

Identification of human resource analytics using machine learning algorithms

Elham Mohammed Thabit A. Alsaadi¹, Sameerah Faris Khlebus², Ashwak Alabaichi³

¹Department of Information Technology, College of Computer Science and Information Technology, University of Kerbala, Karbala, Iraq

²Department of Business Information Technology, College of Business Administration of Informatics, University of Information Technology and Communications, Baghdad, Iraq

³Department of Biomedical Engineering, College of Engineering, University of Kerbala, Karbala, Iraq

Article Info

Article history:

Received Sep 23, 2021

Revised Jul 05, 2022

Accepted Jul 13, 2022

Keywords:

Decision tree classifier

Logistic regression

Machine learning

Random forest

ABSTRACT

Employee attrition is one of the most significant business issues in human resource (HR) analytics. This research aims to identify the most critical elements that contribute to employee attrition. Businesses operate heavily on employee training in order to maximize the returns they will offer to the company in the future. By utilizing the employee information value concept, it has been discovered that employee features such as overtime, the total number of projects and job level have a significant impact on attrition. To find the probability of new employee attrition, various classification algorithms such as decision trees (DT) classifier, logistic regression (LR), random forests (RF), and K-means clustering are used. A comparative analysis of the models with different rating scales is carried out for the highest accuracy. For prediction, four diverse machine learning (ML) algorithms such as LR, RF, DT classifier, and k-nearest neighbors (k-NN) are used. DT classifier outperforms with 97% of accuracy than other techniques. The effects of predictive ML techniques on the employee dataset show that RF evaluation outperforms other ML techniques followed by model of LR for the specific dataset if precision is the preferred metric. Identification of HR is forecasted using ML algorithms on employee data.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Elham Mohammed Thabit A. Alsaadi

Department of Information Technology, College of Computer Science and Information Technology

University of Kerbala, Karbala, Iraq

Email: elham.thabit@uokerbala.edu.iq

1. INTRODUCTION

Human resource (HR) is the organization of a company that is in charge of analysis, recruiting, monitoring and training job candidates, and managing employee benefit programs [1]. It refers to both the people who are working for a company or an organization and the division in charge of dealing with all employee related issues [2], [3]. Employees are among the most important resources in any organization or company. HR are critical in enabling businesses to engage with a business in a socially responsible business atmosphere and a higher improvement in the quality workers in the twenty-first decade [4], [5].

Employee churn is the most serious issue that a company faces, and it has a wide-ranging impact on how the company operates. In this age of intense competition, there are many variables such as salary, working conditions, and so on that lead to employee dissatisfaction [6]. Long working hours, family pressure, work experience, job role, distance traveled, office building, office conveniences, perks, and a variety of other factors may all play a role in employee attrition. It is critical for HR department for recognizing all of

these variables in order to improve employee fulfilment. As the quality has a direct impact on the workers' employee's productivity. Sometimes the employee has no problems in the company, but other companies may offer an improved profile with a better pay package. As a result, the employee may be willing to resign. Retaining a single employee necessitates extensive knowledge in a variety of areas. In this study, we attempt to identify significant contributors to employee attrition [7]. The primary motivation is to design employee turnover and why an employee is leaving the company, which will be beneficial to the company in the future. Until the employee submits his resignation, HR department should devise a strategy [8], [9]. As the primary course of HR management, employee turnover could provide administrators with decision support. In this research, we inability to discover employee turnover using a variety of continuous variables. Employee and job status classes are determined as two factors, with variable value varying based on the estate's nation. We should offer a dataset of employee information to show how these models are work in practice. Lastly, it is demonstrated that the approach proposed in this study is extremely useful to managers when developing succession methods, supporting guidelines, and retirement legislation.

This research comprises a comparative study of different algorithms utilizing model measurement measures such as accuracy, reminder, FN rate, F-measuring, in contrast to the majority of current focusing on a single company-specific algorithm. The reason for using several algorithms is that the special benefits and pitfalls of each algorithm. For example, if the rate of events is low, the logistic model fails, while random forests surpass. When the total attributes contain a large number of important attributes. Decision tree executes linear models when the data is highly non-linear. Through comparative study all the advantages and falls that allow companies to choose the best model for their data would be taken into account [10]. In this paper, a strategy dependent on artificial intelligence (AI) model that gives us the knowledge of worker's turnover of an organization by recognizing its significant variables has been suggested. The referenced model is an all-encompassing way to deal with pick the most noteworthy weighted highlights in two stages. Our primary objective is to diminish the quantity of highlights and take out the most noticeable ones from the component determination process [11].

In circumstances where a dataset requires enormous space for usefulness and might be liable to over fitting. There are two methods to follow so as to gain a reasonable perception and to abstain from overfitting: one is the system of dimensionality decrease and the other is the choice of highlights. We utilized the element choice methodology, in any case, to diminish the quantity of highlights while choosing the most noticeable highlights. Likewise, for most estimations, the precision of the lessened number of specific features and the accuracy of complete features are almost the proportional, prescribing that not all features are on a very basic level critical. Again, with the inconsequential picked features, only two or three figurings like support vector machine (SVM) give insignificantly high precision. We utilized AI situations like scikit-learn [12], Pandas and Matplotlib. The turnover degree is characterized as the association's enlisting and end necessities. For various reasons, a representative may leave the activity. Here the turnover and attrition are the terms of business that are consistently in struggle. In an association, there are various types of turnover. The decrease in the quantity of representatives is commonly known as wearing down. These phrasings can be utilized conversely to examine the information on labor and different advances required for labor planning. This happens when an agent leaves the association trimming down similarly as turnover. Agent trimming down can be depicted as laborer hardship for any of the going with reasons: singular reasons, low work satisfaction, low wages and poor business conditions. The turnover of workers can be isolated into two classes: deliberate and programmed trimming down. The programmed end occurs for several reasons, for example, poor implementation of delegates, business necessities, or when their supervisor terminates the workers [13]. Then again, in willful wearing down, high-performing laborers choose to leave the association, regardless of an exertion by the organization to keep them. Early retirement or work offers from various associations, for example, can achieve stubborn debilitating. While associations that comprehend the noteworthiness of their delegates generally put assets into their workforce by giving liberal getting ready and an unprecedented work environment, they are moreover encountering purposeful wearing out and the loss of proficient authorities. Another issue, the contracting of substitutions, powers critical costs on the client, including the costs of enrolling, securing and getting ready [14]. Anticipating the steady loss of laborers at an association will assist the executives with reacting all the more adequately by reinforcing their inner approaches and techniques. Where skilled specialists with a possibility of leaving can be given different answers for limit their likelihood of stopping, for example, a compensation increment or appropriate preparing. Utilizing models of AI can assist organizations with foreseeing the turnover of workers. Investigators can create and prepare an AI model utilizing chronicled information held in HR divisions that can anticipate workers leaving the organization [15]. These models are set up to dismember the relationship among dynamic and terminated laborer characteristics. This was developed using comprehensive retailer HRIS data by pushing the topic of trimming as a a gathering limit and proving it with stewardship frameworks. This is by distinguishing the XGBoost classifier's predominant precision and different frameworks and identifying the reasoning behind their unique offer [16], [17]. There are several useful algorithms, both qualitative and quantitative,

