaining permission for data gathering after selecting the suitable feature. Patient data is kept under strict confidentiality. Permission to collect data is not readily granted by the hospital administration. I was ultimately granted permission to collect data for my thesis project by the National Institute of Cardiovascular Disease and Ibrahim Cardiac Hospital & Research Institute, after completing the necessary documentation. The problem that I had after collecting data was cleaning it. Real data typically have a large number of missing values. Thus, the difficulty I encountered when conducting my research was data cleaning. The next challenge was handling overfitting issue. Overfitting occurs when model is too trained on training data. So, to handle overfitting I have used 5-fold stratified cross validation.

Chapter 3

Methodology

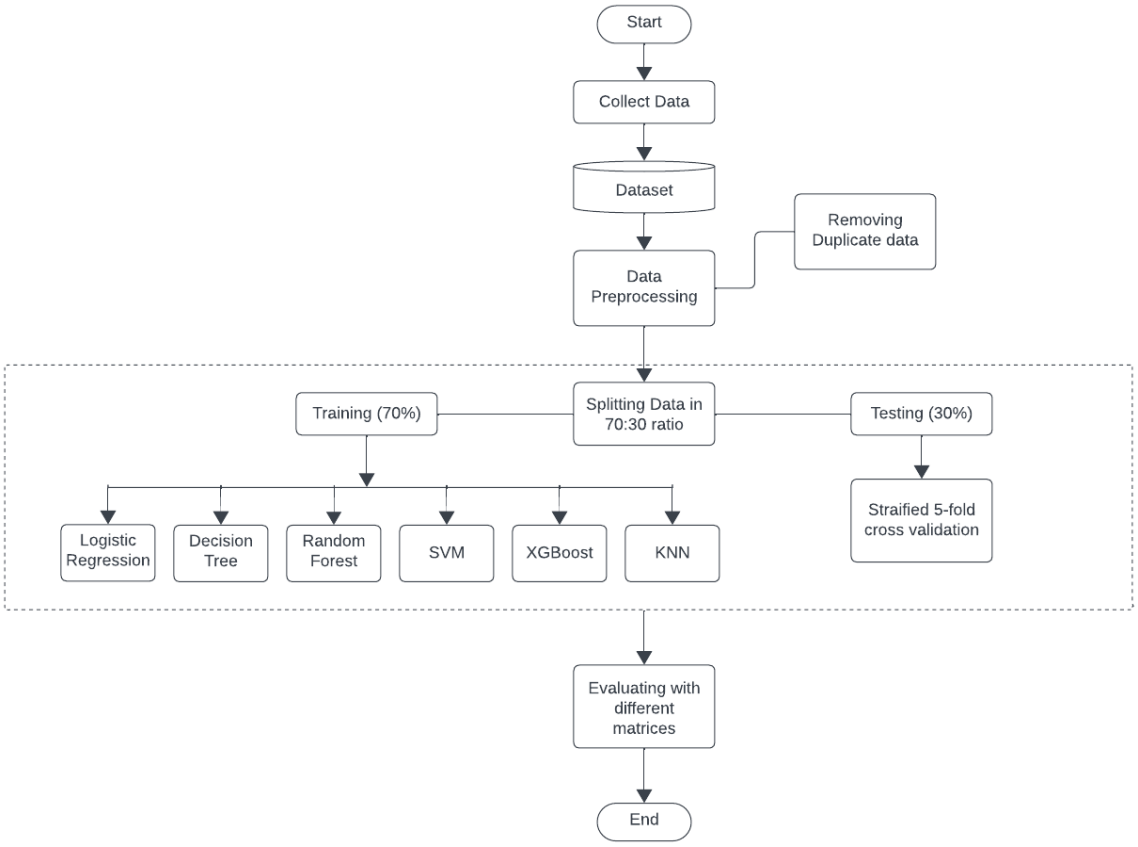


Figure 3. 1: *Working Flow Diagram*

3.1 Data Collection

The majority of the researchers collect data from kaggle and UCI libraries, which is a particularly common technique. Researchers utilize the Kaggle dataset for many reasons. The fact that Kaggle data is readily available and that there are no legal restrictions is one of the primary reasons authors use it. Due to patient privacy concerns, the majority of hospitals do not permit the usage of their patient data. Sensitive information is found in medical records. A few studies in Bangladesh, including Mohammad et al. 2021 [19], used actual patient data that they had collected from hospitals in the Sylhet district of Bangladesh. We attempt to do the same. The **National Institute of Cardiovascular Diseases** and the **Ibrahim Cardiac Hospital & Research Institute** in Dhaka, Bangladesh are two prestigious hospitals from where we obtained our data. There are 8 attributes and 1 target column in the 1019 data that we have collected. The binary value type for the target column is 0 or 1, where 0 indicates that the patient does not have cardiovascular disease and 1 indicates that they have. Our goal is to create a strong model that, given the given characteristics, can reliably determine whether cardiovascular disease is present.

Attribute	Value Type	Variable	Variable Type	Explanation
Age	int	Age	Objective	Age in years
Gender	categorical code	Gender	Objective	Female= 1, Male=2
Systolic Blood Pressure	int	SystolicBP	Examination	Blood Pressure
Diastolic Blood Pressure	int	DiastolicBP	Examination	Blood Pressure
Cholesterol Level	categorical code	CholesterolLevel	Examination	1: Normal, 2: Modeare, 3: High Risk
Glucose Level	categorical code	GlucoseLevel	Examination	1: Normal, 2: Prediabetic, 3: Diabetic
Smoking	binary	Smoking	Subjective	0: Non-Smoker 1: Smoker
Physical Activity	binary	Activity	Subjective	0: Not Active 1: Active
Cardiovascular Disease	binary	Disease	Target Variable	0: Negative 1: Positive

Table 3. 1.1: *Attribute Description*

3.2 Data Pre-Processing

After data collection we have preprocessed our data part. Data that has been gathered might have invalid or missing values. Unorganized, noisy data may be present in the dataset. Hence cleaning the data is necessary. Since noisy, unprocessed data cannot be used to train machine learning models, this step is crucial. As a result, prior to model training, data collection is required. Preprocessed data will provide machine learning models with increased robustness and performance. The dataset consists of 1019 rows and 9 columns (patient records). However, the dataset may contain duplicate records. Thus, the remaining 1001 patient records were used after identical ones were removed. Then we have to find out if our dataset has any missing values or not. After checking we saw that the dataset has no missing values. So our dataset is cleaned. The binary classification is used to investigate if heart disease is present or absent. The patient has cardiovascular disease, as indicated by the result of 1. 0 indicates that the patient does not have cardiovascular disease. We have employed Pandas for data manipulation and used standard scale to scale our data using scikit-learn's StandardScaler.

3.3 Exploratory Data Analysis

As previously stated, our dataset has eight attributes and one target column. These eight characteristics include: age, gender, blood pressure measurements at the diastolic and systolic levels, cholesterol and glucose levels, smoking, physical activity, and cardiovascular disease. The impacts of each dataset attribute will be examined individually in this step.

Age

The Age attribute represents the age of each participant in the study. It is a continuous variable indicating the years of age for each individual. From Figure 3.3.1 and Figure 3.3.2 we can see that the disease increases as the age increases.

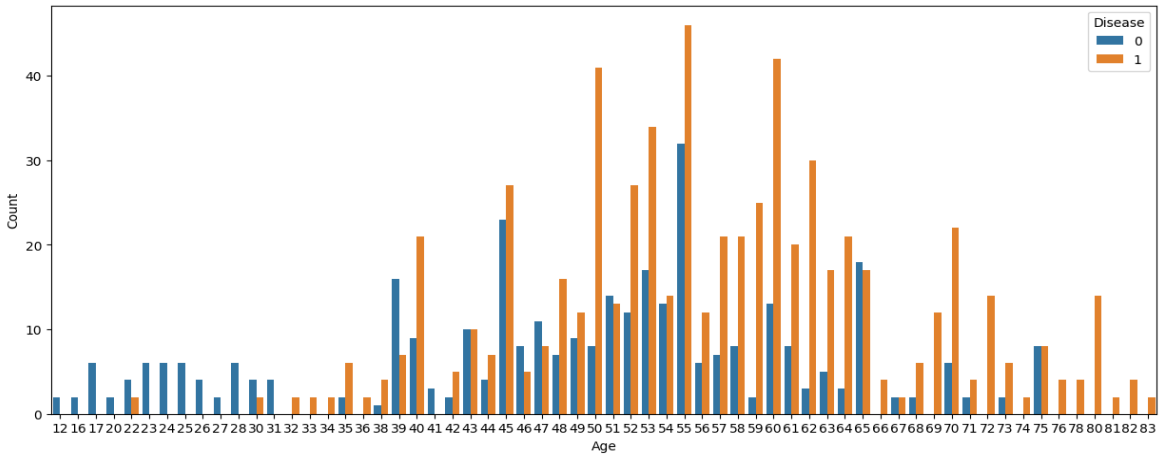


Figure 3.3. 1: *Count plot of Age with respect to Disease*

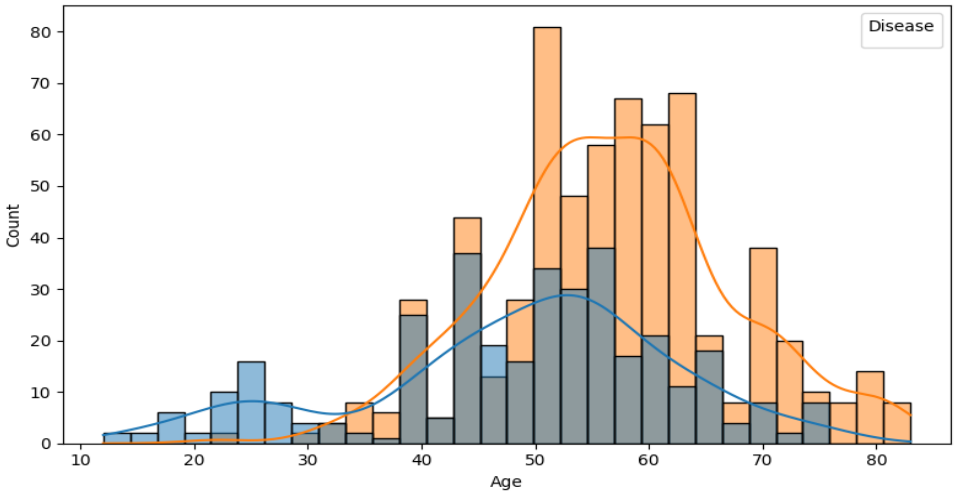


Figure 3.3. 2: *Distribution of Age with respect to Disease*

Gender

The participants' gender is indicated via the Gender attribute. The variable in evaluating is categorical, with 1 being a female and 2 a male.

