Heart Disease Prediction using Supervised Machine Learning Algorithms

Deepak Kumar
Department of Software Engineering
Delhi Technological University
Delhi, India
deepakkumar_2k18se051@dtu.ac.in

Sanjay Patidar
Department of Software Engineering
Delhi Technological University
Delhi, India
sanjaypatidar@dtu.ac.in

Dheeraj Rukwal
Department of Software Engineering
Delhi Technological University
Delhi, India
dheerajrukwal_2k18se055@dtu.ac.in

***Abstract*— Heart disease has become a common cause of death worldwide in recent years. People's way of living changes, dietary habits, office working cultures, and other factors have all played a role in this worrisome problem around the world. The best way to stop this disease is to develop a method that will detect early symptoms and hence save more lives. With the help Machine Learning(ML) algorithms, researchers can predict the likelihood of developing cardiovascular disease in people who are at risk. It is critical to develop a precise and dependable technique to have early disease prediction by automating the task and therefore achieving efficient disease management. Several academics have described their efforts to develop the best feasible technique for predicting heart disease in previous publications. The goal of this study is to compare alternative algorithms for predicting cardiac disease. The results of important data mining techniques are presented in this work, which can be utilized to construct a highly efficient and accurate prediction model that will aid doctors in minimizing the number of people killed by heart disease.**

This study compares the metrics for prediction of heart disease for 6 different machine learning algorithms which are "Logistic Regression" (LR), "Decision Tree" (DT), "Random Forest" (RF), "Support Vector Machine" (SVM), "Gaussian Naïve Bayes" (GNB) and "k-Nearest Neighbor" (kNN).

I. INTRODUCTION

The human heart is one of the most important organs in the human body. It's the device that circulates oxygen-rich blood to different parts of the body. The heart works 24 into 7 to ensure that all other organs get the right amount of oxygen-rich blood, and any interference in its functions would have an adverse effect on other organs' appropriate functioning, which can be catastrophic. Heart disease or cardiovascular disease, is a dangerous medical disorder caused by the heart's failure to perform its circulation functions properly.

If a patient ignores the disease's early symptoms, which appear to be warning signs, the patient will have no time to recover and will eventually die on the spot. A heart attack is the medical term for this. It occurs because the purpose of the arteries is to give oxygen-rich blood to the heart, but plaque forms as a result of fatty and other substances, which disrupts the functioning of a normal artery and converts it into a narrowed coronary artery. As a result, blood flow might be slowed or entirely stopped, as depicted in Fig 1. Controllable risk factors and uncontrolled risk factors are the two types of risk variables that cause coronary artery disease. Diabetes, smoking, obesity or overweight, cholesterol, hypertension, less physical activities are all controllable risk factors. Age, sex, previous medical conditions and history are all uncontrollable risk factors. In the last decade, Heart disease is the top cause for the death of people worldwide According to a WHO report, about 17.9 million people die each year as a result of cardiovascular disorders, with coronary heart disease and brain stroke accounting for 80% of these deaths [1]. A variety of laboratory tests and imaging examinations can be used to identify cardiovascular disease. However, the patient's medical and family history, risk factors, and physical examination are the most important aspects of diagnosis. We can synchronize the results and predict the existence of disease from findings and processes using statistical data. Doctors can make better decisions with the help of automation and deep learning.

Disease prediction automation can establish a single platform from which structured data can be obtained and patients can receive efficient care. As a result, it raises the bar for tailored health care. Computers are taught to discover patterns which disease occurs and translate them into structural data in order to anticipate the disease using AI and machine learning. AI is used to innovate in the areas of operations, revenue cycle and "electronic health records" (EHR). It will be connected with the actual clinical and existing technologies in the future, allowing practitioners to access real-time data at the time of service.

