



Semra TURKAN,  
Esra POLAT,  
Suleyman GUNAY

**CLASSIFICATION OF DOMESTIC  
AND FOREIGN COMMERCIAL BANKS  
IN TURKEY BASED ON FINANCIAL  
PERFORMANCES USING LINEAR  
DISCRIMINANT ANALYSIS,  
LOGISTIC REGRESSION AND ARTIFICIAL  
NEURAL NETWORK MODELS**

**Abstract**

The Data Mining (DM) techniques of linear discriminant analysis (LDA), logistic regression (LR) and artificial neural network (ANN) models are among the multivariate techniques used for predicting the predefined class membership of dependent variables. Hence, the aim of this study is to discuss and illustrate LDA, stepwise LDA, LR, forward LR and four types ANNs and compare these models' correct classification ability. For this purpose, the data of commercial banks operating in Turkey in two pre-defined groups, namely domestic and foreign banks, is used. In this study, the classification performance of ANN models against to LDA and LR is investigated. The ability of these classification methods in classifying the banks correctly is compared in terms of correct classification rates. As the results reveal that ANN (ANN-Prune) outperforms LDA and LR in terms of bank classification accuracy and thus, provide an effective alternative for implementing bank classification.

---

© Semra Turkan, Esra Polat, Suleyman Gunay, 2012.

Turkan Semra, Hacettepe University, Department of Statistics, Ankara, Turkey.  
Polat Esra, Hacettepe University, Department of Statistics, Ankara, Turkey.  
Gunay Suleyman, Hacettepe University, Department of Statistics, Ankara, Turkey.

**Key words:**

Linear Discriminant Analysis, Logistic Regression, Artificial Neural Network.

**JEL:** C380, C450, C190, C100.

## **1. Introduction**

In recent years, both developed and developing countries around the world have relaxed restrictions on foreign banking and most of them now permit more foreign banks to embark on more banking related activities in their domestic markets. This is due to the increasing importance of international trade in goods. Over the last decade, the number of foreign banks has increased in banking sector because of globalization. Several authors have addressed the potential benefits of foreign bank entry for the domestic economy. Foreign competition can enhance the efficiency of the domestic banking sector, improve knowledge and technological skills, provide access to foreign savings (Albayrak, 2009).

As compared to many other countries, the negative effects of the global financial crisis in 2008–2009 on the Turkish banking sector were limited. The reasons behind this are a high capital adequacy ratio, a high asset quality, low currency and liquidity risks thanks to successful risk management and effective public supervision and the good management of the interest, counterparty and maturity risks. The healthy financial structure in the Turkish banking system is mainly the result of the Banking Restructuring Program introduced in Turkey after 2001. Together with the program, measures were taken to keep the risks that can carry under control. The program mainly aimed at improving the financial structure of the domestic banks and Banking Regulatory and Supervisory Authority (BRSA) is responsible from the supervision of domestic, as well as the foreign banks operating in Turkey since then (Sen, 2010).

Data mining (DM) is systematic approach to find underlying structures and hidden relationships in huge databases: It searches for knowledge buried within difficult patterns of relationships in large amount of data. The research regarding DM can be classified into two categories: methodologies and technologies. The main methodologies are data visualization, statistical techniques and deductive database. The related applications using these methodologies can be summarized as classification, prediction, clustering, summarization, data reduction, de-

pendency modeling and sequential analysis. The technology part of DM consists of techniques such as statistical methods, neural networks, decision trees, genetic algorithms and non-parametric techniques. The DM techniques of linear discriminant analysis (LDA), logistic regression (LR) and artificial neural network (ANN) models are among the multivariate techniques used for predicting the predefined class membership of dependent variables. (Albayrak, 2009). Both of LDA and LR are appropriate for the development of linear classification models, i. e. models associated with linear boundaries between the groups (Pohar et al., 2004). LDA and LR techniques have been widely used in bank classification (Albayrak 2009).

