HEALTHCARE ANALYTICS AND PREDICTIONS ON HEART DISEASE DATA

TECHNOLOGIES: BUSINESS INTELLIGENCE

DOMAIN: HEALTHCARE

ABSTRACT:

- Using the dataset provided, we built a simple model for predicting the Heart Disease in patients which can be used in future. Our model revealed a 86% of Accuracy using past data and few major features which can help predicting the heart disease in future.

INTRODUCTION

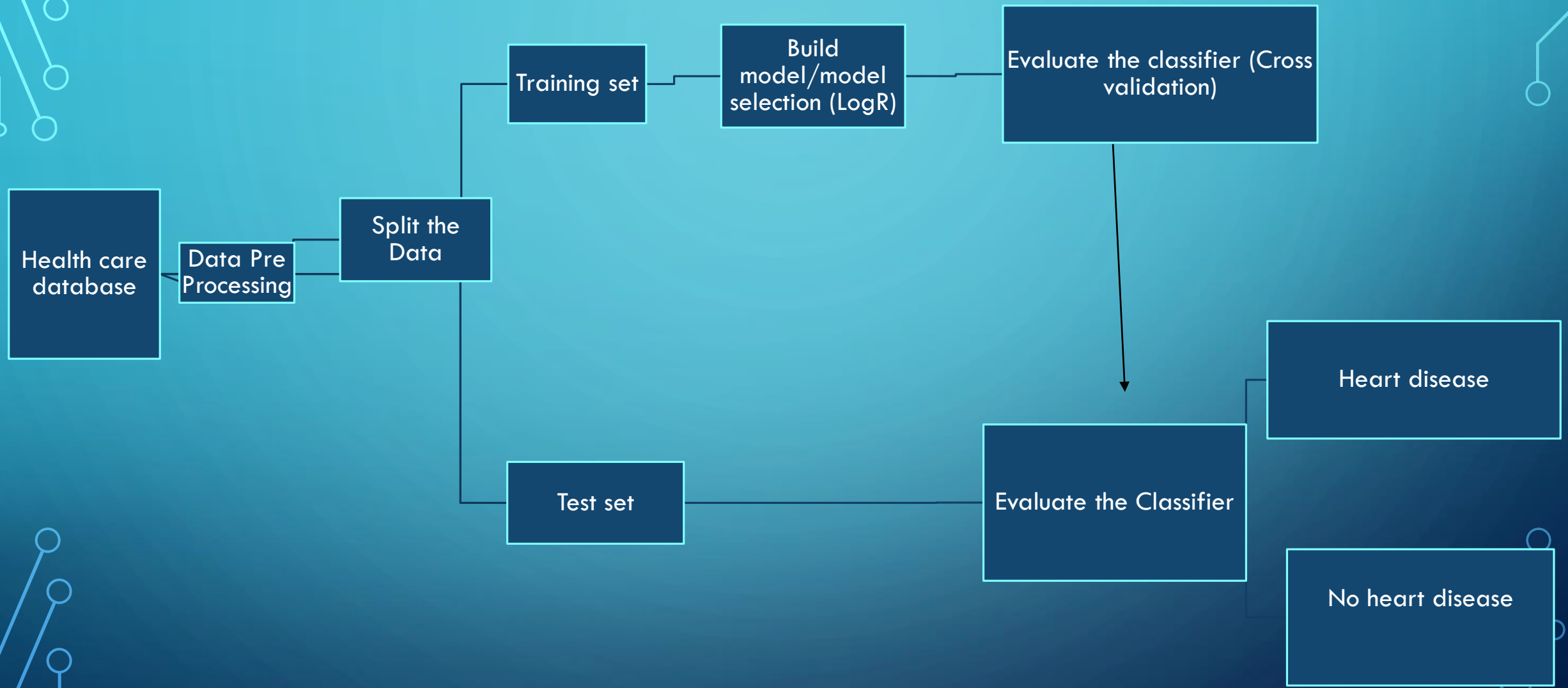
Health is real wealth in the pandemic time we all realized the brute effects of covid-19 on all irrespective of any status. Techniques of data science and predictive analytics can be used to predict the Heart Disease. In this project, we are provided with the Heart disease.csv. This database contains 76 attributes, but all published experiments refer to using asubset of 14 of them. In particular, the Cleveland database is the only one that has been used by ML researchers to this date. The “goal” field refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4. Attribute Information:> 1. age> 2. sex> 3. chest pain type (4 values)> 4. resting blood pressure> 5. serum cholesterol in mg/dl> 6. fasting blood sugar > 120 mg/dl> 7. resting electrocardiographic results (values 0,1,2)> 8. maximum heart rate achieved> 9.exercise induced angina> 10. oldpeak = ST depression induced by exercise relative to rest> 11. the slope of the peak exercise ST segment> 12. number of major vessels (0-3)colored by flourosopy> 13. thal: 3 = normal; 6 = fixed defect; 7 = reversible defect.

