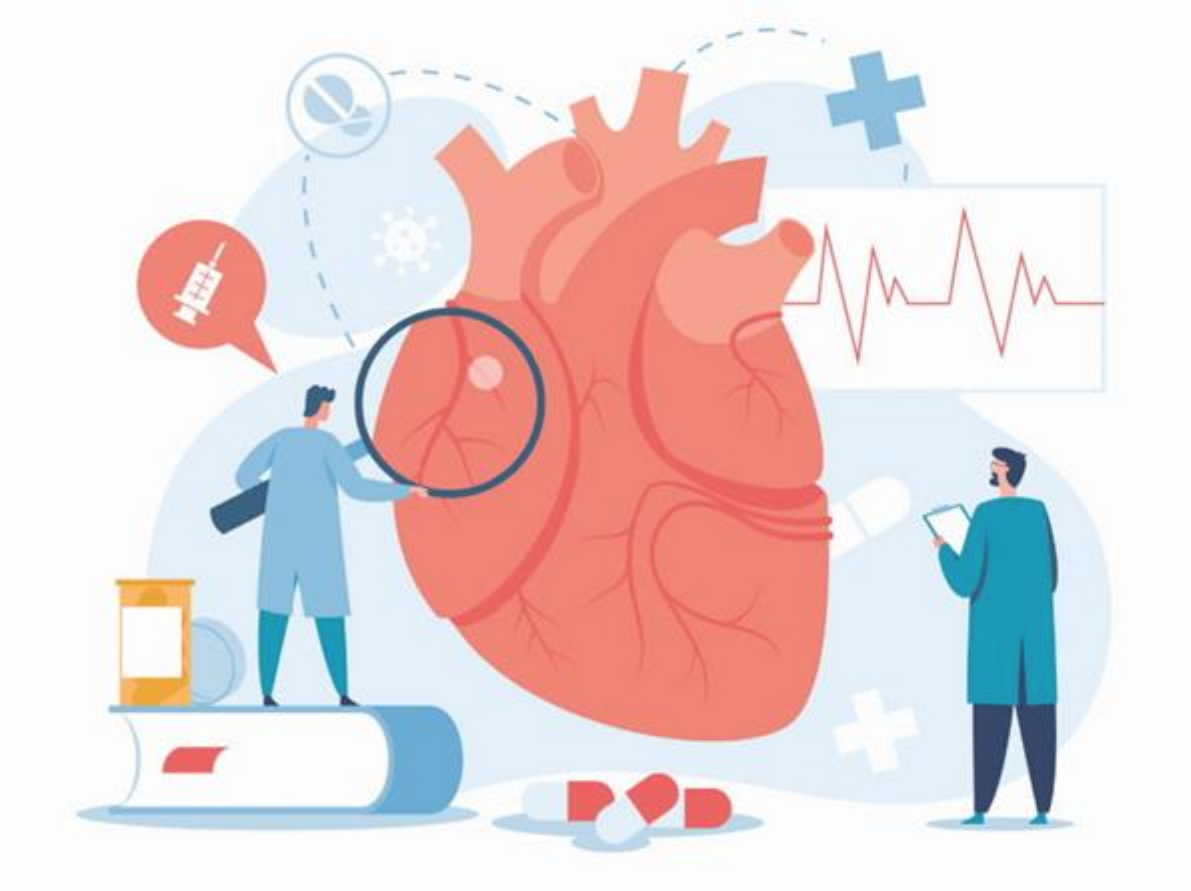
Heart Disease Analysis and Prediction using EDA and ML Classifiers

Introduction

- Around the world, heart disease is regarded as the condition that kills people the fastest
- To reduce heart-related concerns and to protect it from catastrophic hazards, early detection of heart disease is important
- Exploratory data analysis for healthcare purposes aids in disease prediction, better diagnosis, symptom analysis and provision of suitable medications
- The objective is to perform analysis on the heart disease dataset and visualize the same for a better understanding of the dataset and uncover hidden trends
- This will aid in preparing a heart disease prediction model and app which classifies whether a person has a heart disease or not, based on the input of feature variables



