|  |
| --- |
|  |
| Heart Disease Prediction |
| Statement of Work-V1 |

|  |
| --- |
| Bhavna Panwar  100802837 |

**Table of Contents:**

* Introduction………………………………………………………………………………………………………………………3
* Scope of Work……………………………………………………………………………………………………………………3
* Deliverables……………………………………………………………………………………………………………………….3
* EDA……………………………………………………………………………………………………………………………………3
* Statistical Modelling………………….………………………………………………………………………………………9
* Milestone………………………………………………………………………………………………………………………….13
* Gantt Chart……………………………………………………………………………………………………………………….14
* Data Source……………………………………………………………………………………………………………………….14
* Dataset Information………………………………………………………………………………………………………….14
* Dataset Assumptions…………………………………………………………………………………………………………14
* Steps to perform Classification………………………………………………………………………………………….14
* Testing Process………………………………………………………………………………………………………………….15
* Acceptance……………………………………………………………………………………………………………………….15

**Introduction:**

*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 "target" field refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 1(presence).*

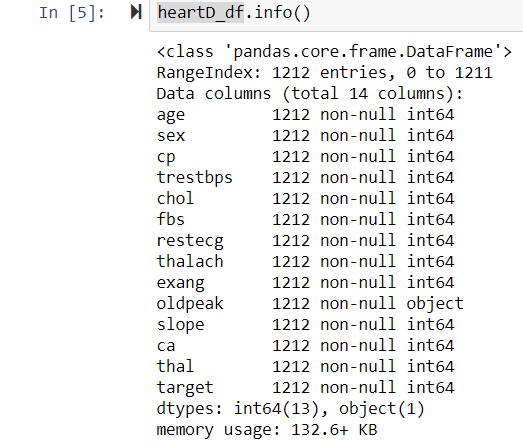
**Scope of Work:**

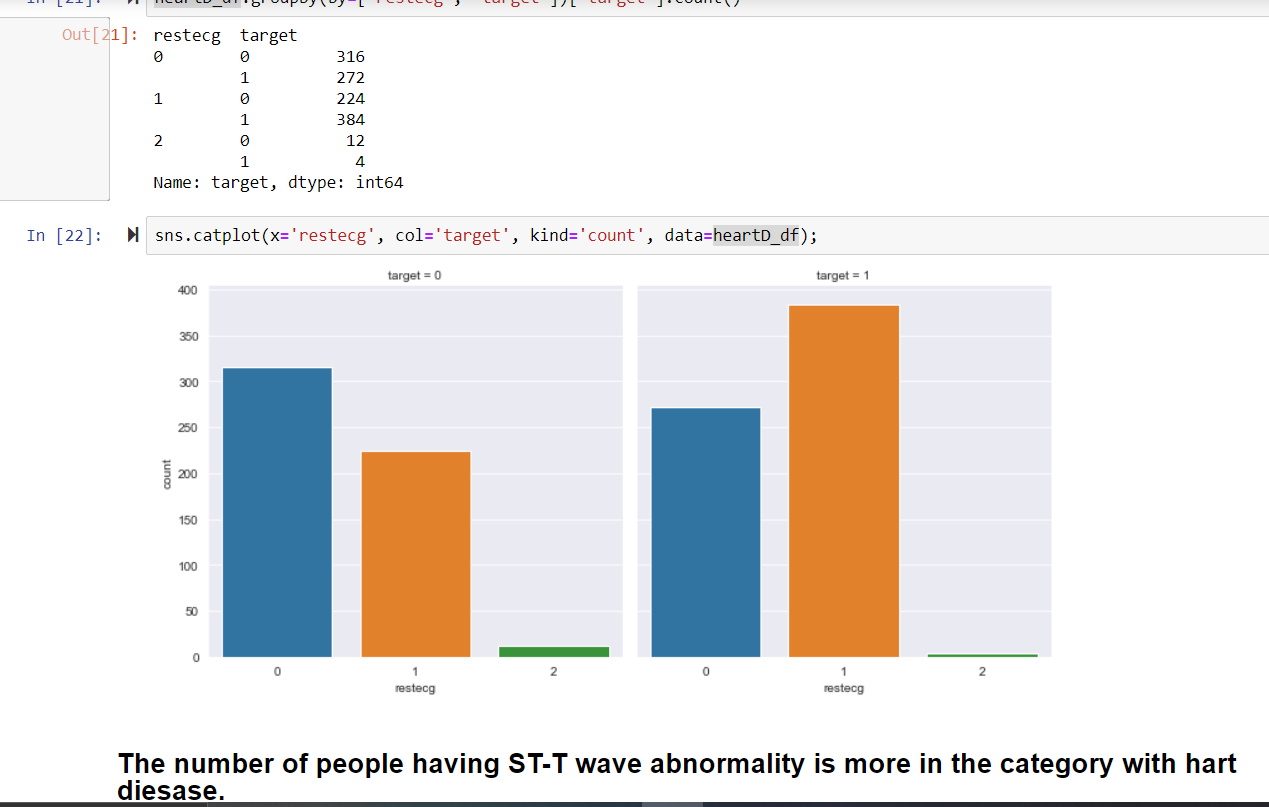
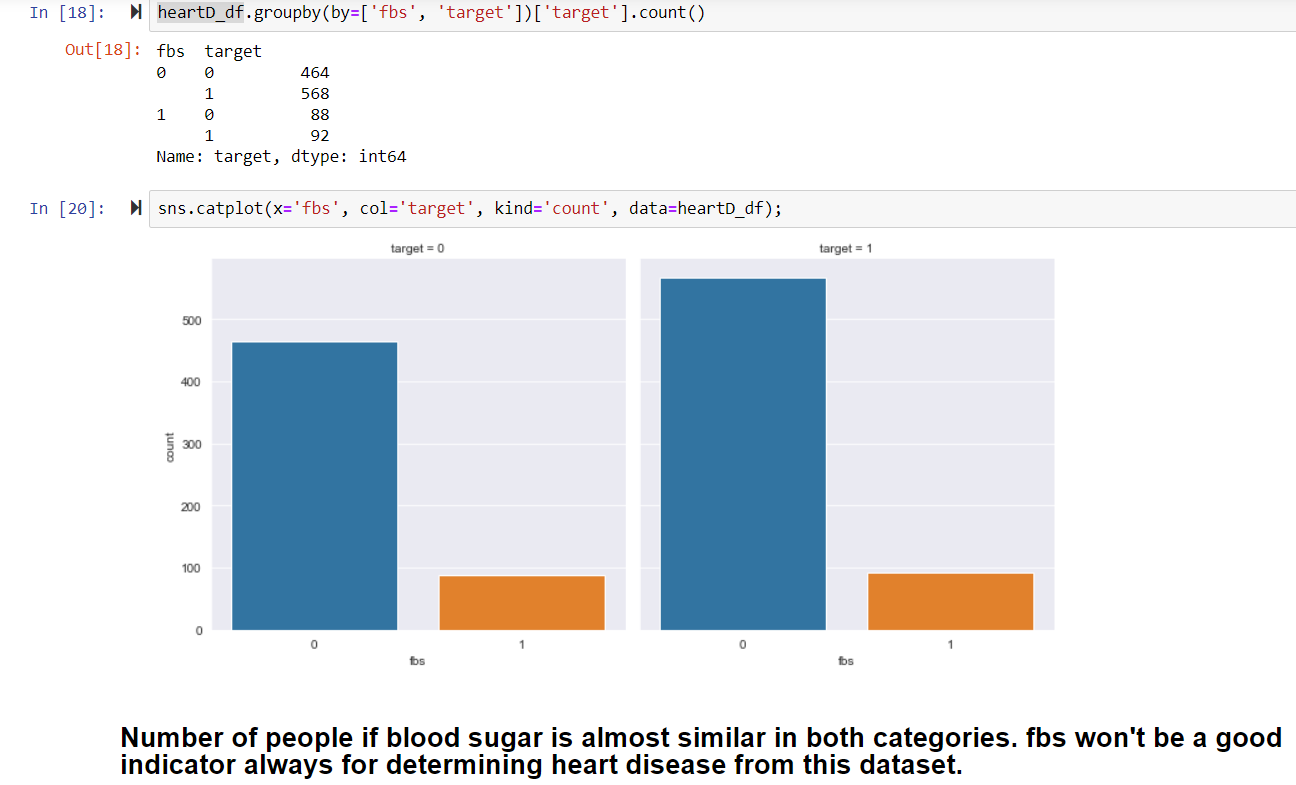
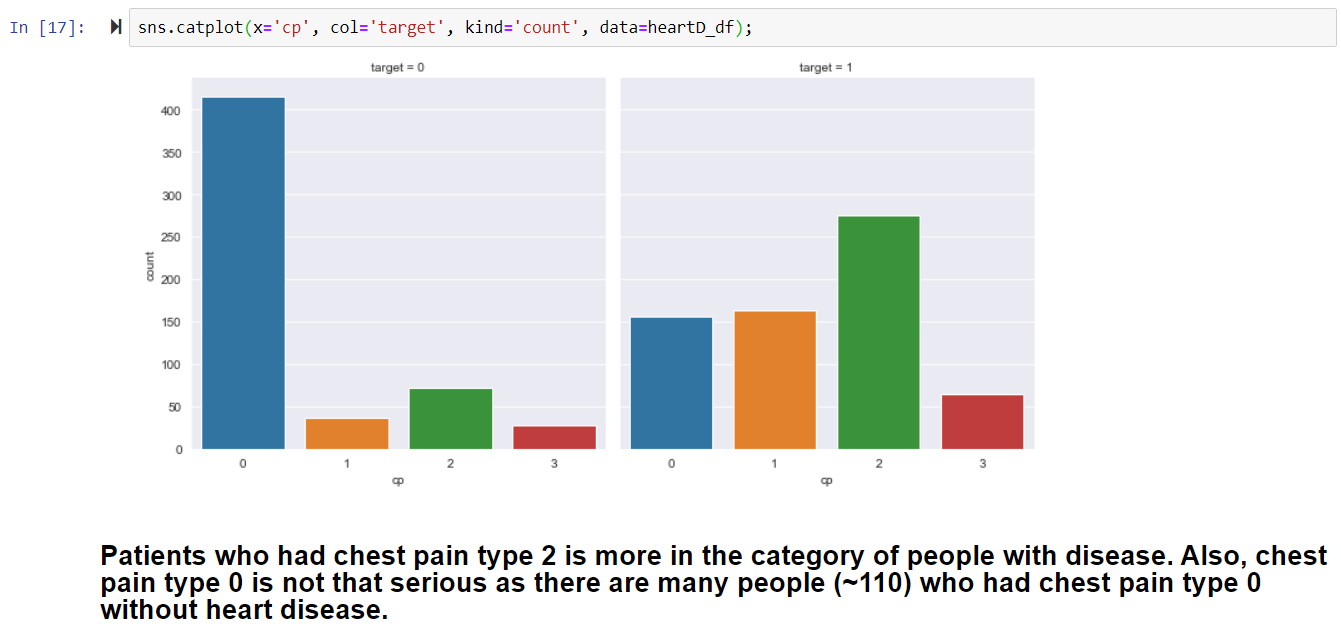
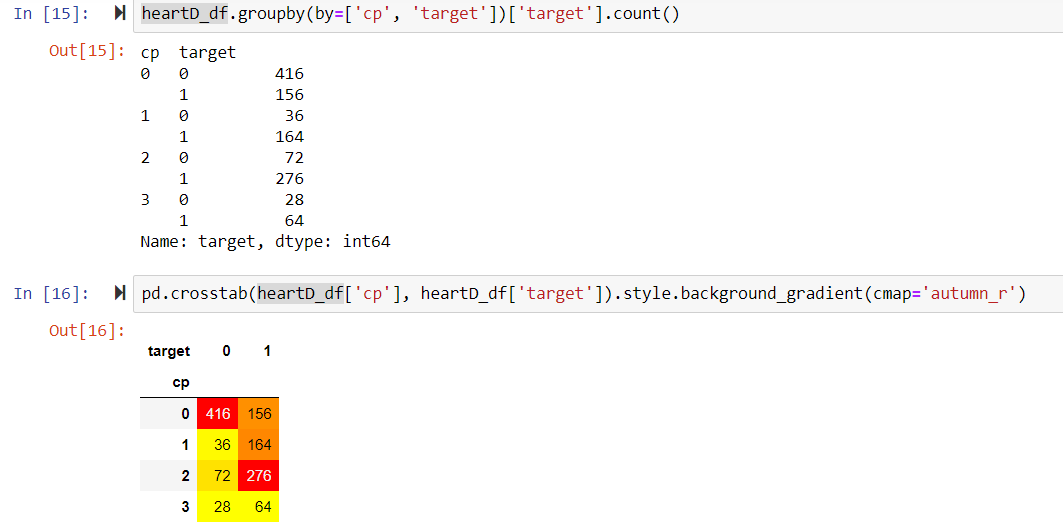
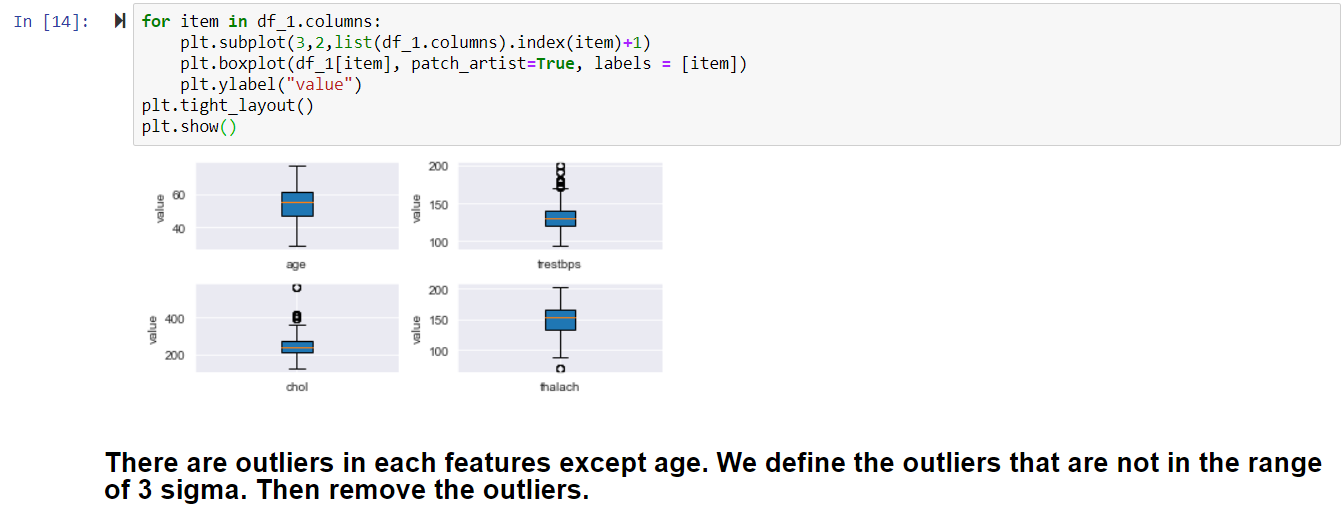
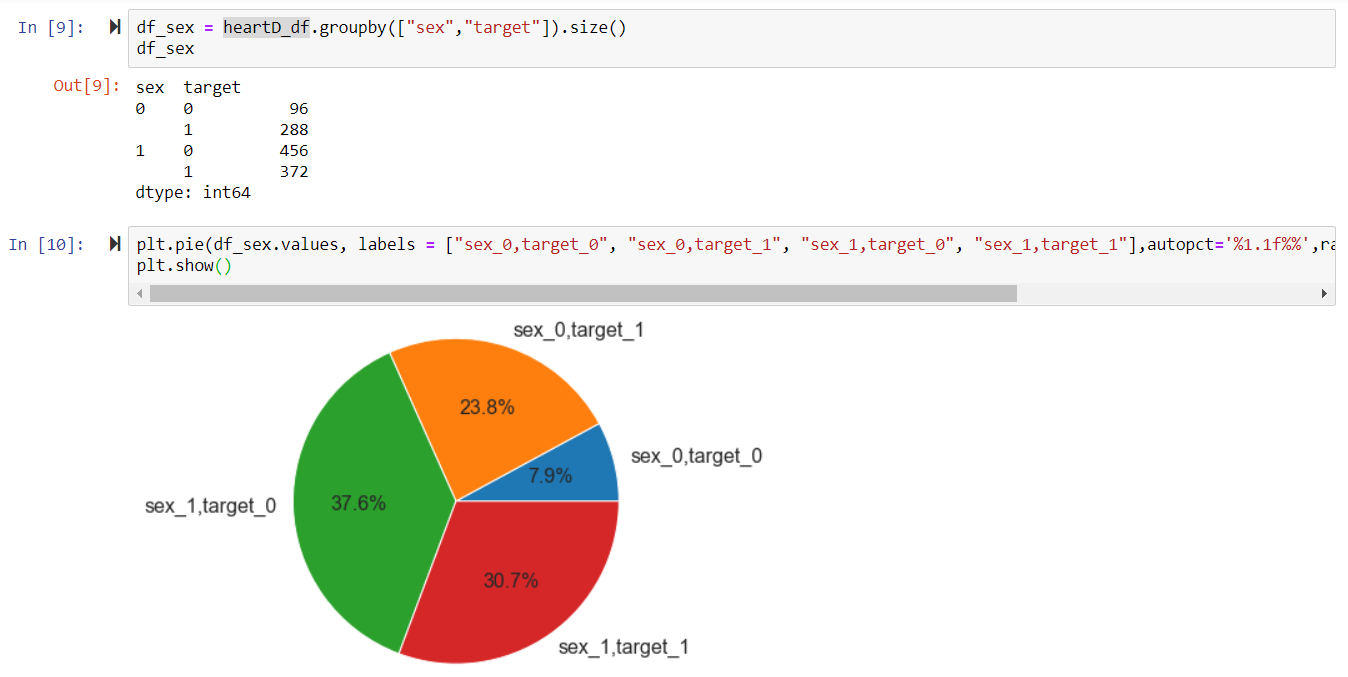
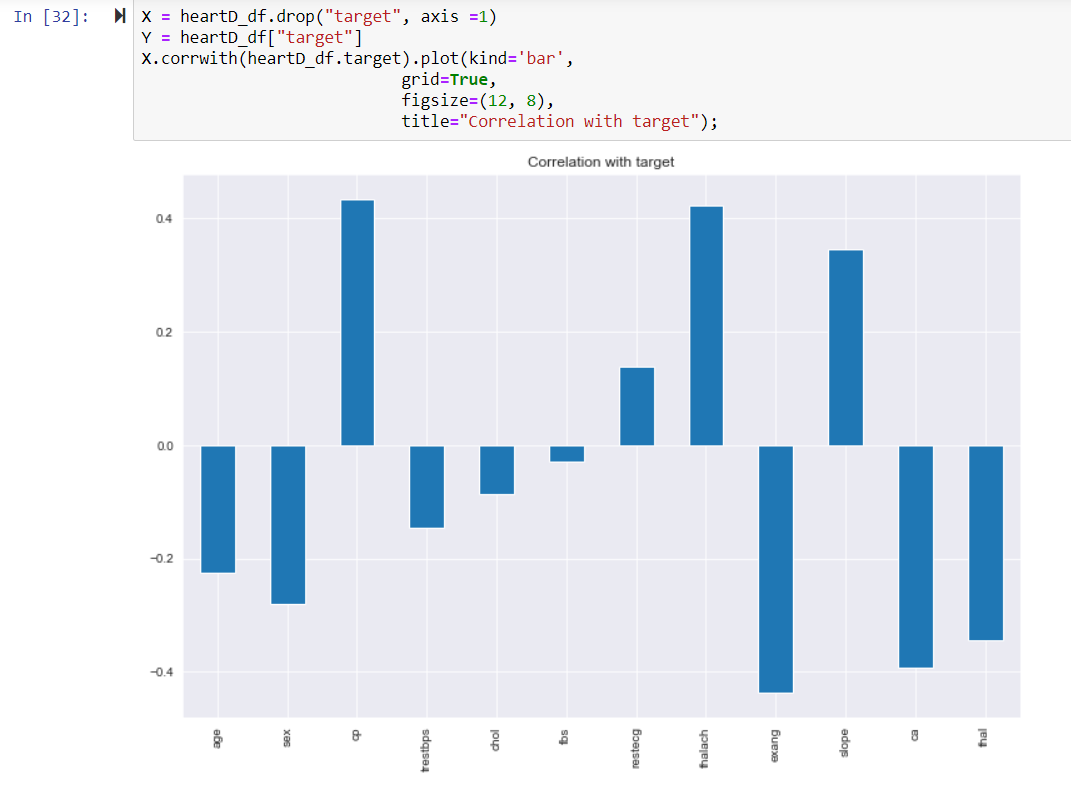
