Journal, Advances in Mathematical & Computational Sciences
Vol. 10 No. 3, September, 2022
www.isteams.net/mathematics-computationaljournal

Journal of Advances in Mathematical & Computational Sciences
An International Pan-African Multidisciplinary Journal of the SMART Research Group
International Centre for IT & Development (ICITD) USA
© Creative Research Publishers
Available online at https://www.isteams.net/ mathematics-computationaljournal.info
CrossREF Member Listing - https://www.crossref.org/06members/50go-live.html

Comparison of Binary Logistic Regression and Perceptron Neural Network Modelling in Terms of Prediction Using Breast Cancer Data

Ohuabunwa, C Humphrey & Ogoke P. Uchenna
Department of Mathematics and Statistics
University of Port Harcourt
Choba, Port Harcourt, Nigeria
Emails: humphrey_ohuabunwa@uniport.edu.ng; uchenna.ogoke@uniport.edu.ng

ABSTRACT

Artificial network modelling are being used in areas of prediction and classification, areas where classical regression models and other related statistical techniques have traditionally been used. This study is aimed at comparing the Logistic Regression Model and the Perceptron Neural Network Modelling to detect the malignancy or benignancy of the tumorous cell of cancer patients more accurately. Data from the Breast Cancer Wisconsin (Diagnostic) was analysed using the Classical Regression Model and the Perceptron model. The study compared both Binary Logistic Model (BLM) and the Perceptron Neural Model (PNM) by their level of accuracy in predicting the breast cancer outcome. The Perceptron model had greater accuracy (98.3%) than the Binary Logistic Model (97.2%) This goes to show that the Perceptron Neural Network Model was a better predictive model as it had a higher accuracy in predicting the cancer model. The study also found that both the perceptron neural networks and logistic regression models can remarkably predict cancer very close to the actual values but the performance of the perceptron neural network model for prediction of cancer was higher and more precise. The study recommends that Perceptron neural network model as a better alternative to the Logistic Regression Model.

Keywords: Logistic Regression, Perceptron, Classical, Neural Network

Ohuabunwa, C.H. & Ogoke, P.U. (2022): Comparison of Binary Logistic Regression and Perceptron Neural Network Modelling in Terms of Prediction Using Breast Cancer Data. Journal of Advances in Mathematical & Computational Science. Vol. 10, No. 3. Pp 111-120
Available online at www.isteams.net/mathematics-computationaljournal.
DOI: dx.doi.org/10.22624/AIMS/MATHS/V10N3P9

1. INTRODUCTION

An artificial intelligence technique known as a neural network instructs computers to analyze data in a manner similar to the way the human brain does, using interconnected nodes, or neurons, in a layered structure liken to the human brain. It develops an adaptive framework that enables computers to continuously learn from their errors and advance. Rosenblatt (1958), developed a probabilistic model for information storage and organization in the brain credited with creating the perceptron. Different comparison techniques have been done using the traditional statistical techniques and the neural networks for prediction and classification problems in various areas of applications (Mukta and Usha (2009), Yesilnacar and Topal (2005) Goetz etal, (2015),).

Bozak and Aybek (2020) identified the classification performance of artificial neural network to be significantly better compared to logistic regression at predicting the science literacy success of the 15-year Turkish students who participated in PISA research carried out in 2015 by using variables like learning time spent on science, test anxiety, environmental awareness, environmental optimism, etc,. A comparison of logistic regression with neural networks has been examined in different ways for prediction (Elif (2016).Some conventional statistical techniques to conduct a thorough analysis and comparison of several neural network models in different areas (Ripley (2014).The basic structure Neural Network is comprised of three distinctive layers, the input layer where the data are introduced to the model and computation of the weighted sum of the input is performed, the hidden layer or layers where data processing takes place, and the output layer where the results of the neural network are produced (Ranzato et al., 2007).The perceptron Neural Networks and Logistic Regression are algorithms related to classification problems where you have a discrete number of possibilities. The Perceptron Neural Networks has a very particular "structure" where you have one input layer, at least one hidden layer and finally an output layer. Both algorithms have a similar problem: to find the best value for their parameters.

