International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.7 March - April – 2023
DOI:10.46647/ijetms.2023.v07i02.017 ISSN: 2581-4621

Early Detection Of Breast Cancer Using Logistic Regression Method

Saheb karan[1] , Abhik Roy Chowdhury[2] , Amit Pal[3] , Susmita Das[4] , Mrs. Sulekha Das[5] , Avijit Kumar Chaudhuri[6]

[1] UG -Computer Science and Engineering, Techno Engineering College Banipur
[2] UG - Computer Science and Engineering, Techno Engineering College Banipur
[3] UG - Computer Science and Engineering, Techno Engineering College Banipur
[4] UG - Computer Science and Engineering, Techno Engineering College Banipur
[5]Assistant Professor, Department of CSE, TEC Banipur.
[6]Assistant Professor, Department of CSE, TEC Banipur.
Corresponding Author Orcid ID :- 0000-0002-0067-5994

❖ ABSTRACT :-

Breast cancer is the most frequently occurring cancer disease in women. It is reported almost 14% of cancers in Indian women are breast cancer. It becomes very crucial to predict breast cancer earlier to minimize the deaths. This research article helps to predict breast cancer earlier and reduce the immature deaths of women in India. In this paper, the authors have used the Logistic Regression method to classify the disease.

The authors simulate the results using logistic regression with 10-fold cross-validations and with a different train-test split of the dataset. The 10-fold cross validations display its potential with almost 94% performance in the research paper. With all features and 90-10 , 80-20,50-50, 66-34 splits, and 10-fold cross-validations the authors achieve 96% accuracy.

we have used different accuracy measures like accuracy, sensitivity, specificity, and kappa statistics to get the novelty of the model.

In this study, the authors use the Wisconsin (Diagnostic) Data Set for Breast Cancer, Created by Dr. William H. Wolberg, General Surgery Dept., University of Wisconsin, Clinical Science Centre, Madison, WI 53792 wolberg@eagle.surgery.wisc.edu available at the UCI ML Repository website.

Keywords—Machine Learning, Logistic Regression, Breast Cancer.

❖ INTRODUCTION :-

Breast cancer is considered a multifactorial disease and the most common cancer in women worldwide [1 , 2] with approximately 30% of all female cancers [3 , 4] (i.e. 1.5 million women are diagnosed with breast cancer each year, and 500,000 women die from this disease in the world). Over the past 30 years, this disease has increased, while the death rate has decreased. However, the reduction in mortality due to mammography screening is estimated at 20% and improvement in cancer treatment is estimated at 60% [5 , 6].
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9175124/#:~:text=The%20proposed%20machine%2Dlearning%20approaches,interventions%20at%20the%20right%20time.

This paper was constructed on Machine learning (ML) algorithms to examine the dataset of 569 cases with breast cancer and thereby explain the results. ML is a subset of artificial intelligence (AI) that is utilize to classify data based on models which have been developed and for predictive analytics, in particular breast cancer. It provides tools via which huge amount of data can be automatically analyzed. In the case of the present study, we utilized ML algorithms and collected a scientific dataset of breast cancer cases from surgery wisc edu . (wolberg@eagle.surgery.wisc.edu)and clarify these data based on various features. Ten (10) real-valued features including: [1] radius (mean of distances from center to points on the perimeter), [2]

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.7 March - April – 2023
DOI:10.46647/ijetms.2023.v07i02.017 ISSN: 2581-4621

❖ LITERATURE REVIEW:-

❖ Machine learning techniques can be beneficial of predicting risk at early stage of breast cancer. For predicting this disease, researchers use different classifiers: - DT,LR, GA,NN, KNM etc.

