Baltic J. Modern Computing, Vol. 9 (2021), No. 1, 49-66
https://doi.org/10.22364/bjmc.2021.9.1.04

Employee Attrition Estimation Using Random Forest Algorithm

Madara PRATT, Mohcine BOUDHANE, Sarma CAKULA

Vidzeme University of Applied Sciences
Cēsu iela 4, Valmiera, LV - 4201, Latvija

madara.pratt@va.lv, mohcine.boudhane@va.lv, sarma.cakula@va.lv

Abstract. Today, almost all companies are concerned about retaining their employees. However, they are not able to recognise the real factors that make them quit their jobs. Many factors could be responsible for that (for example: cultural, financial, etc.). Each company has its way to treat its employees and assure their happiness. But often no measures are taken of the satisfaction rate. As a result, in many cases, employees quit their employment suddenly without an apparent reason. In the last decades, Machine learning (ML) techniques have gained popularity among researchers. It can propose solutions to a wide range of problems. Then, ML learning has the potential to make predictions to anticipate employee attrition. In this paper, the authors compare state-of-the-art solutions for the proposed machine learning algorithms using a real data set sample size of 1469. The results could be used to warn managers in order to change their strategies or behaviour. It could also be used to make recommendations to the managers to add some policies in order to retain their employees in the company. This study aims to present a comparison of different machine learning methods to give a prediction of employees who are likely to leave their company. The data set includes information about the current employees and the employees who had already quit their job with almost 50 valuable information units. This last combines many factors: social, cultural, financial, professional, and relational factors. Six different ML algorithms were used in this paper. Experimental results show that the Random Forest algorithm demonstrated the best capabilities to predict the employees' attrition. The best prediction accuracy was 85.12, that is considered as good accuracy.

Keywords: Data Prediction and Analysis, Employee Attrition, Random Forest, Machine learning

1. Introduction

In past decades technologies have an undeniable impact and have changed every aspect of our lives. ITU (International Telecommunication Union) statistics show that by the end of 2019 93% of the world population lives within a reach of mobile Internet service (WEB (a)). Between 2005 and 2010 the number of people using the Internet is growing on average by 10% per year (WEB (a)). It offers us opportunities as well as challenges. The pandemic in 2020 showed the importance of technologies and proved them to be a very critical part of nowadays life. The technological development and connectedness have created a 24/7 work culture (Piazza, 2007). Communication within companies had become more and more technology-based, which makes it difficult for managers to motivate their employees and assure their satisfaction. Unsatisfied employees can make a decision about leaving the company. Because of the costs of employee hiring, training and acquired intellectual property it is very important to assure low attrition (employee turnover) rate within organizations. There are many factors affecting employee attrition. Satisfaction with salary is only one factor for motivating employees but there are many other aspects for employees to stay within the same company long term (Herzberg, 2003).

The possibilities provided by advanced data analytics provide us with an opportunity to learn employee behaviour effectively and process high amounts of data for more precise results. The aim of this research is to find the best algorithm for studying which people and why are more likely to leave their work and which are the strongest factors.

Within the previous research, the authors had designed an algorithm and tested four different algorithms (Pratt et al., 2020). The two best-performing algorithms were Random Forest and Logistic Regression. As the data set was rather small (n-102), the authors have improved their recent study with a higher number of respondents (n-1469) and a more extensive range of factors involved. The algorithm can be used by organizations to detect the problem areas and guide them to make better decisions and have long-lasting employees. The accuracy provided by data analytics can increase decision making efficiency.

This paper is organized as follows: in the second chapter, we will describe the theoretical background and the previous research. Then, a description of data modeling and processing will be shown in Chapter 3. The theoretical background of configuration of the model and pre-processing and correlation analysis is presented in Chapter 4. In Chapter 5 we describe the proposed model followed by application in Chapter 6. Experimental results that include the comparative study of different ML algorithms will be demonstrated in Chapter 6. Finally, the last chapter concludes the findings.

2. Theoretical background and the previous research

The reasons for employee turnover rate (attrition) are mainly related to their motivation to work and satisfaction measures (Pratt and Cakula, 2021). Employees who are satisfied will less likely decide to leave the company (Coomber and Barriball, 2007; Rusbult and Farrell, 1983). Satisfaction measures are also related to performance. More satisfied employees show higher performance measures (Whetten and Cameron, 2011). According to Herzberg (2003) satisfaction is a result of intrinsic motivational factors such as recognition, professional growth opportunities and a good feeling about the organization (Herzberg, 2003). The factors contributing to dissatisfaction avoidance include effective senior management and supervisor, satisfaction with salary and benefits and good relationships with co-workers. According to the Two-factor theory - by fulfilling extrinsic factors, employees can feel neutral, but not extra satisfied (Herzberg et al., 1959). If the needs of extrinsic factors are met, then employees can get motivated and in turns satisfied by intrinsic factors.

In previous studies turnover prediction has been predicted by using different algorithms. Recommended ones were Decision Tree, Classification and regression trees, Logistic regression, Binomial logit regression, Support Vector Machines, Random forest and Extreme Gradient Boosting (Alao and Adeyemo, 2013; Punnoose and Ajit, 2016; Sisodia et al., 2017). The reason for this many different recommendations might be behind the data set used, specifications in research aims and the volume of data available. For the current research the authors have chosen to test the performance of six algorithms - Logistic Regression, Random Forest, Gaussian NB, Decision Classifier, KNN (Euclidean distance) and Support Vector Machine. The research is an extension of the previous study performed by the authors (Pratt et al., 2020) - improved results and accuracy will be delivered by using a larger data set and more ML algorithms.

3. Data modelling process and data training

3.1. Data modelling process

Organizations can benefit from the use of Prediction Models and get highly accurate predictions based on their data (Fig. 1). The model can help to uncover the factors which contribute to the outcomes the most - has the most effect on the data set. Within this research, the Prediction Model is looking for the factors affecting attrition and get insights employee behaviour by using Prediction Explanations.

The system has two phases - Training Phase and Application Phase (Fig. 1). Expert knowledge is used for developing class labels in the Training Phase.

