HEART DISEASE PREDICTIONS USING SUPERVISED LEARNING

Introduction

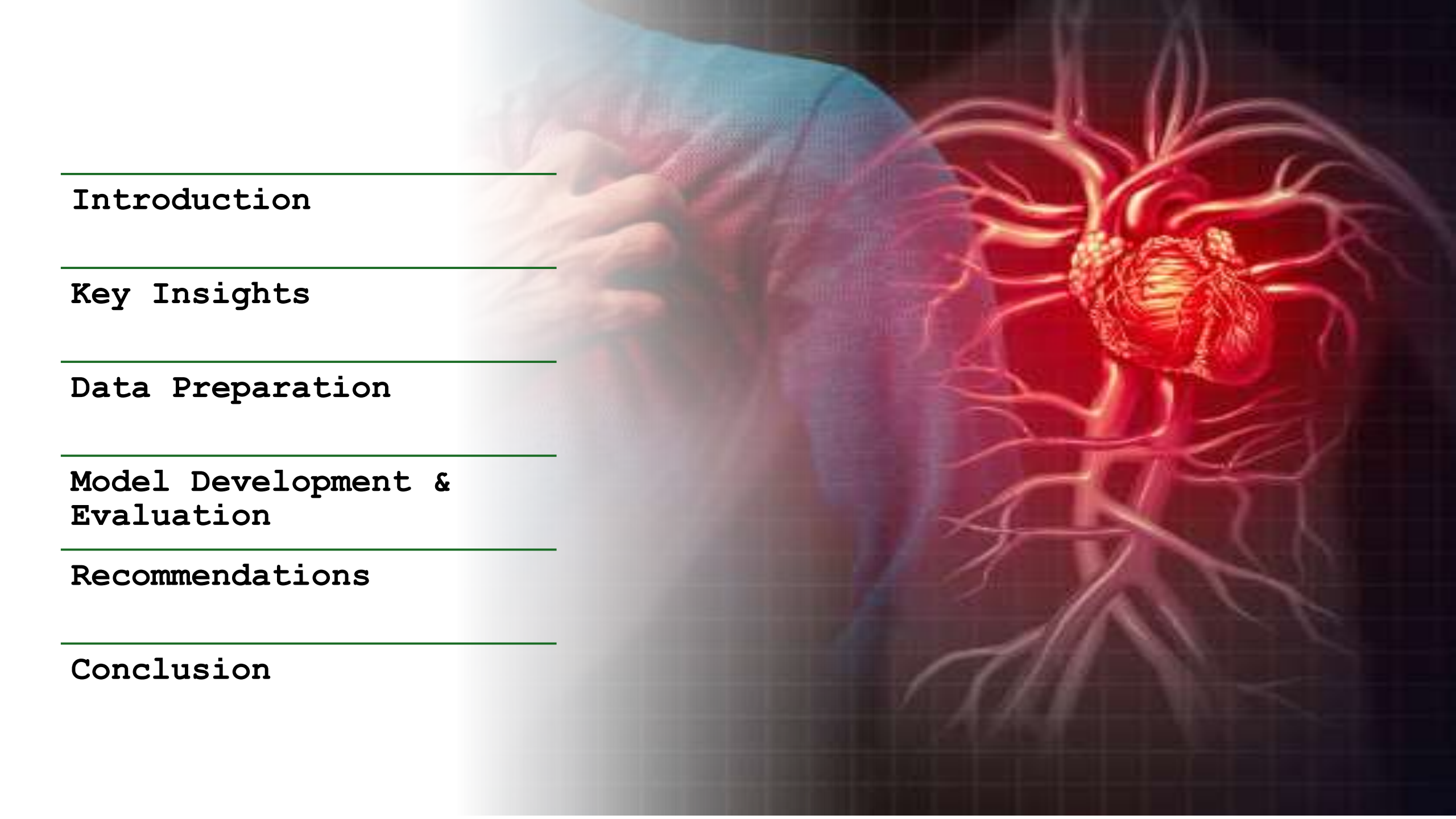
Key Insights

Data Preparation

Model Development & Evaluation

Recommendations

Conclusion



Introduction

- Heart disease remains the leading cause of death globally, responsible for millions of fatalities each year. Despite advancements in medical science, many individuals at risk of heart disease remain undiagnosed until it is too late for effective intervention

- Early and accurate prediction of heart disease can significantly improve patient outcomes by enabling timely medical interventions and lifestyle changes. However, the challenge lies in developing a reliable, efficient, and accessible method to predict heart disease risk using

Introduction

- To address this critical issue, we aim to develop a Machine Learning model that can predict the likelihood of heart disease based on patient data such as age, cholesterol levels, blood pressure, and other relevant health metrics.

