American Journal of Computer Science and Technology
2022; 5(3): 146-154
http://www.sciencepublishinggroup.com/j/ajcst
doi: 10.11648/j.ajcst.20220503.11
ISSN: 2640-0111 (Print); ISSN: 2640-012X (Online)

Heart Disease Prediction Using Machine Learning Techniques

Mohammed Khalid Hossen

Department of Computer Science and Engineering, Sylhet Agricultural University, Sylhet, Bangladesh

Email address:
khalid@sau.ac.bd

To cite this article:
Mohammed Khalid Hossen. Heart Disease Prediction Using Machine Learning Techniques. *American Journal of Computer Science and Technology*. Vol. 5, No. 3, 2022, pp. 146-154. doi: 10.11648/j.ajcst.20220503.11

Received: June 27, 2022; **Accepted**: July 13, 2022; **Published**: July 20, 2022

Abstract: Machine learning and artificial intelligence have been found useful in various disciplines during the course of their development, especially in the enormous increasing data in recent years. It can be more reliable for making better and faster decisions for disease predictions. So, machine learning algorithms are increasingly finding their application to predict various diseases. Constructing a model can also help us visualize and analyze diseases to improve reporting consistency and accuracy. This article has investigated how to detect heart disease by applying various machine learning algorithms. The study in this article has shown a two-step process. The heart disease dataset is first prepared into a required format for running through machine learning algorithms. Medical records and other information about patients are gathered from the UCI repository. The heart disease dataset is then used to determine whether or not the patients have heart disease. Secondly, Many valuable results are shown in this article. The accuracy rate of the machine learning algorithms, such as Logistic Regression, Support vector machine, K-Nearest-Neighbors, Random Forest, and Gradient Boosting Classifier, are validated through the confusion matrix. Current findings suggest that the Logistic Regression algorithm gives a high accuracy rate of 95% compared to other algorithms. It also shows high accuracy for f_1-score, recall, and precision than the other four different algorithms. However, increasing the accuracy rates to approximately 97% to 100% of the machine learning algorithms is the future study and challenging part of this research.

Keywords: Machine Learning, Artificial Intelligence, Heart Disease, Linear Regression, Support Vector Machine, K-Nearest-Neighbors, Random Forest, Decision Tree, Gradient Boosting

1. Introduction

Machine learning (ML) is a part of artificial intelligence (AI) that allows a software application to improve its prediction accuracy without being formally programmed. In order to forecast new output values, machine learning algorithms use historical data as input [1]. Machine learning is a significant and diversified field, and its scope and application are expanding daily. For this reason, machine learning has become a crucial competitive differentiation in many organizations. Machine learning includes supervised, unsupervised, and ensemble learning classifiers that are used to predict and find the accuracy of a dataset. ML algorithms can build a model based on sample data called train data to make a decision or prediction [1, 2].

The use of machine learning methods in the medical industry is the subject of the current study, which mainly focuses on mimicking some human activities or mental processes and recognizing diseases from a variety of inputs [3]. The term “heart disease” refers to a group of conditions that affect the heart. According to World Health Organization reports, cardiovascular diseases are now the leading cause of death worldwide, approximately 17.9 million [4, 5]. Many types of research have been studied and performed with various machine learning algorithms to diagnose heart diseases. According to Ghumbre et al., machine learning and deep learning algorithms are applied to predict heart diseases in the UCI dataset [3]. The authors concluded that machine learning algorithms performed better for this analysis. Machine learning techniques for heart disease prediction are published by Rohit Bharti et al., where the article concluded that different data mining and neural system should be used to find the seriousness of HD among patients [4]. Some analysis has been led to think about the implementation of a predictive data mining strategy on the same dataset [5]. Prediction of heart disease using machine learning is studied by Jee S H et al. in

which training and the testing dataset are performed by using a neural network algorithm [6]. K-Nearest Neighbor algorithm is reviewed to diagnose heart disease by Mai Shouman et al. [7]. Some efficient algorithms have been used to detect HD, which shows results that each algorithm has its strength to register the defined objectives [8]. The supervised network has been applied for HD diagnosis, which is studied by Raihan M et al. [9]. This research idea has been broadened and inspired us worldwide by publishing many articles [10-15].

This article will construct an ML predictive model, which will help analyze heart disease regarding the medical history. Data is collected from the UCI repository with patients' medical records and attributes. This dataset would be utilized to predict whether the patients have heart disease or not. To diagnose the HD dataset, this article considers 14 attributes of a patient. It classifies whether the disease is present or not and can help us diagnose diseases with fewer medical treatments [1, 5]. For this study, this article considers various attributes of patients like age, sex, serum cholesterol, blood pressure, exang, etc. Five different ML algorithms such as Logistic Regression (LR), Support vector machine (SVM), K-Nearest-Neighbors (KNN), Random Forest (RF), and Gradient Boosting Classifier (GBC) are applied for the purpose of classification and prediction of heart disease. Many beneficial results are presented in this article. The attributes of the given dataset are trained under these algorithms. Based on the characteristics of the HD dataset, a comparative analysis of algorithms has been studied regarding the accuracy rate. All the selected ML algorithms are efficient by showing their accuracy, which is greater than 80%. The most efficient algorithm is Logistic Regression (LR), which gives us an accuracy rate of approximately 95%. Finally, Logistic Regression (LR) algorithm will be considered to predict and diagnose for heart disease of a patient.

This article is rearranged sequentially. In section 2, the methodology has been discussed. Various ML algorithms are studied briefly in section 3. Results and analysis are shown in section 4. In the result section, algorithms are compared regarding the confusion matrix. Finally, a conclusion and future scope have been drawn in section 5.