2. METHOD

This study compared the Perceptron neural network model and the Binary logistic regression model in order to determine the best model for the prediction of breast cancer. For the purpose of this study, the radius (mean of distances from center to points on the perimeter), texture (standard deviation of gray-scale values), perimeter, area, smoothness (local variation in radius lengths), compactness (perimeter^2 / area - 1.0), concavity (severity of concave portions of the contour), concave points (number of concave portions of the contour), symmetry, fractal dimension ("coastline approximation" - 1) which served as the explanatory variables were used to predict the presence or absence of the Diagnosis (M = malignant, B = benign). For the basis of comparison, the study will adopt binary logistic regression and perceptron neural network techniques. The data for this study was got from:

https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+wisconsin

Journal, Advances in Mathematical & Computational Sciences
Vol. 10 No. 3, September, 2022
www.isteams.net/mathematics-computationaljournal

2.1 Perceptron Neural Network Model

The Step function or the activation function of the neural network is given as:

$$f(x) = \frac{1}{1+e^{-z}} \tag{1}$$

The Activation function gives an output of l, if $\sum X_i W_i + b \geq 0$ (2)

The Activation function also gives an output of 0, if $\sum X_i W_i + b < 0$ (3)

The weighted sum of the model is represented by $\sum_{i=1}^{n} X_i W_i + b$ (4)

Where X_1 are the Input data, W_i are the associated adjustable weights of the input data
The model is represented diagrammatically as

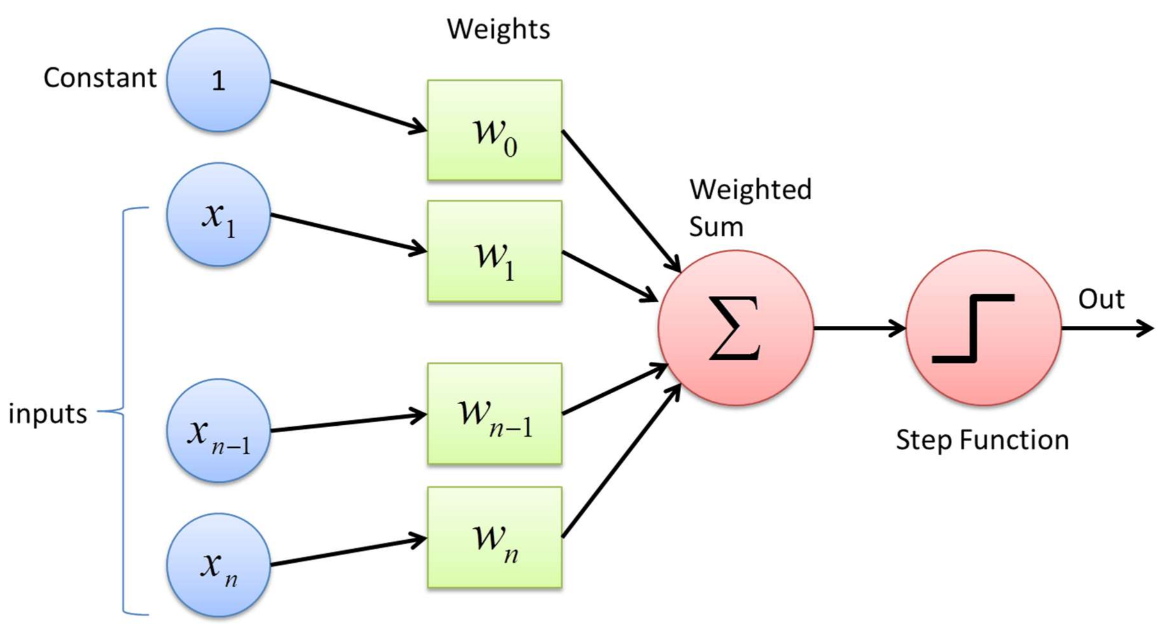

Figure 1: Perceptron Neural Network Model

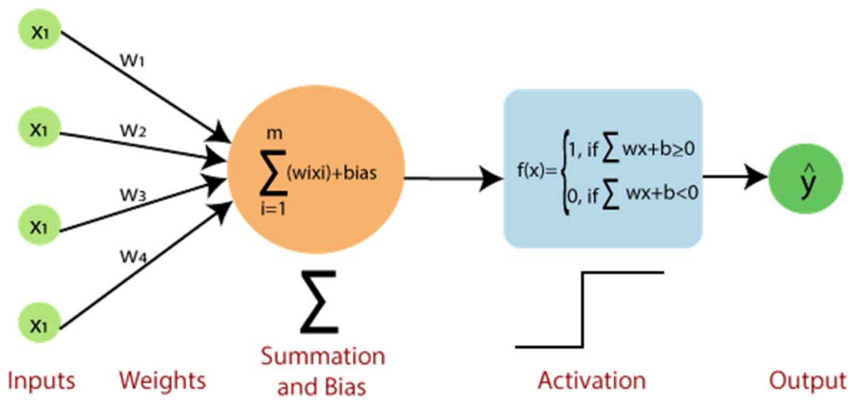

Figure 2: Perceptron Neural Network Model continued.

