Chapter 8
Machine Learning Algorithms with ROC Curve for Predicting and Diagnosing the Heart Disease

R. Kannan and V. Vasanthi

Abstract Heart diseases are now becoming the leading cause of mortality in India with a significant risk of both males and females. According to the Indian Heart Association (IHA), four people die of heart diseases every minute in India and the age-groups are mainly between 30 and 50. The one-fourth of heart failure morality occurs to people less than 40. A day in India nine hundred people dies below the age of 30 due to different heart diseases. Therefore, it is imperative to predict the heart diseases at a premature phase with accuracy and speed to secure the millions of people lives. This paper aims to examine and compare the accuracy of four different machine learning algorithms with receiver operating characteristic (ROC) curve for predicting and diagnosing heart disease by the 14 attributes from UCI Cardiac Datasets.

Keywords Machine learning algorithms · Gradient boosting · ROC curve Heart disease

8.1 Introduction

The explosive growth of health-related data presented unprecedented opportunities for improving health of a patient. Heart disease is the dominant reason for mortality in India, Australia, UK, USA, and so on. Machine learning involves and activates the uncovering new trends in healthcare industries. By using machine learning technique we can conduct the research from different aspects between heart diseased persons and healthy person based on their existing medical considerable datasets. Tremendous approach in this study of all cardiac-related disease classification is done to find the

R. Kannan · V. Vasanthi (✉)
Department of Computer Science, Rathinam College of Arts & Science, Coimbatore, Tamil Nadu, India
e-mail: vasanthi.cs@rathinam.in

R. Kannan
e-mail: dschennai@outlook.com

N. B. Muppalaneni et al., *Soft Computing and Medical Bioinformatics*, SpringerBriefs in Forensic and Medical Bioinformatics
https://doi.org/10.1007/978-981-13-0059-2_8

disguised medical information. It accelerated the establishment of vital knowledge, e.g., patterns, different dimensions for identifying relationships amidst medical factors interconnected with heart diseases.

By using some machine learning techniques, heart disease prediction can be made simple by using various characteristics to find out whether the person suffers from heart attack or not, and it also takes less time to predict and improve the medical diagnosis of diseases with good accuracy and minimizes the occurrence of heart attack. It assists to resolve the hidden reason and diagnose the heart diseases efficiently even with the uncertainties and inaccuracies. This paper emphasizes the machine learning algorithms such as logistic regression, Random Forest, boosted tree, stochastic gradient boosting and support vector machines that are used to confirm the best prediction technique in terms of its accuracy and error rate on the specific dataset.

8.1.1 Heart Diseases

Heart diseases are life-frightening diseases, and it should be contemplating as a global health precedence. Moreover, heart diseases reside a great stress on patients, caretaker, and healthcare systems. For the present, almost 30 million people worldwide are living with the heart diseases such as patients affected by heart dieses because of cholesterol deposits, high blood sugar, poor in hygiene, physically inactive, unhealthy diet, smoking and hand change smoking, being overweight, high blood pressure, viral infection compared with the survival rates worse than any other diseases. Although the survival which is globally 48.9 million, 50 people affected by heart defects by birth defects. The signs and symptoms of 51 heart diseases mentioned below are caused because of above highlighted major reasons.

- Chest pain
- Shortness of breath
- Sweating
- Nausea
- Irregular heartbeat
- Throat or jaw pain
- A cough that won't quit.

