Prediction of Breast Cancer Disease using Machine Learning Algorithms

Muktevi Srivenkatesh

Abstract: Background/Aim: Breast Cancer is the most often identified cancer among women and major reason for increasing mortality rate among women. ***The early strategies for estimating the breast cancer 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 breast cancer data set from kaggle and we have done pre-processing tasks for missing values .We have no missing data values from the considered data set .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 breast cancer 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 breast cancer disease. Results: The machine learning algorithms under study were able to predict breast cancer disease in patients with accuracy between 52.63% and 98.24%. Conclusions: It was shown that Random Forest has better Accuracy (98.24 %) when compared to different Machine-learning Algorithms.***

Keywords: Breast Cancer, 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 Breast Cancer 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 breast cancer malady or not having breast cancer ailment that is obscure.

Machine Learning examines how computers can learn (or improve their exhibition) in view of Breast Cancer 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 Breast Cancer data.

Supervised learning is fundamentally an equivalent word for arrangement.

The supervision in the taking in originates from the named models in the Breast Cancer data collection.

Breast Cancer is very important health issue in women needs to have very much need to take care. There has been a lot of research on cancer diagnosis by using machine learning techniques. We have tests breast cancer which

Includes breast exam, Mammogram Breast and ultrasound Biopsy. As an alternative we can also use Machine Learning techniques for the classification of benign and malignant tumours. The prior diagnosis of Breast Cancer can enhance the prediction and survival rate notably [1], so that patients can be informed to take clinical treatment at the right time.

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.

Breast Cancer and its Symptoms

Different people have different symptoms of breast cancer. Some people do not have any signs or symptoms at all.

Some warning signs of breast cancer are—

- New lump in the breast or underarm (armpit).
- Thickening or swelling of part of the breast.
- Irritation or dimpling of breast skin.
- Redness or flaky skin in the nipple area or the breast.
- Pulling in of the nipple or pain in the nipple area.
- Nipple discharge other than breast milk, including blood.
- Any change in the size or the shape of the breast.
- Pain in any area of the breast.

II. LITERATURE SURVEY

Hiba Asri,Hajar Mousannif,Hassan Al Moatassime,Thomas Noel [2] has discussed breast cancer analysis with different machine learning algorithms: Support Vector Machine (SVM), DecisionTree (C4.5), Naive Bayes (NB) and k Nearest Neighbors (k-NN) on the Wisconsin Breast Cancer (original) datasets is conducted and their performance was compared. Their Experimental results show that SVM gives the highest accuracy (97.13%) with lowest error rate.

Revised Manuscript Received on February 28, 2020.
* Correspondence Author

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

They have executed within a simulation environment and conducted in WEKA data mining tool.

Habib Dhahri,Eslam Al Maghayreh, Awais Mahmood,Wail Elkilani,and Mohammed Faisal Nagi[3] has done study is based on genetic programming and machine learning algorithms that aim to construct a system to accurately differentiate between benign and malignant breast tumors. The aim of their study was to optimize the learning algorithm.

They have applied the genetic programming technique to select the best features and perfect parameter values of the machine learning classifiers. The performance of the proposed method was based on sensitivity, specificity, precision, accuracy, and the roc curves. The have presented study and proves that genetic programming can automatically find the best model by combining feature preprocessing methods and classifier algorithms.

Chang Ming, Valeria Viassolo, Nicole Probst-Hensch, Pierre O Chappuis, Ivo D. Dinov & Maria C. Katapodi[4], their study was to compare the discriminatory accuracy of ML-based estimates against a pair of established methods—the Breast Cancer Risk Assessment Tool (BCRAT) and Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm (BOADICEA) models. Their Results showed that their Predictive accuracy reached 90.17% using ML-adaptive boosting and 89.32% using ML-Markov chain Monte Carlo generalized linear mixed model versus 59.31% with BOADICEA for the Swiss clinic-based sample.

Ch. Shravya, K. Pravalika, Shaik Subhani[5] has studied on the implementation of models using Logistic Regression, Support Vector Machine (SVM) and K Nearest Neighbor (KNN) is done on the breast cancer dataset taken from the UCI repository. With respect to their results of accuracy, precision, sensitivity, specificity and False Positive Rate the efficiency of each algorithm was measured and compared. Their experimental results have shown that SVM is the best for predictive analysis with an accuracy of 92.7%.They infer from our study that SVM is the well suited algorithm for prediction and on the whole KNN presented well next to SVM.

