COVENTRY
UNIVERSITY

Faculty of Engineering, Environment and Computing

School of Computing, Mathematics and Data Science

MSc. Data Science

7150CEM Data Science Project

Predictive Modeling for Early Detection of Cardiovascular Heart Disease Using Machine Learning

Author: Obianuju Okafor

SID: 14281642

Supervisor: Manizheh Montazerian

Submitted in partial fulfilment of the requirements for the Degree of Master of Science in Data Science

Academic Year: 2023/24

Declaration of Originality

I declare that this project is all my own work and has not been copied in part or in whole from any other source except where duly acknowledged. As such, all use of previously published work (from books, journals, magazines, internet etc.) has been acknowledged by citation within the main report to an item in the References or Bibliography lists. I also agree that an electronic copy of this project may be stored and used for the purposes of plagiarism prevention and detection.

Statement of copyright

I acknowledge that the copyright of this project report, and any product developed as part of the project, belong to Coventry University. Support, including funding, is available to commercialise products and services developed by staff and students. Any revenue that is generated is split with the inventor/s of the product or service. For further information please see www.coventry.ac.uk/ipr or contact ipr@coventry.ac.uk.

Statement of ethical engagement

I declare that a proposal for this project has been submitted to the Coventry University ethics monitoring website (https://ethics.coventry.ac.uk/) and that the application number is listed below (Note: Projects without an ethical application number will be rejected for marking)

Signed: Date: 31/07/2024

Please complete all fields.

First Name:	Obianuju
Last Name:	Okafor
Student ID number	14281642
Ethics Application Number	P177676
1st Supervisor Name	Manizheh Montazerian
2nd Supervisor Name	Darren Imrie

This form must be completed, scanned and included with your project submission to Turnitin. Failure to append these declarations may result in your project being rejected for marking.

Abstract

The research explores the application of machine learning in predicting cardiovascular disease (CVD) by utilizing the heart disease dataset from Mendeley. The study's primary objective is to identify the most effective machine-learning model for early CVD prediction. The research also investigates the impact of outliers on model performance by conducting two experiments: one with outliers included and another with outliers removed. The study employs various machine learning algorithms, including K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest, Naive Bayes, and ensemble learning methods. The performance of these models is evaluated using metrics such as accuracy, precision, recall, and F1-score. The research findings indicate that the SVM model, despite performing poorly in the first experiment (with outliers removed), exhibits significantly better performance in the second experiment (with outliers included). The study concludes that while removing outliers can yield fair performance, working with the complete dataset, including outliers, leads to superior outcomes in CVD prediction. The research emphasizes the importance of careful data preprocessing and the selection of appropriate machine learning models for accurate and effective CVD prediction, ultimately contributing to the advancement of early CVD detection.

Table of Contents

Acknowledgements

I would like to thank my supervisor Manizheh Montazerian for her support and guidance during the duration of this research. She was always available to provide feedback and review my work. Her support contributed to the success of this research.

Also, I would like to thank my family and friends for their support and for always motivating me. You are highly appreciated.

1 Introduction

1.1 Background to the Project

The primary cause of death worldwide is cardiovascular illnesses or CVDs. An estimated 17.9 million deaths worldwide in 2019 were attributed to CVDs, accounting for 32% of all deaths. 85% of these fatalities were brought on by heart attacks and strokes. Three-quarters of the 17 million premature fatalities (dead before age 70) attributed to noncommunicable illnesses in 2019 (CVDs) were caused by these disorders.

Including strategies for managing cardiovascular disease in comprehensive health care plans for all is essential to reducing the incidence of cardiovascular disease. Early diagnosis and treatment with medication and therapy are crucial for the effective management of cardiovascular disease (World Health Organization: WHO, 2021).

Medical diagnosis is a classification task in which a physician attempts to determine an issue by examining the values of several attributes. Generally, this type of task is based on multiple algorithms.

Conventional methods like ECG, occultation, and blood pressure measurements are expensive and time-consuming. Machine learning algorithms offer a convenient solution for diagnosing cardiovascular disease by cutting processing time and improving prediction accuracy (Louridi et al., 2019).

Machine Learning is a method of interpreting datasets using computers that learn from experience. Alan Turing's proposal for the first artificially intelligent machine in the 1950s marked the beginning of machine learning applications (Habehh & Gohel, 2021). Machine learning has been utilised for several purposes since its inception, including face detection for security purposes, enhancing productivity and lowering danger in public transit, and more recently, several applications in biotechnology and healthcare. Healthcare and medicine are expected to undergo similar changes as artificial intelligence and machine learning has significantly altered daily life and business procedures. Machine Learning for health informatics has become a multidisciplinary field of study that applies sophisticated computational methods to handle healthcare data. (Habehh & Gohel, 2021, Khan et al., 2019).

To solve issues without the need for specialised programming, machine learning uses a variety of statistical techniques and algorithmic models. A significant portion of feature extraction and data processing is completed before the data is entered into the algorithm since many machine learning models are single-layered. To prevent over- or underfitting of the training dataset and to make correct predictions, these machine learning algorithms would need extensive data preparation if they didn't have the additional layers (Habehh & Gohel, 2021).

1.2 Project Aim

Early identification of cardiovascular disease requires an accurate, effective, and non-invasive predictive model to prevent high morbidity and mortality rates associated with the condition. The research aims to determine the most effective machine-learning model for early cardiovascular disease prediction or identifying individuals at risk after conducting in-depth exploratory analysis and removing and handling patterns or outliers that may be found during this process.

1.3 Project Objectives

The objectives of this study are:

1. To develop a predictive model with the Heart Disease Dataset from the Mendeley website that can precisely identify those at risk of cardiovascular heart disease.
2. To check and contrast how well various machine learning techniques predict CVD.
3. To use cross-validation methods to ensure the model can be used in a wide range of situations.
4. To optimise the high-performing model’s performance through hyperparameter tunning.

1.4 Research Questions

1. How effective is using machine learning techniques in predicting early detection of CVD?
2. Which machine learning model gives the best accuracy in predicting CVD?

1.5 Contribution

1. To demonstrate the significance of early identification of cardiovascular disease (CVD) and develop a reliable predictive model for CVD diagnosis using machine learning.
2. To analyse and compare in depth the accuracy of several models including a hybrid method of combining some of these models.
3. To implement data cleaning through exploratory analysis and data engineering methods.

1.6 Overview of This Report

This report will be arranged in different sections as follows:

Chapter 1 covers the background of the study, report aim, objectives and contributions.

Chapter 2 covers a review of different literature.

Chapter 3 covers the proposed methods applied to the course of the study.

Chapter 4 covers the experimental analysis, results of this research and a comparative analysis.

Chapter 5 covers project schedule, risk management, quality management and legal, social and ethical considerations.

Chapter 6 covers critical analysis

Chapter 7 discusses the research report, conclusions, limitations and future works.

2 Literature Review

2.1 Cardiovascular Disease

Cardiovascular disease is a general term that describes a disease of the heart or blood vessels. It's typically linked to atherosclerosis, or the buildup of fatty deposits inside the arteries, and a higher risk of blood clots. It may also be linked to artery damage in the kidneys, eyes, heart, brain, and heart. Cardiovascular diseases (CVDs) are a group of disorders of the heart and blood vessels.

According to WHO, there are different types of heart diseases included in the table below:

Coronary heart disease	A disease of the blood vessels supplying the heart muscle.
Cerebrovascular disease	A disease of the blood vessels supplying the brain.
Peripheral arterial disease	A disease of blood vessels supplying the arms and legs.
Rheumatic heart disease	Streptococcal bacteria, the cause of rheumatic fever, damage the heart muscle and heart valves.
Congenital heart disease	Anomalies that stem from structural cardiac defects at birth that hinder the heart's capacity to grow and function normally.
Deep vein thrombosis and pulmonary embolism	Blood clots in the veins of the legs have the potential to migrate to the heart and lungs.

Table 1. Types of cardiovascular diseases.

A blockage that stops blood from reaching the heart or brain is the primary cause of heart attacks and strokes, which are typically acute events. The most frequent cause of this is an accumulation of fat deposits on the inside walls of the blood arteries supplying the brain and heart. Bleeding from brain vessels or blood clots are two possible causes of strokes (World Health Organization: WHO, 2021).

According to research by Nagavelli et al. (2022) Cardiovascular heart disease accounts for over 30% of deaths worldwide which makes it a major cause of death in the world and if neglected death rate is predicted to increase by 22 million by 2030. The paper further discusses some causes of CVD which are physical inactivity, unhealthy diet and excessive use of alcohol and tobacco explaining that these factors can be reduced by adopting regular exercise, eating healthy by introducing healthy eating habits, reducing salt intake, eating lots of fruits and vegetables and quitting alcohol and tobacco use, all of these can help manage the risk of heart disease.

2.2 Importance of Early Detection

Due to several contributing risk factors, including diabetes, high blood pressure, excessive cholesterol, an irregular pulse rate, and many more, it is challenging to diagnose heart disease. Because heart illness has a complex nature, it needs to be treated properly. Failure to do so could damage the heart or result in an early death (Mohan et al., 2019). Thus, in healthcare, particularly for CVDs, timely and accurate detection of diseases and determining the vital risk factors are critically important.

Ahdal et al. (2023) research on monitoring cardiovascular problems in heart patients using machine learning also shows that unhealthy diets, excessive alcohol and tobacco intake and irregular exercise are risk factors for CVD, it also shows that heart diseases are one of the major causes of death globally. The researcher further discussed how monitoring CVD problems in heart patients can help manage these deaths while proposing the use of machine learning models to detect these diseases.

Heart disease is a time-sensitive condition and therefore early detection is very important. Many patients' health can be endangered as a result of incorrect diagnoses or trial-and-error processes. Until there are more medical professionals and experts or diagnostic errors are eliminated, these problems will persist. Some patients may receive unnecessary therapy or be admitted to the hospital for chest discomfort. Many undeveloped nations do not have enough specialists. As such an automated system like machine learning technologies, can be of benefit to the medical community in assisting doctors with more accurate and early diagnoses (Dalal et al., 2023).

2.3 Machine Learning in Healthcare

The most important aspect of a person's life is their health. A medical professional gathers clinical information on each patient throughout the healthcare process to diagnose the condition and decide how best to treat it. To treat the health issue, clinical data is essential, and machine learning algorithms can aid in the diagnosis of the illness. To diagnose diseases, provide decision support, provide prognoses, and create the best possible treatment plan for those diseases, machine learning is essential to the healthcare process (Nayyar et al., 2021).

The healthcare industry is increasingly utilizing machine learning to improve decision-making. With over 80% of leaders having AI plans and 86% using ML solutions, machine learning algorithms learn from data and apply it to new data. This approach helps address human errors due to short-term memory limitations. The growing data volume and computer processing capacity have led to the development of various machine learning algorithms, benefiting clinical decision support in fields like clinical medicine, medical imaging, and heart failure management (Alanazi, 2022).

Machine Learning techniques, as opposed to naive Machine Learning and regression methods, are dependable and effective for predicting cardiovascular disease. Over the past ten years, several machine-learning methods have been put forth to predict cardiovascular disease using various parameters, datasets, and methodologies (Taylan et al., 2023). Different machine-learning techniques have been proposed for the prediction of cardiovascular diseases which will be discussed further in this paper.

Javaid et al. (2022) research paper ‘Significance of machine learning in healthcare: Features, pillars and applications’, explained how machine learning works, its applications and features and how it can be of benefit in the healthcare system, it discussed how machine learning as a branch of artificial intelligence can enhance the speed and accuracy of a healthcare practitioner's work. Machine learning can be used to analyse huge amounts of data, detect patterns and trends. As healthcare continues to change as a result of the development of new technology and ideas, machine learning can help medical professionals and healthcare practitioners in some of these scenarios allowing these professionals and healthcare practitioners to make better decisions about patient care.

A large data report and clinical diagnosis of the patient’s cure and treatment can be highly difficult to set up accurately; otherwise, these data can be affected due to insufficient storage or management. Such amount of data requires special means of extraction and processing efficiently, using one of the machine learning applications such as a classifier which can divide the data according to their features; this can be used in medical data analysis or disease identification. In the past years, machine learning has proven to be efficient in making significant developments in disease diagnosis, hospitals have recorded the efficiency of machine learning technologies. Today machine learning can be used several times including in the medical field as it is used in different disciplines from identifying diseases to medical imaging, drug production, predicting diseases and so on (Ahdal et al., 2023).

Two types of machine learning that can play a major role in the healthcare system are supervised learning and unsupervised learning. Supervised learning uses labelled data to train models for tasks such as prediction and classification which has proved successful in fields such as disease risk prediction and medical imaging analysis. Unsupervised learning uses unlabelled data to uncover patterns and relationships and can be used for tasks such as patient grouping, anomaly detection, and feature extraction. As healthcare data grows, machine learning is expected to play an important role in enhancing patient outcomes and advancing medical research by enabling more exact diagnoses, identifying patterns in patient data, streamlining administrative processes, and personalised treatment recommendations (An et al., 2023).

2.4 Overview of Existing Predictive Models

Experts and researchers have conducted several research projects to predict and screen medical data for CVD. Several predictions have been executed in previous studies using different machine learning algorithms of which an overview of different papers has been given below.

In their review, Marimuthu et al. (2018) talked about heart disease and related research findings on the use of machine learning to predict heart disease using data analytics and machine learning approaches. Additionally, they conducted a study on some research publications on the prediction of heart disease and the methodology used. They continued by suggesting the application of more pertinent feature selection techniques to enhance the accuracy of these algorithms' performance and the necessity of combinational and more complicated models to boost the precision of heart disease early detection prediction.

The Khan et al. (2019) survey reviewed about 35 journal articles where machine learning techniques were implemented, the study analysed these journals to get the most used classification techniques and these techniques are Support Vector Machine, Neural Networks, and ensemble classifiers.

Tiwari et al. (2023) using the dataset from the UCI repository used a support vector machine, k-nearest neighbour, and random forest machine learning algorithms to predict early heart disease detection using a reduced number of attributes and majorly focusing on random forest classifier. With 13 attributes random forest entropy and random forest Gini gave 98.49%, these techniques were tested using 10-fold cross-validation.

Pal et al. (2022) research titled “Risk Prediction of Cardiovascular Disease Using Machine Learning Classifier,” suggested that early and automated detection of cardiovascular disease can help reduce the death rate caused by this disease, they proposed two machine learning techniques which multilayer perceptron and k-nearest neighbour to be used in predicting CVD using the UCI dataset. The models gave an accuracy score of 82.47% and 73.77% and an AUC score of 86.41% and 86.21% respectively.

Mohan et al. (2019) proposed hybrid machine-learning techniques for effective heart disease predictions, they used several machine-learning techniques and a hybrid HRFLM model. The proposed hybrid HRFLM model is used combining the characteristics of Random Forest (RF) and Linear Method (LM) and this model produced an accuracy score of 88.4%.

In the prediction of coronary heart disease, an experimental analysis Gonsalves et al. (2019) proposed three machine learning algorithms Support Vector Machine, Naive Bayes and Decision Tree using the South African Heart Disease dataset, the various tasks done to conduct this experiment ranged from pre-processing, classification, regression, feature selection, clustering, association, and visualization. The 10-fold cross-validation method was used to test these machine learning algorithms. All models had an accuracy score of over 70%, with Naive Bayes having the highest at 71.5%. However, it didn't meet the 80% threshold. The study also evaluated sensitivity and specificity scores, with Naive Bayes having the lowest specificity score. The paper was limited to three machine learning algorithms, further research is needed to improve the dataset's performance, including class balancing and exploring other algorithms.

Gulati et al. (2022) applied different methods in the prediction of coronary heart disease using the UCI dataset and had an accuracy score of 86.81% when Naive Bayes was applied, other methods like Multilayer Perceptron, K-star, Decision Tree, J48, Linear regression and they all passed the accuracy threshold score of 80% however J48 had an accuracy score of less than 80% the experiment was achieved by using the following method, inspecting the dataset, pre-processing the data and applying feature selection and then executing the proposed models

To improve the accuracy of the prediction of heart disease risk, an ensemble classification technique was applied as against using one method a combination of different methods was applied. In this project, the ensemble algorithms bagging, boosting, stacking and majority voting were employed for this experiment after applying different methods to check for their accuracy score, a combination of each of them was applied and the proposed model gave an accuracy score of 85.48% using majority voting with the combination Naive Bayes, Bayes Net, Random Forest and Multilayered Perceptron. This was done using the UCI dataset (Latha & Jeeva, 2019).

Sharma and Parmar (2020) use a Deep learning neural network model for predicting heart disease and compare it to some classification algorithms that were also applied in this experiment, these classification algorithms include logistics regression, k-nearest neighbour, support vector machine, Naive bayes and random forest, while they all had high accuracy score of above 80% the deep learning neural network applied had the highest accuracy score of 90.78%, the method applied is the hyper-parameter optimization (Talos).

Bhatt et al. (2023) recommended three models decision tree, multilayer perceptron and XGBoost, using GridSearchCV to hypertune the parameters of these models to optimize the results. The recommended models were used on a dataset from Kaggle. The models' accuracy was as follows: Cross-validation results and without cross-validation, for decision trees are 86.37% and 86.53%, respectively; for XGBoost, they are 86.87% and 87.02%, respectively; for random forests, they are 87.05% and 86.92%, respectively; and for multilayer perceptron, they are 87.28% and 86.94%, respectively. AUC (area under the curve) values for the recommended models are as follows: decision tree: 0.94, XGBoost: 0.95, random forest: 0.95, and multilayer perceptron: 0.95. Based on this research, it was concluded that the accuracy of the multilayer perceptron with cross-validation surpassed that of all previous techniques. At 87.28%, it had the highest accuracy.

Alotaibi (2019) aimed to determine if machine learning (ML) techniques may be applied when predicting the onset of heart failure. The study developed prediction models using a dataset from the Cleveland Clinic Foundation and a variety of machine learning (ML) techniques, including decision trees, logistic regression, random forests, naive Bayes, and support vector machines (SVM). During the model-development process, a 10-fold cross-validation strategy was used. The SVM method came in second at 92.30%, and the decision tree method attained the best accuracy of 93.19% in predicting heart disease. This work highlights the decision tree algorithm as a viable option for future research and sheds light on the potential of machine learning techniques as a useful tool for heart failure illness prediction.

In Louridi et al. (2019) approach to identifying cardiovascular disease using machine learning using the UCI dataset, the study used the mean value to replace missing data in the preprocessing stage to check if there will be an increase in accuracy compared to some other works. The techniques applied in this paper are SVM, Naive Bayes and KNN, there was an increase in accuracy following this approach however SVM had the highest accuracy score of 86.8%. This shows that this method of using mean values at the preprocessing stage can help increase accuracy.

Yılmaz and Yağın (2022) study, where they compared the performance of three machine learning techniques namely support vector machine, random forest and logistic regression using their accuracy, specificity, F1, and sensitivity score. The experiment was done using a dataset from the IEEE database, SVM=SMOTE method was first applied to the dataset, then data preprocessing was carried out on the dataset and hyper-parameter tunning was carried out also on each model to increase accuracy and 10-fold cross-validation. At the end of this experiment, random forest had the highest accuracy, specificity, sensitivity and f1 score. This shows that hyperparameter tuning can help increase accuracy.

Machine learning excels in processing extensive amounts of health records because of its ability to handle highly dimensional data. Some benefits include pattern detection that identifies subtle patterns and correlations in massive datasets that basic analytics might elude. Machine learning quickens the study of large datasets, decreasing the time and resources needed to derive actionable insights (Deepa et al., 2024).

ML models like decision trees, random forest, SVM, Naive Bayes, and neural networks demonstrate the broad application of machine learning in CVD prediction, with each technique providing unique advantages in dealing with diverse data types and issues of CVD risk assessment. Advanced techniques like natural language processing (NLP), ensemble methods and transfer learning enhance the predictive capabilities of models for CVD and contribute to precise and adaptive CVD risk prediction models (Deepa et al., 2024).

Baghdadi et al. (2023) proposed new robust, effective, and efficient machine learning algorithms for predicting CVD based on symptoms, signs, and other patients' information from hospital records with the aim of improving the early prediction of cardiovascular disease. The author conducted exploratory data analysis on the dataset used which is privately sourced with 918 observations and 12 columns, the author proposed advanced machine-learning techniques like XGBoost, Adaboost, Linear Discriminant, LightGBM, Gradient Boosting, Catboost, ExtraTree, K-Nearest Neighbors, SVM, Logistic Regression, Random Forest, with all of these techniques having above 80% accuracy and below 90%, after tuning the Catboost model the author had a high accuracy of 90% and an F1 score of 92%.

In this paper, Biswas et al. (2023) proposed applying different feature selection techniques is the UCI Cleveland dataset to get the best accuracy score from different machine learning models used to predict heart disease in its early stages. These feature selection methods include ANOVA *F* Value, Chi-Square and Mutual Information and the tested models are Logistic regression, Support Vector Machine, KNN, Random forest, Naive Bayes and Decision Tree. In selecting significant features the author was able to get a high accuracy score of 94.51% using support vector machine and linear regression models.

