PERFORMANCE COMPARISON OF CLASSIFICATION TECHNIQUES, ARTIFICIAL NEURAL NETWORK, DISCRIMINANT ANALYSIS & LOGISTIC REGRESSION: APPLICATION “ESTABLISH MORE PRIVATE ACADEMIES OR NOT”

Muhammad Ahsan ul Haq[1*], Irum Sajjad Dar[2] & Qura-tul-ain[3]
[1,2,3] College of statistical & Actuarial Sciences University of Punjab Lahore-Pakistan
[*]ahsanshani36@gmail.com

Abstract: *Academies around the world are well-known for their unique kind of education or knowledge that they provide. The prime aim of study was to investigative classification performance of discriminant analysis (DA), artificial neural networks (ANNs) and logistic regression (LR) methods, while they were applied on data of more private academies established or not. Exploratory factor analysis (EFA) was performing with principal component analysis extraction method. Six factors were extracted which explained 65.582% total variation, factors were namely known as Poor education quality of public sector institutions, Draw backs of academy culture, Career counseling in academies, Competition among academies to survive, Joining academies is a fashion of the day and Benefits of academy culture. Linear discriminant analyses, artificial neural networks and logistic regression are applied in mandate to predict the probability of a specific categorical outcome based upon several independent variables. In discriminant analysis, F-test values are significant for all independent variables, showing the significant difference for supporters and non-supporters. Since a value of Wilks’ lambda is smaller for the predictor “joining academies is a fashion of the day”. LR shows by performing back elimination method just one variables was found highly significant. The classification performance of DA, LR and RBS is 96.5%, 96.1% and 96.5% respectively. So we can conclude that ANNs with Radial Basis System (RBS) have slightly higher discriminative performance than DA and LR. Academies seems to become a fashion in our society.*

Keywords; Factor analysis; classification; discriminant analysis; logistic regression; academies

INTRODUCTION

Statistical methods such as, linear discriminant analysis (DA), logistic regression (LR) and artificial neural networks (ANNs) with Radial Basis system (RBS) are classification and model building methods. These three methods can be used for estimation of associations between a categorical outcome variable and various covariates. Methods, logistic regression, discriminant analysis and artificial neural network are widely implemented practically, especially in social and medical sciences. Practically when we deal regression analysis and our dependent variable is categorical then we are not able to use simple linear or multiple linear regression, especially when dependent variable is binary (dichotomous) then we can use Logistic Regression and the independent variables are of any type like categorical or continuous. LR is mostly applied in social and medical sciences with the outcome variable presence or absence of a disease, success or failure etc. Using the logit transformation, the non-linear or S shaped curve changed into linear line. The LR analysis is based on computing the odds of the dependent variable as the ratio of the probability of presence and the probability of absence.

We discuss in starting Discriminant analysis (DA) similarly use as a classification procedure, which is use to decide which set of items distinguishes between two or more logically arising sets. DA undertakes the same task as multiple linear regression by predicting an outcome. However, linear regression is limited to use when continuous dependent variable while DA can deals with predicting categorical dependent variables with more than two categories of dependent variables. Artificial neural networks (ANNs) with radial basis have been used widely in medical and social sciences fields [1] [2]. This methodology can be used for two major purposes in research are, first one for pattern recognition (classification) and secondly for prediction purposes.

Although throughout the literature the theoretical properties of these methods have been discussed. Classification methods such as discriminant analysis [3] [4] [5] [6] [7], logistic regression [2] [8] [9] [10] [11] [12] and artificial neural network [2] [8] [9] [10] are used these models for categorizing purpose. In recent years published studies reported that artificial neural network method improves prediction in several situations [13]. In contrast, according to [14] [15] logistic regression produced similar results with neural networks. According to [10] artificial neural network outperformed results than logistic regression.