for anticipating staff turnover in order to stabilise a company. However, these methodologies have a number of drawbacks when it comes to predicting staff turnover, including the following:

- 1) Inadequate consideration of data that is uneven. The data on employee turnover included in these studies only represents a small part of the total workforce.
- 2) Inefficient data processing due to high dimensionality. Employees have a variety of feature dimensions, including both static and dynamic ones.
- 3) There is no rating. Because the purpose of the prediction is to reduce employee turnover, we must identify and prioritise its aspects.

The objective of this work is to target the model to recognize the individuals who will leave, with the goal that the organization can intercede, act and forecast the correct fitment for yearning worker, predict whittling down particularly among superior workers and predict how pay esteems will work out [18], [19].

The HR department can utilise the results of our model to design strategy before the employee introduces his resignation. differtn the majority of current research work, which they focus on a single business problem-solving algorithm, this paper compares several algorithms using model evaluation measures including as accuracy, precision and sensitivity. The reason for adopting verouis algorithms is that each has its own set of benefits and drawbacks.

Data preprocessing, feature selection and measurement, the modeling utilising various techniques, and finally evaluation of models using model evaluation metrics are all part of the analytics project as shown in Figure 1. Following evaluation, the best model is used to generate predictions on the data. A raw data set is used for data pre-processing, followed by feature selection and scaling, model building after that, evaluation and model tuning are performed, and finally deployment and monitoring step will be performed.

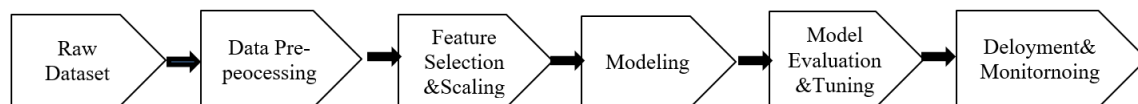


Figure 1. Analytic prosses

– Meaning of the job satisfaction

People bring their mental and physical skills to their work and time. Most people by working attempt to make a diversity in their lives and in others' lives. A pay check is not the only explanation why an individual wants a job. Jobs can be worked to accomplish personal goals. When a job meets the expectations of an individual, he or she also experiences positive emotions. Such optimistic thoughts are job satisfaction.

There are many different ways of defining job satisfaction or employee satisfaction. Many say it's just how satisfied a person is with his or her employment, that is, whether they like the job or particular aspects or facets of jobs, such as the nature of work or supervision. Others think it isn't that easy and require multidimensional psychological reactions to one's job instead. Studies have noted that tests of job satisfaction vary in how much they measure job-related emotions (emotional job satisfaction) or job regard cognitions (cognitive job satisfaction). Often job satisfaction is measured through how well outcomes match, or exceed expectations. It reflects many behaviors linked to that.

2. LITERATURE SURVAY

This section presents some of different studies in the field of HR analytics. As a part of this research work, different reviews are studied to understand the background of the HR system, different approaches. In this section also study the most critical elements that contribute to employee attrition. The literature review provides detailed information on the topic under consideration in this research work along with the gaps in literature.

Zhao *et al.* [20] analyzed HR with supervised methods of machine learning, demonstrated and analyzed within an enterprise for the estimation of employee turnover. Machine learning (ML) is able to identify the factor of human resource. In this analysis, computational tests are conducted with a decision tree system, a random forest technique, a gradient boosting trees technique, an extreme gradient boosting technique, a logistic regression technique, for actual and virtual human resource datasets representing organizations of small, medium, and large employee populations.

Chourey *et al.* [21] analyzed that human resource attrition is nowadays important in the industry. It is the big problem that stands out in all the organizations. Attrition is the steady decline in the numbers of employees by way of retirement, resignation or death. ML technique can predict dataset vary effectively.

Talent retention and professional employee retention is a crucial dilemma for HR manager particularly in the manufacturing industry. This research identified certain complex factors that are main responsible for the turnover of employees in selected organizations. The most appealing environment to make employees stay back in company is the ethical work culture, cordial employee relationship and the execution of organizational policies.

Gao *et al.* [22] focused on human resource analyze the performance of employee turnover. HR analyze the major issue for many companies and businesses. The issue is important because it affects not only the viability of the job but also the quality of the preparation and culture of the enterprise. ML methods analyze the employee performance. This research aims to improve the ability to predict employee turnover, and present a new approach based on algorithm of an improved random forest to tackle this need. Studying offers a new empirical approach which can allow human resource departments to more accurately predict employee turnover and its experimental results afford valuable tools to minimize employee turnover.

