Predictive Analytics in Healthcare

Manas Thapliyal ¹, Dr. Saikat Gochhait ²

Symbiosis Institute of Digital and Telecom Management, constituent of Symbiosis International (Deemed University)

I. Abstract and Introduction:

Breast pathology is one of the most prevalent pathologies in routine practice. The malignant and benign lesions that can lead to morbidity and their masquerade as malignancy, which can be a significant public health threat or concern and the patients' plight. The high incidence of malignant breast cancer, its relatively simple early detection and effective preservative surgery and chemotherapy treatment. Because of this triple assessment involving a clinical, radiological and cytological examination, it started to be generally recognized. The field of healthcare nowadays is also determined by the quality of analysis done during diagnostic tests. It the correctness of the diagnostics tests that matter before treating a patient. Around 2.1 million women are affected each year by breast cancer alone. An estimate of death of around 627k women due to breast cancer in the year 2018. As per experts’ belief, 31% of total breast cancer cases are misdiagnosed. If the analysis of the diagnostic tests is accurate, the patient can be treated for the ailment that he is suffering from and the medicine can be specific and precise too. Including analytics for even the smallest of the tests in healthcare would not only help doctors analyze the data from the relevant tests but also make an accurate diagnosis and in some cases prognosis for the ailment which the patient is suffering or might suffering in the near future.

II. Problem

The female breast cancer in India is as high as 25.8 per 1,00,000 and the death rate is 12.7 per 1,00,000 female. Breast cancer projection in India for 2020 shows the number to reach as high as 17,97,900. The accuracy of the tests needs to improve so that there the chances of the misdiagnosed cases of breast cancer reduces.

III. Literature review:

We also saw a comparison of fine-needle aspiration to core biopsy for breast lesion diagnosis (Mitra and Dey, 2016).

Review of the screening prerequisites for reducing cancer death rates (Mitra and Dey, 2016).

Triple assessment has already gained prominence for breast cancer and the role of fine needle aspiration cytology in triple assessment is significant (Ogbuanya, Anyanwu, Iyare and Nwigwe, 2020). Usage of AI for prediction of breast cancer, by creating a new deep learning model that can predict cancer from a mammogram graph whether a patient is likely to develop cancer 5 years in the future (Simons and Gordon, 2019).

Breast cancer evaluation method compared with the Breast and Ovarian Cancer Incidence Analysis and Carrier Estimation Algorithm models (Ming, Viassolo, Probst-Hensch, Chappuis,

IV. Analysis

The data into consideration has a sample size of 569 patients/instances and the contents were of the patient tested from Breast Cancer, diagnosis showing Malignant or Benign. They are uniquely identified by the ID assigned to them. The parameters are the radius, texture, perimeter, area, smoothness, compactness, concavity, symmetry and fractal dimension. Considering these parameters and comparing the accuracy of logistic regression, KNN Classification, Decision Tree, Discriminant analysis for checking the tumor is malignant or benign. This would help hospitals in accurately analyzing the mass, which is considered for diagnostic tests.

Data analysis is done in IBM SPSS Statistics 23.

Data or Information Description: This set of data or information contains a total of 569 instances. The patient's ID, radius (average distances from the perimeter points to the center point, smoothness (local variance in radius length), texture (S.D. of gray scale values), perimeter, field, concavity (severe concave contour portions), compactness, concave points (no concave contour portions), symmetry, the fractal mass dimension are the attributes. Mean, largest, or "worst" and standard error were calculated from each image, resulting in a total of 30 features. Field 4, for example, is mean texture, field 14 is SE texture, field 24 is worst texture. The feature values are recorded to 4 decimal places.

The dimension of the data: 569 rows x 32 columns.
Missing Attribute values: None
The distribution of class: Benign 357, Malignant 212.

