**Background Information**

There has been a significant increase in the number of patients worldwide suffering from various diseases. Big data and AI solutions play a game-changing role in this field. How amazing it would be that we know before-hand what is happening with our internal organs of the body. Moreover, what are the chances that if we follow the same routine, we can end up having some disease.AI and machine learning can help us predict this better than humans with proper supporting analysis report backing the prediction. This report touches on the basics of working behind AI’s s world. The heart Disease dataset retrieved from the UCI machine learning repository is the backbone for this assignment. The model trains on 14 various attributes of 303 patients having different symptoms. There are numerous visualizations present in this report that provide descriptive and diagnostic analysis of the dataset’s various attributes. The visualizations help understand the dependency of two or more attributes towards each other and contribute to the model’s prediction process. The model is 81.9% accurate in predicting whether the person has the disease or not. The dataset has been trained on two different algorithms, and each algorithm’s performance measures are depicted in this report. The tuning of the model and selecting the best performance model are backed up with the code’s proper explanations and snippets.

**Purpose**

This report’s purpose is to showcase the nature of the dataset briefly and what steps were taken in order to handle and prepare the dataset for the model training phase. The primary goal is to choose the best model with high accuracy and low False-positive rates. Various performance metrics such as accuracy score, recall score, F1 score, confusion matrix are used to select the model wisely, and they are explained through this report.

**Audience**

This report aims to provide our understanding of the problem statement, dataset, and classifiers being used to the primary audience, i.e., Prof. Aruna Dorai. Our utmost effort is to make this report of our assignment more informative and easy to understand when compared with the code. Therefore, fellow students and companions are the secondary audiences for this report. Even with basic knowledge of data science and python programming language, the code can be easily understood as each, and every step is elaborated. This report has provided several visualizations to make it easy to understand and informative to our primary and secondary audience.

1. Dataset Description

The dataset used in this assignment is called as **Heart Disease Data set** and is retrieved from UCI Machine Learning Repository. This dataset originally has **76 attributes** and **303 instances. Still,** for this assignment, a subset of this dataset with only **14 attributes** has been used. The dataset shows different heart disease presence levels from 1 to 4 and 0 for the absence of the disease. Experiments with the Cleveland database have concentrated on simply attempting to distinguish presence (values 1, 2, 3, 4) from absence (value 0). The dataset contains 303 rows of people data with 13 continuous observations of different symptoms.

Features of this dataset are briefly described below: -

* **Age**: the age of the individual in years.
* **Sex**: the gender of the individual using the following format:-

1= male, 0 = female

* **Chest-pain type**: the type of chest-pain experienced by the individual using the following format:-

1 = typical angina, 2 = atypical angina, 3 = non-anginal pain, 4 = asymptotic

* **Blood Pressure**: the resting blood pressure value of an individual in mmHg (unit).
* **Serum Cholesterol**: the serum cholesterol in mg/dl (unit)
* **Fasting Blood Sugar**: compares the fasting blood sugar value of an individual with 120mg/dl. If fasting blood sugar > 120mg/dl then:-

1 (true) else: 0 (false)

