VFAST Transactions on Software Engineering
http://vfast.org/journals/index.php/VTSE@ 2023, ISSN(e): 2309-3978, ISSN(p): 2411-6246
Volume 11, Number 2, April-June 2023 *pp:149-159*

Using Machine Learning Models for The Prediction of Coronary Arteries Disease

Muhammad Bilal*, Naeem Aslam, Ahmad Naeem, Muhammad Kamran Abid

Department of Computer Science, of Engeerning and Technology, Multan, Punjab, Pakistan
*Corresponding author email: 2k19mscs118@nfciet.edu.pk

ABSTRACT

Globally, the leading cause of mortality among both men and women is coronary heart disease. This disease is widely recognized as the primary killer worldwide, and its early detection poses a significant challenge. Given the current state of affairs, it is crucial to promptly identify heart disease in its initial stages to ensure successful patient treatment. Despite numerous attempts by various researchers to develop hybrid and ensemble models for early detection, the desired outcomes have not been achieved. Consequently, the machine learning and algorithmic research community has directed its focus towards improving these methodologies. In this particular study, six supervised machine learning classifiers, namely Random_Forest, extreme gradient boost, Logistic of Regression, Decision_Tree, KNN, and N-Bayes, were employed. The UCI repository dataset was utilized as the sample data, comprising attributes and corresponding values. Data preprocessing techniques were employed to eliminate any missing values. An ensemble model incorporating three algorithms, namely DT (decision-tree), RF (random-forest), and XGB, was constructed. Remarkably, the ensemble model achieved an impressive accuracy rate of 95.33% for predicting coronary heart disease.

KEYWORDS
Machine Learning model, Heart disease, Heart failure, Machine Learning classifiers, Ensemble model, Decision Tree, Logistic regression, XGB, KNN, Naive byes

JOURNAL INFO
HISTORY: Received:May 25, 2023 Accepted: June 27, 2023 Published:June 30, 2023

INTRODUCTION

The body of human is prepared by many organs and each organ has a distinct function. The heart is considered as the central organ of the human being[1] It circulates blood from the heart to extra parts of the body. The human-heart has 4 main functions: pumping oxygenated the blood to the parts of body, pushing hormones and other necessary substances to other organs of the body, it receives deoxygenated blood from the body containing substances. scum and move away. reach the lungs for further oxidation and advancement and also maintain blood pressure. For this purpose, we have three main blood vessels that perform all these functions.

The human-heart is main organ of the body of human. If it doesn't work properly, it will affect other organs of body. In one report confirming that 7(million) people expires from attacks of heart per anum. According to the world health organization statement, about 18 million persons died from this disease in 2k15[2]. 31% of deaths are due to heart disease worldwide each year. Pumping of blood in human is an chief purpose of heart, providing O_2 and nutrient in human and removing extra metabolic wastes from human body. If anemia in the human body, the heart will not work well causes death of the person.[3] Angina arises if there is a momentary injury in plasma to the human-heart, beginning chest pain. Circulatory disease is of two types.

Disease of heart is jam or block your coronary clotting supply paths usually carried out by the development of greasy material called plaque.[4] Coronary vein infection is additionally called coronary illness, ischemic coronary illness, and coronary disease .CHD ensues when plaque builds up in a patient's arteries As plaque continues to build

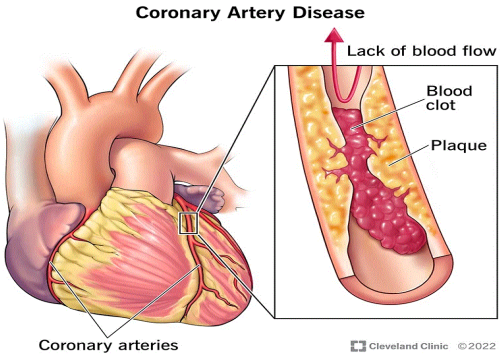

Figure 1 Coronary arteires disease

up, the patient's coronary arteries will narrow over time and reduce blood flow to the heart, thereby reducing the risk of coronary artery disease. enlarged danger of heart outbreak [2] Hereditary heart illness is a over-all term for a diversity of birth defects that affect the normal functioning of the human-heart.[3] as shown in the Figure 1 The term "intrinsic" income that the disorder is present from birth. Intrinsic coronary heart disease is one of the best known forms of birth

defects

Coronary Artery Disease

A heart attack or heart failure is considered a condition in which the tissues and muscles of the heart do not work properly as they should. Regular blood flow creates fluid growth that is so essential to the lungs and hoofs. Fluid growths can cause shortness of breath and enlargement of the feet and feet. Poor blood circulation can cause the skin to appear blue (cyanotic)[5]. This type of condition occurs when the heart's blood vessels narrow and are unable to properly pump blood around the body as shown in the Figure 2

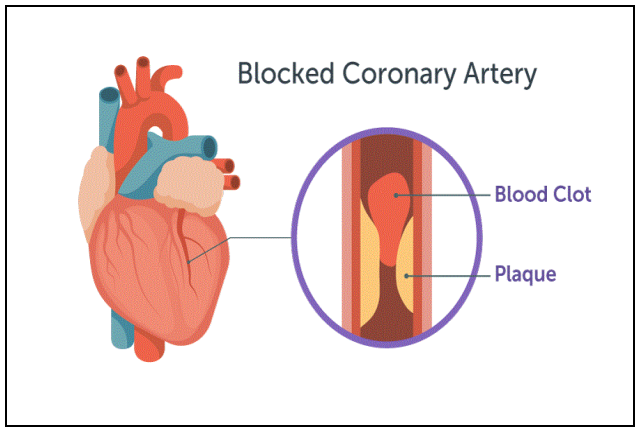

Figure 2 coronary artries

The aim of this study is to advance a classification arrangement that can exactly predict CAD in patients. The problematic reason is choose the Accurate ML model which increase the accuracy of predicting CAD patients The algorithms that will be used for predicting purpose are NB, LR, DT, XGB, . After that, superlative model will undertake an reasonable part where different methods will be studied. Most of the time we feed a different type of data set for the same classifier without feature selection technique, they will give different results and suggest any other classifier better. In this study, we apply six classifiers on 2 dissimilar types on data sets in relations of features and value firstly compute the results without feature selection techniques. We run all 6 algos on data set and compute results and make ensemble of three classifiers which give more accurate results