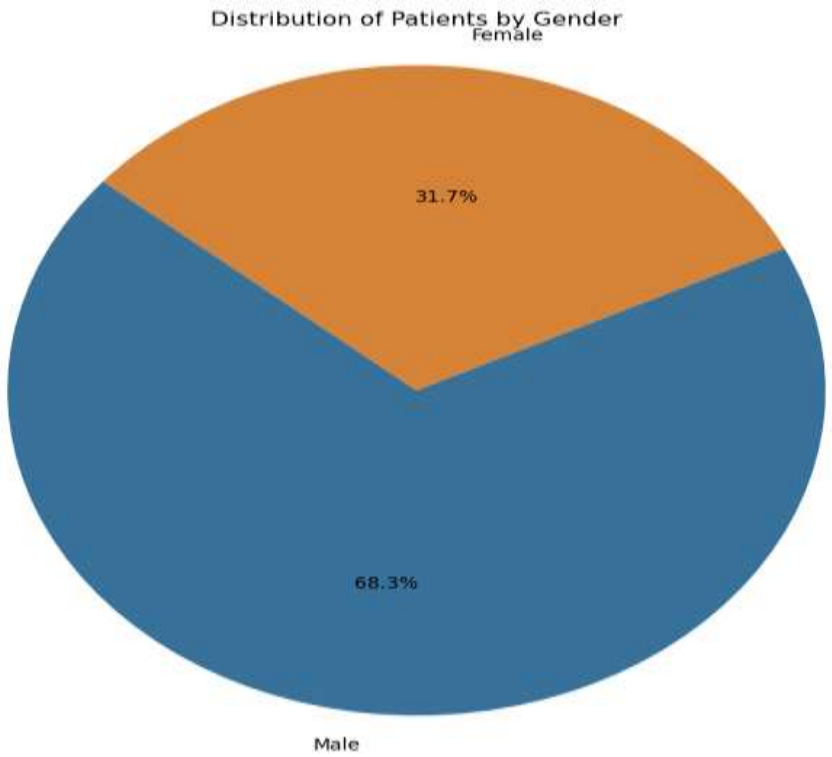
- By leveraging advanced data analytics and machine learning techniques, this project seeks to create a robust predictive tool that can be used by healthcare providers to identify high-risk individuals and take proactive measures to prevent heart disease

