Smart Techno (Smart Technology, Informatic, and Technopreneurship)
Volume 08, Number 01, 2026 pp. 51-62
E-ISSN : 2541-0679
Open Access: https://lppm.primakara.ac.id/jurnal/index.php/smart-techno/

Predicting Employee Attrition Using the Random Forest Algorithm Based on IBM HR Analytics Data

Putu Satya Saputra[1]*, I Putu Gede Abdi Sudiatmika[2], Ni Putu Meiling Utami[3], I Putu Okta Priyana[4]
[1,2,3,4] Bali State Polytechnic, Badung, Indonesia

ARTICLE INFO
Article history:
Initial submission
09-11-2025
Received in revised form
22-11-2025
Accepted 25-11-2025
Available online 27-02-2026

Keywords:
Employee Attrition, Random Forest, HR Analytics, Employee Prediction, Machine Learning

DOI:
https://doi.org/10.59356/smart-techno.v7i1

ABSTRACT

The phenomenon of employee attrition has become a serious challenge for organizations, as it directly affects productivity, recruitment costs, and long-term performance stability. Understanding the factors that lead to employee turnover can no longer rely solely on manual observation; therefore, data-driven approaches are required to identify hidden patterns within workforce data. This study aims to predict employee attrition using the Random Forest algorithm applied to the IBM HR Analytics Employee Attrition & Performance dataset, which consists of 1,470 records and 35 attributes. The research stages include data preprocessing, handling class imbalance using the Synthetic Minority Over-sampling Technique (SMOTE), model training, and performance evaluation using accuracy, precision, recall, F1-score, ROC-AUC, and a confusion matrix. The results indicate that the baseline model without SMOTE exhibits low recall for the attrition class, whereas the application of SMOTE significantly improves model performance, particularly for the minority class, achieving a final accuracy of 83.96%. The most influential features identified are Stock Option Level, MonthlyIncome, and JobSatisfaction. These findings provide a comprehensive understanding of the factors influencing employee attrition and can serve as a foundation for organizations in designing more adaptive and data-driven employee retention strategies.

1. INTRODUCTION

Employees constitute the fundamental foundation of organizational sustainability, as their roles extend beyond the execution of technical tasks to include shaping organizational culture, fostering innovation, and supporting the achievement of long-term strategic visions. Workforce stability reflects the quality of internal management and serves as an indicator of overall organizational health. The phenomenon of employee attrition, or employees leaving an organization, has therefore emerged as a strategic indicator that significantly influences the sustainability of modern organizations.

Employee turnover, particularly voluntary turnover, generates serious consequences for organizations. The loss of experienced employees leads to decreased productivity, increased workloads for remaining staff, and the erosion of institutional knowledge that is often undocumented. Recruitment and training costs for new employees also increase, potentially disrupting the continuity of business processes. Ginting Munthe et al. (2024) emphasize that high turnover rates are directly associated with declining organizational performance and reduced motivation among remaining employees. Decisions to resign are typically influenced by multiple interrelated factors, including personal characteristics, psychological conditions, work environment factors, and organizational policies. Conventional analytical methods such as satisfaction surveys, interviews, and observations frequently fail to capture this complexity. Manual analysis tends to be subjective, time-consuming, and ineffective when applied to large-scale employee data. Moreover, such approaches are reactive in nature, identifying issues only after employees have already decided to leave.

* *Corresponding Author: Putu Satya Saputra: satya@pnb.ac.id*

Contemporary organizations therefore require more systematic, objective, and proactive strategies capable of identifying attrition risks at an early stage. The utilization of historical employee data provides opportunities to uncover hidden patterns that are not readily detectable through human observation. In this context, machine learning technologies offer an alternative approach for modeling employee behavior with greater accuracy. Machine learning algorithms are capable of identifying non-linear relationships among features, recognizing complex patterns, and generating attrition predictions based on prior data experiences (Rama Krishna Debbadi & Obed Boateng, 2025).

Random Forest is one of the most widely used algorithms for classification problems, including employee attrition analysis. This algorithm operates as an ensemble of multiple decision trees trained on randomly selected subsets of data and features. The majority voting mechanism enhances model stability, reduces the risk of overfitting, and produces more consistent predictive performance. Ignatenko et al. (2024) argue that Random Forest is particularly suitable for datasets containing a mixture of numerical, ordinal, and categorical variables. Similarly, Thomas and Kaliraj (2024) demonstrate the algorithm's superiority across various classification contexts due to its strong generalization capability.

