**Heart Disease Prediction**

By - Indu Priya Sharma, Nandha Kishore Sangem, Sai Sujith Chadalavada

**Focus Area:**

Heart Disease and attacks are among the most prominent health issues in the United States. According to the CDC [1], the highest cases of death is due to heart disease in the United States. The United States observes one death every 34 seconds on the account of a ‘cardiovascular disease’. In 2020, according to the statistics, one out of every 5 deaths happening was due to heart disease which accounted for nearly 700,000 deaths. Almost 805,000 people had a heart attack in the United States out of which nearly 25% had already had an attack. So, this issue is a critical risk factor and needs attention. The focus is to predict or detect the chance/likelihood of Heart Disease/ Attack using data analytics and machine learning to assist medical workers and the public.

**Hypothesis:**

The team thinks that exang, age, sex, cholesterol, and heart rate attributes are significant factors that can contribute to heart disease/attack.

**Initial Requirements:**

To ensure fulfilment of the requirements of the project, we had to find a dataset that has the attributes/factors that are responsible for or contribute to a heart attack. The dataset needs to have a sufficient amount of records as well to make sure that the machine learning model has enough for training and testing. Additionally, we needed to also be critical of the fact that the dataset doesn’t have personal records of the patient or affected people, in case we come across any, we need to remove them.

**Dataset and Data Source:**

The data is coming from the repository of UCI Machine Learning [2]. The dataset consists of 75 attributes, but only 14 are utilised. The dataset could also be found on Kaggle [3]. There are 8 categorical and 5 numerical variables along with one target variable. Here is an overview on the attributes that our project would be focusing on:

* **Age**: Numerical Variable; This attribute tells about the age of the patient. This can help us understand the distribution of affected and unaffected patients. This can also be utilised as a categorical variable once we use this as an interval range: under x years, x- y years and so on (depending on the initial exploratory data analysis).
* **Sex**: Categorical Variable; This attribute has two values: 1 and 0 which signifies the two types of gender.
* **Exang**: Categorical Variable; This attribute has two values: 1 and 0. 1 indicates that exercise has caused angina which basically is a condition that results in chest pain. 0 indicates that exercise was not the reason behind angina.
* **Ca**: Categorical Variable: This attribute has values ranging from 0 to 3 which tells about the number of major vessels.
* **Cp**: Categorical Variable: This attribute has values ranging from 1 to 4 which tells about the pain type:  
  ○ 1: Typical Angina   
  ○ 2: Atypical Angina  
  ○ 3: non-Anginal Pain  
  ○ 4: Asymptomatic
* **Trtbps**: Numerical Variable: This variable tells about the resting blood pressure and the units is mm Hg.
* **Chol**: Numerical Variable: This variable tells about the person’s cholesterol which is measured by BMI Sensor and the unit is mg/dl.
* **Fbs**: Categorical Variable; This attribute has two values: 1 and 0. 1 indicates that exercise has caused angina which basically is a condition that results in chest pain. 0 indicates that exercise was not the reason behind angina.
* **Rest\_ecg**: Categorical Variable: This attribute has values ranging from 0 to 2 which tells about the resting electrocardiographic results:   
  ○ 0: Normal Ecg   
  ○ 1: Person has ST-T wave abnormality   
  ○ 2: Person has probable or definite left ventricular hypertrophy by Estes’ criteria
* **Thalachh**: Numerical Variable: This attribute tells about the maximum heart rate noted
* **Target**: Categorical Variable; This attribute has two values: 1 and 0. 1 indicates that the patient has a likelihood of heart attack/disease. 0 indicates that exercise was not the reason behind angina. This will also be utilised as the target variable that can be used for building ML models.