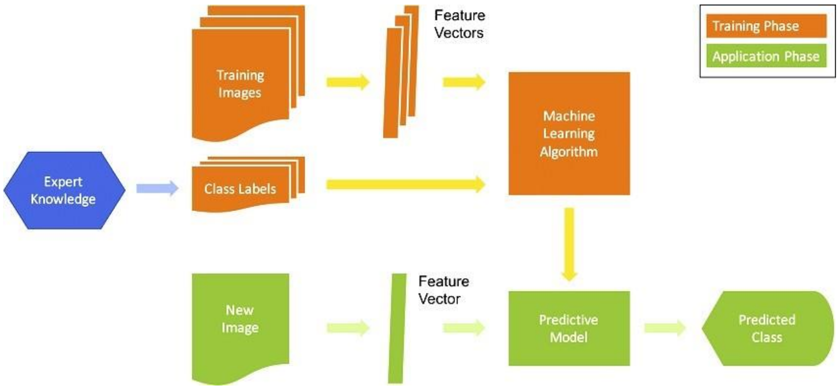


Fig. 1. The model used in the experiment.

Relationships between features are defined according to historical data. Then ML algorithms process the data. Within this phase, we use 80% of the data set. In the Application Phase the rest 20% of data is used for checking the algorithm accuracy. Predictive Model is applied for testing if these employees would leave their organization. By using this model high-impact factors can be recognized to help organizations to focus their strategies and decisions on most relevant issues.

3.2. Data set

Data set used was acquired from an open database “IBM HR Analytics Employee Attrition & Performance” (WEB (b)). The sample is 1470 with a total of 35 attributes. Attributes include several descriptive measures. The key target is ‘’Attrition’’. Measures describing employee motivational factors are pay and benefits, job involvement and training. Also, several satisfaction measures - environment, relationships and job satisfaction.

The main features (attributes) presented in the data set are Age, Attrition Business Travel, Daily Rate, Department, Distance From Home, Education, Education Field, Employee Count, Employee Number, Environment Satisfaction, Gender, Hourly Rate, Job Involvement, Job Level, Job Role, Job Satisfaction, Marital Status, Monthly Income, Monthly Rate, Number Companies Worked, Over18, Over Time, Percent Salary Hike, Performance Rating, Relationship Satisfaction, Standard Hours, Stock Option Level, Total Working Years, Training Times Last Year, Work-Life Balance, Years At Company, Years In Current Role, Years Since Last Promotion, Years With Current Manager.

3.3. Data representation

In this paragraph, some information about the used data set will be presented. Figure 2 illustrates the employees' repartition among the data set (Fig. 2).

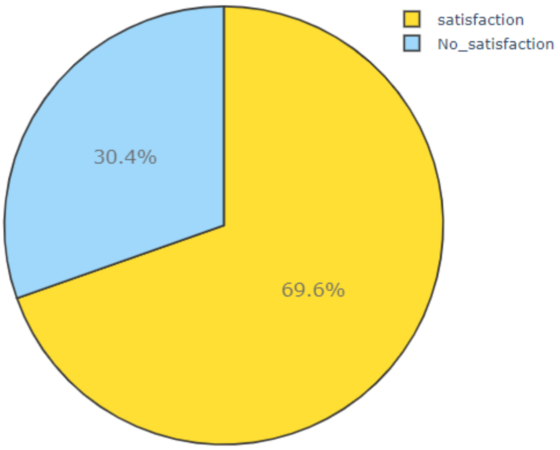


Fig. 2. Employee satisfaction – satisfied (yellow) and dissatisfied (blue).

Within the data set nearly 70% employees are satisfied and 30% of employees are dissatisfied (Fig. 2). Most of the data set (83%) represent employees who are at the time employed within the company and the second part of the data set is historical data from employees who are no longer employed within the company (16%). By having a closer look at our data, more statistics about the company can be extracted. The analysis of quantitative features give insight in general information of the data set (Fig. 3).

One of the descriptives which can give interesting insight is Age (Fig. 3). This figure presents the employees' repartition according to their age. The graph shows that most employees in the company are between 25 and 45 years old. That demonstrates that employees working within the company are rather young. On the other hand, if we compare the same attribute according to the targeted study (attrition rate) it presents another Age distribution (Fig. 4).

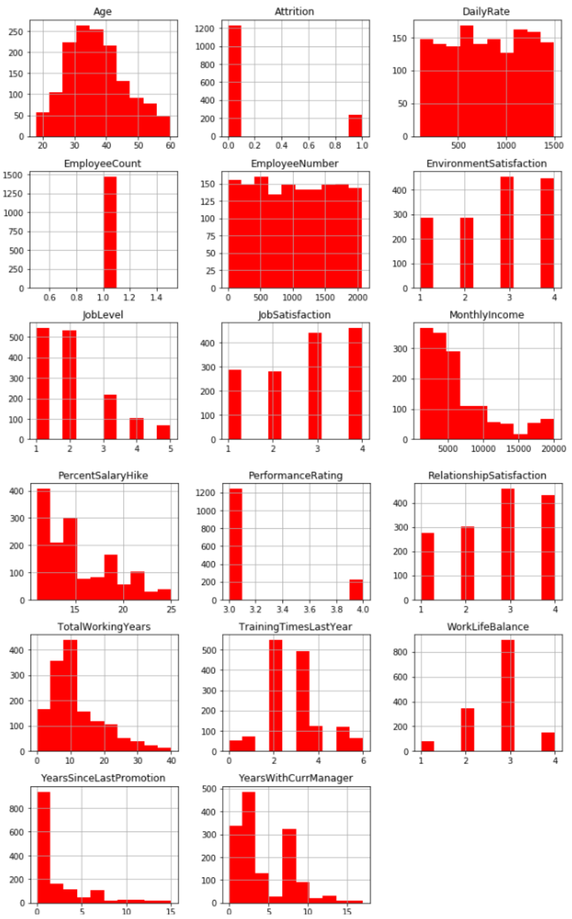


Fig. 3. The analysis of quantitative features.

