Eurasian Academy of Sciences
Eurasian Business & Economics Journal
2021 Volume: 27 S: 41-51
Published Online October 2021 (http://busecon.eurasianacademy.org)
http://dx.doi.org/10.17740/eas.econ.2021.V27-04

EMPLOYEE CHURN ESTIMATION USING MACHINE LEARNING METHODS

Murat Taha BİLİŞİK* **Pınar SARP****

*Doç. Dr., İstanbul Kültür Üniversitesi, İşletme Bölümü, m.bilisik@iku.edu.tr
**Arş. Gör., İstanbul Kültür Üniversitesi, İşletme Bölümü, p.sarp@iku.edu.tr

Received Date:19.08.2021 Revised Date:06.09.2021 Accepted Date:02.10.2021

ABSTRACT

Employee churn is a vital concern for organizations. Many machine learning (ML) based systems have been improved to bring solution to the employee churn problem. The present study tries to predict employee churn using ML algorithms. From this point on, the aim of the study is to predict employee churn for organizations based on mathematical methods using machine learning. Application of information technologies in Human Resources Management in organizations, the data collected is considered as a critical factor in making decisions about employees. The sample of the study consists of 49,653 employees from different sectors. There are almost no studies using machine learning algorithms in management studies. In this sense, it is predicted that the study will fill an important gap in terms of contributing to the literature. In this paper, logistic regression, random forests, artificial neural networks and support vector machine methods were used as classification methods in the research to estimate whether an employee will quit or not. The results are compared with other Machine Learning algorithms. In the analysis, it was seen that the random forests method was found to be the most reliable machine learning with a rate of 98.7715 % in estimating the employee quitting.

Keywords: Machine Learning, Employee Churn, Estimation Methods, Big Data, Benchmarking
JEL-Classification: C13, C45, C51

1. INTRODUCTION

The rise of Machine Learning and algoritms are changing the way the world does business. Employees are valuable assets for organizations. Machine Learning on workforce behavior and attitudes can make managers to forecast behavior, identify valuable talent like never before, match skills to market requirements, retain good employees, and act on proven insights to lead business results. The strength of these analytics lie in its ability to challenge traditional wisdom, influence behavior, enable human resources (HR) and business managers to make and execute smarter and more strategical workplace decisions, and accordingly affect business results.

However, some of employees can be high value creators and the best performers for organizations. Other employees may not exhibit the desired performance level in organizations. Utilizing employee loss analytics will help improve business value from employee attrition information by analyzing employee information and can help an organization address key attrition questions, including:

- Who are the best performers that are at high risk of quitting, and why?
- When are they more likely to leave?
- What kind of proactive decisions could be made to retain employees?
- What is the cost of losing best employees?

Employee retention is related with proactively anticipating and identifying which of your valuable employees are at high risk of leaving and when and why employee leave occurs.
There are some of global companies investing in machine learning. Examples include companies such as AOL, Google, Deloitte, and Pfizer. By harnessing the power of machine learning and forecasting analytics, these companies were able to gain a significant return on investment in the business by optimizing their skill retention activities (Isson, Hariot,2016).
However, when some employees leave an organization, its productivity, project continuity, and growth strategies are largely affected on which the image and turnover of the organization rely on. This also makes other employees leaving the organization (Rashid and Jabar 2016).

There are studies with customer churn prediction analysis. However, there exist almost no studies in the management literature which think about employee churn predict. By the way, with the application of information technologies in Human Resources Management in organizations, the data collected is considered as a critical factor in making decisions about employees. However, the applications on advanced data analytics in the field of Human Resources is still limited in small and medium organizations (Ekawati, 2019). However, Human Resources Analytics is a field of application that needs to be improved (Mishra vd., 2016). In the research conducted by IBM, 700 human resources managers were interviewed. According to the results, less than 25% of managers use advanced analytical methods to predict future results and make decisions (IBM, 2010).

According to a study conducted in 2016-2017, the use of Human Resources Analytics is less than 50%. The development level of HR leaders who apply HR Analytics in the areas of performance management (48%), recruiting and selection (47%), learning and development (44%), employee capacity planning (38%) are still at the basic-intermediate level (CI3PD, 2017).

