# PROCEEDINGS OF
# INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING


DEPARTMENT OF COMPUTER SCIENCE AI AND ML
BHARATA MATA COLLEGE (AUTONOMOUS)
THRIKKAKARA



Co-sponsored by
Anusandhan National Research Foundation (ANRF)



Editors
Dr. John T Abraham, Mr. Harikrishnan P

# Proceedings of

# International Conference on Artificial Intelligence & Machine Learning

Department of Computer Science

Bharata Mata College (Autonomous) Thrikkakara

Co-Sponsored by

Anusandhan National Research Foundation (ANRF)

Editors

Dr John T Abraham Mr Harikrishnan P

Published by: Department of Computer Science
Bharata Mata College Thrikkakara

Editors: **Dr John T Abraham**
HoD & Assistant Professor
Department of Computer Science

**Mr. Harikrishnan P**
Assistant Professor
Department of Computer Science

ISBN: **978-81-973274-9-0**

Month and Year: November 2024

**Disclaimer**

# Predicting Employee Performance Levels Using Machine Learning Algorithms: Enhancing HR Decision-Making through AI

**Dr.Meenu Suresh**[1]
*Assistant Professor, Dept.of Computer Science and IT,*
*Jain University (Deemed-to-be University),*
*Knowledge Park, Kakkanad, Kochi, Kerala 682042*

**Tonny Binoy**[2]
*Assistant Professor, Department of CA & AI,*
*Saintgits College of Applied Sciences,*
*Pathamuttom P. O, Kottayam, Kerala 686532*

**Punnya Sudhakaran**[3]
*Assistant Professor, Dept.of Business Administration,*
*Saintgits College of Applied Sciences,*
*Pathamuttom P. O, Kottayam, Kerala 686532*

**Akhil P Shaji**[4]
*Department of CA & AI*
*Saintgits College of Applied Sciences,*
*Pathamuttom P. O, Kottayam, Kerala 686532*

**Fathima Shemim KS**[5]
*Faculty,Computer Science Department ADVETI ,*
*United Arab Emirates Research Scholar*
*University of Bolton United Kingdom*

**Saritha M S**[6]
*Assistant Professor*
*Department of Artificial Intelligence & Data Science,*
*St. Joseph's College of Engineering & Technology,*
*Palai, Choondacherry P.O, Kottayam 686579*

*Abstract*-**This study presents a machine learning framework to predict employee performance levels, empowering HR professionals with data-driven insights for effective talent management. Leveraging a comprehensive dataset encompassing demographics, job roles, engagement metrics, training history, and historical performance ratings, the research explores multiple algorithms, including LightGBM, XGBoost, XGBoost with SMOTE, and Random Forest. To address class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was implemented, generating synthetic samples to enhance prediction accuracy across all classes. Feature selection and importance analysis identified key performance predictors, such as tenure, engagement scores, work-life balance, and satisfaction levels. Among the evaluated models, Random Forest achieved the highest accuracy (94%) with balanced class performance, making it the preferred choice for deployment. This research underscores the transformative role of machine learning in HR practices, providing actionable insights to design targeted development programs, optimize employee performance, and improve organizational outcomes.**

**Keywords: HR professionals, SMOTE, random forest, optimization**

## I. INTRODUCTION

In today's highly competitive business environment, organizations continuously seek innovative methods to enhance productivity and performance. One critical aspect of achieving these goals is the ability to accurately forecast employee performance. Traditional methods of performance evaluation often rely on subjective assessments, which can lead to inconsistencies or biases, undermining their reliability. To overcome these challenges, the integration of business analytics and machine learning has emerged as a powerful solution [1] .

