Early Prediction of Employee Turnover Using Machine Learning Algorithms

Original Scientific Paper

Markus Atef
Faculty of Management Sciences,
October University for Modern Sciences and Arts (MSA), Giza, Egypt
Business Information Systems Department,
Faculty of Commerce and Business Administration,
Helwan University, Cairo, Egypt
markusdaoud96@outlook.com

Doaa S. Elzanfaly
Information Systems Department, Faculty of Computers and Artificial Intelligence,
Helwan University, Cairo, Egypt
Faculty of Informatics and Computer Science, British University in Egypt, Cairo, Egypt
Doaa.saad@fci.helwan.edu.eg

Shimaa Ouf
Business Information Systems Department, Faculty of Commerce and Business Administration,
Helwan University, Cairo, Egypt
shimaaouf@yahoo.com

Abstract *– Employee turnover is a serious challenge for organizations and companies. Thus, the prediction of employee turnover is a vital issue in all organizations and companies. The present work proposes prediction models for predicting the turnover intentions of workers during the recruitment process. The proposed models are based on k-nearest neighbors (KNN) and random forests (RF) machine learning algorithms. The models use the dataset of employee turnover created by IBM. The used dataset includes the most essential features, which are considered during the recruitment process of the employee and may lead to turnover. These features are salary, age, distance from home, marital status, and gender. The KNN-based model exhibited better performance in terms of accuracy, precision, F-score, specificity (SP), and false-positive rate (FPR) in comparison to the RF-based model. The models predict the average probability percentage of turnover intentions of the workers. Therefore, the models can be used to aid the human resource managers to make precautionary decisions; whether the candidate employee is likely to stay or leave the job, depending on the given relevant information about the candidate employee.*

Keywords*: Prediction Models, Employee Turnover, Machine Learning Algorithms*

1. INTRODUCTION

Employee turnover can be defined as the rate of employees who quit an organization and are substituted by new employees. A high employee turnover rate represents a potentially fatal problem for organizations due to the high costs of separation, vacancy, recruitment, training, and replacement. Moreover, an organization with a high turnover rate eventually becomes understaffed, consequently non-productive and its growth stagnates [1]. Thus, the prediction of employee turnover is a vital procedure for any sustainable organization, where acquiring early information regarding employee turnover status helps organizations to take precautions to such a status.

Artificial intelligence (AI) can be defined as the utilization of the machine instead of human brains to accomplish a required goal or carry out a certain task. Recently, there is an increasing tendency to apply AI in human resource (HR) management [2-3] because using the computational and processing powers of the machine is faster and more accurate than the human brain [4-6]. Nowadays, the AI field is an important aspect and a rapidly growing trend of the technology-driven economy. AI starts to run deeply on the organizational level, affecting some of its structures in some countries. The field of HR has steadily shown interest in the AI technology through baby-steps across the world. An AI-driven HR management will put back the routine jobs and complicated tasks of the human resource personnel;

thus, conserving substantial amounts of time, money, and manpower [7-9].

The literature provides a set of recommendations on how the AI tools and practices could be applied for specific HR management tasks like using machine learning techniques in employee selection [10] and recruitment by information extraction techniques [11, 12].

One of the main branches of artificial intelligence (AI) is machine learning; a scientific development of computer machines, where the machine can learn and adapt based on provided data and experience, without necessarily following programmed instructions. The learning process starts with data analysis of recurring patterns in a dataset. Then, the observed patterns are used to extrapolate a decision or predict an outcome. Various machine learning algorithms can be utilized for turnover prediction such as neural networks, apriori, KNN, extreme gradient boosting, RF, decision tree, logistic regression, support vector machines, etc. [13-17].

Accordingly, the aim of the present research is to construct data-driven prediction models based on machine learning algorithms to predict the likelihood of a candidate quitting his/her job in the future. This would help the organization managers with taking the necessary precautions to diminish the turnover rate. The importance of the present research from a practical standpoint is that as mentioned before, employee turnover is a huge problem, which causes quite a lot of drawbacks. The present work could be a step in diminishing the drawbacks of employee turnover by predicting whether employees will leave the organizations or not before recruitment to help the managers make decisions to diminish the turnover rate. The importance of the present research from an academic point of view is that to the best of our knowledge, there are some studies in the open literature, dealing with the turnover prediction by using machine learning algorithms. But, no academic research attempts have been conducted to tackle the issue at hand, from the same angle, by predicting employee turnover during the recruitment process. The present work's methodology of data-based predictions uniquely stands out amongst several open-literature studies addressing turnover concerns via machine learning. To achieve this goal, the dataset created by IBM was used in the present work. The data was carefully studied and selected to include only the essential features, which should be considered during the recruitment process of the employee and may lead to turnover. The selected features are salary, age, distance from home, marital status, and gender. Models have been built based on KNN and RF algorithms to predict the probability percentage of turnover intention of candidates before recruitment.

2. RELATED WORK

A comparative study involving accuracy and memory utilization of selected algorithms for predicting employee turnover was conducted by Rohit Punnoose et al. [18]. The authors collected the data from the information system used by the human resource department of a retailer with global operations and data from the Bureau of Labor Statistics. Several classification algorithms were applied, namely, extreme gradient boosting (XGBoost), logistic regression, Naïve Bayesian, RF, linear support vector machine, linear discriminant analysis and KNN. The authors have found that XGBoost exhibits the best performance regarding the accuracy and memory utilization.

