Final Project Report
Introduction to Data Analytics

Project Title:
Prediction/Analysis of Heart Disease using various risk factors.

Prepared by:
Chirag Handa (N01604260)
Bhavesh Waghela (N01639685)

AIGC-5000-0NA - Fall 2023
Humber College

1. Problem Statement

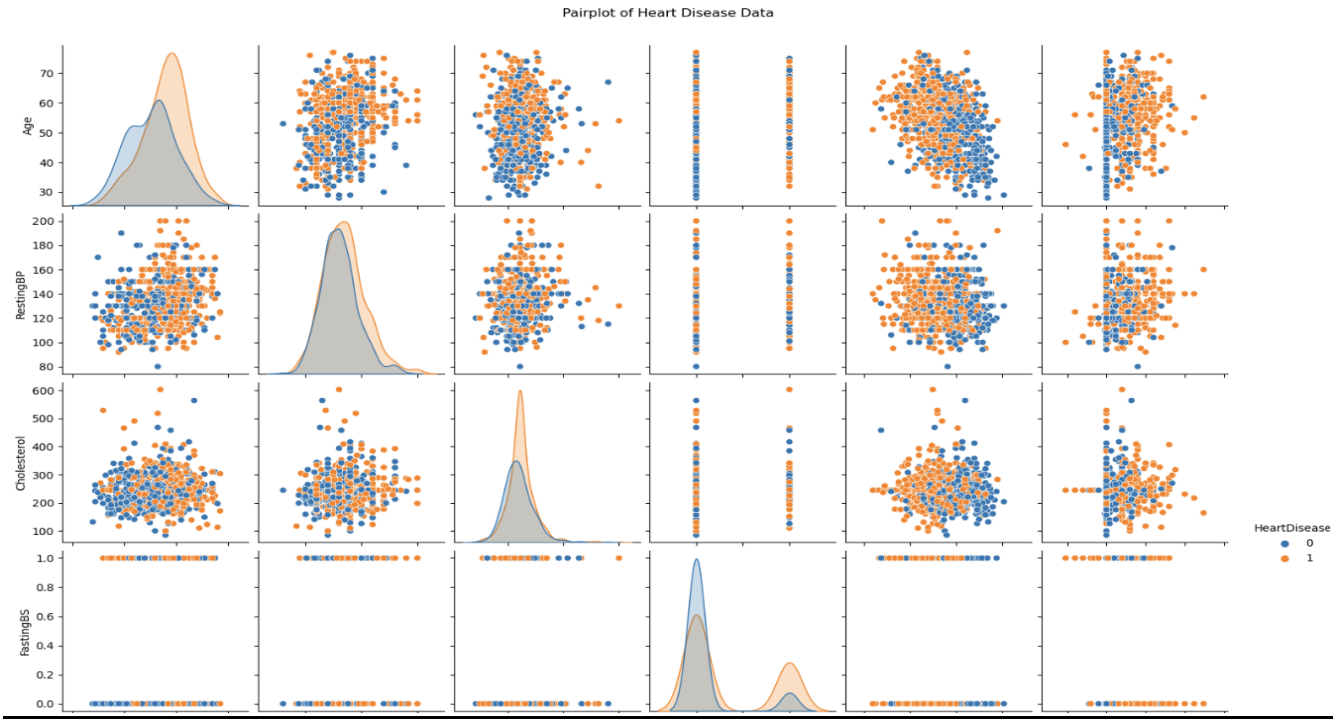
- Prediction/Analysis of Heart Disease using various risk factors.

2. Dataset Description

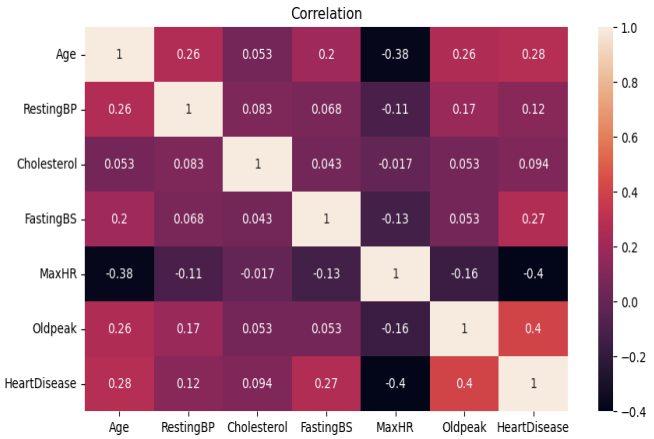
- Heart failure is a common event caused by CVDs and this dataset contains 11 features that can be used to predict a possible heart disease.
- People with cardiovascular disease or who are at high cardiovascular risk (due to the presence of one or more risk factors such as hypertension, diabetes, hyperlipidemia or already established disease) need early detection and management wherein a machine learning model can be of great help.
- The following are the attributes for the heart dataset ie. Age, Sex, ChestPainType, RestingBP, Cholesterol, FastingBS, RestingECG, MaxHR, ExerciseAngina, Oldpeak, ST_Slope, and prediction attribute HeartDisease.

