CARDIOVASCULAR DISEASE PREDICTION: A MACHINE LEARNING-BASED APPROACH

BY

PRANAB CHANDRA BARMAN
ID: 201-15-14151

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

Mr. Md. Aynul Hasan Nahid

Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY
DHAKA, BANGLADESH
July 15, 2024

APPROVAL

This Project titled **"Cardiovascular Disease Prediction: A Machine Learning-Based Approach"**, submitted by Pranab Chandra Barman and 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 July 15, 2024.

BOARD OF EXAMINERS

____

Narayan Ranjan Chakraborty (NRC) — **Board Chairman**
Associate Professor & Associate Head
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

____

Mushfiqur Rahman (MUR) — **Internal Examiner 1**
Senior Lecturer
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

15.7.24
____

Israt Jahan (IJN) — **Internal Examiner 2**
Senior Lecturer
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

15.07.24
____

Dr. Abu Sayed Md. Mostafizur Rahaman (ASMR) — **External Examiner**
Professor
Department of Computer Science and Engineering
Jahangirnagar University

DECLARATION

I hereby declare that this project has been done by me under the supervision of **Mr. Md. Aynul Hasan Nahid, Lecturer, Department of Computer Science and Engineering, Daffodil International University.** I 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:

Mr. Md. Aynul Hasan Nahid

Lecturer
Department of CSE
Daffodil International University

Submitted by:

PRANAB CHANDRA BARMAN

PRANAB CHANDRA BARMAN
ID: 201-15-14151
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First and first, I would like to offer my most sincere gratitude and appreciation to the All-Mighty Allah, whose divine grace has made it possible for me to successfully finish this research project for the last year of my studies.

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

We would like to express our heartiest gratitude to Professor **Dr. Sheak Rashed Haider Noori,** for his kind help in finishing our project and also to other faculty members and the staff of the Department of CSE, Daffodil International University.

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

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

ABSTRACT

Cardiovascular disease is a common condition that frequently results in deadly consequences, affecting a significant number of people in their middle or later years. Globally, cardiovascular diseases are considered the deadliest illnesses, contributing to the highest death rates. Heart palpitations, nausea, and chest pain are common signs of cardiovascular disease. Significant risk factors for cardiovascular disease include age, gender, high blood pressure, stress, improper lifestyle choices, and family history. This study used eight machine learning classifiers to make predictions about cardiovascular disease more accurate. These were support vector machines, random forests, decision trees, gradient boosting, K-nearest neighbors, Gaussian Naive Bayes, MLP, and logistic regression. In the heart disorders dataset, the K-Nearest Neighbors model performed the best, achieving 86% accuracy, 86% precision, 90% recall, an 87% F1-score, and a ROC AUC value of 0.8909 in the cardiovascular disease dataset. In the healthcare sector, machine learning (ML)-based prediction models offer a more efficient way to support patient diagnosis.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

A variety of conditions affecting the cardiovascular system are together known as cardiovascular disease (CVD). This generic term encompasses several different diseases, including coronary artery diseases, which can result in heart attacks, heart failure, stroke, and other vascular disorders. Cardiovascular diseases (CVDs) and morbidity worldwide, accounting for a substantial portion of the global health burden. Early identification and effective care are crucial to enhancing health outcomes for patients and decreasing the societal burden of cardiovascular diseases. Recent advances in machine learning (ML) techniques have presented exciting new possibilities for improving CVD detection through the analysis of various patient data sets.

Cardiovascular Disease (CVD) is the most widely recognized deadly infection worldwide. Compared to other illnesses, it is the primary cause of death for a greater proportion of the population each year. The death toll from cardiovascular disease (CVD) was 17.9 million in 2016. Taking into account that, worldwide, it accounts for 31% of all fatalities. Cardiovascular disease (CVD) accounted for 37% of the 17 million less-thanideal closures (those < 70) in 2015 owing to non-infectious disorders, and 82% of these closures occurred in countries with low yields. Cardiovascular disease (CVD) can be totally avoided by addressing known risk factors, such as tobacco use, unhealthy eating and weight habits, physical inactivity, and hazardous alcohol consumption under circumstances that impact a large population. People with cardiovascular disease (CVD) or those at high risk for the disease (because of a health condition that is well-managed, like hypertension, diabetes, hyperlipidemia, or another risk factor) necessitate a gradual initiation under close monitoring utilizing brief prescription, as determined by the individual. When everything is said and done, the last stages of cardiovascular disease, or CVD, are the development of blood clusters and the expansion conduits. Blood important organs such as the heart, brain, kidneys, and eyes may also be associated with it. Keeping to cardiovascular disease despite the fact that it mortality and disability. In most cases, stressful situations are the root cause of heart attacks and strokes, which manifest as a blockage that prevents blood from reaching the heart or brain. The most widely accepted explanation for this is the rise in greasy stores, which are the most inward vein dividers. Obesity, poor dietary habits, and cigarette use are often the combined causes of cardiovascular disease.

This study, we investigate the feasibility of using machine learning (ML)-based model predictions to diagnose cardiovascular illnesses. Our goal is to create reliable models that can use the abundance of data available, such as the UCI cardiovascular illness dataset, which contains a wide range of biochemical and demographic factor CVDs. Machine learning (ML) uses complex algorithms and computational techniques to make it easier to find complex correlations between parameters in large-scale datasets. By doing so, it will be possible to identify additional risk factors and improve forecast accuracy over current methods.

1.2 Motivation

Investigating machine learning for cardiovascular disease diagnosis makes sense for several reasons:

- Tackling morbidity as well as the two main global causes of death, coronary heart disease and stroke.
- Being aware of the limitations of traditional risk assessment methods regarding screening the whole population and determining the complex causal linkages for Cardiovascular and other diseases.
- Developing accurate, individualized models of prediction that can identify cardiovascular conditions early on with state-of-the-art computational tools and algorithms.
- Early detection of high-risk individuals enables targeted preventative measures and enhanced risk assessment.
- This aims to improve health outcomes and decrease the burden that CVDs place on individuals and healthcare systems by promoting early diagnosis, lifestyle changes, and pharmacological therapy.
- Through the exciting possibilities of machine learning, major gaps in cardiac care can be filled, and more effective management and preventative techniques can be offered.
- Made strides in the realm of cardiac wellness by utilizing the transformative potential of machine learning to offer novel approaches to disease detection and treatment.
- Prioritizing cardiovascular health and lowering the societal and financial impact of CVDs through focused interventions and preventative measures in line with global health efforts and public health objectives.

All things considered, using machine learning to the detection of cardiovascular diseases holds considerable promise for increasing knowledge about healthcare efficiency, decreasing healthcare expenses, and bettering patient outcomes cardiovascular health.

1.3 Research Questions

- **What effects do various feature selection strategies have on the predictive models' ability to detect cardiovascular disease?**

This question intricacy, interpretability, and generalizability of prediction models employed in cardiovascular disease diagnosis are influenced by different feature selection techniques. Although filter techniques examine features outside of the model, they may miss feature interactions and result in lower processing costs. Wrapper techniques select subsets of attributes based on the model's performance; they boost anticipated accuracy but need more processing power. Embedded techniques balance computational efficiency and predictive power by incorporating feature selection into the model training process.

- **Can machine learning algorithms accurately distinguish between various types of cardiovascular disorders?**

In fact, machine learning algorithms can accurately distinguish between various types of cardiovascular diseases by finding patterns and connections in medical data. They can differentiate between problems including coronary artery disease, cardiac failure, arrhythmias, and others based on data from clinical symptoms, medical history, diagnostic tests, and imaging data. Modern and ensemble approaches have the potential to accurately diagnose illnesses by spotting complex links and patterns in the data. However, the precision of these models is affected by a lot of factors, such as the quantity and accuracy of the input information, the additional features used, and the method used. To make these models useful for illness categorization, they need to be refined and validated with multiple sample datasets on an ongoing basis.

- **When it comes to cardiovascular disease diagnosis, which major demographic and physiological variables most affect the accuracy that ML algorithms are anticipated to achieve?**

This inquiry aims to address the impact of demographic and physiological factors accuracy in identifying CVD. This includes comorbidities like diabetes and hypertension as well as age, gender, cholesterol, and blood pressure data from ST_Slope, Oldpeak, ExerciseAngina, and MaxHR. various elements are essential for risk evaluation and diagnosis as well as for evaluating the causes and consequences of various disorders. By examining patterns and correlations in this data, artificial intelligence techniques can be trained to more accurately identify individuals who are at a greater likelihood of cardiovascular events.

- **How can we fix the problem of biased datasets that prevent machine learning algorithms from accurately predicting heart disease?**

To tackle data gaps in forecasting of cardiovascular disorders, these key tactics can be utilized: collective learning, reproducing, cost-sensitive acquiring knowledge, and performance measures adjustment (ROC AUC, F1-score). Increasing data collection, ensuring data quality, and leveraging subject-matter expertise can also help decrease imbalances. When selecting an algorithm and tweaking parameters, it is important to keep the class distribution in mind to avoid creating biased models.

- **How can we fix the problem of biased datasets that limit the applicability of algorithms for learning algorithms for cardiovascular disease prediction?**

Resolving symmetry requires the use of techniques such ensemble approaches, tweaking performance metrics, and reducing the majority class while oversampling the minority. It is crucial to ensure data quality (which necessitates subject knowledge) and to collect a more diversified and representative group of data. Using a methodology and tweaking its parameters should take class distribution into account.

1.4 Rationale of the Study

Many important factors lend credence to the notion that cardiovascular disease prediction models driven by machine learning should be investigated. To start with, machine learning approaches provide a robust set of tools for evaluating complicated and massive datasets, allowing for the discovery of correlations and patterns that traditional statistical methods would overlook. To create more detailed and precise risk combine several patient data sources. These sources may include chromosomal identities, health records, radiological results, biomarkers that are and demographic data. Investigating ML-based prediction models to assess the extent of cardiovascular diseases is warranted for a few compelling reasons. To begin with, machine learning techniques offer a powerful toolkit for exploring large and complex datasets, allowing one to detect nuances in relationships and patterns that traditional statistical methods may miss. Utilizing the computational power in machine learning algorithms, researchers have the potential to build more comprehensive and accurate risk prediction models by merging various patient information, such as genetics characteristics, medical histories, imaging findings, biomarkers, and demographic data. of what danger signs cardiovascular disease (CVD) is a larger goal of diagnostic tools based on machine learning. These models guide future research to better understand the complicated etiology of CVDs and provide focused treatments for populations at risk by identifying novel biomarkers as well risk variables, and disease pathways. Reducing the deaths, morbidity, and related medical expenses of cardiovascular illnesses (CVDs) is why early identification and intervention are crucial components of successful cardiovascular care. In conclusion, the need to improve early detection, overcome the limits of present risk assessment methods, and increase our knowledge of the biology of CVDs motivates research on AI-based prediction models to diagnose cardiovascular disorders. The new devices can revolutionize the treatment of cardiovascular disease by using algorithms that employ machine learning to provide faster, more accurate, and more tailored risk assessments for people at risk of acquiring these conditions.

1.5 Project Management and Finance

The success and longevity of any business depends on meticulously recording and supervising initiatives. Thinking over the study proposal "Cardiovascular Diseases Prediction: A Machine Learning-Based Approach," the answer is obvious. Collecting data, training models, and evaluating the project are all steps that require meticulous planning, collaboration, and execution. The key to an effective and efficient workflow is a project plan that details important milestones, allocates resources wisely, and sets deadlines. At the same time, sufficient funding for hardware, computational resources, and potential research collaborations must be carefully managed for the project to be viable in the long run. The most efficient use of funds while maintaining fiscal responsibility can be achieved with an honest and careful financial strategy. To overcome the study's complexity, allocate resources in a way that supports the research goals, and move the project forward efficiently, the research team needs these integrated approaches to finance and project management as foundational building blocks.

