Prediction of Cardiovascular Disease using Machine Learning Algorithms

Muktevi Srivenkatesh

Abstract: ***Background/Aim:*** *Healthcare is an unavoidable assignment to be done in human life. Cardiovascular sickness is a general class for a scope of infections that are influencing heart and veins. The early strategies for estimating the cardiovascular sicknesses helped in settling on choices about the progressions to have happened in high-chance patients which brought about the decrease of their dangers.* ***Methods:*** *In the proposed research, we have considered informational collection from kaggle and it doesn't require information pre-handling systems like the expulsion of noise data, evacuation of missing information, filling default esteems if applicable and classification of attributes for prediction and decision making at different levels. The performance of the diagnosis model is obtained by using methods like classification, accuracy, sensitivity and specificity analysis. This paper proposes a prediction model to predict whether a people have a cardiovascular disease or not and to provide an awareness or diagnosis on that. This is done by comparing the accuracies of applying rules to the individual results of Support Vector Machine, Random forest, Naive Bayes classifier and logistic regression on the dataset taken in a region to present an accurate model of predicting cardiovascular disease.* ***Results:*** *The machine learning algorithms under study were able to predict cardiovascular disease in patients with accuracy between 58.71% and 77.06%.* ***Conclusions:*** *It was shown that Logistic Regression has better Accuracy (77.06 %) when compared to different Machine-learning Algorithms.*

Keywords: *Cardiovascular disease, Machine Learning Algorithms, Performance Evaluators, toxins*

I. INTRODUCTION

Classification is significant component of data mining. Classification is the way toward finding a model (or capacity) that depicts and recognizes information classes or ideas. The model is inferred dependent on the investigation of a lot of preparing cardiovascular data (i.e., data objects for which the class marks are known).

The model is utilized to foresee the class name of items for which the class name is having the cardiovascular malady or not having cardiovascular ailment that is obscure.

Machine Learning examines how computers can learn (or improve their exhibition) in view of cardiovascular information. The primary research zone is for computer projects to consequently figure out how to perceive complex examples and settle on clever choices dependent on cardiovascular data.

Supervised learning is fundamentally an equivalent word for arrangement. The supervision in the taking in originates from the named models in the cardiovascular preparing data collection.

Cardiovascular malady (CVD) is expanding day by day in this cutting edge world. As per the World Health Organization (WHO), an expected 17 million individuals die every year from cardiovascular ailment, especially respiratory failures and strokes [1]. It is, in this way, important to record the most significant side effects and wellbeing propensities that add to CVD. Different tests are performed before conclusion of CVD, including auscultation, ECG, circulatory strain, cholesterol and glucose.

These tests are regularly long and long when a ient's condition might be basic and the indivpatidual in question must beginning taking prescription quickly, so it gets imperative to organize the tests [2]. A few wellbeing propensities add to CVD. In this way, it is likewise important to know which wellbeing propensities add to CVD. Machine Learning is currently a developing field because of the expanding measure of information. Machine Learning makes it conceivable to secure information from a huge measure of information, which is overwhelming for man and here and there inconceivable [3].

The remaining of the research discussion is organized as follows: Section II briefs Literature , Section III describes brief description of selected machine learning algorithms Section IV describes Patient Data Set and attributes, Section V discusses Proposed Technique ,Section VI Describes Performance measure of classification, Section VII briefs discussion and evaluated Results, and Section VIII determines the Conclusion of the research work and last Section describes References .

A. Cardiovascular disease

Cardiovascular infection, by and large, alludes to conditions that include limited or blocked veins that can prompt a coronary episode, chest torment (angina) or stroke. Other heart conditions, for example, those that influence your heart's muscle, valves or cadence, likewise are viewed as types of coronary illness.

Cardiovascular malady incorporates conditions that influence the structures or capacity of your heart, for example,

- Coronary supply route infection (narrowing of the courses)
- Heart assault.
- Abnormal heart rhythms, or arrhythmias.
- Heart disappointment.
- Heart valve infection.
- Congenital coronary illness.
- Heart muscle sickness (cardiomyopathy)

Revised Manuscript Received on January 22, 2020.

Dr. M. Srivenkatesh, Associate Professor, Department of Computer Science, GITAM Deemed to be University, Visakhapatnam, Andhra Pradesh, India.

Cardiovascular sickness can happen when courses that supply blood and oxygen to your heart muscle and different organs, (for example, the cerebrum and kidneys) become stopped up with greasy material called plaque or atheroma. This procedure is called atherosclerosis.
"Cardio" alludes to the heart, and "vascular" alludes to all the veins in the body. In correlation, coronary illness is increasingly explicit and alludes just to maladies of the heart, for example, coronary conduit sickness, cardiovascular breakdown, heart valve variations from the norm, and anomalous heart rhythms.

II. LITERATURE SURVEY

A.U.Ul Haq,J.P. Li,M.H. Memon,S.Nazir,R.sun[4] has discussed the conjecture of Heart ailment and they have proposed an machine learning based discovering system for heart ailment desire by using heart ailment dataset. They have used seven surely understood machine learning, three-element choice calculations, the cross-approval technique, and seven classifiers execution assessment measurements, for example, characterization precision, particularity, affectability, Matthews' relationship coefficient, and execution time. They have made a system can without a doubt perceive and orchestrate people with coronary ailment from sound people. They have discussed the total of the classifiers, feature assurance figuring, pre-preparing procedures, endorsement technique, and classifiers execution appraisal estimations used in this paper. They have done execution of the proposed system has been endorsed on full features and on a diminished game plan of features. Their features decline influences classifiers execution with respect to exactness and execution time of classifiers. They have proposed machine learning based choice emotionally supportive networks will help the specialists to find heart patients effectively.
S. Krishnan J. Geetha S[5] has made a system that predicts the developing potential results of Heart Disease. Their aftereffects of this system give the chances of happening heart disease to the extent rate. They have considered datasets used are organized similar to therapeutic parameters. Their structure evaluates those parameters using the information mining plan strategy. Their datasets were set up in python programming using two standard Machine Learning Algorithm to be explicit Decision Tree Algorithm and Naive Bayes Algorithm and have exhibited the best estimation among these two to the extent the precision level of heart illness

