Heart Disease Prediction with Ensemble Learning Technique

R. Santosh[1](✉), B. M. M. Tripathi[1], Arempula Sreenivasa Rao[2], and Y. Satwik[3]

[1] Department of Electronics and Communication Engineering, Velagapudi Ramakrishna Siddhartha Engineering College, Kanuru, Vijayawada 520007, India
routus@gmail.com

[2] Department of Electronics and Communication Engineering, Annamacharya Institute of Technology and Sciences, Hyderabad 501512, India

[3] Department of Computer Sciecnce and Engineering, CVR College of Engineering, Hyderabad 501510, India

Abstract. Machine Learning (ML) is a field of science which is proven to be significantly effective and efficient in forecasting diseases and making predictions from analysing the enormous amounts of data produced by various healthcare industries. Several engineers across the world have developed ML algorithms for heart disease prediction in which different accuracies are obtained for the same technique for a given data set. It is in reality, contradictory to say which algorithm will be more beneficial to predicting whether the heart is healthy or unhealthy. A novel approach has been presented to predicting heart disease in which six algorithms have been developed and analyzed for predicting the heart disease efficiently. The automatic and efficient output will be derived depending on the accuracy, sensitivity, specificity, and precision. The Random Forest and Naïve Baye's classifier have proven effective in predicting heart disease from the UCI Cleveland dataset.

Keywords: Machine Learning · Classification · Heart Disease Prediction · Python

1 Introduction

Heart Disease, popularly called Cardiovascular Disease in Medical Field, is among the major causes of increased mortality rate globally [1]. Predicting heart disease is not an easy task, and it involves studying several contributing factors. There are many cases and evidence that Machine Learning (ML) has proven to be very efficient in tackling real-world problems and generating optimal and efficient solutions. Several people focused on improving the accuracy of heart disease prediction and making it efficient using few ML algorithms [2–10]. An hybrid ML model of Random Forest algorithm with Linear ML model produced an output of 88% accuracy [4]. The Support Vector Machine is also significantly effective in that it could have an estimated accuracy of 85% [5]. The model developed as a comparative study on Naïve Baye's Theorem, Decision Tree, K-NN, and

B. Raj et al. (Eds.): ICETE 2023, AER 223, pp. 87–95, 2023.
https://doi.org/10.2991/978-94-6463-252-1_11

Random Forest showed that the accuracy of the output is around 90% for K-NN [6]. It is most contradictory among several workers which algorithm will be more beneficial to predicting whether the heart is healthy or unhealthy. Therefore, the authors have thought it would be useful to take the outcome of the algorithm, which has provided better accuracy for the given data set. In this paper, the authors have provided an approach to generating efficient work from the model by comparing the accuracy of various algorithms. We employed Logistic Regression algorithm, Decision Tree algorithm, K-Nearest Neighbours algorithm, Naïve Baye's classifier, Random Forest algorithm, and Support Vector Machine algorithm to analyze the given datasets. The output is generated from the ML Algorithm, which is having highest accuracy. Python has been used to develop these algorithms, which has proven to be very efficient in developing such applications with the help of its vast number of libraries and functions.

2 Traditional Methodology

The diagnosis of heart diseases can be carried out using the below methods.

i. Electro-cardiogram (ECG or EKG).
ii. Echo-cardiogram.
iii. Holter monitoring.
iv. Catheterization.
v. Computerized Tomography (CT) scan.
vi. Magnetic Resonance Imaging (MRI).

These tests are respectively different from each other. The patient's heart condition is diagnosed based on the clinical analysis of the heart data from respective tests. Manually, the tests are performed on the patient, and after systematic computerization, the reports will be delivered. The significant advantage of the traditional methods is that they are easy to understand and interpret. By considering the small set of attributes of the data, the outcome can be diagnosed accordingly. The biological paradigms are easily calculated and analyzed in traditional approaches. Yet, there are some drawbacks associated with these tests. As per the generalized study, the limitations are

i. It cannot handle the enormous amount of data for patient records.
ii. Any other hospital does not validate specific hospital reports.
iii. Few of the tests are costly and time-taking.

