Available online at http://scik.org

J. Math. Comput. Sci. 11 (2021), No. 4, 3916-3926

https://doi.org/10.28919/jmcs/5812

ISSN: 1927-5307

A COMPARISON BETWEEN CLASSIFICATION STATISTICAL MODELS AND NEURAL NETWORKS WITH APPLICATION ON PALESTINE DATA

AMANI MOUSSA MOHAMED, MAHMOUD A. ABDEL-FATTAH, ABDALLAH SALMAN MOHAMMED ALDIRAWI*

Department of Applied Statistics and Econometrics,

Faculty of Graduate Studies for Statistical, Cairo University, Cairo, Egypt

Abstract: The paper has used labor force as dependent variable which contains two categories (Employment and Unemployment) and 8 independent variables. The results regarding the application of the correct classification technique to assess the accuracy of the three classification methods in predicting the labor force of have shown it was found that Artificial Neural Networks gave the best accuracy in prediction with (82.7%), 79.5% for Discriminant Analysis and (81.6%) for Logistic Regression. Furthermore, ROC curve technique has been applied to evaluate the accuracy of the three classification methods in predicting the labor force. It has been found that Artificial Neural Networks gave the best accuracy in prediction with (85.5%), (72.8%) for Discriminant Analysis and (81.7%) for Logistic Regression. In addition, Artificial Neural Network gave the best results in prediction with 82.7% accuracy, and less error rate with 0.173. Meanwhile, the Discriminant analysis model has shown 79.5% accuracy, and 0.205 error rate. Logistic Regression has shown 81.5% accuracy, 69.8% sensitivity and 0.183 error rate. These results demonstrate that Artificial Neural Network could be the most powerful analytical technique for the variables with two

*Corresponding author

E-mail address: abdullah.s.aldirawi@gmail.com

Received April 4, 2021

categories.

Keywords: classification; artificial neural networks; logistic regression and the discriminant analysis.

2010 AMS Subject Classification: 92B20.

1. INTRODUCTION

The task of classification occurs in a wide range of human activity. At its broadest, the term could cover any context in which some decision or forecast is made based on currently available information. Then a classification procedure is a formal method for repeatedly making such judgments in new situations. It is assuming that the problem concerns the construction of a procedure that will be applied to a continuing sequence of cases, in which each new case must be assigned to one of a set of pre-defined classes based on observed attributes or features. The classification methods are used to categorize certain data of statistical community on different groups based on one or more of the basic properties of these data. The nature of data help or restrict it to choose the best classification method. It is meaningful to address how the analyst can deal with data representing multiple independent variables and a categorical dependent variable, how independent variables can contribute to the discovery of differences in the categories. The assignment of observations or objects into predefined homogenous groups is a problem of major practical and research interest.

Comparison between observations is considered as one of the common methods used, due to the large number of applied phenomena which can be analyzed by way of comparison between observations. There are many methods that can be used to compare between observations such as the Discriminant Analysis (AD) and Logistic Regression (LR). There is another method used for comparison between observations, which is the Artificial Neural Networks (ANN).

The statistical analysis to Labor Force (LF) data used LR, DA and ANN. The data consist of 8 independent variables and one dependent variable with two categories (Employment and Unemployment). The goal is to find the best model according to model selection criteria. Receiver Operating Characteristic (ROC) curve will validate the model.

2. LITERATURE REVIEW

[1] compared Linear Discriminant Analysis (LDA) and Multinomial Logistic Regression (MLR) to make the choice between the two methods easier, and to understand how the two models behave under different data and group characteristics. The performance evaluation was carried out on the same real-world dataset. He also performed a simulation study according to the linear discriminant analysis model, where the multivariate normality is satisfied with the same covariance matrix to examine the group and data characteristics that may affect the performance of LDA and MLR. Both Logistic Regression (LR) and LDA converged in similar results. And estimated the same statistically significant coefficients, and either can be helpful in classifying the class membership of women that diabetics. LR slightly exceeds LDA in the correct classification rate, but when considering sensitivity, specificity and AUC the differences in the AUC were negligibly, thus indicating no discriminating difference between the models. The simulation study examined the impact of changes regarding the sample size, distance between group means, categorization and correlation matrices between the predictors on the performance of each method. Results indicate that the variation in sample size, values of Euclidean distance, different number of categories had similar impact on the result for the two methods, and both methods LDA and MLR show significant improvement in classification accuracy in the absence of multicollinearity among the explanatory variables. [2] aimed at evaluating the performance of the main estimation methods and algorithms for building reliable multinomial logistic regression models. Seven estimation methods and algorithms are compared using different assessment techniques to arrive at a reliable multinomial logistic regression model for a given dataset. The result is that the ridge multinomial regression method proves to be the most reliable method with the highest area under the receiver operating characteristic (ROC), and the lowest error rate for classifying children and identifying significant risk factors on anemia status among all other methods. A detailed description of the results of applying this method to a real dataset from a survey, conducted by the Palestinian Bureau of Statistics to classify children of less than five years of age (2010–2011) according to their anemia status, is illustrated. Ten independent variables from the survey are selected and used to

