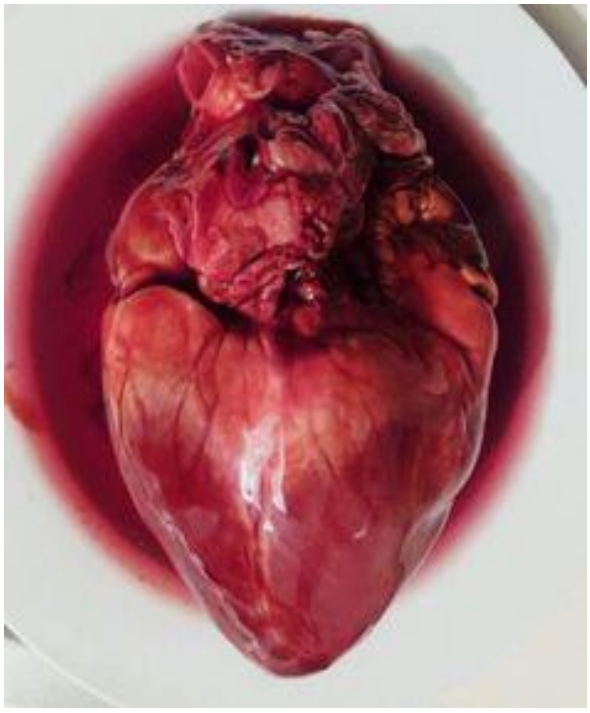
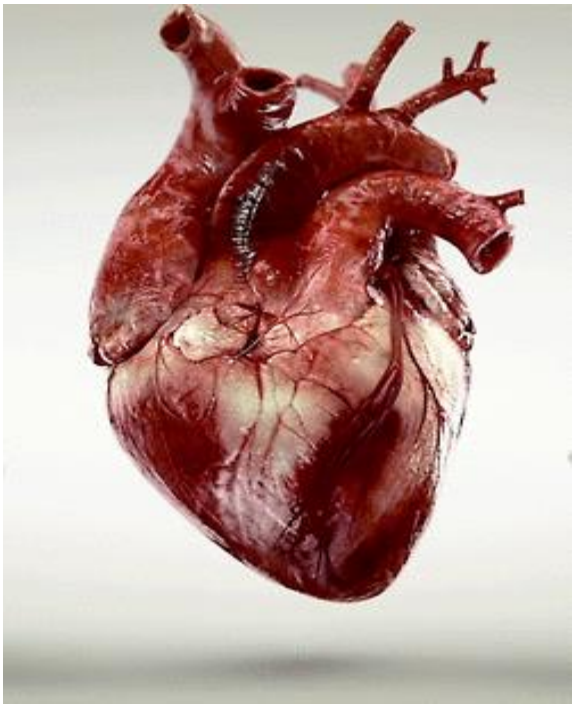
HEART ATTACK PREDICTION ANALYSIS





MADE BY :

SIDHARTH JAIN

JAI19481234

INDEX

INTRODUCTION

Human Heart is the mother of all the organs which supply clean blood to the body and make the system working. Researchers said that our beat around 1000,000 times and the it pump 2,000 gallons of the blood by the body. It's been said that we have 60,000 miles long blood vessels in a human body.

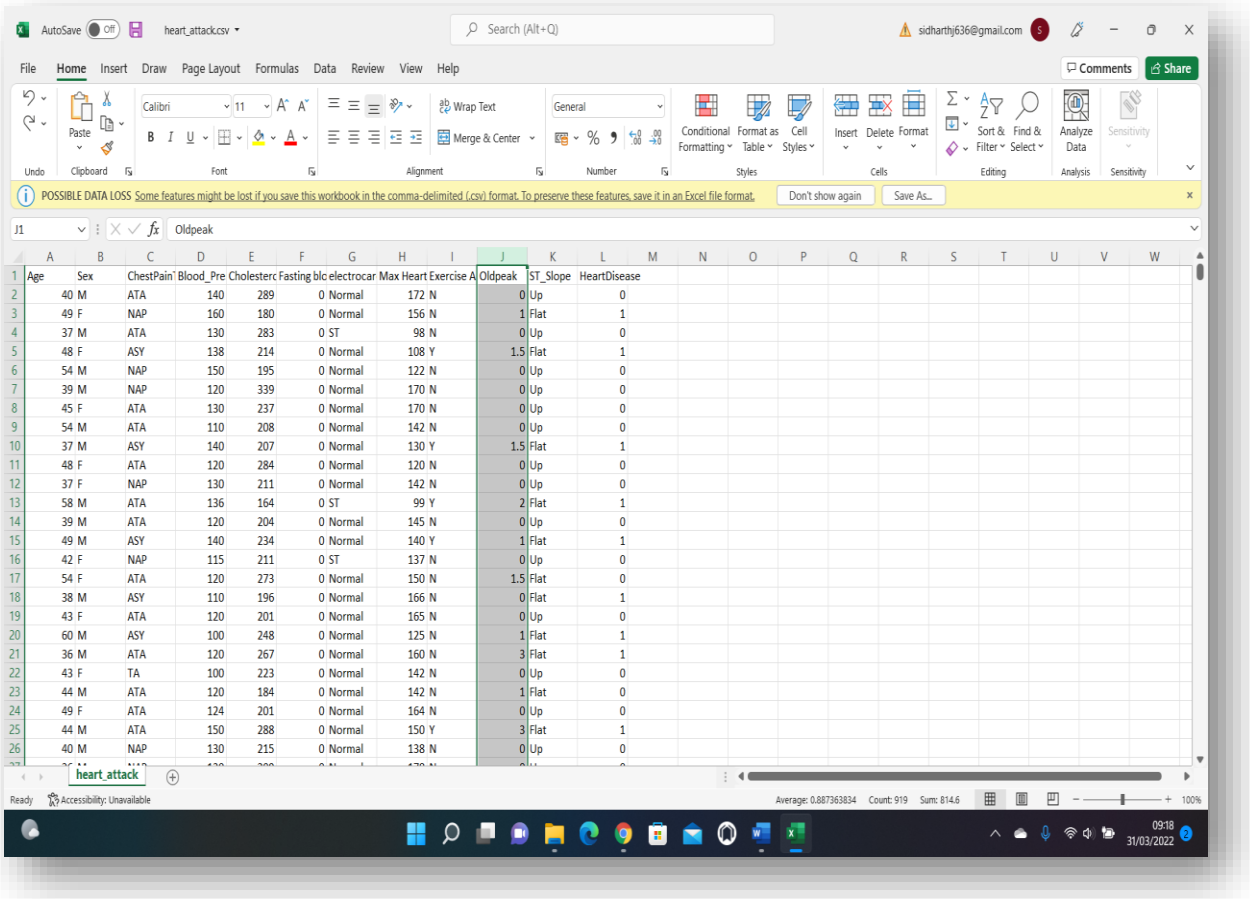
In this report , we are going to discuss about the dataset we have downloaded from this website ('https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset') where the publisher name is Rashik Rahman for and this dataset has been updated in 2021 . We are going to do analysis for different genders on the basis of different factors which are discussed below.

