IJAFIBS, 14 (1) (2026) pp. 173-183

Published by: TRIGIN Publisher

International Journal of Applied Finance and Business Studies

Journal homepage: www.ijafibs.pelnus.ac.id

Analyzing employee turnover risk to strengthen data-driven human resource retention strategies

Feti Arman

Department of Industrial Engineering, Sekolah Tinggi Teknologi Nusantara Lampung, Indonesia

ARTICLE INFO

Article history:

Received Jun 1, 2026
Revised Jun 7, 2026
Accepted Jun 22, 2026

Keywords:

Data-Driven HR;
Employee Attrition;
Employee Retention;
HR Analytics;
Turnover Risk.

ABSTRACT

Employee turnover remains a strategic challenge in human resource management because it affects workforce stability, recruitment costs, productivity, and long-term organizational sustainability. This study aims to analyze employee turnover risk and translate predictive analytics results into data-driven human resource retention strategies. A quantitative predictive analytics approach was applied using the IBM HR Analytics Employee Attrition and Performance dataset, consisting of 1,470 employee records. Data were processed using RapidMiner through attribute selection, nominal-to-binominal transformation, role assignment, nominal-to-numerical conversion, 10-fold cross-validation, and model evaluation. Four models were examined: Naive Bayes, Decision Tree, Random Forest, and Gradient Boosted Trees. The evaluation used accuracy, precision, recall, F1-score, AUC, and confusion matrix, with emphasis on the Attrition = Yes class as the main indicator of turnover risk. The results show that Gradient Boosted Trees provided the most balanced performance, achieving 84.49% accuracy, 52.04% precision, 48.52% recall, 50.22% F1-score, and 0.796 AUC. These findings indicate that predictive analytics can support early identification of employees requiring targeted retention interventions. This study contributes to HR analytics literature by positioning predictive modeling as a managerial decision-support tool for retention planning. Future research should use actual organizational data, longitudinal designs, and explainable AI methods to improve model interpretability and application.

Corresponding Author:

Feti Arman,
Department of Industri Engineering,
Sekolah Tinggi Teknologi Nusantara Lampung,
Pulau Damar St., Sapta Marga Alley, Way Dadi Baru, Sukarame, Bandar Lampung, 35131, Indonesia
Email: fetiarman@sttnlampung.ac.id

1. INTRODUCTION

Employee turnover remains a critical global issue in human resource management because organizations are facing intensified talent mobility, post-pandemic changes in work expectations, digital transformation, and increasing competition for skilled employees (Al Akasheh, Malik, et al., 2024; Hom et al., 2017; Kiran et al., 2024). In this context, turnover is no longer viewed merely as an administrative loss of employees, but as a strategic problem that affects organizational stability, knowledge retention, service continuity, recruitment costs, and long-term workforce planning (Marín Díaz et al., 2023; McCartney & Fu, 2022). The global shift toward data-driven management has also changed how organizations understand turnover, because human resource decisions increasingly require predictive evidence rather than intuition-based judgment (Hamilton & Sodeman, 2020; Minbaeva, 2018; Van den Heuvel & Bondarouk, 2017). Therefore, employee

turnover risk analysis has become an important managerial agenda for strengthening retention strategies and maintaining sustainable organizational performance (Muhammad & Naz, 2022).

Previous studies in the last decade have shown that employee turnover is influenced by multidimensional human resource factors, including job satisfaction, compensation, work environment, workload, career development, employee engagement, and work-life balance (Hassan, 2022; Kiran et al., 2024; Rubenstein et al., 2018). Al Akasheh et al. (2024) reviewed a decade of research on machine learning techniques for employee turnover prediction and found that predictive models have increasingly been used to identify employees who are likely to leave an organization. Marín Díaz et al. (2023) demonstrated that explainable artificial intelligence can support strategic HR decision-making by helping organizations understand the factors behind employee attrition predictions. Kiran et al. (2024) also emphasized that HR analytics and attrition management are closely connected to organizational performance because data-driven HR interventions can improve retention and workforce effectiveness.

