I.J. Information Engineering and Electronic Business, **2026, 2, 89-101**
Published Online on April 8, 2026 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijieeb.2026.02.06

Predicting Employee Attrition: A Machine Learning Approach in Human Resource Analytics

Manjula K. A.
Department of Computer Science, University of Calicut, Kerala, India
Email: manjulaka@uoc.ac.in
ORCID iD: https://orcid.org/0000-0002-8916-9448

Karthikeyan P.*
Department of Management Studies, Kannur University, Kerala, India
Email: karthikeyanp@kannuruniv.ac.in
ORCID iD: https://orcid.org/0009-0005-3161-5109
*Corresponding Author

Santhosh V. A.
TKM Insitute of Management, Kollam, Kerala, India
Email: santhoshva@tkmim.ac.in
ORCID iD: https://orcid.org/0000-0002-5873-4483

Received: 22 October, 2025; Revised: 05 December, 2025; Accepted: 07 January, 2026; Published: 08 April, 2026

Abstract: Employee attrition is an important factor that can affect organizations, both financially and operationally. Human Resource (HR) managers often find it difficult to identify exactly which employees might be planning to leave the organization and what is the root cause for their decision. With the recent advances in computing, Machine Learning (ML) techniques are available for analysing, understanding, and solving complex problems. This study analyses the IBM HR Analytics dataset using ML techniques to predict employee attrition and identify the key factors that influence attrition. Four ML models based on Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting have been used for analysing attrition. It is found that Logistic Regression outperformed all other models in predicting attrition. At the same time, Decision Tree is found to be the weakest among the four techniques. On the analysis of feature importance, it is found that variables related to compensation (Monthly Income), career stage (Total Working Years, Age), and tenure at the organization are among the most significant factors influencing attrition. The insight from this study is expected to help HR managers in developing effective, data-driven strategies to retain their talent in their organization**.**

Index Terms: Employee Attrition, Machine Learning, Logistic Regression, Random Forest, Decision Tree, Gradient Boosting, HR Analytics

1. Introduction

Employee attrition has become a critical concern for modern organizations as it directly affects workforce stability, operational efficiency, and long-term organizational performance. This affects both, their employees as well as the operational efficiency of the organization [1]. Employee attrition is a gradual reduction in the workforce through resignations, layoffs, or retirements, where positions remain unfilled for long periods of time. An increase in employee attrition and turnover affects the smooth functioning of organizations, which can lead to a decrease in productivity. In such cases, organizations need to raise funds to hire new employees and train them. If the root causes of employee attrition can be properly identified, human resource teams can develop suitable employee retention strategies. Traditional human resource methods used for studying attrition among employees, rely mainly on surveys or historical averages, which are not capable of accurately capturing the complex patterns of employee behaviour. Machine learning (ML) models are quite popular and are being increasingly used in making predictions, discovering valuable insights from data, and thus solving complex real world problems, including complex human resource related issues in organisations. This data-driven method of using workforce data to understand, predict, and improve how an organization manages its people is known as Human Resource (HR) Analytics. By using HR Analytics, the raw data on

This work is open access and licensed under the Creative Commons CC BY 4.0 License.

human resource that is available with the organisation is turned into insights for better strategy development. This is achieved by combining the domain knowledge of HR, Statistics and ML within the business context.

Many types of ML techniques are available for analyzing, understanding, and solving complex problems like those related to employee attrition in organisations. These include techniques like Logistic Regression (LR), Decision Tree, Random Forest(RF) and Gradient Boosting (GB). ML based techniques also have the ability to detect complex relationships between variables making them ideal for analysing HR related issues.

From the literature related to studies aimed at understanding and predicting employee attrition, it is found that there are conflicting findings regarding the accuracy of various ML techniques used to study employee attrition. Some studies have found Logistic Regression as be a better technique to predict attrition compared to other ML techniques. Fallucch et. al. [1], in their study have used a number of Machine Learning techniques for predicting attrition among employees with the aim of identifying the most efficient technique. They are of the opinion that the model with the best performance in terms of prediction accuracy was Logistic Regression when compared with other techniques. Guerranti and Dimitri [2] have in their study used a number of ML techniques to predict the likelihood of resignation of an employee. Similar to earlier study, they are also of the opinion that Logistic Regression performs the best compared to other techniques in predicting employee atrition. Likewise Sheth et.al. [3] have also found that Logistics Regression had the best accuracy in predicting employee attrition when compared to other ML techniques.

On the contrary, some studies have found other ML techniques to be better at predicting attrition compared to Logistic Regression. Basha et. al. [4] based on their study on attrition, are of the opinion that models like Random Forest and Gradient Boosting Machines demonstrate superior predictive capabilities to forecast attrition. Alharbi et. al. [5] have used Logistic Regression, Naive Bayes, Nearest Neighbour, Random Forest, Gradient Boosting, AdaBoost, and Logit Boost models to predict employee attrition. They have found that Random Forest, achieves the highest accuracy in identifying employees at risk of leaving. Similarly, Talebi et. al. [6] have carried out a systematic review of 58 studies focused on applying machine learning algorithms to predict employee turnover. They have found that Random Forest emerged as the most widely used technique, achieving high predictive accuracy across multiple studies.

Thus it can be seen that there are conflicting opinion regarding the accuracy of various ML techniques used for predicting attrition among employees. Hence this study intends to demonstrate the application and accuracy of selected ML techniques in understanding and predicting employee attrition. In this study the IBM HR Analytics dataset is used for the above purpose. In this study, four ML techniques, namely Logistic Regression, Random Forest, Gradient Boosting, and Decision Tree, are tested on the above mentioned dataset. This study aims to examine the accuracy of the selected ML techniques in predicting employee attrition. Further this study also aims at understanding the attributes that are most influential for attrition among employees and what are the insights that can be obtained regarding employee attrition from this data.

The remaining sections of this article are structured with the next section discussing on the relevant review of literature, followed by section on methodology. This is followed by the section on results and discussion and section that concludes this article.

2. Literature Review

In this section a brief review of relevant literature related to employee attrition and various ML techniques used for studying employee attrition is discussed.