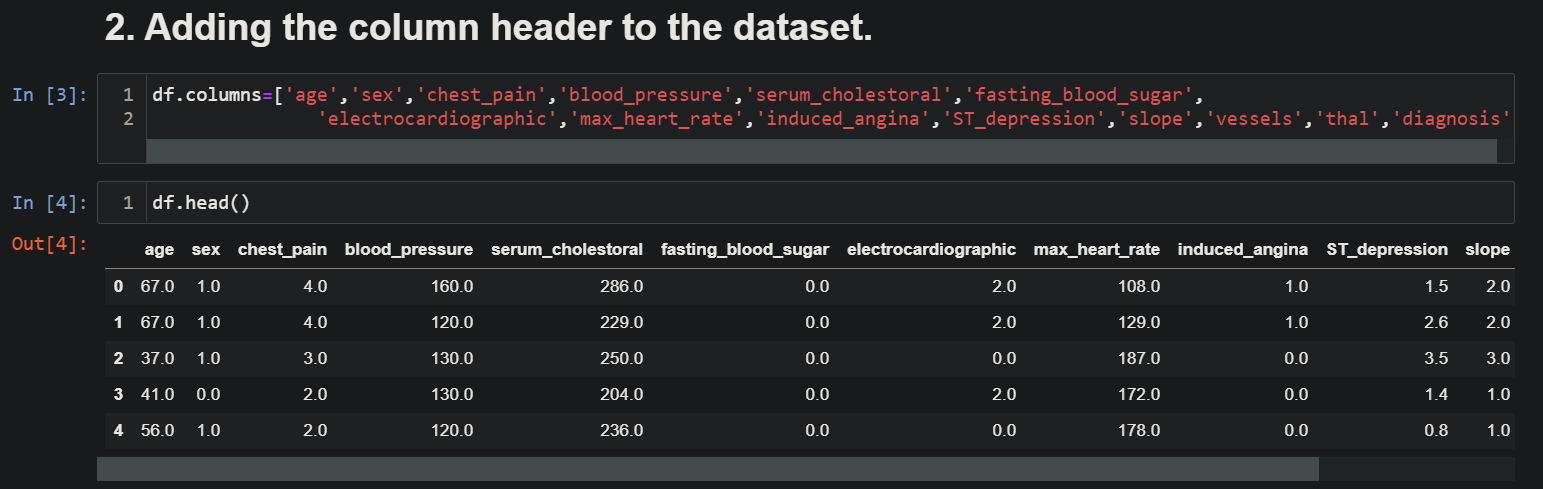
* **Resting ECG***:* displays resting electrocardiographic results in the following format:-

0 = normal, 1 = having ST-T wave abnormality, 2 = left ventricular hypertrophy

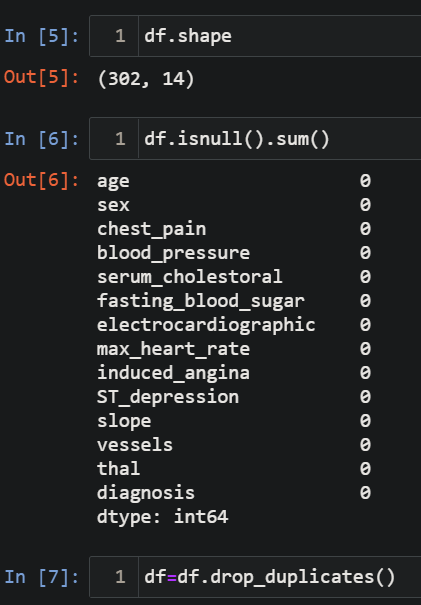
* **Max heart rate achieved:** displays the max heart rate achieved by an individual.
* **Exercise induced angina**: 1 = yes, 0 = no
* **ST depression induced by exercise relative to rest**: displays the value which is an integer or float.
* **Peak exercise ST segment:** 1 = upsloping, 2 = flat, 3 = down sloping
* **The number of major vessels (0–3) colored by fluoroscopy:** displays the value as integer or float.
* **Thal:** displays the thalassemia: 3 = normal, 6 = fixed defect, 7 = reversible defect
* **Diagnosis of heart disease**: Displays whether the individual is suffering from heart disease or not:

0 = absence 1, 2, 3, 4 = present.

