FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 9 No. 1, January, 2025, pp 205 - 210
DOI: https://doi.org/10.33003/fjs-2025-0901-3030

ENHANCING EMPLOYEE ATTRITION PREDICTION: THE IMPACT OF DATA PREPROCESSING ON MACHINE LEARNING MODEL PERFORMANCE

Muhammad Garba, *Musa Usman and Muhammad Saidu

Department of Computer Science, Faculty of Physical Sciences, Kebbi State University of Science & Technology, Aliero. Nigeria.

*Corresponding authors' email: usman.musa91@gmail.com

ABSTRACT

Organizations face a serious problem with employee attrition, which raises expenses and reduces productivity. This study looks at how preprocessing data can help machine learning models forecast employee turnover more accurately. Seven machine learning algorithms—Random Forest, k-Nearest Neighbors (k-NN), XGBoost, Gradient Boosting, Linear Discriminant Analysis (LDA), LightGBM, and Logistic Regression—were used to analyze the 1,470 records in the International Business Machines Human Resources (IBM HR). Employee Attrition dataset. SimpleImputer was used to handle missing values, StandardScaler was used to standardize numerical features, and SelectFromModel was used to choose important features. These actions were essential in improving the accuracy of the model; LDA had the highest accuracy of 87.38%, followed by LightGBM and Logistic Regression, both of which had 87% accuracy. All models' performance metrics were much enhanced by preprocessing; k-NN had the lowest accuracy, at 85.33%. These results demonstrate how important preprocessing is to predictive analytics and how HR management may use it to identify at-risk workers and create successful retention plans.

Keywords: Employee Attrition, Machine Learning, Predictive Analytics, Data Preprocessing, HR Management

INTRODUCTION

The most productive and active research fields are machine learning and data mining. These topics have wide-ranging applications in various industries, including banking, healthcare, education, mobile gaming, security systems, and human resource management. They use a variety of approaches for categorization, clustering, and prediction (Alsheref, Fattoh, & Ead, 2022).

Employee attrition is a significant challenge for organizations in today's rapidly evolving business landscape (Kashyap & Kriti, 2018). Elevated staff attrition not only drives up hiring expenses but also impedes total output and long-term expansion. Organisations are increasingly turning to data-driven solutions to combat employee turnover as the talent war heats (De Winne et al., 2019). By harnessing the capabilities of machine learning, employee attrition has transcended the confines of HR concerns to become a critical business challenge influenced by distinct factors unique to each individual's decision to leave (Marín Díaz et al., 2023).

Human resource management plays a central role in any organization (Najafi-Zangeneh et al., 2021). It encompasses functions such as planning, organizing, staffing, directing, controlling, recruiting, placement, performance appraisal, compensation, and training. Over time, technological advancements have ushered in an era of copious HR-related data, leading to a shift toward data-driven decision-making (Sukmo & Nugroho, 2022). In the realm of human resource management, big data analytics has emerged as an essential tool for achieving superior, cost-effective outcomes.

Studies employing mixed-methods approaches, such as Jayanthi and Prabu, combine quantitative data with qualitative insights to explore employee attrition factors and their organizational impact. Such approaches offer practical recommendations but often fall short in providing detailed data or incorporating contemporary industry trends. Similarly, feature selection techniques in machine learning studies, like those by Sari and Lhaksmana (2022) highlight the effectiveness of specific algorithms, including Recursive Feature Elimination and Random Forest. However, many studies lack robust discussions on data preprocessing, limitations, and biases, which are critical for model reliability and interpretability.

Several studies extend the scope of attrition research by integrating age-related factors and AI techniques. For example, Yousuf Khan (2019) analyzed generational differences in attrition tendencies, while Kapila and Pathak (2018) proposed using ensemble classification and linear regression for proactive management. Despite offering actionable insights, these studies often lack comprehensive literature reviews, detailed methodologies, and discussions on feature engineering. Overall, while the field has made significant strides, future research should address methodological gaps, explore novel approaches, and incorporate diverse evaluation metrics to advance employee attrition prediction and management.

Wardhani and Lhaksmana, (2022)conducted a study on employee attrition prediction. They compared feature selection methods, including information gain, select k-best, and recursive feature elimination, using 10-fold cross-validation. Their findings revealed that recursive feature elimination with 20 selected features achieved the best results, with an accuracy of 0.853 and an AUC score of 0.925. This highlights the significance of machine learning in predicting employee attrition and assisting HR managers in reducing attrition rates. However, the work Lacks a thorough discussion on limitations and potential biases.

