ISSN No:-2456-2165

Hybrid Heart Disease Prediction Model using Machine Learning Algorithm

Ankita Singha[1], Anushka Sikdar[2], Palak Choudhary[3], Pranati Rakshit[4], Sonali Bhattacharyya[5]
[1,2,3] B. Tech Student, [4,5] Associate Professor
Department of Computer Science and Engineering,
JIS College of Engineering, Kalyani, India

Abstract:- Worldwide, machine learning is used in a variety of fields. Machine learning will be crucial in determining whether or not heart disorders will exist. If forecasted long in advance, such information will provide clinicians with crucial intuitions. The majority of our work focuses on applying machine learning algorithms to predict possible heart problems. We tend to compare classifiers such Naive Bayes, logistical Regression, SVM, XGBOOST, Random Forest, etc. during the course of this work. Since it will have a wide range of samples for coaching and confirmatory analysis, Random Forest suggests an ensemble classifier that does hybrid classification by using both strong and weak classifiers. As a result, we analyse planned and existing classifiers like Ada-boost and XG-boost that offer the highest accuracy and prognostication. The best accuracy is provided by XGBOOST (90.6%).

Keywords:- *SVM, Naive Bayes, Random Forest, logistic regression, Ada-boost, XG-boost, Python programming, confusion matrix, and matrix.*

I. INTRODUCTION

The World Health Organization estimates that cardiovascular disease causes 12 million deaths worldwide each year. One of the leading causes of death and disease around the world is cardiovascular disease. One of the most important topics in the area of information analysis is regarded to be the prediction of disorder. Since a few years ago, there has been a rapid increase in the amount of disorder everywhere in the world. Numerous studies are carried out to identify the most prestigious risk factors for cardiovascular disease as well as to precisely anticipate the risk. Cardiovascular disease is also referred to as a silent killer that kills a person without showing any evident signs. The first diagnosis of cardiovascular disease is crucial in helping patients decide whether to adjust their lifestyles and subsequently lowers the problems.

With the use of machine learning, the health care industry's huge volume of data may be used to make decisions and predictions. This study uses machine learning to analyse patient data and categorise whether or not they have cardiovascular disease in order to predict future cardiovascular disease. In this aspect, machine learning techniques are extremely helpful. Even though there are many different ways that cardiovascular disease can manifest, there is a common set of critical risk indicators that can determine whether someone is unquestionably at risk. We may determine that this method is suitable for using to attempt and conduct the prediction of cardiovascular illness by gathering the information from many sources, classifying them under relevant headings, and ultimately examining to make out the necessary knowledge.

Machine learning is unbelievably complicated and the way it works varies counting on the task and the algorithmic program accustomed accomplish it. However, at its core, a machine learning model could be a laptop viewing information and characteristic patterns, so victimization those insights to raised complete its allotted task. Any task that depends upon a group of information points or rules will be automatic victimization machine learning, even those additional complicated tasks like responding to client service calls and reviewing resumes. A Decision Process: normally, machine learning algorithms are accustomed create a prediction or classification. supported some input file, which might be tagged or unlabeled, your algorithmic program can manufacture associate estimate a few patterns within the information.

An Error Function: a blunder perform serves to judge the make out accuracy. If there are best-known examples, a blunder perform will create a comparison to assess the accuracy about our project.

A model improvement process: Weights are modified to reduce the difference between the model estimate and the best-known example if the model performs better with the data points in the coaching set. When a category label is anticipated for a specific example of an input file, this is referred to as classification in machine learning.

A. Supervised Learning

Supervised learning is a type of machine learning in which computers are taught to use carefully "labelled" coaching data and then make predictions about the outcome based on that data. According to the tagged information, some input files have already been labelled with the appropriate output.

Because the supervisor educates the machines to forecast the output correctly, the coaching information given to the machines in supervised learning is effective. It uses a similar idea to how a pupil learns while under the teacher's supervision.

One way to give the machine learning model the "input information input file computer file" in addition to the "right output data" is through supervised learning. The purpose of an algorithmic rule for supervised learning is to

ISSN No:-2456-2165

find a mapping operation to map the input variable (x) with the output variable (y).