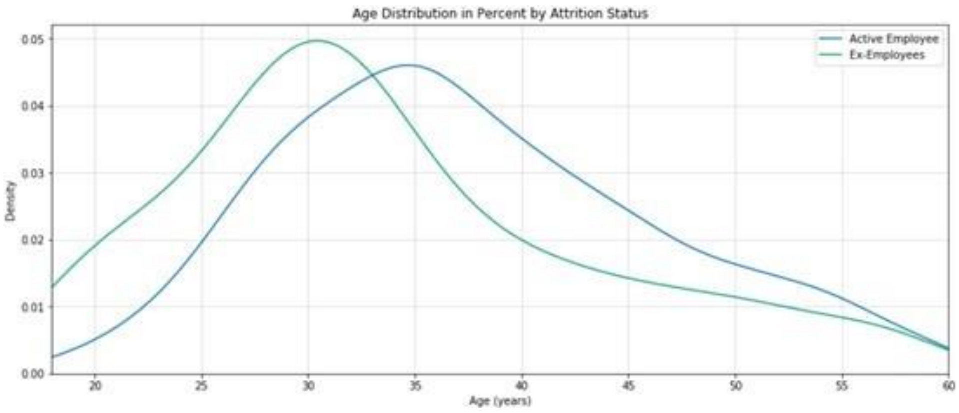


Fig. 4. Employee density according to their age. Blue: Active employees, Green: Ex-employees.

The graph shows that the employees between 25-35 are likely to quit their job (Fig. 4). However, this estimation is not able to show us what were the reasons, and which are the factors responsible for their behavior.

4. Theoretical background for configuration of the model

The objective of this work, which is the objective of developing methods by trees and random forests adapted to data from business employees. This work is also completed by a part where we compare the algorithm used with others in the literature.

Decision trees and random forests are now standard methods in supervised learning. They offer many advantages (wide applicability, ease of use, good performance, etc.) and are now commonly used in many fields, particularly in business management or to make predictions/estimates (Cook et al., 2015; Iswari et al., 2019; Parchande et al., 2019; Torizuka et al., 2018). Before presenting these methods in more detail, we define the mathematical framework within which this research fits.

Let D_n = {(X_1, Y_1),...,(X_n, Y_n)} a training sample, i.e. n independent copies of the pair of random variables (X, Y). The pair (X, Y) is independent of D_n and its law is unknown. Denote by X and Y the measurable spaces in which the random variables X and Y respectively live. In this manuscript, we consider the case X = R_d. The variable X = (X_1, ..., X_d) denotes the vector of explanatory variables and Y is the response variable.

In the context of this article, we consider the case of supervised classification where Y denotes the class with Y = {1, ..., K}, K ≥ 2, and f* is the Bayes classifier (unknown), defined on X by

$$f^* (x) = \text{argmax}_{k\in\{1,\dots,K\}} \mathbb{p}[Y = k \mid X = x] \quad (1)$$

In each context, the problem is to estimate the link between the vector X and the response variable Y, that is, to estimate the function f* from the data of the training sample D_n. An estimator of f* is a measurable function

$$\hat{f} : (X \times (X \times Y))^n \rightarrowtail Y \qquad (2)$$

which, for any new observation x, predicts the value of the response Y by $\hat{f}$ (x, D_n). In the following, we will note for convenience $\hat{f}$ (x). The function $\hat{f}$ is called a prediction rule or a decision rule. A set of reference books deals with the issue of supervised learning, see for example (Hastie et al., 2001).

In many problems in supervised learning, the explanatory variables can have a group structure. The grouping of variables can be natural or well defined in order to capture / model the relationships between the different variables. The explanatory variables can act in groups on the response variable. Thus, the exploitation of such a structure can be very useful to build a prediction rule.

In this work, we are interested in the case where the vector X is structured in *J* known groups. Each one represents a group of employees. We define the j-th group X_j, j = 1, ..., *J*, by:

$$X_j = (X_{j1}, X_{j2}, \dots, X_{jd_j}), \quad (3)$$

where the set $\{j_1, j_2, ..., j_{\mathrm{dj}}\} \subseteq \{1, ..., \mathrm{d}\}$ denotes the d_{j} indices of the explanatory variables belonging to the group j, $d_{\mathrm{j}} \leq$ d. Note that the groups are not necessarily disjoint. The objective is to use this structure to build a prediction rule $\hat{f}$.

4.1. Pre-processing and correlation analysis

In the pre-processing phase, the data is explored and cleaned. Categorical (non-numeric) values were converted to binary fields. Also, non-desirable features were deleted. After exploring the data no missing values were found.

Variables were compared and the correlation hypothesis was tested using Pearson Correlation (Fig. 5). From the analysis of correlations several hypotheses were proposed about the underlying generation of the data set.

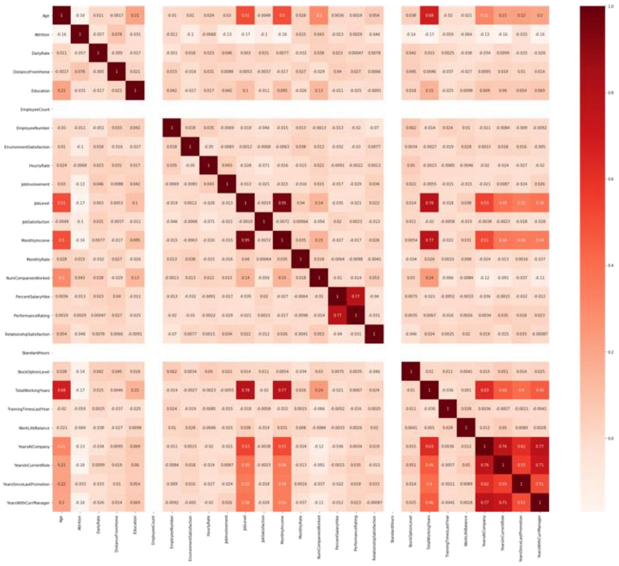


Fig. 5. Correlation analysis between all features presented in the data set.

Table 1. The highest correlation features extracted in the initial analysis

Feature	*CorrelationwithAttrition*
Job Level	-0,17
Total Working Years	-0,17
Age	-0,16
Monthly Income	-0,16
Years in Current Role	-0,16
Years with Current Manager	-0,16
Stock Option Level	-0,14
Years at Company	-0,13
Job Involvement	-0,13
Environment Satisfaction	-0,1
Job Satisfaction	-0,1

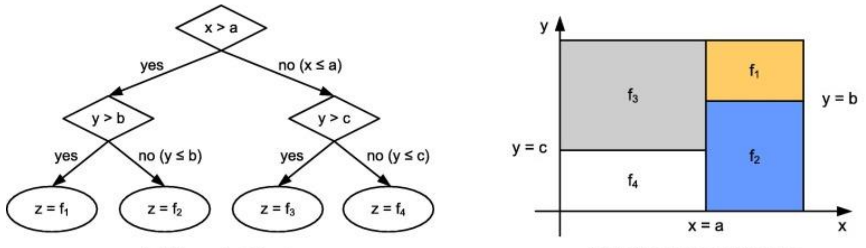
This analysis gave the initial idea of the data set. As the main focus of this study is employee attrition, the highest correlations with attrition measures within the data set were extracted (Table 1).

