Heart disease prediction

Abstract: In this project, data analytic techniques are used in detecting whether a person has a heart disease or not. A lot of people suffer from cardiovascular diseases (CVDs), which even cost people their lives all around the world. Data analytics can be used to detect whether a person is suffering from a cardiovascular disease by considering certain attributes like chest pain, cholesterol level, age of the person, etc. . Classification algorithms based on supervised learning can make diagnosis of cardiovascular diseases easy. Four supervised machine learning algorithms are used in this paper which are, K-Nearest Neighbor (K-NN), Logistic regression, Naïve Bayes and Decision tree.

Keywords: Heart Disease, K-Nearest Neighbor (K-NN), Logistic regression, Naïve Bayes, Decision tree

Introduction: Human body is made up of various organs, all of which have their own functions. Heart is one such organ which pumps blood throughout the body and if it does not do so, the human body can have fatal circumstances. One of the main reasons of mortality today is having a heart disease. So, it becomes necessary to make sure that our cardiovascular system or any other system in the human body for that matter must remain healthy. Unfortunately, people all around the world have been facing cardiovascular diseases. Any technology that can help diagnose these diseases before much damage is done will prove as helpful in saving people's money and more importantly their lives. Data analytic techniques can be useful in predicting heart diseases. Predictive models can be made by finding previously unknown patterns and trends in databases and using the obtained information. Data analytic technique is a technology which can help to achieve diagnosis of heart disease. As an emerging field in science and technology, Data analysis can classify whether a person might be suffering from a heart disease or not.

Data Exploration: The first step is to obtain the data set containing the features of a person suffering from a heart disease and a person who is not suffering from the disease . The data set used in this experiment is taken from a website called https://archive.ics.uci.edu/ml/datasets/Heart+Disease. The programming language used to do the experiment is R. Thirteen attributes are used which are available in the data set. The information of the attributes is available on above website. The next step is to analyze the data. For this, the information of the data set is required. To gather the concise summary of the Data Frame, the summary() function is used on the data set.

```
> str(dataset)
'data.frame':   303 obs. of  14 variables:
 $ age     : num  63 67 67 37 41 56 62 57 63 53 ...
 $ sex     : num  1 1 1 1 0 1 0 0 1 1 ...
 $ cp      : num  1 4 4 3 2 2 4 4 4 4 ...
 $ trestbps: num  145 160 120 130 130 120 140 120 130 140 ...
 $ chol    : num  233 286 229 250 204 236 268 354 254 203 ...
 $ fbs     : num  1 0 0 0 0 0 0 0 0 1 ...
 $ restecg : num  2 2 2 0 2 0 2 0 2 2 ...
 $ thalach : num  150 108 129 187 172 178 160 163 147 155 ...
 $ exang   : num  0 1 1 0 0 0 0 1 0 1 ...
 $ oldpeak : num  2.3 1.5 2.6 3.5 1.4 0.8 3.6 0.6 1.4 3.1 ...
 $ slope   : num  3 2 2 3 1 1 3 1 2 3 ...
 $ ca      : num  0 3 2 0 0 0 2 0 1 0 ...
 $ thal    : num  6 3 7 3 3 3 3 3 7 7 ...
 $ num     : num  0 1 1 0 0 0 1 0 1 1 ...
```

