**Technical Report - Heart Disease**

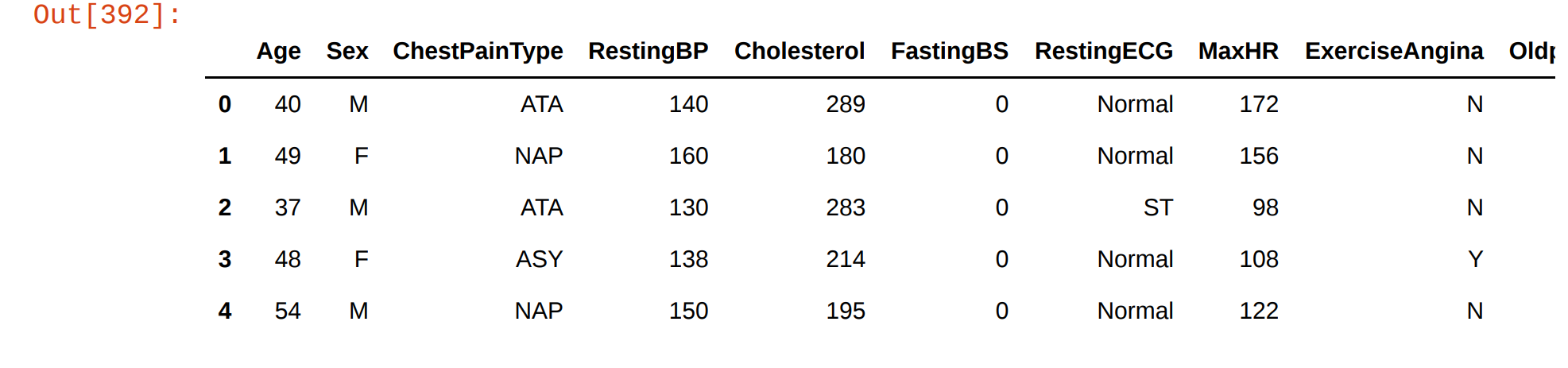
**Problem Statement**

Analyzing the data set (the dataset uploaded to canvas), investigate and evaluate the result and predict the overall performance.

**Solution:**

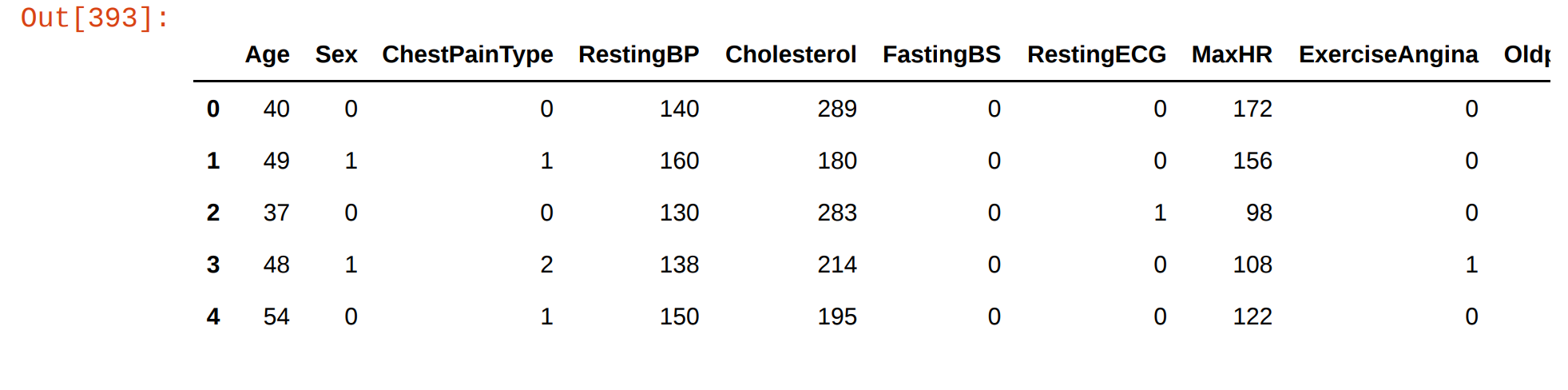
**Steps Involved**:

