PROJECT REPORT
ON
Heart Attack Analysis and Prediction

Submitted by

Ayush Rathi (230920525004)

Deepanshu Mahajan (230920525005)

Shubham Lanjewar (230920525014)

Under the Guidance of

MS. Priti Bhardwaj

CDAC-NOIDA

Candidate Declaration

We hereby declare that the Project Report entitled "**Heart Attack Analysis and Prediction**" System Using Machine Learning done by us under the guidance of Ms. Priti Bhardwaj mam is submitted in partial fulfillment of the requirements for the award of Post graduation diploma in Big Data Analytics.

DATE: 26-02-2024

PLACE: Noida

Acknowledgement

We would like to thank Ms. Priti Bhardwaj Mam for her help through this project. Her suggestions are the keys to the successful completion of this project and to understand the basic of analysis and design of algorithms which is the most important factor behind implementing efficient code. We would also like to thank Mr. Shivam Pandey Sir and Ms. Siddhidatri Nayak Mam for giving us suggestions to improve our project.

Abstract

Day by day the cases of heart diseases are increasing at a rapid rate and it's very Important and concerning to predict any such diseases beforehand. This diagnosis is a difficult task i.e. it should be performed precisely and efficiently. The research paper mainly focuses on which patient is more likely to have a heart disease based on various medical attributes. We prepared a heart disease prediction system to predict whether the patient is likely to be diagnosed with a heart disease or not using the medical history of the patient. We used different algorithms of machine learning such as logistic regression and Random Forest to predict and classify the patient with heart disease. A quite Helpful approach was used to regulate how the model can be used to improve the accuracy of prediction of Heart Attack in any individual. The strength of the proposed model was quiet satisfying and was able to predict evidence of having a heart disease in a particular individual by using Random Forest and Logistic Regression which showed a good accuracy in comparison to the previously used classifier such as Decision Tree etc. So a quiet significant amount of pressure has been lift off by using the given model in finding the probability of the classifier to correctly and accurately identify the heart disease. The Given heart disease prediction system enhances medical care and reduces the cost. This project gives us significant knowledge that can help us predict the patients with heart disease.

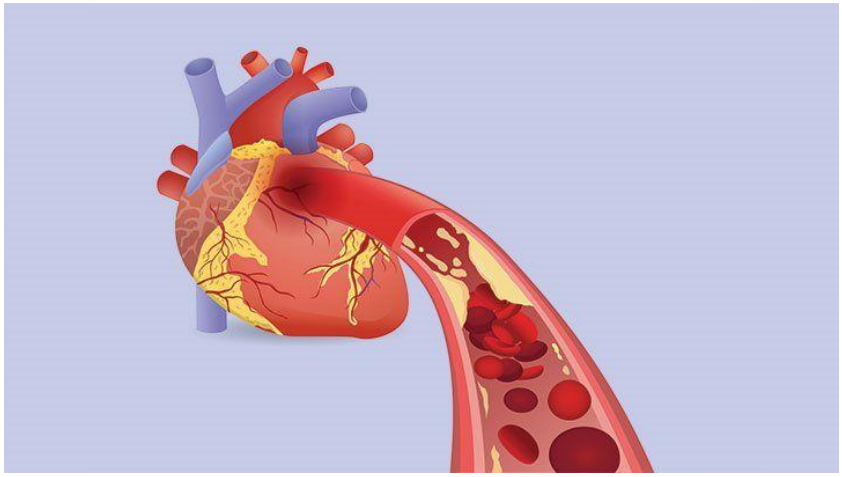
TABLE OF CONTENTS

1. INTRODUCTION

Heart disease remains one of the leading causes of mortality worldwide, with heart attacks, or myocardial infarctions (MI), representing a critical manifestation of this pervasive health concern. A heart attack occurs when blood flow to a section of the heart muscle becomes obstructed, leading to tissue damage or necrosis. Despite advancements in medical science and public health education, the incidence of heart attacks continues to pose a significant burden on individuals, families, and healthcare systems globally.

- The medical name of heart attack is “Myocardial infarction”.
- Heart attack in short; It is the occlusion of the vessel by plaque-like lesions filled with cholesterol and fat.
- The lesion is called abnormal conditions that occur in the organs where the disease is located.
- As a result of the blockage, the blood flow is completely cut off and a heart attack that can lead to death occurs.

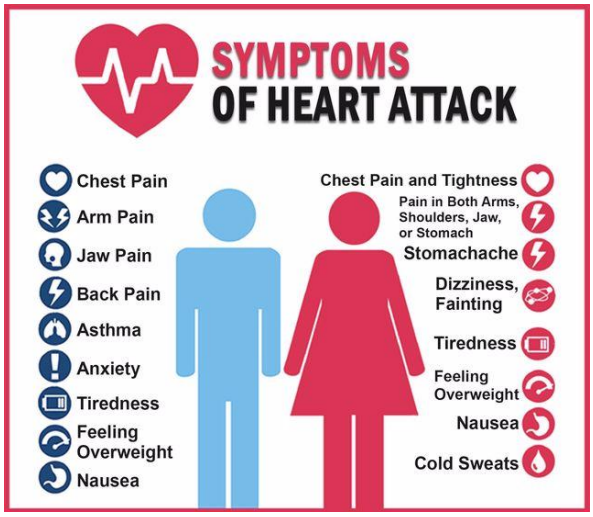
How does a heart attack occur?



- The heart is a powerful pump that pumps blood throughout the body 60-80 times per minute at rest.
- While meeting the blood needs of the whole body, it also needs to be fed and taken blood.
- These vessels that feed the heart itself are called coronary arteries.
- Coronary insufficiency occurs when there is a disruption in the circulation of the coronary arteries.
- The cases of coronary insufficiency vary according to the type, degree and location of the stenosis in the coronary vessels.

- While some patients may have chest pain that occurs only during physical activity and is relieved by rest, sometimes a heart attack may occur as a result of sudden occlusion of the vessels, starting with severe chest pain and leading to sudden death.

What are the symptoms of a heart attack?



So in our project , we want to predict whether patients have heart disease by given some features of users. This is important to medical fields. If such a prediction is accurate enough, we can not only avoid wrong diagnosis but also save human resources. When a patient without a heart disease is diagnosed with heart disease, he will fall into unnecessary panic and when a patient with heart disease is not diagnosed with heart disease, he will miss the best chance to cure his disease. Such wrong diagnosis is painful to both patients and hospitals. With accurate predictions, we can solve the unnecessary trouble. Besides, if we can apply our machine learning tool into medical prediction, we will save human resource because we do not need complicated diagnosis process in hospitals. (though it is a very long way to go.) The input to our algorithm is 13 features with number values. We use several algorithms such as Logistic Regression, Decision Tree, Random Forest, to output a binary number 1 or 0. 1 indicates the patient has heart disease and vice versa.

