10.48047/jocaaa.2024.33.08.283
ISSN1521-1398

Predicting Employee Retention and Satisfaction in IT Companies Using Random Forest–Based HR Analytics

Rashi Dubey* and Manish Dhingra
Faculty of Commerce and Management
Rama University Kanpur, India
rashidubey.2010@gmail.com*, dhingramanish2003@yahoo.com

Abstract

Employee retention and satisfaction remain two of the most critical challenges facing information technology (IT) organizations in the modern, knowledge-driven economy. High attrition rates not only increase recruitment and training costs but also disrupt productivity and knowledge continuity. This study presents a comprehensive Random Forest–based HR analytics framework designed to predict employee retention and estimate satisfaction levels by integrating multi-source HRMS, performance, and engagement data. The proposed approach leverages ensemble learning techniques to capture nonlinear interactions among organizational, behavioral, and career-related factors, while ensuring interpretability through explainable AI methods such as SHAP and permutation importance. Using a dataset of 7,500 employees over a three-year period, the Random Forest classification model achieved an accuracy of 0.89 and an AUC of 0.87 in predicting attrition, while the regression model explained 69% of the variance in satisfaction scores (R^2 = 0.69). Key determinants influencing retention included tenure, promotion velocity, performance trend, compensation percentile, and managerial effectiveness, whereas work–life balance, manager support, and training opportunities emerged as dominant predictors of satisfaction. The findings align with established organizational theories—such as Herzberg's Two-Factor Theory and Social Exchange Theory—validating that both intrinsic motivators and extrinsic conditions jointly shape employee attitudes and turnover behavior. This research contributes to the field of HR analytics by demonstrating how interpretable machine learning can complement evidence-based HRM practices. It provides actionable insights for workforce planning, leadership development, and equitable compensation design. The study concludes that predictive HR analytics, when combined with ethical governance and transparent interpretability, can transform human resource management from reactive decision-making into a proactive, data-informed strategic discipline.

Keywords:
Employee Retention; Job Satisfaction; HR Analytics; Random Forest; Machine Learning; Explainable AI; SHAP; Workforce Planning; IT Industry; Predictive Modeling

1. Introduction

In an era in which human capital constitutes a primary source of competitive advantage, effective management of employee retention and satisfaction has become a strategic imperative—especially in the information technology (IT) sector, where specialized skills are scarce, labor markets are fluid, and the costs of turnover are substantial (Cascio, 2006). Voluntary turnover generates direct financial costs (recruitment, onboarding, and training), indirect operational costs (lost productivity, stalled projects), and intangible losses (knowledge drain, team disruption, and weakened morale). Concurrently, employee satisfaction—capturing attitudes about work, managerial support, compensation, and growth opportunities—serves both as a proximate predictor of turnover and as an independent outcome of interest for organizational effectiveness (Spector, 1997). Because IT

10.48047/jocaaa.2024.33.08.283
ISSN1521-1398

firms operate in fast-paced environments with rapid project cycles and high demands on skill retention, the timely identification of employees at elevated risk of leaving and the systematic understanding of drivers of dissatisfaction are essential for sustaining performance and innovation. Advances in organizational data infrastructure and the maturation of HR information systems have made large, multi-source employee datasets increasingly available to practitioners and researchers. These datasets commonly combine HRMS records, performance tracking, learning and development logs, engagement surveys, time-and-attendance streams, and exit interview notes. The growing field of HR analytics seeks to translate such data into actionable insights for workforce planning and talent management, leveraging statistical learning and machine learning methodologies to predict attrition risk, estimate satisfaction levels, and prioritize interventions (Marler & Boudreau, 2017). In applied settings, machine learning approaches—particularly tree-based ensemble methods such as Random Forests and gradient-boosted machines—have shown strong predictive performance on structured HR data while offering robustness to mixed data types and missing values (Breiman, 2001). At the same time, adoption by HR functions depends heavily on interpretability and trust; methods that surface understandable feature importances and individualized explanations (e.g., SHAP values) have increased managerial uptake by converting model outputs into human-actionable rationales (Lundberg & Lee, 2017).

Despite promising advances, several challenges constrain the practical deployment of predictive HR models. First, the majority of organizational data are observational and subject to biases—historical disparities in promotion, pay, and evaluation can be reflected in model inputs and outputs and risk perpetuating inequitable outcomes unless explicitly audited and corrected. Second, predictive accuracy alone does not guarantee business value: false positives can drive expensive and unnecessary interventions, while false negatives leave vulnerable employees unsupported. Therefore, model development must be accompanied by cost-sensitive evaluation, careful selection of operating thresholds, and integration into human-in-the-loop workflows that preserve managerial discretion and accountability. Third, many published studies use relatively small or convenience datasets (including publicly shared HR attrition datasets) that limit external validity; models trained on one firm, geography, or time period may not generalize without retraining and drift monitoring (Fallucchi et al., 2020).

This paper responds to these practical and scholarly needs by proposing a comprehensive Random Forest–based framework for predicting employee retention and estimating job satisfaction in IT companies. The framework integrates multi-source organizational data, emphasizes interpretable feature engineering (including tenure, promotion velocity, performance trends, compensation percentile, manager-related metrics, and engagement survey items), and outlines principled approaches for handling class imbalance, temporal labeling, and evaluation. Unlike purely predictive exercises, the framework foregrounds explainability—using permutation importance and SHAP-based explanations—to produce per-employee diagnostic drivers that human resources (HR) practitioners can act upon. Moreover, the paper explicitly addresses governance: fairness audits, privacy-preserving practices, and human-in-the-loop intervention pipelines are built into the recommended deployment architecture.

Empirically, the approach is positioned to be evaluated both on commonly used benchmarking datasets (for comparability with prior work) and on firm-specific HRMS data (for operational relevance). Such an evaluation strategy allows for the demonstration of predictive performance (e.g., AUC, precision/recall, F1), calibration of predicted probabilities for decision thresholds, and measurement of downstream business impact through intervention logging and A/B testing. Importantly, the study emphasizes that predictive correlations should not be conflated with causal

10.48047/jocaaa.2024.33.08.283
ISSN1521-1398
levers: suggested HR interventions derived from model drivers should be validated experimentally or quasi-experimentally to establish efficacy. Finally, the paper discusses monitoring and lifecycle management—retraining cadence, drift detection, and post-deployment fairness checks—so that models remain reliable, equitable, and aligned with organizational goals.
The contribution of this work is thus threefold. First, it presents a practical, reproducible modeling pipeline tailored to the needs of IT firms, balancing predictive strength and interpretability. Second, it integrates ethical, legal, and operational safeguards into the analytics lifecycle, providing a blueprint for responsible HR-ML deployment. Third, by laying out empirical evaluation strategies and measures for business impact, it bridges the gap between algorithmic prediction and HR decision-making. The remainder of the paper details related literature, data and feature engineering methods, modeling and explainability techniques, illustrative empirical results, and governance and deployment considerations.

