NITTE MEENAKSHI INSTITUTE OF TECHNOLOGY

(AN AUTONOMOUS INSTITUTION, AFFILIATED TO VISVESVARAYA TECHNOLOGICAL UNIVERSITY,

BELGAUM, APPROVED BY AICTE & GOVT.OF KARNATAKA

PROJECT REPORT

on

HEART DISEASE PERCENTILE PREDICTION

Submitted in partial fulfilment of the requirement for the award of Degree of

Bachelor of Engineering

in

Computer Science and Engineering

Submitted by:

ANUBHAV YADAV	1NT18CS017
KINSHUK CHATURVEDI	1NT18CS076
NEHA V M	1NT18CS106
MUSKAN KHATWANI	1NT18IS202

Under the Guidance of

DR VASANTHAKUMAR G. U
ASSOCIATE PROFESSOR, Dept. of CS&E, NMIT

Department of Computer Science and Engineering
(Accredited by NBA Tier-1)

2021-2022

NITTE MEENAKSHI INSTITUTE OF TECHNOLOGY

(AN AUTONOMOUS INSTITUTION, AFFILIATED TO VISVESVARAYA TECHNOLOGICAL UNIVERSITY, BELGAUM
, APPROVED BY AICTE & GOVT.OF KARNATAKA)

Department of Computer Science and Engineering
(Accredited by NBA Tier-1)

CERTIFICATE

This is to certify that the “**HEART DISEASE PERCENTILE PREDICTION**” is an authentic work carried out by **ANUBHAV YADAV (1NT18CS017)**, **KINSHUK CHATURVEDI (1NT18CS076)**, **NEHA V M (1NT18CS106)** and **MUSKAN KHATWANI (1NT18IS202)** Bonafide students of **Nitte Meenakshi Institute of Technology**, Bangalore in partial fulfilment for the award of the degree of ***Bachelor of Engineering*** in COMPUTER SCIENCE AND ENGINEERING of Visvesvaraya Technological University, Belagavi during the academic year ***2021-22.*** It is certified that all corrections and suggestions indicated during the internal assessment have been incorporated in the report. This project has been approved as it satisfies the academic requirement in respect of project work presented for the said degree.

Internal Guide	**Signature of the HOD**	**Signature of Principal**
____________________	____________________	____________________
Dr. Vasanthakumar G. U	Dr. Sarojadevi H	Dr. H. C. Nagaraj
Associate Professor, Dept. CSE, NMIT, Bangalore	Professor, Head, Dept. CSE, NMIT, Bangalore	Principal, NMIT, Bangalore

Signature of Internal Examiner	**Signature of External Examiner**
______________________________	__________________________

DECLARATION

We hereby declare that

(i) The project work is our original work

(ii) This Project work has not been submitted for the award of any degree or examination at any other university/College/Institute.

(iii) This Project Work does not contain other persons' data, pictures, graphs or other information, unless specifically acknowledged as being sourced from other persons.

(iv) This Project Work does not contain other persons' writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then:
a) their words have been re-written but the general information attributed to them has been referenced;
b) where their exact words have been used, their writing has been placed inside quotation marks, and referenced.

(v) This Project Work does not contain text, graphics or tables copied and pasted from the Internet, unless specifically acknowledged, and the source being detailed in the thesis and in the References sections.

NAME	USN	SIGNATURE
ANUBHAV YADAV	1NT18CS017	
KINSHUK CHATURVEDI	1NT18CS076	
NEHA V M	1NT18CS106	
MUSKAN KHATWANI	1NT18IS202	

Date: 11th of July 2022

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompany the successful completion of any task would be incomplete without the mention of the people who made it possible, whose constant guidance and encouragement crowned our effort with success. I express my sincere gratitude to our Principal **Dr. H. C. Nagaraj**, Nitte Meenakshi Institute of Technology for providing facilities.

We wish to thank our HoD**, Dr. Sarojadevi.H.** for the excellent environment created to further educational growth in our college. We also thank her for the invaluable guidance provided which has helped in the creation of a better project.

I hereby like to thank our **Dr. Vasanthakumar G. U,** Associate Professor, Department of Computer Science & Engineering on his periodic inspection, time to time evaluation of the project and help to bring the project to the present form.

Thanks to our Departmental Project coordinators. We also thank all our friends, teaching and non-teaching staff at NMIT, Bangalore, for all the direct and indirect help provided in the completion of the project.

NAME	USN	SIGNATURE
ANUBHAV YADAV	1NT18CS017	
KINSHUK CHATURVEDI	1NT18CS076	
NEHA V M	1NT18CS106	
MUSKAN KHATWANI	1NT18IS202	

Date: 11th of July 2022

ABSTRACT

Heart disease cases are increasing at an alarming rate, and it's critical and concerning to be able to predict such diseases in advance. This is a difficult diagnosis to make, so it must be done correctly and quickly. The main focus of the research paper is on which patients are more likely to develop heart disease based on various medical characteristics. The project created is a heart disease prediction system that uses the patient's medical history to predict whether or not the patient will be diagnosed with heart disease. To predict and classify patients with heart disease various machine learning algorithms such as Logistic Regression, Random Forest Classifier, Support Vector Classifier, Decision Tree Classifier, and K-Nearest Neighbors Classifier will be employed. To regulate how the model can be used to improve the accuracy of prediction of Heart Attacks in any individual, a very helpful approach was used. The proposed model's strength was quite satisfying, as it was able to predict evidence of heart disease with a minimum score of 86.88 % for Random Forest Classifier, 91.8 % for K-Nearest Neighbors Classifier, 90.16 % for Support Vector Classifier and 88.52 % for Logistic Regression. So, by using the given model to determine the probability of the classifier correctly and accurately identifying heart disease, a significant amount of pressure has been relieved. Given a heart disease prediction system improves medical care while lowering costs.

KEYWORDS: SVM; Random Forest; Logistic Regression; python programming; confusion matrix; correlation matrix.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

1.1 Background

1.2 Brief history of Technology/concept

1.3 Research motivation and Problem statement

1.3.1 Research Motivation

1.3.2 Statement of the Problem

1.4 Research objectives and contributions

1.4.1 Primary objectives

1.4.2 Main contributions

1.5 Organization of the report

1.6 Summary

CHAPTER 2: LITERATURE SURVEY

2.1 Introduction

2.2 Related work

2.3 Study of Tools/Technology

2.4 Summary

CHAPTER 3: SYSTEM REQUIREMENTS SPECIFICATIONS

3.1 General Description

3.1.1 Product Perspective

3.2 System Requirements

3.2.1 Hardware Requirements

3.2.2 Software Requirements

3.2.2.1 Functional Requirements & Non-functional Requirements

3.2.2.2 User Requirements

3.3 Summary

CHAPTER 4: DESIGN

4.1 Architectural Diagram

4.2 Use case Diagram

4.3 Sequence Diagram

4.4 Activity Diagram

CHAPTER 5: IMPLEMENTATION

5.1 Methodology

5.2 Description of Process

CHAPTER 6: TEST CASES

CHAPTER 7: RESULTS

CHAPTER 8: IMPACT OF YOUR PROJECT TOWARDS SOCIETY/ ENVIRONMENT

CHAPTER 9: CONCLUSIONS

CHAPTER 10: SELF ASSESSMENT OF PO-PSO ATTAINMENT

REFERENCES

APPENDIX 1 PLAGIARISM REPORT

LIST OF FIGURES

SL. NO.	FIGURE DESCRIPTION	PAGE NO.
1	4.1 Architectural diagram	20
2	4.2 Use case diagram	23
3	4.3 Sequence Diagram	24
4	4.4 Activity diagram	25
5	5.1 Correlation Matrix	29
6	5.2 Correlation with ‘Target’ Column	30
7	5.3 Importance of Data Pre-processing	30
8	5.4 Null value count for each attribute	31
8	5.4 Duplicate value counts	31
9	5.5 Quartile Range	32
10	5.6 Boxplots showing outliers	33
11	5.7 IQR Score for all attributes	33
12	5.8 Z Score values	34
13	5.9 Conversion of categorical values to numeric	35

14	5.10 Target class count	37
15	5.11 Support Vector Machine Test Result	42
16	5.12 Random Forest Classifier Test Result	43
17	5.13 Logistic Regression Classifier Test Result	45
18	5.14 K-Nearest Neighbor Classifier Test Result	47
19	6.1 Web API Start Page	48
20	6.2 Error message for incorrect inputs	49
21	6.3 Presence of Heart Disease	50
22	6.4 Presence of Heart Disease	51
23	6.5 No Presence of Heart Disease	52
24	7.1 Test Accuracy for different algorithms achieved	53
25	7.2 F1 score for different algorithms achieved	54

LIST OF TABLES

SL. NO.	TABLE DESCRIPTION	PAGE NO.
1	5.1 Attributes of the dataset	28
2	5.2 Confusion Matrix	39
3	7.1 Comparison of accuracies with base papers	53

LIST OF ACRONYMS

SL. NO.	ACRONYM	DESCRIPTION
1	Cp	Chest Pain
2	Trestbps	The person’s resting blood pressure
3	Chol	Cholesterol
4	Fbs	Fasting Blood Sugar
5	Thali	Maximum heart rate achieved
6	Exang	Exercise Induced Angina
7	Ca	Number of major vessels coloured by fluoroscopy
8	Thal	Blood disorder - Thalassemia
9	CVD	Cardiovascular diseases
10	SVM	Support Vector Machine
11	IHDPS	Intelligent Heart Disease Prediction System
12	MITDB arrhythmia	condition in which heart beats with an irregular or abnormal rhythm

CHAPTER 1: INTRODUCTION

Artificial intelligence refers to a machine's capacity to execute tasks that are traditionally handled by humans. It allows machines, in particular computer systems, to mimic the functions of human intelligence. Examples of particular AI programmes include natural language processing, expert systems, machine vision and speech recognition. A branch of artificial intelligence called machine learning makes it possible for programs to improve their likelihood to anticipate events without having to directly build them. Algorithms that use machine learning make predictions about future output values based on historical data.

1.1 BACKGROUND

Modern life's hectic pace results in an unhealthy lifestyle that breeds anxiety and sadness. To deal with an inclination of immerse usage in excessive drinking, smoking and drug usage. All of these factors are the underlying causes of a number of deadly diseases, such as cancer and cardiovascular conditions. A range of disorders that could have an impact on your heart are referred to as cardiovascular diseases, a broad word [1]. 17.9 million deaths globally are attributed to CVDs, the World Health Organization claims. It is the main reason for death for individuals all around the world. Numerous studies have been conducted in an effort to properly identify all risks associated with heart disease and to establish particular risk factors. Since cardiovascular disease frequently results in death without showing any symptoms of the condition, it is known as a "silent killer." It is crucial to anticipate these diseases early so that preventative actions can be undertaken before something terrible occurs.

The term "cardiovascular disease" (CVD) is used to refer to any illness that impacts the system of the heart. Coronary Artery Disease, Cardiovascular Ischemic Attack (Mini-stroke), Cerebral Vascular Disease, and Aortic Disease are the four primary categories of CVDs. Although the actual source of CVDs is yet unclear, various risk factors, such as age, family medical history, increased blood pressure, cigarettes, alcoholism, body mass index (BMI), and cholesterol, etc. are to blame for these illnesses. These variables vary depending on the individual [2]. One of the main causes of CVDs is an unhealthy habit, which includes characteristics including anxiety, age, sexuality. The main goal is to accurately and promptly forecast these disorders so that the rate of death can be decreased by efficient treatment and other preventative measures. By accumulating information from many sources, organizing it under pertinent categories, and then evaluating it to obtain the information we need, the process of creating a projection of heart disease may be considerably improved.

1.2 BRIEF HISTORY OF TECHNOLOGY/CONCEPT

1.2.1 DATA VISUALIZATION

There is a lot of data being produced every day in the modern world. And occasionally, if the data is in its raw format, it may be challenging to examine it for specific trends or patterns. Data visualisation is used to overcome this. Data visualisation makes it simpler to comprehend, observe, and analyse the data by providing a good, organised pictorial depiction of it.

Python offers a variety of libraries with diverse functionalities for displaying data. Each of these libraries has unique features and supports a range of graph types.

Matplotlib and Seaborn are implemented in the project.

1.2.2 PREDICTIVE MODELING

Using data modelling, predictive modelling is a technique for forecasting future results. It's among the best ways for a company to see its future direction and make strategies accordingly. This technique is frequently employed since it typically has good accuracy rates.

It functions by examining recent and historical data and applying what it discovers to a model created to forecast probable events [3]. Almost everything can be predicted with predictive modelling, including TV ratings, a customer's future purchase, credit risk, and business earnings.

Predictive models are not static; they are continually tested or updated to take into account changes in the underlying data. The majority of predictive models operate quickly and frequently finish their computations in real time. In anticipating business results, it takes less time, effort, and money.

1.2.3 CLASSIFIERS AND CLASSIFICATION MODEL

An algorithm known as a classifier in machine learning automatically arranges or categorises data into any number of a group of "classes." Algorithms for machine learning are useful for automating processes that required manual labour in the past. They can reduce costs and time wasted while increasing the effectiveness of businesses.[4].

A classification model is the product of the classifier's machine learning. The classifier is being used to train the model, and the model in turn employs the classifier to categorise the data. Both types of classifiers are supervised and unsupervised. Only unlabeled datasets are supplied to unsupervised machine learning classifiers, and they categorise the data based on patterns, structures, and anomalies found in the data.

In order to categorise material as positive, negative or neutral, classifiers are taught to look for opinion polarity in the text. Training data are facilitated to semi-supervised and supervised classifiers so they can learn how to categorise data into specific groups.

In the project the following 4 major classifiers are considered for analysing the data collected:

1. Random Forest Classifier
2. Support Vector Classifier
3. K-Nearest Neighbors Classifier
4. Logistic Regression.

1.2.4 MACHINE LEARNING IS USED IN MEDICAL FIELD

A key sector of the economy, healthcare provides care to millions of people while also boosting the local economy. The healthcare sector is gaining a lot from artificial intelligence. By lending a hand, information technology is transforming the healthcare sector.[5].

The creation of computing systems with AI capabilities allows them to carry out activities that would ordinarily need human intelligence. These entail demanding duties like making decisions, handling challenging situations, detecting objects, and many other things. The healthcare industry is implementing the advantages of technologies, such as higher degree of precision and high-level computing, which take humans days to solve manually, to improve the services and maintain the data structured.[6].

Artificial intelligence (AI) is being used in a variety of fields, including marketing, banking, the video game industry, and even the performing arts. The healthcare sector is where artificial intelligence is having the biggest influence. By 2030, A PwC analysis estimates that AI will boost the world economy by $15.7 trillion, with the healthcare sector expected to be the most affected.

Applications in the use of machine learning medical field:

- Manages the Medical Data
- Helps in the Medical Diagnosis
- Detects Diseases at an Earlier Stage
- Provides Medical Assistance
- Helps in Decision Making
- Personalization of the Medicine
- Helps Analyse the Errors in Prescriptions

1.3 RESEARCH MOTIVATION AND PROBLEM STATEMENT

1.3.1 RESEARCH MOTIVATION

People all over the world are affected by various cardiovascular diseases. Although there are several machine learning tools available and viable methods for the prediction of cardiovascular disorders, still there is a lack of proper models that could predict the disease more accurately with the higher accuracy and fast response time. As of now, there isn't a reliable automated system that can help in prediction and assistance in diagnosing heart disease. So, the application of the utilising machine learning methods to day-to-day effect of the disease shall be of great achievement. It may boost the living standards of cardiac patients and it will help the patients to prevent the effect up to a great extent.

In India, 78 percent of households have a single earning family member who mostly are men. Men have 70-89% more chances of getting a heart attack. In a middle-class family losing the earning member can lead to a very tough life ahead and economic pressure on the entire family. India is a country with the young people as its backbone [7].

Working on the causes of cardiac problems and developing a model to forecast cardiac disease are the key drivers behind this study's goal of predicting the existence of Myocardial infarction signs.

Additionally, the goal of this work is to determine the classification algorithm that will forecast the aforementioned sickness with high degree of accuracy. This work will be excused by analysing the random forest technique, the K-Nearest Neighbors Classifier, logistic regression, and support vector machines for predicting heart disease and the most accurate algorithm will be considered the best.

1.3.2 PROBLEM STATEMENT

Given a set of data pertaining to patients' clinical parameters, the problem is to clean the dataset and choose an appropriate algorithm which gives high accuracy in predicting the likelihood of diagnosing the patient with cardiovascular disease.

1.4 RESEARCH OBJECTIVES AND CONTRIBUTIONS

1.4.1 PRIMARY OBJECTIVES

The primary objects of the work are:

- to design and build a user-friendly interface that anyone irrespective of their age or gender can use to test the heart health status.
- to pre-process the dataset so to help boost general productivity by allowing for the accurate service possible when making decisions.
- to select a suitable machine learning algorithm for the project with high accuracy of results.
- to specifically mention the percentile chances of heart disease.

1.4.2 MAIN CONTRIBUTIONS

This system utilizes machine learning models for detecting the symptoms of cardiovascular diseases from the given user inputs like age, gender, chest pain, cholesterol, etc. Then classifies it according to a set of predetermined results. The dataset has been cleaned with machine learning techniques in such a way which helps prediction of the results with great accuracy.

i. **Features offered by the assistant:**

- Evaluating a patient's cardiac state depending on the information provided.
- Along with the result the model provides the percentile confirmation on how likely the model is sure of the output.

ii. **Future scope:**

- With the help of more dataset training on lab result variables an option to analyse a patient's lab results and determine the type of the heart condition can be created.
- Based on the heart condition this can be further extended to provide general dos and don’ts instruction for taking care of himself until a diagnosis test is not conducted by a professional.
- With the help of GPS location of the patient, this platform can suggest the Hospitals/Clinics in the vicinity where the treatment can be done for the heart condition.

1.5 ORGANIZATION OF THE REPORT

CHAPTER 1: INTRODUCTION

This chapter gives a detailed introduction of the background, history, and tools used in the project.

CHAPTER 2: LITERATURE SURVEY

This chapter gives information about the related works and study of tools/technology.

CHAPTER 3: SYSTEM REQUIREMENTS SPECIFICATION

This chapter tells us about the software and hardware requirements required for the project.

CHAPTER 4: DESIGN

This chapter gives a detailed idea of the workflow and dataflow of the project.

CHAPTER 5: IMPLEMENTATION

This chapter gives us a detailed structure of the implementation of the model using the appropriate algorithms, libraries and tools.

CHAPTER 6: TESTCASES

This chapter gives us a series of steps on the execution of the model using a set of predefined input data.