2. Methodology

In this section, the method and analysis are described, which is performed in this research work. First of all, the collection of data and selection of relevant attributes are the initial steps in this study. After that, the relevant data is pre-processed into the required format. The given data is then separated into two categories: training and testing datasets. The algorithms are then used, and the given data train the model. The accuracy of this model is obtained by using the testing data. The procedures of this study are loaded by using several modules such as a collection of data, selection of attributes, pre-processing of data, data balancing, and prediction of disease.

2.1. Data Collection

In this article, the dataset is collected from the UCI repository, which is considered in research analysis by the many authors [4, 7]. So, the first step is organizing the dataset from the UCI repository to predict the heart disease and then dividing the dataset into two sections: training and testing. In this article, 80% data has been considered as a training dataset, and 20% dataset is used for testing purposes.

2.2. Dataset and Attributes

Attributes of a dataset are properties of a dataset, which are important to analyze and make a prediction regarding our concern. Various attributes of the patient, like gender, chest pain, serum cholesterol, fasting blood pressure, exang, etc., are considered for predicting diseases. However, the correlation matrix can be used for attribute selection to construct a model.

Table 1. *Attributes used are listed.*

Sl. No.	Attributes	Description	Values
1.	Age	Patients age in years	Continuous
2.	Sex	Sex of subject (male-0, female-1)	Male/Female
3.	CP	Chest pain type	Four types
4.	Trestbps	Resting blood pressure	Continuous
5.	Chol	Serum cholesterol in mg/dl	Continuous
6.	FBS	Fasting blood pressure	< or >120 mg/dl
7.	Restecg	Resting Electrocardiograph	Five values
8.	Thalach	Maximum heart rate achieved	Continuous
9.	exang	Exercise Induced Angina	Yes/No
10.	oldpeak	ST Depression introduced by exer.	Continuous
11.	slope	Slope of Peak Exercise ST segment	up/flat/down
12.	Ca	Number of major vessels	0-3
13.	thal	Defect type	Reversible/Fixed/Normal
14.	Targets	Heart disease	1 (disease), 0 (no disease)

2.3. Pre-processing of Data

We need to clean and remove the missing or noise values from the dataset to obtain accurate and perfect results, known as data cleaning. Using some standard techniques in python 3.8, we can fill missing and noise values, see [16]. Then we need to transform our dataset by considering the dataset's normalization, smoothing, generalization, and aggregation.

Integration is one of the crucial phases in data pre-processing, and various issues are considered here to integrate. Sometimes the dataset is more complex or difficult to understand. In this case, the dataset needs to be reduced in a required format, which is best to get a good result.

2.4. Balancing of Data

Balancing the dataset is necessary to improve the performance of machine learning algorithms. A balanced dataset has the same amount of input samples for each output class (or target class). The imbalanced dataset can be balanced by considering two methods, such as under sampling and over sampling.

2.5. Prediction of Disease

In this article, five different machine learning algorithms are implemented for classification. A comparative analysis of the algorithms has been studied. Finally, this article considers an ML algorithm that gives the highest accuracy rate for heart disease prediction, see Figure 1.

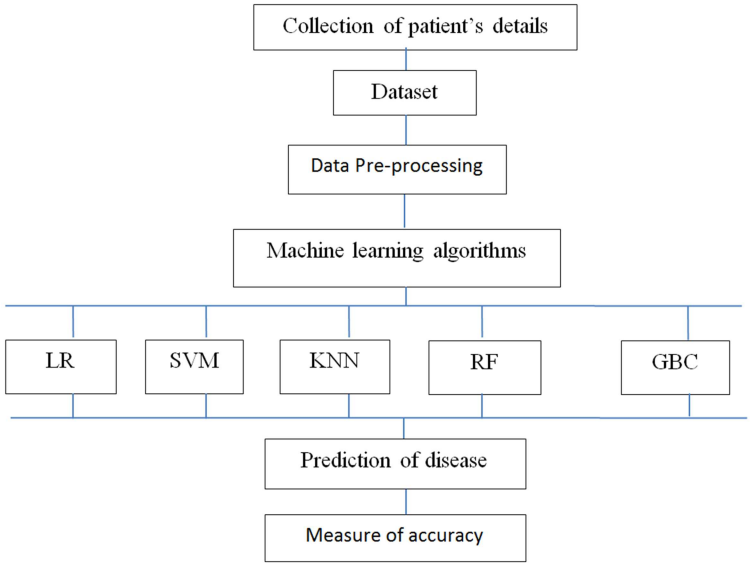

Figure 1. *Architecture of prediction models.*

3. Machine Learning Algorithms

A data analysis technique called machine learning automates the development of analytical models. In this observation, five different algorithms are studied to obtain the accuracy for finding the best one.

3.1. Logistic Regression Model

This ML model is often used for classification and predictive analysis, also known as logit regression [16]. It is also utilized to estimate the discrete values, like the binary outcome, from a collection of independent variables. A binary result means two possibilities will happen: either the event happens (say 1), or it does not happen (say 0).

Here below are the working procedures of the Logistic Regression model.

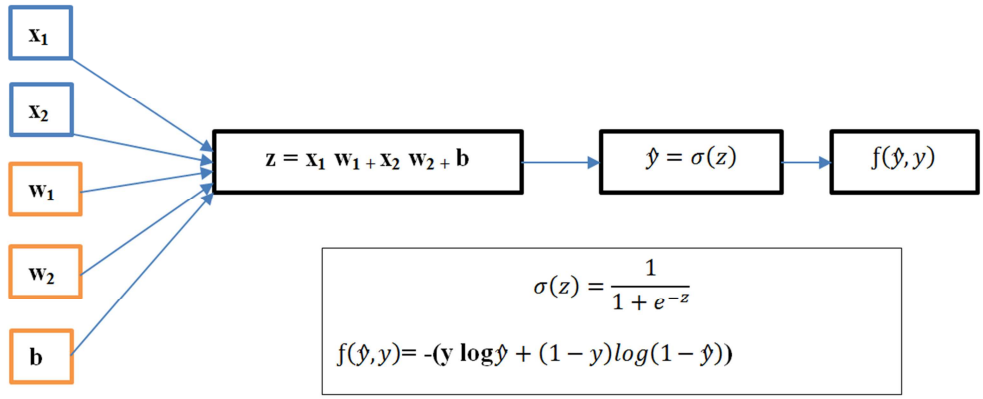

Figure 2. *Logistic Regression model.*

Where z is a function of x1, x2, w1, w2, and b. So, z is a linear equation given to a sigmoid function to predict the output. We calculate the loss to evaluate the performance of this model. In this case, we use the cross-entropy loss

function [17].

