Final Project Report Introduction to Data Analytics

Project Title:
Prediction/Analysis of Heart Attack Chances

Prepared by:
Patel Preet Rajeshkumar (N01511398)

ITE 5201- Winter 2022 Humber College

I. Problem Statement

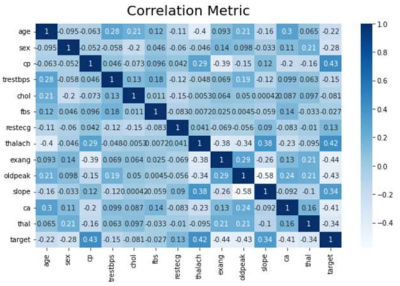
- It predicts whether the patient has less chances or more chances of having heart attack.

2. Dataset Description

- This dataset is used in predicting the patient has less chances of heart attack or more chances.
- It has 14 features such as age, sex, chest pain type, resting blood pressure, serum cholesterol, maximum heart rate achieved etc.
- The Sex field is converted into binary for training the model in which 0 represents male and 1 represents female.
- Chest pain type field is in integer which has values between 0 to 4.
- Target field is also in binary in which 0 shows the less chances of having heart attack and 1 shows the more chances of having heart attack.
- In the dataset, there are dependent variables such as age, sex, chest pain type, resting blood pressure, serum cholesterol, maximum heart rate achieved and the slope which I have used for training the model for predicting the heart attack chances.

3. Dataset Analysis and Observations:

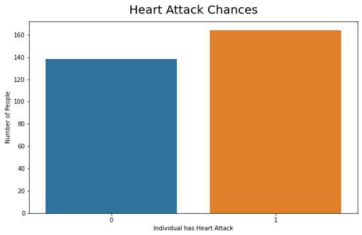
- I have used countplot for univariate and scatterplot for bivariate and Heatmap for finding Correlation Coefficient Rank.



This heatmap represents the information about the association between all features. It provides co-relation index (R).

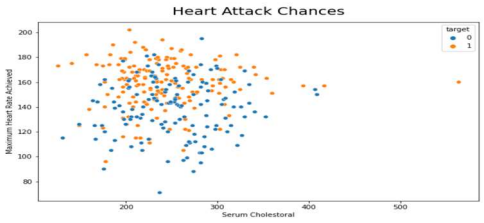
- R=1, Strong Positive relationship
- R=0, Not Linearly correlated
- R=-1, Strong negative relationship

From the heatmap, it can be shown that it is highly dependant on the chest pain, maximum heart rate achieved and slope.



This countplot shows the individuals whether they have less chances of heart attack or more chances of heart attack.

- 0, less chance of heart attack
- 1, more chance of heart attack



- From the Scatter Plot, it can be seen that if the serum cholesterol is in between 200 to 340, there are higher chances of getting the maximum heart rate.

4. Proposed Analytical 'Prediction Model

- I have spilt the data in two parts in which 80% of data is used for training and 20% data for testing.

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X,y,test_size=0.20,random_state=101)
print(X_train.shape, X_test.shape, y_train.shape, y_test.shape)
print(type(X_train),type(y_train))
```

```
(241, 7) (61, 7) (241,) (61,)
<class 'pandas.core.frame.DataFrame'> <class 'pandas.core.series.Series'>
```

- I have used 2 models. First one is Logistic Regression Model and k-NN model.
- For k-NN, I have used Euclidean metric.

- For Logistic Regression, after doing the training and predictions, it shows the output as below:

	Actual	Predicted
162	1	1
8	1	1
89	1	0
154	1	1
201	0	0

- For k-NN Model, after doing the training and predictions, it shows the output as below:

	Actual	Predicted
202	0	0
48	1	1
254	0	0
194	0	1
153	1	0

5. Results and Discussions

- Logistic Regression Model:

	precision	recall	f1-score	support
0	0.91	0.70	0.79	30
1	0.76	0.94	0.84	31
accuracy			0.82	61
macro avg	0.84	0.82	0.82	61
weighted avg	0.84	0.82	0.82	61

```
logistic_reg.score(X_test, y_test)
```

```
0.819672131147541
```

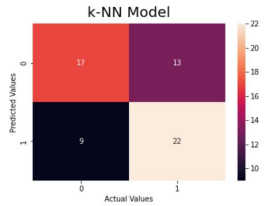
The Confusion metric displays the differentiation between actual values and predicted values.

True Positive: It means actual value and predicted values are positive. In this case, in 21 cases, the patient has low chances and model also predicts that has low chances of heart attack.

False Negative: It means actual value and predicted values are negative. In this case, in 29 cases, the patient has high chances and model also predicts that has high chances of heart attack.

From the classification report, it can be seen that it as accuracy of 82% and it has 0.8196 score.

- K-NN Model:



	precision	recall	f1-score	support
0	0.65	0.57	0.61	30
1	0.63	0.71	0.67	31
accuracy			0.64	61
macro avg	0.64	0.64	0.64	61
weighted avg	0.64	0.64	0.64	61

```
knn.score(X_test, y_test)
```

```
0.639344262295082
```

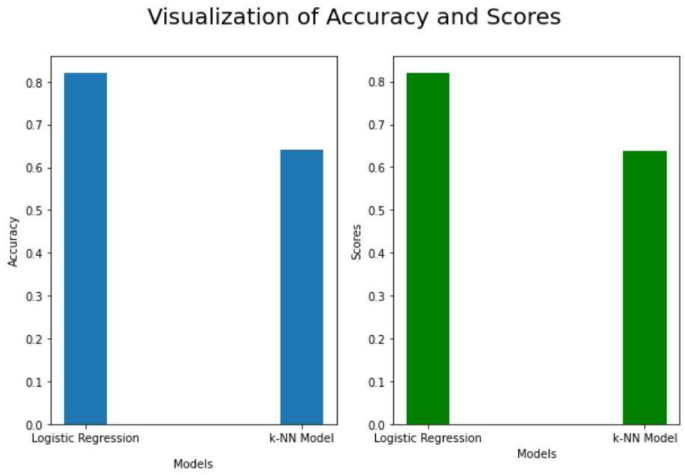
The Confusion metric displays the differentiation between actual values and predicted values.

True Positive: It means actual value and predicted values are positive. In this case, in 17 cases, the patient has low chances and model also predicts that has low chances of heart attack.

False Negative: It means actual value and predicted values are negative. In this case, in 22 cases, the patient has high chances and model also predicts that has high chances of heart attack.

From the classification report, it can be seen that it as accuracy of 64% and it has 0.6393 score.

6. Conclusion:



- From the graph, we can conclude that that **Logistic Regression Model** has higher accuracy as well as score than the **k-NN model**.
- Therefore, for my dataset, the **Logistic Regression Model** is more accurate for the evaluation.