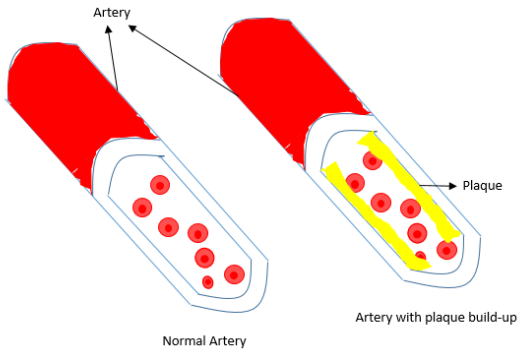

Fig. 1 Plaque formation in arteries

The various algorithms that have been used are:

A. Logistic Regression(LR)

In this algorithm on the basis of independent variables, it is used to predict a categorical dependent variable. In most LR cases, a binary variable is the dependent variable containing data encoded as Yes (1, True, success, etc.) or No (0, False, failure). This logistic function given below is the basis of LR algorithm.

$$P(X) = \frac{1}{1+e^{-g(x)}}$$

Where, g(x) can be represented in a linear form as below with combination of features (x_k) and weights β_k

$$g(x) = x_0 + x_0\beta_0 + x_1\beta_1 + x_2\beta_2 + \cdots + x_3\beta_3 + \varepsilon$$

B. Decision Tree (DT)

A DT is a tree-like model that is used to aid in the development of automated predictive models in both regression and classification applications. It is a type of non-parametric supervised learning algorithm. It uses simple selection rules learned from facts capabilities to build a tree that estimates the value of the predicting variable. The most key attribute is placed at the top for the creation of DT. The training package contains various subsets. The information for the same attribute is contained in each subset. The process goes on till all leaf nodes have not been found. Each leaf node indicates a desired outcome.

The target variables in the classification tree model can take a discrete set of values. In tree structures, on the other hand, leaves indicate class labels and branches represent feature joins that represent class labels. The entropy equation can be found below:

$$E = -\sum_{b=1}^{n} p_{ab}\, log_2 p_{ab}$$

The tree structure, with its various nodes and edges, is well suited for visualizing the interactions of the variables. When there is a monotonic transformation among the features, the decision tree performs well. Nonetheless, decision trees are unable to accommodate linear relationships, and the trees might become unstable at times. When the number of decision tree terminal nodes increases, it becomes increasingly difficult to decipher the entire tree.

C. Random Forest (RF)

Random Forest is a popular technically advanced Supervised Ensemble (combination) classification algorithm. During the training stage, it builds a DT forest by using many datasets. During the testing stage each tree in the forest predicts a class label. A majority vote is used to make the final decision for each set of test data. The Random Forest (RF) method analyses data using several decision trees, gathering predictions from each one, and determining the optimal option. It uses bagging algorithm to deal with missing values of data [2].

D. Support Vector Machine (SVM)

It is a classification algorithm that works upon both non-linear and linear data. Its working is based on finding a suitable hyper plane that will separate values of both classes from one another. When the margins between classes are wider, the model improves. There should be no points in the interior area of a margin. Data points which are nearest to hyper plane and also affect its position are called Support Vectors. design is used for the simulation of real-world situations, uses mathematical functions. Its performance improves as the number of attributes increases [3].

E. Gaussian Naïve Bayes (GNB)

GNB is a variant of the Naïve Bayes and is based on the Bayes theorem. It can be used for both multi-class and binary classification. It follows Gaussian distribution and mostly works upon continuous data. It is a type of probabilistic classifier. The Bayes Theorem is based on the below equation where P(A/B) is the "posterior probability", P(B/A) is the "likelihood probability", P(A) is the "prior probability" and P (B) is the "marginal" probability.

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

Although because it is reliant on assumptions and conditional class independence, accuracy suffers. [4].

F. K-Nearest Neighbour(k-NN)

The K-Nearest Neighbor is a supervised ML algorithm. It categorizes items based on their proximity to one another. It's an example of instance-based learning. The Euclidean distance is used to calculate the distance between an attribute and its neighbors. It makes use of a set of previously used marked point to calculate a new point. Grouping in this algorithm is done by how much some data is resembling to other data. This algorithm is easy to use and very effective for regression, classification and searching. It does get affected by irrelevant information.

