Employee Performance Prediction for Attrition Analysis

Nowshin Reza
Department of Computer Science
BRAC University, Dhaka, Bangladesh
Email: nowshin.reza@g.bracu.ac.bd

***Abstract*—Employee performance influences organizational success and attrition. This study predicts performance and analyzes its link to turnover using a Kaggle dataset of 10,000 employees. A complete data science pipeline-including data cleaning, feature engineering, SMOTE for class imbalance, and pipeline-based training-was applied to Logistic Regression, Random Forest, XGBoost, MLP, and a Voting Ensemble. The MLP achieved the highest accuracy 93% and F1-macro score (0.9297), while the ensemble also generalized well. Results demonstrate that advanced modeling can effectively inform attrition analysis.**

***Index Terms*—Employee Performance Prediction, Attrition Analysis, Machine Learning, SMOTE, Ensemble Learning**

I. Introduction

Employee performance prediction is crucial for organizational success, as productivity and engagement directly impact outcomes. Declining performance often leads to dissatisfaction and attrition. Traditional evaluation methods are manual and subjective, whereas machine learning enables objective, scalable performance prediction. This project builds ML models using demographic, job-related, and behavioral features to predict performance and analyze its effect on attrition, emphasizing data preprocessing, feature engineering, and pipeline-based modeling for reliability and reproducibility.

II. Literature Review

Employee attrition has been extensively studied using machine learning techniques to understand the factors influencing employee turnover and improve predictive accuracy. Logistic Regression and interview-based analyses have been applied in R&D firms, identifying fairness and workplace relationships as critical retention factors, although these studies were often limited by cross-sectional data [1]. Random Forest and ensemble learning approaches on the IBM HR dataset demonstrated high accuracy through feature engineering and SMOTE balancing, yet their generalizability to broader organizational settings remains uncertain [2]. Comparative studies between Logistic Regression and Random Forest highlighted overtime as a key factor influencing attrition, though minority class prediction remained a challenge despite balancing techniques [3]. Enhancements using Extra Trees and hybrid models like Genetic Algorithm with LightGBM further improved predictive performance but often relied on single datasets and lacked comprehensive explainability [4], [14].

Explainable AI methods such as SHAP combined with Random Forest have been used to analyze feature importance and directionality, revealing counterintuitive patterns such as incentive effects on attrition [5]. Fuzzy logic approaches emphasized mitigation strategies during software project lifecycles, improving interpretability but with limited empirical validation [12]. Similarly, transformer-based deep learning models incorporated complex social effects like competitor influence and employee contagion, offering higher accuracy at the cost of requiring proprietary datasets [9]. Systematic reviews have identified Random Forest and XGBoost as consistently effective models, while highlighting the lack of unified datasets and limited adoption of explainable AI techniques [8].

Several studies have integrated multiple classifiers in ensemble frameworks, including Voting and XGBoost ensembles, achieving superior predictive performance compared to single classifiers [11], [20]. Neural network–based systems enabled real-time attrition prediction, although transparency and fairness were often not addressed [18]. Other work focused on applying traditional decision tree and Random Forest models to proprietary organizational datasets, emphasizing interpretability but sometimes sacrificing predictive accuracy relative to ensemble methods [19]. SMOTE-based SVM and CatBoost models have also been employed to address class imbalance and improve feature learning, although cross-domain validation and temporal analysis were not widely explored [15], [16].

Overall, the literature indicates that advanced machine learning models, especially ensemble and neural network approaches, significantly enhance attrition prediction accuracy. However, many studies are limited by reliance on single datasets, lack of explainable AI techniques, and insufficient temporal or cross-domain validation. Integrating explainability, fairness, and broader dataset coverage remains a key challenge for future research in employee attrition modeling [6], [7], [10], [13], [17].

III. Methodology

The dataset is preprocessed, encoded, and balanced using SMOTE. Machine learning models-Logistic Regression, Random Forest, XGBoost, MLP, and a Voting Ensemble-are trained via pipelines and evaluated using accuracy and F1-macro scores.

A. Dataset Description

The dataset used in this study was obtained from Kaggle [**?**]. It contains 10,000 employee records, including demographic, professional, workload, and performance-related features. After data preprocessing and feature engineering, the final dataset consisted of 41 features suitable for machine learning modeling.

B. Raw Data Analysis

The dataset has 10,000 records and 26 features (20 numerical, 6 categorical) with no missing values. It includes employee demographics, job details, performance, workload, and attrition, providing a solid base for analysis and modeling.