Methodology and Implementation

DATA GATHERING

EXPLORATORY DATA ANALYSIS

HEART DISEASE DASHBOARD

DATA PRE-PROCESSING AND SPLITTING THE DATASET

APPLYING MACHINE LEARNING CLASSIFICATION ALGORITHMS

EVALUATING THE EFFECTIVENESS OF ML CLASSIFIERS

HEART DISEASE PREDICTION APP

Exploratory Data Analysis

DATA CLEANING

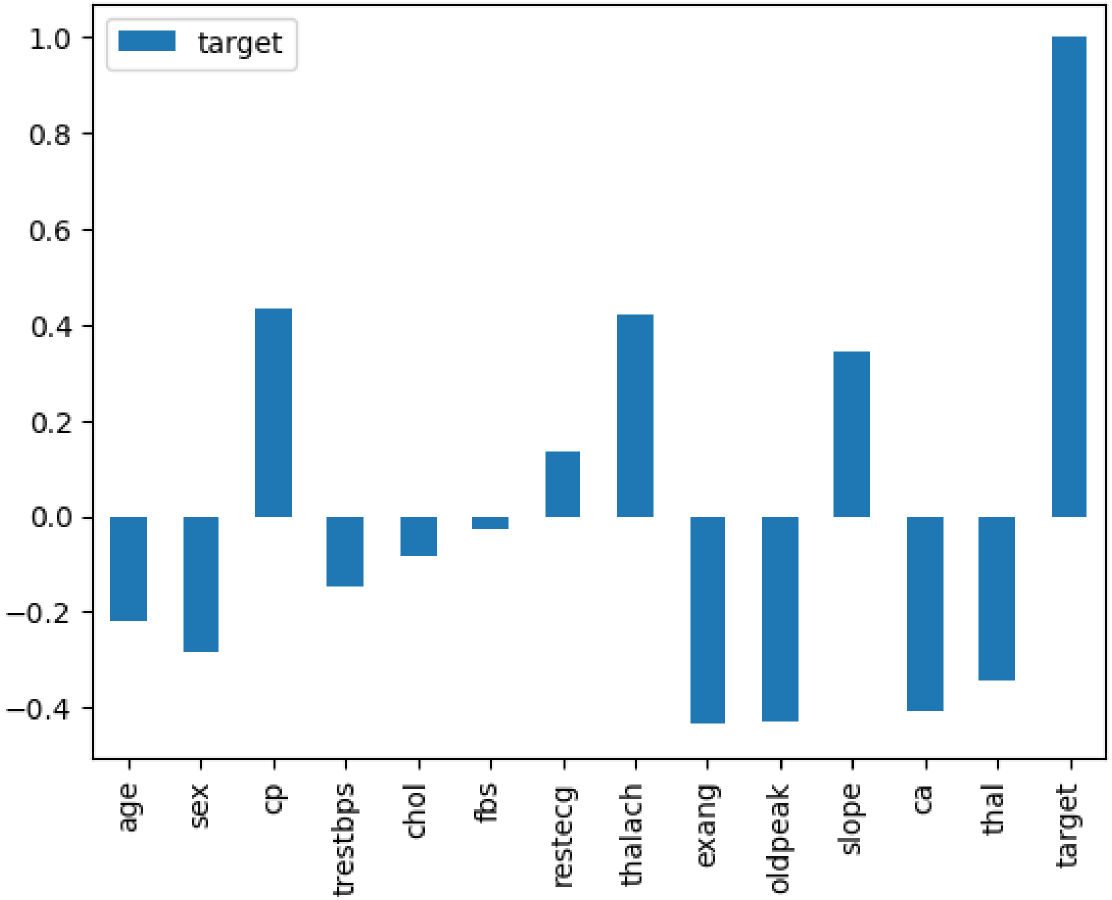
DATA DESCRIPTION

DATA VISUALISATION

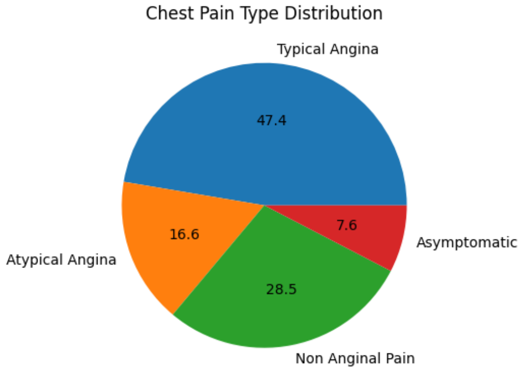
Feature Analysis

The following have the maximum correlation with the target feature in determining the possibility of heart disease.