**Step 1: Loading Dataset –** The sample dataset has been loaded to data frame using **pandas** library.

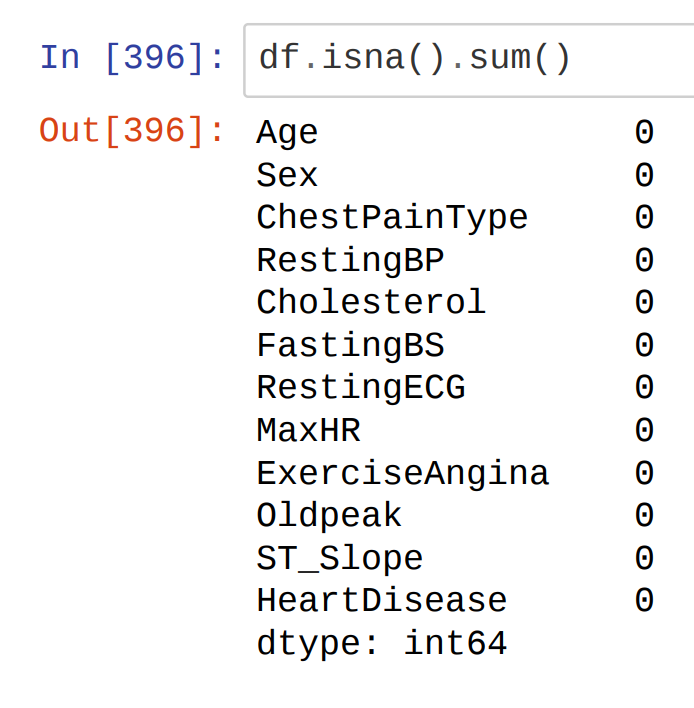
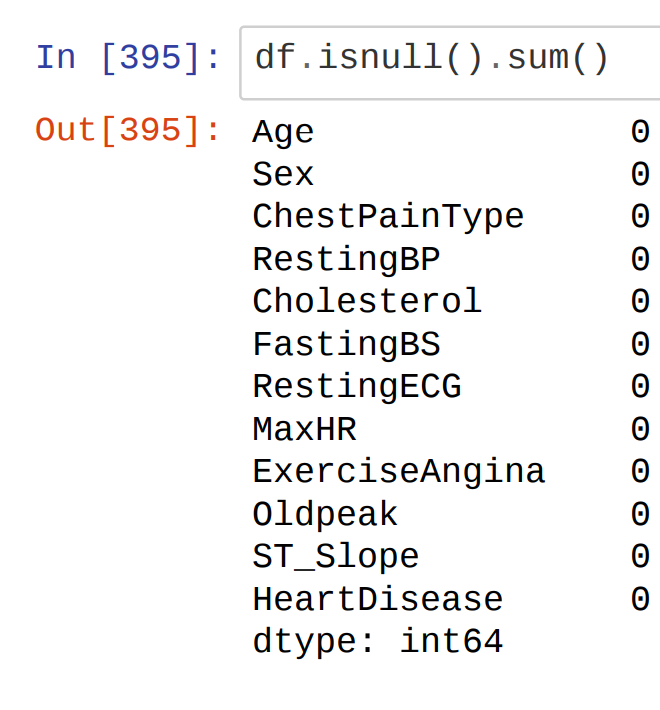


