Logistic regression and artificial neural network models in cervical cancer

Abstract-The objective of this article is the application of Excel, Statgraphics and SPSS to diagnosis of cervical cancer. We set up tables of classification to compare the global correct prediction. The data used comes from 30 cases of cancer of cervix. We proceeded to fit a logistic regression (LR) model and train Artificial Neural Networks (ANNs). The multicollinearity problem, usually present in modelling with numerous predictive variables, was addressed with factor analysis and Pearson Correlations. Based on the percentage of correct classification, the LR was superior to the ANNs. The multicollinearity problem, usually present in modelling with numerous predictive variables, was addressed with factor analysis and Pearson Correlations. Based on the percentage of correct classification, the LR was superior to the ANNs. Based on the percentage of correct classification, the regression logistic was superior to the neural networks for the database of cervical cancer. The Hosmer-Lemeshow test has been used to validate the logistic regression model. The models have reported an acceptable statistical significance for this test.

Keywords: Neural Networks, Pearson correlation, Factorial analysis, Logistic regression, Cervical cancer

Logistic regression and artificial neural network models in cervical cancer

Hernan Oscar Cortez Gutierrez [1][https://orcid.org/0000−0002−1516−5583], Milton Milciades Cortez Gutierrez[2][https://orcid.org/0000−0003−4939−7734], Cesar Ángel Durand Gonzales[1][https://orcid.org/0000−0002−2148−5903], Rubén Dario Mendoza Arenas[1][https://orcid.org/0000−0002−7861−7946], Ana María Yamunaque Morales[1][https://orcid.org/0000−0001−7891−998X], Lucio Arnulfo Ferrer Peñaranda[1][https://orcid.org/0000−0001−7953−925X], Marisol Paola Delgado Baltazar[1][https://orcid.org/0000−0002−0278−9557], Liv Jois Cortez Fuentes Rivera[3][https://orcid.org/0000−0003−2478−712X]

[1]Universidad Nacional del Callao, Perú, hocortezg@unac.edu.pe, cadurandg@unac.edu.pe, rdmendozaa@unac.edu.pe, anyamuniquem@unac.edu.pe, laferrerp@unac.edu.pe, mpdelgadob@unac.edu.pe,
[2]Universidad Nacional de Trujillo, Perú, mcortezgutierrez@yahoo.es.
[3]Universidad Privada San Juan Bautista, Perú, livjoisc@gmail.com.

Abstract-The objective of this article is the application of Excel, Statgraphics and SPSS to diagnosis of cervical cancer. We set up tables of classification to compare the global correct prediction. The data used comes from 30 cases of cancer of cervix. We proceeded to fit a logistic regression (LR) model and train Artificial Neural Networks (ANNs). The multicollinearity problem, usually present in modelling with numerous predictive variables, was addressed with factor analysis and Pearson Correlations. Based on the percentage of correct classification, the LR was superior to the ANNs. The multicollinearity problem, usually present in modelling with numerous predictive variables, was addressed with factor analysis and Pearson Correlations. Based on the percentage of correct classification, the LR was superior to the ANNs. Based on the percentage of correct classification, the regression logistic was superior to the neural networks for the database of cervical cancer. The Hosmer-Lemeshow test has been used to validate the logistic regression model. The models have reported an acceptable statistical significance for this test.

Keywords-Neural Networks, Pearson correlation, Factorial analysis, Logistic regression, Cervical cancer

I. INTRODUCTION

Computer models in medical diagnosis in cervical cancer are being developed to detect the presence of cancer early. We have as a reference [1] that uses logistic regression models to estimate cevical cancer risk factors.