2. Project Overview

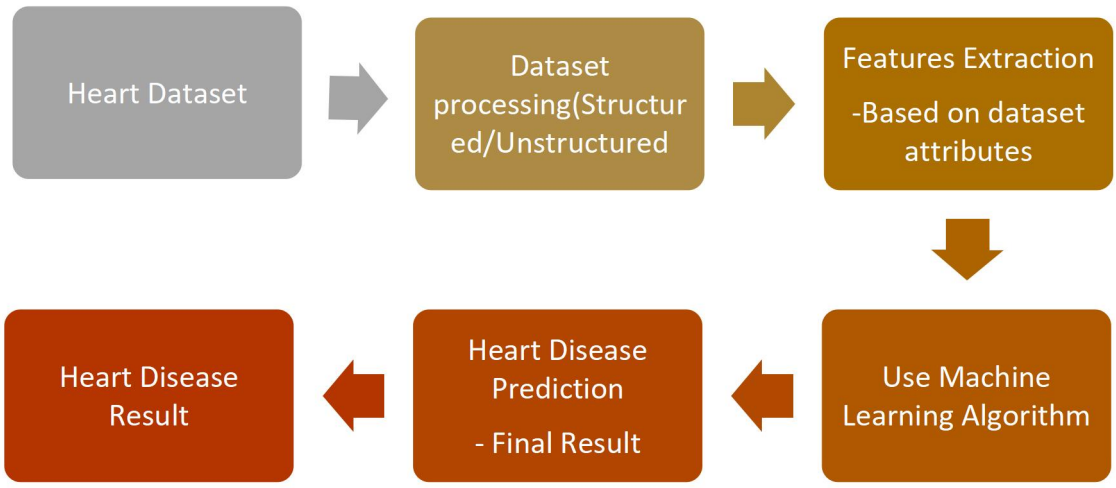
This work analyses the predictive system for heart disease. In this work, medical terms like sex, blood pressure, and cholesterol are used to describe the possibility of heart disease in patients.

Heart disease predictor is designed and developed to explore the path of machine learning. The goal is to predict the health of the patient from collective data to be able to detect configurations at risk for the patient, and therefore, in cases requiring emergency medical assistance, alert the appropriate medical staff of the situation of the latter. We initially have a dataset collecting information of many patients with which we can conclude the results into a complete form and can predict data precisely. The results of the predictions, derived from the predictive models generated by machine learning, will be presented through several distinct graphical interfaces according to the datasets considered.

Data has been collected from Kaggle. Data collection is the process of gathering and measuring information from countless different sources to use the data.

Key components of the project include data collection, pre-processing, data fetching, visualization, and predictive modelling.

3. WORK FLOW OF THE PROJECT



4. DATA COLLECTION/FETCH

The dataset used for this project purpose was the Public Health Dataset and consists of 14 attributes, including the predicted attribute. The “target” field refers to the presence of heart disease in the patient. It is integer-valued 0 = no disease and 1 = disease. Now the attributes which are used in this project purpose are described as follows and for what they are used or resemble:

- Age—age of patient in years.
- sex—(1 = male; 0 = female).
- Cp—chest pain type.
- Trtbps—resting blood pressure (in mm Hg on admission to the hospital). The normal range is 120/80 (if you have a normal blood pressure reading, it is fine, but if it is a little higher than it should be, you should try to lower it. Make healthy changes to your lifestyle).
- Chol—serum cholesterol shows the amount of triglycerides present. Triglycerides are another lipid that can be measured in the blood. It should be less than 170 mg/dL (may differ in different Labs).
- Fbs—fasting blood sugar larger than 120 mg/dl (1 = true, 0 = False). Less than 100 mg/dL (5.6 mmol/L) is normal, and 100 to 125 mg/dL (5.6 to 6.9 mmol/L) is considered prediabetes.
- Restecg—resting electrocardiographic results.

    - Value 0: normal
    - Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
    - Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria

- Thalach—maximum heart rate achieved. The maximum heart rate is 220 minus your age.

• Exang—exercise-induced angina (1 yes). Angina is a type of chest pain caused by reduced blood flow to the heart. Angina is a symptom of coronary artery disease.

• Oldpeak—ST depression induced by exercise relative to rest.

• Slope—the slope of the peak exercise ST segment.

• Ca—number of major vessels (0–3) colored by fluoroscopy.

• Thal—no explanation provided, but probably thallium stress (2 normal; 1 fixed defects; 3 reversible defects).

• Target (T)—no disease = 0 and disease = 1, (angiographic disease status).

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	ca	thal	output
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1298	48	1	2	139	349	0	2	183	1	5.6	2	2	3	1
1299	47	1	3	143	258	1	1	98	1	5.7	1	0	3	0
1300	69	1	0	156	434	1	0	196	0	1.4	3	1	3	1
1301	45	1	1	186	417	0	1	117	1	5.9	3	2	2	1
1302	25	1	0	158	270	0	0	143	1	4.7	0	0	3	0

1303 rows × 14 columns

5. DATA PRE-PROCESSING

Data pre-processing and storage are essential steps in our project to ensure that we work with clean and structured data.Data pre-processing includes cleaning and structuring data to ensure it's ready for analysis.

• The dataset does not have any null values. But many outliers needed to be handled properly, and also the dataset is not properly distributed. Two approaches were used.

• One without outliers and feature selection process and directly applying the data to the machine learning algorithms, and the results which were achieved were not promising.

• But after using the normal distribution of dataset for overcoming the overfitting problem and then applying Isolation Forest for the outlier's detection, the results achieved are quite promising.

• Various plotting techniques were used for checking the skewness of the data, outlier detection, and the distribution of the data. All these preprocessing techniques play an important role when passing the data for classification or prediction purposes.