Machine Learning has been introduced to overcome these limitations to a reasonable extent. Irrespective of the size of the dataset, whether it will be small, large, or even enormous in some cases, the developed model works the same way. The flexibility and the integrity that the ML provides in dealing with the datasets are beneficial and extensive. As per the study, ML can provide significant enhancement in pointing out synaptic communication, which is not perfectly done through traditional methods. These systems are robust and can be fixed for bugs in most cases. The specificity of ML algorithms is that the capabilities of the models are very high that even integrated approaches can be made a part in producing highly accurate and robust applications.

3 Conception of Machine Learning

Machine Learning, as a subject, is very extensive and useful in several applications for data manipulation and data processing. It is used in creating a vast number of applications that potentially lead to the reduction of Human Intervention. ML is an enhanced and advanced study of computer programs that could learn from the experience and the past work without manually altering the way of the work of the programs. ML and DL (Deep Learning) are the subsets of the whole of Artificial Intelligence (AI). In today's world, almost every software application runs on the technology of AIML. Many companies and organizations are consistently working on developing several applications in ML for the enhancement of society in terms of technology. In general, every ML algorithm or model follows several steps for outcome production. These steps in each stage enhance the workflow of the model in generating the output efficiently. ML algorithms work on two methods: The "Supervised Learning" model and The "Unsupervised Learning" model. The seven algorithms used in this work are based on supervised learning.

4 Workflow of the Model

The machine learning algorithms are deployed in Python platform, and the dataset is subjected to the algorithms after necessary pre-processing is done. This subset of all 13 attributes (excluding the target attribute) is considered from the data set that is pre-processed. After the train-test split, the dataset instances are considered to predict the heart disease from the models. The 6 different models (LR, DT, KNN, NB, RF, and SVM) are exclusively trained with the training data and made ready to predict the disease from testing data. Apart from the testing data, a few real-life examples have also been experimented. The accuracy, precision, specificity and sensitivity are calculated for all the algorithms. With the help of Python, the respective scores are recorded, and the output is produced from the model, which is highly accurate among all the algorithms through comparative analysis. The confusion matrix assists in evaluating the model. The confusion matrix generates 4 outcomes, namely TP (True Positive), FP (False Positive), TN (True Negative), and FN (False Negative). Based on these values the accuracy, precision, specificity and sensitivity are calculated.

4.1 Algorithms Deployed

4.1.1 The Logistic Regression Model (LR)

Logistic Regression model is a statistical tool for the analysis and categorical classification which is used to predict the dependent variable on the basis of the functionalities of the set of independent variables, such that the dependent variable is categorized as an outcome. It encases the relationship between the existing independent variables and the dependent variable. Here, Linear Regression model takes multiple inputs and then generates the probability of them falling into one of the two outcome categories. Based upon the historical data, it allows input data for better classification. The simple description of Linear Regression model can be as follows.

Let dependent variables be represented by Y_i in which $Y_i = 0$ denotes that there is an absence of heart disease and Yi = some value indicates that there is a heart disease.

Let $P_r[Y_i = 1 | X_{1i}, X_{2i}, X_{3i},\ldots, X_{ni}] = \pi_i$, $P_r[Y_i = 0 | X_{1i}, X_{2i}, X_{3i},\ldots, X_{ni}] = 1 - \pi_i$ be considered the probabilities of $Y_i = 1$ and $Y_i = 0$ for existing set of given independent variables (X_{1i}, X_{2i},…, X_{ni}). Therefore, the Logistic Regression equation is established as follows

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta X_{1i} + \gamma X_{2i} + \ldots\ldots\ldots\delta X_{ni} \quad (1)$$

The decision tree algorithm's confusion matrix is given in Fig. 1(a), which shows 29 correct test values out of 31. The obtained accuracy for the logistic regression is 93.55%, while Dwivedi et al. [11] have achieved an accuracy of 85%. The calculated specificity, sensitivity and precision are 88.88%, 100% and 86.66%, respectively. The output plot for the predicted and actual target is given in Fig. 1(b).

