PREDICTION OF CARDIOVASCULAR SYSTEM BY DECISION TREE

Sai Shruthi Cherukuri
Department of Information Science
University of North Texas
Texas, USA
saishruthicherukuri@my.unt.edu

Chetan Mylapilli
Department of Information Science
University of North Texas
Texas, USA
chetanmylapilli@my.unt.edu

Tejesh Bathula
Department of Information Science
University of North Texas
Texas, USA
tejeshbathula@my.unt.edu

Joshna Mereddy
Department of Information Science
University of North Texas
Texas, USA
joshnamereddy@my.unt.edu

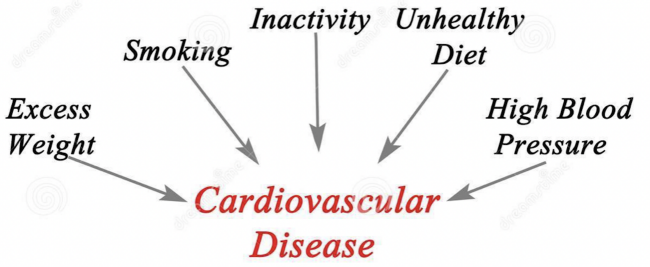
Leela Hari Priya Ginakunta
Department of Information Science
University of North Texas
Texas, USA
leelaharipriyaginakunta@my.unt.edu

***Abstract*—Healthcare is the utmost priority for humanity. According to WHO guidelines, good health is a fundamental right for individuals. Heart disease is a very serious issue and innumerable people suffer from this disease across the world. The heart disease rate is increasing rapidly in today's world due to several issues like a sedentary lifestyle, lack of proper sleep, lack of physical exercise, improper nutritional choices, smoking, and consumption of alcohol. The early detection and treatment of heart disease are challenging due to the lack of advanced equipment in most parts of the world like developing countries. There are many possible signs for heart disease patients that include shortness of breath, swollen feet, and physical body weakness. To avoid such delays in the appropriate treatment of heart disease, computer technology with Machine Learning techniques assists as a support system for early diagnosis of heart disease.**

The objective of our current research is to adopt the Machine Learning algorithm to diagnose heart disease and compare the results with other Machine Learning techniques. The research is an effort to perform an experiment on the Cleveland heart disease dataset. The Cleveland heart disease dataset is extracted from the Kaggle repository while comparing some of the most popular Machine Learning techniques to devise a Machine Learning algorithm that provides the most accurate predictions. We build a classification report and compare our model's Accuracy (ACC), Sensitivity (SEN), Recall, Precision (PRE), and F1 Score (F1) and finalize the best model. In the proposed research, the machine learning models performed are Logistic Regression, Random Forest, K-Neighbours Classifiers, Decision Tree, Naive Bayes, and XGBoost as the results of these classifiers are aggregated using weighted voting classifier which has achieved the accuracy of 87%. So, for predicting cardiovascular disease weighted voting classifier is the perfect classifier for detecting stroke.

I. Introduction

After the human brain, the heart is a vital component of the human body. Heart disease has many different signs and symptoms. These include chest pain, shortness of breath, pressure in the chest, angina, pain, and numbness in the arms and legs, irregular heartbeats, dry or chronic coughing, rashes or other strange patches on the skin, etc. The most common cause of death is cardiovascular disease. Heart disorders come in a variety of forms. They are arrhythmia, coronary artery disease, heart attack, chest pain, stroke, irregular heartbeat, and heart disease at birth [10]. High Blood pressure, cholesterol, and a fast heartbeat are the main contributors to coronary disease. The key contributors to an elevated risk of heart disease include improper lifestyle choices, a poor diet, insufficient exercise, a high BMI, and smoking. And your family's medical history may contribute to heart disease. [13] Use of alcohol and tobacco are individual risk factors for heart disease. Around the world, heart disease can afflict people of all ages.



The leading cause of death worldwide is heart disease. Every year, 18 million individuals worldwide [5] pass away as a result of cardiovascular disorders. WHO [8] estimates that 17.5 million individuals worldwide passed away from heart disease in 2005, accounting for [11] 31% of all fatalities. In addition, the mortality toll is rising steadily every year. By 2030 [3], it's projected to increase to more than 23.6 million.

