INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE

ISSN: 2057-5688

Cardiovascular Disease Forecasting By Using Various Supervised and Unsupervised Machine Learning Algorithms

1. S RamaKrishnaVasamsetti, Research Scholar, Department of Computer Science and Engineering, J.S University, Shikohabad, U.P.
2. Dr. Suribabu Potnuri, Professor, Supervisor, Department of Computer Science and Engineering, J.S University, Shikohabad, U.P.

ABSTRACT

There is a tremendous amount of potential for the use of artificial intelligence in the healthcare sector to improve the overall quality of the services that are offered. This research tries to determine whether or not a person is suffering from cardiovascular disease by predicting whether or not they have the condition. Considering the considerable amount of data that is generated by the health care business, it is imperative that certain approaches be used in order to handle this data. Diseases of the cardiovascular system are the leading cause of death on a worldwide scale. This method makes predictions on the likelihood of the emergence of cardiovascular disease. For the purpose of solving the prediction issue, supervised machine learning techniques are used. Within this particular situation, the dataset that is being used consists of 303 unique entries and incorporates 14 different components. In order to prepare the dataset for the creation of the model, preprocessing is performed. In order to anticipate the accuracy, a number of different machine learning methods are used. These algorithms include Logistics Regression, Naïve Bayes, Support Vector Machine (SVM), K Nearest Neighbor (KNN), Decision Tree (DT), XGBoost, and Neural Networks. According to the results of the experiment, the accuracy rates for the Naïve Bayes, LR, KNN, DT, XGBoost, and Neural Network models were as follows: 85.25%, 85.25%, 81.97%, 67.21%, 81.97%, 85.25%, and 83.61%, respectively.

INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688

Keywords: DecisionTree,Heartdisease,KNN,LogisticRegression,MachineLearning,NaïveBayes, Neural Network, SVM, XGBoost

INTRODUCTION

On a yearly basis, heart disease is responsible for around 12 million deaths throughout the world, as stated by the World Health Organization (WHO). Heart disease is a major contribution to the number of deaths and illnesses that occur all over the globe. In the realm of data analysis, the prediction of cardiovascular disease is a topic that has a significant amount of importance. across the course of the last few years, there has been a significant rise in the prevalence of cardiovascular disease all across the world. Research has been carried out in a number of different ways in order to identify the key risk factors for cardiovascular disease and to appropriately evaluate the overall risk. Heart disease is sometimes referred to as a "silent killer" due to the fact that it might possibly result in a person's death without any obvious symptoms being present. The early diagnosis of heart illness is very important for those who are at a high risk because it allows them to formulate well-informed choices about the modification of their lifestyle, which in turn reduces the likelihood of adverse consequences. The use of machine learning makes the process of generating judgments and projections based on the large amount of data that is created by the healthcare business more simpler. By conducting an analysis of patient data using a machine-learning system that determines whether or not a patient has heart disease, the purpose of this study is to make predictions about the incidence of heart disease in the present and the future. In the context of this specific circumstance, machine learning algorithms could be able to give substantial aid. There is a uniform set of basic risk factors that characterize an individual's vulnerability to the illness, despite the fact that the manifestations of heart disease might vary from person to person. Because it includes gathering data from a variety of sources, classifying it in an appropriate manner, and then analyzing it in order to get the information that is required, this method is an excellent choice for predicting cardiac illness.

MOTIVATIONFORTHEWORK

The development of a prediction model for the course of cardiovascular disease is the primary purpose of this body of study. The primary purpose of this research is to

ISSN: 2057-5688

determine which classification system is the most efficient in determining whether or not a patient is suffering from heart disease. For the purpose of providing support for this research, this study makes use of three separate classification techniques, namely Nave Bayes, Decision Tree, and Random Forests, in order to carry out a comparative analysis and evaluation at various levels of assessment. Although technologies based on machine learning are often used, effectively predicting cardiovascular illnesses is a crucial endeavor that calls for the highest possible degree of accuracy. Consequently, a wide variety of assessment procedures and degrees are used in order to assess the three algorithms. Scientists and medical professionals will be able to benefit from this by developing an improved

PROBLEMSTATEMENT

The most significant obstacle that is linked with heart disease is the discovery of the condition. Despite the fact that there are methods available for forecasting heart disease, these methods are either expensive or ineffective when it comes to precisely measuring the likelihood of heart disease in individuals. The diagnosis of heart disorders at an earlier stage has the potential to reduce the overall ramifications as well as the mortality rate. There are times when it is not possible to monitor a patient on a daily basis in a manner that is both consistent and appropriate. This is because it requires a higher level of intelligence, time, and skill. Furthermore, it is not possible for a physician to take part in continuous conversations with a patient throughout the whole of a twenty-four hour period. Because of the wealth of data that is available to us in the current day, we are able to make use of a variety of machine learning approaches in order to discover patterns that were previously hidden. It is possible to utilize medical data to discover hidden patterns that might be of assistance in the diagnosis of patient health.

.

LITERATURE SURVEY

A number of important academic publications have been produced as a result of recent investigations and research conducted in the disciplines of medical science and machine learning.

The "Efficient Heart Disease Prediction System" that Purushottam and his colleagues

ISSN: 2057-5688

developed included hill climbing as one of its components.

as well as procedures that are based on decision trees. Pre-processing was performed on the Cleveland dataset before proceeding with the categorization of the data collected. When there are missing values in Knowledge Extraction, the data mining technique known as Evolutionary Learning (KEEL), which is publicly accessible, is used to fill them in. A top-down strategy was used in the implementation of decision trees. At each and every level of the examination, a node that has been selected via the use of a hill-climbing algorithm is evaluated. Values and parameters that are associated with confidence. Having a degree of confidence of 0.25 is the bare minimum. 86.7% of the time, the system is accurate.

