2021

Heart Disease Prediction

DR TEHSEEN AHMED JILANI

Mushba Ali Siddiqui – P19101046

Maryum Wasim – P19101035

Iqra Aziz – P19101024

Maliha Sattar – P19101032

Table of Contents

1. Introduction:

In this project, studies are going to check out the impact of different features on human heart. Heart disease is the main cause of deaths for people. Mostly people thought that age is the main factor which causes heart diseases, but in the real world age is not the only factor which causes the heart disease, although there are various factors that causes heart diseases. These factors vary in males and females. Mostly we saw that males are more affected by heart diseases as compared to female. In this project we consider various factors such as blood pressure, heart rate, gender, chest pain, cholesterol, diabetes, smoking, obesity, stress, depression etc. in our project, we have taken the data of heart disease prediction from kaggle. In this data we consider the following features, blood pressure, chest pain, heart rate, gender, fat levels, angina, old level and also the report of ECG. Our dataset contains both numeric and alpha-numeric data. We made our project by using python. In our project we perform clustering, principal component analysis, correlation, and decision tree. On the basis of these techniques we predict results.

2. Literature review:

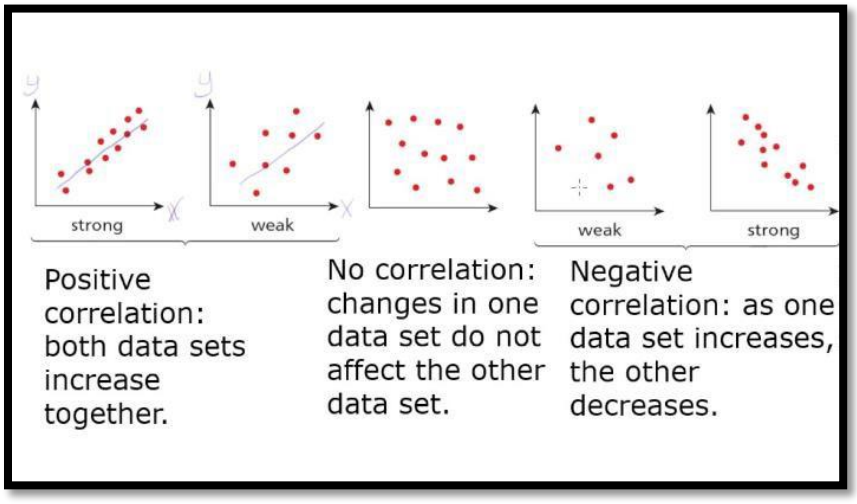
2.1. Correlation:

Correlation is a relationship between variables, by looking at correlation we can get idea how strong, what kind of relationship exists between the correlated variable. One feature of a correlation is direction. By looking positive or negative signs we can tell what is happening in the variables. In correlated data, variation in 1 data set is associated with the change in the values of other data set. If the dependent and independent variable move in same upward direction then they are positively correlated i.e. if one variable increase then the other variable would also be increase same as if they both move in opposite direction i.e. downward then the variables have negative correlation.

Why is correlation important?

Most of the time in dataset the features or the variables have relationship for example supply and demand, temperature and humidity, age and disease etc. by correlation analysis we can predict the degree of relationship in one figure.

