DOCTOR WORKLOAD MANAGEMENT SYSTEM WITH PREDICTING THE POSSIBILITY OF BECOMING HEART PATIENT

Jayasekara J.M.P.N.K

IT20623418

B.Sc. (Hons) Degree in Information Technology Specializing in Software Engineering

Department of Information Technology

Sri Lanka

April 2024

DOCTOR WORKLOAD MANAGEMENT SYSTEM WITH PREDICTING THE POSSIBILITY OF BECOMING HEART PATIENT

Jayasekara J.M.P.N.K

IT20623418

Supervisor - Dr. Kapila Dissanayaka

Co – Supervisor - Ms Bhagyani Chathurika

External Supervisor - Dr Susith Athukorala

B.Sc. (Hons) Degree in Information Technology Specializing in Software Engineering

Department of Information Technology

Sri Lanka

April 2024

DECLARATION PAGE OF THE CANDIDATES & SUPERVISOR

I hereby affirm that this proposal represents my original work. It does not incorporate, without appropriate acknowledgment, any material previously submitted for academic credit towards a degree or diploma at any other university or institution of higher learning. Furthermore, to the best of my knowledge and belief, it does not contain any material previously published or authored by another individual, except where explicit acknowledgment is provided within the text.

Group Member Name	Student ID	Signature
Jayasekara J.M.P.N. K	**IT20623418**	Parami

The aforementioned candidate is currently engaged in research for their undergraduate dissertation under my guidance and supervision.

Signature of the Supervisor: Date: 2024.03.21

(Dr. Kapila Dissanayaka)

ABSTRACT

This research aims to enhance cardiovascular risk assessment through the development of a predictive model utilizing various demographic and physiological factors. Our analysis includes inputs such as age, sex, chest pain type, resting blood pressure, cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, previous peak, slope, number of major vessels, Thalium Stress Test results, and exercise-induced angina. We identified key predictors of heart attack risk, noting that individuals with higher maximum heart rates and lower previous peaks are more susceptible. Logistic regression emerged as the most effective algorithm, achieving an accuracy of 92.97%.

Our findings highlight the complex interplay of factors influencing heart attack risk and the importance of multifactorial assessment in preventive healthcare. The developed model, accessible via mobile application, provides personalized risk assessments while maintaining data confidentiality. This research underscores the potential of predictive analytics to improve preventive strategies and patient outcomes in cardiovascular care.

ACKNOWLEDGEMENT

I extend my sincere gratitude to Dr. Kapila Dissanayaka for his invaluable guidance, steadfast support, and insightful supervision throughout all stages of this research project. His expertise and encouragement have significantly contributed to shaping the trajectory and outcomes of this study.

I am equally indebted to Dr. Bhagayani Chathurika for her indispensable co-supervision, constructive feedback, and continuous encouragement, which significantly enriched the quality of this work.

Special appreciation is extended to Dr. Susith Athukorala of Apeksha Hospital, Dr. Subashini of Ragama Hospital, and Nursing Sister Chandrani Kumari of Monaragala Sirigala Hospital for their generous support and cooperation from the hospital side. Their expertise, assistance, and willingness to facilitate access to resources have been invaluable to the success of this research.

Furthermore, I extend my gratitude to all the members of our research group who tirelessly contributed to various aspects of this project, from data collection to analysis and interpretation. Your dedication and collaborative spirit have been instrumental in overcoming challenges and achieving our research objectives.

Finally, I wish to extend my sincerest gratitude to my family and friends for their steadfast encouragement, understanding, and patience throughout this journey. The completion of this research owes much to the collaborative efforts, support, and encouragement of all those previously mentioned, and for this, I am profoundly thankful.

Table of Contents

List of Figures

List of tables

LIST OF ABBREVIATIONS

Abbreviation	Description
CVD	Cardiovascular diseases
SDLC	Software Development Life Cycle

Table 1 : List of Abbreviations

01. INTRODUCTION

Heart disease, a prevalent global health issue, stands as a significant cause of mortality and morbidity worldwide. Despite advancements in medical science, the burden of cardiovascular diseases (CVDs) continues to escalate, with dire consequences for individuals and economies alike. In light of this, leveraging sophisticated techniques such as machine learning becomes imperative to not only predict but also mitigate the risk associated with heart disease. This research endeavor aims to contribute to this critical domain by developing a logistic regression model to predict heart attack probability and stratify these probabilities into distinct risk levels using color-coded categorization. Additionally, the study seeks to offer tailored instructions based on these risk levels, aligning with WHO National Guidelines to provide comprehensive guidance for individuals at varying risk levels [1].

1.1 Background

Globally, cardiovascular disease (CVD) represents a predominant cause of mortality, responsible for over 70% of all global deaths. Despite its prevalence, a significant portion of these fatalities could be prevented through early detection and intervention. Unfortunately, the economic burden associated with traditional diagnostic methods such as electrocardiograms and CT scans often renders them inaccessible, particularly in low- and middle-income countries. Consequently, there is an urgent need for cost-effective and scalable approaches to predict and manage heart disease risk.

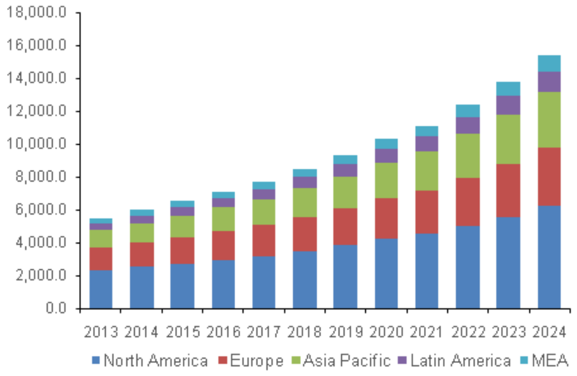


Figure 1 : The impact of heart attacks in the world until 2024

Accordingly, by 2024, a report including the number of deaths due to cardiovascular diseases and the number of patients is shown in the table below based on the association's 2024 heart disease and Stroke Statistics Update [2].

Category	Statistics
1. Heart disease & Stroke	
I. CVD Deaths	931,578 (2021)
II. Leading Cause of Death	Heart disease & stroke
III. CVD Prevalence	127.9 million US adults (48.6%)
IV. CVD Costs	$422.3 billion (2019-2020)
V. Non-Hispanic Black CVD	59.0% (females), 58.9% (males)
VI. CHD Leading Cause	40.3% (CVD deaths)
VII. CVD Health Expenditures	12% of total US health expenditures
VIII. Global CVD Deaths	19.91 million (2021)
2. Coronary Heart Disease	
I. CHD Deaths	375,476 (2021)
II. Heart Attack Incidence	605,000 new attacks annually (US)
III. Average Age at First	Heart Attack: 65.6 years (males), 72.0 years (females)
IV. CHD Death Rate Decline	15.0% (2011-2021)
V. Cost of Heart Disease	$252.2 billion (2019-2020)
3. Stroke	
I. Stroke Deaths	162,890 (2021)
II. Stroke Mortality Rate	41.1 per 100,000 (2021)
III. Stroke Deaths Increase	26.3% (2011-2021)
IV. Global Stroke Deaths	7.44 million (2021)
4. Sudden Cardiac Arrest	
Sudden Cardiac Deaths	20,114 (2021)
Location of OHCA	Home/residence (72.1%), Public settings (17.3%), Nursing homes (10.6%)

Survival Rate (EMS)	9.3% (all OHCA), 14.0% (bystander witnessed), 17.0% (9-1-1 responder witnessed)

Table 2 : 2024 Cardiovascular Health Statistics Overview

Data mining and machine learning methodologies present viable avenues for addressing the complexities inherent in medical data analysis. These techniques facilitate the discernment of latent patterns within extensive datasets, offering promising prospects for predictive modeling in the realm of heart disease. Existing literature has demonstrated the efficacy of such approaches, with reported accuracies reaching as high as 94%. Nevertheless, persistent challenges, notably the constraints imposed by small sample sizes and the specter of overfitting, temper the broad applicability of these findings. Consequently, there exists a compelling imperative for continued research endeavors aimed at mitigating these constraints, thereby fostering the refinement and robustness of predictive models tailored to the domain of heart disease.

1.2 Literature Review

The healthcare sector has experienced notable advancements in data mining and machine learning, particularly within the field of medical cardiology. Given that heart disease continues to rank as a prominent cause of mortality, considerable research efforts have been directed towards enhancing early detection and prevention strategies. Prior investigations have showcased the efficacy of machine learning methodologies in prognosticating cardiovascular diseases (CVDs).

I. **Narain et al. (2016):** Developed a machine-learning-based CVD prediction system to enhance the accuracy of the Framingham risk score. Achieved a high accuracy of 98.57% in forecasting CVD risk, outperforming traditional methods. The proposed system could assist doctors in better treatment planning and early diagnosis.

II. **Alotalibi (2019):** Investigated the utility of machine learning techniques for predicting heart failure disease. Decision tree algorithm achieved an accuracy of 93.19%, showcasing the potential of ML techniques in predicting heart failure disease.

III. **Hasan and Bao (2020):** Compared various algorithms for feature selection in anticipating cardiovascular illness. XGBoost classifier coupled with the wrapper technique provided the most accurate prediction results at 73.74%. Highlighted the importance of feature selection methods in improving prediction accuracy.

IV. **Shah et al. (2020):** Aimed to predict cardiovascular disease using machine learning techniques. Achieved a high accuracy of 90.8% with the KKN model. Highlighted the importance of selecting appropriate models for optimal results in disease prediction.

V. **Drod et al. (2022):** Identified significant risk variables for CVD in patients with metabolic-associated fatty liver disease (MAFLD) using machine learning techniques. Achieved an accuracy of 85.11% in identifying high-risk patients and 79.17% in identifying low-risk patients. Emphasized the utility of ML methods in detecting CVD in MAFLD patients.

Previous research has shown promise in using machine learning for predicting heart diseases. However, limitations such as small datasets and the risk of overfitting have been observed. The current study addresses these limitations by utilizing a large dataset of 70,000 patients and 11 features, reducing the risk of overfitting.

The studies presented highlight the growing interest and success in utilizing machine learning techniques for predicting cardiovascular diseases. From enhancing the accuracy of existing risk assessment tools like the Framingham risk score to identifying significant risk factors in specific patient populations like MAFLD patients, each study contributes to the evolving landscape of cardiac care.

However, these studies also point out some challenges and opportunities for future research. One of the key challenges is the limited dataset size, which can lead to overfitting and may not be representative of the broader population. To address this, larger datasets like the one used in the current study are essential for robust model development and validation.

Another important aspect is the selection of appropriate machine learning algorithms and feature selection methods. While some studies have achieved high accuracies using specific algorithms like KKN or decision trees, others have found success with ensemble methods like XGBoost. Additionally, feature selection plays a crucial role in improving model performance, as demonstrated by Hasan and Bao (2020).

The collective findings underscore the potential of machine learning in revolutionizing cardiac care by enabling early detection, personalized risk assessment, and targeted interventions. By addressing challenges such as dataset size and algorithm selection, future research can further advance the field and ultimately improve patient outcomes.

Study	**Objective**	**Key Findings**	**Outputs**
Narain et al.	Enhance accuracy of Framingham risk score for CVD prediction	Achieved 98.57% accuracy, surpassing traditional methods. Suggested system aids in better treatment planning.	98.57% accuracy in forecasting CVD risk
Alotalibi	Investigate ML techniques for predicting heart failure disease	Decision tree algorithm achieved 93.19% accuracy. Demonstrated potential of ML in predicting heart failure.	85.11% accuracy in identifying high-risk patients
Hasan and Bao	Compare feature selection methods for anticipating cardiovascular illness	XGBoost classifier with wrapper technique provided 73.74% accuracy. Importance of feature selection highlighted.	73.74% accuracy with the XGBoost classifier
Shah et al.	Predict CVD using machine learning techniques	KKN model achieved 90.8% accuracy. Importance of model selection highlighted for optimal prediction results.	90.8% accuracy with the KKN model

Drod et al.	Identify significant risk variables for CVD in MAFLD patients using ML techniques	Achieved 85.11% accuracy in identifying high-risk patients. Emphasized utility of ML in detecting CVD in MAFLD.	85.11% accuracy in identifying high-risk patients

Table 3 : Previous Investigations

1.3 Research Gap

While existing literature has explored various machine learning algorithms for predicting heart disease, there remains a notable research gap in the stratification of predicted probabilities into actionable risk levels. Such stratification is crucial for providing personalized guidance to individuals based on their unique risk profiles. Additionally, prior studies have primarily focused on the development and validation of predictive models without delving into the practical implications of risk assessment and mitigation strategies. Addressing this gap necessitates the development of comprehensive frameworks that not only predict heart attack probability but also offer actionable insights tailored to individuals' risk levels.

The following table presents a comparative analysis between my proposed research idea and prior investigations pertaining to the prediction of heart attacks.

Study	**Final Prediction**	**The prediction was done using a mobile app**	**Risk Levels**	**Risk Levels (According to WHO Guidelines)**	**Risk Levels Colors (According to WHO Guidelines)**	**Details and Instructions According to Risk Levels**
Narain et al.	✓	✗	✓	✗	✗	✗
Alotalibi	✓	✗	✓	✗	✗	✗
Hasan and Bao	✓	✗	✓	✗	✗	✗

Shah et al.	✓	✗	✓	✗	✗	✗
Drod et al.	✓	✗	✓	✗	✗	✗

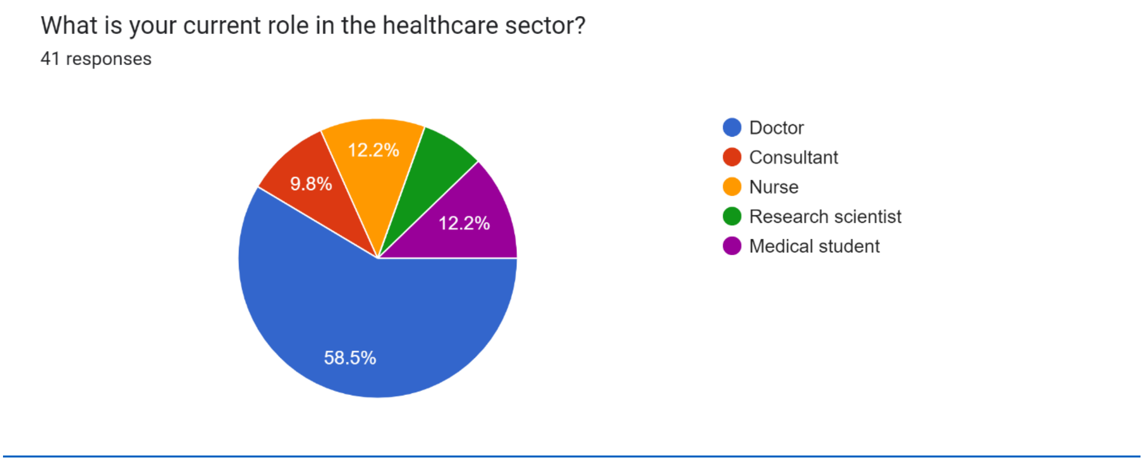
Table 4 : Comparison with previous Investigations

1.4 Research Problem

The research idea for detecting heart attacks before they occur with high accuracy and providing clear instructions to patients involves a multi-faceted approach integrating medical expertise, machine learning, and user interface design. Initially, a comprehensive understanding of the challenges faced in diagnosing heart attacks was gained through surveys distributed among medical professionals at Apeksha Hospital, Ragama Hospital, and Monaragala Sirigala Hospital. These surveys revealed a reliance on either machine diagnostics or the subjective judgment of doctors, often leading to delayed diagnoses. Below are the results of the survey conducted at the hospital I mentioned earlier.

Google form link: Machine Learning in Predicting Heart Attack Risks & Optimizing Doctor Workloads

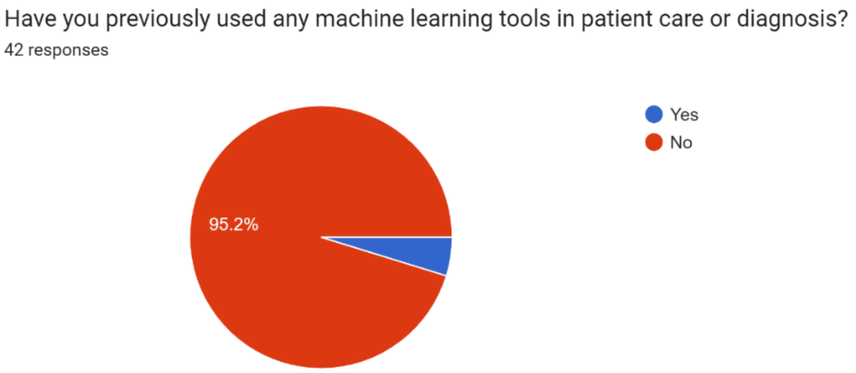
https://docs.google.com/forms/d/e/1FAIpQLSe5f3dY14AQli49MnGOMH1HofcOl4Syy_rYtlSe5139tdTChQ/viewform?usp=sf_link



What is your current role in the healthcare sector?

41 responses

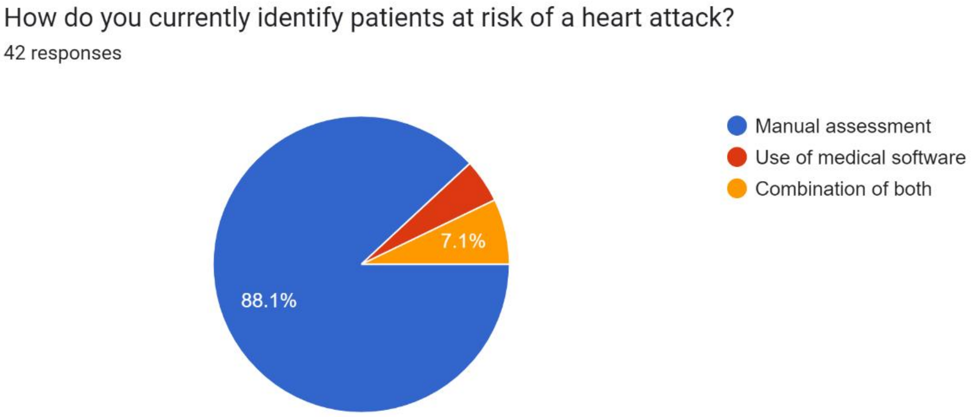
Figure 2 : Google form Question 01



Have you previously used any machine learning tools in patient care or diagnosis?

42 responses

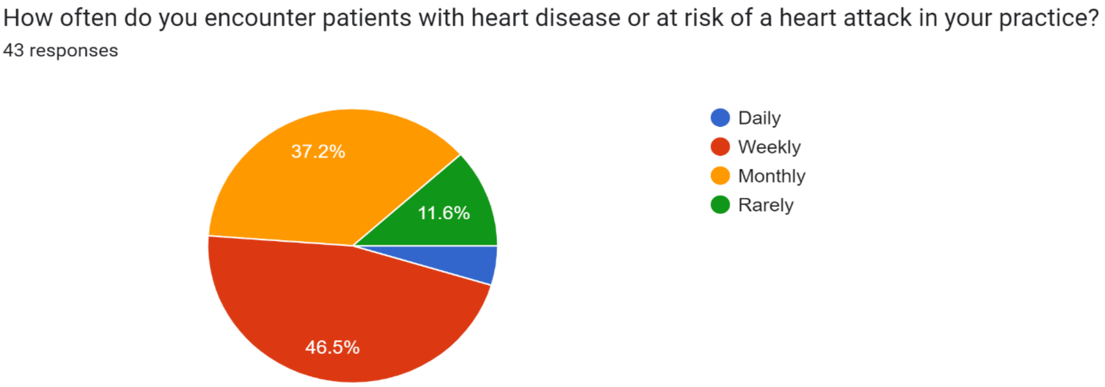
Figure 3 : Google form Question 02



How do you currently identify patients at risk of a heart attack?