Raza et al., (2022)conducted a study on predicting employee attrition using machine learning models with an IBM Watson dataset. They compared SVM, random forest, and KNN models while exploring oversampling and under-sampling techniques to address class imbalance. The research successfully identified significant attrition-related features but lacked comprehensive data information and in-depth result analysis, and also Primarily focused on the F1-score as an evaluation metric, missing the opportunity to include other metrics like ROC-AUC for a more thorough assessment.

Sethy and Kumar Raut, (2022.) conducted a study on predicting employee attrition rates using various machine

learning algorithms based on HR analytics data. The research compared classifiers such as Naive Bayes, Logistic Regression, KNN, SVM, and Random Forest to identify the most accurate predictor. The study found that the Random Forest model achieved the highest accuracy of 85% and suggested a possible association between effort-reward imbalance and employee attrition. However, the work has Limited originality, primarily relying on existing machine-learning techniques without introducing novel approaches.

Davidson and Brindha, (2021)conducted a study on employee attrition in star-category hotels in Tamilnadu, focusing on gender and job designation factors. They employed a descriptive research design and structured questionnaires to gather data from 419 participants. The study found that salary, incentives, and benefits significantly affected male employee attrition, while work-life balance was crucial for females. Attrition drivers differed based on job levels, with top-level employees leaving due to work-life balance, lower-level employees due to salary issues, and middle-level employees due to better opportunities elsewhere. However, limitations included the absence of theoretical frameworks for predictive analytics and personalized retention strategies, reliance on traditional statistical analyses, and the lack of mention of guiding theoretical models.

A study conducted by Pratt et al., (2021)focused on employee attrition prediction using machine learning. They compared six algorithms and found Random Forest to be the most accurate, achieving an 85.12% accuracy rate. The research emphasized understanding attrition factors for organizational retention. However, the study had Limited evaluation metrics, relying mainly on accuracy.

Usha and Balaji, (2021) conducted a study exploring the use of machine learning algorithms to predict employee attrition. They compared the performance of Naïve Bayes, Decision tree, Random Forest, and K-Means, with Naïve Bayes achieving the highest accuracy, while J48 and Random Forest performed best in terms of F1 scores. The research also identified factors strongly correlated with employee retention, including job security, promotion policies, and growth opportunities. However, the study lacked a literature review to contextualize its findings within existing knowledge, and it omitted a discussion of limitations and suggestions for future research.