K.G Dinesh, K.A.raj, K.D.Santhosh, V. M.eswari[6] has talked about heart illness expectation and performed information pre-preparing utilizes strategies like the removal of noisy data, removal of missing data, filling default values if applicable and classification of attributes for prediction and decision making at different levels. Their exhibition of the finding model is acquired by utilizing techniques like order, exactness, affectability and particularity examination. This has proposed a forecast model to anticipate whether people have heart illness or not and to give mindfulness or finding on that. They have done examination by comparing the accuracies of applying rules with the individual consequences of Support Vector Machine, Gradient Boosting, Random backwoods, Naive Bayes classifier and calculated relapse on the dataset taken in a district to display an exact model of foreseeing cardiovascular ailment.
A. Golande, P. Kumar T[7] has talked about heart illness and they have considered both male and female class and this proportion may fluctuate as per the district additionally this proportion is considered for the individuals of age bunch 25-69. This doesn't show that individuals with another age gathering won't be influenced by heart ailments. They have anticipated the reason and heart illness is a significant test these days. They have talked about different calculations and devices utilized for the forecast of heart sicknesses.
Prasad,P. Anjali, S.Adil, N.Deepa[8] has foreseen of heart illnesses using machine learning strategies by bridging the couple of ebb and flow looks into. They have used the calculated regression is utilized and the medicinal services information which arranges the patients whether patients are having heart maladies or not as per the data in the record and created data a model which predicts the patient whether they are having a heart illness or not.
Y. Khourdifi, M.Bahaj [9] has talked about the forecast of coronary illness and abused the Fast Correlation-Based Feature Selection (FCBF) strategy to channel excess highlights so as to improve the nature of coronary illness order. They have done order dependent on various arrangement calculations, for example, K-Nearest Neighbour, Support Vector Machine, Naïve Bayes, Random Forest, and a Multilayer Perception | Artificial Neural Network streamlined by Particle Swarm Optimization (PSO) joined with Ant Colony Optimization (ACO) approaches. Their proposed blended methodology is applied to heart illness dataset and accomplished a greatest order precision of 99.65% utilizing the upgraded model proposed by FCBF, PSO, and ACO.

III. MACHINE LEARNING ALGORITHMS

Machine Learning is modernized learning with for all intents and purposes zero human intervention. It incorporates programming PCs so they gain from the open data sources. The guideline inspiration driving AI is to research and manufacture estimations that can pick up from the past data and make desires on new information data.
The contribution to a learning calculation is preparing information, speaking to understanding, and the yield is any mastery, which typically appears as another calculation that can play out an assignment. The info information to an machine learning framework can be numerical, literary, sound, visual, or sight and sound. The relating yield information of the framework can be a gliding point number.

A. Concepts of Learning

Learning is the way toward changing over understanding into skill or information.
Learning can be comprehensively grouped into three classes, as referenced beneath,

In view of the idea of the learning information and association between the student and the earth.

- Supervised Learning process or Supervised Learning Approach
- Unsupervised Learning process or Unsupervised Learning Approach
- Semi-regulated Learning process or Unsupervised Learning Approach

Correspondingly, there are four classifications of Machine Learning as appeared beneath −

- Supervised learning process/Approach
- Unsupervised learning process/Approach
- Semi-directed learning process/Approach
- Reinforcement learning process/Approach

In any case, the most normally utilized ones are supervised and unsupervised learning

B. Supervised Learning

Machine Learning is normally used in genuine applications, for instance, face and talk affirmation, things or movie proposals, and arrangements assessing. Supervised learning can be moreover requested into two sorts - Regression and Classification.

Regression gets ready on and predicts a reliable regarded response, for example foreseeing land costs.

Characterization endeavours to find the correct class name, for instance, looking at valuable/hostile emotions, male and female individuals, kind and undermining tumors, secure and unbound credits, etc.

Supervised learning includes building machine learning model that depends on named tests

For instance on the off chance that we construct framework to discover of kind of fever dependent on different highlights of patient like temperature ,force of migraine, body agonies, hack and cool, different status parameters of blood to order quiet is having jungle fever, dingo, viral fever, sine flew and so forth .This is the incentive for class mark.

Supervised learning manages taking in a capacity from accessible preparing information. There are many supervised learning calculations, for example, Logistic Regression, Neural systems, Support Vector Machines (SVMs), and Naive Bayes classifiers.

C. Unsupervised Learning

Unaided learning is utilized to recognize inconsistencies, anomalies, for example, extortion or imperfect gear, or to aggregate clients with comparative practices for a business battle. It is something contrary to managed learning. There is no named data here.

When learning information contains just a few signs with no portrayal or names, it is up to the coder or to the calculation to discover the structure of the basic information, to find shrouded designs, or to decide how to depict the information. This sort of learning information is called unlabeled information.

Assume that we have various information focuses, and we need to characterize them into a few gatherings. We may not actually realize what the criteria of order would be. Along these lines, an unsupervised learning algorithms attempts to characterize the given dataset into a specific number of gatherings in an ideal manner.