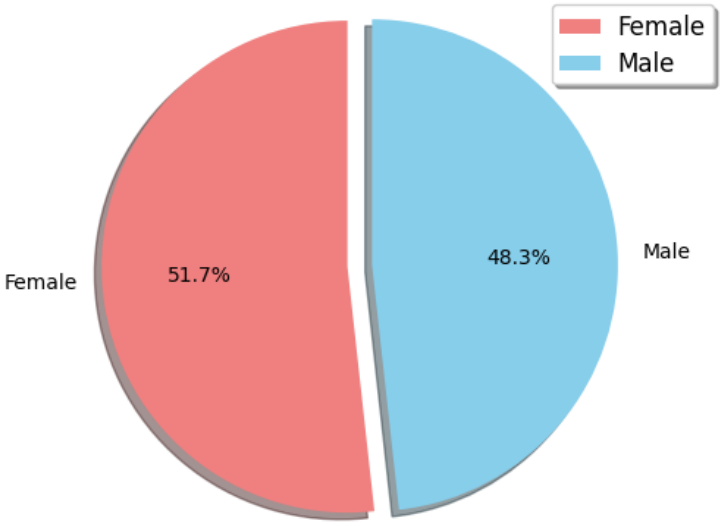


Figure 3.3. 3: *Pie chart of gender distribution*

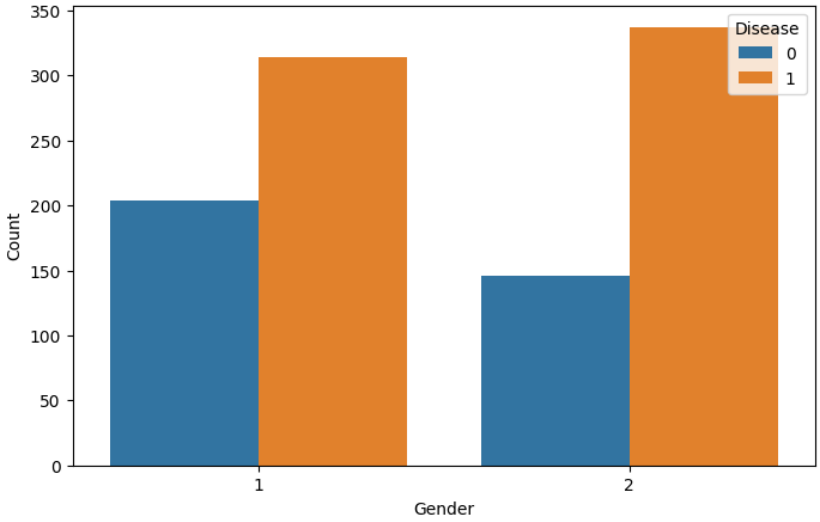


Figure 3.3. 4: *Count plot of Gender with respect to Disease*

We can observe from Figure 3.3.3 that our records for men and women are nearly identical. Therefore, our prediction will be free of prejudice. This distribution of male and female records is an important consideration for our machine learning model development and evaluation. For example, if our model was trained on a dataset that is strongly biased towards one gender, it may not perform as well on more balanced data. Figure 3.3.4 showed that men are often more impacted by CVD than women.

Systolic Blood Pressure

Systolic Blood Pressure is the higher of the two blood pressure values and demonstrates the pressure in the arteries while the heart beats. The unit of measurement for this continuous variable is millimeters of mercury (mmHg). S. A. Mansoor in the financial Express state [22] that a systolic blood pressure reading is considered normal when it is less than 120 mmHg, high when it is 120–129 mmHg, Stage 1 hypertension when it is between 130 and 139 mmHg, and Stage 2 hypertension when it is at least 140 mmHg. Rather than classifying the systolic blood pressure data in our dataset, we employed its continuous value. From Figure 3.3.5 we can see that Systolic Blood Pressure has a significant impact on disease.

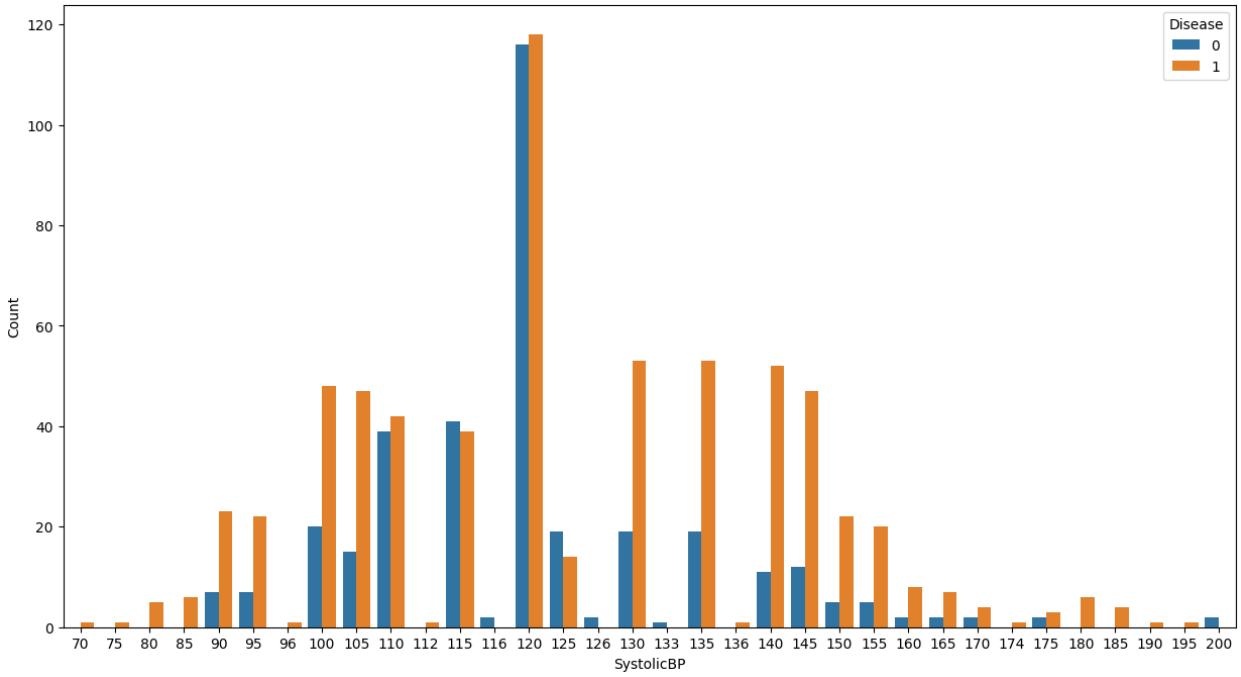


Figure 3.3. 5: *Count plot of Systolic Blood Pressure with respect to Disease*

Diastolic Blood Pressure

Diastolic Blood Pressure is a measurement of the lower blood pressure that shows the pressure within the arteries during a heartbeat's rest. Like systolic blood pressure, its measurement is continuous and expressed in millimeter-Hg. A diastolic blood pressure reading is considered normal when it is less than 120 mmHg, high when it is 120–129 mmHg, Stage 1 hypertension when it is between 130 and 139 mmHg, and Stage 2 hypertension when it is at least 140 mmHg. In our dataset we used continuous value. We didn't categorize the Diastolic Blood Pressure like Systolic Blood Pressure.

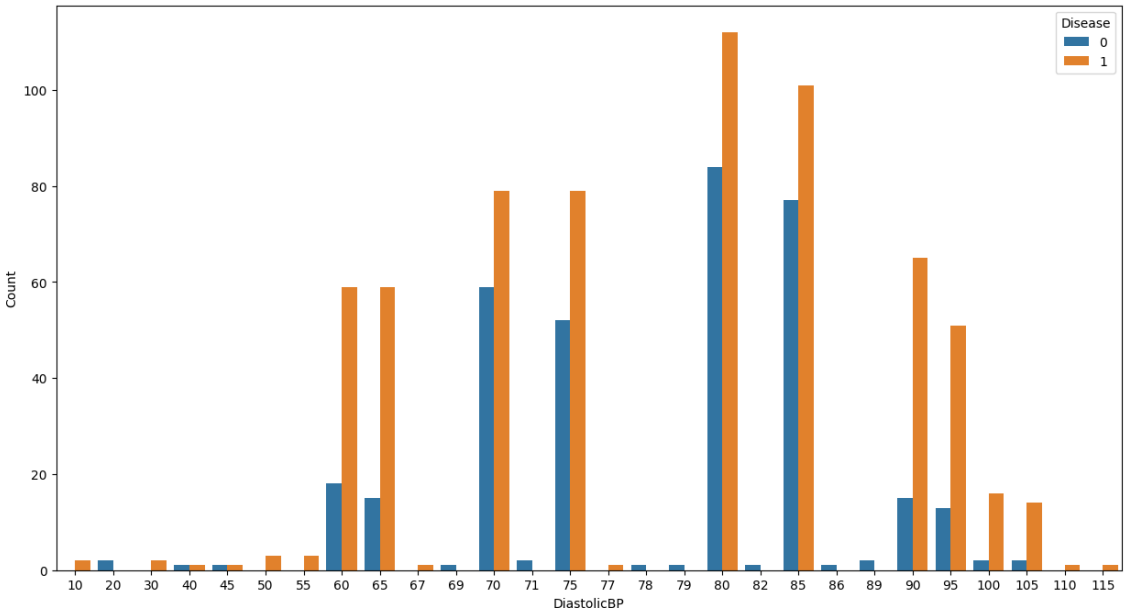


Figure 3.3. 6: *Count plot of Systolic Blood Pressure with respect to Disease*

Cholesterol Level

Another significant risk factor for cardiovascular disease is cholesterol. Rather than use the continuous cholesterol level value in our dataset, we divided the cholesterol level into three categories. There are three ranges: normal starts at 200 mg/dL [milligrams (mg) per deciliter (dL)], 200–239 mg/dL in the moderate range, and 240 mg/dL or more in the high range. Because we have limited data to work with, this classification system simplifies the evaluation of cholesterol levels. Figure 3.3.7 shows the impact of cholesterol level on cardiovascular disease. We can see that with increasing level the risk of CVD is also increasing.

Category of Cholesterol	**Value Range mg/dL**
Normal	<200 mg/dL
Moderate	200-239 mg/dL
High risk	=>240 mg/dL

Table 3.3. 1: *Category of Cholesterol Level*

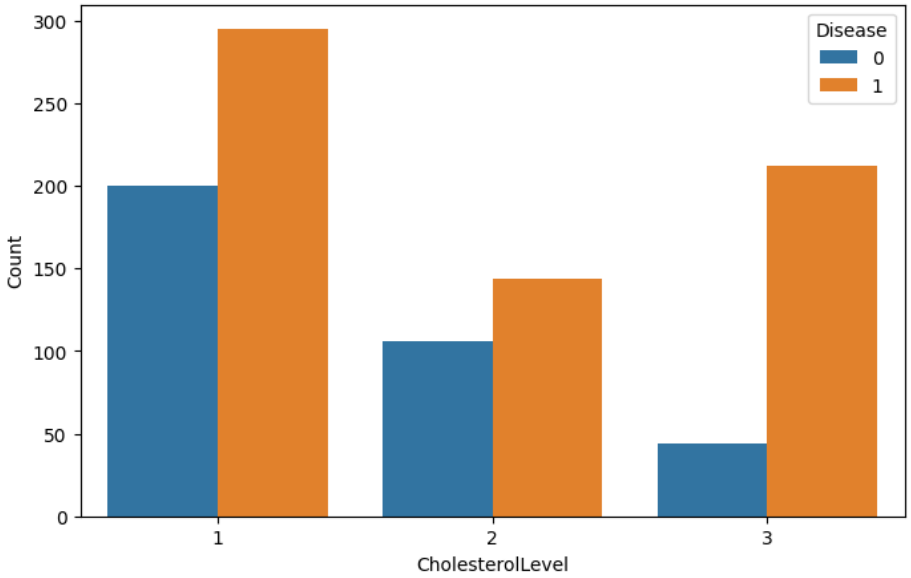


Figure 3.3. 7: *Count plot of Cholesterol Level with respect to Disease*

Glucose Level

The level of glucose (sugar) in the blood is measured by Glucose Level. Similar to the cholesterol level, we classified the glucose level into three groups. Three ranges are identified: the normal range is 3.57-

6.0 mmol/L [millimoles per liter], the prediabetic range is 6.1-6.9 mmol/L, and the diabetic range is 7 mmol/L or higher. From Figure 3.3.8 we can see that glucose level has also a significant impact on CVD.