There are following supervised machine learning algorithms:

- Linear Regression
- Logistical Regression
- Support Vector Machines (SVM)
- Neural Networks
- Random Forest
- Gradient Boosted Trees
- Decision Trees
- Naive Bayes

B. Unsupervised learning

Because, in contrast to supervised learning, we have the "input information input file computer file" but no corresponding output data, unsupervised learning cannot be applied immediately to a regression or classification problem.

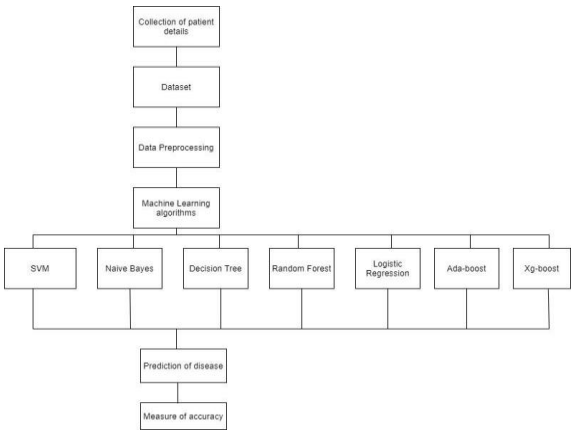

Fig. 1: System Architecture

Unattended learning aims to identify the underlying structure of a dataset, group related pieces of data together, and display that dataset in a very compact manner.

Unsupervised learning "is helpful" in extracting insightful information from the data. Unsupervised learning is far more similar to how a person learns to make assumptions based on their own experiences, which brings it closer to the $64,000 AI. Unattended learning becomes even more critical because unsupervised learning relies on unlabeled and uncategorized input.

In the actual world, we don't always have an input source and an output source, therefore unattended learning is ideal in these situations.

C. Reinforcement learning

Machine learning includes the field of reinforcement learning. It has to do with choosing the right course of action to maximise reward in a very particular circumstance. Different types of programming and machines utilise it to find the most straightforward behaviour or route it should take to deal with a specific situation. Reinforcement learning differs from supervised learning in every manner because in supervised learning, the model is finished because the crucial information is included.

II. RELATED WORK

Utilizing data from the UCI Machine Learning dataset, numerous studies and tests have been conducted to identify cardiac disease. Various data mining approaches have been used to obtain high levels of precision. These strategies are explained as follows:

Avinash Golande and colleagues investigate various Machine Learning techniques that can be applied to the classification and prediction of heart disorders. Research was conducted for the study, and knowledge of the decision tree, k-nearest neighbour, and k-means algorithms, which may be used for classification, was compared. This study comes to the conclusion that the Decision Tree receives accurate predictions. The utmost conclusion was that by combining several strategies and fine-tuning the parameters, it might be made effective.

T. Nagamani et al. developed a system that combined the MapReduce algorithm with data mining principles and practises. For the 45 instances in the experiment testing set, the accuracy obtained from this study was better than the accuracy obtained using a traditional fuzzy artificial neural network. As a result, the usage of linear scaling and dynamic schema increased the algorithm's accuracy.

Fahd Saleh Alotaibi developed a machine learning model by contrasting five alternative methods. The Rapid Miner tool was then used, which generated results that were more accurate than those from the MATLAB and Weka tools. In this investigation and experiment, the classification accuracy predictions of Decision Tree, Logistic Regression, Random Forest, Naive Bayes, and SVM were also contrasted. The findings produced by the decision tree algorithm were the most precise.

Anjan Nikhil Repaka, et al., offered a system for the prediction, accuracy, and output of the disease that provided NB (Na ve Bayesian) strategies for dataset division and the AES (Advanced Encryption Standard) algorithm for the security of data transfer.

Different division algorithms used for heart disease prediction were included in a survey conducted by Theresa Princy, R., et al. In order to classify the data for the survey, Naive Bayes, KNN (K- Nearest Neighbor), Decision Trees, and Neural Networks were utilised. The accuracy of the classifiers was then examined for a variety of present attributes.