2.1 Employee Attrition

For any business organisation, human resource is considered to be of paramount importance for operating its business activities and accomplishing its business goals. Employees are widely recognized as an essential and valuable asset of a company. As it directly affects workforce stability, operational efficiency, and long-term organizational performance, employee attrition has become a critical concern for modern organisations. Attrition, which is defined as the voluntary or involuntary departure of employees, leads to increased costs like recruitment and training costs, disruption in productivity and morale among remaining staff [7]. It is considered that in this era characterized by rapid technological change, evolving reduced job expectations, and heightened competition for talent, understanding the antecedents of attrition is essential for developing evidence-based HR policies [8]. Information regarding employee turnover and its causes provides organisations with insights into the risk factors, predictive indicators, and behavioural patterns that drive employees to leave [9]. Systematic study of attrition enables firms to design targeted interventions that improve satisfaction, engagement, and retention, thereby supporting strategic workforce planning and ensuring sustainable organizational growth [10]. Given its significant implications, employee attrition remains a vital area of inquiry within human resource management. As employee attrition and other employee withdrawal behaviours are considered to be costly for organizations, large number of studies have been conducted to understand the antecedents of this behaviour. Employee attrition has been the subject of attention for scholars and practitioners alike for a century [7]. The objectives of these researches had been to provide actionable information to organizations, so that they can minimize the costs related to this problem and improve the performance of the organisation.

Frye et al. created models to predict employee turnover based on data from the US Office of Personnel Management and the Bureau of Labour Statistics dataset [11]. Their research highlighted those factors such as length of

service, job grade, age, and salary level serve as statistically significant predictors of turnover. Logistic regression provided a high prediction accuracy of over 74%. The authors highlighted the need to combine quantitative and qualitative factors such as job satisfaction and work-life balance to enhance predictive capabilities.

Zhao et al. have done an extensive study making a comprehensive evaluation of ten supervised machine learning algorithms in predicting employee turnover across human resource datasets of varying types [12]. They have used both real data from a US regional bank and also simulated data from IBM. Their datasets were of various sizes representing organisations having small, medium and large number of employees. From their study they have found that on medium and large datasets, tree-based ensemble methods like extreme gradient boosting technique and gradient boosting trees, consistently outperformed other classifier. While traditional methods such as logistic regression were found to be less effective. Neural networks were found to perform strongly. There are of the opinion that techniques should be selected based on the size and complexity of the dataset. For datasets of small size, the key is to try different algorithms and find the best, but for larger datasets, extreme gradient boosting is found to be more reliable.

Al Akasheh et al. have conducted a review of studies conducted on employee turnover prediction using machine learning techniques over the past ten years [13]. They have reviewed fifty-two high quality studies that have been published between 2012 and 2023 in this topic. The authors have found that more than twenty machine learning approaches have been used for predicting attrition. They have reported that more than ninety-six percentage of studies have used supervised learning. The study finds that salary and overtime are the most important predictors of employee attrition, along with factors such as job satisfaction and career development opportunities. The study also identifies some of the challenges faced by studies on attrition, such as unbalanced datasets and lack of interpretability in some complex machine learning models. This indicates that there is a need for development of more explainable AI approaches.

Talebi et al. have conducted a study in which they have reviewed 58 research papers that have used machine learning techniques for predicting employee attrition [6]. They have analysed a wide range of algorithms that have been used to study attrition, including logistic regression, support vector machines, decision trees, random forests, neural networks, and other ensemble methods. They have found that random forest was the most widely used technique and also this technique is had the best accuracy in predicting attrition. They feel this may be due to the reason that random forest technique was good in handling large and complex datasets. Logistic regression was found to be popular due to its interpretability. They have found that factors like job satisfaction, salary, tenure, and age are found to be the most critical predictors of employee attrition. The authors also pointed out that there is a growing trend in using deep learning models for sequential data, along with the integration of explainable AI tools like LIME and SHAP to improve model transparency. This review provides a comprehensive overview of advanced methodologies and highlights the significance of balancing accuracy with interpretability in HR analytics.

2.2 Job Demands Resources (JD - R) Model

A variety of theoretical models have been widely used for studying attrition among employees and each of these offer different perspectives on why employees leave organizations. The Price-Mueller Turnover Model is one of the influential models which focuses on the importance of job satisfaction, pay, organizational commitment, and perceived alternatives in determining turnover decisions [14]. The Unfolding Model of Voluntary Employee Turnover is another model which underscores the role of shocks either in the form of job offers, conflicts or personal events that stimulate different decision-making paths leading to voluntary turnover [15]. Job Embeddedness Theory, meanwhile, focuses on the degree to which individuals are embedded with their job and their community, through fit, links and sacrifice, and that highly embedded employees are less willing to leave [16]. The Job Demands--Resources (JD-R) Model is another common model that is used to describe how over-reaching job demands and inadequate resources result in burnout and withdrawal behaviour [17, 18]. The above models and others together offer extensive information on the psychological, organizational, and behavioural causes of attrition and are central to the current study of employee turnover.

As employee attrition and other employee withdrawal behaviours are considered to be costly for organizations, large number of studies have been conducted to understand the antecedents of this behaviour. Employee attrition has been a subject of interest for scholars and practitioners during the past century [7]. The objectives of these studies have been to provide actionable information to organizations, so that they can minimize costs related to this problem and improve the performance of the organization.

This study is based on the Job Demands - Resources (JD - R) Model. This is a widely used framework for examining and explaining turnover among the employees. According to this model, any job can be defined as a combination of job demands and job resources that have the ability to influence the attitudes and behaviour of employees [17, 18]. Job Demands are aspects of the job that require sustained physical or psychological effort (e.g., workload, overtime, work-life imbalance and stressful work conditions) that can cause strain, fatigue and burnout if excessive. In contrast, job resources are those physical, social, and organizational components that contribute to employees' ability to accomplish work goals, minimize demands, and promote motivation such as supervisor support, compensation, career opportunities, and positive work environment.

The JD-R model suggests two core processes that affect the rate of attrition for employees. The health-impairment pathway proposes that the accumulation of energy link to exhaustion due to long exposure to high job demands that lead to leaving employees with exhaustion and withdrawal behaviour that eventually leads to leaving the organization. The motivational pathway suggests that when jobs have strong resources, they lead to higher levels of engagement,

satisfaction, and commitment to the organization, respectively, lower turnover intentions. Together, these dual mechanisms offer a comprehensive explanation of the process that employees engage in to decide to stay or depart. This framework allows researchers to gain a structured understanding of how stressors and supports interact with each other to influence turnover outcomes, which makes it a strong and theoretically sound basis for exploring employee attrition and for developing employee retention strategies. The framework of this model is given in Fig.1. Numerous studies have applied this model in the study of attrition of employees [19,20,21].