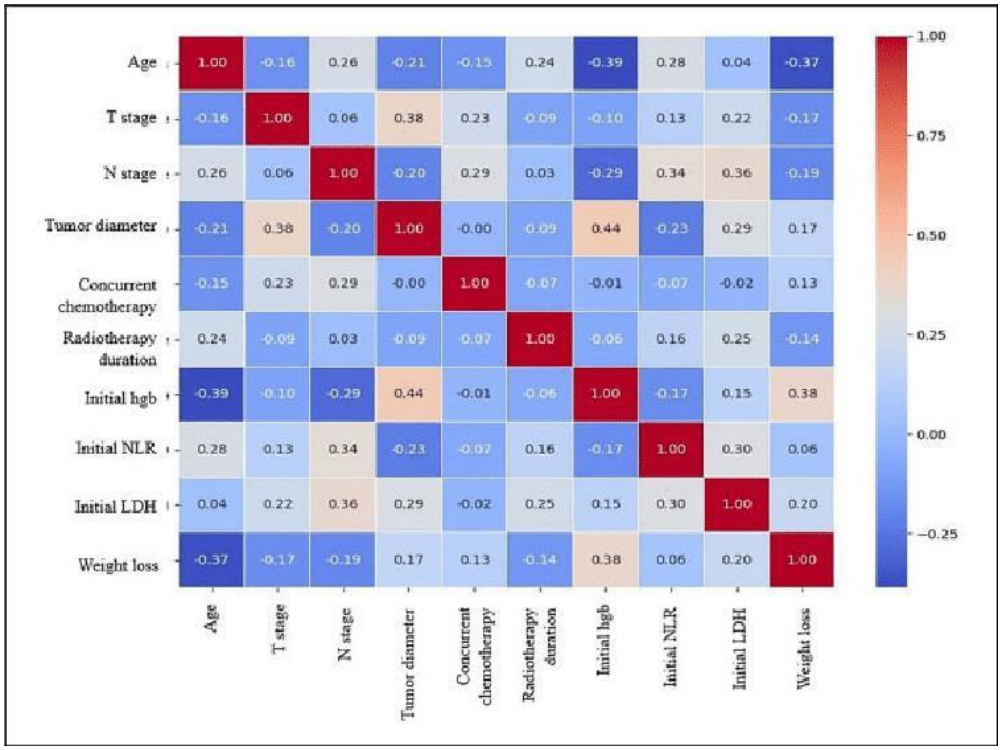
2.2. Types of correlation analysis:



2.3. Heatmap:

Heatmap is a graphical representation of correlation analysis. It shows two dimensional matrix correlation between features, by using colored cells the light color cells shows high correlation while the darker cells show week correlation.

1st dimension can be appeared as row while the 2nd can be appeared as column.



2.4. Normalization of data:

Normalization is a scaling or mapping technique, by using this technique we can normalize our dataset by removing variations. After normalization we can get new range from the existing one. Manage a large variation dataset is not a piece of cake so normalization techniques are used to make the values closer. Following are the techniques of normalization.

2.4.1 Min max normalization:

Min max normalization is the linear transformation of the original dataset. In this technique minimum and maximum value are drawn and each value is replace according to following formula.

$$v' = \frac{v - min_F}{max_F - min_F}(new\_max_F - new\_min_F) + new\_min_F$$

2.4.2. Decimal scaling:

In this technique normalization is done by moving the decimal point of values of the dataset. Tonormalize the data each value is divided by max absolute value in data.

$$X_{new} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$

2.4.3. Z-score normalization:

In this technique normalization is based on mean and standard deviation od dataset.

$$z = \frac{x_i - \mu}{\sigma}$$

Why do we clean data in data analysis?

In data analysis data cleaning done to remove duplicate and obsolete data. It is also used for correcting inaccurate information.

2.5. Principal component analysis:

PCA is a statistical method that is used to summarize the informative content in large dataset on the bases of smaller set of summary catalogues. Visualizing and analyzing large data is difficult, smaller sets can easily be visualize and analyzed. If we have high dimensions it means we have lot information about our data. Large number of variables requirement could increase computational time and some models don't work with large number of data in data set. Including large number of variables in our model increases the complexity as well. To reduce the data in the model we use reduction techniques. PCA is a dimension reduction technique we take linear combinations of our existing variables to create new variables such that the new variables perform better to explain the variation in data. PCA can only be performed on numeric values. The axis we use are the eigen vectors of the covariance measures of the data. To determine the order of the axis we ordered the eigen vectors by the magnitude of their corresponding values.

Eigen value:

Eigen values are the variance of principal components, these values are also known as characteristics values or latent root. These values determine the number of principal components. For visualizing the size of eigenvalues we use scree plot.

Scree plot:

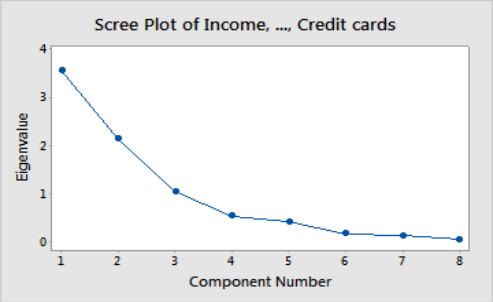
It shows the number of principal components contrasted with its corresponding eigenvalue. This plot ordered the eigenvalues in ascending order. Variance of the principal components and the eigenvalues of covariance matrix are equal.

Proportion:

It is used to determine which principal component explain more variability in data. Proportion is the amount of the variability in the data that each principal component can define.

