Çukurova Üniversitesi Mühendislik Fakültesi Dergisi, 36(4), ss. 913-928, Aralık 2021
Cukurova University Journal of the Faculty of Engineering, 36(4), pp. 913-928, December 2021

Machine Learning Approach for Predicting Employee Attrition and Factors Leading to Attrition

İrem ERSÖZ KAYA[*1] ORCID *0000-0001-5553-3881*
Oya KORKMAZ[2] ORCID *0000-0003-4570-803X*

[1]*Tarsus University, Faculty of Engineering, Computer Engineering Department, Tarsus*
[2]*Tarsus University, Faculty of Applied Sciences, International Trade and Logistics Department, Tarsus*

Geliş tarihi: 07.06.2021 Kabul tarihi: 10.12.2021

Atıf şekli/ How to cite: KAYA, İ.E., KORKMAZ, O., (2021). Machine Learning Approach for Predicting Employee Attrition and Factors Leading to Attrition. Çukurova Üniversitesi, Mühendislik Fakültesi Dergisi, 36(4), 913-928.

Abstract

In this study that aims to prevent the attrition of human resource which is so important for enterprises, as well as to prevent the leave of employment which is the natural result of such attrition, employee attrition and factors causing attrition are tried to be determined by predictive analytics approaches. The sample dataset which contains 30 different attributes of 1470 employees was obtained for the analysis from a database provided by IBM Watson Analytics. In the study, seven different machine learning algorithms were used to evaluate the prediction achievements. The gain ratio approach was preferred in determining the factors causing attrition. The key point of the study was to cope with the imbalanced data through resampling with bootstrapping. Thereby, even in the blind test, prospering prediction performances reaching up to 80% accuracy were achieved in robust specificity without sacrificing sensitivity. Therewithal, the effective factors causing attrition were investigated in the study and it was concluded that the first 20 attributes ranked according to their gain ratio were sufficient in explaining attrition.

Keywords: Employee attrition, Predictive analytics, Machine learning, Feature selection, Data mining

Çalışan Yıpranmasının ve Yıpranmaya Neden Olan Faktörlerin Tahmininde Makine Öğrenimi Yaklaşımı

Öz

İşletmeler için oldukça önemli olan insan kaynağının yıpranmasının ve yıpranmanın doğal sonucu olan işten ayrılmanın önüne geçmek amacıyla yapılan bu çalışmada, yıpranmaya neden olan faktörler tahmine dayalı analitik tekniklerinden biri olan makine öğrenmesi yöntemleri kullanılarak belirlenmeye çalışılmıştır. Analiz için örnek veri seti IBM şirketi Watson Analytics programı kapsamında sunulan bir veri tabanından alınmıştır. Veri seti, 1470 adet çalışanın 30 farklı özniteliğini içermektedir. Çalışmada, tahmin başarısını değerlendirmek amacıyla yedi farklı makine öğrenmesi algoritması kullanılmıştır.

[*]Sorumlu yazar (Corresponding author): İrem ERSÖZ KAYA, *iremer@tarsus.edu.tr*

Yıpranmaya neden olan faktörlerin tespitinde ise kazanç oranı yaklaşımı tercih edilmiştir. Çalışmanın kilit noktası, bootstrap tekniği ile yeniden örnekleme yapılarak sınıfların örnek sayılarının dengelenmesidir. Sonuç olarak, yeniden örnekleme ile makine öğrenmesi yöntemlerinin anlamlı sonuçlar vermesi sağlanmış ve tahmin doğruluk performansı, kör test yapılmasına rağmen %80'ler seviyesine ulaşmıştır. Kazanç oranı ile yapılan öncelik sıralamasında ilk 20'de yer alan özelliğin, yıpranmaya neden olan öncelikli faktörler olabileceği belirlenmiştir.

Anahtar Kelimeler: Çalışan yıpranması, Tahmin analitikleri, Makine öğrenmesi, Öznitelik seçimi, Veri madenciliği

1. INTRODUCTION

Attrition in human resources represents the gradual loss of employees. Employee attrition is an undesirable condition for businesses due to the costs involved. Employee attrition is one of the most important problems affecting employees in the sector [1]. Continuous employee turnover prevents a collective database to be formed in the enterprise. Furthermore, it reduces customer satisfaction levels as the customers always communicate with new staff. On the other hand, this causes an undesirable condition for the enterprise because the employees that quit their jobs may take valuable information, which is a competitive advantage, with them [2]. Therefore, an enterprise must reduce employee attrition as much as possible to maintain its relative advantage. It is vital for an enterprise to find the reasons that cause employee attrition and prevent them [3]. However, employing intuitional methods to this end can be difficult and time-consuming for decision makers, since many factors, such as employees' demographic and working conditions, must be considered. The use of analytical approaches based on prediction can provide administrators with a general idea regarding employee resignation rates, thus ensuring optimal human resource planning [4]. By determining the factors that lead to attrition, administrators can reduce the risk of resignations and take action to preserve high-value talent.

Predictive analytics is a field of study in which data is analyzed using numerical methods, such as data mining, statistics, machine learning, and artificial intelligence, while also making estimations about future events [5]. These methods are utilized in a number of sectors, including human resources, banking, e-commerce, retail, transportation, health and information technology [6]. This study will examine several examples of analytical methods based on prediction that are applied in the field of human resources. Factors that lead to attrition must be determined correctly and precautions must be taken to eliminate them in order to preserve valuable human resources. Numerous studies have already been carried out about the use of such analytical methods to determine the factors that contribute to employee attrition and estimating attrition rates.

In one such study, attrition trends were estimated using k-nearest neighbor and artificial neural network approaches by considering monthly working hours and the number of years spent at the company. The methods achieved prediction performances of 94.32% and 88.83%, respectively [7]. In another study, Alao and Adeyomo pre-classified the personnel in accordance with attrition groups [2]. This study used demographic information and work records of 109 employees who had resigned from a company operating in Nigeria to determine the factors that led to attrition. The authors developed a new predictive model for estimating these factors based on the results of rule clusters and decision tree models. This model applied the repeated incremental pruning technique, a machine learning method, resulting in a classification accuracy rate of 61%. Punnoose and Ajit also conducted a study that examined the problem of employee attrition with the use of machine learning methods. In that study, the naïve Bayes, random forest, support vector machines and k-nearest neighbor methods were used, yielding accuracy rates of 59%, 71%, 52% and 50% respectively [8].

In another study carried out to indicate that employee attrition represents a critical problem for enterprises (especially the resignation of key personnel), the decision tree, logistic regression, support vector machines, k-nearest neighbor, random forest, and naïve Bayes methods were applied to human resource data. Results were then analyzed by applying the feature selection method to the data. The analysis proved that machine learning methods can yield beneficial results in estimating employee attrition rates [9]. Çelik, meanwhile, compared the performance results of these methods by using data mining techniques to estimate attrition rates. A dataset, which was provided by IBM and included 35 different variables of 1,470 sample employees, was used in the study, which employed decision trees and support vector machines, yielding accuracy rates of 84.09% and 91.36% respectively [10].

