Heart Disease Prediction Using Machine Learning

Kundan Kumar Jha – 22MSM40052

M.Sc. Data Science, Department of mathematics,

University Institute of Sciences, Chandigarh University,

Gharuan, Mohali, Punjab-140413

May 2023

Abstract

Throughout their growth, machine learning and artificial intelligence have been shown to be beneficial in a variety of fields, particularly with the recent explosion in data. Making quicker and more accurate decisions in terms of disease forecasts may be possible using this method. As a result, machine learning algorithms are being used more and more to forecast various diseases. Making a model can also assist in bettering the consistency and accuracy of disease reporting by aiding in the visualisation and analysis of diseases. This article has looked into the use of several machine learning algorithms to identify cardiac disease. The research presented in this article revealed a two-step procedure. The heart disease dataset is first converted into the necessary format for machine learning algorithms to run on it. The UCI repository is used to gather patient data, including medical records. The presence or absence of heart disease in the patients is then determined using the heart disease dataset. Second, this essay presents a lot of useful findings. The confusion matrix is used to validate the accuracy rate of machine learning methods including Logistic Regression, Support Vector Machine, K-Nearest-Neighbours, Random Forest, and Gradient Boosting Classifier. According to recent research, compared to other algorithms, the Logistic Regression algorithm provides a high accuracy rate of 95%. It also exhibits higher f1-score, recall, and precision accuracy than the other four algorithms. The difficult element of this research will be to increase the machine learning algorithms' accuracy rates to about 97%–100% in the future.

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

1. Introduction

The human body consists of up of a variety of organs, each of which serves a specific purpose. One such organ that pumps blood across the body is the heart; if it fails to do so, deadly conditions may arise in the human body. Having a cardiac condition is one of the leading causes of death in today's world. As a result, maintaining the health of our cardiovascular system—or any other system of human body, for that matter—becomes imperative. Unfortunately, cardiovascular problems are a problem for people everywhere.

Any technology which can identify these diseases early on and treat them effectively would help individuals save money and, more crucially, their lives. The prediction of heart disorders can benefit from data mining approaches. By searching databases for patterns and trends that had not previously been seen and using the resulting data, predictive models can be created. To mine data is to draw knowledge out of a lot of information. Machine learning is a technique that can assist in diagnosing cardiac disease before a person suffers significant harm. Machine learning, a young branch of science and technology, can determine whether or not a person is likely to have heart disease.

2. Methodology

This section provides a description of the methodology and analysis used in this research project. The first phases in this investigation were the gathering of data and the choice of pertinent attributes. The pertinent data is then pre-processed into the necessary format. The provided data is then divided into training datasets and testing datasets. The model is then trained using the algorithms and the supplied data. The testing data is used to determine the model's correctness. The processes for this study are loaded utilising a number of modules, including data collecting, attribute selection, pre-processing, data balance, and disease prediction.

2.1. Data collection

The dataset was gathered from the UCI repository and is used by numerous authors in their research analyses. In order to predict heart disease, the dataset from the UCI repository is first organised, and the dataset is then divided into two sections: training and testing. In this article, data from 80% of the samples are utilised for training, and data from 20% of the samples are used for testing.

2.2. Dataset and Attributes

The characteristics of a dataset called attributes are crucial for analysis and prediction in relation to our concern. The patient's gender, chest discomfort, serum cholesterol, fasting blood pressure, and other characteristics are taken into account when predicting illnesses. The correlation matrix can be used to choose which attributes to include in a model, though.

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

To get precise and ideal findings, data cleaning, or the removal of missing or noisy values from the dataset, is required. We can fill in missing and noisy numbers using some standard Python 3.8 techniques; for more information. The dataset must next be transformed by taking its normalisation, smoothing, generalisation, and aggregation into account. Multiple difficulties

are taken into consideration here for integration, one of the key stages of data pre-processing. The dataset can occasionally be more complicated or challenging to comprehend. In this instance, it is best to reduce the dataset in the needed format in order to obtain a satisfactory outcome.class (or the intended class). Two approaches, such as undersampling and oversampling, can be used to balance an unbalanced dataset.

2.4. Balancing of Data

To enhance the effectiveness of machine learning algorithms, the dataset must be balanced. An equal number of input samples and output samples make up a balanced dataset.

2.5. Prediction of Disease

Five distinct machine learning algorithms are used in this paper for categorization. These algorithms have been compared in a study. The paper also discusses a machine learning method that has the highest accuracy rate for predicting cardiac disease.