Karande *et al.* [23] suggested that human resource employee turnover in different sectors is now becoming a big problem. This study focused on identifying key characteristics of volunteer employee turnover and how they can be resolved in advance. The question is to figure out whether an employee is going to leave or stay. ML methods analyze the employee performance. Instead of focusing on algorithm of a single classifier, the suggested work will utilize ensemble learning to resolve the problem, combining weak learning algorithms for getting a stronger ensemble model. When the need for scalability and high availability without sacrificing performance, the Apache Cassandra database is the way to go. Cassandra's support for replicating across various datacenters is is best in class, giving your users lower latency and the piece of mind that comes with knowing you can endure regional disruptions. Ensemble model is also used in this research.

Alghamlar and Alabduljabbar [24], a clear picture of the state of the art, describing each typical phase of the data mining process, the variations and similarities across this research, and making additional recommendations as a result, this survey presents a detailed path for using data mining to improve employee attrition. This by using algorithms of ML for evolving a web-based application which assist in predicting the suitability of IT students' skills for the recruitment. The result displays that some hard skills such as (Linux systems, image and video processing, and some programming languages are below average. On the other hand, all soft skills are above average. In addition, the rate of 82% of respondents do not have any IT certification.

Kowsher *et al.* [25] to create a linear function, two shifting vectors dubbed support direction vector (SDV) and support origin vector (SOV) were employed. With both the target class data and the non target class data, these vectors construct a linear function for measuring cosine-angle. When considering target data points, the linear function positions itself in such a way that its angle with target class data is minimised and its angle with non-target class data is maximised. The linear function's positional error has been characterised as a loss function that has been iteratively refined with the gradient descent algorithm. Three different standard datasets have used to demonstrate the acceptability of this strategy. The model's accuracy was comparable to that of a normal supervised classification algorithm.

3. METHODOLOGY

This section provides the theoretical and technical walk-through of the research method used to construct an analytical analysis and prediction of employee data set using python and ML forecast employee turnover and how to identify the successful employee, thereby saving the company's HRM budget on hiring new employees as shown in Figure 2. This chapter explains the collection of methods used to perform the investigation and estimate the accuracy and evaluates the data set. It also includes the techniques used to collect data and interpret data.

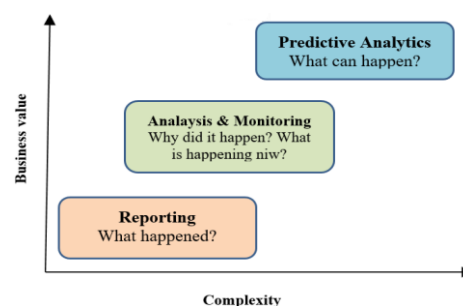


Figure 2. Stages of analytics

Figure 3 displays set of data, the work architecture, recognizes and selects features with variable significance. So many features are performed independently with multiple algorithms previously discussed in order to achieve the individual performance. Then weights are designated for each model to model the learning model of the ensemble and provided to the model of the ensemble. The classifier is categorized, as well as the accuracy is estimated.

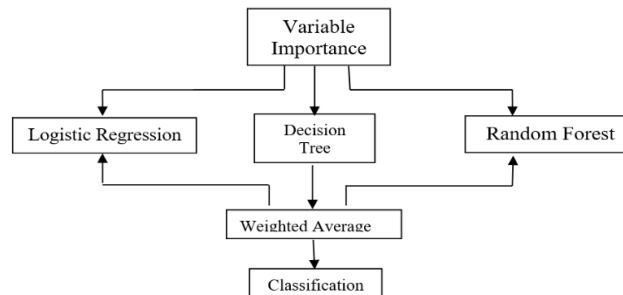


Figure 3. Architecture diagram

3.1. Modelling

Once we've identified important features and data is ready to format, we can proceed the project's predictive analysis section. The most used algorithms are classification and regression trees (CART). Some more unique, and advanced algorithms are also available, such as ridge and lasso, Naïve Bayes, linear discrimination and supporting vector machines used to compare the results of this work. As few algorithms are impacted because numerical factors are different, it is always recommended that the feature is scaled before analysis. Algorithms such as lasso and ridge, while other CART algorithms, are highly influenced due to scaling of the functions.

The modelling begins when the information collected is divided into a training phase and a test set. In the training dataset, algorithms are implemented and tested. The training set is assigned randomly 80% of the data and the test set is assigned 20% of the data. Since our response variable is binary, we will use classification algorithms. The characters "yes" and "no" of the attrition factor are transformed to 1 and 0 for convenience respectively.

3.1.1. Linear discriminant analysis

It is a method used to generalize the linear discriminant of Fisher. It is used in the statistics to find a linear combination of factors to differentiate two or more categories of objects or events for pattern recognition and ML. Basically, the result or the dependent variable can be continuous with a linear discriminant analysis and may contain an effect on another variable and find a sequential combination of characteristics that best characterize or separate the categories 2 or more. Linear discriminant analysis (LDA) is thus used for fine-tune the linear combination of characteristics to the best describe or separate two groups, which indicate if the employee is attrition or not. At beginning, the total data set is partitioned or randomly divided into 80% and 20% training and testing data sets. Figure 4 shows that the receiver operating characteristic (ROC) curve for various classifiers utilizing false positives rate (FPR), and true positive rate (TPR).

