IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

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

1Divyadarshini S, 2Mrs. Maria Sylviaa S

1Student, 2Assistant Professor

1Nirmala College for Women,

2Nirmala College for Women

Abstract

The healthcare field has vast potential for the application of artificial intelligence to improve the quality of services offered. The objective of this work is to predict whether a person has heart disease or not. Health care field has a vast amount of data, for processing those data certain techniques are used. Heart disease is the Leading cause of death worldwide. This System predicts the arising possibilities of Heart Disease. Supervised machine learning techniques are applied for the prediction task. The dataset used here has 303 entries and 14 parameters. The dataset is preprocessed before used to build the model, Logistics Regression, Naïve Bayes, Support Vector Machine (SVM), K Nearest Neighbor (KNN), Decision Tree (DT), XGBoost and Neural Networks are used to predict the accuracy. This experiment results with the accuracy of 85.25%, 85.25%, 81.97%, 67.21%, 81.97%, 85.25% and 83.61% respectively for Naïve Bayes, LR, KNN, DT, XGBoost and Neural Network.

Keywords: Decision Tree, Heart disease, KNN, Logistic Regression, Machine Learning, Naïve Bayes, Neural Network, SVM, XGBoost

I INTRODUCTION

The World Health Organization estimates that heart disease causes 12 million deaths worldwide each year. One of the leading causes of morbidity and mortality among the global population is heart disease. One of the most crucial topics in the data analysis area is predicted cardiovascular disease. Since a few years ago, the prevalence of cardiovascular disease has been rising quickly throughout the world. Many studies have been

carried out in an effort to identify the most important risk factors for heart disease and to precisely estimate the overall risk. Heart disease is also referred to as a silent killer because it causes a person to pass away without any evident signs. With high-risk individuals, an early diagnosis of heart disease is crucial for helping them decide whether to change their lifestyle, which lowers consequences. Making choices and predictions from the vast amounts of data generated by the healthcare sector is made easier with the help of machine learning. By evaluating patient data that uses a machine-learning algorithm to categorize whether a patient has heart disease or not, this study hopes to predict future cases of heart disease. Machine learning methods can be extremely helpful in this situation. There is a common set of basic risk factors that determine whether or not someone will ultimately be at risk for heart disease, despite the fact that heart disease can manifest itself in various ways. We may say that this technique can be very well fitted to accomplish the prediction of heart disease by gathering the data from many sources, classifying them under acceptable categories, and then analyzing to obtain the needed data.

1.1 MOTIVATION FOR THE WORK

The primary reason for conducting this study is to propose a model for predicting the development of heart disease. Also, the goal of this research is to determine the optimum classification method for detecting cardiac disease in a patient. Three different classification procedures, namely Nave Bayes, Decision Tree, and Random Forests, are employed at various levels of evaluations in a comparative investigation and evaluation to support this work. Although these machine learning methods are widely utilised, predicting cardiovascular problems is a crucial task requiring the highest level of accuracy. Thus, a variety of levels and assessment strategy types are used to examine the three algorithms. This will enable scientists and medical professionals to create a better.

1.2 PROBLEM STATEMENT

The main difficulty with heart disease is detecting it. There are tools that can forecast heart disease, but they are either expensive or ineffective at calculating the likelihood of heart disease in a human. The mortality rate and total consequences can be reduced by early identification of heart disorders. Since it takes more intelligence, time, and knowledge, it is not always possible to accurately monitor the patient daily, and a physician cannot discuss with a sufferer for a whole 24 hours. As there is a lot of data available nowadays, we can use a variety of machine learning methods to search for hidden patterns. With medical data, the hidden patterns might be used for health diagnosis.

II LITERATURE SURVEY

Recent studies and research in medical science and machine learning have produced significant papers.

Purushottam et al.[1] proposed an "Efficient Heart Disease Prediction System" using hill climbing and decision tree algorithms. Before classifying data, they preprocessed the Cleveland dataset. Evolutionary Learning (KEEL), an open-source data mining tool, fills missing values for Knowledge Extraction. Top-down decision trees. Each level tests a hill-climbing algorithm-selected node. Confidence parameters and values. Minimum confidence is 0.25. System accuracy is 86.7%.

Santhana Krishnan et al. [2] used decision tree and Naive Bayes algorithms to predict heart disease. Decision tree algorithms use conditions to make True or False decisions. SVM and KNN use vertical or horizontal split conditions depending on dependent variables. But decision trees for a tree-like structure with root node, leaves, and branches based on tree node decisions Decision trees also help emphasize dataset attributes. Cleveland data was used. Methods divide the dataset into 70% training and 30% testing. 91% accuracy. Classification algorithm Naive Bayes follows. It can handle complicated, nonlinear, dependent data, making it suitable for heart disease datasets, which are also complex. 87% accuracy.

Sonam Nikhar et al [3] paper .'s "Prediction of Heart Disease Using Machine Learning Algorithms" explains Naïve Bayes and decision tree classifiers used to predict heart disease. 3 Decision Tree outperforms Bayesian classifier in perceptive data analysis on the same dataset.

