JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

HEART DISEASE PREDICATION USING MACHINE LEARNING CLASSIFICATION MODEL

[1] Dr.Reena shinde 1st, Dr. S. N. shinde 2nd
[1]Assistant Professor, [2]Principal,
[1] Computer Science Department ,
[1] Sinhgad College of Science, Pune, India

ABSTRACT
In today’s era, heart disease (cardiovascular disease) is growing in daily life and due to the effect of heart disease, it is prominent to death. In the health care field, there is a more of data that has been together and from that data we can use definite techniques for data mining and this scheme categorizes all the likelihoods of cardiovascular disease. We are considered in relations of medical parameters and those constraints utilized for data mining and processing these datasets in Python programming and the conclusion of that scheme gives the probabilities of heart disease occuring in terms of the measurement using constructed datasets and in order to determine the accuracy level of heart disease, we can use ten of the most advanced machine learning algorithms, including the Naive Bayes, LR Algorithm, KNN, SVM, DT Algorithm, RF Algorithm, Gradient Boosting Algorithm, Ada Boost Algorithm, ANN, and Passive-Aggressive. The major goal of this study is to use machine learning algorithms to forecast a patient's cardiac condition.

Keywords: Python, ML, LR, KNN, SVM, DT and ANN.

I. INTRODUCTION

The human body's major organ is the heart. It brings blood to every part of our body. The brain and several additional organs will stop operating within a few minutes if it fails to function properly, and the person will die [1]. The medical professionals will not be able to foresee it because it is a difficult undertaking that calls for knowledge and in-depth information. The efficiency of medical care would increase with an automated method for analysis [2]. EDA is a technique for performing advanced data analysis that improves comprehension of a particular dataset, identifies differences, and creates sparse models to test underlying hypotheses [3]. Data mining is mostly used to draw out hidden material from massive databases. KDD is another name for data mining [4].

In order to predict and determine whether a dataset is valid, machine learning uses supervised, unsupervised, and cooperative learning classifiers for a prediction, ML algorithms can build a model built on train data, which are small samples of data [5]. However, because these modifications usually employed tiny number of samples, the results may not be transferable to larger populations when using machine learning algorithms to forecast heart disease. Our research aims to overcome this limitation by utilizing a wider and more varied dataset, which will increase the applicability of the findings [6]. By taking advantage of the intricate relationships between risk factors, machine learning techniques may improve accuracy [7]. Instead, ML techniques enable computers to acquire knowledge on data inputs and make use of statistical analysis to produce values that fall inside a particular range of values. As a result, machine learning makes it simpler for computers to create models from sampling data and automate decision-making processes based on data intakes. ML has benefited all forms of technology [8]. A ML has a vi Heart attacks, chest pain, and strokes can all be caused by cardiovascular disease, which is typically described as having pointed or obstructed blood arteries. The three most prevalent causes of a heart attack are high blood pressure,rapid heartbeat and high cholesterol. The most common heart ailment is a heart attack. [9]. Various data mining, data analysis, and neural network approaches were used to gauge the severity of cardiac disease in individuals. [10].

The goal of this study is to determine whether or not the patient has heart disease in this article. The datasets' records are split up into test set and training set halves. Data mining classification techniques including Naive Bayes, LR, KNN, SVM, Decision Tree, RF, Gradient Boosting, Ada Boost, ANN and Passive-Aggressive are used once the data has been preprocessed. Both the categorization and prediction methods are effectively used in this. Using the programming language, these models are finished.

II. LITERATURE SURVEY

The work on Heart Disease Diagnosis Using ML was done in a study under the direction of Baban.U. Rindhe et al [1]. The complex dataset, which included 303 entries with 14 attributes and was split into training and testing, was employed. The project includes an in-depth study of the patient dataset for heart disease with appropriate data processing. The Support Vector Classifier, Neural Network, and RF Classifier were the three models that underwent training and validation with the highest degree of accuracy.

The research on applying machine learning to predict cardiac disease was done by Rishabh Magar et colleagues [2]. The most successful algorithm was to be chosen based on a number of factors. With an accuracy of 82.89%, they claim that the Logistic Regression algorithm is the most effective of the four. Consequently, a well-designed user interface was used to further affect these four algorithms..

The research on Heart Disease Diagnosis Using Exploratory Analysis of Data was done by R. Indrakumari et al [3]. The dataset, which was split into training and testing, consisted of 303 entries with four attributes. They also started The leading causes of premature mortality and disability are heart attacks and vascular diseases. The key to identifying heart disease is chest pain..