42 responses

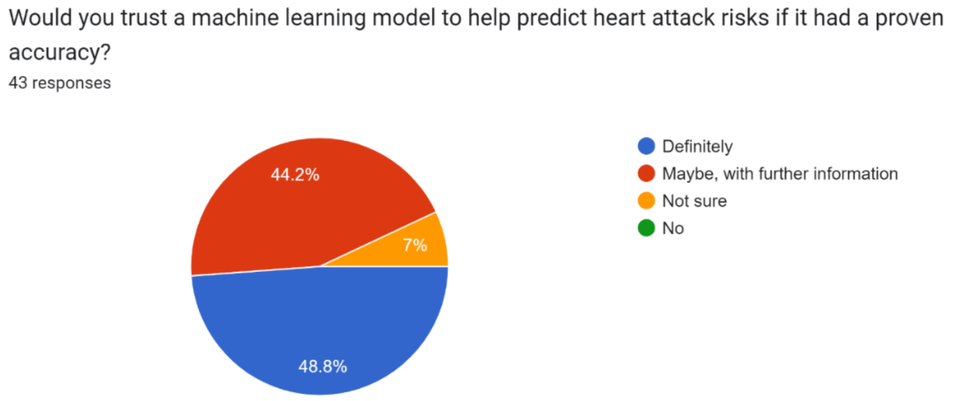
Figure 3 : Google form Question 03



How often do you encounter patients with heart disease or at risk of a heart attack in your practice?

43 responses

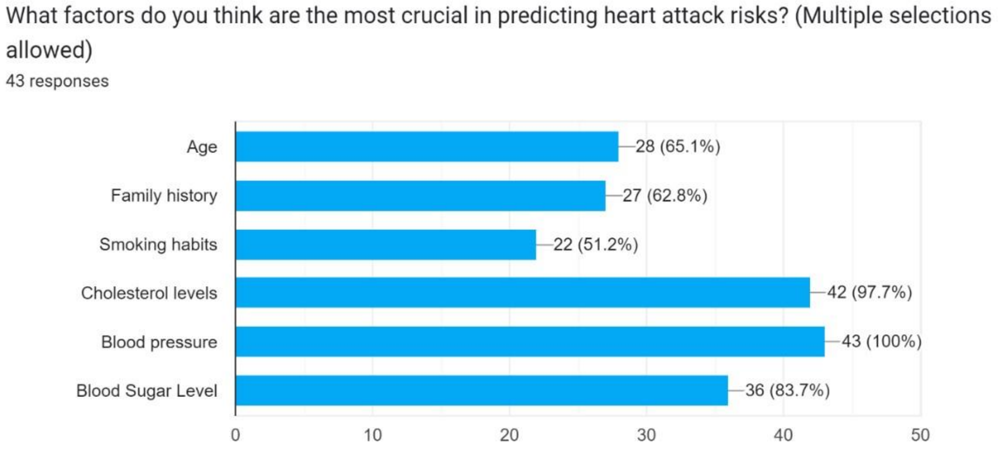
Figure 4 : Google form Question 03



Would you trust a machine learning model to help predict heart attack risks if it had a proven accuracy?

43 responses

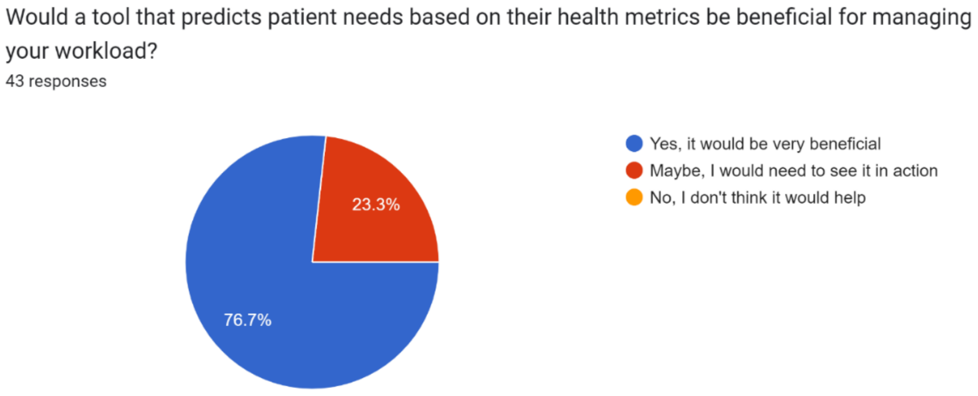
Figure 5 : Google form Question 04



What factors do you think are the most crucial in predicting heart attack risks? (Multiple selections allowed)

43 responses

Figure 6 : Google form Question 05



Would a tool that predicts patient needs based on their health metrics be beneficial for managing your workload?

43 responses

Figure 7 :Google form Question 06

The overarching research problem revolves around the need to enhance the effectiveness of heart disease prediction and management through the integration of machine learning techniques with actionable risk stratification strategies. Specifically, the lack of personalized guidance based on individual risk levels poses a significant challenge in mitigating the burden of heart disease. Additionally, the absence of standardized protocols for translating predictive models' outputs into actionable recommendations further compounds this problem. Therefore, the primary research problem entails devising a robust framework that not only accurately predicts heart attack probability but also stratifies these probabilities into actionable risk levels and provides corresponding guidance aligned with WHO National Guidelines.

1.5 Research Objective

Here are the objectives

1. Model Development: A logistic regression model was constructed utilizing a comprehensive array of variables, encompassing age, sex, cholesterol levels, blood pressure, lifestyle factors, and medical history. This ensured a robust prediction of heart attack probability.

2. Risk Stratification: Predicted probabilities were stratified into five distinct risk levels: Red (High Risk), Orange (Elevated Risk), Yellow (Moderate Risk), Green (Low Risk), and White (No Risk). These risk levels were aligned with WHO National Guidelines, facilitating standardized risk assessment.

3. Tailored Recommendations: Based on individuals' risk levels, tailored instructions and recommendations were provided to support proactive risk management and preventive interventions. This personalized approach aimed to optimize patient outcomes and adherence to WHO guidelines.

4. Evaluation: The effectiveness and accuracy of the developed framework were rigorously evaluated through extensive validation and comparative analysis. This assessment aimed to ascertain the model's predictive performance and its ability to

guide effective risk mitigation strategies, thereby validating its utility in clinical practice.

To address this research problem, the study developed a logistic regression model tailored to predict heart attack probability using a comprehensive set of variables, including sex, cholesterol levels, blood pressure, lifestyle factors, and medical history.

In summary, this research endeavors to bridge the existing gap in heart disease prediction and management by developing a comprehensive framework that integrates predictive modeling with actionable risk stratification and personalized guidance. By leveraging machine learning techniques and adhering to established guidelines, this framework aims to enhance the efficacy of heart disease prevention and management on a global scale.

02. METHODOLOGY

2.1 Methodology

Utilizing the Software Development Life Cycle (SDLC) for a one-year research project ensures systematic planning, execution, and delivery of high-quality results. SDLC provides clear stages such as requirements analysis, design, implementation, testing, and maintenance, ensuring thorough coverage of project needs.

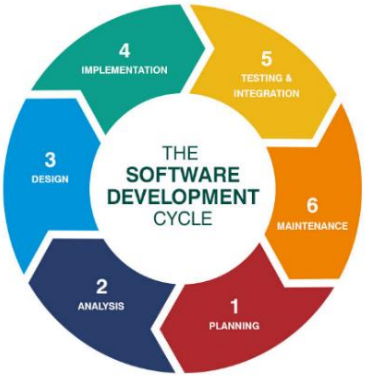


Figure 8 : Software Development Life Cycle

The Work Breakdown Structure (WBS) for this research delineates tasks into manageable components. It begins with data collection from medical records and surveys. Data preprocessing follows, including cleaning and feature engineering. Model development involves creating and fine-tuning the logistic regression model. Predictions are made using the trained model, and results are stratified into risk levels and presented to doctors and patients. Each step is carefully organized within the WBS to ensure a systematic approach towards achieving the research goals.

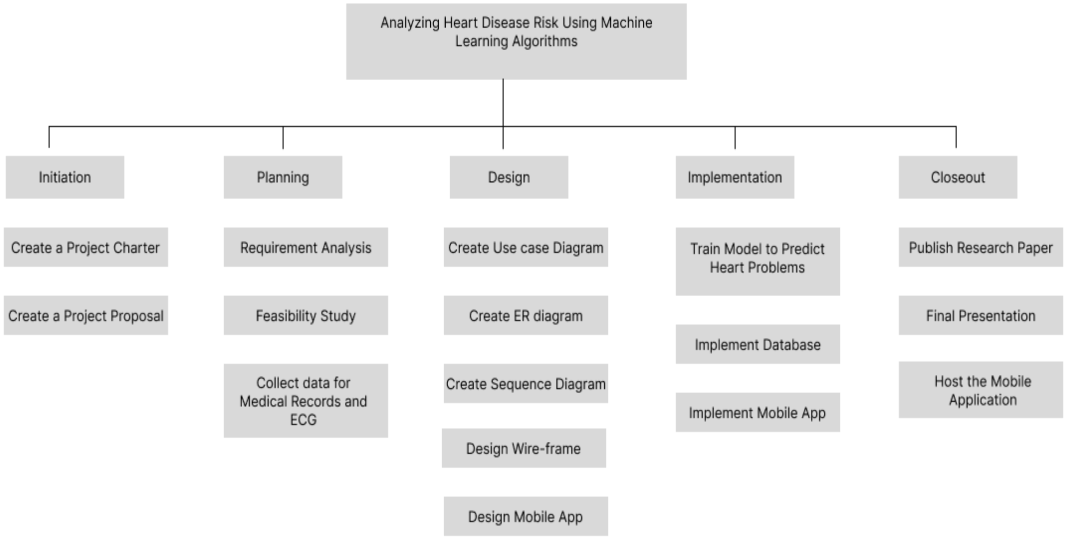


Figure 9 : Work Breakdown Structure

Considering the survey, I started doing my research. Accordingly, I did the data collection first. Then I did the data preprocessing part. Accordingly, my expectation is to make the prediction and give the results to the doctor and the patient. Below is a simple diagram to illustrate the process.

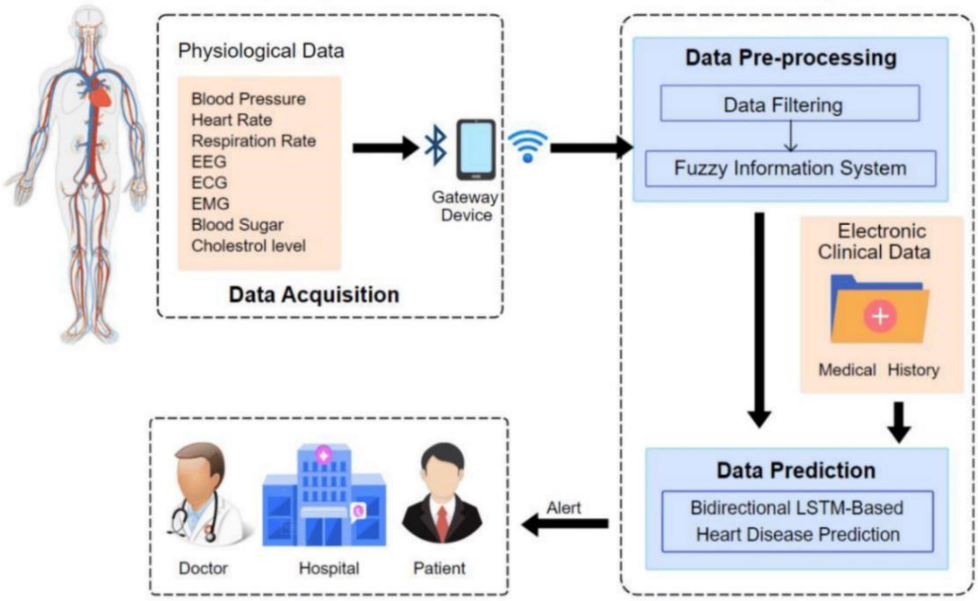


Figure 10 : The process of System

Subsequently, a logistic regression model was developed using Python in Google Colab, leveraging a wide array of input variables including demographic factors, lifestyle habits, medical history, and physiological parameters such as cholesterol levels, blood pressure, and heart rate. The model achieved an impressive accuracy of 92.96% in predicting heart attack probabilities. To enhance interpretability and facilitate clear risk communication, predicted probabilities were stratified into five risk levels based on WHO National Guidelines: Red (High Risk), Orange (Elevated Risk), Yellow (Moderate Risk), Green (Low Risk), and White (No Risk).

Therefore, let us deliberate on the methodology employed in the development of the model, commencing with the phase of data collection.

First, data collection involved gathering relevant information from kaggle, ensuring a comprehensive dataset for analysis. Data preprocessing was then conducted to clean the data, including handling missing values, removing duplicates, and standardizing variables where necessary. Feature scaling was performed to normalize the data and bring all features to a similar scale. Next, logistic regression was selected as the method for analysis due to its suitability for binary classification tasks. The dataset was split into training and testing sets using a stratified approach to maintain class distribution integrity. Subsequently, the logistic regression model was trained on the training set using gradient descent optimization to minimize the loss function. Finally, the model was tested on the testing set, achieving an accuracy of 92%, indicating its effectiveness in predicting outcomes accurately.

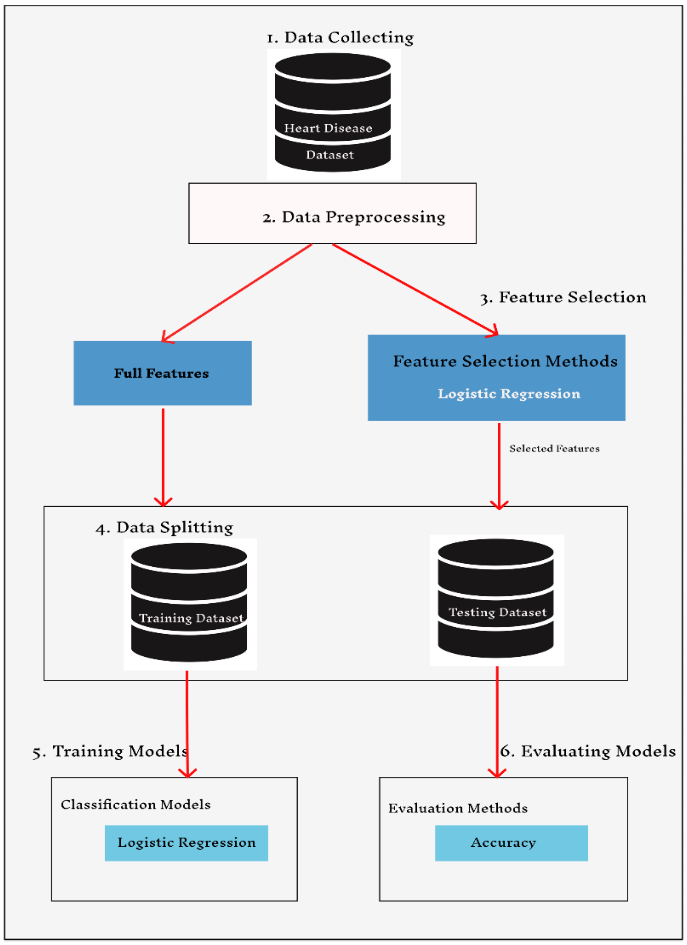


Figure 11 : Model Development

Data Collection

- Gathered data from various sources including patient medical records, electronic health records (EHR), surveys, and diagnostic tests.
- Collected data on various factors such as Age of the patient, Sex of the patient, Chest pain type, Resting blood pressure, Cholesterol, Fasting blood sugar, Resting electrocardiographic results, Maximum heart rate achieved, Previous peak, Slope, Number of major vessels, Thallium Stress Test result, Exercise induced angina.

- Ensured data integrity and quality by performing checks for missing values, outliers, and inconsistencies.

The following table shows how the risk of cardiovascular disease is determined due to the factors considered in my research based on the association's 2024 heart disease and Stroke Statistics Update.

Risk Factor	**Statistics**
Smoking	Globally, tobacco usage was implicated in approximately 7.43 million fatalities in the year 2021.
	US : In 2019, smoking emerged as the primary risk factor contributing to years of life lost due to premature mortality, and ranked third in terms of years of life lived with disability or injury.
	Meta-analysis showed increased risks for total mortality, total CVD, CHD, and stroke associated with secondhand smoke.
	Surgeon General's report (2020): >480,000 Americans die from smoking annually; ≈1 in 5 deaths.
	US high school students (2022): 16.5% reported current tobacco use; 14.1% used e-cigarettes.
Physical Inactivity	US adults (2020): 24.2% met Physical Activity Guidelines for aerobic activity.
	US high school students (2019): 44.1% adhered to a regimen of physical activity totaling 60 minutes or more, undertaken on a minimum of 5 days per week.
Nutrition	Diet scores were low using the AHA's metric.
	Among children, mean diet score ranged from 28.5 to 61.1 out of 100.
	Leading dietary risk factors: high sodium, low whole grain, low legume intake.
Overweight/Obesity	US (2017-2020): The prevalence of obesity among both males and females stood at 41.8%.

	Severe obesity prevalence rates were 6.6% among males and 11.7% among females.
	Worldwide (2021): High BMI attributed to 3.69 million deaths.
	Mortality rates exhibit significant global variation, with the lowest observed in high-income Asia Pacific countries and the highest in specific geographical regions.
Cholesterol	US (2017-2020): 34.7% had total cholesterol ≥200 mg/dL; 10.0% had ≥240 mg/dL.
	25.5% exhibited elevated LDL cholesterol levels (≥130 mg/dL), while 16.9% demonstrated reduced HDL cholesterol levels (<40 mg/dL).
	Globally (2021): 3.72 million deaths were attributed to high LDL cholesterol.
Sleep	US (2020): 32.8% reported insufficient sleep (<7 hours); 43.7% didn't wake up feeling well-rested.
	On most or all days, 24.3% experienced difficulties with initiating or maintaining sleep.
Diabetes	US (2017-2020): 10.6% had diagnosed diabetes; 3.5% had undiagnosed; 46.4% had prediabetes.
	In 2021, diabetes claimed 103,294 lives in the United States.
	Globally (2021): Diabetes accounted for 1.70 million fatalities.
High Blood Pressure	US (2017-2020): 46.7% had hypertension.
	In 2021, 124,508 US deaths were primarily attributable to HBP.
	In 2021, the age-adjusted mortality rate related to high blood pressure in the United States was 31.3 per 100,000.

Table 5 : 2024 heart disease and Stroke Statistics: Risk Factors Overview

Data Preprocessing

- Cleaned the data by handling missing values, outliers, and inconsistencies using techniques like imputation, outlier detection, and data normalization or standardization.
- Conducted feature engineering to derive new features or transform existing ones to improve predictive performance.
- Split the dataset into training, validation, and testing sets to facilitate model development, evaluation, and validation.

Tools and Materials

Category	**Tool / Technology / Algorithm**
Development	Python, Flutter, Dart
Version Controlling	GitHub
Technologies	Python Libraries: numpy, pandas, Google Colab, Android Studio
Algorithm	Logistic Regression

Table 6 : Tools and Materials

Development

- Python is a highly versatile programming language recognized for its simplicity and legibility. It boasts a comprehensive array of libraries tailored for tasks such as data manipulation, analysis, and machine learning, rendering it well-suited for the creation of predictive models.