The IBM HR Analytics Employee Attrition & Performance dataset provides comprehensive and relevant data for employee attrition prediction research. The dataset consists of 1,470 records and 35 attributes encompassing personal characteristics, job history, compensation, job satisfaction, and organizational environment factors. The diversity of features enables a more in-depth analysis of the determinants influencing employees' decisions to stay or leave. The target variable, Attrition, is binary in nature, making it well suited for a classification approach using the Random Forest algorithm. The application of this algorithm extends beyond predictive accuracy, aiming to support evidence-based human resource decision-making. Al-Shammari et al. (2025) emphasize that strategic decisions in human resource management require strong analytical foundations rather than reliance on intuition alone. Predictive models can assist HR practitioners in identifying high-risk employees, enabling early interventions such as engagement enhancement programs, training initiatives, or adjustments to compensation policies.

An additional advantage of Random Forest lies in its ability to generate feature importance measures. Information regarding the most influential features enables organizations to better understand internal factors most strongly associated with attrition decisions. These insights can serve as a basis for designing more targeted retention strategies, improving job satisfaction, enhancing the work environment, and implementing more effective reward systems. This approach aligns with the principles of evidence-based HR articulated by Falletta and Combs (2021), which advocate for the use of empirical evidence to support managerial decision-making.

Despite its potential benefits, the application of machine learning in human resource management requires careful consideration of ethical and fairness issues. Berber and Srećković (2024) caution that algorithms should not serve as the sole basis for decisions that affect individuals' careers and livelihoods. Predictive models must be transparent, accountable, and free from discriminatory bias. Employee data should be managed securely, anonymized where appropriate, and processed in accordance with privacy principles. Clear explanations of model functionality and limitations are essential for maintaining trust among stakeholders.

Overall, a Random Forest–based predictive approach offers significant opportunities for transforming modern human resource management. Organizations can transition from reactive analytical methods toward proactive systems capable of identifying risks early and tailoring retention strategies based on dominant influencing factors. The development of predictive models using the IBM HR Analytics Employee Attrition & Performance dataset contributes to strengthening decision-making processes that are more objective, adaptive, and sustainable.

2. LITERATURE REVIEW

Predicting employee attrition has become a major concern in modern human resource management. Many organizations are increasingly shifting toward data-driven approaches to understand why employees choose to remain with or leave their jobs. Traditional analytical methods often fail to capture the complexity of the underlying factors, making machine learning technologies a relevant solution for uncovering hidden patterns within employee data.

Numerous studies have demonstrated the effectiveness of machine learning algorithms in identifying potential employee turnover. Algorithms such as Decision Tree, Logistic Regression, and Support Vector Machine have been widely applied; however, Random Forest is often regarded as superior due to its ability to handle complex data structures and generate stable predictions. Humairah and Agustina (2024) introduced an entropy-based information gain approach within Random Forest, which was shown to improve classification accuracy on non-normally distributed data. These findings are reinforced by Priantama and Yoga Siswa (2022), who reported that optimized versions of Random Forest can significantly enhance predictive performance across various classification contexts.

The application of Random Forest has also proven effective beyond industrial settings. A study by Gori et al. (2024) on student attrition demonstrated that this algorithm achieved more accurate classification results compared to single-model approaches. These results highlight the flexibility of Random Forest in modeling human behavioral patterns across diverse domains. The adoption of such predictive approaches aligns with the concept of evidence-based HR as articulated by Falletta and Combs (2021), which emphasizes decision-making practices grounded in systematic data analysis. This approach enables organizations to identify turnover risks at an early stage and develop more targeted retention strategies.

Feature selection represents another critical aspect in the development of high-quality predictive models. A comprehensive study by Setiyadi et al. (2024) emphasizes that the selection of relevant features directly influences model performance. Variables such as job satisfaction, salary, overtime frequency, and working environment conditions frequently emerge as dominant factors in explaining employees' decisions to stay or leave. Understanding these variables enhances the interpretability of predictive models and allows the results to be utilized as a foundation for more contextual and informed organizational policies.

Beyond technical considerations, ethical and fairness issues have received increasing attention in the development of machine learning based models. Abdul Rahman (2025) asserts that predictive models should not be used as the sole basis for decision-making due to the potential biases embedded in training data. Jan Melvin Ayu Soraya Dachi and Pardomuan Sitompul (2023) further argue that model evaluation should be conducted comprehensively using metrics such as precision, recall, and confusion matrices to ensure that outcomes are not only numerically accurate but also fair and accountable.

Despite the meaningful contributions of prior studies, several research gaps remain. Most existing studies have not extensively explored the context of global corporations using real-world datasets such as the IBM HR Analytics Employee Attrition & Performance dataset. This dataset represents a multinational corporate environment with high diversity, offering broader and more realistic insights. Additionally, the issue of class imbalance between employees who remain and those who leave is often insufficiently addressed in previous research.