Although previous research has advanced the use of predictive analytics in turnover studies, an important gap remains in translating model results into practical human resource retention strategies (McCartney & Fu, 2022). Many studies still emphasize algorithmic comparison and predictive accuracy, while the managerial meaning of recall, precision, false positives, and false negatives for HR intervention is often insufficiently discussed (Levenson & Fink, 2017). This gap is urgent because turnover data are commonly imbalanced, where employees who leave represent a smaller but strategically important group that may be difficult to detect using accuracy-oriented evaluation alone (Al Akasheh, Malik, et al., 2024). Consequently, employee turnover risk analysis should be positioned not merely as a technical classification task, but as a decision-support mechanism for identifying at-risk employees and designing more targeted retention policies.

The practical gap addressed in this study lies in the limited connection between turnover prediction outputs and the actual implementation of retention strategies in organizations. In many HR analytics studies, prediction results are mainly reported through algorithmic performance indicators such as accuracy, AUC, precision, and recall, but these indicators are not always translated into concrete managerial actions. As a result, organizations may know which model performs better statistically, yet still lack clear guidance on how prediction results should be used to determine retention priorities, design intervention programs, or allocate HR resources efficiently. This gap becomes more critical in turnover prediction because the employees who actually leave often represent a minority class, while their departure may create substantial organizational consequences, including knowledge loss, replacement costs, productivity disruption, and reduced team stability. Therefore, predictive analytics should not stop at identifying turnover probability, but should be interpreted as a decision-support mechanism that helps HR managers determine which employees require further assessment, what type of retention intervention is needed, and how organizational policies can be adjusted based on data-driven evidence.

The novelty of this study lies in its managerial orientation, where machine learning is used as an HR analytics tool to strengthen data-driven human resource retention strategies rather than as the final objective of the research (Margherita, 2022). This study does not only compare predictive models, but also interprets the performance results from the perspective of turnover risk identification, especially by focusing on the Attrition = Yes class as the main indicator of employees who require retention attention (Minbaeva, 2018). By emphasizing accuracy, precision, recall, F1-score, AUC, and confusion matrix interpretation, this study highlights the difference between models that appear statistically strong and models that are managerially useful for early turnover detection (McCartney & Fu, 2022). This approach contributes to the integration of HR analytics, employee turnover risk assessment, and data-driven HR decision-making in the context of human resource management (Fukui et al., 2023).

Based on this background, this study aims to analyze employee turnover risk and translate predictive analytics results into data-driven human resource retention strategies (Marín Díaz et al., 2023). The first research problem is how employee turnover risk can be identified using HR-related variables such as demographic factors, job-related factors, compensation, satisfaction, engagement, and career factors (Margherita, 2022). The second research problem is which

predictive model provides the most relevant performance for supporting HR decision-making when evaluation is interpreted through accuracy, precision, recall, F1-score, AUC, and confusion matrix results (Căvescu & Popescu, 2025). The third research problem is how the results of turnover risk modeling can be used to strengthen retention strategies through workload evaluation, compensation review, career development, employee engagement, and work-life balance management (Al Akasheh, Hujran, et al., 2024).

2. RESEARCH METHOD

Research Design

This study employed a quantitative research design with a predictive HR analytics approach to analyze employee turnover risk and support data-driven human resource retention strategies (McAbee et al., 2017). This design was selected because the study aimed to examine measurable employee-related variables, identify turnover risk patterns, and evaluate the usefulness of predictive modeling for managerial decision-making in human resource management. HR analytics has increasingly been used to support evidence-based HR decisions because it enables organizations to transform employee data into actionable insights for recruitment, retention, performance management, and workforce planning. Recent studies also emphasize that predictive analytics in HRM is valuable when the results are not only interpreted as statistical outputs but also translated into practical managerial interventions for employee retention.

The purpose of this study was to analyze employee turnover risk and determine how predictive model results can strengthen data-driven human resource retention strategies. The research questions addressed in this study were: how can employee turnover risk be identified using HR-related variables, which model provides the most relevant performance for HR decision-making, and how can turnover risk modeling results be interpreted to support employee retention strategies.

Data Source and Human Resource Variables

This study used secondary data obtained from the IBM HR Analytics Employee Attrition and Performance Dataset, which is publicly available through Kaggle. The dataset was designed to explore factors associated with employee attrition, including job role, distance from home, monthly income, education, job satisfaction, work-life balance, overtime, and attrition status. The dataset contains 1,470 employee records and 35 initial attributes, with Attrition as the target variable consisting of two classes: Yes for employees who experienced turnover and No for employees who did not experience turnover.