Jain et al. [19] have carried out research to predict turnover rate using XGBoost. They have found that age, gender, marital status, years at the company, job satisfaction, and distance from home have the most significant effects on turnover among all attributes in the dataset.

Numerous algorithms; namely, logistic regression, gradient boosting classifier, support vector machine, and RF were applied to the IBM dataset prepared by IBM data scientists [20, 21]. After applying the RF classifier, fifteen features were found to be more significant in deciding whether employees quit their jobs or not. Out of the fifteen features, overtime and monthly income exhibit the highest influence on employees to leave their jobs or not. XGBoost was found to have the highest performance (with an AUC of 0.84596) among all the applied algorithms.

Zhao et al. [13] have evaluated the performance of ten supervised machine learning algorithms; namely, RF, gradient boosting trees, XGBoost, support vector machines, decision tree, neural networks, linear discriminant analysis, Naïve Bayesian, logistic regression, support vector machines and KNN on numerous HR datasets. The authors have found that XGBoost is the most reliable algorithm among all the applied algorithms.

Zhang et al. [22] have attempted to find out the most important factors that lead to employee turnover. The authors have found an essential correlation between department and work. Also, they have found that the gender of employees significantly affects turnover. A logistic regression algorithm was applied for predicting the turnover with an accuracy of 87.2%.

Sisodia et al. [23] have carried out an investigation to find out the reasons causing the employee turnover by building models using machine learning algorithms to forecast employee turnover. They used the dataset on Kaggle with ten features. They have found that the main reasons causing high employee turnover rates are time spent with the company, workload, and promotion. The used machine learning algorithms for building the models were decision tree, support vector machine, Naïve Bayesian, KNN, and RF. The accuracy, precision, F-score, recall, specificity, and FPR of the models were compared. In terms of accuracy, F-score, and precision, RF performed better and in terms of recall, the decision tree was better.

3. RESEARCH METHODOLOGY

3.1 RESEARCH FRAMEWORK

The turnover prediction framework is presented in Fig.1. The used methodology comprises several phases; namely, data collection, data cleaning, data selection, data preprocessing, benchmarking the algorithms, and evaluating the predicted outcome.

3.2 DATASET

The dataset was obtained from the Kaggle website (IBM, 2020). The IBM dataset comprises 1470 records with 34 features (6 categorical and 27 numeric), such as monthly salary, experience, distance from home, skills, nature of work, position etc.

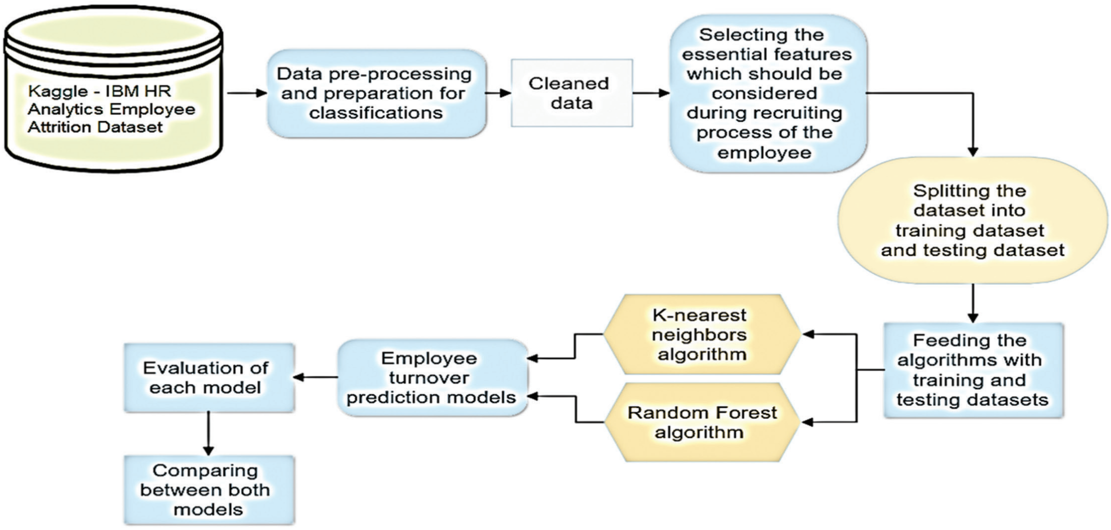

Kaggle - IBM HR Analytics Employee Attrition Dataset
Data pre-processing and preparation for classifications
Cleaned data
Selecting the essential features which should be considered during recruiting process of the employee
Splitting the dataset into training dataset and testing dataset
Feeding the algorithms with training and testing datasets
K-nearest neighbors algorithm
Random Forest algorithm
Employee turnover prediction models
Evaluation of each model
Comparing between both models

Fig. 1. Turnover prediction framework

3.3 DATA PREPROCESSING

Machine Learning algorithms can typically process only numerical input. Hence, the qualitative variables (gender and marital status) were encoded into quantitative variables (One-Hot Encoding) to input an acceptable format for the machine.

Based on the scientific literature [24-26] and logical reasoning, the features in the employee dataset that are essential for the prediction of turnover before recruitment were cherry-picked. This means that the features were selected based on the information provided by the candidates for the jobs and before they are put to work. For example, features such as job involvement, job satisfaction, job level, overtime, relationship satisfaction, and so on cannot be considered during the recruitment process because the candidates for the jobs have not yet been put to work. To fulfill the aim of the present work, all other features were excluded and solely considered salary, distance from home, marital status, age, and gender, which should be considered during the recruiting process of the employee and may lead to turnover in organizations.

