East African Scholars Multidisciplinary Bulletin
Abbreviated Key Title: East African Scholars Multidiscip Bull
ISSN 2617-4413 (Print) | ISSN 2617-717X (Online)
Published By East African Scholars Publisher, Kenya
Volume-9 | Issue-5 | Nov-Dec-2026 | DOI: https://doi.org/10.36349/easjmb.2026.v09i05.002

OPEN ACCESS

Original Research Article

Harnessing Machine Learning for Effective HR Practices: A Multi-Dataset Comparative Evaluation Using Orange Data Mining

Dr. Varad Rajan Bhanage[1]*

[1]Assistant Professor, Department of Management Studies, Vivekanand College, Kolhapur, Maharashtra, India

Article History
Received: 24.09.2026
Accepted: 16.11.2026
Published: 18.11.2026

Journal homepage:
https://www.easpublisher.com

Quick Response Code

Abstract: The growing adoption of Human Resource Analytics has transformed traditional HR practices by enabling organizations to make data-driven decisions related to recruitment, employee retention, performance management, workforce planning, and talent development. However, selecting the most appropriate machine learning (ML) model for a specific HR problem remains a significant challenge for HR practitioners and researchers. This study aims to evaluate the effectiveness of various machine learning algorithms across multiple HR datasets using the Orange Data Mining platform. The research adopts a comparative experimental approach in which diverse HR datasets representing different functional areas of Human Resource Management, such as employee attrition, recruitment and selection, employee performance, compensation, training effectiveness, and workforce management, are analyzed. Multiple machine learning models, including Logistic Regression, Decision Tree, Random Forest, Naïve Bayes, Support Vector Machine, k-Nearest Neighbors, and Neural Networks, are implemented and evaluated within Orange Data Mining. Model performance is assessed using standard evaluation metrics such as classification accuracy, precision, recall, F1-score, Receiver Operating Characteristic (ROC) curves, Area under the Curve (AUC), confusion matrices, and cross-validation results. The findings provide a comprehensive comparison of machine learning models across different HR datasets and reveal that model performance varies significantly depending on dataset characteristics and HR application contexts. The study contributes to both theory and practice by offering empirical evidence on the suitability of machine learning techniques for HR analytics. Furthermore, it proposes a practical framework that can assist HR professionals, analysts, and researchers in selecting appropriate machine learning models for specific HR-related decision-making scenarios. The research demonstrates how user-friendly visual analytics platforms such as Orange Data Mining can facilitate the adoption of advanced machine learning techniques in Human Resource Management, thereby supporting more accurate, efficient, and evidence-based HR practices.

Keywords: Human Resource Analytics, Machine Learning, Orange Data Mining, HR Decision Support, Comparative Model Evaluation, Data-Driven Human Resource Management.

Copyright © 2026 The Author(s): This is an open-access article distributed under the terms of the Creative Commons Attribution **4.0 International License (CC BY-NC 4.0)** which permits unrestricted use, distribution, and reproduction in any medium for non-commercial use provided the original author and source are credited.

INTRODUCTION

The rapid digital transformation of organizations has significantly changed the way Human Resource Management (HRM) functions are performed. Traditional HR practices, which primarily relied on intuition and experience, are increasingly being replaced by data-driven approaches that facilitate objective, timely, and strategic decision-making. Human Resource Analytics (HRA) has emerged as a critical discipline that integrates data analytics, statistical techniques, and machine learning (ML) to support various HR functions, including recruitment, employee retention, performance management, workforce planning, compensation analysis, and talent development (Roul *et al*., 2024; Wang *et al*., 2024).

