Prediction of Heart disease using Machine Learning

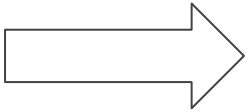
Govarthini G S

Introduction

- Heart disease is one of the leading causes of death globally
- According to WHO, 19 million people die due to Heart disease worldwide every year
- One-third of all global deaths are due to heart disease
- Analyzing a heart disease data can help in understanding its patterns, risk factors and potential preventive measures

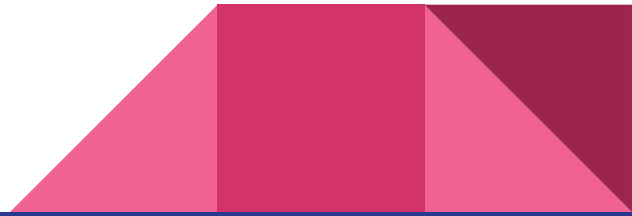
Project Overview

- ***Problem statement***  Heart disease is a leading cause of death due to lack of awareness and early detection. A predictive model can enable early intervention and reduce mortality

- ***Objective***  Develop a machine learning model to predict the likelihood of heart disease in patients based on their various medical attributes

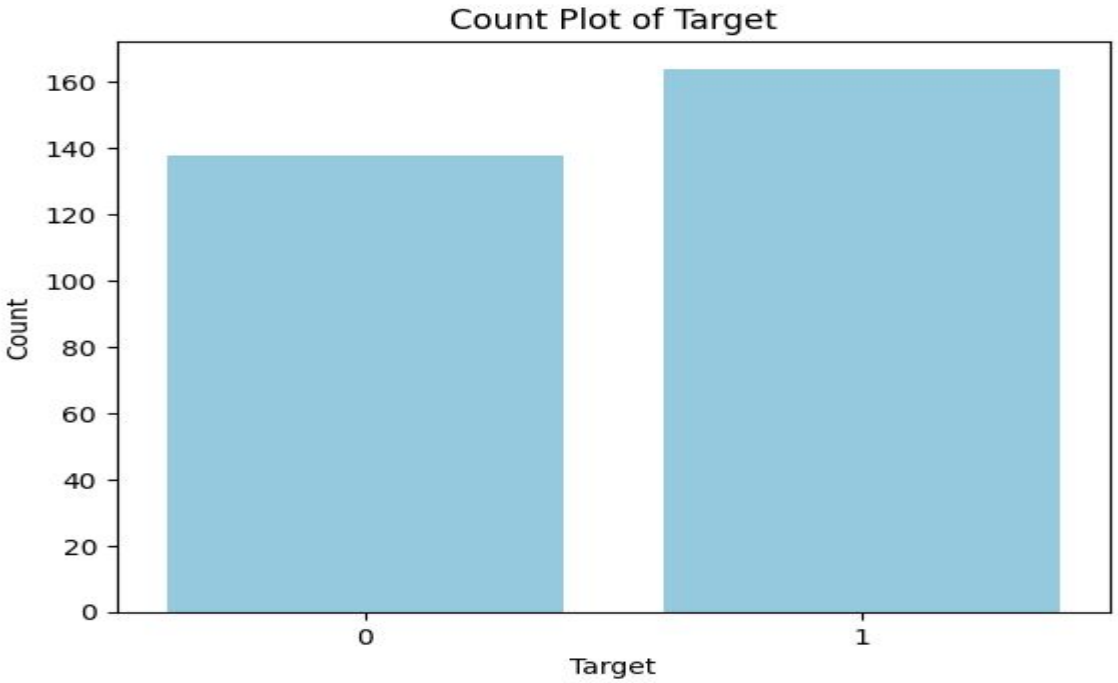
Project Overview

- Exploratory Data Analysis
- Statistical testing (Chi Squared Test)
- Predictive models
- Potential Business case & Conclusion