Ahmed et al. (2023) publication used hybrid machine learning approach in improving heart disease predictions. By utilizing the heart disease dataset from the UCI machine learning repository, the author did a comparative study between support vector machine and K-nearest neighbour algorithms. After preprocessing the data the author proceeded to feature extraction, trained and tested the data and then applied KNN, SVM and a hybrid of both for classification. With this approach, the author had an accuracy score of 81% while the accuracy score for these models individually was lower than 80%.

To predict CVD the effectiveness of cardiovascular disease identification, utilizing the ECG dataset was evaluated for this experiment by (Pandey, 2023). The author explored the use of support vector machines and had a high sensitivity score of 97%. While the accuracy score is unknown and the sensitivity score is high, there is room to use other machine-learning models and select one with the best accuracy.

A study by Hridoy et al. (2023) to predict heart disease using machine learning algorithms examines the accuracy score of different machine learning techniques which includes Naive Bayes, Random Forests, Decision Trees, Logistic Regression, Support Vector Machines, K-Nearest Neighbor and Gradient Boosting. The author pre-processed the data obtained for this experiment including the detection of outliers, feature encoding and normalization, the author went on to use the techniques mentioned above to classify the data. These models had good accuracy scores, precision, recalls and f1 scores however random forest had the highest score of 86.81% and gradient boosting had the lowest 79.12%.

Allheeib et al. (2023) proposed a hybrid method of combining machine learning and deep learning algorithms with artificial intelligence networks-based feature selection techniques to predict heart disease, this proposed framework yielded an accuracy of 98.99% with the application of support vector machine and linear regression classifier.

Jawalkar et al. (2023) proposed training a heart disease dataset with a decision tree-based random forest (DTRF) classifier an ensemble learning method using a stochastic gradient boosting loss optimization technique. For this approach, the researchers gathered a heart disease dataset containing patient records and pre-processed it to remove missing values and duplicate records and then went ahead to encode categorical variables using one-hot encoding and scaling numerical variables before training this dataset using the proposed classifier while the process of stochastic gradient boosting optimization moves the model closer to a more accurate prediction, repeating the process focuses on enhancing the model's predictions based on the errors made in the previous iteration. This approach yielded an accuracy score of 96% using the Cleveland dataset.

In a research study, Chandrasekhar and Peddakrishna (2023) experimented to examine and analyze six major machine learning algorithms' performance on the Cleveland and IEEE Dataport heart disease datasets. These algorithms include random forest, K-nearest neighbour, logistic regression, Naive Bayes, gradient boosting and Adaboost, these techniques were trained individually, hyperparameter tuning was also employed to improve the accuracy of these algorithms, and finally, all algorithms were combined to increase the accuracy of the model using soft voting ensemble method. The author achieved a score of 93.44% and 95% on the Cleveland and IEEE datasets respectively.

Adhishayaa et al. (2023) in the review of cardiovascular disease using machine learning algorithms discussed different algorithms and how they have performed during training based on their accuracy score. The review also employed that based on the various algorithms and how they are trained, it is justifiable to conclude that machine learning algorithms have a large potential for diagnosing CVD. However, the researcher states that machine learning-based systems and techniques have been pretty accurate in heart disease predictions, more research can also be done on the best ensemble of techniques to use for a specific dataset.

In the paper 'Machine learning-based predictive models for detection of cardiovascular diseases', Ogunpola et al. (2024) deployed seven machine learning and deep learning classifiers to improve the accuracy of heart disease predictions. The researchers combined two heart disease datasets from Mendeley and Kaggle. The study aims to see how effective fine-tuning can be and to enhance the accuracy score of these proposed models through hyperparameter tuning. After optimisation, this approach yielded an accuracy score of 98% on the XGBoost model. The models trained in this study are KNN, SVM, Logistic regression, convolutional neural network (CNN), gradient boost, XGBoost and random forest.

The table below shows a comparative study of all the algorithms in the literature review.

Author	Dataset	Techniques Applied	Best Model
Khan et al. (2019)	-	-	SVM, Neural Networks and Ensemble Classifiers
Tiwari et al. (2023)	UCI	SVM, KNN, Random Forest	Random Forest 98.49
Pal et al. (2022)	UCI	Multilayer perceptron, KNN	Multilayer perceptron 82.47%
Mohan et al. (2019b)	UCI	SVM, DT, NB, RF, HRFLM	HRFLM 88.4%
Gonsalves et al. (2019)	South African Heart Disease Dataset	SVM, Naive Bayes, Decision Tree	Naive Bayes 71.5%
Gulati et al. (2022)	UCI	Multilayer perceptron, K-star, DT, J48, Linear regression	Naive Bayes 86.81%
Latha & Jeeva. (2019)	UCI	Ensemble classification	85.48%
Sharma & Parmar. (2020)	UCI	Logistic regression, KNN, SVM, NB, RF, deep learning neural network	Deep learning neural network 90.78%
Bhatt et al. (2023)	Kaggle	Multilayer perceptron, XGBoost	Multilayer perceptron 87.28%
Alotaibi (2019)	Cleveland Clinic Foundation	Logistic regression, SVM, NB, RF, decision tree	Decision tree 93.19% SVM 92.30%
Louridi et al. (2019)	UCI	SVM, NB, KNN	SVM 86.8%

Yılmaz and Yağın (2022)	IEEE Database	SVM, RF, Logistic regression	Random Forest(RF) 92.9%
Baghdadi et al. (2023)	Privately sourced	Advanced techniques, KNN, Catboost, SVM, Logistic regression, RF	Tunned Catboost 90%
Biswas et al. (2023)	UCI	Logistic regression, SVM, KNN, RF, NB, DT	SVM, Linear regression 94.51%
Ahmed et al. (2023)	UCI	KNN, SVM, Hybrid of KNN and SVM	Hybrid method 81%
Pandey (2023)	ECG	SVM	Nil, (sensitivity score 97%)
Hridoy et al. (2023)		NB, RF, DT, SVM, KNN, Gradient boost	RF 86.81%
Allheeib et al. (2023)		Hybrid method of ML and deep learning algorithms	SVM and Linear regression classifiers 98.99%
Chandrasekhar and Peddakrishna (2023)	Cleveland and IEEE individually	KNN, Logistic regression, NB, Gradient boost, Adaboost, ensemble method combining all six algorithms	Tunned Ensemble Method 93.44% and 95% on the datasets respectively
Jawalkar et al. (2023)	Cleveland	DTRF	96% accuracy score
Ogunpola et al. (2024)	Mendeley and Kaggle	KNN, SVM, Logistic regression, CNN, gradient boost, XGBoost, RF	XGBoost Tuned 98.50%

Table 2. Comparative study of the different models in the literature review

Every researcher in the literature has proposed different models and methods for the detection of CVD and there are many more actively working to enhance detection. However, this paper aims to employ in-depth exploratory data analysis and employ some of the models used in these papers while tuning the best model to enhance its accuracy.

3 Methodology

This study is quantitative due to numerical data being analysed to make useful predictions from the dataset. The research design is experimental, and it will be done using machine learning. Figure 2 is a pictorial representation of the proposed methodology.

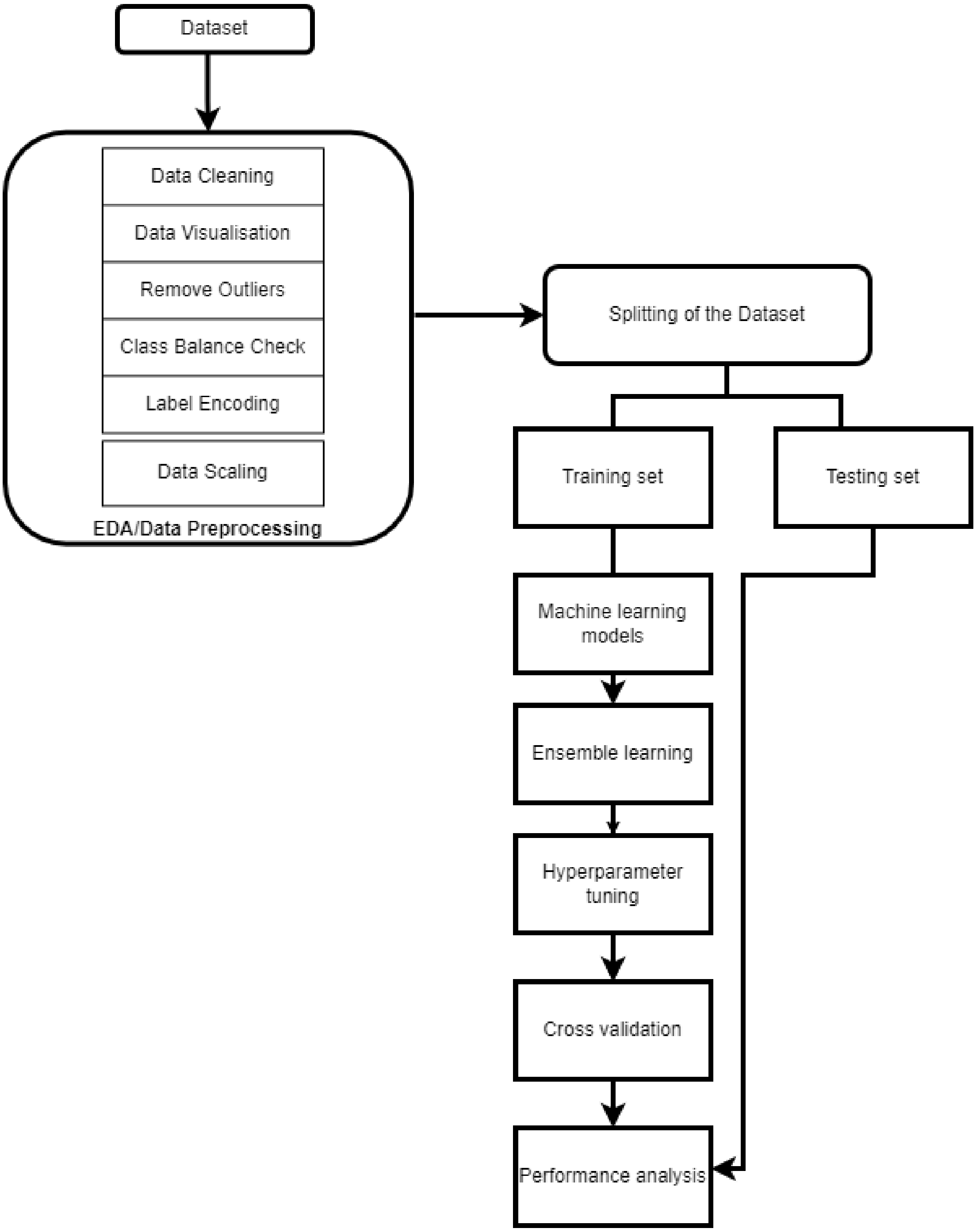


Fig 1. Experiment workflow of the proposed methodology

The first step is to import the data that will be used for the experiment and then in-depth exploratory data analysis is conducted to visualize the data, check for trends or patterns, and check for missing values or outliers. There were no missing values however outliers were detected and removed as the experiment is to see how each model will perform without the outliers. An in-depth analysis was conducted on both the features and target variables. Feature correlation was done to see if any features were highly correlated with each other and then categorical variables were converted to numerical variables using label encoding. Data scaling is also performed to normalise the range of all the variables of the data. All of this is done before the training and performance evaluation.

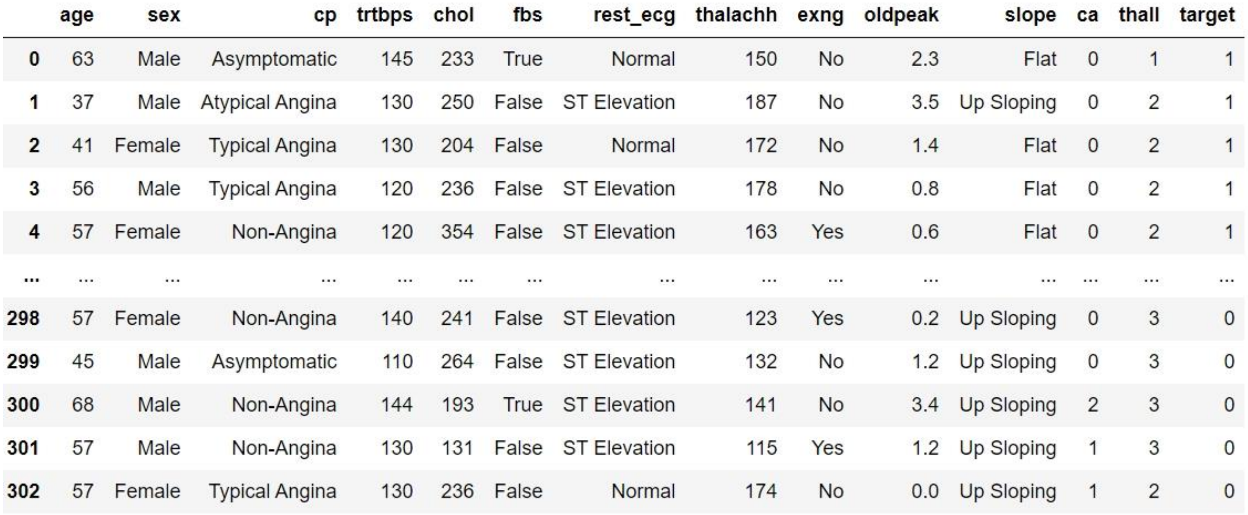
3.1 Dataset and Collection

The data used for this research is secondary data and can be obtained from the Mendeley website at this link https://data.mendeley.com/datasets/yrwd336rkz/1.(Aftab, 2024). The dataset includes 303 instances and 14 variables with one as the dependent variable which is the target variable. The variables have six categorical variables and eight numerical variables and the dependent variable contains two classes zero and one (0 means no heart disease, 1 means presence of heart disease). Below is a table that describes the variables contained in this dataset:

Variable	**Description**
Age	Age of the patient
Sex	Indicates the gender of the patient whether male or female
TRTBPS	Indicates the resting blood pressure
CP	Describe the type of chest pain if it is typical angina, atypical angina, non-angina or asymptomatic
Chol	Measures the serum cholesterol in mg/dl
FBS	Fasting blood sugar. 1 indicates FBS>120mg/dl and 0 otherwise
Rest ECG	Resting electrocardiographic results showing the results of the patients into normal, ST elevation or left ventricular hypertrophy
Thalachh	Maximum heart rate achieved by a patient
Exng	Exercise-induced Angina where 1 indicates true and 0 false

Oldpeak	Slope of the peak exercise (Flat, Up Sloping, down sloping)
CA	Number of major vessels coloured by fluoroscopy ranging from 0 to 3
Thall	Indicates types of thalassemia where 3 means normal, 6 means fixed defect and 7 means reversible defect
Target	Indicating the presence of heart disease 0 means false and 1 means true

Table 3. A summary table of the variables described



	age	sex	cp	trtbps	chol	fbs	rest_ecg	thalachh	exng	oldpeak	slope	ca	thall	target
0	63	Male	Asymptomatic	145	233	True	Normal	150	No	2.3	Flat	0	1	1
1	37	Male	Atypical Angina	130	250	False	ST Elevation	187	No	3.5	Up Sloping	0	2	1
2	41	Female	Typical Angina	130	204	False	Normal	172	No	1.4	Flat	0	2	1
3	56	Male	Typical Angina	120	236	False	ST Elevation	178	No	0.8	Flat	0	2	1
4	57	Female	Non-Angina	120	354	False	ST Elevation	163	Yes	0.6	Flat	0	2	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
298	57	Female	Non-Angina	140	241	False	ST Elevation	123	Yes	0.2	Up Sloping	0	3	0
299	45	Male	Asymptomatic	110	264	False	ST Elevation	132	No	1.2	Up Sloping	0	3	0
300	68	Male	Non-Angina	144	193	True	ST Elevation	141	No	3.4	Up Sloping	2	3	0
301	57	Male	Non-Angina	130	131	False	ST Elevation	115	Yes	1.2	Up Sloping	1	3	0
302	57	Female	Typical Angina	130	236	False	Normal	174	No	0.0	Up Sloping	1	2	0

Fig 2. A Representation of The Dataset Used

3.2 Data Preprocessing

- Data Cleaning: Before the training process, it is important to clean the dataset. As the name implies data cleaning is the process of cleaning data by filling up attributes and missing values, checking and addressing any issues of inconsistencies, identifying and removing outliers and smoothing and levelling noisy data (Fatima et al., 2017). To achieve this, exploratory data analysis was conducted, to properly analyse and visualize the data and check for trends. After exploring the data, outliers were identified with the use of a box plot and removed from the data.
- Removing outliers: Outliers can have a good effect on a dataset and vice versa however this research wants to check if after removing this outlier we can still have a good accuracy score. To accomplish this, the data is converted to a numerical dataset and then an interquartile range (IQR) approach is adopted, where the lower

and upper limits based on the IQR are calculated and the outliers are identified using Boolean arrays, and then the affected rows are removed from the data frame and a new data frame is created without the outliers. See Appendix D. The IQR formula is shown below

IQR = Q3 – Q1 with Q3 being the 3rd quartile and Q1 being the 1st quartile.

*Lower = Q1 – 1.5*IQR*

*Upper = Q3 + 1.5*IQR (Dash et al., 2023)*

After addressing the outliers, a new data frame was created which contains 230 instances and 14 variables.

- Class balance: Class imbalance is a common issue in real-world applications, such as fraud detection or medical diagnosis, where one class is significantly more than the other. This imbalance can lead to biased models that favour the majority class and missing necessary data in the minority class. Class balancing in machine learning addresses the issue of imbalanced datasets and can help improve model performance. Some techniques used for class balancing are undersampling, oversampling and weighted loss (Johnson & Khoshgoftaar, 2019). In this experiment, the classes were fairly balanced and didn't require class balancing.
- Label Encoding: This technique is adopted to change categorical variables to numerical variables. In label encoding, the features are converted into an integer value. This method assigns a unique numerical value to each category within a feature, which makes it easier for machine learning algorithms to process the data (Dahouda & Joe, 2021). For example, if a feature called "letters" has three categories: "A," "B," and "C," label encoding will convert these into numerical values such as 0, 1, and 2 respectively. This transformation helps in handling categorical data that machine learning models cannot interpret in their raw form. (See Appendix D)
- Data Scaling: Data scaling is a part of the pre-processing where the data is scaled or transformed to make an equal contribution of each feature. This involves adjusting the values of numeric columns in the dataset to a common scale, without distorting differences in the ranges of values. Normalized data ensures that the algorithms perform efficiently and effectively, as it helps to avoid issues related to different scales of features, which can lead to biased outcomes. By applying normalization, each feature contributes equally to the model, improving the overall accuracy and robustness of the predictive model. Additionally, scaling helps improve the quality of the data. Normalization techniques were implemented to

standardise the scale of all features, ensuring unbiased model training (See Appendix E).

- Feature Reduction: Feature reduction is an essential part of efficient and effective classification, it ensures only necessary features are used. Various techniques are used, including data transformation, exploiting correlations between neighbouring bands, and identifying the least significant bands. High-dimensional datasets often contain redundant information. Dimension reduction eliminates this by mapping effective information from original features to a lower-dimensional space, thus reducing the impact of irrelevant or duplicated data (Jia et al., 2022). The method applied for this technique is the principle component analysis (PCA) after applying PCA it showed that all the features had their importance hence feature reduction wasn’t used.

3.3 Training and Testing

After preprocessing and normalization of the dataset, the machine learning algorithms are trained. The dataset will be divided into a test set and a training set, and the machine model will be trained using the training dataset with the following machine learning algorithm SVM, Random Forest, Naive Bayes, KNN and Ensemble learning. Hyperparameter tuning will be applied to the highest-performing algorithm(s) to optimize the model’s parameters and a cross-validation method will be applied to ensure that the model can be used in a wide range of situations.

3.4 Machine Learning Models

Choosing the different models to use for the training of this dataset is an essential part of this research. It wasn’t a difficult task as the models were selected based on the aim of this paper and the models used in the literature discussed in Chapter 2 of this study. From the reviewed literature, SVM, RF, NB and KNN were the models that performed well in predicting CVD. However, for this study in-depth exploratory data analysis will be carried out, an ensemble learning approach where combining two or more models to train will be adopted and hyperparameter tuning will be done to improve the accuracy of the highest-performing model.

- Support Vector Machine (SVM): SVM is a popular model due to its various attractive features and excellent performance. SVM is primarily used to separate different classes in the training set using a surface that enhances the margin between them. The formulation encompasses the structural risk minimization principle and is proven to be superior to the traditional empirical risk minimization principle used by conventional machine learning models. Although SVM was first developed to solve classification problems, it is now used for regression problems as well. In solving classification problems, the aim is to partition two classes by a function that is derived from the available examples and the classifier performs better on examples that have not been seen, i.e., it generalises well. Additionally, numerous possible linear classifiers can separate the data, but only one can maximise the margin which is known as the optimal separating hyperplane.

