International Journal of Religion
2024
Volume: 5| Number 10 | pp. 2528 – 2543
ISSN: 2633-352X (Print) | ISSN: 2633-3538 (Online)
ijor.co.uk
DOI: https://doi.org/10.61707/3nb1q484

Comparative Study of Statistical Techniques for Preliminary Diagnosis of Cancer Risk

HERNAN OSCAR CORTEZ GUTIÉRREZ[1], MILTON MILCIADES CORTEZ GUTIÉRREZ[2], VANESSA MANCHA ALVAREZ[3], CÉSAR MIGUEL GUEVARA LLACSA[4], LIV JOIS CORTEZ FUENTES RIVERA[5], CESAR ANGEL DURAND GONZALES[6], GIRADY IARA CORTEZ FUENTES RIVERA[7], BRAULIO PEDRO ESPINOZA FLORES[8] and MIGUEL ANGEL GIL FLORES[9]

Abstract

The purpose of the is to elaborate models for preliminary diagnosis using statistical techniques. We compare two models for the estimation of cervical cancer risks.This article aims to compare predictive models for cervical cancer using machine learning techniques. We set up classification tables to compare the overall correct prediction rates.The data used comes from 30 cases of cervical cancer. We fitted a Logistic Regression (LR) model and trained Artificial Neural Networks (ANNs). The multicollinearity problem, usually present in modeling with numerous predictive variables, was addressed with factor analysis and Pearson Correlations.The LR model and ANN model were evaluated based on their percentage of correct classifications. The LR model achieved an accuracy of 33.33%, while the ANN model achieved an accuracy of 16.67%.Based on the percentage of correct classification, the Logistic Regression model was superior to the Neural Networks for the cervical cancer dataset. This highlights the need for further exploration of different machine learning approaches and data preprocessing techniques to improve predictive performance for cervical cancer risks.

Keywords: *Neural Networks, Python, Prediction, Logistic regression, Cervical.*

INTRODUCTION

The prevalence of cervical cancer in Peru is a significant public health concern. According to the Human Papillomavirus and Related Cancers Fact Sheet 2023, Peru has a population of 12.8 million women aged 15 and above who are at risk of developing cervical cancer. Each year, approximately 4,270 women are diagnosed with cervical cancer, and 2,288 die from the disease. This translates to a crude incidence rate of 25.7 per 100,000 women and a crude mortality rate of 13.8 per 100,000 women. Additionally, the age-standardized incidence rate is 22.2 per 100,000 women, and the age-standardized mortality rate is 11.5 per 100,000 women.

Cervical cancer is the second most frequent cancer among women in Peru, and it is also the second most frequent cancer among women between the ages of 15 and 44. The prevalence of cervical HPV-16/18 infection is estimated to be 6.6% among women in the general population, and 65.9% of invasive cervical cancers are attributed to HPVs 16 or 18.

These statistics highlight the need for effective cervical cancer screening and prevention strategies in Peru. The country has implemented vaccination programs for human papillomavirus (HPV) since 2011, but there is still

[1] Facultad de Ciencias de la Salud, Universidad Nacional del Callao, Email: hocortezg@unacvirtual.edu.pe

[2] Facultad de Ciencias de la Salud, Universidad Nacional del Callao

[3] Facultad de Ciencias de la Salud, Universidad Nacional del Callao
[4] Facultad de Ciencias de la Salud, Universidad Nacional del Callao

[5] Facultad de Ciencias de la Salud, Universidad Nacional del Callao

[6] Facultad de Ciencias de la Salud, Universidad Nacional del Callao

[7] Facultad de Ciencias de la Salud, Universidad Nacional del Callao

[8] Facultad de Ciencias de la Salud, Universidad Nacional del Callao

[9] Facultad de Ciencias de la Salud, Universidad Nacional del Callao

a significant burden of cervical cancer due to various systemic barriers and inequities in access to healthcare services.

Computer models in medical diagnosis in cervical cancer are being developed to detect the presence of cancer early. We have as a reference (Ayer et al. 2010) that uses logistic regression models to estimate breast cancer risk factors.

The diagnostic techniques and cervical pathology of cervical cancer in Peru are crucial aspects of the country's cancer control efforts. Cervical cancer is a significant public health concern in Peru, with an age-adjusted annual incidence rate of 23.2 per 100,000 women, making it the second most frequent cancer among women. The country has implemented various diagnostic methods to detect and manage cervical cancer, including visual inspection with acetic acid (VIA) and the Papanicolaou test (PAP).

Visual inspection with acetic acid (VIA) is a low-cost, non-invasive technique used for cervical screening. It involves applying acetic acid to the cervix, which causes abnormal cells to turn white, allowing for visual detection of lesions. VIA has been shown to be effective in primary-care settings and is widely used in Peru.