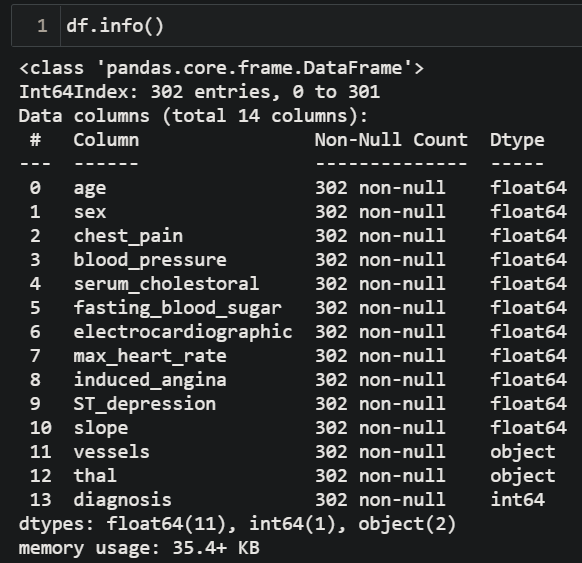
1. Data Understanding & Pre-processing
   1. **Adding the column header to the dataset: -** The Heart Disease data set needed some pre-processing before feeding it into the training model. The first step was to add a column header to the dataset according to its description. The attribute’s values are meaningful to read and interpret. The following image illustrates the statement: -



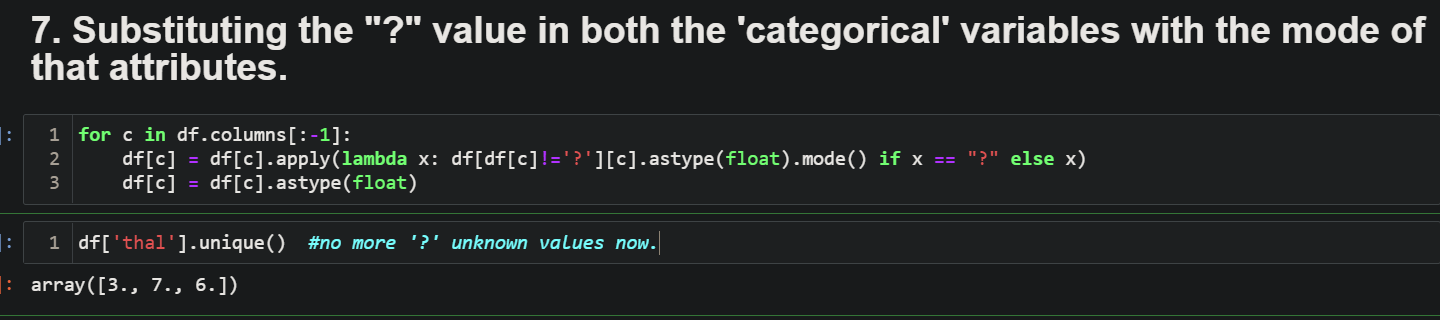
* 1. **Getting some basic information about the dataset: -** To understand the dataset’s shape and whether there are some null values, and whether the dataset has duplicate values or not, the following piece of code illustrates those redundancies: -



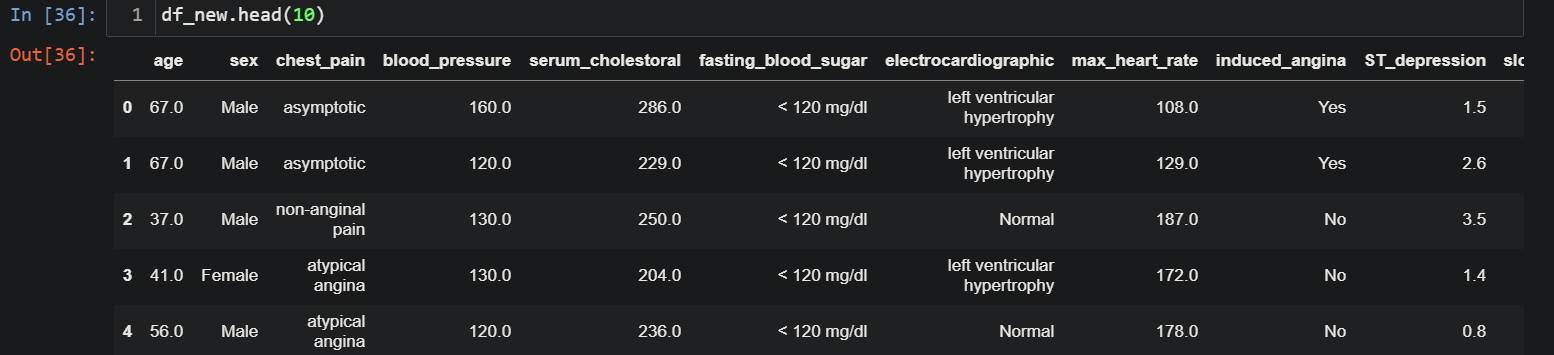
* 1. **Information on the datatypes of attributes: -** The dataset contains a blend of various data types depending on the attributes. In order to find out about it, the following piece of code depicts the data type of each attribute and the memory usage: -



