Acta Informatica Pragensia
2025, Volume 14, Issue 1, 112–127
https://doi.org/10.18267/j.aip.255

PRAGUE UNIVERSITY OF ECONOMICS AND BUSINESS

Article Open Access

Predicting Employee Turnover Using Machine Learning Techniques

Adil Benabou [1], **Fatima Touhami** [1], **My Abdelouahed Sabri** [2]

[1] Multidisciplinary Research Laboratory in Economics and Management, Faculty of Economics and Management, Sultan Moulay Slimane University, Beni Mellal, Morocco
[2] Computer Science Department, Faculty of Sciences, University Sidi Mohamed Ben Abdallah, Fez, Morocco

Corresponding author: Adil Benabou (adil.benabou@usms.ma)

Editorial Record

First submission received:
October 23, 2024

Revision received:
November 13, 2024

Accepted for publication:
December 7, 2024

Academic Editor:
Zdenek Smutny
Prague University of Economics and Business, Czech Republic

This article was accepted for publication by the Academic Editor upon evaluation of the reviewers' comments.

How to cite this article:
Benabou, A., Touhami, F., & Sabri, M. A. (2025). Predicting Employee Turnover Using Machine Learning Techniques. *Acta Informatica Pragensia*, 14(1), 112–127. https://doi.org/10.18267/j.aip.255

Abstract

Background: Employee turnover is a persistent issue in human resource management, leading to significant costs for organizations. This study aims to identify the most effective machine learning model for predicting employee attrition, thereby providing organizations with a reliable tool to anticipate turnover and implement proactive retention strategies.

Objective: This study aims to address the challenge of employee attrition by applying machine learning techniques to provide predictive insights that can improve retention strategies.

Methods: Nine machine learning algorithms are applied to a dataset of 1,470 employee records. After data preprocessing and splitting into training and test sets, the models are evaluated on metrics including accuracy, precision, recall, F1 score and AUC. Model performance is optimized through hyperparameter tuning, using grid search with cross-validation.

Results: Logistic regression achieves the highest accuracy and precision, making it the top-performing model overall. Random forest provides a balanced performance with strong AUC, offering a robust alternative.

Conclusion: Human resources managers and directors should consider using logistic regression or random forest for predictive modelling of employee turnover, as these models have shown strong performance. Future research should employ causal analysis for deeper insights. Real-time monitoring and adaptive prediction could also enhance models, offering a dynamic approach to attrition management.

Index Terms

Human resource management; HRM; Machine learning; Employee attrition; Prediction.

1 INTRODUCTION

In today's competitive economy, organizations face persistent challenges in retaining top talent, with high-performing employees being particularly susceptible to turnover. The examination of both actual and anticipated employee turnover has been a long-standing focus within the disciplines of human resource management and psychology (Mozaffari et al., 2023; Hom et al., 2017). Identifying those individuals at the highest risk of leaving is crucial for decision-makers seeking to minimize the disruptions and costs associated with employee attrition. Accurate predictions of which employees might leave can significantly enhance an organization's ability to develop and implement targeted retention strategies, ultimately reducing the financial burden of hiring and training new staff.

To automate the assessment of employee attrition risk, this study utilizes the IBM HR Analytics Employee Attrition & Performance dataset, a resource widely analysed in the data science community. This dataset has been featured in some studies and competitions, with a significant number of publications exploring its insights through advanced visualizations, statistical analyses and predictive models (e.g., Najafi-Zangeneh et al., 2021; Raza et al., 2022).

In this paper, we train, optimize and evaluate a variety of machine learning models to predict employee attrition. By analysing the outcomes of these models, we aim to provide actionable recommendations for organizations looking to improve their retention strategies based on data-driven insights. The findings of this study contribute to the ongoing discourse on the application of machine learning in human resource management, particularly in the context of predicting employee behaviour and optimizing workforce stability.

The rest of the paper is structured as follows: Section 2 provides a review of the literature on employee attrition prediction using machine learning. Section 3 details the machine learning models and methodologies applied in this study. Section 4 outlines the experimental setup and data preprocessing techniques. Section 5 presents and interprets the results and Section 6 concludes with a discussion of the key findings on the practical and theoretical side.

2 LITERATURE REVIEW

Employee attrition remains a significant challenge in human resource management (HRM), and recent advancements in machine learning have provided new tools for predicting employee turnover. Various studies have employed different machine learning (ML) algorithms, such as random forest, support vector machines (SVM), k-nearest neighbours (KNN), extreme gradient boosting, adaptive boosting, decision tree, neural networks and ensemble methods, with varying degrees of success.

Study of employee turnover has a long-standing tradition in HRM and psychology, primarily focusing on identifying key determinants such as demographic, economic and psychological factors (Fallucchi et al., 2020). Traditionally, regression and ANOVA have been used to analyse these factors, but there has been a notable shift towards employing ML techniques for more robust predictive analytics (Sousa-Poza & Henneberger, 2004).

Recent studies have explored various statistical strategies to forecast employee attrition, emphasizing the role of demographic and job-related attributes. While ML models have demonstrated superior efficiency in processing complex datasets, earlier predictive efforts often focused on variables such as absenteeism and tardiness. However, these approaches sometimes lacked the comprehensive parameters required for more accurate forecasting (Nguyen et al., 2020).

Research comparing naïve Bayes and decision tree algorithms has shown that J48 decision trees generally outperform naïve Bayes in predicting employee departures, achieving an accuracy of 82.7% with a 70% training split, while naïve Bayes reached 81% using the same method (Usha & Balaji, 2019). Similarly, a study by Fallucchi et al. (2020) applied different ML models, including Gaussian naïve Bayes, logistic regression, random forest and others, and indicated that Gaussian naïve Bayes performed best, achieving a recall rate (RR) of 0.54 and minimizing false negatives, making it the most effective model in their analysis for detecting attrition with an accuracy rate 82.5%.

Indeed, other studies have demonstrated that random forest have proven particularly accurate, especially when feature selection is applied to refine the input variables, with factors such as job position and overtime significantly influencing attrition (Chakraborty et al., 2021). In addition, other research utilizing a random forest algorithm for employee attrition prediction reported an accuracy of 85.12%, demonstrating its ability to effectively identify key factors such as monthly income, age, daily rate, total working years and monthly rate as critical predictors of employee turnover (Madara Pratt & Cakula, 2021). Moreover, Krishna and Sidharth (2022) showed that the random forest, combined with SMOTE for class balancing, achieved an impressive accuracy of 99.36% in predicting employee attrition. Feature selection and balancing techniques significantly enhanced the model performance, particularly in addressing class imbalance.

Meanwhile, research comparing ML models, including logistic regression and gradient boosting, found that while simpler models provide reliable baseline predictions, ensemble techniques enhance accuracy and robustness by applying multiple models. logistic regression showed an accuracy rate of 81% in predicting turnover (Najafi-Zangeneh et al., 2021). Similarly, in another study, logistic regression achieved an accuracy of 85% in predicting employee turnover, further supporting its efficacy for binary classification tasks (Ponnuru et al., 2020). Moreover, logistic regression again demonstrated superior performance, achieving an accuracy of 88.43%, outperforming decision trees and random forest (Qutub et al., 2021). In contrast, Lazzari et al. (2022) demonstrated that LightGBM outperformed logistic regression and random forest in predicting employee turnover intention using a Europe-wide dataset, achieving an accuracy of 64.1%.