3.4 APPLIED MACHINE LEARNING ALGORITHMS

The two machine learning algorithms applied in the present work are RF and KNN algorithms. The two algorithms were chosen because most of the open literature vouch for their reliability and accuracy [13,15, 27-28]. For example, Badillo et al. [27] have reported that RF and KNN algorithms demonstrate the best combination of performance and interpretability. Zhao et al. [13] have proved that KNN is accurate and reliable with a small number of features. The novelty of the present work lies within the hybridization technique used to tailor the parameters to better represent the model features in a realistic manner. Moreover, the trial-and-error process in the KNN algorithm was carried out on the cross-K validation rather than the algorithm itself to find the optimal K value for the given dataset as opposed to finding the optimal K for a particular trial and error process. In the case of the RF algorithm, the Random Search algorithm was utilized to narrow down the range of each parameter, then the Grid Search algorithm was applied on the relevant set of parameters only, introducing more novelty to the work.

4. RESULTS AND DISCUSSIONS

4.1 DESCRIPTIVE ANALYSIS

The descriptive analyses were performed on the IBM employee turnover dataset, including the selected features that significantly affect the turnover; namely, age, distance from home, marital status, gender, and monthly income. The HR employee turnover dataset was loaded into MySQL workbench, which is a database management system, and MS Excel and then preprocessed. The gender was converted to 0 and 1 for male and female, respectively. The marital status was converted to 0, 1, and 2 for single, married, and divorced, respectively. The distance unit was converted from a mile to km. Then, the relationships between the employee turnover and the selected features were generated by using MySQL and MS excel. The distributions of the target variable (employee turnover) related to the selected features; namely, age, distance from home, marital status, gender, and monthly income within the dataset are presented in Figs. 2-6. In the IBM dataset, the total number of employees is 1470, from which 237 employees (16%) left the job. Fig. 2 presents the relationship between age and percentage of employee turnover. It can be observed that as the age increases the total turnover percentage or the turnover percentage in the cluster almost linearly decreases. The highest turnover percentage lies within the cluster of 18-24 years (43.7%) meanwhile, the lowest turnover percentage lies within the cluster of 43-48 years (9.1%). Concerning the percentage of total turnover, the highest turnover percentage lies within the cluster of 31-36 years (29.1%) and the lowest turnover percentage lies within the cluster of 55-60 years (4.6%). These findings indicate that younger employees have a more propensity to leave the job.

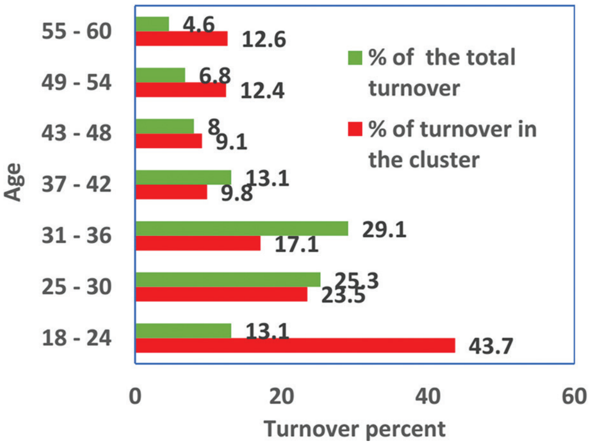


Fig. 2. Age versus percentage of the total turnover and percentage of turnover in the cluster

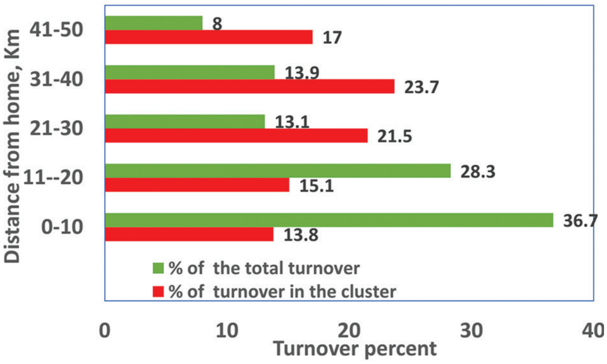


Fig. 3. Influence of distance from home on the percentage of the total turnover and percentage of turnover in the cluster

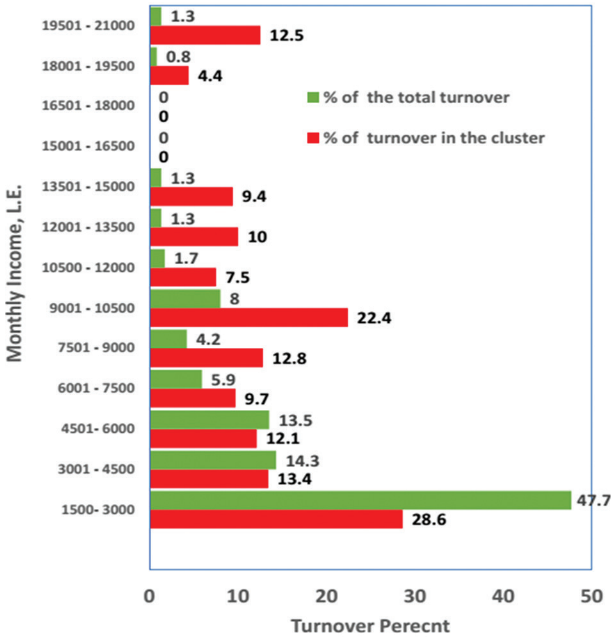


Fig. 4. Influence of monthly income on percentage of the total turnover and percentage of turnover in the cluster