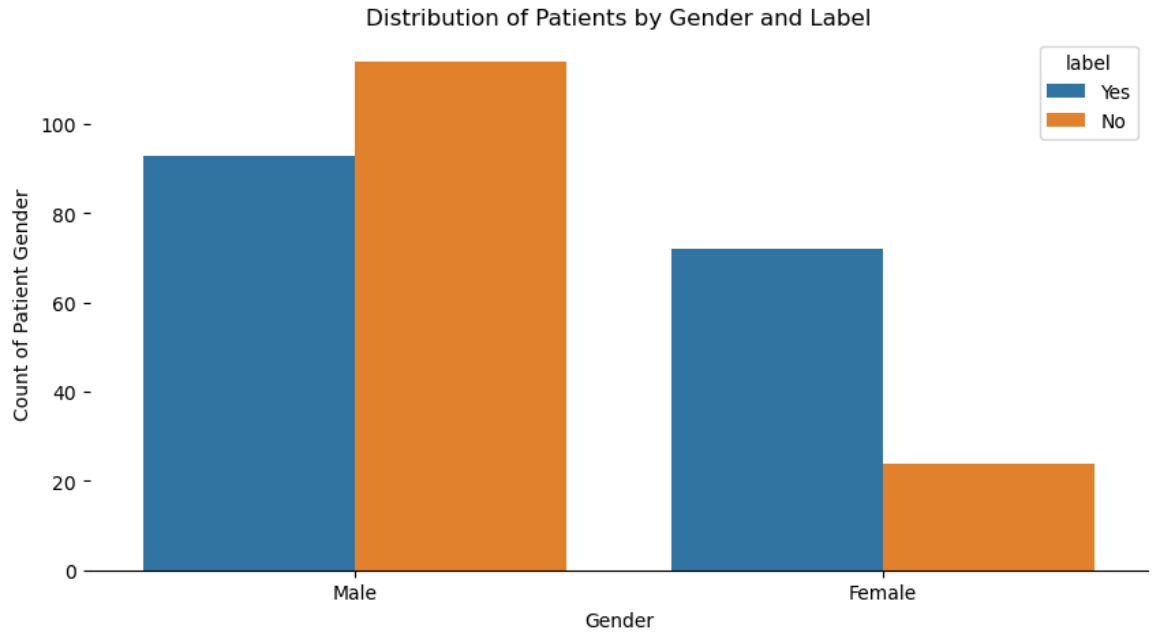
Data Preparation

- The data contains features such as patient age, cholesterol level, blood pressure, type of chest pain, etc., and the target column (presence of heart disease, binary classification: 0 = No, 1 = Yes)

- The data is cleaned for enhanced model performance- including renaming the column for better understanding, checking for missing values, and grouping some variables (e.g. group) to capture non-linear relationships.

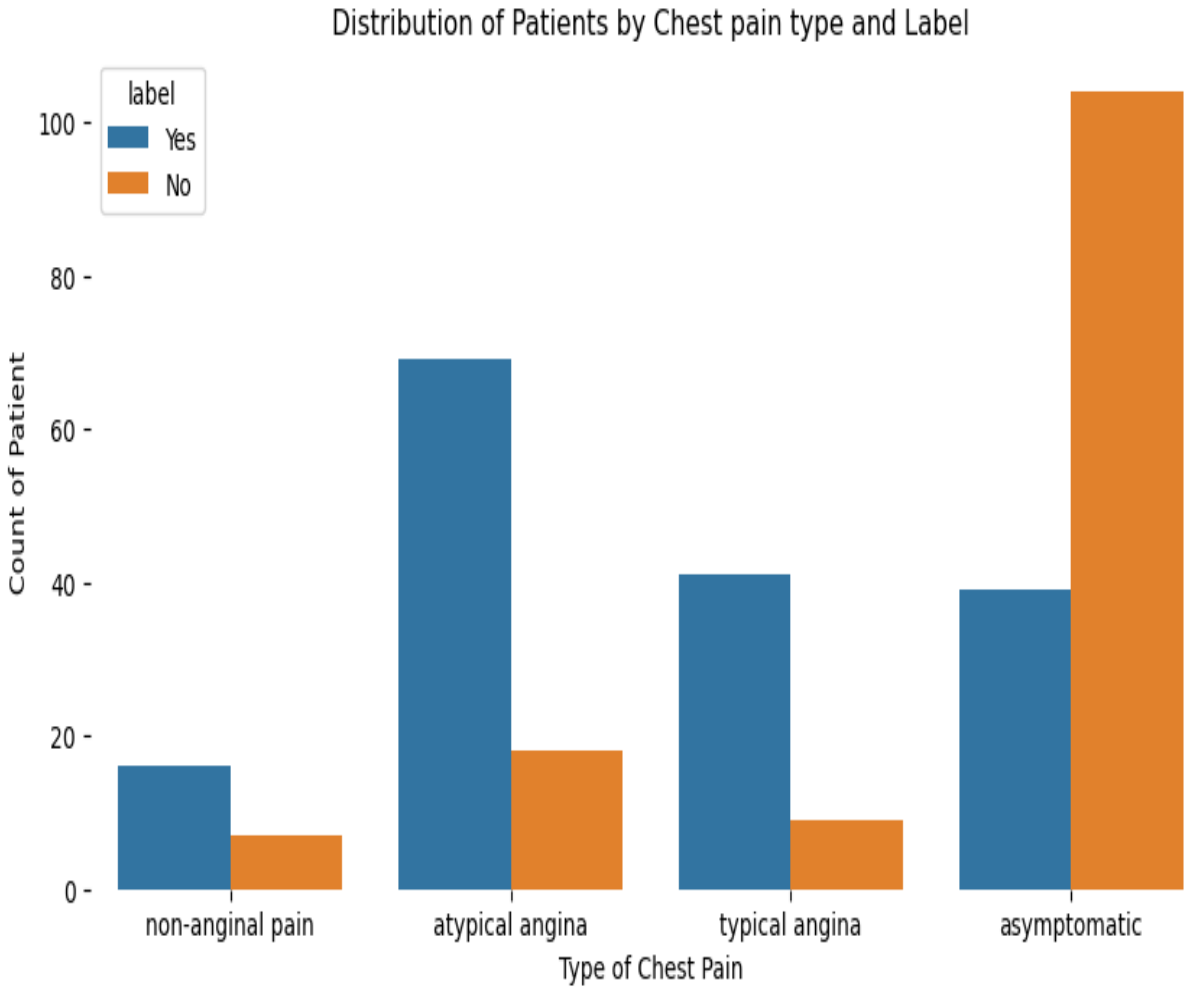
- Normalization: Scaled numerical features