By combining SVM and Naive Bayes classification, Nagaraj M. Lutimath et al. successfully predicted cardiac disease (Support Vector Machine). Mean Absolute Error, Sum of Squared Error, and Root Mean Squared Error were the principal metrics employed in the analysis, and it was established that SVM appeared to be a superior method than Naive Bayes in terms of precision.

After reading the aforementioned studies, the fundamental idea behind the suggested system was to develop a heart disease prediction system with necessary input.

By comparing the accuracy, precision, recall, and f-measure scores of the various classification algorithms, including Decision Tree, Random Forest, Logistic Regression, and Naive Bayes, we were able to determine which classification method would be most effective at predicting heart disease.

III. METHODOLOGY

A. Existing System

The silent killer of heart disease, which is a leading cause of death in people with no outward signs of the condition, is highlighted. The source of mounting worry about the illness and its effects is part of the essence of this sickness. As a result, constant effort is made.

B. Proposed System

Data gathering and the selection of critical attributes are the first steps in the system's operation. The necessary data is then pre-processed into the necessary format. Training and testing data are separated from the whole amount of data. The algorithms are used, and the training data is used to train the model. By analysing the system with the help of the testing data, the precision of the system is discovered. The modules listed below are used to run this system:

- Collection of Dataset
- Selection of attributes
- Data Pre-Processing
- Balancing of Data
- Disease Prediction

a) Collection of datasets:
We first gather a dataset for our algorithm that forecasts cardiac illness. We divide the dataset into training data and testing data after grouping it. The learning of the predicting model uses the training dataset, and the estimation of the predicting model uses the testing dataset. In this project, 70% of the data are used for training, while 30% are used for testing.

Heart Disease UCI is the dataset that was used for this project. There are 76 attributes in the dataset; 14 of them are utilised by the system.

b) **S**election of attributes
The process of choosing appropriate attributes for the prediction system is referred to as attribute or feature selection. By doing this, the system's effectiveness is improved. Numerous patient characteristics, including gender, the kind of chest pain, fasting blood pressure, serum cholesterol, exang, etc., are taken into account for the prediction. In order to choose the attributes for this model, the correlation matrix is used.

c) Pre-processing of data
A crucial first step in creating a machine learning model is data pre-processing. Initial data may not be accurate or in the model's required format, which could lead to misleading results. We frequently alter information during pre-processing so that it fits our specific needs. It won't deal with the dataset's noise, duplication, or missing values. Information pre-processing includes tasks like dataset import, dataset rendering, attribute scaling, etc. Pre-processing data is necessary to improve the model's accuracy.

Fig. 2: Information Pre-processing

d) Balancing of Data
Unbalance datasets would be adjusted in one of two ways: beneath sampling, or (a), and oversampling.

a. beneath Sampling:
By reducing the size of the large category in beneath Sampling, the dataset balance is completed. Once there is enough information, this strategy is taken into consideration.

b. Over Sampling
In this scenario, the dataset balance is accomplished by enlarging the size of the sparse samples. When there is not enough information, this strategy is taken into consideration.

e) Prediction of Disease
SVM, Naive Bayes, Decision Trees, Random Trees, Logistic Regression, Adaboost, and XG-boost are just a few examples of the many machine learning algorithms that are used for classification. Comparative analysis is done between algorithms, and the algorithm that provides the highest accuracy is then used to predict heart disease for patients.

C. Machine Learning Algorithms

Machine learning is a potent technology that is defined as the methodical examination of multiple algorithms that gives systems the potential to mimic human learning processes without the need for programming. Unsupervised learning, supervised learning, and reinforcement learning are the other divisions of machine learning.

ISSN No:-2456-2165

D. Naïve Bayes Algorithm

Naive A supervised learning method called the Thomas Bayes formula that looks for classification problems and is based on the Thomas Bayes theorem. It is mostly used in text categorization, which offers a large training set.

Naive Thomas Bayes Classifier is among the simplest and easiest Classification algorithms that aid in creating quick machine learning models that produce quick predictions.