Logistic regression establishes a relationship between a dichotomous dependent variable, in our case, presence and absence of cervical cancer, and independent variables such as age, biopsy, tobacco and alcohol consumption. Likewise, the following equation is established:

Digital Object Identifier: (only for full papers, inserted by LACCEI).
ISSN, ISBN: (to be inserted by LACCEI).
DO NOT REMOVE

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n$$

Artificial intelligence (AI) has been used in various studies to improve cervical cancer screening and diagnosis. Here are some of the key findings from the search results:

A study [2] discusses the use of AI in cervical cancer screening and diagnosis. The study highlights the benefits of AI-based medical diagnostic applications, including reduced time consumption, reduced need for professional and technical personnel, and no bias owing to subjective factors. The study concludes that AI can be used to improve the accuracy of early diagnosis.

Another study [3] describes the implementation of digital microscopy with AI at a rural clinic to detect atypical cervical smears with a high sensitivity compared with visual sample analysis. The study concludes that the use of AI in cervical cancer screening can provide needed screening to resource-limited areas.

A study [4] reports that an automated dual-stain method using AI improved the accuracy and efficiency of cervical cancer screening compared with cytology (Pap test), the current standard for follow-up of women who test positive with primary human papillomavirus (HPV) screening. The study concludes that their findings serve as an important example for introducing digital pathology and deep learning into clinical practice.

Another study [5] discusses the use of AI-assisted fast screening for cervical high-grade squamous intraepithelial lesion and squamous cell carcinoma diagnosis and treatment planning. The study concludes that the application of AI may provide a new screening method of cervical Pap smear and warrants further validation in a larger population-based study in future work.

A study [6] led by investigators from the National Institutes of Health and Global Good has developed a computer algorithm that can analyze digital images of a woman's cervix and accurately identify precancerous changes that require medical attention. The study concludes that the AI approach, called automated visual evaluation, has the potential to revolutionize cervical cancer screening, particularly in low-resource settings.

In summary, AI has shown promising results in improving cervical cancer screening and diagnosis, including the potential to revolutionize cervical cancer screening in low-resource settings.

II. METHODS

In this research we present two methods, which are shown below.

A. Artificial intelligence

Since memorial times, the study of artificial intelligence has been one of the main topics that have fascinated scientists and philosophers. However, no significant progress has been achieved to date.

Artificial intelligence is divided into two fields, symbolic and sub-symbolic artificial intelligence.

Symbolic artificial intelligence

In symbolic artificial intelligence we must define a problem to be solved and design a system capable of solving the problem following schemes predetermined by the discipline.

Expert systems follow this scheme, introducing a series of logical rules that collect knowledge about a subject, from inference mechanisms similar to those used by human reasoning, to obtain conclusions.

In symbolic artificial intelligence it is said that expert systems follow a top-down scheme since it is necessary to have an approximation of a solution to the problem and to design said approximation.

Thus, we have as an example a symbolic perspective, which consists of the study of human reasoning mechanisms at a high level, that is, how we face a problem, how we approach it, and how we solve it.

A greater understanding of human reasoning implies that the system produced will be more efficient when it comes to solving problems.

Sub-symbolic artificial intelligence

In sub-symbolic artificial intelligence, the design of high-level schemes to solve problems using techniques of the discipline is not implemented, but it starts with generic systems that must be adapted and built so that the system is capable of solving the problem.

His sub-symbolic perspective studies the physical mechanisms that enable us as intelligent beings.

The nervous system is the fundamental mechanism that enables any living being to perform a sophisticated task that is not pre-programmed.

Artificial neural networks

The ideal goal of artificial neural networks is to design machines with parallel processing neural elements, so that the general behavior of the neural network simulates the behavior of animal neural systems.

The perceptron This model is capable of classifying automatically, from a set of examples of different classes.

The information on which the system is based must be constituted by the existing examples of different classes, these examples are known as training patterns, since these are the ones that provide the necessary information for the system to build discriminant surfaces.

The system at the end of the process should be able to determine any new instance, and its corresponding class ideally. However, in most applications there is no such model that classifies any pattern new to it.

Model description

In the network structure there is a set of input cells, as many as are necessary according to the terms of the problem; and one or more output cells. Each of the input cells has connections with all the output cells, and it is these connections that determine the discrimination surface.

