Machines and Algorithms
http://www.knovell.org/mna

Research Article,

Predicting Employee Attrition Using XGB Classifier

Nosheen Aamir [1]

Department of Computer Science, Bahauddin Zakariya University, Multan, 60800, Pakistan
*Corresponding Author: Nosheen Aamir. Email: nosheenamir@gmail.com

Received: 18 June 2025; Revised: 15 July 2025; Accepted: 30 July 2025; Published: 01 August 2025
AID: 004-02-000055

Abstract: Employee attrition is a critical problem in terms of cost and disruption of productivity. Therefore, it is important for organizations to predict which employees are likely to leave. The present paper will be premised on the XGBoost classifier that predicts attrition using the IBM HR Analytics data on Kaggle with 1,470 records of the employees and their demographic, job-related and performance data. The nominal variables were one-hot-coded and the label of the target transformed and stratified during the construction of a training set and a test set assisted in the preparation of the data. Hyper parameter optimization and over sampling techniques as well as feature engineering were chosen to ensure the optimization of the model to deal with the imbalance of the classes. The overall model of the XGBoost was 87.76 and this was reasonable in classifying the employees who remained and those who lapsed. The 'Over Time', 'Monthly Income' and the 'Job Satisfaction' are some of the factors that resulted into high level of impact on attrition. This paper has identified the merits and demerits of machine learning in HR analytics and has uncovered ethical concerns of fairness, transparency and privacy security of employees as applied to the use of predictive models to control the human resource.

Keywords: Machine Learning; Employee Attrition; XGBoost; Binary Classification; Predictive Analytics;

1. Introduction

The use of the data-based decision-making has become one of the cornerstones of the contemporary strategy of the organization and development of the high technology has given the high technology [1,2]. Employee turnover issue is very acute either voluntary in terms of resignation or involuntary by laying off. The high turnover is not merely costly in terms of money, but also it impacts on the cohesion of the group, knowledge transference and long-term productivity [3,4].

Machine learning (ML) has turned out to be an efficient instrument of forecasting employee turnover, identifying the pattern of multidimensional data, and enabling even organizations to engage in the retention [5,6,7]. In particular, the Ensemble algorithm, specifically, the Extreme Gradient Boosting (XGBoost) algorithm could be viewed as one of the most suitable forms of the ML algorithms that could be scaled, more so, were more efficient when used in case of classifications [8,9]. These quantum gains in the average performance of the models can be accomplished by: feature engineering, hyperparam optimization, and hybrid techniques that can enhance the predictive and understandability performance [6].

The problem of employee attrition prediction on XGBoost was one of such binary classification problems that will be discussed in this paper. These are the goals to test the forecasting capacity of the model against the background of the demographic, organizational and performance-based features and to draw the realistic conclusions regarding the HR decision-making. Specifically, the research may be beneficial as it: (i) can be utilized to enhance the quality of the data through a more efficient preprocessing and feature engineering technology; (ii) can examine the outcomes of XGBoost on the base of accuracy, precision, recall, and F1-score; and (iii), can evaluate the importance of the features to be able to present the most significant drivers of attrition that would work in practice by providing practical guidance to the HR specialists [8].

The paper is going to be structured as follows: Section 2 will entail the literature review of the machine learning application to employee attrition prediction. In section 3, the description of the preprocessing steps, method of data description and data set description are provided. Experimental results and discussion are found in section 4. The most significant conclusions, weaknesses, and recommendations are presented in section 5.

2. Literature Review

Predictive capabilities of machine learning (ML) have been massively researched in recent years, and one of the areas is the prediction of employee attrition. The summary of the key investigations that have been performed in the last years of 2021-2025 and the methodology, conclusions, and their relevance to this study will be assessed in this section.

The usefulness of ML has been established to be effective in addressing the HRM issues particularly in estimating employee turnover [1, 2]. Since it has been discovered that HRM predictive analytics helps in improving the decision-making process, it is applied to guide retention and effective operations [3]. The study points out that the demographic characteristics of the age, the time when the job was taken and the job satisfaction are among the most significant influencing the employee turnover [4].

As it is common knowledge, the attrition can be predicted with the use of a variety of ML algorithms. As indicators, one of the researches [5] compared the performance of the ensemble process, and among the others, the XGBoost and random forest provided the following results: the accuracy that could predict an accuracy rate of 89-percent of the data of 20 features using the XGBoost. The deep learning techniques have also achieved as much as 92 per cent but are usually difficult in terms of the contention of the models [6]. By contrast, simple and logistic regression models occupy a relatively desirable underdog position since it is easy to operate, and it is competing with the other two in the accuracy of 85% [7].