This study extends prior work by incorporating feature importance analysis as a means of interpreting model results. Such analysis helps explain which variables exert the greatest influence on employee attrition decisions, enabling the model outcomes to support more effective retention policies. Furthermore, transparency and ethical considerations are emphasized to ensure that the applied technology remains responsible and does not introduce discriminatory practices.

Overall, this research seeks to address the limitations of previous studies through the application of Random Forest, with a focus not only on technical accuracy but also on a deeper understanding of the human factors underlying the data. This approach is expected to assist organizations in developing human resource management systems that are data-driven, ethical, and sustainable.

3. METHOD

This study is systematically designed to develop and evaluate a predictive model capable of classifying employee attrition risk based on historical data. The proposed approach integrates data exploration techniques, feature engineering, and machine learning using the Random Forest algorithm. Each stage of the process is conducted to ensure methodological validity and the accuracy of the predictive outcomes.

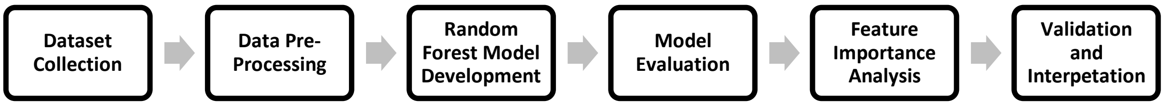
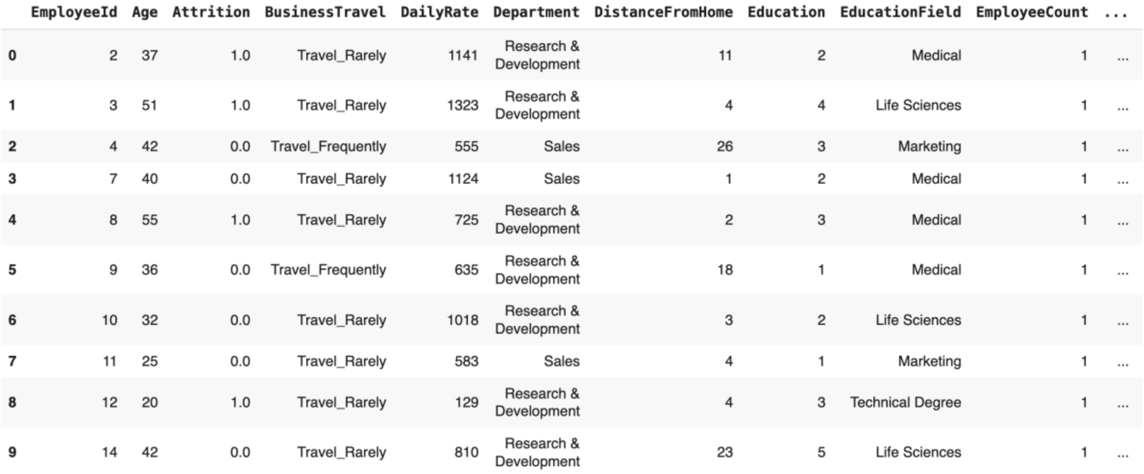

Dataset Collection
Data Pre-Processing
Random Forest Model Development
Model Evaluation
Feature Importance Analysis
Validation and Interpetation

Figure 1. Research Workflow

Figure 1 illustrates the sequence of stages carried out throughout the research. The process begins with data collection and proceeds through to comprehensive interpretation of the model results. Each stage plays a crucial role in developing a predictive system that is not only technically accurate but also relevant and practically applicable within the context of human resource management.

1. Data Source

The dataset used in this study is an open-access dataset provided by IBM, entitled HR Analytics Employee Attrition & Performance.



	EmployeeId	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EmployeeCount	...
0	2	37	1.0	Travel_Rarely	1141	Research & Development	11	2	Medical	1	...
1	3	51	1.0	Travel_Rarely	1323	Research & Development	4	4	Life Sciences	1	...
2	4	42	0.0	Travel_Frequently	555	Sales	26	3	Marketing	1	...
3	7	40	0.0	Travel_Rarely	1124	Sales	1	2	Medical	1	...
4	8	55	1.0	Travel_Rarely	725	Research & Development	2	3	Medical	1	...
5	9	36	0.0	Travel_Frequently	635	Research & Development	18	1	Medical	1	...
6	10	32	0.0	Travel_Rarely	1018	Research & Development	3	2	Life Sciences	1	...
7	11	25	0.0	Travel_Rarely	583	Sales	4	1	Marketing	1	...
8	12	20	1.0	Travel_Rarely	129	Research & Development	4	3	Technical Degree	1	...
9	14	42	0.0	Travel_Rarely	810	Research & Development	23	5	Life Sciences	1	...

