****

**PROJECT WRITE-UP – HEART ATTACK STUDY**

**Submitted on : 02-04-2023**

**Submitted by : Mani Sankar Baskar**

**PROJECT WRITE-UP**

**INTRODUCTION :**

The project that I chose to work upon is “**Factors that play a significant role in increasing the rate of heart attacks”.** The dataset contains 14 columns which are factors that are associated with Heart disease. I am going to do a detailed study on whether those factors influence Heart disease, distribution of patients, use of Statistical tools for inferences, Training the data, testing the data, visualizing the data and so on.

The reason I chose this project is that the data belongs to Healthcare domain for which I absolutely have no clue about. My intention was to gather knowledge on some common terminologies associated with the medical field for Heart diseases and also to have exposure to a data that is completely of numerical type so that I can perform operations like grouping the data, classifying the data and much more.

**TOOLS USED :**

1. Python programming and the coding environment I used is Jupyter labs.
2. Tableau for Visualization

**BUSINESS REQUIREMENT:**

The following are the project tasks:

* Preliminary data inspection - structure of the data, missing values, duplicates, etc
* Removing duplicates, treat missing values if any.
* Perform statistical summary of the data.
* Identifying categorical variables and describing them.
* Occurrence of CVD across Age
* Composition of overall patients with regards to Gender.
* Identifying anomalies in Resting Blood Pressure of the patient and if it has an impact on Heart disease.
* Finding relationship between Cholesterol levels and our target variable
* Study on relationship between peak exercising and the occurrence of a heart attack.
* To check if thalassemia a major cause of CVD
* Other factors that play a role in occurrence of CVD
* Identifying the relationship between variables.

**OBJECTIVES:**

\*To derive the data and visualize it so as to make data more readable.

\*To get better insights on relationship between the variables and its impact.

\*To transform the data for ML models and present the data visualizations and app predictions the most efficient way possible.

\*To analyze the accuracy of predictions between actual and predicted results.

\*To use the model to train similar datasets in future.

\*To support the business to make important decisions based on predictions.

**EXECUTION USING PYTHON PROGRAMMING:**

**MY APPROACH:**

\*Importing Data

\*Data Preprocessing:

a.) Identifying the structure of Data

b.) Finding missing Values

c.) Finding duplicate entries

### d.) Fixing items according to variable description for the dataset

### e.) Checking the distribution of data for Independent features

### f.) Checking the distribution of data for target variable

### g.) Statistical summary of the data

\*Exploratory Analysis

### a.) Identifying the categorical variables and describing them

### b.) Visualising the categorical variables

### c.) Study of variables, relationships between them and their impact on Heart attack.

**\***Test train Split

\*Model building using Logistic Regression

\*Prediction on Train data

\*Prediction on Test data

\*Model Evaluation

a.) Building a confusion matrix

b.) Accuracy score

**SUMMARY :**

* Original dataset had 303 entries with 14 columns. After Preprocessing, our transformed dataset is 296 entries with 14 columns.
* There are more number of Heart disease patients (160) compared to Patients without Heart disease (136).
* Half of the patients in the dataset are below the age – 56
* Average resting blood pressure(in mm Hg) of patients on admission to Hospital is 131.6
* Cholesterol level ranges from 126 till 564. 564 is such an alarming rate when it comes to cholesterol levels.
* Dataset takes into consideration the CVD across youths, matured adults and elderly people. The data points contain minimum age as 29 and maximum age as 77 for this analysis.
* No of female patients are close to 100 and there are more number of Male patients - i.e close to 200.
* Approximately 140 observations in the dataset are the patients with chest pain type as 0.
* Very less number of patients have fasting blood sugar > 120 mg/dl.
* Almost equal distribution of data points for subcategories in restecg(especially 0 and 1).
* Around 80 to 90 patients experience exercise induced angina - i.e discomfort in the chest. This contributes to 30 percent of the patients in the entire dataset.
* There are close to 140 patients each with heart activity under Slope 1 and Slope 2 conditions
* Close to 175 patients doesn’t have any issues with coronary arteries which means that more than 50 percent of observations doesn’t have issues with coronary arteries.
* Around approximately 60 observations have Score 1 - i.e at least 1 coronary artery being affected.
* Thalassemia type 3 is normal and there are less than 20 patients in this category.
* Thalassemia type 6 is more frequent and Fixed defect and there are almost 160 patients who fall under this category.
* Thalassemia type 7 is generally rare but we can see 100 observations under this category.
* It seems that most of the patients with CVD is between the age groups 40 to 60 as the distribution is higher.
* Resting blood pressure is not helping much in diagnosis of heart disease.
* Overall, there are more number of Male patients compared to Female patients for diagnosing the heart disease.
* Males are vulnerable to heart disease. However, the count of male patients without Heart disease is also significantly high.
* Females, on the other hand appears to be more vulnerable to heart disease as the number of female patients without heart disease is comparatively low.
* We have negative correlation between cholesterol and target variable. What this means is that variable ("chol") has no relationship or perhaps negative relationship with variable (target) which points to heart disease.
* Peak exercise refers to the maximum level of physical exertion that a person can sustain during exercise. Here the variable "oldpeak" represents it. Variable "oldpeak" has negative correlation with target variable. This means that the variables are not explaining each other.
* Thalassemia in this case refers to the variable "thal". Since the correlation value is -0.17, there is a negative correlation for it with the variable "target".
* Variable "thalach" has a positive correlation with the variable "target". A positive correlation means that as the values of one variable increase, the values of the other variable tend to increase as well. There may or may not be a linear relationship based on the strength of the correlation that exist between variables.
* There is a weak correlation and not perfectly linear when it comes to relationship between age and resting blood pressure.
* Most of the variables that we see carries a very weak relationship / non-linear relationship with other continuous variables.

**MODEL EVALUATION RESULTS :**

Our Logistic regression model takes into consideration, the Actual vs Predicted results to evaluate the model performance. We used Confusion matrix for our model evaluation.

1.) True positives (TP): The number of instances that are correctly predicted as positive (e.g., the model correctly identifies a disease – the count is 45).

2.) True negatives (TN): The number of instances that are correctly predicted as negative (e.g., the model correctly identifies a non-disease – the count is 34).

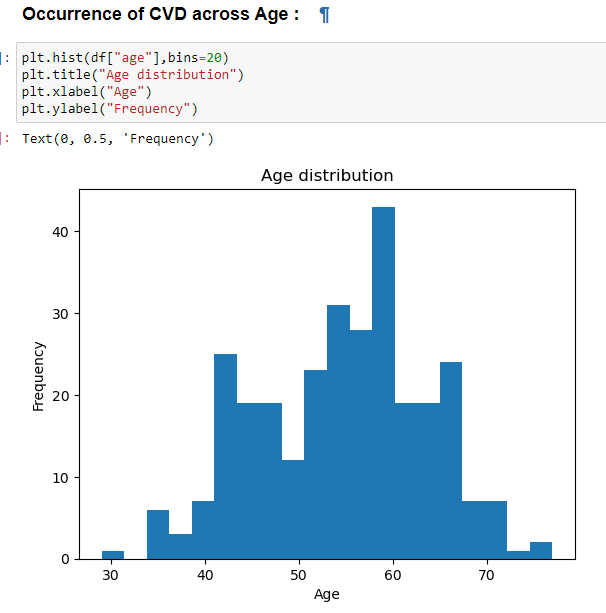
3.) False positives (FP): The number of instances that are incorrectly predicted as positive (e.g., the model predicts a disease when there is none – the count is 7).

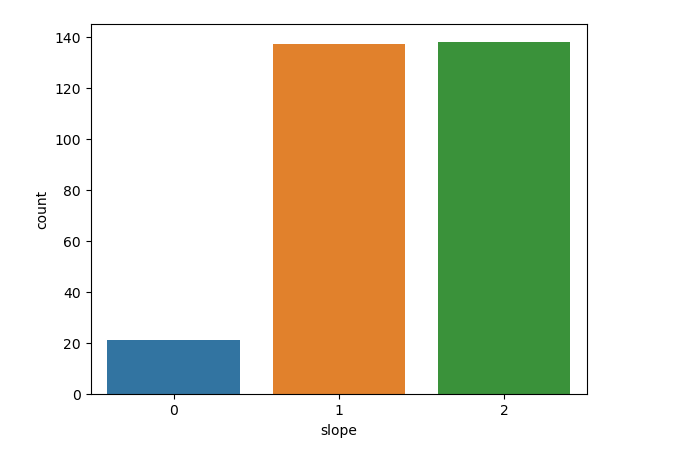
4.) False negatives (FN): The number of instances that are incorrectly predicted as negative (e.g., the model predicts no disease when there is one – the count is 3).