The employee is still the most critical asset in today's knowledge-driven industry. In addition to, employee attrition can become a critical problem due to its effects on organizations' competitive advantage. However, quitting of an employee also poses a problem in terms of cost. Because, the cost of attrition of the employees will be the cost associated with the human resources life cycle, lost information and organizational culture. Since the loss of an employee can affect current projects or services, this is actually only a small fraction of an organization's total cost (Lee et al., 2017). Research on employee quit and / or intention to quit which is voluntary is generally based on survey data, but analyzing the real reason of employee quit and seeing if the employee leaves within a certain time frame requires detailed data. Therefore, in most cases, the intention to leave, not the actual loss of employee, is investigated.
From this point on, the aim of the study is to predict employee churn for organizations based on mathematical methods using machine learning.
Given that employee churn is closely related to the costs associated with losing employees, which are even higher than customer losses in some companies, the literature asks for more attention from researchers in this field (Yiğit, Shourabizadeh, 2017).

2.THEORETICAL FRAMEWORK

2.1.Employee Churn

It is a serious fact how important their employees are for businesses. Employees who have good qualifications and have adopted their jobs are very important for the companies they belong to to perform at a high level, to get ahead of their competitors, to achieve a service level that will

satisfy their customers, in short, in reaching the goals of the company. On the other hand, employees are an important cost factor for their firms. Employee recruitment, training, salary and other benefits, insurance and other legal payments constitute an important cost element for businesses. Loss of qualified manpower for the organization, hiring new employees, their training and adaptation to the working environment are some of these costs. Therefore, companies are able to ensure that their employees are engaged and expected. They should make the necessary effort to ensure that their employees who perform well do not leave their jobs. At this point, it is very important for employees to realize whether they have an intention to leave the job and to take measures to prevent leaving the job (Seyrek, İnal, 2014).

Employee churn is a big issue for the organizations especially when trained, technical and key employees quit for a better position in a competitor organization. It requires time, effort and occurs as financial loss to replace a well trained employee (Yiğit, Shourabizadeh, 2017). This paper, we use past employee data to predict current employee churn.
Employee churn is when individuals working in an organization leave their jobs for various reasons. Employment loss is a serious problem in all sectors, especially high-tech industries.

The cost of employee loss varies between 1.5 and 5 times the annual salary of the employee, depending on how difficult it is to fill the position of the employee (Sesil, 2014). Especially the loss of an experienced employee can result in the loss of the network or other important information. However, the fact that the employee leaving the organization has to meet the workload may also affect other employees. Another point to note is that it takes a certain of time for the new employee to reach a given wished level of expertise and productivity that the last employee had (Lee et al., 2017).

Preventing valuable employees from leaving their jobs remains a priority for Human Resources Department managers, and therefore it is important to understand the factors that lead to their departure (Lindsay et al., 2020). There are too many reasons why the employee leaves the job voluntarily. Positive causes include better offers (job, pay, bonuses, working conditions and facilities, career development, leadership, position, etc.). Negative causes are conflicts with managers or colleagues, perceived shortcomings (appreciation, education, career growth, focus / direction, etc.), underpayment, poor working conditions, etc.

While exit interviews give a good idea of the causes, there is still a need to verify the stated causes by other means, such as work history of employees. Predictive models are beneficial for understanding the root causes of employee loss, planning retention strategies, planning recruitment, and developing team management. Not all employees perform equally well. For example, employees who demonstrate excellence in certain tasks and have specialized technical skills may be more valuable. Therefore, forecasting models of employee loss that focus on accurately determining the loss of "valuable" employees are more useful (Saradhi, Palshikar, 2011).

2.2.Predictive Human Resources Analytics and Machine Learning

Machine learning is the general name of computer algorithms that models a problem according to the data of that problem. The model created with the existing data set and the algorithm used is established to give the highest performance. For this reason, many machine learning methods have been developed and some of them are; k-nearest neighbor algorithm, simple (naive) Bayesian classifier, decision trees, logistic regression analysis, k-means algorithm, support vector machines and artificial neural networks. Some of these approaches are capable of

estimation, some of them are capable of clustering and some of them are capable of classification.