The research on Detection of Heart Disease With ML Algorithms was done by Mr. Santhana Krishnan et al [4]. The dataset was separated into training and testing and had 300 items with 14 attributes. The Decision Tree Classifier diagnosed heart disease patients with an accuracy level of 91%, and the Naive Bayes classifier with an accuracy level of 87%. These two algorithms are used.

The work on Heart Disease Diagnosis Using ML Methods was done in a study led by Mohammed Khalid Hossen et al. [5]. When five methods are used and compared, it becomes clear that the Logistic Regression algorithm has a high accuracy rate. The accuracy rate of the Logistic Regression model is 95%, indicating that in the near future, machine learning algorithms will be used carefully as a pre-defined instrument to hunt for cardiac illnesses.

The research on Heart Disease Estimate Using ML methods was done by Chintan M. Bhatt et al [6]. The dataset was split into training and testing groups and included 70000 entries with 12 attributes. Eighty percent of the dataset was used to train the model, while twenty percent was used to test it. They were effective on the cardiovascular disease dataset and used the MLP, RF, decision tree, and XGBoost algorithms. As a result, the multilayer perceptron (MLP) algorithm obtained the highest cross-validation accuracy of 87.28%, according to the data.

The research on Heart Disease Prevention Using LR Algorithm was done by Bhagyesh Randhawan et al [7]. 304 data will be sampled during this phase of data creation. The information will then be separated into train and test data. they employ the logistic regression technique to generate a variety of data that may be utilized to produce final predictions and The prediction results' accuracy is 85% according to the findings of the data validation.

III. METHEDOLOGY

3.1 PYTHON

The best coding language for creating machine learning models is Python. It promotes versatility, simplicity, and a large library of frameworks. It is a fantastic programming language that is simple to learn, read, and write. This language is free source. There are numerous libraries for Python. It can create complex data structures..

3.2 DATA COLLECTION

The dataset used in this paper was pulled together from the UCI repository and measured for research analysis by several authors [1, 5, 6, 7]. In order to predict heart disease and dataset must first be created from the UCI repository and it must then be divided into two parts: training and testing. In this article, 80% of the data were utilized for training purposes and 20% were used for testing.

3.3 DATASET

We prompted data.dataset in csv. Patient information is used. On Kaggle, the dataset is accessible. 14 characteristics features, 1 output feature, and 270 samples are included in the data set. When making a forecast about our concern, attributes of a dataset are aspects of a dataset that are significant to observe. In order to anticipate diseases, a patient's many characteristics, such as gender, chest discomfort, serum cholesterol, fasting blood pressure, exang, etc., are measured. The correlation matrix, however, can be utilized to choose the attributes for a model.

NO	Attributes	Description

1	**AGE**	Patients age in Year
2	**SEX**	Gender (male=1, women=0)
3	**CPAGE**	Chest Pain Type(0, 1, 2, 3)
4	**TRESTBPS**	BP in mm HG (50-150)
5	**CHOL**	Cholesterol in mg/dl (100-600)
6	**FBS**	Fasting blood sugar>120 (y=1, n=0)
7	**RESTING**	Resting electrocardiographic (y=1, n=0)
8	**THALACH**	Maximum heart rate(71 to 200)
9	**EXANG**	Exercise include agina (y=1, n=0)
10	**OLDPEAK**	St depression (y=1, n=0)
11	**SLOPE**	Slop peak exercise(y=1, n=0)
12	**CA**	No.of major vessels (0-3)
13	**THAL**	Thalassemia (3, 6, 7)
14	**TARGET**	Heart disease (no=1, yes=2)

Table.1 Dataset

IV. MACHINE LEARNING ALGORITHM

4.1 Naïve Bayes

It is constructed using supervised learning theory. It is employed to address classification issues. It is frequently applied to text classification tasks that require large training datasets. A straightforward and effective technique for predictive modeling is naive Bayes. The most efficient classification technique that can handle large, complex, non-linear, dependant data is this model. The nave classifier assumes that the presence of one characteristic in a class does not depend on the presence of any other feature. Nave classifiers have two parts: nave and Bayes. [4]

4.2 Logistic Regression

It is employed in classification where the objective is to forecast the likelihood that a compositional instance will belong to a particular class or not. The data that was partitioned in the previous procedure will be used in this one. [7] When a categorical variable needs to be predicted, we use logistic regression. By 1 or 0, it forecasts the value.

