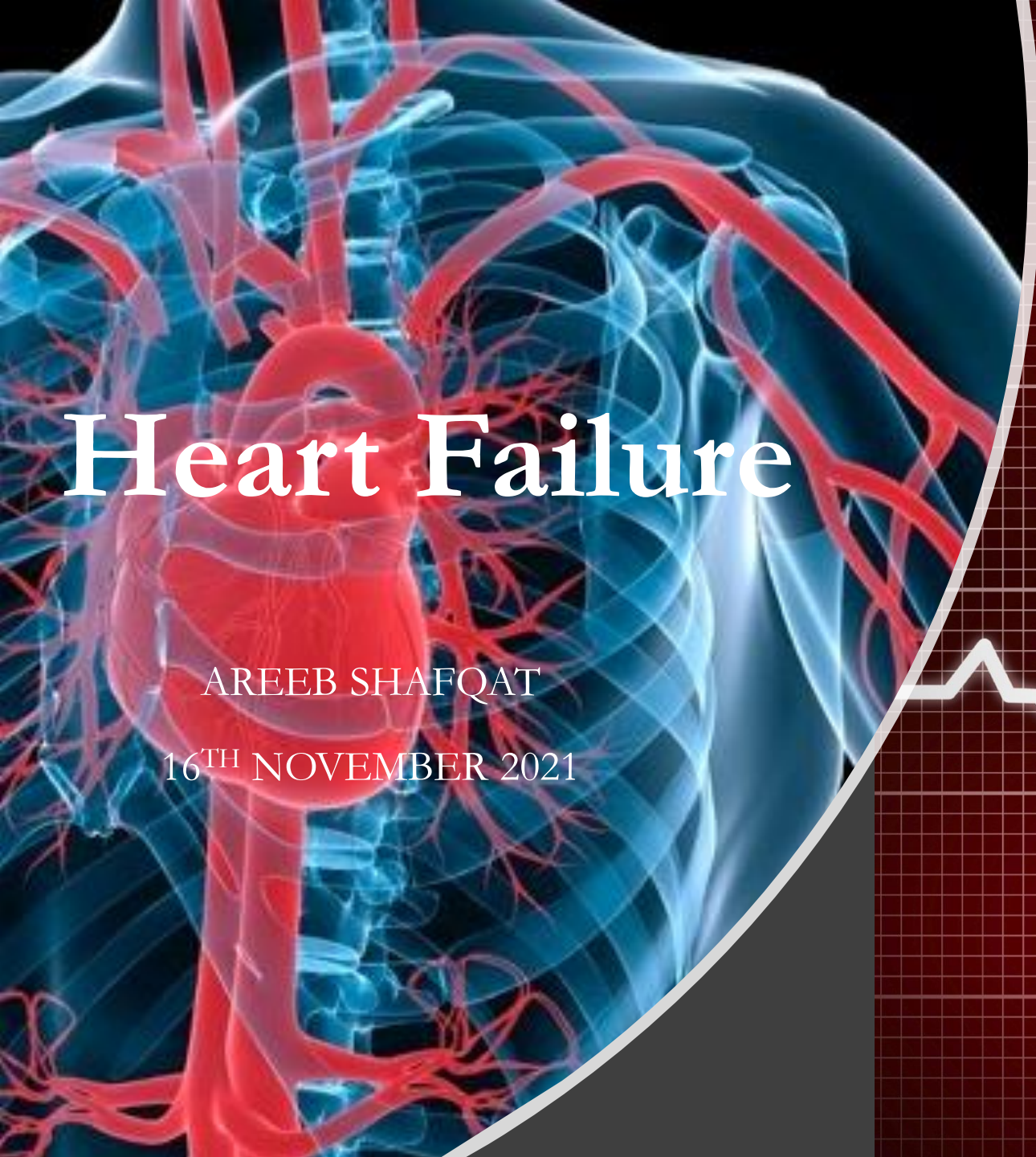


Heart Failure

AREEB SHAFQAT

16TH NOVEMBER 2021

Description of heart dataset

Cardiovascular diseases result in the most deaths worldwide, estimated to be 17.9 million lives each year. This dataset encompasses attributes (918 rows, 12 columns) which can be useful in predicting if someone has a heart disease which are as follows:

- ✓Age of patient (years)
- ✓Gender (M/F)
- ✓Type of chest pain patient has (ATA, NAP, ASY, TA)
- ✓Resting Blood Pressure
- ✓Cholesterol levels
- ✓Resting Electrocardiogram results
- ✓Maximum heart rate of patient
- ✓Fasting Blood Sugar (1 if value >120, 0 otherwise)
- ✓Angina (chest pain) due to exercise
- ✓Old Peak (Slope of peak in exercise relative to rest)
- ✓Slope of maximum or peak exercise (ST_Slope)
- ✓Heart Disease (1 or 0)