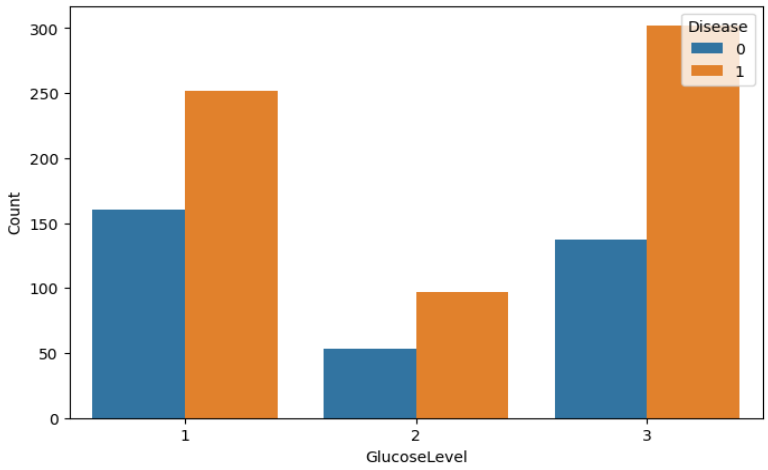


Figure 3.3. 8: *Count plot of Glucose Level with respect to Disease*

Category of Glucose	**Value Range mmol/L**
Normal	3.5-6.0
Prediabetic	6.1-6.9
Diabetic	=>7

Table 3.3. 2: *Category of Glucose Level*

Smoking

The smoking attribute is a behavioral attribute. It indicates the smoking status of a patient where the value is categorical. O signified that the individual does not smoke, whereas 1 denotes that they do. Figure 3.3.9 shows that the majority of our patients with heart disease smoke. Thus, smoking has a significant impact on the detection of cardiovascular disease.

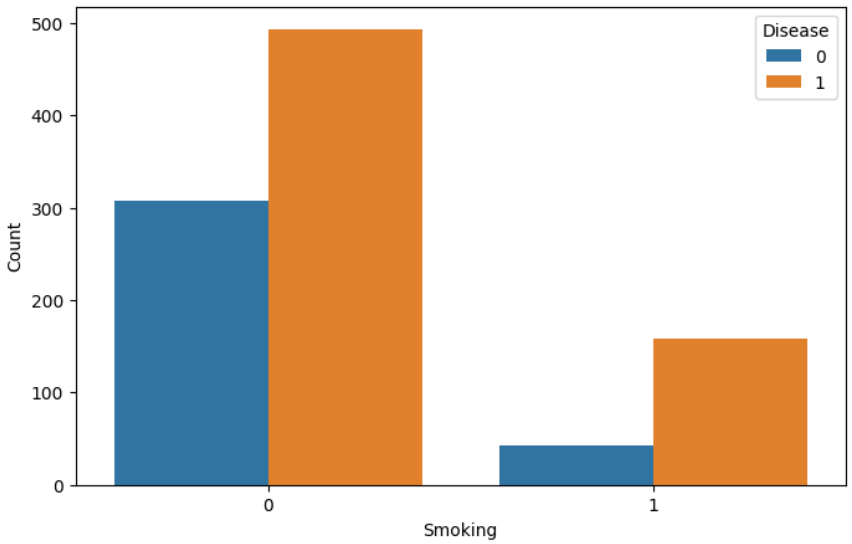


Figure 3.3. 9: *Count plot of Smoking with respect to Disease*

Physical Activity

Like smoking it is also a behavioral attribute which is categorical. A value of 0 indicates that the person is not physically active, whereas a value of 1 indicates that they are. Figure 3.3.10 also shows that physical activity has an impact on CVD.

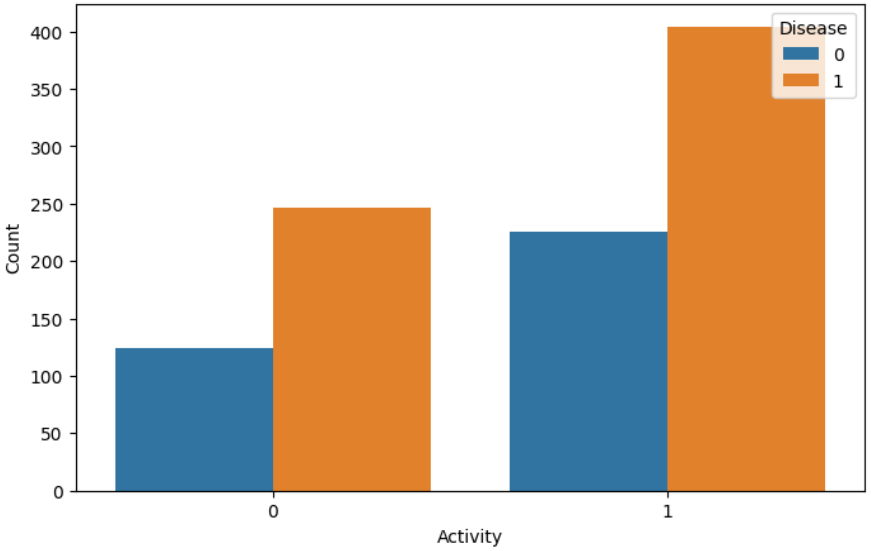


Figure 3.3. 10: *Count plot of Physical Activity with respect to Disease*

Cardiovascular Disease

This column is our goal. We want to determine whether or not a person has cardiovascular disease The target column may have one of two values: 0 or 1. When a person's score is zero, it means they are not cardiac, and when it is one, it means they are. Hence categorization should be our algorithm.

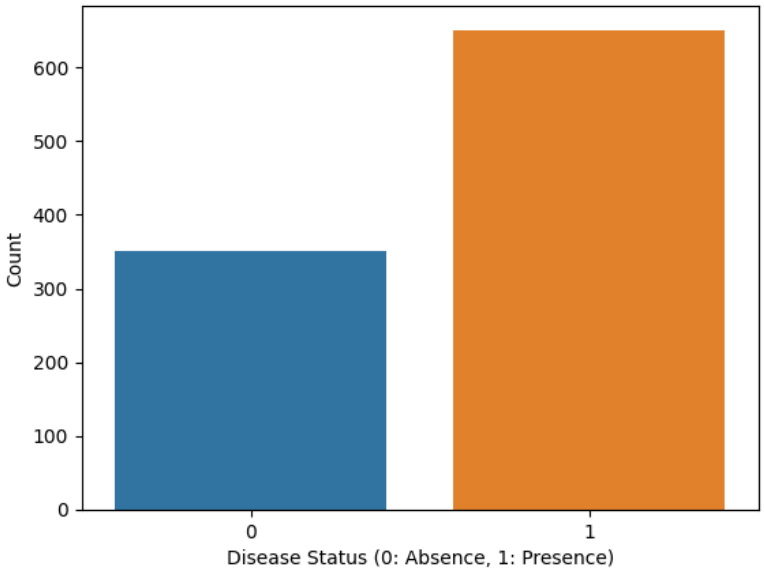


Figure 3.3. 11: *Count plot of Disease*

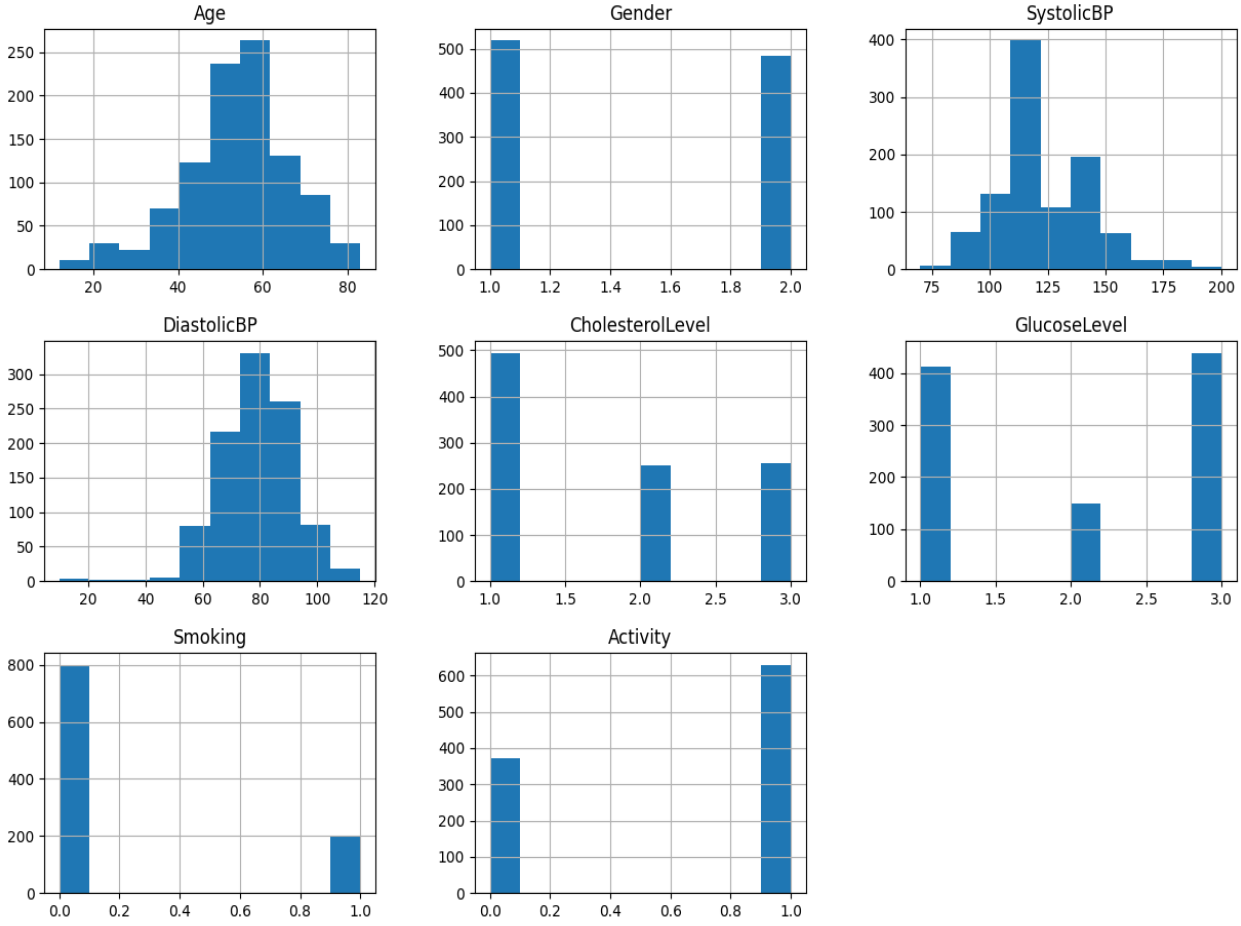


Figure 3.3. 12: *Histogram of the attributes*

3.4 Feature Selection

Feature selection is very important. Selecting relevant features not only improves accuracy but also increases performance. We can determine which features in a dataset are more essential for predicting cardiovascular illness and which features are less relevant by choosing the best feature. So, choosing the best feature is crucial. For feature selection we have applied 2 techniques. ANOVA and correlation coefficient matrix. We can determine which features significantly contribute to the prediction of cardiovascular illness by using ANOVA to evaluate the statistical significance of each feature with respect to the target variable. Furthermore, the correlation coefficient matrix helps identify redundant or strongly linked attributes by focusing on the linear correlations between various parameters. Figure 3.4.1 depicts a boxplot of all attributes. This visualization highlights the distribution and significance of each attribute, emphasizing the most relevant features for predicting cardiovascular disease.