The subject of this study was employee records contained in the dataset, while the object of the study was employee turnover risk based on human resource variables. Since the study used a public secondary dataset, the research location was not limited to a physical organization but was positioned within a data-driven HR analytics research context. The sampling technique used was total sampling, because all available employee records in the dataset were included in the analysis.

Table 1. Human resource variables used in the study

Variable Group	Attributes
Demographic factors	Age, Gender, MaritalStatus, Education, EducationField
Job-related factors	Department, JobRole, JobLevel, BusinessTravel, DistanceFromHome
Compensation factors	MonthlyIncome, DailyRate, HourlyRate, MonthlyRate, StockOptionLevel
Satisfaction factors	JobSatisfaction, EnvironmentSatisfaction, RelationshipSatisfaction
Work engagement factors	JobInvolvement, WorkLifeBalance, OverTime
Career factors	TotalWorkingYears, YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrManager
Target variable	Attrition

The variable grouping in this study was based on human resource management theory and employee turnover literature, which explain that attrition is influenced by individual characteristics, job conditions, compensation, satisfaction, engagement, and career-related experiences. Demographic factors describe employees' personal backgrounds that may shape career expectations, while job-related factors reflect work characteristics, organizational placement,

mobility, and workload. Compensation factors represent financial and reward aspects, whereas satisfaction factors indicate employees' psychological evaluation of their work environment and relationships. Work engagement factors capture involvement, work-life balance, and overtime exposure, which are associated with commitment and potential burnout. Career factors reflect tenure, promotion history, role duration, and managerial relationships as indicators of career development and organizational attachment. Therefore, the variable classification was not arbitrary, but was designed to connect dataset attributes with relevant HRM dimensions for turnover risk analysis. The use of these variables is also consistent with prior attrition studies showing that predictive HR models can support the identification of at-risk employees and improve retention and workforce planning decisions.

Research Procedure

The research procedure was conducted systematically using RapidMiner to ensure that the analysis could be replicated (Van den Heuvel & Bondarouk, 2017). The main workflow consisted of seven stages: dataset input, attribute selection, nominal-to-binominal transformation, role assignment, nominal-to-numerical conversion, cross-validation, and model evaluation. This procedure was designed to ensure that the dataset was properly prepared before the turnover risk modeling process was conducted.

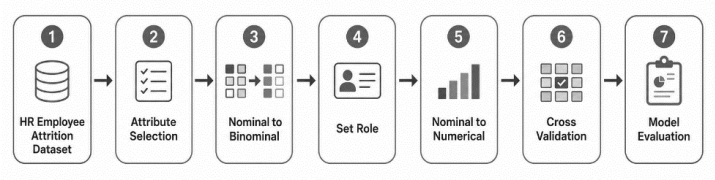

Figure 1. Research workflow

The figure above represents the overall research workflow used in this study. The process begins with importing the HR employee attrition dataset, followed by selecting relevant variables, transforming the target variable, defining the label, converting categorical predictors into numerical form, validating the model through cross-validation, and evaluating the model based on classification performance. This process follows the general logic of predictive analytics, where data preparation, model development, and evaluation are performed sequentially to generate reliable decision-support outputs for HR management.

Data Preprocessing

Data preprocessing was conducted to ensure that the dataset was suitable for turnover risk modeling. The first step was attribute selection by removing variables that did not provide meaningful information for managerial interpretation or predictive analysis. Four attributes were removed, namely EmployeeNumber, EmployeeCount, Over18, and StandardHours. EmployeeNumber was removed because it only functioned as an employee identifier, while EmployeeCount, Over18, and StandardHours were removed because they had constant values and therefore did not contribute to differentiating employee turnover risk.

Table 2. Removed attributes in data preprocessing

Removed Attribute	Reason for Removal
EmployeeNumber	Employee identifier and not relevant for turnover risk interpretation
EmployeeCount	Constant value
Over18	Constant value
StandardHours	Constant value

After attribute selection, the Attrition variable was transformed into a binominal variable and assigned as the label. The Yes class represented employees who experienced turnover, while the No class represented employees who did not experience turnover. Categorical predictor variables such as BusinessTravel, Department, EducationField, Gender, JobRole, MaritalStatus, and OverTime were converted into numerical form using dummy coding through the Nominal to

Numerical operator. This step was necessary to ensure that all models could process the predictor variables consistently.