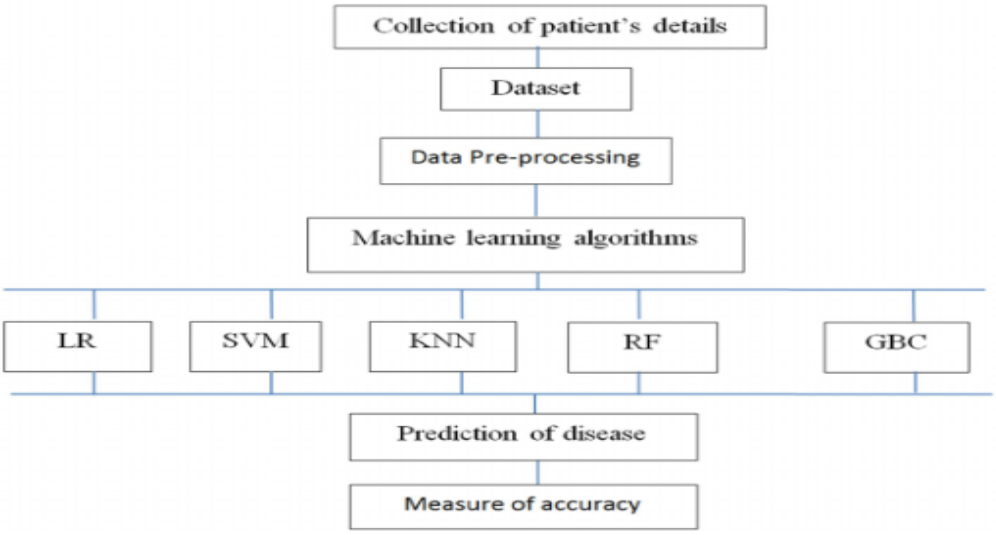

Figure 1: *Architecture of prediction model*

3. Machine Learning Algorithms

Machine learning is a method of data analysis that automates the creation of analytical models. In this observation, five distinct algorithms are examined in order to determine which one is the most accurate.

3.1. Logistic Regression Model

This machine learning (ML) model, also known as logit regression, is frequently used for categorization and predictive analysis. Additionally, it is used to estimate discrete values from a set of independent variables, such as a binary outcome. A binary result means there are just two possible outcomes: either the event occurs (say, 1) or it does not occur (say, 0). The Logistic Regression model's operational processes are listed below. where z depends on the variables x1, x2, w1, w2, and b. In order for a sigmoid function to predict the output, z is a

linear equation. To assess this model's effectiveness, we compute the loss. We apply the cross-entropy loss function in this situation.

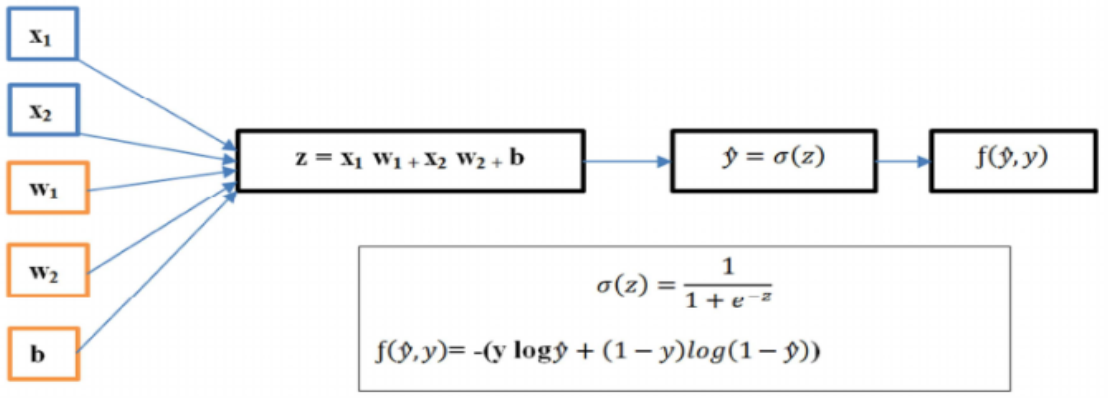

Figure 2: *Logistic Regression Model*

3.2. Support Vector machine (SVM)

The most well-known supervised machine learning algorithm, SVM, is utilised for both regression and classification. However, this approach is generally taken into account for classification issues in ML. The SVM algorithm's goal is to create the best decision boundary or line that can categorise n-dimensional space, allowing us to swiftly assign a new data point to the appropriate category. The hyperplane is the best decision boundary. The extreme vectors that contribute to the hyperplane's creation are chosen via SVM. The extreme vectors are referred to as support vectors, and the support vector machine is the technique that uses them. The decision boundary or hyperplane in the SVM picture below classifies two different categories. (x2, x1), where x1 is the x-axis vector and x2 is the target vector, makes up the training sample dataset.

