Employee Turnover Prediction: Enhancing Workforce Stability through Machine Learning

*

1st Sonal Kasana
Computer Science & Engineering Department
Thapar Institute of Engineering & Technology
Patiala, Punjab
skasana_be22@thapar.edu

2nd Garima Singh
Computer Science & Engineering Department
Thapar Institute of Engineering & Technology
Patiala, Punjab
garima.singh@thapar.edu

***Abstract*—Employee turnover poses significant challenges for organizations, leading to high costs in recruitment, training, and productivity loss. This study applies machine learning (ML) techniques to predict employee attrition, identify the key factors influencing it, and develop data-driven strategies for employee retention. The research process begins with improving data quality by handling missing values, followed by exploratory data analysis (EDA) to highlight the major drivers of turnover. Clustering methods are then used to group employees based on satisfaction and performance levels, providing deeper insights into the characteristics of those likely to leave. To address the imbalance between employees who stay and those who resign, the Synthetic Minority Over-sampling Technique (SMOTE) is employed. Multiple ML models are trained and evaluated using k-fold cross-validation, with the best model selected based on accuracy, precision, recall, and F1-score. Finally, the study offers actionable recommendations for targeted retention strategies aimed at high-risk employees. The findings demonstrate the effectiveness of ML in supporting HR decision-making and promoting workforce stability.**

***Index Terms*—Employee Turnover, Attrition Prediction, Retention Strategies**

I. Introduction

Employees are a company's most valuable resource, and retaining them is critical for long-term success. Yet, employee turnover—whether due to career growth, relocation, retirement, or dissatisfaction—remains unavoidable and costly. When skilled workers leave, organizations face not only recruitment and training expenses but also reduced productivity, disrupted workflows, and loss of expertise. These challenges have intensified in today's global, post-pandemic job market, where opportunities are more accessible than ever [1].

To address this, companies need better ways to understand why employees leave and how to retain them. Machine learning (ML) provides a powerful solution by analyzing employee data to uncover hidden patterns, predict who is likely to leave, and guide proactive retention strategies. Unlike traditional methods, ML can consider multiple factors such as job satisfaction, performance, tenure, promotions, and salary trends to deliver accurate forecasts of attrition [2] [3].

This research explores how machine learning can be applied to predict employee turnover and support HR in designing data-driven retention strategies. The study begins with ensuring data quality, as accurate predictions depend heavily on reliable data. Missing values are identified and treated, and the dataset is prepared for analysis. Next, an Exploratory Data Analysis (EDA) is conducted to uncover the primary factors influencing employee attrition. EDA helps visualize trends and relationships—for example, whether employees with longer working hours are more likely to leave, or whether promotion delays increase the risk of turnover. To better understand the characteristics of employees who leave, clustering techniques are applied. These group employees based on common attributes such as satisfaction levels and performance scores, offering deeper insights into patterns of attrition. One common challenge in turnover prediction is class imbalance—the fact that the number of employees who stay is often much larger than those who leave. This imbalance can make it difficult for models to learn accurately. To address this, the Synthetic Minority Over-sampling Technique (SMOTE) is used. SMOTE balances the dataset by generating synthetic examples of the minority class (employees who leave), ensuring that models can learn from both categories more effectively. A range of ML models is then trained and evaluated using k-fold cross-validation, a method that ensures the results are not biased by the way the data is split. The performance of these models is measured using metrics such as accuracy, precision, recall, and F1-score. The model with the best overall performance is selected for making predictions.

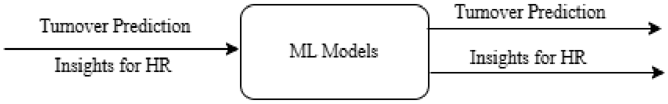

Fig. 1. Research Need and Impact.

This figure 1 shows how machine learning (ML) models help organizations deal with employee turnover. On one side are the needs—predicting which employees may leave and

Identify applicable funding agency here. If none, delete this.

providing useful insights for HR. The ML models process employee data to address these needs, and on the other side, they deliver outcomes: clearer turnover predictions and actionable insights that HR can use to improve retention strategies.

In summary, this study demonstrates how ML can transform HR decision-making by predicting employee attrition more accurately, reducing costs, and improving workforce stability.

II. Literature Survey

Employee turnover, defined as the rate at which employees exit an organization, has been recognized as a significant challenge for companies across the globe. High turnover rates result in increased costs related to recruitment, onboarding, and training, along with decreased productivity and engagement among the remaining workforce. A comprehensive understanding of the factors driving employee attrition is therefore crucial to mitigating these adverse outcomes. Over the years, numerous studies have examined the predictors of turnover, with job satisfaction, salary, and opportunities for career advancement, such as the number of projects, emerging as prominent factors.