Academies all over the world are acknowledged for their special type of education or knowledge that they give. The earliest academy was established in 385 year BC by the Plato, principle of which was to teach philosophy. The Chinese generally refer academies to a private institution build away from cities or town; provide quiet surroundings where scholars could connect in studies and thought without limitations and mature distraction. In Pakistan the idea of academies is entirely reverse. In Pakistan academies instruct identical curriculum suggested by management to public and private segment institute only. Academies go by the same curriculum in Pakistan and prepare students for examination to get good results. Although there are a bunch of academies who teach students in other field of teaching like knowledge of other languages and preparing students for fashion designing, they are still popularly known to prepare students for school/college examination by giving them best guess papers or notes. Every academy provides their guess documents. They teach students to do only selective studies out of the whole syllabus. Education is separated into these stepladders: primary level, secondary education, intermediate and then higher studies which are usually provide by higher institutes like university. But only a short number of the primary conscription enters in the University for Advanced Studies. Students have to manage attendance school/college and academy to study and getting good results. Different books are suggested by academies and schools/colleges.

In this study we organize this study on this pattern, firstly try to give intro about classification techniques and about academies. After that, we try to illustrate logistic regression, artificial neural network and linear discriminant analysis and application of these classification models on data of give support to established more academies in our society or not.

MATERIAL AND METHODS

Exploratory Factor Analysis (EFA)

In multivariate statistics, exploratory factor analysis (EFA) is a method used to discover the underlying structure of a relatively large set of variables. Exploratory factor analyses give meaning full dimensions with the help of highly correlated variables. We require a minimum sample of hundred respondents and minimum three variables in one factor with minimum correlation 0.3 [16].

Logistic Regression (LR)

Logistic regression is a special case of regression modeling when our dependent variable is binary or dichotomous. It is widely used to predict the probability of the presence or absence of a disease, success or failure, or an outcome generally based on discrete, continuous, or categorical independent variables. General form of Logistic regression is as follow $logit\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \sum_{i=1}^{k} \beta_i x_i$, where β_i are estimated parameters which are estimates with maximum likelihood method. For classification purpose a threshold was set to 0.5, then we can classify an object to group 0 if $P_1 < 0.5$ and to group 1 if $P_1 > 0.5$.

Discriminant Analysis (DA)

Discriminant analysis functional form as follows:

$$\text{LDF} = V_1X_1 + V_2X_2 + V_3X_3 + \ldots + V_iX_i + a$$

Where $V_1, V_{2,......,} V_k$ are diagnosis coefficients, $X_1, X_{2,......,} X_i$ are independent variables, “a” is constant. Discriminant analysis is a multivariate technique which focuses on association between categorical dependent variables and multiple independent variables. Simplest form of discriminant analysis is when dependent variable is dichotomous, in this case discriminant function use to classify subjects into two groups. In discriminant analysis similar to factor analysis, variables remain in the model whose diagnosis coefficients comparative significant correlation and greater than 0.3. Assumptions of discriminant analysis have the same form with ordinal regression. The assumptions are; (1) Explanatory variables follows multivariate normal distribution this shows that only continuous variables used in model, (2) all independent variables variance covariance matrix homogeneous which check by using Box’s M statistic and (3) independence.

Artificial neural networks (ANNs)

Artificial neural network (ANNS) is objective and efficient classification method that used in a large number of classification field. A neural network is a computer-intensive, algorithmic procedure for transforming inputs into outputs using a connected network. The neural networks modeling are inspired by the neural activity in the human brain. In statistical applications, the neurons are arranged in a series of layers only connected between neurons of different layer. Considering Given distribution as a default normal distribution of the response variables, linearity correlation and similarity of variance errors are described as some of the classical methods of limitations. Artificial neural network is applied to classify and predict these correlations are not so general linear [17]. Radial Basis Function (RBF) neural network is one ultimate significant type which considered a contender of multilayer perception neural network. In RBF neural network 0.5 threshold was set for classification.