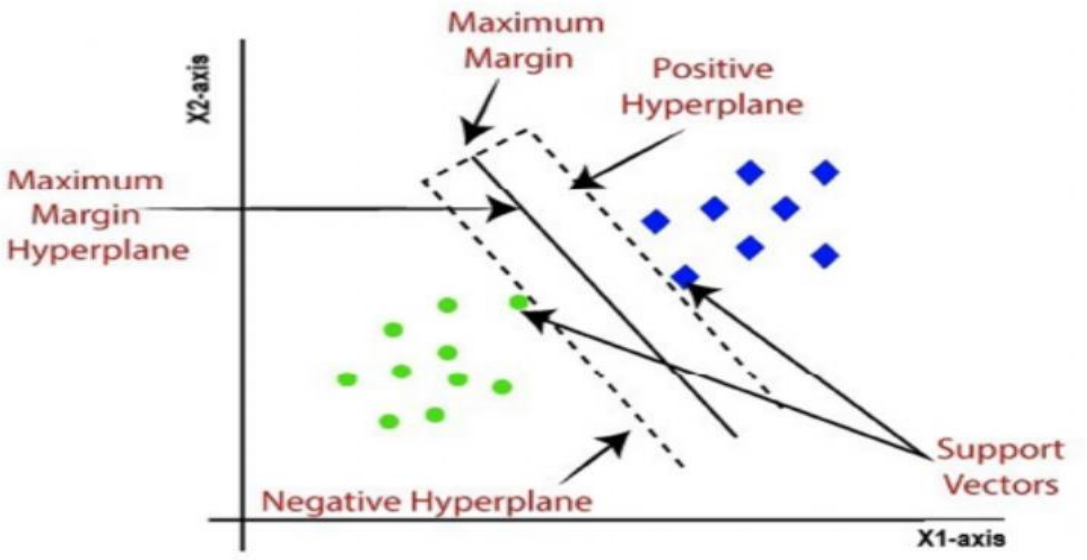

Figure 3: *Support Vector Machine*

3.3. K-Nearest Neighbors (K-NN)

The simplest classification algorithm based on supervised learning methods is K-NN. Although it can be used for regression, the K-NN technique is more frequently used for classification. Based on how similar the preexisting data is stored, a new data point is categorised using the K-NN algorithm. It suggests that when new data appears in an appropriate category, the K-NN algorithm can swiftly categorise it.

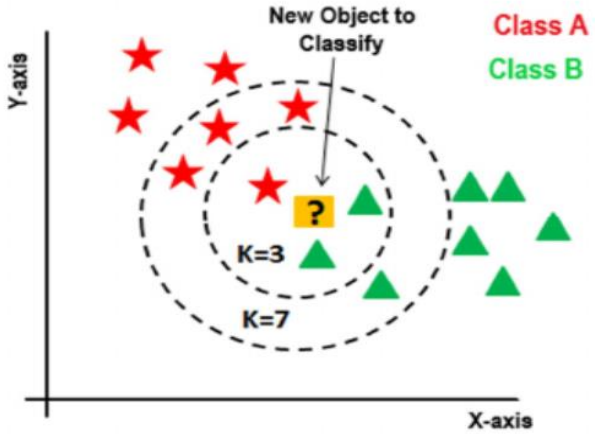

Figure 4: *K-Nearest Neighbors*

The vertical y-axis and horizontal x-axis in this diagram represent independent and dependent variables, respectively, of a function. A straightforward illustration of the K-NN classification method is shown in Figure 4. This algorithm should classify the test sample (Yellow Square with what symbol) as either a green triangle or a red star. The yellow square would become a green triangle when k=3 is taken into account in a little dashed circle since green triangles, not red stars, make up the bulk of numbers in this area. Now, if we take into account k=7, which is in a sizable dashed circle, the yellow square would be red stars since there are four red stars and three green triangles.

3.4. Random Forest

Popular supervised machine learning algorithm Random Forest (RF) is used for classification and regression. However, its primary use is The idea of ensemble learning is the foundation of the RF algorithm. A broad machine learning technique called ensemble learning can be applied to various learning algorithms to improve prediction performance. Therefore, the RF approach builds many decision trees on the data samples, gets the forecast from each tree, and then considers majority vote to arrive at the better option. The ensemble method is said to be superior to a single decision tree because it reduces over-fitting by averaging the outcomes. The abundance of decision trees in RF enables us to obtain accuracy and avoid over-fitting of the issues.

The RF algorithm completes the following processes:

Step 1: a random sample of n numbers is chosen from a given dataset.

Step 2: involves creating a decision tree for each person.