PROJECT OBJECTIVE AND BENEFITS

- The goal of this project is to use techniques of data science to Analyze and Predict the Heart Disease, to Find key metrics and factors and show the meaningful relationships between attributes. We are required to analyze this health and medical data for better future preparation.
- PROJECT AIM: To develop a Heart Disease predicting model using Machine learning

- BENEFITS:
- Detection of Heart Disease
- Gives better insights of patients
- Helps in easy flow for managing resources
- Heart disease is predicted

PROJECT METHODOLOGY



LIFE CYCLE OF THE PROJECT

PHASE 1: HEALTHCARE DATABASE:

Problem Statement: Health is real wealth in the pandemic time we all realized the brute effects of covid-19 on all irrespective of any status. You are required to analyze this health and medical data for better future preparation. This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by ML researchers to this date. The “goal” field refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4. Attribute Information:

PHASE 2: DATA PREPROCESSING:

- Pre-processing refers to the transformations applied to our data before feeding it to the algorithm. Data Preprocessing is a technique that is used to convert the raw data into a clean data set. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis.
- For achieving better results from the applied model in Machine Learning projects the format of the data has to be in a proper manner. Some specified Machine Learning model needs information in a specified format, for example, Random Forest algorithm does not support null values, therefore to execute random forest algorithm null values have to be managed from the original raw data set.
- Another aspect is that the data set should be formatted in such a way that more than one Machine Learning and Deep Learning algorithm are executed in one data set, and best out of them is chosen.

We have been provided with the healthcare data, first step will be importing the pandas which will help us import the data. After importing the data we'll need to do some data cleaning like : Data Shape, Data type, data size, Describing the data, check whether there are any null/missing values in the data. If there any null or missing values present then we will have to delete the values or replace the values with the help of KNN imputation which will take out the average of the surrounding values.

After, do the EDA (Exploratory data analysis) where we plot the graph using matplotlib library which shows the number of Non healthy/Heart disease people are more comparatively to Healthy people. It also shows the number of males having heart disease are more than female.
We have plotted a pie chart which shows that 53% of population from the data are having heart disease and 46% do not have heart disease.

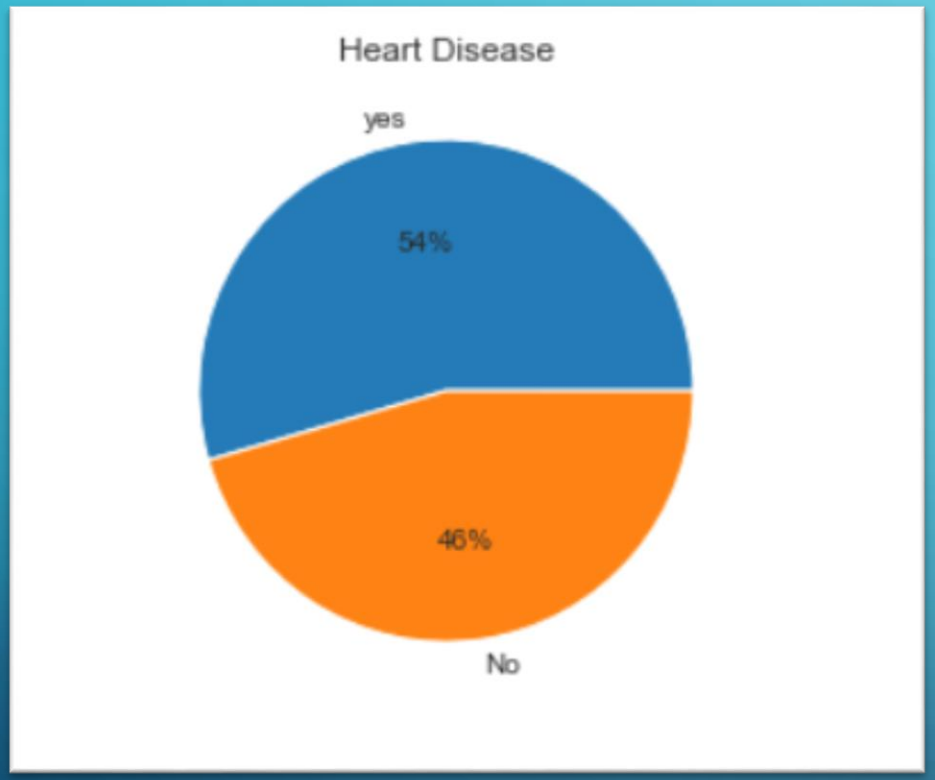
We have also plotted histogram graphs which showed us the distributions among all the features.
With the help of histogram, we concluded that features like: Chest pain, Resting blood pressure, cholesterol , person's maximum heart rate achieved can be major features which can lead us to the detection of heart disease or responsible for getting higher chances of heart disease.

