A Comparison between Classification statistical Models and Neural Networks with Application on Palestine data

Amani Moussa Mohamed[1], Mahmoud A. Abdel-Fattah[2] and Abdallah Salman Mohammed ALdirawi*[3]

[1]*Department of Applied Statistics and Econometrics, Faculty of Graduate Studies for Statistical, Cairo University, Cairo, Egypt.*

[2]*Department of Applied Statistics and Econometrics, Faculty of Graduate Studies for Statistical, Cairo University, Cairo, Egypt.*

[3]*Department of Applied Statistics and Econometrics, Faculty of Graduate Studies for Statistical, Cairo University, Cairo, Egypt.*

Abstract: *There are many possible techniques for classification of data. Multinomial Logistic Regression, Discriminant Analysis and Artificial Neural Networks. Are three techniques that commonly used for data classification. three techniques are applied at Labor Force survey in Palestine in 2019. This study aims to choose the best statistical model for Labor Force in Palestine in 2019 data, through the comparison between Multinomial Logistic Regression, Discriminant Analysis and Artificial Neural Networks on real data set. we used a real data of Labor Force from a survey of labor force which was conducted by Palestinian Central Bureau of Statistics (PCBS) in 2019- 2020. The study sample size had been 22625. The target group was the age group (15- 65) years for both sexes. Labor Force data has 12 variables; the dependent variable is nominal with three categories (Employment, Unemployment and Outside of LF) and 11 independent variables. In this study we compared the three statistical models using different assessment techniques (Cross-validation with half of the observations, sensitivity, accuracy, error rate, and method ROC curves) and obtained the best estimate of accuracy and error rate in order to achieve the best model for the data. These results demonstrate that multinomial logistic regression can be more powerful analytical technique for use than discriminant analysis, and artificial neural networks.*

Keywords: **Classification; Artificial Neural Networks; Logistic Regression and the Discriminant Analysis.**

1. Introduction

A categorical variable (sometimes called a nominal variable) is one that has two or more categories. They represent types of data which may be divided into groups. The categorical variables have no numerical meaning, and there is no intrinsic ordering to the categories. For example, hair color, gender, field of study, college attended, political affiliation, and status of disease infection. Gender is a categorical variable having two categories (male and female) and there is no intrinsic ordering to the categories from highest to lowest. A variable with just two categories is said to be dichotomous, whereas one with more than two categories is described as polytomous [3].

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 be used to 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. For example, we may use quantitative information in predicting who will or will not graduate from college. This would be an example of simple binary classification problems, where the categorical dependent variable can only assume two distinct values. In other cases, there are multiple categories or classes for the categorical dependent variable. For example, when we are ill, we want a doctor to diagnose our disease from our symptoms.

All above are classification problems where we attempt to predict values of a categorical dependent variable from one or more continuous and/or categorical predictor variables. In statistics, it is the process of allocating an observation p in one of several predefined groups or categories. and an ideal classification method, also provides in what distinguishes different classes

from each other. The basic objective is to build a discriminant function that takes the information to summarize the p variables on an indicator that yields the optimal discrimination between the classes, and the goal of classification in this case, also known as supervised pattern recognition [17]. In order to derive the decision rule that yields the optimal discrimination between the classes, one assumes that a training set of pre-classified cases −the data sample− is available, and can be used to determine the sought after rule applicable to new cases. The decision rule can be derived in a model-based approach, whenever a joint distribution of the random variables can be assumed, or in a model-free approach [9].

There are many methods that can be used to compare between observations such as the Discriminant Analysis and Multinomial Logistic Regression. There is another method used for comparison between observations, it is the Artificial Neural Networks. We have statistical analysis labor force using DA, MLR and ANN. The data consist of 11 independent variables and one dependent variable which has three categories (Employment, Unemployment and Outside of LF). The goal is to find the best model according to model selection criteria.