Step 3: An output will be predicted by each decision tree.

Step 4: A majority vote or average was used to determine the outcome.

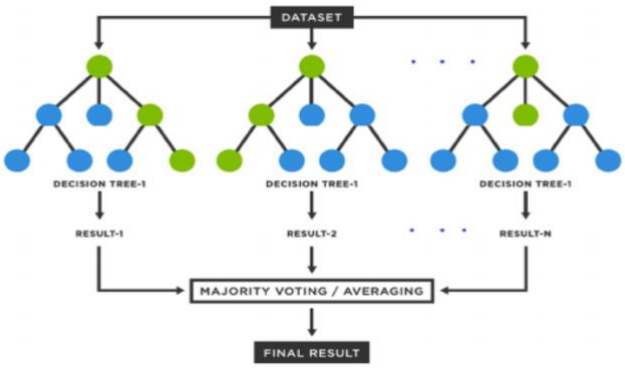

Figure 5: *General Procedure of Random Forest*

3.5. Gradient Boosting

Machine learning techniques like Gradient Boosting (GB) are applied to classification and regression issues among others. In the area of machine learning, it is a potent algorithm. It is well knowledge that bias error and variance error are the two categories into which errors in machine learning algorithms fall. Figure 6, illustrates how GBC aids in sequential bias error minimization in the model. Below is a description of a diagram:

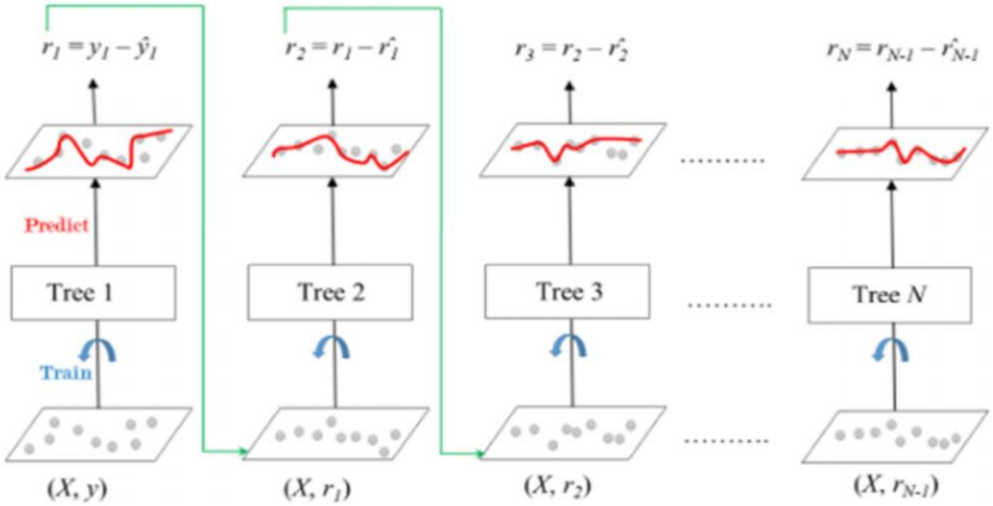

Figure 6: *Diagram Gradient Boosting*

As we can see in Figure 6, the ensemble comprises of N trees. First, Tree 1 is trained using the feature matrix X and the labels Y. The predictions with labels are utilised to determine the

training set residual error r1. Then, Tree 2 is trained with the feature matrix X and the Tree 1's residual errors, r1, as labels. Then, using predictive error, the residual error r2 is determined (see Figure 6).