The study is aimed at estimating employee attrition rates based on analytical approaches and determining the factors that lead to attrition. Results of this study demonstrate that resignation probabilities can be estimated using numerical methods in a rapid and successful manner, thus allowing decision makers to take precautions and reduce personnel turnover. The dataset, provided by IBM Watson, includes demographic features and information on the working conditions of 1,470 sample employees. Samples in the dataset include 30 different attributes classified in accordance with attrition conditions using the Support Vector Machine (SVM), Multilayer Perceptron (MLP), Radial Basis Function (RBF), Random Forest (RF), Bayes classifier (Bayes), k-Nearest Neighbor (kNN) and Repeated Incremental Pruning (JRip) algorithms. Sample numbers of the classes were equalized without changing the total sample number with the resampling method, bootstrapping that is a method used in data mining to balance datasets with a view to increasing success rates. This study focuses on effective factors in the prediction of employee attrition rates. For this purpose, factors sorted in terms of their effectiveness with the Gain Raito (GR) approach were evaluated.

2. MATERIAL AND METHOD

A dataset including information on the demographic features and working conditions of 1,470 sample employees was used in this study with the aim of estimating employee attrition rates. This dataset, provided by IBM Watson, originally includes 34 different independent variables (attributes) as in categorical and continuous structure. However, since four of the attributes have the same values for all samples, they were extracted from the data set. These attributes constitute the independent variables of the data, while dependent variables represent employee attrition with binary classification.

Attributes of the data have values with different scales and codes in accordance with their features. It is a known fact that an attribute with high values hinders the prediction performance of machine learning methods due to suppressing the goal [11]. Therefore, in data mining studies, standardization is sought by pre-processing data using different methods. In the study, min-max normalization was used for data scaling (Equation 1). The technique has been preferred due to the fact that it yields successful results in machine learning methods [12,13].

$$x_{norm} = \frac{x_i - x_{min}}{x_{max} - x_{min}} \quad (1)$$

In Equation 1, x_{norm}, represents the normalized value of x_i data, while min and max indexes represent the lowest and the highest value in data, respectively.

Data may also have an imbalanced class distribution, which can lead to bias in machine learning applications and prevent systems from functioning properly. Therefore, this study used bootstrapping to increase the number of samples in the smaller group. Meanwhile, synthetic data production was not preferred in this study due to the fact that the attributes of the dataset have different structural properties, e.g. some of them have a categorical structure. In synthetic data production, new samples produced by statistical

approaches or according to the values of neighboring samples may result in a value which does not exist among the current categories of the attribute. Therefore, sampling with a replacement was used while balancing the number of groups in order to preserve the general features of the data.

Two different applications were used in this study. In the first stage, sampled and non-sampled datasets were classified with algorithms using machine learning methods, such as the SVM, MLP, RBF, RF, Bayes, kNN and JRip. In the second stage, feature selection was carried out to determine the factors that lead to employee attrition. For this purpose, the attributes was ranked by Gain Ratio approach and the effect of these attributes were evaluated considering the success of machine learning methods.

In the applications, the data was divided into two sets in which two thirds was allocated for training and one third was reserved for testing. The training set was tested by applying 10-fold cross validation, while blind testing classification was executed with the test set. The group distributions of the datasets are given in Table 1.

Table 1. Class distribution of the datasets

Class	**Cross Validation**	**Blind Testing**
Attrition	163	74
Non-attrition	817	416
Total	980	490

The class distribution of the test set was preserved for blind testing, which was carried out to ensure the real success of the system. No resampling was employed for the blind testing dataset.

2.1. Classification Methods

Seven well-known machine learning approaches were employed for the classification in the application phase of the study. The MLP model of artificial neural networks, the Gaussian RBF network, the non-linear SVMs, the Bayesian network, the RF, the kNN and the JRip algorithms were all used in this study due to their different structural properties. All methods were implemented with the Weka software (version 3.8) by using the default parameters for each classifier, provided by the tool.

2.1.1. Multilayer Perceptron Neural Networks

Artificial neural networks, proposed in inspiration by the human mind and the biological neural system, represent one of the oldest and most widespread machine learning methods – one which has the ability to learn with the input data. An artificial network has a structure with three or more layers and interconnected processing units (artificial neurons) connecting the consecutive layers. Each connection in the network is associated with a numerical weight. An artificial neural network in a four-layer structure is shown in Figure 1. In artificial neural networks, intermediate layers are as hidden layers.

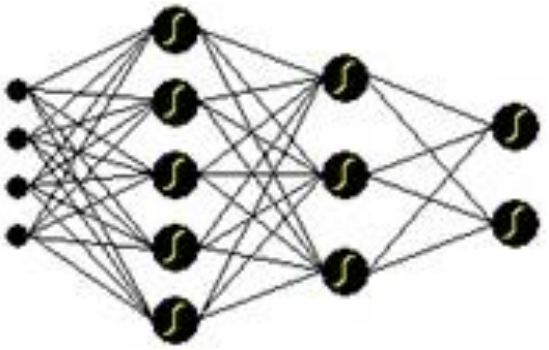

Figure 1. A simple four-layer artificial neural network

In a multilayer feed-forward neural network (multilayer perceptron), which is one of the artificial neural network models, input samples from the previous neurons are processed by calculating their weighted sum (net) and are transmitted to the next layer as an input via an activation function as $f(net)$ (Figure 2). Non-linear functions, such as sigmoid or hyperbolic tangents, are used as activation functions [14]. In the last layer (the output layer), the number of neurons are equal to the dimension of the output vector and an output value ($\hat{y}$) is formed for the processed information from the previous layers.

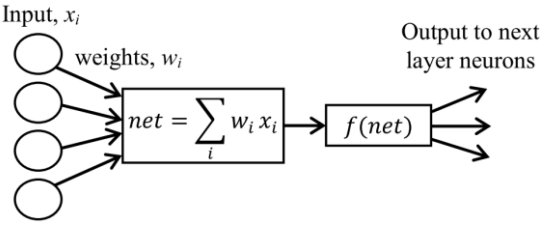

Figure 2. The general model of learning in ANN

The weights (w) are updated until the error arising from the difference between the real output values (y) and estimated output values ($\hat{y}$) reaches minimum [15]. This iterative modification, also known as the training process of the system, characterizes the learning ability of the artificial network. For each input sample given with $y \in R^s$ output vector, the error function (J) at each iteration (t) is formulated as;

$$J(w) = \frac{1}{2}\sum_{k=1}^{s}(y_k - \hat{y}_k)^2 \quad (2)$$

The most common algorithm used for MLP training is the backpropagation algorithm. In the algorithm, the error function is minimized by updating the weight of the layers with a gradient descent approach (Equation 3).

$$w_{kj}^{t+1} = w_{kj}^{t} - \eta \frac{\partial J}{\partial w_{kj}^{t}} \quad (3)$$

where, η represents the learning speed, while w_{kj} represents the weight that connects the k. neuron of the layer to the j. neuron of the previous layer.

