A Comparison of Logistic Regression Model and Artificial Neural Networks in Predicting of Student's Academic Failure

Saeed Hosseini Teshnizi,
Sayyed Mohamad Taghi
Ayatollahi

Department of Biostatistics, Shiraz
University of Medical Sciences,
Shiraz, Iran

Corresponding author: Sayyed Mohammad
Taghi Ayatollahi, Department of
Biostatistics, Faculty of Medicine, Shiraz
University of Medical Sciences, Shiraz, Iran.
Tel/Fax: 00987132349330,
E-mail: ayatollahim@sums.ac.ir

doi: 10.5455/aim.2015.23.296-300

ACTA INFORM MED. 2015 OCT 23(5): 296-300

Received: 21 July 2015 • Accepted: 25 September
2015

ABSTRACT

Background and objective: Artificial Neural Networks (ANNs) have recently been applied in situations where an analysis based on the logistic regression (LR) is a standard statistical approach; direct comparisons of the results, however, are seldom attempted. In this study, we compared both logistic regression models and feed-forward neural networks on the academic failure data set. **Methods:** The data for this study included 18 questions about study situation of 275 undergraduate students selected randomly from among nursing and midwifery and paramedic schools of Hormozgan University of Medical Sciences in 2013. Logistic regression with forward method and feed forward Artificial Neural Network with 15 neurons in hidden layer were fitted to the dataset. The accuracy of the models in predicting academic failure was compared by using ROC (Receiver Operating Characteristic) and classification accuracy. **Results:** Among nine ANNs, the ANN with 15 neurons in hidden layer was a better ANN compared with LR. The Area Under Receiver Operating Characteristics (AUROC) of the LR model and ANN with 15 neurons in hidden layers, were estimated as 0.55 and 0.89, respectively and ANN was significantly greater than the LR. The LR and ANN models respectively classified 77.5% and 84.3% of the students correctly. **Conclusion:** Based on this dataset, it seems the classification of the students in two groups with and without academic failure by using ANN with 15 neurons in the hidden layer is better than the LR model.

Key words: Logistic regression, Artificial Neural Network, Academic failure.

1. INTRODUCTION

Academic failure is one of the main problems of the universities so that it not only wastes the time and resources, but also cause other problems such as have psychological, family and social problems for the university students (1). UNESCO defines it as repeating an educational grade, early dropout or Reduced quality of education (2).

Every year these problems are increasing so that many of students cannot handle their academic courses and eventually leave the university (3).

The results of studies in developing and developed countries showed a many factors internal and external the educational systems effect on the success or failure of students (4).

There are many factors that may be effected on educational performance of students such as high school final grade, matriculation examination, age of admission, gender, economic problems, parental education and etc. (5, 6). An-

other study has pointed to issues (factors) such as socioeconomic status of the family, personality characteristics of student (4).

The ability to classify the student based on influential factors is very important to universities or educational institutions because strategic programs can be planned on improving or maintaining the students' performance during their studies in the university period (6).

For classifying/predicting of binary outcome variable (academic failure), some methods are available such as discriminant analysis, regression techniques, genetic algorithms, different data mining methods, decision tree and artificial neural network models. The general structure of artificial neural network was inspired by neurobiology of the human brain (7).

In theoretical works and more published reports of studies found that ANNs approach compared to tradi-

tional statistical methods such as regression analysis, discriminant analysis have better performance in predicting binary outcomes, especially when the relationship between the dependent and independent variables is complex (8, 9).

The results of a meta-analysis with 28 studies showed that in 36% of them, ANN, in 14% logistic regression method, performed better and in other studies (50% of cases) both modes had a similar performance (10). The number of studies compared ANN and logistic regression and it can be seen that both models perform on about the same level more often than not, with the more flexible neural networks generally outperforming logistic regression in the remaining cases (11).

In this study, we applied logistic regression model and ANN to predict the academic failure based on Effective factors and then compared the ability of each of these models to classify academic failure among students of Medical Sciences.

2. MATERIAL AND METHODS

2.1. Study population

In a cross-sectional study, data collected using a stratified random sampling from 275 undergraduate students in schools of nursing & midwifery and paramedic schools of Hormozgan University of Medical Sciences (HUMS) in the first semester of 2013. Bandar Abbas is the capital of Hormozgan province. This city is located in the south of Iran (north of the Persian Gulf) and it is one of the largest commercial ports in Iran. It is hot and humid in Bandar Abbas.