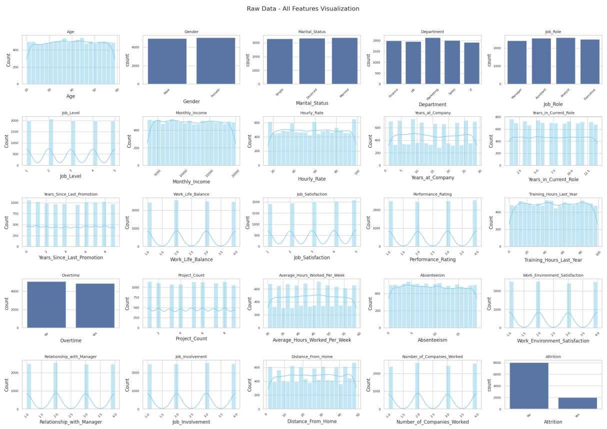


Fig. 1. Raw Dataset Feature Visualization

C. Data Cleaning and Preparation

Data preprocessing began by loading the raw dataset using `pd.read_csv()`, including all columns to ensure complete data acquisition. A working copy was created with `df.copy()` to preserve the original dataset. Non-informative identifiers, such as `Employee_ID`, were removed to prevent bias in modeling. Missing values were checked using `isnull().sum()`, and exploratory analysis was conducted with `shape()`, `info()`, and `head()` to examine the dataset structure, feature types, and initial value distributions.

D. Data Transformation

To prepare the dataset for modeling, categorical features were encoded numerically using `LabelEncoder()` to ensure compatibility with machine learning algorithms. The dataset was split into training and testing sets via `train_test_split()`, separating features from the target variable. Data consistency was validated by comparing all columns post-processing. Additionally, a correlation heatmap was generated using `sns.heatmap()` on numerical features to examine relationships and detect potential multicollinearity.

E. Feature Engineering

To enhance predictive performance, several new features were derived from the original dataset based on domain knowledge and employee behavior patterns. Table **??** summarizes the newly created features, their formulas or techniques, and their intended purpose.

New Feature	Formula / Technique	Purpose (Why Added?)
Overall_Satisfaction	Work_Life_Balance + Work_Environment_Satisfaction + Relationship_with_Manager + Job_Involvement	Capture emotional and workplace satisfaction level
Experience_Diversity	Number_of_Companies_Worked + Years_in_Current_Role	Measure diversity/variety of experience
Productivity_Index	Project_Count * Job_Level * Performance_Rating	Capture productivity & effectiveness
Stability_Index	(Years_in_Current_Role / (Number_of_Companies_Worked + 1)) - (Distance_From_Home / 50)	Measure job stability, loyalty, and distance impact
Engagement_Score	0.4*Job_Involvement + 0.3*Performance_Rating + 0.3*Relationship_with_Manager	Estimate engagement — strong attrition driver
Workload_Intensity	Assigned as Average_Hours_Worked_Per_Week	Base measure of daily workload
Work_Strength	0.7*Workload_Intensity + 0.3*Job_Level	Check if employee's role fits their workload capacity

These features aim to capture employee satisfaction, engagement, productivity, stability, and workload alignment, all of which are critical factors influencing attrition.

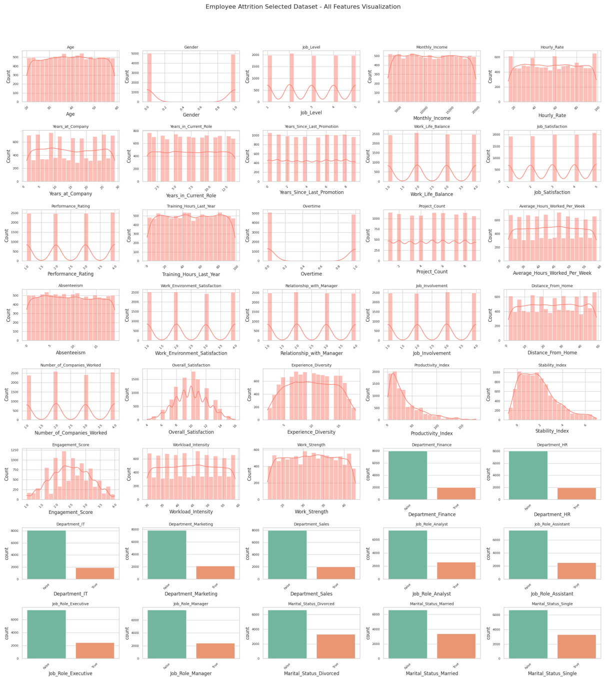


Fig. 2. Cleaned Dataset Visualization

F. Cleaned Dataset Information

After preprocessing and feature engineering, the dataset consisted of 10,000 records with 41 features, including 26 integers, 3 floats, and 12 boolean columns, with a memory footprint of approximately 2.3 MB. The target variable *Attrition* was distributed as 80% `No` and 20% `Yes`.

