J. Indones. Math. Soc.
Vol. 25, No. 3 (2019), pp. 325–335.

PREDICTIVE ANALYSIS OF EMPLOYEE TURNOVER: A COMPARATIVE STUDY USING LOGISTIC REGRESSION AND ARTIFICIAL NEURAL NETWORK

Maria Zefanya Sampe[1], Eko Ariawan[2], and I Wayan Ariawan[3]

**[1]Faculty Member of Business Mathematics,
School of Applied STEM, Universitas Prasetiya Mulya,
Jalan BSD Raya Utama, BSD City, Tangerang 15339, Indonesia,
maria.zefanya@prasetiyamulya.ac.id**

**[2]Faculty Member of Enterprise Software Engineering,
School of Applied STEM, Universitas Prasetiya Mulya,
Jalan BSD Raya Utama, BSD City, Tangerang 15339, Indonesia,
eko.ariawan@pmbs.ac.id**

**[3]Risk Management Section Head, PT Kookmin Best Insurance,
ariawan@kbinsure.co.id Indonesia**

Abstract. Employee turnover is a common issue in any company. A high turnover phenomenon becomes a big problem that will certainly affect the performance of the company. Therefore, measuring employee turnover can be helpful to employers to improve employee retention rates and give them a head start on turnover. A study to analyze for employee loyalty has been carried out by using Logistic Regression (LR) and Artificial Neural Networks (ANN) model. Response variables such as satisfaction level, number of projects, average monthly working hours, employment period, working accident, promotion in the last 5 years, department, and salary level are used to model the employee turnover. Parameters such as accuracy, precision, sensitivity, Kolmogorov-Smirnov statistic, and Mean Squared Error (MSE) are used to compare both models.

Keywords and Phrases: predictive analysis; logistic regression model; artificial neural network; employee loyalty.

2000 Mathematics Subject Classification: 92B20, 62M10.
Received: 01-11-2018, accepted: 01-04-2019.

Abstrak. Pergantian karyawan adalah hal yang biasa di setiap perusahaan. Fenomena pergantian karyawan yang tinggi menjadi sebuah masalah besar yang akan memengaruhi kinerja perusahaan. Suatu studi untuk menganalisis loyalitas karyawan telah dilakukan dengan menggunakan model regresi logistik dan jaringan syaraf tiruan. Peubah-peubah respon seperti tingkat kepuasan, jumlah proyek yang telah dikerjakan, rataan jam kerja bulanan, lama bekerja, kecelakaan kerja, ada tidaknya promosi 5 tahun terakhir, departemen, dan tingkat gaji digunakan untuk memodelkan loyalitas karyawan. Parameter-parameter seperti akurasi, presisi, sensitivitas, statistik Kolmogorv-Smirnov, dan rataan galat kuadrat digunakan untuk membandingkan kedua model.

Kata kunci: predictive analysis; logistic regression model; artificial neural network; employee loyalty.

1. INTRODUCTION

As a company's core asset, employees are one of the determinants of the success or failure of a company. The performance of employees who are loyal to the company is interpreted as an intentional commitment to the best interests of one's boss, even sacrificing some personal interests and other moral duties [9]. Another definition states that loyalty can correspondence with the following expression: relationship of trust, resistance to adoption of opportunistic behavior faced with offers of outside employment [11]. An employee's loyalty can also be associated to job satisfaction representing a match between real and reward. Job satisfaction is also closely related to the behavior of individuals in the workplace [12].

Now days, generally every organization encounter the loyalty problems. The reward system aspect in an organization also plays an important role in increasing employee job satisfaction, higher rewards and satisfied employees in the workplace resulting in higher productivity from business organizations [13]. The level of employee churn is an important aspect that must also be calculated by the company. The high employee churn rate certainly has an impact on employee-churn costs [14]. The company not only provides the employee budget that invested through training, education, and benefits but also prepare extra budgets to ensure that the best employees consistently loyal and willing to contribute in the company.

Researcher also put their attention to employee loyalty in several research context such like working hard, providing higher quality service to customers, reduce intentions to quit and organizational performance. There are two factors that affect employee satisfaction. Company policy factors and work climate are the first factors in employee satisfaction. While the second factor is related to individual characteristics of employees who can be determined from status and seniority[1].