* 1. **Handling the unknown “?” values in the dataset:-** The unknown values present in the dataset needed to be handled before putting the dataset in the training phase. Therefore, as the unknown “?” values were present only in the categorical columns, and after analyzing the values in the column, they were substituted using the following piece of code:-

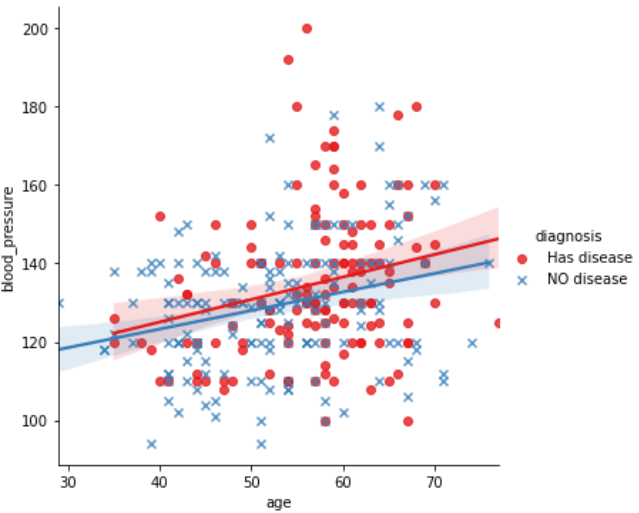


* 1. **Preparing the dataset for EDA and Visualisation: -** The values of each column in the dataset were in the form of numbers. Therefore, to better visualize the dataset and find the relations between attributes, the values are thus converted to a more readable string format for categorical attributes as per the description of the dataset. For example, the dataset values after this step look as below: -

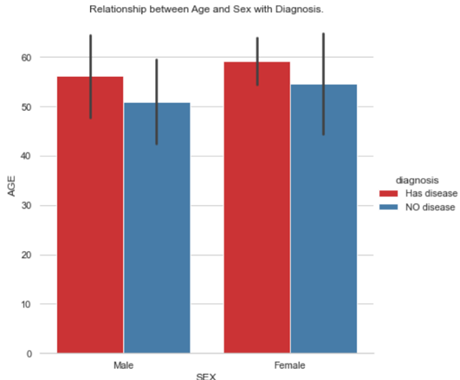


1. EDA with Data Visualisations.

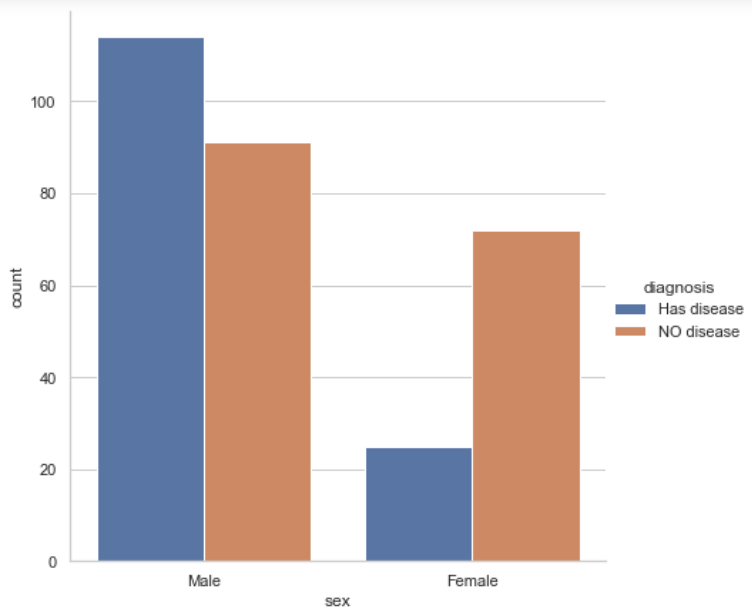
**3.1 Relationship between Age & Blood Pressure: -**