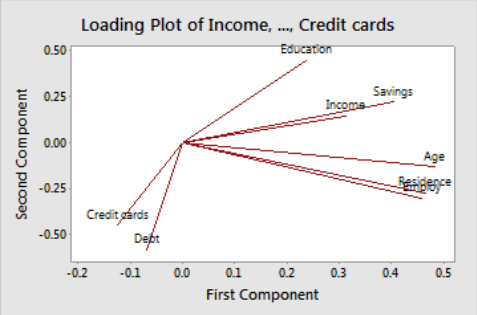
Score plot:

Score plot graph the scores of 2nd pc as opposed to the score of the 1st pc.



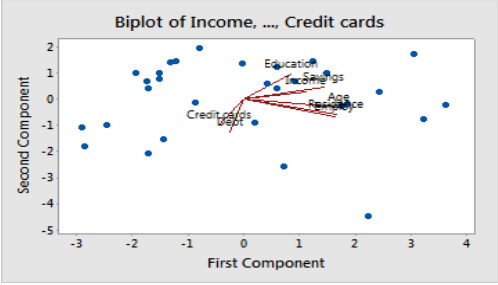
Loading plot:

Loading graph plot the coefficient of every variable of the 1st component against the coefficient ofthe 2nd component. It is used to identify which variables have greatest effect on each component. The range is -1 to +1. Loading close to the range shows the strong influence in component.



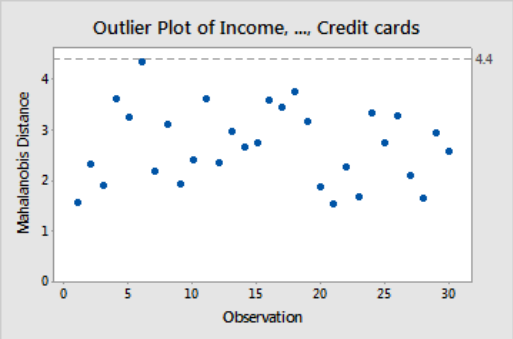
Biplot:

This plot joins the score plot and loading plot.



Outlier plot:

This graph shows the Mahalanob is distance for every observation and a reference line to identify outliers. Outlier plot is used to plot outliers.

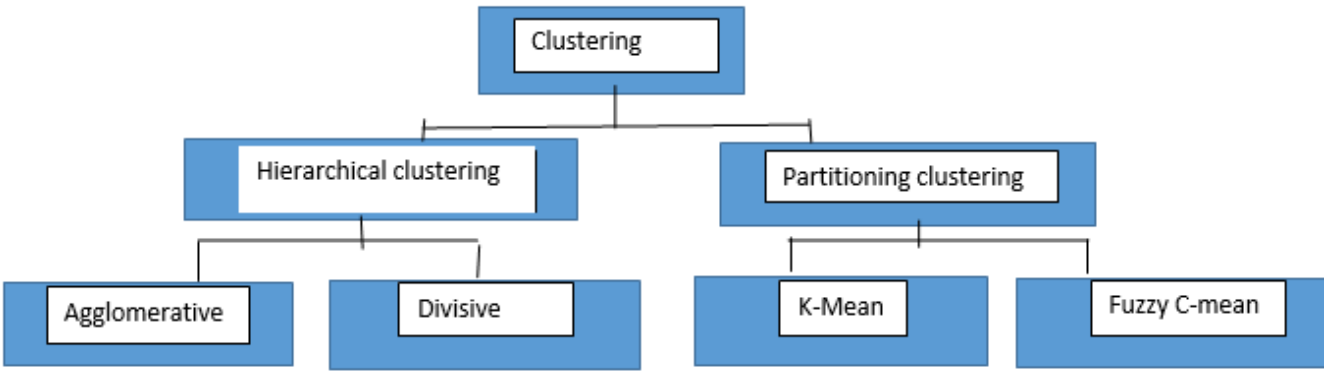


2.6. Clustering:

Clustering is defined as grouping the elements that contains the similar properties.

2.6.1 Types of clustering:

Different types of clustering are:



1. Hierarchical clustering: It is tree like structure.

2. Agglomerative clustering:

Agglomerative clustering is a bottom up approach. Dendogram is the representation of agglomerative clustering.

3. Divisive clustering:

It is the top to down approach, in which large clusters are divided into further sub clusters.