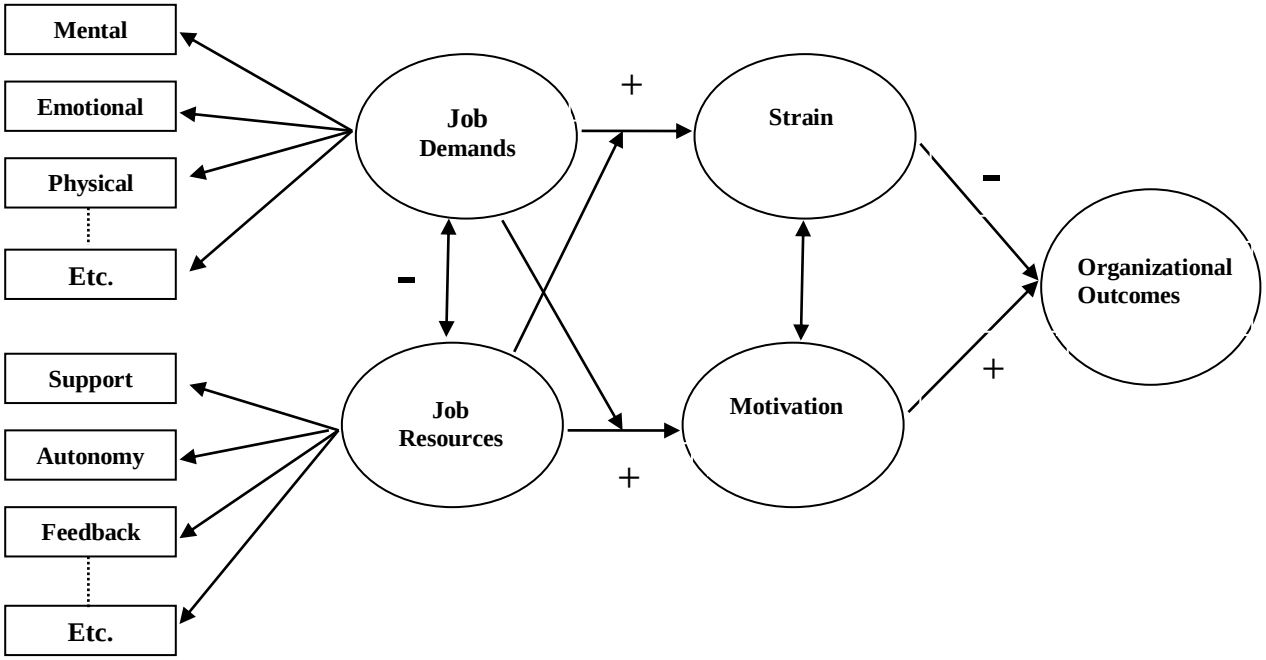


Fig. 1. The Job Demands-Resources model [18].

2.3 Employee Attrition and Machine Learning Techniques

Over the years, large numbers of techniques have been used for examining employee attrition. These methodologies have been continuously evolving over time from basic statistical analysis to recent methodologies using ML techniques. It is considered that the traditional statistics based analysis method used for examining employee attrition does not have a high enough model prediction accuracy and also has the limitation of having a particularly poor performance while analysing data that are complex like those related to employees [22]. With the advances in data analysis and computing techniques, various advanced techniques like the ML techniques are now being widely used in analysing employee attrition. These techniques include Logistic Regression, Decision Tree, Random Forest, Support Vector Machines (SVM), Gradient Boosting, Neural Networks etc…

From the literature it is seen that numerous ML approaches have been applied in studies related to employee attrition. Logistic Regression is one of the most widely used technique because of its simplicity, ease of interpretation and ability to work with small to moderately sized datasets [23]. Another technique is Decision Tree that is believed to present transparent rule-based conclusions, support both numerical and categorical predictors, and reflect transparency that required in managerial decision-making [24]. Random Forests has been reported to perform well on structured tabular data with little or no tuning and resistant to overfitting which makes it effective technique in small datasets [25]. Gradient Boosting algorithms, like XGBoost are routinely more effective than more intricate models on structured/tabular information, especially when data is small [26].

Contrastingly, SVMs are said to be unable to cope with large datasets and datasets with categorical values [27]. This technique involves a lot of parameter tuning which makes this technique less practical to be applied to the type of dataset used in this study. Likewise, it is generally assumed that Neural Networks are highly resource-intensive and demand both massive datasets and significant training to produce consistent results, and they frequently perform poorly with tabular data in comparison with tree-based ensembles [28,29]. So SVM and Neural Networks are not always the best choices, since studies on employee attrition involve cases of small datasets, mixed variables, and interpretability.

The dataset used in this study is the IBM HR Analytics Employee Attrition dataset (1470 records) which is a small, structured, mixed-type tabular dataset with many categorical and ordinal variables. This renders SVM and Neural Networks as a less appropriate modeling choice. In combination with these constraints, the practical and computational demands and the interpretability requirements of HR research makes classical tree based models more suitable to studies of this kind. So this study uses four classical tree based ML techniques, namely Logistic Regression, Random Forest, Gradient Boosting, and Decision Tree for predicting and understanding employee attrition based on the above mentioned dataset.

3. Machine Learning Techniques used for Analysis

3.1 Logistic Regression

Logistic Regression is a statistical method, well-suited for solving many classification problems across a wide range of applications. The popularity of this technique is especially due to its simplicity and interpretability. It is a method that has been extensively employed in the modeling of any situation where the dependent variable is discrete or categorical. Rather than forecasting the precise occurrence of an event, Logistic Regression is used to determine the likelihood of event occurrence, with the anticipated value falling between 0 to 1, which enables researchers to categorize the results with respect to a threshold, which is usually 0.5 [30]. In contrast to the Linear Regression, which presupposes a straight-line dependency between the predictors and the response, the Logistic Regression models the nonlinear relationship with the help of the logistic function, forming a kind of an S-shaped curve that is more capable of describing the behaviour of binary outcomes. Under this model, independent variables are those aspects that drive the event whereas the dependent variable is the event being carried out or not carried out. Logistic Regression is thus quite appropriate to study dichotomous outcomes, which allows the researcher to assess the risk of an occurrence and how the predictor variables play part in changing the probability. It is a strong tool in empirical studies in terms of classification and risk estimation due to its interpretability, flexibility, and the fact that it can deal with various types of predictors.

3.2 Decision Tree

The Decision Tree is a classification method which is widely employed to predict categorical/continuous outcomes by recursively dividing data into homogeneous subsets based on the predictors. Decision trees make predictions using tree-like structures and branches. Decision trees are known as one of the earliest and most interpretable machine learning methods. They are used to construct a tree-like structure of decisions by repeatedly dividing data into branches based on feature values. This approach was formalized by the CART framework [31] and later developed by Quinlan's ID3 and C4.5 algorithms [32,33]. Although decision trees are easy to interpret and implement, they often suffer from overfitting when used alone [34].