CHAPTER 7: RESULTS

This chapter gives us the desired outputs with its applications in the present world.

CHAPTER 8: IMPACT OF YOUR PROJECT TOWARDS SOCIETY/ ENVIRONMENT

This chapter gives us a detailed explanation on the solutions the project offers to solve real-world problems.

CHAPTER 9: CONCLUSIONS

This chapter summarizes the report as a whole, drawing inferences from the entire process about what has been solved, or decided, and impact of those.

CHAPTER 10: REFERENCES

This section gives the details about the sources used in making the project.

CHAPTER 11: SELF ASSESSMENT OF PO-PSO ATTAINMENT

This chapter gives a self-analysis of the project prior to the external peer review.

1.6 SUMMARY

In this chapter the Background of the project with the Brief History of the Technology/ Concepts applied has been discussed. It is aimed to give a detailed discussion on the Research Motivation and the Problem Statement behind the project. The primary objectives of the model and the main contributions have been thoroughly discussed in this section.

CHAPTER 2: LITERATURE SURVEY

2.1 INTRODUCTION

Intensive research made it possible to integrate health with machine learning. There has been considerable research in creating new methods and techniques to diagnose heart health issues using AI algorithms. Utilizing the existing resources while facilitating users to effectively analyse and treat the patient [8].

The goal is to develop a simple system that, by analysing the clinical data of the patients, can predict who is most likely to be diagnosed with cardiovascular illnesses. Assist in the diagnosis of the condition with fewer medical tests and more efficient medications so that those who display heart disease symptoms, it is possible to effectively treat them.

Integrating all the models to work together as the backend for easy interaction with the users, making it possible to achieve a greater user experience. This system utilizes machine learning models for detecting the symptoms of cardiovascular diseases, then classifies it according to a set of predetermined features (Example: Coronary Heart Disease, Stroke, Aortic disease, etc). It then uses an assistance system to provide recommendations on the medical attention the user requires.

2.2 RELATED WORK

This research was inspired by extensive work on applying machine learning techniques to diagnose cardiovascular diseases. This paper includes a brief review of the literature. An accurate prediction of heart disease has been made using a number of methods, such as Random Forest Classifier, KNN, and Logistic Regression. According to the Outcomes, each technique is capable of registering the specified objectives.[9].

The model including Intelligent Heart Disease Prediction System was able to pinpoint the decision boundary by utilizing both the old and new machine learning and deep learning models. It made it simpler to obtain the most fundamental and significant aspects and information related to any heart issue, including a family history. Furthermore, compared to a new, emerging model for identifying cardiovascular disease employing artificial neural networks and other machine learning techniques, the effectiveness of such an IHDPS model was significantly lower. Using an internal implementation approach that makes use of certain neural network techniques, McPherson et al. was able to effectively determine if the test individual had coronary heart disease by identifying the risk variables for the diseases [10].

Neural networks were first used to detect and predict heart disease, increased blood pressure, and other conditions by R. Subramanian et al. [11]. Using the specified health parameters, a Deep Neural Network was created to provide the output that was carried out by the output classifier and approximately contained 120 hidden layers in order to provide an accurate diagnosis of cardiovascular illness if the model is used for the Testing Dataset. For the purpose of diagnosing cardiac disease, supervised networks have been suggested [12]. The model was employed and trained using the previously discovered data and projected the outcome in order to determine the correctness of the provided model when testing was carried out by a physician utilizing unknown data.

Research Paper -1

Title: Heart disease prediction using machine learning algorithms [13]

Year Of Publication: 2020

Author: Harshit Jindal, Sarthak Agrawal, Rishabh Khera, Rachna Jain and Preeti Nagrath

Contributions:

1. Using a dataset of patients' medical histories, including information on chest pain, glucose levels, increased blood pressure, and other symptoms, this method estimates the likelihood that a patient will have cardiovascular disease.

2. The patient benefits from this heart disease detection method since it is based on clinical data concerning the patient's previous heart disease diagnosis. KNN, Random Forest Classifier, and Logistic Regression were used in the creation of the presented model.

3. By cleaning the dataset and utilizing KNN and logistic regression to obtain an average accuracy of 87.5 percent, this study assists in the prediction of individuals who are likely to be diagnosed with heart disease.

4. Additionally, it is determined that KNN's accuracy of 88.52 percent is the greatest of the three algorithms they utilised.

Research Paper -2

Title: Prediction of Cardiovascular Disease Using Machine Learning Algorithms [14]

Year Of Publication: 2018

Author: Dinesh Kumar G, Arumugaraj K, Santhosh Kumar D, Mareeswari V

Contributions:

1. Pre-data analysis is used in the proposed study to remove noisy data, delete data that isn't there, fill in default values when they're available, classify predictive qualities, and make decisions at various levels.

2. Methods including categorization, precision, sensitivity, and relevance studies were used to assess the diagnostic model's efficacy. The prediction model presented in this paper can be used to detect cardiac issues and notify or diagnose the patient.

3. By comparing the effectiveness of the application of rules to the results of each Vector support system, Gradient Boosting, Random Forest, Naive Bayes classifier, and data retrieval from databases taken from the area, this was done in order to display an accurate model of cardiovascular prediction.

4. The accuracy for these models as reported by the authors was as follows: SVM 79.77%, Random Forest 80.89%, and Logistic Regression 86.51%.

Research Paper -3

Title: Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques [15]

Year Of Publication: 2019

Author: Senthil Kumar Mohan, Chandrasegar Thirumalai, And Gautam Srivastava

Contributions:

1. Machine learning algorithms were used in this work to analyse the raw data and provide a unique and original insight into cardiac disease.

2. The suggested Hybrid Random Forest Linear Model approach was created by combining the characteristics of Random Forest and Linear Method. It proved to be highly accurate in predicting heart disease.

3. This research can be conducted in the future using a variety of machine learning methodology combinations to improve prediction methods.

4. New feature selection methods can also be developed in order to gain a better grasp of the crucial elements and boost the precision of heart disease prediction.

Research Paper -4

Title: Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms with Relief and LASSO Feature Selection Techniques. [16]

Year Of Publication: 2021

Author: Pranab Ghosh; Sami Azam; Mir jam Jonkman; Asif Karim; F. M. Javad Mehedi Samrat; Eva Ignatius

Contributions:

1. This study used machine learning to attempt to predict cardiac illness, but it did it in a fresh and improved way, using a larger dataset to train the model. This study demonstrated how the Relief feature selection approach may be used with a variety of machine learning methods to create a strongly linked feature collection.

2. The study also found that Relief feature selection produces accuracy that is significantly greater than comparable work and performs particularly well with high impact features. With 10 characteristics, Relief feature selection was 99.05 percent accurate.

3. The model will eventually be further generalised in order to be robust against datasets with large levels of missing data and to interact with additional feature selection approaches. Another potential strategy is to use Deep Learning algorithms.

Research Paper -5

Title: Predictive Monitoring of Critical Cardiorespiratory Alarms in Neonates Under Intensive Care [17]

Year Of Publication: 2019

Author: Rohan Joshi; Zheng Peng; Xi Long; Loe Feijs; Peter Andriessen

Contributions:

1. In this study, At sub - critical physiological variable thresholds, critical cardiorespiratory alarms can be predicted. The ratio of accurate to inaccurate critical alarm predictions was subject to trade-offs.

2. The both the case and control cohorts divided into training (80%) and test sets, divided by percentage (20 percent). In order to forecast if a red alert or alarms would soon follow a yellow alarm, decision tree-based classifiers were used.

3. Using data from the 2-min window before the occurrence of the yellow alarm, the best classifier was able to predict 26% of the red alerts ahead of time (18.4s, median), but at the cost of a 7% increase in red alarms.

4. It can be implemented safely since alerts that aren't algorithmically anticipated can nevertheless be raised when the threshold is typically crossed, as is the case in everyday clinical practice. rms.

Research Paper -6

Title: Smart Heart Monitoring: Early Prediction of Heart Problems Through Predictive Analysis of ECG Signals [18]

Year Of Publication: 2019

Author: Jiaming Chen; Ali Valehi; Abolfazl Razi

Contribution:

1. In this study, they reviewed the idea that some heart defects progress over time, and as a result, some covert indicators may be present in the patient's ECG signal morphology. This idea had previously been acknowledged by various researchers, and the analysis of the MITDB arrhythmia dataset had confirmed it.

2. ECG signal processing uses a two-step prediction framework in which a global classifier compares the signal to a global reference model to identify significant abnormalities.

3. Following a deviation analysis of the initially seemingly normal signal samples, yellow alarms are triggered by the detection of subtle yet instructive signal morphological distortions in comparison to the baseline for each patient that may be suggestive of impending cardiac issues.

4. It is suggested to increase the symmetry of signals for various abnormality classes using the optimized parameters to help with an accurate deviation analysis.

5. The suggested method achieves a classification accuracy of 96.6 % and offers a unique feature of predictive analysis by cautionary warning messages highlighting the heightened risk of cardiac abnormalities to take the appropriate measures in accordance with doctor's guidance.

CHAPTER 3: SYSTEM REQUIREMENTS SPECIFICATIONS

3.1 GENERAL DESCRIPTION

The current applications needed today must be quick, reliable, and simple to use with a strong user interface and user experience. The best and most recent software technologies are employed to achieve this.

One of the key components of the SDLC is requirement analysis. The best algorithms, data models, and scalability are used in the application, which is built using the best framework and adhering to design and development standards, cross-platform OS, and scalability. Therefore, in order to achieve desired goals, it is essential to have a clear understanding of the resources required in advance.

3.1.1 PRODUCT PERSPECTIVE

This is an application written in multiple technologies, coded in Python, Pandas, Sklearn, Seaborn, Matplotlib. This can be coded in any editor like Visual Studio Code, Atom etc. For the frontend, Sklearn is used for handling registrations and displaying the demographics and handling all sorts of other data. Python is also used for implementing machine learning, for training and deploying models.

The processor required should have a minimum speed of 1.65Ghz and can be higher or equivalent to AMD Ryzen 3 or Intel i3. The minimum RAM required is 2 GB or more according to the size of the dataset and for getting higher performance. Proper Internet Connection required.

The application has features according to which requirements would be made:

- Cross-platform: It can run on any laptop with any operating system.
- The datasets for processing and computing data and getting the resultant sets.

3.2 SYSTEM REQUIREMENTS

3.2.1 HARDWARE REQUIREMENTS

- RAM: 4GB or more.
- Processor: Intel Core i3 or i5 editions.

- CPU: 1.6GHz n above.
- Storage: 5GB
- Internet Connection

3.2.2 SOFTWARE REQUIREMENTS

3.2.2.1 PROGRAMMING LANGUAGE: PYTHON

Python is a powerful, all-purpose, and widely used programming language. Applications for machine learning and web development both use the Python programming language front-line technology in the Software Industry.

The new oil is data. This statement demonstrates how data collection, storage, and analysis are the driving forces behind all contemporary IT systems. Whether it's about making business decisions, predicting the weather, researching the architectures of proteins in biology, or creating a marketing strategy. In each of these cases, the data analysis is based on a multidisciplinary approach that includes mathematical models, statistics, graphs, databases, and, of course, the commercial or scientific reasoning. Therefore, we require a programming language that can accommodate all of the many needs of data science. Python stands out as one such language since it has many built-in capabilities and packages that make it simple to address the needs of data science.

3.2.2.2 MACHINE LEARNING: SKLEARN

Python's Scikit-learn package for electronic learning is quite helpful. The library is built on SciPy (Scientific Python). includes a variety of helpful mathematical modelling and machine learning techniques, such as size reduction, regression, integration, and division. SciPy Toolkits, often known as SciKits, are frequently used for machine learning. A scikit is a specialized toolkit used for particular tasks, such image processing or machine learning. These tasks are carried out using the specialist software packages Scikit-learn and Scikit-image.

The software industry, particularly programmers, love SciKits. One of the foundations of Python-based machine learning is even Scikit-learn. This can be used to prepare and analyse data, develop multiple models, and even produce post-model analysis.

3.2.2.3 DATA PREPROCESSING: NUMPY, PANDAS

i. **NUMPY-** The Python package NumPy is used to manipulate arrays. Additionally, it provides functions for working with matrices, the Fourier transform, and the area of linear algebra. The n-dimensional array is a straightforward but effective data structure offered by the Python package NumPy. Learning NumPy is the first step on every Python data scientist's path because it serves as the cornerstone on which nearly all of the toolkit's capabilities are constructed. Contrary to lists, NumPy arrays are stored in a single continuous location in memory, making it incredibly efficient for processes to access and modify them.

ii. **PANDAS-** The most often used open-source Python library is called Pandas. It is constructed on top of NumPy, a different package that supports multi-dimensional arrays. One of the most popular data wrangling tools is Pandas and it is frequently included in all Python distributions. It integrates nicely with many other data science modules. Pandas allows us to analyze big data and make conclusions based on statistical theories. They provide a range of data manipulation operations, including merging, reshaping, and selecting, in addition to data cleaning and data wrangling functionalities.

3.2.2.4 DATA VISUALIZATION: SEABORN, MATPLOTLIB

i. **SEABORN -** In order to understand the patterns in the data and examine the most effective machine learning or deep learning techniques that a data science enthusiast can employ to achieve high-quality outcomes, visualizations are crucial. These are among the most crucial actions that must be taken for exploratory data analysis (EDA) to get desirable results. A matplotlib-based Python data visualization library is called Seaborn. It offers a sophisticated drawing tool for creating eye-catching and educational statistical visuals.

ii. **MATPLOTLIB -** A Python 2D plotting toolkit called Matplotlib creates publication-quality graphics in a range of physical formats and in cross-platform interactive settings. Four graphical user interface toolkits, four Python scripting languages, the Python and IPython shells, the Jupyter notebook, and web application servers can all use Matplotlib. Matplotlib consists of several plots like line, bar, scatter, histogram etc. Matplotlib is mostly written in python, a few segments are written in C, Objective-C and JavaScript for Platform compatibility. It is open source and can be used freely.

3.2.2.5 CLOUD COMPUTING (KAGGLE, GOOGLE COLLAB)

Through the use of the internet, or "cloud," which is what it is termed in this context, cloud computing enables businesses to access various computing services including databases, servers, software, artificial intelligence, data analytics, etc. These businesses can afford to run their applications in the top data centres in the globe. This guarantees that ambitious and complicated initiatives that would otherwise be highly expensive can be used by small businesses or those in emerging economies. Additionally, this is true for the field of data science. For data scientists, cloud computing has greatly simplified data analytics and data management.

i. An online community of data scientists and machine learning professionals may be found at **KAGGLE**, a division of Google LLC. With Kaggle, people may find and share data sets, analyse and create models in a web-based data science environment, work with other data analysts and machine learning specialists, and take part in contests to solve data science barriers.

ii. A completely cloud-based Jupyter notebook environment called **COLAB** is free. Many well-known machine learning libraries are supported by Colab and are simple to load in the notebook**.** Colab notebooks don't need to be set up, and the ones that are made may be instantly modified by colleagues or friends, allowing them to remark or amend them in the same way that Google Docs pages do. Our personal Colab notebooks that we make are kept in our Google Drive accounts.

3.2.2.1 FUNCTIONAL REQUIREMENTS & NON-FUNCTIONAL REQUIREMENTS

i. FUNCTIONAL REQUIREMENTS:

The application uses the datasets as inputs, computes the normalized resultant sets, and then plots the region's map to display the subregions' demographic priorities. To obtain the proper output and display, a variety of different reinforcement learning models have been applied. Finally, the app offers a text file submission mechanism for patients' medical reports.

ii. NON-FUNCTIONAL REQUIREMENTS:

- **Reliability:** The application should be reliable to plot the correct datasets on the map, failure in doing so may result in delivery of the wrong vaccines.
- **Scalability:** The application is to be scalable to provide world wide availability and the predictions must be correct with any and all regions given the correct datasets.
- **Usability:** The application is built to optimize user interface and user experience, making it easier to use and navigate.
- **Maintainability:** Application maintenance is as crucial as anything to maintain users. There might be some bugs, some unordinary failures that need to be improved. User feedback can be helpful in such situations.
- **Security:** The application has safety of users and the government institutions, so other people cannot get registered in someone else's name, and vaccines cannot be dispatched unless they are from a trusted government official.
- **Adaptability:** The application can run on various OS like Windows, Mac and Linux, while also having full functionality.

3.2.2.2 USER REQUIREMENTS

The users of the system are the public masses as well as the hospitals. The User interface is easy to understand and use. Users can easily upload medical reports and it will calculate heart health.

Some requirements are as follows:

- User should have installed the application.
- They should have a proper internet connection.
- Users should have a medical report.

3.3 SUMMARY

The hardware specifications for this project make it simple to create the application for improved coding standards, maintenance, and library access on any system. An adequate working application can be produced by using the right operating system, memory, CPU processing capacity, a competent framework, and APIs. For browsing and delivering relevant results, a browser and strong Internet access are needed. For the user to comprehend the created sets, they must be accurate and trustworthy.

