Prediction of Cardiovascular Disease Using Machine Learning Algorithms

Dinesh Kumar G
MTech (Software Engineering), SITE
VIT University
Vellore-632014, India
gadinesh.kumar2014@vit.ac.in

Santhosh Kumar D
MTech (Software Engineering), SITE
VIT University
Vellore-632014, India
santhoshkumar.d2014@vit.ac.in

Arumugaraj K
MTech (Software Engineering), SITE
VIT University
Vellore-632014, India
arumugaraj.k2014@vit.ac.in

Mareeswari V
SITE
VIT University
Vellore-632014, India
vmareeswari@vit.ac.in

***Abstract*—Healthcare is an inevitable task to be done in human life. Cardiovascular disease is a broad category for a range of diseases that are affecting heart and blood vessels. The early methods of forecasting the cardiovascular diseases helped in making decisions about the changes to have occurred in high-risk patients which resulted in the reduction of their risks. The health care industry contains lots of medical data, therefore machine learning algorithms are required to make decisions effectively in the prediction of heart diseases. Recent research has delved into uniting these techniques to provide hybrid machine learning algorithms. In the proposed research, data pre-processing uses techniques like the removal of noisy data, removal of missing data, filling default values if applicable and classification of attributes for prediction and decision making at different levels. The performance of the diagnosis model is obtained by using methods like classification, accuracy, sensitivity and specificity analysis. This project proposes a prediction model to predict whether a people have a heart disease or not and to provide an awareness or diagnosis on that. This is done by comparing the accuracies of applying rules to the individual results of Support Vector Machine, Gradient Boosting, Random forest, Naive Bayes classifier and logistic regression on the dataset taken in a region to present an accurate model of predicting cardiovascular disease.**

Keywords-Prediction, Healthcare, Machine learning Algorithms, Cardiovascular disease

I. INTRODUCTION

Nowadays, healthcare is increasing day by day due to lifestyle, hereditary. It creates a lot of data with time. So, the data generated by the health or the survey are getting wasted. But nowadays as the data analytics come into existence. The hospitals and NGOs are making use of the data to generate the useful information from the data. The modern world has cardiovascular disease as its deadliest enemy. This disease affects a person in such a way so that the patients can't be cured as easily as possible. So, diagnosing patients at the right time is the toughest work in medical field. Misunderstanding and wrong diagnosis made by the hospital leads to the bad reputation. India questions that the treatment for this disease is quite tough and can't is reachable by most of the patients. Everyone has different values for Blood pressure, cholesterol, and pulse rate [19]. But per medically proven results the normal values of Blood pressure are 120/80, cholesterol is and pulse rate is 72.

The world health organization reports suggest that greater than 12 million deaths are happening worldwide due to cardiovascular problems. It is a catastrophic disease in India which originates more calamities. The examination of the unhealthiness is a complex mechanism. It should be measured perfectly and precisely. Because lack of experts at some places is resulting in the patients in a hazardous position. Ordinarily, these are diagnosed by the cardiologists (Who generally treat the heart disease patients). It is extremely beneficial if these techniques are combined with the medical information system. This integration of June data taken by survey requires comparison of different machine learning techniques for finding out their suitability for the said job. This paper suggests the different machine learning techniques that are used for forecasting the uncertainty levels of cardiovascular diseases based on the attributes present. The medical datasets used are taken from the research that had been fascinated throughout the world.

Machine learning is an art of mastering system without being explicitly computed. They are used to analyze the analytical arrangement in high dimensional, diverse data sets like heart diseases [13]. They are used in recognition of the arrangements(patterns) that gives support for forecasting and controlling mechanism for analysis and medication.