Another characteristic of the models that is significant to reinforce is preprocessing and preselection. Selection of component of influential features is applicable in order to optimize the precision of forecast and reduce the complexity of calculations as it was proven by [8]. In the same argument, it was determined in a comparative study that decision tree models were more suitable than Naive Bayes which provided 82.7 percent accuracy of percentage split evaluation and logistic regression which also provided 85 percent accuracy [10] at percentage split evaluation.

Despite these advances, the strategies also have certain loopholes in dealing with the problem of the unequal distribution of classes and its progressively interpretable character among HR practitioners. The modern research endeavors to overcome these shortcomings by applying the XGBoost with a more refined preprocessing, hyperparameter optimization, and estimation of the importance of the features with the aim of developing the right forecasts and practical knowledge of variables influencing the staff turnover.

3. Methodology

This section establishes the research methodology that comprises data set, pre-processing, type of model and the measures of evaluation. The framework is formulated in a manner that is strong and they can be reproducible and to learn what factors lead to employee attrition. The steps of the process were computed in such a way that it would provide maximum accuracy to the model and it would not be overly complicated to understand by the HR practitioners. In addition, the methodology is associated with openness and the research can be replicated and generalized in the future.

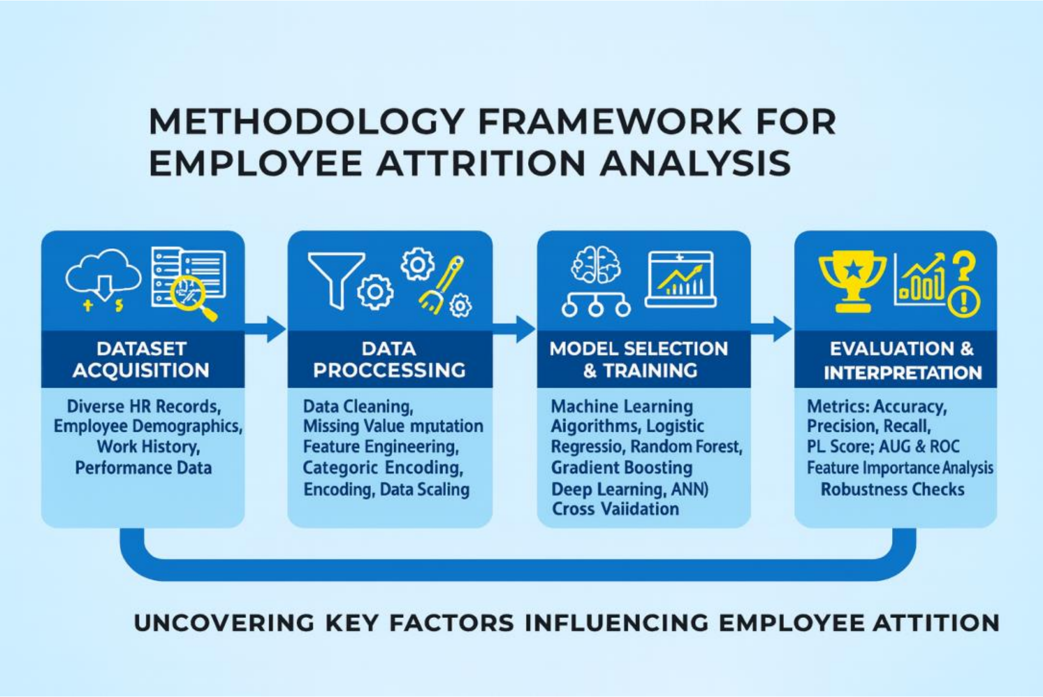
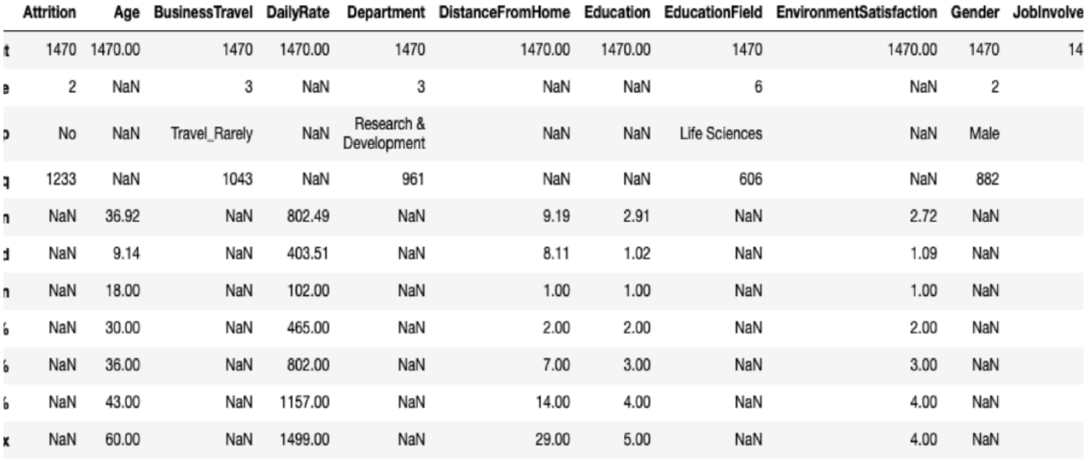