Initial analysis proposes that the most important factors affecting attrition are Job level, Total working years, Age, Monthly income, Years spent in current role and Years spent with the current manager. After data training and the application of the best performing ML algorithm, these results will be compared.

5. Prediction based proposed model

5.1. Classification Tree (CART)

CART is an efficient non-parametric method, simple to implement and usable in both regression and classification. The general principle of CART is to construct a prediction rule by means of recursive and binary partitioning of the data space. The resulting partition can be represented in the form of an easily interpretable binary tree. Binary decision tree and feature space partitioning illustrate the correspondence between a dyadic partition and a binary tree (Fig. 6).



(a) Binary decision tree

(b) Feature space partitioning

Fig. 6. An example of a CART tree in binary classification. Each leaf is associated with the best-represented class.

5.2. Building a CART tree

To construct a CART tree from the data of the training sample D_n, the algorithm proceeds in two steps.

Step 1: Development of a maximal tree.
This step consists of a recursive and dyadic partitioning of the X data space.

Step 2: Pruning and selection of the final tree.
The often too complex maximum tree T_{max} is generally not optimal within the meaning of a chosen performance criterion (for example in classification, classification error). An excessive number of cuts results in a tree that tends to over-adjust. To avoid this, T_{max} is pruned using the minimal cost-complexity pruning method introduced by (Breiman et al., 1984). This process consists in extracting a sequence of sub-trees $\mathrm{T}_{\mathrm{max}}$ by

minimization of the penalized criterion defined for any sub-tree T of T_{max}, noted T ≤ T_{max}, and for all α ϵ $R^{+}$by

$$Ɽ_{\alpha}(T) = Ɽ(T, D_n) + \alpha\ |\widetilde{T}|, \qquad (4)$$

where $\widetilde{T}$ denotes the number of leaves of the tree T and Ɽ(T, D_n) corresponds to the empirical error of the model T estimated from the data of the sample D_n (Breiman et al., 1984).

In resume, the algorithm is described in the following figure (Fig. 7).

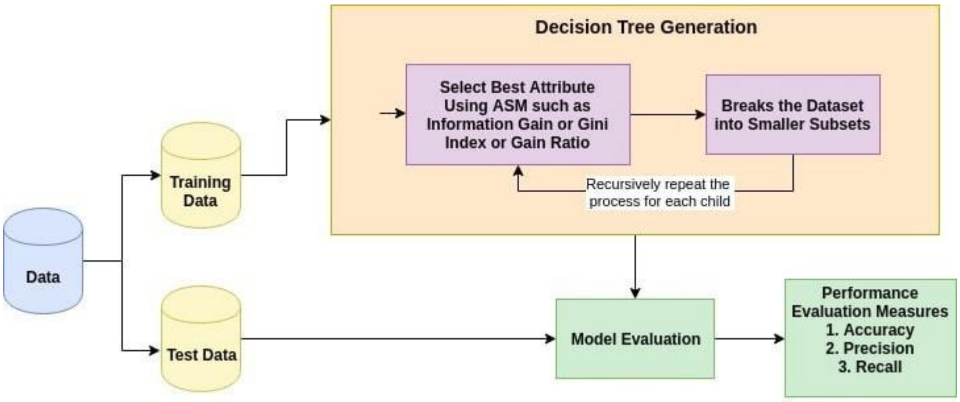

Decision Tree Generation
Select Best Attribute Using ASM such as Information Gain or Gini Index or Gain Ratio
Breaks the Dataset into Smaller Subsets
Recursively repeat the process for each child
Training Data
Data
Test Data
Model Evaluation
Performance Evaluation Measures
1. Accuracy
2. Precision
3. Recall

Fig. 7. CART algorithm process.

The data is divided into two parts – training data and test data. Training data is used for decision tree generation and the test data is used for model evaluation.

5.3. Random forest

This is an approach based on random forest algorithm that consists of aggregating a collection of estimators constructed from bootstrap samples (Fig. 8). A random forest is an aggregation of random trees.

The principle of building a forest is first of all to independently generate a large number (denoted *ntree*) of bootstrap samples $\mathcal{D}_n^1$, ..., $\mathcal{D}_n^{ntree}$ by randomly drawing, for each of them, observations (with or without replacement) in the training sample $D_{n.}$ Then, *ntree* decision trees T^1, ..., T^{ntree} are built from the bootstrap samples $\mathcal{D}_n^1$, ..., $\mathcal{D}_n^{ntree}$ and using a variant of CART. In fact, each tree is here constructed as follows. To split a node, the algorithm chooses randomly and without replacement a number mtry of explanatory variables, then it determines the best cut only according to the selected mtry variables. In addition, the constructed tree is fully developed and is not pruned. The random forest, denoted by$\{T\}_n^1$, is finally obtained by aggregating the *ntree* trees thus constructed. It defines a prediction rule which corresponds to the empirical mean of the predictions in regression and the majority vote in classification. The construction of Breiman's random forests is described by process visualization (Fig. 8).

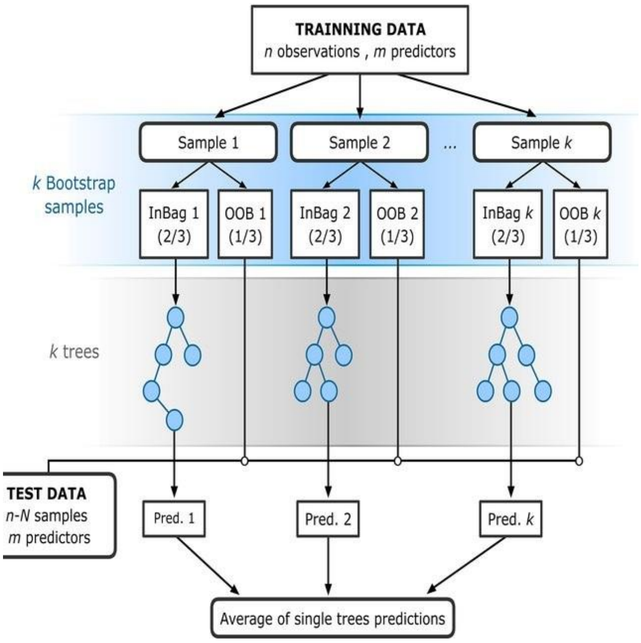


Fig. 8. The description of all process of the used random forest model.