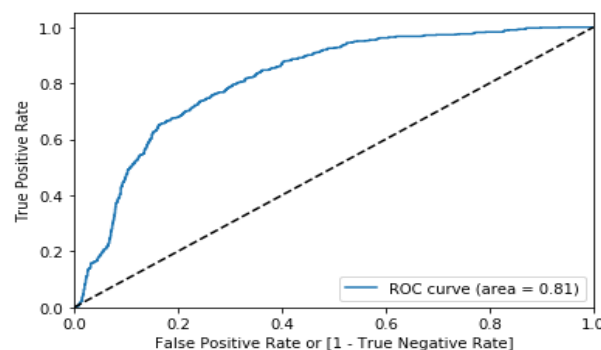


Figure 4. Receiver operating characteristic

The higher region below the curve, the more specifically the classifiers. In this work for plotting a ROC curve probability of employee being classified as leaver and actual class (target variable) will be used. ROC Curve implies different grouping by different coordinates. The node (0,1) for example, implies FPR = 0, TPR = 1. We can get false positive (FP) = 0 and false negative (FN) = 0 according to the TPR and FPR formula. And this is the perfect case because it does all manner of classifications. For node (1,0), meaning FPR = 1, and TPR = 0. Thus, we can obtain true positive (TP) = 0, TN = 0. This is the worst situation in fact, because it prevents all correct classification. We can get FPR=TPR = 0, for the node (0,0), so FP=TP = 0. Therefore, it declares everything to be negative class. Similarly, we can get FPR=TPR = 1 and FP = TP = 1 for node (1,1).

3.1.2. Logistic regression

Logistic regression is a way by which ML is characterized. In this calculation, a logistic regression is used to display the probabilities reflecting possible outcomes of a separate trial.

$$P(\text{churn}|w) = \frac{1}{1+e^{-[w_0+\sum_{i=1}^N w_i x_i]}} \quad (1)$$

$$y = \ln\left(\frac{P(Y=1|x)}{1-P(Y=1|x)}\right) = w_0 + \sum_{i=1}^P w_i x_i \quad (2)$$

x is a vector of independent variables of dimension p and y is the logit (log odds). w_0, L, w_p are model parameters. Logistic regression is a kind of regression that matches the values to logistic function. It is suitable in situations where the dependent variable is categorical. The common form of a model is:

$$P(Y | X^-, W) = \frac{1}{1+e^{-(w_0 + \sum w_i x_i)}} \quad (3)$$

3.1.3. Random forest

The random forest is an ineffable decision tree building block. The decision trees are often regarded as a weak learner as their prediction accuracy is incredibly low. Random forests have been employed to classify the endogenous earthquake sources. A random forest nevertheless collects and incorporates a group of decision makers to achieve particularly powerful program based. This concept, that somehow a collection of weak students is used to create a strong learner, validates the basis for classification techniques, which are often found in machine learning. Random forest samples randomized exercise data to replace each decision tree called classifier. In order to make a single decision, each decision tree returns to a class and then incorporates luggage. Random forests are an accordion of decision-making objects that should be more effective and accurate.

- 1) Pick k usefulness arbitrarily from the all-out m usefulness.
- 2) Where $k < m/2$ is concerned. Calculate the hub d utilizing the best part point among the k highlights.
- 3) Use the right split to partition the hub into little girl hubs.
- 4) Repeat 1 to 3 stages until coming to 1 number of hubs.
- 5) Create woods to assemble n number of trees by rehashing stages 1 to 4 for n number occasions.

3.1.4. Decision tree classifier

A decision tree is utilized as a structure similar to a tree for construction regression and classification models. Decision trees have utilized to categorize hotspot events, it divides set of data into few more subsets while concurrently an associated decision tree is established [15]. The last result is a tree with the required decision nodes. The decision tree is an identification. The other independent variables are the (decision nodes). First of all, the entire data set is splitted randomly into the 80% ratio of a train and the 20% test data. Calculation for decision tree:

- Step 1: having an unfilled N hub.
- Step 2: on the off chance that examples are the entirety of a similar class c , return N as a class c -named leaf hub;
- Step 3: in the event that the characteristic rundown is vacant in the record, at that point return N as a leaf hub set apart with the most widely recognized example name;
- Step 4: pick the check quality, the property in the record between the trait set;
- Step 5: mark hub (N) with characteristic for the checking;
- Step 6: for each realized check property estimation ai ;
- Step 7: grow a branch for the test attribute = ai condition from hub N ;
- Step 8: leave Si alone the example set for which attribute = ai test;
- Step 9: in the event that Si is unfilled, include a leaf hub set apart with the most widely recognized example class;

- Step 10: at that point include the hub returned by the choice tree age (S_i , test-characteristic, quality rundown).

Decision tree model is then applied to the training data set using the “rpart” instruction on the “attrition” based or output variable with all other impartial data set factors. The whole decision tree algorithm on this data set is compiled with the “fancyRpartPlot” command.

Required predictions are made on the test data set utilising the model of train dataset decision tree to predict whether or not a given new employee will attrition based based on probability values ranging from 0 to 1 and then using a specific threshold of 0.5 (in this case), these predictions are cut to 0 and 1. Confusion matrix has been described in Table 1 to Table 3. A decision tree is a type of tree-like structure that is used to develop classification or regression models. Alduayj and Rajpoot [9] classified hotspot occurrences using decision trees. It splits the data set into smaller and smaller number of subsets during developing an accompanying decision tree in stages

The end result is a tree with all of the necessary decision and leaf nodes. A classification or decision is represented by a leaf node (i.e., the goal or dependent variable “attrition” for this data set). The decision nodes are represented by the other independent variables. To begin, the complete dataset is randomly separated or split into train and test datasets in a ratio of 80% train and 20% test. Then the decision tree model is applied to the training data set using the “rpart” command on the dependent or output variable “attrition” along with all other independent variables. The command “fancyRpartPlot” is used to visualise the full decision tree model on this training dataset. Required predictions are made on the test dataset using this decision tree model of the train dataset to predict whether or not a particular new employee will be on the decline based on probability values ranging from 0 to 1 and beyond by taking a threshold Specific to 0.5 (in this case), predictions are divided into two ratings 0 and 1 classifications (i.e. the given employee will not be in attrition).