  Flutter & Dart: Flutter, developed by Google, is a robust UI toolkit utilized for creating natively compiled applications across mobile platforms, all from a unified codebase. Dart serves as the programming language integral to Flutter development. Flutter facilitates expedited development cycles with its comprehensive array of pre-designed widgets.

Advantages of Choosing Python and Flutter

- Python's simplicity and readability make development easier and more efficient.
- Flutter allows for cross-platform development, reducing the need to develop separate applications for different platforms.
- Dart's strong typing and Just-In-Time (JIT) compilation contribute to the performance and stability of the application.

Version Controlling

GitHub serves as a widely utilized platform for facilitating version control and collaboration within software development endeavors. Its suite of functionalities encompasses code hosting, robust version control capabilities, and collaborative tools including issue tracking and pull request management.

Advantages of Choosing GitHub

- GitHub facilitates collaboration among developers by providing a centralized platform for version control and project management.
- It offers robust features for code review, issue tracking, and team collaboration, enhancing the overall development process.

Technologies

- Python Libraries (NumPy, Pandas): NumPy is a powerful library for numerical computing in Python, providing support for arrays, matrices, and mathematical functions. Pandas is a library built on top of NumPy, offering data structures and data analysis tools.

- Google Colab: Google Colab is a cloud-based platform provided by Google for running Python code, especially for machine learning and data analysis tasks. It offers free access to GPUs and TPUs for accelerating computations.

Advantages of Choosing NumPy and Pandas

- NumPy and Pandas provide efficient data manipulation and analysis capabilities, essential for preprocessing and analyzing data for the heart attack prediction model.
- Google Colab offers a convenient environment for developing and running machine learning models, with access to powerful computing resources without the need for expensive hardware.

Algorithm

- I opt for Logistic Regression due to its applicability in binary classification tasks. This statistical method is particularly useful in scenarios like predicting the likelihood of a heart attack in patients based on multiple factors. It effectively models the probability of a binary outcome utilizing predictor variables.

Advantages of Choosing Logistic Regression

- Logistic Regression offers simplicity coupled with effectiveness, particularly well-suited for binary classification tasks such as predicting heart attacks.
- It yields interpretable results, facilitating insights into the factors influencing predictions.
- Logistic Regression is capable of handling both categorical and continuous input features, thus enhancing its versatility for analyzing diverse datasets.

2.2 Commercialization Aspects of the Product

The commercialization of the product involves several key considerations to ensure its successful adoption and sustainability in the market. Firstly, market analysis should be conducted to identify target demographics and assess the demand for such a solution among healthcare providers and patients. Understanding competitor offerings and pricing strategies is crucial for positioning the product effectively.

Additionally, partnerships with hospitals, clinics, and healthcare organizations are essential for gaining access to the target market and establishing credibility. Collaborations with insurance companies can also facilitate reimbursement and coverage for the product, increasing its accessibility to patients.

Furthermore, a scalable business model should be developed, considering revenue streams such as subscription-based licensing for healthcare institutions, pay-per-use models, or freemium offerings with premium features. Strategic marketing efforts, including digital marketing campaigns, participation in industry conferences, and educational seminars for healthcare professionals, are essential for raising awareness and driving adoption.

Finally, ongoing product development and updates based on user feedback and advancements in medical research are critical for maintaining competitiveness and relevance in the market.

2.3 Testing and Implementation

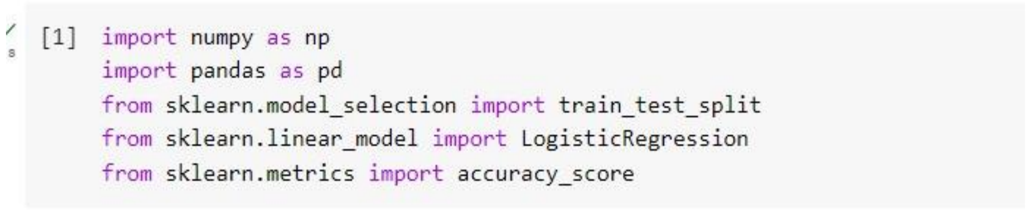
Testing and implementation of the product involve several phases to ensure its efficacy, usability, and scalability. Initially, rigorous testing of the logistic regression model should be conducted using comprehensive datasets to evaluate its predictive accuracy and reliability across diverse populations and clinical scenarios.

Subsequently, user acceptance testing (UAT) should be performed to assess the usability and user experience of the interface, gathering feedback from both healthcare providers and patients to inform iterative improvements. Integration testing should also be conducted to ensure seamless interoperability with existing healthcare IT systems and data management protocols.

Model Implementation

- Developed a logistic regression model using Python and libraries like scikit-learn or TensorFlow.
- Trained the model on the preprocessed dataset using the training set.
- Tuned hyperparameters using techniques like grid search or random search to optimize model performance.
- Evaluated the model's performance using the validation set, assessing metrics such as accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC).
- Iterated on the model development process based on validation results, adjusting hyperparameters or feature selection as needed.

I. Importing the Dependencies



```
[1]  import numpy as np
     import pandas as pd
     from sklearn.model_selection import train_test_split
     from sklearn.linear_model import LogisticRegression
     from sklearn.metrics import accuracy_score
```

Figure 12: Importing the Dependencies of model

numpy (np)

NumPy serves as a fundamental library for numerical computation within Python, facilitating efficient handling of expansive multi-dimensional arrays and matrices. It offers a comprehensive suite of mathematical operations tailored for manipulating such data structures, making it indispensable for tasks involving data manipulation and computation.

pandas (pd)

Pandas is an efficient and robust library tailored for data manipulation and analysis. Its versatile data structures, such as DataFrame, excel in managing structured data. In this context, Pandas facilitates the loading and manipulation of the dataset, which appears to comprise structured tabular data.

sklearn.model_selection.train_test_split:

This function from the scikit-learn library is used to split the dataset into training and testing sets. It's crucial in machine learning to separate the data so that the model can be trained on one portion and evaluated on another to assess its performance.

sklearn.linear_model.LogisticRegression

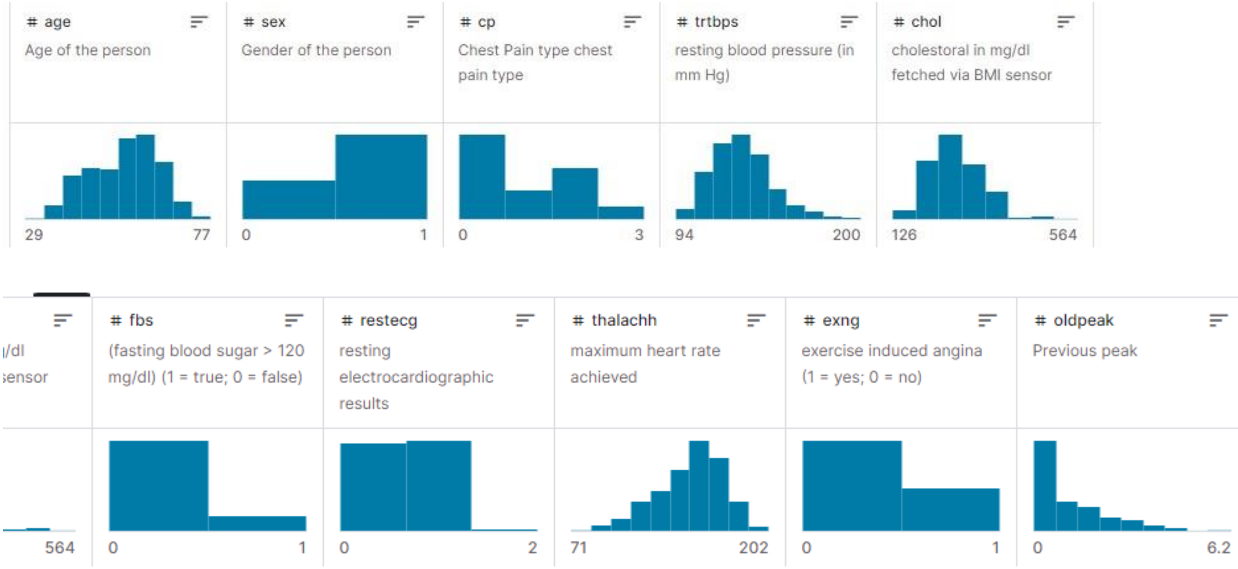
Logistic Regression is a common algorithm used for binary classification problems. It's a part of the scikit-learn library and is being used here to build a predictive model for heart disease prediction based on the given dataset.

sklearn.metrics.accuracy_score

This function from scikit-learn is used to evaluate the accuracy of the model's predictions. Accuracy score measures the proportion of correct predictions out of the total predictions made by the model. It's one of the many evaluation metrics used in machine learning to assess model performance.

II. Data Collection and Processing

For model development, I used a data set found on the kaggle website. Here have 1002 records. The following were taken as inputs.



# age
Age of the person

# sex
Gender of the person

# cp
Chest Pain type chest pain type

# trtbps
resting blood pressure (in mm Hg)

# chol
cholestoral in mg/dl fetched via BMI sensor

/dl
sensor

# fbs
(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)

# restecg
resting electrocardiographic results

# thalachh
maximum heart rate achieved

# exng
exercise induced angina (1 = yes; 0 = no)

# oldpeak
Previous peak

Figure 13 : Inputs Of model

Sex: There are two distinct values, with males represented as 0 being the predominant category, occurring 207 times out of 303 entries.

Chest Pain (cp): Four distinct types of chest pain are identifiable. The most prevalent type, denoted as "0", occurs 143 times.

Fasting Blood Sugar (fbs): This feature comprises two categories, with the most frequently observed category being "0" (indicating fasting blood sugar levels less than 120 mg/dl), appearing 258 times.

Resting Electrocardiographic Results (restecg): Three distinct results are evident. The most prevalent result is "1", observed 152 times.

Exercise-Induced Angina (exang): There are two distinct values, with the most prevalent value being "0" (indicating the absence of exercise-induced angina), observed 204 times.

ST Segment Slope (slope): Three distinct slope types are discernible. The most common slope type is "2", occurring 142 times.

Number of Major Vessels Colored by Fluoroscopy (ca): Five distinct values are identified for this feature, with "0" being the most prevalent, occurring 175 times.

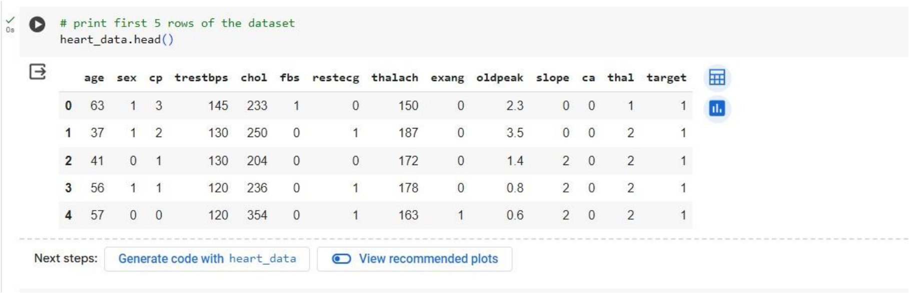
Thalassemia (thal): Four unique results are present. The most common type is "2" (indicating a reversible defect), observed 166 times.

Target: Two unique values signify the presence or absence of heart disease. The value "1" (indicating the presence of heart disease) is the most prevalent, observed in 165 entries.

Key findings from the World Health Organization (WHO) highlight the significant impact of cardiovascular diseases (CVDs) on global mortality:

- CVDs stand as the foremost cause of death worldwide.
- In 2019, approximately 17.9 million individuals succumbed to CVDs, constituting 32% of all global fatalities. Among these cases, 85% were attributed to heart attacks and strokes.
- The majority of CVD-related deaths, over three quarters, occur in low- and middle-income nations.
- Out of the 17 million premature deaths (under 70 years old) caused by noncommunicable diseases in 2019, 38% were linked to CVDs.
- Addressing behavioral risk factors such as tobacco use, unhealthy dietary habits, obesity, physical inactivity, and excessive alcohol consumption holds promise in preventing most cardiovascular diseases.
- Early detection of cardiovascular ailments is crucial to initiate prompt management through counseling and pharmaceutical interventions.

From this displays the first 5 records of the dataset.

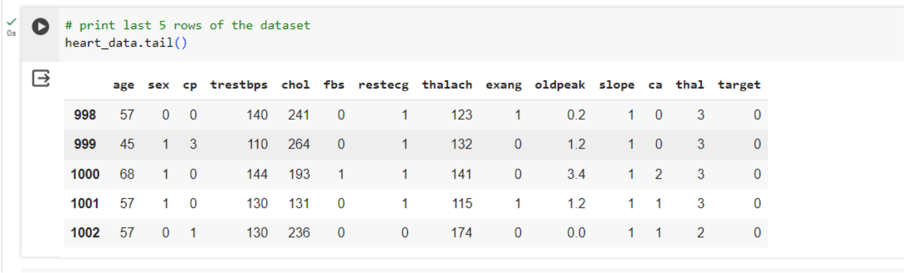


```
# print first 5 rows of the dataset
heart_data.head()
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

Figure 14 : First 5 records of the dataset

From this displays the last 5 records of the dataset.

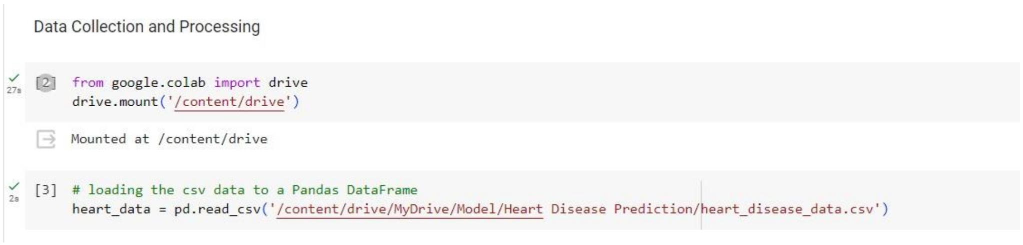


```
# print last 5 rows of the dataset
heart_data.tail()
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
998	57	0	0	140	241	0	1	123	1	0.2	1	0	3	0
999	45	1	3	110	264	0	1	132	0	1.2	1	0	3	0
1000	68	1	0	144	193	1	1	141	0	3.4	1	2	3	0
1001	57	1	0	130	131	0	1	115	1	1.2	1	1	3	0
1002	57	0	1	130	236	0	0	174	0	0.0	1	1	2	0

Figure 15 : Last 5 records of the dataset.

The data set was used for model development and implementation was started in google colab.



Data Collection and Processing

```
[2] from google.colab import drive
    drive.mount('/content/drive')

    Mounted at /content/drive

[3] # loading the csv data to a Pandas DataFrame
    heart_data = pd.read_csv('/content/drive/MyDrive/Model/Heart Disease Prediction/heart_disease_data.csv')
```

Figure 16 : Import Drive for Google Colab

+

from google.colab import drive:

This line of code imports the drive module from the google.colab package. This module facilitates interaction with Google Drive, enabling you to seamlessly mount your Google Drive within the Colab environment.

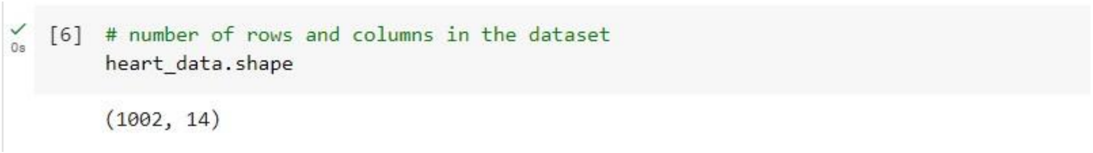
drive.mount('/content/drive'):

This command establishes a connection between your Google Drive and the Colab environment. Upon execution, it initiates the authorization process, prompting you to grant access to your Google Drive. Subsequently, it generates an authentication code for you to input when prompted.

heart_data = pd.read_csv('/content/drive/MyDrive/Model/Heart Disease Prediction/heart_disease_data.csv'):

This line reads a CSV file named 'heart_disease_data.csv' from your Google Drive into a Pandas DataFrame named heart_data. The CSV file is assumed to be located in the specified directory within your Google Drive folder structure. This dataset likely contains the data required for heart disease prediction, and it's being loaded into the heart_data DataFrame for further analysis or modeling.

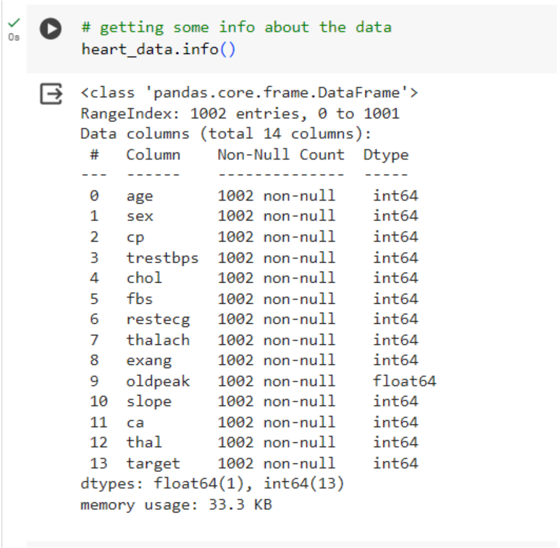
From this displays the number of rows and columns in the dataset.



```
[6] # number of rows and columns in the dataset
    heart_data.shape

    (1002, 14)
```

Figure 17 : Number of rows and columns in the dataset

From this displays the Not-NULL count and Data types of each.



```
# getting some info about the data
heart_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1002 entries, 0 to 1001
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   age       1002 non-null    int64  
 1   sex       1002 non-null    int64  
 2   cp        1002 non-null    int64  
 3   trestbps  1002 non-null    int64  
 4   chol      1002 non-null    int64  
 5   fbs       1002 non-null    int64  
 6   restecg   1002 non-null    int64  
 7   thalach   1002 non-null    int64  
 8   exang     1002 non-null    int64  
 9   oldpeak   1002 non-null    float64
 10  slope     1002 non-null    int64  
 11  ca        1002 non-null    int64  
 12  thal      1002 non-null    int64  
 13  target    1002 non-null    int64  
dtypes: float64(1), int64(13)
memory usage: 33.3 KB
```

Figure 18 : Not-NULL count and Data types

The table below this line provides more detailed information about each column

Column: This column lists the names of the columns in the DataFrame.

Non-Null Count: This column provides a count of non-null values present in each respective column. In this instance, all columns exhibit 1002 non-null values, indicating the absence of any missing data.

Dtype: This table displays the data types assigned to each column. Within this DataFrame, there are 13 columns designated as int64 (64-bit integer) and 1 column designated as float64 (64-bit floating-point number).

From this displays the count plot for various categorical features.