Around the worldwide the chief focus is to care about heart health and its improvement. There is lack of healthcare services and resources for heart disease . paitents and due to this factor heart disease cases increases gradualy yearly detection of heart disease and prediction of early stage heart disease paitent can save human life suffering from disease of heart now a recent time of world many ML algorithms are available to predict these patient with highly accurate ensemble model of Machine learning

1. Which machine learning algorithm was applied well to sort the CAD problem?
2. How to evaluate the performance of classifiers?
3. How you defend your predicting model is more effective then others?
4. Identify the best algorithm with the best performance by assessing the efficiency of a ML classifier
5. To relate the performance of several ML classifiers to advance the accuracy of the ML classifier by using the best fit classifiers
6. To increase the accurateness of previous ML models

This study includes related information on disease of heart. Data preparation data preprocessing

Ensemble models ML algorithms Jupiter notebook for applying ML algorithm have trained and evaluate their performance for the best and more accurate results a hybrid model is developed by combining ML classifiers

- Data cleaning preprocessing been applied for best accurate results
- Data set available at public domain UCI Repository using machine learning models like svm knn rf Dt nb to evaluate results
- after applying all these mentioned ML model developed an ensemble model using these classifiers after evolution of this model create confusion matrix
- comparison of performance analysis with single algorithm with ensemble model using confusion matrix
- using python language and its framework for implementing these classifiers with seaborn library using profiling of panda for finding the co-relation between variables

In the realm of machine learning, various algorithms, also referred to as classifiers, can be employed to assist in making predictions for our project. For our specific undertaking, we endeavor to forecast the number of patients with heart disease and without heart disease. To achieve this, we employ four algorithms on our dataset. The utilization of multiple algorithms allows us to obtain more accurate and dependable predictions. When relying solely on a single algorithm or classifier without any comparative reference, it becomes challenging to ascertain the reliability of the prediction. Although an algorithm may yield a high accuracy rate, it might not necessarily be the optimal or most suitable choice for our specific scenario. Conversely, employing multiple algorithms or classifiers, such as the four we have selected for our project (RF, LG, KNN, XGB, DT, NB), enables us to compare their performance. By observing significant disparities in precision between classifiers, we can identify potential issues. It could be that the algorithm itself is ill-suited for the task or that an error was made in our coding. Consequently, employing a range of algorithms becomes crucial for any prediction-based system.

LITERATURE REVIEW

Literature review or related work is a section in your research paper which describe the previous work and proposed that methodology to solve your problem. In this section we discuss about previous research done by the researcher and check that what kind of techniques and methodology they have used to solve their problem , so that you can used their proposed methodology to solve your work with more efficient manner. In this section we have mainly focused on the similar type of problems and techniques to solve his problem. This section also contains the methods that are used to solve the problems or address the problem that are solved with the similar methodology already used in the previous researches.

With the advent of the contemporary era, machine learning (ML) classifiers have emerged as crucial components in the advancement of heart disease classification models. Researchers have employed diverse machine learning techniques to address health-related challenges. Various methodologies have been utilized to classify heart disease.

This study offers predictive technique for classifying diseases of heart. The risk factors who can control and who can't have been explained in this article. Heart disease prediction is made by a random . The guess correctness found by the organization is around 80 percent. HDPS includes clinical data part, ROC curve part, estimated display part The authors proposed a diabetes prediction system that provides an analysis of diabetes. Two algorithms have been applied, namely Bayesian and K-NN to predict diabetes. [6]author has proposed a model for predicting heart disease by taking samples of 1208 patient record using Naïve Bayes and decision trees. data was taken from UCI repository site

The[7] researcher used a machine learning algorithm to construct a decision_tree. For a small data set, the decision tree does not give accurate results, but Naïve bayes gives more accurate results if the input data is cleaned. The author proposes a data mining model to predict whether patients have heart disease or not. Two types of naive array and data mining algorithm decision trees were used for prediction. These two algorithms were applied to the same data set.

DT displays 91% accurateness and NB displays 87% accurateness [8]. Therefore, paper-based decision trees give better results in this case.

The [9] authors proposed a data mining approach to predict cardiac disease in their paper. The dataset was obtained from the website of the UCI machine learning repository. The authors used four data mining techniques to predict cardiac disease: Nave Bayes, random forest, linear regression, and decision tree. Among these methods, random forest has a high accuracy of 90.16% when compared to others.

The [10] authors compared the accuracy of kNN, decision trees, linear regression, and support vector machine algorithms in predicting heart disease. The UCI repository website provides access to all datasets for prediction. Python is used to implement the algorithm. All algorithms are described in the Jupyter notebook. The authors attained the highest accuracy of 87% using the k-nearest neighbor approach, followed by support vector machines (83%), decision trees (79%), and regression. Among these heart disease prediction methods, linear regression has the highest accuracy (78%).

In their study [11], the authors proposed an application that utilizes multilayer perceptron algorithms to predict heart disease in adolescents. The Cleveland dataset, which can be accessed from the UCI library, was employed, comprising 76 parameters such as chest pain, CT scan, and ECG. The dataset was processed using Python code in the PyCharm engine. Experimental results yielded accuracy, memory, and support values of 0.92, 0.9, and 93 for positive classes, and 0.91, 0.89, and 0.72 for negative classes, respectively. The authors developed a cardiovascular disease prediction model using a hybrid random forest machine learning algorithm with linear mode, achieving an accuracy of 88.7% for CVD prediction. The dataset was obtained from the UCI repository website, specifically the Cleveland dataset, for this proposed study.

Random Forest (RF), a widely used algorithm, is employed for identifying health-related issues and solving regression and classification problems. Instead of using a single decision tree (DT), random forests generate multiple decision trees. The model is trained by creating subsets from the RF training set data, and each subset is used to generate a decision tree. The predictions are made based on the outcome of each individual decision tree [11].

Logistic regression is a supervised machine learning classifier used to solve classification problems [12]. It predicts binary outputs such as "yes" or "no" and is based on probability theory. The dependent variable is evaluated based on the independent variable, resulting in a discrete value of 0 or 1. However, due to its probabilistic nature, logistic regression returns values ranging from 0 to 1.