METHODOLOGY FRAMEWORK FOR EMPLOYEE ATTRITION ANALYSIS
DATASET ACQUISITION
Diverse HR Records, Employee Demographics, Work History, Performance Data
DATA PROCCESSING
Data Cleaning, Missing Value mṛutation Feature Engineering, Categoric Encoding, Encoding, Data Scaling
MODEL SELECTION & TRAINING
Machine Learning Aigorithms, Logistic Regressio, Random Forest, Gradient Boosting Deep Learning, ANN) Cross Vaiidation
EVALUATION & INTERPRETATION
Metrics: Accuracy, Precision, Recall, PL Score; AUG & ROC Feature Importance Analysis Robustness Checks
UNCOVERING KEY FACTORS INFLUENCING EMPLOYEE ATTITION

Figure 1: Phases of analysis process

3.1. Dataset

The dataset that was used in this paper is the IBM HR Analytics Employee Attrition and Performance dataset that was downloaded in the Kaggle repository. The data set will contain 1470 data records on workers that are demographic, organizational and performance based. Some of the valuable features include age, sex, department, job position, overtime, monthly earnings and job satisfaction. The target variable is the attrition that is a binary variable (Yes = 1, No = 0) meaning that the employee has quit the organization or not [1].



	Attrition	Age	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EnvironmentSatisfaction	Gender	JobInvolve
t	1470	1470.00	1470	1470.00	1470	1470.00	1470.00	1470	1470.00	1470	14
e	2	NaN	3	NaN	3	NaN	NaN	6	NaN	2	
p	No	NaN	Travel_Rarely	NaN	Research & Development	NaN	NaN	Life Sciences	NaN	Male	
q	1233	NaN	1043	NaN	961	NaN	NaN	606	NaN	882	
n	NaN	36.92	NaN	802.49	NaN	9.19	2.91	NaN	2.72	NaN	
d	NaN	9.14	NaN	403.51	NaN	8.11	1.02	NaN	1.09	NaN	
n	NaN	18.00	NaN	102.00	NaN	1.00	1.00	NaN	1.00	NaN	
%	NaN	30.00	NaN	465.00	NaN	2.00	2.00	NaN	2.00	NaN	
%	NaN	36.00	NaN	802.00	NaN	7.00	3.00	NaN	3.00	NaN	
%	NaN	43.00	NaN	1157.00	NaN	14.00	4.00	NaN	4.00	NaN	
x	NaN	60.00	NaN	1499.00	NaN	29.00	5.00	NaN	4.00	NaN	

Figure 2: Dataset descriptive statistics

3.2. Data Preprocessing

To ensure data quality and reliable model performance, several preprocessing steps were applied:

3.2.1. Handling Missing Values

Missing values in rows were eliminated in order to ensure consistency and eliminate the chances of bias during model training. Even though data imputation processes can be used to enhance the completeness of a dataset, the research in this study chose to omit the rows because of the low percentage of omissions, thus reducing the possibilities of distortion of data [6].

3.2.2. Encoding Categorical Variables

Categorical variables, such as gender, department, job role and overtime, were encoded using one-hot encoding methods to assign them numbers. Such a strategy facilitated the coding of labels whereby label encoding of the target variable Attrition was applied i.e. the yes answer was coded as 1 and the no answer was encoded as 0 which puts the variable in a binary format that can be utilized in classification processes [2].