The Papanicolaou test (PAP) is a more comprehensive diagnostic method that involves collecting a sample of cells from the cervix and examining them under a microscope for abnormalities. PAP is considered the gold standard for cervical cancer screening, but it requires specialized equipment and trained personnel. In Peru, PAP is commonly used in hospitals and clinics, although there is a need for more widespread implementation and training of healthcare professionals.

Cervical pathology in Peru is characterized by a high prevalence of abnormal cervical cytology, with a significant proportion of women presenting with cervical intraepithelial neoplasia (CIN) and invasive cervical cancer. The most common histological types of cervical cancer in Peru are squamous cell carcinoma and adenocarcinoma.

The diagnostic and management strategies for cervical cancer in Peru are influenced by various factors, including limited access to healthcare services, inadequate training of healthcare professionals, and a lack of resources for screening and treatment. Therefore, there is a need for continued research and development of effective diagnostic and management strategies to address the significant burden of cervical cancer in Peru.

The diagnostic techniques and cervical pathology of cervical cancer in Peru are crucial aspects of the country's cancer control efforts. Cervical cancer is a significant public health concern in Peru, with an age-adjusted annual incidence rate of 23.2 per 100,000 women, making it the second most frequent cancer among women. The country has implemented various diagnostic methods to detect and manage cervical cancer, including visual inspection with acetic acid (VIA) and the Papanicolaou test (PAP).

Visual inspection with acetic acid (VIA) is a low-cost, non-invasive technique used for cervical screening. It involves applying acetic acid to the cervix, which causes abnormal cells to turn white, allowing for visual detection of lesions. VIA has been shown to be effective in primary-care settings and is widely used in Peru.

The Papanicolaou test (PAP) is a more comprehensive diagnostic method that involves collecting a sample of cells from the cervix and examining them under a microscope for abnormalities. PAP is considered the gold standard for cervical cancer screening, but it requires specialized equipment and trained personnel. In Peru, PAP is commonly used in hospitals and clinics, although there is a need for more widespread implementation and training of healthcare professionals.

Cervical pathology in Peru is characterized by a high prevalence of abnormal cervical cytology, with a significant proportion of women presenting with cervical intraepithelial neoplasia (CIN) and invasive cervical cancer. The most common histological types of cervical cancer in Peru are squamous cell carcinoma and adenocarcinoma.

The diagnostic and management strategies for cervical cancer in Peru are influenced by various factors, including limited access to healthcare services, inadequate training of healthcare professionals, and a lack of resources for screening and treatment. Therefore, there is a need for continued research and development of effective diagnostic and management strategies to address the significant burden of cervical cancer in Peru.