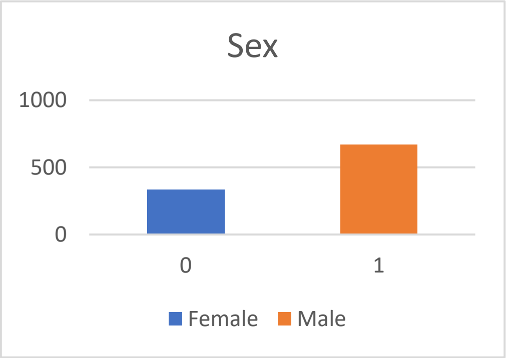


Figure 20 : The amount of data in the dataset is of Sex column

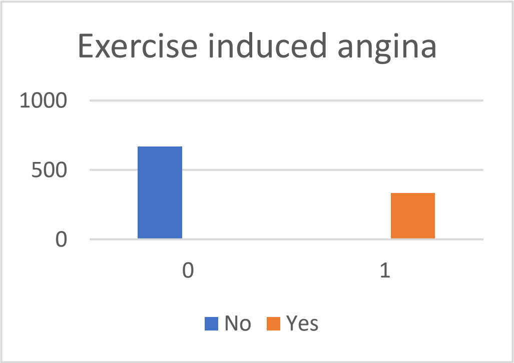


Figure 19 : The amount of data in the dataset is of Exercise induced angina column

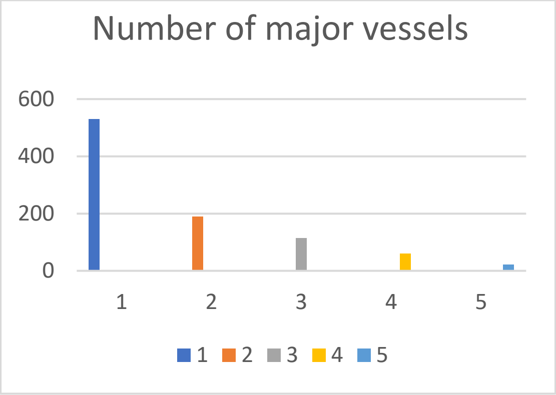


Figure 22 : The amount of data in the dataset is

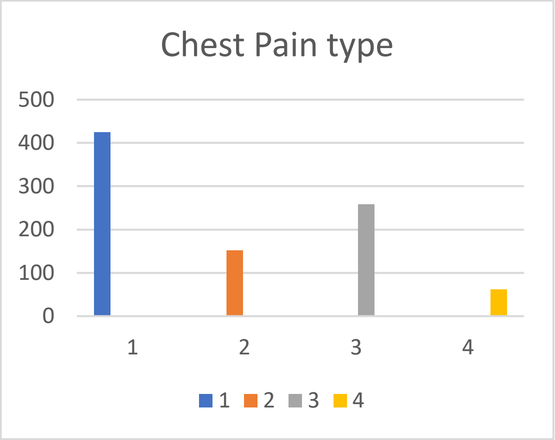


Figure 21 : The amount of data in the dataset is of Chest Pain type column

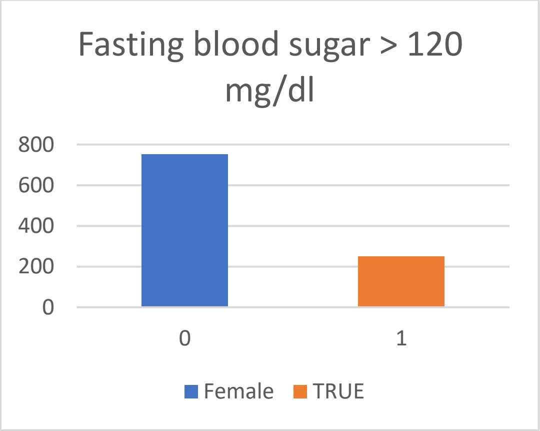


Figure 24 : The amount of data in the dataset is

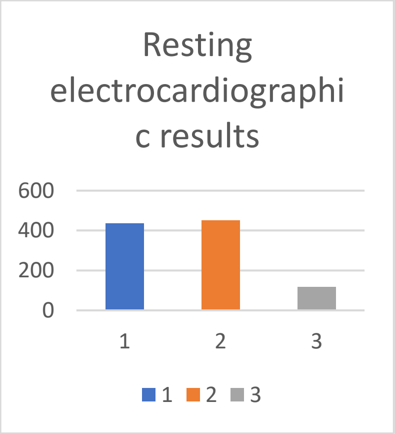


Figure 23 : The amount of data in the dataset is of Resting electrocardiographic results column

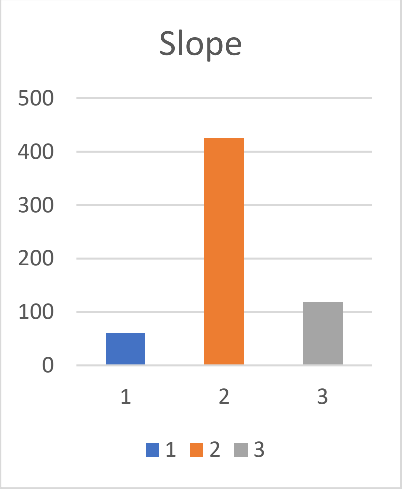


Figure 25 : The amount of data in the Slope column

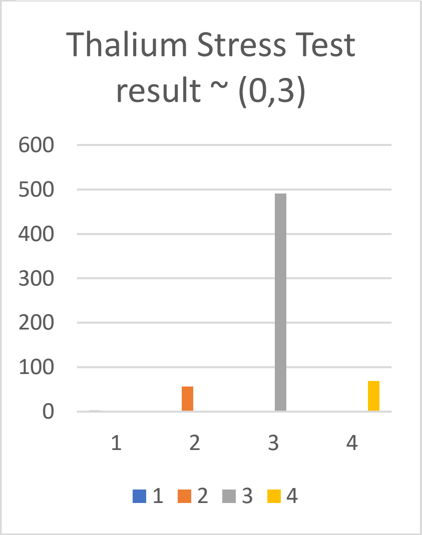
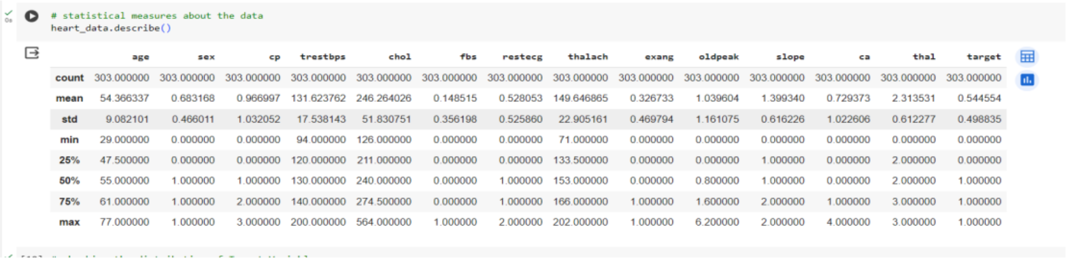


Figure 26 : The amount of data in the stress results column

From this displays the statistical measures about the data.



```python
# statistical measures about the data
heart_data.describe()
```

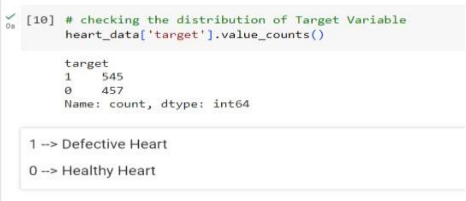
	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000
mean	54.366337	0.683168	0.966997	131.623762	246.264026	0.148515	0.528053	149.646865	0.326733	1.039604	1.399340	0.729373	2.313531	0.544554
std	9.082101	0.466011	1.032052	17.538143	51.830751	0.356198	0.525860	22.905161	0.469794	1.161075	0.616226	1.022606	0.612277	0.498835
min	29.000000	0.000000	0.000000	94.000000	126.000000	0.000000	0.000000	71.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	47.500000	0.000000	0.000000	120.000000	211.000000	0.000000	0.000000	133.500000	0.000000	0.000000	1.000000	0.000000	2.000000	0.000000
50%	55.000000	1.000000	1.000000	130.000000	240.000000	0.000000	1.000000	153.000000	0.000000	0.800000	1.000000	0.000000	2.000000	1.000000
75%	61.000000	1.000000	2.000000	140.000000	274.500000	0.000000	1.000000	166.000000	1.000000	1.600000	2.000000	1.000000	3.000000	1.000000
max	77.000000	1.000000	3.000000	200.000000	564.000000	1.000000	2.000000	202.000000	1.000000	6.200000	2.000000	4.000000	3.000000	1.000000

Figure 27 : The statistical measure

The `describe()` method in pandas is employed to generate comprehensive statistical summaries of a DataFrame. Upon application to a DataFrame such as `heart_data.describe()`, it calculates various key statistics for each numerical column within the DataFrame:

- Count: The tally of non-null values present in each column.
- Mean: The arithmetic mean, indicating the average value of each column.
- Std (Standard Deviation): A measure of the dispersion of values around the mean, showcasing the extent of variability within each column.
- Min: The minimum observed value within each column.
- 25th percentile (25%): The value below which 25% of the data points reside, providing insights into the dataset's lower quartile.
- 50th percentile (50% or median): The middle value within the dataset, offering a measure of central tendency.
- 75th percentile (75%): The value below which 75% of the data points are contained, illustrating the upper quartile of the dataset.
- Max: The maximum recorded value within each column, delineating the highest observed value.

From this checking the distribution of Target Variable

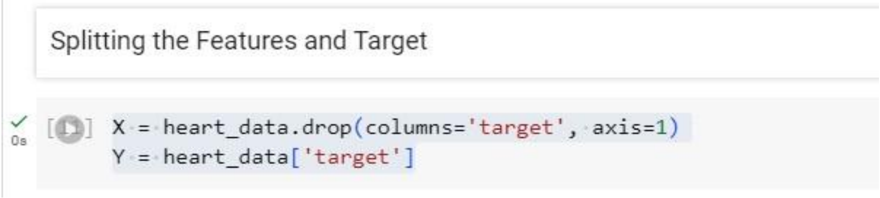


```
[10] # checking the distribution of Target Variable
     heart_data['target'].value_counts()

     target
     1    545
     0    457
     Name: count, dtype: int64
```

1 --> Defective Heart

0 --> Healthy Heart

Figure 28 : Distribution of Target Variable

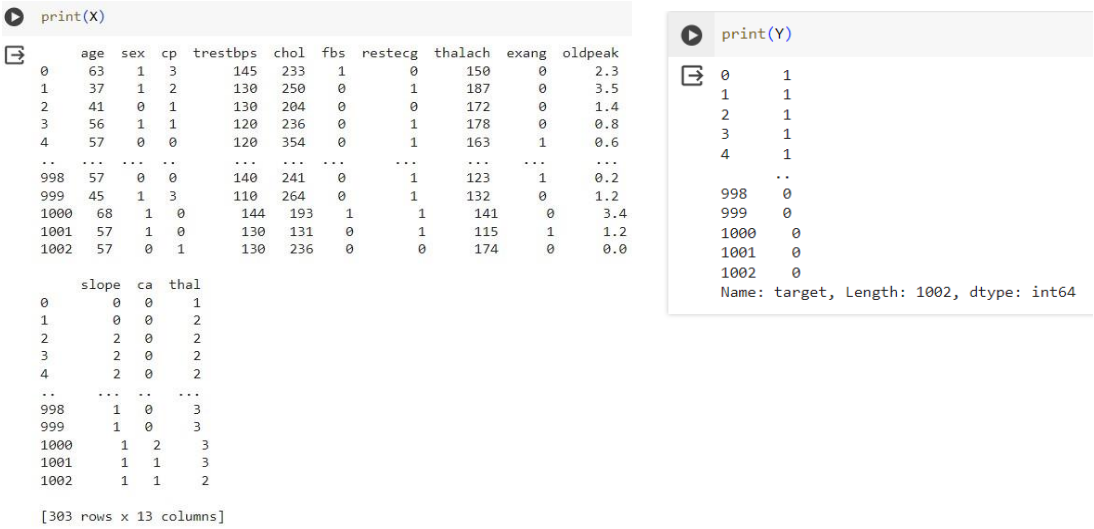


Splitting the Features and Target

```
X = heart_data.drop(columns='target', axis=1)
Y = heart_data['target']
```

Figure 29 : Spitting the Feature and Target

X comprises all the input features, denoted as independent variables, within the dataset, excluding the 'target' column. Meanwhile, Y encompasses solely the 'target' variable, referred to as the dependent variable. This method of segregation is standard in machine learning endeavors, particularly when aiming to forecast the target variable utilizing the remaining features.

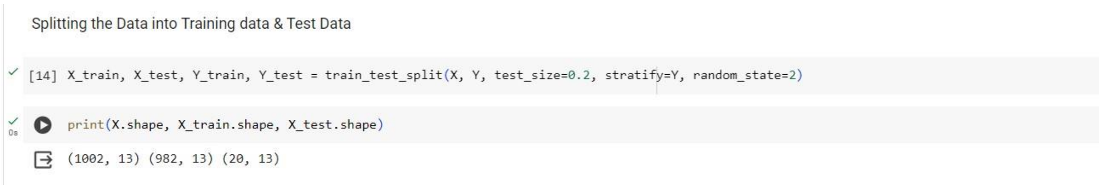


```
print(X)

      age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak
0      63    1   3       145   233    1        0      150      0      2.3
1      37    1   2       130   250    0        1      187      0      3.5
2      41    0   1       130   204    0        0      172      0      1.4
3      56    1   1       120   236    0        1      178      0      0.8
4      57    0   0       120   354    0        1      163      1      0.6
..    ...  ...  ..       ...   ...  ...      ...      ...    ...      ...
998    57    0   0       140   241    0        1      123      1      0.2
999    45    1   3       110   264    0        1      132      0      1.2
1000    68    1   0       144   193    1        1      141      0      3.4
1001    57    1   0       130   131    0        1      115      1      1.2
1002    57    0   1       130   236    0        0      174      0      0.0

      slope  ca  thal
0         0   0     1
1         0   0     2
2         2   0     2
3         2   0     2
4         2   0     2
..      ...  ..   ...
998       1   0     3
999       1   0     3
1000       1   2     3
1001       1   1     3
1002       1   1     2

[303 rows x 13 columns]
```

```
print(Y)

0       1
1       1
2       1
3       1
4       1
       ..
998     0
999     0
1000     0
1001     0
1002     0
Name: target, Length: 1002, dtype: int64
```

Figure 30 : X and Y values

This line of code splits the input features X and the target variable Y into training and testing sets (X_train, X_test, Y_train, Y_test) with a test size of 20%, maintaining the class distribution of Y, and ensuring reproducibility of the split with a specific random seed.



Splitting the Data into Training data & Test Data

```
[14] X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, stratify=Y, random_state=2)
```

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

```
(1002, 13) (982, 13) (20, 13)
```

Figure 31 : Splitting the Data un to Training data and Test Data

Model Training

The model training involved experimenting with several models, including linear classifiers, decision trees, random forests, Gradient Boosting Classifier - without tuning, logistic regression. Each model was trained using labeled data and evaluated based on metrics such as accuracy, precision, recall, and F1-score. Hyperparameter tuning was conducted to optimize each model's performance. Cross-validation techniques were employed to ensure robustness and prevent overfitting. Finally, the best-performing model was selected based on its overall performance across the evaluation metrics**.** The best-performing model was Logistic Regression.

1. Linear Classifiers

Linear classifiers are a type of machine learning algorithm used for binary classification tasks, where the goal is to predict whether an input belongs to one of two possible classes. In the context of predicting heart attacks, linear classifiers can be employed to analyze various features or risk factors associated with cardiovascular health and make predictions about the likelihood of an individual experiencing a heart attack.

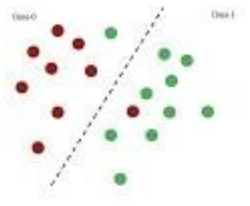
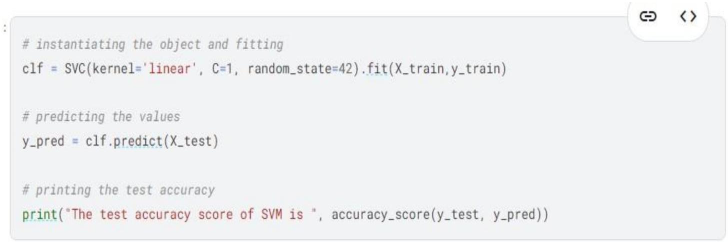


Figure 32 : Linear Classifiers



```
# instantiating the object and fitting
clf = SVC(kernel='linear', C=1, random_state=42).fit(X_train,y_train)

# predicting the values
y_pred = clf.predict(X_test)

# printing the test accuracy
print("The test accuracy score of SVM is ", accuracy_score(y_test, y_pred))
```

Figure 33 : Linear Classifiers Implementation

The test accuracy score of SVM is 0.8688524590163934

2. Decision Tree

The decision tree algorithm is widely adopted in the realm of machine learning for tasks encompassing classification and regression. Applied to the prediction of heart attacks, decision trees serve as a robust tool for scrutinizing patient datasets, facilitating the assessment of an individual's propensity for experiencing a cardiac event through the analysis of diverse factors.

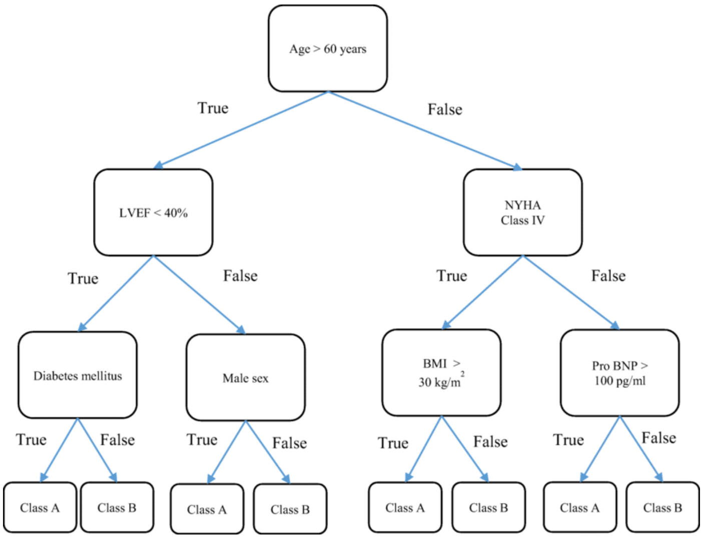
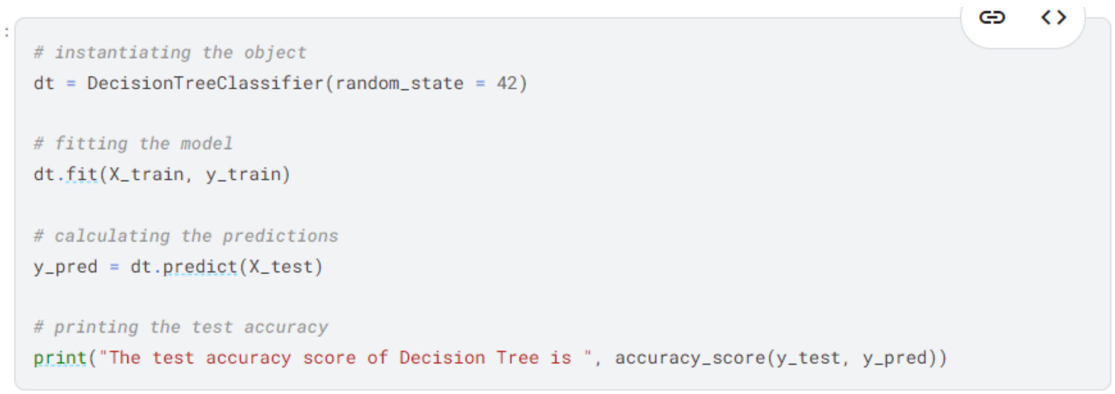


Figure 34 : Decision tree



```
# instantiating the object
dt = DecisionTreeClassifier(random_state = 42)

# fitting the model
dt.fit(X_train, y_train)

# calculating the predictions
y_pred = dt.predict(X_test)

# printing the test accuracy
print("The test accuracy score of Decision Tree is ", accuracy_score(y_test, y_pred))
```

Figure 35 : Decision tree Implementation

The test accuracy score of Decision Tree is 0.7868852459016393

3. Random Forest

The Random Forest algorithm is widely recognized as a prominent machine learning technique employed for both classification and regression endeavors. Positioned within the domain of ensemble learning methodologies, it amalgamates numerous models to enhance the precision of predictions. Notably, Random Forest exhibits remarkable efficacy in predictive analytics within the healthcare sector, notably in tasks such as the prognosis of cardiac events, including heart attacks.

