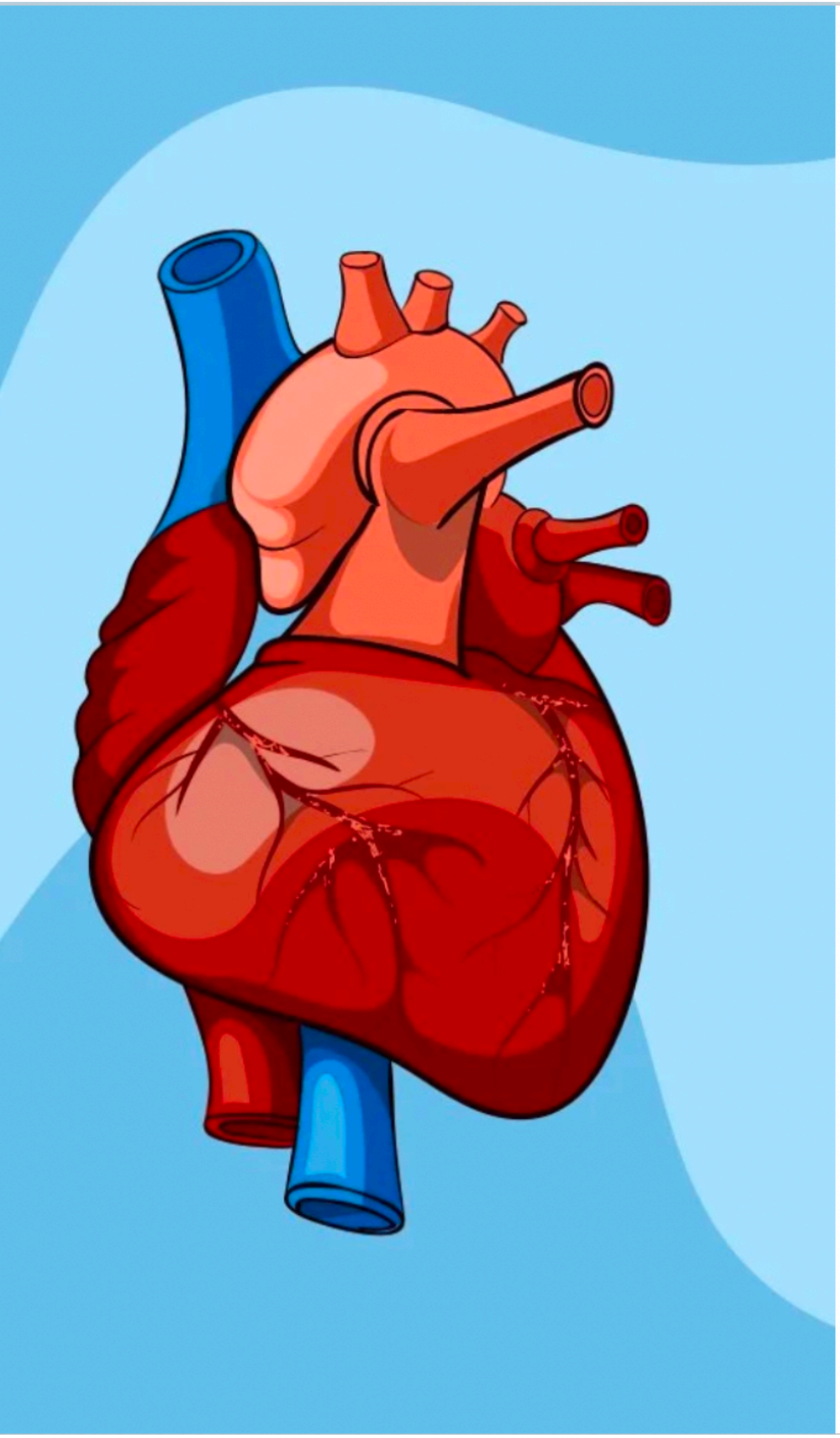


FINAL PROJECT REPORT

GROUP - 38

GROUP MEMBERS:

Rishabh Kanodiya
Simantini Ghosh
Atharv Hejib
Devansh Pratap Singh

1. Problem Statement and Motivation

- Developed machine learning models to predict heart disease based on health measurements and indicators, utilizing a dataset from Kaggle.
- The motivation is to leverage machine learning to improve diagnostic accuracy, assisting medical professionals in timely and effective patient management.
- The impact lies in enhancing patient outcomes and facilitating targeted medical interventions.
- This heart disease dataset from Kaggle Machine Learning Repository provides an opportunity to apply machine learning methods to an important medical domain.
- Heart disease is one of the leading causes of death globally, so being able to accurately diagnose patients and determine who is at higher risk could greatly benefit medical professionals.