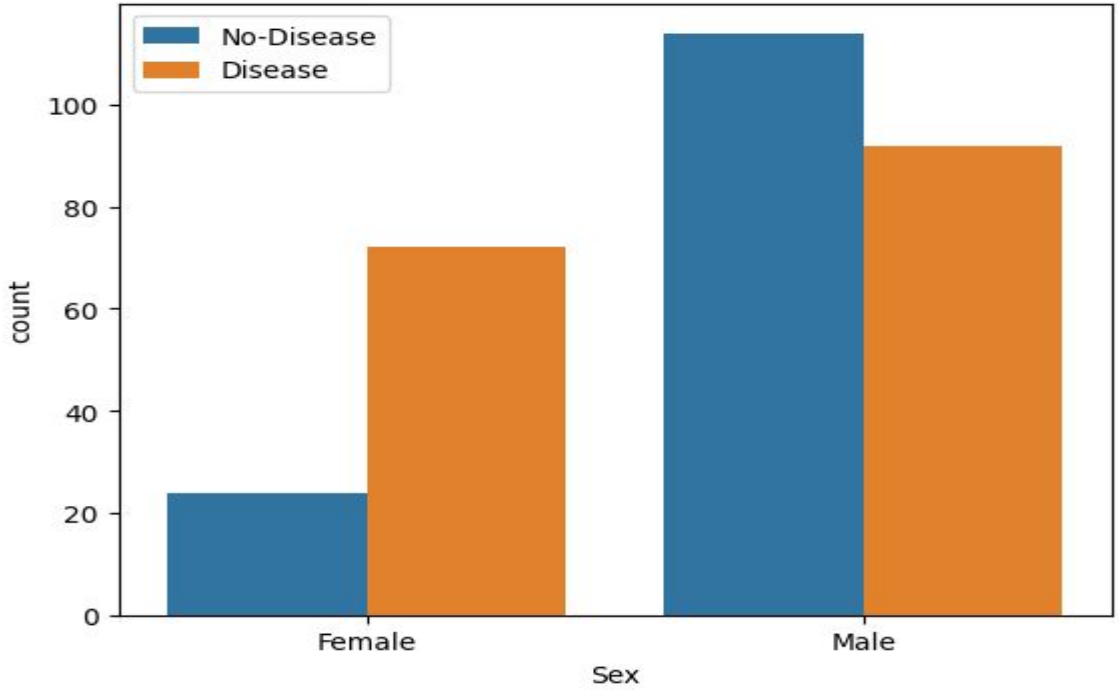
Exploratory Data Analysis

Distribution of target variable

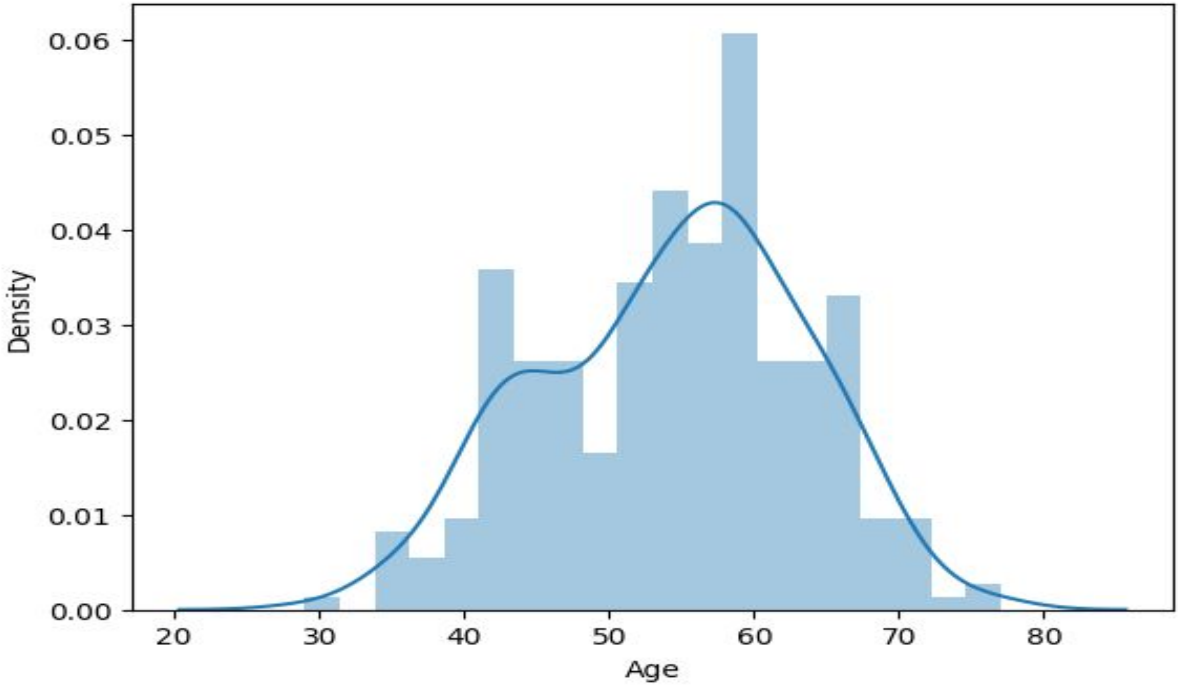