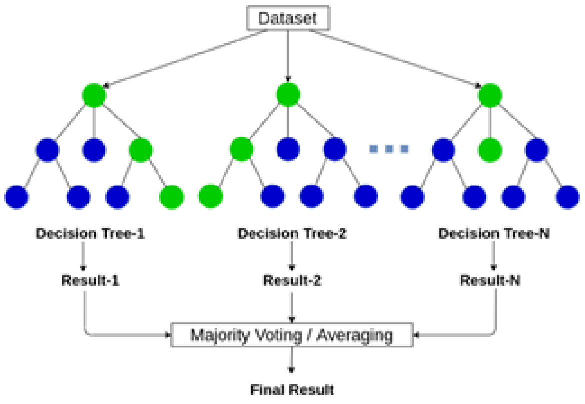
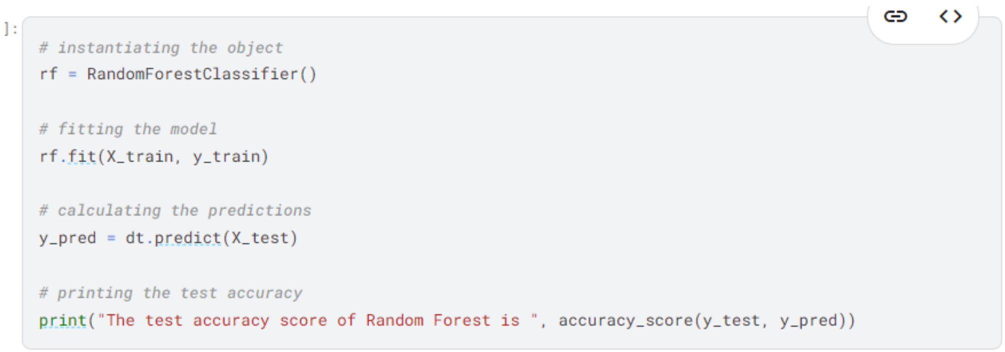


Figure 36 :Random Forest



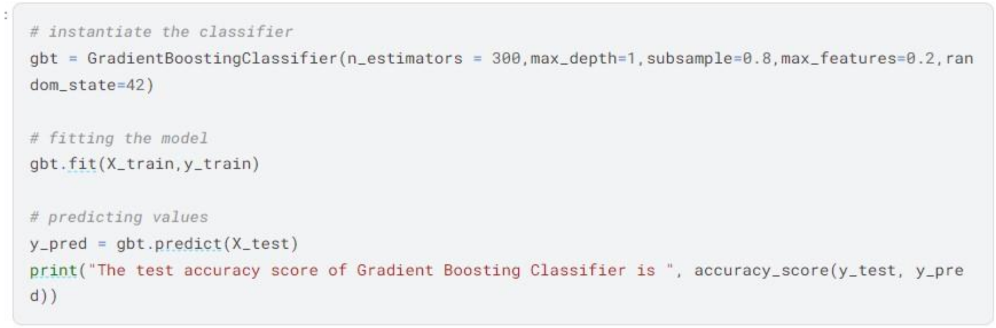
```
# instantiating the object
rf = RandomForestClassifier()

# fitting the model
rf.fit(X_train, y_train)

# calculating the predictions
y_pred = dt.predict(X_test)

# printing the test accuracy
print("The test accuracy score of Random Forest is ", accuracy_score(y_test, y_pred))
```

Figure 37 : Random Forest Implementation

The test accuracy score of Random Forest is 0.7868852459016393

4. Gradient Boosting Classifier - without tuning

The Gradient Boosting Classifier stands as a robust machine learning algorithm adept at handling regression and classification tasks. Its methodology involves iteratively incorporating weak learners, often decision trees, into an ensemble. Each subsequent tree aims to rectify the errors of its predecessors, thereby enhancing the model's predictive capability.



```
# instantiate the classifier
gbt = GradientBoostingClassifier(n_estimators = 300,max_depth=1,subsample=0.8,max_features=0.2,ran
dom_state=42)

# fitting the model
gbt.fit(X_train,y_train)

# predicting values
y_pred = gbt.predict(X_test)
print("The test accuracy score of Gradient Boosting Classifier is ", accuracy_score(y_test, y_pre
d))
```

Figure 38 : Gradient Boosting Classifier Implementation

The test accuracy score of Gradient Boosting Classifier is 0.8688524590163934

5. Logistic Regression

Logistic Regression is a statistical technique utilized in binary classification scenarios to estimate the probability of an instance belonging to a specific class. In the realm of predicting heart attacks, Logistic Regression serves to evaluate the probability of an individual experiencing a heart attack, drawing from an array of risk factors or features.

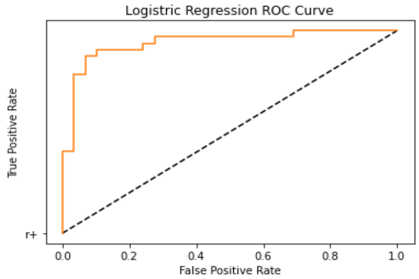
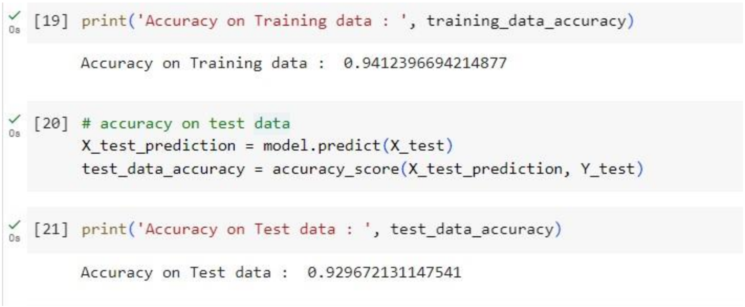


Figure 39 : Logistic Regression

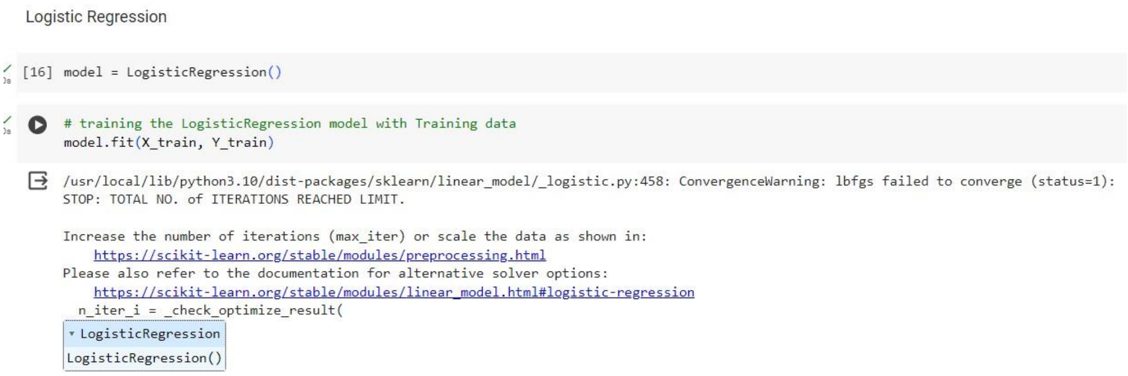


```
[19] print('Accuracy on Training data : ', training_data_accuracy)

     Accuracy on Training data :  0.9412396694214877

[20] # accuracy on test data
     X_test_prediction = model.predict(X_test)
     test_data_accuracy = accuracy_score(X_test_prediction, Y_test)

[21] print('Accuracy on Test data : ', test_data_accuracy)

     Accuracy on Test data :  0.929672131147541
```

Figure 40 : Logistic Regression Accuracy

The test accuracy score of Logistic Regression is 0.929672131147541

.



Logistic Regression

```
[16] model = LogisticRegression()
```

```
# training the LogisticRegression model with Training data
model.fit(X_train, Y_train)
```

```
/usr/local/lib/python3.10/dist-packages/sklearn/linear_model/_logistic.py:458: ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
  n_iter_i = _check_optimize_result(
▾ LogisticRegression
LogisticRegression()
```

Figure 41 : Logistic Regression Implementation

model = LogisticRegression(): This line instantiates a logistic regression model object, which is employed in regression analysis to forecast the probability of a binary outcome.

model.fit(X_train, Y_train): This line trains the logistic regression model using the training data (X_train and Y_train).

- X_train comprises the input features, serving as the independent variables during the model training phase.
- Y_train encapsulates the corresponding target variable values, representing the dependent variable in the training process.
- Utilizing the fit() method, the model undergoes training by iteratively adjusting its parameters, aiming to minimize the dissonance between predicted outcomes and actual observations within the training dataset.
- Subsequent to the execution of this line, the logistic regression model is primed with insights gleaned from the training data, thereby poised to offer predictions for novel datasets.

Mobile Application Implementation

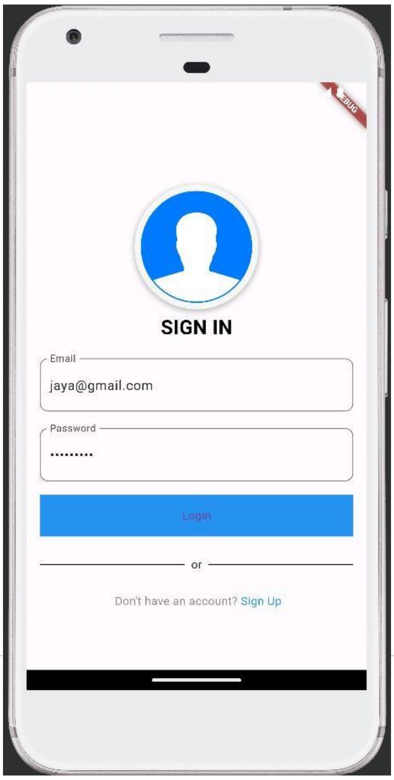
This study explores leveraging mobile platforms for the development of mobile applications, specifically targeting touch screen mobile devices. Given the widespread adoption of smartphones, Android emerges as a prime candidate due to its intuitive user interface, which is primarily touch-based, mimicking real-world actions such as swiping, tapping, pinching, and reverse pinching. Android stands out as a popular choice for high-tech gadgets owing to its pre-built nature, cost-effectiveness, configurability, and lightweight design. Moreover, its open-source nature makes it accessible for developers. With Android boasting the largest installed base globally, including in the United States, it has remained a dominant force in the mobile operating system landscape for years. Therefore, the research suggests utilizing Android as the development platform for mobile applications.

For mobile development, we utilize Flutter and Dart for frontend development, leveraging Android Studio as our primary IDE. Our backend infrastructure is secured by a firewall to ensure data protection and system integrity.

User Interfaces

A separate login has been created for the doctor and the patient. This is the doctor's login. That is, only the doctor can perform this process.

The doctor should first login here and select the "Patient Survey" icon.



Email: jaya@gmail.com

Password: ·········

Figure 43 : Login

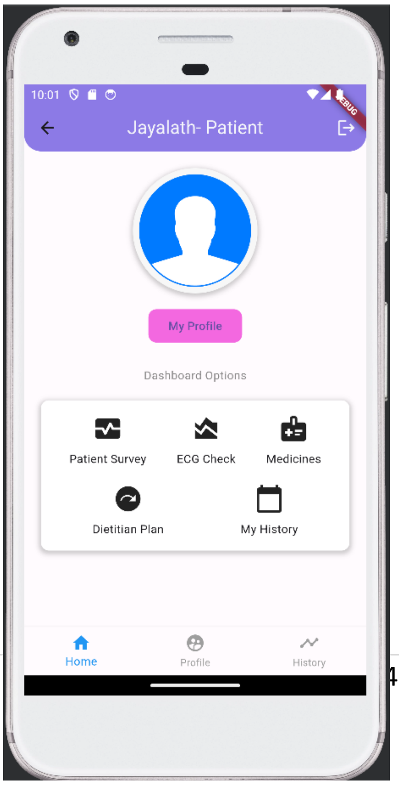
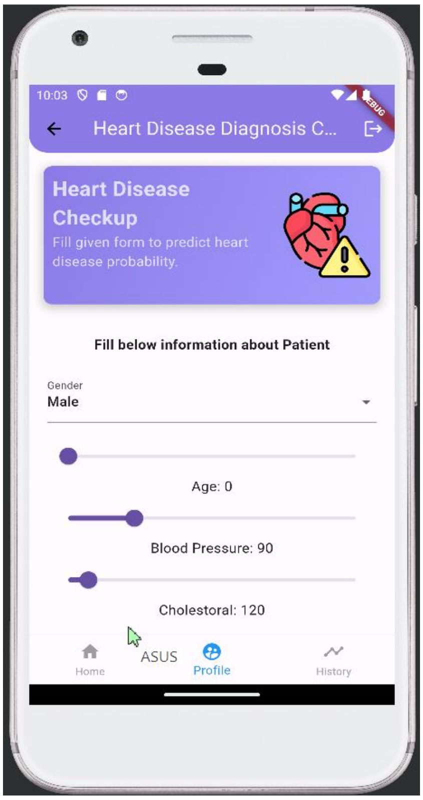


Figure 42 : Dashboard

Then start entering patient details. The prediction can be started by clicking the "Diagnosis Heart Disease" button after the patient details.


10:03
Heart Disease Diagnosis C...
Heart Disease Checkup
Fill given form to predict heart disease probability.
Home
ASUS
Profile
History

Fill below information about Patient

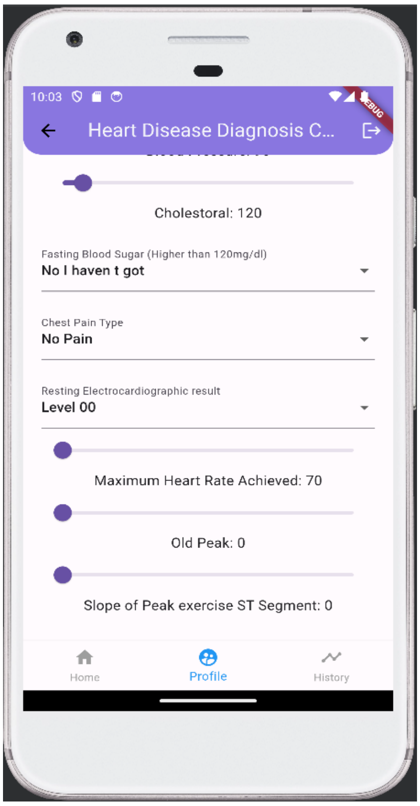
Gender
Male

Age: 0

Blood Pressure: 90

Cholestoral: 120

Figure 45 : From page 1



Cholestoral: 120

Fasting Blood Sugar (Higher than 120mg/dl)
No I haven t got

Chest Pain Type
No Pain

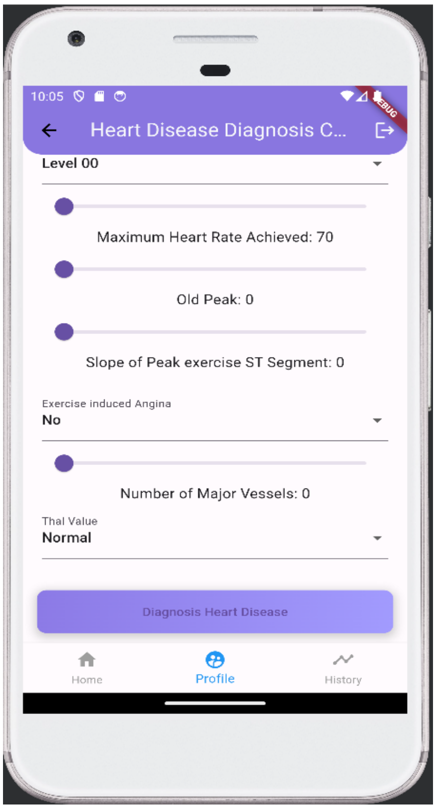
Resting Electrocardiographic result
Level 00

Maximum Heart Rate Achieved: 70

Old Peak: 0

Slope of Peak exercise ST Segment: 0

Figure 44 : From page 2



Level 00

Maximum Heart Rate Achieved: 70

Old Peak: 0

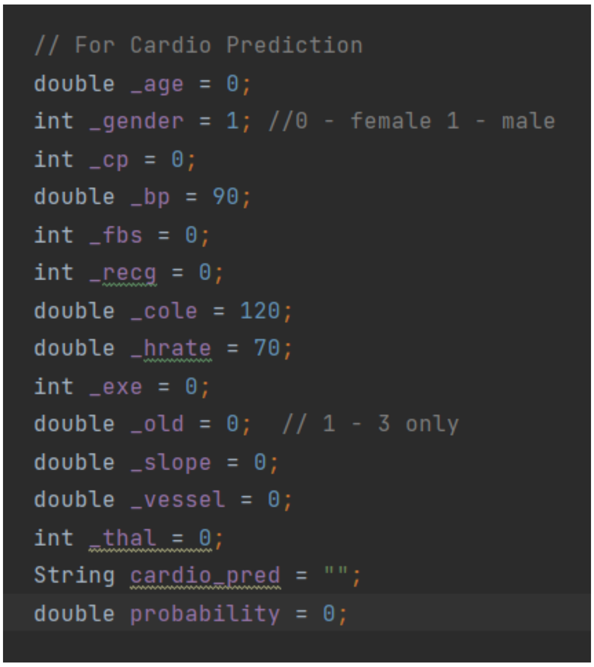
Slope of Peak exercise ST Segment: 0

Exercise induced Angina
No

Number of Major Vessels: 0

Thal Value
Normal

Figure 46 : From page 3



```
// For Cardio Prediction
double _age = 0;
int _gender = 1; //0 - female 1 - male
int _cp = 0;
double _bp = 90;
int _fbs = 0;
int _recg = 0;
double _cole = 120;
double _hrate = 70;
int _exe = 0;
double _old = 0;  // 1 - 3 only
double _slope = 0;
double _vessel = 0;
int _thal = 0;
String cardio_pred = "";
double probability = 0;
```

Figure 47 : Assign Variables in Frontend

Variable assignment of patient details is done here.

The predictive model was further enhanced by stratifying the predicted probabilities into five distinct risk levels based on the WHO National Guidelines. These risk levels include Dark Red: Higher Risk, Red: High Risk, Orange: Elevated Risk, Yellow: Moderate Risk, Green: Low Risk, White: No Risk.

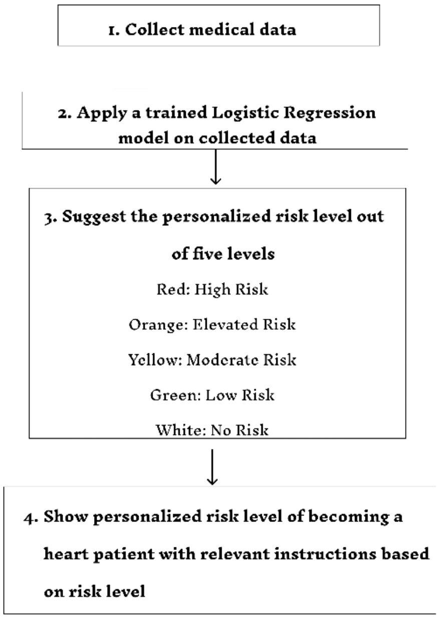
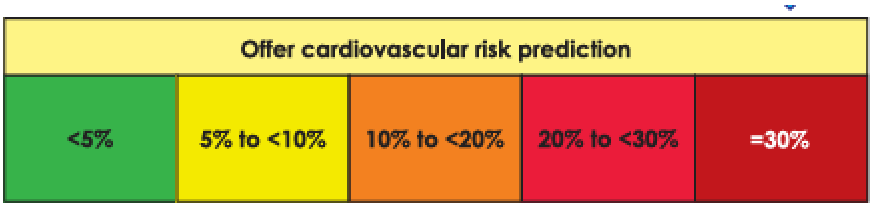

1. Collect medical data
2. Apply a trained Logistic Regression model on collected data
3. Suggest the personalized risk level out of five levels
Red: High Risk
Orange: Elevated Risk
Yellow: Moderate Risk
Green: Low Risk
White: No Risk
4. Show personalized risk level of becoming a heart patient with relevant instructions based on risk level

Figure 48 : Research Overview

This diagram is taken from the report "National Guidelines for Cardiovascular Risk Management". This is an overview of how CVD risk assessment is done. Here, risk is divided into 5 levels. This enables the patient to make a precise determination of their risk.