There are some Turkish studies on domestic and foreign banks operating in Turkey. Isik and Hassan (2002) have examined the effect of bank size, corporate output, and governance, as well as ownership, on the cost and alternative profit efficiencies of Turkish banks by employed stochastic frontier approach. Unsal and Duman (2005) have investigated 32 public, private and foreign banks' performance where located in Turkey with using Factor Analysis. Unsal and Guler (2005) have searched which methods (LDA or LR) are the best to classify the 21 Turkish banks, which failed during the 1997–2003 period, as success and failure banks by using the financial ratios. Gungor (2007) has investigated the factors which affect the bank profitability by using 29 bank data covering 1990–2005 periods for Turkey. Tufan, Vasilescu, Cristea and Hamarat (2007) analyzed the performance of domestic and foreign banks in Turkey by categorizing them as success-failure. They used both the principle component analysis (PCA) and LR. Demirhan (2009) analyzed the change in the sector shares of domestic and foreign banks during the 1990–2007 period, using a panel data model with a qualitative dependent variable. Sen (2010) has investigated which factors are discriminating between domestic and foreign banks operating in Turkey (Sen, 2010; Tufan et al., 2007).

In this study, we discuss and illustrate LDA, stepwise LDA, LR, forward LR and four types ANNs and compare these models' correct classification ability for predicting the group membership of commercial banks into two pre-defined groups, namely domestic and foreign banks, on the basis of financial ratios. In addition, the other purpose of this paper is to identify the distinguishing financial ratios characterizing the operation of domestic and foreign banks in Turkey through stepwise LDA and forward LR. The financial ratios considered in the analysis cover most important aspects of financial performance; including capital ratios, balance sheet ratios, assets quality ratios, income-expenditure ratios, profitability ratios, liquidity ratios and activity ratios.

## 2 . Linear Discriminant Analysis

### 2.1. Two-Groups Linear Discriminant Analysis

LDA focuses on the association between multiple independent variables and a categorical dependent variable by forming a composite of the independent variables. This type of multivariate analysis can determine the composite variables discriminates between two or more preexisting groups of subjects and also can derive a classification model for predicting the group membership of new observations. The simplest type of LDA is two-group LDA which the dependent variable has two groups. In this case, a linear discriminant function (LDF) that passes through the means of the two groups (centroids) can be used to discriminate subjects between the two groups (Antonogeorgos et al., 2009). Two-group LDA is a linear combination of the two or more independent variables that discriminate best between a priori defined groups. Discrimination is achieved by setting weights for each independent variable to maximize the between-group variance to the within-group variance (Albayrak, 2009). LDF is represented by Eq. (1):

$$LDF = b_0 + b_1x_1 + b_2x_2 + \dots + b_px_p \quad (1)$$

Here,  $b_0$  is constant and  $b_1$  to  $b_p$  are the discriminant weights for the  $p$  independent variables. LDA computes the discriminant coefficients and selects the appropriate weights (cut-off score) that will separate the average values of each group, while minimizing the statistical distance of each observation and its own group means (Wongkhamdi and Seresangtakul, 2010). From the LDF, scores can estimate predicted probabilities and predicted group membership for every case on the dependent variable. This approach is based on the rationale that it is more likely that the independent and dependent variables are related as the between groups sum of square is larger relative to within-group sum of squares. Also the ratio of between-group divided by total sum of squares (explained variability) or of within-group divided by total sum of squares (unexplained variability) is used to assess the relationship (Antonogeorgos et al., 2009).

The LDF can also be written in standardized form, in which each variable is adjusted by being subtracted from its mean and divided by its standard deviation (Eq. 2):

$$z = \beta_0 + \beta_1z_1 + \beta_2z_2 + \dots + \beta_pz_p \quad (2)$$

The standardized LDF allows comparing variables measured on different scales (Antonogeorgos et al., 2009). Here,  $z$  is the standardized LDF,  $\beta_1$  to  $\beta_p$  are the standardized coefficients and  $z_1$  to  $z_p$  are the standardized variables. The variables that contribute most to the prediction of group membership are the ones with the largest, standardized regression coefficients. The mean of  $z$  taken

over all observations is zero, because the mean of each variable, when standardized, is zero. Therefore, an object can be classified into one group if its  $z$  score is greater than zero and into the other group if its  $z$  score is less than zero (Albayrak, 2009).