SVM is mostly used when the data is small due to some implementation demands huge training time, this is another reason why the SVM model is a good training model for this study (Cervantes et al., 2020, Roy & Chakraborty, 2023).

- K-Nearest Neighbour (KNN): The KNN algorithm is an instance-based learning method that classifies objects based on how close the training examples are in the feature space. An object is allocated to the class that is most prevalent amongst its k-nearest neighbours as seen in the figure below, where k is a positive integer.

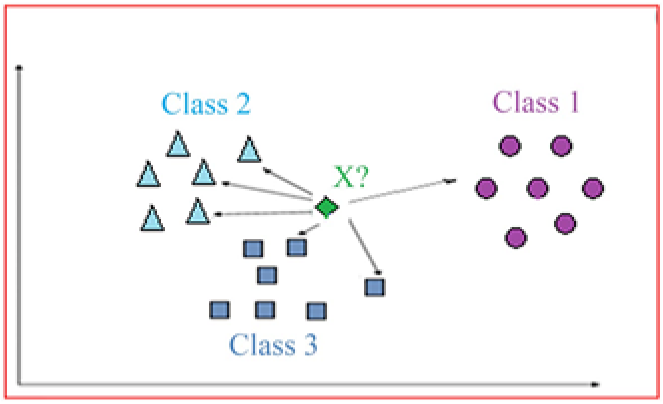


Fig 3. KNN Algorithm Representation (Boateng et al., 2020)

KNN algorithm is applied using Euclidean distance metrics to check for the nearest neighbour. KNN performs well with a large number of training examples. However, this algorithm requires that the value of parameter k that is the number of nearest neighbours and the type of distance to be employed have to be established. To determine the correct K for a dataset, the KNN algorithm will run multiple times with several values of K and the K that minimizes the number of errors encountered is selected while preserving the algorithm’s ability to predict accurately when it is applied to data for which it has no initial contact. The predictions become less reliable as the value of K reduces to 1 but become more reliable, as the value of K is increased, due to majority voting/averaging, and thus, more likely to make more accurate predictions to a certain point. Eventually, mistakes will occur and at this point, it becomes evident that the proper value of K has been exceeded. The value of K is typically an odd number to have a tiebreaker in cases where a majority vote among labels is needed, for example, when selecting the mode in a classification problem (Boateng et al., 2020).

- Random Forest (RF): The RF classifier is made up of several trees, each of which was produced by some form of randomization. Every internal node has a test that splits the space of the data to be classified. Classification takes place when an image is sent down every tree and aggregates the reached leaf distributions. Randomness can be classed into two instances, firstly when subsampling the training data to enable each tree to grow using a distinct subset and secondly when choosing the node tests. The number of trees that are essential for optimal performance increases with the number of predictors. The most effective way to determine how many trees are essential is to compare predictions made by a forest to the ones made by its subset. Once the subsets perform as well as the full forest, it shows that there are enough trees. Random forest performs very well compared to some other classifiers and it is resistant to overfitting. Additionally, random forest is user-friendly because it requires two parameters that is the number of variables in the random subset at every node and the number of trees in the forest and it is highly sensitive to the values of these variables.

  RF is basically a set of decision trees merged where each tree votes on the class assigned to a particular sample with the most frequent answer winning the vote (Boateng et al., 2020).
- Naive Bayes: Naive Bayes is a classification model used for multiclass classification problems. It is known as Naive Bayes because the calculations of the probabilities for each class are computed to make their calculations tractable. Naive Bayes classifiers are formed from Bayesian classification methods. They are based on Baye’s theorem, an equation explaining the relationship of statistical quantities' conditional probabilities. Bayesian classification is more interested in finding the probability of a label given certain observed features (Viet et al., 2021).
- Ensemble learning: The ensemble learning technique is used to merge two or more machine learning models to get a high performance compared to when they are trained individually. Instead of relying on the performance of one model, the

combinations of performances from individual models are merged using a combination rule to make one more accurate prediction.

3.5 Evaluation Metrics

The model evaluation will be obtained by analysing the test set and various evaluation metrics like accuracy, F1-score, sensitivity, and precision will be computed to determine how well the machine has learnt.

Precision is used to check the accuracy of classification models, with a particular focus on the correctness of positive predictions. It calculates the proportion of predicted positive cases that are positive, offering vital insights into the model’s performance. A high precision score suggests that the model is effective at limiting false positives, ensuring that most of its predictions are valid.

$$Precision = TP/(TP + FP)$$

Sensitivity/Recall primarily focuses on the model's capacity to detect all pertinent occurrences inside a given dataset accurately. Recall is frequently employed when the consequences of failing to identify positive cases (false negatives) are significant, such as in medical diagnostics or fraud detection. It quantifies the ratio of properly detected actual positive cases by the model. A high recall score signifies that the model successfully identifies the majority of positive cases, which is crucial in situations when missing positive cases could result in significant consequences.

$$Recall = TP/P = TP/(TP + FN) \quad \text{(Schlosser et al., 2024)}$$

F1 Score is the mean of precision and recall and can be determined by the formula below

$$F1= 2*TP/(2*TP + FP + FN)$$

Accuracy measures the number of correct predictions. That is it measures how a machine learning model predicts an outcome correctly.

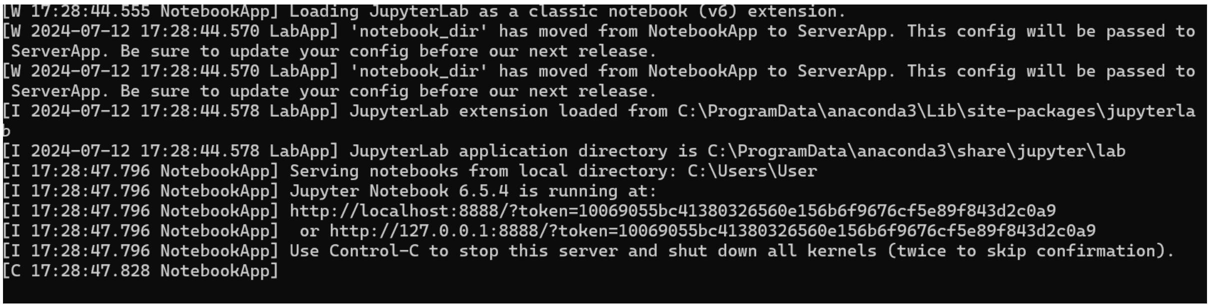
$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \quad \text{(Muntean \& Militaru, 2023)}$$

Hyperparameter tuning is optimizing the model using tuning techniques to enhance its performance.

Cross-validation is a technique for assessing a machine-learning model’s performance. It checks to ensure that the model is good by using the data several times to test the performance of the model.

3.6 Hardware and Implementation

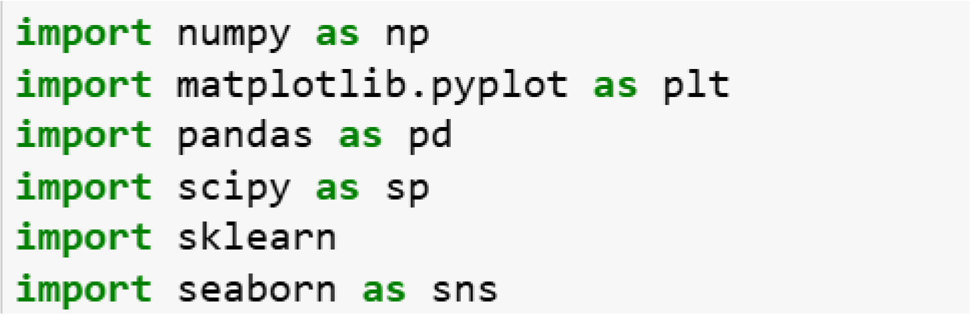
To carry out the practical component of this research, Python programming language and open-source libraries necessary for this research will be employed. Additionally, all the experiments are performed on the Jupyter Notebook environment version 6.5.4 which has been installed.



```
[W 17:28:44.555 NotebookApp] Loading JupyterLab as a classic notebook (v6) extension.
[W 2024-07-12 17:28:44.570 LabApp] 'notebook_dir' has moved from NotebookApp to ServerApp. This config will be passed to
 ServerApp. Be sure to update your config before our next release.
[W 2024-07-12 17:28:44.570 LabApp] 'notebook_dir' has moved from NotebookApp to ServerApp. This config will be passed to
 ServerApp. Be sure to update your config before our next release.
[I 2024-07-12 17:28:44.578 LabApp] JupyterLab extension loaded from C:\ProgramData\anaconda3\Lib\site-packages\jupyterla
b
[I 2024-07-12 17:28:44.578 LabApp] JupyterLab application directory is C:\ProgramData\anaconda3\share\jupyter\lab
[I 17:28:47.796 NotebookApp] Serving notebooks from local directory: C:\Users\User
[I 17:28:47.796 NotebookApp] Jupyter Notebook 6.5.4 is running at:
[I 17:28:47.796 NotebookApp] http://localhost:8888/?token=10069055bc41380326560e156b6f9676cf5e89f843d2c0a9
[I 17:28:47.796 NotebookApp]  or http://127.0.0.1:8888/?token=10069055bc41380326560e156b6f9676cf5e89f843d2c0a9
[I 17:28:47.796 NotebookApp] Use Control-C to stop this server and shut down all kernels (twice to skip confirmation).
[C 17:28:47.828 NotebookApp]
```

Fig 4. Jupyter Notebook Version

Figure 4 shows the version of Jupyter Notebook installed on the system while Figure 5 shows some of the libraries used that are important for this study.



```python
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import scipy as sp
import sklearn
import seaborn as sns
```

Fig 5. Libraries Imported for this study

3.7 Gap Analysis

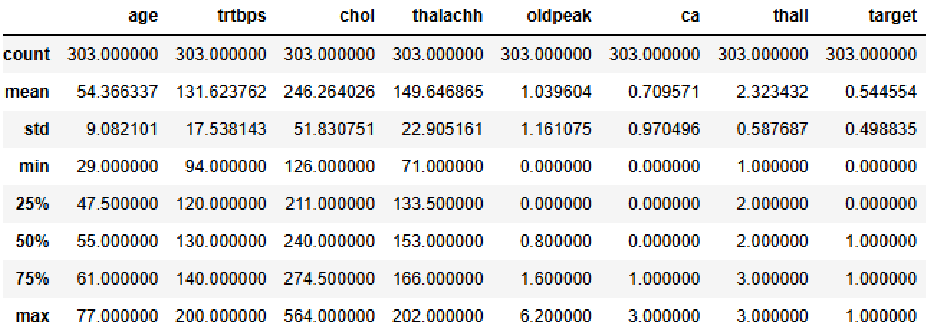
Many researchers have applied different machine learning techniques for predicting cardiovascular disease, and understanding the current state of progress and available results is essential. Different literature was reviewed in this study and all of these papers carried out the experiment in similar ways of not handling outliers. This gap analysis has provided some insights on the effect of outliers in any data.

4 Experimental Analysis and Results

In this chapter, the results obtained from this research will be explained in detail. From the literature, it is observed that not all authors performed an exploratory analysis. Furthermore, the few that did conduct an exploratory data analysis did not do so in depth. This research aims to address this gap by conducting a comprehensive analysis to determine if removing outliers improves the model's training and accuracy scores. To achieve this, two different experiments were conducted: one including outliers and one excluding them. A comparative analysis of the results from these experiments will be discussed towards the end of this chapter. The goal is to provide a thorough understanding of the impact of outliers on model performance and to present a detailed evaluation of the findings. Below are the steps taken for this experiment:

4.1 *Exploratory Data Analysis*

The heart disease dataset comprises a total of 303 instances and 13 features, excluding the target column. Figure 6 provides a summary of the data obtained. The minimum age in this dataset is 29, and the maximum age is 77, with a mean age of 54 years. This indicates that the data encompasses individuals ranging from 29 to 77 years old. Additionally, Table 4 presents the unique values of each variable that contains more than one distinct value. This detailed summary helps in understanding the demographic distribution and variability within the dataset, which is crucial for the subsequent analysis.



	age	trtbps	chol	thalachh	oldpeak	ca	thall	target
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000
mean	54.366337	131.623762	246.264026	149.646865	1.039604	0.709571	2.323432	0.544554
std	9.082101	17.538143	51.830751	22.905161	1.161075	0.970496	0.587687	0.498835
min	29.000000	94.000000	126.000000	71.000000	0.000000	0.000000	1.000000	0.000000
25%	47.500000	120.000000	211.000000	133.500000	0.000000	0.000000	2.000000	0.000000
50%	55.000000	130.000000	240.000000	153.000000	0.800000	0.000000	2.000000	1.000000
75%	61.000000	140.000000	274.500000	166.000000	1.600000	1.000000	3.000000	1.000000
max	77.000000	200.000000	564.000000	202.000000	6.200000	3.000000	3.000000	1.000000

Fig 6 Summary of the data

Variables	**Unique Values**	**Values**
Sex	2	Male, Female
Chest pain (cp)	4	Asymptomatic, Atypical Angina, Typical Angina, Non-angina
Fasting blood sugar (fbs)	2	True, False

Rest Ecg	3	Normal, ST Elevation, Left Ventricular Hypertrophy
exng	2	Yes, No
Slope	3	Flat, Up sloping, Down Sloping
CA	4	0,1,2,3
Thall	3	1,2,3
Target	2	0,1

Table 5. Variables with more than one value

After checking the unique values in the dataset, the next step is to examine the data for any missing values and assess the balance between classes. Upon conducting this analysis, it was found that there were no missing values in the dataset. Additionally, the classes in the target variable were found to be fairly balanced. This balance is illustrated in Figure 7 below. Specifically, out of the 303 individuals in the dataset, 165 have been tested for cardiovascular disease (CVD), while 138 have not. This indicates that the distribution between the two classes is relatively even, which is beneficial for model training and evaluation.

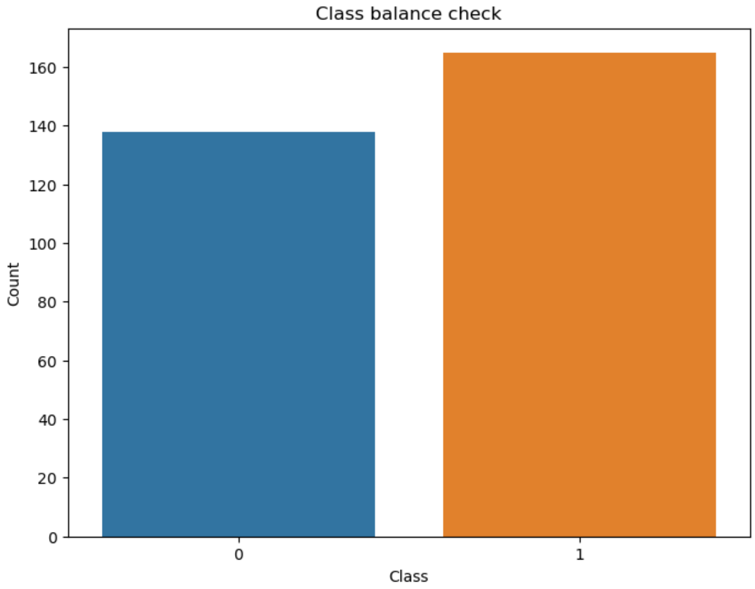


Fig 7. Class Balance Check for 303 instances

Distribution plots and box plots for different variables: The plots below illustrate the heart disease ranges for various variables. Figures 8 specifically show the distribution of age and the presence of cardiovascular disease (CVD). The box plot indicates that individuals diagnosed with cardiovascular disease are predominantly between the ages of 40 and 60, with only one outlier observed in the lower end of this age range. Conversely, individuals without cardiovascular disease fall within the age range of 51 to 62 years, and the plot shows no outliers in this group. This dataset suggests that the majority of people diagnosed with cardiovascular disease are among the younger and older age groups. The visual representation through these plots provides a clear understanding of how age correlates with the prevalence of cardiovascular disease, highlighting the significant age ranges where CVD is most commonly observed.

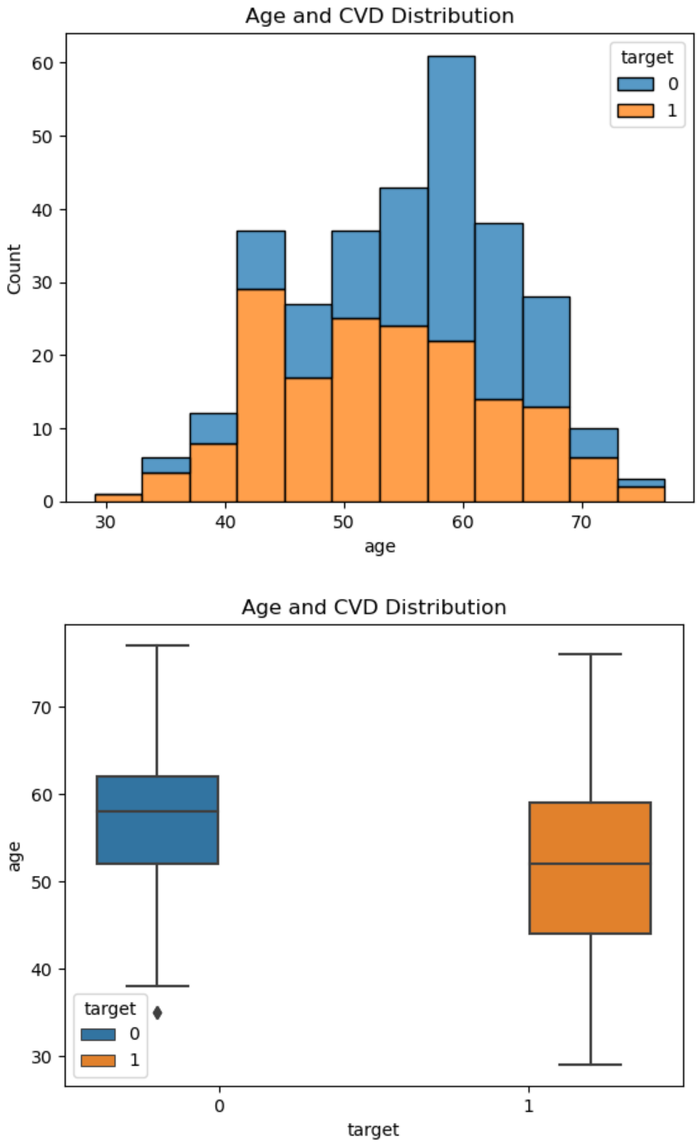


Figure 8. Age and CVD distribution

Figure 9 illustrate that patients with cardiovascular disease (CVD) generally have blood pressure rates ranging from 120 mmHg to 140 mmHg, while patients without cardiovascular disease have blood pressure rates spanning from 120 mmHg to 143 mmHg. This suggests that there is considerable overlap in the blood pressure rates of patients with and without CVD. Furthermore, the box plot analysis indicates that both groups exhibit similar blood pressure distributions. However, it is noteworthy that there are a few outliers present on both sides of the upper margin, indicating some variability in blood pressure readings. This observation is critical as it highlights the potential challenges in distinguishing between patients with and without CVD based solely on blood pressure metrics.