4. Real Analysis

4.1. Analysis of Heart Disease Dataset

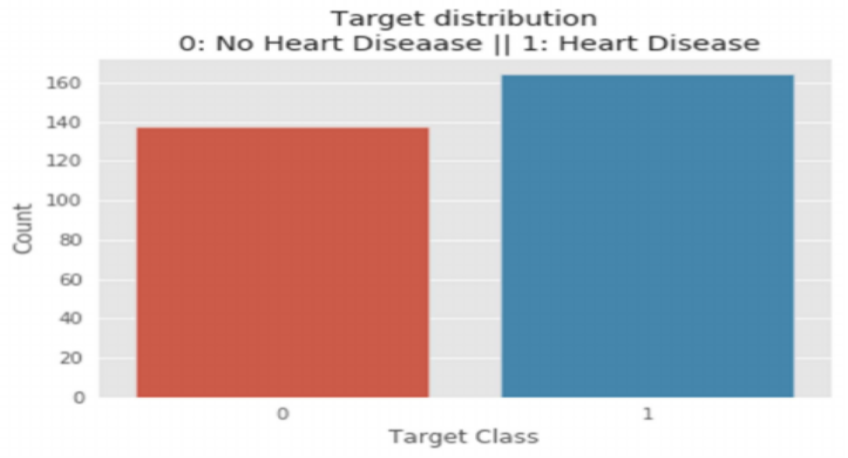

Figure 7: *Target Class*

Analysis of the heart disease dataset's features will be the main emphasis of this research before moving on to examine the performance of machine learning techniques. Figure 7 shows that there are 1025 observations in total for the target qualities, with 499 observations (denoted by 0) showing no heart disease and 526 observations (denoted by 1) showing heart disease. Accordingly, Figure 8(a) shows that the percentage of people without heart disease is 45.7% and the percentage of people with heart disease is 54.3%. It has been demonstrated that the prevalence of heart disease is higher than the prevalence of heart disease-free people. The target feature in Figure 8(b) allows for observation of the HD dataset's sex feature. The female and male numbers in the sex attribute are 312 and 713, respectively. In other words, there are twice as many men as women. Figure 8(b) demonstrates that there are more males than females who suffer from heart disorders. In a similar vein, there is no greater prevalence of cardiac disease in men than in women. Males experience more suffering than females, according to figure 8(b); for further information, see figure 8(b).

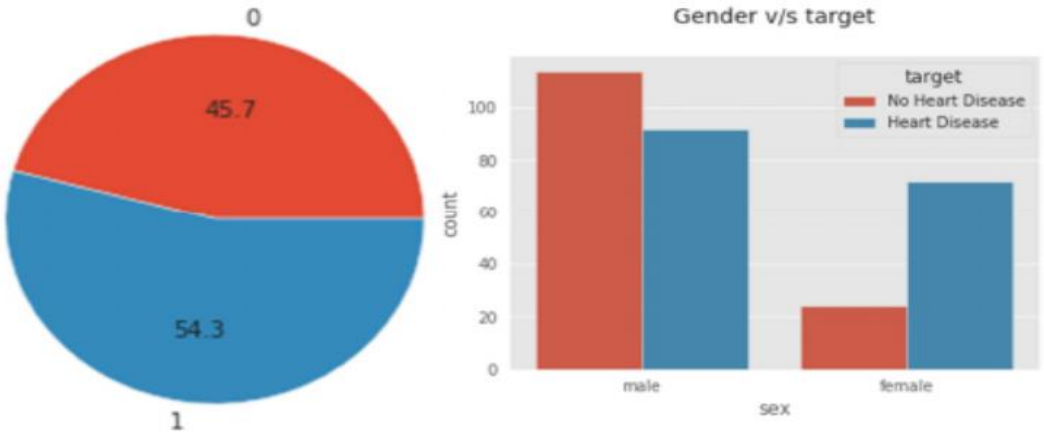

Figure 8: *(a) Percentage of no heart disease and heart disease, (b) Comparison between sex and target feature*

Age and cholesterol are shown to be related to the target feature in Figure 9(a). These dataset features are regarded as random for the experiment. When the cholesterol level is between 200 mg/dl and 300 mg/dl, the trend of having no heart disease increases from 55 to 68. The KDE plot 9(b) is examined for validation and exhibits comparable statistics.

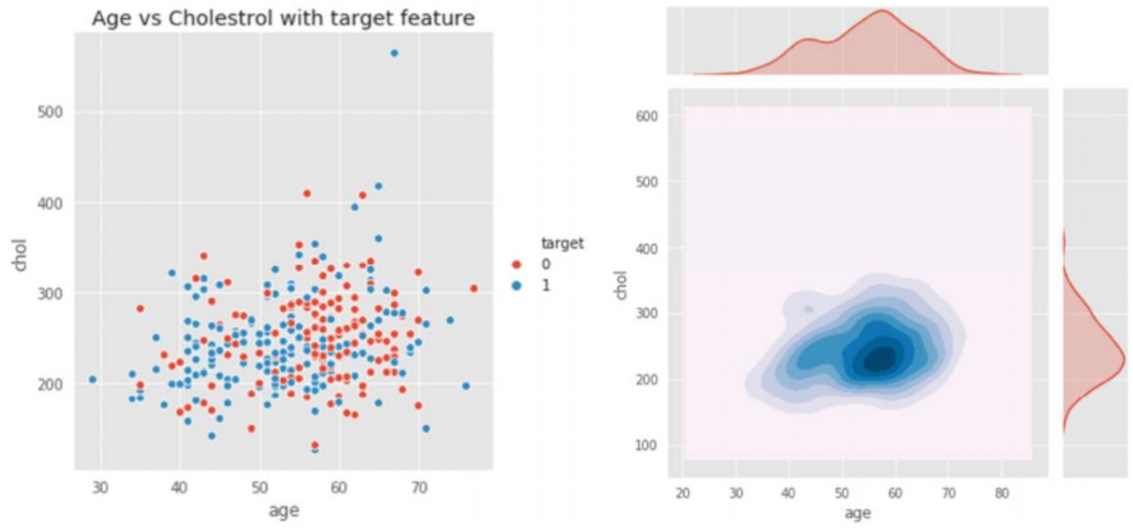

Figure 9: *(a) Age v/s Cholesterol with the target feature, (b) Kernel density estimate (kde) plot of age v/s cholesterol*

