

HEART DISEASE PREDICTION

Submitted by

SARAVANAN S

(1P21CS023)

In partial fulfillment of the requirements for the award of the Degree of

MASTER OF SCIENCE IN COMPUTER SCIENCE

from Bharathiar University, Coimbatore.

Under the Internal Supervision of

Dr. Suganya S. M.C.A., Ph.D.

Associate Professor



SCHOOL OF COMPUTER STUDIES (PG)

RATHINAVEL SUBRAMANIAM COLLEGE OF ARTS AND SCIENCE (AUTONOMOUS)

Sulur, Coimbatore – 641 402.

April 2023.

**RATHNAVEL SUBRAMANIAM COLLEGE OF ARTS AND SCIENCE
(AUTONOMOUS)**

Sulur, Coimbatore – 641 402.

School of Computer Studies (PG)



Register Number: 1P21CS023

Certified bonafide original project work done by SARAVANAN S

Internal Supervisor

HoD

Submitted for the Project Evaluation and Viva voce held on_____

Internal Examiner

External Examiner

CERTIFICATE

CERTIFICATE

This is to certify that the dissertation entitled **HEART DISEASE PREDCTION** submitted to the School of Computer Studies, Rathnavel Subramaniam College of Arts and Science in partial fulfillment of the requirements for the award of the Degree of Master of Science in Computer Science is a record of original project work done by **SARAVANAN S** during the period January-2023-April-2023 of his study in the School of Computer Studies, under my internal supervision and the dissertation has not formed the basis for the award of any Degree/Diploma/Associateship/Fellowship or other similar title to any candidate of any University.

Internal Supervisor

[Dr.S.Suganya]

DECLARATION

DECLARATION

I, **SARAVANAN S** , hereby declare that the project entitled **HEART DISEASE PREDICTION**, submitted to the School of Computer Studies (PG), Rathnavel Subramaniam College of Arts and Science, in partial fulfillment of the requirements for the award of the Degree of Master of Science in Computer Science is a record of original project work done by me during the period Jan 2023 to April 2023 under the internal supervision of **Dr. Suganya S. M.C.A., Ph.D., Associate Professor, Rathnavel Subramaniam College Of Arts and Science (Autonomous)** From Bharathiar University, Coimbatore.

Signature of the Candidate

ACKNOWLEDGEMENT

ACKNOWLEDGEMENT

I express my sincere thanks to our Managing Trustee **Dr. K. Senthil Ganesh MBA (USA), MS (UK), Ph.D.**, for providing us with adequate faculty and laboratory resources for completing my project successfully.

I take this as a fine opportunity to express my sincere thanks to **Dr. T. Sivakumar M.Sc., M. Phil., Ph.D., Principal**, Rathnavel Subramaniam College of Arts and Science (Autonomous) for giving me the opportunity to undertake this project.

I express my sincere thanks to **Dr. P. Navaneetham M.Sc., M.Phil., Ph.D., Director (Administration), School of Computer Studies** for the help and advice throughout the project.

I express my sincere thanks to **Dr. S. Yamini M.Sc., (CC), M. Phil., Ph.D., Director (Academic), School of Computer Studies** for her valuable guidance and prompt correspondence throughout the curriculum to complete the project.

I express my sincere thanks to **Dr. D. Maheswari, M.Sc. (CS), M.Phil., Ph.D., Head and Research Coordinator, School of Computer Studies (PG)** for her support and advice throughout the project.

I express my gratitude to **Dr. Suganya S. M.C.A., Ph.D., Associate Professor, School of Computer Studies (PG)** for her valuable guidance, support, encouragement, and motivation rendered by her throughout this project.

Finally, I express my sincere thanks to all other staff members and my dear friends, dear and near for helping me to complete this project.

SARAVANAN S

ABSTRACT

ABSTRACT

Heart-related diseases or cardiovascular diseases (CVDs) are the main reason for a huge number of death in the world over the last few decades and has emerged as the most life-threatening disease, not only in India but in the whole world. So, there is a need for a reliable, accurate, and feasible system to diagnose such diseases in time for proper treatment. Machine Learning algorithms and techniques have been applied to various medical datasets to automate the analysis of large and complex data. Many researchers, in recent times, have been using several machine learning techniques to help the health care industry and the professionals in the diagnosis of heart-related diseases. Heart is the next major organ comparing to the brain which has more priority in the Human body. It pumps the blood and supplies it to all organs of the whole body. Prediction of occurrences of heart diseases in the medical field is significant work. Data analytics is useful for prediction from more information and it helps the medical center to predict various diseases. A huge amount of patient-related data is maintained on monthly basis. The stored data can be useful for the source of predicting the occurrence of future diseases. Some of the data mining and machine learning techniques are used to predict heart diseases, such as Random Forest . Prediction and diagnosing of heart disease become a challenging factor faced by doctors and hospitals both in India and abroad. To reduce the large scale of deaths from heart diseases, a quick and efficient detection technique is to be discovered. Data mining techniques and machine learning algorithms play a very important role in this area. The researchers accelerating their research works to develop software with the help of machine learning algorithms which can help doctors to decide both prediction and diagnosing of heart disease. The main objective of this research project is to predict the heart disease of a patient using machine learning algorithms.

TABLE OF CONTENTS

TABLE OF CONTENTS

Certificates	iv
Declaration	vi
Acknowledgement	viii
Abstract	x

CHAPTER - I

1. Introduction	1
1.1 Over View of Project	1
1.2 Scope of Analysis	1
1.3 Approach of Data Analysis	2

CHAPTER – II

2. Data Understanding	4
2.1 Data Gathering	4
2.1.1 Context	4
2.1.2 Source	4
2.1.3 Citation	5
2.1.4 Acknowledgements	5
2.2 About Dataset	6
2.3 Structure of The Dataset	7
2.4 Data Description	8

CHAPTER - III

3. Data Preparation	13
3.1 Data Cleaning	13
3.2 Data Cleaning Process In Heart Disease Dataset	14
3.3 Outliers	16

CHAPTER – IV

4. Exploratory Data Analysis	20
4.1 Data Exploration	20

CHAPTER - V	
5. Model Building	38
5.1 Algorithm Description	38
5.2 Data Splitting	39
5.3 Training and Test Data	39
CHAPTER - VI	
6. Evaluation of Model	41
6.1 Performance of Metrics	41
6.2 Performance of Model	42
CHAPTER - VII	
7. Prediction and Inference	44
7.1 Prediction	44
7.2 Inference	44
CHAPTER - VIII	
8. Model Deployment	45
8.1 Importing Relevant Package For Model Deployment	45
8.2 Webpage On Heart Prediction	45
CHAPTER – IX	
Conclusion	47
References	48

Chapter 1: INTRODUCTION

1.1 Overview of project

The objective of heart disease prediction using the machine learning algorithm. This model predicts the heart disease have or not of the people based on their some features like Age, Sex, ChestPainType, RestingBP, Cholesterol, FastingBS, RestingECG, MaxHR, ExerciseAngina, Oldpeak., ST_Slope, HeartDisease. This is a categorical problem which is solved using a Logistic Regression. This model implemented by using the following steps such as :-

1. Import all required libraries
2. load data
3. perform EDA
4. visualize data
5. prepare data
6. split data into training and testing
7. define model
8. test model
9. check accuracy
10. save the model

1.2 Scope of analysis

The scope of heart disease prediction involves identifying individuals who may be at increased risk for developing heart disease in the future. This can include individuals who have not yet been diagnosed with heart disease but may be at risk due to their age, family history, lifestyle factors, or other medical conditions.

Some of the key areas within the scope of heart disease prediction include:

Risk assessment: Conducting a thorough assessment of an individual's risk factors for heart disease, including Age, Sex, ChestPainType, RestingBP, Cholesterol, FastingBS, RestingECG, MaxHR, ExerciseAngina, Oldpeak., ST_Slope, HeartDisease.

Screening tests: Using various screening tests such as blood tests, imaging studies, and electrocardiograms to identify early signs of heart disease.

Predictive modelling: Developing models that can help predict an individual's risk of developing heart disease based on their risk factors and other demographic and clinical variables.

Personalized prevention strategies: Developing personalized prevention strategies based on an individual's risk profile, which may include lifestyle modifications, medication management, and other interventions to reduce their risk of developing heart disease.

Population health management: Developing strategies to manage the health of populations at risk for heart disease, such as developing public health campaigns, promoting healthy lifestyles, and implementing policies to reduce risk factors.

Overall, the scope of heart disease prediction involves early detection and prevention of heart disease through the identification of high-risk individuals and the development of personalized prevention strategies. This can help reduce the burden of heart disease and improve outcomes for those affected by these conditions

1.3 Approach of data analysis

There are several approaches to data analysis in heart disease analysis, depending on the type and nature of the data being analysed. Here are some common approaches:

Descriptive statistics: This involves summarizing and describing the data using statistical measures such as mean, median, mode, variance, standard deviation, and percentiles. This approach can help identify trends and patterns in the data.

Exploratory Data Analysis (EDA): EDA involves analysing and visualizing the data to identify patterns, trends, and relationships between variables. EDA can help in identifying important features or variables that are predictive of heart failure.

Machine learning: This involves using algorithms to automatically learn patterns in the data and make predictions or classifications. Machine learning can be used to develop predictive models for heart disease risk, diagnosis, and treatment.

Model Selection: Model selection involves choosing the appropriate machine learning algorithm for heart failure prediction based on the characteristics of the data and the problem at hand. Common machine learning algorithms used in heart failure prediction include logistic regression, decision trees, random forests, and neural networks.

Interpretability: Interpretability involves understanding how the heart failure prediction models make their predictions. Interpretability techniques such as feature importance ranking, partial dependence plots.

Chapter 2: DATA UNDERSTANDING

2.1 Data Gathering

2.1.1 Context

Cardiovascular diseases (CVDs) are the number 1 cause of death globally, taking an estimated 17.9 million lives each year, which accounts for 31% of all deaths worldwide. Four out of 5 CVD deaths are due to heart attacks and strokes, and one-third of these deaths occur prematurely in people under 70 years of age. Heart failure is a common event caused by CVDs and this dataset contains 11 features that can be used to predict a possible heart disease.

People with cardiovascular disease or who are at high cardiovascular risk (due to the presence of one or more risk factors such as hypertension, diabetes, hyperlipidaemia or already established disease) need early detection and management wherein a machine learning model can be of great help prediction.

2.1.2 Source

This dataset was created by combining different datasets already available independently but not combined before. In this dataset, 5 heart datasets are combined over 11 common features which makes it the largest heart disease dataset available so far for research purposes. The five datasets used for its curation are:

Cleveland: 303 observations

Hungarian: 294 observations

Switzerland: 123 observations

Long Beach VA: 200 observations

Stalog (Heart) Data Set: 270 observations

Total: 1190 observations

Duplicated: 272 observations

Final: 918 Observation

Every dataset used can be found under the Index of heart disease datasets from UCI Machine Learning Repository on the following link: <https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/>.

2.1.3 Citation

[fedesoriano. \(September 2021\). Heart Failure Prediction Dataset. Retrieved \[Date Retrieved\] from https://www.kaggle.com/fedesoriano/heart-failure-prediction.](https://www.kaggle.com/fedesoriano/heart-failure-prediction)

2.1.4 Acknowledgements

Creators:

Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.

University Hospital, Zurich, Switzerland: William Steinborn, M.D.

University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.

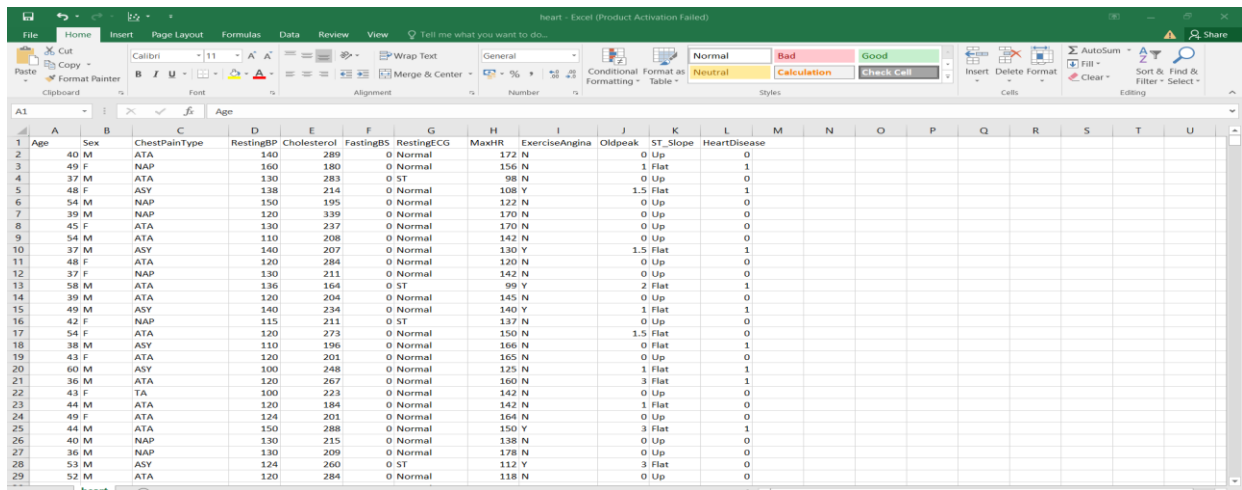
V.A. Medical Centre, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.

Donor:

David W. Aha (aha '@' ics.uci.edu) (714) 856-8779

2.2 About dataset

Data Set



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease									
2	40	M	ATA	140	289	0	Normal	172	N	0	Up	0									
3	49	F	NAP	160	180	0	Normal	156	N	1	Flat	1									
4	37	M	ATA	130	283	0	ST	98	N	0	Up	0									
5	48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1									
6	54	M	NAP	150	195	0	Normal	122	N	0	Up	0									
7	39	M	NAP	120	339	0	Normal	170	N	0	Up	0									
8	45	F	ATA	130	237	0	Normal	170	N	0	Up	0									
9	54	M	ATA	110	208	0	Normal	142	N	0	Up	0									
10	37	M	ASY	140	207	0	Normal	130	Y	1.5	Flat	1									
11	48	F	ATA	120	284	0	Normal	120	N	0	Up	0									
12	37	F	NAP	130	211	0	Normal	142	N	0	Up	0									
13	58	M	ATA	136	164	0	ST	99	Y	2	Flat	1									
14	39	M	ATA	120	204	0	Normal	145	N	0	Up	0									
15	49	M	ASY	140	234	0	Normal	140	Y	1	Flat	1									
16	42	F	NAP	115	211	0	ST	137	N	0	Up	0									
17	54	F	ATA	120	273	0	Normal	150	N	1.5	Flat	0									
18	38	M	ASY	110	196	0	Normal	166	N	0	Flat	1									
19	43	F	ATA	120	201	0	Normal	165	N	0	Up	0									
20	60	M	ASY	100	248	0	Normal	125	N	1	Flat	1									
21	36	M	ATA	120	267	0	Normal	160	N	3	Flat	1									
22	43	F	TA	100	223	0	Normal	142	N	0	Up	0									
23	44	M	ATA	120	184	0	Normal	142	N	1	Flat	0									
24	49	F	ATA	124	201	0	Normal	164	N	0	Up	0									
25	44	M	ATA	150	288	0	Normal	150	Y	3	Flat	1									
26	40	M	NAP	130	215	0	Normal	138	N	0	Up	0									
27	36	M	NAP	130	209	0	Normal	178	N	0	Up	0									
28	53	M	ASY	124	260	0	ST	112	Y	3	Flat	0									
29	52	M	ATA	120	284	0	Normal	118	N	0	Up	0									

Importing the relevant package

```
final.ipynb > # Frequency Distribution of Numerical Columns
Code + Markdown | Run All | Clear All Outputs | Restart | Variables | Outline ... Python 3.10.4

1 import numpy as np
2 import pandas as pd
3 import plotly.express as px
4 import matplotlib.pyplot as plt
5 from plotly.offline import iplot
6
7
8 %matplotlib inline
9 import seaborn as sns
10 from sklearn.naive_bayes import GaussianNB
11 from sklearn import svm
12 from sklearn.ensemble import RandomForestClassifier
13 from sklearn.metrics import accuracy_score
14 from sklearn.model_selection import train_test_split
15 from sklearn.linear_model import LogisticRegression
16 from sklearn.metrics import accuracy_score, confusion_matrix, classification_report

[28] ✓ 0.7s Python

...

1 import pickle
2 import streamlit as st
3 from streamlit_option_menu import option_menu

[4] ✓ 3.0s Python
```

Loading the Data Set

```
1 df=pd.read_csv("heart.csv")
✓ 0.1s Python

1 df
✓ 0.1s Python

Age Sex ChestPainType RestingBP Cholesterol FastingBS RestingECG MaxHR ExerciseAngina Oldpeak ST_Slope HeartDisease
0 40 M ATA 140 289 0 Normal 172 N 0.0 Up 0
1 49 F NAP 160 180 0 Normal 156 N 1.0 Flat 1
2 37 M ATA 130 283 0 ST 98 N 0.0 Up 0
3 48 F ASY 138 214 0 Normal 108 Y 1.5 Flat 1
4 54 M NAP 150 195 0 Normal 122 N 0.0 Up 0
... ..
913 45 M TA 110 264 0 Normal 132 N 1.2 Flat 1
914 68 M ASY 144 193 1 Normal 141 N 3.4 Flat 1
915 57 M ASY 130 131 0 Normal 115 Y 1.2 Flat 1
916 57 F ATA 130 236 0 LVH 174 N 0.0 Flat 1
917 38 M NAP 138 175 0 Normal 173 N 0.0 Up 0

918 rows x 12 columns
```

Data Set Observation Count

```
1 df.shape
✓ 0.1s
(918, 12)
```

The shape of the dataset is (918,12). There are 918 columns and 12 rows

Displaying the first five rows and columns data from the data set

```
1 df.head(5)
✓ 0.1s
```

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
0	40	M	ATA	140	289	0	Normal	172	N	0.0	Up	0
1	49	F	NAP	160	180	0	Normal	156	N	1.0	Flat	1
2	37	M	ATA	130	283	0	ST	98	N	0.0	Up	0
3	48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
4	54	M	NAP	150	195	0	Normal	122	N	0.0	Up	0

This table show first five observations of rows and column in the dataset.