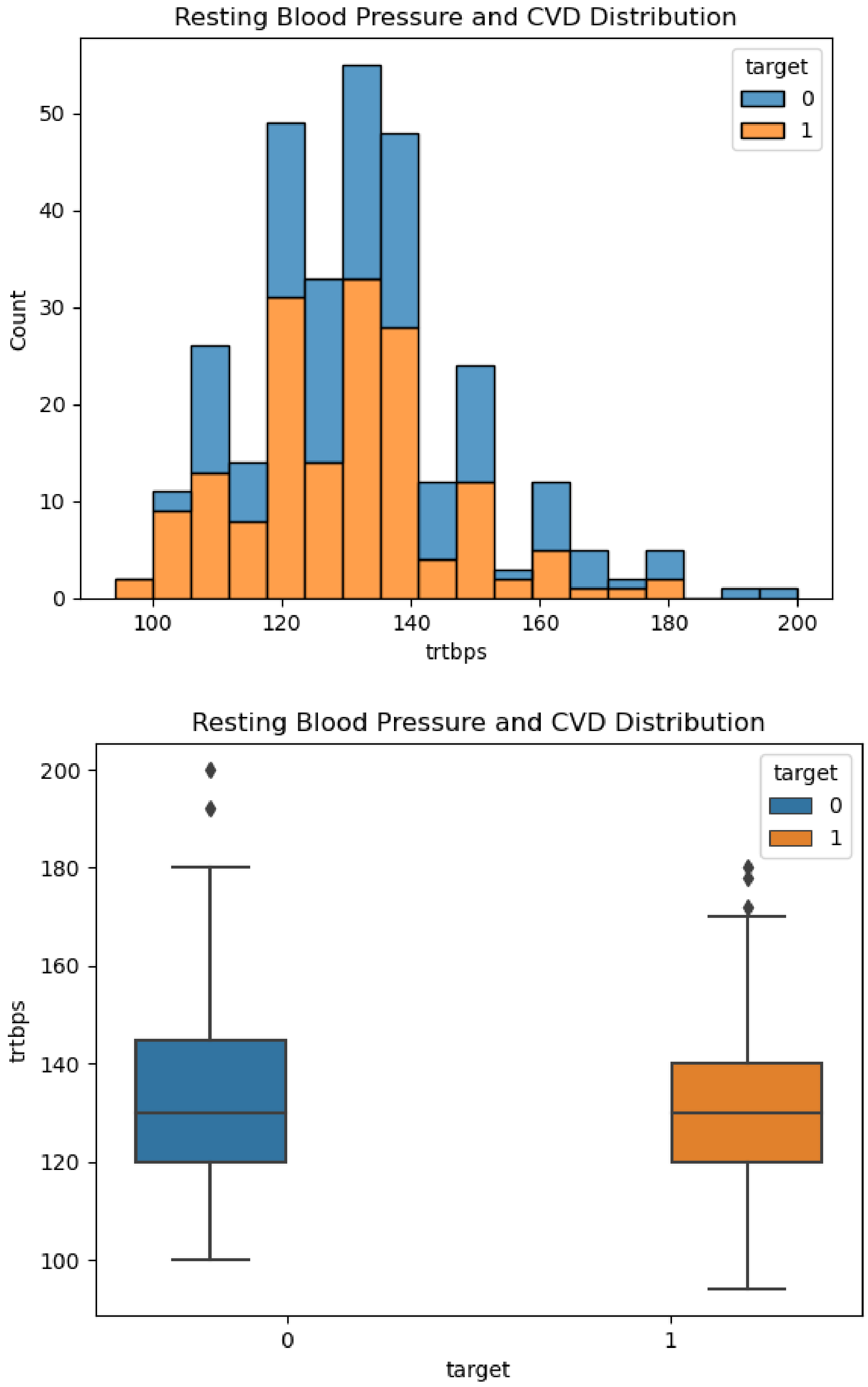
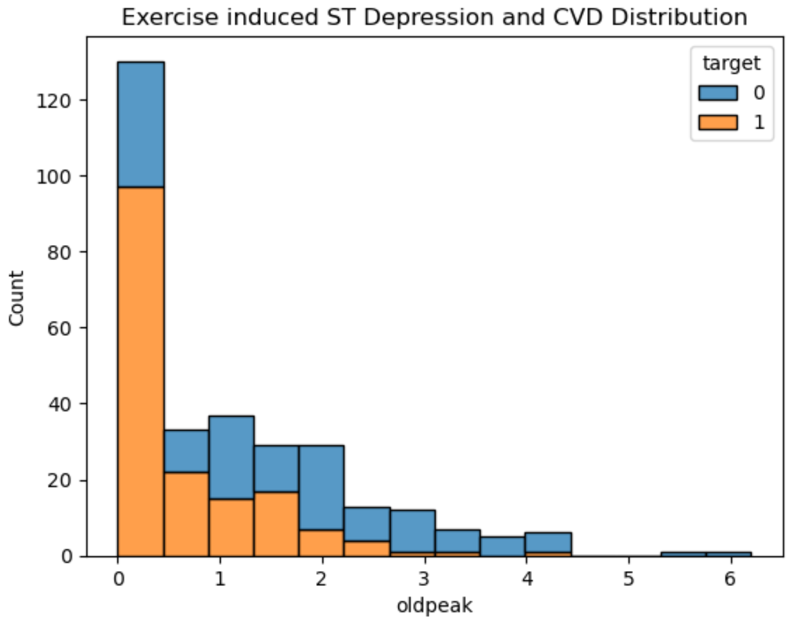


Figure 9. Resting blood pressure and CVD on a box plot

Furthermore, Figure 10 illustrates the ST depression rate in patients diagnosed with cardiovascular diseases and those without such diagnoses. The figures reveal that ST depression is more prevalent among patients with cardiovascular disease compared to those without. Notably, patients with cardiovascular disease exhibit a greater number of upper outliers in ST depression levels than those without the condition. Specifically, the ST depression in patients with cardiovascular disease ranges from 0 to 2 mm, whereas in patients without cardiovascular disease, it ranges from 1 to 2.5 mm. This indicates a broader distribution and higher incidence of elevated ST depression levels in the cardiovascular disease group. The presence of these outliers in ST depression among cardiovascular patients may suggest a more severe or varied manifestation of the disease, highlighting the importance of monitoring this metric closely in clinical assessments and treatment planning.



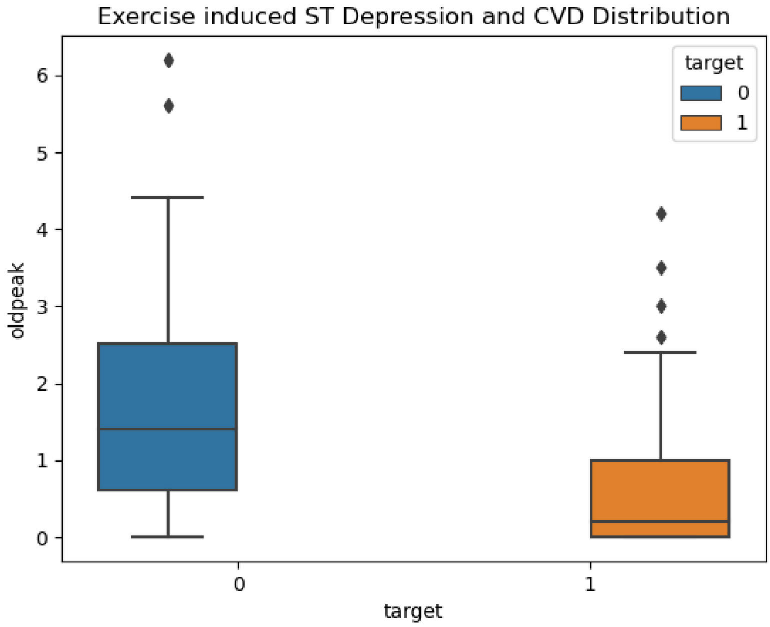


Figure 10. ST depression (oldpeak) and CVD distribution

Additionally, Figure 11 shows that patients with cardiovascular disease have a maximum heart rate per person between 120 and 155 beats per minute, whereas patients without cardiovascular disease have a maximum heart rate per person ranging from 150 to 170 beats per minute. This distinction highlights a notable difference in heart rate patterns between the two groups. Both classes exhibited lower outliers, with patients with cardiovascular disease having the most significant number of outliers. This indicates that the variability in heart rate is more pronounced in this group. Figure 12 illustrates the cholesterol levels among patients. It is evident that patients with cardiovascular disease tend to have higher cholesterol levels compared to those without cardiovascular disease. Both classes displayed upper outliers, but patients without cardiovascular disease had a greater number of outliers. This suggests a broader range of cholesterol levels in the non-CVD group. These figures collectively emphasise the importance of heart rate and cholesterol levels as key indicators in the analysis of cardiovascular disease.

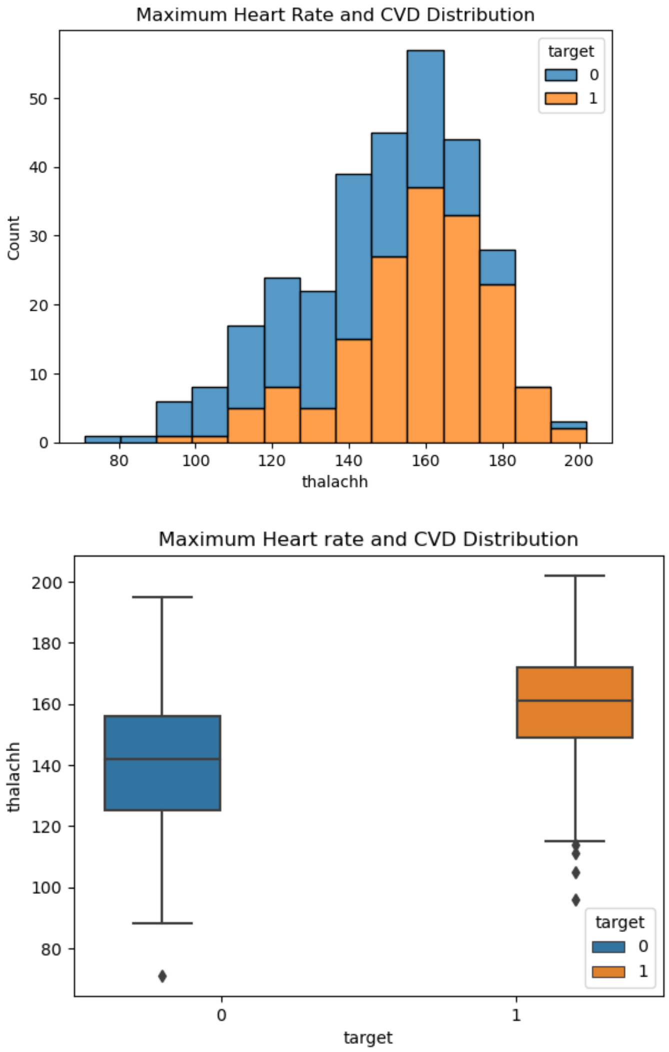


Figure 11. Maximum heart rate and CVD distribution

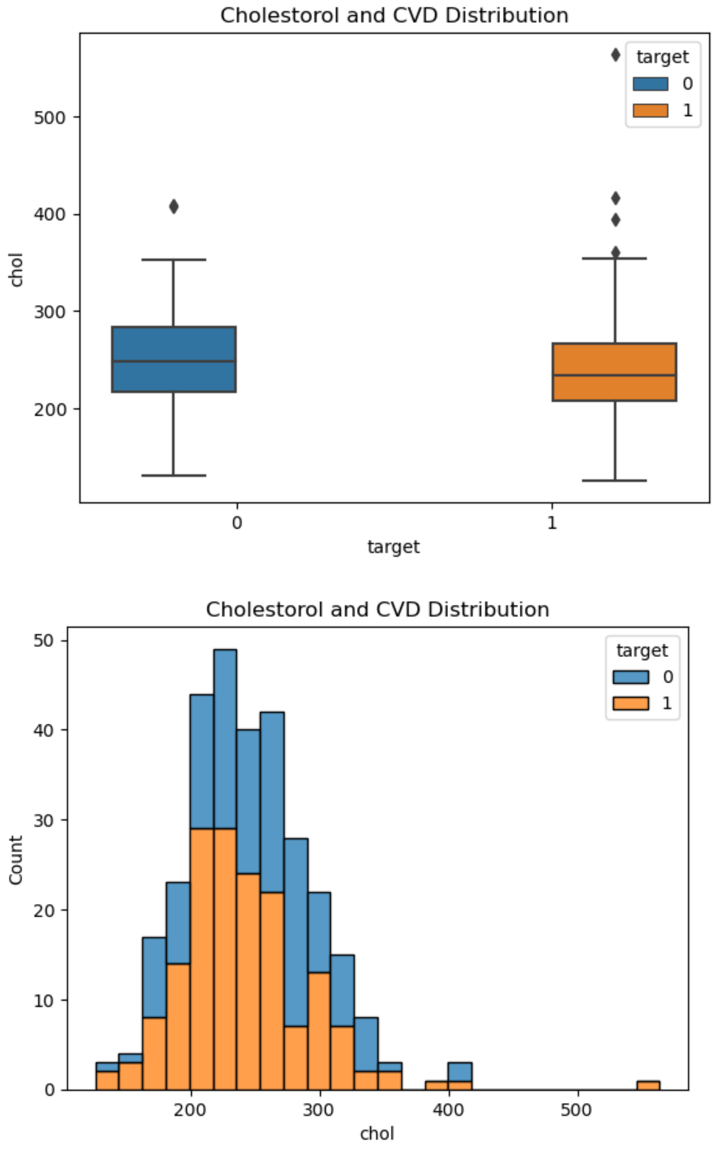


Figure 12. Cholesterol and CVD distribution

Checking the gender that may be more prone to cardiovascular heart disease based on the dataset used was also considered. Figure 13 shows that 75% of the women in this dataset were diagnosed with cardiovascular disease, whereas 44.9% of the men were diagnosed with the condition. According to this data, women are more likely to be diagnosed with cardiovascular disease than men. This significant finding underscores the importance of considering gender differences in cardiovascular health studies. A correlation matrix was also created to examine the relationships between all the variables, as seen in Figure 14. The target variable (CVD) had the strongest positive correlation of 0.42 with maximum heart rate (thalachh), indicating that higher heart rates are associated with a greater likelihood of cardiovascular disease. Also, the strongest negative correlation was -0.43 with ST depression (oldpeak), suggesting that higher levels of ST depression are associated with a lower likelihood of cardiovascular disease. These correlations are important as they will help in identifying key variables and predictors of cardiovascular disease within the dataset, providing valuable insights for further analysis.

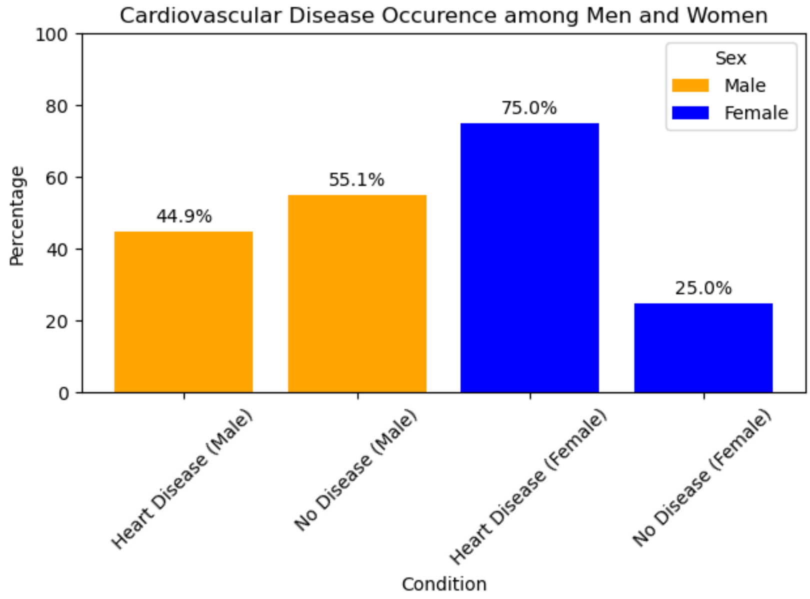


Fig 13. Occurrence of cardiovascular disease among men and women

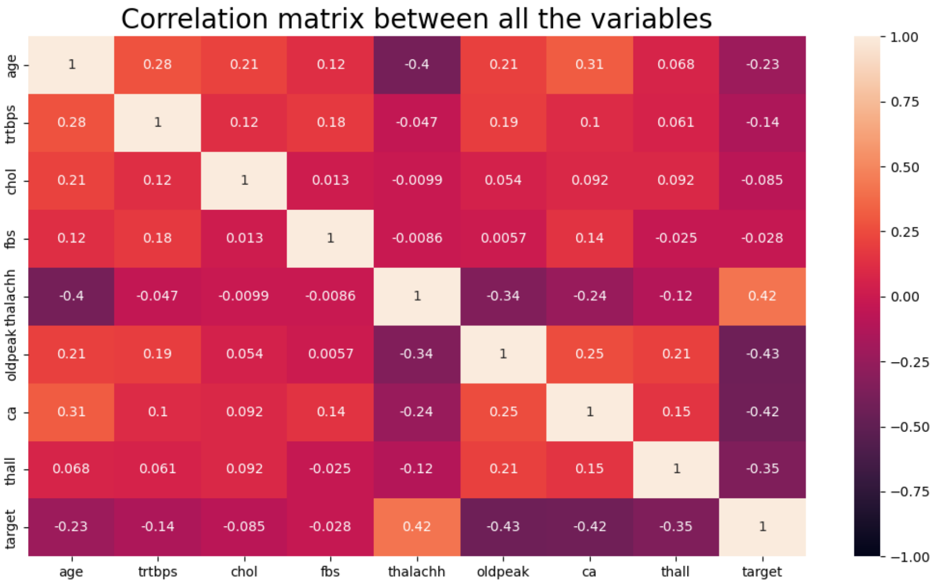


Figure 14. Correlation matrix

The overall insight obtained from exploring the data showed that women are more likely to have heart disease, and individuals over the age of 55 are less likely to have cardiovascular disease. Additionally, individuals with high cholesterol levels are at a higher risk of developing cardiovascular disease. It is evident that all variables play an important role in the diagnosis of cardiovascular disease, highlighting the complexity and multifactorial nature of this health condition.

After thoroughly exploring the data, outliers were identified and removed to ensure the integrity of the study. As initially stated, the research included two experiments: one involving training with the outliers included and the other with the outliers removed. After removal, another class balance check was conducted to ensure that the classes were still balanced, as shown in Figure 15. Interestingly, the results confirmed that the classes remained fairly balanced even after the outliers were removed.

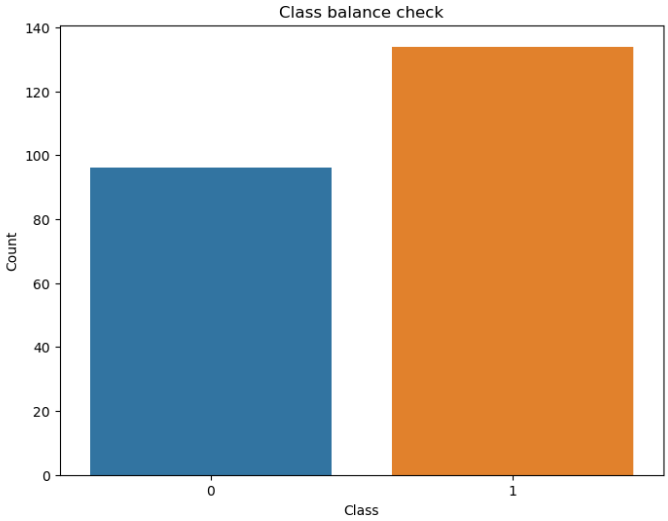


Figure 15. Class balance check for 230 instances

4.2 *Dimensionality Reduction With PCA*

Figure 16 demonstrated that all features had significant importance in the model, and reducing the number of features was unnecessary. Therefore, all features were retained and used for both experiments.

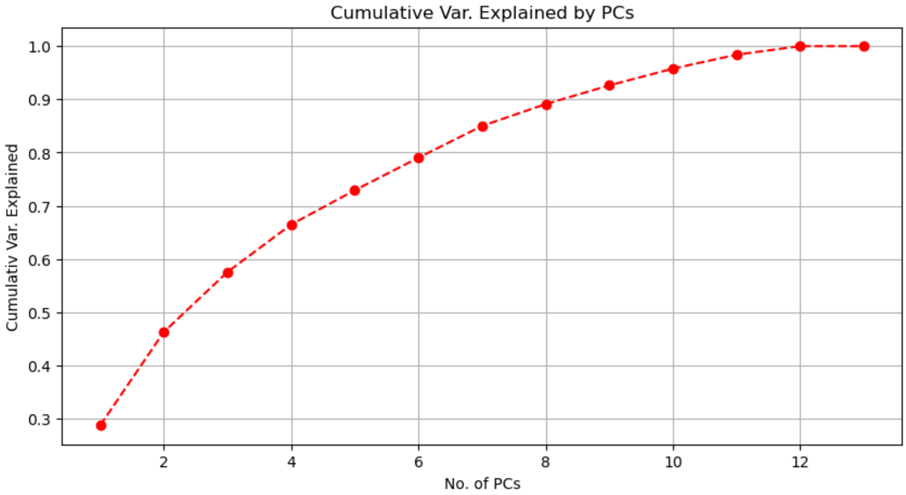


Figure 16. Line graph showing the cumulative variance of the features explained by PCA

4.3 Performance Evaluation

For this experiment, various machine learning algorithms were employed to train the data, including k-nearest neighbour (KNN), support vector machine (SVM), random forest, naive Bayes, and ensemble learning methods combining all four models and the best two performing models. The models were trained using all the features and the data was split using 80% of the data used for the training set and 20% used for the testing set and then hyperparameter tuning was performed on the highest performing model. To assess the performance of these models, these evaluation metrics will be used, namely the accuracy score, F1 score, precision, and recall.

4.3.1 Experiment 1

The first experiment focuses on training the models without the presence of outliers in the dataset. Outliers were identified and removed to ensure the integrity and reliability of the analysis. After this preprocessing step, the dataset was reduced to 230 instances, which is a more manageable and cleaner subset for model training. The impact of outlier removal on the performance of each model is analysed in detail. The results achieved by each machine learning model are discussed below, providing a comprehensive understanding of how each method performs under these conditions. This comparison highlights the effectiveness of different algorithms in handling outlier-free data.