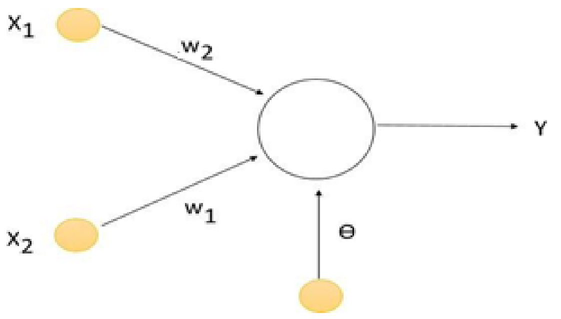


Fig. 1 Perceptron architecture with two inputs and one output.

Where:

- x_1, x_2 correspond to inputs.
- y correspond to output.
- w_1, w_2 correspond to the weights.
- θ correspond to the threshold.

In this scheme, the network output is obtained as follows:

First, we calculate the activation of the cell at the output, through the weighted sum between weights and inputs, as shown in equation (1).

$$y = \sum_{i=1}^{n} w_i x_i \tag{1}$$

The final output is produced by applying an output function to the activation level of the cell.

$$y = F(y, \theta)$$

$$F(s,\theta) = \begin{cases} 1 & , \text{ si } s > \theta \\ -1 & , \text{ otherwise} \end{cases} \tag{2}$$

By passing θ to the other side of the equation, the output can be written in a single equation (3).

$$y = F\left(\sum_{i=1}^{n} w_i x_i + \theta\right) \tag{3}$$

Where F is an independent variable.

$$F(s) = \begin{cases} 1 & , \text{ si } s > \theta \\ -1 & , \text{ otherwise} \end{cases} \tag{4}$$

The output function F is binary and very useful in this method, since, being a classifier, a binary output is translated into a classification into two categories as follows:

If the network produces an output 1, the input belongs to class A.

If the network produces an output -1, the input belongs to class B.

In case of having two neurons at the input, the previous equation becomes:

$$w_1 x_1 + w_2 x_2 + \theta = 0 \tag{5}$$

B. Data base for the comparison of logistic regression and neural network models in the diagnosis of cancer

Factors: Age, cytology, HPV, biopsy, P16/K167, Tobacco; dx = diagnosis. The database is a filtered database of the reference of [7], see Table I.

III. RESULTS

A. The classification for the database of table 1 using Logistic regression

We have in the Table II a 100 of percentage of correct classification.

B. The classification for the database using Neural Networks

In summary we present table III and figures 2 and 3.

TABLE I
DATABASE FOR THE PREDICTION OF CANCER. DATABASE OF [7].

age	cytology	HPV	biposy	P16/K167	tobacco	dx
37	3	3	1	0	1	2,00
31	3	2	1	1	1	1,00

42	5	2	2.5	1	1	3,00
33	4	3	2.5	1	0	3,00
32	0	3	0	0	1	3,00
61	0	2	0	0	0	1,00
39	1	0	2.5	1	1	3,0
39	0	2	2.5	1	0	2,00
33	1	0	1	0	1	2,00
56	0	2	1	0	0	2,0
38	1	3	2.5	1	0	3,0
41	4	2	2.5	1	0	1,00
24	3	2	1	0	1	1,0
39	3	2	2.5	1	1	2,0
31	3	2	2.5	1	1	3,00
44	1	3	0	1	1	3,0
33	0	3	0	1	0	1,0
49	0	2	2.5	1	0	2,0
50	1	0	0	1	0	1,00
34	4	2	1	1	0	1,0
42	0	3	2.5	1	1	2,0
48	5	0	0	0	0	1,0
58	0	3	1	1	1	3,0
35	0	3	1	1	1	1,0
38	1	2	2.5	1	1	3,0
39	0	1	0	0	0	1,0
39	1	2	2.5	1	0	2,0
26	2	3	2.5	1	0	3,00
31	3	3	2	1	0	2,00
35	0	3	2	1	1	2,00

TABLE II

Table of classification of the prediction analyzed using the database of table I of cervical cancer