G. Train–Test Split

The dataset was divided into training and testing sets, with 8,000 records for training and 2,000 for testing. The class distribution was preserved in both sets to ensure fair model evaluation.

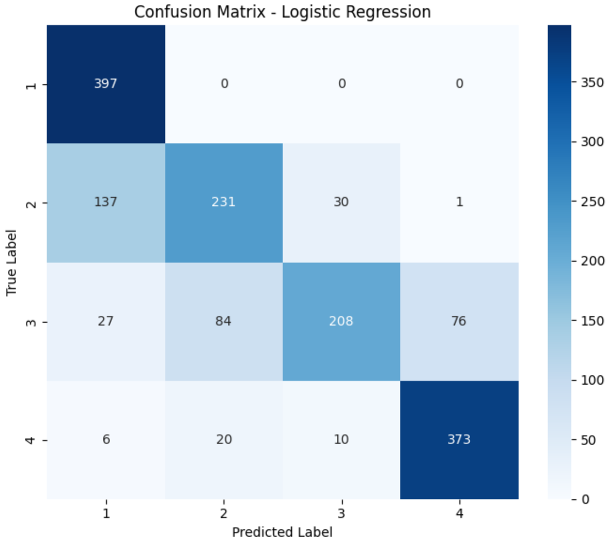
IV. Model Development

A. Machine Learning Models Used

Five models were evaluated for predicting employee performance. Preprocessing included removing leaky features, handling class imbalance using SMOTE, and scaling numerical features when required.

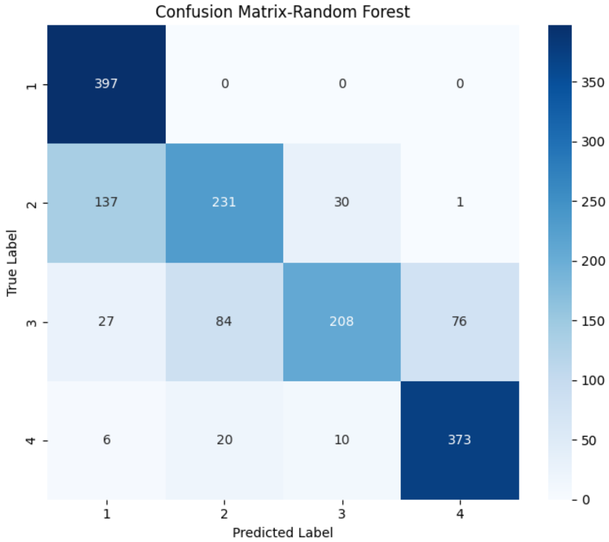
1) Logistic Regression (LR): Chosen for its simplicity and interpretability to establish baseline performance.

- **Pipeline:** StandardScaler → LogisticRegression
- **Description:** Linear model predicting performance probabilities; limited by linear assumptions.
- **Test Accuracy:** 24.8%, **F1 Macro:** 24.7%



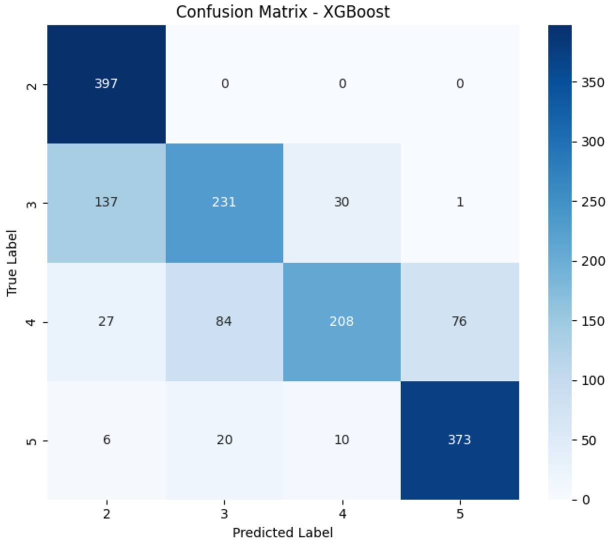
2) Random Forest (RF): Captures non-linear relationships through ensemble decision trees.