The data collection tool was a researcher made questionnaire which contained questions related to the factors affecting student's academic failure as is shown in table1. These questions by reviewing the valid literature and discussion with experts in the Medical Education Development Center (MEDC) of HUMS were identified.

In this study Grade Point Average (GPA) of the previous semester was the output or response variable which represented the performance of a student and other variable as shown in table1 were introduced as input or independent variables. According to the directive of the Ministry of Health and Med-

Variable	Values
Age (Year)	Less than 20 , 20 to 25 , More than 25
Gender	Male , Female
Place of residence	Home , Dormitory
Marital status	Married , Single
School	Nursing & Midwifery , Paramedic
Term	1 , 2 , ..., *10
Grade Point Average(GPA) of high school diploma	Score between 0 to 20
Father's education	Under high school diploma , Diploma , Upper diploma
Mother's education	Under high school diploma , Diploma , Upper diploma
Guest student	Yes , No
Transitional Student	Yes , No
At least once failed of a course	Yes , No
Extracurricular activities	Yes , No
Employment status	Yes , No
Interest in the field	Very much , Much , Moderate , Little , Very little
Grade Point Average (GPA) of the previous semester	Score (between 0 to 20)

Table 1. Student's Characteristics Used for predicting academic failure

ical Education of Iran in 2011, GPA of the previous semester less than 14 is considered as academic failure.

2.2. Artificial Neural Network

Artificial neural networks are one of the most popular models which can be used for prediction and classification a outcome variable. The structure of this model is inspired from neural networks of the human brain.

An ANN consists of a number of computational units called neurons and they are linked with each other by connections. ANN consists of three components or layers: a layer of "input" units is connected to a layer of "hidden" units, which is connected to a layer of "output" units (Fig 1) (12).

Currently the several neural networks have been intro-

Figure 1. Typical a feed forward multilayer perceptron

duced that each application their specific. One of the most common is Multilayer Perceptron (MLP) which is widely used in pattern recognition, prediction and no linear relationship. It is a feed forward neural network with an input layer and output layer and a number of hidden layers (Fig1). One hidden layer is generally sufficient for classifying most data sets and determine the number of hidden layer neurons is an important issue. In fact, network optimization depends on the number of neurons of the hidden layer. However, there is not a rule to select the number of hidden layer neurons, but to prevent of overloading, it is better not a large number (13).

A feed-forward network, the neurons in each layer only connect with the neurons in the next layer. These connections are unidirectional, which means signals or information being processed can only pass through the network in a single direction, from the input layer, through the hidden layer(s) (14).

The MLP takes a vector of real values (independent variables) and computes a linear combination of input data. If the calculated value is greater than a threshold value, the output of the MLP is equal to -1, otherwise it will be equal to +1 (15).

In the MLP with back error propagation, the transfer function in the hidden layer neurons is a nonlinear function such as hardlimit, linear or sigmoid which is able to associate training patterns with outputs. But for simple derivative and related derivatives with function, mostly used sigmoid function (14).

2.3. Statistical analysis

To fit ANN to data, the first step, dataset randomly divided into two sets; one set of 70% (193) cases for training and 30% (82) cases for testing the model. Then, a standard feed-forward, back-propagation neural network with three layers: an input layer, a hidden layer and an output layer. The input layer consisted of 15 neurons; in order to prevent overfitting to training data, the hidden layer contained a different number of neurons such as: 10, 15, 20, 25, 30, 35, 40, 45 and

50 and the output layer contained two neurons. We used the sigmoid function in both hidden layer and output layer for activation. The learning rate and momentum for network training were set respectively 0.25 and 0.9 and the models were run until a minimum average squared error < 0.063 was obtained.

Multivariate logistic regression (with the backward step-wise method) used to perform which variable outcome or dependent variable was academic failure (Score, < 14 and >= 14) and other variable were independent variables. Also Hosmer - Lemeshow goodness-of-fit test was used to compare the observed and predicted values of academic failure.

The Area Under the Receiver Operating Characteristic (AUROC), the indicators of sensitivity, specificity and kappa coefficient used to compare ANNs with different neurons and also for comparing the best ANN with logistic regression curves to evaluate the predictive accuracy of two models. The higher ROC areas indicated a better performance of the models.