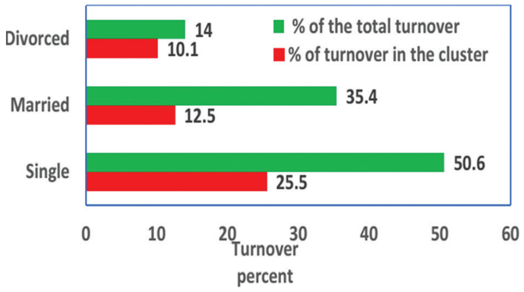


Fig. 5. Martial status versus percentage of the total turnover and percentage of turnover in the cluster

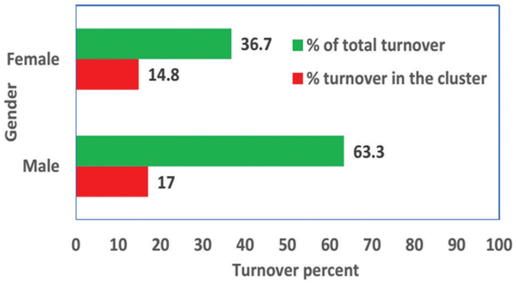


Fig. 6. Gender versus percentage of the total turnover and percentage of turnover in the cluster

Fig. 3 reveals the influence of distance from home on the percentage of employee turnover. The highest turnover percentage lies within the clusters of 21 – 30 and 31 – 40 km. Unexpectedly, the highest percentage (36.7%) of the total turnover lies within the cluster of 0-10 km and the lowest turnover percentage lies within the cluster of 0-10 km. The turnover percentage of employees as a function of salary is presented in Fig.4. The percentage of employee turnover is inversely proportional to the salary with the highest value of the lowest salary cluster (1500-3000 L.E). The employees with salaries between 15000 and 18000 L.E. do not likely show a tendency to leave the job. Fig.5 shows the distribution of turnover percentage related to marital status. Single employees have a higher desire to leave, but divorced employees are more likely to stay in the job as indicated by the highest turnover percentage occurred with the single employees, whereas the lowest turnover percentage with the divorced employees. Fig. 6 reveals the relation between turnover and gender. It is worth noting that female employees show a higher tendency to stay in the job. However, male employees have a strong propensity to leave as they represent 63.3% of the total employee turnover in the dataset.

4.2 FEATURE IMPORTANCE

After analyzing the dataset, the RF algorithm was applied to find the feature importance score, Fig.7.

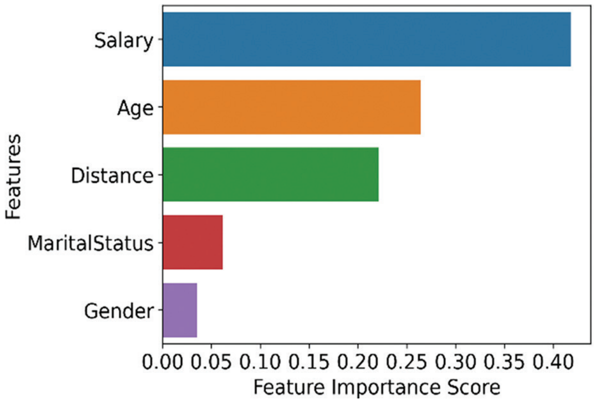
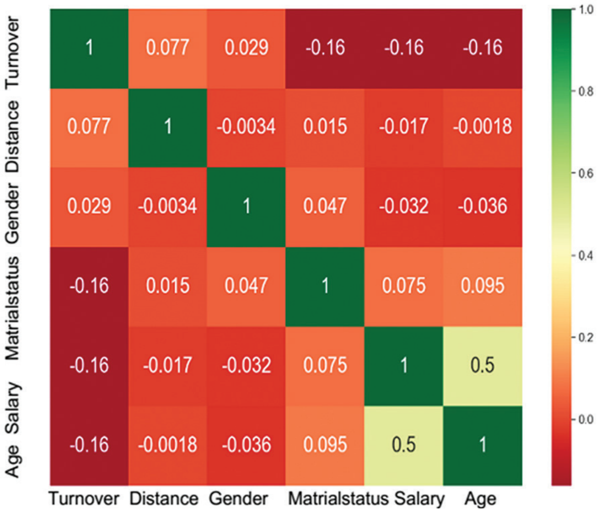


Fig. 7. Feature importance

From Fig. 7, it can be observed that salary exhibits the highest feature affecting the employee turnover and gender has the lowest effect.

4.3 HEATMAP

The heatmap uses a warm-to-cool color spectrum to show the correlations between variables. For obtaining the heatmap, pyplot, pandas, and seaborn libraries were imported and employed. The heat map, demonstrating the correlations between features and target variable (turnover), is presented in Fig. 8. A negative correlation indicates that the alteration of a feature or the target variable is inversely influenced by the other. However, a positive correlation indicates that the variation of a feature or the target variable is directly influenced by the other. The heatmap displays that there is no correlation higher than 0.5, which is between age and salary. The positive correlation between age and salary indicates that as the age increases, the salary increases. The correlation between distance from home and turnover is also positive, indicating that as the distance from home increases, the turnover increases. Moreover, the heatmap reveals a negative correlation between age and turnover, as well as salary and turnover, signifying that as age and salary decrease, the turnover increases (employees are more likely to leave the job).