****

We can observe that as age increases, the blood pressure of a person also increases. Still, the blood pressure remains slightly lower than a person who does not have a disease.

**3.2 Relationship between Age & Sex: -** Visualising the trend between Age of a person and Sex of that person, it can be easily observed from the graph below that females with higher age are more likely to have the disease than males. ****

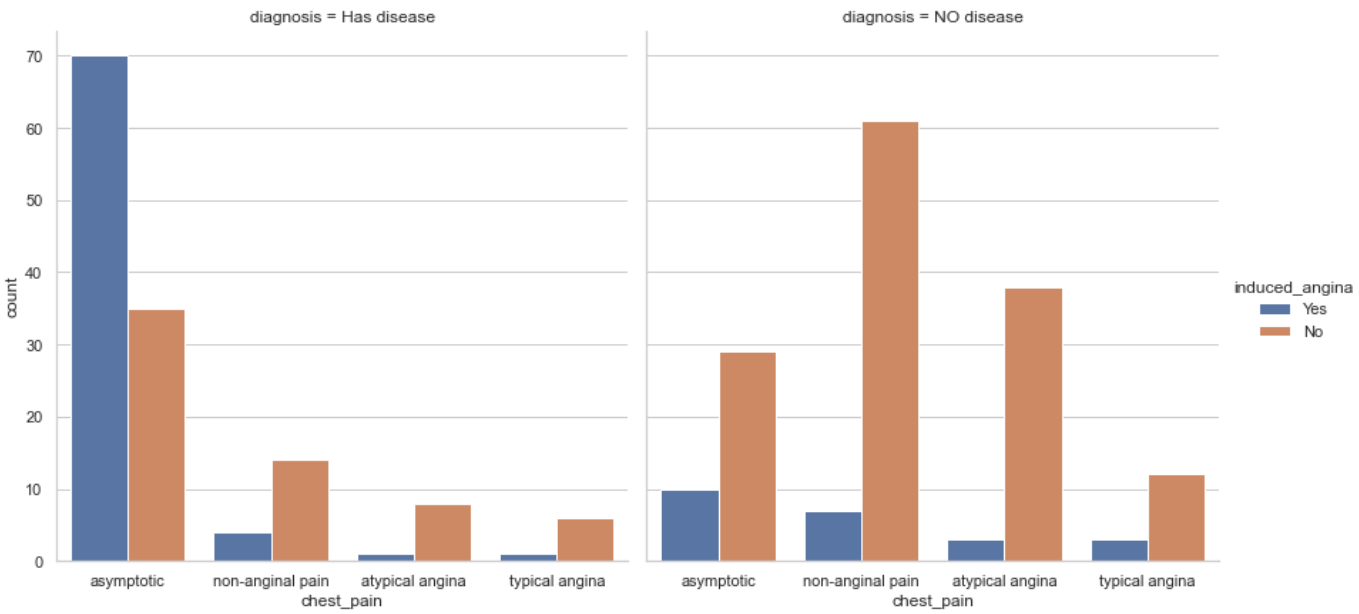
* 1. **Count of people concerning their Sex having a disease or not.**