2. Literature Review

[16] made a comparison between three methods of discrimination has been conducted, which are, the Artificial Neural Networks, the Logistic Model and the Discriminant Function, and that is for classifying observations into the group belonged to, in the manner when some variables don’t follow the normal distribution. That comparison is made for the preference between the three methods. The criterion of misclassified observations ratio (misclassification ratio), has been used as a comparison criterion. The main results of the study; that the suggested models gave similar results concerning significance of the impact and importance of the independent variables involved in the analysis. The variable “family size” is the most important factor for differentiating and discriminating between families’ income concerning the sufficiency, where the variable “existence of university students in the family” has no significant impact on the sufficiency of family income, also the Artificial Neural Networks method obtained the best classification ratio than the Logistic Model and the Discriminant Function method.

[13] aims in his to assess 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 8 regression method proves to be the most reliable method with the highest area under the receiver operating characteristic (ROC), or ROC curve, 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.

[1] the third most serious disease estimated by Word Wide Organization after cancer and cardiovascular disease is the infertility. The advanced treatment techniques is 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.

[15] the problem of accurate medical diagnosis 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 proposes a 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 LF Survey, 2019 where conducting by Palestinian Center Bureau of Statistics (PCBS) used for application of the MLR, DA and ANN. The sample size had been 22625observations of whom 22625 is valid and no missing value. 59.7% residing in the West Bank 40.3% residing in Gaza Strip. Dataset contains 12 variables. We are interested on Labor Force status (1) variable. This variable has been used as a dependent variable in this analysis. It involves three categories (Employment, Unemployment, and Outside of LF). The goal is to find the best model which can describe the relationship between different types of LF and other factors that can be considered as independent variables and have effects on each dependent variable. As a result, we conclude the best model. The target group used in this study was people living in west bank and Gaza governorate in the age group 15 – 65 years in both sexes [14].

Table (1): The explanatory variables

Variable name	Description	Value Lable
HR2	Sex	1. Male 2. Female
Pr1	The Age at last Birthday	15 - 65
Pr2	Attendance in formal Education	1.Currently Attending 2-Attended and left 3-Attended and graduated 4-Never attended
Pr4	Educational Attainment (higher Qualification)	1-Illiterate 2-Can Read and Write 3-Elementary 4-Preparatory 5-Secondary 6-Associatte Diploma 7-BA\ BSc 8-Higher Diploma 9-Master Degree 10-Ph.D
Pr6 -a	Did attendance …. training course attendance (such as training course that managed by ministry of labour, Qalandia institute - (must present certificate at the end of the training course)	1-Currently Attending 2-Attended and graduated 3-Attended and left 4-Never attended
HR5	Refugee Status	1-Registered 2-Not Registered 3-Not Refugee
ID7	Locality Type	-Urban, Rural, Camp
WBGS	Region	1.West Bank 2. Gaza Strip
Marital	Marital Status	1.Never Married 2. Married 3. Other
Industry	INDUSTRY group	1.Agriculture 2. Manufacturing 3. Construction 4. Commerce, Hotels and Restaurants 5. Transport, Storage and Communication 6. Services
HR4	Relationship to the Head of Household	Head, spouse, son\daughter, father\mother, brother\ sister, grand father\mother, grand child, Son Wife\ Daughter Husband, Other relative, Others

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 [17]. The assignment of alternatives observations or objects into predefined homogenous groups is a problem of major practical and research interest [4]. In this paper, the three classification methods namely Multinomial logistic regression (MLR), discriminant analysis (DA) and artificial neural networks (ANN).

4.1 Discriminant Analysis

The best definition of discriminant analysis is a technique which allows the classification of an individual into one of two or more distinctive populations, on the basis of a set of measurements [2].

Discrimination and classification are multivariate techniques concerned with separating distinct sets of objects (or observations) and with allocating new objects (observations) to previously defined groups. Discriminant analysis is rather exploratory in nature. As a separative procedure, it is often employed on a one-time basis in order to investigate observed differences when causal relationships are not well understood. Classification procedures are less exploratory in the sense that they lead to well-defined rules, which can be used for assigning new objects. Classification ordinarily requires more problem structure than discrimination does [10].