MATERIALS AND METHODS

Research Design

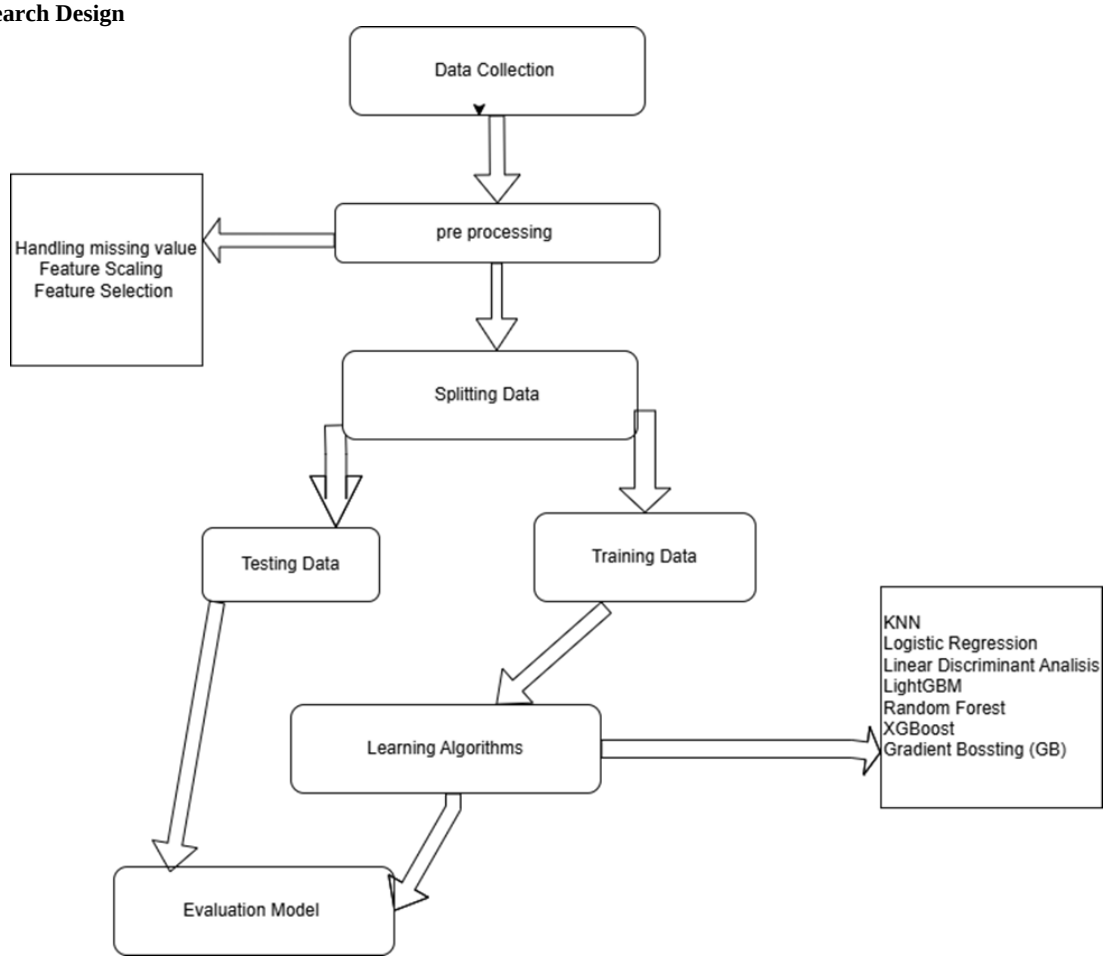

Figure 1: Research Design

This diagram outlines the research process and methodology used in the study. It includes key phases such as data collection from the IBM HR Employee Attrition dataset, preprocessing steps (handling missing values, feature scaling, and feature selection), and the application of machine learning models for predictive analysis. The diagram also illustrates the performance evaluation phase, highlighting the metrics used to assess the accuracy of the models. This visual representation provides a step-by-step overview of the study's approach to examining the impact of data preprocessing on model performance.

Data Collection

The dataset used in this study was obtained from the IBM HR Employee Attrition dataset, consisting of 1470 records for employees, including features such as age, tenure, job satisfaction, department, and whether the employee left the company. The dataset included both numerical and categorical variables. Table 1 describes the attributes (Name and Data type).

The structure of the IBM HR Employee Attrition dataset utilized in the research is shown in this table. With clear column headers identifying each attribute's name and data type (e.g., numerical or categorical), it contains important attributes including age, tenure, job satisfaction, department, and attrition status. These qualities are necessary for properly implementing machine learning models and assessing employee turnover. Their value resides in identifying variables that could affect an employee's decision to remain or depart.

Table 1: Dataset Attribute

Attribute Name	Data Type	Attribute Name	Data Type
Age	INT	Maritalstatus	Object
Attrition	Object	Monthlyincome	INT
Businesstravel	Object	Monthlyrate	INT
Dailyrate	INT	Numcompaniesworked	INT
Department	Object	Over18	Object
Distancefromhome	INT	Overtime	Object
Education	INT	Percentsalaryhake	INT
Educationfield	INT	Performancerating	INT
Employeecount	INT	Relationshipsatisfaction	INT
Employeenumber	INT	Standardshour	INT
Environment Satisfaction	INT	Stockoptionlevel	INT
Gender	Object	Totalworkingyears	INT
Hourlyrate	INT	Trainingtimelastyear	INT
Jobinvolvement	INT	Worklfebalance	INT
Joblevel	INT	Yearinthecompany	INT
Jobrole	Object	Yearsincurrentrule	INT
Jobsatisfaction	INT	Yearssincelastpromotion	INT
Yearwithcurrentmanager	INT		

The bar chart below represents employee attrition statistics where:
"Yes" corresponds to employees who have left (attrited).
"No" corresponds to employees who have not left (retained).