## 2.2. Stepwise Linear Discriminant Analysis

Stepwise LDA is an alternative method to the full model approach. When a number of potential discriminator variables are known, but there is no suggestion as to which would be the best set of variable for forming the discriminant function. Stepwise LDA is a useful technique for selecting the best set of discriminator variables to form the discriminant function. In forward stepwise analysis, all variables are evaluated in the first step to determine which one provides the most significant and unique discrimination between groups. Once this variable has been included in the model, all remaining variables are evaluated to determine which one provides the next best discrimination. The procedure continues until the addition of a new variable does not significantly improve the discrimination between groups (Albayrak, 2009).

## 3. Logistic Regression Analysis

The LR analysis is a statistical technique which has been used for prediction and determining the most influential independent variables on the dependent variable (Abdolmaleki et al., 2004). In LR models, dependent variable is always in categorical form and has two or more levels. Independent variables may be in numerical or categorical form (Albayrak, 2009). LR model, according to independent variables, is a regression model from which the expected values of the dependent variable are obtained as a probability (Tufan et al., 2007). The simplest optimizing method of discrimination is to maximize the posterior probability of correct allocation. To obtain the posterior probability, the logit coefficients could be estimated using the maximum likelihood estimation method (Abdolmaleki et al., 2004).

LR approach is commonly applied to predict membership in two groups using a set of predictors. LR model or logit deals with the binary case, where the dependent variable consists of just two categorical values (Abdolmaleki et al., 2004). If there are two groups, binary logistic regression (BLR) is used (Albayrak, 2009). The BLR model is defined as below:

$$L = \ln(p/1-p) = b_0 + b_1x_1 + b_2x_2 + \dots + b_px_p \quad (3)$$

Where  $p/(1 - p)$  is the ratio of the probability of a success to the probability of a failure and called the odds ratio and  $\ln(p/1 - p)$  the logit transform of  $p$ ;  $x_p$  is the  $p$ th predictor variable and  $b_p$  is the coefficient of the  $p$ th predictor variable. In this equation, the logit transform is used to relate the probabilities of group membership to a linear function of the predictor variables. The parameters of the logistic model ( $b_0$  to  $b_p$ ) are derived by the method of maximum likelihood (Albayrak, 2009). Each regression coefficient describes the size of the contribution of the corresponding independent variable to the dependent variable. The effect of the independent variables on the dependent variable is commonly measured by using the odds ratio of the independent variable, which represents the factor by which the odds of an outcome change for a one-unit change in the independent variable. The odds ratio is estimated by taking the exponential of the coefficient (eg,  $\exp [b_1]$ ) (Ayer et al., 2010).

In BLR model, one can directly estimate the probabilities of an objects occurring as in Eq. 4:

$$p = \frac{1}{1 + e^{-(b_0 + b_1x_1 + \dots + b_px_p)}} \\ 1 - p = \frac{1}{1 + e^{(b_0 + b_1x_1 + \dots + b_px_p)}} \quad (4)$$

The BLR model can be manipulated to define a probability for classifying the objects in to the pre-defined groups. The first step is to arrange Eq. 4 to express the probabilities of group membership in terms of independent variables directly. In the second step, the LR coefficients,  $b_1$  to  $b_p$  and the constant,  $b_0$  are used to define a model for classifying objects into one of the two groups. An object with an equal probability of belonging to the two groups has  $p = 1 - p$ , which means that  $\ln(p/1 - p) = 0$ . In this case, the following equation can be written by using Eq. (3):

$$b_0 + b_1x_1 + \dots + b_px_p = 0 \quad (5)$$

Therefore, an object can be assigned to first group or second group according to the following rules:

Classify into first group if  $b_0 + b_1x_1 + \dots + b_px_p > 0$

Classify into second group if  $b_0 + b_1x_1 + \dots + b_px_p < 0$

These rules are based on a critical probability ( $p_c$ ) value (cut-off value) of 0.5. If it is decided to use a different  $p_c$  value, the following general rules can be applied:

Classify into first group if  $b_0 + b_1x_1 + \dots + b_px_p > \ln(p_c / 1 - p_c)$

Classify into second group if  $b_0 + b_1x_1 + \dots + b_px_p < \ln(p_c / 1 - p_c)$

As mentioned above, LR assumes that a link function (in this case, the logit-transform) can be used to relate the probabilities of group membership to a linear function of the predictor variables. It is also assumed that the observations are independent (Albayrak, 2009). The LR method is relatively robust, flexible, easily used and it lends itself to a meaningful interpretation (Pohar et al., 2004).

In LR model, significant variables can be selected with various methods; forward and backward selection. In forward selection, variables are sequentially added to an «empty» model (i. e., a model with no predictor variables) if they are found to be statistically significant in predicting a dependent variable. In contrast, backward selection starts with all of the variables in the model and the variables are removed one by one as they are found to be insignificant in predicting the outcome (Ayer et al., 2010).

In this study, forward selection is used for LR to determine significant independent variables.

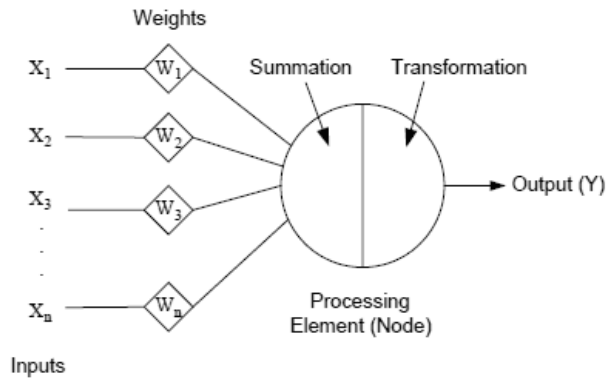
#### 4. Artificial Neural Network

ANNs are mathematical models that have ability to find complex relationships between two data set (Behzadi and Aslaminejad, 2010). ANN models consist of highly interconnected neurons or nodes whose functionality is based on biological neurons. Their overall ability to help predict outcomes is determined by the connections between these neurons. Each neuron receives one or more inputs, processes by those inputs and generates a single output. Main components of information processing in the Neural Networks are: Inputs, Weights, Summation Function (weighted average of all input data going into a processing element (PE)), Transformation function and Outputs. Figure 1 illustrates the general structure of an ANN (Razi and Athappilly, 2005).

The network has three layers: input, hidden and output layer. Network inputs are fed both of the input layer and the output layer and the output layer produces output. The middle layer is called «hidden layer», since it does not communicate with the environment (that is inputs and outputs) directly, however, it allows the ANN to model complex relationships between the input variables and the outcome. All three layers contain a number of neurons that are connected by means of connection weights. Each neuron contains two elements: the summation node and the sigmoid transfer function. The summation node calculates the product of each normalized input value and the weight value. The problem of modeling consists of finding a proper set of connection weights using a suitable optimization algorithm so that the error between predicted and experimental outputs is minimized (Zurada and Lonial, 2005). The procedure of estimating the optimal weights that generate the most reliable outcomes is called learning or training. Training an ANN is analogous to estimating parameters in a LR model. There are several algorithms for training ANNs, the most popular of which is backpropagation (Ayer et al., 2010).

Figure 1

Illustrates the general structure of an ANN



ANN models are used in a variety of applications, for example, nonlinear mapping, data reduction, pattern recognition, clustering, and classification. In this study, three-layer feed-forward neural network with back propagation is used to make a classification application by using the SPSS Clementine 11. In this program, there are many different ways of performing the ANN: Quick, Dynamic, Multiple and Prune (Eriksson, 2010).