Table 1. Confusion matrix and evaluation to decision tree

Predicted	Actual	
	0	1
0	0	1
1	467	13
	667	220
(Model evaluation metrics)		
Accuracy rate	98	
Precision rate	98	
Recall rate	93	
F1-measure	95	

Table 2. Confusion matrix and evaluation to random forests

Predicted	Actual	
	0	1
0	0	1
1	230	19
	450	650
(Model evaluation metrics)		
Accuracy rate	99	
Precision rate	99	
Recall rate	96	
F1-measure	97	

Table 3. Confusion matrix and evaluation to logistic regression

Predicted	Actual	
	0	1
0	0	1
1	290	14
	450	367
(Model evaluation metrics)		
Accuracy rate	80	
Precision rate	83	
Recall rate	92	
F1-measure	87	

The random forest approach which was first advanced by Breiman in 2001, can be classified as an ensemble model [26]. What’s the point of an ensemble? the omnipresent decision tree is the foundation of a random forest. Because of its low prediction accuracy, the decision tree is frequently referred to as a weak learner. Random forests were utilised to classify endogenous seismic sources in a landslide. A random forest, on the other hand, assembles a group (or ensemble) of decision trees and combines their prediction ability to achieve comparatively good predictive performance-strong learner. This notion of bringing together a group of weak learners to generate a strong learner is at the heart of ensemble methods, which are commonly used in machine learning. Bagging is when RF takes a random sample of training data and replaces it with new data before generating each decision tree. Each decision tree returns a class, which is subsequently combined by bagging to provide a unique choice. Random forest is an ensemble of decision trees that is supposed to perform better and, as a result, provide more accuracy.

4. RESULTS AND DISCUSSION

To predict employee turnover, we should predict who will leave. This is a binomial classification problem, where the group of employees has to be splitted into two groups based on some characteristics, one with higher risk of retention and one with lower risk. Commonly used ML algorithms for binomial classification problem are: decision trees, logistic regression, random forest. The following algorithm is used to predict the model accuracy and calculate the confusion matrix. Only a couple of information mining procedures were utilized in the current frameworks for information expectation. Here we additionally utilized the representative informational index highlight determination technique. The representative informational index for the most part contains data about specialists, for example, number of activities, left, pay, level of fulfillment, and so forth. We can pick those fitting highlights from the worker informational index for our

survey by utilizing highlight determination. The steps used in arriving at the results are: 1) dataset is collected. In research, Kaggle dataset is used, 2) predictive model for data visualization and analysis for the attrition is improved, 3) usage of machine algorithms such as random forest, linear regression and decision tree, and 4) finding the best results for the parameters of accuracy, sensitivity and precision.

4.1. Data overview

Kaggle dataset was used where, there are 15,000 specialists in the dataset. Among these, 271 laborers are set apart as turnover, which means as far as the whittling down qualities, they are set apart as “yes”. The turnover proportion is 13.5% and is plainly a lopsided information issue. Here we are showing the main parameters for the prediction and implementation in Table 4, these 10 are the following parameters. It includes all the characteristics or parameters of the employee. After the data set is collected, the data to be modelled must be determined and cleaned up and shaped. The process of data preparation may be followed by choose the metric results and predictor variables, determine how much data can be modelled, cleaned and prepared. For each variable various data types and measuring levels should be identified in the dataset. Data for outliers, missing or incorrect values should then be evaluated. Skew and high cardinality in data should be removed. We use Python with SciPy, NumPy, Pandas, scikit-learn, and Matplotlib in the test and create three apparatuses: one instrument is for RF-based component positioning, one is a visual device for breaking down element factors and target variable, and one is RF-put together demonstrating device based with respect to the weighted F-measure. The dataset we chose contains 10 different attributes such as satisfaction level, last evaluation, average monthly hours, number of projects, time spent in company, left, work accident, promotion in last 5 years, sales, salary. These are represented as a range of values depending on the corresponding feature which is represented in Table 4.

Table 4. Brief of dataset

No.	Feature	Value range
1.	Satisfaction level	0.4 – 0.99
2.	Last evaluation	0.5 – 1
3.	Average monthly hours	96 – 350
4.	Number of projects	2 – 8
5.	Time spent in company	2 – 10
6.	Left	0 – 1
7.	Work accident	0 – 1
8.	Promotion in last 5 years	0 – 1
9.	Sales	Sales, support, technical, IT, marketing, accounting, HR, product mng, RandD
10.	Salary	Low, medium, high

Here the predictive model for the attrition is improved with the help of anaconda-command-prompt to first utilise Python Jupyter. But then to improve the prediction of data visualisation and analysis, here the ML learning and python have been decided. However, it was too time intensive and complicated to download libraries and set up a system, machine-learning, data visualisation and the logic of predictive and decision-making analytics were unlikely. It was therefore a way of achieving the role of ML and data visualization. The following steps are used to pre-process the data to a usable format before using data for the ML algorithm training and evaluating the test data:

- 1) Remove values from null.
- 2) No unique value columns removed.
- 3) Unnecessary columns have been removed.
- 4) Modification of the outdated the unicode transformation format (UTF) for UTF support.
- 5) The weight of each column depended on the columns rearranged.
- 6) Save data that was cleaned in csv format.

Once the important features are identified, and the data is in a model-ready format, we can begin the predictive analytics part of the project. Modeling begins by dividing the available data into a training set and a test set. Algorithms are deployed in the training set and tested in the test set. 80% of the data is randomly assigned to the training set and 20% of the data is assigned to the test set. We will use classification algorithms because our response variable is binary. The ‘yes’ and ‘no’ characters in the attrition variable are converted to 1 and 0 respectively to make it easier.

4.2. Performance matrices

In this paper, Table 5 provides the performance measures used for the methodology assessments, and these entries are used for identifying and measuring classification accuracy. By using the confusion matrix, performance metrics like, accuracy, precision, specificity and sensitivity of the calculated classifier and assembly model and their results as displayed in Table 5. There will be (four cases) if the result is

positive it can be expected to be TP. It is a false negative if it is expected to be a FN. If the result is negative and can be expected to be negative, it is true negative (TN); if the result is positive and can be expected to be positive, it is FP.

