Project Report

Data Collection:

First of all we collect the data from Kaggle website.The data set we select is Heart Disease.In this dataset we have previous data of different things through which we suffer in heart disease.

Problem definition:

The problem is to predict the person have heart disease or not according to the previous data provided in the dataset.

Data preprocessing:

In data preprocessing we have the following steps which are given and explained below:

1> Firstly we import the libraries and upload the csv file in the compiler.The compiler read the csv file which have the following columns in this dataset.

- Age
- Gender
- Cholesterol
- Blood Pressure
- Heart Rate
- Smoking
- Alcohol
- Exercise Hours
- Family History
- Diabetes
- Obesity
- Stress Level
- Blood Sugar
- Exercise Induce Angina
- Chest Pain Type
- Heart Disease

In this data we have 16 columns and 1000 Rows.

2> Remove missing values: check the dataset and remove the missing values from it.

3> Check the duplicate rows: In this dataset we also check the duplicate values which can affect the accuracy of our model in future.

4> After checking duplicate values we check the inconsistent data entries in tha data and correct them if we have the some unconscious entries in the data.

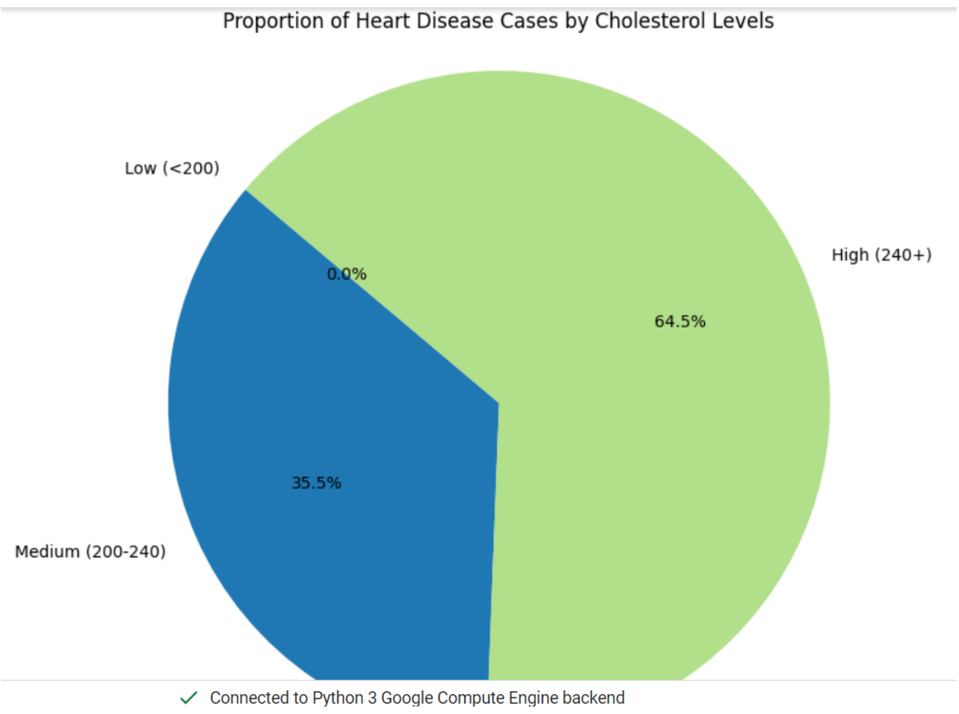
5> And then we can move to the next step which is converting categorical columns into numerical columns.After converting these columns into numerical form check the data set again.

6> In these steps we almost cleaned our data and then draw a correlation heatmap to visualize and the dataset easily.