This paper [2] explores using machine learning, specifically Generalized Linear Models, to predict job involvement based on demographic and work-related factors using IBM Watson Analytics data. The findings enable HR departments to enhance employee engagement and retention strategies. Future research suggests incorporating larger datasets and psychological factors for improved predictions. The paper [3] compares machine learning models, including Logistic Regression and Random Forest, to predict employee attrition using the IBM HR dataset. Logistic Regression achieved the best performance with 88% accuracy and 85% AUC-ROC. Key attrition factors identified include low job involvement, long commutes, and frequent travel requirements. This study [4] uses machine learning to predict employee attrition, with the Extra Trees Classifier achieving 93% accuracy. Key factors identified include monthly income,

ISBN:978-81-973274-9-0 Department of Computer Science Bharata Mata College Thrikkakara

hourly rate, job level, and age. The findings help organizations address attrition by improving critical job-related factors. The proposed machine learning framework predicts employee performance in hiring and appraisals using decision trees and classification algorithms [5]. It combines historical evaluations, employee attributes, and social media data to enhance accuracy and reduce hiring errors.

The study [6] emphasizes skill-specific performance metrics and suggests advanced ML techniques for future improvements. The study compares Naïve Bayes and C4.5 algorithms for predicting employee lateness, finding Naïve Bayes more accurate (83.33%) than C4.5 (72.22%). It identifies "going late to bed" as the primary factor influencing tardiness. The paper [7] highlights the utility of data mining for addressing workplace punctuality issues. LightGBM is an efficient implementation of Gradient Boosting Decision Trees (GBDT), leveraging Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) to improve scalability and training speed. GOSS focuses on high-gradient instances, while EFB reduces feature dimensionality by bundling exclusive features. Experiments show LightGBM achieves similar accuracy to traditional GBDTs like XGBoost but trains up to 20 times faster on large datasets. The paper [8] compares XGBoost, LightGBM, and CatBoost for gradient boosting, assessing accuracy, AUC, and training speed. CatBoost excels in accuracy and AUC, LightGBM in speed, while XGBoost balances performance. It also highlights the impact of hyper-parameter tuning, with CatBoost's defaults often performing best. The author [9] addresses class imbalance challenges using SMOTE to enhance machine learning model performance in predicting employee attributes. It evaluates algorithms like LightGBM, XGBoost, and Random Forest, identifying Random Forest as the most accurate and interpretable model for balanced predictions. The study emphasizes actionable insights for HR optimization. This paper [10] investigates the application of Random Forest for predictive modeling, highlighting its superior accuracy, scalability, and ability to handle high-dimensional data. It underscores the algorithm's effectiveness in feature importance analysis, aiding in reliable and interpretable decision-making processes.

To address class imbalances in employee performance datasets, SMOTE was applied to generate synthetic samples, enabling balanced classification across performance levels in the proposed model. Machine learning models, including LightGBM, XGBoost, and Random Forest, were evaluated, with Random Forest achieving 94% accuracy and proving the most reliable. Feature analysis identified tenure, engagement scores, work-life balance, and satisfaction as key predictors of performance. This research demonstrates the potential of machine learning to enhance HR processes, enabling targeted development programs and fostering organizational effectiveness.

The paper is organized as follows: Section II introduces the working of the model. Section III presents the model output and classification report interpretation, while Section IV displays the visualization, Section V concludes the study and Section VI states the future enhancements.

## II. WORKING OF THE MODEL

The primary goal of our model is to predict the performance score of employees based on a variety of known characteristics (features) from their profiles. Performance Score is a critical metric that helps organizations assess employee contributions, but it's often unknown for new hires or employees undergoing role transitions. By building a machine learning model, we can estimate this score using available data on employee attributes, allowing HR to make proactive and informed decisions.

### *A. Dataset Description*

Our proposed model is trained on a comprehensive dataset containing historical records of employees. This dataset includes:

- Employee Characteristics: Job role, department, tenure, and demographic details.
- Engagement Metrics: Satisfaction scores, engagement levels, and feedback.
- Job-Specific Metrics: Training outcomes, certifications, or promotions.

During training, the model examines how these attributes correlate with the known performance scores. It learns the patterns and relationships that tend to lead to high or low performance, such as:

- The impact of job role or experience on productivity.
- The effect of engagement or satisfaction on overall performance.

Through training, the model essentially "learns" what a high-performing or low-performing profile looks like based on the data.

### *B. Model Architecture*

The model we developed for predicting employee performance is functioning as a classification model. In this HR model, we are predicting performance levels, which are typically represented as distinct classes or categories. For example, employee performance might be classified into categories like:

- High Performer
- Average Performer
- Low Performer

Each of these categories is a class label, making this a multi-class classification problem. The goal of the model is to classify each employee into one of these predefined classes based on various features (like engagement scores, job role, tenure, etc.).

