Predicting Employee Turnover in High-Tech Enterprises Using Machine Learning: Based on the Psychological Contract Perspective

Yiting Zhang*, Ziling Cai[a], Hongyang Fei[b]

School of Economics and Management, Beijing Jiaotong University, Beijing, 100044, China

*yitingzhang0403@163.com, [a]zlingcai@163.com, [b]3194179127@qq.com

Abstract. High-tech enterprises are boosting technological innovation and economic growth in countries worldwide. Compared with general enterprises, high-tech enterprises are characterized by technology-intensive and high employee turnover rates, relying more on human capital, especially researchers with core technical expertise. However, high turnover rates and unexpected departures of key employees place a huge financial burden on enterprises, along with the risk of technology leakage. Therefore, this study establishes a theoretical model of voluntary employee turnover based on psychological contract theory and previous theoretical studies. We also categorize employee turnover characteristics into four dimensions: Individual conditions, Material incentives, Development opportunities, and Environmental support. Given that previous related studies lacked the combination of theory and data-driven methods, this study applies the IBM HR dataset and selects features for each dimension through the PCA method, for which machine learning models are constructed, including logistic regression, random forests, SVMs, decision trees, and XGBoost, and their performances are evaluated. In addition, the importance of different dimensions is analyzed, and it is found that material incentives have the greatest impact on employee turnover.

Keywords: Employee turnover, Psychological contract, Machine learning, Prediction model

1 Introduction

High-tech enterprises are key to global technological innovation and economic growth. Currently, high-tech enterprises face significant challenges, including severe mismatches in human capital and high rates of talent turnover. Exploring the reasons of employee turnover and how to accurately predict the risk of employee turnover are important issues in enterprise human resource management. Employee turnover refers to the process of losing employees and replacing them with new hires, which is divided into voluntary turnover and involuntary turnover. We focus on voluntary turnover in our study. Compared to general enterprises, high-tech enterprises relying heavily on

J. Liao et al. (eds.), *Proceedings of the 2024 2nd International Conference on Digital Economy and Management Science (CDEMS 2024)*, Advances in Economics, Business and Management Research 292,
https://doi.org/10.2991/978-94-6463-488-4_38

human capital, especially research personnel with core technological expertise. Although some talent mobility indeed plays a role in promoting exchange within the enterprise ecosystem, a high turnover rate and unexpected departure of core employees significantly incur costs related to human capital, recruitment, training, and the development of new employees[1].

In previous, some studies have examined the reasons and effects of employee turnover, with some focusing on aspects like professional identity, job satisfaction, and psychological contracts[2]. The psychological contract, which encompasses the informal expectations between employees and organizations beyond formal contracts, stands out for its ability to explain turnover among high-tech enterprise knowledge workers. Violations of these contracts may have a serious impact on these employees' intention to leave, as unmet expectations may have a negative impact on organizational loyalty and attitudes. However, existing research often separates advanced data-driven techniques from management theory, limiting its practical application in managing turnover[3].

Therefore, this study aims to study voluntary employee turnover caused by psychological contract violation by using machine learning and discuss the most critical influencing factors. Our research will start with the theoretical model of employee voluntary turnover shown in Figure 1. The main contributions are as follows:

1. This study proposes the employee mobility path of high-tech enterprises in detail, emphasizing the influence of individual condition dimensions.
2. We combine machine learning technology with psychological contract theory to study the impact of several aspects of the psychological contract on employee turnover. Using principal component analysis for feature engineering based on the theoretical model. Then use different machine learning methods to predict employee turnover.

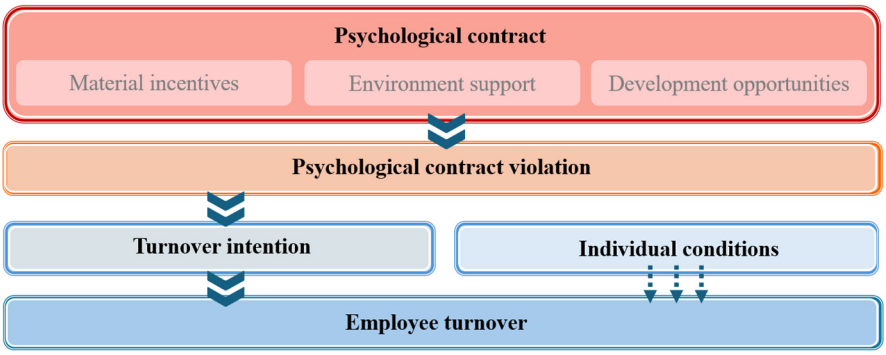

Fig. 1. The Path of Employee Turnover Based on Psychological Contract Violation.

