10.31341/jios.45.2.1
Open Access
UDC 004.032.26:005.96
Original Scientific Paper

A Novel Classification Model for Employees Turnover Using Neural Network to Enhance Job Satisfaction in Organizations

Tarig Mohamed Ahmed *Tmahmad@Kau.edu.sa*
1: Prof - Department of IT,
King Abdul- Aziz University, KSA.
2: Department of Computer Sciences,
University of Khartoum, Khartoum, Sudan

Abstract

The most important challenge facing modern organizations is to keep their employees as valuable assists. Employee turnover is one of these challenges. This paper aims to develop a novel model that can help decision-makers to classify the problem of Employee Turnover. The proposed model is based on machine learning algorithms. The model was trained and tested by using a dataset that consists of 1470 records and 25 features. To develop the research model, many experiments had been conducted to find the best one. Based on implementation results, the Neural Network algorithm is selected as the best one with an Accuracy of 84% and AUC (ROC) of 74%. By validation mechanism, the model is acceptable and reliable to help origination decision-makers to manage their employees in a good manner and setting proactive plans to keep them. Besides the model, three important features should be dealt with carefully as Over Time, Job Level, Monthly Income.

Keywords: Employee Turnover, Job Satisfaction, Machine Learning, Classification.

1. Introduction

Employee turnover is one of the most important challenges facing modern organizations. Employee attrition in terms of deterioration influences the efficiency of the organization [1]. The market world is currently experiencing extreme competition leading to the failure of organizations and businesses that cannot invest optimally in human capital, as the position of these services is based on the new administrative sciences [2]. Employee turnover is defined by the willingness of the employee to quit his current job, whether by his will or the desire of the establishment. Also, it could be defined as the choice to be compensated by a person who belongs to an organization to quit his current work [3]. There are many causes for employee turnover, the most important of which are: lack of a consistent promotional career direction, lack of a sense of justice in measuring work results and lack of a sense of job security [4]. Job satisfaction is a psychological feeling of contentment, satisfaction, and happiness. It satisfies the needs, desires, and expectations with the work itself and the work environment, with confidence, loyalty, and belonging to

work and with the relevant internal and external environmental factors and influences [5]. In another definition, it indicates that job satisfaction is a trend that is the outcome of many beloved and unloved experiences Associated with work and reveal himself with the appreciation of the individual for work and management. Job satisfaction reduces employee turnover [6].

To reduce the ratio of employee turnover, data mining can play an important role by providing a model that works as a decision support system [7]. Data Mining is the process of discovering patterns from large data sets based on methods that intersect both machine learning, statistics, and database systems [8]. In the data mining process, we have the same concept. For data mining, you must first collect data from various sources, prepare it, and store it in one place, as nothing from data mining is related to the process of searching for the data itself [9]. The model in data mining is a computerized representation of real-world observation. Forms are algorithm applications to search, identify, and display any patterns or messages in your data. There are two types of models in data mining: taxonomic, descriptive, or predictive [10].

In this paper, a new model has been proposed. The model will be used as a decision support system that helps the organization managers to take the right decisions for keeping their employees as a valuable asset. The model was built based on classification algorithms. Classification algorithms are a form of data analysis that extracts models that accurately describe important data categories and classifications. For example, a classification form can be built to evaluate loan applications in a bank's loan granting system, dividing the group of applications into two categories, safe and unsafe [11]. By using the model, we can classify employees into two classes: attrition class or not.

The rest of the paper is organized as follows: The next section presents some interesting researches related to the research area as related work. In Section 3, a full description of the proposed model is provided in terms of Framework, Dataset, implementation, and the results. Section 4 concludes the paper and suggesting some points to improve model accuracy as future

2. Related work

There is no doubt that technological discoveries, especially in the field of artificial intelligence, have greatly assisted decision-makers in the decision-making process [12]. They provided them with the ability to analyze their decisions and compare them with similar decisions and show the results that may result from this or that decision. Also, these systems provide advice to administrators through an analysis of Information stored on private databases in decision-making support systems [13]. A decision support system provides organizations with valuable information to achieve their strategic objectives [14].