**Step 2: Data Processing –**

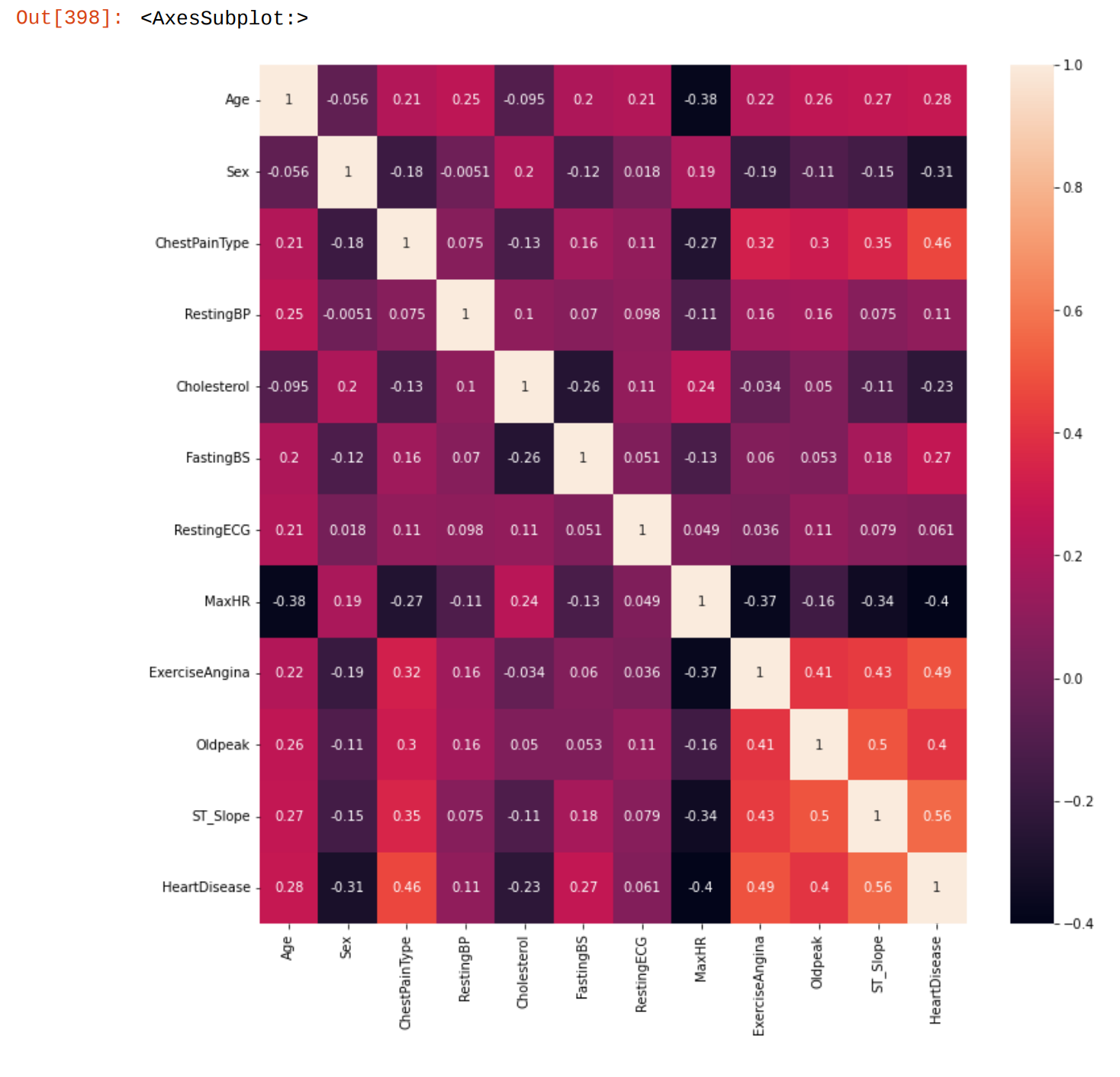
1. Converting categorical attributes to numeric - The main reason to do this is some regression models like Logical Regression needs its Y variable (dependent variable) in the categorical nature. ***In the provided sample data, we have replaced values of the attributes – ChestPainType, RestingECG, ExerciseAngina and ST\_Slope with the categorical values 0,1,2,3..n***



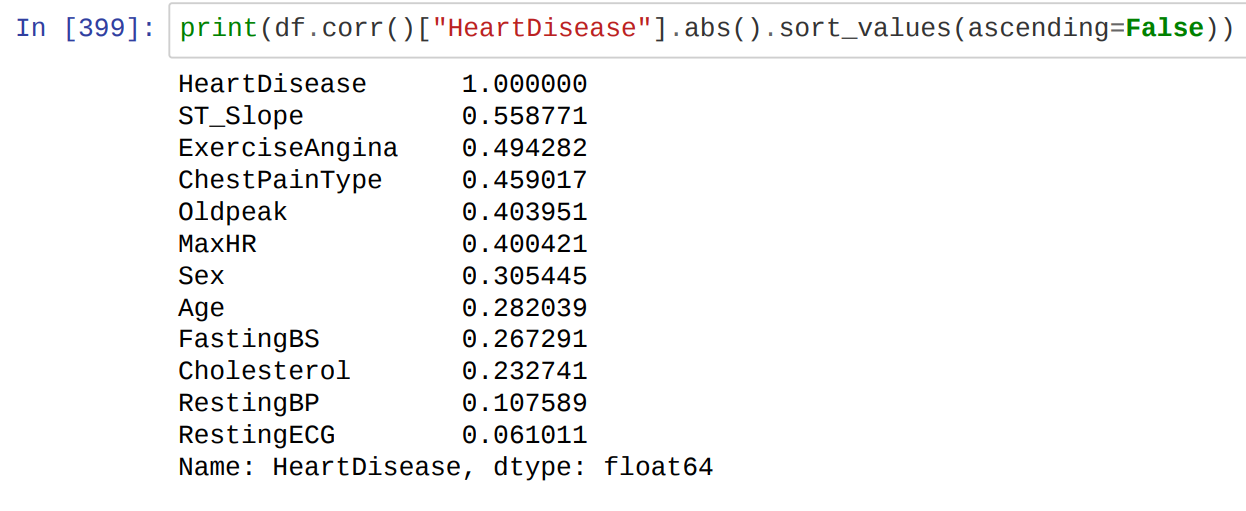
1. Checking for null, missing and non-numeric values. ***There were no such records that had null, missing, non-numeric values in the provided sample data.***



1. Checking the Correlation Between the Attributes – The correlation values greater then **0.7** depicts that the attributes are highly correlated. If in case, the attributes are highly correlated, we can remove either of the attribute. ***In the provided sample data, there were no such values that were greater than 0.7***. ***Thus, none of the attributes were eliminated.***



1. Checking the correlation between dependent attributes with independent attributes – Here, HeartDisease is the dependent attribute. Below is the correlation values for reference.



