UNIVERSITY OF GHANA

COLLEGE OF BASIC AND APPLIED SCIENCES

SCHOOL OF MATHEMATICAL AND PHYSICAL SCIENCES

DEPARTMENT OF STATISTICS AND ACTUARIAL SCIENCE

UNIVERSITY OF GHANA

CLASSIFICATION AND PREDICTION OF HEART DISEASES USING MACHINE LEARNING ALGORITHMS

BY

AKUA SEKYIWAA OSEI-NKWANTABISA

(10711755)

A DISSERTATION SUBMITTED TO THE DEPARTMENT OF STATISTICS AND ACTUARIAL SCIENCE, UNIVERSITY OF GHANA IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE BACHELOR OF ARTS DEGREE IN MATHEMATICS AND STATISTICS

OCTOBER, 2022

DECLARATION

With the exception of citations made by writers whose work has been properly acknowledged, I therefore declare that this project is the outcome of my own research, conducted under the direction of Dr. Louis Asiedu of the Department of Statistics and Actuarial Science, University of Ghana, Legon.

I am therefore responsible for any errors that may be found in this project.

Akua Sekyiwaa Osei-Nkwantabisa	.....	.....
(Student)	Signature	Date

Certified by:

Dr. Louis Asiedu	.....	.....
(Supervisor)	Signature	Date

ACKNOWLEDGEMENT

I would like to begin by expressing my gratitude to the Almighty God for providing me with the skills and strength required to complete my assignment.

Additionally, I want to thank Dr. Louis Asiedu, my boss, for his project-related guidance and assistance.

I am grateful for the encouragement and assistance provided by my family, friends, and colleagues.

Additionally, I would like to thank the teaching assistants, Ms. Ama Konadu Appau and Mr. Andrews Kwasi Boahen, as well as my colleague and friend, Mr. Redeemer Ntumy, for their invaluable help in completing my project successfully. Without their support, I may not have been able to finish my project as successfully.

DEDICATION

I dedicate this study to my supervisor, Dr. Louis Asiedu, and to the Department of Statistics and Actuarial Science for their guidance and support throughout the research process.

Table of contents

LIST OF ABBREVIATIONS

CVDs	Cardiovascular Diseases
HRFLM	Hybrid Random Forest with Linear Model
RF	Random Forest
LM	Linear Method
DT/CART	Decision Tree
SVM	Support Vector Machine
KNN	K-Nearest Neighbor
AB	AdaBoost Classifier
LR	Logistic Regression
ET	Extra Trees Classifier
MNB	Multinomial Naïve Bayes
LDA	Linear Discriminant Analysis
XGB	XGBoost
ANN	Artificial Neural Network
MRMR	Minimal-Redundancy-Maximal-Relevance Feature Selection Algorithm
LASSO	Least Absolute Shrinkage and Selection Operator
NB	Naïve Bayes
LLBFS	Local Learning Based Features Selection
FCMIM	Feature Selection Algorithm
MLP	Multi-Layer Perceptron
AUROC	Area Under Curve Receiver Operating Characteristics
RMSE	Root Mean Squared Error
RHD	Rheumatic heart disease
ML	Machine Learning
UCI	University of California, Irvine
PCA	Principle Component Analysis
ECG	Electrocardiography
TA	Typical angina.
CAD	Coronary Artery Disease
AUC	Area Under the Curve
ROC	Receiver Operating Characteristics
ECG	Electrocardiography
IQR	Interquartile Range

List of Figures

List of Tables

ABSTRACT

Heart disease is a serious worldwide health issue because it claims the lives of many people who might have been treated if the disease had been identified sooner.The leading cause of death in the world is cardiovascular disease, usually referred to as heart disease. Creating reliable, effective, and precise predictions for these diseases is one of the biggest issues facing the medical world today.Although there are tools for predicting heart disease, they are either expensive or challenging to apply for determining a patient's risk.

The best classifier for foretelling and spotting cardiac issues was the aim of this investigation.This experiment examined a range of machine learning approaches, including Logistic Regression, K-Nearest Neighbor, Support Vector Machine, and Artificial Neural Networks, to determine which machine learning algorithm was most effective at predicting heart sickness. One of the most often utilized data sets for this purpose, the UCI heart disease repository provided the data set for this study.The K-Nearest Neighbor technique was shown to be the most effective machine learning algorithm for determining whether a patient has cardiac disease after data analysis using the Python programming language.

It will be beneficial to conduct further future study on the application of additional machine learning algorithms for heart disease prediction.

Keywords- Machine Learning, Logistic Regression, Heart Disease, Cardiovascular Disease, K-Nearest Neighbor, Support Vector Machine and Artificial Neural Networks.

Chapter 1

Introduction

1.1 Background study

The heart and blood arteries are affected by a group of ailments known as cardiovascular diseases (CVDs) and cardiac disorders. These conditions include deep vein thrombosis, coronary heart disease, peripheral arterial disease, and rheumatic heart disease.

One of our body's most important organs is the heart. It is a muscular organ found directly beneath and just to the left of the breastbone. The World Health Organization estimates that CVDs account for 17.9 million deaths annually, making them the leading cause of death worldwide. The majority of these fatalities, which make up 85% of them and happen in low- and middle-income nations, are mostly caused by heart attacks and strokes. Furthermore, persons under the age of 70 account for one-third of these early mortality.

Poor diet, inactivity, smoking, and binge drinking are a few risk factors for heart disease and stroke. These psychological risk factors can manifest physically as obesity, overweight, high blood lipid levels, high blood pressure, and high blood sugar. In order to take preventative actions and avoid the damage that these disorders can cause, early identification of cardiovascular diseases is essential.

Underlying blood vessel issues may go untreated for a long time. The condition typically manifests first as heart attacks and strokes. Heart attack symptoms include pain or discomfort in the middle of the chest, in the arms, left shoulder, elbow, jaw, or back. The person may also have back or jaw discomfort, nausea or vomiting, lightheadedness, or fainting. Common stroke symptoms include trouble speaking or understanding speech, disorientation, difficulty seeing with one or both eyes, numbness on one side of the face, arm, or leg, difficulty walking, and a severe headache with no known reason.

Due to a lack of basic healthcare facilities for early detection and treatment, cardiovascular illnesses often have higher death rates in low- and middle-income nations. People with CVDs and other non-communicable diseases have limited access to adequate, equitable healthcare services that can meet their needs in low- and middle-income countries. Due to the disease's delayed identification, people get CVDs and other non-communicable diseases and pass away early.

It has been demonstrated that lowering salt intake, increasing fruit and vegetable consumption, engaging in regular exercise, quitting smoking, and abstaining from excessive alcohol consumption all reduce the risk of cardiovascular disease. Additionally, recognizing people who are most at risk for CVDs and ensuring they receive the right care will help avoid early mortality. All primary healthcare facilities must have access to non-communicable disease medications and fundamental medical technology to guarantee that individuals in need receive treatment and counseling (World Health Organization, 2021).

Machine learning is becoming more and more popular as its value in different industries increases daily. Among the industries that apply machine learning include manufacturing, retail, healthcare, life sciences, travel and hospitality, financial services, energy, feedstock, and utilities. The healthcare sector is one of these applications' most significant industries (Akhila, Mahalakshmi, & Niriksha, 2022).

The area of machine learning is expanding quickly as a result of the massive volumes of data being collected. Many big data experts predict that the volume of data generated will continue to rise sharply in the future. IDC projects that the world's datasphere will grow to 175 zettabytes by 2025 in its Data Age 2025 research. To put this into context, the stack would have covered two-thirds of the distance between the Earth and the Moon in 2013 if we converted this amount to 128GB iPads. This stack would be 26 times longer in 2025 (Khvoynitskaya, 2020). Given this, it is crucial to comprehend data and obtain insights to comprehend the world of humans. Compared to human doctors, machine learning is more accurate and quicker at diagnosing. In the face of uncertainty, machine learning algorithms create models that make predictions based on data. These techniques train a model to produce accurate predictions using known sets of inputs and known sets of outputs.

1.2 Problem Statement

The biggest issue facing the medical industry today is making accurate and dependable predictions for disease diagnosis and treatment. Although there are techniques for forecasting cardiac disorders, they are occasionally expensive or ineffective at determining an individual's risk. Early detection of cardiovascular diseases (CVDs) can considerably reduce the chance of fatalities and other issues associated to these conditions. Using data mining and machine learning approaches, automation can help solve the issue of low prediction accuracy. Data mining searches through massive data sets using a range of computational technologies in order to identify patterns and predict outcomes. In order to address the issue of poor CVD prediction accuracy, the goal of my project is to design and put into practice a system for heart disease detection and prediction utilizing machine learning techniques.

1.3 Objectives

1.3.1 Main Objectives

Finding the best classifier for predicting and diagnosing cardiac issues is the major goal of this study. Based on factors including gender, age, cholesterol levels, and other medical traits, the goal is to ascertain whether a patient has a cardiovascular disease. The system looks for and extracts helpful insights from prior cardiac data sets to aid in the diagnosis and prevention of heart problems.

1.3.2 Specific Objectives

The specific objective is to compare models to see which one performs better than the other. It specifically seeks to:

- Determine the most effective classification system for cardiac ailments.
- Identify the key criteria for categorizing heart disorders.

1.4 Methodology

The technique used in this work aims to select the best classifier for foretelling and spotting heart issues. This would entail comparing various machine learning methods in order to predict heart disease fast based on many medical parameters, such as gender, age, and cholesterol levels, such as Logistic Regression, K-Nearest Neighbor, and Artificial Neural Networks. The study will extract and reveal hidden knowledge about the ailment through analysis of a historical heart data collection.

This project's initial step is to gather data from a dependable source, like Kaggle. The data will then be used to extract the pertinent values. After preprocessing, the data will be split into training and testing sets. The training data set will be used to test various machine learning algorithms, including Logistic Regression and Random Forest, to see which one is most effective at identifying heart illnesses. The testing data set will be used to group the data into people who have a positive or negative heart disease condition.

The methodology used in my project is shown in Figure 1.1

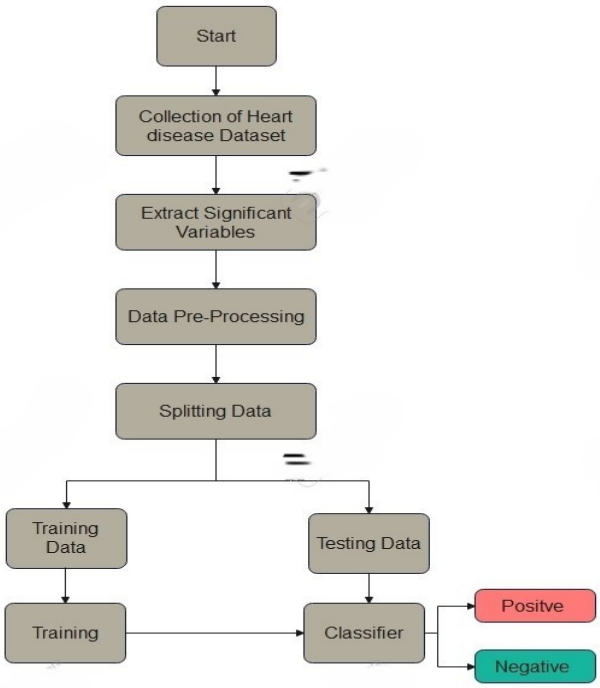

Figure 1.1: Proposed Model

1.4.1 Source of Data

Using a single data set that was obtained from Kaggle, the study's objective is to predict cardiac diseases. The data set was created by combining various Kaggle data points.

The major objective of this work is to find the most effective classifier for the prediction and diagnosis of heart problems by analyzing a sizable dataset on heart disorders. The Long Beach, Switzerland, Hungarian, Cleveland, and Statlog heart disease datasets are combined in one Kaggle-obtained dataset. 1 target variable, 11 attributes, and 1190 records make up the dataset. The method for this study involves classifying and evaluating the data in order to find trends and predict outcomes related to heart disease. These techniques include Logistic Regression and Random Forest.

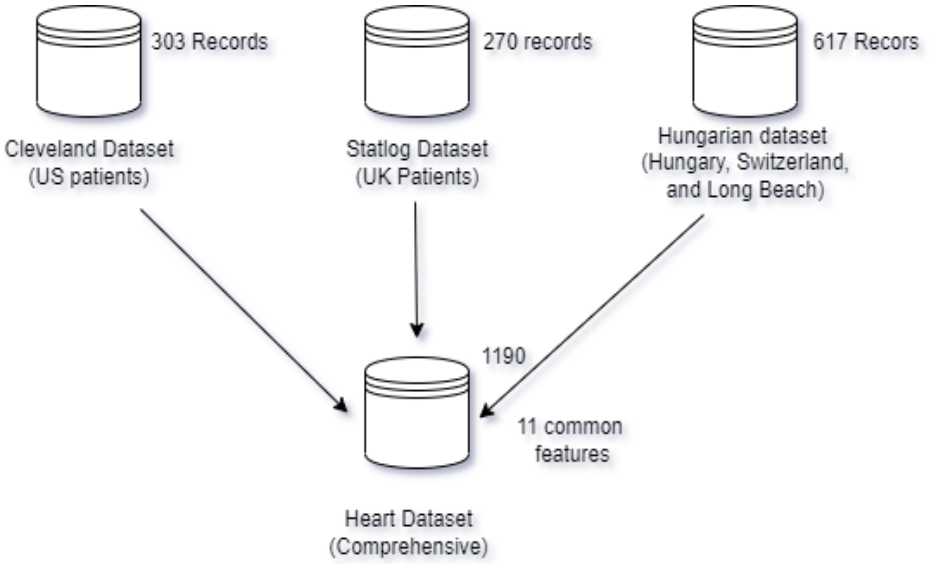

Figure 1.2: Data Curation

1.4.2 Description of Data

Table 1.1: Feature Description

	Feature	Description	Data type
1.	Age	Patients' years of age	Numeric
2.	Sex	Gender of Patient (Male - 1, Female - 0)	Nominal
3.	Chest pain type	Chest pain type 1 = Typical angina 2 = Atypical angina 3 = Non-anginal pain 4 = Asymptomatic	Nominal
4.	Resting BP	Level of blood pressure at resting mode in mm/HG	Numeric
5.	Cholesterol	Serum cholesterol in mg/dl	Numeric
6.	Fasting blood sugar	Fasting blood sugar 0 = Less than 120 mg/dl 1 = More than 120 mg/dl	Nominal
7.	Resting ECG	Resting electrographic result 0 = Normal 1 = Having ST-T wave abnormality 2 = left ventricular hypertrophy	Nominal
8.	Max Heart rate	Maximum heart rate achieved	Numeric
9.	Exercise angina	Exercise induced angina 0 = No 1 = Yes	Nominal
10.	Old peak	Exercise induced ST-depression in comparison with the state of rest	Numeric
11.	ST slope	Slope of the peak exercise ST segment 0 = Normal 1 = Unsloping 2 = Flat 3 = Downsloping	Nominal
TARGET VARIABLE			
12.	Target	It is the target variable that we must forecast; a value of one denotes a patient who is at heart risk, whereas a value of zero indicates a healthy patient	Heart Risk, Nominal

1.4.3 Analytic Technique and Tools

The analytic tool that will be used would be Python and it would be used to classify the various algorithms such as:

- Logistic Regression
- Support Vector Machine
- K Nearest Neighbour (known as Instance-based learning with k parameter in Weka)
- Artificial Neural Networks

1.5 Significance of Study

The heart is a very crucial part of the body; this makes the heart disease prediction system very beneficial globally because it:

- Helps to lower the high global death rate.
- Opens the door to early heart disease diagnosis.
- Reduces the burden on the healthcare facilities.
- Reduces the cost of money spent each year on heart related diseases.

1.6 Organisation of Study

There are five chapters in this project effort.The introduction to the topic in the first chapter covered the background, problem description, objectives, techniques, importance of the study, and organization of the study. Chapter two reviewed the research paper's related papers and gave a theoretical foundation. The machine learning methods, data source, data analysis tools, performance indicators, and certain hyper-parameter tuning models were all described in detail in chapter three. The key conclusions and discussions were addressed in Chapter 4. The project was summarized in chapter five, which also covered the study's weaknesses and offered suggestions for potential fixes.