In [4] study applied machine learning techniques such as logistic regression, decision trees, random forests, and neural networks to predict employee turnover in high-stress sectors. Using factors like demographics, job satisfaction, performance, and stress levels, the models were trained and evaluated with accuracy, precision, recall, and AUC-ROC, showing the potential of ML in improving retention strategies.

A systematic review of 52 studies published between 2012 and 2023 highlights that more than 20 machine learning techniques have been applied to employee turnover prediction, with supervised learning dominating (96% of studies), particularly random forest models. The review further identifies salary and overtime as the most critical factors influencing turnover, providing valuable insights into key predictors and methodological trends in this domain [5].

Research by Varkiani et. al highlight the growing use of data analytics in HR to support data-driven decisions. Employee turnover remains a key challenge, as it negatively affects productivity, performance, and reputation. One study on an Italian financial firm applied machine learning models to predict attrition and used the SHAP algorithm to not only identify important features but also understand their direction of influence. This approach emphasized that knowing a factor's importance alone is not enough—its direction can sometimes be unexpected, making deeper interpretation essential for effective HR decision-making, especially in the evolving post-pandemic job market. [6].

In a related study, Kumar et al. (2023) highlights the efficacy of machine learning models, including Random Forest, Logistic Regression, and Support Vector Machines (SVM), in identifying the key determinants of employee attrition. The authors emphasize the Random Forest model's capability to handle complex, high-dimensional datasets, particularly through its integration with SMOTE to address class imbalances. Their research contributes to the development of proactive strategies for employee retention, emphasizing the importance of predicting turnover to reduce organizational costs and stabilize workforce composition [7].

Furthermore, a study by Valle and Ruz (2015) demonstrates the significance of behavioral factors, especially job satisfaction, in influencing attrition rates. By processing large datasets through machine learning techniques, the authors effectively identify patterns that predict turnover, showcasing how these models can uncover intricate relationships between employee behaviors and their likelihood of leaving [8].

Rombaut and Guerry (2018) investigates voluntary employee turnover through the analysis of HR databases. By employing clustering and classification algorithms, this research identifies key patterns and characteristics of employees predisposed to leaving the organization. The study highlights the utility of machine learning in HR analytics for fostering workforce stability and improving employee retention efforts, ultimately enhancing overall organizational performance [9].

These studies collectively demonstrate that machine learning models enable organizations to predict employee turnover with greater precision. As companies increasingly leverage data-driven insights, predictive analytics will continue to play an essential role in enhancing employee retention and minimizing the associated costs of turnover.

III. Methodology

The methodology follows a structured process that begins with data collection and preprocessing, followed by exploratory data analysis (EDA), machine learning model development, and evaluation. Fig 2 shows pictorial representation of methodology used in the research.

A. Data Collection

The dataset for this research includes historical employee data sourced from the human resources department. The data consists of varios features as depicted in Table I. This dataset serves as the foundation for building machine learning models to predict employee turnover.

B. Data Preprocessing

Data preprocessing is a critical step in preparing the dataset for analysis and modeling. The following steps were undertaken to ensure data quality and suitability for machine learning algorithms:

- Visualization of Numeric Features: Histograms were created using Matplotlib to visualize the distribution of each numeric feature. This helped in understanding the spread and central tendencies of the data.
- Outlier Detection and Handling: The detected outliers and inliers were saved as distinct CSV files, and the total counts of outliers (Number of outliers: 1499) and inliers (Number of inliers: 13500) were reported for further analysis.
- Correlation Analysis:

TABLE I
DESCRIPTION OF DATASET.

Feature	Description
Satisfaction Level	A numerical score indicating the employee's satisfaction with their job.
Last Evaluation	The score from the employee's most recent performance evaluation.
Number of Projects	The total number of projects the employee has worked on.
Average Monthly Hours	The average number of hours the employee works each month.
Time Spent at Company	The total number of years the employee has been with the company.
Work Accident	A binary indicator of whether the employee has experienced a work-related accident (1 for yes, 0 for no).
Left	A binary indicator of whether the employee has left the company (1 for yes, 0 for no).
Promotion in Last 5 Years	The number of promotions the employee has received in the last five years.
Department	The department or sales team the employee belongs to.
Salary	The employee's salary level, typically categorized as low, medium, or high.

