# Analysis of Heart Disease Data Set

Aniketh Nimma Reddy
RMIT University
s3670774@student.rmit.edu.au

Vikrant Yadav
RMIT University
s3676697@student.rmit.edu.au

Date of report: 26th May 2018

# Table of Contents

## Abstract

The analysis aims to find relationship between heart disease and cardiovascular features of the human body. Heart disease is one of the leading cause of death in Australia with one Australian dead every 12 minutes because of it. (3303.0 - Causes of Death, Australia, 2016, 2018). A model to predict heart disease in a patient can streamline the diagnosis process for a doctor and help them in providing the right care needed by a patient as quickly as possible. The classification problem aims to generalise the relationship among the features. The analysis shows presence of chest pain, sex and blood pressure to be influential features.

## Introduction

Heart disease can be influenced by many factors and understanding their relationship can help doctors and hospitals identify patients who are vulnerable to the disease and help them in planning best way to assist them. The data set used for the analysis has been sourced from the UC Irvine Machine Learning Repository (UCI Machine Learning Repository, n.d.). The website hosts several datasets for training machine learning models. The Heart Disease dataset (insert citation) provides unprocessed data of collected in different experiments from Cleveland, Hungary, Switzerland, and the VA Long Beach. All these datasets contain 76 attributes in total. There are processed versions for each of the datasets which have 14 attributes and 303 observations. The focus of this analysis will be on the processed dataset of the experiment conducted in Cleveland. The understanding of features and required literature has been gathered from the description provided alongside that dataset (Aha, 1988).

The features of the dataset contain information about experiment subjects like their age, sex, chest pain type and cholesterol levels. The target feature represents angiographic disease status with label 0 indicating absence of heart disease and 1, 2, 3 and 4 indicating presence of heart disease. Exploratory analysis aims to discover correlation and relation between the data features among themselves and with target feature.

The classification model being developed aims to create a generalised model to predict if a person has a heart disease or not, therefore the target feature will be transformed to absence of heart disease (=0) and presence of heart disease (=1).

## Methodology

First, we clean and explore all the features in the dataset. The target feature is transformed into boolean feature representing absence and presence of heart disease. Missing values were replaced with the most frequent value observed for nominal features and mean value for numerical data. Different classification models are explored to predict presence of heart disease, namely k-nearest algorithm and decision tree. 20% of observations are randomly selected and set aside for testing the model.

For feature selection for k-nearest neighbours algorithm, simple hill climbing technique is used to identify features which are most influential to the target feature. This process employs the entire dataset. Following feature selection, we divide the dataset and set aside 20% of observation of testing the model's performance. The resultant training dataset has 242 observations. Leveraging the high computing capacity relative to size of training set, leave 1-out approach is used to create 242 folds within the dataset and average score from all the folds is considered the average score of the model. For each model the classifier is fitted with value of k ranging from 1 to 50 and average score from all the folds recorded. Other parameters were also tuned to achieve maximum accuracy score and k-nearest neighbour using Manhattan metric was found to give the best accuracy score. The model with highest accuracy was then tested against the test datasets originally set aside which had 61 observations.

Decision tree modelling was performed by setting minimum size of leaf nodes to 10 and like KNN modelling dividing the dataset by 80:20 ratio into training and test dataset.

## Results

The dataset has 303 observations with many interesting cardiovascular features of patients. All of the features are explored to see if they provide any inference which can improve the understanding of

relationship among the features. Following exploratory analysis, we analyse the results of the classification models employed to predict presence of heart disease.

## Exploratory analysis

The dataset has many features relating to age, sex and cardiovascular attributes of the subjects. They are explored in order to provide better understanding.

### Age

The subjects of the experiment have age ranging from 29 to 77 with mean value of 54.438 and standard deviation of 9.038. Figure 1 shows histogram of age of the subjects.