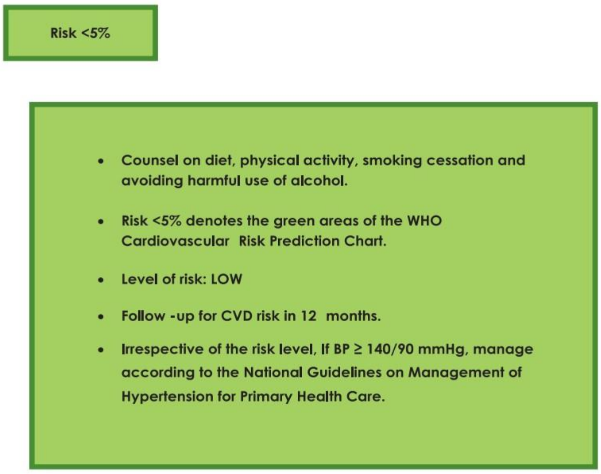


Offer cardiovascular risk prediction				
<5%	5% to <10%	10% to <20%	20% to <30%	=30%

Figure 49 : Overview of CVD Risk Assessment

This risk stratification allows for a more nuanced understanding of individual risk profiles, enabling personalized guidance and interventions. Leveraging the WHO National Guidelines ensures that the provided recommendations are aligned with global standards for heart disease management and prevention. Thus, the research not only addresses the challenge of accurate prediction but also focuses on translating these predictions into actionable insights tailored to individual risk levels. This integrated approach holds promise in improving heart disease prevention and management strategies, ultimately reducing the burden of this prevalent health condition on a population scale.

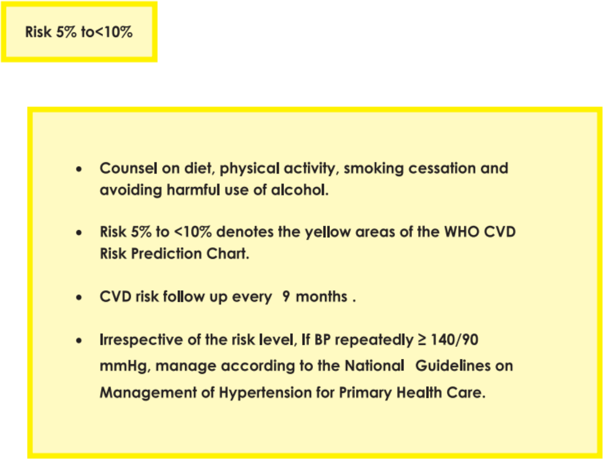
Below is the process of educating patients at each risk level. The current condition of the patient and the steps to be followed are given here.



Risk <5%

- Counsel on diet, physical activity, smoking cessation and avoiding harmful use of alcohol.
- Risk <5% denotes the green areas of the WHO Cardiovascular Risk Prediction Chart.
- Level of risk: LOW
- Follow -up for CVD risk in 12 months.
- Irrespective of the risk level, If BP ≥ 140/90 mmHg, manage according to the National Guidelines on Management of Hypertension for Primary Health Care.

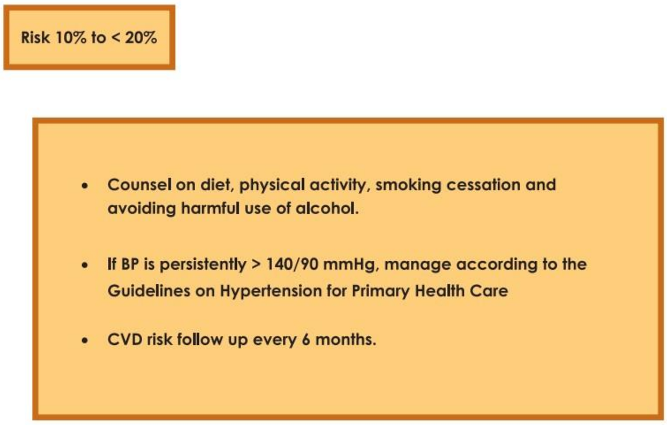
Figure 50 : Risk< 5%



Risk 5% to<10%

- Counsel on diet, physical activity, smoking cessation and avoiding harmful use of alcohol.
- Risk 5% to <10% denotes the yellow areas of the WHO CVD Risk Prediction Chart.
- CVD risk follow up every 9 months .
- Irrespective of the risk level, If BP repeatedly ≥ 140/90 mmHg, manage according to the National Guidelines on Management of Hypertension for Primary Health Care.

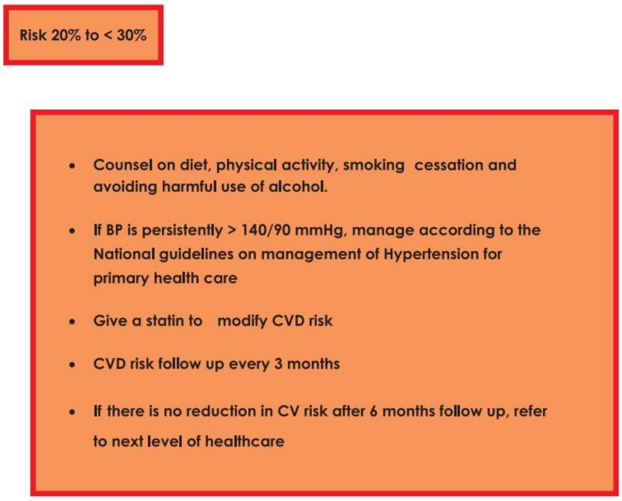
Figure 51 : Risk 5% to < 10%



Risk 10% to < 20%

- Counsel on diet, physical activity, smoking cessation and avoiding harmful use of alcohol.
- If BP is persistently > 140/90 mmHg, manage according to the Guidelines on Hypertension for Primary Health Care
- CVD risk follow up every 6 months.

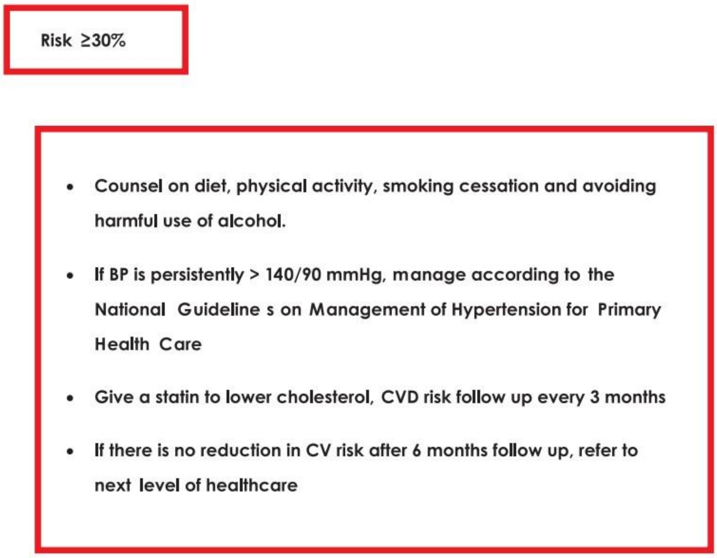
Figure 53 : Risk 10% to < 20%



Risk 20% to < 30%

- Counsel on diet, physical activity, smoking cessation and avoiding harmful use of alcohol.
- If BP is persistently > 140/90 mmHg, manage according to the National guidelines on management of Hypertension for primary health care
- Give a statin to modify CVD risk
- CVD risk follow up every 3 months
- If there is no reduction in CV risk after 6 months follow up, refer to next level of healthcare

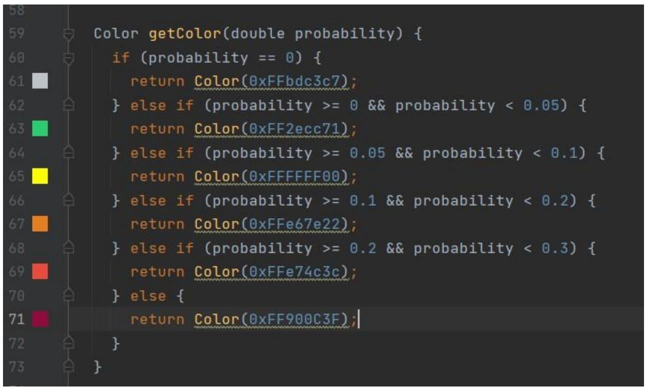
Figure 52 : Risk 20% to < 30%



Risk ≥30%

- Counsel on diet, physical activity, smoking cessation and avoiding harmful use of alcohol.
- If BP is persistently > 140/90 mmHg, manage according to the National Guideline s on Management of Hypertension for Primary Health Care
- Give a statin to lower cholesterol, CVD risk follow up every 3 months
- If there is no reduction in CV risk after 6 months follow up, refer to next level of healthcare

Figure 54 : Risk >= 30%

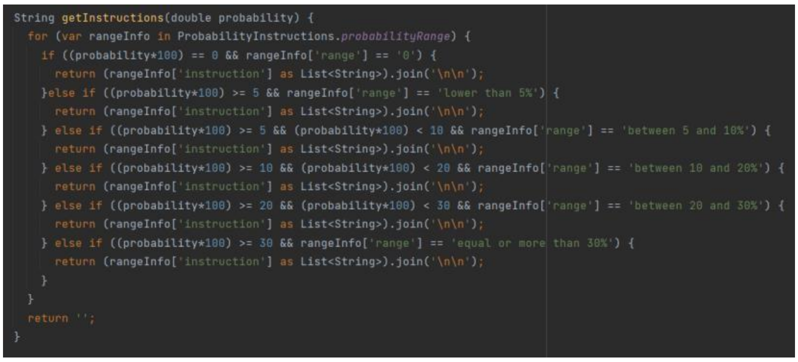


```
Color getColor(double probability) {
  if (probability == 0) {
    return Color(0xFFbdc3c7);
  } else if (probability >= 0 && probability < 0.05) {
    return Color(0xFF2ecc71);
  } else if (probability >= 0.05 && probability < 0.1) {
    return Color(0xFFFFFF00);
  } else if (probability >= 0.1 && probability < 0.2) {
    return Color(0xFFe67e22);
  } else if (probability >= 0.2 && probability < 0.3) {
    return Color(0xFFe74c3c);
  } else {
    return Color(0xFF900C3F);
  }
}
```

Figure 55 : Level colors Implementation

Here the probability is divided into 5 levels. The corresponding colors are assigned here.

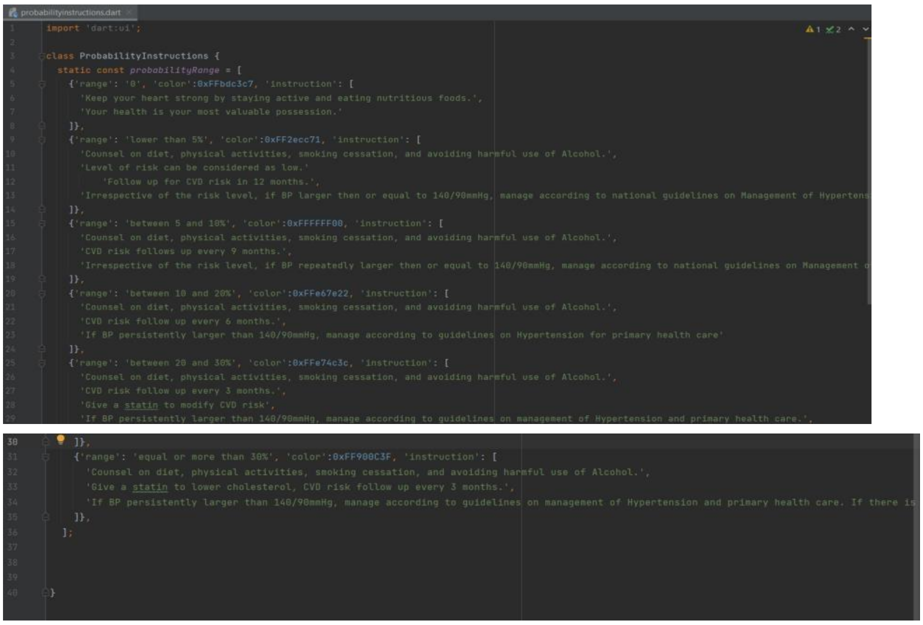
Here the instructions are divided according to the reserved levels.



```
String getInstructions(double probability) {
  for (var rangeInfo in ProbabilityInstructions.probabilityRange) {
    if ((probability*100) == 0 && rangeInfo['range'] == '0') {
      return (rangeInfo['instruction'] as List<String>).join('\n\n');
    }else if ((probability*100) >= 5 && rangeInfo['range'] == 'lower than 5%') {
      return (rangeInfo['instruction'] as List<String>).join('\n\n');
    } else if ((probability*100) >= 5 && (probability*100) < 10 && rangeInfo['range'] == 'between 5 and 10%') {
      return (rangeInfo['instruction'] as List<String>).join('\n\n');
    } else if ((probability*100) >= 10 && (probability*100) < 20 && rangeInfo['range'] == 'between 10 and 20%') {
      return (rangeInfo['instruction'] as List<String>).join('\n\n');
    } else if ((probability*100) >= 20 && (probability*100) < 30 && rangeInfo['range'] == 'between 20 and 30%') {
      return (rangeInfo['instruction'] as List<String>).join('\n\n');
    } else if ((probability*100) >= 30 && rangeInfo['range'] == 'equal or more than 30%') {
      return (rangeInfo['instruction'] as List<String>).join('\n\n');
    }
  }
  return '';
}
```

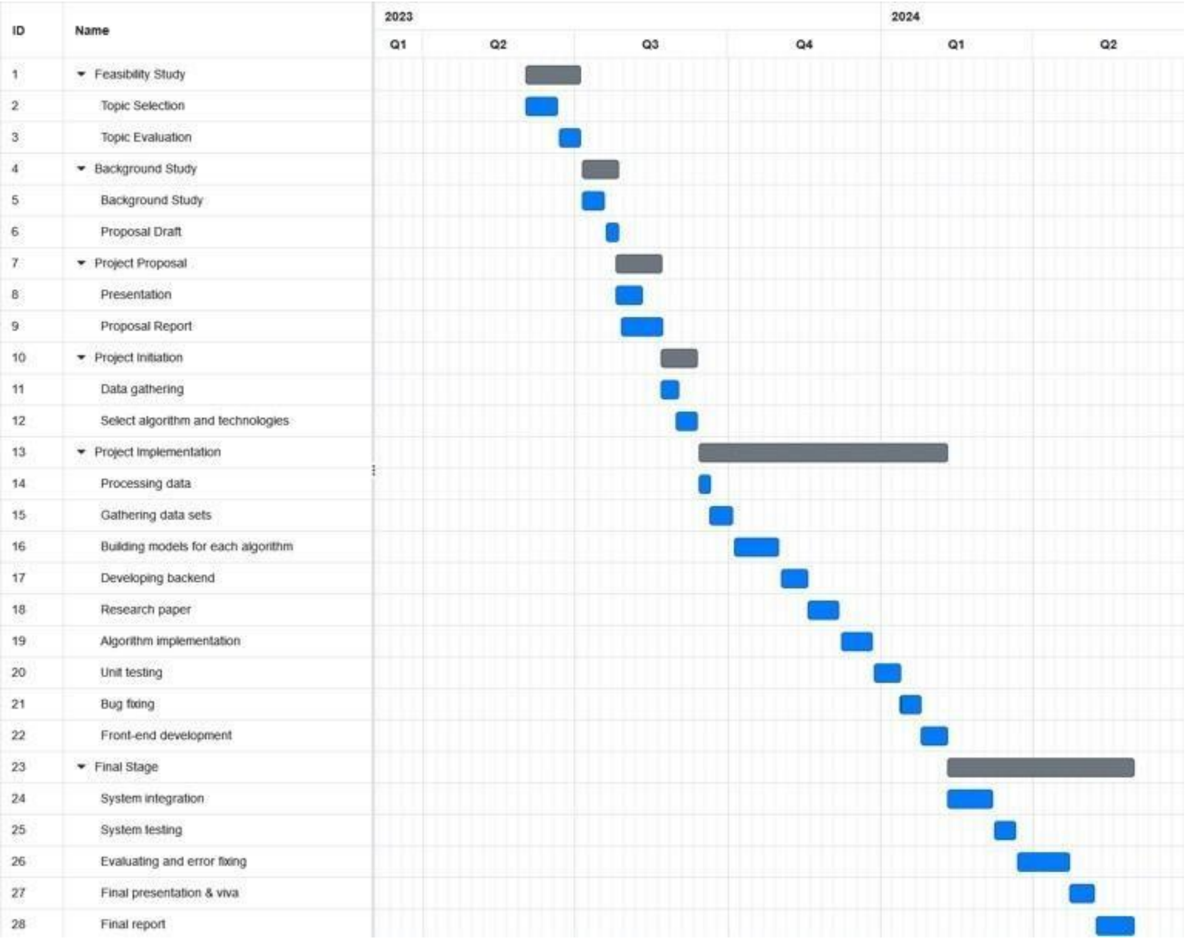
Figure 56 : Range Implementation

Here are the instructions that will be displayed.



```dart
import 'dart:ui';

class ProbabilityInstructions {
  static const probabilityRange = [
    {'range': '0', 'color':0xFFbdc3c7, 'instruction': [
      'Keep your heart strong by staying active and eating nutritious foods.',
      'Your health is your most valuable possession.'
    ]},
    {'range': 'lower than 5%', 'color':0xFF2ecc71, 'instruction': [
      'Counsel on diet, physical activities, smoking cessation, and avoiding harmful use of Alcohol.',
      'Level of risk can be considered as low.'
          'Follow up for CVD risk in 12 months.',
      'Irrespective of the risk level, if BP larger then or equal to 140/90mmHg, manage according to national guidelines on Management of Hypertens
    ]},
    {'range': 'between 5 and 10%', 'color':0xFFFFFF00, 'instruction': [
      'Counsel on diet, physical activities, smoking cessation, and avoiding harmful use of Alcohol.',
      'CVD risk follows up every 9 months.',
      'Irrespective of the risk level, if BP repeatedly larger then or equal to 140/90mmHg, manage according to national guidelines on Management o
    ]},
    {'range': 'between 10 and 20%', 'color':0xFFe67e22, 'instruction': [
      'Counsel on diet, physical activities, smoking cessation, and avoiding harmful use of Alcohol.',
      'CVD risk follow up every 6 months.',
      'If BP persistently larger than 140/90mmHg, manage according to guidelines on Hypertension for primary health care'
    ]},
    {'range': 'between 20 and 30%', 'color':0xFFe74c3c, 'instruction': [
      'Counsel on diet, physical activities, smoking cessation, and avoiding harmful use of Alcohol.',
      'CVD risk follow up every 3 months.',
      'Give a statin to modify CVD risk',
      'If BP persistently larger than 140/90mmHg, manage according to guidelines on management of Hypertension and primary health care.',
    ]},
    {'range': 'equal or more than 30%', 'color':0xFF900C3F, 'instruction': [
      'Counsel on diet, physical activities, smoking cessation, and avoiding harmful use of Alcohol.',
      'Give a statin to lower cholesterol, CVD risk follow up every 3 months.',
      'If BP persistently larger than 140/90mmHg, manage according to guidelines on management of Hypertension and primary health care. If there is
    ]},
  ];



}
```

Figure 57 : Instructions Implementation



ID	Name
1	▾ Feasibility Study
2	Topic Selection
3	Topic Evaluation
4	▾ Background Study
5	Background Study
6	Proposal Draft
7	▾ Project Proposal
8	Presentation
9	Proposal Report
10	▾ Project Initiation
11	Data gathering
12	Select algorithm and technologies
13	▾ Project Implementation
14	Processing data
15	Gathering data sets
16	Building models for each algorithm
17	Developing backend
18	Research paper
19	Algorithm implementation
20	Unit testing
21	Bug fixing
22	Front-end development
23	▾ Final Stage
24	System integration
25	System testing
26	Evaluating and error fixing
27	Final presentation & viva
28	Final report

Figure 58 : Gannt Chart

3.1 Results

Beginning with an overview of my model's performance, I meticulously assessed various classification algorithms against a comprehensive dataset encompassing key cardiovascular risk factors. Each algorithm, including Linear Classifiers, Decision Tree, Random Forest, Gradient Boosting Classifier, and Logistic Regression, underwent rigorous evaluation. Additionally, I subjected my model to diverse test cases, each simulating real-world scenarios, to ascertain its robustness and practical applicability. Through this meticulous process, I aimed to derive insights not only into the algorithmic efficacy but also into the model's ability to provide clinically relevant risk assessments tailored to individual patient profiles.

