Predict Cardiovascular Disease Using Machine Learning Techniques

BY

Esrat Jahan
ID: 212-15-4098

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

Supervised By

Dr. Md. Zahid Hasan
Associate Professor
Department of CSE
Daffodil International University

Co-Supervised By

Saiful Islam
Assistant Professor
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

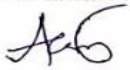
JANUARY 2024

APPROVAL

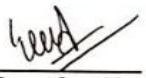
This Project titled "Predict Cardiovascular Disease Using Machine Learning Techniques", submitted by Esrat Jahan to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 16.07.2024.

BOARD OF EXAMINERS

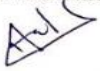
Dr. S.M Aminul Haque — **Board Chairman**
Professor & Associate Head
Department of CSE
Faculty of Science & Information Technology
Daffodil International University



Dr. Arif Mahmud — **Internal Examiner 1**
Associate Professor, Internal Member
Department of CSE
Faculty of Science & Information Technology
Daffodil International University



Md. Sazzadur Ahmed — **Internal Examiner 2**
Assistant Professor, Internal Member
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

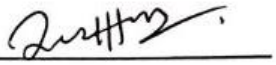


Dr. Md. Arshad Ali — **External Examiner**
Professor, External Member
Department of Computer Science & Engineering
Hajee Mohammad Danesh Science and Technology University

DECLARATION

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

Supervised by:

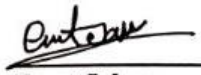


Dr. Md. Zahid Hasan
Associate Professor
Department of CSE
Daffodil International University

Co-Supervised by:

Saiful Islam
Assistant Professor
Department of CSE
Daffodil International University

Submitted by:



Esrat Jahan
ID: 212-15-4098
Department of CSE
Daffodil International University

©Daffodil International University

iii

ACKNOWLEDGEMENT

We extend our sincere gratitude to the Almighty for His blessings, enabling us to successfully complete our final year project.

Our heartfelt thanks go to Dr. Md. Zahid Hasan, Associate Professor in the Department of CSE at Daffodil International University, Dhaka. We appreciate our supervisor's deep knowledge and keen interest in the field of "Machine Learning," which guided us through the project. His unwavering patience, scholarly guidance, continual encouragement, energetic supervision, constructive criticism, valuable advice, and diligent review of multiple drafts at various stages have been instrumental in the completion of this project.

We express our gratitude to the Head of the Department of CSE, for his generous assistance in bringing our project to fruition. Our thanks also go to the other faculty members and staff of the CSE department at Daffodil International University.

Acknowledgment is due to our fellow course mates at Daffodil International University who participated in discussions during the course of our work.

Lastly, we would like to recognize and appreciate the constant support and patience of our parents.

ABSTRACT

The declaration emphasizes the increasing prevalence of physical illnesses, especially cardiovascular disease, and the demand for efficient methods to forecast the illness and increase early detection. The utilization of effective algorithm models is recommended by this study in order to predict the risks related to cardiovascular disease. The three datasets used in the study were obtained from the Kaggle website and the UCI machine learning library. Several algorithm models were used in this work, including Gradient Boosting (GB), K-Neighbors Classifier (KNN), XGB Classifier (XGB), Random Forest (RF), Logistic Regression (LR), and Decision Tree (DT). To assess the performances, ensemble models including Bagging, Boosting, Random Subspace, Stacking, and Voting were also used. For the Hungarian dataset, the conventional model KNN earned the highest accuracy with 80.26%, the LR model achieved the greatest accuracy with 88.52%, and the DT model reached the best accuracy with 72.99%. For these three datasets, we also used five distinct kinds of ensemble methods. The findings showed that for the Hungarian dataset, KNN had the greatest accuracy at 81.26%, for the Cleveland dataset, Bagged GB had the best accuracy at 91.8%, and for the Cardio dataset, GB had the best accuracy at 73.11%. By applying hyperparameter tuning, the best parameters were assigned to each classifier, resulting in more precise cardiovascular disease predictions. Comparing the experimental examination to earlier research, performance was better, with the ensemble model obtaining the greatest accuracy of 91.8%. The study emphasizes how crucial it is to use ensemble methods and sophisticated algorithm models in order to forecast cardiovascular illness properly and enhance real-world results.

Keywords: Cardiovascular *Disease, Correlation, Prediction, Machine Learning, Algorithms.*

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Introduction

The prevalence of cardiovascular disease, a term used to describe a number of heart disorders that have an adverse effect on our daily activities, is rising. Finding the sources of harm during diagnosis is the difficult part. Using datasets, machine learning presents a viable method for predicting the existence of cardiovascular illness. Machine learning algorithms can successfully discover patterns and signs connected with the condition by examining a variety of diagnostic information. These algorithms may create prediction models that take into account variables like genetics, lifestyle decisions, medical history, and biomarkers by learning from past data. Machine learning-enabled early detection can help medical practitioners act quickly, provide the right therapies, and give patients the advice they need. Machine learning offers the power to transform healthcare and enhance patient outcomes by increasing the precision and efficacy of illness diagnosis. A team of researchers has worked together to develop algorithms for machine learning that are meant to identify various human health conditions [1]. Their techniques and precision, however, were inadequate for the diagnosis of cardiovascular disease. We suggest a different strategy to improve the human body's prognostic capacity for sickness detection. Supervised and unsupervised machine learning techniques are the two primary categories. In supervised learning, input-output pairings, or labeled data, are used by models to create outputs from inputs. Models are constructed using the labeled data as the training dataset. Unsupervised learning, on the other hand, builds unlabeled data based on patterns. The incidence of cardiovascular disease is rising significantly and affects a large proportion of the population. This illness is a result of a number of lifestyle factors, including nutrition and activity. According to a Foundation survey from 19 years ago, 190 million people had cardiovascular disease or cardiovascular disease [2]. Despite this, we have persisted in our study and have created the most accurate approach for predicting the existence of cardiovascular disease in both suspect and active cases. The predicted result of machine learning-based cardiovascular disease prediction is a model that can successfully

categorize people into various risk groups, such as low, medium, or high risk, depending on the features they provide. Healthcare providers can identify people who can benefit from treatments and preventative steps to lower their risk of cardiovascular disease by using this important model [3].

1.2 Motivation

The motivation behind conducting the study stems from the urgent need to address the escalating burden of cardiovascular diseases (CVD) worldwide. With CVD being a leading cause of morbidity and mortality globally, there's a critical imperative to enhance our predictive capabilities for early detection and intervention. By harnessing the power of machine learning and integrating multiple datasets encompassing diverse demographic, clinical, and lifestyle factors, the study seeks to develop robust predictive models capable of accurately identifying individuals at high risk of developing CVD [4]. Such predictive analytics hold the promise of enabling personalized preventive strategies and targeted interventions, ultimately contributing to improved patient outcomes and reduced healthcare costs. Moreover, by exploring the comparative performance of various machine learning algorithms and incorporating emerging data sources, the research aims to advance the field of predictive analytics in cardiovascular medicine, paving the way for more effective risk stratification and population-level interventions to combat CVD [5]. Other risk factors that have been linked to cardiovascular disease include high birth weights, alcohol use, and hypertension. Many of the studies that have been done to predict heart illness are not as accurate as would be ideal.

1.2 The rationale of the study

Its goal is to create a method for predicting cardiovascular disease in people. We've noticed a worrying increase in this illness's recent effects on our community. But it soon became clear that diagnostic tools and information are noticeably lacking. Cardiovascular illness is an expensive operation according to patient records in our growing country, which emphasizes the need for more affordable and easily accessible diagnostic options. Using several datasets, the project aims to construct and assess machine learning models that can reliably predict the likelihood of cardiovascular disease (CVD). Cardiovascular diseases,

including heart disease and stroke, are leading causes of mortality globally, emphasizing the critical need for accurate risk prediction and prevention strategies. This work intends to leverage the power of sophisticated computer algorithms to examine a variety of datasets comprising different socioeconomic, clinical, lifestyle, and biological factors linked with CVD risk by utilizing machine learning approaches. The primary goal is to construct predictive models that can effectively identify individuals at high risk of developing CVD, enabling early intervention and personalized preventive measures.

1.4 Research Objective

1. To examine how dietary patterns, particularly unhealthy diets, contribute to the development and progression of cardiovascular diseases.

2. To assess the likelihood and accuracy with which individuals with cardiovascular disease can be identified using the proposed model.

3.To determine the model's effectiveness in predicting early signs and risk factors associated with cardiovascular diseases.

4. To analyze the primary advantages of the proposed model in terms of accuracy, efficiency, and usability in predicting cardiovascular illnesses.

5. To investigate potential practical applications and implications of the study's findings in real-world healthcare settings.

6. To identify and outline the necessary safety measures and protocols to ensure the ethical and safe implementation of the model.

1.5 Project Management and Finance

It is not only reasonably priced but also useful for daily tasks. The assessment of cardiovascular disease is very important to our nation. To use the forecasting process in real-world situations, common tools are required. Although using highly customized tools yields best outcomes and smooth operation, our technique works well and can be used to less sophisticated instruments as well.

1.6 Report Layout

Chapter 2 explores the relevant research that has already been done by other researchers. Prior to starting our investigation, it is imperative that we carefully review the introduction and motivation. As a result, we expound on the introduction, offering a comprehensive explication of the recommended methodology, and the motivation section, clarifying the reasoning for our prediction. Following the conclusion of the Introduction phase, we turned our attention to pertinent research and gathered internal data for our project. After evaluating the machine learning algorithms, we selected for the methodology portion using our dataset, we concluded which one performed the best. After pre-processing, data testing took place, which finally produced the comparative result—which is the one we were after. We go into great detail about this result in our last part, which is called the conclusion.

CHAPTER 2

BACKGROUND

2.1. Preliminaries

Cardiovascular illness is accurately diagnosed by the application of machine learning algorithms. We examine the application and analysis of patient diagnostic reports in this section. A number of models are included, including RF, LR, DT, KNN, XGB, and GB, and they use algorithms to conduct the investigation. In this part, deep learning methods are applied to better enhance the research. Several of us utilized different models in our research; they are detailed in the corresponding section.

2.2. Related works

The impact of cardiovascular diseases (CVDs) on the world's health is substantial, and scientists are increasingly using machine learning approaches to create precise prediction models. An overview of relevant research on machine learning models for CVD prediction is given in this part, with emphasis on important papers, approaches, and conclusions. A key work by Dey et al. (2019) concentrated on applying the random forest method to predict coronary artery disease [2]. The dataset that the researchers used included patient demographic, laboratory, and clinical data. Their model demonstrated the promise of machine learning in CVD prediction with its excellent accuracy, sensitivity, and specificity.

A convolutional neural network (CNN) was investigated by Attia et al. (2019) in another noteworthy study to identify cardiovascular risk variables from electrocardiogram (ECG) data. CNN models showed the promise of deep learning techniques for CVD prediction by effectively identifying patterns in the ECG signals linked to diabetes, hypertension, and smoking status [3].

Using a sizable dataset from electronic health records, researchers created a machine learning model in a work by Krittanawong et al. (2020) [4]. The model used a variety of

clinical characteristics, such as comorbidities, laboratory results, and vital signs, to forecast the likelihood of significant adverse cardiovascular events. The model's excellent accuracy highlights the promise of thorough data integration for CVD prediction.

Techniques for feature selection and dimensionality reduction have also been studied in literature. To find the most useful characteristics for CVD prediction, Khelil et al. (2021) used a genetic algorithm-based feature selection technique [5]. The model obtained better computing efficiency and performance by narrowing the feature space.

For CVD prediction, unsupervised learning methods have been used in addition to conventional supervised learning strategies. For example, clustering algorithms were used in a research by Kotecha et al. (2021) to separate out different heart failure patient profiles from electronic health information [6]. Different clinical features and results were displayed by the discovered phenotypes, underscoring the potential of unsupervised learning in customized treatment approaches.

Additionally, integrating genomic and genetic data has demonstrated potential in enhancing CVD prediction models. Researchers integrated genetic risk scores from genome-wide association studies into a machine learning model in a work by Torkamani et al. (2020) [7]. The model showed improved prediction accuracy by the integration of genetic data with conventional risk variables. The body of research indicates that machine learning algorithms have the capacity to forecast cardiovascular illnesses. These models use a variety of methods, including random forests, neural networks, and clustering approaches, and they make use of a wide range of datasets, including clinical, demographic, genetic, and physiological characteristics. The results emphasize the value of feature selection, thorough data integration, and the possibility of using genetic data to improve prediction accuracy. Notwithstanding the noteworthy advancements, several obstacles persist, such as the requirement for extensive and varied datasets, the comprehensibility of the models, and their validation in actual clinical environments. In order to streamline the integration of machine learning models into clinical settings and enhance the early identification, risk assessment, and individualized treatment of cardiovascular disorders, future research should concentrate on resolving these issues. The state and functionality of

the heart are critical to the timely and correct identification of a variety of cardiac diseases. Electrocardiographs, however, can occasionally yield inaccurate readings, which can result in misdiagnosis, needless invasive operations, or overtreatment. Machine learning approaches have been used to solve this problem. An enormous quantity of historical data has been gathered and used by researchers to create automated systems for the diagnosis of heart disease [8]. The research that investigate the use of machine learning algorithms for heart disease prediction are included in this area.

In order to identify abnormalities in ECG records, Tithi et al. (2018) created a model utilizing supervised machine learning methods, such as DT, LR, and NB. Six machine learning algorithms were used by the researchers to determine if an ECG reading was normal or abnormal and to forecast the possibility that a patient had a certain illness. To avoid abnormalities or repeats, the UCI Machine Learning Repository database was split into training and testing sets. Random Train-Test Split and Cross Validation were two of the strategies utilized. The results showed that the Decision Tree worked best for Myocardial Infarction with an accuracy of 96%, while Logistic Regression had the maximum accuracy of 96% for Right Bundle Branch Block. All algorithms except for Nearest Neighbor showed a 95% accuracy rate for sinus tachycardia. Lastly, for Coronary Artery Disease, Naive Bayes achieved the maximum accuracy of 94% [9].

Regression and classification techniques were applied in data mining by Kavitha et al. (2021), using the Cleveland heart disease dataset. They developed three machine learning models: Decision Tree, Random Forest, and a hybrid model that integrated the two. With an accuracy score of 88.7% for the heart disease prediction model displayed in the testing results, the hybrid model surpassed the other methods [10].

A comparison assessment of popular machine learning algorithms for heart disease prevalence prediction was carried out by Khanna et al. (2015). They tested several categorization strategies using the Cleveland Dataset, which is accessible to the public. The study found that less complicated models outperformed their more complex counterparts in terms of accuracy. Examples of these models are Support Vector Machines with linear kernels and Logistic Regression. Accuracy values of 86.8%, 87.6%, and 89% were attained

by the Logistic Regression, Support Vector Machine, and Generalized Regression Neural Network, respectively [11].

Chang et al.'s (2022) main goal was to employ machine learning methods to build an AI-based system for detecting cardiac disease. They created a dependable Python-based healthcare research application that makes it easier to track and set up several health monitoring apps. Data processing for the program included converting categorical columns and working with categorical variables. The implementation of logistic regression, database collecting, and dataset attribute assessment were the three primary stages of program development. The random forest classifier method was created to more accurately diagnose cardiac conditions. The program achieved an accuracy rate of about 83% over the training data, highlighting the importance of data analysis [12].

A model for predicting cardiac disease was reported by Ansari et al. (2021) utilizing data from the UCI Machine Learning Repository. Because it contained information like blood pressure and pulse rate, the model was more accurate than earlier versions. The variables that shown a high degree of predictive potential were the only ones that the researchers included in their logistic regression model after training it with all the variables. They also employed a Support Vector Machine and a hybrid model that blended logistic regression and principal component analysis. Based on the data, the models with the greatest performance were the logistic regression model with PCA and the logistic regression model with all variables, with 86% accuracy, 68% recall, 69% specificity, 77% precision, and 72% F1-score [13].

In our study, we employed several machine learning classifiers to categorize cardiac illnesses, and these classifiers were found to be suitable for the task at hand. Specifically, decision tree-based algorithms, known as "tree structures," were utilized to run the decision models [14]. In the work conducted by researcher R. Williams et al., various machine learning models including KNN (67.21% accuracy), NB (85.25% accuracy), DT (81.97% accuracy), SVM (81.97% accuracy), and LR (85.25% accuracy) were proposed [15]. Similarly, researcher Nagaelli et al. worked with NB (86% accuracy) and SVM (89.4% accuracy) [16]. Researcher Chanchalani et al., a machine learning model incorporating

SVC (83.70% accuracy), NB (84.50% accuracy), LR (82.35% accuracy), and ABC (88.23% accuracy) was proposed [17]. Additionally, Basha et al. proposed a machine learning model with KNN (85% accuracy), DT (82% accuracy), SVM (82% accuracy), RF (81% accuracy), and NB (80% accuracy) [18]. These models provide valuable insights into the accuracy levels achieved by different classifiers for cardiac illness categorization.