0 - without heart disease
1 - with heart disease

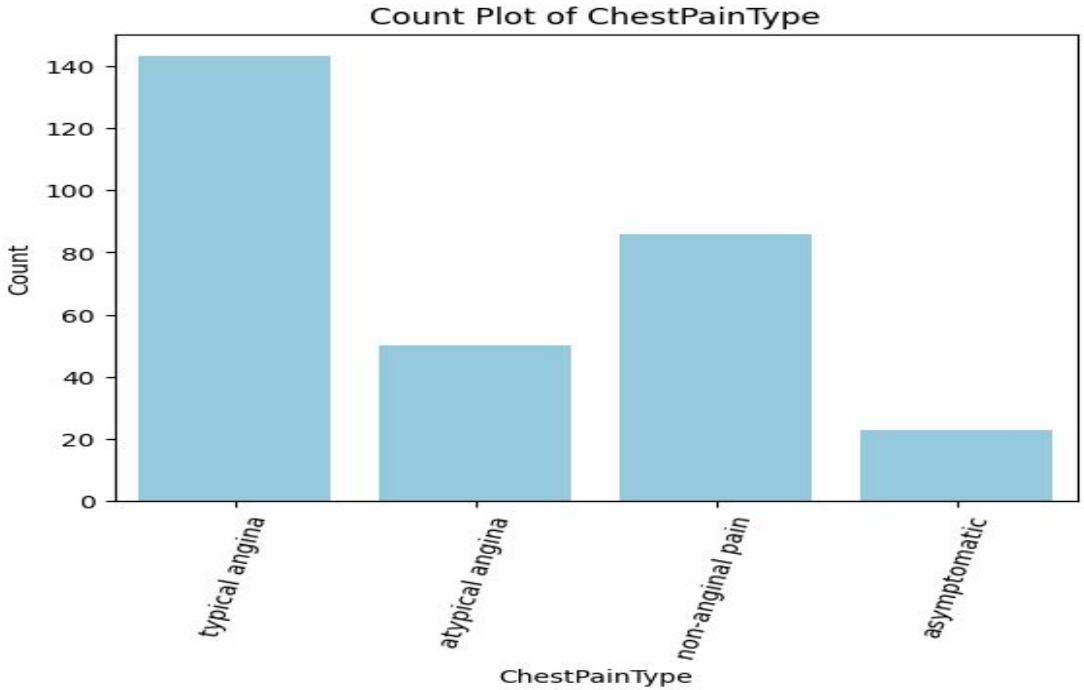
Gender distribution according to the target variable



