Heart Attack Analysis and Prediction

1st Shivam Thakker
Under Prof. Mehul Raval
Ahmedabad University
Ahmedabad, India
shivam.t1@ahduni.edu.in

2nd Devarsh Seth
Under Prof. Mehul Raval
Ahmedabad University
name of organization (of Aff.)
Ahmedabad, India
devarsh.s2@ahduni.edu.in

3rd Pranav Gandhi
Under Prof. Mehul Raval
Ahmedabad University
name of organization (of Aff.)
Ahmedabad, India
pranav.g@ahduni.edu.in

4th Meet Jhaveri
Under Prof. Mehul Raval
Ahmedabad University
name of organization (of Aff.)
Ahmedabad, India
meet.j@ahduni.edu.in

***Abstract*—In this project we are trying to do Heart Attack Analysis and Prediction using dataset which contains features like age, sex, chest pain, Cholestrol, Blood Pressure, Blood Sugar using models such as Logistic Regression, Support Vector Machine, K-Nearest Neighbour and Linear Discriminative Analysis (LDA) for predicting whether there are less chance of heart attack or more chance of heart attack**

***Index Terms*—Heart Attack Prediction, Logistic Regression, KNN, Linear Discriminative Analysis(LDA)**

I. Introduction

Heart Attack prediction has been a spotlight since past few years. People of all the ages starting from 40 yrs are suffering from heart attack. This is a silent killer disease as it is not very contagious but the number of deaths due to it has significantly increased in past few years. Hence to support the medical team with an expertise of Computer science field is a key goal of this research.

We are trying to do heart attack analysis and prediction using Logistic Regression and trying to show it's comparison with other models such as KNN, SVM and LDA. Our data consists of 303 rows and 14 Columns. Out of 14 columns 13 are for features Age, sex, exng, Caa, Cp, trtbps , chol , fbs, $rest\_ecg$, thalachh and 14th column is for Output. Further we have taken steps to remove missing values and outliers from the data. Standardization process is also carried out to remove chance of having high bias in our model. Also we have checked the Correlation of each features with each other and removed the features which had a very low co-relation with each other. We have used this models on our dataset and found out accuracy of each of the Model.

A. Abbreviations and Acronyms

Age: Age of the patient Sex: Sex of the patient exng: Exercise induced angina (1 = yes, 0 = no) caa: Number of major vessels (0-3) cp: Chest pain type

1) Typical angina
2) Atypical angina
3) Non-anginal pain
4) Asymptomatic
5) trtbps: Resting blood pressure
6) chol: Cholesterol in mg/dl
7) fbs: Fasting blood sugar ¿ 120 mg/dl (1 or 0)
8) $rest\_ecg$: Resting electrocardiographic results
9) 0: Normal
10) 1: Having ST-T Wave Abnormality
11) 2: Showing probable or definite left ventricular hypertrophy by Estes' criteria thalachh: Maximum heart rate achieved target: 0 = less chance of heart attack 1 = More chance of heart attack.

II. Procedure

1) Read and Analyse the data

   This step was for better understanding the data. This step showed us that our data has 303 rows and 13 columns.

2) Missing value Analysis

   In this step we used isnull() and then used sum() to find out how many of missing values are there for each features. It turned out that there were no missing values.

3) Unique Value Analysis

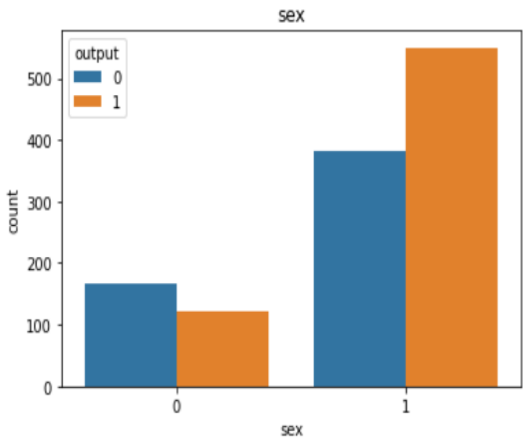
   This step was carried out to check how many types of values does each feature have. It turns out as shown in below image.

Attributes	No. of occurences
Age	50
sex	2
cp	4
trtbps	67
chol	222
fbs	2
restecg	3
thalachh	119
exng	2
oldpeak	53
slp	3
output	2

As we can see in the above image, we found out that sex, cp, fbs, restecg, exng, slp, caa, thall and output are the categorical Features.

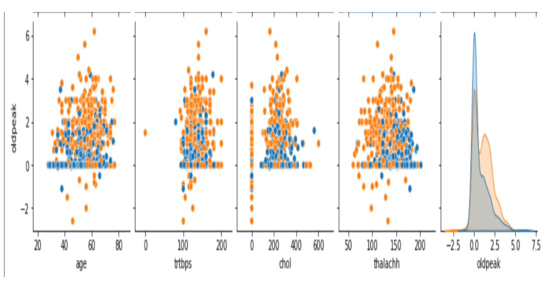
4) Categorical Value Analysis

This step is carried out for better understanding value of which class is there most of the times for a particular Categorical feature. We plotted graphs using sns.countplot(). We plotted a graph of sex vs. count , cp vs. count , fbs vs. count, restecg vs. count , exng vs. count , slp vs. count , caa vs. count, thall vs. count, and output vs. count. Below is one of the output graphs.



5) Numeric Value Analysis

We did Numeric analysis with plotting graphs of numeric features which are age, trtbps, chol, thalachh, oldpeak. Below are the results for it.



