Employee Attrition Prediction Using Machine Learning

**Shriyank DH, Saniya Prasad Mane, Soumyadeep Gosh, Srujana SM**Department of Information Science and Engineering  
BMS Institute of Technology and Management, VTU, India  
Email: {1by22is144,1by22is148,1by22is164,1by22is169}@bmsit.in

# Abstract

Employee attrition poses a significant challenge for organizations, leading to productivity loss and recruitment costs. This study presents a machine learning-based approach to predict employee attrition using classification models such as Random Forest and Support Vector Machine (SVM). We utilized a public IBM dataset containing 10,000 employee records, applying data preprocessing, class balancing using SMOTE, and model evaluation techniques. The Random Forest model achieved superior accuracy and interpretability, making it suitable for deployment in HR analytics. This paper discusses the implementation, evaluation, and key findings. Organizations are increasingly adopting predictive analytics to minimize talent loss. Proactive interventions enabled by predictive models can save costs and boost employee satisfaction. This study presents a structured pipeline and model comparison, highlighting key insights for HR professionals.

# Keywords

Employee attrition, machine learning, Random Forest, SVM, SMOTE, HR analytics.

# I. Introduction

Employee attrition is a critical concern for organizations, especially in highly competitive sectors. Traditional methods such as surveys and interviews are often reactive. Machine Learning (ML) allows for proactive, data-driven approaches to identify potential employee exits. This research uses ML models to predict attrition with high accuracy, aiming to support HR departments with actionable insights. Another significant challenge is dealing with imbalanced datasets where non-attrition instances dominate, leading to skewed results. Studies such as Siandri et al. and Suganthi et al. have recommended integrating explainable AI (XAI) techniques like SHAP and LIME to ensure interpretability. These tools enable HR personnel to trust and act on model recommendations.

# II. Related Work

Several studies have utilized ML techniques for attrition prediction. Bansal et al. [1] used logistic regression identifying job role and overtime as key features. Kumar and Singh [2] highlighted the strength of ensemble models like Random Forest. More recent studies such as Jerly Akku et al. [3], and Tadikonda et al. [4] applied tree-based and boosting algorithms, achieving accuracy over 90%. These efforts highlight the promise of ML in HR analytics but also show limitations like class imbalance and interpretability issues, which our work addresses. Feature selection techniques were used to eliminate redundant features. Model training was performed using 5-fold cross-validation to ensure generalization. Hyperparameters for Random Forest (n\_estimators, max\_depth) and SVM (C, gamma) were tuned using grid search. Performance metrics such as ROC-AUC were also analyzed to evaluate classifier robustness.

# III. Methodology

Our pipeline consists of data loading, preprocessing, feature encoding, data balancing using SMOTE, model training, and evaluation. The IBM HR dataset was cleaned, encoded using LabelEncoder, and balanced using SMOTE to address class imbalance. We applied two models—Random Forest and SVM—to compare performance. Evaluation was done using accuracy, F1-score, and confusion matrices.

# IV. Results and Evaluation

Model performance on the test set showed Random Forest achieving an accuracy of 87.19% and an F1-score of 85.32%, outperforming SVM and KNN. Feature importance analysis identified key predictors like OverTime, MonthlyIncome, and JobRole. Random Forest provided better interpretability and precision in identifying at-risk employees.

# V. Discussion

The proposed system demonstrates scalability and practicality for large organizations. Using SMOTE significantly improved minority class prediction. While SVM performed reasonably well, Random Forest’s feature importance visualization offers critical advantages for real-world HR decision-making.

# VI. Conclusion

This paper presents a practical machine learning framework to predict employee attrition. By addressing class imbalance and evaluating model interpretability, the system becomes more actionable for HR stakeholders. Future work may involve deploying this model in real-time HR platforms and integrating more behavioral data.

# References

1. [1] Bansal et al., “Predicting Employee Attrition using Logistic Regression,” 2019.
2. [2] Kumar and Singh, “Ensemble Learning Approaches for Attrition Prediction,” 2021.
3. [3] J. A. Cherian et al., “Predictive Modeling of Employee Attrition Using Tree‑Based Classifiers,” Feb. 2025.
4. [4] T. K. Rohan et al., “Employee Attrition Using Gradient Boosting,” ICOEI 2025.
5. [5] E. Siandri et al., “Machine Learning in Employee Attrition,” Apr. 2025.
6. [6] P. Suganthi et al., “Employee Attrition Prediction: A Machine Learning Approach,” IJAIA, 2024.

# Background and Motivation

Employee attrition not only leads to financial implications due to rehiring and training but also results in the loss of institutional knowledge and disrupts team cohesion. A high attrition rate can indicate deeper organizational issues such as lack of career progression, dissatisfaction, or poor leadership. Thus, early detection is crucial to intervene and address potential concerns. Modern HR analytics powered by machine learning provides the tools to uncover hidden trends and patterns that are not evident through traditional means.

# Feature Importance Analysis

The Random Forest classifier provided insights into feature significance. Among the top features were OverTime, MonthlyIncome, YearsAtCompany, and JobRole. Employees who regularly worked overtime or had lower income levels showed higher attrition risk. The model's feature importance scores helped visualize the influence of each attribute, allowing HR teams to prioritize interventions.

# Future Work

Future enhancements could include integrating natural language processing (NLP) to analyze employee feedback and surveys. Incorporating real-time data from HR systems can improve prediction timeliness. Furthermore, implementing explainable AI dashboards can empower HR managers to understand individual predictions and strategize accordingly.

# Appendix A: Model Performance Visualization

Figure 1. Accuracy and F1-score comparison of Random Forest, SVM, and KNN models.  
[Insert Graph Here]

# VII. Data Preprocessing Techniques

Preprocessing is a crucial step in any machine learning pipeline. The dataset was carefully cleaned and transformed to ensure optimal performance of the classifiers. Initially, redundant columns such as EmployeeCount and StandardHours were removed. Categorical features including JobRole, Department, and MaritalStatus were encoded using LabelEncoder. Missing values were handled by checking for NaNs and using imputation techniques where necessary. We normalized numerical features to ensure comparability, especially for distance-based models such as KNN (included for comparison).

# VIII. Cross-Validation and Model Tuning

To prevent overfitting and ensure robust model performance, 5-fold cross-validation was performed on the training set. Hyperparameter tuning was conducted using GridSearchCV. For Random Forest, parameters like n\_estimators (number of trees) and max\_depth (maximum tree depth) were tuned. For SVM, the penalty parameter C and the kernel coefficient gamma were optimized. KNN was also evaluated as a baseline, with the number of neighbors k varied from 3 to 15. The final models were selected based on validation performance.

# IX. Comparative Analysis of Models

The Random Forest classifier consistently outperformed other models with respect to accuracy, F1-score, and robustness to noise. SVM showed good classification boundaries in high-dimensional space but struggled slightly with imbalanced data before SMOTE application. KNN had the lowest performance due to its sensitivity to feature scaling and class imbalance. Table 1 summarizes the evaluation metrics for all models.

Table 1. Comparison of classification model performance metrics (accuracy, precision, recall, F1-score).

[Insert Table Here]

# X. Ethical Considerations in Predictive HR Analytics

The use of machine learning in HR processes raises ethical concerns, including data privacy, transparency, and potential bias. While algorithms can uncover hidden patterns, their predictions should not be used as the sole criterion for employment decisions. HR professionals must interpret model outputs within the organizational context and ensure compliance with labor laws and ethical standards. Techniques like SHAP and LIME can be used to provide transparency, making it easier to justify decisions based on model predictions.