classify children according to their anemia status (normal child, mild anemia, moderate anemia, and severe anemia), a reliable multinomial regression model is built, and important risk factors of these anemia statuses are identified. [3] Stated that the third most serious disease estimated by Word Wide Organization after cancer and cardiovascular disease is the infertility. The advanced treatment techniques are the Intra-Cytoplasmic Sperm Injection (ICSI) procedure, it represents the best chance to have a baby for couples having an infertility problem. ICSI treatment is expensive, and there are many factors affecting the success of the treatment, including male and female factors. The paper aims to classify and predict the ICSI treatment results using logistic regression and artificial neural network. For this purpose, data are extracted from real patients and contain parameters such as age, endometrial receptivity, endometrial and myometrium vascularity index, number of embryo transfer, day of transfer, and quality of embryo transferred. Overall, the logistic regression predicts the output of the ICSI outcome with an accuracy of 75%. In other parts, the neural network managed to achieve an accuracy of 79.5% with all parameters and 75% with only the significant parameters. [4] addressed the problem of accurate medical diagnosis that is always urgent for any person. Existing methods for solving the problem of classification of the state of a complex system are considered. The paper proposed method of classification of patients' status in medical monitoring systems using artificial neural networks. The artificial neural networks training method uses bee colonies to simulate less training error. The research purpose is to determine the patient's belonging to a particular class according to the variables of his condition, which are recorded. Examples of using the method to determine the status of patients with urological diseases and liver disease are given. The classification accuracy was more than 80%.

3. DESCRIPTION OF LABOR FORCE DATA

Real data of labor force survey 2019 has been collected by Palestinian Center Bureau of Statistics (PCBS). The sample size is 8953 observations of whom 8953 are valid and no missing value 61.4% of them reside in the West Bank, and 38.6% one resides in Gaza Strip. Data set contains 9 variables. The interest is on labor force variable. This variable has been used as a dependent variable in this

analysis. It involves two categories (Employment and Unemployment). The goal is to find the best model which can describe the relationship between different types of labor force and other factors that can be considered as independent variables and have effects on each dependent variable. The target group used in this study is people living in West Bank Gaza Strip in the age 15 – 65 years. The dependent variable has two categories (Employment and Unemployment).

Table (1): Distribution of labor force status

Category	Number	Percent
Employment	6224	69.5
Unemployment	2729	30.5
Total	8953	100.0

Table (1) demonstrates that 69.5 % of persons are Employment, 30.5% are unemployment.

4. CLASSIFICATION METHODS

In Statistical classification we attempt to predict values of a categorical dependent variable from one or more continuous and/or categorical predictor variables. It is the process of allocating an observation (i) in one of several predefined groups or categories. An ideal classification method provides in what distinguishes different classes from each other. It deals with rules of case assignment to categories or classes, and the goal of classification, is to provide a model that yields the optimal discrimination between several classes in terms of predictive performance [5]. The assignment of alternatives observations or objects into predefined homogenous groups is a problem of major practical and research interest. In this chapter, the three classification methods namely discriminant analysis (DA), logistic regression (LR) and artificial neural networks (ANN).

4.1 LOGISTIC REGRESSION

Logistic regression is a statistical modeling technique designed for binary response variables, for which the response outcome of each subject is a “success” or “failure.” Binary data are the most common form of categorical data, and the most popular model for binary data is logistic regression

model [6].

Logistic regression (sometimes called the logistic model or logit model) is used for prediction of the probability of occurrence of an event by fitting data to a logistic function. Similar to other forms of regression analysis, it makes use of one or more predictor variables that may be either numerical or categorical. It is used extensively in the medical and social sciences fields, as well as marketing applications. It allows one to predict a discrete outcome, such as group membership, from a set of variables that may be continuous, discrete, dichotomous, or a mix of any of these. In binary logistic model, the dependent or response variable is dichotomous, such as presence absence or success/failure. We will present statistical inference for the binary logistic model parameters.

In binary logistic regression models, is usually dichotomous, that is, the dependent variable can take the value 1 with a probability of success θ, or failure with the value 0 with probability $1-\theta$ .This type of variable is called a Bernoulli (or binary) variable. Although not as common and not discussed in this treatment, applications of logistic regression have also been extended to cases where the dependent variable is of more than two cases, known as multinomial or polytomous [7] use the term polychotomous.