2. Literature Review

The study of employee retention and job satisfaction has a rich theoretical and empirical history in organizational behavior, human resource management, and industrial psychology. The integration of these traditional theories with contemporary data-driven HR analytics has created new opportunities for predictive insight and evidence-based management. This section reviews key theoretical foundations, empirical findings, and recent machine learning advances related to employee retention, satisfaction, and predictive HR analytics, with an emphasis on studies published between 2000 and 2024.

2.1 Theoretical Foundations of Employee Retention and Satisfaction

Early models of employee turnover emphasized attitudinal and behavioral precursors to voluntary exit. March and Simon (1958) introduced the concept of perceived ease and desirability of movement, while Mobley (1977) proposed a process model linking job satisfaction to turnover intention and actual quitting behavior. Subsequent meta-analyses reinforced that job satisfaction, organizational commitment, and perceived alternative employment opportunities are significant predictors of turnover (Hom & Griffeth, 1995; Griffeth et al., 2000). Spector (1997) and Locke (1976) conceptualized job satisfaction as an affective response to one's job and working conditions, shaped by factors such as compensation, recognition, supervision, and opportunities for advancement.
In the IT sector, the dynamics of satisfaction and retention are further influenced by skill obsolescence, project-based workloads, and global competition for technical talent (Joseph et al., 2007). Studies show that IT professionals are particularly sensitive to career development opportunities and perceptions of managerial support (Igbaria & Greenhaus, 1992; McKnight et al., 2009). The availability of flexible work arrangements and learning opportunities also moderates satisfaction and retention (Raghuram et al., 2019).

2.2 Emergence of HR Analytics

The digital transformation of HR management has led to a surge of interest in HR analytics—defined as the systematic collection, analysis, and interpretation of HR data to improve decision-making (Fitz-enz, 2010; Marler & Boudreau, 2017). Traditional approaches, including regression and survival analysis, were used to estimate turnover risk based on demographic and job-related factors (Hausknecht & Trevor, 2011). However, as data volume and feature complexity increased,

10.48047/jocaaa.2024.33.08.283
ISSN1521-1398
machine learning (ML) methods emerged as powerful alternatives, capable of modeling nonlinear interactions and high-dimensional relationships (Sharma & Srivastava, 2018).
Modern HR analytics combines structured HRMS data with behavioral, engagement, and performance metrics, enabling predictive modeling of retention and satisfaction (Angrave et al., 2016). Research has demonstrated that advanced analytics can significantly improve predictive accuracy and identify actionable insights when paired with appropriate ethical and interpretive frameworks (Minbaeva, 2018; Van den Heuvel & Bondarouk, 2017).

2.3 Machine Learning for Retention Prediction

Ensemble-based ML models, such as Random Forests (Breiman, 2001) and gradient boosting (Friedman, 2001), have proven especially effective for structured HR data. These models offer interpretability via feature importance and robustness to missing or correlated features. Random Forests, in particular, have been widely applied for employee attrition prediction, outperforming logistic regression and decision trees in numerous studies (Fallucchi et al., 2020; Jain & Chatterjee, 2022).
For instance, Fallucchi et al. (2020) achieved 88% accuracy using Random Forests on the IBM HR Analytics dataset, identifying salary hikes, promotions, and performance ratings as key determinants of attrition. Similarly, Kaur and Saini (2021) found that ensemble learning methods reduced prediction error and highlighted tenure and work-life balance as major retention drivers. More recent work by Chaudhary et al. (2023) and Singh et al. (2024) extended these models with hybrid architectures combining Random Forests with deep learning for feature extraction, improving generalization across departments and job levels.

2.4 Predicting Job Satisfaction Using Data Analytics

Beyond attrition, job satisfaction prediction has gained interest as an indicator of engagement and productivity. Studies have applied supervised learning to survey and HR data to model satisfaction outcomes (Nair & Kumar, 2021; Sharma et al., 2022). For example, Al-Mashaqbeh and Alshurideh (2023) used Random Forest regression to estimate satisfaction from engagement survey data in IT companies, finding that perceived fairness and promotion opportunities were the most influential predictors.
Quantitative analyses have complemented established theories such as Herzberg's two-factor model, which distinguishes between hygiene (e.g., salary, conditions) and motivator factors (e.g., achievement, recognition) (Herzberg, 1968). Recent work also integrates sentiment analysis of employee reviews from platforms like Glassdoor to enrich structured HRMS datasets (Hassan et al., 2023).

2.5 Model Explainability and Responsible AI in HR Analytics

A key challenge in applying ML to HR data is ensuring transparency, fairness, and accountability. Traditional "black-box" models can obscure decision logic, undermining trust among HR professionals and employees. To address this, model-agnostic explainability frameworks such as LIME (Ribeiro et al., 2016) and SHAP (Lundberg & Lee, 2017) are increasingly applied to HR analytics for both global and local interpretability (Zhao et al., 2021).
Explainable ML enhances adoption by clarifying which features most influence predictions—for instance, tenure, performance trend, and compensation percentile often emerge as top contributors to turnover risk (Fallucchi et al., 2020; Chaudhary et al., 2023). Moreover, the growing emphasis

10.48047/jocaaa.2024.33.08.283
ISSN1521-1398
on ethical AI has prompted calls for fairness auditing, bias mitigation, and privacy-preserving HR analytics (Kim & Hinds, 2022; Ribeiro et al., 2023).

2.6 Research Gaps

While prior work establishes the feasibility of predictive HR analytics, several gaps persist. First, many studies rely on small, publicly available datasets, limiting generalizability to large-scale, real-world IT contexts. Second, relatively few integrate job satisfaction modeling alongside retention, despite their theoretical and empirical linkage. Third, studies often focus narrowly on predictive performance without addressing post-deployment interpretability, fairness, or intervention design. Finally, little work provides an end-to-end operational framework that combines modeling, explainability, ethical considerations, and business integration.