The main aim of the research is to predict cardiovascular disease by implementing a classification model, [10] which serves as a primary classifier. A decision tree is a supervised machine-learning technique used in classification and regression modeling. The kind of target variables is another consideration while choosing an algorithm. The decision tree employs a number of algorithms, including ID3, C4.5, CART, and CHAID. After building the Decision tree classifier if our model is overfitted we have an optimization technique called Pruning. Pruning is a process of deleting unnecessary nodes from a tree in order to get the optimal decision tree. A too-large tree increases the risk of overfitting, and a small tree may not capture all the important features of the dataset. Therefore, a technique that decreases the size of the learning tree without reducing accuracy is known as Pruning. There are mainly two types of tree pruning technology used: cost complexity pruning and reduced error pruning. Pruning reduces the complexity of the final classifier, and hence improves predictive accuracy by the reduction of overfitting.

Moreover, in the research along with the Decision Trees other Machine Learning algorithms like Logistic Regression, Random Forest, K-Nearest Neighbors, and Naive Bayes are being compared for the prediction of heart disease and to calculate the performance of all the algorithms in this model.

Related works

Vijeta Sharma proposed [11] the main goal is to analyze various machine learning techniques and based on a person's unique characteristics and warning signs, determine whether a person can have heart disease or not. Researchers have used SVM, Random forest, and naive Bayes methods and calculated precision, recall, and ROC. The model developed with [11] SVM gives 98% accuracy which is 8% greater than the Naïve Bayes and approximately 13% greater than the Decision tree. In the same way, the model built with Random Forest gives the best prediction result with 99% accuracy, which is itself more accurate than our second-best SVM model for heart disease prediction. [9] Shamsher Bahadur Patel proposed the most interesting and challenging task is the pre-diction of heart disease via data mining. The lack of professionals and the high rate of cases with incorrect diagnoses have made the development of a quick and effective detection system necessary. Moreover, the decision tree has the highest accuracy of 99.2% when compared with Naïve Bayes and Classification clustering. Amita Malav [8] utilized advanced data mining techniques to provide an effective hybrid algorithmic technique for predicting heart disease. [8] Researchers utilized a hybrid Kmeans clustering algorithm and artificial neural network to predict heart disease. A hybrid model achieved a high accuracy rate of 97% when compared with Naïve Bayes and KNN.

Vedha Krishna Yarasuri has proposed a machine learning model which predicts the presence of severe diseases and the risk they cause to patients. This successful method can assist doctors in identifying patients with a higher risk of heart failure to ensure timely treatment. [14] In the research they have calculated the Accuracy score, Concordant-index value (C-index), F1 score, Recall score, and Precision score of 5 different machine learning algorithms such as logistic regression, support vector machine (SVM), k-nearest neighbor (KNN), decision trees and random forest. [14] random forest outperformed all the other four algorithms by having an average test accuracy of 90.29%, C-index value of 0.9402, precision of 0.9039, F1 of 0.9020, and recall score of 0.9039. [5] A. Lakshmanarao suggested a classifier model using feature selection and ensemble learning techniques that can predict heart disease. Researchers have applied sampling techniques such as Random Over sampling, Synthetic Minority Oversampling (SMOTE), and Adaptive synthetic sampling approach (ADASYN). random oversampling given the accuracy of 99% with stacking classifier, SMOTE sampling given an accuracy of 93%, and ADASYN sampling technique given an accuracy of 91% with stacking classifier. Mohd Zubir Suboh [12] has suggested a pc based system that classifies normal and abnormal heart sound signals from the heart sound audio files. Researchers used the Auscultation technique to detect heart problems to date. The segmentation process is done to get heart sound samples and moreover, feature extraction such as cross-correlation is done to find the difference between the samples. This system [12] calculates the specificity, sensitivity, and screening accuracy and resulted in 96.3%, 92.59%, and 94.44%.

Minhaz Uddin Emon has proposed prediction of heart disease at its early stages will be possible with the help of machine learning models. With the help of the variables, they [2] predicted whether the patient is going to effect by heart disease or not in the future. In their research [2] got the accuracy of 97%. Respectively In order to predict heart disease at its early stages by the physicians Rahul Katarya suggested the machine learning models which predicts with higher accuracy will be helpful for that. For prediction of heart disease what [2] did but he used the different dataset in the research. Also for predicting the heart diseases at early stages the [4] considered machine learning models like ANN, Random Forest, Support Vector Machine, Naïve Bayes and choose the best machine learning model . In their research author [4] suggested to follow the Support vector machine accuracy which predicted with 95% of accuracy. Also for the evaluation author [4] considered K-fold cross validation with 10 splits but the accuracy is for that is not as best as Support vector machine's accuracy. Respectively Author Damir Imamovic predicted the mortality rate of cardiovascular disease based of past events with the help of data mining techniques. For prediction of mortality rate author[3] performed 3 machine learning models Neural netwroks, Logistic regression and Decision Tree classifier. In their research [3] used rapid minor tool and for the evaluation to consider which method is predicting well they considered the accuracy in which neural network performed well with 74.34%. The main drawback of research is the accuracy scores of all the models are low.

