Predicting whether a person is suffering from a heart related disease or not

B AASHISH CHANDRA

October 2nd 2020.

1. Introduction

1.1. Background

Starting right from sports to fitness clubs, medical science to engineering, there are a lot of fields where in data science has stepped into. All the recent advancements in all the fields have been somewhere and somehow being related to data science. So applied data science has recently become a great tool to discover new insights. Data Science has been in limelight for around ten years or ever fewer than ten years, but the impact it has been having on most of the global giants has absolutely been remarkable. In so high huge scope, the field I chose to apply my data science skills, and build a project, for which this document is going to be a report is medical science. The model which has been built is basically a classification model. A whole lot of people keep suffering from a whole lot of diseases. So early prediction of the disease could therefore help the medical experts provide proper medical treatments and therefore could help them overcome the diseases. The model built completely focuses on this aspect.

1.2. Problem Statement

Having defined the field I chose, it is now time to define the problem statement. The problem statement I framed up thus is prediction of whether a person is suffering from any kind of heart related disease or not by a model. The model that has been built is largely dependent on the data that has been feed into it in the form of training of the model as well as it's testing. The data that has been used to perform the tasks related to training & testing is a multifaceted one and includes a lot of features including demographic data along with certain medical science related features. So in short this project basically comprised of developing a prediction model that would help people know whether they are suffering from any kind of heart related diseases.

1.3. Scope of Interest

As this project basically focuses on applying data science techniques into medical sciences, the scope of this data science project would mostly be limited to medical experts! Doctors would be thrilled to know that there is a model that would predict the well-being level of a person's heart with some mere numeric values.

2. Data Collection, Preparation & Cleaning

2.1. Data Collection (Process & Source)

The dataset I have used to train the model, test the model is basically from kaggle, which has a great collection of datasets for most of the data science projects, irrespective of the field we chose or opt for. The link for the dataset is https://www.kaggle.com/ronitf/heart-disease-uci. The dataset is relatively small but was good enough for model to be trained and tested. The shape of the dataset is something like (14, 303) meaning it comprised of fourteen features which included the independent variables i.e. attributes and a target variable, which was the final outcome of the project, and comprised of 303 data entries, i.e. records of 303 people who mostly had differing attributes all together. The further in-detailed description of the dataset is as follows,

- Age, describing the age of the person, having continuous values.
- Sex, describing gender of the person, having discrete values i.e. one for male and 0 for female.
- Chest Pain, describing the type of the chest pain having discrete values i.e. zero for typical angina, one for atypical angina, two for non-anginal pain, three for asymptomatic.
- Blood Pressure, describing the blood pressure of the person at that particular instance, having continuous values.
- Cholesterol, describing the cholesterol levels of the person at that particular instance, having continuous values.
- Blood Sugar, describing if the person is diabetic or not having discrete values i.e. one if he is diabetic, zero otherwise.
- Electro cardiac Measurement at Rest, having discrete values i.e. zero if normal, one if there's ST-T wave abnormality, two if showing probable or definite left ventricular hypertrophy by Estes' criteria.
- Maximum Heart Rate, describing the maximum heart rate of the person having continuous values.
- Exercise induced Angina, describing whether it has been caused due to exercise, having discrete values i.e. one if yes and zero otherwise.
- ST Depression Induced, describing the ST depression of the person, having continuous values.
- Slope of Peak Exercise, having discrete values i.e. zero if it is up slopping, one if it is flat and two if it is down slopping.
- Number of Major Vessels, having discrete values, zero if one, one if two, two if three, and three if all four.
- Thalassemia, having discrete values again, and one if normal, two if high, and three if too high.

The final feature is the Prediction that would be returned by the model trained, i.e. a one, if the person is suffering from a heart related disease, or a zero otherwise.