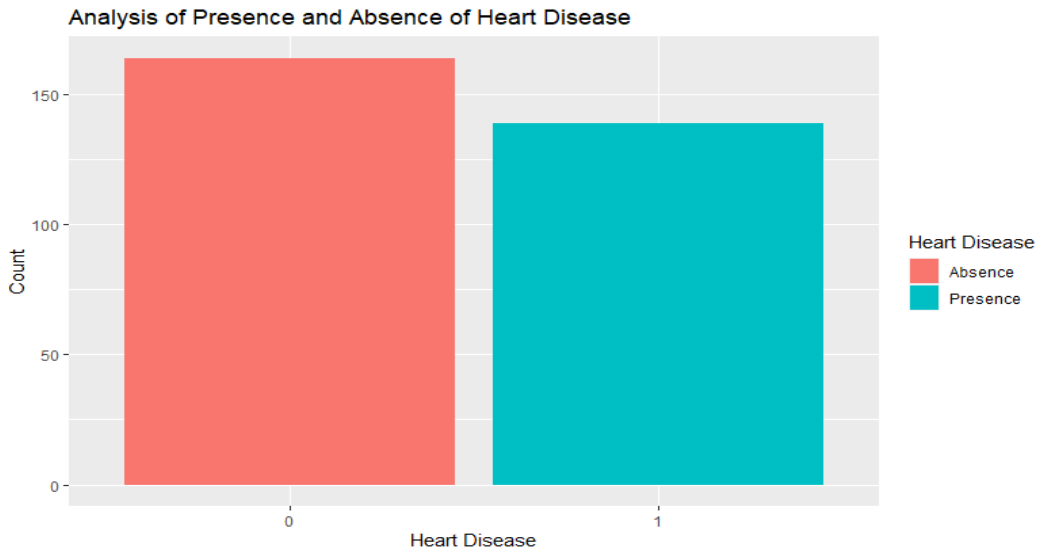
Fig: Concise structure of the dataset

The summary() function is used to retrieve some statistical information of the data set like mean of the values of the attributes used. An attribute named class is taken whose value is 1, if the patient is suffering from a heart disease, or 0, if the person is not suffering from any heart disease.

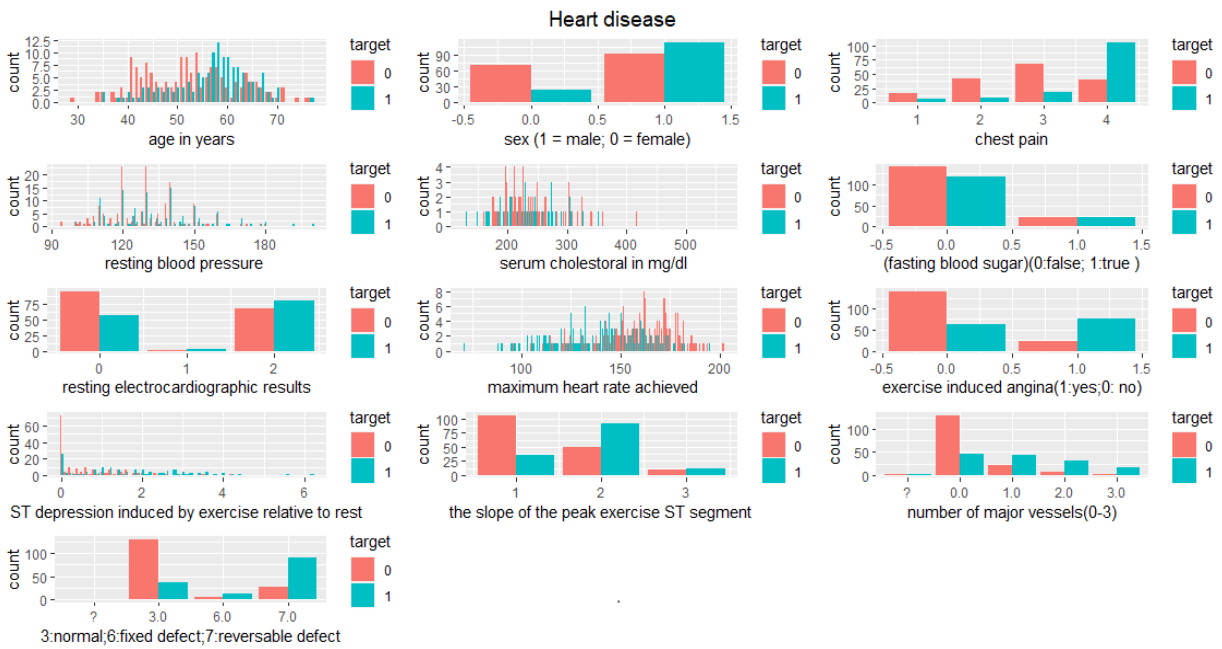
```
> summary(dataset)
      age             sex               cp            trestbps         chol            fbs        
 Min.   :29.00   Min.   :0.0000   Min.   :1.000   Min.   : 94.0   Min.   :126.0   Min.   :0.0000  
 1st Qu.:48.00   1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:120.0   1st Qu.:211.0   1st Qu.:0.0000  
 Median :56.00   Median :1.0000   Median :3.000   Median :130.0   Median :241.0   Median :0.0000  
 Mean   :54.44   Mean   :0.6799   Mean   :3.158   Mean   :131.7   Mean   :246.7   Mean   :0.1485  
 3rd Qu.:61.00   3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:140.0   3rd Qu.:275.0   3rd Qu.:0.0000  
 Max.   :77.00   Max.   :1.0000   Max.   :4.000   Max.   :200.0   Max.   :564.0   Max.   :1.0000  

    restecg          thalach         exang           oldpeak         slope             ca        
 Min.   :0.0000   Min.   : 71.0   Min.   :0.0000   Min.   :0.00   Min.   :1.000   Min.   :0.0000  
 1st Qu.:0.0000   1st Qu.:133.5   1st Qu.:0.0000   1st Qu.:0.00   1st Qu.:1.000   1st Qu.:0.0000  
 Median :1.0000   Median :153.0   Median :0.0000   Median :0.80   Median :2.000   Median :0.0000  
 Mean   :0.9901   Mean   :149.6   Mean   :0.3267   Mean   :1.04   Mean   :1.601   Mean   :0.6722  
 3rd Qu.:2.0000   3rd Qu.:166.0   3rd Qu.:1.0000   3rd Qu.:1.60   3rd Qu.:2.000   3rd Qu.:1.0000  
 Max.   :2.0000   Max.   :202.0   Max.   :1.0000   Max.   :6.20   Max.   :3.000   Max.   :3.0000  
                                                                                  NA's   :4       

      thal            num        
 Min.   :3.000   Min.   :0.0000  
 1st Qu.:3.000   1st Qu.:0.0000  
 Median :3.000   Median :0.0000  
 Mean   :4.734   Mean   :0.4587  
 3rd Qu.:7.000   3rd Qu.:1.0000  
 Max.   :7.000   Max.   :1.0000  
 NA's   :2
```