<table>
<thead>
<tr>
<th>Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>N Valid</td>
</tr>
<tr>
<td>Missing</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
</tr>
<tr>
<td>Benign (0)</td>
</tr>
<tr>
<td>Malignant (1)</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>

30 Features are a lot of data when the dataset is huge in size and it would take more computing power to analyze if the number of cases increases. Hence, we have done Principal Component Analysis of those 30 features or variables.
The method of extraction used for this case is Principal Component Analysis. Factors from 7 to 30 have Eigenvalues less than one and hence can be ignored. In the above, table information only up to component 7 is shown and 8 to 30 have not been displayed.

As per our observation from the above table, 44.27% variance explained by factor 1 alone. 88.759% variance explained by 6 factors mentioned in the table also having Eigenvalues greater than 1. The conclusion from the above table is that 6 factors were extracted and 24 factors were dropped. The rotation form used for this main component analysis is Direct Oblimin, with Kaiser Normalization and 14 iterations of rotation converged. Analysis of the pattern matrix indicates that as shown below, the variables or features come under a specific factor. We have sorted the features according to the size and suppressed the values less than 0.5.
Factor 1 having the highest number of features grouped which is 12. Some features like symmetry_se, symmetry_worts, and symmetry_mean are having a negative impact on factor 6.
The component correlation matrix is not a unit matrix and hence we can go ahead with the rotation that we have considered.

Using the dimension reduction technique, we have reduced the 30 variable data set to 6 variable data sets which explain 88.759% of the total variance. To determine if the patient has a malignant or benign tumor in the breast, we should apply the necessary algorithms to these factors.

1. Logistic Regression

   The equation for the logistic regression model is given by the following:
   \[ P(Y) = \frac{1}{1 + e^{-(a_0 + a_1X_{1j} + a_2X_{2j} + \ldots + a_nX_{nj})}} \]

   We estimate the probability of \( Y \) being from \( X \). Equation value ranges from 0 to 1. A value close to 1 means that \( Y \) is very likely to occur and very unlikely to occur if the value is close to 0. When using all 6 factors the Hosmer and Lemeshow test was showing 0.95 value which means the model is not significant. Therefore, we have not considered the 2\(^{nd}\) factor for this logistic regression as the factor was not significant.
When values in the above table are updated in the equation and factor value for the specific patient is entered, we would get the diagnosis according to this model. After analyzing the classification table, we observe that there are 6 patients who have a benign tumor but according to this model they are tested as having a malignant tumor and 7 patients who had a malignant tumor were tested as having a benign tumor. The model shows an overall accuracy of 97.7%.

2. K- Nearest Neighbour Classification

The value of k is 3, meaning 3 nearest neighbors were taken into consideration for this model. The overall accuracy of this model is 94.20%.
3. Decision Tree for predictive analysis of the data gives the below-mentioned results.

<table>
<thead>
<tr>
<th>Risk</th>
<th>Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Estimate</td>
<td>Std. Error</td>
</tr>
<tr>
<td>.104</td>
<td>.013</td>
</tr>
</tbody>
</table>

Growing Method: CHAID
Dependent Variable: Diagnosis_New

The number of nodes and terminal nodes is 7 and 6 respectively; and decision tree depth is 1. The overall accuracy for this model is 89.6%.

4. Discriminant Analysis
A statistical technique used for classification of observations forming groups without overlaps, which are based on scores or values of the variables or factors into consideration.
From the above table, it is observed that the model is significant as the value of Sig. is less than 0.05, Chi-Square is high. The overall accuracy of the model is 96.48%.

V. Conclusion:

Therefore, we can conclude that we can use the above-mentioned algorithms to boost the accuracy of the diagnostic tests and thus reduce the number of patients misdiagnosed. When data on these models be trained would be large enough, they would be able to able to diagnose whether the patient has a malignant tumor or is it benign more accurately.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression</td>
<td>97.70 %</td>
</tr>
<tr>
<td>KNN Classification</td>
<td>94.20 %</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>89.60 %</td>
</tr>
<tr>
<td>Discriminant Analysis</td>
<td>96.48 %</td>
</tr>
</tbody>
</table>

References:


Date Source:
https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29