4. K-means clustering:

K-Means clustering is an unsupervised and hard clustering approach in data mining. In this technique data is divided into clusters according to the number of clusters define by the user. In k- Mean clustering each observation belongs to the clusters with nearest means. Following are the steps for finding k mean clusters.

Step: 01

Select the number of clusters 'K' you want to identify in your dataset.

Step: 02

Randomly select distinct data point.

Step: 03

Allocate two centroid randomly.

Step: 04

Determine the distance between each randomly assigned centroid.

Step: 05

Determine the actual centroid of that clusters.

The process of calculating distance and repositioning continues until the final cluster is obtained.

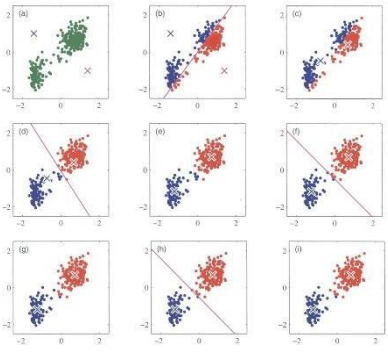
Distances that can be supported by K-Mean are;

Euclidean distance. $d(p,q) = \sqrt{(p_1 - q_1)2 + (p_2 - q_2)2}$

Manhattan distance. $d = \sum_{i=1}^{n} |x_i - y_i|$

A squared Euclidian measure. $d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$

Cosine distance measure. $\cos\theta = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\| \cdot \|\vec{b}\|}$



2.7.Decision tree:

Decision tree is one of the data mining technique which is used to create classification models in the form of tree like structure. It is a supervise learning technique which means that the result is already known. It is useful for both numerical and categorical data. It is used to solve regression and classification problem.

The difficult part is the identification of the attributes for the root node in each level which isknown as attribute selection. Following are the attribute selection measures.

1-Information gain

2-Gini index

1-Information gain:

When a node is used in a decision tree to partition the training instance into smaller subset theentropy changes.

- Entropy :
  Entropy is the uncertainty measure of the variable. the higher the entropy the more the informative content.

2-Gini index:

It is a metrics to measure how often a randomly chosen element would be incorrectly identified.

2.8.Naïve bayes.

Naïve Bayes algorithm is a popular machine learning algorithm. It is a probabilistic classifier, based on probability model. Naïve Bayes algorithm is based on Bayes theorem. It is an efficient algorithm for classification problem. This algorithm is suitable for real time prediction. This algorithm can be built using Gaussian, multinomial and Bernoulli distribution. This algorithm is scalable and can easily be implemented on large data set. It helps to calculate the posterior probability P(c/x) using the prior probability of class P(c), the prior probability of predictor P(x), and the probability of predictor given class, also called as likelihood P(x/c).

The posterior probability can be calculated by using formula:

$$P(c/x) = (P(x/c) * P(c)) / P(x)$$

2.8.1. Working of Naive Bayes algorithm.

Step 01: Create frequency table using dataset.

Step 02: Create a likelihood table by calculating the probabilities on the basis of attributes.

Step 03: Then evaluate the posterior probability using the naïve bayes equation for each class.

Areas where naïve Bayes algorithm is used,

1-Real time prediction.

2-Multi class prediction.

3-Recommendation system.

4-Text classification.

3. Data analysis:

First of all we have cleaned the data by removing alpha numeric categories and after cleaning thedata we plot the cleaned data by using pair plot command. To make our data representation betterwe have made histogram on the cleaned data which also shows the overlapping of data, which is useful for the data analysis. Then we found the correlation of all attributes which shows how much variables are correlated with each other. Correlation shows how much the particular factor affectedthe other factor.