3.2. Support Vector Machine (SVM)

SVM is the most popular supervised machine learning algorithm, which is used for classification as well as regression [23]. Although, we primarily consider this algorithm for classification problems in ML. The purpose of the SVM algorithm is to construct the optimum decision boundary or line that can divide n-dimensional space into classes so that we can quickly put the new data point in the correct category. This optimal decision boundary is known as hyperplane [23]. SVM selects the extreme vectors that help to create the hyperplane. The extreme vectors are known as support vectors, and the algorithm with support vectors is called the support vector machine. Here below is a figure of SVM, where the decision boundaries or hyperplane classifies two different categories. The training sample dataset is (x2, x1), where x1 is the x-axis vector, and x2 is the target vector, see figure 3.

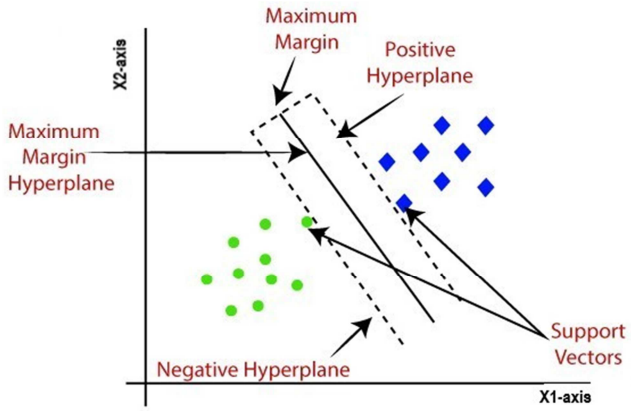

Figure 3. *Support Vector Machine.*

3.3. K-Nearest Neighbors (K-NN)

K-NN is the most straightforward classification algorithm based on supervised learning techniques. However, the K-NN algorithm can also be used for regression but is mostly used for classification [18]. A new data point is classified by using the K-NN algorithm depending on how similar the existing data is stored. It indicates that the K-NN algorithm can quickly classify new data when it appears in a suitable category, see Figure 4.

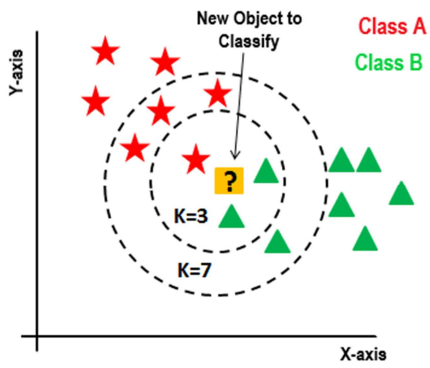

Figure 4. *K-Nearest Neighbors.*

Here, the horizontal x-axis and vertical y-axis are independent and dependent variables of a function, respectively. Figure 4 is a simple example of the K-NN classification algorithm. The test sample (Yellow Square with what symbol) should be classified as either a green triangle or a red star in this algorithm. When k=3 is considered in a small dash circle, the yellow square would be a green triangle because the majority number in this region is green triangles, not red stars. Now, if we consider k=7, which is in a large dash circle, then the yellow square would be red stars because the number of red stars is four and the green triangles are 3. So, It can conclude that the majority vote in a specific region is important here, see Figure 5.

3.4. Random Forest

Random Forest (RF) is a popular supervised machine learning algorithm used for both classifications and regression. However, it is mainly used in classification problems. RF algorithm is based on the concept of ensemble learning. Ensemble learning is a general machine learning procedure that can be used for multiple learning algorithms to seek better predictive performance [2, 19]. So, the RF technique creates several decision trees on the data samples, obtains the prediction from each tress, and finally gets the better solution by considering the majority voting. It is noted that the ensemble method is better than a single decision tree because it mitigates the over-fitting by averaging results. The large number of decision trees in RF helps us to get the accuracy and prevent over-fitting of the problems. The following procedures are completed by RF algorithm, see also figure 6:

Step 1: First, n numbers of the random sample are selected from a given dataset.

Step 2: A decision tree will be constructed for every individual.

Step 3: Each decision tree will predict an output.

Step 4: Final result has come through a majority voting or averaging.

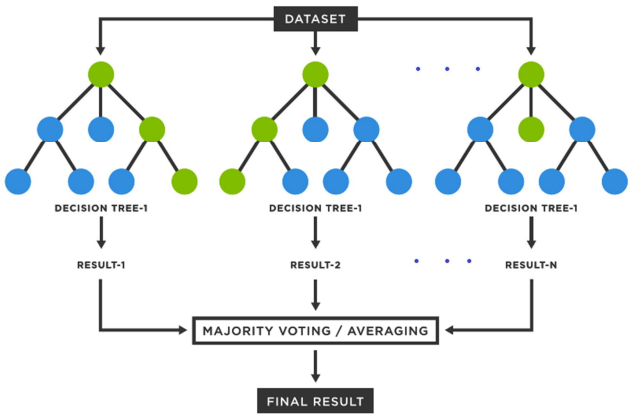

Figure 5. *General procedure of Random Forest.*

3.5. Gradient Boosting

Gradient Boosting (GB) is a machine learning technique

that is used in classification and regression problems like others. It is a powerful algorithm in the field of machine learning [21]. As is well known, the errors are classified into two categories in machine learning algorithms: Bias error and Variance error. GBC helps us to minimize bias error sequentially in the model, see Figure 6. A diagram is described as follows below,