- **Discriminant Functions**

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:

$$D = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + . . . + b_kX_k. \qquad (1).$$

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. When (the number of independent variables)

= 2, the classification boundary is a Straight line. Every individual on one side of the line is classified as Group 1, on the other side, as Group 2. When n = 3, the classification boundary is a two-dimensional plane in three-dimensional space, the classification boundary is generally an n-1 dimensional hyper plane in n space [12].

4.2 Multinomial Logistic Regression

Multinomial logistic regression (MLR) is used to predict categorical placement in or the probability of category membership on a dependent variable based on multiple independent variables. The independent variables can be either dichotomous (i.e., binary) or continuous. MLR is a simple extension of binary logistic regression that allows for more than two categories of the dependent or outcome variable. Like binary logistic regression, MLR uses maximum likelihood estimation to evaluate the probability of categorical membership the basic principle of multinomial logistic regression is similar to that for logistic regression (LR), in that it is based on the probability of membership of each category of the dependent variable. LR was proposed in the late 1960s and early 1970s [5], and it became routinely available in statistical packages in the early 1980s.There are two models of logistic regression to include binary LR and MLR. Binary logistic regression is typically used when the dependent variable is dichotomous, and the independent variables are either continuous or categorical variables. Logistic regression is best used in this condition. When the dependent variable is not dichotomous and is comprised of more than two cases, an MLR can be employed. Also referred to as logit regression, MLR has very similar results to binary logistic regression [8].

- **The General Multinomial Logistic Model**

A multinomial logistic model classifies d-dimensional real-valued input vectors x ∈ R^d into one of k outcomes c ∈ {0, . . ., k − 1} using k − 1 parameter vectors $\beta_0, \ldots, \beta_{k-2} \in R^d$, (Carpenter , 2008)

$$p(c|x,B) = \begin{cases} \dfrac{\exp\,(B_C X)}{Z_X} & if\ c\ < K-1 \\ \dfrac{1}{Z_X} & if\ c = K-1 \end{cases} \qquad (2)$$

where the linear predictor is inner product: $\beta_c.x = \sum_{i<d} \beta_{c,i}.x_i$

The normalizing factor in the denominator is the partition function: $Z_x = 1 + \sum_{c<k-1} \exp(\beta_c.x)$.

4.3 Artificial Neural Networks

Neural Networks became a common solution for a wide variety of problems in many fields, (ANN) is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain. The most important features of the human brain is the ability to learn from the past, according to a complex system of sending and receiving electrical pulses between neurons. This fact has prompted many researchers and led to the establishment of the cognitive sciences, known as artificial intelligence and building the network. An artificial neural network is composed of many artificial neurons that are linked together according to a specific network architecture. The objective of the neural network is to transform the inputs into meaningful outputs. [18].

Definition: 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 [6].

- **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 17 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 the area of image processing.

Modelling: A neural network is a powerful data modeling tool that is able to 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.

5. Methods Evaluation

Table (2): Confusion matrix for two classes (Positive and Negative)

Confusion Matrix		Target	
		Negative	Positive
Model	Negative	**TN**	**FP**
	Positive	**FN**	**TP**

As we can see in table (2), True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN), are the four different possible outcomes of classification prediction for a two-class case with classes “1” (“yes”) and “0” (“no”). A False Positive is when the outcome is incorrectly classified as “yes” (or “positive”), when it is in fact “no” (or “negative”). A false negative is when the outcome is incorrectly classified as negative when it is in fact positive. True Positives and True Negatives are obviously correct classifications. Below formulae were used to calculate sensitivity, specificity, and accuracy. [11].