Because it is a probabilistic classifier, it predicts based on the likelihood that an object will exist. Spam filtration, Sentimental analysis, and categorising articles are some examples of Nave Thomas Bayes algorithm applications in style.

It is a classification method that relies on the Bayes Theorem and the assumption of predictor independence. Simply put, a Naive Thomas Bayes classification presupposes that the presence of one particular feature in a larger class is unrelated to the presence of another feature.

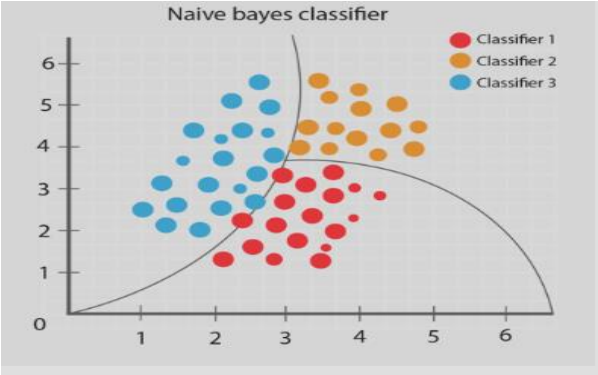

Fig. 3: Naïve Thomas Bayes Classifier

E. Support Vector Machine(SVM) :

One of the most well-known supervised learning algorithms, Support Vector Machine, or SVM, is used for both classification and regression problems. Nevertheless, it mostly addresses Classification challenges in machine learning.

The SVM algorithmic rule's objective is to create the simplest line or call boundary that will divide an n-dimensional space into categories, allowing us to quickly assign fresh data to the appropriate class in the future. A hyperplane's SVM selects the intense points and vectors that help create the border known as a best call boundary.

F. Decision Tree Algorithm

Decision trees are a supervised learning method that can be used to find classification and regression problems, however they are typically most frequently used to find classification problems. Internal nodes in the classifier's tree-like structure stand in for a dataset's possibilities, branches for the decision-making process, and each leaf node for the final classification outcome. There are two nodes in an option tree: the choice node and the leaf node.

While Leaf nodes are the result of these choices and don't have any additional branches, Decision nodes desire to build any call and have numerous branches. the choices or examinations are made in accordance with the options in the given dataset. It's a graphical representation of all possible decisions or solutions to a problem that are supported by the current situation. It is called a call tree because, like a tree, it starts with a base node and grows by adding more branches to form a structure resembling a tree. We frequently use the CART algorithmic programme, which stands for Classification and Regression Tree algorithmic programme, to generate trees.The call tree simply poses a question then supports the response (Yes/No). It does not further divide the tree into subtrees.

The supervised machine learning algorithm family includes the Decision Tree algorithmic programme. It is frequently used for both a classification and a regression drawback.

The objective of such a algorithmic programme is just to make predictions of the value of a target variable. To do this, a decision tree is used, in which the interior node of the tree contains diagrams of the characteristics and the leaf node corresponds to a category label.

The goal to keep in mind when creating a machine learning model is to use the simplest algorithmic programme for a given dataset and problem. There are several machine learning algorithms. The following list includes the two justifications for using the decision tree:

a) The branch is followed, and a subsequent node is jumped to using the (l dataset) attribute and comparison support.
   After comparing the attribute value with the opposing sub-nodes for each subsequent node, the algorithmic programme proceeds on. The process is carried out until the tree's leaf node is reached. Using the following algorithm, the entire procedure is frequently easier to comprehend:
   - Step-1: S advises starting the tree at the base node, which holds the entire dataset.
   - Step-2: Take note of the dataset's most basic attribute, which is Attribute Choice Live (ASM).
   - Step-3: Subsets of the S that include possible values for the most basic properties should be created.
   - Step-4: Create the tree node of your choice that has the most basic characteristic.
   - Step-5: Recursively develop new call trees by using the subsets of the step 3-created dataset. Continue using this technique until you can no longer categorise the nodes, at which point you will declare the final node to be a leaf.