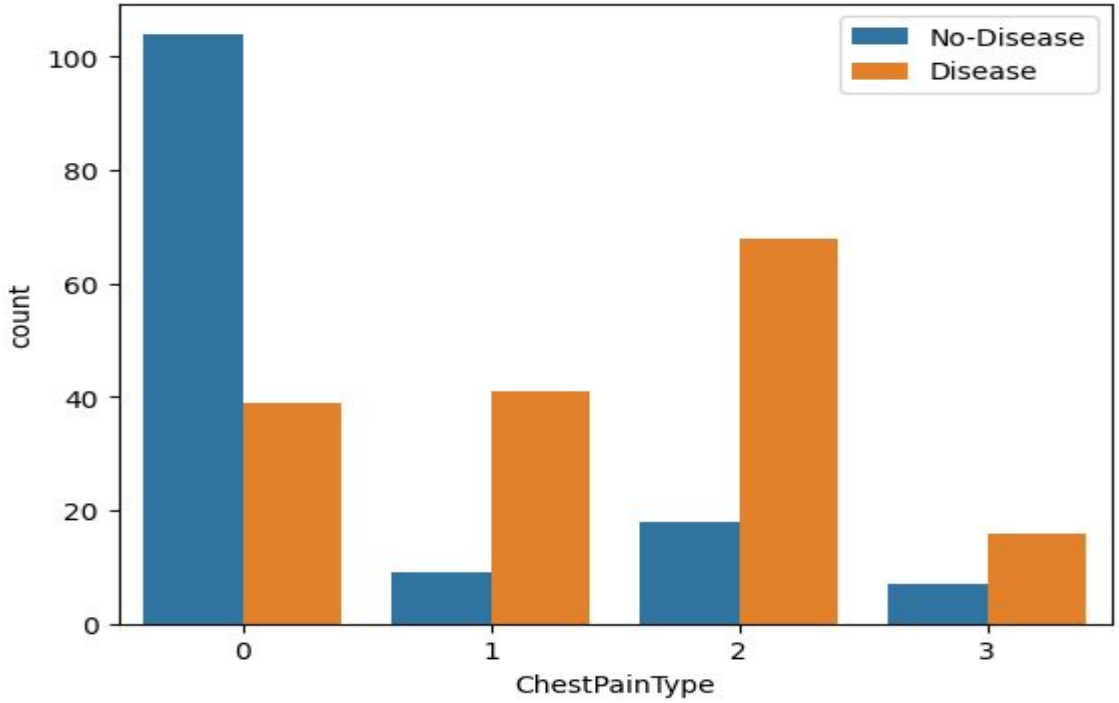
Age Distribution in the dataset



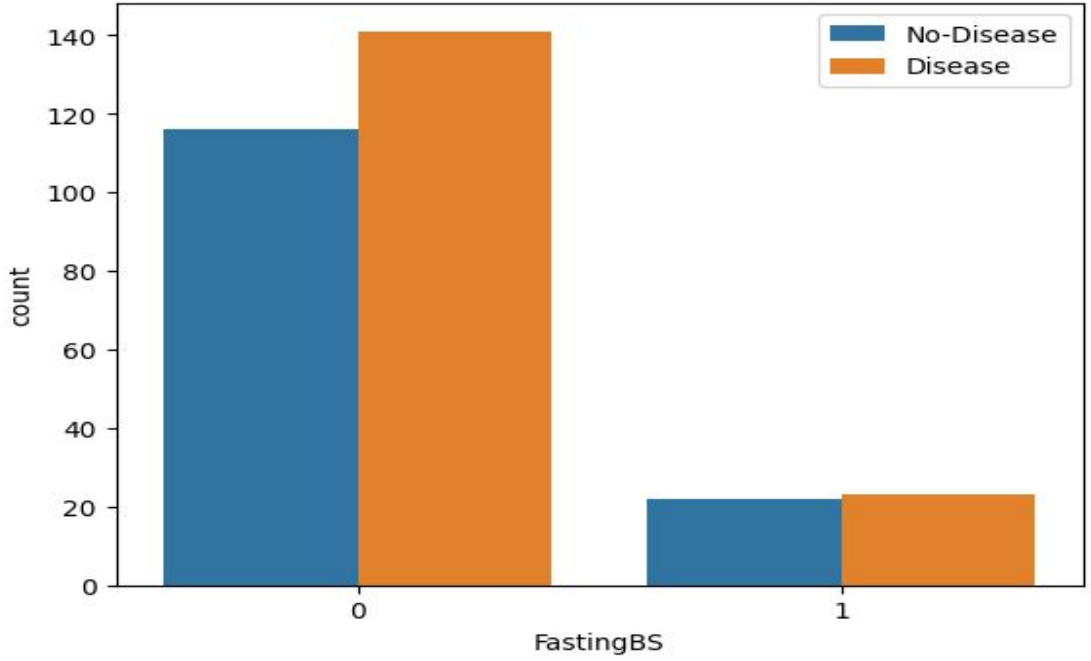
Checking pain types



Chest pain distribution as per the target variable

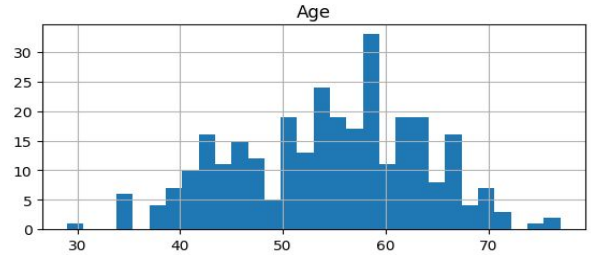