In Classification models, the size of the training set [1] is one of the key factors that influence a classification model's accuracy. Numerous studies have shown large datasets are

required to train models and can be costly for high quality labels. Meriem Benhaddi presented online classification models – Very fast Decision tree and Extremely Fast Decision tree to increase the accuracy measure for small training dataset. The parameter "nmin" value of 100 is utilized to increase the accuracy of the model to 78% for small dataset. [13] proposed to forecast cardiac disease using the irregular forest method and to improve forecasting accuracy by incorporating machine learning algorithms and evaluating Decision tree performance against KNN. The accuracy will grow as the number of independent and dependent variables increases. The experiment is conducted repeatedly to acquire data on different scales of accuracy rate. In prediction of heart diseases, accuracy of Decision tree (86.7%) and KNN (82.5%). Decision Tree has a higher accuracy level compared to other models. [10] stated that Machine learning is widely used to forecast heart diseases, as predictive outcomes are far more accurate than any other technology. The main objective of authors is to present a smartphone application in predicting stages of heart diseases in advance with the use of machine learning models. Authors' research work related to cardiac disease medical problems, finally accuracy rate is higher than other research.Decision tree, Random Forest, and k-nearest neighbors algorithms used in the research. The most precise Machine learning model, Decision tree obtained 85% accuracy.

The goal to predict an accurate blood pressure value is proposed using the Gradient Boosting Decision Tree(GBDT) algorithm by Bing Zhang[15]. For the accuracy range of measurement errors, selected (5, +5) mmHg, which means that if the difference between the predicted blood pressure and the true blood pressure is larger than 5 was treated as an error. The mean error of the GBDT algorithm is compared with the Least Squares, RidgeRegression, elastic net, KNN and lasso. Based on the overall accuracy rate and average absolute error evaluation, results of GBDT algorithm yielded the best performance with the highest accuracy of 65% with the training time being less than 0.5s. Machine learning algorithms are capable of diagnosing cardiovascular disease, which lowers the death rate by reducing the error in factual outcomes and prediction. Abid Isaq proposed [6] machine learning algorithms which helps medical healthcare professionals to analyze data, to make correct and precise diagnostic conclusions .To prevent overfitting, the dataset splitted into 70:30 ratio for training and the test data. The complete collection of features in the heart failure clinical record dataset has been subjected to a comparative examination of Decision Tree, Adaptive boosting classifier, Logistic Regression. Experimental results demonstrate that the Decision Tree outperforms other models and achieved an accuracy value of 0.9262 with Synthetic Minority Out-Sampling Technique in prediction of patient's survival.Building a machine learning model to aid in the detection of heart disease could lead to early detection and potentially life-saving treatment. A Cleveland heart illness data set was examined by Yu Lin [7] and preprocessed the dataset. In the process of model training, algorithms(Logistic Regression, K-nearest Neighbors, Decision Tree) were applied, and the decision tree was determined to be the optimal model after hyperparameter tuning and cross-validation as it outperformed other models with superior scores of accuracies of 0.848, f1 score of 0.829, PRC-AUC of 0.909, and ROC-AUC of 0.917.

II. The methodology

1. Data Collection

The Cleveland Heart Disease Dataset is considered, which is available in [16] Kaggle repository. Dataset consists of 303 rows and 14 columns. The column "num" is the target variable out of the 14 variables. The description for each attribute is displayed below:

Attributes	Description
age	age in years 4
sex	sex (1=male ; 0= female)
cp	chest pain type - Value 0: typical angina Value 1: atypical angina Value 2: non-anginal pain Value 3: asymptomatic
trestbps	resting blood pressure
chol	serum cholestoral in mg/dl
fbs	(fasting blood sugar ¿ 120 mg/dl) (1 = true; 0 = false)
restecg	resting electrocardiographic results Value 0: normal Value 1: having ST-T wave abnormality Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
thalach	maximum heart rate achieved
exang	exercise induced angina (1 = yes; 0 = no)
oldpeak	ST depression induced by exercise relative to rest
slope	the slope of the peak exercise ST segment Value 0: upsloping Value 1: flat Value 2: downsloping
ca	number of major vessels (0-3) colored by flourosopy
thal	0 = normal 1 = fixed defect 2 = reversable defect and the label
num	0 = no disease, 1 = disease