6. Application

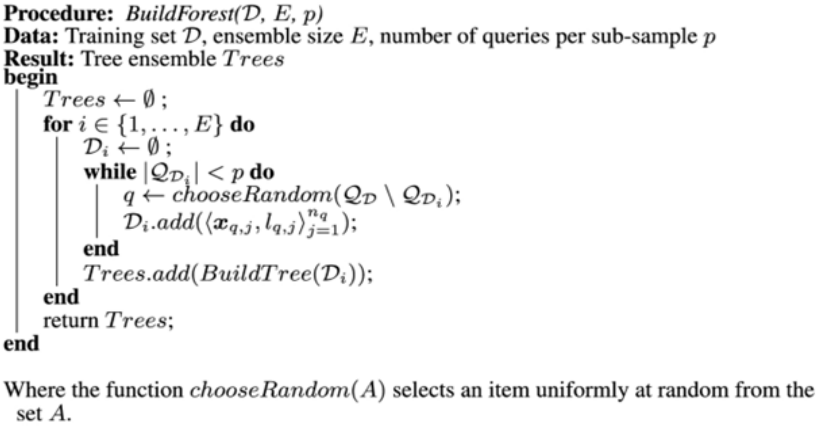
6.1. The proposed algorithm

The proposed algorithm has several parameters:

- The number of trees in the forest. Its default value is 500. Note that this parameter is not really a parameter to calibrate in the sense that a larger value of this parameter will always lead to more stable predictions than a smaller value of this parameter.
- The number *mtry* of variables chosen for the division of each node. Its default value is *mtry* = *pd* in classification. This is arguably the most important parameter to calibrate as it can greatly influence the performance of the forest.
- The minimum number of *nodesize* observations below which a node is no longer split. The default value for this parameter is *nodesize* = 1 for classification. Usually, this setting is left at its default value.
- The number of observations per year in each bootstrap sample. By default, each bootstrap sample contains an = n observations drawn with replacement in the initial sample Dn.
-

Several authors have been interested in the choice and influence of these parameters (Biau and Scornet, 2016; Genuer and Poggi, 2017). In general, the default values of the parameters work well. Indeed, there are few theoretical results available for the Breiman random forests. We can nevertheless cite a major result recently established, focusing on the convergence of random forests in the additive model (Scornet et al., 2015).

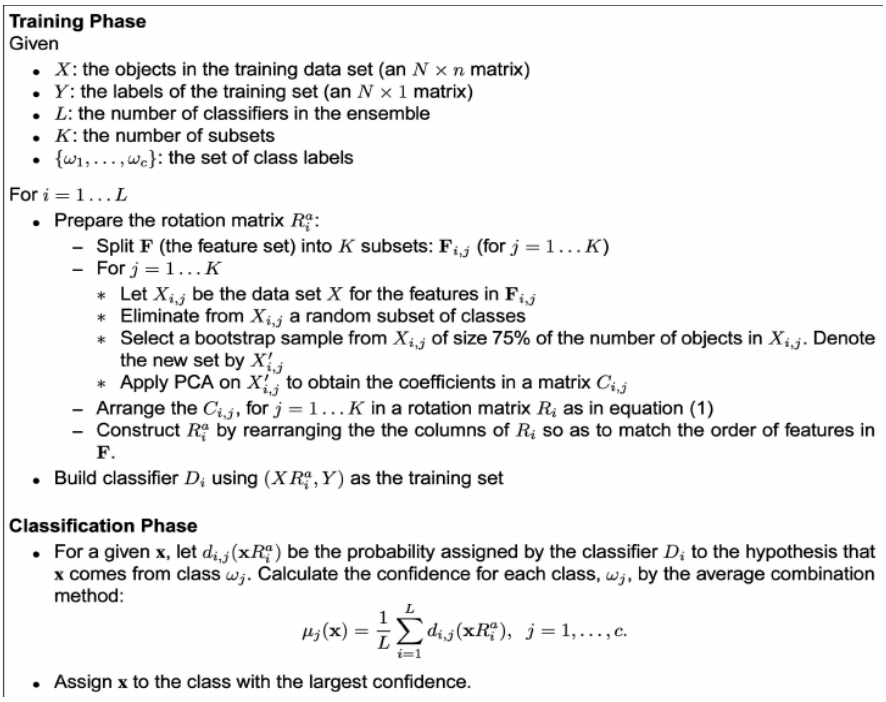
Theoretical guarantees have also been obtained for simplified versions of the method (Wager, 2014). A summary of the main theoretical results is illustrated in the first algorithm before it is updated (Fig. 9).



```
Procedure: BuildForest($\mathcal{D}$, $E$, $p$)
Data: Training set $\mathcal{D}$, ensemble size $E$, number of queries per sub-sample $p$
Result: Tree ensemble $Trees$
begin
    $Trees \leftarrow \emptyset$ ;
    for $i \in \{1, \ldots, E\}$ do
        $\mathcal{D}_i \leftarrow \emptyset$ ;
        while $|\mathcal{Q}_{\mathcal{D}_i}| < p$ do
            $q \leftarrow chooseRandom(\mathcal{Q}_{\mathcal{D}} \setminus \mathcal{Q}_{\mathcal{D}_i})$;
            $\mathcal{D}_i.add(\langle \boldsymbol{x}_{q,j}, l_{q,j} \rangle_{j=1}^{n_q})$;
        end
        $Trees.add(BuildTree(\mathcal{D}_i))$;
    end
    return $Trees$;
end
```

Where the function $chooseRandom(A)$ selects an item uniformly at random from the set A.

Fig. 9. The used model based on random forest algorithm.



