



江南大学  
JIANGNAN UNIVERSITY

***Machine Learning Final Project Report***

<b>Student ID:</b>	<b>5035190144</b>
<b>Student Name:</b>	<b>Abdulkadir Duran Adan</b>
<b>Major:</b>	<b>Computer science and technology</b>
<b>Class</b>	<b>BCS1901</b>

# *Heart disease prediction Machine Learning Web App*

## ***Abstract***

In this paper/report, I will train a model for the task of heart disease prediction using Machine Learning. I will use the Logistic Regression algorithm in machine learning to train a model to predict heart disease.

Predicting and diagnosing heart disease is the biggest challenge in the medical industry and relies on factors such as the physical examination, symptoms and signs of the patient.

Due to the increasing use of advanced technology and data collection, I have created this system to predict heart disease using machine learning algorithms. Heart disease is leading cause of death in many parts of the world. In particular, in this type of disease, the heart is cannot pump the required amount of blood to the other parts of the human body to perform proper blood circulation.

**Keywords.** Logistic regression, predict, heart disease

## ***1. Introduction***

Heart Disease (including Coronary Heart Disease, Hypertension, and Stroke) remains the No. 1 cause of death in the many parts of the world. The Heart Disease and Stroke Statistics—2019 Update from the American Heart Association indicates that:

- ⇒ 116.4 million, or 46% of US adults are estimated to have hypertension. These are findings related to the new 2017 Hypertension Clinical Practice Guidelines.
- ⇒ On average, someone dies of CVD every 38 seconds. About 2,303 deaths from CVD each day, based on 2016 data.

⇒ On average, someone dies of a stroke every 3.70 minutes. About 389.4 deaths from stroke each day, based on 2016 data.

To solve this problem using advanced technological solutions & Machine learning, I have come up with a Heart Disease Prediction System using the Logistic Regression Machine Learning algorithm. This system takes the data entered by the user and then it uses a statistical approach by employing probabilistic & optimization techniques to draw out a result based on past datasets. This evaluation technique aims at helping doctors & users to detect heart disease so that it can be prevented & cured, thereby saving many lives.

### ***Advantages***

- ❖ User can easily get the heart disease prediction on single click The system is free and available at any time
- ❖ System is kept online in order to serve people 24x7
- ❖ The Heart Disease Prediction web application is an end user support and online consultation project
- ❖ The application is fed with various details and the Heart disease associated with those details and it allows user to share their heart related issues. It then processes user specific details to check for various cancer disease that could be associated with the inputs received from user.
- ❖ Accurate prediction

## ***2. Algorithm***

### ***Logistic Regression Machine Learning algorithm***

*In this project, I used Logistic Regression Machine Learning algorithm to train the model for the task of Heart Disease Prediction using python language.*

Logistic Regression is a statistical and machine-learning technique classifying records of a dataset based on the values of the input fields. It predicts a dependent variable based on one or more set of independent variables to predict outcomes. It can be used both for binary classification and multi-class classification and. I used *Logistic Regression for this model* to give the most accurate heart disease prediction.

Logistic regression is most frequently applied for classification, primarily two-category issues (that is, there are only two types of output, each representing one category), and can indicate the probability of occurrence of each classification event.

Logistic regression model is shown below:

$$\text{Prob}(Y = 1) = \frac{e^z}{1 + e^z}$$

Where Y refers to binary dependent variable (Y is equal to 1 if event happens; Y=0 otherwise),

e stands for the foundation of natural logarithms and Z means  $Z = \beta_0 + \beta_1 x_1 + \beta_2 + \dots + \beta_p p$

with constant  $\beta_0$ , coefficients  $\beta_j$  and predictors  $X_j$  for p predictors ( $j=1,2,3,\dots,p$ )

### 3. Experiments.

#### 3.1 Dataset description

This dataset consists of 13 features and a target variable. It has 8 nominal variables and 6 numeric variables. This dataset is available on Kaggle and here is the link for the dataset: [Heart](#)

The detailed description of all the features are as follows:

no	feature	description	Type
1	Age	Age in years	Numeric
2	Sex	Sex (1 = male; 0 = female)	Nominal
3	CP	chest pain type  Value 1: typical angina -- Value 2: atypical angina -- Value 3: non-anginal pain -- Value 4: asymptomatic	Nominal
4	trestbps	resting blood pressure anything above 130-140 is generally of concern	Numeric
5	chol	Serum cholestrol in mg/dl (Numeric)	Numeric
6	fbs	Fasting blood sugar > 120 mg/dl	Nominal
7	restecg	Result of electrocardiogram while at rest are represented in 3 distinct values 0 : Normal 1: Abnormality in ST-T wave 2: Left ventricular hypertrophy	nominal
8	thalach	Maximum heart rate achieved	Numeric

9	exang	Exercise induced angina	nominal
10	oldpeak	exercise-induced ST depression vs. rest looks at heart stress during exercise an unhealthy heart will stress more.	Numeric
11	slope	The slope of the peak exercise ST segment Nominal - - Value 1: upsloping -- Value 2: flat -- Value 3: downsloping	nominal
12	ca	number of major vessels (0-3) stained by fluoroscopy: the more blood movement the better, so people with ca equal to 0 are more likely to have heart disease.	Numeric
13	thal	thallium stress result: People with a thal value of 2 (defect corrected: once was a defect but ok now) are more likely to have heart disease.	nominal
14	target	It is the target variable which we have to predict 1 means patient is suffering from heart risk and 0 means patient is normal.	nominal

## 3.2 implementation

### 3.2.1 displaying sample entries of dataset so that we can see what we are working on

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.30	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.50	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.40	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.80	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.60	2	0	2	1

### 3.2.2 Exploratory Data Analysis (EDA)

Before training the logistic regression we need to observe and analyze the data to see what we are going to work with. The goal here is to learn more about the data and become a topic expert on the dataset you are working with.