II.LITERATURE SURVEY

The data which is recognized can be utilized by the social insurance directors to show signs of improvement administrations. Coronary illness was the most significant explanation behind casualties in the nations like India, United States. Machine learning calculations like a Logistic relapse, irregular woods, angle boosting, and Support vector machine and order calculations, for example, Naive Bayes encounters various types of heart-related issues. These calculations can be utilized to upgrade the information stockpiling for viable and legitimate purposes. [1]

In this paper [2], it is presumed that albeit most analysts are utilizing diverse classifier methods, for example, Neural system, SVM, KNN and twofold discretization with Gain Ratio Decision Tree in the conclusion of coronary illness, applying Naïve Bayes and Decision tree with data pick up counts gives better outcomes in the finding of coronary illness and better exactness when contrasted with different classifiers.

The expectation of cardiovascular illness by methods for a few machine learning calculations is going on [2]. Many research papers have executed different machine learning calculations, for example, Naive Bayes, Random Forest, Gradient boosting, Logical Regression and Support Vector Machine for anticipating cardiovascular illness [1]. In [10], it depicts how these machine learning calculations are utilized to foresee the pneumonia ailment.

In paper [9], it unmistakably tells about the headway of the Support Vector Machines used to look at and to separate learning from extensive informational collections consequently. In [4], Naive Bayes assumes a noteworthy part of the conclusion of ailment. Discovering precision and better execution of calculations are made in the paper [1]. Paper [5] demonstrates the forecast of survival rates and mistake rates. Thus [8] depicts to discover the death rate and exactness among the calculations by finding the death rate. [6] Shows the conceivable outcomes of utilizing the diverse calculation for coronary illness forecast and to give the write about that. By utilizing less number of characteristics, the framework can foresee the coronary illness with cutting-edge advancements and with various machine learning calculations [7].

In the present world, there are numerous logical innovations which help specialists in taking clinical choices however they won't be precise. Coronary illness expectation framework can help therapeutic experts in anticipating the condition of the heart, in view of the clinical information of patients nourished into the framework. [18] Specialists may some of the time neglect to take precise choices while diagnosing the coronary illness of a patient, in this manner coronary illness forecast frameworks which utilize machine learning calculations aid such cases to get exact outcomes. [3] There are many instruments accessible which utilize expectation calculations yet they have a few blemishes. A large portion of the instruments can't deal with huge information and most are not brought together, not conveyed on cloud and consequently not open on the web [12]. There are numerous doctor's facilities and social insurance businesses which gather colossal measures of patient information which ends up plainly hard to deal with as of now existing frameworks. [14]

In this paper, conceivable outcomes of anticipating the hazard levels of patients for a colossal informational collection and sending the application on cloud stage where specialists and patients can sign in with a one of a kind ID made by them. [4] Doctors can transfer the patient reports and patients, then again, can see the reports on their portable workstations or PCs [20]. A gathering of patient reports will be kept up online in the cloud.

III.DATASETS AND DESCRIPTION

Data Source

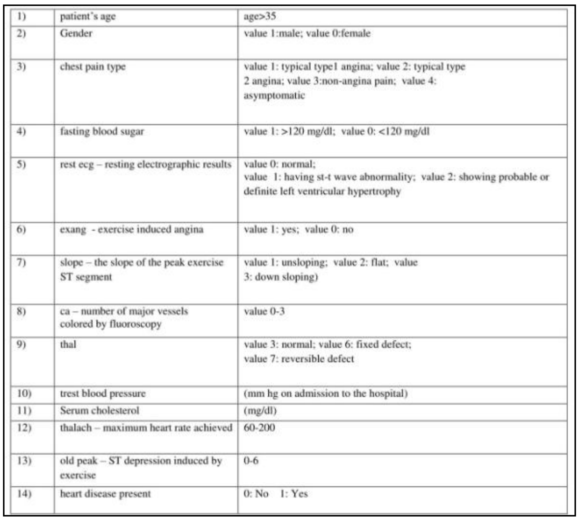
Healthcare databases have collected a significant amount of patient's records. The term heart disease circulates on various conditions which are harmful to the human heart. Cardiovascular disease is one of the deadliest diseases in nature. The term "Cardiovascular Disease" deals with the situation by which the heart and the blood veins are affected and the result by which the blood pumping and circulation of blood takes place in the body. Records were obtained from the Cleveland, Hungarian, Switzerland, Long Beach VA heart disease database (UCI machine Learning Repository). Datasets segregate the patterns related to the disease. The records classify into two datasets: training dataset and testing dataset. Sum of 920 records along with 76 attributes related to the medical was obtained. The following table (Table 1) shows the list of 14 attributes on which the system is working.