Data Visualization:

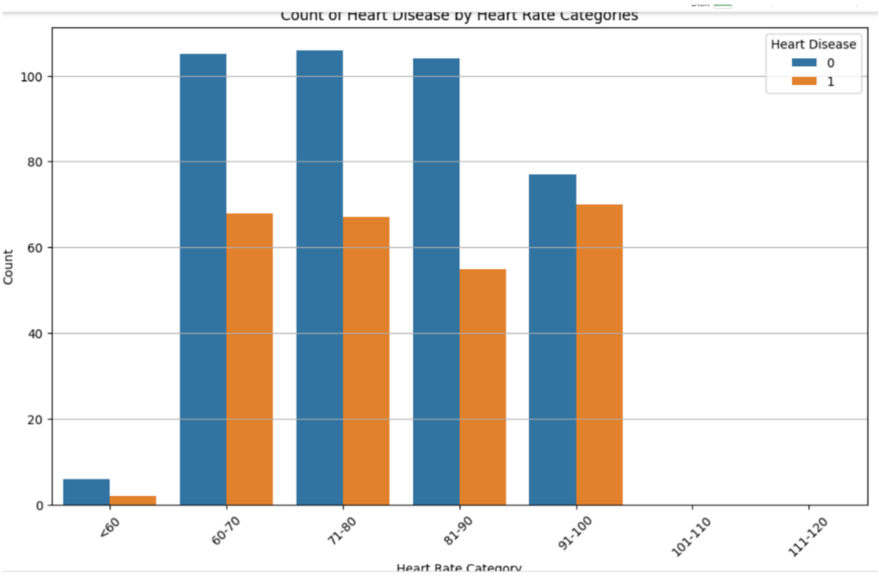
Once the dataset is cleaned apply Data Visualization on this code.In visualization we compare different columns with the target column Heart Disease which we want to predict.Here is the some columns result in the form of graph to analyze it.