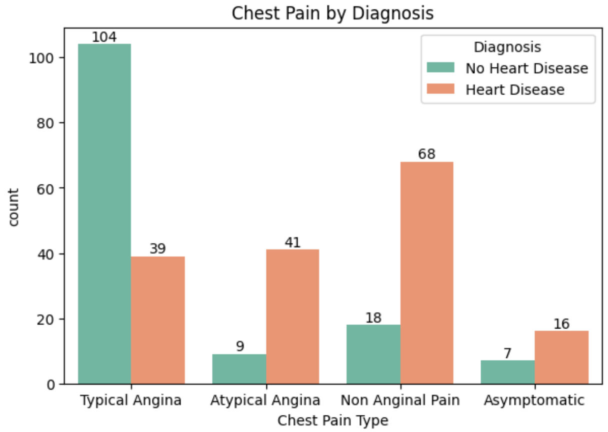
1. cp-Chest Pain Type
2. thalach-Max heart rate
3. exang-Exercise induced angina,
4. oldpeak-ST-depression
5. ca-Number of blood vessels colored



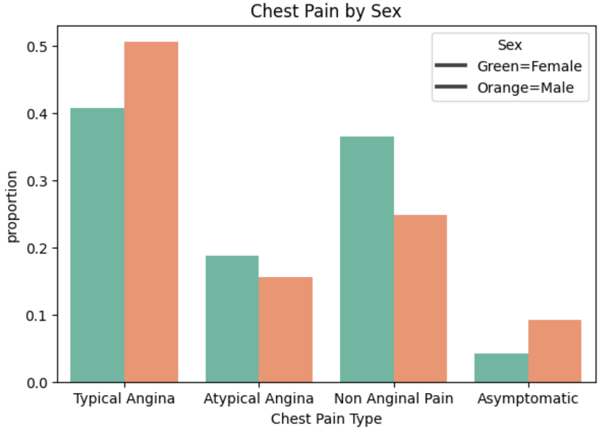
Type of Chest Pain



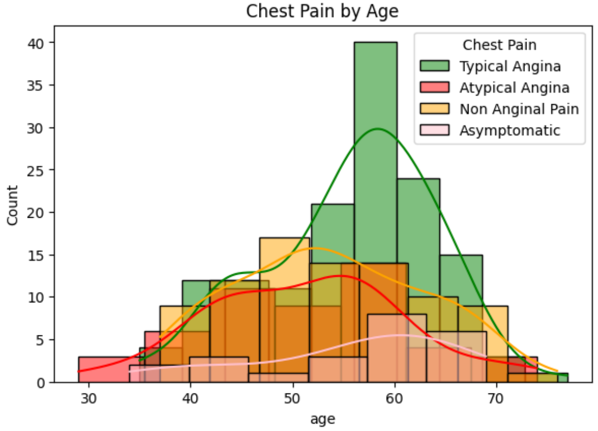
Chest Pain Type Distribution



Chest Pain by Diagnosis



Chest Pain by Sex



Chest Pain by Age

Inference

- Patients with typical angina pain tend to not have heart disease.
- Non-anginal pain shows the strongest relation with the possibility of heart disease.
- Most of the females have either typical angina or non-anginal pain and most of the males have typical anginal pain.
- Mode value of typical angina pain, atypical angina pain and asymptomatic is approximately 60 years while the mode value of non-anginal pain is around 50 years, in terms of age.

