**REAL TIME RESEARCH PROJECT REPORT**

**ON**

**“HEART HEALTH ANALYSIS AND PREDICTION”**

Submitted by

J Poojitha(23J41A66M1)

**BACHELOR OF TECHNOLOGY**

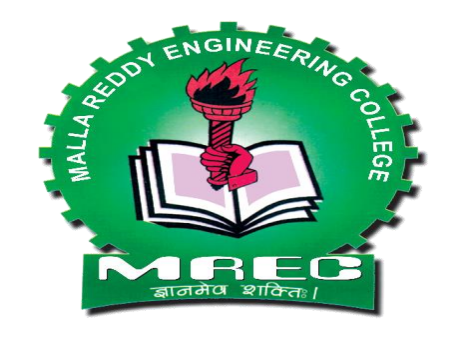
**In**

**COMPUTER SCIENCE AND ENGINEERING – AIML**

Under the Supervision of

**Dr. S Shiva Prasad**

Head of Department – Data Science



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING – DATA SCIENCE

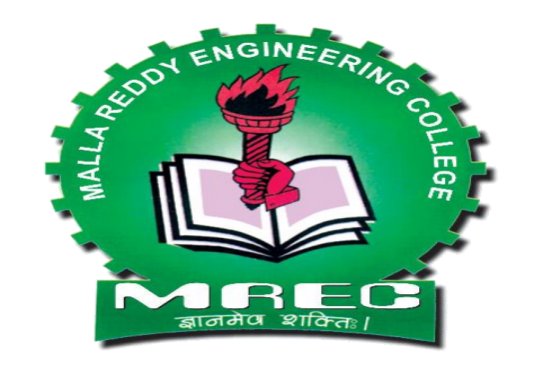
MALLA REDDY ENGINEERING COLLEGE

An UGC Autonomous Institution,(Approved by AICTE ,New Delhi & Affiliated to JNTUH,Hyderabad) Maisammaguda,Secunderabad,Telangana,India500100

**JULY-2024**

**MALLA REDDY ENGINEERING COLLEGE**

**Department of Computer Science and Engineering – Data Science**



CERTIFICATE

This is to certify that the project report entitled **“ Heart Health Analysis and Prediction”** is the bonafide record of project work carried out under my supervision by **K.Manisha** **(22J41A67G3)** during the academic year 2023-2024 for Bachelor of Technology in Computer Science and Engineering – Data Science of Malla Reddy Engineering College, Maisammaguda.The results embodied in this project report have not been submitted to any other University or Institute for the of any Degree or Diploma.

*Head of the Department Signature of Project Guide*

**Dr. S Shiva Prasad Dr. S Shiva Prasad**

HOD, Professor HOD,Professor

Department of Data Science Department of Dat Science

Malla Reddy Engineering College Malla Reddy Engineering College

**-----------------**

**External Examiner**

# 

# DECLARATION

I hereby declare that the project report entitled “**Heart Health Analysis and prediction**”. has been written by me and has not been submitted either in part or whole for the award of any degree, diploma or any other similar title to this or any other university.

K.Manisha

**22J41A67G3**

Date:

Place: Malla Reddy Engineering College

ACKNOWLEDGEMENT

It gives me a great sense of pleasure to acknowledge the assistance and cooperation I have received from several persons while undertaking this B.Tech.Second year Realtime Project Report. I owe special debt of gratitude to **Dr. S. Shiva Prasad** Department of Computer Science & Engineering-Data Science , for his constant support and guidance throughout the course of my work. His sincerity, thoroughness and perseverance have been a constant source of inspiration by me.

I also take the opportunity to acknowledge the contribution of **Prof. Dr.S.Shiva Prasad, HOD, Professor, Department** of Computer Science & Engineering (Data Science) , for his full support and assistance during the development of the project.

I also do not like to miss the opportunity to acknowledge the contribution of all faculty members of the department for their kind assistance and cooperation during the development of my project. Last but not the least, I acknowledge my friends for their contribution in the completion of the project.

I wish to express my thanks to god for always showering good vibes on me and finally thanks to our family for the love and affection overseas and forbearance and cheerful depositions, which are vital for sustaining effort, required for completing this work.

K.Manisha

22J41A67G3

disease are crucial medical tasks to ensure correct classification, which helps cardiologists provide proper treatment to the patient. Machine learning applications in the medical niche have increased as they can recognize patterns from data. Using machine learning to classify cardiovascular disease occurrence can help diagnosticians reduce misdiagnosis.. Models such as random forest (RF), decision tree classifier (DT), multilayer perceptron (MP), and XGBoost (XGB) are used. GridSearchCV was used to hypertune the parameters of the applied model to optimize the result.

The proposed model is applied to a real-world dataset of 70,000 instances from Kaggle. Models were trained on data that were split in 80:20 and achieved accuracy as follows: decision tree: 86.37% (with cross-validation) and 86.53% (without cross-validation), XGBoost: 86.87% (with cross-validation) and 87.02% (without cross-validation), random forest: 87.05% (with cross-validation) and 86.92% (without cross-validation), multilayer perceptron: 87.28% (with cross-validation) and 86.94% (without cross-validation).

The proposed models have AUC (area under the curve) values: decision tree: 0.94, XGBoost: 0.95, random forest: 0.95, multilayer perceptron: 0.95. The conclusion drawn from this underlying research is that multilayer perceptron with cross-validation has outperformed all other algorithms in terms of accuracy. It achieved the highest accuracy of 87.28%.

It manifests when the buildup of arterial plaque obstructs the circulation of blood to the heart or brain, potentially resulting in a stroke or heart attack. Early identification of heart disease is needed to reduce mortality rates and improve decision-making in the prevention and treatment of high-risk individuals.

Table of Contents

[Candidate’s Declaration i](#_TOC_250022)

Plagiarism Certificate ii

[Acknowledgement iii](#_TOC_250021)

[Table of Contents iv](#_TOC_250020)

[List of Figures v](#_TOC_250019)

[Abstract v](#_TOC_250017)i

1. INTRODUCTION 1
   1. [Introduction 1](#_TOC_250016)
      1. Recommendation system work mechanism 2
      2. Types of recommendation systems 3
      3. Genre based filtering… 4
      4. Types of Genre based algorithms 6
   2. [Problem Statement 7](#_TOC_250015)
   3. [Objectives 8](#_TOC_250014)
   4. [Methodology… 8](#_TOC_250013)
      1. Cosine similarity algorithm 9
      2. Libraries 11
      3. Kaggle 15
      4. API 16
      5. TMDB API 16
   5. [Organization 17](#_TOC_250012)
2. LITERATURE SURVEY 19
3. SYSTEM DESIGN & DEVELOPMENT… 24
   1. [Technologies Implementation 24](#_TOC_250011)
   2. [Functional Requirements 28](#_TOC_250010)
   3. [Non-functional Requirements 30](#_TOC_250009)
4. PERFORMANCE ANALYSIS……………………………………………….33

4.1 Design of problem statement……………………………………………..33

4.2 Algorithm / pseudo code of the project problem……………………..33

4.3 Screenshots of the various stages of the project………………………34

1. CONCLUSION 38
   1. Conclusions 38
   2. [Application of the Minor Project 3](#_TOC_250002)8
   3. [Future Scope 38](#_TOC_250001)

[References 40](#_TOC_250000)

**Chapter 1 . INTRODUCTION**

Cardiovascular diseases have been the most common cause of death worldwide over the last few decades in developed as well as underdeveloped and developing countries. Early detection of cardiac diseases and continuous supervision of clinicians can reduce the mortality rate. However, it is not possible to monitor patients every day in all cases accurately and consultation of a patient for 24 hours by a doctor is not available since it requires more sapience, time, and expertise.

Every day, the average human heart beats around 100,000 times, pumping 2,000 gallons of blood through the body. Inside your body, there are 60,000 miles of blood vessels. The signs of a woman having a heart attack are much less noticeable than the signs of a man. In women, heart attacks may feel uncomfortable squeezing, pressure, fullness, or pain in the center of the chest. It may also cause pain in one or both arms, the back, neck, jaw, stomach, shortness of breath, nausea, and other symptoms.

Men experience typical symptoms of heart attack, such as chest pain, discomfort, and stress. They may also experience pain in other areas, such as arms, neck, back, and jaw, and shortness of breath, sweating, and discomfort that mimics heartburn. It’s a lot of work for an organ which is just like a large fist and weighs between 8 and 12 ounces.

## 1.1 Motivation

In terms of motivation, many aspects contributed to direct us to the area of health and the use of physiological sensors to enhance health care. One major trigger that made us realize this, is the very important fact that we all dealing with saving lives and trying to increase the average life span of humans. Many heartbreaks and sorrows happen every day when family members pass away due to heart disease, there is no reason why anyone should not have the opportunity to live long healthy life surrounded by his loved ones because of insufficient monitoring of health. Next, we stress the importance of scenario that anyone would much rather avoid.