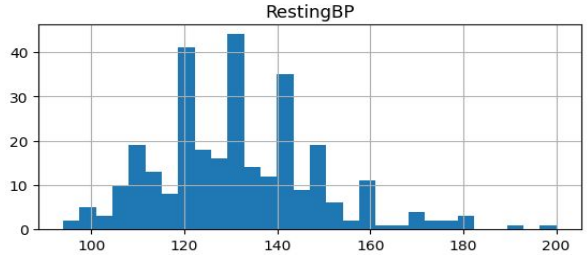
Fasting blood sugar distribution according to target variable

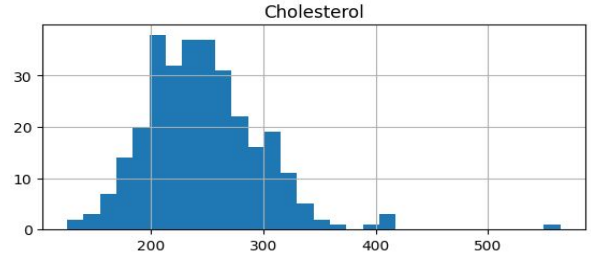


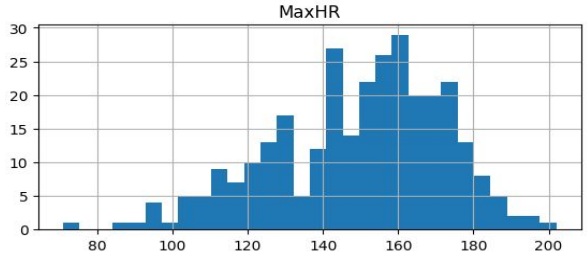
0 - without heart disease
1 - with heart disease