	Turnover	Distance	Gender	Matrialstatus	Salary	Age
Turnover	1	0.077	0.029	-0.16	-0.16	-0.16
Distance	0.077	1	-0.0034	0.015	-0.017	-0.0018
Gender	0.029	-0.0034	1	0.047	-0.032	-0.036
Matrialstatus	-0.16	0.015	0.047	1	0.075	0.095
Salary	-0.16	-0.017	-0.032	0.075	1	0.5
Age	-0.16	-0.0018	-0.036	0.095	0.5	1

Fig. 8. Correlation heatmap

4.4 BUILDING OF THE MODEL

4.4.1 KNN-based model

The idea of KNN is to determine the K data points in the training data that are closest to the new instance and categorizes this new instance by a majority vote of its K neighbors. The KNN algorithm is usually denoted as K- Nearest Neighbor classification because it takes more than one neighbor into account. The dataset present in the CSV data was loaded into a DataFrame, which was split into the feature's matrix and the target values vector. Then, the categorical features were one- hot-encoded (OHE) by mapping each categorical feature to a vector basis in the n-space, where n is the cardinality of the feature set. OHE is essential for the machine to interpret the categorical data.

Sklearn implemented the model selection class, the K- Nearest neighbor classifier class, and the score matrix class. The model selection class was used for splitting the dataset into two sets; the training set and the testing set.

Moreover, the Sklearn library defined cross K-validation and grid search classes, which were used for K-folding and hyperparameter tuning. The K -Nearest neighbor classifier class carried out the KNN algorithm by using the training and testing sets. The score metrics class generated the algorithm's classification report and calculated the confusion matrix with the accuracy score per model training instance. The data was randomly split into two sets, X and Y, where X represented the feature matrix and Y was the target variable (turnover); the size of the training set was α, and the testing set was 1- α, where α was kept at 0.84. It should be mentioned that the target variable (turnover) is a binary representation of No (84%) and Yes (16%) outcomes. Thus, the dataset was split and kept at 84% and 16% for training and test datasets, respectively. Random splitting percentages were avoided because it can change the percentages of the classes present in the training and test datasets from that in the original to conform to the original distribution. Each set was further split into two sets where X was split into X-train, X-test, and Y was split into Y-train, Y-test. Before running the algorithm, the sets must be normalized to minimize the outlier effect on the dataset. In other words, feature scaling was employed to eliminate biases towards the outlier feature values. The feature scaler used for standardization on this data was the z-score method. It should be here mentioned that data normalization is one of the crucial issues in ML because unnormalized data allows one feature to entirely dominate the other features. The most common techniques for normalization are Min-max and Z-score. Min-max normalization ensures that all features will have an equal influence on that data despite the outliers. The Z-score technique handles outliers but does not generate normalized data with the same scale. Thus, in the present work, Z-score normalization was selected to be applied to the dataset.

The grid search algorithm was used for obtaining the optimal K value by performing the cross K-validation algorithm on the original standardized dataset. Finally, the model based on the KNN algorithm ran on the training and testing sets with the chosen K value, generating the classification report, confusion matrix, and accuracy score. The model based on the KNN algorithm ran several times to obtain the mean accuracy and graphs, which were generated based on the classification report, correspondingly.

Cross-validation is considered the most used technique to evade overfitting, diminish the bias of sampling data and guarantee model error randomness; thus, in the present work, ten-fold cross-validation was applied. Cross-validation randomly divided the dataset into ten-fold subsets, where each subset was used as a training dataset once and as a testing dataset nine times. This process was repeated iteratively ten times to ensure that each subdivision is used as the training set once. In each repetition, a different part was selected as the training set. Finally, the average prediction error was obtained by measuring the mean accuracy of the ten iterations performed on each validation set.

The Cross-validation result performed above was used for the grid search algorithm's parameters to determine the optimal K value; this process is known as hyperparameter tuning. The grid search algorithm exhaustively searched for the hyperparameters on all the datasets using cross-validation as a metric. Unlike the mean error method used to obtain the K value as conducted above, grid search guarantees a distinct K value relative to the cv metric. For instance, a cv value of 5 consistently provided an optimal K value of 13, meanwhile, a cv value of 10 provided an optimal K value of 12. As expected, the cross-validation results for each subset revealed low variance in its accuracy score, compared with the other subsets' accuracy scores, signifying a low bias of sampling data; hence, the effect of overfitting was diminished. It should be mentioned that the algorithm runs iteratively using the features (salary, age, distance from home, gender, and marital status) of the candidate employee before recruitment to predict the turnover intentions of the candidate by obtaining the average probability percentage of whether the candidate employee is likely to stay or leave the job, depending on the given features of the candidate employee. It is crucial to run the algorithm several times because a single run is susceptible to false positives or false negatives, which should be mitigated. An independent optimal K value was calculated for each run on the randomly split datasets. However, in the case of the full dataset being used as a training set for the prediction during the recruitment process, the optimal K is calculated only once.