Accuracy, which is the proportion of true results (both true positives and true negatives) in the population, can be obtained by the following formula:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \qquad (3)$$

Sensitivity (equivalent to the true positive rate) It is the proportion of positive cases that are well detected by the test. In other words, the sensitivity measures how the test is effective when used on positive individuals. The test is perfect for positive individuals when sensitivity is 1, equivalent to a random draw when sensitivity is 0.5. If it is below 0.5, the test is counter-performing, and it would be useful to reverse the rule so that sensitivity is higher than 0.5 (provided that this does not affect the specificity). The mathematical definition is given by:

$$\text{Sensitivity(True Positive Rate)} = \frac{TP}{(TP) + (FN)} \qquad (4)$$

The error rate of all classifications is calculated by the following formula:

$$\text{Error Rate} = \frac{FN + FP}{TP + TN + FP + FN} \qquad (5)$$

- **ROC Curve**

ROC (Receiver Operating Characteristic) analysis is being used as a method for evaluation and comparison of classifiers [7].

The ROC gives complete description of classification accuracy as given by the area under the ROC curve. The ROC curve originates from signal detection theory; the curve shows how the receiver operates the existence of signal in the presence of noise. The ROC curve plots the probability of detecting true signal (sensitivity) and false signal (1 – specificity) for an entire range of possible cut points. The sensitivity and specificity of a classifier also depend on the definition of the cut-off point for the probability of predicted classes. In many situations, not all misclassifications have the same consequences, and misclassification costs must be considered. A ROC curve demonstrates the tradeoff between true positive rate and false positive rate in binary classification problems. To draw a ROC curve, the true positive rate TP rate and the false positive rate FP rate are needed. TP rate determines the performance of a classifier or a diagnostic test in classifying positive cases correctly among all positive samples available during the test. FP rate, on the other hand, defines how many incorrect positive results, which are negative, there are among all negative samples available during the test. Because TP rate is equivalent to sensitivity and FP rate is equal to (1 – specificity), the ROC graph is sometimes called the sensitivity vs (1 - specificity) plot. The area under the ROC curve has become a particularly important measure for evaluating classifiers‟ performance because it is the average sensitivity over all possible specificities [4].

6. Results

Table 3 shows the comparison between the classification methods based on Sensitivity, accuracy, error rate, Area under ROC curve and Area under ROC curve.

Table (3): Sensitivity, accuracy, error rate, Area under ROC curve and Area under ROC curve

	Sensitivity	**accuracy**	**Error Rate**	**Area under ROC curve**	**Correct classification**
MLR	0.969	0.822	0.178	0.910	82.2%
DA	0.941	0.791	0.208	0.606	79.2%
ANN	0.965	0.815	0.184	0.880	81.6%

The results of these comparisons showed that the application of the correct classification technique to assess the accuracy of the three classification methods in predicting the of labor force we found that Multinomial

Logistic Regression gave best accuracy in prediction is (82.2%) ,with 79.2% for Discriminant Analysis and with (81.6%) for Artificial Neural Networks. In addition that the ROC curve technique was applied to assess the accuracy of the three classification methods in predicting the labor force we found that Multinomial Logistic Regression gave best accuracy in prediction is (91%), with 60.6% for Discriminant Analysis and with (88%) for Artificial Neural Networks. In addition, that we found that Multinomial Logistic Regression gave the best in prediction is from where the accuracy of the Multinomial Logistic Regression 82.2%, sensitivity 96.6% and less error rate 0.178. While it gave the Discriminant analysis model accuracy 79.1% sensitivity 94.1% and error rate 0.208. It also gave artificial neural network accuracy 81.5% sensitivity 96.5% and error rate 0.184.

7. Conclusions

In this paper, we have used three different classification methods, MLR, DA and ANN. Using different assessment techniques in order to achieve to the best model that represents the dataset of Labor Force. We compared the performance of DA, MLR 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 82.2% for MLR model compared with 79.2% for DA and Artificial Neural Networks gave accuracy in prediction (81.6%). In addition, that the area under the ROC curve is 91 % for MLR and 60.6% for DA and 88% for ANN.