Chapter 2

Literature Review

2.1 Introduction

Machine learning is a technique for automatically identifying patterns in data.Supervised, unsupervised, and reinforcement learning are the three main categories of machine learning. A technique called supervised learning, commonly referred to as supervised machine learning, uses labeled training data to learn how to predict results for new data.Support vector machines, decision trees, logistic regression, and naive bayes classifiers are a few supervised machine learning techniques. It is essential to identify and anticipate heart disorders early because doing so can lower mortality rates and overall consequences. As a result, various studies on the topic of heart disease have been carried out using data mining and machine learning approaches.

2.2 Related works

In order to improve the precision of the prediction of cardiovascular disease, Mohan et al. (2019) suggested a unique method that makes an effort to identify crucial variables by utilizing machine learning techniques. Using the UCI heart disease data set, the accuracy of the recommended strategy was compared to various machine learning techniques. A unique technique known as Hybrid Random Forest with Linear Model (HRFLM) was developed for this investigation. In their hybrid HRFLM technique, the authors combined the traits of Random Forest (RF) and Linear Method (LM). HRFLM surpassed Decision Tree (DT), Random Forest (RF), and Linear Method (LM) in terms of the number of attributes and prediction error, demonstrating that it is quite accurate in predicting heart disease. In order to improve the accuracy of heart disease prediction and gain a deeper understanding of the key features, new feature-selection strategies can be developed, claim Mohan et al. (2019).

The study by Dhai et al. (2021) used data analytics to predict cardiac disease. The two portions of their study were pre-processing, where they identified the most important features, and machine learning techniques, where they identified which algorithm offered the highest level of accuracy. A structured data set from Algeria, namely the Mohand Amokrane EHS Hospital ex CNMS in Algiers, comprising 1200 rows and 20 columns. Three techniques were selected and used by (Dhai, Abdelkamel, & Tahar, 2021). These were neural networks, support vector machines (SVM), and K-nearest neighbor (KNN). After research, it was shown that neural networks deliver the best results. In their conclusion, Dhai et al. (2021) stated that their method may be improved by adding deep learning algorithms, additional methods of attribute selection, and even larger data sets.

The accuracy of machine learning algorithms for predicting cardiac illnesses using approaches such as k-nearest neighbor, decision tree, linear Regression, and Support Vector Machine was examined by Achana et al. (2020) using the UCI repository data set for training and testing (SVM). They used Jupyter Notebook, a tool from Anacoda, to carry out these analyses. It has been established that K-Nearest Neighbor is superior to other techniques. The following ones were Support Vector Machine, Decision Tree, and Linear Regression. The conclusion of their article said that "In the Future, More Machine Learning Approach Will Be Used For the Best Analysis Of Heart Diseases And For Early Prediction Of Diseases So That The Rate Of The Death Cases Can Be Minimized By The Awareness About The Diseases." (Archana & Rakesh, 2020).

Several machine learning techniques were employed by Saboor et al. (2022) in their study to identify and forecast human cardiac disease. The algorithms' performance was then evaluated using a variety of metrics, such as classification accuracy, sensitivity, specificity, and F measure, on the heart disease data set. Nine (9) machine learning classifiers were applied to the final data set both before and after the hyper parameter tweaking of the classifiers. These included the AdaBoost Classifier (AB), the Logistic Regression (LR), the Extra Trees Classifier (ET), the Multinomial Nave Bayes (MNB), the Decision Trees (CART), the Support Vector Machine (SVM), the Linear Discriminant Analysis (LDA), the Random Forest (RF), and the XGBoost (XGB). Online repositories such as the Cleveland heart disease data set, Z-Alizadeh Sani data set, Statlog Heart, Hungarian Long Beach VA, and Kaggle Framingham data set were the sources of the data sets used by Saboor et al. (2022). In conclusion, a variety of classifiers were employed to forecast the development of heart disease, with the Support Vector Machine emerging as the most reliable. In this study article by Saboor et al. (2022), some disadvantages include the fact that the functioning of the earlier proposed systems is significantly lowered if the size of the data set is raised. Additionally, the classifier prediction accuracy increases as the size of the data set grows, but beyond a certain size, the accuracy of the classifier prediction decreases.

Haq et al. (2018) created a machine learning-based diagnosis method for heart disease prediction using a data set of heart disease. Cross-validation, three feature selection techniques, seven well-known machine learning algorithms, and metrics for classifier performance assessment such as classification accuracy, specificity, sensitivity, Mathews' correlation coefficient, and execution time were all used. This study employed the Cleveland heart disease data collection from 2016, which is popular among researchers. A hybrid intelligent machine learning-based prediction system was proposed for the diagnosis of cardiac illness. The seven well-known classifiers Logistic Regression, K Nearest Neighbor, Artificial Neural Network (ANN), Support Vector Machine (SVM), Naive Bayes, Decision Tree, and Random Forest were used to select the critical features using the three feature selection algorithms Relief, Minimal-Redundancy-Maximum-Relevance Feature Selection Algorithm (MRMR), and Least Absolute Shrinkage and Selection Operator (LASSO). Logistic regression with 10-fold cross-validation demonstrated the best accuracy when it was selected by the FS algorithm Relief. However, in terms of specificity, the MRMR algorithm surpassed SVM (linear) with feature selection. The authors conclude that more research is required. (Haq, Li, Memon, Nazir, & Sun, 2018), to improve the efficiency of these predictive classifiers for the diagnosis of heart disease by using various feature selection algorithms and optimization strategies.

In their research work, Li et al. (2020) proposed a machine learning-based approach for diagnosing cardiac illness that is both accurate and efficient. Support Vector Machine (SVM), Logistic Regression (LR), Artificial Neural Network (ANN), K-Nearest Neighbor (K-NN), Nave Bayes (NB), and Decision Tree (DT) are some of the classification algorithms. Standard features selection algorithms, such as Relief, Minimal Redundancy Maximal Relevance (MRMR), Least Absolute Shrinkage Selection Operator (LASSO), and Local Learning Based Features Selection (LLBFS) have also been used. The data set used was the Cleveland Heart Disease. According to their paper, the specificity of ANN classifier is best on Relief FS algorithm as compared to the specificity of MRMR, LASSO, LLBFS and FCMIM feature selection algorithms (Li et al., 2020). The performance of the diagnosis system could be negatively impacted by irrelevant features, which would lengthen computation times. As a result, a unique touch was introduced to the study. The novel approach uses feature selection algorithms to choose the right features that increase classification accuracy and shorten the diagnosis system's processing time. In order to enhance the functionality of a prediction system for the identification of heart disease, more feature selection algorithms and optimization approaches will be used in the future, according to Li et al. (2020).

In the proposed publication by Obasi et al. (2019), the authors developed a machine learning-based system that uses patient medical information to identify and forecast heart problems in patients. The proposed solution was built on the use of already-existing techniques including Random Forest (RF), Logistic Regression (LR), and Naive Bayee, which provided a decision support system for medical professionals to recognize and predict cardiac problems. Three data sets were used: the Cardiovascular illness data set from Kaggle, the Framingham Heart study data set from Kaggle, and the Cleveland heart disease data set from the UCI machine learning library.The system, which forecasts patients' risk of heart disease, was deployed on the RStudio platform. Following data analysis, it was discovered that Random Forest outperformed Logistic Regression and Nave Bayes. Since additional qualities (risk factors) were included in their model compared to earlier

works, their study paper stood out since it improved and expanded the system by identifying and foretelling heart disease in people who have a risk factor that was not taken into account in previous studies.

The goal of Kumar et al(2018) .'s research is to do predictive analysis on cardiac illnesses using data mining and machine learning algorithms, examine the various mining and machine learning techniques used, and make judgments about whether approaches are effective and beneficial. Four algorithms—the Decision Tree (J48) approach, Naive Bayes, K-Nearest Neighbor, and Support Vector Machine—were applied to predict heart illness. The data sets were provided through the UCI Machine Learning Repository, and Weka was the preferred data mining program. The J48 tree technique was found to be the best precise and quick classifier for heart disease prediction after analyzing the trial results (Kumar, Koushik, & Deepak, 2018). Given the limited success in developing predictive models for heart disease patients, they found more combinatorial and more sophisticated models would be required to improve the accuracy of predicting the early beginning of heart disease.

The goal of Nishadi's (2019) study was to use logistic regression to determine the most important risk factors for heart disease and to predict overall risks. The data set was obtained from the Kaggle website, and JupyterLab's Python environment was used for data analysis. In conclusion, men are more vulnerable to heart disease than women are, according to the results of logistic regression. The model employed is more sensitive than specific. He continued by saying that adding more data would help the model.

Researchers employ a variety of machine learning techniques to analyze enormous volumes of complex healthcare data, which helps healthcare practitioners forecast cardiac illness, according to Ramesh et al research from the year 2022. The performance of various techniques was validated using the UCI heart disease data sets. They used supervised learning techniques such Naive Bayes, Support Vector Machines, Logistic Regressions, Decision Tree Classifiers, Random Forests, and K-Nearest Neighbors for their investigation. The various Machine Learning techniques were put into practice utilizing Python programming and the Anaconda package. The KNN classification algorithm surpassed the other parameters at the conclusion of the data analysis.The research conducted by Ramesh et al. (2022) can be extended to different biological disorders and the new classification models can be included for complex time series data sets.

In their study article, Katarya et al. (2020) noted that machine learning has demonstrated successful outcomes when making judgments and predictions from a large set of data supplied by the health care industry. Artificial Neural Networks (ANN), Decision Trees (DT), Random Forests (RF), Support Vector Machines (SVM), Naive Bayes (NB), and the K-Nearest Neighbor (KNN) algorithm were some of the supervised learning approaches employed in this prediction of heart disease. In their conclusion, Katarya et al. (2020) stated that using search algorithms to choose features will be preferable in the future, and that machine learning approaches will produce better predictions for heart disease.

In order to determine which machine learning algorithm produces the best results, Jonnavithula et al. (2022) compared the accuracy of various machine learning strategies. The "heart.csv" data collection was utilized in this study to determine whether or not a person has a heart condition. Artificial Neural Network (ANN), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), Naive Bayes (NB), and K-Nearest Neighbor (KNN) algorithms were the machine learning methods employed. The most precise forecast is given by an artificial neural network after viewing the results of various methods (Jonnavithula, Jha, Kavitha, & Srinivasulu, 2022).The use of a small data set in this study had its limitations, and the authors concluded that using a larger data set might improve the performance and accuracy of the algorithms.

In their work, Vardhan et al. (2022) established a technique for detecting the existence of cardiac disease using clinical data collected from subjects. The main objective was to create a predictive model for heart disease using a variety of characteristics. Different machine learning classification strategies were also tested and assessed using traditional performance metrics like accuracy in order to compare various machine learning algorithms. For this experiment, data from the UCI Heart Disease Data Collection were used. Machine learning methods such Random Forest, Support Vector Machine, K-Nearest Neighbor, Decision Tree, Artificial Neural Networks, Logistic Regression, and Naive Bayes were used to compare how well they predicted cardiac diseases.The Random Forest algorithm was shown to be the most reliable strategy for predicting cardiac illness, with an accuracy rate of 90.16 percent, according to Vardhan, Reddy, and Umamaheswari (2022). In their conclusion, Vardhan et al. (2022) recommended using large data sets in trials in the future to increase the reliability of their findings and help doctors better anticipate cardiac disease.

In their research paper, Robini et al. (2021) presented a machine learning-based system that would be quick and accurate at diagnosing cardiac disease. On the data set, which consists of 303 occurrences and 14 attributes, the classification was performed using machine learning algorithms including Naive Bayes, Decision Tree, and Random Forest Algorithm. The best classification method was the Random Forest approach. In order to further improve performance, Robini et al. (2021) concluded that "In the future, it is conceivable to give expansions or modifications to the suggested clustering and classification methods employing intelligent agents. Other combinations, such as artificial intelligence, soft computing, and other clustering methods, can be employed in addition to the tested combination of data mining approaches to increase accuracy (Rohini, Keerthika, Pavithra, Sandhiya, & Vedha, 2021).

The most essential and sought-after machine learning method was attempted to be implemented in the study by Praveen et al. (2019) in order to predict the presence of cardiac disease in a patient. To determine whether a patient has the condition, machine learning methods like the Decision Tree Classifier and Support Vector Machine were employed. Python and the Weka Tool, which was utilized to manipulate the data collection, were the programming languages employed in this paper. They mentioned in their conclusion that, based on the data set utilized for the study, they could employ machine learning algorithms to anticipate particular results. When used on the data set, the decision tree approach had a greater prediction accuracy than the Support Vector Machine.

The primary goal of Ufumaka's research is to compare the performance of Multi-Layer Perceptron (MLP) and Support Vector Machine(SVM), as well as to classify the presence of heart disease utilizing fine-tuning hyperparameters. This study specifically employed the Cleveland data set from the UCI heart disease data set. The work was implemented using the Python programming language. The performance criteria Accuracy, Recall, Precision, F1-Score, and Area Under Curve Receiver Operating Characteristics (AUROC) were used to assess the optimized prediction model. In summary, the MLP outperforms SVM in terms of accuracy, precision, and F1-score but trails SVM in terms of recall and AUROC (Ufumaka, 2020).

The UCI heart disease machine learning repository data set was used in the Nagaraj et al. (2019) paper to predict whether a person has a heart condition. Patients with cardiac problems were examined using naive Bayes classification and support vector machines. The data analysis were conducted using the R programming language. It was discovered that the Nave Bayes classification was inaccurate, whereas the Support Vector Machine was more accurate. Using the consistency metrics Root Mean Squared Error (RMSE), it was discovered that women are more susceptible to heart disease than men. In the future, it is advised by Nagaraj et al. (2019) to research various machine learning approaches including deep learning, association rule analysis, and genetic algorithms to predict the accuracy with appropriate performance criteria.

The goal of the Yash et al. (2020) paper was to determine the most effective machine learning model for predicting various cardiac diseases. They evaluated the accuracy of each algorithm and chose the one that would produce the most accurate results. All of the methods utilized in this work were supervised learning algorithms, including the Gaussian Classifier, Decision Tree, K-Nearest Neighbor, and Random Forest. For this investigation, the Python and Spyder programming languages were employed. The K-Nearest Neighbor approach produced the best results in the end, outperforming all other algorithms (Yash, Shruti, & Shubham, 2020).

In the research paper put forth by Serkalem et al. (2022), an effort was made to use Machine Learning (ML) approaches to detect rheumatic heart disease (RHD) based on patient history, symptoms, and pathological results of patients with heart disease. Based on the features discovered from the training data set, a machine learning classification model was created and utilized to identify the presence and absence of RHD (Serkalem, Kula, Beakal, & Azene, 2022). The Cardiac Center of Ethiopia provided the data collection that was used to develop the RHD classification model.For the purpose of detecting heart illnesses, various machine learning methods like K-Nearest Neighbor, Support Vector Machine, Logistic Regression, and Random Forest were applied. The Anaconda environment and the Juypter Notebook tool with Python language were used to analyze the data. For the purpose of RHD prediction, the Support Vector Machine classifier was advised due of its high accuracy. The use of a small data set was the study's only drawback, and it was noted that the availability of larger clinical data and the inclusion of additional features could have enhanced the study's ability to identify and detect Rheumatic Heart Disease risk levels and produce more trustworthy results.

Since heart disease is a leading cause of death, Sujay et al. (2020) used a variety of machine learning algorithms, including Logistic Regression, Decision Trees, K-Nearest Neighbors, Support Vector Machines, and Artificial Neural Networks, to develop a predictive model that can tell whether a person has heart disease or not. The UCI machine learning repository's data set was used for this heart disease prediction, and data analysis was done using the python programming language. Artificial neural networks had the highest degree of accuracy, demonstrating the superiority of deep learning techniques for categorizing cardiac disorders." The model proposed in this study can be applied to biotechnological devices for automated and quicker diagnosis of heart ailments, which will help the medical sector grow and take the next important step toward automation, the scientists concluded. (Sujay, Ritesh, Mahendra, Siddharth, & Manjusha, 2020).

With the aid of different machine learning methods, such as K-Nearest Neighbors, Support Vector Machine, Logistic Regression, Nave Bayes, Decision Tree, and Random Forest, Nikita et al. (2021) suggested a model to predict patients' chances of developing heart disease. The data analysis for this study utilizes the Python Library. For predicting heart diseases, K-Nearest Neighbor and the Random Forest approach procedure were effective machine learning techniques.