Solo learning calculations are very amazing assets for examining information and for recognizing examples and patterns. They are most ordinarily utilized for bunching comparative contribution to consistent gatherings. Solo learning calculations incorporate K-implies, Random Forests, and Hierarchical bunching, etc.

D. Semi-supervised Learning

In the event that some learning tests are marked, yet some other are not named, at that point it is semi-supervised learning. It utilizes a lot of unlabeled data for preparing and a modest quantity of named data for testing. Semi-regulated learning is applied in situations where it is costly to get a completely named dataset while progressively pragmatic to mark a little subset.

E. Reinforcement Learning

Here learning data gives input with the goal that the framework acclimates to dynamic conditions so as to accomplish a specific goal. The framework assesses its exhibition dependent on the input reactions and responds in like manner.

A. Supervised Learning Algorithms

K-Nearest Neighbour Algorithm

K-closest neighbors (KNN) algorithm is a kind of supervised machine learning algorithms which can be utilized for both classification as well as regression predictive issues.

- Lazy learning calculation − KNN is a lazy learning algorithm since it doesn't have a specific training phase and uses all the data for training while classification.
- Non-parametric learning calculation − KNN is additionally a non-parametric learning algorithm calculation since it doesn't expect anything about the fundamental data.

K-closest neighbors (KNN) calculation utilizes 'highlight closeness' to anticipate the estimations of new data points which further implies that the new data point will be assigned a value based on how closely it matches the points in the training set. We can comprehend its working with the assistance of following advances –

Stage 1 − For executing any algorithm, we need dataset. So during the initial step of KNN, we should stack the preparation just as test information.

Stage 2 − Next, we have to pick the estimation of K for example the closest data points. K can be any whole number.

Stage 3 − For each point in the test information do the accompanying –

- 3.1 − Calculate the separation between test data and each row of training data with the help of any of the following methods namely:

Euclidean, Manhattan or Hamming distance. The most ordinarily utilized strategy to compute separation is Euclidean.

- **3.2** − Now, based on the distance value, sort them in ascending order.
- **3.3** − Next, it will choose the top K rows from the sorted array.
- **3.4** − Now, it will assign a class to the test point based on most frequent class of these rows.
- 3.4 − Now, it will appoint a class to the test point dependent on the most successive class of these columns.

Stage 4 − End

Support Vector Machines

Support vector machines (SVMs) are amazing yet adaptable administered machine learning algorithms which are utilized both for classification and regression. SVMs have their one of a kind method for execution when contrasted with other machine learning algorithms. Of late, they are very famous as a result of their capacity to deal with various continuous and categorical variables.

A SVM model is essentially a portrayal of various classes in a hyperplane in multidimensional space. The hyperplane will be created in an iterative way by SVM with the goal that the mistake can be limited. The objective of SVM is to partition the datasets into classes to locate a most extreme peripheral hyperplane

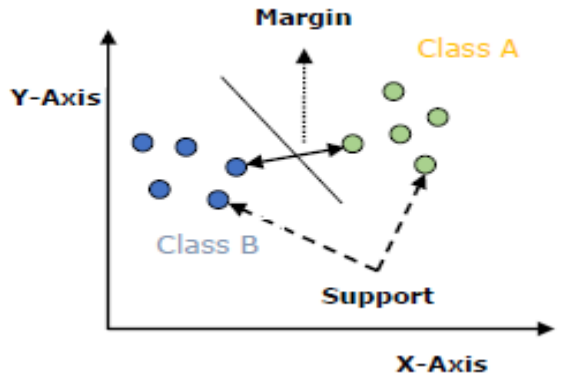

Fig.1. Support Vector Machines

- Support Vectors – Data indicates that are nearest the hyperplane is called support vectors. Isolating line will be characterized with the assistance of these data points .
- Hyperplane − As we can find in the above outline, it is a choice plane or space which is isolated between a lot of articles having various classes.
- Margin − It might be characterized as the gap between two lines on the data points of different classes . It tends to be determined as the opposite good ways from the line to the help support vectors. Huge edge is considered as a decent edge and little edge is considered as a terrible edge.

The fundamental objective of SVM is to separate the datasets into classes to locate a most extreme minor hyperplane (MMH) and it very well may be done in the accompanying two stages −

- First, SVM will produce hyperplanes iteratively that isolates the classes in most ideal manner.
- Then, it will pick the hyperplane that isolates the classes effectively.

Logistic Regression

Linear Regression isn't constantly fitting on the grounds that the data may not fit a straight line yet in addition the straight line esteems can be more prominent than 1 and under 0 .Thus ,they surely can't be utilized as the likelihood of event of the objective class **.**Under these circumstances logistic regression is used . Instead fitting data into straight line logistic regression uses logistic curve.

Simple Logistic Regression

Output = 0 or 1, Hypothesis => Z = WX + B hΘ(x) = sigmoid (Z)

Sigmoid Function

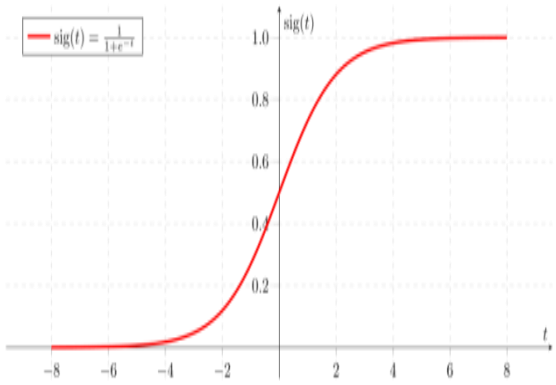

Fig. 2. Sigmoid Activation Function

If 'Z' goes to infinity, Y(predicted) will become 1 and if 'Z' goes to negative infinity, Y(predicted) will become 0.