Table 5. Performance metrics & their definition

Metric	Equation	Definition
Accuracy	$(TP + TN) / (P+N)$	Ratio of the total number of predictions which are correct
Sensitivity	$TP / (TP + FN)$	Ratio of the positive cases which are correctly identified
Specificity	$TN / (FP / TN)$	Ratio of the negative cases which are correctly identified
Precision	$TP / (TP + FP)$	Ratio of the predicted positive cases which are correct

4.3. Matrix of confusion for different methodologies

This matrix describes the Predictive Analysis methodology on HR data, in order to analyze the use of predictive analysis for HR. The following chosen matrix for prediction is employee turnover, because of its high importance for organization. Thus, goal of the thesis is to predicting the employee turnover. Recall, F1 performance, and accuracy are all infuriating matrix metrics. (Positives 1) and (negatives 0) will occur when we divide a sample into two parts. There will be (four cases) if the result is positive it can be expected to be positive (TP). It is a false negative if it is expected to be a negative (FN). If the result is negative and can be expected to be negative, it is TN; If the result is positive and can be expected to be positive, it is FP.

Figure 5 to Figure 7 predicting the following algorithms used here such as decision tree, logistic regression, random forest tree. The accuracy provided by the decision tree is 97%. The accuracy provided by random forest algorithm is 98%. The accuracy provided by the logistic regression is 78%. Hence, there calculate and get the best accuracy from random forest algorithm of the employee data set.

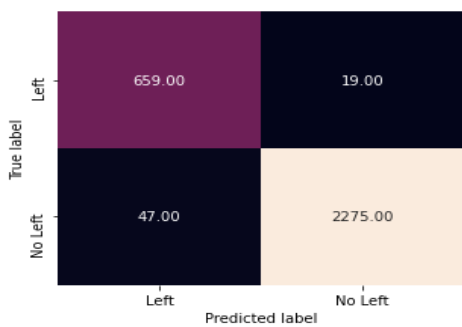


Figure 5. Decision tree matrix

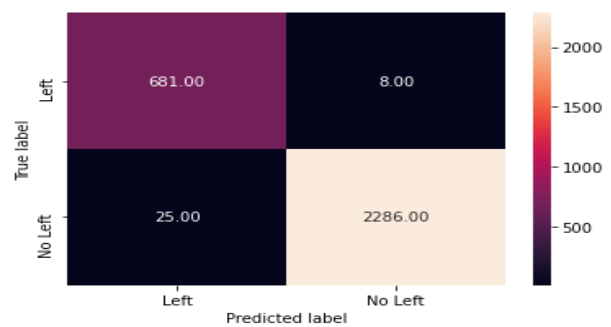


Figure 6. Random forest matrix

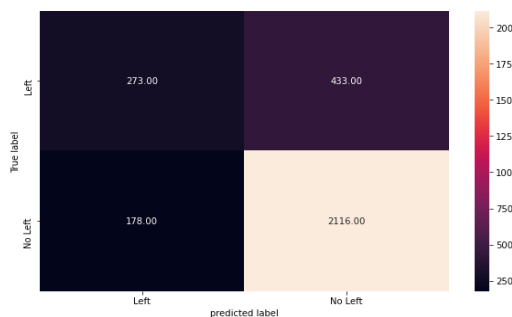


Figure.7. Logistic regression

To conclude, random forest classifier performed better than other tested models, on the Kaggle sample datasets. Therefore, random forest model has been used to predict employee turnover in company. Prediction results have been saved in flat file, including prediction probabilities for each employee. In addition, for interpretation of the decision path, decision trees were represented as graphs. Furthermore, feature importance has been evaluated to better understand variables that influenced the decision most.

Such influencers are monthly income, satisfaction, last evaluation, and left, all in correlation with over time. In this work, data preparation has been done on HR datasets as shown in Figure 8. Example dataset was used for demonstrating methods used for cleaning and preparing data. First of all, employee data from multiple sources has been gathered and unified. Then missing values has been imputed, outliers have been removed and skew has been reduced. Afterwards parameters for selected ML algorithms have been tuned using different methodologies. Later, on pre-processed employee data different ML algorithms have been applied and algorithm with best prediction accuracy has been selected.

From Table 6, above outcomes describes, compared to the other model, the model has various TP and TN. The accuracy obtained by using decision tree classifier with a 97% efficiency outperforms other mining techniques. Considering the sensitivity parameter, decision tree classifier has the highest value. The effects of predictive ML techniques on the employee dataset show that random forest evaluation outperforms other ML techniques followed by model of logistic regression for the specific dataset if precision is the preferred metric. Identification of human resources is forecasted using ML algorithms on employee data. This suggests that the prediction of the employees interest in maintaining their job or not was better estimated by this method compared to the other techniques. Several implementations are inferior to other frameworks in consistency, reliability, and specialization. We conclude from this that the model's consistency is stronger than other models.

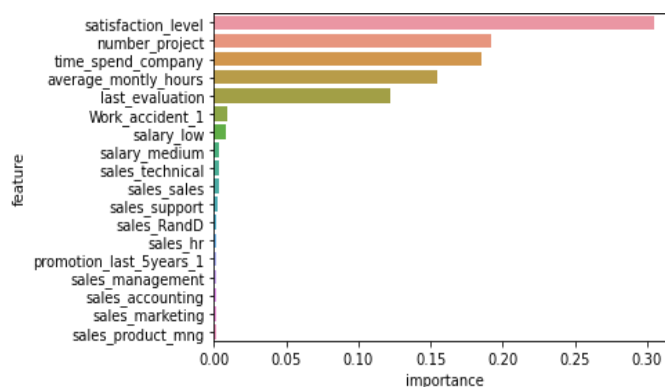


Figure.8. Analysis the parameters

Table 6. Performance metrics

Algorithms	Precision	Sensitivity	Accuracy
Logistic regression	83	81	78%
Random forest	99	82	97%
Decision tree	98	84	98%

5. CONCLUSION

There are several useful algorithms, both qualitative and quantitative, for anticipating staff turnover in order to stabilise a company. However, these methodologies have a number of drawbacks when it comes to predicting staff turnover, such as the following: Inadequate consideration of inconsistent data. The data provided in these studies regarding employee turnover only reflects a small portion of the entire workforce. Inefficient data processing due to high dimensionality. There are a range of feature dimensions, both static and dynamic, that characterize employees. In addition, there is no evaluation, given that the aim of the forecast is to reduce employee turnover, we must define and prioritize its components.