These investigations show how well machine learning systems can detect cardiac problems. The models evaluate datasets and produce precise predictions by utilizing a variety of algorithms and strategies, including supervised learning, regression, and classification. The results highlight how machine learning might enhance the identification and diagnosis of cardiac disease [19].

2.3. Comparative Analysis and Summary

The analysis showed in Table 2.1.

Table 2.1: Comparative analysis

Authors	Methodology	Dataset	Results
R. Williams et al.	Machine Learning Algorithms	Accomplish Cleveland Heart disease dataset (303 instances and 14 attributes) and statlog heart dataset [266 instances and 14 attributes] by collecting it	KNN (67.21% accuracy), NB (85.25% accuracy), DT (81.97% accuracy), SVM (81.97% accuracy), and LR (85.25% accuracy)
Nagaelli et al.	Machine Learning Algorithms	The presented approach is evaluated using 38 actual data recordings of electrocardiogram (ECG) signals featuring high-frequency components obtained from the PhysioNet database.	NB (86% accuracy) and SVM (89.4% accuracy)
Chanchalani et al.	Machine Learning Algorithms	The dataset, available on Kaggle since 1988, amalgamates data from Long Beach V, Switzerland, Hungary, and Cleveland. It comprises 76 columns, including 14 selected attributes.	SVC (83.70% accuracy), NB (84.50% accuracy), LR (82.35% accuracy), and ABC (88.23% accuracy)
Basha et al.	Machine Learning Algorithms	The dataset, available on Kaggle.	KNN (85% accuracy), DT (82% accuracy), SVM (82% accuracy), RF (81% accuracy), and NB (80% accuracy)
Our proposed method	Machine Learning Algorithms and Ensemble classifiers	The dataset, available on Kaggle.	Bagged GB had the best accuracy at 91.8%, for the Cardio dataset.

2.4. Scope of the Problem

The task at hand included rationalizing and expediting the process of diagnosing cardiovascular illness. Our goal was to get the highest accuracy possible with our recommended model, taking into account the many machine learning research that are associated with it. Our objective was to implement the concept by using basic technologies to simplify the diagnosis of coronary artery disease and so minimize complexity, even if there was not much room for improvement in the present approach.

2.5 Challenges

Three well-known heart disease datasets were combined to create this curated dataset. The first dataset, which was gathered via Kaggle, is the largest heart disease dataset currently available for research purposes. It has 70000 records with 11 distinct variables. These were obtained at the time of the patient's informational interview and medical checkup. Thirteen similar traits are shared by 303 and 293 cases, respectively, in the second and third datasets. Its curation is based on three datasets: Heart-related Data (Kaggle Dataset), Cleveland (Repository for UCI Machine Learning), (UCI Machine Learning Repository) in Hungarian. It was essential to manually check the dataset after data collection was finished to look for any missing information. With this dataset, our accuracy stands out as unmatched compared to others, indicating a degree of precision never previously attained.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Research Subject and Instrument

To maximize the accuracy of the dataset, we applied many hybrid models and algorithms. This process needed a number of tools, including efficient setup tools that used the greatest GPUs on the market. Python was the primary programming language utilized, and Jupyter Notebook, Google Colab, and Anaconda were among the tools used. These browser-based platforms made it possible for Python applications to be developed and executed without any issues.

3.2 Data Collection Procedure

Three well-known heart disease datasets were combined to create this curated dataset. The first dataset, which was gathered via Kaggle, is the largest heart disease dataset currently available for research purposes. It has 70000 records with 11 distinct variables. These were obtained at the time of the patient's informational interview and medical checkup. Thirteen similar traits are shared by 303 and 293 cases, respectively, in the second and third datasets [21]. The three datasets used for its curation are:

- Cardio Data (Kaggle Dataset)
- Cleveland (UCI Machine Learning Repository)
- Hungarian (UCI Machine Learning Repository)

Every characteristic included in the dataset was thought to be significant in the prediction of cardiovascular disease. Two conditions (0 and 1) were used to classify the patients in the dataset. When the value is 1, it means that the patient has cardiovascular disease; when the value is 0, it means that the patient does not have cardiovascular disease. We computed the dataset's distribution under these two criteria. To facilitate additional research and the development of prediction models, this data offers a summary of the dataset and the patient distribution according to the status of cardiovascular disease. Figure 3.1 provides a graphic

representation of the ratio for Hungarian dataset, Figure 3.2 shows the ratio for Cardio dataset and Figure 3.3 shows the ratio for Cleveland Dataset. The training set and the test set are the two separate portions of the dataset. We decided on an 80-20 split for this partitioning, allocating 80% of the data to the training set and keeping the remaining 20% for the test set.

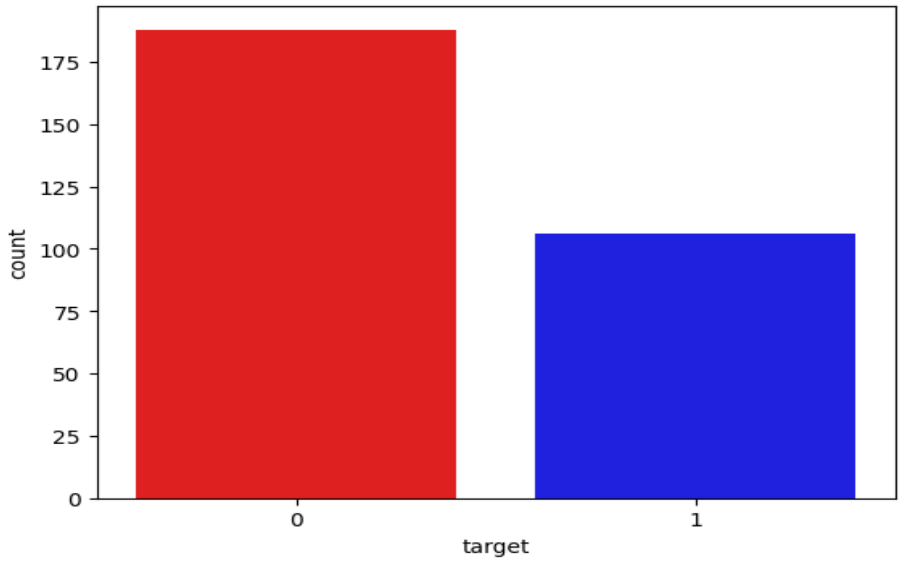


Figure 3.1: Count plot of Hungarian Dataset

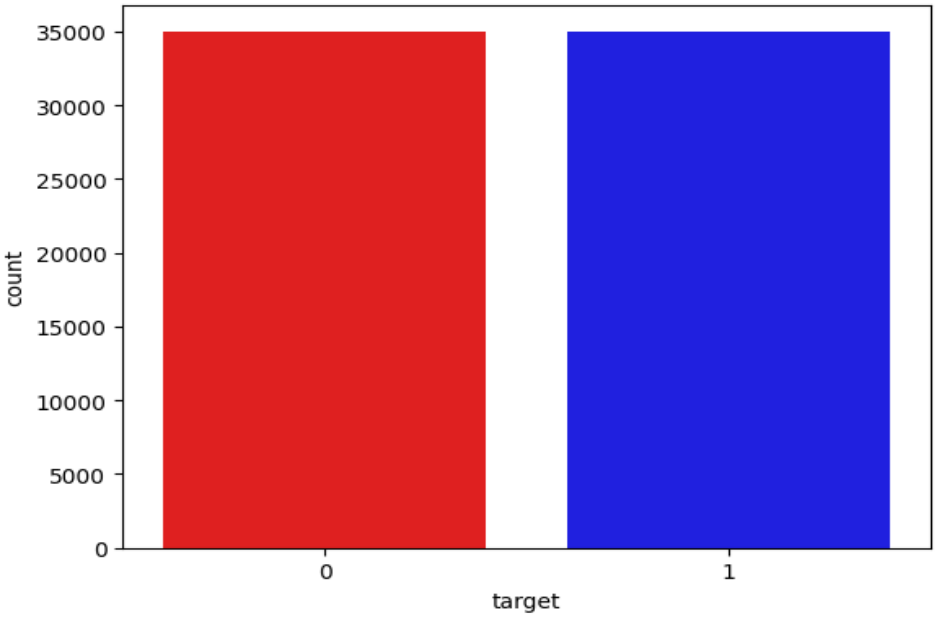


Figure 3.2: Count plot of Cardio Dataset

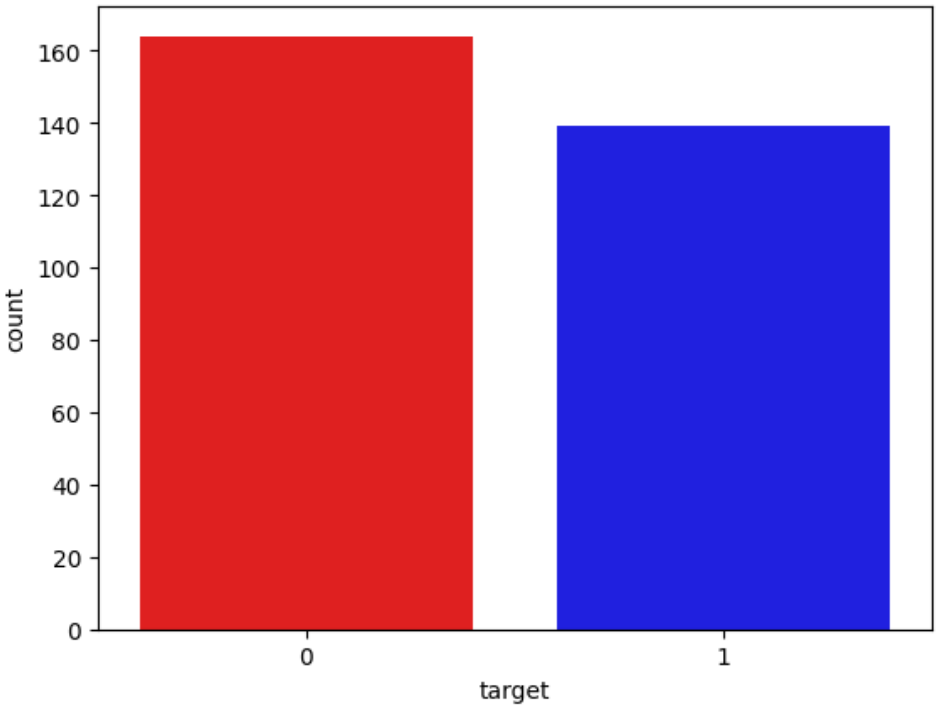


Figure 3.3: Count plot of Cleveland Dataset

3.2.1 Categorical Data Encoding

The process of converting nominal values from categorical data is known as the "data encoding system of categorical method". The category encoding approach becomes crucial in machine learning since input and output are often numerical. Using the categorical data encryption approach in our research, which involved columns like target column, was crucial.

3.2.2 Missing Value Imputation

Imputation is the process of substituting values from another dataset with research-derived values to fill in the blanks or missing data. That being said, the fact that our dataset had two null values is notable and promising. We have replaced 0 in null values to handle them.

3.2.3 Handling Imbalanced Data

Through the process of incrementally adding new samples to the dataset, the original data's category distribution is altered. The representation of minority data is improved when the

entire dataset is used as input. We used SMOTE technique to balance the imbalance dataset Hungarian. The after result of SMOTE technique is shown below count plot in Figure 3.4.

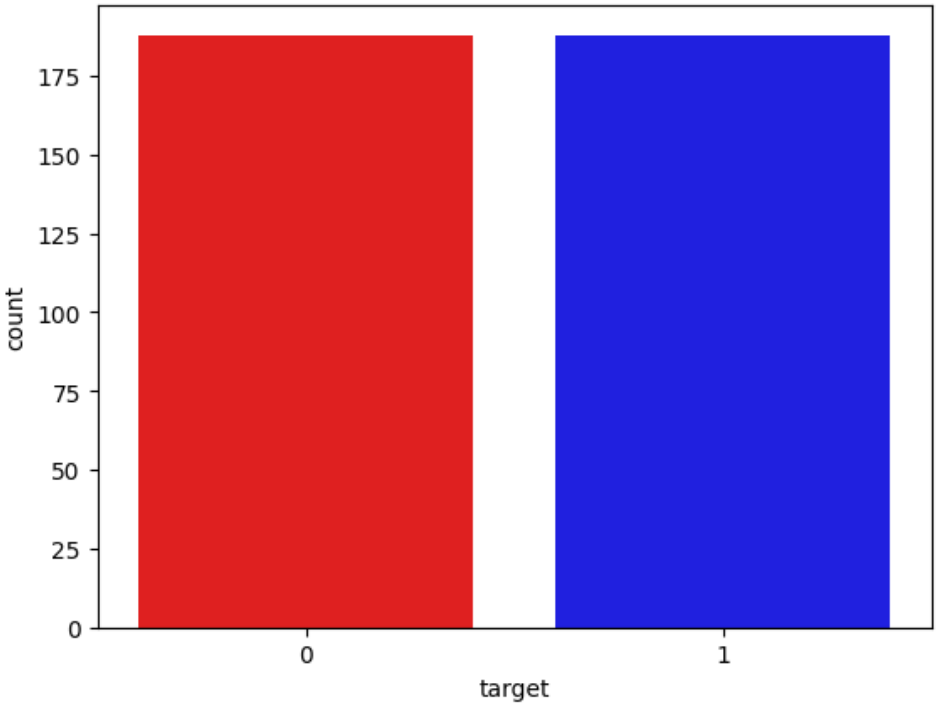


Figure 3.4: Count plot of Cleveland Dataset

3.2.4 Feature Scaling

It is a technique for homogenizing a range of dissimilar independent datasets. The Standard Scalar measure scales all relevant data to the interval [0, 1] when there are no local variables. However, in the event that local variables are available, it scales the relevant data to the range [-1, 1].

3.3 Statistical Analysis

In our research, the analysis stage is essential to the creation and assessment of the algorithms we have used. Our method consists of a series of processes, beginning with data collection and pre-processing, which includes actions such as filling in missing numbers and fixing discrepancies. After extracting insights using exploratory data analysis approaches, machine learning models are developed and assessed using appropriate metrics. This comprehensive analysis contributes to the overall success of our study by ensuring that the dataset is ready and reliable for algorithm creation and assessment.

3.4 Proposed Methodology

Random Forest (RF), Logistic Regression (LR), Decision Tree (DT), K-Neighbors Classifier (KNN), XGB Classifier (XGB) and Gradient Boosting Classifier (GB) were used in our proposed model.

3.4.1 Traditional Classifiers

Logistic Regression

Modeling the likelihood of a binary result using one or more predictor variables, logistic regression is a method of statistics used for binary classification. Typically coded as 0 or 1, logistic regression predicts the chance that an instance belongs to a certain category, as opposed to linear regression's prediction of a continuous result. This is accomplished by converting the linear sum of the given input variables into an amount between 0 and 1 using the logistic function, commonly referred to as the sigmoid function. The model provides coefficients that indicate the influence of each predictor and calculates the likelihood that the dependent event will occur [21-23]. In order to predict the existence of diseases, consumer behavior, and other things, logistic regression is often utilized in the social sciences, medical, and machine learning domains.

Random Forest

For problems involving regression and classification, an ensemble learning technique called a random forest is commonly employed. It functions by building many decision trees in training and producing the mean prediction (regression) or mode of grouping (classification) of each individual tree. The "forest" is constructed using a process known as bagging (Bootstrap Aggregating), in which a decision tree is taught on each of the dataset's many subsets that are produced via replacement. This method averages out variances and biases from individual trees, which helps to decrease overfitting and increase the model's generalization [24-26]. The usefulness of random forests in producing precise forecasts without requiring substantial parameter adjustment is well-known, as is their robustness and capacity to handle big datasets with increased dimensionality.

Gradient Boosting

For regression and classification applications, gradient boosting is a potent machine learning approach that creates an ensemble of weak learners, usually decision trees. In order to fix the mistakes produced by earlier models, iteratively adding new models while concentrating on the data points that are most difficult to forecast correctly is how it operates. Every new model is trained to gradually minimize the total prediction error by reducing the residual errors of the aggregate ensemble. Because gradient boosting uses a sequential technique, it may produce very precise prediction models. This makes it very useful for managing complicated datasets and identifying minute trends. To avoid overfitting and maximize performance, nevertheless, meticulous hyperparameter adjustment is necessary [27-29].