3.2.3. Train-Test Split

To test the possibility of the generalization, the data was separated into training and testing (20 and 80) sets. The stratified sampling was used to even out the classes of the non-attrition and attrition cases. It was also seeded randomly to enhance comparability and reproducibility or experiment [5, 8].

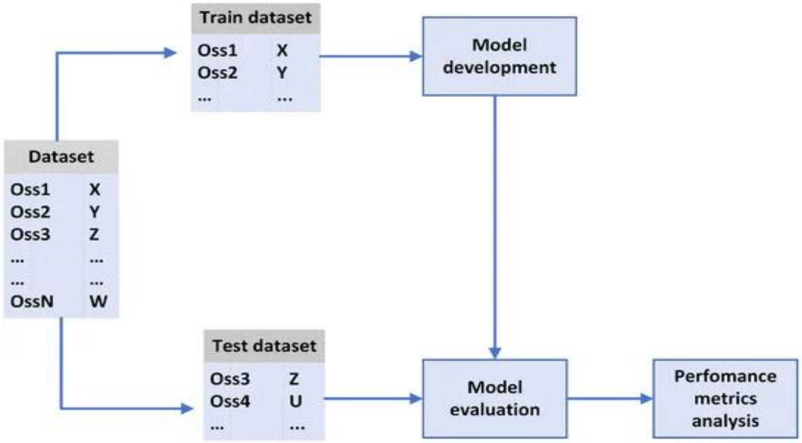


Figure 3: Data preprocessing

3.2.4. Model Selection

The use of Extreme Gradient Boosting (XGBoost) classifier as the primary model is due to its efficiency, scaling, and imbalance data features [2,5]. The XGBoost has already been applied in the area of employee attrition, and the research has demonstrated that it is a great predictor and that it can succeed in a broad spectrum of organizational data [3,6].The XGBoost has been compared with the Logistic Regression, Decision Trees and Random Forests, and Support Vector Machines (SVMs), which were also considered in the studies [3,5,7]. The second comparative framework ensures that the predictive capability of XGBoost is validated against the other simpler and established classifiers and enhances the legitimacy of the model choice [1,3,5].

3.2.5. Model Training Hyperparameter Tuning

The initial parameters used to train XGBoost classifier were default. Significant hyperparameters including the learning rate, max depth, the number of estimators and subsample were optimized using grid search and cross validation in order to have more predictive accuracy and reduced overfitting [2,5]. In the training of a model, feature importance scores were also acquired in order to find out the most important predictors of employee departure. It is also interesting to note that the characteristics such as OverTime, JobSatisfaction and MonthlyIncome were significant and yielded useful data on the organizational variables that had the strongest relationships with the employee turnover [6,11,12].