## 1.2 Problem Statement

Heart disease remains a significant global health challenge, with early detection crucial for effective intervention. This project aims to develop a predictive model using machine learning techniques to assess the risk of heart disease based on diverse health metrics and patient data. By analyzing factors such as demographics, medical history, lifestyle choices, and diagnostic results, the model seeks to provide accurate risk assessments. The goal is to empower healthcare professionals with a tool that enables early identification of individuals at high risk, facilitating timely intervention and personalized treatment strategies. Through comprehensive data analysis and model development, this project aims to improve proactive healthcare management and enhance patient outcomes.

## 1.3 Objectives

* Collect and prepare a comprehensive dataset including demographics, medical history, lifestyle factors, and clinical measurements relevant to heart health.
* Perform exploratory data analysis (EDA) to understand the distribution and relationships of key variables with respect to heart disease.
* Select and engineer features that significantly impact heart disease risk, using methods like correlation analysis and feature importance techniques.
* Develop predictive models using machine learning algorithms such as logistic regression, decision trees, random forests, SVM, and neural networks.
* Evaluate model performance using metrics like accuracy, precision, recall, F1-score, and AUC.
* Ensure model interpretability to provide insights into factors influencing heart disease risk.
* Deploy the model into clinical practice or healthcare systems for real-time risk.
* Monitor and maintain model performance to ensure accuracy and effectiveness over time.Top of FormBottom of Form

## 1.4 Scope

**Data Collection:** Gather diverse datasets including demographics, medical history, lifestyle factors, and clinical measurements related to heart health.

**Risk Prediction:** Develop models to predict the likelihood of heart disease based on collected data, aiming for early detection and intervention.

**Feature Selection:** Identify key factors influence heart health outcome statistical analysis and machine learning techniques.

**Model Implementation:** Utilize predictive algorithms to build robust models for assessing heart disease risk, ensuring accuracy and reliability.

**Application:** Deploy models in clinical settings to assist healthcare professionals in making informed decisions and improving patient care strategies.

**CHAPTER 2 : LITERATURE REVIEW**

The literature review in a heart disease prediction study serves to provide a comprehensive overview of existing research, methodologies, and findings related to the prediction of heart disease. It helps establish context for the current study, identifies gaps in the literature, and justifies the need for the new research. Here's a structured approach for a literature review in the context of heart disease prediction. By following this structured approach, the literature review becomes a cohesive narrative that informs the reader about the state of the art in heart disease prediction, establishes the foundation for the current study, and provides a roadmap for further exploration in the field. With the rise of deep learning, Convolutional Neural Networks (CNNs) and Recurrent Neural Networks . have been utilized to analyze medical images such as echocardiograms or ECGs to predict heart diseases. In a 2018 study, Acharya et al. employed CNNs to classify ECG beats to detect cardiac arrhythmias, showcasing deep learning's potential for heart disease prediction.

Bo Jin, Chao Che et al. (2018) proposed a “Predicting the Risk of Heart Failure With EHR Sequential Data Modeling” model designed by applying neural network. This paper used the electronic health record (EHR) data from real-world datasets related to congestive heart disease to perform the experiment and predict the heart disease before itself.

Aakash Chauhan .(2018) presented “Heart Disease Prediction using Evolutionary Rule Learning”. This study eliminates the manual task that additionally helps in extracting the information (data) directly from the electronic records. To generate strong association rules, we have applied frequent pattern growth association mining on patient’s dataset. This will facilitate (help) in decreasing the amount of services and shown that overwhelming majority of the rules helps within the best prediction of coronary sickness. [2]. Ashir Javeed, Shijie Zhou et al. (2017) designed “An Intelligent Learning System based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection”. This paper uses random search algorithm (RSA) for factor selection and random forest model for diagnosing the cardiovascular disease. This model is principally optimized for using grid search algorithmic program.

**2.1 Related data**

The detection and prediction of heart disease an important issue. For this reason, a lot of work has been done in this area. We divide existing work into two categories: The first presents approaches that select the most relevant patient by features selection, and the second is to explore the learning algorithms that offers high accuracy.

**2.2 Feature Selection**

The choice of features is one of the major challenges to train a predictive learning algorithm. Several studies have been carried out and we present the most recent works in the literature. The authors made a kind of comparison of the different Machine Learning algorithms on two classes of attributes. The first one contains 10 attributes and the second one contains 14 attributes. The problem with this study is that the authors did not mention the attributes treated. The authors of developed the same approach as but with only six attributes (Age, Sex, Using Machine Learning for Heart Disease Prediction Blood Pressure, Heart Rate, Diabetes, Hyper cholesterol

In recent years, the healthcare industry has seen a significant advancement in the field of data mining and machine learning. These techniques have been widely adopted and have demonstrated efficacy in various healthcare applications, particularly in the field of medical cardiology. The rapid accumulation of medical data has presented researchers with an unprecedented opportunity to develop and test new algorithms in this field. Heart disease remains a leading cause of mortality in developing nations and identifying risk factors and early signs of the disease has become an important area of research. The utilization of data mining and machine learning techniques in this field can potentially aid in the early detection and prevention of heart disease.

The purpose of the study described by Narain et al. (2016) is to create an innovative machine-learning-based cardiovascular disease (CVD) prediction system in order to increase the precision of the widely used Framingham risk score (FRS). With the help of data from 689 individuals who had symptoms of CVD and a validation dataset from the Framingham research, the proposed system—which uses a quantum neural network to learn and recognize patterns of CVD—was experimentally validated and compared with the FRS. The suggested system’s accuracy in forecasting CVD risk was determined to be 98.57%, which is much greater than the FRS’s accuracy of 19.22% and other existing techniques. According to the study’s findings, the suggested approach could be a useful tool for doctors in forecasting CVD risk, assisting in the creation of better treatment plans, and facilitating early diagnosis.

The primary drawback of the prior research is its limited dataset, resulting in a high risk of overfitting. The models developed may not be appropriate for large datasets. In contrast, we utilized a cardiovascular disease dataset consisting of 70,000 patients and 11 features, thereby reducing the chance of overfitting.FIG.1 presents a concise review of cardiovascular disease prediction studies performed on large datasets, further reinforcing the effectiveness substantial dataset.

**2.3 About Data Set**

* This is a multivariate type of dataset which means providing or involving various mathematical or statistical variables, and multivariate numerical data analysis.
* It is composed of 14 attributes which are age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise-induced angina, old peak-ST depression induced by exercise relative to rest, the slope of the peak exercise ST segment, number of major vessels and Thalassemia.
* This database includes 76 attributes, but all published studies relate to using a subset of 14 of them. One of the major tasks of this dataset is to predict based on the attributes of a patient whether that particular person has heart disease or not. The other is the experimental task to diagnose.

# Chapter 3. METHODOLOGY

## 3.1 Data Collection

A comprehensive dataset that includes relevant features such as age, gender, cholesterol levels, blood Collect pressure, family history, lifestyle habits, and medical history.

**3.2 Data Preprocessing**

Clean the data by handling missing values, outliers, and duplicates. Normalize or standardize numerical features. Encode categorical variables using techniques like one-hot encoding. Split the data into a training set, a validation set, and a test set.

**3.3 Future Selection and Engineering**

Identify relevant features through exploratory data analysis (EDA) and domain knowledge. Create new features if necessary. Apply dimensionality reduction techniques like Principal Component Analysis (PCA) if the dataset is large.

**3.4 Model Evaluation**

Assess the models using appropriate evaluation metrics, such as accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUC-ROC). Consider the clinical relevance of these metrics in the context of heart disease prediction.

**3.5 Future Research**

Identify potential areas for future research and improvements in heart disease prediction. Remember that research in healthcare and predictive modeling should prioritize patient safety and ethical considerations. Ensure that your research adheres to relevant regulations and guidelines for medical data use.