Checking the Distribution of the Data :

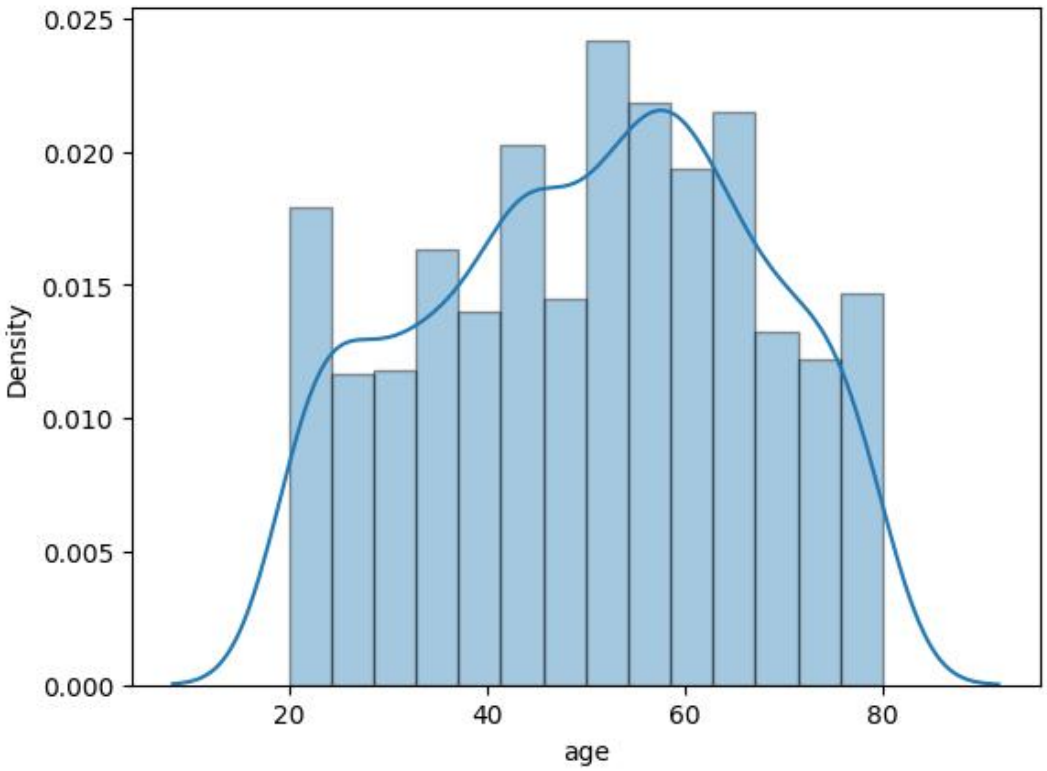
• The distribution of the data plays an important role when the prediction or classification of a problem is to be done.

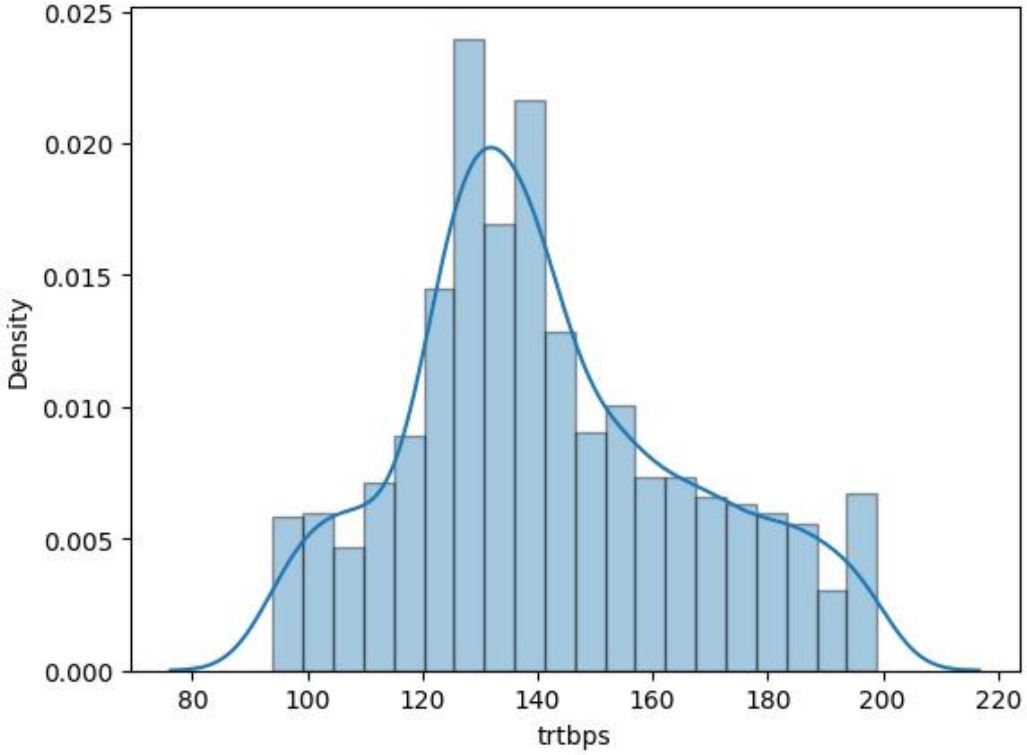
• We see that the heart disease occurred 54.46% of the time in the dataset, whilst 45.54% was the no heart disease.

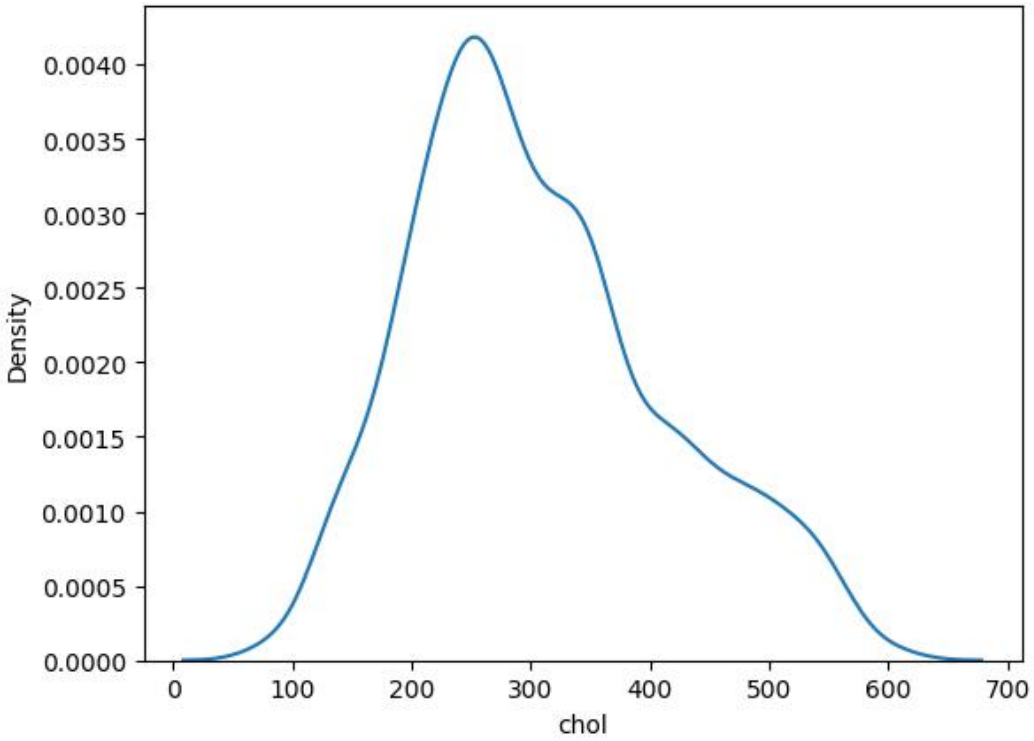
• So, we need to balance the dataset or otherwise it might get overfit. This will help the model to find a pattern in the dataset that contributes to heart disease.