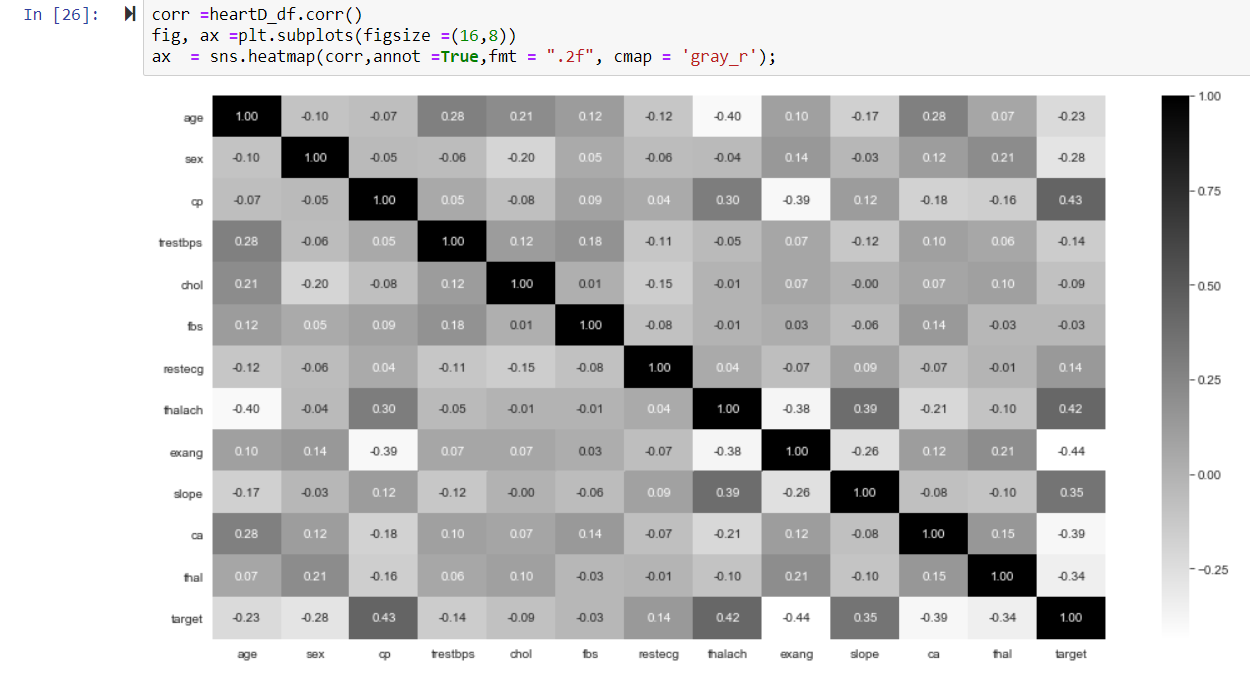
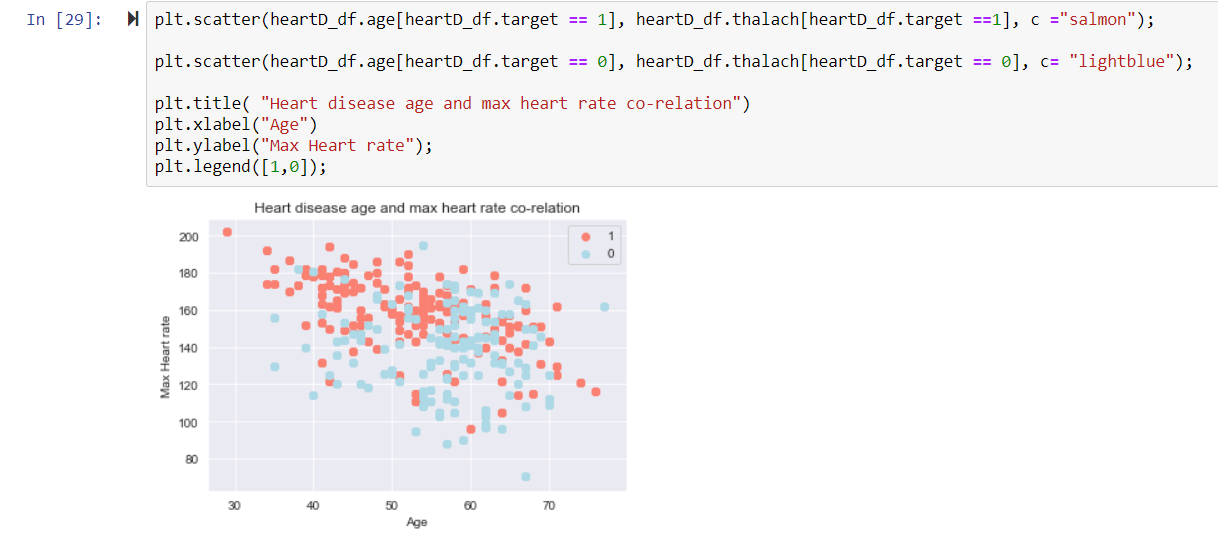
The objective of this project is to understand and analyze the factors that influence the heart disease. By using different machine learning regression models and comparing the results, we are going to achieve the best model which is more accurate and effective.