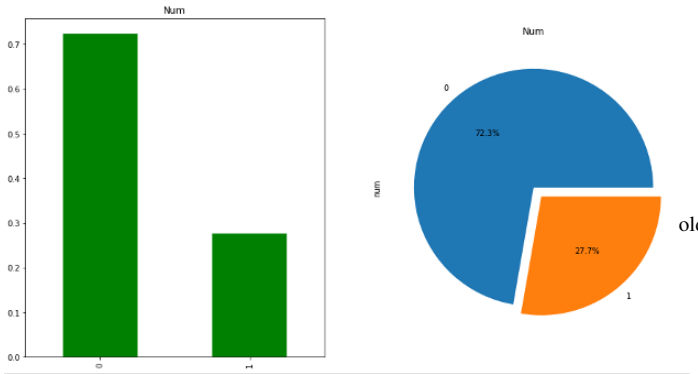
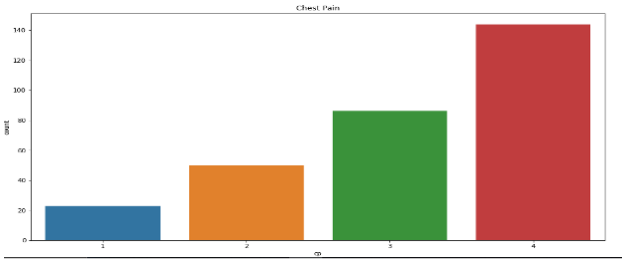
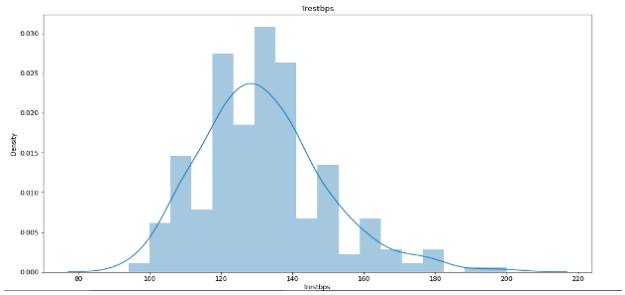


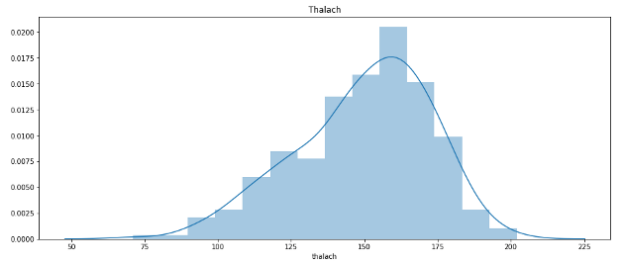
figure 1 we can infer that the angiographic disease value : 0 is about 72.3% and value : 1 is about 27.7%



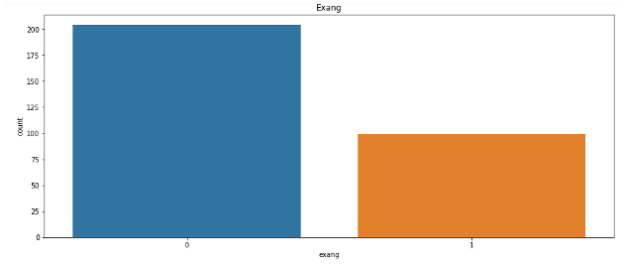
Chest pain: we can infer the persons who are effected with heart disease we can observe that most of the people are suffering with asymptomatic type of chest pain.