Analysis of Data

This phase has the major jobs of performing data pre-processing such as data cleaning, data integration, filling of missing values, removing redundant data as the dataset contains missing values and redundant data. It leads to fault prediction.

Operating Environment

The language R provides a stage for performing statistical computation and graphical representation, especially for data analysis. Due to the collection of packages which helps in statistical computation and graphical representation; a user can make the quick analysis of data and graphical representation which leads to the development of prediction system for given application. R provides an open source software which makes the best compatibility in UNIX and Windows [7] For prediction results, R offers a better outcome compared to other languages. Although heart disease can happen in various structures, there is a typical arrangement of center hazard factors that impact whether somebody will, at last, be in danger of heart disease or not. These are the basic characteristics that ought to be checked to know whether the heart disease will come or not [15].



1)	patient's age	age>35
2)	Gender	value 1:male; value 0:female
3)	chest pain type	value 1: typical type1 angina; value 2: typical type 2 angina; value 3:non-angina pain; value 4: asymptomatic
4)	fasting blood sugar	value 1: >120 mg/dl; value 0: <120 mg/dl
5)	rest ecg – resting electrographic results	value 0: normal; value 1: having st-t wave abnormality; value 2: showing probable or definite left ventricular hypertrophy
6)	exang - exercise induced angina	value 1: yes; value 0: no
7)	slope – the slope of the peak exercise ST segment	value 1: unsloping; value 2: flat; value 3: down sloping)
8)	ca – number of major vessels colored by fluoroscopy	value 0-3
9)	thal	value 3: normal; value 6: fixed defect; value 7: reversible defect
10)	trest blood pressure	(mm hg on admission to the hospital)
11)	Serum cholesterol	(mg/dl)
12)	thalach – maximum heart rate achieved	60-200
13)	old peak – ST depression induced by exercise	0-6
14)	heart disease present	0: No 1: Yes

Table 1: Attributes to be collected

IV.PROPOSED SYSTEM

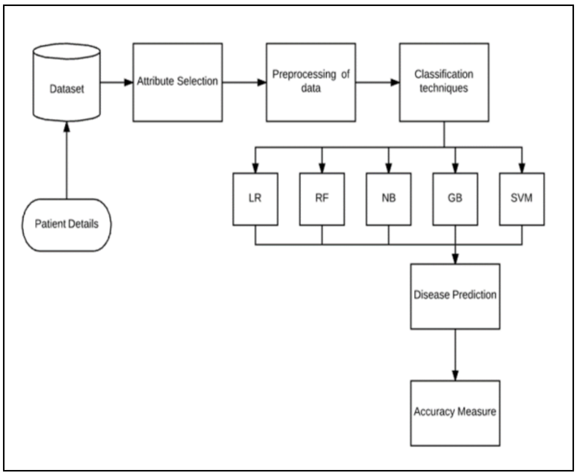


Figure 1: Proposed System

The above figure (figure 1) shows functioning of the system is described step by step

Fig. 1. The dataset contains the details of the patients.

Fig. 2. Attribute selection takes the attributes which are useful for the prediction of the heart disease.

Fig. 3. After identifying the data from the available resources, they are further selected for processing which includes data cleaning, removal of noise i.e. missing data

Fig. 4. Different classification algorithms are performed on the preprocessed data to find the chance of getting heart disease.

Fig. 5.

Fig. 6. It also finds the accuracy of the algorithms and compares the accuracy among all the algorithms.