Training Phase

Given

- X: the objects in the training data set (an $N \times n$ matrix)
- Y: the labels of the training set (an $N \times 1$ matrix)
- L: the number of classifiers in the ensemble
- K: the number of subsets
- $\{\omega_1, \ldots, \omega_c\}$: the set of class labels

For $i = 1 \ldots L$

- Prepare the rotation matrix R_i^a:
  - Split $\mathbf{F}$ (the feature set) into K subsets: $\mathbf{F}_{i,j}$ (for $j = 1 \ldots K$)
  - For $j = 1 \ldots K$
    - $*$ Let $X_{i,j}$ be the data set X for the features in $\mathbf{F}_{i,j}$
    - $*$ Eliminate from $X_{i,j}$ a random subset of classes
    - $*$ Select a bootstrap sample from $X_{i,j}$ of size 75% of the number of objects in $X_{i,j}$. Denote the new set by $X'_{i,j}$
    - $*$ Apply PCA on $X'_{i,j}$ to obtain the coefficients in a matrix $C_{i,j}$
  - Arrange the $C_{i,j}$, for $j = 1 \ldots K$ in a rotation matrix R_i as in equation (1)
  - Construct R_i^a by rearranging the the columns of R_i so as to match the order of features in $\mathbf{F}$.
- Build classifier D_i using (XR_i^a, Y) as the training set

Classification Phase

- For a given $\mathbf{x}$, let $d_{i,j}(\mathbf{x}R_i^a)$ be the probability assigned by the classifier D_i to the hypothesis that $\mathbf{x}$ comes from class ω_j. Calculate the confidence for each class, ω_j, by the average combination method:

$$\mu_j(\mathbf{x}) = \frac{1}{L}\sum_{i=1}^{L} d_{i,j}(\mathbf{x}R_i^a), \quad j = 1, \ldots, c.$$

- Assign $\mathbf{x}$ to the class with the largest confidence.

Fig. 10. The proposed algorithm.

6.2. Model configuration

The model is using two phases: training phase and classification phase (testing phase).

Forward the proposed approach of updated algorithm (Fig. 9) has been illustrated in Figure 10.

The model validation technique used for this particular data set was splitting the data set 80:20. 80% of data was used for training and 20% was held out. After validating the data and training the models' optimal configuration these models were tested on 20% holdout sample. Computing configuration to train this model consisted of 8GB RAM and a Intel core i5-8250U CPU 1.80 GHz CPU.

Importance of the parameters:

Random forests are generally more efficient than simple decision trees but have the disadvantage of being more difficult to interpret. In order to overcome this, several indices of the importance of the variables are defined. These scores make it possible to establish a hierarchy of explanatory variables based on the importance in relation to answer Y. The method of random forests mainly offers two criteria: the importance of Gini and the importance by permutation.

The Gini importance rating approximates the importance score proposed in CART, except that it does not use substitution cuts. This index is defined from the impurity criterion (4) used during the construction of a tree. The importance of a variable is first assessed on each tree in the forest. Thus, for a given tree, it corresponds to the overall reduction in impurity, that is to say the weighted sum of the reductions in impurity induced when the variable is used to cut a node from said tree. The Gini importance of a variable is then defined by the average (over all trees in the forest) of the overall impurity reductions. In other words, the permutation importance index is based on the idea that an explanatory variable can be considered important in predicting the Y response if breaking the link between this variable and the Y response deteriorates the quality of the prediction (Breiman, 2001). In this sense, random permutations of the values of the variable are used to mimic the breaking of this link. Formally, the calculation of the measure of importance by permutation for a variable X_j (with j = 1, ..., d) consists first of all in defining the out-of-bag (OOB) sample associated with each sample bootstrap.

These steps are repeated on all the trees in the forest. The importance index then corresponds to the average over all trees of the increase in error:

$$\mathcal{I}_{\text{perm}}(X_j\,,\,\{T^b\}_1^{\texttt{ntree}}) \;=\; \frac{1}{\texttt{ntree}}\sum_{b=1}^{\texttt{ntree}} \mathcal{R}(T^b,\bar{\mathcal{D}}_n^b) - \mathcal{R}(T^b,\bar{\mathcal{D}}_n^{bj}), \qquad (5)$$

If the random permutation of the j-th variable induces a large increase in the error then $\mathcal{J}_{\text{perm}}$ $(X_j, \{T^b\}_n^1)$ is large and the variable is considered important. Conversely, if the perturbations do not affect the error, then the permutation importance index of X_j is close to zero and the variable is considered unimportant in predicting the response Y.

6.3. Prediction accuracy of the proposed method

The Random Forest confusion matrix shows results for predicted and actual values (Fig. 11).

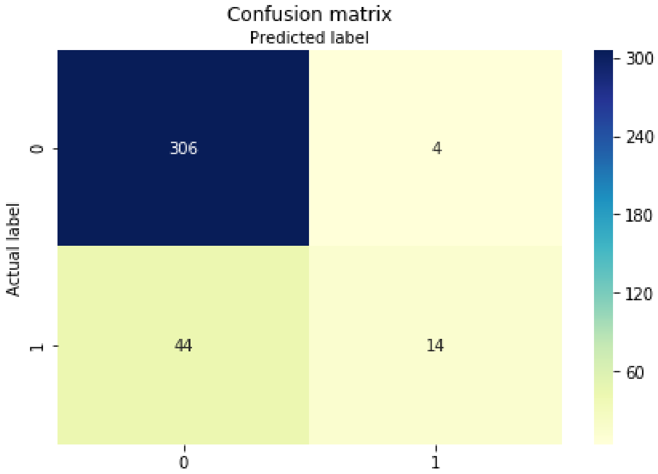


Fig. 11. Confusion matrix (Random forest).

True Positive (right lower corner) – employees (14) who left the company and were correctly identified by ML algorithm. True Negative (upper left corner) – the number (306) of employees who did not leave the company and were correctly determined by the algorithm. False Negative (lower-left corner) are the employees (44) that have left the company, although the algorithm predicted they would stay. And finally, the False Positive (upper right corner) are employees (4) who remained in the company, but the algorithm said they would leave.

6.4. Comparison of the results

For data training, the authors are compared their results with five different statistical modelling approaches. The proposed ML algorithms are:

Logistic Regression – one of the most commonly used multivariate analysis models. This predictive technique aims to build a model allowing to explain the values taken by a qualitative target variable from a set of quantitative or qualitative explanatory variables (Gutiérrez et al., 2010).

Gaussian NB (Naive Bayes Classifier) –a simple and well-performing classification technique. This technique doesn’t require a large data sample. It performs classification based on probabilities with a condition (assumption) that these variables are conditionally independent of each other (Punnoose and Ajit, 2016).

Decision Tree Classifier – a supervised ML technique used for classifying instances by sorting them based on feature values (Alao and Adeyemo, 2013).