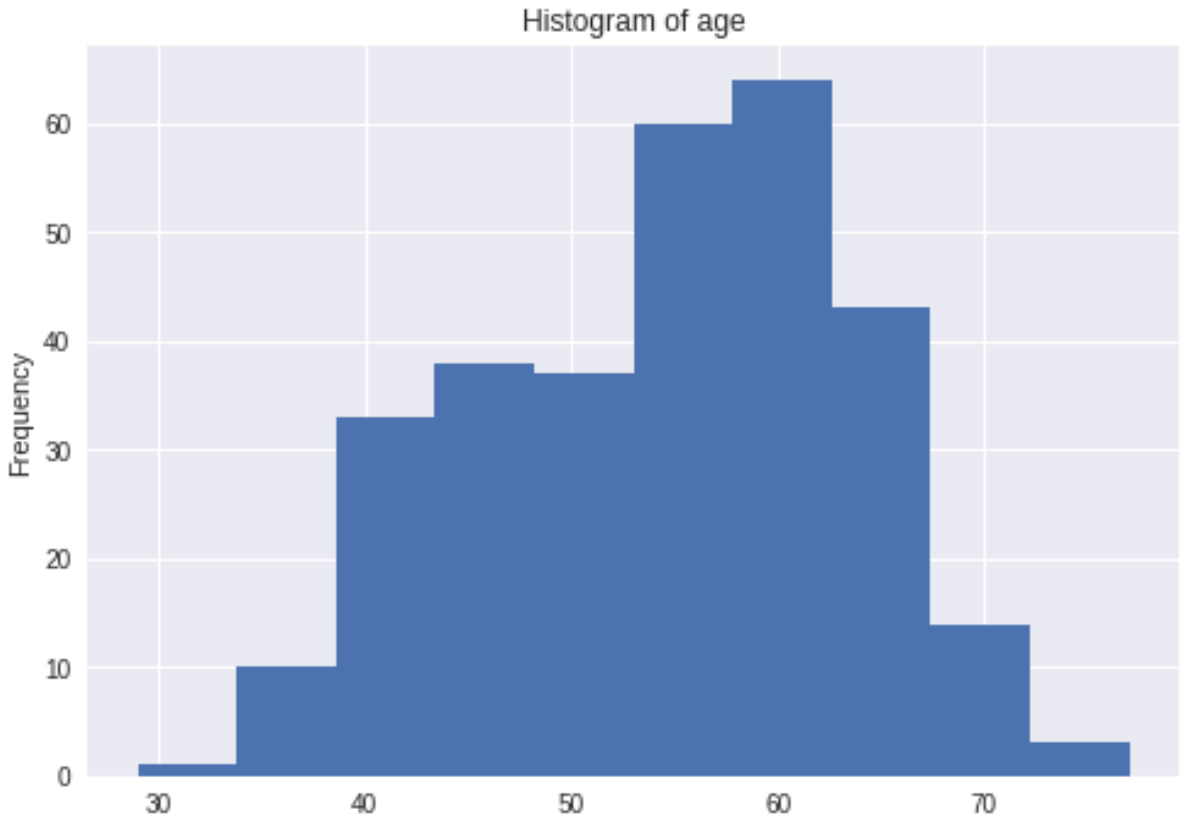


*Figure 1Histogram of age*

### Sex

The sex of subjects has been given as a numerical value with 1 representing male and 0 representing female subjects. From Figure 2, it can be observed that 67.99% of the subjects are make and 32.01% of subjects are female.

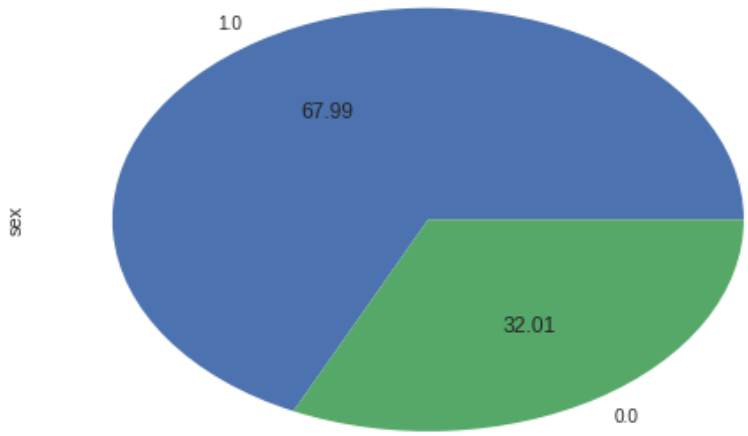


*Figure 2Percentage distribution of sex of subjects*

## Chest pain

The dataset categorises chest pain into four categories, typical angina (1), atypical angina (2), non-anginal pain (3) and asymptomatic (4). Figure 3 shows that most of the chest pain types are asymptomatic, followed by non-anginal pain.