In "Prediction of Heart Disease Using Machine Learning," Aditi Gavhane et al.[4] used a multi-layer feed - forward neural neural network algorithm to prepare and test datasets. This methodology has one input and one output layer and one or more hidden layers between them. Hidden layers connect input nodes to output nodes. This linkage has spontaneous weights. The bias input is weighted according to need, and the nodes can be convolutional or feedback.

Avinash Golande et al. [5] proposed "Heart Disease Prediction Using Effective Machine Learning Techniques" using a few data gathering techniques to help doctors distinguish cardiovascular disease. k-nearest neighbor, decision tree, and Naïve Bayes are used. Packing calculation, part thickness, consecutive negligible standardizing, neural systems, straight Kernel self arranging guide, and SVM are other characterization-based methods (Bolster Vector Machine).

Lakshmana Rao et al. [6] suggested "Machine Learning Techniques for Heart Disease Prediction" with more major contributors. Thus, diagnosing heart disease is difficult. Data gathering and neural nets are used to assess heart disease severity.

Abhay Kishore et al. [7] suggested "Heart Attack Prediction Using Deep Learning" to predict heart-related infectious diseases using convolutional and recurrent neural systems. Deep learning and data mining produce the most accurate and error-free model. This study is a strong heart attack prediction model reference.

Senthil Kumar Mohan et al. [8] proposed "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques" to enhance cardiovascular problem accuracy. The prediction model for heart disease with hybrid random forest with linear model uses KNN, LR, SVM, and NN algorithms to improve exhibition with 88.7% precision (HRFLM).

Anjan N. Repaka et al. [9] proposed a model that compares the prediction performance of two classification models to previous work. Our proposed method predicts risk percentages more accurately than other models, according to experiments.

"Heart Disease Prediction using Evolutionary Rule Learning" by [10] Aakash Chauhan et al. Electronic records reduce manual data retrieval. Services are reduced and a large number of rules predict heart disease best. Frequent pattern growth association mining on patient data generates strong associations.

Existing System:

Tools and methods are regularly tested to meet current health needs. Machine learning can help. Heart disease can take many forms, but there are core risk factors that determine whether someone will develop it. We can draw conclusions by gathering data from various sources, classifying it under appropriate headings, and then analysing it. This method can predict heart disease well.

Disadvantages

1. The system is mostly manual, requiring experienced doctors to predict.
2. Manual processes take time.

Method

The system works by collecting data and selecting key attributes. Then data is preprocessed. Training and testing data are separated. Algorithms and training data train the model. Test data determines system accuracy.

Advantages

1. Automatic kidney disease prediction.
2. Computer-assisted methods are faster.
3. Its accuracy helps patients receive timely medical care to cure and save their lives.

III METHODOLOGY

These modules are implemented in this system.

1.) Collection of Dataset

2.) Attributes selection

3.) Data Pre-Processing

4.) Balancing of Data

5.) Disease Prediction

3.2.1 Collection of dataset

Our heart disease prediction system starts with a dataset. We divided the dataset into training and testing data. Prediction model learning and evaluation use training and testing datasets. This project uses 70% training and 30% testing data. This project used Heart Disease UCI. 14 of the dataset's 76 attributes are system-relevant.

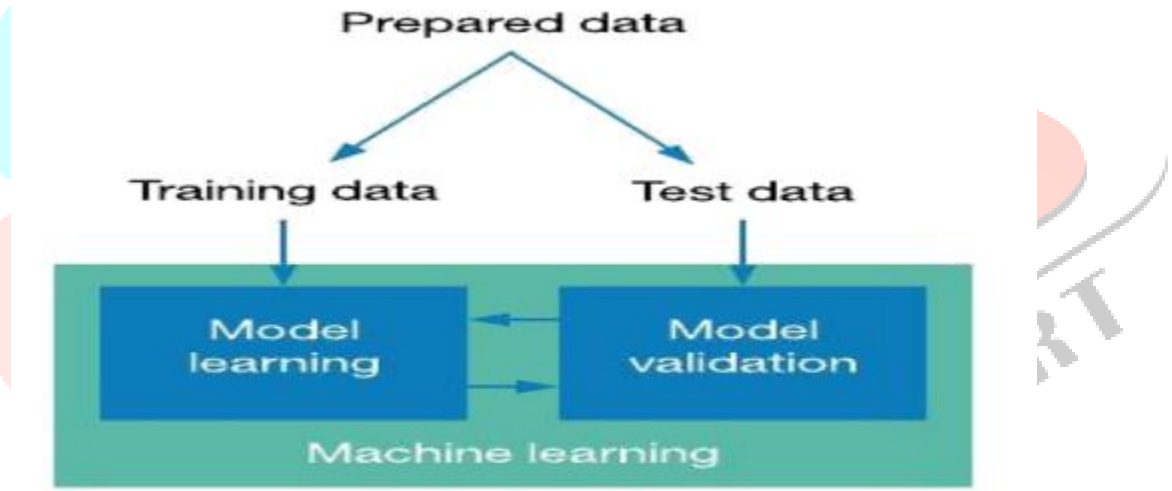

Figure: Collection of Data

3.2.2 Attributes selection

Attribute or feature selection involves choosing appropriate attributes for the prediction system. For system efficiency. Gender, chest pain type, fasting blood pressure, serum cholesterol, exang, etc. are used to predict. This model selects attributes using the correlation matrix.