As mentioned previously, the independent or predictor variables in logistic regression can take any form. That is, logistic regression makes no assumption about the distribution of the independent variables. They do not have to be normally distributed, linearly related or of equal variance within each group. The relationship between the predictor and response variables is not a linear function in logistic regression, instead, the logistic regression function is used, which is the logit transformation of θ.

$$\theta = \frac{e^{(\alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k)}}{1 + e^{(\alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k)}} \qquad (1)$$

Where a= the constant of the equation and, β= the coefficient of the predictor variables.

An alternative form of the logistic regression equation can be obtained as:

$$\text{logit}(y) = \ln\left(\frac{p}{1-p}\right) = \alpha + \beta X_1 + \ldots + \beta X_k \qquad (2)$$

The goal of logistic regression is to correctly predict the category of outcome for individual cases using the most parsimonious model. To accomplish this goal, a model is created that includes all

predictor variables that are useful in predicting the response variable. Several different options are available during model creation. Variables can be entered into the model in the order specified by the researcher or logistic regression can test the fit of the model after each coefficient is added or deleted, in a procedure called a stepwise regression analysis.

4.2 DISCRIMINANT ANALYSIS

Discriminant function analysis is used to classify individuals into the predetermined groups. It is a multivariate analogue of analysis of variance and can be considered as a posterior procedure of multivariate analysis of variance, if discriminant function analysis is effective for a set of data, the classification table of correct and incorrect estimates will yield a high percentage correct. Multiple discriminant function analysis (sometimes called canonical variate analysis) is used when there are three or more groups. In DA multiple quantitative attributes are used to discriminate single classification variable, Discriminant analysis (in the broad sense) is a very powerful statistical tool for many types of analyses [8,9].

The goal of discriminant analysis is to construct the model based on observational unit ‘s variation. Based on the discriminant model the classification of new observational units into the groups or categories will be conducted. Some authors [10,11]. indicate that goals of a discriminant analysis are to construct a set of discriminants that may be used to describe or characterize group separation based upon a reduced set of variables, to analyze the contribution of the original variables to the separation, and to evaluate the degree of separation.

A discriminant function, also called a canonical root, is a latent variable which is created as a linear combination of discriminating (independent) variables, the form of the equation or function is:

$$D = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_kX_k \qquad (3)$$

Where: D = discriminant function.

X 's = independent variables. b 's = discriminant coefficients or weights.

The coefficients, or weights (b), are estimated so that the groups differ as much as possible on the values of the discriminant function. This occurs when the ratio of between-group sum of squares

to within-group sum of squares for the discriminant scores is at a maximum. When the criterion variable has two categories, the technique is known as two-group discriminant analysis. When three or more categories are involved, the technique is referred to as multiple discriminant analysis.

4.3 ARTIFICIAL NEURAL NETWORKS

A neural network (NN) is a computer-intensive, algorithmic procedure for transforming inputs into desired outputs using highly connected networks of relatively simple processing units (neurons or nodes). Neural networks are modeled after the neural activity in the human brain. The three essential features, then, of an NN are the basic computing units (neurons or nodes), the network architecture describing the connections between the computing units, and the training algorithm used to find values of the network parameters (weights) for performing a particular task. The computing units are connected to one another in the sense that the output from one unit can serve as part of the input to another unit. Each computing unit transforms an input to an output using some prespecified function that is typically monotone, but otherwise arbitrary. This function depends on constants (parameters) whose values must be determined with a training set of inputs and outputs. Network architecture is the organization of computing units and the types of connections permitted. In statistical applications, the computing units are arranged in a series of layers with connections between nodes in different layers, but not between nodes in the same layer. The layer receiving the initial inputs is called the input layer. The final layer is called the output layer. Any layers between the input and output layers are called hidden layers [12].

A regression model in which the responses are nonlinear functions of inputs through layers of connected hidden variables, originally by treating biological neurons as binary thresholding devices. They are flexible models useful for discrimination and classification and are implanted by a computerized "black-box" trained by a training data set [13].

- **APPLICATION AREAS OF ARTIFICIAL NEURAL NETWORKS:**

Application areas of ANN can be technically divided into the following categories: **Classification and diagnostic:** ANN have been applied in the field of diagnosis in medicine, engineering, and manufacturing.

Pattern recognition: ANN have been successfully applied in recognition of complex patterns such as: speech recognition, handwritten character recognition and a lot of other applications in image processing.

Modelling: A neural network is a powerful data modeling tool that can capture and represent complex input/output relationships. The true power and advantage of neural networks lies in their ability to represent both linear and non-linear relationships and in their ability to learn these relationships directly.

Forecasting and prediction: ANN have shown high efficiency as a predictive tool by looking at the present information and predict what is going to happen.

Estimation and Control: ANN have been successfully applied in the field of automatic control in system identification, adaptive control, parameter estimation and optimization and a lot of other applications in this field.