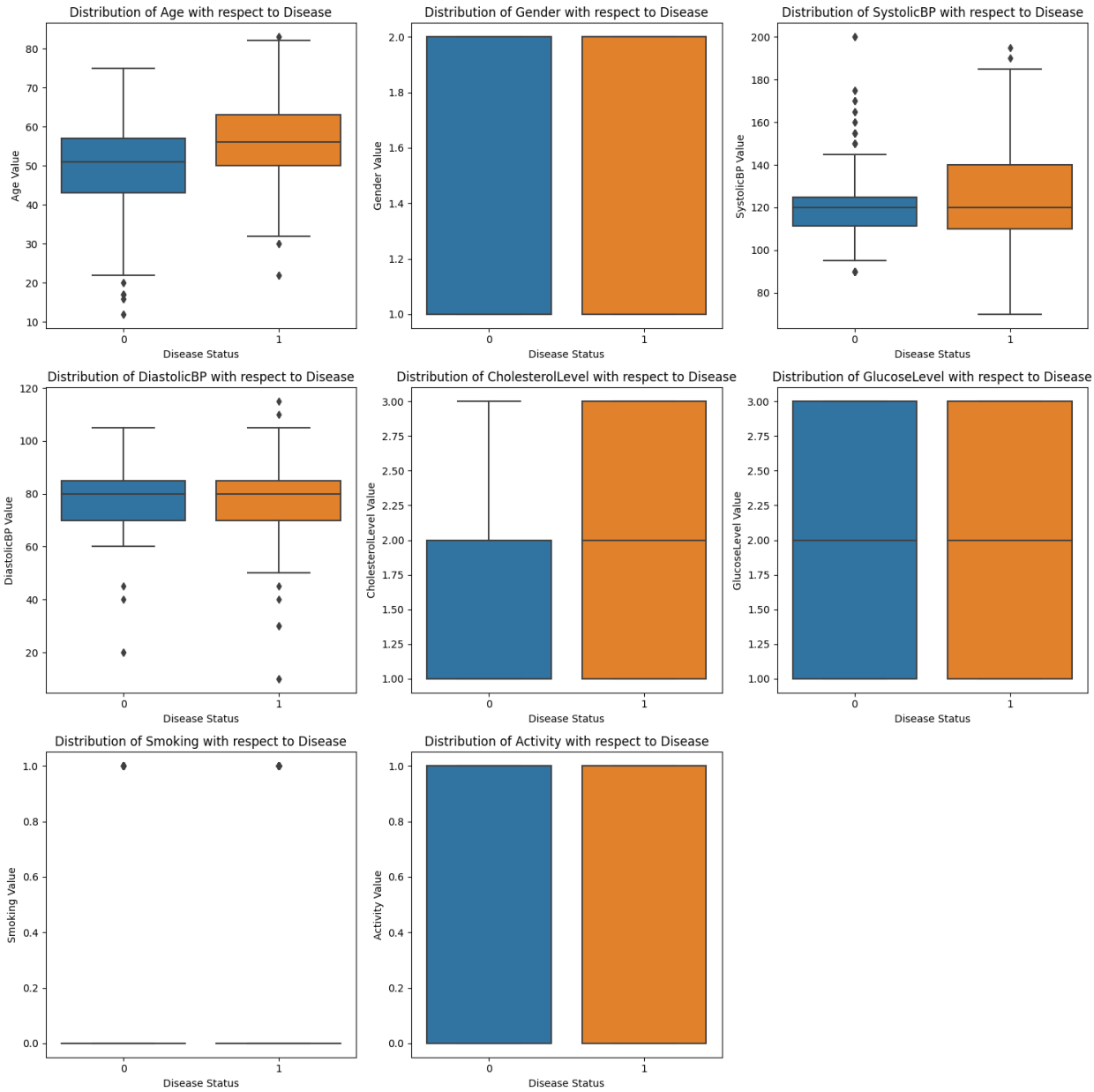


Figure 3.4. 1: *Boxplot showing the distribution of attributes with respect to Disease*

A. ANOVA

In our study, one of the feature selection techniques we employed was analysis of variance (ANOVA). ANOVA is highly useful when working with target variables that are categorical, such as the presence or absence of cardiovascular disease in our dataset, as it assesses the significance of the mean differences between groups. Using the ANOVA technique, each feature is assigned a score that represents its significance for the target variable's prediction. An attribute's perceived importance in group differentiation increases with its ANOVA score. In Table 3.4.1, the score is displayed.

Feature	**Score**
Age	94.15
Cholesterol Level	24.41
Smoking	13.83
Gender	9.92
Glucose Level	5.96
Systolic BP	4.56
Diastolic BP	0.40
Physical Activity	0.36

Table 3.4. 1: *ANOVA Ranking Score*

Figure. 3.4.1 box plot illustrates how these ANOVA scores are distributed visually. Higher scoring features have a greater predictive power for cardiovascular illness, which helps us identify the most important characteristics for our machine learning model.

B. Correlation Coefficient

We have used the Correlation Coefficient Matrix to find the linear relationship between two variables measured by the correlation coefficient, which also indicates its direction and strength. In Fig 3.4.2 we found that the strongest positive connection (0.31) with age, suggesting that the chance of cardiovascular illness tends to rise with age which we also saw in Figure 3.3.1 and Figure 3.3.2. Furthermore, there are moderately positive connections between smoking (0.14) and cholesterol level (0.18). Conversely, Physical Activity (-0.02) shows a modest negative connection, suggesting that a slight reduction in the risk of cardiovascular disease is related with higher levels of physical activity.

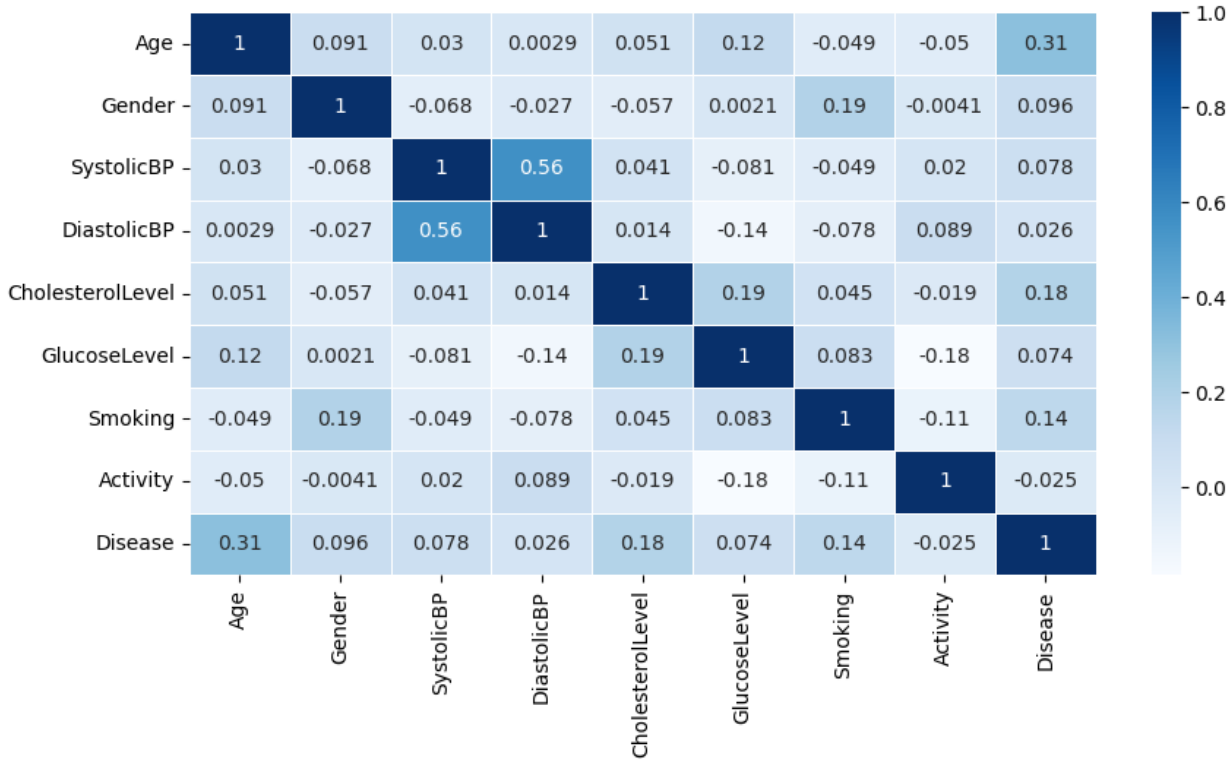


Figure 3.4. 2: *Heatmap of Correlation Matrix*

We decided to keep all attributes in our cardiovascular disease prediction model even after noticing low scores of some attributes. This choice was made since it is recognized that every attribute adds to the total complexity of predicting CVD, and that leaving out any component could result in the loss small patterns or possible interdependencies. Even though some attributes individually have lesser predictive value, our approach promotes a thorough analysis, taking into account all available information to ensure an accurate evaluation of potential factors for cardiovascular disease.

3.5 Data Splitting

Data splitting, which divides a dataset into a training set and a testing set, is an essential step in machine learning. After being trained on the training set, the machine learning model is able to recognize relationships and patterns in the data. Testing set is used to evaluate the model's performance on unknown data, which it hasn't seen in training. The cardiovascular disease dataset was divided into two sections: a training set comprising 70% and a testing set comprising 30%. The testing set is used to evaluate the models' performance after they have been trained on the training set.

3.6 Training Model

Since we know that classification algorithms are used when the output variable is categorical—that is, when it has two classes—such as Yes-No, Male-Female, True-False, and so forth—we will use supervised classification machine learning techniques for our research. We must forecast whether a patient has cardiovascular disease or not for our model to work. Thus, we have a categorical output variable. We will thus employ six categorization algorithms in order to forecast the onset of cardiovascular illness. The scikit-learn Python library was used to conduct the experiments in this work. On our 1001 samples, we applied the following algorithms: Random Forest, XGB (Extreme Gradient Boosting), Decision Tree (DT), SVM (support vector machine), KNN (K-Nearest Neighbor), and Logistic Regression.