V. VARIOUS MACHINE LEARNING ALGORITHMS

A. Logistic Regression

Logistic regression is well known for binary classification and it is one of the most efficient machine learning algorithms. Due to its simplicity, which has its application on a wide range of problems and provides suitable solutions [11]. It works on the dependent variable which is categorical. The variables are binary dependent variables such as 0s or 1s, pass or fail etc. If the variables are having more than one outcomes, then multinomial logistic regression or if there are ordered multiple categories then uses ordinal logistic regression. The logistic function as,

$$P = (y = 1|X) = \frac{1}{1 + e^{-wa}}$$

Where e is the numerical constant Euler's number and a is an input we put into the function. The roc curve (figure2) for the logistic regression.

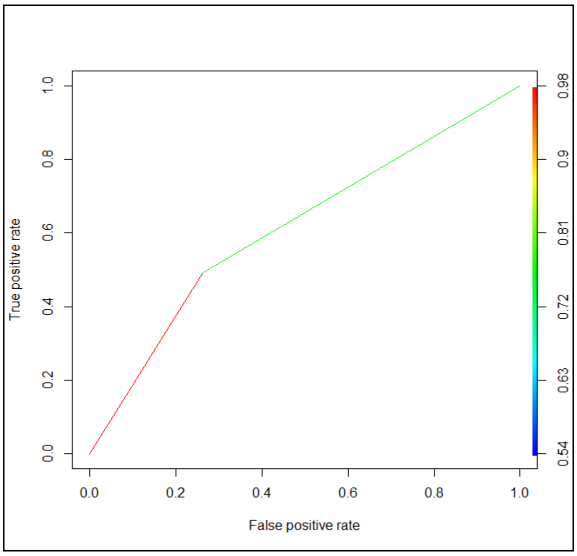


Figure 2: ROC Curve

B. Naive Bayes

Navies Bayes classifier is one of best classification algorithm in machine learning which uses the Bayesian algorithm. Naive Bayes classification algorithm is strongly scalable, which require variables linear in the form of predictor variables in a problem statement [17]. It is similar for classification and regression and makes tough competition with SVM. It identifies the specialty of the patients related to the disease. It shows the probability of each input attribute for the predictable state and provides the probability of event occur. A conditional probability is the likelihood of some conclusion A, given some evidence/observation, B, where a dependence relationship exists between A and B. This probability is denoted as P (A|B) where, P(A) is the probability of event A, P(B) is the probability of event B, P(B|A) is the probability of event B with the condition that event A has taken place.

$$P(A|B) = \frac{(P(B|A) * P(B))}{P(A)}$$

C.Random Forest

Random forest is a machine learning algorithm used for classification and regression. It creates decision trees for each attribute. It corrects the overfitting to their training set. It also avoids the missing values, outliers by following the steps of data analysis, data pre-processing. It is kind of machine learning method where the weak models are combined to form a dynamic model. The random forest tree (figure 3) shows the multiple decision trees that are linked to the system.

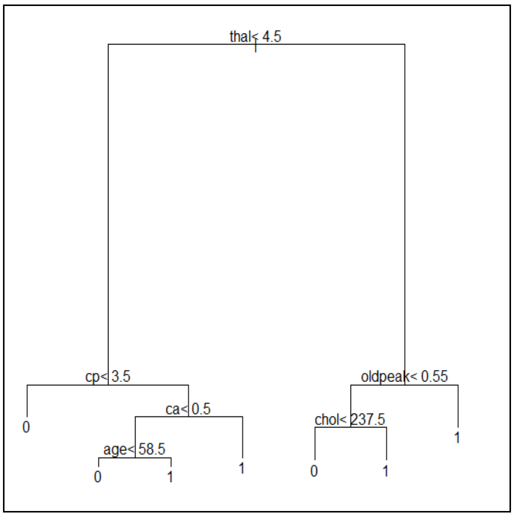


Figure 3: Random forest tree

D.Support Vector Machine

Support Vector Machine is an algorithm which is used in machine learning for classification and regression techniques. It is regularly used as the classification techniques due to its efficiency when compared with the other algorithms. This technique plots a hyperplane for every attribute as a coordinate that is present in the dataset [16]. Classification is performed by identifying the hyperplane that divides one class with the other class. It builds a model which assigns new example to the other, making it a non- probabilistic binary linear classifier.

