Early Detection Of Breast Cancer Using Logistic Regression Method

Saheb karan1, Abhik Roy Chowdhury2, Amit Pal3, Susmita Das4, Mrs. Sulekha Das5, Avijit Kumar Chaudhuri6

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]
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 :-

**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, Symmetry_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 collects 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 affected in Breast Cancer. There are 30 features each of which the author considers as independent variables and one Outcome which is count as dependent variable.

<table>
<thead>
<tr>
<th>Sl no.</th>
<th>Attribute</th>
<th>Description</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Radius mean</td>
<td>Average distance between perimeter points and center</td>
<td>14.127291739894561</td>
</tr>
<tr>
<td>2</td>
<td>Texture mean</td>
<td>Gray scale’s (Magnitude) standard deviation</td>
<td>19.289648506151185</td>
</tr>
<tr>
<td>3</td>
<td>Perimeter mean</td>
<td>Size(average) of the central tumor</td>
<td>91.9690339191564</td>
</tr>
<tr>
<td>4</td>
<td>Area mean</td>
<td></td>
<td>654.8891036906856</td>
</tr>
<tr>
<td>5</td>
<td>Smoothness mean</td>
<td>Mean Local variation in radius length</td>
<td>0.09636028119507901</td>
</tr>
<tr>
<td>6</td>
<td>Compactness mean</td>
<td>((Average of perimeter)^2/area)-1.0</td>
<td>0.1043409841827768</td>
</tr>
<tr>
<td>7</td>
<td>Concavity mean</td>
<td>Mean Severity of concave part of the contour</td>
<td>0.08879931581722325</td>
</tr>
<tr>
<td>8</td>
<td>Concave points mean</td>
<td>The mean number of concasactionsion of the contour</td>
<td>0.04891914586994723</td>
</tr>
<tr>
<td>9</td>
<td>Symmetry mean</td>
<td></td>
<td>0.18116186291739902</td>
</tr>
<tr>
<td>10</td>
<td>Fractal dimension mean</td>
<td>Mean of ‘coastline approximation’ - 1</td>
<td>0.06279760984182771</td>
</tr>
<tr>
<td>11</td>
<td>Radius se</td>
<td>The standard error for the mean length from center to perimeter</td>
<td>0.4051720562390162</td>
</tr>
<tr>
<td>12</td>
<td>Texture se</td>
<td>The standard error for the standard deviation of gray scale</td>
<td>1.2168534270650262</td>
</tr>
<tr>
<td>13</td>
<td>Perimeter se</td>
<td></td>
<td>2.866059226713528</td>
</tr>
<tr>
<td>14</td>
<td>Area se</td>
<td></td>
<td>40.337079086116034</td>
</tr>
<tr>
<td>15</td>
<td>Smoothness se</td>
<td>Standard fault for local difference in radius length</td>
<td>0.00704097891036907</td>
</tr>
<tr>
<td>16</td>
<td>Compactness se</td>
<td>Standard error for the ((perimeter)^2/area)-1.0</td>
<td>0.02547813884007029</td>
</tr>
<tr>
<td>17</td>
<td>Concavity se</td>
<td>Normal error for the severity of concave section of the contour</td>
<td>0.03189371634446394</td>
</tr>
</tbody>
</table>
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.

![Confusion Matrix Diagram](image)

*Fig 1 : Overview of a confusion matrix*
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 :-

\[
y_f = \frac{e^{(b_0 + b_1*x_1 + b_2*x_2 + \ldots + b_n*x_n)}}{1 + e^{(b_0 + b_1*x_1 + b_2*x_2 + \ldots + b_n*x_n)}}
\]

Where ,

e=Exponential constant

\(y_f\) = Predicted outcome

\(b_0\) = 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**
B) The confusion Matrix :-
Now let’s take,
TP= TRUE POSITIVE
TN= TRUE NEGATIVE
FP= FALSE POSITIVE
FN= FALSE NEGATIVE
Now,
Accuracy = \frac{TP+TN}{TP+TN+FP+FN}

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

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

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

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

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

<table>
<thead>
<tr>
<th>Accuracy of 90% Data as Training Data or(0.9)</th>
<th>92.85714285714286</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy of 80% Data as Training Data or(0.8)</td>
<td>95.57522123893806</td>
</tr>
<tr>
<td>Accuracy of 66% Data as Training Data or(0.66)</td>
<td>94.79166666666666</td>
</tr>
<tr>
<td>Accuracy of 50% Data as Training Data or(0.5)</td>
<td>95.75971731448763</td>
</tr>
</tbody>
</table>