A comparative analytical strategy was used in the research study by Ufumaka (2021) to determine which algorithm works better under the specified conditions. Data from the University of California, Irvine (UCI) machine learning database were used in the study. Support Vector Machine, K-Nearest Neighbor, Logistic Regression, Naive Bayes, Random Forest, and Gradient Boosting Ensemble Method were utilized as supervised machine learning models for categorization. In addition, these algorithms' performance was assessed using common performance indicators like precision, recall, and F1-score. The Support Vector Machine proved to be the most effective algorithm, although he stated that it should be noted that each of the used algorithms can perform better than the others in some situations. Ufumaka (2021) concluded by advising that future research studies should concentrate more on enhancing the precision of these models by the use of hyper parameter tuning and larger data sets on other ensemble methodologies.

Ozichi et al. (2019) use a trained model from supervised learning to forecast a system for heart diseases. K Nearest Neighbor (KNN), Support Vector Machine (SVM), and Decision Tree (DT) methods were utilized as the three machine learning algorithms to train the data set. To forecast heart illnesses using various machine learning methods, MATLAB2018 was utilized as the programming language. When the Principle Component Analysis (PCA) was enabled, the K Nearest Neighbor method had the greatest and best accuracy.The study was limited by the use of only heartbeat measurements without correlating electrocardiography (ECG), chest X-ray, and echocardiography because the heartbeat may not always match these other measurements. According to Ozichi et al. (2019), larger characteristics should be included in the heart rate data sets to improve training in a wider demographic sector. Additionally, for a better assessment of the work, other features and readings could be added.

In the Shafiul et al. (2020) paper, different machine learning algorithms, such as Naive Bayes (NB), Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM), and Logistic Regression, are evaluated using a certified data set for the prediction of cardiac disease (LR). In order to forecast the possibility of acquiring heart disease, the study examines the effects of Principle Component Analysis (PCA) on the accuracy of machine learning algorithms and establishes the relationships between the various features that can be employed to do so. Random Forest is the strongest predictor for cardiac problems, according to the analysis.In their conclusion, Shafiul et al. (2020) stated that they would "strive to use the primary data as much as possible in Bangladesh in the future and employ deep learning with some ensemble classification techniques find out the risk of heart diseases and try to draw some conclusions and make some recommendations." (Shafiul, Abu, & Humayan, 2020).

Chandu et al. (2022) proposed model aims to forecast the risk of developing cardiopathy using machine learning methods. Their research focuses on the prediction of cardiopathy using patient data sets and patient data for whom they intended to forecast the likelihood of the development of cardiovascular illness. The machine learning techniques utilized were Linear Regression, Naive Bayes, Support Vector Machine, K-Nearest Neighbors, Decision Tree, and XG Boost. The data set was obtained from Kaggle. The machine learning technique using decision trees performed the best at predicting heart disease.In their conclusion, Chandu et al. (2022) stated that their project can be used in other real-world scenarios or in other medical diagnostics to evaluate large volumes of data and find the risk factors associated with various diseases. Due to the small sample size, their biggest drawback was the difficulty in applying their findings to cardiac disease. They hope to expand the use of their methodology in the future and analyze additional diseases using various feature selection methodologies.

In the research work by Akanksha et al. (2017), they compared the competency of four supervised machine learning algorithms in terms of the accuracy they were able to forecast the development of heart disease. The four machine learning techniques used are boosting trees, random forests, support vector machines, and logistic regression. The Cleveland heart disease data set, which is accessible at UCI, was the one used. R was the chosen programming language. The best model in terms of accuracy turned out to be the logistic regression, while three models using various tuning parameters fared only somewhat worse. Data preparation approaches, including the use of outliers, variances, model selection, and model tuning parameters, among others, can be used to make a number of changes, according to Akanksha et al. (2017) (Akanksha, Shubham, & Prof., 2017).

The study by Nashif et al. (2018) used machine learning to forecast approaching heart disease and presented an early concept for a cloud-based heart disease prediction system. Machine learning techniques were examined using programming software called WEKA, a Java-based Open Access Data Mining Platform.The suggested technique was then tested using two popular open-access databases, employing 10-fold cross-validation to examine how well the heart disease performed.(Nashif, Raihan, Islam, & Imam, 2018). The Cleveland Heart Disease data collection served as their data source, while Simple Logistic Regression, Random Forest, Artificial Neural Networks, Support Vector Machine, and Naive Bayes were among the machine learning methods employed. A smartphone-based application for identifying and forecasting heart disease risk level was developed after se-

lecting the machine learning algorithm with the best accuracy and performance. Support Vector Machine was shown to be the best effective algorithm for use in the prediction of heart illnesses at the conclusion of the study. They suggested that going forward, they should concentrate more on creating a dedicated server and database for the patient monitoring app, which, if authorized, would be accessible through the Android Play Store. As a result, the application can be installed and used to anticipate heart disease by any patient or physician from anywhere in the world.(Nashif, Raihan, Islam, & Imam, 2018).

Using three crucial feature selection techniques and seven supervised machine learning algorithms, Suneeta et al. (2021) proposed a prediction. The Cleveland data set was downloaded from the UCI repository and used in the experiments. Logistic Regression (LR), Artificial Neural Network (ANN), Decision Tree (DT), Supervised Vector Machine (SVM), Naive Bayes (NB), Random Forest (RF), and K-Nearest Neighbor (KNN) are the seven supervised machine learning methods that are employed. The filter method, wrapper method, and embedding method are the feature selection techniques that are used. Spyder software running on Windows was used to run the simulations. The Random Forest algorithm produced the best results at the end of the analysis.

Fernandes et al. (2022) used a data collection and six machine learning algorithms to predict cardiac problems in order to evaluate which model is the most successful. The data collection was examined using Python and the Rapid Miner program. The Cleveland Heart disease data set from the UCI library was the one that was used. Decision Tree, K-Nearest Neighbor, Naive Bayes, Support Vector Machine, Logistic Regression, and Neural Networks. Finally, since it has the highest accuracy, Logistic Regression is the best model for categorizing this exam, followed by Naive Bayes.

In their work, Nikhil et al. (2022) propose employing a variety of machine learning techniques, including Logistic Regression, Naive Bayes, Support Vector Machine, K-Nearest Neighbors, Random Forest, and Extreme Gradient Boost, to predict cardiac disease. Based on elements that were taken out of the data set, these machine learning algorithms were utilized to forecast a person's likelihood of developing heart disease. (Nikhil, Sreedevi, & Ahmad, 2022). In this study, the UCI machine learning repository and a data collection gathered from Kaggle were both employed as data sources. The intrinsic libraries of the Python programming language, including numpy and matplotlib, were used. Support When applied to the first data set, Vector Machine had the best test accuracy, whereas Random Forest method had the highest accuracy when applied to the second data set. The accuracy obtained by Random Forest when the two data sets were combined was the highest. In their conclusion, Nikhil et al. (2022) mentioned that they might look at deep learning algorithms in the future to see how they perform.

In the study paper by Karthiga et al. (2022), a machine learning method is used to forecast the severity of heart illnesses in the near future. K-Nearest Neighbor Algorithm and Tuned -K-Nearest Neighbor Algorithm were the two methods employed. Python programming was used, and the data set was retrieved from the UCI repository. The Tuned K-Nearest Neighbor showed the highest accuracy for predicting heart illnesses once the analysis was completed.

A program was created by Prathamesh et al. (2022) to forecast a person's risk of developing heart disease. To determine the likelihood of having a heart condition or not, a number of machine learning techniques like Support Vector Machine, Naive Bayes, and Logistic Regression were tested. The computer language Python was utilized, and the data source used included the medical histories of 70.000 patients that were gathered through Kaggle. Out of the three, the Logistic Regression had the best accuracy. The Android application was made using Java and FlaskAPI, which the developers hope can significantly aid in the rapid diagnosis of cardiac illness and have a significant global impact.

The study by Mursal et al.(2020) outlines a statistical model of heart illness that, based on the patient's basic health history, would aid coroners and cardiac practitioners in predicting heart diseases. Logistic Regression Classifier, K-Nearest Neighbors Classifier, and Random Forest Classifier were the three different machine learning classifier models that were used. Since the University College Irvine (UCI) data collection contains all the necessary attributes, it was chosen. After the analysis, Logistic Regression produced the best results. Mursal et al. (2020) suggested that the data set can be further extended to include new attributes in order to diversity. Additionally, the addition of additional data sets would improve prediction accuracy. "Centered on diseases and algorithms, a comparative study of the performance of this model would be facilitated," they concluded, and I paraphrase.(Mursal, Adnan, Hiba, Sanam, & Kanwal, 2020).

By using machine learning approaches to identify significant features, Galla et al. (2020) suggested a narrative method that attempts to increase the accuracy of cardiovascular disease prediction. They classified cardiac disease using the data set from the UCI repository using Python and Pandas Operations. Decision Trees, Language Model, Random Forest, and Support Vector Machine were the machine learning models employed. At the conclusion of the analysis, the Random Forest Classification provided better findings and would be excellent at correctly predicting whether a patient had heart disease or not.

Gradient Boosting, Decision Tree, Random Forest, and Logistic Regression were among the Supervised Machine Learning algorithms utilized in Sujay et al. (2020) research to develop a model for forecasting Myocardial Infarction. The UCI Machine Learning Repository and the Framingham data set were used as the data sets. Python was the coding language employed. The best machine learning method for predicting cardiac illnesses is the gradient boosting classifier because it had the highest percentage.

In their article from 2021, Kwakye et al. discuss a comparative method for categorizing data sets related to coronary heart disease using machine learning algorithms. Using the Framingham data set, their study developed and evaluated a number of machine learning-based classification models. K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Decision Tree (CART), Logistic Regression (LR), Naive Bayes (NB), and Random Forest (RF) were the machine learning methods employed. Cross-Validation Accuracy and the Area Under the Curve (AUC) of the Receiver Operating Characteristics (ROC) were used to assess these models' performance. The initial data were transformed by balancing the classes, which was done using the Synthetic Minority Oversampling Technique (SMOTE). The Logistic Regression model outperformed the other models in predicting coronary heart illnesses once hyper-parameter tweaking was completed. Fu-

ture studies will "study missing values and outliers using imputation techniques and other methods that avoid the use of a single value, such as the mean for data cleaning and preprocessing," according to Kwakye et al. (2021). (Kwakye & Dadzie, 2021). Additionally, they want to construct unified models by combining the best performing models into an ensemble, which they hope will enhance prediction performance.

The goal of Nishadi's study (2019) is to determine the most important predictors of heart illnesses and forecast overall risks using only the logistic regression machine learning method. The data set was obtained from Kaggle, and all data analysis was done on Python using JupyterLab. In conclusion, it was found through the use of logistic regression that men are more susceptible to heart disease than women. The accuracy of the model was 87%, which is excellent. It is advised that more data be used in future studies.

A model for identifying cardiovascular illnesses using machine learning algorithms was put up by Ahmad et al.(2020). Machine learning algorithms like Support Vector Classifier, K-Nearest Neighbors Classifier, Random Forest Classifier, and Decision Tree Classifier were utilized to train the data set. With the use of the Kaggle data collection, Python programming was used to train the models. They used an API called Flask as an application programming interface. Following investigation, two machine learning algorithms—Decision Tree and Random Forest—had the same level of precision.

Using data mining classification technique, the Patel et al. (2022) model analyses the medical parameters in the data sets. The cardiac condition was collected from the UCI Machine Learning Repository and was programmed in Python. The three machine learning methods used were Logistic Regression, Decision Trees, and Random Forests. The results showed that the Logistic Regression had the highest accuracy, at 92 percent. In order to obtain important features to forecast whether a patient has a cardiac condition given the input they provide, Patel et al. (2022) developed a website using HTML, CSS, JAVASCRIPT, and Django for the backend and the frontend, respectively.They concluded by saying, "The system we are proposing is more advanced and affordable than the ones that are already available as we have put a prior records and doctor and hospitals details features on our website."(Patel, Patange, Patil, & Kapoor, 2022).

Chapter 3

Methodology

3.1 Source of Data/Data Acquisition

This suggested model's data came from the UCI Machine Learning Repository as its data source. The Heart Disease Data Set was employed, and it has been a popular global resource for students, instructors, and researchers looking for machine learning data sets. David Aha and graduate students at UC Irvine produced the data set in 1987. Four datasets from different institutions make up the heart disease data set, including the following:

1. Cleveland - Cleveland Clinic Foundation
2. Hungary - Hungarian Institute of Cardiology, Budapest
3. Switzerland – University Hospital, Zurich, Switzerland
4. Long Beach VA – V.A Medical Centre, Long Beach, CA.

The Heart Disease Data Set has 76 attributes and is made up of four databases from different universities. Only 12 of these traits—including the projected trait—are utilized in this investigation.The target field in the data set indicates whether a patient has heart disease or not, with 0 signifying no disease and 1 signifying the presence of a disease. The form of the data collection is (1190,12).

Table 3.1: Feature Description

	Feature	Description	Data type
1.	Age	Patients' years of age	Numeric
2.	Sex	Gender of Patient (Male - 1, Female - 0)	Nominal
3.	Chest pain type	Chest pain type 1 = Typical angina 2 = Atypical angina 3 = Non-anginal pain 4 = Asymptomatic	Nominal
4.	Resting BP	Level of blood pressure at resting mode in mm/HG	Numeric
5.	Cholesterol	Serum cholesterol in mg/dl	Numeric
6.	Fasting blood sugar	Fasting blood sugar 0 = Less than 120 mg/dl 1 = More than 120 mg/dl	Nominal
7.	Resting ECG	Resting electrographic result 0 = Normal 1 = Having ST-T wave abnormality 2 = left ventricular hypertrophy	Nominal
8.	Max Heart rate	Maximum heart rate achieved	Numeric
9.	Exercise angina	Exercise induced angina 0 = No 1 = Yes	Nominal
10.	Old peak	Exercise induced ST-depression in comparison with the state of rest	Numeric
11.	ST slope	Slope of the peak exercise ST segment 0 = Normal 1 = Unsloping 2 = Flat 3 = Downsloping	Nominal
TARGET VARIABLE			
12.	Target	It is the target variable that we must forecast; a value of one denotes a patient who is at heart risk, whereas a value of zero indicates a healthy patient	Heart Risk, Nominal

Table 3.2: Heart Disease Data Set

number	**age**	**sex**	···	**stslope**	**target**
0	40	1	···	1	0
1	49	0	···	2	1
2	37	1	···	1	0
3	48	0	···	2	1
···	···	···	···	···	···
···	···	···	···	···	···
···	···	···	···	···	···
1186	68	1	···	2	1
1187	57	1	···	2	1
1188	57	0	···	2	1
1189	38	1	···	1	0

3.2 Machine Learning Techniques

3.2.1 K-Nearest Neighbor

The K-Nearest Neighbors (K-NN) algorithm is a well-known, basic, and fundamental supervised machine learning solution for classification and regression problems even though it is primarily utilized for classification applications. The K-NN algorithm allocates a new data point to the category that is most similar to the existing categories by assuming similarity between the new data point and the existing data points. All of the available data is stored by the algorithm, which also categorizes fresh data based on similarities. This shows that the K-NN algorithm can quickly categorize fresh data into the proper category (JavaTpoint-KNN, 2011-2021).

K-NN uses the straight-line distance popularly known as the Euclidean distance. The formula is shown below.

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

x, y = two points in Euclidean n-space
x_i, y_i = Euclidean vectors, beginning at the space's origin
(initial point)
n = n-space

Since the K-NN method is non-parametric, it makes no fundamental assumptions on the distribution of the data. This method is frequently referred to as a "lazy learner" because it does not immediately learn from the training set but instead keeps the data and does classification when making a prediction.