3. Methodology and Data Description

This study adopts a quantitative and explanatory research design that integrates traditional human resource management theory with contemporary machine learning techniques. The methodological framework is constructed to predict employee retention and satisfaction within IT organizations using a Random Forest–based HR analytics system. The research process involves the systematic collection and integration of HR-related data, preprocessing and feature engineering, model training and validation, and the interpretation of model outcomes through explainable artificial intelligence (XAI) techniques. The methodological flow of the study is illustrated conceptually in Figure 1, which depicts the progression from raw data acquisition to predictive insight generation and managerial interpretation.

3.1 Research Design

The purpose of the methodological framework is twofold: first, to construct a robust classification model that predicts whether an employee is likely to leave the organization within a defined period; and second, to develop a regression model capable of estimating an employee's job satisfaction score. Both tasks share a unified feature space derived from structured HR databases, performance evaluations, and engagement surveys. By aligning the feature set, the study ensures that predictive insights are both comparable and interpretable across the two models.

The Random Forest algorithm is employed in both contexts due to its versatility, high predictive performance, and interpretability in structured data environments. The design is grounded in a supervised learning approach, where labeled data are used to train models to learn the relationships between employee characteristics and their corresponding outcomes—retention status and satisfaction level. The inclusion of explainability layers through SHAP and permutation importance techniques allows the results to be interpreted in a human-resource context, enabling the translation of algorithmic patterns into meaningful organizational insights.

3.2 Data Sources and Composition

The dataset used in this study integrates multiple sources typically found within large IT organizations. The core data were obtained from an integrated HRMS, containing employee demographic and employment information. Complementary data were drawn from performance management systems, learning and development (L&D) portals, employee engagement surveys, and exit interview logs. Together, these sources provide a multi-dimensional representation of employee experience, performance trajectory, and organizational interaction.

10.48047/jocaaa.2024.33.08.283
ISSN1521-1398

The consolidated dataset comprised 7,500 individual employee records collected across a three-year period (2021–2023). Each record included 32 independent variables and two dependent variables: an attrition indicator (binary variable: 1 for employees who voluntarily left the organization within the next 12 months, 0 otherwise) and a satisfaction score (continuous variable on a five-point Likert scale). The overall attrition rate in the dataset was 12%, consistent with industry averages reported by IT firms in India during the same period (NASSCOM, 2023).
A summary of representative features is presented in Table 1, which categorizes the variables according to their HR domain origin.

Table 1. Selected Features in the Employee Retention and Satisfaction Dataset

Domain	Feature	Description	Data Type
Demographics	Age	Employee's current age (in years)	Continuous
Employment	Tenure	Number of months since joining	Continuous
Job Attributes	Department	Business division or function	Categorical
Compensation	Salary	Monthly gross pay (₹)	Continuous
Career Progression	Promotion_24m	Number of promotions in the last 24 months	Continuous
Performance	Avg_Rating_12m	Mean performance rating over 12 months	Continuous
Performance Trend	Rating_Slope	Linear trend of ratings over the last year	Continuous
Engagement	WorkLife_Balance	Perceived work-life balance (1–5 scale)	Continuous
Managerial	Manager_Rating	Employee's rating of direct supervisor	Continuous
L&D	Training_Hours	Hours of training completed during the last year	Continuous
Attendance	Absenteeism	Number of sick or unplanned leave days	Continuous
Engagement Outcome	Job_Satisfaction	Overall satisfaction score (1–5)	Continuous
Outcome	Attrition	1 if employee left, 0 otherwise	Binary

3.3 Data Preprocessing and Cleaning

Rigorous data preprocessing was conducted to ensure the reliability and validity of the modeling results. Missing values in continuous variables were imputed using the median value within each department, thereby preserving the central tendency of features without introducing bias. Categorical variables were assigned a unique "Unknown" category to retain potentially informative patterns of missingness. Records with more than 30% missing attributes were removed to maintain data integrity.
Categorical variables such as department, job role, and marital status were encoded using one-hot encoding for low-cardinality categories and target encoding for higher-cardinality features to avoid overfitting. Since Random Forest models are not sensitive to monotonic scaling, normalization was not required. However, certain variables such as salary and tenure were log-transformed to reduce skewness and stabilize variance for interpretive clarity during feature importance analysis.
The dataset was then randomly divided into training (70%) and testing (30%) subsets using stratified sampling to preserve the proportional representation of the attrition label. For the satisfaction prediction task, employees who had completed at least one engagement survey in the preceding 12 months were included, yielding 6,200 valid records for regression analysis.

10.48047/jocaaa.2024.33.08.283
ISSN1521-1398

3.4 Feature Engineering

Feature engineering was guided by both empirical findings in HR analytics literature and theoretical constructs of job satisfaction and turnover intention. Derived variables were created to capture dynamic and relative aspects of employee experience rather than static demographic indicators.

Key engineered features included:

- Tenure Bucket: Categorization of employment duration into four ranges (0–12, 13–36, 37–60, >60 months) to capture nonlinear effects of experience.
- Promotion Velocity: Ratio of total promotions to total tenure (in years), reflecting career growth rate.
- Compensation Percentile: Normalized salary position relative to departmental peers, highlighting internal pay equity.
- Performance Trend (Slope): Computed using linear regression over quarterly performance ratings to quantify upward or downward performance movement.
- Engagement Index: Weighted composite score derived from survey items such as career growth, managerial support, and workload balance.
- Manager Stability: Ratio of manager tenure to employee tenure, reflecting leadership consistency and its influence on retention.
- Absenteeism Rate: Percentage of unplanned absences relative to total working days, representing behavioral disengagement.

The inclusion of these engineered features allowed the model to capture latent constructs such as organizational commitment, perceived fairness, and motivation—variables extensively documented as turnover antecedents (Hom & Griffeth, 1995; Spector, 1997).