LITERATURE REVIEW

To use Python, follow the instructions of Colab - Google

https://colab.research.google.com/github/RFajardoMonzon/MachineLearningCourse/blob/master/Logistic_regression_Regresi%C3%B3n_Log%C3%ADstica.ipynb

```
# Importing necessary libraries for data handling, machine learning, and visualization
import pandas as pd  # For handling data in DataFrame format
from sklearn.model_selection import train_test_split  # For splitting data into training and testing sets
from sklearn.preprocessing import StandardScaler  # For standardizing features
from sklearn.linear_model import LogisticRegression  # For Logistic Regression model
from sklearn.metrics import classification_report, accuracy_score, confusion_matrix  # For evaluating the model
import tensorflow as tf  # For working with TensorFlow and building Neural Networks
from tensorflow.keras.models import Sequential  # For initializing a neural network model
from tensorflow.keras.layers import Dense  # For creating layers in a neural network
import matplotlib.pyplot as plt  # For plotting graphs and charts
import seaborn as sns  # For advanced data visualization
import numpy as np  # For numerical operations
# Define the dataset using a dictionary, where each key represents a feature and values are lists of data points
data = {
    "age": [37, 31, 42, 33, 32, 61, 39, 39, 33, 56, 38, 41, 24, 39, 31, 44, 33, 49, 50, 34, 42, 48, 58, 35, 38, 39, 39, 26, 31, 35],
    "cytology": [3, 3, 5, 4, 0, 0, 1, 0, 1, 0, 1, 4, 3, 3, 3, 1, 0, 0, 1, 4, 0, 5, 0, 0, 1, 0, 1, 2, 3, 0],
    "HPV": [3, 2, 2.5, 3, 3, 2, 0, 2, 0, 2, 3, 2, 2, 2, 2.5, 3, 0, 2, 0, 2, 2.5, 0, 3, 3, 2.5, 1, 2.5, 3, 2, 3],
    "biopsy": [1, 1, 2.5, 2.5, 0, 0, 2.5, 2.5, 1, 1, 2.5, 2.5, 1, 2.5, 2.5, 0, 0, 2.5, 0, 1, 2.5, 0, 1, 1, 2.5, 0, 2.5, 2.5, 2, 2],
    "P16/K167": [0, 1, 1, 1, 0, 0, 1, 1, 0, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1],
    "tobacco": [1, 1, 1, 0, 1, 0, 1, 0, 1, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 1, 0, 1, 1, 1, 0, 0, 0, 0, 1],
    "dx": [2.0, 1.0, 3.0, 3.0, 3.0, 1.0, 3.0, 2.0, 2.0, 2.0, 3.0, 1.0, 1.0, 2.0, 3.0, 3.0, 1.0, 2.0, 1.0, 1.0, 2.0, 1.0, 3.0, 1.0, 3.0, 1.0, 2.0, 3.0, 2.0, 2.0]
}
# Create a DataFrame from the dictionary
df = pd.DataFrame(data)
# Separate the features (X) and the target variable (y)
X = df.drop("dx", axis=1)  # Features
y = df["dx"]  # Target variable
# Split the dataset into training and testing sets
# test_size=0.2 means 20% of the data will be used for testing, and 80% for training
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
# Standardize the features to have a mean of 0 and standard deviation of 1
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)  # Fit and transform the training data
X_test_scaled = scaler.transform(X_test)  # Only transform the testing data
# Initialize and train the Logistic Regression model
lr_model = LogisticRegression(max_iter=1000)  # max_iter=1000 means the model will iterate up to 1000 times for optimization
lr_model.fit(X_train_scaled, y_train)  # Train the model with the training data
# Predict the target variable for the test data using the Logistic Regression model
y_pred_lr = lr_model.predict(X_test_scaled)
# Evaluate the Logistic Regression model
print("Logistic Regression Classification Report:")
print(classification_report(y_test, y_pred_lr))  # Print detailed classification report
print("Logistic Regression Accuracy:", accuracy_score(y_test, y_pred_lr))  # Print the accuracy score
# Compute the confusion matrix for the Logistic Regression model
cm_lr = confusion_matrix(y_test, y_pred_lr)
# Initialize and train the Artificial Neural Network (ANN) model
ann_model = Sequential()  # Initialize the model
ann_model.add(Dense(12, input_dim=X_train_scaled.shape[1], activation='relu'))  # Add input layer and first hidden layer
ann_model.add(Dense(8, activation='relu'))  # Add second hidden layer
ann_model.add(Dense(1, activation='sigmoid'))  # Add output layer
# Compile the ANN model with loss function, optimizer, and metrics
ann_model.compile(loss='binary_crossentropy', optimizer='adam', metrics=['accuracy'])
# Train the ANN model with the training data
ann_model.fit(X_train_scaled, y_train, epochs=100, batch_size=5, verbose=0)
# Predict the target variable for the test data using the ANN model
y_pred_ann = (ann_model.predict(X_test_scaled) > 0.5).astype(int)
# Evaluate the ANN model
print("ANN Classification Report:")
print(classification_report(y_test, y_pred_ann))  # Print detailed classification report
print("ANN Accuracy:", accuracy_score(y_test, y_pred_ann))  # Print the accuracy score
# Compute the confusion matrix for the ANN model
cm_ann = confusion_matrix(y_test, y_pred_ann)
```

```
# Compare the performance of both models based on accuracy
accuracy_lr = accuracy_score(y_test, y_pred_lr)
accuracy_ann = accuracy_score(y_test, y_pred_ann)
print(f"Logistic Regression Accuracy: {accuracy_lr * 100:.2f}%")
print(f"Artificial Neural Network Accuracy: {accuracy_ann * 100:.2f}%")
# Plotting the results
# Plot Confusion Matrices for both models
fig, ax = plt.subplots(1, 2, figsize=(14, 5))  # Create subplots
sns.heatmap(cm_lr, annot=True, fmt='d', ax=ax[0], cmap='Blues')  # Plot confusion matrix for Logistic Regression
ax[0].set_title('Logistic Regression Confusion Matrix')  # Set title for the plot
ax[0].set_xlabel('Predicted')  # Set x-axis label
ax[0].set_ylabel('Actual')  # Set y-axis label
sns.heatmap(cm_ann, annot=True, fmt='d', ax=ax[1], cmap='Blues')  # Plot confusion matrix for ANN
ax[1].set_title('ANN Confusion Matrix')  # Set title for the plot
ax[1].set_xlabel('Predicted')  # Set x-axis label
ax[1].set_ylabel('Actual')  # Set y-axis label
plt.show()  # Display the plots
# Plot Histograms for the features to visualize distributions
df.hist(bins=10, figsize=(14, 10), grid=False)  # Plot histograms
plt.suptitle('Feature Distributions', fontsize=16)  # Set the title for the plot
plt.show()  # Display the plot
# Scatter plot matrix to see pairwise relationships between features
pd.plotting.scatter_matrix(df, figsize=(15, 10), diagonal='kde')  # Plot scatter matrix
plt.suptitle('Scatter Matrix of Features', fontsize=16)  # Set the title for the plot
plt.show()  # Display the plot
# Plotting the likelihood for each parameter in Logistic Regression
coefficients = lr_model.coef_[0]
features = X.columns
likelihoods = np.exp(coefficients) / (1 + np.exp(coefficients))
plt.figure(figsize=(10, 6))
plt.bar(features, likelihoods, color='skyblue')
plt.xlabel('Features')
plt.ylabel('Likelihood')
plt.title('Likelihood for Each Parameter in Logistic Regression')
```

```
plt.xticks(rotation=45)

plt.show()
```