II. LITERATURE REVIEW

S. Musfiq Ali et.al conducted a research on the Cleveland database , with 10-fold cross validation and achieved a highest accuracy of 91.2% for GNB[5].

Abdullah et. al developed a Data Mining model to increase the accuracy for heart Disease Prediction, the model was based on RF classifier [6]. Sonam Nikhar et al.[7] employed used Cleveland Dataset with 303 instances for prediction model, although only 19 attributes were used for the GNB, DT technique. They also discovered that Decision Tree has a higher accuracy than the Nave Bayes Classifier. Ravindra Yadav et al.[8] deployed ML approach for the Cardio Vascular Disease Prediction Survey, which included the DT, GNB , Neural Networks , Deep Learning and SVM. The decision tree's conclusion is generated using ID3, CART Cpercent.0,CYT , and J48 .

Devansh Shah et.al,[9] used Cleveland database with 303 instances and 14 attributes for heart disease prediction. K-NN algorithm had the highest accuracy. Archana Singh et al. [10] employed the Cleveland dataset resulting in the following results: Linear Regression: 78 percent, DT: 79 percent, SVM: 83 percent, and K-NN: 87 percent accuracy.

PushkalaV et.al, used Cleveland dataset and found GNB gives the maximum accuracy of 91 % [11]. This study work by Jaymin Patel et al. [12] compares remarkable class algorithms in pursuit of superior performance in heart Disease Prediction utilising WEKA .

Using standard performance matrices, Kavitha et al. [13] tested several machine learning techniques. The performance of the RF, KNN, SVM and NB algorithms was assessed. Dataset was taken from the UCI repository to conduct this analysis. They only looked at accuracy, precision, and recall when evaluating their models, whereas in our study, we looked at harmonic mean also known as F1-Score of recall and precision.

Rovin Dbritto et. al[14], evaluated three different machine learning models accuracy on large dataset, which are Naïve Bayes, SVM and LR. This paper evaluates the performance for a large dataset for 6 different Machine Learning Models.

Punith H B et al. [15] did a comparative analysis of ML Algorithms on 6 approaches which are NB, SVM, DT, RF, k-NN and Linear Regression using only 9 Variables. The highest accuracy was achieved in Random Forest Algorithm. This study uses 13 variables to predict and has done one-hot encoding on some of the categorical independent variables.

In 2013, Shamsher B. P. et. al[16] used ML Techniques for heart disease. The objective of our works increase the accuracy using one-hot encoding.

A survey was conducted by Kavitha B S, M.Siddapa to predict cardiac disease. The majority of the data in this study came from the Cleveland repository, and multiple machine learning classifiers were employed to create a prediction model for cardio vascular disease prediction. According to the study, Random Forest algorithms outperformed other models in terms of accuracy [17].

N. Komal et.al [18] did a study to predict accuracy using various ML algorithms. In their study they found that RF Algorithm had the maximum accuracy with 85.71%. Aditi Gavhane et. al [19] created a model for the forecasting of cardiac disease using the MLP (Multilayer Perceptron) machine learning method, which provides predictive results of a user leading to CAD.

By employing logistic regression as a result discriminant purpose, Robert Detrano produced experimental results that demonstrated precise classification of cardiac illnesses, with an accuracy of about 77 percent [20].

M A Rizvi et. al [21] their research highlights current modern disease prediction methodologies. All the researches that we studied have not used one-hot encoding in the data-preprocessing step. Our research does one-hot encoding and compares the performance for each classifier in various metrics.

III. PROPOSED SOLUTION

The block diagram for the proposed methodology has been depicted in Fig 2. The key components include Data collection, Data preprocessing, Data splitting and Performance Evaluation.