**Questions:**

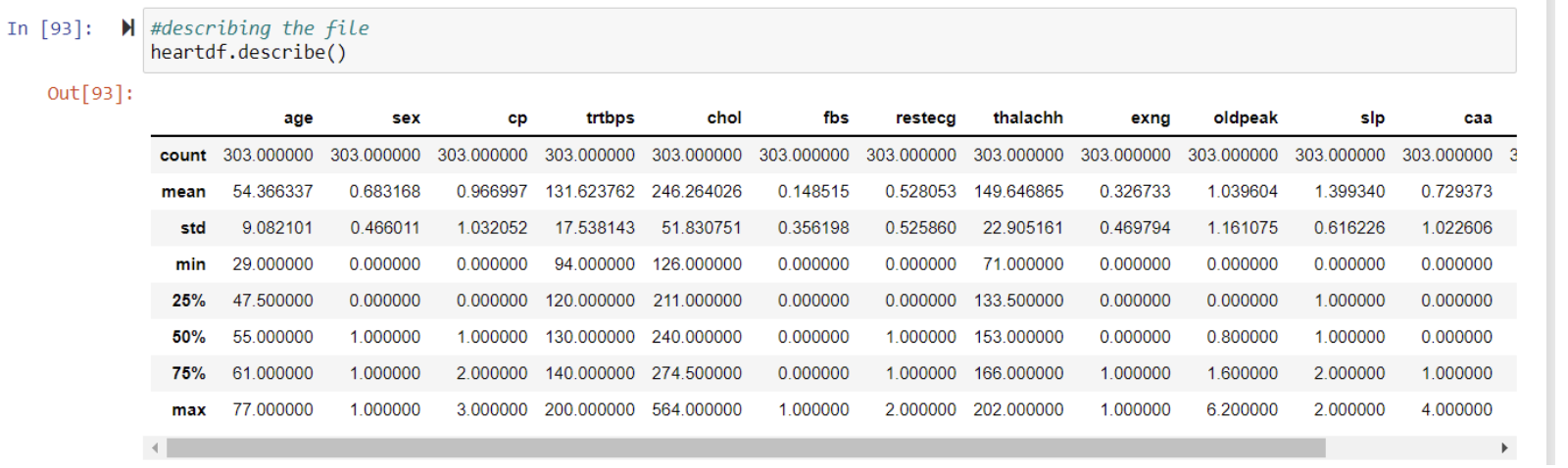
1. Some of the essential questions we wish to answer during our project:
2. What is the most critical factor that may cause Heart Attack/ disease?
3. What factor can reduce the likelihood of Heart Attack/ disease?
4. What insights can be gained using this dataset that can assist Medical experts?
5. What model will help in accurately predicting future likelihood?
6. How the heart rate of a person will have an impact on the chances of getting a heart attack?
7. How does the age factor affect the raise of heart attacks in a person?
8. Which sex category is mostly affected by heart attacks?

**Data Cleaning**

The data seems to have no missing values and the quality of the data looks fine. The data can be normalised as the numerical attributes have different units. This again can be done after the exploratory data analysis.

**Exploratory Data Analysis**

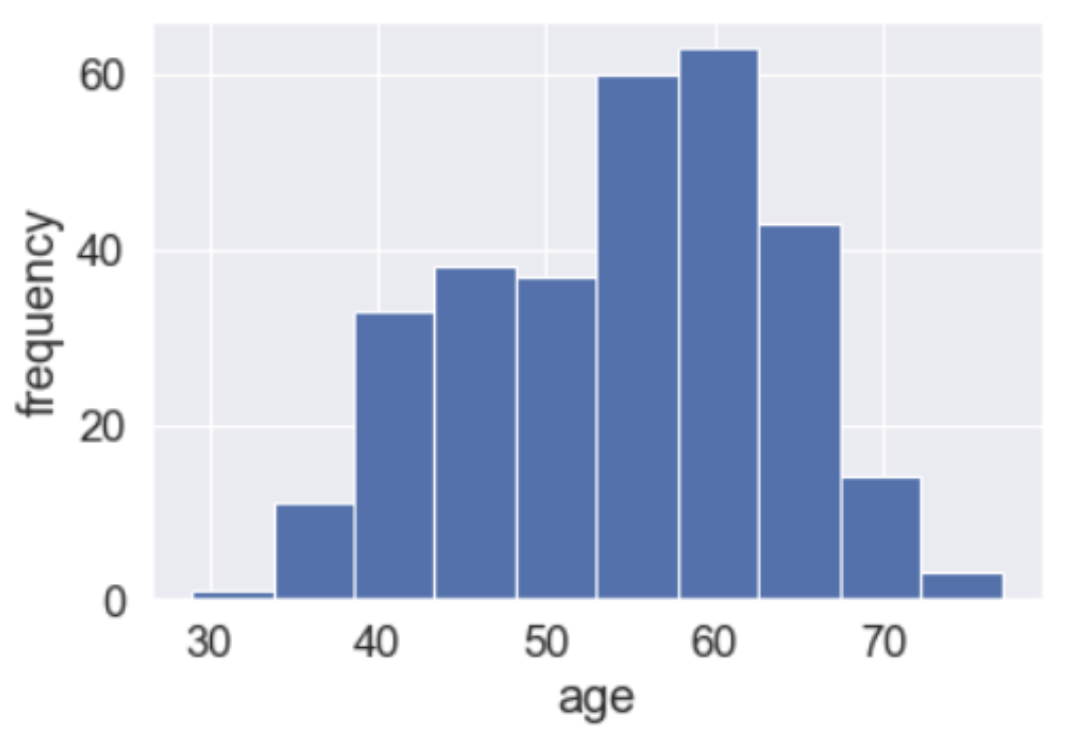
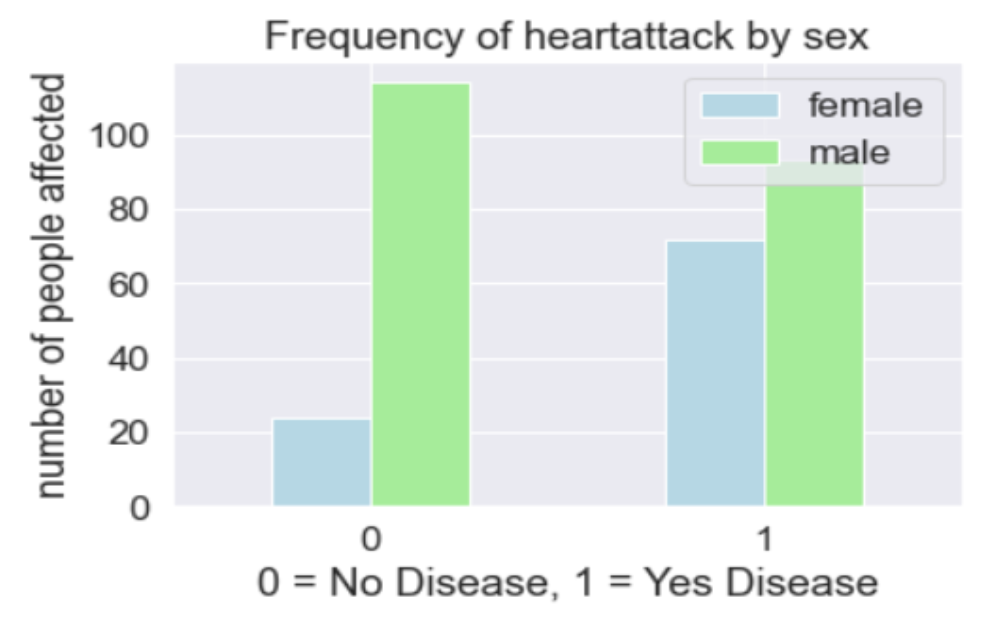
Basic exploration of data has been performed using Python to find various statistics of the obtained data. Initially, the distribution of data was taken and parameters which we considered are observed.