After plotting the histogram, we plotted Barplots with the help of seaborn library. Bar plots helps to estimate the central tendency for numeric variables with the height of each rectangles and it provides some indication of uncertainty.
Barplots helped us in knowing the possibilities of having heart disease as per the feature.

- we observed that females have the higher possibilities in order to get heart disease.
- In chest pain Type 2 (atypical agina(bar1)) has the possibility to have heart disease.
- Exercise induced angina has the higher possibility of having heart disease.

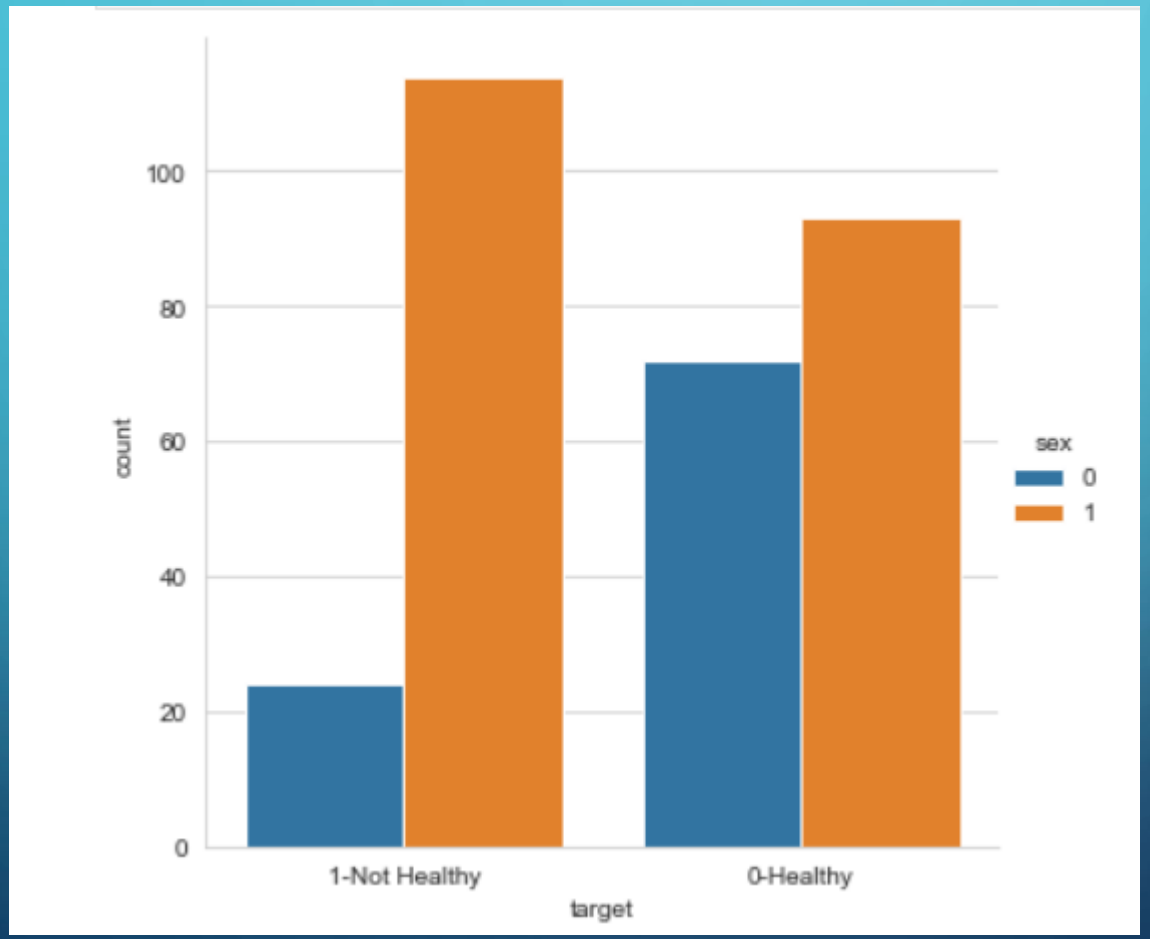
EXPLORATORY DATA ANALYSIS

The dataset was important in Python and calculations were performed using python. We plot the following figures:

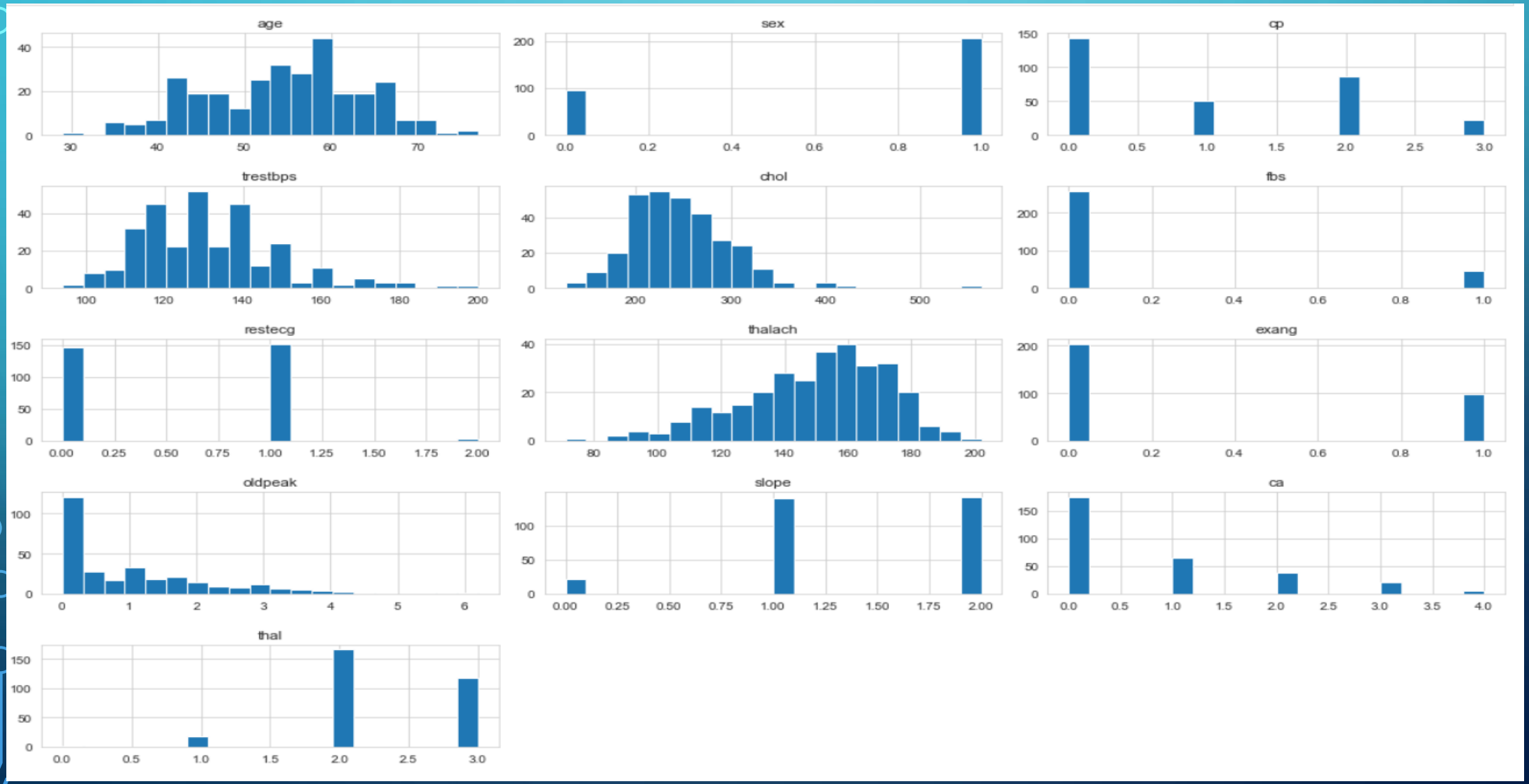


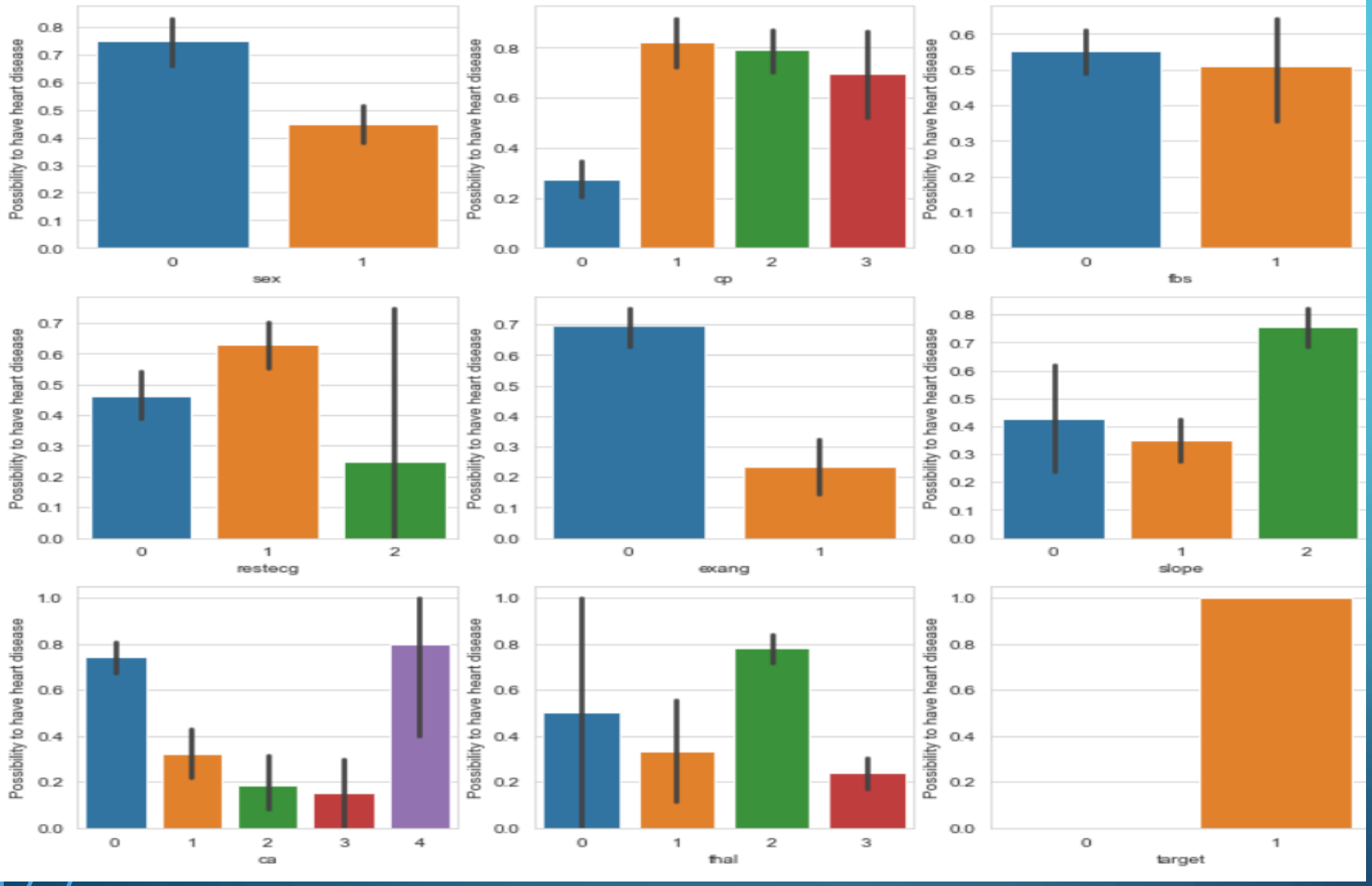
The Pie chart shows the percentage of patients having Heart Disease(yes) and patients who do not have heart disease (no).