**3.6 Data Source**

The dataset utilized in this study, as described in , comprises 70,000 patient records with 12 distinct features, as listed in . These features include age, gender, systolic blood pressure, and diastolic blood pressure. The target class, “cardio,” indicates whether a patient has cardiovascular disease (represented as 1) or is healthy (represented as 0)

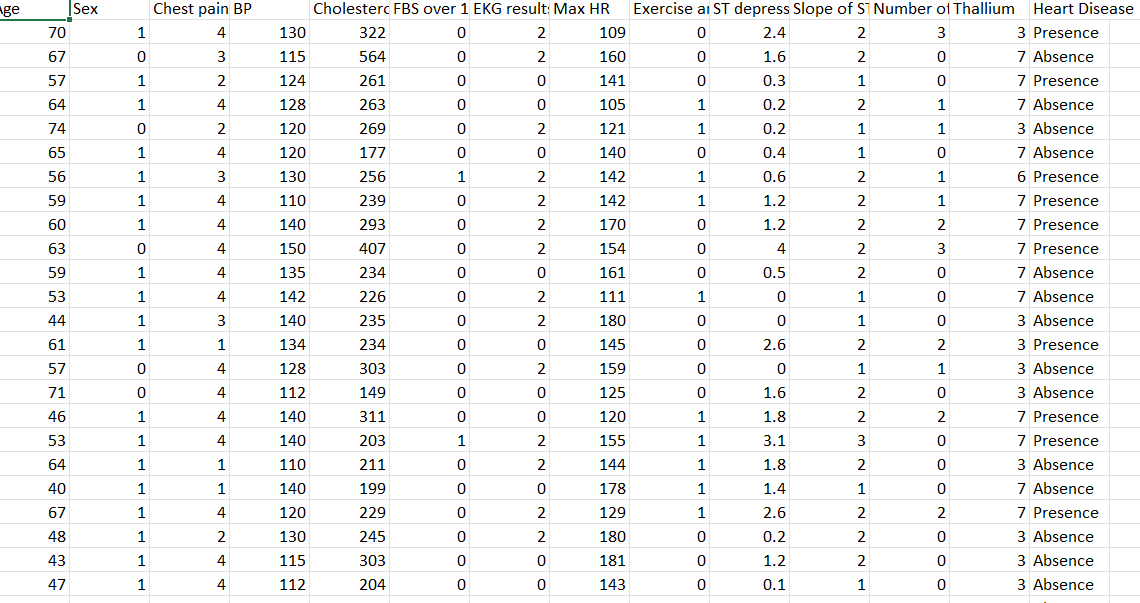
****

Fig 1:Data source

**3.6.1 Data Splitting**

Data is splitted into training and testing data.25 % data is used for testing purpose while75 % data is used for training purpose. We performed data normalization for removing nan values.

**3.7 Feature Model**

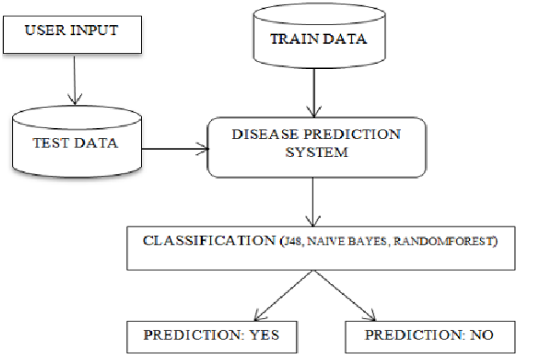
****

Fig 2: Data base design

**3.8 Model Evaluation and Validation**

* **Cross-Validation:** Dividing data into subsets for training and testing, ensuring models generalize well to unseen data.
* **Performance Metrics:** Assessing model performance using metrics such as accuracy, precision, recall, F1-score, and area under the ROC curve (AUC).

**Chapter 4.PROPOSED SYSTEM**

**4.1 Data Integration**

Integrate diverse datasets including demographics, medical history, lifestyle factors and clinical measurements relevant to heart health.

**4.2 Machine Learning Models**

Develop and optimize machine learning algorithms such as logistic regression, decision trees, random forests, and neural networks for accurate prediction of heart disease risk.

**4.3 Real-time Prediction**

Enable the system to provide real-time risk assessment and predictions, integrated into clinical workflows to support timely interventions.

**4.4 Interpretability**

Ensure that the models provide explainable results, allowing healthcare professionals to understand the rationale behind risk predictions.

**4.5 Security and Privacy**

Implement robust security measures to protect patient data and comply with healthcare regulations.

**4.6 Maintenance and Updates**

Establish procedures for ongoing maintenance, updates, and improvements based on new medical research, technological advancements, and feedback from healthcare professionals.

**4.7 Real Life Applications**

**4.7.1 Clinical Decision Support**:

Healthcare professionals can use these models as decision-support tools during patient consultations. By inputting patient data into the model, clinicians can obtain risk scores and recommendations for further evaluation or treatment.

**4.7.2 Public Health Initiatives**:

Public health authorities can utilize predictive models to identify populations a high risk of heart disease and implement targeted prevention strategies, such as educational campaigns, screening programs, or policy interventions.

**4.7.3 Remote Monitoring**:

Remote monitoring devices equipped with heart disease prediction algorithms can continuously monitor individuals at risk and alert them or their caregivers of any significant changes or warning signs, enabling timely medical intervention.

**4.7.4 Personalized Medicine**:

Predictive models can facilitate the shift towards personalized medicine by enabling healthcare providers to tailor treatment plans based on an individual's risk profile and genetic position to heart disease.

**4.8 How Data Help In Business**

**4.8.1 Healthcare Providers**:

Hospitals and clinics can use these models to assess the risk of heart disease in patients during routine check-ups. This can lead to early detection and intervention, ultimately improving patient outcomes and reducing healthcare costs.

**4.8.2 Insurance Companies**:

Insurance companies can utilize these models to assess the risk of heart disease in their policyholders. By identifying high-risk individuals, they can offer targeted interventions or wellness programs to mitigate the risk and claims.

**4.8.3 Pharmaceutical Companies**:

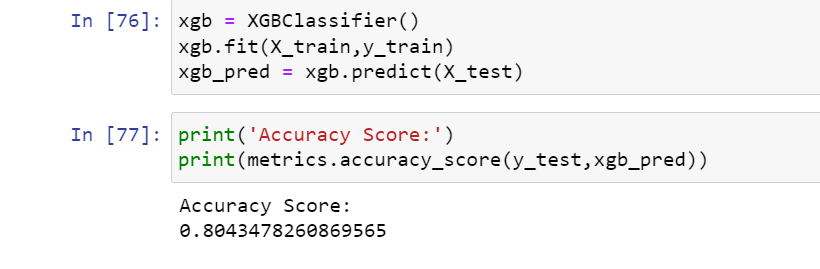
Pharmaceutical companies can use predictive models to identify potential candidates for clinical trials of new drugs aimed at preventing or treating heart disease. This can streamline the drug development process and bring new treatments to market more efficiently.

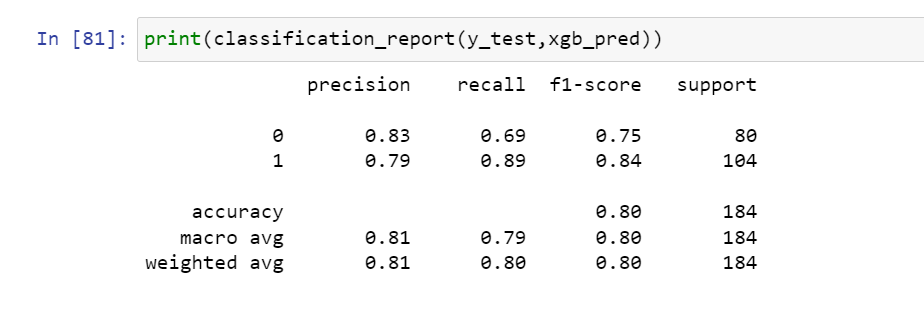
**4.8.4 Healthtech Startups**:

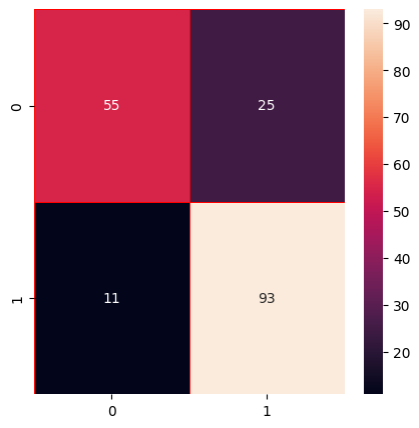
Startups focused on digital health and wellness can develop applications or wearable devices that utilize heart disease prediction models to provide personalized health recommendations to users. This can empower individuals to take proactive steps toward preventing heart disease.

**4.9 Algorithms Used**

**1.XGBOOST Classifier**

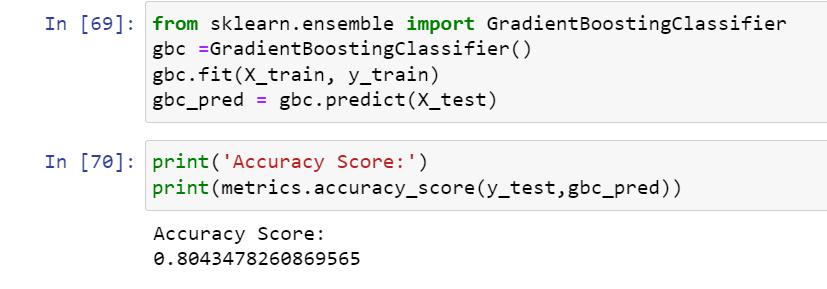


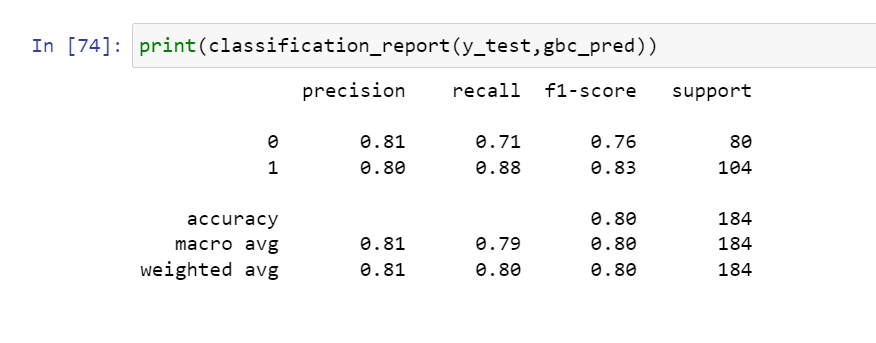


Fig 3:confusion matrix

After applying the Xgboost classifier the confusion matrix True positive and True Negative has increased from the previous model.