4.3 Support Vector Machine

It is available in both linear and non-linear variants. SVM uses a training set and a test set, which are two distinct datasets. [1] Classifiers provide more accurate results. Due to the decision function's utilization of a subset of training points, SVM classifies with relatively little memory usage..

4.4 K-Nearest Neighbor(KNN)

When new data and existing data are similar, the KNN algorithm accepts that similarity and places the new data in the category that is most similar to the existing categories. The KNN algorithm classifies a new data point based on how similar the stored previous data is to it. [5] KNN is a method of lazy learning.

4.5 Decision Tree

It is based on clear-cut problems with clear-cut answers, such as Yes/No, True or false, 1 or 0, and it employs the decision tree classification technique to handle datasets. This model's results are different from those of the KNN as well as SVM models. Based on the condition associated with the dependent variables, the output consists of vertical as well as horizontal line splitting [4].

4.6 Random Forest

Both regression as well as classification can be done with it. To address a challenging issue and enhance the functionality of the model, it is classified. As implied by the name, this technique takes into account numerous decision trees before producing an output. [1] The results of the random forest algorithm are extremely accurate.

4.7 Gradient Boosting

It is a technique that stands out for its precision as well as speed of prediction, especially for large and complicated datasets. Machine learning algorithms divide errors into two categories: bias error and variance error [2]. Bias is a method's propensity to repeatedly learn the incorrect thing by ignoring all the information in the input.

4.8 Passive Aggressive

If the prediction is accurate, keep the model and don't make any adjustments, and the example's data isn't enough to justify making a number of changes to the model's passive components. Make model modifications if the prediction turns out to be inaccurate.

4.9 Ada Boosting:

Machine learning is being used to solve classification and regression issues. With all classifiers assigning more weight to the data points that are misclassified, the basic idea was to train the weak classifier using the training dataset. The model with the lowest accuracy is given a lower weight, while the weak model with the highest accuracy is given the highest weight..

4.10 Artificial neural network (ANN)

They are suitable for taking on complex and large-scale machine learning models since they are at the basis of deep learning and are versatile, powerful, and easily accessible. An adaptable system called a neural network learns by using reliable nodes or neurons in a covered design that resembles the human brain.

V ATTRIBUTE CORRELATION BY USING HEAT MAP

The kind of pain in the chest (cp), exertion-induced angina (exang), depression of the ST brought on by exercise in comparison to rest (old peak), and the gradient of the highest point during activity are all shown on the correlation plot. The cardiac condition (target) is directly connected with the ST segment (slope), total number of primary vessels (0-3) colored by fluoroscopy (ca), and halassemia(thal). Additionally, we observe that the relationship between heart illness and maximum heart rate is inverse .

