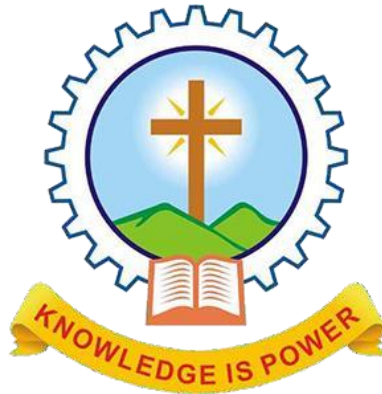


**MAR ATHANASIOUS COLLEGE OF ENGINEERING**  
**(Affiliated to APJ Abdul Kalam Technological University, TVM)**  
**KOTHAMANGALAM**



**Department of Computer Applications**

Main Project Report

# **HEART DISEASE PREDICTION**

Done by

**Rejitha Ramesh**

**Reg No: MAC20MCA-2017**

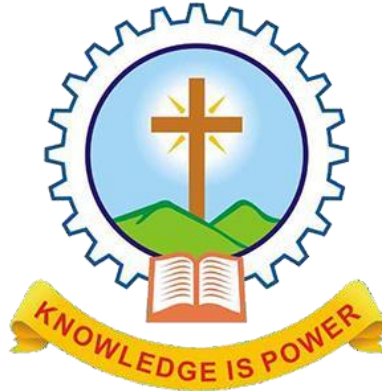
Under the guidance of

**Prof. Nisha Markose**

**2020-2022**

**MAR ATHANASIOUS COLLEGE OF ENGINEERING**  
**(Affiliated to APJ Abdul Kalam Technological University, TVM)**  
**KOTHAMANGALAM**

**CERTIFICATE**



**HEART DISEASE PREDICTION**

Certified that this is the bonafide record of project work done by

**Rejitha Ramesh**

**Reg No: MAC20MCA-2017**

During the academic year 2020-2022, in partial fulfilment of requirements for  
award of the degree,

**Master of Computer Applications**

**of**

**APJ Abdul Kalam Technological University**

**Thiruvananthapuram**

**Faculty Guide**

Prof. Nisha Markose

**Head of the Department**

Prof. Biju Skaria

**Project Coordinator**

Prof. Biju Skaria

**External Examiner**

## **ACKNOWLEDGEMENT**

First and foremost, I thank God Almighty for his divine grace and blessings in making all this possible. May he continue to lead me in the years to come.

I am also grateful to Prof. Biju Skaria, Head of Computer Applications Department and my project guide Prof. Nisha Markose, Associate Professor, Department of Computer Applications for her valuable guidance and constant supervision as well as for providing necessary information regarding the Main project & also for his support.

I am highly indebted to our project coordinator Prof. Biju Skaria, Associate Professor, Department of Computer Applications for his guidance and support.

I profusely thank other Professors in the department and all other staffs of MACE, for their guidance and inspirations throughout my course of study. No words can express my humble gratitude to my beloved parents who have been guiding me in all walks of my journey. My thanks and appreciations also go to my friends and people who have willingly helped me out with their abilities.

## ABSTRACT

The heart is one of the main parts of the human body after the brain. The primary function of the heart is to pump blood to the whole body parts. Any disorder that can lead to disturbing the functionality of the heart is called heart disease. Different factors can raise the risk of heart failure. Medical scientists have classified those factors into two different categories; one of them is risk factors that cannot be changed, and another one is risk factors that can be changed. Family history, sex, age comes under risk factors that cannot be changed. High cholesterol, smoking, physical inactivity, high blood pressure all these come under risk factors.

Heart disease is a significant issue, so there is a need for diagnosis or prediction of heart disease. There are several methods to diagnose heart disease among them Angiography is the trending method which is used by most of the physicians across the world. The conventional methods may give erroneous results because these conventional methods are performed by humans. To avoid these errors and to achieve better and faster results, we need an automated system. The proposed system is meant for the same. The algorithm which is proposed in the project is support vector machine.

The proposed project uses the dataset named “framingham.csv”. It has features such as age, gender, blood pressure, diabetes, heart rate, blood glucose level, etc. The dataset contains 4135 records. The dataset used in this project to predict the heart disease is downloaded from <https://www.kaggle.com/datasets/captainozlem/framingham-chd-preprocessed-data>

# List of Tables

- 1. About the Dataset ..... 6
- 2. Description of dataset..... 7
- 3. Dataset..... 8

# List of Figures

1. Handling Missing Values .....	9
2. Prediction Features.....	10
3. Data Visualization .....	.11
4. Gender v/s TenYearCHD .....	.12
5. Age v/s TenYearCHD .....	12
6. CurrentSmoker v/s TenYearCHD .....	13
7. BPMeds v/s TenYearCHD .....	13
8. Heart Rate v/s TenYearCHD.....	14
9. SVM.....	16
10. Flowchart .....	17
11. Project Pipeline.....	19
12. SVM Splitting.....	28
13. SVM Training.....	29
14. Prediction.....	30
15. Evaluation of SVM.....	31
16. Home Page.....	33
17. Option Page .....	34
18. Page for Entering Values.....	35
19. Output Prediction.....	36
20. Prediction Report.....	37
21. Git History .....	38

# Contents

<b>1 INTRODUCTION</b>	<b>1</b>
<b>2 SUPPORTING LITERATURE</b>	<b>2</b>
2.1 Literature Review .....	2
2.2 Findings and Proposals .....	5
<b>3 SYSTEM ANALYSIS</b>	<b>6</b>
3.1 Analysis of Dataset .....	6
3.1.1 About the Dataset .....	6
3.1.2 Explore the Dataset .....	7
3.2 Data Preprocessing .....	9
3.2.1 Data Cleaning .....	9
3.2.2 Analysis of Feature Variables .....	10
3.2.3 Analysis of Class Variables .....	10
3.3 Data Visualization .....	11
3.4 Analysis of Algorithms .....	14
3.5 Project Pipeline... ..	19
3.6 Feasibility Analysis .....	21
3.7 System Environment .....	23
3.7.1 Software Environment .....	23
3.7.2 Hardware Environment .....	27

<b>4</b>	<b>SYSTEM DESIGN</b>	<b>28</b>
4.1	Model Building .....	28
4.1.1	Model Planning.....	28
4.1.2	Training .....	29
4.1.3	Testing .....	29
<b>5</b>	<b>RESULT AND DISCUSSION</b>	<b>31</b>
<b>6</b>	<b>MODEL DEPLOYMENT</b>	<b>32</b>
<b>7</b>	<b>GIT HISTORY</b>	<b>38</b>
<b>8</b>	<b>CONCLUSION</b>	<b>39</b>
<b>9</b>	<b>FUTURE WORK</b>	<b>40</b>
<b>10</b>	<b>APPENDIX</b>	<b>41</b>
10.1	Minimum Software Requirements .....	41
10.2	Minimum Hardware Requirements.....	41
<b>11</b>	<b>REFERENCES</b>	<b>42</b>



# 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.

Machine learning proves to be effective in assisting in making decisions and predictions from the large quantity of data produced by the health care industry. This project aims to predict future Heart Disease by analyzing data of patients which classifies whether they have heart disease or not using 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 analyzing to extract the desired data we can say that this technique can be very well adapted to do the prediction of heart disease.

## **2. SUPPORTING LITERATURE**

### **2.1 Literature Review**

[1] Rahul Katarya and Polipireddy Srinivas “Predicting Heart Disease at Early Stages using Machine Learning: A Survey”. 2020 International Conference on Advanced Computing and Communication System (ICACCS).

The heart is one of the main parts of the human body after the brain. The primary function of the heart is to pumping blood to the whole body parts. Any disorder that can lead to disturbing the functionality of the heart is called heart disease. Heart disease is a very critical issue in the present growing world. So, there is a need for an automated system to predict heart disease at earlier stages. So that it will be useful for the physician to diagnose the patients efficiently, and it will be useful to the people also because they can track their health issues by using this automated system. Some of the expert automated systems were summarized in this paper. Feature selection and prediction, these two are essential for every automated system. By choosing features efficiently, we can achieve better results in predicting heart disease. We have summarized some algorithms which are useful while selecting the features, like hybrid grid search algorithm and random search algorithm, etc. So, in the future, it is better to use search algorithms for selecting the features and then applying machine learning techniques for prediction will give us better results in the prediction of heart disease.

[2] Cincy Raju, Philippsy E, Siji Chacko, L Padma Suresh and Deepa Rajan S “A Survey on Predicting Heart Disease using Data Mining Techniques”. 2018 IEEE Conference on Emerging Devices and Smart Systems (ICEDSS).

Heart disease is the type of disease that involves the heart or blood vessels. It is one of the most-flying diseases of the modern world. The diagnosis of the heart disease should be accurately and correctly. The motivation of this paper is to develop an efficacious treatment using data mining techniques that can help remedial situations. Further data mining classification algorithms like decision trees, neural networks, Bayesian classifiers, Support vector machines, Association Rule, K- nearest neighbour classification are used to diagnosis the heart diseases. Among these algorithms Support Vector Machine (SVM) gives best result. The various heart disease prediction techniques are discussed and analyzed in this paper. The data mining techniques used to predict heart diseases are discussed here. Heart disease is a mortal disease by its nature. This disease makes several problems such as heart attack and death. In the medical domain, the significance of data mining is perceived. Various steps are taken to apply pertinent techniques in the disease prediction. The research works with effective techniques that are done by different researchers were studied in this paper. From the comparative study we can conclude that Support Vector Machine (SVM) technique is an efficient method for predicting heart disease. It gives good accuracy by observing various research papers.

[3] J.Neelaveni,M.S.Geetha Devasana “Heart Disease Prediction Using Machine using Machine Learning Algorithm”. 2020 International Conference on Electrical and Electronics Engineering (ICE3-2020).

Heart plays significant role in living organisms. Diagnosis and prediction of heart related diseases requires more precision, perfection and correctness because a little mistake can cause fatigue problem or death of the person, there are numerous death cases related to heart and their counting is increasing exponentially day by day. To deal with the problem there is essential need of prediction system for awareness about diseases. As the definition of machine learning, it learns from the natural phenomenon, natural things so in this project we uses the biological parameter as testing data such as cholesterol, Blood pressure, sex, age, etc. and on the basis of these, comparison is done in the terms of accuracy of algorithms such as in this project we have used SVM algorithm. Heart is one of the essential and vital organ of human body and prediction about heart diseases is also important concern for the human beings so that the accuracy for algorithm is one of parameter for analysis of performance of algorithms. Accuracy of the algorithms in machine learning depends upon the dataset that used for training and testing purpose. When we perform the analysis of algorithms on the basis of dataset and on the basis of confusion matrix, we find KNN is best one. For the Future Scope more machine learning approach will be used for best analysis of the heart diseases and for earlier prediction of diseases so that the rate of the death cases can be minimized by the awareness about the diseases.

## 2.2 Findings and Proposals

Heart disease is a significant issue, so there is a need for diagnosis or prediction of heart disease there are several methods to diagnose heart disease among them Angiography is the trending method which is used by most of the physicians across the world. The conventional methods may give erroneous results because these conventional methods are performed by humans. To avoid these errors and to achieve better and faster results, we need an automated system.

The referred papers are concentrated on Heart disease prediction using various machine learning methods. Heart Disease Prediction Using Machine using Machine Learning Algorithm Predicting Heart Disease at Early Stages using Machine Learning: A Survey [1] Some of the expert automated systems were summarized in this paper. Feature selection and prediction, these two are essential for every automated system. By choosing features efficiently, we can achieve better results in predicting heart disease. A Survey on Predicting Heart Disease using Data Mining Techniques [2] The various heart disease prediction techniques are discussed and analysed in this paper. Heart Disease Prediction Using Machine using Machine Learning Algorithm [3] Accuracy of the algorithms in machine learning depends upon the dataset that used for training and testing purpose.

Heart disease is a significant issue, so there is a need for diagnosis or prediction of heart disease there are several methods to diagnose heart disease among them Angiography is the trending method which is used by most of the physicians across the world. The conventional methods may give erroneous results because these conventional methods are performed by humans. To avoid these errors and to achieve better and faster results, we need an automated system. The proposed system is meant for the same. The algorithm which is proposed in the project is support vector machine.

### 3. SYSTEM ANALYSIS

#### 3.1 Analysis of Dataset

##### 3.1.1 About the Dataset

The dataset used for Heart disease prediction. The dataset is “framingham.csv”. This dataset consists of various attributes like gender, age, current smoker, BP meds, heart rate...etc. The dataset provides the patients information. It includes over 4,000 records and 16 attributes. Variables each attribute is a potential risk factor. There are both demographic, behavioral and medical risk factors.

Source	Description
<a href="https://www.kaggle.com/datasets/captainozlem/framingham-chd-preprocessed-data">https://www.kaggle.com/datasets/captainozlem/framingham-chd-preprocessed-data</a>	The dataset is “framingham.csv”. The dataset is downloaded from reliable source kaggle.com. It contains sufficient parameters to predict the model.

Table 1: About the Dataset

### 3.1.2 Explore the Dataset

The dataset contains 16 features. The features are gender, age, education, current smoker, cigs per day, BP meds, prevalent stroke, prevalent Hyp, diabetes, tot chol, sys BP, dia BP, BMI, heart rate, glucose and 10 year risk of coronary heart disease (CHD). It contains more than 1000 rows.

Attribute	Type	Description
Gender	Numerical	Gender of the patient.
Age	Numerical	Age of the patient.
Education	Numerical	No further information provided.
Current Smoker	Numerical	Whether or not the patient is a current smoker.
Cigs Per Day	Numerical	The number of cigarettes that the person smoked on average in one day.
BP Meds	Numerical	Whether or not the patient was on blood pressure medication.
Prevalent Stroke	Numerical	Whether or not the patient had previously had a stroke.
Prevalent Hyp	Numerical	Whether or not the patient was hypertensive.
Diabetes	Numerical	Whether or not the patient had diabetes.
Tot Chol	Numerical	Total cholesterol level of patient.
Sys BP	Numerical	Systolic blood pressure of patient.
Dia BP	Numerical	Diastolic blood pressure of patient.
BMI	Numerical	Body Mass Index of patient.
Heart Rate	Numerical	Heart rate of patient.
Glucose	Numerical	Glucose level of patient.
10 year risk of coronary heart disease (CHD)	Numerical	Binary: “1”, means “Yes”, “0” means “No”

Table 2: Description of dataset

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	gender	age	education	currentSm	cigsPerDay	BPMeds	prevalentS	prevalentH	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
2	1	39	4	0	0	0	0	0	0	195	106	70	26.97	80	77	0
3	0	46	2	0	0	0	0	0	0	250	121	81	28.73	95	76	0
4	1	48	1	1	20	0	0	0	0	245	127.5	80	25.34	75	70	0
5	0	61	3	1	30	0	0	1	0	225	150	95	28.58	65	103	1
6	0	46	3	1	23	0	0	0	0	285	130	84	23.1	85	85	0
7	0	43	2	0	0	0	0	1	0	228	180	110	30.3	77	99	0
8	0	63	1	0	0	0	0	0	0	205	138	71	33.11	60	85	1
9	0	45	2	1	20	0	0	0	0	313	100	71	21.68	79	78	0
10	1	52	1	0	0	0	0	1	0	260	141.5	89	26.36	76	79	0
11	1	43	1	1	30	0	0	1	0	225	162	107	23.61	93	88	0
12	0	50	1	0	0	0	0	0	0	254	133	76	22.91	75	76	0
13	0	43	2	0	0	0	0	0	0	247	131	88	27.64	72	61	0
14	1	46	1	1	15	0	0	1	0	294	142	94	26.31	98	64	0
15	0	41	3	0	0	1	0	1	0	332	124	88	31.31	65	84	0
16	0	39	2	1	9	0	0	0	0	226	114	64	22.35	85	NA	0
17	0	38	2	1	20	0	0	1	0	221	140	90	21.35	95	70	1
18	1	48	3	1	10	0	0	1	0	232	138	90	22.37	64	72	0
19	0	46	2	1	20	0	0	0	0	291	112	78	23.38	80	89	1
20	0	38	2	1	5	0	0	0	0	195	122	84.5	23.24	75	78	0
21	1	41	2	0	0	0	0	0	0	195	139	88	26.88	85	65	0
22	0	42	2	1	30	0	0	0	0	190	108	70.5	21.59	72	85	0
23	0	43	1	0	0	0	0	0	0	185	123.5	77.5	29.89	70	NA	0
24	0	52	1	0	0	0	0	0	0	234	148	78	34.17	70	113	0
25	0	52	3	1	20	0	0	0	0	215	132	82	25.11	71	75	0
26	1	44	2	1	30	0	0	1	0	270	137.5	90	21.96	75	83	0
27	1	47	4	1	20	0	0	0	0	294	102	68	24.18	62	66	1
28	0	60	1	0	0	0	0	0	0	260	110	72.5	26.59	65	NA	0
29	1	35	2	1	20	0	0	1	0	225	132	91	26.09	73	83	0
30	0	61	3	0	0	0	0	1	0	272	182	121	32.8	85	65	1

Table 3: Dataset



## 3.2 Data Preprocessing

### 3.2.1 Data Cleaning

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled. If data is incorrect, outcomes and algorithms are unreliable, even though they may look correct. There is no one absolute way to prescribe the exact steps in the data cleaning process because the processes will vary from dataset to dataset. But it is crucial to establish a template for your data cleaning process so you know you are doing it the right way every time.

In this project, the dataset contains missing values

- Handling missing values

The dataset contains missing values in 5 attributes 'Education', 'BPMed', 'glucose', 'totChol' and 'CigsPerDay'. The missing values are represented by a not available (N/A). Here mean value is taken to fill the missing value. First we replace the not available (N/A) using 0. Then replace 0 using mean value.

```
[ ] df.isna().sum()
null = df[df.isna().any(axis=1)]
null
```

```
[ ] import numpy as np
df['BPMed'] = df['BPMed'].replace(np.NaN, df['BPMed'].mean())
```

```
[ ] df[4000:4237]
```

	gender	age	education	currentSmoker	cigsPerDay	BPMed	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
4000	1	46	4.0	1	20.0	0.000000	0	0	0	200.0	110.0	72.0	28.61	70.0	75.0	0
4001	0	58	1.0	0	0.0	0.000000	0	1	0	385.0	165.0	95.0	41.66	82.0	91.0	0
4002	0	46	3.0	0	0.0	0.000000	0	0	0	277.0	122.5	77.5	27.42	63.0	77.0	0
4003	0	53	1.0	1	10.0	0.000000	0	0	0	366.0	116.0	83.0	27.87	68.0	NaN	0
4004	1	39	1.0	1	20.0	0.000000	0	0	0	186.0	126.0	67.0	22.04	63.0	72.0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4232	1	68	1.0	0	0.0	0.000000	0	1	0	176.0	168.0	97.0	23.14	60.0	79.0	1
4233	1	50	1.0	1	1.0	0.000000	0	1	0	313.0	179.0	92.0	25.97	66.0	86.0	1
4234	1	51	3.0	1	43.0	0.000000	0	0	0	207.0	126.5	80.0	19.71	65.0	68.0	0
4235	0	48	2.0	1	20.0	0.029615	0	0	0	248.0	131.0	72.0	22.00	84.0	86.0	0

Figure 1: Handling Missing Values

### 3.2.2 Analysis of Feature Variable

```
[ ] df.dtypes

gender          int64
age             int64
education       float64
currentSmoker   int64
cigsPerDay      float64
BPMeds          float64
prevalentStroke int64
prevalentHyp    int64
diabetes        int64
totChol         float64
sysBP           float64
diaBP           float64
BMI             float64
heartRate       float64
glucose         float64
TenYearCHD      int64
dtype: object
```

Figure 2: Prediction Features

This dataset consists of various attributes like gender, age, current smoker, BP meds, heart rate...etc. The dataset provides the patients information. It includes over 4,000 records and 16 attributes. Variables each attribute is a potential risk factor. There are demographic, behavioral and medical risk factors.

### 3.2.3 Analysis of Class Variable

The first dataset contains TenYearCHD is the class variable. The class variable contains mainly 2 values, they are 0 and 1.

## 3.2 Data Visualizations

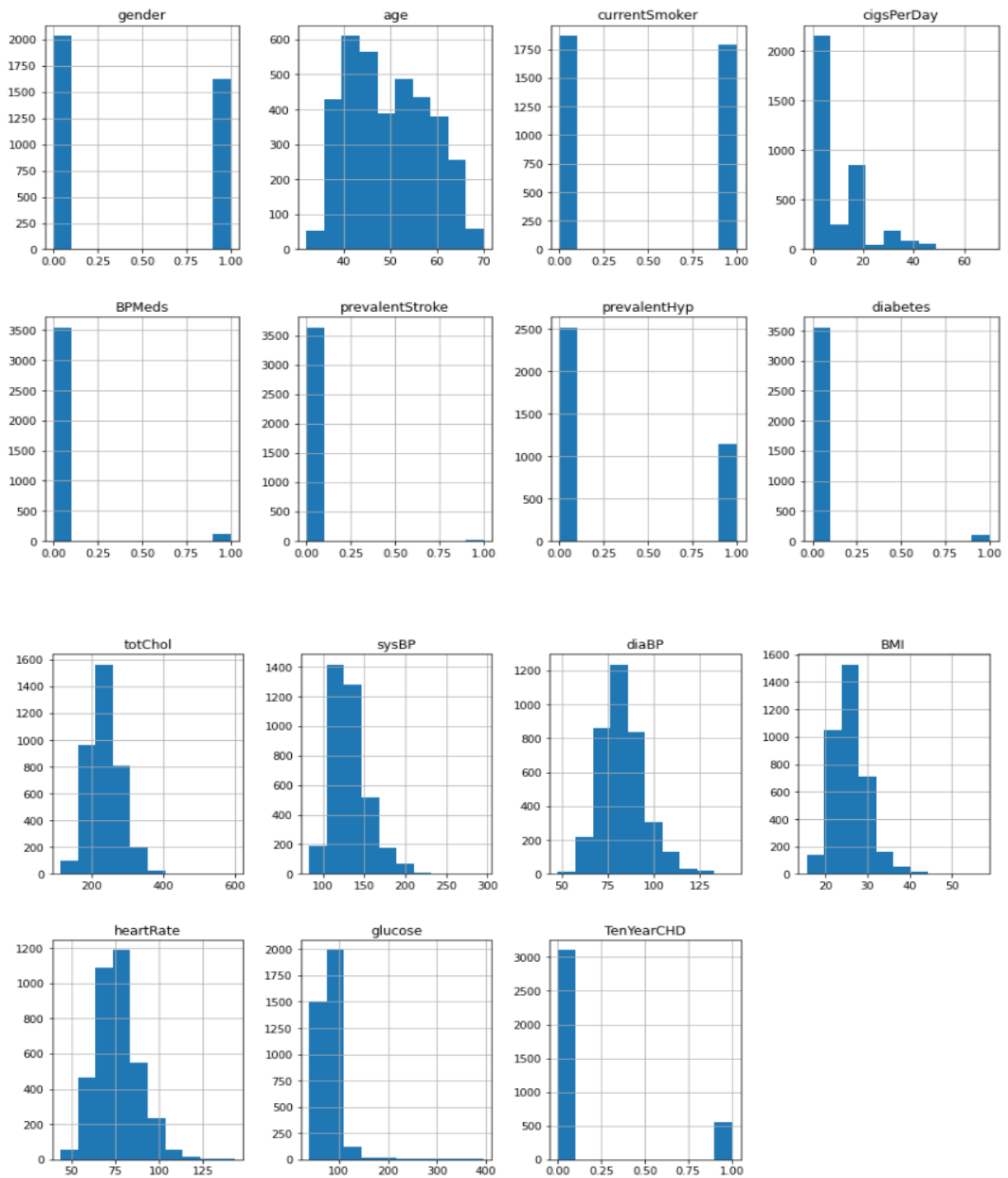


Figure 3: Data Visualization

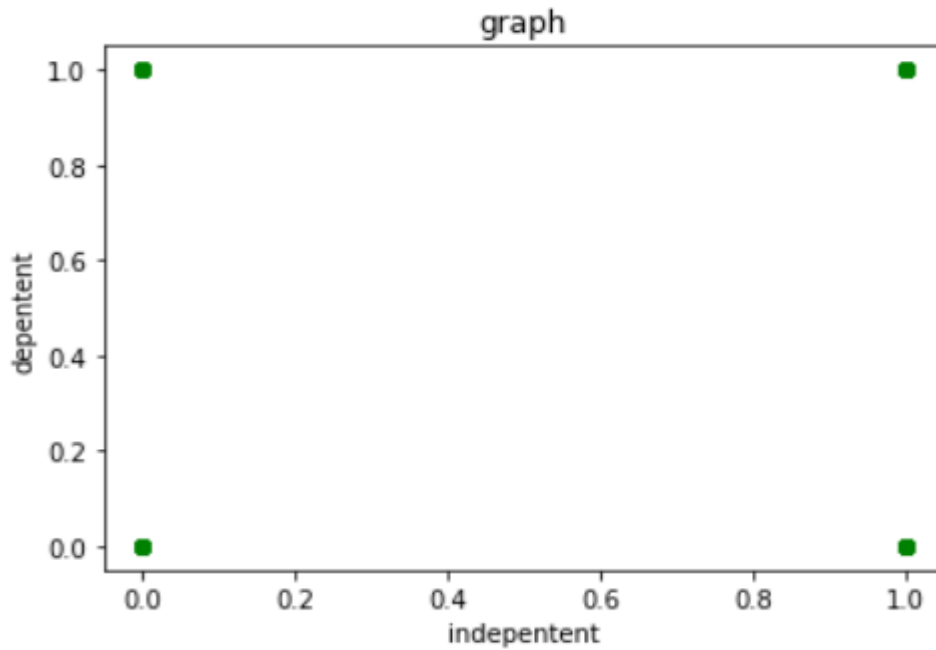


Figure 4: Gender v/s TenYearCHD

From the above graph, we can identify that both female and male go through any of the label.

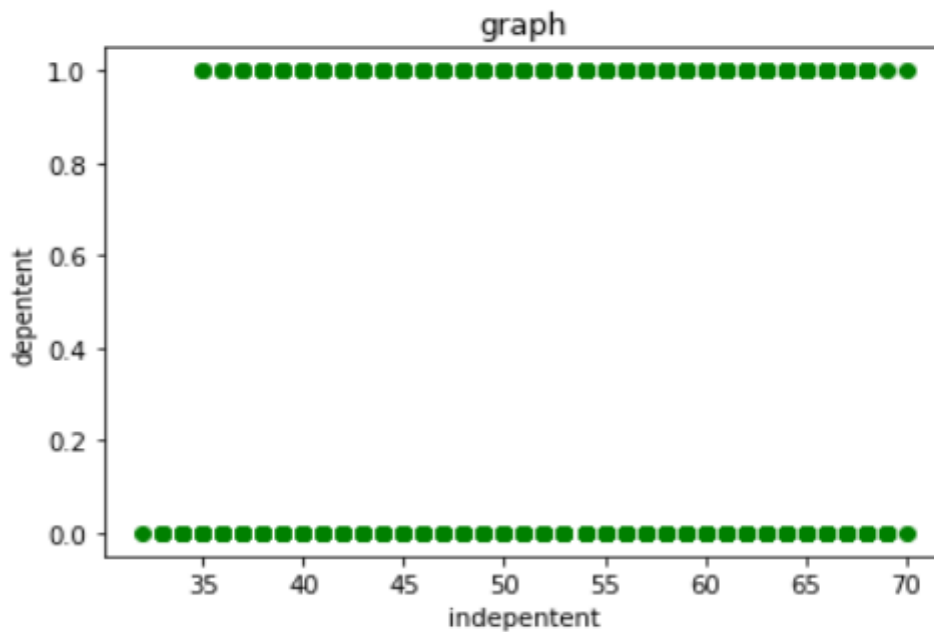


Figure 5: Age v/s TenYearCHD

From this, we can identify that within the range 0-35 will not be affected in the class of no-heart disease. 35-70 can have heart disease. So, age is an important factor.

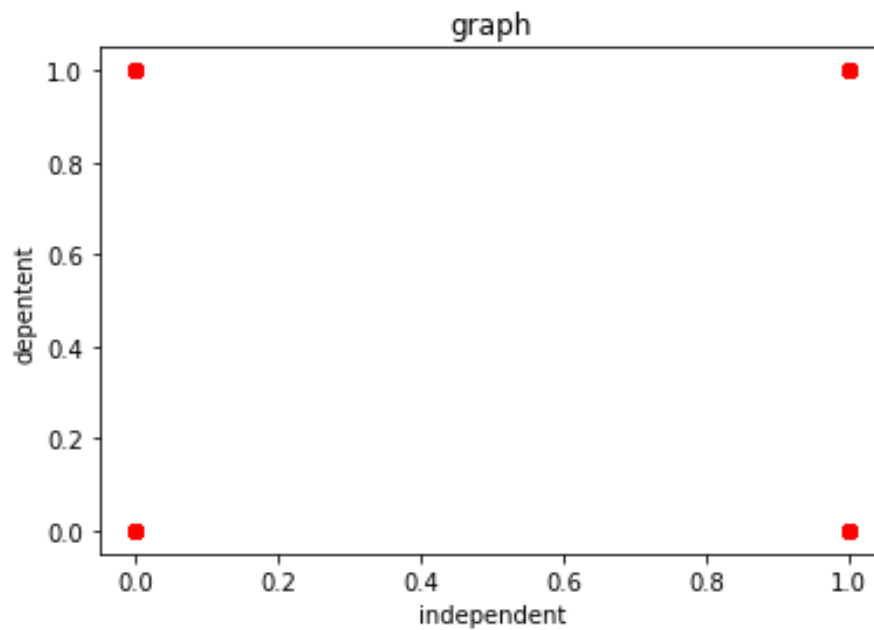


Figure 6: CurrentSmoker v/s TenYearCHD

From this, we can identify that CurrentSmoker go through any of the label.

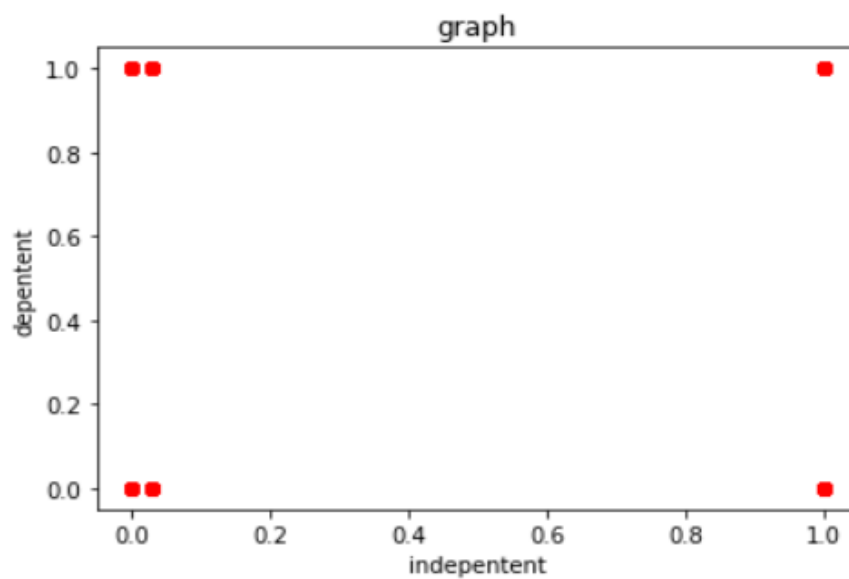


Figure 7: BPMeds v/s TenYearCHD

From this, we can identify that BPMeds go through any of the label.

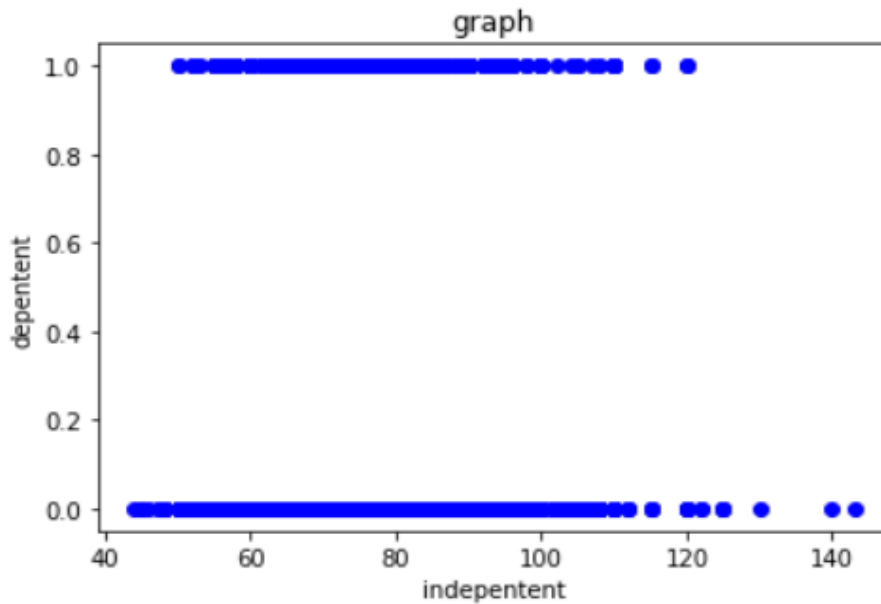


Figure 8: Heart Rate v/s TenYearCHD

From this, we can identify that within the range 40-50 will not be affected in the class of no-heart disease. 50-130 can have heart disease. So, heart rate is an important factor.

### 3.4 Analysis of Algorithm

The proposed model predicts Heart disease. The machine learning approach is used for heart disease prediction. Machine learning (ML) is defined as the study of computer programs that leverage algorithms and statistical models to learn through inference and patterns without being explicitly programmed. ML algorithms learn over experience and improve automatically. It finds techniques, trains models, and uses the learned approach to determine the output automatically. Machine learning systems can also adjust themselves to a changing environment.

Various standard machine learning algorithms are used to predict Heart Disease. Among the selected algorithms, the Support Vector Machine provided the best accuracy. Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a

hyperplane. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.

Support vector machines (SVMs) are powerful yet flexible supervised machine learning algorithms which are used both for classification and regression. But generally, they are used in classification problems. In the 1960s, SVMs were first introduced but later they got refined in 1990. SVMs have their unique way of implementation as compared to other machine learning algorithms. Lately, they are extremely popular because of their ability to handle multiple continuous and categorical variables.

The followings are important concepts in SVM:

**Support Vectors** - Data Points that are closest to the hyperplane are called support vectors. Separating line will be defined with the help of these data points.

**Hyperplane** - As we can see in the above diagram, it is a decision plane or space which is divided between a set of objects having different classes.

**Margin** - It may be defined as the gap between two lines on the closest data points of different classes. It can be calculated as the perpendicular distance from the line to the 12 support vectors. Large margin is considered as a good margin and small margin is considered as a bad margin.

Types of SVM:

SVM can be of two types:

- **Linear SVM:** Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier.
- **Non-linear SVM:** Non-Linear SVM is used for non-linearly separated data, which means if a dataset cannot be classified by using a straight line, then such data is termed as non-linear data and classifier used is called as Non-linear SVM classifier.

The objective of the support vector machine algorithm is to find a hyperplane in an Ndimensional space (N - the number of features) that distinctly classifies the data points.

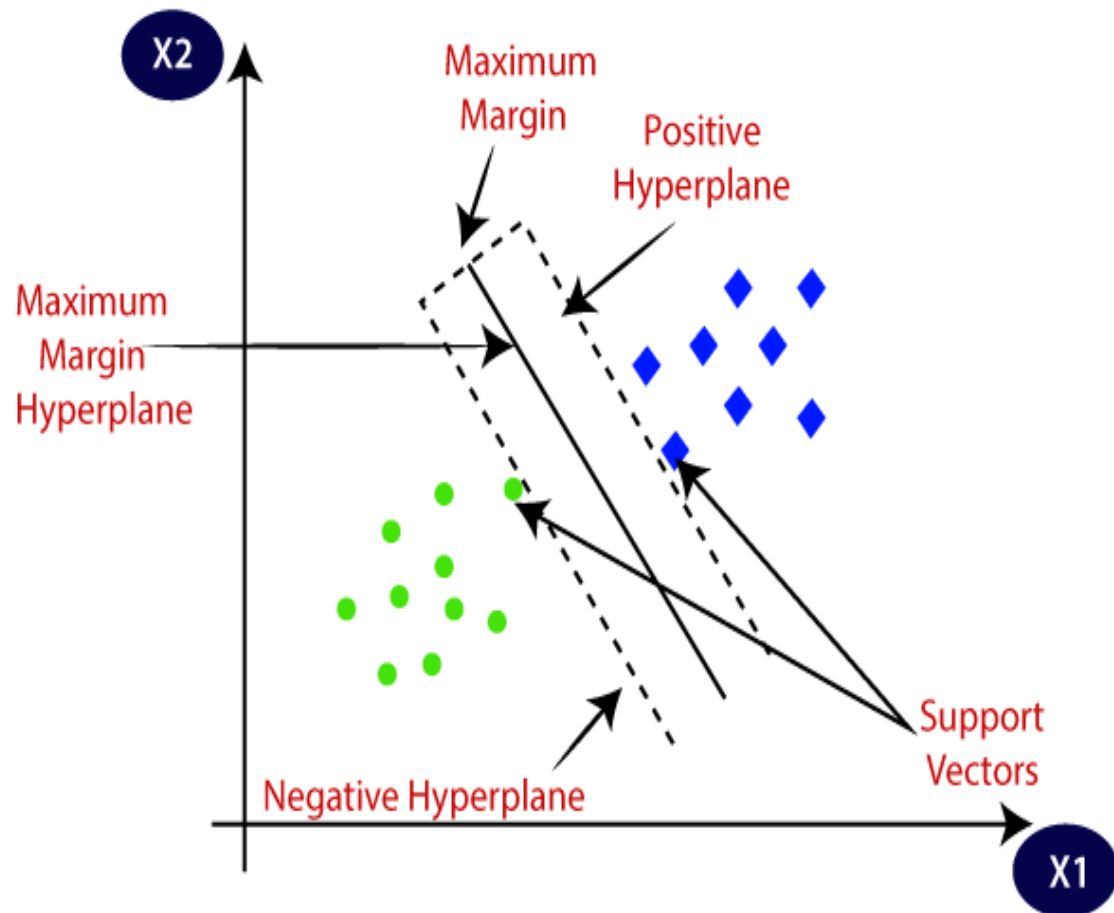


Figure 9: SVM



## Flowchart

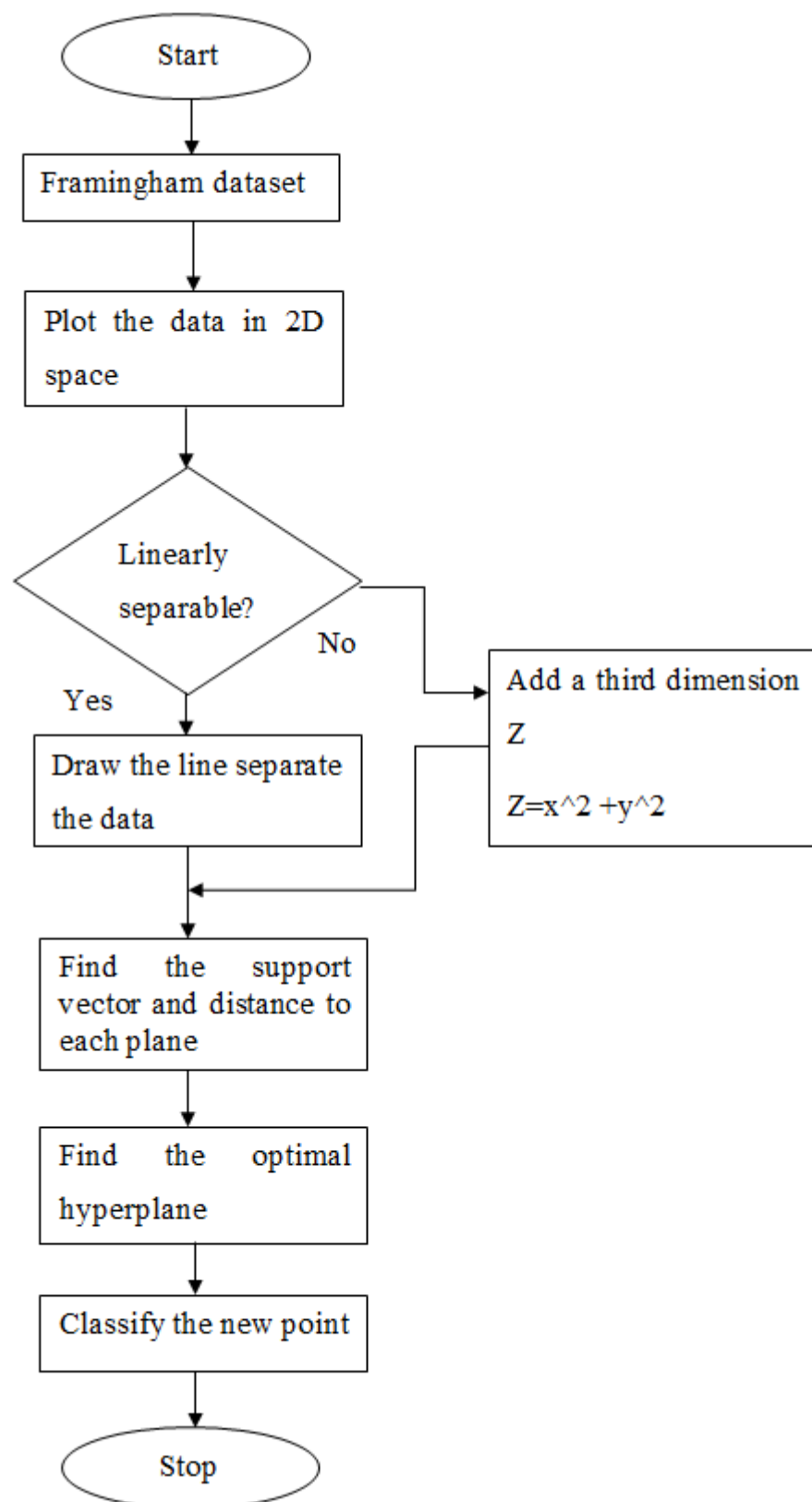


Figure 10: Flowchart

## **Features and characteristics about the algorithm**

- SVM is directed study model that classifies by separating the objects using a hyperplane.
- It can be used for both classification and regression.
- The hyperplanes are drawn with the help of the margins.
- The main goal is to maximize the distance between the hyperplane and the margin.
- The main advantage of SVM is that it can distinguish linear and nonlinear objects.

### **The advantages of support vector machines are:**

- Effective in high dimensional spaces.
- Still effective in cases where the number of dimensions is greater than the number of samples.
- Uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.
- Versatile: different kernel functions can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels.

### **The disadvantages of support vector machines include:**

- If the number of features is much greater than the number of samples, avoid over-fitting in choosing Kernel functions and regularization term is crucial. SVMs do not directly provide probability estimates, these are calculated using an expensive five-fold cross-validation.

### 3.5 Project Pipeline

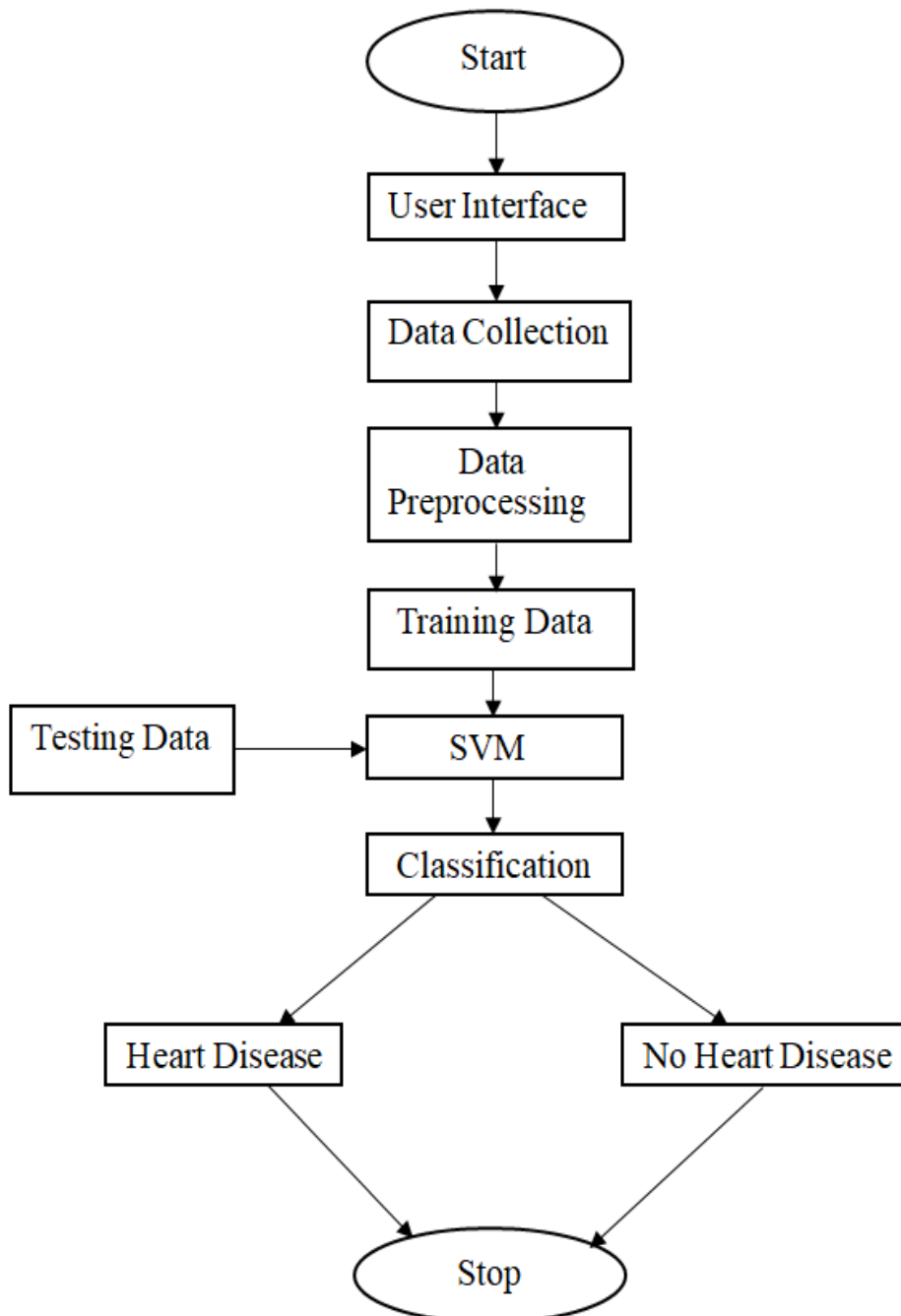


Figure 11: Project pipeline

Here in this project, there are mainly 2 purposes Heart Disease prediction and find the stage of Heart Disease.

First prediction is based on machine learning. For this, we need a dataset then the dataset is split into training data and test data. In preprocessing step, the training data goes through the data preprocessing methods for clearing the missing data and for remaining consistent. We also build a model using the SVM algorithm and then the preprocessed training data and the test data are given to the model for prediction. Using machine learning technology, the model predicts the suitable result.

### 3.6 Feasibility Analysis

Feasibility study is made to see if the project on completion will serve the purpose of the organization for the amount of work, effort and the time that spend on it. Feasibility study lets the developer foresee the future of the project and the usefulness. A feasibility study of a system proposal is according to its workability, which is the impact on the organization, ability to meet their user needs and effective use of resources. Thus when a new application is proposed it normally goes through a feasibility study before it is approved for development. There are three aspects in the feasibility study portion of the preliminary investigation

- Technical Feasibility
- Economic Feasibility
- Operational Feasibility

#### ➤ **Technical Feasibility**

Technical study is a study of hardware and software requirements. All the technical issue related to the proposed system is dealt during feasibility stage of preliminary investigation. The web application is suitable for different web browsers, because which is platform independent.

Data keeping capacity of the proposed equipment to be used for the system are enough. There is no need to develop any hardware to use this system. The heart disease prediction is systems that run on windows operating system which is available in most of the systems these days. So there is no need to install any bulk software for using this application. So there is no need for using other special equipment's, thus the system is technical feasible and non-intrusive.

#### ➤ **Economic Feasibility**

Economic analysis is the most frequently used method for evaluating the effectiveness of a candidate system. Heart disease prediction system will be cost effective and budgetary constraints, it would be cheap and quick to implement. There isn't any extra requirement of peripheral or software for development of system as it can be completed with the available resource. There is no need of special equipment to use this system. Also doesn't need bulk writing. So it is economically feasible.

### ➤ **Operational Feasibility**

Operational feasibility is the measure of how well a proposed system solves the problems, and takes advantage of the opportunities identified during scope definition and how it satisfies the requirements identified in the requirements analysis phase of system development. Heart Disease prediction is easy to operate because it only uses simple steps to predict the whether a patient has Heart Disease and find stage of heart disease. The application is simple for user. Because there is separate forms for prediction. So there is no any complicated steps to use this system.

### 3.7 System Environment

System environment specifies the hardware and software configuration of the new system. Regardless of how the requirement phase proceeds, it ultimately ends with the software requirement specification. A good SRS contains all the system requirements to a level of detail sufficient to enable designers to design a system that satisfies those requirements. The system specified in the SRS will assist the potential users to determine if the system meets their needs or how the system must be modified to meet their needs.

#### 3.7.1 Software Environment

Front End	: Python
IDE	: Internet Explorer
Operating System	: Windows

##### 1. Python

Python is an interpreted high-level programming language for general-purpose programming. Created by Guido van Rossum and first released in 1991, Python has a design philosophy that emphasizes code readability, notably using significant whitespace. It provides constructs that enable clear programming on both small and large scales. Python features a dynamic type system and automatic memory management. It supports multiple programming paradigms, including object-oriented imperative, functional and procedural, and has a large and comprehensive standard library. Python interpreters are available for many operating systems. It has a wide range of applications from Web development (like: Django and Bottle), scientific and mathematical computing (Orange, SymPy, NumPy) to desktop graphical user Interfaces (Pygame, Panda3D). Python is a widely used high-level programming language for general-purpose programming. Apart from being an open-source programming language, python is a great object-oriented, interpreted, and interactive programming language. Python combines remarkable power with very learn syntax. It has modules, classes, exceptions, very high-level dynamic data types, and dynamic typing. There are interfaces to many systems calls and libraries, as well as to various windowing systems. New built-in modules are easily written in C

or C++ (or other languages, depending on the chosen implementation). Python is also usable as an extension language for applications written in other languages that need easy-to-use scripting or automation interfaces. Python is also usable as an extension language for applications written in other languages that need easy-to-use scripting or automation interfaces. Few simple reasons are:

- It's simple to learn. As compared to C, C++ and Java the syntax is simpler and Python also consists of a lot of code libraries for ease of use
- Though it is slower than some of the other languages, the data handling capacity is great
- Open Source! – Python along with R is gaining momentum and popularity in the Analytics domain since both of these languages are open-source. Capability of interacting with almost all the third-party languages and platforms

## **Python Libraries:**

### **NumPy**

NumPy is the fundamental package for scientific computing with Python. It contains:

- A powerful N-dimensional array object.
- Sophisticated (broadcasting) functions.
- Tools for integrating C/C++ and Fortran code.
- Useful linear algebra, Fourier transform, and random number capabilities

Besides its obvious scientific uses, NumPy can also be used as an efficient multi-dimensional container of generic data. Arbitrary data-types can be defined. This allows NumPy to seamlessly and speedily integrate with a wide variety of databases.

### **Matplotlib**

Matplotlib is an amazing visualization library in Python for 2D plots of arrays. Matplotlib is a multi-platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack. It was introduced by John Hunter in the year 2002.

One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib consists of several plots like line, bar, scatter,



histogram etc. Matplotlib is a plotting library for creating static, animated, and interactive visualizations in Python. Matplotlib can be used in Python scripts, the Python and IPython shell, web application servers, and various graphical user interface toolkits like Tkinter, awxPython, etc

## **PHP**

PHP is an open-source server-side scripting language that many devs use for web development. It is also a general-purpose language that you can use to make lots of projects, including Graphical User Interfaces (GUIs).

## **Tensorflow**

Tensorflow is an open source library developed by Google primarily for deep learning application.

## **Keras**

Keras is a powerful and easy-to-use free open source Python library for developing and evaluating deep learning models. It wraps the efficient numerical computational libraries.

## **Os**

The OS module in python provides functions for interacting with the operating system.

## **2. HTML**

HTML stands for Hypertext Markup Language, and it is the most widely used language to write Web Pages. HTML is a mark-up language for structuring and presenting content for the World Wide Web. HTML5 is the fifth revision of the HTML standard (created in 1990 and standardized as HTML4 as of 1997). Its core aims have been to improve the language with support for the latest multimedia while keeping it easily readable by humans and consistently understood by computers and devices (web browsers, parsers, etc).

- It is a very easy and simple language. It can be easily understood and modified.
- It is very easy to make an effective presentation with HTML because it has a lot of formatting tags.
- It is a markup language, so it provides a flexible way to design web pages along with the text.

- It facilitates programmers to add a link on the web pages (by html anchor tag), so it enhances the interest of browsing of the user.
- It is platform-independent because it can be displayed on any platform like Windows, Linux, and Macintosh, etc.
- It facilitates the programmer to add Graphics, Videos, and Sound to the web pages which makes it more attractive and interactive.
- HTML is a case-insensitive language, which means we can use tags either in lower-case or upper-case.

### **3. CSS**

CSS stands for Cascading Style Sheets. It is a style sheet language which is used to describe the look and formatting of a document written in markup language. It provides an additional feature to HTML. It is generally used with HTML to change the style of web pages and user interface.

### **4. Github**

Git is an open-source version control system that was started by Linus Torvalds the same person who created Linux. Git is similar to other version control systems Subversion, CVS, and Mercurial to name a few. Version control systems keep these revisions straight, storing the modifications in a central repository. This allows developers to easily collaborate, as they can download a new version of the software, make changes, and upload the newest revision. Every developer can see these new changes, download them, and contribute. Git is the preferred version control system of most developers since it has multiple advantages over the other systems available. It stores file changes more efficiently and ensures file integrity better. If you're interested in knowing the details, the Git Basics page has a thorough explanation on how Git works.

The social networking aspect of GitHub is probably its most powerful feature, allowing projects to grow more than just about any of the other features offered. Project revisions can be discussed publicly, so a mass of experts can contribute knowledge and collaborate to advance a project forward.

## 5. Visual Studio Code

Visual Studio Code is a streamlined code editor with support for development operations like debugging, task running, and version control. It aims to provide just the tools a developer needs for a quick code-build-debug cycle and leaves more complex workflows to fuller featured IDEs, such as Visual Studio IDE

### 3.7.2 Hardware Environment

RAM	: 8 GB
Disk space	: 256 GB
Processor	: RYZEN Core i3
Display	: 1920 * 1080

## 4. SYSTEM DESIGN

Despite many solutions that have been recently proposed, any disorder that can lead to disturbing the functionality of the heart is called heart disease. Different factors can raise the risk of heart failure. Medical scientists have classified those factors into two different categories; one of them is risk factors that cannot be changed, and another one is risk factors that can be changed. gender, age comes under risk factors that cannot be changed. High cholesterol, smoking, physical inactivity, high blood pressure all these come under risk factors.

### 4.1 Model Building

#### 4.1.1 Model Planning

The system model is developed using algorithm. The proposed model predicts whether the patient is suffering from Heart disease or not. The dataset provides the patients information. It includes over 4,000 records and 16 attributes. Before splitting into training and testing data, we have to check the data. If the dataset is to be cleaned. It is necessary to clean the dataset. The dataset contains missing values. So we first preprocess the dataset.

Then the dataset is divided into training and testing dataset in certain ratio and set a random state. Then the training data is trained with the algorithm to make prediction.

```
[ ] from sklearn import model_selection  
    X_train, X_test, y_train, y_test = model_selection.train_test_split(X, y, test_size = 0.2, random_state=5)
```

Figure 12: SVM Splitting

### 4.1.2 Training

The data is trained using SVM algorithm. For this we have to import SVM classifier for fitting the model.

```
[ ] from sklearn.svm import SVC
    svm_clf = SVC()
    svm_clf.fit(X_train, y_train)
    y_pred = svm_clf.predict(X_test)
    print(y_pred)
```

Figure 13: SVM Training

### 4.1.3 Testing

Software testing is a critical element of software quality assurance and represents ultimate view of specification, design and code generation. Once the source code has been generated the program should be executed before the customer gets it with the specific intend of finding and removing all errors, test must be conducted systematically and test must be designed using disciplined techniques.

In machine learning model, after training the data with training dataset. Then we have to find the accuracy of the algorithm. Here we have the accuracy of the algorithm by changing the random state, features etc. From the analysis we have find that we have to predict the system using the parameters gender, age, education, current smoker, cigs per day, BP meds, prevalent stroke, prevalent Hyp, diabetes, tot chol, sys BP, dia BP, BMI, heart rate, glucose and 10 year risk of coronary heart disease (CHD). Because they gave more accuracy. We get an accuracy of 85.92% by using this algorithm. Then we have to predict the label using any constant value.

```
[ ] import numpy as np
    data=svm_clf.predict(np.array([[1, 46,1, 20.0, 0.000000, 0, 0, 0, 200.0, 110.0, 72.0, 28.61, 70.0, 75.0 ]]))
    print(data)

[0]
```

Figure 14: Prediction

## 5. RESULTS AND DISCUSSION

In the machine learning model, we just compare to algorithm and find out that SVM is the best algorithm to predict the Heart disease. The we train the various features using the algorithm and find out that gender, age, education, current smoker, cigs per day, BP meds, prevalent stroke, prevalent Hyp, diabetes, tot chol, sys BP, dia BP, BMI, heart rate, glucose and 10 year risk of coronary heart disease (CHD) are the good parameters for predicting the model.

The model has an accuracy of 85.92% using the parameters. The accuracy, confusion matrix and classification report are:

```
[ ] from sklearn.metrics import classification_report, confusion_matrix, accuracy_score
    print("accuracy_score: \n",accuracy_score(y_test,y_pred))
    print("confusion matrix: \n",confusion_matrix(y_test,y_pred))
    print("classification report: \n",classification_report(y_test,y_pred))
```

```
accuracy_score:
 0.8592896174863388
confusion matrix:
[[628  0]
 [103  1]]
classification report:
              precision    recall  f1-score   support

     0       0.86         1.00         0.92         628
     1       1.00         0.01         0.02         104

 accuracy                   0.86         732
```

Figure 15: Evaluation of SVM

## 6. MODEL DEPLOYMENT

After testing, the system “Heart Disease Prediction” is ready for implementation. Implementation is the stage of the project when the theoretical design is turned into a working system. Implementation is the process of bringing a newly developed system or revised into an operational one. The new system and its components are to be tested in a structured and planned manner. There are some challenges faced while implementing the software.

The implementation stage of a project is often very complex and time consuming. This involves careful planning, investigation of the current system and constraints of implementation, and training the operating users in the changeover procedures before the system is set up and running. So, “Heart Disease Prediction” is easy to implement. It would be very easy to run also.

The Heart Disease Prediction system implemented successfully. The system predicts the Heart disease prediction using various test results and predict the stage.



## UI Design

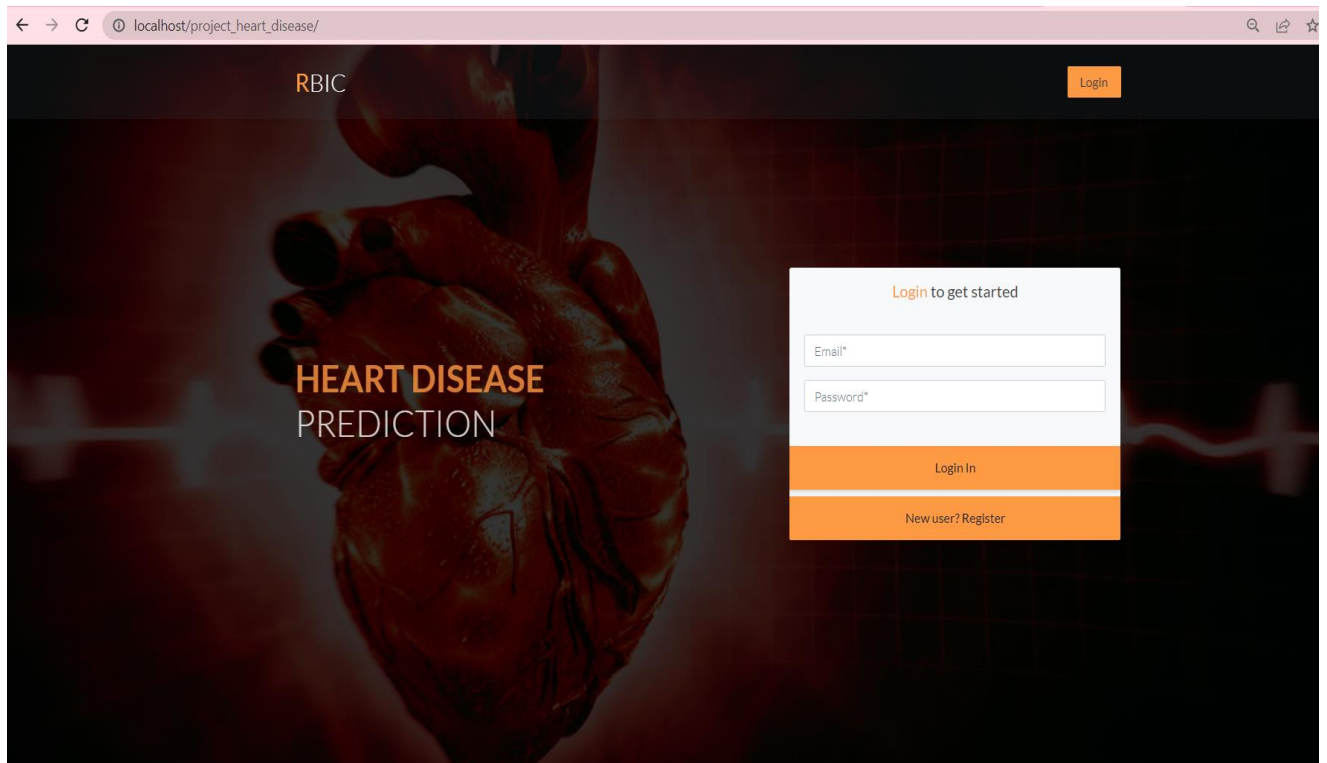


Figure 16: Home page

This is the home page for the system. When we open the system it will leads to the home page and page contain login option.

The screenshot shows a web browser at the URL `localhost/project_heart_disease/home.php?id=3`. The page has a header with the 'RUBIC' logo and 'Patient' and 'Logout' buttons. The main title 'Heart Disease Prediction' is centered. Below it is a form with 14 input fields arranged in two columns. The left column contains: 'Age' (text), 'Gender' (dropdown), 'Current Smoker' (dropdown), 'Cigs Per Day' (text), 'BP Meds' (dropdown), 'Prevalent Stroke' (dropdown), and 'Prevalent Hyp' (dropdown). The right column contains: 'Diabetes' (dropdown), 'Total Cholesterol Level' (text), 'Systolic Blood Pressure' (text), 'Diastolic Blood Pressure' (text), 'Body Mass Index' (text), 'Heart Rate' (text), and 'Glucose Level' (text). At the bottom of the form is a large orange 'Predict' button. The footer is dark gray and contains the copyright notice 'Copyright 2022 © DevCRUD' and a row of social media icons.

Field Name	Field Type
Age	Text
Gender	Dropdown
Current Smoker	Dropdown
Cigs Per Day	Text
BP Meds	Dropdown
Prevalent Stroke	Dropdown
Prevalent Hyp	Dropdown
Diabetes	Dropdown
Total Cholesterol Level	Text
Systolic Blood Pressure	Text
Diastolic Blood Pressure	Text
Body Mass Index	Text
Heart Rate	Text
Glucose Level	Text

Figure 17: Option page

This page contains a form that contains 14 options for entering the test results and a predict button for submit.

The screenshot shows a web browser at the URL `localhost/project_heart_disease/home.php?id=3`. The page has a header with the RUBIC logo and buttons for 'Patient' and 'Logout'. The main title is 'Heart Disease Prediction'. Below it is a form with two columns of input fields. The first column contains: a text field with '62', a dropdown menu with 'Male', a dropdown menu with 'Yes', a text field with '26', a dropdown menu with 'Yes', a dropdown menu with 'Yes', and a dropdown menu with 'Yes'. The second column contains: a dropdown menu with 'Yes', a text field with '244', a text field with '98', a text field with '96', a text field with '57.34', a text field with '80', and a text field with '180'. At the bottom of the form is an orange 'Predict' button. The footer is dark grey and contains the text 'Copyright 2022 © DevCRUD' and social media icons for Facebook, Twitter, Google+, YouTube, Instagram, and LinkedIn.

62	Yes
Male	244
Yes	98
26	96
Yes	57.34
Yes	80
Yes	180

Predict

Copyright 2022 © DevCRUD

f t g+ y i in

Figure 18: Page for Entering Values

The user is asked to provide details like age, gender, whether he is a current smoker or not, number cigarettes person smokes per day, blood pressure details, details of prevalent stroke and hyper tension, whether the person is diabetic or not, total cholesterol level, systolic and diastolic blood pressure value, body mass index, heart rate and blood glucose level.

localhost/project\_heart\_disease/home3.php

RUBIC Result Logout

### Heart Disease Prediction

Age	Diabetes
Gender	Total Cholesterol Level
Current Smoker	Systolic Blood Pressure
Cigs Per Day	Diastolic Blood Pressure
BP Meds	Body Mass Index
Prevalent Stroke	Heart Rate
Prevalent Hyp	Glucose Level

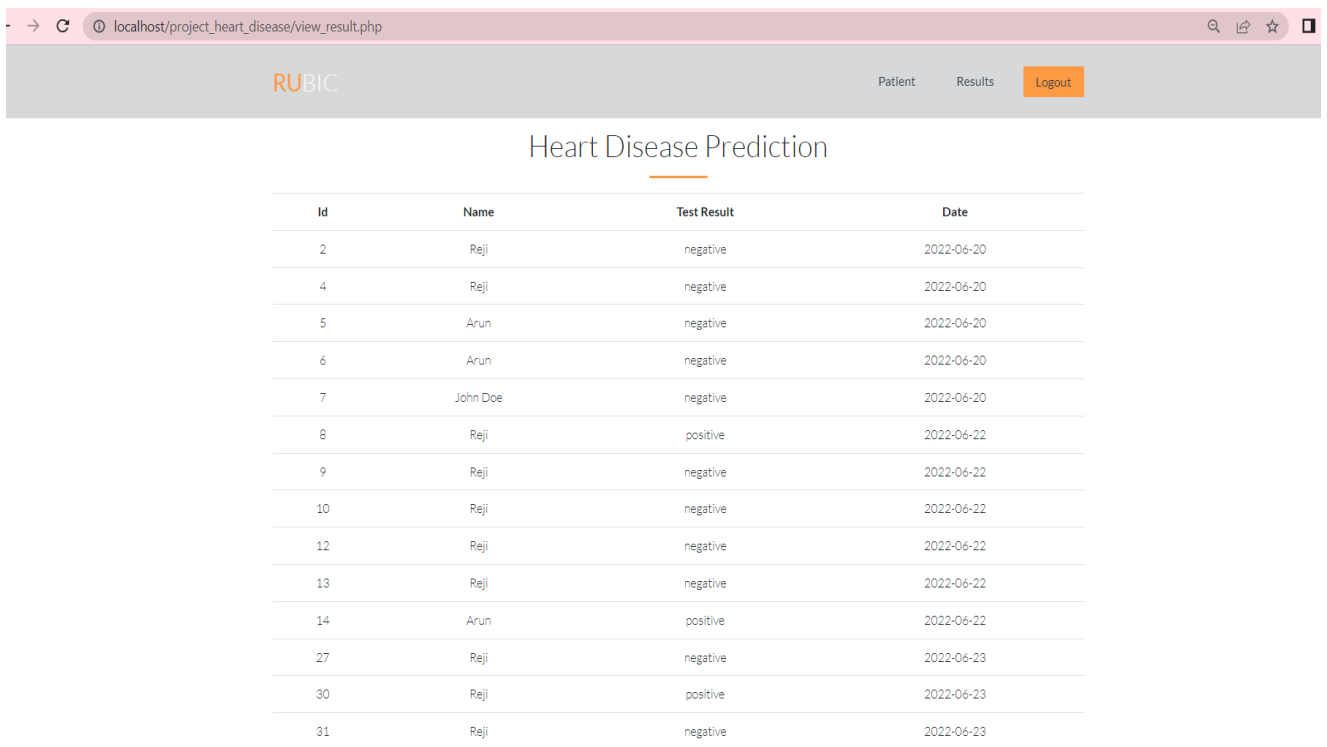
Predict

## OUTPUT

You Have Heart Disease

Figure 19: Output Prediction

Here is also forms that contain an option for uploading features and make prediction. This form is to predict the Heart disease. Using the values entered, the system predicts whether he/she has heart disease. The output will be 'you have heart disease' if the person has heart disease.



Id	Name	Test Result	Date
2	Reji	negative	2022-06-20
4	Reji	negative	2022-06-20
5	Arun	negative	2022-06-20
6	Arun	negative	2022-06-20
7	John Doe	negative	2022-06-20
8	Reji	positive	2022-06-22
9	Reji	negative	2022-06-22
10	Reji	negative	2022-06-22
12	Reji	negative	2022-06-22
13	Reji	negative	2022-06-22
14	Arun	positive	2022-06-22
27	Reji	negative	2022-06-23
30	Reji	positive	2022-06-23
31	Reji	negative	2022-06-23

Figure 20: Prediction Report

The figure shows report of the predictions made by the system.

## 7. GIT HISTORY

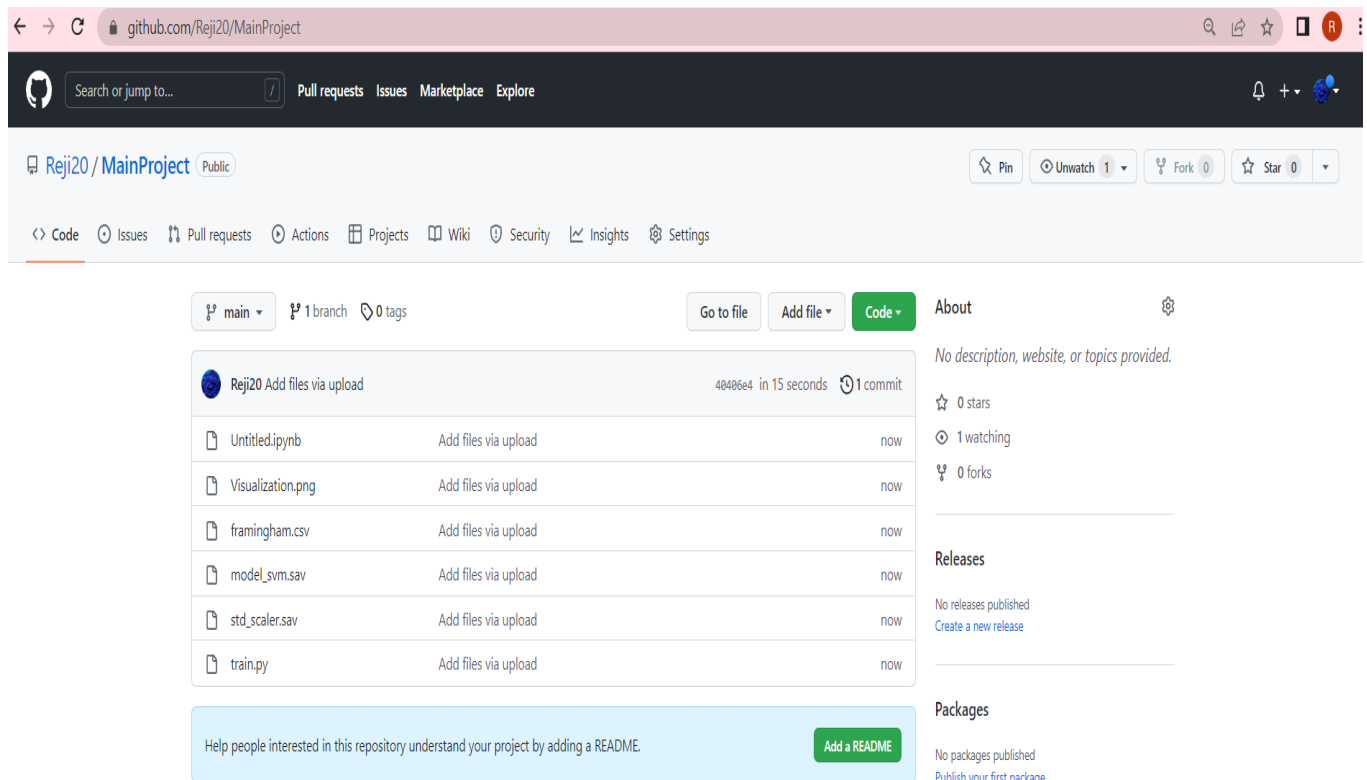


Figure 21: Git History

## 8. CONCLUSION

The heart is one of the main parts of the human body after the brain. The primary function of the heart is to pumping blood to the whole body parts. Any disorder that can lead to disturbing the functionality of the heart is called heart disease. Different factors can raise the risk of heart failure. Medical scientists have classified those factors into two different categories; one of them is risk factors that cannot be changed, and another one is risk factors that can be changed.

Heart diseases are a major killer in India and throughout the world, application of promising technology like machine learning to the initial prediction of heart diseases will have a profound impact on society. The early prognosis of heart disease can aid in making decisions on lifestyle changes in high-risk patients and in turn reduce the complications, which can be a great milestone in the field of medicine. The number of people facing heart diseases is on a raise each year. This prompts for its early diagnosis and treatment. The utilization of suitable technology support in this regard can prove to be highly beneficial to the medical fraternity and patients. In this paper, the machine learning algorithm used to measure the performance is SVM and Extreme Gradient Boosting applied on the dataset.

For accurate prediction, machine learning algorithms SVM were implemented and tested on the given datasets . The model gives 85.92% accuracy and predict the results.

## 9. FUTURE WORK

For the Future Scope more machine learning approaches will be used for best analysis of the heart diseases and for earlier prediction of diseases so that the rate of the death cases can be minimized by the awareness about the diseases. To avoid these errors and to achieve better and faster results, we need an automated system. The system contain interface for predicting Heart Disease and predicting output.

The future scope includes Prediction of the diseases specifically like Unstable angina, Cardiomyopathy, Coronary Artery Disease (CAD), Heart Arrhythmias, Heart attack, Heart failure, Valve disease, High blood pressure, Congenital heart conditions, Inherited heart conditions etc.. It can also be developed as a mobile application.



## **10. APPENDIX**

### **10.1 Minimum Software Requirements**

Browser : Chrome or Internet Explorer  
Operating System : Windows

### **10.2 Minimum Hardware Requirements**

Hard disk capacity : 150 GB  
RAM : 4 GB  
Processor : Intel Core i3 preferred  
Display : 1366\*768

## 11. REFERENCES

1. Rahul Katarya, Polipireddy Srinivas, “Predicting Heart Disease at Early Stages using Machine Learning: A Survey”, International Conference on Advanced Computing and Communication System (ICACCS), 2020.
2. Cincy Raju, Philipsey E, Siji Chacko, L Padma Suresh, Deepa Rajan S, “A Survey on Predicting Heart Disease using Data Mining Techniques”, IEEE Conference on Emerging Devices and Smart Systems (ICEDSS), 2018.
3. Archana Singh, Rakesh Kumar, “Heart Disease Prediction Using Machine Learning Algorithms”, International Conference on Electrical and Electronics Engineering (ICE3), 2020.
4. [https://www.tutorialspoint.com/php/php\\_introduction.htm](https://www.tutorialspoint.com/php/php_introduction.htm)
5. [https://www.w3schools.com/python/python\\_intro.asp](https://www.w3schools.com/python/python_intro.asp)