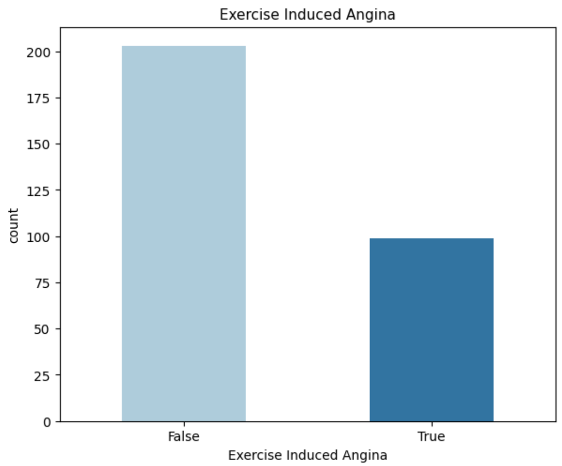
Exercise Induced Angina

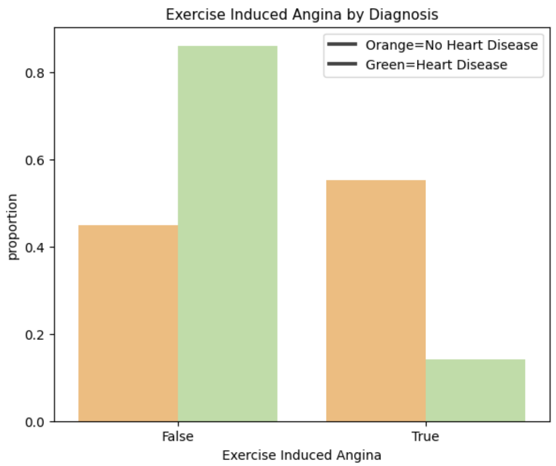
Inference

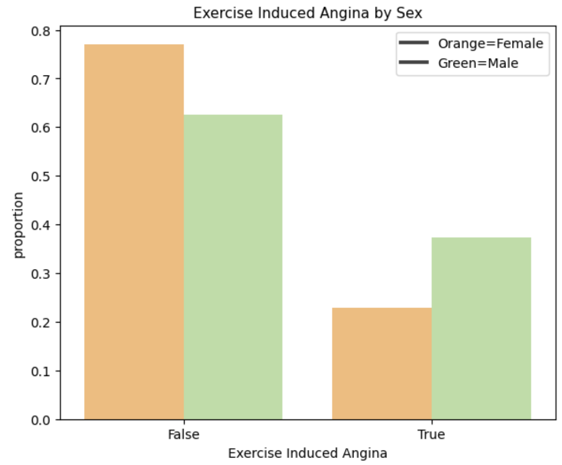
- Exercise-induced angina correlates more to not having heart disease instead of having heart disease.

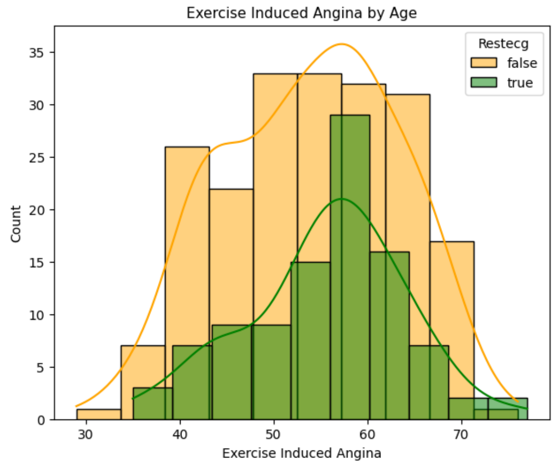
- males are more likely to feel exercise induced angina than females.

- People having exercise-induced angina show a more left skewed age distribution than people not having it.