Figure: Correlation matrix

3.2.3 Pre-processing of Data

Data pre-processing is crucial to machine learning model creation. Initial data may not be clean or in the model's format, resulting in misleading results. Data pre-processing formats data. Handles dataset noise, duplicates, and missing values. Data pre-processing includes imports, splits, attribute scaling, etc. Data preprocessing improves model accuracy.

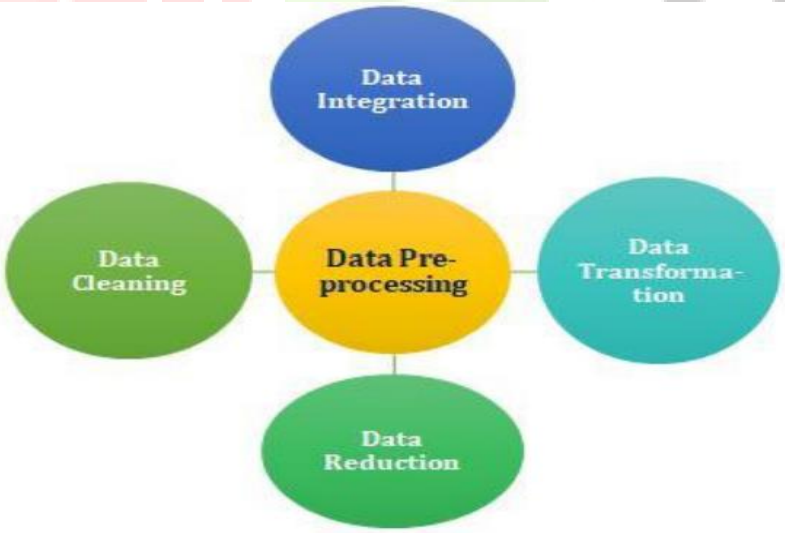

Figure: Data Pre-processing

3.2.4 Data balancing

Two methods balance unbalanced datasets. Under- and over-sampling Sampled: Under Sampling balances datasets by shrinking the ample class. Data is sufficient for this process. (b) Over Sampling: Over Sampling balances datasets by expanding scarce samples. When data is scarce, this is considered.

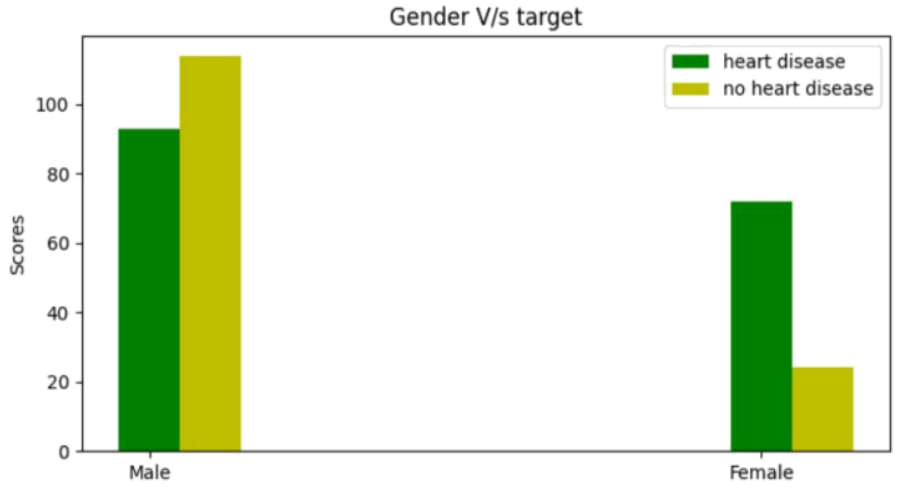

Figure: Data Balancing

3.2.5 Prediction of Disease

Classification uses SVM, KNN, Naive Bayes, Decision Tree, Random Tree, Logistic Regression, Neural network, and Xg-boost. The algorithm with the highest accuracy predicts heart disease.

