Engineering and Scientific International Journal (ESIJ)
Volume 13, Issue 2, April – June 2026
ISSN 2394-7187(Online)
ISSN 2394 - 7179 (Print)

HR Insight Analytics: An Intelligent Platform for Employee Performance Evaluation and Attrition Risk Prediction using Random Forest and Workforce Visualization

Pasupuleti Sai Deepthi*[1], Mr. S. Manjunath Reddy[2]

[1]*Student, Department of Computer Applications, Viswam Engineering College, Madanapalle, Andhra Pradesh.*
[2] *Assistant Professor, Department of Computer Applications, Viswam Engineering College, Madanapalle, Andhra Pradesh.*

Abstract — HR Insight Analytics is an end-to-end platform designed to predict employee performance and attrition risk, addressing the high costs and disruptions caused by unexpected workforce turnover. Attrition is influenced by multiple factors, including compensation dissatisfaction, limited career growth, poor work-life balance, workload stress, and weak managerial relationships. The platform leverages a synthetic dataset of 15,000 employee records with 34 attributes covering demographics, job roles, compensation, engagement levels, and attendance patterns. The system operates through four stages: a Python-based engine generates realistic workforce data; a Random Forest classification model with advanced feature engineering predicts attrition with high accuracy using cross-validation; data is structured in a normalized MySQL database for efficient storage and querying; and interactive dashboards present insights on attrition trends, performance distribution, and risk factors. The platform delivers actionable insights by assigning attrition probabilities, risk categories, and retention priorities to employees, along with tailored intervention strategies, enabling organizations to proactively manage workforce challenges and improve retention outcomes.

Keywords: *Employee Attrition Prediction; HR Analytics; Machine Learning (Random Forest); Workforce Performance Analysis; Predictive Modeling.*

1. Introduction

Human capital is the most critical and irreplaceable resource in any organization, encompassing expertise, knowledge, and relationships that drive performance and innovation. Employee attrition leads to significant direct and indirect costs, including recruitment, training, productivity loss, and reduced team morale. Studies indicate that replacing an employee can cost between 50% and 200% of their annual salary, making attrition a major financial concern for organizations. Traditionally, attrition management has been reactive, relying on resignation notices or visible disengagement signals, often when it is too late for effective intervention. HR Insight Analytics addresses this challenge by applying a Random Forest model to employee data to predict attrition risk in advance. It enables HR teams to identify at-risk employees, implement targeted retention strategies, and allocate resources efficiently. The platform integrates with a MySQL database and visualization tools like Power BI for actionable insights. Built using Python and its data science ecosystem, the system is scalable, reproducible, and accessible. A synthetic dataset simulates real-world workforce patterns while ensuring data privacy and enabling practical model development.

1.1 Project Objectives

The development of HR Insight Analytics is guided by key objectives across technical, analytical, and operational domains. The primary objective is to build a fully automated end-to-end analytics pipeline that processes employee data, performs feature engineering, trains a machine learning model, stores predictions in a relational database, and generates visualization dashboards with minimal manual intervention. Additional objectives include achieving high predictive accuracy in identifying at-risk employees, developing advanced feature engineering incorporating derived metrics, designing a normalized relational database schema for efficient storage, and creating multiple stakeholder-facing dashboards presenting insights on attrition patterns, performance, and retention strategies. Finally, the platform is intended to be scalable, maintainable, and adaptable for real-world deployment.

2. Literature Survey

Employee attrition and workforce analytics have been widely studied across organizational behavior and machine learning domains. Early foundational studies by Hom et al. (2012), Mitchell et al. (2001), and Mobley (1977) explored the psychological and behavioral aspects of employee turnover, identifying job satisfaction, organizational commitment, and job embeddedness as key determinants. Cotton and Tuttle (1986) further strengthened this understanding through a meta-analysis highlighting multiple factors influencing turnover decisions [6,7,8,17]. With the advancement of data analytics, researchers have increasingly applied machine learning techniques to predict employee attrition. Breiman (2001) introduced the Random Forest algorithm, which has become a widely used classification method due to its robustness and accuracy [1]. Chawla et al. (2002) proposed SMOTE to address class imbalance, a common issue in attrition datasets [5]. Studies such as Jain

