PROGRESSIVE EDUCATION SOCIETY’S

MODERN COLLEGE , PUNE

‘ HEART ATTACK PREDICTION ’

CERTIFICATE

This is to certify that the project report entitled ' **Heart Attack Prediction** ' is being submitted by **Jeet Agrawal(2002895), Divyam Machale (2003383) , Akshay Kedari(2002901)** as a partial fulfilment for the award of the degree of the Master of Science (M.Sc.). This is a record of bonafide work carried out by them under supervision and guidance.

Prof. Vidya Sangale Mam	Prof. P.G. Dixit
Project Guide	**Head Dept. of Statistics**

Place: Pune Date:

ACKNOWLEDGEMENT

We have satisfaction upon completion of our project work 'Heart Attack Prediction' at the Department of Statistics of Modern College, Pune during the academic year 2021-2022.

We are thankful to Prof. Vidya Sangale Mam for providing all the necessary help, guidance and co-operation required for the completion of this project work.

Also, the teaching staff of Department of Statistics for their valuable support and cooperation during completion of this project.

Index

MOTIVATION

A major challenge facing healthcare organizations (hospitals, medical centers) is the provision of quality treatments that are effective. Poor clinical decisions can lead to disastrous consequences which are therefore unacceptable. Hospitals must also minimize the cost of clinical tests. They can achieve these results by employing support systems.

Most hospitals today employ some sort of hospital information systems to manage their healthcare or patient data. These systems typically generate huge amounts of data which take the form of numbers, text, charts and images. Unfortunately, these data are rarely used to support clinical decision making.

There is a wealth of hidden information in these data that is largely untapped. This raises an important question: How can we turn data into useful information that can enable healthcare practitioners to make intelligent clinical decisions?
This is the main motivation for this research.

ABSTRACT

Heart is considered as the second most important organ after brain. Any disturbance in Heart leads to Whole body disturbed. Diseases including heart disease as major are the result of the changes happening to us on daily basis and major that heat disease is among the among the top five killer disease worldwide. Therefore, predicting the disease at the right time and right moment is the important. Data mining is a basic and primary process in defining and discovering useful information and finding hidden patterns from large databases. Data mining and Machine Learning techniques find its use in medical sciences in resolving real health related issues by prediction and diagnosis of various diseases. This paper aims at analyzing three machine learning algorithms comparatively for heart disease prediction viz., support vector machine, decision tree and random forest.

Keywords: Heart Disease, Machine Learning, SVM, Deacision Tree, Random Forest ,XGBoost.

INTRODUCTION

In human body, the heart is a muscular organ which is located at the center of the chest and points slightly towards the left that collects deoxygenated blood from all parts of the body and takes this deoxygenated blood to lungs wherein it is converted into oxygenated blood, releasing carbon dioxide. Thus, oxygenated blood is then transported from lungs to all parts of the body.

The heart pumps an approximate of 72,00 liters of blood throughout the body in 24 hours and it beats around 3 billion beats in a life time. The heart not only circulates blood, but also circulates other substances like: nutrients from digestion are delivered to all cells of the body, hormones which are produced from tissues are distributed to all cells of the body . The circulatory system carries waste substances which are finally passed to bladder. The important functioning of heart includes the interstitial fluid pumping from blood into the extracellular space .

From the above discussion, it is clear that heart is an important organ in the body .so, any disruption in heart leads to disruption in the whole body.

The factors that lead to malfunctioning in heart allude to any serious abnormal condition of the heart or blood vessel (arteries, veins) called heart disease. The various kinds of heart disease are :

- **Coronary Artery Disease** – refers to formation of cholesterol plaque which causes hardening or narrowing of the coronary artery, (which supplies blood to the heart).
- **Cardiomyopathy** – refers to disease of the heart muscle because of several reasons
- **Angina** – in which chest pain causes due to less blood flow to a part of the heart muscle.
- **Valvular Heart Disease** – refers to disease that affects one or more of the four valves of the heart
- **Congenital Heart Disease** – heart structure malformation at birth
- **Cerebrovascular Disease** – refers to disease of the blood vessels that supply blood to the brain.

- **Rheumatic Heart Disease** – disease causing damage to the heart muscles and valves due to rheumatic fever.
- **Heart attack** – is permanent damage to the part of the heart muscle to which blood supply was cut off
- **Heart failure**- is heart's pumping power has decreased causing heart fail.

The above given heart diseases are only some of the several heart diseases that affect both women and men. These can be recognized by deeply observing the common factors causing the heart diseases

STATISTICAL STATEMENT

With reference to the data set attached with the report, the objective is to observe the effect of various factors like age, cholesterol level, blood sugar level ,etc on heart using various Machne Learning Classification Algorithms like Naïve Bayes , KNN , Decision Tree(CART) , SVM(support vector machine) Also, to obtain the best classification model for predicting the heart disease in a person.

It is of interest to check the association between the any two categorical using chi square test of independence

Dataset Description

Name	Type	Description
Age	Continuous	Age Age in years
Sex	Discrete	0 = female 1 = male
Cp	Discrete	Chest pain type: 1 = typical angina, 2 = atypical angina, 3 = non-anginal pain 4 =asymptom
Trestbps	Continuous	Resting blood pressure (in mm Hg)
Chol	Continuous	Serum cholesterol in mg/dl
Fbs	Discrete	Fasting blood sugar>120 mg/dl: 1-true 0=False
Exang Continuous Maximum heart rate achieved	Discrete	Exercise induced angina: 1 = Yes 0 = No
Thalach	Continuous	Maximum heart rate achieved
Old peak ST	Continuous	Depression induced by exercise relative to rest
Slope	Discrete	The slope of the peak exercise segment : 1 = up sloping 2 = flat 3 = down sloping
Target	Discrete	Diagnosis classes: 0 = No Disease 1= likely have heart disease

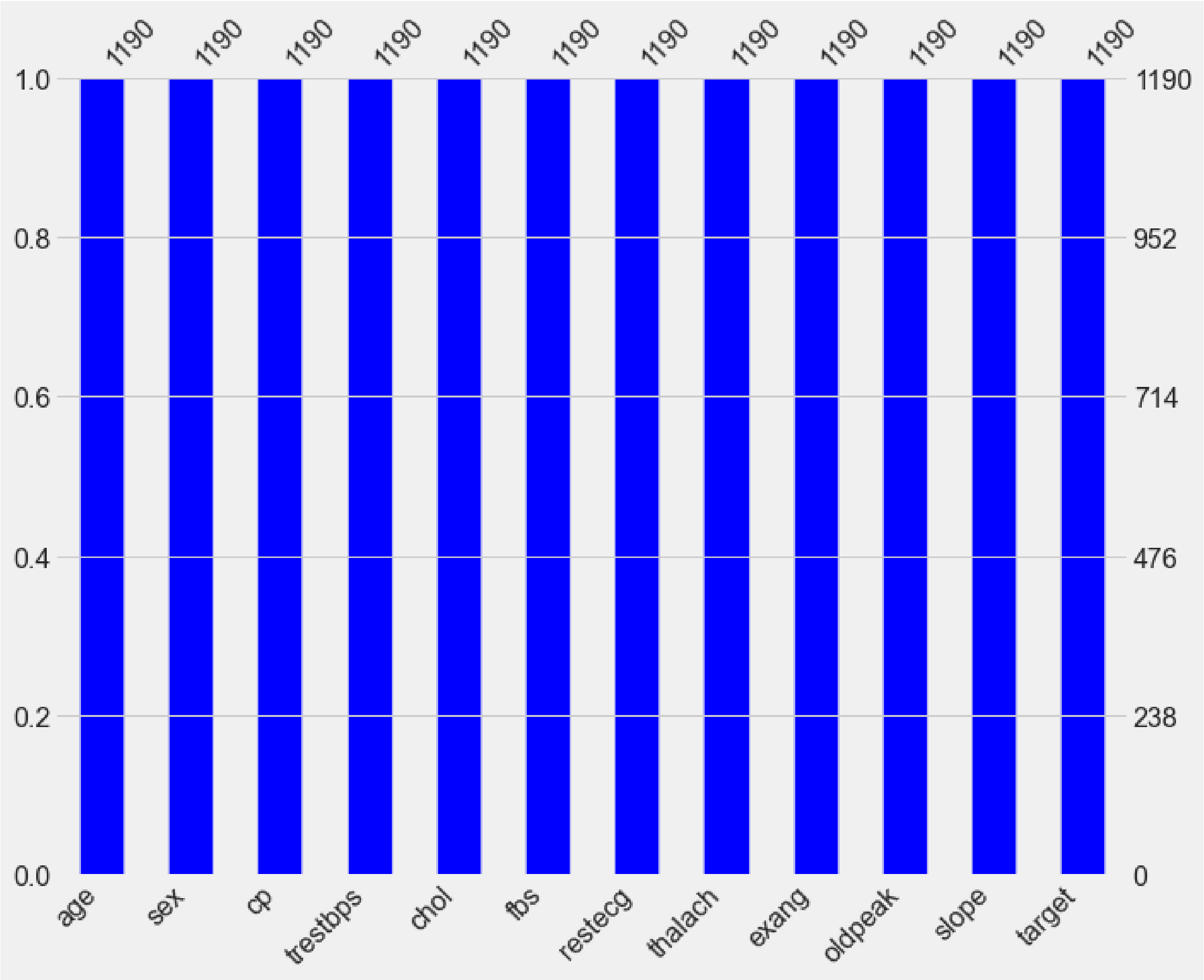
The dataset used for this research purpose was the Public Health Dataset. It contains 76 attributes, including the predicted attribute, but all published experiments refer to using a subset of 12 of them. The "**target" field refers to the presence of heart disease in the patient. It is integer-valued 0 no disease and 1 disease.**

Data

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	Slope	target
0	40	1	2	140	289	0	0	172	0	0.0	1	0
1	49	0	3	160	180	0	0	156	0	1.0	2	1
2	37	1	2	130	283	0	1	98	0	0.0	1	0
3	48	0	4	138	214	0	0	108	1	1.5	2	1
4	54	1	3	150	195	0	0	122	0	0.0	1	0
...	...	...	...	...	...	...	...	...	...	...	...	...
1185	45	1	1	110	264	0	0	132	0	1.2	2	1
1186	68	1	4	144	193	1	0	141	0	3.4	2	1
1187	57	1	4	130	131	0	0	115	1	1.2	2	1
1188	57	0	2	130	236	0	2	174	0	0.0	2	1
1189	38	1	3	138	175	0	0	173	0	0.0	1	0