3.1.Cleaned data:

```
]: cleaned_df.sample(5)
```

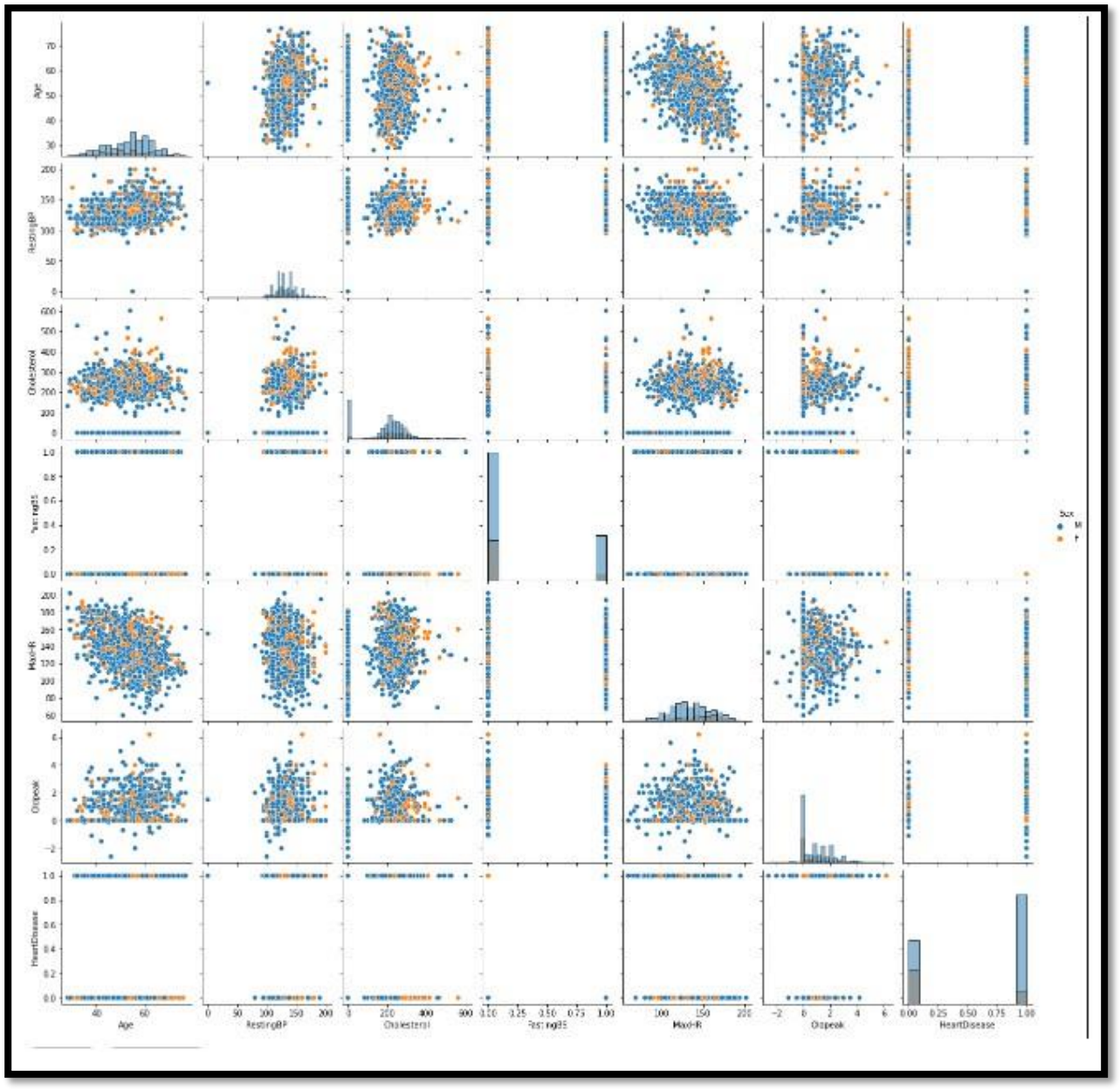
	Age	RestingBP	Cholesterol	FastingBS	MaxHR	Oldpeak	HeartDisease
541	76	104	113	0	120	3.5	1
522	50	144	349	0	120	1.0	1
563	55	135	204	1	126	1.1	1
3	48	138	214	0	108	1.5	1
240	55	145	326	0	155	0.0	0

Interpretation:

Cleaned data means, removal of alpha numeric data from the given dataset. Which helpsin plotting of data in a good way.

3.2. Pair Plot using histogram:

In the below plot we have used pairplot command to represent our data. In diagonal we make histogram which gives us better understanding of data and show overlapping of data. In the below plot we compared all the features with each other, which shows how one feature affect the other. As in the dataset 'Gender' is an independent feature, on behalf of this we made this plot.