Several previous studies have shown that turnover is the consequence of work dissatisfaction a combination of factors which include pay, recognition, and career development opportunities.... etc. In one of the previous studies, data were collected from primary and secondary sources, they used the questionnaire to analyze factors

affecting employee turnover and found there were three main factors that influence the decision to resign: internal factors, personal and external [19].

Due to stressful work environment retention of employees in IT sector is low. These stressful situations stimulate employees to keep changing their jobs. Job satisfaction is directly associated with the continuance of the employment specifically in IT sector Jan et al., (2016). It is cumbersome process to replace tech workers as it has negative effects on customer services and information security. In compare to other professions IT professionals receives more job solicitation per week.

A previous study conducted several in-depth interviews and surveys with leaders and managers who lose their employees in the department and proved that there is no unique factor that leads to turnover but is a set of factors that influenced the employee and led to the decision to resign. The salary is an important factor that causes turnover; other factors found are recognition, identification of responsibilities, and career opportunities. This set of factors indicates that the employee has found other job offers or dissatisfaction resulting in his resignation, which means that employees are going through a series of stages before deciding to change the jobs [20, 29 , 31].

Data Mining Model is one of the important models that help the decision-makers by discovering knowledge behind data. Knowledge has different forms such as predictive or descriptive models. This paper focuses on predictive models. Many types of research have been conducted in this area. This section presents some of them.

IBM Team M. Watson Singh et al. conducted brilliant research on the turnover cycle. The research proposed a system that defines the causes of the turnover and predicts future employee turnover. They also attempted to quantify the cost of attrition and recommend the name of the employee for the retention process, comparing the discrepancy between the Expected cost of attrition before retention period (EACB) and the Expected cost of attrition after retention period (EACA) [15].

L. Carlos et al. put forward a proposal for a Chronic problem with college forecasts for Brazilian Universities Heartlessness. H2O software was used as a data extraction tool and deep learning and attempted to predict the dropout cases Identify and initiate student attrition profiles [16]. K. Dejaeger et.al suggested a centric production on income by Maximum benefit estimation using a maximal fraction of the Best forecasted consumer loss rates in retention [17].

For another research, Choudhary et.al outline the implementation of the methodology of logistics regression based on the demographic data for independent personnel as well as current staff Create a probability algorithm to estimate the turnover of the workforce. This will come later Equation used to compare the probability of turnover with the present one Employees set. The high-risk cluster was established after the assessment to figure out the reasons for this, and the action plan was named to reduce the risk [18].

Also, research studies predicted a voluntary turnover rate by applying cluster analysis which considered an unsupervised learning method. It was found the results indicate that the high turnover trend circle was primarily caused by a lack of inner fidelity identification, leadership, and management [14]. The Predictive Model can be created by various methods of data mining.

Data mining defined as "the process of sorting through large data sets to identify patterns and establish relationships to solve problems through data analysis. Data mining tools allow enterprises to predict future trends" [15] Data mining Techniques can be useful for Human Resource Managers in identifying factors influencing employees to predict turnover.

There are many data mining algorithms for classification such as Decision Tree, SVM, Random Frost, Neural Network, and Naïve Bayes. In this paper, a brief description is provided for Neural Network and Naïve Bayes Algorithms.

Neural networks are a series of algorithms that are closely modeled after the human brain and programmed to identify patterns. They view sensory data by means of a form of raw input system interpretation, marking, or clustering [21] [30]. The trends they know are empirical, stored in vectors that need to be converted into all real-world data, be it images, voice, text, or time series. Neural networks allow one to build and identify clusters. On top of the data you store and handle, you might think of them as a layer of clustering and grouping. They help group unlabeled data according to differences between example inputs and identify data when they have to work on a labeled dataset. (Neural networks may also derive characteristics that are fed into other clustering algorithms and Classification; and you can think of deep neural networks as components of broader machine-learning systems including instruction, classification, and regression reinforcement algorithms [22]. Figure 1 presents Neural Network Algorithm.