Turnover Risk Modeling and Evaluation

Turnover risk modeling was conducted using four classification models: Naive Bayes, Decision Tree, Random Forest, and Gradient Boosted Trees (Iparraguirre-Villanueva et al., 2024; Krishna et al., 2022). These models were selected because they are frequently used in employee attrition prediction and HR analytics studies, and each model provides different analytical strengths. Naive Bayes was used as a baseline probabilistic model (Zhang, 2016), Decision Tree was used because of its interpretability (Zhao et al., 2021), Random Forest was used as an ensemble model with stable predictive capability (Ignatenko et al., 2024), and Gradient Boosted Trees was used as a boosting-based model to improve predictive performance through sequential learning. Studies on employee turnover prediction show that machine learning models are increasingly used to identify attrition risk and support employee retention decisions.

Model evaluation was conducted using 10-fold cross-validation to test model performance more reliably across different training and testing partitions. The evaluation metrics included accuracy, precision, recall, F1-score, AUC, and confusion matrix because predictive HR studies require evaluation measures that can capture both overall model performance and the ability to detect actual turnover cases (Iparraguirre-Villanueva et al., 2024). Since the purpose of this study was to identify employees at risk of turnover, the interpretation of precision, recall, and F1-score focused on the Attrition = Yes class. This focus is important because, in human resource management, failing to detect employees who are actually at risk of leaving may reduce the effectiveness of retention interventions.

Table 3. Model evaluation metrics

Metric	Function in This Study
Accuracy	Measures overall prediction correctness
Precision	Measures the correctness of turnover-risk predictions
Recall	Measures the ability to detect actual turnover cases
F1-score	Balances precision and recall for turnover-risk identification
AUC	Measures the model's ability to distinguish turnover and non-turnover cases
Confusion matrix	Identifies true positive, false positive, false negative, and true negative cases

The best model was selected not only based on accuracy but also on its managerial relevance for HR decision-making (Wang, 2024). In this study, a useful turnover risk model was defined as a model that could provide a balanced ability to identify employees at risk while minimizing misleading predictions. Therefore, model interpretation emphasized how predictive outputs could support data-driven retention strategies, including workload evaluation, compensation review, career development planning, employee engagement programs, and work-life balance management (Tessema, 2025).

3. RESULTS AND DISCUSSIONS

Employee Attrition Distribution

The first stage of the results focuses on the distribution of employee attrition as the basis for understanding turnover risk in the dataset. The IBM HR Analytics Employee Attrition and Performance dataset used in this study consists of 1,470 employee records, with the target variable classified into two categories: employees who experienced turnover and employees who did not experience turnover. The distribution shows that 1,233 employees, or 83.88% of the total data, were categorized as Attrition = No, while 237 employees, or 16.12%, were categorized as Attrition = Yes. This pattern indicates that the dataset is imbalanced, as the number of employees who remained in the organization is substantially higher than the number of employees who left.

Table 4. Distribution of employee attrition

Attrition Status	Number of Employees	Percentage
No	1,233	83.88%
Yes	237	16.12%
Total	1,470	100.00%

The imbalance in employee attrition distribution has important implications for turnover risk analysis. From a managerial perspective, the smaller proportion of employees who experienced turnover does not reduce the strategic importance of this group. Instead, it highlights the need for a more careful analytical approach because employees who leave the organization represent a critical risk group related to recruitment costs, productivity disruption, knowledge loss, and retention planning. If model evaluation only relies on overall accuracy, the analysis may be biased toward the majority class, namely employees who did not experience turnover. Therefore, this study interprets model performance by paying special attention to the Attrition = Yes class.

This distribution also supports the relevance of data-driven HR decision-making. Since turnover cases represent only 16.12% of the dataset, human resource managers need predictive tools that can help detect employees at risk more systematically. The attrition distribution becomes the empirical foundation for the next analysis stage, where several predictive models are evaluated not only based on general accuracy but also based on their ability to identify actual turnover cases through precision, recall, F1-score, AUC, and confusion matrix interpretation.

Turnover Risk Model Performance

The second stage of the results presents the performance comparison of four predictive models used to identify employee turnover risk, namely Naive Bayes, Decision Tree, Random Forest, and Gradient Boosted Trees (Chen et al., 2021). The evaluation was interpreted by focusing on the Attrition = Yes class because the main objective of this study is to identify employees who are at risk of leaving the organization. Therefore, precision, recall, and F1-score were calculated based on the turnover class, while accuracy and AUC were reported from the model evaluation results in RapidMiner.