1.6 Expected Output

Machine learning is a component of our subject matter. We utilize only eight of the numerous types of algorithms for machine learning. Decision Tree, Gradient Boosting, K-nearest neighbor, Support Vector Machine, and Random Forest, Gaussian Nave Bayes, MLP, and Logistic Regression comprise these models. This investigation anticipates the creation of an intelligent medical system that is both effective and reliant on machine learning techniques. The system will ease the process of identifying the patient's cardiac condition by utilizing these algorithms and methodologies. The top-performing algorithms are as follows: Decision Tree (82) and Logistics Regression (85%), Gaussian Naive Bayes (85%), Support Vector Machine (87%), Extreme Gradient Boosting (86%), and Random Forest (88%). When compared to other algorithms, Random Forest produces the best outcome. The goal of this study is to improve diagnostic capabilities by analyzing and implementing machine learning methodologies. Integrating the prediction system in conjunction with other systems, like a SACS that takes context into account, reinforces the health care system's safety fundamentals.

1.7 Report Layout

This study report's format was intentionally designed to support a logical progression of concepts, procedures, findings, and conclusions.

Chapter 1: Introduction

The study outlining the background, and expected results, providing a roadmap for the entire report.

Chapter 2: Background

Provides a comprehensive background, outlining terminology, significant works, comparative analysis, and scope, ensuring readers understand the complexity of the topic.

Chapter 3: Methodology

Details research methodologies, data collection techniques, machine learning model architecture, utilizing visualizations, statistical analysis.

Chapter 4: Experimental Result

Presents a comprehensive investigation of the model’s results.

Chapter 5: Impact on Society

Examines the societal, environmental, and ethical implications of machine learning models, addressing ethical issues, and proposing a sustainability report intention.

Chapter 6: Summary, Conclusion, Recommendation, and Implications for Future Research

Summarizes the research, highlighting significant findings, providing conclusions, practical applications, and potential future research opportunities, offering closure and stimulating further exploration.

CHAPTER 2

BACKGROUND

2.1 Preliminaries

In order to lay the groundwork for the main ideas and methods used in this study on "Cardiovascular Disease Prediction: A Machine Learning-Based Approach," it is necessary to define important terms and provide some preliminary information. Interpretability is essential in healthcare applications. As a result, doctors might learn to trust the algorithm's predictions and follow its advice to the letter. The reader should be well-versed in those fundamental concepts in order to grasp the study's complexities and the accuracy of contemporary methods for identifying cardiovascular illnesses.

2.2 Related Works

One study Herdian, C., Widianto, (2024) proposed heart disease is an important global health concern since it causes a significant number of deaths globally. Predictive methods, particularly those incorporating machine learning, aim to reduce mortality rates by forecasting the likelihood of heart disease. Using Supervised Learning likelihood heart disease occurrence based on specific characteristics like Exercise Angina, ST Slope, and Chest Pain Type. By using the K-Nearest Neighbors technique after Hyperparameter Tuning GridSearch, this model was able to attain an appropriateness level of 85.23% (data testing) and 82.61% (data validation) using Hyperparameter and Feature Engineering strategies [1].

Another work Naser, Marwah Abdulrazzaq, (2024) In this work, we will look at how machine learning (ML) can help tackle the problems caused by cardiovascular diseases (CVD), the biggest killer on a global scale. Problems with delayed treatment and incorrect diagnosis are driving research into artificial intelligence (AI) approaches, especially ML algorithms, to improve disease detection. To offer a thorough evaluation of ML's potential in CVD prediction, covering topics such as types of CVD, feature selection, model evaluation, data preprocessing, evaluation metrics, recent trends, and suggestions for future research [2].

In Alrais, A. I. (2024) heart disease emphasizes the significance of early identification to improve outcomes because it is the leading cause of death. This study employs various machine learning methods to categorize datasets pertaining to heart disease and ascertain if a person possesses the illness. These methods include naïve Bayes, multilayer perceptron, random forest. The dataset is primarily utilized for training purposes with 80% coming 20% for testing. The multilayer perceptron technique achieves over 84% accuracy, while the decision tree strategy performs badly, with accuracy only reaching 79% [3].

The author of Louridi, N. (2021) aimed to enhance the efficacy of machine learning models using data pretreatment by normalization, which substitutes the average rather than disregarding it, value each feature when data is lacking. As part of the assessment, we compared the kNN, NB, and SVM models. An accuracy of 82.8% was reached by the linear kernel support vector machine (SVM) [4].

Similarly, Mohan et al. (2018) recommended (with an accuracy of 81.7% using the UCI Cleaveland dataset) for the purpose of predicting cardiac events using hybrid machine learning approaches. [5].

Dinesh et al. (2018) proposed employing ML algorithms to predict CVD using logistic regression (83.5%) after examining four UCI heart disease datasets [6].

The author of In Kusuma, (2018) developed a method to identify heart problems by combining iris scans with a smartphone. Feature construction relied the thresholding procedure yielded an accuracy performance value of 83.3% for the test data. These studies show that iridology is a reliable diagnostic tool for treating coronary artery disease. [7].

The author of Mohammad (2019) have created a model to predict cardiovascular diseases using many feature combinations. The UCI-hosted Cleveland database datasets database for machine learning. The results indicate that the suggested model can predict heart diseases with an accuracy of 84.4% [8].

This author of Princy et al. (2020) makes use of six ML classification algorithms to forecast the occurrence of cardiovascular illnesses. The suggested models have been tested using the cardiovascular disease[9].

The study found that decision trees were the most accurate classifiers (2020). When fed the patient's medical history, DT mode had a 78% success rate in predicting the patient's ailment [10].

The author of to improve forecast accuracy, more work needs to be done. Because of this, Bashir et al. (2019) research combines multiple datasets linked to heart disease in the experiment and focuses on a features selection to boost accuracy. Accuracy has been enhanced, according to the results [11].

Sabab et al. (2016) offers a way to improve the accuracy of heart illness diagnoses and classification models by feature selection techniques. The models employed by the authors to detect cardiovascular disease had accuracy rates of 85.8%, 84.80%, and 79.9%, respectively [12].

The author of the optimal prediction technique for heart illness is identified by Zunaidi et al. (2020) using the Heart illness Dataset and five categorization methods Predicting early heart disease with the best accuracy, according to the data, is Multilayer Perceptron Neural Networks (85.03%) [13].

Similarly, with an accuracy of 80.17 percent, Khemphila et al. (2011) used Multi-Layer Perceptron (MLP) to diagnose cardiac illness using the same dataset [14].

Kavitha et al. (2021) uses a combination of decision trees, random forests, and a third algorithm that combines the two to forecast cardiac disease. Cleveland Heart Disease was the dataset utilized. According to the hybrid model was the most accurate, scoring 85.7% [15].

Princy et al. (2021) was determined using data mining techniques considering several factors. Increases in both the number of characteristics and features were associated with improvements in accuracy, according to the scientists. An accuracy rating of 80.6% was shown by the inquiry's results [16].

In a similar vein, Gawali et al. (2016used data mining techniques such as Naive Bayes and K-Means to determine the probability of cardiac disease. The study produced an 84% accuracy rate [17].

In a similar vein, Maiga (2018) offers four methods for machine learning that account for blood pressure, BMI, and cholesterol levels when identifying CVD. The proposed study applies Naïve Bayes, KNN, Analyzing the UCI health record dataset using logistic regression and Random Forest. The most accurate prediction values from the random forest classifier were 80% specificity, 65% sensitivity, and 73% accuracy [18].

Mohan et al. (2019) finds the pertinent aspects of the dataset linked to utilizing machine learning methods for heart illness. Hybrid HRFLM is the result of combining the features of Random Forest, with an accuracy percentage of 85.7%, it was clear that the outcome was spot on [19].

Additionally, Arabasadi et al. (2017) describe a method for the purpose of detecting and diagnosing anomalies in the coronary arteries utilizing trained on the by initializing the network's weights using genetic algorithms, the proposed strategy improved the network's performance by 10% [20].

Latha and associates (2019) Provide a unique method for enhancing the classification models' performance that combines features resembling to produce accurate prediction results. Using enfeebling approaches increased the week classifiers' accuracy by 7%. The suggested method makes use of classifiers including Bagging, boosting, stacking, majority voting, Random Forest, C4.5, MLP, and Naïve Bayes are all included. The most significant improvement in accuracy occurred just before the majority vote [22].

2.3 Comparative Analysis and Summary

The linked works discussed in Section 2.2 are summarized and compared in this section.

Table 2.3.1: Previous Analysis Table

Author	**Titles**	**Classifier Model**	**Accuracy**
Herdian, C., Widianto, (2024)	A Case Study on Heart Disease Prediction	K-Nearest Neighbors	85%
M.Mijwil, Alaa K. Faieq(2024)	Early Detection of Cardiovascular Disease Utilizing Machine Learning Techniques	Decision tree	84%
Louridi, N. (2021)	Identification of cardiovascular diseases using machine learning	SVM	82%
Dinesh et al. (2018)	Prediction of cardiovascular disease using machine learning algorithms	logistic regression	81%
Rajib KumarHalder(2020)	Cardiovascular Disease dataset	DT	78%
Maiga et al. (2018)	Comparison of Machine Learning Models in Prediction of Cardiovascular Disease	Naïve Bayes, KNN, Random Forest, and logistic regression	73%
Princy et al. (2021)	Human heart disease prediction system using data mining techniques	Naïve Bayes, ID3, and neural networks	80.6%
Mohan et al. (2019)	Effective heart disease prediction using hybrid machine learning techniques	Linear Method (LM) and Random Forest (RF)	83.7%.

2.4 Scope of the Problem

There are a lot of potential problems when I start a new job. able to write novels using fresh algorithms and data collection tools. The magnitude of the problem is reflected in the challenges associated with accurately and promptly diagnosing cardiovascular illnesses. Despite medical advancements, traditional risk assessment methods remain imprecise. Machine learning proves to be promising for personalized interventions and sickness prediction by analyzing massive volumes of healthcare data. But there are still concerns, such the model's interpretability, ethical dilemmas, and complicated data. Interdisciplinary cooperation will be required to get over these challenges and produce transparent, trustworthy predictive models that are simple to apply in clinical practice. In the end, this will lead to better patient outcomes. And people can receive the appropriate results. For people to use with ease.

2.5 Challenge

One of the main challenges in applying machine learning to the diagnosis of cardiovascular illness is ensuring the validity and interpretability of predictive models. Healthcare data is complicated due to its heterogeneity and dimensionality, which makes it difficult to design and validate models. Moreover, maintaining impartiality and avoiding prejudice in the process of making accurate forecasts remains a challenging task. Data protection and other ethical concerns like patient consent make model adoption more challenging. To get past these challenges, interdisciplinary collaboration, innovative algorithmic solutions, and stringent validation procedures are required. There are a few challenges that need to be solved to completely apply machine learning to transform cardiovascular care and improve patient outcomes.

CHAPTER 3

RESEARCH METHODOLOGY

Cardiovascular disease prediction as a whole is described and illustrated in Figure 3.