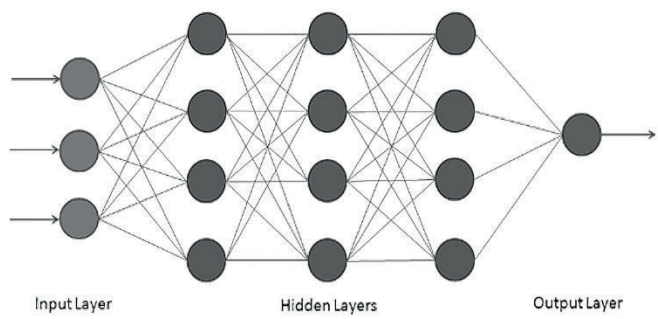

Figure 1. Neutral Network Algorithm

Naive Bayes algorithm is a technique of classification based on the theorem of Bayes with an assumption of freedom of predictors [24]. A Naive Bayes classifier believes, in basic words, that the inclusion of a specific function in a class is irrelevant to some other function. Naive Bayes model is simple to construct and particularly useful for incredibly large data sets. Naive Bayes is considered to outperform even extremely advanced methods of classification, in addition to simplicity. Bayes Theorem provides a way for P(c), P(x), and P(x) to measure posterior likelihood. Look at the equation underneath Fig. 2:

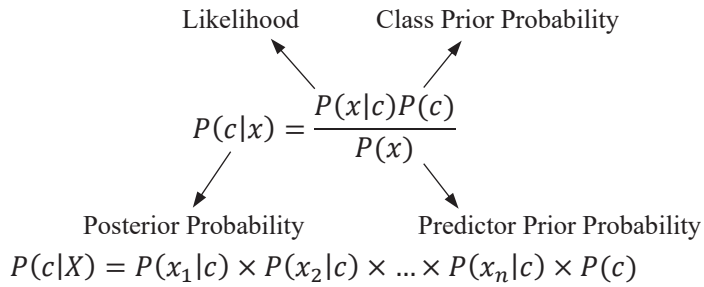

Figure 2. Naïve Bayes Algorithm

3. Method and Techniques

The purpose of this research is to save organization costs by use information available in human resource management to predict employee turnover and also, to improve retention of valuable employees. In this way, the internal resources of the organization will be maintained as much as possible. The experimental method was used to develop an accurate machine learning model. Based on this, machine learning classification algorithms were used for multiple experiments to find the more suitable algorithm for the proposed model. This research answers two questions: First, how can an organization detect the expected turnover of employees. Second, what are the features that should be focused on by the organization for the detected group? Fig 3 shows the research model components terms the methodology that was used to develop the proposed model.

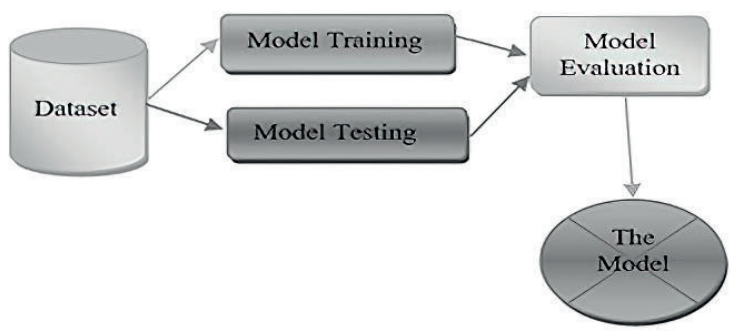

Figure 3. Model Components

The dataset contains HR employee records. The model is trained and evaluated by means of the dataset and classification algorithms. Based on the test results, the best algorithm is selected to be used to identify knowledge about the new staff. The model should help decision-makers make the best decision to keep their employees.