2.2 Logistic Regression Model

Using Logistic Regression technique, we can obtain the probability that a patient will be have breast cancer is:
Let y_i be the response of the i^{th} breast cancer patient randomly selected having cancer (y_i=1) or not (y_i=0) for i=1, 2, ..., n.

Now let

$$Y_i = \begin{cases} 1, if\ the\ ith\ patient\ is\ breast\ cancer\ \ positive\ or \\ \qquad 0\ if\ negative \\ \qquad for\ i = 1, 2, \ldots\ \ n \end{cases} \qquad (4)$$

Then a multiple logistic regression model regressing the probability that the ith subject responds positive to the condition under study on the independent variables,$x_{i1},\ x_{i2}, \ldots, x_{ik}$

is $P(y_i = 1/x_{i1}, x_{i2}, \ldots, x_{ik} = \ p_{ix} = \frac{1}{1+\ e^{-\beta_0+\sum_{i=1}^{k}\beta_j x_{ij}+e_j}}$ (5)

Where $\beta_{j's}$ are the regression coefficients and $e_{i's}$ are error term uncorrelated with $x_{ij's}$ for i= 1,2,...,k.

The estimate of the probability of positive response is thereby given below as

$$\hat{p}_{ix}(1) = \frac{1}{1+\ e^{-b_0+\sum_{i=1}^{k} b_j x_{ij}+e_j}} \qquad (6)$$

For s i=1, 2, ..., n

3. RESULTS AND DISCUSSION

3.1 Logistic Regression Results

The Dependent Variable Encoding Table 3.1 displays how the values for Benign(B) and Malignant(M) cancer were coded. This is important for classification in the logistic regression.

Journal, Advances in Mathematical & Computational Sciences
Vol. 10 No. 3, September, 2022
www.isteams.net/mathematics-computationaljournal

Table 3.1: Dependent Variable Encoding

Original Value	Internal Value
B	0
M	1

Table 3.2 Classification Table without the independent variables (Block 0: Beginning Block)

	Observed		Predicted		
			Y diagnosis		Percentage Correct
			B	M	
Step 0	Ydiagnosis	B	357	0	100.0
		M	212	0	.0
	Overall Percentage				62.7
a. Constant is included in the model.					
b. The cut value is .700					

Table 3.3: Variables in the Logistic Equation

Variables in the Equation									
		B	S.E.	Wald	Df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 7[a]	radius_se	-15.566	4.110	14.343	1	.000	.000	.000	.001
	compactness_se	81.887	28.273	8.388	1	.004	365551603538054900000000000000000.000	313812862374.659	4.258E+59
	radius_worst	-1.440	.293	24.213	1	.000	.237	.133	.420
	texture_worst	-.371	.080	21.373	1	.000	.690	.590	.808
	smoothness_worst	-56.047	23.419	5.727	1	.017	.000	.000	.000
	concavity_worst	-9.283	3.914	5.625	1	.018	.000	.000	.200
	concave points_worst	-47.162	18.113	6.780	1	.009	.000	.000	.000
	Constant	52.653	9.677	29.608	1	.000	7363393270807389000000.000		
a. Variable(s) entered on step 7: concavity_worst.									

Table 3.4: Omnibus Tests of Model Coefficients

		Chi-square	Df	Sig.
Step 1	Step	719.035	30	.000
	Block	719.035	30	.000
	Model	719.035	30	.000

Table 3.4 contains statistics that measure goodness-of-fit to assist in determining if the model accurately captures the data. The model fit is examined using the Omnibus Tests of Model Coefficients. If the model is significant, this indicates that the fit is much better than the null model, indicating that the model is demonstrating a good fit which is also displayed by our model for cancer (p <.000).