Table 5. Turnover risk model performance

Model	Accuracy	Precision	Recall	F1-score	AUC
Naive Bayes	68.71%	30.87%	75.95%	43.91%	0.763
Decision Tree	83.61%	45.83%	9.28%	15.42%	0.631
Random Forest	84.42%	68.18%	6.33%	11.59%	0.795
Gradient Boosted Trees	84.49%	52.04%	48.52%	50.22%	0.796

The results show that each model produced different performance characteristics, confirming that employee turnover prediction should not be assessed only through accuracy but also through metrics that reflect the ability to detect the turnover class (Al Akasheh, Malik, et al., 2024). Naive Bayes achieved the highest recall of 75.95%, indicating that this model was the most sensitive in detecting employees who actually experienced turnover. However, its precision was only 30.87%, meaning that many employees predicted as turnover risks were actually non-turnover employees. This pattern suggests that Naive Bayes is useful as an early warning model, but less precise for targeted HR intervention.

Decision Tree and Random Forest produced high accuracy values of 83.61% and 84.42%, respectively. However, both models showed very low recall for the turnover class, at 9.28% and 6.33%. This indicates that although these models performed well in recognizing the majority class, namely non-turnover employees, they were less effective in detecting actual turnover cases. In the context of HR decision-making, this is problematic because the failure to identify at-risk employees may reduce the effectiveness of retention strategies.

Gradient Boosted Trees produced the most balanced performance, with the highest accuracy of 84.49%, the highest F1-score of 50.22%, and the highest AUC of 0.796. Although its recall was lower than Naive Bayes, Gradient Boosted Trees provided a better balance between precision and recall. Therefore, this model is more relevant for supporting data-driven HR decision-making because it can identify turnover risk more proportionally while reducing the tendency toward excessive false predictions.

Selection of the Most Relevant Model for HR Decision-Making

The selection of the most relevant model in this study was not determined solely by the highest accuracy, but by the model's ability to support human resource decision-making in identifying employees at risk of turnover (Levenson & Fink, 2017). In the context of employee

retention, a model with high accuracy may still be less useful if it fails to detect employees who actually experience turnover. Therefore, the interpretation of model performance emphasized the balance between accuracy, precision, recall, F1-score, and AUC, especially for the Attrition = Yes class.

Based on the model performance results, Gradient Boosted Trees was selected as the most relevant model for HR decision-making. This model achieved the highest accuracy of 84.49%, the highest F1-score of 50.22%, and the highest AUC of 0.796. Although Naive Bayes produced the highest recall of 75.95%, its precision was only 30.87%, indicating that many employees predicted as at risk of turnover were actually not turnover cases. This condition may lead HR managers to allocate retention interventions too broadly and less efficiently.

Decision Tree and Random Forest showed high accuracy values of 83.61% and 84.42%, respectively, but their recall values for the turnover class were very low. Decision Tree only reached 9.28% recall, while Random Forest only reached 6.33% recall. These findings indicate that both models were less effective in identifying employees who actually experienced turnover. From a managerial perspective, this is risky because employees who should receive early retention attention may not be detected by the system.

Table 6. Basis for selecting the most relevant model for HR decision-making

Model	Main Strength	Main Limitation	HR Decision-Making Relevance
Naive Bayes	Highest recall	Low precision	Useful for early screening
Decision Tree	High accuracy	Very low recall	Limited for turnover detection
Random Forest	Highest precision	Very low recall	Too conservative in identifying turnover
Gradient Boosted Trees	Most balanced performance	Moderate recall	Most relevant for retention strategy

Thus, Gradient Boosted Trees was considered the most relevant model because it provided a more balanced performance for turnover risk identification. For HR managers, this balance is important because retention strategies require not only the ability to detect employees at risk, but also reasonable precision to avoid unnecessary or misdirected interventions.

Confusion Matrix Analysis of Turnover Risk

The confusion matrix analysis was conducted on the selected model, namely Gradient Boosted Trees, to evaluate how well the model identified employees who were at risk of turnover. Since the main focus of this study is the Attrition = Yes class, the interpretation emphasizes the model's ability to correctly detect employees who actually experienced turnover and the potential managerial risk of incorrect predictions.