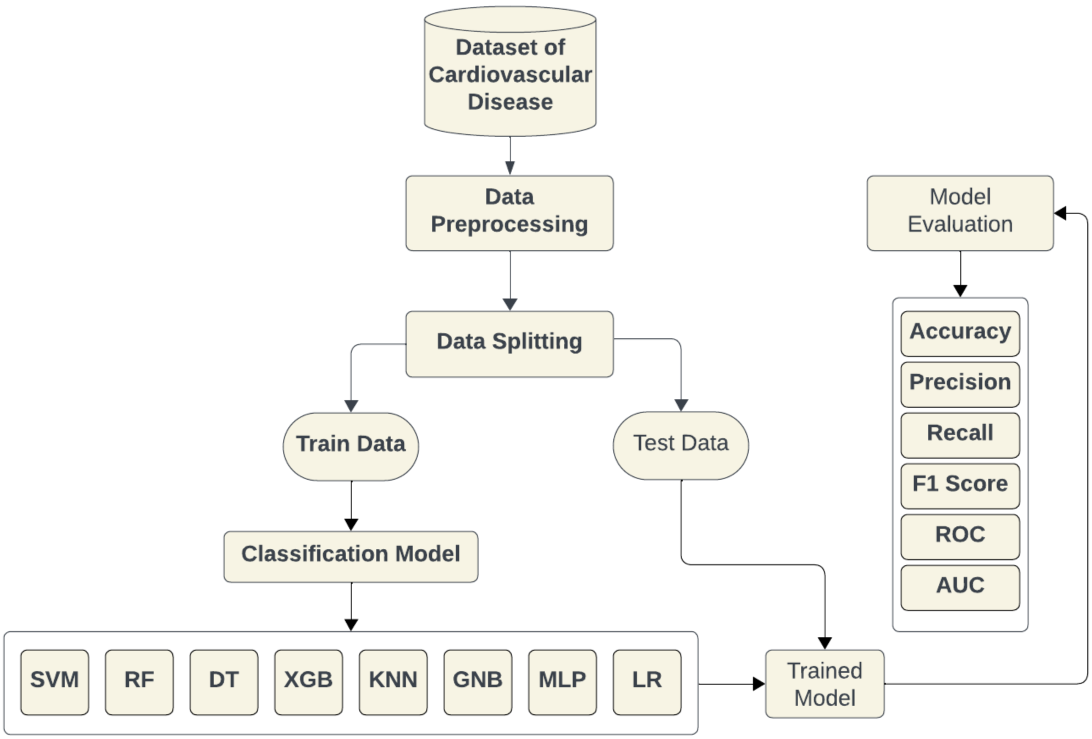


Figure 3: Detailed method of Cardiovascular Disease Prediction

3.1 Research Subject and Instrumentation

Research Subject:

We will discuss the detailed research approach in this section. Numerous strategies can be employed to solve any analysis. First, we gathered the survey information. Next, a machine learning algorithm is chosen. We have mentioned that we are employing eight distinct machine learning techniques, therefore before we can create the model and fit the algorithm, we need to decide how we will collect data. After that, using this data collection, the model is trained. That is the fundamental idea behind feature selection. After the data is fitted into different machine learning algorithm models and trained using a training data set, only a significant portion of the data required to evaluate our model is available. Next, the model's accuracy is assessed. We've been providing an overview with our standard process flow chart, but we'll go over some of the algorithms using equations and diagrams to help you understand how they operate. This is a summary of the complete research project process, including the study endeavor.

Instrumentation:

This was used a wide variety of equipment and devices to carry out our investigation. Because of its popularity in data analysis and machine learning, Python was the primary language of choice. To build our machine learning, we used well-known libraries like NumPy, Seaborn, Pandas, and Python. These libraries' extensive collections of tools and functions made data preliminary processing, training models, and evaluation easier. Python and a variety of tools and libraries for data analysis and machine learning were important to our study's subjects—cardiovascular illness and instrumentation.

3.2 Data Collection Procedure

In this section, the data set was obtained from Kaggle and slightly altered to satisfy our needs. Every procedure then needs to use the dataset. The eight methods we utilize all behave differently, as we have shown earlier, therefore preprocessing the data is required before it can be correctly fitted into our model.

3.3 Statistical Analysis

1. In total, there are twelve columns.
2. The total number of rows is 918.
3. This model was trained on 80% of the available data, with the remaining 20% being used for testing.
4. A CSV file is saved with the dataset.

3.4 Proposed Methodology

Support Vector Machine is one of the classification methods that will be used in this study, The following methods are included: Decision Tree, Logistic Regression, K-Nearest Neighbors, Gradient Boosting, MLP, and Random Forest. Deploying are also necessary for finding the best classifier for cardiovascular disease detection. Figure 3 above outlines the processes that will be followed in the research.

3.4.1 Data Preprocessing

In terms of data preprocessing involves Outlier, remove outlier, correlation analysis, Standard Scaler, and validating data to ensure it's accurate, consistent, and ready for analysis or machine learning, enhancing model performance and reliability.

3.4.2 Split train and test data

Based on the data study, we split the time between training and testing in a 80/20 split to help with cardiovascular disease prediction.

3.4.3 Classification Model

We have discussed the Classification Model as depicted in 3.5.

3.4.4 Model Evaluation

In classification problems, the method of determining which metrics are most appropriate for assessing a given classifier's performance about a specific dataset is contingent upon the number of things like expected results and class balance. A classifier may be evaluated using only one performance indicator and not any others, or the opposite may occur. As a result, the generic performance measurement of the classifier lacks a single, well-defined indicator. The accuracy, precision, recall, F1 score, and area under the curve (AUC-ROC) are some of the metrics that will be computed to evaluate the algorithms' performance on the test data that are chosen for this study.

The following four groups provide the metrics used here:

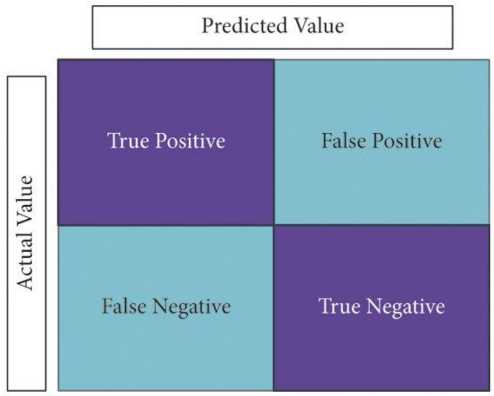


Figure 4: Block diagram of the confusion matrix.

True Positives (TP): In this scenario, both the actual and predicted classes of the instance were 1.

False Positives (FP): an instance when the model predicted a true value of 1 and the true value was 0 (False).

True Negatives (TN): an instance when both the true class and the model prediction were false.

False Negatives (FN): an instance when the model predicted 0 (False) but the true class was 1 (True).

Accuracy – the mean number of right predictions is a measure of accuracy. On the other hand, the imbalanced dataset makes this less than robust.

$$\text{Accuracy} = \frac{\text{Correct pridiction}}{\text{Total pridiction}}$$

Precision – known as positive predictive value, evaluates a model's accuracy in classifying data. When it comes to imbalanced datasets and multi-class classification, this matrix is strong.

$$\text{Precision} = \frac{\text{True Positive}}{(\text{True Positive} + \text{False Positive})}$$

Recall – This statistic evaluates how well a model identifies true positives relative to all true positives.

$$\text{Recall} = \frac{\text{True Positive}}{(\text{True Positive} + \text{False Negative})}$$

F-score – popularly known as the balanced F-score or F-measure. is the average of the two metrics, recall and precision, given a weight.

$$\text{F-score} = 2 * \frac{(\text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall}))}$$

Cross-Validation: To prevent overfitting, this method divides data into two sets: training and testing.

Hyperparameter Tuning: Using techniques like GridSearch, model parameters can be optimized to increase performance.

3.5 Implementation Requirements

Support Vector Machine

Features can be classified by locating the plane of motion which reduces the distance between them across the support vectors. Support vector machines, also known as SVMs, can handle all complex in addition to linear classifiers with kernel functions by projected information onto a higher-dimensional space. During training, it minimizes a cost function to determine the hyperplane parameters. SVMs are resistant to overfitting and perform well in high-dimensional domains; nonetheless, regularization and suitable kernel selection are required for optimal performance. Generally, may be computationally costly for very big datasets, but they perform better if there are more attributes than samples. SVMs can handle nonlinear data, prevent overfitting, and be helpful in high-dimensional domains, among other advantages. Despite this, SVMs are widely used because of their versatility and effectiveness in handling difficult classification problems across a range of industries, including image recognition, text classification, and biomedical information technology.

Random Forest

Random Forest is becoming more and more common within the categorization application field as a result of its ensemble learning technique. It creates many decision trees during training and uses various characteristic and data sets to train them. This randomization improves generalization and reduces overfitting. Random Forest creates many bootstrapping versions based on the original data using a technique called bagging. It also chooses at random a subset of the attributes that each branching point among the trees has. Every plant makes an individual forecast of the class label during prediction, with a majority vote determining the final prediction. You can modify the maximum depth, feature subsets, and number of plants generated using the Random Forest settings. It also provides an attribute relevance score that helps identify the qualities that have the biggest impact on the categorization process. Random forest methods are across different data types, scalability, and ease of use.

Decision Tree

One type of machine learning technique that divides the characteristic set into segments and assigns a feature label to each is the decision-tree classifier. Starting from the initial node, the method partitions the data as efficiently as feasible after choosing the feature that maximizes a specific criterion (such as entropy or Gini impurity). This loop keeps going until something breaks it, like when a predefined depth, sample count a node decreased. Decision trees are simple to comprehend and can be utilized in both regression and classification. However, they could overfit, particularly if the plant's depth is not suitably restricted. One way to reduce this and simplify the tree's structure is by pruning. Despite their tendency to overfit, decision trees are widely used because of their simplicity, ease of use, adaptability in handling both numerical and categorical data, and straightforward interpretation. They are very helpful in industries aside from marketing, finance, and medicine because of these features.

Extreme Gradient Boosting

The accuracy and effectiveness of modern ensemble learning methods, such as the Extreme Gradient Boosting Classification algorithm (XGBoost), are well known. XGBoost sequentially adds decision trees, one at a time, to an ensemble, focusing each tree on the errors made by all the trees that came before it. This method is based on the ideas of gradient boosting. By encouraging model simplicity and avoiding overfitting by the application for the first as well as second normalization processes, it improves on ordinary gradient boosting. Utilizing node split optimization, XGBoost's depth-wise tree building enhances loss reduction and speeds up convergence. In order to facilitate hyperparameter adjustment and avoid overfitting, it offers early halting and cross-validation. In addition, the XGBoost exhibits high parallelizability, facilitating the effective utilization in multiple threads and decentralized computing to facilitate training on extensive datasets. Its adaptability allows it to regression, ranking classification. Thanks to its excellent performance and scalability, XGBoost is a popular option for both practical and academic applications. It has won multiple machine learning competitions and is extensively used in a variety of sectors.

K-Nearest Neighbors (KNN)

The K-Nearest Neighbors (KNN) Classifier is a straightforward yet effective technique for solving classification problems. It is predicated on the notion that data points that share similar properties usually fall into the same class. Using a distance metric, usually the Euclidean distance, CNN predicts newly added to decide its class, the new data point's K nearest neighbor’s vote. No assumptions on the distribution of the data are made by KNN as it is not a parametric method. However, because the entire training dataset must be stored for prediction, it is memory-intensive for large datasets. A straightforward and effective technique for handling classification problems. It is predicated on the notion that when two data points share comparable characteristics, they usually fall into the same class. CNN uses a distance metric typically the Euclidean distance. The class of the newly gained point is then determined by voting on its K nearest neighbors. KNN does not assume anything about the distribution of the data because it is not a technique that is descriptive. However, because the entire training dataset must be kept generating predictions, it may be memory-intensive for large data sets.

Gaussian Naive Bayes

The Gaussian Naive Bayes model is a popular probabilistic classifier that does a great job with continuous variable variables. It is used in classification-related applications. It uses the Bayes theorem and is predicated on the notion that attributes are not only developed independently but also uniformly across all classes. As training progresses, GaussianNB calculates the variance and mean of each feature according on each class. When classifying a new event, it considers both the prior chance for each class and the statistical probability determined using the Bayes theorem. This computation is finished by dividing the calculated probability for the actual value of features by the Gaussian distribution of each class. The category with the highest chance of success is considered before the predicted category is designated. Because GaussianNB has a straightforward model structure and minimal parameter estimation requirements, it can perform well on huge datasets.

Multi-Layer Perceptron

Its component parts are a layer for input, a result layers & a disguised layer or layers. Every neuron in the layer that lies next to it is fully coupled to every other neuron using a stimulation mechanism to the sum of its input weights. To provide output predictions during training, MLP uses feedforward propagation to distribute input information throughout a network. Then, it employs reverse propagation to ascertain the strength of neural connections. MLP's architecture and adaptability allow it to recognize complicated non-linear connections in data. The system's effectiveness is determined by how well-tuned its activation functional decision-making, per-layer neurons measure, secret layer weigh in, and hyperparameters are. Regularization methods such as dropout regularization and L2 regularization can help minimize overfitting. Because MLP classifiers can and understand complex patterns, they are of fields bioinformatics, photo among audio recognition, and more. State-of-the-art outcomes are frequently achieved if MLPs are applied to problems with classification, given the computational complexity and requirement for adjustment.