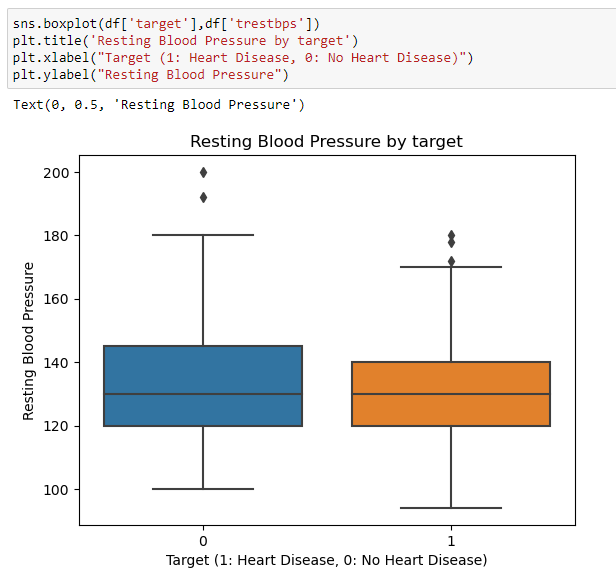
**MODEL ACCURACY :**

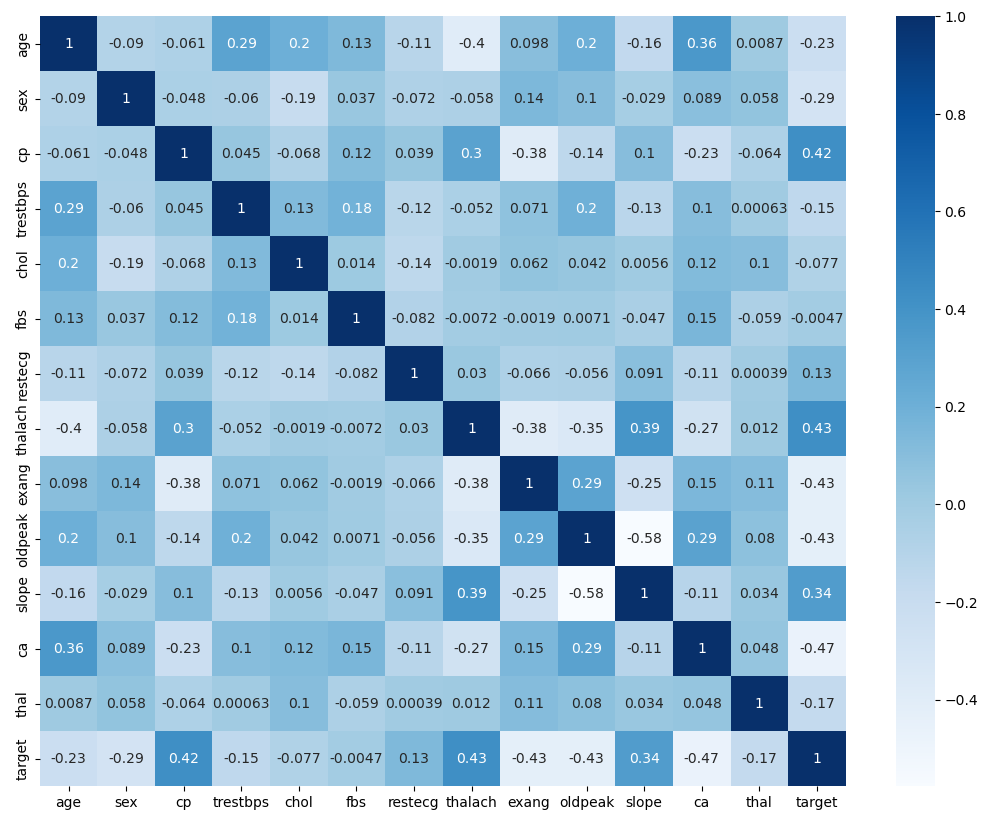
The Accuracy for our Logistic regression is 88.7 percentage. This means – Our model was able to predict 88.7 percentage of entries correctly. Hence, I can conclude, that the model is a very good model to be used for similar datasets in future.