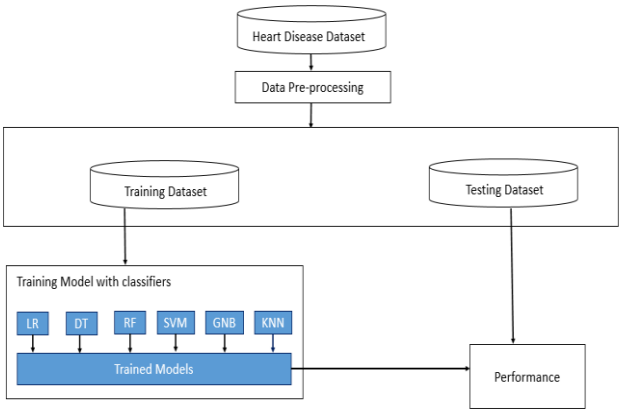

Fig. 2 Block Diagram of the proposed System

A. Data Collection

The dataset that was used in this study is available at kaggle[22] . The dataset dates to 1988 and consists a combination of four datasets which are from Long Beach V, Switzerland, Hungary and Cleveland. In total it contains 75 Attributes, excluding the predicted attribute. All the researches use a subset of 14 of them. The “target” field is the predicted attribute making a total of 76 columns. The dataset consists of 1025 patients with 713 male and 312 females. The description of attributes is given in Table III.

B. Data Pre-processing

The collected dataset contained no missing values. The dataset consists of 5 attributes with continuous numerical values which are age, trestbps, chol, thalach and oldpeak. There are also 8 categorical columns excluding the “target” column. These are exang, fbs, sex, ca, cp, restecg, thal, and slope. Out of these 8 categorical columns 3 are binary categorical values which are exang, fbs and sex. The rest are ordinal categorical variables.

One-Hot encoding was done on these variables. This means converting categorical data into numerical form. In one hot encoding a set of binary variables in a particular order represent the integer encoded variables.

TABLE I. BEFORE ONE-HOT ENCODING

Index in dataset	restecg
0	1
1	0
6	2

TABLE II. AFTER ONE-HOT ENCODING

Index in dataset	restecg_0	restecg_1	restecg_2
0	0	1	0
1	1	0	0
6	0	0	1

TABLE III. DESCRIPTION OF THE ATTRIBUTES

S/ No	Attribute	Description	Values
1	age	Age of the patient in years	Continuous [29-77]
2	sex	Patient's sex	1 : male 2 : female
3	cp	Chest pain type	0 : asymptomatic 1 : atypical angina 2 : non-angina pain 3 : typical angina
4	trestbps	Patient's resting blood pressure (mm Hg on admission to the hospital)	Continuous [94-200]
5	chol	Patient's cholesterol measurement in mg/dl	Continuous [126-564]
6	fbs	Patient's fasting blood sugar	1 : > 120 mg/dl 0 : <= 120 mg/dl
7	restecg	Resting electrocardiographic results	0 : normal 1 : having_ST_T wave abnormal 2 : left ventricular hypertrophy
8	thalach	Patient's maximum heart rate achieved	Continuous [71-202] bpm
9	exang	Exercise included angina	0 : no 1 : yes
10	oldpeak	Depression in ST brought by exercise that is relative to rest	Continuous [0-6.2]
11	slope	ST segment's peak exercise slope.	0 : down sloping 1 : flat 2 : up sloping
12	ca	Count of major vessels that have been colored with fluoroscopy	0-4 value
13	thal	A blood disorder called thalassemia value	0 : Null 1: represents a defect that is fixed. In this condition in some parts of the heart there is no blood flow 2 : blood flow is normal 3: defect is reversible.
14	target	Is the heart disease present	0 : No 1 : Yes

For example the "restecg" variable has three possible values 0, 1 or 2, basically three categories are present. 3 binary variables would be needed to depict the categories. A "1" value is used for that particular category and "0" for the other categories in a particular row. For example Table I would be converted into Table II.