The prediction of heart illness was accomplished by Santhana Krishnan and colleagues via the use of decision tree and Naive Bayes algorithms. In order to generate binary verdicts, which may be either True or False, decision tree algorithms make use of the conditions. Both SVM and KNN use split conditions that are either vertical or horizontal, depending on the variables that are reliant on them. One kind of structure that is similar to trees is called a decision tree. It is made up of a root node, branches, and leaves within its structure. All of these components are established by the decisions that are taken at each node of the tree. In addition to this, decision trees are used to attract attention to the properties of the dataset. It was decided to utilize the data from Cleveland. The dataset is divided into two sets: a training set that contains seventy percent of the data, and a testing set that contains thirty thousand percent of the data. Accomplished a level of accuracy of 91%. For the classification process, the Naive Bayes method is used. Because it is able to analyze complex, non-linear, and interconnected data, it is well suited for processing heart disease datasets, which are also complex in their own right. A level of accuracy that is 87%.

The research paper titled "Prediction of Heart Disease Using Machine Learning Algorithms" authored by Sonam Nikhar and her colleagues offers a comprehensive description of the use of Naïve Bayes and decision tree classifiers in the process of predicting heart disease. When it comes to the processing of perceptual data utilizing the same dataset, the Decision Tree method displays greater performance when

ISSN: 2057-5688

compared to the Bayesian classifier. A multi-layer feed-forward neural network approach was used by Aditi Gavhane and colleagues in their research project named "Prediction of Heart Disease Using Machine Learning" in order to generate and assess datasets. The following layers make up this technology: a single input layer, a single output layer, and one or more hidden layers that are positioned in between the two levels. All of the connections between the input nodes and the output nodes are provided by the hidden layers. There are weights that are intrinsic to these linkages. While the nodes may either be convolutional or feedback, the weights of the input with bias are assigned depending on the requirements of the situation.

A study article titled "Heart Disease Prediction Using Effective Machine Learning Techniques" was recently published by AvinashGolande and colleagues. This work makes use of a variety of data collecting techniques in order to assist medical professionals in distinguishing between various types of cardiovascular sickness. k-nearestneighbor, decision tree, and Naïve Bayes are some of the algorithms that are used in this process. Packing calculation, component thickness, sequential negligible standardization, neural systems, straight Kernel self-arranging guidance, and Support Vector Machine (SVM) are some of the other methodologies that involve characterization-based methods. The application of "Machine Learning Techniques for Heart Disease Prediction" was suggested by Lakshmana Rao and colleagues, along with other notable contributions. Consequently, the procedure of detecting heart illness is a difficult one. In order to determine the degree of cardiac disease, data collection methods and neural networks are used in the process of conducting the evaluation

Through the use of convolutional and recurrent neural networks, Abhay Kishore and his colleagues suggested the utilization of "Heart Attack Prediction Using Deep Learning" in order to anticipate heart-related viral diseases. The most accurate and perfect model may be obtained with the use of deep learning and data mining. The findings of this study provide a reliable reference for predicting heart attacks.

Senthil Kumar Mohan and his colleagues submitted a research article with the title "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques" with

ISSN: 2057-5688

the intention of enhancing the precision with which cardiovascular problems may be predicted. An algorithm that combines a linear model with a hybrid random forest is used in the model that predicts cardiac disease. The K-Nearest Neighbors (KNN) method, the Logistic Regression (LR) algorithm, the Support Vector Machine (SVM) technique, and the Neural Network (NN) algorithm are brought together in this model. With an accuracy of 88.7% (HRFLM), the technique accomplishes a remarkable breakthrough in terms of improvement.

A model was developed by Anjan N. Repaka and colleagues that compares the prediction capacity of two different categorization models to any previous study that has been conducted. The results of the experiments demonstrate that the method that we have proposed is more accurate than the models that have been used in the past when it comes to estimating risk percentages.

"Heart Disease Prediction using Evolutionary Rule Learning" is the title of the work that was written by Aakash Chauhan and his colleagues. Records that are stored electronically make the process of manually accessing data more efficient. Additionally, there is a reduction in the number of services available, as well as a plethora of restrictions that properly predict cardiac illness. Using patient data to perform frequent pattern growth association mining results in the formation of strong associations.

EXISTINGSYSTEM:

In order to guarantee that they are in accordance with the most recent health regulations, tools and procedures are subjected to regular testing. Making use of machine learning might prove to be advantageous. There is a wide variety of ways in which heart disease may present itself; nonetheless, there are underlying risk factors that determine the possibility of a person acquiring heart disease. In order to arrive at conclusions, it is necessary to first gather data from a wide variety of sources, then classify that data into appropriate categories, and last analyze the data. An very high degree of accuracy is shown by this method when it comes to predicting heart illness.

Disadvantages

1. Since the system is mostly dependent on human input, it is essential to have the knowledge and experience of

ISSN: 2057-5688

experienced physicians in order to make correct predictions.

2. Work that is done manually takes a lot of time.

Method

In order for the system to function, it must first collect data and then determine the essential properties. The next step is to do pre processing on the data. Several distinct sets are created from the training data and the testing data respectively. It is via the use of algorithms and training data that the model is trained. It is the outcomes of the tests that will indicate whether or not the system is proper.

Advantages

Prediction of renal disease using automated approaches.

The speed of computer-assisted techniques is shown to be higher.

Patients are able to get quick medical care because to the accuracy of this technology, which ultimately leads to their recovery and survival and ultimately.