1190 rows × 12 columns

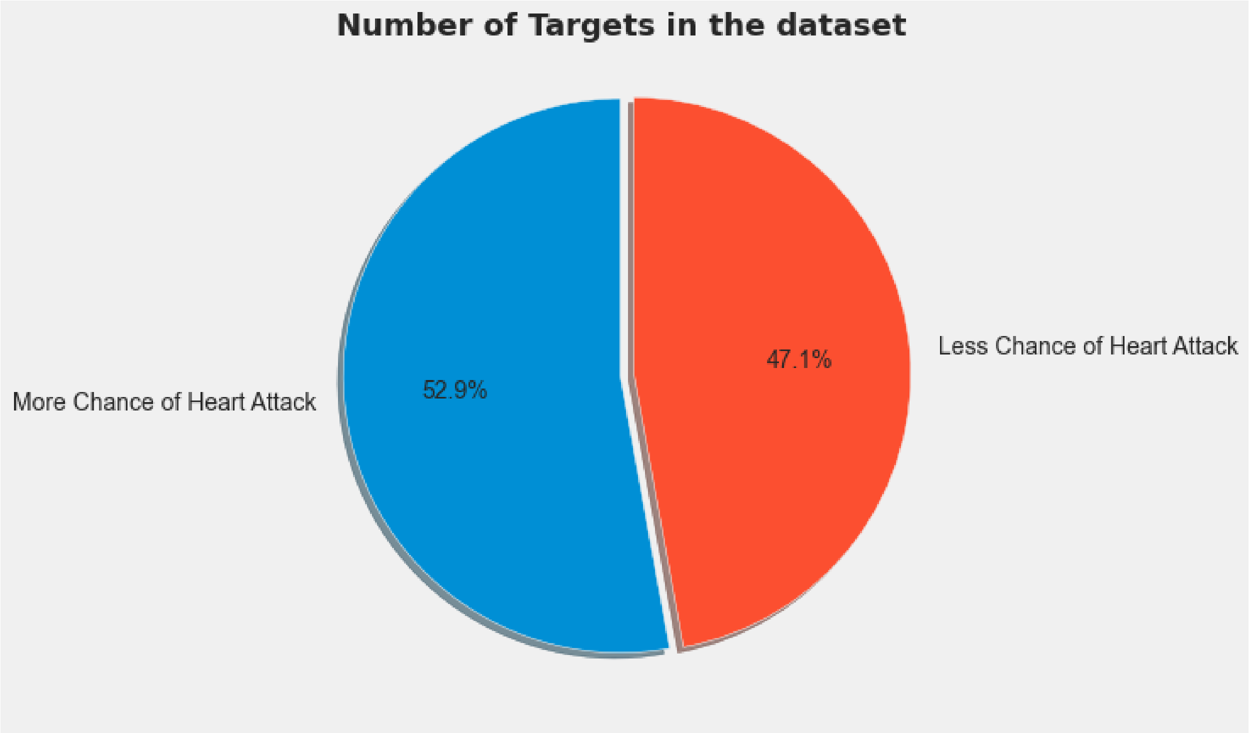
❖ Checking The Missing Values in our data



📌 From above plot we can observe that there are no missing values present in our dataset.

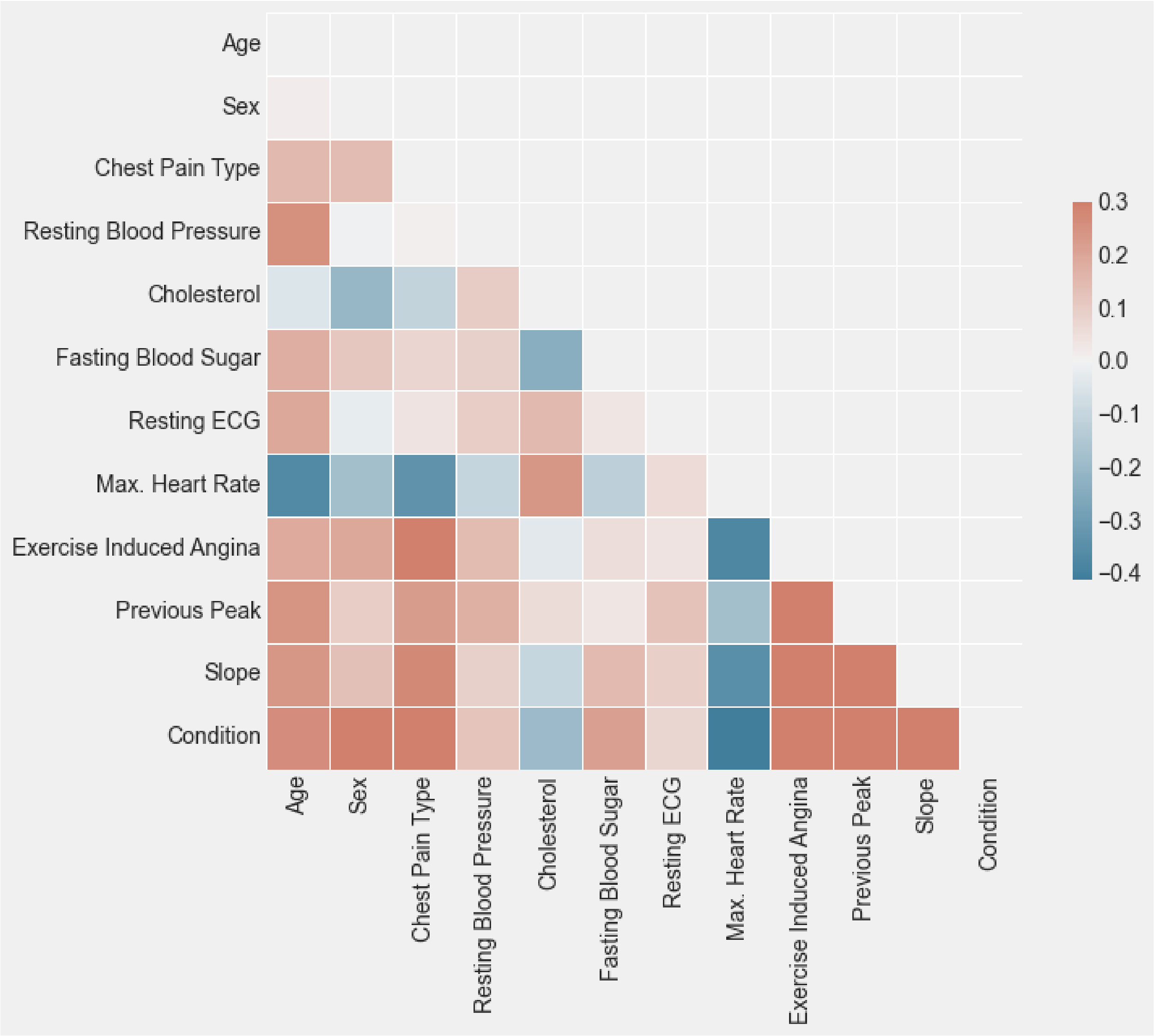
Exploratory Data Analysis

❖ Pie Chart :

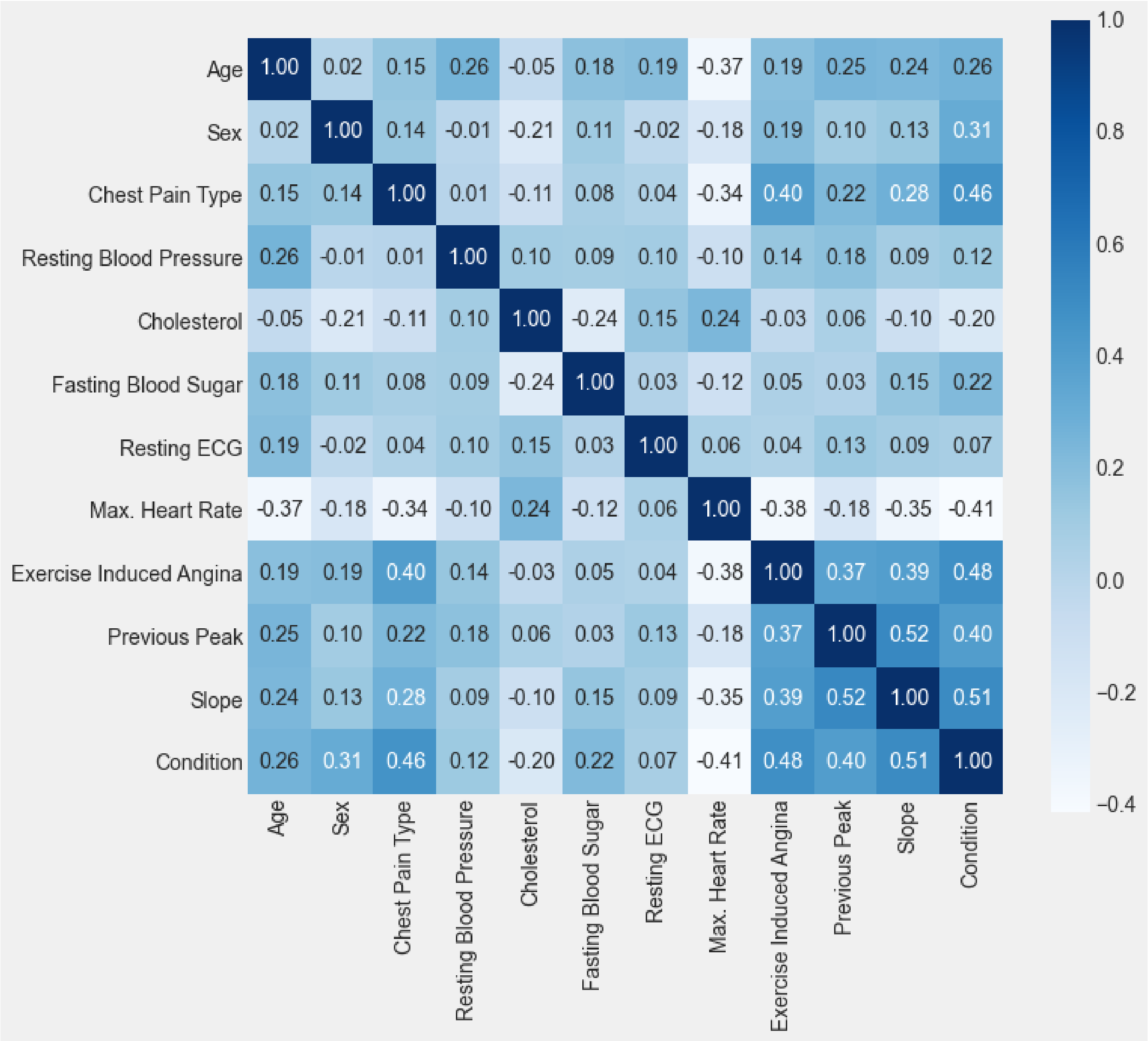


📌 **From the above pie chart, we can see that we have relatively more people who have more chances of having a Heart Attack.**

❖ Heat Map Correlation



With correlation coefficients :