Mandeep Rana, Pooja Chandorkar, Alishiba Dsouza , Nikahat Kazi [6] their aim is to classify whether the breast cancer is benign or malignant and predict the recurrence and non-recurrence of malignant cases after a certain period. They have used machine learning techniques such as Support Vector Machine, Logistic Regression, KNN and Naive Bayes. These techniques are coded in MATLAB using UCI machine learning depository. They have compared the accuracies of different techniques and observed the results. They have found SVM most suited for predictive analysis and KNN performed best for our overall methodology.

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 Machine Learning 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 a 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.

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.

A. 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.

B. 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.

SSSSWhen 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.

C. 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.

D. 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

1. 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.

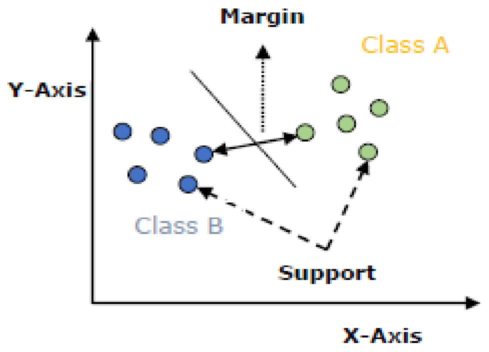
- **3.2** − Now, based on the distance value, sort them in ascending order.
- **3.3** − Next, it will choose the top K rows from the sorted array.
- **3.4** − Now, it will assign a class to the test point based on most frequent class of these rows.
- 3.4 − Now, it will appoint a class to the test point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

Stage 4 − End

2. 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 hyper plane in multidimensional space. The hyper plane 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 hyper plane



.

- ➢ Support Vectors – Data indicates that are nearest the hyper plane is called support vectors. Isolating line will be characterized with the assistance of these data points .
- ➢ Hyper plane − 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 hyper plane (MMH) and it very well may be done in the accompanying two stages –

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

3. 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

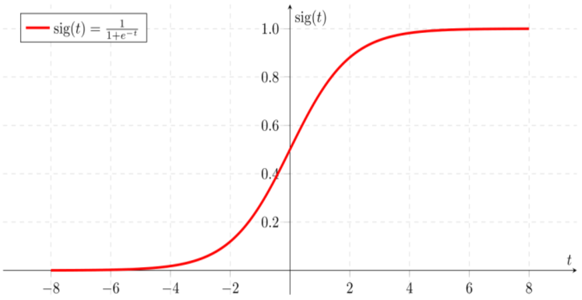


Figure 2: Sigmoid Activation Function

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.

4. Naive Bayes

1. 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.

5. 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 569 cases with 32 attributes was amassed for the Breast Cancer data set from kaggle. The attribute "diagnosis" described as the measurable and zero indicates patients are not having Breast cancer(B=Benign) and one indicates patients are having Breast Cancer (M = Malignant).Table I suggests the attributes values of Breast Cancer data set .The data set having consists of 569 tuples out of which 357 no breast cancer (B=Benign) cases and 212 breast cancer yes cases .

Table 1: Breast Cancer Data Set

Serial Number	**Attribute**	
1	ID	Identification Number
2	Diagnosis	The diagnosis of breast tissues (M = malignant, B = benign)
3	Radius_ mean	Mean of distances from center to points on the perimeter
4	Texture_ mean	Standard deviation of gray-scale values
5	Perimeter_ mean	Mean size of the core tumor
6	Area_mean	Area Mean
7	Smoothness_mean	Mean of local variation in radius lengths
8	Compactness_ mean	Mean of perimeter^2 / area - 1.0
9	Concavity_ mean	Mean of severity of concave portions of the contour
10	concave points_ mean	Mean for number of concave portions of the contour
11	Symmetry_ mean	symmetry mean
12	Fractal_dimension_mean	mean for "coastline approximation" - 1
13	Radius_se	standard error for the mean of distances from center to points on the perimeter
14	Texture_se	standard error for standard deviation of gray-scale values
15	Perimeter_se	perimeter_se
16	Area_se	Area_se
17	Smoothness_se	standard error for local variation in radius lengths
18	Compactness_se	standard error for perimeter^2 / area - 1.0