In the past, when an employee were hired they would be stay until their retirement. But nowadays the phenomenon is changing. The loyalty of an employee may depend on the corporate downsizing condition, company relocation, benefit they get and many more. As there are many factors affecting to employee loyalty, the study of the loyalty dimension can be reviewed from internal and external aspect. The internal dimension is more related to emotional aspect such like feelings of caring, of affiliation and of commitment. The external dimension is more related with the way loyalty manifests itself. This aspect is comprised of the behaviors that display the emotional component and is the part of loyalty that changes the most. On this research we use several variable such like the organization, satisfaction, evaluation index, project, average working time, spent time, salary level and more as the parameter input for Logistic Regression and Artificial Neural Network. The results presented in this paper are a comparison between the level of accuracy, precision, sensitivity, Statistics KS, and MSE for the logistic regression model and ANN.

2. THEORETICAL PRELIMINARIES

2.1. Logistic Regression

The relationship between the predictor variable and the response variable is not linear, then it is called non-linear regression [5]. One form of non-linear regression is logistic regression consisting of predictor variables which might be discrete, continuous, dichotomous or a combination of the three response variable. For the response variable, the output form of is a dichotomous variable or binomial distribution that represents the occurrence or not of an event that is stated in the number 0 or 1.

2.1.1. Logistic Regression Model

Logistic regression will use the link function for the binomial distribution. The linear equation for the binomial distribution using GLM as follows:

$$\ln\left(\frac{np}{1-np}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p. \tag{1}$$

Logistic regression is a *Odinary Least Square* approach on dichotomy qualitative dependent variable with minimum scale of 0 to express no characteristic and 1 with characteristic. Logit function for logistic regression as follows:

$$g(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n \tag{2}$$

with

$$\begin{aligned} g(x) &= \text{logit estimation} \\ \beta_i &= \text{parameter for constant variables} \\ x_i &= \text{predictor variable for } i, \text{where } i = 1, \cdots, n. \end{aligned}$$

2.1.2. Logit Function Model

Logit function is a transformation result of ln from the probability ratio from *odds ratio* described as:

$$g(x) = \ln\left(\frac{p}{1-p}\right). \tag{3}$$

The $g(x)$ function is linear to the parameter and $-\infty < g(x) < \infty$. The probability of p on interval 0 to 1 is based on the logit function transformation. Logit transformation used for creating linear function from the parameters implemented on logistic regression model. Logit function model expressed as follows:

$$\prod_{i=1}^{N} \left(\frac{p_i}{1-p_i} \right)^{y_i} (1-p_i)^{n_i}. \tag{4}$$

Furthermore the value of p_i can be obtained by:

$$p_i = \frac{e^{\sum\limits_{n=0}^{N} x_{ik}\beta_k}}{1 + e^{\sum\limits_{n=0}^{N} x_{ik}\beta_k}} \tag{5}$$

with $i = 1, \cdots, N$ and $k = 1, \cdots, K$.

2.2. Artificial Neural Network

Artificial neural network concept uses the form of how neural works in brain. Every neuron consists of body cell, dendrite and axon [6]. A neuron in brain is a biological cell with special ability to process information. Inside the body cell there is a nucleus which contains the characteristics information that can be inherited. Nucleus contains plasma material that produce nutrient needed by the neuron.

Neuron process information by receiving a signal from other neuron through dendrite and transmit the signal to body cell using axon all the way to synapses. This part has the main function to join each neuron. Electrical pulses strength in synapses depends on the neuron transmitter amount filling the synapses.

Artificial neural network is a system consist of many processing element connected parallel each other to control specific function defined by the network structure, type of the network and node function [2]. Neural network imitate brain ability to process information with the following condition:

(i). Artificial neural network have the knowledge from learning process.
(ii). Strength neuron connection is known as synapses weight to store the information.

Artificial neural network consist of three layer [3]. The first layer is input layer, second layer is hidden layer and the last layer is output layer. Sometimes the hidden layer does not exist. Raw information acquired by the neuron into the network is an unit input activity. Activity on every hidden layer connected by weight on every node. The activity on hidden layer ends up to output layer.

3. DATA AND METHODS

3.1. Data Overview

The data used for this research is extracted from Kaggle HR data [8]. 14,999 observations are loaded into the system. These data will be divided into 2 data sets: training and testing data set with composition of 80% and 20%, respectively. 12,000 observations will be chosen randomly to train the model and the rest will be used to test the model. The data contains 10 variables of employee information such as:

(1). **Status**
Status is the independent or target variable of the models consisting of 2 values where 0 means the employee will stay and 1 means the employee will leave.