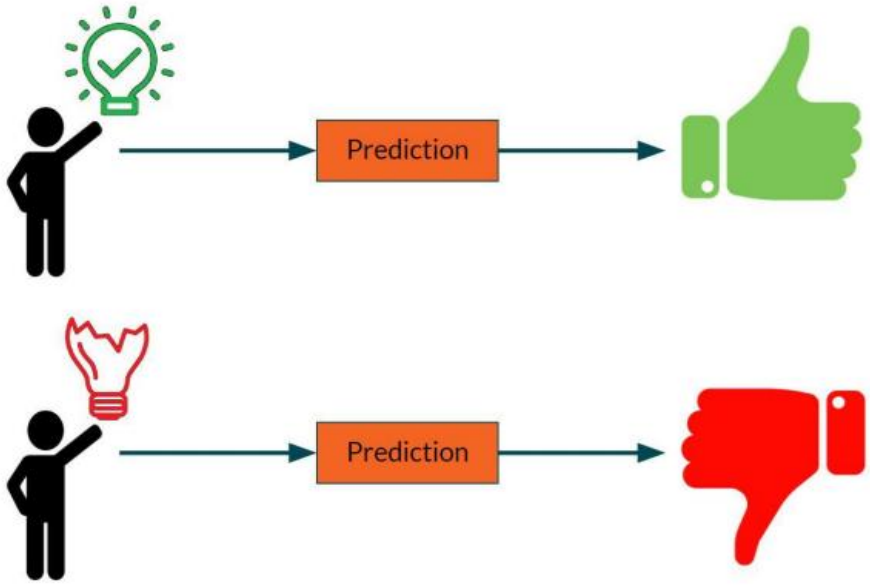

Figure: Prediction of Disease

IV MACHINE LEARNING

Classification in machine learning predicts a class label for a given input data example. Assisted Study Supervised learning uses well-labeled training data to predict output. Labelled data has the correct output already attached. In supervised learning, training data helps machines predict output correctly. It uses the same concept as a teacher-supervised student. Supervised learning feeds the machine learning model input and output data. A supervised learning algorithm finds a mapping function to map x(input variable) to y(output variable) .

Self-taught Unlike supervised learning, unsupervised learning only has input data and no output data. Unsupervised learning finds a dataset's structure, groups similar data, and compresses it. • Unsupervised learning provides data insights. • Unsupervised learning is closer to real AI because it mimics how humans learn to think through experience. • Unsupervised learning is important because it uses unlabeled and uncategorized data. • Unsupervised learning is needed when input data does not match output data in real life.

Reinforcement Machine learning includes reinforcement learning. It's about taking appropriate action to maximise reward. It helps software and machines choose the best action in a given situation. In supervised learning, the training data includes the answer key, so the model is trained with the correct answer. In reinforcement learning, there is no answer, but the reinforcement agent decides how to complete the task. It learns from experience without a training dataset.

4.1 ALGORITHMS

4.1.1 SUPPORT VECTOR MACHINE (SVM):

SVM is a popular Supervised Learning algorithm for classification and regression. It is mostly used for Machine Learning classification problems. The SVM algorithm creates the best line or decision boundary to divide n-dimensional space into classes so we can easily place new data points in the correct category in the future. Hyperplanes are best decision boundaries. SVM selects hyperplane-creating extreme points/vectors. Support vectors are extreme cases, so the algorithm is called SVM. SVMs are versatile supervised machine learning algorithms used for classification and regression. They're mostly used for classification. In 1990, SVMs were refined. SVMs implement differently than other machine learning algorithms. They're popular now because they can handle multiple continuous and categorical variables.

SVM's key concepts are:

- Support vectors Support vectors are hyperplane-close data points.
- These data points will define separating line.
- Hyperplane - The above diagram shows a decision plane or space divided between objects of different classes.
- Margin—The gap between two lines on the closest data points of different classes. The line-to-support vector perpendicular distance can be calculated.
- Small margins are bad and large margins are good.

Types of SVM:

Linear SVM: Linear SVM classifiers are used for data that can be classified into two classes using a straight line.

Non-linear SVM: Non-linear SVM classifiers are used for datasets that cannot be classified using a straight line. The support vector machine algorithm finds a hyperplane in an N-dimensional space (N - the number of features) that classifies data points.

Support vector machines are memory-efficient, effective in high-dimensional spaces, and still effective when the number of dimensions is greater than the number of samples.

Versatile: decision function kernels can be specified. Custom kernels are available. Support vector machines have drawbacks: Avoid overfitting when choosing Kernel functions and regularisation terms if the number of features is much greater than the number of samples. SVMs use five-fold cross-validation to estimate probability.