4.4.2 RF algorithm-based model

The basic methodology of RF algorithm is to merge numerous algorithms, which is known as ensemble learning, to resolve a complex problem and enhance the model's performance. RF is based on tree algorithms, where it generates a set of decision trees from a subset of training data and collectively takes the prediction of all the trees instead of one decision tree. Then, a vote is conducted from each tree to conclude a prediction based on the most voted for the outcome. The final output is the prediction, which has the most votes. RF utilizes so-called "Bagging", which means that the consecutive forthcoming trees do not depend on the preceding trees. The accuracy and performance of a model based on RF could be enhanced by increasing the number of trees. This helps subside the effect of data overfitting. RF is a robust algorithm because each node is split depending on some prediction variables and probability functions, where the best subset of trees is split. Like the model based on the KNN algorithm, the CSV data was loaded into a DataFrame, which was split into the feature's matrix and the target values vector. Then, One-Hot-Encoding (OHE) was applied to the categorical features, where each feature was mapped to a vector basis in the n-space. In order to avoid feature bias, where a feature value prominently influences the data, feature scaling (Z-score normalization) was applied to the dataset.

The model selection class, the RF classifier class, and the score metrics class are implemented by Sklearn. The dataset was split into the training and testing sets by the model selection class. The dataset was split into the X-set with a size of α% and the Y-set with a size of 100% - α%, representing the feature matrix and the target variable (turnover), respectively.

The grid search algorithm is typically used to obtain the optimal hyperparameters, which yield the best performance for a model. Alternatively, a random search algorithm would find parameters yielding accurate results on average at risk of occasional non-optimal parameters. Hence, the random search algorithm has a high variance. Improving the random search algorithm results would require several iterations at least proportional to the set's cardinality, which could be inefficient on a set of a large size. On the other hand, the grid search requires a certain range of parameters to exhaustively search for the parameters yielding the most accurate results. Hypothetically, the range of parameters provided to the grid search algorithm could be impractically large. Taking advantage of both algorithms, multiple iterations of the random search algorithm on the dataset were used to prune the parameter ranges later provided to the grid search algorithm. Thus, eliminating the random search algorithm's high variance while maintaining efficiency for the grid search algorithm.

The cross-validation class and grid search class for K-folding and hyperparameter tuning were then applied to the dataset. The K-folding further divided each X- set and Y-set into K training and K testing sets, respectively. Then, the RF classifier was fit using the training and testing sets. Finally, the classification report, the confusion matrix, and the accuracy score were generated per model training instance using the score metrics class. Unlike the model based on the KNN algorithm, the grid search for hyperparameters of the RF algorithm is less methodical since the RF parameters should be tailored for a given dataset to obtain optimal results. Thus, a process of trial and error was necessary to narrow down the grid search range for each parameter. It is important to maintain a balance between the algorithm's efficiency and precision by keeping each parameter range concise. It follows that running the grid search for each RF iteration is inefficient since it increases the algorithm's runtime exponentially. In other words, the grid search of the RF algorithm was applied only once. In order to achieve an accurate prediction without sample bias, the RF algorithm should be applied multiple times with random data splits, where the prediction is the mean prediction probability of all iterations.

4.5 EVALUATION OF THE ALGORITHMS

For evaluating the performance of the present trained models, including accuracy, recall, precision, the area under the curve (AUC) of the receiver operating characteristics curve (ROC) and F-score, 'metrics' module from 'Sklearn' was used. Accuracies of the algorithms by ten-fold cross-validation are given in Table 1 and results of the KNN and RF-based models are presented in Table 2. The AUC-ROC curves are presented in Fig. 9. The AUC under the ROC is used to compare the accuracies of the models, where a higher AUC indicates better performance of the model.

Table 1. Prediction accuracies of KNN and RF-based models by 10-fold cross-validation

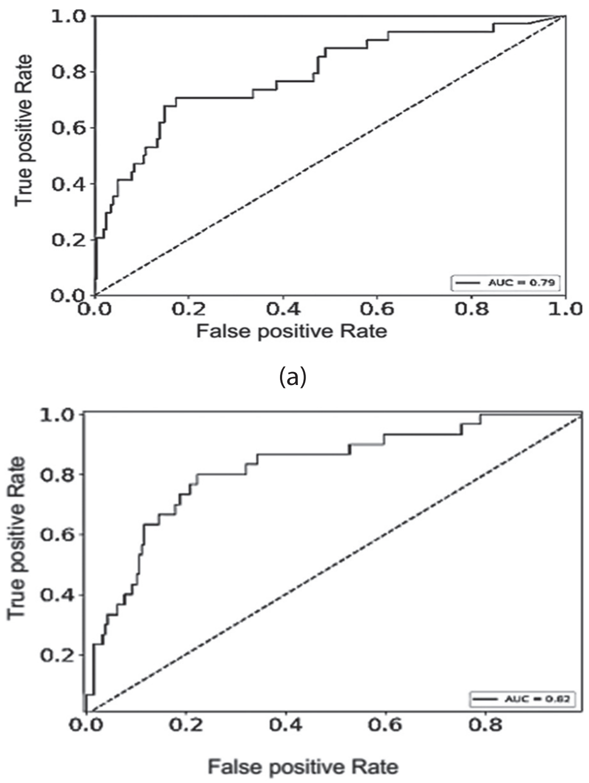
Accuracy	KNN	RF
Fold1	0.823	0.728
Fold2	0.857	0.782
Fold3	0.81	0.803
Fold4	0.844	0.776
Fold5	0.816	0.789
Fold6	0.823	0.789
Fold7	0.837	0.776
Fold8	0.837	0.789
Fold9	0.837	0.796
Fold10	0.864	0.81
Average	0.835	0.784