K-Nearest

A straightforward, non-parametric, and user-friendly technique for classification and regression problems in machine learning is the k-Nearest Neighbors (k-NN) classifier. It functions by finding the k instances in the training dataset that are the closest to a given input (neighbors), and then predicting the output based on the average value (for regression) or majority class (for classification) of these neighbors. Distance measures like Euclidean distance are commonly used to quantify similarity. Even though k-NN is straightforward, it can be effective, particularly for short datasets or in situations where the decision border is extremely irregular. It is, however, sensitive to the choice of k and the distance measure and computationally costly for big datasets. For data to perform better, feature selection and data scaling must be done correctly [30-31].

XGB Classifier

The XGBoost (Extreme Gradient Boosting) classifier is a powerful and efficient implementation of the gradient boosting framework, specifically designed for supervised learning tasks. It excels in predictive performance by combining an ensemble of weak prediction models, typically decision trees, to create a robust predictive model. XGBoost offers several key advantages, including handling missing values, incorporating

regularization to prevent overfitting, and leveraging advanced tree learning algorithms that optimize speed and performance. Its scalability and flexibility make it a popular choice for various applications, from structured/tabular data to more complex domains, offering superior accuracy and efficiency in comparison to many traditional machine learning algorithms [29].

3.4.2 Ensemble Methods of Machine Learning

For bagging, boosting, stacking, voting, and random subspace, we used ensemble models [32-36].

Bagging

Bootstrap Aggregating, or Bagging, is a strategy for ensemble learning that aims to increase the precision and resilience of machine learning models. It entails using several subsets of the training data produced by bootstrap sampling—random sampling with replacement—to train numerous base models, usually decision trees. Every model undergoes separate training, and for classification tasks, the predictions are aggregated by a majority vote, and for regression tasks, through averaging. Bagging, as opposed to using a single model, produces more consistent and dependable performance by combining the predictions of many models, which lowers variance and helps prevent overfitting. One of the most well-known uses of this technique is Random Forest, which expands bagging by adding more randomization to the characteristics chosen for each split in the trees.

Boosting

An effective machine learning method that builds a strong overall model by combining several weak classifiers is called a boosting ensemble classifier. Boosting is based on the sequential training of classifiers, where each new model learns from the mistakes made by the prior ones. This is usually accomplished by giving misclassified cases larger weights, which incentivizes the subsequent classifier in the series to fix the errors. AdaBoost and Gradient Boosting are two popular boosting methods that iteratively modify the weights of the classifiers and training data. Boosting improves the final classifier's accuracy and resilience by combining the predictions of all the models, frequently beating the

performance of individual models, especially in situations with complicated patterns and noisy data.

Stacking

An effective machine learning method that mixes many base models to increase prediction accuracy is the stacking ensemble classifier. This approach involves training several learning algorithms on the same dataset so they can each generate unique predictions. The final prediction is then produced by a second-level meta-model using these predictions as input attributes. By leveraging the strengths and compensating for the weaknesses of the individual base models, stacking often results in superior performance compared to any single model. This approach effectively reduces overfitting and bias, making it particularly useful in complex tasks where no single model performs optimally across all scenarios.

Voting

A machine learning model called a voting ensemble classifier aggregates the predictions of several different classifiers to increase overall resilience and accuracy. The way the ensemble functions is by using a voting mechanism, which can be either soft or hard voting, to aggregate the predictions of its component models. Hard voting selects the class with the majority of votes as the final forecast, with each classifier voting for a particular class label. During the soft voting process, the class with the greatest average probability is selected by averaging the projected probabilities of each class from all classifiers. This approach leverages the strengths of diverse models, reducing the likelihood of errors and increasing the performance, especially in complex and noisy datasets. By pooling the decisions of several models, a voting ensemble can achieve better generalization compared to any individual model alone.

Random Subspace

By merging many classifiers, each trained on a different subset of the feature space, a random subspace ensemble classifier is a machine learning approach that improves the performance and durability of classification models. Random subspace techniques use a

random subset of features for each classifier rather than utilizing the complete set of features, producing a variety of models that can capture various facets of the data. This variety enhances generalization and lessens overfitting. By aggregating the predictions of these diverse classifiers, usually through methods like majority voting or averaging, the ensemble often achieves better performance than individual classifiers, particularly in high-dimensional spaces or when dealing with noisy or sparse data. This approach leverages the strength of ensemble learning to handle complex patterns and improve predictive accuracy.

3.4.3 Flow chart of proposed methodology:

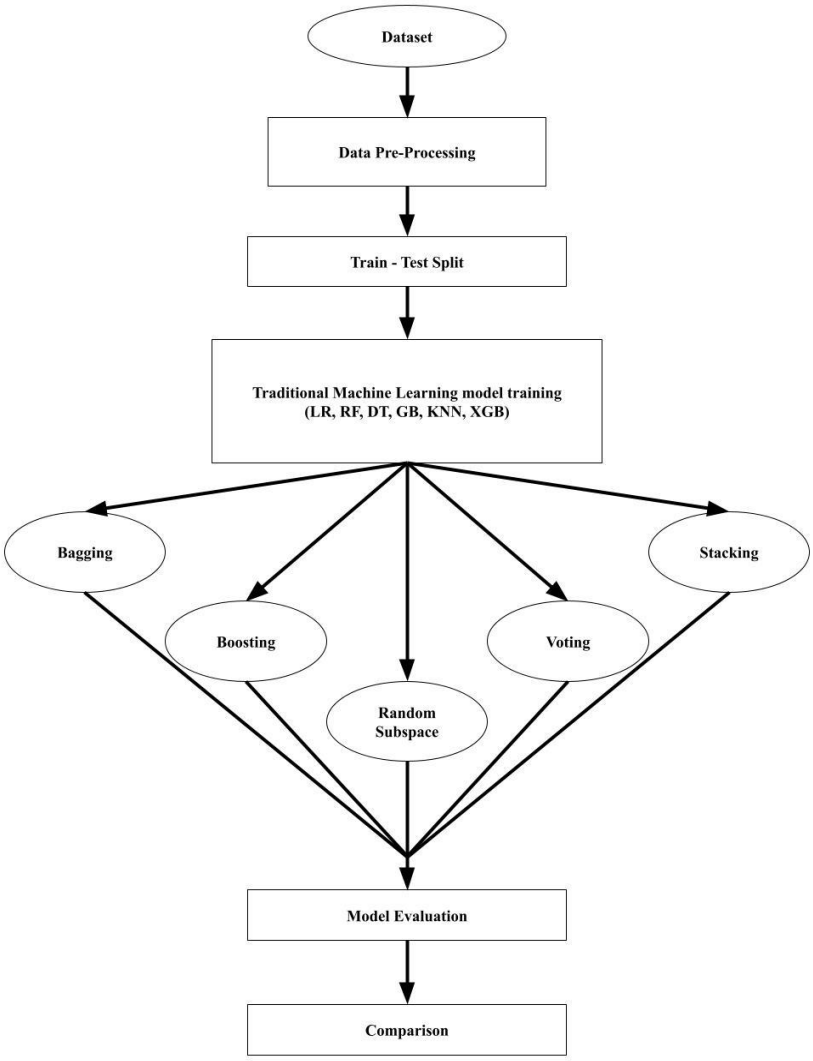


Figure 3.5: Methodology of Cardiovascular Disease Prediction

Figure 3.5 The research utilizes three datasets sourced from Kaggle and the UCI machine learning library. First, the dataset we selected was put into use. By locating and updating any inaccurate or missing numbers, we guaranteed the integrity of the data. Next, we employed a range of algorithms and assessed their performance. It employs a variety of algorithm models, including Gradient Boosting (GB), K-Neighbors Classifier (KNN), XGB Classifier (XGB), Random Forest (RF), Logistic Regression (LR), and Decision Tree (DT), to predict cardiovascular disease risks. To enhance prediction accuracy, the study

also incorporates ensemble models such as Bagging, Boosting, Random Subspace, Stacking, and Voting.

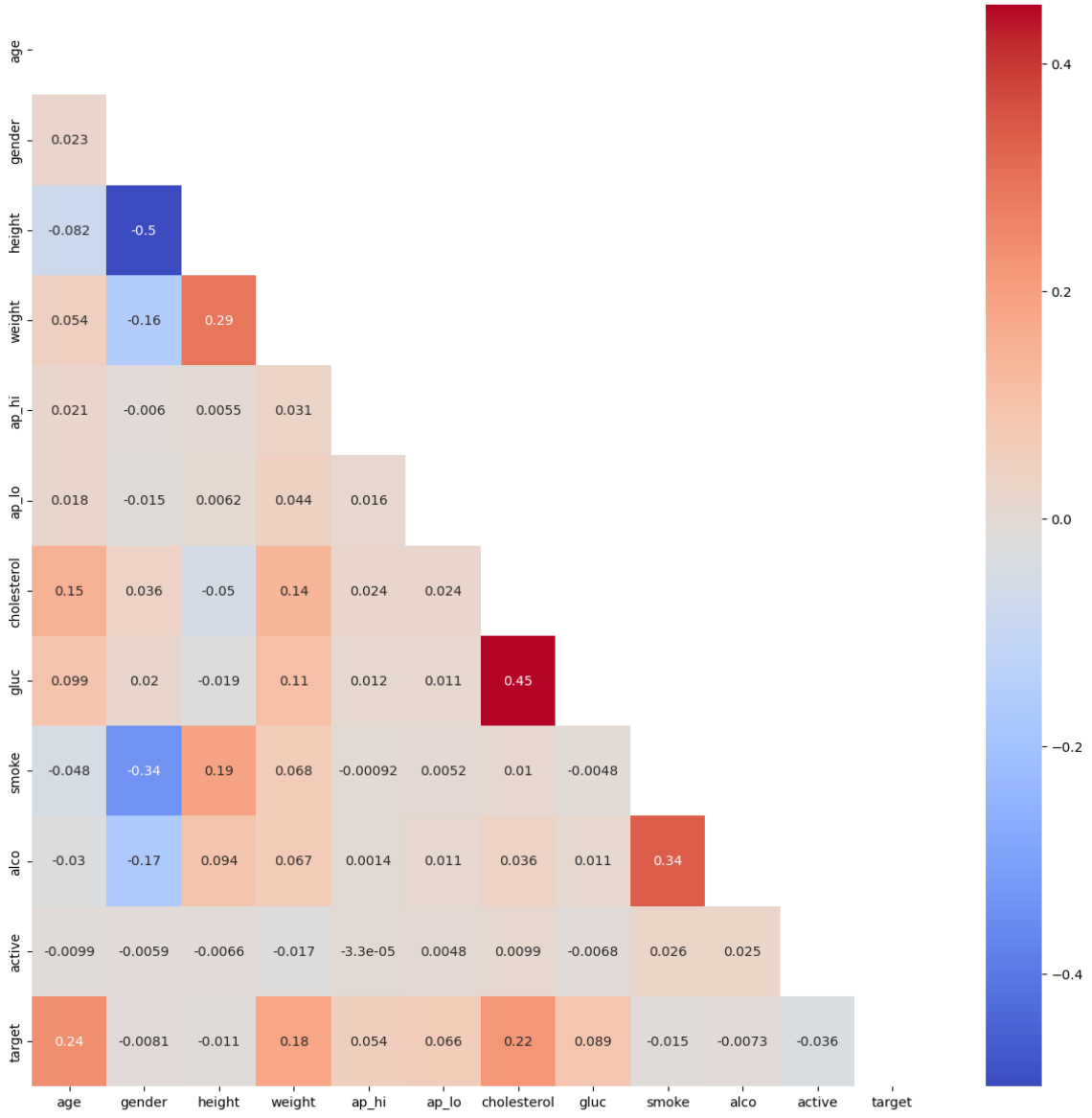


Figure 3.6: Correlated Features of Cardio Dataset

A correlation subplot illustrates the finding of interdependencies between variables that change how they relate to one other. The likelihood that one may be anticipated from the other increases with the degree of interdependence between the variables. This methodology denotes a more profound comprehension of the dataset, augmenting our capacity to discern pivotal elements [37]. Every connected quality that was associated with

the anticipated characteristic "target" was displayed in Fig. 3.6, 3.7 and 3.8 for Cardio dataset, Cleveland dataset and Hungarian dataset.