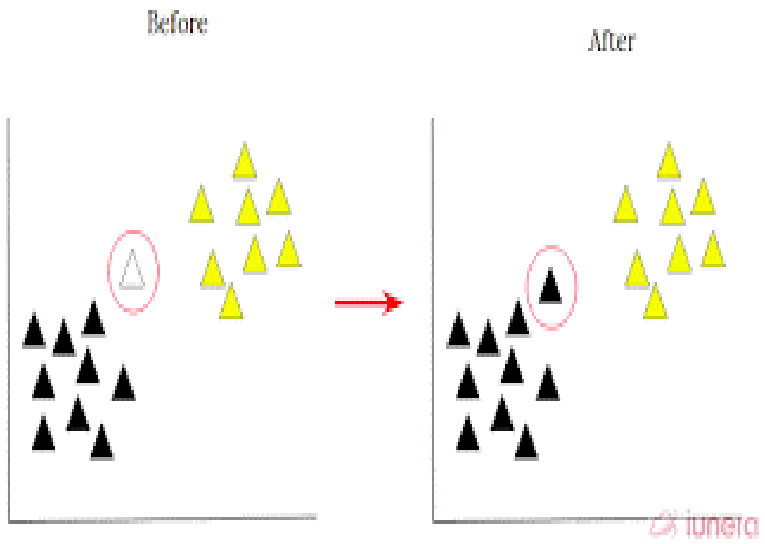

Figure 3.1: Image of K-NN before and after

Advantages

1. The K-NN algorithm is straightforward and uncomplicated to implement.

2. It doesn't require building a model, adjusting multiple parameters, or making additional assumptions.

3. The algorithm is adaptable and can be applied to a variety of tasks including classification, regression, and search (Onel, 2018).

Disadvantages

1. The K-NN technique has the drawback that, occasionally, figuring out the value of K can be difficult.

2. Another drawback is that the computation cost is relatively high, as the distance between data points must be calculated for all training samples (JavaTpoint-KNN, 2011-2021).

3.2.2 Logistic Regression

The widely used machine learning method of logistic regression is used to predict a categorical dependent variable from a set of independent inputs. A logistic regression model's output is typically a categorical or discrete value, such as "Yes" or "No," "0" or "1," "True" or "False," etc. It does, however, provide a probabilistic value that ranges from 0 to 1. (JavaTpoint-LR, 2011-2021).

A useful tool in a variety of fields, including marketing, finance, and healthcare, is logistic regression. The fact that this machine learning system can assign probabilities and categorize fresh data using both continuous and discrete data sets makes it especially helpful.

It is crucial to take into account the following premises while using Logistic Regression on a set of data:
1. There must be a categorical dependent variable.
2. There shouldn't be much association between the independent variables (multi-collinearity).

Advantages

1. Implementing logistic regression is comparatively easy and, in some circumstances, extremely effective.

2. One of the advantages of Logistic Regression is that it does not require extensive computational resources.

3. It is primarily used to establish relationships between different features.

4. Additionally, Logistic Regression models are relatively easy to update as new data becomes available, unlike Decision Tree or Support Vector Machine algorithms (Gunturu, Cherukuri, S., Koruprolu, & Kesuboyina, 2017-2021).

Disadvantages

1. Logistic Regression creates linear decision boundaries.

2. For Logistic Regression to be effective, a significant degree of relationship between the independent variables must be avoided (multi-collinearity).

3. Finding complex correlations using logistic regression can be difficult.

4. It is preferable to avoid using logistic regression when there are less data than features because this could result in overfitting.

3.2.3 Support Vector Machine

The supervised machine learning technique Support Vector Machine (SVM) is frequently used for classification and regression issues. Classification is its major application, despite the fact that it can be used for either type of problem. SVM tries to construct a hyperplane that can partition an n-dimensional space into different classes, making it simple to classify new data points for subsequent classifications.

Finding a hyperplane in an n-dimensional space that can successfully divide multiple classes is the aim of SVM. The decision boundary, also known as the hyperplane, is formed by choosing the most extreme points or vectors, referred to as support vectors, that best characterize the boundary between classes. The algorithm's use of support vectors gives it its name (Javatpoint-SVM, 2021).

Face detection, Image Classification and Text categorization are some of the uses of the SVM algorithm.

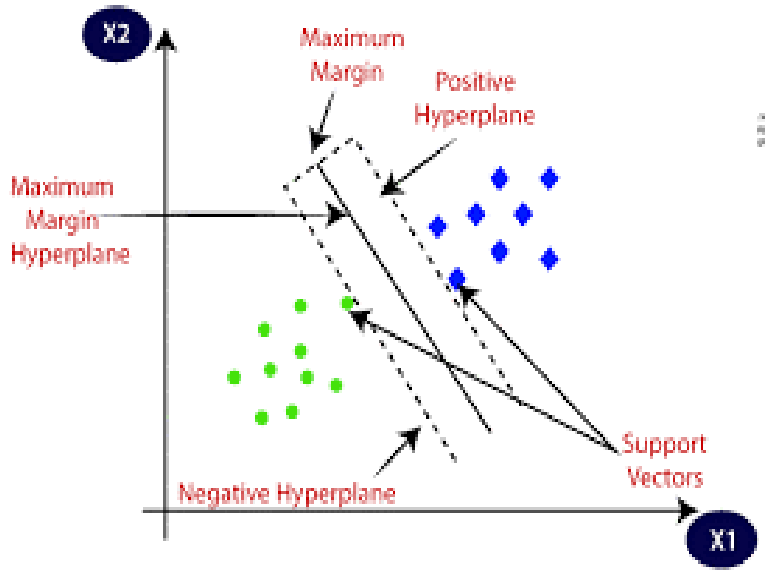

Figure 3.2: Labelled parts of Support Vector Machine

Advantages

1. Support vector machines, or SVMs, perform well in multidimensional spaces.

2. When the number of dimensions is more than the number of samples, SVMs are quite helpful.

3. SVMs are memory efficient.

4. SVMs are versatile and can be used for a variety of tasks, including face detection, image classification, and text categorization.

Disadvantages

1. Support Vector Machines (SVMs) do not directly output probability estimates. Instead, they utilize a computationally intensive method called five-fold cross-validation to calculate them.

2. SVMs may not perform optimally when the dataset contains a significant amount of overlap between the target classes. (Source: Developers, 2007-2022)

3.2.4 Artificial Neural Networks

The computer systems known as Artificial Neural Networks (ANNs), often referred to as Neural Networks or Neural Nets, are modeled after the organic neural networks present in animal brains. Artificial neurons, the interconnected units that make up these systems, imitate the behavior of neurons in biological brains.

An artificial neural network is made up of three main components: inputs, outputs, and transfer functions (ANN). The input units receive changed weights and input values during the network's training phase. The output is determined by a known class, and the weight is adjusted based on the discrepancy between the anticipated output and the actual class (Haq, Li, Memon, Nazir, & Sun, 2018).

Advantages

1. Artificial Neural Networks (ANNs) possess the ability to perform parallel processing.

2. ANNs are able to store data throughout the entire network.

3. ANNs can function with incomplete knowledge.

4. ANNs have a distributed memory structure.

5. ANNs exhibit fault tolerance.(Javatpoint-ANN, 2011-2021).

Disadvantages

1. Artificial Neural Networks (ANNs) have a dependency on specific hardware.

2. Identifying and troubleshooting issues within ANNs can be challenging.

3. The duration of the training process for ANNs is often uncertain

4. ANNs can exhibit unpredictable behavior.(Javatpoint-ANN, 2011-2021).

3.3 Analytic Tools

Different analytic tools were utilized in this project, all of which are open-source and free to use.
The Google Colab platform was employed, utilizing the Python programming language (version 3.9). The libraries utilized for data analysis in Python include:

- Matplotlib 3.2.2
- Numpy 1.21.6
- Pandas 1.3.5
- Scikit-learn 1.0.2
- Tensorflow

3.3.1 Python

Dynamic semantics are used in Python, an interpreted, object-oriented, high-level programming language.

Python is ideal for designing applications fast and for using scripting to connect already-existing components. because it features high-level internal data structures, dynamic type support, and dynamic binding.

Python's simple syntax prioritizes readability, making it simple to learn and reducing the cost of program maintenance. Code reuse and software modularity are further supported by Python's support for modules and packages.(Python Programming Documentation, 2001-2002).

3.3.2 Google Colab

Free to use, Colab is a cloud-based Jupyter notebook environment. It allows users to collaborate online while working with code and data.

Similar to how documents are modified in Google Docs, your team members can simultaneously edit the notebooks you create using Colab. Furthermore, there is no setup necessary before utilizing it.

Numerous well-known machine learning libraries are supported by Colab and are simple to load in your notebook.(Google Colab Documentation, 2022).

3.3.3 Matplotlib

Matplotlib is a cross-platform data visualization and graphical charting package for Python and its numerical extension NumPy. It serves as a strong open-source alternative

to MATLAB. The application programming interfaces (APIs) provided by matplotlib allow programmers to include graphs in their GUI programs (Matplotlib Documentation, 2022).

3.3.4 Numpy

A crucial Python package for scientific computing is called NumPy.It provides a wide range of tools for conducting quick operations on arrays, such as multidimensional array objects, derived objects like masked arrays and matrices, and routines for discrete Fourier transforms, basic linear algebra, fundamental statistical operations, and random simulation (Numpy Documentation, 2008-2022).

3.3.5 Pandas

The Pandas open source Python library is widely used for data science, data analysis, and machine learning. Pandas is a powerful data management tool that is built on top of the Numpy library, which supports multi-dimensional arrays. It is frequently included in all Python distributions, including those from commercial vendors like ActiveState's ActivePython and those bundled with operating systems. It smoothly integrates with other data science modules (Pandas Documentation, 2022).

3.3.6 Scikit-learn

A well-known and dependable Python library for machine learning is Scikit-Learn, sometimes referred to as Sklearn. It offers a consistent user interface for several tools and methods, including dimensionality reduction, clustering, regression, and classification. The library is mostly written in Python and is based on the principles of NumPy, SciPy, and Matplotlib (Scikit-Learn Documentation, 2022).

3.3.7 TensorFlow

TensorFlow, an open-source library developed by Google, supports deep learning applications. Although it was later modified for deep learning, its original purpose was to manage huge numerical computations. Tensors, multi-dimensional arrays, and data flow graphs with nodes and edges are both used in the operation of TensorFlow. TensorFlow code can be distributed more easily across a cluster of GPU-powered machines by employing graph-based execution.

3.4 Performance Evaluation Metrics

Evaluating the performance of a machine learning model is one of the crucial steps in its development. To evaluate the efficacy or caliber of the model, evaluation measures, sometimes referred to as performance metrics, are employed. These metrics enable us to assess how well the model processed the supplied data. The performance of the model can be enhanced by changing the hyper parameters. Performance indicators gauge a machine learning model's ability to generalize to new or unexplored data.

There are five performance evaluation metrics used namely;

- Confusion Matrix
- Precision
- Classification Accuracy
- Recall/Sensitivity
- F1 –Score/F Measure

3.4.1 Confusion Matrix

The confusion matrix is a matrix for evaluating the performance of categorization models given a certain set of test data. The matrix cannot be calculated until the true values of the test data are known. Even while the matrix itself is simple to understand, some of the verbiage used to describe it might be. It also goes by the name of "error matrix" since it displays errors in the model's performance as a matrix. If there are two prediction classes for the classifiers, the matrix is a 2x2 table. The matrix divides the projected values and the actual values into two different dimensions. The values that the model predicts are the predicted values, whereas the values that are actually present in the provided data are the actual values.

The table below illustrates the confusion matrix.

Table 3.3: Description of Confusion Matrix

N = Total predictions	**Actual: No**	**Actual: Yes**
Predicted : No	True Negative (TN)	False Positive (FP)
Predicted : Yes	False Negative (FN)	True Positive (TP)

From the above table:

- True Negative: Both the real or actual value and the model's prediction of "no" were negative.
- True Positive: Both the true or actual value and the model's prediction of yes were accurate.
- False Negative: The model predicted no, but the true value was Yes, another name for it is the Type-II Error.

- False Positive: Despite what the model predicted, the outcome was negative. A Type-I Error is another term for it.

With the aid of the confusion matrix, the other performance metrics like Precision, Accuracy, Recall, F1- Score and Support could be measured.

3.4.2 Precision

Precision may be measured in terms of the number of accurate outputs the model generated or the proportion of accurately predicted positive classes that actually materialized. You can find it by using the following formula:

$$Precision = \frac{TP}{TP + FP}$$

3.4.3 Classification Accuracy

When evaluating the accuracy of a classification task, it is one of the most crucial factors to take into account. It provides information on how frequently the model predicts the outcome properly. It can be computed by dividing the total number of predictions made by the classifier by the percentage of correct predictions. The equation reads as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

3.4.4 Recall/Sensitivity

The percentage of all affirmative classes that our model correctly predicted is known as recall. There has to be a big recall.

$$Recall = \frac{TP}{TP + FN}$$

3.4.5 F1-Score/F-Measure

It is challenging to compare two models that have a high recall but a low precision. F-score can therefore be used in this situation. This score can be used to evaluate recall and precision simultaneously. When recall and precision are equal, the F-score is at its highest. You can find it by using the following formula:

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

3.4.6 Support

The term "support" refers to how many actual instances of a given class there are in a data set. An unequal distribution of support in the training data, which can also imply issues with the reported classifier scores, may be a sign that stratified sampling or rebalancing is necessary. It's important to keep in mind that support is consistent across models but actually refers to the evaluation process.

3.5 Hyper-Parameter Tuning/Optimization

There are several design options accessible when creating a machine learning model to specify the model architecture. As the ideal model design may not always be known for a certain model, it is crucial to weigh different choices. In a real machine learning manner, the computer should be able to investigate and choose the optimum model architecture on its own. By altering the factors that influence the model architecture, also referred to as hyper parameters, hyper parameter tuning is the process of determining the best model architecture.

The Grid Search approach for hyper parameter tuning is one way to find the optimum machine learning algorithm for a particular problem (Jordan, 2017).

3.5.1 Grid Search

Grid search is a frequent method for adjusting hyperparameters. For each combination of the provided hyper parameter values, a model is created, evaluated, and the design that yields the best results is chosen (Jordan, 2017). 80% of the data were utilized for training, and 20% were used for testing.

Chapter 4

Results from Data Analysis and Interpretation

4.1 Introduction

The results of data analysis utilizing several machine learning models, such as K-Nearest Neighbors, Support Vector Machine, Logistic Regression, and Artificial Neural Network, are presented in this chapter (Tensorflow). Using the UCI data to forecast cardiac illness, these models' performance was assessed. Hyper parameter adjustment was also used to improve the models' ability to accurately predict cardiac disease in patients under particular circumstances. 80 percent of the data was used for training, while 20 percent was used for testing. The data contained no missing values, and Python was the study's chosen programming language.

4.2 Graphical Summary of UCI Data set

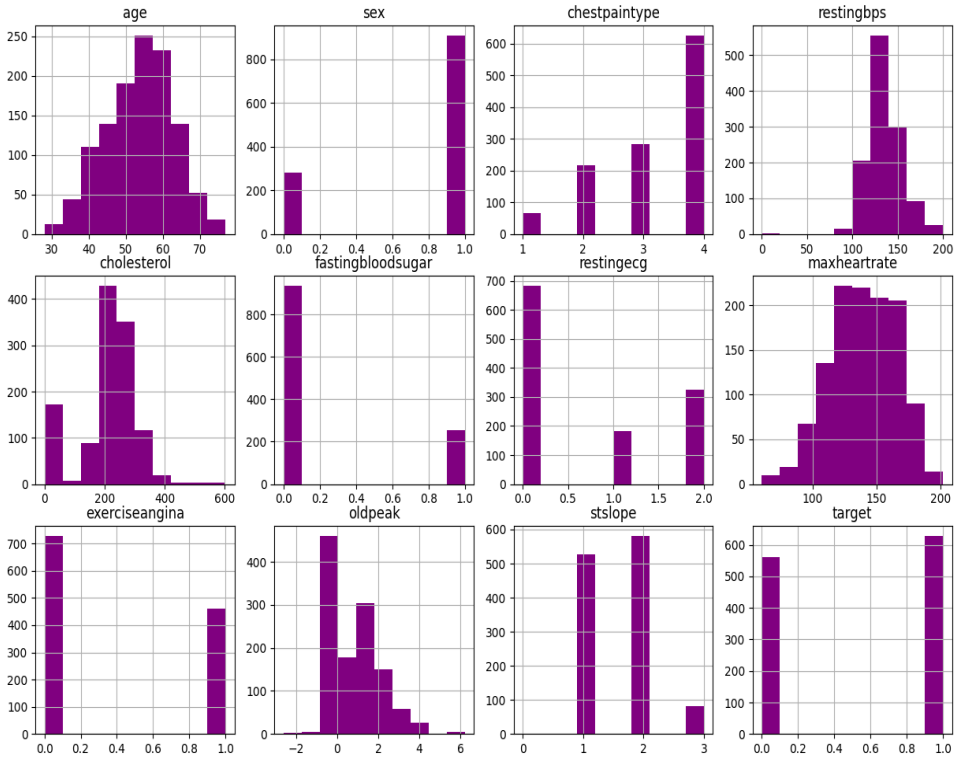

Figure 4.1: Subplots of the features of the Data sets

The Figure 4.1 above is a graphical representation of the UCI data set used for predicting heart diseases. There were 12 features:

- Figure 4.1: a) Distribution of age of patients.
- Figure 4.1: b) The sex of patients.
- Figure 4.1: c) The chest pain type of patients.
- Figure 4.1: d) The level of blood pressure at resting mode in mm/Hg.
- Figure 4.1: e) The level of cholesterol in mg/dl of patients.
- Figure 4.1: f) The fasting blood sugar of patients.
- Figure 4.1: g) The resting electrographic results of patients.
- Figure 4.1: h) The maximum heart rate achieved by patients.
- Figure 4.1: i) The exercise angina of patients.
- Figure 4.1: j) The old peak of patients.
- Figure 4.1: k) The slope of the peak exercise ST segment.
- Figure 4.1: l) The target variable.