Table:- Confusion Matrix & Corresponding Result

For 90% of data
Confusion Matrix:-

\begin{pmatrix}
4 & 4 \\
0 & 48 \\
\end{pmatrix}

Accuracy:- 0.9285714285714286
Sensitivity:- 1.0
Specificity:- 0.9230769230769231
Kappa:- 0.6315789473684212

For 80% of data
Confusion Matrix:-

\begin{pmatrix}
8 & 5 \\
0 & 100 \\
\end{pmatrix}

Accuracy:- 0.9557522123893806
Sensitivity:- 1.0
Specificity:- 0.9523809523809523
Kappa:- 0.7390300230946885

For 66% of data
Confusion Matrix:-

\begin{pmatrix}
32 & 81 \\
2 & 150 \\
\end{pmatrix}

Accuracy:- 0.9479166666666666
Sensitivity:- 0.9411764705882353
Specificity:- 0.9493670886075949
Kappa:- 0.8328690807799441

For 50% of data
Confusion Matrix:-

\begin{pmatrix}
70 & 7 \\
5 & 201 \\
\end{pmatrix}

Accuracy:- 0.9575971731448764
Sensitivity:- 0.9333333333333333
Specificity:- 0.9663461538461539
Kappa:- 0.8920739846183182
Table: - 10-fold cross-validation Accuracy

<table>
<thead>
<tr>
<th>TEST CASE</th>
<th>ACCURACY RATE (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>94.73684210526315</td>
</tr>
<tr>
<td>2</td>
<td>92.98245614035088</td>
</tr>
<tr>
<td>3</td>
<td>92.98245614035088</td>
</tr>
<tr>
<td>4</td>
<td>89.47368421052632</td>
</tr>
<tr>
<td>5</td>
<td>94.73684210526315</td>
</tr>
<tr>
<td>6</td>
<td>96.49122807017544</td>
</tr>
<tr>
<td>7</td>
<td>92.98245614035088</td>
</tr>
<tr>
<td>8</td>
<td>92.98245614035088</td>
</tr>
<tr>
<td>9</td>
<td>92.98245614035088</td>
</tr>
<tr>
<td>10</td>
<td>92.85714285714286</td>
</tr>
</tbody>
</table>

Table: - 10-fold cross-validation Results

<table>
<thead>
<tr>
<th>TEST CASE</th>
<th>Confusion Matrix</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-56 Test Data</td>
<td>25 3 0 29</td>
<td>0.9473684210526315</td>
<td>1.0</td>
<td>0.90625</td>
<td>0.8945095619987661</td>
</tr>
<tr>
<td>114-170 Test Data</td>
<td>25 4 0 28</td>
<td>0.9298245614035088</td>
<td>1.0</td>
<td>0.875</td>
<td>0.8595095619987661</td>
</tr>
<tr>
<td>228-284 Test Data</td>
<td>31 2 1 23</td>
<td>0.9473684210526315</td>
<td>0.96875</td>
<td>0.92</td>
<td>0.95426230907073</td>
</tr>
<tr>
<td>342-398 Test Data</td>
<td>30 2 2 23</td>
<td>0.9298245614035088</td>
<td>0.9375</td>
<td>0.92</td>
<td>0.85714285714286</td>
</tr>
<tr>
<td>455-512 Test Data</td>
<td>4 4 0 49</td>
<td>0.9298245614035088</td>
<td>1.0</td>
<td>0.9245283018867925</td>
<td>0.632258064516129</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TEST CASE</th>
<th>Confusion Matrix</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>57-113 Test Data</td>
<td>23 2 2 30</td>
<td>0.9298245614035088</td>
<td>0.92</td>
<td>0.9375</td>
<td>0.8575</td>
</tr>
<tr>
<td>171-227 Test Data</td>
<td>27 4 2 24</td>
<td>0.8947368421052632</td>
<td>0.9310344827586207</td>
<td>0.85714285714286</td>
<td>0.7891491985203453</td>
</tr>
<tr>
<td>285-341 Test Data</td>
<td>31 1 1 1</td>
<td>0.9649122807017544</td>
<td>0.96875</td>
<td>0.96</td>
<td>0.92875</td>
</tr>
<tr>
<td>399-455 Test Data</td>
<td>4 4 0 49</td>
<td>0.9298245614035088</td>
<td>1.0</td>
<td>0.9245283018867925</td>
<td>0.632258064516129</td>
</tr>
<tr>
<td>513-569 Test Data</td>
<td>4 4 0 48</td>
<td>0.9285714285714286</td>
<td>1.0</td>
<td>0.9230769230769231</td>
<td>0.6315789473684212</td>
</tr>
</tbody>
</table>

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 recorded 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 is proposed to predict the breast cancer results of this database. 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 :-