1. KNN: After training and testing with KNN, below is how the model performed.

	Precision	Recall	F1-score	Accuracy at 5 neighbours
Class 0	0.83 (83%)	0.48 (48%)	0.61 (61%)	
Class 1	0.68 (68%)	0.92 (92%)	0.78 (78%)	
Weighted average	0.75 (75%)	0.72 (72%)	0.70 (70%)	
				0.72 (72%)

Table 6. Evaluation Metrics for KNN

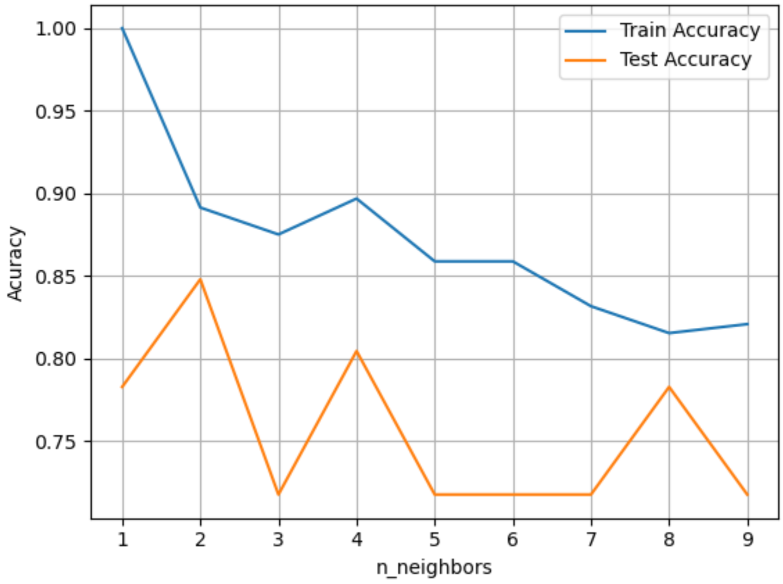


Figure 17. Training and testing accuracy for the KNN model

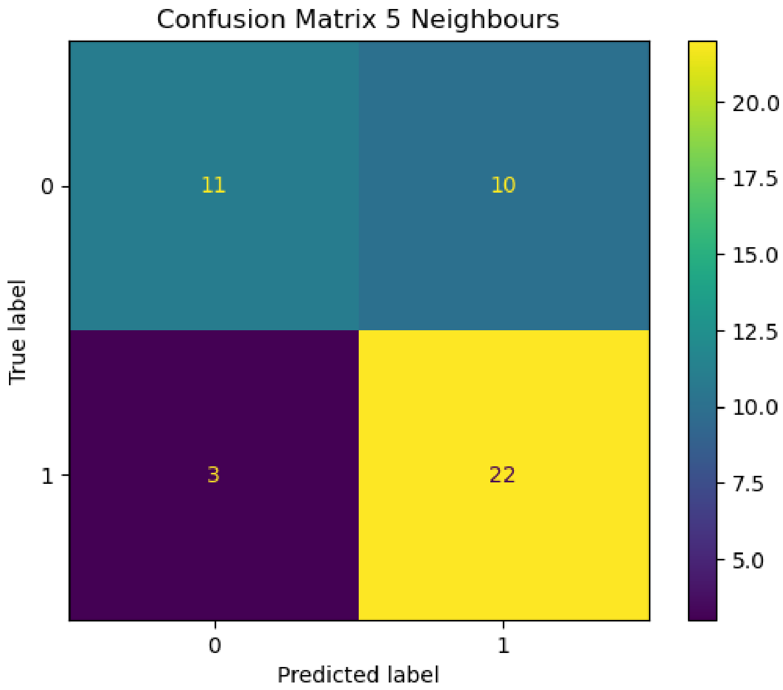


Figure 18. KNN confusion matrix

Precision: During prediction, when the model predicted cases with no CVD, it was correct 83% of the time and when it predicted cases with CVD, it was correct 68% of the time. Thus, while the model demonstrates a higher precision for no CVD, indicating more reliability when predicting this class, its precision for cases with CVD is lower, showing more frequent misclassifications in this category.

Recall: the model was able to identify 92% of the cases with CVD and 48% of the cases without CVD, indicating a higher rate of identifying positive cases and missing a significant portion of negative cases. This shows that the model is more effective in the detection of CVD cases than cases without CVD.

The F1-score shows that the model performed better at identifying and predicting CVD cases compared to cases without CVD.

From the confusion matrix in Figure 18, out of 46 test sets the model was able to predict correctly 11 cases of class 0 and 22 cases of class 1. Despite the model being able to achieve a higher number of correct predictions, it didn’t perform well in predicting class 0.

2. Random forest: below is how the model performed.

	Precision	Recall	F1-score	Accuracy at 5 estimators
Class 0	0.92 (92%)	0.52 (52%)	0.67 (67%)	
Class 1	0.71 (71%)	0.96 (96%)	0.81 (81%)	
Weighted average	0.80 (80%)	0.76 (76%)	0.75 (75%)	
				0.74 (74%)

Table 7. Evaluation Metrics for Random Forest

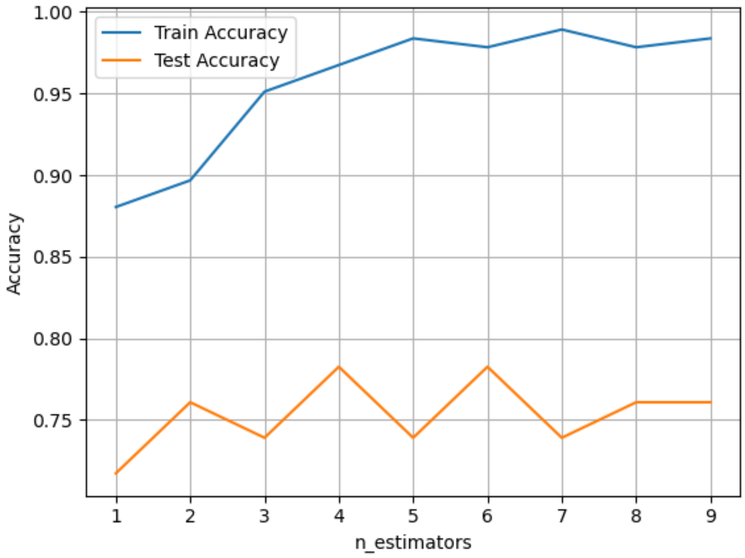


Figure 19. Training and testing accuracy for the Random Forest model

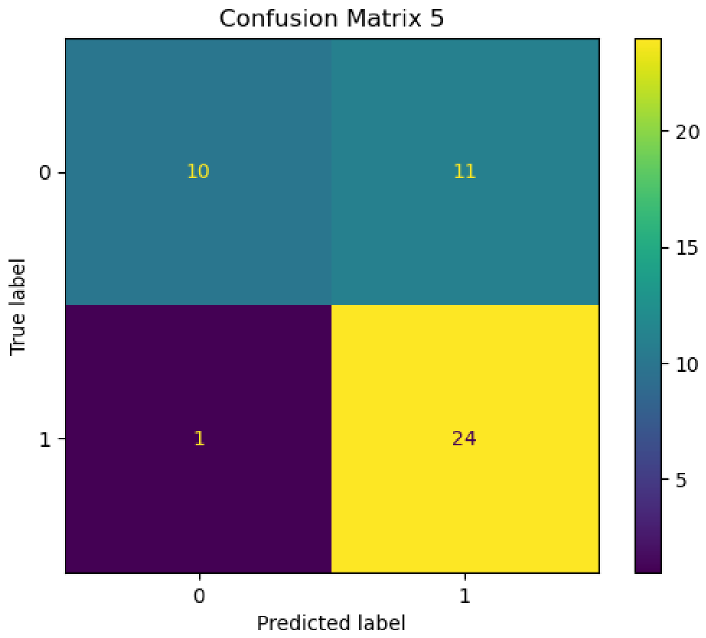


Figure 20. Random Forest Confusion Matrix

Precision: During prediction, when the model predicted cases with no CVD, it was correct 92% of the time and when it predicted cases with CVD, it was correct 71% of the time. Thus, while the model demonstrates a higher precision for no CVD, the model did perform very well in predicting both classes.

Recall: the model was able to identify 96% of the cases with CVD and 52% of the cases without CVD, which once again indicates a higher rate of identifying positive cases and missing some portion of negative cases, showing that the model is more effective in detecting CVD cases than cases without CVD.

The F1-score shows that the model performed better at identifying and predicting CVD cases compared to cases without CVD. This gives a balanced measure of performance.

The confusion matrix in Figure 20 shows that the model was able to predict correctly 10 positive cases and 24 negative cases. However, failed to identify 11 positive cases which is not a good performance and can be problematic seeing that the data used for this experiment is for medical diagnosis.

3. Support vector machine (RBF):

	Precision	Recall	F1-score	Accuracy
Class 0	0.79 (79%)	0.52 (52%)	0.63 (63%)	
Class 1	0.69 (69%)	0.88 (88%)	0.77 (77%)	
Weighted average	0.73 (73%)	0.72 (72%)	0.71 (71%)	
				0.72 (72%)

Table 8. Evaluation Metrics for SVM

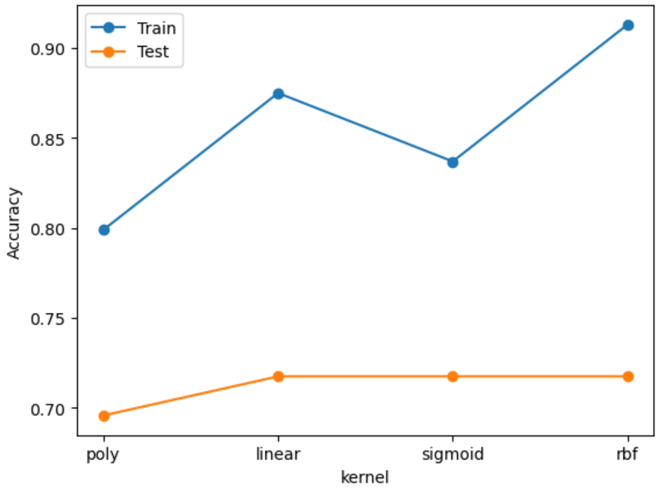


Figure 21. Training and testing accuracy for the SVM model

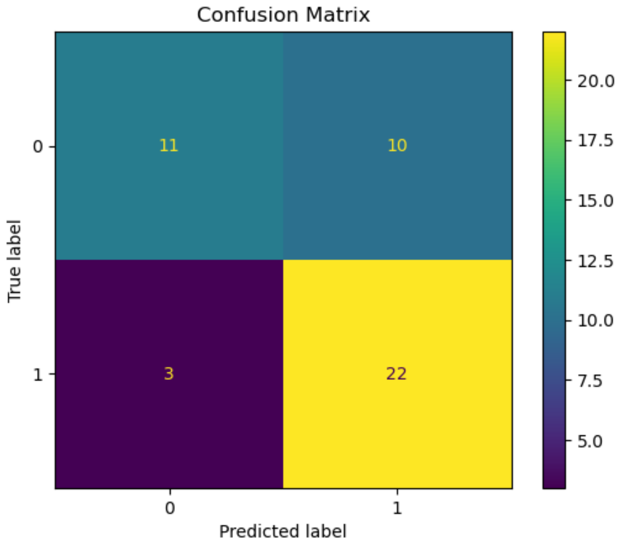


Figure 22. SVM Confusion Matrix

The result shows that during prediction, when the model predicted cases with no CVD, it was correct 79% of the time and when it predicted cases with CVD, it was correct 69% of the time. The model performed well in predicting both classes given their precision score however in predicting class 1 did not perform as well as predicting class 0.

Recall: the model was able to identify 88% of the cases with CVD and 62% of the cases without CVD, which once again indicates a higher rate of identifying positive cases, showing that the model is more effective in detecting CVD cases than cases without CVD.

The F1-score shows that the model performed well in identifying and predicting CVD cases compared to cases without CVD. However, the weighted average score gives a balanced measure of performance.

The SVM confusion matrix shows that the model was able to predict correctly 11 positive cases and 22 negative cases. However, failed to identify 10 positive cases which is not a very good performance and can be an issue seeing that the data used for this experiment is for medical diagnosis.

4. Naive Bayes:

	Precision	Recall	F1-score	Accuracy
Class 0	0.81 (81%)	0.62 (62%)	0.70 (70%)	
Class 1	0.73 (73%)	0.88 (88%)	0.80 (80%)	
Weighted average	0.77 (77%)	0.76 (76%)	0.76 (76%)	
				0.76 (76%)

Table 9. Evaluation Metrics for Naive Bayes Model

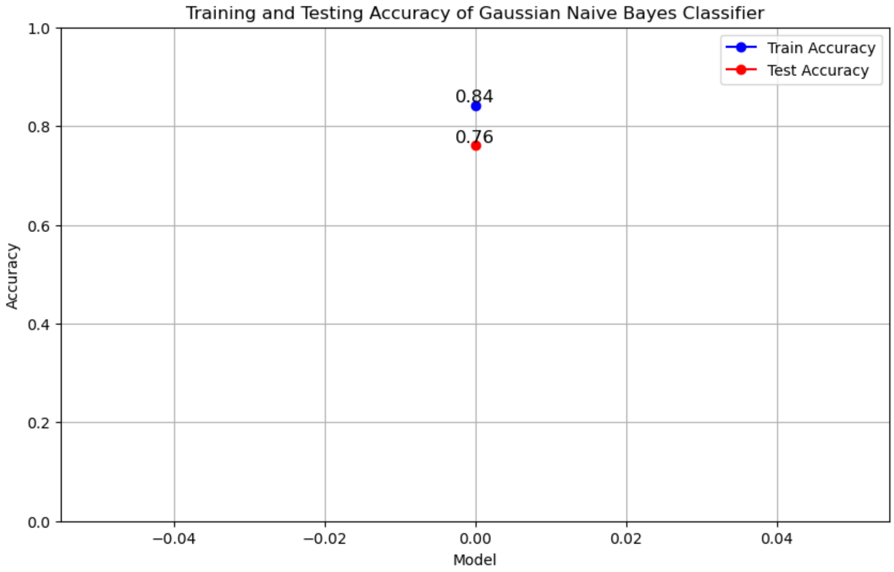


Figure 23. Training and testing accuracy for the Naive Bayes Model

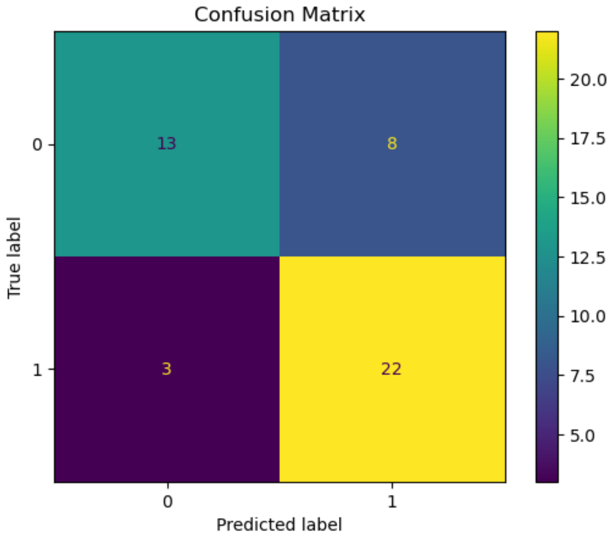


Figure 24. Naive Bayes Model Confusion Matrix

The precision score shows that during prediction, when the model predicted cases with no CVD, it was correct 81% of the time and when it predicted cases with CVD, it was correct 73% of the time. The model performed well in predicting both classes given their precision score.

Recall: the model was able to identify 88% of the cases with CVD and 62% of the cases without CVD, which once again indicates a higher rate of identifying positive cases, showing that the model is more effective in detecting CVD cases than cases without CVD.

The F1-score shows that the model performed well in identifying and predicting CVD cases compared to cases without CVD. However, the weighted average score gives a balanced measure of performance.

The Naive Bayes confusion matrix shows that the model was able to predict correctly 13 positive cases and 22 negative cases. However, failed to identify 8 positive cases which is not a very bad performance compared to other models.

Looking at the evaluation metrics of these models including their accuracy score, the models performed poorly even if Naive Bayes performed better than the rest with an accuracy score of 76% it still performed poorly. To boost the scores of these models, combining the best two and all four of them is another step taken in this experiment, which is called the ensemble learning method, another step taken is to tune the best-performing models to see if the scores will increase.

5. Ensemble learning using a voting classifier was implemented in this experiment, all four models were combined. Additionally, the two best-performing models were combined which are random forest with a 74% accuracy score and naive bayes

with a 76% accuracy score. The combined result of these experiments is shown below:

	Precision	Recall	F1-score	Accuracy
Class 0	0.80 (80%)	0.57 (57%)	0.67 (67%)	
Class 1	0.71 (71%)	0.88 (88%)	0.79 (79%)	
Weighted average	0.75 (75%)	0.74 (74%)	0.73 (73%)	
				Train accuracy: 92% Test Accuracy: 74%

Table 10. Ensemble learning of all four models combined

	Precision	Recall	F1-score	Accuracy
Class 0	0.82 (82%)	0.67 (67%)	0.74 (74%)	
Class 1	0.76 (76%)	0.88 (88%)	0.81 (81%)	
Weighted average	0.79 (79%)	0.78 (78%)	0.78 (78%)	
				Train accuracy: 94% Test Accuracy: 78%

Table 11. Ensemble learning of random forest and naive bayes

The results shown in Table 10 and Table 11 reveal that the ensemble learning approach, when using all four models, yielded accuracy scores similar to the individual models, which were relatively low. However, when combining only naive Bayes and random forest, the accuracy score improved significantly, reaching 78%. This improvement highlights the effectiveness of using ensemble learning to enhance model performance. Moreover, the precision and recall scores were higher in the ensemble model compared to the individual models. Specifically, the ensemble model was able to correctly predict the absence of cardiovascular disease (class 0) 80% of the time, and it correctly identified the presence of cardiovascular disease (class 1) 88% of the time.

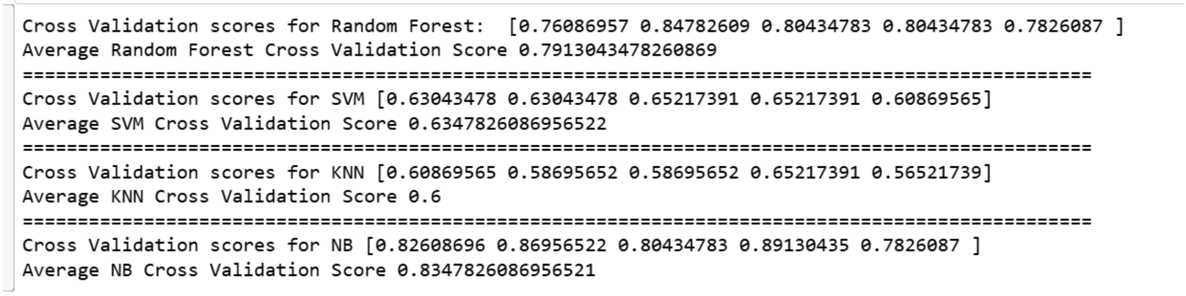
6. Hyperparameter tuning on the best-performing models Random Forest and Naive Bayes using random search:

Tuned Model	Best Parameters	Precision	Recall	F1-Score	Accuracy
Random Forest	Best Parameters: {'n_estimators': 25, 'max_leaf_nodes': 6, 'max_features': 'log2', 'max_depth': 6}	Class 0: 71% Class 1: 92% Weighted Avg: 84%	Class 0: 88% Class 1: 79% Weighted Avg: 83%	Class 0: 79% Class 1: 85% Weighted Avg: 83%	78.26%
Naive Bayes	Best Parameters: {'var_smoothing': 0.01}	Class 0: 48% Class 1: 96% Weighted Avg: 84%	Class 0: 91% Class 1: 69% Weighted Avg: 74%	Class 0: 62% Class 1: 80% Weighted Avg: 76%	76.09%

Table 12. Hyperparameter tuning results

From the result shown in Table 12, these tuned models worked better at predicting and identifying cases with CVD compared to cases without CVD and they did much better, However, Naive Bayes accuracy score remained the same even after tuning but Random Forest performed better with 4.26%.

7. Cross-validation: Initially cross-validation was adopted for this experiment to enable us to confirm the performances of these models, however during the experiment, looking at the graphs and results of each model, overfitting was observed from the high performance obtained from the training set and low result from the test set showing that is not generalizing well with the data hence another need for cross-validation as cross-validation gives a more realistic performance of the model. 5-fold cross-validation was deployed for this experiment. The cross-validation score is shown in Figure 25 below.



```
Cross Validation scores for Random Forest:  [0.76086957 0.84782609 0.80434783 0.80434783 0.7826087 ]
Average Random Forest Cross Validation Score 0.7913043478260869
==================================================================================================
Cross Validation scores for SVM [0.63043478 0.63043478 0.65217391 0.65217391 0.60869565]
Average SVM Cross Validation Score 0.6347826086956522
==================================================================================================
Cross Validation scores for KNN [0.60869565 0.58695652 0.58695652 0.65217391 0.56521739]
Average KNN Cross Validation Score 0.6
==================================================================================================
Cross Validation scores for NB [0.82608696 0.86956522 0.80434783 0.89130435 0.7826087 ]
Average NB Cross Validation Score 0.8347826086956521
```

Figure 25. Cross-validation score for all four models

The cross-validation scores for each model are relatively consistent excluding KNN which is fairly consistent and Naive Bayes which have relatively high values. This consistency shows that each model's performance is stable within different subsets of the data. The models are training well however they are not performing as effectively on the test data and the difference between the training accuracy, test accuracy and the average score from the cross-validation indicates that the models may be overfitting which can attributed to the reduced size of the dataset used for this experiment after removing outliers. Notably, Naive Bayes achieved a better score of 0.834 indicating that the model is generally performing well but there might be some variability that caused the drop in performance on the initial test data.