RESULTS AND DISCUSSION

The results found from sample 256, (66%) female students, average age of respondents was 16 years and average family income was 17000. In our sample (52%) were metric, (34%) intermediate and (14.1%) graduate students. Majority (19.5%) of students’ father education was graduation but mother (37%) was metric. For further analysis, we firstly perform factor analysis then classification methods which discussed in earlier section.

Exploratory factor analysis was used with PCA and Kaiser’s principles. Six factors were extracted from our original data, Poor Education Quality of Public Sector Institutions, Drawbacks of Academy Culture, Career counseling in academies, Competition among Academies to survive, Joining Academies is a Fashion of the Day, and benefits of academy culture. These extracted variables were used in logistic regression, discriminant analyses and artificial neural networks after computing summated scores. Table-1 shows the total variation of all factors and individual factors, also reliability analysis for overall data and individual factors. The value of Cronbach’s Alpha was 0.840 for all variables which is considered as very good. Further detailed information exists in below Table-1.

require for discriminant, logistic regression and artificial neural network analysis. Firstly we use discriminant analyses to identify students’ opinion to give support establishing more private academies. For identification dependent variable is give support to establishing more private academies, which has two categories (Supporters 0, Non-supporters 1). Poor Education Quality of Public Sector Institutions, Drawbacks of Academy Culture, Career Counseling in Academies, Competition among Academies to Survive, Joining Academies is a Fashion of the Day, and benefits of academy culture are used as independent variables. Overall significance of fitting discriminant analysis is checked with help of Wilks’ lambda which shows highly significant at (p<0.001). Furthermore for the purpose of assumption of homogeneity of covariance matrices between groups we see the value of

Table-1: Results of the Factor Analysis

Factor(s)	Explained Variation (%)	Cronbach's Alpha	No. of Items
Poor Education Quality of Public Sector Institutions	22.632	0.953	10
Drawbacks of Academy Culture	11.777	0.915	5
Career counseling in Academies	10.456	0.929	4
Competition among Academies to Survive	7.258	0.852	3
Joining Academies is a Fashion of the Day	7.071	0.860	3
Benefits of Academy Culture	6.388	0.840	3
Total	**65.582**	**0.840**	**28**

Box's M Test. Since the (p=0.250) indicating that the covariance matrices are similar for the two groups. Note that discriminant analysis is robust enough to the violation of assumption of homogeneity of covariance matrices.

In section-3 we describe about the nature of variables that

Furthermore Table-2 which shows the group means for each of the independent variables. In profiling the two groups, we can first identify the six variables, with the largest differences in the group means (Poor Education Quality of Public Sector Institutions, Drawbacks of Academy Culture, Career Counseling in Academies, Competition among Academies to Survive, Joining Academies is a Fashion of the Day, Poor Education Quality of Public Sector Institutions and hardworking). Table-2 also shows the Wilks' lambda and univariate ANOVA used to assess the significance between the means of the independent variables for the two groups. The Wilks' lambda and univariate F values represent the separate or univariate effects of each variable. These tests indicate that the all six independent variables are significant univariate differences between the groups. By examining the group differences we identify variables as the most logical set of candidates for entry into the discriminant analysis. From our review of group differences, we saw that all independent variables have significance difference between groups. We can see that F-tests values are significant for all the independent variables, indicating that Supporters and Non-supporters differ. Since the value of Wilks' Lambda is smaller for the predictor Joining Academies is a Fashion of the Day**,** so it's more important to the Discriminant function. Furthermore, classification of the model for supporters and non-supporters was assessed as much as (96.5%) for original and (96.1%) for Cross-validated discriminant analysis.