Fig: Statistical summary of the dataset

Data Visualization: Now the data set is to be checked that is it balanced or not. This is done using ggplot().



After looking at the plot, it can be concluded that the data is quite balanced. We can also use ggplot to compare different dependable attributes with independent attribute of the data set:



Here target 0 means absence of heart disease and target 1 means presence of heart disease.

Observations from the above plot:

Age and sex: Males age around 60 and above have high chance of heart disease.

cp {Chest pain}: People with cp 2, 3, 4 are more likely to have heart disease than people with cp 1.

restecg {resting EKG results}: People with a value of 2 (showing probable or definite left ventricular hypertrophy by Estes' criteria) are more likely to have heart disease.

exang {exercise-induced angina}: people with a value of 0 (No ==> angina induced by exercise) have more heart disease than people with a value of 1 (Yes ==> angina induced by exercise)

slope {the slope of the ST segment of peak exercise}: People with a slope value of 2 (flat slope: signs of an unhealthy heart) are more likely to have heart disease than people with a slope value of 1 (Upsloping: best heart rate with exercise) or 3 (downsloping: minimal change (typical healthy heart)).

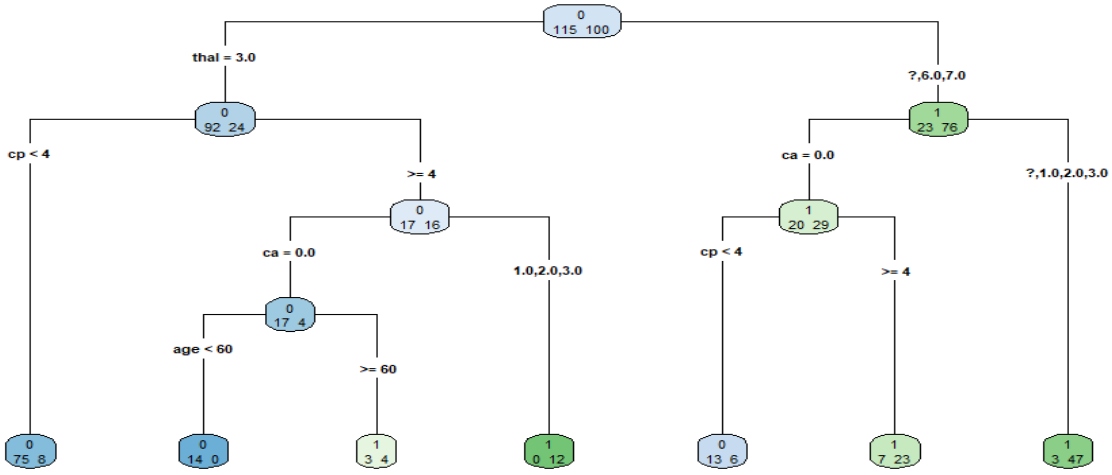
ca {number of major vessels (0-3) stained by fluoroscopy}: the more blood movement the better, so people with ca equal to 0 are more likely to have heart disease.

thal {thalium stress result}: People with a thal value of 7 are more likely to have heart disease.