Yahia et al. (2021) conducted a comprehensive study that uniquely combined both machine learning and deep learning approaches to predict employee attrition across three datasets: medium (similar to ours: IBM HR Attrition), large and real-world. Their research revealed that the voting classifier (VC) outperformed all other models, achieving 99% accuracy on the real-world dataset, 96% on the large dataset and 98% on the medium dataset, marking it as the most effective model for HR analytics according to the authors.

In another study, deep learning models outperformed traditional machine learning methods, achieving over 94% accuracy and F1 scores, demonstrating their capability to capture complex patterns within large datasets, which conventional methods might overlook (Arqawi et al., 2022). The study also highlighted the random forest as the best-performing ML model, achieving 92.55% accuracy. Confirmed by the study of Reddy et al. (2023), which found that transformer-based models performed better than traditional ML models, such as random forest and decision tree, in predicting employee attrition, highlighting their effectiveness in handling structured and imbalanced datasets.

Raza et al. (2022) emphasized the role of monthly income, hourly rate, job level and age as significant factors influencing employee attrition. Their study applied multiple machine learning techniques, which achieved a high accuracy of 93% in predicting attrition, outperforming other state-of-the-art models, highlighting its effectiveness in identifying key factors leading to employee turnover.

Nevertheless, there has been a shift towards more advanced modelling approaches in human resources predictive analytics, with a focus on using machine learning and data mining techniques to support human resources teams (Nijjer & Raj, 2020). Most studies employ classification models to identify turnover predictors and often test multiple ML models to find the most effective one (see, e.g., Fallucchi et al., 2020; Jain et al., 2021; Gabrani & Kwatra, 2018). However, despite using similar datasets, there is no consensus on which model is the best for predicting employee turnover. Table 1 below provides a comprehensive overview of key studies that have utilized ML models for predicting employee attrition, highlighting the best-performing models and their corresponding accuracy metrics across different datasets.

Table 1. *Recent related studies.*

Study	Dataset	Models used	Best model	Accuracy
Najafi-Zangeneh et al. (2021)	IBM HR Analytics Employee Attrition	Random forest, logistic regression, naïve Bayes, KNN, decision tree	Logistic regression	81%
Ponnuru et al. (2020)		Logistic regression	Logistic regression	85%
Reddy et al. (2023)		Random forest, transformer	Transformer	Non
Madara Pratt and Cakula (2021)		Random forest, logistic regression, KNN, SVM, decision tree, Gaussian naïve Bayes	Random forest	85.12%
Raza et al. (2022)		Extra trees classifier, SVM, logistic regression, decision tree	Extra trees classifier	93%
Fallucchi et al. (2020)		Gaussian NB, Bernoulli NB, logistic regression, k-nearest neighbours, decision tree, random forest, SVC, linear SVC	Gaussian naïve bayes	82.5%
Qutub et al. (2021)		Decision tree, random forest, logistic regression, Adaboost, gradient boosting	Logistic regression	88.43%
Chakraborty et al. (2021)		Random forest, gradient booster, SVM, KNN, naïve Bayes, logistic regression	Random forest	90.20%

Study	Dataset	Models used	Best model	Accuracy
Arqawi et al. (2022)		Random forest, extra trees, LightGBM, gradient boosting, label propagation, Adaboost, MLP, KNN, deep learning	Deep learning model Random forest	94.52% 92.55%
Usha and Balaji (2019)		J48 (C4.5 DT), naïve Bayes, EM (expectation maximization), k-means	J48 (C4.5 DT)	82.77%
Yahia et al. (2021)	IBM HR Analytics Employee Attrition, HR Analysis (kaggle), Real HR	Decision tree, random forest, logistic regression, SVM, XGBoost, DNN, LSTM, CNN, vote classifier	Vote classifier	98%, 96%, 99%
Krishna and Sidharth (2022)	HR Analysis (kaggle)	Random forest	Random forest	99.36%
Liu et al. (2020)	Real-world data from an aluminium company	Logistic regression, KNN, decision tree, random forest, gradient boosting	Random forest	83%
Lazzari et al. (2022)	Europe-wide employee survey dataset	Random forest, LightGBM, logistic regression, decision trees, XGBoost, KNN, TabNet	LightGBM	64.1%

Table 1 summarizes findings from various studies that show both random forest and logistic regression as highly effective in predicting employee attrition. While random forest often shows robust performance across datasets, logistic regression is frequently highlighted for its accuracy. These results demonstrate the versatility of both models in handling complex datasets.

3 MACHINE LEARNING MODELS

In this study, nine ML algorithms are employed to predict employee attrition, each selected for its strengths in handling classification problems and the specific characteristics of our dataset. These models include decision tree, random forest, LightGBM, logistic regression, XGBoost, AdaBoost, SVM, K-nearest neighbours (KNN) and extra trees classifier (ETC).

Decision tree (DT) is a classification method that builds a model using "if-then" rules, creating an easily interpretable tree structure by splitting data based on significant features (Gerdes, 2013). The Information Gain is used to select the feature that best splits the data:

$$Gain(S, A) = H(S) - \sum_{v \in Values} \frac{|s_v|}{|S|} H(s_v)$$

where:

- S is the global dataset,
- Sv is the subset of S where the attribute A takes the value v,
- |S| is the total number of examples in S,
- $H(Sv)$ is the entropy of the subset Sv.

This calculation makes it possible to identify the attribute offering the best separation of classes in the dataset (Gupta et al., 2017).

$$\mathrm{H(Sv)} = -\sum_{i=1}^{c} P_i \log_2(P_i)$$

The entropy $H(Sv)$ represents the uncertainty in the data, with pi being the proportion of examples of the class i in Sv. For example, if for the attribute *Attrition*, 84% of employees in a subset do not leave (class "*No*") and 16% leave (class "*Yes*"), the entropy of this subset is calculated to assess the purity of this division.

In the decision tree model, *MonthlyIncome, OverTime, TotalWorkingYears and YearsAtCompany* were key factors that influenced attrition. Employees with lower monthly income and frequent overtime showed a greater tendency to leave, particularly in roles with lower job satisfaction. Using R functions such as *rpart* and *varImp*, we applied Information Gain and Gini Impurity to rank these features, allowing the model to identify factors most strongly linked to employee attrition.

Random forest (RF) is an ensemble method that builds multiple decision trees using random subsets of data, combining their outputs to improve accuracy and reduce overfitting (Xia, 2020). The model calculates the Gini Impurity for each split:

$$Gini(j) = 1 - \sum_{i=1}^{c} P_i^2$$

where:

- j is the node being evaluated,
- pi represents the proportion of elements of the class i in the node j,
- c is the total number of classes.

The Gini impurity quantifies the degree of class heterogeneity within a node. For instance, if a node comprises 84% of employees who leave and 16% who stay, the Gini impurity for this node would be relatively high, reflecting an imperfect class separation.

In the random forest model, *MonthlyIncome, Age, TotalWorkingYears* and *DailyRate* emerged as essential predictors. Employees with lower monthly income or higher overtime showed a greater likelihood of leaving. The model identified these features through functions such as *randomForest::randomForest* for training and *importance* to rank feature significance.

Logistic regression (LR) is a statistical model for binary classification, predicting the probability of an outcome based on input features using the logistic function (Saidi et al., 2021).

$$P\left(Attrition = \frac{1}{x}\right) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n)}}$$

where:

- P(Attrition=1|x) represents the probability that an employee will leave the company,
- β_0 is the intercept term, serving as a baseline value,
- βi are the coefficients for each predictor variable xi, indicating the impact of individual features on attrition likelihood.