Table 2. Results of KNN (a) and RF(b) based models

Average Confusion matrix values: [[193 5] [33 5]]
Average accuracy score : 0.84
Average precision score: [No :0.853, Yes: 0.481]
Average recall score: [No: 0.973, Yes: 0.126]
Average F1 score: [No: 0.909, Yes: 0.196]
Average support score: [No: 198, Yes: 38]
(a)
Average Confusion matrix values: [[179 19] [29 9]]
Average accuracy score: 0.8
Average precision score: [No: 0.861, Yes: 0.333]
Average recall score: [No: 0.903, Yes: 0.249]
Average F1 score: [No:0.882, Yes: 0.281]
Average support score: [No: 198, Yes: 38]
(b)

The present study shows that the KNN-based model exhibits better prediction performance in terms of precision, accuracy, F-score, FPR, and SP in comparison to the RF-based model, Table 2. Table 3 presents a comparison between KNN- and RF-based models developed in the present work and those reported in the literature using the same dataset. It should be mentioned that the number of selected features (5) is low, which could have a negative effect on the performance of the model. Despite that, the developed models in the present work show comparable performance in terms of accuracy or better performance in terms of AUC and F-score when compared with those reported in the literature using a much higher number of features. This can be attributed to the fact that the methodology present in the current research is unprecedented due to the hybridization technique used to tailor the parameters to better represent the model features in a realistic manner. Typically, the KNN algorithm's optimal parameters are obtained through the iteration yielding the least margin of error from multiple trial and error processes. In this paper's KNN algorithm, the trial-and-error process was carried out on the cross-K validation rather than the algorithm itself. This technique consistently finds the optimal K value for a given dataset as opposed to finding the optimal K for a trial and error process Similarly, the RF algorithm conceptually requires a wider span of trial-and-error processes for each parameter. Unlike the typical RF method, this research paper employs the Random Search algorithm to narrow down the range of each parameter, then applies the Grid Search algorithm on the relevant set of parameters only. The conclusion inferred via this methodology is supported by the empirical observations from various experiments on the parameters, especially for the RF, given the specific set of features fulfilling the objective of this research.

Finally, the constructed model based on the KNN algorithm developed in the present work to predict the turnover probability of employees before recruitment could be a useful decision support tool to help the HR managers of the organizations during the recruitment process. This is because the constructed model based on the KNN algorithm reveals high performance to be implemented for a real-life business.



(a)

(b)

Fig. 9 AUC-ROC of KNN algorithm (a) and RF algorithm (b) based models

Table 3. Comparison between the KNN and the RF-based model in the present work and those reported in the literature

Algorithm	Accuracy	AUC	Precision	Recall	F-score	FPR	SP	Ref.
KNN	0.84	0.79	0.47	0.12	0.187	0.025	0.974	present work
RF	0.80	0.82	0.333	0.249	0.281	0.095	0.904	present work
KNN	0.867	-	0.38	0.23	-	-	-	[29]
RF	0.879	-	0.45	0.22	-	-	-	[29]
KNN	0.852	-	0.551	0.09	0.15	-	0.994	[30]
RF	0.861	-	0.658	0.132	0.194	-	0.991	[30]
KNN	0.832	0.52	0.070	0.384	0.119	-	-	[14]
RF	0.85	0.58	0.183	0.619	0.282	-	-	[14]

5. Conclusions

There is a positive correlation between distance from home and employee turnover, signifying that as the distance from home increases, the employee turnover increases. However, there is a negative correlation between age and employee turnover, indicating that as age decreases, employees are more likely to leave the job. The percentage of employee turnover is inversely proportional to the salary. Single employees show a higher desire to leave, but divorced employees are more likely to stay in the job, where the highest turnover percentage occurs in the single employees and the lowest turnover percentage occurs in the divorced employees. Female employees have more tendency to stay in the job. However, male employees have a strong propensity to leave. Salary exhibits the highest feature affecting the turnover of employees and gender has the lowest effect.

The KNN-based model exhibits better prediction performance in terms of accuracy, precision, F-score, FPR, and SP in comparison to the RF-based model.

A data-driven prediction model of the turnover probability of employees before recruitment is constructed to predict the probability percentage of the likelihood of an employee quitting or staying. The constructed model could be a useful decision support tool to help the HR managers during the recruitment process.

6. References

[1] J. Berengueres, G. Duran, D. Castro, "Happiness, an Inside Job? Turnover Prediction Using Employee Likeability, Engagement and Relative Happiness", Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Sydney, Australia, 31 July - 3 August 2017, pp. 509–516.

[2] S. Pandey, P. Khaskel, "Application of AI in Human Resource Management and Gen Y"s Reaction", International Journal of Recent Technology and Engineering, Vol. 8, 2019. pp. 10325-10331.

[3] G. Ginu, M. Thomas. "Integration of Artificial Intelligence in Human Resource", International Journal of Innovative Technology and Exploring Engineering, Vol. 9, 2019, pp. 5069-5073.

[4] S. Kar, S. Srihari, "Impact of Digital HR Practices on Strategic Human Capital Management and Organizational Performance in IT Sector", International Journal of Advance Research in Computer Science and Management Studies, Vol. 6, No. 12, 2018, pp. 17-25.

[5] V. Yawalkar, "A Study of Artificial Intelligence and Its Role in Human Resource Management", International Journal of Research and Analytical Reviews, Vol. 6, 2019, pp. 20-24.

[6] P. Merlin, R. Jayam, "Artificial Intelligence in Human Resource Management", International Journal of Pure and Applied Mathematics, Vol.119, No 14, 2018, pp. 1891-1895.