Naive Bayes, built on Bayes theorem, is considered the most efficient probabilistic machine learning classifier for solving classification problems [3]. The Naive Bayes classifier assumes the occurrence of one feature is independent of the occurrence of another feature. It is commonly used for binary classification and predicting historical outcomes. Its effectiveness relies on Bayes theorem.

K-nearest neighbors (KNN) used for both regression and classification tasks, with a primary focus on classification [13]. It is a slow learning algorithm as it lacks a specific training step and uses each data item for both training and classification. It does not make any assumptions about background information, making it a non-parametric classifier. During the training stage, KNN stores the dataset, and when new information is encountered, it assigns it to a class similar to the new information.

KNN starts by selecting the k-nearest neighbors as the closest points. Using the Euclidean distance approach, it calculates the distance between the test data and the training data of the k neighbors. These distance values are then arranged in ascending order, and the relevant components in each classifier are counted. The test data is assigned a class based on the majority class among the k neighbors [13].

The DT classifier [14] is a tree-like structure of a supervised machine learning classifier that is recommended for solving classification problems with nonlinear data. It predicts the value of the target class based on decision policies. The decision tree resembles a tree-like architecture, where inner nodes represent features or attributes, branches represent decisions, and leaf nodes represent outcomes.

Another study [15], author employed ML algos such as LR,DT, and NB which identify early stages case The dataset are utilized to analysis, & the results showed that Gaussian Naive Bayes achieved 76% accuracy, Decision Tree achieved 79.31% accuracy, and linear regression achieved 82.75% accuracy.

Thalliam and Peek [16] utilized neural network modeling to classify heart disease. Researchers have employed different classifier methods for disease prediction, and the most accurate classifiers were selected based on the obtained results. Sudhaa conducted a similar study in their research article.

In the prediction of heart disease, researchers [17] utilized several ML classifiers including Naive Bayes, decision trees, and neural networks. Different datasets from various patients were utilized, and these classifiers were employed to predict and mitigate the occurrence of heart disease.

In the evaluation of heart patients' dataset, other contemporary ML classifiers such as Classifiers, Decision Tree, JRIP, Naive Bayes, Gradient Stochastic, SVM, and K-nearest Neighbor were applied, and the results were compared to choose the best classifier [18].

Senthi Mohaan [19] developed a model based on hybrid approach, combined Random Forest model with the Linear Model, achieved an accuracy of 89% for heart disease. This research study suggested a new hybrid system.

In [20], the dataset related to heart disease was used to create a multi-layer artificial neural network perception. Along with the suggested approach, a program was employed to identify heart illness based on predetermined symptoms such as age, gender, and maximum heart rate. This method allowed for a more accurate system. A new hybrid system was proposed in this research article.

Data mining and ML classifiers were utilized based on various risk factor attributes, and the results showed that using more attributes led to better accuracy for some ML classifiers such as Genetic Algorithm, Decision Tree, K-nearest Neighbor, and Naive Bayes [21].

In another study [22], the author proposed a new genetic algorithm that utilized the backpropagation technique. The research included 13 heart-related risk factors such as blood pressure, cholesterol, and gender. The suggested method yielded better results for predicting heart disease.

To increase the accuracy of the system, the Random Forest model was combined with data mining techniques [23]. The experiment's results demonstrated that the suggested method can more accurately forecast the disease in its earliest stages.

The decision tree algorithm was applied in [24], where heart-related risk factors that are not typical were examined. The suggested model also employed a boosting strategy to improve system performance.

In [25], the authors employed various ML methods including logistic regression, random forest (RF), neural networks (NN), and boosting machines. They utilized 378,256 patient records from the United Kingdom and evaluated the results based on risk factor values. According to the AUC data, random forests improved accuracy by up to 1.7%, neural networks by 3.6%, and logistic regression by 3.2%.

The author of [26] presented ANN with backpropagation to identify cardiac disease in its early stages. They used a dataset with 13 risk indicators and achieved a proposed accuracy of 95%.

Ashook Kummar [27] applied different ML classifiers such as SVM, logistic regression, Naive Bayes, and ANN. Based on the experiment, logistic regression showed the highest accuracy on the UCI dataset.

[28] proposed a novel decision support system that combines a support vector machine with a genetic algorithm to predict sickness in its early phases. The Cleveland heart disease dataset was used, which contained variables including age, blood sugar level, blood pressure, and maximum heart rate. A novel system was proposed in this paper.

Arabas diet [29] suggests that by increasing the initial weight using a general method, neural networks can improve their accuracy by 10%.

Yil_maz [30] chose the least square support machine with a binary decision tree technique to more accurately identify heart disease.

This article [31] shows that heart disease can be diagnosed using market data. Every significant risk factor for heart disease was considered, resulting in an increased accuracy rate of 82%.

In this research study [32], ensemble computer learning techniques were employed, and four separate datasets collected from different medical centers were utilized. Detecting and training for heart disease were performed using a boosting strategy. The adaptive boosting algorithm provided varying levels of accuracy on different datasets.

Josh et al. [33] recommended a heart disease prediction model based on ensemble classifiers. Machine learning algorithms such as random forest (RF), SVM, AdaBoost, logistic regression (LR), bagging, and voting ensemble were employed. The main goal was to evaluate

performance in terms of sensitivity, specificity, and accuracy. The suggested system achieved an accuracy rate of 87%. Different classifiers perform differently depending on various factors, and Naive Bayes also provided 84% accuracy.

In [34], Mega Sahhi proposed the use of data mining techniques to detect heart disease. Various classifiers, including SVM, Naive Bayes, Association Rule, Decision Tree, KNN, and ANN, were implemented using the WEEKA Tool. Are shown in the table 1.

.