Displaying the last five rows and columns data from the data set

```
1 df.tail(5)
✓ 0.1s
```

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
913	45	M	TA	110	264	0	Normal	132	N	1.2	Flat	1
914	68	M	ASY	144	193	1	Normal	141	N	3.4	Flat	1
915	57	M	ASY	130	131	0	Normal	115	Y	1.2	Flat	1
916	57	F	ATA	130	236	0	LVH	174	N	0.0	Flat	1
917	38	M	NAP	138	175	0	Normal	173	N	0.0	Up	0

This table show last five observations of rows and column in the dataset.

2.3 Structure of the data set

```
1 df.info()
✓ 0.1s

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 918 entries, 0 to 917
Data columns (total 12 columns):
#   Column                Non-Null Count  Dtype
---  ---                -
0   Age                   918 non-null   int64
1   Sex                   918 non-null   object
2   ChestPainType         918 non-null   object
3   RestingBP             918 non-null   int64
4   Cholesterol            918 non-null   int64
5   FastingBS             918 non-null   int64
6   RestingECG            918 non-null   object
7   MaxHR                 918 non-null   int64
8   ExerciseAngina        918 non-null   object
9   Oldpeak               918 non-null   float64
10  ST_Slope              918 non-null   object
11  HeartDisease          918 non-null   int64
dtypes: float64(1), int64(6), object(5)
memory usage: 86.2+ KB
```

This is the structure of the dataset. It shows the what are the datatypes are occurred in dataset

2.4 Data Description

```
1 df.describe()
```

	Age	RestingBP	Cholesterol	FastingBS	MaxHR	Oldpeak	HeartDisease
count	918.000000	918.000000	918.000000	918.000000	918.000000	918.000000	918.000000
mean	53.510893	132.396514	198.799564	0.233115	136.809368	0.887364	0.553377
std	9.432617	18.514154	109.384145	0.423046	25.460334	1.066570	0.497414
min	28.000000	0.000000	0.000000	0.000000	60.000000	-2.600000	0.000000
25%	47.000000	120.000000	173.250000	0.000000	120.000000	0.000000	0.000000
50%	54.000000	130.000000	223.000000	0.000000	138.000000	0.600000	1.000000
75%	60.000000	140.000000	267.000000	0.000000	156.000000	1.500000	1.000000
max	77.000000	200.000000	603.000000	1.000000	202.000000	6.200000	1.000000

Age: Maximum age is 77 and the minimum age is 28, the mean is 53.51 and 75 percent of the age is below 60. RestingBP: Maximum value of RestingBP is 200 and the minimum is 0 and 75 percent of the BP is below 140. Cholesterol: Maximum value of Cholesterol is 603 and the minimum is 0 and 75 percent of the Cholesterol is below 267. FastingBS: Maximum value of FastingBS is 1 and the minimum is 0 and 75 percent of the FastingBS is below 0. MaxHR: Maximum value of MaxHR is 202 and the minimum is 60 and 75 percent of the MaxHR is below 156. Oldpeak: Maximum value of Oldpeak is 6.2 and the minimum is -2.6 and 75 percent of the Oldpeak is below 1.5.

The heart failure prediction dataset contains 918 observations with 12 attributes

This data frame contains the following columns:

Age:

Age of the patient [years]

Sex:

Sex of the patient [M: Male, F: Female]

Chest Pain Type:

Chest pain type [TA: Typical Angina,

ATA: Atypical Angina,

NAP: Non-Anginal Pain,

ASY: Asymptomatic]

Resting BP:

Resting blood pressure [mm Hg]

Cholesterol:

serum cholesterol [mm/dl]

Fasting BS:

Fasting blood sugar [1: if Fasting BS > 120 mg/dl, 0: otherwise]

Resting ECG:

Resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]

Max HR:

maximum heart rate achieved

[Numeric value between 60 and 202]

Exercise Angina:

Exercise-induced angina [Y: Yes, N: No]

Old peak:

ST [Numeric value measured in depression]

ST_Slope:

the slope of the peak exercise ST segment [Up: up sloping, Flat: flat, Down: down sloping]

Heart Disease:

[1: heart disease, 0: Normal]

Detail Review in Dataset**Age:**

Age of the patient [years]

Range [28-77]

Sex: Sex of the patient [M: Male, F: Female]

Range: [M=79%, F=21%]

Chest Pain Type:

Range [ASY=54%, NAP=22%, OTHER=24%]

Angina= Angina is **chest pain caused by reduced blood flow to the heart muscles**. It's not usually life threatening, but it's a warning sign that you could be at risk of a heart attack or stroke. Angina is a symptom of coronary artery disease. Angina is also called angina pectoris.

Chest pain type [TA: Typical Angina This is the one type of chest pain type and this is the fundable one,

ATA: Atypical Angina This is the one type of chest pain type but is not follow criteria of chest pain,

NAP: Non-Anginal Pain, Non cardiac chest pain is defined as recurring pain in your chest typically, behind your breast bone and near your heart that is not related to your heart.,

ASY: Asymptomatic This is the one type of chest pain type. A person with CAD may be asymptomatic (i.e., not have any symptoms)]

Resting BP:

Resting blood pressure [mm Hg]

Range: [0 – 200]

Normal :80/120 mm/hg

Cholesterol:

Serum: A person's serum cholesterol level represents the amount of total cholesterol in their blood.

serum cholesterol [mm/dl]

Range = 0 – 603

Normal: Less than 200mg/l

Fasting BS:

Fasting: Fasting means without food

blood sugar [1: if Fasting BS > 120 mg/dl, 0: otherwise]

Resting ECG:

Resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]

Range: [Normal = 60% LVH = 20% Other = 19%]

Normal: If the test is normal, it should show that your heart is beating at an even rate of 60 to 100 beats per minute

LVH: Irregular blood circulation (impact of depression)

Exercise Angina:

Exercise Angina: It usually happens during activity (exertion) and goes away with rest or angina medication. For example, pain that comes on when you're walking uphill or in the cold weather may be angina

Exercise include angina [Y=Yes, N=No]

Range: [Y:40% , N:60%]

Old Peak:

ST [Numeric value measured in depression]

Old peak refers to an exercise stress test parameter that measures the maximum ST segment depression during or after exercise. This value is expressed in millimetres and is a measure of the severity of ischemia or heart disease.

Range: [-2.6, 6.2]

ST_ Slope:

The slope of the peak exercise ST segment [Up: up sloping, Flat: flat, Down: down sloping]

Range: [Flat:43%, Up:43%, Other:7%]

slope of the peak exercise ST segment can provide important information about the health of the heart and its blood supply.

An upward-sloping (or up sloping) ST segment during exercise can indicate a healthy blood flow to the heart and is considered a normal finding.

A flat ST segment during exercise may suggest that there is an insufficient blood supply to the heart, which could be due to a blockage in the coronary arteries. This finding is considered abnormal and may require further evaluation.

A downward-sloping (or down sloping) ST segment during exercise is also considered abnormal and may indicate a significant blockage in the coronary arteries, which could increase the risk of a heart attack

Heart Disease:

output class [1: heart disease, 0: Normal]

CHAPTER III: DATA PREPARATION

3.1 Data Cleaning

Data cleaning or data preparation is an essential part of statistical analysis. In this part mostly we are checking the null values, inconsistency of data, and outliers. One of the biggest challenges when it comes to utilizing Machine Learning data is Data Cleaning. Although data cleaning may not be mentioned too often, it is very critical for the success of Machine Learning applications. This will help you to yield better results from your machine learning functions. Also, remove duplicate categorization from your data list and streamline your data. It is good to reduce the data you are handling. A downsized dataset can help you generate more accurate results. The main aim of Data Cleaning is to identify and remove errors & duplicate data, to create a reliable dataset. This improves the quality of the training data for analytics and enables accurate decision-making. Data cleansing is a time-consuming process and most data scientists spend an enormous amount of time enhancing the quality of the data. However, there are various methods to identify and classify data for data cleansing.

The data set is an important asset in any data analysis and model-building process. Generally, 80% of the time s scientists are utilized in data cleaning and manipulation, whereas 20% of the time are utilized in analysis and modelling. **‘Data cleaning’** is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate, or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data. To perform the data analytics properly we need a variety of data cleaning methods. Data cleaning depends on the type of data set. Data cleaning will help us to increase the accuracy of the model. The data cleaning and validation steps undertaken for any data science project are implemented using a data pipeline. Each stage in a data pipeline consumes input and produces output. The main advantage of the data pipeline is that each step is small, self-contained, and easier to check.

3.2 Data cleaning process in Heart Disease dataset

Finding the Null Values in Whole Date Set

```
1 df.isna().sum()
[1]
· Age      0
  Sex      0
  ChestPainType  0
  RestingBP  0
  Cholesterol  0
  FastingBS  0
  RestingECG  0
  MaxHR     0
  ExerciseAngina  0
  Oldpeak   0
  ST_Slope  0
  HeartDisease  0
  dtype: int64
```

This heart disease dataset not have a NA values.

Viewing the column name

```
1 df.columns
[12] ✓ 0.0s
... Index(['Age', 'Sex', 'ChestPainType', 'RestingBP', 'Cholesterol', 'FastingBS',
         'RestingECG', 'MaxHR', 'ExerciseAngina', 'Oldpeak', 'ST_Slope',
         'HeartDisease'],
         dtype='object')
```

Above this table shows what are the columns are occurred in the dataset. There are 12 columns are occurred

Finding the Duplicate Values in Whole Date Set

```
1 df.duplicated()
0      False
1      False
2      False
3      False
4      False
...
913     False
914     False
915     False
916     False
917     False
Length: 918, dtype: bool

checking the duplicate

1 df.duplicated().sum()
0
```

This dataset there is no duplicate values

Finding the Categorical Values in Whole Date Set

```
1 discrete_cols = df.select_dtypes(include=['object']).columns.tolist()
✓ 0.0s

1 discrete_cols
✓ 0.0s

['Sex', 'ChestPainType', 'RestingECG', 'ExerciseAngina', 'ST_Slope']
```

Above this table shows how many categorical columns are presented in dataset.

Finding the Numerical Values in Whole Date Set

```
1 numeric_cols = df.select_dtypes(include=[np.number]).columns.tolist()[:-1]
✓ 0.0s

1 numeric_cols
✓ 0.0s

['Age', 'RestingBP', 'Cholesterol', 'FastingBS', 'MaxHR', 'Oldpeak']
```

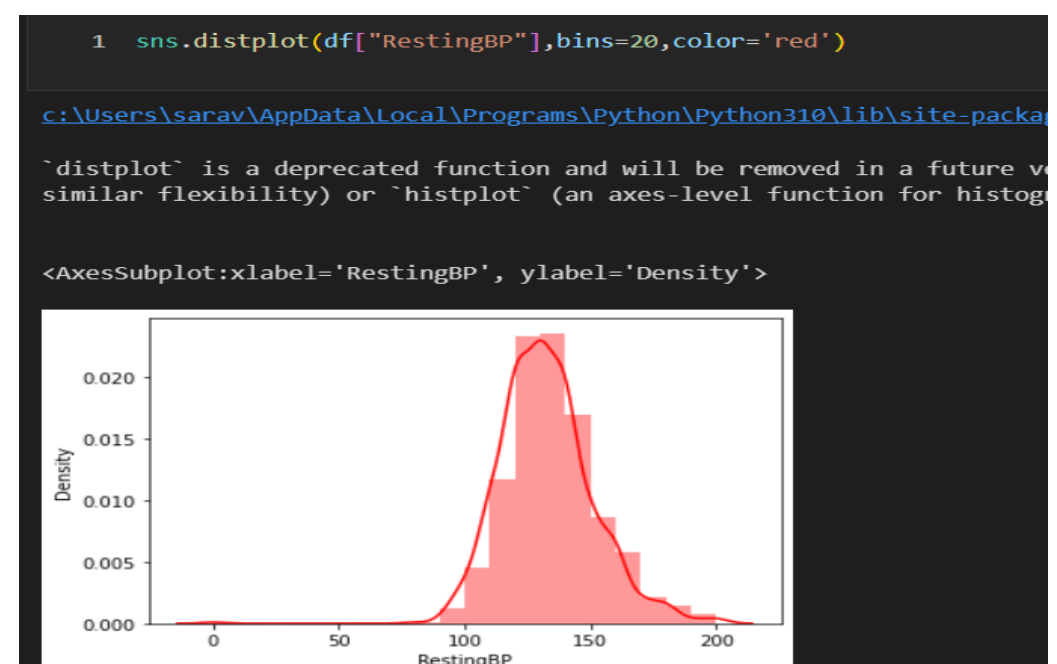
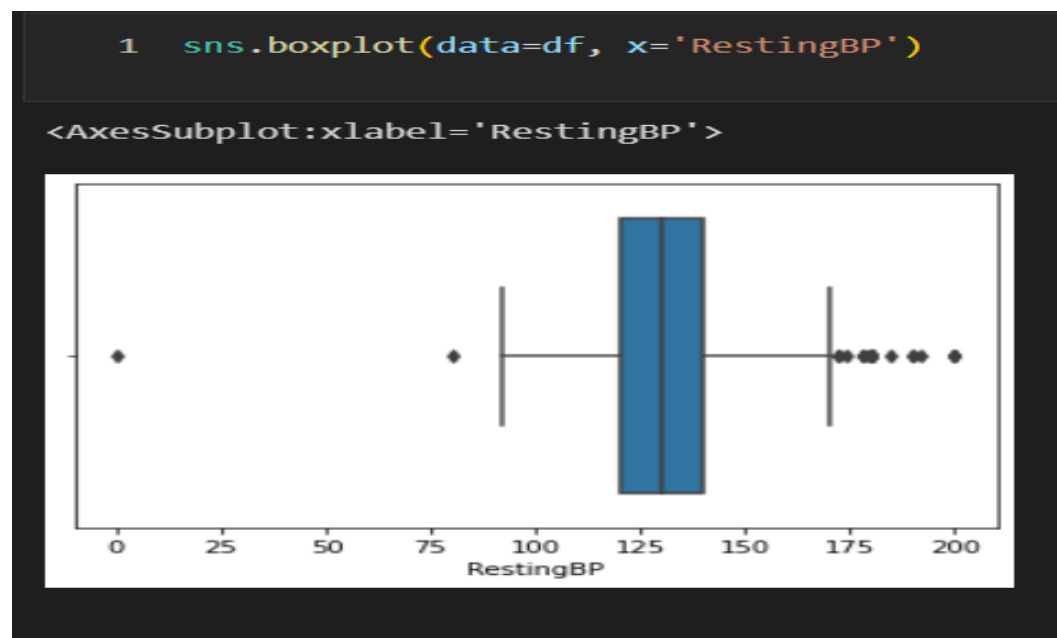
Above this table shows how many numerical columns are presented in dataset.

