European Journal of Molecular & Clinical Medicine
ISSN 2515-8260 Volume 7, Issue 03, 2020

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 Nwigwe1, 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,

Dinov and Katapodi, 2019).Application of predictive analysis and comparison of models for breast cancer survivability (Jhajharia, Verma and Kumar, 2016).

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 isconsidered 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 to4 decimal places.

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

Statistics

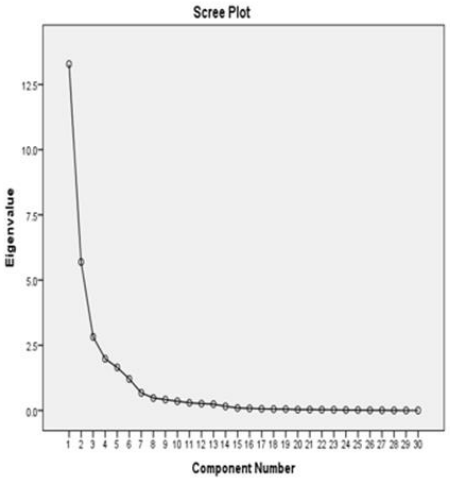
N	Valid	569
	Missing	0

	Frequency	Percent	Valid Percent	Cumulative %
Benign (0)	357	62.7	62.7	62.7
Malignant (1)	212	37.3	37.3	100.0
Total	569	100.0	100.0	

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.

Total Variance Explained

Factors	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings[a]
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total
1	13.282	44.272	44.272	13.282	44.272	44.272	11.124
2	5.691	18.971	63.243	5.691	18.971	63.243	7.554
3	2.818	9.393	72.636	2.818	9.393	72.636	3.004
4	1.981	6.602	79.239	1.981	6.602	79.239	3.736
5	1.649	5.496	84.734	1.649	5.496	84.734	5.597
6	1.207	4.025	88.759	1.207	4.025	88.759	4.493
7	.675	2.251	91.010				



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 valuesless than 0.5.

Pattern Matrix*

	Component					
	1	2	3	4	5	6
area_mean	.969					
radius_mean	.943					
area_worst	.940					
perimeter_mean	.935	1				
radius_worst	.925					
area_se	.916					
perimeter_worst	.911					
radius_se	.865					
perimeter_se	.840					
concavepoints_mean	.749					
concavepoints_worst	.583					
concavity_mean	.551					
concavity_se		.938				
compactness_se		.904				
fractal_dimension_se		.903				
concavepoints_se		.730	2			
concavity_worst		.519				
fractal_dimension_worst		.501				
compactness_mean						
smoothness_se			.762			
texture_se			.654	.538	3	
compactness_worst						
texture_worst				.983		
texture_mean				.931	4	
smoothness_worst					.975	
smoothness_mean					.860	5
fractal_dimension_mean					.503	
symmetry_se						-.827
symmetry_worst						-.818
symmetry_mean						-.771

6

a. Rotation converged in 14 iterations.

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.

Component Correlation Matrix

Component	1	2	3	4	5	6
1	1.000	.226	-.125	.254	.146	-.130
2	.226	1.000	.080	.156	.356	-.381
3	-.125	.080	1.000	-.083	-.159	-.087
4	.254	.156	-.083	1.000	.106	-.113
5	.146	.356	-.159	.106	1.000	-.322
6	-.130	-.381	-.087	-.113	-.322	1.000

Extraction Method: Oblimin with Kaiser Normalization

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 the Y is very likely to occur and very unlikely to occur if the value is close to 0 Y. 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 2nd factor for this logistic regression as the factor was not significant.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	81.799[a]	.692	.944

a. Estimation ended at the number 10 iteration, because estimates of parameters changed by less than.001.

Cox & Snell R Square value ranges from 0 to 0.7 and 0 to 1 is the range for Nagelkerke R square. Both conditions are satisfied as observed from the above table.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	25.773	8	.001

The model's significance is demonstrated by testing Hosmer and Lemeshow. As the value of Sig. is less than 0.05, hence we can conclude that model is significant.

Classification Table

Observed		Predicted		
		Diagnosis_New		Percentage Correct
		0	1	
Diagnosis_New	0 (Benign)	351	6	98.3 %
	1 (Malignant)	7	205	96.7 %
Overall %				97.7 %

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	FAC1_1	9.721	1.541	39.792	1	.000	16669.187
	FAC3_1	-1.186	.348	11.589	1	.001	.306
	FAC4_1	2.812	.479	34.401	1	.000	16.639
	FAC5_1	2.333	.476	24.031	1	.000	10.311
	FAC6_1	-.612	.304	4.050	1	.044	.542
	Constant	-.514	.308	2.793	1	.095	.598

a. Factor(s) entered on step 1: FAC1_1, FAC3_1, FAC4_1, FAC5_1, FAC6_1.

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

Case Processing Summary

		N	Percent
Sample	Training	384	67.5%
	Holdout	185	32.5%
Valid		569	100.0%
Excluded		0	
Total		569	

Diagnosis_New * Predicted Value for Diagnosis_New Cross tabulation

		Predicted Value for Diagnosis_New		Total	Percentage Correct
		0	1		
Diagnosis_New	0	348	9	357	97.47%
	1	24	188	212	88.67%
Total		372	197	569	94.20%

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.