The following graph shows the count of male and females who are healthy and not healthy.
It is observed that females are more Healthier than male.



THE HISTOGRAMS CONCLUDES AGE, CHOLESTEROL, RESTING BLOOD PRESSURE, AND MAXIMUM HEART RATE ACHIEVED PLAYS THE MAJOR ROLE IN DETECTION OF HEART DISEASE.





- Here we can observe that females have the higher possibilities in order to get heart disease.
- In chest pain Type 2 (atypical agina(bar1)) has the possibility to have heart disease.
- Exercise induced angina has the higher possibility of having heart disease.

The above Barplot shows the possibilities of having heart disease as per the features.

PHASE 3: SPLITTING THE DATA: **Split the Dataset With scikit-learn's train_test_split()**

Splitting dataset is essential for an unbiased evaluation of prediction performance.
We split the data into 80% Training set and 20% testing set.

After splitting the dataset we tried Linear regression, Logistic regression, Random forest and SVM.
Linear regression gave 53% of accuracy score.
Logistic regression gave 86%
Decision tree gave 77%
Random forest gave 84%
SVM gave 86% same as logistic regression.

PHASE 4: BUILD MODEL /MODEL SELECTION

Asper the accuracy of the models we selected logistic regression with 86%. It is mused to predict the data value based on observation of datasets. So it will be suitable for healthcare analysis.

Logistic regression can be binomial, ordinal or multinomial. Binomial or binary logistic regression deals with situations in which the observed outcome for a dependent variable can have only two possible types, "0" and "1" .

PHASE 5: EVALUATION (CROSS VALIDATION)

SVM Classification Report

	precision	recall	f1-score	support
0	0.90	0.80	0.85	35
1	0.84	0.93	0.88	41
accuracy			0.87	76
macro avg	0.87	0.86	0.87	76
weighted avg	0.87	0.87	0.87	76

Accuracy: 86.84%

Random Forest Classification Report

	precision	recall	f1-score	support
0	0.81	0.86	0.83	35
1	0.87	0.83	0.85	41
accuracy			0.84	76
macro avg	0.84	0.84	0.84	76
weighted avg	0.84	0.84	0.84	76

Accuracy: 84.21%

DT Classification Report

	precision	recall	f1-score	support
0	0.74	0.80	0.77	35
1	0.82	0.76	0.78	41
accuracy			0.78	76
macro avg	0.78	0.78	0.78	76
weighted avg	0.78	0.78	0.78	76