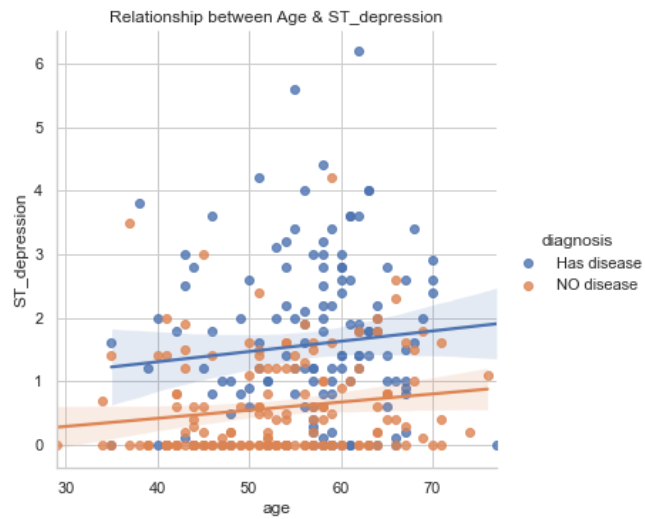
It is interesting to observe that though the females are less in number than males who encounter disease, as observed earlier, the females’ age tends to be on the higher side than men who are having these diseases.

**3.4 Count of people having chest pain complaints along with induced angina and resulting in disease.**

Interestingly, people suffering from induced angina and having asymptotic chest pain and having the disease are significantly more than people who do not have induced angina. Thus, we can conclude that induced angina along with asymptotic chest pain both play an important factor in determining whether the person will have a disease or not.



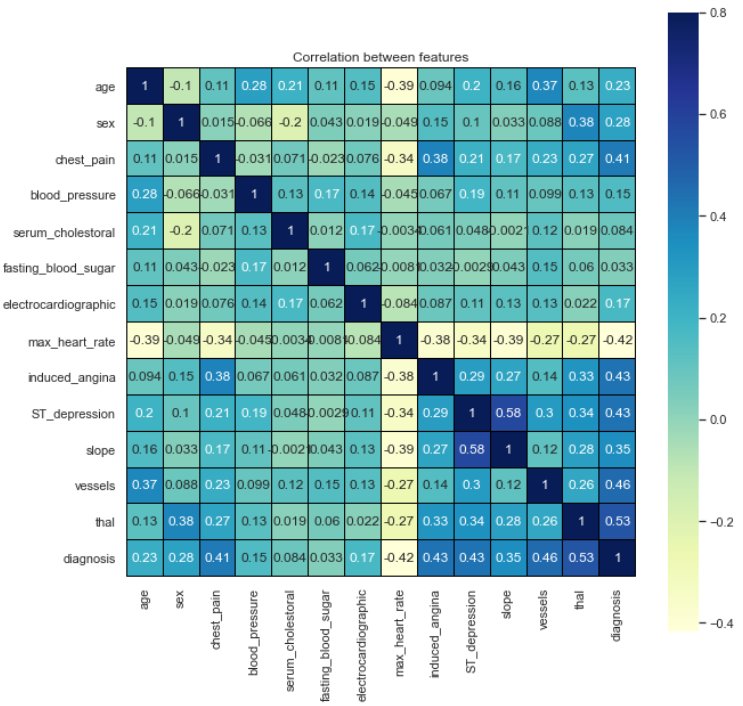
* 1. **Relationship between Age & ST depression with Diagnosis.**

****

We can observe that as the age of person increases, there is a slight increase in ST depression. The blue line indicates that ST depression on the higher side has more probability that person will have heart disease.