Table 1. Review of most research related articles

Year	Reference	Title of Paper	Classifier	Accuracy%
2018	[6]	Estimate of HD using ensemble learning and Subdivision Swarm Optimization	Naïve Bayes K-nearest neighbor	82.6% accuracy
2019	[11]	Improving the accuracy of prediction of HD risk based on ensemble organization practices	Naïve Bayes	76% accuracy
2019	[15]	ML-based coronary artery disease diagnosis: A inclusive review	Ensemble model	86.93%
2019	[14]	HD prediction using machine learning analytical approach and Random forest algorithm	Random forest model	76% accuracy
2020	[8]	A data driven approach for predicting the heart disease by using logistic regression	logistic regression	83.12% accuracy
2020	[16]	A innovative intelligence system based on machine learning system for coronary heart disease prediction	XGBOOT	82% accuracy
2020	[12]	(CHD) calculation using efficient ANN Model	KNN	87% accuracy
2021	[13]	Design and Application of CAD prediction using NB	Naïve Bayes	79 % Accuracy
2021	[17]	Effective ML based CAd calculation using Logistic regression	Logistic regression	77% accuracy
2022	[18]	CAD calculation using supervised type ML algorithm	RF, LG, SVM, KNN gives below then 88 % accuracy but Naïve Bayes Gives 85% accuracy.	88% and 85%

METHODOLOGY

The way that research is conducted is known as research methodology. It describes many methods and approaches that are employed to locate particular data regarding any research topics. The basic purpose of methodology is to specify the procedures for conducting research, the proposed methodology is shown in the Fig 3.

Dataset Collection & description

A data set is a collection of recorded values from different ethnicities and locations. Several experiments were conducted on these data values to obtain the required results.

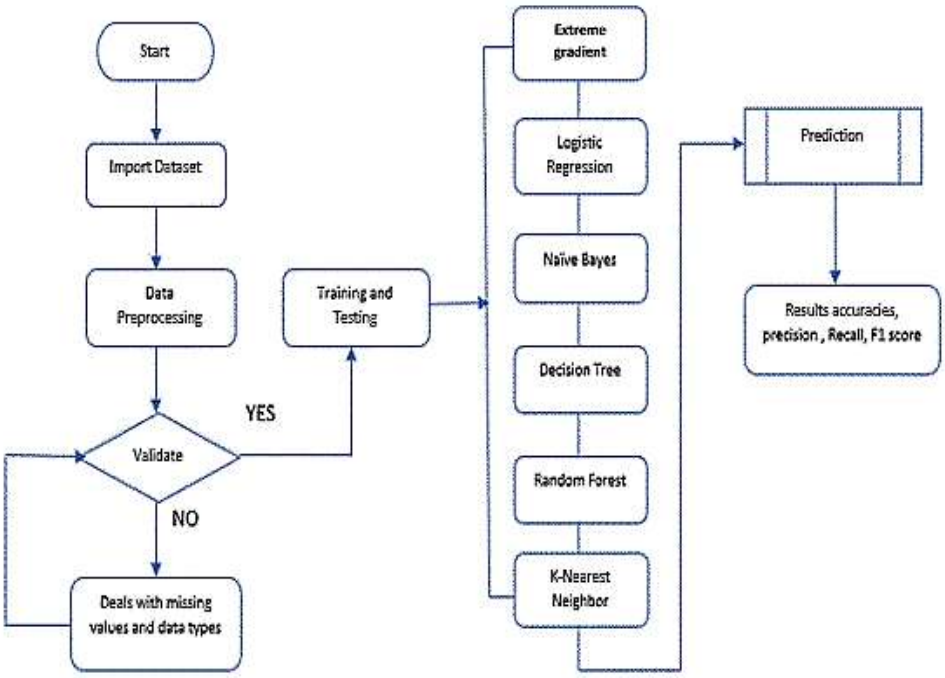

Figure 3. Research Methodology

Several researchers have used the Cleveland dataset in their studies. we also used it with another type of dataset obtained from the UCI dataset repository, having 14 attributes with 1024 different patients record

Attributes Description

This dataset is obtained from UCI repository having 14 attributes with 1024 different patients record are shown in table 2.

Table 2. Attribute Description

Attributes	Description	Values
1-Age	Age of paitent	Age values
2-Gender	Gender	1-Male-gender, 0=Female-gender
3- C_P	Chest-pain	1 = typ-ang, 2 = atyp-Ang, 3 = non-ang
4-trest_bps	Blood-pressure in rest	BP values
5-CHOL,	Dietary fat	Cholesterol values
6-F-B-S	Sugar in blood	1 = present, 0 = Absent
7-restecg	Echocardiography at rest	1=Abnormal-ECG, 0=Normal
8-oldpeak	Exercise related to rest	Different values
9-slope	ST depression slope	0,1,2 represent different slopes
10-CA	Vessels	0,1,2,3 values represent how your arteries are effected
11-thal	Thalassemia	0,1,2,3 represent colored Fluoroscopy vessels
12- thalach	Patient maximum heart rate	Heart rate values
13-exang	Angina with exercise	1=present, 0=absent
14-target	Heart Disease	1= effected person, 0= healthy

Remove Null Values

In this section first we remove the missing value from the dataset as shown in the Figure 4.

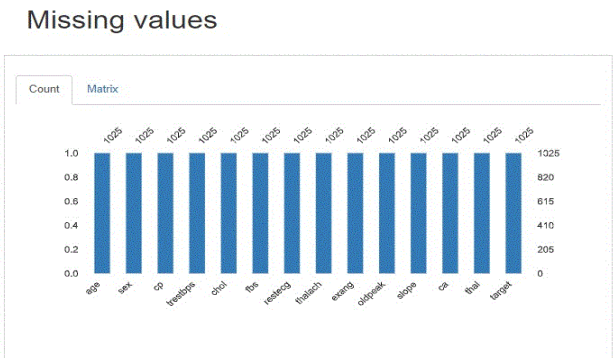

Figure 4 Missing values

Dataset Preprocessing

Data set preprocessing is the procedure of changing raw data keen on an understandable arrangement. The whole process is divided into different parts like data cleaning, data normalization, removing null values from the dataset and finally dividing the data into 2 portions, one is the testing dataset and the other is the data set. the second is the testing data set. These steps are important to transform the data into a valuable representation and increase the productivity of the model.

Min-Max-Scaler technique is used to shift features to a certain range to get the best results from the classifier. This range is the distinction between the initial maximum and the initial minimum. It subtracts the min value of the function and divides by the range.