3.3 Outliers

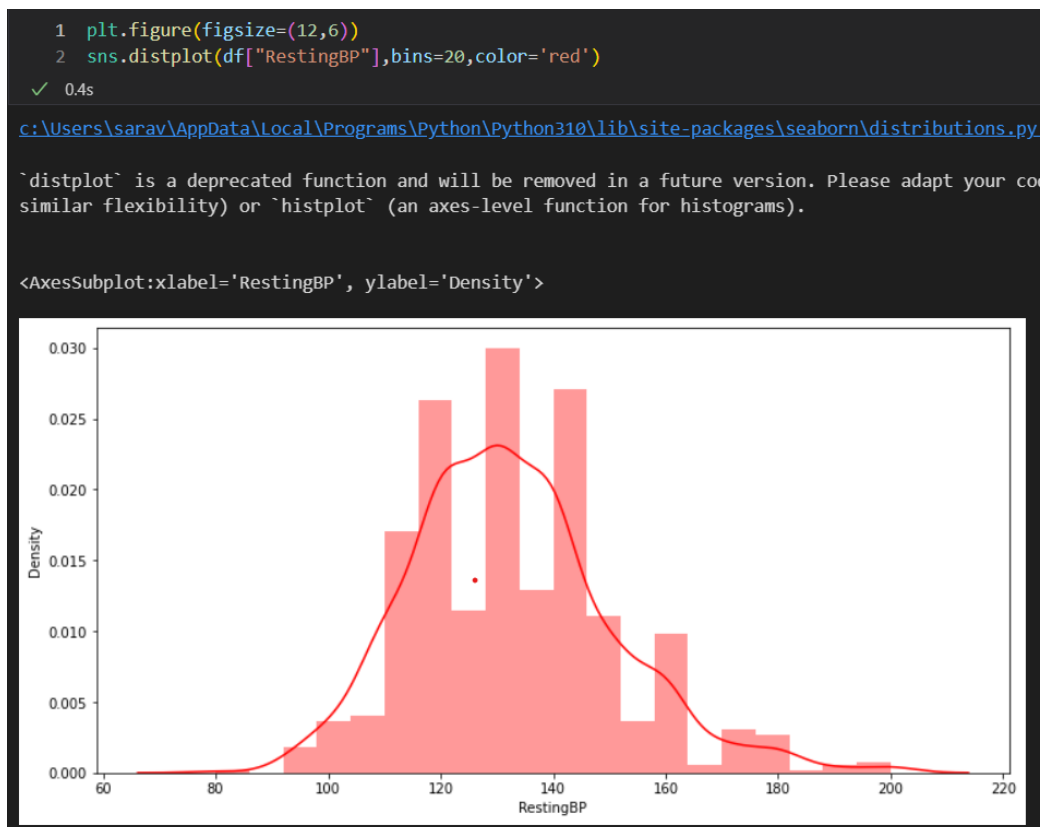
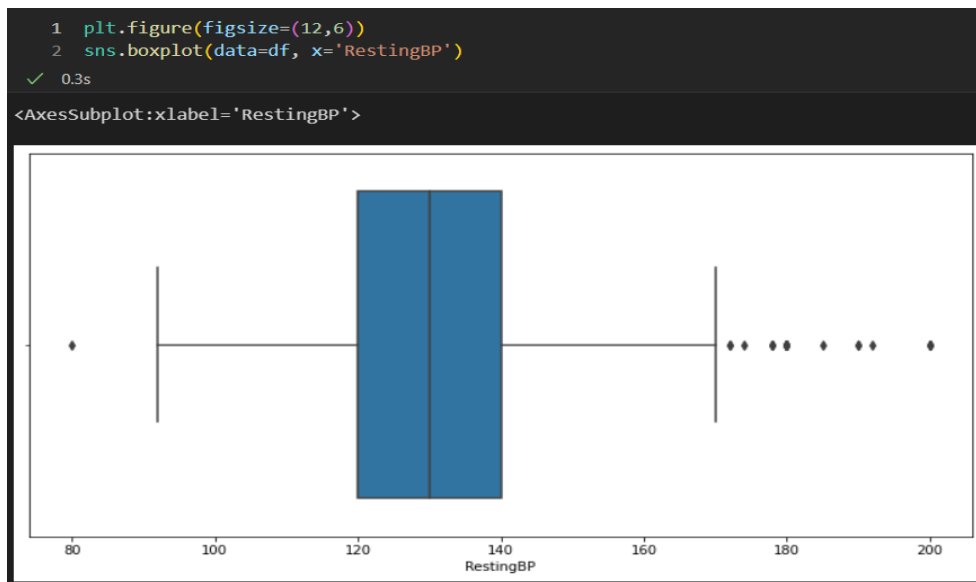
In this dataset the variables like RestingBP, Cholesterol have values 0 which is not Biologically possible so it is treated as an outliers and has been changed to mean value of the variable because after checking the skewness it has left, Right skewness

RestingBP

Before changing to median



After changing to median



The RestingBP have the zero value in the dataset. In medically this is not a correct one so I am changed to the median value in the respective column.

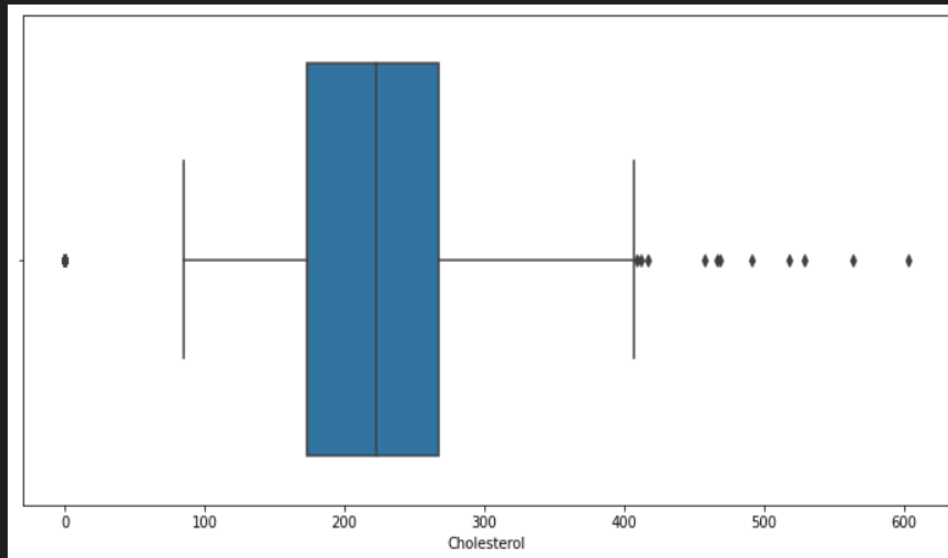
Cholesterol

Before changing to mean

```
1 plt.figure(figsize=(12,6))
2 sns.boxplot(data=df, x='Cholesterol')
3
```

✓ 0.4s

<AxesSubplot:xlabel='Cholesterol'>



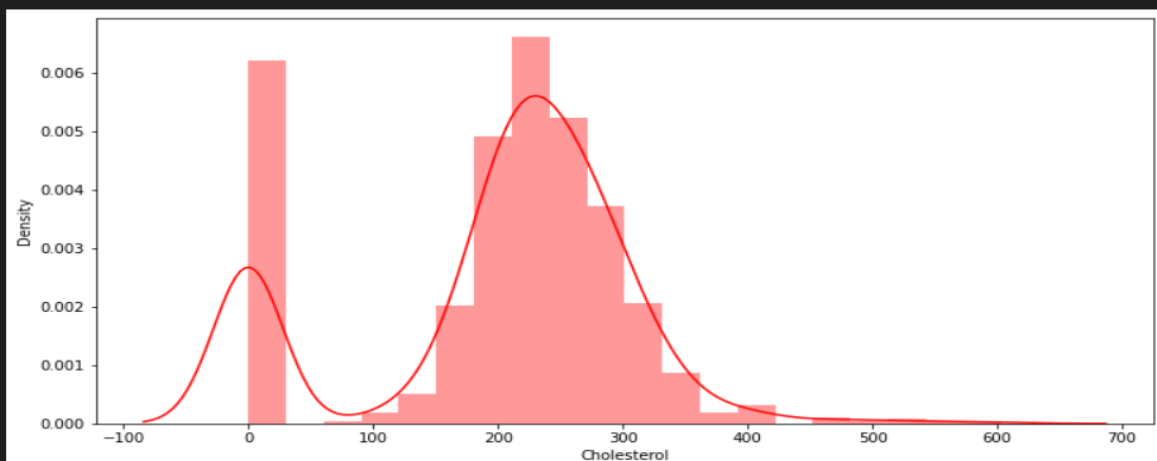
```
1 plt.figure(figsize=(12,6))
2 sns.distplot(df["Cholesterol"],bins=20,color='red')
```

✓ 4.2s

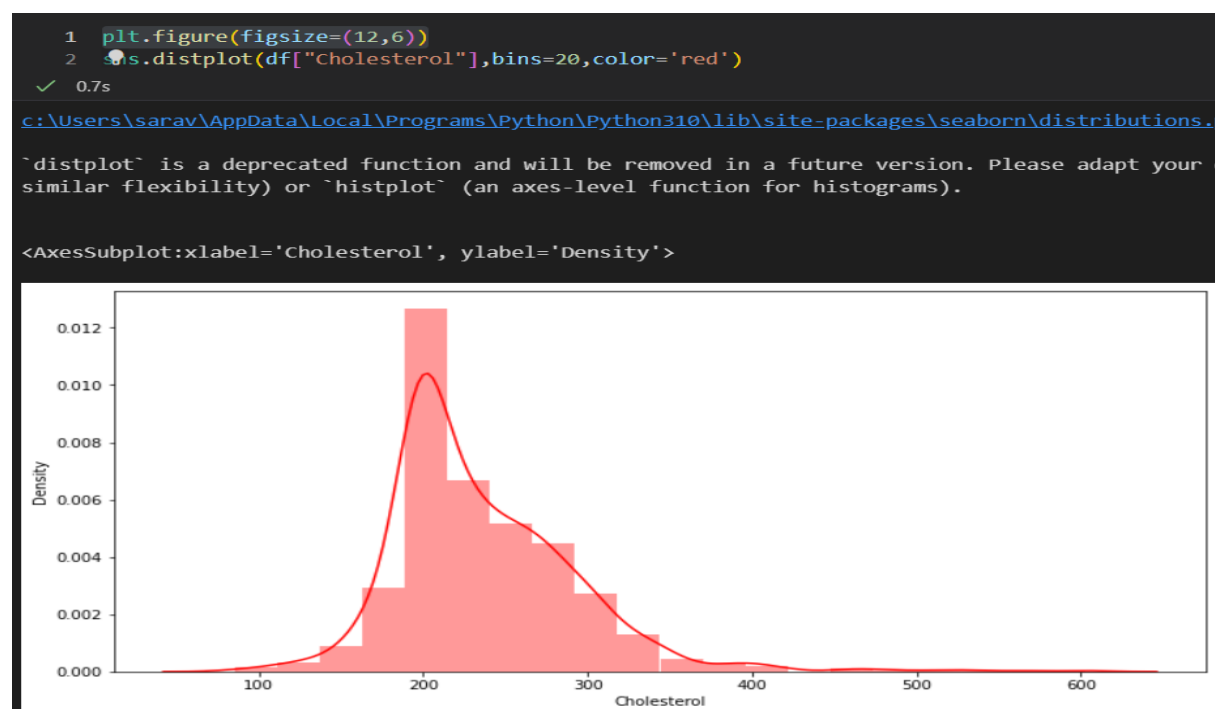
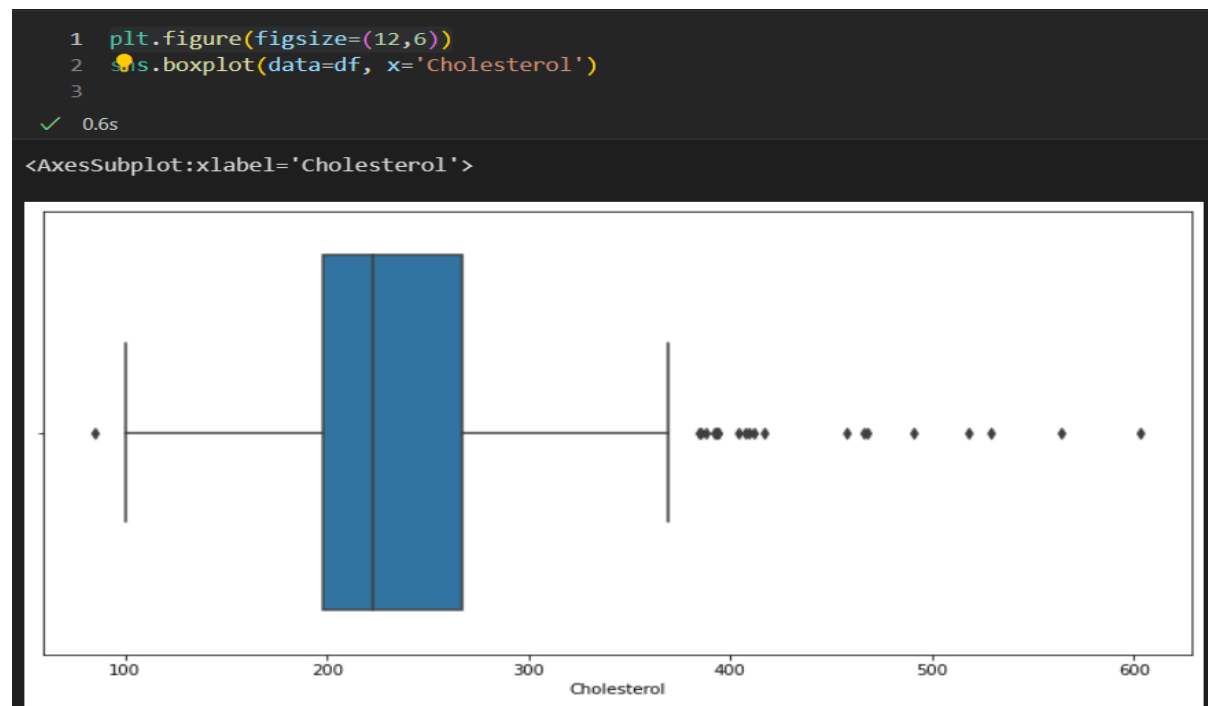
c:\Users\sarav\AppData\Local\Programs\Python\Python310\lib\site-packages\seaborn\distributions.p

`distplot` is a deprecated function and will be removed in a future version. Please adapt your c
similar flexibility) or `histplot` (an axes-level function for histograms).

<AxesSubplot:xlabel='Cholesterol', ylabel='Density'>



After changing to mean



The Cholesterol have the zero value in the dataset. In medically this is not a correct one so I am changed to the mean value in the respective column.

CHAPTER IV: EXPLORATORY DATA ANALYSIS

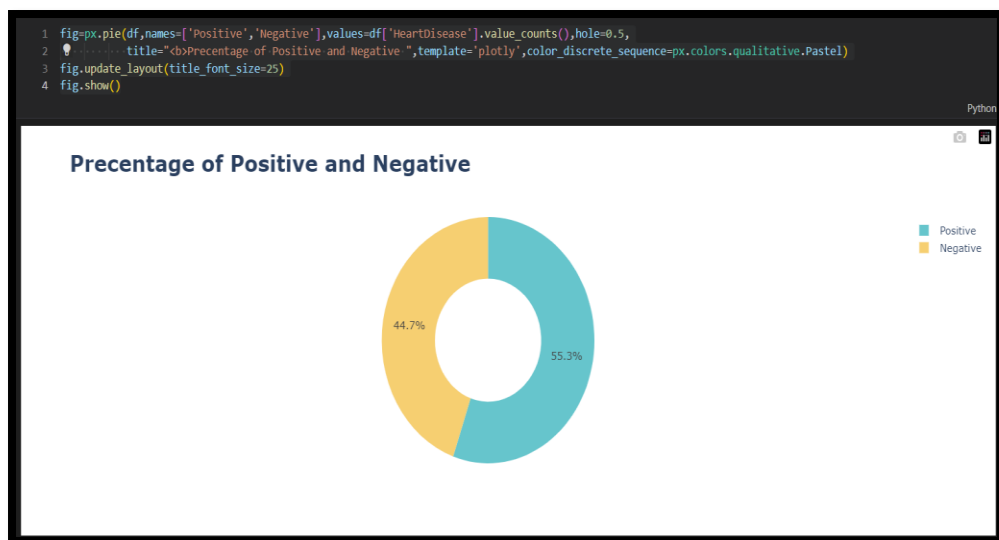
4.1 Data Exploration

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

EDA is primarily used to see what data can reveal beyond the formal modelling or hypothesis testing task and provides a better understanding of data set variables and the relationships between them. It can also help determine if the statistical techniques you are considering for data analysis are appropriate. Originally developed by American mathematician John Tukey in the 1970s, EDA techniques continue to be a widely used method in the data discovery process today.