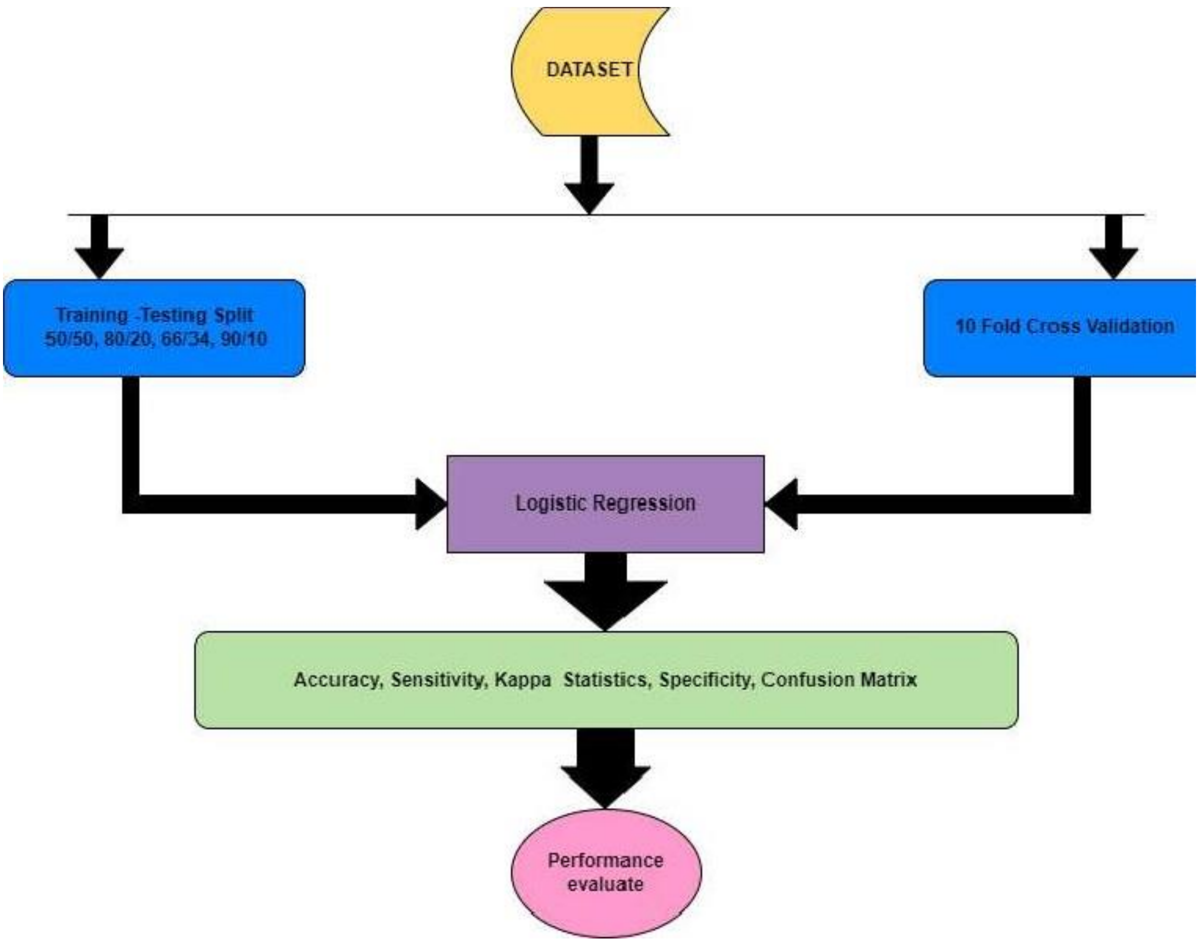
- The fuzzy laws which were used by Keles et al. (2011) [24] and created a method, achieved 97% accuracy.
- Kim et al. (2012) [25] used the SVM technique using BC dataset having 679 records that include clinical, and pathological data types. Here the accuracy was 99 %.

- Kharya et al. (2014) [26] developed a probabilistic method for forecasting BC utilizing Naive Bayes Classifiers. This paper includes 65.5 % of stable cases and 34.5 % of malignant cases. The method showed a precision of 93 %.

- Lavanya and Rani (2012) [27] organized data on the BC. This approach is based on CART and bagging schemes. Pre-processing which was used to improve the collection of features and efficiency and it showed the improved accuracy of the classification.