1) Logistic Regression

Cardiovascular disease likelihood can be predicted using logistic regression. Utilizing a logistic model, one can determine a person's likelihood of having CVD by summing the weighted input variables. This total is converted by the sigmoid function into a probability score, ranging from 0 to 1, that represents the possibility of developing CVD. Sigmoid/ Logistic function is:

$$g(z) = \frac{1}{1 + e^{-z}} \quad (1)$$

The output of the sigmoid function must be between 0 and 1. If the probability is above a threshold (usually we consider 0.5), then it is classified as cardiac (1). Otherwise, it is classified as non-cardiac (0).

2) Support Vector Machine

Support Vector Machine uses hyperplane (decision boundary) to separate different classes. For our CVD prediction SVM generates a decision border that optimally separates instances linked with cardiovascular disease. By carefully placing the hyperplane produced by support vector machines (SVM) to traverse the feature space, it will be possible to classify instances based on relevant factors like age, gender, blood pressure, cholesterol, physical activity, and others. In order to define the decision boundary, SVM makes sure that crucial cases are prioritized by choosing extreme points or support vectors. The distance between the hyperplane and support vectors influences prediction. From Table 4.4.1 and Table 4.4.2 unlike other model SVM performed better in training data reducing overfitting shows that SVM is effective in capturing complex relationships in patient data.

3) Decision Tree

Decision Tree predicts cardiovascular disease (CVD) by following a hierarchical structure. The dataset is divided using important attributes such as age, systolic and diastolic blood pressure, cholesterol, diabetes, and smoking, starting at the root node. Internal nodes evaluate particular circumstances pertaining to these attributes in order to serve as decision points. The final prediction for CVD are stored in leaf nodes and might be either positive or negative. Prediction methods are defined by explicit decision rules formed by the routes from root to leaf nodes. Binary conditions are present in category features, while thresholds exist in numerical features. Until certain requirements are satisfied, this recurrent process keeps going. The model's explicit decision structure makes it easy to interpret and provides insights into the factors impacting cardiovascular disease predictions. For instance, if a person's age is below a certain threshold and their systolic blood pressure falls within a specific range, the prediction may be Negative for CVD; otherwise, it could be Positive for CVD.

4) Random Forest

The Random Forest classification algorithm, which is based on Decision Trees, works by grouping multiple trees together to produce what is figuratively called a "forest" of Decision Trees. Regarding the prediction of cardiovascular disease (CVD), every tree in the Random Forest evaluates the input data on its own and generates a prediction. A majority voting technique is then used to generate the final predictions for a specific data point; the class label with the most votes is regarded as the overall prediction. Every single piece of data is assessed by every tree in the Random Forest in order to begin the prediction process. This thorough analysis enables the model to examine several features and take into account different input data components. Random Forest's power is found in its capacity to identify complex relationships and patterns in the dataset.

5) XGB (Extreme Gradient Boosting)

"Extreme Gradient Boosting," or "XGBoost," uses an ensemble learning approach to function as a cardiovascular disease (CVD) prediction model. In order to get an overall prediction that is more reliable and accurate, predictions from several weak models are combined iteratively. The method starts

with a first model, and the next steps concentrate on improving predictions by dealing with residuals, or the differences between expected and actual results. A new weak model, frequently a decision tree, is introduced with each iteration and trained to forecast the residual errors from the prior models. The ensemble's overall forecasts are updated by combining and weighting the predictions from various models. This iterative process goes on, with an increasing focus on fixing previous mistakes. The final CVD predict is generated by combining the weak model predictions. Because of its excellent performance in predicting CVD and its ability to handle imbalanced data and incorporate regularization approaches like L1 and L2 regularization, XGBoost is a recommended option for effective and scalable machine learning model training.

6) KNN (K-Nearest Neighbor)

Another useful technique in our cardiovascular disease prediction models is K-Nearest Neighbors (KNN). It does not instantly learn from the training set; instead, it saves the dataset and processes it when it comes time to classify. In CVD prediction, this algorithm efficiently manages data. As new data arrives, KNN categorizes it into a subset similar to the already stored data. The classification of a data point in CVD prediction is determined by a majority vote from its closest neighbors. The selection of the distance measure and the number of neighbors (k) significantly impacts KNN's performance. In our approach, we employed the elbow method to determine the optimal number of neighbors, setting it to k=18 for effective CVD prediction.

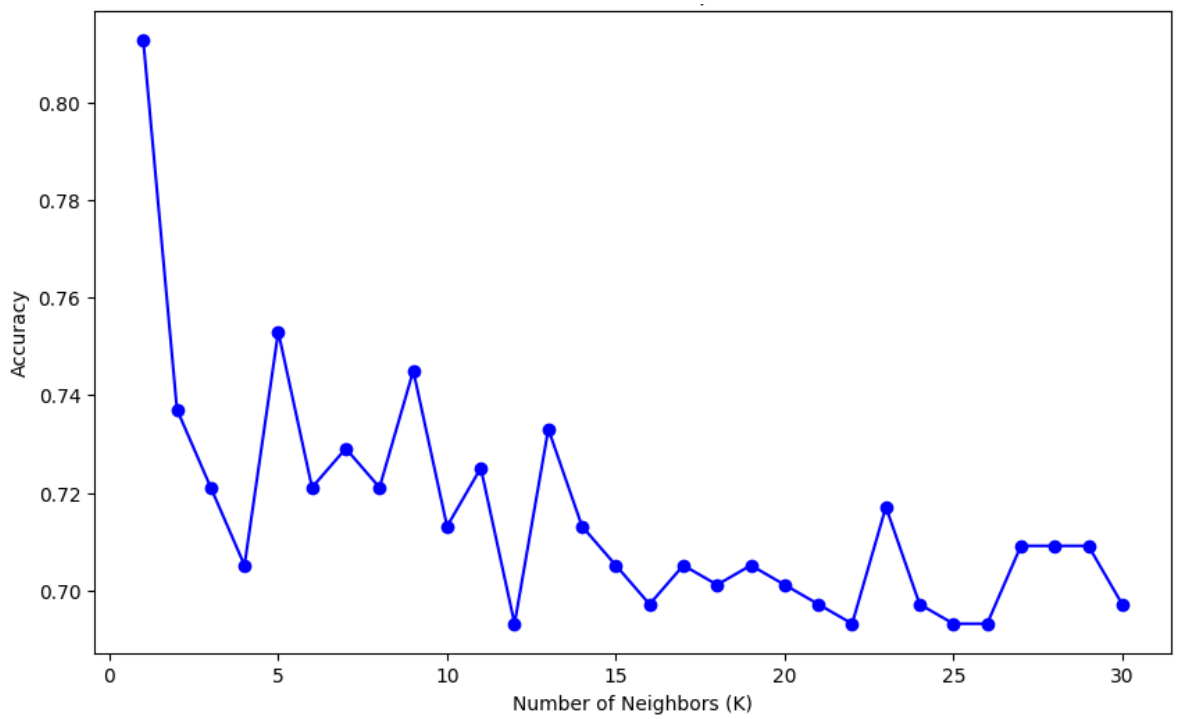


Figure 3.6. 1: *Elbow method in KNN*

Chapter 4

Result & Discussions

4.1 Introduction

This chapter presents a thorough examination of the machine learning models used to predict cardiovascular disease. I have explored the complexities of my evaluation process in this chapter, providing insight to evaluate the effectiveness of our predictive models. I have showed the numerical and visual representation of important metrics, including accuracy, precision, recall, F1 score and ROC curve, through a thorough examination of performance evaluation, providing an open perspective of my models' effectiveness in handling the prediction of cardiovascular disease. Furthermore, this chapter makes it easier to do a comparison with other finished studies in the field of cardiovascular disease prediction. I have completed all of my work using Google Colab.

4.2 Evaluation Metrics

I have already discussed that I have used different evaluation metric to assess the performance of my model. When evaluating the effectiveness of machine learning models, evaluation metrics are crucial techniques. My model's performance has been assessed using f1, accuracy, precision, and recall. They are essential for determining how well our model is working and whether it satisfies the objectives and specifications of our goal. These matrices are described below-

- **Precision**

The ratio of accurately anticipated positive cases to all expected positive cases is known as precision. High precision in the context of CVD denotes a low risk of false positives and trustworthy identification of true positive cases.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

- **Recall**

Sensitivity, or recall, measures how well the model captures all of the real positive cases of cardiovascular disease (CVD). It is the proportion of all actual positive CVD cases to all accurately anticipated positive CVD cases.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

- **F1 Score**

The harmonic mean of recall and precision is the F1 score. It offers a balance between recall and precision, particularly when there are unequal class sizes in cardiovascular disease (CVD) prediction.

$$\text{F1 Score} = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (4)$$

- **Accuracy**

The overall accuracy of the model's predictions is measured by accuracy. It is the proportion of accurately predicted observations to all observations, including true positives and true negatives.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (3)$$

Where,

$TP = True\ Positive$ (Total number of patients correctly identified that they have CVD).
$FP = False\ Negative$ (Total number of patients incorrectly identified that they have CVD).
$TN = True\ Negative$ (Total number of patients who correctly identified that they have no CVD).
$FN = False\ Negative$ (Total number of patients incorrectly identified that they have no CVD).

- **AUC-ROC Curve**

One of the statistical techniques for data classification that is most frequently employed is the receiver operating characteristic curve (ROC). It makes it easier for us to comprehend how the model makes decisions at different degrees of certainty. The curve is composed of two lines: one shows how often the model properly identifies positive instances (true positives), while the other shows how frequently it wrongly identifies negative instances as positive (false positives). This graph aids in our comprehension of the model's functionality. It is crucial for accurately predicting cardiovascular disease. The ROC curve's area under the curve (AUC) shows how well a classifier can differentiate between several classes. The AUC indicates how well the model can distinguish between positive and negative categories.