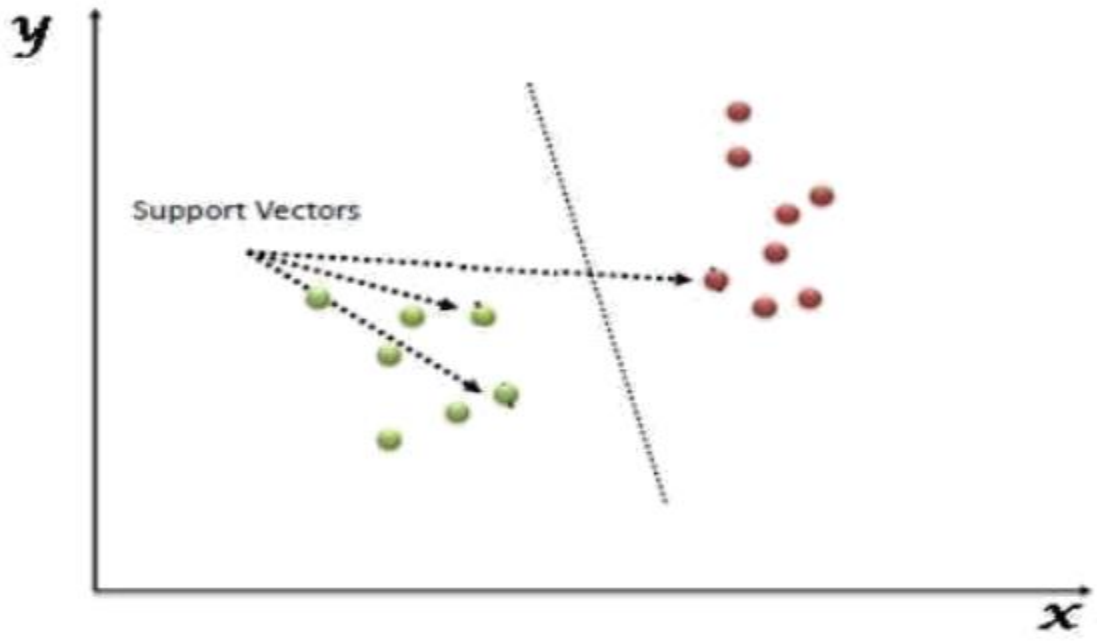

Figure: Support Vector Machine

4.1.2 **NAIVE BAYES ALGORITHM:**

Bayes theorem-based supervised learning algorithm Naive Bayes solves classification problems. Text classification with a high-dimensional training dataset uses it most. Naive Bayes Classifier is one of the simplest and most effective classification algorithms for building fast models using machine learning that can make assumptions. It predicts based on object probability as a probabilistic classifier. Spam filtration, sentiment analysis, and article classification use Naïve Bayes Algorithm. Bayes' Theorem-based classification assumes estimator independence. A Naive Bayes classifier assumes that a class's features are unrelated. Naive Bayes is simple to build as well as helpful for large data sets. Naive Bayes outperforms even complex classification methods. Naive Bayes algorithm:

Naive: It presumes that one feature is independent of others. Apples are red, round, and sweet. Each characteristic independently identifies it as an apple.

It's called Bayes because it uses Bayes' Theorem. Bayes: Bayes' theorem, also known as Bayes' Rule or Bayes' law, determines the probability of a hypothesis with prior knowledge. Conditional distribution determines.

The formula for Bayes' theorem is given as:

$$P(A|B)=\frac{P(B|A)P(A)}{P(B)}$$

Posterior probability: hypothesis A on observed event B. Likelihood probability (P(B|A)) is the evidence that a hypothesis is true.

- Prior Probability: Hypothesis before evidence.
- Marginal Probability: Evidence Probability is P(B).

Naive Bayes model types:

Below are three Naive Bayes Models:

The Gaussian model assumes features have a normal distribution. The model assumes Gaussian sampling for continuous predictors.

Multinomial: The Multinomial Naïve Bayes classifier is used for multinomial data. Document classification, such as sports, politics, education, etc., is its main use. Word frequency predicts the classifier.

• Bernoulli: Like the Multinomial classifier, but the predictor variables are independent Booleans variables. Whether a document contains a word. This model is popular for document classification.

4.1.3 **DECISION TREE ALGORITHM**

The Decision Tree is a supervised learning method that can solve classification and regression problems, but it's best for classification. A tree-structured classifier, nodes in the network represent dataset characteristics, branches depict decision rules, and leaf nodes represent outcomes. Decision Trees have two nodes: Decision and Leaf. Decision nodes make decisions and have multiple branches, while leaf nodes output those decisions and have no branches. Decisions and tests are based on dataset features. It shows all potential solutions to a problem/decision according to specified situations. It's termed a Decision Tree because it begins with a root of the tree and branch offices out like a tree. The Classification and Regression Tree algorithm (CART) builds trees. Based on the answer (Yes/No), a Decision Tree splits the tree into subtrees. The Decision Tree Algorithm is supervised machine learning. It can classify and regress. This algorithm uses a decision tree to predict the value of a target variable. The leaf node represents a class label, and the internal nodes represent attributes. When creating a machine learning model, remember to choose the best algorithm for the dataset and problem.

The Decision Tree has two benefits:

- Because of their tree-like structure, decision trees are easy to understand.
- Understanding the logic is easy because it has tree structure.
- Identifying each level's root node attribute in Decision Tree is difficult in attribute selection.

Two attribute selection measures are popular:

1. Information Gain: Decision Tree nodes partition training instances into smaller subsets, changing entropy. Information gain measures entropy change. Entropy is a random variable's uncertainty and an arbitrary collection's impurity. Entropy increases information content.

2. Gini Index: Measures how often a randomly chosen element is misidentified. A lower Gini index attribute is preferred. Sklearn defaults to "gini" for Gini Index.