The ENDES (Encuesta Demográfica y de Salud Familiar) is Peru's Demographic and Family Health Survey, which collects data on various health indicators, including cervical cancer screening. Here are the key findings on cervical cancer trends in Peru from 2015 to 2020 based on ENDES data:

From 2015 to 2019, the average number of Pap smears performed annually was 59,171.5 ± 8,898.7. However, in 2020, only 16,273 (4.58%) Pap tests were conducted, with a monthly mean of 1,356.1 ± 684.2 (95% CI: 149.7 to 2,861.9), representing a 76.7% reduction in cervical cancer screening during the COVID-19 pandemic.

In 2020, 78.5% of women reported ever having a Pap smear, with higher rates among those aged 35-44 years.

The prevalence of high-risk HPV infection, a major risk factor for cervical cancer, ranges from 17.7% among cytologically normal women in Lima to 23.4% in rural areas of Peru.

Cervical cancer incidence in Peru is 23.2 per 100,000 women, with a mortality rate of 10.2 per 100,000 women, significantly higher than the rest of South America.

Barriers to effective cervical cancer prevention in Peru include low screening rates, poor follow-up of abnormal results, and misclassification of Pap smears.

In conclusion, while cervical cancer screening rates in Peru have been suboptimal, the COVID-19 pandemic has further exacerbated the situation, leading to a dramatic decline in Pap smear testing in 2020. Addressing barriers to screening and implementing effective prevention strategies, such as HPV testing and vaccination, are crucial to reducing the burden of cervical cancer in Peru.

Multiple 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:

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

Artificial intelligence (AI) has been used in various studies to improve cervical cancer screening and diagnosis.

Logistic regression is a statistical technique used to model the relationship between a categorical dependent variable and one or more independent variables. It is commonly used in various fields, including medicine, social sciences, and business, to predict the likelihood of an event occurring based on specific conditions. This essay will discuss the key concepts and applications of binary and multinomial logistic regression, highlighting their differences and similarities.

Binary logistic regression is used when the dependent variable is dichotomous, meaning it can only take on two values (e.g., 0/1, yes/no, or present/absent). This type of regression is particularly useful for predicting the probability of an event occurring or not occurring based on a set of predictor variables. The model is written as:

$$log\ \left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X$$

where P is the probability of the event occurring, β_0 is the intercept, β_1 is the coefficient for the independent variable X, and log is the natural logarithm.

Multinomial logistic regression, on the other hand, is used when the dependent variable is nominal with more than two categories. This type of regression is useful for modeling the relationship between a categorical dependent variable with multiple categories and one or more independent variables. The model is written as:

$$log\left(\frac{P_i}{P_j}\right) = \beta_0 + \ \beta_1 X$$

}where Pi and Pj are the probabilities of the ith and jth categories, respectively, and $\beta 0$ and $\beta 1$ are the intercept and coefficient for the independent variable X, respectively.

Both binary and multinomial logistic regression models rely on certain assumptions to ensure the accuracy of the results. These assumptions include the independence of observations, no multicollinearity among the predictor variables, and a sufficient number of observations to support the model. The output of the models typically includes the coefficients, standard errors, Wald statistics, and p-values, which are used to determine the significance of the relationships between the variables. The coefficients can be interpreted as the change in the log odds of the event occurring for a one-unit change in the independent variable, while the odds ratios provide the relative change in the odds of the event occurring for a one-unit change in the independent variable

2

Binary logistic regression has been widely used in various fields to predict the likelihood of an event occurring. For instance, it has been applied to predict the likelihood of a patient developing a certain disease based on demographic and medical factors. Multinomial logistic regression, on the other hand, has been used to model the relationship between a categorical dependent variable with multiple categories and one or more independent variables. For example, it has been used to predict the likelihood of a customer choosing a particular product based on demographic and marketing factors.