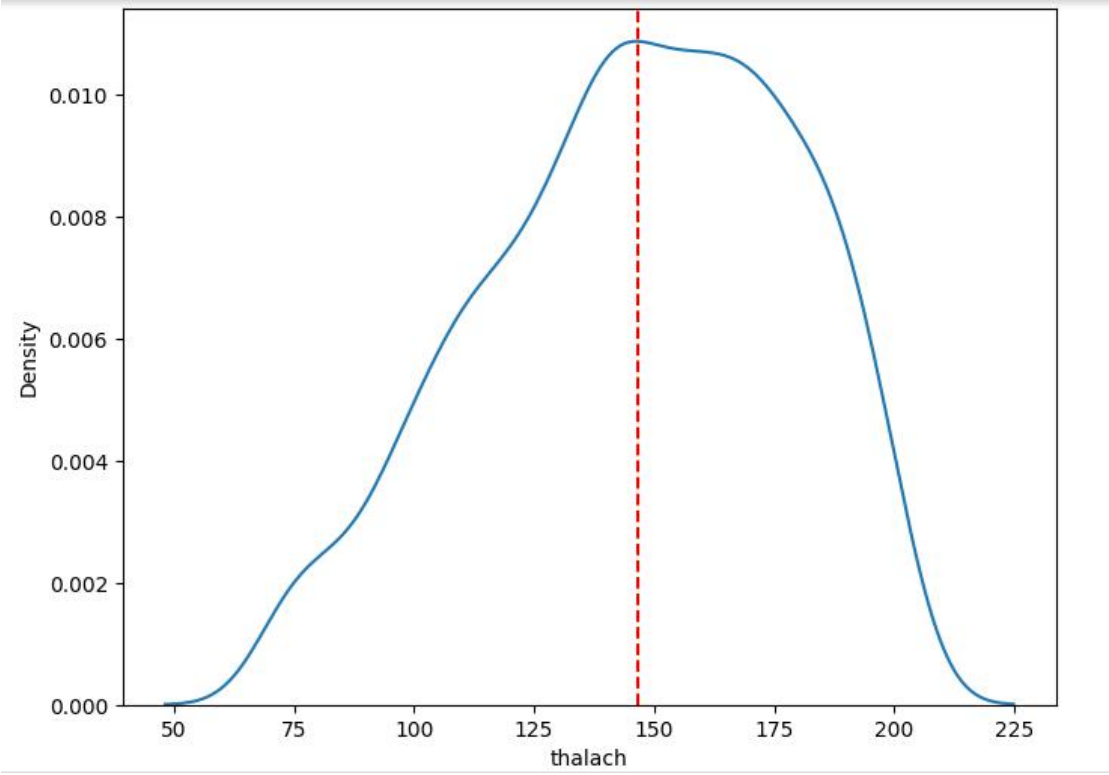
Checking the Skewness of the Data

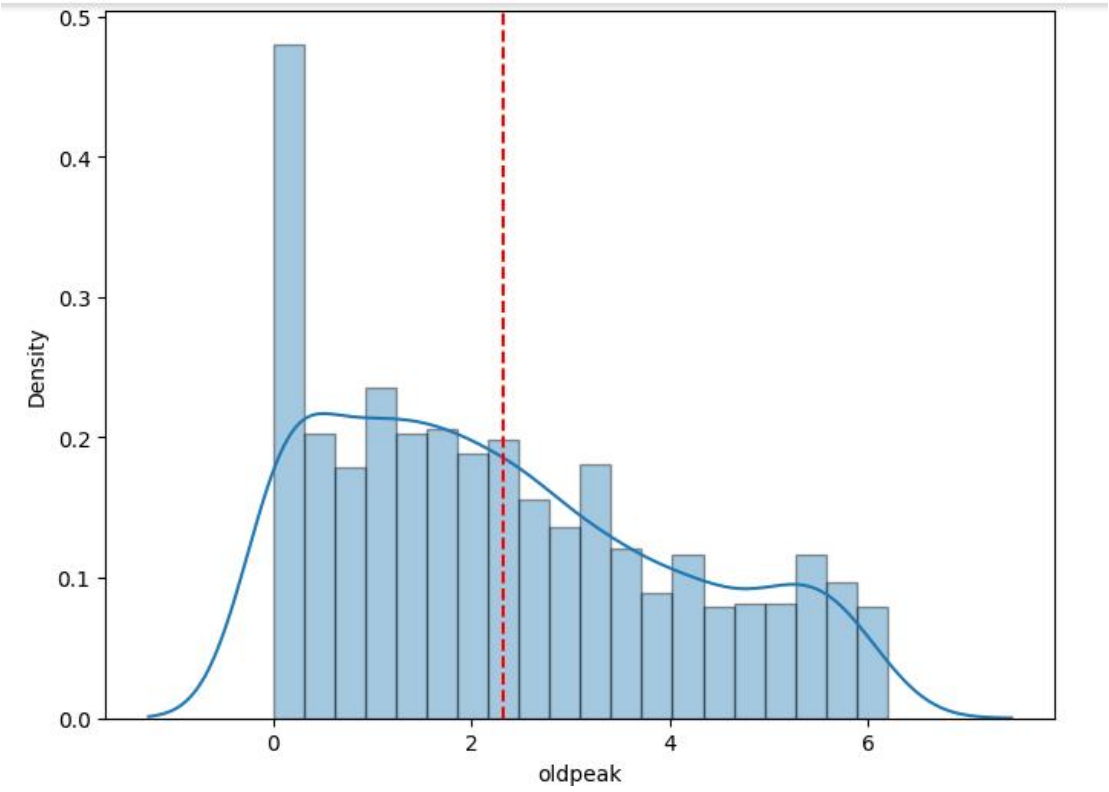
For checking the attribute values and determining the skewness of the data (the asymmetry of a distribution), many distribution plots are plotted so that some interpretation of the data can be seen. Different plots are shown, so an overview of the data could be analyzed. The distribution of age and sex, the distribution of chest pain and trestbps, the distribution of cholesterol and fasting blood, the distribution of ecg resting electrode and thalach, the distribution of exang and oldpeak, the distribution of slope and ca, and the distribution of thal and target all are analyzed and the conclusion.