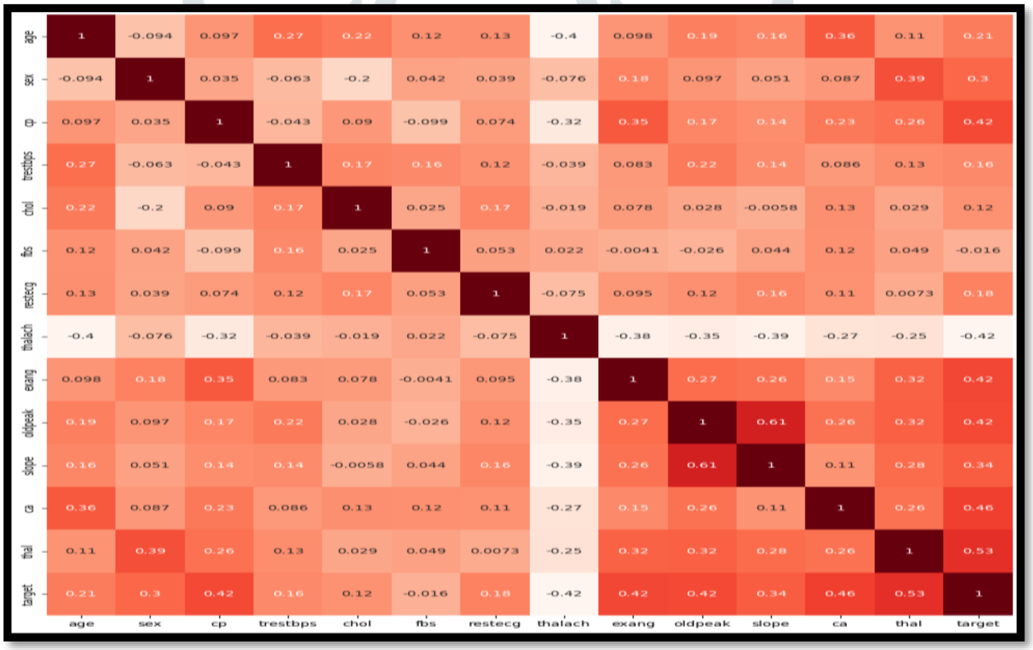


fig.1 Attribute Correlation

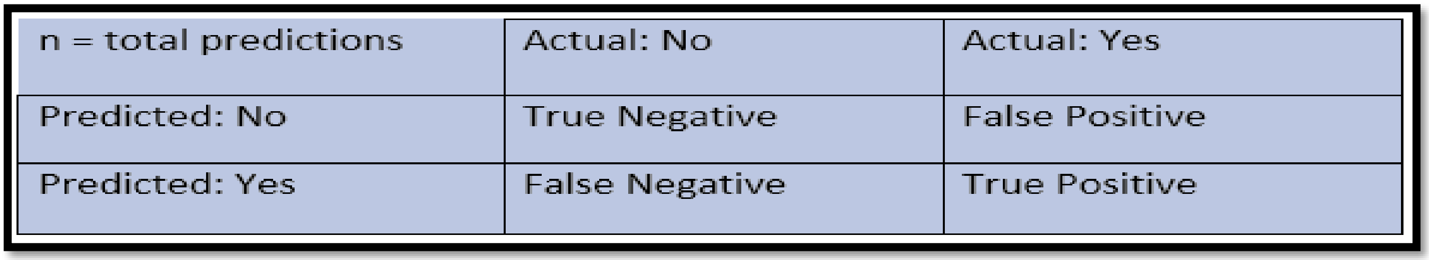
The correlations between the following characteristics can also be seen.:

- The age and the number of main vessels (0–3) that are fluoroscopically colored
- The old peak (old peak) of the exercise-induced ST depression and the slope (slope) of the exercise-induced ST segment peak.
- Exercise-induced angina, a kind of chest pain (cp).

- Age and maximum heart rate (thalch).

5.1 Confusion Matrix in Machine Learning

It is a matrix used to describe how well categorization models perform given a known set of test data. It can only be established if the real test data values are known. The predicted as well as actual values and the total number of forecasts are separated into two dimensions in the matrix. Actual values are the real values for the provided data, whereas projected values are the values that the model predicts.



n = total predictions	Actual: No	Actual: Yes
Predicted: No	True Negative	False Positive
Predicted: Yes	False Negative	True Positive

Table.2 Confusion Matrix

The cases listed in the table above include

True Negative: Prediction If not, actual value was also false.

True Positive: If the value was predicted, it was also true.

False Negative: Actual value was yes when it was predicted to be no. Type II mistake.

False Positive : Predicted If yes, the true value was false. Type-I blunder.

5.2 Calculations using Confusion Matrix

- **Accuracy:** The correctness of categorized issues must be determined by taking into account crucial factors. It forecasts a precise outcome**.**

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

- **Error Rate:** It clearly makes incorrect predictions. The ratio of inaccurate predictions to all of the predictions made by the classifier can be used to calculate error rate.

$$\boxed{\text{ERROR RATE}=FP+FN/TP+FP+FN+TN}$$

- **Precision:** Precision refers to the percentage of results that are proper, while accuracy refers to how many correct outputs the model produced..

$$\boxed{\text{Precision}=TP/TP+FP}$$

- **Recall:** It is clearly defined as the percentage of all positive classes that match the classifications that the model we developed accurately predicted. There must be the greatest possible recall.