In 2018, Peru had an age-standardized cervical cancer mortality rate of 11.5 per 100,000 women, ranking it 2nd among all cancers affecting women of all ages and 1st among women aged 15-44

Cervical cancer incidence in Peru was 22.2 per 100,000 women in 2018, with 4,270 new cases diagnosed annually.

HPV prevalence in Peru was 6.6% among women with normal cytology, 27.3% among those with low-grade cervical lesions, 53.1% with high-grade lesions, and 65.9% of invasive cervical cancers were attributed to HPV 16/18.

From 2008 to 2017, cervical cancer mortality in Peru showed a significant downward trend, decreasing from 11.62 to 9.69 per 100,000 women (APC = −2.2, 95% CI: −4.3, −0.1, $p<0.05$)

Mortality rates remained highest in the rainforest region, declining from 34.16 in 2008 to 17.98 in 2017 (APC = −4.3, 95% CI: −7.2, −1.3, $p<0.01$).

Cervical cancer screening coverage in Peru was 58% among women aged 25-65 ever screened, 61% screened in the last 3 years, and 75% of those aged 30-49 screened in the last 5 years as of 2019.

The main factors contributing to the high cervical cancer mortality rates in Peru are:

Socioeconomic disparities and lack of access to health care services, particularly in the rainforest region, which has the highest mortality rates. The rainforest region has a high density of indigenous women who face barriers to accessing screening and treatment. Low education and knowledge about cervical cancer prevention in underserved regions, resulting in poor adherence to screening and treatment. Cervical cancer screening coverage in Peru was only 58% among women aged 25-65 ever screened as of 2019

HPV prevalence is high in Peru, with 65.9% of invasive cervical cancers attributed to HPV 16/18

However, HPV vaccination coverage remains low. Underreporting of cervical cancer deaths, especially in regions with poor health systems like the rainforest department of Loreto where the death registration omission rate is as high as 78%.

Cervical cancer disproportionately affects women of reproductive age in Peru, with it ranking as the 2nd most frequent cancer in women aged 15-44. Late-stage diagnosis and lack of access to treatment in this age group contributes to high mortality.

In this section, the review of theoretical and empirical literature related to the variables of this research was addressed, such as cervical cancer, artificial neural network and risks estimation.

A study (Xou et al. 2022) 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 (Holmström et al. 2021) 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 (National Cancer Institute 2020) 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 (Wang et al. 2021) 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 (National Cancer Institute 2019) 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.

METHODOLOGY AND MODEL

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.

.

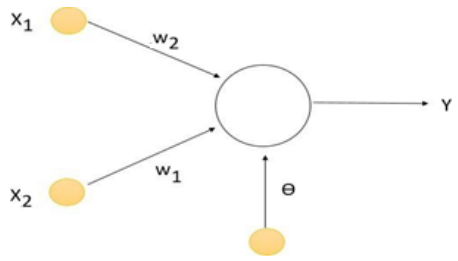


Figure 1. Perceptron architecture with two inputs and one output.

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.

$$y=\sum_{i=1}^{n} [\![w_i x_i \quad (1)]\!]$$

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

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

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

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

$$y=F(\sum_{i=1}^{n} [\![w_i x_i+]\!] \ \theta) \quad (3)$$

Where F is an independent variable.

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

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:

$$w_1 x_1+w_2 x_2+\theta=0 \quad (5)$$

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 (Viñas 2015).

Table 1. Degrees assigned to the DIAGNOSTIC variable

DIAGNÓSTICO	GRADO ASIGNADO
NEGATIVO	0
CIN I	1
CIN II	2
CIN II-III	2,5
CIN III	3
ADENOCARCINOMA	4
CA ESCAMOSO	4

Table 2. Database for the prediction of cancer. Filtered database of the reference of the author Olga Viñas Aparicio

age	Cytology	HPV	biopsy	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

RESULTS

RESULTS USING PYTHON

Logistic Regression Classification Report:

	precision	recall	f1-score	support
1.0	0.00	0.00	0.00	1
2.0	0.50	0.33	0.40	3
3.0	0.50	0.50	0.50	2
accuracy			0.33	6
macro avg	0.33	0.28	0.30	6
weighted avg	0.42	0.33	0.37	6

Logistic Regression Accuracy: 0.3333333333333333

1/1 [==============================] - 0s 65ms/step

ANN Classification Report:

	precision	recall	f1-score	support
1.0	0.17	1.00	0.29	1
2.0	0.00	0.00	0.00	3
3.0	0.00	0.00	0.00	2
accuracy			0.17	6
macro avg	0.06	0.33	0.10	6
weighted avg	0.03	0.17	0.05	6

ANN Accuracy: 0.16666666666666666

Logistic Regression Accuracy: 33.33%

Artificial Neural Network Accuracy: 16.67%

/usr/local/lib/python3.10/dist-packages/sklearn/metrics/_classification.py:1344: UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero_division` parameter to control this behavior.