***Corresponding Author:** Dr. Varad Rajan Bhanage
Assistant Professor, Department of Management Studies, Vivekanand College, Kolhapur, Maharashtra, India

Machine learning has become one of the most influential technologies in HR analytics because of its ability to discover hidden patterns, classify employee behavior, and generate predictive insights from large and complex datasets. Organizations increasingly employ ML techniques to predict employee attrition, identify high-performing candidates, evaluate employee performance, detect absenteeism patterns, and optimize workforce planning. These predictive capabilities enable HR professionals to make informed decisions, improve organizational efficiency, and gain a competitive advantage (Heidemann *et al*., 2024; Kim *et al*., 2024).

Recent studies indicate that HR analytics has evolved from descriptive reporting toward predictive and prescriptive analytics, where machine learning plays a central role. Advances in big data technologies, artificial intelligence, and cloud computing have accelerated the adoption of ML-based HR solutions across organizations of different sizes. Consequently, researchers have devoted considerable attention to developing predictive models that improve HR decision-making and organizational performance (Sakib *et al*., 2024; Donthu *et al*., 2024; Sangu *et al*., 2024).

Despite the growing popularity of machine learning in HR analytics, selecting an appropriate ML algorithm remains a major challenge. Numerous classification algorithms, including Logistic Regression, Decision Trees, Random Forest, Naïve Bayes, Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), Neural Networks, and Gradient Boosting methods, have been applied to HR datasets. However, their predictive performance often varies according to dataset characteristics such as sample size, feature distribution, class imbalance, missing values, and data complexity. Consequently, a model that performs well on one HR dataset may not necessarily provide superior performance on another (Heidemann *et al*., 2024; Tanasescu *et al*., 2024).

Another important challenge is the practical adoption of machine learning techniques by HR professionals. Although sophisticated programming frameworks such as Python and R offer extensive analytical capabilities, many HR practitioners have limited programming expertise. Visual data mining platforms such as Orange Data Mining address this limitation by providing an intuitive drag-and-drop interface that enables users to implement, compare, visualize, and interpret multiple machine learning algorithms without extensive coding knowledge. Such platforms facilitate experimental analysis and improve the accessibility of advanced analytics for business professionals and academic researchers.

Existing literature predominantly focuses on developing predictive models for individual HR applications such as employee attrition, recruitment, or performance evaluation. Most studies evaluate only one or two datasets and often compare a limited number of machine learning algorithms. Consequently, the findings are generally restricted to specific HR problems and do not provide comprehensive guidance regarding the suitability of different ML models across diverse HR domains. Furthermore, relatively few studies have employed a unified experimental framework using multiple HR datasets and a consistent evaluation methodology to compare various machine learning algorithms.

To address this research gap, the present study conducts a comprehensive comparative evaluation of multiple machine learning classification models using several HR-related datasets representing different functional areas of Human Resource Management. The study utilizes Orange Data Mining as the experimental platform to implement and compare various machine learning algorithms. Model performance is evaluated using widely accepted performance metrics, including classification accuracy, confusion matrix, precision, recall, F1-score, Receiver Operating Characteristic (ROC) curve, Area Under the Curve (AUC), and cross-validation results. By systematically comparing algorithmic performance across multiple datasets, this research seeks to identify the relative strengths and limitations of different ML models under varying HR data conditions.

The findings of this study are expected to contribute to both academic research and professional practice. From a theoretical perspective, the study extends the existing body of knowledge by providing empirical evidence regarding the comparative effectiveness of machine learning models in HR analytics across multiple datasets. From a practical perspective, the findings offer valuable guidance to HR professionals, business analysts, researchers, and practitioners in selecting appropriate machine learning algorithms for different HR applications. Additionally, the study demonstrates the usefulness of Orange Data Mining as an accessible and effective platform for conducting machine learning experiments in Human Resource Analytics, thereby promoting the wider adoption of data-driven HR decision-making.

Objectives:

1. To apply different machine learning models to multiple HR datasets using Orange Data Mining.
2. To compare the performance of machine learning models using evaluation metrics such as Accuracy, Confusion Matrix, ROC Curve, Precision, Recall, and F1-Score.
3. To identify the most suitable machine learning model for different HR datasets to support effective HR decision-making.