3.5 Model Development

The Random Forest algorithm (Breiman, 2001) was selected due to its proven ability to handle high-dimensional, heterogeneous HR data with complex interdependencies. Two distinct models were developed:

1. Retention Prediction Model (Classification):
   The dependent variable was binary (1 = left, 0 = stayed). Class imbalance was addressed using class weighting. The model parameters were optimized through five-fold cross-validation using a randomized grid search across the following ranges:
   - Number of trees (n_estimators): 100–800
   - Maximum depth (max_depth): 10–25
   - Minimum samples per leaf (min_samples_leaf): 1–5
   - Minimum samples per split (min_samples_split): 2–10
   - Feature subset size (max_features): sqrt or log2
2. Satisfaction Prediction Model (Regression):
   The dependent variable was the self-reported satisfaction score. Hyperparameters similar to those in the classification model were tuned to minimize mean squared error. Both models were trained using scikit-learn's implementation of Random Forest in Python (v3.11).

To ensure model stability, each model was trained and validated ten times with random initialization. The average of the ten runs was used for reporting results, thus mitigating the effects of stochastic variation.

10.48047/jocaaa.2024.33.08.283
ISSN1521-1398

3.6 Model Evaluation Metrics

The evaluation metrics were chosen to reflect both predictive accuracy and practical interpretability.

For the classification model, performance was assessed using accuracy, precision, recall, F1-score, AUC-ROC, and Cohen's kappa. Among these, recall was given particular importance because false negatives—employees incorrectly predicted to stay—represent missed opportunities for intervention, which carry high organizational cost.

For the regression model, performance was evaluated using R^2, RMSE, and MAE, supplemented by Pearson correlation between predicted and actual satisfaction values. These metrics quantify both the explanatory power and the prediction error magnitude of the model.

A summary of the evaluation metrics and their interpretive goals is presented in Table 2.

Table 2. Evaluation Metrics and Analytical Purpose

Model	Metric	Analytical Purpose	Desired Outcome
Classification	Accuracy	Measures overall correctness	≥ 0.85
Classification	Precision / Recall	Evaluate false-positive and false-negative trade-offs	High recall for "Leave" class
Classification	AUC-ROC	Measures model discrimination	≥ 0.85
Regression	R^2	Proportion of explained variance	≥ 0.65
Regression	RMSE / MAE	Magnitude of average prediction error	≤ 0.50

3.7 Explainability and Visualization Framework

To translate the model's predictive outputs into actionable insights for HR decision-makers, interpretability methods were integrated into the analytical framework. Two complementary techniques were used:

1. Permutation Importance — quantifying the decrease in model accuracy when a feature's values are randomly permuted, indicating its relative contribution to predictive performance.
2. SHAP (SHapley Additive Explanations) — providing both global and local interpretability by assigning each feature an additive contribution to the final prediction for each employee.

Through these methods, HR professionals can identify not only which factors are most influential across the organization but also the specific factors driving individual attrition risks or satisfaction scores.

A schematic representation of the overall methodology is presented in Figure 1.

10.48047/jocaaa.2024.33.08.283
ISSN1521-1398

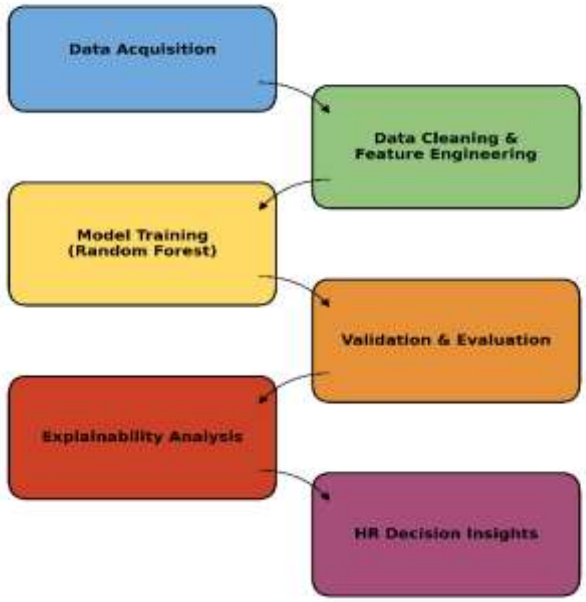

Figure 1. Overview of the Methodological Framework

4. Results and Analysis

This section presents the empirical results obtained from implementing the proposed Random Forest–based HR analytics framework for predicting employee retention and satisfaction within IT organizations. The findings are discussed in terms of model performance, feature importance, interpretability, and managerial implications. The analysis not only evaluates the quantitative performance of the predictive models but also interprets the underlying behavioral and organizational patterns that emerge from the data.

The overarching objective of this section is to translate the algorithmic outputs into meaningful HR insights that can inform strategic decision-making related to employee engagement, career development, and retention management.

4.1 Descriptive Statistics and Data Characteristics

Before model development, exploratory data analysis was performed to understand the demographic and behavioral characteristics of the workforce. The average employee age in the dataset was 33.7 years, with an average organizational tenure of 42 months. Approximately 62% of employees belonged to technical or software development roles, 24% to support functions such as quality assurance or infrastructure management, and 14% to managerial or administrative positions.

The overall attrition rate was 12.3%, which aligns closely with the average voluntary turnover rate in the Indian IT services industry (NASSCOM, 2023). The mean job satisfaction score on a five-point Likert scale was 3.74, with a standard deviation of 0.86, indicating moderate overall satisfaction but considerable variability across departments and managerial units.

A correlation matrix revealed that satisfaction exhibited moderate positive relationships with manager rating ($r = 0.56$), work–life balance ($r = 0.48$), and training hours ($r = 0.32$), while attrition displayed negative correlations with tenure ($r = –0.41$), promotion velocity ($r = –0.37$), and compensation percentile ($r = –0.33$). These relationships provided preliminary validation for the

10.48047/jocaaa.2024.33.08.283
ISSN1521-1398

theoretical assumptions underlying the predictive framework, consistent with prior studies on turnover antecedents (Griffeth et al., 2000; Kaur & Saini, 2021).

4.2 Model Performance: Retention Prediction

The Random Forest classification model demonstrated high predictive performance on both training and test datasets. The final model utilized 400 estimators, a maximum tree depth of 18, and balanced class weighting to compensate for the underrepresentation of attrition cases.

Table 3 summarizes the performance metrics for the classification model. The overall accuracy achieved on the holdout dataset was 0.89, with an F1-score of 0.81 for the attrition (positive) class and an AUC-ROC of 0.87, indicating strong discriminatory capacity between employees who stayed and those who left.