	Age	Sex	Chest Pain Type	Resting Blood Pressure	Cholesterol	Fasting Blood Sugar	Resting ECG	Max. Heart Rate	Exercise Induced Angina	Previous Peak	Slope	Condition
Age	1.00	0.02	0.15	0.26	-0.05	0.18	0.19	-0.37	0.19	0.25	0.24	0.26
Sex	0.02	1.00	0.14	-0.01	-0.21	0.11	-0.02	-0.18	0.19	0.10	0.13	0.31
Chest Pain Type	0.15	0.14	1.00	0.01	-0.11	0.08	0.04	-0.34	0.40	0.22	0.28	0.46
Resting Blood Pressure	0.26	-0.01	0.01	1.00	0.10	0.09	0.10	-0.10	0.14	0.18	0.09	0.12
Cholesterol	-0.05	-0.21	-0.11	0.10	1.00	-0.24	0.15	0.24	-0.03	0.06	-0.10	-0.20
Fasting Blood Sugar	0.18	0.11	0.08	0.09	-0.24	1.00	0.03	-0.12	0.05	0.03	0.15	0.22
Resting ECG	0.19	-0.02	0.04	0.10	0.15	0.03	1.00	0.06	0.04	0.13	0.09	0.07
Max. Heart Rate	-0.37	-0.18	-0.34	-0.10	0.24	-0.12	0.06	1.00	-0.38	-0.18	-0.35	-0.41
Exercise Induced Angina	0.19	0.19	0.40	0.14	-0.03	0.05	0.04	-0.38	1.00	0.37	0.39	0.48
Previous Peak	0.25	0.10	0.22	0.18	0.06	0.03	0.13	-0.18	0.37	1.00	0.52	0.40
Slope	0.24	0.13	0.28	0.09	-0.10	0.15	0.09	-0.35	0.39	0.52	1.00	0.51
Condition	0.26	0.31	0.46	0.12	-0.20	0.22	0.07	-0.41	0.48	0.40	0.51	1.00

📌 From the above correlation matrix, we can see that the **correlation between features is less.**

Range	Strength of association
0	No association
0 to ±0.25	Negligible association
±0.25 to ±0.50	Weak association
±0.50 to ±0.75	Moderate association
±0.75 to ±1	Very strong association
±1	Perfect association

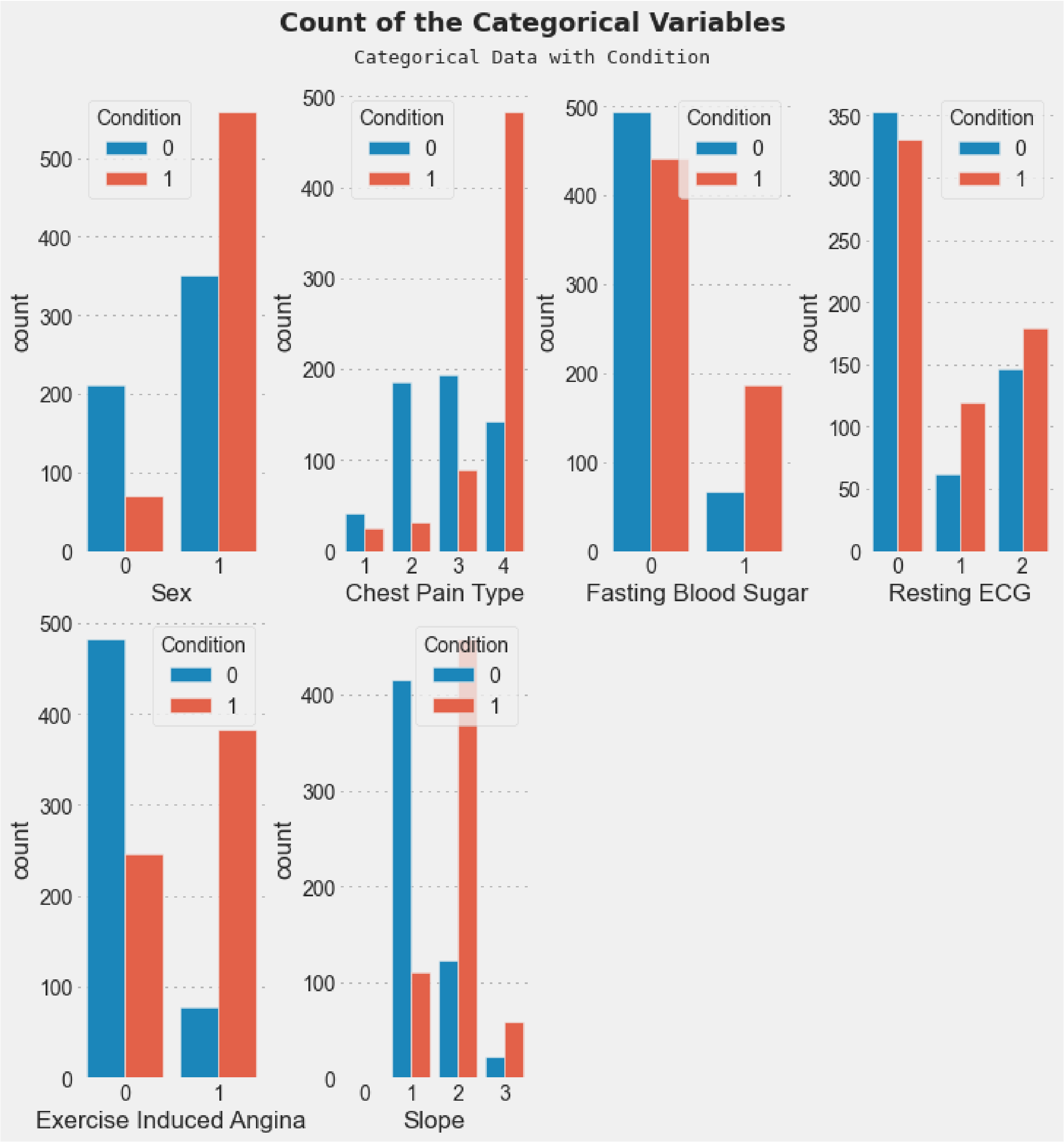
📌 **oldpeak** i.e.ST depression induced by exercise relative to rest ('ST' relates to positions on the ECG plot) with **the slope of the peak** exercise ST segment and **the slope of the peak** exercise ST segment with **Condition(Target Variable)** are the **moderately correlated features** in our dataset; Correlation Coefficient of **0.52 and 0.51** respectively.

📌 Our features have a lot of negative correlation coefficient indicating that two **individual variables have a statistical relationship such that generally move in opposite directions from one another.**

❖ Count Plot :

count plot is a graphical display to show **the number of occurrences or frequency for each categorical data using bars**

Count Plot of categorical variables with condition(target variable) :-



From above categorical plots we can see that:

📌 **In sex, Male(1) has the highest number of people who have more chances of suffering from a heart attack.**

📌 **In Chest Pain, Type 4 has the highest number of people who have more chances of suffering from a heart attack.**

📌 **Fasting Blood Sugar and Resting ECG doesn't have much difference in their respective conditions.**

📌 **In Exercise-Induced Angina, Type 0 has the highest number of people who have less chances from suffering a heart attack.**

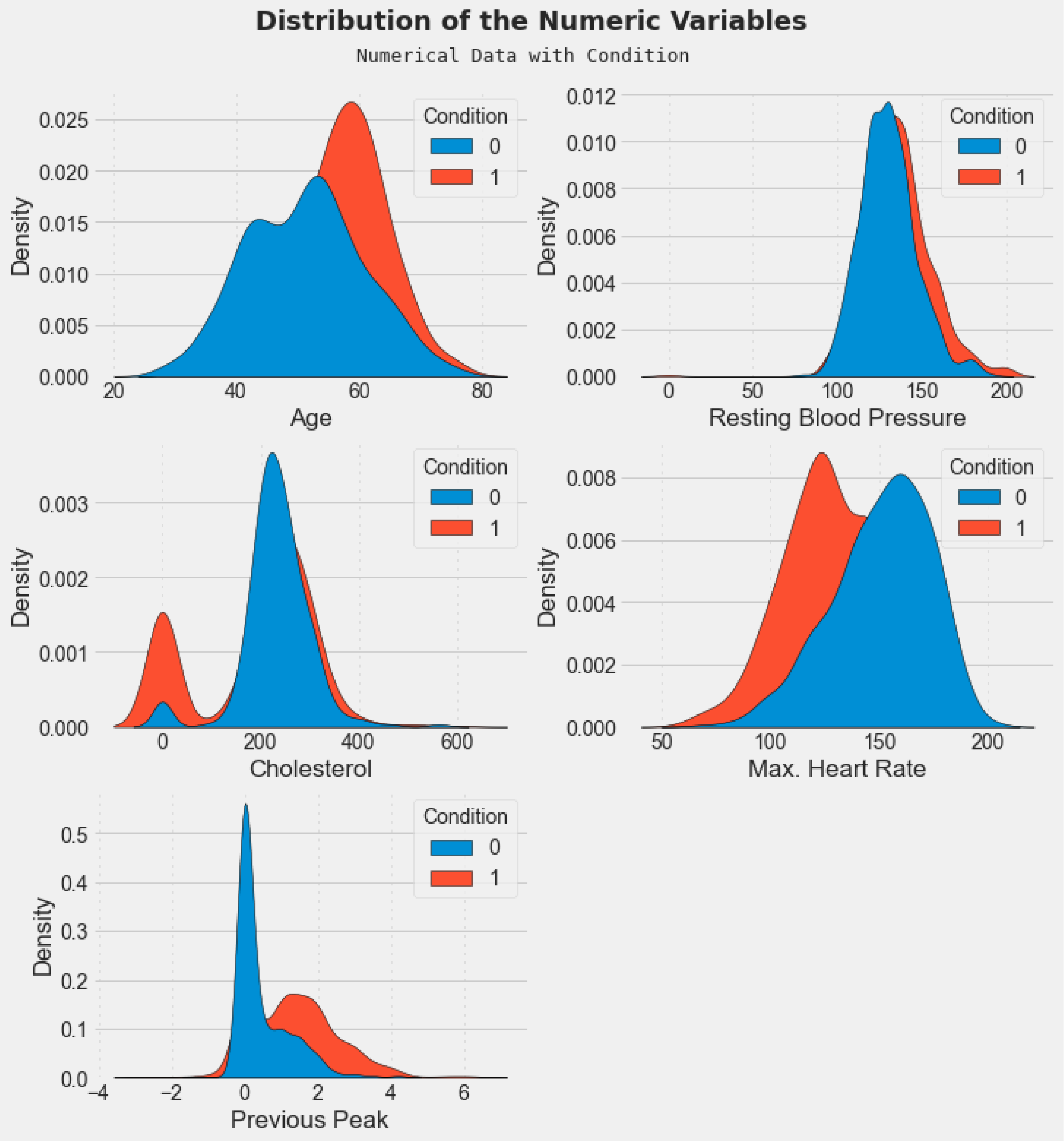
📌 **Slope has Type 2 which shows people who are likely to suffer from a heart attack.**

❖ Distribution Plot :

A Distplot or distribution plot, depicts the variation in the data distribution.

They help us detect skewness of the dataset (Skewness refers to **a distortion or asymmetry that deviates from the symmetrical bell curve, or normal distribution**).

Distribution of numeric variables with condition (target variable):



- **Positively skewed distribution**, the data is more bent towards the lower side, the average value will be more than the middle value.
- **Negatively skewed distribution**, the average value will be less than the middle value.

📌 **Positively skewed distribution**

The distribution of **Resting blood pressure and Previous peak is positively skewed** which concludes that most of the patients have rbp and previous peak around some mean value with few outlier patients have those values are towards higher range .