In the logistic regression model, *MonthlyIncome, OverTime, Department* and *MaritalStatus* were identified as critical features. For instance, the coefficient for *MonthlyIncome* (β_{income}) was negative, indicating that higher income levels reduce the probability of attrition. Using R functions such as *glm* for model training and summary to examine feature significance, the logistic regression model provided insights into how factors such as income and marital status affect attrition risk, enabling HR teams to better target retention efforts.

LightGBM is a fast, scalable gradient boosting framework that excels at handling large, high-dimensional datasets by growing tree's leaf-wise for improved accuracy.

$$L(y, \hat{y}) = \sum_{i=1}^{n} [\hat{y}_i \log(\hat{y}_i) + (1 - \hat{y}_i) \log(1 - \hat{y}_i)]$$

where:

- yi represents the actual class label (0 for staying, 1 for leaving),
- $\hat{y}_i$ is the predicted probability of attrition for the i-th employee.

In the LightGBM model, *MonthlyIncome, Age, OverTime* and *DailyRate* emerged as critical features, for example younger employees with fewer years at the company being more likely to leave. By utilizing functions such as *lgb.train* for model training and *lgb.importance* to identify key predictors, LightGBM effectively captured non-linear relationships between factors such as age and company tenure.

XGBoost is another gradient boosting algorithm that builds trees sequentially, excelling in accuracy, speed and memory efficiency for predicting employee turnover (Ahmetoglu & Das, 2022).

$$Obj(\theta) = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{k=1}^{T} \Omega(f_k)$$

where:

- $L(y_i, \hat{y}_i)$ represents the loss function measuring the difference between the actual label yi and the predicted value $\hat{y}i$ for the i-th observation,
- $\Omega(f_k)$ is the regularization term for the k-th tree f_k, controlling model complexity to prevent overfitting,
- T denotes the total number of trees.

In this study, *MonthlyIncome, Age, OverTime* and *DailyRate* were highlighted as key features. Using *xgb.train* for model training and *xgb.importance* to identify feature importance, XGBoost effectively captured these nuanced patterns in employee data.

AdaBoost (AB) builds a strong classifier by combining several weak learners, such as decision trees, and focuses on the errors made by previous learners (Giorgio et al., 2023). The loss function minimized by AdaBoost is:

$$L(y, \hat{y}) = e^{-yf(x)}$$

where:

- y is the actual class label (with values +1 for positive and -1 for negative cases),
- $f(x)$ is the combined prediction from the ensemble of weak learners for the given input x.

In the Adaboost model, *MonthlyIncome, MonthlyRate, DailyRate* and *HourlyRate* were significant predictors, using *adabag::boosting* for model training and *importanceplot* to evaluate feature impact.

Support vector machine (SVM) separates classes by finding the hyperplane with the maximum margin, effectively minimizing classification errors, though they can be computationally demanding (Brants, 2006).

$$\min_{\omega} \frac{1}{2} \|\omega\|^2 + c \sum_{i=1}^{n} \xi_i$$

where:

- ω represents the weights of the hyperplane, defining its orientation,
- $||\omega||^2$ controls the margin width, promoting a larger margin for better separation,
- C is the regularization parameter that balances margin maximization and classification error,
- ξi are slack variables that allow some misclassification to handle non-linearly separable data.

The key predictors in this study included *TotalWorkingYears, OverTime, YearsAtCompany* and *MonthlyIncome*. The *varImp* function was used to assess feature importance, calculating the overall importance by averaging the contributions across classes.

K-nearest neighbours (KNN) classifies data based on the majority class among the k-nearest neighbours, effective for small datasets but less efficient for larger ones (Xiong & Yao, 2021). The distance between points is calculated using Euclidean Distance:

$$\boldsymbol{d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}}$$

where:

- x and y are the feature vectors of two data points,
- n is the number of features,

- $(x_i - y_i)^2$ calculates the squared difference between corresponding features of x and y.

In this model, *DistanceFromHome* and *MonthlyIncome* were identified as significant predictors of attrition. Using the *caret::train* function to train the model and a custom permutation-based function, *calculate_permutation_importance*, to assess feature importance, the impact of each feature was measured by calculating the reduction in accuracy when its values were randomly shuffled.

Extra trees classifier (ETC) constructs random decision trees, improving accuracy and speed while avoiding overfitting by using random splits and the Gini index for feature selection (Baby et al., 2021).

$$\boldsymbol{Split(feature) \sim u(min, max)}$$

where:

- u(min, max) denotes a uniform distribution between the minimum and maximum values of the feature,
- the split points are chosen randomly within this range, reducing overfitting and enhancing model robustness.

In this model, *MonthlyIncome, Age, OverTime* and *DailyRate* were identified as key predictors of attrition. Using *ranger::ranger* for model training and extracting feature importance with *ranger_model$variable.importance,* the extra trees classifier was able to highlight patterns, such as higher attrition risk for employees at lower job levels.

4 METHODS

4.1 Dataset used

This study utilizes a publicly available dataset from IBM HR Analytics Employee Attrition & Performance, see Data availability statement. It contains 1,470 records, each representing an employee with 35 features encompassing both numerical and categorical data (see Table 2).

Table 2. *Dataset features.*

Feature name	**Type**	**Feature name**	**Type**
Age	Numeric	MonthlyRate	Numeric
DistanceFromHome	Numeric	NumCompaniesWorked	Numeric
Attrition	Categorical	Over18	Categorical
Education	Categorical	OverTime	Categorical
EducationField	Categorical	PercentSalaryHike	Numeric
EmployeeCount	Numeric	PerformanceRating	Numeric
EmployeeNumber	Numeric	RelationshipSatisfaction	Categorical
EnvironmentSatisfaction	Categorical	DailyRate	Numeric
Gender	Categorical	StandardHours	Numeric
HourlyRate	Numeric	StockOptionLevel	Categorical
JobInvolvement	Categorical	TotalWorkingYears	Numeric
JobLevel	Categorical	TrainingTimesLastYear	Numeric
JobRole	Categorical	WorkLifeBalance	Categorical
Department	Categorical	YearsAtCompany	Numeric
BusinessTravel	Categorical	YearsInCurrentRole	Numeric
JobSatisfaction	Categorical	YearsSinceLastPromotion	Numeric
MaritalStatus	Categorical	YearsWithCurrManager	Numeric
MonthlyIncome	Numeric		

The dataset was split into two categories (see Figure 1): 1,233 records for employees who stayed with the company ("No" attrition) and 237 records for those who left ("Yes" attrition).