**2.Gradient Boosting Classifier**





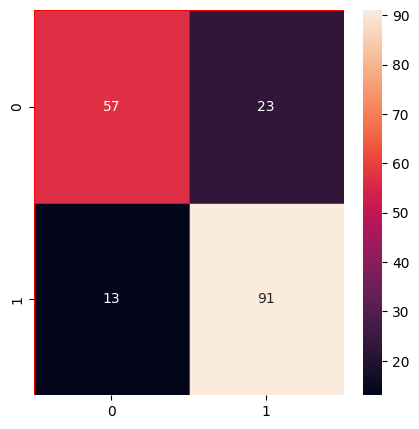


Fig 4

Gradient Boosting has performed better than all models till now with an accuracy of **80.43**%.

**Chapter 5.MODEL AND TRAINING**

**5.1 Correlation Matrix Missing of missing values**

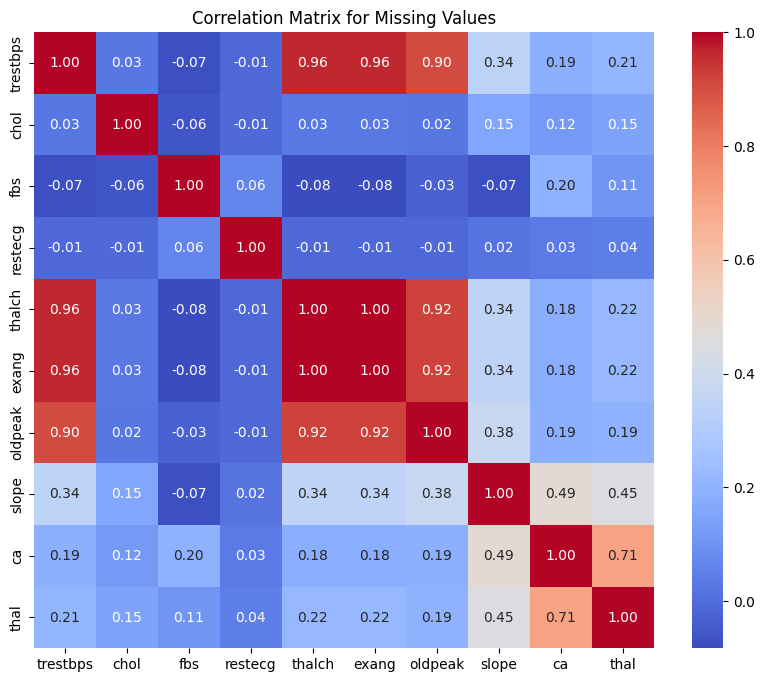


Fig 5

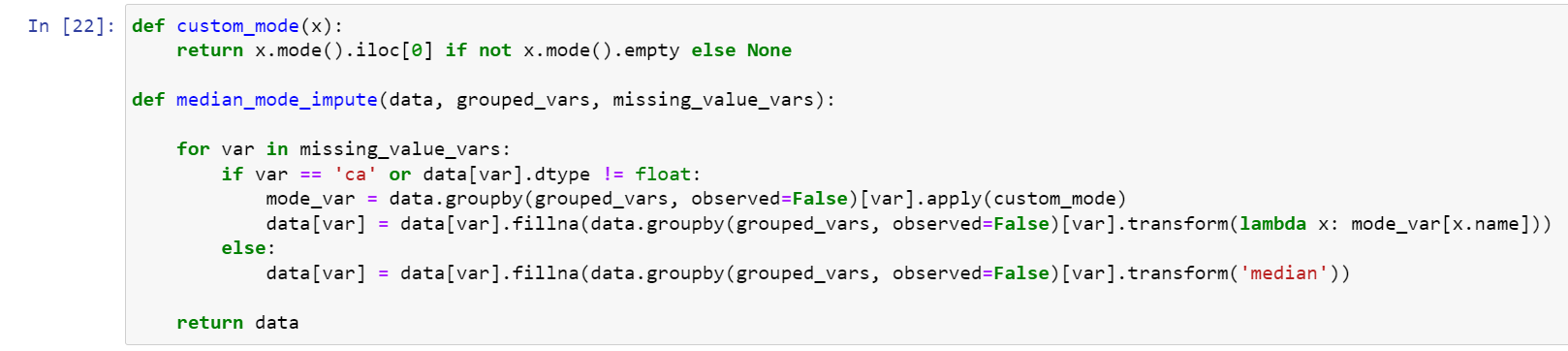
From the heatmap above, we observe a strong relationship of missing values between t halch and t restbps, exang and trestbps, oldpeak and t restbps etc.

Once again, the pattern of missing values among variables does not appear random.

As we mentioned above, the dataset includes 15 variables. However, at least variables have missing values.

Hence, we will apply 2 imputation methods (Median/Mode imputation and Random Forest imputation) to fill in the missing values.

**5.1.1 Imputing The Missing Values**



We will start by trying the simplest imputation method, which is Median/Mode Imputation, to fill in missing values .

we will fill in the missing values by inputting the median value if the feature is numerical. For categorical features, we will use the mode value to replace the missing values.

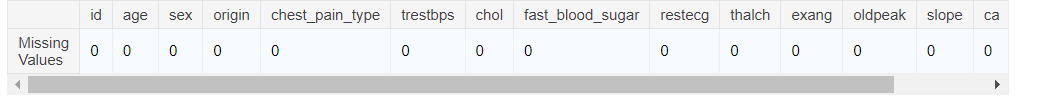
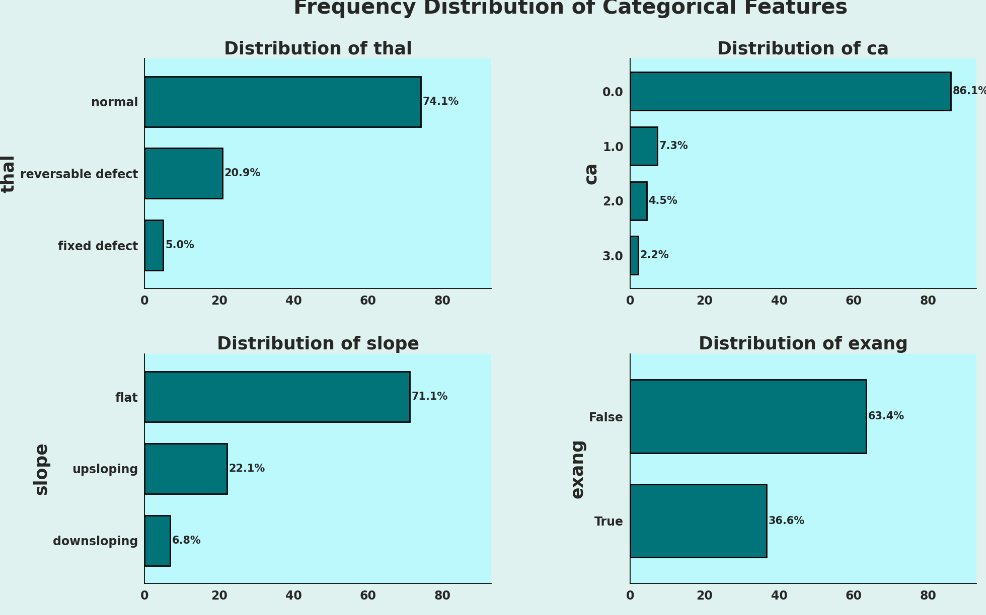