Min-Max-Scaler provides a range for each feature from 0 to 1. In the split step of the dataset preprocessing we have to divided the data_set keen on two portions in the initial part we generate the training dataset of 80% and the second part checks the data set to 20%. First, we train our machine learning classification models by training a dataset and the remaining 20% are applied to detect CAD in its early stages

Selection of Algorithm

In machine learning, we can use different algorithms also known as classifiers to help us make predictions for our project. Here in our project, we aim to predict the number of patients with heart disease and the number of patients without heart disease by running four algorithms on our data set. The reason we will use four is that it will allow us to get better and more reliable predictions. Because if we use an algorithm or a classifier and we have nothing else to compare, we cannot say it is a reliable prediction, because it can give us very high accuracy, but this algorithm may not be the best or most suitable. to use for our script. While if we use more than one algorithm or classifier in our case four of them we can compare them with each other and if we find that one classifier provides gives us precision not even in the shadow of another algorithm provided we can figure out that something is wrong. Maybe the algorithm itself isn't right for the job, or we made a mistake in our coding. Therefore, the use of some algorithm is essential for any prediction-based system. Now the algorithms we have chosen to use in our project are:

Random Forest (RF)

Very popular RF is designed to diagnose health related problems. It is used to solve regression and classification problems. Random forests generate multiple decision trees instead of a single DT. First, we train the model so that from the RF training set data, first generate a subset, and from the subset, the algorithm generates a decision tree. For each decision tree, a new subset is produced using the training-set. Built on the outcome of separately decision tree, it predicts the required outcome.[11]

If the number of trees in FR = T, then the number of votes received from these trees is m, so

$V_m = \sum I (y_t == m)$

$t=1$

From the Y_T We can estimate the T_{th} no of tree, and if the condition is meet, then the function (Y_T==m) get the value 1 other wise 0. At the end RF makes the final prediction based on most no of votes are shown in the Figure 5.

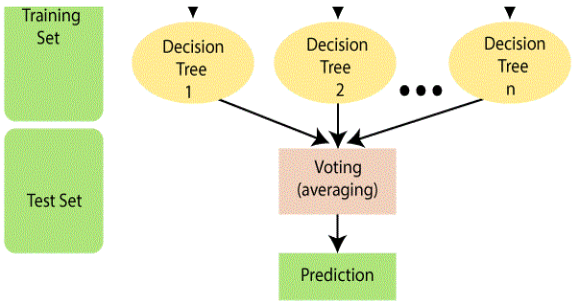

Figure 5 random forest

Logistic Regression

Logistic classifier is a type of (supervised) ML classifier applied to problem solving related to classification [12]. It is used to predict binary production as YES/NO. It is also based on the concept of probability. Here we calculate the ressult of the in-dependent variable based on the dependent variable. So value must be discrete or exactly like 0, 1. but since it is based on probability, so it gives values from 0 to 1. Logistic and linear both regression are alike, the key change is that logistic-R solves sorting problem and linear-R solves deterioration problem.

We use an S like shaped bend to predict binary(0,1) which is called the logistic function., and this S-shaped curve indicates whether a patient suffering from CAD or no.

Logistic.function also known as Sigmond function is used to convert regression line to decision limit, also mapping predicted values to probability. It maps two values in the range 0 and 1. In this function, values above hold tend to be 1 and values below hold tend to 0.

Naive Bayes

Naive bayes is the most efficient probabilistic ML classifier, built on Bayes theorem, used to solve classification problem [3]. The NB-classifier supposes the

occurrence of a particular section in a session unconnected to the occurrence of another element. Naive Bayes being used for the binary classification and source prediction of historical outcomes, as shown in the Figure 6. Since it depends on Bayes theorem, its equation is represented by

P/A|B = P(B|A) P(A)/P(B)

P/A|B = the prospect of probability A happening, given chance B has happened

P/B|A = the prospect of probability B happening, given chance A has happened

P-A = the prospect of probability A

P-B = the prospect of probability B

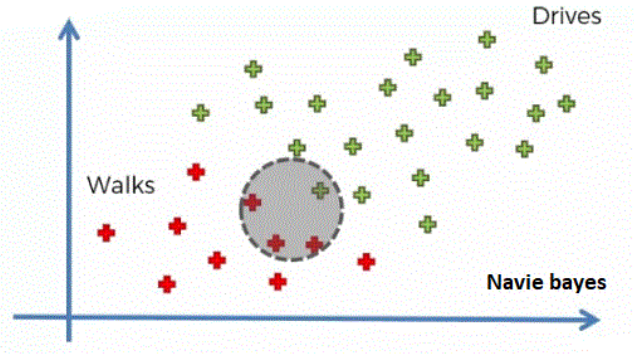

Figure 6. Naïve Bayes

Decision Tree

Classifier-DT is a graph-like tree of a supervised machine learning classifier recommended to solve classification problems by nonlinear data [14] as shown in the Figure 7. It predicts the value of the target class based on decision policies. AS is a tree like construction anywhere inner-nodes characterize elements or attributes, subdivisions represent conclusion and leaf-nodes signify marks. We have several types of nodes in the decision tree. At the beginning, we have a root node representing the entire sample dataset. Decision nodes are used to make decisions & leaf nodes are the results of those decisions. The parent nodes are the base nodes and the child-nodes are children of the parental node.

In decision_tree, we rank the hypothetical dataset by starting the computation from the root node. Our proposed continues until it reaches the leaf node of the tree.

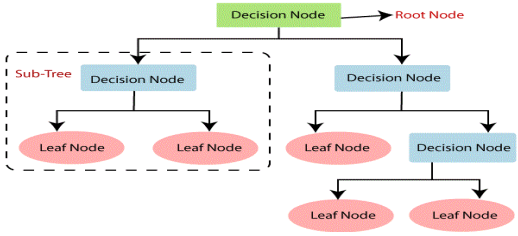

Figure 7 Decesion Tree

Extreme gradient boosting

XGB [19] is a tree like technique in the supervised machine learning type . It can be used for both classification and regression problems, but all formulas and examples in this narration relate to classification problems. Gradient-boost decision trees are implemented using XGBoost technology. In C++, this library was created. It is a kind of software library that was primarily created to increase model performance and speed. In recent years, it has dominated applied machine learning. Many Kaggle Competitions are dominated by XGBoost models. Decision trees are created in sequence in this method. Weights are significant in XGBoost. Each independent variable is given a weight before being put into the decision tree that forecasts outcomes. The variables are subsequently put into the second decision tree with an enhanced weight for variables that the tree incorrectly anticipated. These distinct classifiers and predictors are then combined to produce a robust and accurate model.