- **Heart disease cases by cholesterol level**

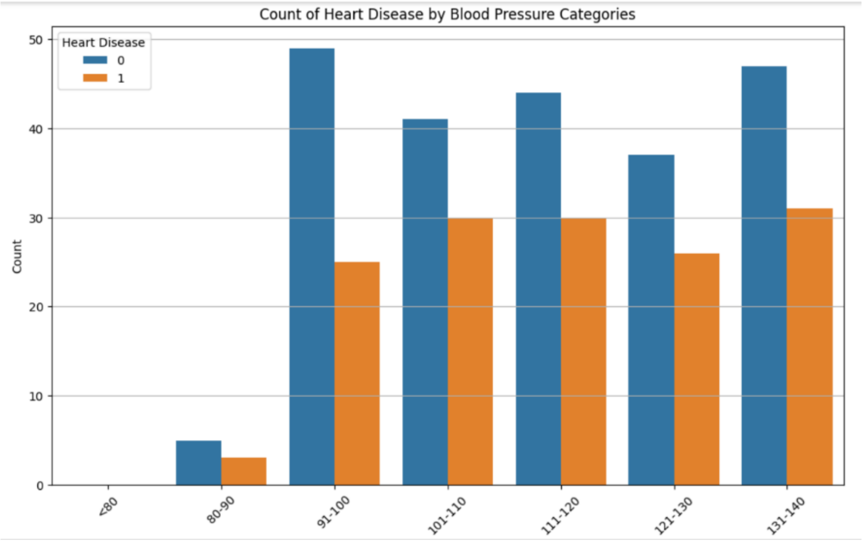


✓ Connected to Python 3 Google Compute Engine backend

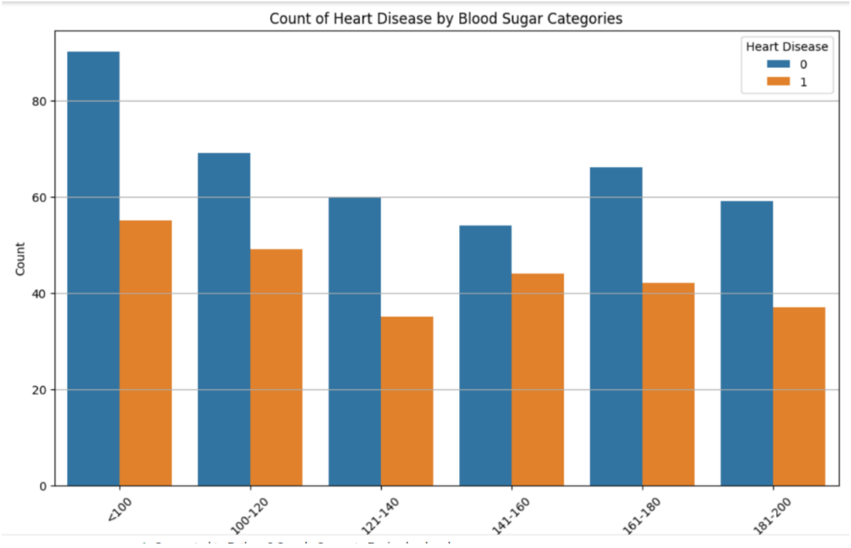
- **Heart disease case by Heart Rate**



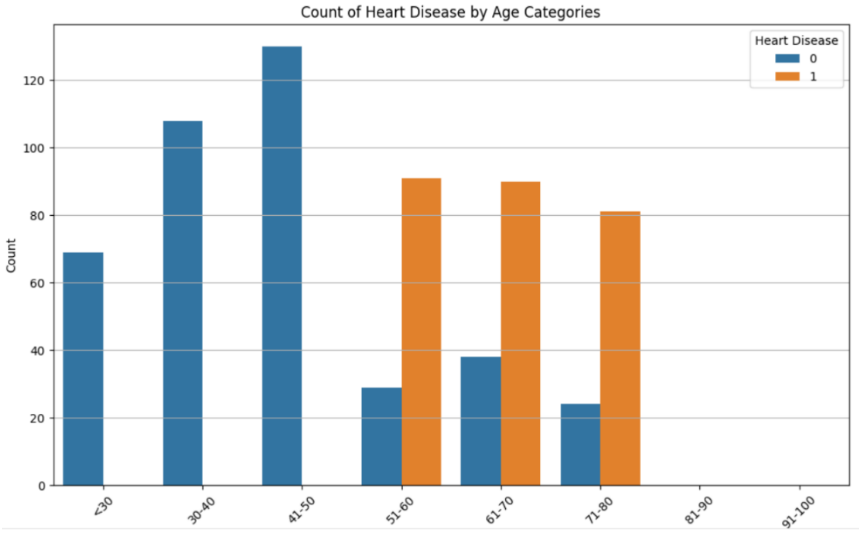
- **Heart disease case by Blood Pressure**



- **Heart disease case by Bood Sugar**



- **Heart disease case by Age**



Through this visualization technique we easily understand the relation of other columns with the target column.

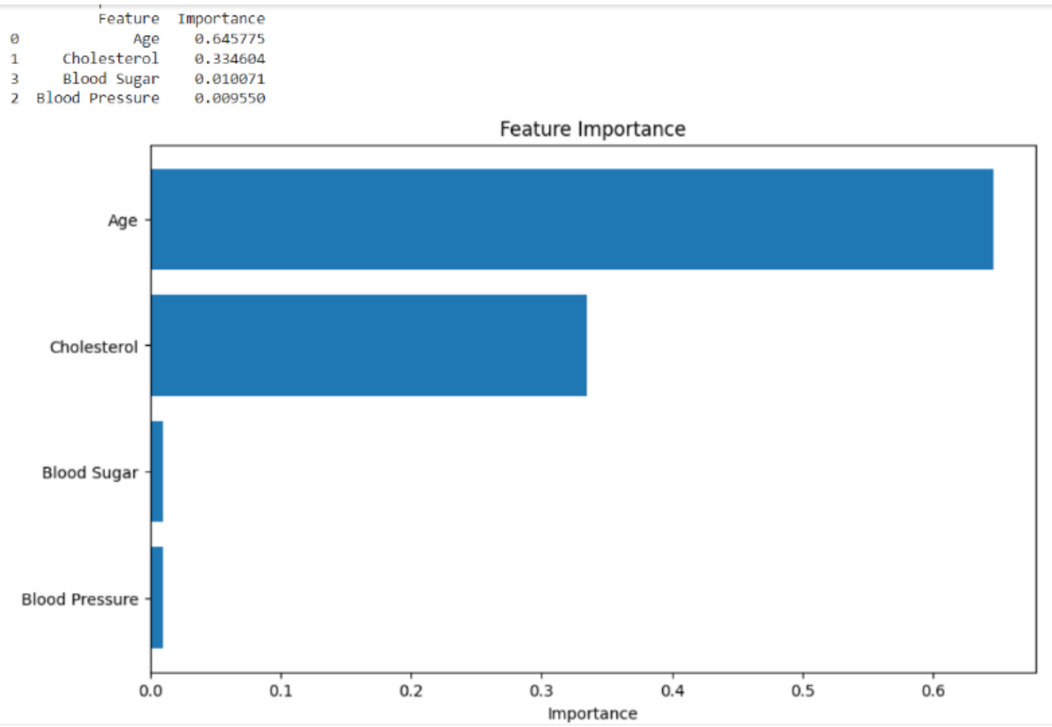
Outlier Analysis:

The next step is to do the outlier analysis of the give data.The purpose of this step is to check and handle any outlier in the given data.There are two method to detect and handle the outliers.So we can use in this code IQR method to handle the outliers.