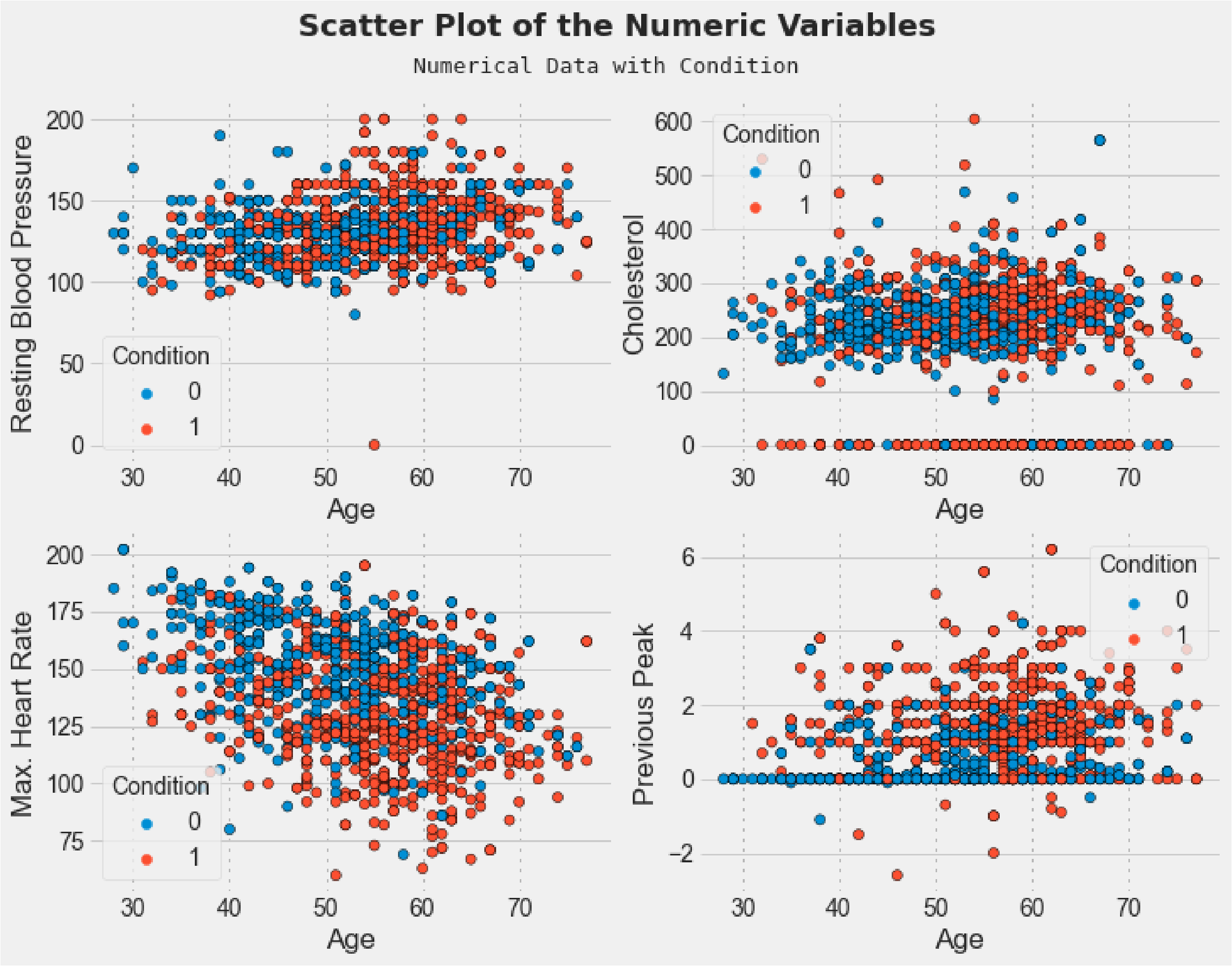
📌 **Negatively skewed distribution**

Heart rate is negatively skewed implies most of the patients have heart rate towards higher side .

📌 Where **age and cholesterol are giving us bimodal distribution as it has 2 peaks.**

❖ Scatter Plot :

Scatter plots' primary uses are to observe and show **relationships between two numeric variables ,** Scatter plots can also show if there are **any unexpected gaps** in the data and if there are any **outlier points**.



📌 From the above plot, we can see that the **relationship between Age and different numerical features** in our dataset **with Condition.** We can also see few **Outliers** in our plot.

❖ Outliers :

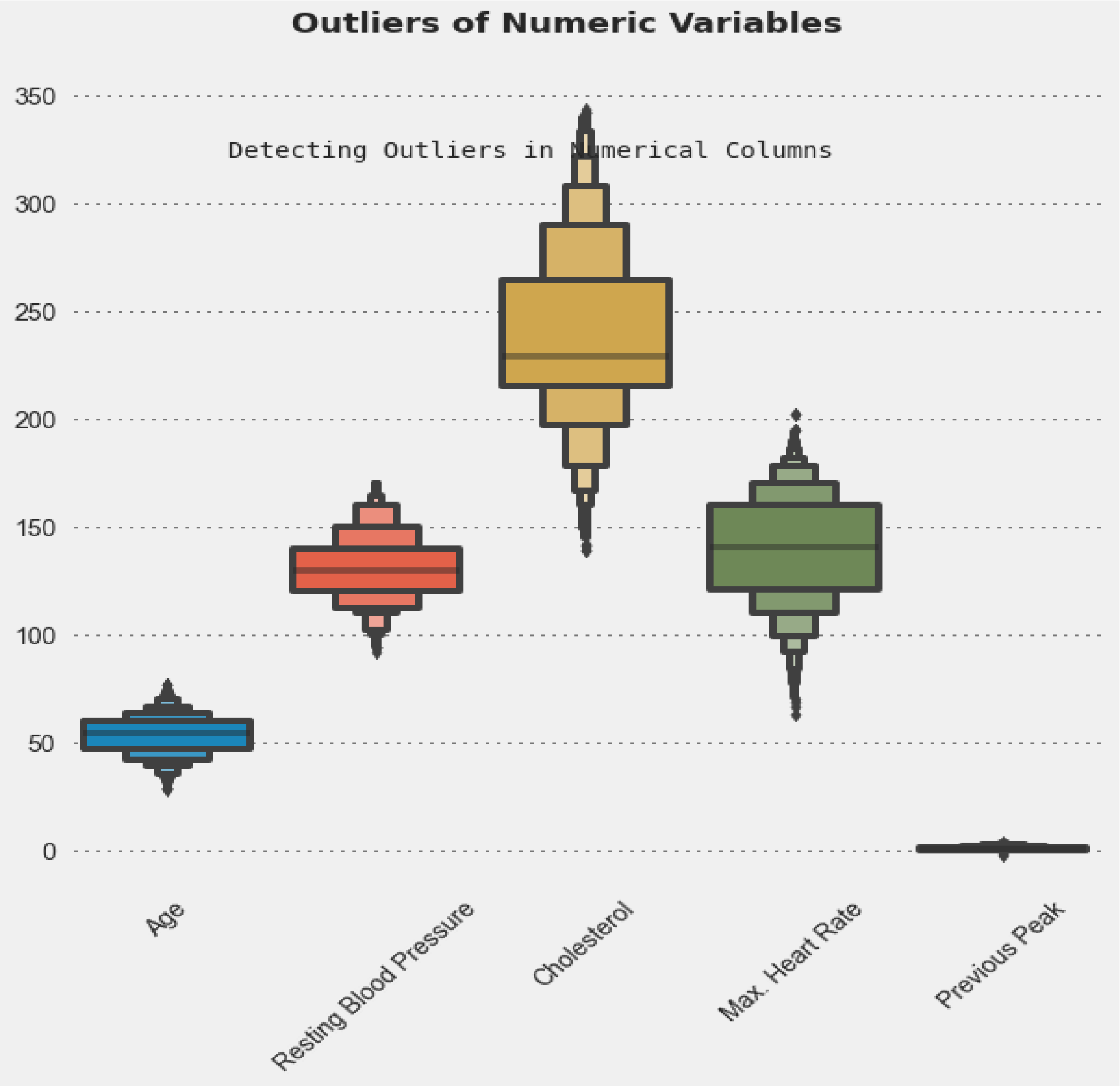
📌 **An outlier is an observation that lies an abnormal distance from other values in a random sample from a population.**

Detection

📌 **Here we have used Boxen Plot to detect the outliers of each features in our dataset, where any point above or below the whiskers represent an outlier. This is also known as "Univariate method" as here we are using one variable outlier analysis.**

❖ Boxen Plot :

The Boxen plot is **very similar to box plot**, except for the fact that **it plots different quartile values**. By plotting different quartile values, we are able to understand the **shape of the distribution particularly in the head end and tail end.**

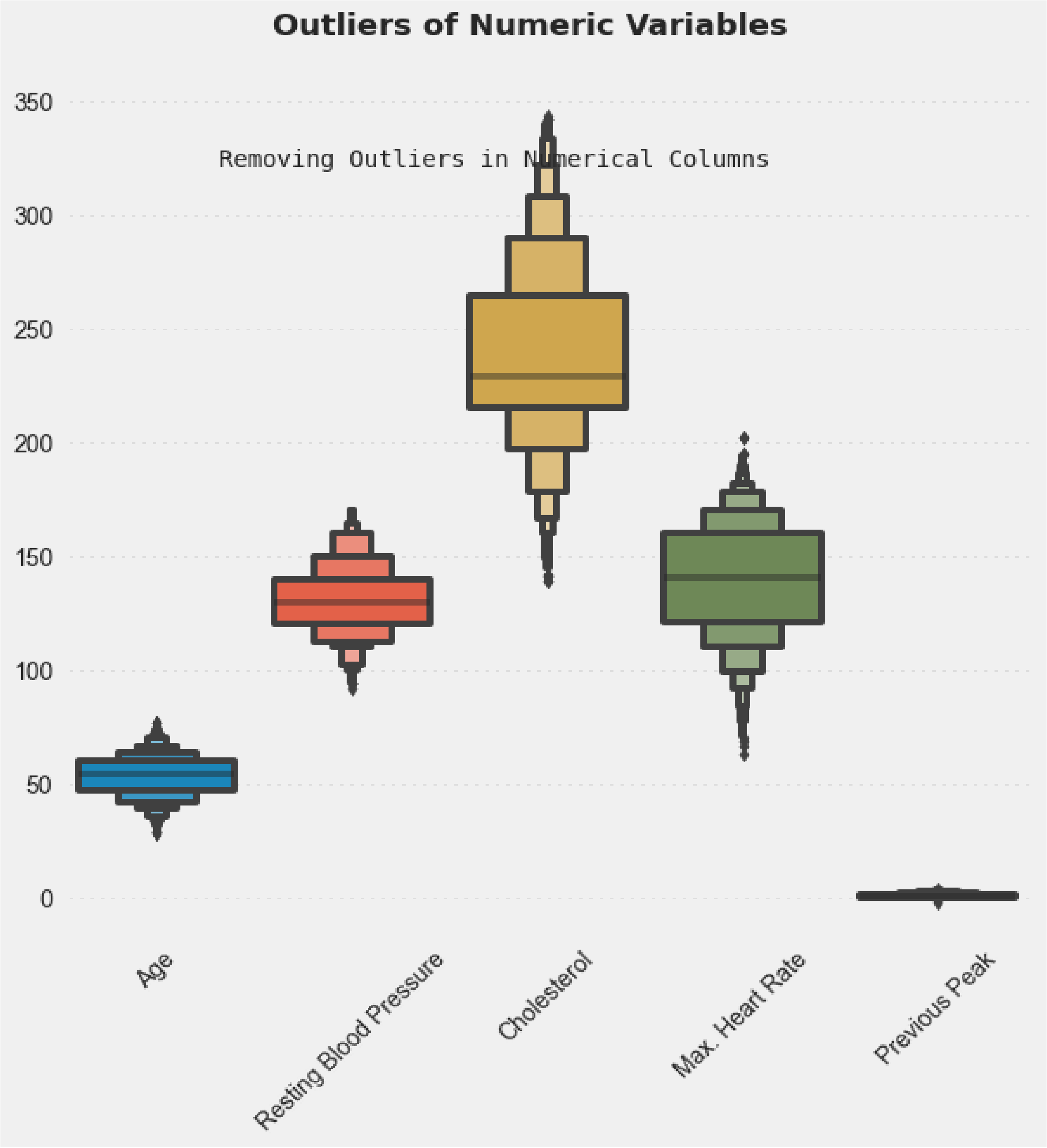


Removal

📌 After detecting, we are using **Median Imputation** to take care of outliers. In this technique, we replace the extreme values with median values.