3.3 Random forest

Breiman introduced the Random Forest algorithm [25], which is an ensemble learning method. It works by combining multiple decision trees with random feature selection and bagging to achieve high accuracy and robustness. It is seen that the generalization error of random forests converges as the number of trees increases, thereby reducing the risk of overfitting. The Random Forest method is more robust to noise and outliers. Since the method provides internal estimates of error and variable importance, it is highly interpretable for practical applications. Random forests can be considered a valuable tool for predictive modelling in diverse domains, especially due to its robustness and ability to handle high volumes of data, including employee turnover prediction. Ensemble learning approaches can capture complex patterns in workforce data and can increase the accuracy of turnover prediction. Empirical results have shown that random forests perform competitively with boosting methods such as Adaboost [35] and gradient boosting [36]. Adaboost is one of the earliest boosting algorithms that iteratively combines weak classifiers, adjusting weights to focus on misclassified instances. Gradient boosting extends boosting by using gradient descent optimization to minimize the selected loss function, often achieving higher accuracy, but can be more sensitive to noise.

3.4 Gradient Boosting

Gradient Boosting is a highly sophisticated ensemble ML method. It constructs an effective predictive model, through sequential combination of large numbers of weak learners, which are usually shallow decision trees. Gradient Boosting builds trees sequentially as opposed to a single tree to categorize the results by creating trees that are designed to minimize the errors of its predecessors and thus increasing the overall accuracy [36]. Gradient Boosting is effective in modeling of complex, nonlinear relationships where unlike linear or single-tree models, it only captures limited patterns, thereby minimizing the loss through gradient descent. The model is capable of achieving high performance on structured tabular datasets through this approach which enables the model to refine its predictions incrementally. Gradient Boosting has become one of the most effective predictive techniques used in applied research since it can work with mixed types of data, requires minimum pre-processing and predictive power.

The observations from previous research clearly indicate that the prediction of employee turnover or attrition has progressed from conventional interpretable models, such as Logistic Regression, to sophisticated ensemble techniques like Random Forest and Gradient Boosting. Ensemble techniques are known for their consistently superior performance over single classifiers, especially while working with larger and more complex datasets. Logistic Regression continues to be popular due to its transparency and ease of interpretation. Decision Trees offer straightforward and rule-based impactful insights. Random Forest and Gradient Boosting are considered to be leading methods because they are able to manage non-linear relationships, high-dimensional data, and imbalanced class distributions with enhanced accuracy. Based on these findings, the current study applies Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting to evaluate their effectiveness and accuracy in forecasting employee attrition. In this way, a comparison between traditional, interpretable methods and cutting-edge ensemble techniques will be possible. This type of analysis

will provide both theoretical insights and practical recommendations for HR analytics applications. This study is expected to be helpful to HR practitioners by providing them with insight and evaluation of the traditional methods and recent ML based methods.

Further this study also intends to understand the most important attributes or features that influence employee attrition based on the data used for this study. According to Zhou and Hooker [37], single decision trees suffer from the drawback of high variance and instability, and in some situations, they are too simple to capture complex signals in the data. These researchers are of the opinion that ensemble methods like Random Forest and Gradient Boosting correct these weaknesses through bagging and boosting and so are better options. They also state that feature importance is natively and reliably implemented for Random Forests and Gradient Boosting in scikit-learn. Wang [38] is of the opinion that Random Forests ranks features based on their contribution to predictive accuracy and is less prone to over fitting compared to single decision trees. Saarela et al. [39] in their study conclude that logistic regression coefficients are model-dependent and yield different importance rankings compared to tree-based models. Hence understand the importance of various features on the attrition of employees and feature ranking ensemble methods like Random Forest and Gradient Boosting are used in this study.

4. Methodology

The methodology followed in this study is represented in Fig.2, and the process followed is discussed below.



Fig. 2. Methodological Process for this Study

4.1 Dataset

This study makes use of a publicly available dataset, the IBM HR Analytics Attrition Dataset [40]. This is a HR dataset created by IBM for demonstration and educational purposes and is widely used benchmark dataset for studying employee turnover, predictive modelling, and HR analytics. This dataset has 1,470 employee records with 35 attributes. It is a mixed tabular dataset comprising of categorical, ordinal and continuous, variables. The target dependent variable is Attrition, which is categorical variable (Yes/No). Age, Department, Job Role, Monthly Income, Overtime, Work Life Balance, Years at Company and Job Satisfaction are a few of the other attributes present and are of categorical, ordinal and continuous, variables types.

4.2 Data Pre-Processing

Data pre-processing was carried out to prepare the dataset for analysis. The target attribute Attrition which was a categorical variable was converted into binary values, where Yes was encoded as 1 and No as 0. Non-informative columns such as Employee Number and Standard Hours were removed. Categorical features were encoded using one-hot encoding, and numerical features were normalized with Standard Scaler to ensure comparability across variables, making it suitable for logistic regression. To address the problem of class imbalance, a condition where one class (majority class) contains significantly more observations than the other (minority class), the Synthetic Minority Over-sampling Technique (SMOTE) [41] was used. This is a commonly used technique in machine learning because imbalances in data often causes predictive models to become biased toward the majority class, resulting in poor detection of the minority class, which is usually the class of primary interest. Traditional oversampling methods simply duplicate minority samples, which can lead to overfitting. In contrast, SMOTE algorithm first finds the nearest neighbour of the upsampled data and calculates the distance between the two. The distance is then multiplied by an arbitrary numerical value which is then added into the original sample to create new synthetic instances.

4.3 Data Splitting

In the next step, the dataset was split into two sets. One for training (70%) and the other for testing (30%), with stratification based on the attrition ratio to maintain class balance.

4.4 Machine Learning Models

The Four classification models has been considered in this study, and they were implemented using the scikit-learn library in Python. Logistic Regression was used as an interpretable linear baseline model. A Decision Tree Classifier was used as a simple non-linear learner that is able to manage feature interactions. To improve the accuracy of prediction, a Random Forest Classifier consisting of 400 decision trees was utilized as an ensemble method. Also, a Gradient Boosting Classifier was used, which builds models sequentially by boosting weak learners to enhance overall accuracy.

4.5 Performance Metrics

This research work employed multiple standard performance metrics to evaluate the performance of the models. In the first instance the performance of the models under metrics like Accuracy, Precision, Recall and F1 Score were calculated. The Accuracy metric gives the measure of overall proportion of correct predictions. The Precision metric refers to the ability of the model to correctly identify positive cases and it is the proportion of true positives among all instances predicted as positive. Recall evaluates the model's effectiveness in capturing all actual positive instances. The F1-score, as the harmonic mean of precision and recall, provides a balanced measure of predictive performance. In addition, Next the confusion matrix, a widely used performance evaluation tool that summarizes the predictions made by a model relative to the actual outcomes is used to evaluate the models. Also, Receiver Operating Characteristic (ROC) curve analysis is done to enable a comparative evaluation of the models across different threshold settings. Finally, feature importance plots are constructed for the Random Forest and Gradient Boosting classifiers to identify the most crucial variables influencing the employee attrition prediction.