Table 3.5: Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	32.405[a]	.717	.979

a. Estimation terminated at iteration number 14 because parameter estimates changed by less than .001.

In Table 3.5, the Nagelkerke's R^2, a modified version of the Cox & Snell R-square that changes the scale of the statistic to encompass the entire range from 0 to 1, is what is typically employed. In this instance, the findings show that the predictor variables in the model can account for 97.9% of the change in the dependent variable.

Table 3.6 Classification Table for Logistic Regression

	Observed		Predicted		
			Y diagnosis		Percentage Correct
			B	M	
Step 1	Y diagnosis	B	208	4	98.1
		M	12	345	96.6
	Overall Percentage				97.2

a. The cut value is .700

Table 3.6 shows how effectively the model is able to predict the cancer category well. To determine how much the predictor variables enhance the model, we may contrast this with the Classification Table 3.2 displayed for Block 0 with 62.7%. Overall, 97.2% of instances were correctly classified by the model (also known as the Percentage Accuracy in Classification (PAC)).

Table 3.7: Variables in the Binary Logistic Regression Equation

	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
							Lower	Upper
x1	-9.593	20.429	0.221	1	0.639	0	0	1.6702E+13
x2	0.36	0.592	0.37	1	0.543	1.433	0.45	4.568
x3	-1.318	2.928	0.202	1	0.653	0.268	0.001	83.211
x4	0.15	0.116	1.675	1	0.196	1.162	0.926	1.457
x5	-39.808	183.835	0.047	1	0.829	0	0	1.56E+139
x6	-71.351	96.3	0.549	1	0.459	0	0	9.62E+50
x7	60.86	99.784	0.372	1	0.542	2.6986E+26	0	2.33E+111
x8	153.157	207.396	0.545	1	0.46	3.28E+66	0	1.12E+243
x9	27.106	75.054	0.13	1	0.718	5.9163E+11	0	4.55E+75
x10	333.527	247.092	1.822	1	0.177	7.06E+144	0	.
x11	49.93	52.315	0.911	1	0.34	4.8333E+21	0	1.64E+66
x12	0.167	3.897	0.002	1	0.966	1.182	0.001	2454.406
x13	-8.811	5.775	2.328	1	0.127	0	0	12.275
x14	0.389	0.436	0.796	1	0.372	1.475	0.628	3.465
x15	219.727	259.792	0.715	1	0.398	2.67E+95	0	.
x16	-181.599	152.681	1.415	1	0.234	0	0	1.25E+51
x17	-253.363	140.672	3.244	1	0.072	0	0	5081058386
x18	780.455	366.633	4.531	1	0.033	.	7.3955E+26	.
x19	35.388	210.222	0.028	1	0.866	2.3367E+15	0	2.04E+194
x20	-164.082	425.283	0.149	1	0.7	0	0	5.52E+290
x21	-0.083	6.929	0	1	0.99	0.921	0	727183.311
x22	0.524	0.475	1.22	1	0.269	1.689	0.666	4.284
x23	1.258	1.012	1.546	1	0.214	3.518	0.484	25.553
x24	-0.018	0.062	0.083	1	0.773	0.982	0.869	1.11
x25	-2.95	101.301	0.001	1	0.977	0.052	0	8.83E+84
x26	-26.429	31.37	0.71	1	0.4	0	0	1.6775E+15
x27	49.485	33.436	2.19	1	0.139	3.0971E+21	0	8.94E+49
x28	57.421	72.992	0.619	1	0.431	8.6616E+24	0	1.17E+87
x29	33.214	47.037	0.499	1	0.48	2.6586E+14	0	2.90E+54
x30	81.05	121.24	0.447	1	0.504	1.5829E+35	0	2.51E+138
Constant	-37.074	51.339	0.521	1	0.47	0		

Table 3.7 is the most important table in the binary regression analysis. B (Beta) which is also known as the regression coefficient is the expected change in Log Odds and for every unit change in the B (Beta), there is an Exp(B) change in the likelihood of the result. If the B (Beta) coefficient is negative, the outcome variable will decrease by the B (Beta) coefficient value for each unit rise in the predictor variable. Thus variable x1 with B= -9.593 will decrease the outcome or classifier (M or B) by 9.593 units while x2 will increase the outcome or classifier (M or B) by 0.36 units. The study found that only variable x18 with Beta (B = 780.455) and p-value (p = 0.033) was statistically significant at 95% confidence level while all other variable were statistically non-significant.