Table 2 shows the comparison between the classification methods based on Sensitivity, accuracy, error rate, Area under ROC curve and Area under ROC curve.

	Sensitivity	Accuracy	Error Rate	Area under ROC curve	Correct classification
LR	0.574	0.698	0.183	0.817	81.6%
DA	0.559	0.795	0.205	0.728	79.5%
ANN	0.595	0.827	0.173	0.858	82.7%

The results of these comparisons showed that Artificial Neural Network can predict better than Logistic Regression and Discriminant Analysis. This is justified by correct classification of 82.7% the ANN model, 81.6% the LR model and 79.5% the DA model in the analysis. By using the ROC curve the area under the curve of the three statistical methods, the area under the curve for the LR is (81.7%), The area under curve for the DA is (72.8%) and the area under curve for ANN is (85.5%). By using accuracy and error rate for artificial neural network are 82.7% and 17.3% respectively. The accuracy and error rate for discriminant analysis are 79.5% and 20.5%

respectively. The accuracy and error rate for logistic regression analysis are 69.8 and 18.3 % respectively.

Thus, the results we have from Artificial Neural Network model are better than the results of Logistic Regression and Discriminant Analysis model for this data set.

5. CONCLUSIONS

This paper used three different classification methods Logistic Regression, Discriminant Analysis and Artificial Neural Network. Using different assessment techniques to achieve to the best model that represents the dataset of Labor Force. We compared the performance of DA, LR and ANN on LF data. The sample size has the most s obvious impact on the difference between and the errors it makes in prediction three methods. The three methods are different in results. Correct classification is 81.6% for LR model compared with 79.5% for DA and Artificial Neural Networks gave accuracy in prediction (82.7%). In addition, that the area under the ROC curve is 81.7 % for LR and 72.8% for DA and 85.8% for ANN. The model means that anyone (observation) in Palestinian region (West bank and Gaza Strip) can answer 8 questions (independent variables) and the age is between (15- 65). LR and DA and ANN model can classify it into one of three groups (Employment and Unemployment) with misclassification 18.3% and 20.5 % and 17.3 % respectively. Thus, the results we have from Artificial Neural Network model are better than the results of Logistic Regression and Discriminant Analysis model for this data set.

CONFLICT OF INTERESTS

The author(s) declare that there is no conflict of interests.

REFERENCES

[1] M.F. Al-Jazzar, A Comparative Study Between Linear Discriminant Analysis and Multinomial Logistic Regression in Classification and Predictive Modeling, Thesis, Department of Statistics, Faculty of Economics and Political Administrative, Al -Azhar University Gaza, 2012.

[2] M.K. Okasha, M.A.M. Shehada, Classification of anemic Palestinian children using the multinomial logistic regression model, Int. J. Adv. Res. 4 (2016), 560-573.

[3] Z. Abbas, A. Saad, M. Ayache, C. Fakih, Applications of Logistic Regression and Artificial Neural Network for ICSI Prediction, Int. Arab J. Inform. Technol. 16 (2019), 557-564.

[4] V. Strilets, N. Bakumenko, S. Chernysh, M. Ugryumov, V. Donets, Application of Artificial Neural Networks in the Problems of the Patient's Condition Diagnosis in Medical Monitoring Systems, in: M. Nechyporuk, V. Pavlikov, D. Kritskiy (Eds.), Integrated Computer Technologies in Mechanical Engineering, Springer International Publishing, Cham, 2020: pp. 173–185.

[5] A. Agresti, Building and applying logistic regression models. In: Agresti A (ed) Categorical data analysis, 2nd edn. Wiley, Hoboken, (2002), pp 211–266.

[6] B.G. Tabachnick, L.S. Fidell, J.B. Ullman, Using Multivariate Statistics. 3d ed. HarperCollins, New York, 1996.

[7] R.B. Burns, R.A. Burns, Business research methods and statistics using SPSS, Sage, Thousand Oaks, (2009).

[8] G.C.J. Fernandez, Discriminant analysis, a powerful classification technique in data mining. In: Proceedings of the SAS users international conference, (2002), pp 247–256.

[9] M. Savić, D. Brcanov, S. Dakić, Discriminant analysis applications and software support, Manage. Inform. Syst. 3 (2008), 29-33.

[10] N.H. Timm, ed., Applied multivariate analysis, Springer, New York, 2002.

[11] Y. Dodge, The Oxford dictionary of statistical terms. Oxford University Press, Oxford, (2003).

[12] R.A. Johnson, D.W. Wichern, Applied multivariate statistical analysis, 5th ed, Prentice Hall, Upper Saddle River, 2007.

[13] PCBS, User guide, Labor Force Survey 2019, Palestinian Central Bureau of Statistics, (2019).