The neural network development software used in this study was MATLAB 2010 and the statistical analyses, including descriptive statistics, logistic regression analyses and kappa statistic were performed by the IBM SPSS, version 22.0.

3. RESULTS

From 275 students, 77 (28%) suffered academic failure (GPA<14); 183 cases (66.5%) were female, 183 (66.5%) were less than 20 years old, and 216 (75.5%) were single. For 176 (64%) students education of the father and for 219 (79.6%) education of mothers were under high school diploma; 174 (63.3%) studied in paramedical school, 198 (72%) were unemployment, and 216 (78.5%) students resided in dormitory. 196 (76.3%) of them the GPA of diploma high school was greater than 16, 229 (82.3%) students were studying between second to fifth semester, 13 (4.7%) were guest student from other universities, and 8 (2.9%) were transferred from other universities. Also, 252 (91.6%) were interested in their field of study in moderate to very much, 94 (34.2%) had extracurricular activities in their college, 102 (37%) at least once fail of a course.

To select the most appropriate ANN, the nine perceptron models with 16 neurons in the input layer, one neuron in

Number of neurons	% Train Accuracy	% Test Accuracy	Incorrect Prediction	Az (±SD)*	Kappa statistic
ANN(10)	90.4	80.15	19.85	0.84±0.03	0.45
ANN(15)	86.8	79.10	17.90	0.89±0.02	0.59
ANN(20)	72.9	80	27.1	0.83±0.03	0.41
ANN(25)	80.5	78.91	21.09	0.85±0.03	0.46
ANN(30)	6.1	75.8	24.2	0.81±0.03	0.29
ANN(35)	78.9	75.33	24.67	0.83±0.03	0.37
ANN(40)	82.4	78.21	21.79	0.88±0.02	0.47
ANN(45)	82	75	25	0.83±0.03	0.47
ANN(50)	80.4	79	21	0.88±0.02	0.42

Table 2. Accuracy, incorrect prediction, Roc Curve Area and kappa for comparison ANNs models. *Az =AUROC and SD=Standard Deviation

output layer and different numbers of neurons in the hidden layer were used (Table2). Comparison of indicators such as train accuracy, test accuracy, incorrect prediction, AUROC,

and kappa statistics showed that the ANN with 15 neurons in the hidden layer, compared with other neural networks, had a better performance. Therefore, the desirable neural network that should be compared with the logistic regression model was a neural network with 15 neurons in the hidden layer.

Model	Az (±SD)	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
LR	0.55±0.04	0.22	0.47	0.62
ANN (15)	0.89±0.02	0.00*	0.85	0.93

Table 3. Test UAROC for Logistic regression and ANN (15), *P<0.0001

The UAROC for ANN (15) and logistic models separately were compared with the reference AUROC (Table 3). The results showed that ANN (15) model was significantly more than logistic regression (p<0.001).

Model	Failure	No Failure	Kappa statistic
Logistic Regression	Failure	29 (37.70)	0.45
	No Failure	14 (7.70)	
ANN with 15 Neuron	Failure	49 (63.30)	0.60
	No Failure	15 (7.60)	

Table 4. Classification of students based on logistic regression and ANN (15) models

The LR and ANN models classified 77.5% and 84.3% of students respectively, with and without academic failure correctly. The LR have sensitivity and specificity were 67.3% and 79.3%, respectively, and the ANN had a sensitivity and specificity of 76.6% and 86.7%, respectively. Also, the kappa statistics for LR was 0.452, showing that the emerged classification may be due to chance (P=0.215) and this statistic for the ANN was 0.60 which was significant; showing that the emerged classification was not due to chance (p<0.001) (Table 4).

Females were two times more than men, married students 3.3 more than singles, students with grade point (GPA) of diploma greater than or equal to sixteen 1.4 times more than others, students how were employed 2.3 times more than the unemployed subjects, students of nursing and midwifery 5.7 times more than paramedics school and students with at least once failed of a courses 2 times more than students without it was at risk of academic failure (Table 5). According to neural network results, five variables, i.e. grade point average of diploma, number of crash course (repeated), age, father's education, and sex had the greatest effect on how the network classifies the students.

