Heart Failure Prediction

using Machine Learning Algorithms

by M Shaik Kamal Afridi
916387844

1. About Data Set:

- Cardiovascular diseases (CVDs) are the leading cause of death globally, resulting in the deaths 17.9 millions of people each year, representing about 31% of all fatalities. Heart attacks and strokes account for four in every five Cardiovascular deaths, with one-third among these fatalities occurring before the age of 70. CVDs are a common cause of cardiac arrest, and this dataset contains 11 variables that would be used to predict heart disease.
- Patients having heart disease or who are under high stroke risk (due to the existence of one or more conditions such as heart disease, diabetes, hyperlipidemia, or previously existing condition) require early identification and care, which can be aided by a machine learning model.
- **There are <u>918</u> observations and <u>12</u> columns in the dataset.**

2. Data Collection:

- This data set is sourced from Kaggle.
- URL: https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction
- This dataset is a hybrid dataset built from various sources listed below.

Source:

This dataset was created by combining different datasets already available independently but not combined before. In this dataset, 5 heart datasets are combined over 11 common features which makes it the largest heart disease dataset available so far for research purposes. The five datasets used for its curation are:

- Cleveland: 303 observations
- Hungarian: 294 observations
- Switzerland: 123 observations
- Long Beach VA: 200 observations
- Stalog (Heart) Data Set: 270 observations
  - Total: 1190 observations Duplicated: 272 observations
  - Final dataset: 918 observations

Every dataset used can be found under the Index of heart disease datasets from UCI Machine Learning Repository on the following link: https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/

3. Data Pre-Processing:

- There are 0 null values and 0 duplicated values. So, we can consider this as a clean dataset.
- As there are 3 different data types in our dataset i.e., float64(1), int64(6), object (5), I had to implement <u>label encoding</u> to convert non-numeric data to numeric data.
- <u>Feature scaling</u> is used in this project to equalize the range of variables. It is done during the data pre-processing stage.

- Features should be normalized such that no feature is unnecessarily big (centering), and all features are on the same scale (scaling).

- K-NN for example, are sensitive to feature transformations since they rely on dist ances or similarities between data samples. As a result, it is advantage for solving a system of equations, least squares, or other problems where rounding mistakes might cause major problems.

4. Data Dictionary:

Our dataset is in a CSV file format consisting of 12 attributes. As listed in detail about each attribute:

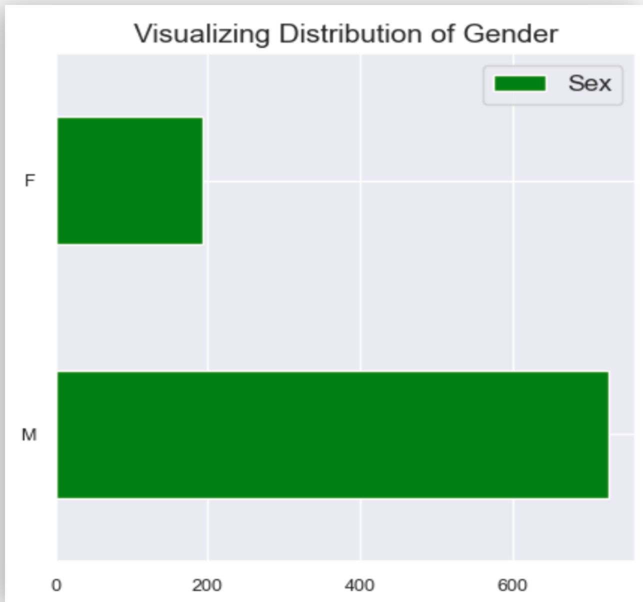
Age	age of the patient [years]
Sex	Gender of the patient [M: Male, F: Female]
ChestPain	chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
RestingBP	resting blood pressure [mm Hg]
Cholesterol	serum cholesterol [mm/dl]
FastingBS	fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
RestingECG	resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
MaxHR	maximum heart rate achieved [Numeric value between 60 and 202]
ExerciseAngina	exercise-induced angina [Y: Yes, N: No]
Oldpeak	oldpeak = ST [Numeric value measured in depression]
ST_Slope	the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
HeartDisease	output class [1: heart disease, 0: Normal]