2 Theoretical Preparation

2.1 Psychological Contract and the Causes of Employee Turnover

In today's dynamic market, employees value economic benefits, emotional support, and career growth. Wang[4] et al. introduced a three-dimensional theory for Chinese employees, focusing on material incentives, environmental support, and development opportunities, a view confirmed by Rousseau and others[5]. This study adopts this framework to analyze factors influencing employee turnover.

Material incentives (salary, bonuses, overtime, etc.) play a crucial role in retaining employees. When material incentives are not in line with employee expectations, it will lead to negative emotions such as dissatisfaction. Environmental support includes work environment and interpersonal relationships. Positive environmental support can enhance work focus and loyalty and reduce the risk of turnover. Development opportunities (training, career advancement, etc.) can enhance employee skills and organizational value. The above three dimensions address employees' basic needs, emotional support and self-development from a psychological perspective. Analyzing six recent studies from the past five years reveals three main turnover factors (Table 1). However, most studies focus on specific factors and lack a comprehensive perspective, which may lead to overlooking interactions and similarities between factors. Therefore, our analysis combines the three dimensions of material incentives, environmental support, and development opportunities.

To comprehensively consider the impact of individual differences and corporate practices on employee turnover, we introduce a fourth dimension—individual conditions. An analysis based on six studies showed that 83.33% took into account personal conditions such as gender, age, position and attitude towards the organization(Table 2). Thus, we creates a comprehensive model incorporating material incentives, environmental support, development opportunities, and individual conditions to deeply explore the turnover mechanism in high-tech enterprises.

2.2 PCA in Feature Engineering

Feature engineering can improve data quality and make it more suitable for the model. Pourkhodabakhsh et al. compared various feature extraction methods and it was finally found that the factors identified by the mutual information algorithm showed superior performance in most algorithms[13]. Although effective, it sometimes reduces feature information and affects the accuracy of the model. Principal component analysis (PCA) is a well-known technique for reducing feature dimensionality, which effectively combines data attributes, retains most of the information and improves data representation. The study by Zhu et al. Regarding bearing data[6] and Kong et al. Medical quality assessment[7] demonstrated the effectiveness of PCA in feature fusion, emphasizing its ability to minimize information loss. Therefore, under the guidance of the theoretical model, this paper applies PCA for feature fusion to enhance data analysis.

2.3 Machine Learning Approaches for Predicting Employee Turnover

Machine learning algorithms can learn complex data relationships to effectively predict employee turnover. Random forest was found to outperform other classic algorithms in predicting turnover through confusion matrix evaluation[8]. In terms of practical HR management, Juvitayapu highlighted the excellent performance and interpretability of extreme gradient boosting methods[9]. To tackle imbalanced datasets, Raza et al. introduced the SMOTE technique along with the Extra Trees Classifier (ETC) method, achieving better results than prior approaches[10] . Similarly, the FATPNN model of Xue et al. incorporates a weighted probabilistic loss function for data balancing and shows strong applicability on various datasets. [11].

Our study uses psychological contract theory to define feature dimensions and principal component analysis to extract features. Based on this, we will predict employee attrition by building machine learning models such as including Logistic Regression, Random Forest, SVM, XGBoost, and Decision Trees. The next sections will go over each of the machine learning models used to predict employee attrition in the study.

Table 1. Key factor dimensions statistics in the literature related to employee turnover.

Literature	The top three key factors of importance	Involving dimensions		
		Material incentives	Environment support	Development opportunities
Xia et al[11]	Monthly income, Monthly interest rate, Environmental satisfaction	√	√	
Najafi-Zangeneh et al[12]	Overtime, Job position, Years after the last promotion	√		√
Pourkhodabakhsh et al[13]	Years with the current manager, Work life balance, Distance from home, Years at the company		√	
Raza et al[10]	Monthly income, Hourly wage, Work level, Age	√		√
Chang et al[14]	Job satisfaction, Management style adaptability, Career opportunities		√	√
Srivastava et al[15]	Employee satisfaction, Appraisal rating, Employee CTC level	√		

Table 2. Statistics on Dimensions of Individual Conditions Involved.