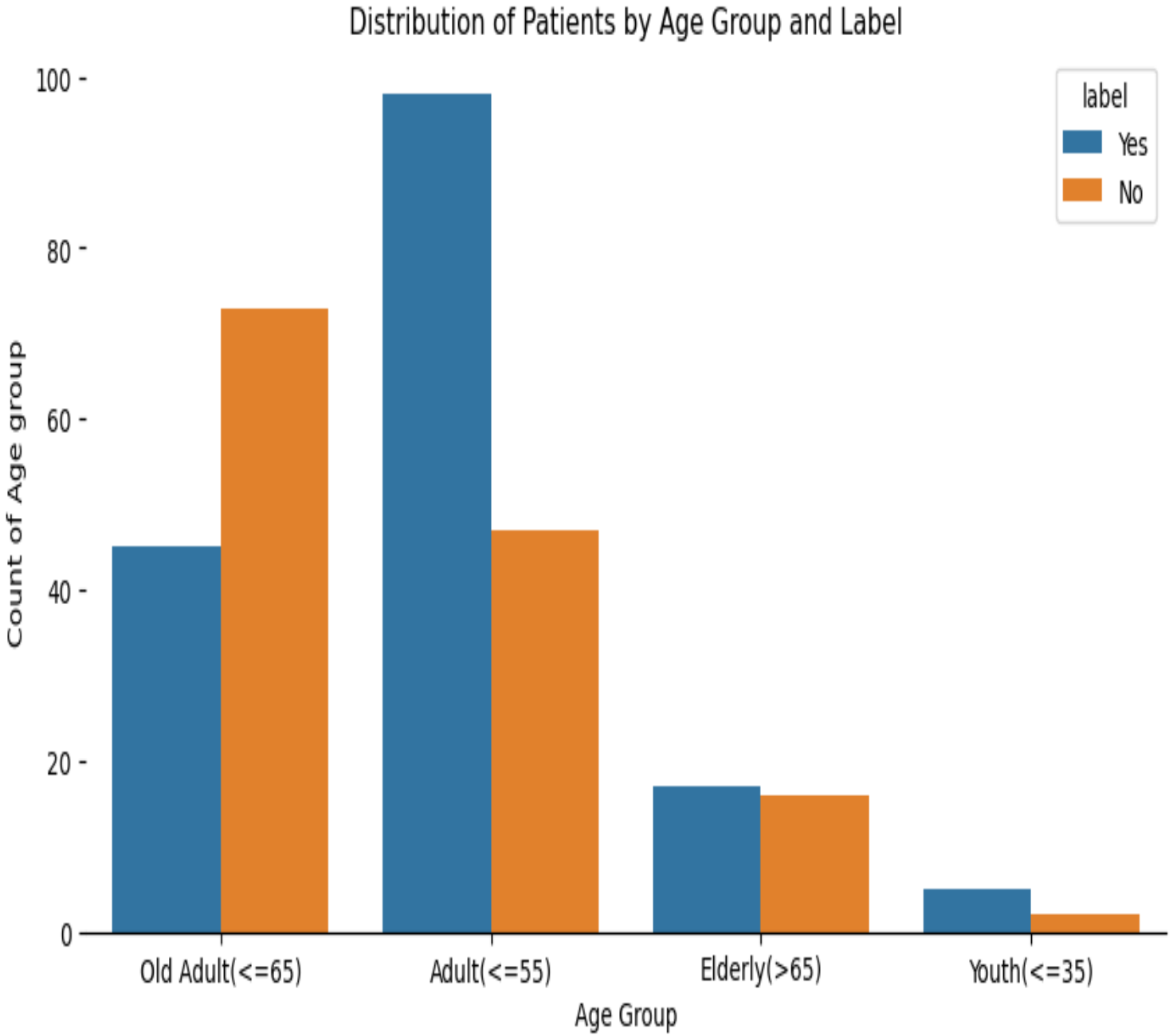
Key Insights

- Using exploratory analysis to explore key insights from the dataset.
- Male to-female distribution of patients is 68% and 32%, respectively.
- A significant number of female patients developed heart disease compared to males, comparing the gender categories.



Key Insights

Patients with any type of chest pain – atypical, typical, and non-anginal chest pain develop heart disease, compared to a few who came down with heart disease in the asymptomatic category



Key Insights

Patients with any type of chest pain – atypical, typical, and non-anginal chest pain develop heart disease, compared to a few who came down with heart disease in the asymptomatic category

Model Development

- **Preprocessing**