0.025
0.020
0.015
0.010
0.005
0.000
Density
20
40
60
80
age


0.025
0.020
0.015
0.010
0.005
0.000
Density
80
100
120
140
160
180
200
220
trtbps


0.0040
0.0035
0.0030
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
Density
0
100
200
300
400
500
600
700
chol


0.010
0.008
0.006
0.004
0.002
0.000
Density
50
75
100
125
150
175
200
225
thalach



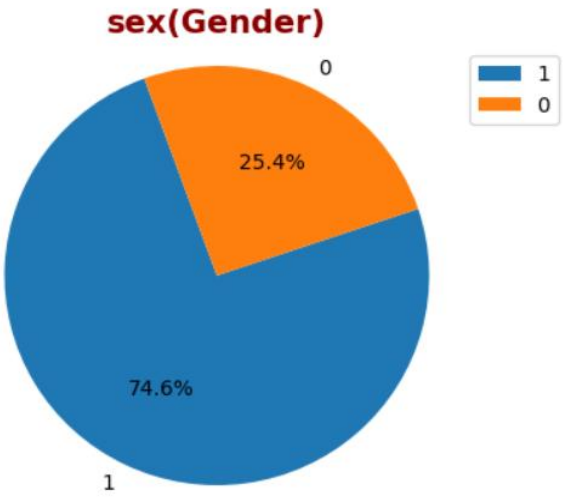
6. EXPLORATORY DATA ANALYSIS

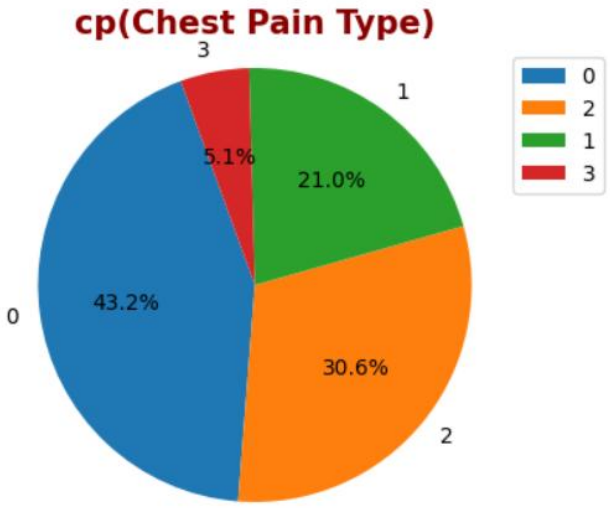
Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods.

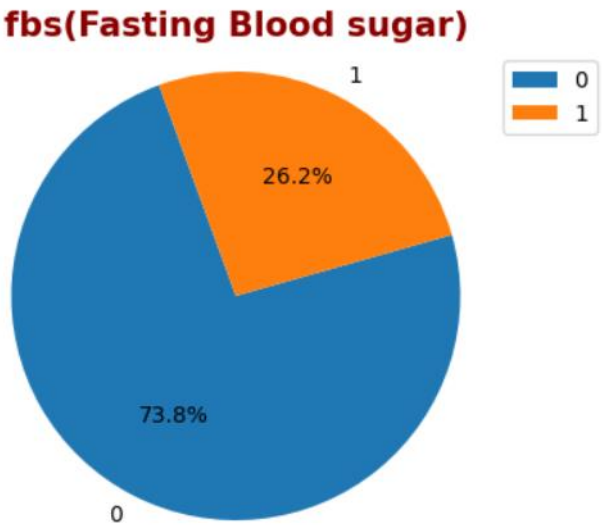
EDA helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions.

The main purpose of EDA is to help look at data before making any assumptions. It can help identify obvious errors, as well as better understand patterns within the data, detect outliers or anomalous events, find interesting relations among the variables.

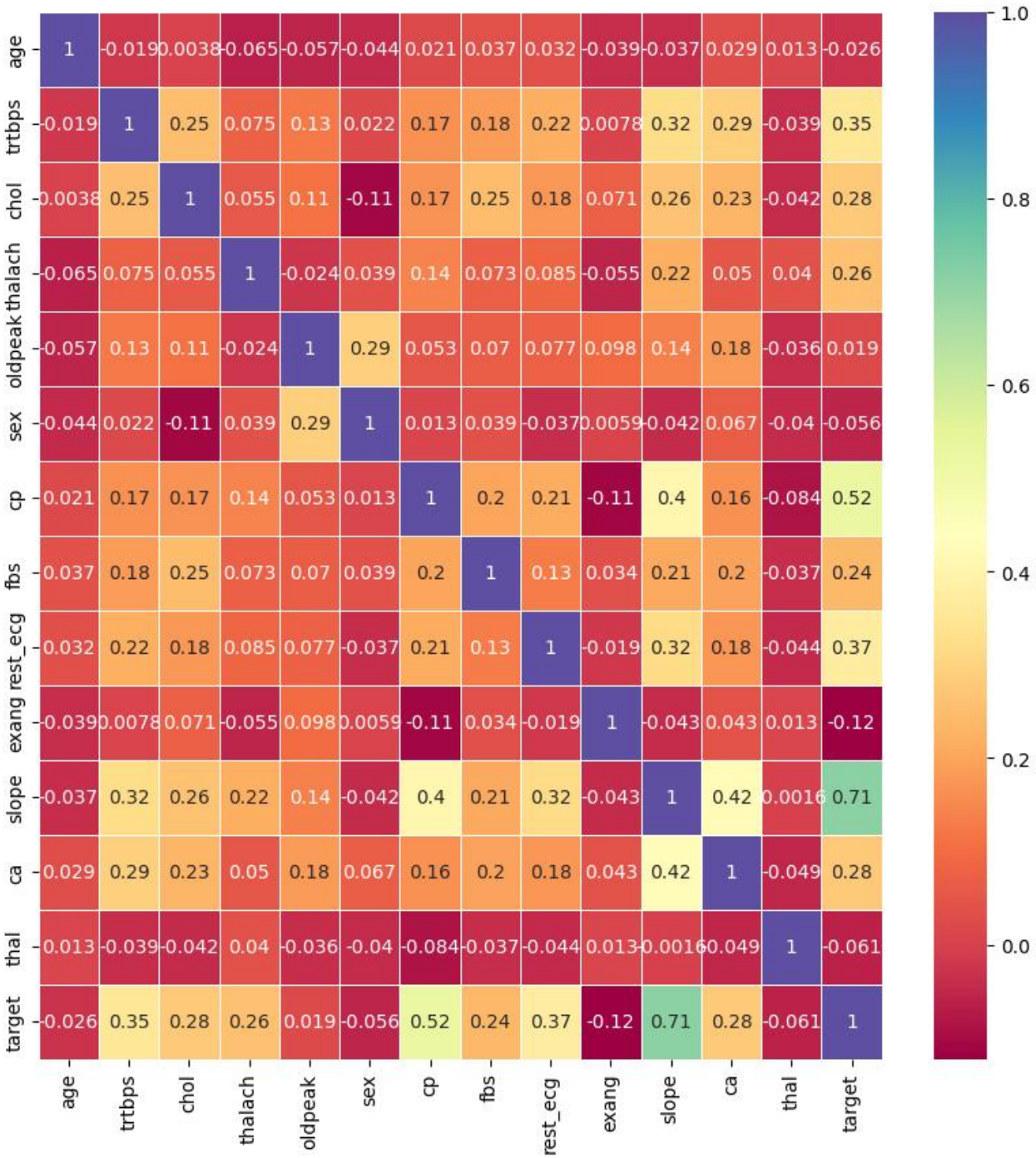
Categorical Variables (Analysis with Pie Chart)