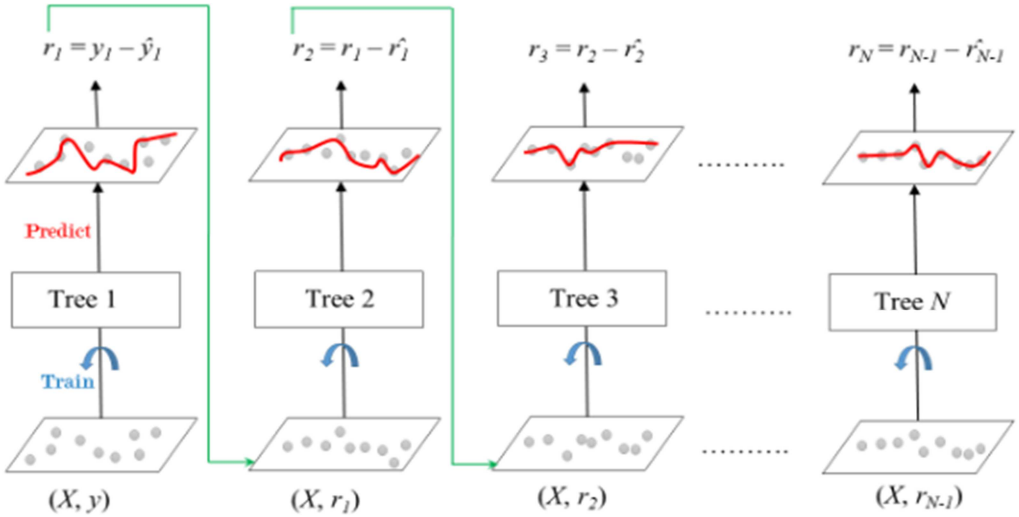

Figure 6. *Diagram Gradient Boosting (Source [22]).*

As we can see that the ensemble consists of N trees; see Figure 6. First of all, the feature matrix X and the labels y are used to train Tree 1. To calculate the training set residual error r1, the predictions labeled are used. Then, Tree 2 is trained using feature matrix X and residual errors r1 of Tree 1 as labels. The residual error r2 is then calculated by using predictive error, see Figure 6.

4. Result Analysis

4.1. Analysis of Heart Disease Dataset

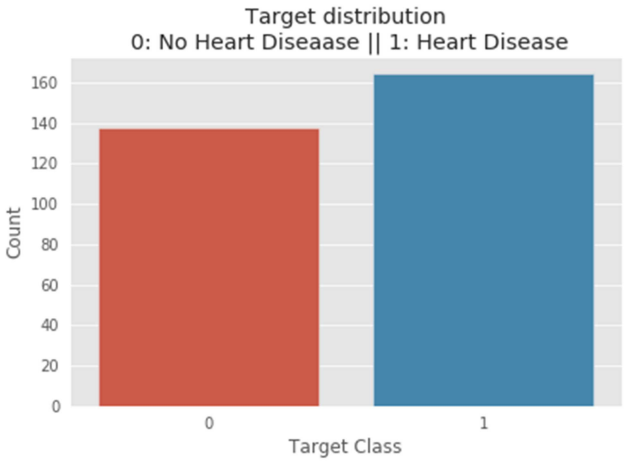

Figure 7. *Target class.*

Before going to study the performance of considering machine learning algorithms in this research, analysis of the features of the heart disease dataset will be focused on here. The total number of observations in the target attributes is 1025, where not having heart disease 499 (denoted by 0) and having heart disease 526 (represented by 1), see Figure 7. So, the percentage of not having heart disease is 45.7%, and the percentage of having heart disease is 54.3%, see Figure 8(a). It is shown that the rate of heart disease is more than the rate of no heart disease. In Figure 8(b), the sex feature of the HD dataset is observed through the target feature. In sex attribute, the female and male numbers are 312 and 713, respectively. So, the male number is more than double of female number. We can see in this figure 8(b) that the number of heart diseases in males is higher than in females. Similarly, no heart disease among males is higher than in females. Figure 8(b) concludes that male is sufferer than female; for more information, see figure 8(b).

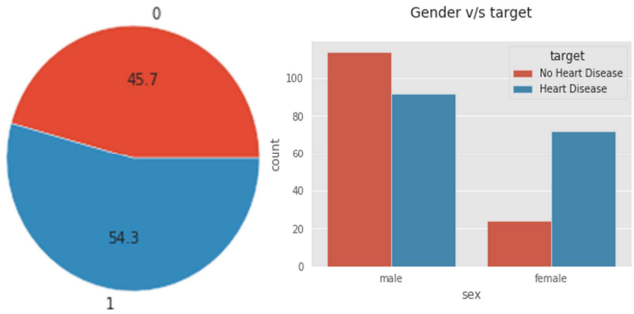

Figure 8. *(a) Percentage of no heart disease and heart disease, (b) Comparison between sex and target feature.*

Figure 9(a) shows a relationship between age and cholesterol with the target feature. For an experiment, these features from the dataset are considered randomly. The trend of no heart disease is higher from 55 to 68 when the cholesterol level is between 200 mg/dl and 300 mg/dl. For validation, the KDE plot 9(b) is studied and shows similar statistics.

