Advances and Applications in Mathematical Sciences
Volume 21, Issue 7, May 2022, Pages 3785-3794

EFFICIENCY OF MACHINE LEARNING ALGORITHMS IN PROPHECY OF CORONARY ILLNESS: A PERFORMANCE ANALYSIS

A. P. LAVANYA

Assistant Professor
Department of Computer Science and Engineering
Sona College of Technology
Salem, Tamilnadu, India
E-mail: lavanya.cse@sonatech.ac.in

Abstract

Heart disorders, also known as cardiovascular diseases, have been the main source of death in late many years and have emerged as the most life-threatening disease. To automate the examination of big and complicated data, machine learning methods and techniques have been used to a variety of medical datasets. There are several supervised learning algorithms available, including SVM, Naive Bayes, Logistic Regression, Decision Tree, Random Forest, Light GBM, XG Boost. The algorithms were applied to compare the performance and accuracy of the ML algorithms in coronary illness prophecy.

1. Introduction

One compelling cause of death worldwide is cardiovascular diseases. Heart disease is associated with a variety of symptoms, making it more challenging to identify it quickly and accurately. In today's healthcare, the key difficulty is to provide high-quality services and precise diagnoses. The proposed research aims to detect these heart illnesses early on to lessen catastrophic consequences. Most of the information in a medical database is discrete. As a result, making decisions using discrete data becomes a difficult and time-consuming process. Machine Learning succeeds at processing huge, sensible datasets. The significance of machine learning in detecting hidden discrete patterns and analysing the information is critical. Following data

2020 Mathematics Subject Classification: 34Bxx, 76-10, 80A30.
Keywords: Logistic Regression, Decision Tree, Random Forest, Light GBM.
Received December 13, 2021 Accepted January 14, 2022.

analysis, machine learning techniques contribute to the prediction and early diagnosis of cardiac disease. To allocate the weight to each attribute, doctor's knowledge is required. The characteristic with the greatest impact on disease prediction is given more weight. As a result, it appears reasonable to attempt to use the expertise and experience of several specialists gathered in databases to aid the diagnosis process. It also gives healthcare providers with an additional source of information to help them make decisions. This proposed work deals with analysing the efficiency of various ML techniques in for forecasting heart disorders.

2. Literature Survey

In [1] Mamun Ali et al. presented the performance analysis of various supervised machine learning algorithms in predicting heart disease. Parthiban et al. [12] analysed coronary illness in diabetic patients utilizing programmed learning techniques. Utilizing the Naive Bayes calculation gives 74% exactness. SVM gives the most noteworthy precision of 94.60%. A. Malav et al. [5] propose a compelling mixture algorithmic methodology for anticipating coronary illness, to decide and extricate obscure information about coronary illness utilizing the crossbreed technique joining the *K*-means clustering calculation and the artificial neural network. The proposed model accomplishes a precision of 97%. In [7], the dynamic course of coronary illness is successfully analysed by the Random Forest calculation. In [10] view of the likelihood of decision help, coronary illness is anticipated. As a result, the author inferred that the decision tree produces better accuracy and some of the time the exactness is comparable in Bayesian. In [9] Jaymin et al. presented J48 tree technique worked better for coronary disease prediction by comparing J48, LM Tree algorithm and RF algorithm. In [2] Manoj Diwakar et al. discussed the use of machine learning and image synthesis to detect heart disorder. In [4] Shylaja et al. compared various classification and clustering algorithms in predicting the heart disease and found that SVM worked well out of ANN, SVM, Decision Tree, RIPPER, Naive Bayes and KNN.

3. Experimental Environment

3.1 Dataset. In this proposed idea, a coronary illness dataset was

engaged to complete our expected model. The source of the dataset is Kaggle which consists of 303 individual's information. There are 14 attributes in this dataset. "Age, Sex, Chest-pain type, Resting Blood Pressure, Serum Cholesterol, Fasting Blood Sugar, Resting ECG , Max heart rate achieved, Exercise induced angina, ST depression induced by exercise relative to rest, Peak exercise ST segment, Number of major vessels (0-3) colored by flourosopy, Thal, Diagnosis of heart disease" [1]

3.2 Algorithms and Techniques

3.2.1 Support Vector Machine SVM is one of the supervised machine learning techniques that can be utilized for classification and regression. Every information is plotted as a point in n-dimensional space (where n is the quantity of features you have), with the worth of each element being the worth of a specific direction in the SVM calculation. Then, at that point, we achieve grouping by finding the hyper-plane that obviously recognizes the two classes.