- Checking relationship between the feature variables through heat map


age 1 -0.019 0.0038 -0.065 -0.057 -0.044 0.021 0.037 0.032 -0.039 -0.037 0.029 0.013 -0.026
trtbps -0.019 1 0.25 0.075 0.13 0.022 0.17 0.18 0.22 0.0078 0.32 0.29 -0.039 0.35
chol 0.0038 0.25 1 0.055 0.11 -0.11 0.17 0.25 0.18 0.071 0.26 0.23 -0.042 0.28
thalach -0.065 0.075 0.055 1 -0.024 0.039 0.14 0.073 0.085 -0.055 0.22 0.05 0.04 0.26
oldpeak -0.057 0.13 0.11 -0.024 1 0.29 0.053 0.07 0.077 0.098 0.14 0.18 -0.036 0.019
sex -0.044 0.022 -0.11 0.039 0.29 1 0.013 0.039 -0.037 0.0059 -0.042 0.067 -0.04 -0.056
cp 0.021 0.17 0.17 0.14 0.053 0.013 1 0.2 0.21 -0.11 0.4 0.16 -0.084 0.52
fbs 0.037 0.18 0.25 0.073 0.07 0.039 0.2 1 0.13 0.034 0.21 0.2 -0.037 0.24
rest_ecg 0.032 0.22 0.18 0.085 0.077 -0.037 0.21 0.13 1 -0.019 0.32 0.18 -0.044 0.37
exang -0.039 0.0078 0.071 -0.055 0.098 0.0059 -0.11 0.034 -0.019 1 -0.043 0.043 0.013 -0.12
slope -0.037 0.32 0.26 0.22 0.14 -0.042 0.4 0.21 0.32 -0.043 1 0.42 0.0016 0.71
ca 0.029 0.29 0.23 0.05 0.18 0.067 0.16 0.2 0.18 0.043 0.42 1 -0.049 0.28
thal 0.013 -0.039 -0.042 0.04 -0.036 -0.04 -0.084 -0.037 -0.044 0.013 -0.0016 0.049 1 -0.061
target -0.026 0.35 0.28 0.26 0.019 -0.056 0.52 0.24 0.37 -0.12 0.71 0.28 -0.061 1
age trtbps chol thalach oldpeak sex cp fbs rest_ecg exang slope ca thal target
1.0 0.8 0.6 0.4 0.2 0.0

7. MODELING and PREDICTION

Predictive modeling is a statistical approach that analyzes data patterns to determine future events or outcomes. It's an essential aspect of predictive analytics, a type of data analytics that involves machine learning and data mining approaches to predict activity, behavior, and trends using current and past data.

In this project we applied 3 machine learning models to analysis the data and predict the outcome.

Firstly we are splitting our dataset into train and test data.

```
[140] from sklearn.model_selection import train_test_split

[141] x=df_copy.drop(["target"],axis=1)
      y=df_copy[["target"]]

[142] x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.2,random_state=3)

[143] x_train.head()
```

	age	chol	thalach	trtbps_winsorize	oldpeak_winsorize_sqrt	sex_1	cp_1	cp_2	cp_3	exang_1	slope_1	slope_2	slope_3	ca_1	ca_2	ca_3	ca_4	thal
748	-1.307300	348	-0.830970	0.975044	-0.203632	1	0	0	0	0	1	0	0	0	0	0	0	
1096	-1.002154	248	-1.426656	-0.693032	0.900826	1	0	0	0	0	0	0	0	1	0	0	0	
1098	-0.635978	390	0.078235	-0.339198	1.108304	1	1	0	0	0	0	0	1	0	0	1	0	
836	1.133868	200	-0.486099	0.317923	1.108304	1	0	0	1	0	0	0	1	0	0	1	0	
550	0.096372	133	-1.677471	-1.400701	-1.088274	1	0	0	0	1	1	0	0	0	0	0	0	

```
[145] print(f"X_train:{x_train.shape[0]}")
      print(f"X_test:{x_test.shape[0]}")
      print(f"Y_train:{y_train.shape[0]}")
      print(f"Y_test:{y_test.shape[0]}")

      X_train:1041
      X_test:261
      Y_train:1041
      Y_test:261
```

Now applying the various Machine Learning models on our dataset:

1. Logistic Regression Algorithm

Logistic regression is a supervised machine learning algorithm that accomplishes binary classification tasks by predicting the probability of an outcome, event, or observation. The model delivers a binary or dichotomous outcome limited to two possible outcomes: yes/no, 0/1, or true/false.

```
[146] from sklearn.linear_model import LogisticRegression
      from sklearn.metrics import accuracy_score
```

```
[147] log_reg=LogisticRegression()
      log_reg
```

```
LogisticRegression()
```