DOI: **10.30726/esij/v13.i2.2026.1320033**

ISSN 2394-7187(Online)
ISSN 2394 - 7179 (Print)

and Nayyar (2018) and Alao and Adeyemo (2013) demonstrated the effectiveness of XGBoost and Decision Trees in predicting employee turnover [19,20].The development of tools and libraries has significantly supported HR analytics implementation. Pedregosa et al. (2011) introduced Scikit-learn, enabling efficient machine learning modeling, while McKinney (2010) and Harris et al. (2020) contributed through Pandas and NumPy [2,3,11]. Visualization tools such as Matplotlib (Hunter, 2007) and platforms like Power BI (Microsoft, 2024) enhance the interpretability of analytical results [12,16]. Recent research by Zhao et al. (2018) emphasizes the growing importance of HR analytics in strategic decision-making [9].

3. System Proposal

3.1 Existing System

Existing employee attrition management approaches range from traditional methods to modern analytical tools. Exit interview analysis provides qualitative insights but is retrospective and often influenced by response bias. Employee engagement surveys offer quantitative insights but suffer from low participation and infrequent administration. HRIS-based reporting analyzes historical turnover rates but lacks predictive capability and individual-level risk assessment. Advanced proprietary platforms like Workday, SAP SuccessFactors, and Oracle HCM use machine learning for prediction, but are expensive and less customizable. Managerial intuition, while context-rich, is subjective, inconsistent, and not scalable.

3.2 Proposed System

HR Insight Analytics proposes a comprehensive, automated, and data-driven workforce analytics platform integrating machine learning, database systems, and visualization tools into a four-stage pipeline: synthetic data generation, predictive modeling, relational database storage, and dashboard visualization. As illustrated in Fig. 1, the system architecture spans five layers from orchestration to visualization.

Table 1. HR Insight Analytics System Architecture Layers

Layer	Component & Function
Orchestration	main.py — Coordinates pipeline: Data Generation → ML Training → DB Storage → Visualization
Data Source	hr_data_generator.py — Generates 15,000 synthetic records (34 attributes) → HR_Employee_Dataset.xlsx
Analytics	hr_prediction_model.py — Feature Engineering → Random Forest → Cross-Validation → Predictions
Storage	mysql_storage.py — MySQL HRInsight_Analytics DB with 7 normalized tables
Visualization	visualization_dashboard.py — 6 Professional Dashboards (Executive, Department, Attrition, Performance, Retention, Predictions)

Table 2. Pipeline Flow — Stages and Outputs

Stage	Process & Output
START	python main.py — Pipeline timer starts; orchestration layer initializes
Stage 1	Data Generation — generate_hr_dataset(15000) → HR_Employee_Dataset.xlsx (3 sheets)
Stage 2	ML Pipeline — load → preprocess → feature engineer → train RF → cross-validate → predict → HR_Attrition_Predictions.xlsx + model.pkl
Stage 3	MySQL Storage — create_db → create_tables → insert_records → 7 normalized tables
Stage 4	Dashboards — generate_all_dashboards → 6 JPEG dashboard files in /dashboards/
END	Pipeline Complete — Execution summary printed; total time reported

The platform is built on a synthetic dataset of 15,000 employee records with 34 attributes representing demographic, organizational, compensation, engagement, and behavioral factors. In the modeling stage, feature engineering introduces derived variables including income-per-experience ratio, promotion staleness index, composite satisfaction score, and overwork flag, collectively improving predictive accuracy. A Random Forest classifier is trained using 75% dataset split with five-fold cross-validation and evaluated using accuracy, precision, recall, F1-score, and ROC-AUC metrics.

The system integrates with a MySQL database designed using a normalized seven-table schema, as detailed in Table 1 above. The visualization layer produces six professional dashboards using Matplotlib, presented through the pipeline summarized in Table 2.

3.2.1 Limitations of the Proposed System

While the platform demonstrates strong capabilities, certain limitations exist. It relies on synthetic rather than real organizational data, which may not fully capture company-specific culture or local workforce dynamics. The system requires retraining on actual employee data for accurate real-world deployment. Database integration currently uses hardcoded MySQL credentials, posing security risks in production environments. Generated dashboards are static JPEG files, limiting interactivity. Additionally, the sequential pipeline execution may encounter performance issues with very large datasets and lacks direct real-time integration with live BI tools.