📌 It is represented by the formula **IQR = Q3 − Q1**. The lines of code below calculate and print the interquartile range for each of the variables in the dataset.

📌 It is advised to not use mean values as they are affected by outliers.



Checking The Normality of Dataset

We need to test whether a sample of observations comes from a normal distribution.

When we'd like to test whether or not a single variable is normally distributed, we can create a q-q plot to visualize the distribution .However, when we'd like to test whether or not *several* variables are normally distributed as a group we must perform a **multivariate normality test**.

For testing multivariate normality, we have used the **Henze-Zirkler (HZ) test,**

The Henze-Zirkler test has a good overall power against alternatives to **normality and works for any dimension and sample size.**

The Henze-Zirkler Multivariate Normality Test determines whether or not a group of variables follows a multivariate normal distribution. The null and alternative hypotheses for the test are as follows:

H_0 (null): The variables follow a multivariate normal distribution.

H_a (alternative): The variables *do not* follow a multivariate normal distribution.

To perform this test in Python we can use the **multivariate_normality()** function from the **pingouin** library.

Output obtained:

```
HZResults(hz=3.2008380572189306, pval=0.0, normal=False
)
```

as **p value is less than 0.05** so we are unable to accept H0 and hence variables **do not follow a multivariate normal distribution.**

❖ Chi-square Test Of Independence :

The Chi-square test of independence determines if there is a significant relationship between two **categorical (i.e. nominal or ordinal) (with two or**

more mutually exclusive groups) variables. Hence, the **null hypothesis is that no relationship exists between the two variables.**

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

- **Non-parametric;** does not require assumptions about population parameters
- **Compares difference** in population proportions between groups.

- **Contingency table of observed** values is required.

It is a test to indicate if the variables are related (i.e. correlated).

1. **H0 (Null Hypothesis) = The 2 variables to be compared are independent.**
2. **H1 (Alternate Hypothesis) = The 2 variables are dependent.**

Here we have used **`chi2_contingency`** function which computes the **chi-square statistic** and **p-value** for the hypothesis test of independence of the observed frequencies in the **contingency table** observed.

➢ First we check that is there any relation between **chest pain type and heart disease (target variable) or not**

Contingency table :

Condition	0	1
Chest Pain Type		
1	41	25
2	185	31
3	193	90
4	142	483

The degrees of freedom (df):

$df=(r-1)\times(c-1)$df=(r−1)×(c−1)

here degrees of freedom = 3 × 1 = 3

we get ,

```
critical=7.815, stat=334.419
```

and ,

```
significance=0.050, p=0.000
```

as p-value that is **less than** to our significance **level i**ndicates that ,

a **relationship exists between the categorical variables.**

So **chest pain type and heart disease (target variable) are dependent on each other.**

- ➢ Now we'll see if is there any relationship between **fasting blood sugar level and heart disease (target variable) or not**

Contingency table :

Condition	0	1

Fasting Blood Sugar		
0	494	442
1	67	187

The degrees of freedom (df):

df=(r−1)×(c−1)df=(r−1)×(c−1)

here degrees of freedom = 1 × 1 = 1

we get ,

```
critical=3.841, stat=54.824
```

and ,

```
significance=0.050, p=0.000
```

as p-value that is **less than** to our significance **level i**ndicates that ,

a **relationship exists between the categorical variables.**

So we can say that ,

Fasting blood Sugar Level and heart disease (target variable) are dependent on each other.

❖ Data Preprocessing :

Train-Test Spli Evaluation :

The **train-test split** is a technique for evaluating the **performance of a machine learning algorithm.**

It can be used for **classification** or regression problems and can be used for any supervised learning algorithm.(In **supervised learning**, input data is provided to the **model along with the output**. In unsupervised learning, only input data is provided to the model. The goal of supervised learning is **to train the model** so that it can predict the output when it is given new data.)

The procedure involves taking a dataset and dividing it into **two subsets**. The first subset is used to fit the model and is referred to as **the training dataset**. The second subset is not used to train the model; instead, the input element of the dataset is provided to the model, then predictions are made and compared to the expected values. This second dataset is referred to as the **test dataset**.

- ➢ **Train Dataset: Used to fit the machine learning model.**
- ➢ **Test Dataset: Used to evaluate the fit machine learning model**.

The objective is to estimate the performance of the machine learning model on new data: data not used to train the model.

- ❖ Here we have taken a **test size as 0.2** which means **20 percent** of the dataset will be allocated to **the test set** and **80 percent** will be allocated to the **training set**.

After splitting we get ,

```
Number transactions x_train dataset:  (952, 11)
Number transactions y_train dataset:  (952,)
Number transactions x_test dataset:  (238, 11)
Number transactions y_test dataset:  (238,)
```

We can see that 952 examples (80 percent) are allocated to the training set and 238 examples (20 percent) are allocated to test set , as we specified.

❖ Feature Scaling With Standardscalar Function :

- The goal of applying Feature Scaling is to make sure features are on almost the same scale so that each feature is equally important and make it easier to process by most ML algorithms.

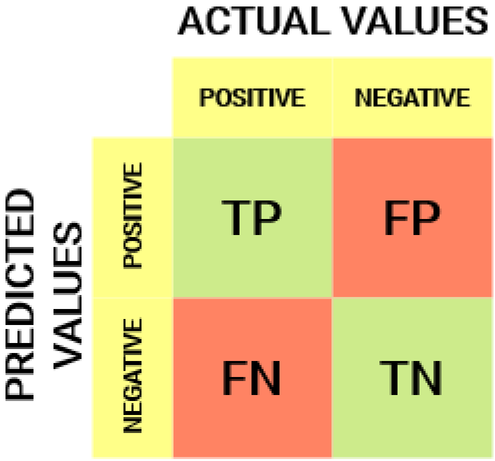
- StandardScaler standardizes a feature by subtracting the mean and then scaling to unit variance.

- StandardScaler results in a distribution with a standard deviation equal to 1 StandardScaler does distort the relative distances between the feature values.

❖ Confusion Matrix :

A **Confusion matrix** is an N x N matrix used for evaluating the **performance of a classification mode**l, where N is the number of target classes. The matrix **compares the actual target values with those predicted by the machine learning model.**

For a binary classification problem, we would have a 2 x 2 matrix as shown below with 4 values:



		ACTUAL VALUES	
		POSITIVE	NEGATIVE
PREDICTED VALUES	POSITIVE	TP	FP
	NEGATIVE	FN	TN

- **True Positive (TP) :**

The predicted value matches the actual value

- **True Negative (TN)**

The predicted value matches the actual value

- **False Positive (FP) – Type 1 error**

The predicted value was falsely predicted

The actual value was negative but the model predicted a positive value

Also known as the Type 1 error

- **False Negative (FN) – Type 2 error**

The predicted value was falsely predicted

The actual value was positive but the model predicted a negative value

Also known as the Type 2 error

- **Accuracy :**

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

- **Precision :** Precision tells us how many of the correctly predicted cases actually turned out to be positive

$$Precision = \frac{TP}{TP + FP}$$

❖ Logistic Regression :

📌 **Logistic Regression** assumes a Gaussian distribution for the numeric input variables and can model binary classification problems. You can construct a logistic regression model using the LogisticRegression class.

Confusion matrix :

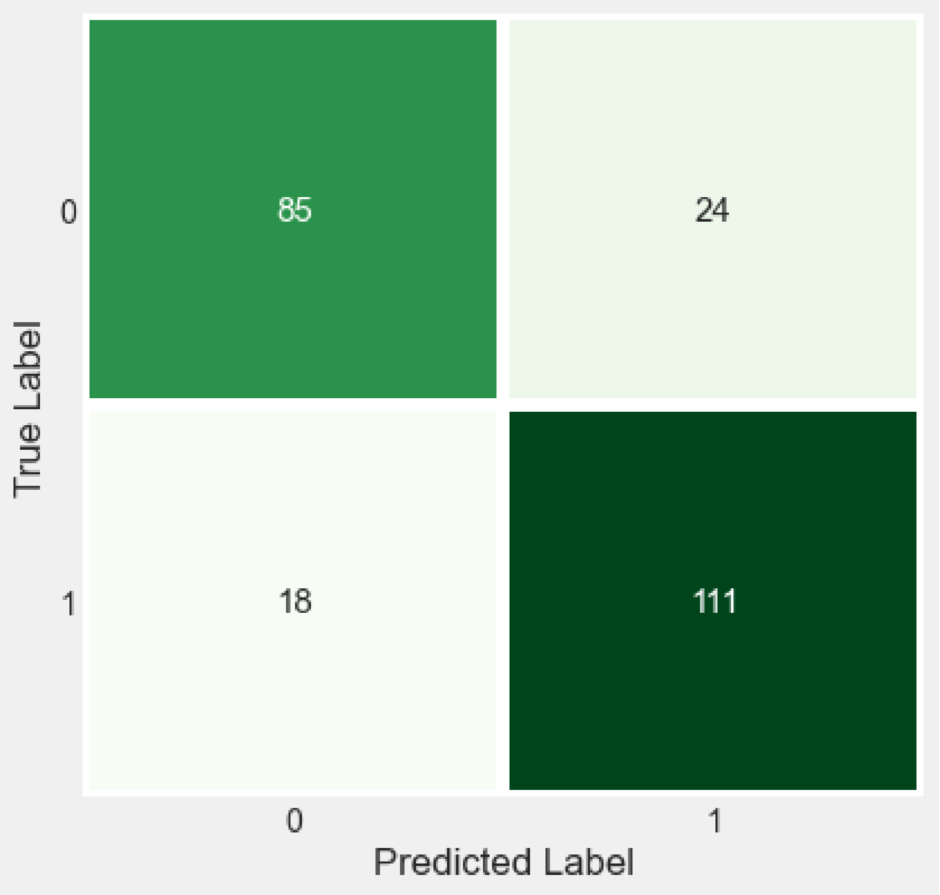
```
precision    recall  f1-score   support

           0       0.83      0.78      0.80       109
           1       0.82      0.86      0.84       129

    accuracy                           0.82       238
   macro avg       0.82      0.82      0.82       238
weighted avg       0.82      0.82      0.82       238

Accuracy Score:  0.8235294117647058
```

Visualizing Confusion matrix :