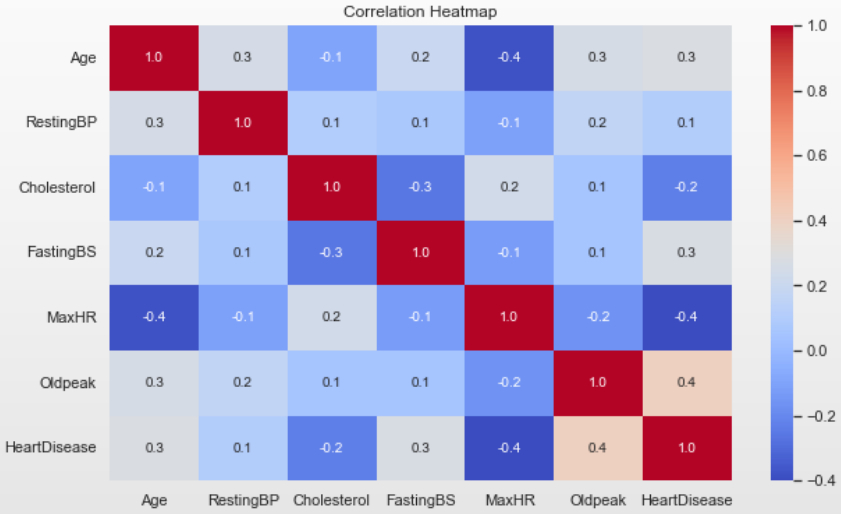
Data Exploration Strategy

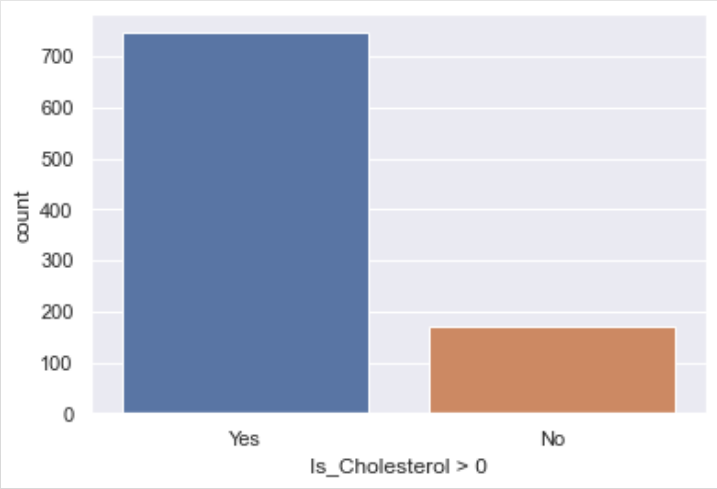
The initial steps for finding key insights from the data are:

- Finding typical features of patients such as what ages, genders, type of chest pains are likely to indicate or result in a heart disease
- Plot bar and scatter charts to find trends between variables
- Using box plots on variables to find outliers and descriptive statistics for patients with heart disease
- Searching for trends in the data to find useful information about heart disease patients

Data Cleaning Steps

- Removing any duplicated and null values, if any.
- Finding any correlations between independent variables to see if any multicollinearity exists
- Removing any outliers through calculating absolute value of z-scores of column values and retaining only a z below 3 standard deviations
- Cholesterol column was removed due to there being several 0 values which acted as replacement for missing values.