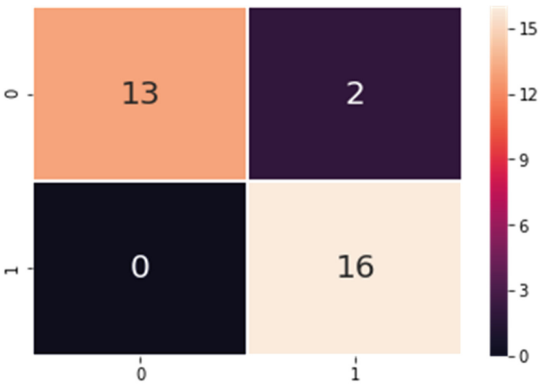

Fig. 1(a)

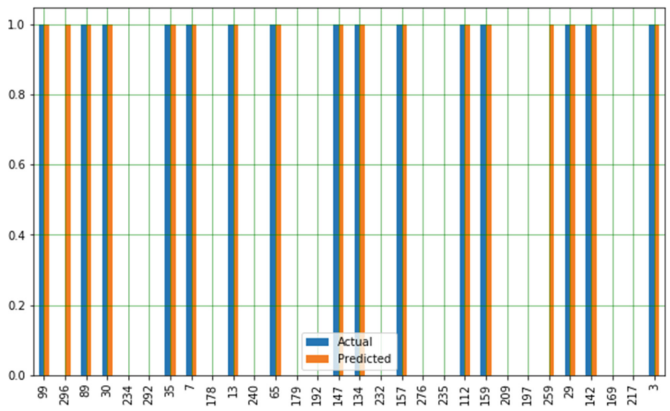

Fig. 1(b)

Fig. 1. (a) confusion matrix (b) data frame of actual and predicted values of Logistic Regression

4.1.2 Decision Tree (DT)

Decision trees are tree-structured algorithms used for both classification and regression categories of problems. The decision tree focuses on generating the output based on the majority and the entropy of the data attributes. When the raw data is given to the model from a given dataset, the algorithm will classify which class the instance belongs to regarding the target variable. Considering the functionalities of the attributes and also based on the input data here, the algorithm divides the data into several subsets.

Initially, the target variable is chosen from the dataset, and then the information gain (IG) is calculated using the following relation

$$IG = -\frac{P}{(P + N\ log\left(P/(P+N)\right)} - \frac{N}{P + N\ log\left(N/P+N\right)} \tag{2}$$

where P is for the positive instance and N for the negative instance.

For the remaining variables, the entropy is calculated accordingly. Based on the classification and the calculations of IG and Entropy, the attributes are structured in a tree format. Thus, the output will be generated from the tree through respective operations for retrieval.

The obtained accuracy is 93.54%, whereas Patel et al. [6] have got accuracy for the decision tree is 80.263%. The specificity, sensitivity and precision are 88.8%, 100% and 86.6%, respectively.

4.1.3 K – Nearest Neighbours (KNN)

Out of many simplest Supervised Machine Learning algorithms, K- Nearest Neighbors algorithm is often used in many applications. It is very widely used for the classification of data. In K-NN, the training data is not learned but memorized. Based on the similarities of the data and features, it stores all the cases that are available, classifies the set of new instances, and generates the outputs accordingly. It is based on feature similarity. The process of choosing the proper figure to justify the value of K is called Parameter Tuning. This is employed to improve the accuracy.

To choose a value of K, one of the following is followed as per the dataset.

- sqrt[n], where n represents the total number of data points available.
- Confusion between the two classes of data can be avoided by choosing the value of K as an odd figure.