The **age** feature has a normal distribution shape which shows the distribution of ages of patients. It is neither skewed to the left nor right to make the predictions bias. About 480 people are between the age of 50 to 60 years from figure 4.1: a).

The **sex** feature is made up of male and females. Ones were used to represent males and Females were zeros. From figure 4.1: b) it shows that the males dominate the females which makes it quiet skewed to the males. The males are approximately 900 while the females are approximately 300.

The **chest pain type** feature was classified into four parts that were;

1 =Typical angina, 2= Atypical angina, 3= Non-anginal pain and
4=Asymptomatic

Typical angina (TA) is characterized by chest pain that is relieved by rest or nitroglycerin and is triggered by physical exertion or emotional stress. However, both at rest and under stress, patients with non-obstructive coronary artery disease (CAD) who are female or old may have atypical symptoms.(Ahmed et al., 2017).

A type of pain that can occur in the chest is called **Atypical angina**, commonly referred to as chest pain that does not meet the criteria for typical angina. Atypical chest pain does not start in the sternum and can radiate to other regions of the body, unlike typical angina, which is defined by a pressure or squeezing-like feeling that typically arises from a lack of oxygen-rich blood to the heart muscle. This kind of chest pain is frequently characterized as happening in situations where the pain is not angina.(Atypical Chest Pain Documentation, 2022).

If a chest pain lasts longer than 30 minutes or less than 5 seconds, intensifies with inhalation, can be brought on by a single trunk or arm movement, can be brought on by local finger pressure, can be brought on by bending forward, or can be immediately relieved by lying down, it is unlikely to be **Non-angina** (Constant, 1990).

A heart attack that is **Asymptomatic**, also known as silent myocardial ischemia (SMI), causes a brief stoppage of blood flow to a section of the heart and may cause scarring and damage to the heart muscle despite having few, no, or unrecognized symptoms.(Loskot & Novotny, 1990).

As seen in Figure 4.1c, many people who have heart disease and pass away from it might not be aware that they have it because there aren't many symptoms. This group of persons belongs to the Asymptomatic category, which has the largest population at roughly 600+. This emphasizes the significance of researching and determining the best machine learning algorithm for categorizing and forecasting cardiac disease. There are roughly 250, 230, and 50 members in the non-anginal, atypical angina, and typical angina categories, respectively.

The blood pressure level at rest, expressed in mm/Hg, is what the term **"The Resting Blood Pressure"** refers to. Less than 120/80 mmHg is considered to be a normal blood pressure reading. According to the Low Blood Pressure Documentation, 1998–2022, a

person has low blood pressure (hypotension) if their reading is less than 90/60 mm/Hg and high blood pressure (hypertension) if it is between 120 and 129 systolic and less than 80 diastolic (High Blood Pressure Documentation, 1998-2022). Figure 4.1d demonstrates that a significant portion of people, more than 1000, had a resting blood pressure of around 100 mm/Hg, indicating that they are likely to have high blood pressure, a major risk factor for heart disease.

In mg/dl, the **The Cholesterol** feature is expressed. All of the cells in the body contain cholesterol, a waxy, fatty-like substance.In order to produce hormones, vitamin D, and substances that aid in food digestion, it is essential. All of the cholesterol that the body need can be produced. But animal-derived foods like cheese, meat, and egg yolks also contain cholesterol (Cholesterol Documentation, 2020). Figure 4.1e indicates that individuals with a healthy heart have a cholesterol level under 200 mg/dl, or around 270+. People who have a cholesterol level between 200 and 239 and weigh more than 400 are considered to be at danger. Finally, persons with cholesterol levels of 240 or greater are regarded as being at a dangerous level, which includes 400+ people.

A popular technique for determining the amount of glucose (sugar) in the body in milligrams per deciliter (mg/dl) is the **Fasting Blood Sugar (FBS)** test. It is employed in the diagnosis of gestational diabetes, diabetes, and prediabetes. The FBS Test Documentation for 2022 specifies that a normal FBS level is 99mg/dl or below. Prediabetes is defined as blood sugar levels between 100 and 125 mg/dl, and high blood sugar, which is an indication of diabetes, is defined as levels of 126 mg/dl or higher.FBS levels were categorized in Figure 4.1f as 0s and 1s, where 0 denotes a level below 120 mg/dl and 1 denotes a level beyond 120 mg/dl. According to this classification, it seems that the majority of the people in the data had normal blood sugar levels, whereas 16 people may have diabetes-related symptoms.

The Resting Electrocardiography (ECG) test assesses the heart's electrical activity. A score of 0 denotes a normal outcome, a score of 1 indicates an ST-T wave irregularity, and a score of 2 denotes left ventricular hypertrophy. In the medical community, a resting ECG score between [-0.5, 0.4] and [2.45, 1.8] is seen as normal, while between [1.4, 2.5] and 0 is regarded as hypertrophy. The majority of patients have normal outcomes, followed by those who have hypertrophy and subsequently ST-T wave anomalies, as shown in Figure 4.1g.

The maximum number of beats a person's heart can pump each minute under extreme stress is indicated by **The Max Heart Rate** function (Waehner, 2022). The maximal heart rate reached was around 120 beats per minute, as shown in Figure 4.1h. Age has an impact on a person's maximal heart rate. A person's current age should be deducted by 220 to determine their maximal heart rate.

Exercise-induced chest pain, commonly known as **The Exercise Angina** characteristic, is a specific type of chest discomfort brought on by decreased blood supply to the heart. Angina Documentation claims that it is a notional feature denoted by 0s and 1s, where 0 stands for "No" and 1 for "Yes." The majority (450 or more) of a sample of over 700 participants were found to have exercise-induced angina.

The Old Peak feature is a measurement of exercise-induced ST-T depression compared to the person's resting state. A low risk is indicated when the range of old peak is less than 2, a moderate risk is indicated when it falls between 1.5 and 4.2, and a high risk is indicated when it is greater than 2.55. According to figure 4.1j, the majority of people in the sample have a low risk, with Old Peak values below 2.

The ST Slope feature is divided into four parts, with 0 representing "Normal", 1 representing "Unsloping", 2 representing "Flat", and 3 representing "Down sloping." According to Figure 4.1 k, there are no individuals with a normal ST slope, while the majority of people have a flat slope, followed by unsloping and then down sloping.

Based on the aforementioned factors, the **Target Variable** feature forecasts whether a person will have a cardiac condition or not. Since it can either be true or false, this makes it a binary feature. A score of 0 indicates that the person is healthy, whereas a value of 1 indicates that the person has cardiac disease. Figure 4.1 l shows that cardiac disease affects the majority of people, however the difference is not great.

4.2.1 Statistical Description of the Data Set

Table 4.1: Descriptive Statistics of the Data Set

	Age	**Sex**	**Chest pain type**	**Resting bps**	**Cholesterol**	**Fasting blood sugar**
Count	1190	1190	1190	1190	1190	1190
Mean	53.720	0.763	3.233	132.154	210.364	0.213
Std	9.358	0.425	0.935	18.369	101.421	0.410
Min	28	0	1	0	0	0
25%	47	1	3	120	188	0
50%	54	1	4	130	229	0
75%	60	1	4	140	269.75	0
Max	77	1	4	200	603	1

	Resting ecg	**Max heart rate**	**Exercise angina**	**Old peak**	**St slope**	**Target**
Count	1190	1190	1190	1190	1190	1190
Mean	0.698	139.733	0.387	0.923	1.624	0.529
Std	0.870	25.517	0.487	1.086	0.610	0.499
Min	0	60	0	-2.6	0	0
25%	0	121	0	0	1	0
50%	0	140.5	0	0.6	2	1
75%	2	160	1	1.6	2	1
Max	2	202	1	6.2	3	1

The descriptive statistics for the data set are displayed in the table above. These statistics, which can be divided into measures of central tendency and measures of dispersion, give an overview of the data. The mean, median, and mode are examples of central tendency measures that give an understanding of the data's center of gravity. The degree of the data's dispersion is shown by measures like the standard deviation, variance, and interquartile range (IQR).

It is impossible to categorize the nominal parameters, such as sex, chest pain kind, fasting blood sugar, exercise angina, st slope, and the goal variable. Since there are no missing values, the count for each feature is the same because the count column displays the number of rows in the data set.

All of the individuals in the data set are roughly 54 years old on average. The data set is divided into two halves by the median, which is also known as the 50th percentile. The median age in this situation is similarly 54. The interval between the 75th percentile (upper quartile) and the 25th percentile (lower quartile), or the interquartile range, is 13 years (60-47). The data set has a maximum age of 77 years and a minimum age of 28 years. The standard deviation is a metric that expresses how far a group of data values might vary from their mean.The age feature's standard deviation in this instance is 9 years, which shows that the data points are in close proximity to the mean.

The Resting blood pressure feature has the minimum value to be 0 mm/Hg and the maximum value to be 200 mm/Hg. The mean is approximately 132 mm/Hg. The 75th percentile is 140mm/Hg and the 25th percentile is 120 mm/Hg so to obtain the interquartile range is 140-120=20 mm/Hg. The median for this feature is 130 mm/Hg with a standard deviation of 18.369 mm/Hg which is low so it's close to the mean.

The minimum and maximum values for the cholesterol characteristic are 0 mg/dl and 603 mg/dl, respectively. The interquartile range is 269.75 - 188 = 81 mg/dl because the 75th percentile is 269.75 mg/dl and the 25th percentile is 188 mg/dl. The mean blood sugar level is 210.364 mg/dl, while the standard deviation is 101.421 mg/dl. 269.75mg/dl is the value of the median.

The Maximum heart rate feature has a minimum heart rate at 60 and a maximum heart rate of 202. The mean for the maximum heart rate is 139.733 with a standard deviation of 25.517. The 25th percentile ,50th percentile and 75th percentile have values 121, 140.5 and 160 maximum heart rate respectively. The interquartile range is 160 -121 = 39.

The mean and standard deviation of the Old Peak feature are both 0.92. This feature's minimum and maximum values are -2.6 and 6.2, respectively. There is a 0.6 median. The interquartile range is 1.6 since the 75th percentile is 1.6 and the 25th percentile is 0, respectively.

4.3 Heart Disease Frequency for Sex

Figure 4.2: Heart Disease Frequency for Sex

As seen in Figure 4.2, the frequency of heart disease is displayed by sex. The green bars represent females while the purple bars represent males. The chart indicates that males are more likely to develop heart disease, with approximately 550 cases shown. In contrast, the number of females with heart disease is significantly lower, with less than 100 patients. Additionally, the chart shows that there are approximately 350 males without heart disease and 205 females without heart disease.

4.3.1 A Heart Disease Frequency for Sex(Trained Data set)

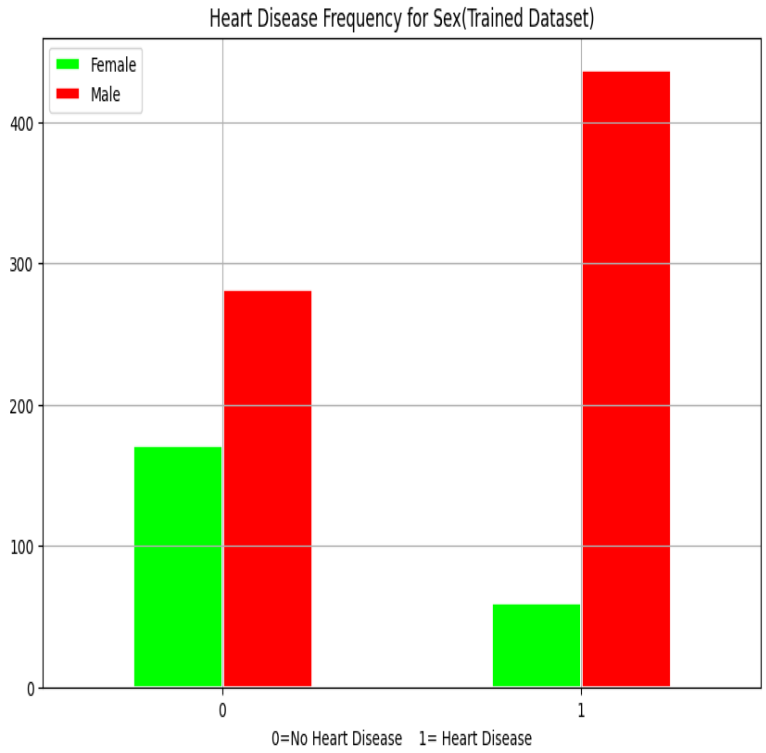

Figure 4.3: Heart Disease Frequency for Sex (Trained Data set)

Figure 4.3 uses only the data set that was trained to an average of 80% to show the frequency of heart disease by gender. Males are represented by the red bars, and females by the green bars. With about 440 cases recorded, the bar chart shows that men are more likely to have heart disease. Comparatively, there are fewer than 100 cases of heart disease among women. Additionally, there are roughly 280 males and 180 females without cardiac disease in the general population. According on this data, it's possible that the model has a bias in favor of men.

4.3.2 A Heart Disease Frequency for Sex(Test Data set)

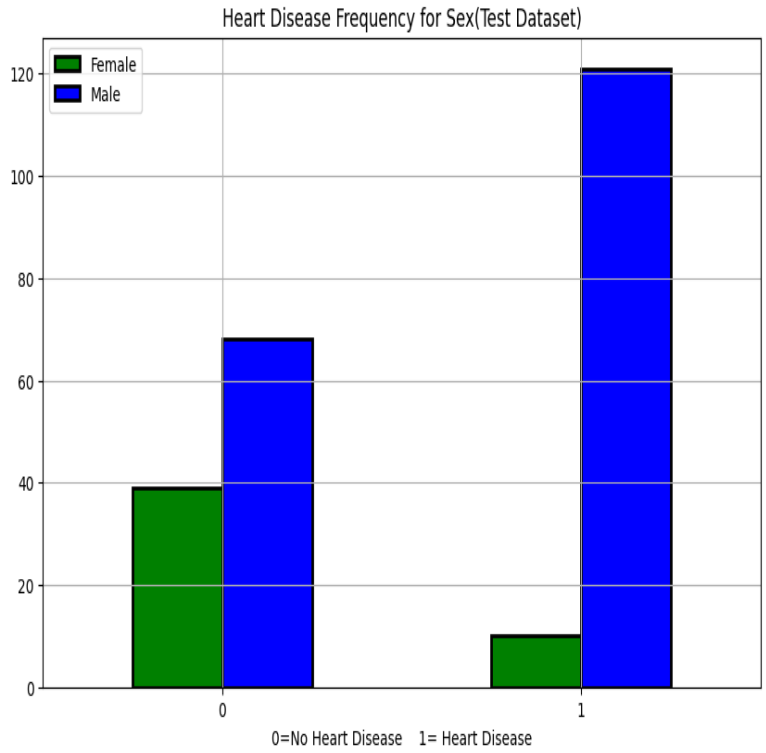

Figure 4.4: Heart Disease Frequency for Sex (Test Data set)

Figure 4.4 illustrates the frequency of heart disease by gender, using 20% of the test data set. The bars are colored green for females and blue for males. The chart indicates that males are more likely to have heart disease, with approximately 121 males affected. In contrast, the number of females with heart disease is significantly lower, with less than 10 patients. Additionally, the chart shows that there are approximately 70 males and 40 females who do not have heart disease.