**Fig1- File description and distribution (statistical analysis)**

We believe that heart attack is dependent on the factors such as age, sex, cholesterol and heart rate(thalachh). So, we have done the analysis on each of the factor individually. Based on the analysis performed on the sex & occurrence of heart attack, we observed that males highly suffered the cardiac arrest also, it provides the information who haven’t suffered the cardiac arrest based on sex. it shows us about the number of people who got heart attack based on gender. If we look at the graph there are 24 females whose report shows that they had a heart attack as it is false positive, there are 114 men who had heart attack these are false negative. On the other hand, we can see that there are 93 males who had heart attack (true positive). 74 females had heart attack (true negative).

Other important factor which we focused is, age. We have performed analysis and observed that people of the age group around 60, who suffered cardiac arrest have the highest frequency. In this we check the age distribution in our dataset, and we can conclude that most of the population is of range 40 to 65.



**Fig2- Left: Frequency of heart attack by sex; Right: Frequency of heart attack by age**

In the next step, we perform the analysis based on the heart rate (thalachh) and the age. As per the report, the occurrence of the cardiac arrest is highly probable with increase in the heart rate. Maximum people below the age group of 53 had heart attack, while the age group from 55 to 70 did not show many signs of disease. The people who had heart attack had high heart rate ranging from 140 to 190. The maximum it had gone to 200 and is as low as ranging below 100.

We also believed that the type of exercise which causes angina, denoted by parameter (exang) as an important variable as per the dataset there are 4 different types of chest pains. As per the report, we observe that the chest pain type 0 is the reason causing the cardiac arrest when compared to the other types. Here let us understand what cp and its values are