The model means that anyone (observation) in Palestinian region (West bank and Gaza Strip) can answer 11 questions (independent variables) and the age is between (15- 65). MLR and DA and ANN model can classify it into one of three groups (Employment, Unemployment and Outside LF) with misclassification 17.8 % and 20.8 % and 18.4 % respectively. These results demonstrate that Multinomial logistic regression can be more powerful analytical technique than Discriminant Analysis, Artificial Neural Networks.

Reference

[1] Abbas Zeinab, Saad Ali, Ayache Mohammad, and Fakih Chadi (2019). " Applications of Logistic Regression and Artificial Neural Network for ICSI Prediction “. The International Arab Journal of Information Technology, Vol. 16, No. 3A, Special Issue

[2] Afifi, a. A. and Clark, V. (1984)." Computer-aided multivariate Analysis", Belmont, California: Lifetime Learning Publications.

[3] Agresti Alan. (2002). Categorical Data Analysis. 2nd Edition. University of Florida. New York: Wiley

[4] Bradley, A. P. (1997)."The use of the area under the roc curve in the evaluation of machine learning algorithms", Pattern Recognition, Jul, 30(7): pp.1145-59.

[5] Cabrera, A. F. (1994). "Logistic regression analyses in higher education: An applied perspective". In J. C. Smart (Ed.), Higher Education Handbook of Theory and Research (Vol. 10, pp.225-256). New York: Agathon Press.

[6] Dodge, Y. (2003)."The Oxford Dictionary of Statistical Terms", Oxford University, p. 282.

[7] Ferri, C., Flach, P. & Hernandez-Orallo, J. (2002). "Learning Decision Trees Using the Area under the ROC Curve", Nineteenth International Conference on Machine Learning (ICML 2002), Morgan Kaufmann, pp. 46-139.

[8] Jat D. Singh, Dhaka Poonam and Limbo Anton (2018). "Applications of statistical techniques and artificial neural networks: A review".A review, Journal of Statistics and Management Systems, 21:4, 639-645.

[9] Joaquim,P. ,Marques de Sá.(2007). Applied Statistics Using SPSS, STATISTICA, MATLAB and R. 2nd Edition. Springer-Verlag Berlin.

[10] Johnson, R. A., and Wichern, D. W. (2007). " Applied Multivariate Statistical Analysis".

[11] Kumari, Milan and Godara, Sunila (2011). "Review of Data Mining Classification Models in Cardiovascular Disease Diagnosis", International Journal of Computer Sci ence and Technology, Vol. 2, Issue

[12] Morrison, D. (1969), "Interpretation of Discriminant Analysis" Journal of Marketing Research". 6 156-63.

[13] Okasha, Mahmoud K. and Abu Samra, Ashraf I. (2016). "Classification Methods for Hypertension Patients' Data in Palestine"; Proceedings of the 25th Annual International Conference on Statistics and Modeling in Human and Social Sciences; Department of Statistics; Faculty of Economics and Political Science; Cairo University; Cairo, Egypt; March 25-28,.

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

[15] Strilets V., Bakumenko N., Chernysh S., Ugryumov M.,and Donets V. (2020)" Application of Artificial Neural Networks in the Problems of the Patient's Condition Diagnosis in Medical Monitoring Systems". In: Nechyporuk M., Pavlikov V., Kritskiy D. (eds) Integrated Computer

Technologies in Mechanical Engineering. Advances in Intelligent Systems and Computing, vol 1113. Springer, Cham

[16] Suleiman. Ali (2015). "Comparison Between discriminant analysis, Binary logistic regression and neural networks for Classification Observation (Case Study: factors affecting the adequacy of family income) ". Ph.D. Thesis. Sudan University of Science and Technology.

[17] Wehrens ,Ron .(2010). Chemometrics with R Multivariate Data Analysis in the Natural Sciences and Life Sciences. Springer.

[18] Zurada ,(1992), "Introduction to Artificial Neural Systems",west publishing company.