Literature	Whether involving individual conditions	Involved proportion
Xue et al[11]	×	
Najafi-Zangeneh et	√	33.33%
Pourkhodabakhsh et	√	50%
Raza et al[10]	√	25%
Chang et al[14]	√	33.33%
Srivastava et al[15]	√	33.33%

3 Methodology

3.1 Analysis Methodology

Figure 2 presents the framework for building and testing models in this study. The success of machine learning predictions heavily depends on data quality, requiring proper data collection, preprocessing, and feature engineering. Given the complexity of employee turnover causes, this study will adjust the data features to better fit the models. To ensure that the model is effective and avoid overfitting, the data is divided into three groups: training set, validation set, and test set. The training set is used for model building, the validation set is used to check the model generalization ability, and the test set is used to evaluate the performance of the best model. The final model selection will be based on accuracy, precision, recall, and F-score.

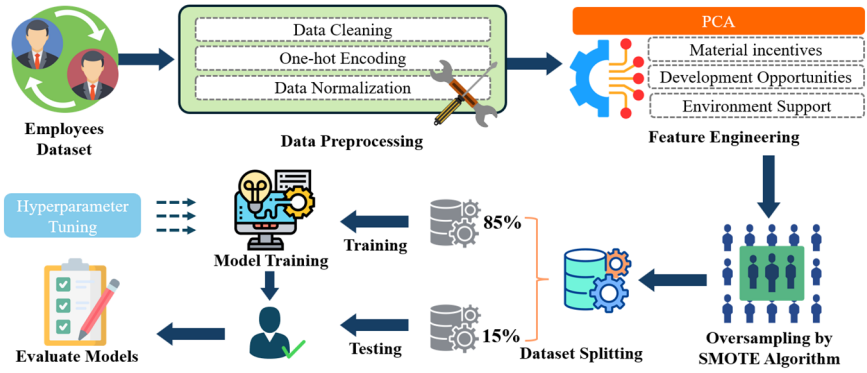

Fig. 2. The experimental framework for building and testing models in this study.

3.2 Data Source and Description

We conduct experiments on a public turnover dataset from IBM[16] published in Kaggle's Competition. This dataset is ideal for research on predicting employee turnover given its high correlation with relevant factors. The IBM HR dataset includes 34

relevant characteristics of personal work and life information from 1470 employees. Specifically, the target field refers to "Attrition", which is an integer value of 0- current employee and 1- former employee.

In our study on predicting employee turnover, we address the complexity of datasets with multidimensional features and diverse correlations by categorizing features into four main dimensions: material incentives, environmental support, development opportunities, and individual conditions, based on the model shown in Figure 1 and detailed in Table 3. By applying Principal Component Analysis (PCA) to each dimension separately, we refine our dataset for more accurate predictions, enabling a broad analysis from a generalized perspective.

3.3 Machine Learning Approaches for Predicting Employee Turnover

Before embarking on feature engineering, it's essential to conduct exploratory data analysis to understand the dataset's structure and distribution. Gender distribution reveals that male and female employees constitute 49.2% and 50.7%. By age distribution, 27.4% of employees are under 30 years old, 63.6% are between 30 and 50 years old, and 8.9% are above 50 years old. Besides, 'EmployeeCount', 'EmployeeNumber', 'Over18', and 'StandardHours' have been identified as irrelevant to employee turnover and are thus removed, keeping 30 relevant features for analysis. Most machine learning algorithms require numerical inputs, categorical variables must also be expressed as numerical data. The categorical features within the dataset are mapped into numerical format via one-hot encoding and label encoding methodologies. In the next step, we normalize the training data using the estimated mean and standard deviation for each variable. This normalization aims to mitigate the impact of features' relative magnitudes on the efficacy of machine learning model training.

3.4 PCA for Feature Engineering

The relationship between the 30 retained attributes is relatively complex, and using them directly for prediction will produce a large amount of data noise, which may adversely affect the prediction performance of the model. In this study, feature engineering is applied to minimize the dimensionality of the features.

Principal Component Analysis (PCA), a method that transforms correlated variables into a set of linearly uncorrelated variables through orthogonal transformations, is used to simplify the high-dimensional data, preserving essential features for predictive modeling. However, PCA's challenge lies in the interpretability of the transformed data[17].