Fig 6

**5.2 Frequency Distribution Of Categorial Features**

****

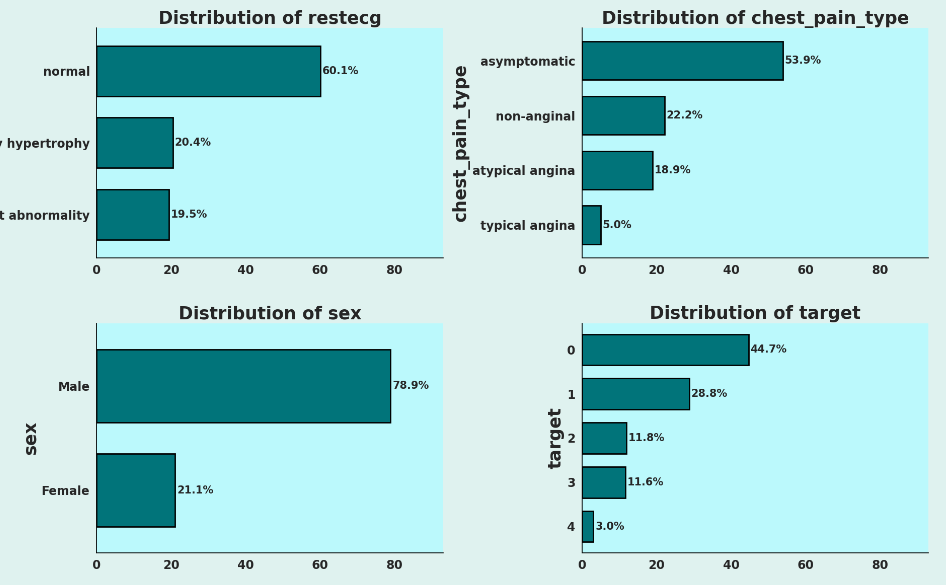
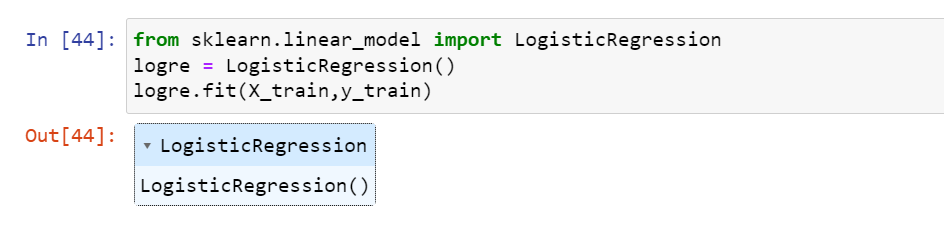
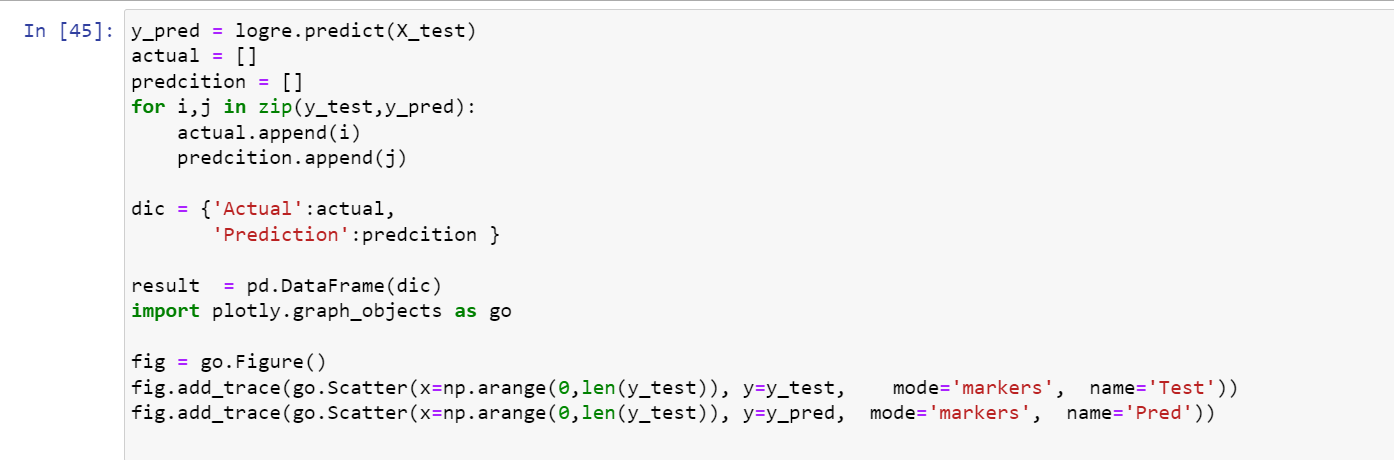


Fig 7

**5.2.1 Machine Learning Model**

**Logistic Regression**





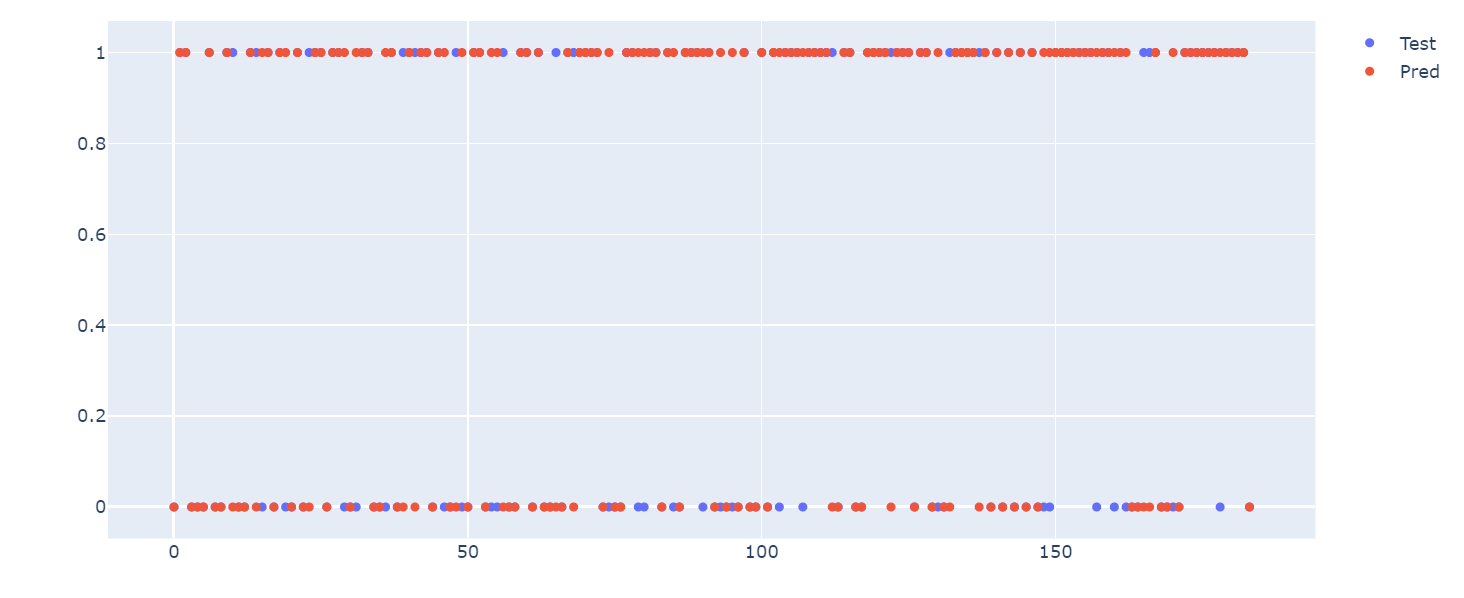
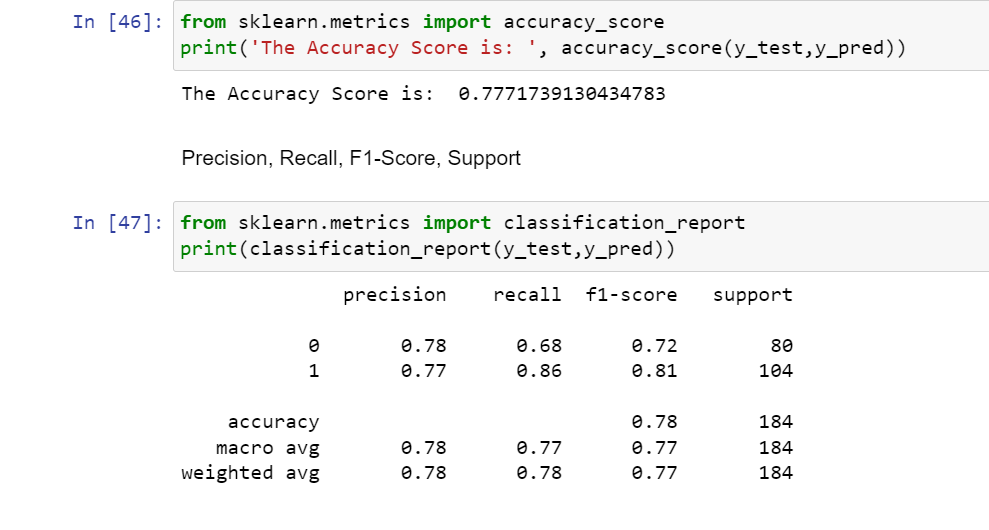


Fig 8

In the above figure, the red dots represent the predicted values that are either 0 or 1 and the blue line & and dot represent the actual value of that particular patient. In the places where the red dot and blue dot do not overlap are the wrong predictions and where both dots overlap those are the right predicted values.