This type of regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes.

In basic words, the dependent variable is double in nature having information coded as either 1 (represents achievement/yes) or 0 (represents disappointment/no).

Scientifically, a calculated this model predicts P(Y=1) as an element of X. It is one of the

Mathematically, a logistic regression model predicts P(Y=1) as a function of X. It is one of the simplest ML algorithms that can be used for various classification problems .In our example

Sorts of Logistic Regression

For the most part, strategic regression implies twofold calculated regression having paired objective factors, however there can be two additional classes of target factors that can be anticipated by it. In view of that number of classifications, Logistic regression can be separated into following sorts −

Parallel or Binomial

In such a sort of arrangement, a needy variable will have just two potential sorts either 1 and 0. For instance, these factors may speak to progress or disappointment, yes or no, win or misfortune and so on.

Multinomial
In such a sort of arrangement, subordinate variable can have at least 3 potential unordered sorts or the sorts having no quantitative hugeness. For instance, these factors may speak to "Type A" or "Type B" or "Type C".
Ordinal
In such a sort of characterization, subordinate variable can have at least 3 potential arranged sorts or the sorts having a quantitative centrality. For instance, these factors may speak to "poor" or "great", "generally excellent", "Superb" and every classification can have the scores like 0,1,2,3.
Numerically, a strategic relapse model predicts P(Y=1) as a component of X. It is one of the least difficult ML calculations that can be utilized for different characterization issues.
Regression Models

- Binary Logistic Regression Model − The most straightforward type of strategic regression is parallel or binomial calculated regression in which the objective or ward variable can have just 2 potential sorts either 1 or 0.
- Multinomial Logistic Regression Model − another valuable type of calculated regression is multinomial strategic regression in which the objective or ward variable can have at least 3 potential unordered sorts for example the sorts having no quantitative hugeness.

Naive Bayes

The Bayes Rule and Naïve Bayes Classification

The Bayes Rule is a method for going from P(X|Y), known from the preparation dataset, to discover P(Y|X).
What occurs if Y has multiple classes? we process the likelihood of each class of Y and let the most elevated success.
P(X/Y)= P(X ∩ Y)/P(Y) [P(Evidence/Outcome) (Known from Training Data)]
P(Y/X)= P(X ∩ Y)/P(X) [P(Outcome/Evidence) (To be Predicted for Test Data)]
Naïve Bayes calculations are an arrangement method dependent on applying Bayes' hypothesis with a solid supposition that every one of the indicators is autonomous to one another. In basic words, the assumption is that the nearness of a component in a class is autonomous to the nearness of some other element in a similar class
In Bayesian portrayal, the rule interest is to find the back **probabilities** for instance the probability of a name given some watched features, $(L \mid features)$. With the help of Bayes speculation, we can express this in quantitative structure as seeks after −
P(L|features)=P(L)P(features|L)/P(features)
Here, $(L \mid features)$ is the posterior probability of class.
(L) L) is the earlier probability of class.
$(features|L)$ is the likelihood which is the probability of marker given class.
$(features)$ is the earlier probability of pointer.

Random Forest

Random forest is a supervised learning which is utilized for both classifications just as regression. In any case, be that as it may, it is principally utilized for classification issues. As we realize that a forest is comprised of trees and more trees implies progressively robust forest. So also, arbitrary random forest algorithm makes choice trees on data samples and afterward gets the forecast from every one of them lastly chooses the best solution by methods for casting a vote. It is an outfit strategy which is superior to anything a solitary choice tree since it decreases the over-fitting by averaging the outcome.

Random Forest Algorithm

- Step 1 − First, start with the choice of random samples from a given dataset.
- Step 2 − Next, this calculation will build a choice tree for each example. At that point, it will get the forecast outcome from each choice tree.
- Step 3 − In this progression, casting a ballot will be performed for each anticipated outcome.
- Step 4 − At last, select the most casted a ballot forecast result as the final prediction result.

IV. PATIENT DATA SET

The complete of 1090 cases with ten attributes was amassed for the cardiovascular data set from kaggle. The attribute "Cardiovascular" described as the one indicates people having Cardiovascular Disease and Zero Indicates no cardiovascular disease .Table I suggests the attributes values of cardiovascular disease data set .The data set having 539 cardiovascular diseases no cases and 551 cardiovascular yes cases**.**

Table 1: Cardiovascular Data Set

Serial Number	Attribute	Remarks
1	ID	ID number
2	Age	in Days
3	Gender	1-women ,2-Men
4	Height	In Cent Meter
5	Weight	Kilo Grams
6	Systolic Blood Pressure	Systolic Blood Pressure
7	Diastolic Blood Pressure	Diastolic Blood Pressure
8	Cholesterol	1-Normal ,2-Above Normal, 3-Well Above Normal
9	Glue	1-Normal ,2-Above Normal, 3-Well Above Normal
10	Smoke	Whether Patient Smokes or Not
11	Alco	Binary Feature
12	Active	Binary Feature
13	Cardiovascular	Target Variable

V. PROPOSED TECHINQUE

The principle destinations of this examination are to propose a technique that can create best Machine Learning algorithm for prediction of cardiovascular disease. We have considered various machines learning algorithms and their various performance metrics have compared.

A. Selection

We have considered kidney data set from Kaggle .We have considered 13 attributes of cardiovascular data set as stated in section IV .They are 1090 tuples in this data set and this set having 551 yes (having cardiovascular disease)cases and 539 no cases(not having cardiovascular disease) .