8.1.2 Machine Learning (ML)

Machine learning is the technique to allow the computers to learn and predict automatically for achieving the difficult jobs whose processes cannot be simply described by humans, for example, self-driving car, Netflix movies rating, Amazon online recommendations, diagnosis in medical imaging, autonomous robotic

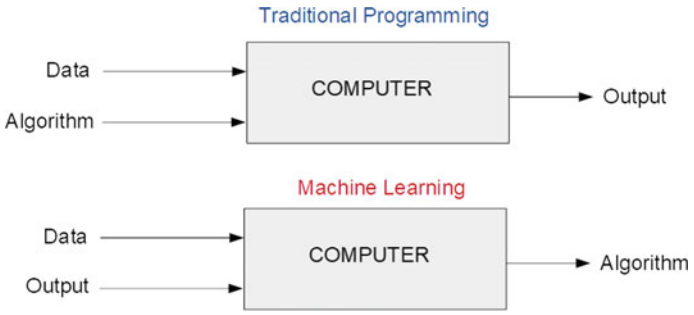

Fig. 8.1 Overview of machine learning

surgery. The machine learnings are of three different types such as supervised learning, unsupervised learning, and reinforcement learning. The following Fig. 8.1 shows the contradiction between the traditional programming and machine learning to provide the best results.

8.1.3 Supervised Machine Learning Algorithms

Supervised learning is the one type of machine learning algorithms used to learn and predict with labeled training data. The training data contains the set of training examples, and each example has the input object and the desired output value. The supervised learning algorithms examine the training data, produce the complete function, which can be used for mapping new examples, and correctly determine the class labels for hidden examples. Here the below list shows the supervised machine learning algorithms [1].

- Linear regression
- Logistic regression
- Decision tree
- Support vector machine (SVM)
- Random Forest
- And so on.

8.1.4 Unsupervised Machine Learning Algorithms

In contrast, unsupervised machine learning algorithms are used to learn and predict without labeled training data. Here the below list shows the unsupervised machine learning algorithms [1].

- Clustering
- Anomaly detection
- Approaches for learning latent variable models
- And so on.

8.1.5 Reinforcement Machine Learning Algorithms

Reinforcement machine learning algorithms differ from supervised machine learning and unsupervised machine learning. Reinforcement machine learning algorithms allow technologies and software agents to automatically determine the ideal behavior within a specific context in order to maximize its performance. Simple reward response is required for the agent to learn which action is best; this is known as the reinforcement signal [2].

8.2 Materials and Methods

In this section, we have introduced the heart disease Cleveland dataset and explained the four different machine learning models with ROC curve for the heart disease predictions with diagnoses.

8.2.1 The Cleveland Dataset

Machine learning models need a certain amount of data to lead an adequate algorithm. We can collect the necessary datasets from repository of healthcare industries and third-party data sources. It is enabling comparative effectiveness in the research done by producing unique and powerful machine learning algorithms [3].

In this paper, we have obtained 303 records with 14 set of variables and divided the data into training (70%) and testing (30%) from the Cleveland dataset. Percentage of heart disease should not be the same in training and testing data. We have listed the 14 attributes (Table 8.1) below.

The variable we want to predict is Num with value 0: <50% diameter narrowing and value 1: >50% diameter narrowing. We assume that every value with 0 means heart is normal for patient and 1,2,3,4 means heart disease.

Table 8.1 Heart disease attributes

Variable name	Description
Age	Age in years
Sex	Sex, 1 for male, 0 for female
CP	Chest pain type (1 = typical angina; 2 = atypical angina; 3 = non-angina pain; 4 = asymptomatic)
Trestbps	Resting blood pressure
Chol	Serum cholesterol in mg/dl
Fps	Fasting blood sugar larger 120 mg/dl (1 true)
Restecg	Resting electrocardiographic results (1 = abnormality, 0 = normal)
Thalach	Maximum heart rate achieved
Exang	Exercise-induced angina (1 yes)
Oldpeak	ST depression induce. Exercise relative to rest.
Slope	Slope of peak exercise ST
CA	Number of major vessel
Thal	No explanation provided, but probably thalassemia
Num	Diagnosis of heart disease (angiographic disease status) 0 (<50% diameter narrowing) 1 (>50% diameter narrowing)