* 1. **Heatmap to find the correlations between the attributes.**

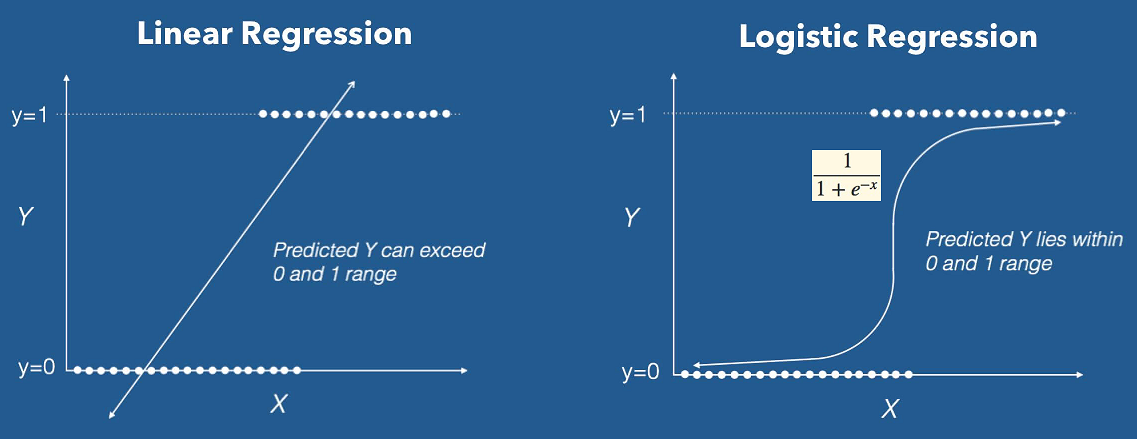
The following Heatmap shows us the correlation of different attributes in the dataset. The columns "chest\_pain", "induced\_angina"," ST\_depression", "vessels" & "thal" are highly correlated to the target attribute "diagnosis" than other attributes of the dataset.

****

1. Model training & Evaluation
   1. **Logistic Regression: -** Logistic Regression is a technique used for solving the classification problem, and classification is nothing but a problem of identifying to which of a set of categories a new observation belongs, based on a training dataset containing observations (or instances) whose categorical membership is known. For example, to predict: Whether an email is spam (1) or not (0) or, Whether the tumor is malignant (1) or not (0).

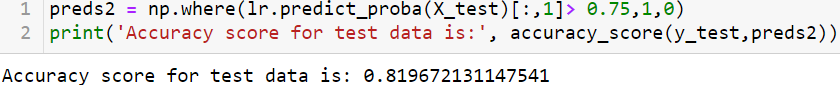
* Sigmoid function or logistic function is used as hypothesis function for Logistic Regression.
* Below is a figure showing the difference between linear Regression and logistic Regression; also, notice that logistic Regression produces a logistic curve, which is limited to values between 0 and 1.
* Logistic Regression outputs the probabilities of a specific class. Those probabilities can be converted into class predictions.
* The logistic function has some nice properties:

1. Takes on an “s” shape
2. The target output falls between 0 and 1.

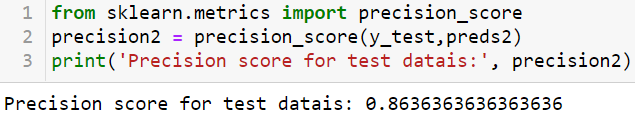


* 1. **Evaluation metrics: -** After training the model, certain evaluation metrics were selected to check the accuracy of the model.