$$\boxed{\text{Recall}=TP/TP+FN}$$

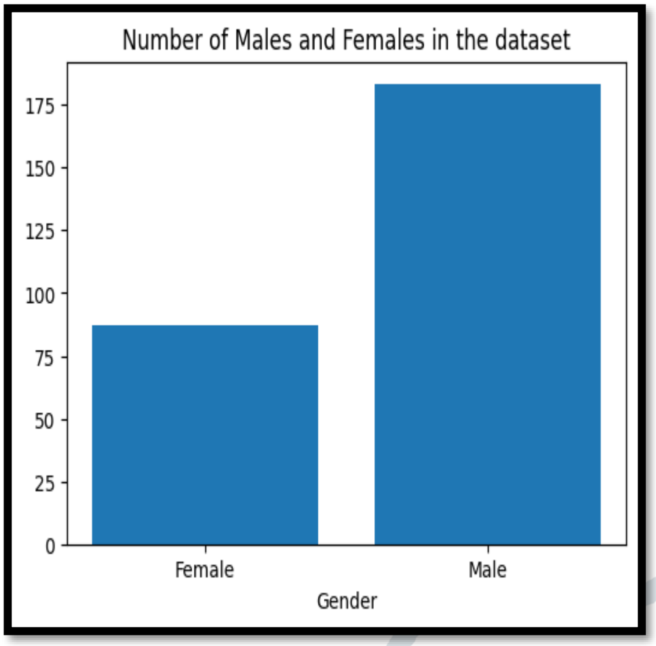
- **F1-Score:** We can simultaneously calculate the recall as well as accuracy .The F1-score is at its highest if recall equals precision.

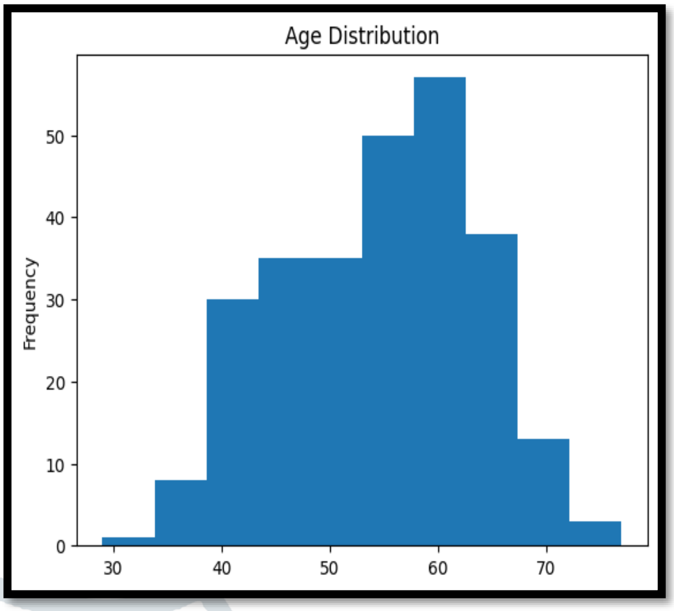
$$\boxed{\text{F1 Score}=2*\text{Recall}*\text{Precision}/\text{Recall}=\text{Precision}}$$

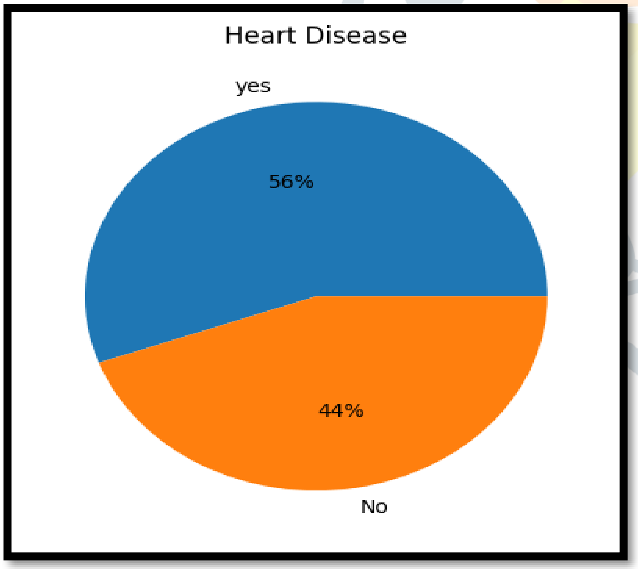
VI RESULT AND DISCUSSION

This study aims to determine whether the patient has cardiac disease or not. Both a training set and a test set of records are included in the datasets. The classification system, which included Nave Bayes, LR,KNN, SVM, Decision Tree, RF, Gradient Boosting, Ada