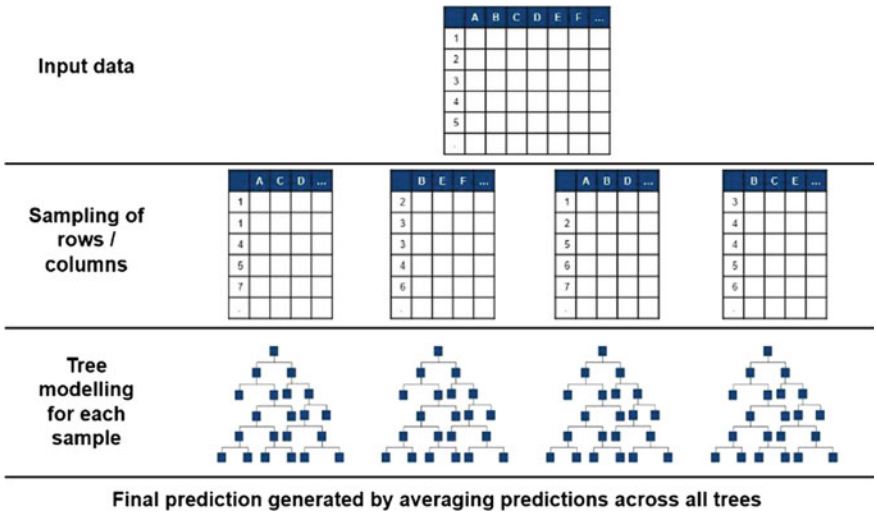

Fig. 8.2 Decision tree and Random Forest

8.2.2 Random Forests

Random Forest is a supervised machine learning algorithm. We can see it from its name, which creates a forest by random tree. It contains a direct relationship between the number of trees in the forest (Fig. 8.2 shows the relationship) and the

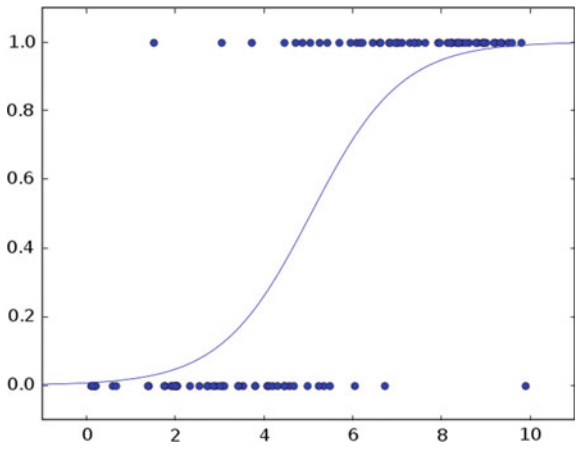

Fig. 8.3 Logistic regression

results it can get [4]. It can be used for both classification and regression tasks. Over fitting is one critical problem that may make the results worse, but for Random Forest algorithm, if there are enough trees in the forest, the classifier will not over fit the model. The third advantage is the classifier of Random Forest can handle missing values, and the last advantage is that the Random Forest classifier can be modeled for categorical values.

8.2.3 Logistic Regression

Logistic regression algorithm is a regression and classification method for evaluating the dataset in which it contains one or more independent variables that conclude an outcome. The outcome is measured with a divided variable (in which can be two possible outcomes) [5]. The following Fig. 8.3 shows the two possible outcomes from logistic regression.

8.2.4 Gradient Boosting

Gradient boosting is one of the best supervised machine learning algorithms for regression and classification problems. Gradient boosting algorithm is the form of an ensemble of weak prediction models likely decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function [6]. The following Fig. 8.4 shows the gradient boosting algorithms.