Logistic Regression

We can better understand the relationship between independent factors and outcome likelihood. Despite the name, it is essentially an equation that uses the function of sigmoid activation to translate the input data into options between 0 and 1. To determine the variables (weights) for training, it optimizes the likelihood function. The choice of boundary, which operates on the basis that instances above a certain quantity belong to one category and those below to the other, is what separates the two classes. Logistic regression is a practical, interpretable, and computationally efficient technique for data that can be split linearly. It is widely used in many different industries, including marketing, finance, and healthcare, where readability and simplicity are desired.

CHAPTER 4

EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Experimental Setup

Here, this will talk about the experimental results. We have covered one of our five algorithms in detail. Now, we can assess how effective those algorithms were. We'll also evaluate the accuracy of each of the three approaches for cardiovascular disease prediction. We used Scikit-Learn to compute performance metrics, Matplotlib to visualize data, and Pandas for data management and preprocessing. The datasets utilized for the assignments contained 918 pieces of data.

4.2 Experimental Results & Analysis

This project's result analysis includes a thorough investigation of how well machine learning algorithms perform in terms of optimizing prediction accuracy for cardiovascular disease. The analysis comprised a thorough assessment of a few factors, such as feature significance, model evaluation, algorithm selection, data preprocessing, and result interpretation. The findings provide useful information for accurate cardiovascular disease prediction and shed light on the pros and cons of different techniques.

We can see that in our dataset 410 are not cardiovascular disease attacks on the other hand 508 are affected by cardiovascular disease.

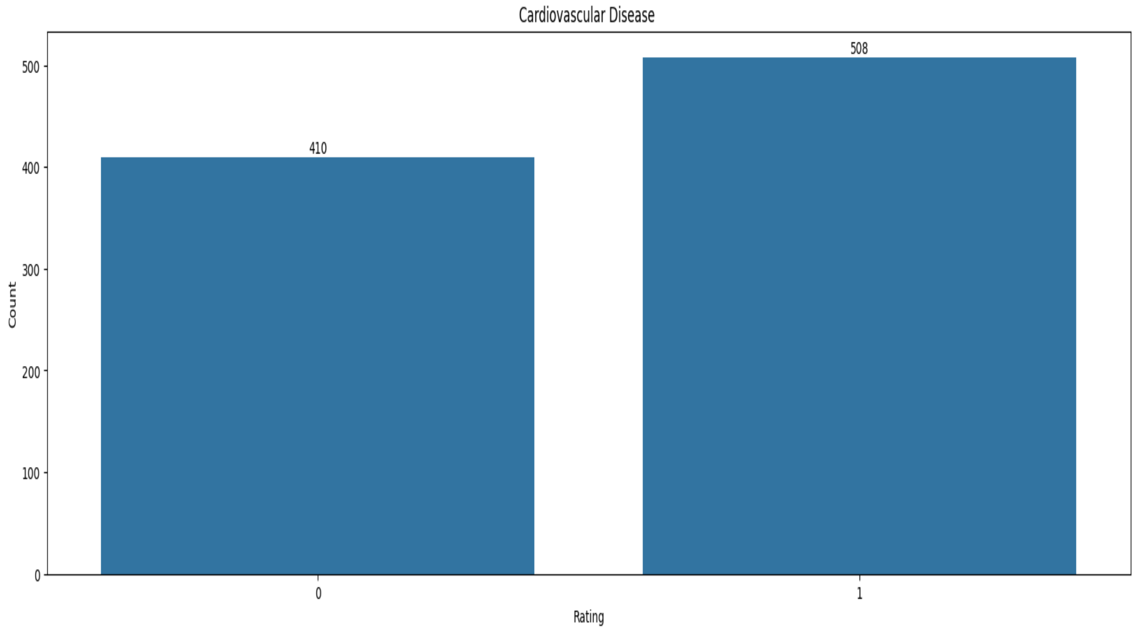


Figure 4.2.1: Cardiovascular disease.

Among them there were 725 males and 193 females.

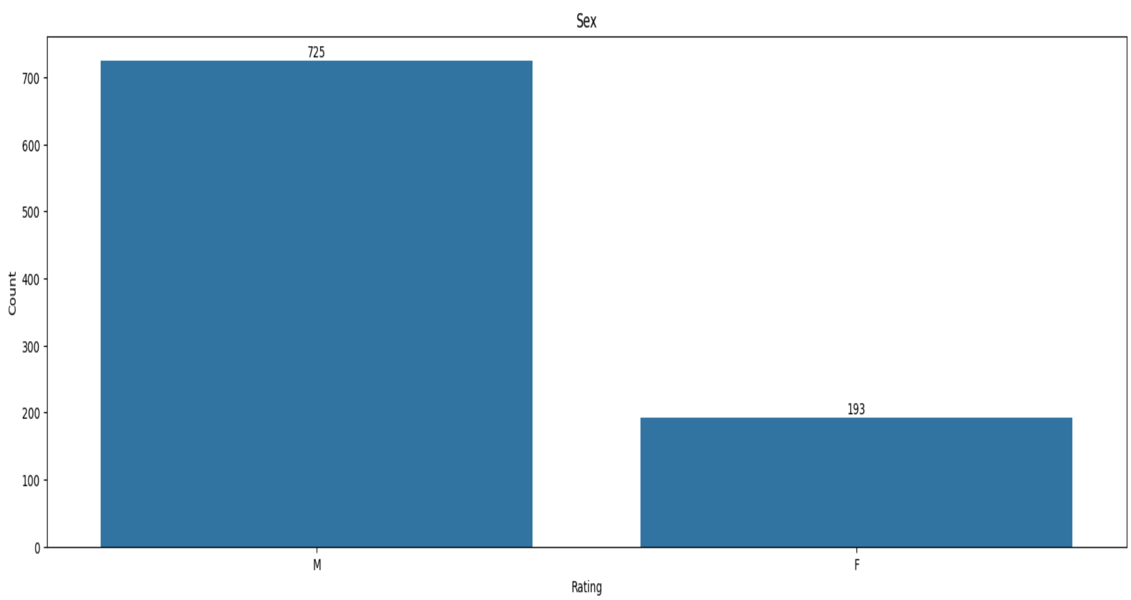


Figure 4.2.2: Bar chart of Sex.

There are 4 types of chestpain: Atypical Angina (173), Non-Anginal Pain (203), Asymptomatic (496) and Typical Angina(46).

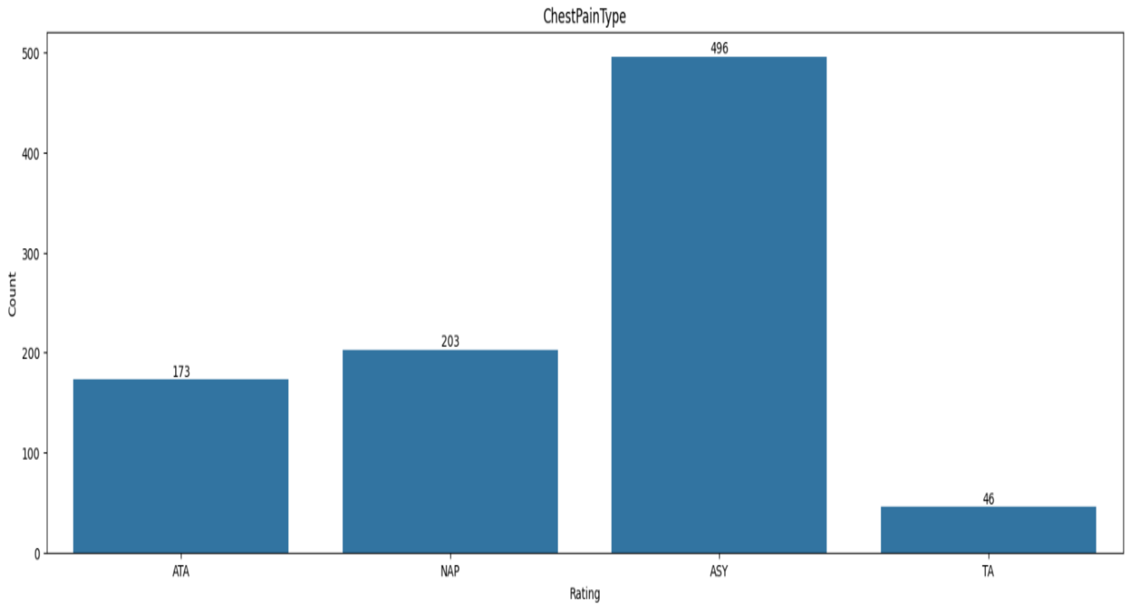


Figure 4.2.3: Bar chart of chestpain.

RestingECG does not present a clear-cut category that highlights cardiovascular disease patients. All the 3 values consist of high number of heart disease patients.

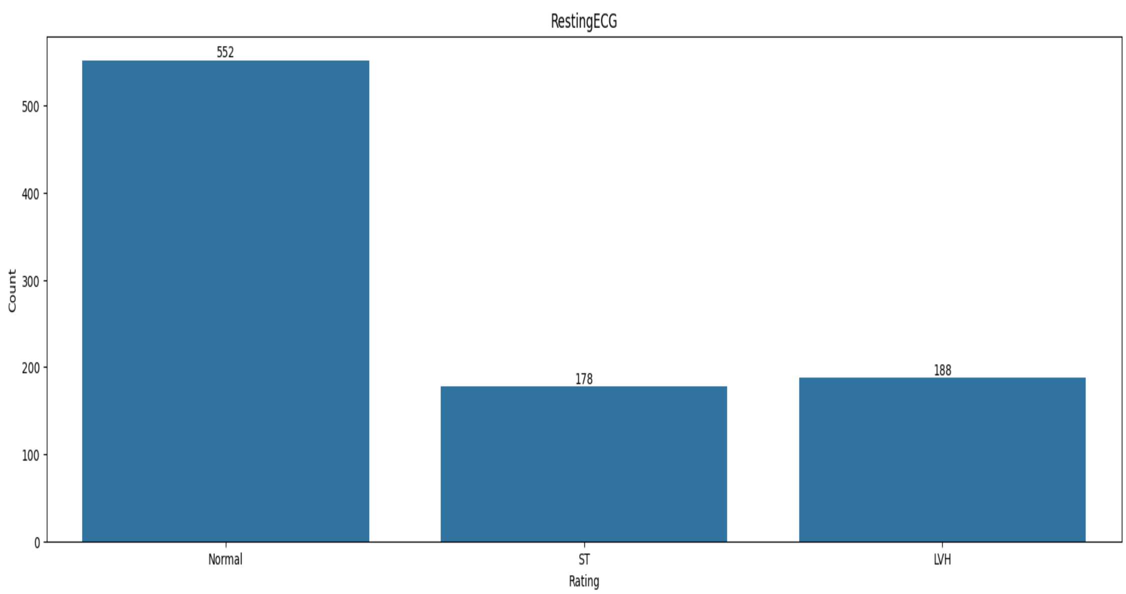


Figure 4.2.4: Bar chart of RestingECG.

The likelihood of being diagnosed with cardiovascular illnesses is high, medium, and low, respectively, when the three states are flat, down, and up.