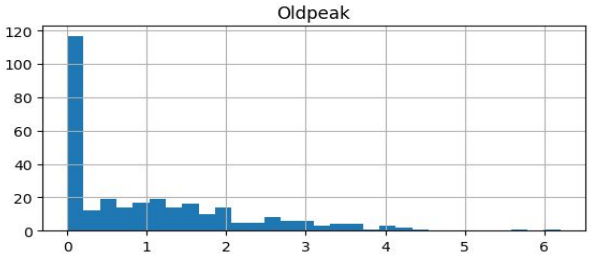
Distribution of continuous features

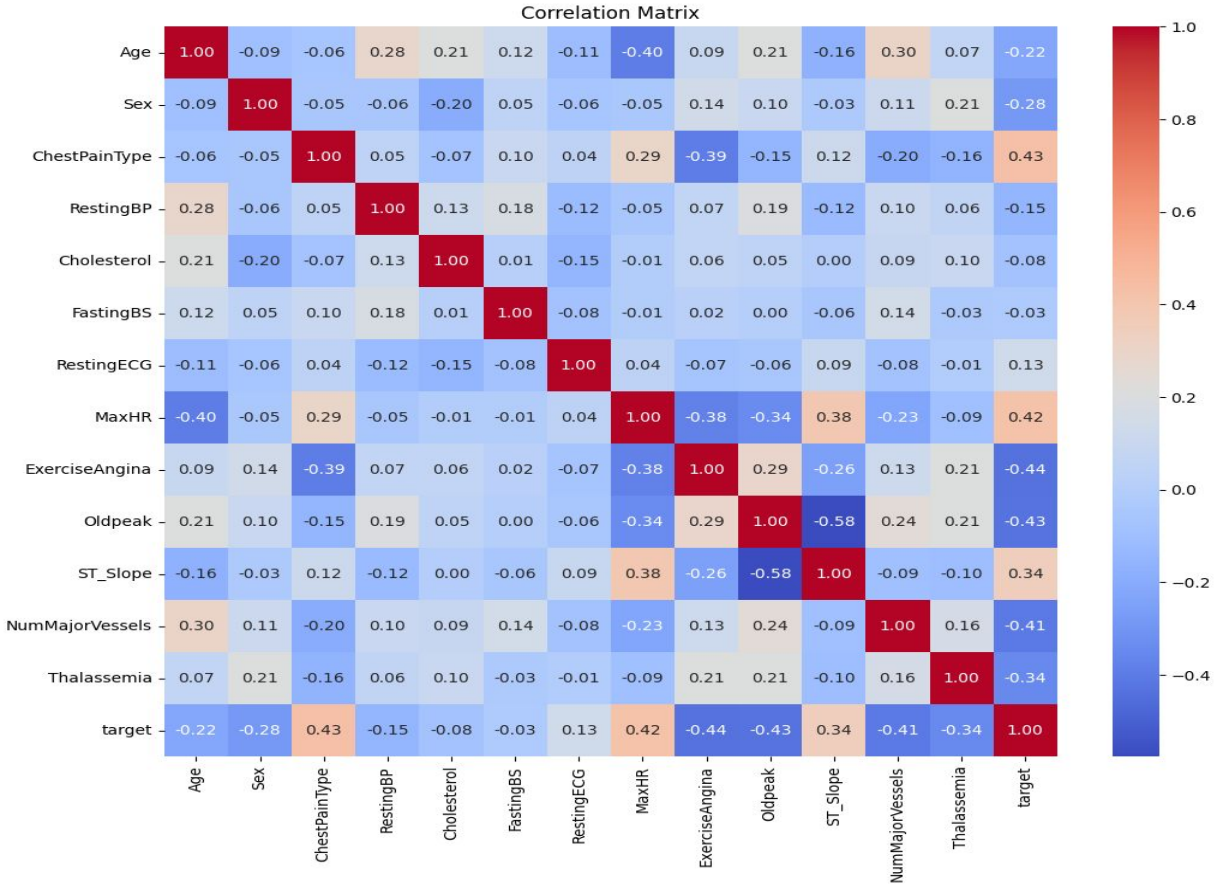











Correlation Matrix
Age 1.00 -0.09 -0.06 0.28 0.21 0.12 -0.11 -0.40 0.09 0.21 -0.16 0.30 0.07 -0.22
Sex -0.09 1.00 -0.05 -0.06 -0.20 0.05 -0.06 -0.05 0.14 0.10 -0.03 0.11 0.21 -0.28
ChestPainType -0.06 -0.05 1.00 0.05 -0.07 0.10 0.04 0.29 -0.39 -0.15 0.12 -0.20 -0.16 0.43
RestingBP 0.28 -0.06 0.05 1.00 0.13 0.18 -0.12 -0.05 0.07 0.19 -0.12 0.10 0.06 -0.15
Cholesterol 0.21 -0.20 -0.07 0.13 1.00 0.01 -0.15 -0.01 0.06 0.05 0.00 0.09 0.10 -0.08
FastingBS 0.12 0.05 0.10 0.18 0.01 1.00 -0.08 -0.01 0.02 0.00 -0.06 0.14 -0.03 -0.03
RestingECG -0.11 -0.06 0.04 -0.12 -0.15 -0.08 1.00 0.04 -0.07 -0.06 0.09 -0.08 -0.01 0.13
MaxHR -0.40 -0.05 0.29 -0.05 -0.01 -0.01 0.04 1.00 -0.38 -0.34 0.38 -0.23 -0.09 0.42
ExerciseAngina 0.09 0.14 -0.39 0.07 0.06 0.02 -0.07 -0.38 1.00 0.29 -0.26 0.13 0.21 -0.44
Oldpeak 0.21 0.10 -0.15 0.19 0.05 0.00 -0.06 -0.34 0.29 1.00 -0.58 0.24 0.21 -0.43
ST_Slope -0.16 -0.03 0.12 -0.12 0.00 -0.06 0.09 0.38 -0.26 -0.58 1.00 -0.09 -0.10 0.34
NumMajorVessels 0.30 0.11 -0.20 0.10 0.09 0.14 -0.08 -0.23 0.13 0.24 -0.09 1.00 0.16 -0.41
Thalassemia 0.07 0.21 -0.16 0.06 0.10 -0.03 -0.01 -0.09 0.21 0.21 -0.10 0.16 1.00 -0.34