	B	S.E.	Wald	Sig.	Exp (B)	95.0% C.I. for EXP (B)	
						Lower	Upper
Sex	.72	.34	4.40	.036	2.06	1.05	4.04
Marital	-1.20	.45	7.13	.008	.30	.13	.73
GPA.Diploma	.326	.079	17.12	.000	1.38	1.19	1.62
Employment	-.85	.38	5.08	.024	.43	.20	.89
School	1.73	.35	25.00	.000	5.66	2.87	11.17

Table 5. Odds ratios and coefficients of binary logistic regression analysis of factors associated with academic failure

Results of ANN (15) showed that respectively at least once failed of a course, GPA of high school diploma, type of school (Nursing & Midwifery or paramedical) Father's education were five effective factors on academic failure.

Figure 2. Factors affecting on academic failure in the order of their importance in ANN (15)

4. DISCUSSION

In this study, after changes in the number of hidden layer neurons, an MLP with 15 neurons in the hidden layer was the best structure among nine ANN models and then it was chosen to compare with LR model. But the result showed that by increasing the number of neurons in hidden layer cannot certainly effected on the performance of the neural network.

Significantly, the AUROC for ANN (15) model (89%) was considerably higher than LR model (55%). Also sensitivity for two models showed that the classification accuracy of academic failure in ANN (76.6%) was more accurate than the LR model (67.3%). So far, many studies in order to compare ANN and LR models in various fields such as medicine, economics, agriculture, etc. have been conducted but there are a few studies in the educational field. In a study almost the same, results showed that the correct classification in predicting graduate students for the ANN (93.3%) was higher than the discriminant analysis (81.5%) (16). Also in other studies showed that ANN model is far better than the LR model (7, 11, 17-20). Although most studies indicated ANNs are a technique alternative to conventional statistical methods for predicting, but there were a few studies which showed LR model have the same or better performance (10, 11). In addition the results of a meta-analysis study showed that in 36% of cases ANN and 14% LR performed better and in others both models functioned well (10).

In general ANN methods as semi-parametric methods have many advantages such as, allow a large number of variables in the model, no need to assumptions such as normality and..., finding the models despite missing data, detection of complex and nonlinear relationship between independent and dependent variables (21).

Although in theory and practical studies have been hinted ANN have better performance than statistical methods, but it has some disadvantage such as, the accuracy of the results depends largely on the size of the training set, requires the initialization and adjustment of many individual parameters to optimize the classification performance, standardized coefficients and odds ratios corresponding to each independent variable cannot be easily calculated, weights are generated in

a neural network analysis, but their interpretation is difficult, the weights may be influenced by the program used to generate them (22).

Logistic regression model is mainly influenced by the sample size, number of independent variables, potential multicollinearity and missing (23). Unfortunately, in this study the number of variables relative to the sample size is large and missing data observed, In this case we used ANN beside LR for academic failure data set. The ANN model indicated respectively five factors, at least fail of a course, GPA of high school diploma, type of school (nursing& midwifery or paramedical), father's education and interest in the field had the greatest impact on academic failure. Only GPA of high school diploma and school were significant variables common in ANN and LR models. Also, some studies noted the effect of these factors on academic failure (4) but other indicators such as school related, family related, student related and community related were the most important factors effect on academic failure (24) which in this study it was not possible to collect them and this was one of the limitations of our study.

Although the results of this study and previous studies showed that the ANNs has better performance, but it seems to perform a meta-analysis is a useful way to identify model with better performance for binary data.

5. CONCLUSION

In general, the results of this study showed that among 9 ANNs, an ANN with 15 neurons in the hidden layers had better performance. In comparison with the conventional LR model, the ANN model in the study was more accurate in predicting academic failure and had higher overall performance indices. Therefore, based on the results of other and academic failure data, it seems that for classification of a dichotomous dependent variable, artificial neural network methods are appropriate to be used.

Acknowledgments

The authors would like to thank Dr. Nasrin Shokrpour at Center for Development of Clinical Research of Nemazee Hospital for editorial assistance.