Figure 2. Dataset

The dataset consists of 1,470 rows and 35 columns encompassing personal information, job-related attributes, compensation details, work environment factors, and employee attrition status. All data are tabular in nature and are structurally represented using a combination of numerical, ordinal, and categorical features.

2. Data Preprocessing

Prior to model implementation, the data underwent a preprocessing stage to ensure optimal input quality for the predictive model. This stage included the following steps:

- **Categorical variable encoding:** Binary variables such as Yes/No were transformed into 1/0, while other categorical attributes, including Gender, OverTime, MaritalStatus, and Department, were converted into numerical representations using one-hot encoding.

- **Feature and target separation:** The classification target was defined as the Attrition column, while the remaining attributes were used as predictor variables.
- **Training and testing data split:** The dataset was divided into 80% training data and 20% testing data to evaluate the model's generalization performance.

```
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
import pandas as pd

df = pd.read_csv("employee_data.csv")
df['Attrition'] = df['Attrition'].map({'Yes': 1, 'No': 0})
df_encoded   =   pd.get_dummies(df.drop('EmployeeNumber',   axis=1),
drop_first=True)

X = df_encoded.drop('Attrition', axis=1)
y = df_encoded['Attrition']

X_train,   X_test,   y_train,   y_test   =   train_test_split(X,   y,
test_size=0.2, random_state=42)
```

The IBM HR Analytics Employee Attrition & Performance dataset consists of 1,470 records and 35 attributes covering personal characteristics, compensation, work environment, job satisfaction, and employee attrition status. The class distribution of the target variable indicates a substantial imbalance, with 1,233 employees (83.87%) classified as Attrition = No and only 237 employees (16.13%) classified as Attrition = Yes. This imbalance has the potential to reduce the model's ability to accurately identify the minority class, thereby necessitating the application of data balancing techniques.

Data quality was assessed prior to the modeling process. All attributes in the dataset were found to contain no missing values; therefore, imputation was not required. Outlier detection was performed on numerical variables, namely MonthlyIncome, Age, DailyRate, and YearsAtCompany, using boxplot visualization and the interquartile range (IQR) method. The identified outliers were not removed, as they remain within the natural distribution of employees in a large corporate environment and are considered valid representations of real-world conditions. Removing these outliers could potentially eliminate important patterns related to attrition risk.

Data preprocessing was conducted through several stages to ensure data readiness before model training. Categorical variables, including Gender, MaritalStatus, BusinessTravel, JobRole, and Department, were transformed into numerical form using one-hot encoding. Binary variables, namely Attrition and OverTime, were converted into 1 and 0. The EmployeeNumber attribute was removed due to its lack of predictive value. Following the encoding process, the data were separated into features (X) and target (y) variables and subsequently split into 80% training data and 20% testing data. Class imbalance was addressed using the Synthetic Minority Oversampling Technique (SMOTE), which generated a balanced number of minority-class samples relative to the majority class within the training data. The SMOTE process was applied exclusively to the training set to prevent information leakage (data leakage).

3. Random Forest Model Development

The Random Forest algorithm was employed as the primary classification technique. This model consists of an ensemble of decision trees that operate in parallel to improve predictive accuracy and reduce the risk of overfitting. The model was trained using the training dataset (X_train, y_train) and evaluated on the testing dataset (X_test, y_test).

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report,
confusion_matrix

rf_model = RandomForestClassifier(n_estimators=100,
random_state=42)
rf_model.fit(X_train, y_train)

y_pred = rf_model.predict(X_test)
```

To enhance model performance, a simple hyperparameter tuning experiment was conducted on the Random Forest algorithm. The *n_estimators* parameter was evaluated across several values (100, 200, and 300). The experimental results indicate that increasing the number of trees tends to improve prediction stability; however, it does not yield a significant improvement in overall accuracy or recall for the minority class. These findings suggest that model performance is more strongly influenced by data balancing through SMOTE than by the number of estimators. Nevertheless, hyperparameter adjustment provides additional validation that the model configuration was evaluated under multiple experimental scenarios.

4. Model Evaluation

Model evaluation was conducted using accuracy, precision, recall, F1-score, confusion matrix, and ROC-AUC to ensure a comprehensive assessment of model performance. These metrics are particularly important given the class imbalance present in this study, as accuracy alone is insufficient to adequately represent the model's predictive effectiveness.

- Precision & Recall: These metrics provide class-specific performance measurements. *Precision* describes the proportion of correctly predicted positive instances among all instances classified as positive.

$$Precision = \frac{TP\ (True\ Positive)}{TP\ (True\ Positive) + TP\ (True\ Positive)} \qquad (1)$$

Recall describes the model's ability to correctly identify actual positive instances,

$$Recall = \frac{TP\ (True\ Positive)}{TP\ (True\ Positive) + FN\ (False\ Negative)} \qquad (2)$$

- F1-score : The harmonic mean of precision and recall

$$F1\ score = 2\ x\ \frac{Precision\ x\ Recall}{Precision + Recall} \qquad (3)$$

- *Accuracy* : The proportion of correct predictions relative to the total number of predictions

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (4)$$

- *Confusion Matrix* It is used to evaluate the accuracy of the model's classification based on the obtained values
  - TP (*True Positive*)
  - TN (*True Negative*)
  - FP (*False Positive*)
  - FN (*False Negative*)

- ROC-AUC: This metric measures the model's ability to distinguish between positive and negative classes through the Receiver Operating Characteristic (ROC) curve. An Area Under the Curve (AUC) value closer to 1 indicate