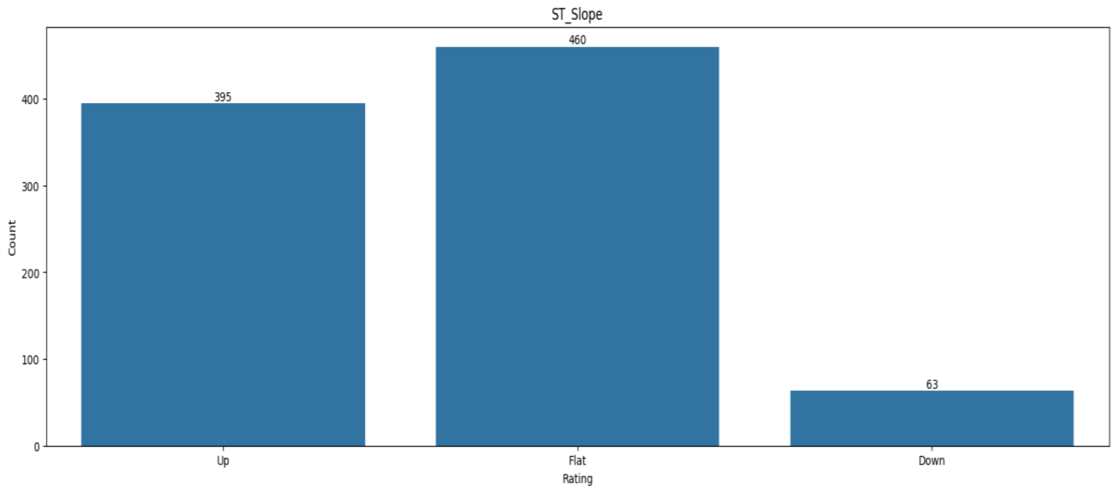


Figure 4.2.5: Bar chart of ST_Slope.

Here we have checked Outlier dataset.

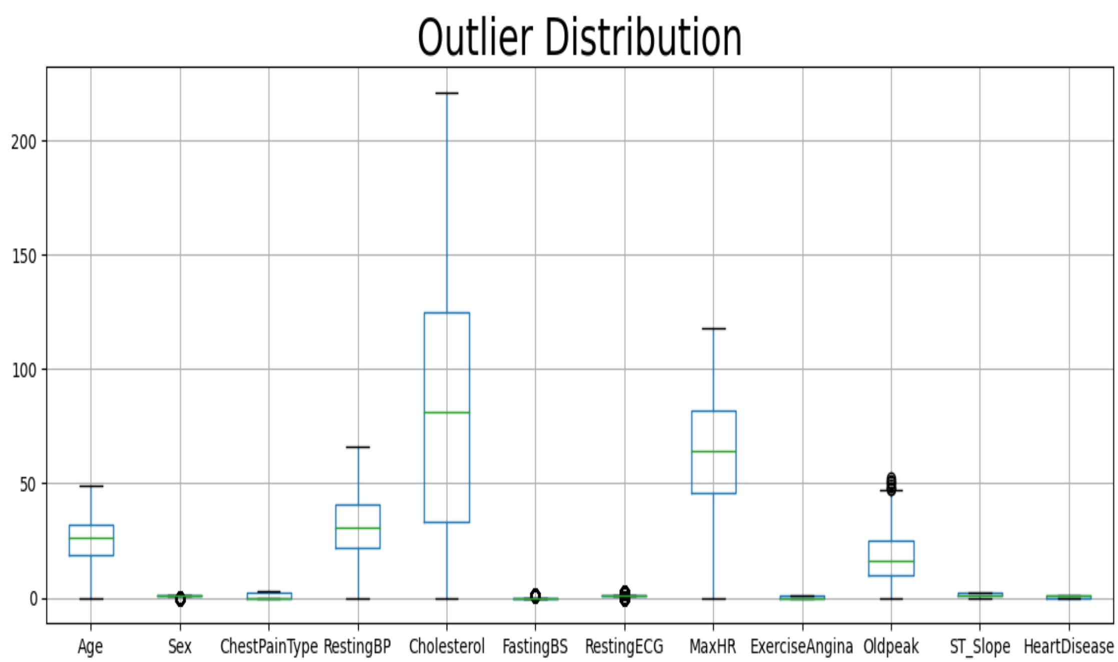


Figure 4.2.6: Outlier of dataset.

After that I check the outlier and remove the outlier value.

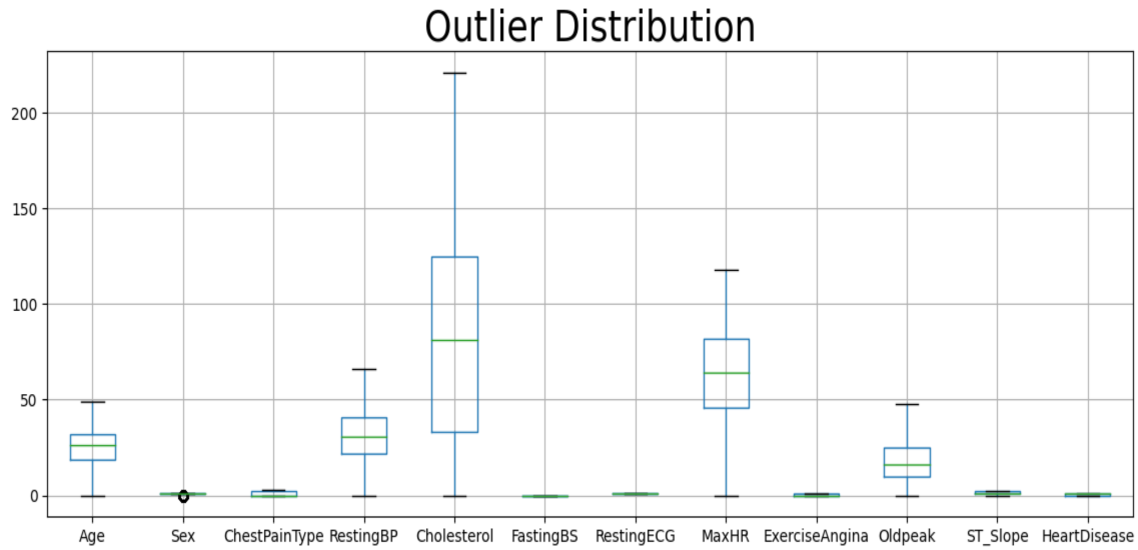


Figure 4.2.7: After removing outlier.

A heatmap of the correlation matrix can be generated to visualize the relationships between variables within the cardiovascular disease dataset. This data visualization technique is highly effective in revealing insights about the correlations between the 12 aspects of the dataset. In a correlation heatmap, which uses colors to graphically represent correlations between variables in a dataset, warmer colors suggest stronger positive correlations while blue colors indicate stronger negative correlations.

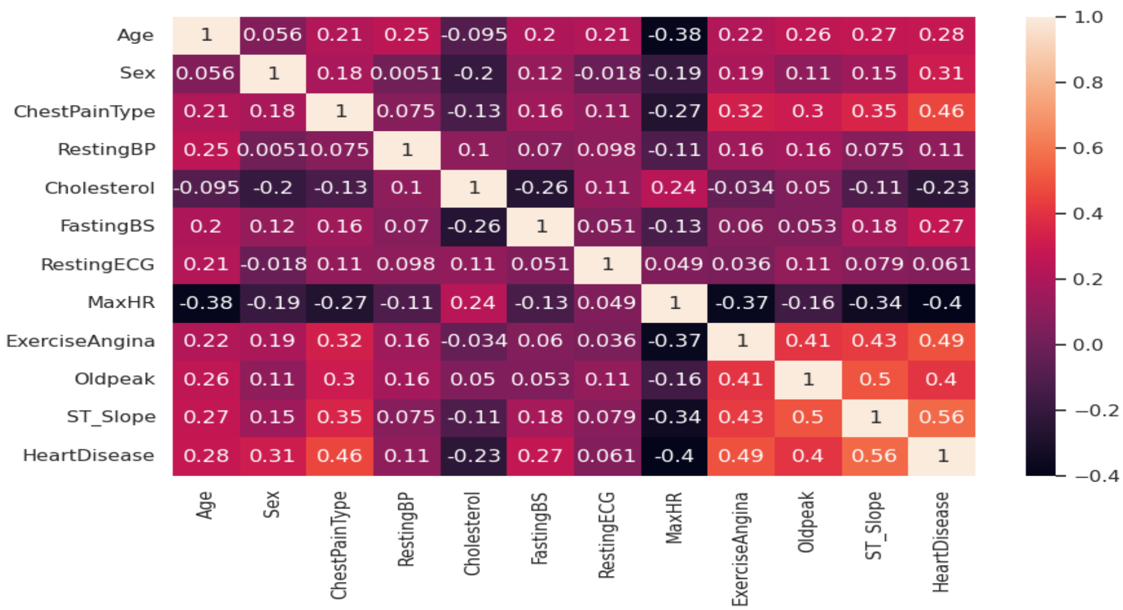


Figure 4.2.8: Correlation Heatmap.

The support vector classifier gets 157 data points accurate and 27 data points erroneous, as shown in Figure 4. Figure 5 displays the (ROC) curve for the predictions. The (AUC) for a support vector machine classifier is 89%.

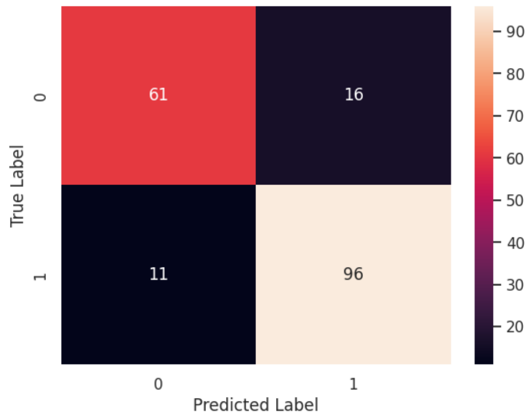


Figure 4: Confusion matrix of SVM classifier.

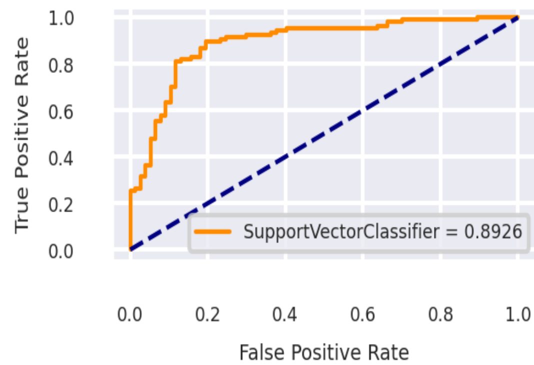


Figure 5: ROC curve for SVM

As seen in Figure 6, the Random Forest classifier gets 157 data points accurate and 27 data points erroneous. Pictured in Figure 7 is the (ROC) curve that represents the forecasts. An AUC of 91% is achieved with a Random Forest classifier.

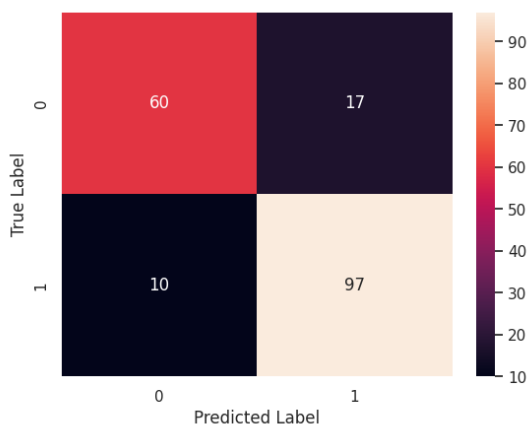


Figure 6: Confusion matrix of Random Forest.

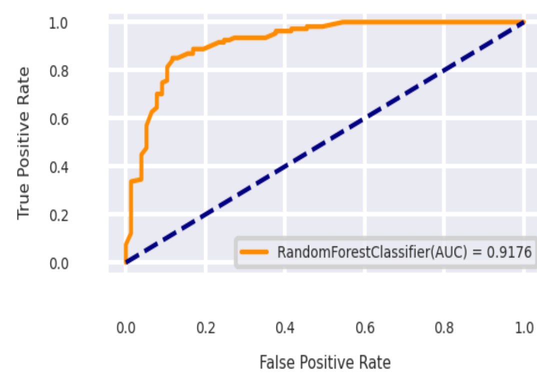
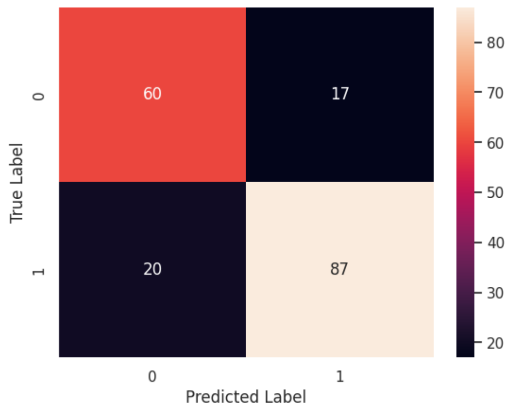


Figure 7: ROC curve RF

The Decision Tree classifier gets 147 predictions accurate and 37 wrong, as shown in Figure 8. Pictured in Figure 9 is the (ROC) curve that represents the forecasts. The (AUC) for a decision tree classifier is 79%.