- Kumar et al. (2013) [28] used a dataset containing 699 patient studies in their research paper, and the training constitutes 499 records and 200 for testing. Here, 241 or 34.5 % had BCs, while the 458 or 65.5 %. were non-cancerous. Here applying NB and SVM algorithm. Here achieved the accuracy of 94.5%

➢ ***Architectural Design :-***



International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.7 March - April – 2023
DOI:10.46647/ijetms.2023.v07i02.017 ISSN: 2581-4621

❖ METHODOLOGY :-

➢ ***Research Method*** : As mentioned earlier, we have used Logistic Regression (LR), a statistical technique for regression analysis. Our first work was to find the independent variables which were making impact on the single dependent variable. Now as we have found the independent variables, namely- Radius_mean, Texture_mean, Perimeter_mean, Area_mean,, Smoothness_mean, Compactness_mean, Concavity_mean, Cancave points_mean, Symmentry_mean Fractal_dimension, mean , and so on and the dependent variable, namely- Outcome (y). We now construct a stepwise logistic relation between them.

➢ ***Description of dataset :-***

The author collect the dataset from *wolberg@eagle.surgery.wisc.edu*. No missing values are there in the dataset. The dataset contains data of 569 patients where 212 patients are suffering from Breast Cancer and rest are not effected in Breast Cancer. There are 30 features each of which the author consider as independent variables and one Outcome which is count as dependent variable.

Sl no.	Attribute	Description	Mean
1	Radius mean	Average distance between perimeter points and center	14.127291739894561
2	Texture mean	Gray scale's (Magnitude) standard deviation	19.289648506151185
3	Perimeter mean	Size(average) of the central tumor	91.96903339191564
4	Area mean		654.8891036906856
5	Smoothness mean	Mean Local variation in radius length	0.09636028119507901
6	Compactness mean	((Average of perimeter)^2/area)-1.0	0.1043409841827768
7	Concavity mean	Mean Severity of concave part of the contour	0.08879931581722325
8	Concave points mean	The mean number of concasectionsion of the contour	0.048919145869947236
9	Symmetry mean		0.18116186291739902
10	Fractal dimension mean	Mean of 'coastline approximation' - 1	0.06279760984182771
11	Radius se	The standard error for the mean length from center to perimeter	0.4051720562390162
12	Texture se	The standard error for the standard deviation of gray scale	1.2168534270650262
13	Perimeter se		2.866059226713528
14	Area se		40.337079086116034
15	Smoothness se	Standard fault for local difference in radius length	0.007040978910369072
16	Compactness se	Standard error for the ((perimeter)^2/area)-1.0	0.025478138840070295
17	Concavity se	Normal error for the severity of concave section of the contour	0.03189371634446394

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.7 March - April – 2023
DOI:10.46647/ijetms.2023.v07i02.017 ISSN: 2581-4621

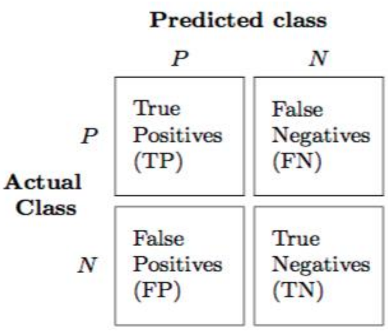
18	Concave points se	The standard error is the number of the concave part of the contour	0.011796137082601058
19	Symmetry se		0.02054229876977152
20	Fractal dimension se	Ordinary error for of ‘coastline approximation’ - 1	0.0037949038664323383
21	Radius worst	“worst” or greater mean value for the mean distance between middle and perimeter point	16.269189806678387
22	Texture worst	“worst” or largest mean value for standard deviation of gray scale	25.67722319859401
23	Perimeter worst		107.26121265377863
24	Area worst		880.5831282952546
25	Smoothness worst	“worst” or biggest mean value for local radius length differences	0.1323685940246047
26	Compactness worst	“worst” or largest mean value for ((perimeter)^2/area)-1.0	0.25426504393673127
27	Concavity worst	“worst” or greater mean value for the severity of concave portion of the contour	0.2721884833040421
28	Concave points worst	“worst” or largest mean value for number of concave portion of the contour	0.11460622319859404
29	Symmetry worst		0.2900755711775047
30	Fractal dimension worst	“worst” or biggest mean value for of ‘coastline approximation’ - 1	0.08394581722319859