CHAPTER 4: DESIGN

4.1 ARCHITECTURAL DESIGN

The following figure shows the proposed system's architecture. The following are the main elements of the architecture:

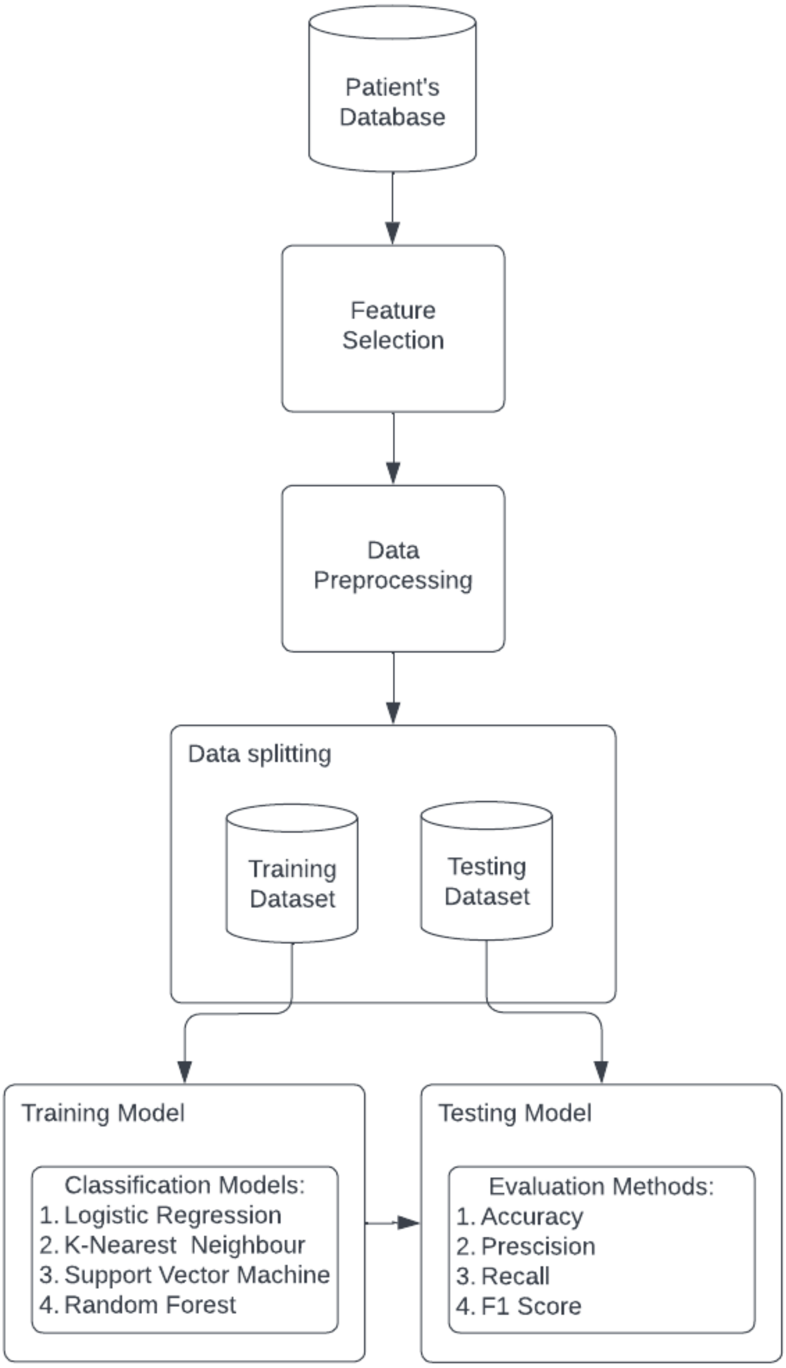


Figure 4.1 Architectural design diagram

4.1.1 Patient’s Database

The Hungarian Institute of Cardiology in Budapest, University Hospitals in Zurich, Basel, and Long Beach, California, as well as the Cleveland Clinic Foundation's Heart Disease Databases, all contributed to the dataset, which was first imported for this system's study. This dataset contains 76 attributes with 303 entries.

4.1.2 Feature Selection

Although there are 76 attributes in this database, all published trials only use a subset of 14 of them that are linked to heart disease. Age, sex, cp, trestbps, chol, fbs, restecg, thalach, Exang, oldpeak, slope, ca, and thal are the attributes included in the chosen new dataset. The final predicted attribute will be specified in ‘num’. The attributes are further elaborated in Table 5.1.

4.1.3 Data Pre-processing

Pre-processing is a key phase in the knowledge discovery process. Data from the real world is frequently erratic and loud. Data processing techniques like data cleaning etc help in overcoming these drawbacks. Standardization of the dataset helps in classifying the data which further makes the data to smoothly allow algorithms to execute with efficient results. To carry out standardization, standard scalar function is used. This helps in bifurcating the data into classes. Then a variable will be created that is ‘num’ which will hold the predicted attribute.

4.1.4 Data Splitting

It is usually done to prevent overfitting. In this case, a machine learning model fits its training data too well and fails to fit further data reliably. The original data in a machine learning model is classified into three or four groups. The training set, development set, and testing set are the three most commonly used sets:

- The **training set** is part of the data used to train the model. The model should look at and learn from the training set, improving any of its limitations.
- The **dev set** is a data set of examples used to modify learning process parameters. It is also called the opposite verification set or model verification. This data set aims to measure model accuracy and can assist in model selection.

- The **testing set** is the piece of data examined in the final model and compared to the preceding data sets. The testing set serves as an assessment of the final mode and algorithm.

Data should be sorted so that the data sets have the highest amount of training data. For example, data may be divided into 80-20 or 70-30 training sessions against experimental data. The exact rate depends on the data, but the 70-20-10 training, dev and test division is suitable for small data sets. In this model 80-20 split is used.

4.1.5 Training model

In the training part, 4 classifiers will be trained and for analysing the best among them. The training is done on the basis of the dataset input to the system. Each time the model is trained, the number of iterations, etc., the system's efficiency can be enhanced. The whole dataset provided which consists of 13 attributes and 80% of dataset entries will help the model undergo training. The input layer, hidden layer, and output layer make up the three layers of the neural network we design here.

4.1.6 Testing model

Testing will be conducted so as to determine whether the model that is trained is providing the desired output. As the data is entered for testing, the .csv file will be retrieved to crosscheck and then compare and the results of the newly entered data will be generated. Based on the manner in which the model is trained using the dataset, the user will input values of his choice to the attributes specified and the results will be generated as to whether there is a risk of heart disease or not.

4.2 USE CASE DIAGRAM

The following figure shows the proposed project's use case. The major components of the use case diagrams are as follows:

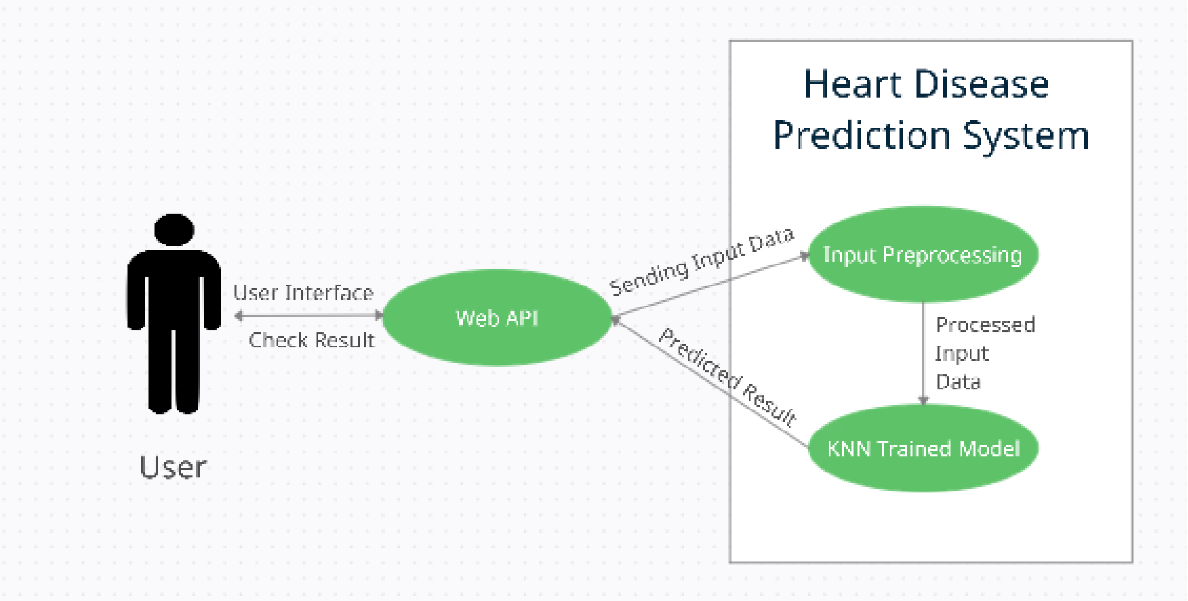


Figure 4.2 Use case diagram

4.2.1 User

The user can use the Web API to access the Heart Disease Prediction System, which offers input options that must be filled out because they are the essential data for the system's accurate outcome prediction. Once the result is predicted by the heart disease prediction it will be published on the Result section on the Web API for users to access it.

4.2.2 Web API

It is a bridge between the User and the Heart Disease Prediction System. It has multiple input fields for taking various inputs from the user at a time. The result for the input filled will be shown in the output filled provided on the end of Web API.

4.2.3 Input Pre-processing

This is an important step where the input values will undergo standardization for acting as input for the KNN classifier trained model to process it for predicting the result.

4.2.4 KNN Trained Model

The trained model which is being used for predictions is trained with the K-Nearest Neighbor classifier.

4.3 SEQUENCE DIAGRAM

The below sequence diagram depicts interaction details as how operations are carried out in this project. It depicts how items interact within the framework of a cooperation. The vertical axis of the graphic is used to indicate time, the messages that are sent when, and the order in which they are sent in this heart disease prediction model. This is time-focused and visually depicts the order of the interaction.

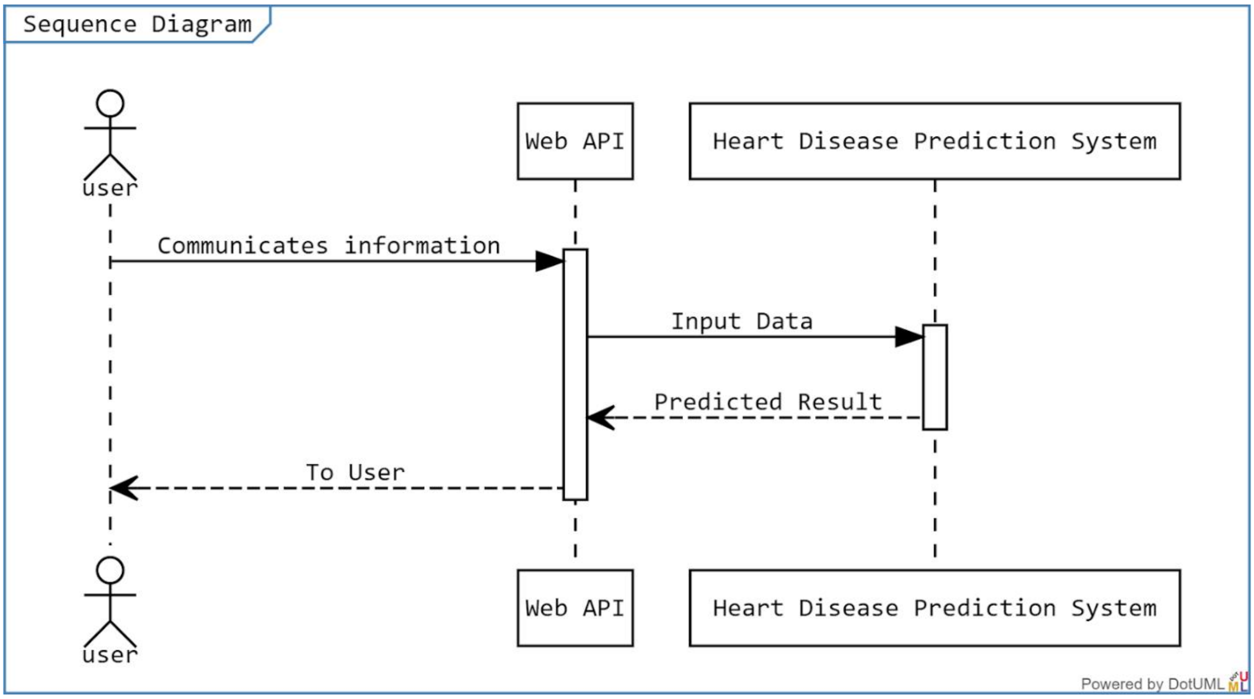


Figure 4.3 Sequence Diagram

4.4 ACTIVITY DIAGRAM

The below diagram depicts the basic flow of the Heart Disease Prediction System. Initially, input data is taken from the user. If the age of the user is less than or equal to 0, then the inputted data is an error, which needs to be rectified. So, if that condition is true then an error message will be projected in the output screen, and the user will be redirected to input the data again.

Once the data is correctly filled then the input data undergoes pre-processing (i.e., creation of dummies and standardization). The processed input is fed to a trained model which is trained by the K-Nearest Neighbor classifier. After all the processes are successfully completed, result is obtained in the form of classifier i.e. Yes or No with the appropriate message and the prediction percentage computed by the model for the respective input.

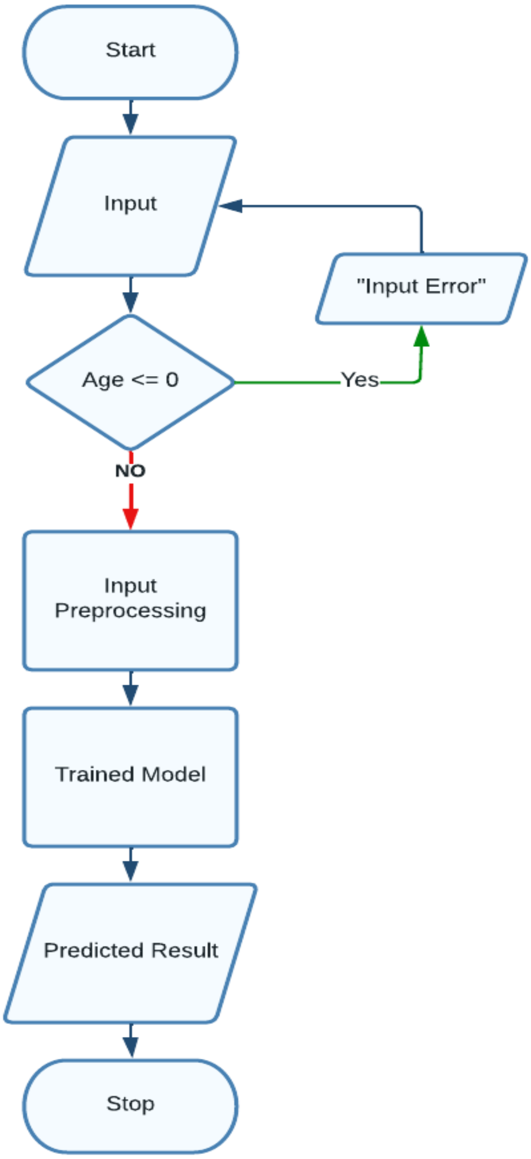


Figure 4.4 Activity diagram

CHAPTER 5: IMPLEMENTATION

5.1 METHODOLOGY

Heart disease has been referred to as a silent killer since it can lead to death even in the absence of overt symptoms. Increased worry about the sickness and its implications is brought on by the nature of the illness. To forecast the occurrence of this fatal disease, further attempts are being done various tools and procedures are constantly tested to meet the demands of contemporary health. In this regard, machine learning approaches can be useful. Given the fact that heart disease can present itself in a variety of ways, there are a number of important health conditions that can predict whether a person would eventually be at risk of having cardiovascular events or not. “Prevention is better than cure” as the famous saying goes, early diagnosis and monitoring can be beneficial in preventing and reducing the risk of death from heart disease.

System performance begins with data collection and selection of key features. The required data is pre-processed into the required format before being subdivided into training and testing data. The models are deployed by using the algorithm and the training data. Utilizing test data, the model is tested to determine its accuracy. The aforementioned blocks are used to execute this programme:

1) Data Collection
2) Attribute Selection
3) Data Pre-Processing
4) Data Balancing
5) Disease Prediction

5.1.1 DATA COLLECTION

An attack of the heart (cardiovascular illnesses) happens when the flow of blood to the heart muscle is suddenly cut off. According to WHO data, 17.9 million people die from heart attacks each year. According to the medical study, this heart condition is primarily caused by human lifestyle. In addition to this, there are a number of important factors that can indicate whether or not someone will experience a heart attack. Every attribute has a numerical value.

Although there are 76 attributes in this database, all published experiments only mention using a portion of 14. The Cleveland database in particular is the only one that ML researchers have used. The "goal" field alludes to the patient's having heart disease. It has an integer value between 0 (absence) and 1. The Cleveland database has been used for experiments that have mostly focused on separating presence from absence.

Creators:

- Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D
- University Hospital, Zurich, Switzerland: William Steinbrunn, M.D
- University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D
- V.A. Medical Centre, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.

Donor:

David W. Aha (aha@ics.uci.edu) (714) 856-8779

SL NO.	ATTRIBUTE	DESCRIPTION	TYPE
1	Age	Age of patient	Numerical
2	Sex	Patient’s gender (M-0, F-1)	Nominal
3	Cp	Type of chest discomfort	Nominal
4	Trestbps	Blood pressure at rest (hospital admission values in mmHg, ranging from 94 to 200)	Numerical
5	Chol	Mg/dl of serum cholesterol (ranges from 126 to 564)	Numerical
6	Fbs	glucose level at fasting > 120 mg/dl (false-0,true-1)	Nominal
7	Resting	electrocardiogram recorded when at rest (0 to 1)	Nominal
8	Thali	reached maximum heart rate (71 to 202)	Numerical
9	Exang	Angina caused by exercise (yes-1, no-0)	Nominal
10	Oldpeak	ST Depression during exercise in relation to the quantity of rest received (0 to .2)	Numerical
11	Slope	The slope of the ST workout segment's peak (0 to 1)	Nominal
12	Ca	Fluoroscopically coloured major vessels (0 to 3)	Numerical
13	Thal	Defect type (0 to 2)	Nominal
14	Target	Heart Disease	Nominal

Table 5.1 Attributes of the dataset

5.1.2 ATTRIBUTE SELECTION

Feature selection or attribute selection refers to the choice of suitable system-relevant qualities. This is done to improve performance of the system. Various patient characteristics such as gender, type of cp, fbs, chol, Exang, etc. are chosen for forecasting. The attributes for the models are selected using a correlation matrix.