LITERATURE REVIEW

Table 1: Literature Review

Authors (Year)	Research Focus	ML Models / Technique	Major Findings	Research Gap
Roul *et al*., (2024)	Systematic review of HR Analytics	Literature Review	HR analytics is rapidly evolving toward predictive analytics.	No comparative evaluation of ML models on multiple datasets.
Wang *et al*., (2024)	HR analytics implementation	Conceptual Framework	Organizational readiness influences successful ML adoption.	Did not experimentally compare ML algorithms.
Sakib *et al*., (2024)	Bibliometric analysis of HR Analytics	Bibliometric Analysis	Significant growth in AI and ML applications in HR.	No empirical model evaluation.
Heidemann *et al*., (2024)	ML performance vs explainability	Multiple ML Algorithms	Predictive accuracy must be balanced with model interpretability.	Limited dataset diversity.
Tanasescu *et al*., (2024)	Employee performance prediction	Supervised ML	ML improves employee performance prediction.	Focused on one HR application.
Kim *et al*., (2024)	HR practices and firm performance	Machine Learning	HR practices significantly affect organizational outcomes.	Did not compare multiple classifiers.
Nguyen *et al*., (2024)	Machine learning in HRM	Various ML Models	ML supports HR decision-making.	Limited experimental comparison.
Aburayya *et al*., (2023)	HR intelligence using ML	Classification Models	ML enhances HR intelligence systems.	Single-domain application.
Downes *et al*., (2023)	HR analytics in employee selection	Predictive Analytics	Analytics improves recruitment decisions.	Focused only on recruitment.
Donthu *et al*., (2024)	Big Data and HR Analytics	Review	Big data strengthens strategic HR decisions.	No algorithm comparison.
Sangu *et al*., (2024)	AI in HR decision-making	AI & ML	AI improves workforce management efficiency.	No multi-dataset evaluation.
Present Study	Comparative evaluation using Orange Data Mining	Decision Tree, Random Forest, Logistic Regression, SVM, Naïve Bayes, k-NN, Neural Network, etc.	Evaluates multiple ML models on multiple HR datasets using common evaluation metrics (Accuracy, ROC, AUC, Precision, Recall, F1-score, Confusion Matrix, Cross-validation).	Addresses the lack of a unified, multi-dataset comparative evaluation framework and provides practical guidance for HR professionals.

Research Gap Statement

Existing studies typically focus on applying a single machine learning model to a specific HR problem, such as employee attrition or recruitment prediction. Limited research has systematically compared multiple machine learning algorithms across diverse HR datasets using a unified evaluation framework. This study addresses this gap by conducting a multi-dataset comparative assessment of machine learning models and identifying their relative suitability for different HR analytics applications.

MATERIALS AND METHODS

Research Design

This study employed a quantitative, comparative experimental research design to evaluate the performance of various machine learning classification algorithms across multiple Human Resource (HR) datasets. The primary objective was to examine how different machine learning models perform when applied to HR-related problems using a common analytical framework. A comparative design was selected because it facilitates the evaluation of multiple algorithms under identical experimental conditions, allowing meaningful comparisons based on standardized performance metrics. The experimental analysis was conducted using Orange Data Mining, an open-source visual analytics platform that enables machine learning model development through a workflow-based interface. Orange was selected because it supports data preprocessing, model training, validation, visualization, and performance evaluation without requiring extensive programming expertise.

Data Sources

The study utilized multiple publicly available HR datasets collected from widely recognized open-data repositories. The datasets represent diverse functional areas of Human Resource Management to ensure broader applicability of the findings. These domains include Employee Attrition, Recruitment and Selection,

Employee Performance, Compensation and Salary Analysis, Employee Training and Development, and Workforce Planning. The datasets vary in terms of sample size, number of attributes, class distribution, and data complexity, enabling the evaluation of machine learning algorithms under different HR scenarios.