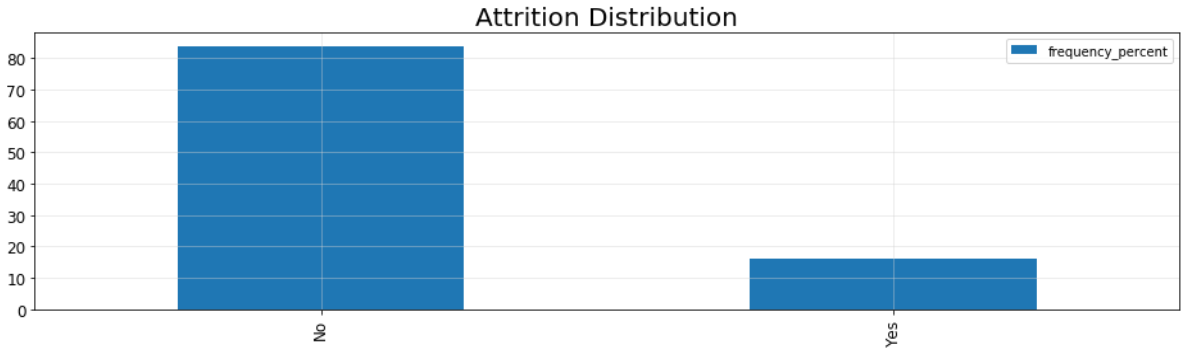

Figure 2: Attrition Distribution

The distribution of the dataset's employees according to their attrition status is shown in this bar chart. Employees who left the company (attrited) are represented by the "Yes" category, whilst those who remained (retained) are represented by the "No" category. The percentage of each group is shown in the graphic, which offers a fundamental understanding of the class imbalance in the dataset. This disparity emphasizes how crucial data preparation methods—like addressing class imbalance—are to obtaining precise predictive modeling.

Data Preprocessing

An essential step in getting the dataset ready for machine learning algorithms was data preparation. The following steps were included in this phase: Before processing Information, It is advised that you perform data pre-processing before using any machine learning algorithms on a dataset. Cleaning up the data before doing more analysis is known as pre-processing. It involves several procedures, including addressing missing values, feature extraction, and feature selection(Ahmad et al., 2023).

Handling Missing Values

SimpleImputer from the sklearn.impute module in Python was used to address missing data in the dataset by employing distinct imputation strategies for numerical and categorical variables. For numerical columns, missing values are replaced with the mean of the respective column, ensuring that all numeric data is retained. For categorical columns, missing

values are imputed using the most frequent value (mode) within each column, preserving categorical consistency. Once imputation is completed, the dataset is divided into predictor variables (X) and the target variable (Attrition). This method ensures that the dataset remains complete and suitable for further analysis or model development (Batista & Monard, 2003).

Feature Scaling

The term "feature scaling" refers to the approaches or strategies used to arrange the feature value range within a similar scale or to normalize the range of independent variables in our data (Ahmad et al., 2023). In this research study, we apply standardization to the numerical features using StandardScaler(). It scales the data to have a mean of 0 and a standard deviation of 1, ensuring that all numerical columns are on the same scale. This is important for models that are sensitive to feature scales, improving model efficiency and performance. The fit_transform() function calculates the mean and standard deviation for each column and then transforms the data accordingly.

Feature Selection

A key component of machine learning's data preparation and model construction is featuring selection. It entails selecting a subset of the original collection of features' most important traits (attributes) (Ahmad et al., 2023). We used SelectFromModel to perform feature selection with a random forest classifier. It initializes the classifier with 100 trees and uses the median feature importance to determine which features to keep. It then fits this model to the data and transforms the feature set to include only the selected features based on their importance.

Classification Algorithms

Seven machine learning algorithms were employed to predict employee attrition. The choice of algorithms was based on their proven effectiveness in classification tasks and their ability to handle complex datasets. The following models were trained and evaluated:

Random Forests: An ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes (classification) (Punnoose & Xavier, 2016).

k-Nearest Neighbor (k-NN): A simple algorithm that stores all available cases and classifies new cases based on a similarity measure (Cunningham & Delany, 2021).

XGBoost: An optimized gradient boosting framework that uses tree-based learning algorithms (Arif Ali et al., 2023).

Gradient Boosting (GB): A regression technique that builds models sequentially, each new model correcting errors made by the previous ones (Otchere et al., 2022).

Linear Discriminant Analysis (LDA): A statistical method used to find a linear combination of features that best separates two or more classes of objects (Rahamneh et al., 2023).

LightGBM: A gradient-boosting framework that uses tree-based learning algorithms, designed for high-performance and distributed systems (Ke et al., 2017).

Logistic Regression: is a fundamental predictive modelling technique widely used in HR analytics (Ponnuru, 2020).