To maintain interpretability, we categorize features into four dimensions: material incentives, environmental support, development opportunities, and individual conditions. Based on theoretical considerations (Figure 1), apply PCA within these dimensions, making sure that features belonging to the same dimension are extracted together as a principal component. This ensures each principal component retains clear interpretive value. Notably, due to the unique and specific nature of each feature, the individual conditions dimension were not affected by PCA and were normalized and included directly in the new dataset. This approach enhances the interpretability of the main

components and provides a comprehensive perspective on how these dimensions influence employee turnover.

3.5 Data Balancing

Due to the imbalance in the ratio of former employees to current employees in our dataset, we employ the SMOTE algorithm for oversampling to generate synthetic instances. SMOTE identifies the nearest neighbors of a small number of samples, calculates the distance between them, and creates new instances by interpolating that distance with the original data.

3.6 Model Selection

This study utilizes several machine learning algorithms to predict employee turnover:

Logistic Regression: A straightforward model for binary classification, predicting turnover by analyzing variables. It's widely used due to its simplicity and efficiency in binary classification problems.

Decision Tree: Uses data-derived decision rules for classification or regression, notable for its clear, interpretable model structure.

Random Forest: An ensemble method that improves prediction accuracy and robustness by combining multiple decision trees.

Support Vector Machine(SVM): Efficiently classifies by finding the best boundary between data classes, suitable for high-dimensional data.

XGBoost: A scalable gradient-boosting algorithm prized for its efficacy in regression and classification tasks. It sequentially constructs decision trees, combining their outputs for precise results.

Table 3. Statistics on Dimensions of Individual Conditions Involved.

Individual conditions	Material incentives	Environment support	Development opportunities
Age	Daily Rate	Environment Satisfaction	Job Involvement
Department	Hourly Rate	Relationship Satisfaction	Job Level
Education	Monthly Income		Training Times Last Year
Education Field	Monthly Rate		Years At Company
Gender	Over Time		Years In Current Role
Num Companies Worked	Percent Salary Hike		Years Since Last Promotion
Total Working Years	Performance Rating		Years With Current-Manager
Work Life Balance	Stock Option Level		Business Travel

Individual conditions	Material incentives	Environment support	Development opportunities
Job Role Marital Status Job Satisfaction Distance From Home			

4 Results and Analysis

This study used two approaches to predict employee turnover: one normalized the data and used all features directly, and the other applied normalization and PCA for feature engineering, based on a turnover model. The PCA-based method outperformed the first and demonstrated higher accuracy compared to similar studies.

4.1 Using the First Approach (Without Feature Engineering)

In this method, after normalizing the dataset and balancing it with SMOTE, all 30 features are retained for model input. Using this approach, SVM outperforms other models with an 87.8% accuracy, followed by Random Forest at 86.9%, XGBoost at 86.4%, Logistic Regression at 75.6%, and Decision Tree at 74.7%. These results are obtained through 10-fold cross-validation and optimal parameter identification via grid search for hyperparameter tuning.

4.2 Using the Second Approach (Using PCA for Feature Engineering)

In this approach, the dataset is standardized, and the 30 features are divided into four dimensions material incentives, environmental support, development opportunities, and individual conditions respectively by principal component analysis for feature selection. Using the extracted principal components to form a new data set as the input of the model reduces the impact of data noise to a certain extent.

Comparing the performance of different machine learning models, the SVM model performed the best among all indicators, such as accuracy (0.951), F1 score (0.953), precision (0.924), and recall (0.984). In the case of high-dimensional datasets, SVM can maximize the inter-category spacing by constructing optimal hyperplanes for efficient data classification. Besides, XGBoost and Random Forest also performed encouragingly, achieving an accuracy rate of 91.9% and 91.1%. The performance comparison of other machine learning models is shown in Table 4.

In this study, we used XGBoost to analyze the impact of various dimensions on employee turnover. XGBoost determines feature importance by evaluating each feature's contribution to model performance, using 'gain' as a key metric to measure accuracy improvements from features in splits. Results show that 'Material Incentives' are most

significant, comprising 26.94% of the total importance, followed by 'Development Opportunities' and 'Environmental Support' at 20.88% and 12.56%, respectively. These findings underscore the importance of career growth and a supportive work environment in predicting turnover. Although the impact of 'Individual Conditions' is relatively minor, it still plays a crucial role. Consistent with previous research, factors like job satisfaction, work-life balance, and lifestyle—categorized under 'Individual Conditions'—are critical in identifying potential turnover, highlighting their significant weight in our analysis. As show in figure 3.