**5.2.2 Model Evaluation**



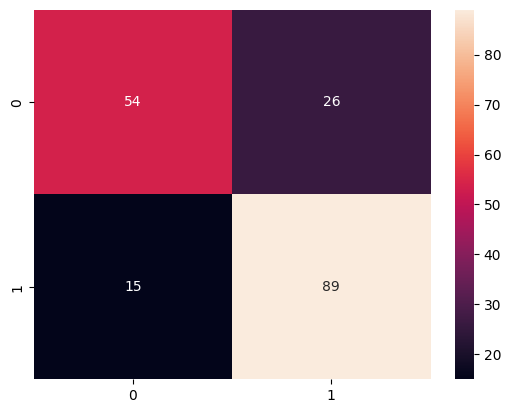


Fig 9

The logistic regression has given an accuracy of **77.71**%.From the confusion matrix, we can say the model can classify whether the disease is present or not. But

False Positives and False Negatives are also high to reduce this we will fit another classification model.

A ROC curve, or receiver operating characteristic curve, is like a graph that shows how well a classification

model performs.

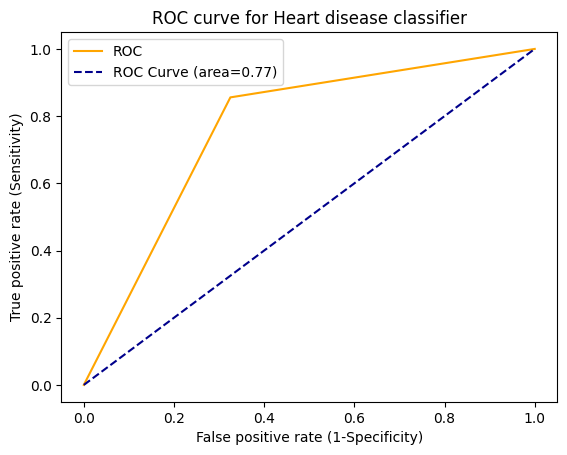
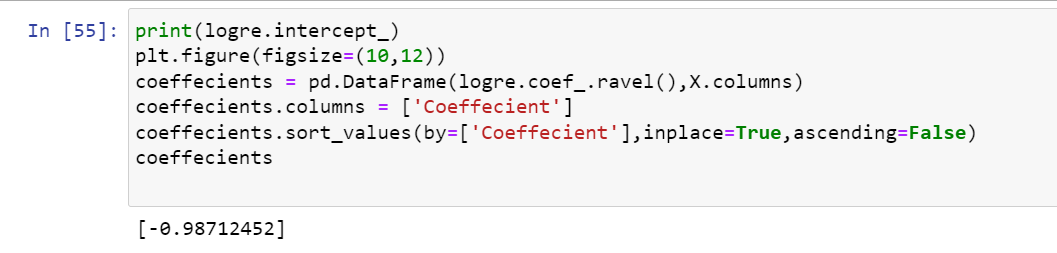


Fig 10

**5.2.3 Coefficients**



Linear Regression calculates the total outcome by summing up the weighted sum of the different features.

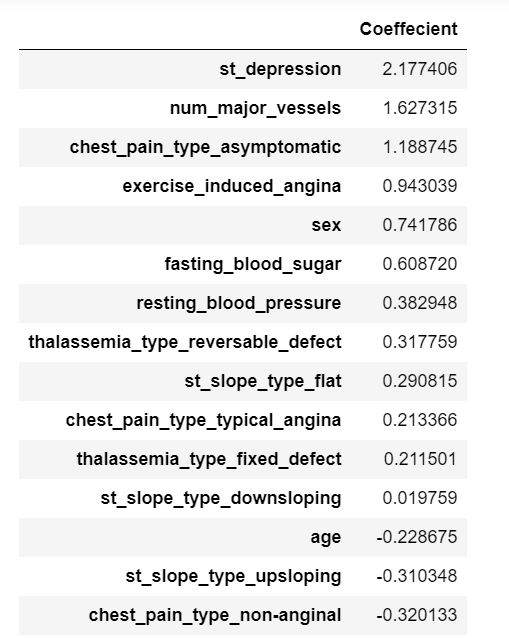


Fig 11

**5.2.4 Heart Disease Prevalence By Sex**

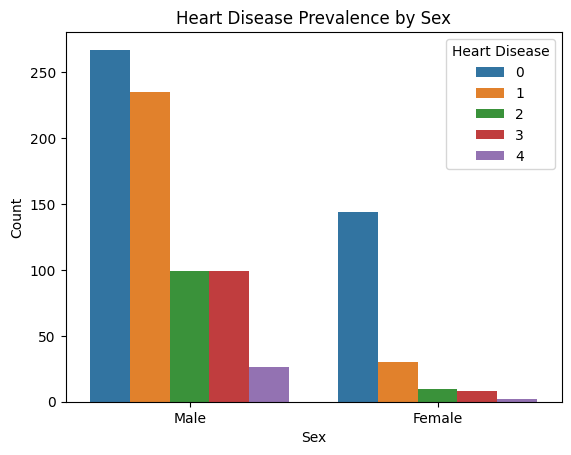


Fig 12

We can notice that men are more susceptible to heart disease at all levels.

**5.2.5 Relationship Between Cholestrol Levels and Heart Disease**

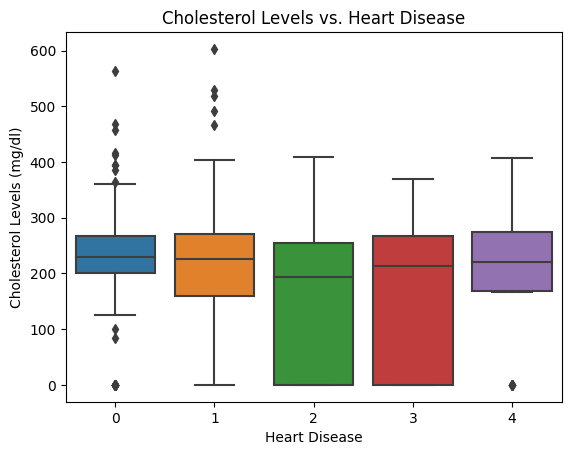


Fig 13

The box plot illustrates cholesterollevels across five heart diseasecategories, showing medianvalues, range variability, and outliers.

Categories 1 to 4 have similar medians, but the spread and outliers differ, with category

showing the most variability.

**5.3 Maximum Heart Rate And Heart Disease**

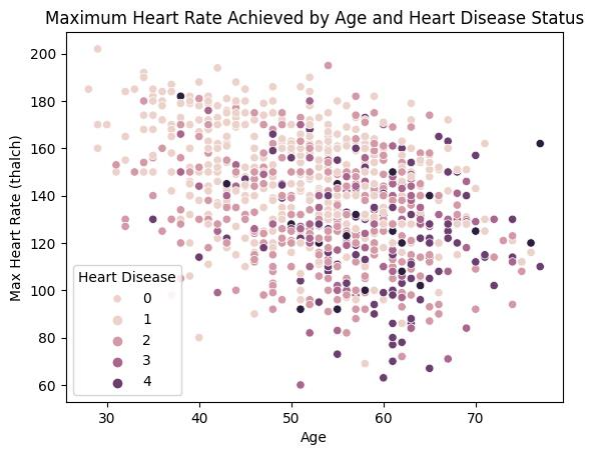


Fig 14

he plot shows a **negative correlation** where the maximum heart rate tends to decrease as age increases.

**5.3.1 Distribution of chest pain among patients**

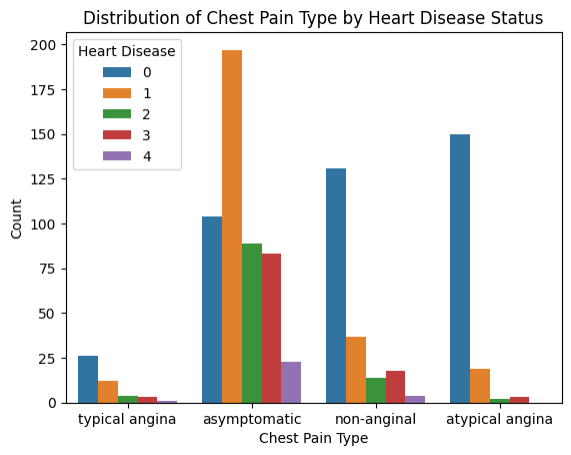


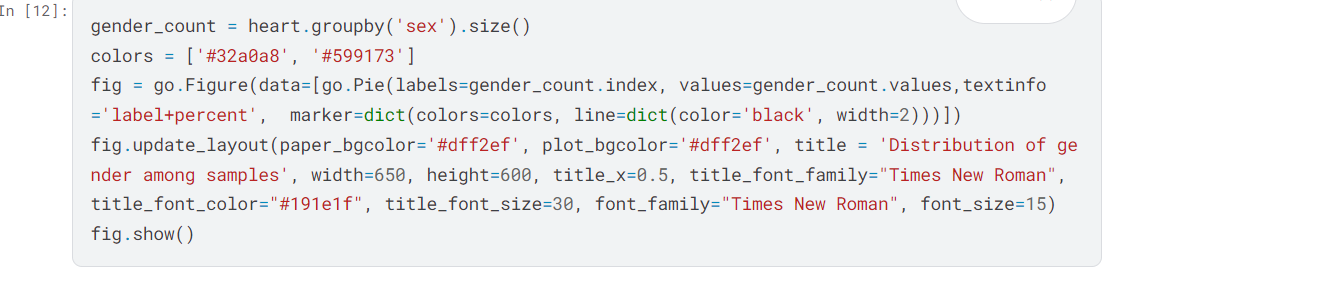
Fig 15

**Asymptomatic** is the most common type of chest pain across all heart

disease statuses except for status 0, where 'typical angina' is more prevalent.