2.1.2. Radial Basis Function Networks

RBF networks are a type of feed-forward artificial neural networks. Its difference from the MLP lies in the fact that the radial basis activation functions are used in the hidden layer and a non-linear cluster analysis is conducted on the direct inputs of the input layer [16]. Therefore, the hidden layer is the most important component of the network.

Each neuron in the hidden layer represents a center and a radial basis function form of the distance between the input samples, and these centers constitute the activation function ($\emptyset$). The measure of the distance of the input vector to the centers is mostly calculated with the Euclidian norm ($\|.\|_2$). A Gauss function is commonly preferred in the literature as an activation function, given in Equation 4 [17].

$$\varnothing(x, c_j) = exp\left(\frac{\left(\left\|x - c_{j2}\right\|\right)^2}{2\sigma^2}\right) \quad (4)$$

where, x represents the input vector of the network, σ the width parameter and c_j the radial basis centers. Various methods are proposed in the literature for the selection of the center vectors, like selecting the center vectors among the samples, allocating each sample as the center vector or determining the center vectors by unsupervised learning methods, such as k-means and self-organizing maps [18].

Net values, calculated by using the weights between two layers and the function outputs of the hidden layer, give the output values as the learning model as shown in Figure 2. Thereby, the output ($\hat{y}_k$) of an RBF network that has a hidden layer with N neuron is found by;

$$\hat{y}_k = \sum_{j=1}^{N} w_{kj} \varnothing_j(x, c_j) \quad (5)$$

(3)

where, w_{kj} represents the weight that connects the k. neuron of the layer to the j. neuron of the hidden layer.

2.1.3. Support Vector Machines

SVM is a machine learning algorithm which aims to find the hyperplane with the maximum distance to two classes [19]. This hyperplane (decision boundary) is defined by $\langle w, x_i \rangle + b = 0$ in a linearly separable sample space that consists of $(x_i, y_i), i = 1,2, \ldots, n$ data pair, which includes the $x_i \in R^d$ input vectors and the class values $y_i \in \{+1, -1\}$ of the samples. In this equation, b is the bias value and w is the weight vector (normal

for the hyperplane). The boundary points of two classes that are at the maximum distance to the hyperplane are referred to as support vectors (Figure 3). The margin between the parallel boundaries formed by the support vectors is defined by $2/\|w\|$. In this regard, the most proper hyperplane is obtained by maximizing the margin between the boundaries subject to the linear constraints of $y_i(\langle w, x_i\rangle + b) \geq 1$ [20].

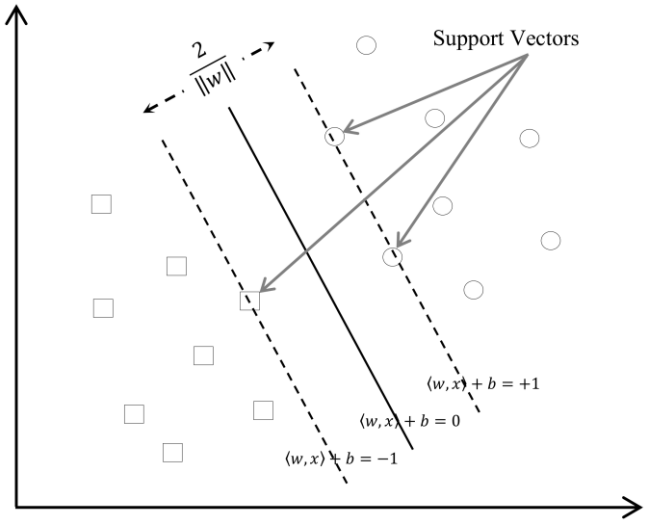

Figure 3. Decision boundary and support vectors of SVM

In real-world situations, datasets mostly contain outlier observations that cause noise. Constraints are released with a soft-margin approach that is used for outlier points in the created margin or on the wrong side of the margin [21]. According to this, an optimal hyperplane with the largest margin is achieved by minimizing the objective function within the constraints of $y_i(\langle w, x_i\rangle + b) \geq 1 - \xi_i$ and $\xi_i > 0$ (Equation 6).

$$\min_{w,b} \left\{ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \right\} \tag{6}$$

where, ξ_i represents the outlier variables used for calculating the error at the (x_i, y_i) point, while C represents the correction parameter that controls the balance between the margin maximization and error minimization. The quadratic optimization problem represented in Equation 6 is transformed into a dual form via Lagrange multipliers technique. Subsequently, it is simplified by using Karush-Kuhn-Tracker conditions [22]. The objective function transformed into dual form is defined in Equation 7.

$$\max_{a,w,b} \left\{ \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \left\langle x_i, x_j \right\rangle) \right\} \tag{7}$$

where, α_i represents the Lagrange multipliers and is defined at the $0 \leq \alpha_i \leq C$ interval. Furthermore, according to Karush-Kuhn-Tracker condition, there is a constraint for Equation 7 given by $\sum_{i=1}^{n} y_i\alpha_i = 0$. In the course of solving the problem, many α_i take the values of zero, while all training samples of $\alpha_i > 0$ are defined as support vectors (Sv). According to this, the weight vector (w^*) for the best solution can be given as follows;

$$w^* = \sum_{i \in Dv} \alpha_i y_i x_i \tag{8}$$

In cases where the data cannot be linearly separated, SVM changes the data into a linearly-separable form by carrying the input space to a

higher dimensional space with the help of the kernel function, $K(x_i, x) = \emptyset(x_i) \cdot \emptyset(x)$. The decision function obtained by using the solution of the objective function, along with the kernel trick, is shown in Equation 9.

$$f(x) = sgn\left(y_i \alpha_i K(x_i, x) + b\right) \quad (9)$$

The classification for each x value is conducted via calculating the kernel function of x. One of the most commonly used and successful kernel functions is the RBF kernel [23].

2.1.4. Random Forest

Random Forest is an ensemble-learning algorithm that has decision trees at its basis. In the ensemble-learning approach, a single prediction model is created by gathering more than one classifier. The classifiers used in the RF are decision trees. These decision trees combine and form the decision forest [24].

In the RF approach, N number of sub-sample data is created by using a bootstrap technique to grow each tree in the forest. Two thirds of each sample is used to create training data (in-bag), which is needed to create the model. The remaining third is then used to create a test set (out-of-bag/OOB), which is used to measure the model's performance. Then, an unpruned classification or regression tree is created for the in-bag dataset of each sub-sample data by using the CART algorithm [25]. For this, m number of random variables $(m < p)$ is selected out of all variables (p) at each node. The most suitable one for branching out is determined with information gain. The best cut-off value for the variable is calculated using the Gini index, and each node is split into two sub-branches (Equation 10). All these processes are conducted for all nodes until a leaf node is achieved.

$$\sum \sum_{j \neq i} (f(S_i, T)/|T|)\left(f\left(S_j, T\right)/|T|\right) \quad (10)$$

where, T represents the training dataset, S is the class of the randomly selected sample and $(f(S_i, T)/|T|)$ gives the probability of the selected sample belonging to class S_i.