4.3.2 Experiment 2

With all the observations from the first experiment, a second experiment with the outliers included was carried out to check if the models would perform better. This experiment has the complete data of 303 instances and 13 columns.

1. KNN:

	Precision	Recall	F1-score	Accuracy at 5 neighbours
Class 0	0.76 (76%)	0.86 (86%)	0.81 (81%)	
Class 1	0.86 (86%)	0.75 (75%)	0.80 (80%)	
Weighted average	0.81 (79%)	0.80 (80%)	0.80 (80%)	
				78%

Table 13. KNN Evaluation Metrics for Experiment 2

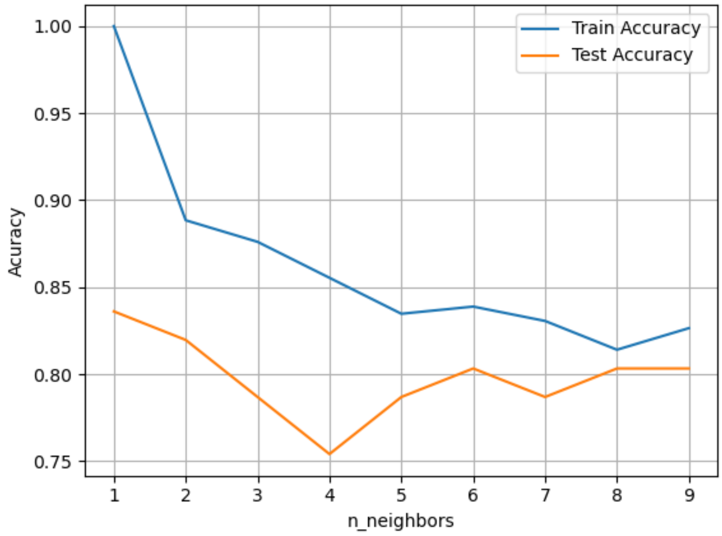


Figure 26. Training and testing accuracy for the KNN Experiment 2

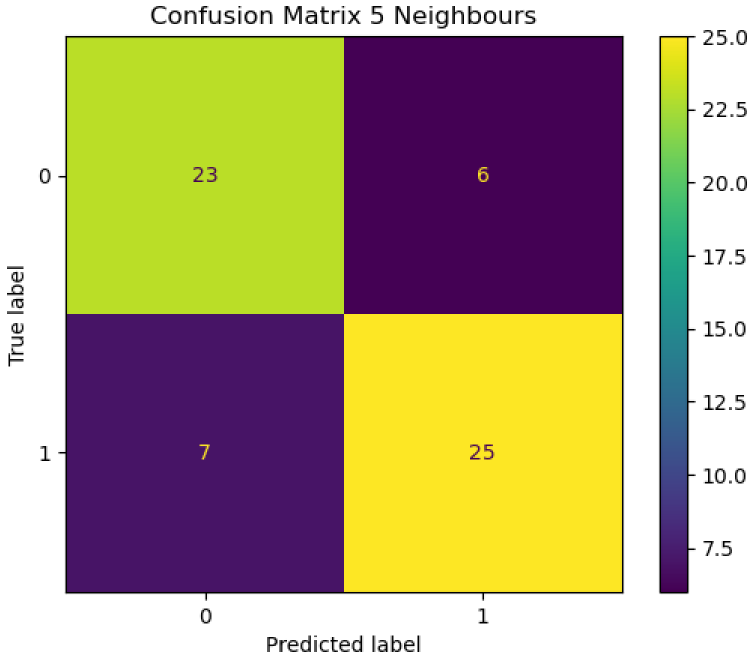


Figure 27. KNN Confusion Matrix for Experiment 2

The precision score reveals that when the model predicted cases with no CVD, it was correct 76% of the time and when it predicted cases with CVD, it was correct 86% of the time.

Recall: the model was able to identify 75% of the cases with CVD and 86% of the cases without CVD. This shows that the model is more effective in the detection of cases without cardiovascular disease than cases with cardiovascular disease.

The F1-score shows that the model performed well at identifying and predicting both cases with CVD and cases without CVD.

From the confusion matrix in Figure 27, 23 instances of class 0 were correctly predicted as class 0 while 6 instances of class 0 were incorrectly predicted as class 1 and 25 instances of class 1 were predicted correctly as class 1 while 7 instances of class 1 were incorrectly predicted as class 0. The model performed well in predicting and identifying however, the accuracy score is not high enough but did better with outliers included.

2. Random forest: below is how the model performed.

	Precision	Recall	F1-score	Accuracy at 5 estimators
Class 0	0.77 (77%)	0.93 (93%)	0.84 (84%)	
Class 1	0.92 (92%)	0.75 (75%)	0.83 (83%)	
Weighted average	0.85 (85%)	0.84 (84%)	0.84 (84%)	
				0.81 (81%)

Table 7. Evaluation Metrics for Random Forest

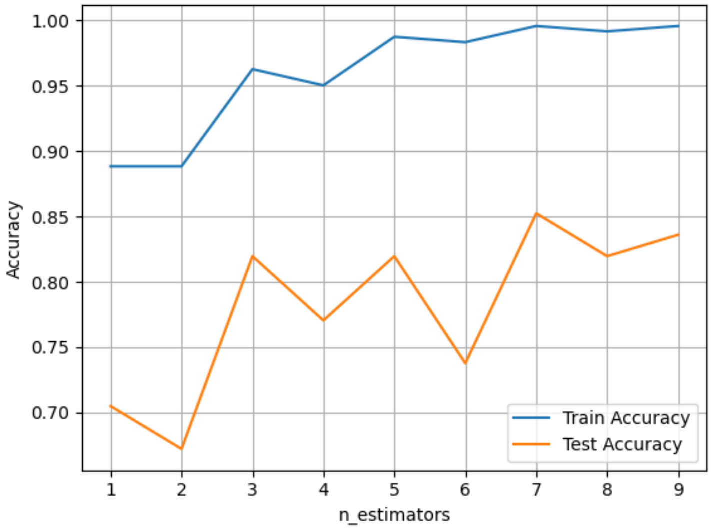


Figure 28. Training and testing accuracy for the Random Forest model for Experiment 2

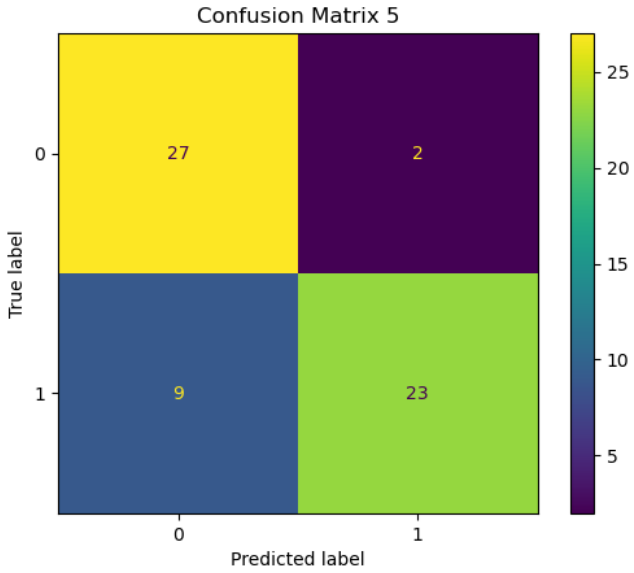


Figure 29. Random Forest Confusion Matrix for Experiment 2

Precision: During prediction, when the model predicted cases with no CVD, it was correct 77% of the time and when it predicted cases with CVD, it was correct 92% of the time. Thus, while the model demonstrates a higher precision for cases with CVD, the model performed fairly in predicting cases without CVD.

Recall: the model was able to identify 75% of the cases with CVD and 93% of the cases without CVD, which indicates a higher rate of identifying cases without CVD, showing that the model is more effective in detecting non-CVD cases than cases with CVD.

The F1-score reveals that the model performed well at identifying and predicting CVD cases compared to cases without CVD showing a balanced measure of performance.

The confusion matrix in Figure 29 shows that the model was able to predict correctly 27 positive cases and 23 negative cases. However, failed to correctly identify 2 positive cases and 9 negative cases which is not a bad performance. This model was also able to achieve a high accuracy score of 81%.

3. Support vector machine (RBF):

	Precision	Recall	F1-score	Accuracy
Class 0	0.84 (84%)	0.93 (93%)	0.89 (89%)	
Class 1	0.93 (93%)	0.84 (84%)	0.89 (89%)	
Weighted average	0.89 (89%)	0.89 (89%)	0.89 (89%)	
				0.88 (88%)

Table 14. Evaluation Metrics for SVM Experiment 2

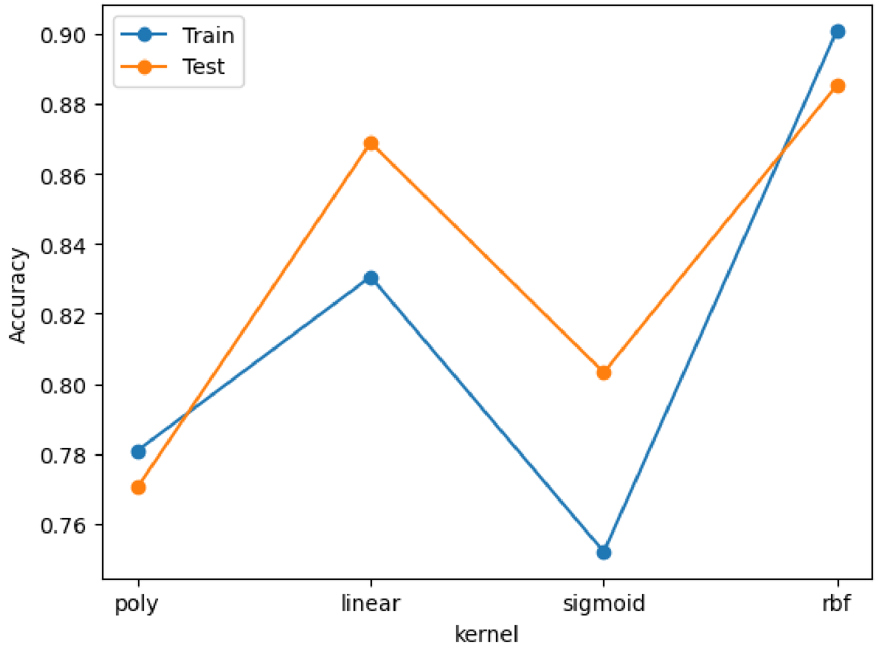


Figure 30. Training and testing accuracy for the SVM model Experiment 2

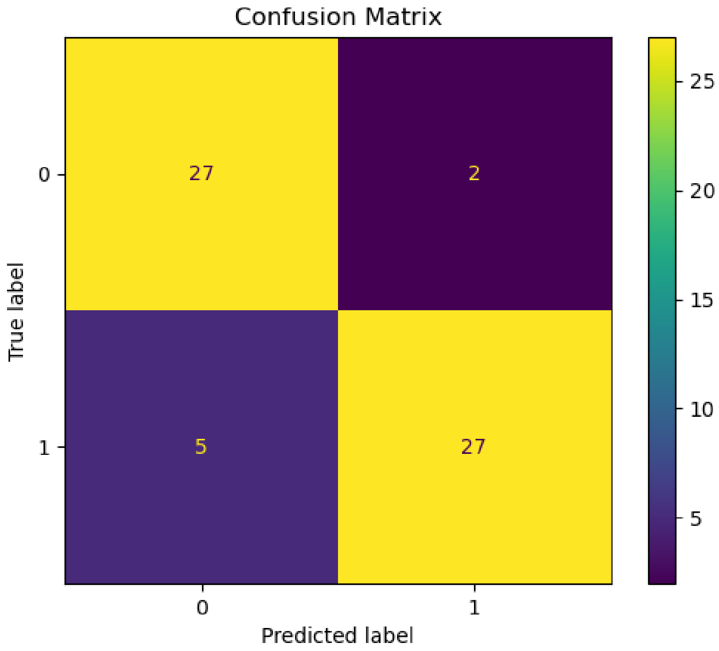


Figure 31. SVM Confusion Matrix Experiment 2

The result in Table 14 shows that during prediction, when the model predicted cases with no CVD, it was correct 84% of the time and when it predicted cases with CVD, it was correct 93% of the time. The model performed well in predicting both classes given their precision score.

Recall: the model was able to identify 84% of the cases with CVD and 93% of the cases without CVD, showing that the model is more effective in detecting non-CVD cases than cases without CVD.

The F1-score shows that the model performed well in identifying and predicting CVD cases and cases without CVD.

From the confusion matrix in Figure 31, 27 instances of class 0 were correctly predicted as class 0 while 2 instances of class 0 were incorrectly predicted as class 1 and 27 instances of class 1 were predicted correctly as class 1 while 5 instances of class 1 were incorrectly predicted as class 0. Overall this model performed better than the other models with the precision score, recall, f1-score and a higher accuracy score of 88%.

4. Naive Bayes:

	Precision	Recall	F1-score	Accuracy
Class 0	0.74 (74%)	0.90 (90%)	0.81 (81%)	
Class 1	0.88 (88%)	0.72 (72%)	0.79 (79%)	
Weighted average	0.82 (82%)	0.80 (80%)	0.80 (80%)	
				0.80 (80%)

Table 15. Evaluation Metrics for Naive Bayes Model Experiment 2

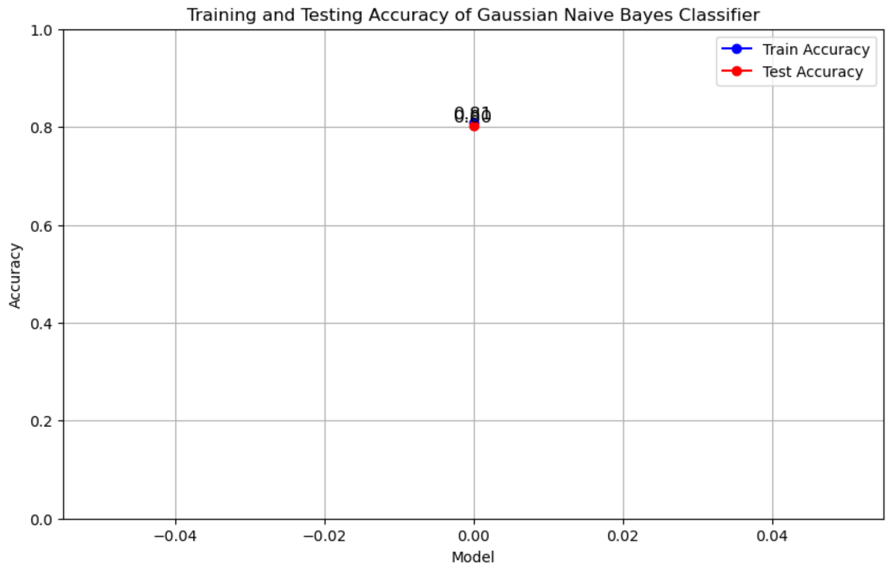


Figure 32. Training and testing accuracy for the Naive Bayes Model Experiment 2

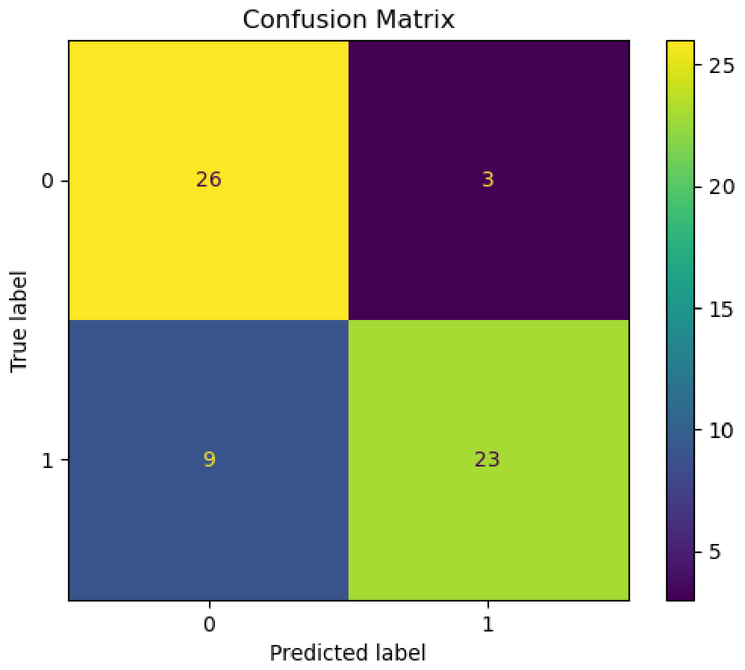


Figure 33. Naive Bayes Model Confusion Matrix Experiment 2

The precision score reveals that when the model predicted cases with no CVD, it was correct 74% of the time and when it predicted cases with CVD, it was correct 88% of the time. The model performed well in predicting both classes given their precision score but was better at predicting cases with CVD.

Recall: the model was able to identify 72% of the cases with CVD and 90% of the cases without CVD, which once again indicates a higher rate of identifying negative cases, showing that the model is more effective in detecting non-CVD cases than cases with CVD.

The F1-score shows that the model performed well in identifying and predicting both cases.

The confusion matrix shows that the model was able to predict correctly 26 positive cases and 23 negative cases. However, identified incorrectly 9 positive cases as negative cases and 3 negative cases as positive cases which is not a very bad performance compared to other models.

5. Ensemble learning: The combined result of these experiments is shown below:

	Precision	Recall	F1-score	Accuracy
Class 0	0.79 (79%)	0.93 (93%)	0.86 (86%)	
Class 1	0.93 (93%)	0.78 (78%)	0.85 (85%)	
Weighted average	0.86 (86%)	0.85 (85%)	0.85 (85%)	
				Train accuracy: 90% Test Accuracy: 85%

Table 16. Ensemble learning of all four models combined for experiment 2

	Precision	Recall	F1-score	Accuracy
Class 0	0.77 (77%)	0.93(93%)	0.84 (84%)	
Class 1	0.92 (92%)	0.75 (75%)	0.83 (83%)	
Weighted average	0.85 (85%)	0.84 (84%)	0.84 (84%)	
				Train accuracy: 94% Test Accuracy: 83%

Table 17. Ensemble learning of random forest and SVM for experiment 2

The results shown in Tables above reveal that the ensemble learning approach, when using all four models, yielded high accuracy scores however SVM alone outperformed these approaches. Moreso, the precision and recall scores were higher in the ensemble model compared to the individual models again excluding SVM.

6. Hyperparameter tuning on the two best-performing models Random Forest and Support Vector Machine using random search:

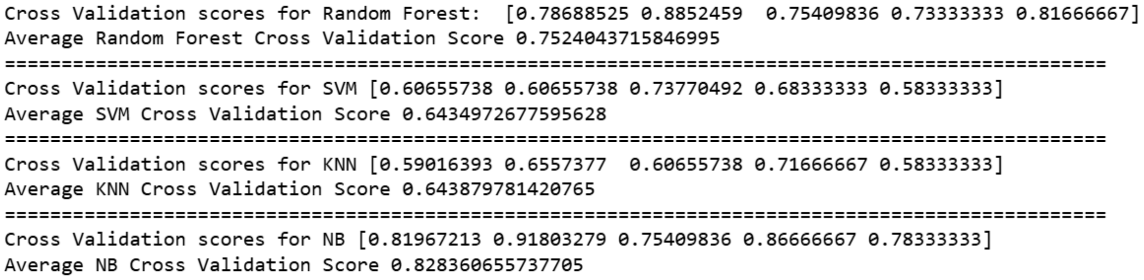
Tuned Model	Best Parameters	Precision	Recall	F1-Score	Accuracy
Random Forest	Best Parameters: {'n_estimators':	Class 0: 93%	Class 0: 79%	Class 0: 86%	83.61%

	150, 'max_leaf_nodes': 9, 'max_features': 'sqrt', 'max_depth': 3}	Class 1: 78% Weighted Avg: 86%	Class 1: 93% Weighted Avg: 85%	Class 1: 85% Weighted Avg: 85%	
SVM	Best Parameters: {'kernel': 'rbf', 'gamma': 0.01, 'C': 100}	Class 0: 74% Class 1: 88% Weighted Avg: 82%	Class 0: 90% Class 1: 72% Weighted Avg: 80%	Class 0: 81% Class 1: 79% Weighted Avg: 80%	86.89%

Table 18. Hyperparameter tuning results for experiment 2

Table 18 shows random forest tuned performed better and SVM performed well however there was a discrepancy across all evaluation metrics compared to its initial results.