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.

Data scientists can use exploratory analysis to ensure the results they produce are valid and applicable to any desired business outcomes and goals. EDA also helps stakeholders by confirming they are asking the right questions. EDA can help answer questions about standard deviations, categorical variables, and confidence intervals. Once EDA is complete and insights are drawn, its features can then be used for more sophisticated data analysis or modelling, including machine learning.



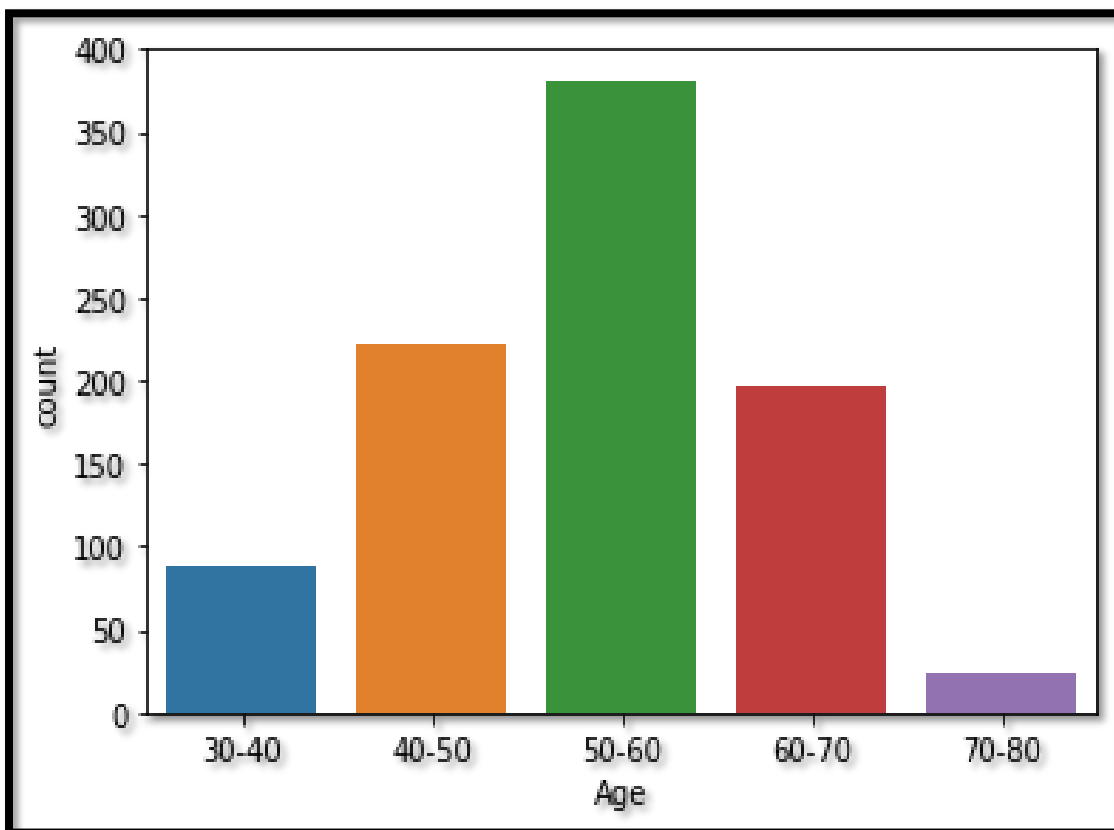
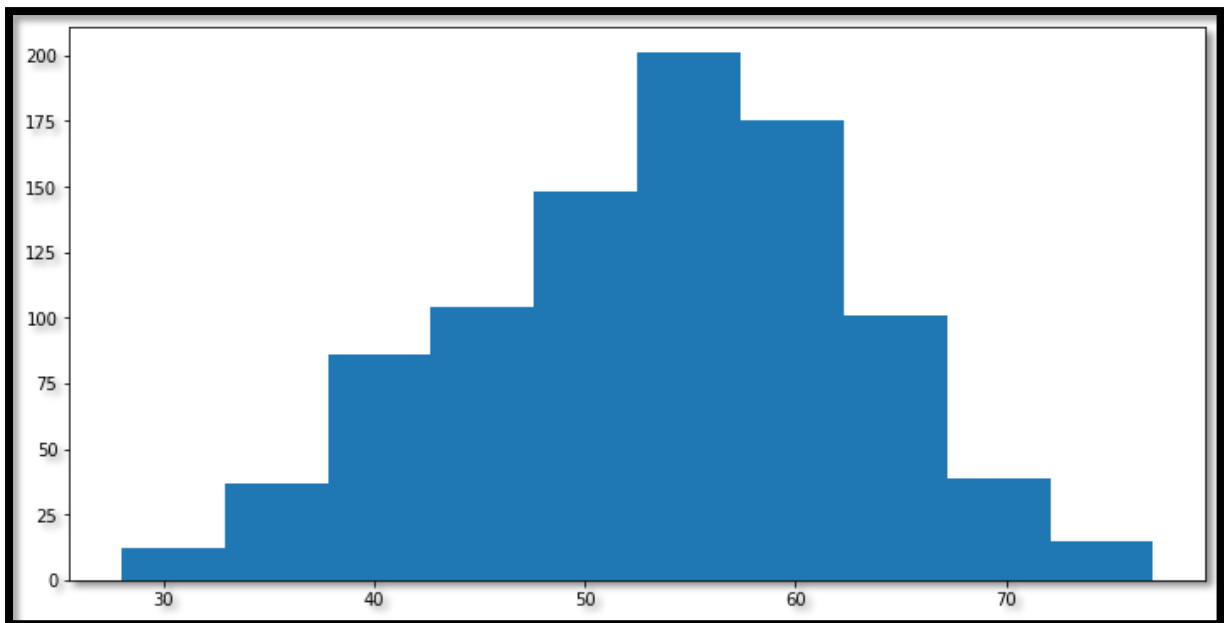
This is the pie chart is used to easily understand the how many percentages are there in positive case and negative case. Finally 44.7% in negative case and 55.3% in positive case occurs in dataset

Heat map



Each cell in the heat map represents the correlation coefficient between two variables. The color of the cell indicates the strength of the correlation, with blue indicating a negative correlation, yellow indicating a positive correlation, and green indicating no correlation. The shade of the color represents the magnitude of the correlation coefficient, with darker shades representing stronger correlations

Age

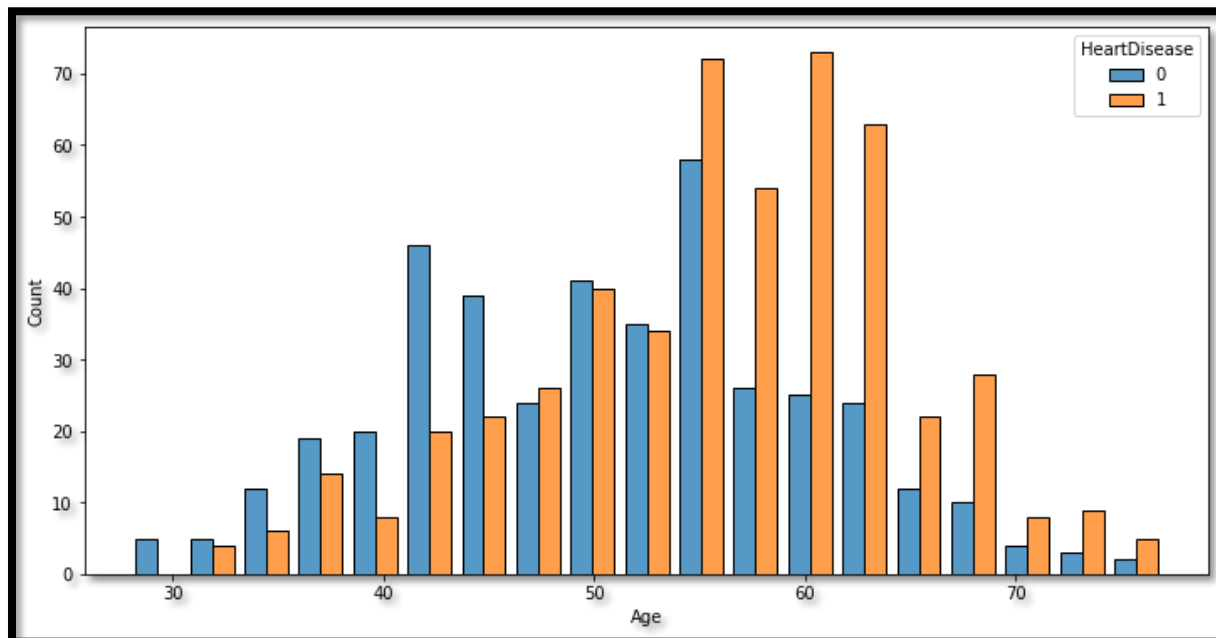


```
1 fage = df.loc[(df["Age"] >= 40) & (df["Age"] <= 70)]  
  
1 fage.count()
```

Age	814
Sex	814
ChestPainType	814
RestingBP	814
Cholesterol	814
FastingBS	814
RestingECG	814
MaxHR	814
ExerciseAngina	814
Oldpeak	814
ST_Slope	814
HeartDisease	814

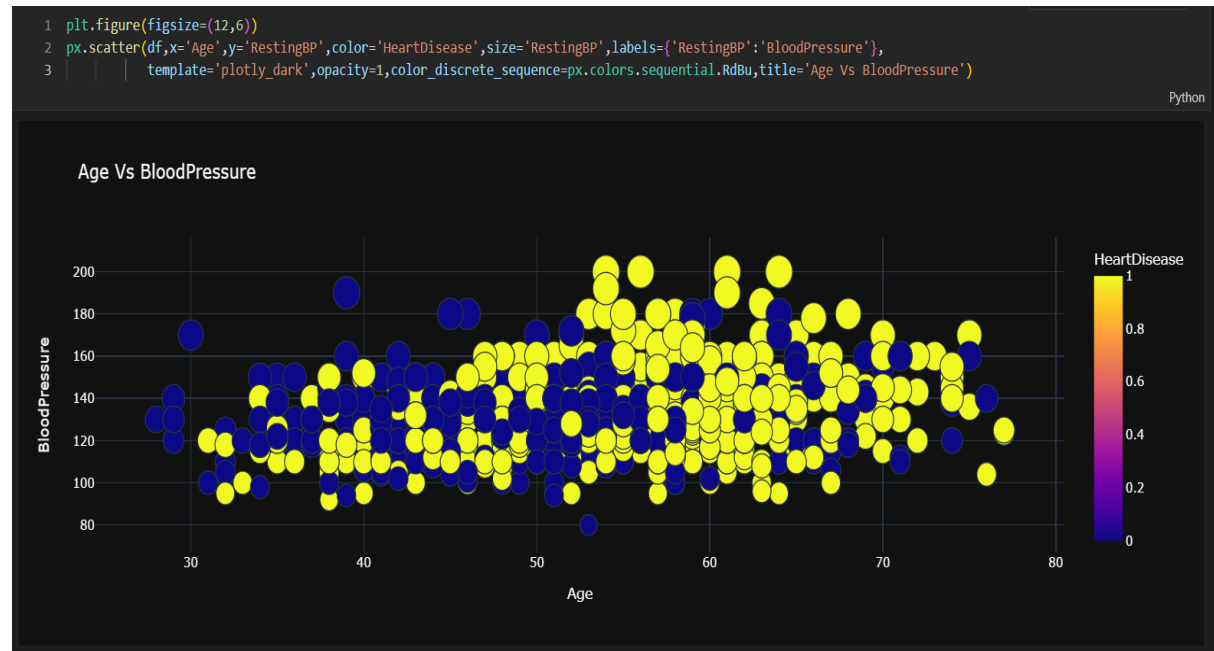
dtype: int64

In the above plot shows that people in the dataset are age between 28 to 77. But most of them are in age group 40 to 70



This plot shows that most of the patient are affected by heart disease are above 50.

Age vs RestingBP



Above the chart shows the who are all having the blood pressure more than 120 and age between 40 – 70 these category peoples are the most affected in heart disease

Age vs MaxHR



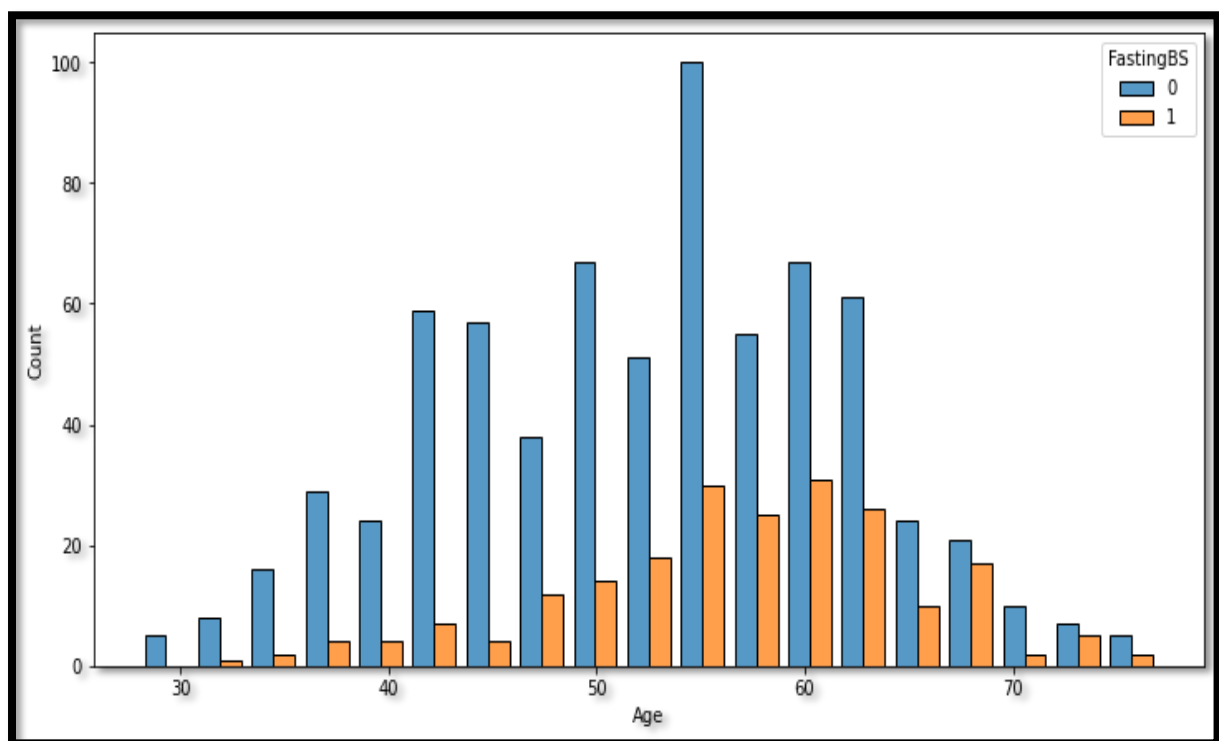
Above the chart shows the who are all having the Heartrate more than 100 and age between 40 – 70 these category peoples are the most affected in heart disease

```
1 ghr = df.loc[df["MaxHR"] >= 100 ]

1 ghr.count()

Age            847
Sex            847
ChestPainType  847
RestingBP      847
Cholesterol    847
FastingBS      847
RestingECG     847
MaxHR          847
ExerciseAngina 847
Oldpeak        847
ST_Slope       847
HeartDisease   847
dtype: int64
```

Age vs FastingBS

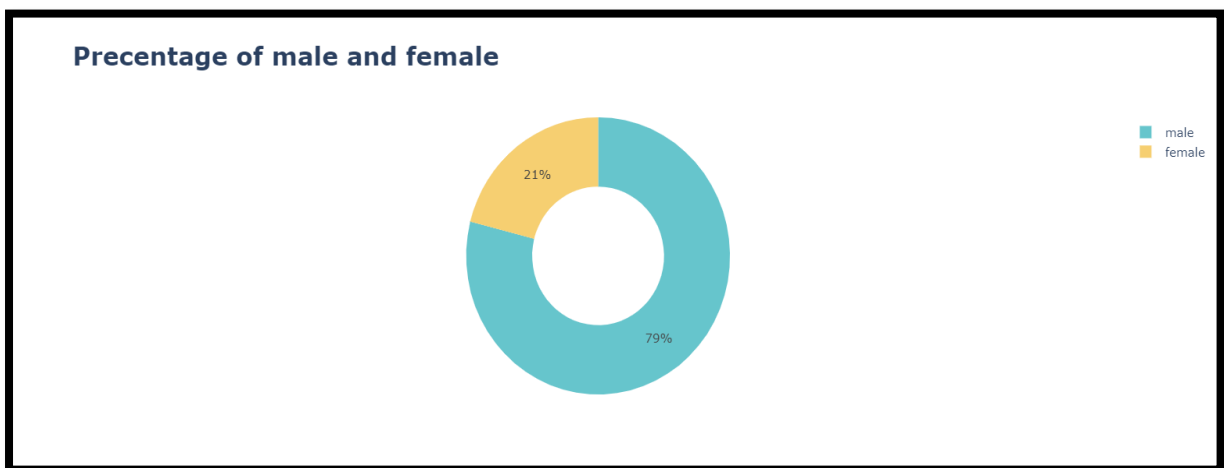


The above plot shows that fasting blood sugar is normal for all the age category people in the dataset.