5. Results and Discussion

The results of the implemented classification models are presented in this section as tables and graphs along with a discussion on these. Model performance is evaluated on the test dataset using standard metrics, including accuracy, precision, recall, F1-score, and ROC-AUC. In addition, confusion matrices, ROC curves, and feature importance plots are provided to enable a comprehensive assessment of predictive effectiveness.

5.1 Model Performance Metrics

The following Table 1 represents model performance on the test dataset based on the various performance metrics.On the comparison of performance of the four models as per the metrics used for performance evaluation, it is seen that Logistic Regression performs better than the other models. It has the highest accuracy (88%), proving it as effective and accurate technique for the employee attrition prediction. In the case accuracy, Gradient Boosting (84%) and Random Forest (83%) models follow Logistics Regression model. Decision Tree (74%) is found to have the least accuracy among the models considered. In the case of Recall value, Logistics Regression (0.40) is the best. This metric for other models, Gradient Boosting (0.23), Decision Tree (0.23) and Random Forest (0.14) shows poor performance. This low recall values may be due to their limited ability to capture actual attrition cases. In case of precision also, Logistics Regression (0.72) has highest value, and the values for this metric for Gradient Boosting (0.53), Random Forest (0.43) and Decision Tree (0.23) are much lower. The F1 score for Logistic Regression (0.52) is the highest followed by Gradient Boosting (0.33), Decision Tree (0.23) and Random Forest (0.21).

Table 1. Comparison of machine learning models for employee attrition prediction

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.88	0.725	0.408	0.523
Gradient Boosting	0.844	0.531	0.239	0.33
Random Forest	0.832	0.435	0.141	0.213
Decision Tree	0.748	0.23	0.239	0.234

5.2 Confusion Matrix

A confusion matrix is a widely used performance evaluation tool in classification problems, summarizing the predictions made by a model relative to the actual outcomes. This section describes the comparison of the confusion matrices corresponding to the four models. Table 2 shows the values for these models.

Table 2. Confusion Matrix

Model	TN	FP	FN	TP
Logistic Regression	359	11	42	29
Gradient Boosting	355	15	54	17
Random Forest	357	13	61	10
Decision Tree	313	57	54	17

Attrition is a minority class in the dataset used for the study. This is due to the reason that there are far fewer employees leaving than stay in the organisation. The confusion matrix provides a clearer assessment of performance than accuracy alone, highlighting the model's ability to correctly detect employees who may leave. The confusion matrix helps in evaluating how effectively a model distinguishes between employees who leave and those who stay. True Positives (TP) represent employees who actually left and were correctly identified, while True Negatives (TN) are those who stayed and were correctly classified. False Positives (FP) occur when the model predicts an employee will leave even though they remain, and False Negatives (FN) represent employees who leave but are incorrectly predicted to stay.

From Table 2 and Fig. 3, it can be seen that the best balance is observed in case of Logistic Regression model. This model has correctly predicted the most "stay" cases (TN = 359) and also captured the highest number of "leavers" (TP = 29). It still missed 42 leavers (FN), but this was fewer compared to other models. From this, Logistic Regression model is found to be the model with the best prediction accuracy among the four models used in this study as this model gives fewer mistakes and more correct predictions employee attrition.

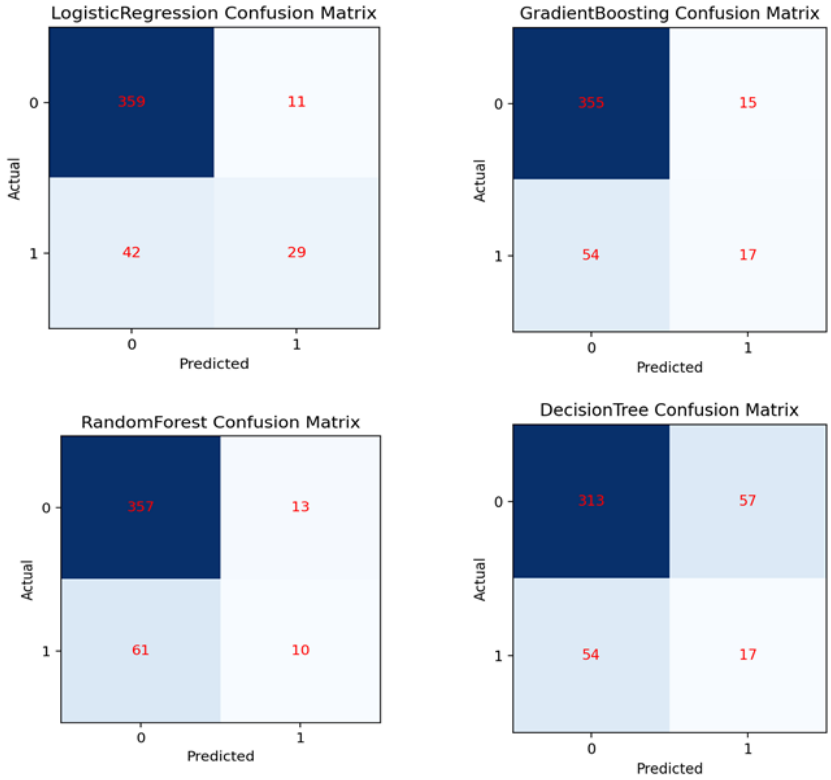


Fig. 3. Confusion Matrix

5.3 ROC Curve Analysis

Receiver Operating Characteristic (ROC) Curve analysis is another widely used method for evaluating the performance of classification models. This examines the ability of models to distinguish between positive and negative classes. The ROC curve plots the True Positive Rate (TPR or Sensitivity) on the y-axis against the False Positive Rate (FPR) on the x-axis across various classification thresholds [42]. Each point on the curve represents a different threshold used to convert predicted probabilities into class labels. A model that performs well will achieve a high TPR while maintaining a low FPR, producing a curve that bows toward the upper-left corner of the graph. The Area Under the Curve (AUC) is used as a single numerical summary of performance, with values closer to 1 indicating strong discriminatory ability and values near 0.5 suggesting no better performance than random guessing. ROC analysis is particularly valuable for imbalanced datasets, such as employee attrition, because it provides insight into how effectively a model identifies minority-class instances across different threshold settings. Fig. 4 shows the ROC curve for the various techniques used in this study.

From the Fig. 4 it can be seen that the highest AUC is for Logistics Regression (0.82), This is followed by Gradient boosting and Random Forest which have AUC of 0.78. Decision Tree is found to have the smallest AUC (0.54). Based on the graph and AUC values it can be concluded that Logistic Regression is the best technique compared to other techniques used in the study. The performance of Decision Tree is not satisfactory compared to other techniques.