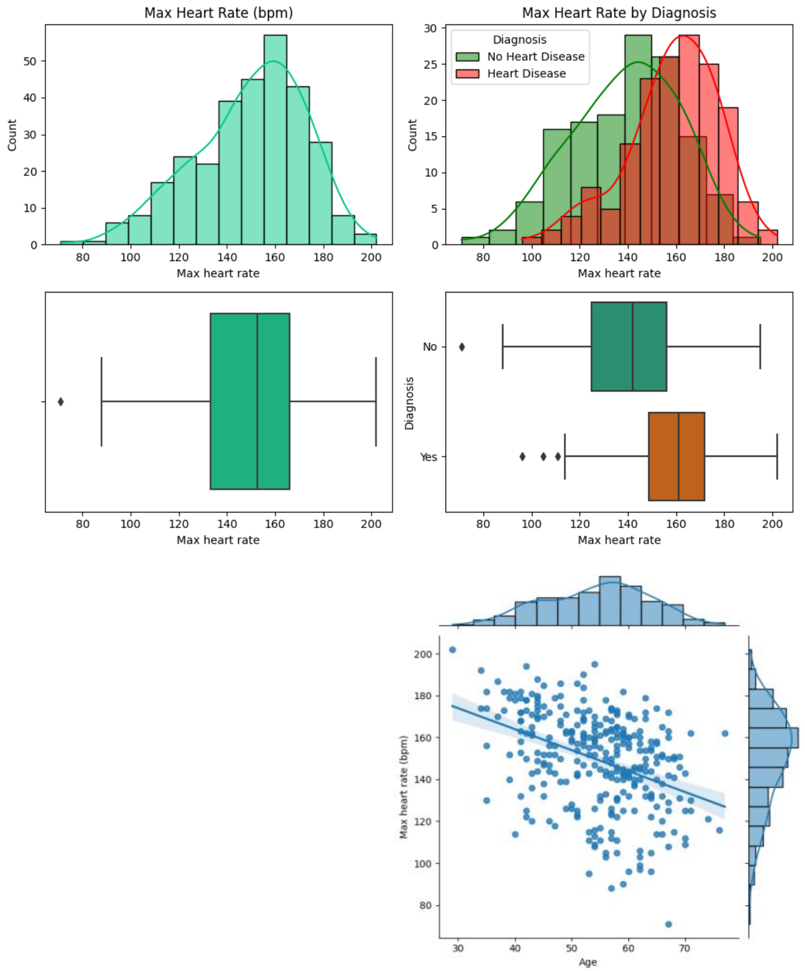

Exercise Induced Angina


Exercise Induced Angina by Diagnosis


Exercise Induced Angina by Sex


Exercise Induced Angina by Age

Maximum Heart Rate



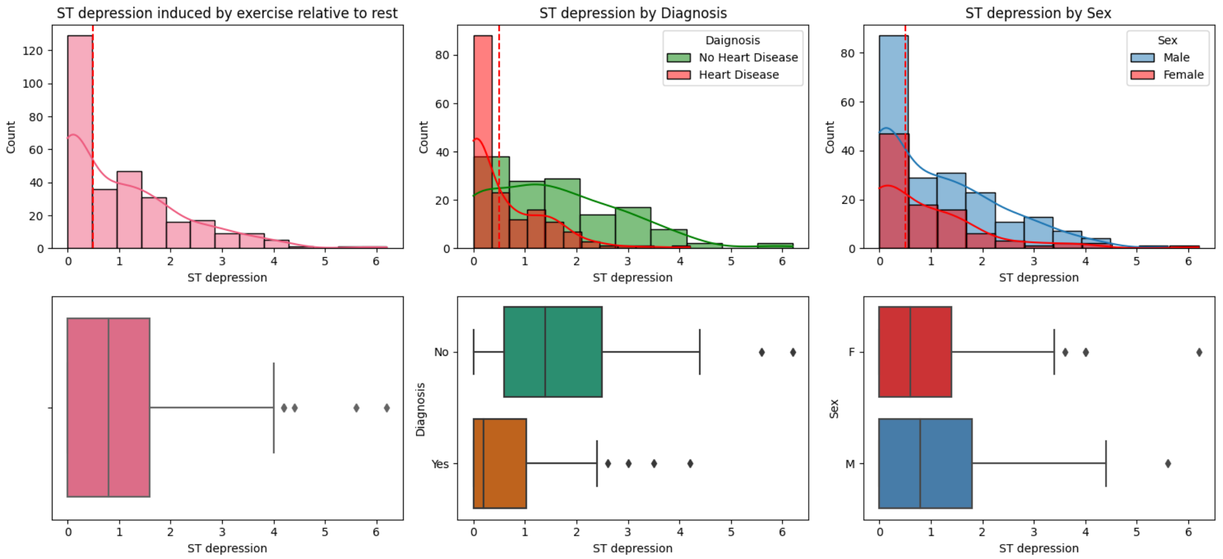
Inference

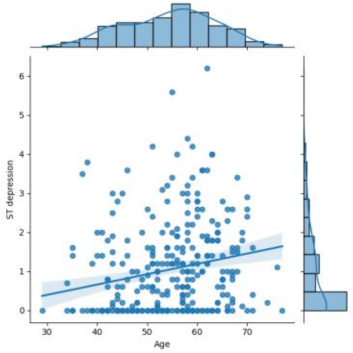
- Some subjects have too low max heart rate though the distribution is left-skewed
- A Higher max heart rate corresponds to more chances of having heart disease.
- Females mostly have a max heart rate between the range of 150 to 200 bpm while males have max heart rate from 140 to 200 bpm.
- age has a negative correlation with maximum heart rate as with age, heart rate decreases.