The K value will be compared with the nearby the distance calculated for different attributes

$$d = \sqrt{(x_{21} - x_{11})^2} + (x_{22} - x_{12})^2 + \ldots\ldots\ldots \tag{3}$$

The confusion matrix and data frame show that the KNN algorithm predicts 21 correct values out of 31, and the accuracy is 67.7%, which is significantly less compared to all techniques. Since KNN is a distance-based algorithm, it doesn't go easy to work with when there is a large amount of data. The cost to measure the distance from a new point to an existing attribute is very high, which will degrade the accuracy of an algorithm.

Also, it doesn't work well with the high number of dimensions. In comparison, Patel et al. [6] got the KNN accuracy of 90.789%, and Dwivedi et al. [11] obtained an accuracy of 80%. The calculated sensitivity, specificity, and precision in this work are 72%, 65%, and 53.33%, respectively.

4.1.4 Naïve Baye's Classifier Model (NB)

The Naïve Baye's Theorem/ Algorithm is a Machine Learning algorithm used for classification problems. It presumes that the occurrence of a specific feature or outcome is independent of the event of other elements or attributes. This algorithm defines the contribution of the probability of an object with specific outcomes. Baye's Theorem is stated in the following equation

$$P(y/X) = \frac{P(X/y)P(y)}{P(X)} \tag{4}$$

P(X/Y) denotes the posterior probability, P(X) denotes the class prior probability, P(Y) denotes the predictor prior probability, P(Y/X) denotes the likelihood, probability of predictor.

$$P(X/y) = P(x1/y) * P(x2/y) * .....P(x2/y) \tag{5}$$

$$P(y/X)\,\alpha\, P(X/y)P(y) \tag{6}$$

$$P(y/X)\,\alpha\, P(y)\prod_{i=1}^{n} P(x_i/y) \tag{7}$$

$$P(y/X) = \arg\max_{y}\,[P(y)\prod_{i=1}^{n} P(x_i/y)] \tag{8}$$

Naïve Baye's classifier gives the best accuracy of 96.77%. The confusion matrix and data frame plot show the 30 correct predicted values out of 31 test values and only one false negative value. This algorithm is best suitable to predict the heart disease due to all the input attributes are independent and only two output variables, whether 0 or 1. So it is easy to classify and predict the correct values for the Naive Baye's theorem. Table 1 shows Naive Baye's classifier gets 93.55% sensitivity, 100% specificity, and 100% precision.

4.1.5 Random Forest Model (RF)

Random Forest model is an algorithm that works by building a set of multiple decision trees in the training phase and the outcome, as the decision of the majority of the trees in set is chosen by the model as the final decision. The following steps are exercised for constructing multiple nodes during the training phase of the model.

Step-1: Select 'K' data points randomly from the existing training set.

Step-2: Building a set of decision trees that are related to the set of training data points.
Step-3: Choose how many decision trees the model needs to build.
Step-4: Repeat 1 & 2.
Step-5: For a new set of data or instances, find the predicted outcomes of each decision tree. Relate the novel data points to the category that bags the majority of the outcomes and accuracy.

The final decision is made based on the outcome of the majority of the outcomes of the trees.

The accuracy of RF is 96.77%, in which 30 predicted outcomes are correct out of 31 tested values. The confusion matrix shows only one false positive value. Hence the precision and specificity are 93.33% and 94.11%, respectively. The sensitivity of this algorithm is 100%. However, Patel et al. [6] got an accuracy of 86.84% for the RF algorithm. The analysis shows RF is suitable for the heart disease prediction due to RF will take the output of high majority outcomes of decision trees, whether 0 or 1.