As we are moving forward toward our final model, a few steps need to be followed in LR

STEP – 1 : *Check 1* :-

1) Cross-Validation or one can say out-of-sample testing, is a method where we test and train various parts of the data individually and calculate the accuracy of the model in practice. Here we divided the dataset into 10 paths, each time we select a part out of the 10 as the testing data and the remaining part as training parts.

2) Confusion matrix is also unknown as an error matrix in a table that shows the overall performance of an algorithm or a clarification model. In the field of statistical analysis, a confusion matrix shows a set of test data for which the values are true or not.



[P=With Cancer
N=Without Cancer]

Fig 1 : Overview of a confusion matrix

DOI:10.46647/ijetms.2023.v07i02.017 ISSN: 2581-4621

3) Accuracy must be calculated for our model if means how precisely or how close the measured value reflects the originals.
4) Specificity must be calculated. It refers to the test accuracy at identifying the probability of a negative test, provided the condition is absent.
5) Sensitivity refers to the test accuracy in identifying the probability of a positive test, provided the condition is present.
6) Kappa is the ratio of the proportion of times the raters agree(adjusted for agreement by chance) to the maximum proportion of times the raters could have agreed (adjusted for agreement by chance)

STEP 3 :- *Selecting the suitable method*

We first develop the stepwise logistic relations between the dependent & independent variables then we split the data set into four fractions as 90\10, 80\20 , 66\34, 50\50 as the train test splitting followed by the 10-fold cross validations method.

STEP 4 *:- Developing Equation of LR and Confusion Matrix* :-

The logistic regression Model :- The logistic regression [7] is fairly a generalization of a binary model. In general, logistic regression model is used to find the probability of an existing class such as yes or no based on the observation of a dataset.

A) It can be defined as a classification problem, where the output or target variable (y) is dependent on the given values or inputs (X) in a dataset
The model of logistic regression can be represented as :-

$$yf = \frac{e^{(b0+b1*x1+b2*x2+\cdots..........bn*xn)}}{1+e^{(b0+b1*x1+b2*x2+\cdots..........+bn*xn)}}$$

b0 = y'- (b1*X1'+b2*X2'+b3*X3'+.....bn*Xn')
Where ,
e=Exponential constant
yf = Predicted outcome
b0 = bias or intercept term
b_1= coefficient of the first controlled variable
b_2= coefficient of the second controlled variable
b_3 = coefficient of the third controlled variable
b_4 = coefficient of the fourth controlled variable and so on
x_1 = radius_mean
x_2 =texture_mean
x_3 = perimeter_mean
x_4 =area_mean and so on
In the case of b_1, $\bar{x}$ is the mean of radius. In the case of b_2, $\bar{x}$ is the mean of texture.. In the case of b_3, $\bar{x}$ is the mean of perimeter. In the case of b_4, $\bar{x}$ is the mean of area , and so on we find mean vale .

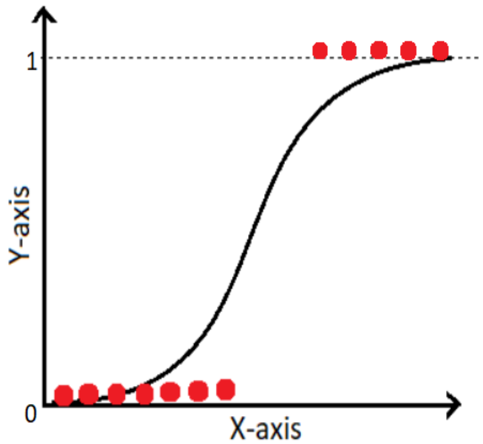


Fig2: Logistic regression graph

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.7 March - April – 2023
DOI:10.46647/ijetms.2023.v07i02.017 **ISSN: 2581-4621**