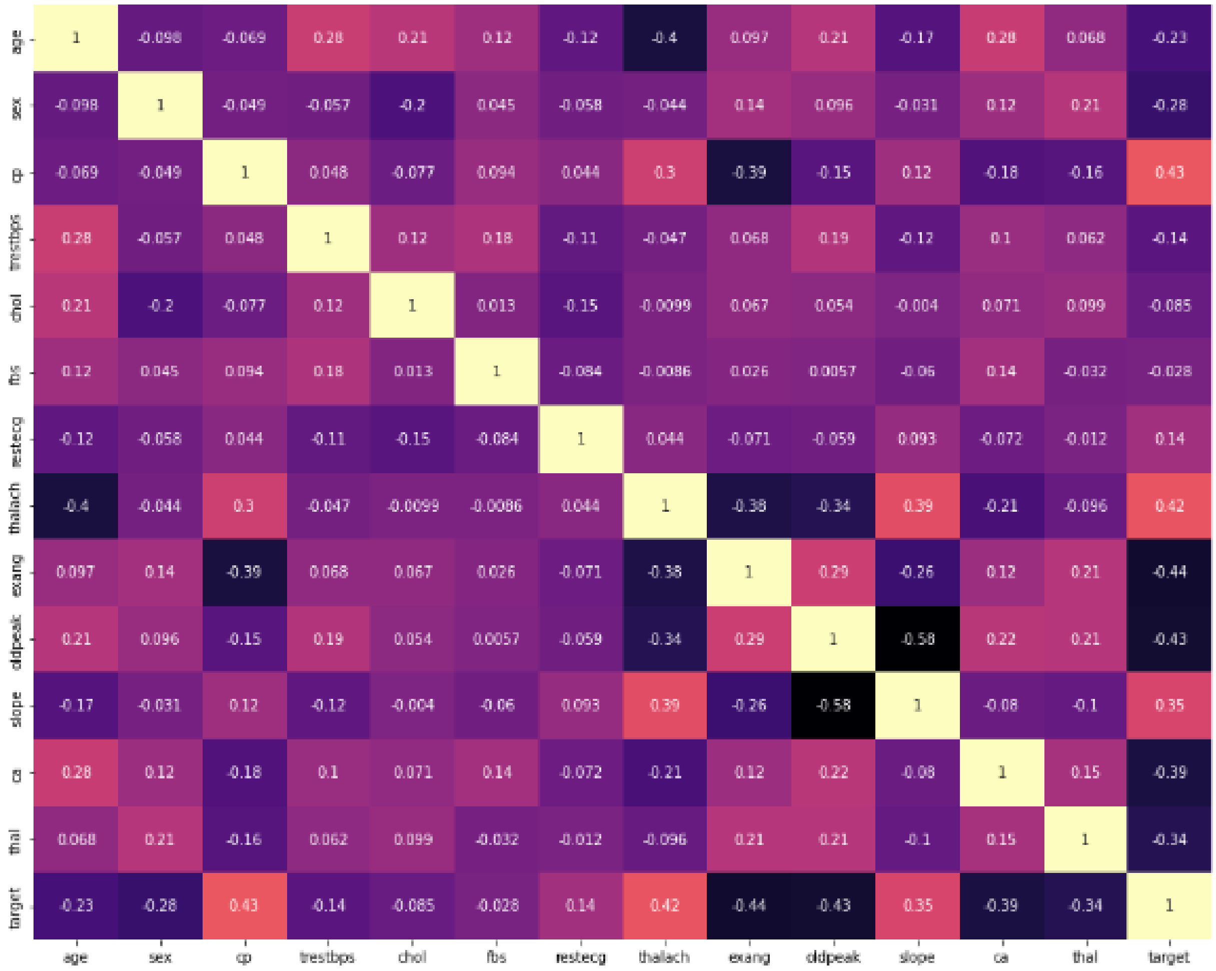


Figure 5.1 Correlation Matrix

It can be seen from the correlation matrix that there isn't a single attribute that is highly correlated with the target.

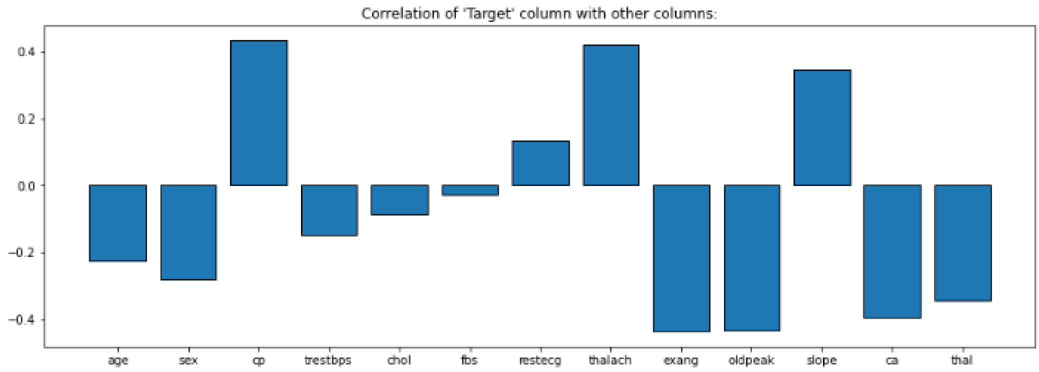


Figure 5.2 Correlation with ‘Target’ Column

Similarly, the correlation between independent variables is also not high. Which means there is no single attribute which has its major impact in deciding the result. So, all these attributes will be considered in this project just to ensure that no attribute is left out without consideration.

5.1.3 DATA PRE-PROCESSING

A crucial step in training a machine learning model is data pre-processing. At first, the data might not be accurate or might not be in the model's needed format, which could lead to false findings. In the data pre-processing, the data is converted into the desired format first and then for dealing with sounds, duplicates, and non-database values. Data pre-processing has functions such as importing data sets, separating data sets, attribute measurements, etc. Preliminary data processing is required to improve model accuracy.

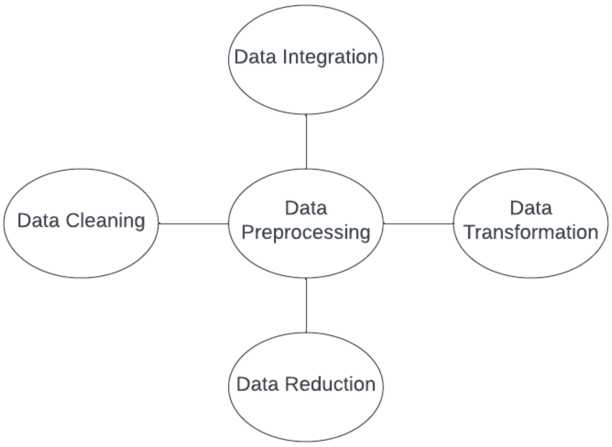
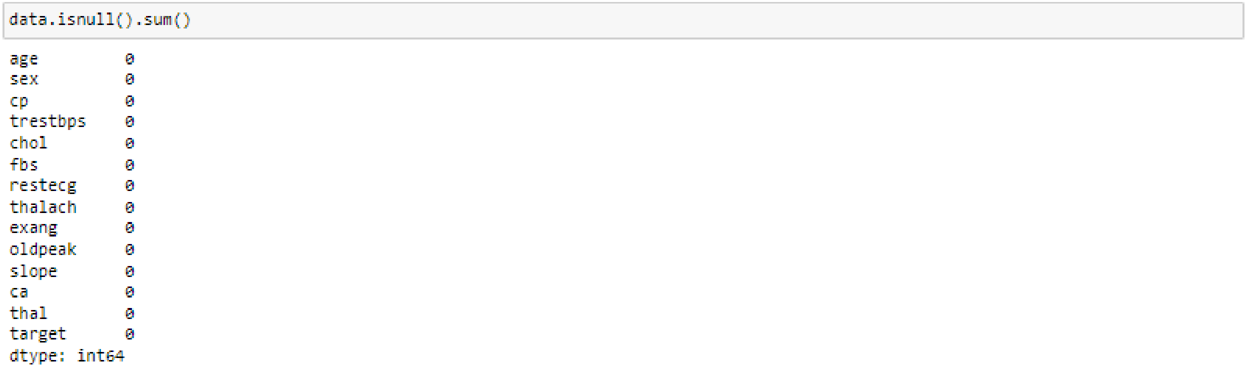


Figure 5.3 Importance of Data Pre-processing

5.1.3.1 Handling Null values

When you don't have data stored for specific variables or participants, you have missing data, also known as missing values. Various factors might cause data loss, including incorrect data entry, device failures, lost files, and more. It can weaken a study's statistical power and provide skewed estimates, which can result in false findings.



```
data.isnull().sum()
```

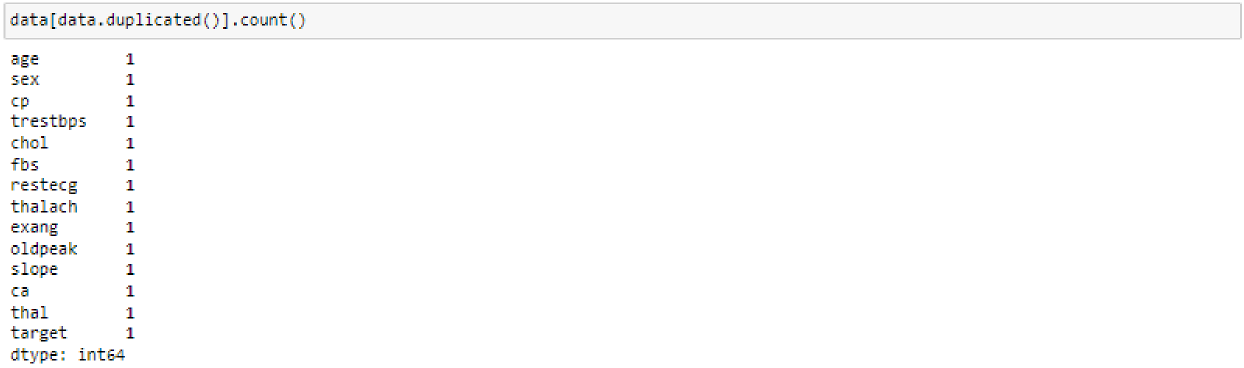
```
age         0
sex         0
cp          0
trestbps    0
chol        0
fbs         0
restecg     0
thalach     0
exang       0
oldpeak     0
slope       0
ca          0
thal        0
target      0
dtype: int64
```

Figure 5.4 Null value counts for each attribute

It can be observed that the dataset has no null values.

5.1.3.2 Check for duplicate values

A duplicate value is one where every single value in at least one row is the exact same as every single value in every other row. The appearance of the cell, not the underlying value stored in the cell, determines how duplicate values are compared. Additionally, duplicate entries can ruin the split between train, validation and test sets in cases where identical entries are not all in the same set. This can lead to biased performance estimates that will lead to disappointing models in production.



```
data[data.duplicated()].count()
```

```
age         1
sex         1
cp          1
trestbps    1
chol        1
fbs         1
restecg     1
thalach     1
exang       1
oldpeak     1
slope       1
ca          1
thal        1
target      1
dtype: int64
```

Figure 5.5 Duplicate value counts

The dataset had one duplicate value and it was removed for further processing.

5.1.3.3 Check for outliers

An observation that appears to differ significantly from other observations in the sample is known as an outlier. It's crucial to identify probable outliers for the following reasons:

- An outlier might represent flawed data. For instance, an experiment might not have been properly run or the data might have been coded wrongly. If an outlying point is found to be incorrect, the outlying value should be eliminated from the analysis (or corrected if possible).
- It might not always be feasible to tell whether an outlying point contains bad data. In addition to being the result of random fluctuation, outliers can also point to important scientific phenomena. In any case, it is not appropriate to simply discard the outlier observation. However, the use of robust statistical approaches may need to be taken into consideration if the data contains considerable outliers.

3 techniques for finding out outliers in the dataset are used, those are:

i. USING BOXPLOT

This technique finds the data set's maximum and minimum values. The data set's median is at the 50% percentile. The median of the data between the minimum and maximum is in the first quartile, while the median of the data between the maximum and minimum is in the third quartile. The values that fall outside of the (1.5*interquartile range) of the 25 or 75 percentiles will be considered outliers.

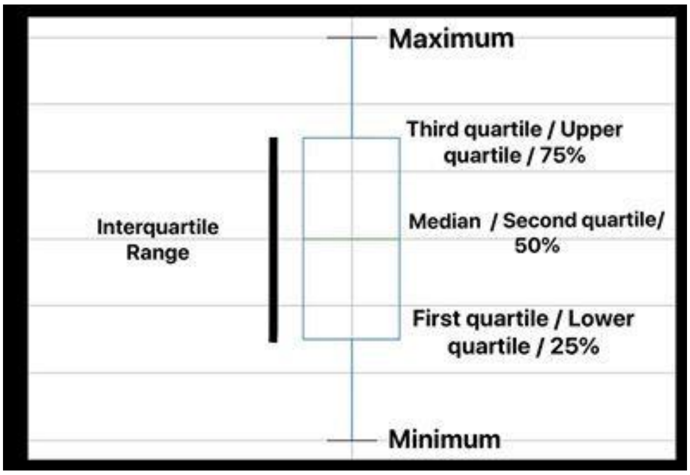

Maximum
Third quartile / Upper quartile / 75%
Interquartile Range
Median / Second quartile/ 50%
First quartile / Lower quartile / 25%
Minimum

Figure 5.6 Quartile Range

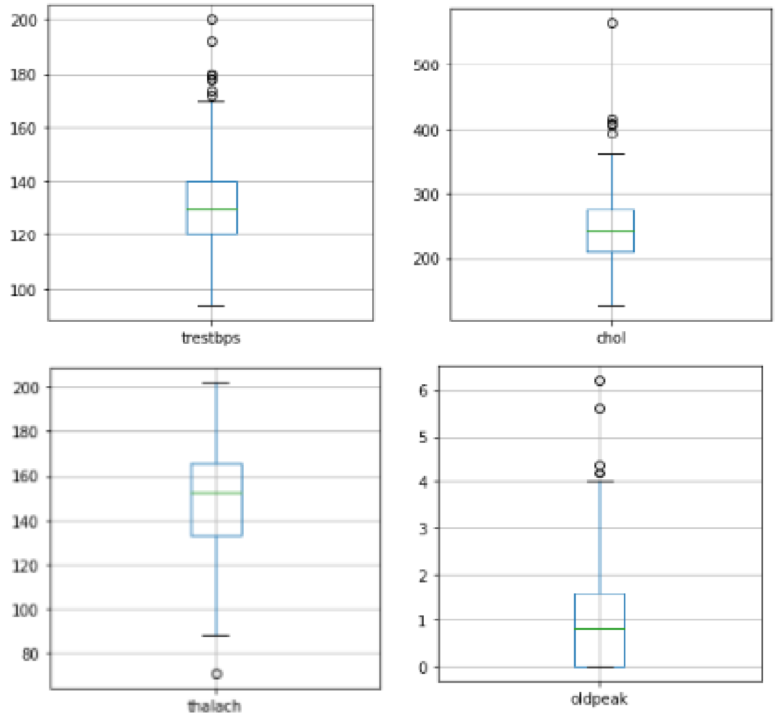


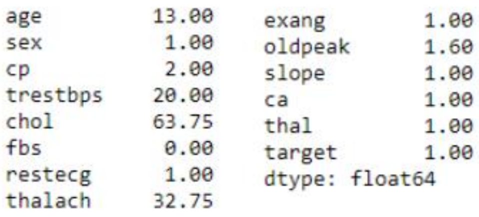
Figure 5.7 Boxplots showing outliers

After considering the boxplots for all the attributes it can be noticed that the columns trestbps, chol, thalach and oldpeak have outliers.

ii. USING IQR SCORE

This method also uses the quartile range with the formula:

IQR = Third Quartile - First Quartile



```
age          13.00    exang        1.00
sex           1.00    oldpeak      1.60
cp            2.00    slope        1.00
trestbps     20.00    ca           1.00
chol         63.75    thal         1.00
fbs           0.00    target       1.00
restecg       1.00    dtype: float64
thalach      32.75
```

Figure 5.8 IQR Score for all attributes

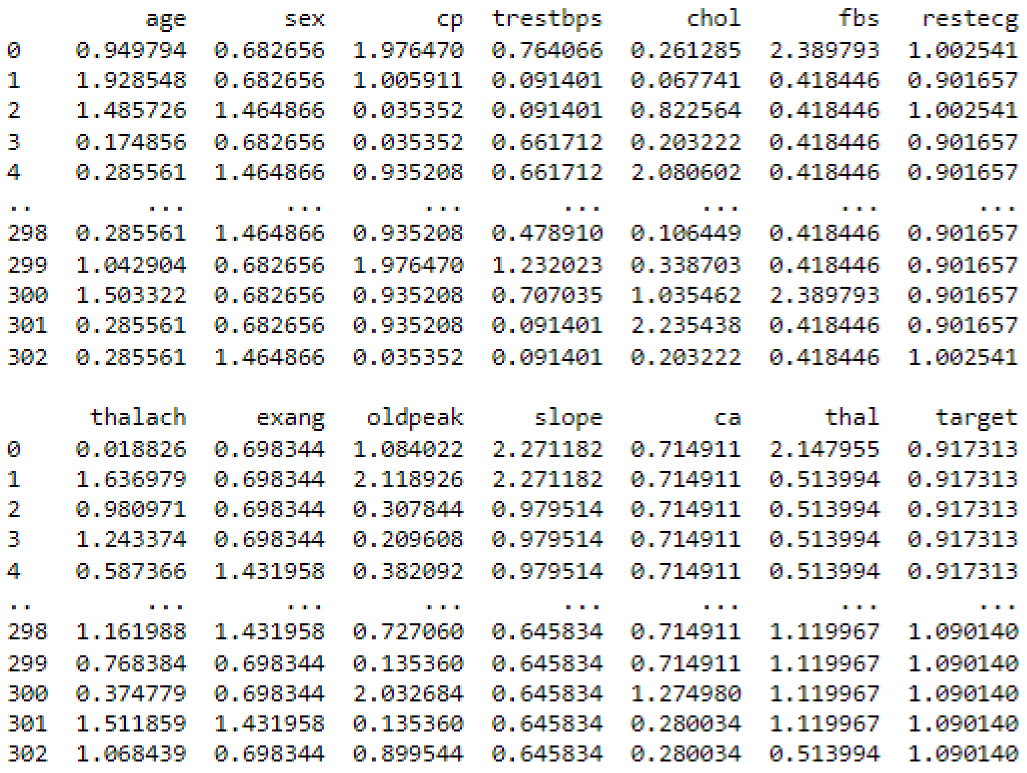
And from the above output it can be noticed that the column trestbps, chol and thalach have outliers.

iii. USING Z SCORE

When the data has a normal distribution, Z-scores help identify how odd an observation is. Z-scores are the standard deviations of the ranges of values above and below the mean. A Z-score of 2 indicates that the observation is 2 standard deviations above the mean, whereas a Z-score of -2 indicates that the observation is two standard deviations below the mean. A value equal to the mean has a Z-score of zero. It is computed using the following formula:

$$Z\ Score\ =\ \frac{X\ -\ \mu}{\sigma}$$

From below output it is found that the columns: 3, 4, 7 and 9 have outliers.