19	Concavity_se	standard error for severity of concave portions of the contour
20	concave points_se	standard error for number of concave portions of the contour
21	symmetry_se	symmetry_se
22	Fractal_dimension_se	standard error for "coastline approximation" - 1
23	Radius_worst	"worst" or largest mean value for mean of distances from center to points on the perimeter
24	Texture_worst	"worst" or largest mean value for standard deviation of gray-scale values
25	Perimeter_worst	Perimeter_worst
26	Area_worst	Area_worst
27	smoothness_worst	"worst" or largest mean value for local variation in radius lengths
28	Compactness_worst	"worst" or largest mean value for perimeter^2 / area - 1.0
29	Concavity_worst	"worst" or largest mean value for severity of concave portions of the contour
30	concave points_worst	"worst" or largest mean value for number of concave portions of the contour
31	symmetry worst	symmetry worst
32	fractal_dimension_worst	"worst" or largest mean value for "coastline approximation" - 1

V. PROPOSED TECHINQUE

The principle destinations of this examination are to propose a technique that can create best Machine Learning algorithm for prediction of Breast Cancer disease.

We have considered various machines learning algorithms and their various performance metrics have compared.

1. Selection

We have considered Breast Cancer data set from Kaggle .We have considered 32 attributes of Breast cancer data set as stated in section IV . The data set having consists of 569 tuples out of which 357 no breast cancer (B=Benign) cases and 212 breast cancer (M= Malignant) yes cases .

2. Pre-processing and Transformation

The breast cancer 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 breast cancer data set does not have any missing values .

3. 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.

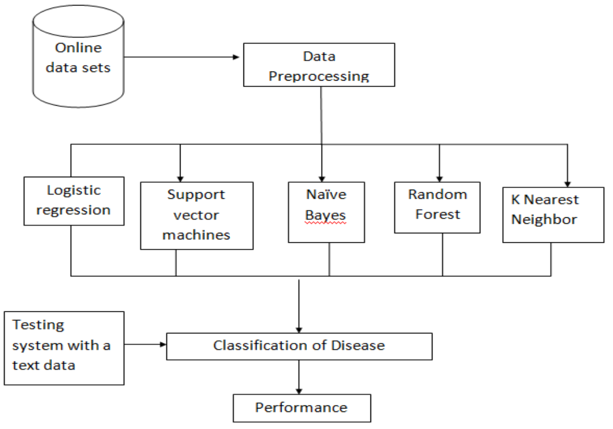


Figure 3: Architecture diagram of Breast Cancer with Machine Learning Algorithms

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 breast cancer tuples that were precisely named by the classifier,
- True negatives are the negative breast cancer tuples that were precisely set apart by the classifier.
- False positives are the negative breast cancer tuples that were erroneously set apart as positive tuples
- False negatives are the positive breast cancer tuples that were incorrectly stamped negative tuples

- A confusion matrix for positive and negative tuples is as follows

Predicted Class

Table 2: Components of Confusion Matrix

		Yes	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 breast cancer tuples for the considered data set is as follows

Table 3: Confusion Matrix of Various Algorithms

Name of the algorithm	Classes	Breast Cancer = yes	Breast Cancer = no
K-Nearest Neighbour	Breast Cancer = yes Breast Cancer =no	27 15	5 10
	Total	42	15
Support Vector Machines	Breast Cancer = yes Breast Cancer =no	50 13	6 8
	Total	63	14
Logistic Regression	Breast Cancer = yes Breast Cancer =no	2 0	30 25
	Total	2	55
Naive Bayes	Breast Cancer = yes Breast Cancer =no	30 25	2 0
	Total	55	2
Random Forest	Breast Cancer = yes Breast Cancer =no	32 1	0 24
	Total	33	24

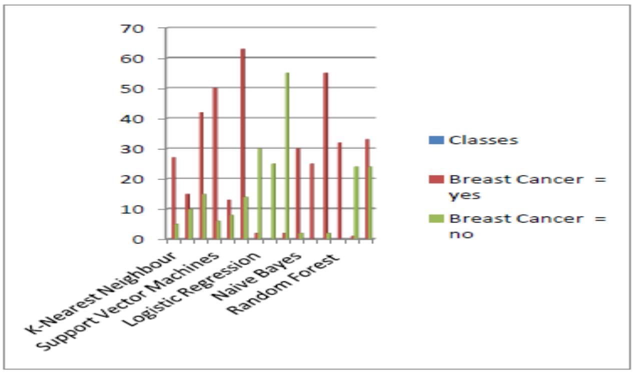


Figure 4: Graphical Presentation of various algorithms

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 4: 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 5: Results of Precision, Recall, F1-Score for various algorithms with breast cancer data set