Learning strategies in these methods; It is examined in three groups as supervised, unsupervised and reinforced (reinforced). With the model created in supervised learning, a group of input values In return, it is aimed to learn the relationship between them by giving their target values and to produce the outputs closest to the target values. The best model obtained also uses the closest output for the new input values.
In unsupervised learning, on the other hand, only the relationship between the input values is tried to be revealed without the target values. With the help of this relationship(s), values close to each other are grouped, that is, clustering is done. A new entry will belong to whichever of these clusters it is associated with. In the reinforcement learning method, instead of an advisor to give the target output, a criterion that evaluates the output as good or bad against the given input is used (Atalay, Çelik, 2017).

Predictive HR analytics is the systematic application of forecasting models using inferential statistics to existing HR or causal factors driving key HR-related performance indicators. Ability to access accurate information about current capabilities is critical for businesses. This analytics can help deliver the right workforce, the right recruitment, declining turnover, a solid skill line, a committed team, and the promise of achieving financial and strategical goals (Jain, Maitri, 2018).

Machine learning algorithms are growing rapidly and successfully implemented in various fields such as the recommendation system (Tarnowska et al. 2019). Estimation methods such as artificial intelligence and machine learning are widely used in areas such as mathematics, computers, health and etc.

Lots of machine learning (ML) based algorithms have been developed to bring solution to the employee churn problem. Machine learning includes techniques such as data / text mining, pattern matching, prediction, visualization, meaning analysis, sensitivity analysis, network and cluster analysis, multivariate statistics, graph analysis, simulation, complex event processing, neural networks (Dolatabadi, 2017). With these methods, accurate estimates can be made about the data collected about employees.

Machine learning is the broad name of computer algorithms which analyze a problem according to the data of that problem. The model created with the current data set and the algorithm used is set up to give the highest performance. For this reason, many machine learning methods have been developed, some of which are; k-nearest neighbor algorithm, simple (naive) Bayes classifier, decision trees, logistic regression analysis, k-means algorithm, support vector machines and artificial neural networks. Some of these approaches have the ability to predict and estimate, some to cluster and some to classify (Atalay, Çelik, 2017).

2.3.Logistic Regression

Logistic regression is a similar model to linear regression, but it works with the binomial response variable. In logistic regression, it is easier to use more than two explanatory variables at the same time. Although seemingly insignificant, this property is important in terms of the influence of numerous explanatory variables on the response variable. Logistic regression models the probability of an outcome based on individual preferences. One of the alternative methods developed for estimation that defines the relationship between dependent and

independent variables is Logistic Regression Analysis. In recent years, logistic regression analysis has come to the forefront with its ease of use and easy interpretation, and it has been widely used in many applications in the field of social sciences (Gök, Özdemir, 2011).

2.4.Random Forests

Random forests have emerged as serious competitors to cutting-edge methods such as retrofitting (Freund & Shapire, 1996) and supporting vector machines (Shawe-Taylor & Cristianini, 2004). It is a fast and easy to apply analysis. According to Breiman's approach, each tree in the collection is first randomly selected on each node. It is created by selecting a small group of input coordinates to be split and secondly by calculating the best division based on these features in the training set. The tree is grown to maximum size without pruning using the CART methodology (Breiman et al., 1984). This subspace randomization scheme is mixed with bagging for resampling, the training data set is changed each time a new tree is grown. Although the procedure seems easy, it contains many different factors that make it difficult to analyze (Biau, 2012). The graphical representation of random forests used as estimation method is as follows.