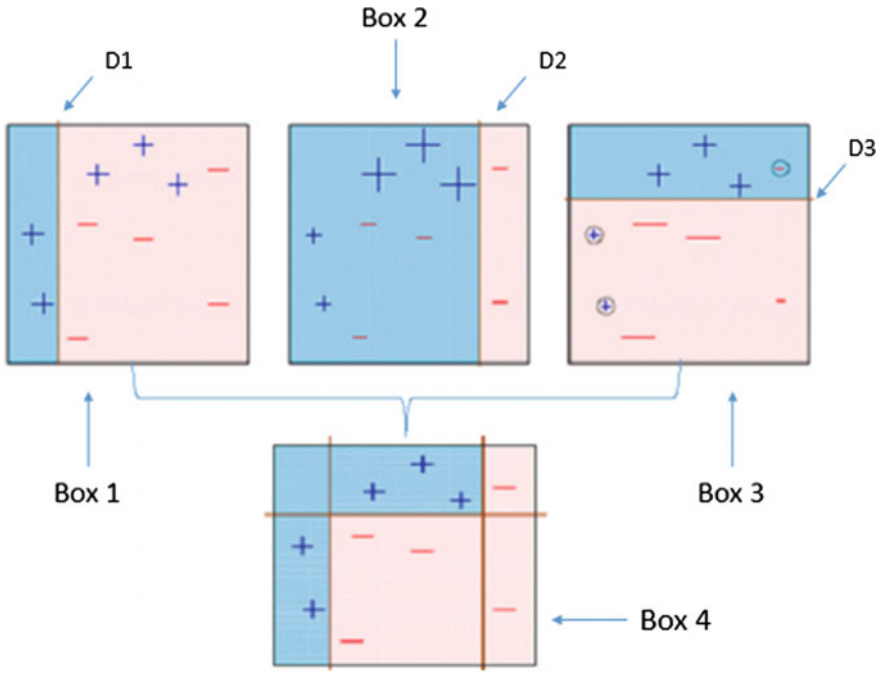

Fig. 8.4 Gradient Boosting

8.2.5 Support Vector Machine

Support vector machine (SVM) is a supervised machine learning algorithm. SVM can be used for both the classification and regression problems. However, it is common for classification problems [7]. In this algorithm, we plot each data item as a point in *n*-dimensional space (where *n* is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyperplane that differentiates the two classes very well (Fig. 8.5 shows an example).

8.2.6 ROC Curve

The receiver operating characteristic (ROC) plot is a most popular measure for evaluating classifier performance [7]. The ROC plot is based on two basic evaluation measures—specificity and sensitivity. Specificity is a performance measure of the negative part, and sensitivity is a performance measure of the positive part. The majority of machine learning models produce some kind of scores in addition to

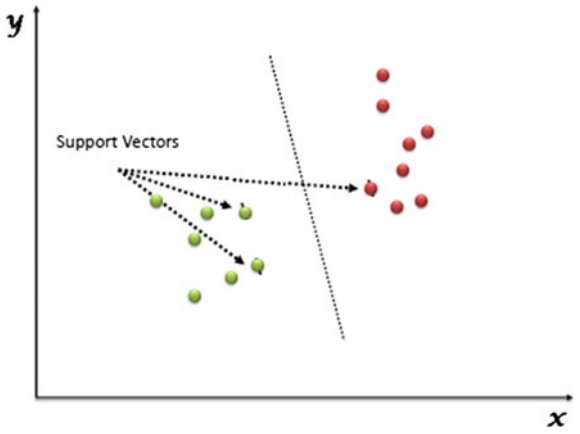

Fig. 8.5 Support vector machine

predicted labels. These scores can be discriminant values, posterior probabilities, and so on. Model-wide evaluation measures are calculated by moving threshold values across the scores.

8.2.7 Software Tools

As time goes on, there are best tools and software that are available to predict the heart diseases in the market. Here, we used the most favorable open-source software name called as "R programming" for this paper. R is the most popular statistical computing and visualization software package in the worldwide, which used in number of growing healthcare industries, commercial, government organizations, academics, and research.