4.3 Performance Evaluation

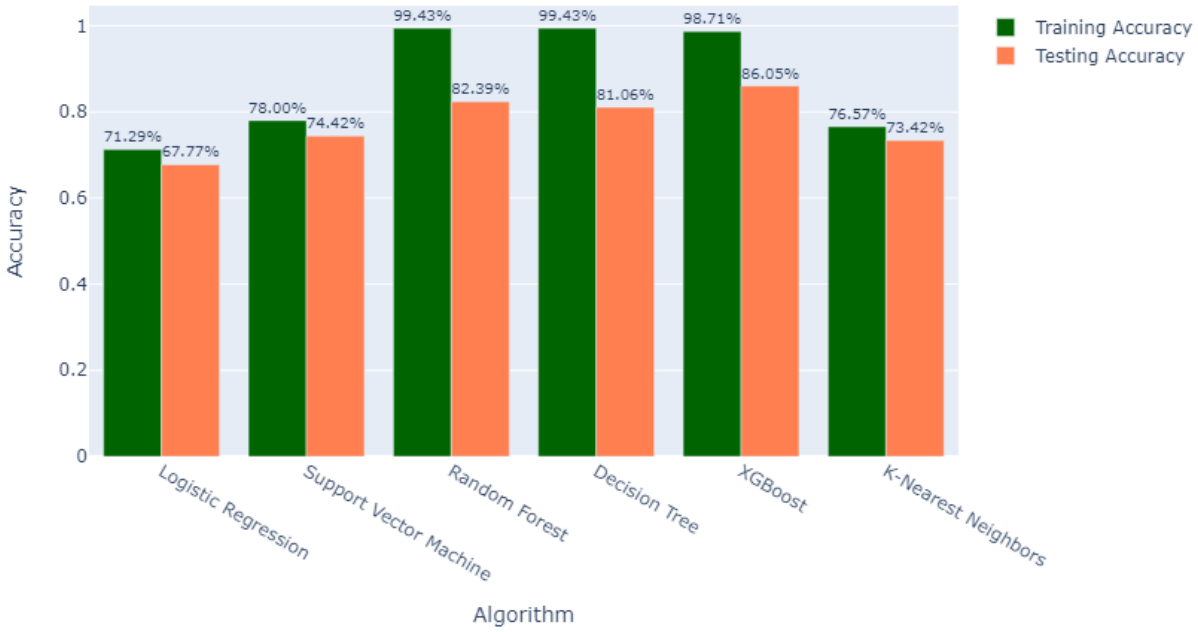


Figure 4.3. 1: *Training and Testing Accuracy plot without 5-fold Cross Validation*

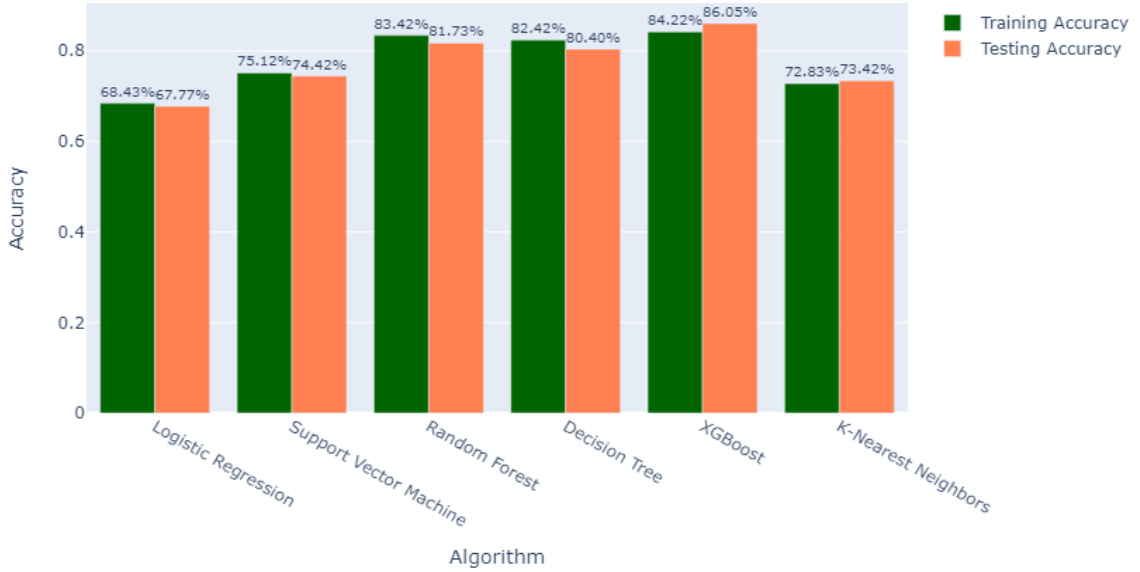


Figure 4.3. 2: *Training and Testing Accuracy plot with 5-fold Cross Validation*

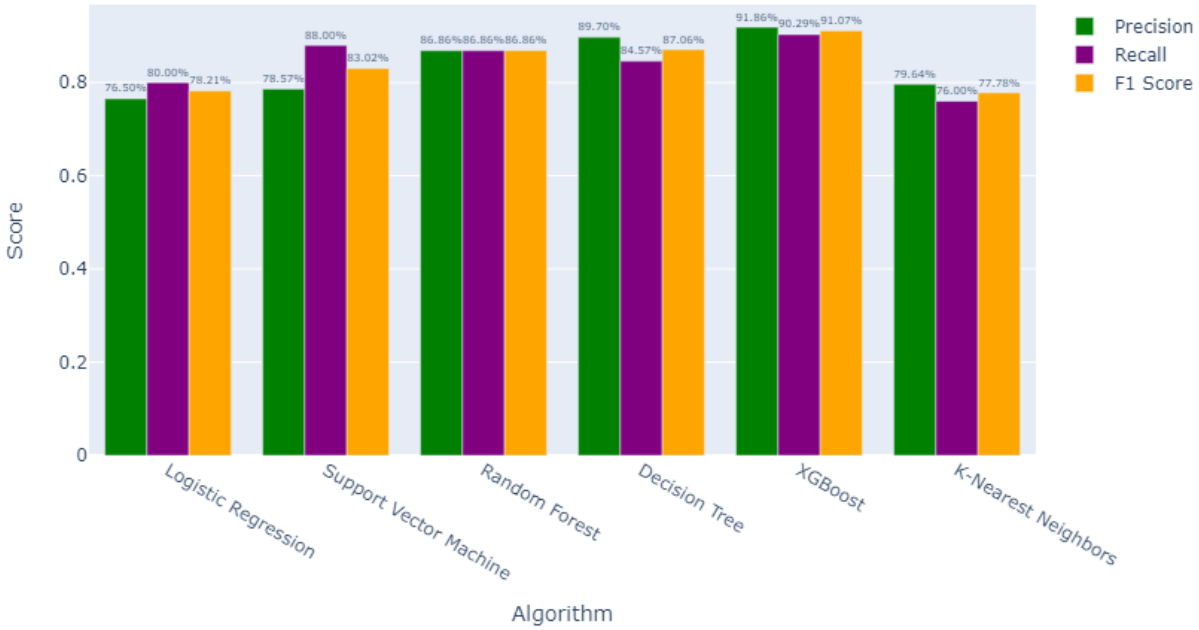


Figure 4.3. 3: *Precision, Recall and F1 Score plot without 5-fold Cross Validation*

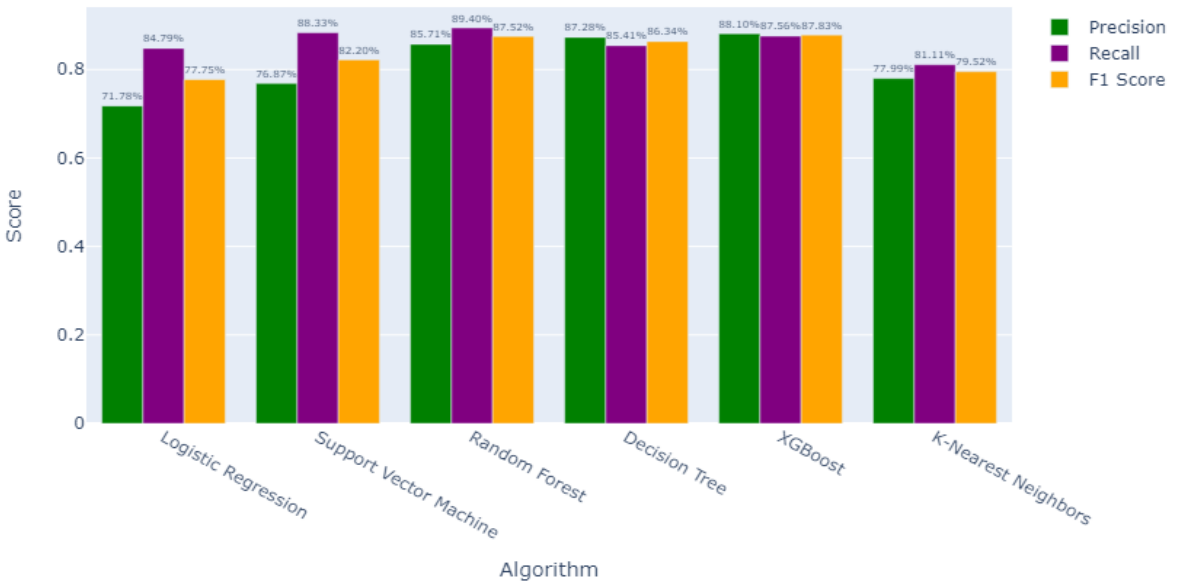


Figure 4.3. 4: *Precision, Recall and F1 Score plot with 5-fold Cross Validation*