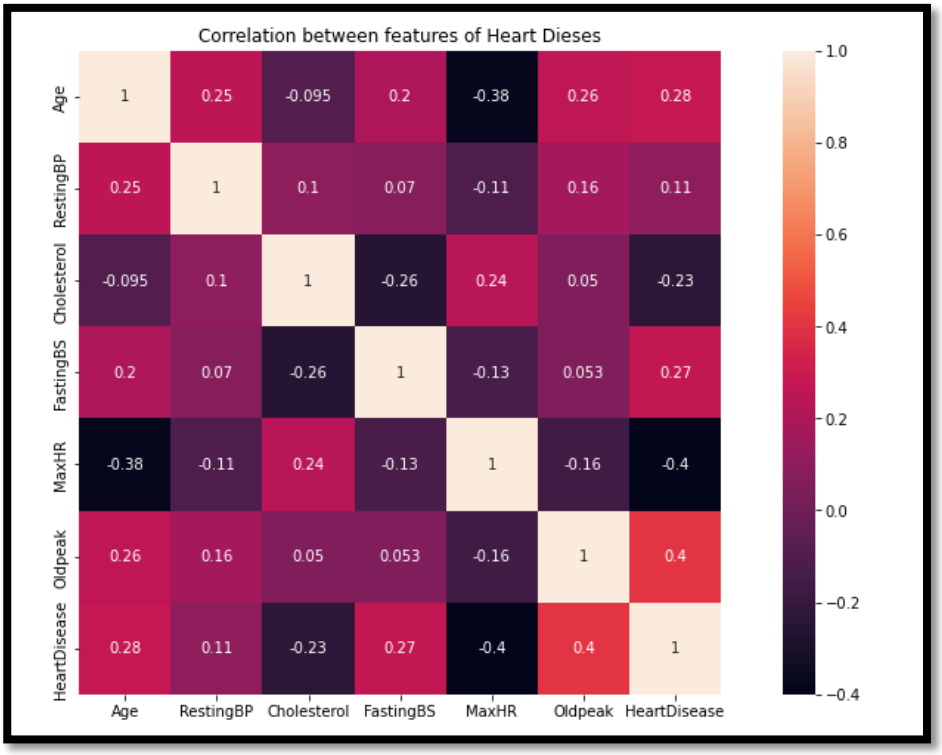


Interpretation:

From the above plot we can conclude that, as men's age increased, the cholesterol level in few ofthem becomes high than women. At young age males heart rate is more affected as compare to women. Men's are more affected by heart diseases.

3.3 Heatmap:

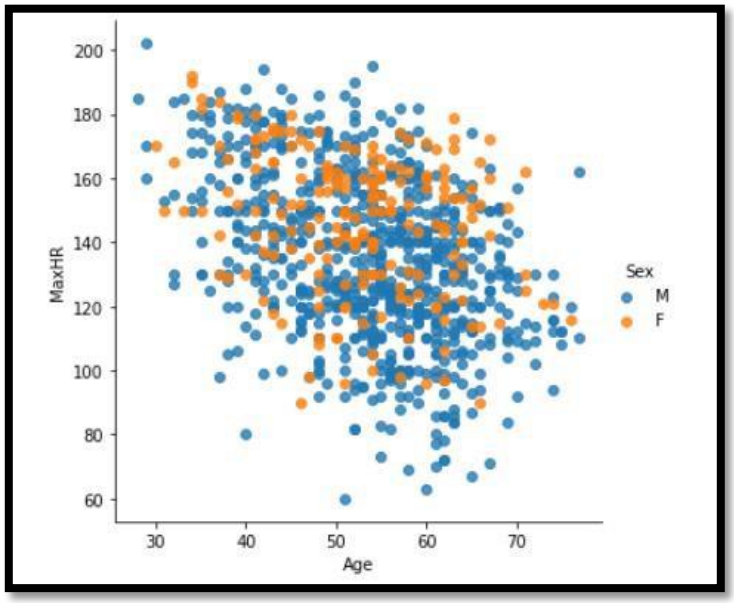
Heatmap is made after finding correlation of the cleaned dataset. It shows that how much our datais correlated with each other. Correlation gives us better understanding of data.



Interpretation:

From the above heatmap we can conclude that, the light color boxes shows that how muchthe particular features are correlated with each other. Older people have high chances of heart diseases and by increasing age there are many chances that one can have a heart disease. Dark color boxes shows that particular features are not much correlated with eachother. As age does not affect the heart beat rate, people can have heart beat issue at any age. Maximum heart rate does not lead to heart disease.