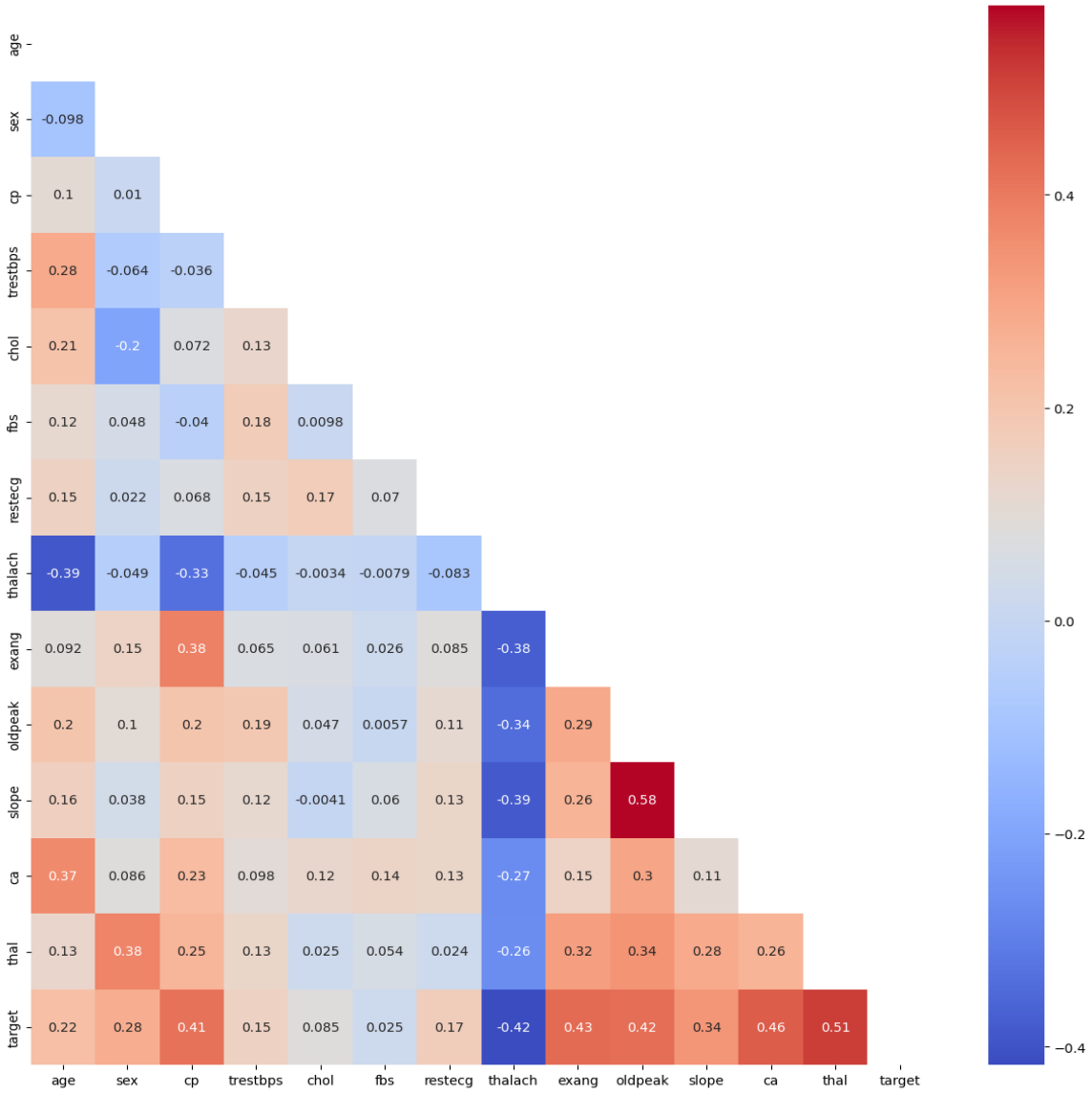


Figure 3.7: Correlated Features of Cleveland Dataset

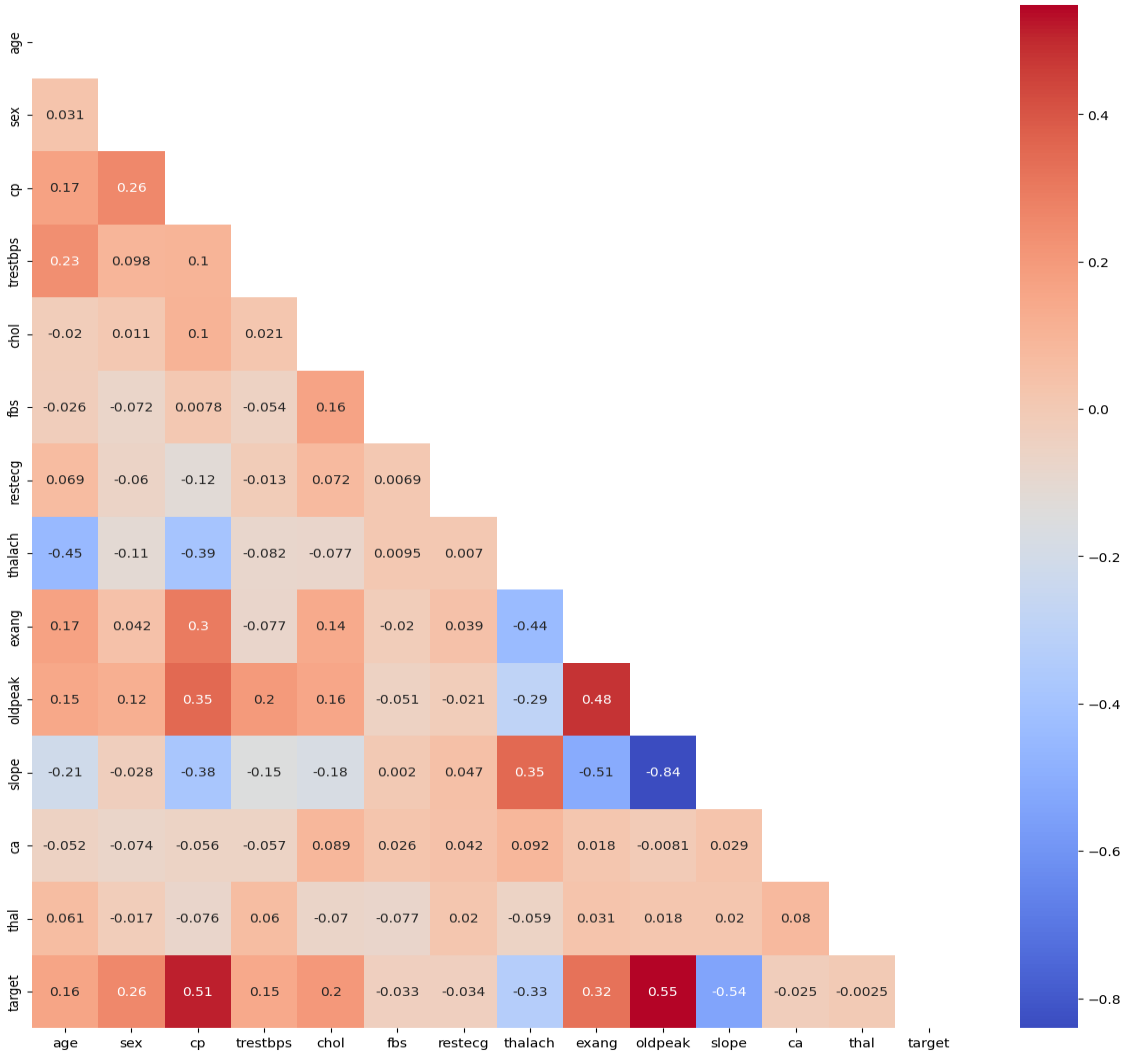


Figure 3.8: Correlated Features of Hungarian Dataset

3.5 Implementation Requirements

We need data sources for the model's research and training in order to put our suggested model into practice. The dataset is thoroughly cleaned to guarantee efficient operation. A variety of filtering methods are applied to tidy up the dataset. Following that, steps for data preparation are completed, such as the Standard Scaler Transform and the conversion of categorical to numerical data. Of the dataset, 20% is utilized for testing and the remaining 80% is used for training. The results are evaluated once algorithms are implemented. To increase prediction accuracy, ensemble techniques including Bagging, Boosting, Stacking, Voting, and Random subspace are applied. The results of these ensemble algorithms are evaluated thoroughly. Over parameter is the output of optimized by applying tuning.

Decisions are taken after a comprehensive evaluation of the applied models and outcome analysis. The learning process then begins with data analysis, and is followed by the application of the model learning and fitting the predictions approach. To get the best accuracy, the models are then put through voting, stacking, bagging, and boosting. To facilitate well-informed judgments during the learning process, the model with the highest accuracy, precision, recall, and F-1 score is chosen.

CHAPTER 4

EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Experimental Setup

The supervised learning technique used in this work is based on training and assessment. Using the training datasets, a classification model is constructed as part of the procedure. The result is then obtained by applying the developed model to the testing dataset. In the sections that follow, the machine-learning algorithm will be covered in more detail.

4.2 Experimental Result & Analysis

It was now essential to evaluate the models that were already in place. We used performance assessment metrics to gauge our suggested model's effectiveness. These techniques use data that hasn't been seen to assess overall performance. This section contains an analysis report that was created using our targeted coronary artery disease dataset and the experimental outputs of machine learning models. First, the dataset we selected was put into use. By locating and updating any inaccurate or missing numbers, we guaranteed the integrity of the data. Next, we employed a range of algorithms and assessed their performance. For classic methods like Random Forest (RF), Logistic Regression (LR), Decision Tree (DT), K-Neighbors Classifier (KNN), XGB Classifier (XGB), and Gradient Boosting Classifier (GB), confusion matrices were produced. We also experimented with several ensemble methods, evaluating their effectiveness using confusion matrices.

Accuracy

The percentage of correct predictions generated from testing data is covered by the measure. Comparing the expected results with the actual measurements, based on a single variable, yields the accuracy. This basic measurement methodology is one of the easiest ways to evaluate a model because it mostly concentrates on deliberate mistakes. Making sure that models are accurate is an essential part of what I do.

$$Accuracy = \frac{TruePositive + TrueNegative}{TruePositive + FalsePositive + TrueNegative + FalseNegative}$$

Precision

The accuracy, or the percentage of positively predicted observations that really occurred, is examined in this section. Precision is the actual percentage of all cases that were correctly predicted to be true. It's crucial to remember that a high recall rate might be deceiving for any kind of model.

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive}$$

Recall

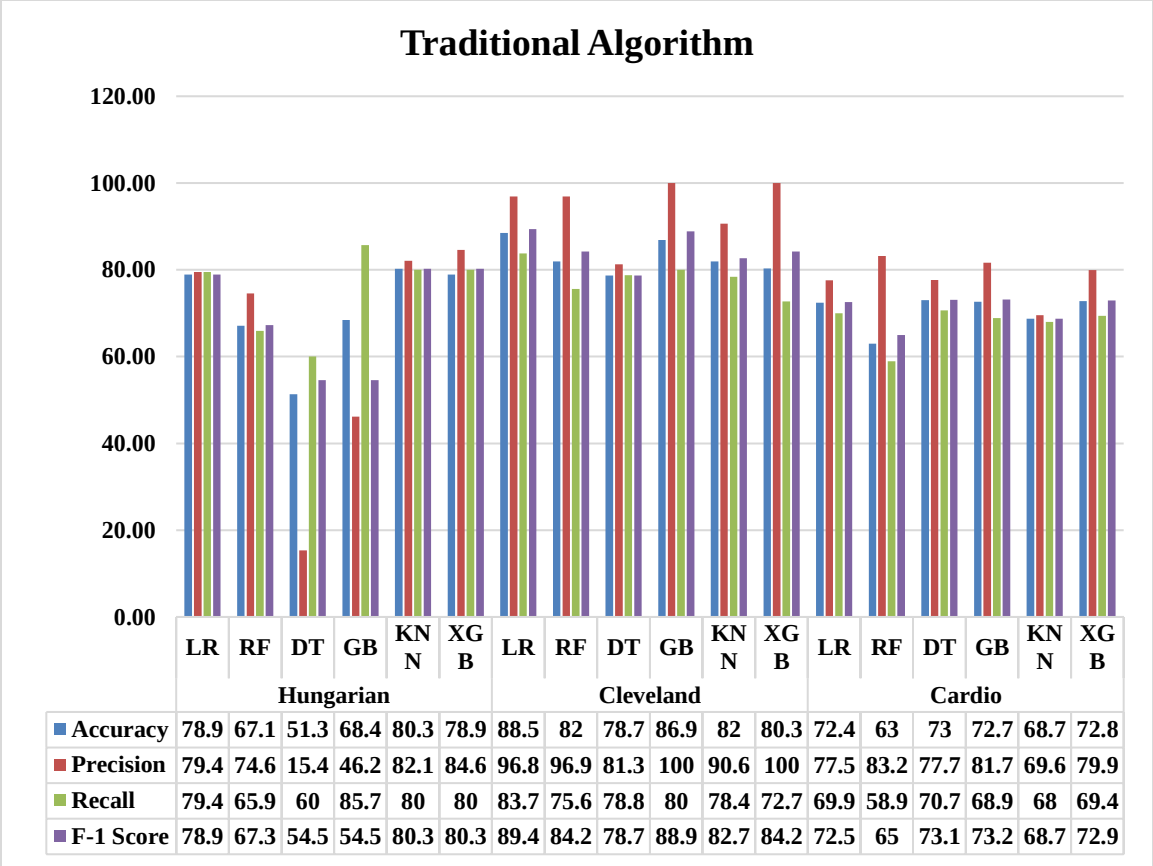
This has to do with the expectedly positive ratio of data from a model. It's important to remember, though, that great accuracy can occasionally be deceptive. In this context, recall—typically defined as the ratio of all positive labels to expected positives—is an important parameter.

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative}$$

F-1 Score

Recall and accuracy ratios are two important metrics that are highlighted in this section along with the harmonic and accuracy techniques of recall. It is mentioned that the model could not function properly if the mean of the harmonic measurements is low.

$$F - 1\ Score = 2 * \frac{Recall * Precision}{Recall + Precision}$$



	LR	RF	DT	GB	KNN	XGB	LR	RF	DT	GB	KNN	XGB	LR	RF	DT	GB	KNN	XGB
	Hungarian						Cleveland						Cardio					
Accuracy	78.9	67.1	51.3	68.4	80.3	78.9	88.5	82	78.7	86.9	82	80.3	72.4	63	73	72.7	68.7	72.8
Precision	79.4	74.6	15.4	46.2	82.1	84.6	96.8	96.9	81.3	100	90.6	100	77.5	83.2	77.7	81.7	69.6	79.9
Recall	79.4	65.9	60	85.7	80	80	83.7	75.6	78.8	80	78.4	72.7	69.9	58.9	70.7	68.9	68	69.4
F-1 Score	78.9	67.3	54.5	54.5	80.3	80.3	89.4	84.2	78.7	88.9	82.7	84.2	72.5	65	73.1	73.2	68.7	72.9

Figure 4.1: Results of Traditional Classifiers

The comparison analysis of classifier outputs across three datasets—Hungarian, Cleveland, and Cardio—reveals varying performances among different algorithms shows in Figure 4.1. Logistic Regression (LR) consistently demonstrates competitive accuracy across datasets, particularly excelling in the Cleveland dataset with an accuracy of 88.52%, precision of 96.87%, recall of 83.78%, and an F-1 score of 89.40%. Random Forest (RF) exhibits strong precision scores, notably in the Hungarian and Cleveland datasets, yet its performance in terms of accuracy and recall varies. Decision Trees (DT) generally show lower accuracy and precision compared to other algorithms across datasets. Gradient Boosting (GB) achieves the highest precision of 99.99% and recall of 85.71% in the Cleveland dataset but demonstrates inconsistency in other datasets. K-Nearest Neighbors (KNN) yields relatively stable performance across datasets, with the highest accuracy of 81.96% and precision of 90.62% in the Cleveland dataset. Extreme Gradient Boosting (XGB) showcases competitive accuracy and precision, especially

notable in the Hungarian and Cleveland datasets, with the highest precision of 99.99% in the Cleveland dataset. Overall, the choice of algorithm should be tailored to specific dataset characteristics and performance metrics of interest, with LR and XGB emerging as strong contenders across datasets for balanced performance in accuracy, precision, recall, and F-1 score. Figure 4.2, Figure 4.3 and Figure 4.4 shows the traditional algorithm AUC-ROC curve for Hungarian dataset, Cleveland dataset and Cardio dataset.

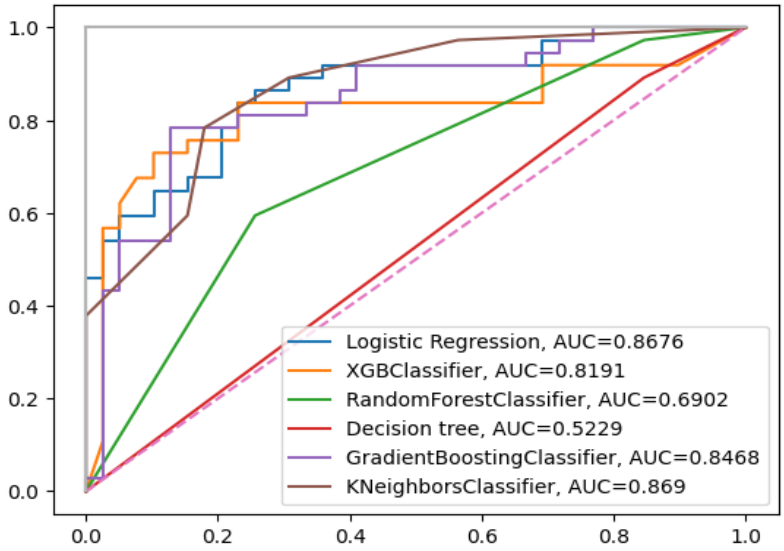


Figure 4.2: Analysis of Traditional algorithms for Hungarian dataset (AUC-ROC)

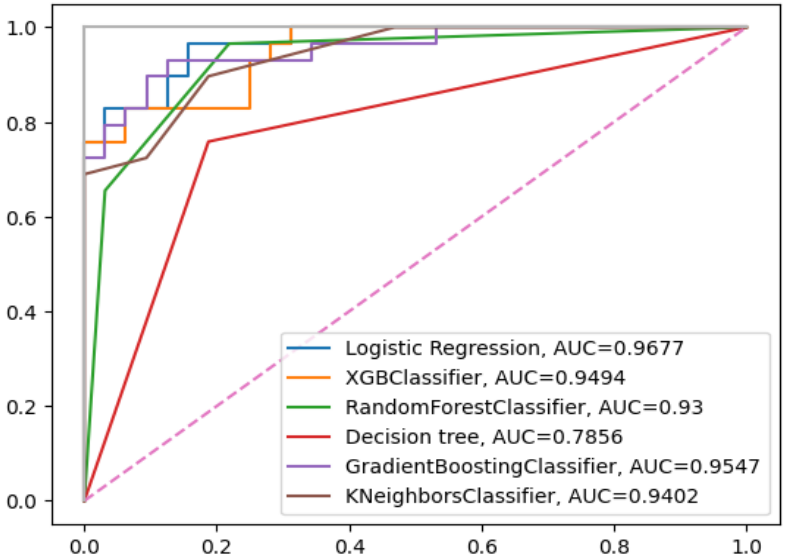


Figure 4.3: Analysis of Traditional algorithms for Cleveland dataset (AUC-ROC)