2.2. Data Preparation & Cleaning

The amount of work I had to do as a part of data preparation and cleaning has been very small. The dataset I used had already applied all the techniques related to data preparation. All the categorical value based columns excluding the target were converted to numerical value based columns performing the one hot encoding operation. Many unwanted features that wouldn't help in prediction were removed and the number of features including the target came down to fourteen which is generally the number of features a good dataset has. All the data redundancy issues were taken care of properly. I then had to rename all the features to proper and meaningful names and after importing the .csv file as and into a data frame it looked like the following,

	Age	Sex	Chest Pain Type	Blood Pressure	Cholestrol	Blood Sugar	Electrocardiac Measurement at Rest	Maximum Heart Rate	Exercise indused Angina	ST Depression Induced	Slope of Peak Exercise ST Segment	Number of Major Vessels	Thalassemia	Prediction
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

After basic data analysis was performed the above data frame was split to separate the variables that were used to predict, and the target variable which looked like the following,

	Age	Sex	Chest Pain Type	Blood Pressure	Cholestrol	Blood Sugar	Electrocardiac Measurement at Rest	Maximum Heart Rate	Exercise indused Angina	ST Depression Induced	Slope of Peak Exercise ST Segment	Number of Major Vessels	Thalassemia
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2

	Prediction
0	1
1	1

3. Exploratory Data Analysis

There are around five continuous value based columns in the data set and therefore the boxplots were used to show to relationship between each of them and the target variable i.e. prediction.