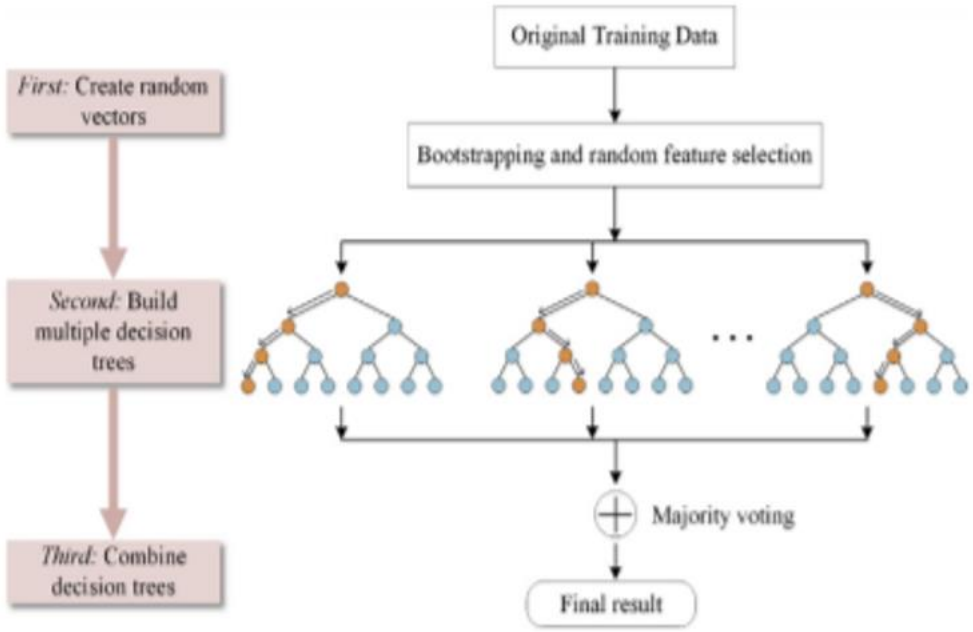

Figure 1. Illustration of the random forest method.

2.5.Artificial Neural Networks

Artificial Neural Network (ANN) is a machine learning algorithm that simulates the behavior of a human neuron. The human brain consists of millions of neurons linked together by a special structure known as a synapse. Synapses let neurons process signals from one to another. The Artificial Neural Network applies this behavior to a large number of interconnected processing units which work together to process information and produce meaningful outcomes. The process of training and adjusting the weight of these connections to achieve the required general behavior is known as the learning process. The ability to automatically learn from current data to generate forecasts is the extraordinary feature that makes this algorithm very interesting. It also has the ability to apply confidential insights into confidential relationships (Strohmeier & Piazza, 2013).

There are three kinds of layers in neural networks: input, hidden, and output layers. The general layout or architecture of a network is presented in Figure 2. Traditionally, neural networks had a very simple structure with only input and output layers, these are called single layer neural networks. Neural networks with multiple hidden layers are called multilayer neural networks or deep neural networks. Most of the modern neural networks used in practical applications are in general deep neural networks. Each entry node has a link to all nodes in the next hidden layer. Network training includes two different steps: feed forward propagation and back propagation. Training of a network begins with feed-forward propagation, where inputs and correct outputs from training data are introduced into the neural network. In the end, the neural network produces outputs based on inputs and a random results of weights. Next, the outputs from the neural network are compared with the actual outputs and the error is found out. Backpropagation is the process where the weights are updated according to the error computation in each node and adjust the weights accordingly to bring down the error (Kim 2017).

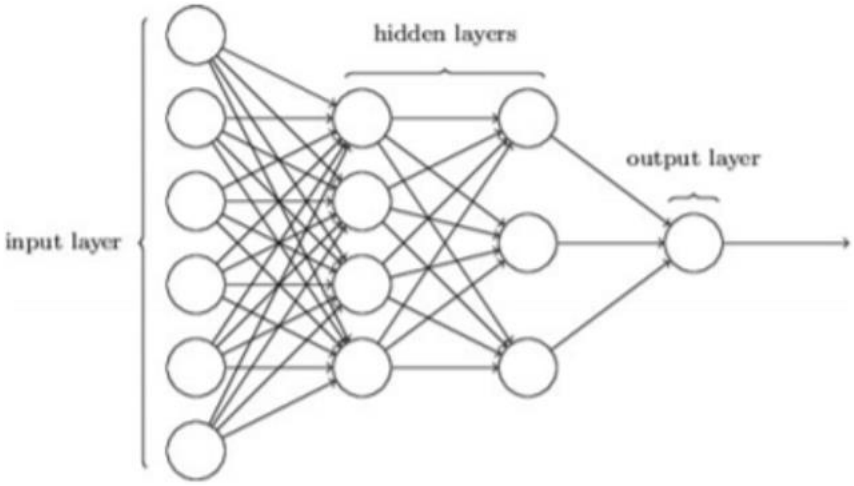

Figure 2. Structure of deep neural network or multilayer perceptron.