At the end of the training, each branch of the tree represent the parameters of the split function determined in accordance with the x input vector, while each leaf gives the distribution of the y output variable. In RF, the independence of each tree is ensured by randomly choosing a subset from the training data. In the testing stage, each sample of the OOB data goes through each tree with respect to the split function and classified. According to the classification results, an OOB error rate is calculated for each tree in the forest, and then each tree is endowed with an inversely proportional weight in accordance with this error rate [26]. Each decision tree votes in compliance with its weight for the classification of each sample in the dataset. All the prediction values (votes) of the trees are gathered and the class that gets the most votes is assigned to that sample.

2.1.5. Bayes Networks

The Bayes network is a machine learning method that presents variables and the probabilistic relationships of the variables with a directed acyclic graph (DAG). The Bayes network is considered as one of the strongest methods of interpretation and inference, due to its ability to graphically depict the complex relations between variables while also numerically representing them [27].

A Bayes network contains many nodes and edges (arrows) that connect these nodes according to their probabilistic relationships. In the network, the nodes with arrows directed towards them are called child nodes, while nodes with arrows directed away from them are called parent nodes. Nodes without children are leaf nodes and nodes without parents are root nodes. Each node in the network corresponds to a random variable (attribute) and is associated with a conditional probability distribution of the random variables given as its parents. The directed edges between the nodes represents a probabilistic dependency between the variables. The variables without

parents have a marginal probability distribution. A Bayes network sample with five variables is shown in Figure 4.

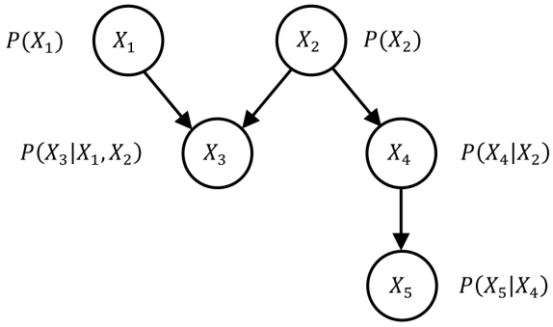

Figure 4. The sample of a Bayesian network

Any variable in a Bayes network is independent of all variables except its descendants, given its parents [28]. Thus, the joint probability distribution in a Bayes network is derived by the chain rule;

$$P(X_1,\ldots,X_d)=\prod_{i=1}^{d}P\left(X_i|ebeveyn\left(X_i\right)\right) \qquad (11)$$

where, d represents the number of variables in the network, while X_i represents the random variables.

Learning in Bayes networks is conducted with the prediction of the graphical structure of the network and the parameters of the joint probability distributions. The process of determining the structure of the network, which is defined as structural learning, includes the identification of the variables and their relationships within the model. Predictions can be performed by proper inferences considering the optimum parameters and the network structure, which are constructed as a result of the learning process carried out with many different algorithms [29].

2.1.6. k-Nearest Neighbor

First introduced by Fix and Hodges in 1951, the 'nearest neighbor' approach is one of the simplest and oldest machine learning techniques that depends on linear supervised pattern recognition [30]. In this technique, the model is fed with a training set, and this training set is used to classify objects. According to this, an unknown sample of the prediction set is classified pursuant to the majority class of its k nearest neighbor. The value of k has a great effect on the accuracy of kNN, and the optimum value of the parameter can be found by using cross-validation during the adjustment of the model [31].

2.1.7. Repeated Incremental Pruning

JRip, a rule based machine learning method, was proposed by Cohen as an optimized version of the IREP learning algorithm [32]. JRip applies the Repeated Incremental Pruning to Produce Error Reduction (RIPPER) technique which is a rule learning in Java. The algorithmm is based on the creation of a rule set that includes all positive samples. In the algorithm, decision rules are created in the form of IF-THEN statements for each class of the training set and then pruned [33].

JRIP, initially, partitions the samples of the training set into two subsets as growing and pruning. Later, an initial rule set is produced from the samples of the growing set using heuristic methods. The overgrown rule set is then repeatedly simplified by pruning any single condition or any single rule. At each stage of the simplification, the pruning which provides the greatest reduction in error in the pruning set is performed. The simplification ends when the error starts to increase.

A rule set starts as an empty set and the rules are gradually added to the rule set until the total description length of the new rule set is d bits larger than the smallest description length of the previous rule set or there are no positive examples left [34]. A rule is grown by adding conditions to the rule until none of the negative samples are covered i.e. until it achieves 100% accuracy. During the growing process, all the values of each attribute are tested and the condition that has the highest information gain is selected. In the sequel, JRip pruning either changes the place of the individual conditions or reorganizes them with reduced error pruning, in order to increase the accuracy of the rules [33].

2.2. Feature Selection

The Gain Ratio (GR) approach, used for ranking attributes according to priorities in data mining, was employed in feature selection. The GR approach is widely used due to its suitability for datasets with multi-valued attributes which have many probable values [35]. It is an improved extension of the information gain measure.

Information gain is a measure that gives the amount of knowledge gained about a random variable (Y) from another random variable (X) by using the entropy model [36]. Entropy (H) is defined as the measure of a system's unpredictability (Equation 12).

$$H(Y) = -\sum_{y \in Y} p(y) log_2 \left(p(y) \right) \quad (12)$$

The conditional entropy of Y given the value of attribute X is shown as;

$$H(Y \setminus X) = -\sum_{x \in X} p(x) \sum_{y \in Y} p(y \setminus x) log_2 \left(p(y \setminus x) \right) \quad (13)$$

With reference to Equation 13, the information gain and the GR are given in Equations 14 and 15, respectively.

$$\text{Information Gain} = \text{H}(\text{Y}) - \text{H}(\text{Y}\setminus\text{X}) \quad (14)$$

$$\text{GR} = \frac{\text{Information Gain}}{\text{H(X)}} \quad (15)$$

For the feature selection phase of the study, three new datasets were created by selecting the first 20, 15 and 10 attributes ranked with GR. The machine learning methods were then reapplied to the reorganized datasets with the selected attributes.