In this paper, identification of human In this paper, identification of human resource is forecasted using ML algorithms on employee data. Employee dataset has been collected from Kaggle. The emphasis is on utilizing various ML algorithms and mixtures of several recognizes the importance for efficient and effective employee attrition reduction using best ML algorithms, as employee attrition is one of the most crucial business problems. On the kaggle employee data, following algorithms has been implemented such as logistic regression, decision tree classification, linear discriminant analysis and random forest. We discovered that the accuracy obtained by using a random forest analysis model with a 99% efficiency outperforms mining techniques. As a result, the effects of predictive ML techniques on the employee dataset show that random forest evaluation outperforms other ML techniques followed by model of logistic regression for this specific dataset if precision is the preferred metric. The biggest drawback of random forest is that it can become very slow and ineffective for real-time predictions if there are too many trees. Generally, these

algorithms are quick to learn but take a long time to make predictions after they have been taught. The applications of this method is used to determine the efficiency of staff members, relationship of retirement behaviors on employee turnover, like, job content, lateness and absenteeism, length of service, and demographics. It has the application of the logit, and probit models for voluntary employee turnover predictions and to explore different personnel and job factors influencing the voluntary turnover of employees. It is used to investigate employee attrition using multiple decision tree algorithms and to correlate data mining techniques to calculate employee inconvenience. In random forest, the future study can be on improving the accuracy by utilising different attribute split measurements, different combine functions, or both. Increasing the diversity of base classifiers is a continuous process of quality improvement that will improve accuracy. As a result, discovering new approaches to achieve diversity will undoubtedly be a research topic in the future. For enhancing the accuracy of random forest classifiers, out of bag (OOB) estimates, proximity computation, and variable importance features can be used more extensively. The random forest technique generates a large number of classification trees, each of which is independent of the others. As a result, random forest is an excellent option for parallel processing. Furthermore, data mining is typically done on very large datasets, and random forest can handle datasets with a lot of predictors. As indicated in section 4, each parallel random forest implementation is tailored to a particular platform or language. As a result, there is room for a generalised random forest parallel algorithm. Business data is dispersed around the globe due to the geographical spread of business and the world's connection to the Internet. As a result, developing a distributed random forest algorithm is an important future research topic.




REFERENCES

- [1] J. Y. Yong, M. Y. Yusliza, T. Ramayah, C. J. C. Jabbour, S. Sehnem, and V. Mani, "Pathways towards sustainability in manufacturing organizations: Empirical evidence on the role of green human resource management," *Business Strategy and the Environment*, vol. 29, no. 1, pp. 212–228, 2020, doi: 10.1002/bse.2359.
- [2] M. M. Abdeldayem and S. H. Aldulaimi, "Trends and opportunities of artificial intelligence in human resource management: Aspirations for public sector in Bahrain," *International Journal Of Scientific and Technology Research*, vol. 9, no. 1, pp. 3867–3871, 2020. [Online]. Available: <http://www.ijstr.org/final-print/jan2020/Trends-And-Opportunities-Of-Artificial-Intelligence-In-Human-Resource-Management-Aspirations-For-Public-Sector-In-Bahrain.pdf>
- [3] A. K. Mishra, "Assessment of human resource capacity of construction companies in Nepal," *Journal of Advanced Research in HR and Organizational Management*, vol. 5, no. 4, pp. 14–25, 2018. [Online]. Available: https://www.researchgate.net/profile/AnjayMishra/publication/329776127_Assessment_of_Human_Resource_Capacity_of_Construction_Companies_in_Nepal/links/5c1a32bd299bf12be38a6b56/Assessment-of-Human-Resource-Capacity-of-Construction-Companies-in-Nepal.pdf
- [4] S. Kraus, C. Palmer, N. Kailer, F. L. Kallinger, and J. Spitzer, "Digital entrepreneurship: A research agenda on new business models for the twenty-first century," *International Journal of Entrepreneurial Behavior and Research*, vol. 25, no. 2, pp. 353–375, 2018, doi: 10.1108/IJEBR-06-2018-0425.
- [5] S. Al-Qudah, A. M. Obeidat, H. Shrouf, and M. A. Abusweilem, "The impact of strategic human resources planning on the organizational performance of public shareholding companies in Jordan," *Problems and Perspectives in Management*, vol. 18, no. 1, pp. 219–230, 2020, doi: 10.21511/ppm.18(1).2020.19.
- [6] A. Gatto, "A pluralistic approach to economic and business sustainability: A critical meta-synthesis of foundations, metrics, and evidence of human and local development," *Corporate Social Responsibility and Environmental Management*, vol. 27, no. 4, pp. 1525–1539, 2020, doi: 10.1002/csr.1912.
- [7] S. Basnyat and C. S. C. Lao, "Employees' perceptions on the relationship between human resource management practices and employee turnover: A qualitative study," *Employee Relations*, vol. 42, no. 2, pp. 453–470, 2020, doi: 10.1108/ER-04-2019-0182.
- [8] S. N. Khera and Divya, "predictive modelling of employee turnover in indian it industry using machine learning techniques," *Vision*, vol. 23, no. 1, pp. 12–21, 2019, doi: 10.1177/0972262918821221.
- [9] S. S. Alduayj and K. Rajpoot, "Predicting employee attrition using machine learning," in *2018 International Conference on Innovations in Information Technology (IIT)*, 2018, pp. 93–98, doi: 10.1109/INNOVATIONS.2018.8605976.
- [10] A. M. Karminsky and R. N. Burekhin, "Comparative analysis of methods for forecasting bankruptcies of Russian construction companies," *Journal Business Informatics*, vol. 13, no. 3, pp. 52–66, 2019, doi: 10.17323/1998-0663.2019.3.52.66.
- [11] H. Zhang, L. Xu, X. Cheng, K. Chao, and X. Zhao, "Analysis and prediction of employee turnover characteristics based on machine learning," in *2018 18th International Symposium On Communications And Information Technologies (ISCIT)*, 2018, pp. 371–376, doi: 10.1109/ISCIT.2018.8587962.
- [12] G. Gabrani and A. Kwatra, "Machine learning based predictive model for risk assessment of employee attrition," in *International Conference on Computational Science and Its Applications*, 2018, pp. 189–201, doi: 10.1007/978-3-319-95171-3_16.
- [13] D. K. Srivastava and P. Nair, "Employee attrition analysis using predictive techniques," in *International Conference On Information And Communication Technology For Intelligent Systems*, 2017, pp. 293–300, doi: 10.1007/978-3-319-63673-3_35.
- [14] A. M. E. Sikaroudi, R. Ghousi, and A. E. Sikaroudi, "A data mining approach to employee turnover prediction (case study: arak automotive parts manufacturing)," *Journal Industrial System Engineering*, vol. 8, no. 4, pp. 106–121, 2015.
- [15] S. Sajjadiani, A. J. Sojourner, J. D. K. -Mueller, and E. Mykerezzi, "Using machine learning to translate applicant work history into predictors of performance and turnover," *Journal of Applied Psychology*, vol. 104, no. 10, 2019, doi: 10.1037/apl0000405.
- [16] C. Leung, A. Law, and O. Sima, "Towards privacy-preserving collaborative gradient boosted decision trees," UC Berkeley, 2019. [Online]. Available: https://people.eecs.berkeley.edu/~kubitron/courses/cs262a-F19/projects/reports/project6_report.pdf
- [17] R. S. Shankar, J. Rajanikanth, V. V. Sivaramaraju, and K. V. S. S. R. Murthy, "Prediction Of Employee Attrition Using Datamining," in *2018 IEEE International Conference On System, Computation, Automation And Networking (ICSCA)*, 2018, pp. 1–8, doi: 10.1109/ICSCAN.2018.8541242.
- [18] L. Alaskar, M. Crane, and M. Alduailij, "Employee turnover prediction using machine learning," in *International conference on computing*, 2019, pp. 301–316, doi: 10.1007/978-3-030-36365-9_25.