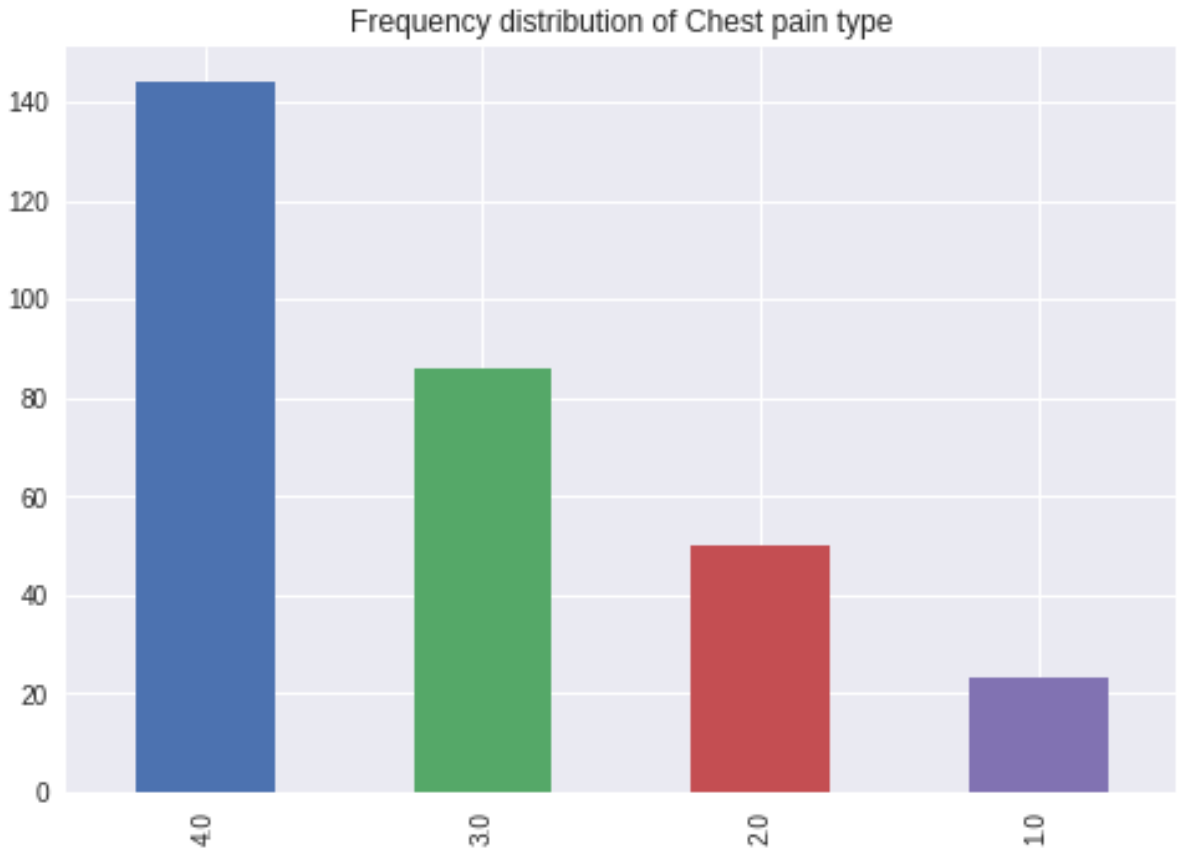


*Figure 3Frequency distribution of chest pain types*

## Resting blood pressure

Resting blood pressure if the blood pressure recorded by hospitals when the patient is admitted. Values are recorded in mmHg.

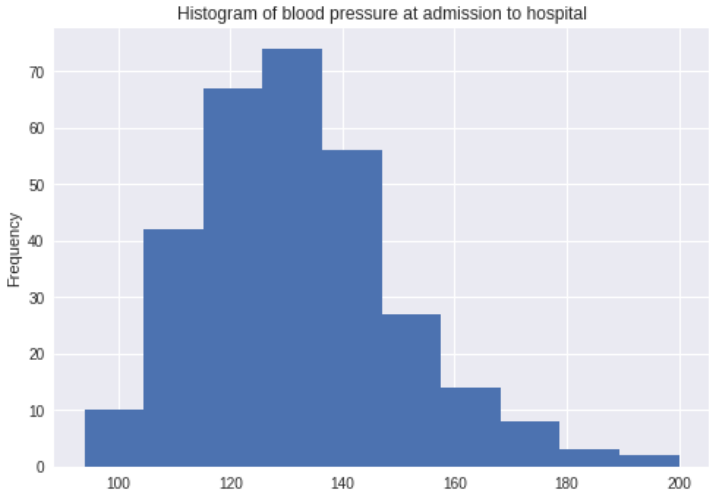


*Figure 4Histogram of resting blood pressure*

The histogram shows patients to be having higher than normal blood pressure. The histogram in Figure 4 seems to be skewed to right.

## Cholesterol

Cholesterol levels are also regarded as good indicator to cardiovascular health of a person. The cholesterol levels of patients range from 126mg/dl to 564mg/dl with mean value of 246.693mg/dl and standard deviation of 51.776.