5. Data Analysis:

Important questions that this project can answer:

I. Which Gender is more prone to heart diseases?

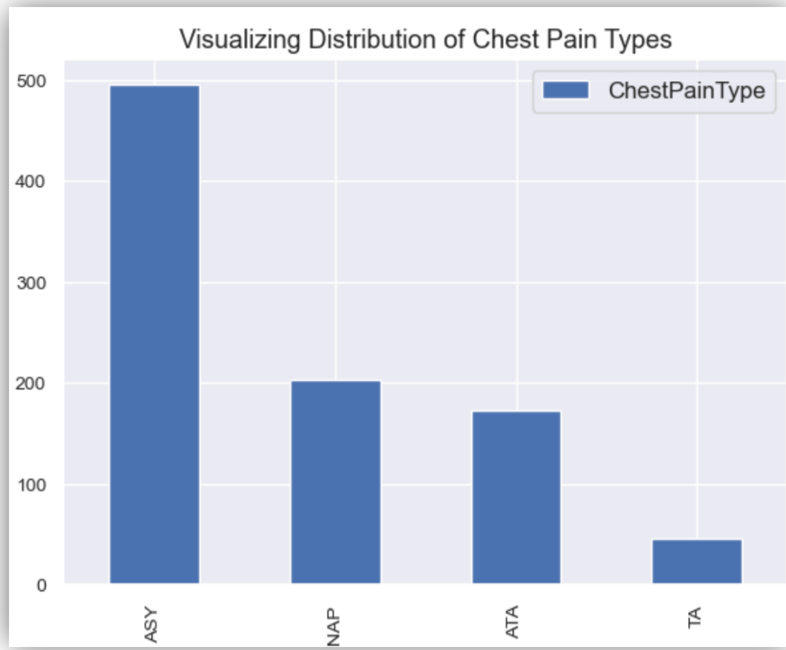
Although we have a greater number of Male samples in Dataset,



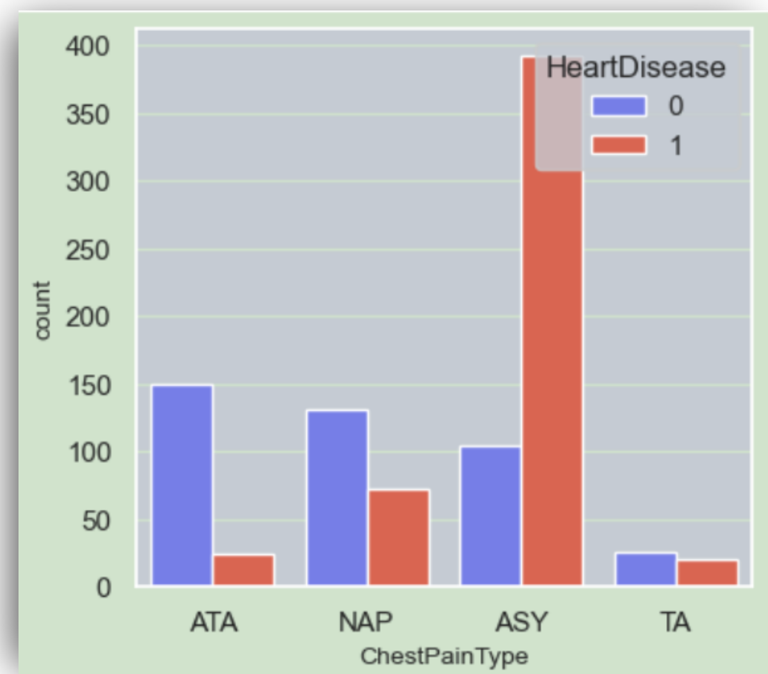
we can still conclude that Male patients are more prone to heart disease than Female patients.

II. **Which is most common type of Chest pain among patients?**

We see that the most common type of Chest Pain in Heart Patients is <u>Asymptomatic</u> with around 400 patients.