.B***) The confusion Matrix :-***
Now let's take,
TP= TRUE POSITIVE
TN= TRUE NEGATIVE
FP= FALSE POSITIVE
FN= FALSE NEGATIVE
Now,

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

$$\text{Sensitivity} = \frac{TP}{TP+FN}$$

$$\text{Specificity} = \frac{TN}{TN+FP}$$

$$Po = \frac{TP+TN}{TP+TN+FP+FN}$$ [po = relative observed agreement among raters]

$$Pe = \frac{((TP+FN)*(TP+FP)+(FP+TN)*(FN+TN))}{(TP+TN+FP+FN)^2}$$ [pe= the hypothetical probability of chance agreement.]

$$\text{Kappa} = \frac{(po-pe)}{(1-pe)}$$

❖ RESULTS AND DISCUSSION :-

After analysing this model . we get the results that are given below .

Table:- Accuracy of difference between Actual data and Calculated data

Accuracy of 90%Data as Training Data or(0.9)	92.85714285714286
Accuracy of 80%Data as Training Data or(0.8)	95.57522123893806
Accuracy of 66%Data as Training Data or(0.66)	94.79166666666666
Accuracy of 50%Data as Training Data or(0.5)	95.75971731448763

Table:- Confusion Matrix & Corresponding Result

For 90% of data	**For 80% of data**
Confusion Matrix:- $\begin{matrix} 4 & 4 \\ 0 & 48 \end{matrix}$ Accuracy:- 0.9285714285714286 Sensitivity:-1.0 Specificity:- 0.9230769230769231 Kappa:- 0.6315789473684212	Confusion Matrix:- $\begin{matrix} 8 & 5 \\ 0 & 100 \end{matrix}$ Accuracy:- 0.9557522123893806 Sensitivity:-1.0 Specificity:- 0.9523809523809523 Kappa:- 0.7390300230946885
For 66% of data	**For 50% of data**
Confusion Matrix:- $\begin{matrix} 32 & 81 \\ 2 & 150 \end{matrix}$ Accuracy:- 0.9479166666666666 Sensitivity:- 0.9411764705882353 Specificity:- 0.9493670886075949 Kappa:- 0.8328690807799441	Confusion Matrix:- $\begin{matrix} 70 & 7 \\ 5 & 201 \end{matrix}$ Accuracy:- 0.9575971731448764 Sensitivity:- 0.9333333333333333 Specificity:- 0.9663461538461539 Kappa:- 0.8920739846183182

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.7 March - April – 2023
DOI:10.46647/ijetms.2023.v07i02.017 ISSN: 2581-4621