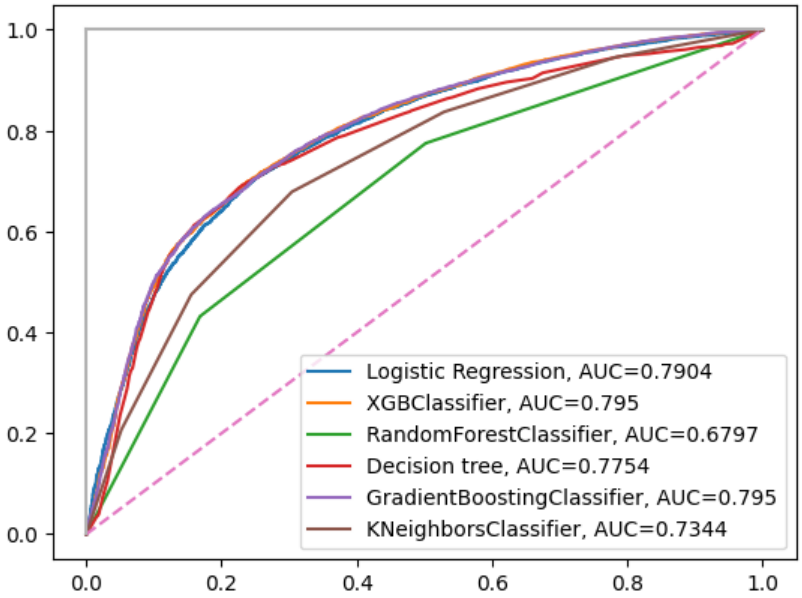
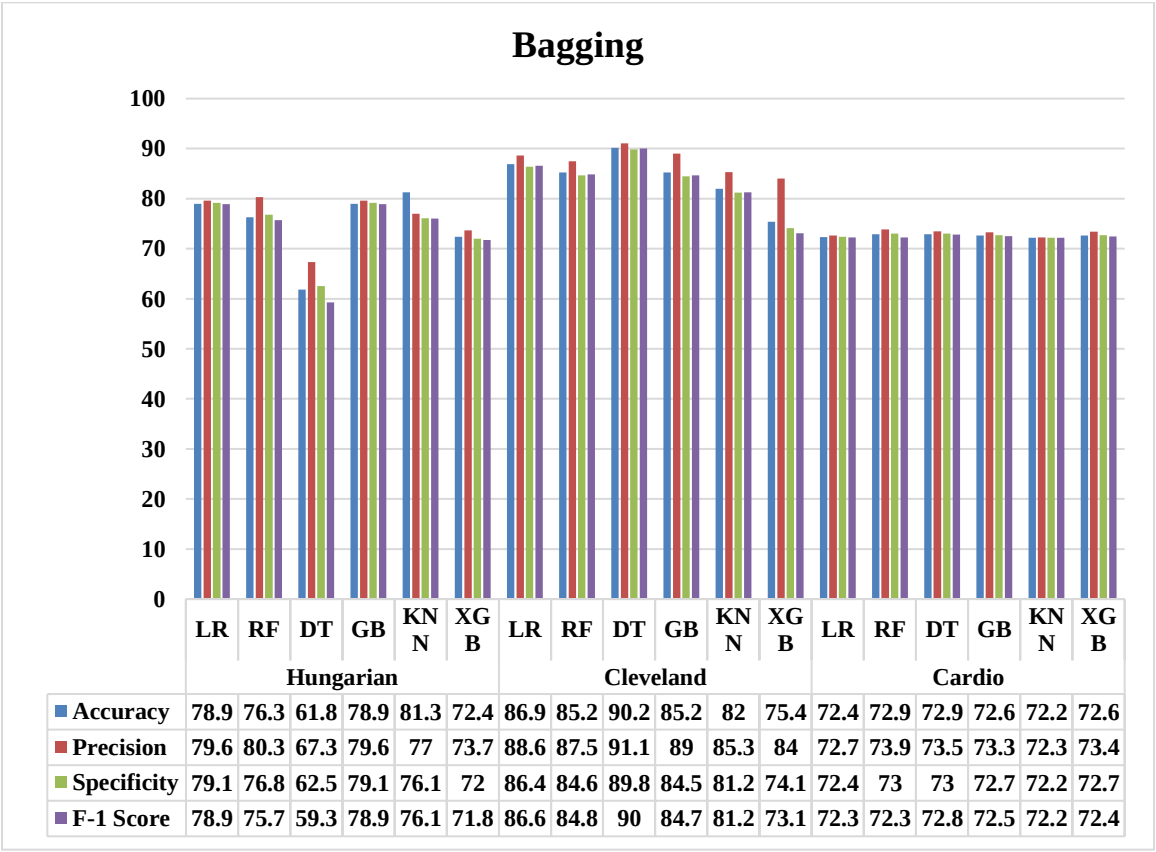


Figure 4.4: Analysis of Traditional algorithms for Cardio dataset (AUC-ROC)



	Hungarian						Cleveland						Cardio					
	LR	RF	DT	GB	KNN	XGB	LR	RF	DT	GB	KNN	XGB	LR	RF	DT	GB	KNN	XGB
Accuracy	78.9	76.3	61.8	78.9	81.3	72.4	86.9	85.2	90.2	85.2	82	75.4	72.4	72.9	72.9	72.6	72.2	72.6
Precision	79.6	80.3	67.3	79.6	77	73.7	88.6	87.5	91.1	89	85.3	84	72.7	73.9	73.5	73.3	72.3	73.4
Specificity	79.1	76.8	62.5	79.1	76.1	72	86.4	84.6	89.8	84.5	81.2	74.1	72.4	73	73	72.7	72.2	72.7
F-1 Score	78.9	75.7	59.3	78.9	76.1	71.8	86.6	84.8	90	84.7	81.2	73.1	72.3	72.3	72.8	72.5	72.2	72.4

Figure 4.5: Experimental Results of Bagging

The comparison analysis for bagging of these algorithms across the Hungarian, Cleveland, and Cardio datasets highlights notable performance variations shows in Figure 4.5. In the Hungarian dataset, Random Forest (RF) demonstrates the highest accuracy of 76.31% and precision of 80.31%, followed closely by K-Nearest Neighbors (KNN) with an accuracy of 81.26% and precision of 76.98%. In the Cleveland dataset, Decision Trees (DT) showcase the highest performance across all metrics, with an accuracy of 90.16%, precision of 91.05%, and the highest F-1 score of 90.03%. In the Cardio dataset, all algorithms perform relatively similarly, with Logistic Regression (LR) achieving the highest accuracy of 72.35% and precision of 72.65%. Overall, bagging techniques, particularly Random Forest and Decision Trees, demonstrate strong performances in various datasets, indicating their effectiveness in improving predictive accuracy and robustness through ensemble methods. Figure 4.6, Figure 4.7 and Figure 4.8 shows the Bagging algorithm AUC-ROC curve for Hungarian dataset, Cleveland dataset and Cardio dataset

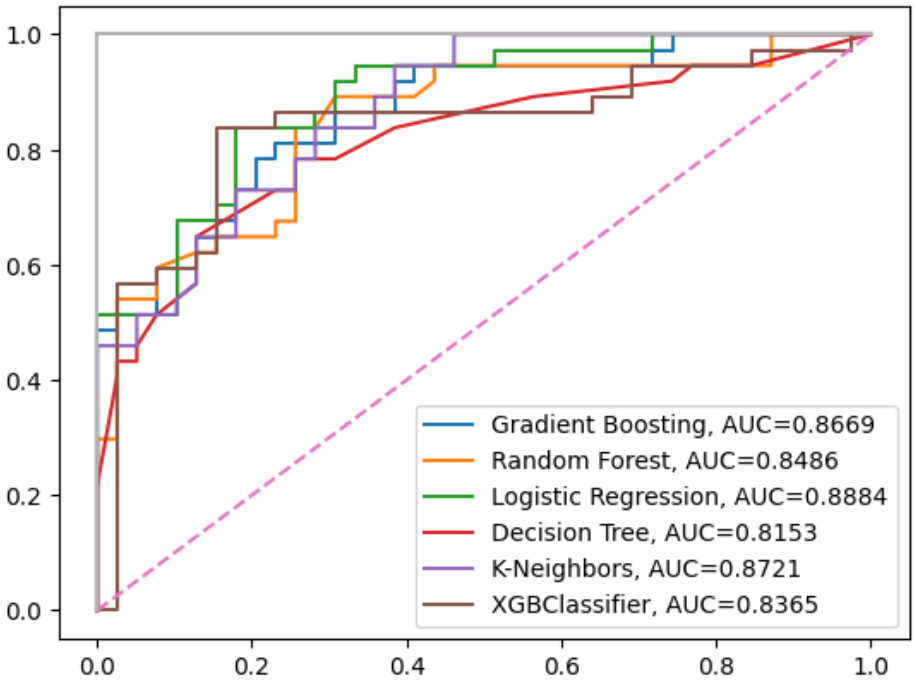


Figure 4.6: Analysis of Bagging for Hungarian dataset (AUC-ROC)

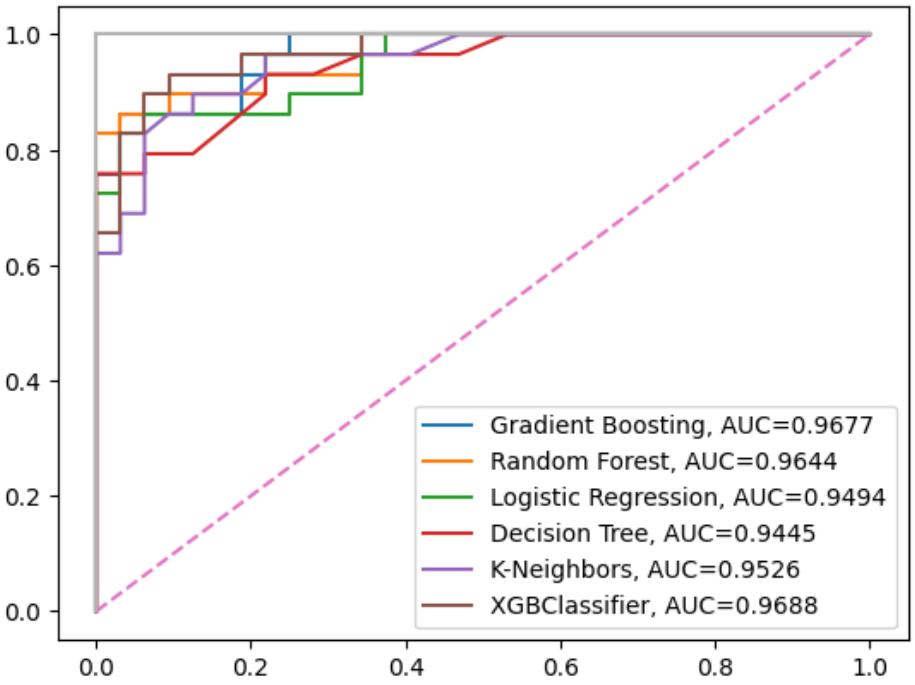


Figure 4.7: Analysis of Bagging for Cleveland dataset (AUC-ROC)

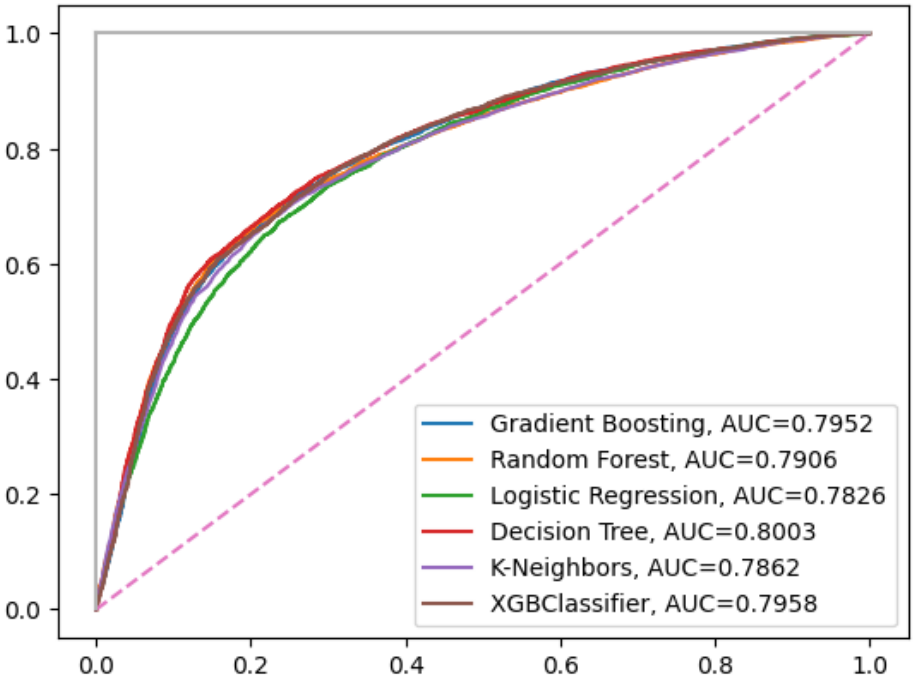
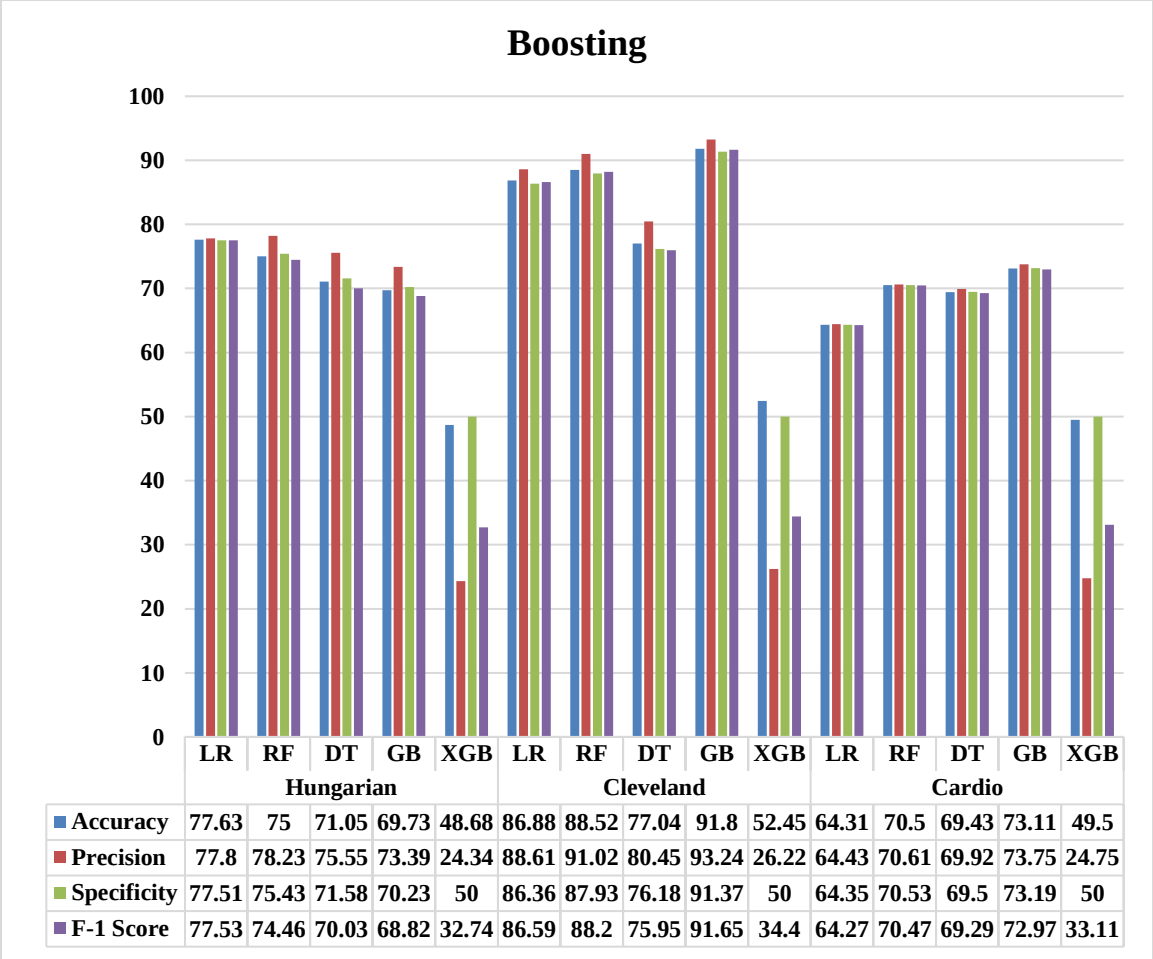


Figure 4.8: Analysis of Bagging for Cardio dataset (AUC-ROC)



	Hungarian					Cleveland					Cardio				
	LR	RF	DT	GB	XGB	LR	RF	DT	GB	XGB	LR	RF	DT	GB	XGB
Accuracy	77.63	75	71.05	69.73	48.68	86.88	88.52	77.04	91.8	52.45	64.31	70.5	69.43	73.11	49.5
Precision	77.8	78.23	75.55	73.39	24.34	88.61	91.02	80.45	93.24	26.22	64.43	70.61	69.92	73.75	24.75
Specificity	77.51	75.43	71.58	70.23	50	86.36	87.93	76.18	91.37	50	64.35	70.53	69.5	73.19	50
F-1 Score	77.53	74.46	70.03	68.82	32.74	86.59	88.2	75.95	91.65	34.4	64.27	70.47	69.29	72.97	33.11