Exercise-induced ST Depression

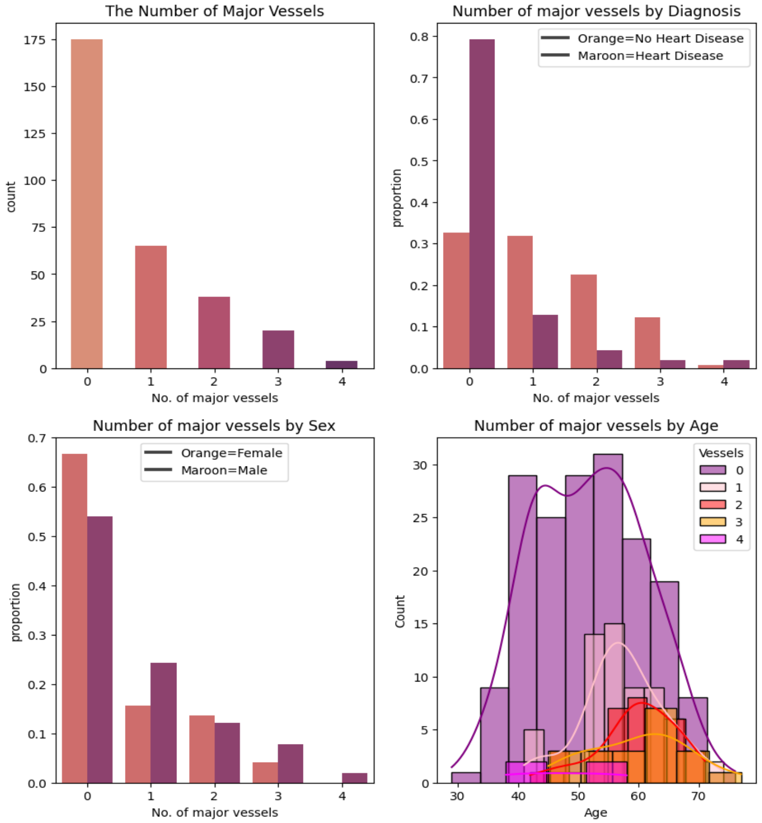
Inference

- It is a right skewed distribution with majority of the values being less than 1.
- No heart disease condition has values widely distributed while heart disease condition is tightly right skewed with several outliers .
- Both males and females show a right skewed distribution ranging between same values.
- There is a positive correlation between age and ST depression.