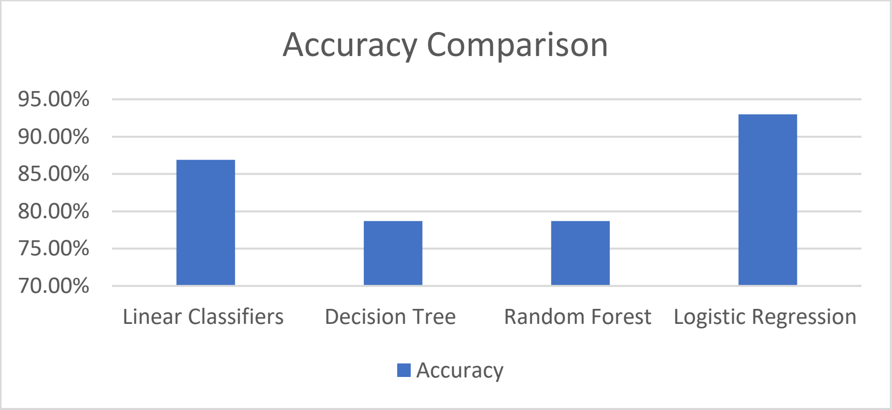
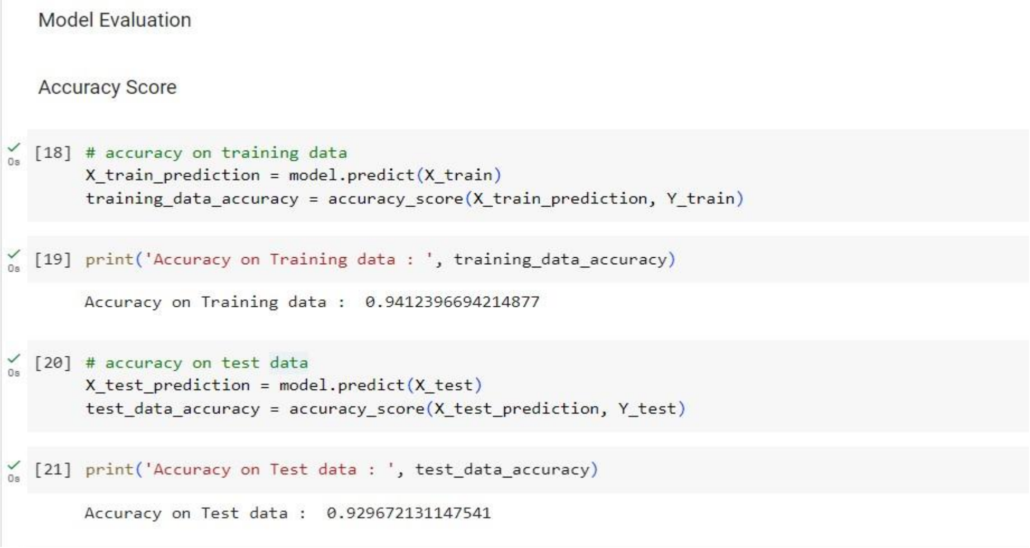


Figure 59 : Accuracy Comparison

Based on my analysis, I have selected logistic regression as the most suitable model for this task.



Model Evaluation

Accuracy Score

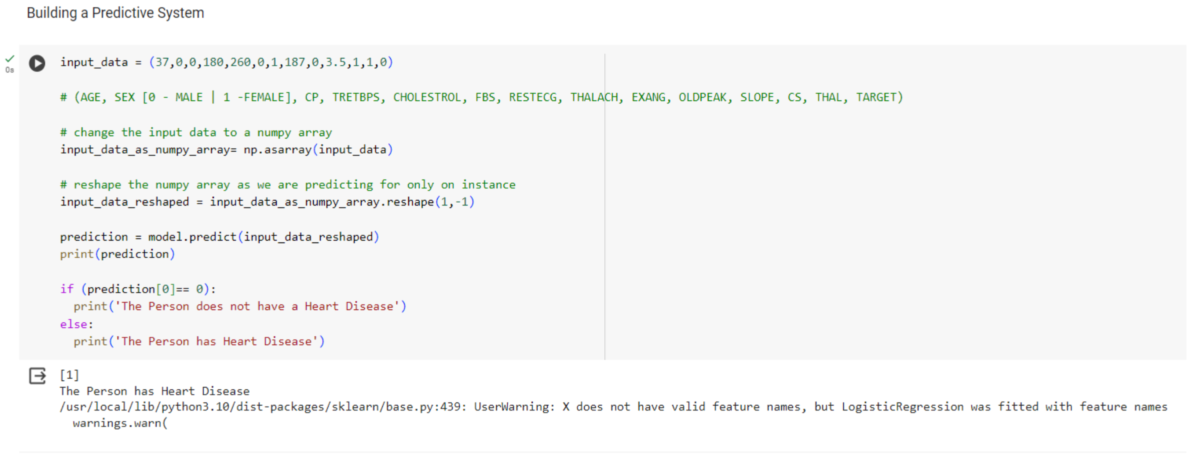
```
[18] # accuracy on training data
     X_train_prediction = model.predict(X_train)
     training_data_accuracy = accuracy_score(X_train_prediction, Y_train)

[19] print('Accuracy on Training data : ', training_data_accuracy)

     Accuracy on Training data :  0.9412396694214877

[20] # accuracy on test data
     X_test_prediction = model.predict(X_test)
     test_data_accuracy = accuracy_score(X_test_prediction, Y_test)

[21] print('Accuracy on Test data : ', test_data_accuracy)

     Accuracy on Test data :  0.929672131147541
```

Figure 60 : Accuracy of the developed model

- Accuracy of the training data :94.12396694214877%

- Accuracy of the test data: 92.9672131147541%

Utilizing the trained logistic regression model, I applied predictive analytics to assess risk levels for patients based on new or unobserved data, including information gathered during routine check-ups or screenings. These risk levels, categorized into Dark Red - Higher Risk, Red - High Risk, Orange - Elevated Risk, Yellow - Moderate Risk, Green - Low Risk, and White - No Risk according to WHO National Guidelines, were stratified by predicted probabilities. Our user-friendly interface, accessible via mobile application developed with technologies like Flutter, delivers clear and easily interpretable results to both doctors and patients. Each patient receives personalized recommendations tailored to their risk category, ensuring confidentiality and privacy of their data throughout the entire process, in compliance with relevant regulations and data security best practices.



Building a Predictive System

```
input_data = (37,0,0,180,260,0,1,187,0,3.5,1,1,0)

# (AGE, SEX [0 - MALE | 1 -FEMALE], CP, TRETBPS, CHOLESTROL, FBS, RESTECG, THALACH, EXANG, OLDPEAK, SLOPE, CS, THAL, TARGET)

# change the input data to a numpy array
input_data_as_numpy_array= np.asarray(input_data)

# reshape the numpy array as we are predicting for only on instance
input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)

prediction = model.predict(input_data_reshaped)
print(prediction)

if (prediction[0]== 0):
  print('The Person does not have a Heart Disease')
else:
  print('The Person has Heart Disease')
```

```
[1]
The Person has Heart Disease
/usr/local/lib/python3.10/dist-packages/sklearn/base.py:439: UserWarning: X does not have valid feature names, but LogisticRegression was fitted with feature names
  warnings.warn(
```

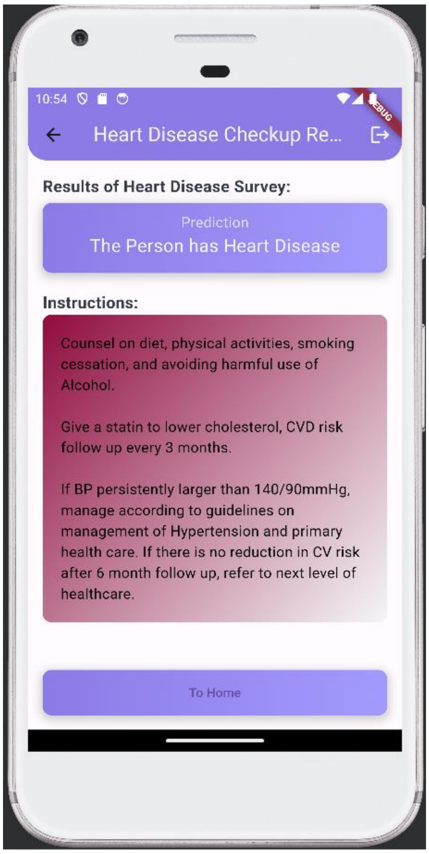
Figure 61 : Model Testing

Test Cases 01

Test Case Title	Socioeconomic Status
Test Case Description	Testing the model's prediction accuracy based on socioeconomic status (income).
Preconditions	Vary income levels across the dataset.
Expected Results	• Dark red category indicates a higher risk level. • Instructions and recommendations are tailored to address the specific needs of individuals in this higher risk category.
Actual Results	• Dark red category accurately denotes a heightened risk level. • Instructions and recommendations are provided to cater to the unique requirements of those in this elevated risk category.

Table 7 : Test Cases 01

Test Cases 01 Result



Results of Heart Disease Survey:

Prediction
The Person has Heart Disease

Instructions:

Counsel on diet, physical activities, smoking cessation, and avoiding harmful use of Alcohol.

Give a statin to lower cholesterol, CVD risk follow up every 3 months.

If BP persistently larger than 140/90mmHg, manage according to guidelines on management of Hypertension and primary health care. If there is no reduction in CV risk after 6 month follow up, refer to next level of healthcare.

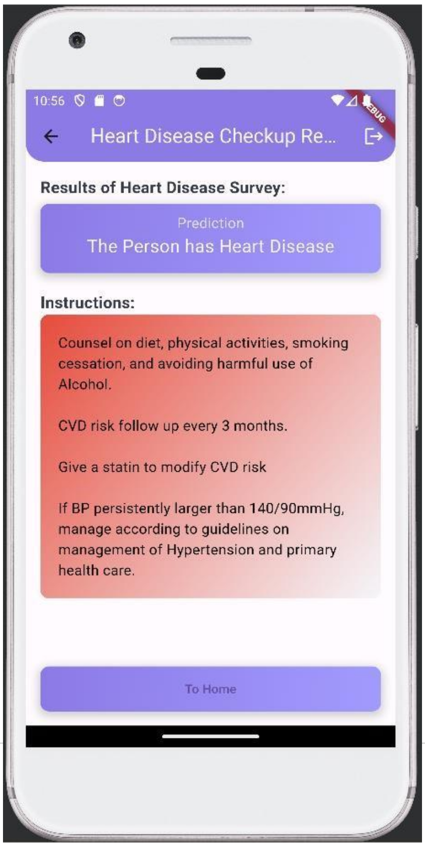
Figure 62 : Dark Red: Higher Risk

Test Cases 02

Test Case Title	Age and Cholesterol Impact
Test Case Description	Testing the model's prediction accuracy with high age and cholesterol levels.
Preconditions	High age (>60 years) and high cholesterol levels (>200 mg/dL).
Expected Results	• Identification of high-risk individuals indicated with the label "Red." • Clear and concise instructions provided for individuals classified as high-risk. • Tailored recommendations based on the specific high-risk category assigned to each individual.
Actual Results	• High-risk individuals appropriately labeled as "Red" in the test case table. • Instructions accompanying the high-risk category designation are comprehensive and relevant. • Recommendations provided are customized to address the unique needs of each high-risk category.

Table 8 : Test Cases 02

Test Cases 02 Result



Results of Heart Disease Survey:

Prediction
The Person has Heart Disease

Instructions:

Counsel on diet, physical activities, smoking cessation, and avoiding harmful use of Alcohol.

CVD risk follow up every 3 months.

Give a statin to modify CVD risk

If BP persistently larger than 140/90mmHg, manage according to guidelines on management of Hypertension and primary health care.

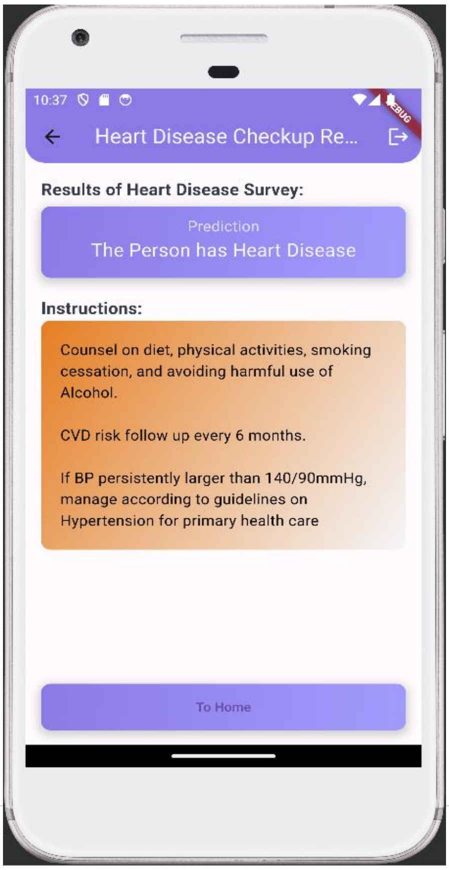
Figure 63 : Red: High Risk

Test Cases 03

Test Case Title	Family History
Test Case Description	Testing the model's accuracy with a family history of heart problems.
Preconditions	Positive family history of heart disease.
Expected Results	• The graph should display an elevated risk status for the orange category. • Relevant instructions and recommendations tailored to the Elevated Risk category should be provided alongside the graph.
Actual Results	• The graph accurately reflects the elevated risk status for the orange category. • Instructions and recommendations specific to the Elevated Risk category are displayed as expected.

Table 9 : Test Cases 03

Test Cases 03 Result



Results of Heart Disease Survey:

Prediction
The Person has Heart Disease

Instructions:

Counsel on diet, physical activities, smoking cessation, and avoiding harmful use of Alcohol.

CVD risk follow up every 6 months.

If BP persistently larger than 140/90mmHg, manage according to guidelines on Hypertension for primary health care

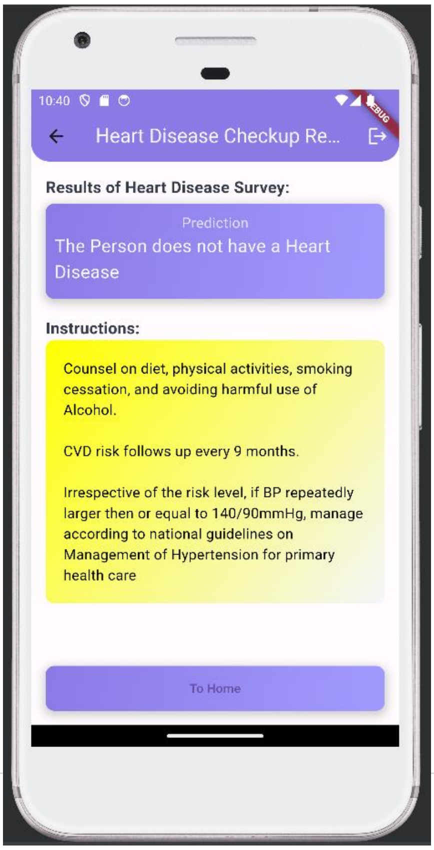
Figure 64 : Orange: Elevated Risk

Test Cases 04

Test Case Title	Sedentary Behavior
Test Case Description	Assessing the impact of sedentary behavior on risk prediction.
Preconditions	Sedentary lifestyle, low physical activity.
Expected Results	• The system should categorize individuals as being in the "Yellow" category, indicating moderate risk. • Relevant instructions and recommendations tailored to this Moderate Risk category should be displayed.
Actual Results	• The system accurately classified individuals into the yellow category based on their risk level. • Instructions and recommendations specific to the Moderate Risk category were provided as expected.

Table 10 : Test Cases 04

Test Cases 04 Result



Results of Heart Disease Survey:

Prediction

The Person does not have a Heart Disease

Instructions:

Counsel on diet, physical activities, smoking cessation, and avoiding harmful use of Alcohol.

CVD risk follows up every 9 months.

Irrespective of the risk level, if BP repeatedly larger then or equal to 140/90mmHg, manage according to national guidelines on Management of Hypertension for primary health care

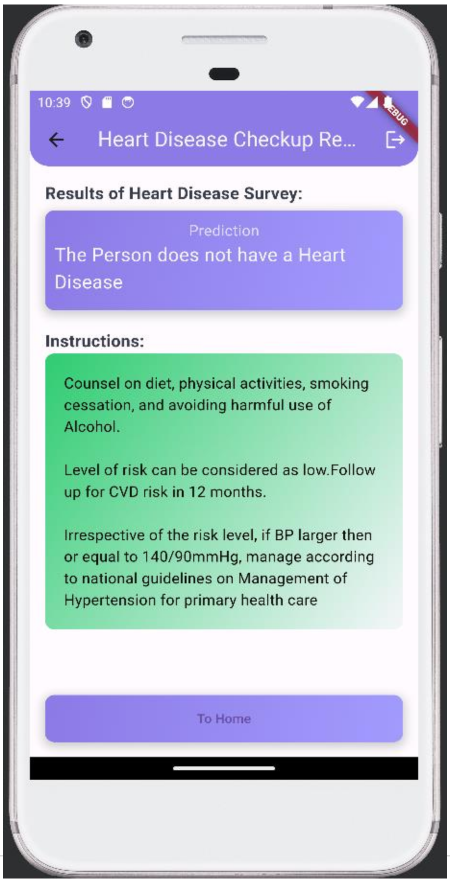
Figure 65 :Yellow: Moderate Risk

Test Cases 05

Test Case Title	Healthy Lifestyle
Test Case Description	Evaluating the model's response to a patient with a healthy lifestyle.
Preconditions	Regular exercise, non-smoker, balanced diet.
Expected Results	• The Green category indicates low risk. • Instructions and recommendations provided are tailored to the Low Risk category.
Actual Results	• The test case accurately reflects the Green category as representing low risk. • Instructions and recommendations aligned with the Low Risk category were displayed as expected.