Thus based on the analysis of the selected performance metrics, the confusion matrix and ROC curve analysis with respect to the four models used in this study, it can be inferred that in all instances Logistics Regression is the model that exhibits best performance in predicting employee attrition among the models consider for this study. This is similar to results that have been reported by other similar studies on attrition [1,3]. It is also noted that the Decision Tree model showed the weakest performance in most of the metrics among the models used, and this probably may be due to overfitting and limited ability to generalize to new data. These findings suggest that, for structured HR datasets, simpler interpretable models such as Logistic Regression can deliver superior predictive performance compared to more advanced ensemble methods. Ensemble techniques like Random Forest and Gradient Boosting are powerful, and their advantages are more evident in complex or high-dimensional datasets. Thus, model selection should be guided by both dataset characteristics and the need for interpretability in HR analytics.

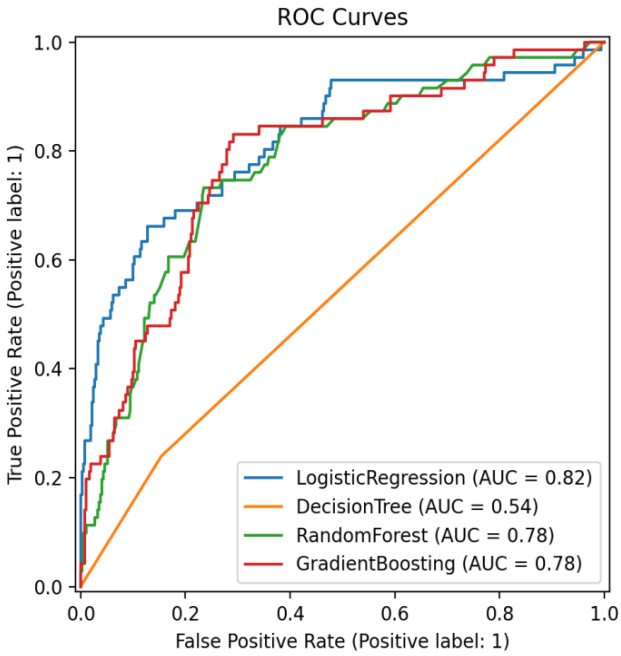


Fig. 4. ROC Curves.

5.4 Feature Importance

Feature importance is a key interpretability measure in machine learning that identifies how strongly each predictor variable contributes to a model's decision-making process. In classification problems such as employee attrition analysis, feature importance helps determine which factors have the greatest influence on whether an employee is likely to leave or stay. Higher importance values indicate that a variable plays a larger role in predicting the outcome, whereas lower values suggest limited influence. In this study the feature importance analysis was conducted using two ensemble methods Random Forest and Gradient Boosting. Table 3 shows the importance of the top ten ranked features that influence attrition based on the two techniques used for analysis.

Table 3. Comparison of Feature Importance

Rank	Gradient Boosting (GB)	Importance (GB)	Random Forest (RF)	Importance (RF)
1	MonthlyIncome	0.124	MonthlyIncome	0.082
2	TotalWorkingYears	0.097	Age	0.067
3	OverTime_Yes	0.094	TotalWorkingYears	0.065
4	Age	0.08	DailyRate	0.049
5	DailyRate	0.058	HourlyRate	0.047
6	YearsWithCurrManager	0.056	MonthlyRate	0.044
7	StockOptionLevel	0.055	YearsAtCompany	0.041
8	NumCompaniesWorked	0.047	DistanceFromHome	0.04
9	EnvironmentSatisfaction	0.046	YearsWithCurrManager	0.039
10	WorkLifeBalance	0.045	OverTime_Yes	0.039

As per the Gradient Boosting model, monthly income is the most important factor influencing attrition. This is followed by total working years which shows the experience, overtime which denotes workload, age, daily rate that shows the remuneration, years with current manager that denotes the supervisor etc… With respect to Random Forest model, monthly income is the most important factor similar to Gradient Boosting model. The other factors as per importance in the case of Random Forest are age, total working years which shows the experience, daily rate, hourly rate and monthly rate etc… The top twenty factors that influence attrition as per Gradient Boosting model and Random Forest model are shown in Fig. 5.

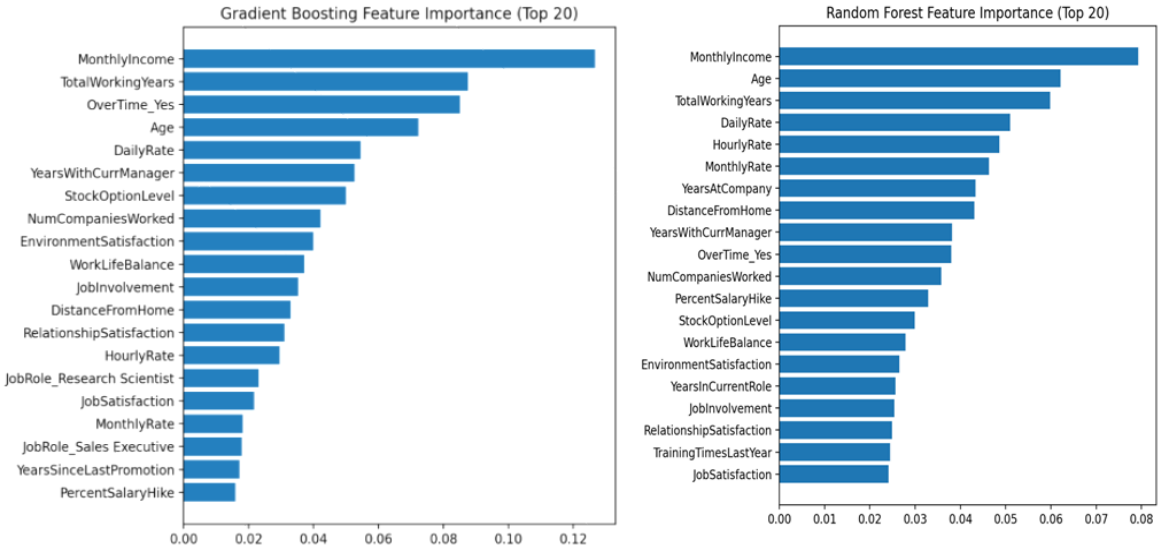


Fig. 5. Feature Importance Corresponding to Gradient Boosting and Random Forest

It is evident from the figures and tables, that both Gradient Boosting and Random Forest models identified Monthly Income, Age, and Total Working Years as dominant predictors. Also, the models diverged while predicting secondary factors. Gradient Boosting prioritized workload and incentive-related variables such as Overtime, Stock Option Level, and Work-Life Balance, whereas Random Forest highlighted demographic and structural factors like Daily/Hourly Rate, Distance from Home, and Years at Company. From this complementary observation, it can be understood that employee attrition is not determined by a rigid set of factors, rather it is influenced by a blend of financial, demographic, and experiential determinants.