4.3.3 A Statistical Relationship Between Age and Max Heart Rate

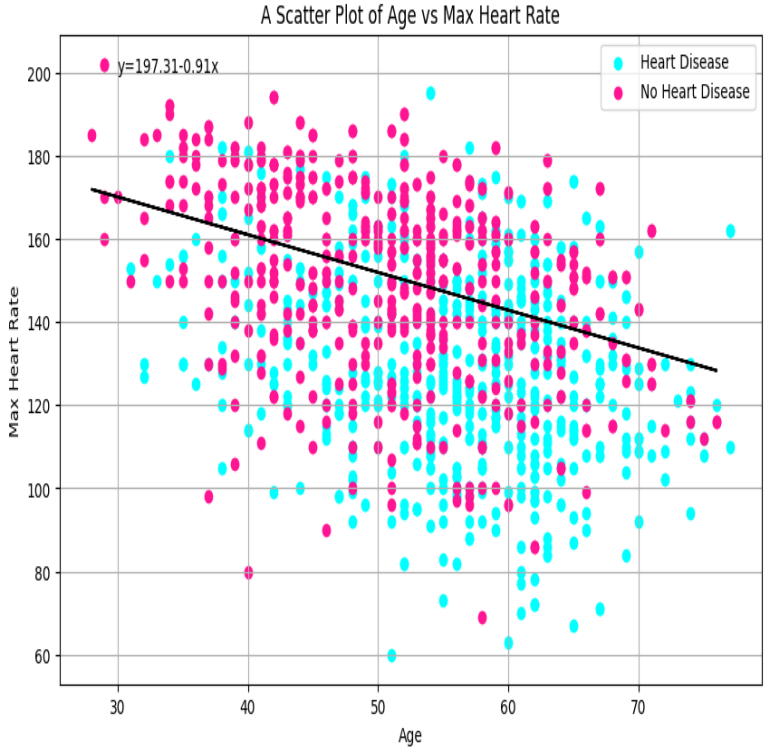

Figure 4.5: A Statistical Relationship Between Age and Max Heart Rate

The graph shows the statistical correlation between a patient's age and maximal heart rate. Age and maximal heart rate have a negative association, according to the line of best fit, which is depicted in black. This implies that heart rate declines with age. The line of best fit has an equation of 197.31 - 0.91x, which supports the negative association. Patients with heart illness are represented by the sea blue color, while those without heart disease are represented by the pink color. In conclusion, the graph illustrates how an individual's age and maximal heart rate affect their likelihood of receiving a heart disease diagnosis.

4.3.4 A Heart Frequency per Chest Pain Type

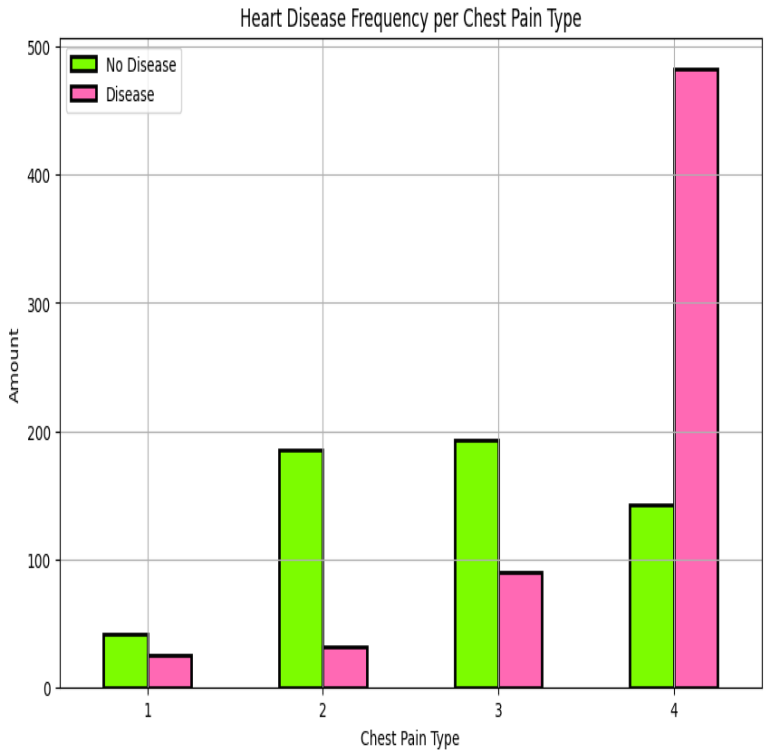

Figure 4.6: Heart Disease Frequency per Chest Pain Type

The four different types of chest discomfort and the proportion of people with or without heart disease are depicted in a bar graph in Figure 4.6. The four different categories of chest pain include asymptomatic, non-anginal, and classic angina. There are roughly 480 people with cardiac disease and 130 without it in the Asymptomatic group. Those who have typical angina are the group with the lowest number. There are roughly the same numbers of people with atypical angina and non-anginal pain who do not have heart disease. Pink denotes the presence of heart illness in the patient, while green denotes the absence of heart disease.

4.4 Correlation Matrix

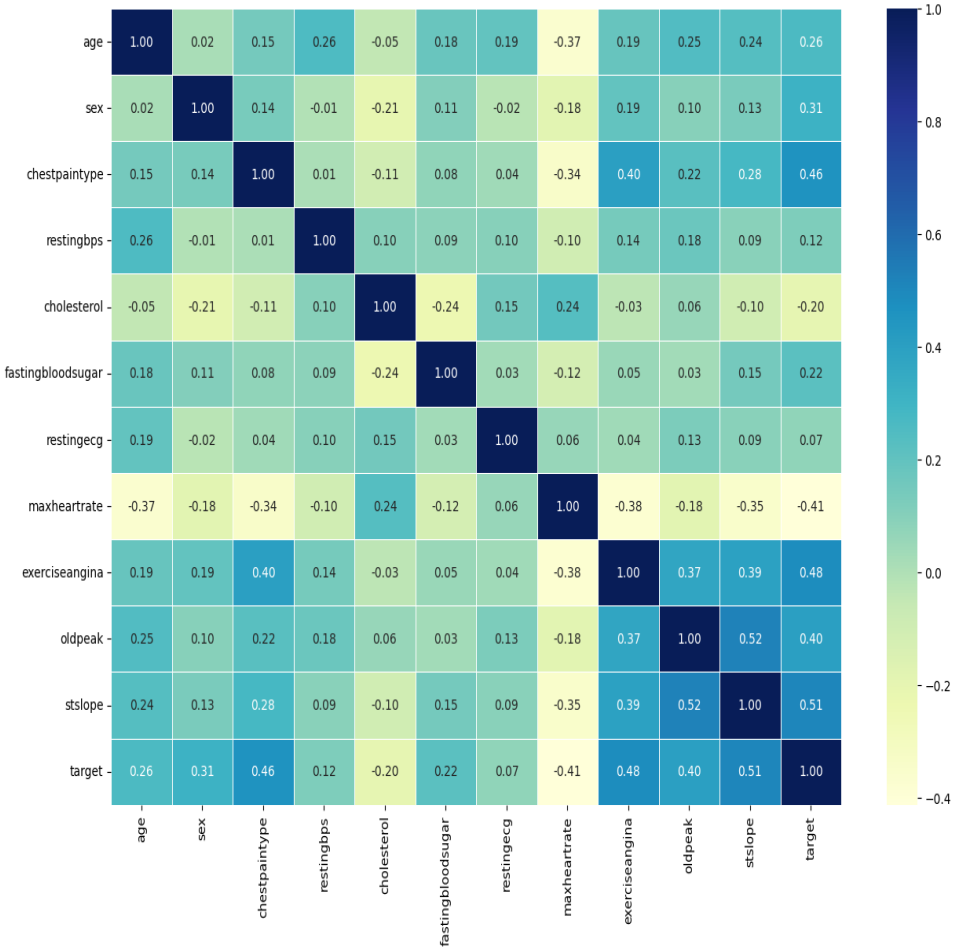

Figure 4.7: Correlation Matrix

The degree and direction of the linear link between two variables are determined using the correlation matrix displayed in Figure 4.7. The SAS Annotated Output Documentation for Process CORR, 2021 The matrix's values vary from -1 to 1, where -1 denotes a perfect negative correlation, 0 a perfect zero correlation, and 1 a perfect positive correlation. Because every variable that is related to itself will always have a correlation of 1, the graph's main diagonal elements all have a correlation of 1.

4.5 Performance of Machine Learning Algorithms

4.5.1 Before Hyper-Parameter Tuning

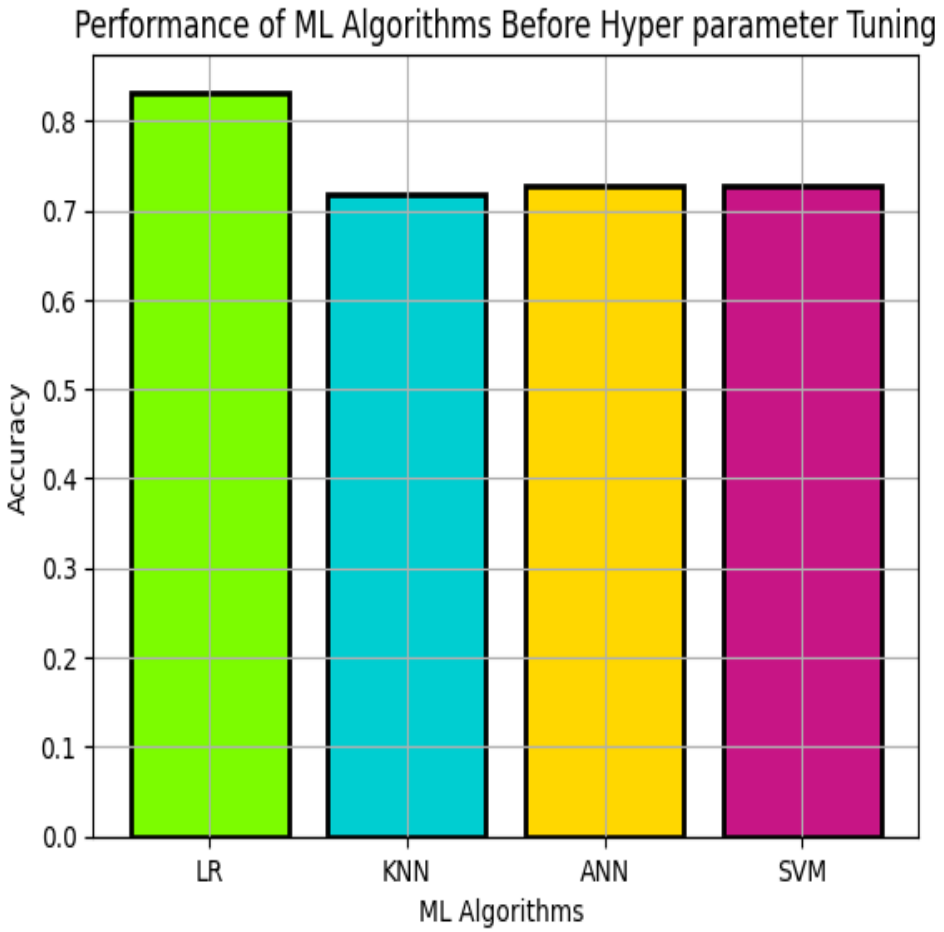

Figure 4.8: Performance of Machine Learning Algorithms Before Hyper-Parameter Tuning

Before applying hyperparameter tweaking to the data, Figure 4.8 shows the output of four machine learning algorithms. In order to keep the findings consistent, a seed of 42 was utilized. An 80 percent training set and a 20 percent testing set were created from the data. The algorithm for predicting cardiac illnesses with the highest accuracy, almost 83%, was logistic regression. The K-Nearest Neighbors and Support Vector Machine algorithms came next, with accuracy rates of 72% and 73%, respectively. The accuracy of the deep learning technique known as the Artificial Neural Network was 73%.

4.5.2 After Hyper-Parameter Tuning

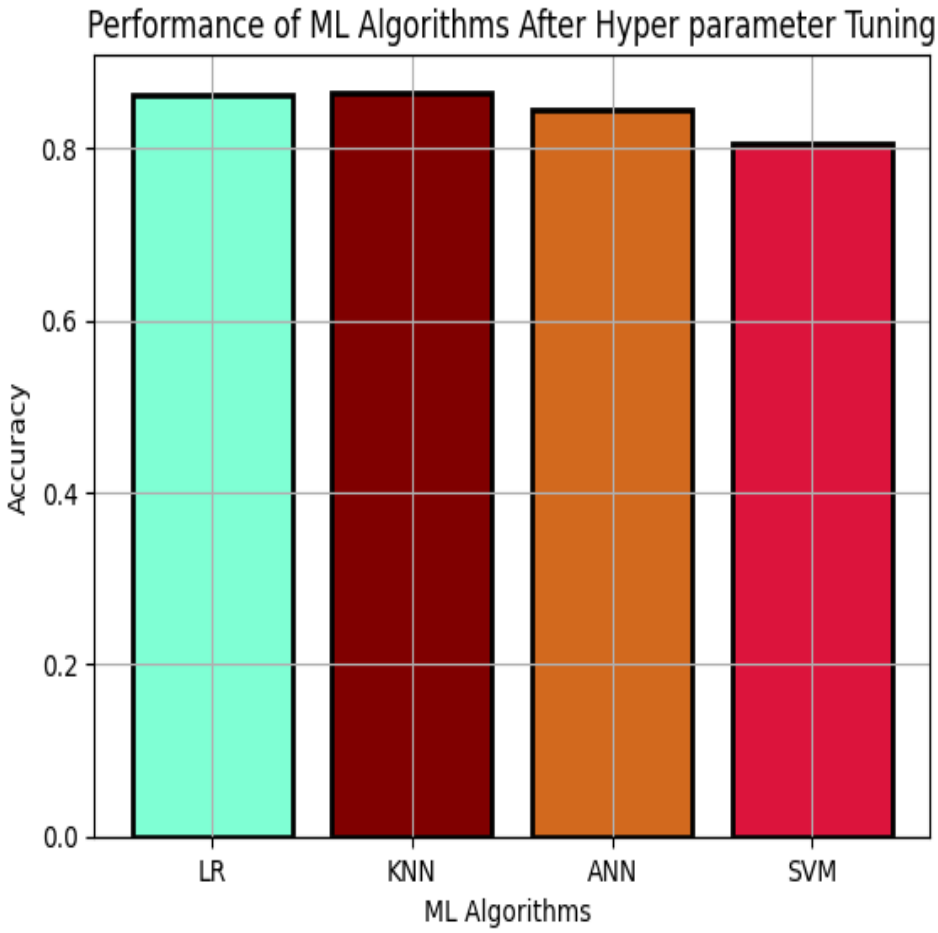

Figure 4.9: Performance of Machine Learning Algorithms After Hyper-Parameter Tuning

The performance was significantly enhanced after applying several machine learning methods to the data set for hyper-parameter tuning. The Logistic Regression, Support Vector Machine, and K-Nearest Neighbors algorithms were tuned using the GridSearchCV technique. The Sequential model in TensorFlow was used to fine-tune the artificial neural network. In order to attain the best outcomes for each algorithm, many parameters were changed.

The Logistics Regression best parameters were;
['C': 100, 'penalty': 'l2', 'solver': 'newton-cg']

The K-Nearest Neighbor best parameter were;
['metric': 'manhattan', 'n_neighbors': 15, 'weights': 'distance']

The Support Vector Machine best parameters were;
['C': 50, 'gamma': 'scale', 'kernel': 'poly']

The Artificial Neural Network best parameter was;
['epochs': 40]

Figure 4.9 shows that, with the use of hyper parameter adjustment, all algorithms have an accuracy of over 80%. The Support Vector Machine (SVM) performs well for predicting cardiac disease in patients despite having the lowest accuracy, which is 81%. While Logistic Regression has an accuracy of 86 percent, Artificial Neural Networks (ANN) have an accuracy of 84 percent. The K-Nearest Neighbors algorithm has an 87 percent accuracy rate.