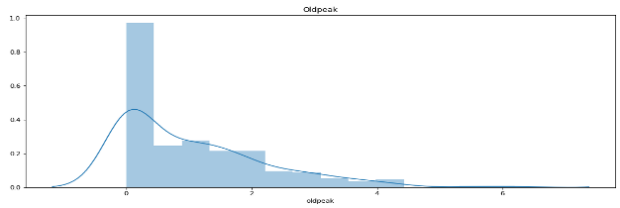
trestbps: resting blood pressure (in mm Hg on admission to the hospital)



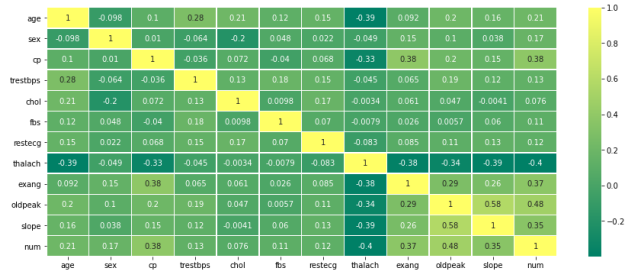
thalach: maximum heart rate achieved)



exang: exercise induced angina (1 = yes; 0 = no)



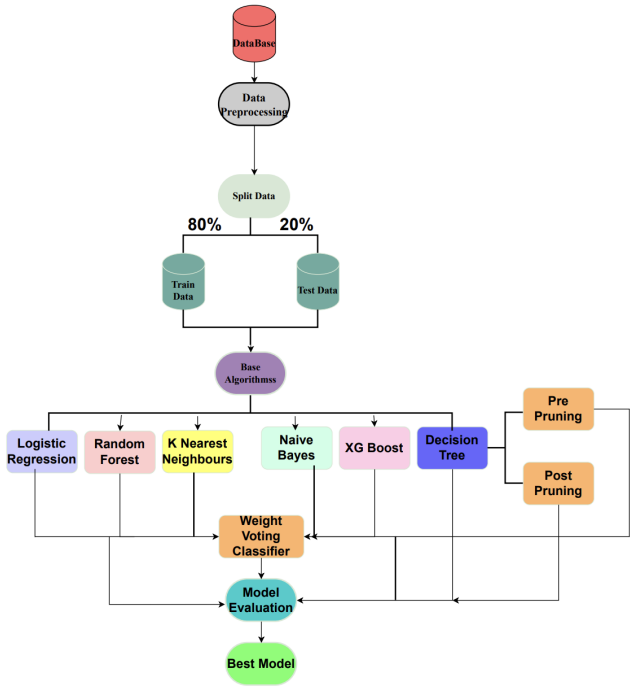
oldpeak: ST depression induced by exercise relative to rest)



heatmap infers how variables are correlated with each other.

Data Preprocessing:

Preparing raw data and to be appropriate for a machine learning model is termed as [3] data preprocessing. The process refers to a collection of techniques for improving the quality of the raw data, such as missing value imputation and outlier elimination. Real data are frequently erratic and incomplete. Dealing with missing values for some features is one of the major issues during data preprocessing. Eliminating samples with missing values, replacing them with mean or median values, or to manually enter them are a few solutions to this problem. The dataset has no specific null values and duplicate values. Instead the dataset consists of some special characters ("?") in "thal" and "ca" variables, which are removed as they are not affecting the data. Normalization[17] is the process to transform all columns to the same scale in the dataset to enhance the model's training stability and performance. MinMax scaling technique is implemented. The method subtracts the minimum value from the maximum value of each column and divided by range. Dataset is rescaled in each feature and their value lies between [0, 1]. Principal component analysis, or PCA, is a technique for reducing the number of dimensions in large data sets by condensing a large collection of variables into a smaller set that retains the majority of the large set's information.



Feature Selection:
Feature selection [10] is the significant step in using a machine learning model.A dataset can contain a wide variety of variables. However, some variables do not directly affect the outcome of the prediction. By eliminating those extra variables, we can improve the model's performance. All features from the Heart Disease Cleveland dataset are selected as every variable is significant to predict the target variable (num).

Splitting Dataset:
The fundamental goal of splitting a dataset[12] is to prevent the model from underfitting or overfitting. The most typical method of splitting a dataset is into two sections. Training Dataset - The machine learning model is trained using the training dataset. Using the training dataset, the machine learning model finds the data's hidden patterns. Testing Dataset - The machine learning model's primary objective is to predict the result of unobserved data. The data is splitted into 80% training set and 20% testing set.

Methods:

Logistic Regression :
Logistic regression is a well-known machine learning technique that belongs to the Supervised Learning method. It is utilized to forecast the categorical dependent variable from a collection of independent variables. A categorical dependent variable's outcome is predicted using logistic regression. As a result, the outcome should be discrete or a categorical value. It can either be 0 or 1, true or False, Yes or No, and so on, rather than giving the precise values as 0 and 1, it delivers the probabilistic values that fall between 0 and 1. It is a statistical technique for describing and explaining the correlation between one or more nominal, ordinary, dependent binary variables or ratio-level independent variables.

$$P = e^{ax+b}/1 + e^{ax+b}$$

Where P is the probability
e is the base of the natural logarithm
a and b are parameters of the model.