EDA helps us find answers to some important questions such as: What question (s) are you trying to solve? What kind of data do we have and how do we handle the different types? What is missing in the data and how do you deal with it? Where are the outliers and why should you care? How can you add, change, or remove features to get the most out of your data?

### 3.2.2.1 Checking missing entries in the dataset columnwise

```
age      0
sex      0
cp       0
trestbps 0
chol     0
fbs      0
restecg  0
thalach  0
exang    0
oldpeak  0
slope    0
ca       0
thal     0
target   0
dtype: int64
```

there are no missing entries in the dataset and that is perfect

```
age : [63 37 41 56 57 44 52 54 48 49 64 58 50 66 43 69 59 42 61 40 71 51 65 53
46 45 39 47 62 34 35 29 55 60 67 68 74 76 70 38 77]

sex : [1 0]

cp : [3 2 1 0]

trestbps : [145 130 120 140 172 150 110 135 160 105 125 142 155 104 138 128 108 134
122 115 118 100 124 94 112 102 152 101 132 148 178 129 180 136 126 106
156 170 146 117 200 165 174 192 144 123 154 114 164]

chol : [233 250 204 236 354 192 294 263 199 168 239 275 266 211 283 219 340 226
247 234 243 302 212 175 417 197 198 177 273 213 304 232 269 360 308 245
208 264 321 325 235 257 216 256 231 141 252 201 222 260 182 303 265 309
186 203 183 220 209 258 227 261 221 205 240 318 298 564 277 214 248 255
207 223 288 160 394 315 246 244 270 195 196 254 126 313 262 215 193 271
268 267 210 295 306 178 242 180 228 149 278 253 342 157 286 229 284 224
206 167 230 335 276 353 225 330 290 172 305 188 282 185 326 274 164 307
249 341 407 217 174 281 289 322 299 300 293 184 409 259 200 327 237 218
319 166 311 169 187 176 241 131]

fbs : [1 0]

restecg : [0 1 2]

thalach : [150 187 172 178 163 148 153 173 162 174 160 139 171 144 158 114 151 161
179 137 157 123 152 168 140 188 125 170 165 142 180 143 182 156 115 149
146 175 186 185 159 130 190 132 147 154 202 166 164 184 122 169 138 111
145 194 131 133 155 167 192 121 96 126 105 181 116 108 129 120 112 128
109 113 99 177 141 136 97 127 103 124 88 195 106 95 117 71 118 134
90]

exang : [0 1]

oldpeak : [2.3 3.5 1.4 0.8 0.6 0.4 1.3 0. 0.5 1.6 1.2 0.2 1.8 1. 2.6 1.5 3. 2.4
0.1 1.9 4.2 1.1 2. 0.7 0.3 0.9 3.6 3.1 3.2 2.5 2.2 2.8 3.4 6.2 4. 5.6
2.9 2.1 3.8 4.4]

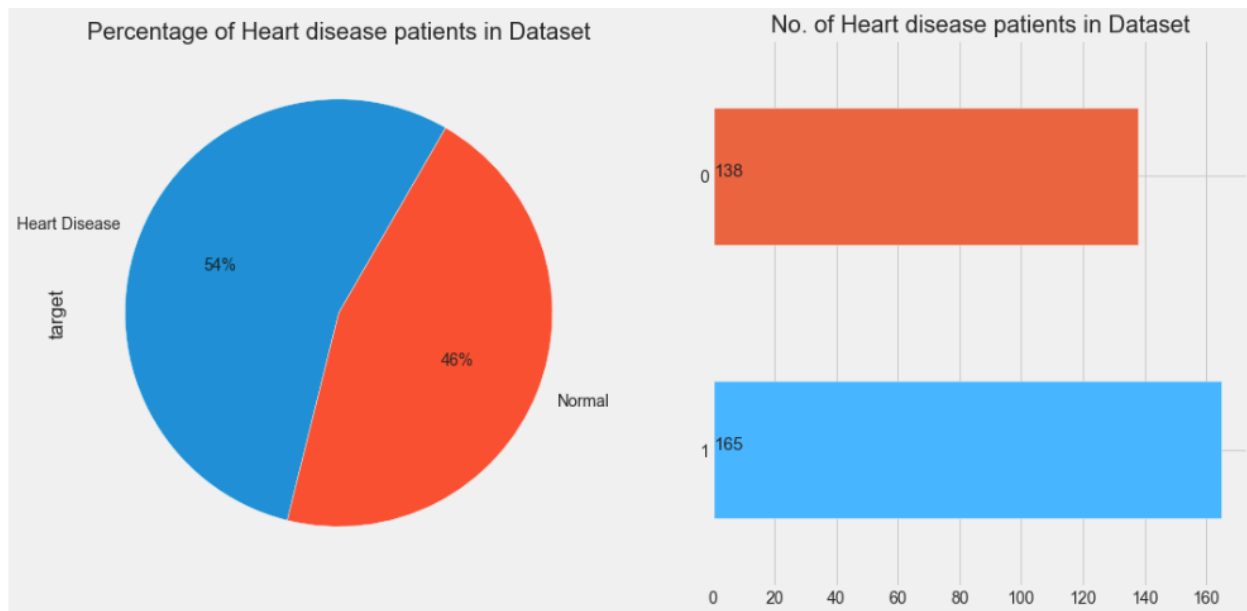
slope : [0 2 1]

ca : [0 2 1 3 4]

thal : [1 2 3 0]

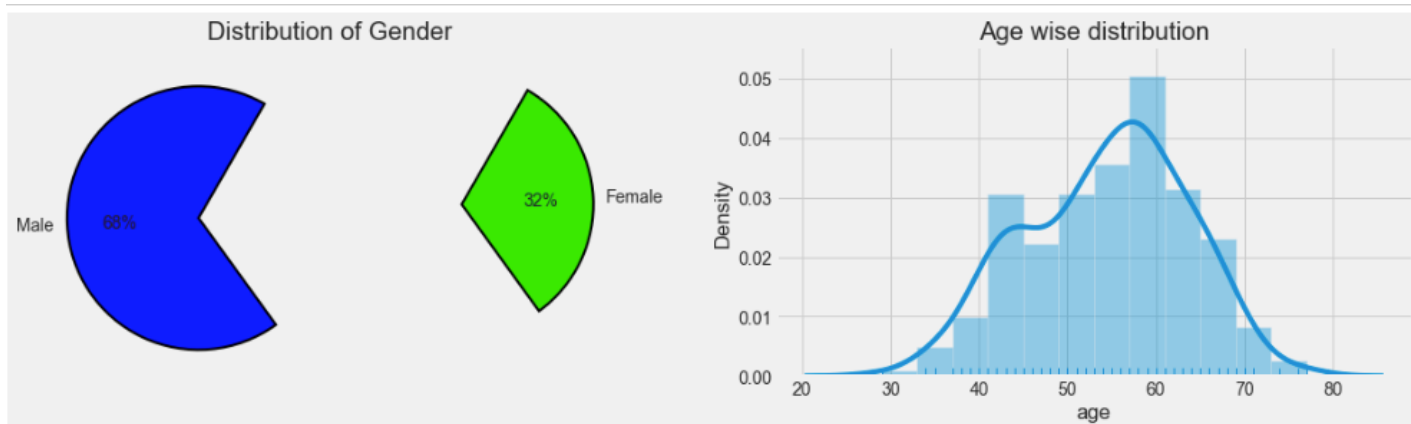
target : [1 0]
```

### 3.2.2.2 Distribution of Heart disease (target variable)



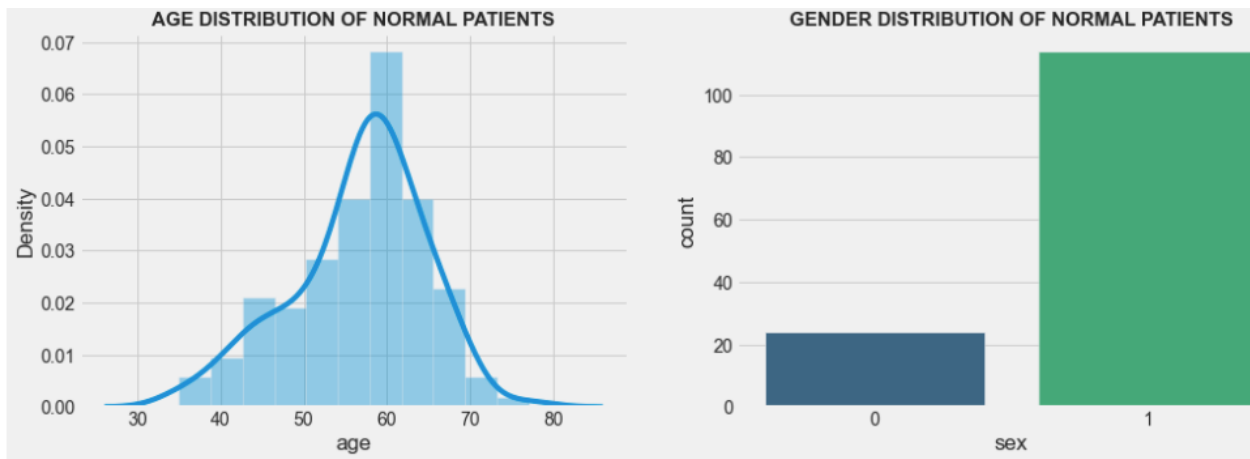
We have 165 people with heart disease and 138 people without heart disease, so our problem is balanced.

### 3.2.2.3 Checking Gender & Age wise Distribution (target variable)

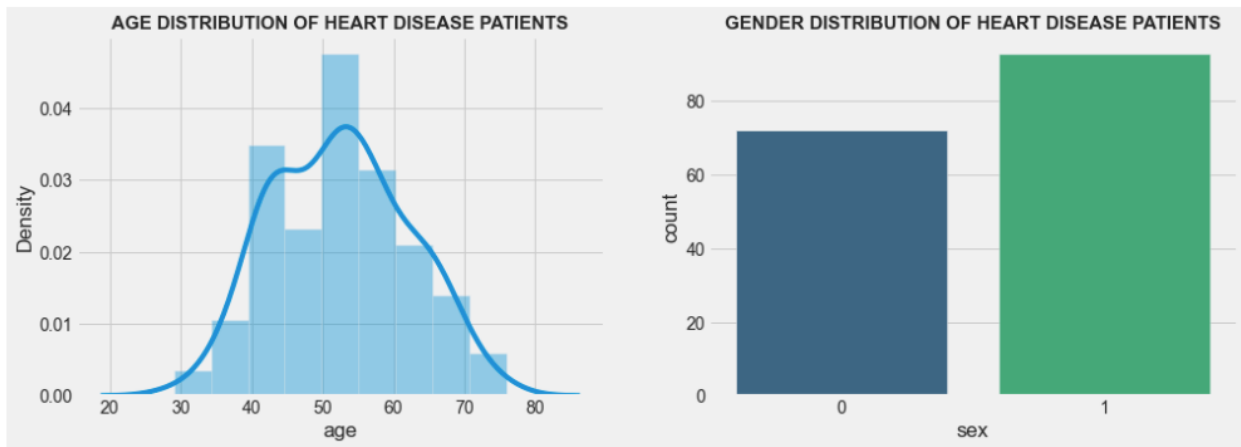


As we can see from above plot, in this dataset males percentage is way too higher than females where as average age of patients is around 55.

### 3.2.2.4 AGE DISTRIBUTION OF NORMAL PATIENTS

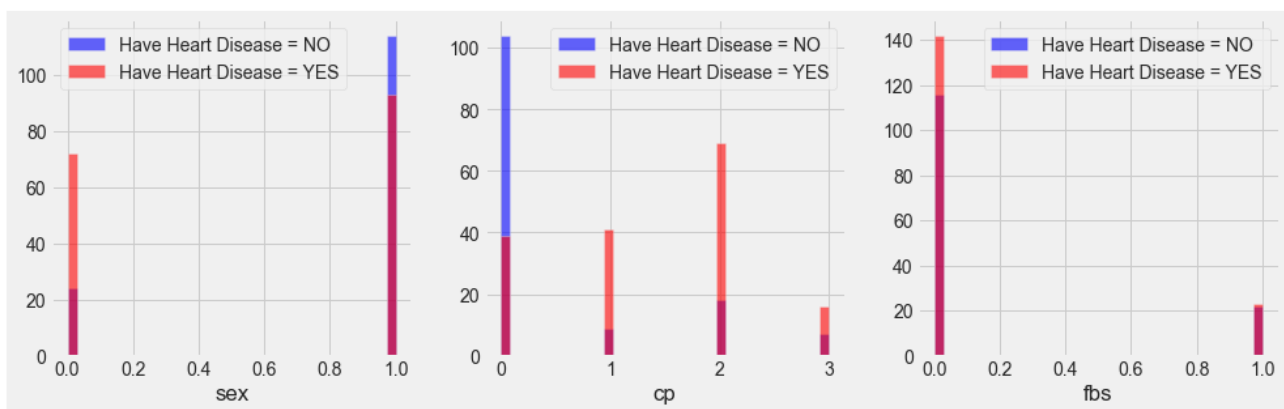


### 3.2.2.5 GENDER DISTRIBUTION OF HEART DISEASE PATIENTS

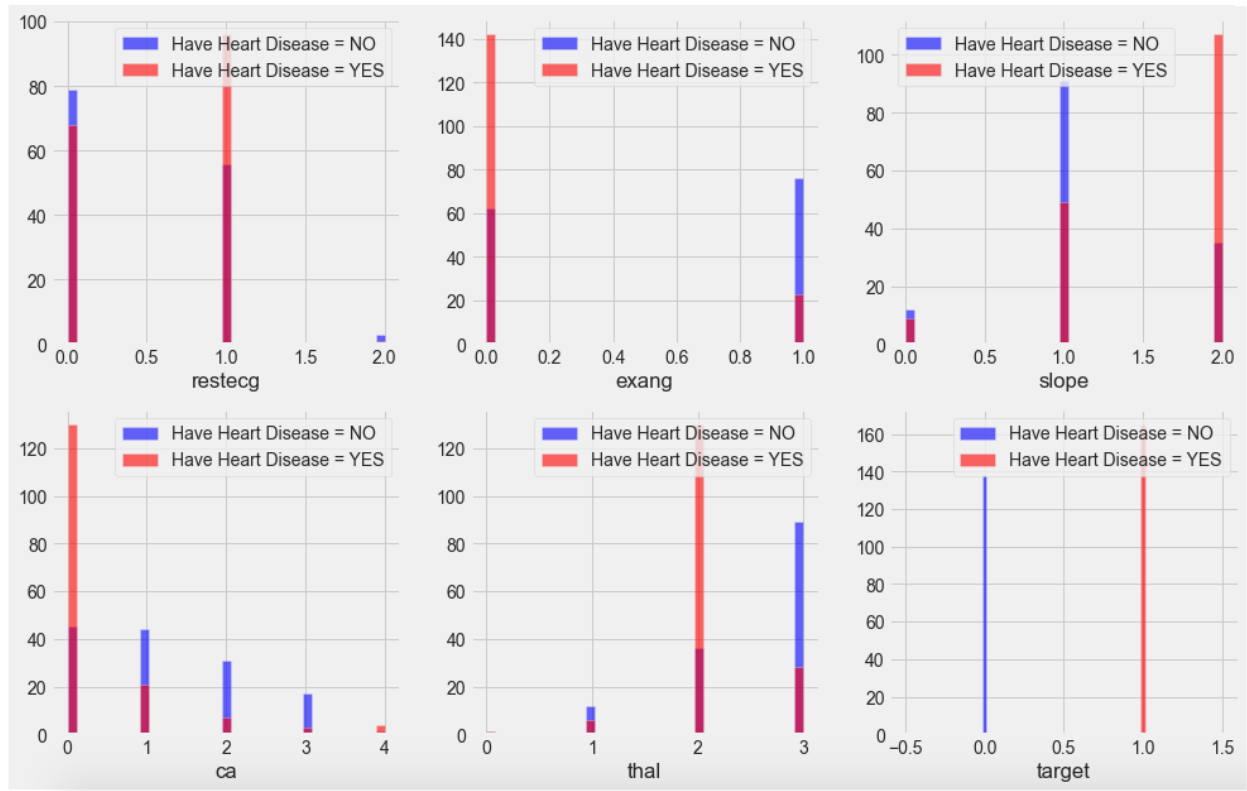


As we can see from above plot more patients accounts for heart disease in comparison to females whereas mean age for heart disease patients is around 58 to 60 years.

### 3.2.2.6 Distribution of all features such as sex, Chest Pain Type, Distribution of Rest ECG of patients



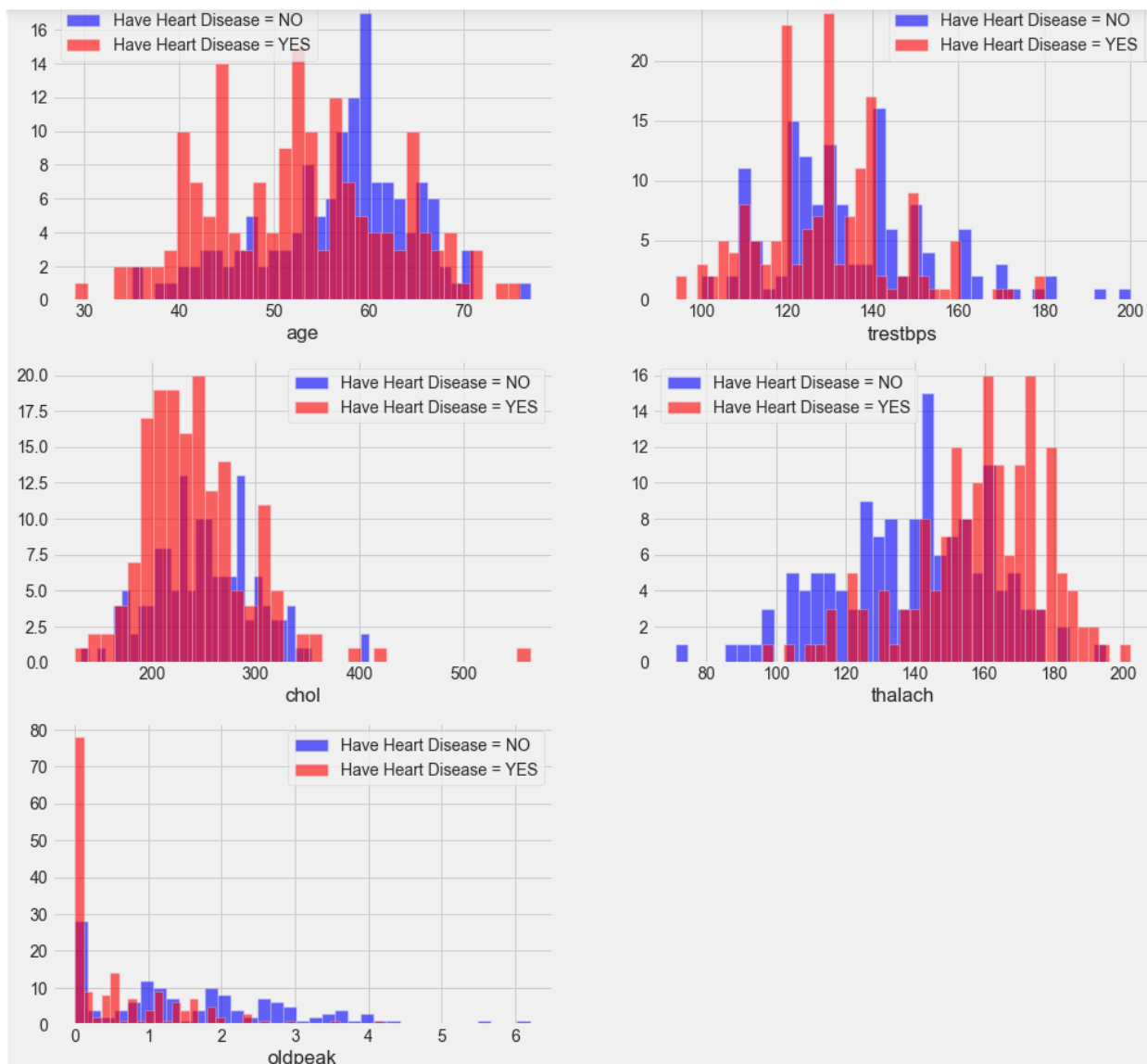




### Observations from the above plot:

1. cp {Chest pain}: People with cp 1, 2, 3 are more likely to have heart disease than people with cp 0.
2. restecg {resting EKG results}: People with a value of 1 (reporting an abnormal heart rhythm, which can range from mild symptoms to severe problems) are more likely to have heart disease.
3. exang {exercise-induced angina}: people with a value of 0 (No ==> angina induced by exercise) have more heart disease than people with a value of 1 (Yes ==> angina induced by exercise)
4. slope {the slope of the ST segment of peak exercise}: People with a slope value of 2 (Downsloping: signs of an unhealthy heart) are more likely to have heart disease than people with a slope value of 0 (Upsloping: best heart rate with exercise) or 1 (Flatsloping: minimal change (typical healthy heart)).
5. ca {number of major vessels (0-3) stained by fluoroscopy}: the more blood movement the better, so people with ca equal to 0 are more likely to have heart disease.

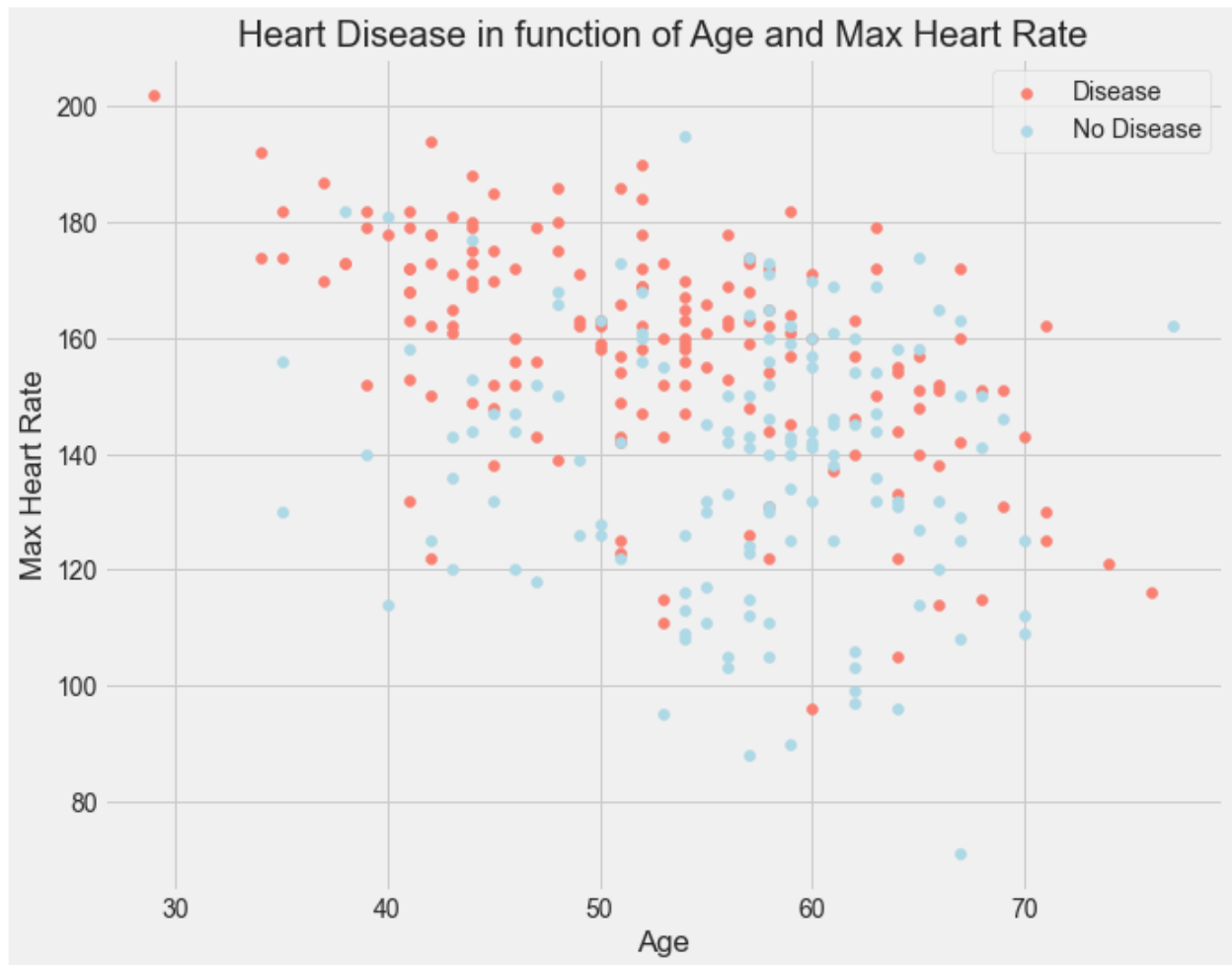
6. `thal` {thallium stress result}: People with a `thal` value of 2 (defect corrected: once was a defect but ok now) are more likely to have heart disease.



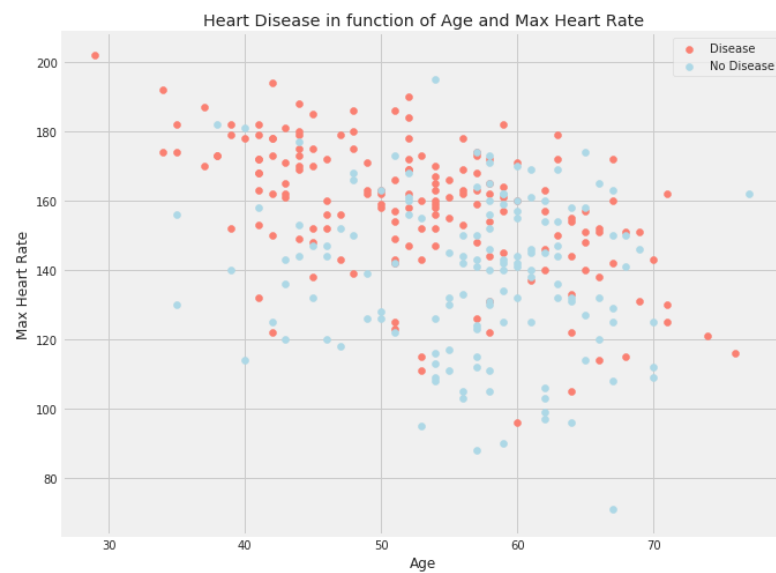
### Observations from the above plot:

1. `trestbps`: resting blood pressure anything above 130-140 is generally of concern
2. `chol`: greater than 200 is of concern.
3. `thalach`: People with a maximum of over 140 are more likely to have heart disease.
4. the old peak of exercise-induced ST depression vs. rest looks at heart stress during exercise an unhealthy heart will stress more.

### 3.2.2.7 Heart Disease in function of Age and Maximum Heart Rate achieved



### 3.2.2.8 Heart Disease in function of Age and Maximum Heart Rate achieved

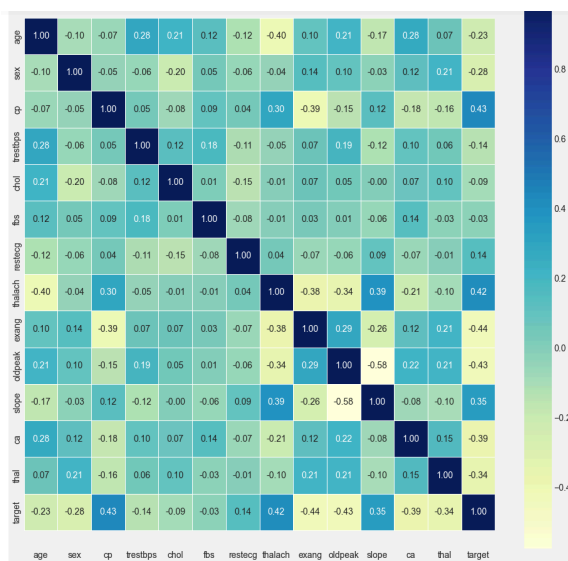


### 3.2.2.9 Distribution of Numerical features



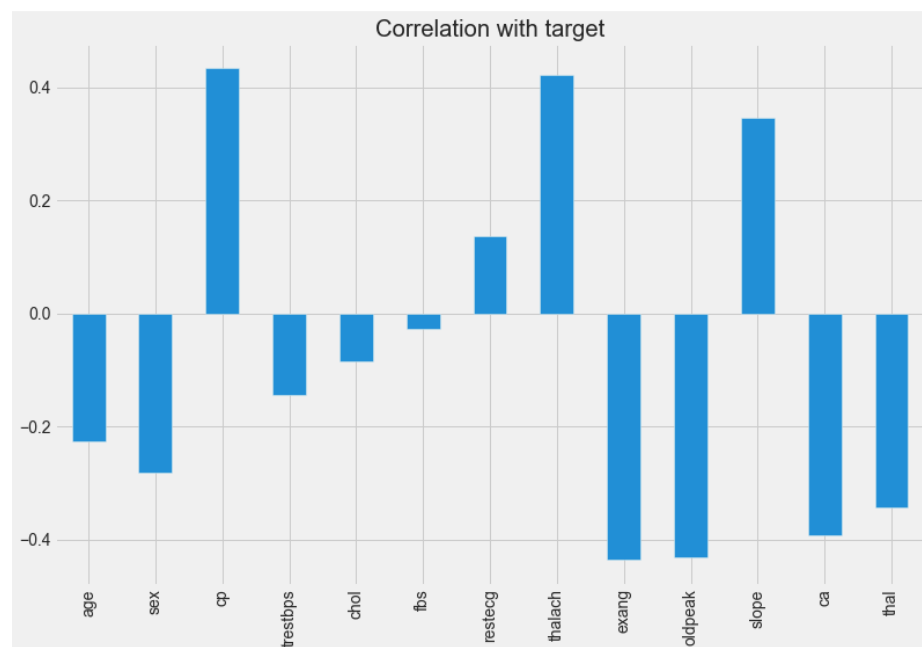
From the above plot it is clear that as the age increases chances of heart disease increases

### 3.2.2.9 Correlation Matrix



Correlation between each characteristic value According to the correlation between the eigenvalues combination with the most significant features was selected for data analysis, and different data mining techniques were used to test the selected combination.

### **Correlation with target**



### **Observations from correlation:**

1. fbs and chol are the least correlated with the target variable.
2. All other variables have a significant correlation with the target variable.

## **3.2.3 Train Test Split and Data Processing**

### **Applying Logistic Regression**

Now, I will train a machine learning model for the task of heart disease prediction. I will use the logistic regression algorithm as I mentioned at the beginning of the article. But before training the model I will first define a helper function for printing the classification report of the performance of the machine learning model:

First I split the data into training and test sets. I will split the data into 70% training and 30% testing. Then I train the machine learning model and print the classification report of our logistic regression model:

Train Result:

=====

Accuracy Score: 86.79%

CLASSIFICATION REPORT:

	0	1	accuracy	macro avg	weighted avg
precision	0.879121	0.859504	0.867925	0.869313	0.868480
recall	0.824742	0.904348	0.867925	0.864545	0.867925
f1-score	0.851064	0.881356	0.867925	0.866210	0.867496
support	97.000000	115.000000	0.867925	212.000000	212.000000

Confusion Matrix:

```
[[ 80  17]
 [ 11 104]]
```

Test Result:

=====

Accuracy Score: 86.81%

CLASSIFICATION REPORT:

	0	1	accuracy	macro avg	weighted avg
precision	0.871795	0.865385	0.868132	0.868590	0.868273
recall	0.829268	0.900000	0.868132	0.864634	0.868132
f1-score	0.850000	0.882353	0.868132	0.866176	0.867776
support	41.000000	50.000000	0.868132	91.000000	91.000000

Confusion Matrix:

```
[[34  7]
 [ 5 45]]
```

	Model	Training Accuracy %	Testing Accuracy %
0	Logistic Regression	86.792453	86.813187

As we can see the model performs very well of the test set as it is giving almost the same accuracy in the test set as in the training set. The training accuracy is 86.8%.

### 3.3 the web app

In this section, we deploy our modal as machine learning web app using flask framework.

I use the modal we train so the system will the data given by the user and the system will predict and give the result using the dataset we gave.

#### 3.3.1 the web user interface

#### Home page:

ML Medical App
Home
About
Heart
Documentation

### Background/ Problem Statement

Heart Disease (including Coronary Heart Disease, Hypertension, and Stroke) remains the No. 1 cause of death in the many parts of the world. The Heart Disease and Stroke Statistics—2019 Update from the American Heart Association indicates that:

- 116.4 million, or 46% of US adults are estimated to have hypertension. These are findings related to the new 2017 Hypertension Clinical Practice Guidelines.
- On average, someone dies of CVD every 38 seconds. About 2,303 deaths from CVD each day, based on 2016 data.
- On average, someone dies of a stroke every 3.70 minutes. About 389.4 deaths from stroke each day, based on 2016 data.


To predict this disease and solve the problem using advanced technological solutions, machine Learning algorithms, I have created this Heart Disease Prediction Web App System using the Logistic Regression Machine Learning algorithm. This web app takes a statistical techniques to draw out a result based on trained model. This web app is intended:

1. to help doctors detect heart disease at an early stage where it can be cured, thereby saving many lives.
2. To Help people who cannot afford medical fees as the system is free and available every time.

### Advantages

- User can easily get the heart disease prediction on single click
- The system is free and available at any time
- System is kept online in order to serve people 24x7
- The Heart Disease Prediction web application is an end user support and online consultation project
- The application is fed with various details and the Heart disease associated with those details and it allows user to share their heart related issues. It then processes user specific details to check for various cancer disease that could be associated with the inputs received from user.
- Accurate prediction

### Project Contributor




#### About page:

ML Medical App
Home
About
Heart
Documentation

## Heart Disease Prediction using Logistic Regression algorithm in machine learning

### Introduction to Heart Disease

#### Overview



Heart disease describes a range of conditions that affect your heart. Diseases under the heart disease umbrella include blood vessel diseases, such as coronary artery disease; heart rhythm problems (arrhythmias); and heart defects you're born with (congenital heart defects), among others. The term "heart disease" is often used interchangeably with the term "cardiovascular disease." Cardiovascular disease generally refers to conditions that involve narrowed or blocked blood vessels that can lead to a heart attack, chest pain (angina) or stroke. Other heart conditions, such as those that affect your heart's muscle, valves or rhythm, also are considered forms of heart disease. Many forms of heart disease can be prevented or treated with healthy lifestyle choices.

**Heart diseases include**

- vessel disease, such as coronary artery disease
- Irregular heartbeats (arrhythmias)
- Heart problems you're born with (congenital heart defects)
- Disease of the heart muscle
- Heart valve disease

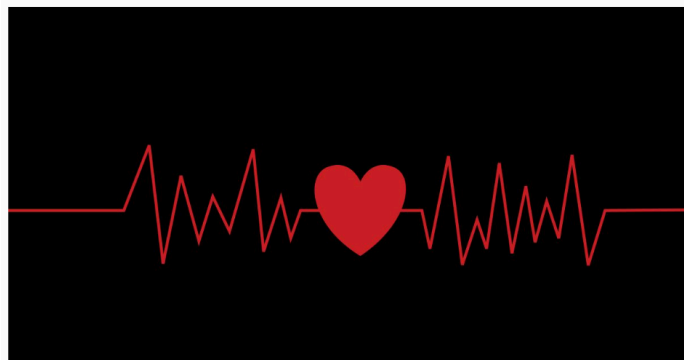
#### Symptoms

- Chest pain, chest tightness, chest pressure and chest discomfort (angina)
- Shortness of breath

## Heart disease prediction page:

[ML Medical App](#) [Home](#) [About](#) [Heart](#) [Documentation](#)

### Welcome To The Heart Disease Predict Health App



#### Enter The Following Features

Chest Pain Type

Resting Blood Pressure (in mm Hg)

Serum Cholesterol in mg/dl

Fasting Blood Sugar

Resting Electro-cardiographic Result

Thalach (the maximum heart rate acheived)

Exercise Induced Angina

Predict

## Documentation page:

Name: Abdulkadir Duran Adan

Stu\_ID: 5035190144

### Machine Learning Final Project

Heart Disease Prediction using Logistic Regression algorithm in machine learning

Programming language: Python

#### 1. Importing Packages (Libraries)

```
In [202]: import warnings
warnings.filterwarnings('ignore')
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
%matplotlib inline
sns.set_style('whitegrid')
plt.style.use('fivethirtyeight')
```

#### 2. Importing and Loading Dataset

```
In [203]: data_frame = pd.read_csv("heart.csv")
```

#### 3. displaying sample entries of dataset so that we can see what we are working on

```
In [204]: data_frame.head()
```

```
Out[204]:
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.30	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.50	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.40	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.80	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.60	2	0	2	1

#### 4.Data Cleaning & Preprocessing

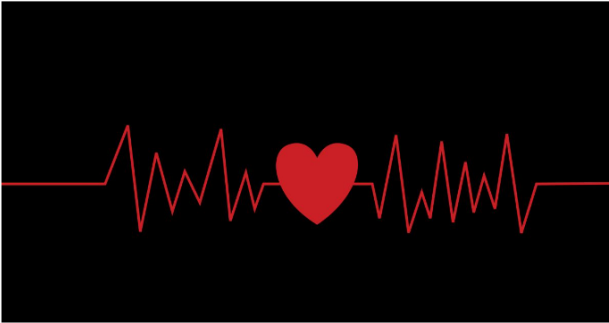


### 3.3 testing the web app

#### User input data:

ML Medical App Home About Heart Documentation

### Enter The Following Features



Chest Pain Type  
Typical Angina

Resting Blood Pressure (in mm Hg)  
144

Serum Cholesterol in mg/dl  
130

Fasting Blood Sugar  
Fasting Blood Sugar > 120 mg/dl

Resting Electro-cardiographic Result  
having ST-T wave abnormality

Thalach (the maximum heart rate acheived)  
150

Exercise Induced Angina  
Yes

Predict

#### The system response/result

**Sorry! your are suffering from heart disease, please consult a doctor as soon as possible.**

[Return to Home](#)

#### User input data:

## Enter The Following Features



Chest Pain Type

Typical Angina

Resting Blood Pressure (in mm Hg)

130

Serum Cholesterol in mg/dl

90

Fasting Blood Sugar

Fasting Blood Sugar &lt; 120 mg/dl

Resting Electro-cardiographic Result

Normal

Thalach (the maximum heart rate achieved)

82

Exercise Induced Angina

Yes

Predict

***The system response/result*****Congrates. Your heart is healthy.****Return to Home**

## **4. Conclusion**

In this paper, we have trained a model for the heart disease prediction task and then we have created machine learning web app in which users can use any time for free. We have used logistic regression which gave us the most accuracy.

In this paper, logistic regression models were used to explore the feasibility of predicting heart disease. Experiments were conducted using the data set provided by kaggle and the results were evaluated. Significant features of logistic regression models affecting heart disease were found: Age, sex, cp, chol, restecg, oldpeak, slope, ca, thal.

The main influencing factors and logistic regression technology are used to establish the prediction model, and the accuracy of the model is compared with the model proposed in the existing research. According to the test results, the classification model proposed is highly accurate and has certain research value.

## **5. References.**

- [1] Mohammad Shafenoor Amin, Yin Kia Chiam, Kasturi Dewi Varathan. Identification of significant features and data mining techniques in predicting heart disease.
- [2] Cincy Raju, Philipsey E, Siji Chacko, L Padma Suresh, Deepa Rajan S, A Survey on Predicting Heart Disease using Data Mining Techniques.
- [3] Sana Bharti, Dr.Shaliendra Narayan Singh, Analytical Study of Heart Disease Prediction Comparing With Different Algorithms.
- [4] Montu Saw, Tarun Saxena, Sanjana Kaithwas, Rahul Yadav, Nidhi Lal. Estimation of Prediction for Getting Heart Disease Using Logistic Regression Model of Machine Learning.
- [5] K.Srinivas, Dr.G.Raghavendra Rao, Dr. A.Govardhan. Analysis of Coronary Heart Disease and Prediction of Heart Attack in Coal Mining Regions Using Data Mining Techniques..
- [6] World Health Organization (WHO) [https://www.who.int/health-topics/cardiovascular-diseases#tab=tab\\_1](https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1)