Name of the algorithm	Averages	Precision	Recall	F1-Score	Support
K-Nearest Neighbour		0.64	0.84	0.73	32
		0.67	0.40	0.50	25
	Micro Average	0.65	0.65	0.65	57
	Macro Average	0.67	0.62	0.61	57
	Weighted Average	0.65	0.65	0.63	57
Support Vector Machines		0.79	0.89	0.84	56
		0.57	0.38	0.46	21
	Micro Average	0.75	0.75	0.75	77
	Macro Average	0.68	0.64	0.65	77
	Weighted Average	0.73	0.75	0.74	77
Logistic Regression		1.00	0.06	0.12	32
		0.45	1.00	0.62	25
	Micro Average	0.47	0.47	0.47	57
	Macro Average	0.73	0.53	0.37	57
	Weighted Average	0.76	0.47	0.34	57
Naive Bayes		0.55	0.94	0.69	32
		0.00	0.00	0.00	25
	Micro Average	0.53	0.53	0.53	57
	Macro Average	0.27	0.47	0.34	57
	Weighted Average	0.31	0.53	0.39	57
Random Forest		0.97	1.00	0.98	32
		1.00	0.96	0.98	25
	Micro Average	0.98	0.98	0.98	57
	Macro Average	0.98	0.98	0.98	57
	Weighted Average	0.98	0.98	0.98	57

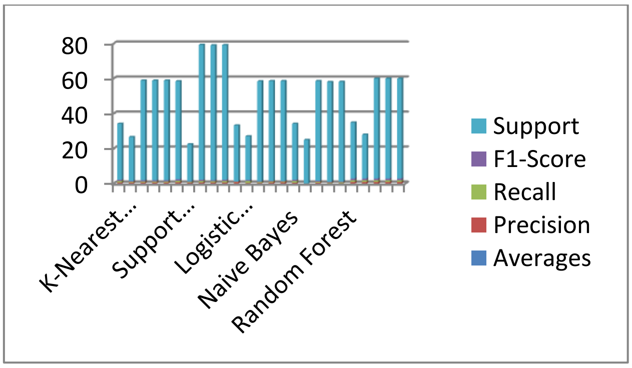


Figure 5: Comparison of Micro, Macro, aSSnd Weighted Average of various algorithms

Table 6: Accuracy Measure for Breast Cancer Dataset

Name of the Algorithm	Correctly Classified instances (%)	Incorrectly Classified instances (%)
K-Nearest Neighbour	64.0	35.08
Support Vector Machines	75.32	24.67
Logistic Regression	47.36	52.63
Naive Bayes	52.63	47.36
Random Forest	98.24	1.75

Table 7: Accuracy Measure for Breast Cancer Dataset

Name of the Algorithm	Kappa Statistics	Mean Absolute Error
K-Nearest Neighbour	0.25	0.35
Support Vector Machines	0.30	0.24
Logistic Regression	0.05	0.52
Naive Bayes	-0.06	0.47
Random Forest	0.96	0.01

Table 8: Accuracy Measure for Breast Cancer Dataset

Name of the Algorithm	Root Mean Squared Error	Relative Absolute Error (%)	Root Relative Square Error(%)
K-Nearest Neighbour	0.59	71.25	50.09
Support Vector Machines	0.49	62.49	40.29
Logistic Regression	0.72	106.87	75.14
Naive Bayes	0.68	96.18	67.62
Random Forest	0.13	3.56	2.50

A. Correctly and Incorrectly Classified Instances:

Correctly classified instances mean the sum of True Positives and True Negatives of breast cancer data set tuples. Similarly, incorrectly classified instances means the sum of false positive and False Negatives of breast cancer data sets. The total number of correctly breast cancer data instances divided by total number of breast cancer data instances gives the accuracy.

