**Heart Disease UCI:**

The term “Heart Disease” is often used interchangeably with the term “Cardiovascular Disease”. Cardiovascular disease generally refers to conditions that involve narrowed or blocked blood vessels that can lead to a heart attack, chest pain (angina) or stroke. Other heart conditions, such as those that affect your heart’s muscle, valves or rhythm, also are considered forms of heart disease.

Prediction of cardiovascular disease is regarded as one of the most important subjects in the section of clinical data analysis. The amount of data in the healthcare industry is huge. Data mining turns the large collection of raw healthcare data into information that can help to make informed decisions and predictions.

Heart Disease is a major concern to be dealt with. But it is difficult to identify heart disease because of several contributory risk factors such as diabetes, high blood pressure, high cholesterol, abnormal pulse rate, and many other factors. Due to such constraints, scientists have turned towards modern approaches like Data Mining and Machine Learning for predicting the disease.

The dataset used in this article is the Cleveland Heart Disease dataset taken from the UCI repository. (<https://www.kaggle.com/ronitf/heart-disease-uci>)

Link to dataset: [heart.csv](https://www.kaggle.com/ronitf/heart-disease-uci/download/DtBEBYWYc4oADuSll7Tu%2Fversions%2Fio5hOzMZw7oE0LCEiuBm%2Ffiles%2Fheart.csv?datasetVersionNumber=1)

There are 14 columns in the dataset, which are described below.

1. Age: displays the age of the individual.
2. Sex: displays the gender of the individual using the following format :  
   1 = male  
   0 = female
3. Chest-pain type: displays the type of chest-pain experienced by the individual using the following format :  
   1 = typical angina  
   2 = atypical angina  
   3 = non - anginal pain  
   4 = asymptotic
4. Resting Blood Pressure: displays the resting blood pressure value of an individual in mmHg (unit)
5. Serum Cholestrol: displays the serum cholesterol in mg/dl (unit)
6. Fasting Blood Sugar: compares the fasting blood sugar value of an individual with 120mg/dl.  
   If fasting blood sugar > 120mg/dl then : 1 (true)  
   else : 0 (false)
7. Resting ECG : displays resting electrocardiographic results  
   0 = normal  
   1 = having ST-T wave abnormality  
   2 = left ventricular hyperthrophy
8. Max heart rate achieved: displays the max heart rate achieved by an individual.
9. Exercise induced angina :  
   1 = yes  
   0 = no
10. ST depression induced by exercise relative to rest: displays the value which is an integer or float.
11. Peak exercise ST segment :  
    1 = upsloping  
    2 = flat  
    3 = downsloping
12. Number of major vessels (0–3) colored by flourosopy: displays the value as integer or float.
13. Thal : displays the thalassemia :  
    3 = normal  
    6 = fixed defect  
    7 = reversible defect
14. Diagnosis of heart disease: Displays whether the individual is suffering from heart disease or not:   
    0 = absence  
    1, 2, 3, 4 = present.

**Original Analysis Case Study Part -1**

Graph Analysis part of the case study to predict Heart Disease in patients:

1. Load the data from the “train.csv” file into a Data Frame.
2. Display the dimensions of the file.
3. Display the first 5 rows of data so you can see the column headings and the type of data for each column.
   * Notice that target variable is represented as a 1 or 0
   * Notice that there is no missing data.
   * The target variable will be the “target” and the other variables will be the “features”.
4. Some questions that might help to predict heart disease in patients:
   * What do the variables look like? For example, are they numerical or categorical data? If they are numerical, what are their distribution; if they are categorical, how many are they in different categories?
   * Are the numerical variables correlated?
   * Are the distributions of numerical variables the same or different among patients with heart disease and without heart disease? How other variables like chol, age, and cp impact the result?
   * How the result varies in male and female?
5. Look at summary information about the data (total, mean, min, max, freq, unique, etc.) Does this present any more questions for you? Does it lead you to a conclusion yet?
6. Make some histograms of your data (“A picture is worth a thousand words!”)

From the histograms we can notice that

* Most number of patients are between 50-60 years of age
* Resting blood pressure is 120 for maximum number of patients.
* Serum cholestoral in mg/dl for maximum number of patients is between 200-300.
* Most number of patients reported chest pain type 0.

1. Make some bar charts for variables with only a few options.

From the bar charts we can conclude that:

* There is more number of data available for male patients
* Patients with presence of heart disease are more.
* More patients are with number of major vessels colored by fluoroscopy is 0.
* Fasting blood sugar > 120 mg/dl for fewer number of patients.

1. To see if the data is correlated, make some Pearson Ranking charts

From Pearson Ranking chart we can conclude that:

* There is some positive correlation between age, trestbps and age, chol.
* There is some negative correlation between age, cp and chol, cp.
* There is no correlation between cp and trestbps.

1. Use Parallel Coordinates visualization to compare the distributions of numerical variables between passengers that survived and those that did not survive.

From Parallel Coordinates visualization we can conclude that:

* More number of male patients have the presence of heart disease.
* As the chol value increases the presence of heart disease also increased.
* As the number of major vessels colored by fluoroscopy increases the presence of heart disease also increased.
* Age doesn't have much impact on the presence of heart disease.

1. Use Stack Bar Charts to compare passengers who survived to passengers who didn’t survive based on the other variables.

From the stacked bar charts we can conclude that:

* The percentage of presence of heart disease is more in female patients.
* Chest pain type 2 had more number of heart disease patients.
* Heart disease is not dependent on fasting blood sugar
* More number of heart disease patients does not have exercise induced angina.

**Original Analysis Case Study Part -2**

1. Looking at information of heart disease risk factors led me to the following:
   1. Find any correlation between the variables and reduce some of the features.

I couldn't see any strong correlation between the variables. But slope and oldpeak has some negative correlation. So I have decided to keep all the variables in the dataset.

* 1. Find any missing values and fill them.

No missing data in the data set. So we can move to the next step.

1. Look for skewed data and transform it.
2. Convert categorical data into numbers.

'fbs':{ 1: 'true', 0: 'false'}, 'target': {0:"No disease", 1:"Disease", 'sex': {1: 'male', 0: 'female'}

**Original Analysis Case Study Part -3**

14. Training - Split the data into two sets: Training and Testing.

15. Evaluation – Here we are trying to predict if a patient has heart disease or not. I am going to use logistic regression for prediction. Further I would like to like to use different algorithms and compare the results.

a. Metrics for the evaluation:

i. Confusion Matrix: Accuracy of the model is 79%.

ii. Precision, Recall & F1 score:

Precision: 79.5%

Recall: 83.8%

F1 score: 81.6%

iii. ROC curve: Plotted the ROC curve which are above the dotted line which concludes my model is pretty much better than the randomly guessed.

**Original Analysis Case Study Part -4**

16. I have used different classifiers for my analysis and compared the results.

* + Accuracy of Logistic Regression model is: 81.96%
  + Accuracy of DecisionTreeClassifier is: 70.49%
  + Accuracy of RandomForestClassifier is: 78.6%

All the ROC curves are above the dotted lines which signifies the predicted values are much better than the randomly guessed values.

17. Finally I have created a Binary Classification Problem using keras and got the model accuracy as 77%.

From my analysis I can conclude that Logistic Regression model is best for this dataset for which maximum accuracy was achieved. Further the dataset is small if we can train the model with larger data, we can hopefully get better results.