3. Dataset Analysis and Observations

- For dataset analysis I have used pairploit for univariant and bivariant analysis and Heatmap for finding Correlation Coefficients Rank using the above listed columns(few columns are visualized below).
- **Observations:**



From the pairplot analysis it is difficult to understand the relationship between the variables so we would consider a heatmap and correlation matrix for further analysis.



The heatmap plot using spearman method which denotes the relationship between two data variables and return its Correlation Coefficients Rank

- R = 1 Strong positive relationship
- R = 0 Not linearly correlated
- R = -1 Strong negative relationship

From the heatmap I conclude several notable associations in the dataset. Age demonstrates a moderate positive correlation with heart disease, indicating an increased likelihood of cardiovascular issues with age.

Resting blood pressure and maximum heart rate exhibit a moderate negative correlation, implying that higher resting blood pressure may coincide with a lower maximum heart rate during exercise.

Oldpeak, representing ST depression induced by exercise, shows a moderate positive correlation with heart disease, suggesting that greater ST depression is linked to an increased likelihood of cardiovascular issues.

4. Proposed Analytical/Prediction Model

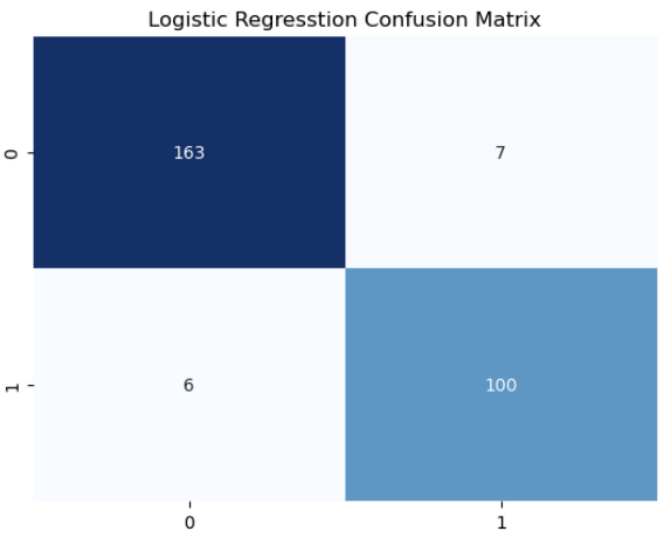
- As above analysis, the dataset has dependent variables for calculating heart disease.
- For predicting the dependent variable, I have used Logistic Regression and Random Forest model.
- For both the models I have processed the dataset encoded the classification variables, transformed the dataset for converting the features to standard scaler, applied the model and retrieved confusion matrix for prediction analysis.

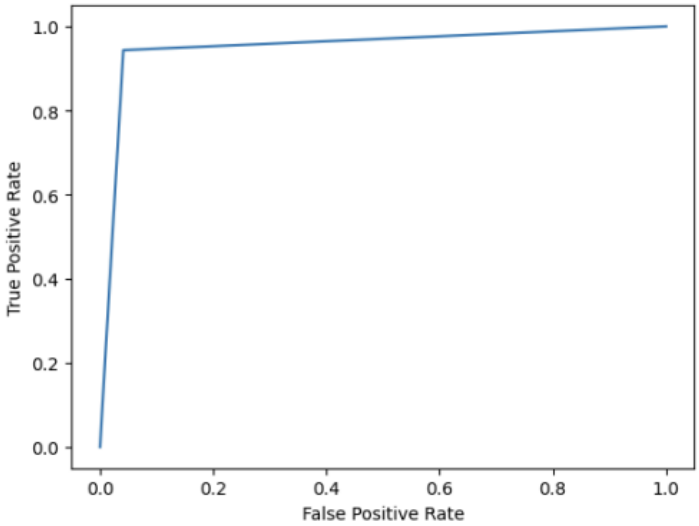
5. Results and Discussions

- Model Performance: Both the models performed well but the best performance that we received is from Logistic Regression model with an accuracy of 95.29% and True Positives classified for 0 = 163 and 1 = 100 respectively.
- Visualization: Confusion Matrix , Classification Report and the ROC Curve for the Logistic Regression Model has been attached for reference.

```
Accuracy =  95.29  %

Text(0.5, 1.0, 'Logistic Regresstion Confusion Matrix')
```





	precision	recall	f1-score	support
0	0.96	0.96	0.96	170
1	0.93	0.94	0.94	106
accuracy			0.95	276
macro avg	0.95	0.95	0.95	276
weighted avg	0.95	0.95	0.95	276

6. Conclusion

- From the Correlation Matrix we can observe that MaxHR (Max Heart Rate) has a highly negative correlation with heart disease. This means that with a high Heart Rate, a person is at a greater risk of getting a heart disease.
- Stress is a big factor in determining Heart Rate so we should try to remain as calm and stress-free as possible for a healthy life.