If the performance score column has values such as 0, 1, 2, and 3, these are effectively classes in the classification model. Each number corresponds to a distinct performance level, even if these levels aren't labeled with descriptive names. For instance:

- 0 might represent one level (e.g., "Low Performance"),
- 1 another level (e.g., "Below Average"),
- 2 a higher level (e.g., "Average"),
- 3 the highest level (e.g., "High Performance").

Even without explicit labels, the Random Forest model will learn to classify instances into these numeric levels based on the patterns in the features.

*C. Work Flow*

*1. Data Loading and Target Definition*

The code loads the dataset and defines the target variable. Features (X) are all columns except the target, while target (y) is the Performance Score.

*2. Encoding Categorical Features*

Categorical features are label-encoded to convert them into a numerical format that the Random Forest model can understand.

*3. Handling Class Imbalance with SMOTE*

SMOTE (Synthetic Minority Over-sampling Technique) is used to address class imbalance by generating synthetic samples for underrepresented classes. This is essential for ensuring that the model doesn't become biased toward majority classes. After applying SMOTE, the data is split for the training purpose.

*4. Train-Test Split*

The code splits the resampled data into training and testing sets.

*5. Random Forest Model and Hyperparameter Tuning*

A Random Forest Classifier is initialized, and GridSearchCV is used for hyperparameter tuning. GridSearchCV finds the best parameters by evaluating model performance using cross-validation on the training set. The code defines a parameter grid to try different values for:

- n_estimators: Number of trees.
- max_depth: Maximum depth of each tree.
- min_samples_split: Minimum samples needed to split a node.
- class_weight: Balances classes or not.

## III. MODEL OUTPUT AND CLASSIFICATION REPORT INTERPRETATION

The proposed model is evaluated using classification-specific metrics such as accuracy, precision, recall, F1-score, and the confusion matrix. Table 1 shows the accuracy rate of different models along with the proposed model. These metrics are tailored to classification tasks and helps to assess how well the model performs across different performance classes 0, 1, 2, and 3. The interpretation of the output are:

- The precision, recall, and F1-score metrics are provided for each numeric class (0, 1, 2, 3), indicating how well the model predicts each level of performance.
- Overall accuracy is high (93.81%), which shows that the model is quite accurate in classifying employees into these numeric performance levels.

Table.1 ACCURACY OF DIFFERENT MODELS

| Model | Accuracy |
|---|---|
| LightGBM Model<br>Random Forest Model without SMOTE | 93.70<br>78.67 |
| Random Forest Model with SMOTE (Proposed) | 93.81 |
| XGBoost Model with SMOTE and Class Weighting | 93.75 |
| XGBoost Model without SMOTE | 78.67 |

## IV. VISUALIZATION

This section presents the key visualizations generated during the analysis of the Random Forest model. These visualizations provide insights into model performance, feature importance, and evaluation metrics.

*A. Confusion Matrix*

A confusion matrix is plotted to visualize how well the model performs on each class, showing where it makes correct and incorrect predictions and is depicted in Fig 1

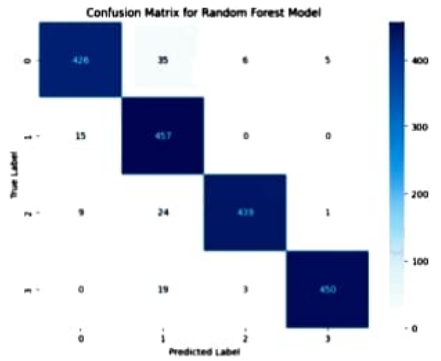


Fig 1. Confusion Matrix

### *B. Cross-Validation Accuracy*

The Fig 2 demonstrates the cross-validation accuracy scores for the Random Forest model. A boxplot summarizes the distribution of accuracy scores across multiple validation folds, with the majority of scores tightly concentrated near the upper bound of the accuracy range. However, an outlier with lower accuracy is noticeable, which indicates potential variations in model performance on specific folds.