Table 2: HR Datasets Used in the Study

Dataset Domain	Purpose	Target Variable
Employee Attrition	Predict employee turnover	Attrition (Yes/No)
Recruitment	Candidate selection	Hiring Decision
Performance Management	Employee performance classification	Performance Rating
Compensation	Salary prediction/classification	Salary Category
Training	Training effectiveness	Training Outcome
Workforce Planning	Employee retention/workforce status	Workforce Status

Data Preprocessing

Before model development, each dataset underwent preprocessing to improve data quality and ensure consistency across experiments. The preprocessing activities such as Identification and treatment of missing values, Removal of duplicate records where applicable, Detection and correction of inconsistent data entries, Conversion of categorical variables into machine-readable formats, Normalization or scaling of numerical attributes where required, Selection of relevant predictor variables, and Definition of the target class for classification are involved in the study. These preprocessing steps were performed using the preprocessing widgets available in Orange Data Mining.

Machine Learning Models

The study compared several supervised machine learning classification algorithms commonly used in HR analytics. These models were selected because of their widespread application, predictive capability, and availability within Orange Data Mining. The evaluated algorithms are Logistic Regression, Decision Tree, Random Forest, Naïve Bayes, Support Vector Machine (SVM), k-Nearest Neighbors (k-NN) and Neural Network. Each algorithm was trained and tested independently on every HR dataset using identical preprocessing procedures.

Experimental Procedure

The experimental workflow consisted of the following stages:

1. Import HR dataset into Orange Data Mining.
2. Perform data preprocessing and cleaning.
3. Select predictor variables and target variable.
4. Apply multiple machine learning classification algorithms.
5. Validate models using cross-validation.
6. Compare model performance using standardized evaluation metrics.
7. Identify the most suitable model for each HR dataset.
8. Interpret the results and compare performance across datasets.

The same workflow was followed for all datasets to ensure consistency and fairness during model comparison.

Performance Evaluation Metrics

Model performance was assessed using multiple evaluation metrics rather than relying solely on classification accuracy. This approach provides a comprehensive understanding of each model's predictive capability. The evaluation measures are Classification Accuracy, Precision, Recall (Sensitivity), F1-Score, Area Under the ROC Curve (AUC), Receiver Operating Characteristic (ROC) Curve, Confusion Matrix, and Cross-validation Score. These metrics collectively evaluate both the correctness of predictions and the ability of the models to distinguish between different classes.

Cross-Validation Strategy

To obtain reliable and unbiased performance estimates, k-fold cross-validation was employed. The datasets were partitioned into ten equal subsets. During each iteration, nine subsets were used for model training while the remaining subset was used for testing. This process was repeated until every subset had served as the testing set once. The final performance score for each algorithm was calculated as the average across all folds. The use of cross-validation minimizes the risk of overfitting and provides a more robust estimate of model performance.

Software Environment

The experimental analysis was conducted using Orange Data Mining, an open-source visual machine learning platform. The software provides integrated tools for data preprocessing, classification, model evaluation, visualization, and comparative analysis through an interactive workflow environment. Graphical outputs, including Confusion Matrices, ROC Curves, Lift Curves, Precision–Recall Curves, Calibration Plots, and Model Comparison results, were generated directly within the software.