target -0.22 -0.28 0.43 -0.15 -0.08 -0.03 0.13 0.42 -0.44 -0.43 0.34 -0.41 -0.34 1.00
Age Sex ChestPainType RestingBP Cholesterol FastingBS RestingECG MaxHR ExerciseAngina Oldpeak ST_Slope NumMajorVessels Thalassemia target
1.0 0.8 0.6 0.4 0.2 0.0 −0.2 −0.4

Statistical Testing

Chi Square Test to access the relationship between categorical features and the target variable

- For features Sex, ChestPainTypes, Exercise Angina,ST_slope,Num Major Vessels,Resting ECG and Thalassemia, the p-value are 0.00
- For Fasting Blood Sugar, p-value is 0.76

All the categorical features are strongly associated with the target variable except Fasting Blood Sugar!!

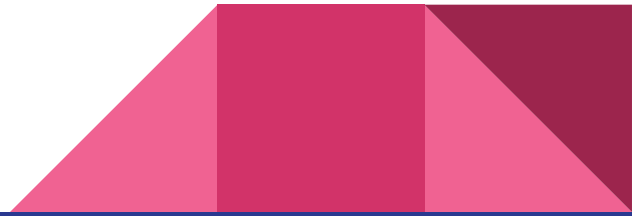
Predictive Models

Data selection & Preparation

- Data cleaning (checking for duplicates and dropping them), String formatting (formatted the column names for better consistency)
- Data Exploration and Visualization
- Pre-processed the data in order to implement the Machine Learning models

Feature Engineering & Selection

- **Train test split**: Setting our target column (Target) and pre-selecting the rest of the features
- Normalize all the values by using the **MinMax Scaler**.