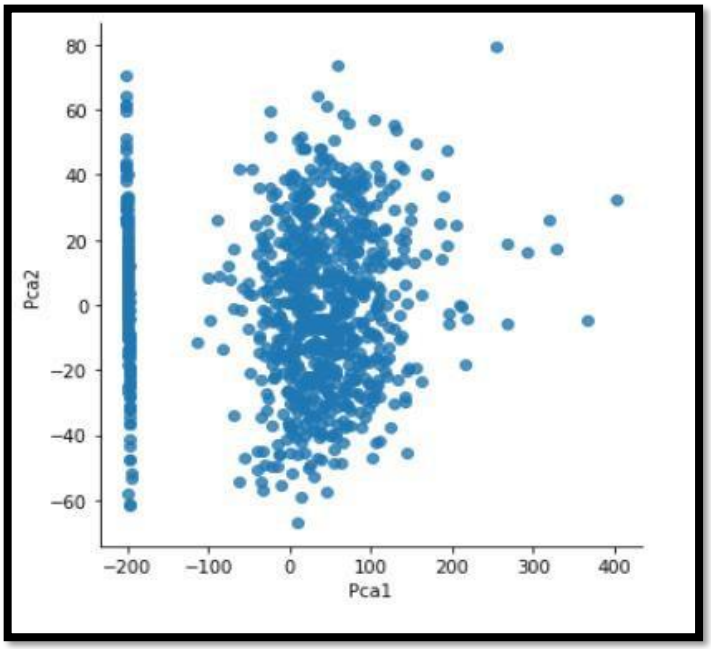
3.4 Lmplot:



Interpretation:

The above plot is made for plotting heart rate ratio by age using gender as an independent variable. The above plot clearly shows that men's heart rate is more affect at the young age as compared to women's. Heart rate affects male the most.

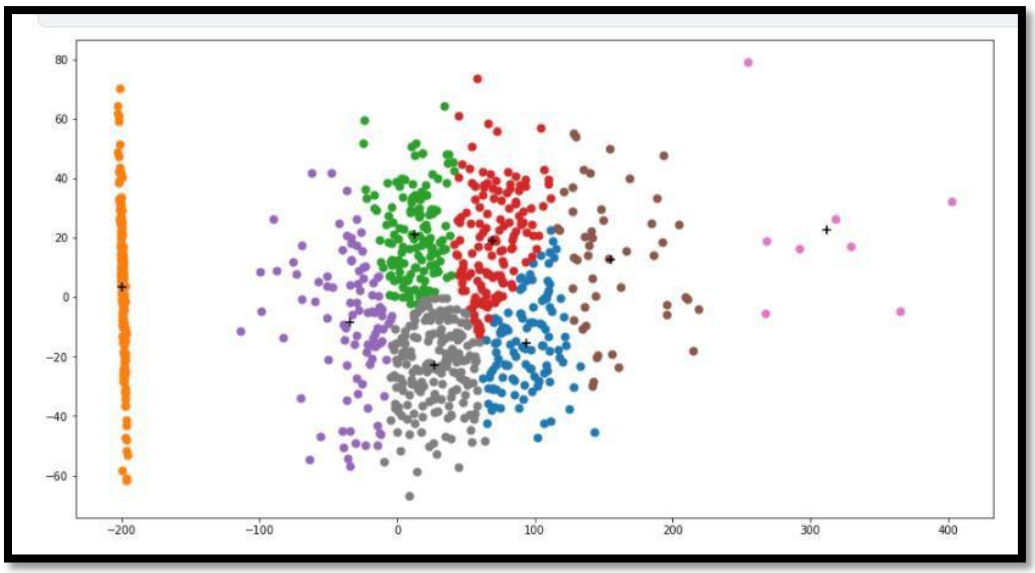
3.3 PCA plot:



Interpretation:

PCA plots are very useful to work on the reduce dataset. PCA used to reduce dimension of the dataset. Above PCA plot showed that most of the variables are heavily correlated to each other, but also it showed that some of the data is not.