(2). **Current satisfaction level**
Satisfaction levels are expressed on a scale of 0 to 1 which indicates the level of employee satisfaction from low to high.

(3). **Last evaluation level**
The previous evaluation of the employee satisfaction levels that vary from 0 to 1.

(4). **Number projects**
The number of projects that have been done by employees.

(5). **Average monthly working hours**
The average employees working time per month.

(6). **Employment period**
This refers to how many years employees have worked at the company.

(7). **Working accident**
The number of work accidents experienced during employment period.

(8). **Promotion in the last 5 years**
Indicator of employee promotion where 0 if the employee has not been promoted in the last 5 years, otherwise 1.

(9). **Department**
The departments in the company such as Accounting, Sales, Human Resources, Technical, Support, Management, IT, Product Manager, Marketing, and R & D.

(10). **Salary level**
Salary classifications are categorized into three levels: low, medium, and high.

The summary of the variables can be seen in the charts and tables below.

Table 1. Statistics Descriptive of Some Variables

Parameter	Satisfaction Level	Last Evaluation	Number Project	Average Monthly Hours	Employment Period
Minimum	0.09	0.36	2.00	96.00	2.00
1st Quartile	0.44	0.56	3.00	156.00	3.00
Median	0.64	0.72	4.00	200.00	3.00
Mean	0.61	0.72	3.80	201.05	3.50
3rd Quartile	0.82	0.87	5.00	245.00	4.00
Maximum	1.00	1.00	7.00	310.00	10.00

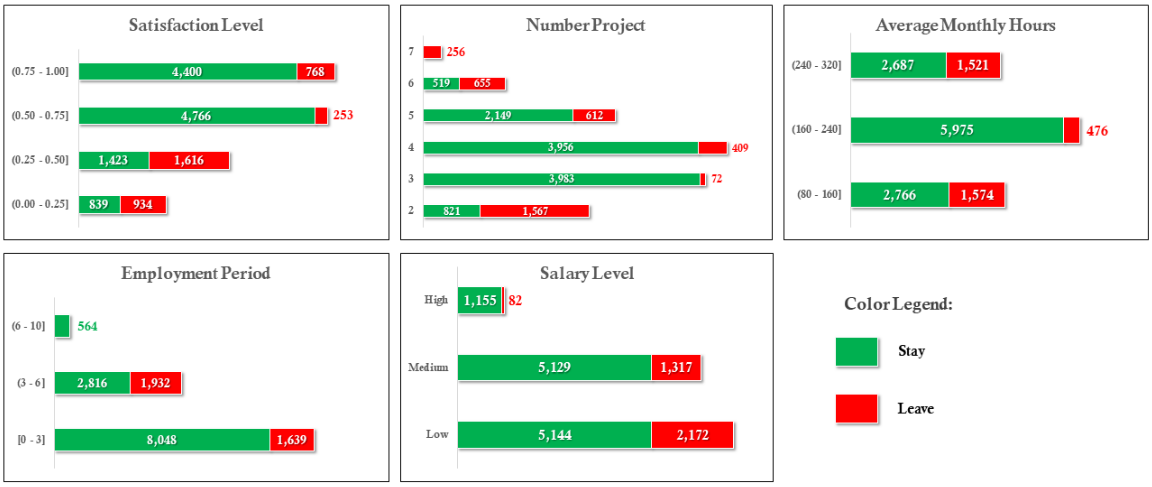

Figure 1. Chart for Some of Dependent Variables

Table 2. Summary of Promotion Last 5 Years

Promotion Last 5 Years	**Stay**	**Leave**	**Total**
No	11,128	3,552	14,680
Yes	300	19	319
Total	**11,428**	**3,571**	**14,999**

Table 3. Summary of Work Accidents

Work Accident	**Stay**	**Leave**	**Total**
No	9,428	3,402	12,830
Yes	2,000	169	2,169
Total	**11,428**	**3,571**	**14,999**

3.2. Model Structure

The turnover of an employee will be predicted using 2 popular classification models: Logistic Regression (LR) and Artificial Neural Network (ANN). In order to make a fair comparison, both models will be trained using the same training data sets and tested using the same testing data sets. In the same spirit, LR and ANN will use 9 input variables since LR has the lowest Akaike Information Criterion (AIC) value when using all of the dependent variables. In LR model selection, the lower AIC value suggests the better at measuring of model fit.