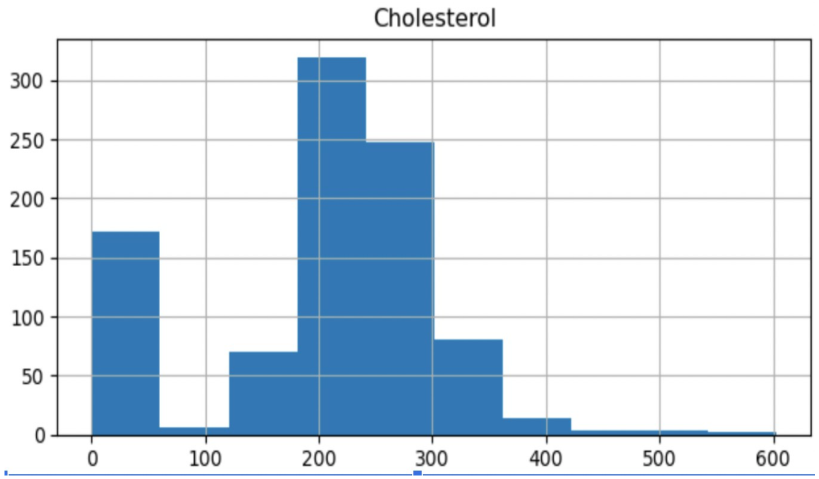
2. Code Explanation/Analysis

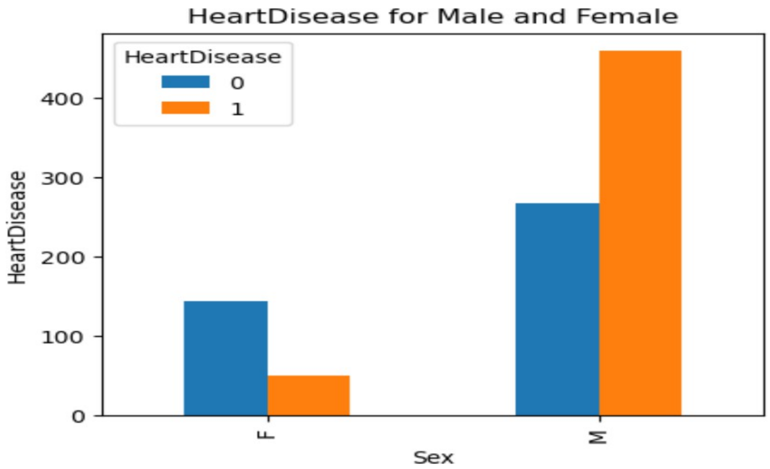
◆ Data Extraction and Feature Engineering

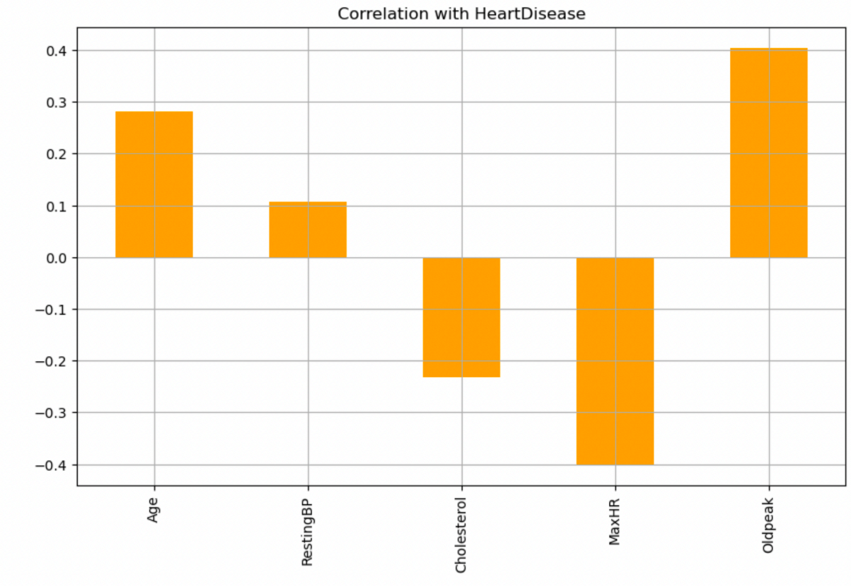
- Our dataset has 12 features which include age, sex, chest pain type, Cholestrol, MaxHR etc of 918 patients.
- Initially, we organized the complete dataset into halves based on patient demographics and medical history, with the patient_id serving as a foreign key. This approach allowed us to gather comprehensive information about each patient before determining whether they have a heart disease.
- We structured the data into two separate schemas: one for patient demographics and another for medical history tests and reports, forming distinct tables. Subsequently, we populated the respective tables by distributing the dataset both row-wise and column-wise.
- To consolidate the data, we created a unified dataframe through an inner join, combining information from both tables. We then split the entire dataset into the predictor variables (X) using the iloc function and isolated the response variable into (y).
- To handle categorical data, we employed one-hot encoding using the ColumnTransformer. This transformation facilitated the conversion of categorical variables into a format suitable for machine learning analysis.
- We applied Random Forest and Logistic Regression methods for predicting if a patient has Heart Disease or not.