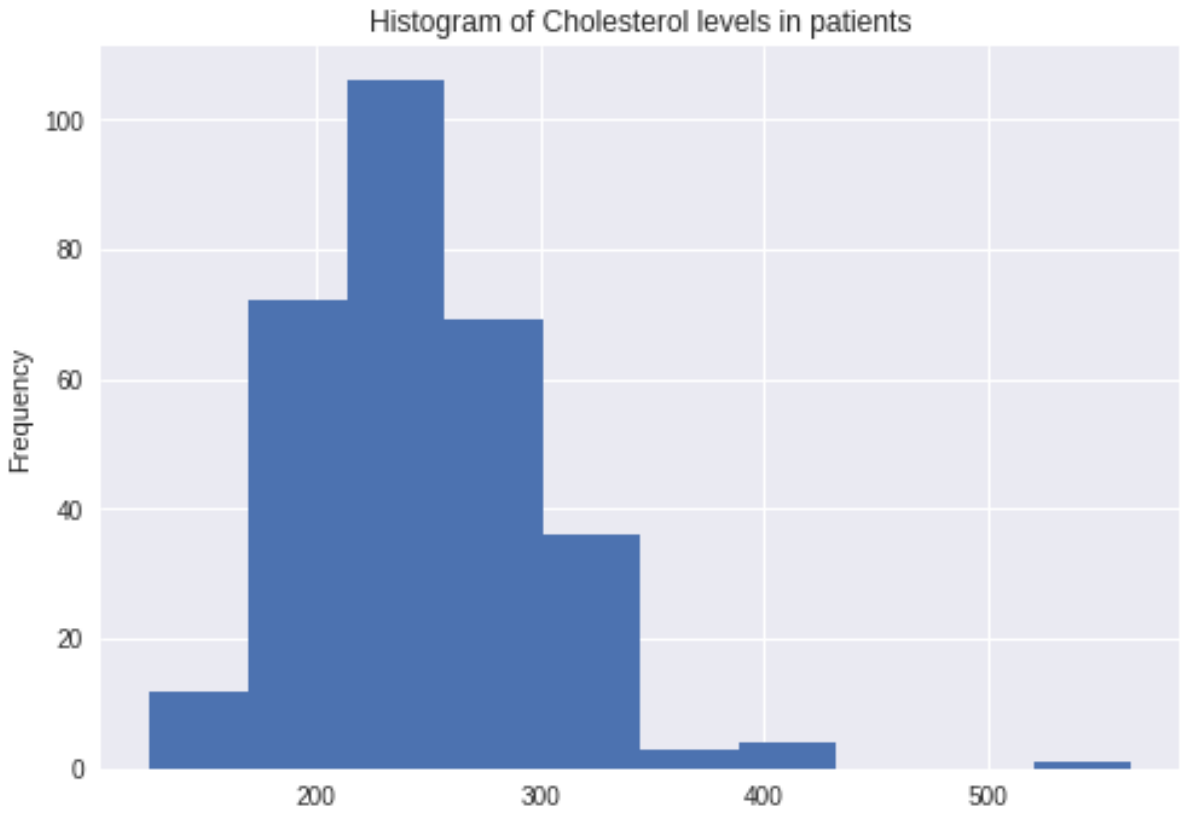


*Figure 5Histogram of Cholesterol level of patients*

## Fasting blood sugar

Fasting blood sugar is a categorical feature with value 1 if the blood sugar is greater than 120mg/dl and 0 otherwise. Figure 6 shows that 14.85% of admitted patients have blood sugar level greater than 120mg/dl.

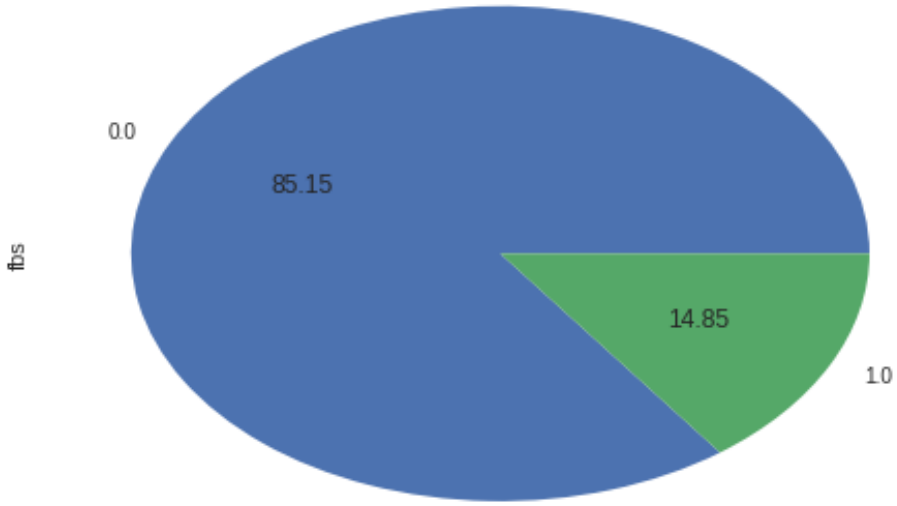


*Figure 6Percentage distribution of fasting blood sugar*

## Resting electrocardiographic result

The dataset provides result of electrocardiographic when patients are admitted and categorises them into three categories of normal (0), having ST-T wave abnormality (1) and showing probable or definite left ventricular hypertrophy (2). Figure 7 shows that there almost equal counts of patients with normal values and patients having STT wave abnormality.