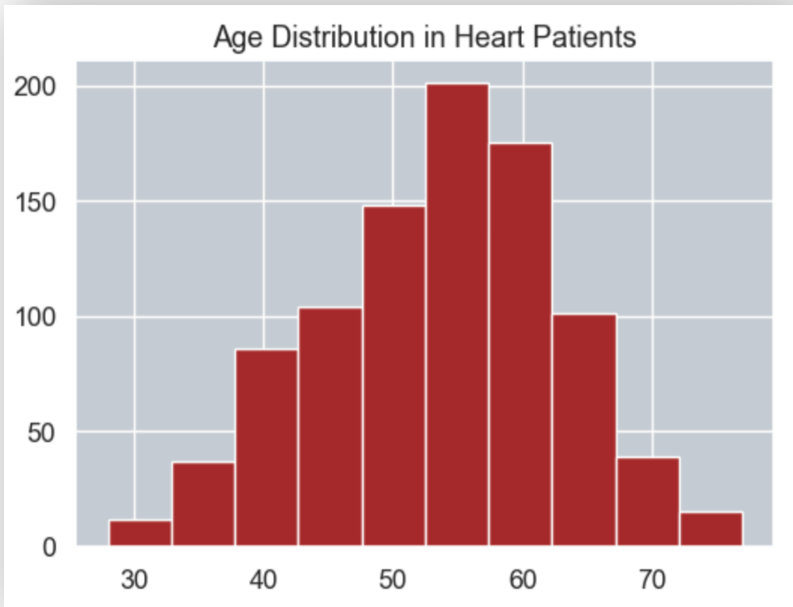


And It's a common symptom in group who suffered heart disease ('1').



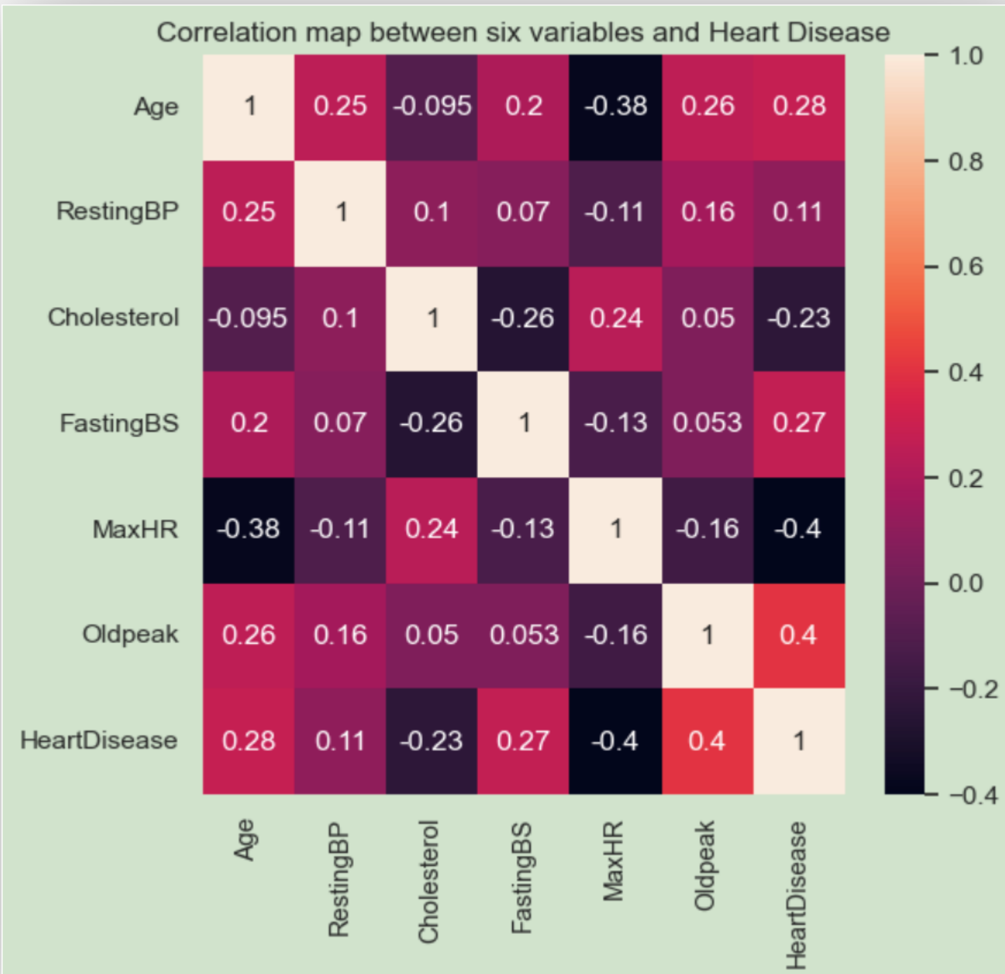
III. Which age group of patients are more vulnerable to heart disease?