CONFLICT OF INTEREST: NONE DECLARED

REFERENCES

- Pan WD. *Causal attributions and affective reactions to academic performance of Chinese students in Hong Kong*. 1993, The Chinese University of Hong Kong.
- Changizy Ashtyani S, Shamsi M, Mohammadbeygi A. *Frequency of educational decline and some effective factors of student's opinion in Arak University of Medical Sciences, 2009*. Arak Medical University Journal. 2010; 12(4): 24-33.
- Pagani L. et al. *Effects of poverty on academic failure and delinquency in boys: A change and process model approach*. Journal of Child Psychology and Psychiatry. 1999; 40(8): 1209-1219.
- Iqbal Chohan B, Amin Qadir S. *Academic Failure at Primary Level: A Qualitative Approach to Primary Education in Pakistan*. Journal of Research & Reflections in Education (JRRE). 2013; 7(1).
- Oladokun V, Adebajo A, Charles-Owaba O. *Predicting students' academic performance using artificial neural network: A case study of an engineering course*. The Pacific Journal of Science and Technology. 2008; 9(1): 72-79.
- Rios G. et al. *Predicting Early Students with High Risk to Drop Out of University using a Neural Network-Based Approach*. in ICCGI 2013, *The Eighth International Multi-Conference on Computing in the Global Information Technology*. 2013.
- Heydari ST, Ayatollahi SMT, Zare N. *Comparison of Artificial Neural Networks with Logistic Regression for Detection of Obesity*. Journal of Medical Systems. 2012; 36(4): 2449-2454.
- Hornik K, Stinchcombe M, White H. *Multilayer feedforward networks are universal approximators*. Neural networks. 1989; 2(5): 359-366.
- BK R, Srivatsa S. *Evaluation of Logistic Regression and Neural Network Model With Sensitivity Analysis on Medical Datasets*. International Journal of Computer Science and Security (IJCSS). 2011; 5(5): 503.
- Song JH. et al. *Comparative analysis of logistic regression and artificial neural network for computer-aided diagnosis of breast masses*. Academic Radiology. 2005; 12(4): 487-495.
- Dreiseitl S, Ohno-Machado L. *Logistic regression and artificial neural network classification models: a methodology review*. Journal of biomedical informatics. 2002; 35(5): 352-359.
- Ayer T. et al. *Comparison of Logistic Regression and Artificial Neural Network Models in Breast Cancer Risk Estimation 1*. Radiographics. 2010; 30(1): 13-22.
- Popescu MC. et al. *Multilayer perceptron and neural networks*. WSEAS Transactions on Circuits and Systems. 2009; 8(7): 579-588.
- Jiang J, Trundle P, Ren J. *Medical image analysis with artificial neural networks*. Computerized Medical Imaging and Graphics. 2010; 34(8): 617-631.
- Cybenko G. *Approximation by Superpositions of a Sigmoidal function (Article)*. Mathematics of.
- Oancea B, Dragoescu R, Ciucu S. *Predicting students' results in higher education using neural networks*. in the *Proceedings of the International Conference on applied Information and Communication Technologies*. 2013.
- Green M. et al. *Comparison between neural networks and multiple logistic regression to predict acute coronary syndrome in the emergency room*. Artificial intelligence in medicine. 2006; 38(3): 305-318.
- Kazemnejad A, Batvandi Z, Faradmal J. *Comparison of artificial neural network and binary logistic regression for determination of impaired glucose tolerance/diabetes*. EMHJ. 2010; 16(6): 615-620.
- Pao HT. *A comparison of neural network and multiple regression analysis in modeling capital structure*. Expert Systems with Applications. 2008; 35(3): 720-727.
- Parsaeian M. et al. *Comparison of logistic regression and artificial neural network in low back pain prediction: second national health survey*. Iranian journal of public health. 2012; 41(6): 86.
- Taşdelen B. et al. *Artificial neural network analysis for prediction of headache prognosis in elderly patients*. Turkish Journal of Medical Sciences. 2009; 39(1): 5-12.
- Gan C. et al. *Consumer choice prediction: Artificial neural networks versus logistic models*. 2005: Lincoln University. Commerce Division.
- Ducher M. et al. *How to limit screening of patients for atheromatous renal artery stenosis in two-drug resistant hypertension?* J Nephrol. 2005; 18(2): 161-165.
- Tyler JH, Lofstrom M. *Finishing high school: Alternative pathways and dropout recovery*. The future of children. 2009; 19(1): 77-103.