3.1. Boxplots

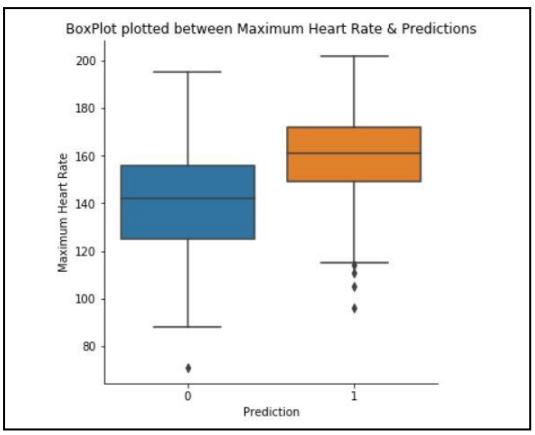


Figure 3.1.1

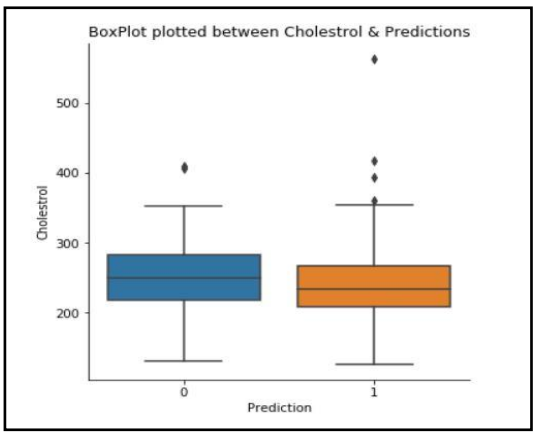


Figure 3.1.2

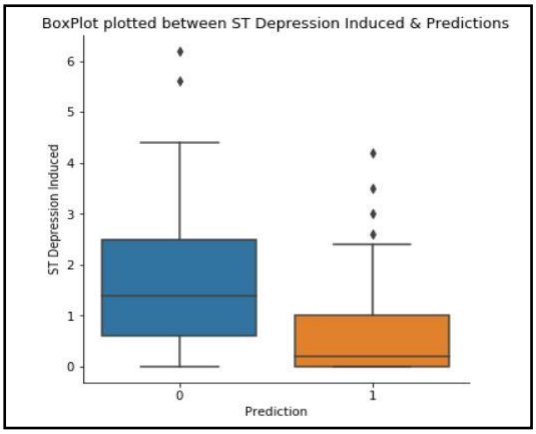


Figure 3.1.3

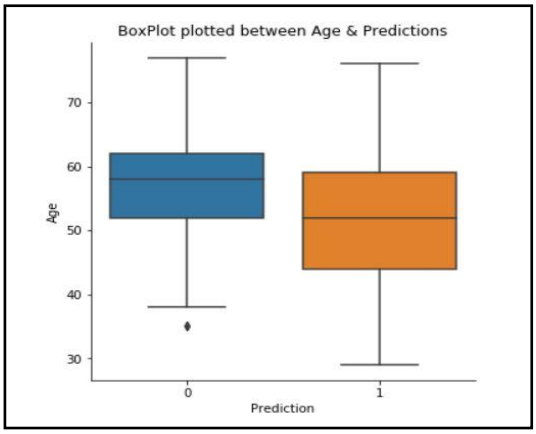


Figure 3.1.4

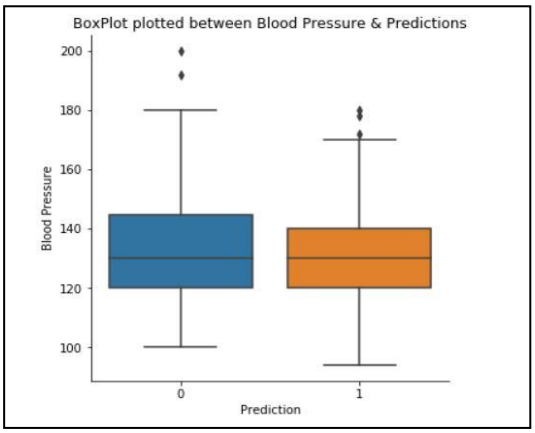


Figure 3.1.5

As you can see there are five boxplots in total. The first plot i.e. Figure 3.1.1 shows the relationship between the maximum heart rate and the target variable. When observed closely we get to know that the medians, maximums, minimums, upper quartile, and the lower quartile differ by a good proper difference, and we observe there are also a few outliers. As the difference is relatively good, the maximum heart rate can be regarded as a good independent variable to predict the target variable. The second plot i.e. Figure 3.1.2 shows the relationship between cholesterol and the target variable. Here we observe there is no great difference between the medians, maximums, minimums, upper quartile, and the lower quartile and therefore it could be said that this is not a proper independent variable that could solely be used for prediction. Even in this plot there are a few outliers. The third plot i.e. Figure 3.1.3 shows the relationship between the ST depression induced and the target variable. Here we observe again as in the previous case, there is a good relative difference between medians, maximums, minimums, upper quartile, and the lower quartile. As again in the previous case it can be said that it is a good prediction variable or a good independent variable that could solely be used to predict the target variable. This plot has some outliers as well. Moving on the plot four i.e. Figure 3.1.4, this shows the relationship between age and the target variable. As we observe there is a good relative difference between the minimums, medians, and the lower quartiles but there is no great difference between the upper quartiles and

the maximums. There are a very few (almost none) number of outliers in this plot meaning there is no case of any exceptions. Coming to the last plot the fifth one i.e. Figure 3.1.5 we observe there is no difference between the medians, maximums, minimums, lower quartiles and the upper quartiles, and this is not a significant contributor in prediction of the target variable. There are a good number of outliers in this plot as well.