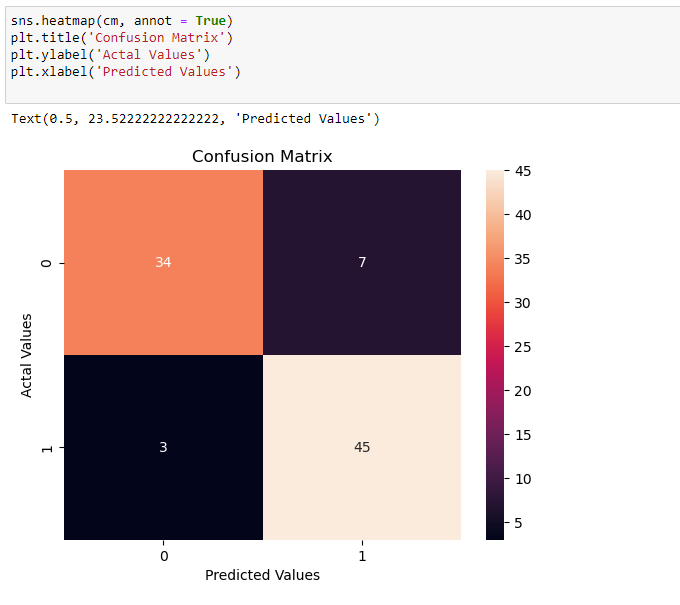
**SCREENSHOTS OF VISUALIZATIONS :**











**TABLEAU VISUALIZATION:**

**PROJECT LINK :**

https://public.tableau.com/app/profile/mani.shankar1721/viz/Capstoneproject-Heartdisease/Dashboard1

**VISUALISATIONS ACROSS VARIABLES :**

\* Bar graph to visualize the patient’s distribution based on three factors – Age, Sex, Heart disease.

\* Pie chart to explain Chest pain type in percentage terms for both patients with and without heart disease.

\* Bar graph to visualize the patient counts with different cholesterol levels – for patients with and without heart disease.

\* Tree map to explain the weightage of patients across different Blood pressure levels – for both patients with and without heart disease.

\* Plotted bar graphs to study the Thalassemia type for patients – with and without heart disease.

\* Scatterplot matrix to identify the relationship between variables – for patients with and without heart disease.

**INSIGHTS :**

* 1. There are more number of Male patients across patients with Heart disease and without Heart disease.
  2. Most of the patients admitted to the hospital belong to the Age group 41-60.
  3. For Patients without Heart disease – Almost 34.4 percentage of patients have Type 0 for Chest Pain
  4. For Patients with Heart disease – Almost 21.96 percentage of patients have Type 2 Chest Pain. It appears that majority of the contributions for Heart disease patients are from Type 1 and Type 2.
  5. For Patients with and without Heart Disease – Majority of them are with Cholesterol levels between 201 to 300. Hence, it is difficult to determine / conclude that Cholesterol levels as a reason for Heart attacks.
  6. Resting Blood pressure for patients with and without Heart disease doesn’t create an impact / plays a vital role in predicting Heart attacks. However, the rate of Blood pressure between ranges 100 to 140 is considerably higher for patients with heart disease.
  7. Counts of Patients with fixed defect is significantly higher for those with heart disease than the ones without heart disease. Patients with fixed defect usually have the red blood cells always be smaller and less able to carry oxygen. It appears that Thalassemia could be a contributor to heart disease. However, there could also be other factors associated with heart disease.
  8. While analyzing the relationship between the variables using Scatterplot matrix, it is evident that there is no / very weak relationship between the variables.

**POINTS TO NOTE IN TABLEAU DASHBOARD:**

* Male vs Female - This filter is applied to the entire Dashboard. User will be able to see the visualisation across all the plots based on this selection.
* All the other filters are added at Sheet level and not applied to all the worksheets using the data source.
* Header, Background fill and Tile adjustment done in Dashboard
* Imported image to justify the dataset on Heart disease study.
* Automatic Layout setting and hence the visualisation will auto-adjust its resolution based on user’s device.
* Navigation button for Scatterplot matrix to shift to that worksheet.
* Download button enabled so that the user can even download it in a desired format.

**SCREENSHOTS OF VISUALIZATIONS:**