Random Forest :
The random forest is composed of several decision trees which tends as a group to predict the outcome. The dataset will be divided among the trees in this method. Every independent and unique tree divides the dataset into subgroups and makes predictions. This technique determines the conclusion according to the predictions of decision trees. It anticipates by taking the average of the output from multiple trees. The precision of the results improves as the amount of trees grow. It can eliminate draw backs in a decision tree algorithm and can bring down the dataset overfitting and enhances the accuracy.

K-Nearest Neighbour:
K-nearest neighbor usually denoted as KNN, is a distance-based technique for determining the distance between numerous data points in feature space. It is primarily focused on determining the most appropriate k value to use. The dataset is composed of arithmetic points representing every feature space. This technique could fit the dataset by selecting the appropriate value of K. To achieve reliable results, we must identify appropriate K values. In KNN, the distance is calculated using two methods, those are Euclidean distance and Manhattan distance. It is extremely simple to configure and is commonly utilized as a standard for much more sophisticated classifiers like Support Vector Machines (SVM) and Artificial Neural Networks (ANN). Regardless of its structure, k-NN outdoes more effective classifiers in a wide range of applications, along with forecasting, data reduction, and genetic factors.

Euclidean Distance

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(qi - pi)^2}$$

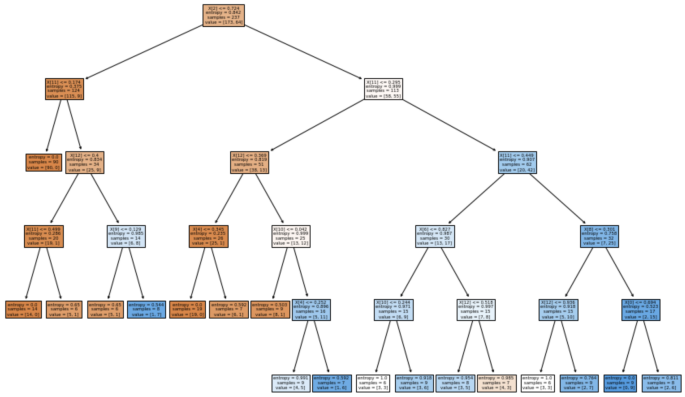
Where p,q = two points in Euclidean n-space
qi,pi = Euclidean vectors, starting from the origin of the space
N = n-space

Naive Bayes:
It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. The Naive Bayes Classifier is utilized when the input data is highly dimensional. In computer vision applications, the Naive Bayes method is extremely useful. This is based on the Bayes theorem's principle of conditional probability.

$$P(c|x) = P(x|c)P(c)/P(x)$$

Where P(c,x) = Posterior probability
P(x,c) = Likelihood
P(c) = Class prior probability
P(x) = Predictor prior probability

Decision Trees:
A decision tree is a tree-structured algorithm. In a tree-based algorithm each internal node represents a condition on a single dataset's characteristics. Every branch in the tree represents the result of the experiment performed on every branch. A class label is attached to every single leaf node. It constructs a binary tree on the basis of the input and output features by splitting syllables at every node in accordance with the function mentioned for every input characteristic.



XGBoost:
One decision tree is constructed at a time by XGBoost in order to account for all relevant data and fill in any gaps that may exist. This makes it easier for programmers to combine the decision tree algorithm and gradient algorithms for better outcomes. The gradient of the outcomes is taken into consideration while developers construct the decision trees since they compute and add the results for the following tree. Even though decision trees take time, this aids developers in getting a sense of the outcomes. The calculation is quicker and the precision is more precise than with Random Forest because the gradient of the data is taken into account for each tree. Because of this, developers now rely more on XGBoost than Random Forest. In comparison to other decision tree algorithms, XGBoost is the most sophisticated. The algorithm's total number of leaves is not taken into account by XGBoost. The algorithm performs better with more leaves in the decision tree if the model predictability is poor. As a result, the bias is reduced and the outcomes are entirely dependent on the data used in the algorithm.

Weight Voting Classifier:
A Voting classifier is a machine learning estimator that trains various base models and predicts on the basis of aggregating the findings of each base estimator. In the following research, aggregated Logistic Regression, Random Forest, K Nearest Neighbours, Naive Bayes, and Decision Tress and aggregated in the final result.