3.2. Countplots

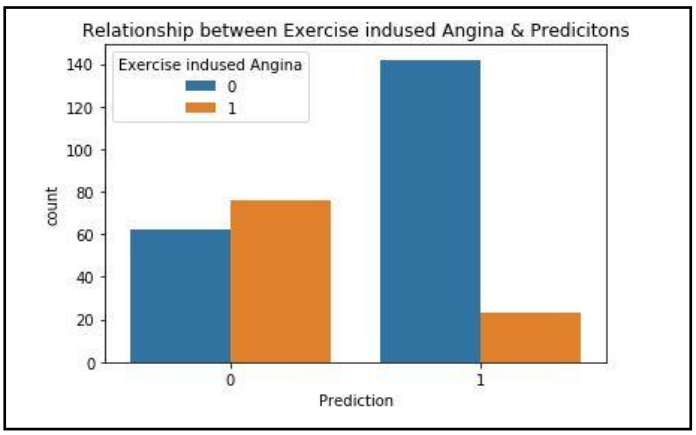


Figure 3.2.1

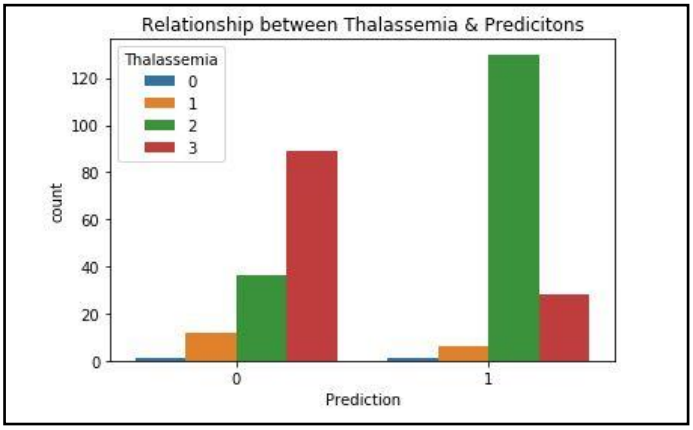


Figure 3.2.2

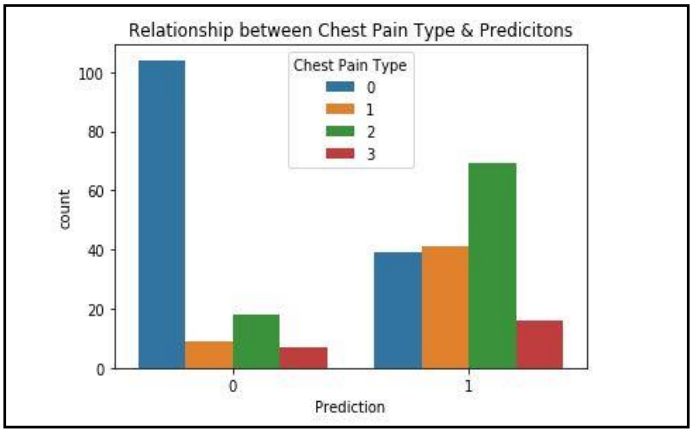


Figure 3.2.3

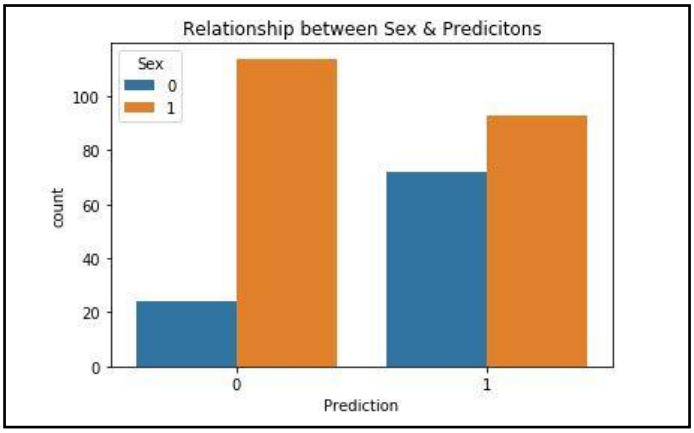


Figure 3.4.4

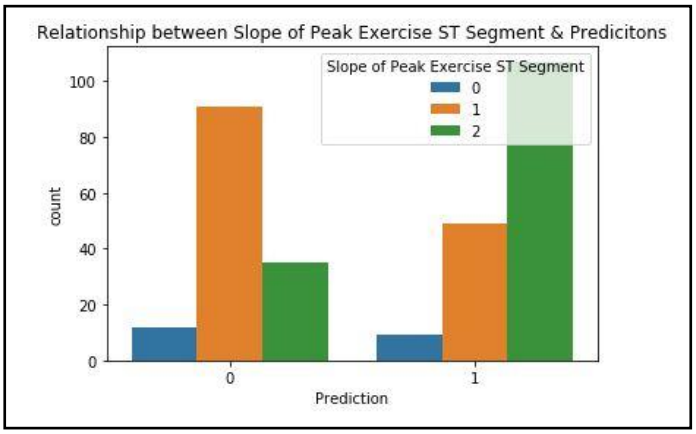


Figure 3.2.5

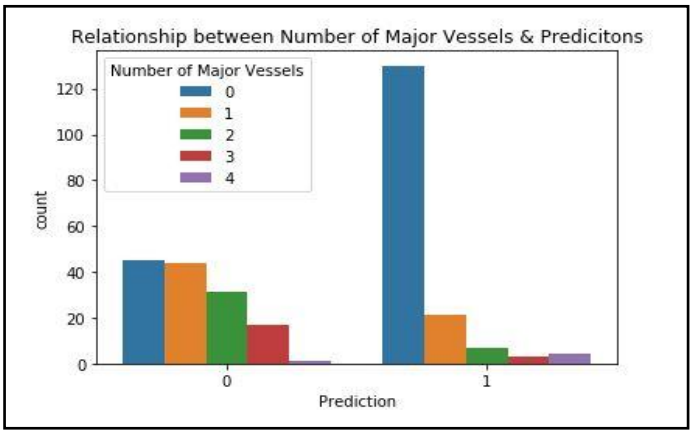


Figure 3.2.6

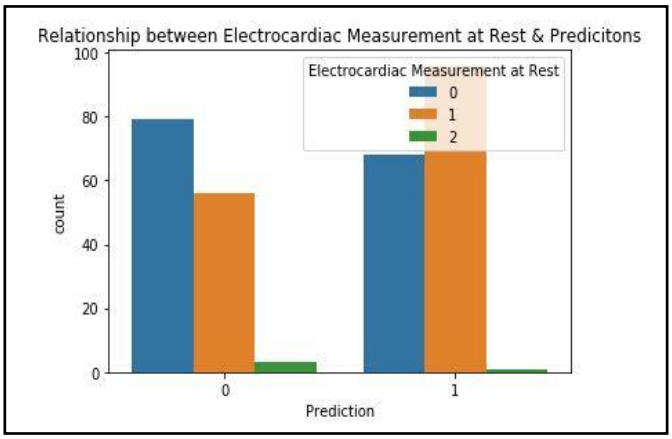


Figure 3.2.7

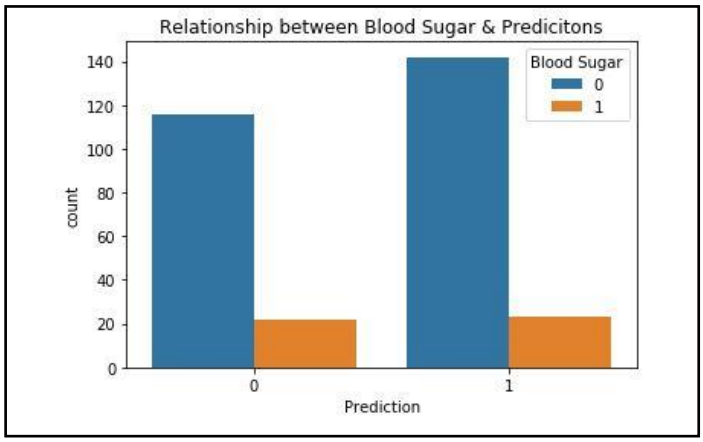


Figure 3.2.8

These are the Countplots that are used to plot and show the relationship between two categorical variables i.e. it shows the count. The first plot i.e. Figure 3.2.1 shows the relationship between exercises induced angina and the target variable. The insight we could draw from it is, if the person isn't having angina induced due to exercise he may have huge probability of suffering from a heart related disease. The second plot i.e. Figure 3.2.2 shows the relationship between thalassemia and the target variable. Here if we observe carefully, we get to know that the person with the thalassemia value of two would most probably be suffering from a heart related disease. The third plot i.e. Figure 3.2.3 shows the relationship between chest pain type and the target variable. The insight we could draw from here is the person with chest pain type zero would not be suffering from a heart related disease most probably. Moving on to plot four i.e. Figure 3.2.4, this shows the relationship between sex and the target variable. As the plot is equally distributed (near equal) we cannot draw any insights here. Coming to plot five i.e. Figure 3.2.5, this shows the relationship between slope of peak exercise ST segment and the target variable. The insight we could draw is if the person has a one as the value for slope of peak exercise ST segment then most probably he isn't suffering from any disease and if the person has a three as the value for slope of peak exercise ST segment then most probably he is suffering from one. Moving on to plot six i.e. Figure 3.2.6, this shows the relationship between number of major vessels and the target variable. The insight we could draw here is, if the person has a zero as the value then most probably, he is suffering from a heart related disease. Moving on to plot seven i.e. Figure 3.2.7, this shows the relationship between electro cardiac measurement at rest and the target variable. As the plot is equally distributed (near equal) we cannot draw any insights here. Now, the last plot i.e. Figure 3.2.8 shows the relationship between blood sugar and the target variable. As the plot is equally distributed (near equal) we cannot draw any insights here as well.