METHODOLOGY

This system does contain the implementation of these modules as part of its functionality.

The Accumulation of Datakits

Choosing the characteristics to be used

Third, the Pre-Processing of Data

The balance of data

Prediction of Disease Status

Collectionof dataset

Utilizing a dataset is the first step in the process of putting our heart disease prediction system into functioning. A number of different sets were created from the dataset in order to facilitate training and testing sessions. The use of training and testing datasets is an integral part of the process of constructing and evaluating prediction models. The data for this project is divided as follows: 70% is used for training, and 30% is used for testing. As part of this investigation, the Heart Disease UCI dataset was used. When it comes to the 76 attributes

ISSN: 2057-5688

that make up the dataset, only 14 are regarded to be system-relevant.

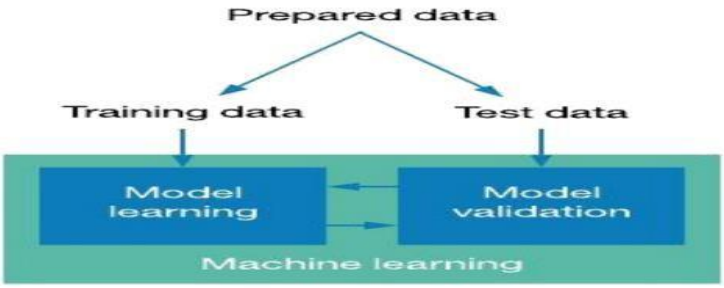

Figure: Collection of Data

Attributesselection

When it comes to the prediction system, attribute or feature selection refers to the process of picking qualities that are appropriate for the system. to maximize the efficiency of the system. A number of factors, including gender, the kind of chest pain, blood pressure when fasting, serum cholesterol levels, and the existence of exercise-induced angina, are taken into consideration when making predictions. For the purpose of termining attribute selection, this technique makes use of the correlation matrix.

INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688

Figure: Correlation matrix

Pre-processingofData

The construction of machine learning models requires the inclusion of data pre-processing as a key component. There is a possibility that the original data is tainted or not in the format that is wanted, which might result in erroneous results. The method known as "data pre-processing" is responsible for organizing and structuring the data. Controls the noise in the dataset, gets rid of duplicates, and deals with missing values. The process of importing data, separating datasets, and scaling characteristics are all examples of jobs that fall under the category of "data pre-processing." The accuracy of the model is improved by the use of data preprocessing.

INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688

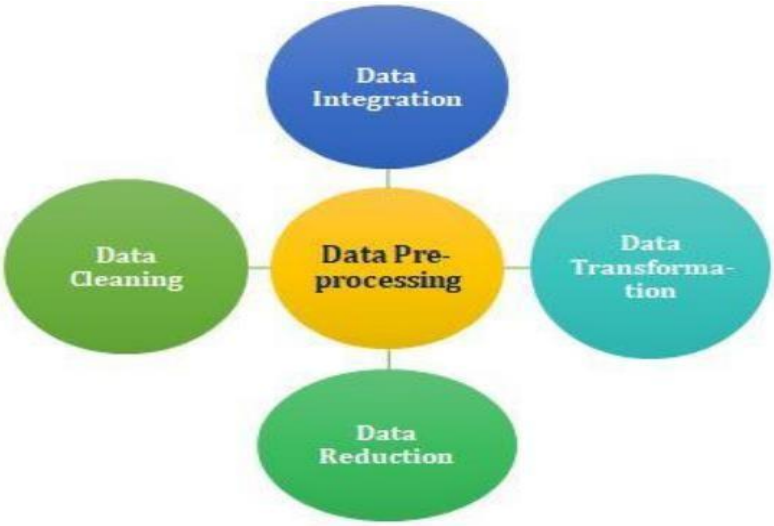

Figure: Data Pre-processing

Data balancing

When it comes to addressing the problem of unbalanced datasets, there are two different approaches that may be used. Using both under sampling and oversampling I have gathered a tiny piece to serve as an example that is representative: Datasets are equalized by the use of under sampling, which reduces the dominating class. The information that was supplied is sufficient for this process. [b] Excessive sampling: Over Sampling is a technique that enhances limited samples in order to equalize datasets. This is considered to be the case when there is a restricted quantity of data accessible.

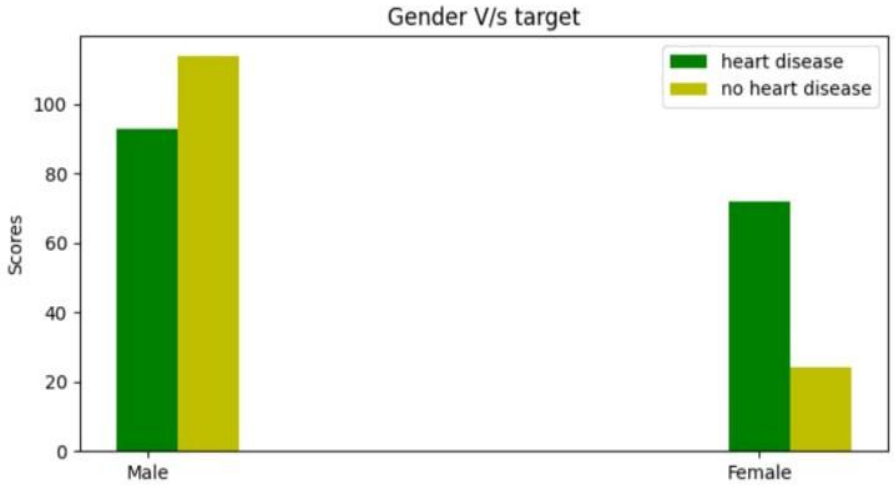

Figure: Data Balancing

INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688

PredictionofDisease

A wide variety of machine learning techniques are used in the classification process. These algorithms include Support Vector Machines (SVM), k-Nearest Neighbors (KNN), Naive Bayes, Decision Trees, Random Trees, Logistic Regression, Neural Networks, and Xg-boost. The algorithm that is able to attain the highest degree of accuracy is the one that is able to produce estimates about heart disease.