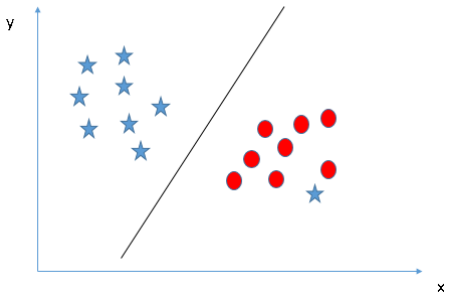

Figure (i). SVM.

3.2.2 Naive Bayes. The classification algorithm is based on Bayes Theorem with the assumption of predictor independence. A Naive Bayes classifier, in basic terms, accepts that the presence of one component in a class is disconnected to the presence of some other element.

Bayes Theorem.

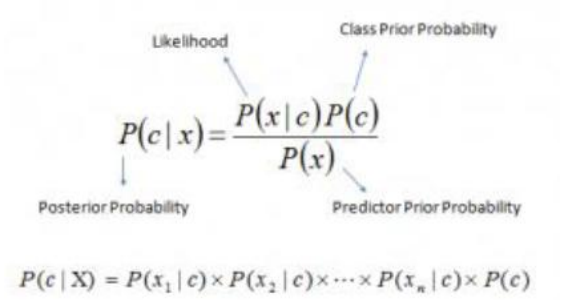

Figure (ii). Bayes Theorem.

We can ascertain the probability of A happening assuming B has effectively happened utilizing Bayes theorem. The evidence is B, and the hypothesis is A. In this situation, the predictors/features are supposed to be independent. That is, the presence of one attribute has no bearing on the other. As a result, it is said to be naïve.

3.2.3 Logistic Regression. One of the supervised machine learning techniques, Logistic regression may be used to estimate the likelihood of a given class or occurrence. When the data is linearly separable and the result is binary or dichotomous, this process is employed. The sigmoid function is used in logistic regression to model the data.

$$f(x) = \frac{1}{1 + e^{-(x)}} \tag{1}$$

3.2.4 Decision Tree. The ID3 (by Quinlan) algorithm is the basic method used in decision trees. The ID3 method uses a top-down, greedy technique to create decision trees. The procedure of the algorithm are as follows: -Choose the best attribute A Assign A as the NODE's decision attribute (test case) create a new descendant of the NODE for each value of A Assign each descendent node leaf to the training examples. If all of the instances have been correctly classified, STOP; otherwise, iterate over the new leaf nodes.

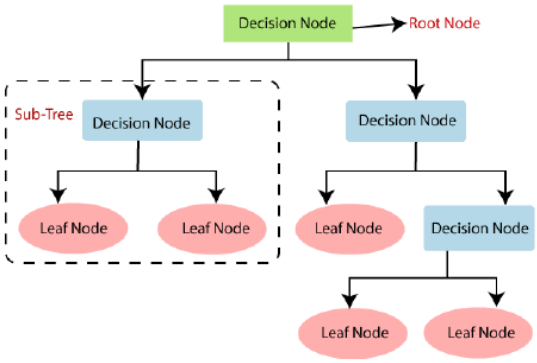

Figure (iii). Decision Tree.

3.2.5 Random Forest. Random Forest is one of the supervised learning algorithms that is habitually used to take care of classification and regression. It develops decision trees from different examples, involving most of decisions in favour of order and the middle for regression. One of the important qualities of this algorithm is that it can deal with informational indexes including both continuous and categorical values, as in classification and regression separately. For classification problem, it gives better outcomes.

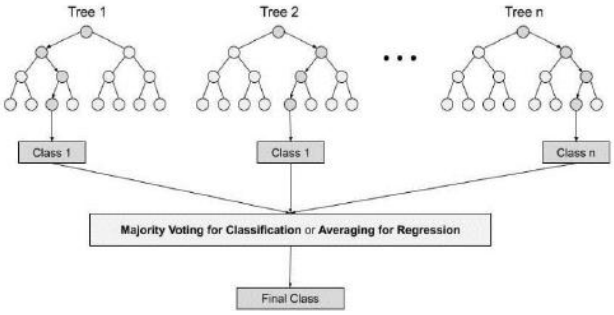

Figure (iv). Random Forest.

3.2.6 XG Boost. Ensemble learning is a method of machine learning that integrates the predictions of multiple models. Boosting algorithms distinguish from other ensemble learning techniques in that they transform a series of weak models into increasingly powerful models. Gradient boosting techniques select how to develop a more powerful model based on the gradient of a loss function that represents the model's performance. The XG Boost algorithm is

a gradient boosting algorithm, which is a common ensemble learning technique. The XG Boost algorithm works with decision trees, which are models that build a graph that analyses the input using various "if" statements (vertices in the graph). The following "if" condition and eventual prediction are influenced by whether the "if" condition is satisfied. To develop a stronger model, XG Boost the Algorithm gradually adds more and more "if" conditions to the decision tree.