DOI: 10.30726/esij/v13.i2.2026.1320033

ISSN 2394-7187(Online)
ISSN 2394 - 7179 (Print)

4. Implementation

HR Insight Analytics is structured into five main Python modules and one orchestration script, each handling a specific stage of the analytics pipeline. The modules include data generation (hr_data_generator.py), machine learning (hr_prediction_model.py), database integration (mysql_storage.py), visualization (visualization_ dashboard. py), and orchestration (main.py). The dependency structure is hierarchical, ensuring each module relies only on the outputs of the previous stage.

4.1 Module Description

- Data Generation Module: Generates a synthetic dataset of employee records using statistical distributions and fixed random seeds for reproducibility. It creates demographic, organizational, and behavioral attributes, along with attrition labels based on multiple weighted risk factors such as satisfaction levels, overtime, compensation, and promotion history. Output is saved as an Excel file with multiple sheets.
- Machine Learning Module: Handles preprocessing, feature engineering, model training, and prediction. Creates derived features like income-to-experience ratio, promotion staleness, satisfaction score, and overwork indicators. A Random Forest classifier is trained using a 75–25 split with cross-validation and evaluated using standard metrics, generating attrition probabilities, risk categories, and feature importance rankings.
- MySQL Storage Module: Manages database connectivity and storage using MySQL. Creates a normalized schema with seven tables including employee details, performance metrics, predictions, and analytics summaries. Data is inserted using secure, parameterized queries, ensuring integrity and efficient querying.
- Visualization Dashboard Module: Generates six dashboards using Matplotlib — Executive Overview, Department Analysis, Attrition Factors, Performance Analysis, Retention Strategy, and Prediction Results — presenting key insights on attrition trends, risk distribution, and feature importance.
- Orchestration Module: Coordinates the entire pipeline by executing each stage sequentially. Manages data flow between modules, tracks execution time, and generates a final summary of outputs including datasets, predictions, database storage status, and dashboard files.

5. Results and Discussion

The HR Insight Analytics platform demonstrates strong performance in predicting employee attrition and generating actionable workforce insights. As shown in Fig. 1, the system's overall analytical workflow delivers measurable outcomes across all four pipeline stages. The Random Forest model achieved high overall accuracy with balanced precision and recall scores, indicating reliable classification of both attrition and non-attrition cases. Particular emphasis was placed on recall for high-risk employees, ensuring that most at-risk individuals were correctly identified. The ROC-AUC score confirmed the model's strong discriminative ability in distinguishing between employees likely and unlikely to leave.

Feature importance analysis revealed that job satisfaction, work-life balance, overtime, income level, and promotion history were the most influential predictors of attrition. Derived features like income-per-experience ratio and promotion staleness significantly improved model performance by capturing underlying organizational dynamics. The overwork flag showed a strong association with higher attrition probability, highlighting the impact of burnout on employee retention. The system successfully categorized employees into five risk levels — Very Low, Low, Medium, High, and Very High — enabling targeted intervention strategies. High and Very High-risk employees were assigned higher retention priority scores and recommended actions such as compensation adjustments, workload balancing, or career development opportunities, allowing HR teams to allocate resources efficiently. At the aggregate level, department-level analysis identified specific departments with higher attrition rates, enabling focused policy interventions. The Attrition Factors dashboard confirmed relationships between key variables and attrition risk, while the Performance Analysis dashboard highlighted the distribution of high-performing employees and their associated risk levels. Integration with the MySQL database ensured structured storage and easy retrieval of prediction results, enabling seamless connectivity with business intelligence tools.

6. Conclusion and Future Enhancement

6.1 Conclusion

HR Insight Analytics successfully demonstrates an end-to-end workforce analytics platform that transforms employee data into actionable insights for attrition prediction and performance analysis. The system integrates synthetic data generation, machine learning modeling, database storage, and visualization into a unified pipeline, ensuring both technical efficiency and practical usability. The Random Forest classifier achieved strong performance with balanced handling of class imbalance and improved accuracy through feature engineering. The five-tier risk classification and retention priority scoring provide meaningful segmentation, allowing HR teams to focus on high-risk employees. The MySQL database integration ensures structured data storage and seamless connectivity with business intelligence tools.