```
from sklearn.metrics import roc_auc_score, plot_roc_curve
import matplotlib.pyplot as plt
import seaborn as sns

print(confusion_matrix(y_test, y_pred))
print(classification_report(y_test, y_pred))
```

5. Feature Importance Analysis

Random Forest has an advantage in measuring feature importance, which reflects the relative influence of each variable on the prediction outcomes. These values help identify the most relevant features in employee attrition classification, such as OverTime, MonthlyIncome, and JobSatisfaction.

```
import numpy as np

importances = rf_model.feature_importances_
indices = np.argsort(importances)[-10:]

plt.figure(figsize=(10, 6))
plt.title("Top 10 Feature Importances")
plt.barh(range(len(indices)), importances[indices], align="center")
plt.yticks(range(len(indices)), [X.columns[i] for i in indices])
plt.xlabel("Importance Score")
plt.tight_layout()
plt.show()
```

6. Validation and Interpretation

The evaluation was conducted not only based on numerical accuracy but also on the extent to which the model is fair, accountable, and free from bias. The model evaluation results were carefully examined to avoid complete reliance on automated decision-making. Feature importance visualizations provide a strong interpretative basis for supporting data-driven policy recommendations.

5. RESULT AND DISCUSSION

1. Model Evaluation Results Before SMOTE

Initial testing was conducted on the original dataset without class balancing. The results indicate that the model exhibits a strong tendency to classify employees as "staying" (the majority class). The Random Forest model prior to the application of SMOTE achieved an accuracy of 82.31%. However, performance on the attrition class (*Attrition = 1*) was considerably lower.

Table 1. Precision, Recall, F1-Score, And Support Values Before Smote

Class	Precision	Recall	F1Score	Support
Staying(0)	0.85	0.90	0.90	247
Leaving(1)	0.31	0.85	0.13	47

The very low recall value (8.51%) indicates that the model is almost unable to identify employees who actually leave the organization. Support represents the number of actual samples for each class present in the test dataset during model evaluation.

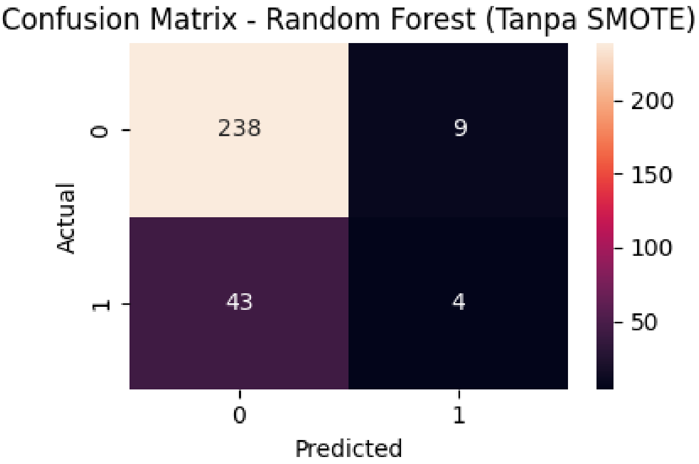


Figure 3. Confusion Matrix After SMOTE

The confusion matrix results before the application of SMOTE indicate that only four employees who left the organization were correctly identified by the model. The ROC-AUC value of 0.7673 suggests that the model exhibits a moderate level of discriminative ability.

2. Model Evaluation Results After SMOTE

Model training using the Random Forest algorithm demonstrates satisfactory performance in classifying employee attrition risk after the application of SMOTE. The evaluation was conducted on the test dataset, which comprised 20% of the total dataset and had undergone class balancing through SMOTE.

The model achieved an accuracy of 83.96%, indicating a high overall level of predictive correctness. However, the primary focus extends beyond overall accuracy to the classification performance of employees who left the organization (Attrition=1), which represents the minority class and holds greater strategic importance in the context of employee retention policies.

Table 2. Precision, Recall, F1-Score, and Support Values After SMOTE

Class	Precision	Recall	F1Score	Support
Staying(0)	0.88	0.93	0.91	176
Leaving(1)	0.54	0.39	0.45	36

The evaluation metrics employed include accuracy, precision, recall, F1-score, confusion matrix, and ROC-AUC, as described in the methodology section. All of these metrics are also consistently reported in the results section. The ROC-AUC value of the baseline model is 0.7673, indicating a moderate level of discriminative capability. For the model after the application of SMOTE, ROC-AUC was also calculated to ensure consistency in the evaluation process, thereby maintaining alignment between the methodology and the reported results.