Predicted

Observed	NEGATIVE	CIN I	CIN II-III	Correct percentage
Negative	10	0	0	100,0 %
CIN I	0	10	0	100,0 %
CIN II-III	0	0	10	100,0 %
Global percentage	33,3 %	33,3 %	33,3 %	100,0 %

TABLE III
Classification table for the database of table I using Back propagation

		Predicted			
Example	Observed	NEGATIVE	CIN I	CIN II-III	Correct percentage
Training	NEGATIVE	4	2	1	57, 1 %
	CIN I	1	5	0	83, 3 %
	CIN II-III	4	1	1	16, 7 %
	Global percentage	47, 4 %	42, 1 %	10, 5 %	52, 6 %
Test	NEGATIVE	3	1	0	75, 0 %
	CIN I	0	3	0	100, 0 %
	CIN II-III	0	2	0	0, 0 %
	Global percentage	33, 3 %	66, 7 %	0, 0 %	66, 7 %

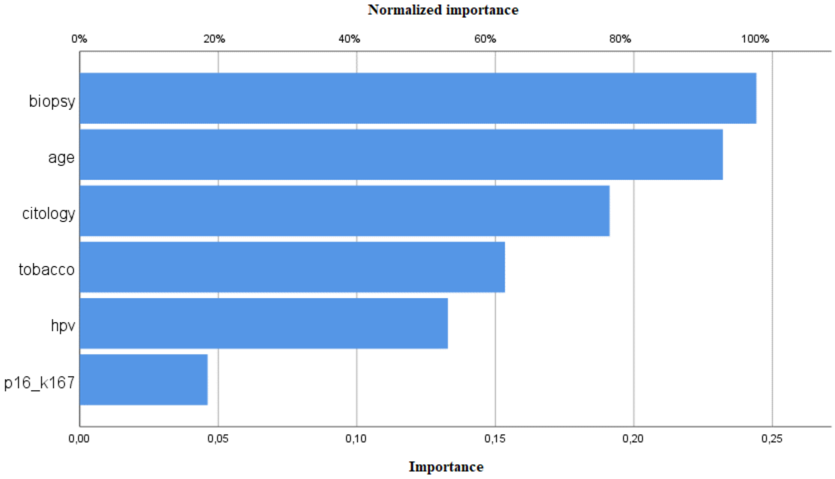
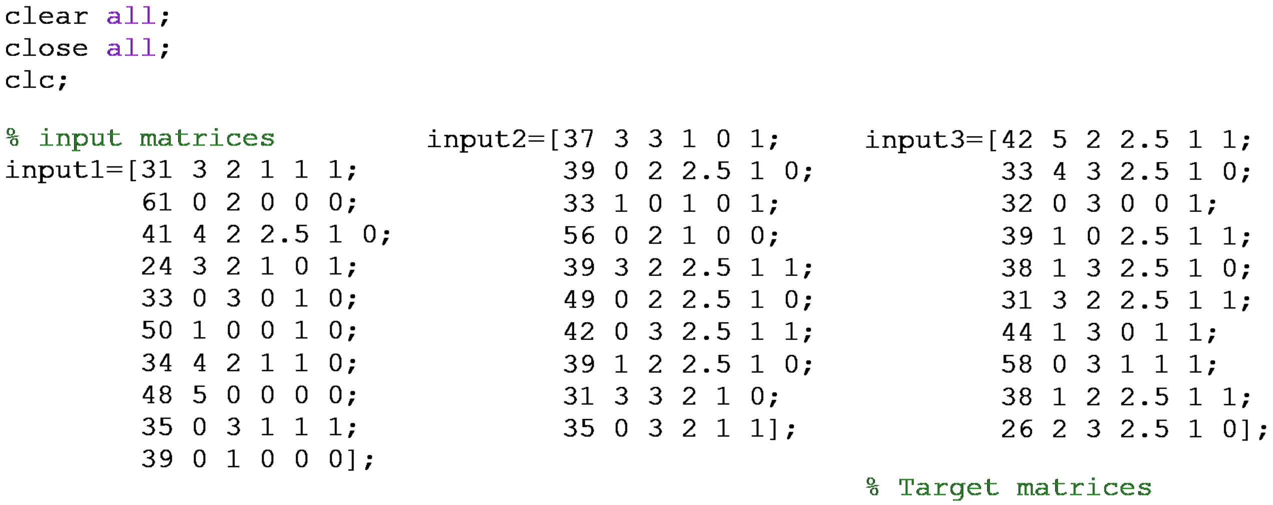