Figure 10 depicts the association of the features. The correlation plot's main objective is to identify the characteristics' positive and negative correlations. It does, however, presuppose that figure 10 is complicated in order to obtain the strong and weak association. This article included an additional figure 10 to effectively obtain these relationships. Figure 10 demonstrates the positive correlation between three features—cp, thalach, and slope—and the target features.

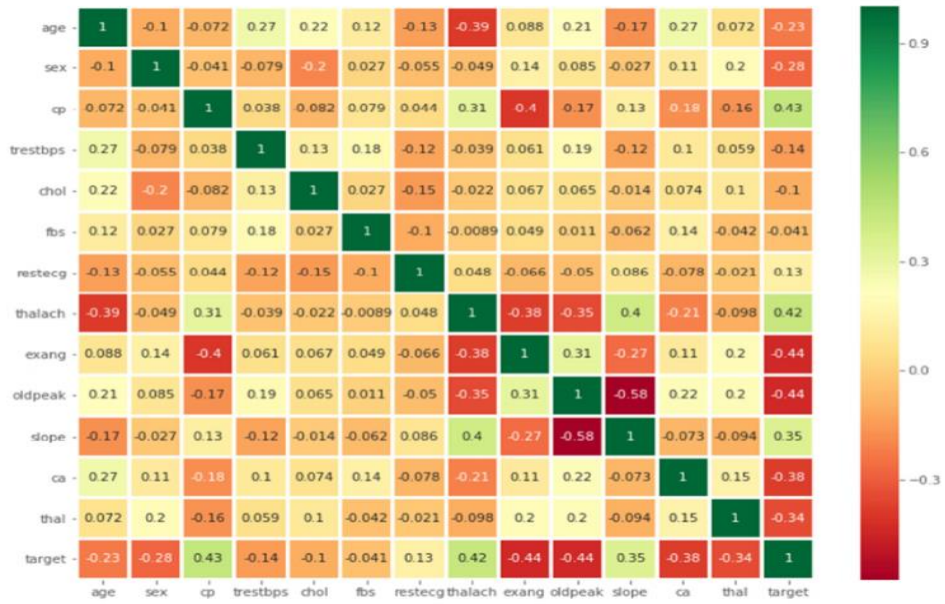

***Figure 10**: Correlation matrix of the attributes.*

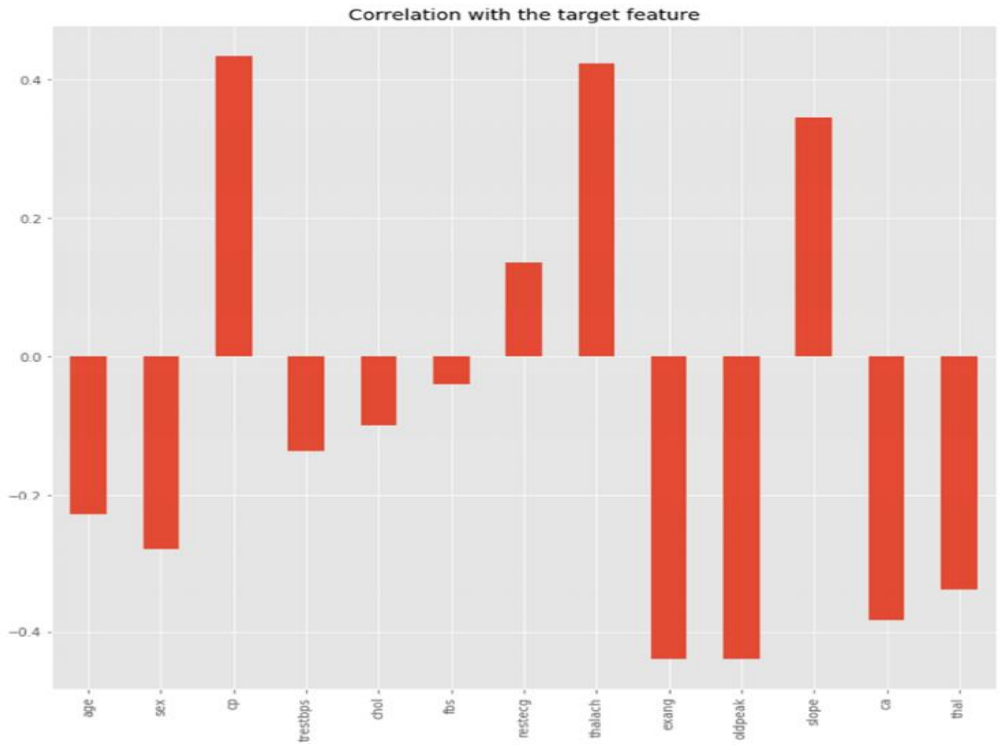

***Figure 11:** Correlation with the target feature*

Two strong relationships between the target feature and cp and slope are statistically examined. Figure 12(a) shows that there is no heart disease when the cp level is greater than 350, but heart disease is more likely to persist when the cp is between 200 and 250. Additionally, as shown in figure 12(a), there is no sickness when the slope is in the range of 300 slope-1 350. For slope-2, however, there is a cardiac condition in 300 slope-2 350.