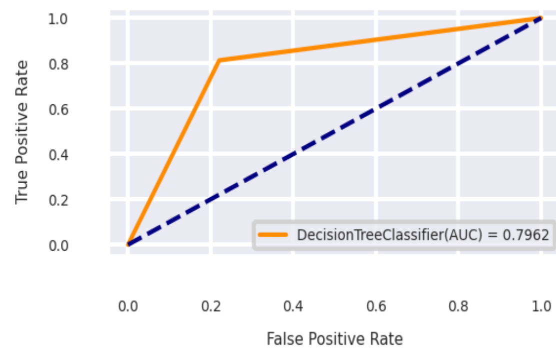
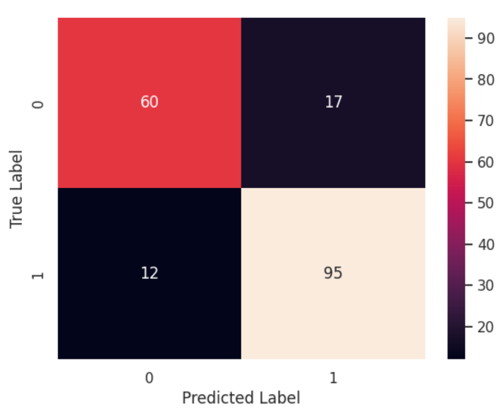


Figure 8: Confusion matrix of Decision Tree Classifier. Figure 9: ROC curve Decision Tree

Figure 10 shows that out of 155 data points, the Extreme Gradient Boosting classifier gets 29 predictions accurate and 29 points wrong. Pictured in Figure 11 is the (ROC) curve that represents the forecasts. The (AUC) for a classifier that uses Extreme Gradient Boosting is 90%.



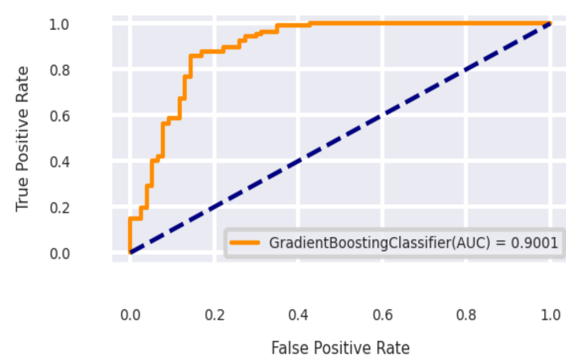


Figure 10: Confusion matrix of Extreme Gradient Boosting Classifier. Figure 11: ROC curve XGB

The K-Nearest Neighbors classifier gets 157 data points accurate and 27 data points incorrect, as shown in Figure 12. The (ROC) curve for the predictions is displayed in Figure 13. The (AUC) for a K-Nearest Neighbors classifier is 89%.

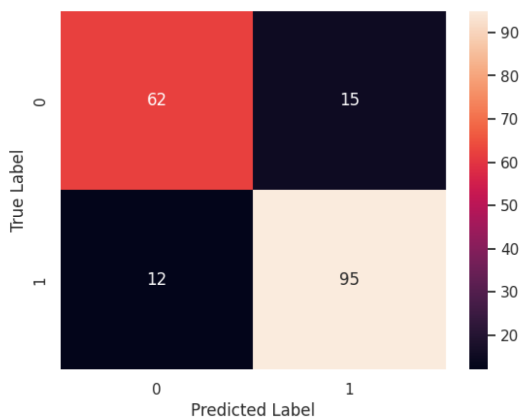


Figure 12: Confusion matrix of K-Nearest Neighbors Classifier.

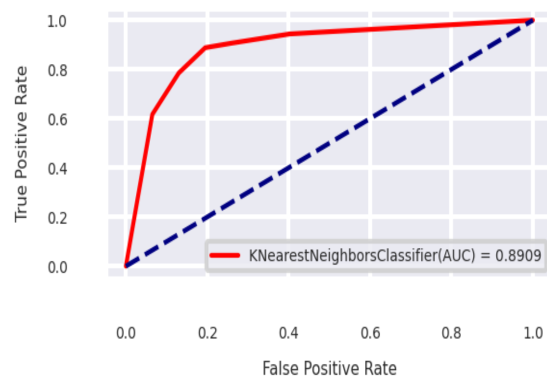


Figure 13: ROC curve K-Nearest Neighbors

The Gaussian Naïve Bayes classifier gets 153 data points accurate and 31 data points erroneous, as shown in Figure 14. The (ROC) curve for the predictions is displayed in Figure 15. Gaussian Naïve Bayes classifier has an AUC of 89%.

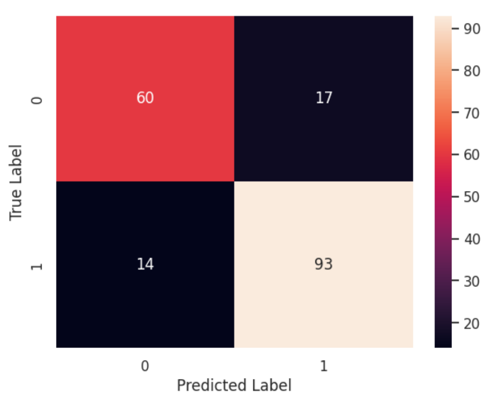


Figure 14: Confusion matrix of Gaussian Naïve Bayes Classifier.

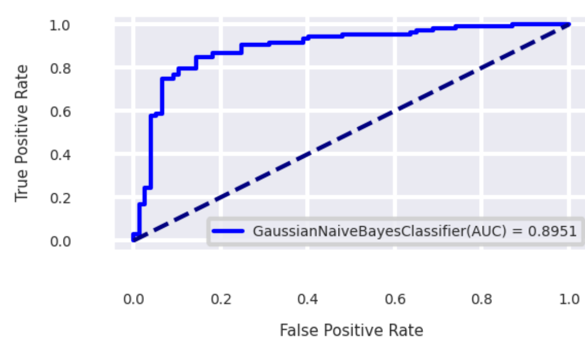
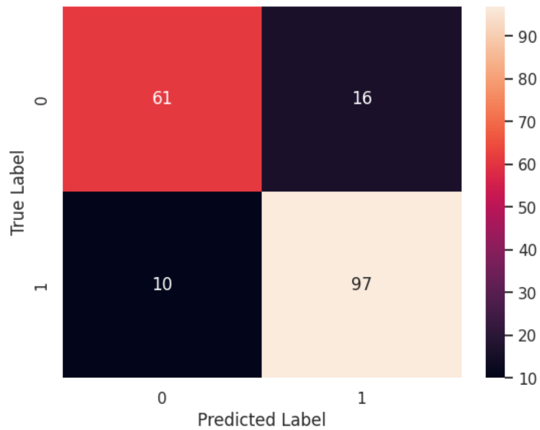


Figure 15: ROC curve Gaussian Naïve Bayes

The Multi-Layer Perceptron classifier correctly predicts 26 out of 158 data points, as shown in Figure 16. Figure 17 shows the predicted (ROC) curve. With a Multi-Layer Perceptron classifier, you may expect an AUC of 90%.



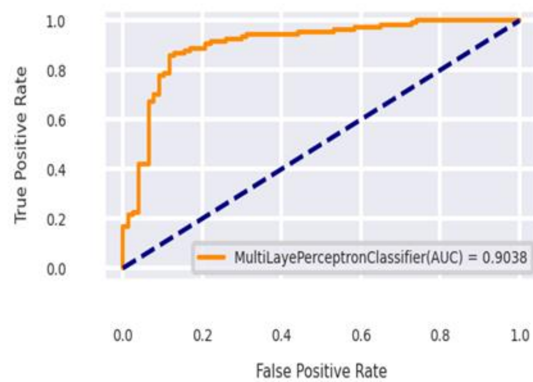
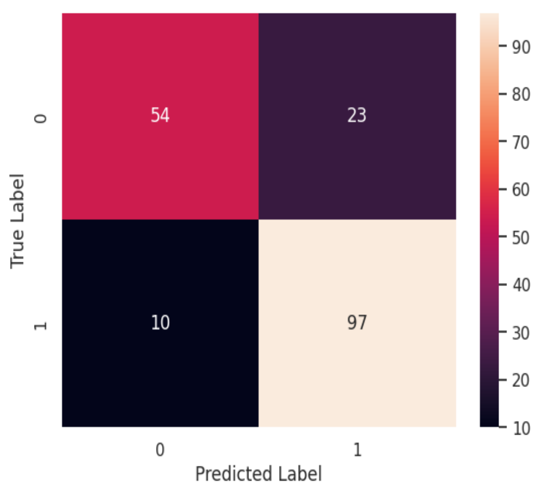


Figure 16: Confusion matrix of Multi-Layer Perceptron Classifier.

Figure 17: ROC curve Multi-Layer Perceptron

The Logistic Regression classifier gets 151 data points right and 33 data points erroneous, as shown in Figure 18. Pictured in Figure 19 is the (ROC) curve that represents the forecasts. An 89% (AUC) is achieved via a Logistic Regression classifier.



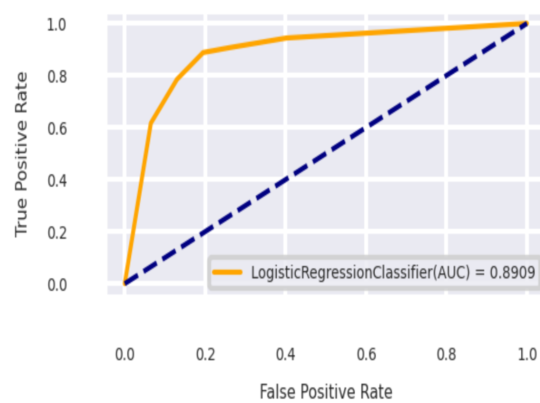


Figure 18: Confusion matrix of Logistic Regression Classifier.

Figure 19: ROC curve Logistic Regression.

Table 4.1.1: Algorithms for Cardiovascular Disease Prediction.

Algorithms	**Accuracy**	**F1 Score**	**Precision**	**Recall**
SVM	85%	87%	85%	85%
RF	85%	88%	85%	90%
DT	79%	82%	83%	81%
XGB	84%	86%	84%	88%
KNN	86%	87%	86%	88%
GNB	83%	85%	84%	86%
MLP	85%	88%	85%	90%
LR	82%	85%	80%	90%

Table 4.2.2: Previous Comparison of results Cardiovascular Disease Prediction.

References	**Algorithms**	**Accuracy**
Herdian, et al. [1]	K-nearest neighbor	85%
Mijwil, et al. [3]	Decision Tree	84%
Louridi, et al. [4]	Support Vector Machine	82%
Dinesh KG, et al. [5]	Logistic Regression	81%
J. Maiga, et al. [18]	Naive Bayes	73%
R. J. P. Princy, et al. [9]	Logistic Regression	80.6%
Mohan S, et al. [6]	Random Forest	83.7%
Proposed Model	**K-nearest neighbor (KNN)**	**86%**

From table 4.1.2 the previous accuracy is 82% in K-nearest neighbor (KNN) (best) [1]. This paper 83% in Decision Tree [2], 84% in SVM [3], 82% in Logistic Regression [4] 82% in Random Forest(best) etc. Maximums are same accuracy But, in my paper, highest performance model is 86% K-Nearest Neighbors.

However, according to table 4.1.1, K-Nearest Neighbor is the top algorithm, boasting an F1 score of 87% and an accuracy of 86%. Among all the algorithms that were analyzed, this one had the greatest F1 score and the highest accuracy.