K-Nearest Neighbor

[13]KNN relate to the family of (SM)learning classifiers, which is run to solve problems related to regression and classification, but mainly focus on classification. KNN is a slow learning algorithm because it has no specific preparation step and uses each data item to train while classifying. (KNN) has no parametric classifier because this one marks no expectations about background information. KNN classifier at the preparatory stage just stores the dataset and when it finds new information then at that point it organizes that information into a class that looks like the new information.

KNN starts its work by choosing its k-neighbors as the nearest-points. it compute the space between the test-data and the training-data of the K neighbors using the new method term Euclidean distance. Now these distance values being arrange in ascending order and count the number of relevant elements in each classifier, then it assigns a class to the test score according to the most continuous class of the lines this.From confusion matrix, we compute four performance evaluation metrics

Machine Learning Algorithms Results

After applying all the machine learning classifier on the dataset we get this results shown in the table 3.

Table 3. Machine Learning Algorithms Results

Classifier Name	Accuracy	Precision	Recall	F1 score
Extreme gradient boost	94.43%	92%	94%	95.00%
Logistic regression	86.34%	85%	84%	86%
Naive Bayes	85.36%	88%	84%	81%
Decision tree	94.36%	92%	95%	97%
Random Forest	94.36%	98%	94%	91%
k- Nearest	87.80%	88%	86%	84%

Result and discussion

In this study of research we practical about six ML classifiers , on UCI datasets and to get our required results. ML classifiers which are use for experiments are

Random forest, K-nearest-neighbor, LR, DT, and NB, XGB . We choose dataset according to their attributes and values. dataset contains 14 risk factors and having values in Boolean forms, We used three classifier which give maximum accuracy so that we select the necessary factors and derive our results with enhance accuracy.

Here in our experiment first we applied proposed classifiers on the dataset without ensemble technique and derived their accuracies alongside by (precision-recall-F1 score). After computing results , we applied the ensemble techniques and check that our accuracy increases or not by applying these techniques

Model Accuracy Graph

Accuracy graph of our selected algorithms on dataset is shown in theFigure 7.

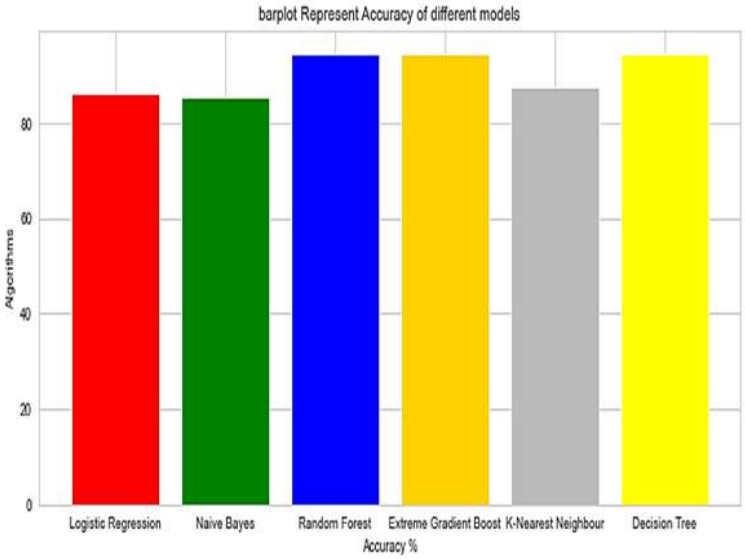

Figure 7 Model Accuracy Graph

Our Ensemble model ML algorithms

Ensemble model

As we discus in result and discussion our esenmble model is shown in figure 7.

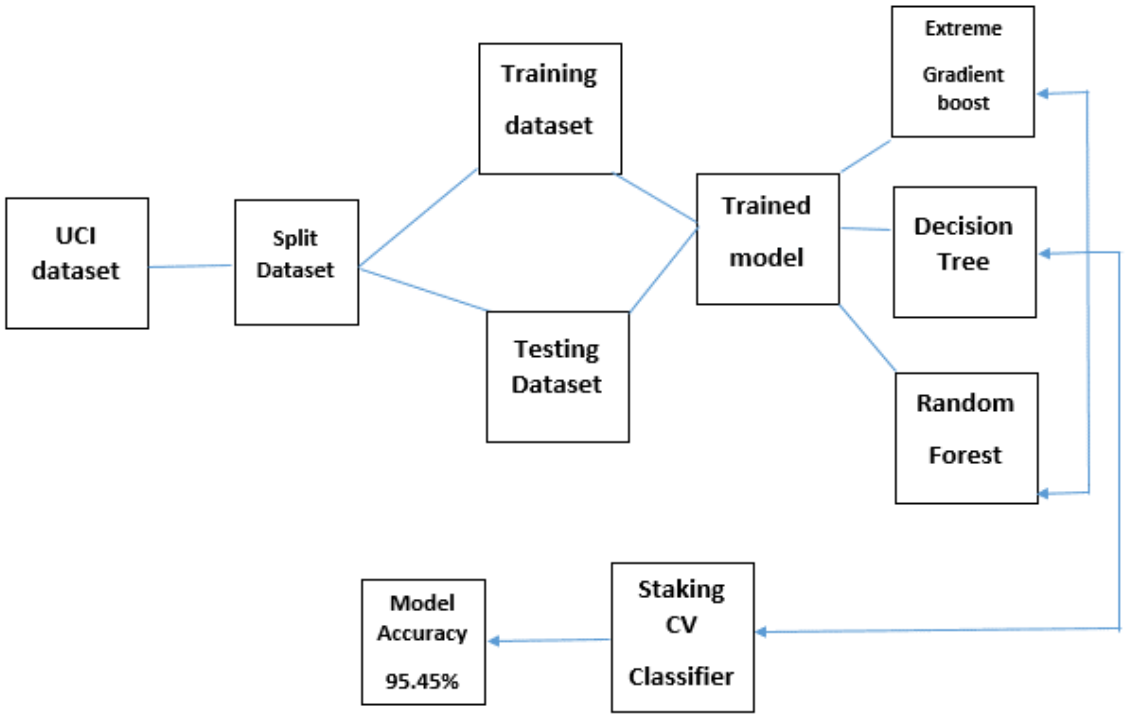

figure 7 Ensemble model diagram

Comparison b/w ensemble model vs other ML algorithm

As we compare other machine learning models with our esenmble model in tabular form shown in the Table 4.

