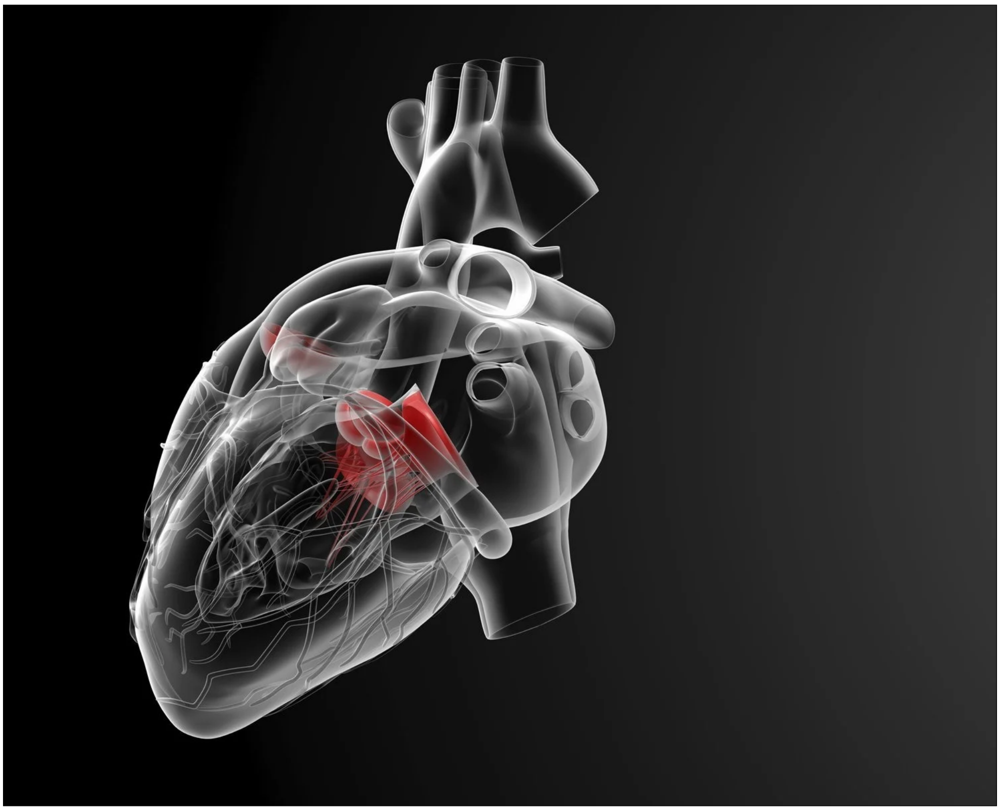
HEART DISEASE PREDICTION



Project Report

Prepared for: Machine Leaning Bootcamp

Prepared by: Meelan Bandara, GTN Technology

9 May 2023

Dataset

Dataset to be used Is from the heart disease dataset collection @https://archive.ics.uci.edu/ml/datasets/Heart+Disease

For the convenience of usage the css version of the same dataset available @https://www.kaggle.com/datasets/aavigan/cleveland-clinic-heart-disease-dataset?select=processed_cleveland.csv was used

Following overall properties can be identified off the considered dataset:

- Number of features - 13
- Number of target variables - 1
- Number of classes in the target variable - 5
- Number of data rows - 303
- Availability of missing/corrupted values - True
- Number of categorical features - 8
- Number of numerical features - 5

Data Preparation

process/pcsv.ipynb and process/pcsv.py holds the data exploration functionality

1. In columns “ca” (4 times) and “thal” (2 times) a usage of “?” Character can be identified.
2. Assumed the publishers used “?” For missing values.
3. Replaced the 6 places with “?” With the mode value of the considered column.
4. The distribution of numerical columns were not standardised.
5. 5 numerical data columns were normalised using the standard normalisation according to,

$$x_{\text{stand}} = \frac{x - \text{mean}(x)}{\text{standard deviation } (x)}$$

6. From the 8 categorical columns, 3 features were identified as binary features.
7. The remaining 5 variables were one hot encoded using the pandas library.

Data Visualisation

- Pandas profiling package was used to generate a visual summary of the dataset.
- After activating the environment the report can be generated by typing “python data-profile.py”
- The exported data visualisations can be viewed at “data-profile.html”
- Scatter plots were drawn between each feature and the target variable.
- Scatter plots are available in the folder “scatter plots”
- Box plots were drawn for each feature column to identify outliers.
- Box plots are available in the folder “box plots”
- Bar charts were drawn to identify the distribution of the categories of categorical features with respective to the target variable.
- Bar charts are available in the folder “bar charts”

Observations Drawn

Followings are the observations drawn by the data exploration phase:

- Non of the features show significantly high correlation between themselves.
- “ca”, “chol”, “cp”, “oldpeak”, “thalach”, “trestbps” columns shows the possibility of having outliers.
- From bar charts,
  - 0 “ca” probable to have no heart disease
  - 0 “exang” probable to have no heart disease
  - 0 “thal” probable to have no heart disease

Classification Approaches

Logistic regression cannot be directly used to achieve the classification task as the algorithms doesn't inherently support multi class classifications. Instead Multinomial logistic regression can be successfully employed to achieve the classification decision for this scenario (or one vs rest method can also be applied).

As the data falls into multi class classification with a single label, following four classification approaches were used, (Achieved accuracies are also presented)

- Multinomial logistic regression - 61.09%
- Extreme Gradient Boosting - 57.1%
- Naive Bayes - 58.78%
- K - Nearest Neighbours - 56.13%

5 Fold cross validation was used to generate a more generalised accuracy result.

Full classification pipeline can be executed by the command “python app.py” after activating the environment.