KNN (Euclidean distance) – this is a superwised learning method. The learning database is made up of N “input-output” pairs. To estimate output associated with a new input ‘x’, the KNN method takes into account the ‘k’ training samples whose training sample is closest to the new input ‘x’ (Lu, Tong, and Chen, 2015).

Support Vector Machine (SVM) – this generalization of linear classifiers can be applied to a large number of fields. These large margin separators are a set of supervised learning techniques intended to solve discrimination and regression problems. Previous research show, the performance of support vector machines is in the same order, or even better, than that of a neural network or a Gaussian mixture model (Hastie et al., 2001).

The performance of the algorithms was analysed and compared (Table 2). Based on AUC score the Logistic Regression (LR) and Random Forest (RF) has the highest scores.

Table 2. Performance comparison between all ML algorithms used in this study

$Algorithm$	$ROCAUCMean$	$ROCAUCSTD$	$AccuracyMean$	$AccuracySTD$
Gaussian NB	77.84	4.16	76.42	3.63
KNN	59.94	7.39	80.67	4.16
Decision Tree	60.20	7.40	75.51	4.21
SVM	50.00	0.00	83.76	2.71
Logistic regression	**80.85**	4.79	75.42	5.11
The proposed	80.84	**4.70**	**85.12**	**2.70**

When algorithm accuracy mean score is compared, the proposed model, SVM and KNN have the highest scores (Table 2). Box plot of processed data visualizes medians, minimum and maximum values, interquartile range and outliers (the points which are going beyond the upper and lower boundaries) (Fig. 12). The same as in Table 2, in the box plot, we can see that Random Forest, SVM and KNN algorithms have the highest medians and therefore algorithm performance.

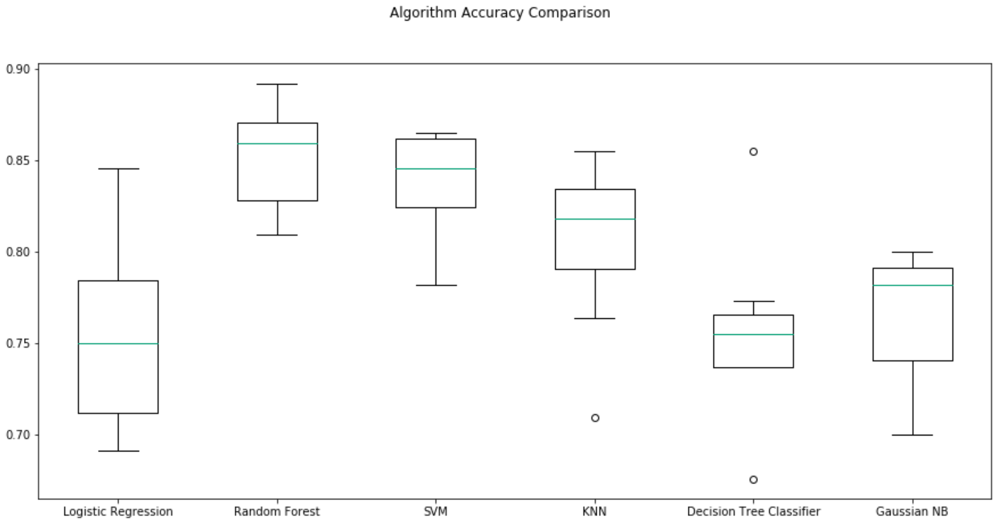


Fig. 12. Box plot for algorithm accuracy comparison of different ML algorithms.

The proposed algorithm correctly classified 320 employees and falsely 48 employees. The accuracy this algorithm is 85,12%. These results are compliant with the authors' previous research results, where Random Forest was the most effective algorithm for the given data set with an accuracy of 95,24% following with Logistic Regression with the score of 80,95% (Pratt et al., 2020). That is due to the application of Iperm (eq), that are able to compute perturbations that affect generate errors in predictions. It helps more to select and reject unimportant features in the model.

7. Discussion

The Random Forest is the most efficient algorithm for the data set and we have used this ML algorithm to detect the most important features responsible for employee attrition (turnover) rate (Fig. 13).

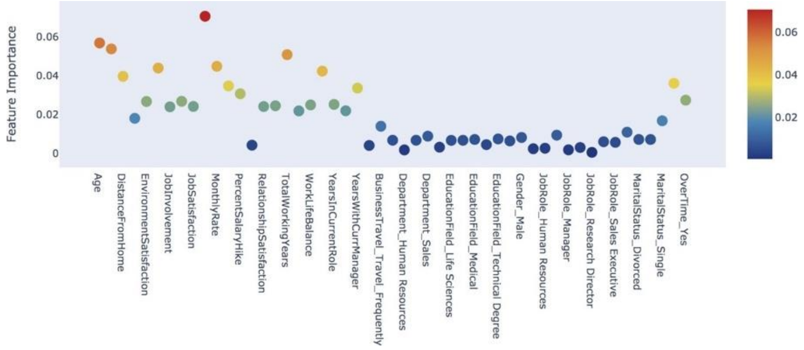


Fig. 13. Feature importance estimated by Random forest algorithm.

According to the proposed algorithm, the most important features are Monthly Income, Age, Daily Rate, Total Working Years and Monthly Rate. Only five of these top 10 features were recognized as high importance within Pearson Correlation results (Table 1). These common factors are Age, Monthly Income, Total Work Years, Years at the Company and Years with Current Manager. It shows the importance and efficiency of ML Algorithm in predicting the attrition and having more precise results. The value calculated by the proposed algorithm is displayed in Table 3.

Table 3. Performance comparison between all ML algorithms used in this study

nr.	*Features*	*CorrelationwithAttrition*
0	Monthly Income	0.070741
1	Age	0.056962
2	Daily Rate	0.053886
3	Total Working Years	0.050907
4	Monthly Rate	0.044915
5	Hourly Rate	0.044034
6	Years At Company	0.042443
7	Distance From Home	0.039759
8	Number Companies Worked	0.034798
9	Years With Current Manager	0.033677

The most significant coefficients are Monthly Income (0,07), Age (0,056), Daily Rate (0,053) and Total Working Years (0,05). When compared to the previous research conducted by the authors, these results differ in emphasizing the importance of salary (Income, Daily/Monthly Rate) and age (Pratt et al., 2020). These were not factors with high correlation; rather, factors of importance were all related to motivation and behaviour of the management. The higher number of respondents have increased the accuracy of the algorithm, and the broader range of featured included have added more insight into the problem of attrition. With using five algorithms and more substantial data set, the proposed model repeatedly is proven the best fit for calculating employee attrition.