Table 4.Comparison b/w ensemble model vs ML algorithms

Classifier Name	Accuracy	Precision	Recall	F1 score
Extreme gradient boost	94.43%	92%	94%	95.00%
Logistic regression	86.34%	85%	84%	86%
Naive Bayes	85.36%	88%	84%	81%
Decision tree	94.36%	92%	95%	97%
Random Forest	94.36%	98%	94%	91%
k- Nearest	87.80%	88%	86%	84%
	Our proposed	**Ensemble model**		
Accuracy of Model	95.45%	95%	93%	95%

CONCLUSION AND FUTURE WORK

In this research study, we have utilized six machine learning algorithms on a dataset from the UCI repository, without implementing any feature selection technique. Our research was conducted using the Python programming language with Jupiter Notebook. Initially, we applied our proposed classifiers to the datasets without employing feature selection and calculated various metrics including accuracy, precision, recall, and F1 score. Subsequently, we combined three classifiers, namely Random Forest, Extreme Gradient Boosting, and Decision Tree, to evaluate the accuracies and compare the results with our ensemble model. Remarkably, our model yielded superior accuracy, achieving a rate of 95.33%. In our research study, we specifically focused on numeric datasets and three classifier techniques. However, it is worth noting that we also explored the availability of alphanumeric datasets and feature selection techniques. Therefore, future researchers have the opportunity to apply alphanumeric datasets along with these techniques, as well as explore novel feature selection methods to further enhance the accuracy of their models.

REFRENCES

[1]. N. L. Fitriyani, M. Syafrudin, G. Alfian, and J. Rhee, "HDPM: An Effective Heart Disease Prediction Model for a Clinical Decision Support System," IEEE Access, vol. 8, pp. 133034–133050, 2020, doi: 10.1109/ACCESS.2020.3010511.

[2]. M. Abdar, W. Książek, U. R. Acharya, R. S. Tan, V. Makarenkov, and P. Pławiak, "A new machine learning technique for an accurate diagnosis of coronary artery disease," Comput. Methods Programs Biomed., vol. 179, 2019, doi: 10.1016/j.cmpb.2019.104992.

[3]. A. Kondababu, V. Siddhartha, B. B. Kumar, and B. Penumutchi, "A comparative study on machine learning based heart disease prediction," Mater. Today Proc., no. xxxx, pp. 1–5, 2021, doi: 10.1016/j.matpr.2021.01.475.

[4]. K. H., J. H., and G. J., "Diagnosing Coronary Heart Disease using Ensemble Machine Learning," Int. J. Adv. Comput. Sci. Appl., vol. 7, no. 10, pp. 30–39, 2016, doi: 10.14569/ijacsa.2016.071004.

[5]. S. Pouriyeh, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia, and J. Gutierrez, "A comprehensive investigation and comparison of Machine Learning Techniques in the domain of heart disease," Proc. - IEEE Symp. Comput. Commun., no. Iscc, pp. 204–207, 2017, doi: 10.1109/ISCC.2017.8024530.

[6]. I. Yekkala, S. Dixit, and M. A. Jabbar, "Prediction of heart disease using ensemble learning and Particle Swarm Optimization," Proc. 2017 Int. Conf. Smart Technol. Smart Nation, SmartTechCon 2017, pp. 691–698, 2018, doi: 10.1109/SmartTechCon.2017.8358460.

[7]. D. Sivabalaselvamani, D. Selvakarthi, L. Rahunathan, S. N. Eswari, M. Pavithraa, and M. Sridhar, "Investigation on Heart Disease Using Machine Learning Algorithms," 2021 Int. Conf. Comput. Commun. Informatics, ICCCI 2021, 2021, doi: 10.1109/ICCCI50826.2021.9402390.

[8]. I. D. Mienye, Y. Sun, and Z. Wang, "An improved ensemble learning approach for the prediction of heart disease risk," Informatics Med. Unlocked, vol. 20, p. 100402, 2020, doi: 10.1016/j.imu.2020.100402.

[9]. D. Krishnani, A. Kumari, A. Dewangan, A. Singh, and N. S. Naik, "Prediction of Coronary Heart Disease using Supervised Machine Learning Algorithms," IEEE Reg. 10 Annu. Int. Conf. Proceedings/TENCON, vol. 2019-Octob, pp. 367–372, 2019, doi: 10.1109/TENCON.2019.8929434.

[10]. M. Kavitha, G. Gnaneswar, R. Dinesh, Y. R. Sai, and R. S. Suraj, "Heart Disease Prediction using Hybrid machine Learning Model," Proc. 6th Int. Conf. Inven. Comput. Technol. ICICT 2021, pp. 1329–1333, 2021, doi: 10.1109/ICICT50816.2021.9358597.

[11]. C. B. C. Latha and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," Informatics Med. Unlocked, vol. 16, no. June, p. 100203, 2019, doi: 10.1016/j.imu.2019.100203.