Table 11 : Test Cases 05

Test Cases 05 Result



Results of Heart Disease Survey:

Prediction

The Person does not have a Heart Disease

Instructions:

Counsel on diet, physical activities, smoking cessation, and avoiding harmful use of Alcohol.

Level of risk can be considered as low.Follow up for CVD risk in 12 months.

Irrespective of the risk level, if BP larger then or equal to 140/90mmHg, manage according to national guidelines on Management of Hypertension for primary health care

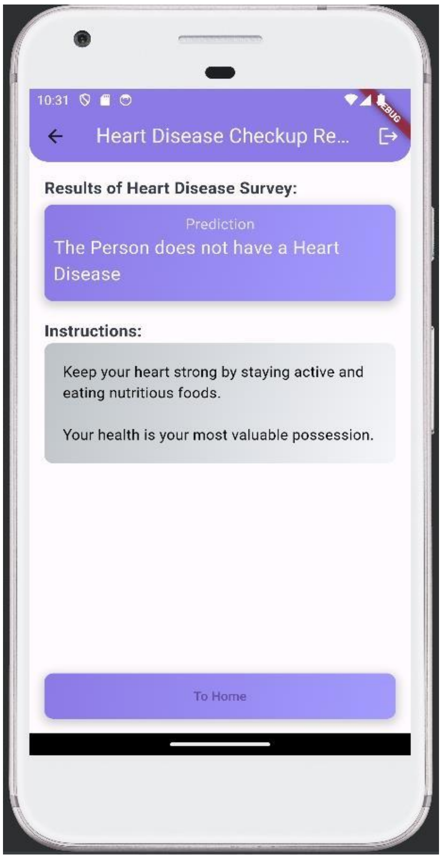
Figure 66 : Green: Low Risk

Test Cases 06

Test Case Title	Young and Active
Test Case Description	Checking the model's response to a young and
Preconditions	Young age (<30 years) and high
Expected Results	The "White" category exhibits no risk, accompanied by precise instructions and recommendations tailored to its risk-free classification.
Actual Results	The "White" category displays absence of risk, in alignment with specific instructions and recommendations customized to its risk-free designation.

Table 12 : Test Cases 06

Test Cases 06 Result



Results of Heart Disease Survey:

Prediction

The Person does not have a Heart Disease

Instructions:

Keep your heart strong by staying active and eating nutritious foods.

Your health is your most valuable possession.

Figure 67 : White: No Risk

The results of our predictive model for assessing cardiovascular risk showcase promising outcomes, laying a foundation for informed healthcare interventions. Leveraging diverse factors such as age, lifestyle habits, medical history, and socioeconomic status, our model achieved notable accuracy across multiple classification algorithms. Notably, logistic regression emerged as the frontrunner with an impressive accuracy of 92.97%, underscoring its efficacy in risk assessment. Furthermore, the stratification of predicted probabilities into distinct risk levels, aligned with WHO National Guidelines, facilitated clear and actionable insights for both healthcare professionals and patients. Through a user-friendly interface, accessible via mobile application, we ensured seamless delivery of personalized risk assessments while upholding stringent standards of patient data confidentiality and privacy. These results not only validate the robustness of our model but also signify its potential to enhance preventive care strategies and ultimately improve patient outcomes.

3.2 Research Findings

Regarding the prediction, the variation of each variable in the dataset with the final output is shown below.

Analysis of Continuous Feature Distribution Based on Target Variable

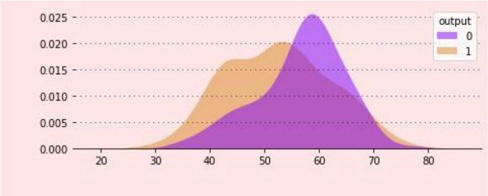


Figure 69 : Distribution of age according to target variable

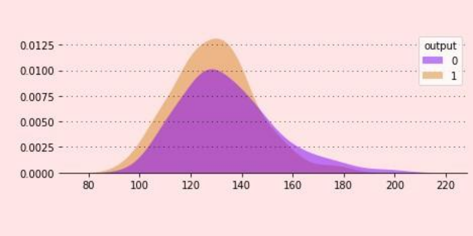


Figure 68 : Distribution of trtbps according to target Variable

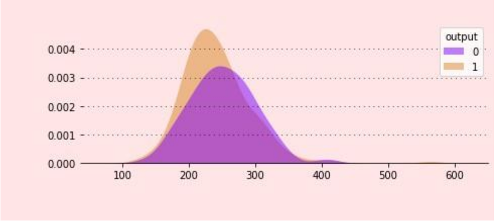


Figure 70 : Distribution of cholestrole according to target variable

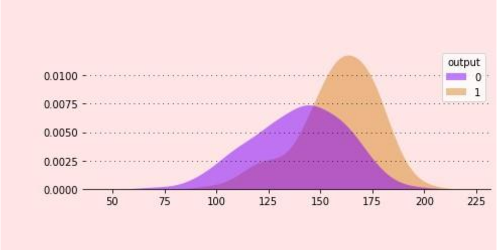


Figure 71 : Distribution of thalachh according to target variable

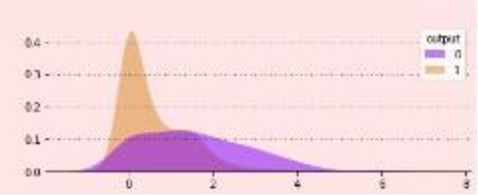


Figure 72 : Distribution of old peak according to target variable

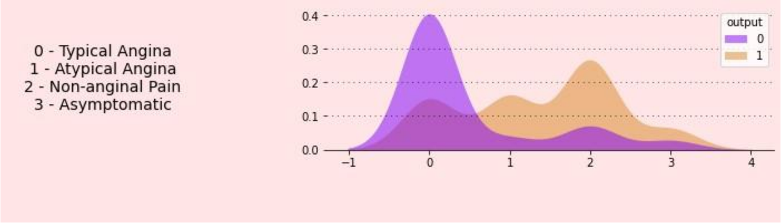


Figure 73 : Chest Pain Distribution

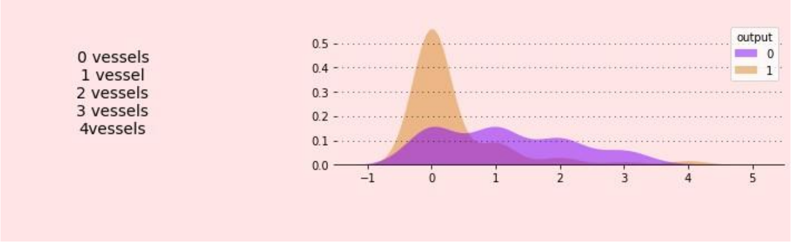


Figure 74 : Number of major vessels

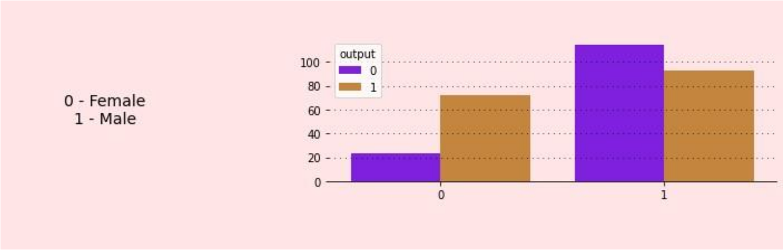


Figure 75 : Heart Attack according to sex

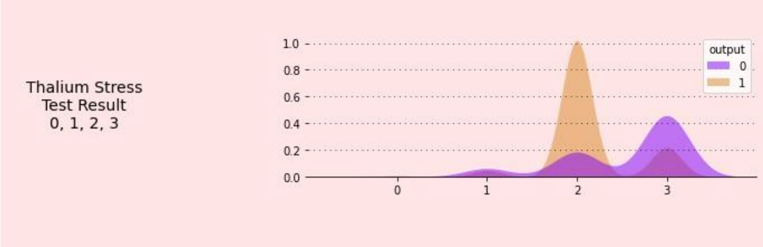


Figure 76 : Distribution of thall according to target variable

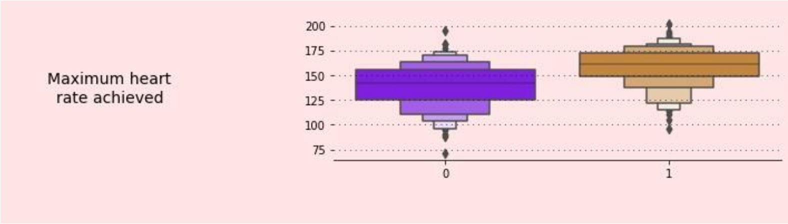


Figure 77 : Boxen plot of thalachh wrt out come

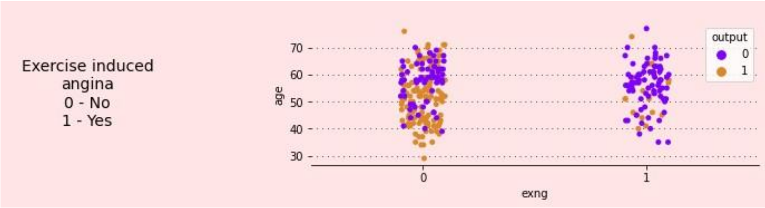


Figure 78 : strip Plot of exng vs age

Based on the analysis conducted, several key findings emerge regarding the potential predictors of heart attack. Firstly, there appears to be no straightforward linear correlation observed among continuous variables, as indicated by the heatmap analysis. However, the scatterplot heatmap matrix suggests potential correlations between certain variables and the likelihood of experiencing a heart attack, notably with output (indicating heart attack) and variables such as cp (chest pain type), thalachh (maximum heart rate achieved), and slp (slope of the peak exercise ST segment).

Contrary to intuition, the distribution plot of age with respect to the likelihood of experiencing a heart attack does not show a clear trend, suggesting that age alone may not be a reliable indicator. However, individuals with higher maximum heart rates achieved, as depicted in the distribution plot of thalachh with respect to output, appear to have an increased likelihood of experiencing a heart attack. Additionally, those with lower previous peak achieved, as illustrated in the distribution plot of oldpeak with respect to output, exhibit a higher probability of experiencing a heart attack.

Further analysis reveals specific demographic and physiological factors associated with a heightened risk of heart attack. Individuals experiencing non-anginal chest pain (Chest pain type = 2), those with no major vessels affected (Number of major vessels = 0), males (sex = 1), individuals with thallium stress test result value of 2 (Thallium Stress Test result = 2), and those without exercise-induced angina (Exercise induced angina = 0) are all identified as having a significantly increased likelihood of experiencing a heart attack. These findings underscore the complex interplay of various factors in determining an individual's susceptibility to cardiovascular events and highlight the importance of considering multiple variables in assessing heart attack risk.

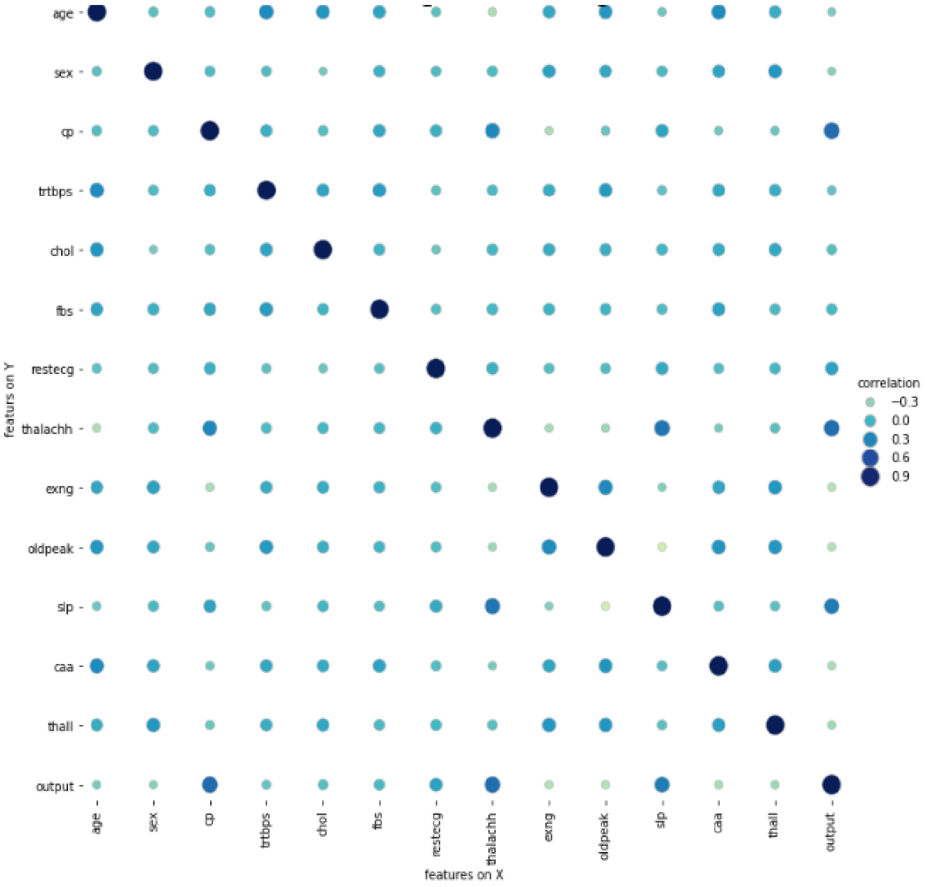


Figure 79 : Scatterplot heatmap of data frame

3.3 Discussion

The findings from our research provide valuable insights into the intricate interplay of various demographic and physiological factors in assessing cardiovascular risk. By analyzing a comprehensive dataset encompassing parameters such as age, sex, chest pain type, resting blood pressure, cholesterol levels, fasting blood sugar, electrocardiographic results, maximum heart rate achieved, previous peak, slope, number of major vessels affected, Thallium Stress Test results, and exercise-induced angina, we have been able to develop a predictive model with promising accuracy.

The importance of this research lies in its potential to revolutionize preventive healthcare strategies. By accurately identifying individuals at higher risk of cardiovascular events, healthcare professionals can intervene early with targeted interventions, including lifestyle modifications, medication regimens, and regular monitoring. This proactive approach not only reduces the burden on healthcare systems but also improves patient outcomes by mitigating the incidence and severity of cardiovascular diseases.

Furthermore, the incorporation of advanced classification algorithms, such as logistic regression, underscores the sophistication and reliability of our predictive model. The high accuracy achieved by our model demonstrates its robustness in risk assessment, providing both healthcare providers and patients with confidence in its predictive capabilities.

Moreover, the user-friendly interface of our model, accessible through mobile applications, ensures widespread adoption and seamless integration into clinical practice. This accessibility, coupled with stringent safeguards for patient data confidentiality and privacy, enhances patient engagement and trust in the healthcare system.

Our research not only validates the effectiveness of our predictive model but also signifies its potential to transform preventive cardiology by enabling early identification and intervention for individuals at risk of cardiovascular events. By harnessing the power of data analytics and machine learning, we have taken a significant step towards personalized and proactive healthcare, ultimately leading to improved patient outcomes and reduced cardiovascular morbidity and mortality.

04. CONCLUSION

our research represents a significant step forward in understanding and predicting cardiovascular risk. Through the analysis of a comprehensive dataset encompassing various demographic and physiological factors, we have identified key predictors associated with the likelihood of experiencing a heart attack. Our findings highlight the complex nature of cardiovascular risk assessment, with factors such as age, sex, chest pain type, and physiological indicators playing crucial roles.

Despite the absence of a straightforward linear correlation among continuous variables, our analysis revealed potential correlations between certain variables and the likelihood of a heart attack. Specifically, individuals with higher maximum heart rates achieved and lower previous peak values exhibit increased susceptibility to heart attacks, while factors such as non-anginal chest pain, the absence of major vessel involvement, male sex, specific thallium stress test results, and absence of exercise-induced angina are also associated with elevated risk.

These findings underscore the importance of considering multiple factors in assessing cardiovascular risk and designing effective preventive strategies. By leveraging predictive modeling techniques, such as logistic regression, we have developed a robust predictive model with high accuracy, paving the way for personalized risk assessment and early intervention.

Moving forward, our research has significant implications for preventive cardiology and public health. By identifying individuals at higher risk of cardiovascular events, healthcare professionals can implement targeted interventions aimed at reducing risk factors and improving patient outcomes. Additionally, our model's user-friendly interface and adherence to strict data privacy standards ensure its accessibility and trustworthiness among both healthcare providers and patients.

Overall, our research contributes to the growing body of knowledge aimed at reducing the burden of cardiovascular disease worldwide. Through continued research and innovation, we can further refine our predictive models and preventive strategies, ultimately leading to better cardiovascular health outcomes for individuals across diverse populations.

05. REFERENCES

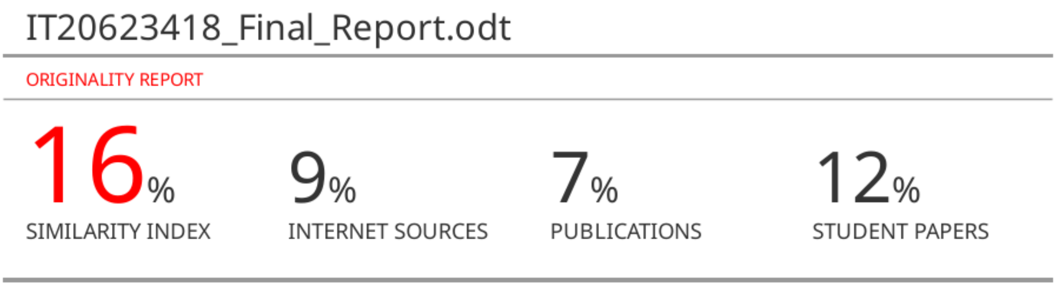
[1] D. S. C. Wikramasinghe and D. V. Kumarapeli, "National Guideline for Cardiovascular Risk Management (Total cardiovascular risk assessment approach) for Primary Health Care Providers," 2019. [Online]. Available: https://www.ncd.health.gov.lk/images/pdf/circulars/National_Gulidline_for_Cardiovascular_Risk_Management.pdf.

[2] M. SS, A. AW and A. ZI, "American Heart Association," 24 January 2024. [Online]. Available: https://www.heart.org/-/media/PHD-Files-2/Science-News/2/2024-Heart-and-Stroke-Stat-Update/2024-Statistics-At-A-Glance-final_2024.pdf.

06. GLOSSARY

1. Cardiovascular Risk: The likelihood of experiencing heart-related issues or diseases such as heart attack or stroke.
2. Demographic Factors: Characteristics of individuals such as age, sex, and socioeconomic status.
3. Physiological Factors: Biological indicators or measurements related to bodily functions, such as blood pressure, cholesterol levels, and heart rate.
4. Predictive Model: A mathematical algorithm or statistical tool used to predict future outcomes based on input variables.
5. Logistic Regression: A statistical method used to model the probability of a binary outcome, such as the presence or absence of a heart attack.
6. Chest Pain Type: Different classifications of chest pain, often indicative of various heart-related conditions.
7. Thallium Stress Test: A diagnostic test used to evaluate the blood flow to the heart during physical activity, helping to detect coronary artery disease.
8. Exercise-Induced Angina: Chest pain or discomfort experienced during physical exertion, typically due to reduced blood flow to the heart.
9. Risk Assessment: The process of evaluating potential risks or hazards to determine the likelihood and severity of adverse events.
10. Preventive Cardiology: The branch of cardiology focused on identifying and managing risk factors to prevent cardiovascular diseases.

07. APPENDICES



IT20623418_Final_Report.odt

ORIGINALITY REPORT

16%	9%	7%	12%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

Figure 80 : Appendices