3.2.7 Light BGM. It is a gradient boosting system that utilizes tree-based learning strategies, which are viewed as an extremely powerful algorithm. It is thought to be a speedy handling algorithm. While the trees of different calculations grow evenly, the Light GBM algorithm develops upward, and that implies it develops leaf-wise while different calculations develop level-wise. To develop, Light GBM chooses the leaf with the best misfortune. While extending a similar leaf, it can diminish misfortune in excess of a level insightful technique. As a result of its process power and capacity to convey discoveries rapidly, LightGBM is named "Light". It utilizes less memory to work and can deal with enormous volumes of information.

4. Performance Evaluation Metrics

The dataset was subjected to seven classification algorithms in order to determine the best performing method based on accuracy and other statistical characteristics. For each algorithm, a confusion matrix was plotted to calculate the following measures:

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \tag{2}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \tag{3}$$

$$\text{F Meassure} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

$$\text{Sensitivity} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \tag{5}$$

$$\text{Specificity} = \frac{\text{True Negative}}{\text{True Negative} + \text{False Positive}} \tag{6}$$

Sensitivity is the metric that assesses a model's capacity to anticipate true positive of each accessible classification.

The classifier's ability to correctly discriminate negative outcomes is known as specificity.

The percentage of successfully categorized occurrences by a classifier is known as accuracy. To compare the efficiency of different algorithms, different statistical metrics were utilized, such as precision, recall, and *f*-measure. Precision is a good statistic to use when the proposed machine learning model needs to be validated based on the projected and actual results. It ascertains the extent of expected positives that end up being true positive. Therefore, TP and FP esteems are important. While deciding the quantity of positives that might be sensibly expected, recall is a significant assessment measurement that addresses the proportion of positives that are precisely classified. The TP and FP values are utilized to compute recall. The F-Measure value is a number going from 0 to 1 that shows genuinely critical precision and recall estimations. For a classifier, F-Measure strikes a trade-off among precision and recall.

5. Result and Discussion

A few standard exhibition measurements like precision, recall, accuracy have been considered for the calculation of execution viability of this model. The Confusion matrix obtained for all the seven classification algorithms were shown below:

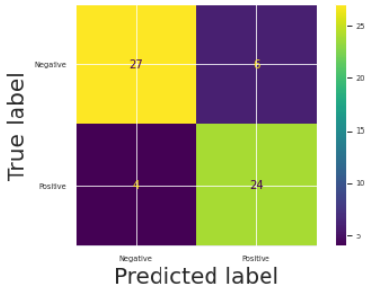

Figure (v). SVM.

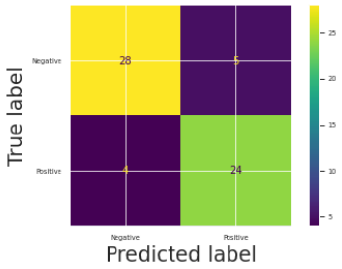

Figure (vi). Naive Bayes.

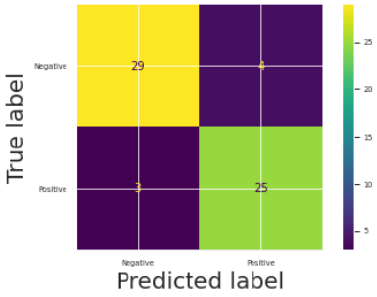

Figure (vii). Logistic.

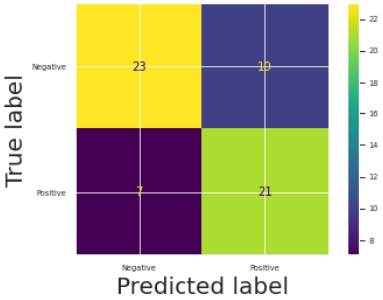

Figure (viii). Decision Tree.

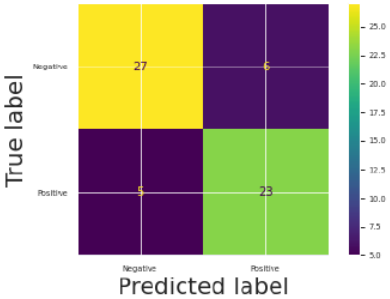

Figure (ix). Random Forest.

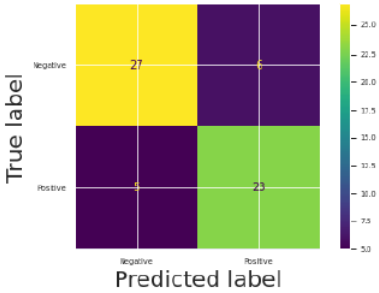

Figure (x). XG Boost.