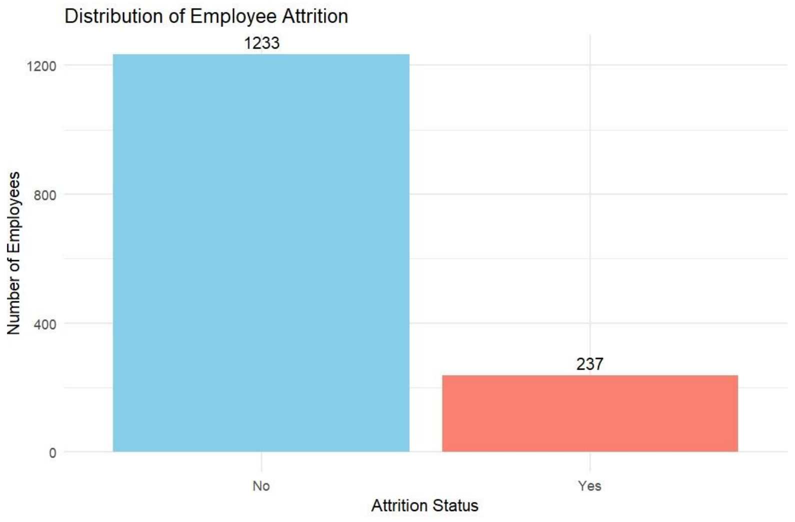

Figure 1. *Distribution of employee attrition.*

4.2 Data preparation and model development workflow

In developing the predictive model, we follow a conventional approach, starting with preparing the input data. The initial data, comprising employee information, are fed into the machine learning algorithm. These data are then used to train the algorithm, allowing it to learn patterns and relationships within the dataset. Once the model is trained, it is evaluated to ensure its effectiveness. This involves assessing its accuracy in predicting outcomes and if necessary, adjusting the model for better performance. Finally, the refined model is applied to new data to generate predictions about employee attrition, providing actionable insights for HR management.

4.2.1 Data preparation

Data preprocessing is a critical step to ensure that the data fed into the ML models are clean, consistent and in a format that maximizes the model learning potential. This process includes:

Handling missing values: Missing values were identified and managed using imputation techniques or by excluding records with substantial gaps. To remove rows with missing data, the *na.omit()* function was used, which excludes any row containing an NA value. In cases where imputation was suitable, the *impute()* function was applied to estimate missing values based on other available data, allowing a more complete dataset without losing potentially valuable records.

Encoding categorical variables: Categorical variables were converted into numerical equivalents using one-hot encoding, creating binary columns for each category. The *model.matrix()* function in R facilitated this transformation, allowing the models to effectively interpret categorical information in a numeric format.

Feature scaling: Feature scaling was applied to ensure that all variables had a consistent impact on the analysis, preventing any single feature from dominating due to its range. Using the *scale()* function in R, each variable was adjusted to have a mean of zero and a standard deviation of one, bringing them to a common scale. This step was especially helpful for models such as logistic regression, which can be sensitive to differences in feature magnitudes, allowing balanced contributions from all features.

Feature selection: Irrelevant or redundant features were removed to streamline the model and improve accuracy. This step was informed by analysing feature importance metrics, especially those provided by tree-based methods such as random forest. Using functions such as *randomForest()* in R, feature importance scores were obtained, highlighting the variables that had the most predictive value. Features with lower importance were subsequently removed to focus on the most relevant predictors, enhancing both the model efficiency and interpretability.

4.2.2 System implementation

The analysis was conducted using R version 4.0.2 in RStudio, utilizing libraries such as *caret*, *randomForest*, *xgboost*, *e1071*, *ggplot2* and others for data manipulation, machine learning model development and result visualization. The computations were carried out on a system equipped with an Intel Core i7 processor, 32GB RAM and a 512GB SSD.

The dataset was split into training (80%) and testing (20%) subsets with the *createDataPartition()* function from the *caret* package. Cross-validation and hyperparameter tuning were performed through grid search using the *train()* function. Models were evaluated using key metrics: accuracy, precision, recall and area under the curve (AUC), ensuring a thorough assessment of their ability to handle class imbalance in employee attrition prediction.

4.3 Performance metrics

Given the imbalanced nature of the dataset, it was essential to use evaluation metrics that provide a balanced and accurate assessment of model performance. Five distinct metrics were employed for a comprehensive evaluation of the models.

Accuracy: Reflects the overall correctness of the model by measuring the proportion of true positive (TP) and true negative (TN) predictions out of the total number of instances.

$$Accuracy = \frac{TP + TN}{N}$$

Precision: Indicates how accurately the model identifies positive cases, calculated as:

$$Precision = \frac{TP}{TP + FP}$$

Recall (sensitivity): This metric measures the model's ability to capture all actual positive cases.

$$Recall = \frac{TP}{TP + FN}$$

F1 score: Balances precision and recall by calculating their harmonic mean, making it especially useful for evaluating performance on imbalanced datasets.

$$F1\ score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Receiver operating characteristic (ROC) curve and area under the curve (AUC): The ROC curve illustrates the trade-off between sensitivity (true positive rate) and specificity (false positive rate) across different decision thresholds. The AUC provides a single value to summarize the overall performance of the model, with higher values indicating better performance.

$$AUC = \int_0^1 TPR(FPR)d(FPR)$$

Here, the true positive rate (TPR) measures the proportion of correctly identified positive cases, with its function being (Vázquez-Diosdado et al., 2024):

$$TPR = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP)\ +\ False\ Negatives\ (FN)}$$

Furthermore, the false positive rate (FPR) quantifies the proportion of negative cases incorrectly classified as positive, with its function being (Vázquez-Diosdado et al., 2024):

$$FPR = \frac{False\ Positives\ (FP)}{False\ Positives\ (FP)\ +\ True\ Negatives\ (TN)}$$

5 RESULTS

In this study, we evaluated the performance of nine ML models in predicting employee attrition. Each model was tested and compared based on five key metrics. Table 3 provides a comprehensive comparison of these models, highlighting the strengths and weaknesses of each approach.

Table 1. *Performance comparison of machine learning models in employee attrition prediction.*

Model	Accuracy	Precision	Recall	F1 score	AUC
Logistic regression	0.877	0.877	0.684	0.769	0.839
Random forest	0.857	0.850	0.692	0.763	0.845
Decision tree	0.846	0.846	0.763	0.802	0.694
LightGBM	0.871	0.160	0.697	0.260	0.500
XGBoost	0.863	0.863	0.694	0.769	0.830
AdaBoost	0.857	0.857	0.692	0.766	0.838
SVM	0.870	0.870	0.963	0.914	0.135
KNN	0.809	0.819	0.786	0.802	0.605
Extra trees	0.846	0.840	0.692	0.759	0.155

Table 3 shows that logistic regression achieves a high accuracy (0.877) and precision (0.877), making it a strong model overall. However, its recall (0.684) suggests that it may miss some true positives, which could be critical in attrition prediction. The SVM model shows exceptional recall (0.963) and high precision (0.870), indicating that it is very effective in identifying true positives, although its AUC is notably lower (0.135), which limits its overall reliability. LightGBM, despite having good accuracy (0.871), struggles with precision (0.160), suggesting a tendency to generate many false positives. The random forest provides a balanced performance with decent accuracy (0.857), precision (0.850), recall (0.692) and a competitive AUC (0.845), making it a well-rounded choice. AdaBoost also performs well with strong precision (0.857) and a high AUC (0.838), indicating its robustness. Meanwhile, the decision tree and KNN models each achieve F1 scores of 0.802, indicating a strong balance between precision and recall. This balance makes them reliable choices for identifying attrition cases, particularly in situations where maintaining accuracy across metrics is crucial. In contrast, extra trees show lower F1 and AUC values, which could reduce its effectiveness for consistently predicting attrition.