*Quick:* This method uses rules of thumb and characteristics of the data to choose an appropriate shape (topology) for the network.

*Dynamic:* This method creates an initial topology but modifies the topology by adding and/or removing hidden units as training progresses.

*Multiple:* This method creates several networks of different topologies (the exact number depends on the training data). These networks are then trained in a pseudo-parallel fashion. At the end of training, the model with the lowest root mean square error is presented as the final model.

*Prune:* This method starts with a large network and removes (prunes) the weakest units in the hidden and input layers as training proceeds. This method is usually slow, but it often yields better results than other methods.

In this study, these four methods are used to perform the ANN for determining which of them has the best predictive ability.



## 5. Application

The aim of this application is to classify domestic and foreign commercial banks according to the financial performances. The data used in this study involves 14 domestic and 17 foreign commercial banks operating in Turkey in 2009. The variables used in this study involve financial ratios as a basis to classify a bank's performance. There are many ratios in financial tables which show the success level of banks. However, there is no consensus on the issue of which one of these ratios should be taken into consideration when making judgments (Erdoğan, 2008). Hence, in this application one or more ratios are selected from seven main performance ratios groups in order to decrease possibility of multicollinearity among variables. Firstly, the high correlated variables are determined according to correlation matrix and these variables are omitted. Afterwards, the nineteen ratios, computed from the balance and income tables, are selected to measure the performance of domestic and foreign banks. These financial ratios are given as below.

### Capital Ratios

**S1:** Shareholders' Equity / (Amount Subject to Credit Risk + Market Risk + Operational Risk); **S2:** On Balance-sheet FC Position / Shareholders' Equity; **S3:** Net on Balance-sheet Position / Total Shareholders' Equity

### Balance Sheet Ratios

**B1:** TC Assets / Total Assets; **B2:** FC Assets / FC Liabilities; **B3:** TC Loans and Receivables / Total Loans and Receivables; **B4:** Funds Borrowed / Total Assets.

### Assets Quality Ratios

**A1:** Financial Assets (Net) / Total Assets; **A2:** Loans under follow-up (net) / Total Loans and Receivables; **A3:** Consumer Loans / Total Loans and Receivables

### Income-Expenditure Ratios

**G1:** Net Interest Income / Total Assets; **G2:** Interest Income / Total Expenses

### Profitability Ratios

**K1:** Net Profit (Losses) / Total Assets; **K2:** Net Profit (Losses) / Paid-In Capital

### Liquidity Ratios

**L1:** Liquid Assets / Total Assets; **L2:** Liquid Assets / Short-Term Liabilities

### Activity Ratios

**AC1:** Reserve for Employee Termination Benefit / Number of Personnel (Thousand TRY); **AC2:** Total Operating Income / Total Assets; **AC3:** Net Operating Income(Loss) / Total Assets

Although omitting the most correlated variables according to correlation matrix, there is still high correlated variables among the rest of them. Therefore, it is decided to use factor analysis for avoiding multicollinearity problem and keeping a better balance between the number of variables and the sample size. It should be mentioned that at the beginning of the application, it is tried to apply the factor analysis before omitting the high correlated financial ratios only by considering the correlation matrix. However, in this case, many factors are obtained and these factors couldn't explain the variance sufficiently. So that in order to obtain less factors explaining the variance more highly, it is preferred to apply factor analysis to the data after omitting the high correlated financial ratios. As the result of factor analysis, nineteen financial ratios are collected in five factors and these factors explain 84% of total variance. These five factors are given as below.