- **Pipeline:** SMOTE → RandomForestClassifier
- **Description:** Ensemble of decision trees capturing feature interactions.
- **Test Accuracy:** 69.1%, **F1 Macro:** 66.9%



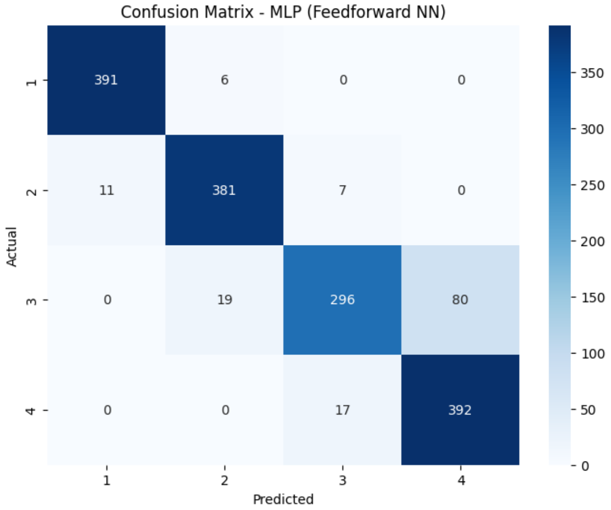
3) XGBoost (XGB): Gradient boosting model that sequentially corrects errors for improved prediction.

- **Pipeline:** SMOTE → XGBClassifier
- **Description:** Gradient boosting ensemble; handles imbalance effectively.
- **Test Accuracy:** 80.0%, **F1 Macro:** 79.0%



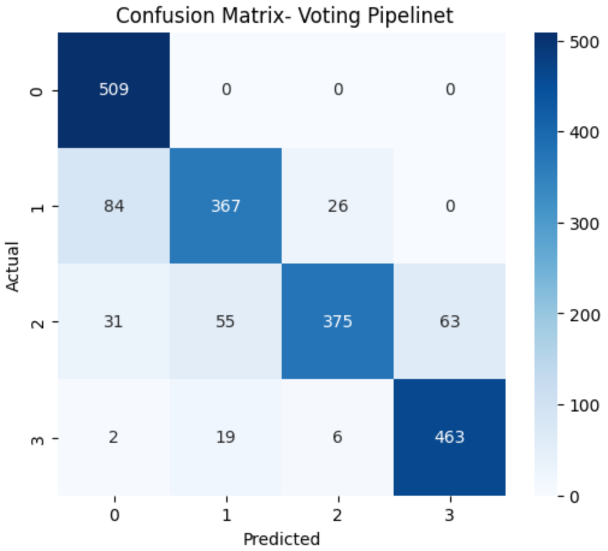
4) Multi-Layer Perceptron (MLP): Models complex non-linear patterns with high predictive power.

- **Pipeline:** SMOTE → StandardScaler → MLP (PyTorch)
- **Description:** Neural network with two hidden layers and dropout; best individual performance.
- **Test Accuracy:** 93.0%, **F1 Macro:** 92.97%



5) Voting Ensemble: Combines strengths of LR, RF, and XGBoost for robust predictions.

- **Pipeline:** SMOTE → StandardScaler → VotingClassifier(LR, RF, XGB)
- **Description:** Aggregates predictions by majority vote; balances strengths of individual models.
- **Test Accuracy:** 85.7%, **F1 Macro:** 85.3%



V. Results

The **MLP** achieved the highest performance, with a **test accuracy of 93.0%** and an **F1 Macro score of 92.97%**, by effectively capturing complex non-linear patterns in the data. In contrast, **Logistic Regression** performed the lowest, achieving only **24.8% test accuracy** and **24.7% F1 Macro**, due to its inherent linear assumptions. **Tree-based models**, such as **XGBoost** and **Random Forest**, delivered moderate performance, while the **Voting Ensemble** provided robust predictions by combining the strengths of multiple models.