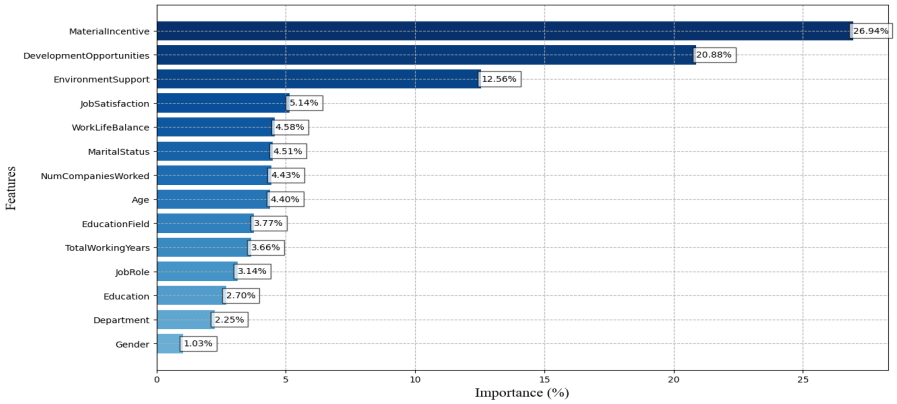

Fig. 3. The importance weights of the features of employee turnover by XGBoost.

Table 4. Performance of each model.

Model	Accuracy	Precision	Recall	F-score
LR	0.768	0.769	0.769	0.769
RF	0.911	**0.923**	0.898	0.910
SVM	**0.925**	0.882	**0.980**	**0.928**
DT	0.808	0.790	0.832	0.810
XGB	0.919	0.933	0.903	0.918

5 Conclusion

We develop a model based on psychological contract theory to study employee turnover in high-tech companies, focusing on material incentives, development opportunities, environmental supports, and personal conditions. Using PCA on 30 features of the IBM HR dataset improved the accuracy of our model, with PCA-enhanced models such as SVM, XGBoost, and Random Forest outperforming traditional methods. This approach is validated by superior performance metrics compared to previous studies, demonstrating the superiority of our dimensional analysis and PCA in minimizing data noise and enhancing data correlation for turnover prediction.

Some studies using the same dataset have employed machine learning to predict employee turnover[18][19]. Table 5 compares the best performing models from previous

studies. Despite differences in data processing and sampling, our model stands out in terms of prediction accuracy and performance metrics, highlighting the effectiveness of dimensional partitioning and PCA in reducing data noise and maintaining relevant data.

In addition, this study provides recommendations on which factors should be considered to predict employee turnover. In the realm of psychological contract dimensions that organizations can address through management and policy adjustments, the 'Material incentives' dimension (overtime, monthly income, etc.) significantly outweighs the 'Development Opportunities' and 'Environmental Support' in terms of its impact on turnover. Therefore, companies should prioritize evaluating and strengthening material incentives in their strategies to reduce employee turnover and give due consideration to employees' feedback in these areas to effectively meet their needs and expectations. Among the identified dimensions, 'Job Satisfaction' and 'Work Life Balance' showed greater influence, in line with other previous studies.

This study presents a machine learning model to predict employee turnover, using PCA for feature selection based on psychological contract theory, thereby enhancing prediction accuracy. However, there are still some limitations. Though we enhance the accuracy of prediction turnover, there is still ambiguity of principal components from dimensions, for instance, like 'Material Incentives', making specific factor impacts, such as overtime versus salary, unclear. Besides, 'Individual Conditions' not fully encompassing all personal factors influencing turnover. Future research could explore more comprehensive factors through qualitative analysis and advanced techniques like deep learning for deeper insights into turnover dynamics.

Table 5. Performance of previous models using the same dataset as this study.

Reference	Best Model	Performance
Guerranti,F. ,& Dimitri G.M. (2022)[20]	Logistic Regression	Accuracy: 0.880 F-score: 0.312 AUC-ROC: 0.850
Alduayj, S. S., & Rajpoot, K. (2018)[18]	Random Forest	Accuracy: 0.914 Precision: 0.925 Recall: 0.893 F-score: 0.909
Fallucchi et al (2020)[19]	SVM	Accuracy: 0.879 Precision: 0.665 Recall: 0.247 Specificity: 0.978 F-score: 0.358
This study	PCA-SVM	Accuracy: 0.925 Precision: 0.882 Recall: 0.980 F-score: 0.928

Acknowledgment

The author would like to thank Professor Wang Xindi from Beijing Jiaotong University for her guidance on this work. This work has received support from Beijing Jiaotong University and data support from the Kaggle platform.