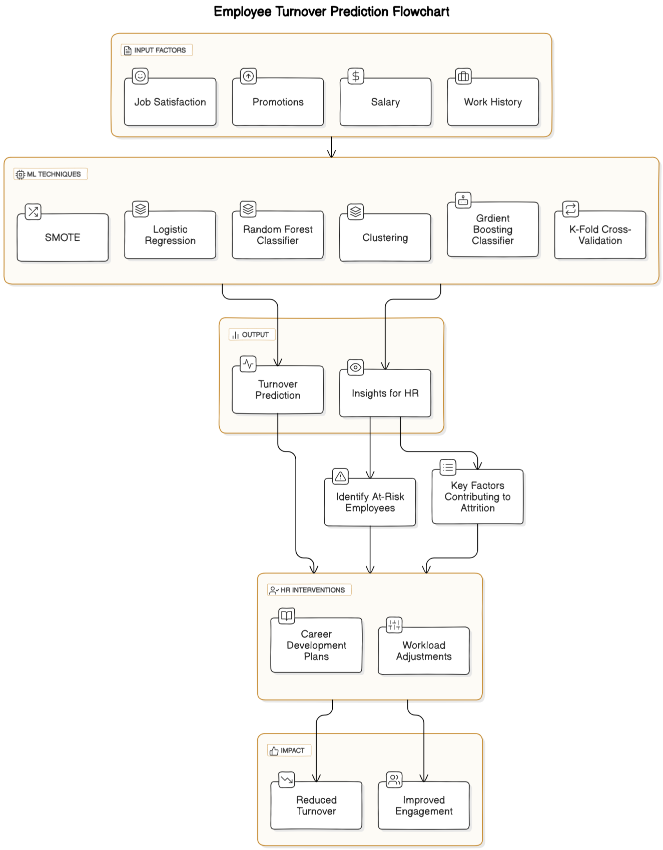

Fig. 2. Employee Turnover Prediction Methodology.

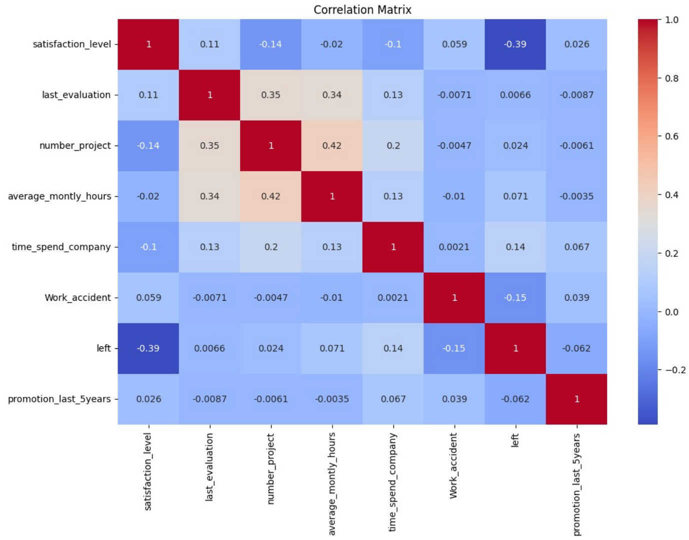

Fig. 3. Co-relation matrix to analyze potential relationships.

The correlation heatmap Fig 3 revealed the following key insights: The correlation matrix reveals that employee satisfaction level has the strongest negative correlation with turnover (-0.39), indicating that lower satisfaction is closely linked to higher attrition. Work accidents (-0.15) and lack of promotions (-0.062) show weaker negative associations with turnover, suggesting only a minor impact. Variables related to workload—such as number of projects, last evaluation, and average monthly hours—are positively correlated with each other (0.34–0.42), reflecting that employees handling more projects tend to work longer hours and receive higher evaluations. However, these workload factors exhibit only weak correlations with turnover, implying they are not direct predictors. Overall, the analysis highlights employee satisfaction as the most influential factor in predicting turnover, while other attributes contribute marginally.

- Satisfaction level had a negative correlation with turnover, confirming that dissatisfied employees are more likely to leave.
- Promotions within the last five years were moderately negatively correlated with turnover, indicating that promotions reduce attrition.
- High evaluations were positively correlated with leaving, suggesting that recognition alone does not ensure retention.
- Heavy workloads were positively associated with turnover, showing that overburdened employees are at greater risk of resigning.

- Data Normalization: The combined dataset underwent standardization using the StandardScaler method. This normalization process ensured that all features have a mean of zero and a standard deviation of one, which is essential for effective clustering and modeling.

C. Handling Class Imbalance

The dataset showed a strong imbalance, with far more employees staying than leaving. Such imbalance reduces a model's ability to correctly identify attrition cases. To address this, the Synthetic Minority Over-sampling Technique (SMOTE) was applied. This balanced the dataset by generating synthetic samples of the minority class, which improved fairness in training and increased the reliability of precision, recall, and F1-scores.