❖ K-Nearest Neighbors :

📌 The **k-nearest neighbors (KNN)** algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems

Confusion matrix :

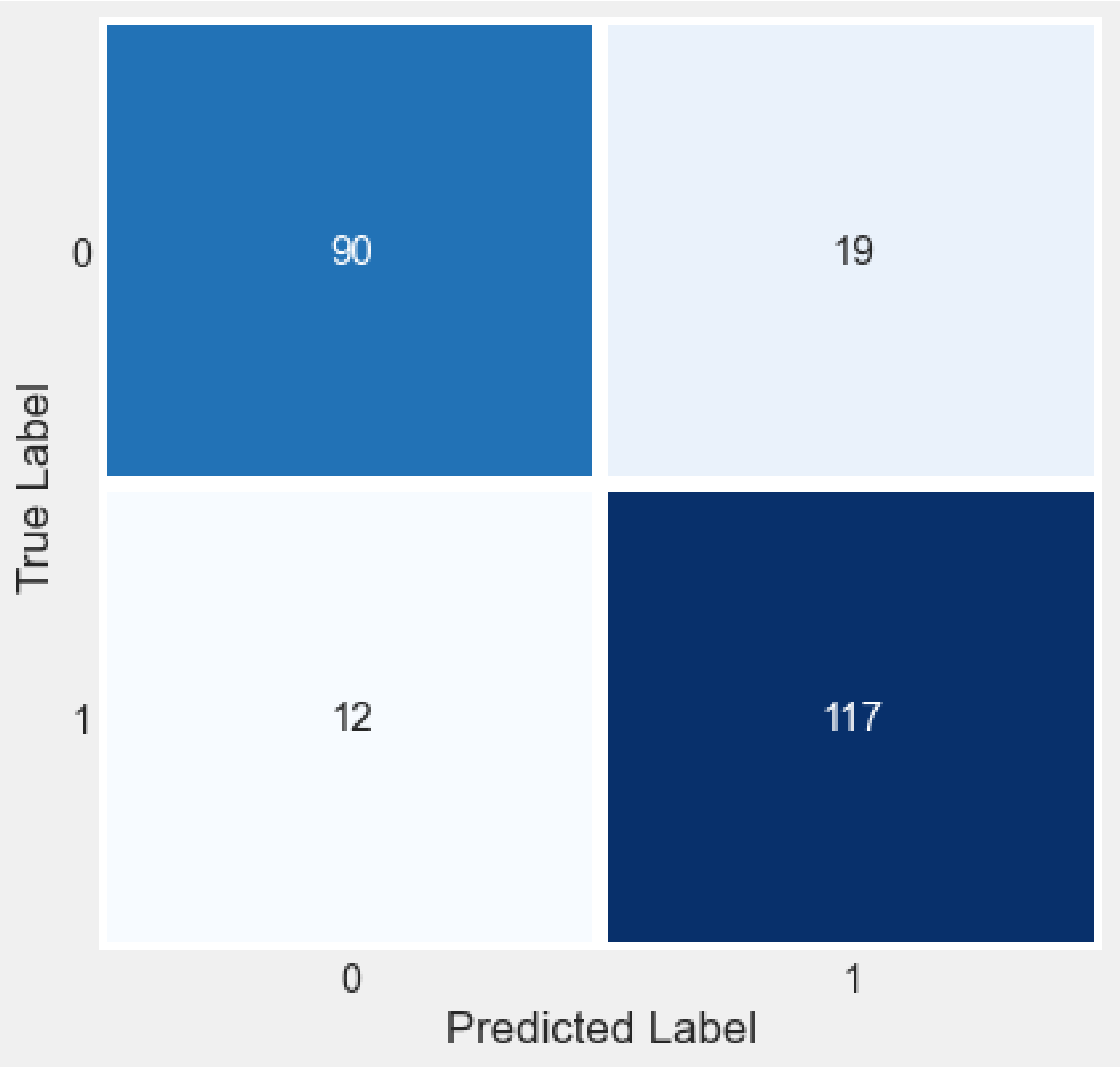
```
precision    recall  f1-score   support

           0       0.88      0.83      0.85       109
           1       0.86      0.91      0.88       129

    accuracy                           0.87       238
   macro avg       0.87      0.87      0.87       238
weighted avg       0.87      0.87      0.87       238

Accuracy Score:  0.8697478991596639
```

Visualizing Confusion matrix :



❖ Naive Bayes Classifier :

📌 **A Gaussian Naive Bayes** algorithm is a special type of NB algorithm. It's specifically used when the features have continuous values. It's also assumed that all the features are following a gaussian distribution i.e, normal distribution.

Confusion matrix :

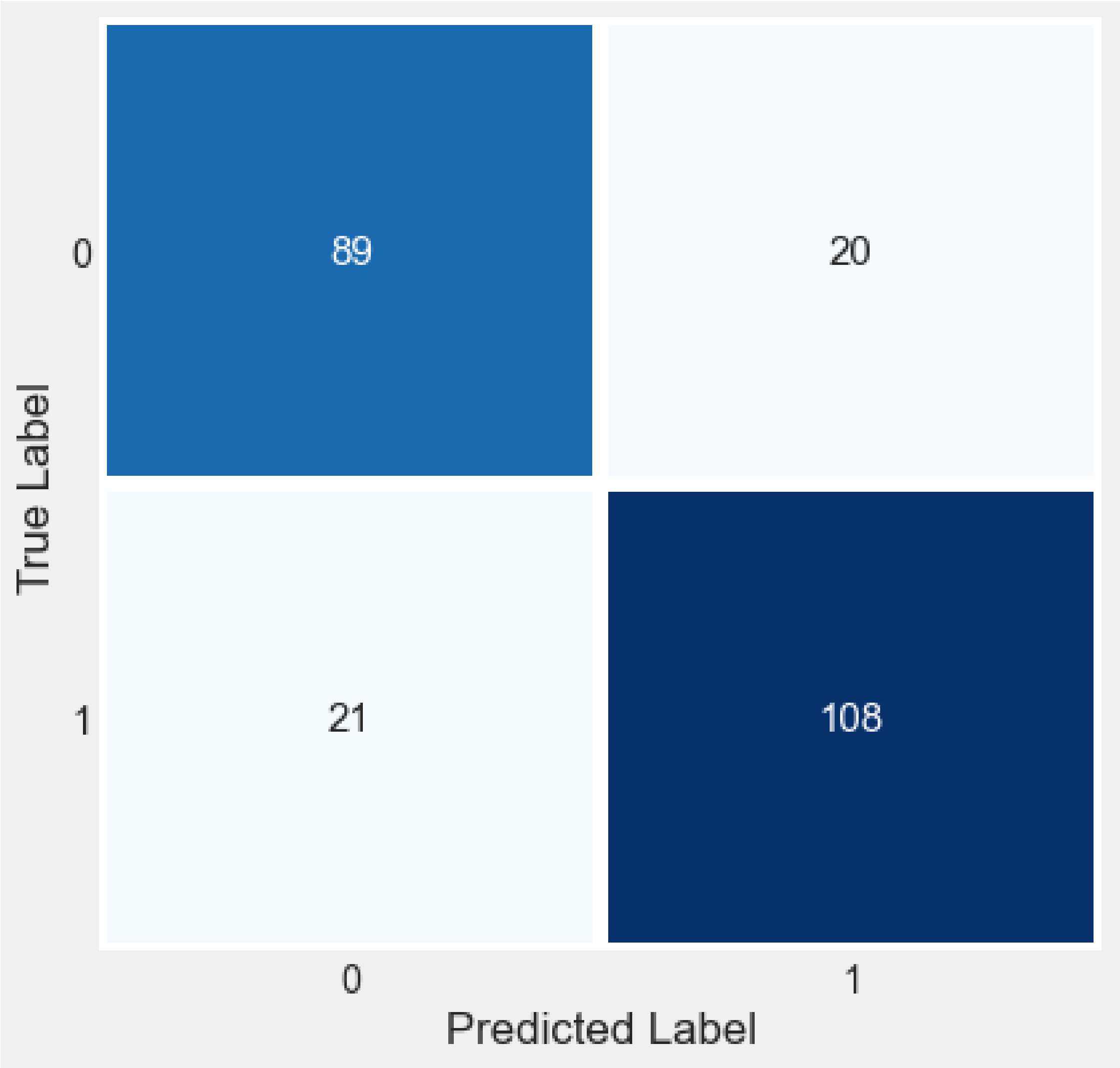
```
precision    recall  f1-score   support

           0       0.81      0.82      0.81       109
           1       0.84      0.84      0.84       129

    accuracy                           0.83       238
   macro avg       0.83      0.83      0.83       238
weighted avg       0.83      0.83      0.83       238

Accuracy Score:  0.8277310924369747
```

Visualizing Confusion matrix :



❖ Decision Tree (CART) :

📌 **Classification and Regression Trees (CART** or just decision trees) construct a binary tree from the training data. Split points are chosen greedily by evaluating each attribute and each value of each attribute in the training data in order to minimize a cost function (like the Gini index). You can construct a CART model using the DecisionTreeClassifier class

Confusion matrix :

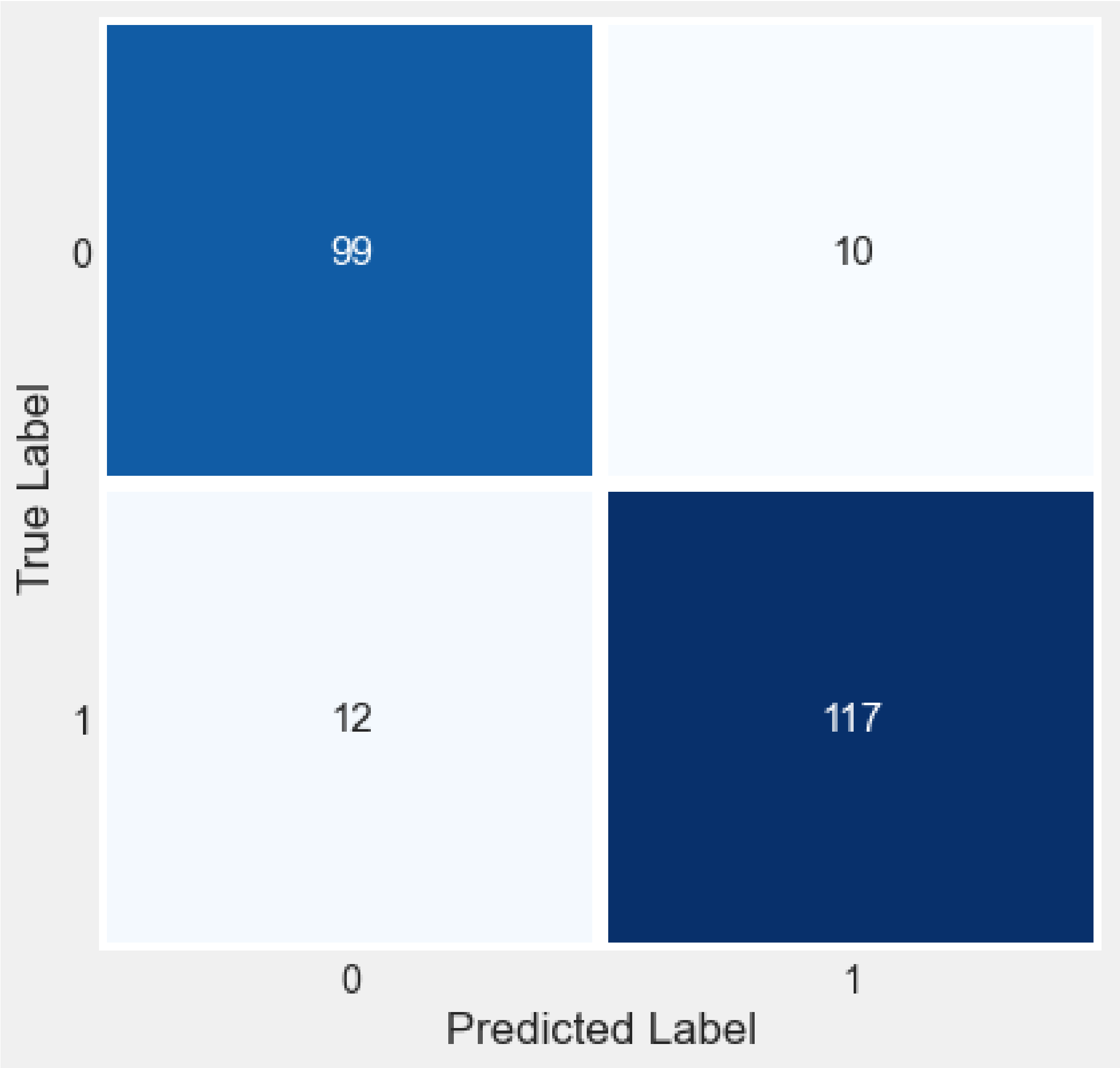
```
precision    recall  f1-score   support

           0       0.89      0.91      0.90       109
           1       0.92      0.91      0.91       129

    accuracy                           0.91       238
   macro avg       0.91      0.91      0.91       238
weighted avg       0.91      0.91      0.91       238

Accuracy Score:  0.907563025210084
```