Splitting the dataset: Randomly split the dataset into training and testing set. The testing set is essential to validate our results. For splitting the data, we will use the caTools Package. The package contains sample.split command to split the data with a split ratio of 0.75. This means we'll put 75% of the data in the training set, which we'll use to build the model, and 25% of the data in the testing. There are 215 training samples and 88 testing samples.

Algorithms and Techniques Used

A . Decision tree: Given a data of attributes together with its classes, a decision tree produces a sequence of rules that can be used to classify the data. rpart() returns a Decision tree created for the data. It allows to grow the whole tree using all the attributes present in the data. The tree obtained using rpart.plot is:



Checking the accuracy using a confusion matrix by comparing predictions to actual classifications.

```
> confusionMatrix(data = as.factor(heart_pred),
+                 reference = as.factor(test_data$num))
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 39  9
         1 10 30

               Accuracy : 0.7841
                 95% CI : (0.6835, 0.8647)
    No Information Rate : 0.5568
    P-Value [Acc > NIR] : 7.302e-06

                  Kappa : 0.5637

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.7959
            Specificity : 0.7692
         Pos Pred Value : 0.8125
         Neg Pred Value : 0.7500
             Prevalence : 0.5568
         Detection Rate : 0.4432
   Detection Prevalence : 0.5455
      Balanced Accuracy : 0.7826

       'Positive' Class : 0
```

In the above table, the accuracy is 0.78. So, the accuracy to predict heart disease using decision tree is 78%.

B . K-Nearest Neighbors (KNN): K-Nearest Neighbors algorithm is a non-parametric method used for regression& classification. The principle behind nearest neighbor method is to find a predefined number (k value) of training samples closest in distance to the new point & predict label from known label. Accuracy with different k values is shown below:

```
[1] "Accuracy = k =  1  =  71.5909090909091"
[1] "Accuracy = k =  2  =  73.8636363636364"
[1] "Accuracy = k =  3  =  77.2727272727273"
[1] "Accuracy = k =  4  =  76.1363636363636"
[1] "Accuracy = k =  5  =  78.4090909090909"
[1] "Accuracy = k =  6  =  77.2727272727273"
[1] "Accuracy = k =  7  =  76.1363636363636"
[1] "Accuracy = k =  8  =  78.4090909090909"
[1] "Accuracy = k =  9  =  75"
[1] "Accuracy = k =  10  =  76.1363636363636"
[1] "Accuracy = k =  11  =  77.2727272727273"
[1] "Accuracy = k =  12  =  78.4090909090909"
[1] "Accuracy = k =  13  =  79.5454545454545"
[1] "Accuracy = k =  14  =  78.4090909090909"
[1] "Accuracy = k =  15  =  77.2727272727273"
[1] "Accuracy = k =  16  =  78.4090909090909"
[1] "Accuracy = k =  17  =  81.8181818181818"
[1] "Accuracy = k =  18  =  78.4090909090909"
[1] "Accuracy = k =  19  =  81.8181818181818"
[1] "Accuracy = k =  20  =  82.9545454545455"
[1] "Accuracy = k =  21  =  82.9545454545455"
[1] "Accuracy = k =  22  =  82.9545454545455"
[1] "Accuracy = k =  23  =  82.9545454545455"
[1] "Accuracy = k =  24  =  81.8181818181818"
[1] "Accuracy = k =  25  =  81.8181818181818"
```

In the above table, k with 20,21,22,23 values have high accuracy. Now prediction is done with k =23 on test data and accuracy is checked with confusion matrix.

```
> confusionMatrix(data = as.factor(pred_knn),
+                 reference = as.factor(test_data$num ))
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 45 11
         1  4 28

               Accuracy : 0.8295
                 95% CI : (0.7345, 0.9013)
    No Information Rate : 0.5568
    P-Value [Acc > NIR] : 5.771e-08

                  Kappa : 0.6482

 Mcnemar's Test P-Value : 0.1213

            Sensitivity : 0.9184
            Specificity : 0.7179
         Pos Pred Value : 0.8036
         Neg Pred Value : 0.8750
             Prevalence : 0.5568
         Detection Rate : 0.5114
   Detection Prevalence : 0.6364
      Balanced Accuracy : 0.8182

       'Positive' Class : 0
```

In the above table, the accuracy is 0.8295 with less p value. So, the accuracy to predict heart disease using Knn algorithm is 83%.