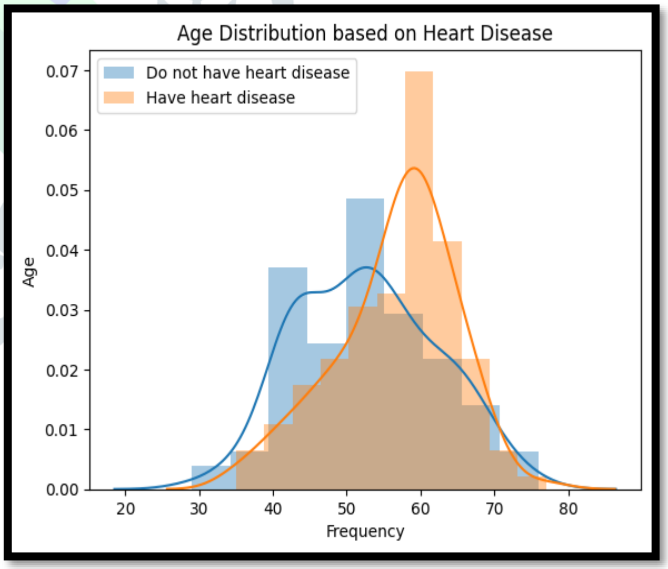
Boost, ANN, and Passive-Aggressive, was used after the data had been preprocessed. The results of the categorization models created using Python programming are shown in this part. Results are produced using both training as well as testing dataset.







TARGET VALUE (NO=1, YES=2)
fig.2 Heart Disease



AGE DISTRIBUTION BY USING TARGET
fig.3 Age Frequency

AGE DISTRIBUTION: (0-100)
fig.4 Age Distribution[0-100]

SEX: GENDER (MALE=1, WOMEN=0)
fig.5 Gender Count[M=1,W=0]

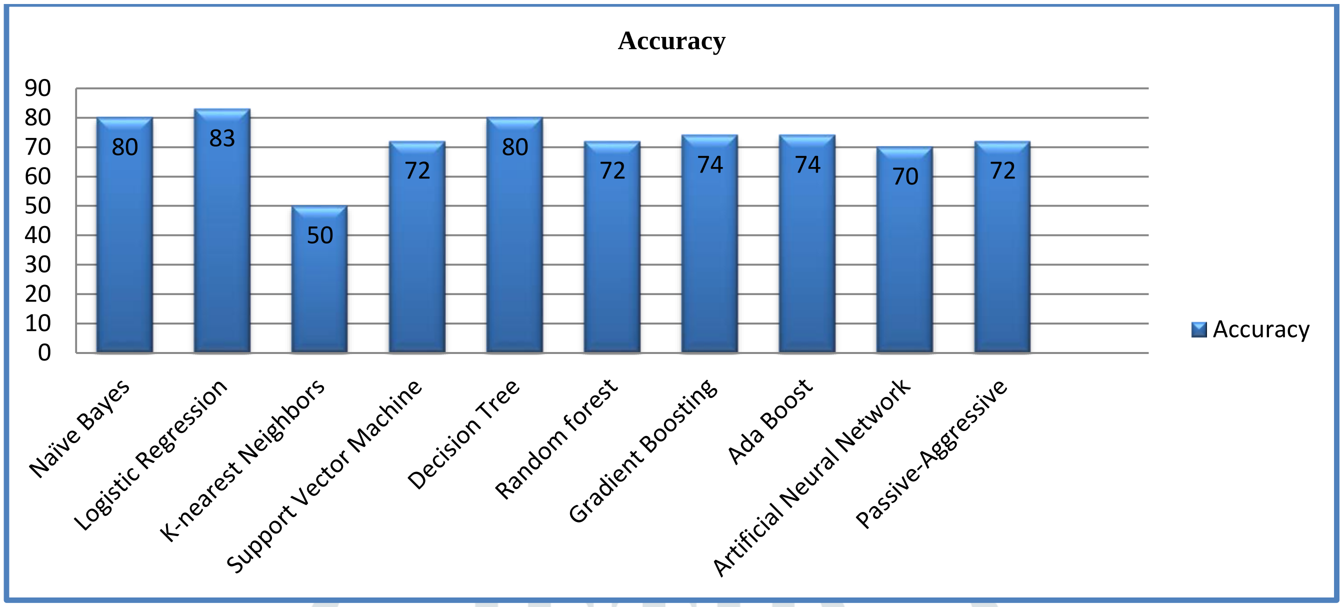


Fig 6. Accuracy measure of algoritham in percentage

The result in Fig. 6 shows that the algorithm with the highest accuracy is the Logistic Regression model. Heart disease patients were accurately predicted by the logistic regression model 83% of the time.