We could observe that adults around the age of 60 seem to be the most vulnerable to heart disease.



IV. Which attributes influence heart disease?

Here we can observe that AGE and FASTING BLOOD SUGAR are the two variables that have the greatest impact on heart disease. While Max Heart Rate seems to have the least impact.



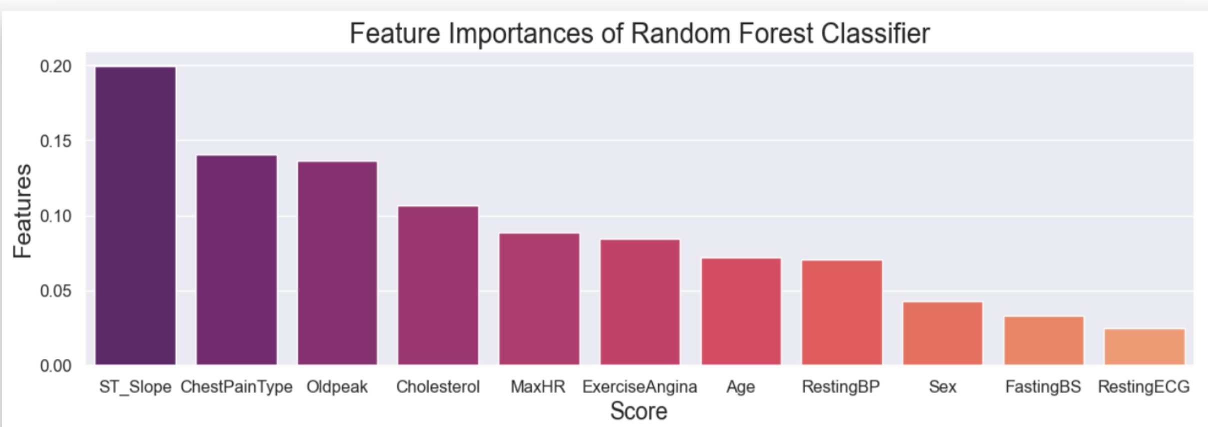
V. Which Features make biggest impact on Prediction of Target Variable using ML Models?

We can see ST_Slope, ChestPainType, Oldpeak are major features impacting model prediction.

As RandomForest classifier is giving best score, let's see what features are influencing most for the decision

```
ft = pd.Series(rf.feature_importances_, index = X_train.columns).sort_values(ascending = False)
print(ft)
```

```
ST_Slope          0.199108
ChestPainType     0.140650
Oldpeak           0.136305
Cholesterol       0.106668
MaxHR             0.088653
ExerciseAngina    0.084797
Age               0.072079
RestingBP         0.070365
Sex               0.043241
FastingBS         0.033259
RestingECG        0.024875
dtype: float64
```



6. References:

- Scikit learn Documentation https://scikit-learn.org/
- Seaborn Documentation https://seaborn.pydata.org/
- Pandas Documentation https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.plot.barh.html
- geeksforgeeks.org/understanding-logistic-regression/
- StackOverFlow Forums
- https://www.analyticsvidhya.com/blog/2021/04/beginners-guide-to-decision-tree-classification-using-python/
- https://machinelearningmastery.com/robust-scaler-transforms-for-machine-learning/
- https://plotly.com/python/templates/
- http://man.hubwiz.com/docset/Seaborn.docset/Contents/Resources/Documents/generated/seaborn.color_palette.html