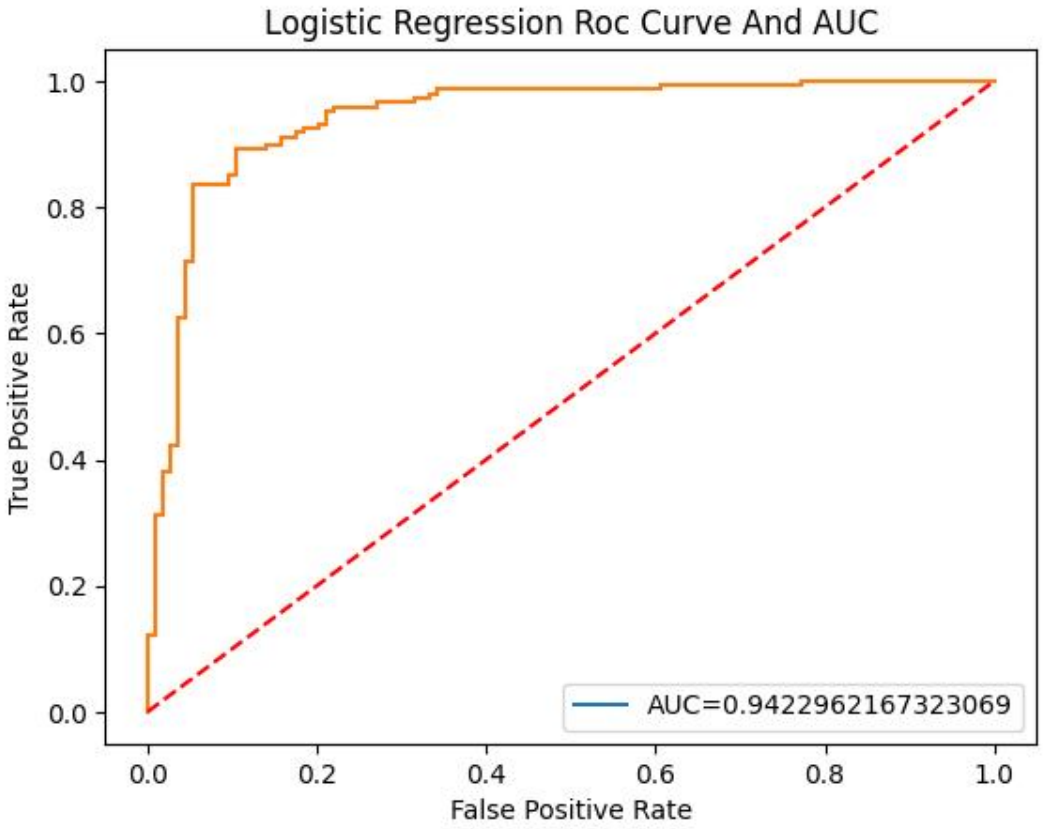
```
log_reg.fit(x_train,y_train)
```

```
LogisticRegression()
```

```
[149] y_pred=log_reg.predict(x_test)
```

```
[151] accuracy=accuracy_score(y_test,y_pred)
      print("Test Accuracy:{}".format(accuracy))
```

```
Test Accuracy:0.8735632183908046
```




```
[135] from sklearn.metrics import classification_report,confusion_matrix
```

```
[136] print(classification_report(y_test,y_pred_log))
```

```
                   precision    recall  f1-score   support

                0       0.88      0.82      0.85       114
                1       0.87      0.91      0.89       147

         accuracy                           0.87       261
        macro avg       0.87      0.87      0.87       261
     weighted avg       0.87      0.87      0.87       261
```

```
[137] print(confusion_matrix(y_test,y_pred_log))
```

```
[[ 94  20]
 [ 13 134]]
```

2. Decision Tree

A decision tree is a non-parametric supervised learning algorithm for classification and regression tasks. It has a hierarchical tree structure consisting of a root node, branches, internal nodes, and leaf nodes. Decision trees are used for classification and regression tasks, providing easy-to-understand models.

A decision tree is a hierarchical model used in decision support that depicts decisions and their potential outcomes, incorporating chance events, resource expenses, and utility. This algorithmic model utilizes conditional control statements and is non-parametric, supervised learning, useful for both classification and regression tasks. The tree structure is comprised of a root node, branches, internal nodes, and leaf nodes, forming a hierarchical, tree-like structure.

```
[219] from sklearn.tree import DecisionTreeClassifier
```

```
[224] dec_tree=DecisionTreeClassifier(random_state=3)
```

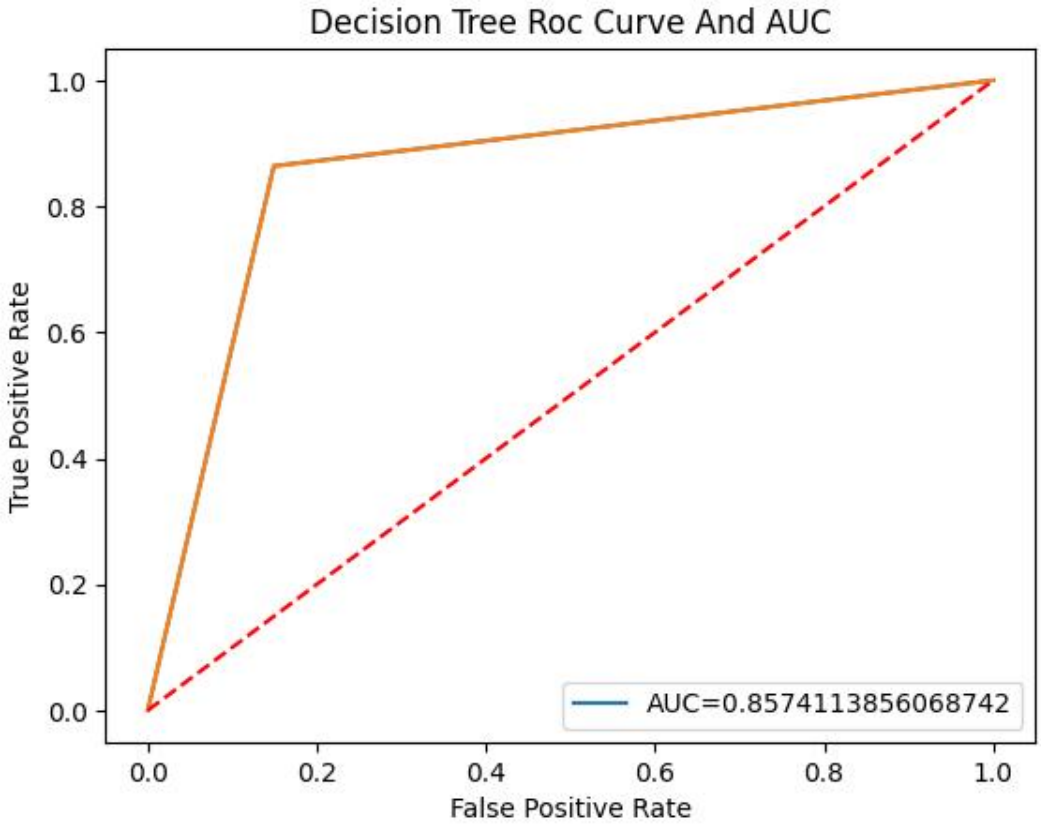
```
[225] dec_tree.fit(x_train,y_train)
```

```
DecisionTreeClassifier(random_state=3)
```

```
[226] y_pred=dec_tree.predict(x_test)
```

```
print("The test accuracy score of Decision Tree is:",accuracy_score(y_test,y_pred))
```

```
The test accuracy score of Decision Tree is: 0.8582375478927203
```




```
print(classification_report(y_test,y_pred_dec))
```