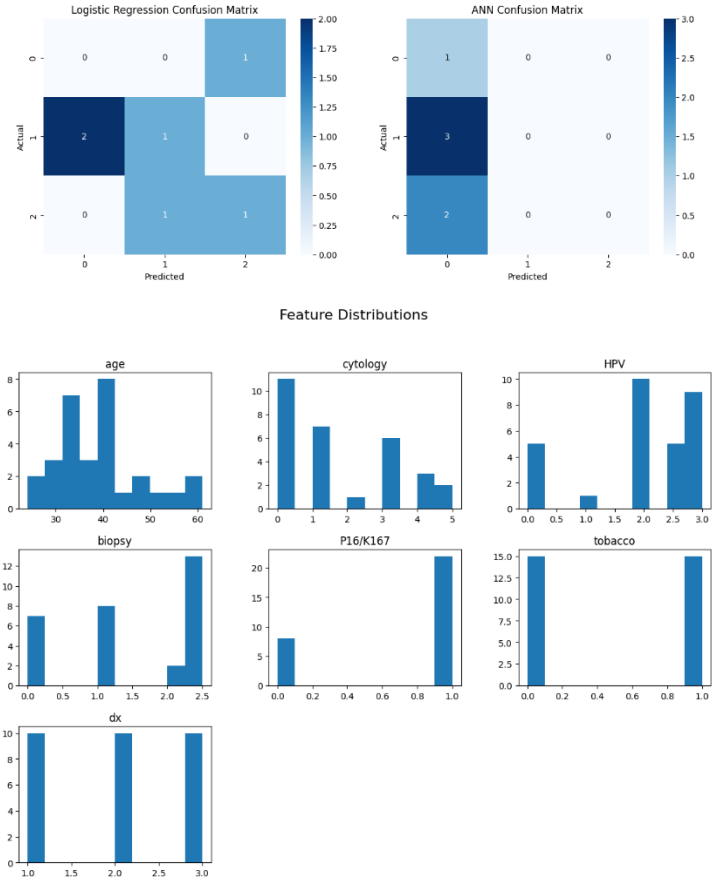
_warn_prf(average, modifier, msg_start, len(result))

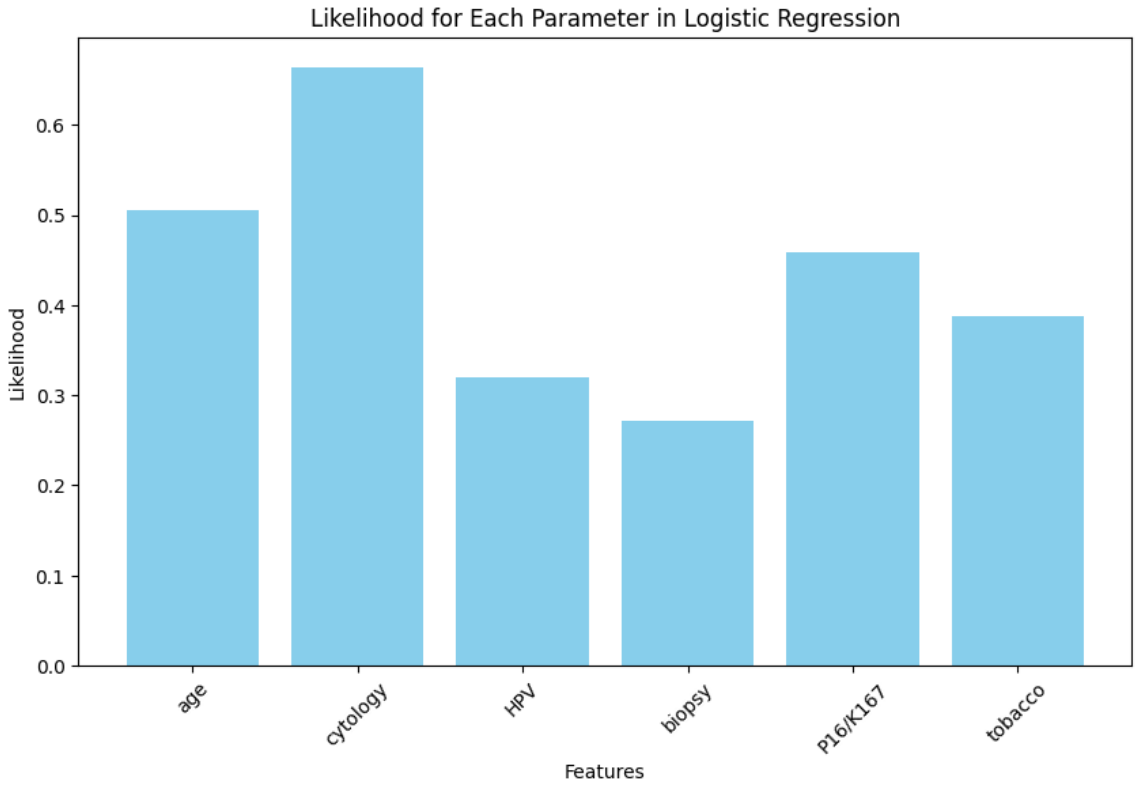
/usr/local/lib/python3.10/dist-packages/sklearn/metrics/_classification.py:1344: UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero_division` parameter to control this behavior.