B. Pre-processing and Transformation

The cardiovascular dataset is set up in Comma Separated Document format (CSV) from Excel File. Different things required are the expulsion of right qualities for missing records, copy records evacuate pointless information field, standard information position, adjust information in a convenient way and so on. The considered cardiovascular data set do not have any missing data values for different attributes.

C. Performance Evaluation

The performance evaluation of various machine learning algorithms like correctly classified instances, incorrectly classified instances, kappa statistic, Mean absolute error (MAE), Root Mean square error (RMSE),Relative Absolute Error, Root Relative Square Error are to be discussed. We are about to do calculation of True positive rate, False positive rate Precision, Recall, F-Measure and confusion matrix of various considered machine learning algorithms.

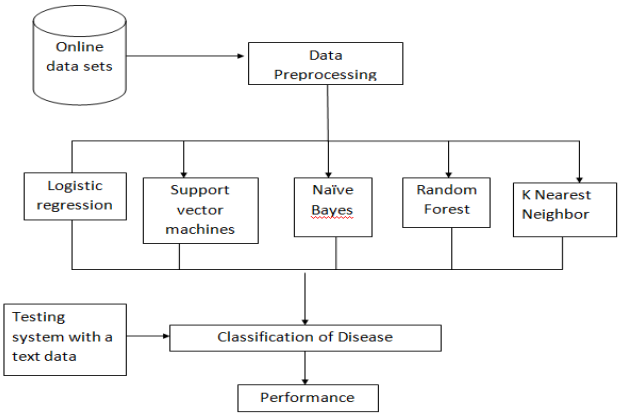

Fig .3. Architecture diagram of cardiovascular disease prediction system

VI. PERFORMANCE MEASURES FOR CLASSIFICATION

One can use following execution measures for the request and figure of imperfection slanted module as shown by his/her own need. Confusion Matrix: The confusion matrix is used to measure the introduction
of two class issue for the given instructive record. The right corner to corner parts TP (True positive) and TN (True Negative) adequately describe Instances similarly as FP (false positive) and FN (false negative) wrongly request Instances. Confusion Matrix Correctly Classify Instance TP+TN Incorrectly Classify Instances.

- True positives imply the positive cardiovascular tuples that were precisely named by the classifier,
- True negatives are the negative cardiovascular tuples that were precisely set apart by the classifier.
- False positives are the negative cardiovascular tuples that were erroneously set apart as positive tuples
- False negatives are the positive cardiovascular tuples that were incorrectly stamped negative tuples
- A confusion matrix for positive and negative tuples is as follows

Table II: Components of Confusion Matrix

		Predicted Class: Yes	Predicted Class: No	
Actual Class	Yes	True Positives(TP)	False Negatives(FN)	P
	No	False Positives(FP)	True Negatives(TN)	N
		P Complement	N Complement	P+N

A confusion matrix for positive and negative cardiovascular tuples for the considered data set is as follows

Table III: Confusion Matrix of Various Algorithms

Name of the algorithm	Classes	Cardiovascular disease = yes	Cardiovascular disease = no
K-Nearest Neighbour	Cardiovascular disease = yes	33	24
	Cardiovascular disease = no	21	31
	Total	54	55
Support Vector Machines	Cardiovascular disease = yes	38	19
	Cardiovascular disease = no	12	40
	Total	50	59
Logistic Regression	Cardiovascular disease = yes	42	15
	Cardiovascular disease = no	10	42
	Total	52	57
Naive Bayes	Cardiovascular disease= yes	51	6
	Cardiovascular disease = no	27	25

	Total	78	31
Random Forest	Cardiovascular disease = yes	40	17
	Cardiovascular disease = no	11	41
	Total	51	58

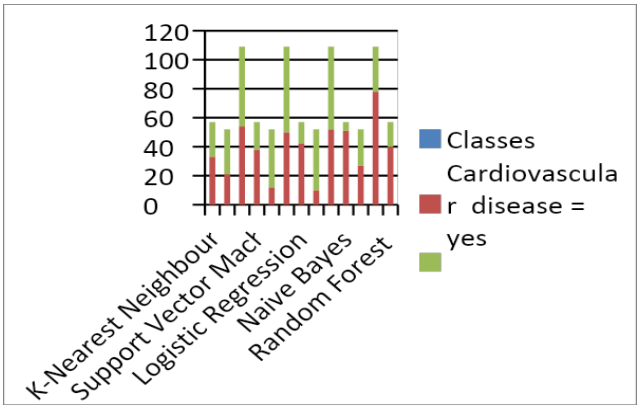

Fig. 4. Graphical Presentation of various algorithms

The accuracy of a classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier. That is,

Table IV: Various Measurements Formula

Measure	Formula
Accuracy, Recognition Rate	TP+TN / P+N
Error ,Misclassification Rate	FP+FN / P+N
Sensitivity, True Positive rate, Recall	TP / P
Specificity, True Negative Rate	TN / N
Precision	TP / TP+FP
F, F1, F-score, Harmonic mean of precision and recall	2* Precision* Recall / Precision +Recall

Table V: Results of Precision, Recall, F1-Score for various algorithms with cardiovascular data set