Factor 1	S1, B2, L1, L2, K1, G1, AC2, AC3
Factor 2	B1, B3, A2
Factor 3	S2, S3, B4, G2
Factor 4	A3, K2
Factor 5	A1, AC1

DM techniques of LDA, LR and ANN, which are built in SPSS Clementine 11, are used to develop classification for predicting the group membership of commercial banks into two pre-defined groups, namely domestic and foreign banks. In this case, LDA, stepwise LDA, LR, forward LR and four types ANN models are developed using five factors as independent variables and the status of the banks (1 for domestic banks and 2 for foreign banks) as dependent variable. Stepwise LDA and forward LR analysis are used to see the most important variables. The results of stepwise LDA and forward LR analysis are given in Table 1 and Table 2, respectively.

As the result of stepwise LDA and forward LR analysis, the same variable is taken into model. As seen from Table 1 and Table 2, Factor 3 is found the most important factor that affects the bank performance significantly. Thus, S2, S3, B4, and G2 variables are most important financial ratios in classifying the banks to pre-defined classes, namely domestic and foreign.

Table 3 shows the summary of each model's predictive ability that measured by percentage of correct classification. As shown in Table 3, ANN-Prune model has the best predictive ability with a percentage of correct classification of 87.1. In addition, the subsequent second best models in terms of predictive ability are LDA, forward LR and ANN-Quick models that they have the same correct classification rate of 80.65%.

Table 1

**Stepwise discriminant analysis results**

	Linear Discriminant Analysis	
	Sutstructure Matrix	CDF
Constant		
Factor 1	0.057	
Factor 2	-0.085	
Factor 3	1.000	1.173*
Factor 4	0.072	
Factor 5	-0.082	
Model Summary a) Wilks' Lamda = 0.702 (0.001)		
Statistics b) Canonical Corr. = 55%		
CDF: Canonical Discriminat Function; *: 0.05		

Table 2

**Logistic regression analysis results**

	Logistic Regression			
	B	Wald	Sig.	ExpB
Constant	0.857	2.052		
Factor 1				
Factor 2				
Factor 3	-3.313	5.898	0.015*	0.036
Factor 4				
Model Summary a) Nagelkerke R <sup>2</sup> = 54%				

Table 3

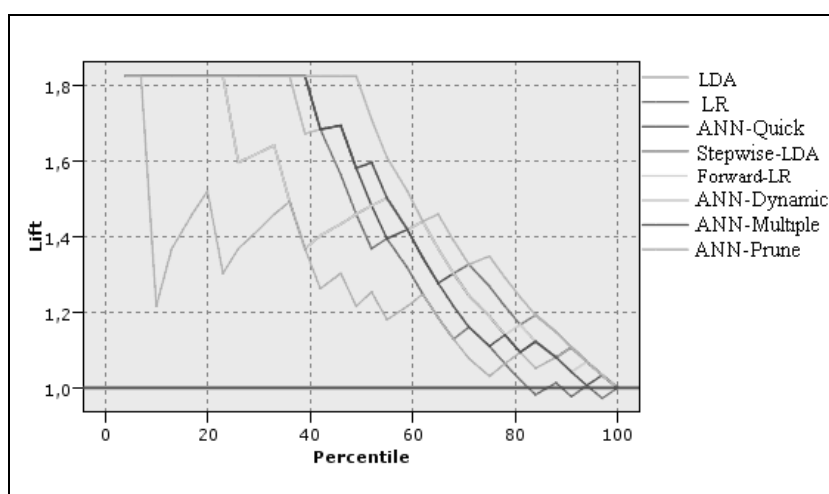
**Correct classification rates of models**

Model	Correct Classification Rates
LDA	80.65%
LR	74.19%
Stepwise LDA	70.97%
Forward LR	80.65%
ANN-Quick	80.65%
ANN -Dynamic	54.84%
ANN -Multiple	77.42%
ANN-Prune	87.1%

The lift chart of LDA, stepwise LDA, LR, forward LR and four types ANN models is obtained as in Figure 2. Lift chart shows how much better the model performs compared to the baseline model. Therefore, lift charts are visual tools for measuring the model performance. These charts consist of the lift curves for each model and the baseline (the horizontal line). The greater area between the lift curve and the baseline, the better model (Wah, 2006). Hence, Figure 2 shows that ANN-Prune is the best model.