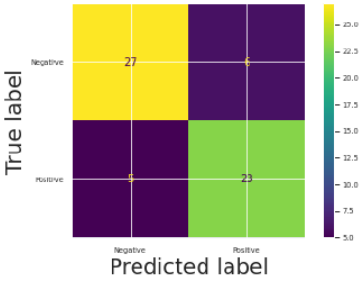

Figure (xi). Light GBM.

Table (i). Precision, Recall and F-Measure.

Classification algorithm	Precision	Recall	F-Measure
SVM	0.80	0.85	0.82
Naïve Bayes	0.82	0.85	0.84

Logistic Regression	0.86	0.89	0.87
Decision Tree	0.67	0.75	0.71
Random Forest	0.79	0.82	0.80
XGBoost	0.80	0.89	0.84
LightGBM	0.85	0.85	0.85

Table (ii). Sensitivity, Specificity and Accuracy

Classification algorithm	Precision	Recall	F Measure
SVM	0.85	0.81	0.83
Naïve Bayes	0.85	0.84	0.85
Logistic Regression	0.89	0.87	0.88
Decision Tree	0.75	0.69	0.72
Random Forest	0.82	0.81	0.81
XGBoost	0.89	0.81	0.85
LightGBM	0.85	0.87	0.86

6. Conclusion

Coronary illness is a dangerous condition that can lead to fatal complications including heart attacks. The significance of information mining and AI procedures could be utilized to decide the presence of infection because of its true capacity for precise illness expectation rate. To explore the viability techniques in prediction of heart disease, we utilized a heart disease dataset. Finding that logistic regression algorithm performed extremely well with 88% accuracy. The goal of the study was to identify the best machine learning approaches among a number of well-known and straightforward algorithms, and it was observed that, as certainly for this dataset, they scored well. Although the use of machine learning technologies is still in its early stages, it appears that it could be a valuable addition to patient care.

References

[1] Md Mamun Ali, et al., Heart disease prediction using supervised machine learning algorithms: performance analysis and comparison, Computers in Biology and Medicine 136 (2021), 104-672.

[2] Diwakar, Manoj, et al., Latest trends on heart disease prediction using machine learning and image fusion, Materials Today: Proceedings 37 (2021), 3213-3218.

[3] S. Mohan, C. Thirumalai and G. Srivastava, Effective heart disease prediction using hybrid machine learning techniques, IEEE Access 7 (2019), 81542-81554.

[4] S. Shylaja and R. Muralidharan, Comparative analysis of various classification and clustering algorithms for heart disease prediction system, Biometrics Bioinf 10(4) (2018), 74-77.

[5] Malav, K. Kadam and P. Kamat, Prediction of heart disease using *k*-means and artificial neural network as a hybrid approach to improve accuracy, International Journal of Engineering and Technology 9(4) (2017).

[6] Sowmiya and P. Sumitra, Analytical study of heart disease diagnosis using classification techniques, in Proc. IEEE Int. Conf. Intell. Techn. Control, Optim. Signal Process. (INCOS), Mar. (2017), 1-5.

[7] M. Fatima and M. Pasha, Survey of machine learning algorithms for disease diagnostic, Journal of Intelligent Learning Systems and Applications 9(01) (2017), 16.

[8] T. Vivekanandan and N. C. S. N. Iyengar, Optimal feature selection using a modified differential evolution algorithm and its effectiveness for prediction of heart disease, Comput. Biol. Med. 90 (2017), 125-136.

[9] Patel, Jaymin, Dr. Tejal Upadhyay and Samir Patel, Heart disease prediction using machine learning and data mining technique, Heart Disease 7(1) (2015), 129-137.

[10] C. Rajeswari, B. Sathiyabhama, S. Devendiran and K. Manivannan, Bearing fault diagnosis using wavelet packet transform, hybrid PSO and support vector machine, Procedia Engineering 97(1) (2014), 1772-1783.

[11] A. T. Sayad and P. P. Halkarnikar, Diagnosis of heart disease using neural network approach, Int. J. Adv. Sci. Eng. Technol. 2 (2014), 88-92.

[12] G. Parthiban and S. K. Srivatsa, Applying machine learning methods in diagnosing heart disease for diabetic patients, International Journal of Applied Information Systems 3(7) (2012), 2249-0868.

[13] M. Marimuthu, G. Vidhya, J. Dhaynithi, G. Mohanraj, N. Basker, S. Theetchenya and D. Vidyabharathi, Detection of Parkinson's disease using Machine Learning Approach. Annals of the Romanian Society for Cell Biology (2021), 2544-2550.

[14] K. C. Tan, E. J. Teoh, Q. Yu and K. C. Goh, A hybrid evolutionary algorithm for attribute selection in data mining, Expert Systems with Applications 36(4) (2009), 8616-8630.

[15] https://medium.com/analytics-vidhya/heart-disease-prediction-using-knn-algorithm-be78f800e2a9.