Table 2. *Key metric comparison across machine learning models.*

Metric	Best model	Value
Accuracy	Logistic regression	0.877
Precision	Logistic regression	0.877
Recall	SVM	0.963
F1 Score	SVM	0.914
AUC	Random forest	0.845

Table 4 highlights that logistic regression performs well with both high accuracy (0.877) and precision (0.877), making it effective at identifying true positives and minimizing false positives. Support vector machines stand out with the highest recall (0.963) and the top F1 score (0.914), making it particularly strong in capturing true attrition cases. This balance suggests that SVM is highly effective in situations where identifying all attrition cases is critical. The random forest leads in AUC (0.845), demonstrating a strong discriminative power across different decision thresholds, which enhances its reliability in classifying both attrition and non-attrition cases.

5.1 Performance comparison

The following figures illustrate the performance metrics of machine learning models in predicting employee attrition.

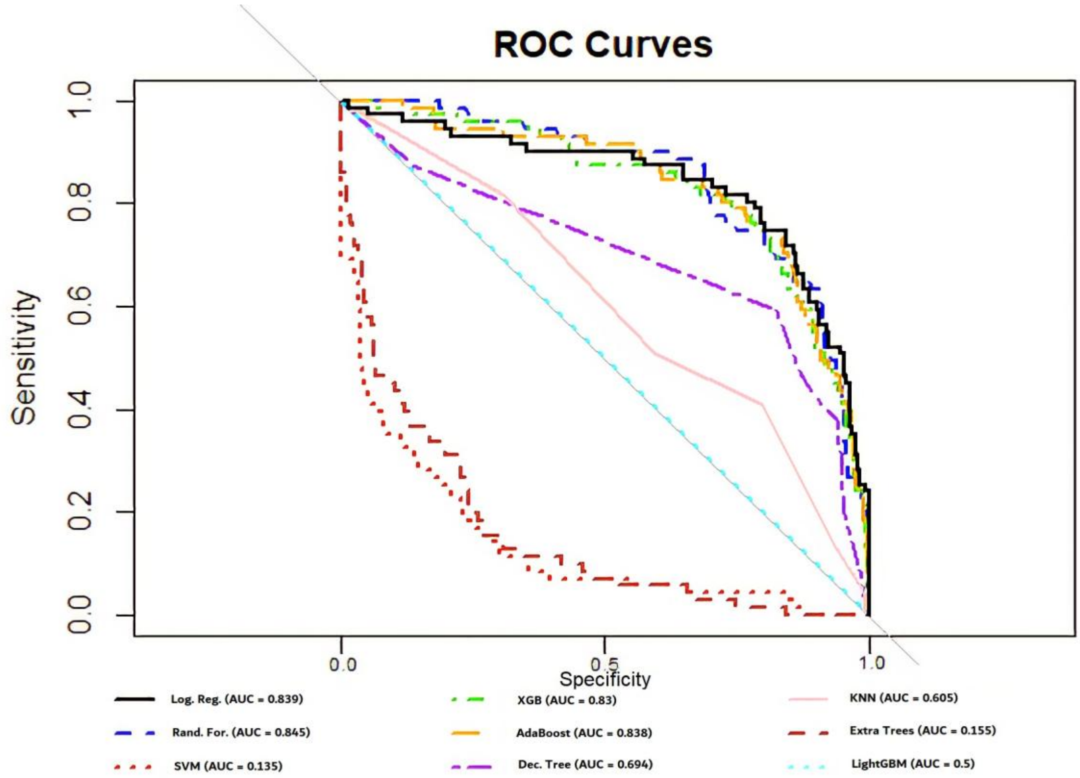

***Figure 2**. ROC curves for machine learning models.*

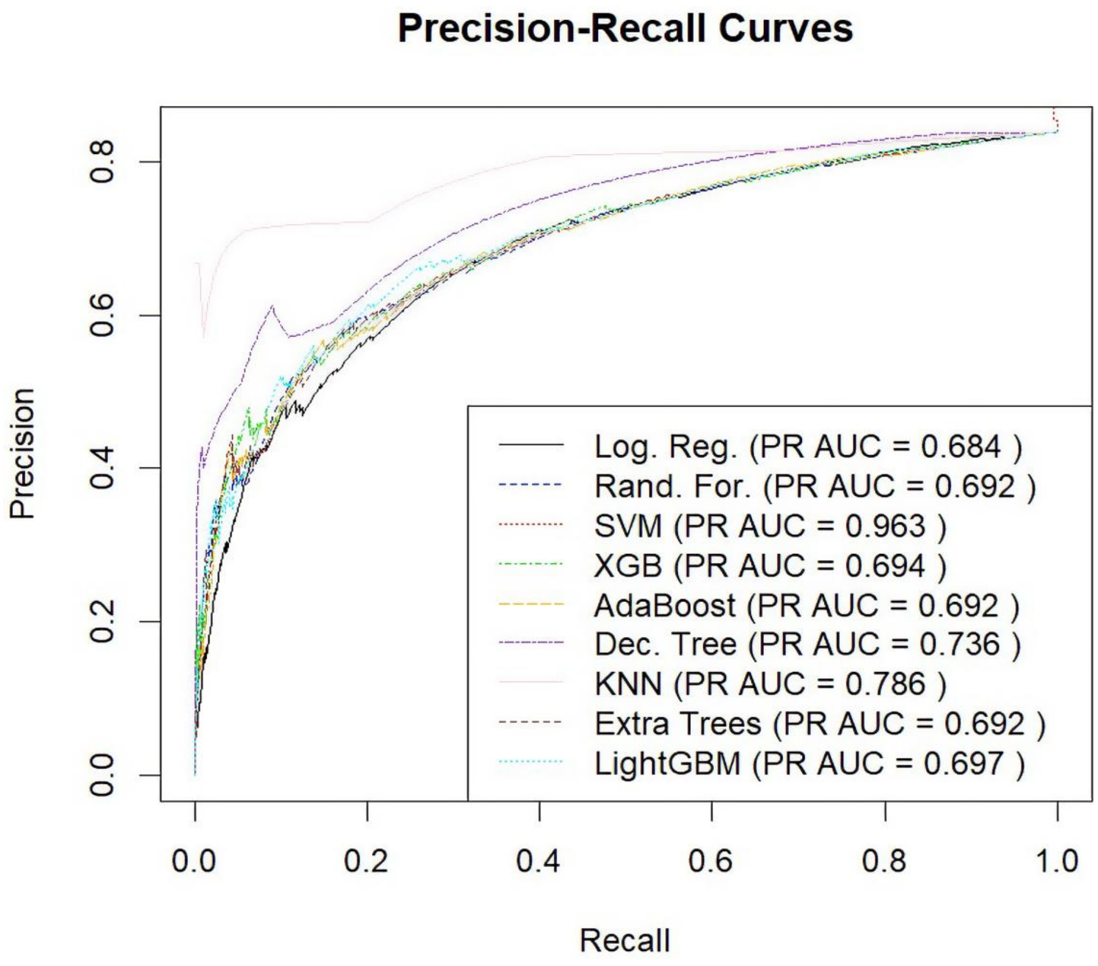

***Figure 3.** Employee attrition prediction.*

Figure 2 shows the ROC curves for ML models. The random forest leads with the highest AUC (0.845), indicating a strong discriminative power. Logistic regression follows with an AUC of 0.839, while XGBoost and AdaBoost also perform well, with AUCs of 0.830 and 0.838, respectively. The decision tree and KNN models show moderate performance with AUCs of 0.694 and 0.605. In contrast, SVM (0.135), extra trees (0.155) and LightGBM (0.500) demonstrate weaker discriminative abilities. These curves highlight the trade-offs between sensitivity and specificity across models.