Feature Engineering:

In this method we use feature importance technique to identify most influential features.we also select the some features for the model training.These are the feature and their importance and the graphical representation of feature importance is also shown:

- **Feature Importance**



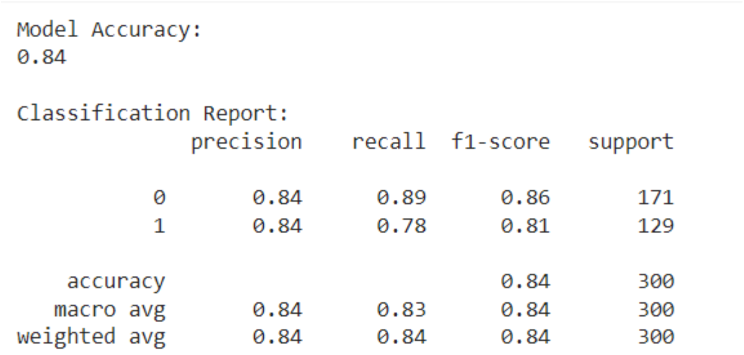
```
          Feature  Importance
0             Age    0.645775
1     Cholesterol    0.334604
3     Blood Sugar    0.010071
2  Blood Pressure    0.009550
```

Model Selection and Accuracy:

Applying model on the dataset is the main work to predict the any dataset.Thats why, we can apply different types of classification models on the data and select some of the models which are giving accuracy correct and predict our model easily and understandably.Here are some models having their accuracy and graph which can be applied on the data.

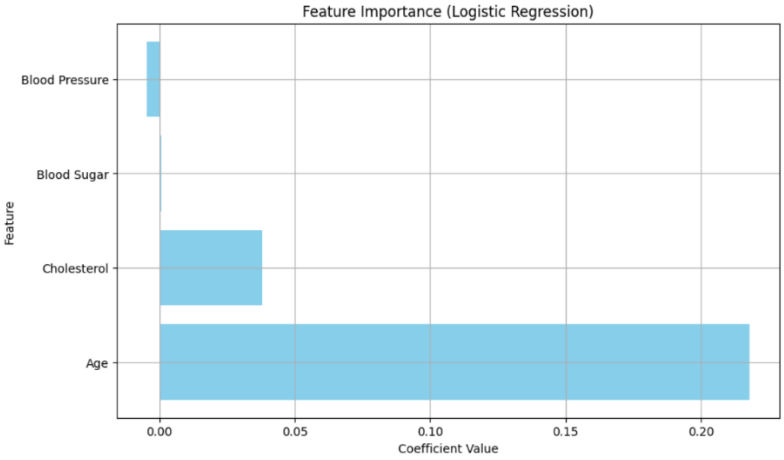
- **Sklearn logistic Reggression**

  Accuracy:



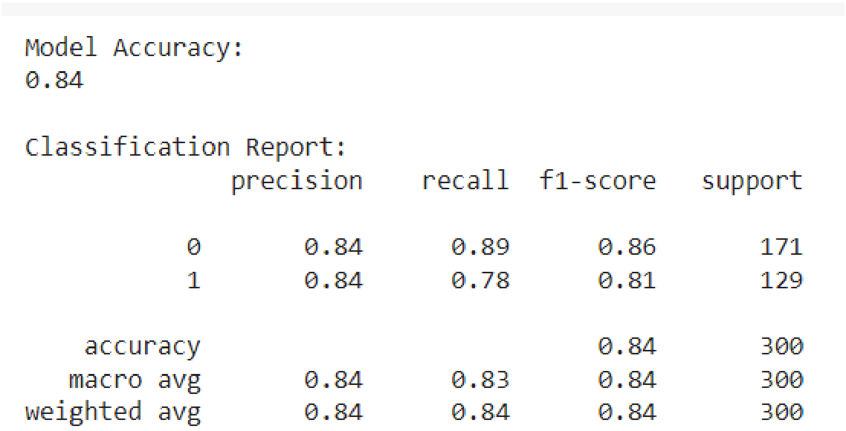
```
Model Accuracy:
0.84

Classification Report:
              precision    recall  f1-score   support

           0       0.84      0.89      0.86       171
           1       0.84      0.78      0.81       129

    accuracy                           0.84       300
   macro avg       0.84      0.83      0.84       300
weighted avg       0.84      0.84      0.84       300
```