**Non-anginal** pain is notably frequent in heart disease status 4, while 'atypical angina' is relatively less common across all states.

**5.3.2 Pie Chart of Gender Distribution**

****

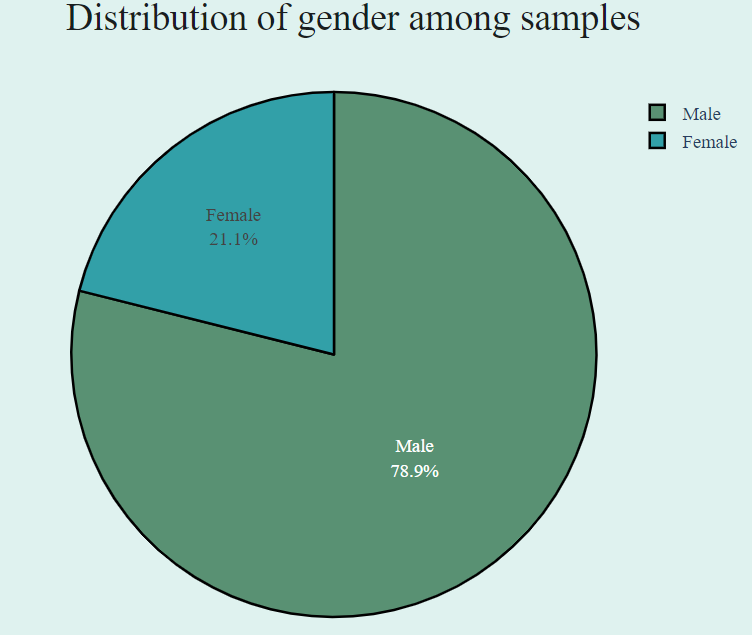
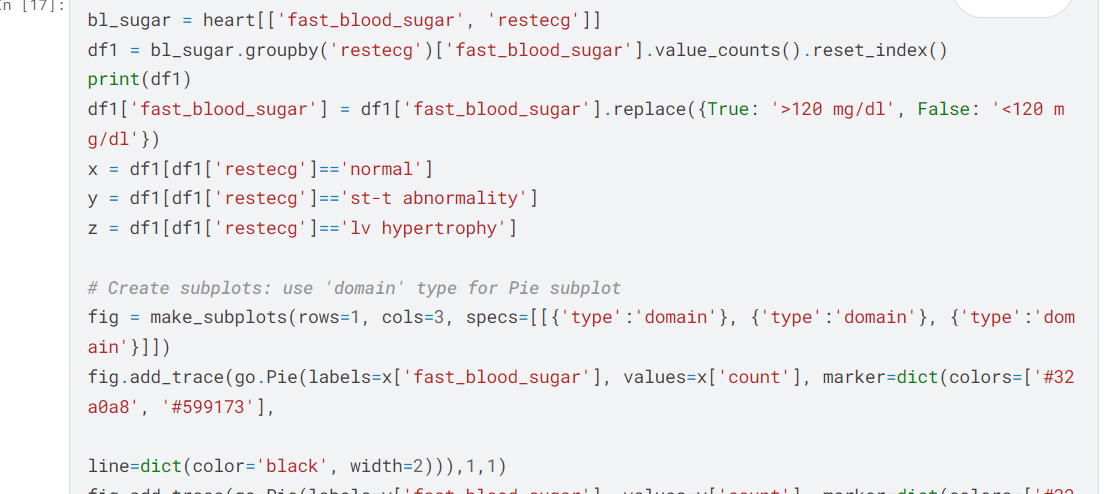
****

Fig 16

**5.3.3 Donut Chart of Fasting Sugars**

****

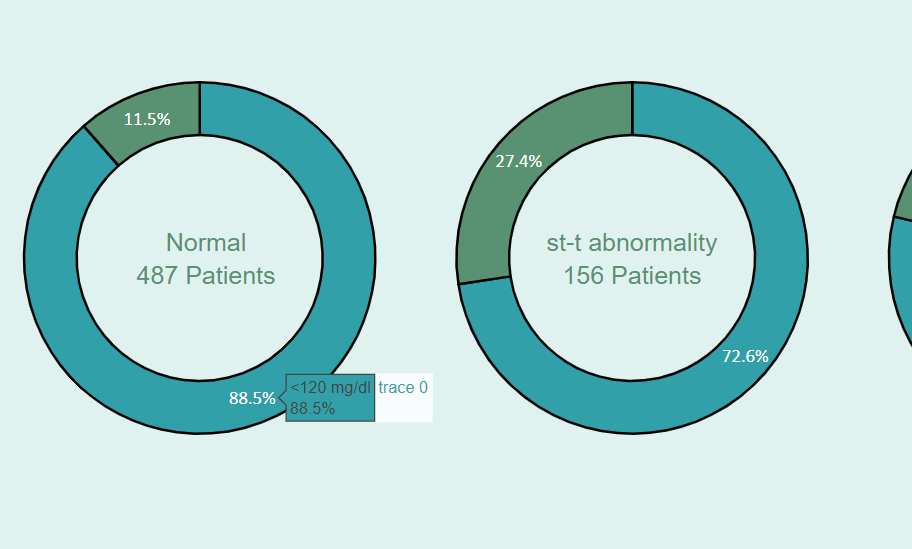
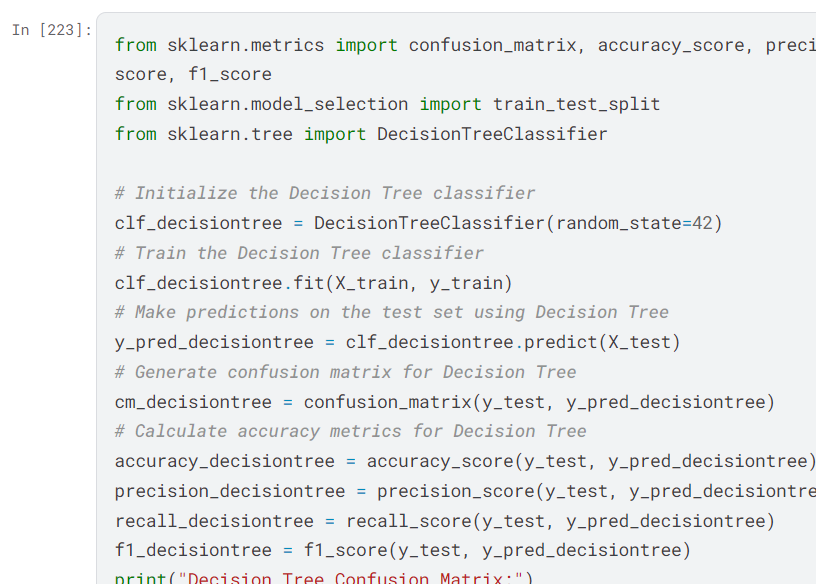
****