Performance Evaluation Matrics

A confusion matrix was used to assess the research's performance. A visual table called a confusion matrix is used to evaluate the effectiveness of a classification algorithm. A confusion matrix is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known. It allows visualization of the performance of an algorithm by comparing predicted values with actual values (Xu et al., 2020). In the matrix, the actual target instances and the anticipated cases are represented by each column and row, respectively. The following table displays the confusion matrix along with the computed recalls, accuracy, and precision.

Table 2: Performance Evaluation Matrix

True positive	False Negative
False Positive	True Negative

True Positive (TP): Predicted positive and positive.
True Negative (TN): Predicted negative and negative.
False Positive (FP): Predicted positive but negative (Type I error).
False Negative (FN): Predicted negative but positive (Type II error).

RESULTS AND DISCUSSION

Performance Results for the classifier

This section provides a detailed analysis of the experimental outcomes obtained by applying data mining techniques to the IBM HR Analytics dataset. The study utilized seven selected machine learning algorithms: Random Forest, k-Nearest Neighbors (k-NN), Logistic Regression, XGBoost, Gradient Boosting Classifier (GBC), Linear Discriminant Analysis (LDA), and LightGBM. These algorithms were chosen for their robust classification capabilities and diverse approaches to handling data structures.

The dataset used in this study comprised 1,470 instances and 35 attributes. Proper preprocessing techniques were applied to ensure data quality, including handling missing values, addressing class imbalance, feature selection, and feature scaling. The training and testing of the models revealed the following accuracies:

Logistic Regression: 87%
LDA: 87.38%
k-NN: 85.33%
LightGBM: 87%
GBC: 86%
XGBoost: 86%
Random Forest: 86%

These results highlight that LDA emerged as the best-performing algorithm, achieving an accuracy of 87.38%, while k-NN recorded the lowest accuracy of 85.33%. Figure 3 summarizes the performance of the seven algorithms, showcasing the effectiveness of proper preprocessing techniques in achieving competitive results.

```
final = pd.DataFrame(model_metrics,
                     columns = ['Accuracy', 'APS', 'F1', 'Recal', 'Auc'],
                     index = model_name)
final.sort_values(by = 'Accuracy', ascending = False)
```

	Accuracy	APS	F1	Recal	Auc
LogRegression	0.874505	0.689696	0.509396	0.408084	0.686024
LDA	0.873759	0.680429	0.510787	0.412957	0.687559
Lgbm	0.871495	0.715239	0.458633	0.342525	0.657749
GBC	0.866958	0.708576	0.425295	0.309856	0.641415
Xgboost	0.863934	0.663156	0.445270	0.342746	0.653355
RandomForesst	0.861672	0.868398	0.227618	0.154817	0.569025
KNN	0.853362	0.712987	0.249716	0.154485	0.570936

Figure 3: Result performance of the algorithms with proper pre-processing techniques

Comparison with Related Works

To provide a broader context for the study, the performance of the selected algorithms was compared to results from related works in the literature. Table 3 illustrates the comparative analysis, showcasing how this study's findings align with or differ from previous research.

Table 3: The performance of other related works compared to this study

S/N	Author	Method	Dataset	Result %
1	(Alsheref et al., 2022)	Logistic Regression	IBM	82.00
2	Yang & Islam, (2020)	Random Forest	IBM	84.00
3	(Dutta et al., 2020)	KNN	IBM	83.74

Compared to these studies, this research demonstrated superior accuracy across most algorithms, likely due to the comprehensive preprocessing techniques employed. For example, Logistic Regression, which achieved 87% accuracy in this study, outperformed Alsheref et al.'s (2022) result of 82%. Similarly, LDA's performance surpasses related algorithms used in comparable studies, establishing the robustness of this study's methodology.

Performance Evaluation

This research was conducted based on data mining which focused on the impact of pre-processing techniques on models' performance. which include four main pre-processing techniques which are; Handling missing Values, Class imbalance, feature Selection and Feature Scaling. The outcome results were in line with the related literature. The research also provides a review of the related literature which was conducted by other researchers. In the final result of the research, after an experiment on the IBM data was performed, the best-performing classifier was the LDA with 87.38 % accuracy. The lowest score of accuracy goes to KNN with 85.33% accuracy.

CONCLUSION

The experiment was carried out on IBM fictional dataset, which had 35 attributes and 1470 instances. Using a variety of machine learning methods, this study divided the dataset into two categories: attrition and not attrition. After the investigation, the study identifies the impact of various pre-processing techniques on the model performance. Moreover, the impact of preprocessing techniques, including handling missing values, feature selection, and feature scaling, was examined in detail. These techniques were found to significantly influence model performance, with certain methods demonstrating improvements in predictive accuracy and interpretability across different models.