4.2.1 Selective Feature Based on Correlation Metrix

We saw that ChestPainType(0.46), ExerciseAngina(0.49), Oldpeak(0.4) and ST_Slope(0.56) by analyzing these features and found that the correlation between them is high, then with these features, we got this accuracy by running using the same algorithm.

Table 4.2.3: Performance of Selective Feature Based on Correlation Metrix Algorithms for Cardiovascular Disease Prediction.

Algorithms	**Accuracy**	**F1 Score**	**Precision**	**Recall**
SVM	78%	80%	84%	82%
RF	79%	83%	82%	82%
DT	78%	83%	78%	80%
XGB	77%	81%	79%	80%
KNN	82%	85%	84%	84%
GNB	78%	82%	81%	81%
MLP	79%	80%	85%	83%
LR	77%	76%	88%	81%

Table 4.2.4: Previous Comparison of results Selective Feature Based on Correlation Metrix Algorithms for Cardiovascular Disease Prediction.

References	**Algorithms**	**Attribute**	**Accuracy**
Herdian, et al. [1]	KNN	7	85%
Mijwil, et al. [3]	Decision Tree	14	84%
Louridi, et al. [4]	SVM	16	82%
Dinesh KG, et al. [5]	Logistic Regression	12	81%
J. Maiga, et al. [18]	Naive Bayes	14	73%
R. J. P. Princy, et al. [9]	Neural Networks	12	80.6%
Mohan S, et al. [6]	RF	14	83.7%
This Study	**K-nearest neighbor (KNN)**	**4**	**82%**

This Study found that by analyzing the features ChestPainType, ExerciseAngina, Oldpeak, ST_Slope. This found that the correlation between them is high, then by reducing the features less than them, we got better accuracy using the same algorithm.

4.3 Discussion

Here, the outcomes are shown, and the outputs and accuracies of the classifiers are examined. When their hyperparameters were changed, all the models operated properly; nevertheless, the top model is thought to have the highest accuracy. The K-nearest neighbor (KNN) Classifier tied for the highest accuracy in this investigation with an accuracy of 86%. KNN is a non-parametric method that predicts values by using a dataset's closest neighbors. This allows it to handle complex decision boundaries. Another advantage is the ability to effectively manage multi-class categorization problems. The algorithm is easy to implement because of its simplicity. incredibly adaptable The KNN algorithm requires all data to be kept in memory. Future estimates from the algorithm will automatically include new examples and data points without the need for human interaction. The only other parameters required for training the KNN algorithms was, and the length measure you wish to employ for evaluation, other from a few hyperparameters. The features ST_Slope, Oldpeak, ChestPainType, and ExerciseAngina were investigated in this study. Since no attributes were excluded in the previous analysis, my findings are superior.We improved the algorithm's performance by decreasing the features less than in the previous study. A strong association between the features was found in this investigation.

CHAPTER 5

IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

5.1 Impact on Society

Models have the potential to totally change the way healthcare is provided by enabling the prompt identification of those that might be at risk of acquiring cardiovascular diseases. As a result, patient outcomes are enhanced and the frequency of CVD-related events, such as strokes and heart attacks, is decreased. The decentralized nature of healthcare delivery will enable more people to take part in early detection programs, which will particularly benefit disadvantaged groups and those with limited access to healthcare. This customized approach not only increases treatment effectiveness but also increases patient happiness and involvement in the healthcare process. Furthermore, the use of AI in CVD forecasting offers tremendous cost savings for healthcare systems by reducing the cost of treating advanced stages of illness. Early intervention and preventive care reduce the need for expensive medical facilities and procedures, increasing the effectiveness by which medical facilities are distributed and lessening the financial burden on individuals and society at large. Individual differences in risk factors, genetics, and lifestyle choices can also be addressed by customized therapy plans and treatments developed from machine learning-derived predictive models. Our knowledge of cardiovascular medicine is expanded by research and developments cardiovascular illness, in addition to the direct benefit to patients. To better comprehend future investigations and clinical trials aiming at generating tailored treatments and medications in CVD management and prevention, this work finds novel biomarkers, risk factors, and prediction indicators related with CVDs. All things considered, using computer learning-based forecasting models for the diagnosis of CVD is an essential first step toward increased global welfare, closed healthcare gaps, and improved health outcomes.

5.2 Impact on Environment

The environmental implications for diagnosing CAD may be significant, even though they are primarily indirect. One important aspect that may contribute to increased energy use and carbon emissions is the computational resources required to train and use these models. Energy-intensive computer operations are required due to large datasets and complicated algorithms. Consequently, the environment may be impacted by machines learning-driven CVD detection mechanisms because of higher energy needs and associated greenhouse gas emissions from data centers or computer equipment. However, in addition to their effects on the environment, models based on machine learning for predicting for CVD diagnosis may have benefits for society. Because these models make it easier to identify and treat CVDs early on, they may reduce the overall cost of illness on healthcare systems. By doing this, the environmental effects of treating CVDs at advanced stages may be lessened. Additionally, machine learning algorithms' scalability and accessibility can support remote monitoring and healthcare initiatives, minimizing the need for patients to visit medical facilities and, as a result, reduce carbon emissions associated with transportation. Additionally, as machine learning techniques progress, initiatives are made to improve processing infrastructure and create more energy-efficient algorithms to decrease the adverse environmental impacts of these technologies. Through adding efficiency and sustainability elements into the development and deployment using models based on machine learning for predicting during CVD screening, healthcare stakeholders can maximize societal benefits and minimize environmental impact.

5.3 Ethical Aspects

In our study, Important ethical issues are brought up using machine learning algorithms for predicting to diagnose cardiovascular illnesses (CVDs), and our study must carefully consider these issues. Developing and implementing these models with clarity and simplicity in mind is essential if we hope to foster trust amongst patients and medical providers. It should be possible for the public to obtain comprehensive explanations of the variables influencing the forecasts made by the models along with the procedures followed in their creation. To safeguard confidential patient information and safeguard sensitive medical data, strict compliance with data protection laws and robust security measures are essential. Reducing inefficiencies in the data used to train the machines and the algorithms themselves is necessary to prevent disparities in health care outcomes between different demographic groups. Ethical guidelines should prioritize fairness and equity to ensure that predictive models do not inadvertently exacerbate or sustain pre-existing gaps in access to medical care. In addition, it is imperative to maintain human supervision and accountability when employing predictive models to avoid an over-reliance on methods for making decisions and to ensure that healthcare professionals are free to assess data and make knowledgeable clinical decisions. By addressing these ethical concerns, machine learning-derived algorithmic models for CVD detection can equitable delivery.

5.4 Sustainability

The long-term sustainability illnesses, or CVDs, is contingent upon several interrelated aspects, such as the environment, society, and economics. From an environmental perspective, these models must be made less harmful to the environment and use less resources. The training and implementation of machine learning methods typically need a significant amount of computing power, which can increase energy consumption and release of greenhouse gases. However, the negative effects on the environment can be reduced by utilizing energy-efficient computing hardware and developing better algorithms. Furthermore, the integration of green computing concepts might enhance the sustainability of CVD detection systems developed from machine learning. Two such instances of these ideas are the use of alternate power sources and energy-efficient equipment for data centers. For these tactics to be financially feasible, they need to be both scalable and inexpensive. Even while infrastructure and technology expenditures may initially be substantial, early detection and avoidance of CVDs could save a significant amount of money for healthcare organizations. Based on algorithmic learning, prediction models can reduce the need for expensive therapies and prolonged hospital stays associated with severe illness stages, which benefits both patients and medical staff. Scalability of these models also ensures equitable treatment utilization for CVD diagnostics and maximizes societal value in a wide range of healthcare scenarios. Fair resource distribution, trust, and accessibility are essential elements of machine learning-driven forecasting models for CVD diagnosis in terms of sustainability. Healthcare workers must manage a variety of ethical concerns, such as guaranteeing responsibility, transparency, and patient privacy, in order to build confidence with their patients. Eliminating inefficiencies in training data and algorithms is critical to improving the equity of CVD diagnosis and treatment, hence averting disparities in healthcare outcomes. Programs that increase health literacy and empower individuals to take an active role in their health care may also contribute to the development of more environmentally friendly healthcare systems by promoting preventative measures and reducing reliance on costly treatments. Furthermore, the long-term dependability of machine learning-based forecasted systems for CVD detection may be enhanced by ongoing research and collaboration. These models' viability and usefulness are mostly dependent on continuous attempts to improve algorithm performance, data quality, and model accessibility. Additionally, interdisciplinary cooperation between healthcare professionals, data scientists, legislators, and other community members is crucial to addressing complicated healthcare issues and ensuring a fair distribution of benefits coming from CVD detection programs. Encouragement of an open, honest, and creative culture can optimize sustainability in the long run, this will enhance mental health and health results.

In conclusion, future-proofing machine learning-based forecasting methods for cardiovascular disease detection need a holistic approach that considers social fairness, fiscal viability, and environmental effect. Players in the healthcare sector should include sustainable elements in the development and implementation of these models to create strong healthcare systems that prioritize long-term health results, fair treatment, and environmental responsibility.

CHAPTER 6

CONCLUSION & FUTURE WORK

6.1 Summary of the Study

The research on " Cardiovascular Disease Prediction: A Machine Learning-Based Approach" the discovery of effective therapies for cardiovascular diseases (CVDs), a significant global public health concern, depends on early and accurate detection. I investigate the possibility of developing predictive models, which includes clinical and demographic information from 918 patients, we examine several classification strategies to differentiate between people who have and do not have heart disease. Preprocessing the dataset aids in our initial understanding of feature distributions and correlations. This deals with feature scaling and missing value handling. The analysis of the exploratory data comes next. The study's findings will affect patient satisfaction and healthcare accessibility worldwide respiratory disease diagnosis. All things considered, the findings of this study advance our knowledge of the diagnosis of respiratory disorders and have significant implications for enhancing patient treatment and accessibility globally.

6.2 Conclusion

In this paper, to identify cardiac illness, eight classification techniques are used, one among each from the 12 variables included in the data set. This collection of techniques includes Decision Tree, Gradient Boosting, K-nearest neighbor (KNN), the support vector machine (SVM), Gaussian Nave Bayes, MLP Classifier, and Logistic Regression. Additionally, the accuracy of these algorithms is determined by utilizing Sensitivity and Specificity. The KNN (k-nearest-neighbor) methods can be inferred to have an accuracy rate of up to 86% based on the available data. Furthermore, a front-end system was developed that displayed the probability of heart disease upon data entry based on the 12 attributes. The relevance of the probability ranges is also described, and the findings are derived from the methods used. Further studies can use different ensemble methods to these algorithms, resulting in additional parameter settings and better performance.

6.3 Implication for Further Study

For studies to come, with the information provided here, we hope to develop a prediction system that will improve healthcare while cutting expenses. The usefulness of this work might be enhanced by conducting more investigations into the many open questions. Conditions other than cardiovascular disease, such as diabetes, should be considered if there is a clinical correlation between them. We also plan to employ AI to find a correlation between air pollution and heart disease. Healthcare system security may be enhanced by the integration.

References

[1] Herdian, C., Widianto, S., Ginting, J. A., Geasela, Y. M., & Sutrisno, J. (2024). The Use of Feature Engineering and Hyperparameter Tuning for Machine Learning Accuracy Optimization: A Case Study on Heart Disease Prediction. In Engineering Applications of Artificial Intelligence (pp. 193-218). Cham: Springer Nature Switzerland.

[2] Naser, M. A., Majeed, A. A., Alsabah, M., Al-Shaikhli, T. R., & Kaky, K. M. (2024). A Review of Machine Learning's Role in Cardiovascular Disease Prediction: Recent Advances and Future Challenges. Algorithms, 17(2), 78.

[3] Mijwil, M., Faieq, A. K., & Aljanabi, M. (2024). Early Detection of Cardiovascular Disease Utilizing Machine Learning Techniques: Evaluating the Predictive Capabilities of Seven Algorithms. Iraqi Journal For Computer Science and Mathematics, 5(1), 263-276.