[7] R. Jesuthasan, "HR's New Role: Rethinking and Enabling Digital Engagement", Strategic HR Review, Vol. 16, No. 2, 2017, pp. 60-65.

[8] R. Rathi, "Artificial Intelligence and the Future of HR Practices", International Journal of Applied Research, Vol. 4, No. 6, 2018, pp. 113-116.

[9] Q. Jia, Y. Guo, R. Li, Y. R. Li, Y. W. Chen, "A Conceptual Artificial Intelligence Application Framework in Human Resource Management", Proceedings of the 18th International Conference on Electronic Business, Guilin, China, 2-6 December 2018, pp.106-114.

[10] C. Chen-Fu, C. Li-Fei, "Data Mining to Improve Personnel Selection and Enhance Human Capital: A Case Study in High-technology Industry", Expert Systems with Applications, Vol. 34, 2008, pp.280-290.

[11] T. Kaczmarek, M. Kowalkiewicz, J. Pikorski, "Information Extraction from CV", Proceedings of the 8th International Conference on Business Information Systems, 20-22 April 2005, Poznan, Poland. pp. 185-189.

[12] H. Zeng, "Adaptability of Artificial Intelligence in Human Resources Management in this Era", International Journal of Science, Vol. 7, 2020, pp. 271-276.

[13] Y. Zhao, M. Hryniewicki, F. Cheng, B. Fu, X. Zhu, "Employee Turnover Prediction with Machine Learning: A Reliable Approach", Proceedings of the Intelligent System Conference, London, UK, 1 September 2018, pp. 737-758.

[14] H. Sri, A. Varaprasad, L. V. N. Sujith, "Early Prediction of Employee Attrition", International Journal of Scientific & Technology Research", Vol. 9, No. 3, 2020, pp. 3374-3397.

[15] P. M. Usha, "An Analysis of the Use of Machine Learning for Employee Attrition Prediction – A Literature Review", Journal of Information and Computational Science, Vol. 10, 2020, pp.1429-1438.

[16] D. Shawni, S. Bandyopadhyay, "Employee Attrition Prediction Using Neural Network Cross Validation Method", International Journal of Commerce and Management Research, Vol. 6, 2016, pp. 80-85.

[17] T. Aniket, D. Motwani, "Employee Churn Rate Prediction and Performance Using Machine Learning", International Journal of Recent Technology and Engineering, Vol. 8, 2019, pp. 824-826.

[18] R. Punnoose, P. Ajit, "Prediction of Employee Turnover in Organizations Using Machine Learning Algorithm", International Journal of Advanced Research in Artificial Intelligence, Vol. 5, No. 9, 2016, pp. 22-26.

[19] R. Jain, A. Nayyar, "Predicting Employee Attrition using XGBoost Machine Learning Approach", Proceedings of the International Conference on System Modeling & Advancement in Research Trends, 23-24 November 2018, pp. 113-120.

[20] J. Sukhadiya, H. Kapadia, M. D'silva, "Employee Attrition Prediction Using Data Mining Techniques", International Journal of Management, Technology and Engineering, 2018, pp. 2249-7455.

[21] S. Ponnuru, G. Merugumala, S. Padigala, R.Vanga, B. Kantapalli,"Employee Attrition Prediction Using Logistic Regression", International Journal for Research in Applied Science and Engineering Technology, Vol.8, 2020, pp. 2871-2875.

[22] H. Zhang, L. Xu, X. Cheng, K. Chao, X. Zhao, "Analysis and Prediction of Employee Turnover Characteristics Based on Machine Learning", Proceedings of the 18th International Symposium on Communications and Information Technologies, Bangkok, Thailand, September 2018, pp. 371-376.

[23] D. Sisodia, S. Vishwakarma, A. Pujahari, "Evaluation of Machine Learning Models for Employee Churn Prediction", Proceedings of International Conference on Inventive Computing and Informatics, Coimbatore, India, 23-24 November 2017, pp. 1016-1020.

[24] C. A. Al Mamun, M. Hasan, "Factors Affecting Employee Turnover and Sound Retention Strategies in Business Organization: A Conceptual View" Problems and Perspectives in Management, Vol. 15, No. 1, 2017, pp.63-71.

[25] D. Verma, R. Chaurasia, "A Study to Identify the Factors Affecting Employee Turnover in Small Scale Industries", International Journal of Engineering Sciences & Research Technology, 2016, pp. 639-652.

[26] N. Govindaraju, "Demographic Factors Influence on Employee Retention", International Journal of Engineering Studied and Technical Approach, Vol. 7, 2018, pp.10-20.

[27] S. Badillo, B. Banfai, F. Birzele, I. Davydov, L. Hutchinson, T. Kam-Thong, J. Siebourg-Polster, B. Steiert, J. Zhang, "An Introduction to Machine Learning", Clinical pharmacology & therapeutics, Vol. 107, No 4, 2020, pp. 871-885.

[28] G. Xiang, J. Wen, C. Zhang, "An Improved Random Forest Algorithm for Predicting Employee Turnover", Mathematical Problems in Engineering, Vol 2019.

[29] I. Onuralp, H. Shourabizadeh, "An Approach for Predicting Employee Churn by Using Data Mining", Proceedings of the International Artificial Intelligence and Data Processing Symposium, Malatya, Turkey, 16-17 September 2017, pp. 1-4.

[30] F. Francesca, M. Coladangelo, R. Giuliano, E. Luca, "Predicting Employee Attrition Using Machine Learning Techniques", Computers, Vol. 9, No. 4, 2020.