Our dataset was scaled using StandardScaler. Numerical input variables scaled to the normal (standard) ranges improve the performance of several machine learning methods. It is done by first subtraction of the mean and then division by the standard deviation to every variable. After the following operations have been performed the Standard Deviation becomes one and the mean equals to zero.

The process of standardization to variable is done as follows:

Standardization:

$$y = \frac{(x - \mu)}{\sigma}$$

With Mean:

$$\mu = \frac{1}{N}\sum_{i=1}^{N}(x_i)$$

And Standard Deviation:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

IV. RESULTS AND ANALYSIS

The performance metrics that have been used are as follows.

A. Accuracy

Accuracy is the ratio of total number of correct predictions to the total number of predictions made by the model.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

B. Precision

Out of the total predicted positive values, how many are true positives (actual positives) and not false positives

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

C. Recall/Sensitivity

Out of the total actual positive values, how many are correctly predicted as positive values and not false negatives.

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

D. F1-Score / Harmonic Mean

The F1 score is the harmonic mean calculated between Precision and Recall

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall}$$

In these Metrics,
"TP" denotes "True Positives" (True predicted to be true). FP denotes "False Positive" (False predicted to be True).FN denote "False Negative" (True predicted to be False).TN denotes "True Negative" (False predicted to be False)

Receiver Operator Characteristic (ROC) is used for evaluation of binary classification problems. It is curve that is plotted between "True Positive Rate" (TPR) and "False Positive Rate"(FPR) at various threshold values. The AUC or

Area under curve is the synopsis of the ROC curve. Fig 4 depicts the ROC-AUC curve of the various classifiers.

TABLE IV. COMPARISON OF METRICS FOR DIFFERENT APPROCAHES

S/No.	Classifier	Accuracy (%)	Precision (%)	Recall (%)	F-1 Score (%)
1.	Logistic Regression	82.92	78.33	91.26	84.30
2.	Decision Tree	95.12	94.28	96.11	95.19
3.	Random Forest	98.53	100	97.08	98.52
4.	Support Vector Machine	87.31	84.07	92.23	87.96
5.	Gaussian Naïve Bayes	74.63	68.34	92.23	78.51
6.	K- Nearest Neighbors	81.95	77.96	89.32	83.25

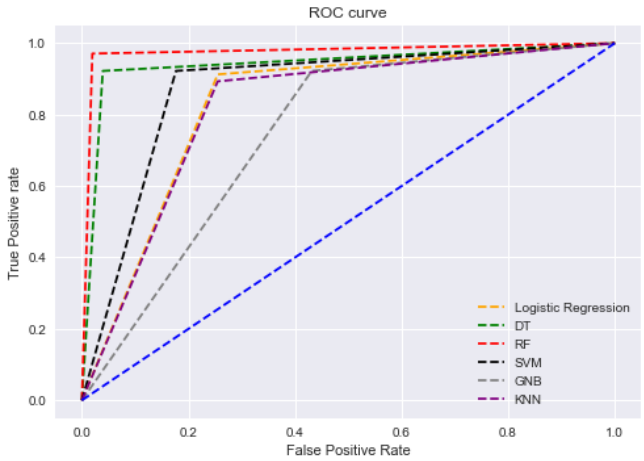

Fig. 3 ROC-AUC curve for each classifier

V. CONCLUSION

The aim of this study was to compare the performance of various supervised machine learning algorithms for the prediction of Heart Disease. 6 Machine learning classifiers were used that were Logistic Regression, Decision Tree, Random Forest, Support Vector Machine , Gaussian Naïve Bayes and k-Nearest Neighbors. Many prior researches on the same topic were analyzed. In this study it was found out that random Forest algorithm performed the best with 98.53% accuracy and GNB the worst with 74.63% accuracy. It was mainly due to one-hot encoding that was done which helped Random Forest in making better informed decisions. Further work that can be done in this area is increasing the accuracy using hyper-parameter tuning and using a larger dataset.