E. Gradient Boosting

It is a machine learning technique for regression and classification techniques which as a group of weak prediction models that are like that of decision trees. In this, gradient boosting technique it provides a variable importance of the attribute that is related to predict the heart disease in this dataset. The following figure (figure 4) shows the variable importance of heart failure prediction with the help of the boosted tree.

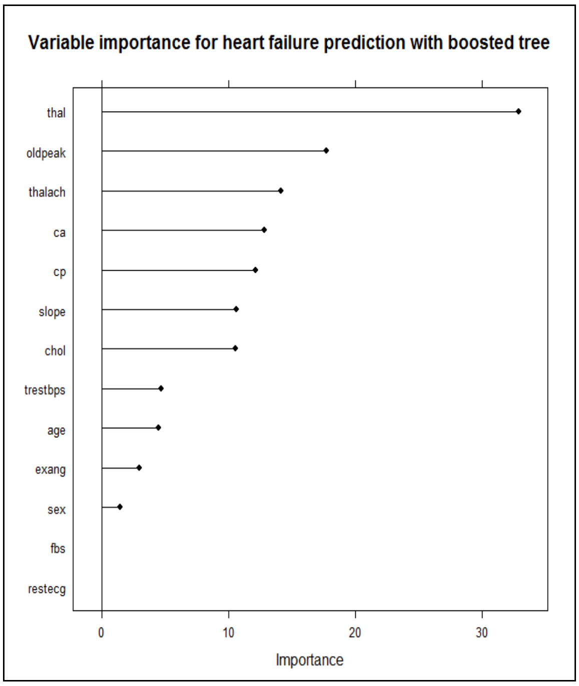


Figure 4: Variable importance graph

F. Accuracy Module

The module predicts the accuracy by using machine learning algorithms. This module takes the maximum accuracy generated by the algorithms which predict the maximum chances of getting a cardiovascular disease. In this, each algorithm provides different accuracy rate for taken attributes which is the cause of the cardiovascular disease. You can calculate the accuracy of your model with:

From Confusion Matrix Specificity and Sensitivity can be derived as illustrated below:

$$\left.\begin{array}{l} True\ Negative\ Rate, specificity = \dfrac{P}{P+Q} \\ False\ Positive\ Rate, 1 - specificity = \dfrac{Q}{Q+P} \end{array}\right\} sum\ to\ 1$$

$$\left.\begin{array}{l} True\ Positive\ Rate, sensitivity = \frac{R}{R+S} \\ False\ Negative\ Rate = \frac{S}{S+R} \end{array}\right\} sum\ to\ 1$$

Where, P=Number of true negatives, Q=Number of false positives and R=Number of true positives, S=Number of false negatives.

Sensitivity is the approach that identify the people those with the cardiovascular disease (true positive rate) and specificity is the approach that identify the people those without the cardiovascular disease (true negative rate).