2.6.Support Vector Machine

Support vector machines are statistical and machine learning algorithms whose primary purpose is prediction. Gaussian can be applied for continuous, binary and categorical outputs similar to logistic and multinomial regression (Guenter, Schonlau, 2016). SVM can also make nonlinear classification. In this case, data points are mapped using a nonlinear function using kernel number in a highly dimensional space (Cortes et al., 1995). SVM can be run on classification and regression problems. The basic idea in the SVM regression method is to find the linear discriminant function that reflects the characteristics of the available training data as closely as possible and conforms to the statistical learning theory. Similar to classification, kernel functions are used in regression to handle nonlinear situations (Çomak, 2008)

3.METHODS

Churn analysis and forecasting is widely studied for customer churn, since it is a badly known issue and results in revenue loss. Employee churn is a huge concern for organization, however to forecast employee churn is rather more difficult than customer churn forecast. Employee churn includes issues such as efforts and time to get the replacement and retraining, financial loss, customer dissatisfaction and many more. Therefore, for smooth running of an organization, the key is to keep its valuable workforce (Yiğit, Shourabizadeh, 2017).

We used a wide range of data mining algorithms including logistic regression, random forests, artificial neural networks and support vector machine methods.
The first step in our approach is data selection. In this step, we utilize to predict employee attrition by using Human Resources Employee Attrition dataset provided by kaggle. The dataset includes employee information such as demographics, age, length of service, position etc. There were totally 8 attributes about employees. Table I shows the attributes of data and their type and definition.

No	Attribute	Data Type
1	Age	Numeric
2	Length of Service	Numeric
3	City Name	Categorical
4	Department Name	Categorical
5	Store Name	Categorical
6	Gender	Categorical
7	Status Year	Numeric
8	Business Unit	Categorical

Table 1. HR Dataset Features

3.1.Data

The employee attrition dataset on kaggle.com.tr has been selected for analysis. The knowledge of 49.653 employees was used to make predictions. In the dataset, there are 8 atttibutes (age, service period, city name, department name, company name, gender, status year and business unit) related to the employees. The dataset in csv format was analyzed in Weka 3.8.5 version.

3.2.Comparison of Classification Methods

In this part, we compare the classification methods. We want to find out how reliable a classification algorithm is by measuring accuracy, precision, and F measure in the test set. In order to make a well prediction n in the dataset, we used 80% as training set and 20% as test set. Table II demonstrates detailed information about the distribution of the datasets after spliting into two parts.

Dataset	Percentage	Number of Employees
Train	%80	39.722
Test	%20	9.931

Table II. Train - Test Datasets

We start with simple Logistic Regression classification methods for employee churn prediction. Then, we try to more complex methods as Support Vector Machines (SVM), Random Forests and Artificial Neural Networks.

Figure 3. demonstrates the comparison of classification methods in terms of accuracy metric on the test dataset.

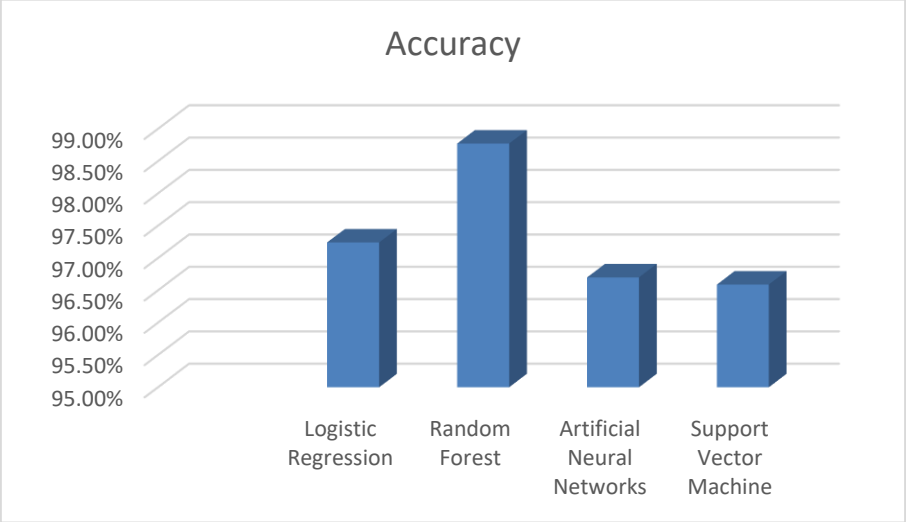

Figure 3. Comparison of Classification Methods According to Accuracy

Figure 4. shows the comparison of classification methods in accordance with precision metric on the test dataset.