8.3 Experimental Evaluation

We have applied the four machine learning algorithms such as logistic regression, Random Forest, stochastic gradient boosting, and support vector machine with the attributes taken from the UCI heart diseases dataset. Eventually, we predicted the four different results from different tuning parameters and compared the best model of each machine learning algorithm with ROC curve. Here Table 8.2 shows the results comparison of ACU and accuracy between models.

The results for the compression of ACU and accuracy between the machine learning models represented as graphical visualization are shown in Table 8.2 and Fig. 8.6.

Table 8.2 Comparison of ACU and accuracy between models

Algorithms	ACU	Accuracy
Logistic Regression	0.9161585	0.8651685
Random Forest	0.8953252	0.8089888
Stochastic gradient boosting	0.9070122	0.8426966
Support vector machine	0.882622	0.7977528

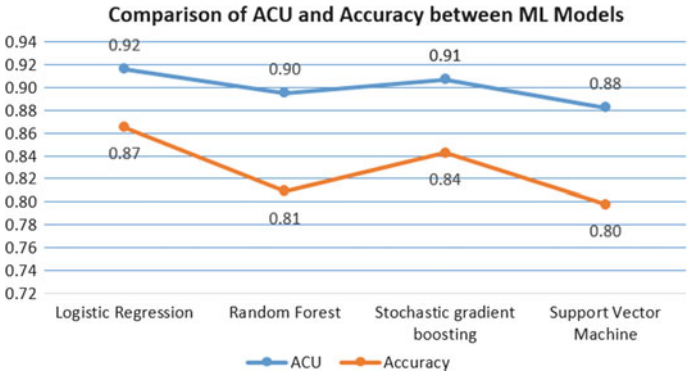

Fig. 8.6 Visualizing the results of ACU and accuracy models

8.4 Results and Conclusion

Fourteen predictor variables from the UCI heart disease dataset are used to predict the diagnosis of heart disease. The performance of the four different machine learning algorithms, such as logistic regression, stochastic gradient boosting, Random Forest, and support vector machines are compared with the accuracy obtained from them.

Thirty percent of the data is hold out as a testing dataset that is not seen during the training stage of the data. During the training of boosted and support vector machines, tenfold cross-validation is used to maximize the ROC (parameter tuning) and select the appropriate models. A comparison of the area under the ROC and the accuracy of the model predictions shows that logistic regression performs best and it can predict with 0.87% of accuracy.

8.5 Future Work

This paper has summarized the methods for heart disease prediction along with innovative machine learning algorithms. As a future work, we have planned to predict the heart diseases using tensor flow of deep learning algorithms with more dataset. This deep learning tensor flow will automate and increase the process of prediction in terms of speed.

References

1. What is Machine Learning? A definition. www.expertsystem.com/machine-learning-definition/
2. Elshaw M, Mayer NM (2008) Reinforcement learning edited by Cornelius Weber
3. Heart Disease Data Set. http://archive.ics.uci.edu/ml/datasets/heart+Disease
4. How Random Forest Algorithm Works in Machine Learning. https://medium.com/@Synced/how-random-forest-algorithm-works-in-machine-learning-3c0fe15b6674
5. The Pennsylvania State University, 'STAT 504 | Analysis of Discrete Data'. https://onlinecourses.science.psu.edu/stat504/node/149
6. Boosting Machines. https://github.com/ledell/useR-machine-learning-tutorial/blob/master/gradient-boosting-machines.Rmd
7. Introduction to the ROC (Receiver Operating Characteristics) plot. https://classeval.wordpress.com/introduction/introduction-to-the-roc-receiver-operating-characteristics-plot
8. Practical Machine Learning, https://www.coursera.org/learn/practical-machine-learning
9. Dr. Brownlee J (2017) Master machine learning algorithms, eBook 2017. https://machinelearningmastery.com
10. Daumé III H, A course in machine learning. https://ciml.info
11. Understanding Support Vector Machine algorithm. https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code