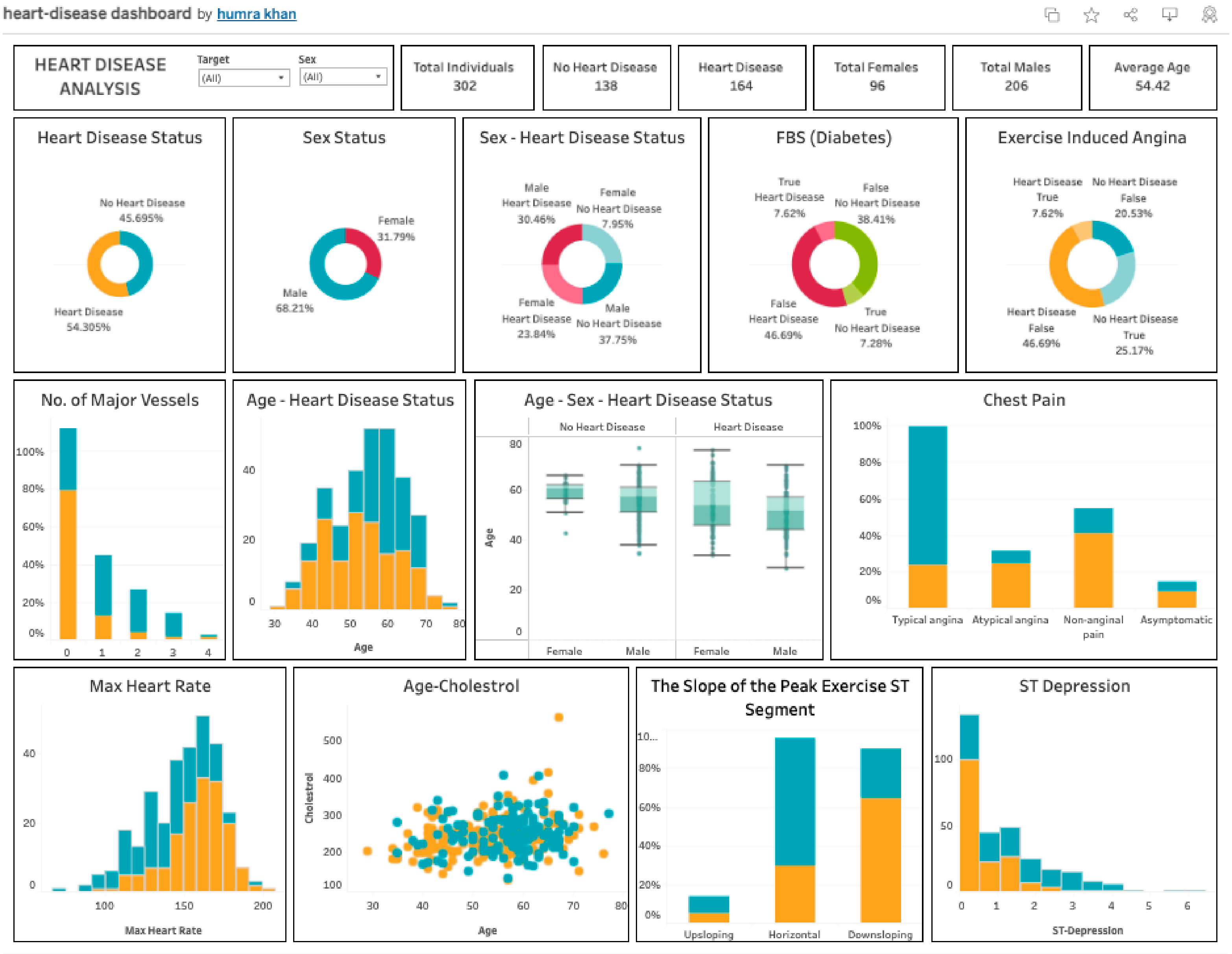
Number of Major Vessels



Inference

- This shows the number of vessels colored. Vessels not colored have clots.
- Most of the subjects have clots. These people are more exposed to having a heart disease than people who do not have clots.
- Males tend to have less clots than females.
- people having clots have ages mostly above 50 years.

Heart Disease Tableau Dashboard



heart-disease dashboard by humra khan

HEART DISEASE ANALYSIS

Target (All)

Sex (All)

Total Individuals 302

No Heart Disease 138

Heart Disease 164

Total Females 96

Total Males 206

Average Age 54.42

Data Pre-processing and Splitting the Dataset

As raw data is unusable for data analysis and prediction, pre-processing the data increases the accuracy of ML algorithms while also enhancing the quality of the data

Encoding the categorical variables becomes an essential step to convert the categorical variables to numbers so that the model can interpret and extract useful information because the majority of machine learning models only take numerical variables

Dataset is then divided into training set and testing set. A known output is part of the training set, and the model is built using this data in order to later generalize it to other data. 80% of the data in this study are used for training.

Applying Machine Learning Classification Algorithms

The classification algorithms used in the proposed work are Logistic Regression, Support Vector Machine, K- nearest Neighbour, Decision Tree, Random Forest and Gradient Boosting.

Logistic Regression: To estimate probabilities, LR uses a logistic function, commonly referred to as the sigmoid function.

Support Vector Machine: In order for SVM to function, a hyperplane or group of hyperplanes must be built in the feature space.

K-Nearest Neighbour: KNN is a supervised, "lazy learning" algorithm that uses "instance-based learning" or non-generalizing learning

Decision Tree: it categorizes the instances by sorting the tree's leaf nodes from root to leaf.

Random Forest: This ensemble classification method fits several decision tree classifiers. As a result, it improves control and forecast accuracy.

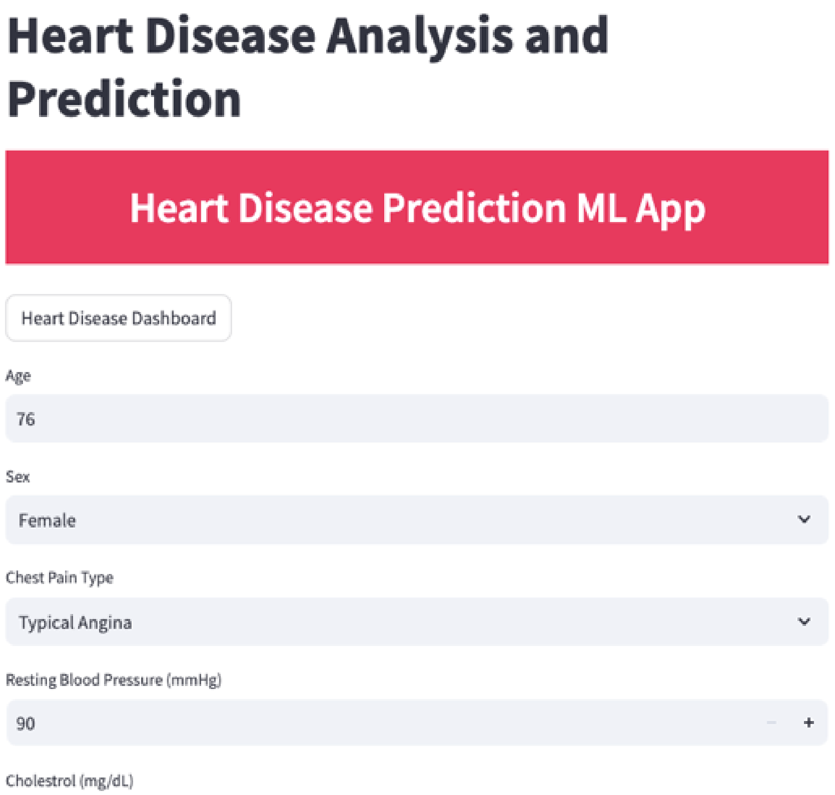
Gradient Boosting: Gradient Boosting creates a final model by combining several weak learners to construct a stronger predictor. This ensemble method is based on the iterative improvement of the model through loss function minimization.