3.1 HR Employee Dataset

The dataset includes a fictional data set created by IBM data scientists. The dataset has 1470 observations that refer to employees with 25 attributes. This data set is obtained from IBM Data Science Community [25].

The attrition attribute has been selected to be used for the classification model. It is very important to understand the dataset before conducting the analysis. One of the options is feature selection. The selection of features can be defined as a process that selects a minimum subset of M features from the initial set of N features so that the feature space is minimized optimally according to a certain evaluation criterion. The amount of function N decreases as the dimensional of a domain extends. It is typically intractable to find the right subset of functions [23]. The table. 1 gives an overview of the dataset variables.

Variables	Category	Frequency(Percentages)
Environment Satisfaction	Low	284(19.3%)
	Medium	287(19.5%)
	High	453(30.8%)
	Very High	446(30.3%)
Job Involvement	Low	83(5.6%)
	Medium	375(25.5%)
	High	868(59%)
	Very High	144(9.8%)
Job Satisfaction	Low	289(19.7%)
	Medium	280(19%)
	High	442(30.1%)
	Very High	459(31.2%)
Performance Rating	Excellent	1244(84.6%)
	Outstanding	226(15.4%)
Relationship Satisfaction	Low	276(18.8%)
	Medium	303(20.6%)
	High	459(31.2%)
	Very High	432(29.4%)
Work Life Balance	Bad	80(5.4%)
	Good	344(23.4%)
	Better	893(60.7%)
	Best	153(10.4%)
Education	Below College	170(11.6%)
	College	282(19.2%)
	Bachelor	572(38.9%)
	Master	398(27.1%)
	Doctor	48(3.3%)

Table1. Descriptive statistics for variables

For ranking the important attributes, two methods were used: Information Gain and Chi-Square. Information gain is a synonym for Kullback – Leibler divergence in information theory and machine learning; the sum of knowledge obtained regarding a random variable or signal from another random variable being measured [26]. However, the term is often used synonymously with reciprocal knowledge in the sense of decision trees, which is the conditional predicted value of the Kullback – Leibler variance of the univariate distribution of the probability of one variable from the conditional distribution of this variable to the other. The information gain is obtained by the following equation:

$$\mathrm{IG}(T, a) = \mathrm{H}(T) - \mathrm{H}(T|a),$$

where $\mathrm{H}(T|a)$ is the contradictional entropy of T given the value of attribute a.

For machine learning, the Chi-Square test is important, and how this test makes a difference. The selection of features is a significant problem in machine learning, where we will have many features in line and have to select the best features to create the model [27]. Through checking the relationship between the features, the chi-square test lets you solve the question of feature selection. The chi-Square is obtained by the following formula:

$$x_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where:

$c =$ Defrees of freedom

$O =$ Observed value(s)

$E =$ Expected value(s)

The two methods are agreed on four attributes Over Time, Job Level, Monthly Income and Years at Company. These attributes have a strong effect on building the model. Table 2 shows the ranking attribute by using two methods: Information Gain and Chi-Squair.

Features	**Rank(Information Gain)**	**Rank(Chi-Squair)**
Over Time	0.04	63.845
Job Level	0.036	44.669
Monthly Income	0.03	41.795
Years at Company	0.028	37.09

Table 2. Ranking Features

Based on the results in table 2, these four features are reflecting very important indicators for keeping employees satisfied with their jobs and they play a significant role to reduce employee turnovers.

3.2 Proposed Model development

To develop the research model, many experiments have been conducted using five classification algorithms in order to reach the best one. These algorithms are as follows: Decision Tree, SVM, Random Frost, Neural Network, and Naïve Bayes. The Cross-Validation 10 method was used for testing the models.