VII CONCLUSION AND FUTURE WORK

The ten machine learning algorithms used for data mining in this paper These algorithms were useful on the dataset to calculate the likelihood that a patient has heart disease and were examined with classification models.

To analyze the best method in terms of accuracy, these 10 algorithms are realistic to the same dataset. With an accuracy level of 83%, the logistic regression model predicted the heart disease patient. The best and most effective logistic regression approach for handling medical datasets, according to all observations.

In the future, more diseases may be predicted or identified using the computed system and machine learning classification algorithm. For the automation of heart disease analysis, incorporating some other machine learning techniques, the work can be enhanced or extended.

VIII REFERENCES

[1] Baban.U. Rindhe1 , Nikita Ahire2 , Rupali Patil3 , Shweta Gagare4 , Manisha " **Heart Disease Prediction Using Machine Learning**", International Journal of Advanced Research in Science, Communication and Technology (IJARSCT) Volume 5, Issue 1, May 2021.

[2] Rishabh Magar, Rohan Memane, Suraj Raut ,Prof. V. S. Rupnar ,Computer Department, MMCOE, Pune, India," **HEART DISEASE PREDICTION USING MACHINE LEARNING**", Journal of Emerging Technologies and Innovative Research (JETIR) June 2020, Volume 7, Issue 6.

[3] R.Indrakumari ,T. Poongodi,Soumya ranjan Jena "**Heart Disease Prediction using Exploratory Data Analysis**", International Conference on Smart Sustainable Intelligent Computing and Applications under ICITETM2020, 173 (2020) 130-139.

[4] Mr.Santhana Krishnan.J PG Student Department of Computer Applications Anna University, BIT Campus Tiruchirapalli-24 Dr.Geetha.S " **Prediction of Heart Disease Using Machine Learning Algorithms.**" International conference on innovation in information and communication technology (ICIICT)2019

[5] Mohammed Khalid Hossen Department of Computer Science and Engineering, Sylhet Agricultural University, Sylhet, Bangladesh, "**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.

[6] Chintan M. Bhatt 1,* , Parth Patel 1 , Tarang Ghetia 1 and Pier Luigi Mazzeo 2,* 1 Department of Computer Science and Engineering, School of Technology, India 2 Institute of Applied Sciences and Intelligent Systems, National Research Council of Italy, 73100 Lecce, "**Effective Heart Disease Prediction Using Machine Learning Techniques**", MDPI, 2023, 16, 88. https:// doi.org/10.3390/a1602008.

[7] Bhagyesh Randhawan1 , Ritesh Jagtap2 , Amruta Bhilawade3 , Durgesh Chaure4 1, 2, 3, 4Department of Information Technology, Parvatibai Genba Moze College of Engineering, Pune, "**Heart Disease Prediction Using Logistic Regression Algorithm**", International Journal for Research in Applied Science & Engineering Technology (IJRASET) ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538 Volume 10 Issue IV Apr 2022.

[8] P. Shashwith*1, B. Sai Prasad*2, CH. Shashi Kumar Reddy*3, N. Abhinav Krishna*4, K. Ramesh*5 *1,2,3,4 Student, *5Assistant Professor, Teegala Krishna Reddy Engineering College, Hyderabad, Telangana, India, " **HEART DISEASE PREDICTION BY USING MACHINE LEARNING ALGORITHMS**", International Research Journal of Modernization in Engineering Technology and Science, Volume:04/Issue:06/June-2022.

[9] Malavika G1 , Rajathi N2 , Vanitha V3 and Parameswari P4 1 PG Scholar, Department of Information Technology, Kumaraguru College of Technology, Coimbatore, India. 2,3Professor, 4 Assistant Professor (SRG), " **Heart Disease Prediction Using Machine Learning Algorithms**", Biosc.Biotech.Res.Comm. Special Issue Vol 13 No 11 (2020) Pp-24-27.

[10] BhaveshDhande1 , Kartik Bamble2 , Sahil Chavan3 , Tabassum Maktum4 1,2,3,4Ramrao Adik Institute of Technology, D Y Patil Deemed to be University, Navi Mumbai, India, "**Diabetes & Heart Disease Prediction Using Machine Learning**", ICACC-2022, ITM Web of Conferences 44, 03057 (2022).