G. RANDOM FOREST ALGORITHM

A supervised learning algorithmic software called Random Forest may exist. It's an extension of machine learning classifiers that uses sacking to improve call Tree performance. It combines tree predictors, and trees are

ISSN No:-2456-2165

focused on a randomly sampled vector. All trees have a consistent distribution.

On randomly chosen knowledge samples, Random Forests generate decision trees, obtain predictions from each tree, and then vote on the best option. Additionally, it gives a clear indication of how important the function is.

Random Forest may be a classifier that uses a number of call trees on various subsets of the supplied dataset and uses the average to improve the prognosis accuracy of that dataset. The random forest predicts the final output by using predictions from all trees and supported by the majority votes of forecasts, as opposed to just one call tree.

The greater variety of trees inside the forest results in greater accuracy and avoids the issue of overfitting.

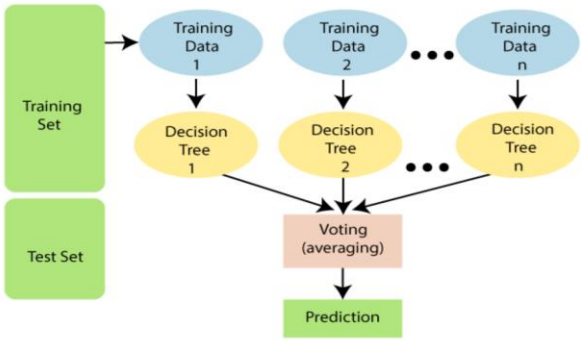

Fig. 3: Random Forest Algorithm

H. LOGISTIC REGRESSION ALGORITHM

One of the most popular Machine Learning algorithms that falls under the superior education approaches is logistic regression. It is passed down anticipating a particular variable quantity using a particular collection of independent variables.

In supply regression, we typically deal with a "Shaped supply operation, that predicts 2 most values" rather than fitting a regression curve (0 or 1).

The curve from the supply function shows the likelihood of anything, such as whether or not the cells are cancerous or not, whether or not a mouse is heavy or not supported by its weight, etc.

Because of its ability to categorise fresh data using distinct and continuous datasets, logistic regression may be a crucial machine learning strategy.

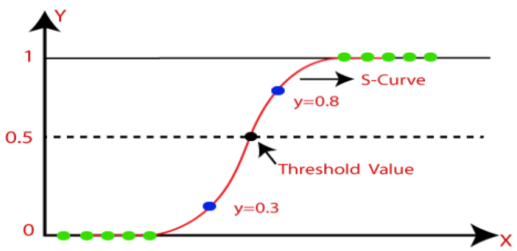

Fig. 4: Logistic Regression

It is tough to capture advanced relationships exploitation supply regression. Additional powerful and complicated rules like Neural Networks will simply shell this algorithm.

I. ADABOOST ALGORITHM

The first tremendously invigorating formula created with binary classification in mind was called ADABOOST. The acronym ADABOOST stands for "adaptive Boosting," and it is a well-known boosting approach that combines several "weak classifiers" into one "strong classifier."

At first, ADABOOST randomly chooses a coaching set.

By selecting the coaching set that supported the accurate forecast of the previous coaching, it iteratively trains the ADABOOST machine learning model.

It gives incorrectly classified observations a higher weight so that they will have a higher chance of being correctly classified in the following round.

Additionally, it shifts the responsibility to the trained classifier in accordance with the classifier's improvement with each iteration. Additional accurate classifiers may be given more weight.

This process is repeated till the entire coaching. The best boosted classifier is created by adding these classifiers in a weighted manner. Every classifier's accuracy is measured using weights.

Changing weights AdaBoost requires a learning formula that takes into consideration the weighted input instances, meaning that the loss perform should give heavier examples more weight.

By selecting the coaching set that supported the accurate forecast of the previous coaching, it iteratively trains the ADABOOST machine learning model.

It gives incorrectly classified observations a higher weight so that they will have a higher chance of being correctly classified in the following round.

Additionally, it places responsibility on the trained classifier in each iteration in accordance with the classifier's validity. Additional accurate classifiers may be given more weight.