To evaluate the model, a confusion matrix was used. The confusion matrix, also known as the error matrix, is the problem of mathematical classification in the field of machine learning [28]. The uncertainty matrix is a table frequently used to describe a classification model's output (or "workbook") on a collection of test data recognized for the real values. It allows the algorithm's output to be visualized. different parameters were used as mentioned in fig. 5:

		Prediction Class: Positive	Prediction Class: Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error	Sensitivity $\frac{TP}{(TP+FN)}$
	Negative	False Positive (FP) Type I Error	True Negative (TN)	Specificity $\frac{TN}{(TN+FP)}$
		Prediction $\frac{TP}{(TP+FP)}$	Negative Predictive Value $\frac{TN}{(TN+FN)}$	Accuracy $\frac{TP+TN}{(TP+TN+FP+FN)}$

Figure 5. Confusion matrix attributes

Many experiments have been conducted to build the model in three phases: phase one the model was trained by using classification algorithms (Decision Tree, SVM, Random Frost, Neural Network, and Naïve Bayes). In phase two, the model has been tested by using Cross-Validation 10. Finally, the model has been evaluated based on the confusion matrix to select the best algorithm. Table 3. Shows the implementation results.

Model	**AUC**	**CA**	**F1**	**Precision**	**Recall**
Tree	0.577	0.791	0.78	0.771	0.791
SVM	0.475	0.727	0.718	0.711	0.727
Random Forest	0.704	0.836	0.819	0.812	0.836
Neural Network	**0.755**	**0.85**	**0.819**	**0.823**	**0.85**
Naive Bayes	0.7416	0.8116	0.8114	0.8112	0.8116

Table 3. Confusion Matrix Results

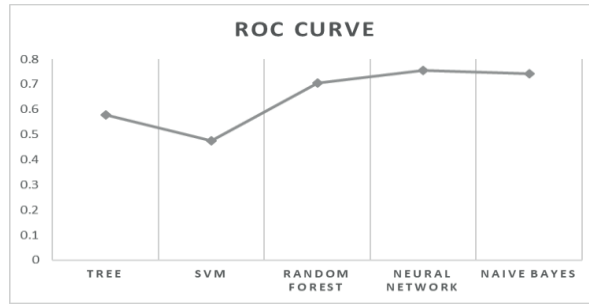

Figure 6. Model ROC curve

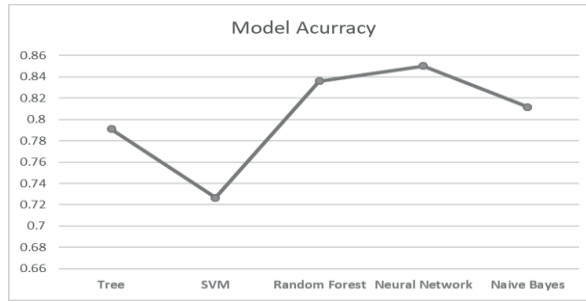

Figure 7. Model Accuracy

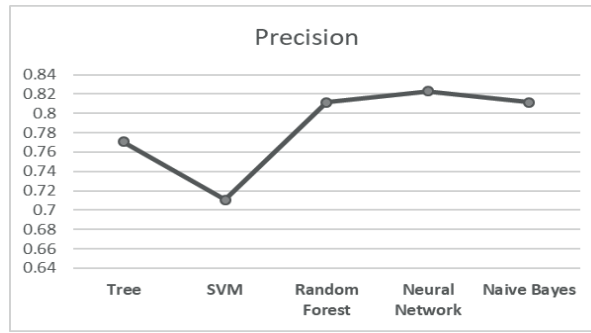

Figure 8. Model Precision

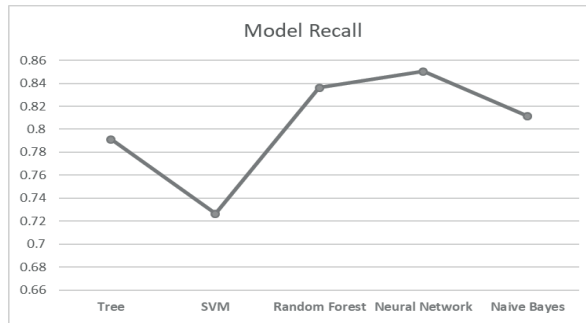

Figure 9. Model Recall

3.3 Results Discussion

As mentioned in table 2 The accuracy of five algorithms: Tree, Random Frost, Neural Network, VSM and Naïve Bayes were above 80%. The accuracy is obtained by the equation mentioned in fig. 5. But, the accuracy is not a best factor to select the best algorithm and that because the tendency of the classification attribute "Attrition" is very high as shown in figure 10.