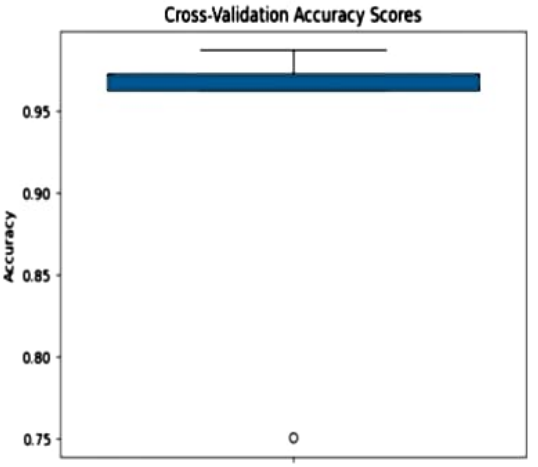


Fig 2. Cross-Validation Accuracy

### *C. Classification Report Heatmap*

The classification report heatmap (Fig 3) provides a comprehensive summary of precision, recall, and F1-scores for each class. The model achieves high performance metrics across all classes, with Class 0 achieving precision of 0.95 and recall of 0.90, and Class 3 achieving a precision of 0.99. These metrics confirm the model's strong generalization ability across diverse classes.

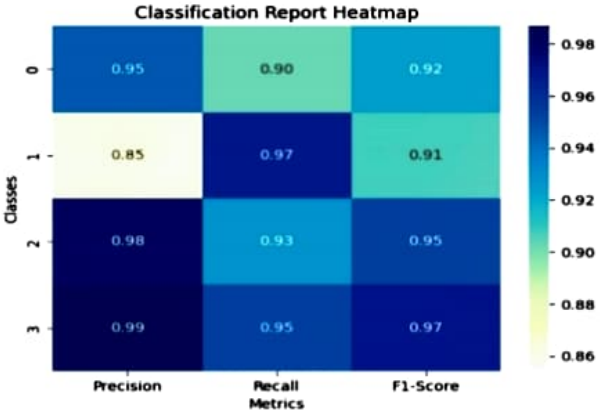


Fig 3. Classification Report Heatmap

### *D. Multi-Class ROC Curve*

The Fig 4 presents the multi-class Receiver Operating Characteristic (ROC) curves along with the Area Under the Curve (AUC) for each class. All classes show AUC values close to 1.0, indicating excellent discriminative power. Classes 2 and 3 achieve perfect AUC scores of 1.0, further reinforcing the model's strong classification performance.

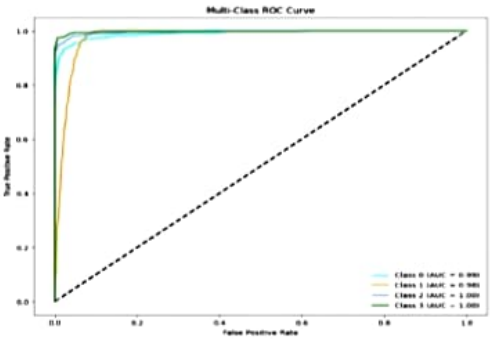


Fig 4. Multi-Class ROC Curve

### *E. Feature Importance*

The Fig 5 ranks the top 10 most important features identified by the Random Forest model. The feature "Tenure (Years)" emerges as the most influential, followed by "Age" and "Satisfaction Score." These features significantly contribute to the model's predictive power, providing actionable insights into the key determinants of the target variable.