**Step 3: Data cleaning –** Having clean data will ultimately increase overall accuracy of the model and allow for the highest quality information in the decision-making. Benefits include removal of errors when multiple sources of data are at play.

***In our sample data, the data values that are outside the below mentioned ranges are cleaned for the respective attributes. Replacement was done with the mean value of the range of the sample data.***

|  |  |
| --- | --- |
| ATTRIBUTES | RANGE |
| Age | 0 to 100 |
| RestingBP | 60 to 180 |
| Cholesterol | 40 to 500 |
| MaxHR | 0 to 200 |
| Oldpeak | -2.55 to 4.2 |

**Formula:**

**Mean = total / n**

where,

total = summation of all the values of the respective attribute excluding the noise / values out of the range

n = total no of records

**Step 4: Exploratory Data Analysis -** Performing initial investigations on the sample data so as to discover patterns, to spot anomalies, and to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

|  |  |
| --- | --- |
| 1. Analyzing Heart Disease variable – *From the below plot we can conclude that the given sample data has* ***greater rate of heart disease.*** |  |
| 1. Analyzing Sex variable wt HeartDiease - *From the plot we can conclude that* ***"Male" (0) in the given dataset has greater heart disease than "Female" (1)*** |  |
| 1. Analyzing ChestPainType variable wrt HeartDisease - *From the plot we can conclude that patient with* ***ChestPainType "ASY" (2)*** *in the given dataset* ***has greater heart disease than remaining ChestPainType.*** |  |
| 1. Analysing RestingBP variable wrt HeartDisease - *From the plot we can conclude in the given dataset there is* ***higher distribution of people******with BP level between 130-140 which is higher and correlated to heart disease.*** |  |
| 1. Analyzing FastingBS variable wrt HeartDisease *- From the plot we can conclude that patient with FastingBS (1) in the given dataset has* ***greater heart disease than without FastingBS.*** |  |
| 1. Analyzing RestingECG variable wrt HeartDisease *- From the plot we can conclude that patient with RestingECG "ST" (1) in the given dataset has* ***greater heart disease than remaining RestingECG.*** |  |
| 1. Analyzing MaxHR variable wrt HeartDisease - *From the plot we can conclude in the given dataset there is* ***higher distribution of people with MaxHR level between 110-130 which is higher and correlated to heart disease*.** |  |
| 1. Analyzing ExerciseAngina variable wrt HeartDisease - *From the plot we can conclude that* ***patient with ExerciseAngina in the given dataset has greater heart disease than without ExerciseAngina.*** |  |
| 1. Analyzing Oldpeak variable wrt HeartDisease *- From the plot we can conclude in the given dataset there is* ***higher distribution of people with Oldpeak level between 0.7 to 1.1 which is lower and negatively correlate to heart disease.*** |  |
| 1. Analysing ST\_Slope variable wrt HeartDisease - *From the plot we can conclude that patient with ST\_Slope "Flat" (1) in the given dataset has* ***greater heart disease than remaining ST\_Slope.*** |  |

1. Train Test split - Splitting the data into training and test data.

**Step 5 -** Training the model or model fitting.

**Different Algorithms that can be used:**

1. **Logistic Regression:** Logistic regression estimates the probability of an event occurring, such as voted or didn't vote, based on a given dataset of independent variables. Since the outcome is a probability, the dependent variable is bounded between 0 and 1.
2. **Naïve Bayes:** Naive Bayes uses a similar method to predict the probability of different class based on various attributes. This algorithm is mostly used in text classification and with problems having multiple classes.
3. **K Nearest -** The k-nearest neighbors’ algorithm, also known as KNN or k-NN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point.
4. **Decision Tree -** Decision tree is a tree based algorithm used to solve regression and classification problems. An inverted tree is framed which is branched off from a homogeneous probability distributed root node, to highly heterogeneous leaf nodes, for deriving the output. Regression trees are used for dependent variable with continuous values and classification trees are used for dependent variable with discrete values.

***For Model Fitting, below are the algorithms with their respective accuracy % computed on the given sample data.***

|  |  |
| --- | --- |
| Algorithm | Accuracy % |
| Logistic Regression | 80.98 % |
| Naive Bayes | 81.52 % |
| K Nearest Neighbors | 65.76 % |
| Decision Tree | 78.8 % |

**From the accuracy we conclude that Naive Bayes is the best fit to our model. Other algorithms can be used as alternatives for refinement opportunity.**