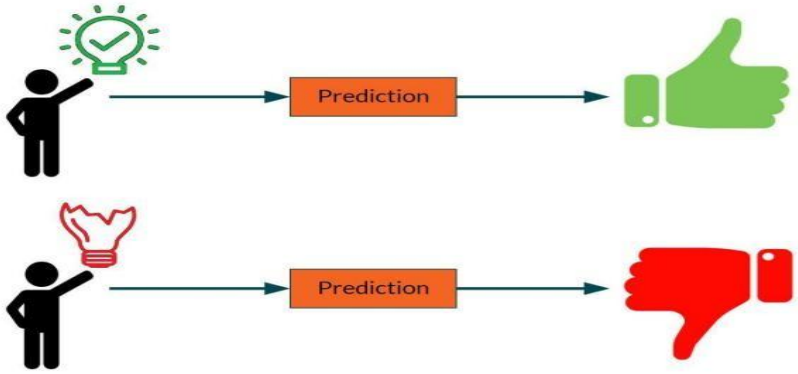

Figure: Prediction of Disease

MACHINE LEARNING

In the field of machine learning, classification refers to the process of predicting a certain class label for a given input data sample. Tutoring services When it comes to forecasting output, supervised learning makes use of training data that has been precisely categorized. Information that already has the appropriate output associated with it is referred to as labeled data. For the purpose of effectively predicting output, supervised learning is dependent on training the data. It operates according to the same premise as a pupil who is directed by a teacher. When using supervised learning, you are responsible for supplying the machine learning model with both input and output data. In the process of supervised learning, the objective is to identify a mapping function that establishes a connection between the input variable, x, and the output variable, y.

ISSN: 2057-5688

Teaching oneself The difference between supervised and unsupervised learning is that the former depends only on input data, while the latter does not have any output data that corresponds to it. The process of unsupervised learning involves determining the underlying structure of a dataset, grouping together data points that are linked to one another, and reducing the complexity of the data. Information that is useful may be gleaned from data via unsupervised learning. Because it simulates the process by which individuals gain thinking skills via experience, unsupervised learning is a method that comes very near to providing an accurate representation of artificial intelligence. As it makes use of data that is neither labeled or classified, unsupervised learning is an extremely important technique. When there is a discrepancy between the data that is entered and the data that is produced in real-world circumstances, it is vital to do so.

Enhancing Strength The field of machine learning includes a component known as reinforcement learning. The implementation of appropriate methods to maximize the achievement of benefits is the primary emphasis of this endeavor. In a certain situation, it makes it easier for software and robots to choose the most appropriate course of action to take. In the process of supervised learning, the training data is comprised of the answer key, which enables the model to be trained using the correct response. In the case of reinforcement learning, there is no solution that has been predefined; rather, the reinforcement agent is responsible for independently determining the greatest possible course of action to achieve the target that has been presented. It is able to gain information by experience learning, without depending on a training dataset that has already been established.

NAIVEBAYESALGORITHM:

A Bayes' theorem-based supervised learning approach is described here. There is an algorithm known as Naive Bayes that is capable of successfully addressing classification challenges. The majority of the time, text classification is accomplished using high-dimensional training datasets. The Naive Bayes Classifier is a straightforward and very

ISSN: 2057-5688

effective classification method that is used in the field of machine learning for the purpose of constructing faster models that may be subject to assumptions. It functions as a probabilistic classifier by making predictions based on the likelihood of the objects being considered. Applications of the Naive Bayes Algorithm include the filtering of spam, the study of sentiment, and the classification of articles. Classification based on Bayes' Theorem is predicated on the premise that estimators are independent of one another. The notion that the properties of a class are independent of one another is the foundation upon which a Naive Bayes classifier is built. The Naive Bayes algorithm is not only easy to develop, but it also has advantages when used to large datasets. The Naive Bayes algorithm is superior to even the most complex classification methods. The algorithm known as Naive Bayes: Poor in experience: This presupposes that one attribute is not affected by any other elements that may be present. Apples are defined by their red hue, their spherical form, and their flavor, which is characterized by a delightful sweetness. Apples are also known as a fruit. Each of the apple's distinct characteristics functions as a separate identification for the company.

Based on the fact that it makes use of Bayes' Theorem, the name "Bayes" was given to the instrument. The mathematical formula known as Bayes' theorem, which is also widely referred to as Bayes' rule or Bayes' rule, is used to determine the likelihood of a hypothesis based on information that has been gained in the past. The conditional distribution is a statistical tool that is used to determine the likelihood of an event taking place in the presence of another event that has already taken place.

Taking into account the fact that event B was witnessed, the posterior probability is the chance that hypothesis A is correct. The degree of evidence that supports the truth of a claim is represented by the likelihood probability, which is denoted by the symbol P(B|A).

The initial probability that is given to a hypothesis before any evidence is taken into consideration is referred to as the prior probability theory.

ISSN: 2057-5688

Marginal likelihood: Additionally known as P(B), this refers to the likelihood of evidence.

NaiveBayesmodeltypes:

Below is a list of the three Naive Bayes models that are available:

The Gaussian model is based on the assumption that particulars are distributed according to a normal distribution. Under the assumption that continuous predictors are drawn from a Gaussian distribution, the model is constructed.