**Deliverables:**

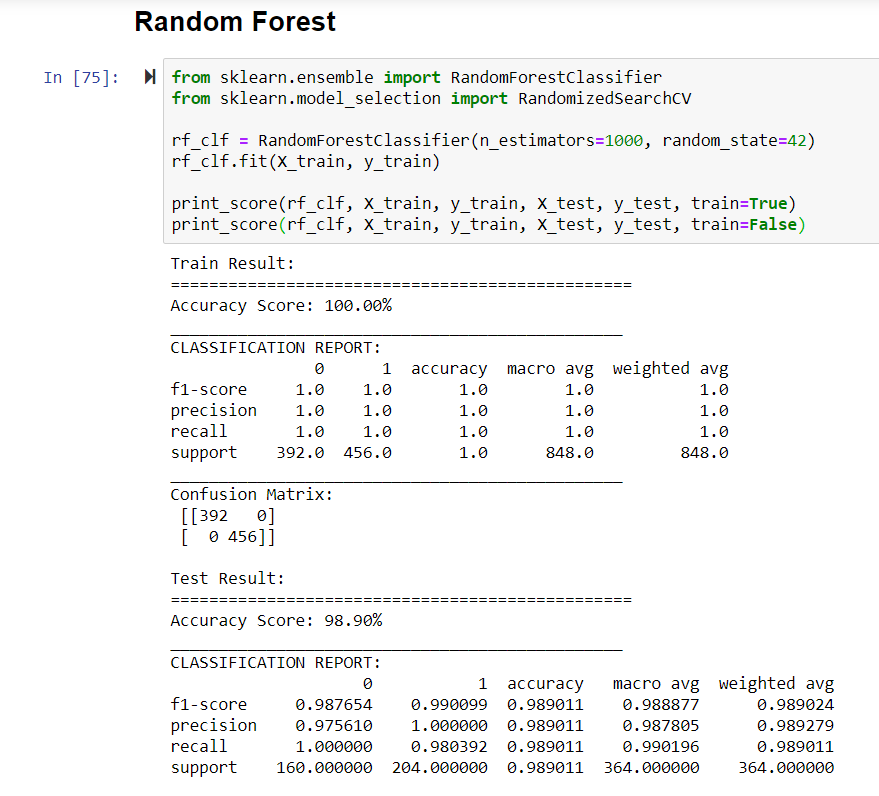
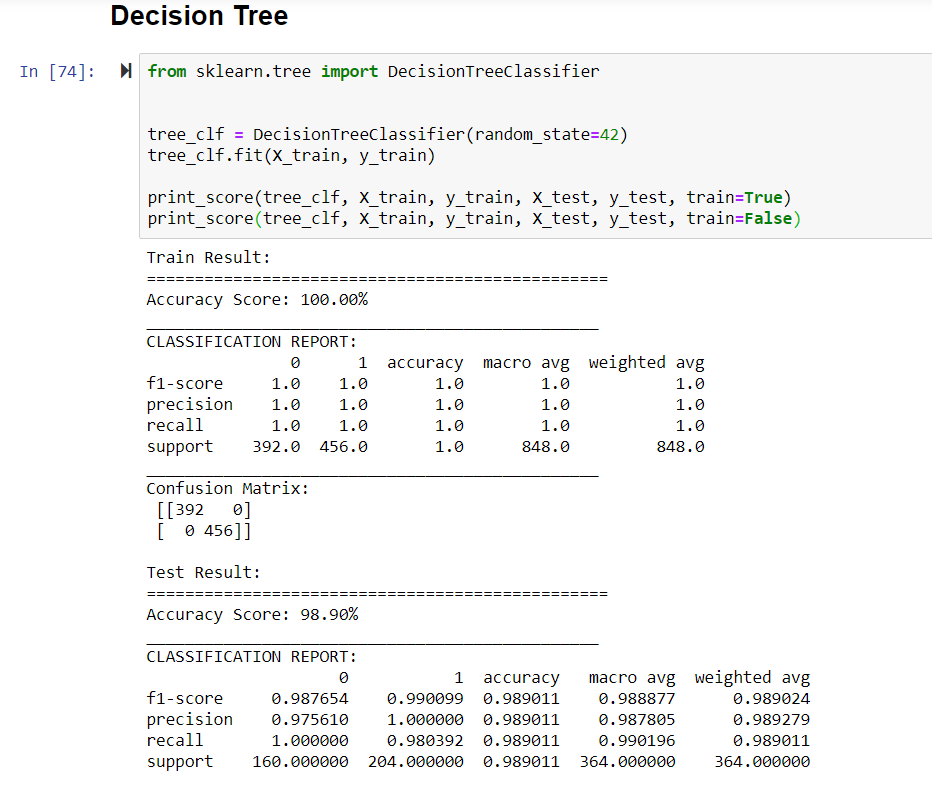
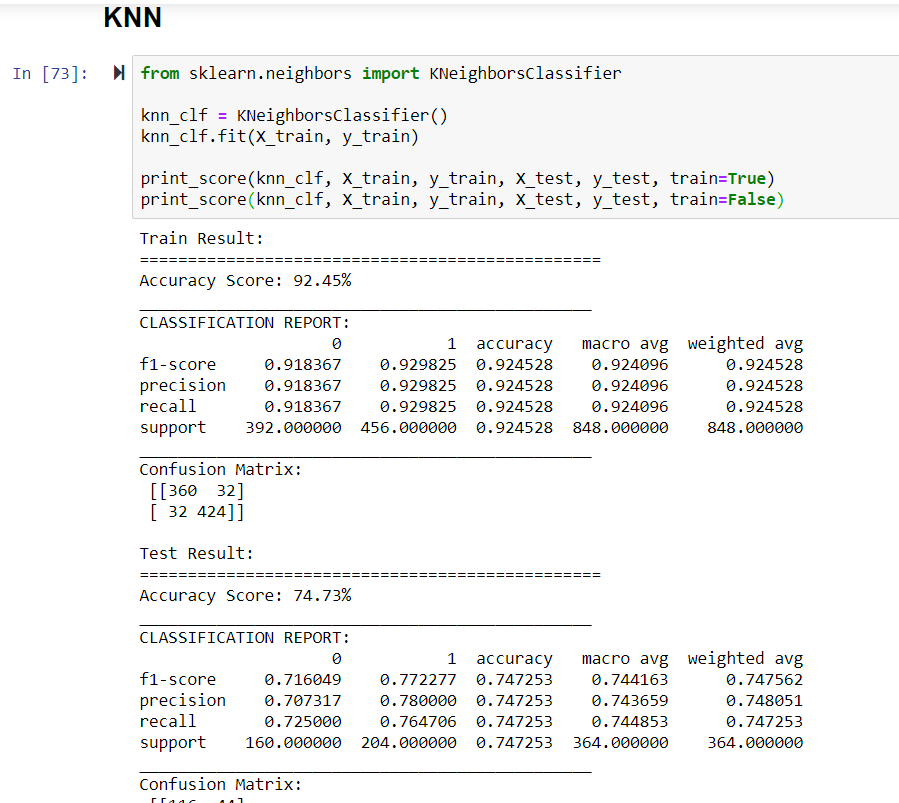
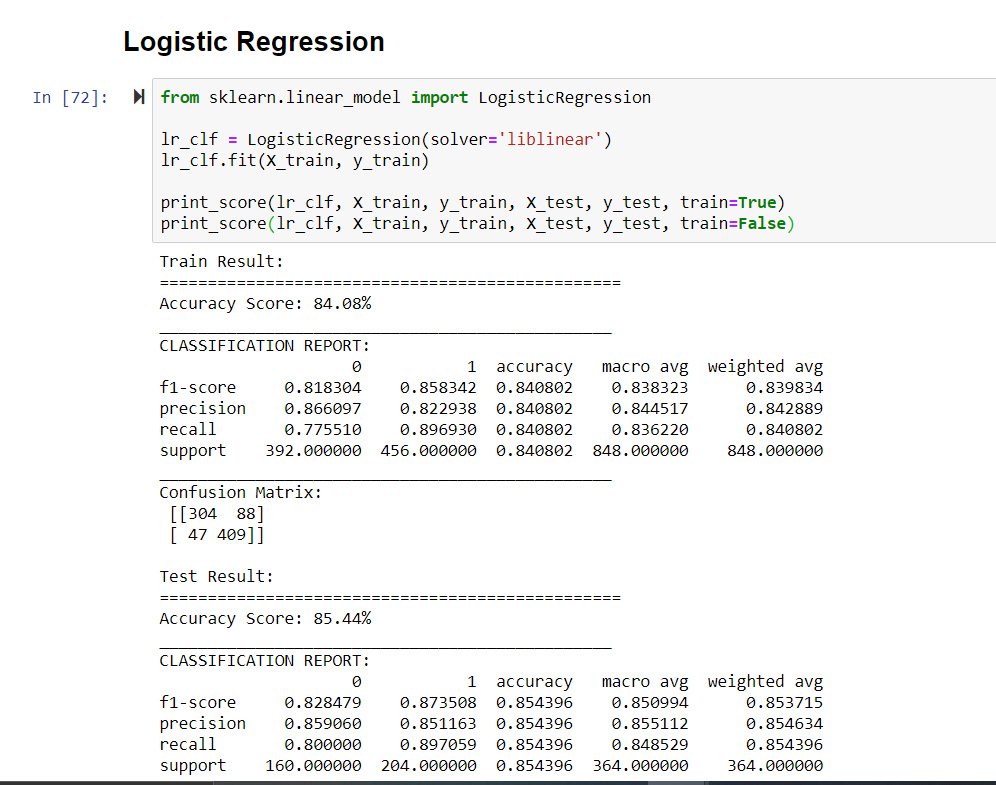
At the end, there will be a python file and a document outlining the entire project that can be used to define variables that influence the heart disease and predicting the best effective model. The application which will ask for the user value and will predict the absence or presence of heart disease.