J. Xgboost algorithm

Gradient Boosted call trees are implemented in part by XG-boost. It is a type of code library that was created primarily to improve model performance and speed. Call trees are generated using this approach in consecutive kinds. The weights that are used in XG-boost are crucial. Any or all of the independent variables are given weights before being placed into the choice tree that forecasts outcomes. The weight of variables that the tree incorrectly predicted in advance is increased, and these variables are subsequently sent to the second call tree. These distinct classifiers/predictors are then combined to provide a robust

ISSN No:-2456-2165

and more accurate model. Regression, classification, ranking, and user-generated.

- The Power of XGBoost

The beauty of this powerful algorithm lies in its scalability, which drives superfast learning through parallel and distributed computing and offers efficient memory usage.

IV. RESULTS & DISCUSSION

After learning the machine learning algorithms, we are getting outputs represented in tabular Form.

Rhythm	Accuracy
XGBoost	90.6%
SVM	82.5%
Logistic Regression	83.5%
Random Forest	90.2%
Naive Bayes	76.9%
Decision Tree	89.3%
Adaboost	83.4%

Table 1: Accuracy comparison of algorithms

The highest accuracy/precision is given by the XGBOOST algorithm.

V. PERFORMANCE ANALYSIS

In this project, numerous machine learning algorithms like SVM, Naive mathematician, call Tree, Random Forest, provision Regression, ADABOOST, XG-boost square measure accustomed predict heart condition. Heart condition UCI dataset, incorporates a total of seventy-six attributes, out of these solely fourteen attributes square measure thought for the prognosis of heart condition. Numerous attributes of the patient like gender, hurting kind, abstinence force per unit area, body fluid cholesterol, heart disease etc. Square measure thought of for this project. Any algorithm that provides the simplest Accuracy must be used for specific algorithmic programmes. For the forecast of the gastrointestinal ailment, that is considered. Numerous analysis criteria, including accuracy, confusion matrix, precision, recall, and f1-score, have been considered for analysing the experiment.

VI. CONCLUSION AND FUTURE SCOPE

Application of promising technology, such machine learning, to the initial prognosis of heart problems can have a significant influence on society. Heart diseases are a major cause of death in India and around the world. The first cardiac problem prognosis will help in making decisions about behaviour adjustments in high-risk patients and progressively reduce the complications, which may be an excellent milestone in the medication industry. The number of people suffering from heart disease is increasing yearly. This helps in both early detection and treatment. The medical community and patients will find the use of suitable technology support in this regard to be of great benefit. SVM, call Tree, random forest, naive Thomas bayes, logistical regression, reconciling boosting, and extreme gradient boosting are just a few of the seven distinct types of machine learning algorithms that will be tested in this research as they are applied to the dataset.

Seventy-six possibilities make up the dataset that contains the expected characteristics that lead to cardiac conditions in patients, and fourteen crucial options that are useful for assessing the system are selected among them. One prediction model was created, as evidenced by the comparison of the seven machine learning methods' accuracies. Therefore, it is intended to apply a variety of analysis metrics, with XGBOOST providing the best accuracy (90.6%).

ACKNOWLEDGMENT

We acknowledge the persons who helped us in doing this present work.

REFERENCES

[1.] Soni J, Ansari U, Sharma D & Soni S (2011). anticipating data analytics for medical diagnosis: a form for predicting heart condition. International Journal of pc Applications, 17(8), 43-8

[2.] Dangare C S & Apte S S (2012). Improved study of heart condition prediction system using data analytics classification techniques. International Journal of pc Applications, 47(10), 44-8.

[3.] Ordonez C (2006). Association rule discovery with the train and test approach for predicting heart disease. IEEE Transactions on IT in Biomedicine, 10(2), 334-43.

[4.] Shinde R, Arjun S, Patil P & Waghmare J (2015). An intelligent heart disease predictive system using k-means clustering and Naïve Bayes algorithm. International Journal of CS and Information Technologies, 6(1), 637-9.

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

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

[7.] Jabbar M A, Deekshatulu B L & Chandra P (2013, March). Prediction of heart disease using lazy associative classification. In 2013 International Mutli-Conference on Automation, Computing, Communication.