Accuracy: 77.63%

LogistiC Regression Classification Report

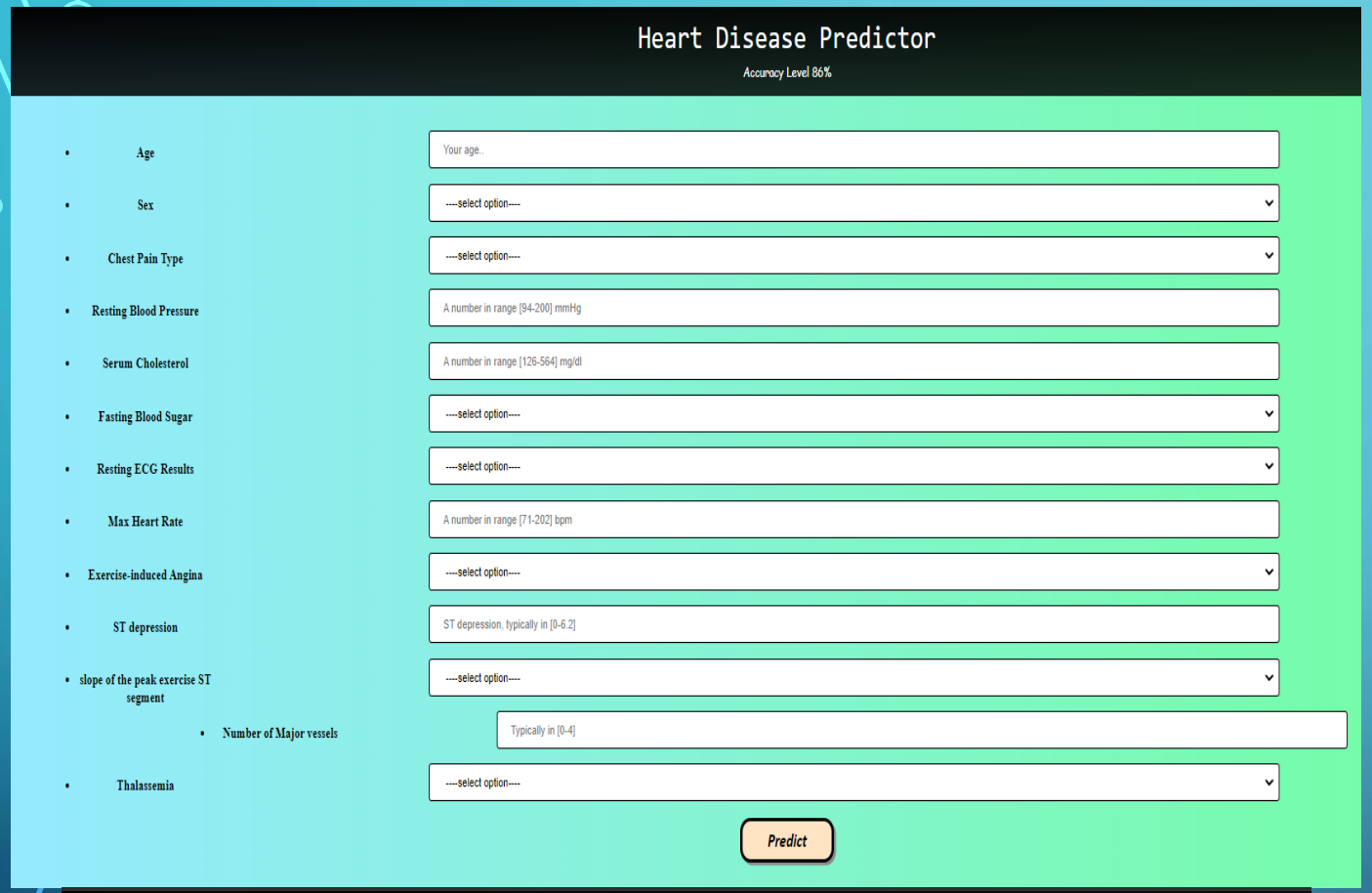
	precision	recall	f1-score	support
0	0.90	0.80	0.85	35
1	0.84	0.93	0.88	41
accuracy			0.87	76
macro avg	0.87	0.86	0.87	76
weighted avg	0.87	0.87	0.87	76

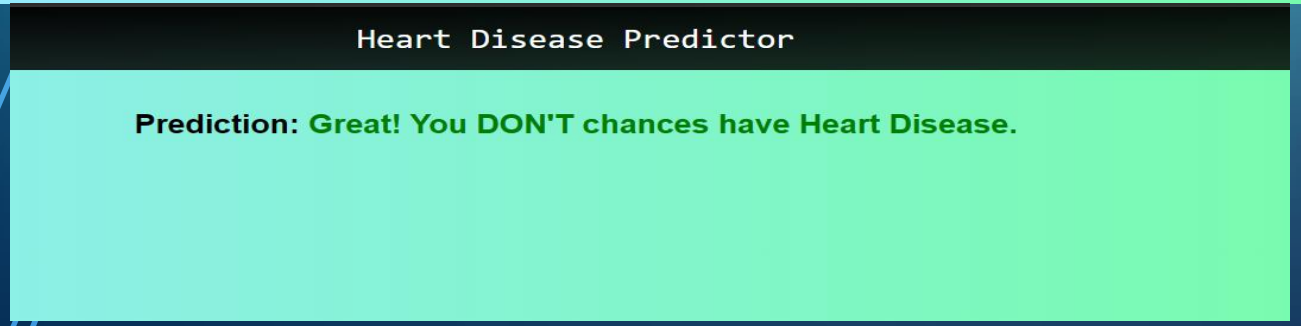
Accuracy: 86.84%

After cross validating our models manually we can se logistic regression has highest accuracy level as comparative to others. Cross validation is a technique which calculates and gives the list of all the model's accuracy. It can be the easiest way to know accuracy % of the model.

PHASE 6: EVALUATE THE CLASSIFIER ON TESTING SET: *Once the model is selected, the model is fitted on the test set to predict the outcome. After fitting the model on the test, we can save the model in pikle file and now it can be used in our Heart disease predicting API.*

END RESULT





This is how the end result looks.