Name of the algorithm	Averages	Precision	Recall	F1-Score	Support
K-Nearest Neighbour		0.61	0.58	0.59	57
		0.56	0.60	0.58	52
	Micro Average	0.59	0.59	0.59	109
	Macro Average	0.59	0.59	0.59	109
	Weighted Average	0.59	0.59	0.59	109
Support Vector Machines		0.76	0.67	0.71	57
		0.68	0.77	0.72	52
	Micro Average	0.72	0.72	0.72	109
	Macro Average	0.72	0.72	0.72	109
	Weighted Average	0.72	0.72	0.72	109
Logistic Regression		0.81	0.74	0.77	57
		0.74	0.81	0.77	52
	Micro Average	0.77	0.77	0.77	109
	Macro Average	0.77	0.77	0.77	109
	Weighted Average	0.77	0.77	0.77	109
Naive Bayes		0.65	0.89	0.76	57
		0.81	0.48	0.60	52
	Micro Average	0.70	0.70	0.70	109
	Macro Average	0.73	0.69	0.68	109
	Weighted Average	0.73	0.70	0.68	109
Random Forest		0.78	0.70	0.74	57
		0.71	0.71	0.75	52
	Micro Average	0.74	0.74	0.74	109
		0.75	0.75	0.75	109
		0.75	0.74	0.74	109

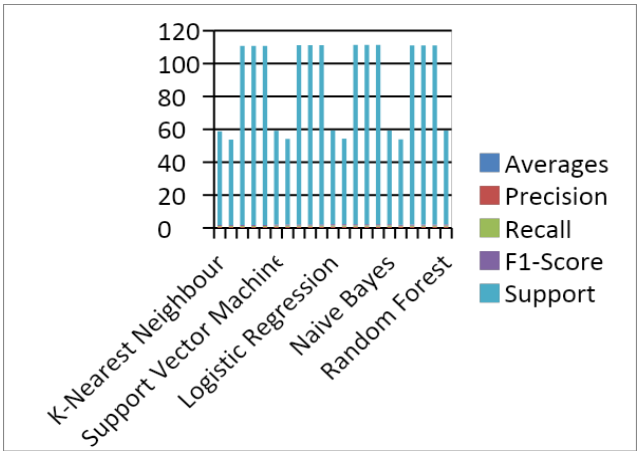

Fig. 5. Comparison of Micro, Macro, and Weighted Average of various algorithms

Table V1: Accuracy Measure for Cardiovascular Disease Dataset

Name of the Algorithm	Correctly Classified instances (%)	Incorrectly Classified instances (%)
K-Nearest Neighbour	58.71	41.28
Support Vector Machines	71.55	28.44
Logistic Regression	77.06	22.93
Naive Bayes	69.72	30.27
Random Forest	74.31	25.68

Table V1I: Accuracy Measure for Cardiovascular Disease Dataset

Name of the Algorithm	Kappa Statistics	Mean Absolute Error
K-Nearest Neighbour	0.17	0.41
Support Vector Machines	0.43	0.28
Logistic Regression	0.54	0.22
Naïve Bayes	0.38	0.55
Random Forest	0.48	0.25

Table VIII: Accuracy Measure for Cardiovascular Disease Dataset

Name of the Algorithm	Root Mean Squared Error	Relative Absolute Error (%)	Root Relative Square Error(%)
K-Nearest Neighbour	0.64	82.74	58.46
Support Vector Machines	0.53	51.48	40.27
Logistic Regression	0.47	45.96	38.47
Naive Bayes	0.55	60.67	42.87
Random Forest	0.50	51.48	36.37

A. Correctly and Incorrectly Classified Instances: Correctly classified instances mean the sum of True Positives and True Negatives of cardiovascular data set tuples. Similarly, incorrectly classified instances means the sum of false positive and False Negatives of cardiovascular data sets. The total number of correctly cardiovascular data instances divided by total number of cardiovascular data instances gives the accuracy.

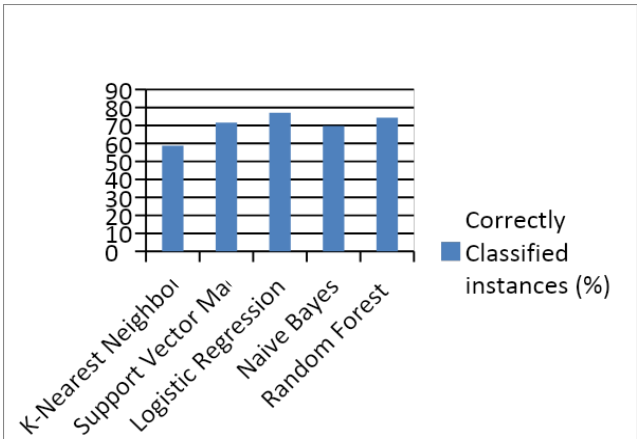

Fig.6. Comparison of correctly classified Instances for various algorithms

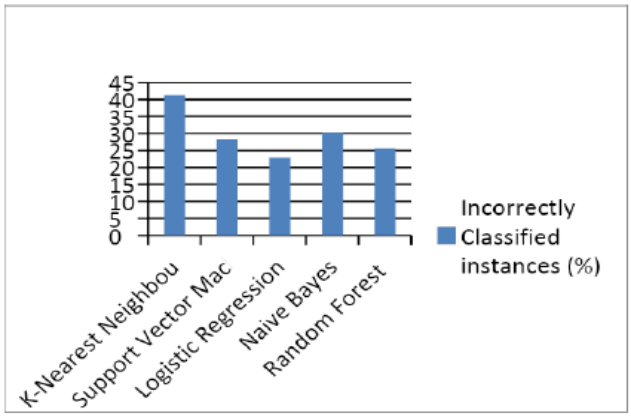

Fig.7. Comparison of incorrectly classified Instances for various algorithms

B. Kappa Statistics

Kappa Statistic: The kappa measurement is a proportion of how intently the cardiovascular data instances characterized by the machine learning classifier coordinated the cardiovascular data named as ground truth, controlling for the exactness of an irregular classifier as estimated by the normal precision.