4.6 Performance Evaluations of Machine Learning Algorithms

According to these four (4) measures, the machine learning algorithms' performance was assessed:
- Accuracy
- Precision
- Recall
- F1-Score

Table 4.2: Performance Metrics of the Machine Learning Algorithms

Machine Learning Algorithms	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Logistic Regression	0.86	0.87	0.88	0.87
Support Vector Machine	0.81	0.81	0.84	0.83
K - Nearest Neighbors	0.87	0.86	0.90	0.88
Artificial Neural Networks	0.84	0.84	0.88	0.86

The four machine learning algorithms' performance indicators were examined using Accuracy, Precision, Recall, and F1-Score. Table 4.2's findings reveal that the K-Nearest Neighbor algorithm delivered the greatest results, with accuracy rates of 87%, precision rates of 86%, recall rates of 90%, and F1-Score rates of 88%.

The accuracy, precision, recall, and F1-Score of the Logistic Regression algorithm were all determined to be 86 percent, 87 percent, and 88 percent respectively, making it perform as well as the K-Nearest Neighbor technique.

The Support Vector Machine had an accuracy of 81%, which was not as high as the Artificial Neural Networks' accuracy of 86%. The Precision of the Support Vector Machine was 81% with a recall of 84% and F1-Score of 83%.

Last but not least, the deep learning algorithm Artificial Neural Networks had an accuracy of 84 percent, precision of 84 percent, recall of 88 percent, and F1-Score of 86 percent.

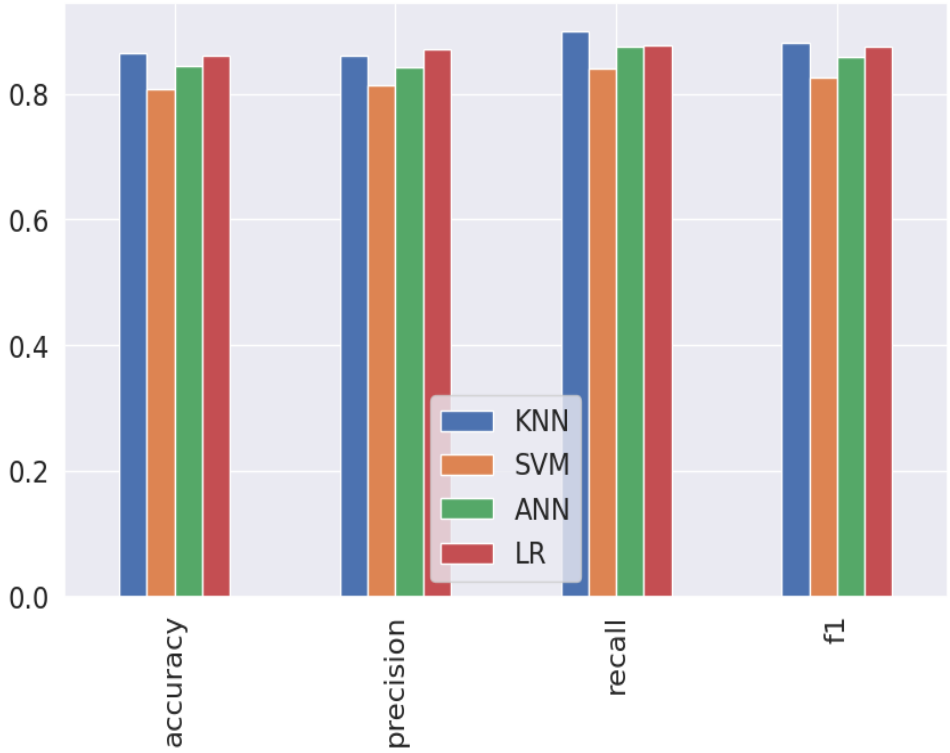

Figure 4.10: A Graph of the Performance Metrics of the Machine Learning Algorithms

The data in Table 4 are shown graphically in the graph in Figure 4.10.It compares various machine learning algorithms using different performance metrics. As shown by this graph and other analyses, the K-Nearest Neighbor algorithm appears to be the most effective option for predicting heart diseases based on the given dataset. This is due to its superior performance after hyperparameter tuning and its strong performance across multiple metrics.

4.7 Confusion Matrix

Machine learning algorithms can more correctly predict patients' heart problems with the help of hyper-parameter tweaking. It is important to keep in mind that improving the algorithms by themselves is insufficient to determine the performance of the model. This is because there's a chance that the sample size will be uneven, which could bias the results. A useful tool for providing a more detailed evaluation of the model's performance is the confusion matrix. The confusion matrices for a number of machine learning techniques, including K-Nearest Neighbors, Support Vector Machine, Logistic Regression, and Artificial Neural Networks, are shown in the diagrams below.

4.7.1 Logistic Regression

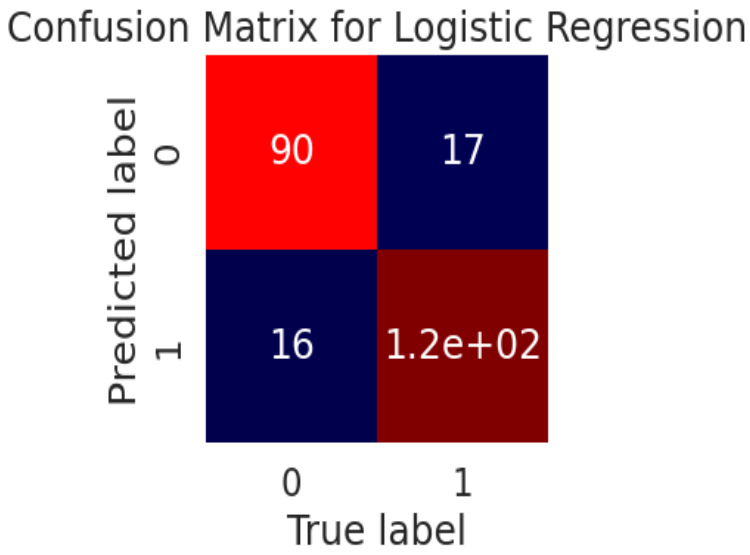

Figure 4.11: Confusion Matrix for Logistic Regression

The confusion matrix of the Logistic Regression model can be seen in Figure 4.11, and it offers more details than simply the model's precision. The true labels for individuals without heart disease in the first column are 90 + 16 = 106 and for those with heart disease are 115 + 17 = 132. The sum of the boxes, which is equal to 106 + 132 = 238 samples, represents the total number of samples analyzed.

There were 90 people tested who were appropriately diagnosed as not having heart disease whereas there were 120 people who actually have the condition. When someone is diagnosed with heart disease when they truly have the condition, this is referred to as a Type 1 error and is also referred to as "False Positives." In this instance, the inaccuracy had an impact on 17 patients. False negatives, or Type II errors, occur when a patient is misdiagnosed as having cardiac disease despite not having the condition. In this instance, the inaccuracy had an impact on 16 patients.

Table 4.3: Classification Report for Logistic Regression

	precision	**recall**	**f1-score**	**support**
0	0.85	0.84	0.85	107
1	0.87	0.88	0.87	131
accuracy			0.86	238
macro avg	0.86	0.86	0.86	238
weighted avg	0.86	0.86	0.86	238

The classification report for the Logistic Regression algorithm, which contains precision, recall, f1-score, accuracy, support, macro average, and weighted average, may be derived from Figure 4.11. Precision, recall, and F1 Score are the three often used measures to rate the model's quality.

From Table 4.3, we can see that:

Precision is the proportion of patients who actually had the cardiac disease that the model had predicted they would have. Precision in this instance is 87%.

Recall is the proportion of patients for whom the algorithm's diagnosis of heart disease was accurate. Recall in this instance is 88%.

The F1-score for the Logistic Regression is 87 percent, which is quite close to 1, showing that the model is successful at predicting whether a patient has a cardiac condition or not.

Support shows how many patients are included in each class of the data set. 131 people in this example had heart disease, compared to 107 patients who did not.

The accuracy for the Logistic Regression is 86%.

4.7.2 K-Nearest Neighbor

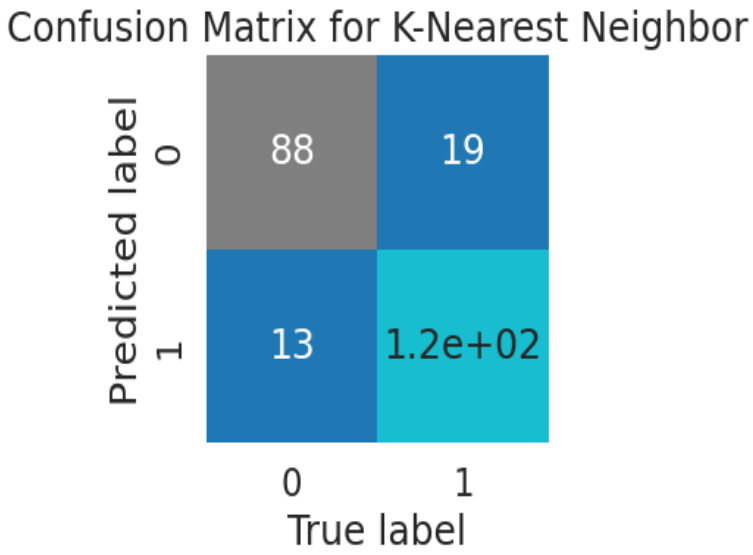

Figure 4.12: Confusion Matrix for K- Nearest Neighbor

Figure 4.12's confusion matrix for the K-Nearest Neighbor technique provides more information than just the model's precision. According to the first column of the real label, 137 persons had heart disease whereas just 101 did not. The total number of samples that were examined was 238.

88 of the persons tested were appropriately diagnosed as not having heart disease, whereas 120 were correctly identified as having heart disease. 19 people who were actually determined to have heart disease but were misdiagnosed as not having it experienced the Type 1 error, often known as "False Positives." 13 patients who did not have heart disease but were given a diagnosis of having it experienced type II mistake, also known as "False Negatives."

Table 4.4: Classification Report for K-Nearest Neighbor

	precision	**recall**	**f1-score**	**support**
0	0.87	0.82	0.85	107
1	0.86	0.90	0.88	131
accuracy			0.87	238
macro avg	0.87	0.86	0.86	238
weighted avg	0.87	0.87	0.87	238

The K-Nearest Neighbor algorithm's classification report, which contains precision, recall, f1-score, accuracy, support, macro average, and weighted average, may be derived from Figure 4.12. Precision, recall, and F1 Score are the three often used measures to rate the model's quality.

From Table 4.4, we can see that:

Precision is 86%, meaning that only 86% of the patients that the model predicted would have heart disease actually had it.

Recall is 90%, meaning that only 90% of those who actually had heart disease were correctly predicted by the algorithm.

The K-Nearest Neighbor's F1-score is 88 percent, which is quite close to 1, indicating that the model is very effective at determining whether a patient has heart disease or not.

Support provides information on the number of patients in each data set class. 131 people had heart disease, compared to 107 persons who did not.

The K-Nearest Neighbor method has an accuracy rate of 87%.

4.7.3 Support Vector Machine

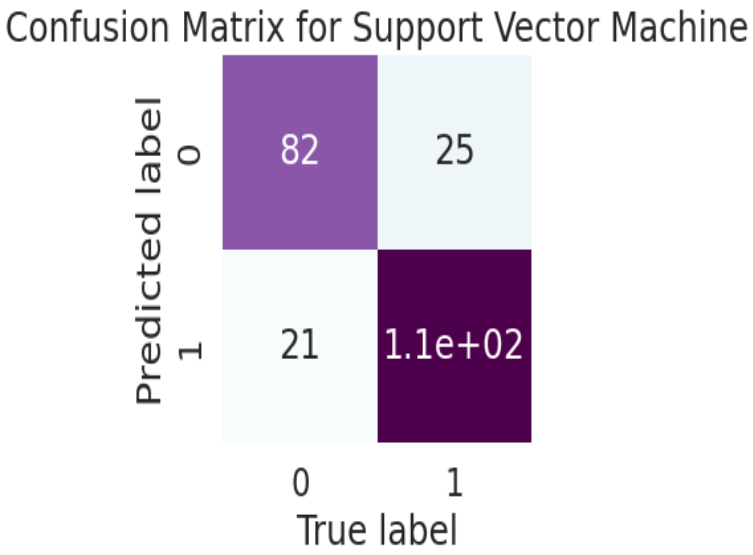

Figure 4.13: Confusion Matrix for Support Vector Machine

The Support Vector Machine's confusion matrix, which is depicted in Figure 4.13, gives more details than just the model's accuracy. The genuine label's first column for people without heart disease is 103 (82+21), while the true label's first column for people with heart disease is 135 (25+110). There were 238 samples analyzed in all (103+135).

In contrast to the 110 persons who had the condition and were accurately diagnosed as such, 82 people who were tested and did not have heart disease were correctly identified as such. 25 patients experienced a type 1 error, sometimes referred to as "False Positives," which refers to those who genuinely have heart disease but were misdiagnosed with it. There were 21 patients who experienced a type II error, sometimes known as "False Negatives," in which the diagnosis of cardiac disease was made despite the absence of the condition.

Table 4.5: Classification Report for Support Vector Machine

	precision	**recall**	**f1-score**	**support**
0	0.80	0.77	0.78	107
1	0.81	0.84	0.83	131
accuracy			0.81	238
macro avg	0.81	0.80	0.80	238
weighted avg	0.81	0.81	0.81	238

The classification report for the Support Vector Machine can be seen in Figure 4.13 and includes statistics like precision, recall, f1-score, accuracy, support, macro average, and weighted average. Precision, recall, and F1 Score are the three often used measures to assess the model's quality.

According to Table 4.5, precision shows that only 81% of the patients with heart disease who the model predicted to have it actually did. Recall indicates that the algorithm only properly predicted 84% of those who actually had heart disease. The Support Vector Machine's F1-score is 83 percent, which is quite near to 1, demonstrating the model's great accuracy in determining whether a patient has heart disease or not.

According to Support, there were 107 people without heart disease and 131 patients with the condition, making up each class of the data set.

The Support Vector Machine has an accuracy rate of 81%.

4.7.4 Artificial Neural Networks

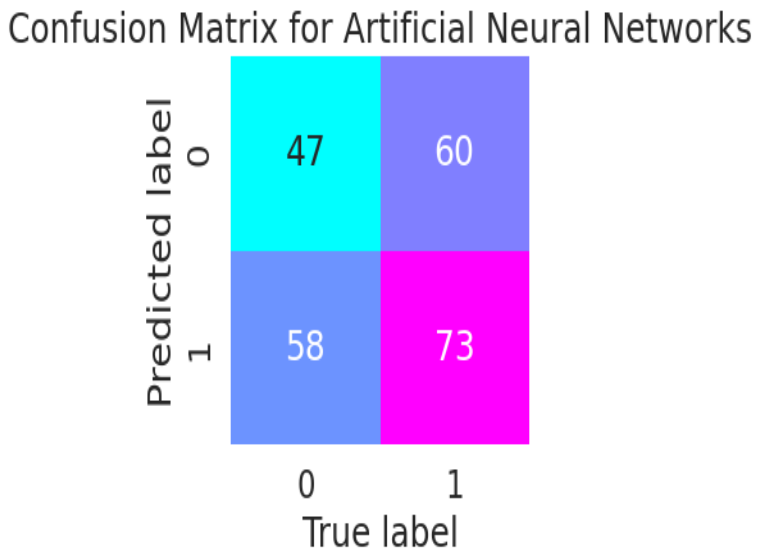

Figure 4.14: Confusion Matrix for Artificial Neural Networks

Figure 4.14's representation of the Artificial Neural Network's confusion matrix reveals more details about the model than just its correctness. The real label's first column for people without heart disease is 105 (47+58), while the column for people with heart disease is 133 (60+73). There were 238 samples analyzed in all (133+105).

47 individuals underwent testing, were found to be free of heart disease, and had their diagnoses confirmed. Similarly, 73 patients had a heart disease diagnosis that was accurate. 58 patients had the condition misdiagnosed (Type II error or "False Negatives") while 60 had it incorrectly classified as not having it (Type 1 error or "False Positives").

Table 4.6: Classification Report for Artificial Neural Networks

	precision	**recall**	**f1-score**	**support**
0	0.45	0.44	0.44	107
1	0.55	0.56	0.55	131
accuracy			0.50	238
macro avg	0.50	0.50	0.50	238
weighted avg	0.50	0.50	0.50	238

The classification report for the artificial neural network, which includes statistics like precision, recall, f1-score, accuracy, support, macro average, and weighted average, may be derived from Figure 4.14. Precision, recall, and F1 Score are the three measures that are most frequently used to assess the model's level of quality.

Table 4.6 demonstrates that the model's accuracy is 55%, which means that only 55% of the patients who the model predicted would develop heart disease actually did. Only 56% of individuals who actually had heart disease were properly predicted by the algorithm, according to the model's recall, which stands at 56%. The Artificial Neural Network's F1-score is 55 percent, which is fairly near to 1, indicating that the model does a respectable job of predicting whether a patient has heart disease or not.

Support tells us how many patients belonged to each class of the data set. 107 patients did not have heart disease while 131 patients had the disease.

The accuracy for the Artificial Neural Network is 50%. This confirms the initial assumption that was made in Figure 8 about the bias in the data causing the Artificial Neural Network to underperform (overfitting)."

Chapter 5

Summary, Conclusions, Limitations and Recommendations

5.1 Introduction

This chapter provides an extensive summary of the research and makes important conclusions. The goal of this study was to choose the top machine learning algorithm for quickly recognizing heart issues. The chapter is divided into four sections: summary, study limitations, recommendations, and conclusion.

5.2 Summary

In this study, K-Nearest Neighbors, Support Vector Machine, Logistic Regression, and Artificial Neural Networks were utilized to classify and predict heart illness in patients. The performance of the K-Nearest Neighbor approach was enhanced by hyper-parameter tweaking. Performance measures like as accuracy, recall, and f1-score were used to evaluate the models. The models' accuracy significantly improved after hyper-parameter tweaking. The distribution of the data was skewed since more men than women had heart disease, but it also suggested that men might be more susceptible to the condition than women. The study also found a link between patient age and maximum heart rates that was unfavorable. K-Nearest Neighbor was the most effective model overall, despite the precision rate of 87 percent for Logistic Regression.

5.3 Conclusions

Heart disease is the leading cause of death worldwide. Despite efforts by organizations and medical professionals to identify and treat the ailment, it may be difficult to recognize some early symptoms, which can lead to preventable fatalities.

K-Nearest Neighbors, Support Vector Machine, Logistic Regression, and Artificial Neural Networks were employed in this study as four machine learning techniques to aid in the identification of heart illness. When the data were split into 20% for testing and 80% for training, Logistic Regression showed the highest accuracy at 83 percent before to hyper-parameter tuning. The best technique was found to be K-Nearest Neighbors after hyper-parameter tuning with GridSearchCV. Python programming and data from the UCI machine learning repository were used in the study. The Artificial Neural Network, a deep learning approach, was found to be overfitted as a result of the small and biased data set.

Precision, recall, accuracy, and f1-score were just a few of the performance criteria where K-Nearest Neighbors excelled. However, it was discovered that the precision of logistic regression was particularly strong.

In conclusion, the best machine learning model for recognizing and classifying heart illness is the K-Nearest Neighbors approach.

5.4 Recommendations

The research's conclusions lead to the following recommendations:

- I advise modeling these machine learning algorithms with a larger data set because they work best when given more data, especially the Artificial Neural Network.

- The number of male and female patients should be balanced in the data set to avoid skewing the results.

- I advise investigating the application of other machine learning algorithms for heart disease prediction in future studies.

5.5 Limitations

This study had a number of shortcomings, such as:

- The size of the data set employed had an impact on how well the artificial neural network performed (TensorFlow).
- The available computational power was not sufficient to try a wider range of hyper parameters with GridSearchCV to improve the model.
- The bias in the data prevented some models from performing as expected.
- Since the majority of the data came from guys, removing some of the males from the data set to avoid bias would have made the data set too small for analysis.

References

[1] Ahmad, E., Tiwari, A., & Kumar, A. (2020). *Cardiovascular Diseases (CVDs) Detection using Machine Learning Algorithms. International Journal for Research in Applied Science & Engineering Technology (IJRASET, 2321-9653).*

[2] Ahmed, A., Leong, D., Merz, C., Wei, J., Handberg, E., Shufelt, C.& Cook-Wiens, G. C. (2017). *Typical angina is associated with greater coronary endothelial dysfunction but not abnormal vasodilatory reserve. National Libraty of Medicine, 886-891.*

[3] Akanksha, G., Shubham, P., & Prof., K. (2017).*An Evaluation of Supervised Machine Learning Algorithms for Heart Disease Diagnosis. International Research Journal of Computer Science(IRJCS), 33-42.*

[4] Akhila, M., Mahalakshmi, N., & Niriksha, N. (2022). *Prediction of Heart Disease and Diabetes using Machine Learning. International Journal of Innovative Technology and Research (IJITR), 16.*

[5] Angina Documentation. (1998-2022).*Mayoclinic. Retrieved from https://www.mayoclinic.org/diseases-conditions/angina/symptoms-causes/syc-20369373*

[6] Archana, S., & Rakesh, K. (2020). *Heart Disease Prediction Using Machine Learning Algorithms. International Conference on Electrical and Electronics Engineering(ICE3), 452-457.*

[7] Atypical Chest Pain Documentation. (2022). *Ppschicago. Retrieved from https://www.ppschicago.com/pain-management/chest-pain/atypical-chest-pain/*

[8] Chandu, D., Ch, S., Darshan, V., Dereddy, P., & Karthik, M. (2022).*Predicting the Risk of having Heart Disease using Machine Learning Techniques. International Research Journal of Engineering and Technology (IRJET).*

[9] Cholesterol Documentation. (2020, December 10).*Medlineplus. Retrieved from https://medlineplus.gov/cholesterol.html*

[10] Constant, J. (1990).*The diagnosis of nonanginal chest pain . National Library of Medicine, 187-192.*

[11] Deepika, S. (2019, July 19). *Pluralsight. Retrieved from https://www.pluralsight.com/guides/interpreting-data-using-descriptive-statistics-python*

[12] Developers, S.-L. (2007-2022). *Scikit-Learn. Retrieved from https://scikit-learn.org/stable/modules/svm.html*

[13] Dhai, E., Abdelkamel, A., & Tahar, K. (2021).*Using Machine Learning for Heart Disease Prediction. Research Gate, 70-81.*

[14] Fasting Blood Sugar Test Documentation. (2022). *https://my.clevelandclinic.org/health/diagnostics/21952-fasting-blood-sugar. Retrieved from https://my.clevelandclinic.org/health/diagnostics/21952-fasting-blood-sugar*

[15] Fernandes, A., & Gonc¸alo, J. L. (2022). *Heart Disease Prediction and Classifcation using Machine Learning.*

[16] Galla, S. S., Munaga, M., Manchuri, S., & Rajalakshmi. (2020). *Heart Disease Prediction Using Machine Learning Techniques. International Research Journal of Engineering and Technology (IRJET).*

[17] Google Colab Documentation. (2022). *Tutorialspoint. Retrieved from https://www.tutorialspoint.com/googlecolab /whatisgooglecolab.html*

[18] Gunturu, D., Cherukuri, S., Koruprolu, N., & Kesuboyina, H. (2017-2021). *Heart Disease Prediction Using Machine Learning Algorithms.*

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

[20] High Blood Pressure Documentation. (1998-2022). *Mayoclinic. Retrieved from https://www.mayoclinic.org/diseases-conditions/high-blood-pressure/symptoms-causes/syc-20373410*

[21] Javatpoint-ANN. (2011-2021). *JavaTpoint. Retrieved from https://www.javatpoint.com/artificial-neural-network*

[22] JavaTpoint-KNN. (2011-2021). *Javatpoint. Retrieved from https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning*

[23] JavaTpoint-LR. (2011-2021). *Javatpoint. Retrieved from https://www.javatpoint.com/logistic-regression-in-machine-learning*

[24] Javatpoint-SVM. (2011-2021).*Javatpoint. Retrieved from https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm*

[25] Jonnavithula, S. K., Jha, A. K., Kavitha, M., & Srinivasulu, S. (2022). *Role of machine learning algorithms over heart diseases prediction. AIP Conference Proceedings, 040013.*

[26] Jordan, J. (2017, Nov 2).*jeremyjordan. Retrieved from https://www.jeremyjordan.me/hyperparameter-tuning/*

[27] Karthiga, A. S., & Mary, M. S. (2022). *A predictive analysis of heart diseases using machine learning algorithms. International Journal of Health Sciences, 6(S3), 3108-3125. Retrieved from https://doi.org/10.53730/ijhs.v6nS3.6309*

[28] Katarya, R., & Srinivas, P. (2020). *Predicting heart disease at early stages using machine learning: a survey. International Conference on Electronics and Sustainable Communication Systems (ICESC), 302-305.*

[29] Khvoynitskaya, S. (2020, January 30th). *The future of big data: 5 predictions from experts for 2020-2025. Retrieved from Itransition: https://www.itransition.com/blog/the-future-of-big-data*

[30] Kumar, M. N., Koushik, K. V., & Deepak, K. (2018).*Prediction of heart diseases using data mining and machine learning algorithms and tools. International Journal of Scientific Research in Computer Science, Engineering and Information Technology,, 887-898.*

[31] Kwakye, K., & Dadzie, E. (2021).*Machine Learning-Based Classification Algorithms for the Prediction of Coronary Heart Diseases.*

[32] Li, J. P., Haq, A. U., Din, S. U., Khan, J., Khan, A., & Saboor, A. (2020). *Heart disease identification method using machine learning classification in e-healthcare. IEEE Access, 8, 107562-107582.*

[33] Loskot, F., & Novotny, P. (1990). *Asymptomatic myocardial ischemia. National Library of Medicine, 370-373*

[34] Low Blood Pressure Documentation. (1998-2022). *Mayoclinic. Retrieved from https://www.mayoclinic.org/diseases-conditions/low-blood-pressure/symptoms-causes/syc-20355465*

[35] Matplotlib Documentation. (2022).*Activestate. Retrieved from https://www.activestate.com/resources/quick-reads/what-is-matplotlib-in-python-how-to-use-it-for-plotting/*

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

[37] Mursal, F., Adnan, A., Hiba, R., Sanam, N., & Kanwal, A. (2020). *Heart Disease Prediction using Machine Learning Algorithms. International Conference on Computational Sciences and Technologies.*

[38] Nagaraj, M., Lutimath, C. C., & Pol, B. S. (2019).*Prediction of Heart Disease using Machine Learning. International Journal of Recent Technology and Engineering (IJRTE), 2277-3878.*

[39] Nashif, S., Raihan, M., Islam, M., & Imam, M. (2018).*Heart Disease Detection by Using Machine Learning Algorithms and a Real-Time Cardiovascular Health Monitoring System. World Journal of Engineering and Technology, 854-873.*

[40] Nikhil, B., Sreedevi, G., & Ahmad, H. (2022) *Using machine learning to Predict Heart Disease. WSEAS TRANSACTIONS on BIOLOGY and BIOMEDICINE, 2224-2902.*

[41] Nikita, S., Shashank, G., & C., R. (2021). *Comparison of Various Machine Learning Algorithms for Heart Disease Prediction. Journal of Emerging Technologies and Innovative Research(JETIR), 2349-5162.*

[42] Nishadi, A. (2019).*Predicting Heart Diseases In Logistic Regression Of Machine Learning Algorithms By Python Jupyterla. International Journal of Advanced Research and Publications.*

[43] Nishadi, A. T. (2019).*Predicting heart diseases in logistic regression of machine learning algorithms by Python Jupyterlab. International Journal of Advanced Research and Publications, 1-6.*

[44] Numpy Documentation. (2008-2022). *Numpy. Retrieved from https://numpy.org/doc/stable/user/whatisnumpy.html*

[45] Obasi, T., & Shafiq, M. O. (2019). *Towards comparing and using Machine Learning techniques for detecting and predicting Heart Attack and Diseases. IEEE, 2393-2402.*

[46] Onel, H. (2018, September 10).*Towards Data Science. Retrieved from https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761*

[47] Ozichi, N., Segun, A., Ayodeji, I., Madamidola, O. A., & Aderibigbe, T. (2019). *Predictive System for Heart Disease Using a Machine Learning Trained Model. International Journal of Computer (IJC), 140-152.*

[48] Pandas Documentation. (2022). *Activestate. Retrieved from https://www.activestate.com/resources/quick-reads/what-is-pandas-in-python-everything-you-need-to-know/*

[49] Patel, M., Patange, R., Patil, C., & Kapoor, A. (2022). *Predicting Heart Disease Using Machine Learning Algorithms. International Research Journal of Engineering and Technology.*

[50] Prathamesh, K., Pratik, P., Kaustubh, L., & Rovina, D. (2022). *Heart Disease Prediction using Machine Learning. International Research Journal of Engineering and Technology(IRJET).*

[51] Praveen, K. R., Sunil, K. R., Balakrishnan, S., Syed, M. B., & Ravi, K. (2019).*Heart Disease Prediction Using Machine Learning Algorithm. International Journal of Innovative Technology and Exploring Engineering (IJITEE), 2278-3075.*

[52] Proc CORR — SAS Annotated Output Documentation. (2021). *Stats.oarc.ucla. Retrieved from https://stats.oarc.ucla.edu/sas/output/proc-corr/*

[53] Python Programming Documentation. (2001-2002). *Python. Retrieved from https://www.python.org/doc/essays/blurb/*

[54] Ramesh, T. R., Lilhore, U. K., Poongodi, M., Simaiya, S., Kaur, A., & Hamdi, M. (2022). *Predictive Analysis of Heart Diseases with Machine Learning Approaches .Malaysian Journal of Computer Science, 132-148.*

[55] Respository, U. M. (2017, February 26).*Retrieved from http://archive.ics.uci.edu/ml/about.html*

[56] Rohini, M., Keerthika, A., Pavithra, N., Sandhiya, S., & Vedha, S. S. (2021). *Heart Disease Identification using Machine Learning. International Journal of Scientific Research in Science, Engineering and Technology, 166-170.*

[57] Saboor, A., Usman, M., Ali, S., Samad, A., Abrar, M., & Ullah, N. (2022).*A Method for Improving Prediction of Human Heart Disease Using Machine Learning Algorithms. Mobile Information Systems.*

[58] Scikit-Learn Documentation. (2022).*Tutorialspoint. Retrieved from https://www.tutorialspoint.com/scikitlearn/scikitlearnintroduction.html*

[59] Serkalem, N., Kula, K., Beakal, G., & Azene, D. (2022). *Rheumatic Heart Disease Detection Using Machine Learning Techniques.*

[60] Shafiul, A., Abu, R., & Humayan, K. R. (2020). *An Experimental Study of Various Machine Learning Approaches in Heart Disease Prediction. International Journal of Computer Applications, 0975-8887.*

[61] Sujay, D., Ritesh, S., Mahendra, K. G., Siddharth, S. R., & Manjusha, P. (2020). *Heart Disease Detection using Core Machine Learning and Deep Learning Techniques: A Comparative Study. International Journal on Emerging Technologies, 531-538.*

[62] Suneeta, R., & Vinayak, S. (2021, FEB).*Comparative Analysis of Feature Selection Based Machine Learning Methods for Heart Disease Prediction. International Journal of Information Technology and Electrical Engineering, 10(1), 41-48.*

[63] Suraj, G., Aditya, S., Upadhyay, S., & Pawan, C. (2021). *A Machine Learning Approach for Heart Attack Prediction. International Journal of Engineering and Advanced Technology (IJEAT).*

[64] TensorFlow Documentation. (2009-2022). *simplilearn. Retrieved from https://www.simplilearn.com/tutorials/deep-learning-tutorial/what-is-tensorflow*

[65] Ufumaka, I. (2020).*Machine Learning Approach for Heart Disease Prediction*

[66] Ufumaka, I. (2021). *Comparative Analysis of Machine Learning Algorithms for Heart Disease Prediction. International Journal of Scientific and Research Publications, 2250-3153.*

[67] Waehner, P. (2022, September 29). *Understanding Your Maximum Heart Rate .*

[68] Wikipedia. (2022, October 15). *Wikipedia. Retrieved from https://en.wikipedia.org/wiki/Artificialneuralnetwork*

[69] World Health Organisation (WHO), W. H. (2021, June 11). *Cardiovascular Diseases. Retrieved from Who: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)*

[70] Yash, R., Shruti, G., & Shubham, M. (2020). *Research on Machine Learning Algorithm on Heart Disease Prediction. International Journal of Engineering Applied Sciences and Technology, 489-493.*