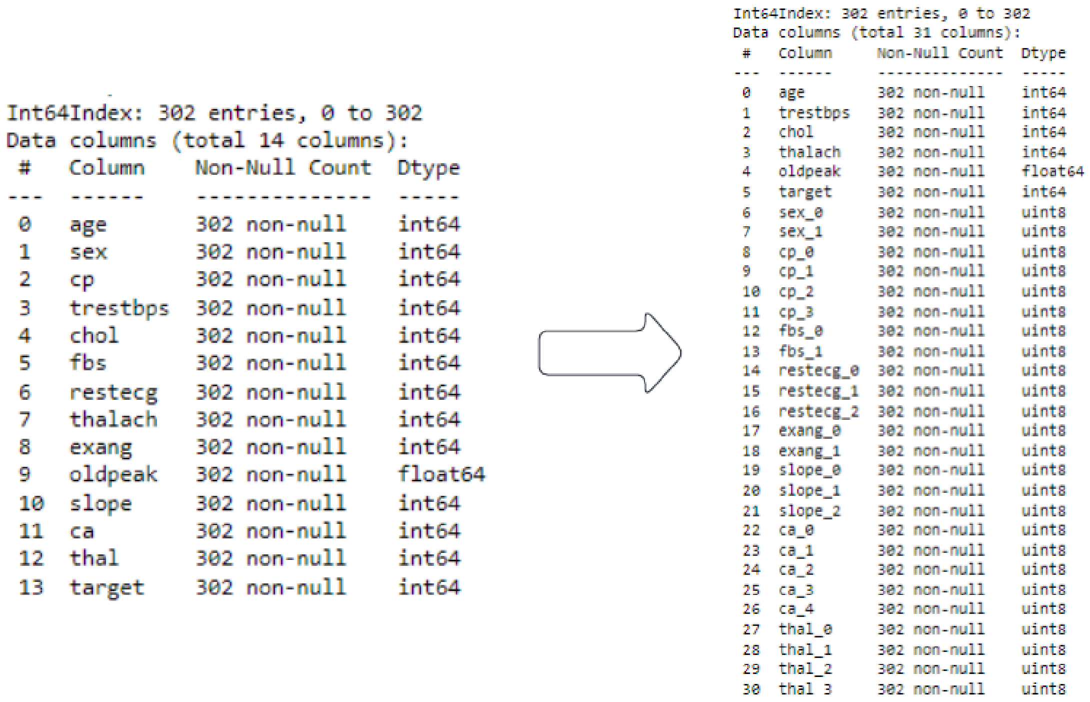
```
          age       sex        cp  trestbps      chol       fbs   restecg
0    0.949794  0.682656  1.976470  0.764066  0.261285  2.389793  1.002541
1    1.928548  0.682656  1.005911  0.091401  0.067741  0.418446  0.901657
2    1.485726  1.464866  0.035352  0.091401  0.822564  0.418446  1.002541
3    0.174856  0.682656  0.035352  0.661712  0.203222  0.418446  0.901657
4    0.285561  1.464866  0.935208  0.661712  2.080602  0.418446  0.901657
..        ...       ...       ...       ...       ...       ...       ...
298  0.285561  1.464866  0.935208  0.478910  0.106449  0.418446  0.901657
299  1.042904  0.682656  1.976470  1.232023  0.338703  0.418446  0.901657
300  1.503322  0.682656  0.935208  0.707035  1.035462  2.389793  0.901657
301  0.285561  0.682656  0.935208  0.091401  2.235438  0.418446  0.901657
302  0.285561  1.464866  0.035352  0.091401  0.203222  0.418446  1.002541

      thalach     exang   oldpeak     slope        ca      thal    target
0    0.018826  0.698344  1.084022  2.271182  0.714911  2.147955  0.917313
1    1.636979  0.698344  2.118926  2.271182  0.714911  0.513994  0.917313
2    0.980971  0.698344  0.307844  0.979514  0.714911  0.513994  0.917313
3    1.243374  0.698344  0.209608  0.979514  0.714911  0.513994  0.917313
4    0.587366  1.431958  0.382092  0.979514  0.714911  0.513994  0.917313
..        ...       ...       ...       ...       ...       ...       ...
298  1.161988  1.431958  0.727060  0.645834  0.714911  1.119967  1.090140
299  0.768384  0.698344  0.135360  0.645834  0.714911  1.119967  1.090140
300  0.374779  0.698344  2.032684  0.645834  1.274980  1.119967  1.090140
301  1.511859  1.431958  0.135360  0.645834  0.280034  1.119967  1.090140
302  1.068439  0.698344  0.899544  0.645834  0.280034  0.513994  1.090140
```

Figure 5.9 Z Score values

From the above three methods, it can be noticed that the state and confirm present are outliers in the columns trestbps, chol, thalach and oldpeak.

5.1.3.4 Handling Categorical Values

All input and output variables must be numbers for ML models. This means that in order to fit and assess a model, categorical data must first be converted to integers. Here, hot encoding is applied to that one. Each category value is transformed into a new categorical column in one-hot, and each column is given a binary value of 1 or 0. A binary vector is used to represent each integer value. The index is designated with 1, and all of the values are zero.



```
Int64Index: 302 entries, 0 to 302
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       302 non-null    int64
 1   sex       302 non-null    int64
 2   cp        302 non-null    int64
 3   trestbps  302 non-null    int64
 4   chol      302 non-null    int64
 5   fbs       302 non-null    int64
 6   restecg   302 non-null    int64
 7   thalach   302 non-null    int64
 8   exang     302 non-null    int64
 9   oldpeak   302 non-null    float64
 10  slope     302 non-null    int64
 11  ca        302 non-null    int64
 12  thal      302 non-null    int64
 13  target    302 non-null    int64
```

```
Int64Index: 302 entries, 0 to 302
Data columns (total 31 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   age        302 non-null    int64
 1   trestbps   302 non-null    int64
 2   chol       302 non-null    int64
 3   thalach    302 non-null    int64
 4   oldpeak    302 non-null    float64
 5   target     302 non-null    int64
 6   sex_0      302 non-null    uint8
 7   sex_1      302 non-null    uint8
 8   cp_0       302 non-null    uint8
 9   cp_1       302 non-null    uint8
 10  cp_2       302 non-null    uint8
 11  cp_3       302 non-null    uint8
 12  fbs_0      302 non-null    uint8
 13  fbs_1      302 non-null    uint8
 14  restecg_0  302 non-null    uint8
 15  restecg_1  302 non-null    uint8
 16  restecg_2  302 non-null    uint8
 17  exang_0    302 non-null    uint8
 18  exang_1    302 non-null    uint8
 19  slope_0    302 non-null    uint8
 20  slope_1    302 non-null    uint8
 21  slope_2    302 non-null    uint8
 22  ca_0       302 non-null    uint8
 23  ca_1       302 non-null    uint8
 24  ca_2       302 non-null    uint8
 25  ca_3       302 non-null    uint8
 26  ca_4       302 non-null    uint8
 27  thal_0     302 non-null    uint8
 28  thal_1     302 non-null    uint8
 29  thal_2     302 non-null    uint8
 30  thal 3     302 non-null    uint8
```

Figure 5.10 Conversion of categorical values to numeric

Hence, we end up getting 31 columns from 14 columns, where columns sex, cp, fbs, recteg, Exang, slope, ca and thal are the columns which get converted during this phase.

5.1.3.5 Standardization

It entails setting the variable's centre at 0 and the variance standard at 1. The process entails dividing by the standard deviation after subtracting the mean of each observation. As a result of standardisation, the features are rescaled to have the characteristics of a normal distribution with a mean of 0 and a standard deviation from the mean of 1. StandardScalar (), a Scikit-Learn transformer, is employed in this.

When continuous independent variables are measured at several scales, the idea of standardisation becomes relevant. Algorithms that use distance measurements, like K-Nearest-Neighbors, also employ rescaling (KNN). This indicates that these factors do not contribute equally to the analysis.

In circumstances when the data has a Gaussian distribution, standardisation can be useful. This need not necessarily be the case, though. Geometrically speaking, it shrinks or extends the points depending on whether std is 1 and converts the data to the mean vector of the original data to the origin. We can see that the distribution's shape is unaffected because we are simply altering the mean and standard deviation to a standard normal distribution, which is still normal.

Outliers have no impact on standardisation because there is no predetermined range of converted features.

5.1.4 DATA BALANCING

Unbalanced data sets with class A having 90 observations and class B having 10 observations are frequently encountered in the real world. Balanced dataset is one of the conditionings for machine learning, or at least near to being balanced. In simple terms, this is done in order to give each class equal precedence. Unbalanced datasets can be balanced using one of two methods.

i. **Under Sampling:** It achieves dataset balance by decreasing the size of the sample class. When the amount of data is sufficient, this technique is examined.

ii. **Over Sampling:** It achieves dataset balance by increasing the size of the sparse samples. When the amount of data available is insufficient, this procedure is explored.

In this case the data is almost balanced which can be seen from the figure below.

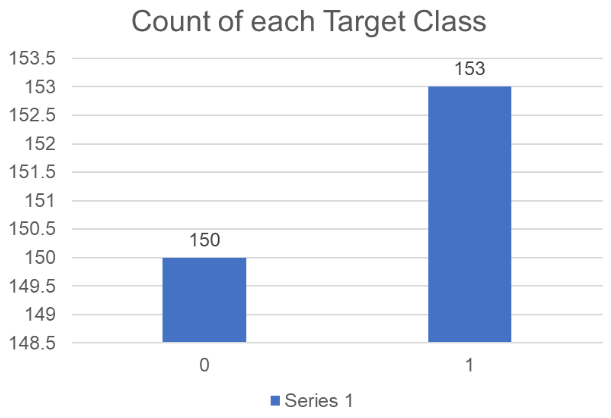


Figure 5.11 Target class count

5.1.5 DISEASE PREDICTION

Machine learning algorithms such as Logistic Regression, SVM, K-Nearest Neighbor and Random Forest are used for segmentation. All these models are analysed and compared for selecting an algorithm which provides high accuracy and hence used in this model for predicting heart disease.

5.2 DESCRIPTION OF PROCESS

When referring to classification in the context of machine learning, we are referring to the model prediction problem where a specific sample of input data is used to anticipate the class label.

i. Supervised Learning

Machines are trained using labelled "data" in the process of "supervised learning," which then allows them to predict outcomes using the knowledge gained. Labeled data refers to input data that has already been assigned the appropriate output. In supervised learning, the machine-provided training data serves as a supervisor, instructing the machines on how to accurately estimate output. The goal of the supervised learning method is to identify a feature map to distinct inputs (x) and outputs (y). A machine learning model is subjected to supervised learning, which involves giving correct input data and output data.

ii. Unsupervised learning

Contrary to supervised reading, unsupervised learning uses input data but no matching output data, hence it can't be employed directly in split issues. Unsupervised learning aims to identify a fundamental database structure, gather data based on similarities, then display that database in a compacted way.

1. Unsupervised learning is advantageous for drawing meaningful conclusions from the data.
2. Unsupervised learning is more like how humans learn to think from their own experiences, which brings it closer to true artificial intelligence.
3. Unsupervised learning uses uncategorized and unlabeled data, which increases its importance.
4. It isn’t always had input data with equivalent output in the actual world, therefore unsupervised learning is needed to handle these situations.

iii. Reinforcement learning

Machine learning includes reinforcement learning as one of its elements. It involves acting in a way that will maximise benefits in a certain circumstance. Computer programmes and other applications utilise it to decide what action to take or what behaviour is best in a particular circumstance. In contrast to supervised learning, where the solution key is present in the training data and allows the model to be taught with the appropriate response, reinforcement learning lacks this information and instead relies on the reinforcement agent to determine how to carry out the task at hand. It must gain knowledge through experience when there isn't a training dataset available.

5.2.1 EVALUATION MODELS

In this research, cardiovascular disease is predicted using a variety of machine learning methods, including Random Forest, SVM, K-NN and Logistic Regression. To predict the better among them we need some standards for comparing between them. There are several evaluation measures available for the experiment, including accuracy, precision, recall, specificity, confusion matrix, PR curve, ROC curve, and f1-score. Whichever algorithm provides the highest accuracy will be used to predict heart disease.

The commonly used top 5 methods:

i. Confusion Matrix

- It is a matrix utilized to assess how well the separation models perform given a certain set of test data. Only after the test data's actual values are known can it be determined. Although the matrix itself is simple to understand, some of the related terminology might be. It is also referred to as an error matrix since it displays model performance errors as a matrix. Below are a few characteristics of the Confusion matrix:

- The matrix is 2*2 table for classifiers with two prediction classes, 3*3 tables for classifiers with three prediction classes, and more.
- The observed and predicted values, and also the overall number of predictions, are split into two dimensions in the matrix.
- Predicted values are those obtained from the model, whereas observed values are the actual values for the provided data.

observe the table below:

N = Total no. of Predictions	**Actual: No**	**Actual: Yes**
Predicted as: No	True Negative	False Negative
Predicted as: Yes	False Positive	True Positive

Table 5.2 Confusion Matrix

The table above has the following cases:

- True Negative (TN): The model predicted no, and the actual value was also No.
- True Positive (TP): The model predicted yes, and the actual value was also .yes
- False Negative (FN): known as type II error. The model predicted no, but the actual value was Yes.
- False Positive (FP): known as I-type error. The model predicted Yes, but the actual number was No.

ii. Precision

It is the number of positive results given a model or in all the positive categories that accurately predicted a model, how many of these were correct. It is computed using the following formula:

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

iii. Recall

This model's accuracy is described as the output of the total number of classes. Recovery should be as high as possible. It is computed using the following formula:

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

iv. F1 Score

It is difficult to compare two models that have poor comprehension but strong memory or vice versa. Therefore, for this reason, F1 points are used. This result helps us to assess memory and accuracy at the same time. The F1 score is higher if the memory is equal to accuracy. It is computed using the following formula:

$$F1\ Score = \frac{2 * Precision * Recall}{Precision + Recall}$$

v. Accuracy

It is one of the crucial factors in figuring out how accurate a classification problem is. It specifies how frequently the model predicts the right result. The number of accurate predictions produced by the classifier divided by the total number of predictions made by the classifiers may be used to compute it. The following is the formula:

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

5.2.2 CLASSIFICATION MODELS

5.2.2.1 Support Vector Machine (SVM)

SVM, or Support Vector Machine, is a common Supervised Learning technique for scheduling and re - training issues. In machine learning, distribution problems are where it is most commonly utilised.

Finding a hyperplane in an n-dimensional space (where n is the number of features) that clearly classifies the data points is the goal of the support vector machine algorithm. The hyperplane is called the best decision limit. SVM identifies the additional vectors that aid in the formation of the hyperplane. The approach is known as a "vector support machine" because these extreme examples are known as supporting vectors.

SVMs are strong yet flexible guided machine learning tools that are utilised for both modification and recovery. However, they are more commonly used in separate issues. In comparison to other machine learning algorithms, SVMs have their own operating system.

The following are important concepts in SVM -

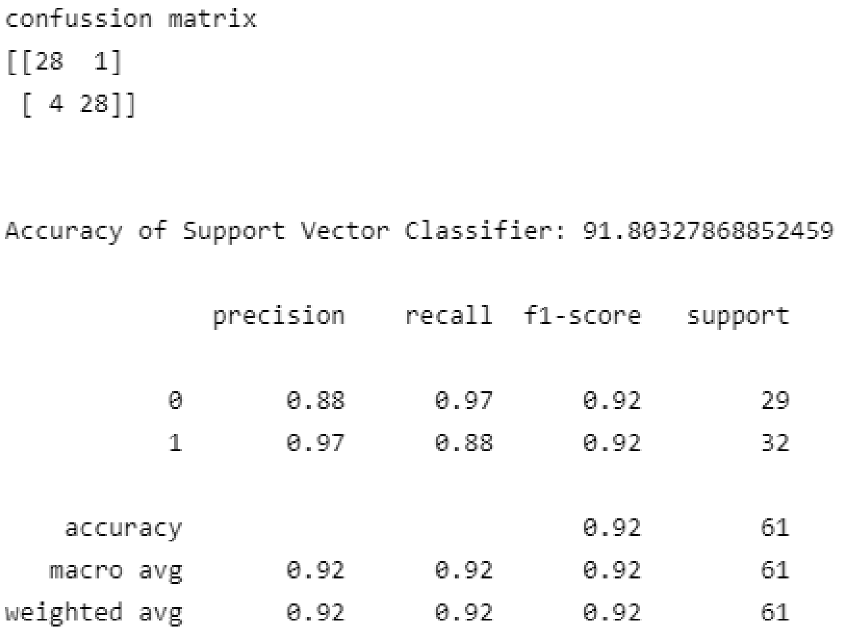
1. **Support Vectors -**Support vectors are the data points that are closest to the hyperplane. These data points will be used to generate the separation line.
2. **Hyperplane -** In SVM, it is a decision border that distinguishes the two classes.
3. **Margin -** It is defined as the distance between two lines drawn on the nearest data points for distinct classifications. It is computed using the perpendicular distance between the line and the support vectors. A big margin is preferred.

Advantages of SVM:

- Effective in places with high dimensions.
- When the number of samples is greater than the number of dimensions, the method remains effective.
- Using fewer training points in the decision function conserves memory as well.
- Versatile: For the decision function, various kernel functions can be defined. Common kernels are given, however custom kernels can also be specified.

Disadvantages of SVM:

- When choosing Kernel functions, avoid overfitting if samples are more than the amount of features, and the regularisation term is crucial. SVMs use a time-consuming five-fold cross-validation approach instead of directly estimating probabilities.



```
confussion matrix
[[28  1]
 [ 4 28]]


Accuracy of Support Vector Classifier: 91.80327868852459

              precision    recall  f1-score   support

           0       0.88      0.97      0.92        29
           1       0.97      0.88      0.92        32

    accuracy                           0.92        61
   macro avg       0.92      0.92      0.92        61
weighted avg       0.92      0.92      0.92        61
```

Figure 5.11 Support Vector Machine Test Result

5.2.2.2 Random Forest Classifier

Using the best among smaller prediction sets that are randomly picked from the node itself, Random Forests splits nodes. O (M (dnlogn)), where M is the number of trees developing, n is the number of events, and d is the amount of the data, is the most difficult worst-case scenario for the study of Random Forests.

In addition to regression, it may be used for classification. It serves as a good indicator of the feature's relevance. It may be used to detect fraud, categorize dependable loan candidates, and foresee illnesses.

Its foundation is the idea of ensemble learning, which entails combining a large number of classifiers to tackle a challenging issue and increase the model's efficiency. The accuracy increases and the risk of overfitting decreases as the number of trees in the forest increases.