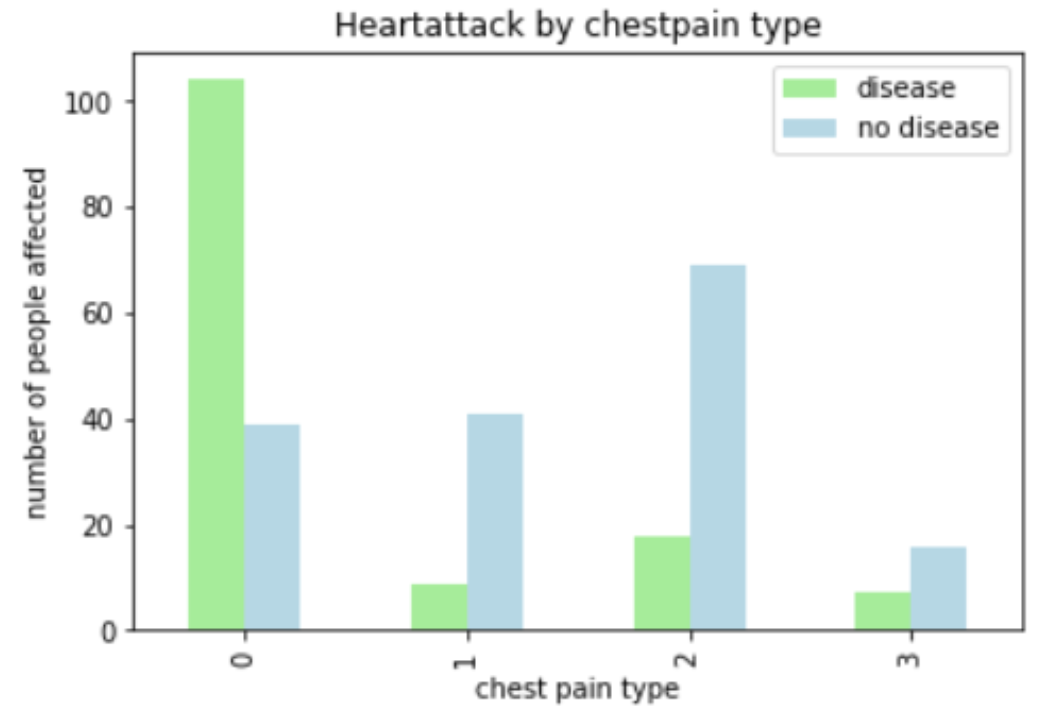
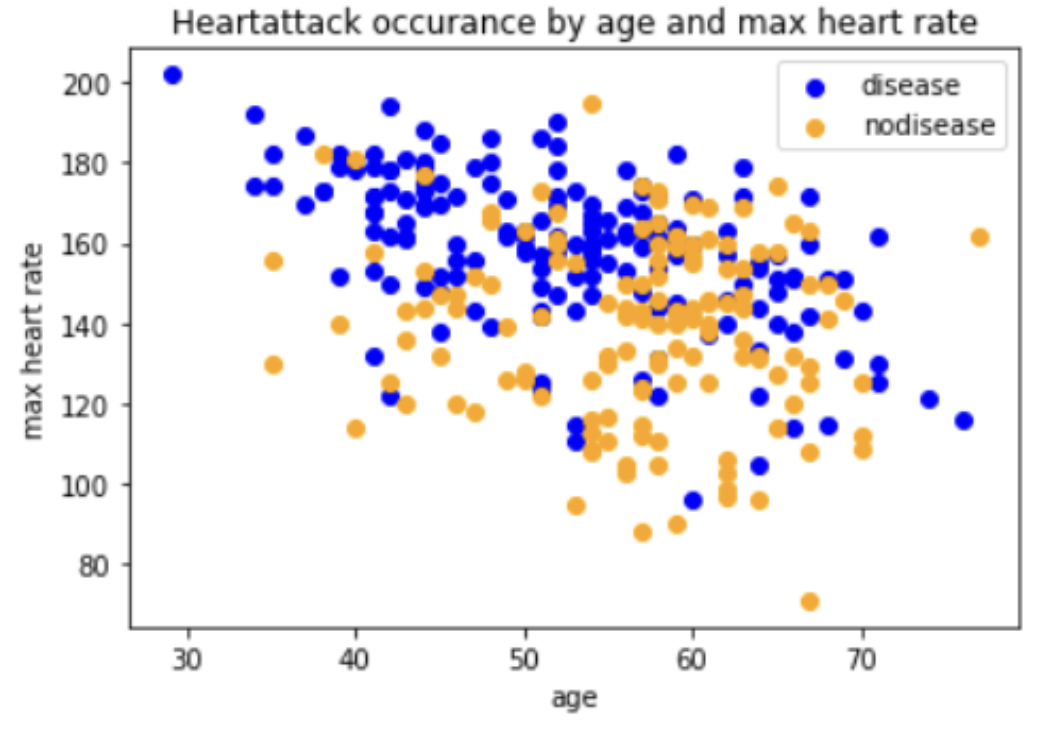
cp 0: typical angina - chest pain due to excess physical/emotional stress