  - Normalized continuous features: To ensure that all continuous features are on a similar scale (range of 0 and 1), which helps in faster convergence of gradient-based optimization algorithms and prevents features with larger ranges from dominating the model training.

- **Identify the input and the target or label column**

- **Data Splitting**

  - 80% training, 20% testing

  - The training set (80% of the data) is used to train the machine learning models. It allows the models to learn the underlying patterns and relationships in the data

  - The testing set (20% of the data) is used to evaluate the performance of the trained models. It provides an unbiased estimate of model accuracy by testing on data that the model hasn't seen during

- **Model Selection**

Logistic Regression

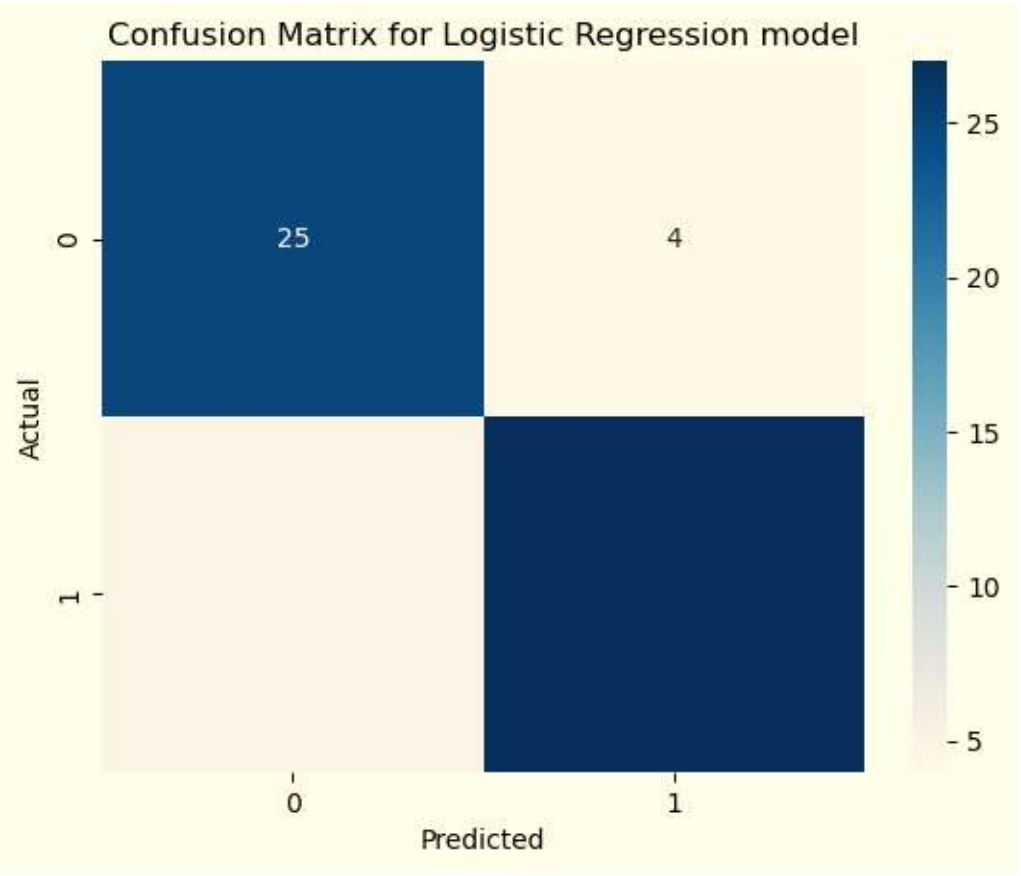
- Effective model for binary classification tasks
- Provides interpretable coefficients that can help understand the relationship between features and the target variable