8. Conclusion

The word is changing, and so is our work setting. The development of technologies gives us an opportunity to improve our data analysis and aim for more precise predictions. Motivating and satisfying employees in the new work environment is challenging, and so is keeping attrition rate low. Having loyal and long-lasting employees are essential to most of the organizations. In this paper, the authors have presented the experiment of finding the best-fit algorithm for employee behaviour measures within the given context.

The study is based on the author's previous research and is improved with a higher amount of data. It supplements the author's previous research and conclusion that the proposed model is the most appropriate algorithm for data set describing employee satisfaction and attrition with an accuracy of 85,12%.

After data training and the use of the algorithms, the results were improved, and several different features were highlighted. The most important features are Monthly Income, Age, Daily Rate, Total Working Years and Monthly Rate. These results are different from previous research analysis which didn't suggest age or salary of being high importance factors, rather motivational factors and management being the highest importance. These new results are a valuable addition to existing data and will be researched more in detail in future.

Acknowledgement

This research has been supported by a grant from the European Regional Development Fund project “The Study of Computer Vision Algorithms for Underwater Fish Inspection” No. 1.1.1.2/VIAA/2/18/348 within the Activity 1.1.1.2 “Postdoctoral Research Aid”.

References

Alao, D., Adeyemo, A. B. (2013). Analyzing employee attrition using decision tree algorithms. *Computing, Information Systems, Development Informatics and Allied Research Journal*, ***4***(1), 17–28.

Biau, G., Scornet, E. (2016). A random forest guided tour. *Test*, ***25***(2), 197–227.

Breiman, L. (2001). Random forests. *Machine Learning*, ***45***(1), 5–32.

Breiman, L., Friedman, J., Stone, C. J., Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Cook, A., Wu, P., Mengersen, K. (2015). Machine learning and visual analytics for consulting business decision support. *2015 Big Data Visual Analytics (BDVA)*, 1–2. IEEE.

Coomber, B., Louise Barriball, K. (2007). Impact of job satisfaction components on intent to leave and turnover for hospital-based nurses: A review of the research literature. *International Journal of Nursing Studies*, ***44***(2), 297–314. https://doi.org/10.1016/j.ijnurstu.2006.02.004

Genuer, R., Poggi, J.-M. (2017). *Arbres CART et Forêts aléatoires, Importance et sélection de variables*. HAL archieves, HAL Id: hal-01387654.

Gutiérrez, P. A., Hervás-Martínez, C., Martínez-Estudillo, F. J. (2010). Logistic regression by means of evolutionary radial basis function neural networks. *IEEE Transactions on Neural Networks*, ***22***(2), 246–263.

Hastie, T., Tibshirani, R., Friedman, J. (2001). The elements of statistical learning: Springer series in statistics. Springer, New York, NY.

Herzberg, F., Mausner, B., Snyderman, B. (1959). The motivation to work. 2. Ed. New York : John Wiley & Sons.

Herzberg, F.. (2003). One More Time: How Do You Motivate Employees? *Harvard Business Review*, ***81***(1), 87–96.

Iswari, N. M. S., Budiardjo, E. K., Santoso, H. B., Hasibuan, Z. A. (2019). E-Business Application Recommendation for SMEs based on Organization Profile using Random Forest Classification. *2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, 522–527.

Lu, S., Tong, W., Chen, Z. (2015). Implementation of the KNN algorithm based on Hadoop. *Proceedings of 2015 International Conference on Smart and Sustainable City and Big Data (ICSSC).* 123-126. DOI: 10.1049/cp.2015.0265

Parchande, S., Shahane, A., Dhore, M. (2019). Contractual Employee Management System Using Machine Learning and Robotic Process Automation. *2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA)*, 1–5.

Piazza, C. F. (2007). 24/7 workplace connectivity: A hidden ethical dilemma. *Business and Organizational Ethics Partnership, Markkula Center for Applied Ethics, Santa Clara University, Santa Clara, CA,* ***21****, 1-16.*

Pratt, M., Boudhane, M., Cakula, S. (2020). Predictive Data Analysis Model for Employee Satisfaction Using ML Algorithms. . In Advances on Smart and Soft Computing (pp. 143-152). Springer, Singapore. DOI:10.1007/978-981-15-6048-4_13

Pratt, M., Cakula, S. (2021). Motivation in a Business Company Using Technology-Based Communication. In *Artificial Intelligence in Industry 4.0. Studies in Computational Intelligence*: ***928*** (pp. 15–30). DOI:10.1007/978-3-030-61045-6_2

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. DOI:10.14569/ijarai.2016.050904

Rusbult, C. E., Farrell, D. (1983). A longitudinal test of the investment model: The impact on job satisfaction, job commitment, and turnover of variations in rewards, costs, alternatives, and investments. *Journal of Applied Psychology*, ***68***(3), 429–438. DOI:10.1037/0021-9010.68.3.429

Scornet, E., Biau, G., Vert, J.-P. (2015). Consistency of random forests. *The Annals of Statistics*, ***43***(4), 1716–1741.

Sisodia, D. S., Vishwakarma, S., Pujahari, A. (2017). Evaluation of machine learning models for employee churn prediction. *2017 International Conference on Inventive Computing and Informatics (ICICI)*, 1016–1020. DOI:10.1109/ICICI.2017.8365293

Torizuka, K., Oi, H., Saitoh, F., Ishizu, S. (2018). Benefit segmentation of online customer reviews using random forest. *2018 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, 487–491. IEEE.

Wager, S. (2014). Asymptotic theory for random forests. *ArXiv Preprint ArXiv:1405.0352*.

WEB (a) Measuring digital development. Facts and Figures 2019. https://www.itu.int/en/mediacentre/Documents/MediaRelations/ITU%20Facts%20and%20Figures%202019%20-%20Embargoed%205%20November%201200%20CET.pdf

WEB (b) IBM HR Analytics Employee Attrition & Performance. https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset

Whetten, D. A., Cameron, K. S. (2011). Developing management skills. 8. Ed, Peason, Prantice Hall.

Received January 15, 2021, revised March 12, 2021, accepted March 13, 2021