7. Cross-validation gives a more realistic performance of the model. The cross-validation score is shown in Figure 34 below.



```
Cross Validation scores for Random Forest:  [0.78688525 0.8852459  0.75409836 0.73333333 0.81666667]
Average Random Forest Cross Validation Score 0.7524043715846995
=================================================================================================
Cross Validation scores for SVM [0.60655738 0.60655738 0.73770492 0.68333333 0.58333333]
Average SVM Cross Validation Score 0.6434972677595628
=================================================================================================
Cross Validation scores for KNN [0.59016393 0.6557377  0.60655738 0.71666667 0.58333333]
Average KNN Cross Validation Score 0.643879781420765
=================================================================================================
Cross Validation scores for NB [0.81967213 0.91803279 0.75409836 0.86666667 0.78333333]
Average NB Cross Validation Score 0.828360655737705
```

Figure 34. Cross-validation score for all four models for experiment 2

The cross-validation scores for random forest and naive bayes show the models are performing well and that of KNN and SVM show very low scores. These models from the result in Table show inconsistent performance across all 5 folds.

4.4 Comparative Analysis Of Experiment 1 And Experiment 2

Experiment 1					
Models	Precision	Recall	F1-score	Accuracy	Avg Cross-val-score
RF	80	76	75	74	79
KNN	75	72	70	72	63
SVM	73	72	71	72	60
NB	77	76	76	76	83
RF+KNN+SVM+NB	75	74	73	74	-
RF+NB	79	78	78	78	-
RF tuned	84	83	83	78.26	-
NB tuned	84	74	76	76.09	-
Experiment 2					
KNN	81	80	80	78	75
RF	85	84	84	81	64
SVM	89	89	89	88	64
NB	82	80	80	80	82
RF+KNN+SVM+NB	86	85	85	85	-
RF+SVM	85	84	84	83	-
RF tuned	86	85	85	83	-
SVM tuned	82	80	80	86	-

Table 19. Summary table of the results of each model on both experiments

Table 19 shows the results of the models and classifiers on the dataset. It gives the performances of various models measured in terms of precision, recall, F1-score and accuracy. Comparing the results of the proposed technique which is experiment 1 against experiment 2, experiment 1 had an overall fair performance however experiment 2 outperformed it. In the overall experiment, the highest performing model is SVM although it performed poorly in the first experiment it performed significantly well in experiment two. Additionally using the confusion matrix in section 4.2 of this chapter experiment 2 performed better in the prediction of CVD than experiment 1. This demonstrates that removing outliers can still give a fair performance but working with them gives a better performance.

4.5 Observations From the Experiment

- In section 4.2, the test and train accuracy of the models in experiment 1 and random forest in experiment 2, notice that there is a significant amount of difference between them.
- Considering the cross-validation score gotten also, experiment 1 shows that models are training well but are not performing effectively on the data. Experiment 2 has a similar issue, but it also shows some inconsistencies across the folds in each model and with only naive bayes performing better than the others that performed well initially.
- Notice that in experiment 2, although the ensemble learning approach yielded high scores, SVM did better on its own.

All of these could be an issue of overfitting, and variance in the dataset or can also be how best these models could perform. However, for certainty and clarity, it is better to investigate these issues.

5 Project Management

5.1 Project Schedule

Project management is important when trying to achieve a goal. This project has different aspects and these steps have been set to a particular timeline to enable proper work efficiency and planning. Figure 35 shows the project schedule represented in a Gantt chart. Each task in the chart has a start and end date. The timeline on these charts was strictly adhered to, although there were some adjustments due to unplanned challenges faced during implementation. For instance, during the course of this project, additional time was allocated during the design and implementation phases to ensure all aspects met the required standards, and more recent literature had to be included. Some time was also taken off due to health reasons, which caused a slight delay in the original schedule (See Appendix A).

Additionally, during documentation, some challenges arose which have been taken note of, these challenges require more investigation and certain methods to be implemented, however, due to time these investigations could not be implemented.

5.2 Risk Management

This research is limited to training and testing the heart disease dataset on the Mendeley website using machine learning algorithms. However, during this project work, some factors could limit the carrying out of this project.
They include:

- Data loss: data loss can occur if the laptop used for this project experiences any form of issue. To mitigate this risk other storage options have been put in place to store data used throughout this study like cloud backup and if there is dataset loss, since the dataset used is publicly available it can be downloaded again. This risk did not materialise.
- Time: The time constraint prevents having an in-depth study and analysis of the subject matter; however, the study has been limited to enable as much study as possible. This risk materialised as there were some unforeseen problems and findings made during evaluation and reporting, to manage this, the aim of the project was first achieved, these observations were taken note of and suggestions on what can be done have been taken note of as a future work.

5.3 Quality Management

The following steps were adopted to ensure this project is of good quality:

- Use of good software for programming.
- In-depth research to gain a good knowledge of previous work done.
- Importation of the correct libraries for the methods adopted.
- Regular meetings with the supervisor and regular feedback.
- Implementation of correct evaluation metrics.
- Good and quiet working environment to ensure a high level of concentration.

5.4 Social, Legal, Ethical and Professional Considerations

The study uses patient healthcare data to create predictive models for CVD, adhering to ethical and legal principles, ensuring privacy, confidentiality, and consent, and adhering to GDPR rules is important, although the data is publicly sourced, the data has no sensitive information and using this dataset has not raised any issues. For credibility, the appropriate sources have been cited and given due credit.

6 Critical Appraisal

The research design incorporates machine learning methods and provides a comprehensive approach to addressing the research questions and aim of the study. The methodology chosen for this study is relevant to the research objectives ensuring that the data obtained is used for analysis. The application of different methods made the research complex, while the methodology is robust and produced different insights there was a time limitation that restricted exploring some of these insights. The data collection was not difficult as the data used is a publicly sourced data available for research works. Ethical guidelines of consent, excluding sensitive information are followed to protect patient's rights and appropriate citing and referencing was done to acknowledge the source of the dataset. Although the dataset is good for study and helped greatly in this experiment, it is small, and the size affected a part of the experiment.

A thorough analysis was carried out using Python programming and appropriate evaluation metrics were adopted to analyse and interpret the results of the experiment. The experiment yielded some unexpected results which made interpretation of the overall result challenging requiring more research to avoid misinterpretation which can be time-consuming. Despite these challenges, the results were presented in detail, clear and concise. As earlier stated, some results were unexpected and yielded some more insights which suggest the need for further research for a better understanding of these outcomes. Overall, the research provides clear insights into achieving the aim of the study and contributing to the advancement of knowledge in machine learning. The experience has added to my development as a data scientist equipping me with skills such as research skills, analytical skills, and critical thinking and has enhanced my problem-solving skills as these skills are important for my success in the future.

7 Conclusions

7.1 Achievements

Considering the main aim of the study is to find the model that will effectively predict CVD disease or identify those at risk after conducting in-depth exploratory analysis which involves handling patterns or removing outliers if found, it can be concluded that the project successfully attempted this approach fairly in experiment 1. Even though it didn’t achieve very high scores it was able to predict cases of CVD. With the precision score for each model ranging from 70 to 80. Random Forest achieved a precision score of 80% and Naive Bayes achieved a precision score of 77%, when optimized using hyperparameter tuning they both achieved a high precision score of 84%, this shows that although they may not have performed well in the accuracy score, they were able to predict these cases which is fairly good.

Comparing it to experiment 2 which shows how different machine learning techniques predict CVD, the models performed well and achieved better scores after optimization. SVM had the highest precision score and accuracy score or 89% and 88% respectively and the second-best performing model was the ensemble method of combining the four models (RF, SVM, NB & KNN) with a precision score and accuracy score of 86% and 85% respectively.

In conclusion, the results obtained from Table 19 show that optimizing models can enhance the performance of these models.

7.2 Limitation and Future Work

Although the work achieved some of its objectives there is always room for improvement which can be recommended for future work. With the result obtained from experiments 1 and 2 in Table 19, there were some concerns raised that the limited time of the project did not allow it to be investigated. Firstly, for experiment 1, the proposed method showed fair results however from checking the difference between the train accuracy, test accuracy and cross-validation score overfitting was observed and confirmed by employing cross-validation. It is suggested that outliers should not be removed as it shows it had a huge impact on this dataset. Secondly, experiment 2 showed that the Random forest model did not test well and this might be an issue of overfitting also, SVM-tuned had a decrease in the result, the SVM model without tuning performed better than the ensemble learning approach and optimization approach and with cross-validation Naive Bayes seemed to perform better than the other models after cross-validation. To investigate these concerns, this study suggests implementing other ensemble learning approaches such as stacking, boosting etc., and performing cross-validation in hyperparameter tuning to curb the overfitting if any and produce the best score for each model.

8 Student Reflections

This project was quite tasking and challenging however it was an opportunity to apply the knowledge I had gained during my master's program to solving practical problems and it was also an opportunity to push myself to enhance my skills and knowledge which I can apply in my future career. This project has also taught me the importance of time management, completing a big part of this project with the amount of time allocated was challenging but this was achievable because of proper planning.

One major challenge I faced during this time was in the practical part of this project and analysing the outcomes because of the inconsistencies I found in the results due to the approach I applied. However, this challenge opened a whole world of knowledge through research and with the help I gained I was able to handle the inconsistencies, interpret my results properly and finish my project successfully. Although there are more things to consider I am satisfied with the extent I was able to achieve within this time and I look forward to doing more research in the future.

Bibliography and References

Adhishayaa, P. V., Gomathi, V., & Mahendran, K. (2023). Review On Cardiovascular Disease Prediction Using Machine Learning Algorithm. *2023 International Conference on Computer Communication and Informatics (ICCCI)*. https://doi.org/10.1109/iccci56745.2023.10128403

Aftab, R. S. (2024). Prediction of heart attack. *Mendeley Data*. https://doi.org/10.17632/yrwd336rkz.1

Ahdal, A. A., Rakhra, M., Rajendran, R. R., Arslan, F., Khder, M. A., Patel, B., Rajagopal, B. R., & Jain, R. (2023). Monitoring cardiovascular problems in heart patients using machine learning. *Journal of Healthcare Engineering*, *2023*, 1–15. https://doi.org/10.1155/2023/9738123

Ahmed, R., Bibi, M., & Syed, S. (2023). Improving Heart Disease Prediction Accuracy Using a Hybrid Machine Learning Approach: A Comparative study of SVM and KNN Algorithms. *International Journal of Computations, Information and Manufacturing*, *3*(1), 49–54. https://doi.org/10.54489/ijcim.v3i1.223

Alanazi, A. (2022). Using machine learning for healthcare challenges and opportunities. *Informatics in Medicine Unlocked*, *30*, 100924. https://doi.org/10.1016/j.imu.2022.100924

Allheeib, N., Kanwal, S., & Alamri, S. (2023). An intelligent heart disease prediction framework using machine learning and deep learning techniques. *International Journal of Data Warehousing and Mining*, *19*(1), 1–24. https://doi.org/10.4018/ijdwm.333862

Alotaibi, F. S. (2019). Implementation of machine learning model to predict heart failure disease. *International Journal of Advanced Computer Science and Applications*, *10*(6). https://doi.org/10.14569/ijacsa.2019.0100637

An, Q., Rahman, S., Zhou, J., & Kang, J. J. (2023). A Comprehensive review on machine learning in healthcare industry: Classification, restrictions, opportunities and challenges. *Sensors*, *23*(9), 4178. https://doi.org/10.3390/s23094178

Baghdadi, N. A., Abdelaliem, S. M. F., Malki, A., Gad, I., Ewis, A., & Atlam, E. (2023). Advanced machine learning techniques for cardiovascular disease early detection and diagnosis. *Journal of Big Data*, *10*(1). https://doi.org/10.1186/s40537-023-00817-1

Bhatt, C., Patel, P., Ghetia, T., & Mazzeo, P. L. (2023). Effective heart disease prediction using machine learning techniques. *Algorithms*, *16*(2), 88. https://doi.org/10.3390/a16020088

Biswas, N., Ali, M. M., Rahaman, M. A., Islam, M., Mia, M. R., Azam, S., Ahmed, K., Bui, F. M., Al-Zahrani, F. A., & Moni, M. A. (2023). Machine Learning-Based model to predict heart disease in early stage employing different feature selection techniques. *BioMed Research International*, *2023*, 1–15. https://doi.org/10.1155/2023/6864343

Boateng, E. Y., Otoo, J., & Abaye, D. A. (2020). Basic Tenets of Classification Algorithms K-Nearest-Neighbor, Support Vector Machine, random Forest and Neural network: a review. *Journal of Data Analysis and Information Processing*, *08*(04), 341–357. https://doi.org/10.4236/jdaip.2020.84020

Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., & Lopez, A. (2020). A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*, *408*, 189–215. https://doi.org/10.1016/j.neucom.2019.10.118

Chandrasekhar, N., & Peddakrishna, S. (2023). Enhancing Heart Disease Prediction Accuracy through Machine Learning Techniques and Optimization. *Processes*, *11*(4), 1210. https://doi.org/10.3390/pr11041210

Dahouda, M. K., & Joe, I. (2021). A Deep-Learned embedding technique for categorical features encoding. *IEEE Access*, *9*, 114381–114391. https://doi.org/10.1109/access.2021.3104357

Dalal, S., Goel, P., Onyema, E. M., Alharbi, A., Mahmoud, A., Algarni, M. A., & Awal, H. (2023). Application of machine learning for cardiovascular disease risk prediction. *Computational Intelligence and Neuroscience*, *2023*, 1–12. https://doi.org/10.1155/2023/9418666

Dash, C. S. K., Behera, A. K., Dehuri, S., & Ghosh, A. (2023). An outliers detection and elimination framework in classification task of data mining. *Decision Analytics Journal*, *6*, 100164. https://doi.org/10.1016/j.dajour.2023.100164

Deepa, R., Sadu, V. B., C, P. G., & Sivasamy, A. (2024). Early prediction of cardiovascular disease using machine learning: Unveiling risk factors from health records. *AIP Advances*, *14*(3). https://doi.org/10.1063/5.0191990

Fatima, A., Nazir, N., & Khan, M. G. (2017). Data Cleaning in Data Warehouse: A survey of data pre-processing techniques and tools. *International Journal of Information*

Technology and Computer Science, *9*(3), 50–61. https://doi.org/10.5815/ijitcs.2017.03.06

Gonsalves, A. H., Thabtah, F., Mohammad, R. M. A., & Singh, G. (2019b). Prediction of Coronary Heart Disease using Machine Learning. *ICDLT '19: Proceedings of the 2019 3rd International Conference on Deep Learning Technologies*. https://doi.org/10.1145/3342999.3343015

Gulati, S., Guleria, K., & Goyal, N. (2022). Classification and Detection of Coronary Heart Disease using Machine Learning. *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*. https://doi.org/10.1109/icacite53722.2022.9823547

Habehh, H., & Gohel, S. (2021). Machine learning in healthcare. *Current Genomics*, *22*(4), 291–300. https://doi.org/10.2174/1389202922666210705124359

Hridoy, K. M., Akash, S. I., Dipti, F. A., Hasan, M., Ovi, J. A., Al-Imran, M., & Jahan, S. (2023). Heart Disease Prediction Using Machine Learning Algorithms. *2023 4th International Conference on Big Data Analytics and Practices (IBDAP)*. https://doi.org/10.1109/ibdap58581.2023.10271997

Jawalkar, A. P., Swetcha, P., Manasvi, N., Sreekala, P., Aishwarya, S., Bhavani, P. K. D., & Anjani, P. (2023). Early prediction of heart disease with data analysis using supervised learning with stochastic gradient boosting. *Journal of Engineering and Applied Science*, *70*(1). https://doi.org/10.1186/s44147-023-00280-y

Javaid, M., Haleem, A., Singh, R. P., Suman, R., & Rab, S. (2022). Significance of machine learning in healthcare: Features, pillars and applications. *International Journal of Intelligent Networks*, *3*, 58–73. https://doi.org/10.1016/j.ijin.2022.05.002

Jia, W., Sun, M., Lian, J., & Hou, S. (2022). Feature dimensionality reduction: a review. *Complex & Intelligent Systems*, *8*(3), 2663–2693. https://doi.org/10.1007/s40747-021-00637-x

Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, *6*(1). https://doi.org/10.1186/s40537-019-0192-5

Khan, Y., Qamar, U., Yousaf, N., & Khan, A. (2019). Machine Learning Techniques for Heart Disease Datasets. *ICMLC '19: Proceedings of the 2019 11th International Conference on Machine Learning and Computing*. https://doi.org/10.1145/3318299.3318343

Latha, C. B. C., & Jeeva, S. (2019). Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. *Informatics in Medicine Unlocked*, *16*, 100203. https://doi.org/10.1016/j.imu.2019.100203

Louridi, N., Amar, M., & Ouahidi, B. E. (2019). Identification of Cardiovascular Diseases Using Machine Learning. *2019 7th Mediterranean Congress of Telecommunications (CMT)*. https://doi.org/10.1109/cmt.2019.8931411

Marimuthu, M., Abinaya, M., Hariesh, K. S., Madhankumar, K., & Pavithra, V. (2018). A Review on Heart Disease Prediction using Machine Learning and Data Analytics Approach. *International Journal of Computer Applications*, *181*(18), 20–25. https://doi.org/10.5120/ijca2018917863

Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access*, *7*, 81542–81554. https://doi.org/10.1109/access.2019.2923707

Muntean, M., & Militaru, F. (2023). Metrics for evaluating classification algorithms. In *Smart innovation, systems and technologies* (pp. 307–317). https://doi.org/10.1007/978-981-19-6755-9_24

Nagavelli, U., Samanta, D., & Chakraborty, P. (2022). Machine Learning Technology-Based Heart Disease Detection Models. *Journal of Healthcare Engineering*, *2022*, 1–9. https://doi.org/10.1155/2022/7351061

Nayyar, A., Gadhavi, L., & Jhanjhi, N. Z. (2021). Machine learning in healthcare: review, opportunities and challenges. In *Elsevier eBooks* (pp. 23–45). https://doi.org/10.1016/b978-0-12-821229-5.00011-2

Ogunpola, A., Saeed, F., Basurra, S., Albarrak, A. M., & Qasem, S. N. (2024). Machine Learning-Based Predictive Models for detection of cardiovascular diseases. *Diagnostics*, *14*(2), 144. https://doi.org/10.3390/diagnostics14020144

Pal, M., Parija, S., Panda, G., Dhama, K., & Mohapatra, R. K. (2022). Risk prediction of cardiovascular disease using machine learning classifiers. *Open Medicine*, *17*(1), 1100–1113. https://doi.org/10.1515/med-2022-0508

Pandey, S. (2023). The Cardiovascular Disease Prediction using Machine Learning. *Buana Information Technology and Computer Sciences :*, *4*(1), 24–27. https://doi.org/10.36805/bit-cs.v4i1.3060

Roy, A., & Chakraborty, S. (2023). Support vector machine in structural reliability analysis: A review. *Reliability Engineering & Systems Safety*, *233*, 109126. https://doi.org/10.1016/j.ress.2023.109126

Schlosser, T., Friedrich, M., Meyer, T., & Kowerko, D. (2024). A Consolidated Overview of Evaluation and Performance Metrics for Machine Learning and Computer Vision. *Tobias Schlosser, Michael Friedrich, Trixy Meyer, and Danny Kowerko–*

Junior Professorship of Media Computing, Chemnitz University of Technology, *9107*.