There are two factors that should be considered when building ANN model, namely: number of hidden layers used and number of neurons in each hidden layer. In **Introduction to Neural Networks for Java** by Jeff Heaton [15], the author suggests that one hidden layer is sufficient for most of the problems since problems that require two hidden layers are rarely encountered and extremely hard to train. Hence, for the practical reason and simplicity, one hidden layer will be used in the ANN structure. He also adds that the number of neurons in the hidden layer should be between the size of the input layer and the size of the output layer. Furthermore, he suggest the number of neurons should be 2/3 the size of the input layer plus the size of the output layer. In this case, the number of neurons is between 1 and 9 and it should be 7 neurons. Hence for this reason, 7 neurons will be used in the ANN and in order to obtain a better ANN model, 3 and 5 neurons will also be used for the comparison purpose.

3.3. Model Comparison

The model comparison will use the metrics: Accuracy, Precision, and Sensitivity, where they are commonly found in a confusion matrix. There are 4 conditions in a confusion matrix such as True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). The interpretation of all conditions can be easily understood by seeing the table below.

TABLE 4. Confusion Matrix Table

	Predicted Value	
Actual Value	Yes	No
Yes	TP	FN
No	FP	TN

Therefore, using the information in the table above, the three metrics are formulated as:

(i). **Accuracy:** $\dfrac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FN}+\mathrm{FP}+\mathrm{TN}}$

(ii). **Precision:** $\dfrac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$

(iii). **Sensitivity:** $\dfrac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$.

Two additional metrics will be used to have a better comparison, i.e.

(i). **Mean Square Error (MSE)** is the average squared difference between the actual values and predicted values;

(ii). **Statistics Kolmogorov Smirnov** is a non parametric test to determine whether the actual values and predicted values come from the same distribution.

Monte Carlo cross-validation method is used to make robust comparison metrics by repeating the simulation several times and then the results are averaged over the total number of simulations. The process simulation is presented in the flow chart below.

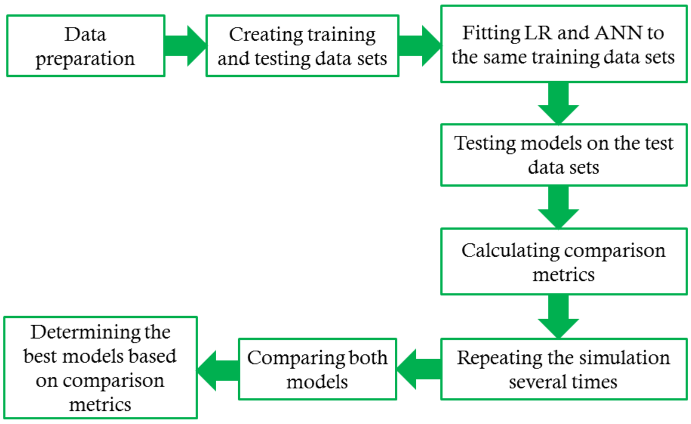
Data preparation
Creating training and testing data sets
Fitting LR and ANN to the same training data sets
Testing models on the test data sets
Calculating comparison metrics
Repeating the simulation several times
Comparing both models
Determining the best models based on comparison metrics

FIGURE 2. Flowchart of Simulation

4. DISCUSSION

4.1. ANN Model Selection

In order to obtain the best ANN model, the simulation runs 3 ANN models, namely:

(i). **ANN3** is the ANN model that uses 1 hidden layer and 3 neurons in the hidden layer;
(ii). **ANN5** is the ANN model that uses 1 hidden layer and 5 neurons in the hidden layer;
(iii). **ANN7** is the ANN model that uses 1 hidden layer and 7 neurons in the hidden layer.

The simulation results are presented in the table and chart below.