cp 1: atypical angina - this not realted to heart

cp 2: non-aginan pain - pain in chest not caused due to heart/heartattack

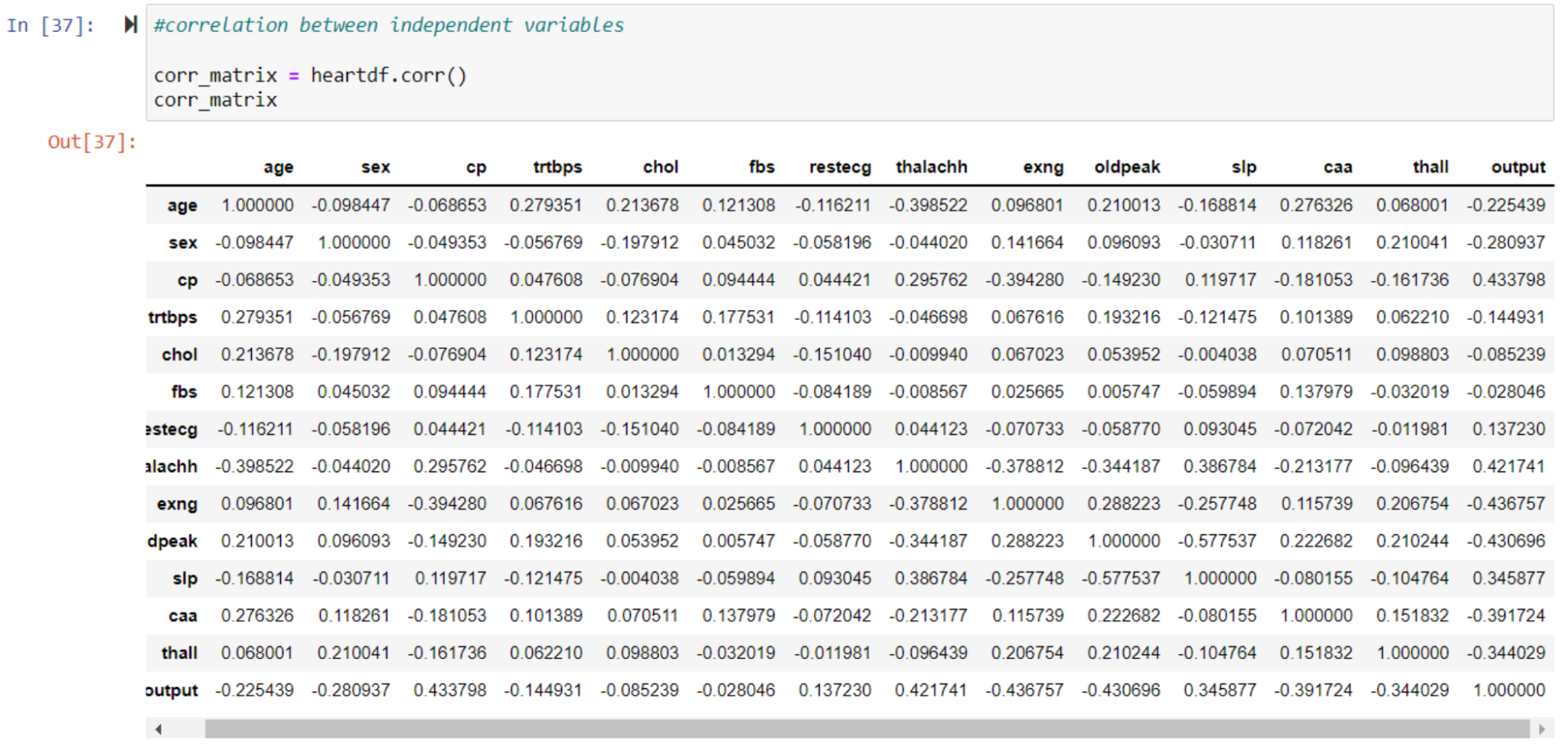
cp 3: asymptomatic - pain not showing any signs of having a disease

The occurrence of heart attack due to typical angina pain is relatively higher and due to cp3 is the least.



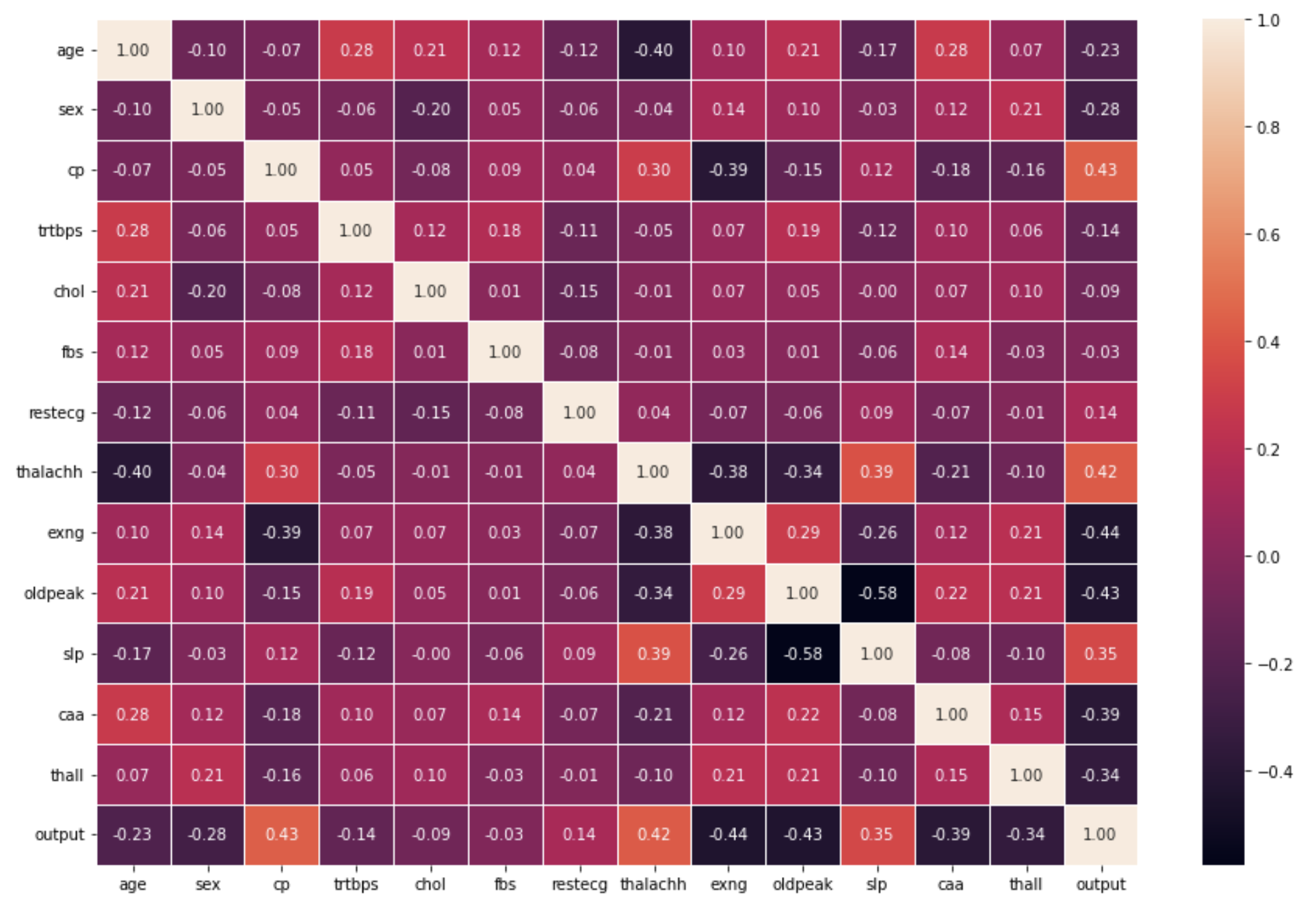
**Fig3- Left: Heart attack occurance by age and max heart rate; Right: Heart attack by chestpain type**