C . Naïve Bayes: Naive Bayes is a Supervised Non-linear classification algorithm in R Programming. Naive Bayes algorithm based on Bayes' theorem with the assumption of independence between every pair of features. Naive Bayes classifiers work well in many real-world situations. The e1071 package is used for Naïve Bayes classification. Prediction is done on test dataset and for model evaluation confusionMatrix() is used. The result is shown below:

```
> confusionMatrix(data = as.factor(pred_bayes),
+                 reference = as.factor(test_data$num ))
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 45 12
         1  4 27

               Accuracy : 0.8182
                 95% CI : (0.7216, 0.8924)
    No Information Rate : 0.5568
    P-Value [Acc > NIR] : 2.152e-07

                  Kappa : 0.6237

 Mcnemar's Test P-Value : 0.08012

            Sensitivity : 0.9184
            Specificity : 0.6923
         Pos Pred Value : 0.7895
         Neg Pred Value : 0.8710
             Prevalence : 0.5568
         Detection Rate : 0.5114
   Detection Prevalence : 0.6477
      Balanced Accuracy : 0.8053

       'Positive' Class : 0
```

On studying the confusion matrix of the model, the accuracy is 0.8182. So, the accuracy to predict heart disease using Naïve Bayes is 82%.

D . Logistic Regression: Logistic Regression is a statistical and machine-learning technique classifying records of a dataset based on the values of the input fields. It predicts a dependent variable based on one or more set of independent variables to predict outcomes. Logistic regression belongs to a class of models called the Generalized Linear Models (GLM) which can be built using the glm() function. After applying glm() select attributes with less p-value which are significant to predict the heart disease. The summary table of the same is shown below:

```
> summary(logistic_reg)

Call:
glm(formula = num ~ sex + cp + trestbps + thalach + exang + slope +
    ca, family = "binomial", data = train_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7043  -0.5907  -0.1802   0.4368   2.3524

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -10.12414    3.02685  -3.345 0.000823 ***
sex           1.61806    0.46805   3.457 0.000546 ***
cp            0.93769    0.23282   4.028 5.64e-05 ***
trestbps      0.02321    0.01210   1.917 0.055191 .
thalach      -0.01885    0.01114  -1.693 0.090511 .
exang         1.10063    0.46910   2.346 0.018964 *
slope         1.06614    0.35139   3.034 0.002413 **
ca            1.38732    0.29772   4.660 3.17e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 297.01  on 214  degrees of freedom
Residual deviance: 157.23  on 207  degrees of freedom
AIC: 173.23

Number of Fisher Scoring iterations: 6
```

On studying the summary of the model, it is evident that both the coefficients are significant since their p-values are small and also the AIC and deviance values have dropped down when compared to earlier, which is a good thing. The final step is to evaluate the efficiency of the model by making predictions on test data and find the accuracy of the model using confusion matrix. The confusion matrix table is shown below:

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 43 10
         1  6 29

               Accuracy : 0.8182
                 95% CI : (0.7216, 0.8924)
    No Information Rate : 0.5568
    P-Value [Acc > NIR] : 2.152e-07

                  Kappa : 0.6277

 Mcnemar's Test P-Value : 0.4533

            Sensitivity : 0.8776
            Specificity : 0.7436
         Pos Pred Value : 0.8113
         Neg Pred Value : 0.8286
             Prevalence : 0.5568
         Detection Rate : 0.4886
   Detection Prevalence : 0.6023
      Balanced Accuracy : 0.8106

       'Positive' Class : 0
```

On studying the confusion matrix of the model, the accuracy is 0.8182. So, the accuracy to predict heart disease using Logistic regression is 82%.

Conclusion: In this project, it is clear that most data analytic techniques perform well in the prediction of heart diseases, the results showed a great accuracy standard for producing a better estimation result. Accuracy on different models is shown below:

DATA ANALYTIC TECHNIQUES	ACCURACY
Decision Tree	78%
KNN	83%
Naïve Bayes	82%
Logistic Regression	82%