Support represents the number of actual samples in each class within the test dataset. This value indicates how many instances truly belong to the “No Attrition” (0) and “Attrition” (1) classes at the time of evaluation. A larger support value for a given class implies a greater contribution to the overall model evaluation. In this study, the “No Attrition” class has a support of 176, while the “Attrition” class has a support of 36, further confirming the presence of class imbalance in the dataset.

The recall value for the Attrition = 1 class is 0.39, indicating that the model is able to correctly identify only 39% of employees who actually leave the organization. This result reflects a common challenge in predicting minority classes, even after the application of SMOTE. In contrast, the precision value for the attrition class is 0.54, meaning that 54% of the instances predicted as “leaving” are correct. Beyond overall accuracy, the model was evaluated using precision, recall, F1-score, confusion matrix, and ROC-AUC in accordance with the methodological framework. The resulting ROC-AUC value suggests that the model demonstrates good capability in distinguishing the majority class, while still leaving room for improvement in the classification of the minority class. These findings reinforce the notion that dataset imbalance continues to affect the model’s sensitivity to attrition prediction, despite the application of SMOTE.

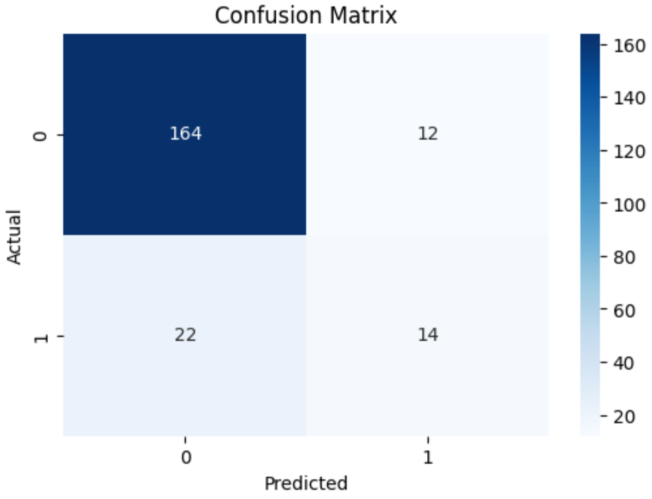


Figure 4. Confusion Matrix After SMOTE

The application of SMOTE was performed exclusively on the training data to prevent data leakage, thereby preserving the original imbalanced distribution in the test dataset. Consequently, despite the use of SMOTE, Figure 4 still reflects an imbalance in the number of actual samples across classes in the test set (support). This condition is expected, as SMOTE does not modify the test data but instead assists the model in learning the patterns of the minority class more effectively. The prediction results shown in Figure 4 indicate that the model is able to detect a greater number of attrition cases after the application of SMOTE, even though the test set distribution remains imbalanced.

Table 3. Predicted and Actual Values After SMOTE

	Predicted Staying	**Predicted Leaving**
Actual Staying	164	12
Actual Leaving	22	14

The model successfully classified the majority of employees who remained, but it still faces challenges in detecting employees who are likely to leave, a phenomenon generally influenced by complex psychosocial factors that are not directly observable through quantitative data.

Random Forest provides the capability to evaluate the relative influence of each feature on the final predictions. Feature importance visualization identified the top 10 features that are most determinant in modeling employee attrition.

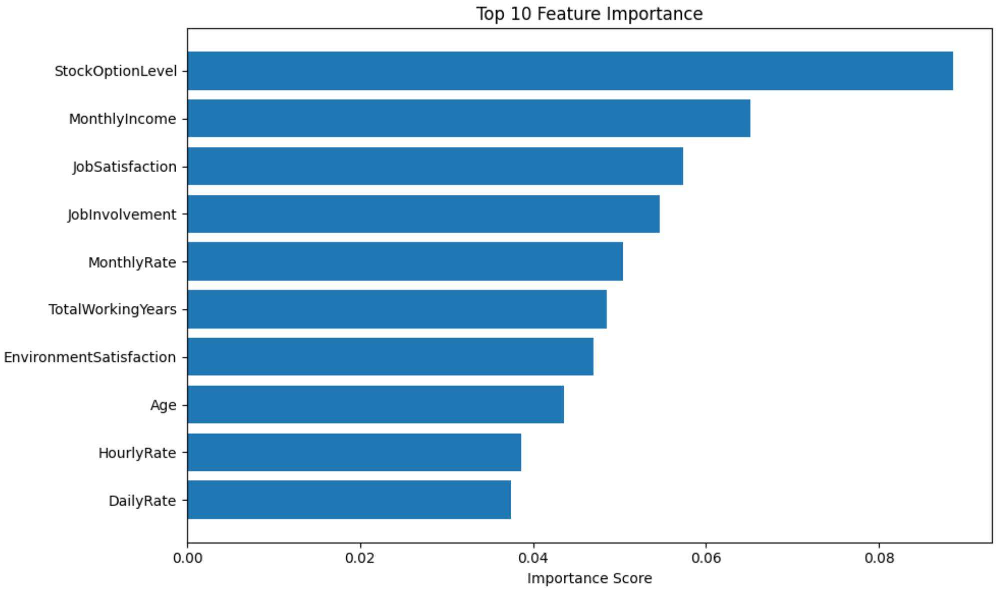


Figure 5. Top 10 Feature Importances

The StockOptionLevel feature emerged as the most significant variable. This finding indicates that employees who receive stock options tend to exhibit higher loyalty and are less likely to leave the organization. Equity-based financial incentives appear to play a role in fostering a sense of ownership and long-term commitment.