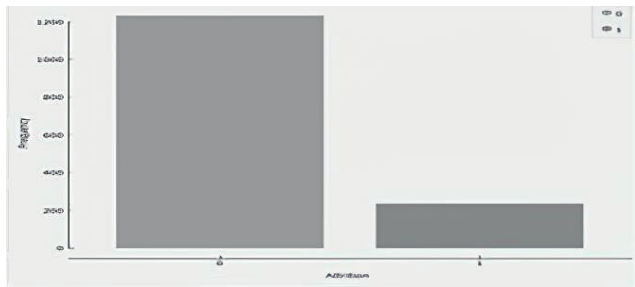

Figure 10. Classification Attribute: Attrition

In this case, the model depends on t Area Under Carve(AUC) Parameter and also called Receiver Operating Characteristics(ROC). AUC means "Space under the ROC Curve," i.e. AUC measures the whole two-dimensional space under the whole ROC curve (think integral calculus) from (0,0) to (1,1). Fig. 11 presents AUC for the five proposed algorithms [29].

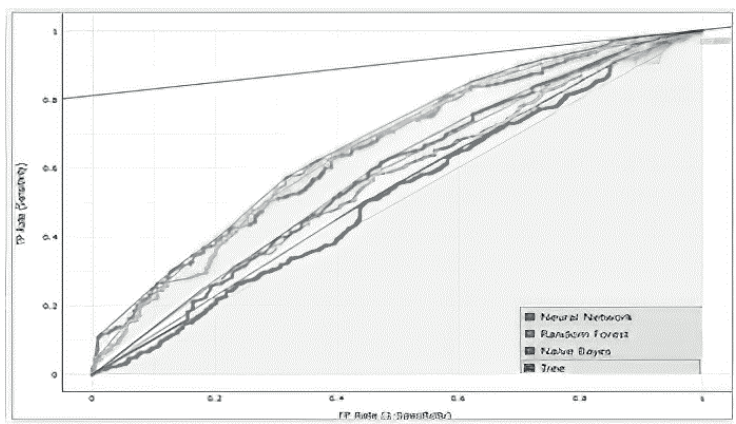

Figure 11. presents AUC for the five proposed algorithms

To select the best algorithms among the five ones, it is based on the AUC rate. The acceptable rate between 0.5 to 1. The better ones, the near to 1. Based on this fact, the neural network algorithm is the best one to be used by the model to classify the employees. In this way, the top management of the organizations can prepare plans to keep their employees as valuable assets. Figure 12. present AUC for the Neural Network.