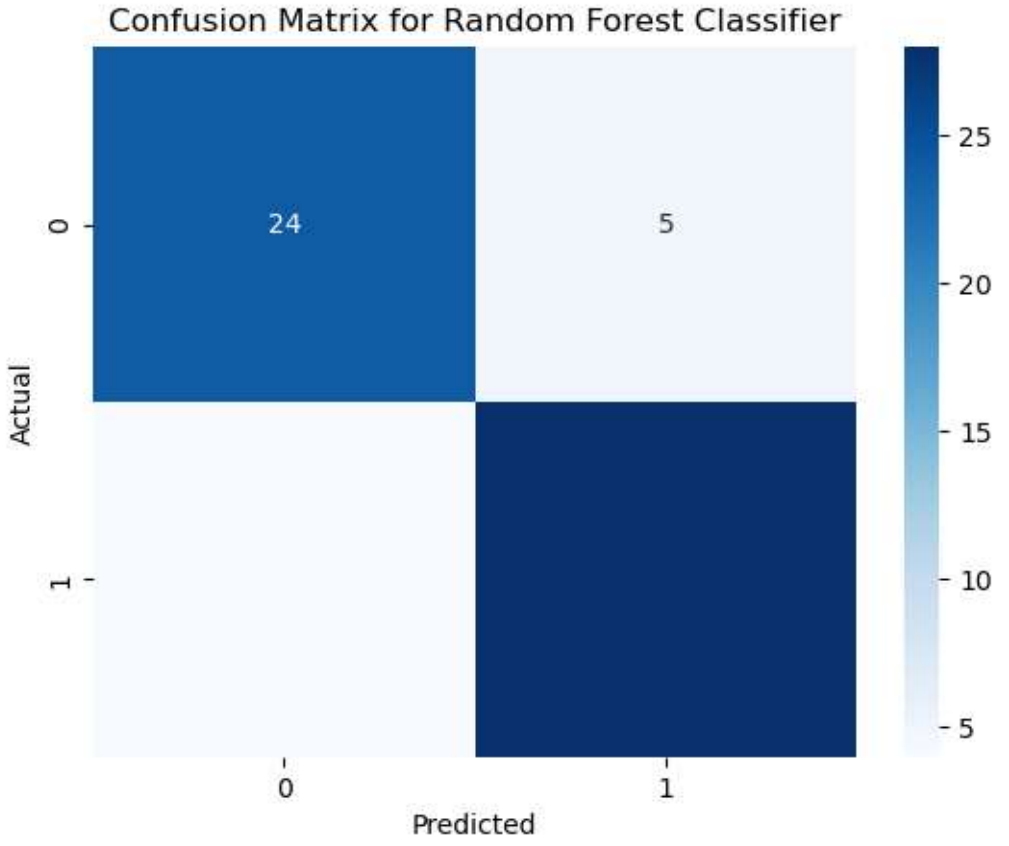
Random Forest Classifier

- Improves predictive accuracy by combining the results of multiple decision trees
- It constructs a multitude of decision trees during training and outputs the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

Evaluation

Metrics	Logistic Regression Model	Random Forest Classifier
Accuracy	0.85	0.85
Precision	0.87	0.84
Recall	0.84	0.87
F1- Score	0.86	0.86


Confusion Matrix for Logistic Regression model
25
4
Actual
0
1
0
1
Predicted
25
20
15
10
5


Confusion Matrix for Random Forest Classifier
24
5
Actual
0
1
0
1
Predicted
25
20
15
10
5

- **Benefits**

- Patients with a likelihood of getting heart disease can be detected early

- intervention methods can immediately be conducted to prevent the actual occurrence or further damage.

Conclusion

Random Forest Classifier has a better evaluation matrix

- Recall = 87%
- Precision = 84%

This is because it will detect many cases of heart attack. However, there may few cases of false prediction