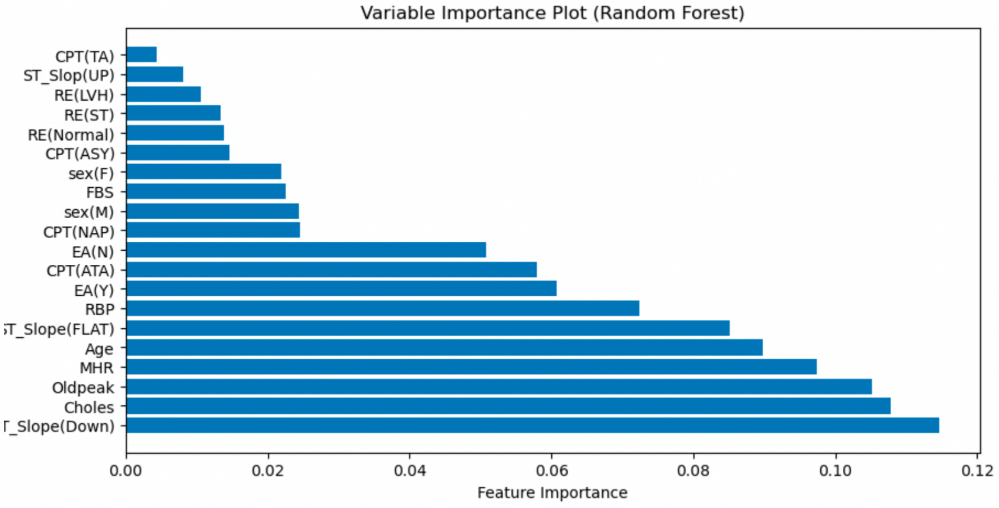
◆Exploratory data analysis

- In this dataset, the predictors like Cholesterol, RestingBP, MaxHR exhibit a uniform and continuous distribution across their respective ranges, as indicated by the histograms.
- This characteristic suggests that there is diversity and ample variation in the data for each parameter. Having such a broad spread of values can be advantageous for model training, as it provides the algorithm with a rich and comprehensive set of inputs.
- A diverse range of predictor values allows the model to capture different patterns and relationships within the data, potentially enhancing its ability to generalise well to unseen instances. This wide spread of parameters is generally considered favorable for robust and effective machine learning model training.



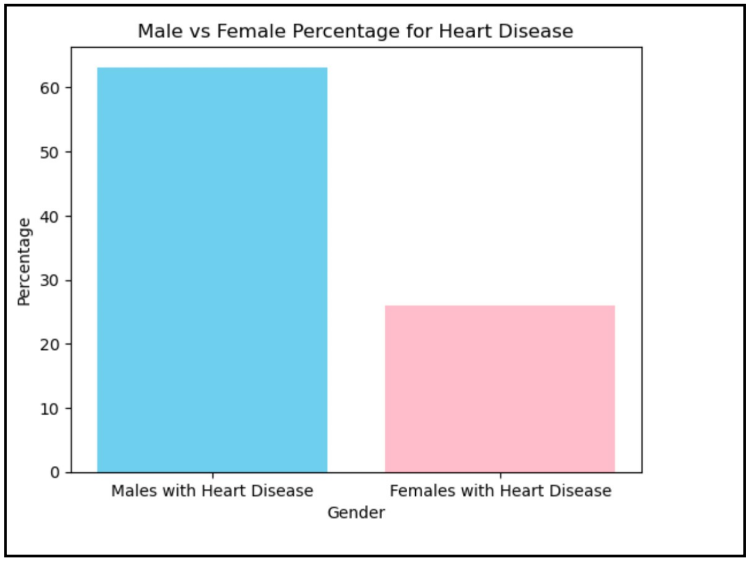


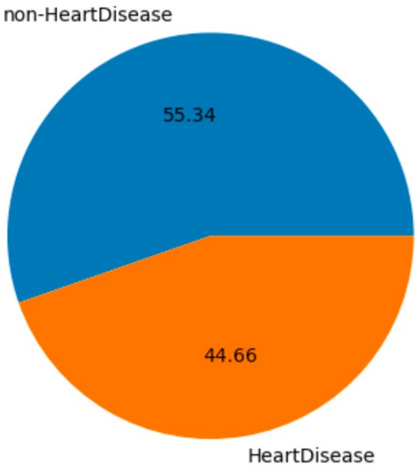




3. Results and Findings

- Our analysis of the heart disease data set using logistic regression and random forest produced promising results in predicting the likelihood of heart disease in patients. The model achieved respectable accuracy with logistic regression at 85.32% and random forest at 87.5%.
- Analysis of our study data revealed gender-based differences in the prevalence of heart disease in men than in women. Specifically, 63% of men in the database were diagnosed with heart disease, compared to 25% of women.
- Our models also predicted the same on test data as via Logistic regression and Random forests we came to see that males in general have a higher percentage of heart disease as compared to females.
- These findings highlight the importance of gender in assessing and managing heart health. The predictive capabilities of our models, especially Random Forests, show that they can help healthcare providers in the early detection and intervention of patients at risk of heart disease.
- This research lays the groundwork for future research into personalised and effective strategies to prevent and manage heart disease.


Male vs Female Percentage for Heart Disease
Percentage
Males with Heart Disease
Females with Heart Disease
Gender


non-HeartDisease
HeartDisease