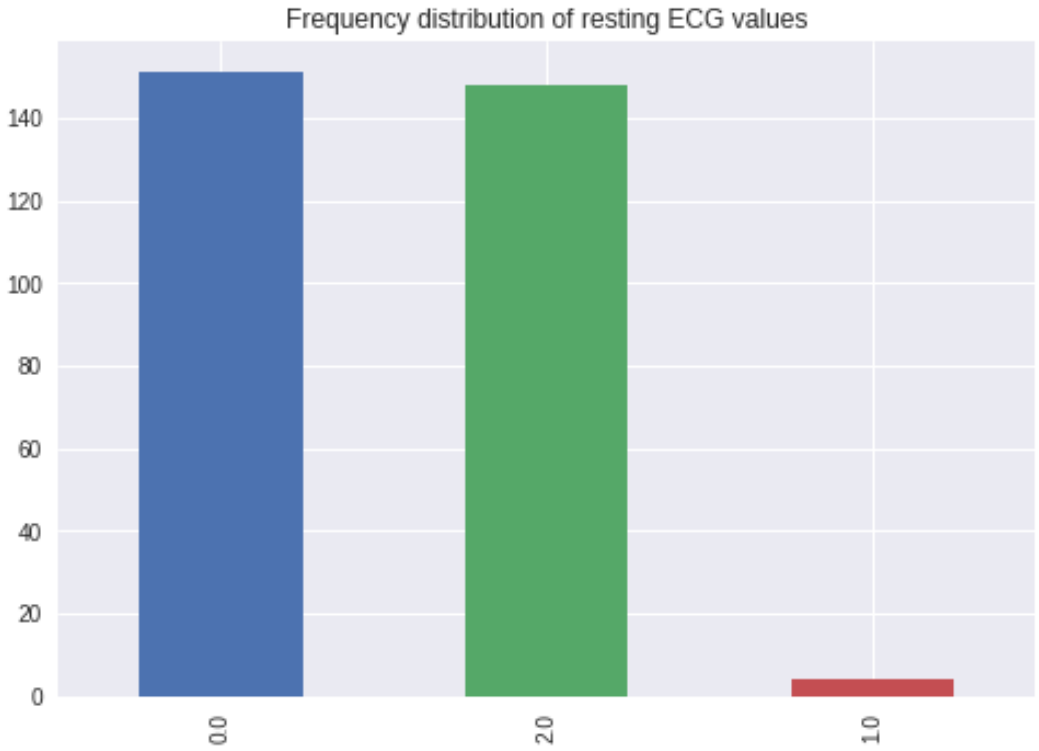


*Figure 7Frequency distribution of resting ECG*

## Maximum Heart rate achieved

The maximum heart rate achieved is the highest heart rate achieved by patients during care. The histogram shows higher values of heart rate than those recorded at time of admission of hospital.

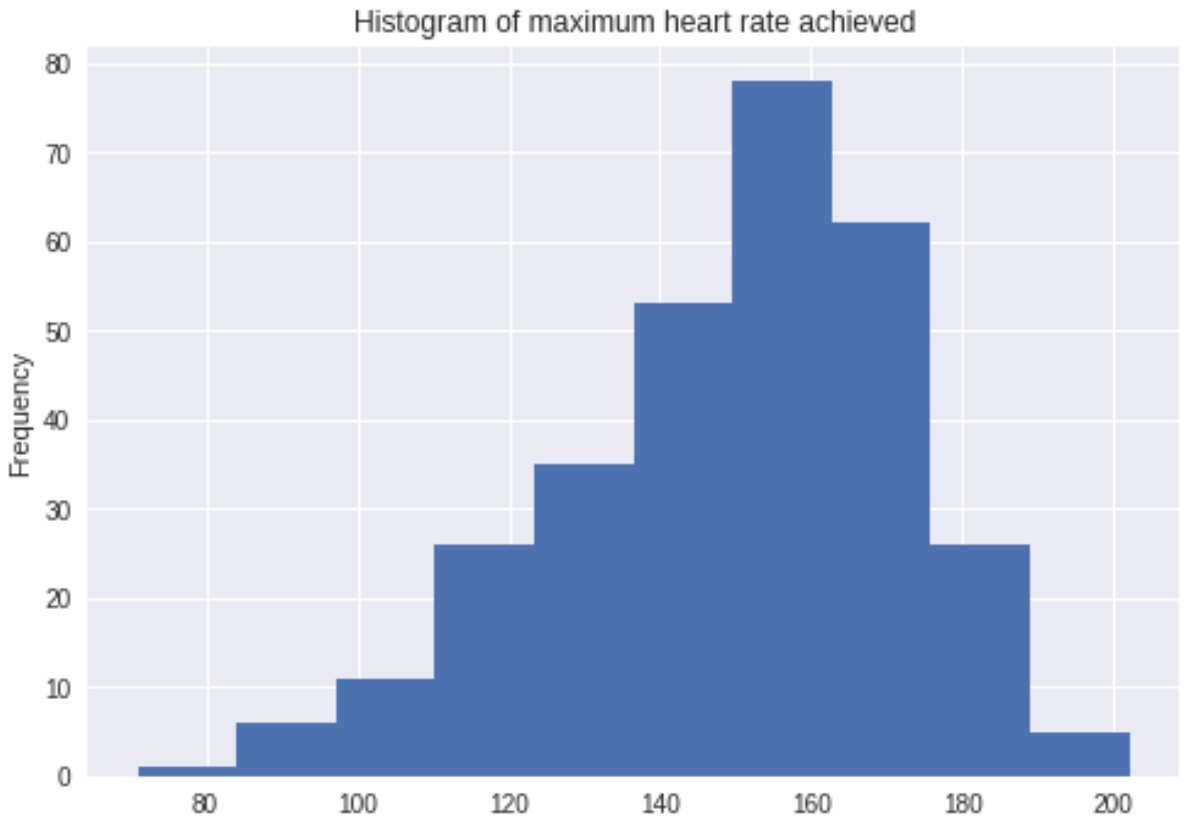


*Figure 8Histogram of maximum heart rate achieved*

## Heart defect

Type of heart defect have been categorised into three categories, normal (3), fixed defect (6) and reversable defect (7). Figure 9 shows majority of admitted patients have no defect in heart (normal) followed by people with reversible defect.