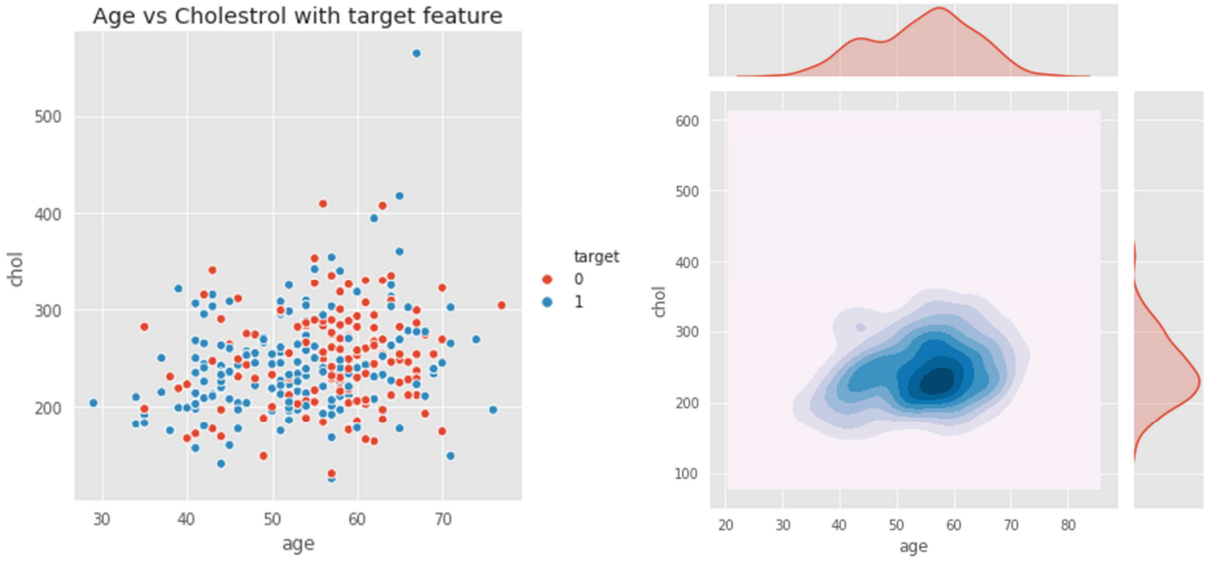

Figure 9. *(a) Age v/s Cholesterol with the target feature, (b) Kernel density estimate (kde) plot of age v/s cholesterol.*

The correlation of the features is drawn in figure 10. The main purpose of the correlation plot is to define the positive and negative correlation between the features. However, it assumes that figure 10 is complex for getting the strong and weak correlation. For this reason, this article added another figure 11 to obtain these correlations efficiently. In figure 11, we can see that three features like cp, thalach, and slope positively correlate with target features.