Fig. 2 Importance of the risk factors analyzed using the database of Table I of cervical cancer.



```
clear all;
close all;
clc;

% input matrices
input1=[31  3  2  1  1  1;
        61  0  2  0  0  0;
        41  4  2  2.5  1  0;
        24  3  2  1  0  1;
        33  0  3  0  1  0;
        50  1  0  0  1  0;
        34  4  2  1  1  0;
        48  5  0  0  0  0;
        35  0  3  1  1  1;
        39  0  1  0  0  0];

input2=[37  3  3  1  0  1;
        39  0  2  2.5  1  0;
        33  1  0  1  0  1;
        56  0  2  1  0  0;
        39  3  2  2.5  1  1;
        49  0  2  2.5  1  0;
        42  0  3  2.5  1  1;
        39  1  2  2.5  1  0;
        31  3  3  2  1  0;
        35  0  3  2  1  1];

input3=[42  5  2  2.5  1  1;
        33  4  3  2.5  1  0;
        32  0  3  0  0  1;
        39  1  0  2.5  1  1;
        38  1  3  2.5  1  0;
        31  3  2  2.5  1  1;
        44  1  3  0  1  1;
        58  0  3  1  1  1;
        38  1  2  2.5  1  1;
        26  2  3  2.5  1  0];

% Target matrices
```

Fig. 3 Risk factors analyzed using the MATLAB-Excel of Table I of cervical cancer.

IV. CONCLUSIONS

A classification table can be calculated by creating algorithms as done by the author. Likewise, in MATLAB we have elaborated a similar program to compare all the results according to the figure 3.

We have obtained results of prediction according to the figure 2.

BIBLIOGRAPHIC REFERENCE

[1] T. Ayer, F. Chhatwal, O. Alagoz, Ch. Kahn, R. Woods and E. Burnside, "Comparison of Logistic Regression and Artificial Neural network Models in Breast Cancer Risk Estimation", *Informatics in Radiology*, 2010.

[2] X. Hou, G. Shen, L. Zhou, Y. Li, T. Wang and X. Ma, "Artificial Intelligence in Cervical Cancer Screening and Diagnosis",*Front Oncol*, 2022.

[3] O. Holmström, N. Lundin, H. Kaingu, N. Mbuuko, J. Mbete, F. Kinyua, S. Törnquist, M. Muinde, L. Krogerus, M. Lundin, V. Diwan and J. Lundin, "Point-of-Care Digital Cytology With Artificial Intelligence for Cervical Cancer Screening in a Resource-Limited Setting", *JAMA Netw Open*, 2021.

[4] National Cancer Institute, "AI dual-stain approach improved accuracy, efficiency of cervical cancer screening in NCI study", *NCI Press Release*, 2020.

[5] C-W. Wang, Y-A. Liou, Y-J. Lin, C-C. Chang, P-H. Chu, Y-C. Lee, C-H. Wang, T-K. Chao, "Artificial intelligence-assisted fast screening cervical high grade squamous intraepithelial lesion and squamous cell carcinoma diagnosis and treatment planning", *Sci Rep* 11, 16244, 2021.

[6] National Cancer Institute, "AI approach outperformed human experts in identifying cervical precancer", *NCI Press Release*, 2019.

[7] O. Viñas, "Red Neuronal artificial como modelo predictivo en una unidad de patología cervical", *Universidad de Valladolid*, Tesis doctoral, 2015.