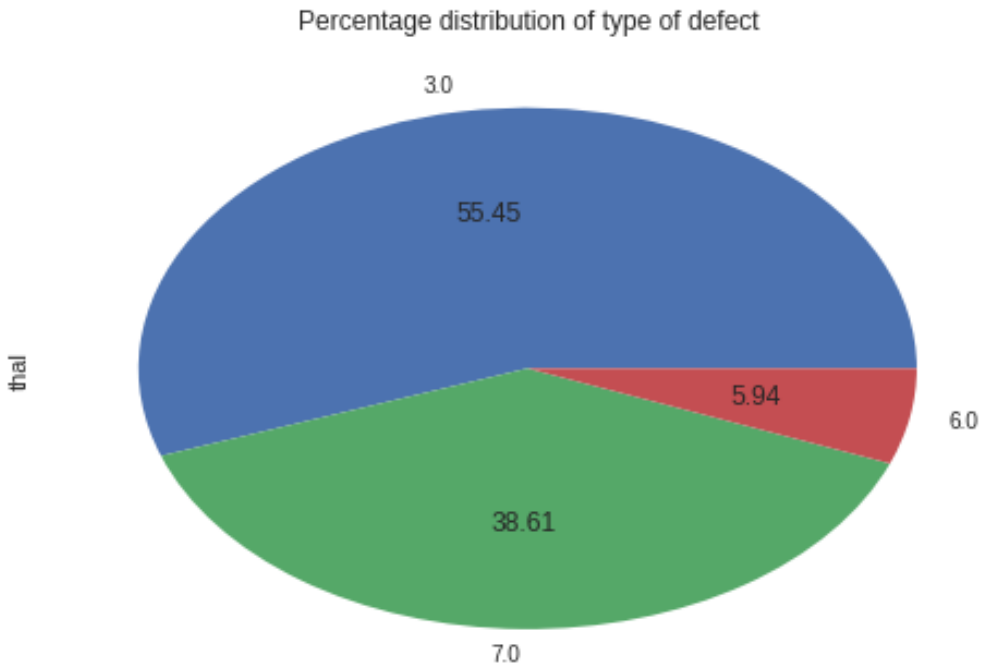


*Figure 9Percentage distribution of type of defect*

## Heart disease (target feature)

The target feature represents presence or absence of heart disease. The Cleveland experiment has categorised it as a multiclass feature but for the scope of this analysis, it has been reduced to a Boolean feature with 0 indicating absence of any heart disease and 1 indicating presence of hear disease. Percentage distribution of the feature can be seen in Figure 10

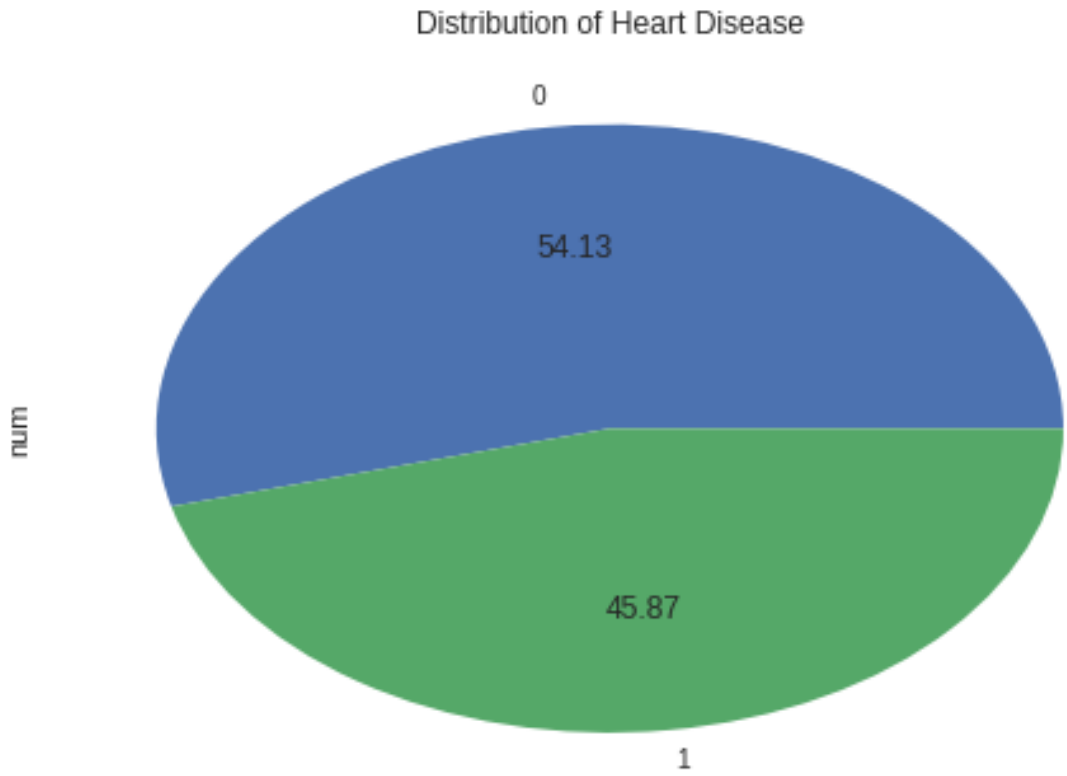


*Figure 10Pie chart of target feature*

## Scatter matrix

The scatter matrix shown in Figure 11 shows a scatter matric of all features with blue points representing absence of heart disease and red points representing presence of heart disease. Patterns are visible between target feature and features like age, sex and blood pressure at rest.