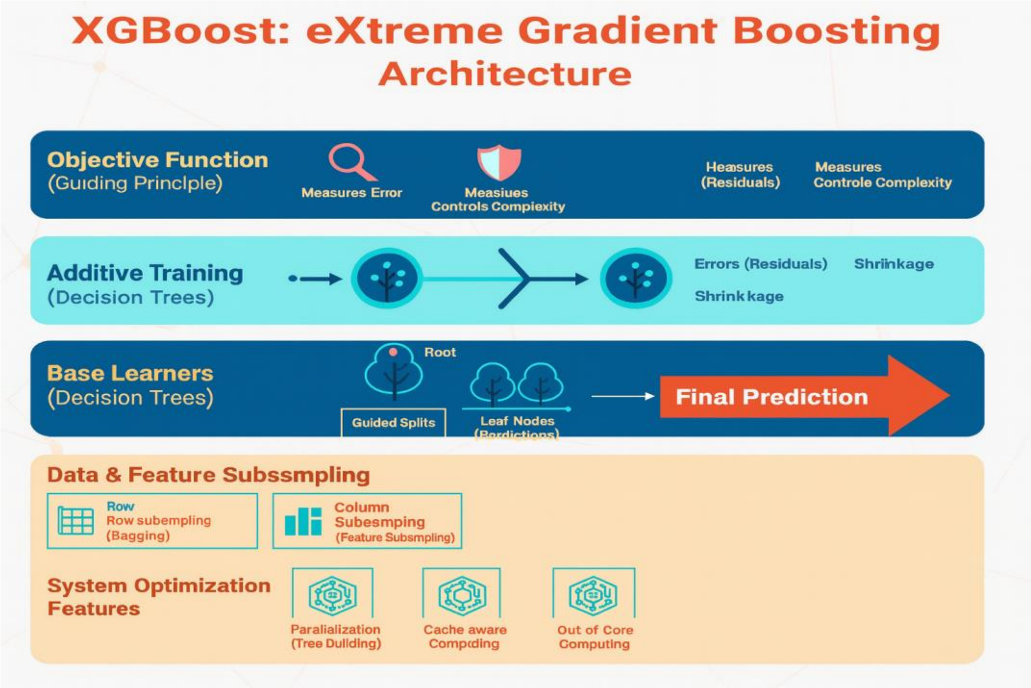


Figure 4: XGBoost Architecture

3.2.6. Feature Importance Analysis

Parameters learned by XGBoost classifier include the following: learning rate, maximum depth, the number of estimators and sub sample available [2,5] with default parameters and trained on grid search and cross-validation. The Python packages such as Scikit-learn, XGBoost API, and Pandas made it possible to ensure the reproducibility of the results because the feature importance was checked on the basis of the OverTime_Yes, JobSatisfaction, and MonthlyIncome as important predictors of attrition. The greater the working hours the greater the risk of turnover, the greater the job satisfaction the lesser the risk of turnover, the greater the competitive income the greater the retention [6,11,12]. These are the ones that are consistent with the previous studies and can be implemented in HR activities like managing workloads, job satisfaction programs, and fair remunerations.

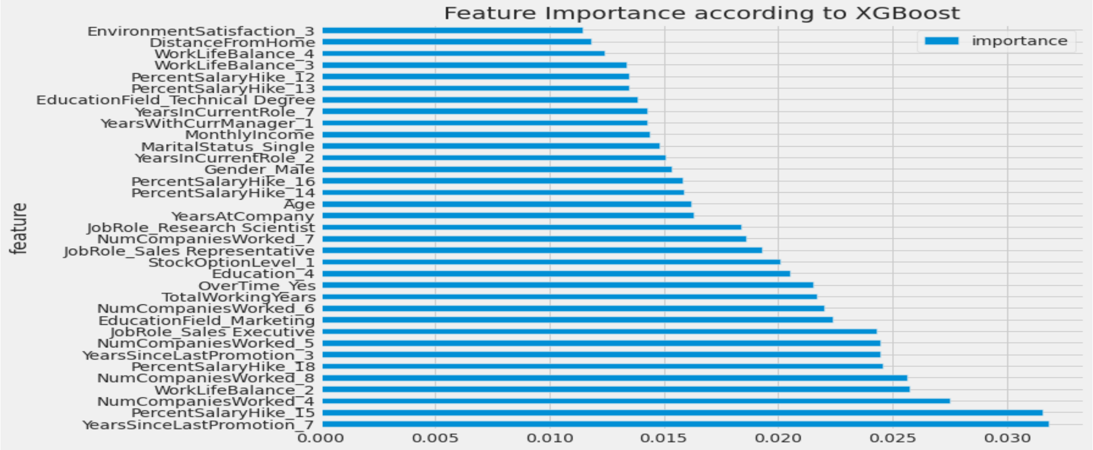


Figure 5: Feature importance score

3.2.7. Evaluation Metrics

In order to check the stability of the XGBoost model and practicability of the model, the following standard measures of classification were used:

- **Accuracy:** The mean percentage of appropriately classified instances that gives a relative estimate of the model performance.
- **Precision:** This metric is evaluated as a measurement of how the model predicts positive instances and also measures the false positives.
- **Recall:** This is an evaluation used to determine how the model predicts all the true positive instances.
- **F1-Score**: Gives a more specific picture of precision and remembrance, and especially with unbalanced data.
- **Confusion matrix:** A row and a column in a confusion matrix will depict true positives, true negatives, false positives and false negatives.

4. Results and Discussion

The high performance of the model was exhibited by the classifier having the training accuracy of 99.87 percent and the test accuracy of 87.76 percent, which was trained using XGBoost. The large value of the training accuracy indicates that the model is doing well in learning patterns with the help of the training data, but the rather small value of the testing accuracy indicates the difficulty in transferring the results to the data that is not observed. The measures of evaluation are condensed in Table 1.