Decision Tree algorithms include:

1. IDichotomiser 3 (ID3): Information Gain determines which attribute is used to classify the current subset of data. Recursively calculate data information gain for each tree level.

2. C4.5: ID3's successor. This algorithm classifies using information gain or gain ratio. It handles continuous and missing attribute values, improving on the ID3 algorithm.

3. Classification and Regression Tree (CART): This dynamic learning algorithm produces regression and classification trees depending on the dependent variable.

Working: Decision Trees predict dataset classes by starting at the root node. This algorithm compares the root attribute values to the record (real dataset) attribute, follows the branch, and jumps to the next node. The algorithm compares the attribute value of the next node to the other sub-nodes and continues. It continues until the tree's leaf node.

Algorithm:

Step 1: Start the tree with the root node, S, which has the entire dataset.

Step-2: Use Attribute Selection Measure to find the best dataset attribute (ASM).

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

Step 4: Create the best attribute Decision Tree node.

Step-5: Recursively create new decision trees from step-3 dataset subsets.

Continue until you can no longer classify the nodes and call the final node a leaf node.

4.1.4 **RANDOM FOREST ALGORITHM**

Random Forest is supervised. Machine learning classifiers with bagging improve Decision Tree performance. It combines tree predictors and depends on an independently sampled random vector. Tree distribution is uniform. Instead of splitting nodes by variables, Random Forests splits them by the best predictor subset randomly chosen from the node. In the worst case, learning with Random Forests takes O(M(dnlogn)), where M is the number of growing trees, n is the number of instances, and d is the data dimension. It can classify and regress. The most flexible and user-friendly algorithm. Forests have trees. Forests are stronger with more trees. Random Forests create Decision Trees on randomly selected data samples, make predictions, and vote on the best solution.

It also indicates feature importance well. Recommendation engines, image classification, and feature selection use Random Forests. It can identify loyal loan applicants, fraud, and diseases. It underpins the Boruta algorithm, which selects key dataset features. Supervised learning algorithm Random Forest is popular. It solves ML classification and regression problems. Ensemble learning uses multiple classifiers to solve complex problems and improve model performance. "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset," as the name suggests. The random forest predicts the final output by taking the prediction from each decision tree and calculating the majority vote. More trees in the forest improve accuracy and prevent overfitting.

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, two assumptions for a better Random forest classifier:

- The dataset's feature variable should have actual values so the classifier can predict accurate results rather than guesses.
- Tree predictions must be uncorrelated.

Algorithm Steps:

It works in four steps:

- Randomize a dataset.
- Each sample's Decision Tree should predict.
- Vote for each prediction.
- The most-voted prediction is the final prediction.

Advantages:

- Random Forest can handle large datasets with high dimensionality
- improve model accuracy, and
- prevent overfitting.

Disadvantages:

- Random Forest is not better for regression than classification.

4.1.5 **LOGISTIC REGRESSION ALGORITHM**

Logistic regression, a Supervised Ml algorithm, is famous. It predicts classification dependent variables using different factors. Logistic regression predicts categorical dependent variable output. Thus, the result has to be discrete. Instead of 0 and 1, it helps give predictive values between 0 as well as 1. Logistic Regression has similarities to Linear Regression but used differently. Linear regression solves regression problems, while logistic regression classifies them. Logistic regression forecasts two maximum values using a "S"-shaped logistic function instead of a regression line (0 or 1). The logistic function curve predicts whether cells are cancerous, mice are obese according to their weight, etc. Logistic Regression classifies and provides probabilities for continuous and discrete datasets.

Advantages:

Logistic Regression, which is among the easiest machine learning algorithms, can improve training efficiency in some cases. This algorithm's model training doesn't require much computational resources either. The trained weights indicate feature importance. Positive or negative association is also given. Logistic Regression can determine feature relationships. Unlike Decision Tree or Support Vector Machine, this algorithm can easily update models with new data. Stochastic gradient descent updates. Logistic Regression provides classification and well-calibrated probabilities. This is better than designs that only classify. We can infer which training examples are more accurate for the formulated problem if one has a 95% probability for something like a class and another has a 55% probability.

Disadvantages:

Logistic regression uses independent features to predict precise probabilistic outcomes. On high-dimensional datasets, this may cause the model to be over-fit on the training set, overstating its accuracy and preventing it from predicting accurate test set results. When the model is trained on little data with many features, this happens. To avoid overfitting on high-dimensional datasets, consider regularisation methods (but this makes the model complex). High regularisation factors may common complaint the training data. Logistic regression's linear decision surface precludes solving nonlinear problems. Real-world data is rarely linearly separable. Thus, increasing the amount of characteristics makes data linearly separable in higher dimensions, transforming nonlinear features.