_warn_prf(average, modifier, msg_start, len(result))

/usr/local/lib/python3.10/dist-packages/sklearn/metrics/_classification.py:1344: UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero_division` parameter to control this behavior.

_warn_prf(average, modifier, msg_start, len(result).





The classification for the database of figure 6 using Logistic regression

We have in the Table 2 a 100 of percentage of correct classification.

Table 3. Table of classification of the prediction analyzed using the database of figure 6 of cervical cancer

		Predicted		
Observed	NAGATIVE	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%

The classification for the database using Neural networks

Table 4. Classification table for the database of figure 4 using Back propagation

		Predicted			
Example	Observed	NAGATIVE	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%
Tests	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%

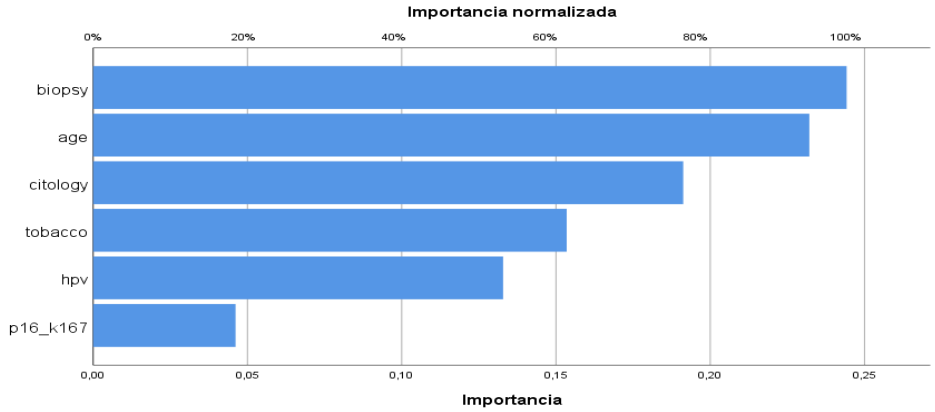


Figure 1. Importance of the risk factors analyzed using the database of Table 1 of cervical cancer.

```
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
```

Figure 2. Risk factors analyzed using the MATLAB-Excel of Table 1 of cervical cancer

DISCUSSIONS AND CONCLUSIONS

Logistic regression and artificial neural networks are two widely used techniques for making predictions in the field of medicine. In the context of predicting cancer in Peru, both methods have been applied and compared to determine their effectiveness.

Logistic regression is a statistical model that is used to predict the probability of a binary outcome based on a set of independent variables. It assumes a linear relationship between the independent variables and the logarithm of the odds of the outcome. Logistic regression has been preferred for statistical models where the outcome variables are dichotomous (binary).

On the other hand, artificial neural networks are composed of interconnected layers of nodes, which can learn from data and make predictions. They have the ability to model complex non-linear relationships between variables.

In a study conducted in Peru, the performance of logistic regression and artificial neural networks was compared for predicting skin cancer in dogs. The variables analyzed included age, sex, breed, sun exposure, albinism, and dermatitis. The results showed that the backpropagation neural network technique with cross-validation outperformed the logistic regression model, with an accuracy of 89.6% for the neural network compared to 84% for logistic regression.

Another study compared logistic regression and artificial neural networks for predicting child labor in Peru. The results of this study were not provided in the given search results.

Additionally, studies conducted in other countries have also compared the performance of logistic regression and artificial neural networks for predicting various cancer-related outcomes. These studies have shown that artificial neural networks can outperform logistic regression in certain scenarios, particularly when dealing with complex non-linear relationships.