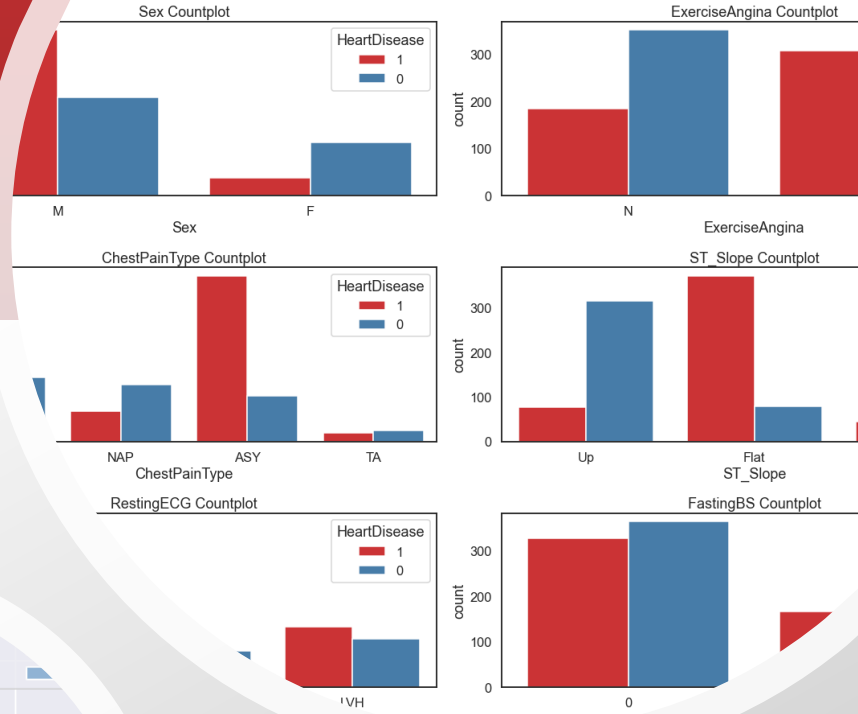


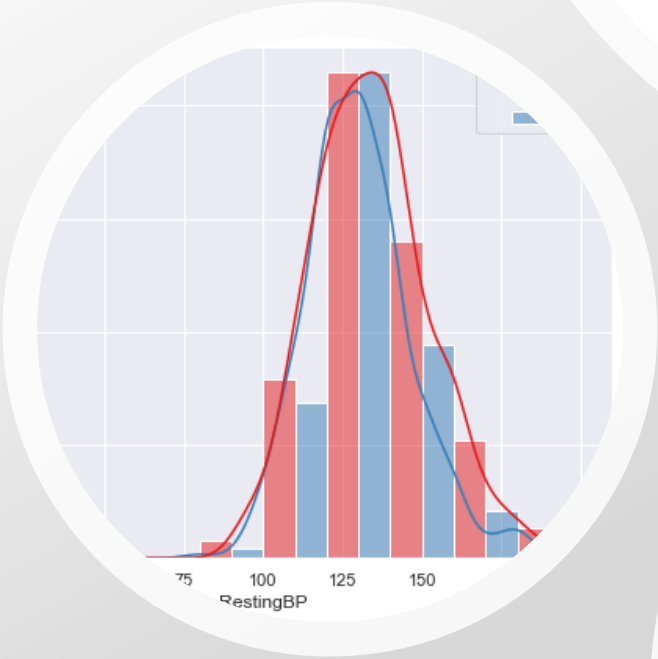


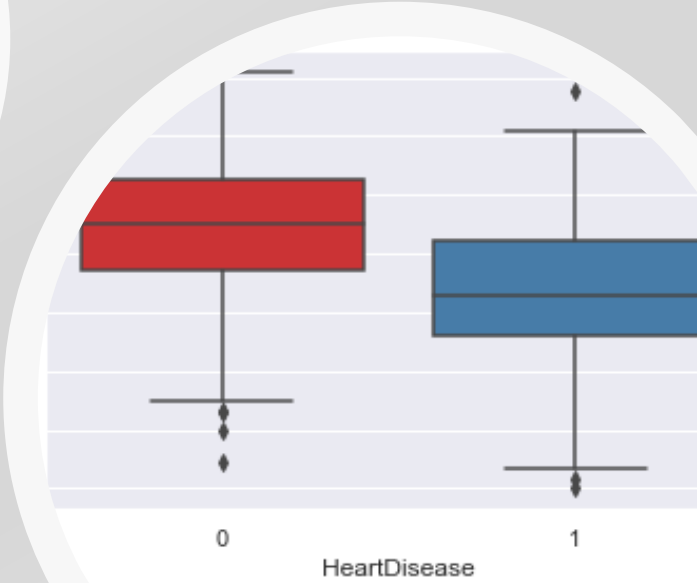
*NB – No feature engineering required

Key Findings from EDA

- Heart disease is more in males than in females
- Patients with ASY type of chest pain, exercise angina, a normal resting ECG, a flat ST slope and high fasting blood sugar levels indicate higher risk of heart disease
- The age of around 60, maximum heart rare approximately 120 and resting blood pressure of 125 are where heart diseases are common.
- Patients with lower ages and high fasting blood pressure are prone to have heart disease







Formulating Hypotheses

HP I – A male patient with asymptomatic chest pains while also being affected by exercise Angina is predicted to have a heart disease 90% of the time.

HP II – A female patient below the age of 30 with a maximum heart rate below 100 is predicted to not have a heart disease 95% of the time.

HP III – With resting blood pressure below 110 and age below 25 does not have a heart disease 80% of the time

Significance testing for HP I

Hypothesis Test:

- Null: These type of patients have a heart disease 90% of the time
- Alternative: There is not a 90% chance of this type of patient having a heart disease
- Choose a significance level of 5%
- Test this with a binomial distribution with a probability of 0.9

Analysis:

- 259 male patients in total have ASY (asymptomatic) chest pain type in addition to Exercise Angina.
- Out of these 259, 239 have a heart disease

Significance testing for HP I

Results:

- Choose a test statistic ~B(259, 0.9)
- Using 1 - Binom.cdf(239, 259, 0.9), we obtain a p value of 0.0884 > 0.05
- This implies that the result is not significance enough and therefore, we reject the null
- To conclude, there is insufficient evidence that there is a 90% chance of a male patient with exercise Angina and asymptomatic chest pain to have a heart disease.

Next Steps and Conclusions

Build a Machine Learning model:

- Train test split the data and use classification models such as logistic regression
- Find features which impact heart disease the most and keep only those for better accuracy

Data Quality and Enhancement

- The dataset is mostly balanced between having a heart disease and not having one
- It contains no null values and duplicates and hence, relatively clean although outliers were present
- There seems to be a lack of features where for example, the Cholesterol column was not used due to containing several 0 value and perhaps, this could be improved upon

Thank You!