It can be seen from Table 3 and Fig. 5 that both Gradient Boosting and Random Forest models have identified monthly income as the most influential predictor. This indicates that compensation plays a central role in employees' decisions to stay or leave. This finding aligns with HR literature emphasizing that financial rewards are key job resources that enhance retention. Total working years is seen to have high importance in both the models, suggesting that employees with limited experience or early-career employees are more likely to resign than experienced employees who will stay longer. Overtime emerges as a major predictor in the Gradient Boosting model, reflecting the role of workload and job demands in increasing turnover risk, while age holds strong influence in both models, indicating that attrition patterns differ across career stages.

Other predictors such as daily rate, years with current manager, stock option level, and number of companies worked also contribute significantly in Gradient Boosting, revealing the combined effects of compensation structures, managerial relationships, and career mobility on attrition. In the Random Forest model, variables like hourly rate, monthly rate, years at company, and distance from home appear among the top predictors, demonstrating that both financial and contextual job characteristics shape employee decisions.

6. Conclusion & Future Work

In this study, employee attrition was studied applying selected ML techniques on the IBM HR Analytics Attrition Dataset. Four ML techniques – Logistic Regression, Gradient Boosting, Random Forest and Decision Tree was used for analysing the data. The models based on these four techniques were evaluated for prediction precision using certain performance metrics. The metrics that was used for this were Accuracy, Precision, Recall, F1 Score, Confusion Matrix Receiver Operating Characteristic curve analysis. It was seen that all these metrics showed that the model based on Logistic Regression techniques was the best model in predicting attrition. It was also seen that while the models based on Gradient Boosting and Random Forest model were good at prediction, the model based on the Decision Tree technique had the least accuracy.

Feature importance analysis was conducted to identify the factors that had largest influence on attrition. For this, two ML techniques Gradient Boosting and Random Forest were as it was seen that ensemble techniques were best for analysing feature importance. It was found that factors related to compensation and career stage (e.g., income, age, total working years) dominated in both models. However, in the case of secondary factors, there was difference among these two models. While the Gradient Boosting model highlighted workload and incentive-related factors (e.g., overtime, stock options, work–life balance), Random Forest model emphasized demographic and structural features (e.g., tenure, commuting distance). Altogether, these findings suggest that a hybrid approach need to be adopted in analysing problems like attrition. While the interpretable models will give transparency, the ensemble methods will be useful for identifying the fine-grained patterns in the data. Such hybrid approach will offer the most comprehensive framework for HR practitioners to predict and effectively manage employee attrition.

This research highlights how machine learning can be useful in predicting employee attrition and identifying the factors that influence it. Organizations, with the help of these techniques, can predict attrition and also identify the primary factors that cause attrition. This will help in understanding attrition effectively and develop appropriate retention strategies to avoid attrition. The dataset used is in this study is synthetic, and hence it might affect how widely applicable these findings are. Further research works can also be undertaken in similar methodology to address other HR related challenges, such as predicting promotions, analysing absenteeism, or modelling career progression

Author Contributions Statement

Manjula K. A. – Data Curation and Software Implementation: Handled data acquisition, dataset preprocessing, and analysis.

Karthikeyan P. – Conceptualization, Methodology, Constructed the overall framework and writing.

Santhosh V. A. – Conceptualization, Methodology and Review.

All authors have read and agreed to the published version of the manuscript.

Conflict of Interest Statement

The authors declare no conflicts of interest.

Funding Declaration

This research was not supported by the grant.

Data Availability Statement

This study analyzed publicly available datasets. The results obtained and datasets can be found at https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset

Ethical Declarations

This study does not involve any human and/or animal subjects.

Acknowledgments

We sincerely thank the reviewers for their professional evaluation and valuable recommendations, which have contributed to improving the quality of this work.

Declaration of Generative AI in Scholarly Writing

In this article AI and AI-assisted technologies have not been used for the writing process.

Abbreviations

The following abbreviations are used in this manuscript:

HR – Human Resource
ML – Machine Learning

References

[1] F.Fallucchi, M. Coladangelo, R. Giuliano, and E. W. De Luca, "Predicting employee attrition using machine learning techniques," Computers, vol. 9, no. 4, p. 86, 2020.

[2] Guerranti F, Dimitri GM. A Comparison of Machine Learning Approaches for Predicting Employee Attrition. Applied Sciences. 2023; 13(1):267.

[3] Sheth, Kalgi, Jaynil Patel, and Jaiprakash Verma. "Machine learning-based investigation of employee attrition prediction and analysis." In Emerging Technology Trends in Electronics, Communication and Networking: Select Proceedings of the Fourth International Conference, ET2ECN 2021, pp. 221-238. Singapore: Springer Nature Singapore, 2022.

[4] M. S. Amzad Basha, O. Varikunta, A. U. Devi and S. Raja, "Machine Learning in HR Analytics: A Comparative Study on the Predictive Accuracy of Attrition Models," 2024 2nd International Conference on Device Intelligence, Computing and Communication Technologies (DICCT), Dehradun, India, 2024, pp. 475-480.