VI.DATA VISUALISATION

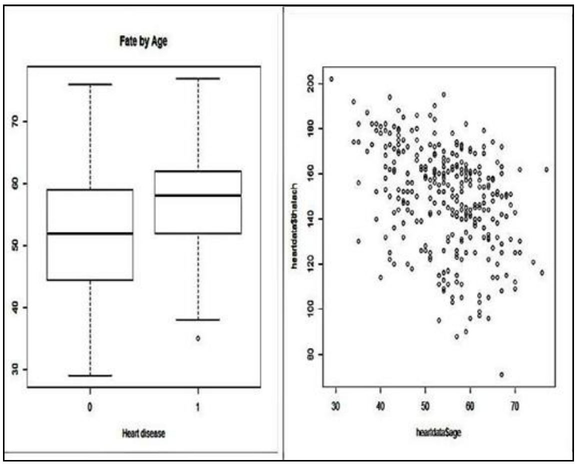


Figure 5: Data Visualisation based on attributes

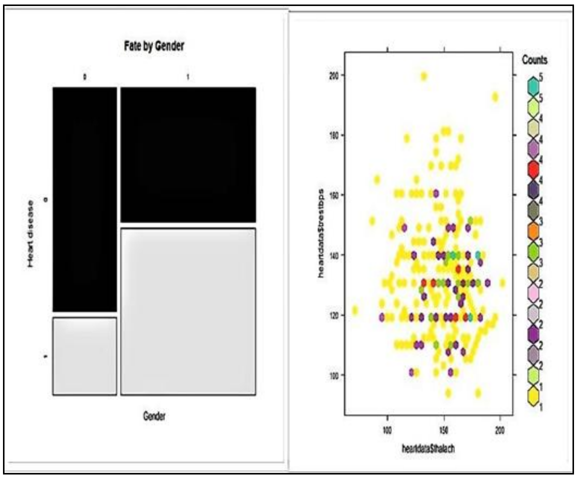


Figure6: Data visualizations based on attributes

The data visualization in this system expresses a powerful visualization about the data (14 attributes) that contribute to the prediction. The above figure (figure 5 and figure 6) shows the boxplot, mosaic plot, scatterplot visualizations of all the attribute relationships and their nature.

VI.RESULTS AND DISCUSSION

In this section, the outputs and the accuracies generated are reviewed and the results are displayed. In this mostly the algorithm which has the highest accuracy gives the more accurate result. The following table (Table 2) shows the accuracy generated by each algorithm as follows:

Table 2: Comparison of Accuracies

ALGORITHM NAME	ACCURACY	OVERALL ACCURACY
Logistic regression	0.9161585	0.8651685
Random forest	0.8953252	0.8089888
Naïve Bayes	0.9095528	0.8426966
Gradient boosting	0.9070122	0.8426875
SVM	0. .882622	0.7977528

The accuracy arises from the attributes collected from the datasets. These accuracies can also be increased with the help of better and datasets and better computational systems. Based on the accuracy produced by the algorithms the best algorithm which gives accuracy is chosen for finding the prediction. The following figure (figure 7), shows the comparison of the performance of all algorithms.

The user interface is designed in such a way to use effectively by the user. The design of the user interface for the Cardiovascular disease prediction system follows below.

The following figure (figure 8) shows the probability of heart disease for a patient based on the algorithms. Here the 14 attributes that get to the patient produces the appropriate result that is predicted with the help of algorithms. The following figure 9, shows the medical range value for the probabilistic value, and it has live theme selector where the theme will be changed at any time. There are nearly 15 themes that are associated with the help of shiny.

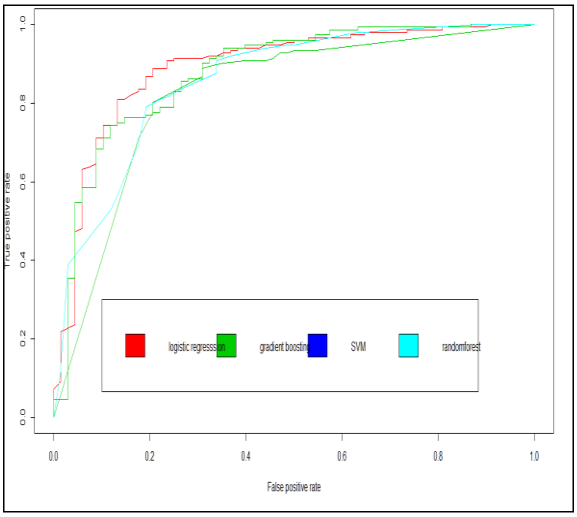
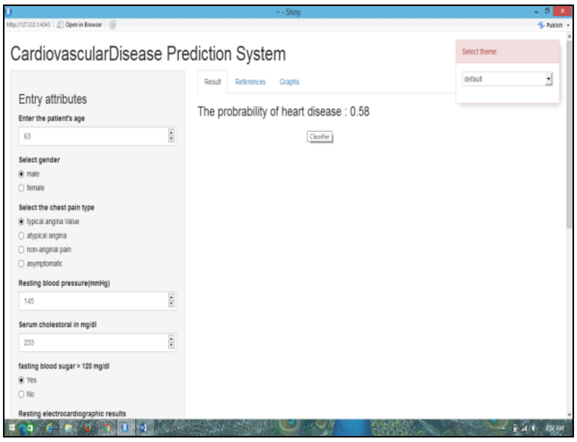