Table 7. Confusion matrix of gradient boosted trees

Actual / Predicted	Predicted Yes	Predicted No
Actual Yes	115	122
Actual No	106	1127

Based on Table 7, the Gradient Boosted Trees model correctly classified 115 employees who actually experienced turnover as Predicted Yes. These cases represent true positive predictions, meaning that the model successfully detected employees who were at risk of leaving. The model also correctly classified 1,127 employees who did not experience turnover as Predicted No, indicating strong capability in identifying retained employees.

However, the model produced 122 false negative cases, where employees who actually experienced turnover were predicted as not at risk. From a human resource management perspective, this is an important concern because false negative cases represent employees who may not receive early retention intervention. The model also produced 106 false positive cases, where employees who did not leave were predicted as having turnover risk. Although false positives may lead to broader HR monitoring, they are generally less harmful than false negatives because they still allow managers to provide preventive engagement, workload evaluation, or career support.

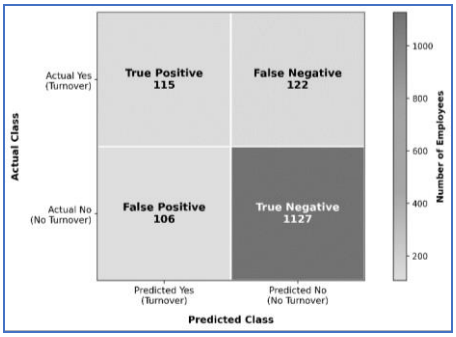

Figure 2. Confusion matrix visualization of gradient boosted trees

Overall, the confusion matrix indicates that Gradient Boosted Trees provides a balanced foundation for turnover risk analysis. The model is not perfect in detecting all turnover cases, but it offers useful information for prioritizing HR attention and supporting data-driven retention strategies.

Managerial Implications for Data-Driven Retention Strategy

The findings provide managerial implications for strengthening employee retention strategies through a data-driven HR approach. The selection of Gradient Boosted Trees shows that turnover risk analysis should be interpreted not only as a technical prediction result, but also as a managerial tool for identifying employees who require early attention. With 84.49% accuracy, 50.22% F1-score, 0.796 AUC, and 48.52% recall, the model offers a balanced basis for HR decision-making, although it can detect only nearly half of actual turnover cases. Therefore, predictive analytics should function as an early warning system rather than the sole basis for retention decisions. Employees predicted as high risk should be prioritized for targeted interventions such as workload evaluation, engagement surveys, career discussions, compensation review, and work-life balance support. At the same time, employees not flagged by the model should still be monitored through routine HR practices. Thus, predictive analytics should strengthen, not replace, managerial judgment, while HR managers remain responsible for validating model results and designing interventions according to organizational context.

From a retention strategy perspective, the model's results suggest that HR managers need to pay close attention to employees predicted as having turnover risk. These employees can be prioritized for further assessment through employee engagement surveys, workload evaluation, career discussions, compensation review, and work-life balance monitoring, as retention strategies are strongly connected to HR practices, compensation systems, employee experience, and career development opportunities (Hassan, 2022). The existence of 115 true positive cases indicates that predictive analytics can help identify employees who are actually at risk of turnover, allowing HR departments to design preventive interventions before resignation occurs.

However, the 122 false negative cases also show that predictive models should not replace managerial judgment. Employees who are not detected as at risk may still experience dissatisfaction, career stagnation, excessive workload, or lack of organizational attachment. Therefore, turnover risk modeling should be integrated with regular HR monitoring, supervisor feedback, and employee development programs. In this context, data-driven retention strategy means using predictive results as an early warning system while maintaining human-centered decision-making. The findings support the view that HR analytics can strengthen retention management by helping organizations allocate retention efforts more systematically, efficiently, and strategically.

Discussion

The findings of this study confirm that employee turnover risk analysis should not be interpreted merely through overall model accuracy, but through the model's ability to support meaningful HR decision-making. The distribution of the dataset shows that employees who experienced turnover represented only 16.12% of the total sample, while employees who did not experience turnover represented 83.88%. This imbalance explains why several models obtained

high accuracy but showed weak ability to detect actual turnover cases. In the context of human resource management, this finding is important because the minority class, namely employees who leave, is precisely the group that requires greater managerial attention.