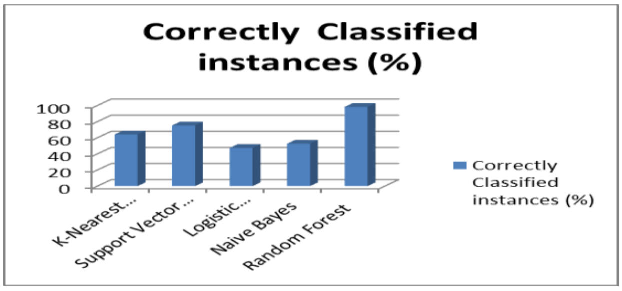


Figure 6: Comparison of correctly classified instances of various algorithms

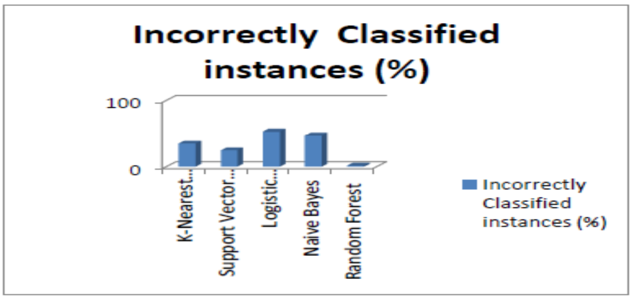


Figure 7: Comparison of Incorrectly Classified Instances of various Algorithms

B. Kappa Statistics

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

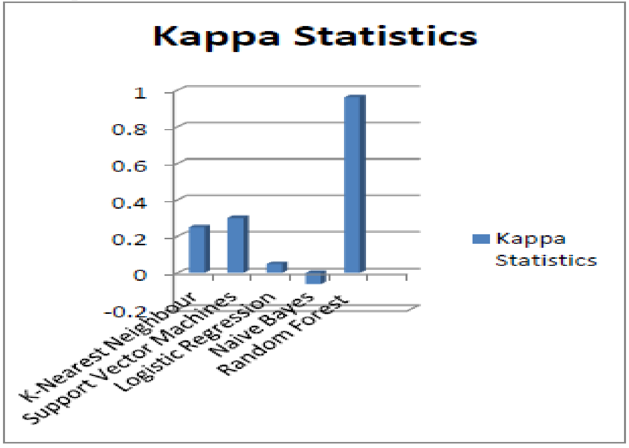


Figure 6: Comparison of kappa statistic of various

C. Mean Absolute Error

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

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

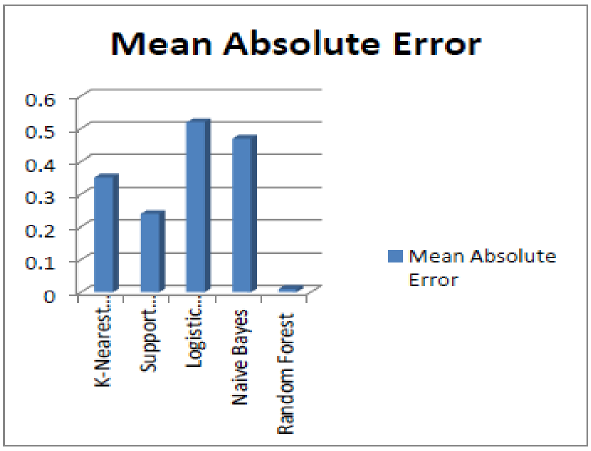


Figure 7: Comparison of Mean Absolute Error of various algorithms

Retrieval Number: D1866029420/2020©BEIESP
DOI: 10.35940/ijitee.D1866.029420
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

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.