[5] G. M. Alharbi, H. M. Alharthi, N. K. Alharthi, S. M. Alotaibi and H. Meshref, "Predicting Employee Attrition Using Machine Learning Approaches," 2025 Intelligent Methods, Systems, and Applications (IMSA), Giza, Egypt, 2025, pp. 650-655.
[6] H. Talebi, A. Khatibi Bardsiri, and V. K. Bardsiri, " Machine Learning Approaches for Predicting Employee Turnover: A Systematic Review," Engineering Reports 7, no. 8, 2025.
[7] Hom, Peter W., Thomas W. Lee, Jason D. Shaw, and John P. Hausknecht. "One hundred years of employee turnover theory and research." Journal of applied psychology 102, no. 3 (2017): 530.
[8] Holtom, Brooks C., Terence R. Mitchell, Thomas W. Lee, and Marion B. Eberly. "5 turnover and retention research: a glance at the past, a closer review of the present, and a venture into the future." The Academy of Management Annals 2, no. 1 (2008): 231-274.
[9] Hausknecht, John P., and Charlie O. Trevor. "Collective turnover at the group, unit, and organizational levels: Evidence, issues, and implications." Journal of management 37, no. 1 (2011): 352-388.
[10] Allen, David G., Phillip C. Bryant, and James M. Vardaman. "Retaining talent: Replacing misconceptions with evidence-based strategies." Academy of management Perspectives 24, no. 2 (2010): 48-64.
[11] Frye, Alex, Christopher Boomhower, Michael Smith, Lindsay Vitovsky, and Stacey Fabricant. "Employee attrition: what makes an employee quit?." SMU Data Science Review 1, no. 1 (2018)
[12] Y. Zhao, M. K. Hryniewicki, F. Cheng, B. Fu, and X. Zhu, "Employee turnover prediction with machine learning: A reliable approach," in Proc. SAI Intelligent Systems Conf., Cham: Springer International Publishing, Sep. 2018, pp. 737–758.
[13] M. Al Akasheh, E. F. Malik, O. Hujran, and N. Zaki, "A decade of research on machine learning techniques for predicting employee turnover: A systematic literature review," Expert Systems with Applications, vol. 238, p. 121794, 2024.
[14] Price, J. L., & Mueller, C. W. Absenteeism and turnover of hospital employees. (1986): JAI Press.
[15] Lee, Thomas W., and Terence R. Mitchell. "An alternative approach: The unfolding model of voluntary employee turnover." Academy of management review 19, no. 1 (1994): 51-89.
[16] Mitchell, Terence R., Brooks C. Holtom, Thomas W. Lee, Chris J. Sablynski, and Miriam Erez. "Why people stay: Using job embeddedness to predict voluntary turnover." Academy of management journal 44, no. 6 (2001): 1102-1121.
[17] Demerouti, Evangelia, Arnold B. Bakker, Friedhelm Nachreiner, and Wilmar B. Schaufeli. "The job demands-resources model of burnout." Journal of Applied psychology 86, no. 3 (2001): 499.
[18] Bakker, Arnold B., and Evangelia Demerouti. "The job demands-resources model: State of the art." Journal of Managerial Psychology 22, no. 3 (2007): 309-328.
[19] Bakker, Arnold, Evangelia Demerouti, and Wilmar Schaufeli. "Dual processes at work in a call centre: An application of the job demands–resources model." European Journal of work and organizational psychology 12, no. 4 (2003): 393-417.
[20] Qin, Xin, Peter Hom, Minya Xu, and Dong Ju. "Applying the job demands–resources model to migrant workers: Exploring how and when geographical distance increases quit propensity." Journal of occupational and organizational psychology 87, no. 2 (2014): 303-328.
[21] Carlson, John R., Dawn S. Carlson, Suzanne Zivnuska, Ranida B. Harris, and Kenneth J. Harris. "Applying the job demands resources model to understand technology as a predictor of turnover intentions." Computers in Human Behavior 77 (2017): 317-325.
[22] Chung, Doohee, Jinseop Yun, Jeha Lee, and Yeram Jeon. "Predictive model of employee attrition based on stacking ensemble learning." Expert Systems with Applications 215 (2023).
[23] Hosmer Jr, David W., Stanley Lemeshow, and Rodney X. Sturdivant. Applied logistic regression. John Wiley & Sons, 2013.
[24] Breiman, Leo, Jerome Friedman, Richard A. Olshen, and Charles J. Stone. Classification and regression trees. Chapman and Hall/CRC, 2017.
[25] Breiman, Leo. "Random forests." Machine learning 45, no. 1 (2001): 5-32.
[26] Chen, Tianqi. "XGBoost: A Scalable Tree Boosting System." Cornell University (2016).
[27] Hsu C-W, Chang C-C, Lin C-J. A practical guide to support vector classification. Department of Computer Science and Information Engineering, National Taiwan University, Taipei, 2016.
[28] Shavitt, Ira, and Eran Segal. "Regularization learning networks: deep learning for tabular datasets." Advances in neural information processing systems 31 (2018).
[29] Gorishniy, Yury, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. "Revisiting deep learning models for tabular data." Advances in neural information processing systems 34 (2021): 18932-18943.
[30] Jin, Soo-Bong, and Jong-Woo Lee. "Study on accident prediction models in urban railway casualty accidents using logistic regression analysis model." Journal of the Korean Society for Railway 20, no. 4 (2017): 482-490
[31] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, Classification and Regression Trees. Belmont, CA, USA: Wadsworth, 1984.
[32] J. R. Quinlan, "Induction of decision trees," Machine Learning, vol. 1, no. 1, pp. 81–106, 1986.
[33] J. R. Quinlan, C4.5: Programs for Machine Learning. San Mateo, CA, USA: Morgan Kaufmann, 1993.
[34] K. Thajeel, K. Samsudin, S. J. Hashim, and F. Hashim, "Machine and deep learning-based XSS detection approaches: A systematic literature review," J. King Saud Univ.-Comput. Inf. Sci., vol. 35, Jul. 2023, Art. no. 101628.
[35] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," in Proc. European Conf. Computational Learning Theory, Berlin, Heidelberg: Springer, pp. 23–37, Mar. 1995.
[36] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001
[37] Zhou, Zhengze, and Giles Hooker. "Unbiased measurement of feature importance in tree-based methods." ACM Transactions on Knowledge Discovery from Data (TKDD) 15, no. 2 (2021): 1-21.
[38] Wang, Y. Analysis of Feature Importance Based on Random Forest for Stock Selection. In Proceedings of the 1st International Conference on E-commerce and Artificial Intelligence (ECAI 2024), 2024 pages 374-379
[39] Saarela, Mirka, and Susanne Jauhiainen. "Comparison of feature importance measures as explanations for classification models." SN Applied Sciences 3, no. 2 (2021): 272.
[40] IBM, "IBM HR Analytics Employee Attrition & Performance," Kaggle, 2017. [Online].

[41] Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. "SMOTE: synthetic minority over-sampling technique." Journal of artificial intelligence research 16 (2002): 321-357.
[42] Fawcett, Tom. "An introduction to ROC analysis." Pattern recognition letters 27, no. 8 (2006): 861-874.

Authors' Profiles



Manjula K. A. is presently an Associate Professor in the Department of Computer Science at the University of Calicut. She completed her MCA and PhD in Computer Science, and has postgraduate teaching experience of more than 20 years. She has published widely through journal articles, book chapters, and invited talks. Her research areas span contemporary aspects of Computer Science to the Business Analytics domain, and she continues to be actively interested in academic and professional activities.



Karthikeyan P. is an Assistant Professor in the Department of Management Studies at Kannur University, with over 20 years of postgraduate teaching experience. He holds MBA, MSc, and PhD (Management) degrees. His academic and research interests include Finance and Business Analytics.



Santhosh V. A. currently Director, TKM Institute of Management, Kollam, Kerala, is an accomplished academic administrator, professor, trainer, psychotherapist, and author. A Ph.D. holder from the University of Kerala with over 15 years of post-doctoral experience, Dr. Santhosh has guided five research scholars to the completion of their PhDs, and published extensively in reputed journals, edited volumes, and national and international conferences. He has authored the acclaimed book "Hallucinations of Differently Abled Living Beings", co-edited two volumes, and contributed several scholarly chapters.

How to cite this paper: Manjula K. A., Karthikeyan P., Santhosh V. A., "Predicting Employee Attrition: A Machine Learning Approach in Human Resource Analytics", International Journal of Information Engineering and Electronic Business(IJIEEB), Vol.18, No.2, pp. 89-101, 2026. DOI:10.5815/ijieeb.2026.02.06