The comparison of model performance shows that Naive Bayes achieved the highest recall for the turnover class at 75.95%, indicating strong sensitivity in detecting employees who actually experienced turnover. However, its precision was only 30.87%, which means that the model produced many false alarms. This finding suggests that Naive Bayes can be useful as an early screening tool, but may be less efficient for targeted retention intervention. In contrast, Decision Tree and Random Forest produced higher accuracy values of 83.61% and 84.42%, but their recall values for turnover cases were very low, at 9.28% and 6.33%. This confirms that high accuracy does not always reflect managerial usefulness in turnover risk analysis, especially when the model tends to classify most employees into the majority non-turnover class.

The selection of Gradient Boosted Trees as the most relevant model is supported by its balanced performance. The model achieved 84.49% accuracy, 52.04% precision, 48.52% recall, 50.22% F1-score, and 0.796 AUC. Although its recall was lower than Naive Bayes, it provided a better balance between detecting actual turnover cases and reducing inaccurate risk predictions. This balance is essential for HR managers because retention strategies require both early detection and efficient allocation of organizational resources. The confusion matrix also shows that the model correctly identified 115 turnover cases and 1,127 non-turnover cases, but still missed 122 employees who actually experienced turnover. These false negative cases should become a key concern because they represent employees who may leave without being detected by the HR system.

These results are consistent with previous studies emphasizing that HR analytics and predictive modeling can strengthen data-driven HR decision-making when interpreted beyond technical performance metrics. The contribution of this study lies in connecting predictive turnover analysis with managerial retention implications, consistent with the view that HR analytics should generate actionable business and human capital insights rather than merely statistical outputs (Hamilton & Sodeman, 2020). Rather than positioning machine learning as the final objective, this study uses the model results as a basis for identifying retention priorities, such as workload evaluation, compensation review, career development, employee engagement, and work-life balance improvement. Therefore, data-driven retention strategy should be understood as a combination of predictive analytics and human-centered managerial judgment. This approach enables HR departments to use data not only to predict turnover, but also to design more proactive, targeted, and evidence-based retention interventions.

4. CONCLUSION

This study concludes that employee turnover risk analysis is a strategic foundation for strengthening data-driven human resource retention strategies. Turnover should not be viewed merely as the number of employees leaving the organization, but as a risk pattern reflected in HRM-related variables such as job characteristics, compensation, satisfaction, work engagement, tenure, and career development. Among the four models evaluated, Gradient Boosted Trees was identified as the most relevant model for HR decision-making because it produced the most balanced performance, with 84.49% accuracy, 52.04% precision, 48.52% recall, 50.22% F1-score, and 0.796 AUC. These results indicate that high accuracy alone is insufficient if a model cannot detect employees who are genuinely at risk of leaving. Practically, predictive analytics can help organizations move from reactive retention practices toward proactive and evidence-based interventions by prioritizing at-risk employees for workload evaluation, compensation review, career development planning, engagement improvement, and work-life balance support. However, the recall value also shows that predictive models should function as an early warning tool rather than a single decision-making instrument. Therefore, HR managers should integrate predictive analytics with regular monitoring, supervisor feedback, and human-centered judgment. This study contributes by positioning predictive analytics as a managerial decision-support tool, although future research should use actual organizational data, longitudinal designs, psychological variables, and explainable AI methods to improve interpretability and practical relevance.

REFERENCES

Al Akasheh, M., Hujran, O., Malik, E. F., & Zaki, N. (2024). Enhancing the prediction of employee turnover with knowledge graphs and explainable AI. *IEEE Access*, *12*, 77041–77053. https://doi.org/10.1109/ACCESS.2024.3404829

Al Akasheh, M., Malik, E. F., Hujran, O., & Zaki, N. (2024). A decade of research on machine learning techniques for predicting employee turnover: A systematic literature review. *Expert Systems with Applications*, *238*, 121794. https://doi.org/10.1016/j.eswa.2023.121794

Căvescu, A. M., & Popescu, N. (2025). Predictive analytics in human resources management: Evaluating AIHR's role in talent retention. *AppliedMath*, *5*(3), 99. https://doi.org/10.3390/appliedmath5030099

Chen, H., Hu, S., Hua, R., & Zhao, X. (2021). Improved naive Bayes classification algorithm for traffic risk management. *Eurasip Journal on Advances in Signal Processing*, *2021*(1). https://doi.org/10.1186/S13634-021-00742-6

Fukui, S., Rollins, A. L., Salyers, M. P., & Rapp, C. A. (2023). Applying machine learning to human resources data to predict employee turnover. *Human Service Organizations: Management, Leadership & Governance*, *47*(3), 207–217.