TABLE 5. Comparison of ANN Models

Model	Accuracy	Precision	Sensitivity	KS Stat	MSE
ANN3					
Mean	0.95573	0.90772	0.91083	0.86553	0.04427
SD	0.01692	0.05595	0.01234	0.01391	0.01692
ANN5					
Mean	0.96571	0.94277	0.91237	0.86776	0.03429
SD	0.00818	0.02876	0.01224	0.00896	0.00818
ANN7					
Mean	0.96959	0.95513	0.91553	0.87031	0.03041
SD	0.00403	0.01485	0.01046	0.00829	0.00403

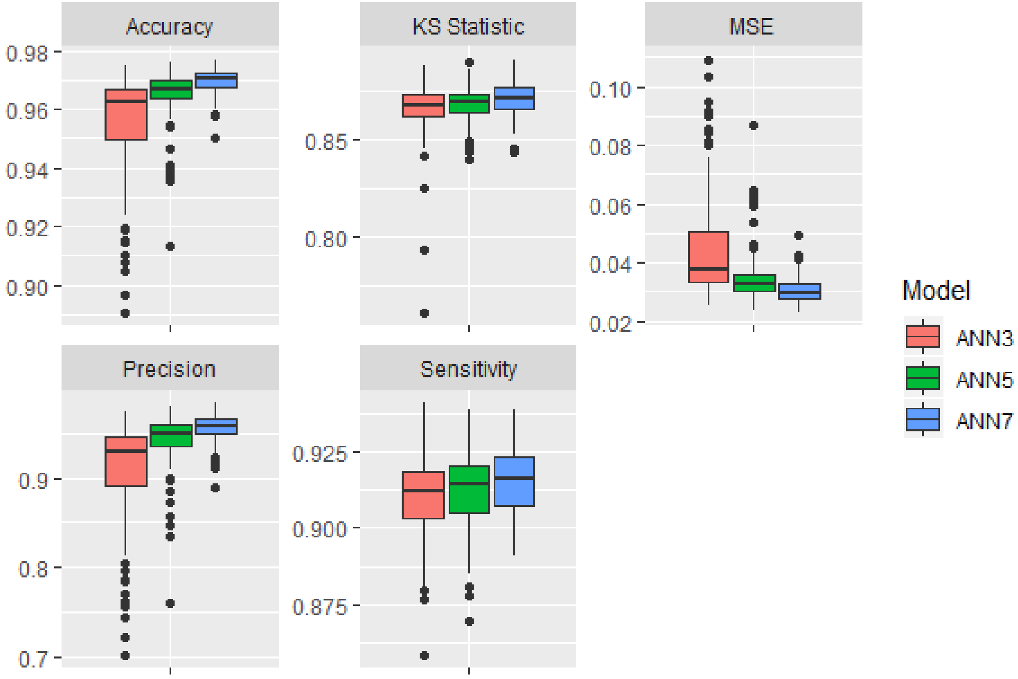

Figure 3. Boxplot Comparison of ANN Models

The t-test comparisons to the results show that the 3 ANN models are statistically significant. Therefore, it is obvious that ANN7 is the best ANN model since it has the smallest MSE value and the biggest Accuracy, Precision, Sensitivity, and KS Statistic values than ANN3 and ANN5.

4.2. Comparison of LR and ANN7 Models

Knowing that ANN7 is the best ANN model, then it will be compared to LR to determine the best model between them. Using the same process, the following results are obtained from the simulation.

Table 6. Comparison of LR and ANN Model

Model	Accuracy	Precision	Sensitivity	KS Stat	MSE
LR					
Mean	0.76860	0.51297	0.72092	0.61721	1.03436
SD	0.01951	0.02950	0.04106	0.03156	0.05574
ANN7					
Mean	0.96959	0.95513	0.91553	0.87031	0.03041
SD	0.00403	0.01485	0.01046	0.00829	0.00403