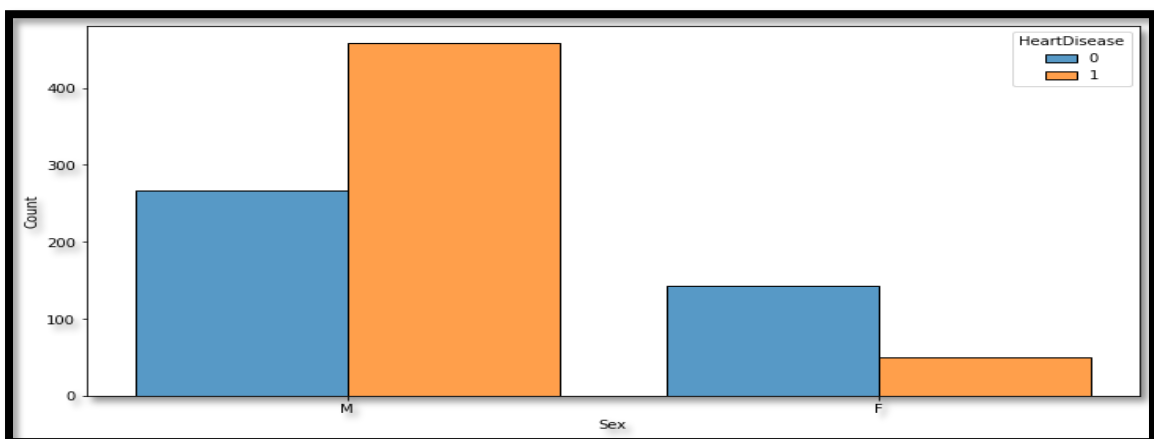
Gender



In the given dataset count of men are higher than woman

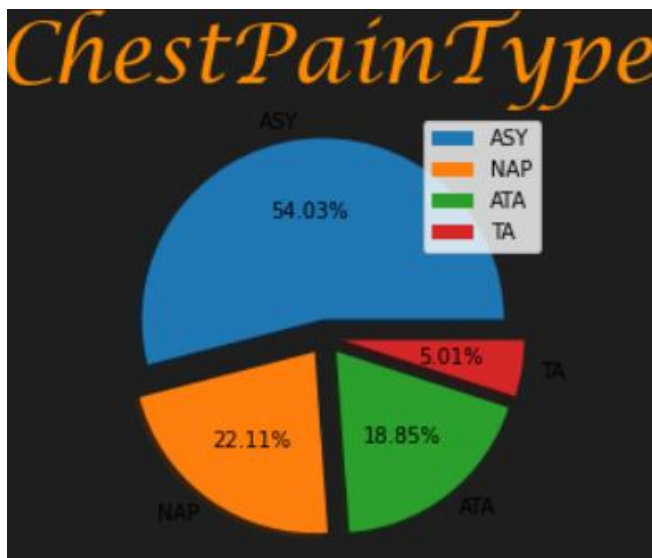


This is the pie chart is used to easily understand the percentage of men are 79% and 21% are women in dataset.

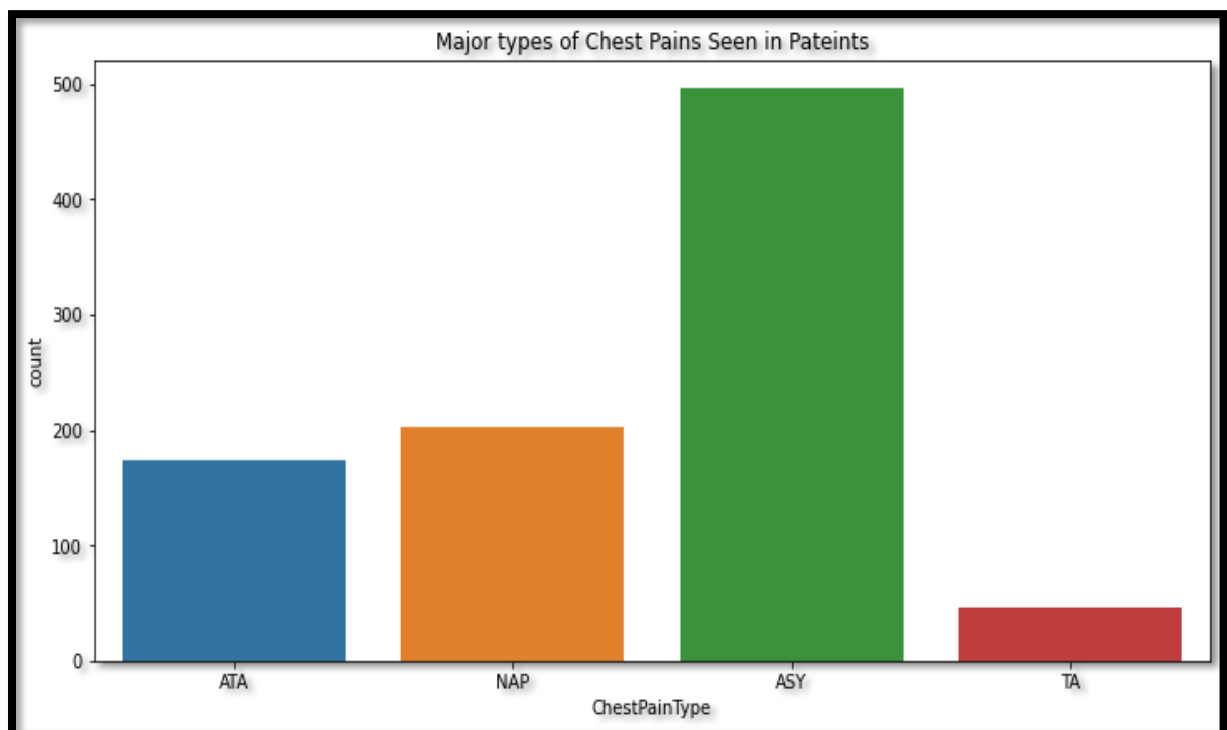


The above plot shows that men are more affected by heart disease than women

ChestPianType



The pie chart shows the Percentage of chest pain type ASY is 54.03%, chest pain type NAP is 22.11%, chest pain type ATA 18.85% and chest pain type TA is 5.01%

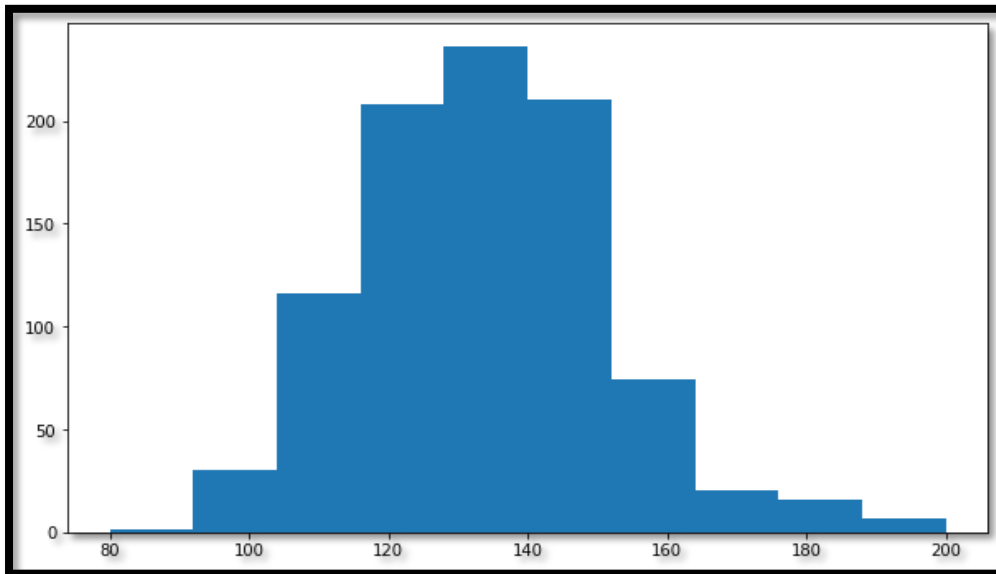


The most of the people are affected by chest pain type ASY in data set

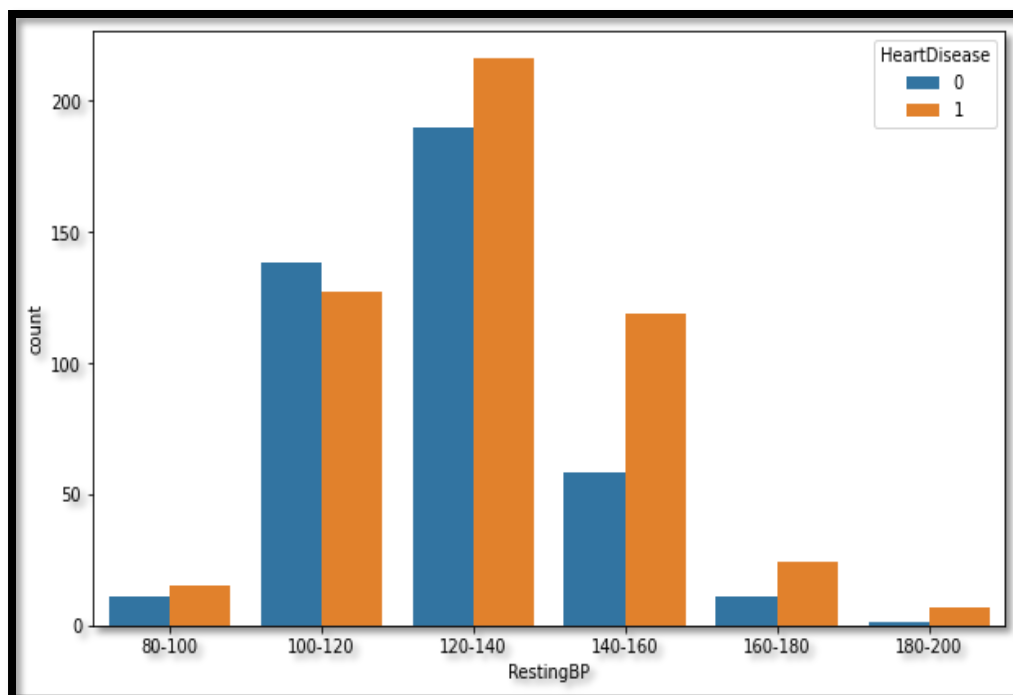
ASY = A silent heart attack is a heart attack that has few, if any, symptoms or has symptoms not recognized as a heart attack. A silent heart attack might not cause chest pain or shortness of breath, which are typically associated with a heart attack

RestingBP

Blood pressure is the pressure of blood pushing against the walls of your arteries. Arteries carry blood from your heart to other parts of your body. Your blood pressure normally rises and falls throughout the day. The normal range of blood pressure is 80 – 120.



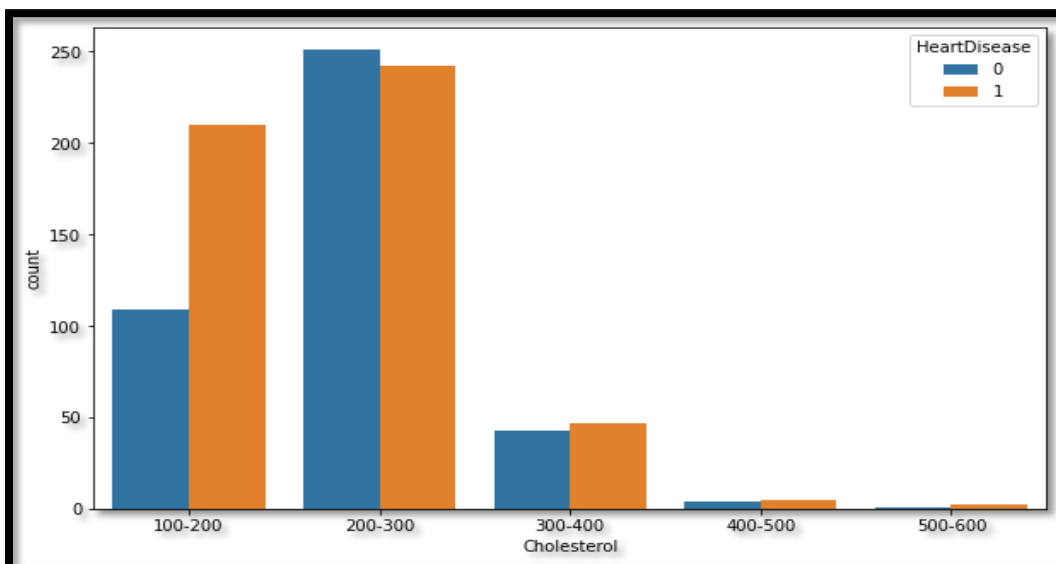
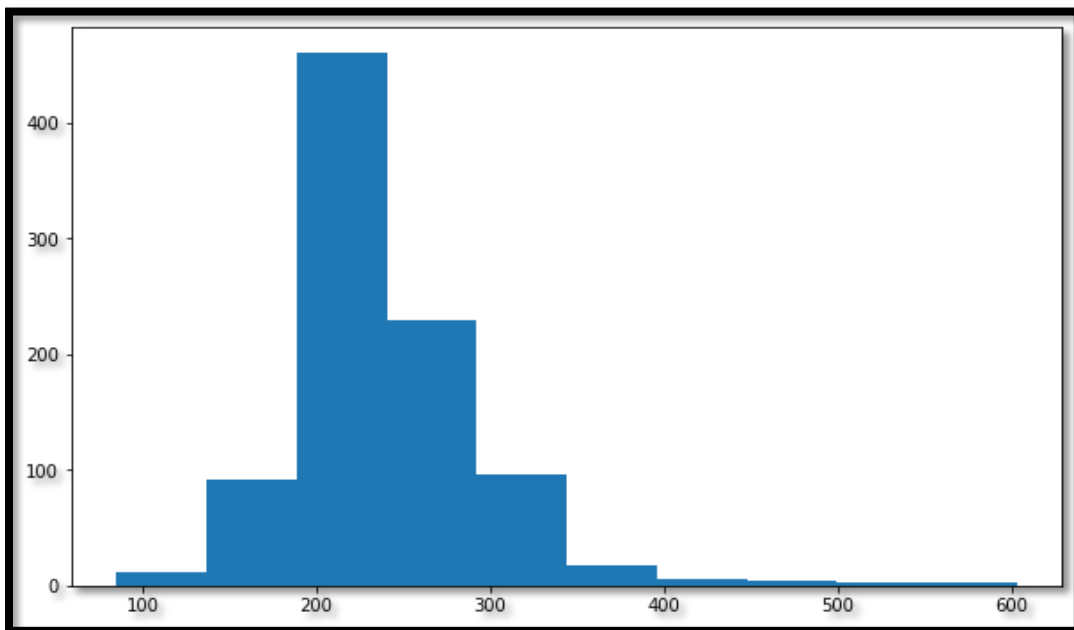
In this plots shows that distribution between 120 - 150