6.2 Future Enhancement

Future improvements can further enhance the platform's accuracy, usability, and real-world applicability.

DOI: 10.30726/esij/v13.i2.2026.1320033

ISSN 2394-7187(Online)
ISSN 2394 - 7179 (Print)

These include implementing ensemble models combining Gradient Boosting, Logistic Regression, and SVM for improved prediction accuracy; using automated hyperparameter tuning techniques such as Bayesian optimization; integrating SHAP analysis for better interpretability of individual predictions; connecting with real-time HR systems like Workday and SAP SuccessFactors for continuous data updates; developing interactive web-based dashboards using Streamlit or Dash; enabling automated model retraining and performance monitoring; incorporating NLP-based sentiment analysis from employee feedback; and expanding the system to include performance prediction, promotion readiness, and workforce planning.

References

[1] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
[2] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
[3] W. McKinney, "Data Structures for Statistical Computing in Python," Proc. 9th Python in Science Conf., 2010, pp. 51–56.
[4] S. M. Lundberg and S. I. Lee, "A Unified Approach to Interpreting Model Predictions," Advances in Neural Information Processing Systems, vol. 30, pp. 4765–4774, 2017.
[5] N. V. Chawla et al., "SMOTE: Synthetic Minority Over-sampling Technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.
[6] P. W. Hom et al., "Reviewing Employee Turnover: Focusing on Proximal Withdrawal States," Psychological Bulletin, vol. 138, no. 5, pp. 831–858, 2012.
[7] T. R. Mitchell et al., "Why People Stay: Using Job Embeddedness to Predict Voluntary Turnover," Academy of Management Journal, vol. 44, no. 6, pp. 1102–1121, 2001.
[8] J. L. Cotton and J. M. Tuttle, "Employee Turnover: A Meta-Analysis and Review," Academy of Management Review, vol. 11, no. 1, pp. 55–70, 1986.
[9] Y. Zhao et al., "Enterprise Human Resources Analytics: A Literature Review," Journal of Organizational Behavior, vol. 39, no. 8, pp. 1048–1060, 2018.
[10] Society for Human Resource Management, "SHRM Benchmarking Report: Human Capital," Alexandria, VA, USA, 2022.
[11] C. R. Harris et al., "Array Programming with NumPy," Nature, vol. 585, pp. 357–362, 2020.
[12] J. D. Hunter, "Matplotlib: A 2D Graphics Environment," Computing in Science and Engineering, vol. 9, no. 3, pp. 90–95, 2007.
[13] G. van Rossum and F. L. Drake, Python 3 Reference Manual. Scotts Valley, CA, USA: CreateSpace, 2009.
[14] M. Bohanec et al., "Decision Making with Machine Learning and Rule-Based Classifiers," Expert Systems with Applications, vol. 71, pp. 242–257, 2017.
[15] MySQL AB, "MySQL 8.0 Reference Manual," Oracle Corporation, 2023.
[16] Microsoft Corporation, "Power BI Desktop Documentation," 2024.
[17] W. H. Mobley, "Intermediate Linkages in the Relationship between Job Satisfaction and Employee Turnover," Journal of Applied Psychology, vol. 62, no. 2, pp. 237–240, 1977.
[18] J. L. Price, The Study of Turnover. Ames, IA, USA: Iowa State University Press, 1977.
[19] R. Jain and A. Nayyar, "Predicting Employee Attrition Using XGBoost Machine Learning Approach," Proc. Int. Conf. System Modeling and Advancement in Research Trends, 2018, pp. 113–120.
[20] D. Alao and A. B. Adeyemo, "Analyzing Employee Attrition Using Decision Tree Algorithms," Computing, Information Systems, Development Informatics and Allied Research Journal, vol. 4, no. 1, pp. 17–28, 2013.

DOI: **10.30726/esij/v13.i2.2026.1320033**