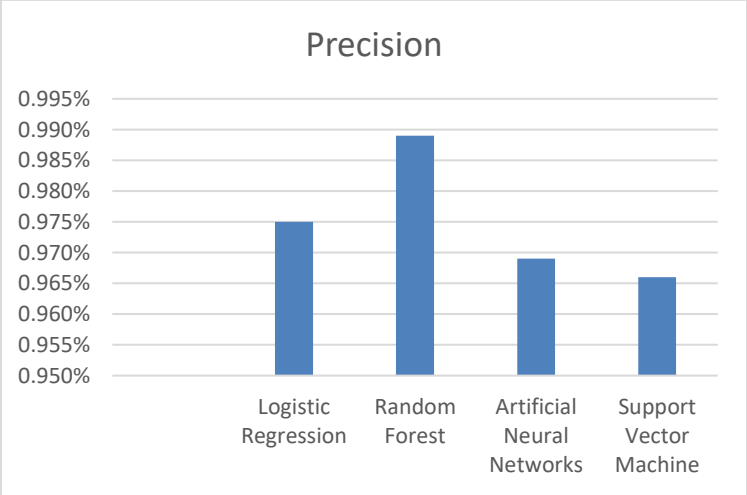

Figure 4. Comparison of Classification Methods According to Precision

The summary of the analysis is given in table 3.

Method	Accuracy	Precision	**F-Measure**
Logistic Regression	%97,24	0,975	**0,986**
Random Forecast	%98,77	0,989	**0,994**
Artificial neural networks	%96,70	0,969	**0,983**
Support Vector Machine	%96,59	0,966	**0,983**

Table III. The Summary Of Analysis

In the analysis, it was seen that the random forests method was found to be the most reliable machine learning with a rate of 98.77 % in estimating the employee churn.

4.PRACTICAL IMPLICATIONS

Many companies just concentrate on their employee turnover rates, but that can leave out essential information. You need to know if your company is losing employees, why you are losing them, and what you can do to keep them. Artificial intelligence and machine learning algorithms play a very important role for organizations. Many organizations make significant investments in data science teams to take advantage of all the value AI has to offer. Our research provides practical recommendations. The concept of employee is extremely important in small and medium-sized businesses. The employee churn of the invested employee means cost and time loss for the organization. Losses can be prevented, especially by using the data collected from the employees in the analysis. A number of strategies can be developed to retain employees, especially those who intend to quit. Such a machine learning model could be used to design employee retention plans targeted towards retaining valuable employees for Human Resources Department. This knowledge would aid HR managers in taking preventive action on time. By learning the reasons of employee quitting at your organization and having an overview of your attrition rates, managers can employ long-term workforce planning strategies to manage attrition so that it doesn't hurt businesses.

5.LIMITATIONS AND FUTURE RESEARCH

Although our study provides important insights, some limitations could be noted. The sample limits the generalizability of the study, which was conducted in a specific national context, Canada firms in general and the Vancouver district in particular. It is important to say that readers should be cautious when generalizing the results found to different cultural contexts.

In this research, employee churn was tried to be predicted by using data sets. Researchers who want to work in this field can make estimations using data sets on management issues about performance evaluation, training and development.

In addition, researchers who want to work on this subject in the future can make case studies on how the variables discussed affect their turnover. In addition, future researchers can try to make predictions by finding demographic information of employees in different sectors. It can be examined the employee churn in different sectoral context.

6.CONCLUSION

Employee churn is a great cost for businesses. Employee Churn is similarly painful for an organization because of lost time and effort in replacement finding and training. Therefore, the aim of the study is to predict employee churn with the highest reliability. The major contribution of this research is the demonstration that machine learning techniques could be used to build reliable and accurate predictive models for employee churn. Because management studies, make surveys as a data collection method. Since the results are subjective and relative, it is not objective in that the results may vary depending on the researcher conducting the research. However, in this study, objectivity has been tried to be obtained using mathematical algorithms. Thus, it can be concluded that by thinking of the important features of each respective employee and classifying as quitting candiadate or not, it can help the organizations to make a more productive, cost-effective, and timely retention policies to stop employees from quitting their jobs.