Non-Linearly Separable Data:

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

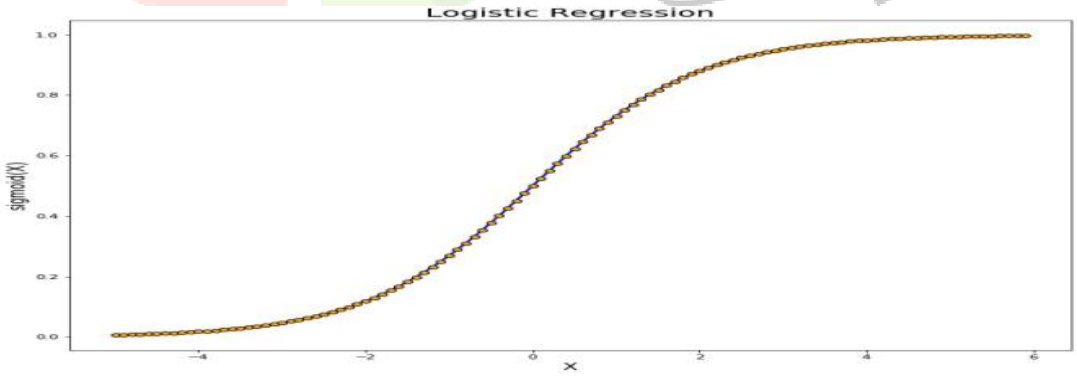

Figure: Logistic Regression

4.3.6 **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 randomly chooses a training 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. Each iteration, it weights the trained classifier based on accuracy. Accuracy is rewarded.

5. This process repeats until all training data fits correctly or the maximum number of estimators is reached.

6. "Vote" on all your learning algorithms to classify.

Advantages:

Adaboost is easier to use than SVM algorithms and requires less parameter tweaking. Adaboost can be used with SVM, but theoretically, Adaboost applications don't overfit because the parameters aren't optimised jointly and the learning process is slowed by estimation stage-wise. This link helps math. Adaboost can also improve weak image/text classifiers and cases.

Disadvantages:

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

4.1.7 **XGBOOST ALGORITHM**

XG-boost implements gradient-boosted decision trees. It's a software library that speeds up models. This algorithm sequentially creates decision trees. Weights affect XG-boost. The decision tree predicts results by weighting all independent variables. Variables predicted wrong by the tree are weighted and fed to the second decision tree. These classifiers/predictors form a powerful and accurate model. It performs regression, classification, ranking, and user-defined prediction.

Regularization: L1 (Lasso Regression) and L2 (Ridge Regression) regularisation in XG-boost prevents overfitting. Thus, XG-boost is called GBM's regularised form (Gradient Boosting Machine). XG-boost receives two regularisation hyper-parameters (alpha and lambda) from Scikit Learn. L1 regularisation uses alpha and L2 regularisation uses lambda.

Parallel Processing: XG-boost uses parallel processing to outperform GBM. Multiple CPU cores run the model. Scikit Learn's nthread hyper-parameter parallelism processing. nthread specifies CPU cores. The algorithm will detect all cores if you don't specify nthread.

XG-boost can handle missing values. XG-boost tries both the left and right hand split when a node has a missing value and learns which one causes more loss. It then applies the same to testing data.

Cross Validation: XG-boost lets users run cross-validations at each boosting iteration, making it easy to find the optimal number of iterations in a single run. Unlike GBM, which requires a grid-search and tests only a few values.

Effective Tree Pruning: When a GBM splits a node and loses, it stops. It's greedy. XG-boost splits up to the max depth and then prunes the tree backwards to remove splits with no positive gain.

