**Vivekanand Education Society's Institute of Technology**
**Hashu Advani memorial Complex Collector's Colony**
**R C Marg, Chembur, Mumbai 400074**

**DEPARTMENT OF INFORMATION TECHNOLOGY**


MINI PROJECT REPORT

ON

# "Heart Disease Diagnosis"

**B.E. (Information Technology)**

*SUBMITTED BY*

**Prerna Peswani - 43**
**Shasmita Raveendran - 48**
**Jai Rohra - 51**

*UNDER THE GUIDANCE OF*

**PROF. Dimple Bohra**

**(Academic Year: 2022-2023)**

**Mumbai University**
**Vivekanand Education Society's Institute Of Technology, Mumbai**

**DEPARTMENT OF INFORMATION TECHNOLOGY**


# *Certificate*

This is to certify that project entitled

**"Heart Disease Diagnosis"**

Prerna Peswani - 43
Shasmita Raveendran - 48
Jai Rohra - 51

have satisfactorily carried out the project work, under the head - Data Science Lab at Semester VII of BE-IT in Information Technology as prescribed by the Mumbai University.

**Prof. Dimple Bohra**
**Subject Teacher**

**External Examiner**

**Dr.(Mrs.)Shalu Chopra**
**H.O.D**

**Dr.(Mrs.)J.M.Nair**
**Principal**

Date: / /2022
Place: VESIT, Chembur

# *LO Mapping*

LO1: To apply reasoning for a problem in an uncertain domain.

LO2: To discuss the solution after building a Cognitive application.

LO3: To familiarize the students with the basics of Fuzzy Logic and Fuzzy Systems.

LO4: To familiarize the students with Learning Architectures and Frameworks.

LO5: To define and apply metrics to measure the performance of various learning algorithms.

LO6: To enable students to analyze data science methods for real world problems.

# *Declaration*

I declare that this written submission represents my ideas in my own words and where other's ideas or words have been included, I have adequately cited and referenced the original source. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

- - - - - - - - - - -

**(Signature)**

Mr/Ms
**Prerna Peswani - 43**
**Shasmita Raveendran - 48**
**Jai Rohra - 51**
B.E. INFT

# Contents

# List of Figures

**Abstract**

Machine Learning is used across many ranges around the world. The healthcare industry is no exclusion. Machine Learning can play an essential role in predicting presence/absence of locomotors disorders, Heart diseases and more. Such information, if predicted well in advance, can provide important intuitions to doctors who can then adapt their diagnosis and dealing per patient basis. We work on predicting possible Heart Diseases in people using Machine Learning algorithms. In this project we perform the comparative analysis of classifiers like decision tree, Naïve Bayes, Logistic Regression, SVM and Random Forest and we propose an ensemble classifier which perform hybrid classification by taking strong and weak classifiers since it can have multiple number of samples for training and validating the data so we perform the analysis of existing classifier and proposed classifier like Ada-boost and XG-boost which can give the better accuracy and predictive analysis.

**Keywords-** ***SVM; Naive Bayes; Decision Tree; Random Forest; Logistic Regression; Adaboost; XG-boost; python programming; confusion matrix; correlation matrix.***

# Chapter 1

# Introduction

## 1.1 Introduction

According to the World Health Organization, every year 12 million deaths occur worldwide due to Heart Disease. Heart disease is one of the biggest causes of morbidity and mortality among the population of the world. Prediction of cardiovascular disease is regarded as one of the most important subjects in the section of data analysis. The load of cardiovascular disease is rapidly increasing all over the world from the past few years. Many researches have been conducted in attempt to pinpoint the most influential factors of heart disease as well as accurately predict the overall risk. Heart Disease is even highlighted as a silent killer which leads to the death of the person without obvious symptoms. The early diagnosis of heart disease plays a vital role in making decisions on lifestyle changes in high-risk patients and in turn reduces the complications.

Among all fatal diseases, heart attack diseases are considered as the most prevalent. Medical practitioners conduct different surveys on heart diseases and gather information of heart patients, their symptoms and disease progression. Increasingly are reports about patients with common diseases who have typical symptoms.

In this fast moving world people want to live a very luxurious life so they work like a machine in order to earn lot of money and live a comfortable life therefore in this race they forget to take care of themselves, because of this there food habits change their entire lifestyle change, in this type of lifestyle they are more tensed they have blood pressure, sugar at a very young age and they don't give enough rest for themselves and eat what they get and they even don't bother about the quality of the food if sick the go for their own medication as a result of all these small negligence it leads to a major threat that is the heart disease The term 'heart disease' includes the diverse diseases that affect heart. The number of people suffering from heart disease is on the rise (health topics, 2010). The report from the World Health Organization shows us a large number of people that die every year due to heart disease all over the world. Heart disease is also stated as one of the greatest killers in Africa.

Machine learning proves to be effective in assisting in making decisions and predictions from the large quantity of data produced by the healthcare industry. This project aims to predict future Heart Disease by analyzing data of patients which classifies whether they have heart disease or not using a machine-learning algorithm. Machine

Learning techniques can be a boon in this regard. Even though heart disease can occur in different forms, there is a common set of core risk factors that influence whether someone will ultimately be at risk for heart disease or not. By collecting the data from various sources, classifying them under suitable headings finally analysing to extract the desired data we can say that this technique can be very well adapted to do the prediction of heart disease.

## 1.2 Problem Statement

Heart disease can be managed effectively with a combination of lifestyle changes, medicine and, in some cases, surgery. With the right treatment, the symptoms of heart disease can be reduced and the functioning of the heart improved. The predicted results can be used to prevent and thus reduce cost for surgical treatment and other expenses. The overall objective of the project will be to predict accurately with few tests and attributes the presence of heart disease. Attributes considered form the primary basis for tests and give accurate results more or less. Many more input attributes can be taken but our goal is to predict with few attributes and faster efficiency the risk of having heart disease. Decisions are often made based on doctors' intuition and experience rather than on the knowledge rich data hidden in the data set and databases. This practice leads to unwanted biases, errors and excessive medical costs which affect the quality of service provided to patients. Data mining holds great potential for the healthcare industry to enable health systems to systematically use data and analytics to identify inefficiencies and best practices that improve care and reduce costs. According to (Wurz Takala, 2006) the opportunities to improve care and reduce costs concurrently could apply to as much as 30

## 1.3 Objectives of the project

- To develop a heart prediction system
- To develop a system that can discover and extract hidden knowledge associated with diseases from a historical heart data set.
- To develop a system that aims to exploit data mining techniques on medical data sets to assist in the prediction of heart diseases.
- To provide a new approach to concealed patterns in the data and help avoid human biases. to effectively predict if the patient suffers from heart disease.

## 1.4 Functionalities

The working of the system starts with the collection of data and selecting the important attributes. Then the required data is preprocessed into the required format. The data is then divided into two parts training and testing data. The algorithms are applied and the model is trained using the training data. The accuracy of the system is obtained by testing the system using the testing data. This system is implemented using the following modules.

1. Collection of Dataset
2. Selection of attributes
3. Data Pre-Processing
4. Balancing of Data
5. Disease Prediction

## 1.5 Scope

Here the scope of the project is that integration of clinical decision support with computer-based patient records could reduce medical errors, enhance patient safety, decrease unwanted practice variation, and improve patient outcome. This suggestion is promising as data modeling and analysis tools, e.g Data mining, have the potential to generate a knowledge-rich environment which can help to significantly improve the quality of clinical decisions. This project aims to predict future Heart Disease by analyzing data of patients which classifies whether they have heart disease or not using a machine-learning algorithm. By collecting the data from various sources, classifying them under suitable headings finally analysing to extract the desired data we can say that this technique can be very well adapted to do the prediction of heart disease.

# Chapter 2

# Analysis and Design

## 2.1 Analysis of the system

Heart disease is even being highlighted as a silent killer which leads to the death of a person without obvious symptoms. The nature of the disease is the cause of growing anxiety about the disease its consequences. Hence continued efforts are being done to predict the possibility of this deadly disease in prior. So that various tools techniques are regularly being experimented with to suit the present-day health needs. Machine Learning techniques can be a boon in this regard. Even though heart disease can occur in different forms, there is a common set of core risk factors that influence whether someone will ultimately be at risk for heart disease or not. By collecting the data from various sources, classifying them under suitable headings finally analysing to extract the desired data we can conclude. This technique can be very well adapted to the do the prediction of heart disease. As the well-known quote says "Prevention is better than cure", early prediction its control can be helpful to prevent decrease the death rates due to heart disease.

## 2.2 Design of the proposed system

In machine learning, classification refers to a predictive modeling problem where a class label is predicted for a given example of input data. Supervised learning is the type of machine learning in which machines are trained using well "labelled" training data, and on the basis of that data, machines predict the output. The labelled data means some input data is already tagged with the correct output. Supervised learning is a process of providing input data as well as correct output data to the machine learning model. The aim of a supervised learning algorithm is to find a mapping function to map the input variable(x) with the output variable(y). Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.

```
LogisticRegression()
```

Model Evaluation

Accuracy Score

```
# accuracy on training data
X_train_prediction = model.predict(X_train)
training_data_accuracy = accuracy_score(X_train_prediction, Y_train)
```

```
print('Accuracy on Training data : ', training_data_accuracy)
```

```
Accuracy on Training data :  0.8512396694214877
```

```
# accuracy on test data
X_test_prediction = model.predict(X_test)
test_data_accuracy = accuracy_score(X_test_prediction, Y_test)
```

```
print('Accuracy on Test data : ', test_data_accuracy)
```

```
Accuracy on Test data :  0.819672131147541
```

### 2.2.1 Architecture/ Block Diagram

The system architecture gives an overview of the working of the system. The working of this system is described as follows: Dataset collection is collecting data which contains patient details. Attributes selection process selects the useful attributes for the prediction of heart disease. After identifying the available data resources, they are further selected, cleaned, made into the desired form. Different classification techniques as stated will be applied on preprocessed data to predict the accuracy of heart disease. Accuracy measure compares the accuracy of different classifiers.


### 2.2.2 Algorithms Used

Logistic regression is one of the most popular Machine Learning algorithms,which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas logistic regression is used for solving the classification problems. In Logistic regression, instead of fitting a regression line, we fit an "S"shaped logistic function, which predicts two maximum values (0 or 1). The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on itsweight, etc.

Advantages:

Logistic Regression is one of the simplest machine learning algorithms and is easy to implement yet provides great training efficiency in some cases. Also due to these reasons, training a model with this algorithm doesn't require high computation power.

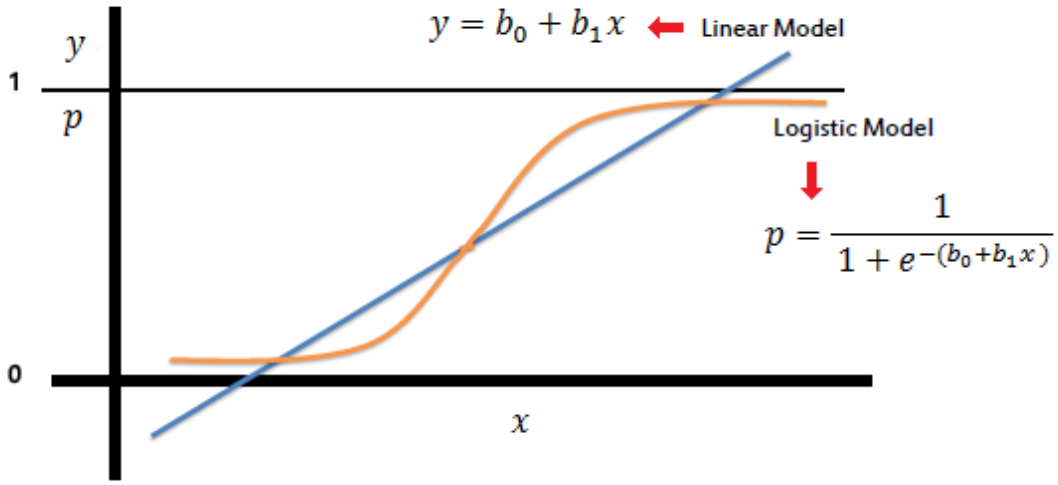
The predicted parameters (trained weights) give inference about the importance of each feature. The direction of association i.e. positive or negative is also given. So we can use Logistic Regression to find out the relationship between the features. This algorithm allows models to be updated easily to reflect new data, unlike Decision Tree or Support Vector Machine. The update can be done using stochastic gradient descent. Logistic Regression outputs well-calibrated probabilities along with classification results. This is an advantage over models that only give the final classification as results. If a training example has a 95another has a 55training examples are more accurate for the formulated problem.

Disadvantages:

Logistic Regression is a statistical analysis model that attempts to predict precise probabilistic outcomes based on independent features. On high dimensional datasets, this may lead to the model being over-fit on the training set, which means overstating the accuracy of predictions on the training set and thus the model may not be able to predict accurate results on the test set. This usually happens in the case when the model is trained on little training data with lots of features. So on high dimensional datasets, Regularization techniques should be considered to avoid overfitting (but this makes the model complex). Very high regularization factors may even lead to the model being under-fit on the training data. Non linear problems can't be solved with logistic regression since it has a linear decision surface. Linearly separable data is rarely found in real world scenarios. So the transformation of non linear features is required which can be done by increasing the number of features such that the data becomes linearly separable in higher dimensions.

Non-Linearly Separable Data:

It is difficult to capture complex relationships using logistic regression. More powerful and complex algorithms such as Neural Networks can easily outperform this algorithm



### 2.2.3 Details of Hardware and Software

The software and hardware requirements are as follows

Hardware requirements:
Processer : Any Update Processer
Ram : Min 4GB
Hard Disk : Min 100GB

Software requirements:
Operating System : Windows family
Technology : Python3.8
IDE : Jupiter notebook

## 2.3 Tools and Datasets

Tools:

Python:

Python is a high-level, general-purpose programming language. Its design philosophy emphasizes code readability with the use of significant indentation. Python is dynamically-typed and garbage-collected. It supports multiple programming paradigms, including structured, object-oriented and functional programming.

Pandas:

pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free software released under the three-clause BSD license.

Numpy:

NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.

Scikit-learn:

Scikit-learn is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support-vector machines.

Flask:

Flask is a micro web framework written in Python. It is classified as a microframework because it does not require particular tools or libraries. It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions.

Dataset:

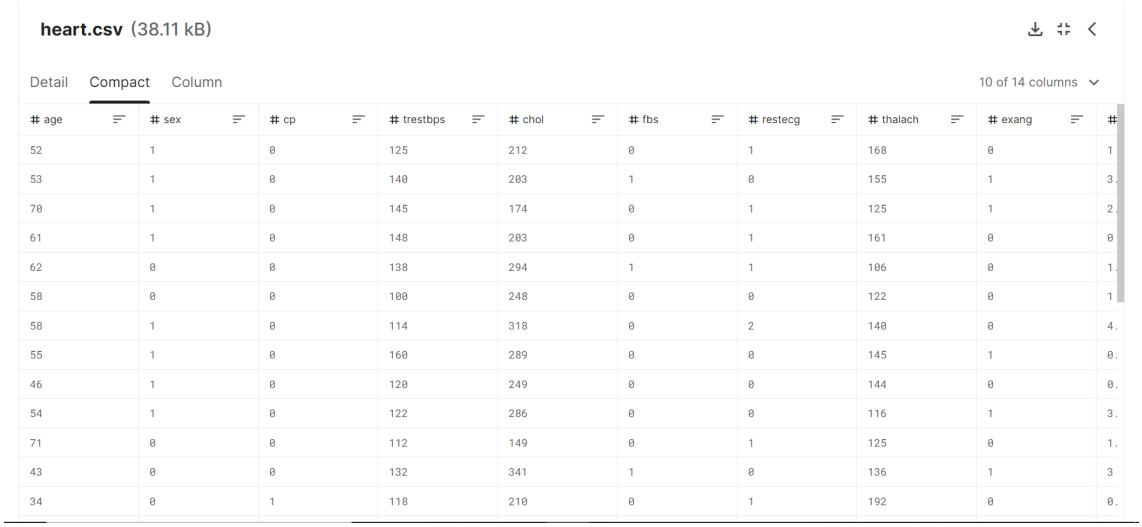
Heart Disease Dataset

This data set dates from 1988 and consists of four databases: Cleveland, Hungary, Switzerland, and Long Beach V. It contains 76 attributes, including the predicted

attribute, but all published experiments refer to using a subset of 14 of them. The "target" field refers to the presence of heart disease in the patient. It is integer valued 0 = no disease and 1 = disease.

Attribute Information:

age
sex
chest pain type (4 values)
resting blood pressure
serum cholestoral in mg/dl
fasting blood sugar ¿ 120 mg/dl
resting electrocardiographic results (values 0,1,2)
maximum heart rate achieved
exercise induced angina
oldpeak = ST depression induced by exercise relative to rest the slope of the peak exercise ST segment
number of major vessels (0-3) colored by flourosopy
thal: 0 = normal; 1 = fixed defect; 2 = reversable defect



**heart.csv** (38.11 kB)

Detail | Compact | Column — 10 of 14 columns

| # age | # sex | # cp | # trestbps | # chol | # fbs | # restecg | # thalach | # exang | # |
|---|---|---|---|---|---|---|---|---|---|
| 52 | 1 | 0 | 125 | 212 | 0 | 1 | 168 | 0 | 1 |
| 53 | 1 | 0 | 140 | 203 | 1 | 0 | 155 | 1 | 3. |
| 70 | 1 | 0 | 145 | 174 | 0 | 1 | 125 | 1 | 2. |
| 61 | 1 | 0 | 148 | 203 | 0 | 1 | 161 | 0 | 0 |
| 62 | 0 | 0 | 138 | 294 | 1 | 1 | 106 | 0 | 1. |
| 58 | 0 | 0 | 100 | 248 | 0 | 0 | 122 | 0 | 1 |
| 58 | 1 | 0 | 114 | 318 | 0 | 2 | 140 | 0 | 4. |
| 55 | 1 | 0 | 160 | 289 | 0 | 0 | 145 | 1 | 0. |
| 46 | 1 | 0 | 120 | 249 | 0 | 0 | 144 | 0 | 0. |
| 54 | 1 | 0 | 122 | 286 | 0 | 0 | 116 | 1 | 3. |
| 71 | 0 | 0 | 112 | 149 | 0 | 1 | 125 | 0 | 1. |
| 43 | 0 | 0 | 132 | 341 | 1 | 0 | 136 | 1 | 3 |
| 34 | 0 | 1 | 118 | 210 | 0 | 1 | 192 | 0 | 0. |

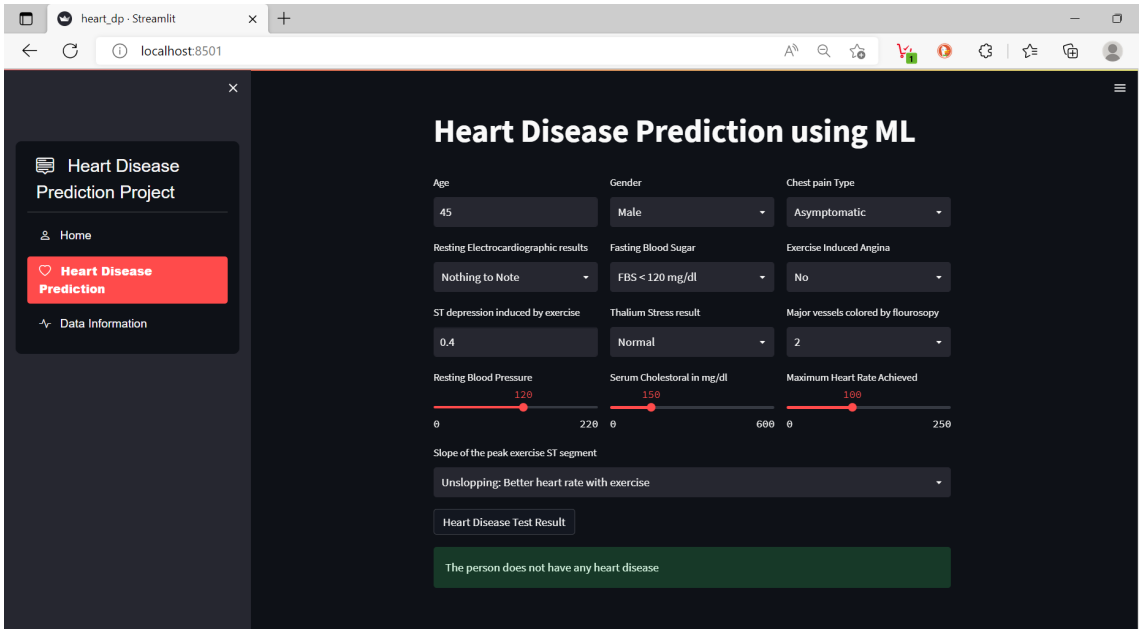
### 2.3.1 Experimental Results (Code and GUI)

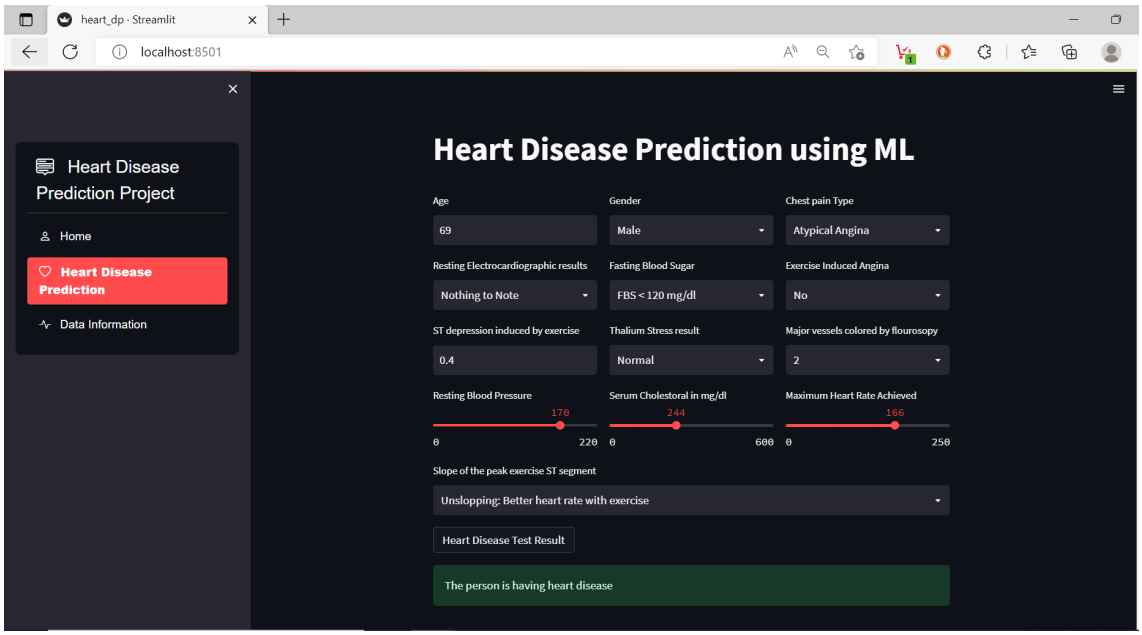
Code:

```
# Importing Libraries:
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score


# Splitting the Features and Target
X = heart_data.drop(columns='target', axis=1)
```

```
Y = heart_data['target']

# Train Test Split:
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, stratify=Y
print(X.shape, X_train.shape, X_test.shape)

# Logistic Regression:
model = LogisticRegression()

# training the Logistic Regression model with Training data
model.fit(X_train, Y_train)

# Building a Predictive System
input_data = (62,0,0,140,268,0,0,160,0,3.6,0,2,2)

# change the input data to a numpy array
input_data_as_numpy_array= np.asarray(input_data)

# reshape the numpy array as we are predicting for only on instance
input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)

prediction = model.predict(input_data_reshaped)
print(prediction)

if (prediction[0]== 0):
  print('The Person does not have a Heart Disease')
else:
  print('The Person has Heart Disease')
```

GUI:




heart_dp - Streamlit
localhost:8501
Heart Disease Prediction Project
Home
Heart Disease Prediction
Data Information
Heart Disease Prediction using ML
Age
69
Gender
Male
Chest pain Type
Atypical Angina
Resting Electrocardiographic results
Nothing to Note
Fasting Blood Sugar
FBS < 120 mg/dl
Exercise Induced Angina
No
ST depression induced by exercise
0.4
Thalium Stress result
Normal
Major vessels colored by flourosopy
2
Resting Blood Pressure
170
0
220
Serum Cholestoral in mg/dl
244
0
600
Maximum Heart Rate Achieved
166
0
250
Slope of the peak exercise ST segment
Unslopping: Better heart rate with exercise
Heart Disease Test Result
The person is having heart disease

# Chapter 3

# Conclusion and Future Work

The proposed system is GUI-based, user-friendly, scalable, reliable and an expandable system. The proposed working model can also help in reducing treatment costs by providing Initial diagnostics in time. The model can also serve the purpose of training tools for medical students and will be a soft diagnostic tool available for physicians and cardiologists. General physicians can utilize this tool for initial diagnosis of cardiac-patients. There are many possible improvements that could be explored to improve the scalability and accuracy of this prediction system. As we have developed a generalized system, in future we can use this system for the analysis of different data sets. The performance of the health's diagnosis can be improved significantly by handling numerous class labels in the prediction process, and it can be another positive direction of research. In DM warehouses, generally, the dimensionality of the heart database is high, so identification and selection of significant attributes for better diagnosis of heart disease are very challenging tasks for future research.

# References


[1] Dangare C S Apte S S (2012). *'Improved study of heart disease prediction system using data mining classification techniques.'* International Journal of Computer Applications, 47(10), 44-8

[2] Soni J, Ansari U, Sharma D Soni S (2011). *'Predictive data mining for medical diagnosis: an overview of heart disease prediction'* International Journal of Computer Applications, 17(8), 43-8

[3] Ordonez C (2006). *Association rule discovery with the train and test approach for heart disease prediction* IEEE Transactions on Information Technology in Biomedicine, 10(2), 334-43.