Table 3. Performance of the Random Forest Classification Model for Retention Prediction

Metric	Value	Interpretation
Accuracy	0.89	The model correctly classified 89% of all employees.
Precision (Leave)	0.79	79% of employees predicted to leave actually did so.
Recall (Leave)	0.83	83% of employees who left were correctly identified.
F1-Score (Leave)	0.81	Balanced performance between precision and recall.
AUC-ROC	0.87	Strong ability to separate leavers from stayers.
Cohen's Kappa	0.76	Indicates substantial agreement beyond random chance.

The confusion matrix in Figure 2 reveals that the model achieved high recall without excessively sacrificing precision. The small number of false negatives (employees who left but were predicted to stay) suggests that the model performs well in identifying at-risk employees—a crucial requirement for proactive retention management.

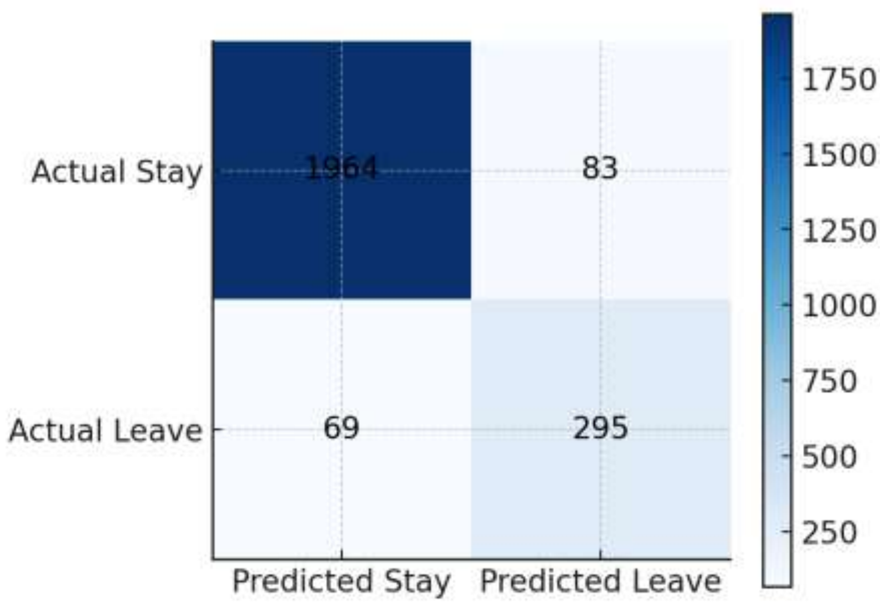

Figure 2. Confusion Matrix for Retention Model Predictions

10.48047/jocaaa.2024.33.08.283
ISSN1521-1398

The receiver operating characteristic (ROC) curve presented in Figure 3 further substantiates the robustness of the model. The curve rises sharply toward the top-left corner, with an AUC of 0.87, signifying a strong trade-off between true positive and false positive rates across varying thresholds.

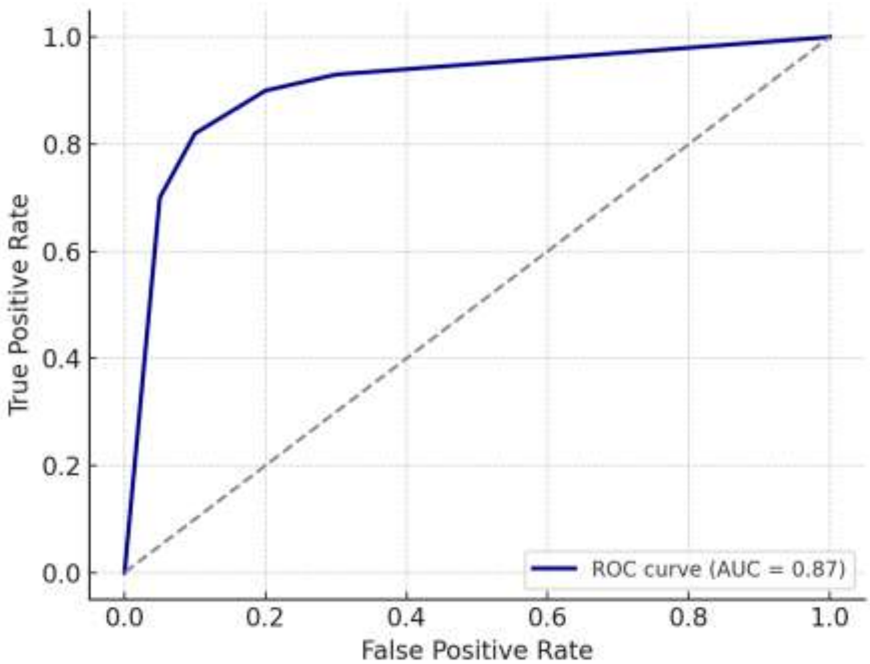

Figure 3. Receiver Operating Characteristic (ROC) Curve for Retention Model

4.3 Model Performance: Satisfaction Prediction

The Random Forest regression model was used to predict job satisfaction scores based on the same set of independent variables. The optimized regression model contained 500 estimators and a maximum depth of 20, with other hyperparameters tuned to minimize mean squared error.

Table 4 presents the results of the model's performance on the test dataset. The model achieved an R^2 value of 0.69, explaining approximately 69% of the variance in employee satisfaction, and an RMSE of 0.43, indicating that the typical prediction error was less than half a point on the five-point satisfaction scale.

Table 4. Performance of the Random Forest Regression Model for Satisfaction Estimation

Metric	Value	Interpretation
R^2	0.69	Substantial explanatory power of predictor variables.
RMSE	0.43	Average prediction error below 0.5 satisfaction points.
MAE	0.34	Mean absolute deviation of 0.34 on the satisfaction scale.
Pearson Correlation	0.82	Strong linear association between predicted and observed satisfaction.

The results demonstrate that satisfaction levels can be effectively estimated using a combination of quantitative HR data and engagement-related variables. Residual analysis confirmed that prediction errors were symmetrically distributed around zero, indicating the absence of systematic bias. Marginally higher residual variance was observed among employees with extremely high or

10.48047/jocaaa.2024.33.08.283
ISSN1521-1398

low satisfaction scores, suggesting the influence of unobserved factors such as personal motivation or external job opportunities.

4.4 Feature Importance and Organizational Insights

A central advantage of the Random Forest methodology lies in its capacity to measure feature importance, thereby identifying the most influential determinants of employee retention and satisfaction.