Algorithm Steps: involves four steps:

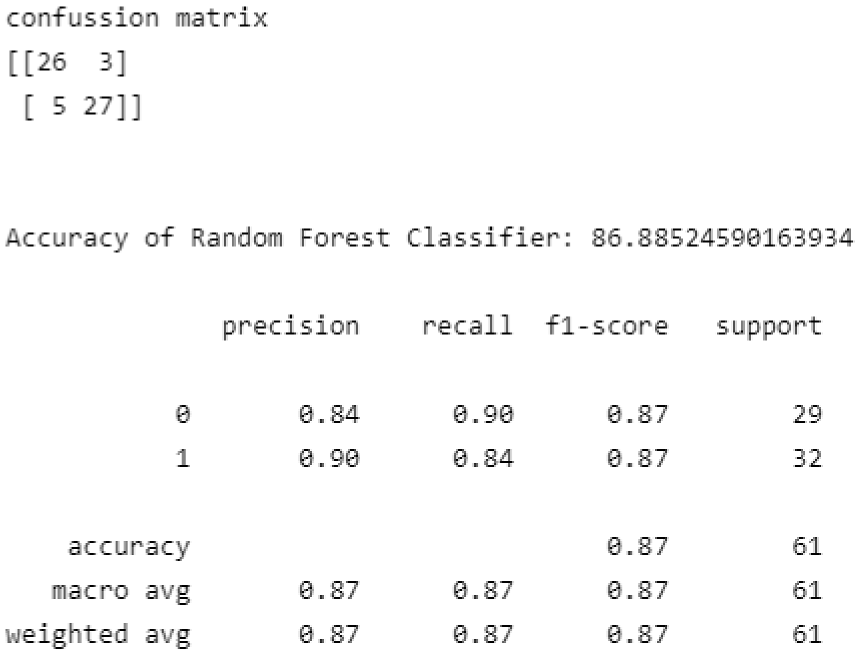
1. Select at random from the available dataset.
2. For every sample, create a Decision Tree and get a prediction from it.
3. For each predicted result, cast a vote.
4. As the final forecast, choose the prediction result with the most votes.

Advantages:

- Random Forest can perform problems involving both classification and regression.
- It can handle huge datasets with high dimensionality.
- It improves model accuracy and avoids the overfitting problem.

Disadvantages:

- While both classification and regression tasks may be performed with Random Forest, regression is not its strongest strength.



```
confussion matrix
[[26  3]
 [ 5 27]]


Accuracy of Random Forest Classifier: 86.88524590163934

              precision    recall  f1-score   support

           0       0.84      0.90      0.87        29
           1       0.90      0.84      0.87        32

    accuracy                           0.87        61
   macro avg       0.87      0.87      0.87        61
weighted avg       0.87      0.87      0.87        61
```

Figure 5.12 Random Forest Classifier Test Result

5.2.2.3 Logistic Regression Classifier

Logistic regression is a common machine learning approach that falls under the category of supervised learning. From a collection of independent factors, it is used to forecast the categorical dependent variable.

Logistic regression is used to predict the output of a dependent categorical variable. A discrete or categorical value must consequently be the result. It may be True or False, Yes or No, 0 or 1, and so on, but rather than giving the exact values like 0 and 1, it displays the probability values that fall between 0 and 1.

Logistic regression and linear regression are fairly similar, with the exception of how they are used. Regression issues are resolved using linear models, and classification issues are resolved using logistic models.

Rather than a regression line, we fit a "S"-shaped logistic function that predicts two maximum values in logistic regression (0 or 1). Depending on a cat's weight, the logistic function curve may be used to predict a variety of outcomes, such as whether the cells are malignant or not, if the cat is overweight, and more.

As it can produce probabilities and categorise new data using both continuous and discrete datasets, logistic regression is a crucial machine learning technique.

Advantages:

- One of the most fundamental machine learning methods is logistic regression and it is simple to construct while providing excellent training efficiency in some circumstances. Because of these factors, training a model with this technique does not necessitate a large number of computational resources.

- The projected parameters (trained weights) infer the significance of each characteristic. It also specifies if the relationship is positive or negative. As a conclusion, we may use Logistic Regression to verify how the characteristics are related to one another.

- In contrast to Decision Tree and Support Vector Machine, models may be easily updated using this method to account for new data. Stochastic gradient descent can be used to update the model.

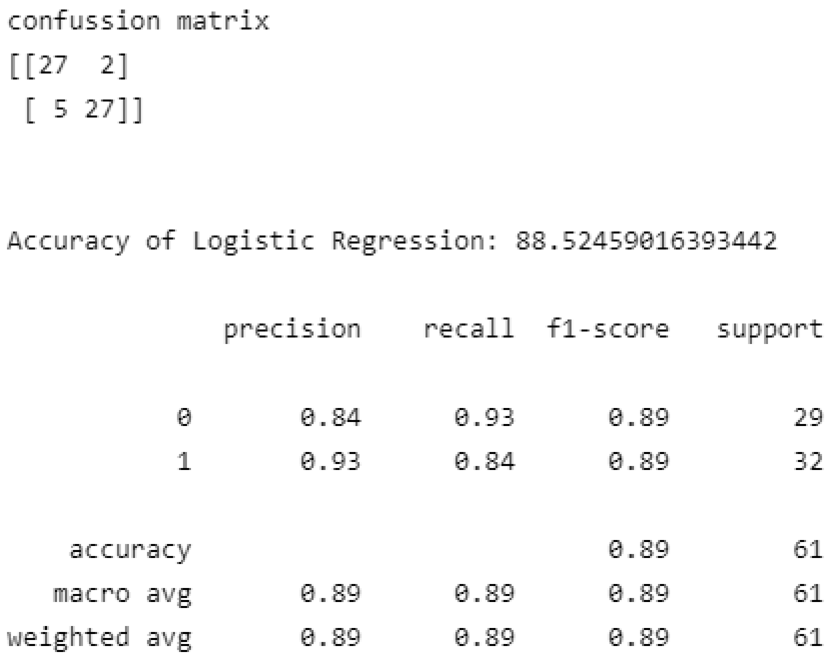
- Logistic Regression produces calibrated probability as well as classification results.

Disadvantages:

- A statistical analysis method known as logistic regression employs independent data to precisely forecast probability outcomes. On high-dimensional datasets, this might result in the model being over-fit on the training set, which would overestimate the accuracy of predictions and prevent the model from accurately predicting results on the test set. Regularization techniques should be investigated to prevent overfitting on high-dimensional datasets (but this makes the model complex). The model might potentially underfit the training set of data if the regularisation variables are extremely high.

- Logistic regression cannot tackle nonlinear issues because it has a linear decision surface. In reality, linearly separable data is rare. In order to make the data linearly separable in higher dimensions, nonlinear properties must be converted, which may be done by adding more features.

- Non-Linearly Separable Data: Complex associations are difficult to represent with logistic regression. This approach is readily outperformed by more powerful and complicated algorithms, such as Neural Networks.



```
confussion matrix
[[27  2]
 [ 5 27]]


Accuracy of Logistic Regression: 88.52459016393442

              precision    recall  f1-score   support

           0       0.84      0.93      0.89        29
           1       0.93      0.84      0.89        32

    accuracy                           0.89        61
   macro avg       0.89      0.89      0.89        61
weighted avg       0.89      0.89      0.89        61
```

Figure 5.13 Logistic Regression Classifier Test Result

5.2.2.4 K-Nearest Neighbor Classifier

A fundamental machine learning technique that makes advantage of supervised learning is K-Nearest Neighbor. The K-NN approach groups the new instance into the category that is most similar to the existing categories by assuming familiarity between the new case/data and previous cases.

The K-NN method uses similarity to categorise new data points while preserving all previously gathered data. This suggests that when new data is received, it can be swiftly categorised using the K-NN approach.

The K-NN approach is most frequently applied in classification jobs. Since it is non-parametric, no assumptions are made on the underlying data. It is also known as a lazy learner algorithm since it saves the training set before acting on it during classification rather than learning from it immediately.

The K-NN method simply saves the dataset and categorises it into a category that is relatively similar to the incoming data during the training phase.

Algorithm Steps: It works in five steps:

1. First, choose the Kth neighbour.
2. Next, determine the K number of neighbours' Euclidean distance.
3. Based on the calculated Euclidean distance, find the K nearest neighbours.
4. Among these k neighbours, total the amount of data points in each category.
5. To the category with the most neighbours, assign the new data points.

Selecting optimum value of K:

Here are some things to keep in mind while choosing the value of K in the K-NN algorithm:

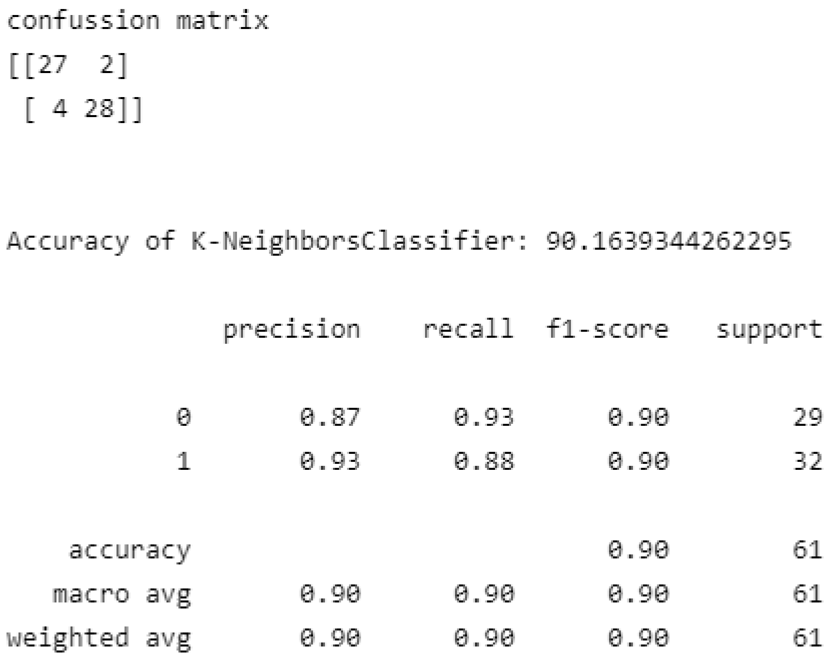
- There is no specific technique to identify the optimal value for "K," therefore we must experiment with many values to get the best one. The most popular K value is 5.
- K might have a variable value, such as K=1 or K=2, which could lead to outlier effects in the model.
- Although they could cause some issues, large values for K are preferred.

Advantages of KNN Algorithm:

- Implementing it is easy.
- It can withstand noisy training data.
- If there is a lot of training data, it could work better.

Disadvantages of KNN Algorithm:

- It is always necessary to identify the value of K, which might be difficult at times.
- The calculation cost is considerable since the distance between data points for all training samples is calculated.

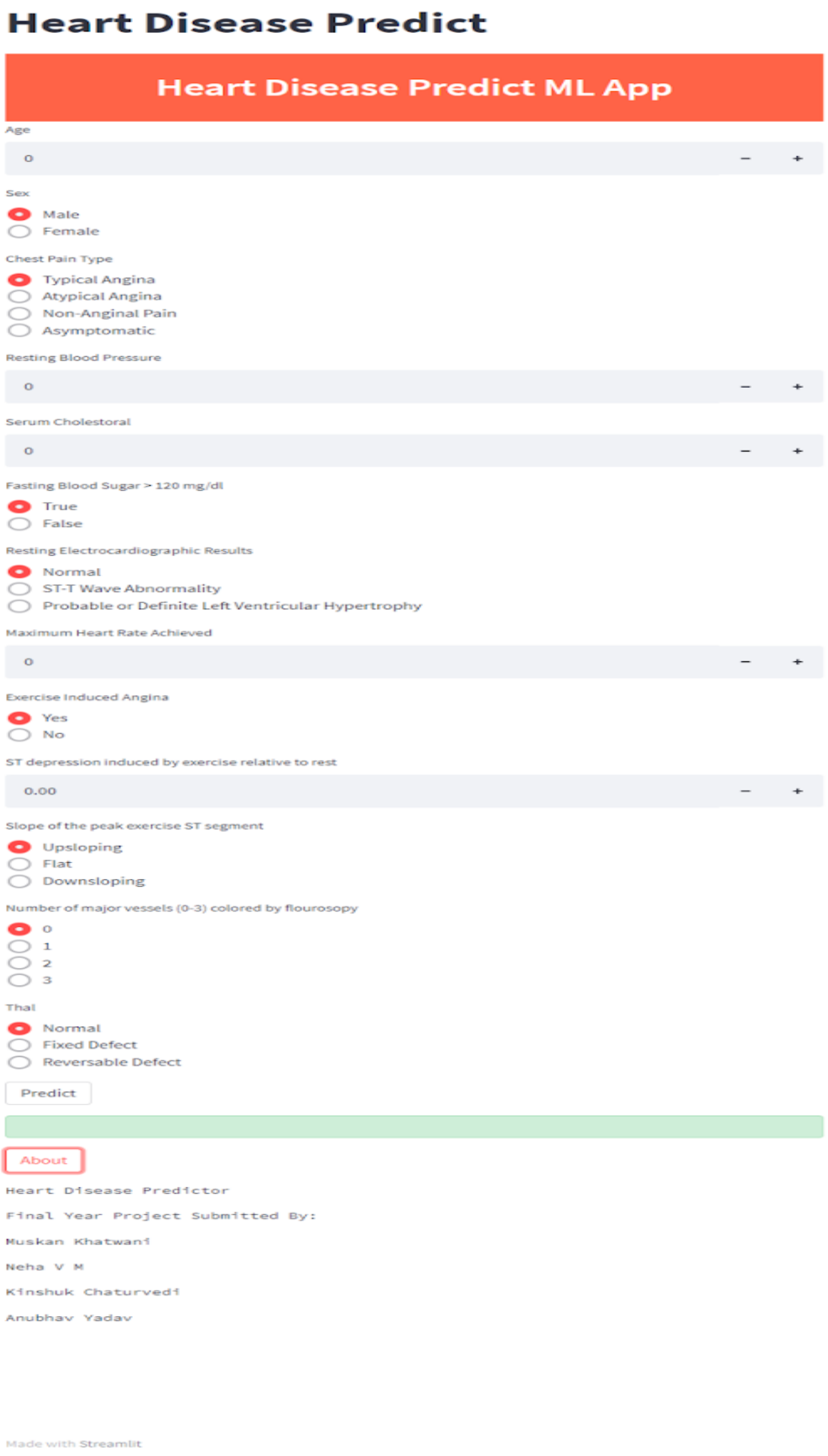


```
confussion matrix
[[27  2]
 [ 4 28]]


Accuracy of K-NeighborsClassifier: 90.1639344262295

              precision    recall  f1-score   support

           0       0.87      0.93      0.90        29
           1       0.93      0.88      0.90        32

    accuracy                           0.90        61
   macro avg       0.90      0.90      0.90        61
weighted avg       0.90      0.90      0.90        61
```

Figure 5.14 K-Nearest Neighbor Classifier Test Result

CHAPTER 6: TEST CASES

i. The ideal Web API looks like shown in the figure below:



Heart Disease Predict

Heart Disease Predict ML App

Age

0 − +

Sex

● Male
○ Female

Chest Pain Type

● Typical Angina
○ Atypical Angina
○ Non-Anginal Pain
○ Asymptomatic

Resting Blood Pressure

0 − +

Serum Cholestoral

0 − +

Fasting Blood Sugar > 120 mg/dl

● True
○ False

Resting Electrocardiographic Results

● Normal
○ ST-T Wave Abnormality
○ Probable or Definite Left Ventricular Hypertrophy

Maximum Heart Rate Achieved

0 − +

Exercise Induced Angina

● Yes
○ No

ST depression induced by exercise relative to rest

0.00 − +

Slope of the peak exercise ST segment

● Upsloping
○ Flat
○ Downsloping

Number of major vessels (0-3) colored by flourosopy

● 0
○ 1
○ 2
○ 3

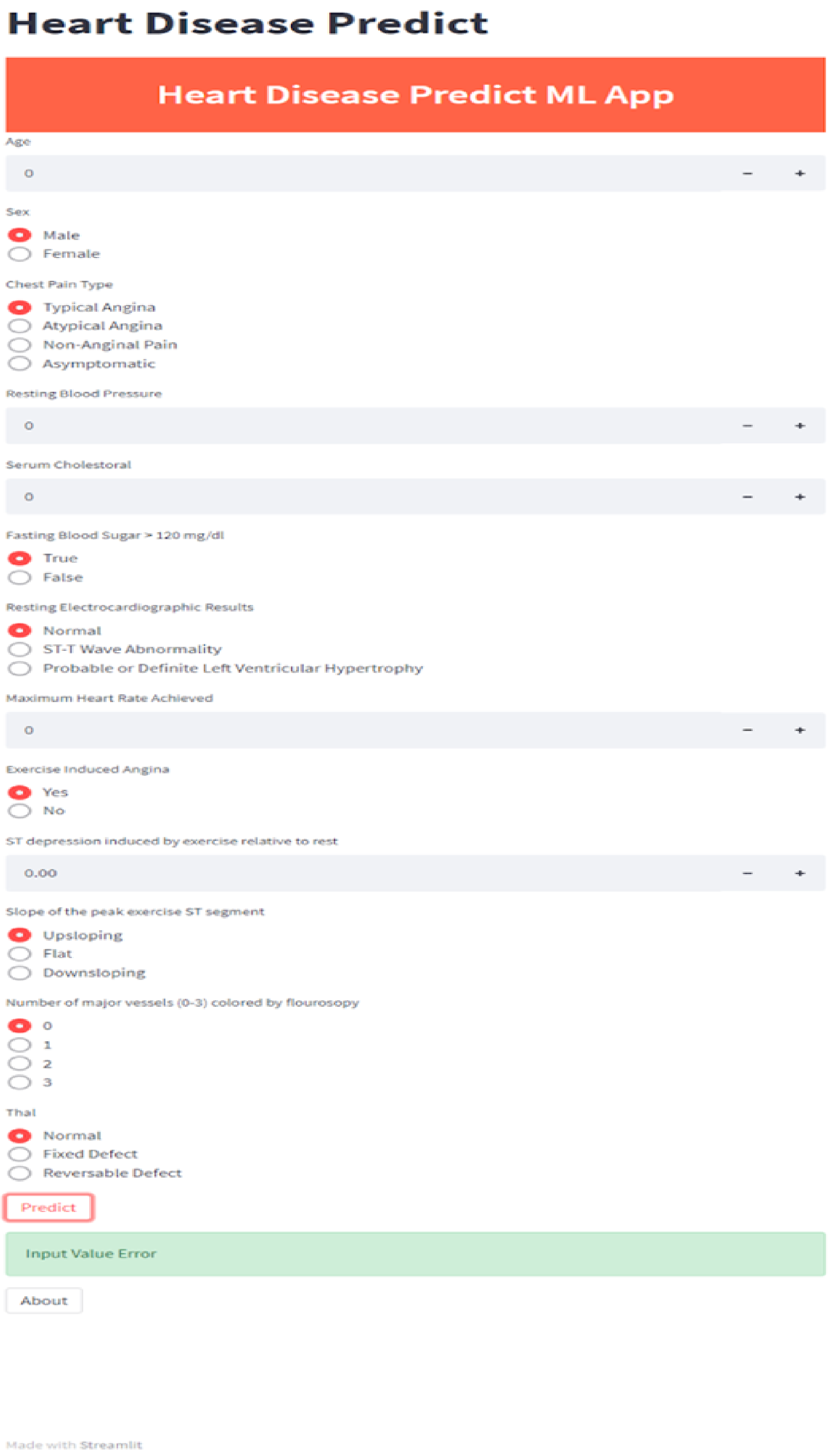
Thal

● Normal
○ Fixed Defect
○ Reversable Defect

```
Heart Disease Predictor
Final Year Project Submitted By:
Muskan Khatwani
Neha V M
Kinshuk Chaturvedi
Anubhav Yadav
```

Figure 6.1 Web API Start Page

ii. Test Case 1: When inputs are not filled according to the standard format, the Web API gives an error message as shown in the figure below.