Graph:



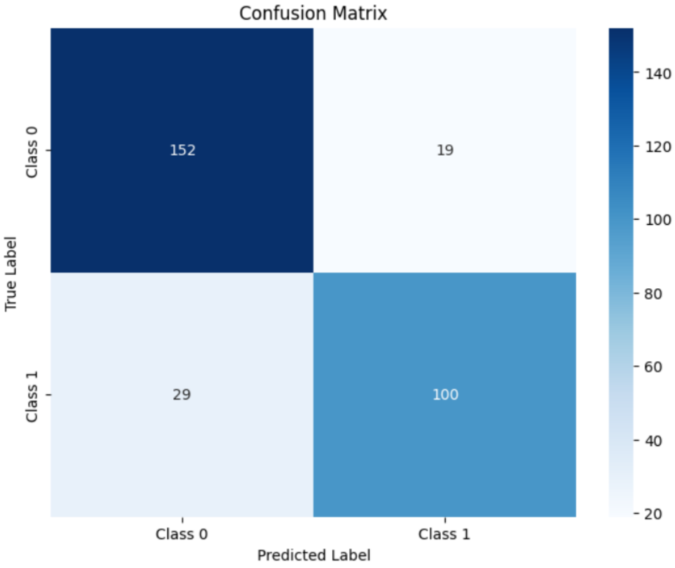
- ## Support Vector Machine

Accuracy:



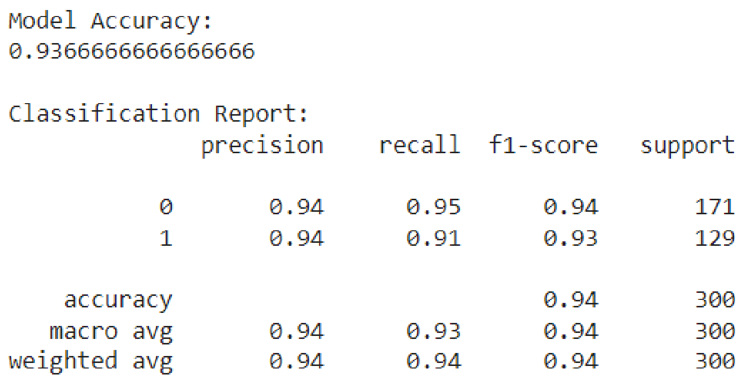
```
Model Accuracy:
0.84

Classification Report:
              precision    recall  f1-score   support

           0       0.84      0.89      0.86       171
           1       0.84      0.78      0.81       129

    accuracy                           0.84       300
   macro avg       0.84      0.83      0.84       300
weighted avg       0.84      0.84      0.84       300
```

Graph:



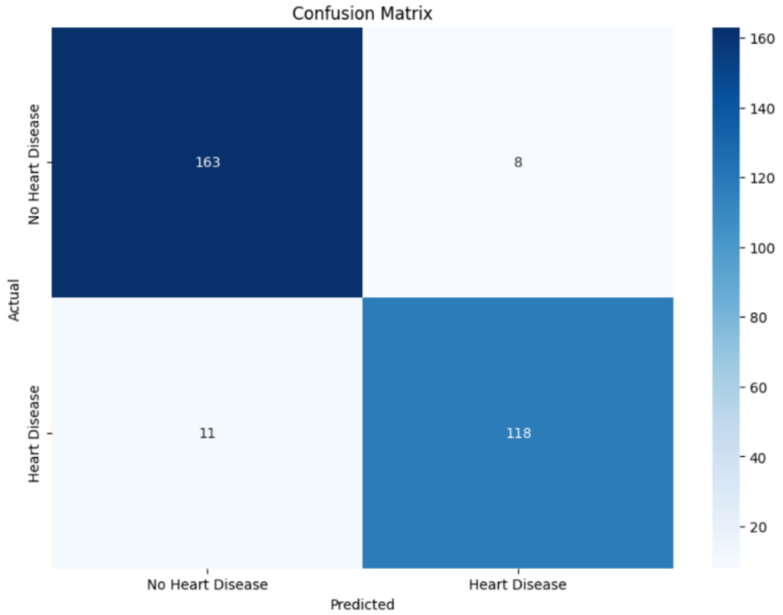
- ## Sklearn k-nearest neighbors

Accuracy:



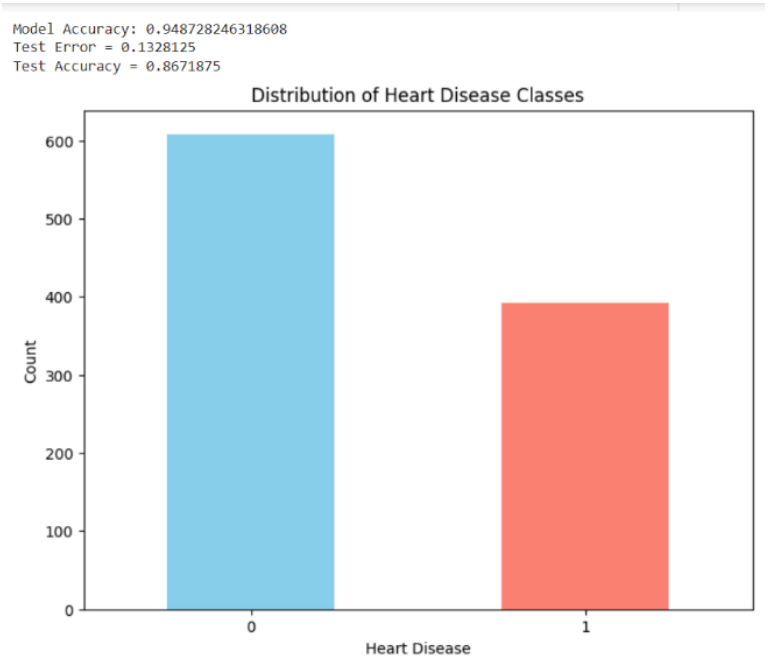
```
Model Accuracy:
0.9366666666666666

Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.95      0.94       171
           1       0.94      0.91      0.93       129

    accuracy                           0.94       300
   macro avg       0.94      0.93      0.94       300
weighted avg       0.94      0.94      0.94       300
```

Graph:



- ## Pyspark Logistic Reggression

Accuracy and Graph:



```
Model Accuracy: 0.948728246318608
Test Error = 0.1328125
Test Accuracy = 0.8671875
```

Evaluation Metrices:

Evaluation metrice is the method to measure the accuracy,precision,F1 score and confusion metrix.In evaluation metrices their include the total prediction of the model accuracy,precison etc.we can judge from this that our models prediction is correct or not.

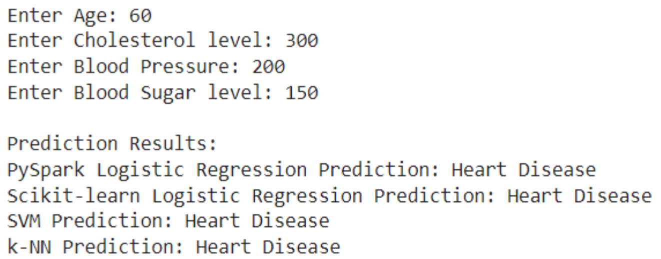
Interprtablity:

Through this method in which we can understand the accuracy and prediction of the model.The purpose of this is to make the model more understandable for the user.