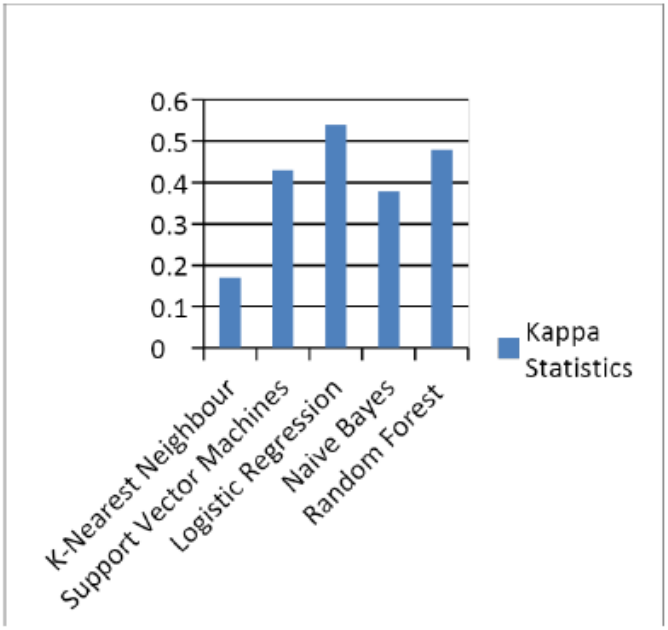

Fig.8. Comparison of various algorithms for Kappa Statistics

C. Mean Absolute Error

Mathematical representation of mean absolute error (MAE) is the mean cardiovascular test instances of the absolute difference between predicted and actual results.

$$MAE = \frac{1}{N}\sum_{j=1}^{n}|yi - y'i|$$

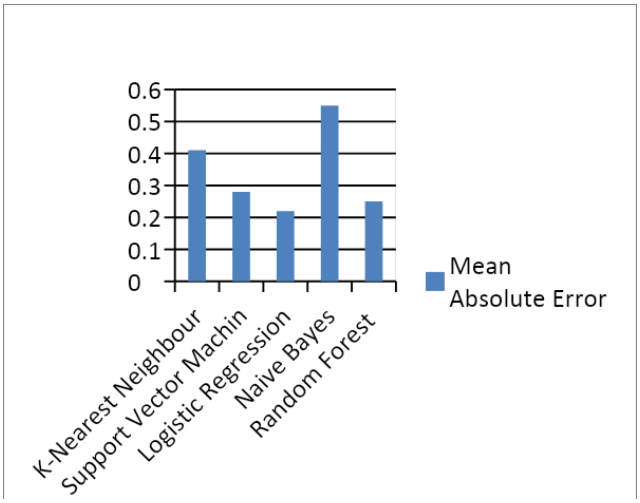

Fig.9. Comparison of Mean Absolute Error of various algorithms

D. Root Mean Squared Error

The size of root mean squared error (RMSE) is determined and It's the square base of the normal of squared contrasts among anticipated and genuine outcomes.

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(yi - y'i)}$$

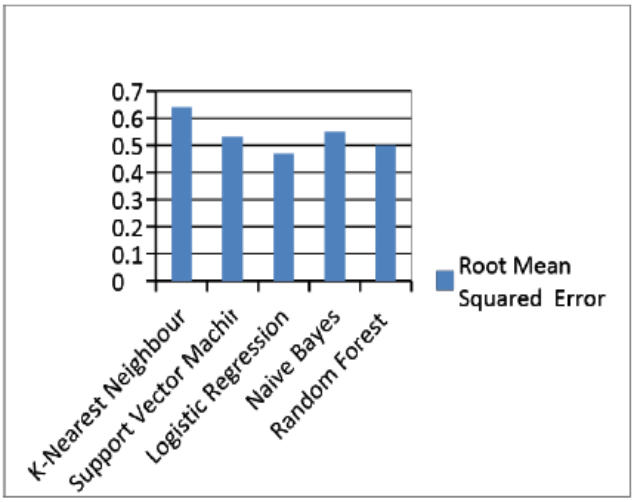

Fig.10. Comparison of Root Mean Squared Error of various Algorithms

E. Root Absolute Error

.It is the root of Absolute Error. It is one of the important performances Measure for machine learning algorithms.

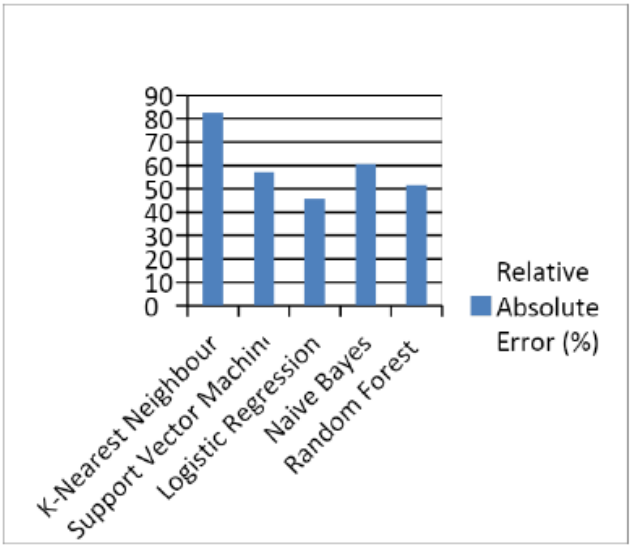

Fig.11. Comparison of Relative Absolute Error for various Algorithms

F. Root Relative Squared Error It is the root of relative squared Error. It is also one of the important performances Measure for machine learning algorithms.