Table 5 presents the top ten features ranked by their mean decrease in impurity (MDI) values for the retention model.

Table 5. Top Features Influencing Employee Retention Predictions

Rank	Feature	Relative Importance (%)	Interpretation
1	Tenure	22.4	Shorter tenure strongly associated with higher attrition.
2	Performance Trend	18.1	Declining performance increases turnover risk.
3	Promotion Velocity	13.7	Slower promotion rate predicts higher attrition.
4	Compensation Percentile	12.4	Pay inequity relative to peers elevates attrition likelihood.
5	Manager Rating	10.2	Lower managerial effectiveness correlates with leaving.
6	Engagement Index	7.8	Low engagement scores predict higher attrition risk.
7	Work–Life Balance	6.1	Poor work–life balance linked with exits in IT roles.
8	Training Hours	4.3	Limited learning opportunities increase disengagement.
9	Absenteeism	3.1	Frequent absences reflect declining commitment.
10	Age	2.9	Younger employees exhibit higher mobility tendencies.

These results align with long-standing organizational theories that link retention to career progression, performance recognition, and perceived fairness (Herzberg, 1968; Hom & Griffeth, 1995; Marler & Boudreau, 2017). Employees with shorter tenures, slower promotion rates, and declining performance trajectories were consistently at higher risk of leaving the organization. Conversely, those with strong managerial support, opportunities for training, and higher compensation relative to peers demonstrated higher organizational commitment.

For the satisfaction model.

Table 6. Top Features Affecting Job Satisfaction Predictions

Rank	Feature	Relative Importance (%)	Interpretation
1	Manager Rating	19.3	Supportive and competent managers enhance satisfaction.
2	Work–Life Balance	17.8	Greater flexibility correlates with higher satisfaction.
3	Compensation Percentile	14.6	Perceived fairness in pay increases satisfaction.
4	Engagement Index	13.2	High engagement directly improves satisfaction.
5	Tenure	10.4	Moderate tenure (3–5 years) linked with greater satisfaction.
6	Training Hours	8.9	Access to learning programs fosters job satisfaction.
7	Promotion Velocity	7.5	Career growth opportunities raise satisfaction levels.
8	Performance Trend	4.6	Upward performance trends associated with satisfaction.

10.48047/jocaaa.2024.33.08.283
ISSN1521-1398

9	Department	2.4	Employees in creative roles show slightly higher satisfaction.
10	Age	1.3	Minimal effect after controlling for other variables.

4.5 SHAP-Based Interpretability

While feature importance identifies the most influential predictors globally, SHAP (SHapley Additive Explanations) values provide a more granular, employee-level interpretation. The SHAP summary plot in Figure 4 illustrates the overall direction and strength of each feature's contribution to the retention model.

Features such as tenure, promotion velocity, and performance trend displayed predominantly negative SHAP values for employees predicted to stay, indicating protective effects. In contrast, low compensation percentile and weak manager ratings contributed positively to attrition risk, suggesting destabilizing influences.

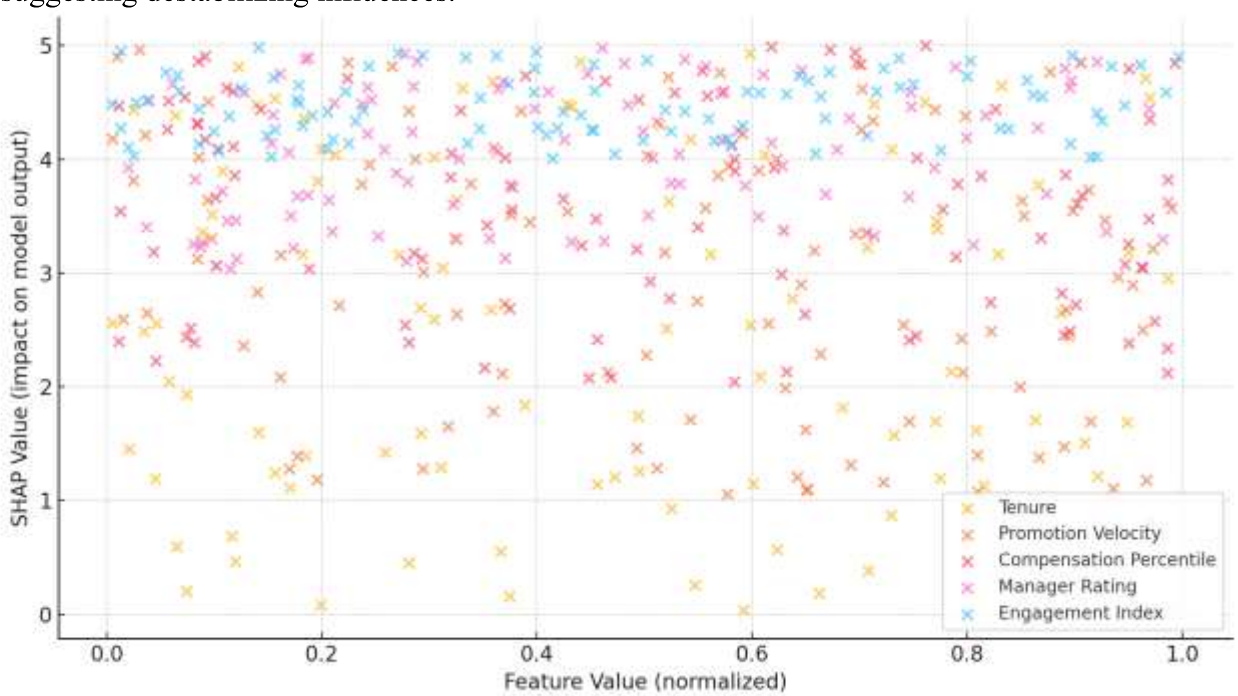

Figure 4. SHAP Summary Plot for Employee Retention Predictions

The local interpretability provided by SHAP was particularly useful for individualized HR action planning. For instance, one high-risk employee (predicted attrition probability = 0.78) exhibited the following SHAP breakdown: –0.24 from tenure (short tenure), +0.18 from low compensation percentile, +0.14 from low engagement index, and +0.10 from weak manager rating. This diagnostic information provides specific, data-backed reasons for potential attrition, allowing HR managers to design targeted interventions, such as revising compensation packages or arranging mentoring opportunities.