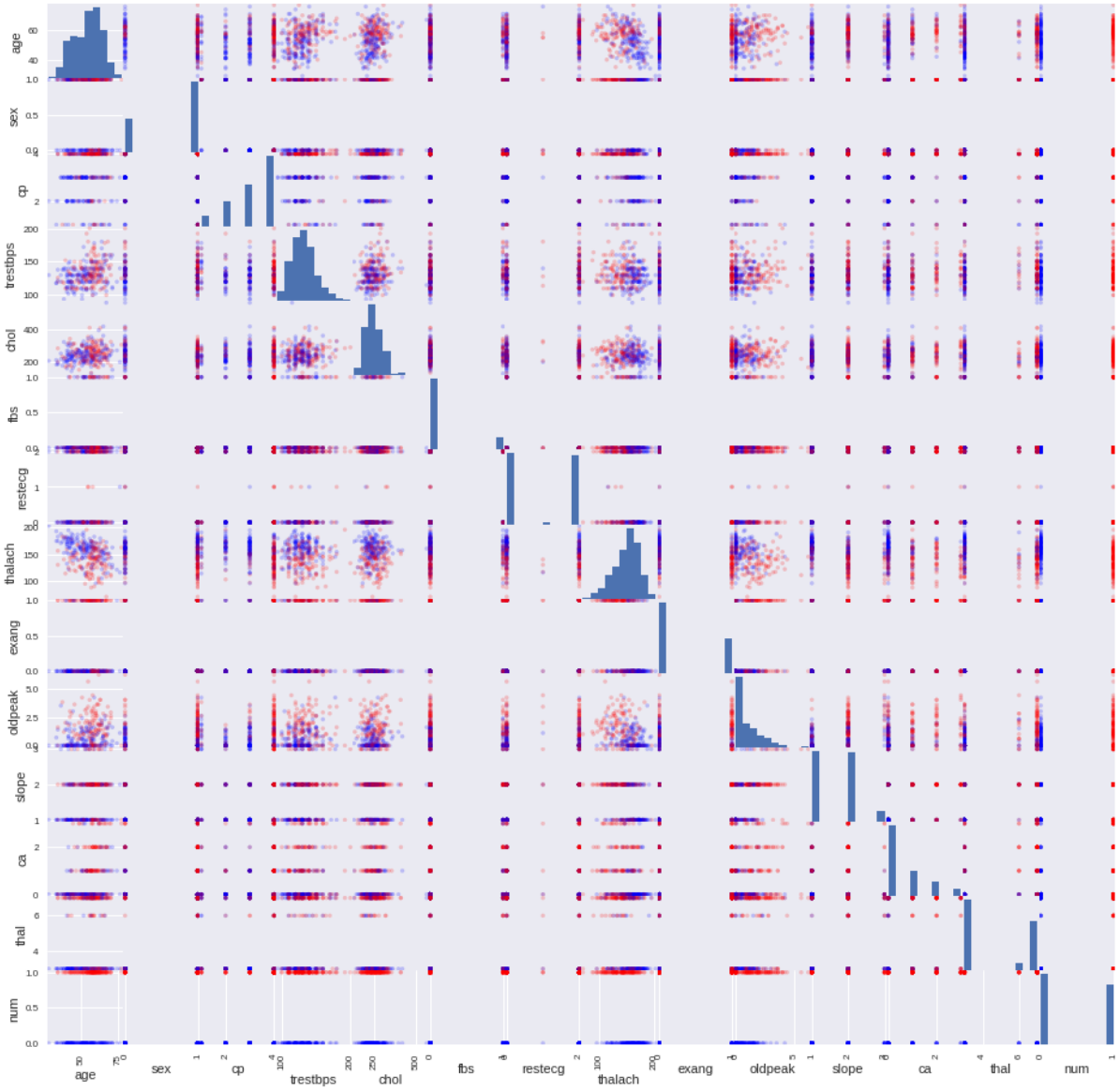


*Figure 11Scatter matrix of all features*

## Data modelling

The aim of classification model is to correctly predict the target feature. Two classification models, k-nearest neighbour and decision tree are used for this purpose.

### k-nearest neighbour

Before applying the classification model, it is must that features are selected which will be used for training the model. Simple hill climbing approach is used to identify give provide a model with maximum score. For this purpose, the entire dataset is used. The approach identified 8 features to be used in the KNN model namely, chest pain, sex, resting electrocardiographic result, type of heart defect, ST depression induced by exercise, exercise induced angina, number of vessels coloured by fluoroscopy and fasting blood sugar.

After testing different hyperparameter tuning, Manhattan distance metric has provided the best result. Figure 12 shows accuracy values against various values of k. The accuracy values have been

determined by calculating mean value of all the accuracy value obtained using leave 1-out approach. From the plot we can see that the simplest model with highest accuracy is obtained at k=7. Now this model is validated against the test data.