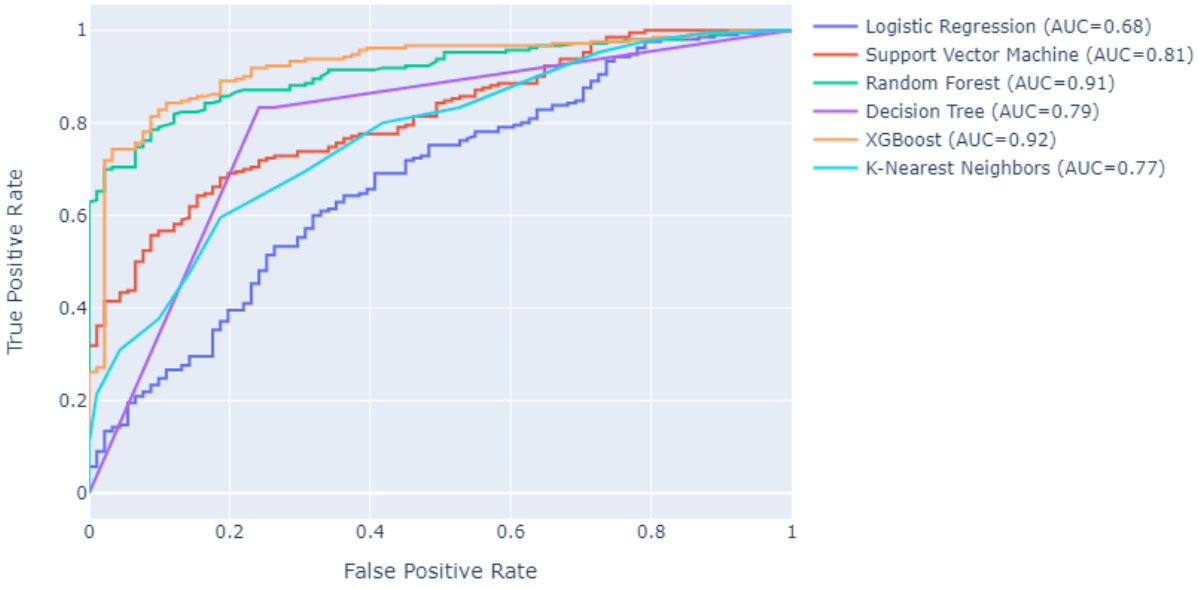


Figure 4.3. 5: *Comparison of ROC curve without 5-fold Cross Validation*

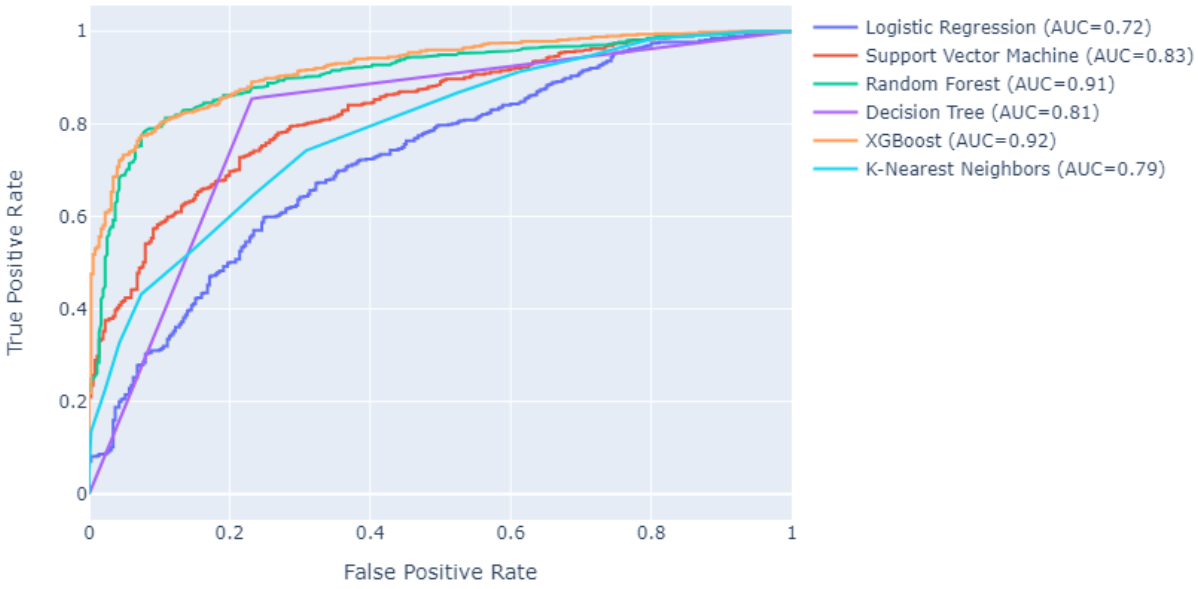


Figure 4.3. 6: *Comparison of ROC curve with 5-fold Cross Validation*

30% of the data in my dataset were test sets, and the remaining 70% were training sets. I have compared the accuracy, precision, recall, f1 score and ROC with and without stratified 5-fold cross validation. My model has only been coded and trained on a Google Colab notebook. I have utilized the Python Seaborn package and Matplotlib for the metrics visualization. My program's objective is to increase the testing accuracy as much as possible and training accuracy but carefully maintain the difference of training and testing accuracy for avoiding overfitting problem.

4.4 Result discussion of the model

I used an 11th generation Intel(R) Core (TM) i7-1165G7 @ 2.80GHz computer with 16 GB of RAM for my model prediction using Google Colab. The dataset was cleaned and preprocessed to decrease its 1019 rows and 9 attributes to 1001 rows and 9 attributes. This research employed the following algorithms: ANN, XGBoost, KNN, decision tree, random forest, logistic regression, support vector machine, and random forest. As I previously discussed, this study included a number of performance metrics, including area under the ROC curve, F1 score, accuracy, precision, and recall. 30% of the dataset was used to test the model and 70% was used to train the model. This experiment was performed based on 5-fold cross-validation. Each classification algorithm was implemented into two cases. First, they used all of the attributes without 5-fold cross validation and then with 5-fold cross validation. The number of decision trees formed in the random forest ensemble will be 300 for K-Nearest Neighbors K=18. I have also compared ROC curve and AUC for both cases. If a classifier's AUC approaches 1, we can be certain that it can correctly distinguish between all Positive and Negative class points. Conversely, if the AUC had been near zero, the classifier would have predicted all Positives as Negatives and all Negatives as Positives. The model's ability to differentiate between positive and negative classifications improves with increasing AUC. So, when evaluating different classifiers we would prefer the model with a higher AUC score, as it indicates better performance in terms of discrimination. Below Table 4.4.1 and Table 4.4.2 Shows the comparison of some evaluation metrices before and after 5-fold cross validation. Graphical representation of the metrices for each algorithm in both cases is shown on Figure 4.3.1, Figure 4.3.2, Figure 4.3.3 and Figure 4.3.4.

The several machine learning classifiers that were employed to determine whether cardiovascular disease was present in the dataset are shown in Table 4.4.1. In this case, I did not use 5-fold cross

validation. XGBoost, logistic regression, decision trees, random forests, support vector machines, and k-nearest neighbors were some of these classifiers. The findings showed that the maximum training and testing accuracy of 98.93% and 87.65% were attained by the XGB (Extreme gradient boosting) method, which also had the highest precision, recall, and F1 score of 0.92, 0.90, and 0.91.

No.	Model	Training Accuracy (%)	Testing Accuracy (%)	Precision	Recall	F1 Score
1.	XGB	98.71%	87.05%	0.90	0.89	0.89
2.	Decision Tree	99.43%	81.06%	0.88	0.83	0.85
3.	Random Forest	99.43%	82.39%	0.86	0.88	0.87
4.	Support Vector Machine	78%	74.42%	0.78	0.8	0.82
5.	K-Nearest Neighbors	76.57%	73.42%	0.81	0.8	0.80
6.	Logistic Regression	71.29%	67.77%	0.76	0.71	0.77

***Table 4.4. 1**: Accuracy of model without considering 5-fold Cross validation*

I have noticed that in almost every classifier there is a noticeable difference between training and testing accuracy. Almost all the classifier's training accuracy is higher than the testing accuracy. We know that if training accuracy gets very high than testing accuracy it shows overfitting issues. When a model is unable to generalize and fits too closely to the training dataset, it is said to be overfitting. So, it needs to be removed. To address this concern, I have applied 5-fold cross-validation to my models, resulting in a notable reduction in overfitting issues.

No.	Model	Training Accuracy (%)	Testing Accuracy (%)	Precision	Recall	F1 Score
1.	XGB	84.22%	86.05%	0.88	0.87	0.87
2.	Decision Tree	82.42%	80.40%	0.87	0.85	0.86
3.	Random Forest	83.42%	81.73%	0.85	0.89	0.87
4.	Support Vector Machine	75.12%	74.42%	0.76	0.88	0.82
5.	K-Nearest Neighbors	72.83%	73.42%	0.77	0.81	0.79
6.	Logistic Regression	68.43%	67.77%	0.71	0.84	0.77

***Table 4.4. 2**: Accuracy of model with considering 5-fold Cross validation*

In Table 4.4.2, the impact of Stratified 5-fold cross-validation is evident as XGBoost demonstrates enhanced accuracy in both training (84.22%) and testing (86.05%). Furthermore, precision, recall, and F1 score metrics are notably higher compared to other models, registering at 0.88, 0.87, and 0.87, respectively. The application of 5-fold cross-validation has effectively mitigated overfitting, as

evidenced by the reduced disparity between training and testing accuracy. Cross-validation ability to effectively address overfitting highlights my model's increased reliability and generalizability.

Based on the performance evaluation metrics presented in Table 4.4.1 and 4.4.2, it can be said that using 5-fold cross validation improved the performance of each classification algorithm. According to the aforementioned experimental results, utilizing cross validity enhances performance and in terms of cardiovascular disease prediction, XGBoost classifier performed better.

As we already know, a model with an AUC near 1 is quite effective and has good separation powers, meaning it can distinguish between instances that are positive and negative. Conversely, an AUC nearing 0 suggests a less effective model, indicating challenges in accurately separating classes. This concise assessment of AUC provides a quick and insightful summary of the model's discriminatory power in my study. In my analysis of various classifiers, XGBoost (Extreme Gradient Boosting) stands out with the highest AUC of 0.92 which we can see in Figure 4.3.5 and Figure 4.3.6. This AUC value reflects its superior ability to distinguish between positive and negative instances, making it the most effective classifier among those evaluated in my study.

4.5 Comparison to some previously completed works

Authors	**Approach**	**Best Model**	**Accuracy**	**Dataset**
Weicheng et al., (2021) [10]	5-fold cross-validation, Feature integration techniques, Correlation coefficient for feature extraction	Support Vector Machine (SVM)	71.48%	Svetlana Ulianova research group (70,000 data points, 12 attributes)
Syed et al., (2020) [11]	Deep learning techniques, Machine learning algorithms (KNN, DT, SVM), Tensor Flow Keras	Artificial Neural Network (ANN)	85.24%	Kaggle
Atharv et al., (2020) [12]	Feature engineering (BMI), Outlier and missing value handling	Decision Tree	73.13%	Kaggle
Ibashisha et al.,(2020) [13]	Data Pre-processing, Feature extraction, K-fold cross-validation for finding the optimal result	K-Nearest Neighbor (KNN)	72.91%	Kaggle
Vansh et al.,(2021) [14]	Machine learning model integrated into iOS application	XGBoost	72.70%	Kaggle
Abdul et al.,(2022) [15]	Prediction with Various feature integration techniques, also incorporates correlation	Support Vector Machine (SVM)	96.72%	Z-Alizadeh Sani dataset,cleveland,statlogHungary, Switzerland,

	coefficient-based feature extraction (Hyperparameter tuning, Classifier sensitivity analysis)			
Arsalan et al.,(2023) [16]	Simple random sampling, Exploratory analysis, Experimental output analysis, Confusion matrix, Recursive operating characteristic curve, Algorithm optimization and comparison	Random Forest (RF)	85%	Lady Reading Hospital and Khyber Teaching Hospital, Pakistan
Khairul et al., (2022) [18]	Feature selection improved performance. 5 fold validation applied on training data.	XGBoost	73.72% & 81.14%	Kaggle
Mohammad et al. 2021 [19]	Data collection through interviews, Cross-validation to remove noisy input, Model testing with eight qualities	SVM	91%	Sylhet region of Bangladesh
Md et al. 2022 [20]	Feature selection using info-gain; Null value handling with mean imputation; Comparison of 13-feature and 10-feature datasets; Results reported in confusion matrices and accuracy.	RF	95.63%	Kaggle
M.Shahiduzzaman et al. 2022 [21]	Voting ensemble classifiers	KNN	75%	Kaggle

Table 4.5. 1: *Comparison with other related works*

Chapter 5

Conclusion

Cardiovascular disease is a serious health issue that is becoming more deadly these days. The early identification of CVD can save a great deal of lives. In this situation, a machine learning model may be useful. A machine learning model may identify a disease rapidly, contributes in early detection and appropriate treatment and lowers the mortality rate. I have attempted to forecast CVD using several ML models. I have used real world patient information from two hospitals in Dhaka, Bangladesh, for better performance. The major goal of this work was to properly diagnose cardiovascular illness by utilizing many machine learning models. I have preprocessed my collected dataset by removing duplicate data. After using Stratified K-Fold cross-validation, where the number of folds for cross-validation is five, I saw that my machine learning model performed better. Especially helpful for unbalanced datasets, stratified K-fold guarantees that every fold retains the same distribution of target classes as the original dataset. In addition, classifiers are shown using the Receiver Operating Characteristic (ROC) curve. Classifier performance is evaluated using the area under the ROC curve (AUC). All things considered, this work offers a thorough analysis of several classifiers using Stratified K-Fold cross-validation and presents the results using ROC curve metrics, accuracy, precision, recall, and F1 score. XGBoost (XGB) particularly stands out, demonstrating superior performance after Stratified K-Fold cross-validation, with training accuracy at 84.22% and testing accuracy at 86.05%. Moreover, precision, recall, and F1 score metrics for XGB are notably higher compared to other models. These findings emphasize the effectiveness of XGBoost in this classification task, showcasing its robustness and potential for practical deployment.