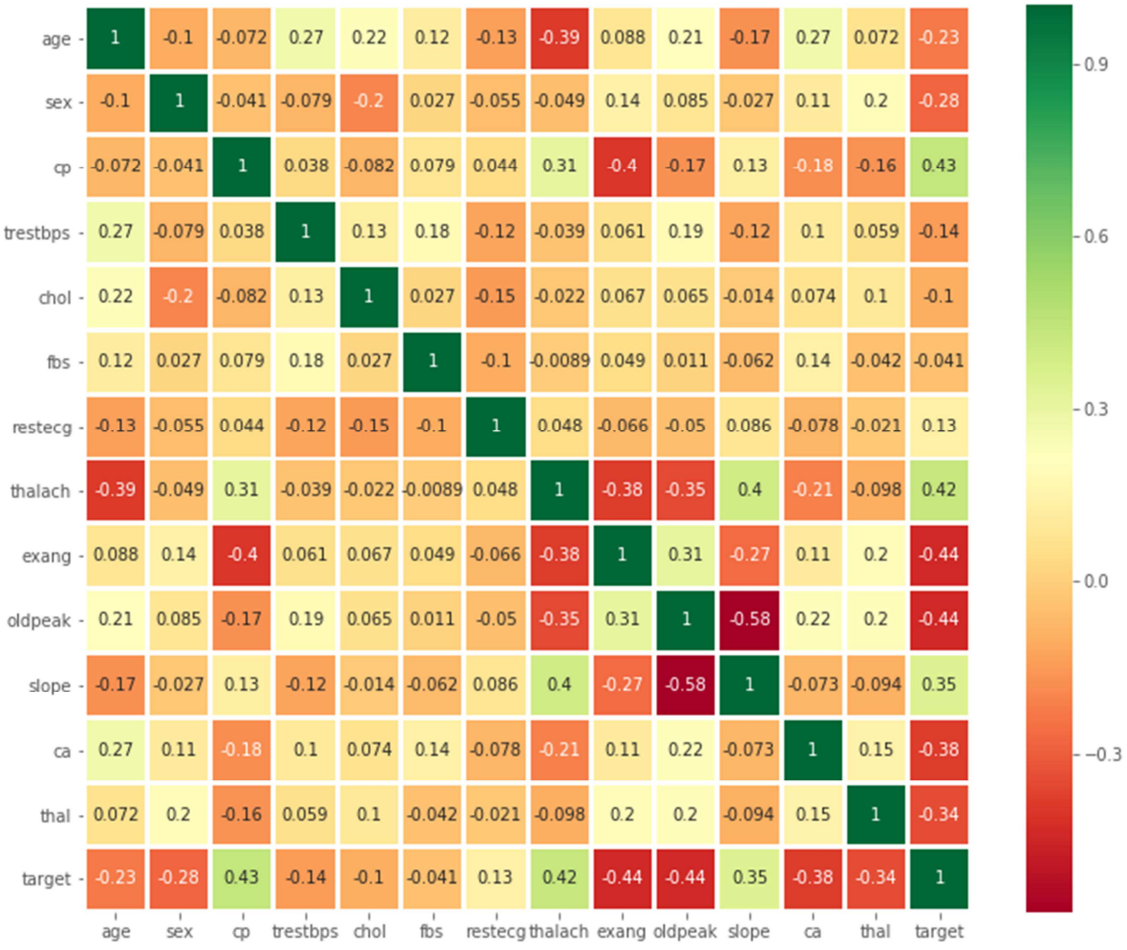

Figure 10. *Correlation matrix of the attributes.*

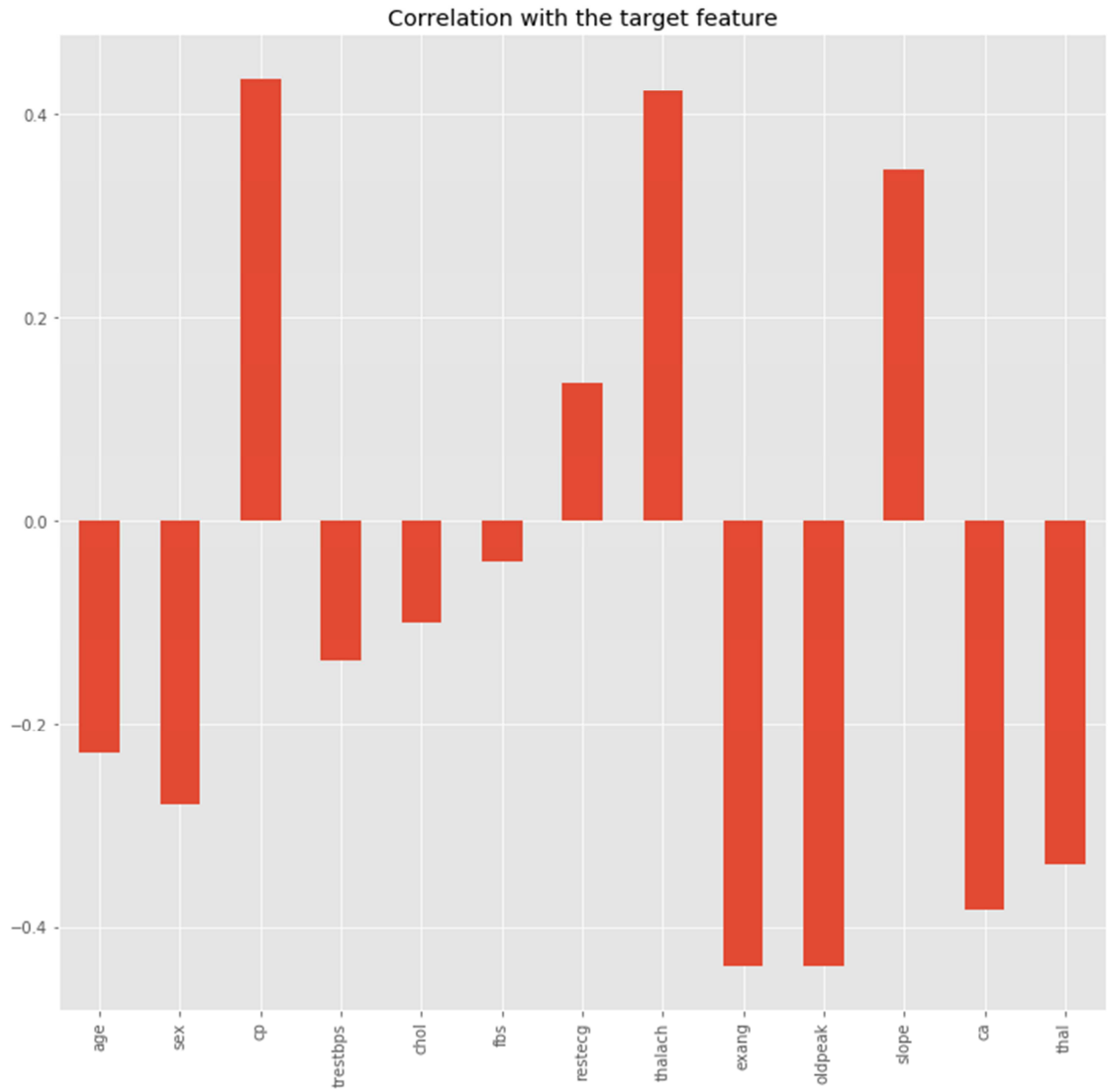

Figure 11. *Correlation with the target feature.*

Two strong correlations by cp and slope with target feature are studied statistically. As we can see in figure 12(a), there is no heart disease when the cp level is more than 350; however, heart disease is sustained more when the cp is between 200 and 250. In addition, when the slope is in 300 < slope-1 < 350, it shows that there is no disease, see figure 12(a). In contrast, for slope-2, there is a heart disease in 300 < slope-2 < 350.