MonthlyIncome and JobSatisfaction further reinforce the notion that financial satisfaction and emotional attachment to the job are important factors in employees' retention decisions. TotalWorkingYears and EnvironmentSatisfaction reflect experience and the quality of the work environment, which together influence the likelihood of resignation.

The model demonstrates the ability to map employee attrition risk profiles based on historical data. Although recall for the attrition class remains suboptimal, the insights gained from feature importance analysis provide a solid foundation for policy improvements. Organizations can prioritize retention programs based on high-impact variables such as stock incentives, competitive salaries, and strategies to enhance job satisfaction.

This approach supports the creation of a more proactive and data-driven HR ecosystem. Predictive systems serve not only as risk-monitoring tools but also as strategic navigators that enable fair, evidence-based decisions with tangible impacts on employee loyalty.

6. CONCLUSION

The application of the Random Forest algorithm in predicting employee attrition demonstrates that a data-driven approach can provide a reasonably accurate risk mapping. The model achieved an accuracy of 83.96%, indicating that the majority of predictions were made correctly. Although the model's performance on the attrition class remains suboptimal in terms of recall, the system was still able to consistently identify a number of critical cases.

Feature importance contributed significantly to the interpretation of the results. Factors such as stock option levels, monthly income, job satisfaction, and job engagement emerged as the most influential variables. These findings reinforce the understanding that financial motivation and affective conditions play a major role in employees' decisions to stay or leave the organization.

The developed model functions not only as a classification tool but also as a foundation for evidence-based decision-making in human resource management. Once critical features are identified, retention strategies can be formulated more specifically. Such an approach enables organizations to proactively prevent attrition rather than merely reacting to it.

Machine learning-based predictive modeling provides opportunities for developing adaptive early-warning systems that continuously learn from historical patterns. However, the use of such algorithms must remain within an ethical, transparent framework and should not replace professional judgment. The system has significant potential to be integrated into modern HR management processes, particularly in organizations seeking to foster a data-driven work culture and sustain long-term employee relationships.

7. ACKNOWLEDGEMENT

The authors would like to express their gratitude to IBM for providing the HR Analytics Employee Attrition & Performance dataset used in this study. Sincere appreciation is also extended to colleagues and fellow faculty members for their support, guidance, and constructive feedback throughout the research process. Thanks are further given to family members and all individuals who provided encouragement and motivation, enabling the successful completion of this study.

8. REFERENCES

Abdul Rahman. (2025). Urgensi Pembuatan Model Prediktif Dalam Tata Kelola Bisnis. Jurnal Ilmiah Multidisiplin Ilmu, 2(2), 78–87. https://doi.org/10.69714/jvtm6q52

Al-Shammari, M., Al Bin Ali, F., AlRashidi, M., & Albuainain, M. (2025). Big Data and Predictive Analytics for Strategic Human Resource Management: A Systematic Literature Review. International Journal of Computing and Digital Systems, 17(1), 1–9. https://doi.org/10.12785/ijcds/1571015706

Berber, A., & Srećković, S. (2024). When something goes wrong: Who is responsible for errors in ML decision-making? AI & SOCIETY, 39(4), 1891–1903. https://doi.org/10.1007/s00146-023-01640-1