[12]. C. Guo, J. Zhang, Y. Liu, Y. Xie, Z. Han, and J. Yu, "Recursion Enhanced Random Forest with an Improved Linear Model (RERF-ILM) for Heart Disease Detection on the Internet of Medical Things Platform," IEEE Access, vol. 8, pp. 59247–59256, 2020, doi:

10.1109/ACCESS.2020.2981159.
[13]. V. Shorewala, "Early detection of coronary heart disease using ensemble techniques," Informatics Med. Unlocked, vol. 26, no. July, p. 100655, 2021, doi: 10.1016/j.imu.2021.100655.
[14]. M. Abdar, E. Nasarian, X. Zhou, G. Bargshady, V. N. Wijayaningrum, and S. Hussain, "Performance improvement of decision trees for diagnosis of coronary artery disease using multi filtering approach," 2019 IEEE Int. journel Comput. Commun. Syst. ICCCS 2019, no. Dm, pp. 26–30, 2019, doi: 10.1109/CCOMS.2019.8821633.
[15]. R. Alizadehsani et al., "Machine learning-based coronary artery disease diagnosis: A comprehensive review," Comput. Biol. Med., vol. 111, p. 103346, 2019, doi: 10.1016/j.compbiomed.2019.103346.
[16]. W. Chang, Y. Liu, X. Wu, Y. Xiao, S. Zhou, and W. Cao, "A New Hybrid XGBSVM Model: Application for Hypertensive Heart Disease," IEEE Access, vol. 7, pp. 175248–175258, 2019, doi: 10.1109/ACCESS.2019.2957367.
[17]. X. Zhu, J. Chu, K. Wang, S. Wu, W. Yan, and K. Chiam, "Prediction of rockhead using a hybrid N-XGBoost machine learning framework," J. Rock Mech. Geotech. Eng., vol. 13, no. 6, pp. 1231–1245, 2021, doi: 10.1016/j.jrmge.2021.06.012.
[18]. G. N. Ahmad, S. Ullah, A. Algethami, H. Fatima, and S. M. H. Akhter, "Comparative Study of Optimum Medical Diagnosis of Human Heart Disease Using Machine Learning Technique with and Without Sequential Feature Selection," IEEE Access, vol. 10, pp. 23808–23828, 2022, doi: 10.1109/ACCESS.2022.3153047.
[19]. X. Zhu, J. Chu, K. Wang, S. Wu, W. Yan, and K. Chiam, "Prediction of rockhead using a hybrid N-XGBoost machine learning framework," J. Rock Mech. Geotech. Eng., vol. 13, no. 6, pp. 1231–1245, 2021, doi: 10.1016/j.jrmge.2021.06.012.
[20]. G. N. Ahmad, S. Ullah, A. Algethami, H. Fatima, and S. M. H. Akhter, "Comparative Study of Optimum Medical Diagnosis of Human Heart Disease Using Machine Learning Technique with and Without Sequential Feature Selection," IEEE Access, vol. 10, pp. 23808–23828, 2022, doi: 10.1109/ACCESS.2022.3153047.
[21]. J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan, and A. Saboor, "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare," IEEE Access, vol. 8, no. Ml, pp. 107562–107582, 2020, doi: 10.1109/ACCESS.2020.3001149.
[22]. J. Ma, Z. Yu, Y. Qu, J. Xu, and Y. Cao, "Application of the xgboost machine learning method in pm2.5 prediction: A case study of shanghai," Aerosol Air Qual. Res., vol. 20, no. 1, pp. 128–138, 2020, doi: 10.4209/aaqr.2019.08.0408.
[23]. J. H. Friedman, ''Greedy function approximation: A gradient boosting machine,'' Ann. Statist., pp. 1189–1232, Oct. 2001.
[24]. 24] A. Sharaff and H. Gupta, ''Extra-tree classifier with metaheuristics approach for email classification,'' in Proc. Adv. Comput. Commun. Com- put. Sci. Singapore: Springer, 2019, pp. 189–197. [
[25]. A. Pérez, P. Larrañaga, and I. Inza, ''Supervised classification with condi- tional Gaussian networks: Increasing the structure complexity from naive Bayes,'' Int. J. Approx. Reasoning, vol. 43, no. 1, pp. 1–25, Sep. 2006.
[26]. B. Schölkopf, C. Burges, and V. Vapnik, ''Incorporating invariances in support vector learning machines,'' in Proc. Int. Conf. Artif. Neural Netw. Berlin, Germany: Springer, 1996, , pp. 47–52.
[27]. Z. Ahmed, K. Mohamed, S. Zeeshan, and X. Dong, ''Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine,'' Database, vol. 2020, Jan. 2020.
[28]. R. Gupta, ''Recent trends in coronary heart disease epidemiology in India,'' Indian heart J., vol. 60, no. 2, pp. B4–B18, 2008.
[29]. X. Liu, X. Wang, Q. Su, M. Zhang, Y. Zhu, Q. Wang, and Q. Wang, ''A hybrid classification system for heart disease diagnosis based onthe RFRS method,'' Comput. Math. Methods Med., vol. 2017, pp. 1–11, Jan. 2017.
[30]. S. Mokeddem, B. Atmani, and M. Mokaddem, ''Supervised feature selec- tion for diagnosis of coronary artery disease based on genetic algorithm,'' 2013, arXiv:1305.6046.
[31]. A. M. Usman, U. K. Yusof, and S. Naim, ''Cuckoo inspired algorithms for feature selection in heart disease prediction,'' Int. J. Adv. Intell. Inform., vol. 4, no. 2, pp. 95–106, Jul. 2018.
[32]. A. U. Haq, J. Li, M. H. Memon, M. Hunain Memon, J. Khan, and S. M. Marium, ''Heart disease prediction system using model of machine learning and sequential backward selection algorithm for features selec- tion,'' in Proc. IEEE 5th Int. Conf. Converg. Technol. (I2CT), Mar. 2019, pp. 1–4.
[33]. A. Javeed, S. S. Rizvi, S. Zhou, R. Riaz, S. U. Khan, and S. J. Kwon, ''Heart risk failure prediction using a novel feature selection method for
[34]. feature refinement and neural network for classification,'' Mobile Inf. Syst., vol. 2020, pp. 1–11, Aug. 2020.
[35]. D. C. Yadav and S. Pal, ''Prediction of heart disease using feature selection and random forest ensemble method,'' Int. J. Pharmaceutical Res., vol. 12, no. 4, pp. 56–66, Oct. 2020.
[36]. R. Aggrawal and S. Pal, ''Sequential feature selection and machine learning algorithm-based patient's death events prediction and diagnosis in heart disease,'' Social Netw. Comput. Sci., vol. 1, no. 6, pp. 1–16, Nov. 2020.?