Metric	**Training Set (%)**	**Testing Set (%)**
Accuracy	99.87	87.76
Precision (Class 0)	99.94	90.00
Recall (Class 0)	99.91	97.00
F1-Score (Class 0)	99.92	93.00
Precision (Class 1)	99.70	59.00
Recall (Class 1)	99.80	26.00
F1-Score (Class 1)	99.75	36.00

Table 1: Model Evaluation Metrics

5. Conclusion

As analyzed in this paper machine learning and XGBoost, in particular, will prove handy when predicting employee attrition. The model provided a training and testing ratio of 99.87 and 87.76 respectively and the important predictors in the model were OverTime_Yes, JobSatisfaction and MonthlyIncome among others.

Even though the overall performance is high, however, there is an issue with anticipating the minority class (those who leave) and it is founded on the necessity to regulate the imbalance in classes in future employment. Finally, the study is noteworthy because it demonstrates the possibility of changing the employee retention process with the help of machine learning to present the primary causes of turnover and the opportunity to make active decisions of the human resources, as well as to address ethical concerns, including data privacy, and equity.More research is needed to understand how the model will help in other fields, use modern tools, including deep learning or ensemble models, and rationalize ethical considerations, including data privacy, and fairness.

Funding Statement: Author has not received any funds from external source.

Conflicts of Interest: Author has no conflicts of interest.

Data Availability: The IBM HR Analytics dataset used in this study is available publicly.

References

[1] Analytics Vidhya. (2021). Employee Attrition Prediction - A Comprehensive Guide. Retrieved from https://www.analyticsvidhya.com/blog/2021/11/employee-attrition-prediction-a-comprehensive-guide/

[2] Fallucchi, Francesca, Marco Coladangelo, Romeo Giuliano, and Ernesto William De Luca. "Predicting employee attrition using machine learning techniques." *Computers* 9, no. 4 (2020): 86.

[3] Raza, Ali, Kashif Munir, Mubarak Almutairi, Faizan Younas, and Mian Muhammad Sadiq Fareed. "Predicting employee attrition using machine learning approaches." *Applied Sciences* 12, no. 13 (2022): 6424.

[4] Patil, Harsh, and Prabha Kadam. "Machine Learning Applications in Human Resource Management: Predicting Employee Turnover and Performance." *The Voice of Creative Research* 7, no. 2 (2025): 295-301.

[5] Alshiddy, Muneera Saad, and Bader Nasser Aljaber. "Employee attrition prediction using nested ensemble learning techniques." *International Journal of Advanced Computer Science and Applications* 14, no. 7 (2023).

[6] Sari, Sindi Fatika, and Kemas Muslim Lhaksmana. "Employee attrition prediction using feature selection with information gain and random forest classification." *Journal of Computer System and Informatics (JoSYC)* 3, no. 4 (2022): 410-419.

[7] Ponnuru, S. R., G. K. Merugumala, Srinivasulu Padigala, Ramya Vanga, and Bhaskar Kantapalli. "Employee attrition prediction using logistic regression." *International Journal for Research in Applied Science and Engineering Technology* 8, no. 5 (2020): 2871-2875.

[8] Ali, Zeravan Arif, Ziyad H. Abduljabbar, Hanan A. Tahir, Amira Bibo Sallow, and Saman M. Almufti. "eXtreme gradient boosting algorithm with machine learning: A review." *Academic Journal of Nawroz University* 12, no. 2 (2023): 320-334.

[9] Pristyanto, Yoga, Zulfikar Mukarabiman, and Anggit Ferdita Nugraha. "Extreme gradient boosting algorithm to improve machine learning model performance on multiclass imbalanced dataset." *JOIV: International Journal on Informatics Visualization* 7, no. 3 (2023): 710-715.

[10] Usha, P. M., and N. V. Balaji. "A comparative study on machine learning algorithms for employee attrition prediction." In *IOP Conference Series: Materials Science and Engineering*, vol. 1085, no. 1, p. 012029. IOP Publishing, 2021.

[11] Al-Suraihi, Walid Abdullah, Siti Aida Samikon, Al-Hussain Abdullah Al-Suraihi, and Ishaq Ibrahim. "Employee turnover: Causes, importance and retention strategies." *European Journal of Business and Management Research* 6, no. 3 (2021): 1-10.

[12] Tnay, Evelyn, Abg Ekhsan Abg Othman, Heng Chin Siong, and Sheilla Lim Omar Lim. "The influences of job satisfaction and organizational commitment on turnover intention." *Procedia-Social and Behavioral Sciences* 97, no. 201-208 (2013): 3-8.