Visualizing Confusion matrix :



❖ Random Forest :

📌 **Random Forests** is an extension of bagged decision trees. Samples of the training dataset are taken with replacement, but the trees are constructed in a way that reduces the correlation between individual classifiers. Specifically, rather than greedily choosing the best split point in the construction of each tree, only a random subset of features are considered for each split. You can construct a Random Forest model for classification using the RandomForestClassifier class.

Confusion matrix :

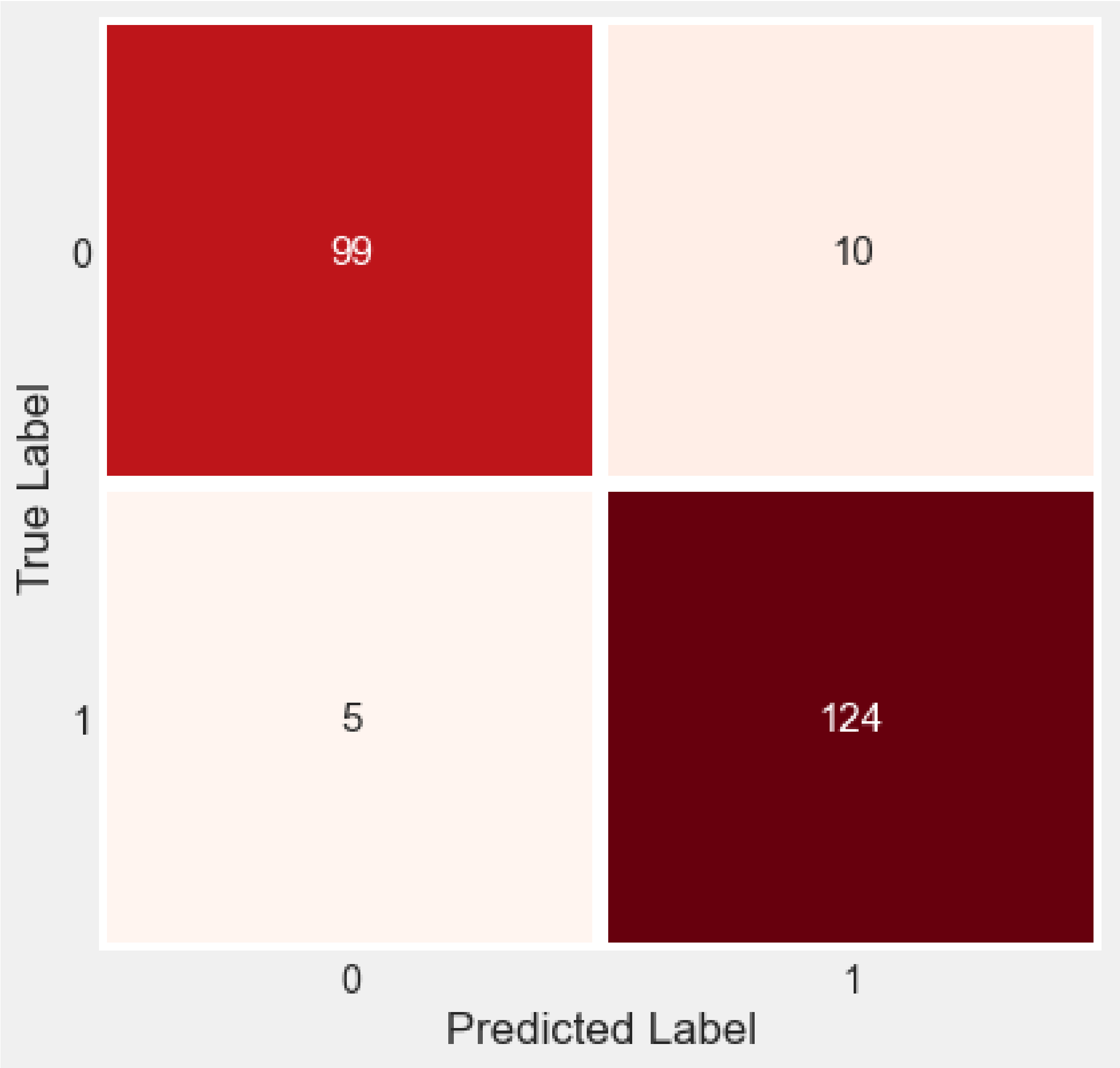
```
precision    recall  f1-score   support

           0       0.95      0.91      0.93       109
           1       0.93      0.96      0.94       129

    accuracy                           0.94       238
   macro avg       0.94      0.93      0.94       238
weighted avg       0.94      0.94      0.94       238

Accuracy Score:  0.9369747899159664
```

Visualizing Confusion matrix :



❖ SVM :

📌 **Support Vector Machines (or SVM)** seek a line that best separates two classes. Those data instances that are closest to the line that best separates the classes are called support vectors and influence where the line is placed. SVM has been extended to support multiple classes Of particular importance is the use of different kernel functions via the kernel parameter .A powerful Radial Basis Function is used by default. You can construct an SVM model using the SVC class.

Confusion matrix :

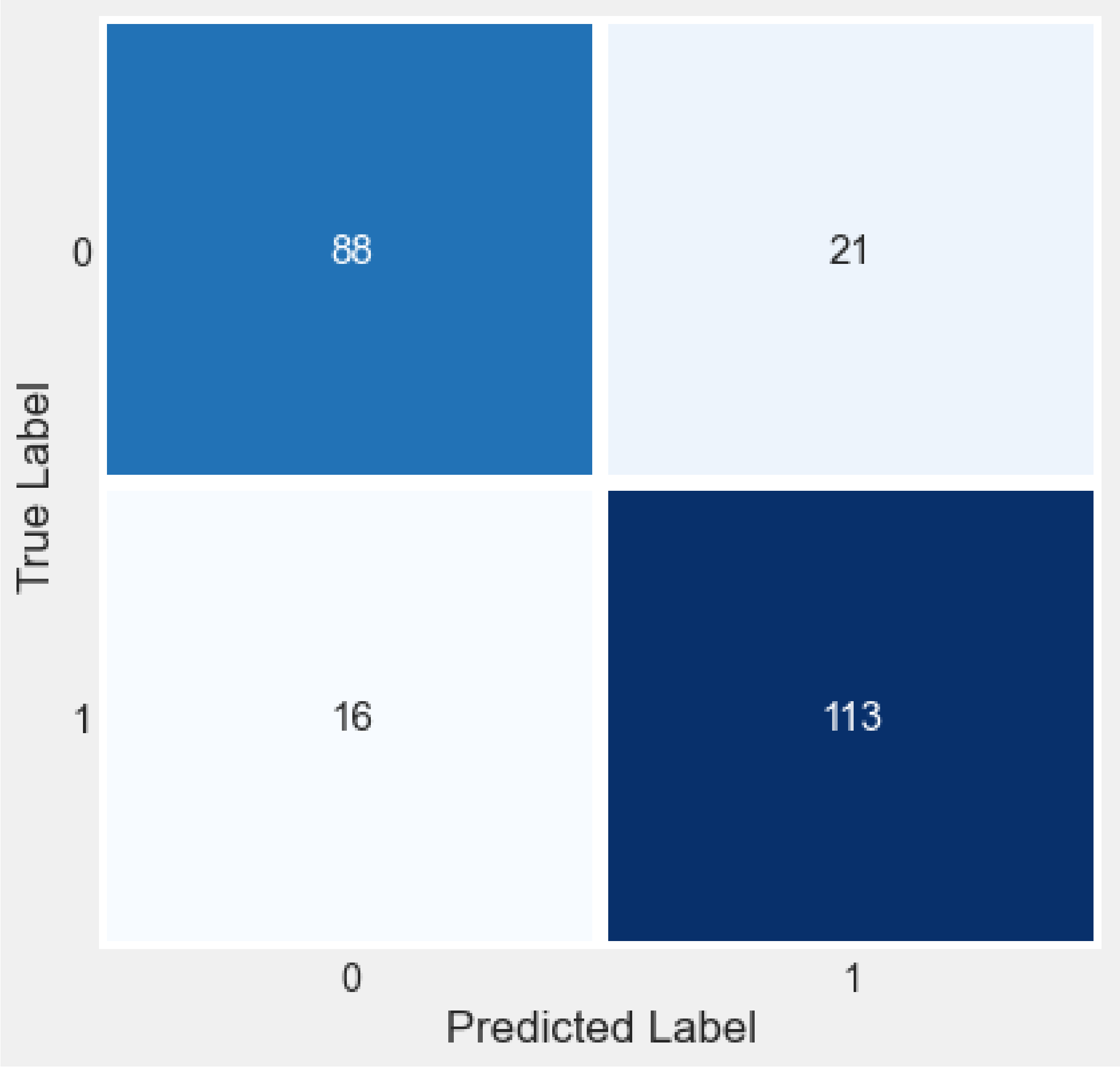
```
precision    recall  f1-score   support

           0       0.85      0.81      0.83       109
           1       0.84      0.88      0.86       129

    accuracy                           0.84       238
   macro avg       0.84      0.84      0.84       238
weighted avg       0.84      0.84      0.84       238

Accuracy Score:  0.844537815126050
```

Visualizing Confusion matrix :



❖ XGBoost :

📌 **XGBoost stands for Extreme Gradient Boosting**, it is a performant machine learning library based on the paper Greedy Function Approximation: A Gradient Boosting Machine, by Friedman. XGBoost implements a Gradient Boosting algorithm based on decision trees.