REFERENCES

- Atalay M and E Çelik (2017) Büyük Veri Analizinde Yapay Zekâ ve Makine Öğrenmesi Uygulamaları. Mehmet Akif Üniversitesi Sosyal Bilimler Enstitüsü Dergisi 9(22), 1309-1387.
- Atkinson AB (1970) On the measurement of inequality. Journal of Economic Theory 2 (3), 244–263.
- Biau G (2012) Analysis of a Random Forests Model. Journal of Machine Learning Research 13, 1063-1095.
- Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone. 1984. Classification and regression trees. Monterey, Calif., U.S.A.: Wadsworth, Inc.
- CIPD (2017) HR outlook-winter. https://www.cipd.co.uk/knowledge/strategy/hr/outlook-reports.
- Cortes C and V Vapnik (1995) Support-Vector Networks. Machine Learning 20(3), 273-297.
- ÇOMAK, E. (2008). Destek Vektör Makinelerinin Etkin Eğitimi Ġçin Yeni Yaklaşımlar, Doktora Tezi, Selçuk Üniversitesi Fen Bilimleri Enstitüsü, Konya.
- Dolatabadi SH (2017) Designing of Customer and Employee Churn Prediction Model Based on Data Mining Method and Neural Predictor. Procedia- The 2nd International Conference on Computer and Communication Systems.
- Ekawati AD (2019) Predictive Analytics In Employee Churn: A Systematıc Literature Review. Journal of Management Information and Decision Sciences 22(4), 387-397.
- Freund Y and R Shapire (1996) Experiments with a new boosting algorithm. Procedia-13th International Conference, 148–156.
- Guenther N and M Schonlau (2016) Support vector machines. The Stata Journal 16(4), 917-937.
- Isson, JP and JS Harriott (2016) Big Data and People Analytics. People Analytics in the Era of Big Data, 331-356. Wiley, New Jersey.
- Kim P (2017) Neural Network and Classification. MATLAB Deep Learning,81-102. Appress, Berkeley, CA.
- Lee TW, P Hom, BE Marion, JIL Junchao and TR Mitchel (2017) On the Next Decade of Research in Voluntary Employee Turnover. The Academy of Management Perspectives 31(3), 201-221.
- Lindsay S, C Sheehan and HD Cieri (2020) The Influence of Workgroup Identification on Turnover Intention and Knowledge Sharing: The Perspective of Employees In Subsidiaries. The International Journal of Human Resources Management 31(3), 432-455.
- Matel E, F Vahdatikhaki, S Hosseinyalamdary, T Evers and H Voordjik (2019) An artificial neural network approach for cost estimation of engineering services. International Journal of Construction Management 1-15.
- Mishra SN, DR Lama and Y Pal (2016) Human Resource Predictive Analytics (HRPA) For HR Management In Organizations. International Journal of Scientific & Technology Research 5(5), 3-5.
- Rashid TA, and AL Jabar (2016) Improvement on predicting employee behaviour through intelligent techniques. IET Networks 5(5),136–142.
- Saradhi VV and GK Palshikar (2011) Expert Systems with Applications Employee Churn Prediction. Expert Systems with Applications 38(3), 1999-2006.
- Shawe J and NC Taylor (2004) Kernel Methods for Pattern Analysis. Cambridge

University Press, Cambridge.

- Strohmeier S and F Piazza (2013) Domain Driven Data Mining in Human Resource Management: A Review of Current Research. Expert Systems with Applications 40(7), 2410-2420.
- Tarnowska K, ZW Ras, L Daniel (2019) Customer Attrition Problem. In K Tarnowska (ed) Recommender system for improving customer loyalty, 113-122. Springer International Publishing, USA.
- Yiğit Oİ and H Shourabizadeh (2017) An Approach for Predicting Employee Churn by Using Data Mining. Procedia, International Artificial Intelligence and Data Processing Symposium, 1-5.