REFERENCES

[1] T. R. H. Michael D Seckeler, "The worldwide epidemiology of acute rheumatic fever and rheumatic heart disease," Clinical Epidemiology, pp. 3-67, 2011.

[2] Y. Y. W. a. J. Z. Liu, "New machine learning," International Conference on Information Computing and Applications, pp. 246-252, 2012.

[3] D. M. N. P. a. A. Mythili T., IJCA, vol. 68, p. 16, 2013.

[4] S. P. S. K. B. Devansh Shah, "Heart Disease Prediction using Machine Learning Techniques," Springer Nature, 2020.

[5] M. I. K. ,. M. A. I. M. S. Musfiq Ali, "Heart Disease Prediction Using Machine Learning Algorithms," Dhaka, 2019.

[6] R. R. Abdullah AS, "A data mining model for predicting the coronary heart disease using random forest classifier.," *International Journal of Computer Applications® (IJCA)*, 2012.

[7] Y. T. V. Lafta R, "Intelligent Recommender System based on Short Term Risk Prediction for Heart Disease patients.," in *International Conference on Web Intelligence and Intelligent Agent Technology.*, Singapore, 2015.

[8] h t t p s : / / w w w . i j c r t . o r g / p a p e r s / IJCRT1807425.pdf (13 November 2020)

[9] S. P. S. K. B. Devansh Shah, "Heart Disease Prediction using Machine Learning Techniques," *Springer Nature Singapore Pte Ltd*, 2020.

[10] R. K. Archana Singh, "Heart Disease Prediction using Machine Learning Algorithms," in *International Conference on Electrical and Electronics Engineering*, 2020.

[11] A. T. &. A. S. A. PushkalaV, "Comparative Study of Heart Disease Prediction Using Machine Learning," *International Journal of Innovations in Engineering and Technology (IJIET)*, vol. 124, no. 10, 2019.

[12] J. T. D. &. P. S. Patel, "Heart disease prediction using machine learning and data mining technique.," pp. 129-137.

[13] S. a. S. A. Mukherjee, "Intelligent Heart Disease Prediction using Neural Network.," *International Journal of Recent Technology and Engineering*, 2019.

[14] A. S. a. V. J. Rovin Dbritto, "Comparative Analysis of Accuracy on Heart Disease Prediction using Classification Methods," *International Journal of Applied Information Systems*, vol. 11, no. 2, 2016.

[15] V. H. K. a. U. E. S. Punith H B, "Comparitive Analysis of Machine Learning Algorithms in the Study of Heart Disease Prediction," International Journal of Engineering Research & Technology (IJERT), 2020.

[16] P. K. Y. a. D. D. P. Shamsher Bahadur Patel, "Predict the Diagnosis of Heart Disease Patients Using Classification Mining Techniques," Journal of Agriculture and Veterinary Science, vol. 4, no. 2, 2013

[17] K. B. S. a. D. M. Siddappa, "A Survey on Machine Learning Techniques to Predict HeartDisease," *International Journal of Computer Science & Communication*, pp. 48-53, 2020.

[18] N. K. S. G. S. P. D. K. Kumar, "Analysis and prediction of cardio vascular disease using machine learning classifiers," in *6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, 2020.

[19] A. K. G. P. I. &. D. K. Gavhane, "Prediction of heart disease using machine learning.," in *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, 1275-1278, 2018.

[20] R. Detrano, W. Steinbrunn and M. Pfisterer, "International application of a new probability algorithm for the diagnosis of coronary artery disease," *American Journal of Cardiology*, vol. 64, no. 3, pp. 304-310, 1987

[21] H. &. R. M. A. Sharma, "Prediction of heart disease using machine learning algorithms: A survey.," *International Journal on Recent and Innovation Trendsin Computing and Communication*, vol. 5, no. 8, pp. 99-104, 2017.

[22] Heart Disease Dataset | Kaggle