- [19] F. Fallucchi, M. Coladangelo, R. Giuliano, and E. W. De Luca, "Predicting employee attrition using machine learning techniques," *Computers*, vol. 9, no. 4, p. 86, 2020, doi: 10.3390/computers9040086.
- [20] Y. Zhao, M. K. Hryniewicki, F. Cheng, B. Fu, and X. Zhu, "Employee turnover prediction with machine learning: a reliable approach," in *Proc. of SAI Intelligent Systems Conference*, 2018, pp. 737–758, doi: 10.1007/978-3-030-01057-7_56.
- [21] A. Chourey, S. Phulre, and S. Mishra, "A survey paper on employee attrition prediction using machine learning techniques," *Journal of Interdisciplinary Cycle Research*, vol. 11, no. 12, pp. 199–202, 2019. [Online]. Available: <http://www.jicrjournal.com/gallery/24-jicr-december-2247.pdf>
- [22] X. Gao, J. Wen, and C. Zhang, "An improved random forest algorithm for predicting employee turnover," *Mathematical Problems in Engineering*, pp. 1–12, 2019, doi: 10.1155/2019/4140707.
- [23] S. Karande, A. Shelake, Sivagami M., and S. Shopia, "Prediction of employee retention using cassandra and ensemble learning," *Journal of Information Organization*, vol. 9, no. 4, pp. 134–140, 2019. [Online]. Available: https://www.dline.info/jio/fulltext/v9n4/jiov9n4_3.pdf
- [24] M. Alghamlas And R. Alabduljabbar, "Predicting the suitability of it students' skills for the recruitment in saudi labor market," in *2019 2nd International Conference on Computer Applications and Information Security (ICCAIS)*, 2019, pp. 1–5, doi: 10.1109/CAIS.2019.8769577.
- [25] M. Kowsher, I. Hossen, A. Tahabilder, N. J. Prottasha, K. Habib, and Z. R. M. Azmi, "Support Directional Shifting Vector: A Direction Based Machine Learning Classifier," *Emerging Science Journal*, vol. 5, no. 5, pp. 700–713, 2021, doi: 10.28991/esj-2021-01306.
- [26] G. Biau and E. Scornet, "A random forest guided tour," *Test*, vol. 25, no. 2, pp. 197–227, 2016, doi: 10.1007/s11749-016-0481-7.

BIOGRAPHIES OF AUTHORS






Elham Mohammed Thabit A. Alsaadi    BSc. in Computer Science from Al-Mustansiriya University -Iraq, MSc. in Real Time Systems from UK. Ph.D. from College of Information Technology at Babylon University-Iraq. She is a lecturer in College of Computer Science & Information Technology, Karbala University. Her research interests include: Artificial Intelligence, Computer Vision, Image Processing, Database, System Analysis, and E- business. She can be contacted at email: elham.thabit@uokerbala.edu.iq.



Sameerah Faris Khlebus    MSc. in Data Security from department of data security. She is a lecturer in University of Information Technology & Communications. Her research interests include: data security, artificial intelligence, information systems and information technology. She can be contacted at email: sameerah.alradh@uoitc.edu.iq and sameera_alradhi@yahoo.com.



Ashwak Alabaichi    Msc. in steganalysis. Ph.D. in cryptography from Computing of School, Universiti Utara Malaysia, Kedah, Malaysia. She is a lecturer in Department of Biomedical Engineering, College of Engineering, University of Kerbala, Iraq. Her research interests include cryptography, Image Processing, Stegoanalysis, and Data security. She can be contacted at email: ashwaq.mahmood@uokerbala.edu.iq.