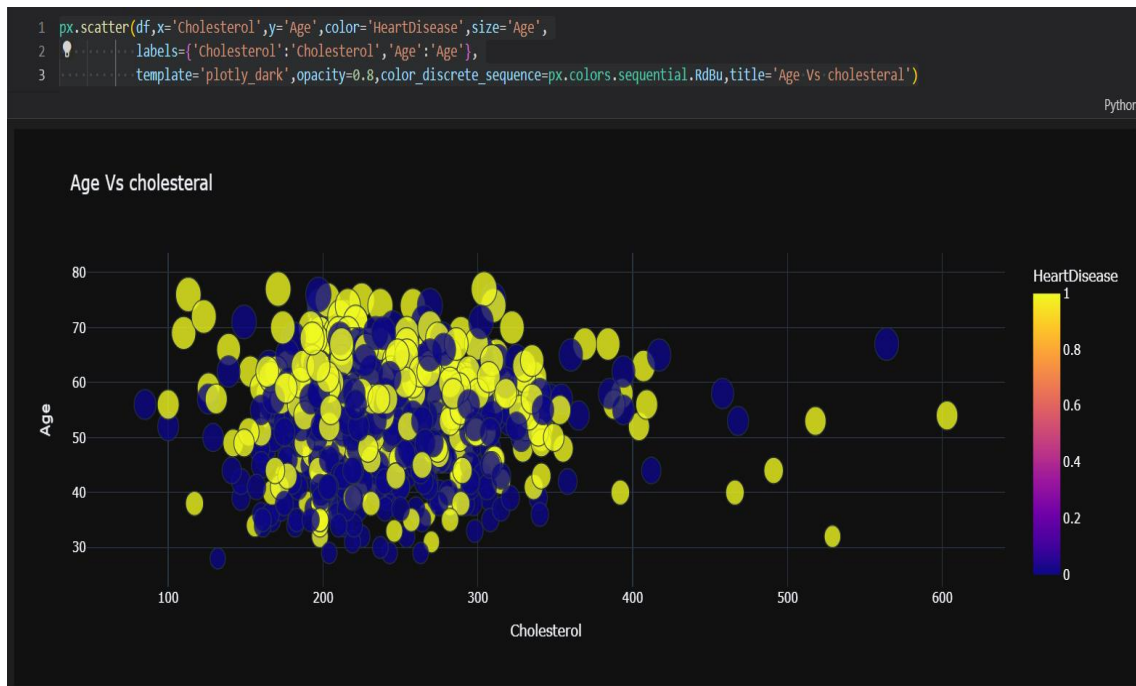
The people who have blood pressure range above 120 which is higher than the normal range are affected by HeartDisease

Cholesterol

The total serum cholesterol level is the amount of cholesterol in the blood. A high serum cholesterol level is a concern, because it raises your risk of heart disease. When people talk about getting their cholesterol checked or finding out their cholesterol levels, they are usually referring to serum cholesterol levels. The normal cholesterol is less than 200.



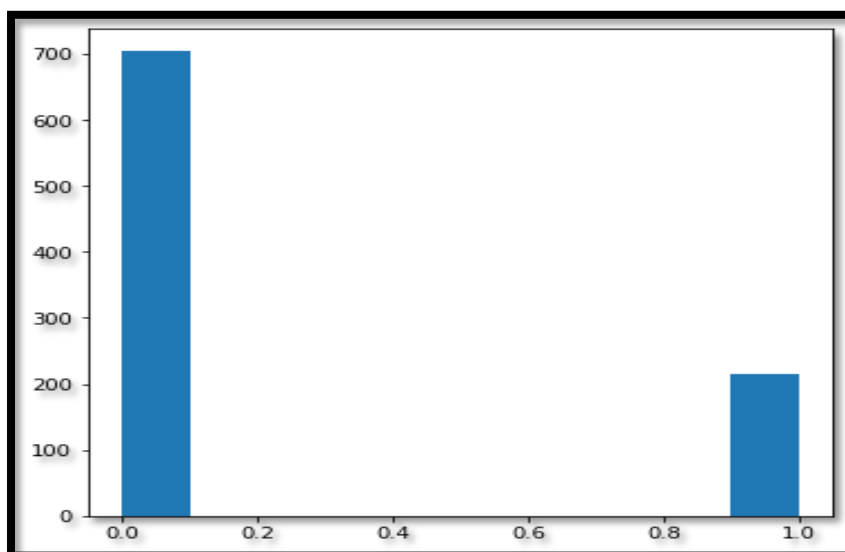
Above plot shows that the people who have cholesterol range above 200 are having high possibility to affect the heart disease



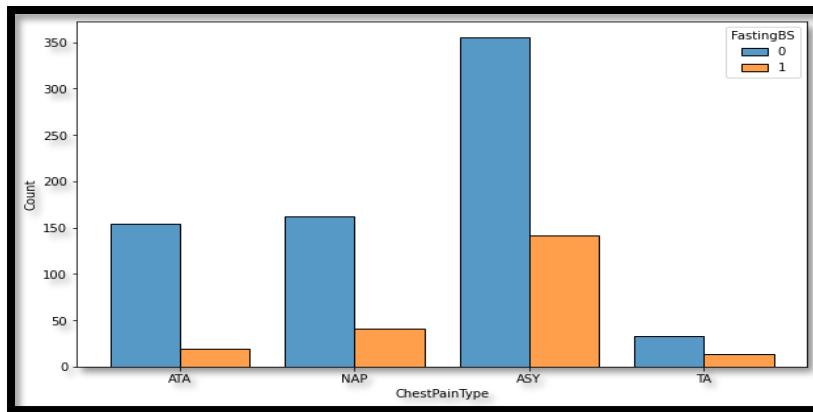
In This plot shows who all are having the cholesterol range above 200 and age between 40-70 having a more chance to affected to the heart disease

FastingBS

Blood sugar tends to peak about an hour after eating and declines after that. Fasting blood sugar levels are measured several hours after eating, as this gives a more accurate view of a person's glucose levels. In most cases, fasting blood sugar levels should be below 5.7%.

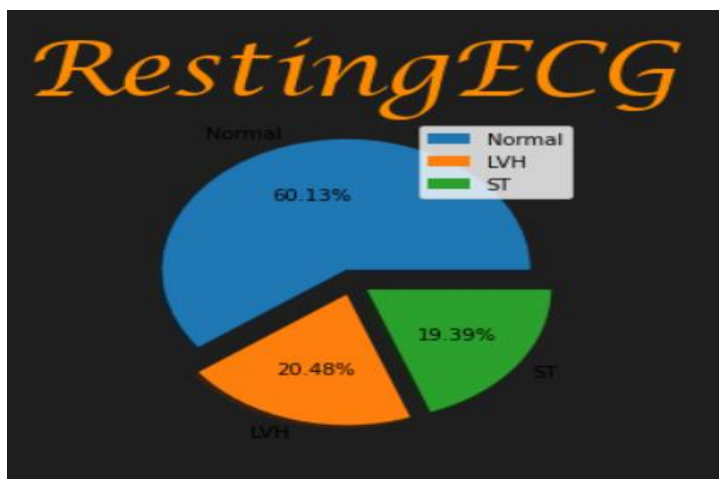


This plots shows the people are who have FastingBS is lower than the people don't have FastingBS

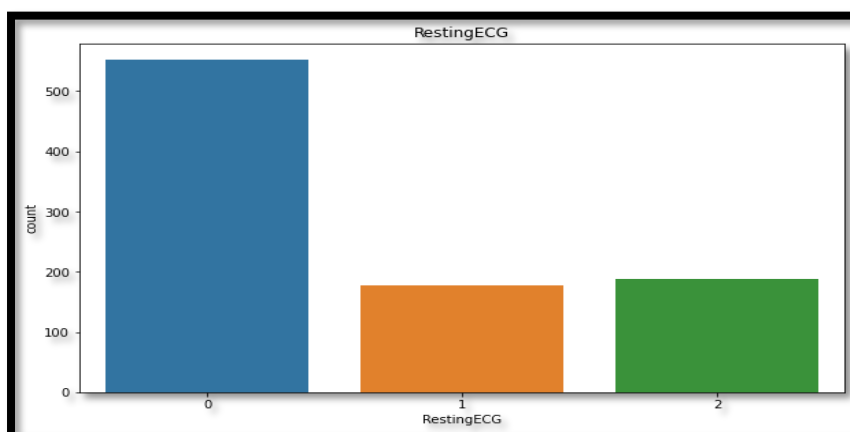


This plot shows that people having the chest pain type ASY having are affected with FastingBS comparably other chest pain type like ATA, NAP, TA.

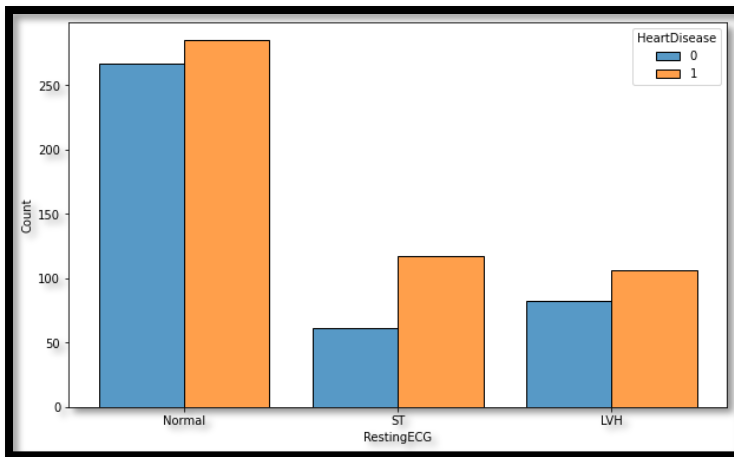
RestignECG



The pie chart shows the Percentage of RestingECG Normal is 60.03%, LVH is 20.48%, ST is 19.39.

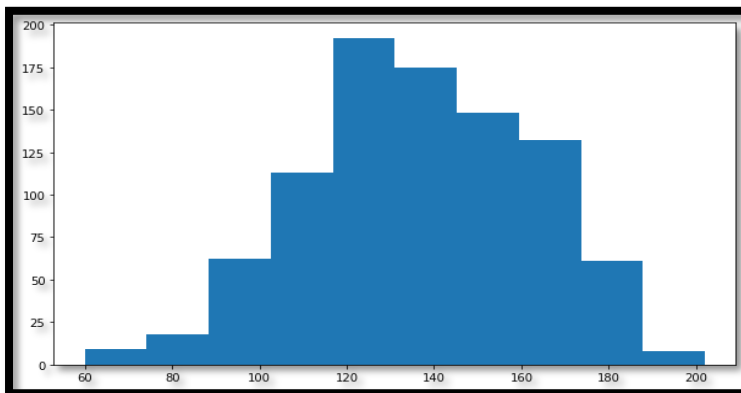


In this charts shows most of peoples RestingECG is normal stage compared to other RestingECG types like ST and LVH

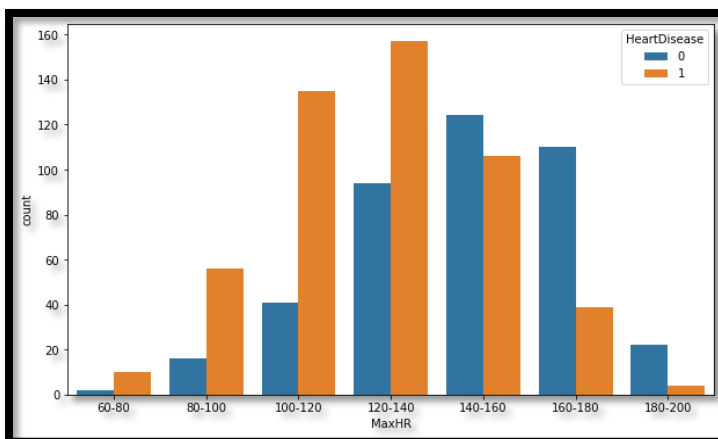


In this chart shows people who have and have not heart disease their ECG level is normal so it is not major component.

MaxHR

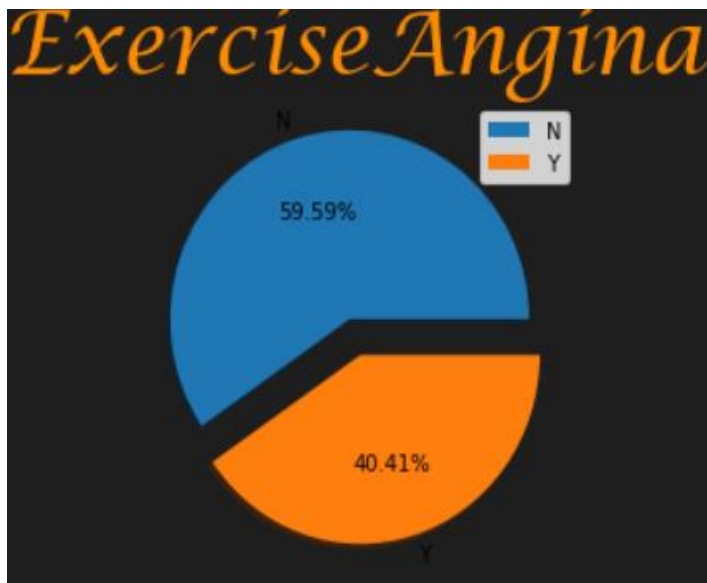


In this his plot shows that maximum range of MaxHR lies between 110 to 170.

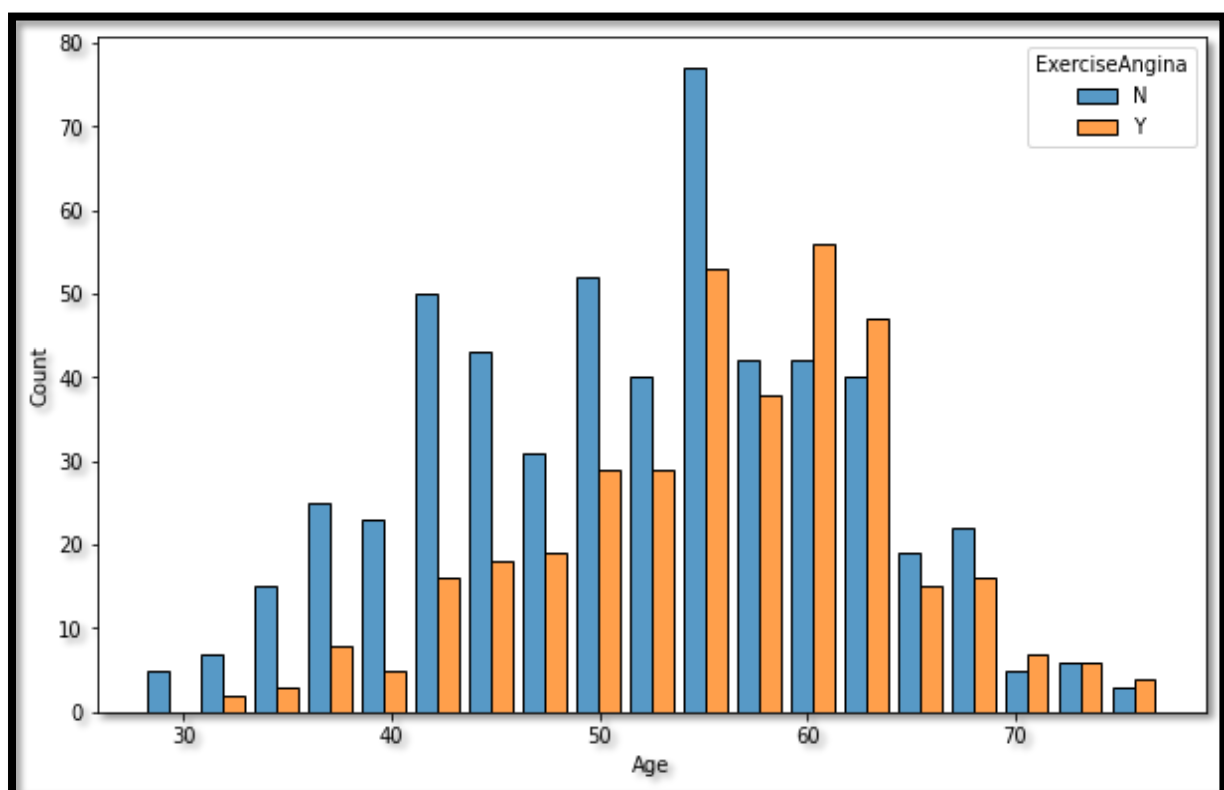


In this plot shows that people who are affected by heart disease having MaxHR lies between 100 – 160. The normal heart rate is less than 100.