Figure 7: Comparison of performance of algorithms



CardiovascularDisease Prediction System

Select theme: default

Entry attributes

Enter the patient's age: 63

Select gender
● male
○ female

Select the chest pain type
● typical angina Value
○ atypical angina
○ non-anginal pain
○ asymptomatic

Resting blood pressure(mmHg): 145

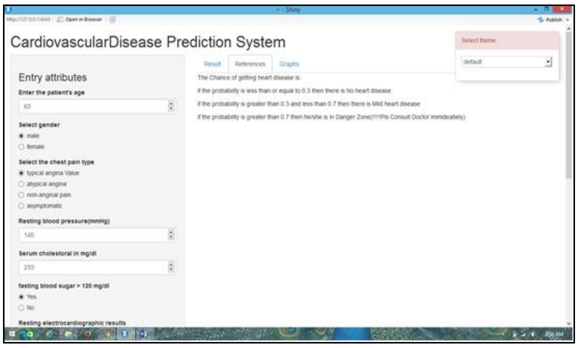
Serum cholestoral in mg/dl: 233

fasting blood sugar > 120 mg/dl
● Yes
○ No

Resting electrocardiographic results

The probrability of heart disease : 0.58

Figure 8: Front-end view of the system



CardiovascularDisease Prediction System

Select theme: default

Entry attributes

Enter the patient's age: 63

Select gender
● male
○ female

Select the chest pain type
● typical angina Value
○ atypical angina
○ non-anginal pain
○ asymptomatic

Resting blood pressure(mmHg): 145

Serum cholestoral in mg/dl: 233

fasting blood sugar > 120 mg/dl
● Yes
○ No

Resting electrocardiographic results

The Chance of getting heart disease is:

if the probability is less than or equal to 0.3 then there is No heart disease

if the probability is greater than 0.3 and less than 0.7 then there is Mild heart disease

if the probability is greater than 0.7 then he/she is in Danger Zone(!!!!Pls Consult Doctor immideately)

Figure 9: Results and Discussion

VII.CONCLUSION

This paper contributes the correlative application and analysis of distinct machine learning algorithms in the R software which gives an immediate mechanism for the user to use the machine learning algorithms in R software for forecasting the cardiovascular diseases. This is non-ethical study aims to use available machine learning techniques in R software. Future work includes different ensemble methods of these algorithms which can advance to better performance with more parameter settings for these algorithms.

REFERENCES

[1]. Jaymin Patel, Prof.Tejal Upadhyay, Dr.Samir Patel "Heart disease prediction using Machine learning and Data Mining Technique" Volume 7.Number1 Sept 2015-March 2016.

[2]. G.Parthiban, S.K.Srivasta "Applying Machine learning methods in Diagnosing Heart disease for Diabetic Patients" International Journal of Applied Information Systems (IJAIS) – ISSN: 2249-0868 Foundation of Computer Science FCS, New York, USA Volume 3– No.7, August 2012

[3]. Thenmozhi.K and Deepika.P, Heart Disease Prediction using classification with different decision tree techniques. International Journal of Engineering Research & General Science, Vol 2(6), pp 6-11, Oct 2014.