For the dataset we have chosen, the correlation between the parameters have been plotted. Here’s the data of correlation between the parameters:



**Fig4- Correlation**

Below is the heat map, which shows the correlation between the parameters present in the dataset. We can see that there isn’t much correlation for each pair.

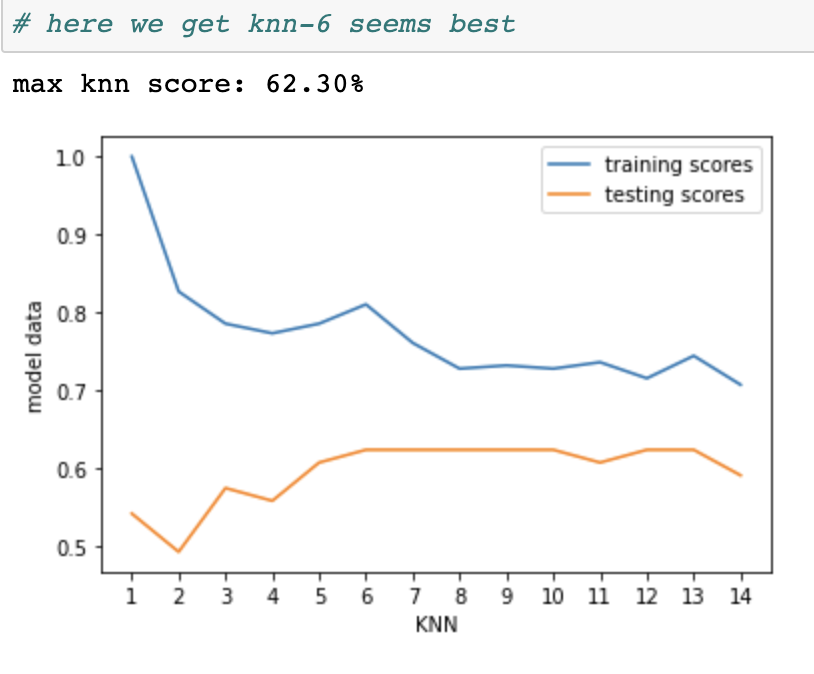


**Fig5- Correlation Heat Matrix**

**Modelling & Algorithms:**

**KNN**

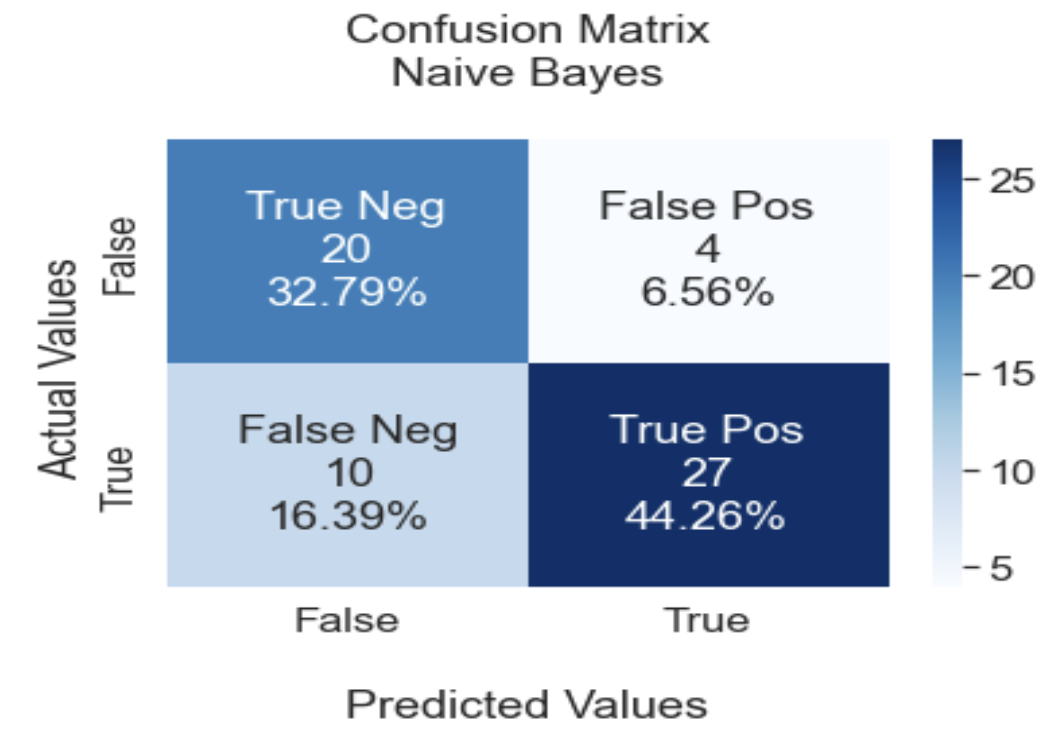