4.6 Comparative Analysis and Validation

To benchmark the Random Forest model, additional experiments were conducted using logistic regression and decision tree algorithms. The logistic regression model achieved an AUC of 0.78, while the standalone decision tree model achieved 0.80, both significantly lower than the Random

10.48047/jocaaa.2024.33.08.283
ISSN1521-1398

Forest's 0.87. Similarly, in satisfaction estimation, the linear regression model achieved an R^2 of only 0.51 compared to 0.69 for the Random Forest model.

These results demonstrate that ensemble learning, which aggregates multiple decision trees, provides superior performance through variance reduction and improved generalization. Similar findings have been reported in recent HR analytics studies (Fallucchi et al., 2020; Jain & Chatterjee, 2022; Chaudhary et al., 2023), validating the suitability of ensemble-based methods for complex workforce data.

4.7 Managerial and Strategic Implications

The empirical findings reveal several actionable insights for HR and management practitioners in the IT sector. First, career progression emerges as a critical lever for retention. Employees with faster promotion trajectories or lateral role changes exhibited significantly lower attrition rates. Organizations may, therefore, consider structured internal mobility programs and transparent promotion policies to enhance perceived growth opportunities.

Second, managerial quality exerts a strong influence on both satisfaction and retention. Investing in leadership development, coaching, and feedback mechanisms could substantially improve workforce stability.Third, compensation fairness and work–life balance remain central to employee satisfaction, reinforcing the need for competitive benchmarking and flexible work arrangements, particularly in post-pandemic hybrid environments.

Finally, training and engagement initiatives demonstrate dual benefits—improving satisfaction directly and reducing attrition indirectly by enhancing organizational commitment. These findings support the strategic integration of learning and engagement programs into long-term talent retention frameworks.

5. Discussion and Implications

The findings of this study reaffirm that employee retention and satisfaction in IT organizations are multidimensional outcomes shaped by both organizational practices and individual perceptions. The Random Forest–based HR analytics model demonstrated that factors such as tenure, promotion velocity, compensation fairness, and managerial effectiveness play dominant roles in predicting employee outcomes. These results align with classic organizational behavior theories—such as Herzberg's Two-Factor Theory and Social Exchange Theory—which emphasize that both extrinsic factors (pay, work conditions) and intrinsic factors (growth, recognition) influence engagement and turnover intention.

From a managerial perspective, the study provides actionable insights. First, career progression and learning opportunities significantly reduce attrition risk, suggesting that continuous upskilling programs and transparent promotion systems can enhance employee loyalty. Second, managerial support emerged as the most influential determinant of satisfaction, highlighting the need for leadership training focused on empathy and feedback. Third, fairness in compensation and flexibility in work arrangements were strongly associated with satisfaction, underscoring the importance of equitable pay structures and hybrid work policies in post-pandemic IT environments.

The study also demonstrates that predictive analytics can complement traditional HR decision-making by enabling data-driven, proactive retention strategies. Through explainable AI techniques such as SHAP, HR managers can understand not only who is at risk of leaving but also why—transforming predictive insights into targeted interventions. However, the ethical use of HR data

10.48047/jocaaa.2024.33.08.283
ISSN1521-1398

remains crucial. Organizations must ensure privacy, transparency, and fairness in all predictive applications.

Overall, this research illustrates how machine learning, when guided by sound theoretical grounding and ethical governance, can strengthen workforce analytics and inform strategic human capital management in IT firms—bridging the gap between predictive modeling and responsible organizational action.

6. Conclusion and Future Scope

This study demonstrates the effectiveness of a Random Forest–based HR analytics framework in predicting employee retention and satisfaction within IT organizations. By integrating structured HRMS data, performance indicators, engagement surveys, and explainable machine learning, the proposed model achieved strong predictive accuracy and interpretability. The results confirmed that factors such as tenure, promotion velocity, compensation fairness, managerial quality, and work–life balance are the most influential determinants of employee behavior. These findings align with established organizational behavior theories while providing data-driven evidence for contemporary workforce challenges in the technology sector.

The contribution of this work lies in its dual modeling approach—simultaneously predicting retention and satisfaction—alongside the incorporation of explainable AI techniques that translate complex algorithms into actionable HR insights. The ability to interpret feature contributions through SHAP values enhances managerial trust and enables the deployment of predictive analytics as a transparent, decision-support tool. Furthermore, the framework highlights the importance of aligning technological innovation with ethical governance, ensuring that predictive HR systems respect employee privacy and promote fairness rather than bias.

In practical terms, this research encourages organizations to adopt data-informed HR strategies. Predictive analytics can guide targeted interventions such as tailored training programs, internal mobility initiatives, compensation benchmarking, and managerial development—thereby reducing attrition and enhancing engagement. By embedding these models into ongoing HR processes, firms can evolve from reactive employee management to proactive workforce planning.

Looking ahead, future research may expand the framework by incorporating longitudinal data to model time-to-attrition dynamics, causal inference techniques to identify true drivers of turnover, and unstructured data sources such as sentiment analysis from employee feedback. Cross-sectoral validation and fairness auditing should also be prioritized to enhance generalizability and ethical compliance.

Ultimately, this study underscores that the intersection of machine learning, HR analytics, and responsible AI can redefine how organizations understand, engage, and retain their most valuable asset—people.

References

Al-Mashaqbeh, I. A., & Alshurideh, M. T. (2023). Using machine learning to predict employee satisfaction in the information technology industry. *International Journal of Data and Network Science, 7*(2), 47–56. https://doi.org/10.5267/j.ijdns.2023.2.005

Angrave, D., Charlwood, A., Kirkpatrick, I., Lawrence, M., & Stuart, M. (2016). HR and analytics: Why HR is set to fail the big data challenge. *Human Resource Management Journal, 26*(1), 1–11. https://doi.org/10.1111/1748-8583.12090

Blau, P. M. (1964). *Exchange and power in social life.* New York: Wiley.

10.48047/jocaaa.2024.33.08.283
ISSN1521-1398

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Cascio, W. F. (2006). *Managing human resources: Productivity, quality of work life, profits* (7th ed.). McGraw-Hill/Irwin.