III. The Results

The results section will discuss the results of the test data. Results section mainly concentrate on the metrics like Accuracy, F1- Score, Precision, Recall, Roc_Auc Score. Among the data of 303 instances 60 instances has

been used for testing purposes.

Evaluation Metrics:
There are several techniques for assessing model performance. The ratio of cases that are correctly categorized is accuracy. The definition of accuracy is

$$Accuracy = TP + TN/TP + FP + FN + TN$$

However, accuracy won't give a decent indication of model performance if the class distribution is unbalanced. Sensitivity, Specificity, Precision, and F-measure are regularly used evaluation metrics that are appropriate in this situation. Recall, another name for sensitivity, calculates the proportion of real positives that are accurately identified.

$$Sensitivity = TP/TP + FN$$

The ratio of the actual negatives that are accurately identified is known as specificity.

$$Specificity = TN/TN + FP$$

The proportion of true positives to predicted positives is measured by precision, often known as the positive predictive value (PPV).

$$Precision = TP/TP + FP$$

The harmonic mean of recall and precision is known as F-Measure.

$$F - measure = 2 * Precision * Recall/Recall + Precision$$

The results of our experiment indicate that the voting classifier which is a combination of DecisionTreeClassifier, GaussianNB, LogisticRegression, RandomForestClassifier, KNeighborsClassifier outperformed Logistic Regression, Random Forest, KNN Gaussian Naive Bayes in terms of performance. Voting Classifier model has an accuracy of 87%, which is 3-5% better than Naive Bayes, KNN, Random Forest and roughly 6-12% better than Decision tree. Decision tree base model has produced results with an accuracy of 73%. For better understanding pre pruning and post pruning was performed on the Decision tree which generated an accuracy of 75% for pre pruning and 72% for post pruning. Unfortunately, we were unable to identify a decision tree that worked with our data.

	Model	Test_Accuracy	Train Accuracy	Roc_Auc_Score	F1 Score	Precision Score	Recall Score
0	Logistic Regression	83	87	88	74	74	74
1	Random Forest	77	100	90	59	67	53
2	Knn	80	89	83	68	68	68
3	Naive Bayes	82	86	82	74	67	84
4	XGB	77	100	87	59	67	53
5	Decision Tree Base Model	73	100	69	58	58	58
6	Decision Tree pre prune 1	73	92	72	60	57	63
7	Decision Tree pre prune 2	77	82	79	59	67	53
8	Decision Tree post prune	75	83	71	59	61	58
9	Weight Voting Classifier	87	90		75	71	79

The above table represents the final results of the research.

Table-1: Comparision study of references

Authors	Techniques used by authors	Accuracy
Vedha Krishna Yarasuri[14]	Logistic regression	90.29%
A. Lakshmanarao[5]	SMOTE, ADASYN	93%, 91%
Tamara Islam Meghla[2]	SGD	65%
Amita Malav[8]	Naïve Bayes	88%
Radhika Baskar[13]	KNN	82.5%
Elmir Babovic [3]	Neural Networks,	74%

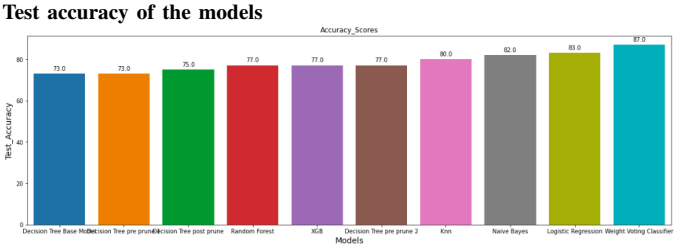
Table-1 represents the results of Authors from references.

Table-2: Results of the paper:

Models	Accuracy
Logistic Regression	83%
Random Forest	77%
KNN	80%
Naive Bayes	82%
XGB	77%
Decision Tree Base Model	73%
Decision Tree pre prune 1	73%
Decision Tree pre prune 2	77%
Decision Tree post prune	75%
Weight Voting Classifier	87%

Table-2 represents the results of the research.

Test accuracy of the models



IV. Conclusion

In the modern, expanding world, heart disease is a very serious problem. Consequently, a system that can anticipate cardiac disease at an early stage is needed. By using this automated method, people may keep track of their health difficulties, which will be helpful for both the doctor and the patient in terms of accurate diagnosis. Research work has been employed with the machine learning classifiers to find out the stroke accurately in a person. Weight voting classifier has considered features like age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal which helped to get the highest accuracy of about 87% when compared to all the accuracies used in research. As a result weight voting classifier can be considered for the prediction of cardiovascular disease in a person. As well this is also helpful in predicting heart disease at early stages. As a result if the disease is diagnosed it will be helpful for the doctor and the person affected with the disease to take precautions and cure the disease before as soon as possible. Therefore, in the future, it is preferable to voting classifiers to predict outcomes will provide us better outcomes for heart disease prediction.