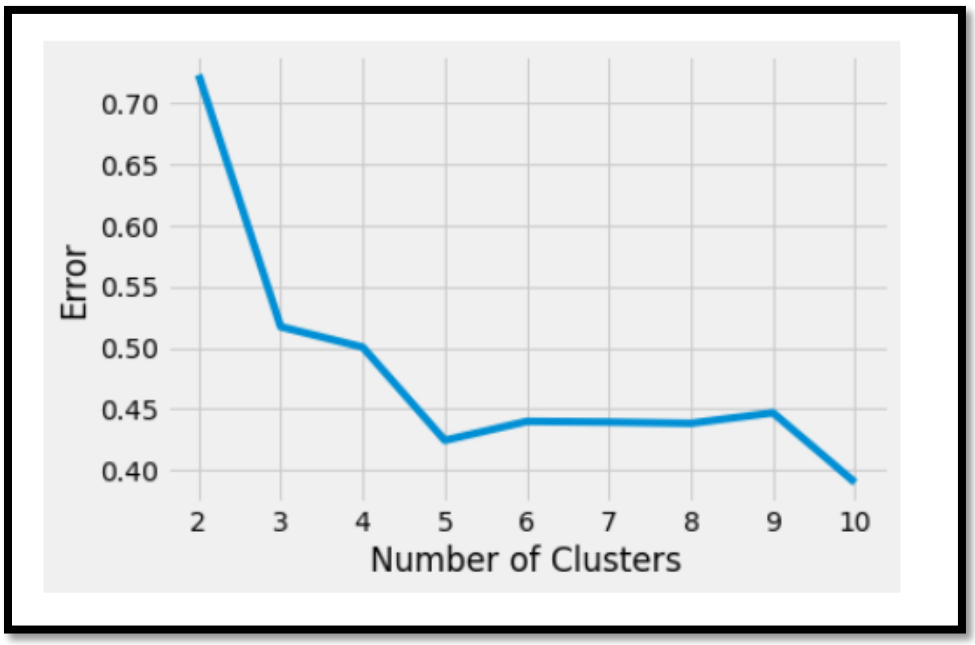
3.6 Clusters using K-means:



Interpretation:

By using k-means clustering, we made clusters of our datasets and also showed their centroids. It also shows that how much data is similar to each other and its centroid shows that how far or close each point in the cluster from centroid.

3.7 Silhouette Score:



Interpretation:

As silhouette score is used to measure how close the point lies within the cluster or in its neighbor cluster. Also it show the distance between the clusters by taking average or means of the values.

3.8 Decision Tree:

3.8.1 Transformation of data:

We have transformed our dataset into completely in numeric data by using enable loader command, means 0,1 and 2 values are assigned to them, like in old peak feature 1 is assigned to 'Flat' and 2 is assign to ' Up'

.

	Age	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
0	12	1	41	147	0	1	98	0	10	2	0
1	21	2	55	40	0	1	82	0	20	1	1
2	9	1	31	141	0	2	25	0	10	2	0
3	20	0	39	72	0	1	34	1	25	1	1
4	26	2	49	53	0	1	48	0	10	2	0
...	...	...	...	...	...	...	...	...	...	...	...
913	17	3	14	122	0	1	58	0	22	1	1
914	40	0	45	51	1	1	67	0	42	1	1
915	29	0	31	9	0	1	41	1	22	1	1
916	29	1	31	94	0	0	100	0	10	1	1
917	10	2	39	35	0	1	99	0	10	2	0

Accuracy:

As decision tree also shows how much our dataset is accurate. By using Gini coefficient, we have found our data is 70.65% accurate. Gini coefficient shows the inequality in a dataset. Low values shows equality and high values shows inequality or deviation within the data. As 70% showed that our dataset have inequality or deviations.

3.8.2 Naïve bayes.

Accuracy.

By using naïve bayes we have calculated the accuracy of the dataset. It shows that our data is 90 % accurate. Naïve bayes use probabilistic approach to show the accuracy of the dataset.

4. **Result:**

From the above observations we can conclude that, older people have high chances of heart diseases and by increasing age there are many chances that one can have a heart disease. As the dataset showed that features are not correlated with each other. As age does not affect the heart beat rate, people can have heart beat issue at any age. Maximum heart rate does not lead to heart disease.

The results is made for plotting heart rate ratio by age using gender as an independent variable. The results clearly shows that men's heart rate is more affected at the young age as compared towomen's. Heart rate affects male the most. Also from the results we can conclude that, as men's age increased, the cholesterol level in few of them becomes high than women. Men's are more affected by heartdiseases.

By using naïve bayes we have calculated the accuracy of the dataset. It shows that our data is 90 % accurate. Naïve bayes use probabilistic approach to show the accuracy of the dataset.

Reference:

1-Heart Failure Prediction Dataset | Kaggle