Logistic Regression (LR)

To asses' classification of observations by logistic regression, six predictor variables were entered into model and one dependent variable. Firstly logistic regression analysis was performed with enter method which identified five variables such as; poor education quality of public sector institutions (p=0.921), drawbacks of academy culture (p=0.079), career counseling in academies (p=0.569), competition among academies to survive (p=0.384), benefits of academy culture (p=0.358) insignificant and only one variable joining academies is a fashion of the day (p=0.000) found significant. Overall model classification value is 96.9%. Secondly LR applied with backward method, significance value (p=0.989) of Hosmer-Lemsho test indicated a high precision of the method. Furthermore, (Negelkerke R square = 0.886) statistic indicate the degree of explained variation of dependent variable (supporters, non-supporters). Secondly logistic regression applied with backward LR values of Hosmer-Lemsho and Negelkerke R square test are (p=0.574; 0.866) and found only one significant variable in model. The overall classification power of the method is 96.1% (see table-3).

Academies all over the world are well known institutions which can become a well source to provide better education. We discuss earlier it is a first type study which conducted on academies data in Pakistan. In this study, we want to compare the performance of well-known and popular models such as discriminant analysis, artificial neural networks and logistic regression. In this study, six factors were extracted by using factor analysis method with

Varimax rotation and principal component extraction technique and explained 65.582% total variation. The individual factor variation was 22.632, 11.777, 10.456, 7.258, 7.071 and 6.388 respectively. We found the reliability of all items (0.840) and individual factors such as; 0.953, 0.915, 0.929, 0.852, 0.860 and 0.840 respectively. Further we use these factors as independent variables in logistic, discriminant and artificial neural networks methods with a binary dependent variable give support to establishing more private academies (yes or no).

Discriminant analysis results shows that all independent variables are significant. We can see that F-test values are significant for all the independent variables, indicating that supporters and non-supporters differ. We can see value of Wilk's lambda is very low for predictor variables "joining academies is a fashion of the day". Logistic regression modeling shows only one variable is significant in the model with backward elimination method "joining academies is a fashion of the day". The result of artificial neural networks with RBS the power of discriminate Supporters and Non-supporters was 96.6% which is slightly higher than discriminant and logistic regression 96.5% and 96.1% respectively.

Table-2: Group Descriptive Statistics and test of equality in the two-group Discriminant Analysis

Independent Variables	Means		Wilks' Lambda	F	Sig.
	Supporters	Non-Supporters			
Poor Education Quality of Public Sector Institutions	45.5318	35.6944	0.927	19.884	0.000**
Drawbacks of Academy Culture	22.9864	20.2500	0.977	5.915	0.016*
Career Counseling in Academies	19.8409	12.9722	0.816	57.457	0.000**
Competition among Academies to Survive	13.5045	11.1944	0.965	9.272	0.003**
Joining Academies is a Fashion of the Day	15.8409	7.3611	0.284	639.39	0.000**
Poor Education Quality of Public Sector Institutions	13.5773	11.3056	0.958	11.137	0.001**

*p-value< 0.05; **p-value < 0.01

Table-3: Models Classification for Supporters & Non-Supporters

Models	Observed groups	Predicted groups		Percentage correct
		Supporters	Non-supporters	
Discriminant analysis	Supporters	212	8	220
	Non-supporters	1	35	36
	Overall			**96.5%**
Logistic regression	Supporters	214	6	97.7%
	Non-supporters	4	32	91.7%
	Overall			**96.1%**
ANNs (RBF)	Supporters	166	2	98.8%
	Non-supporters	5	17	77.3%
	Overall			**96.6%**

We conclude that in our society academies becomes a need and they contribute a bunch because students feel public sector institutions provide poor quality of education. We can say it’s become a necessary to join academies for better career counseling and necessary for surviving in this society. Academies these days looks like become a fashion of the day.

REFERENCES

[1] Lisboa, P. J. (2002). A review of evidence of health benefit from artificial neural networks in medical intervention. *Neural networks, 15*(1), 11-39.

[2] Bourdès, V., Bonnevay, S., Lisboa, P., Defrance, R., Pérol, D., Chabaud, S., Négrier, S. (2010). Comparison of artificial neural network with logistic regression as classification models for variable selection for prediction of breast cancer patient outcomes. *Advances in Artificial Neural Systems, 2010*.