Confusion matrix :

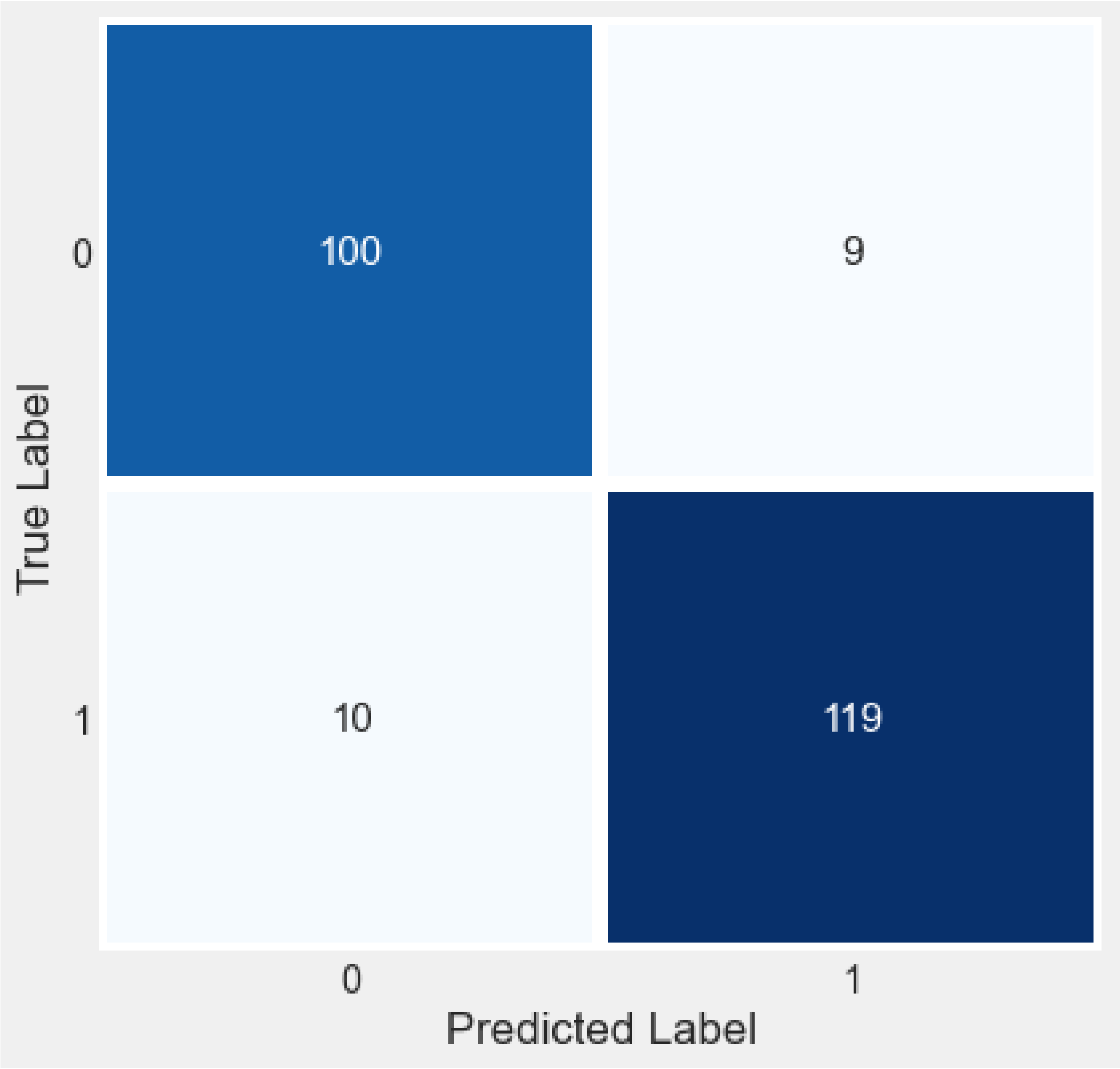
```
precision    recall  f1-score   support

           0       0.91      0.92      0.91       109
           1       0.93      0.92      0.93       129

    accuracy                           0.92       238
   macro avg       0.92      0.92      0.92       238
weighted avg       0.92      0.92      0.92       238

Accuracy Score:  0.9201680672268907
```

Visualizing Confusion matrix :



- **Dataframe of different ML classification models and respective accuracy percentages :**

	Model	Accuracy Percentage
1	Logistic Regression	82.3529
2	K-Nearest Neighbors	86.9747
3	Naive Bayes Classifier	82.7731
4	Decision Tree (CART)	90.7563
5	Random Forest	93.6974
6	SVM	84.4537
7	XGBoost	92.0168

📌 From above we can observe that the **Random Forest classifier model** is giving us the **maximum accuracy of almost 93 percent** followed by the **XGBoost** which is giving us the accuracy of **92 percent**

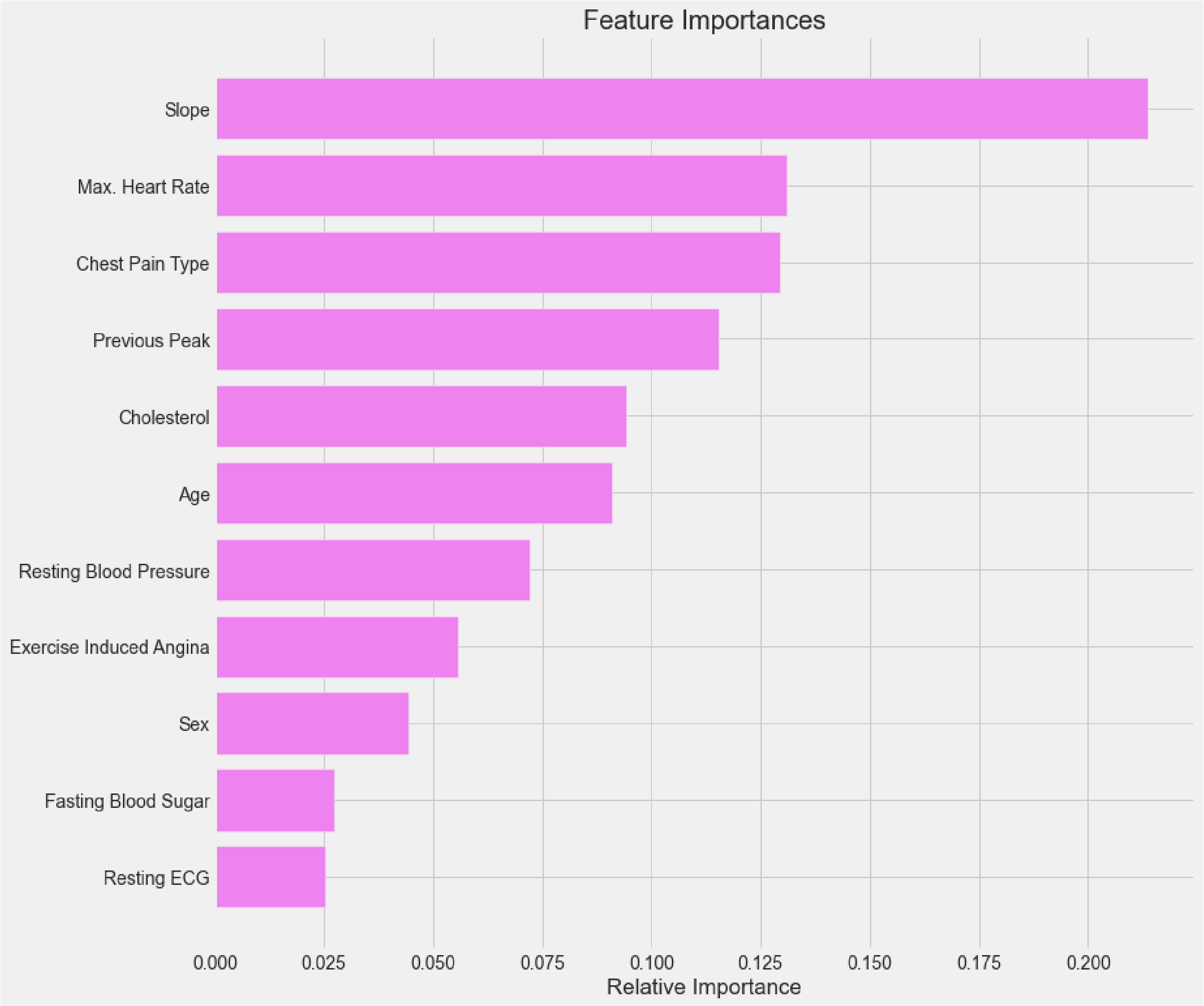
📌 Hence we will use **Random Forest classifier** for the **prediction of Heart Attack.**

📌Also now we will fit the **Random Forest classifier** model for performing **feature importance** which will give us the most important feature in our prediction model.

❖ Feature Importance :

Feature importance refers to techniques that **assign a score to input features** based on how useful they are at predicting a target variable.

Most importance scores are calculated by a predictive model that has been fit on the dataset , (**as Random Forest model gives the maximum accuracy we will use it as a predictive model**). Inspecting the importance score provides insight into that specific model and which features are the most important and least important to the model when making a prediction.



📌 From above we can observe that the **slope of the peak exercise ST segment is the most important feature.**

And **chest pain is the second most important feature** in our data.

Conclusion

Heart diseases have become more and more frequent among people. Therefore, predicting the disease before becoming infected decreases the risk of death . The main contribution of this study was a **comparison of various ML algorithms for prediction of Heart Disease** at the early stages. It is based on the application of Machine Learning algorithms, of which we have chosen the 7 most used algorithms (**Logistic Regression , Gaussian NB, Random Forest , KNN , Decision Tree (CART) ,SVM and XGBoost**), on our data set, where we had very good results.

We arrived at **93% of accuracy with Random Forest** . The strong point of our study, we tested the stability of the algorithm on different sizes of our data set, we noticed at the end that **Random Forest** gives the best results. Also, we made a study on the **features importance** , to see which is the most important feature for predicting heart disease which comes out to be **slope of the peak exercise ST segment**, also we used the **correlation matrix to detect the dependencies between the attributes.**

The future scope of this study can be extended by using **deep learning, fuzzy etc.** and by **taking large primary data** in order to find the exact patterns of knowledge.

References

1. Improving the Accuracy for Analyzing Heart Diseases Prediction Based on the Ensemble Method.

   Link : https://doi.org/10.1155/2021/6663455

2. R. Kavitha and E. Kannan et al. “An Efficient Framework for Heart Disease Classification using Feature Extraction and Feature.

3. “A PROGNOSIS ON CARDIAC INFARCTION USING IMPLIED DATA MINING CLASSIFICATION ALGORITHMS” by S.Padma , K.Yasudha

4. Himanshu Sharma,M A Rizvi Prediction of Heart Disease using Machine Learning Algorithms: A Survey (August 2017)

5. Blake, C.L., Mertz, C.J.: “UCI Machine Learning Databases”, http://mlearn.ics.uci.edu/databases/heart-disease/, 2004.