[4] Louridi, N., Amar, M., & El Ouahidi, B. (2019, October). Identification of cardiovascular diseases using machine learning. In 2019 7th mediterranean congress of telecommunications (CMT) (pp. 1-6). IEEE.

[5] Dinesh, K. G., Arumugaraj, K., Santhosh, K. D., & Mareeswari, V. (2018, March). Prediction of cardiovascular disease using machine learning algorithms. In 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT) (pp. 1-7). IEEE.

[6] Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. IEEE access, 7, 81542-81554.

[7] Kusuma, F. D., Kusumaningtyas, E. M., Barakbah, A. R., & Hermawan, A. A. (2018, October). Retracted: Heart Abnormalities Detection Through Iris Based on Mobile. In 2018 International Electronics Symposium on Knowledge Creation and Intelligent Computing (IES-KCIC) (pp. 152-157). IEEE.

[8] Amin, M. S., Chiam, Y. K., & Varathan, K. D. (2019). Identification of significant features and data mining techniques in predicting heart disease. Telematics and Informatics, 36, 82-93.

[9] Princy, R. J. P., Parthasarathy, S., Jose, P. S. H., Lakshminarayanan, A. R., & Jeganathan, S. (2020, May). Prediction of cardiac disease using supervised machine learning algorithms. In 2020 4th international conference on intelligent computing and control systems (ICICCS) (pp. 570-575). IEEE.

[10] Jinjri, W. M., Keikhosrokiani, P., & Abdullah, N. L. (2021, July). Machine learning algorithms for the classification of cardiovascular disease-A comparative study. In 2021 International Conference on Information Technology (ICIT) (pp. 132-138). IEEE.

[11] Bashir, S., Khan, Z. S., Khan, F. H., Anjum, A., & Bashir, K. (2019, January). Improving heart disease prediction using feature selection approaches. In 2019 16th international bhurban conference on applied sciences and technology (IBCAST) (pp. 619-623). IEEE.

[12] Sabab, S. A., Munshi, M. A. R., & Pritom, A. I. (2016, December). Cardiovascular disease prognosis using effective classification and feature selection technique. In 2016 international conference on medical engineering, health informatics and technology (MediTec) (pp. 1-6). IEEE.

[13] Zunaidi, W. H. A. W., Saedudin, R. R., Shah, Z. A., Kasim, S., Seah, C. S., & Abdurohman, M. (2018). Performances analysis of heart disease dataset using different data mining classifications. International Journal on Advanced Science, Engineering and Information Technology, 8(6), 2677-2682.

[14] Khemphila, A., & Boonjing, V. (2011, August). Heart disease classification using neural network and feature selection. In 2011 21st international conference on systems engineering (pp. 406-409). IEEE.

[15] Kavitha, M., Gnaneswar, G., Dinesh, R., Sai, Y. R., & Suraj, R. S. (2021, January). Heart disease prediction using hybrid machine learning model. In 2021 6th international conference on inventive computation technologies (ICICT) (pp. 1329-1333). IEEE.

[16] Thomas, J., & Princy, R. T. (2016, March). Human heart disease prediction system using data mining techniques. In 2016 international conference on circuit, power and computing technologies (ICCPCT) (pp. 1-5). IEEE.

[17] Taneja, A. (2013). Heart disease prediction system using data mining techniques. Oriental Journal of Computer science and technology, 6(4), 457-466.

[18] Maiga, J., & Hungilo, G. G. (2019, October). Comparison of machine learning models in prediction of cardiovascular disease using health record data. In 2019 international conference on informatics, multimedia, cyber and information system (ICIMCIS) (pp. 45-48). IEEE.

[19] Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. IEEE access, 7, 81542-81554.

[20] Haq, A. U., Li, J. P., Memon, M. H., Nazir, S., & Sun, R. (2018). A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. Mobile information systems, 2018(1), 3860146.

[21] Arabasadi, Z., Alizadehsani, R., Roshanzamir, M., Moosaei, H., & Yarifard, A. A. (2017). Computer aided decision making for heart disease detection using hybrid neural network-Genetic algorithm. Computer methods and programs in biomedicine, 141, 19-26.

[22] Latha, C. B. C., & Jeeva, S. C. (2019). Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. Informatics in Medicine Unlocked, 16, 100203.

Appendix A

Course Outcomes, Complex Engineering Problems (EP) and Complex Engineering Activities (EA) Addressing

Title:

CARDIOVASCULAR DISEASE PREDICTION: A MACHINE LEARNING-BASED APPROACH

Student ID: [201-15-14151]

CO Description for FYDP

CO	CO Descriptions	PO
Phase -I		
CO1	Integrate recently gained and previously acquired knowledge to identify a c**ardiovascular Disease Prediction** problem for the Final Year Design Project (FYDP)	**PO1**
CO2	Analyze different aspects of the goals in designing a solution for this FYDP	**PO2**
CO3	Explore diverse problem domains through a literature review, delineate the issues, and establish these goals for the FYDP	**PO4**
CO4	Perform economic evaluation and cost estimation and employ suitable project management procedures throughout the development life cycle of the FYDP	**PO11**
Phase -II		
CO5	Design and develop technical solutions and system components or processes that meet specified requirements, ensuring compliance with public health and safety standards, as well as considering cultural, socioeconomic, and environmental factors in this FYDP	**PO3**
CO6	Choose and apply appropriate methodologies, resources, and contemporary engineering and IT technologies to address complex engineering processes, encompassing prediction and modeling, while adhering to relevant constraints in this FYDP	**PO5**
CO7	Analyze societal, health, safety, legal, and cultural considerations, along with associated responsibilities, in the context of professional engineering practice and the resolution of this problem, employing logical reasoning guided by contextual understanding.	**PO6**

CO8	Comprehend and evaluate the enduring sustainability and impact of professional engineering endeavors in addressing intricate engineering challenges within social and environmental frameworks.	**PO7**
CO9	Implement ethical principles and adhere to professional standards and norms in this FYDP	**PO8**
CO10	Capable of operating proficiently both individually and as a team member or leader across diverse teams and interdisciplinary settings in this FYDP.	**PO9**
CO11	Proficiently communicate with the engineering community and broader society regarding complex engineering endeavors, including the ability to comprehend and generate comprehensive reports and design documentation, as well as provide and receive clear instructions throughout this FYDP.	**PO10**
CO12	Acknowledge the importance of self-directed and life-long learning within the evolving landscape of technology and possess the readiness and capability to engage in lifelong learning endeavors.	**PO12**

Addressing CO (1 to 8), Knowledge Profile (K), Attainment of Complex Engineering Problems (EP), and Attainment of Complex Engineering Activities (EA)

Addressing CO (1 to 8), Knowledge Profile (K), Attainment of Complex Engineering Problems (EP):

SN	EP Definition	Attainment	CO	Justification (with Knowledge Profile)	References
1.	**EP1: Depth of Knowledge required**	Yes	CO1, CO2, CO3, CO5, CO6, CO7 and CO8	This project demonstrates fundamental engineering (K3) principles by employing machine learning and classification tasks. The project shows specialized knowledge (K4) through computer-assisted diagnosis. This relies on better cardiovascular disease prediction across several attributes. This is made possible by transfer learning with advanced K-nearest neighbor, decision trees, support vector machines, logistic regression, naive bayes, and random forests.	**Page no:** [16] **Section:** [3.2] **Page no:** [1-2] **Section:** [1.1]

				The project applies Practice engineering and design (K5), using the testing procedure depicted in the illustration. The study uses a machine learning model to solve engineering and technology procedures (K6).	**Page no:** [16] **Section:** [3.2(B)] **Page no:** [17] **Section:** [3.4]
				This project ensures to **K8 (Research Literature)** to enhance machine learning-based cardiovascular disease prediction by displaying an extensive understanding of current methods and integrating results from prior research.	**Page no:** [9-13] **Section:** [2.2, 2.3]
2.	**EP2: Range of Conflicting Requirements**	Yes	CO2, and CO7	This project addresses **EP-2** given the inadequacy of conventional methods for cardiovascular detection, the challenges of combining learning through transfer using machine learning are taken into consideration. It addresses problems in comprehending geographic distributions through comparisons and proposes improved diagnostic approaches.	**Page no:** [24,14] **Section:** [2.4, 2.5]

3.	**EP3: Depth of analysis required**	Yes	CO2, and CO6	This project addresses EP-3 by with the objective of enhancing cardiovascular disease prediction, we reviewed numerous studies' findings and found that machine learning was the most effective.	**Page no:** [24-35] **Section:** [4.2]
4.	**EP4: Familiarity of Issues**	Yes	CO8	This endeavor's multidisciplinary strategy extends beyond the realms of computer science and engineering to impact the diagnosis and treatment of respiratory disorders, including cardiovascular disease prediction. Innovations in healthcare and everyday health practices are aided by it, suggesting.**EP-4**.	**Page no:** [9-14] **Section:** [2.2, 2.4]
5.	**EP5: Extends of application codes**	No	CO5	N/A	N/A
6.	**EP6: Extends of stakeholders involved and**	No	CO8	N/A	N/A

7.	**EP7: Interdependence**	Yes	CO5	This project's comprehensive approach consists of multiple elements including data gathering, statistical analysis, and suggested methodology to guarantee comprehensive resolutions to difficult medical diagnosis scenarios **EP-7**.	**Page no:** [16-19] **Section:** [3.2, 3.3, 3.4]

Addressing CO11 with Complex Engineering Activities (EA) [Some or all of the following]:

SN	EA Definition	Attainment	CO	Justification	References
1.	**EA1: Range of resources**	Yes	CO11	This study makes use of a variety of resources, such as GPUs, annotated datasets, sophisticated computing infrastructure, and ethical considerations, to guarantee methodical research and contributes to the field of machine learning-based cardiovascular disease prediction.	**Page no:** [20-23] **Section:** [3.5]
2.	**EA2: Level of interaction**	No		N/A	N/A

3.	**EA3: Innovation**	No		N/A	N/A
4.	**EA4: Consequences for society and the environment**	Yes		This project enhances the ecology through the effective use of computing resources, upholds ethical standards for patient privacy, and enhances healthcare through the creation of improved detection systems—all of which have a positive impact on society.	**Page no:** [37-40] **Section:** [5.1, 5.2, 5.4]
5.	**EA-5: Familiarity**	Yes		This project combines basic terminology and a comprehensive comparative analysis with machine learning to present a novel approach for the prediction of heart disease. This approach builds on earlier findings and offers new insights into the subject.	**Page no:** [9-12] **Section:** [2.1, 2.3]

Addressing CO (4, 9, 10, and 12):

SN	COs	Attainment	Justification	References
1	CO4	Yes	This project addresses **CO4** by combining effective project management with financial supervision to guarantee careful planning, fair resource allocation, and estimated costs for maximum resource use during the research project.	**Page no:** [7] **Section:** [1.6]
2	CO9	Yes	The project supports responsible information sharing and a healthier society by implementing cutting-edge healthcare technologies in an ethical manner that follows all regulations,	**Page no:** [38-39] **Section:** [5.3]

			puts patient privacy first, gets their informed consent, and records all research in a transparent manner **CO9**.	
3	CO10	No	N/A	N/A
4	CO12	Yes	Extensive data collection, rigorous statistical evaluation, method development, and exact experimental results and analysis exhibit an approach to staying current and examining solutions for dealing with contemporary challenges. All these characteristics reflect the study's commitment to CO12, or lifelong learning, and the dynamic character of technology.	**Page no:** [16-23, 24-32] **Section:** [3.2, 3.3, 3.4, 3.5, 4.2]

Plagiarism check:

Final Defense for Cardiovascular Disease Machine Learning.pdf

ORIGINALITY REPORT

23%	13%	4%	17%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	10%
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	7%
3	Submitted to Grambling State University Student Paper	<1%
4	"Smart Trends in Computing and Communications", Springer Science and Business Media LLC, 2023 Publication	<1%
5	dokumen.pub Internet Source	<1%
6	Submitted to University of East London Student Paper	<1%
7	link.springer.com Internet Source	<1%
8	Submitted to University of Westminster Student Paper	<1%