2.3. Performance Assessment

Estimating employee attrition is a binary-classification problem, and there are plenty of measures that have been used to assess the performance of predictive models in such problems. The widely preferred measures to quantify the prediction power of the methods can be given as sensitivity (*Sens*), specificity (Spec), accuracy (*Acc*), receiver operating characteristic (ROC), and area under ROC curve (AUC) values. The AUC value is the area underneath the ROC curve, in which the sensitivity (true positive rate) is plotted in function of 1-Specificity (false positive rate) for different threshold points of a parameter. *Acc*, *Sens* and *Spec* indices are defined by the following equations;

$$Accuracy\left(Acc\right) = \frac{TP + FN}{TP + FP + TN + FN} \quad (16)$$

$$Sensitivity\left(Sens\right) = \frac{TP}{TP + FN} \quad (17)$$

$$Specificity\left(Spec\right) = \frac{TN}{TN + FP} \quad (18)$$

where TP (true positive) is the number of correctly classified positives; TN (true negative) is the number of correctly classified negatives; FP (false positive) is the number of incorrectly classified positives; and FN (false negative) is the number of incorrectly classified negatives. Accordingly, sensitivity and specificity represent the fraction of correctly identified samples as attrition and non-attrition, respectively. Since all three measures are critically affected by the relative frequency of the target, they are not suitable for the isolated evaluation [37]. A sensitivity of less than 50% but a specificity of more than 80% demonstrates an under-prediction of a predictor which has the tendency of predicting non-attrition more than attrition. Furthermore, the AUC is also biased towards negative examples when data are imbalanced with few positives relative to negatives [38].

When the properties of the traditional measures are considered, an unbiased measure is required for evaluating an imbalanced data. Therefore, probability excess (*ProbEx*) was proposed as an

unbiased measure for evaluating the performance of prediction by Yang et al. [39]. Probability excess is independent of the relative class frequencies by means of the evaluation of sensitivity and specificity values in cooperatively with sensitivity + specificity − 1, that can be graphed by a plot of sensitivity versus specificity. It is defined by the following equation;

$$Probability\,Excess\left(ProbEx\right)=\frac{TPxTN-FPxFN}{\left(TP+FN\right)x\left(TN+FP\right)} \quad (19)$$

The values of greater than 0.5 reveal an acceptable prediction performance in probability excess criteria. Here the value of 1 is also an indicator of a perfect predictor.

3. RESULTS AND DISCUSSION

In this study, employee attrition was intended to predict considering several factors and the factors that are effective in the attrition were evaluated. For this purpose, a dataset which was provided by IBM including 30 attributes relating to the demographic and working conditions of 1,470 employees was used in the study. In the research, employee attrition was evaluated based on the prediction results which was obtained by applying seven different machine learning algorithms to the dataset. To this end, firstly, the dataset was partitioned into two parts as training and testing. Next, 10-fold cross validation was applied to training dataset to evaluate the prediction performances. Finally, blind testing was executed in order to validate the reliability of the prediction success of the methods.

The implementation phase of the study was carried out in two parts. In the first part, prediction performances of the methods were obtained as a result of the applications executed with sampled and non-sampled datasets. In the second part, feature selection was performed with the sampled dataset. At this stage, the factors provoking employee attrition were evaluated in terms of their effects on the prediction success.

3.1. Estimating Employee Attrition

In the application phase of the study, the raw state of the normalized data was used. The prediction performance of the machine learning methods was appraised by the average results of the 10-fold cross validation applied to the training set, and later attested with blind testing. The results obtained from the applications with the raw data are presented in Table 2.

Table 2. The prediction performance results for the raw training dataset

Method	Cross Validation				Blind Testing			
	Sens	***Spec***	***Acc***	**AUC**	***Sens***	***Spec***	***Acc***	**AUC**
SVM	0.33	0.98	0.87	0.66	0.37	0.98	0.88	0.67
MLP	0.42	0.91	0.83	0.77	0.51	0.94	0.87	0.80
RBF	0.36	0.96	0.86	0.84	0.38	0.97	0.88	0.80
RF	0.14	0.99	0.85	0.79	0.20	0.99	0.87	0.81
Bayes	0.42	0.88	0.80	0.74	0.40	0.91	0.83	0.74
kNN	0.35	0.91	0.82	0.63	0.30	0.87	0.79	0.59
JRip	0.26	0.94	0.82	0.61	0.28	0.95	0.85	0.62

When the results reported in Table 2 are examined, it can be seen that the *Sens* value is too low for all methods, and the values reveal that employee attrition was not successfully predicted. On the other hand, the classification accuracy of non-attrition state was reached over 90%. The situation is called as the tendency of predicting non-attrition i.e. *under-prediction*. The distribution of the *probEx* values which are represented by red dots on the triangular graph given in Figure 5 and Figure 6 also show this situation. All of the red dots are found on the left-hand side of both graphs. This is due to the fact that the samples of attrition are rather few compared to the other class. As is known, imbalanced distribution of data results in bias during the systems' learning.

In the second part of the machine learning applications, the runs were repeated on a recreated training dataset in which the samples of the rare class were increased via bootstrapping while the blind testing data was preserved unchanged to ensure the correct comparisons of the reliability of the methods. The results are provided in Table 3.

Table 3. The prediction performance results for the resampled training dataset

Method	Cross Validation				Blind Testing			
	Sens	***Spec***	***Acc***	**AUC**	***Sens***	***Spec***	***Acc***	**AUC**
SVM	0.87	0.78	0.82	0.82	0.73	0.76	0.76	0.75
MLP	0.94	0.92	0.93	0.93	0.61	0.83	0.80	0.77
RBF	0.82	0.76	0.79	0.87	0.72	0.78	0.77	0.80
RF	0.97	0.94	0.96	0.99	0.45	0.92	0.85	0.79
Bayes	0.70	0.70	0.70	0.80	0.69	0.73	0.72	0.74
kNN	0.97	0.89	0.93	0.94	0.46	0.80	0.75	0.63
JRip	0.89	0.82	0.85	0.88	0.70	0.70	0.70	0.72

The results of both cross validation and blind testing show that elimination of the imbalance in data distribution by resampling led to a considerable increase in the success of attrition state prediction. The *probEx* values of the resampled data (blue dots) shown in both Figure 5 and Figure 6 verify the argument as well.

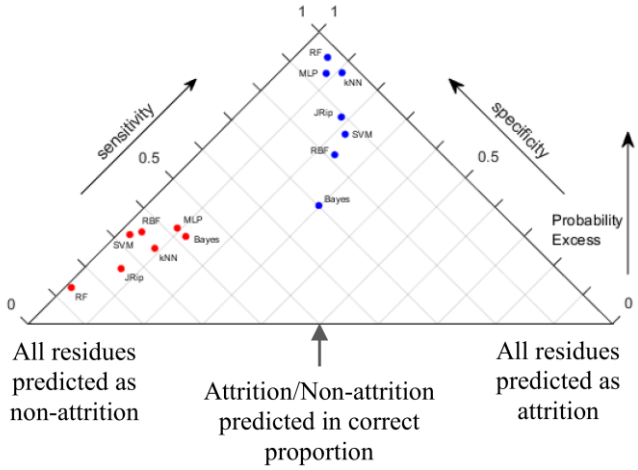

Figure 5. The cross validation performance of the seven methods on both raw data (red dots) and resampled data (blue dots)