Risk

Estimate	Std. Error
.104	.013

Growing Method: CHAID

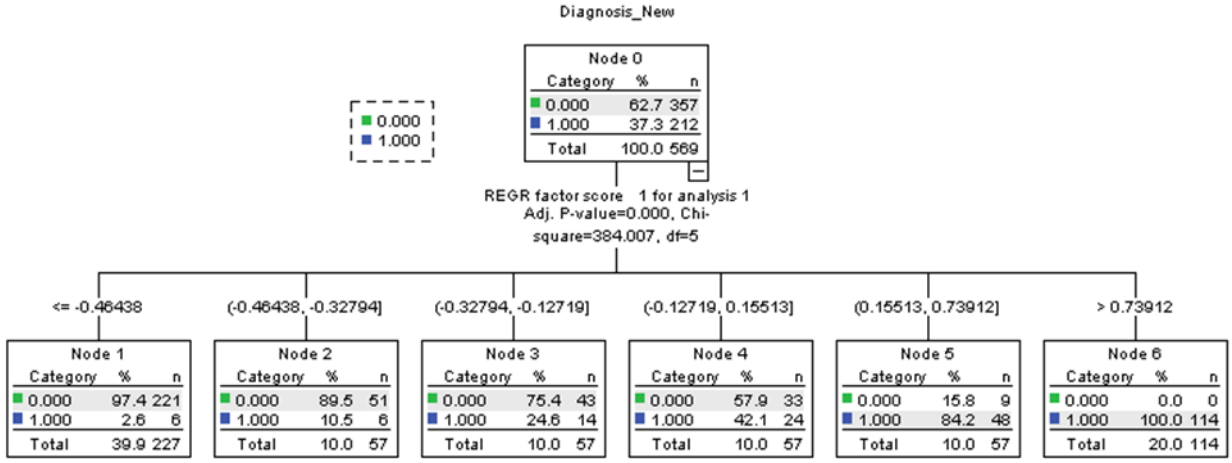
Dependent Variable: Diagnosis_New

Classification

	Predicted		
Observed	0	1	Percent Correct
0	348	9	97.5%
1	50	162	76.4%
Overall Percentage	69.9%	30.1%	89.6%

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, whichare based on scores or values of the variables or factors into consideration.

Box's Test of Equality of Covariance Matrices

Log Determinants

Diagnosis New	Rank	Log Determinant
0	6	-3.917
1	6	-.829
Pooled within-groups	6	-1.777

Log Determinants

Diagnosis New	Rank	Log Determinant
0	6	-3.917
1	6	-.829
Pooled within-groups	6	-1.777

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	2.400[a]	100.0	100.0	.840

a. The study used the first 1 canonical discriminant functions.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	.294	690.226	6	.000

Diagnosis_New * Predicted Group for Analysis 1 Cross tabulation

Count

		Predicted Group for Analysis 1		Total	Percentage Correct
		0	1		
Diagnosis_New	0	355	2	357	99.43%
	1	18	194	212	91.50%
Total		373	196	569	96.48%

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 be able to diagnose whether the patient has a malignant tumor or is it benign more accurately.

Algorithm	Accuracy
Logistic Regression	97.70 %
KNN Classification	94.20 %
Decision Tree	89.60 %
Discriminant Analysis	96.48 %

References:

1. Breast Cancer. (2020). World Health Organization. Retrieved from https://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/
2. Mitra, S., &Dey, P. (2016). Fine-needle aspiration and core biopsy in the diagnosis of breast lesions: A comparison and review of the literature. *CytoJournal*, *13*, 18. https://doi.org/10.4103/1742-6413.189637
3. Bleyer, A., & Welch, H. G. (2012). Effect of three decades of screening mammography on breast-cancer incidence. *The New England journal of medicine*, *367*(21), 1998–2005. https://doi.org/10.1056/NEJMoa1206809
4. Gupta S. (2016). Breast cancer: Indian experience, data, and evidence. *South Asian journal of cancer*, *5*(3), 85–86. https://doi.org/10.4103/2278-330X.187552

5. Malvia, S., Bagadi, S. A., Dubey, U. S., &Saxena, S. (2017). Epidemiology of breast cancer in Indian women. *Asia-Pacific journal of clinical oncology*, *13*(4), 289–295. https://doi.org/10.1111/ajco.12661
6. Ogbuanya, A. U., Anyanwu, S. N., Iyare, E. F., &Nwigwe, C. G. (2020). The Role of Fine Needle Aspiration Cytology in Triple Assessment of Patients with Malignant Breast Lumps. *Nigerian journal of surgery : official publication of the Nigerian Surgical Research Society*, *26*(1), 35–41. https://doi.org/10.4103/njs.NJS_50_19
7. Simons, A. and Gordon, R., (2019, May 7). Using AI to Predict Breast Cancer and Personalize Care. MIT News, p. 1. Retrieved from https://news.mit.edu/search?keyword=breast%20cancer
8. Ming, C., Viassolo, V., Probst-Hensch, N., Chappuis, P. O., Dinov, I. D., &Katapodi, M. C. (2019). Machine learning techniques for personalized breast cancer risk prediction: comparison with the BCRAT and BOADICEA models. *Breast cancer research : BCR*, *21*(1), 75. https://doi.org/10.1186/s13058-019-1158-4
9. Jhajharia, S., Verma, S., and Kumar, R. (2016). Predictive Analytics for Breast Cancer Survivability: A Comparison of Five Predictive Models. In Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies (ICTCS '16). Association for Computing Machinery, New York, NY, USA, Article 26, 1–5. DOI:https://doi.org/10.1145/2905055.2905084

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