	precision	recall	f1-score	support
0	0.83	0.85	0.84	114
1	0.88	0.86	0.87	147
accuracy			0.86	261
macro avg	0.86	0.86	0.86	261
weighted avg	0.86	0.86	0.86	261

```
[161] print(confusion_matrix(y_test,y_pred_dec))
```

```
[[ 97  17]
 [ 20 127]]
```

3. Random Forest

A Random Forest is like a group decision-making team in machine learning. It combines the opinions of many “trees” (individual models) to make better predictions, creating a more robust and accurate overall model.

Random Forest Algorithm widespread popularity stems from its user-friendly nature and adaptability, enabling it to tackle both classification and regression problems effectively.

The algorithm’s strength lies in its ability to handle complex datasets and mitigate overfitting, making it a valuable tool for various predictive tasks in machine learning.

```
from sklearn.ensemble import RandomForestClassifier
```

```
[176] random_forest=RandomForestClassifier(random_state=5)
```

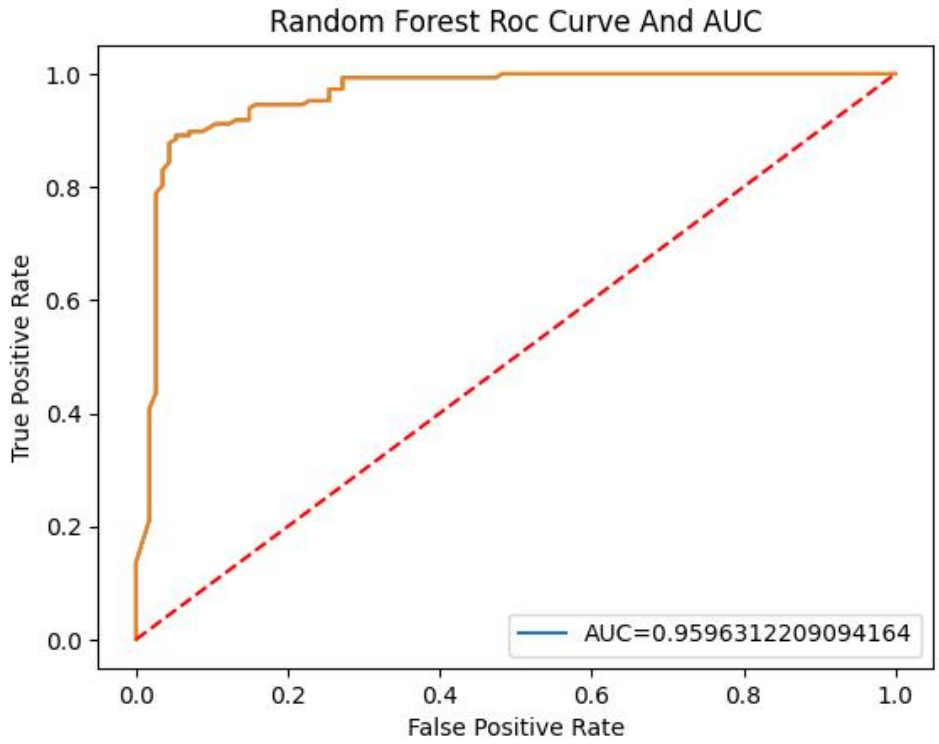
```
[177] random_forest.fit(x_train,y_train)
```

```
RandomForestClassifier(random_state=5)
```

```
[178] y_pred=random_forest.predict(x_test)
```

```
[179] print("The accuracy score of Random Forest is",accuracy_score(y_test,y_pred))
```

```
The accuracy score of Random Forest is 0.896551724137931
```




```
[169] print(classification_report(y_test,y_pred_random))
```

	precision	recall	f1-score	support
0	0.88	0.88	0.88	114
1	0.91	0.91	0.91	147
accuracy			0.90	261
macro avg	0.90	0.89	0.89	261
weighted avg	0.90	0.90	0.90	261

```
[170] print(confusion_matrix(y_test,y_pred_random))

[[100  14]
 [ 13 134]]
```

8. Conclusion

The activities we carried out within the scope of the project are as follows:

1. Within the scope of the project, we first made the data set ready for Exploratory Data Analysis(EDA)
2. We performed Exploratory Data Analysis(EDA).
3. We analyzed numerical and categorical variables within the scope of univariate analysis by using Distplot and Pie Chart graphics.
4. Within the scope of bivariate analysis, we analyzed the variables among each other using FacetGrid, Count Plot, Pair Plot, Swarm plot, Box plot, and Heatmap graphics.
5. We made the data set ready for the model. In this context, we struggled with missing and outlier values.
6. We used three different algorithms in the model phase.
7. We got 87% accuracy and 94% AUC with the Logistic Regression model.
8. We got 85% accuracy and 85% AUC with the Decision Tree Model.
9. And we got 89% accuracy and 95% AUC with the Random Forest Classifier Model.
10. When all these model outputs are evaluated, we prefer the model we created with the Random Forest Algorithm, which gives the best results.

9. REFERENCES

- "Cardiovascular Diseases (Cvds)". Who.Int, 2020,

  https://www.researchgate.net/publication/333888974_Effective_Heart_DiseasePrediction_Using_Hybrid_Machine_Learning_Technique

- "Logistic Regression". En.Wikipedia.Org, 2020,

  https://en.wikipedia.org/wiki/Logistic_regression

- "Understanding Random Forest". Medium, 2020,

  https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

- "Understanding Decision Tree". https://www.analyticsvidhya.com/blog/2021/08/decision-tree-algorithm/

- Source Of Datasets

  https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset