Mapana – Journal of Sciences
2024, Vol. 23, Special Issue 3, 65-80
ISSN 0975-3303|https://doi.org/10.12723/mjs.sp3.4

Prediction of Heart Disease Using Machine Learning for Framingham Dataset

Debangana Mahapatra* and Hemlata Joshi*

Abstract

Cardiovascular Diseases (CVDs) are considered a silent killer as many people fall prey in the hand of death without any prior warning. This work represents various attributes, habits, parameters related to human beings and their connection to heart disease. Machine learning model based algorithms such as Naive Bayes, Decision Tree, K-nearest Neighbor, Logistic Regression, Recurrent Neural Network, Random Forest algorithm have been evaluated for Cardiovascular Disease (CVD) prediction. It was found that among all the algorithms evaluated, logistic regression performs best in predicting cardiovascular disease.

Keywords—Logistic Regression, Random Forest, KNN, Decision Tree, Naïve Bayes, Heart Disease Prediction, Neural Network.

I. INTRODUCTION

One of the most prevalent illnesses, cardiovascular disease (CVD) is the leading cause of mortality and according to the World Health Organization (WHO), 32% of the world's population dies from cardiovascular disease (CVD). CVDs are a group of disorders of the heart and blood vessels and cerebrovascular disease. The populations affected by heart diseases are mostly in low and middle-income countries (LMIC), around 80% of these deaths typically happen at younger ages than in countries with higher incomes.

* Department of Statistics and Data Science, CHRIST (Deemed to be University), Bangalore, India, debangana.mahapatra@stat.christuniversity.in, hemlata.joshi@christuniversity.in

To prevent and reduce the death linked to CVD, timely diagnosis and prediction are very important. But it is such a difficult task for accurate diagnosis of heart diseases. Decision making in the health care area needs expertise and sometimes for experts it becomes difficult to reach a conclusion considering limitations on human knowledge and innumerable attributes. The traditional methods of treatments are more or less dependent on physicians' intuition and biases. Now with the progress of IOT and body sensors, a human life can always be under supervision with minimal human indulgence and infinite capability. Huge amounts of captured data throughout these sensors can be tapped and processed using computing systems and it can help to reach more accurate decisions.

Most commonly used ML algorithms which have shown promising results such as Logistic Regression, Random forest, KNN, Recurrent Neural Network (RNN) and decision tree have been evaluated through the dataset. The results and successive performance of each model has been discussed in the appropriate section.

We contributed in the following ways:

Section I includes introduction of CVD. Further the section II discusses the review of the literatures helped for this study, the section III gives the data descriptions of the data used for the study and section IV gives the methodology parts. Section V and VI contains the results and discussion, conclusion respectively. And the last section gives all the literature helped for the study.

II. LITERATURE REVIEW

Dewan et.al (2015) [1] proposed a hybrid technique of Neural Network and Genetic Algorithm. But the problem is genetic algorithm converges very slower, computation time is very high. Mai showman and Tim Turner (2012) [2] used a single data mining technique with neural network ensemble. Training an Ensemble takes more time though the accuracy might be good. Veera Krishna et al. suggested PLS-DA algorithm outperforms all but it is very difficult to use in real time. Lakshmi et.al (2013) [3] used a two stage classification model to reach to conclusion. In the first stage, data gathered by sensors are put into different models and prediction is done. In second stage electrocardiogram classification using CNN and RNN leads to more accurate decision. Basheer, s., Alluhaidan, A.s., & Bivi, M. A.

(2021) [4] A remote monitoring system has been used to detect the onset of cardiac illnesses using a hybrid fuzzy-based decision tree algorithm. The structure would aid in enhancing model presentation for classification if irrelevant structures were removed from the data. Dehkordi et.al (2019) [5] proposed a model which predicts on prescription data. This method is an ensemble of many methods. The accuracy they claimed is low compared with other ML algorithms. Soni et.al (2011) [6] proposed a decision tree with Genetic Algorithm which resulted in accuracy of 99.2%. Chitra and Sreenivasan (2013) [7] proposed a cascaded neural network which has one set of cached neurons with the hidden layer. The results of the network (accuracy 85%) was better than SVM based classifier (82%).

III. DATA DESCRIPTION

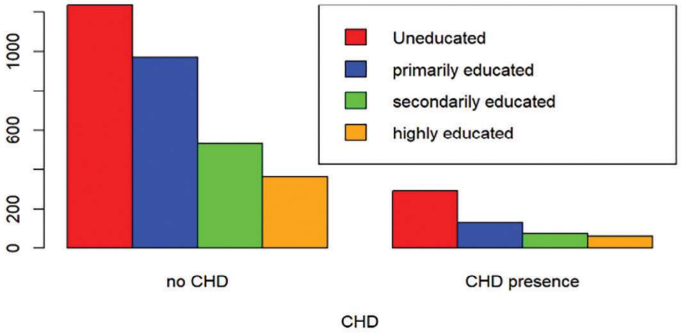
The dataset used here, Framingham Heart Study-Cohort (FHS-Cohort), is obtained through 32 clinical examination from the year 1948 and follow-up till 2018 from residents, aged 28-62 years of Framingham, Massachusetts. The participants were examined in every two years and the disease conditions under investigations were Coronary disease, stroke, hypertension, heart failure and arterial disease. The dataset has been collected from website (https://biolincc.nhlbi.nih.gov/studies/framcohort), the people who live in the Massachusetts town of Framingham. Here the main objective of the work is to predict whether the patients will develop heart disease in between 10 years. The dataset has 4238 rows and 16 columns. Where first 15 columns (factors) is a potential risk factor to develop heart disease. The factors are described in the table 1.

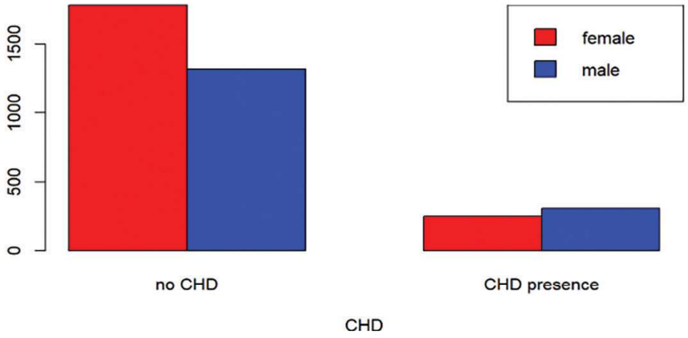
Table 1: Descriptions of the variables used for CVD

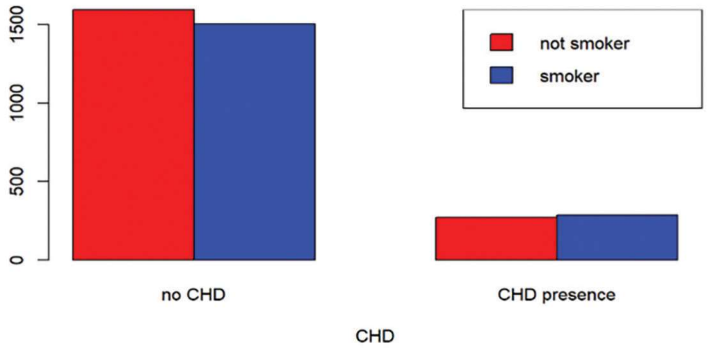
Factors	Type of variables
Male	Nominal attribute. Two factors are 0 (female) and 1 (male).
Age	Age of the person (Continuous variable).
Education	Education of the person (Categorical variable). 4 factors are 1 (uneducated), 2 (with primary education), 3 (with secondary education), and 4 (with higher education).
Current Smoker	Whether the person is a smoker (1) or not (0), Nominal (categorical) variable.
Cigs Per Day	The number of cigarettes per day that the person takes.

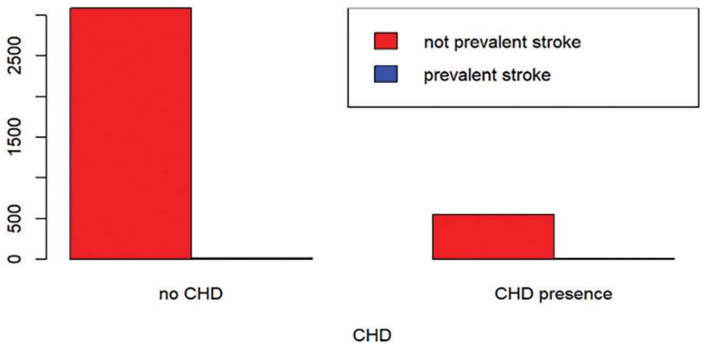
BP Meds	Whether the person is under the medication of blood pressure (1) or not (0).
Prevalent Stroke	Whether the person had previously stroked (1) or not (0).
Prevalent Hyp	Whether the persons has hypertension (1) or not (0).
Diabetes	Whether the person has diabetes (1) or not (0).
Tot chol	Total cholesterol level of the person.
Sys BP	Systolic blood pressure of that person.
Dia BP	Diastolic blood pressure of that person.
BMI	Body Mass Index of the person.
Heart Rate	Heart rate of the person.
Glucose	Glucose Level in blood.
10 Year CHD	Risk of CHD in between 10 years. ("0" means "no" and "1" means "yes").

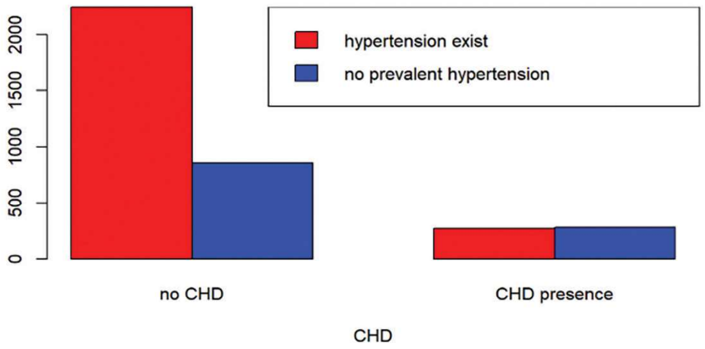
In this work, first exploratory data analysis (EDA) is done to find the graphical representation of the data and descriptive analysis. For graphical representation of categorical variables, bar plot is used.






0
500
1000
1500
not smoker
smoker
no CHD
CHD presence
CHD


0
500
1500
2500
not prevalent stroke
prevalent stroke
no CHD
CHD presence
CHD


0
500
1000
1500
2000
hypertension exist
no prevalent hypertension
no CHD
CHD presence
CHD

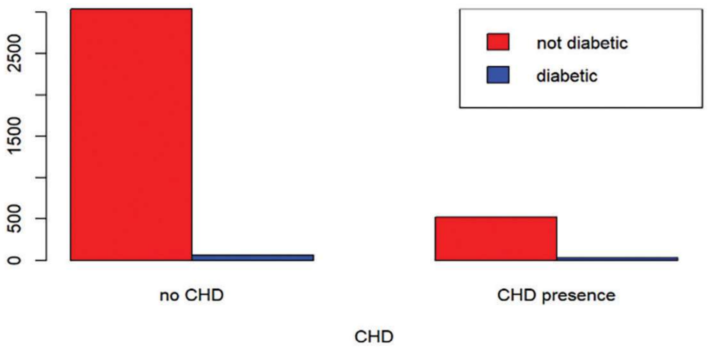


Fig. 1. Graphical Representation of the variables under study

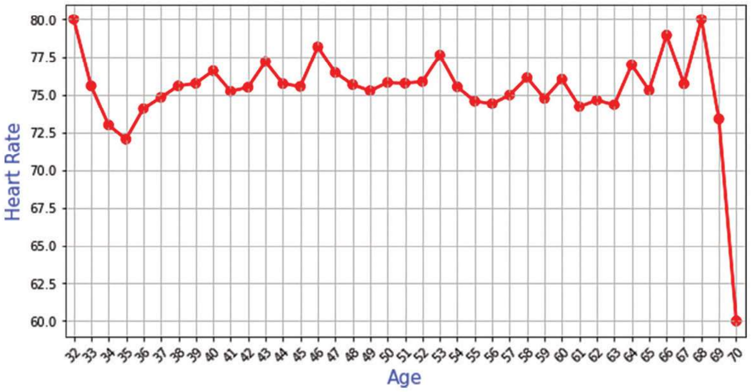


Fig. 2. Fluctuations of The Heart Rate (bpm) over the different age groups

After EDA, the data has been split into two different groups as train data (70%) and test data (30%) and fit into respective models. Train data is the larger subset of the original dataset, which is used to train model.

Table 2. Descriptive Analysis for Continuous Variables

Attributes	Minimum	Maximum	Mean	Median	1st quartile	3rd quartile
Age	32	70	49.56	49	42.00	56.00
Cig Per Day	0.000	70.00	9.022	0.000	0.000	20.00
Total Cholesterol	113.0	600.0	236.9	234.0	206.0	263.2
Systolic BP	83.5	295.0	132.4	128.0	117.0	144.0

Diastolic BP	48.00	142.50	82.91	82.00	75.00	90.00
BMI	15.54	56.80	25.78	25.38	23.08	28.04
Heart Rate	44.00	143.0	75.73	75.00	68.00	82.00
Glucose	40.00	394.0	81.86	78.00	71.00	87.00

Table 3. Descriptive Analysis for Discrete Variables.

Attributes	Types of Attributes	Population
Gender	Female (0)	55.7%
	Male (1)	44.3%
Education	Uneducated (1)	41.7%
	with primary education (2)	30.3%
	with secondary education (3)	16.5%
	with higher education (4)	11.5%
Current Smoker	Not smoker (0)	51%
	Smoker (1)	49%
BP Meds	Not under BP medication (0)	96%
	Under BP medication (1)	4%
Prevalent Stroke	Not previously stroked (0)	99%
	Previously stroked (1)	1%
Prevalent Hypertensive	Not Hypertensive (0)	69%
	Hypertensive (1)	31%
Diabetes	Non diabetic (0)	97%
	Diabetic (1)	3%
Ten Years CHD	Not developing CHD (0)	85%
	Developing CHD (1)	15%

IV. METHODOLOGY

This work represents the analysis of various types of machine learning algorithms. These are KNN algorithm, Logistic Regression, Random Forest Classification, Decision tree, Gaussian NB, Neural Network. Classification is a supervised machine learning technique where model is fully trained by the training data and is evaluated on the test data.

A. Logistic Regression

Logistic Regression is a technique of classification and predictive analysis. It gives the probability of an event occurring, such as developing a disease (1) or not developing that disease (0). In logistic regression, a logit transformation is needed on the odds, i.e.

probability of getting heart disease is divided by the probability of not getting the heart diseases. This is known as the log odds. The formula is:

Logit function:

$$ln\ (\frac{\pi}{1-\pi})=\beta_0+\beta_1x_1+\ \beta_2x_2\ ....+\ \beta_kx_k \qquad (1)$$

Here β_0, β_1, β_2, ……., β_k are the regression coefficient. In this equation, ᴨ denotes the probability of getting the heart disease. The odds ratio value $(\frac{\pi}{1-\pi})$ in logistic regression can be used to determine the impact of each variable on the target variable.

The parameters are estimated via maximum likelihood estimation (MLE). For binary classification when the probability less than 0.5 (default value) will predict 0 and when the probability greater than 0.5 will predict 1.

Logistic regression is used to estimate the relationship between dependent variables and one or more independent variables where the dependent variable is a categorical variable formed as true (1) or false (0).

B. Neural Network

A neural network consists of layers and each layer consist of nodes of shown in the following Fig. 3. The number of Layers and nodes is dependent on the type of input data. The input side nodes are mapped to independent variables and these nodes are connected to next layer's nodes either fully or in a dropped out fashion. Whenever the value of a node crosses a certain threshold of that layer, it propagates the value to the next level. It's called as forward propagation. Each path of connection does not have equal contribution for activating any node. The paths are connected in a weighted average fashion and these weights are decided by the network in backpropagation algorithm.

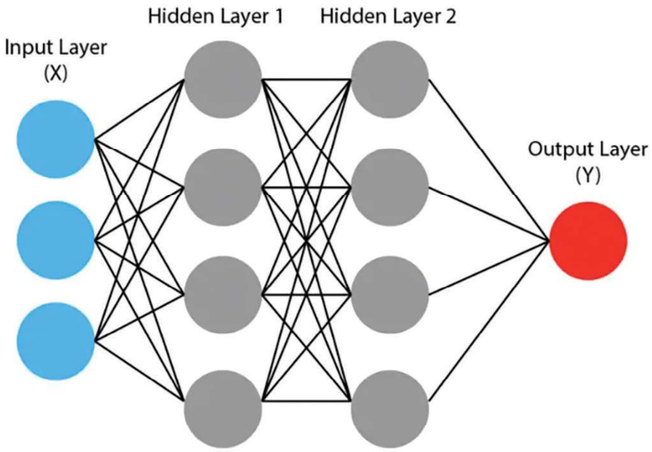


Fig. 3. Neural Network Architecture

C. Random Forest classification

In order to increase the predicted accuracy of the dataset, the random forest classification technique uses a number of decision trees on different subsets of the dataset. Each tree in RF predicts a class label for a sample. The major predicted label for a test is decided as final after voting. This goes on for each sample point.

D. KNN classification

It is a non-parametric machine learning classifier. First the distance must be defined before the classification can be made. The distance can be evaluated by Euclidean distance, Hamming distance, Manhattan distance etc. The k value specifies how many neighbors will be examined to classify a particular spot. The instance will be given the same class when k=1. Lower k value has the higher variance but low bias, and the larger k value indicates lower variance and low bias. If Datasets have missing values, the algorithm can estimate for those values in a process as missing data imputation.

E. Naïve Bayes Classification

Naïve Bayes Classifications are based on Bayes theorem. There is an assumption of this classification that there is no dependence between the features i.e. the value of a particular feature is independent of any other feature.

The proposed classification algorithm is detailed in the following figure 4. At the end of classification it can be predicted that a person is prone to develop a heart disease or not in coming 10 years.

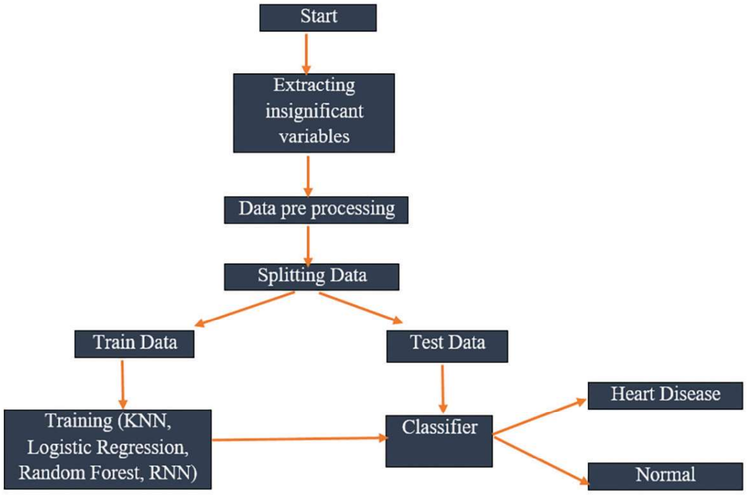

Start
Extracting insignificant variables
Data pre processing
Splitting Data
Train Data
Test Data
Training (KNN, Logistic Regression, Random Forest, RNN)
Classifier
Heart Disease
Normal

Fig. 4. Proposed Algorithm

Confusion matrix is a 2×2 matrix. It summarises the number of correct predictions of the model.

Table 4. Confusion Matrix

Predicted Actual	**Yes**	**No**
Yes	True positives (TP)	False negatives (FN)
No	False positives (FP)	True Negatives (TN)

Type I error and type II error can also be used interchangeably when referring to false positives and false negatives respectively.

From the confusion matrix table we can also obtain the accuracy i.e.

$$\text{Accuracy} = \frac{TP + TN}{Total\ number\ of\ observations}$$

Which tells that what percentage of the model is correct.

$$\text{Precision} = \frac{TP}{TP + FP}.$$

Precision estimates out of all yes predictions, how many of them were correct.

Specificity = $\frac{TN}{TN + FP}$

The level of specificity measures how accurate the model performed at predicting actual “no” events.

Sensitivity = $\frac{TP}{TP + FN}$

Sensitivity evaluates the model’s accuracy in predicting actual “yes” events.

F1-SCORE

Sometimes when we have to deal with imbalanced data set then F1-Score is used.

F1-Score = $\frac{Precision \ \times Specificity}{Precision + Specificity}$

ROC CURVE

It generates the probability values instead of binary 0/1 values. ROC curve provides good overview between TP rate and FP rate for binary classifier using different probability thresholds.

V. RESULTS AND DISCUSSION

Table 5 gives the results obtained from logistic regression where the study variable is the risk of developing the CHD in between 10 years.

Table 5. Results from Regression Model

Attributes	**Coefficients**	**P-Value**
Intercept	-8.258321	< 2e-16 ***
Male	0.534871	1.14e-06 ***
Age	0.062166	< 2e-16 ***
Primarily Educated	-0.192060	0.11970
Secondarily Educated	-0.193891	0.19661
Highly Educated	-0.059869	0.71609
Current Smoker	0.072388	0.64421
Cig Per Day	0.018020	0.00384 **
Under BP Medication	0.165049	0.48148
Prevalent Stroke Exist	0.704569	0.15167
Prevalent Hypertensive	0.233855	0.09065.
Diabetic Patient	0.026308	0.93367
Total Cholesterol	0.002369	0.03590 *

Systolic BP	0.015451	5.05e-05 ***
Diastolic BP	-0.004095	0.52506
BMI	0.005149	0.68716
Heart Rate	-0.003007	0.47533
Glucose	0.007212	0.00125 **

From the table 5, ‘.’, ‘*’, ‘**’, ‘***’ marked attributes are significant for 0.1, 0.05, 0.01, 0.001 level of significance respectively.

The odds ratio value in logistic regression can be used to determine the impact of each variable on the target variable. As for example, for the male variable is having a coefficient value of 0.534871 with a reference value 1 and the odds ratio value is 1.7072274346 which means that for male patients, the odds of getting heart disease 1.7072274346 times the female odds or it can be said the tendency of men to heart disease is higher than women. For the age variable with a coefficient value of 0.062166, it is found that the odds ratio value is 1.0641394116 which means that for the age variable there will have a significant increase. On the other hand, for the variable education with reference education values are 2, 3 and 4. Education2 with an estimated coefficient of -0.192060295 will have odds of 0.8252571085, while education3 with an estimated coefficient of -0.193890867 will have odds of 0.8237477974 and education4 with an estimated factor of 0.9418880223 will have odds in the amount of 0.9418880223. Similarly for the other variables.

We checked the dataset for others classifications - Random Forest, Decision tree, Gauss naïve Bayes, KNN classification. It involves that a building like as a tree. Each node in this tree represents a test. The leaf nodes represents the class levels. As Random Forest classification is the combination of more than one decision tree, it works in a better way. Decision trees and random forest are combined with each other and applied here. The Random forest classifier is initialized with n_estimators = 7, random_state = 11, max_depth = 5. Overfitting is less likely with random forest, and accuracy is significantly higher than decision trees. Uncorrelated decision trees have a significant influence on the accuracy of random forests. Through bootstrapping, random forests are able to produce decision trees that are not associated. The depth of decision tree classifier is 6 and the criterion, parameter which decides impurity of a split will be measured, selected “entropy”.

We choose k = 9, for KNN classification. The model might be considered overly specific and poorly generalized. It also has a tendency to be noise-sensitive. This model performs well on train data but may not be a good predictor for unobserved data. Performance as training progresses is function used for fully connected layers is Rectified Linear (ReLu).

As the prediction is a binary classification, last layer activation is chosen as Sigmoid. Reported accuracy for test set for this three layer neural network is 84.59%.

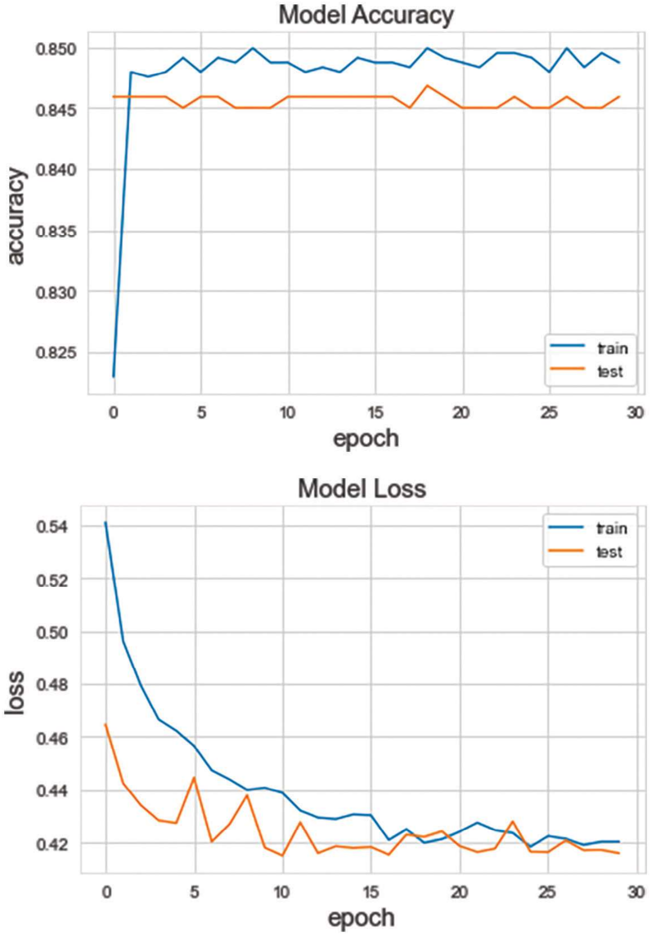


Fig. 5. Loss and accuracy vs Epoch for three layer sequential model

Results from all the classical ML algorithms are listed in the following Table 6 where it is clearly evident Logistic Regression turns out to be best performer out of all. The analysis of highly important features shows that clinicians can focus more on certain causes of CVD.

Table 6. Comparison of Performance of Algorithms

Algorithm	**Accu-racy**	**Precision (for 0)**	**Precision (for 1)**	**f1-score (for 0)**	**f1-score (for 1)**	**Recall (for 0)**	**Recall (for 1)**
Logistic Regression	84.86	0.85	0.64	0.92	0.08	1.00	0.04
Naïve Bayes	82.86	0.87	0.40	0.90	0.29	0.94	0.23
KNN	82.95	0.86	0.47	0.91	0.15	0.98	0.09
Decision Tree	83.13	0.86	0.35	0.91	0.16	0.96	0.11
Random Forest	84.59	0.85	0.62	0.92	0.06	1.00	0.03

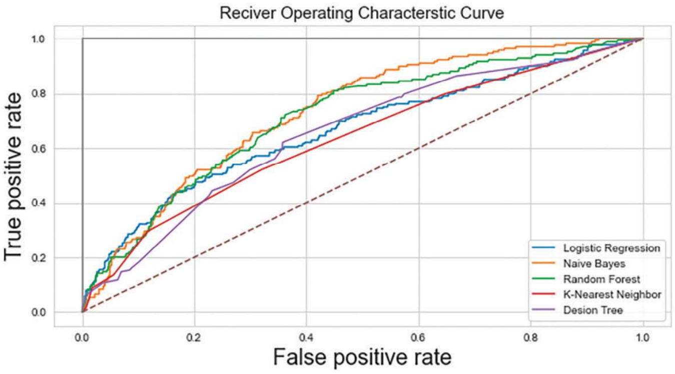


Fig. 6. ROC curve of various algorithms used

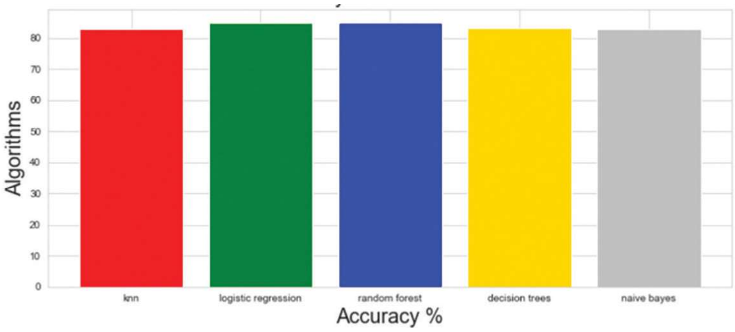


Fig. 7. Accuracy of different algorithms

VI. CONCLUSION

In this study using a training dataset of 70% and a testing dataset of 30%, the LR (logistic regression) classifier had an accuracy rate of 84.86%. It is shown that the logistic regression model gave the better accuracy among Random Forest, KNN, Naïve Bayes, Neural Network, and Decision Tree. Logistic regression can be used to predict the risk of developing CHD in a period of 10 years. Use of more training data ensures the higher chances of the model to predict accurately whether the person will develop heart disease or not.

Table 7 gives the estimated significant coefficient (only for the significant variables) obtained using logistic regression.

Table 7. Significant Variables Fro Logistic Regression

Attributes	Coefficients	P-Value
Intercept	-8.258321	< 2e-16
Male	0.534871	1.14e-06
Age	0.062166	< 2e-16
Cig Per Day	0.018020	0.040277
Total Cholesterol	0.002369	0.03590
Systolic BP	0.015451	5.05e-05
Glucose	0.007212	0.00125

And it can help medical practitioners to diagnose their patients more accurately and death rate due to cardiovascular disease can be greatly reduced if we take care of all significant variables. Integration of the algorithms with the healthcare smart system will automate the process and unburden the load on medical practitioners.

REFERENCES

[1] Dewan, A., & Sharma, M. (2015). Prediction of heart disease using a hybrid technique in data mining classification. In 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom) (704-706). IEEE.

[2] Shouman, M., Turner, T., & Stocker, R. (2012). Using data mining techniques in heart disease diagnosis and treatment. In 2012 Japan-Egypt Conference on Electronics, Communications and Computers (173-177). IEEE.

[3] Lakshmi, K. R., Krishna, M. V., & Kumar, S. P. (2013). Performance comparison of data mining techniques for predicting of heart disease survivability. International Journal of Scientific and Research Publications, 3(6), (1-10).

[4] Basheer, S., Alluhaidan, A. S., & Bivi, M. A. (2021). Real-time monitoring system for early prediction of heart disease using Internet of Things. Soft Computing, 25(18), (12145-12158).

[5] Dehkordi, S. K., & Sajedi, H. (2019). Prediction of disease based on prescription using data mining methods. Health and Technology, 9, (37-44).

[6] 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), (43-48).

[7] Chitra, R., & Seenivasagam, V. (2013). Heart disease prediction system using supervised learning classifier. Bonfring International Journal of Software Engineering and Soft Computing, 3(1), (01-07).