Table:- 10-fold cross-validation Accuracy

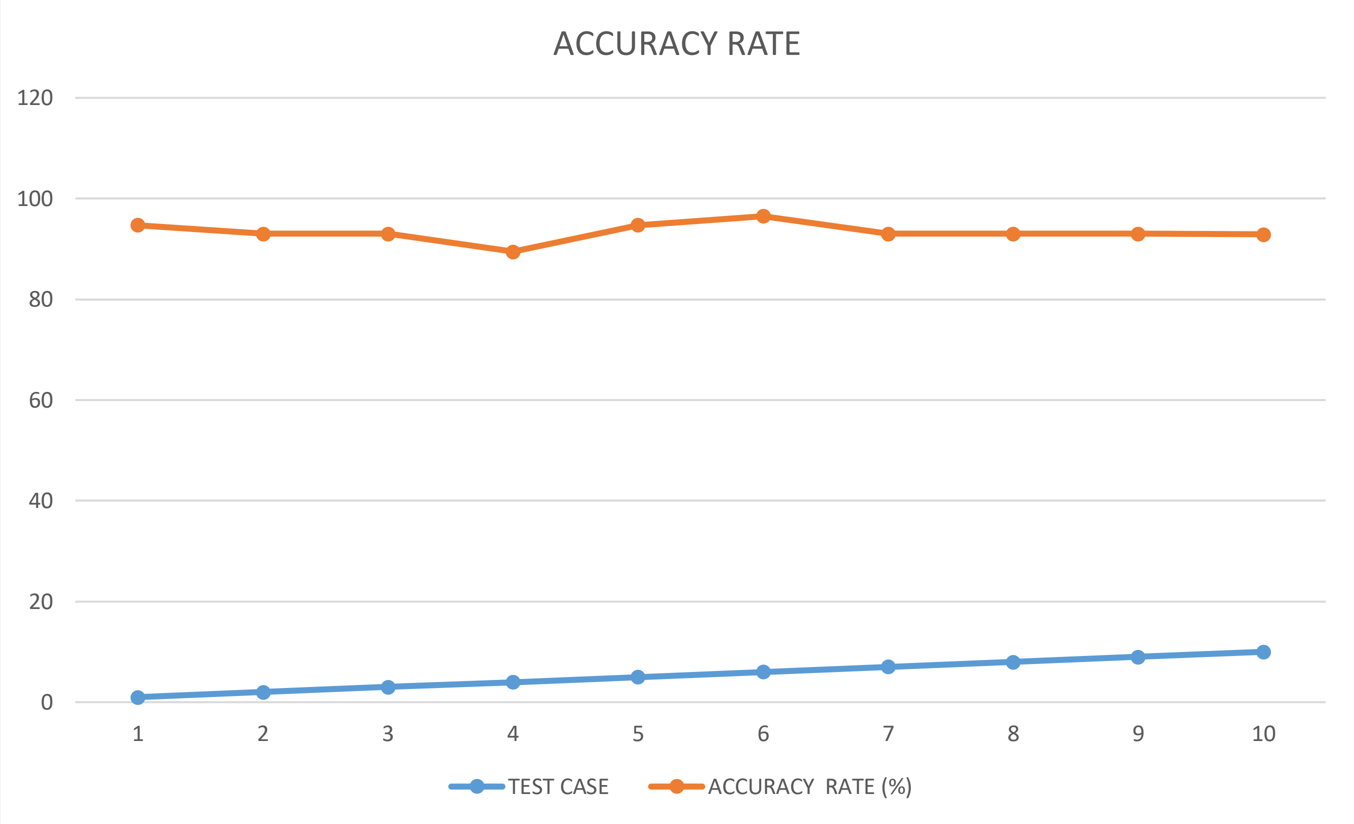
TEST CASE	ACCURACY RATE (%)
1	94.73684210526315
2	92.98245614035088
3	92.98245614035088
4	89.47368421052632
5	94.73684210526315
6	96.49122807017544
7	92.98245614035088
8	92.98245614035088.
9	92.98245614035088.
10	92.85714285714286

❖ **Table:- 10-fold cross-validation Results**

0-56 Test Data	**57-113 Test Data**
Confusion Matrix:- 25 3 0 29 Accuracy :- 0.9473684210526315 Sensitivity :- 1.0 Specificity :- 0.90625 Kappa :- 0.8945095619987661	Confusion Matrix:- 23 2 2 30 Accuracy :- 0.9298245614035088 Sensitivity :- 0.92 Specificity :- 0.9375 Kappa :- 0.8575
114-170 Test Data	**171-227 Test Data**
Confusion Matrix: - 25 4 0 28 Accuracy :- 0.9298245614035088 Sensitivity :- 1.0 Specificity :- 0.875 Kappa :- 0. 85995085995086	Confusion Matrix:- 27 4 2 24 Accuracy :- 0.8947368421052632 Sensitivity :- 0.9310344827586207 Specificity :- 0.8571428571428571 Kappa :- 0.7891491985203453
228-284 Test Data	**285-341 Test Data**
Confusion Matrix:- 31 2 1 23 Accuracy :- 0.9473684210526315 Sensitivity :- 0.96875 Specificity :- 0.92 Kappa :- 0.95426230907073	Confusion Matrix: 31 1 1 24 Accuracy :- 0.9649122807017544 Sensitivity :- 0.96875 Specificity :- 0.96 Kappa :- 0.92875
342-398 Test Data	**399-455Test Data**
Confusion Matrix:- 30 2 2 23 Accuracy :- 0.9298245614035088 Sensitivity :- 0.9375 Specificity :- 0.92 Kappa :- 0.8575	Confusion Matrix:- 4 4 0 49 Accuracy :- 0.9298245614035088. Sensitivity :- 1.0 Specificity :- 0.9245283018867925 Kappa :- 0.632258064516129
455-512 Test Data	**513-569 Test Data**
Confusion Matrix:- 4 4 0 49 Accuracy :- 0.9298245614035088. Sensitivity :- 1.0 Specificity :- 0.9245283018867925 Kappa :- 0.632258064516129	Confusion Matrix: 4 4 0 48 Accuracy :- 0.9285714285714286 Sensitivity :- 1.0 Specificity :- 0.9230769230769231 Kappa :- 0.6315789473684212