	Features	Description
Categorical Data	electrocardiographic results LVH = 1 Normal=2 ST=3	Normal = Normal ST = having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
	Sex M =1 F=2	M= MALE F= FEMALE
	Cheat Pain : ASY = 1 ATA =2 NAP =3 TA =4	ASY : asymptomatic ATA: atypical angina NAP: non-anginal pain TA: typical angina
	Exercise Y =1 N =0	Y= YES N= NO
	ST Slope peak exercise Up =1 Flat =0 DowNO = 2	the slope of the peak exercise ST segment — 0: UP; 1: flat
Numerical Data	Age	The Age of the Person
	Blood Pressure	The persons resting blood pressure(admission to hospital.)
	Cholesterol	The persons cholesterol measured in mg/dl
	Fasting blood sugar	The person's fasting blood sugar (> 120 mg/dl, 1 = true; 0 = false)
	Max Heart Rate	Maximum Heart Achieved
	Heart Disease	Heart disease (1 = no, 0= yes)

PRE-PROCESSING :

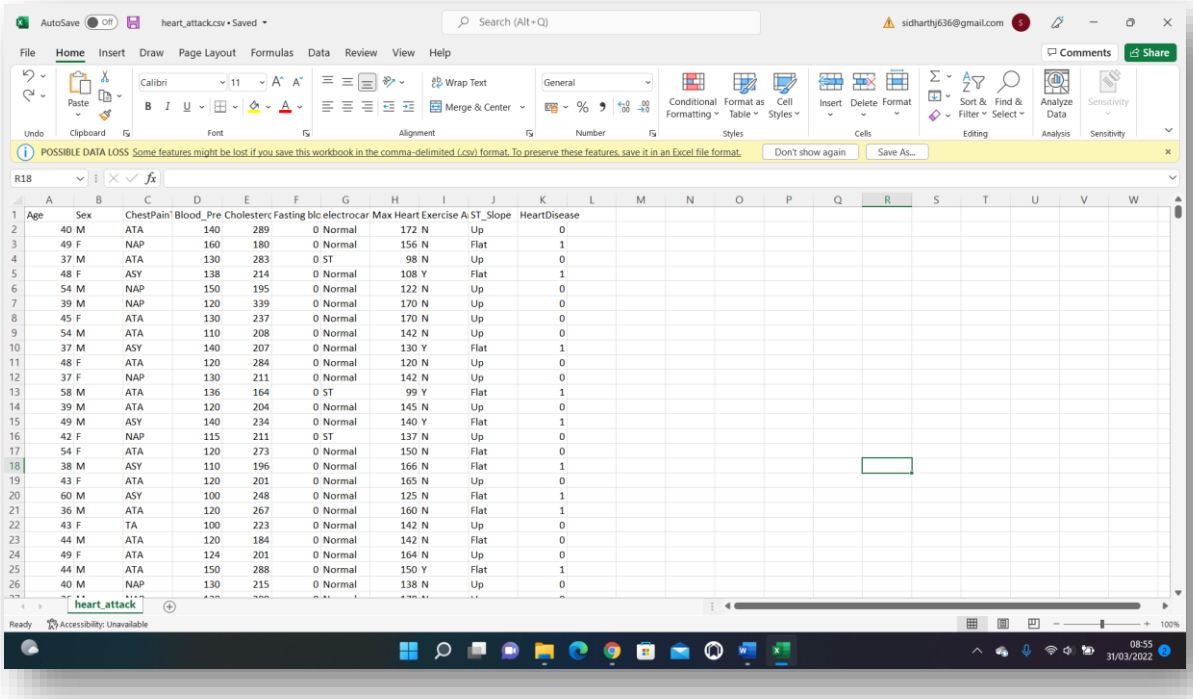
MANUALLY:

Step 1: First and Foremost, i have deleted the oldpeck column because it is not of my use anymore. As shown in image below.



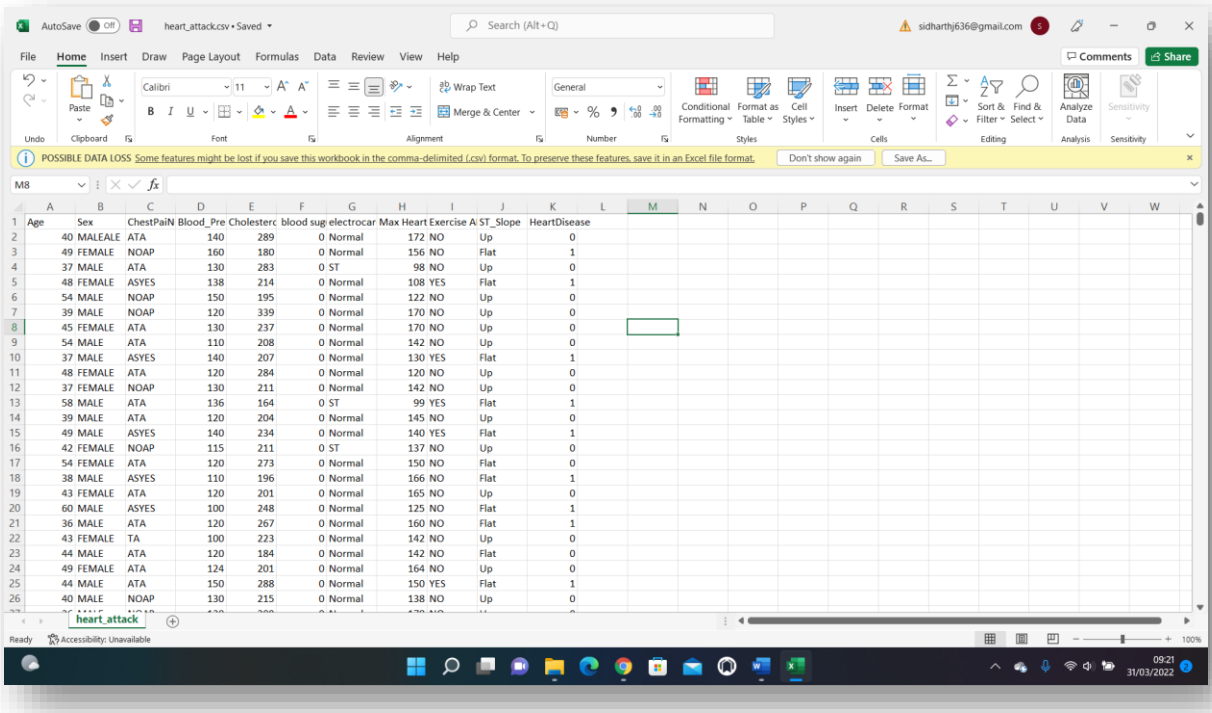
Age	Sex	ChestPain	Blood_Pre	Cholesterc	Fasting blc	electrocar	Max Heart	Exercise A	Oldpeak	ST_Slope	HeartDisease
40	M	ATA	140	289	0	Normal	172	N	0	Up	0
49	F	NAP	160	180	0	Normal	156	N	1	Flat	1
37	M	ATA	130	283	0	ST	98	N	0	Up	0
48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
54	M	NAP	150	195	0	Normal	122	N	0	Up	0
39	M	NAP	120	339	0	Normal	170	N	0	Up	0
45	F	ATA	130	237	0	Normal	170	N	0	Up	0
54	M	ATA	110	208	0	Normal	142	N	0	Up	0
37	M	ASY	140	207	0	Normal	130	Y	1.5	Flat	1
48	F	ATA	120	284	0	Normal	120	N	0	Up	0
37	F	NAP	130	211	0	Normal	142	N	0	Up	0
58	M	ATA	136	164	0	ST	99	Y	2	Flat	1
39	M	ATA	120	204	0	Normal	145	N	0	Up	0
49	M	ASY	140	234	0	Normal	140	Y	1	Flat	1
42	F	NAP	115	211	0	ST	137	N	0	Up	0
54	F	ATA	120	273	0	Normal	150	N	1.5	Flat	0
38	M	ASY	110	196	0	Normal	166	N	0	Flat	1
43	F	ATA	120	201	0	Normal	165	N	0	Up	0
60	M	ASY	100	248	0	Normal	125	N	1	Flat	1
36	M	ATA	120	267	0	Normal	160	N	3	Flat	1
43	F	TA	100	223	0	Normal	142	N	0	Up	0
44	M	ATA	120	184	0	Normal	142	N	1	Flat	0
49	F	ATA	124	201	0	Normal	164	N	0	Up	0
44	M	ATA	150	288	0	Normal	150	Y	3	Flat	1
40	M	NAP	130	215	0	Normal	138	N	0	Up	0

Step2: I have deleted all the duplicate rows from the dataset and make it normalise. Moreover, there are some values missing from the dataset I just deleted from and the dataset perfect and checked all the names and values to avoid error or consistency in the analysis.



Age	Sex	ChestPain	Blood_Pre	Cholesterc	Fasting blc	electrocar	Max Heart	Exercise A	ST_Slope	HeartDisease
40	M	ATA	140	289	0	Normal	172	N	Up	0
49	F	NAP	160	180	0	Normal	156	N	Flat	1
37	M	ATA	130	283	0	ST	98	N	Up	0
48	F	ASY	138	214	0	Normal	108	Y	Flat	1
54	M	NAP	150	195	0	Normal	122	N	Up	0
39	M	NAP	120	339	0	Normal	170	N	Up	0
45	F	ATA	130	237	0	Normal	170	N	Up	0
54	M	ATA	110	208	0	Normal	142	N	Up	0
37	M	ASY	140	207	0	Normal	130	Y	Flat	1
48	F	ATA	120	284	0	Normal	120	N	Up	0
37	F	NAP	130	211	0	Normal	142	N	Up	0
58	M	ATA	136	164	0	ST	99	Y	Flat	1
39	M	ATA	120	204	0	Normal	145	N	Up	0
49	M	ASY	140	234	0	Normal	140	Y	Flat	1
42	F	NAP	115	211	0	ST	137	N	Up	0
54	F	ATA	120	273	0	Normal	150	N	Flat	0
38	M	ASY	110	196	0	Normal	166	N	Flat	1
43	F	ATA	120	201	0	Normal	165	N	Up	0
60	M	ASY	100	248	0	Normal	125	N	Flat	1
36	M	ATA	120	267	0	Normal	160	N	Flat	1
43	F	TA	100	223	0	Normal	142	N	Up	0
44	M	ATA	120	184	0	Normal	142	N	Flat	0
49	F	ATA	124	201	0	Normal	164	N	Up	0
44	M	ATA	150	288	0	Normal	150	Y	Flat	1
40	M	NAP	130	215	0	Normal	138	N	Up	0

Step 3: I have changed the naming for the columns and the values for example M= male , F= Female , N= NO , Y = YES , NOORMAL = NORMAL , FEMAALTE= FLAT and so on and the perfect dataset in shown below.



Age	Sex	ChestPaiN	Blood_Pre	Cholesterc	blood sug	electrocar	Max Heart	Exercise A	ST_Slope	HeartDisease
40	MALEALE	ATA	140	289	0	Normal	172	NO	Up	0
49	FEMALE	NOAP	160	180	0	Normal	156	NO	Flat	1
37	MALE	ATA	130	283	0	ST	98	NO	Up	0
48	FEMALE	ASYES	138	214	0	Normal	108	YES	Flat	1
54	MALE	NOAP	150	195	0	Normal	122	NO	Up	0
39	MALE	NOAP	120	339	0	Normal	170	NO	Up	0
45	FEMALE	ATA	130	237	0	Normal	170	NO	Up	0
54	MALE	ATA	110	208	0	Normal	142	NO	Up	0
37	MALE	ASYES	140	207	0	Normal	130	YES	Flat	1
48	FEMALE	ATA	120	284	0	Normal	120	NO	Up	0
37	FEMALE	NOAP	130	211	0	Normal	142	NO	Up	0
58	MALE	ATA	136	164	0	ST	99	YES	Flat	1
39	MALE	ATA	120	204	0	Normal	145	NO	Up	0
49	MALE	ASYES	140	234	0	Normal	140	YES	Flat	1
42	FEMALE	NOAP	115	211	0	ST	137	NO	Up	0
54	FEMALE	ATA	120	273	0	Normal	150	NO	Flat	0
38	MALE	ASYES	110	196	0	Normal	166	NO	Flat	1
43	FEMALE	ATA	120	201	0	Normal	165	NO	Up	0
60	MALE	ASYES	100	248	0	Normal	125	NO	Flat	1
36	MALE	ATA	120	267	0	Normal	160	NO	Flat	1
43	FEMALE	TA	100	223	0	Normal	142	NO	Up	0
44	MALE	ATA	120	184	0	Normal	142	NO	Flat	0
49	FEMALE	ATA	124	201	0	Normal	164	NO	Up	0
44	MALE	ATA	150	288	0	Normal	150	YES	Flat	1
40	MALE	NOAP	130	215	0	Normal	138	NO	Up	0

USING PYTHON COMMANDS:

Step 1: if there is no Null Values in the data .

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 918 entries, 0 to 917
Data columns (total 9 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Age                             918 non-null    int64
 1   Sex                             918 non-null    int64
 2   ChestPain                       918 non-null    int64
 3   Blood_Pressure                  918 non-null    int64
 4   Cholesterol                     918 non-null    int64
 5   electrocardiographic results    918 non-null    int64
 6   Max Heart Rate                  918 non-null    int64
 7   ST_Slope                        918 non-null    int64
 8   HeartDisease                    918 non-null    int64
dtypes: int64(9)
memory usage: 64.7 KB
```

Step2: Describe the dataset with all the mean values and std etc.

	Age	Sex	ChestPain	Blood_Pressure	Cholesterol	electrocardiographic results	Max Heart Rate	ST_Slope	HeartDisease
count	918.000000	918.000000	918.000000	918.000000	918.000000	918.000000	918.000000	918.000000	918.000000
mean	53.510893	0.789760	1.781046	132.396514	198.799564	1.989107	136.809368	0.567538	0.553377
std	9.432617	0.407701	0.956519	18.514154	109.384145	0.631671	25.460334	0.618959	0.497414
min	28.000000	0.000000	1.000000	0.000000	0.000000	1.000000	60.000000	0.000000	0.000000
25%	47.000000	1.000000	1.000000	120.000000	173.250000	2.000000	120.000000	0.000000	0.000000
50%	54.000000	1.000000	1.000000	130.000000	223.000000	2.000000	138.000000	0.000000	1.000000
75%	60.000000	1.000000	3.000000	140.000000	267.000000	2.000000	156.000000	1.000000	1.000000
max	77.000000	1.000000	4.000000	200.000000	603.000000	3.000000	202.000000	2.000000	1.000000

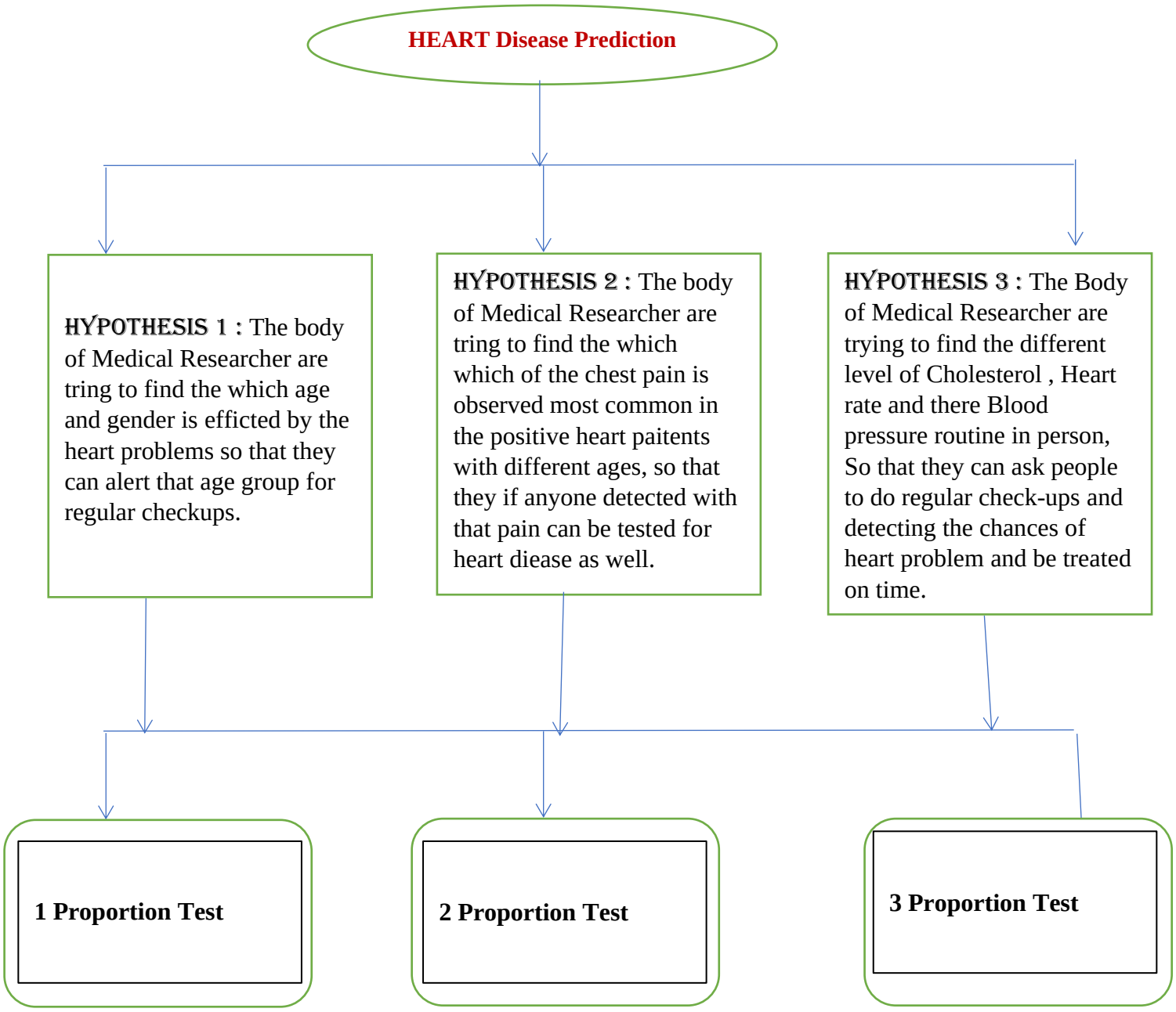
Step 3: Dividing data into categorical and numerical list.

```
catagorocal data

['Sex', 'ChestPain', 'electrocardiographic results', 'ST_Slope', 'HeartDisease']
numerical data


['Age', 'Blood_Pressure', 'Cholesterol', 'Max Heart Rate']
```

HYPOTHESIS :


HEART Disease Prediction
HYPOTHESIS 1 : The body of Medical Researcher are tring to find the which age and gender is efficted by the heart problems so that they can alert that age group for regular checkups.
HYPOTHESIS 2 : The body of Medical Researcher are tring to find the which which of the chest pain is observed most common in the positive heart paitents with different ages, so that they if anyone detected with that pain can be tested for heart diease as well.
HYPOTHESIS 3 : The Body of Medical Researcher are trying to find the different level of Cholesterol , Heart rate and there Blood pressure routine in person, So that they can ask people to do regular check-ups and detecting the chances of heart problem and be treated on time.
1 Proportion Test
2 Proportion Test
3 Proportion Test

PROPOSED SOLUTION:

To find the solution of this problems proposed in hypothesis can be achieved by **KNN** algorithm. It is used to solve the classification and regression problems. It find the nearest neighbour in certain area around the point for prediction and check which attribute is closest to that so that it can predict it. Which comes under supervised learning technique.

The diagram below describe how does it work.

Step 1: load the dataset

Step 2: Initialize K to your chosen number however by default if k=5.

Step 3: Calculate the distance between the actual point and the predicted point.

Step 4: Add the distance and the index of the data point to the ordered list.

Step 5: pick the first k values from the list and find the least distance from the predicted point.

Step 6: get the label of the nearest K value.

Hypothesis 1: we have the results in zero and one which represents yes or no if the person has heart disease then it will be one if he doesn't have heart disease then it will be 0 and we have some values like age and gender through which we can differentiate how many males and females of which each category are affected most by the heart diseases. So that if you are giving any age and gender it can calculate the nearest neighbour and can predict whether they have heart disease or not.

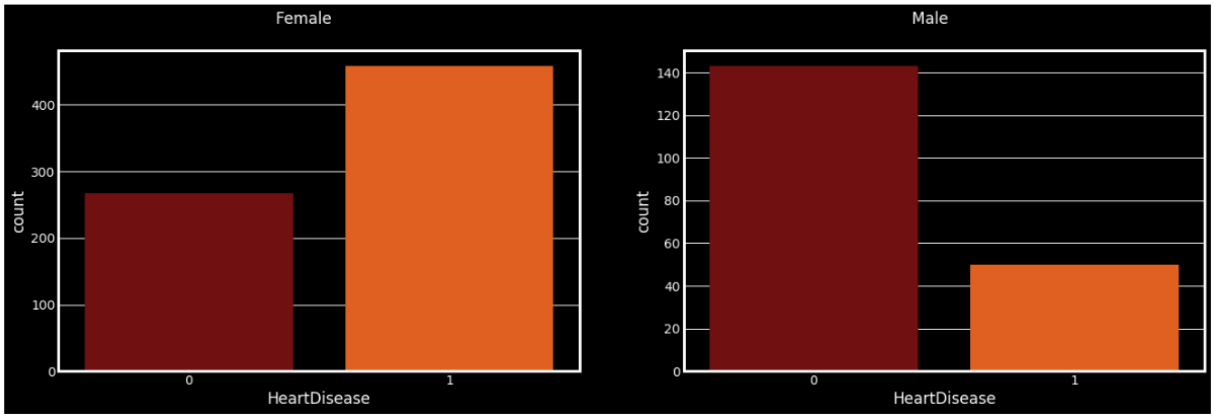
Step 1: calculate the number ages in male= 1 and female =0 with the target value 0=negavtive and 1= positive .

```
Percent of '1' at high risk of heart attack = 63.17 %
 '1' average high-risk age = 56 yrs

Percent of '0' at high risk of heart attack = 25.91 %
'0' Average high-risk age = 56 yrs
```

Step 2: the number of positive heart problem and people with no heart problem in different genders where 1 = positive and 0= negative.

```
Female Value Counts:
1    458
0    267
Name: HeartDisease, dtype: int64
Male Value Counts:
0    143
1     50
Name: HeartDisease, dtype: int64
```



Step3: Training the dataset and checking Acuracy

```
0.625
```

Step 4: Giving raw data and getting the output.

	Age	Sex
0	15	1

```
No Disease , You can healty But have regular check-ups
```

Hypothesis 2: It is going to check from the four different chest pains which is the most common and highly affected with heart disease so that if you're giving and he does mean it can identify by calculating the nearest neighbour and can predict changes of the person is affected with heart disease or not.

Step 1: Finding all the different Chest Pains using head which gives top 5 of all the different number of chest pains.

	index	Age	Sex	ChestPain	Blood_Pressure	Cholesterol	electrocardiographic results	Max Heart Rate	ST_Slope	HeartDisease
0	3	48	0	1	138	214	2	108	0	1
1	8	37	1	1	140	207	2	130	0	1
2	13	49	1	1	140	234	2	140	0	1
3	16	38	1	1	110	196	2	166	0	1
4	18	60	1	1	100	248	2	125	0	1

	index	Age	Sex	ChestPain	Blood_Pressure	Cholesterol	electrocardiographic results	Max Heart Rate	ST_Slope	HeartDisease
0	0	40	1	2	140	289	2	172	1	0
1	2	37	1	2	130	283	3	98	1	0
2	6	45	0	2	130	237	2	170	1	0
3	7	54	1	2	110	208	2	142	1	0
4	9	48	0	2	120	284	2	120	1	0

	index	Age	Sex	ChestPain	Blood_Pressure	Cholesterol	electrocardiographic results	Max Heart Rate	ST_Slope	HeartDisease
0	1	49	0	3	160	180	2	156	0	1
1	4	54	1	3	150	195	2	122	1	0
2	5	39	1	3	120	339	2	170	1	0
3	10	37	0	3	130	211	2	142	1	0
4	14	42	0	3	115	211	3	137	1	0

	index	Age	Sex	ChestPain	Blood_Pressure	Cholesterol	electrocardiographic results	Max Heart Rate	ST_Slope	HeartDisease
0	20	43	0	4	100	223	2	142	1	0
1	88	43	1	4	120	291	3	155	0	1
2	118	35	0	4	120	160	3	185	1	0
3	119	34	1	4	140	156	2	180	0	1
4	165	46	1	4	140	272	2	175	0	1

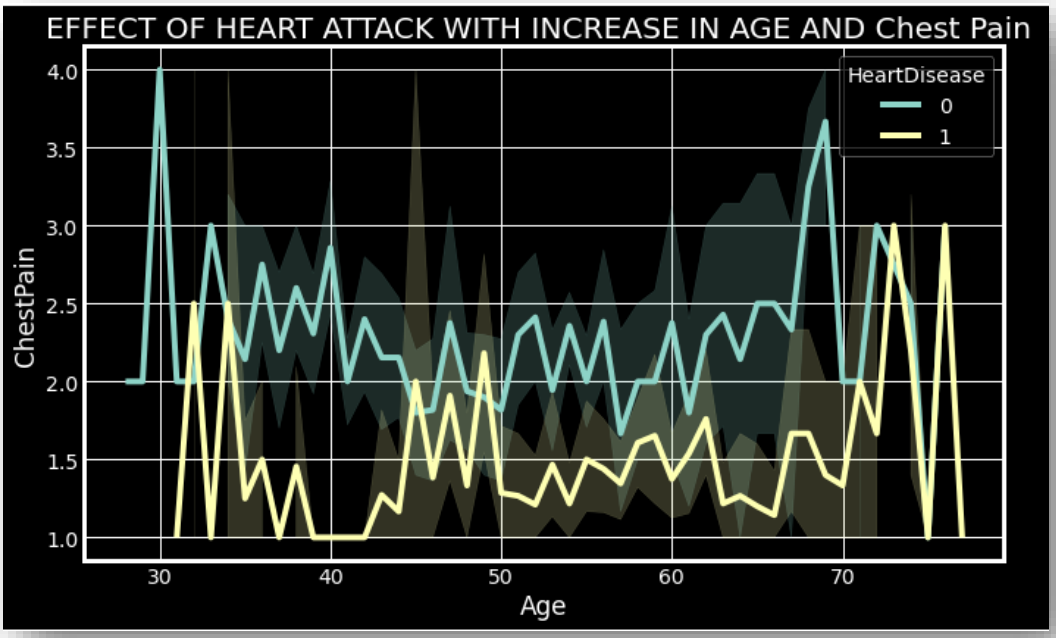
Step 2: Finding the percentage of different ages affected with the different kind of chest pain and highlighting the most affected age.

```
Percent of '1' at high risk of heart attack = 42.7 %
Asymptomatic average high-risk age = 56 yrs

Percent of '2' at high risk of heart attack = 2.61 %
Atypical-Angina Average high-risk age = 56 yrs
Percent of '3' at high risk of heart attack = 7.84 %
 Non-Anginal Pain average high-risk age = 57 yrs

Percent of '4' at high risk of heart attack = 2.18 %
Typical Angina Average high-risk age = 55 yrs
```

Step 3: Plotting the chest pain among different ages with regard to target value of o and 1 which is positive and negative heart problems.



Step3: Training the dataset and checking Acuracy

```
0.7554347826086957
```

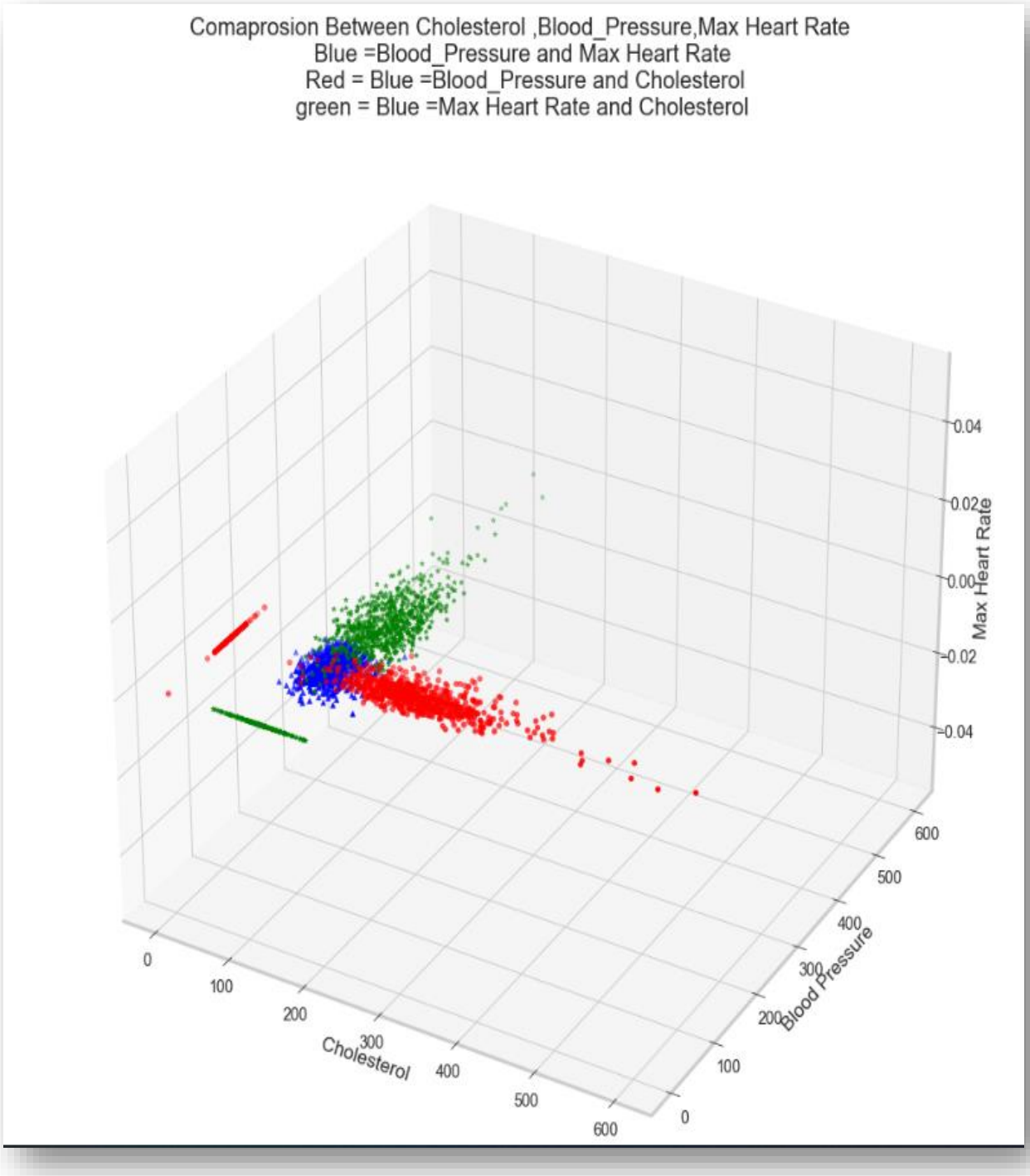
Step 4: Giving raw data and getting the output.

	Age	Chest Pain
0	32	1

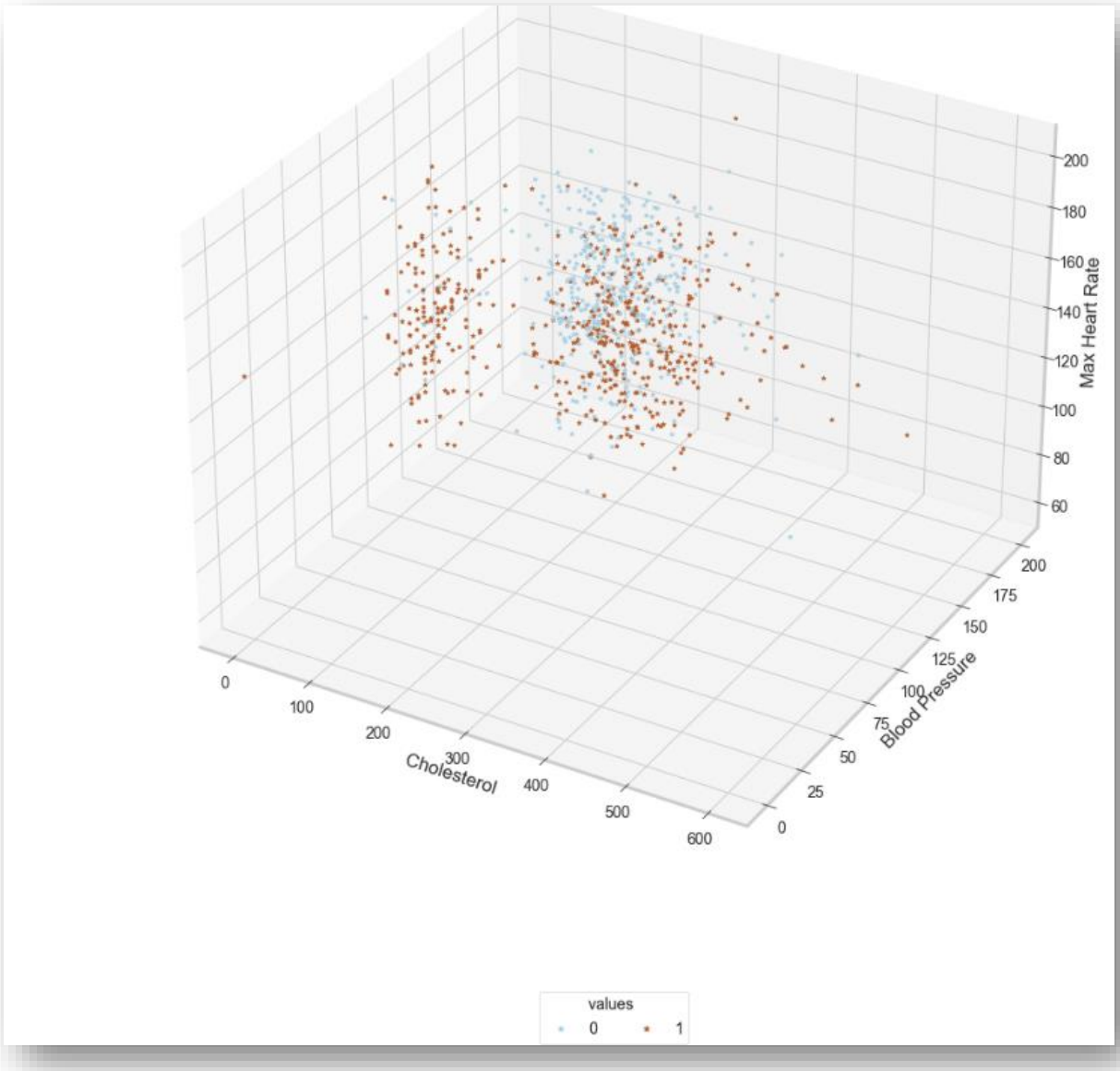
```
***Warning***
There is chance of Heart Disease Contact Your Doctor
```

Hypothesis 3: we are going to take people who are doing exercise and people who are not doing exercise and what is their cholesterol an heart rate level moreover they are affected with the heart disease or not so that if you read any patients report which is near to that values which are positive heart disease then we can give them a better treatment on time.

Step 1: Finding the relation between different categories through graph as shown below



Step 2 : Now Plotting relation between cholesterol , blood pressure and max heart rate with Target value 0 and 1 where 0 is no heart problem and 1 means have heart problems.



Step 3: Training the dataset and checking Acuracy

```
0.7119565217391305
```

Step 4: Giving raw data and getting the output.

	Cholesterol	Blood_Pressure	Max Heart Rate
0	1	175	0

```
No Disease , You can healty But have regular check-ups
```

Reflection :

As I coursework two I have made some assumption with hypothesis 3 I have to change it because if I was using chest pain then it was not producing the appropriate output .However , I have to change it to max_heart_rate and compare values with target value 0 and 1.

In the second hypothesis I have added age through which I can compare chest pain with the target values which was missing in coursework2.

However, apart from this i have achieved all the requirement which were stated in coursework 2 and proper implementation.

References :

[1]https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761

[2] https://www.alliedacademies.org/articles/prediction-of-heart-disease-using-knearest-neighbor-and-particle-swarm-optimization.html#:~:text=Supervised%20algorithms%20are%20used%20for,a%20large%20number%20of%20features.

[3]http://rstudio-pubstatic.s3.amazonaws.com/318411_18399592759841f2a151e445adb851c7.html

[4] https://www.youtube.com/watch?v=_xfCq9mxrwM&ab_channel=CodewithMarcus