*Figure 12Accuracy of KNN model for different values of k*

On testing the validity of model against the test data, the classification error rate obtained is 0.163. The confusion matrix obtained is given in Table 1. There are 8 false positives and 2 false negatives.

| | |
|---|---|
| 33 | 2 |
| 8 | 18 |

*Table 1Confusion matrix of KNN model*

The classification report provides precision, recall and f1-score of the model's prediction for both the classes and model itself.

| | Precision | recall | F1-score | Support |
|---|---|---|---|---|
| 0 | 0.80 | 0.94 | 0.87 | 35 |
| 1 | 0.90 | 0.69 | 0.78 | 26 |
| Avg / total | 0.85 | 0.84 | 0.83 | 61 |

*Table 2Classification report of KNN mode*

Table 2 shows the KNN model has a precision of 0.85 and recall of 0.84 with F1-score of 0.83.

## Decision tree

The decision tree model is shown in Figure 13. The model was generated by setting minimum sample size of leaf nodes as 10.

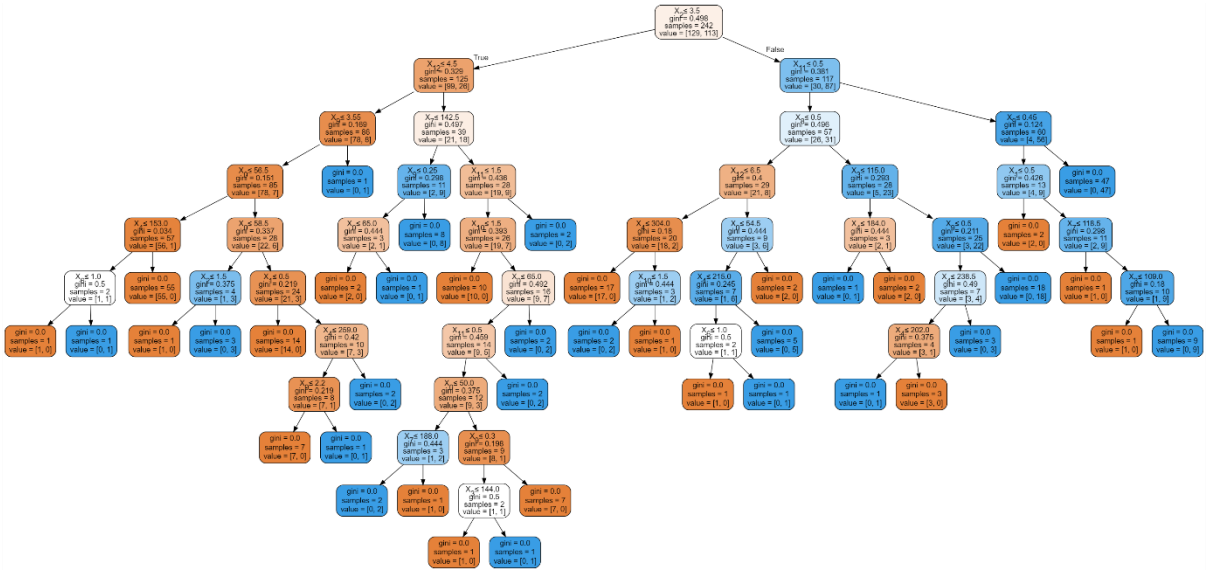

*Figure 13Decision tree model*

Table 3 and Table 4 show the confusion matric and classification report of decision tree. The model has classification error rate of 0.21.

| | |
|---|---|
| 30 | 5 |
| 8 | 18 |

*Table 3Confusion matrix of decision tree model*

| | Precision | recall | F1-score | Support |
|---|---|---|---|---|
| 0 | 0.81 | 0.71 | 0.76 | 35 |
| 1 | 0.67 | 0.77 | 0.71 | 26 |
| Avg / total | 0.75 | 0.74 | 0.74 | 61 |

*Table 4Classification report of decision tree model*

## Discussion

The analysis helps in identifying features which are influential in determination of presence of heart disease. Contrary to initial intuition there were attributes more influential than age of the patient in determining presence of heart disease. This provides more strength to the belief that heart disease are related to blood pressure and cholesterol levels of a person.

## Conclusion

The classification models provide interesting insight into relationship of cardiovascular attributes and presence of heart disease. This knowledge can be used in making the diagnosis process more efficient and provide care to susceptible patients at the right time. The KNN model was able to provide precision of 0.85 with F1-score of 0.83. Additional research should be performed in modelling the risk of having a heart disease by analysing eating habits and how much people exercise so it can be used people themselves to monitor their health.

## References

*3303.0 - Causes of Death, Australia, 2016*. (2018, March 20). Retrieved from Australian Bureau of Statistics: http://www.abs.gov.au/ausstats/abs@.nsf/Lookup/by%20Subject/3303.0~2016~Main%20Features~Australia%27s%20leading%20causes%20of%20death,%202016~3

Aha, D. (1988, July 22). *Heart Disease Names*. Retrieved from https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/heart-disease.names

*UCI Machine Learning Repository*. (n.d.). Retrieved from UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/index.php