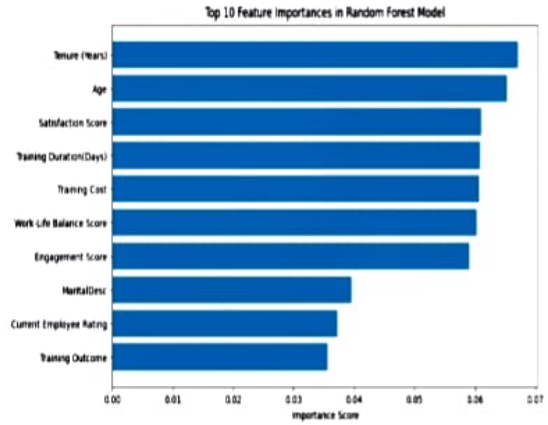


Fig 5. Feature Importance

## V. CONCLUSION

A machine learning-based framework for predicting employee performance levels using historical HR data is developed. The Random Forest model, enhanced with the Synthetic Minority Oversampling Technique (SMOTE), emerged as the most effective model, achieving an impressive accuracy of 93.81% and consistently high performance across precision, recall, and F1-scores for all classes. This result highlights the model's robustness and ability to handle imbalanced data effectively, making it a valuable tool for HR decision-making. By leveraging this model, organizations can make proactive, data-driven decisions to enhance employee development, improve engagement, and optimize workforce planning.

## VI. FUTURE ENHANCEMENTS

While the current model demonstrates strong predictive performance, there are several opportunities for improvement and expansion:

### *A. Incorporation of Real-Time Data:*

Integrating real-time performance data, such as ongoing project metrics and employee feedback, could enhance the model's accuracy and responsiveness.

### *B. Exploring Advanced Algorithms:*

Future studies could explore ensemble methods like XGBoost with hyperparameter optimization or deep learning models to capture more intricate patterns in large and complex datasets.

### *C. Broader Feature Set:*

Incorporating additional features, such as external factors (e.g., market trends or economic conditions) and employee sentiment analysis, to improve predictive accuracy further.

### *D. Deployment and Real-World Testing:*

Implementing the model in a real-world HR system and evaluating its performance on unseen data streams to refine its robustness and practical applicability.

### *E. Cross-Organizational Application:*

Generalizing the model to work across multiple organizations or industries by developing a standardized framework adaptable to various datasets and HR structures.

## REFERENCES

[1] A. Tambde and D. Motwani, "Employee Churn Rate Prediction and Performance Using Machine Learning," vol. 8, no. 2, 2019.

[2] Y. Choi and J. W. Choi, "A study of job involvement prediction using machine learning technique," *IJOA*, vol. 29, no. 3, pp. 788–800, May 2021, doi: 10.1108/IJOA-05-2020-2222.

[3] F. Guerranti and G. M. Dimitri, "A Comparison of Machine Learning Approaches for Predicting Employee Attrition," *Applied Sciences*, vol. 13, no. 1, p. 267, Dec. 2022, doi: 10.3390/app13010267.

[4] A. Raza, K. Munir, M. Almutairi, F. Younas, and M. M. S. Fareed, "Predicting Employee Attrition Using Machine Learning Approaches," *Applied Sciences*, vol. 12, no. 13, p. 6424, Jun. 2022, doi: 10.3390/app12136424.

[5] A. A. Mahmoud, T. Al Shawabkeh, W. A. Salameh, and I. Al Amro, "Performance Predicting in Hiring Process and Performance Appraisals Using Machine Learning," in *2019 10th International Conference on Information and Communication Systems (ICICS)*, Irbid, Jordan: IEEE, Jun. 2019, pp. 110–115. doi: 10.1109/IACS.2019.8809154.

[6] J. L. D. Mercara, "Prediction of Employees' Lateness Determinants using Machine Learning Algorithms," *IJATCSE*, vol. 9, no. 1, pp. 779–783, Feb. 2020, doi: 10.30534/ijatcse/2020/111912020.

[7] G. Ke *et al.*, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree".

[8] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A comparative analysis of gradient boosting algorithms," *Artif Intell Rev*, vol. 54, no. 3, pp. 1937–1967, Mar. 2021, doi: 10.1007/s10462-020-09896-5.

[9] A. Saad Hussein, T. Li, C. W. Yohannese, and K. Bashir, "A-SMOTE: A New Preprocessing Approach for Highly Imbalanced Datasets by Improving SMOTE:," *IJCIS*, vol. 12, no. 2, p. 1412, 2019, doi: 10.2991/ijcis.d.191114.002.

[10] A. B. Wild Ali, "Prediction of Employee Turn Over Using Random Forest Classifier with Intensive Optimized Pca Algorithm," *Wireless Pers Commun*, vol. 119, no. 4, pp. 3365–3382, Aug. 2021, doi: 10.1007/s11277-021-08408-0.