Hamilton, R. H., & Sodeman, W. A. (2020). The questions we ask: Opportunities and challenges for using big data analytics to strategically manage human capital resources. *Business Horizons*, *63*(1), 85–95. https://doi.org/10.1016/j.bushor.2019.10.001

Hassan, Z. (2022). Employee retention through effective human resource management practices in Maldives: Mediation effects of compensation and rewards system. *Journal of Entrepreneurship, Management and Innovation*, *18*(2), 137–173. https://doi.org/10.7341/20221825

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

Ignatenko, V., Surkov, A., & Koltcov, S. (2024). Random forests with parametric entropybased information gains for classification and regression problems. *PeerJ Computer Science*, *10*. https://doi.org/10.7717/PEERJ-CS.1775

Iparraguirre-Villanueva, O., Guevara, J., & Sierra-Liñan, F. (2024). Employee attrition prediction using machine learning models. *Proceedings of the LACCEI International Multi-Conference for Engineering, Education and Technology*.

Kiran, P. R., Chaubey, A., & Shastri, R. K. (2024). Role of HR analytics and attrition on organisational performance: A literature review leveraging the SCM-TBFO framework. *Benchmarking: An International Journal*, *31*(9), 3102–3129. https://doi.org/10.1108/BIJ-06-2023-0412

Krishna, S., Sidharth, A., & Thirukkumaran, M. (2022). HR analytics: Employee attrition analysis using random forest. *International Journal of Performability Engineering*, *18*(4), 275–281. https://doi.org/10.23940/ijpe.22.04.p5.275281

Levenson, A., & Fink, A. (2017). Human capital analytics: Too much data and analysis, not enough models and business insights. *Journal of Organizational Effectiveness: People and Performance*, *4*(2), 145–156. https://doi.org/10.1108/JOEPP-03-2017-0029

Margherita, A. (2022). Human resources analytics: A systematization of research topics and directions for future research. *Human Resource Management Review*, *32*(2), 100795. https://doi.org/10.1016/j.hrmr.2020.100795

Marín Díaz, G., Galán Hernández, J. J., & Galdón Salvador, J. L. (2023). Analyzing employee attrition using explainable AI for strategic HR decision-making. *Mathematics*, *11*(22), 4677. https://doi.org/10.3390/math11224677

McAbee, S. T., Landis, R. S., & Burke, M. I. (2017). Inductive reasoning: The promise of big data. *Human Resource Management Review*, *27*(2), 277–290. https://doi.org/10.1016/j.hrmr.2016.08.005

McCartney, S., & Fu, N. (2022). Bridging the gap: Why, how and when HR analytics can impact organizational performance. *Management Decision*, *60*(13), 25–47. https://doi.org/10.1108/MD-12-2020-1581

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

Muhammad, G., & Naz, F. (2022). A moderating role of HR analytics between employee engagement, retention and organisational performance. *International Journal of Business Environment*, *13*(4), 345–357.

Rubenstein, A. L., Eberly, M. B., Lee, T. W., & Mitchell, T. R. (2018). Surveying the forest: A meta-analysis, moderator investigation, and future-oriented discussion of the antecedents of voluntary employee turnover. *Personnel Psychology*, *71*(1), 23–65. https://doi.org/10.1111/peps.12226

Tessema, S. A. (2025). The effect of human resource analytics on organizational performance. *Systems*, *13*(2), 134. https://doi.org/10.3390/systems13020134

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), 127–148. https://doi.org/10.1108/JOEPP-03-2017-0022

Wang, Y. (2024). Human resource analytics and data-driven human resource management. *Human Resource Management Review*.

Zhang, Z. (2016). Naïve Bayes classification in R. *Annals of Translational Medicine*, *4*(12), 241–241. https://doi.org/10.21037/ATM.2016.03.38

Zhao, L., Lee, S., & Jeong, S. P. (2021). Decision Tree Application to Classification Problems with Boosting Algorithm. *Electronics 2021, Vol. 10, Page 1903*, *10*(16), 1903. https://doi.org/10.3390/ELECTRONICS10161903