Chapter 6

Future Work & References

6.1 Future Work

I intend to use cutting-edge deep learning algorithms—such as convolutional neural networks (CNNs) and artificial neural networks (ANNs)—in my future work in order to take advantage of their capacity for intricate pattern recognition. In comparison to a normal BMI, obesity was reported by Sadia et al. to be related with a significantly higher risk of cardiovascular morbidity and death as well as a shorter lifespan [23]. So, BMI is a risk factor for CVD. But in my dataset, it is missing. Given the current limitation of missing BMI data, a significant focus in my future work plan is to collect and incorporate BMI information into the dataset. This effort aims to enrich the dataset, providing a more comprehensive health profile for improved model performance. In order to further increase the dataset, I also plan to collect data from more than two hospitals in Bangladesh. Increasing both the dataset size and including BMI data will enhance the model's ability to learn diverse patterns, improve generalization, and contribute to a more robust healthcare predictive model.

6.2 References

[1] Cardiovascular diseases. (2019, June 11). https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1

[2] Molla, W. B. H. A. M. A. M. (2022, September 29). Cardiovascular diseases: A rising concern for young people. The Daily Star. https://www.thedailystar.net/supplements/world-heart-day-2022/news/cardiovascular-diseases-rising-concern-young-people-3130526

[3] Islam, A. M., Mohibullah, A. K. M., & Paul, T. (2016). Cardiovascular disease in Bangladesh: a review. Bangladesh Heart Journal, 31(2), 80-99.

[4] Hanif, A. A. M., Hasan, M., Khan, M. S. A., Hossain, M. M., Shamim, A. A., Hossaine, M., ... & Mridha, M. K. (2021). Ten-years cardiovascular risk among Bangladeshi population using non-laboratory-based risk chart of the World Health Organization: Findings from a nationally representative survey. Plos one, 16(5), e0251967.

[5] World Bank Open Data. (n.d.). World Bank Open Data. https://data.worldbank.org/indicator/SP.POP.TOTL?locations=BD

[6] Chowdhury, M. Z. I., Haque, M. A., Farhana, Z., Anik, A. M., Chowdhury, A. H., Haque, S. M., ... & Turin, T. C. (2018). Prevalence of cardiovascular disease among Bangladeshi adult population: a systematic review and meta-analysis of the studies. Vascular health and risk management, 165-181.

[7] Khanam, F., Hossain, M. B., Mistry, S. K., Afsana, K., & Rahman, M. (2019). Prevalence and risk factors of cardiovascular diseases among Bangladeshi adults: findings from a cross-sectional study. Journal of epidemiology and global health, 9(3), 176.

[8] Mondal, R., Ritu, R. B., & Banik, P. C. (2023). Prevalence of cardiovascular disease and its associated factors among middle-aged type-2 diabetic subjects: a cross-sectional study in selected hospitals in Bangladesh. Journal of Xiangya Medicine, 8.

[9] Ramalingam, V. V., Dandapath, A., & Raja, M. K. (2018). Heart disease prediction using machine learning techniques: a survey. International Journal of Engineering & Technology, 7(2.8), 684-687.

[10] Sun, W., Zhang, P., Wang, Z., & Li, D. (2021). Prediction of cardiovascular diseases based on machine learning. ASP Transactions on Internet of Things, 1(1), 30-35.

[11] Pasha, S. N., Ramesh, D., Mohmmad, S., & Harshavardhan, A. (2020, December). Cardiovascular disease prediction using deep learning techniques. In IOP conference series: materials science and engineering (Vol. 981, No. 2, p. 022006). IOP Publishing.

[12] Nikam, A., Bhandari, S., Mhaske, A., & Mantri, S. (2020, December). Cardiovascular disease prediction using machine learning models. In 2020 IEEE Pune Section International Conference (PuneCon) (pp. 22-27). IEEE.

[13] Marbaniang, I. A., Choudhury, N. A., & Moulik, S. (2020, December). Cardiovascular disease (CVD) prediction using machine learning algorithms. In 2020 IEEE 17th India Council International Conference (INDICON) (pp. 1-6). IEEE.

[14] Kedia, V., Regmi, S. R., Jha, K., Bhatia, A., Dugar, S., & Shah, B. K. (2021, April). Time Efficient IOS Application For CardioVascular Disease Prediction Using Machine Learning. In

2021 5th International Conference on Computing Methodologies and Communication (ICCMC) (pp. 869-874). IEEE.

[15] Saboor, A., Usman, M., Ali, S., Samad, A., Abrar, M. F., & Ullah, N. (2022). A method for improving prediction of human heart disease using machine learning algorithms. Mobile Information Systems, 2022.

[16] Khan, A., Qureshi, M., Daniyal, M., & Tawiah, K. (2023). A Novel Study on Machine Learning Algorithm-Based Cardiovascular Disease Prediction. Health & Social Care in the Community, 2023.

[17] Rashme, T. Y., Islam, L., Jahan, S., & Prova, A. A. (2021, July). Early prediction of cardiovascular diseases using feature selection and machine learning techniques. In 2021 6th International Conference on Communication and Electronics Systems (ICCES) (pp. 1554-1559). IEEE.

[18] Fahim, K. E., Yassin, H., Amin, M. H., Dewan, P. D., & Islam, A. (2022, September). Detection of Cardiovascular Disease of Patients at an Early Stage Using Machine Learning Algorithms. In 2022 International Conference on Healthcare Engineering (ICHE) (pp. 1-6). IEEE.

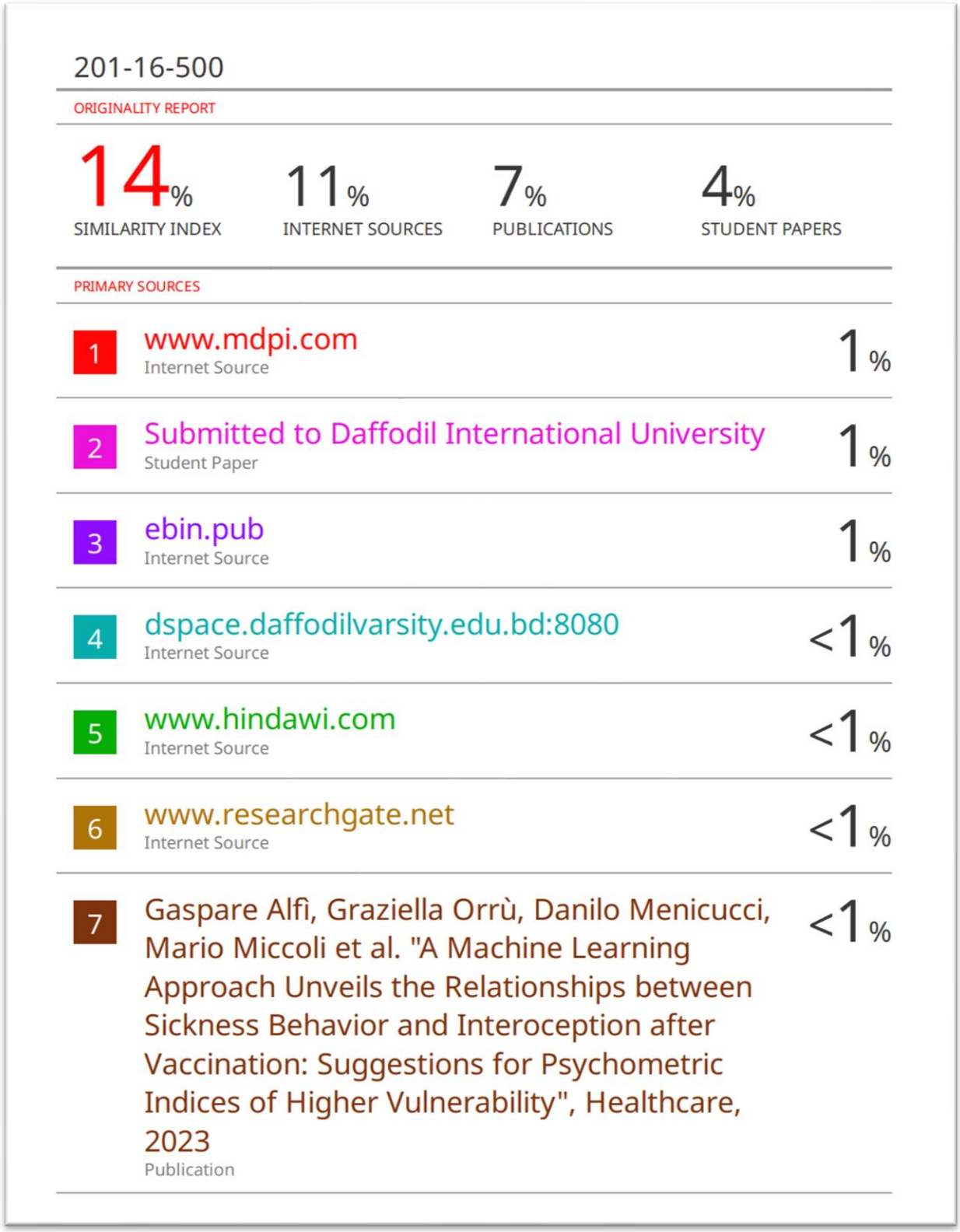
[19] Chowdhury, M. N. R., Ahmed, E., Siddik, M. A. D., & Zaman, A. U. (2021, April). Heart disease prognosis using machine learning classification techniques. In 2021 6th International Conference for Convergence in Technology (I2CT) (pp. 1-6). IEEE.

[20] Nayeem, M. J. N., Rana, S., & Islam, M. R. (2022). Prediction of Heart Disease Using Machine Learning Algorithms. European Journal of Artificial Intelligence and Machine Learning, 1(3), 22-26.

[21] Shahiduzzaman, M., Biswas, N. H., Momin, M., & Sikdar, R. (2022, February). Prognosis of Cardiovascular Disease Using Machine Learning Procedures. In 2022 International Conference on Advancement in Electrical and Electronic Engineering (ICAEEE) (pp. 1-6). IEEE.

[22] Issue-I, T. A. (2017, November 15). Blood pressure range. The Financial Express. https://thefinancialexpress.com.bd/views/letters/blood-pressure-range-1510760557

[23] Khan, S. S., Ning, H., Wilkins, J. T., Allen, N., Carnethon, M., Berry, J. D., ... & Lloyd-Jones, D. M. (2018). Association of body mass index with lifetime risk of cardiovascular disease and compression of morbidity. JAMA cardiology, 3(4), 280-287.



201-16-500

ORIGINALITY REPORT

14%	11%	7%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.mdpi.com Internet Source	1%
2	Submitted to Daffodil International University Student Paper	1%
3	ebin.pub Internet Source	1%
4	dspace.daffodilvarsity.edu.bd:8080 Internet Source	<1%
5	www.hindawi.com Internet Source	<1%
6	www.researchgate.net Internet Source	<1%
7	Gaspare Alfì, Graziella Orrù, Danilo Menicucci, Mario Miccoli et al. "A Machine Learning Approach Unveils the Relationships between Sickness Behavior and Interoception after Vaccination: Suggestions for Psychometric Indices of Higher Vulnerability", Healthcare, 2023 Publication	<1%