DOI:10.46647/ijetms.2023.v07i02.017 ISSN: 2581-4621

❖ ***10-fold cross-validation graph :-***



CONCLUSION :-

In this paper , Logistic regression (LR) statistical technical has been used to develop a breast cancer predictor. The overall data has been divided into two paths referred as train-test following up with 10-fold cross-validation and developing the confusion matrix. The recoeded accuracy for the 90/10 , 80/20, 66/34 and ,50/50 train-test-split are 92.85%, 95.57% , 94.79% , 95.75% respectively . This model in proposed to predict the breast cancer results of this data base. we made a relationship between the dependent variable and the independent variable after that we perform a confusion matrix where we compare the actual target values with those predicted by the machine learning model. After checking the confusion matrix we move to the Cross Validation where we find the accuracy of 10 sub-list elements and we also find the Confusion Matrix of each Sub-list. we predict the accuracy as well as sensitivity, and specificity for user choice test data and the 10 sub-list. This type of project may help in the future to find any kind of prediction from any data field.

REFERENCES :-

[1] M. U. Sarwar, M. K. Hanif, R. Talib, A. Mobeen, and M. Aslam, "A survey of big data analytics in healthcare," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, pp. 355-359, 2017.
[2] S. Maldonado, J. López, A. Jimenez -Molina, and H. Lira, "Simultaneous feature selection and heterogeneity control for SVM classification: An application to mental workload assessment," *Expert Syst. Appl.*, vol. 143, pp. 112988, 2020.
[3] T. Sridevi and A. Murugan, "A novel feature selection method for effective breast cancer diagnosis and prognosis," *Int. J.Comput. Appl.*, vol. 88, 2014.
[4] L. Breiman, "Random forests," *Mach. Learn*, vol. 45, pp. 5-32, 2001.
[5] G. Cavallaro, M. Riedel, M. Richerzhagen, J. A. Benediktsson, and A. Plaza, "On understanding big data impacts in remotely sensed image classification using support vector machine methods," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 8, pp. 4634-4646, 2015.