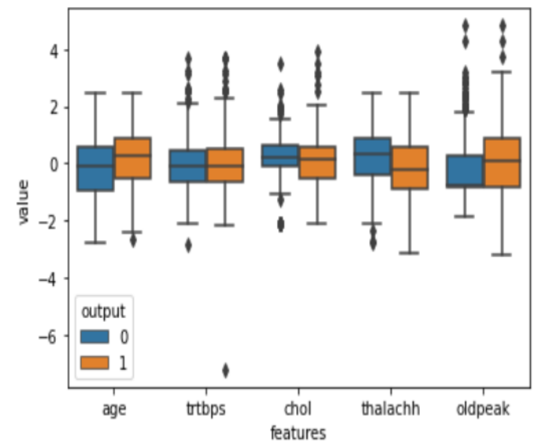
In the above image, we can see that there is a lot of overlap of data points which indicates that we should further carry out correlation Analysis also to better understand the data and remove some features if necessary.

6) Standardization

We carried out Standardization using StandardScalar() and .fit transform()

7) Box plot analysis

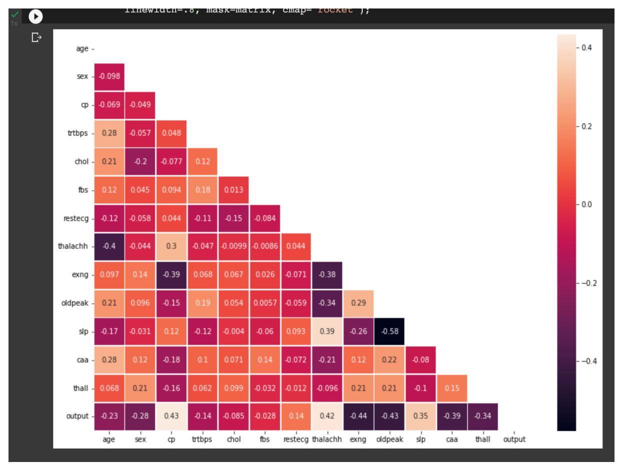
We will plot the figure using sns.boxplot() with giving it x = "features" , y = "value" and hue = "output". Below is the output graph.



As we can see in the above graph, the height of the box is showing interquartile range and line in the middle of the box is median. And the top and bottom line is whiskers. The points which are lying outside the whiskers are known as Outliers.

8) Coorelation Analysis

This step is carried out to find correlation between features. Using sns.heatmap() and by passing corr() we got the correlation graph as shown below.



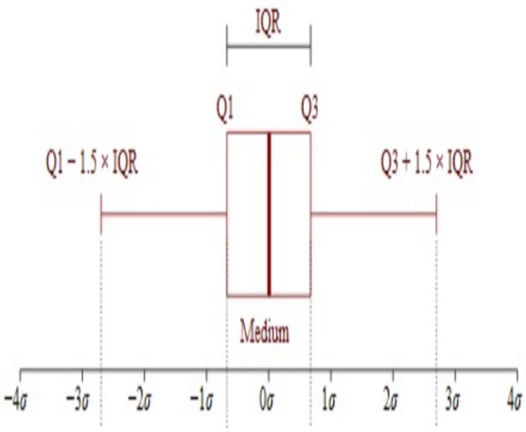
9) Dropping Uncorrelated Features

Using df.drop(), we are dropping features which have correlatedness less than 0.15. Below is the output of categorical variables which have correlatedness less than 0.15 will have output as True. As we can see in the above, restecg and fbs will be dropped using df.drop(). Similarly, we also do the same for Numeric Features.

10) Outlier Detection

- Here we are going to first find q1 and q3 using q1=np.percentile(currentItem, 25) and q3 = np.percentile(currentItem, 75)
- So we get interquartile range iqr = q3 - q1
- Then we will find the upperlimit and lowerlimit using formulas:- upperLimit = q3 + 2.5 * iqr and lowerLimit = q1 - 2.5 * iqr
- Now the points which are less than lowerLimit and more than upperLimit are outliers which we will

remove.



11) Modeling
   - Splitting into test and train data
     Using train-test-split function and by passing testsize = 0.2, we are splitting data set into 80 percent as train data and 20 percent of it as test data.
   - Logistic Regression
     Here we are using LogisticRegression() and by passing xTrain and yTrain data to it we are training our model. We will then find y predicted probability using $predict\_proba(xTest)$. Accuracy here is 90 percent
   - KNN
     Using KNeighborsClassifier() we also trained our KNN model on same xTrain and yTrain and found out the accuracy. Accuracy in case of KNN is 83 percent.
   - LDA
     Using LDA() and passing number of components equal to 1, we also trained our LDA model on same xTrain and yTrain and found out the accuracy. Accuracy here is 87 percent.

   Below given is the table of accuracies

Model	Accuracy in percentage
Logistic Regression	90
KNN	83
LDA	87

III. Conclusion

Logistic regression works best in our case because our output label has only two classes. LDA is close to it, but it would work well than logistic regression if our output label would have more than two classes. Also KNN is not good comapared to LDA and Logistic Regression both.

References

1. Hidayet, Takci. (2018). Improvement of Heart Attack Prediction by the Feature Selection Methods. Turk J Elec Eng Comp Sci, 26, 1-10
2. Sharma, H., Rizvi, M. A. (2017). Prediction of heart disease using machine learning algorithms: A survey. International Journal on Recent and Innovation Trends in Computing and Communication, 5(8), 99-104.
3. T.Obasi and M. Omair Shafiq, ”Towards comparing and using Machine Learning
4. S.S.Yadav, S. M. Jadhav, S. Nagrale and N. Patil, ”Application of Machine Learning for the Detection of Heart Disease,” 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), 2020, pp. 165-172, doi: 10.1109/ICIMIA48430.2020.9074954.