References

1. Alsheref, F.K. , Fattoh, I.E. , & Ead, W.M. (2022). Automated Prediction of Employee Attrition Using Ensemble Model Based on Machine Learning Algorithms. Computational Intelligence and Neuroscience, 2022, 1-9.
2. He, Zhenhua, et al. (2023). How psychological contract violation impacts turnover intentions of knowledge workers? The moderating effect of job embeddedness. Heliyon, 9, 114409.
3. Kurniawaty, K., Ramly, M., & Ramlawati. (2019). The effect of work environment, stress, and job satisfaction on employee turnover intention. Management Science Letters, 9(6), 877–886.
4. Zhu X M, Wang Z M. (2005). A study on the psychological contract structure of knowledge-based employees in the context of China. Scientific Research, 01, 118-122.
5. Rousseau, D. M. , & Tijoriwala, S. A. . (1998). Assessing psychological contracts: issues, alternatives and measures. Journal of Organizational Behavior, 19, 679-695.
6. Zhu, W. , Ni, G. , Cao, Y. , & Wang, H. . (2021). Research on a rolling bearing health monitoring algorithm oriented to industrial big data. Measurement, 185, 110044.
7. Kong, G. , Jiang, L. , Yin, X. , Wang, T. , Xu, D. , Yang, J. ,& Hu, Y. (2018). Combining principal component analysis and the evidential reasoning approach for healthcare quality assessment. Annals of Operations Research,271(2),679-699.
8. Chakraborty, Raj, et al. (2021). Study and prediction analysis of the employee turnover using machine learning approaches. 2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON). IEEE.
9. Juvitayapun, T. (2021). Employee Turnover Prediction: The impact of employee event features on interpretable machine learning methods. 2021 13th International Conference on Knowledge and Smart Technology (KST).
10. Raza, A. , Munir, K. , Almutairi, M. , Younas, F. ,& Fareed, M.M.S. (2022). Predicting Employee Attrition Using Machine Learning Approaches. Applied Sciences, 12(13), 6424.
11. Xue, X. , Sun, X. , Wang, H. , Zhang, H. ,& Feng, J. (2022). Neural network fusion with fine-grained adaptation learning for turnover prediction. Complex & Intelligent Systems, 9(3), 3355-3366.
12. Najafi-zangeneh, S. , Shams-gharneh, N. , Arjomandi-nezhad, A. ,& Zolfani, S.H. (2021). An Improved Machine Learning-Based Employees Attrition Prediction Framework with Emphasis on Feature Selection. Mathematics, 9(11), 1226.
13. Pourkhodabakhsh, N. , Mamoudan, M.M. ,& Bozorgi-amiri , A. (2022). Effective machine learning, Meta-heuristic algorithms and multi-criteria decision making to minimizing human resource turnover. Applied Intelligence, 53(12), 16309-16331.
14. Chang, V. , Mou, Y. , Xu, Q.A. ,& Xu, Y. (2022). Job satisfaction and turnover decision of employees in the Internet sector in the US. Enterprise Information Systems,17(8).

15. Srivastava, P. R. & Eachempati, P. (2021). Intelligent Employee Retention System for Attrition Rate Analysis and Churn Prediction: An Ensemble Machine Learning and Multi-Criteria Decision-Making Approach. Journal of Global Information Management (JGIM), 29(6), 1-29.
16. IBM. IBM HR Analytics Employee Attrition & Performance; Kaggle: San Francisco, CA, USA, 2017.
17. Johnstone, I., & Lu, A. (2009). On Consistency and Sparsity for Principal Components Analysis in High Dimensions. Journal of the American Statistical Association, 104, 682 - 693.
18. Alduayj, S. S., & Rajpoot, K. (2018). Predicting Employee Attrition using Machine Learning. 2018 International Conference on Innovations in Information Technology (IIT), Al Ain, United Arab Emirates, 93-98.
19. Fallucchi, F., Coladangelo, M., Giuliano, R., & William De Luca, E. (2020). Predicting Employee Attrition Using Machine Learning Techniques. Computers, 9(4), 86.
20. Guerranti, F. ,& Dimitri G.M. (2022). A Comparison of Machine Learning Approaches for Predicting Employee Attrition. Applied Sciences, 13(1), 267.