4.1.6 Support Vector Machine Model (SVM)

Support Vector Machine model is a Supervised Machine Learning algorithm, and also is a popularly used classification technique. SVM is applicable and effective for the data that is linearly separable. SVM classifies two classes of the data using a typical Hyperplane. The hyperplane should possess the largest margin value in a high dimensional space to have the access to separate the given data into respective classes. This margin between the two classes is to represent the longest valued distance between the nearest data points of these classes. The sequence or the steps in the implementation of SVM are as follows,

Step 1: Loading the dataset and pre-processing.
Step 2: Exploring the data of the dataset.
Step 3: Splitting the data into classes.
Step 4: Employing the model.
Step 5: Evaluation of the model.

The accuracy for the supporting vector machine is 93.55% in which it is predicted 29 correct outcomes out of 32, and only two false positives have come. The sensitivity for this SVM is 100%, and specificity and precision are 88.88% and 86.66%, respectively. The SVM may be suitable for the heart disease prediction due to fewer classifications such as the healthy or unhealthy heart.

The three major steps have been taken after the execution of all algorithms

Step 1: Compare and get the maximum accuracy score
Ex: predictionout = algolist[max(algo_list.keys())]
Step 2: Maximum accuracy score to prediction function of particular ML algorithm
Ex: algo_list = {score_lr = prediction_lr}.
Step 3: Execution of ML algorithm for the given input data
Ex: prediction_lr = lr.predict(inputdata_reshaped)

Table 1. Output Scores of the algorithms

Algorithm	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)
Logistic Regression	93.55	100	88.88	86.66
Naïve Baye's Theorem	96.77	93.75	100	100
K - Nearest Neighbours	67.74	72	65	53.33
Decision Tree	93.55	100	88.88	86.66
Artificial neural network	84.62	89.47	81.13	77.27
Support Vector Machine	93.55	100	88.88	86.66
Random Forest	96.77	100	94.11	93.33

5 Conclusion

The random forest and Naive Baye's algorithm show good accuracy (96.77%) for the UCI data set. However, this work will suit any other data set due to the output is from the best ML algorithm, which has high accuracy. This approach can be trusted because the obtained accuracy is greater than 95%, which offers the best outcome for the input attributes. Using these techniques and computer-predicted values, we can classify the disease fast with reduced cost. This kind of application finds use in medical and biomedical fields for computerized and faster studies of diseases.

References

1. Seckeler MD, Hoke TR. The worldwide epidemiology of acute rheumatic fever and rheumatic heart disease. Clin Epidemiol. 2011;3:67.
2. Ramalingam VV, Dandapath A, Raja MK. Heart disease prediction using machine learning techniques: a survey. Int J Eng Technol. 2018;7(2.8):684–7.
3. Patel J, TejalUpadhyay D, Patel S. Heart disease prediction using machine learning and data mining technique. Heart Dis. 2015;7(1):129–37.
4. Fatima M, Pasha M. Survey of machine learning algorithms for disease diagnostic. J Intell Learn Syst Appl. 2017;9:1–16.
5. Senthilkumar M, Chandrasegar T, Gautam S. Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. IEEE Access. 2019;7: 81542–81554.
6. Sandhya Y. Prediction of Heart Diseases using Support Vector Machine. Int. J. Res. App. Sci. & Engg. Tech. 2020;8: 126–135.
7. Devansh S, Samir P, Santosh Kumar B. Heart Disease Prediction using Machine Learning Techniques. SN Computer Science. 2020;1:345, 1–6.
8. Durairaj M, Revathi V. Prediction of heart disease using back propagation MLP algorithm. Int. J. Sci. Technol. Res., 2015;8:235–239.
9. Krishnaiah V, Narsimha G, Subhash N. Heart disease prediction system using data mining techniques and intelligent fuzzy approach: A review. Int. J. Comput. Appl. 2016;136:43–51.
10. Kumar PS, Anand D, Kumar VU, Bhattacharyya D, Kim TH. A computational intelligence method for effective diagnosis of heart disease using genetic algorithm. Int. J. Bio-Sci. Bio-Technol. 2016; 8, 363–372.

11. Dwivedi AK. Performance evaluation of different machine learning techniques for prediction of heart disease. Neural Comput Appl. 2018;29(10):685–693.