Sharma, S., & Parmar, M. (2020). Heart Diseases Prediction using Deep Learning Neural Network Model. *International Journal of Innovative Technology and Exploring Engineering*, *9*(3), 2244–2248. https://doi.org/10.35940/ijitee.c9009.019320

Taylan, O., Alkabaa, A. S., Alqabbaa, H. S., Pamukçu, E., & Leiva, V. (2023). Early Prediction in Classification of Cardiovascular Diseases with Machine Learning, Neuro-Fuzzy and Statistical Methods. *Biology*, *12*(1), 117. https://doi.org/10.3390/biology12010117

Tiwari, K., Ansari, S. A., & Kushwaha, G. (2023). Earlier heart disease prediction system using machine learning. In *Algorithms for intelligent systems* (pp. 191–204). https://doi.org/10.1007/978-981-19-7892-0_16

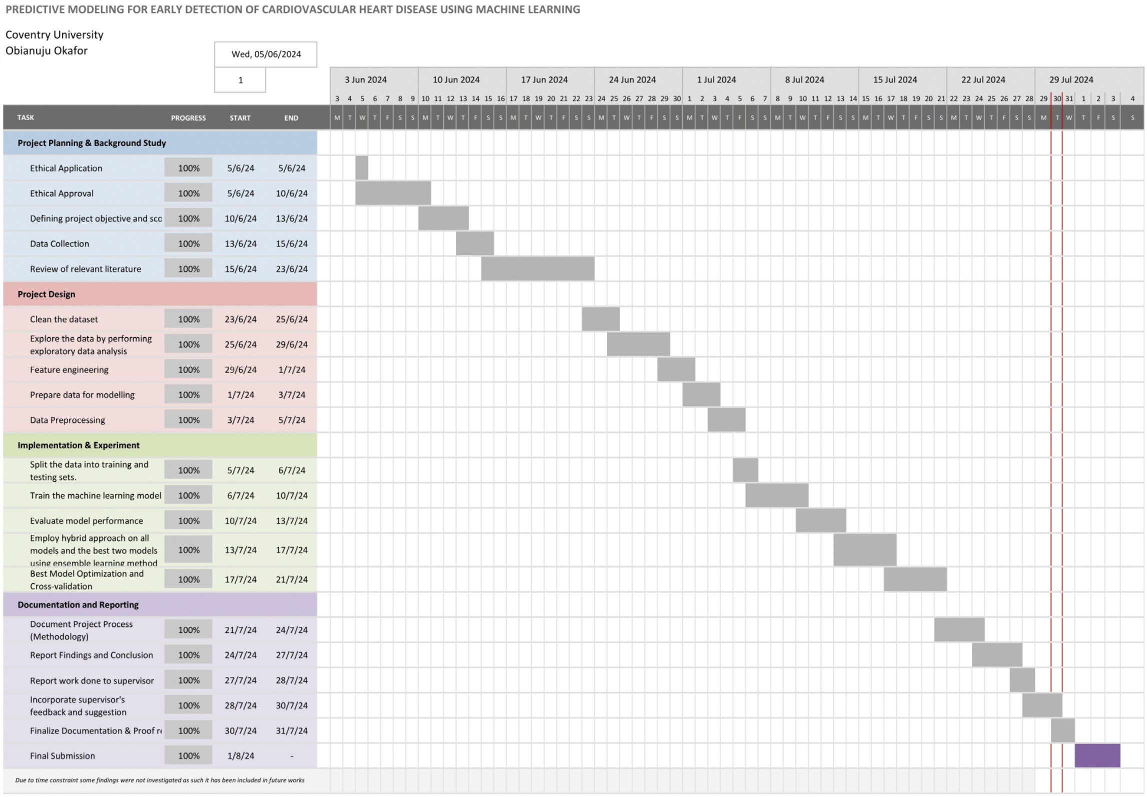
Viet, T. N., Minh, H. L., Hieu, L. C., & Anh, T. H. (2021). THE NAIVE BAYES ALGORITHM FOR LEARNING DATA ANALYTICS. *Indian Journal of Computer Science and Engineering*, *12*(4), 1038–1043. https://doi.org/10.21817/indjcse/2021/v12i4/211204191

World Health Organization: WHO. (2021, June 11). *Cardiovascular diseases (CVDs)*. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

Yılmaz, R., & Yağın, F. H. (2022). Early detection of coronary heart disease based on machine learning methods. *Medical Records-international Medical Journal*, *4*(1), 1–6. https://doi.org/10.37990/medr.1011924

Zebari, R., Abdulazeez, A., Zeebaree, D., Zebari, D., & Saeed, J. (2020). A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction. *Journal of Applied Science and Technology Trends*, *1*(2), 56–70. https://doi.org/10.38094/jastt1224

Appendix A – Project Gantt Chart



PREDICTIVE MODELING FOR EARLY DETECTION OF CARDIOVASCULAR HEART DISEASE USING MACHINE LEARNING

Coventry University
Obianuju Okafor

Wed, 05/06/2024

1

TASK	PROGRESS	START	END
Project Planning & Background Study			
Ethical Application	100%	5/6/24	5/6/24
Ethical Approval	100%	5/6/24	10/6/24
Defining project objective and sco	100%	10/6/24	13/6/24
Data Collection	100%	13/6/24	15/6/24
Review of relevant literature	100%	15/6/24	23/6/24
Project Design			
Clean the dataset	100%	23/6/24	25/6/24
Explore the data by performing exploratory data analysis	100%	25/6/24	29/6/24
Feature engineering	100%	29/6/24	1/7/24
Prepare data for modelling	100%	1/7/24	3/7/24
Data Preprocessing	100%	3/7/24	5/7/24
Implementation & Experiment			
Split the data into training and testing sets.	100%	5/7/24	6/7/24
Train the machine learning model	100%	6/7/24	10/7/24
Evaluate model performance	100%	10/7/24	13/7/24
Employ hybrid approach on all models and the best two models using ensemble learning method	100%	13/7/24	17/7/24
Best Model Optimization and Cross-validation	100%	17/7/24	21/7/24
Documentation and Reporting			
Document Project Process (Methodology)	100%	21/7/24	24/7/24
Report Findings and Conclusion	100%	24/7/24	27/7/24
Report work done to supervisor	100%	27/7/24	28/7/24
Incorporate supervisor's feedback and suggestion	100%	28/7/24	30/7/24
Finalize Documentation & Proof r	100%	30/7/24	31/7/24
Final Submission	100%	1/8/24	-

Due to time constraint some findings were not investigated as such it has been included in future works

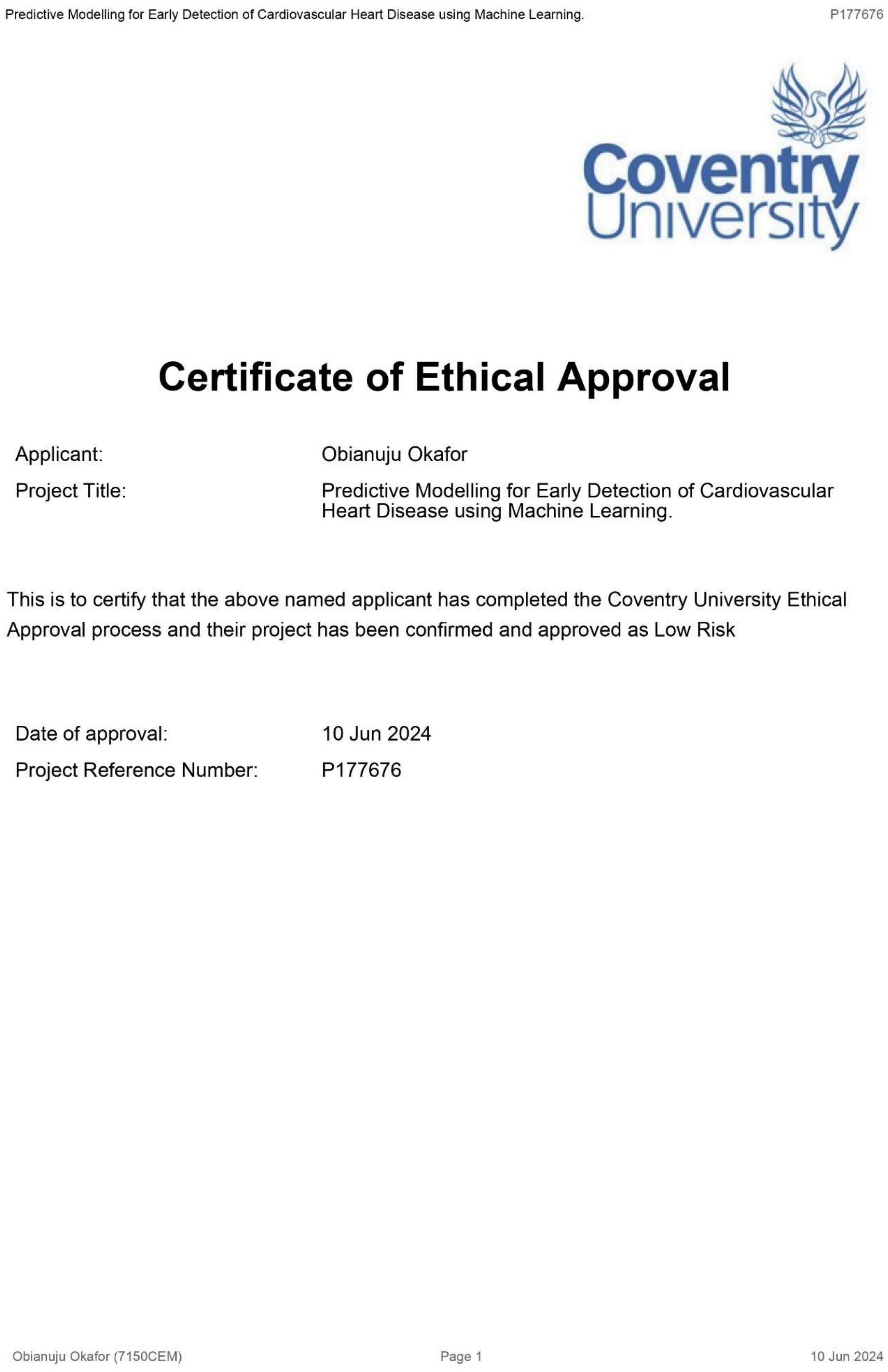
Figure 35. Project schedule represented in a Gantt chart.

Appendix B – Interim Progress Report and Meeting Records

Meeting No	Update Report	Date
1	• Introductions • Introducing the project, the aims and objective • Meeting time agreement Next work to do: start working on Chapter 1 and the literature review before the next meeting.	30/05/2024
2	• Update report on work done on my project proposal via email	03/06/2024
3	• Progress report on the literature review • Gathering resources that can help with reviewing • Discussion on previous papers related to the project • Discussion on the proposed methodology for the project Next work to do: submit the work we have started, round up the literature review and start working on our methodology	13/06/2024
4	Progress report via email of work done between 14/06/2024 to 27/06/24 • Discussion on work done • Data Collection • Data Exploration • Model Analysis Next work to do: Include existing gap in the introduction, combine Chapter 1 and Chapter 2	27/06/2024

5	Unable to attend due to health reasons	08/06/2024
6	Progress report on work done between 27/06/24 to 15/07/24 and feedback from my supervisor Supervisor feedback: she sent additional resources to be added to the literature review and ensured that I am following the template provided on Aula. Next work to do: finish up and submit the methodology	16/07/2024
7	Progress report on the literature review Submitted the methodology and the updated literature. Next work to do: start and finish the experiment and compile the reports	19/07/2024
8	Meeting via teams • Discussed the extent of my experiment and the challenges I incurred Next work to do: compile the results from the experiment.	25/07/2024
9	Final report update	31/07/2024

Appendix C – Certificate of Ethics Approval



Predictive Modelling for Early Detection of Cardiovascular Heart Disease using Machine Learning. P177676

Certificate of Ethical Approval

Applicant:	Obianuju Okafor
Project Title:	Predictive Modelling for Early Detection of Cardiovascular Heart Disease using Machine Learning.

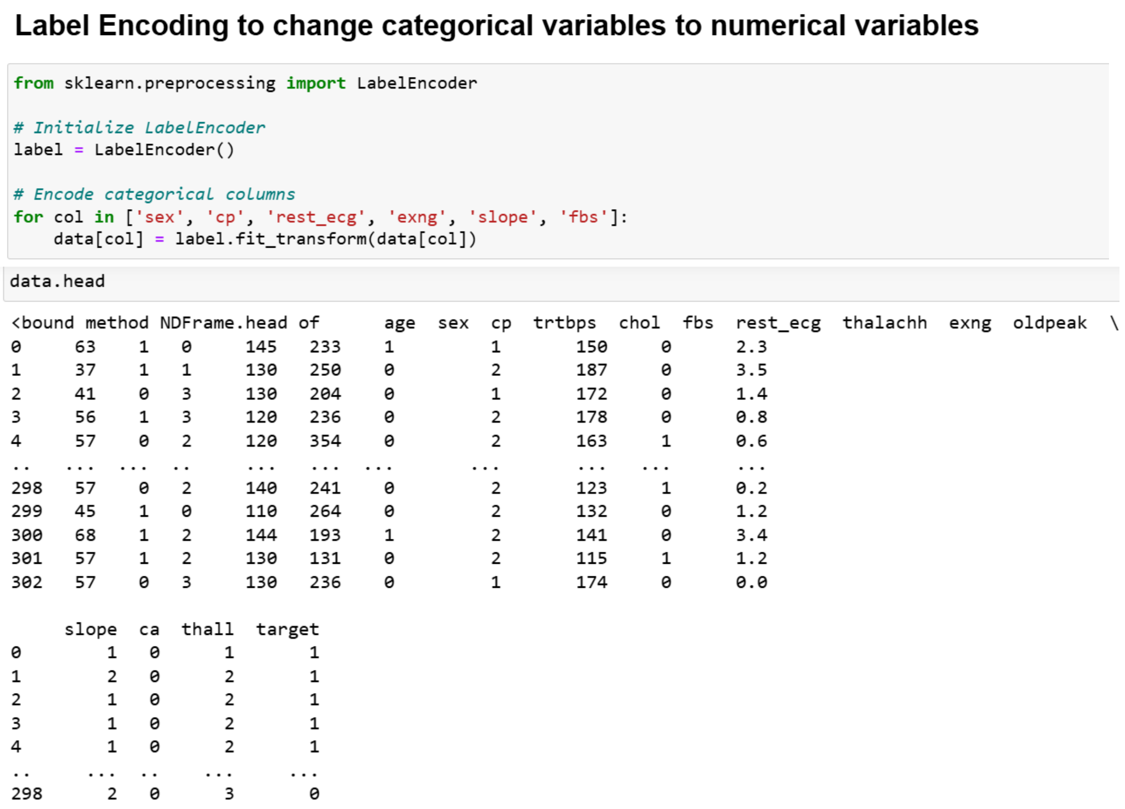
This is to certify that the above named applicant has completed the Coventry University Ethical Approval process and their project has been confirmed and approved as Low Risk

Date of approval:	10 Jun 2024
Project Reference Number:	P177676

Obianuju Okafor (7150CEM) Page 1 10 Jun 2024

Appendix D – Pictorial Representation of Label Encoding and Handling of Outliers

Label Encoding



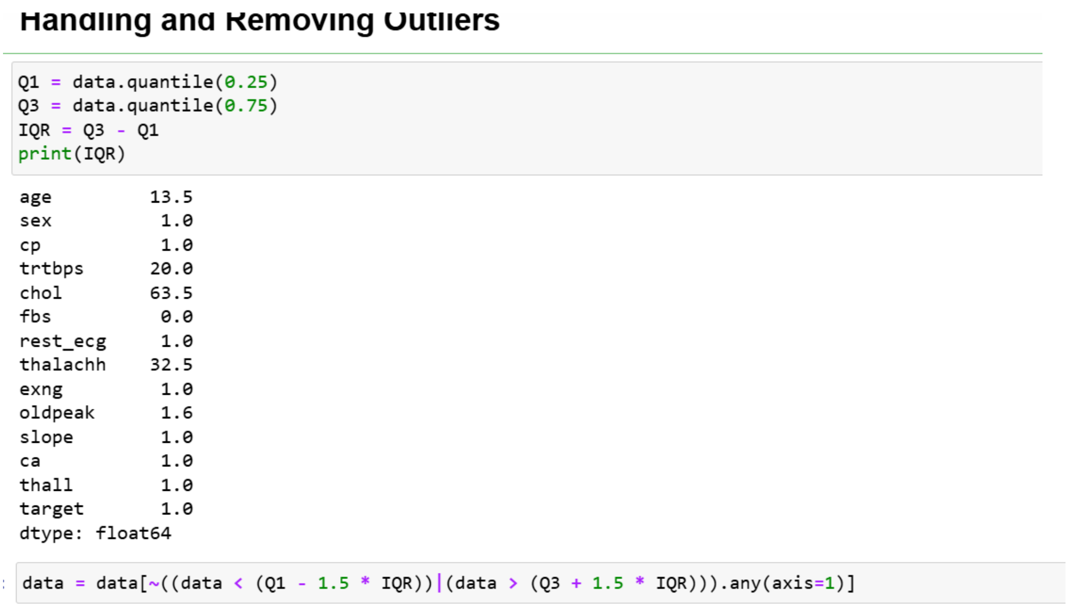
Label Encoding to change categorical variables to numerical variables

```
from sklearn.preprocessing import LabelEncoder

# Initialize LabelEncoder
label = LabelEncoder()

# Encode categorical columns
for col in ['sex', 'cp', 'rest_ecg', 'exng', 'slope', 'fbs']:
    data[col] = label.fit_transform(data[col])
```

```
data.head
```

```
<bound method NDFrame.head of      age  sex  cp  trtbps  chol  fbs  rest_ecg  thalachh  exng  oldpeak  \
0     63    1   0     145   233    1         1       150     0      2.3
1     37    1   1     130   250    0         2       187     0      3.5
2     41    0   3     130   204    0         1       172     0      1.4
3     56    1   3     120   236    0         2       178     0      0.8
4     57    0   2     120   354    0         2       163     1      0.6
..   ...  ...  ..     ...   ...  ...       ...       ...   ...      ...
298   57    0   2     140   241    0         2       123     1      0.2
299   45    1   0     110   264    0         2       132     0      1.2
300   68    1   2     144   193    1         2       141     0      3.4
301   57    1   2     130   131    0         2       115     1      1.2
302   57    0   3     130   236    0         1       174     0      0.0

     slope  ca  thall  target
0        1   0      1       1
1        2   0      2       1
2        1   0      2       1
3        1   0      2       1
4        1   0      2       1
..     ...  ..    ...     ...
298      2   0      3       0
```

Handling Outliers



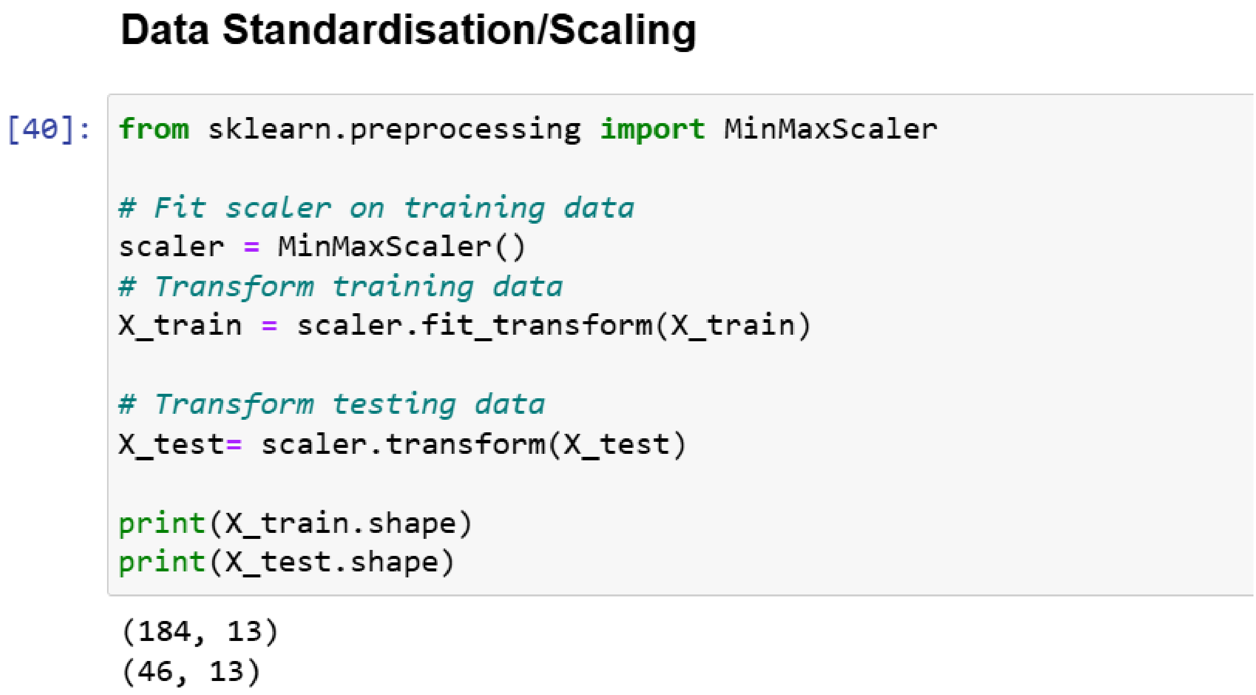
Handling and Removing Outliers

```
Q1 = data.quantile(0.25)
Q3 = data.quantile(0.75)
IQR = Q3 - Q1
print(IQR)
```

```
age         13.5
sex          1.0
cp           1.0
trtbps      20.0
chol        63.5
fbs          0.0
rest_ecg     1.0
thalachh    32.5
exng         1.0
oldpeak      1.6
slope        1.0
ca           1.0
thall        1.0
target       1.0
dtype: float64
```

```
data = data[~((data < (Q1 - 1.5 * IQR))|(data > (Q3 + 1.5 * IQR))).any(axis=1)]
```

Appendix E – Pictorial Representation of Data Scaling



Data Standardisation/Scaling

```
[40]: from sklearn.preprocessing import MinMaxScaler

      # Fit scaler on training data
      scaler = MinMaxScaler()
      # Transform training data
      X_train = scaler.fit_transform(X_train)

      # Transform testing data
      X_test= scaler.transform(X_test)

      print(X_train.shape)
      print(X_test.shape)
```

```
(184, 13)
(46, 13)
```

Appendix F – Project Code Link

https://github.com/Uju024/Predictive-Modeling-for-Early-Detection-of-Cardiovascular-Heart-Disease-Using-Machine-Learning