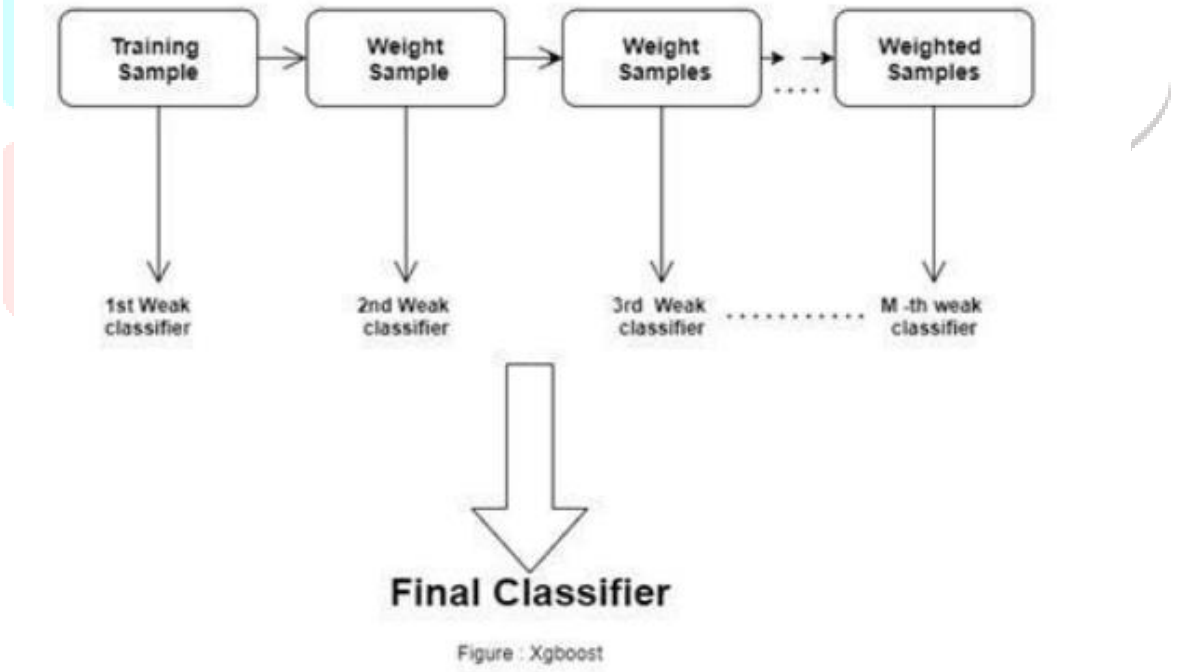

Figure : Xgboost

V EXPERIMENTAL ANALYSIS

5.1 SYSTEM CONFIGURATION

5.1.1 Hardware requirements:

Processor : Any Update Processor

Ram : Min 4GB

Hard Disk : Min 100GB

5.1.2 Software requirements:

Operating System : Windows family

Technology : Python3.7

IDE : Jupyter notebook

5.1.3 Implementation – Coding

VI RESULTS AND DISCUSSION

OUtputs

VII CONCLUSION AND FUTURE WORK

Machine learning can predict heart disease, a major killer in India and around the world. Early heart disease prognosis can help high-risk patients make lifestyle changes and reduce complications, a medical milestone. Heart disease rates are rising. This necessitates early diagnosis and treatment. Medical professionals and patients can benefit from appropriate technology support. This paper measures performance using SVM, Decision Tree, K nearest neighbours, Naïve Bayes, Logistic Regression, Neural netwrok, and Extreme Gradient Boosting on the dataset. The dataset contains 76 features that predict heart disease in patients, and 14 important features are chosen to evaluate the system. The author gets less system efficiency if all features are considered. Attribute selection boosts efficiency. Selecting n features for model evaluation improves accuracy. The dataset's almost-correlated features are removed. Efficiency drops when all dataset attributes are considered. The seven machine learning methods' accuracy is compared to generate a prediction model. Thus, confusion matrix, accuracy, precision, recall, and f1-score should be used to predict disease efficiently. Logistic regression, Extreme gradient boosting Naive bayes has 85.25% accuracy.

APPENDIX

Python

Python, created by Guido Van Rossum in 1991, is an interpreted, high-level, general-purpose programming language that emphasises code readability with its extensive use of white space. Its language constructs and object-oriented approach help programmers write logical code for small and large projects. Python is garbage-collected. It supports procedural, object-oriented, and functional programming.

Sklearn

Sklearn is Python's most powerful machine learning library. It offers efficient Python interfaces for classification, regression, clustering, and dimensionality reduction. NumPy, SciPy, and Matplotlib underpin this Python library.

Numpy

NumPy is a Python library that supports large, multi-dimensional arrays and matrices and a large set of high-level mathematical functions. Jim and others developed Numeric, the predecessor to NumPy. Travis modified Numeric to create NumPy in 2005. Open-source NumPy has many contributors.

Librosa

Python package Librosa analyses music and audio. Librosa is used to generate music using LSTMs and recognise speech. 67 It supplies music information retrieval system components. Librosa visualises audio signals and extracts features using signal processing.

Matplotlib

Matplotlib plots in Python and NumPy. It provides an object-oriented API for embedding plots into applications using Tkinter, wxPython, Qt, or GTK. A procedural "pylab" interface based on a statemachine (like OpenGL) that mimics MATLAB is discouraged.

Seaborn

Matplotlib-based Python data visualisation library Seaborn. It draws attractive and informative statistical graphics. Python library Seaborn creates statistical graphics. Seaborn is a Python data visualisation library based on matplotlib. Seaborn relies on visualisation to explore and comprehend data.

SciPy

SciPy includes optimization, linearalgebra, integration, interpolation, special functions, FFT, signal and image processing, ODE solvers, and other science and engineering tasks. SciPy, EuroSciPy, and SciPy.in are conferences for users and developers of these tools (in India). Enthought founded the US SciPy conference and sponsors many international conferences and hosts the website. Scientific computation library SciPy uses NumPy. It adds optimization, stats, and signal processing functions.

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.

[5] 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. BMJ open, 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, Computing,Communication, Control and Compressed Sensing (iMac4s) (pp. 40- 6). IEEE.

[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] Patel S & Chauhan Y (2014). Heart attack detection and medical attention using motion sensing device - 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 heart attack prediction by the feature selection methods. Turkish Journal of Electrical Engineering & Computer Sciences, 26(1), 1-10.

[19] Ankita Dewan and Meghna Sharma, "Prediction of heart disease using a hybrid technique in data mining 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.