[4]. Igor Kononenko "Machine learning for medical diagnosis: history, state of art& perspective" Elsevier - Artificial intelligence in Medicine, Volume23, Aug 2001

[5]. Gregory F. Cooper,*, Constantin F. Aliferis", Richard Ambrosino, John Aronisb, Bruce G. Buchanan, Richard Caruana', Michael J. Fine, Clark Glymour", Geoffrey Gordon", Barbara H. Hanusad, Janine E. Janoskyf, Christopher Meek", Tom Mitchell", Thomas Richardson", Peter Spirtes" An evaluation of machine-learning methods for predicting pneumonia mortality"-Elsevier Feb 1997

[6]. Sana Bharti, Shailendra Narayan Singh" Analytical study of heart disease comparing with different algorithms": Computing, Communication & Automation(ICCCA),2015InternationalConference.

[7]. B.Dhomse Kanchan, M.Mahale Kishore "Study of Machine learning algorithms for special disease prediction using principal of component analysis" Global Trends in Signal Processing, Information Computing and Communication(ICGTSPICC),2016InternationalConference

[8]. Matjaz Kuka, Igor Kononenko, Cyril Groselj, Katrina Kalif, JureFettich" Analysing and improving the diagnosis of ischaemic heart disease with machine learning" Elsevier -Artificial intelligence in Medicine, Volume23, May 1999.

[9]. Geert Meyfroidt, Fabian Guiza, Jan Ramon, Maurice Brynooghe" Machine learning techniques to examine large patient databases"-Best practice &ReasearchClinicalAnaesthesiology, Elsevier**Volume 23 (1) – Mar 1, 2009.**

[10]. Gregory F.Cooper, Constantin F.Aliferis, Richard Ambrosino"An evaluation of Machine learning methods for predicting pneumonia mortality"-Elsevier, 1997.

[11]. Sanjay Kumar Sen" Predicting and Diagnosing of Heart Disease Using Machine Learning Algorithms"-International Journal of Engineering And Computer Science ISSN:2319-7242Volume6Issue 6 June 2017.

[12]. Abhishek Taneja" Heart Disease Prediction SystemUsing Data Mining Techniques"-Vol.6, No(4) December 2013.

[13]. Animesh Hazra, Subrata Kumar Mandal, Amit Gupta,Arkomita Mukherjee and Asmita Mukherjee" Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review"-Advances in Computational Sciences and Technology ISSN 0973-6107, Volume10, Number7(2017).

[14]. Beant Kaur, Williamjeet Singh" Review on Hear Disease Prediction System using Data Mining Techniques"-International Journal on Recent and Innovation Trends in Computing and Communication Volume:2 Issue:10, October 2014.Transl. J. Magn. Japan, vol. 2, pp. 740-741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[15]. Sonam Nikhar, A.M. Karandikar" Prediction of Heart Disease Using Machine Learning Algorithms"- Vol-2 Issue-6, June 2016.

[16]. S. U. Ghumbre and A. A. Ghatol, "Heart Disease Diagnosis Using Machine Learning Algorithm," Advances in Intelligent and Soft Computing Proceedings of the International Conference on Information Systems Design and Intelligent Applications

2012 (INDIA 2012) held in Visakhapatnam, India, January 2012, pp. 217–225, 2012.

[17]. Meherwar Fatima, Maruf Pasha" Survey of Machine Learning Algorithms for Disease Diagnostic"- Journal of Intelligent Learning System and applications, 2017.

[18]. Younus Ahmad Malla, Mohammad Ubaidullah Bokari" A Machine Learning Approach for Early Prediction of Breast Cancer"- International Journal of Engineering And Computer Science, Volume6, Issue5, May 2017.[

[19]. B. D. C. N. Prasad, P. E. S. N. Krishna Prasad, and Y. Sagar, "A Comparative Study of Machine Learning Algorithms as Expert Systems in Medical Diagnosis (Asthma)," Advances in Computer Science and Information Technology Communications in Computer and Information Science, pp. 570–576, 2011.

[20]. Heart Disease Forecasting System using K-Mean Clustering Algorithm with PSO and other Data Mining Methods Shilna S1, Navya EK2 ISSN(P): 2349-3968, ISSN (O): 2349-3976. Volume III, Issue IV, April- 2016.