Finally, the practical implications of this research for HR management are profound. By leveraging predictive analytics, organizations can identify key predictors of attrition, tailor retention efforts, and foster a positive work environment conducive to employee satisfaction and retention. These insights offer valuable guidance for HR practitioners seeking to develop data-driven retention strategies that align with organizational goals.

Future research could investigate different machine learning algorithms to determine their effectiveness in improving prediction accuracy, and exploring how they handle class imbalance, feature selection, and data preprocessing to optimize results for specific datasets and domains.

REFERENCES

Ahmad, S., Iliyasu, U., & Jamilu, B. A. (2023). ENHANCED PREDICTIVE MODEL FOR SCHISTOSOMIASIS. *FUDMA JOURNAL OF SCIENCES*, *7*(3), 288–292. https://doi.org/10.33003/fjs-2023-0703-801

Alsheref, F. K., Fattoh, I. E., & Mead, W. (2022). Automated Prediction of Employee Attrition Using Ensemble Model Based on Machine Learning Algorithms. *Computational Intelligence and Neuroscience*, *2022*. https://doi.org/10.1155/2022/7728668

Davidson, R. G., & Brindha, Dr. (2021, June 8). *Inspecting the Impact of Various Factors Influencing Employee Attrition in Hotel Industry*. https://doi.org/10.4108/eai.7-6-2021.2308612

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (n.d.). *LightGBM: A Highly Efficient Gradient Boosting Decision Tree*. https://github.com/Microsoft/LightGBM.

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). https://doi.org/10.3390/math9111226

Otchere, D. A., Ganat, T. O. A., Ojero, J. O., Tackie-Otoo, B. N., & Taki, M. Y. (2022). Application of gradient boosting regression model for the evaluation of feature selection techniques in improving reservoir characterisation predictions. *Journal of Petroleum Science and Engineering*, *208*. https://doi.org/10.1016/j.petrol.2021.109244

Pratt, M., Boudhane, M., & Cakula, S. (2021). Employee attrition estimation using random forest algorithm. *Baltic Journal of Modern Computing*, *9*(1), 49–66. https://doi.org/10.22364/BJMC.2021.9.1.04

Ponnuru, S. R. (2020). Employee Attrition Prediction using Logistic Regression. *International Journal for Research in Applied Science and Engineering Technology*, *8*(5), 2871–2875. https://doi.org/10.22214/ijraset.2020.5481

Punnoose, R., & Xlri -Xavier, C. (2016). Prediction of Employee Turnover in Organizations using Machine Learning Algorithms A case for Extreme Gradient Boosting. In *IJARAI) International Journal of Advanced Research in Artificial Intelligence* (Vol. 5, Issue 9). www.ijarai.thesai.org

Rahamneh, A. A. A. L., Jresat, S. S., Zubaidi, F., & Al-Hawary, S. I. S. (2023). Using the Linear Discriminant Analysis Method to Classify Types of Bowels and Esophageal cancer in Jordan. *Information Sciences Letters*, *12*(3), 1299–1305. https://doi.org/10.18576/isl/120320

Raza, A., Munir, K., Almutairi, M., Younas, F., & Fareed, M. M. S. (2022). Predicting Employee Attrition Using Machine Learning Approaches. *Applied Sciences (Switzerland)*, *12*(13). https://doi.org/10.3390/app12136424

Sethy, A., & Kumar Raut, A. (n.d.). EMPLOYEE ATTRITION RATE PREDICTION USING MACHINE LEARNING APPROACH. *Turkish Journal of Physiotherapy and Rehabilitation*, *32*(3). www.turkjphysiotherrehabil.org

Usha, P. M., & Balaji, N. V. (2021). A comparative study on machine learning algorithms for employee attrition prediction. *IOP Conference Series: Materials Science and Engineering*, *1085*(1), 012029. https://doi.org/10.1088/1757-899x/1085/1/012029

Xu, J., Zhang, Y., & Miao, D. (2020). Three-way confusion matrix for classification: A measure driven view. *Information Sciences*, *507*, 772–794. https://doi.org/10.1016/j.ins.2019.06.064

Wardhani, F. H., & Lhaksmana, K. M. (2022). Predicting Employee Attrition Using Logistic Regression with Feature Selection. *Sinkron*, *7*(4), 2214–2222. https://doi.org/10.33395/sinkron.v7i4.11783