The precision-recall curves in Figure 3 show the trade-off between precision and recall across nine ML models. The SVM model achieved the highest AUC-Recall rate (RR) of 0.963, indicating superior performance in balancing precision and recall. K-nearest neighbours followed with an AUC-PR of 0.786, showing relatively strong performance. XGBoost and LightGBM demonstrated moderate effectiveness with AUC-PR of 0.694 and 0.697, respectively. The logistic regression, random forest, AdaBoost and extra trees models all had similar AUC-PR values around 0.684 to 0.692, suggesting comparable performance. The decision tree model displayed an AUC-PR of 0.736, reflecting a moderate balance between precision and recall. These results emphasize the importance of selecting a model that aligns with the specific needs of the project, particularly in managing the trade-offs between false positives and false negatives.

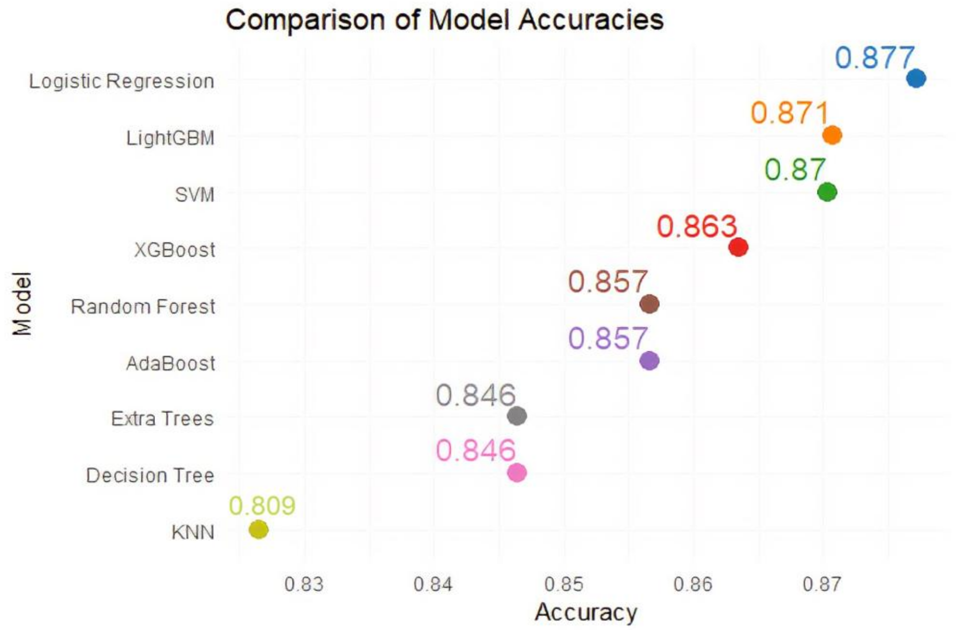

Figure 4. *Comparison of model accuracy.*

The plot in Figure 4 shows that the logistic regression model achieved the highest accuracy at 0.877, demonstrating its strong performance in correctly classifying the data. LightGBM and support vector machines followed closely with accuracies of 0.871 and 0.870, respectively. XGBoost also performed well with an accuracy of 0.863, while the random forest and AdaBoost models both recorded an accuracy of 0.857. The decision tree and extra trees classifier models showed lower accuracies at 0.846 each. Notably, the KNN model exhibited the lowest accuracy at 0.809, indicating that it may be less suitable for this particular task.

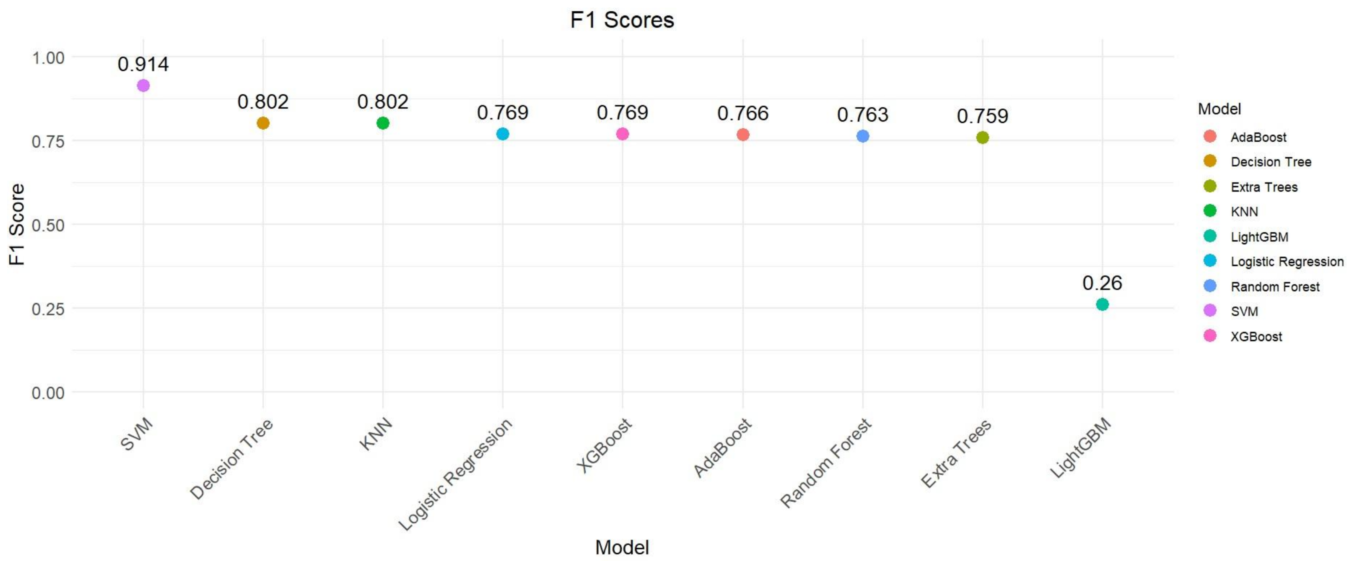

Figure 5. *F1 scores across machine learning models.*

In Figure 5, support vector machines achieved the highest F1 score at 0.914, indicating its strong balance between precision and recall, making it highly effective for predicting attrition cases. The decision tree and K-nearest neighbours models also performed well, each with an F1 score of 0.802, showcasing their ability to reliably identify cases of attrition. logistic regression, XGBoost and AdaBoost demonstrated moderate performance, with F1 scores of 0.769, 0.769 and 0.766, respectively, suggesting that these models offer a balanced approach but with slightly lower effectiveness compared to SVM. Meanwhile, extra trees classifier and random forest had F1 scores of 0.759 and 0.763, respectively, reflecting a reasonable balance but with limited precision-recall synergy. LightGBM, with an F1 score of 0.26, showed comparatively lower effectiveness, indicating a higher rate of false positives and reduced reliability in this context.

5.2 Performance analysis

The evaluation of the machine learning models on the IBM dataset (1,470 records, 16.1% attrition) highlighted logistic regression as the best-performing model, with the highest accuracy (87.7%) and precision, correctly identifying 208 out of 237 employees who had left. However, it missed some attrition cases due to lower recall (0.684). The random forest provided a balanced performance with an accuracy of 85.7% and an AUC of 0.845, correctly identifying 164 out of 237 employees who had left, making it a robust alternative for HR decision-making, particularly for balancing true positives and false positives. Support vector machines showed exceptional recall (0.963) but struggled with a low AUC (0.135), limiting its reliability. LightGBM demonstrated strong accuracy (87.1%) but had poor precision, indicating a high rate of false positives. While support vector machines are useful in scenarios where capturing all potential attrition cases is critical, logistic regression remains the top choice due to its superior accuracy and precision.