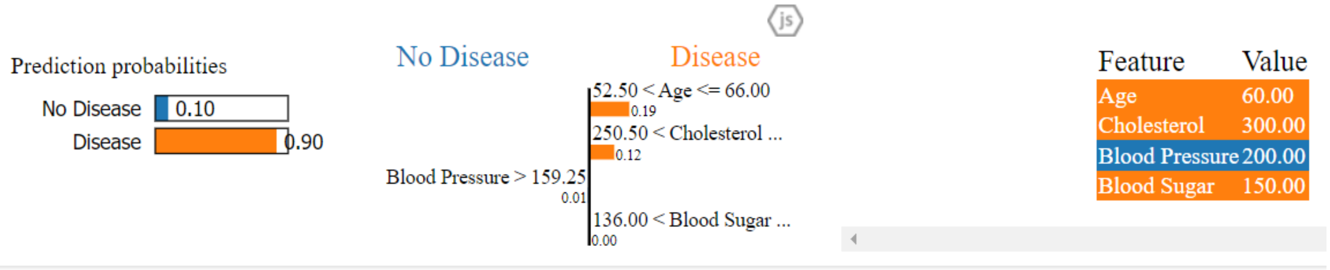
Final Result:

The final result for the model is that we can give the input to the model and the model predict it.In this model by analysing graphs accuracy and by understanding all the things we can conclude that when a user give the input of these things Age,Cholestrol,Blood Pressure,Blood sugar level then the model predict the output that the user have heart disease or not.For example here the result of inputing some data in the model to predict heart disease.

Prediction of Heart Disease



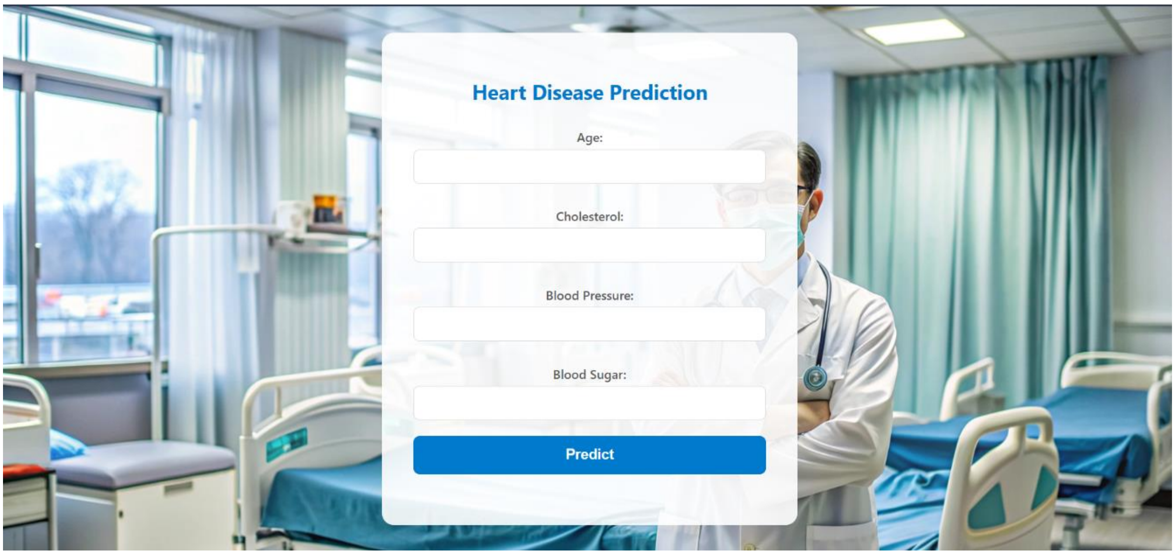
```
Enter Age: 60
Enter Cholesterol level: 300
Enter Blood Pressure: 200
Enter Blood Sugar level: 150

Prediction Results:
PySpark Logistic Regression Prediction: Heart Disease
Scikit-learn Logistic Regression Prediction: Heart Disease
SVM Prediction: Heart Disease
k-NN Prediction: Heart Disease
```



Feature	Value
Age	60.00
Cholesterol	300.00
Blood Pressure	200.00
Blood Sugar	150.00

Front End View:

In front end view same like we put four inputs and the model predict the result that the person have heart disease or not.Here is the front end picture of the model.



Innovation:

This is the heart disease prediction base model.Through this model we can check the person having heart disease or not.We can also make a website or app and lounched it.On this app or website the user can check that he has heart disease or not by inputing two to three inputs about their health.