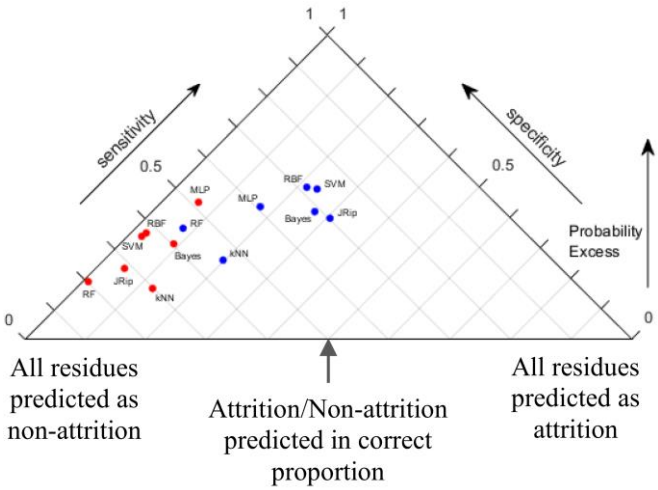

Figure 6. The blind testing performance of the seven methods on both raw data (red dots) and resampled data (blue dots)

Especially, in Figure 5 which shows the cross validation performance of the methods, it can be clearly seen that the blue dots are distributed over the middle of the pyramid. This is due to the fact that there is no great difference between the sensitivity and specificity. This indicates that a relatively balanced prediction accuracy has been achieved on the cross-validation of the methods.

When the success rates of the methods' blind testing are evaluated, it is also observed that the *Sens* value rise above 60% and more balanced estimates are obtained in all methods except RF and kNN (Figure 6). At this point, it was determined that the *Sens* value of the RF method, which has the highest *Acc* value at 85%, stayed below 50%, thus failing to achieve a significant estimation. The RBF and SVM are especially notable in terms of their balanced prediction performances among all the methods. With the purpose of comparing the performance of the seven machine learning methods in blind testing, the ROC curve shown in Figure 7 was plotted. When the curves in the figure are examined, it can be seen that the RBF method has the highest prediction success. The lowest prediction success was obtained by using the kNN method.

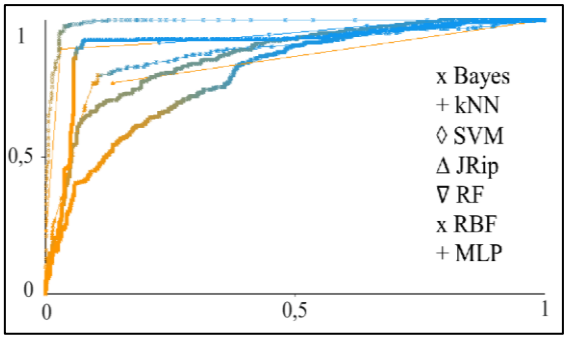

Figure 7. ROC curve for blind testing performance of the seven methods on resampled data

Considering the similar prior studies, it is observed that the prediction performance of machine learning methods attains an accuracy of 95% [7, 10]. Similarly, in this study, the classification performance achieved with 96% accuracy rate. Moreover, it is of great importance to perform both a cross validation and blind testing to ensure reliability of the prediction results. *Acc* results of up to 88% were achieved in accordance with blind testing, which was carried out to measure real success. However, *Sens* and *Spec* values must also be considered in the evaluation process. In order to make a meaningful interpretation on the prediction performance of the methods, these values should be relatively close each other and over 50%. Therefore, in the study, the sample number of the rare class was increased by resampling in order to provide a significant improvement in prediction success. Thereby, successful results reaching up to 80% were achieved in robust specificity without sacrificing sensitivity.

3.2. Determining Employee Attrition Factors

In the feature selection phase of the study, the training and testing data were rearranged by selecting the first 20, 15 and 10 attributes, respectively, ranked according to their discrimination capability with the GR approach. At this stage, the resampled dataset which has a balanced class distribution was used since it enables to obtain significant results. The machine learning methods were reapplied to the arranged datasets with respect to the selected attributes in order to investigate the factors that cause employee attrition.

The first 20 attributes ranked by GR and taken into account for the evaluation are as follows: monthly income, job level, total working years, overtime, stock option level, age, years with current manager, job involvement, marital status, job role, years at company, years in current role, education field, work life balance, environment satisfaction, job satisfaction, business travel, distance from home, years since last promotion, and department. Application results for the first 20, 15 and 10 attributes are given in Tables 4, 5 and 6, respectively.

In the light of the results provided in the Table 4, it can be said that the prediction accuracy of the methods was preserved in general, while the *Sens* values of the methods were increased at least 1% except for MLP and Bayes (Table 4).

Consequently, the balanced prediction results of above 70% were achieved.

Table 4. The prediction performance results for the datasets arranged with the first 20 attributes

Method	Cross Validation				Blind Testing			
	Sens	***Spec***	***Acc***	**AUC**	***Sens***	***Spec***	***Acc***	**AUC**
SVM	0.86	0.76	0.81	0.81	0.74	0.74	0.74	0.74
MLP	0.93	0.88	0.91	0.92	0.61	0.78	0.76	0.76
RBF	0.84	0.77	0.80	0.86	0.73	0.76	0.75	0.80
RF	0.97	0.93	0.95	0.99	0.50	0.91	0.85	0.81
Bayes	0.70	0.70	0.70	0.80	0.69	0.73	0.72	0.74
kNN	0.96	0.90	0.93	0.93	0.47	0.80	0.76	0.64
JRip	0.90	0.79	0.84	0.87	0.73	0.69	0.70	0.74

For the first 15 attributes with the highest GR, similar to the first 20 attributes, the prediction accuracy of the methods was preserved, however there is a slight decrease in *Sens* values (Table 5).

Table 5. The prediction performance results for the datasets arranged with the first 15 attributes

Method	Cross Validation				Blind Testing			
	Sens	***Spec***	***Acc***	**AUC**	***Sens***	***Spec***	***Acc***	**AUC**
SVM	0.78	0.78	0.78	0.78	0.73	0.77	0.77	0.75
MLP	0.94	0.86	0.90	0.90	0.60	0.73	0.71	0.72
RBF	0.79	0.77	0.78	0.85	0.73	0.78	0.77	0.78
RF	0.97	0.91	0.94	0.99	0.47	0.91	0.84	0.79
Bayes	0.70	0.71	0.70	0.80	0.68	0.73	0.72	0.74
kNN	0.97	0.90	0.93	0.94	0.42	0.82	0.76	0.61
JRip	0.87	0.79	0.83	0.87	0.68	0.74	0.73	0.72

Results of the application on the first 10 attributes are provided in Table 6, where it is observed that the prediction accuracy of employee attrition was significantly decreased according to the situation in which no attribute was selected.