3.2 Perceptron Neural Network Results

Table 3.8 shows how effectively the model is able to predict the cancer category well for the testing and training categories. The overall percentage accuracy in classification (PAC) for training model is 98.3% while that of the testing model is 98.1%.

Table 3.8: Perceptron Neural Network Classification Table

Sample	Observed	Predicted		
		B	M	Percent Correct
Training	B	139	3	97.9%
	M	3	263	98.9%
	Overall Percent			98.3%
Testing	B	67	3	95.7%
	M	0	91	100.0%
	Overall Percent			98.1%
Dependent Variable: Y diagnosis				

3.3: Comparison of Binary Logistics Regression and Perceptron Neural Network Models

Table 3.9: Comparison of Logistics Regression and Perceptron Neuron Network Models

		Predicted		
Models	Observed	Benign	Malignant	Correct Predicted Percentage
Logistic Regression	Benign	208	4	98.1%
	Malignant	12	345	96.6%
Overall Percentage				97.2%
PNN (Training)	Benign	139	3	97.9%
	Malignant	3	263	98.9%
Overall Percentage				98.3%
PNN (Testing)	Benign	67	3	95.7%
	Malignant	0	91	100%
Overall Percentage				98.1%

Table 3.9 compares the Logistic Regression model and Perceptron Neural Network Models.The findings showed that the Perceptron Neural Network Model performed better in terms of predicting accurately the malignancy or benignancy of the tumorous cell of cancer patients. It showed a 98.3% and 98.1% training and testing accuracy as compared to 97.2% predictive accuracy of the logistic regression model. This finding is an accordance with the study of Abdolmaleki *et al* (2004) in his study on comparison of logistic regression and neural network models in the outcome of biopsy in breast cancer from MRI. They revealed in their study that the performance of ANN is better than the logistic regression model when all input and/or variables are similar.

4. CONCLUSION

The study compared both Logistic Regression Model and the Perceptron Neural Network Model by their level of accuracy in predicting the outcome, the Perceptron Neural Network Model had a greater accuracy than the logistic regression model by accurately predicting 98.3% and 98.1% instances in both the training and testing dataset while the logistic model predicted 97.2% accurately. This goes to show that the Perceptron Neural Network Model was a better predictive model as it had a higher accuracy in predicting the cancer model.

REFERENCES

1. Abdolmaleki, P., Gity, M., Tahmasebi, M., and Rohandeh, M. (2004). Majid,1382,Logistic Reversion
2. Scrunity of Breast Carcinomas From Mammographic Evaluation. 11th Iranian conference on
3. Biomedical Engineering.
4. Bozak, A. and Aybek, E.C. (2020). Comparison of artificial neural networks and logistic regression Analysis in PISA science literacy success prediction. *International Journal of Contemporary Educational Research,* 7(2), 99-111.
5. Elif, B. (2016). Using Neural Network and Logistic Regression Analysis to Predict Prospective
6. Mathematics Teachers' Academic Success upon Entering Graduate Education. Educational
7. Sciences: Theory & Practice 16(3), 943-964.
8. Goetz,J.N; Brenning, A, Petschko,H and Leopold, P (2015).Evaluating machine learning and Statistical prediction techniques for landslide susceptibility modeling, Computers & Geosciences. 81, 1-11.
9. Mukta, P and Usha A. K, (2009). Neural networks and statistical techniques: A review of applications, *Expert Systems with Applications.* 36(1) 2-17.
10. Ranzato, M. A., Huang, F. J., Boureau, Y. L., and LeCun, Y. (2007). Unsupervised learning of invariant feature hierarchies with applications to object recognition. In *2007 IEEE conference on computer vision and pattern recognition*. 1-8. IEEE.
11. Ripley, B. D. (2014). Neuronic Links and related methods for classification. *Journal of the Royal Statistical Society, Series B (Methodological), 56*(3), 409–456.

12. Rosenblatt, F. (1958). Of biological memory. *Matematika*, *5*(6), 18-31.
13. Yesilnacar, E and Topal, T. (2005). Landslide susceptibility mapping: A comparison of logistic
14. regression and neural networks methods in a medium scale study, Hendek region (Turkey),
15. Engineering Geology. 79 (3–4), 251-266.