Evaluating The Effectiveness Of ML Classifiers

The performance of the six models is assessed using a confusion matrix and all pertinent metrics, such as, accuracy, precision, recall and F1- score.

It is assessed that random forest model has the highest accuracy of 85.2% along with precision, recall and F1 score.

Heart Disease Prediction App



Heart Disease Analysis and Prediction

Age

76

Sex

Female

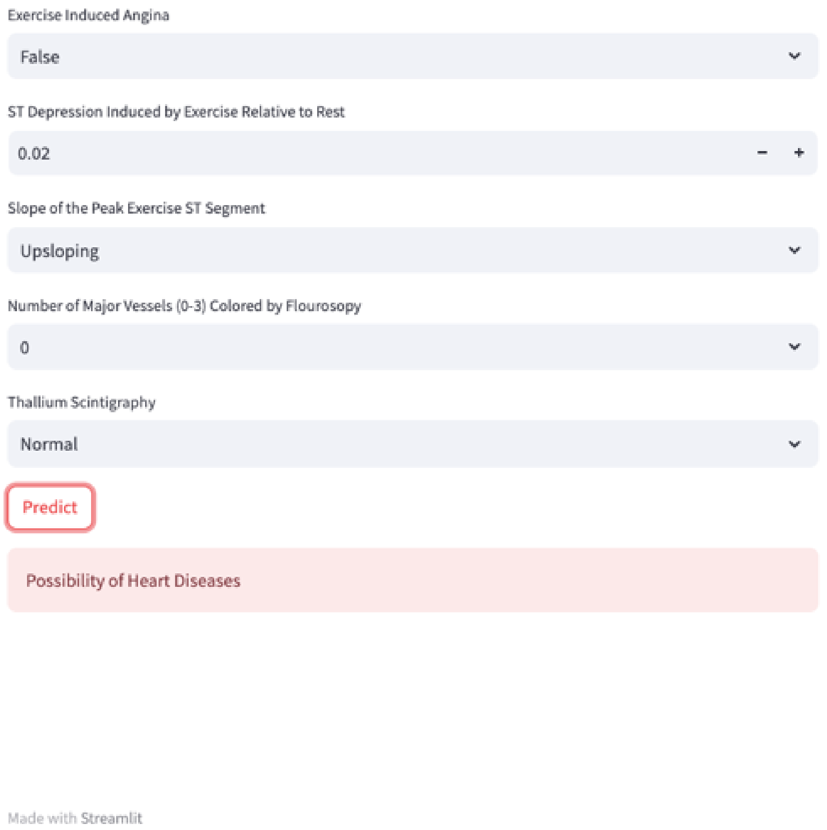
Chest Pain Type

Typical Angina

Resting Blood Pressure (mmHg)

90

Cholestrol (mg/dL)



Exercise Induced Angina

False

ST Depression Induced by Exercise Relative to Rest

0.02

Slope of the Peak Exercise ST Segment

Upsloping

Number of Major Vessels (0-3) Colored by Flourosopy

0

Thallium Scintigraphy

Normal

Possibility of Heart Diseases

Conclusion

Other ailments can also benefit from the use of exploratory data analysis and algorithms, especially as more accurate medical datasets are created in the future. To put it another way, AI-based approaches help medical systems diagnose and anticipate illnesses by maximising the utilisation of various resources.