[3] Way, T. W., Sahiner, B., Chan, H. P., Hadjiiski, L., Cascade, P. N., Chughtai, A., Kazerooni, E. (2009). Computer-aided diagnosis of pulmonary nodules on CT scans: improvement of classification performance with nodule surface features. *Medical physics, 36*(7), 3086-3098.

[4] Shah, S. K., McNitt-Gray, M. F., Rogers, S. R., Goldin, J. G., Suh, R. D., Sayre, J. W., et al. (2005a). Computer aided characterization of the solitary pulmonary nodule using volumetric and contrast enhancement features. *Academic Radiology, 12(10)*, 1310–1319.

[5] Shah, S. K., McNitt-Gray, M. F., Rogers, S. R., Goldin, J. G., Suh, R. D., Sayre, J. W., et al. (2005b). Computer-aided diagnosis of the solitary pulmonary nodule. *Academic Radiology, 12(5)*, 570–575.

[6] Aoyama, M., Li, Q., Katsuragawa, S., Li, F., Sone, S., & Doi, K. (2003). Computerized scheme for determination of the likelihood measure of malignancy for pulmonary nodules on low-dose CT images. *Medical Physics, 30(3)*, 387–394.

[7] Shiraishi, J., Abe, H., Engelmann, R., Aoyama, M., & Doi, K. (2003). Computer-aided diagnosis to distinguish benign from malignant solitary pulmonary nodules on radiographs: ROC analysis of radiologists’ performance. *Radiology, 227(2)*, 469–474.

[8] Rahman, A., Nesha, K., Akter, M., & Uddin, M. S. G. (2013). Application of artificial neural network and binary Logistic regression in detection of Diabetes status. *Science, 1*(1), 39-43.

[9] Mansour, R., Eghbal, K., & Amirhossein, H. (2013). Comparison of Artificial Neural Network, Logistic Regression and Discriminant Analysis Efficiency in Determining Risk Factors of Type 2 Diabetes.

[10] Chen, H., Zhang, J., Xu, Y., Chen, B., & Zhang, K. (2012). Performance comparison of artificial neural network and logistic regression model for differentiating

lung nodules on CT scans. *Expert Systems with Applications, 39*(13), 11503-11509.

[11] Antonogeorgos, G., Panagiotakos, D. B., Priftis, K. N., & Tzonou, A. (2009). Logistic regression and linear discriminant analyses in evaluating factors associated with asthma prevalence among 10-to 12-years-old children: divergence and similarity of the two statistical methods. *International journal of pediatrics, 2009*.

[12] Maroco, J., Silva, D., Rodrigues, A., Guerreiro, M., Santana, I., & de Mendonça, A. (2011). Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. *BMC research notes, 4*(1), 299.

[13] Lisboa, P. J., Wong, H., Harris, P., & Swindell, R. (2003). A Bayesian neural network approach for modelling censored data with an application to prognosis after surgery for breast cancer. *Artificial intelligence in medicine, 28*(1), 1-25.

[14] Ergün, U., Serhatlıoğlu, S., Hardalaç, F., & Güler, İ. (2004). Classification of carotid artery stenosis of patients with diabetes by neural network and logistic regression. *Computers in biology and medicine, 34*(5), 389-405.

[15] Tafeit, E., Möller, R., Sudi, K., & Reibnegger, G. (1999). The determination of three subcutaneous adipose tissue compartments in non-insulin-dependent diabetes mellitus women with artificial neural networks and factor analysis. *Artificial intelligence in medicine, 17*(2), 181-193.

[16] Anderson, T. W., & Rubin, H. (1956). Statistical Inference in Factor Analysis. *Bekerly Symposium on Mathematical Statistics and Probability, 5*, 111-150.

[17] Hagan, M. T., & Demuth, H. B. (1999). *Neural networks for control.* Paper presented at the American Control Conference, 1999. Proceedings of the 1999.