DOI:10.46647/ijetms.2023.v07i02.017 ISSN: 2581-4621

[6] D. J. Hand, H. Mannila, and P. Smyth, *Principles of Data Mining: Adaptive Computation and Machine Learning*. ISBN:026208290X, 2001.
[7] H. L. Chen, B. Yang, J. Liu, and D. Y. Liu, "A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis," *Expert Syst. Appl.*, vol. 38, pp. 9014-9022, 2011.
[8] B. Yuan and X. Ma, "Sampling+ reweighting: Boosting the performance of AdaBoost on imbalanced datasets," in *The 2012 international joint conference on neural networks (IJCNN)*, IEEE, June 2012, pp. 1-6.
[9] B. Yuan and X. Ma, "Sampling+ reweighting: Boosting the performance of AdaBoost on imbalanced datasets," in *The 2012 international joint conference on neural networks (IJCNN)*, IEEE, June 2012, pp. 1-6.
[10] W. Kim, K. S. Kim, J. E. Lee, D. Y. Noh, S. W. Kim, Y. S. Jung, and R. W. Park, "Development of novel breast cancer recurrence prediction model using support vector machine," *J. Breast Canc.*, vol. 15, pp. 230-238, 2012.
[11] M. U. Sarwar, M. K. Hanif, R. Talib, A. Mobeen, and M. Aslam, "A survey of big data analytics in healthcare," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, pp. 355-359, 2017.
[12] L. Breiman, "Random forests," *Mach. Learn*, vol. 45, pp. 5-32, 2001.
[13] P. Hamsagayathri and P. Sampath, "Performance analysis of breast cancer classification using decision tree classifiers," *Int. J. Curr. Pharm. Res.*, vol. 9, pp. 19-25, 2017.
[14] B. Xue, M. Zhang, W. N. Browne, and X. Yao, "A survey on evolutionary computation approaches to feature selection," *IEEE Trans. Evol. Comput.*, vol. 20, pp. 606-626, 2015.
[15] B. Bhasuran, G. Murugesan, S. Abdulkadhar, and J. Natarajan, "Stacked ensemble combined with fuzzy matching for biomedical named entity recognition of diseases," *J Biomed. Informat.*, vol. 64, pp. 1-9, 2016.
[16] A. K. Chaudhuri, D. Sinha, and K. S. Thyagaraj, "Identification of the recurrence of breast cancer by discriminant analysis," in *Emerging technologies in data mining and information security*, Singapore: Springer, 2019, pp. 519-532.
[17] S. L. Ang, H. C. Ong, and H. C. Low, "Classification Using the General Bayesian Network," *Pertanika J. Sci. Technol.*, vol. 24, 2016.
[18] S. Kumari and M. Arumugam, "Application of bio-inspired krill herd algorithm for breast cancer classification and diagnosis, "*Indian J. Sci. Technol.*, vol. 8, pp. 30, 2015.
[19] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *ICML*, vol. 96, July 1996, pp. 148-156.
[20] R. Sikora, "A modified stacking ensemble machine learning algorithm using genetic algorithms," in *Handbook of research on organizational transformations through big data analytics*, IGI Global, 2015, pp. 43-53.
[21] S. K. Trivedi and S. Dey, "A study of ensemble based evolutionary classifiers for detecting unsolicited emails, in October 2014, pp. 46-51,in *Proceedings of the 2014 conference on research in adaptive and convergent systems*.
[22] In 1992, pp. 196-202,F. Wilcoxon, "Individual comparisons by ranking methods," in *Breakthroughs in statistics*, New York: Springer.
[23] K. Sivakami and N. Saraswathi, "Mining big data: breast cancer prediction using DT-SVM hybrid model," *Int. J. Sci. Eng.Appl. Sci.*, vol. 1, pp. 418-429, 2015.
[24] A. Keleş, A. Keleş, and U. Yavuz, "Expert system based on neuro-fuzzy rules for diagnosis breast cancer," *Expert Syst. Appl.*,vol. 38, pp. 5719-5726, 2011
[25] W. Kim, K. S. Kim, J. E. Lee, D. Y. Noh, S. W. Kim, Y. S. Jung, and R. W. Park, "Development of novel breast cancer recurrence prediction model using support vector machine," *J. Breast Canc.*, vol. 15, pp. 230-238, 2012.
[26] S. Kharya, S. Agrawal, and S. Soni, "Naive Bayes classifiers: a probabilistic detection model for breast cancer," *Int. J. Comput.Appl.*, vol. 92, pp. 0975-8887, 2014.
[27] S. Kharya, S. Agrawal, and S. Soni, "Naive Bayes classifiers: a probabilistic detection model

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.7 March - April – 2023
DOI:10.46647/ijetms.2023.v07i02.017 ISSN: 2581-4621

for breast cancer," *Int. J. Comput. Appl.*, vol. 92, pp. 0975-8887, 2014
[28] G. R. Kumar, G. A. Ramachandra, and K. Nagamani, "An efficient prediction of breast cancer data using data mining techniques," *Int. J. Innov. Eng. Technol.*, vol. 2, pp. 139, 2013.