Research Framework

The overall research framework followed in this study is illustrated below:

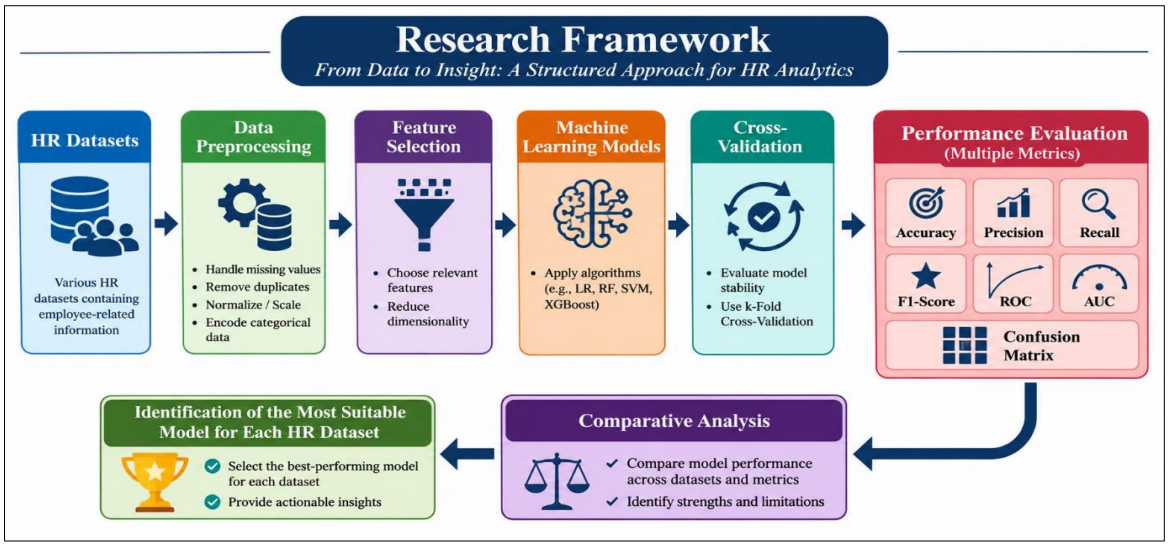


Figure 1: Research Framework

RESULTS AND DISCUSSION

Receiver Operating Characteristic (ROC) Curve Analysis:

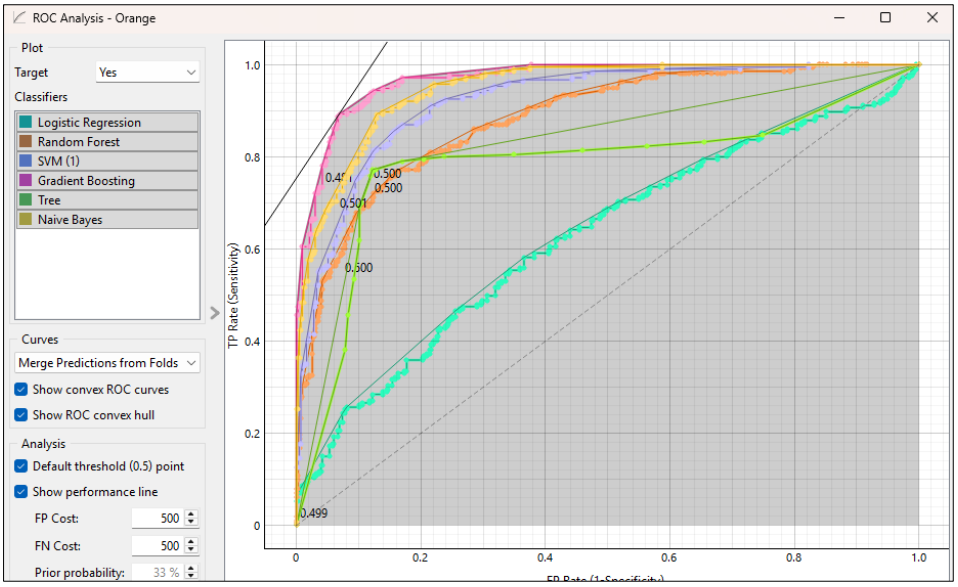


Figure 2: ROC Curve Analysis

Figure 2 presents the Receiver Operating Characteristic (ROC) curves for the six machine learning classifiers evaluated in this study. The ROC curve is a widely accepted performance measure for binary classification problems because it illustrates the trade-off between the True Positive Rate (Sensitivity) and the False Positive Rate (1 − Specificity) across different classification thresholds. A classifier whose ROC curve approaches the upper-left corner demonstrates superior discriminative ability, whereas a curve close to the diagonal reference line indicates performance comparable to random classification.