References

[1] Mariam Benllarch, Salah El Hadaj, and Meriem Benhaddi. Improve extremely fast decision tree performance through training dataset size for early prediction of heart diseases. In *2019 International Conference on Systems of Collaboration Big Data, Internet of Things & Security (SysCoBIoTS)*, pages 1–5. IEEE, 2019.

[2] Minhaz Uddin Emon, Maria Sultana Keya, Tamara Islam Meghla, Md Mahfujur Rahman, M Shamim Al Mamun, and M Shamim Kaiser. Performance analysis of machine learning approaches in stroke prediction. In *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pages 1464–1469. IEEE, 2020.

[3] Damir Imamovic, Elmir Babovic, and Nina Bijedic. Prediction of mortality in patients with cardiovascular disease using data mining methods. In *2020 19th International Symposium INFOTEH-JAHORINA (INFOTEH)*, pages 1–4. IEEE, 2020.

[4] Rahul Katarya and Polipireddy Srinivas. Predicting heart disease at early stages using machine learning: a survey. In *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pages 302–305. IEEE, 2020.

[5] A Lakshmanarao, A Srisaila, and T Srinivasa Ravi Kiran. Heart disease prediction using feature selection and ensemble learning techniques. In *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*, pages 994–998. IEEE, 2021.

[6] Yu Lin. Prediction and analysis of heart disease using machine learning. In *2021 IEEE International Conference on Robotics, Automation and Artificial Intelligence (RAAI)*, pages 53–58. IEEE, 2021.

[7] Yu Lin. Prediction and analysis of heart disease using machine learning. In *2021 IEEE International Conference on Robotics, Automation and Artificial Intelligence (RAAI)*, pages 53–58. IEEE, 2021.

[8] Amita Malav, Kalyani Kadam, and Pooja Kamat. Prediction of heart disease using k-means and artificial neural network as hybrid approach to improve accuracy. *International Journal of Engineering and Technology*, 9(4):3081–3085, 2017.

[9] Shamsher Bahadur Patel, Pramod Kumar Yadav, and DP Shukla. Predict the diagnosis of heart disease patients using classification mining techniques. *IOSR Journal of Agriculture and Veterinary Science (IOSR-JAVS)*, 4(2):61–64, 2013.

[10] Tharanga Peiris et al. Heart disease stages prediction using machine learning. In *2022 8th International Conference on Big Data and Information Analytics (BigDIA)*, pages 504–511. IEEE, 2022.

[11] Vijeta Sharma, Shrinkhala Yadav, and Manjari Gupta. Heart disease prediction using machine learning techniques. In *2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, pages 177–181. IEEE, 2020.

[12] Mohd Zubir Suboh, Muhyi Yaakop, Mohd Syazwan Md Yid, Mohd Azlan Abu, and Imran Mohammad Sofi. Heart valve disease screening system—pc based. In *2017 International Conference on Engineering Technology and Technopreneurship (ICE2T)*, pages 1–4. IEEE, 2017.

[13] Guna Sekhar Reddy Thummala and Radhika Baskar. Prediction of heart disease using decision tree in comparison with knn to improve accuracy. In *2022 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES)*, pages 1–5. IEEE, 2022.

[14] Vedha Krishna Yarasuri, Dhumsapuram Saikrishna Reddy, Puligundla Sai Muneesh, Ramabhotla Venkata Sai Kaushik, Thupalli Nanda Vardhan, and KL Nisha. Developing machine learning models for cardiovascular disease prediction. In *2022 2nd Asian Conference on Innovation in Technology (ASIANCON)*, pages 1–6. IEEE, 2022.

[15] Bing Zhang, Jiadong Ren, Yongqiang Cheng, Bing Wang, and Zhiyao Wei. Health data driven on continuous blood pressure prediction based on gradient boosting decision tree algorithm. *IEEE Access*, 7:32423–32433, 2019.

[16] https://www.kaggle.com/datasets/cherngs/heart-disease-cleveland-uci

[17] https://deepchecks.com/glossary/normalization-in-machine-learning/