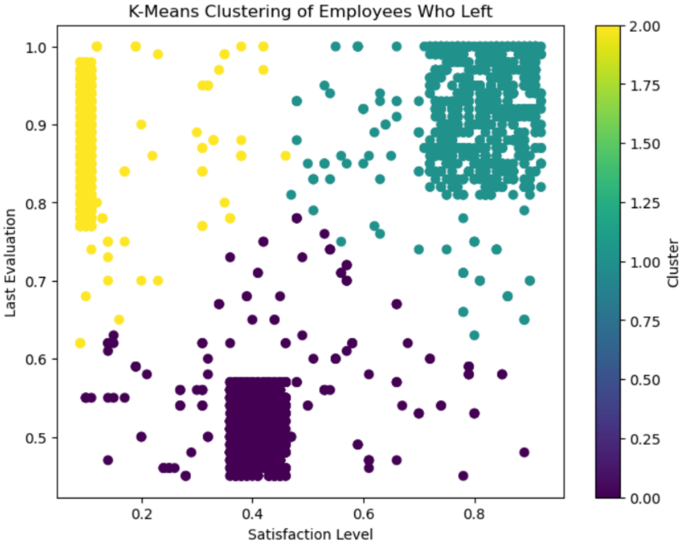

Fig. 4. Clustering of employees who left based on their satisfaction and evaluation.

IV. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was carried out to identify patterns and relationships between employee attributes and turnover. This stage provided clarity on the most influential factors driving attrition and also guided the preparation of data for modeling.

A. Feature Distributions

The distribution of numerical features such as satisfaction level, last evaluation, number of projects, average monthly hours, and tenure revealed clear patterns:

- Employees who left typically reported low satisfaction levels, mostly ranging between 0.1 and 0.4.
- High evaluation scores (above 0.8) were more common among employees who resigned, indicating that even top performers tend to leave.
- Employees working more than 250 hours per month showed a higher likelihood of turnover.
- Attrition peaked among employees with 2–4 years of tenure, reflecting early-career mobility.

V. Clustering Employees Based on Satisfaction and Evaluation

The clustering analysis shown in Fig 4 successfully identifies three distinct groups of employees based on their satisfaction levels and performance evaluation scores, as follows:

1) Cluster 1 (Yellow): This group consists of employees who have low satisfaction but high performance evaluation scores. These employees are likely at high risk of leaving the company, despite their strong performance. This suggests that their dissatisfaction may stem from factors unrelated to their job performance, such as company culture, work-life balance, or lack of growth opportunities.
2) Cluster 2 (Teal): Employees in this cluster exhibit both high satisfaction and high performance evaluation scores. These individuals are performing well and appear satisfied with their work, making them less likely to leave. Retaining this group should be a priority for the company, as they represent the high-performing and engaged workforce.
3) Cluster 3 (Purple): This cluster comprises employees with mixed levels of satisfaction and performance. There is more variability in this group, with some employees being satisfied and others showing dissatisfaction, as well as differing performance levels.

Clustering employees based on satisfaction and evaluation scores Fig 4 produced three distinct groups:

1) Low satisfaction, high evaluation — top performers but at the highest risk of leaving.
2) High satisfaction, high evaluation — stable, satisfied employees least likely to leave.
3) Mixed satisfaction and evaluation — moderate-risk group that requires closer HR monitoring.

A. Key Findings

Overall, the EDA confirmed that satisfaction level is the most important predictor of turnover, followed by promotion history and workload intensity. Clustering analysis further segmented employees into meaningful groups, highlighting where HR retention efforts should be concentrated.

VI. Results and Outcomes

The study followed a structured workflow that combined exploratory analysis, preprocessing, clustering, and supervised learning models to predict employee turnover. The results and their implications are summarized below.

A. Model Performance

Three machine learning models were trained and evaluated: Logistic Regression, Random Forest, and Gradient Boosting. To ensure robustness and avoid bias from a single data split, k-fold cross-validation was used, allowing each model to be tested across multiple folds of the dataset.