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

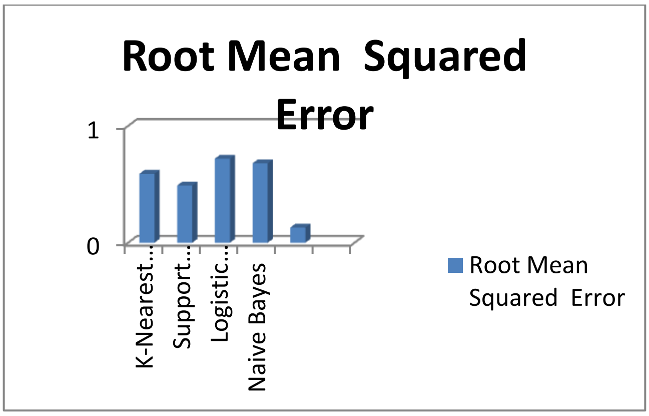


Figure 8: Comparison of Root Mean Squared Error of various algoritham

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

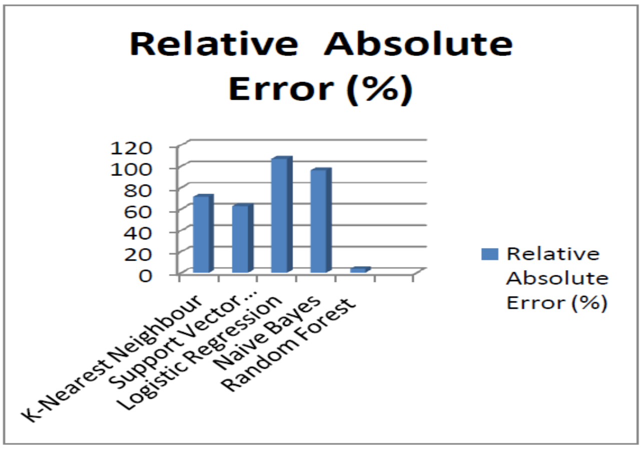


Figure 9: Comparison of Relative Absolute Error

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

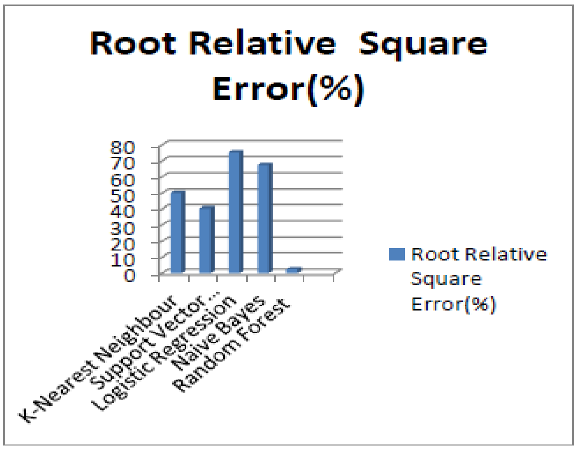


Figure 10: Comparison of Root Relative Square Error of various Algorithms

VII. DISCUSSION AND RESULTS

In this assessment, we applied Machine Learning Algorithms on Breast Cancer data set to foresee patients who have interminable breast cancer 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 above tables 6,7,8 ,We have showed that Random Forest have highest accuracy when compared to remaining algorithms .

Table 9: Accuracy Measure for Breast Cancer Dataset

Name of the Algorithm	Correctly Classified instances (%)	Incorrectly Classified instances (%)
Random Forest	98.24	1.75

Table 10: Accuracy Measure for Breast Cancer Dataset

Name of the Algorithm	Kappa Statistics	Mean Absolute Error
Random Forest	0.96	0.01

Table 11: Accuracy Measure for Breast Cancer Dataset

Name of the Algorithm	Root Mean Squared Error	Relative Absolute Error (%)	Root Relative Square Error(%)
Random Forest	0.31	3.56	2.50

Random Forest has highest number of correctly classified instances that is 98.24% and it has lees number of in correctly classified instances that is 1.75% and 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 Random Forest indicators scored the most reduced qualities (MAE = 0.01) (RMSE = 0.31, RAE =3.56%, RRSE=2.50%) 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 breast cancer disappointment infection, and patients who are not experiencing this illness. Re-enactment results demonstrated that and Random Forest classifier demonstrated its exhibition in foreseeing with best outcomes regarding precision and least execution time.

REFERENCES

1. Yi-Sheng Sun, Zhao Zhao, Han-Ping-Zhu,"Risk factors and Preventions of Breast Cancer" International Journal of Biological Sciences.

2. Hiba Asri,Hajar Mousannif,Hassan Al Moatassime,Thomas Noel, Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis, The 6th International Symposium on Frontiers in Ambient and Mobile Systems (FAMS 2016), Procedia Computer Science 83 (2016) 1064 – 1069.
3. Habib Dhahri,Eslam Al Maghayreh, Awais Mahmood,Wail Elkilani,and Mohammed Faisal Nagi, Automated Breast Cancer Diagnosis Based on Machine Learning Algorithms, Journal of Health Care Engineering , Volume 2019 .
4. Chang Ming, Valeria Viassolo, Nicole Probst-Hensch, Pierre O Chappuis, Ivo D. Dinov & Maria C. Katapodi, Machine learning techniques for personalized breast cancer risk prediction: comparison with the BCRAT and BOADICEA models, Breast Cancer Research volume 21, Article number: 75 (2019).
5. Ch. Shravya, K. Pravalika, Shaik Subhani,Prediction of Breast Cancer Using Supervised Machine Learning Techniques, International Journal of Innovative Technology and Exploring Engineering , Volume-8 Issue-6, April 2019.
6. MandeepRana, PoojaChandorkar, AlishibaDsouza, "Breast cancer diagnosis and recurrence prediction using machine learning techniques", International Journal of Research in Engineering and Technology Volume 04, Issue 04, April 2015.

AUTHORS PROFILE

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