In conclusion, both logistic regression and artificial neural networks have been used for predicting cancer-related outcomes in Peru and other countries. While logistic regression is a simpler and more interpretable method, artificial neural networks have shown superior performance in some cases, especially when dealing with complex data. The choice between the two methods depends on the specific problem, the available data, and the desired level of interpretability.

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 3.

ACKNOWLEDGEMENTS

The authors would like to thank everyone, just everyone.

REFERENCES

Abuelo, C. E., Levinson, K. L., Salmeron, J., Sologuren, C. V., Fernandez, M. J. V., & Belinson, J. L. (2014). The Peru cervical cancer screening study (PERCAPS): The design and implementation of a mother/daughter screen, treat, and vaccinate program in the Peruvian jungle. Journal of Community Health, 39(3), 409–415. https://doi.org/10.1007/s10900-013-9786-6

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

Becerra-Canales, B., Campos, M., Atuncar-Deza, S., & Cáceres-Yparraguirre, H. (2023). Prevalence and factors associated with cervical cancer preventive screening in a Peruvian region. Medwave, 23(08), e2709–e2709. https://doi.org/10.5867/medwave.2023.08.2709

Bendezu-Quispe, G., Soriano-Moreno, A. N., Urrunaga-Pastor, D., Venegas-Rodríguez, G., & Benites-Zapata, V. A. (2020). Asociación entre conocimientos acerca del cáncer de cuello uterino y realizarse una prueba de Papanicolau en mujeres

peruanas. Revista Peruana de Medicina Experimental y Salud Publica, 37(1), 17–24. https://doi.org/10.17843/rpmesp.2020.371.4730

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

Hou X., Shen G., Zhou L., Li Y., Wang T. and Ma X. (2022). Artificial Intelligence in Cervical Cancer Screening and Diagnosis. Front Oncol.

ICO/IARC Information Centre on HPV and Cancer. (2023, octubre 3). Peru Human Papillomavirus and Related Cancers, Fact Sheet 2023. Hpvcentre.net. https://hpvcentre.net/statistics/reports/PER_FS.pdf

Instituto Nacional de Estadística e Informática. INEI. (2021). Perú enfermedades noi transmisibles y transmisibles 2020. Gob.pe. https://proyectos.inei.gob.pe/endes/2020/SALUD/ENFERMEDADES_ENDES_2020.pdf

León-Nakamura, C., & Yábar-Berrocal, A. (2019). Características del tamizaje para cáncer cérvico-uterino en 08 establecimientos de salud, Lima Metropolitana 2017. Revista de la Facultad de Medicina Humana, 19(1), 1–5. https://doi.org/10.25176/rfmh.v19.n1.1788

LibGuides: Statistics Resources: Multinomial Logistic Regression. (s. f.). https://resources.nu.edu/statsresources/Multinomiallogistic

Ministerio de Salud (2017). Guía de práctica clínica para la prevención y manejo del cáncer de cuello uterino. Gob.pe. https://bvs.minsa.gob.pe/local/MINSA/4146.pdf

National Cancer Institute (2019). AI approach outperformed human experts in identifying cervical precancer. NCI Press Release.

National Cancer Institute (2020). AI dual-stain approach improved accuracy, efficiency of cervical cancer screening in NCI study. NCI Press Release.

Norma Técnico-Oncológica para la prevención, detección y manejo de la lesiones premalignas del cuello uterino a nivel nacional. (s/f). https://www.irennorte.gob.pe/. Recuperado el 31 de mayo de 2024, de https://www.irennorte.gob.pe/pdf/doctec/d0003.pdf

Thoumi, A., Bond, S. J., Dotson, M. E., Krieger, M., Garcia, P. J., & Ramanujam, N. (2021). Policy considerations to promote equitable cervical cancer screening and treatment in Peru. Annals of global health, 87(1), 116. https://doi.org/10.5334/aogh.3442

Torres-Roman, J. S., Ronceros-Cardenas, L., Valcarcel, B., Arce-Huamani, M. A., Bazalar-Palacios, J., Ybaseta-Medina, J., La Vecchia, C., & Alvarez, C. S. (2021). Cervical cancer mortality in Peru: regional trend analysis from 2008–2017. BMC Public Health, 21(1). https://doi.org/10.1186/s12889-021-10274-1

Viñas O. (2015). Red Neuronal artificial como modelo predictivo en una unidad de patología cervical. Universidad de Valladolid, Tesis doctoral.

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

Colab – Google (2024) Clasificación Logística. Adquirido 15 de junio del 2024. https://colab.research.google.com/github/RFajardoMonzon/MachineLearningCourse/blob/master/Logistic_regression_Regresi%C3%B3n_Log%C3%ADstica.ipynb.