The multinomial known as the multinomial When it comes to dealing with multinomial data, the Naïve Bayes classifier is particularly intended to handle it. The basic objective of document classification is to classify documents into various classifications, such as those pertaining to sports, politics, education, and other related subjects. On the basis of the frequency of the words, the classifier may be predicted.

There are some similarities between the Bernoulli classifier and the Multinomial classifier; however, the Bernoulli classifier employs independent Boolean variables as predictor variables. Finding out whether or not a certain word is included in a recorded document. The usefulness of this approach in the classification of documents is already well-known and well recognized.

DECISION TREE ALGORITHM

The Decision Tree is a method for supervised learning that may solve problems involving classification as well as regression; however, it is more effective when used to classification tasks. Tree-structured classifiers are a kind of network in which nodes reflect the characteristics of the dataset, branches represent the decision rules, and leaf nodes indicate the results of the classification. The nodes that make up a decision tree are divided into two categories: decision nodes and leaf nodes. On the other hand, leaf nodes are in charge of distributing the decisions that have been made, and they do not have any branches. Decision nodes are responsible for making choices and have several branches. The properties of the dataset are responsible for determining decisions and tests. The application presents all of the potential answers to an issue or option depending on the conditions that have been specified. This structure is referred to as a "Decision Tree" due to the fact that it begins with a root node and then branches out in a tree-like fashion. It is possible to

ISSN: 2057-5688

create trees using the CART technique for the purposes of classification and regression. By dividing the tree into subtrees according to the answer (Yes/No), a Decision Tree may be created. The Decision Tree Algorithm is an example of another kind of machine learning that is supervised. Additionally, it is able to classify numerical quantities and make predictions about them. In order to make a prediction about the value of a target variable, this method makes use of a decision tree. A particular class label is represented by the leaf node, while additional properties are represented by the interior nodes. In the process of building a machine learning model, it is of the utmost importance to choose the approach that is the most appropriate by taking into consideration the features of the dataset as well as the nature of the problem that is being addressed.

TheDecisionTreehastwo benefits:

As a result of their tree-like structure, decision trees are simple to understand. Furthermore, the hierarchical tree structure of decision trees makes it possible to comprehend the logic being presented.

During the process of attribute selection, it might be difficult to determine which attribute corresponds to the root node of each level in a Decision Tree.

Twoattributeselectionmeasuresare popular:

Information Gain: The nodes of the Decision Tree take training examples and split them into smaller groups, which results in a change in the entropy. The measurement of the change in entropy is known as information gain. The term "entropy" refers to the degree of impurity in a collection that does not have any particular order or structure among its components, as well as the measure of uncertainty that is connected with a random variable. It is possible to estimate the amount of disorder or randomness that exists inside a system using entropy. The term "entropy" refers to the degree to which a system gets more disorganized or unpredictable as it increases. The quantity of information that is housed inside the system has likewise increased, which corresponds to the rise in entropy that has already occurred.

The Gini Index is a statistical measure that estimates the frequency with which an

ISSN: 2057-5688

element that is randomly picked is identified wrongly. A personality trait that has a lower Gini index is preferred. When it comes to computing the Gini Index, Sklearn uses the term "gini" as its default criteria.

DecisionTreealgorithms include:

The Information Gain algorithm is employed by the IDichotomizer 3 (ID3) algorithm in order to determine the feature that is applied in the process of classifying the current subset of observations. For each level of the tree, use recursion to compute the information gain that has been gained. It is the successor of ID3 that is C4.5. For the purpose of categorization, this approach employs either the information gain or the gain ratio. As a result of the technique's ability to successfully handle both continuous and missing attribute data, it offers improvements over the ID3 method.

CART, which stands for the Classification and Regression Tree, is a dynamic learning approach that builds regression and classification trees depending on the dependent variable.

Working:

By starting at the root node, decision trees are able to generate predictions about the classes that are included inside a dataset. A comparison is made between the attribute values of the root and the attribute values of the record in the real dataset using this technique, which evaluates the root's attribute values. Following that, it begins to follow the branch that corresponds to it and then carries on to the subsequent node. A comparison is made between the attribute value of the upcoming node and the value of the sub-nodes that are already there, and the algorithm then moves on appropriately. In order to reach the last node of the tree, the procedure will continue till it is finished.

Algorithm:

Step1:Startthetreewiththerootnode, S,whichhastheentire dataset.

Step-2:UseAttributeSelectionMeasureto findthebestdatasetattribute(ASM).

Step-3: Divide S into subsets with best attribute values.

Step4: Createthebest attributeDecision Treenode.

Step-5:Recursivelycreatenewdecision treesfrom step-3dataset subsets.

ISSN: 2057-5688

Continueuntilyoucannolongerclassifythenodesandcall the finalnode aleaf node.

- **RANDOMFORESTALGORITHM**

Random Forest is an example of a kind of learning method that is supervised. The performance of Decision Tree classifiers may be improved with the use of bagging, a method that is utilized in machine learning. There are three predictors that are used, and it is dependant on a vector that is independent and randomly picked. There is a uniform distribution of trees across the area. The process of splitting nodes in Random Forests is distinct from the process of separating them according to variables. As an alternative, they divided the nodes by randomly picking the subset of the node that had the best predictors. The temporal complexity of learning with Random Forests is O(M(dnlogn)), where M indicates the number of growing trees, n represents the number of occurrences, and d represents the data dimension. Additionally, it is able to classify numerical quantities and make predictions about them. a method that is not only simple to implement but also very flexible. Forested areas are made up of trees. There is a correlation between increased tree density and increased resilience and robustness in forests. Random Forests are used to create Decision Trees by using randomly chosen data examples, make predictions, and then aggregate votes in order to discover the best possible decision.