Heart Disease Predict

Heart Disease Predict ML App

Age

0 − +

Sex

● Male
○ Female

Chest Pain Type

● Typical Angina
○ Atypical Angina
○ Non-Anginal Pain
○ Asymptomatic

Resting Blood Pressure

0 − +

Serum Cholestoral

0 − +

Fasting Blood Sugar > 120 mg/dl

● True
○ False

Resting Electrocardiographic Results

● Normal
○ ST-T Wave Abnormality
○ Probable or Definite Left Ventricular Hypertrophy

Maximum Heart Rate Achieved

0 − +

Exercise Induced Angina

● Yes
○ No

ST depression induced by exercise relative to rest

0.00 − +

Slope of the peak exercise ST segment

● Upsloping
○ Flat
○ Downsloping

Number of major vessels (0-3) colored by flourosopy

● 0
○ 1
○ 2
○ 3

Thal

● Normal
○ Fixed Defect
○ Reversable Defect

Predict

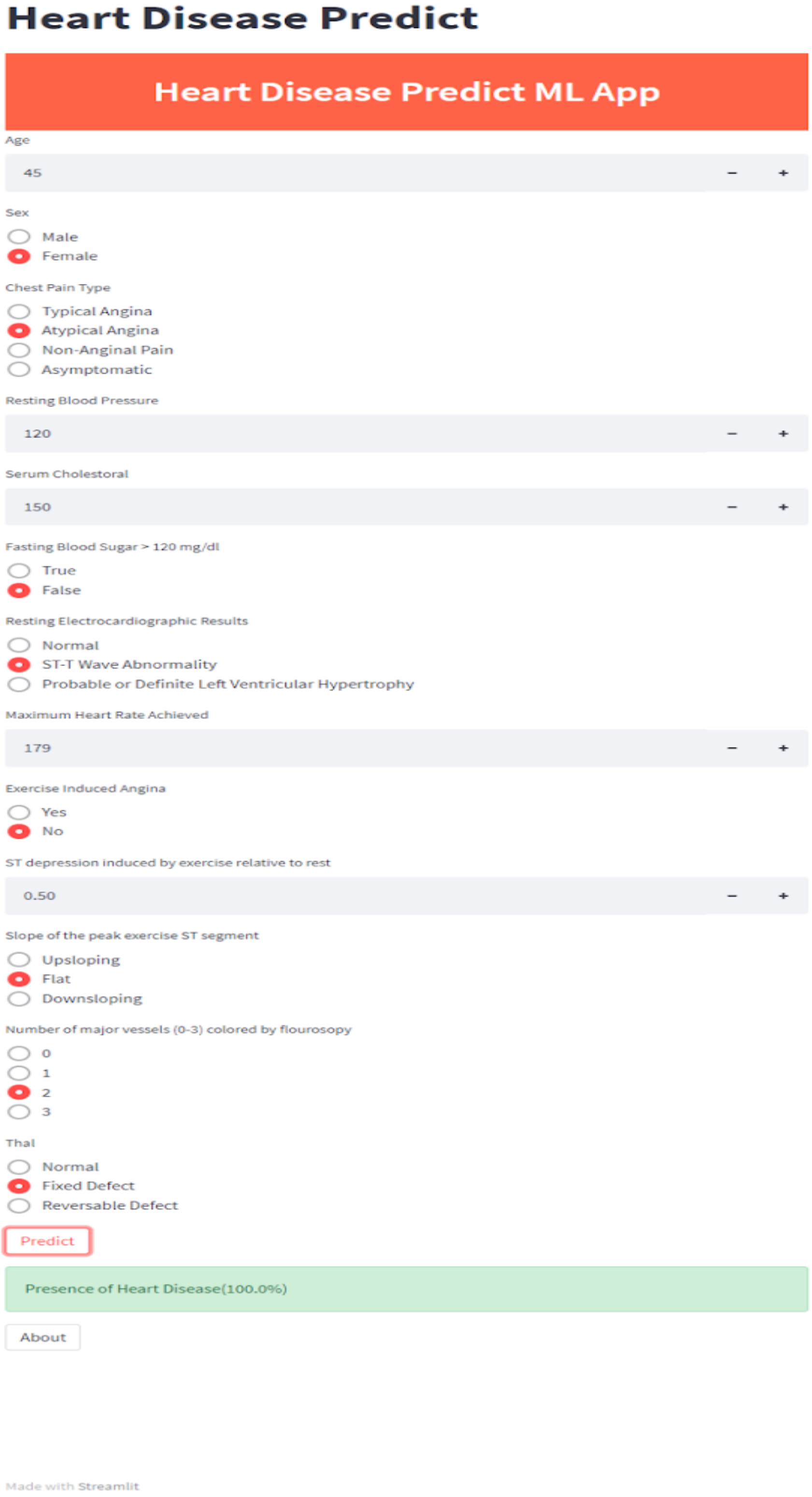
Input Value Error

About

Made with Streamlit

Figure 6.2 Error message for incorrect inputs

iii. Test Case 2: For a heart disease patient the output will be a confirmatory message as seen in the figure below.



Heart Disease Predict

Heart Disease Predict ML App

Age

45 − +

Sex

○ Male
● Female

Chest Pain Type

○ Typical Angina
● Atypical Angina
○ Non-Anginal Pain
○ Asymptomatic

Resting Blood Pressure

120 − +

Serum Cholestoral

150 − +

Fasting Blood Sugar > 120 mg/dl

○ True
● False

Resting Electrocardiographic Results

○ Normal
● ST-T Wave Abnormality
○ Probable or Definite Left Ventricular Hypertrophy

Maximum Heart Rate Achieved

179 − +

Exercise Induced Angina

○ Yes
● No

ST depression induced by exercise relative to rest

0.50 − +

Slope of the peak exercise ST segment

○ Upsloping
● Flat
○ Downsloping

Number of major vessels (0-3) colored by flourosopy

○ 0
○ 1
● 2
○ 3

Thal

○ Normal
● Fixed Defect
○ Reversable Defect

Predict

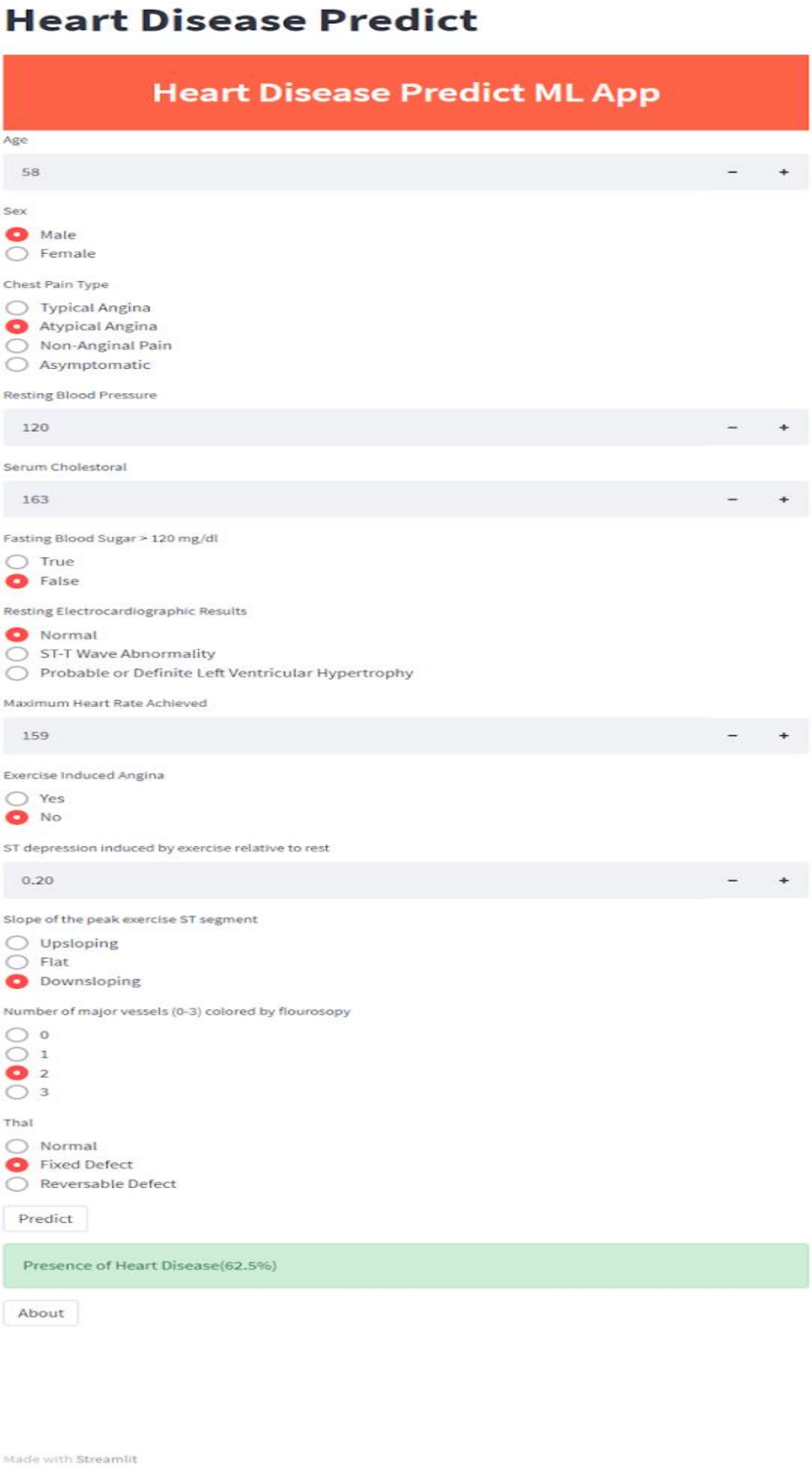
Presence of Heart Disease(100.0%)

About

Made with Streamlit

Figure 6.3 Presence of Heart Disease

iv. Test Case 3: For another heart disease patient the output will be a confirmatory message as seen in the figure below.



Heart Disease Predict

Heart Disease Predict ML App

Age

58 − +

Sex

● Male
○ Female

Chest Pain Type

○ Typical Angina
● Atypical Angina
○ Non-Anginal Pain
○ Asymptomatic

Resting Blood Pressure

120 − +

Serum Cholestoral

163 − +

Fasting Blood Sugar > 120 mg/dl

○ True
● False

Resting Electrocardiographic Results

● Normal
○ ST-T Wave Abnormality
○ Probable or Definite Left Ventricular Hypertrophy

Maximum Heart Rate Achieved

159 − +

Exercise Induced Angina

○ Yes
● No

ST depression induced by exercise relative to rest

0.20 − +

Slope of the peak exercise ST segment

○ Upsloping
○ Flat
● Downsloping

Number of major vessels (0-3) colored by flourosopy

○ 0
○ 1
● 2
○ 3

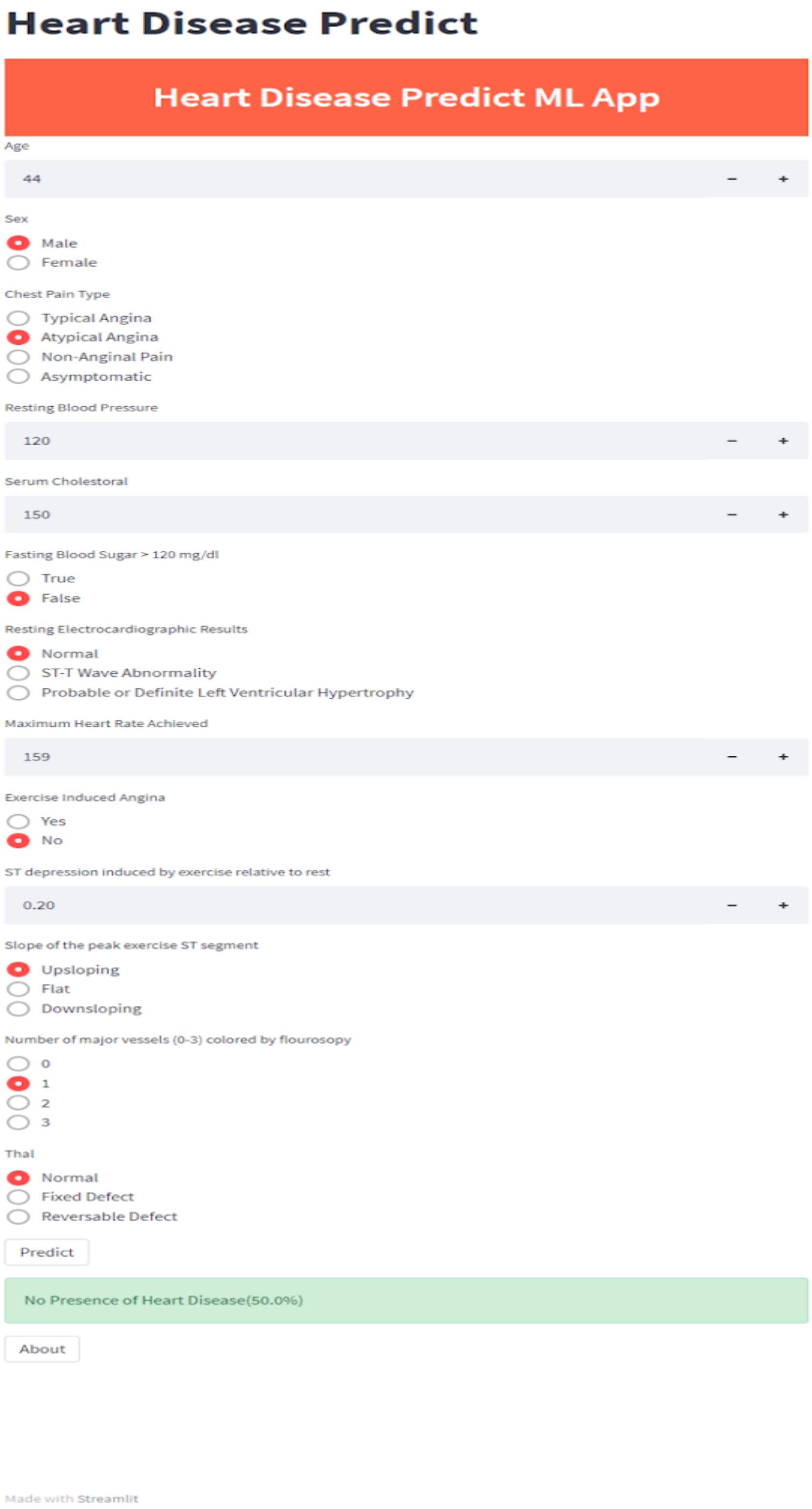
Thal

○ Normal
● Fixed Defect
○ Reversable Defect

Presence of Heart Disease(62.5%)

Figure 6.4 Presence of Heart Disease

v. Test Case 4: For a patient with non-heart disease condition the result is shown in the figure below.



Heart Disease Predict

Heart Disease Predict ML App

Age

44

Sex

● Male
○ Female

Chest Pain Type

○ Typical Angina
● Atypical Angina
○ Non-Anginal Pain
○ Asymptomatic

Resting Blood Pressure

120

Serum Cholestoral

150

Fasting Blood Sugar > 120 mg/dl

○ True
● False

Resting Electrocardiographic Results

● Normal
○ ST-T Wave Abnormality
○ Probable or Definite Left Ventricular Hypertrophy

Maximum Heart Rate Achieved

159

Exercise Induced Angina

○ Yes
● No

ST depression induced by exercise relative to rest

0.20

Slope of the peak exercise ST segment

● Upsloping
○ Flat
○ Downsloping

Number of major vessels (0-3) colored by flourosopy

○ 0
● 1
○ 2
○ 3

Thal

● Normal
○ Fixed Defect
○ Reversable Defect

No Presence of Heart Disease(50.0%)

Figure 6.5 No Presence of Heart Disease

CHAPTER 7: RESULTS

The project includes the necessary data pre-processing and examination of the client data for heart disease. The following maximum scores were obtained after training and testing four models:

- Random Forest Classifier (86.88%)
- K Nearest Neighbors Classifier (91.8%)
- Support Vector Classifier (90.16%)
- Logistic Regression (88.52%)

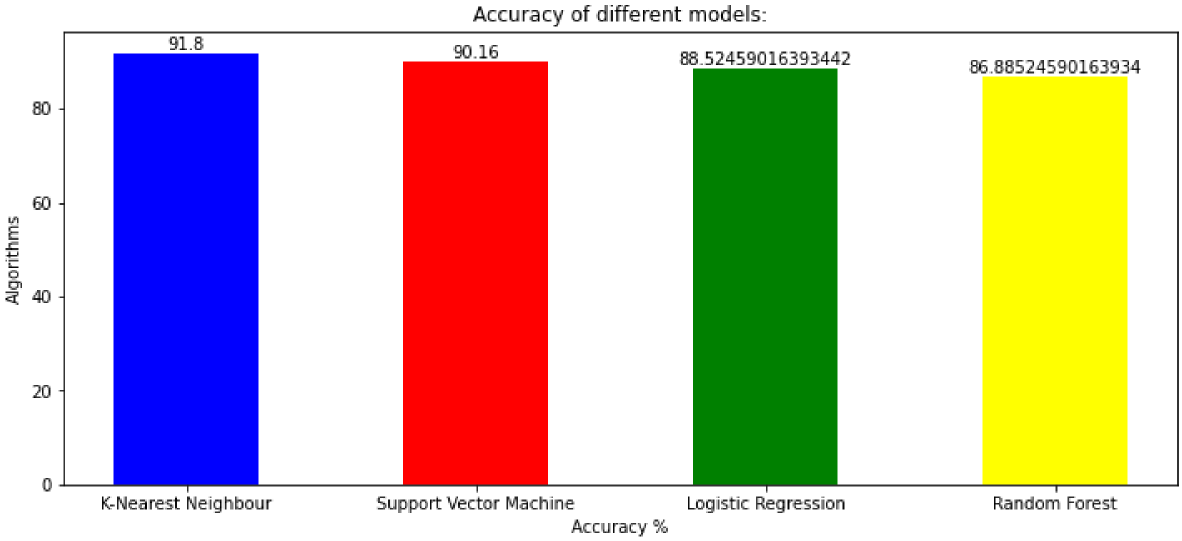


Figure 7.1 Test Accuracy for different algorithms achieved

Comparing the results with previous works:

Models\Results (%)	[13]	[14]	The proposed method
Logistic Regression	87.5	86.51	88.52
KNN	88.52	N/A	91.80
SVM	N/A	79.77	90.16
Random Forest	85.1	80.89	86.88

Table 7.1: Comparison of accuracies with base papers

It can be observed that among these models from all papers, K Nearest Neighbors Classifier performs the best.