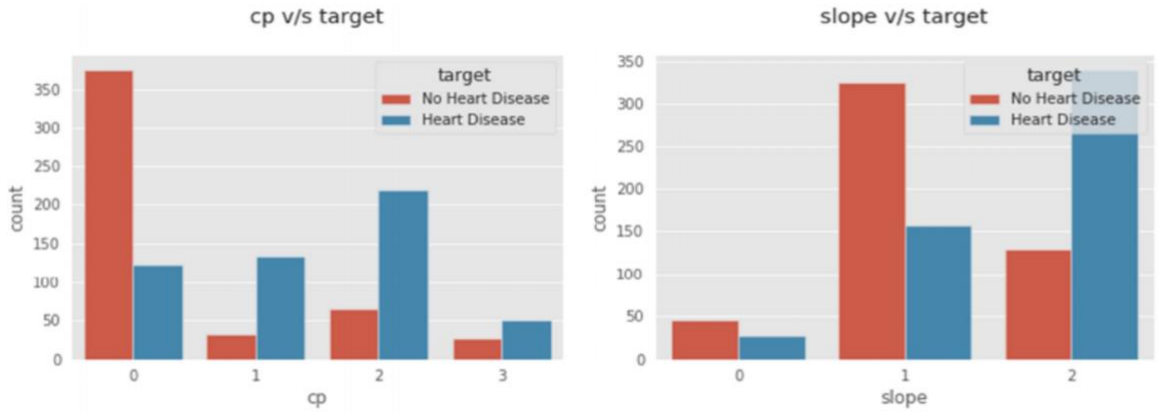

Figure 12: *(a) cp v/s target, (b) slope v/s target.*

4.2. Performance Analysis

In order to predict heart disease, a number of machine learning techniques, including Logistic Regression (LR), Support vector machine (SVM), k-Nearest-Neighbors (KNN), Random Forest Classifier (RF), and Gradient Boosting Classifier (GBC), are investigated in-depth in this work. Each algorithm's accuracy rate has been measured, and the algorithm with the highest accuracy is chosen. The accuracy rate is the proportion of valid predictions to all available datasets. You can spell it out as,

$$\text{Accuracy} = (TP+TN)/(TP+FP+FN+TN)$$

Where, TP: True Positive

TN: True Negative

FP: False Positive

FN: False Negative

We can determine the better algorithm by taking the accuracy rate into account after running the machine learning algorithms on the dataset for training and testing. With the aid of a confusion matrix, the accuracy rate is computed. The Logistic Regression algorithm provides us with the highest accuracy when compared to other ML techniques, as demonstrated in Table 2.