Furthermore, it emphasizes the value of characteristics in an engaging and effective manner. Feature selection, image categorization, and recommendation engines are all examples of applications that make use of Random Forests. As well as being able to identify sicknesses and fraudulent activities, it is also able to identify trustworthy loan applicants. In order to function properly, the Boruta technique requires and encourages the identification of significant attributes within a dataset. A significant number of people use the Random Forest supervised learning approach. The difficulties associated with machine learning classification and regression are successfully addressed by it.

ISSN: 2057-5688

Ensemble learning makes use of a number of different classifiers in order to tackle complex problems and improve the overall performance of the model. The Random Forest classifier is made up of a large number of decision trees that have been trained on various subsets of the dataset. Increasing the accuracy of its predictions is accomplished by the classifier via the process of averaging the predictions of these trees. Through the process of collecting the forecasts from each decision tree and finding the majority vote, the random forest algorithm is able to provide a prediction on the ultimate result. One way to improve accuracy and reduce the likelihood of overfitting is to increase the number of trees that are present in the forest.

Assumptions:

The random forest uses multiple trees to predict the dataset class, so some decision trees may predict correctly and others may not. All trees predict the correct output.

Thus,twoassumptionsforabetterRandom forest classifier:

- The dataset's feature variable should have actual values so the classifier can predict accurateresults rather than guesses.
- Treepredictions mustbe uncorrelated.

AlgorithmSteps:

Itworksinfoursteps:

- Randomizeadataset.
- Eachsample'sDecision Treeshouldpredict.
- Voteforeach prediction.
- Themost-voted predictionis thefinal prediction.

Advantages:

- RandomForestcanhandlelargedatasetswithhigh dimensionality
- improvemodelaccuracy,and
- preventoverfitting.

Disadvantages:

- RandomForestisnotbetterfor regressionthanclassification.

LOGISTICREGRESSION ALGORITHM

The approach of logistic regression, which is included in the category of supervised machine learning, is well recognized. As part of the classification process, it employs a number of criteria in order to create

ISSN: 2057-5688

predictions about dependent variables. The output of a categorical dependent variable may be predicted accurately with the application of a statistical technique known as logistic regression. In light of this, the result must be entirely unique. In place of binary values, which are either 0 or 1, it offers predicted values that fall somewhere in the range of 0 to 1. There are some parallels between Logistic Regression and Linear Regression; nevertheless, Logistic Regression is used in a different way. The statistical technique known as linear regression is used to address problems associated with regression, one of which is the prediction of a continuous outcome variable. For classification issues, on the other hand, logistic regression is a statistical approach that is utilized. The objective of this strategy is to forecast the likelihood of an event happening. Instead of using a linear regression line, logistic regression makes use of a sigmoidal logistic function in order to make predictions about two alternative outcomes: either 0 or 1. On the basis of the weight of mice, the logistic function curve may be used to make predictions on the possibility of cells developing into cancerous cells or mice becoming overweight. A classification procedure known as logistic regression may be used to assign labels and compute probability for both continuous and discrete datasets. Logistic regression can be used to assign labels.

Advantages:

One of the most straightforward approaches to machine learning, known as logistic regression, has the potential to improve the effectiveness of training in specific circumstances. There is no major need for processing resources when it comes to the training of the model for this approach. The training weights are used as indications of the relevance of each characteristic in the training process. In addition to this, there is the possibility of a link that is either positive or negative. The ability to discover correlations between features is one of the capabilities of logistic regression. Not like the Decision Tree or Support Vector Machine methods, this specific technique offers the benefit of being able to easily update models with new data. This is a significant advantage. Stochastic gradient descent is the method that is used to make updates. The Logistic Regression technique provides a mechanism for the development of well-calibrated probability as well as a strategy for categorizing data. Compared to

ISSN: 2057-5688

designs that just classify, this is superior. We could be able to draw the conclusion that the first training example is more accurate for the problem at hand if it has a probability of 95% for a certain class while another example has a probability of 55% for the same class.

Disadvantages:

For the purpose of making accurate predictions about probabilistic outcomes, logistic regression makes use of independent variables. This might result in the model being too fitted to the training set for datasets that have a high number of dimensions, which would result in the model's accuracy being exaggerated and its ability to reliably predict outcomes on the test set being hindered. During the training process, the following happens when the model is trained on a small dataset that has a high number of features. It is recommended to make use of regularization methods in order to reduce the problem of overfitting for datasets that include a high number of dimensions; nevertheless, this may lead to an increase in the complexity of the model under consideration. One of the potential consequences of having high regularization factors is that the training data may become overfit. Because logistic regression uses a linear decision surface, it is unable to tackle problems that include nonlinear relationships. When it comes to real-world data, there are very few cases in which information can be cleanly separated into different groups by utilizing a straight line. Because of this, increasing the number of attributes causes the data to become linearly separable in higher dimensions, which in turn causes the nonlinear properties to change.

Non-LinearlySeparableData:

Logistic regression struggles with complex relationships. Neural Network models can easily outperform this procedure.

INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688

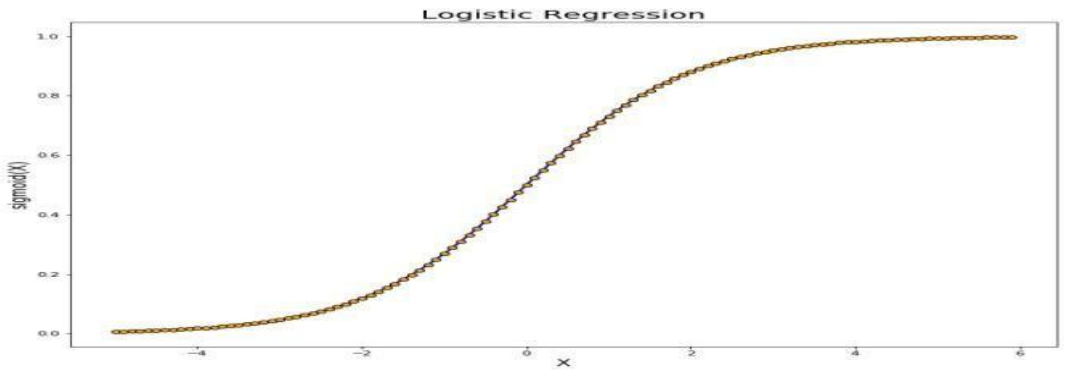

Figure: Logistic Regression

➢ ADABOOST ALGORITHM

First effective binary classifier boosting algorithm was Adaboost. Adaptive Boosting (Adaboost) is a prevalent boosting method that merges multiple "weak classifiers" into a single "strong classifier."

Algorithm:

1. Adaboost randomlychooses atraining subset.
2. It iteratively trains the Adaboost machine learning model by selecting the training set based on the last training's accurate assessment.
3. It gives wrongly classified observations more weight to increase their classification probability in the next iteration.
4. Eachiteration,itweightsthetrained classifierbasedonaccuracy.Accuracyis rewarded.
5. This process repeats until all training data fits correctly or the maximum number of estimators isreached.
6. "Vote"onallyourlearningalgorithmstoclassify.

Advantages:

Comparatively speaking, Adaboost algorithms are easier to use and need less tweaking of parameters than SVM methods. The usage of Adaboost in combination with SVM is a possibility. However, in principle, Adaboost applications do not experience overfitting since the parameters are not optimized concurrently and the learning process is slowed down by the estimate stage-wise. This prevents the problem from occurring. Because of this relationship, mathematics may benefit from it. In a variety of contexts, Adaboost has the

ISSN: 2057-5688

potential to improve the performance of image and text classifiers that are struggling.

Disadvantages:

Adaboost boosts by learning. Adaboost vs. Random Forest examples require high-quality data. It issensitive to outliers and noise in data and must be removed before use. XG-boost is faster.

XGBOOSTALGORITHM

In order to make use of gradient-boosted decision trees, the software package known as XG-boost is used. This software library was developed with the purpose of improving the overall performance of models. The decision trees that are produced by this method are generated in a sequential fashion. The XG-boost algorithm is affected by weights. When it comes to making predictions, the decision tree employs a method that involves giving weights to each of the independent variables used. Weights are applied to the variables that were incorrectly categorized, and then those weights are fed into the second decision tree. The robustness and accuracy of this model is provided by these classifiers and predictors. Among the tasks that the system is able to do are regression, classification, ranking, and user-defined prediction requirements.

In order to prevent overfitting, XG-boost makes use of regularization, namely L1 (Lasso Regression) and L2 (Ridge Regression) regularization. There is a common misconception that XG-boost is the regularized form of GBM, which is an abbreviation that stands for Gradient Boosting Machine. There are two regularization hyperparameters that are offered by Scikit Learn for XGBoost. These hyperparameters are alpha and lambda. In L1 regularization, the parameter alpha is used, but in L2 regularization, the value lambda is utilized. The performance of XG-boost is superior than that of GBM since it makes use of parallel processing. The model is designed to run on a large number of CPU cores. It is possible to do parallel processing with Scikit Learn thanks to the nthread hyper-parameter. The number of CPU cores that are going to be used is determined by the "nthread" option. In the event

ISSN: 2057-5688

that the nthread option is not supplied, the method will automatically detect all of the processor cores.

Missing values may be handled using XGBoost, which has this capacity. When XGBoost comes across a node that is lacking a value, it utilizes an approach that involves exploring both the left hand split and the right hand split that is available. After then, it decides which of the two splits will result in a greater loss. Subsequently, it applies the same method to the data that is being tested.

The functionality of cross-validation in XG-boost enables users to do assessments throughout each iteration of the boosting process. This simplifies the process of identifying the optimal

.

number of repetitions to perform within a single run. In contrast to the Gradient Boosting Machine (GBM), which requires a grid search and analyzes just a select few variables, this method does not need grid searches.

When a Gradient Boosting Machine (GBM) attempts to split a node and is unsuccessful, it will terminate. This is an efficient method of tree pruning. The conduct that it displays is one of greed. The data is partitioned using the XGBoost method up to the maximum depth, and then splits that do not result in any substantial improvement are eliminated thereafter.