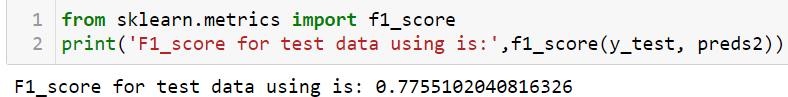
1. **Accuracy Score: -**



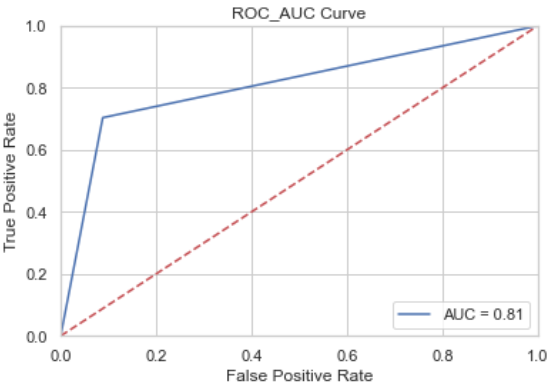
1. **Precision Score: -**

****

1. **F1 Score: -**

****

**4.ROC\_AUC curve: -**

****

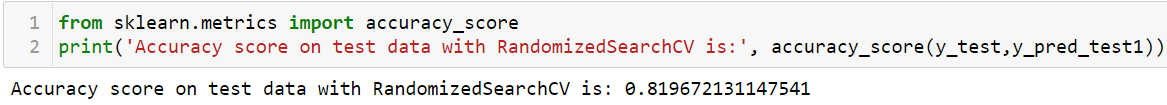
**4.3 Random Forest Classifier: -** In Random Forest, we grow multiple trees instead of a single tree in the CART model. To classify a new object based on attributes, each tree gives a classification, and we say the tree “votes” for that class. The forest chooses the classification having the most votes (over all the trees in the forest). In the case of Regression, it takes the average of outputs by different trees.

Random Forest is a versatile machine learning method capable of performing both regression and classification tasks. It also undertakes dimensional reduction methods, treats missing values, outlier values, and other essential data exploration steps, and does a fairly good job. It is a type of ensemble learning method where a group of weak models combines to form a powerful model.

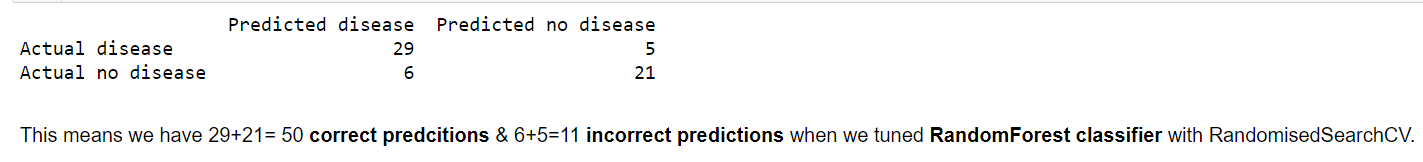
* One of the keys behind the random forest is that each tree trains on random samples of the data points.
* The samples are drawn with *replacement* (known as bootstrapping) which means that some samples will be trained on in a single tree multiple times (we can also disable this behavior if we want).
* The idea is that by training each tree on different samples, although each tree might have high variance with respect to a particular set of the training data, overall, the entire forest \_\_will have \_\_low variance.
* This procedure of training each individual learner on different subsets of the data and then averaging the predictions is known as bagging, short for bootstrap aggregating.

**4.4 Evaluation metrics:-** After training the model now with RandomForest Classifier and using Randomised Search CV for model tuning, certain evaluation metrics were selected to check the accuracy of the model.

**1. Accuracy score:-**

****

**2.Confusion matrix:-**



1. **Conclusion**

The Heart Disease dataset used in this assignment was prepared, cleansed, analyzed, and visualized to get some actionable insights from the data. It was trained on two different algorithms, and parameters were tuned to get better accurate results. Therefore, after thorough analysis, it can be concluded that a person have following traits is most likely to have a disease:-

• Age > 50

• Sex = male

• Chest pain = Asymptotic

• Blood pressure > 112

• Serum cholesterol > 166

• Fasting blood sugar = 0

• Electrocardiographic = 2

• Max heart rate > 114

• ST depression about 2

• Induced angina

• Slope >=2

• Vessels about 1.6

• Thal > 4

1. **References**

• http://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/

• https://archive.ics.uci.edu/ml/datasets/heart+Disease

• https://www.slideshare.net/CharlesVestur/building-a-performing-machine-learning-model-from-a-to-z

• https://scikit-learn.org/stable/modules/generated/sklearn.linear\_model.LogisticRegression.html

• https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

• https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5