Falletta, S. V., & Combs, W. L. (2021). The HR analytics cycle: a seven-step process for building evidence-based and ethical HR analytics capabilities. Journal of Work-Applied Management, 13(1), 51–68. https://doi.org/10.1108/JWAM-03-2020-0020

Ginting Munthe, R., Susan, M., & Sulungbudi, B. M. (2024). The Role of Internal Marketing in Building Organizational Commitment and Reducing Turnover Intention Affecting the Improved Performance of Life Insurance Agents in Indonesia. Aptisi Transactions on Technopreneurship (ATT), 6(1), 56–71. https://doi.org/10.34306/att.v6i1.387

Gori, T., Sunyoto, A., & Al Fatta, H. (2024). Preprocessing Data dan Klasifikasi untuk Prediksi Kinerja Akademik Siswa. Jurnal Teknologi Informasi Dan Ilmu Komputer, 11(1), 215–224. https://doi.org/10.25126/jtiik.20241118074

Humairah, P., & Agustina, D. (2024). Stock Price Index Prediction Using Random Forest Algorithm for Optimal Portfolio. Jurnal Varian, 8(1), 113–124. https://doi.org/10.30812/varian.v8i1.4276

Ignatenko, V., Surkov, A., & Koltcov, S. (2024). Random forests with parametric entropy-based information gains for classification and regression problems. PeerJ Computer Science, 10, e1775. https://doi.org/10.7717/peerj-cs.1775

Jan Melvin Ayu Soraya Dachi, & Pardomuan Sitompul. (2023). Analisis Perbandingan Algoritma XGBoost dan Algoritma Random Forest Ensemble Learning pada Klasifikasi Keputusan Kredit. Jurnal Riset Rumpun Matematika dan Ilmu Pengetahuan Alam, 2(2), 87–103. https://doi.org/10.55606/jurrimipa.v2i2.1470

Mohamed, M., Abdullah, A., Zaki, A. M., Rizk, F. H., Eid, M. M., & El-Kenway, E. M. El. (2024). Advances and Challenges in Feature Selection Methods: A Comprehensive Review. Journal of Artificial Intelligence and Metaheuristics, 7(1), 67–77. https://doi.org/10.54216/JAIM.070105

P, P., L R, S., & V, K. (2024). Forecasting Student Attrition Using Machine Learning. 2024 4th Asian Conference on Innovation in Technology (ASIANCON), 1–7. https://doi.org/10.1109/ASIANCON62057.2024.10838214

Priantama, Y., & Yoga Siswa, T. A. (2022). Optimasi Correlation-Based Feature Selection Untuk Perbaikan Akurasi Random Forest Classifier Dalam Prediksi Performa Akademik Mahasiswa. JIKO (Jurnal Informatika Dan Komputer), 6(2), 251. https://doi.org/10.26798/jiko.v6i2.651

Rama Krishna Debbadi, & Obed Boateng. (2025). Optimizing End-To-End Business Processes by Integrating Machine Learning Models with Uipath for Predictive Analytics and Decision Automation. International Journal of Science and Research Archive, 14(2), 778–796. https://doi.org/10.30574/ijsra.2025.14.2.0448

Salari, M., Sadati, S. M., Sedaghat, A., Abbasi, B., Zamanpour, S. A., Khodashahi, R., & Davoudi, M. (2025). Evaluating the Application of Machine Learning in Predicting the Mortality of Hospitalized COVID-19 Patients Using the Confusion Matrix and the Matthews Correlation Coefficient. Archives of Clinical Infectious Diseases, 20(2). https://doi.org/10.5812/archcid-150150

Salman, H. A., Kalakech, A., & Steiti, A. (2024). Random Forest Algorithm Overview. Babylonian Journal of Machine Learning, 2024, 69–79. https://doi.org/10.58496/BJML/2024/007

Setiyadi, P., Prayogi, M. N., & Solichin, A. (2024). OPTIMALISASI PREDIKSI KEHILANGAN KARYAWAN MENGGUNAKAN TEKNIK RFE, SMOTE, DAN ADABOOST. JIPI (Jurnal Ilmiah Penelitian Dan Pembelajaran Informatika), 9(4), 2131–2145. https://doi.org/10.29100/jipi.v9i4.5642

Thomas, N. S., & Kaliraj, S. (2024). An Improved and Optimized Random Forest Based Approach to Predict the Software Faults. SN Computer Science, 5(5), 530. https://doi.org/10.1007/s42979-024-02764-x