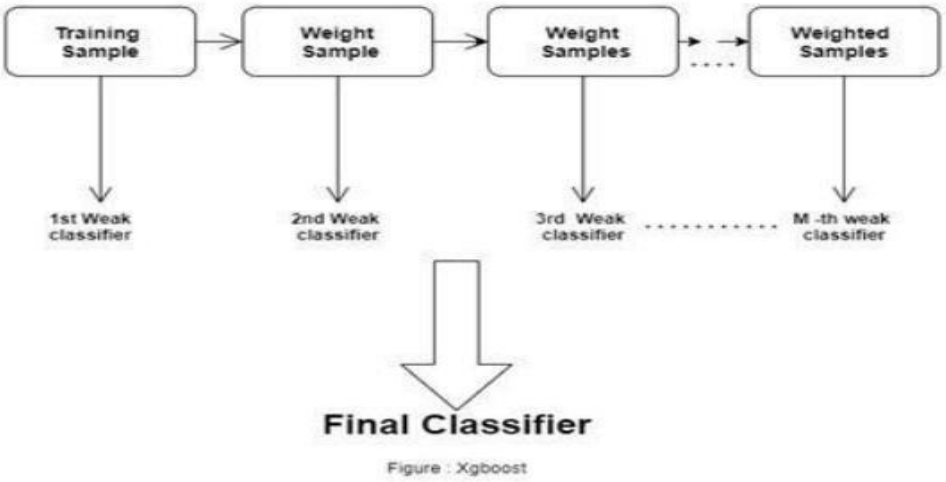

Figure : Xgboost

INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688

EXPERIMENTALANALYSIS

A. SYSTEMCONFIGURATION

- **Hardware requirements:**

Processor:AnyUpdateProcessor Ram : Min 4GB

HardDisk:Min100GB

- **Softwarerequirements:**

OperatingSystem:Windowsfamily

Technology : Python3.7

IDE:Jupyter notebook

- Implementation– Coding

CONCLUSIONANDFUTUREWORK

A major cause of mortality in India and throughout the world, heart disease may be predicted with the use of machine learning, which has the power to do so. An early detection of heart illness may be of assistance to high-risk individuals in making changes to their lifestyle and limiting the implications of such changes. This would be a tremendous accomplishment in the area of medicine. There is a growing number of people who are suffering from heart disease. Identification and treatment must be administered as soon as possible. It is possible that both patients and medical staff may benefit from the provision of appropriate technology help. The objective of this research is to assess the effectiveness of several machine learning techniques, such as Support Vector Machines (SVM), Decision Trees, K Nearest Neighbours, Naïve Bayes, Logistic Regression, Neural Networks, and Extreme Gradient Boosting, when applied to a specific dataset. There are 76 factors in the dataset that are used to predict the development of heart disease in people. For the purpose of evaluating the system, a subset of 14 relevant variables is chosen. Although all of the system's features are taken into consideration, the author notices a drop in the system's overall efficiency. Efficient operation is improved by attribute selection. An improvement in the accuracy of model evaluation may be achieved by optimizing the selection of n features. The properties of the dataset that would have been strongly connected have been removed. When all of the characteristics of the dataset are taken into consideration, the efficiency of the process drops. In order to construct a prediction model, a comparison is done between the accuracy of seven different techniques to

ISSN: 2057-5688

machine learning. For this reason, it is advised that the confusion matrix, accuracy, precision, recall, and f1-score be used in order to accurately anticipate disease. The techniques for machine learning that are known as logistic regression and extreme gradient boosting are two examples. 85.25% is the level of accuracy that the Naive Bayes model has.

REFERENCES

[1] Soni J, Ansari U, Sharma 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

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

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

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

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

ICCRDA 2020 IOP Conf. Series:Materials Science and Engineering 1022 (2021) 012072 IOP Publishing doi:10.1088/1757-899X/1022/1/012072 9

[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. BMJopen, 4(5), e005025.

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

[8] Jabbar M A, Deekshatulu B L & Chandra P (2013, March). Heart disease prediction using lazy associative classification. In 2013 International MutliConference on Automation,

ISSN: 2057-5688

Computing,Communication, Control and Compressed Sensing (iMac4s) (pp. 40- 6). IEEE.

ISSN: 2057-5688

[9] 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.

[10] Folsom A R, Prineas R J, Kaye S A & Soler J T (1989). Body fat distribution and self-reported prevalence of hypertension, heart attack, and other heart disease in older women. International journal of epidemiologyy, 18(2), 361-7. 69

[11] Chen A H, Huang S Y, Hong P S, Cheng C H & Lin E J (2011, September). HDPS: Heart disease prediction system. In 2011 Computing in Cardiology (pp. 557- 60). IEEE.

[12] Parthiban, Latha and R Subramanian. "Intelligent heart disease prediction system using CANFIS and genetic algorithm." International Journal of Biological, Biomedical and Medical Sciences 3.3 (2008).

[13] Wolgast G, Ehrenborg C, Israelsson A, Helander J, Johansson E &Manefjord H (2016). Wireless body area network for heart attack detection [Education Corner]. IEEE antennas and propagation magazine, 58(5), 84-92.

[14] PatelS&ChauhanY(2014).Heartattackdetectionandmedicalattentionusingmotionsensingdevice - kinect. International Journal of Scientific and Research Publications, 4(1), 1-4. [15] Piller L B, Davis B R, Cutler J A, Cushman W C, Wright J T, Williamson J D & Haywood L J (2002). Validation of heart failure events in the Antihypertensive and Lipid Lowering Treatment to Prevent Heart Attack Trial (ALLHAT) participants assigned to doxazosin and chlorthalidone. Current controlled trials in cardiovascular medicine

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

[17] A. Aldallal and A. A. A. Al-Moosa, "Using Data Mining Techniques to Predict Diabetes and Heart Diseases", 2018 4th International Conference on Frontiers of Signal Processing (ICFSP), pp. 150-154, 2018, September.

[18] Takci H (2018). Improvement of

ISSN: 2057-5688

heart attack prediction by the feature selection methods. Turkish Journal of Electrical Engineering & Computer Sciences, 26(1), 1-10.

[19] AnkitaDewanandMeghnaSharma,"Predictionofheartdiseaseusingahybridtechniqueindatamining classification", 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)

[20] Aditya Methaila, Prince Kansal, Himanshu Arya and Pankaj Kumar, "Early heart disease prediction using data mining techniques", Computer Science & Information Technology Journal, pp. 53-59, 2014.