Performance was compared using accuracy, precision, recall, and F1-score. Table II and Fig. 5 summarize the results.
Analysis:

TABLE II
Model Performance on Employee Turnover Prediction

Model	**Accuracy (%)**	**Precision**	**Recall**	**F1-Score**
Logistic Regression	91	0.89	0.87	0.88
Random Forest	98	0.97	0.98	0.98
Gradient Boosting	96	0.95	0.94	0.95

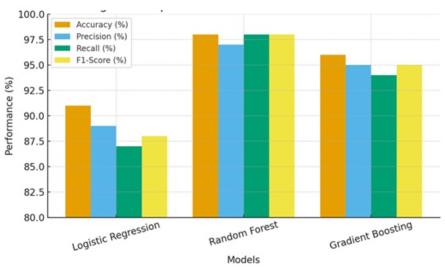

Fig. 5. Comparison of model performance metrics (Accuracy, Precision, Recall, and F1-Score) for Logistic Regression, Random Forest, and Gradient Boosting.

- **Random Forest** achieved the best results, with 98% accuracy and an F1-score of 0.98. High precision (0.97) and recall (0.98) confirmed its ability to correctly identify most attrition cases while minimizing false predictions.
- **Gradient Boosting** reached 96% accuracy and an F1-score of 0.95. Its probability estimates made it useful for ranking employees by risk, although it was slightly less balanced than Random Forest.
- **Logistic Regression** achieved 91% accuracy with an F1-score of 0.88. While less accurate, its interpretability allowed a clearer understanding of how features influence turnover risk.

Based on these results, Random Forest was selected as the best-performing model.

B. Key Outcomes

The main outcomes of the analysis are:

- Satisfaction level is the strongest predictor of turnover.
- Promotions significantly reduce attrition, confirming career growth as a key retention factor.
- Heavy workloads are linked to higher turnover risk.
- High evaluation but low satisfaction identifies top performers most vulnerable to leaving.
- Random Forest, evaluated under k-fold cross-validation, outperformed all other models.

VII. Conclusion

The results confirm that machine learning can effectively support HR teams in predicting turnover and designing focused interventions. By combining accuracy, precision, recall, and F1-score, the models provide both breadth and balance in evaluating performance. Clustering analysis complements this by segmenting employees into groups that highlight where HR should direct retention strategies.

Among the tested models, Random Forest achieved the highest accuracy (98%) and balanced precision–recall performance, making it the most effective approach. Gradient Boosting also performed strongly with 96% accuracy, offering fine-grained probability scores that can support targeted employee retention strategies. Logistic Regression, while less accurate, provided balanced and interpretable results. Overall, ensemble methods, particularly Random Forest, are highly effective for turnover prediction when combined with appropriate resampling techniques.

Finally, this study demonstrates that ML-based prediction systems can enhance HR decision-making and directly contribute to workforce stability. When integrated into HR dashboards, such models allow real-time monitoring of attrition risk, reduce costs associated with hiring and training, and help preserve institutional knowledge through proactive retention policies.

References

[1] Alaskar, L., Crane, M., & Alduailij, M. (2019, December). Employee turnover prediction using machine learning. In International conference on computing (pp. 301-316). Cham: Springer International Publishing.

[2] Al Akasheh, Mariam, et al. ”Enhancing the prediction of employee turnover with knowledge graphs and explainable AI.” IEEE Access 12 (2024): 77041-77053.

[3] Haque, M., Paralkar, T. A., Rajguru, S., Goyal, A. A., Patil, T., & Upreti, K. Featuring Machine Learning Models to Evaluate Employee Attrition: A Comparative Analysis of Workforce Stability-Relating Factors.

[4] Adeusi, K. B., Amajuoyi, P., & Benjami, L. B. (2024). Utilizing machine learning to predict employee turnover in high-stress sectors. International Journal of Management & Entrepreneurship Research, 6(5), 1702-1732.

[5] M. Al Akasheh, E. F. Malik, O. Al Hujran, and N. M. Zaki, “A decade of research on machine learning techniques for predicting employee turnover: A systematic review,” Expert Systems with Applications, vol. X, 2024.

[6] S. M. Varkiani, Predicting employee attrition and explaining its determinants, Expert Systems with Applications, 2025.

[7] Kumar, P., Gaikwad, S. B., Ramya, S. T., Tiwari, T., Tiwari, M., & Kumar, B. (2023). Predicting employee turnover: a systematic machine learning approach for resource conservation and workforce stability. Engineering Proceedings, 59(1), 117.

[8] Valle, M. A., & Ruz, G. A. (2015). Turnover prediction in a call center: behavioral evidence of loss aversion using random forest and naïve bayes algorithms. Applied Artificial Intelligence, 29(9), 923-942.

[9] I. Safiulina and H. Rabii, “Predicting employee turnover in consulting firms: A machine learning approach to multi-parameter satisfaction modeling,” in Proc. Int. Conf. Research in Human Resource Management, vol. 1, no. 1, pp. 35–51, 2024.