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

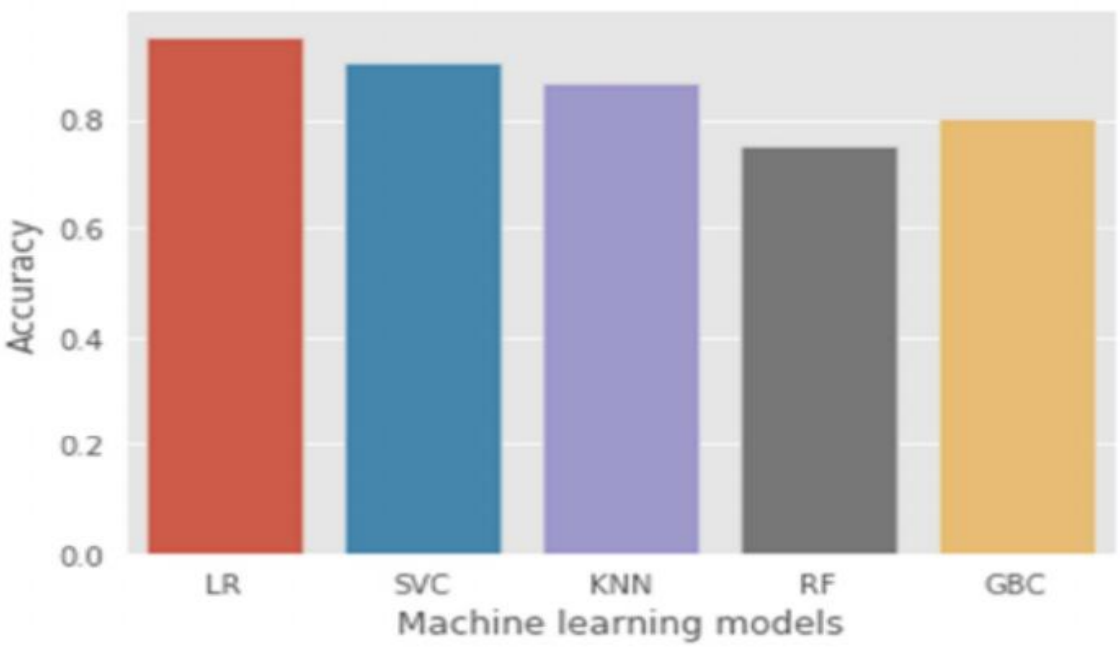

Figure 13: *Accuracy comparison of machine learning algorithms by bar diagram.*

Through the use of the confusion matrix and f1-score, this has been more thoroughly investigated using the LR machine learning algorithm. Figure 14's confusion matrix reveals that 95% of the anticipated value is accurate. Figure 15 illustrates how the f1-score is calculated.

$$f1 = 2PR/P+R$$

where, Precision, $P=TP/TP+FP$, Recall $R = TP/TP+FN$

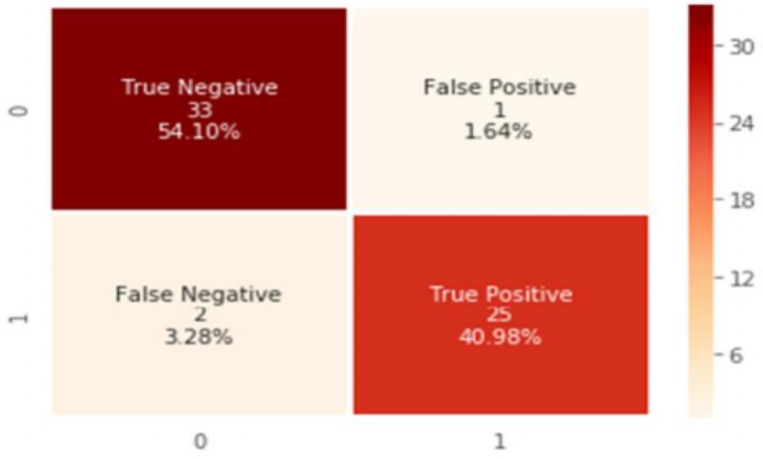

Figure 14: *Confusion matrix of LR algorithm.*

	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: *f1-score, precision, and recall of LR algorithm.*

5. Conclusion

The human heart is an essential organ, but because the prevalence of heart disease is rising globally, there is a serious concern about this condition. Therefore, if we have a model that can anticipate the early stages of cardiac disease, we can manage this disease. Therefore, we need to develop a machine learning model that can be more precise and aid in heart disease diagnosis with less uncertainty and expense. It may serve as a fundamental method for determining heart health. This article concentrates on the heart disease prognosis based on the confusion matrix's accuracy rate for this reason. As a result, the statistics of the specified algorithms are used to validate the statistics among machine learning algorithms and estimate the accuracy rate of the confusion matrix. When five methods are compared, it is discovered that the Logistic Regression algorithm is chosen due to its high accuracy rate performance. The accuracy rate of the Logistic Regression model is 95%, indicating that in the near future, machine learning algorithms will be viewed as a pre-defined instrument to find cardiac disorders. For the Logistic Regression, other statistics like the f1-score, recall, and precision rate have been estimated as 95%, 95%, and 95%, respectively. These estimated numbers point to the algorithm's highest level of accuracy. These results imply that machine learning systems

can efficiently pick up on disease forecasts. This type of research could be expanded to diagnose further illnesses. For a more thorough investigation, we might further examine the data's prior history and integrate different machine learning approaches. This study may also have other uses in the future, including the ability to forecast numerous diseases as well as cardiovascular disease, diabetes, breast cancer, tumours, and other diseases.

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=10943611.

[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