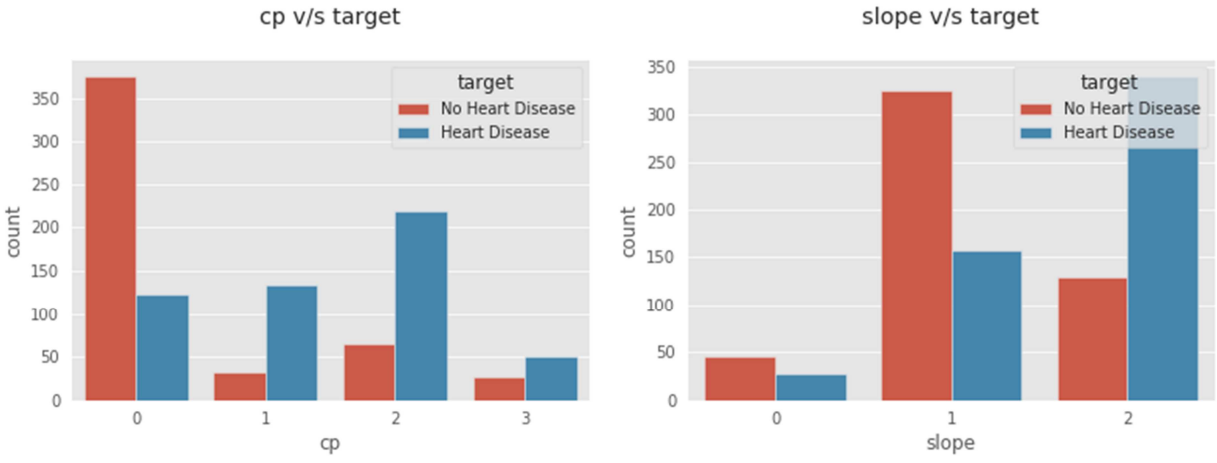

Figure 12. *(a) cp v/s target, (b) slope v/s target.*

4.2. Performance Analysis

In this article, various machine learning algorithms like Logistic Regression (LR), Support vector machine (SVM), k-Nearest-Neighbors (KNN), Random Forest Classifier (RF), and Gradient Boosting Classifier (GBC) are studied broadly to predict the heart disease. The accuracy rate of each algorithm has been measured, and selects the algorithm with the highest accuracy. The accuracy rate is a correct prediction ratio to the total number of given datasets. It can be written as,

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

Where, TP: True Positive
TN: True Negative
FP: False Positive
FN: False Negative

After performing the machine learning algorithms for training and testing the dataset, we can find the better algorithm by considering the accuracy rate. The rate of accuracy is calculated with the support of a confusion matrix. As shown in Table 2, the Logistic Regression algorithm gives us the best accuracy to compare with other ML algorithms.

Table 2. *Accuracy comparison of algorithms.*

Algorithms	Accuracy
Logistic Regression (LR)	0.95
Support vector machine (SVM)	0.90
K-Nearest-Neighbors (KNN)	0.87
Random Forest Classifier (RF)	0.79
Gradient Boosting Classifier (GBC)	0.80

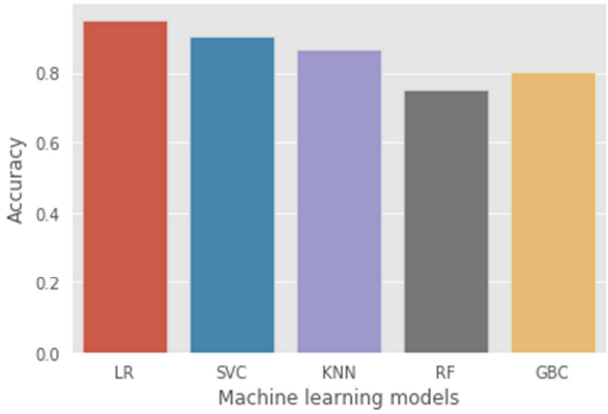

Figure 13. *Accuracy comparison of machine learning algorithms by bar diagram.*

This has been studied more on the LR machine learning algorithm through confusion matrix and f_1-score. The confusion matrix shows that the correct predicted value is 95%, see figure 14. f_1-score is calculated by, which is shown in figure 15,

$$f_1 = 2 * \frac{1}{\frac{1}{P}+\frac{1}{R}} = \frac{2PR}{P+R}$$

Where, Precision, $P = \frac{TP}{TP+FP}$, and Recall $R = \frac{TP}{TP+FN}$.

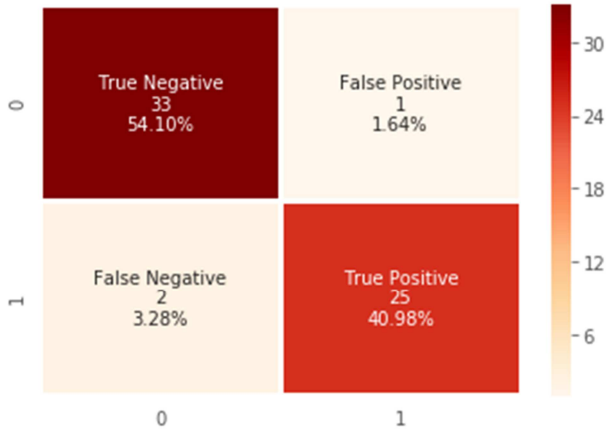

Figure 14. *Confusion matrix of LR algorithm.*

	precision	recall	f1-score	support
0	0.94	0.97	0.96	34
1	0.96	0.93	0.94	27
avg / total	0.95	0.95	0.95	61

Figure 15. *f_1-score, precision, and recall of LR algorithm.*

5. Conclusion and Future Scope

The heart is a vital organ in the human body; however, heart disease is a major concern in the world because this disease is increasing day by day. So, we can handle this disease if we have a model which can predict the initial condition of heart disease. So, we need to create a machine learning model that can be more accurate and help to diagnose heart disease with less doubt and cost. It can be a primary technique for knowing the condition of the heart. For this reason, this article focuses on the heart disease prediction based on the accuracy rate of the confusion matrix. Following this idea, the statistics of the given algorithms are used to estimate the accuracy rate of confusion matrix and validated the statistics among the machine learning algorithms. When five algorithms are compared, it is found that Logistic Regression algorithm is selected regarding the performance of high accuracy rate. The accuracy rate of Logistic Regression model is 95%, which indicates that machine learning algorithm will be considered as a pre-defined tool to seek heart diseases in the near future. Other statistics such as f_1-score, recall, and precision rate have been calculated for Logistic Regression as 95%, 95%, and 95%, respectively. These estimated values suggest the highest accuracy of this algorithm.

These findings suggest that machine learning algorithms can effectively learn about the disease predictions. We may extend this kind of study to diagnose other diseases. We may also analyze the past history of data and combine other

machine learning techniques for better study. Other possible further applications of this study can include such as, cardiovascular disease prediction, diabetic prediction, breast cancer prediction, tumor prediction, and multiple disease predictions.