Given the class imbalance in our dataset by 1,233 instances of “No” attrition and 237 instances of “Yes” attrition, a baseline was established using a majority class classifier. This classifier, predicting all the cases as “No” attrition, achieved an accuracy of 83%, with a precision of 0.83, a recall of 1.0 and an F1 score of 0.91 for the “No” class. These metrics provide a basic benchmark, setting a minimum standard that machine learning models must surpass to demonstrate meaningful predictive capability.

In comparison, the top-performing models such logistic regression, random forest and support vector machines, all outperformed this baseline. Logistic regression achieved an accuracy of 87.7% for predicting “Yes” attrition cases, demonstrating an improvement over the baseline 83% accuracy and addressing cases missed by the majority class classifier. Random forest showed similar gains, with an accuracy of 85.7% and an F1 score of 0.763, effectively identifying instances of attrition. Support vector machines achieved an accuracy of 87% and the highest F1 score of 0.914, further illustrating the advantages of using ML models to detect employees at-risk of leaving. These results confirm that these models provide substantial improvements in predictive accuracy and recall over the simplistic majority class approach, supporting more nuanced HR decision-making.

6 DISCUSSION

6.1 Theoretical contribution

This study provides a comprehensive comparison of machine learning models for predicting employee attrition, contributing to the academic understanding of model effectiveness in this area. Logistic regression achieved the highest accuracy at 87.7%, demonstrating its strong suitability for binary classification in human resources data. The random forest followed closely with an accuracy of 85.7% and an AUC of 0.845, reflecting its consistent performance across multiple metrics. These findings align with prior studies; for instance, Madara Pratt and Cakula (2021) reported an accuracy of 85.12% on the same dataset as ours, while Chakraborty et al. (2021) achieved 90.2% accuracy by applying feature selection. Additionally, Liu et al. (2020) demonstrated the ability of random forest to predict turnover with 83% accuracy on industrial data, and Krishna & Sidharth (2022) reached 99.36% accuracy after applying SMOTE to handle class imbalance. These studies underscore the adaptability of random forest across datasets with mixed categorical and numerical variables.

Our analysis also highlighted LightGBM, which achieved an accuracy of 87.1% and an AUC-PR of 69.7% but struggled with an F1 score of 0.26, likely due to class imbalance. This contrasts with the findings of Lazzari et al. (2022), where LightGBM reached a significantly higher AUC-PR of 64.1% on a larger dataset, indicating that its performance could improve with more data and tailored feature engineering. These results show that LightGBM is relatively close to our Random Forest results in terms of accuracy but struggles with precision, as indicated by its lower F1 score.

In our study, logistic regression achieved the highest accuracy at 87.7%, outperforming all the other models tested. This result is consistent with findings from similar research using the IBM HR Analytics dataset. For instance, Ponnuru et al. (2020) reported an accuracy of 85% with logistic regression, while Najafi-Zangeneh et al. (2021) found 81% accuracy on the same dataset. These recurring results highlight logistic regression as a dependable option for predicting employee attrition. However, this model did show some limitations in recall (0.684) and F1 score (0.769), indicating challenges in identifying all employees at risk of leaving. Qutub et al. (2021) observed a similar trend, noting that even with a high accuracy of 88.43% on a comparable dataset, logistic regression struggled with precision and recall, likely due to class imbalance. This pattern suggests that, while logistic regression serves as a reliable baseline, enhancements through feature engineering or class balancing could improve its ability to capture attrition cases more effectively.

In addition, several studies have highlighted the effectiveness of deep learning for predicting employee attrition. Arqawi et al. (2022) found that deep learning models outperformed traditional algorithms, achieving 94.52% accuracy. Yahia et al. (2021) similarly demonstrated that CNN and LSTM models reached accuracies of up to 99%, surpassing random forest and logistic regression. Reddy et al. (2023) also showed that transformer-based models excelled in handling structured and imbalanced datasets, further emphasizing the growing relevance of deep learning as datasets become more complex.

However, this study provides a comprehensive analysis of machine learning models for predicting employee attrition, uniquely characterized by its clear explanations of mathematical functions and detailed examination of feature importance for each model. By identifying and interpreting the most influential variables, this research supports HR professionals in understanding the factors that contribute to turnover risks. Additionally, it compares the findings with prominent studies in this domain, placing its results within the broader landscape of attrition research. The study also highlights practical R functions commonly used in machine learning, offering readers concrete tools for replicating and adapting these analyses to their own datasets. This approach enhances the study's methodological transparency and makes it a valuable resource for organizations and researchers aiming to refine their predictive strategies in human resources.

6.2 Practical implications

For HR managers, these insights offer valuable guidance on model selection; for example, logistic regression, with its accuracy of 87.7%, emerges as a reliable choice for identifying employees likely to stay, making it suitable for broader HR use cases. On the other hand, the random forest shines in its balanced performance across accuracy, precision and recall, making it adaptable to scenarios where consistent metrics across different measures are essential.

In contexts where a high recall rate is necessary, as in cases focused on identifying most potential attrition cases, support vector machines offer a recall of 0.963, though its AUC of 0.135 suggests some limitations in its general reliability. LightGBM also showed strong accuracy at 87.1%, but its lower precision indicates that it may be best utilized in combination with other models to reduce false positives in HR scenarios.

The study underscores how class imbalance can affect model effectiveness, suggesting that HR managers might consider techniques such as SMOTE to enhance model performance, particularly in datasets where attrition is underrepresented. By evaluating nine distinct models, including traditional, ensemble and newer methods, this research provides HR practitioners with an informed basis for choosing models tailored to their organizational needs and patterns in employee engagement and turnover.

7 CONCLUSION

This study demonstrated the effectiveness of various machine learning models in predicting employee turnover, with logistic regression achieving the highest accuracy. Future research should explore the integration of contextual features, such as external labour market conditions, to enhance model sensitivity to external factors influencing turnover. Additionally, expanding the analysis across multiple organizations and sectors could improve generalizability and adaptability to different workplace environments. Employing causal analysis algorithms, such as causal graphs, could also provide insights into underlying factors driving attrition. Finally, incorporating real-time monitoring and adaptive prediction systems would allow models to continuously refine predictions as new data become available, offering a dynamic approach to attrition management. This paper contributes to a deeper understanding of the role of machine learning in predicting employee turnover, highlighting both the strengths and challenges of various models and proposing avenues for further exploration.

ADDITIONAL INFORMATION AND DECLARATIONS

Conflict of Interests: The authors declare no conflict of interest.

Author Contributions: A.B.: Conceptualization, Methodology, Data analysis, Writing – Original draft preparation, Software, Writing – Reviewing and Editing. F.T.: Supervision, Validation. M.A.S.: Conceptualization, Validation.

Statement on the Use of Artificial Intelligence Tools: The authors declare that they didn't use artificial intelligence tools for text or other media generation in this article.

Data Availability: The authors used a dataset available from https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset (IBM HR Analytics Employee Attrition & Performance dataset).

REFERENCES

Ahmetoglu, H., & Das, R. (2022). A comprehensive review on detection of cyber-attacks: Data sets, methods, challenges, and future research directions. *Internet of Things*, 20, 100615. https://doi.org/10.1016/j.iot.2022.100615

Arqawi, S., Rumman, M., Zitawi, E., Abunasser, B., & Abu-Naser, S. (2022). Predicting Employee Attrition and Performance Using Deep Learning. *Journal of Theoretical and Applied Information Technology*, 100, 6526–6536.