The KNN model is being implemented to aid with the analysis of the dataset using python. As KNN is the supervised model, we have chosen output column as target, which is Y and variables to be explanatory as X. We have split the data into training data and test data. 6 Neighbors seems to be the best performing one.



**Fig6- KNN**

**Naive Bayes**

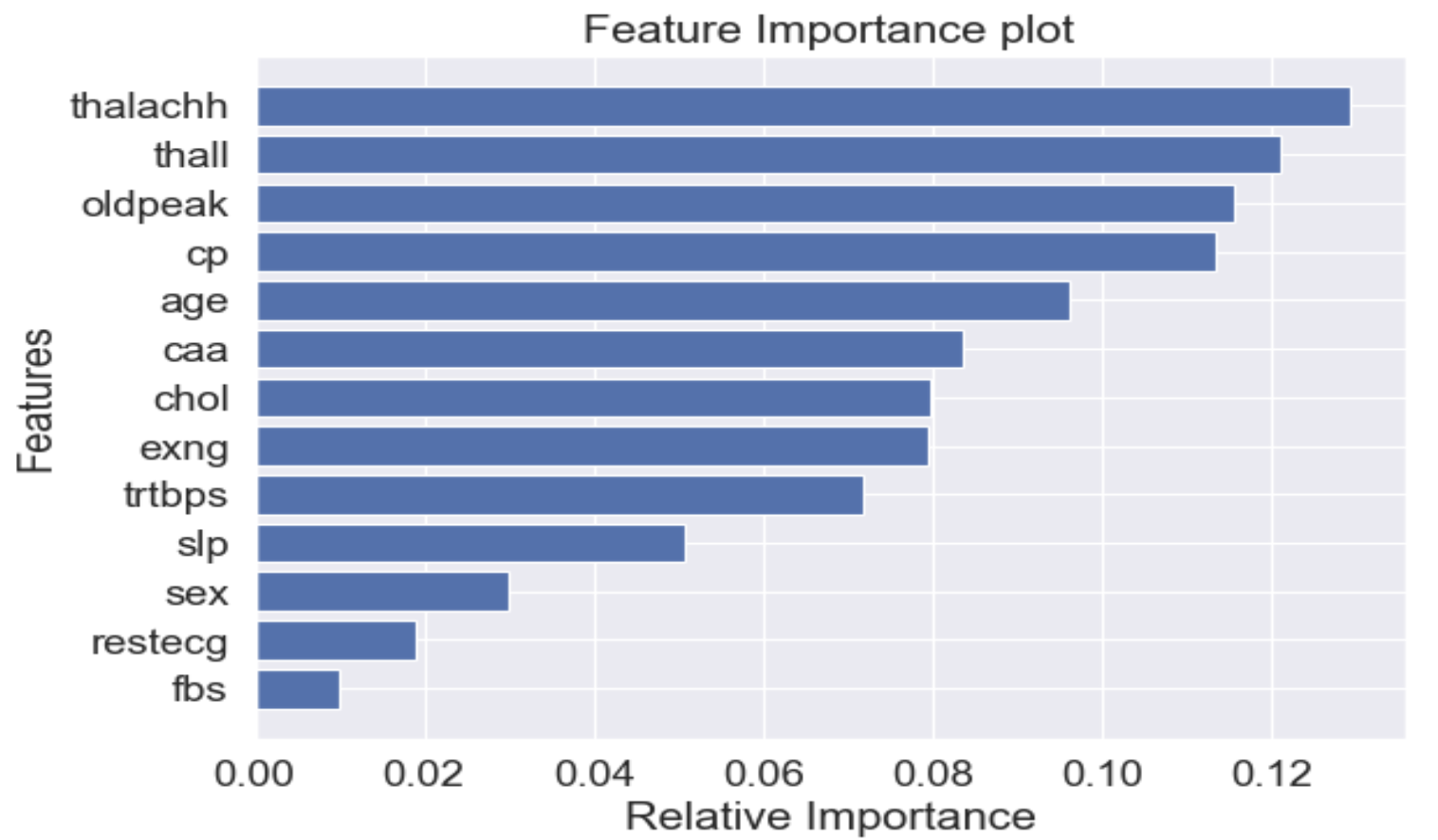
We have used a classifier algorithm to predict and a confusion matrix to measure the performance of this algorithm for the obtained predicted values and actual values. Here is the confusion matrix:



**Fig7- Confusion Matrix Naive Bayes**

**Random Forest**

Using this model, we have measured the accuracy and tried to identify the most important feature in the chosen dataset using feature importance plot.

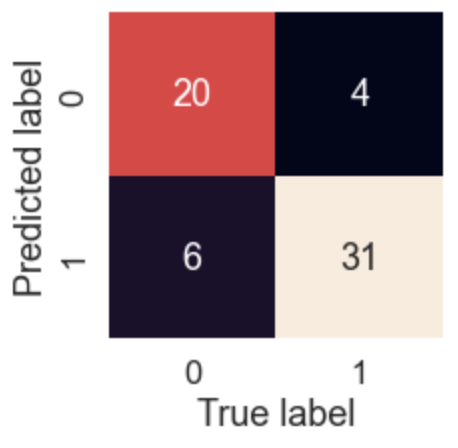


**Fig8- Feature Importance Plot Random Forest**

Through this, we observed that the parameter “thalachh” is an important variable based on which the output is dependent on, when using Random Forest Model.

**Logistic Regression:**

We also utilized Logistic Regression as another classification model. We were able to get an accuracy of 83.48% by using the Grid Search based Logistic Regression model. Below is the confusion matrix for the testing set

.

**Fig9- Confusion Matrix Naive Bayes**

**Results:**

Below is the review of the results of the top 3 Machine learning model:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **F1 Score** | **Precision** | **Recall** |
| **Logistic** | 83.5% | 85.73% | 81.83% | 90.03% |
| **Random Forest** | 80% | 81% | 82% | 80% |
| **Naive Bayes** | 77.05% | 79.41% | 87.10% | 72.97% |

**Table1- Model Performance**

As we can observe that in terms of accuracy the best performing model is the Logistic Regression Model with the accuracy of 83.5% followed by Random Forest and Naive Bayes at 80% and 77.05% respectively. In our case, we want to ensure that the model is able to classify more instances of likelihood of heart diseases hence, we are also considering Recall as one of our key parameter to evaluate the model results. Hence again Logistic Regression performed considerably better than the other models with recall of slightly over 90%.

**Approach:**

Prior to moving ahead with testing the hypothesis that we had formulated we wanted to get more insight about the data, so we first did the exploratory data analysis that included statistical analysis and visualizations. We also did data processing in order to ensure that there is no future inconvenience during the machine learning phase. Since the target variable was categorical, we wanted to choose models that can perform well for binary classification hence we chose the above models. Also, we wanted to make sure we test multiple models prior to settling on one for evaluation.

**Effectivity:**

If we consider the table mentioned under the results, we can see that logistic was able to forecast heart disease likelihood accurately with the percentage being 83.5%. Also as we considered Recall as our key metric, we can also conclude that the algorithm is able to flag 90.03% of the time for heart disease likelihood.

**Lessons Learned:**

There are several lessons learned from our project. First would be that thalachh and thal are two most important factors in comparison to our assumption in the null hypothesis that exang, age, sex, cholesterol, and heart rate attributes would be the significant factors that would contribute more to heart disease/attack. The variable importance plot can be utilized by the doctors and potential patients to ensure that they are on the top of this. Doctors can keep an eye on this even during the regular health check ups. Also, it is wise to clean and process the data prior to machine learning models to ensure correct fit. As a team we also got to learn a lot about what is going on in the world in terms of severity of heart disease and how many people are suffering from it, it was an eye opener. We learned how to approach a formulated hypothesis. We were able to hone our skills in Python and Excel during the course of our project and learned to effectively conduct exploratory data analysis as well.

**References:**

[1] Heart Disease Facts | cdc.gov. (2022, July 15). Centers for Disease Control and Prevention. Retri eved September 18, 2022, from<https://www.cdc.gov/heartdisease/facts.htm>

[2] UCI Machine Learning Repository: Heart Disease Data Set. (n.d.). Retrieved September 18, 2022, from<https://archive.ics.uci.edu/ml/datasets/Heart+Disease>

[3] Kaggle: Your Home for Data Science. (n.d.). Retrieved September 18, 2022, from [https://www.k aggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset](https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset)+