The key factors from the dataset leading to heart diseases from the dataset are:

1. Exercise Induced Angina
2. Chest Pain
3. ST Depression during exercise in relation to the quantity of rest received
4. Reached maximum heart rate
5. Fluoroscopically coloured major vessels

The F1 score for the four models are:

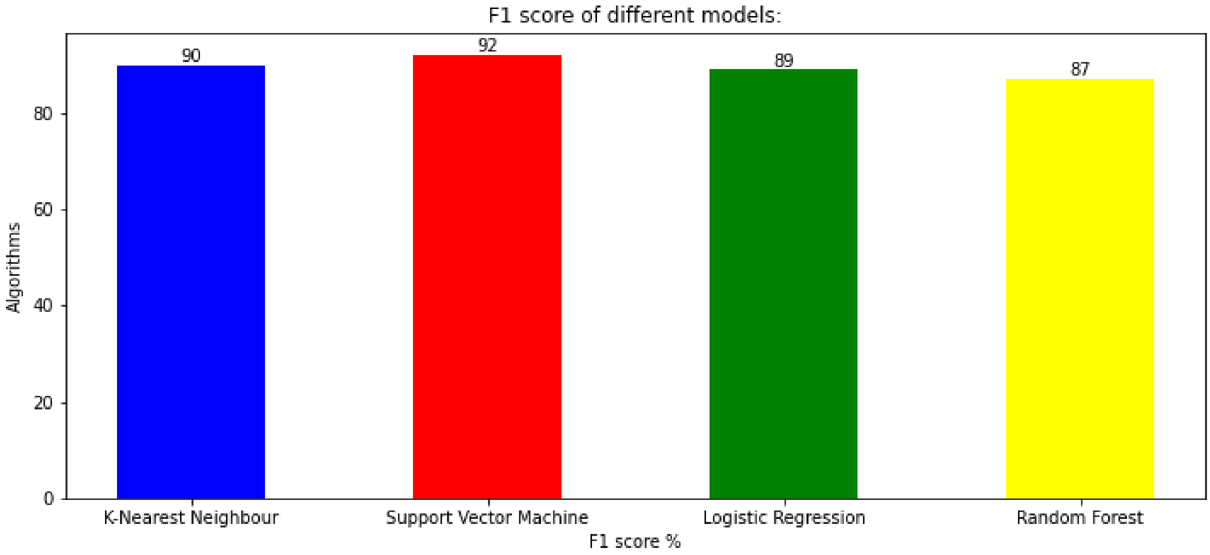


Figure 7.2 F1 score for different algorithms achieved

Even F1 score for the K-Nearest Neighbor classifier, for the given dataset is 90%.

The accuracy of the K-Nearest Neighbor classifier is discovered to be higher than other algorithms when using the machine learning technique for training and testing. The precision equation is applied to the number count of TP, TN, FP, and FN to get a value. It is concluded that the K-Nearest Neighbor classifier is the best with 91.80% accuracy. Hence for this model using the **K-Nearest Neighbor classifier** is preferred.

CHAPTER 8: IMPACT OF YOUR PROJECT TOWARDS SOCIETY/ ENVIRONMENT

Heart disease is one of many illnesses that can be fatal, and it has received considerable attention in medical studies. Heart disease diagnosis is a difficult process that can provide automated predictions about a patient's heart state to improve the effectiveness of subsequent treatment. Heart disease is typically diagnosed based on the patient's physical examination, signs, and symptoms. The risk of heart disease is influenced by a number of variables, including smoking, obesity, family history, body cholesterol, high blood pressure and inactivity.

The delivery of better services at reasonable prices is a significant problem for healthcare institutions, like hospitals and health centres. The provision of high-quality care necessitates accurate patient diagnosis and efficient treatment delivery.

To overcome these issues this project proposes an affordable solution to the situation where each and every person with the internet can access this webpage for knowing their heart condition and can decide whether to refer to a specialist or not. With this technology each and every one will have basic access to medical facility which miserably lacks in developing countries like India.

CHAPTER 9: CONCLUSIONS

In India as well as the rest of the globe, heart disease is the main cause of death in human beings. The application of innovative technology, such as machine learning as an advanced cardiac disease predictor, will have a big impact on society. Recommendations on lifestyle for high-risk individuals can be made with the use of early heart disease predictors, which will help to prevent complications, which could mark a significant turning point in medical history. Every year, there is an increase in the number of people who are developing heart related diseases. The use of appropriate technology help in this regard may be highly beneficial to patients and other medical professionals. Logistic Regression, Random Forest, SVM and K-Nearest Neighbor are four separate machines whose learning methods are employed in this project to evaluate performance.

The dataset with 76 features contains expected characteristics that cause heart disease in people. When all 76 features are considered, the system's efficiency is less as most of the features are irrelevant. Attribute selection is used to improve efficiency. 14 appropriate features are selected to evaluate a model that provides more accuracy because linking certain features to the dataset is almost similar and they are removed. These selected features are further preprocessed to increase the overall productivity and allow the machine learning algorithms for high quality information while making decisions.

All four machine learning algorithms are assessed based on their accuracies and out of which one is selected. In order to evaluate all of the algorithms, it is expected to use numerous assessment metrics, including the confusion matrix, accuracy, precision, f1-score, recall and select one which predicts the disease efficiently. When these four are compared, the K-Nearest Neighbor model has the most accuracy of 91.80.%. Therefore, for predicting heart disease, the K-Nearest Neighbor classifier is used in the model.

CHAPTER 10: SELF ASSESSMENT OF PO-PSO ATTAINMENT

PROGRAM OUTCOMES (PO)	JUSTIFICATION
PO1. Engineering Knowledge:	We have been able to apply the knowledge of science, mathematics and engineering like machine learning, cloud, visualization and problem solving.
PO2. Problem Analysis:	People face this problem, where they are unable to access affordable basic medical facilities.
PO3. Design Development of solutions:	To build a simple and easy to use Heart Disease Prediction System for early prognosis of heart disease.
PO5. Modern Tools:	Google Collab, Kaggle, Jupyter notebook,
PO6. The Engineer and society:	We learnt how to handle a project professionally, maintaining safety and social responsibilities.
PO8. Ethics:	We were able to handle responsibilities and norms of ethical engineering practices and applied the same in the project.
PO9. Individual and Teamwork:	We performed our individual tasks and were able to collaborate as a team better.
PO10. Communication:	We were able to communicate and present our ideas with each other effectively. This is what made it possible for us to collaborate our work together and build a project.

PO11: Project management and finance:	We used some open-source technologies and some resources that were for students at minimum charges. The team members were supportive and managed various disciplines in the project
PO12. Life-long learning:	We will be able to apply what we have learnt from this project in other real-life scenarios as well.

PROGRAM SPECIFIC OUTCOMES (PSO)	**JUSTIFICATION**
PSO1. Professional Skills:	We have been able to apply the knowledge of machine learning, visualization and problem-solving ability to various tasks including data cleaning, training, evaluating and analysing. Outcomes and results were presented proficiently.
PSO2. Problem Solving Skills:	We were able to change our tools and technologies when required in the project. We shifted to better and newly released versions of some technologies and were able to make changes in the project accordingly.
PSO3. Ethics and career development:	We were able to handle responsibilities and norms of ethical engineering practices and applied the same in the project. There was effective communication between the team and evaluators and among the team members as well. We took their feedback and incorporated them in the project.

REFERENCES

[1] Soni J, Ansari U, Sharma D & Soni S (2011). Predictive data mining for medical diagnosis: an overview of heart disease prediction. International Journal of Computer Applications, 43-8.

[2] Dangare C S & Apte S S (2012). Improved study of heart disease prediction system using data mining classification techniques. International Journal of Computer Applications, 44-8.

[3] Ordonez C (2006). Association rule discovery with the train and test approach for heart disease prediction. IEEE Transactions on Information Technology in Biomedicine, 334- 43.

[4] Shinde R, Arjun S, Patil P & Waghmare J (2015). An intelligent heart disease prediction system using k-means clustering and Naïve Bayes algorithm. International Journal of Computer Science and Information Technologies, 6(1), 637-9.

[5] Bashir S, Qamar U & Javed M Y (2014, November). An ensemble-based decision support framework for intelligent heart disease diagnosis. In International Conference on Information Society (i-Society 2014) (pp. 259-64). IEEE.

[6] Jee S H, Jang Y, Oh D J, Oh B H, Lee S H, Park S W & Yun Y D (2014). A coronary heart disease prediction model: the Korean Heart Study. BMJ Open, 4(5), e005025.

[7] Ganna A, Magnusson P K, Pedersen N L, de Faire U, Reilly M, Ärnlöv J & Ingelsson E (2013). Multilocus genetic risk scores for coronary heart disease prediction. Arteriosclerosis, thrombosis, and vascular biology, 33(9), 2267-72.

[8] Jabbar M A, Deekshatulu B L & Chandra P (2013, March). Heart disease prediction using lazy associative classification. In 2013 International Multi-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s) (pp. 40- 6). IEEE.

[9] Dangare Chaitrali S and Sulabha S Apte. "Improved study of heart disease prediction system using data mining classification techniques." International Journal of Computer Applications 47.10 (2012): 44-8.

[10] Soni Jyoti. "Predictive data mining for medical diagnosis: An overview of heart disease prediction." International Journal of Computer Applications 17.8 (2011): 43-8.

[11] Chen A H, Huang S Y, Hong P S, Cheng C H & Lin E J (2011, September). HDPS: Heart disease prediction system. In 2011 Computing in Cardiology (pp. 557-60). IEEE.

[12] Parthiban, Latha and R Subramanian. "Intelligent heart disease prediction system using CAN FIS and genetic algorithm." International Journal of Biological, Biomedical and Medical Sciences 3.3 (2008).

[13] Jindal H, Agrawal S, Khera R, Jain R & Nagrath P (2020). Heart disease prediction using machine learning algorithms. 1st International Conference on Computational Research and Data Analytics (ICCRDA 2020).

[14] Kumar G Dinesh; K Arumugaraj; Kumar D Santhosh; V Mareeswari (2018). Prediction of Cardiovascular Disease Using Machine Learning Algorithms. 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT).

[15] Mohan K S; Thirumalai C; Srivastava G (2019). Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. IEEE Access (Volume: 7).

[16] Ghosh P; Azam S; Jonkman J M; Karim A; Samrat F. M. J M; Ignatius E (2021).Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms with Relief and LASSO Feature Selection Techniques. IEEE Access (Volume: 9).

[17] Joshi R; Peng Z;Long X; Feijs L; Andriessen P (2019). Predictive Monitoring of Critical Cardiorespiratory Alarms in Neonates Under Intensive Care.IEEE Journal of Translational Engineering in Health and Medicine (Volume: 7).

[18] Chen J; Valehi A; Razi A (2019). Smart Heart Monitoring: Early Prediction of Heart Problems Through Predictive Analysis of ECG Signals. IEEE Access (Volume: 7).

plagiarism_18PCS_36_.pdf

by KINSHUK CHATURVEDI

Submission date: 11-Jul-2022 11:23AM (UTC+0530)
Submission ID: 1869073085
File name: 1744_KINSHUK_CHATURVEDI_plagiarism_18PCS_36__4001_753773691.pdf (1.27M)
Word count: 11735
Character count: 62362

plagiarism_18PCS_36_.pdf

ORIGINALITY REPORT

8%	6%	6%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	towardsdatascience.com Internet Source	1%
2	ijariie.com Internet Source	<1%
3	paper.ijcsns.org Internet Source	<1%
4	researchr.org Internet Source	<1%
5	M Sowmiya, Rekha B Banu, A Dhaksana, A Priyadharshini. "An Integrated Vessel Segmentation and Machine Learning approach for abnormal vasculature detection in Retinal Images", 2021 IEEE Madras Section Conference (MASCON), 2021 Publication	<1%
6	Senthilkumar Mohan, Chandrasegar Thirumalai, Gautam Srivastava. "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques", IEEE Access, 2019 Publication	<1%

7	nasdag.github.io Internet Source	<1 %
8	www.hindawi.com Internet Source	<1 %
9	Dinesh Reddy Vemula, Mahesh Kumar Morampudi, Sonam Maurya, Ashu Abdul, Md. Muzakkir Hussain, Ilaiah Kavati. "Enhanced resource provisioning and migrating virtual machines in heterogeneous cloud data center", Journal of Ambient Intelligence and Humanized Computing, 2022 Publication	<1 %
10	www.wovo.org Internet Source	<1 %
11	Pronab Ghosh, Sami Azam, Mirjam Jonkman, Asif Karim et al. "Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms with Relief and LASSO Feature Selection Techniques", IEEE Access, 2021 Publication	<1 %
12	Ana Koren, Marko Jurčević, Ramjee Prasad. "Comparison of Data-Driven Models for Cleaning eHealth Sensor Data: Use Case on ECG Signal", Wireless Personal Communications, 2020 Publication	<1 %

13	"Advanced Machine Learning Technologies and Applications", Springer Science and Business Media LLC, 2021 Publication	<1 %
14	doaj.org Internet Source	<1 %
15	www.pubfacts.com Internet Source	<1 %
16	prms.ase.ro Internet Source	<1 %
17	des-ouwe.fun Internet Source	<1 %
18	hdl.handle.net Internet Source	<1 %
19	pracoval-szeretem.com Internet Source	<1 %
20	Bipasa Mukherjee, Shreya Priyadarshini Roy, Vergin Raja Sarobin. "Application of Machine Learning Algorithm for Cardiovascular Disease Detection", 2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2021 Publication	<1 %
21	Cao Truong Tran, Mengjie Zhang, Peter Andreae, Bing Xue, Lam Thu Bui. "An effective and efficient approach to classification with	<1 %

incomplete data", Knowledge-Based Systems, 2018
Publication

22 "ICT Systems and Sustainability", Springer Science and Business Media LLC, 2022
Publication <1%

23 Mohammed Nasir Uddin, Rajib Kumar Halder. "An ensemble method based multilayer dynamic system to predict cardiovascular disease using machine learning approach", Informatics in Medicine Unlocked, 2021
Publication <1%

24 Priyanka Padhiyar, Rashmin Prajapati. "A Review On Anomalous Movement Fraudulent Detection in Exam Hall", 2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS), 2022
Publication <1%

25 FARMAN HASSAN, Auliya Ur Rahman, Ali Javed, Ali Alhazmi, Majed Alhazmi. "CNN-CardioAssistant: Deep Convolutional Neural Network and Recursive Feature Elimination Method for Heart Disease Detection", Research Square Platform LLC, 2022
Publication <1%

26 Rony Chowdhury Ripan, Md. Moinul Islam, Hamed Alqahtani, Iqbal H. Sarker. "Effectively predicting cyber ‐ attacks through isolation <1%

	forest learning - based outlier detection", SECURITY AND PRIVACY, 2022 Publication	
27	www.ijirset.com Internet Source	<1 %
28	ijarcce.com Internet Source	<1 %
29	ieeexplore.ieee.org Internet Source	<1 %
30	Www.tutorialspoint.com Internet Source	<1 %
31	scholar.ppu.edu Internet Source	<1 %
32	www.mdpi.com Internet Source	<1 %
33	yes-great.com Internet Source	<1 %
34	Archana Singh, Rakesh Kumar. "Heart Disease Prediction Using Machine Learning Algorithms", 2020 International Conference on Electrical and Electronics Engineering (ICE3), 2020 Publication	<1 %
35	eprints.qut.edu.au Internet Source	<1 %

36	ijisrt.com Internet Source	<1 %
37	mldata.org Internet Source	<1 %
38	studentsrepo.um.edu.my Internet Source	<1 %
39	www.ijraset.com Internet Source	<1 %
40	"Computational Intelligence in Data Mining—Volume 2", Springer Science and Business Media LLC, 2016 Publication	<1 %
41	Günther Grabner, Andrew L. Janke, Marc M. Budge, David Smith, Jens Pruessner, D. Louis Collins. "Chapter 8 Symmetric Atlasing and Model Based Segmentation: An Application to the Hippocampus in Older Adults", Springer Science and Business Media LLC, 2006 Publication	<1 %
42	www.coursehero.com Internet Source	<1 %
43	"Soft Computing in Industrial Applications", Springer Science and Business Media LLC, 2011 Publication	<1 %

44	Butch Quinto. "Next-Generation Big Data", Springer Science and Business Media LLC, 2018 Publication	<1 %
45	Fathania Firwan Firdaus, Hanung Adi Nugroho, Indah Soesanti. "Deep Neural Network with Hyperparameter Tuning for Detection of Heart Disease", 2021 IEEE Asia Pacific Conference on Wireless and Mobile (APWiMob), 2021 Publication	<1 %
46	S.K.L. Sameer, P. Sriramya. "Improving the Accuracy for Prediction of Heart Disease by Novel Feature Selection Scheme using Decision tree comparing with Naive-Bayes Classifier Algorithms", 2022 International Conference on Business Analytics for Technology and Security (ICBATS), 2022 Publication	<1 %
47	certrofisio.com.br Internet Source	<1 %
48	github.com Internet Source	<1 %
49	medium.com Internet Source	<1 %
50	tojqi.net Internet Source	<1 %

51 Ebrahim Mohammed Senan, Ibrahim Abunadi, Mukti E. Jadhav, Suliman Mohamed Fati. "Score and Correlation Coefficient-Based Feature Selection for Predicting Heart Failure Diagnosis by Using Machine Learning Algorithms", Computational and Mathematical Methods in Medicine, 2021 <1%

Publication

52 Rohan Joshi, Zheng Peng, Xi Long, Loe Feijs, Peter Andriessen, Carola Van Pul. "Predictive Monitoring of Critical Cardiorespiratory Alarms in Neonates Under Intensive Care", IEEE Journal of Translational Engineering in Health and Medicine, 2019 <1%

Publication

53 Jiaming Chen, Ali Valehi, Abolfazl Razi. "Smart Heart Monitoring: Early Prediction of Heart Problems Through Predictive Analysis of ECG Signals", IEEE Access, 2019 <1%

Publication

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		