ExerciseAngina

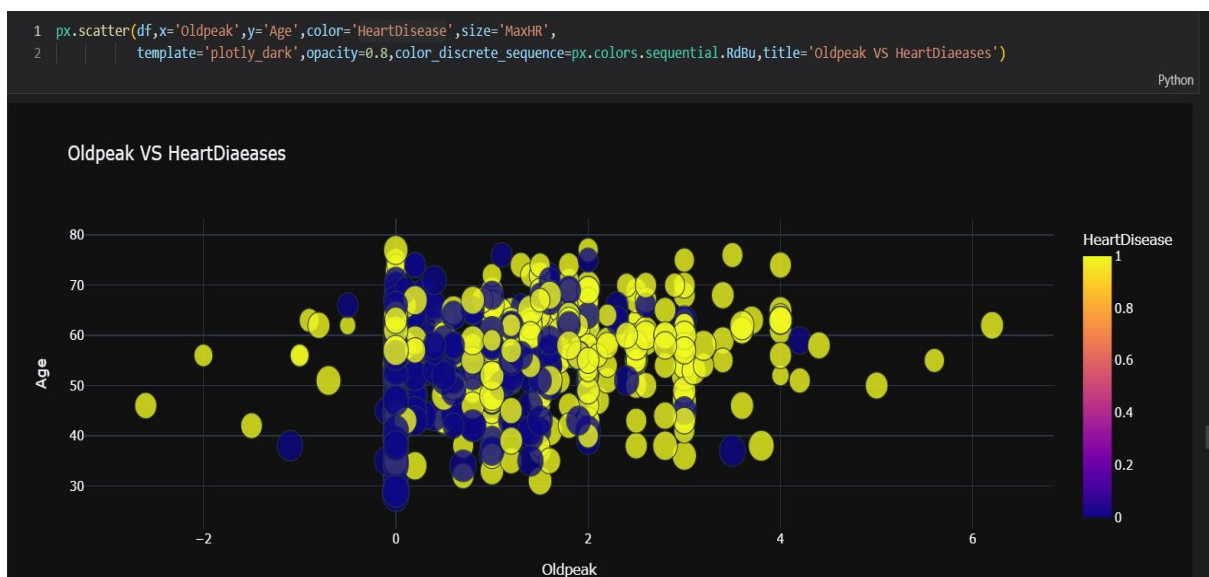
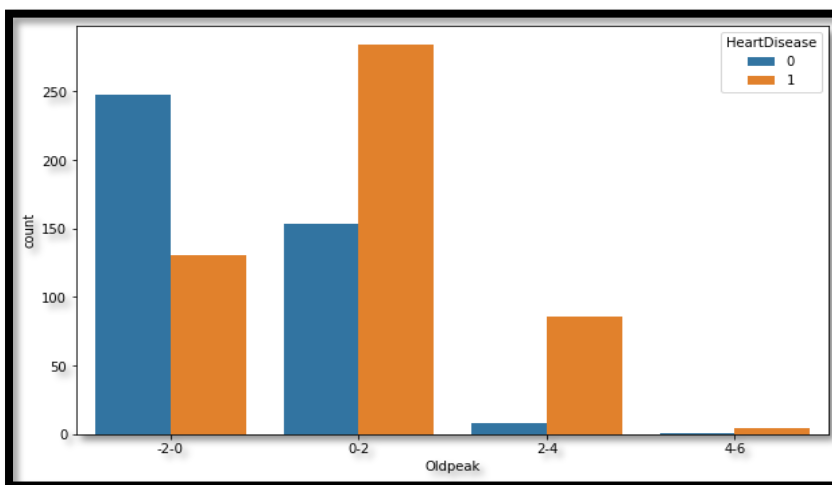
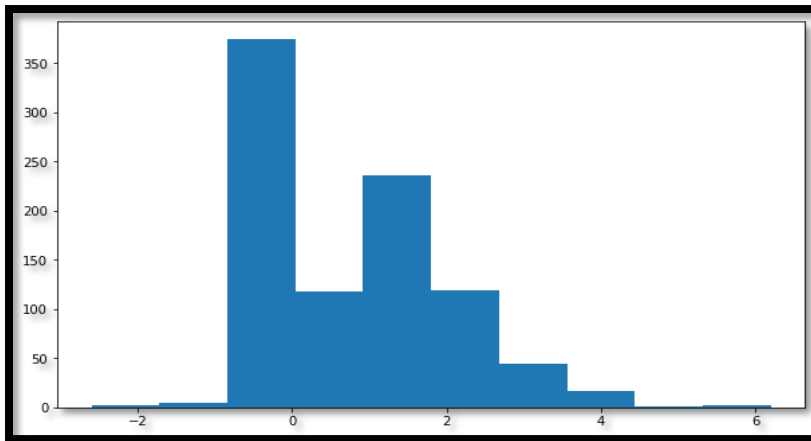


This pie charts shows that people who are affected with Exercise Angina in percentage of 40.41% While who don't have Exercise Angina are in percentage of 59.59%.



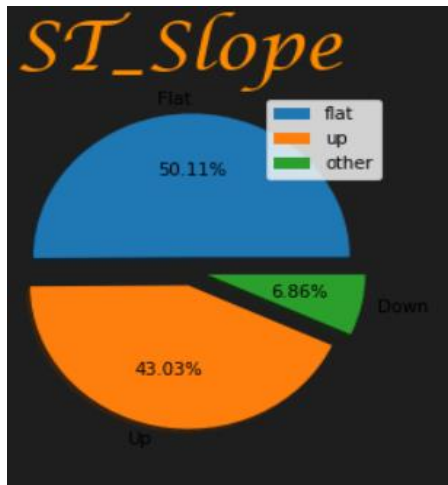
This plot shows that the people in the age of 55 to 65 are affected with Exercise Angina are higher than the other age group of people

Old peak

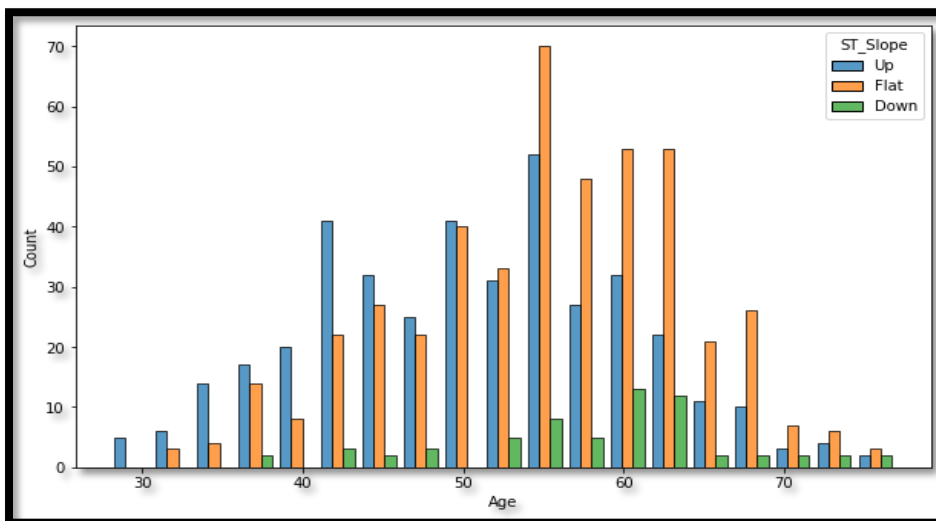


The above plot shows the people who are affected with heart disease their old peak second are above 1 second

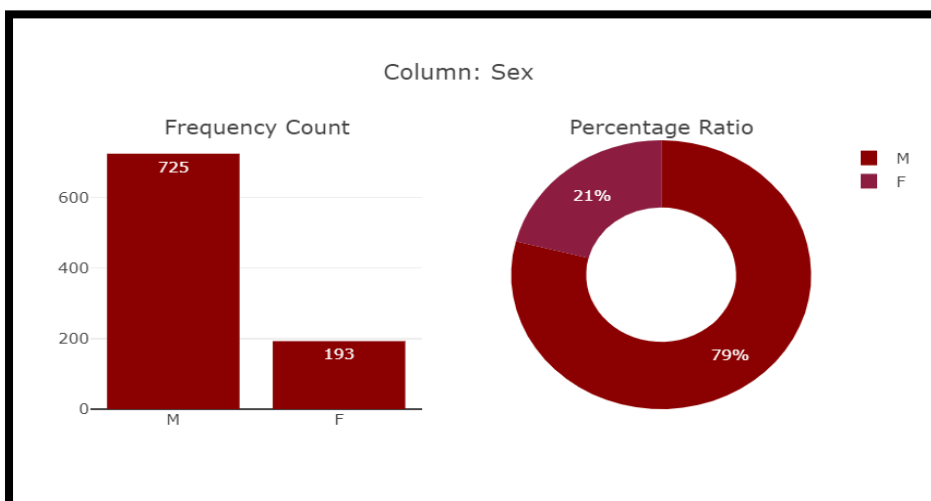
ST-Slope

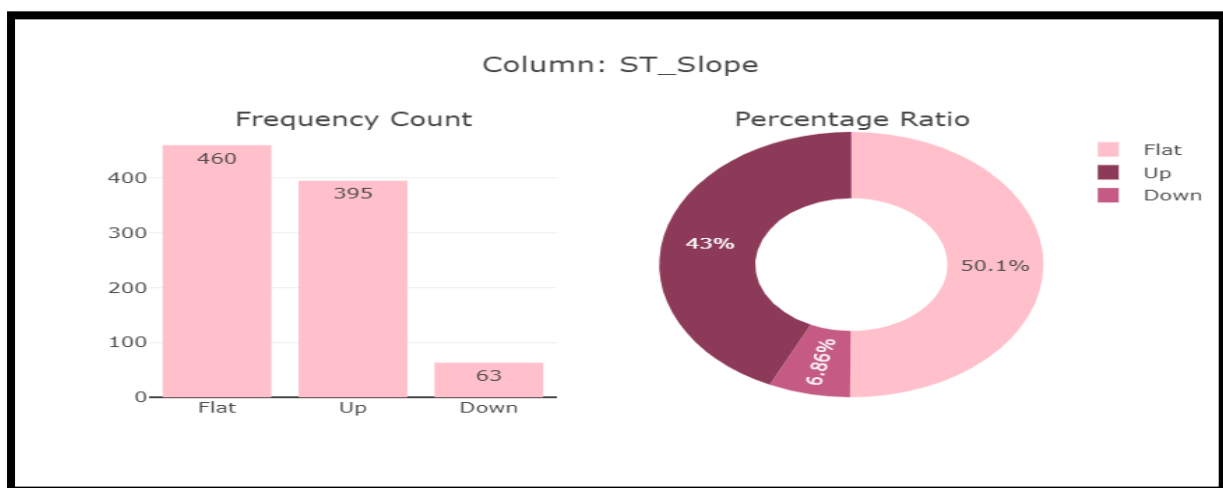
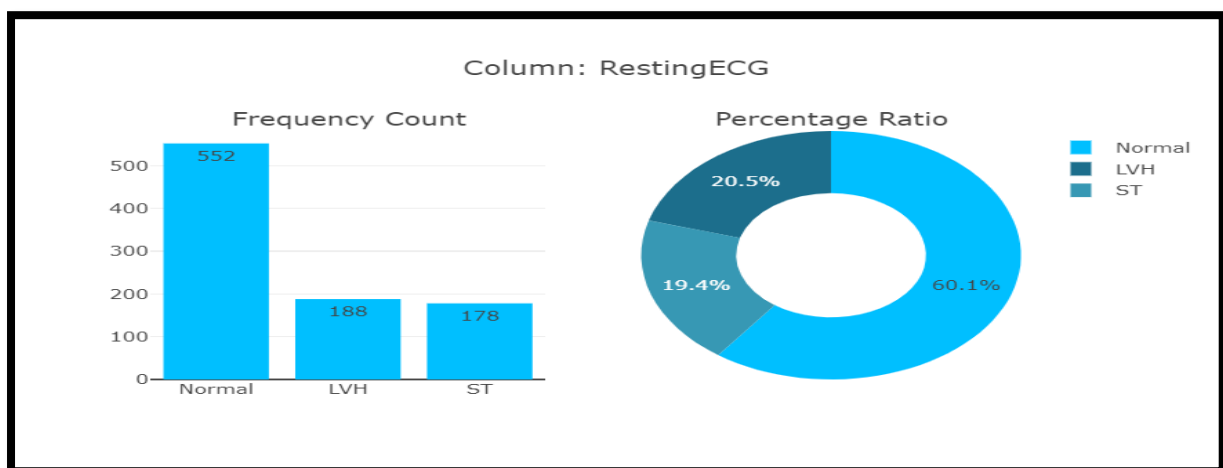
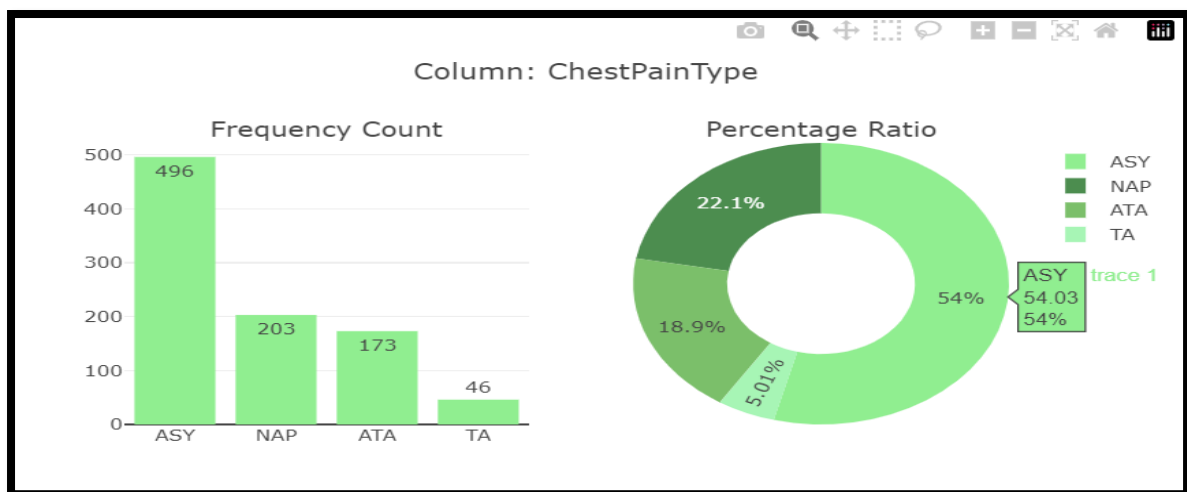


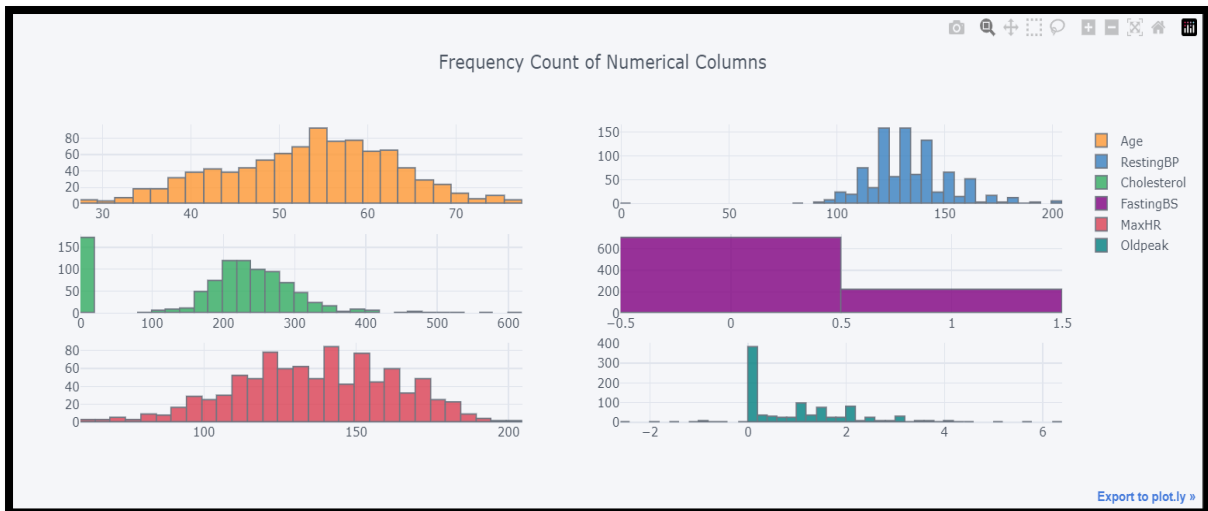
In this pie chart shows that ST _Slope FLAT is 50.11% , UP is 43.03%, OTHER 6.86%.