Figure 4.9: Experimental Results of Boosting

The comparative analysis for boosting algorithms across the Hungarian, Cleveland, and Cardio datasets illustrates distinct performance trends shows in Figure 4.9. In the Hungarian dataset, Gradient Boosting (GB) emerges as the top performer, exhibiting the highest accuracy of 69.73%, precision of 73.39%, specificity of 70.23%, and an F-1 score of 68.82%. Conversely, Extreme Gradient Boosting (XGB) demonstrates the lowest performance among algorithms in this dataset, with notably low accuracy and precision. In the Cleveland dataset, GB showcases outstanding performance, achieving the highest accuracy of 91.8%, precision of 93.24%, specificity of 91.37%, and an F-1 score of 91.65%. Random Forest (RF) also exhibits strong performance in Cleveland, particularly in terms of accuracy and precision. In the Cardio dataset, GB continues to demonstrate robust performance, achieving competitive accuracy and precision scores, closely followed by RF. XGB once again underperforms in this dataset compared to other

algorithms. Overall, boosting techniques, especially Gradient Boosting, prove effective in enhancing predictive performance across various datasets, highlighting their utility in classification tasks. Figure 4.10, Figure 4.11 and Figure 4.12 shows the Boosting algorithm AUC-ROC curve for Hungarian dataset, Cleveland dataset and Cardio dataset

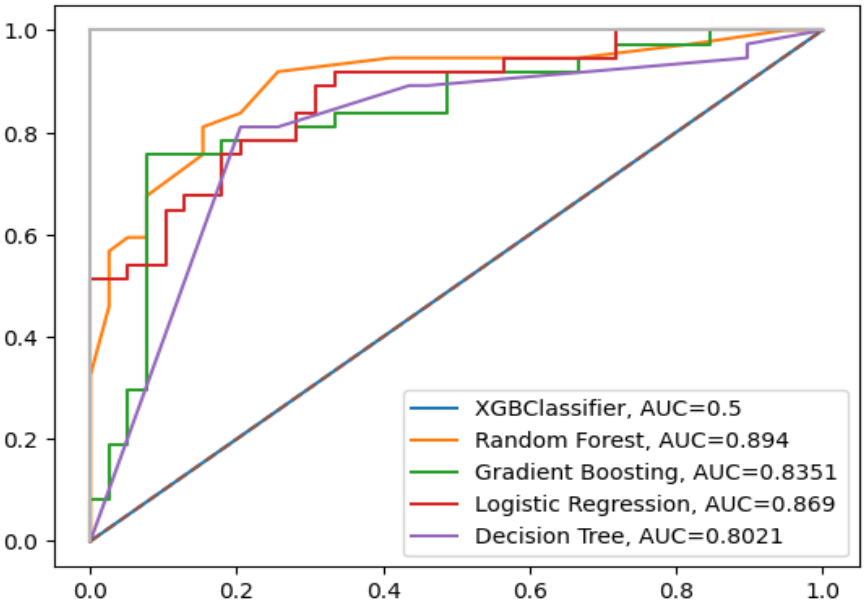


Figure 4.10: Analysis of Boosting for Hungarian dataset (AUC-ROC)

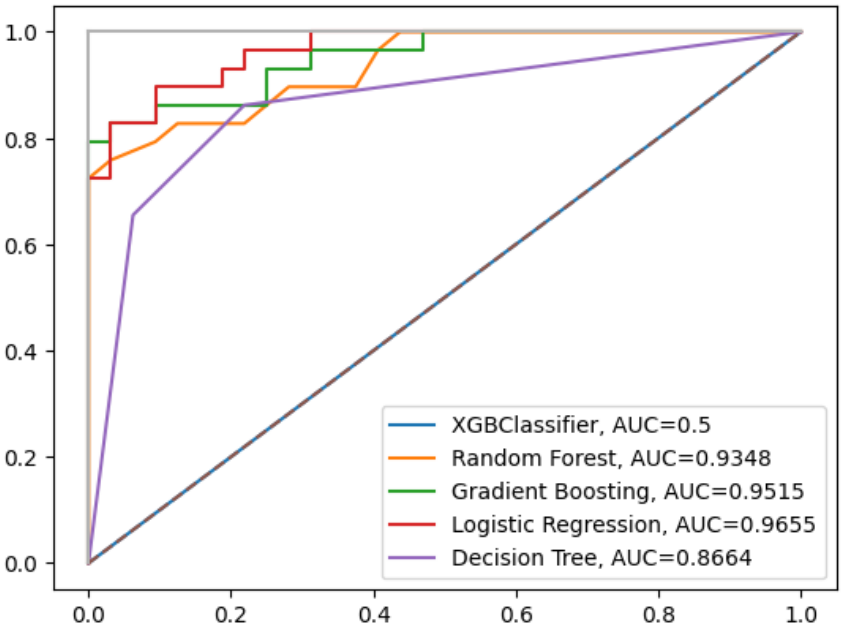


Figure 4.11: Analysis of Boosting for Cleveland dataset (AUC-ROC)

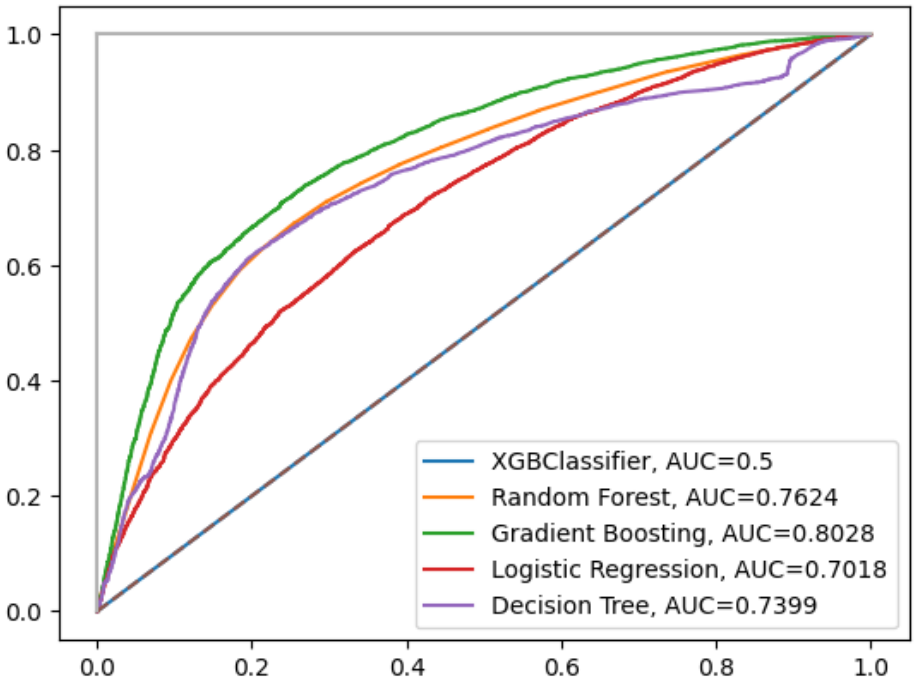
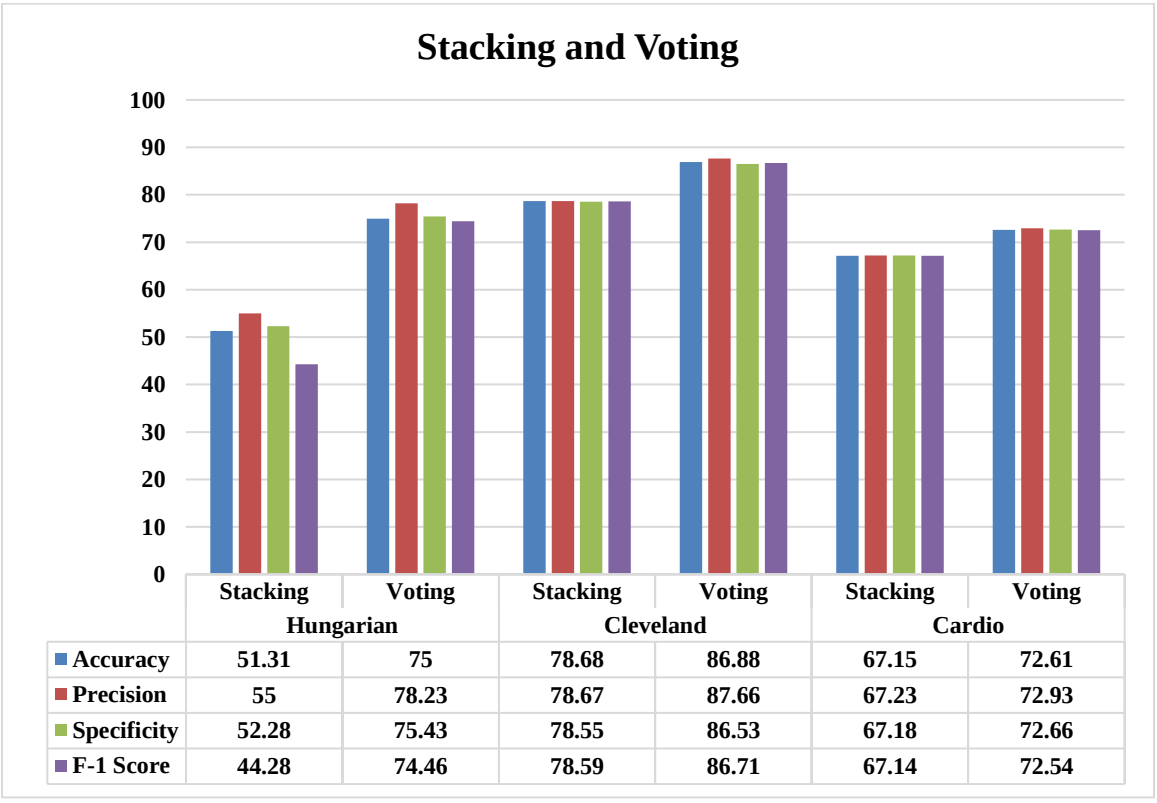


Figure 4.12: Analysis of Boosting for Cardio dataset (AUC-ROC)



	Hungarian		Cleveland		Cardio	
	Stacking	Voting	Stacking	Voting	Stacking	Voting
Accuracy	51.31	75	78.68	86.88	67.15	72.61
Precision	55	78.23	78.67	87.66	67.23	72.93
Specificity	52.28	75.43	78.55	86.53	67.18	72.66
F-1 Score	44.28	74.46	78.59	86.71	67.14	72.54

Figure 4.13: Analysis of Boosting (AUC-ROC)

The comparative analysis between stacking and voting ensemble methods across the Hungarian, Cleveland, and Cardio datasets reveals distinctive performance characteristics shows in Figure 4.13. In the Hungarian dataset, while voting achieves a significantly higher accuracy of 75%, stacking lags behind with an accuracy of 51.31%. However, it's worth noting that stacking demonstrates higher precision and specificity compared to voting, indicating its potential in capturing specific patterns within the data despite its lower overall accuracy. Conversely, voting showcases balanced performance across multiple metrics, including precision, specificity, and F-1 score. In the Cleveland dataset, both stacking and voting demonstrate competitive performances, with voting slightly outperforming stacking across all metrics. However, the differences in performance between the two methods are relatively minor. In the Cardio dataset, a similar pattern emerges, with voting showing slightly higher accuracy and precision compared to stacking. Overall, while voting tends to achieve higher overall accuracy, stacking may offer advantages in capturing specific patterns or characteristics within the data, as indicated by its higher precision and specificity in certain cases. Choosing between stacking and voting would depend on the specific objectives of the classification task and the trade-offs between accuracy and other performance metrics. Figure 4.14, Figure 4.15, Figure 4.16, Figure 4.17, Figure 4.18 and Figure 4.19 shows the Stacking algorithm AUC-ROC curve, Voting AUC-ROC curve for Hungarian dataset, Cleveland dataset and Cardio dataset.

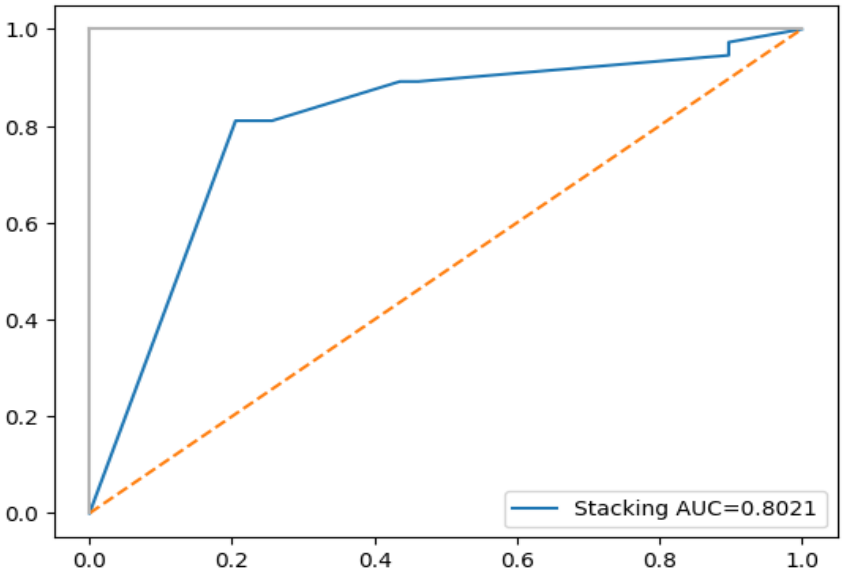


Figure 4.14: Analysis of Stacking for Hungarian dataset (AUC-ROC)

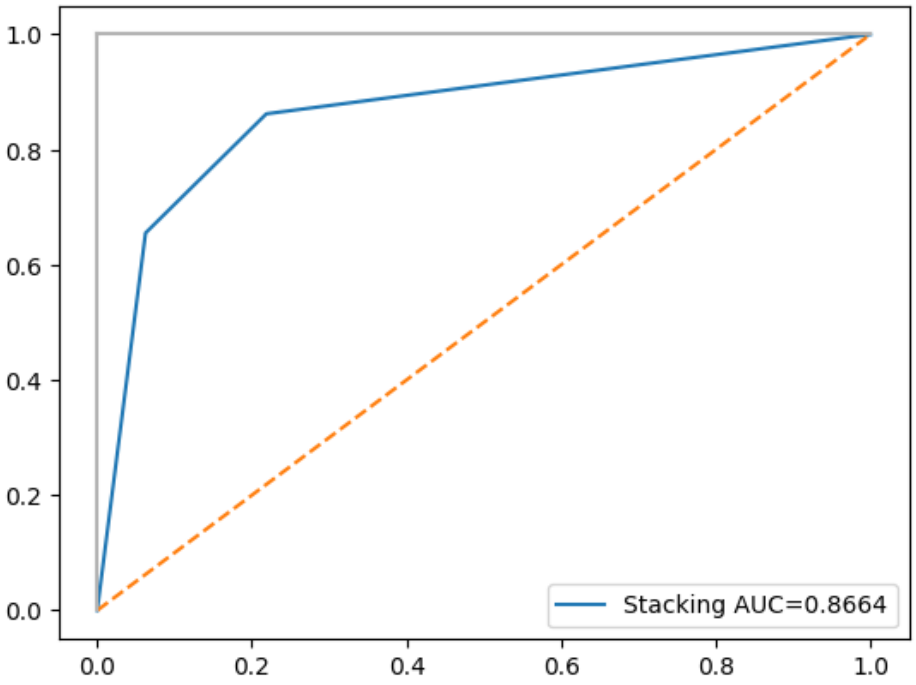


Figure 4.15: Analysis of Stacking for Cleveland dataset (AUC-ROC)

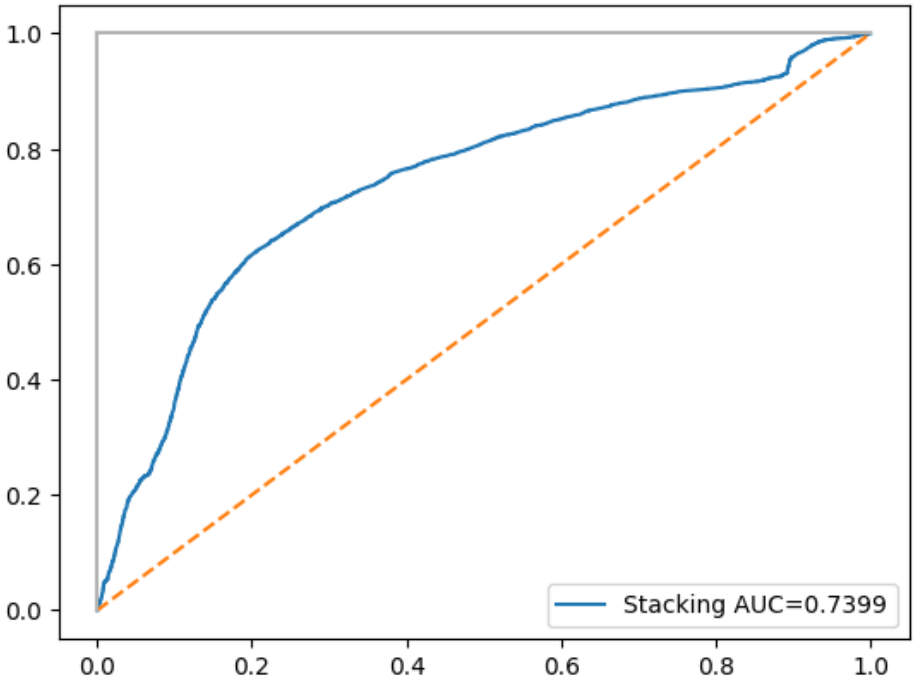


Figure 4.16: Analysis of Stacking for Cardio dataset (AUC-ROC)

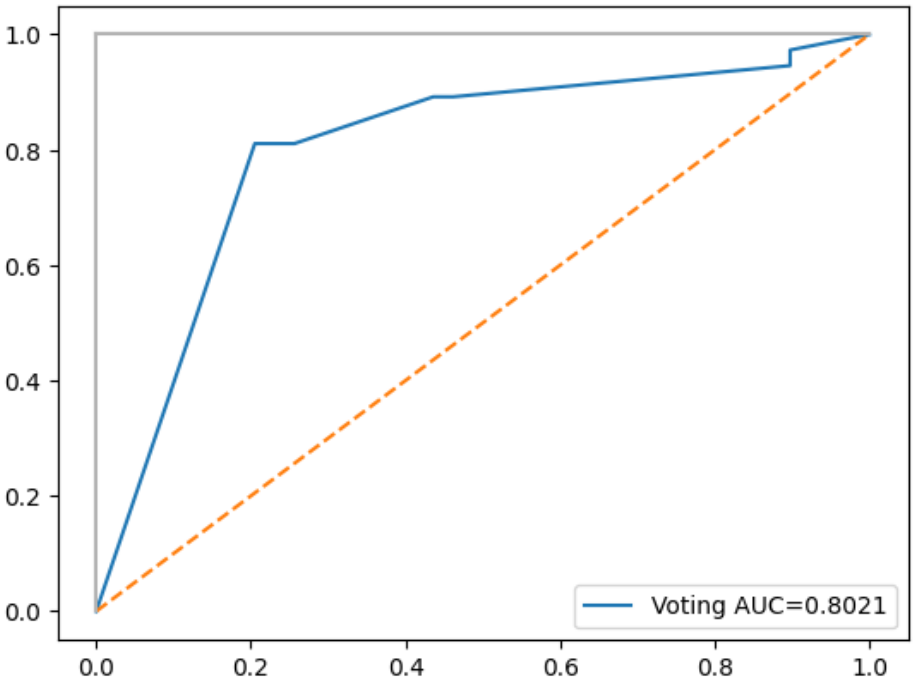


Figure 4.17: Analysis of Voting for Hungarian dataset (AUC-ROC)

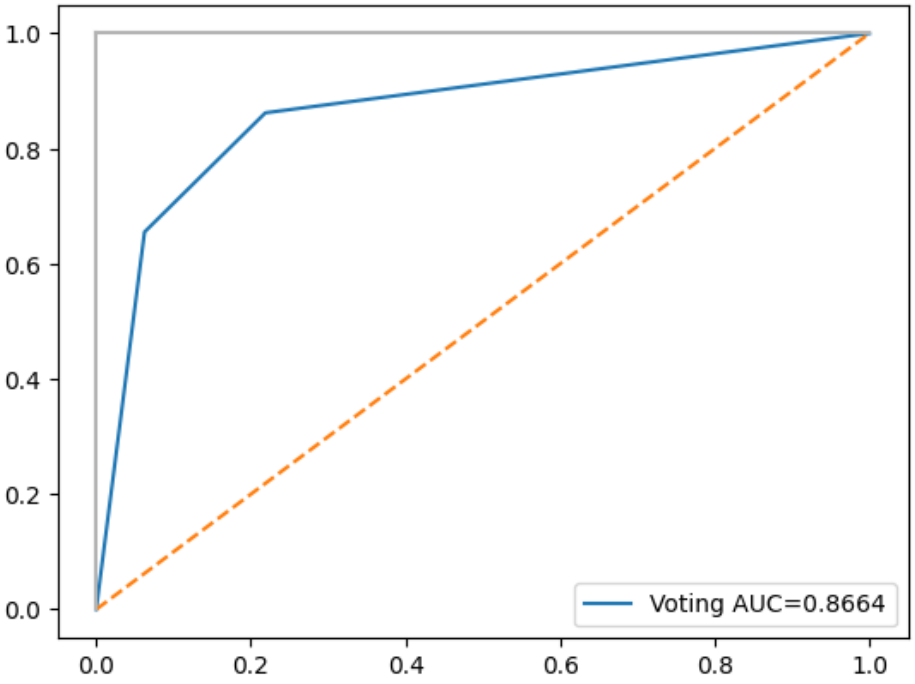


Figure 4.18: Analysis of Voting for Cleveland dataset (AUC-ROC)

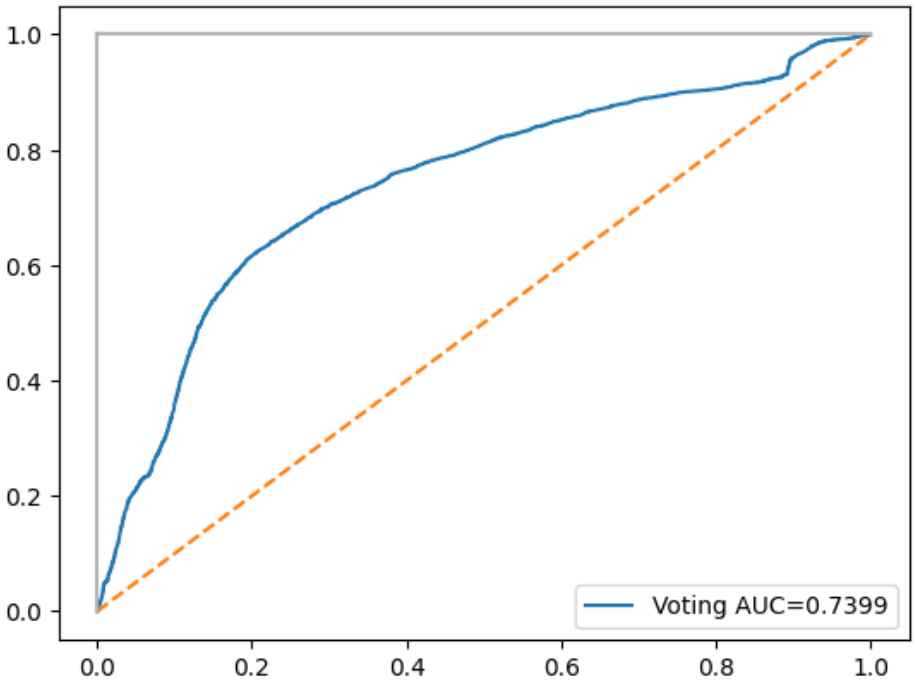
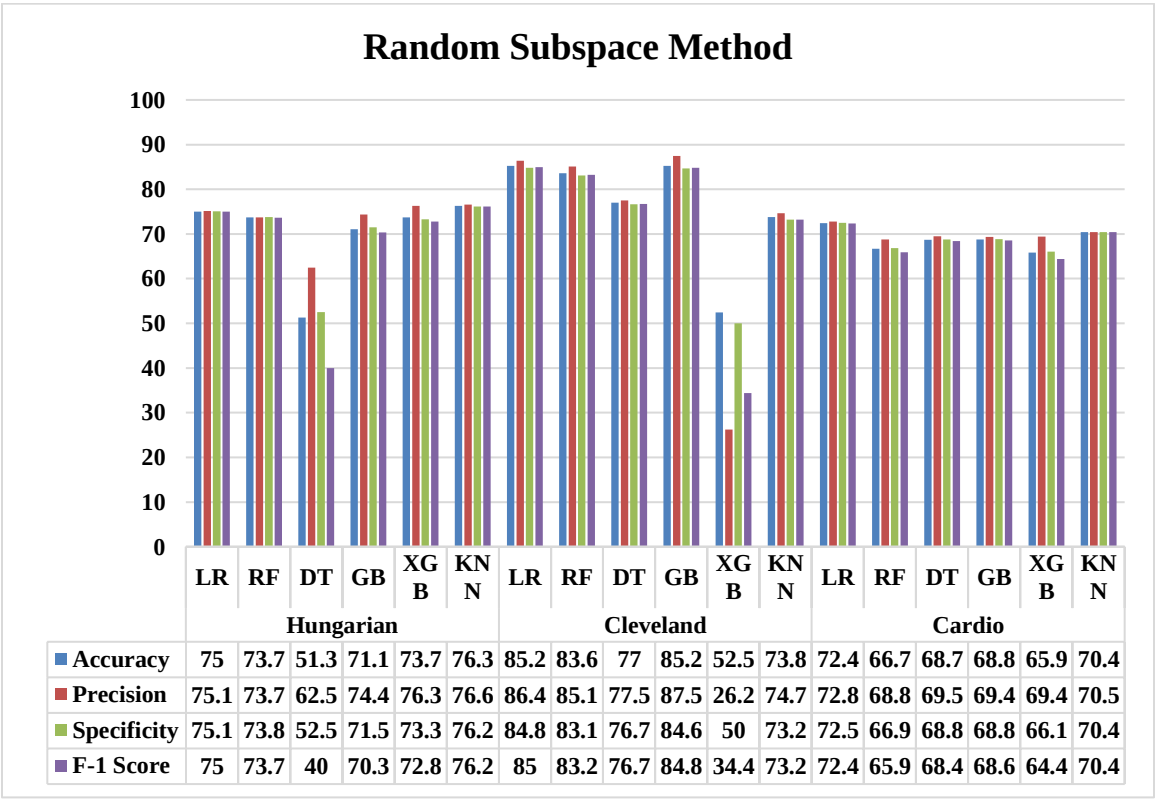


Figure 4.19: Analysis of Voting for Cardio dataset (AUC-ROC)



	Hungarian						Cleveland						Cardio					
	LR	RF	DT	GB	XGB	KNN	LR	RF	DT	GB	XGB	KNN	LR	RF	DT	GB	XGB	KNN
Accuracy	75	73.7	51.3	71.1	73.7	76.3	85.2	83.6	77	85.2	52.5	73.8	72.4	66.7	68.7	68.8	65.9	70.4
Precision	75.1	73.7	62.5	74.4	76.3	76.6	86.4	85.1	77.5	87.5	26.2	74.7	72.8	68.8	69.5	69.4	69.4	70.5
Specificity	75.1	73.8	52.5	71.5	73.3	76.2	84.8	83.1	76.7	84.6	50	73.2	72.5	66.9	68.8	68.8	66.1	70.4
F-1 Score	75	73.7	40	70.3	72.8	76.2	85	83.2	76.7	84.8	34.4	73.2	72.4	65.9	68.4	68.6	64.4	70.4

Figure 4.20: Experimental Results of Random Subspace

The comparative analysis for random subspace ensemble methods across the Hungarian, Cleveland, and Cardio datasets highlights varying performances among different algorithms shows in Figure 4.20. In the Hungarian dataset, K-Nearest Neighbors (KNN) emerges as the top performer, achieving the highest accuracy of 76.31%, precision of 76.6%, specificity of 76.16%, and an F-1 score of 76.16%. Notably, Decision Trees (DT) show lower performance compared to other algorithms, with an accuracy of 51.31% and a precision of 62.5%. In the Cleveland dataset, Logistic Regression (LR) and Gradient Boosting (GB) demonstrate strong performances, with LR achieving an accuracy of 85.24% and GB achieving an accuracy of 85.24%, both showcasing precision scores above 86%. However, Extreme Gradient Boosting (XGB) performs notably lower in this dataset compared to other algorithms. In the Cardio dataset, LR and KNN exhibit the highest performance, with LR achieving an accuracy of 72.42% and KNN achieving an accuracy of 70.42%. Random Forest (RF) shows relatively lower performance across all datasets compared to other algorithms. Overall, the random subspace ensemble method yields varied performances across different datasets and algorithms, with KNN and LR often demonstrating competitive performances across datasets. Figure 4.21, Figure 4.22 and Figure 4.23 shows the traditional algorithm AUC-ROC curve for Hungarian dataset, Cleveland dataset and Cardio dataset.

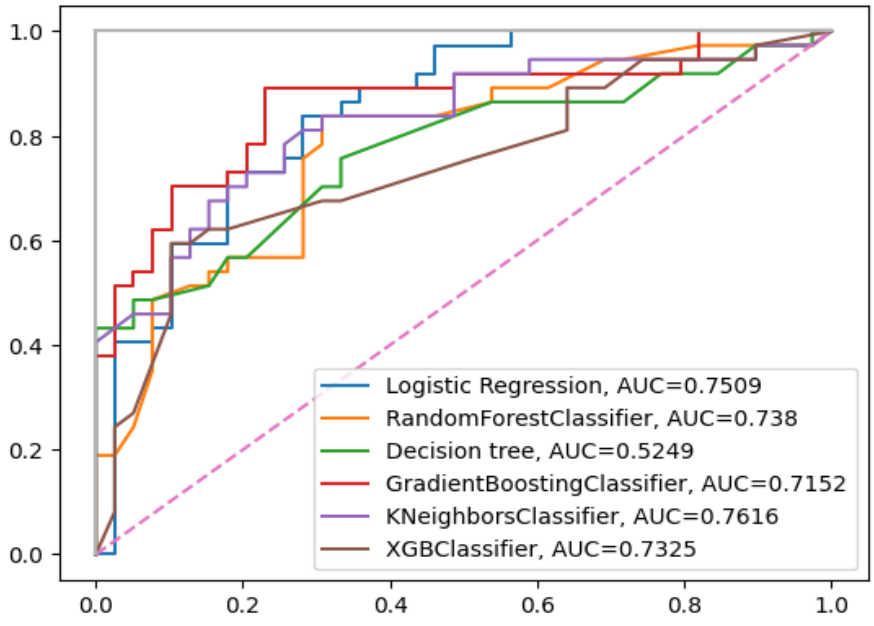


Figure 4.21: Analysis of Random Subspace for Hungarian dataset (AUC-ROC)

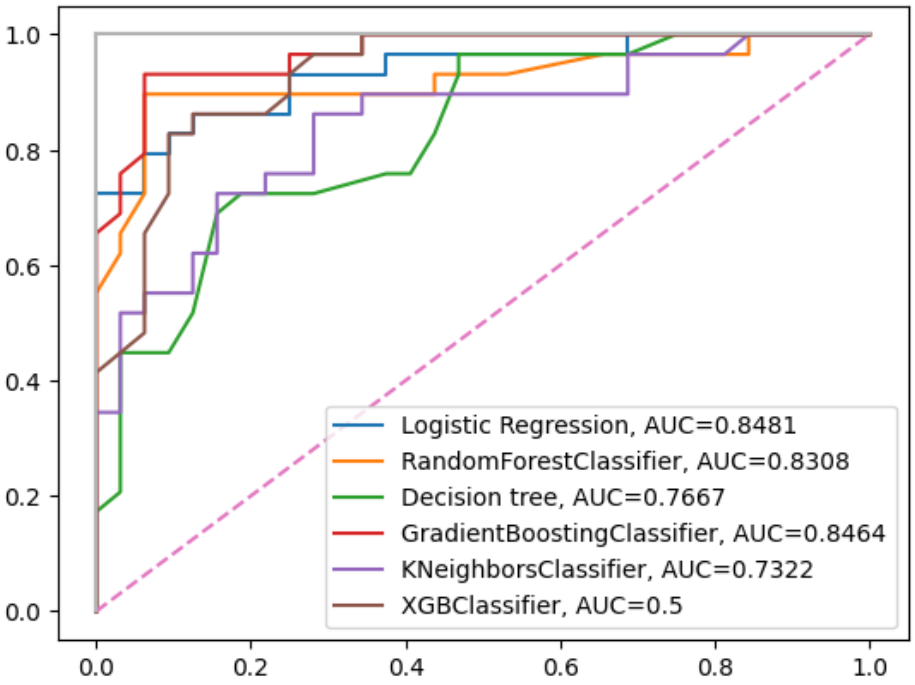


Figure 4.22: Analysis of Random Subspace for Cleveland dataset (AUC-ROC)

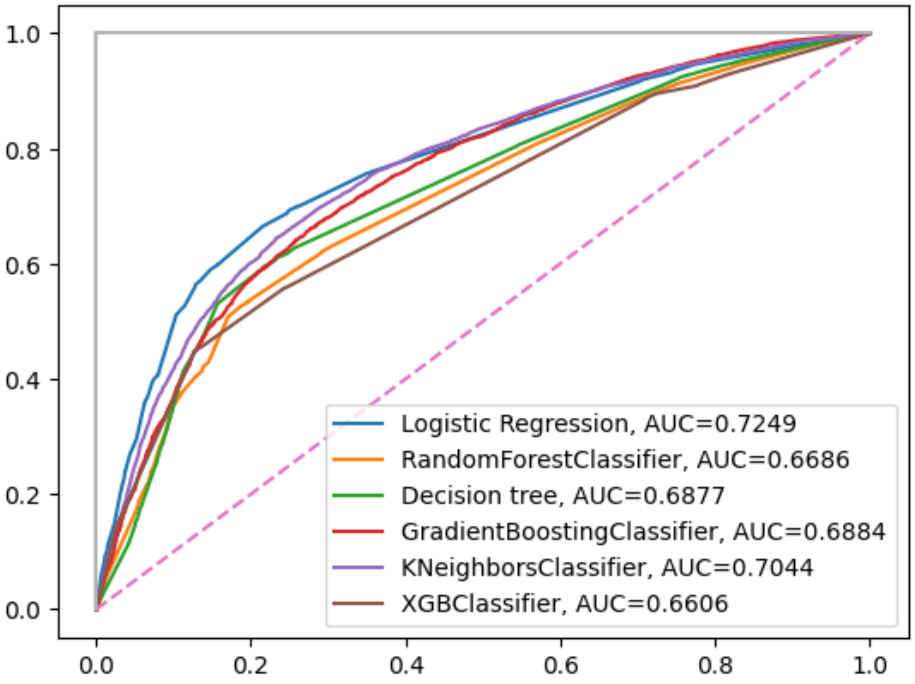


Figure 4.23: Analysis of Random Subspace for Cardio dataset (AUC-ROC)

4.3 Discussion

Several algorithm models were used in this work, including GB, KNN, XGB, RF, LR, and DT. To assess the performances, ensemble models including Bagging, Boosting, Random Subspace, Stacking, and Voting were also used. For the Hungarian dataset, the conventional model KNN earned the highest accuracy with 80.26%, the LR model achieved the greatest accuracy with 88.52%, and the DT model reached the best accuracy with 72.99%. For these three datasets, we also used five distinct kinds of ensemble methods. The findings showed that for the Hungarian dataset, KNN had the greatest accuracy at 81.26%, for the Cleveland dataset, Bagged GB had the best accuracy at 91.8%, and for the Cardio dataset, GB had the best accuracy at 73.11%. By applying hyperparameter tuning, the best parameters were assigned to each classifier, resulting in more precise cardiovascular disease predictions. Comparing the experimental examination to earlier research, performance was better, with the ensemble model obtaining the greatest accuracy of 91.8%. The study emphasizes how crucial it is to use ensemble methods and sophisticated algorithm models in order to forecast cardiovascular illness properly and enhance real-world results.

CHAPTER 5

IMPACT ON SOCIETY, ENVIRONMENT, AND SUSTAINABILITY

5.1 Impact on Society

The strategy we propose has many benefits, both in terms of the economy and society. Our model is designed to explore and detect the essential features of a patient with cardiovascular disease using a real-world dataset. By providing information regarding the frequency of cardiovascular illness and suggesting preventative strategies, this project has advantages for society. By use of precise diagnosis and routine examinations, early intervention may be suggested, increasing people's awareness of possible health hazards. Our method simplifies and accurately predicts disease, as seen by its quicker processing times and lower compilation needs. Our approach employs cutting-edge diagnostic techniques and comprehensive data analysis to identify the underlying causes of cardiovascular disease. We hope that society will accept and apply this suggested course of action.

5.2 Impact on Environment

Our suggested paradigm's streamlined diagnostic processes are especially useful in distant areas. By applying our proven technique, we can significantly reduce the time and complexity required to detect cardiovascular disease. The approach is uncomplicated and has no negative effects, thus the environment gains from its simplicity. Patients may now receive evaluations for cardiovascular illness without having to fly to large cities. The patient's diagnostic report and the prediction model, which also forecasts potential outcomes, might work well together. The inexpensive cost of identifying heart disease allays worries over the cost of in-person treatments. It is simple enough for people of all skill levels to use. Our methodology helps to enhance the social and economic environments by making it clear whether a patient has cardiovascular disease. We are optimistic that our approach will constitute a major breakthrough in medical scientific technology if it is put into practice.

5.3 Ethical Aspects

It's critical to set up ethical protections prior to system deployment in order to stop sensitive data, such as diagnostic reports and personal information, from being disclosed. Future research projects and the diagnosis and treatment of cardiovascular illness may benefit from the practical uses of our suggested methodology. We recognize that the problem at hand has worldwide ramifications, impacting not just certain places or regions but the entire earth. The proposed approach enables people with cardiovascular illness or others who are aware of the problem to predict how it could influence their life.

5.4 Sustainability Plan

We are optimistic that our proposed model may be easily included into the global cardiovascular disease diagnostic and research technologies. We think that women who are at risk of cardiovascular disease would benefit most from our suggested strategy, which would help them predict how likely it is that they would have the ailment. We are motivated and prepared to offer our support to rural areas if given the resources and chances to put it into practice. We believe that the paradigm we have proposed will be both practical and long-lasting, making a significant worldwide contribution to the progress of cardiovascular disease research and diagnostics.

CHAPTER 6

SUMMARY, CONCLUSION, RECOMMENDATION, AND IMPLICATION FOR FUTURE RESEARCH

6.1 Summary of the Study

In this interesting paper, we use algorithms to assess the effect rate on people. Our methodology allows us to accurately predict the future, even though people may wrongly think they should be on the lookout for cardiovascular disease. People may determine if they are going to be impacted or not by using the diagnostic approach that our prediction system employs. Our approach makes it easier to identify different phases of cardiovascular illness quickly, and it can also be useful for diagnostic authorities. We have employed a variety of widely used algorithms that are simple to use, need minimal training, and have good accuracy.

6.2 Conclusion

The modern society we live in is open to everybody and has sophisticated technology. The recommended technology is quick and easy to use, making it accessible to a worldwide audience. We want to make the process of predicting Cardiovascular disease as easy as possible so that everyone may take advantage of our state-of-the-art algorithms. We pledge to guarantee the idea's feasibility, and in the future, we'll add more features and focus on subjects that are more commonly discussed. This objective is stated rather effectively.

6.3 Implication for Further Study

Death is a natural part of being human, and different illnesses have an impact on our day-to-day living. Cardiovascular illness is a problem that affects many people, yet new developments in treatment and diagnostic tools provide chances for recovery. Since we live in a developing nation, we have access to more sophisticated and accurate medical

technology. These modern tools have made the process of diagnosing heart disease simpler and more efficient. We hope that by attempting something different, more individuals would choose to support their community as we have. We wish to introduce additional algorithms in the future after working on a few to increase performance.

6.4 Limitations

Because we are mortal, we are susceptible to a wide range of illnesses. Many health issues affect our everyday lives, and although some of us may have access to necessary recovery tools and efficient treatments, many others still struggle with ailments like cardiovascular disease. Living in a developing nation, however, has allowed us to benefit from advances in diagnostic and treatment technology, which are always changing to become more precise and dynamic. These developments provide promise for better healthcare results and a higher standard of living for those suffering from a variety of illnesses. We want to continue working on this project if the community accepts our suggested methods. It's crucial to remember that our model is now only being used with a small dataset, thus the assessment may differ from other research approaches.

REFERENCE

[1] Dey, D., Slomka, P. J., Leeson, P., Comaniciu, D., Shrestha, S., Sengupta, P. P., & Marwick, T. H. (2019). Artificial intelligence in cardiovascular imaging: JACC state-of-the-art review. Journal of the American College of Cardiology, 73(11), 1317-1335.

[2] Attia, Z. I., Noseworthy, P. A., Lopez-Jimenez, F., Asirvatham, S. J., Deshmukh, A. J., Gersh, B. J., ... & Friedman, P. A. (2019). An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. The Lancet, 394(10201), 861-867.

[3] Krittanawong, C., Virk, H. U. H., Bangalore, S., Wang, Z., Johnson, K. W., Pinotti, R., ... & Tang, W. W. (2020). Machine learning prediction in cardiovascular diseases: a meta-analysis. Scientific reports, 10(1), 16057.

[4] Khalil, A., Fulop, T., & Berrougui, H. (2021). Role of paraoxonase1 in the regulation of high-density lipoprotein functionality and in cardiovascular protection. Antioxidants & redox signaling, 34(3), 191-200.

[5] Laad, M., Kotecha, K., Patil, K., & Pise, R. (2022). Cardiac Diagnosis with Machine Learning: A Paradigm Shift in Cardiac Care. Applied Artificial Intelligence, 36(1), 2031816.

[6] Muse, E. D., Chen, S. F., & Torkamani, A. (2021). Monogenic and polygenic models of coronary artery disease. Current cardiology reports, 23, 1-12.

[7] Jindal, H., Agrawal, S., Khera, R., Jain, R., & Nagrath, P. (2021). Heart disease prediction using machine learning algorithms. In IOP conference series: materials science and engineering (Vol. 1022, No. 1, p. 012072). IOP Publishing.

[8] Tithi, S. R., Aktar, A., & Aleem, F. (2018). Machine Learning Approach for ECG Analysis and predicting different heart diseases (Doctoral dissertation, BRAC University).

[9] 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.

[10] Khanna, D., Sahu, R., Baths, V., & Deshpande, B. (2015). Comparative study of classification techniques (SVM, logistic regression and neural networks) to predict the prevalence of heart disease. International Journal of Machine Learning and Computing, 5(5), 414.

[11] Chang, V., Bhavani, V. R., Xu, A. Q., & Hossain, M. A. (2022). An artificial intelligence model for heart disease detection using machine learning algorithms. Healthcare Analytics, 2, 100016.

[12] Ansari, M. F., Alankar, B., & Kaur, H. (2021). A prediction of heart disease using machine learning algorithms. In Image Processing and Capsule Networks: ICIPCN 2020 (pp. 497-504). Springer International Publishing.

[13] Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. Neurocomputing, 415, 295-316.

[14] Williams, R., Shongwe, T., Hasan, A. N., & Rameshar, V. (2021, October). Heart disease prediction using machine learning techniques. In 2021 international conference on data analytics for business and industry (ICDABI) (pp. 118-123). IEEE.

[15] Nagavelli, U., Samanta, D., & Chakraborty, P. (2022). Machine learning technology-based heart disease detection models. Journal of Healthcare Engineering, 2022.

[16] B. Akkaya, E. Sener and C. Gursu, "A Comparative Study of Heart Disease Prediction Using Machine Learning Techniques," 2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), Ankara, Turkey, 2022, pp. 1-8, doi: 10.1109/HORA55278.2022.9799978.

[17] Basha, N., PS, A. K., & Venkatesh, P. (2019, December). Early detection of heart syndrome using machine learning technique. In 2019 4th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT) (pp. 387-391). IEEE.

[18] Bhowmick, A., Mahato, K. D., Azad, C., & Kumar, U. (2022, June). Heart Disease Prediction Using Different Machine Learning Algorithms. In 2022 IEEE World Conference on Applied Intelligence and Computing (AIC) (pp. 60-65). IEEE.

[19] Karthick, K., Aruna, S. K., Samikannu, R., Kuppusamy, R., Teekaraman, Y., & Thelkar, A. R. (2022). Implementation of a heart disease risk prediction model using machine learning. Computational and Mathematical Methods in Medicine, 2022.

[20] Rajib Kumar Halder. (2020). Cardiovascular Disease Dataset. IEEE Dataport. https://dx.doi.org/10.21227/7qm5-dz13.

[21] Lahoura, V., Singh, H., Aggarwal, A., Sharma, B., Mohammed, M. A., Damaševičius, R., ... & Cengiz, K. (2021). Cloud computing-based framework for breast cancer diagnosis using extreme learning machine. Diagnostics, 11(2), 241.

[22] Gladence, L. M., Karthi, M., & Anu, V. M. (2015). A statistical comparison of logistic regression and different Bayes classification methods for machine learning. ARPN Journal of Engineering and Applied Sciences, 10(14), 5947-5953.

[23] "Logistic Regression for Machine Learning", Accessed: August 6, 2021, Available: https://www.capitalone.com/tech/machine-learning/what-is-logistic-regression/

[24] Ghosh, P., Karim, A., Atik, S. T., Afrin, S., & Saifuzzaman, M. (2021). Expert cancer model using supervised algorithms with a LASSO selection approach. International Journal of Electrical and Computer Engineering (IJECE), 11(3), 2631.

[25] Nahar, N., & Ara, F. (2018). Liver disease prediction by using different decision tree techniques. International Journal of Data Mining & Knowledge Management Process, 8(2), 01-09.

[26] Aljahdali, S., & Hussain, S. N. (2013). Comparative prediction performance with support vector machine and random forest classification techniques. International journal of computer applications, 69(11).

[27] Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. Artificial Intelligence Review, 54, 1937-1967.

[28] Drucker, H., Cortes, C., Jackel, L. D., LeCun, Y., & Vapnik, V. (1994). Boosting and other ensemble methods. Neural computation, 6(6), 1289-1301.

[29] Pasha, M., & Fatima, M. (2017). Comparative Analysis of Meta Learning Algorithms for Liver Disease Detection. J. Softw., 12(12), 923-933.

[30] Wang, Y., Jha, S., & Chaudhuri, K. (2018, July). Analyzing the robustness of nearest neighbors to adversarial examples. In International Conference on Machine Learning (pp. 5133-5142). PMLR.

[31] Sharma, A., & Suryawanshi, A. (2016). A novel method for detecting spam email using KNN classification with spearman correlation as distance measure. International Journal of Computer Applications, 136(6), 28-35.

[32] Zhou, Z. H. (2012). Ensemble methods: foundations and algorithms. CRC press.

[33] Drucker, H., Cortes, C., Jackel, L. D., LeCun, Y., & Vapnik, V. (1994). Boosting and other ensemble methods. Neural computation, 6(6), 1289-1301.

[34] Lemmens, A., & Croux, C. (2005, January). Bagging and boosting classification trees to predict churn. Insight from the US Telecom industry. In Marketing Science Conference (INFORMS2005).

[35] Islam, R., Beeravolu, A. R., Islam, M. A. H., Karim, A., Azam, S., & Mukti, S. A. (2021, December). A Performance Based Study on Deep Learning Algorithms in the Efficient Prediction of Heart Disease. In 2021 2nd International Informatics and Software Engineering Conference (IISEC) (pp. 1-6). IEEE.

[36] Tajmen, S., Karim, A., Hasan Mridul, A., Azam, S., Ghosh, P., Dhaly, A. A., & Hossain, M. N. (2022, July). A machine learning based proposition for automated and methodical prediction of liver disease. In Proceedings of the 10th International Conference on Computer and Communications Management (pp. 46-53).

[37] "What is Correlation in Machine Learning?", Accessed: August 6, 2020, Available: https://medium.com/analytics-vidhya/what-is-correlation-4fe0c6fbed47.

Cardiovascular Disease

ORIGINALITY REPORT

10%	8%	2%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	4%
2	Submitted to Daffodil International University Student Paper	2%
3	dokumen.pub Internet Source	<1%
4	Submitted to Kaplan Professional Student Paper	<1%
5	Laio Oriel Seman, Stefano Frizzo Stefenon, Viviana Cocco Mariani, Leandro dos Santos Coelho. "Ensemble learning methods using the Hodrick–Prescott filter for fault forecasting in insulators of the electrical power grids", International Journal of Electrical Power & Energy Systems, 2023 Publication	<1%
6	Submitted to Monash University Student Paper	<1%
7	Saul Beltozar-Clemente, Orlando Iparraguirre-Villanueva, Félix Pucuhuayla-Revatta, Joselyn	<1%