Baby, D., Devaraj, S. J., Hemanth, J., & M, A. R. (2021). Leukocyte classification based on feature selection using extra trees classifier: atransfer learning approach. *Turkish Journal of Electrical Engineering and Computer Sciences*, 29(SI-1), 2742–2757. https://doi.org/10.3906/elk-2104-183

Brants, T. (2006). Part-of-Speech Tagging. In *Encyclopedia of Language & Linguistics*, (pp. 221–230). Elsevier. https://doi.org/10.1016/B0-08-044854-2/00952-4

Chakraborty, R., Mridha, K., Shaw, R. N., & Ghosh, A. (2021). Study and Prediction Analysis of the Employee Turnover using Machine Learning Approaches. In *2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON).* IEEE. https://doi.org/10.1109/GUCON50781.2021.9573759

Fallucchi, F., Coladangelo, M., Giuliano, R., & Luca, E. W. (2020). Predicting Employee Attrition Using Machine Learning Techniques. *Computers*, 9(4), 86. https://doi.org/10.3390/computers9040086

Gabrani, G., & Kwatra, A. (2018). Machine Learning Based Predictive Model for Risk Assessment of Employee Attrition. In *Computational Science and Its Applications – ICCSA 2018* (pp 189–201). Springer. https://doi.org/10.1007/978-3-319-95171-3_16

Gerdes, M. (2013). Decision trees and genetic algorithms for condition monitoring forecasting of aircraft air conditioning. *Expert Systems with Applications*, 40(12), 5021–5026. https://doi.org/10.1016/j.eswa.2013.03.025

Giorgio, A. d., Cola, G., & Wang, L. (2023). Systematic review of class imbalance problems in manufacturing. *Journal of Manufacturing Systems*, 71, 620–644. https://doi.org/10.1016/j.jmsy.2023.10.014

Gupta, B., Rawat, A., Jain, A., Arora, A., & Dhami, N. (2017). Analysis of various decision tree algorithms for classification in data mining. *International Journal of Computer Applications*, 163(8), 15–19. https://doi.org/10.5120/ijca2017913660

Hom, P. W., Lee, T. W., Shaw, J. D., & Hausknecht, J. P. (2017). One hundred years of employee turnover theory and research. *Journal of Applied Psychology*, 102(3), 530–545. https://doi.org/10.1037/apl0000103

Jain, N., Tomar, A., & Jana, P. K. (2021). A novel scheme for employee churn problem using multi-attribute decision making approach and machine learning. *Journal of Intelligent Information Systems*, 56, 279–302. https://doi.org/10.1007/s10844-020-00614-9

Krishna, S., & Sidharth, S. (2022). HR Analytics: Employee Attrition Analysis using Random Forest. *International Journal of Performability Engineering*, 18(4), 275–281. https://doi.org/10.23940/ijpe.22.04.p5.275281

Lazzari, M., Alvarez, J. M., & Ruggieri, S. (2022). Predicting and explaining employee turnover intention. *International Journal of Data Science and Analytics*, 14(3), 279–292. https://doi.org/10.1007/s41060-022-00329-w

Liu, L., Akkineni, S., Story, P., & Davis, C. (2020). Using HR Analytics to Support Managerial Decisions: A Case Study. In *Proceedings of the 2020 ACM Southeast Conference*, (pp. 168–175). ACM. https://doi.org/10.1145/3374135.3385281

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

Mozaffari, F., Rahimi, M., Yazdani, H., & Sohrabi, B. (2023). Employee attrition prediction in a pharmaceutical company using both machine learning approach and qualitative data. *Benchmarking: An International Journal*, 30(10), 4140–4173. https://doi.org/10.1108/BIJ-11-2021-0664

Najafi-Zangeneh, S., Shams-Gharneh, N., Arjomandi-Nezhad, A., & Zolfani, S. H. (2021). An Improved Machine Learning-Based Employees Attrition. *Mathematics*, 9(11), 1226. https://doi.org/10.3390/math9111226

Nguyen, T. D., Pham, L. D., Crouch, M., & Springer, M. G. (2020). The correlates of teacher turnover: An updated and expanded Meta-analysis of the literature. *Educational Research Review*, 31, 100355. https://doi.org/10.1016/j.edurev.2020.100355

Nijjer, S., & Raj, S. (2020). *Predictive Analytics in Human Resource Management*. Routledge India. https://doi.org/10.4324/9781003026822

Ponnuru, S. R., Merugumala, G. K., Padigala, S., Vanga, R., & Kantapalli, B. (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

Qutub, A., Al-Mehmadi, A., Al-Hssan, M., Aljohani, R., & Alghamdi, H. S. (2021). Prediction of employee attrition using machine learning and ensemble methods. *International Journal of Machine Learning and Computing*, 11(2), 110–114. https://doi.org/10.18178/ijmlc.2021.11.2.1022

Raza, A., Munir, K., Almutairi, M., & Younas, F. (2022). Predicting Employee Attrition Using Machine Learning. *Applied Sciences*, 12(13), 6424. https://doi.org/10.3390/app12136424

Reddy, V., Taurani, S. K., Chandrashekhar, A., & Shravya, D. (2023). Big data-based framework for prediction of employee attrition by using deep data people analytics. *Journal of Informatics Education and Research*, 3(2). https://doi.org/10.52783/jier.v3i2.432

Saidi, A., Othman, S. B., Dhouibi, M., & Saoud, S. B. (2021). FPGA-based implementation of classification techniques: A survey. *Integration*, 81, 280–299. https://doi.org/10.1016/j.vlsi.2021.08.004

Sousa-Poza, A., & Henneberger, F. (2004). Analyzing Job Mobility with Job Turnover Intentions: An International Comparative Study. *Journal of Economic Issues*, 38(1), 113–137. https://doi.org/10.1080/00213624.2004.11506667

Usha, P.M., & Balaji, N. V. (2019). Analysing Employee attrition using machine learning. *Karpagam JCS*, 13, 277–282.

Vázquez-Diosdado, J. A., Doidge, C., Bushby, E. V., Occhiuto, F., & Kaler, J. (2024). Quantification of play behaviour in calves using automated ultra-wideband location data and its association with age, weaning and health status. *Scientific Reports*, 14, Article number 8872. https://doi.org/10.1038/s41598-024-59142-z

Xia, Y. (2020). Chapter Eleven – Correlation and association analyses in microbiome study integrating multiomics in health and disease. In *Progress in Molecular Biology and Translational Science*, (pp. 309–491). AP. https://doi.org/10.1016/bs.pmbts.2020.04.003

Xiong, L., & Yao, Y. (2021). Study on an adaptive thermal comfort model with K-nearest-neighbors (KNN) algorithm. *Building and Environment*, 202, 108026. https://doi.org/10.1016/j.buildenv.2021.108026

Yahia, N. B., Hlel, J., & Colomo-Palacios, R. (2021). From Big Data to Deep Data to Support People Analytics for Employee Attrition Prediction. *IEEE Access*, 9, 60447–60458. https://doi.org/10.1109/ACCESS.2021.3074559

Acta Informatica Pragensia is published by the Prague University of Economics and Business, Czech Republic | eISSN: 1805-4951