The above plot shows that ST_Slope FLAT is higher when the age is above 55.







According to above plots and table.

Male records are far more as compare to Female records.

More than 50% Chest Pain type is ASY.

Around 60% RestingECG is Normal.

Approx 60% doesn't have Exercise Angina.

50% of the ST_Slope is Flat.

Min Age is 28 and Max Age is 77, while the Avg. Age is between 50-60 years.

RestingBP is between 80 – 200

Max Heart Rate is 60 – 202

CHAPTER V: MODEL BUILDING

5.1 Algorithm Description

Random forest is an ensemble method that combines multiple decision trees to improve predictive accuracy and reduce overfitting. The algorithm starts by specifying the number of decision trees to include in the ensemble, as well as the maximum depth for each tree to prevent overfitting. For each decision tree, a random subset of the training examples is selected (with replacement) to create a bootstrap sample, and a random subset of the features is selected (without replacement) to use as potential split criteria at each node. The decision tree is then built using a recursive algorithm that selects the feature that results in the best split according to a predefined criterion, such as information gain or Gini impurity. This process is repeated for each decision tree in the ensemble, resulting in a set of uncorrelated models that can be combined to improve overall predictive accuracy. When making a prediction on new input data, each decision tree in the ensemble is used to make a prediction, and the final prediction is determined by taking the majority vote of the individual tree predictions.

Random forest algorithms have three main hyper parameters, which need to be set before training. These include node size, the number of trees, and the number of features sampled. From there, the random forest classifier can be used to solve for regression or classification problems.

The random forest algorithm is made up of a collection of decision trees, and each tree in the ensemble is comprised of a data sample drawn from a training set with replacement, called the bootstrap sample. Of that training sample, one-third of it is set aside as test data, known as the out-of-bag sample, which we'll come back to later. Another instance of randomness is then injected through feature bagging, adding more diversity to the dataset and reducing the correlation among decision trees. Depending on the type of problem, the determination of the prediction will vary. For a regression task, the individual decision trees will be averaged, and for a classification task, a majority vote—i.e. the most frequent categorical variable—will yield the predicted class. Finally, the oob sample is then used for cross-validation, finalizing that prediction.

5.2 Data Splitting

Label Encoding

There are some features which are in object datatype, we have to convert those to numeric data. For that we have to use encoding. There are various methods for encoding categorical data, including one-hot encoding, label encoding, and target encoding. I used label encoding here.

```
1 df.replace({"Sex":{"M":1,'F':0}},inplace=True)
2
✓ 0.4s

1 df.replace({"ChestPainType":{"TA":0,'ATA':1,'NAP':2,'ASY':3}},inplace=True)
✓ 0.1s

1 df.replace({"RestingECG":{"Normal":0,'ST':1,'LVH':2}},inplace=True)
✓ 0.1s

1 df.replace({"ExerciseAngina":{"N":0,'Y':1}},inplace=True)
2
✓ 0.1s

1 df.replace({"ST_Slope":{"Down":0,'Flat':1,'Up':2}},inplace=True)
✓ 0.0s
```

```
1 x = df.drop(columns='HeartDisease', axis=1 )
2 y = df['HeartDisease']
✓ 0.1s
```

5.3 Training and Test Data

A set of m training examples, where each example represents a patient and consists of a set of input features (such as age, gender, blood pressure, cholesterol level, etc.) and a corresponding binary label indicating whether the patient has heart disease or not.

The training data should be representative of the underlying population and should be chosen randomly to avoid bias. It is recommended to use a diverse set of patient demographics, as well as a range of feature values to ensure that the model can generalize to different patient populations.

Test Data:

A set of n test examples, where each example represents a new patient and consists of the same set of input features as the training data.

The test data should also be representative of the underlying population and should be chosen randomly to avoid bias. It is important to ensure that the test data does not overlap with the training data to avoid overfitting.

During the training process, the model learns the relationship between the input features and the binary label indicating whether the patient has heart disease or not. The trained model is then evaluated on the test data to assess how well it generalizes to new, unseen patients. The performance of the model is typically measured using metrics such as accuracy, precision, recall, and F1-score. In addition, it may be useful to evaluate the model using ROC curves and AUC metrics to assess its ability to discriminate between positive and negative cases.

```
1 x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=2)
✓ 0.1s
```

```
1 print( x_train.shape,y_train.shape, x_test.shape,y_test.shape)
✓ 0.0s
(734, 11) (734,) (184, 11) (184,)
```

```
1 forest= RandomForestClassifier(n_estimators =40, random_state = 0)
✓ 0.1s
```

[+ Code](#) [+ Markdown](#)

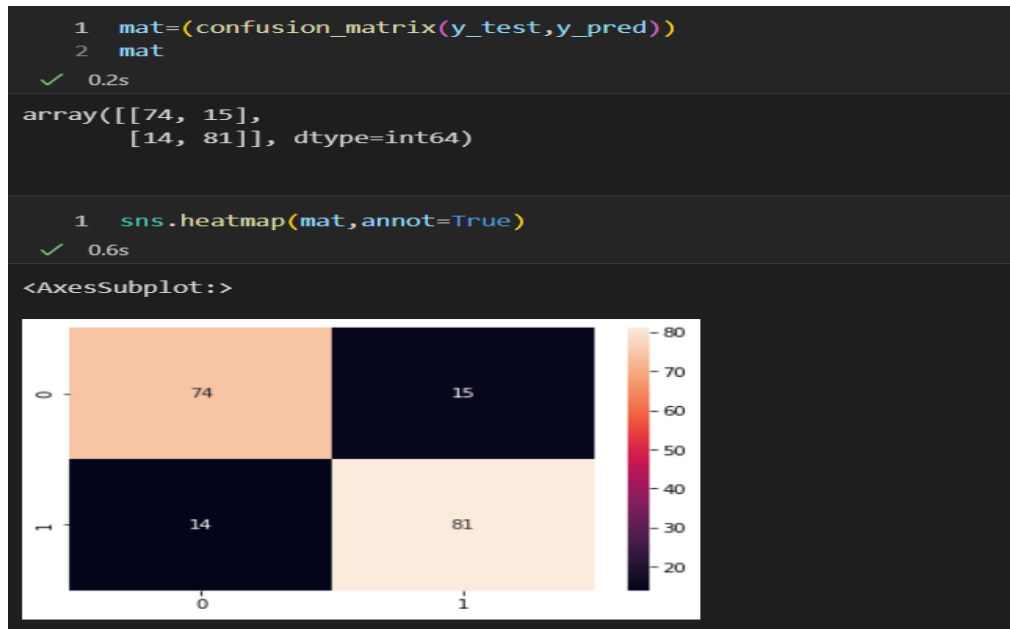
```
1 forest.fit(x_train,y_train)
✓ 0.5s
```

```
▼ RandomForestClassifier
RandomForestClassifier(n_estimators=40, random_state=0)
```

CHAPTER 6: EVALUATION OF MODEL

6.1 Performance of metrics

Confusion Matrix:



A confusion matrix is a table that shows the counts of true positives, false positives, true negatives, and false negatives for each class in a classification problem. It is useful for evaluating the performance of a machine learning model by comparing the predicted values with the actual values.

The `confusion matrix()` function from the `scikit-learn` library is used to create the confusion matrix. It takes two parameters: the actual target values and the predicted target values. In this case, `y_test` contains the actual target values, and `y_pred` contains the predicted target values generated by the machine learning model.

The resulting matrix, stored in the variable "mat", has the actual target values as the rows and the predicted target values as the columns. The diagonal values of the matrix represent the number of correctly classified instances for each class, while the off-diagonal values represent the number of misclassified instances.

By analysing the values in the confusion matrix, we can calculate various performance metrics of the machine learning model such as accuracy, precision, recall, and F1-score. These

metrics help to evaluate the effectiveness of the model in classifying instances into their respective classes

Mean Squared Error

MSE measures the average squared difference between the predicted and actual values. A lower value of MSE indicates better model performance.

```
1 mse = mean_squared_error(y_test, y_pred)
2
```

```
1 mse
```

```
0.15760869565217392
```

6.2 Performance of model

```
1 print("Accuracy:",forest.score(x_test,y_test))
✓ 0.1s
```

```
Accuracy: 0.842391304347826
```

```
1 print(classification_report(y_test,y_pred))
✓ 0.1s
```

	precision	recall	f1-score	support
0	0.84	0.83	0.84	89
1	0.84	0.85	0.85	95
accuracy			0.84	184
macro avg	0.84	0.84	0.84	184
weighted avg	0.84	0.84	0.84	184

Accuracy for Random Forest Regression model

The Random Forest model is best-fit because accuracy in Random Forest Regression model is 84percentage. and Precision is one indicator of a machine learning model's performance – the quality of a positive prediction made by the model accuracy is 84%.

CHAPTER 7: PREDCTION AND INFERENCE

7.1 Prediction

The model used for prediction is Random Forest. This model predict with the accuracy score of 84%

```
1 input_data= (44,1,1,150,288,0,0,150,1,3,1)
2 input_data_as_numpy_array = np.asarray(input_data)
3 input_data_reshape = input_data_as_numpy_array.reshape(1,-1)
4 prediction = forest.predict(input_data_reshape)
5 print(prediction)
6
7
8 if (prediction[0]== 0):
9     print('have not heart problem')
10 else:
11     print('have problem')
✓ 0.4s
[1]
have problem
```

The input data are given from the dataset. They are heart disease prediction dataset contains 12 columns. Now excluded the target column, remaining all respective columns are considered.

7.2 Inference:

1. Age: In the above plot shows that people in the dataset are age between 28 to 77. But most of them are in age group 40 to 70

2. Sex: Men are more affected by heart disease than women

3. ChestPainType: The most of the people are affected by chest pain type ASY in data set

ASY = A silent heart attack is a heart attack that has few, if any, symptoms or has symptoms not recognized as a heart attack. A silent heart attack might not cause chest pain or shortness of breath, which are typically associated with a heart attack

4. Blood Pressure: The people who have blood pressure range above 120 which is higher than the normal range are affected by HeartDisease

5. cholesterol: who all are having the cholesterol range above 200 and age between 40-70 having a more chance to affected to the heart disease

CHAPTER 8: MODEL DEPLOYMENT

8.1 Importing the relevant package for model deployment

```
1 import pickle
2 import streamlit as st
3 from streamlit_option_menu import option_menu
```

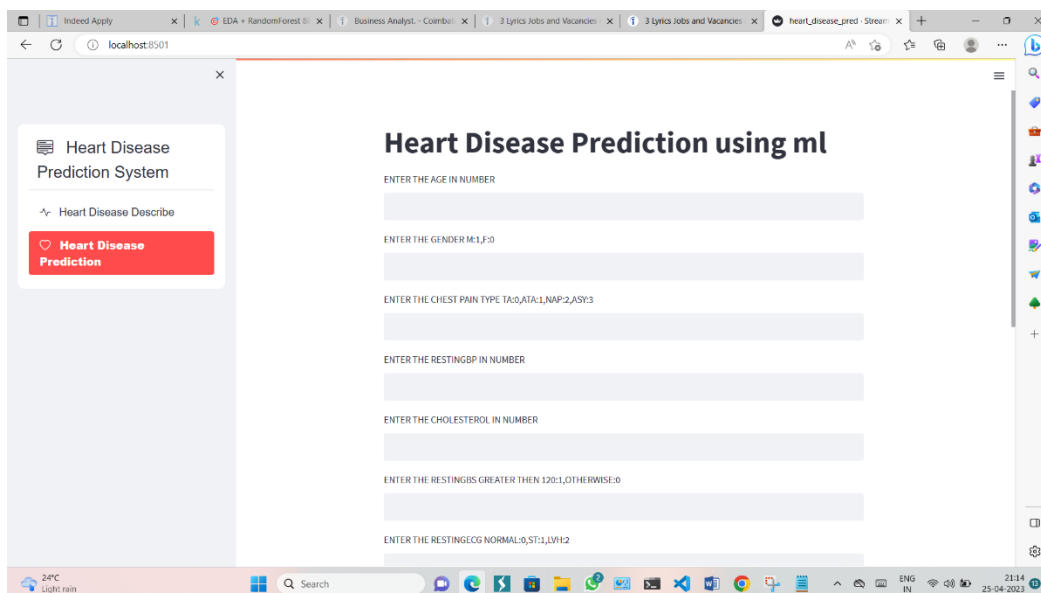
Pickle:

Pickle in Python is primarily used in serializing and deserializing a Python object structure. In other words, it's the process of converting a Python object into a byte stream to store it in a file/database, maintain program state across sessions, or transport data over the network.

Streamlit:

Streamlit is a promising open-source Python library, which enables developers to build attractive user interfaces in no time. Streamlit is the easiest way especially for people with no front-end knowledge to put their code into a web application: No front-end (html, js, css) experience or knowledge is required.

8.2 webpage on heart prediction



Machine Learning model is software that can learn and respond through experiences. Machine Learning has a wide range of applications in various fields of technology and science. Many top companies give preference to machine learning as one of the most important domains. For example, Amazon, Netflix, Facebook, YouTube all these companies use Machine Learning algorithms to improve customer experience. Since Machine Learning is a much established. The ultimate goal of the Machine Learning model is to make use of provided data and make better predictions. There are some steps to build a machine learning model

One of the most important and final steps in building a Machine Learning project is Model deployment. There are many frameworks available for deploying the Machine learning model on the web. Some of the most used Python frameworks are Django and Flask. But these frameworks require a little knowledge of languages such as HTML, CSS, and JavaScript. So, a new framework known as Streamlit was introduced to deploy the Machine Learning model without the need of having the knowledge of Front End Languages. It is quite easy to deploy using Streamlit.

CHAPTER 9: CONCLUSION

Heart diseases are one of the major concerns of society and the number of people affected by these diseases is increasing day by day and it is important to find a solution to this problem.

It is difficult to manually determine the odds of getting heart disease based on risk factors. But with the help of data analytics and machine learning models, we can determine these diseases and have a better chance of treating it. This project predicts people with cardiovascular disease by extracting the patient medical history that leads to a fatal heart disease from a dataset that includes patients' medical history such as chest pain, sugar level, blood pressure, etc. This Heart Disease detection system assists a patient based on his/her clinical information of them been diagnosed with a previous heart disease. The algorithms used in building the given model are Random Forest Classifier. The accuracy of our model is 84%. Use of more training data ensures the higher chances of the model to accurately predict whether the given person has a heart disease or not . By using these, computer aided techniques we can predict the patient fast and better and the cost can be reduced very much. There are a number of medical databases that we can work on as these Machine learning techniques are better and they can predict better than a human being which helps the patient as well as the doctors

REFERENCES

1. Kaggle: -<https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction>
2. Medium:-<https://medium.com/analytics-vidhya/evaluating-a-random-forest-model-9d165595ad56>
3. Iopscience:-<https://iopscience.iop.org/article/10.1088/1757-899X/1022/1/012046>
4. National library of medicine:-
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9206502/>
5. Towards science:-<https://towardsdatascience.com/heart-disease-prediction-73468d630cfc>
6. Shiksha.com:-<https://www.shiksha.com/online-courses/articles/heatmap-in-seaborn/>