Figure 2

**Lift chart of models**



## 6. Summary and Conclusions

In this study, the banks operating in Turkey in 2009 are classified by using LDA, stepwise LDA, LR, forward LR and four types ANN models in terms of their financial ratios. These models are applied to the data and compared to get the best prediction model for correct classification of the commercial banks as to their pre-defined groups. The results show that ANN-Prune model has the best predictive ability with a percentage of correct classification of 87.1. In addition, LDA, forward LR and ANN-Quick models are the second best models giving similar results in terms of correct classification rate. In conclusion, the study reveals that ANN-Prune is the best model for correctly classification of banks as domestic or foreign. Furthermore, Stepwise LDA and forward LR are used to de-

termine discriminating factors among domestic and foreign banks. The analyses show that these two models give the same results in terms of most important financial ratios in classifying the banks. The most important financial ratios for classification are «On Balance-sheet FC Position / Shareholders' Equity», «Net on Balance-sheet Position / Total Shareholders' Equity», «Funds Borrowed / Total Assets» and «Interest Income / Total Expenses» according to stepwise LDA and forward LR analyses.

### Bibliography

1. Albayrak, A. S. (2009), «Classification of Domestic and Foreign Commercial Banks in Turkey Based on Financial Efficiency: A Comparison of Decision Tree, Logistic Regression and Discriminant Analysis Models», *Suleyman Demirel University The Journal of Faculty of Economics and Administrative Sciences*, 14(2):113–139.
2. 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*.
3. Ayer, T. – Alagoz, O. – Kahn, C. E. – Woods R. W (2010), «Comparison of Logistic Regression and Artificial Neural Network Models in Breast Cancer Risk Estimation», *Informatics in Radiology*.
4. Behzadi, M. R. B. – Aslaminejad, A. A. A. (2010), «Comparison of Neural Network and Nonlinear Regression Predictions of Sheep Growth», *Journal of Animal and Veterinary Advances*, 9: 16.
5. Erdogan, B. E. (2008), «Bankruptcy Prediction of Turkish Commercial Banks Using Financial Ratios», *Applied Mathematical Sciences*, 2(60): 2973–2982.
6. Eriksson, B.G. – Sundh, V. (2010), «Prediction of seven years survival by artificial neural network and logistic regression: A comparison of results from medical and social data among 70-years old in Göteborg, Sweden», *Manuscript*.
7. Razi, M. A. – Athappilly, K. A. (2005), «Comparative Predictive Analysis of Neural Networks (NNs), Nonlinear Regression and Classification and Regression Tree (CART) Models», *Expert Systems with Applications*, 29: 65–74.
8. Sen, A. B. (2010), «Factors that Discriminate Between Domestic and Foreign Banks Operating in Turkey», *Marmara Üniversitesi İ. İ. B. F. Dergisi*, 28:1, 445–462.

9. Tufan, E. – Hamarat B. – Cristea, M. – Vasilescu, L. G. (2007), «Evaluation Of Turkish Domestic and Foreign Banks By Using Financial Ratios» *11th International Finance and Banking Symposium*, 814–829.
10. Wah, Y. B. (2006), «Some Applications of Data Mining», *Paper Presented at the National Statistics Conference*, Putrajaya International Convention Centre, Putrajaya, Malaysia.
11. Wongkhamdi T. – Seresangtakul P. A. (2010), «Comparison of Classical Discriminant Analysis and Artificial Neural Networks in Predicting Student Graduation Outcomes», *Proceedings of the Second International Conference on Knowledge and Smart Technologies*.
12. Zurada, J. – Lonial, S. (2005), «Comparison of the Performance of Several Data Mining Methods for Bad Debt Recovery in the Healthcare Industry», *The Journal of Applied Business Research*, 21: 2.