Table 6. The prediction performance results for the datasets arranged with the first 10 attributes

Method	Cross Validation				Blind Testing			
	Sens	*Spec*	*Acc*	AUC	*Sens*	*Spec*	*Acc*	AUC
SVM	0.72	0.81	0.77	0.77	0.609	0.79	0.76	0.69
MLP	0.83	0.81	0.82	0.85	0.61	0.79	0.76	0.72
RBF	0.75	0.77	0.76	0.82	0.70	0.76	0.75	0.77
RF	0.97	0.90	0.94	0.98	0.41	0.88	0.81	0.74
Bayes	0.74	0.71	0.72	0.79	0.65	0.73	0.72	0.74
kNN	0.46	0.81	0.75	0.64	0.465	0.81	0.75	0.64
JRip	0.83	0.76	0.79	0.81	0.53	0.82	0.78	0.68

When the results are evaluated in general, it can be interpreted that the first 20 attributes have adequate explanatory power in predicting the attrition since the accuracy results of the methods obtained with these attributes did not change compared to those obtained without the selection of the attributes, and even the prediction success of the class which has the small sample size has increased. On the other hand, when the attribute number was reduced to 15, a slight decrease in success values is observed. Accordingly, it is possible to say that the five extracted features or some of these are among the factors affecting attrition.

When similar previous studies investigating factors affecting employee attrition are examined, the attributes similar to those found in this study become prevalent. For example, in a study conducted by Alduayj and Rajpoot, the most influential factors in terms of attrition are sorted as marital status, years with current manager, stock option level, business travel, job role, job involvement, job satisfaction and environment satisfaction [40]. In another study in which different machine learning methods were applied to the dataset provided by IBM Watson in order to predict employee attrition, it is found that job role is among the dominant factors causing attrition [41]. In yet another study using machine learning techniques, the data collected from an insurance company was evaluated and an attribute selection method was applied to these data, which are grouped in three categories: customer, product and vehicle. In customer category, age and gender were determined to be among the effective factors [42]. In a similar study carried out with machine learning methods, the five most affecting factors on attrition were reported as overtime, monthly income, daily rate, age, and total working years [43].

4. CONCLUSION

Employee attrition is an important problem in today's business world. Because the personnel turnover rate in enterprises is increasing as a natural result of attrition. The leaving of such personnel, who have been recruited as a result of a great deal of endeavor and effort, causes the loss of

cost and time for enterprises. In addition, employee attrition negatively affects the career success of managers and leads to an increase in the number of occupational accidents caused by recruited personnel. Therefore, it is very important to predict attrition. Estimating the number of personnel to leave the job, which will occur as a natural consequence of attrition will help managers to make human resource planning more accurately. As managers make the human resources planning more accurately, the businesses will become sustainable and gain a competitive advantage.
It is a difficult and laborious process to determine attrition through traditional research techniques. In addition, the estimation performance can be low as it is an evaluation with multiple factors. Therefore, in the study, attrition and the factors that cause attrition were intended to be determined by using predictive analytical approaches. For this purpose, seven different machine learning methods were applied to the IBM Watson dataset containing the demographic characteristics and working conditions of 1470 employees and the predictive success of the methods were evaluated by comparing the results. Additionally, in order to increase the predictive performance of the methods, the class distribution in the dataset was balanced with resampling by using the technique of bootstrapping and the results were compared by repeating the applications. Using the resampling technique, the predictive success of the attrition class which is less in the dataset was increased and thus, in the blind test, balanced prediction performances were obtained with an accuracy level of 80%. Among the methods used, the most successful ones were found to be RBF and SVM in terms of balanced Sens and Spec values and high success rates.

Later in the study, 30 features were ranked in terms of explanatory power using the gain ratio approach, and the application results obtained with the first 20, 15, and 10 features were interpreted to determine the factors that affect attrition. Based on the prediction results, it was concluded that the first 20 factors were sufficient in explaining attrition. It was inferred that the determined factors show similarities to those found in the studies in the literature. In addition, despite blind testing was used in the study, comparatively more robust results were obtained with high prediction accuracy rates and balanced prediction performances.

The difficulty of analyzing and interpreting big data by end users, here managers, who lack technical knowledge, raises the need to design systems that allow clear and understandable interpretation of the data rather than expressing it in numbers. As a future work, a system comprising of a combinatorial machine learning approach in which several feature selection algorithms are hierarchically included, could provide an increase in success for determining the factors leading attrition. Such a system that performs forecasting with the successful machine learning methods having preprocessed the data with bootstrapping in the case of imbalanced distribution and determines important factors by the feature selection methods found by trying out different algorithms, would provide support to human resources managers in decision-making processes.

As a conclusion, it was shown that it is possible to use machine learning techniques as a business analytics, successfully. The study will benefit the management and production of new generation human resource policies in terms of predicting employee attrition more accurately and more easily, as well as providing information to help reduce the staff turnover rate as a result of taking measures to eliminate the factors causing attrition by identifying them.

5. REFERENCES

1. Sridhar, G.V., Vetrivel, S., Venugopal, S., 2018. Employee Attrition and Employee Retention-challenges & Suggestions. 2018 International Conference on Economic Transformation with Inclusive Growth-2018, Chennai, India, 1-16.
2. Alao, D., Adeyemo, A.B., 2013. Analyzing Employee Attrition Using Decision Tree Algorithms. Computing, Information Systems & Development Informatics Journal, 4(1), 17-28.

3. Srivastava, D.K., Nair, P., 2017. Employee Attrition Analysis Using Predictive Techniques. 2017 International Conference on Information and Communication Technology for Intelligent Systems, Ahmedabad, India, 293-300.

4. Raman, R., Bhattacharya, S., Pramod, D., 2019. Predict Employee Attrition by Using Predictive Analytics. Benchmarking: An International Journal, 26(1), 2-18.

5. Gandomi, A., Haider, M., 2015. Beyond the Hype: Big Data Concepts, Methods and Analytics. International Journal of Information Management, 35(2), 137-144.

6. Zhao, W., Pu, S., Jiang, D., 2020. A Human Resource Allocation Method for Business Processes Using Team Faultlines. Applied Intelligence, 50, 2887-2900.

7. Yedida, R., Reddy, R., Vahi, R., Jana, R.J., Gv, A., Kulkarni, D., 2018. Employee Attrition Prediction, arXiv:1806.10480, https://arxiv.org/ftp/arxiv/papers/1806/1806.10480.pdf

8. Punnoose, R., Ajit, P., 2016. Prediction of Employee Turnover in Organizations Using Machine Learning Algorithms. International Journal of Advanced Research in Artificial Intelligence, 5(9), 22-26.

9. Shankar, R.S., Rajanikanth, J., Sivaramaraju, V.V., Murthy, K.VSSR., 2018. Prediction of Employee Attrition Using Datamining. 2018 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN), Pondicherry, India, 335-342.

10. Çelik, U., 2019. Estimation of Employee Attrition in Business Life Balance with Data Mining Methods. Journal of Management and Economics Research, 17(1), 63-76.

11. Sevilla, J., 1997. Importance of Input Data Normalization for the Application of Neural Networks to Complex Industrial Problems. IEEE Transactions on Nuclear Science, 44(3), 1464 – 1468.