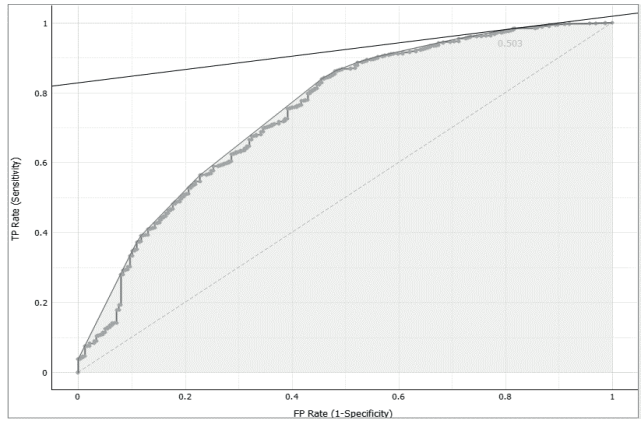

Figure 12. AUC for the Neural Network

4. Conclusion and future work

In this paper, a novel data mining model has been developed. By using feature selection methods: Information Gain and Chi-Square, the most important four features have been extracted from the dataset. These features are over time, job level, salary, and years in the organization. As one of the important results of this research, these features should be planned carefully to keep organizations their employees as valuable assets. The proposed model is based on machine learning algorithms. Classification algorithms were used to implement the model such as Decision Tree, SVM, Random Frost, Neuronal Network, and Naïve Bayes. The model was trained and tested by using a dataset that consists of 1470 records and 25 features. To develop the research model, many experiments had been conducted to find the best one. Based on implementation results, the Neural Network algorithm is selected as the best one with an Accuracy of 84% and AUC (ROC) of 74%. By validation mechanism, the model is acceptable and reliable to help decision-makers to manage their employees in a good manner. The could be used to classify current employees if they will be turned over or not. This point answers the first research question. Figure 13 presents the model with results. This model could be used by HR departments to predict employees into two groups: expected turnover employees or not. For the first group, they may set plans to improve the features: Over Time, Job Level, Monthly Income. This point answers the second research question.

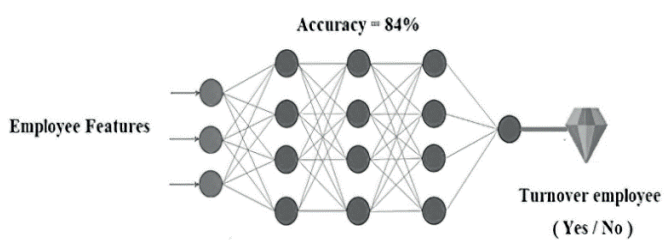

Figure 13. The Proposed model

As future work of this research, the dataset needs to cover the social information about the employees and not only information about the organization. In this way, the efficiency of the model will be improved.

References

[1] Mobley, William H., et al. "Review and conceptual analysis of the employee turnover process." sychological bulletin 86.3 (1979): 493.

[2] Ganotakis, Panagiotis. "Founders' human capital and the performance of UK new technology based firms." Small Business Economics 39.2 (2012): 495-515.

[3] Felps, Will, et al. "Turnover contagion: How coworkers' job embeddedness and job search behaviors influence quitting." Academy of Management Journal 52.3 (2009): 545-561.

[4] Kowske, Brenda J., Rena Rasch, and Jack Wiley. "Millennials'(lack of) attitude problem: An empirical examination of generational effects on work attitudes." Journal of Business and Psychology 25.2 (2010): 265-279.

[5] Božović, Jelena, Ivan Božović, and Isidora Ljumović. "Impact of HRM practices on job satisfaction of employees in Serbian banking sector." Management: Journal of Sustainable Business and Management Solutions in Emerging Economies 24.1 (2019): 63-77.

[6] Andreyko, Tammy A. Principal leadership in the accountability era: Influence of expanding job responsibilities on functional work performance, stress management, and overall job satisfaction. University of Pittsburgh, 2010.

[7] Sauter, Vicki L. Decision support systems for business intelligence. John Wiley & Sons, 2014

[8] Chaudhuri, Surajit. "Data mining and database systems: Where is the intersection?." IEEE Data Eng. Bull. 21.1 (1998): 4-8.

[9] Provost, Foster, and Tom Fawcett. Data Science for Business: What you need to know about data mining and data-analytic thinking. " O'Reilly Media, Inc.", 2013.

[10] Han, Jiawei, Jian Pei, and Micheline Kamber. Data mining: concepts and techniques. Elsevier, 2011.

[11] Hamid, A. Jafar, and Tarig Mohammed Ahmed. "Developing prediction model of loan risk in banks using data mining." Machine Learning and Applications: An International Journal (MLAIJ) 3.1 (2016).