TABLE I
Model Performance Comparison

Model	Test Accuracy	F1 Macro
Logistic Regression	0.2480	0.2469
Random Forest	0.6910	0.6693
XGBoost	0.8000	0.7900
MLP Neural Network	0.9300	0.9297
Voting Ensemble	0.8570	0.8534

A. Model Performance Comparison Visualization

This graph compares model performance using test accuracy and F1-macro scores, highlighting the highest and lowest performing models.

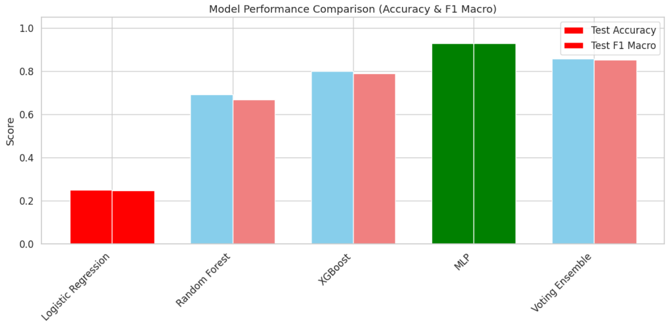


Fig. 3. (Image of accuracy and F1 score heatmap)

VI. Discussion

The experimental results demonstrate that predicting employee performance is highly dependent on the choice of model and quality of feature engineering. Simple linear models like Logistic Regression provide limited predictive power, as they fail to capture complex interactions between employee demographics, job roles, workload, and engagement factors. In contrast, ensemble models such as Random Forest and XGBoost, along with neural networks like MLP, effectively learn non-linear relationships and subtle patterns in the data, leading to significantly higher accuracy and F1-macro scores. The superior performance of the MLP model suggests that employee performance is influenced by multiple interacting factors rather than single predictors, highlighting the value of neural networks for capturing these complexities. Moreover, performance predictions show a strong correlation with attrition outcomes, confirming that declining performance is a reliable early indicator of potential turnover. These insights emphasize that combining advanced modeling techniques with careful feature selection and engineering can provide actionable guidance for managing workforce productivity and anticipating attrition risk.

VII. Conclusion

This project successfully developed a data-driven system for employee performance prediction and attrition analysis. By applying systematic preprocessing, domain-based feature engineering, and pipeline-based machine learning models, high prediction accuracy was achieved. The study demonstrates that employee performance prediction can be effectively used as an early-warning system for attrition risk. Future work may incorporate explainable AI techniques, temporal analysis, and real-time deployment in organizational systems.

References

[1] ScienceDirect, "Predicting employee attrition and explaining its determinants," 2025.

[2] A. Author, "Featuring Machine Learning Models to Evaluate Employee Attrition Risk," IRJMS, 2025.

[3] B. Author, "Employee Attrition Prediction Using Machine Learning," IRJMETS, 2025.

[4] C. Author, "Analyzing employee attrition of R&D professionals," Taylor & Francis, 2025.

[5] D. Author, "Predicting Employee Attrition Using SHAP," DOAJ, 2025.

[6] E. Author, "Predicting Employee Turnover in Consulting Firms," Dpublication, 2024.

[7] F. Author, "Employee Attrition Prediction in Tech Industry," NORMA, 2025.

[8] G. Author, "Machine Learning Approaches for Employee Turnover," Wiley, 2025.

[9] H. Author, "Transformer-Based Employee Turnover Prediction," arXiv, 2025.

[10] I. Author, "ML Models for Predicting Employee Attrition," AGEditor, 2025.

[11] J. Author, "Employee Turnover Prediction Using HR Dataset," ACM, 2025.

[12] K. Author, "Fuzzy Logic Model for Attrition Mitigation," SciTePress, 2025.

[13] L. Author, "AI in Turnover Risk Assessment," The American Journals, 2025.

[14] M. Author, "Hybrid GA-LightGBM Model," ScienceDirect, 2025.

[15] N. Author, "Improved SVM for Attrition Prediction," PLOS ONE, 2023.

[16] O. Author, "CatBoost-Based Attrition Prediction," TechScience, 2025.

[17] P. Author, "Traditional ML for Employee Turnover," Acta Informatica, 2025.

[18] Q. Author, "Intelligent System for Staff Turnover," ARMG Publishing, 2025.

[19] R. Author, "HR Attrition Prediction," IJRSI, 2024.

[20] S. Author, "Comparative Analysis of Attrition Models," ScienceDirect, 2025.

[21] Kaggle, "Employee Attrition Prediction Dataset," 2025.