References

[1] Wikipedia contributors. (2022, June 22). Machine learning. In *Wikipedia, The Free Encyclopedia*. Retrieved 06:31, June 26, 2022, from https://en.wikipedia.org/w/index.php?title=Machine_learning &oldid=1094363111.

[2] Victor Chang, Vallabhanent Rupa Bhavani, Ariel Qianwen Xu, MA Hossain. An artificial intellegence model for heart disease detection using machine learning. Healthcare Analytics, volume 2, November 2022, 100016. https://doi.org/10.1016/j.health.2022.100016.

[3] Ghumbre, S. U., & Ghatol, A. A. (2012). Heart disease diagnosis using machine learning algorithm. In Proceedings of the International Conference on Information Systems Design and Intelligent Applications 2012 (INDIA 2012) held in Visakhapatnam, India, January 2012 (pp. 217-225). Springer, Berlin, Heidelberg.

[4] Rohit Bharti, Aditya Khamparia, Mohammed Shabaz, Gaurav Dhiman, Sagar pande, and Parneet Singh. Prediction of Heart Disease Using a combination of Machine Learning and Deep learning. Hindawi Computational Intelligence and Neuroscience, Volume 2021, Article ID 8387680, 11 pages. https://doi.org/10.1155/2021/8387680.

[5] Khaled Mohamed Almustafa. Prediction of heart disease and classifiers sensitivity analysis. Almustafa BMC Bioinfirmatics (2020) 21: 278. https://doi.org/10.1186/s12859-020-03626-y.

[6] Jee S H, Jang Y, Oh D J, Oh B H, Lee S H, Park S W & Yun Y D (2014), A coronary heart disease prediction model. The Korean Heart Study. BMJ open, 4 (5), e005025.

[7] Mai Shouman, Tim Turner, and Rob Stocker. Applying k-Nearest Neighbour in diagnosis heart disease patients.. International Journal of Information and Education Technology, vol. 2, No. 3, June 2012.

[8] Ganna A, Magnusson P K, Pedersen N L, de Faire U, Reilly M, Arnlov J & Ingelsson E (2013). Multilocus genetic risk scores for coronary heart disease prediction. Arteriosclerosis, thrombosis, and vascular biology, 33 (9), 2267-72.

[9] Raihan M, Mondal S, More A, Sagor M O F, Sikder G, Majumder M A & Ghosh K (2016, December). Smartphone based ischeme heart disease (heart attact) risk prediction using clinical data and data mining approaches, a prototype design. 19th International conference on Computer and Information Technology (ICCIT) (pp. 299-303). IEEE.

[10] Acharya U R, Fujita H, Oh S L, Hagiwara Y, Tan J H & Adam M (2017). Application of deep convolutional neural network for automated detection of myocardial infarction using ECG signals. Information Sciences, 415, 190-8.

[11] Takci H (2018). Improvement of heart attack prediction by the feature selection methods. Turkish Journal of Electrical Engineering & Computer Sciences, 26 (1), 1-10.

[12] Brown N, Young T, Gray D, Skene A M & Hampton J R (1997). Inpatient deaths from acute myocardial infarction, 1982-92: analysis of data in the Nottingham heart attack register, BMJ, 315 (7101), 159-64.

[13] Soni J, Ansari U, Sharman D & Soni S (2011). Predictive data mining for medical diagnosis: an overview of heart disease prediction. International Journal of Computer Applications, 17 (8), 43-8.

[14] Bashir S, Qamar U & Javed M Y (2014, November). An ensemble-based decision support framework for intelligent heart disease diagnosis. International Conference on Information Sociaty (i-Sociaty 2014) (pp. 259-64). IEEE.

[15] Ordonez C (2006). Associate rule discovery with the train and test approach for heart disease prediction. IEEE Transaction on Information Technology in Biomedicine, 10 (2), 334-43.

[16] Wikipedia contributors. (2022, June 21). Logistic regression. In *Wikipedia, The Free Encyclopedia*. Retrieved 06:36, June 26, 2022, from https://en.wikipedia.org/w/index.php?title=Logistic_regression&oldid=1094256072.

[17] Wikipedia contributors. (2022, June 1). Linear regression. In *Wikipedia, The Free Encyclopedia*. Retrieved 06:39, June 26, 2022, from https://en.wikipedia.org/w/index.php?title=Linear_regression &oldid=1091044459.

[18] Wikipedia contributors. (2022, June 4). K-nearest neighbors algorithm. In *Wikipedia, The Free Encyclopedia*. Retrieved 06:40, June 26, 2022, from https://en.wikipedia.org/w/index.php?title=K-nearest_neighbors_algorithm&oldid=1091525121.

[19] Wikipedia contributors. (2022, June 20). Random forest. In *Wikipedia, The Free Encyclopedia*. Retrieved 06:41, June 26, 2022, from https://en.wikipedia.org/w/index.php?title=Random_forest&oldid=1094130824.

[20] Wikipedia contributors. (2022, June 15). Decision tree learning. In *Wikipedia, The Free Encyclopedia*. Retrieved 06:42, June 26, 2022, from https://en.wikipedia.org/w/index.php?title=Decision_tree_learning&oldid=1093316444.

[21] Wikipedia contributors. (2022, June 24). Gradient boosting. In *Wikipedia, The Free Encyclopedia*. Retrieved 06:43, June 26, 2022, from https://en.wikipedia.org/w/index.php?title=Gradient_boosting &oldid=1094845596.

[22] nikki2398. (02 Sep, 2020). ML–Gradient Boosting. https://www.geeksforgeeks.org/ml-gradient-boosting/.

[23] Wikipedia contributors. (2022, June 20). Support-vector machine. In *Wikipedia, The Free Encyclopedia*. Retrieved 06:51, June 26, 2022, from https://en.wikipedia.org/w/index.php?title=Support-vector_machine&oldid=1094109362.