4.Predictive Modeling

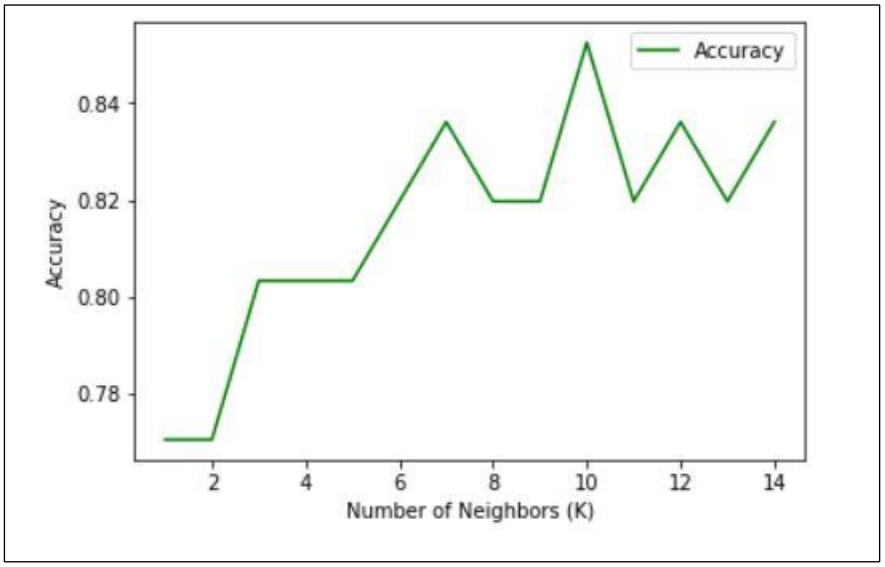
On the basic note this being a classification model, it needs to be built on classification algorithms. So, here for model training & testing the algorithms which I chose were classification algorithms as it goes. The model had been built on multiple classification models and accuracy of the model when those algorithms were under the box were noted.

4.1. Classification Models

The standard classification algorithms which are most widely used and accepted by most of the data scientists were put into model training and testing. The model was trained and tested using the following four classification algorithms.

4.1.1 K-Nearest Neighbors,

One of the most basic yet efficient classification algorithms used to build dependable classification models. So, the dataset was split into train & test sets using a built-in function. The training set consisted of eighty percent of the whole dataset, and twenty percent as the test set. As the value of K needs to be provided by the model-trainer itself, I first tried to figure out the best value for K, i.e. figure out when the model performs best depending on the accuracy. The plot was something like the following,



As we can see, the best accuracy obtained was when the value of K was set to ten. Therefore, ultimately the model was trained and tested using the built-in libraries, packages and functions. The model built using the K-Nearest Neighbors gave out the following accuracy scores, F1-score, and confusion matrix score to be precise.

```
Model-1's Jaccard Accuracy:  0.8524590163934426
Model-1's F1-Score Accuracy:  0.853660164156861
```

4.1.2 Decision Trees,

One of the easy to use yet efficient classification algorithms used to build dependable classification models. So, the dataset was split into train & test sets using a built-in function. The training set consisted of eighty percent of the whole dataset, and twenty percent as the test set. The model was built, trained, and tested and it gave out the following accuracy scores, F1-score, and confusion matrix to be precise. The criterion used here was entropy.

```
Model-2's Jaccard Accuracy:  0.8524590163934426
Model-2's F1-Score Accuracy:  0.8507010812696197
```

4.1.3 Support Vector Machines (SVM),

One of the easy to use yet efficient classification algorithms used to build dependable classification models. So, the dataset was split into train & test sets using a built-in function. The training set consisted of eighty percent of the whole dataset, and twenty percent as the test set. The model was built, using the built-in functions obviously, then trained, and tested and it gave out the following accuracy scores, F1-score, and confusion matrix to be precise. The kernel used here was the rbf and gamma was set to auto.

```
Model-3's Jaccard Accuracy:  0.8688524590163934
Model-3's F1-Score Accuracy:  0.8699169682776241
```

4.1.4 Logistic Regression,

One of the most dependable yet efficient classification algorithms used to build proper classification models. So, the dataset was split into train & test sets using a built-in function. The training set consisted of eighty percent of the whole dataset, and twenty percent as the test set. The model was built, using the built-in functions obviously, then trained, and tested and it gave out the following accuracy scores, F1-score, and confusion matrix, and log loss to be precise. The probability of the prediction was also figured out. The solver was set to liblinear and C was set to 0.01.

```
Model-4's Jaccard Accuracy:  0.8852459016393442
Model-4's F1-Score Accuracy:  0.8848609284270637
Model-4's Log Loss Accuracy:  0.4363789674264379
```

4.Future Scope

The model was trained and tested on a dataset that had entries that could expand itself as time goes by. Even the columns i.e. the attributes could be even more convincing and could increase in number therefore leading to easy and better predictions. The model was trained and tested on different standardized classification algorithms. The best could be therefore chosen, an API could be designed, connected to a web application, therefore making it easy for patients and users as well. There's a lot left to do beyond imaginations. 😊

5.Final Conclusions

This is a report that is being made for the project which is a part of applied data science capstone project. In my study I learnt a lot, about how different features could contribute to predicting the whether a person is suffering from a heart related disease. This project would further be taken up and with certain improvements would definitely become a good, efficient, dependable working model....