[12] Agrawal, Ajay, Joshua Gans, and Avi Goldfarb. Prediction machines: the simple economics of artificial intelligence. Harvard Business Press, 2018.

[13] Dean Jr, James W., and Mark P. Sharfman. "Does decision process matter? A study of strategic decision-making effectiveness." Academy of management journal 39.2 (1996): 368-392.

[14] Sauter, Vicki L. Decision support systems for business intelligence. John Wiley & Sons, 2014.

[15] M. Singh et al., "An Analytics Approach for Proactively Combating Voluntary Attrition of Employees," 2012 IEEE 12th International Conference on Data Mining Workshops, pp. 317-323, Brussels, 2012.

[16] L. C. B. Martins, R. N. Carvalho, R. S. Carvalho, M. C. Victorino and M. Holanda, "Early Prediction of College Attrition Using Data Mining," 16th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 1075-1078, Cancun, 2017.

[17] K. Dejaeger, W. Verbeke, D. Martens, J. Hur, and B. Baesens, "New insights into churn prediction in the telecommunication sector: A profit driven data mining approach," European Journal of Operational Research, vol. 218, no. 1, pp. 211–229, 2012.

[18] C. K. Choudhary, R. Khare, D. Kaloya, and G. Gupta, "Employee attrition risk assessment using logistic regression analysis" 2nd IIMA International Conference on Advanced Data Analysis, Business Analytics and Intelligence ,Ahmedabad, 2011.

[19] Al-Habil, W. I., A. Allah, and M. Shehadah. "Factors Affecting the Employees' Turnover at the Ministry of High Education in Gaza Governorates-Case study: North and West Gaza Directorates of Education." Arts Social Sci J 8.304 (2017): 2.

[20] Lee Liu, Jaime. "Main causes of voluntary employee turnover a study of factors and their relationship with expectations and preferences." (2014).

[21] Rogers, Steven K., et al. "Neural networks for automatic target recognition." Neural networks 8.7-8 (1995): 1153-1184.

[22] Arulkumaran, Kai, et al. "A brief survey of deep reinforcement learning." arXiv preprint arXiv:1708.05866 (2017).

[23] Kohavi, Ron, and George H. John. "Wrappers for feature subset selection." Artificial intelligence 97.1-2 (1997): 273-324.

[24] Tang, Bo, Steven Kay, and Haibo He. "Toward optimal feature selection in naive Bayes for text categorization." IEEE transactions on knowledge and data engineering 28.9 (2016): 2508-2521.

[25] IBM Watson Analytics. "Sample Data: HR Employee Attrition and Performance." In: (September 2015). url: https://www.ibm.com/communities/analytics/watson-analytics-blog/hr-employee-attrition/.McKinley Stacker, I.V.: IBM waston analytics. Sample

data: HR employee attrition and performance [Data file]. Retrieved from https://www.ibm.com/communities/analytics/watson-analytics-blog/hr-employee-attrition/ (2015)

[26] Onchiri, Sureiman. "Conceptual model on application of chi-square test in education and social sciences." Global Journal of Art and Social Science Education 1.1 (2013): 16-26.

[27] Tharwat, Alaa. "Classification assessment methods." Applied Computing and Informatics (2018).

[28] Ilies, Remus, Kelly Schwind Wilson, and David T. Wagner. "The spillover of daily job satisfaction onto employees' family lives: The facilitating role of work-family integration." Academy of Management Journal 52.1 (2009): 87-102.

[29] Crespi Vallbona, Montserrat, and Oscar Mascarilla i Miró. "Job satisfaction. The case of information technology (IT) professionals in Spain." Universia Business Review, 2018, vol. 58, num. 2, p. 36-51 (2018).

[30] Jan, N. Akbar, A. Nirmal Raj, and A. K. Subramani. "Employees' Job Satisfaction in Information Technology Organizations in Chennai City-An Empirical Study." Asian Journal of Research in Social Sciences and Humanities 6.4 (2016): 602-614.