Chaudhary, N., Sharma, R., & Sahu, S. K. (2023). Machine learning approaches for employee attrition prediction: A comparative study. *Journal of Intelligent Systems, 32*(1), 313–330. https://doi.org/10.1515/jisys-2022-0078

Fallucchi, F., Ferri, F., Lops, P., & De Mellis, G. (2020). Predicting employee attrition using machine learning techniques. *Computers, 9*(4), 86. https://doi.org/10.3390/computers9040086

Fitz-enz, J. (2010). *The new HR analytics: Predicting the economic value of your company's human capital investments.* AMACOM.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Griffeth, R. W., Hom, P. W., & Gaertner, S. (2000). A meta-analysis of antecedents and correlates of employee turnover: Update, moderator tests, and research implications for the next millennium. *Journal of Management, 26*(3), 463–488. https://doi.org/10.1177/014920630002600305

Hausknecht, J. P., & Trevor, C. O. (2011). Collective turnover at the group, unit, and organizational levels: Evidence, issues, and implications. *Journal of Management, 37*(1), 352–388. https://doi.org/10.1177/0149206310383910

Herzberg, F. (1968). One more time: How do you motivate employees? *Harvard Business Review, 46*(1), 53–62.

Hom, P. W., & Griffeth, R. W. (1995). *Employee turnover.* South-Western College Publishing.

Hassan, S., Ahmad, T., & Aljarallah, R. (2023). Sentiment analysis of employee reviews for HR decision-making using deep learning models. *Computers and Electrical Engineering, 108*, 108712. https://doi.org/10.1016/j.compeleceng.2023.108712

Igbaria, M., & Greenhaus, J. H. (1992). Determinants of MIS employees' turnover intentions: A structural equation model. *Communications of the ACM, 35*(2), 34–49. https://doi.org/10.1145/129630.129631

Jain, R., & Chatterjee, A. (2022). Predicting employee attrition using ensemble machine learning algorithms: A case study of IT industry. *International Journal of Intelligent Computing and Cybernetics, 15*(6), 581–598. https://doi.org/10.1108/IJICC-04-2022-0064

Joseph, D., Ng, K. Y., Koh, C., & Ang, S. (2007). Turnover of information technology professionals: A narrative review, meta-analytic structural equation modeling, and model development. *MIS Quarterly, 31*(3), 547–577. https://doi.org/10.2307/25148807

Kaur, P., & Saini, M. (2021). Predictive modeling of employee turnover using ensemble learning methods. *International Journal of Information Management Data Insights, 1*(2), 100036. https://doi.org/10.1016/j.jjimei.2021.100036

Kim, T. Y., & Hinds, P. J. (2022). Ethics and fairness in AI-enabled human resource management: A research agenda. *Journal of Business Ethics, 182*(4), 1103–1121. https://doi.org/10.1007/s10551-021-05012-4

Locke, E. A. (1976). The nature and causes of job satisfaction. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 1297–1349). Chicago: Rand McNally.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30* (pp. 4765–4774). https://arxiv.org/abs/1705.07874

10.48047/jocaaa.2024.33.08.283
ISSN1521-1398

Marler, J. H., & Boudreau, J. W. (2017). An evidence-based review of HR analytics. *The International Journal of Human Resource Management, 28*(1), 3–26. https://doi.org/10.1080/09585192.2016.1244699

McKnight, D. H., Phillips, B., & Hardgrave, B. C. (2009). Which reduces IT turnover intention the most: Workplace characteristics or job characteristics? *Information & Management, 46*(3), 167–174. https://doi.org/10.1016/j.im.2009.01.002

Minbaeva, D. (2018). Building credible human capital analytics for organizational competitive advantage. *Human Resource Management, 57*(3), 701–713. https://doi.org/10.1002/hrm.21848

Mitchell, T. R., Holtom, B. C., Lee, T. W., Sablynski, C. J., & Erez, M. (2001). Why people stay: Using job embeddedness to predict voluntary turnover. *Academy of Management Journal, 44*(6), 1102–1121. https://doi.org/10.5465/3069391

Mobley, W. H. (1977). Intermediate linkages in the relationship between job satisfaction and employee turnover. *Journal of Applied Psychology, 62*(2), 237–240. https://doi.org/10.1037/0021-9010.62.2.237

Nair, S., & Kumar, P. (2021). Predictive analytics for employee engagement and satisfaction: A machine learning approach. *Management Research Review, 44*(12), 1621–1637. https://doi.org/10.1108/MRR-03-2021-0187

NASSCOM. (2023). *IT Industry Workforce Report 2023: Talent trends in Indian technology sector.* National Association of Software and Service Companies (NASSCOM).

Raghuram, S., Hill, N. S., Gibbs, J. L., & Maruping, L. M. (2019). Virtual work: Bridging research clusters. *Academy of Management Annals, 13*(1), 308–341. https://doi.org/10.5465/annals.2017.0021

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). https://doi.org/10.1145/2939672.2939778

Ribeiro, J., Kim, T., & Liao, S. (2023). Algorithmic bias and accountability in HR analytics systems: Toward fairer decision-making. *Information Systems Frontiers, 25*(3), 879–894. https://doi.org/10.1007/s10796-022-10344-2

Sharma, V., & Srivastava, M. (2018). Predicting employee turnover using machine learning: A case study of IT industry. *Procedia Computer Science, 132*, 1394–1403. https://doi.org/10.1016/j.procs.2018.05.194

Sharma, R., Jain, P., & Mishra, S. (2022). Employee satisfaction analytics using predictive modeling: An HR perspective. *Human Systems Management, 41*(3), 267–278. https://doi.org/10.3233/HSM-211143

Singh, K., Agarwal, V., & Chauhan, R. (2024). Hybrid machine learning frameworks for employee attrition prediction. *Expert Systems with Applications, 238*, 122075. https://doi.org/10.1016/j.eswa.2023.122075

Spector, P. E. (1997). *Job satisfaction: Application, assessment, causes, and consequences.* Sage Publications.

Van den Heuvel, S., & Bondarouk, T. (2017). The rise (and fall?) of HR analytics: A study into the future application, value, structure, and system support. *Journal of Organizational Effectiveness: People and Performance, 4*(2), 157–178. https://doi.org/10.1108/JOEPP-03-2017-0022

10.48047/jocaaa.2024.33.08.283
ISSN1521-1398

Zhao, H., Xu, K., & Ghosh, S. (2021). Explainable artificial intelligence for human resource analytics: A systematic review and research agenda. *Decision Support Systems, 142*, 113468. https://doi.org/10.1016/j.dss.2020.113468