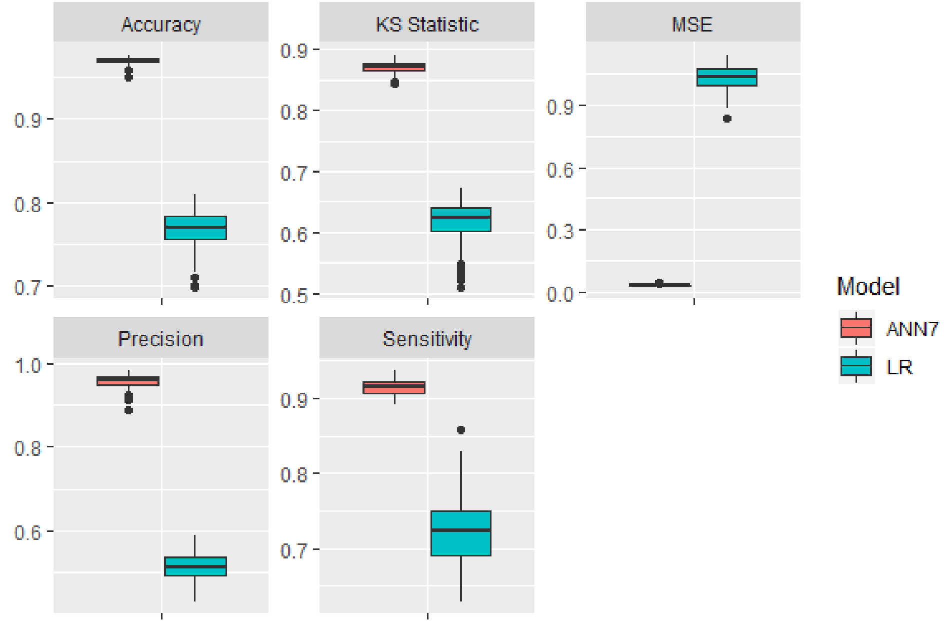

Figure 4. Boxplot Comparison of LR and ANN Models

Based on the simulation results above, ANN7 is obviously better than LR by significant margins. The simulation seems to show that the HR data set is best modelled by ANN of any kind.

5. CONCLUSION

The results from the simulation show that ANN7 is the best ANN model. It seems that the higher the number of neurons, the better the ANN model. The overall performance of LR and ANN show that the ANN model fits the data better than the LR model and therefore yield significantly better prediction results.

REFERENCES

[1] Baron and Byrne, "Social Psychology: Understanding Human Interaction", USA, (1994). Heights Allyn & Bacon Inc.

[2] Darpa, "DARPA Neural Network Study Final Report, Lincoln Laboratory, Massachusetts Institute of Technology",(1989).

[3] Fausett, Laurene, "Fundamental of Neural Networks: Architectures Algorithms and Application, Prentice Hall Inc",(1994).

[4] Galton, Francis, "Family Likeness in Stature", Proceedings of Royal Society, London, **40** (1886), pp. 42-72.

[5] Gujarati, D and Porter, D.C, "Basic Econometrics", McGraw-Hill, New York, (2009),Fifth Edition, pp. 527.

[6] Jain, Anil K, “Artificial Neural Network: A Tutorial Michigan State University, Proceeding of 1996 IEEE”, (1996), 31-44.

[7] Jong, P.D and Heller, G.Z, “Generalized Linear Model for Insurance Data”, Cambridge University Press, New York, (2008), pp. 35-36.

[8] https://www.kaggle.com/playtimez/predict-employee-leave/data

[9] Elegido, J. M., “Does it make sense to be a loyal employee?” Journal of Business Ethics, **116(3)** (2013), 495–511.

[10] Seema Mehta,Tarika Singh, S.S. Bhakar, Brajesh Sinha, “Employee Loyalty towards Organization–A study of Academician”, International Journal of Bussiness and Management and Economy, **1** (2010), ISSN: 2229-6247.

[11] Dutot, C., “Contribution to representations of personal loyalty to the company: convergence and divergence elements between workers and employers. The case of the workers of two metallurgical industries of the Country of Retz”, Institute of Business Administration (Poitiers), (2004), http://www.theses.fr/en/2004POIT4003.

[12] Davis, K.Y. and Newstrom, J.W., “Comportamien to Humano en al Trabjo: Comportamien to Organizational”, McGraw-Hill, Mexico City, 10th ed, (1999)

[13] Abugre, J.B., “Perceived Satisfaction in Sustained Outcomes of Employee Communication in Ghanaian Organizations”, Journal of Management Policy and Practice, **12** (7) (2011), 37-49.

[14] Yigit, Ibrahim Onuralp and Shourabizadeh, Hamed, “An Approach for Predicting Employee Churn by Using Data Mining”, Conference: International Artificial Intelligence and Data Processing Symposium’17, Turkey: Malatya, (2017)

[15] Heaton, Jeff, “Introduction to Neural Networks for Java”, Heaton Research, Inc, (2008), 2nd Edition