12. Zhang, Y-P., Qiqige, W., Zheng, W., Liu, S., Zhao, C., 2016. gDNA-Prot: Predict DNA-Binding Proteins by Employing Support Vector Machine and a Novel Numerical Characterization of Protein Sequence. Journal of Theoretical Biology, 406, 8-16.

13. Christo, V.R.E., Nehemiah, H.K., Minu, B., Kannan, A., 2019. Correlation-based Ensemble Feature Selection Using Bioinspired Algorithms and Classification Using Backpropagation Neural Network. Computational and Mathematical Methods in Medicine, 7398307, 1-17.

14. Wang, Z., Fu, Y., Huang, T.S., 2019. Signal Processing. Deep Learning Through Sparse and Low-rank Modeling, San Diego, USA: Academic Press, 121-142.

15. Duda, R.O., Hart, P.E., Stork, D.G., 2000. Pattern Classification. John Wiley & Sons, New York, USA, 688.

16. Raitoharju, J., Kiranyaz, S., Gabbouj, M., 2016. Training Radial Basis Function Neural Networks for Classification via Class-specific Clustering. IEEE Transactions on Neural Networks and Learnıng Systems, 27(12), 2458-2471.

17. Schwenker, F., Kestler, H.A., Palm, G., 2001. Three Learning Phases for Radial-basis-function Networks. Neural Networks, 14, 439-458.

18. Faris, H., Aljarah, I., Mirjalili, S., 2017. Evolving Radial Basis Function Networks Using Moth–flame Optimizer. Samui, P., Sekhar, S., Balas, V.E., (Ed.), Handbook of Neural Computation, San Diego, USA: Academic Press, 537-550.

19. Cortes, C., Vapnik, V., 1995. Support-Vector Networks. Machine Learning, 20, 273-297.

20. Battineni, G., Chintalapudi, N., Amenta, F., 2019. Machine Learning in Medicine: Performance Calculation of Dementia Prediction by Support Vector Machines (SVM). Informatics in Medicine Unlocked, 16:100200, 1-8.

21. Awad, M., Khanna, R., 2015. Support Vector Machines for Classification. Awad, M., Khanna, R., (Ed.). Efficient Learning Machines, Berkeley, CA: Apress, 39-66.

22. Ibrikçi, T., Üstün, D., Ersöz Kaya, I., 2012. Diagnosis of Several Diseases by Using Combined Kernels with Support Vector Machine. Journal of Medical Systems, 36(3), 1831-1840.

23. Öztürk, G., Çimen, E., 2019. Polyhedral Conic Kernel-like Functions for SVMs, Turkish

Journal of Electrical Engineering & Computer Sciences, 27, 1172-1180.

24. Breiman, L., 2001. Random Forests. Machine Learning, 45(1), 5-32.

25. Pal, M., 2005. Random Forest Classifier for Remote Sensing Classification. International Journal of Remote Sensing, 26(1), 217-222.

26. Winham, S.J., Freimuth, R.R., Biernacka, J.M., 2013. A Weighted Random Forests Approach to Improve Predictive Performance. Statistical Analysis and Data Mining, 6(6), 496-505.

27. Chan, A.P.C., Wong, F.K.W., Hon, C.K.H., Choi, T.N.Y., 2018. A Bayesian Network Model for Reducing Accident Rates of Electrical and Mechanical (E&M) Work. International Journal of Environmental Research and Public Health, 15(11):2496, 1-19.

28. Carson, E., Cobelli, C., 2014. Modelling Methodology for Physiology and Medicine. Elseiver, Waltham, USA, 588.

29. Ruz, G.A., Araya-Diaz, P., 2018. Predicting Facial Biotypes Using Continuous Bayesian Network Classifiers. Complexity, (4075656), 1-14.

30. Fix, E., Hodges, J.L., 1951. Discriminatory Analysis-nonparametric Discrimination: Consistency Properties. Project No. 2-49-004, Report No. 4, Contract No. AF 41(128)-31, USAF School of Aviation, Randolph Field, Texas.

31. Lu, L., Zhu, Z., 2014. Prediction Model for Eating Property of Indica Rice. Journal of Food Quality, 37, 274-280.

32. Cohen, W.W., 1995. Fast Effective Rule Induction. 1995 Twelfth International Conference on Machine Learning, California, 115-123.

33. Rezapour, M., Zadeh, M.K., Sepehri, M.M., 2013. Implementation of Predictive Data Mining Techniques for Identifying Risk Factors of Early AVF Failure in Hemodialysis Patients. Computational and Mathematical Methods in Medicine, 2013 (Article ID: 830745), 1-8.

34. Du, J., 2010. Iterative Optimization of Rule Sets, Master's Thesis. Technische Universitat Darmstadt, Fachbereich Informatik, Darmstadt, 72.

35. Witten, I.H., Frank, E., 2005. Data Mining: Practical Machine Learning Tools and Techniques. Elsevier Inc., San Francisco, USA, 525.

36. Chen, J., Li, Q., Wang, H., Deng, M., 2020. A Machine Learning Ensemble Approach Based on Random Forest and Radial Basis Function Neural Network for Risk Evaluation of Regional Flood Disaster: A Case Study of the Yangtze River Delta. China, International Journal of Environmental Research and Public Health, 17(1), 49, 1-21.

37. Kaya, I.E., Ibrikci, T., Ersoy, O.K., 2011. Prediction of Disorder with New Computational Tool: BVDEA. Expert Systems with Applications, 38, 14451-14459.

38. Carrington, A.M., Fieguth, P.W., Qazi, H., Holzinger, A., Chen, H.H., Mayr, F., Manuel, D.G., 2020. A New Concordant Partial AUC and Partial C Statistics for Imbalanced Data in the Evaluation of Machine Learning Algorithms. BMC Medical Informatics and Decision Making, 20 (4), 1-12.

39. Yang, Z.R., Thomson, R., McNeil, P., Esnouf, R.M., 2005. RONN: The Bio-Basis Function Neural Network Technique Applied to the Detection of Natively Disordered Regions in Proteins. Bioinformatics, 21, 3369–3376.

40. Alduayj, S.S., Rajpoot, K., 2018. Predicting Employee Attrition Using Machine Learning. IIT 2018: 13th International Conference on Innovations in Information Technology, Al Ain, United Arab Emirates, 93-98.

41. Bhuva, K., Srivastava, K., 2018. Comparative Study of the Machine Learning Techniques for Predicting the Employee Attrition. International Journal of Research and Analytical Reviews, 5(3), 568-577.

42. Paredes, M., 2018. A Case Study on Reducing Auto Insurance Attrition with Econometrics, Machine Learning, and A/B Testing. 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), Turin, Italy, 410-414.

43. Sukhadiya, J., Kapadia, H., D'silva, M., 2018. Employee Attrition Prediction Using Data Mining Techniques. International Journal of Management, Technology And Engineering, 8(X), 2882-2888.