Fig 17

**5.4 Decision Tree Confusion Matrix**

****

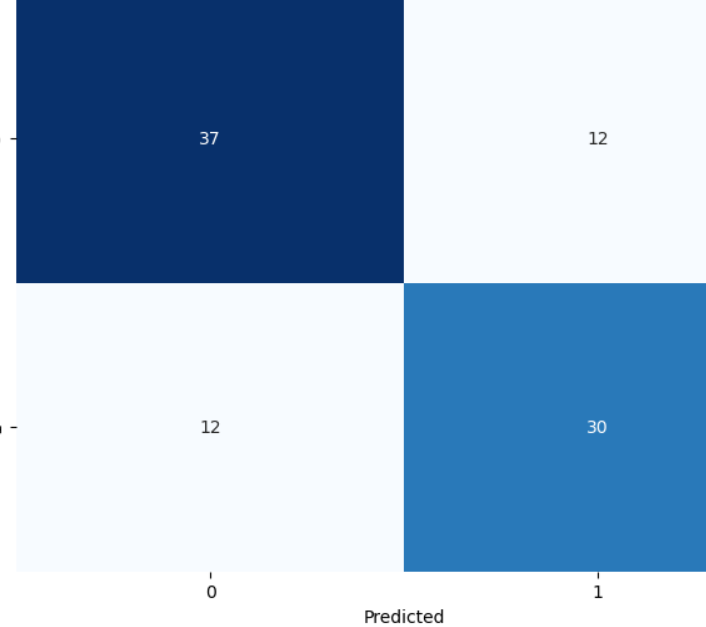
****

Fig 18

**5.4.1 SVM RBF Kernel Confusion Matrix**

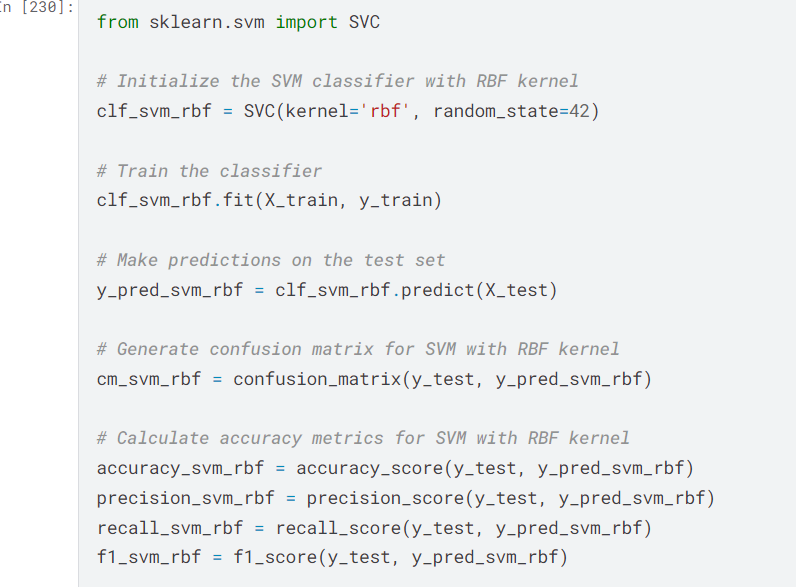
****

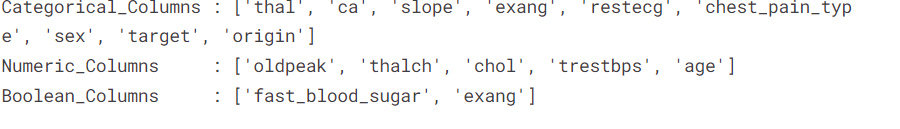
Fig 19

**A screenshot of a graph

Description automatically generated**

**5.5 Separating Features On Data Types**

****



**5.6 Pie Chart Of States Data**

****

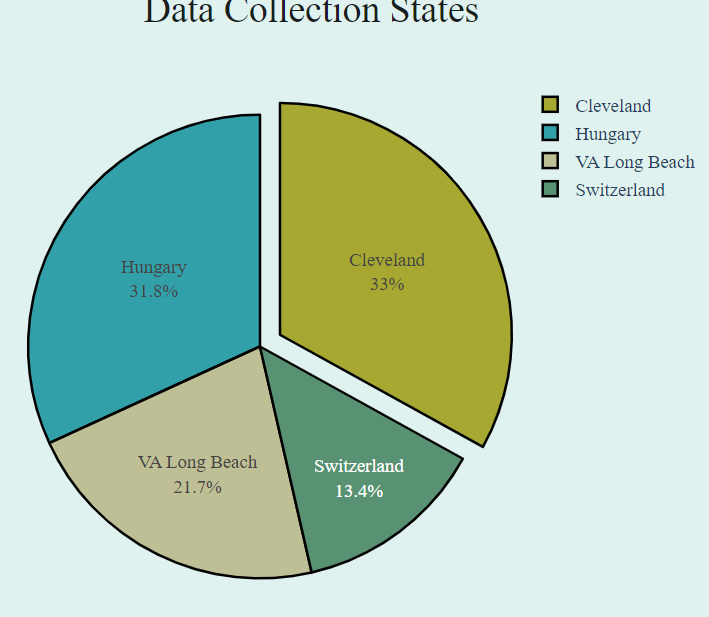
****

Fig 20

**5.7 Checking Outliers**

****

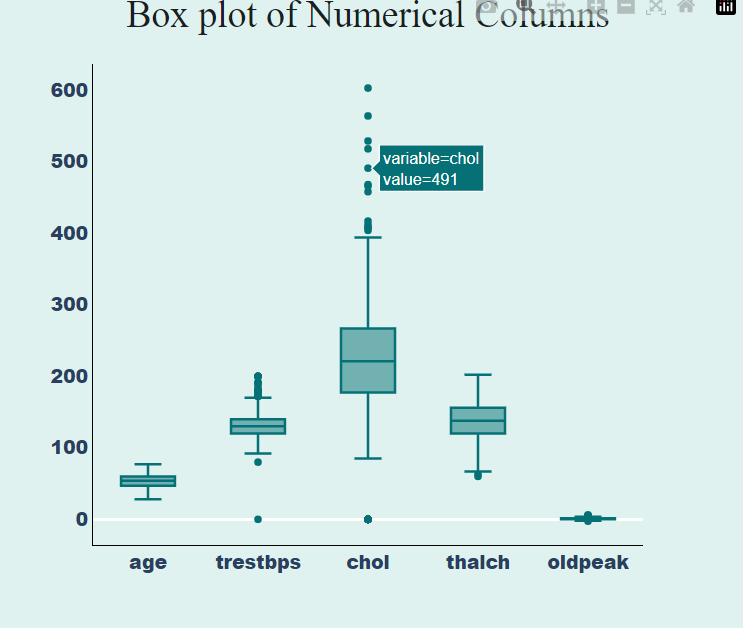
****

Fig 21

**FUTURE SCOPE**

Future research could focus on addressing the limitations of this study by comparing the performance of the k-modes clustering algorithm with other commonly used clustering algorithms, such as k-means or hierarchical clustering [**[36](https://www.mdpi.com/1999-4893/16/2/88" \l "B36-algorithms-16-00088)**], to gain a more comprehensive understanding of its performance. Additionally, it would be valuable to evaluate the impact of missing data and outliers on the accuracy of the model and develop strategies for handling these cases. Furthermore, it would be beneficial to evaluate the performance of the model on a held-out test dataset in order to establish its generalizability to new, unseen data. Ultimately, future research should aim to establish the robustness and generalizability of the results and the interpretability of the clusters formed by the algorithm, which could aid in understanding the results and support decision making based on the study’s findings.

**C0NCLUSION**

The primary objective of this study was to classify heart disease using different models and a real-world dataset. The k-modes clustering algorithm was applied to a dataset of patients with heart disease to predict the presence of the disease. The dataset was preprocessed by converting the age attribute to years and dividing it into bins of 5-year intervals, as well as dividing the diastolic and systolic blood pressure data into bins of 10 intervals. The dataset was also split on the basis of gender to take into account the unique characteristics and progression of heart disease in men and women.

The elbow curve method was utilized to determine the optimal number of clusters for both the male and female datasets. The results indicated that the MLP model had the highest accuracy of 87.23%. These findings demonstrate the potential of k-modes clustering to accurately predict heart disease and suggest that the algorithm could be a valuable tool in the development of targeted diagnostic and treatment strategies for the disease. The study utilized the Kaggle cardiovascular disease dataset with 70,000 instances, and all algorithms were implemented on Google Colab. The accuracies of all algorithms were above 86% with the lowest accuracy of 86.37% given by decision trees and the highest accuracy given by multilayer perceptron, as previously mentioned.

Despite the promising results, there are several limitations that should be noted. First, the study was based on a single dataset and may not be generalizable to other populations or patient groups. Furthermore, the study only considered a limited set of demographic and clinical variables and did not take into account other potential risk factors for heart disease, such as lifestyle factors or genetic predispositions. Additionally, the performance of the model on a held-out test dataset was not evaluated, which would have provided insight on how well the model generalizes to new, unseen data. Lastly, the interpretability of the results and the ability to explain the clusters formed by the algorithm was not evaluated. In light of these limitations, it is recommended to conduct further research to address these issues and to better understand the potential of k-modes clustering for heart disease prediction.

**REFERENCES**

1.Babu, S., Vivek, E., Famina, K., Fida, K., Aswathi, P., Shanid, M., Hena, M.: Heart disease diagnosis using data mining technique. In: 2017 International conference Using Machine Learning for Heart Disease Prediction 13 of Electronics, Communication and Aerospace Technology (ICECA).

2. Cai, J., Luo, J., Wang, S., Yang, S.: Feature selection in machine learning: A new perspective. Neurocomputing 300, 70–79 (2018).

3. Dangare, C.S., Apte, S.S.: Improved study of heart disease prediction system using data mining classiﬁcation techniques. International Journal of Computer Applications 47(10), 44–48 (2012).

4. Fang, X., Hodge, B.M., Du, E., Zhang, N., Li, F.: Modelling wind power spatial-

temporal correlation in multi-interval optimal power ﬂow: A sparse correlation

matrix approach. Applied energy 230, 531–539 (2018).