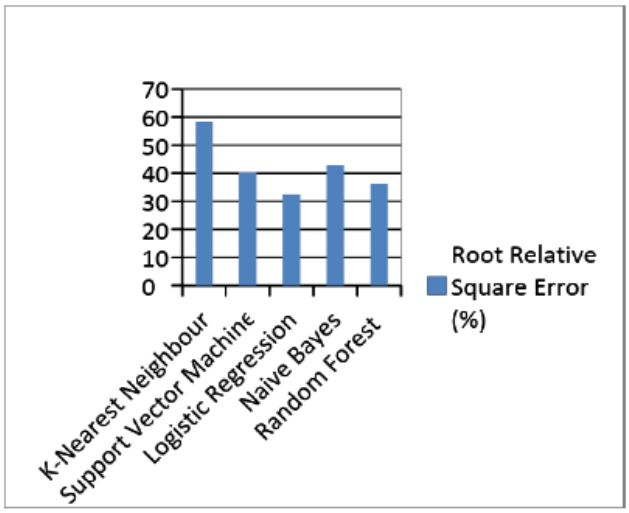

Fig.12. Comparison of Root Relative Square Error for various Algorithms

VII. DISCUSSION AND RESULTS

In this assessment, we applied Machine Learning Algorithms on cardiovascular disease dataset to foresee patients who have interminable cardiovascular ailment, and the individuals who are not debilitated, in light of the information of each characteristic for every patient. Our objective was to think about various arrangement models and characterize the most productive one. Our examination was made based on five calculations positioned among the K-Nearest Neighbour, Support Vector Machines, Logistic Regression, Naive Bayes, Random Forest. From the tables 6,7,8 ,we have the following results among the five stated comparison algorithms

Table 9: Accuracy Measure for Cardiovascular Disease Dataset

Name of the Algorithm	Correctly Classified instances (%)	Incorrectly Classified instances (%)
Logistic Regression	77.06	22.93

Table 10: Accuracy Measure for Cardiovascular Disease Dataset

Name of the Algorithm	Kappa Statistics	Mean Absolute Error
Logistic Regression	0.54	0.22

Table 11: Accuracy Measure for Cardiovascular Disease Dataset

Name of the Algorithm	Root Mean Squared Error	Relative Absolute Error (%)	Root Relative Square Error(%)
Logistic Regression	0.47	45.96	32.47

Logistic Regression has highest number of correctly classified instances that is 77.06 and it has lees number of in correctly classified instances that is 22.93 when compared to remaining four algorithms

Concerning estimation of indicators, the estimations of Mean total error(MAE), Root Mean Square error(RMSE), Relative Absolute error(RAE), Root relative square error (RRSR) demonstrated that Logistic Regression indicators scored the most reduced qualities (MAE = 0.22) (RMSE = 0.47, RAE =45.96%, RRSE = 32.47) trailed by different calculations .

VIII. CONCLUSION

As end, the use of information digging systems for prescient examination is significant in the wellbeing field since it enables us to confront ailments prior and accordingly spare individuals' lives through the expectation of fix. In this work, we utilized a few learning calculation K-Nearest Neighbour, Support Vector Machines, Logistic Regression, Naive Bayes, Random Forest to foresee patients with constant cardiovascular disappointment infection, and patients who are not experiencing this illness. Re-enactment results demonstrated that Logistic Regression classifier demonstrated its exhibition in foreseeing with best outcomes regarding precision and least execution time.

REFERENCES

1. "The Atlas of Heart Disease and Stroke",[online].http://www.who.int/cardiovascular_diseases/resources/atlas/en/
2. J. S. Rumsfeld, K. E. Joynt, and T. M. Maddox, "Enormous information examination to improve cardiovascular consideration: guarantee and difficulties", Nature Reviews Cardiology, Vol.13, No.6, pp.350, 2016.
3. W. Dai, T. S. Brisimi, W. G. Adams, T. Mela, V. Saligrama, and I. C. Paschalidis, "Forecast of hospitalization because of heart sicknesses by managed learning techniques", International Journal of Medical Informatics, Vol.84, No.3, pp.189–197, 2015.
4. A.U.Haq,J.P.Li,M.H.H.Memon,S. Nazir,R.sun "A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms", Mobile Information Systems, Volume 2018.
5. S.Krishnan J. Geetha S, "Forecast of Heart Disease Using Machine Learning Algorithms",First International Conference on Innovations in Information and Communication Technology ,2019.
6. K.G.Dinesh, K A.garaj, K.D.Santhosh, V M.eswari. "Forecast of Cardiovascular Disease Utilizing Machine Learning Algorithms", 2018 International Conference on Current Trends towards Converging Technologies ,2018
7. A.Golande, P.kumar T, "Coronary illness Prediction Using Effective Machine Learning Techniques", International Journal of Recent Technology and Engineering, ISSN: 2277-3878, Volume-8, Issue-1S4, June 2019.
8. Prasad,P.Anjali, S.Adil, N.Deepa, "Coronary illness Prediction utilizing Logistic Regression Algorithm utilizing Machine Learning", International Journal of Engineering and Advanced Technology, ISSN: 2249 – 8958, Volume-8, Issue-3S, February 2019
9. 9].Y. Khourdifi, M. Bahaj , "Coronary illness Prediction and Classification Using Machine Learning Algorithms Optimized by Particle Swarm Optimization and Ant Colony Optimization", International Journal of Intelligent Engineering and Systems, Vol.12,No.1,2019.

AUTHOR PROFILE

Dr. M. Srivenkatesh, working as Associate Professor, Department of Computer Science, GITAM Deemed to be University, Visakhapatnam, Andhra Pradesh, India .He has published Eleven International Journal Papers. His research interest includes Data mining, Machine Learning, Software Engineering, Cloud Computing and Rough Sets. Nine Research Scholars are working for their PhD in Computer Science under his guidance.