Model Testing

Basic classification Machine Learning models:

- ❏ Logistic Regression
- ❏ K Neighbors Classifier
- ❏ Support Vector Classifier
- ❏ Decision Tree Classifier

Basic ensemble approaches:

- ❏ Random Forest Classifier
- ❏ Gradient Boosting Classifier

Metrics Used

The below are the following metrics used to compare:

- ❏ Accuracy
- ❏ Precision
- ❏ Recall
- ❏ F1-score

Finally Classification report is also generated!!

Model Optimization

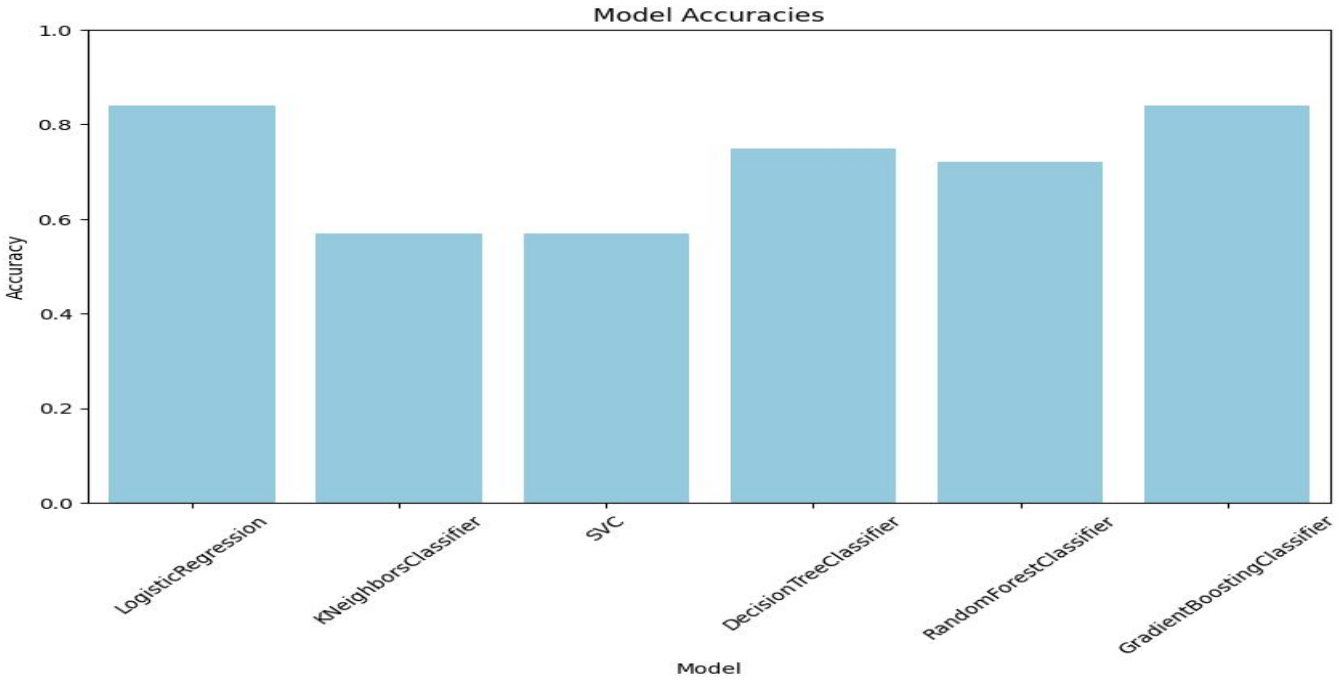
- Hyperparameter tuning technique: GridSearchCV
- *Gradient Boosting Classifier performed good*
- ***The final Metrics***: *Accuracy 84%*

  Precision 79%

  Recall 90%

  F1-score 84%

Model Comparison



Potential Business Case

"Prevention is better than cure"

- Healthcare providers and Health insurance companies can use this predictive model to identify high-risk patients
- By proactively managing these high-risk individuals, healthcare providers can prioritize care to prevent severe outcomes
- Insurers can reduce claims costs through early intervention

Thank you!!

Govarthini G S