**EDA:**





**Statistical Modeling:**



**Milestone:**

|  |  |
| --- | --- |
| Milestone | Estimated Delivery Date |
| Statement of Work | 11/06/2020 |
| Data Acquisition and Understanding | 11/23/2020 |
| Modelling | 11/23/2020 |
| Prototyping | 11/23/2020 |
| Deployment | 12/18/2020 |

**Gantt Chart:**

**Data Source:**

https://www.kaggle.com/ronitf/heart-disease-uci

**Dataset Information :**

This dataset gives the information related to heart disease. Datasets contain 14 columns; target is the class variable which is affected by other 13 columns. Here the aim is to classify the target variable to (disease\non disease) using different machine learning algorithm and find out which algorithm suitable for this dataset.

**Attribute Information**

age

sex

chest pain type (4 values)

resting blood pressure

serum cholesterol in mg/dl

fasting blood sugar > 120 mg/dl

resting electrocardiographic results (values 0,1,2)

maximum heart rate achieved

exercise induced angina

oldpeak = ST depression induced by exercise relative to rest

the slope of the peak exercise ST segment

number of major vessels (0-3) colored by fluoroscopy

thal: 3 = normal; 6 = fixed defect; 7 = reversable defect

**Dataset Assumptions:**

As the dataset is clean and every variable is important in our analysis so there are no assumptions required.

**Steps to perform the classification**

The steps that will be followed to perform the classification are-

Import the dataset

Clean the dataset.

Data Visualization, in order to get the general idea about the data like the distribution of individual columns as well as determining the relationship between two variables

Preprocess the dataset. Preprocessing will include converting the string values into numeric format and converting the data which does not follow any ranking into One hot encoding. Also, preprocessing will include removal of outliers and scaling the data.

Model Fitting, various types of classification models will be fitted. Some of the models are-

Logistic Regression

Decision Tress

Support Vector Machines

Naive Bayes

Random Forest Classifier

Extreme Gradient Boost

K-Nearest Neighbor

Model Selection, After the best model is selected, the next step will be moving towards model evaluation.

After the model has been trained on out training data, the model will be tested on testing data, which is separated from the training data. The model will be evaluated on various metrices such as accuracy, precision, recall, and f1-score. If the model is performing well on testing data, the model will be deployed on production, otherwise, one must rework on the data.

**Testing Process:**

**After the model has been trained on out training data, the model will be tested on testing data, which is separated from the training data. The model will be evaluated on various metrices such as accuracy, precision, recall, and f1-score. If the model is performing well on testing data, the model will be deployed on production, otherwise, one must rework on the data.**

**Acceptance:**

Date:01-12-2020

I Bhavna Panwar, a student at Durham College, consent to and accept the terms stated in this Statement of Work by initialing and signing each page below.

*(Durham College)*

By: Bhavna Panwar