Among the evaluated models, Gradient Boosting exhibited the best discriminative performance. Its ROC curve consistently remained closest to the upper-left corner of the graph, indicating that the model achieved high sensitivity while maintaining a relatively low false positive rate. This observation is consistent with its AUC value of 0.968, demonstrating excellent capability to distinguish between the positive and negative classes. The Naïve Bayes classifier also showed outstanding classification performance. Its ROC curve closely followed that of Gradient Boosting throughout most threshold values and achieved an AUC of 0.948. The proximity of the curve to the upper-left corner indicates that Naïve Bayes effectively balanced sensitivity and specificity, making it a strong alternative for HR prediction tasks.

The Support Vector Machine (SVM) produced another strong ROC curve with an AUC of 0.920. The

curve remained substantially above the diagonal reference line, indicating good classification capability. Although its performance was slightly inferior to Gradient Boosting and Naïve Bayes, the model demonstrated consistent discrimination between the two target classes. The Random Forest classifier achieved moderate predictive performance with an AUC of 0.881. Its ROC curve indicated satisfactory classification ability but remained below those of Gradient Boosting, Naïve Bayes, and SVM over most threshold values. This suggests that, for the present HR dataset, Random Forest did not exploit the underlying data patterns as effectively as the best-performing models.

The Decision Tree model generated a comparatively lower ROC curve with an AUC of 0.783. Although the curve remained above the diagonal reference line, its greater distance from the upper-left corner indicates reduced discriminative capability. The model therefore exhibited acceptable but comparatively weaker classification performance. Among all evaluated classifiers, Logistic Regression produced the weakest ROC curve, with an AUC of 0.627. Its curve remained considerably closer to the diagonal reference line than the other models, indicating limited ability to distinguish between positive and negative observations. This result suggests that the linear decision boundary employed by Logistic Regression was insufficient to capture the complex relationships present in the HR dataset.

Overall, the ROC analysis demonstrates clear differences in classifier performance. The superior ROC curves obtained by Gradient Boosting, Naïve Bayes, and Support Vector Machine indicate that these algorithms possess stronger discriminatory power than Logistic Regression, Decision Tree, and Random Forest for the analyzed HR data. These findings are consistent with the comparative evaluation results obtained from AUC, classification accuracy, precision, recall, F1-score, and Matthews Correlation Coefficient, thereby confirming the robustness of the selected models. The results further suggest that ensemble learning techniques, particularly Gradient Boosting, are more effective in capturing complex patterns within HR datasets. Consequently, Gradient Boosting appears to be the most appropriate algorithm for supporting predictive Human Resource Analytics, while Naïve Bayes and Support Vector Machine may serve as reliable alternatives depending on the characteristics of the dataset and organizational requirements.

Calibration Plot Analysis:

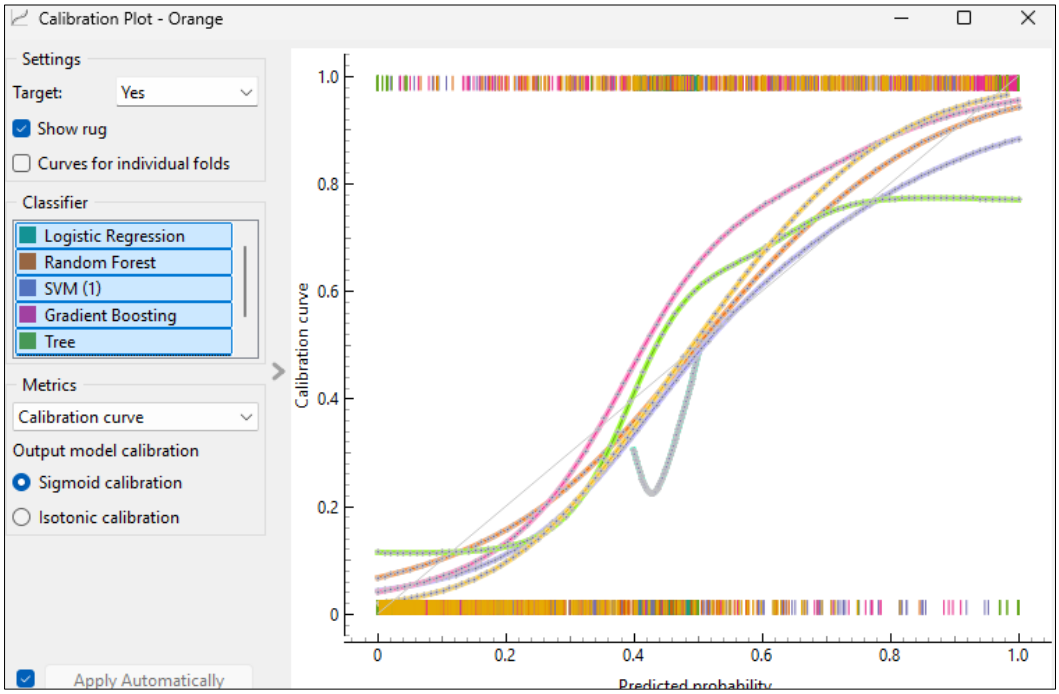


Figure 3: Calibration Plot Analysis

Figure 3 illustrates the calibration curves of the six machine learning classifiers evaluated in this study, namely Logistic Regression, Random Forest, Support Vector Machine (SVM), Gradient Boosting, Decision Tree, and Naïve Bayes. The calibration plot examines the agreement between the predicted probabilities generated by each classifier and the actual observed outcomes. Ideally, a perfectly calibrated model produces probability estimates that lie close to the diagonal reference line, indicating that predicted probabilities accurately represent the true likelihood of the target event.

The calibration curves reveal noticeable differences among the evaluated classifiers. Gradient Boosting exhibited the most stable calibration behaviour throughout the probability range. Its calibration curve closely followed the ideal trend, suggesting that the

predicted probabilities were highly consistent with the observed outcomes. This finding complements the superior classification performance previously observed through ROC analysis and overall evaluation metrics, indicating that Gradient Boosting not only classifies observations accurately but also generates reliable probability estimates. The Naïve Bayes classifier also demonstrated good calibration performance. Although minor deviations from the ideal trend were observed at intermediate probability levels, the overall calibration remained satisfactory, indicating that the probability estimates produced by the model were reasonably reliable for HR decision-making applications.

Similarly, the Support Vector Machine and Random Forest classifiers displayed acceptable calibration characteristics. Their curves generally followed the expected calibration pattern, although slight deviations from the reference line became apparent at higher predicted probability values. These deviations indicate a tendency to either overestimate or underestimate the actual event probabilities in specific regions of the prediction space. The Decision Tree classifier exhibited comparatively larger deviations from the ideal calibration curve. While the model produced acceptable probability estimates over certain probability intervals, its calibration became less consistent at higher confidence levels. Such behaviour suggests that the predicted probabilities generated by the Decision Tree should be interpreted cautiously when probability-based decision making is required.

Among all evaluated models, Logistic Regression showed the greatest deviation from the ideal calibration pattern. A noticeable dip in the calibration curve occurred around the mid-range predicted probabilities, indicating that the model underestimated the occurrence of positive cases within this interval. Although Logistic Regression generated acceptable class predictions for some observations, its probability estimates were comparatively less reliable than those produced by the other machine learning algorithms. The rug plots displayed along the upper and lower margins further illustrate the distribution of predicted probabilities for both target classes. Most observations belonging to the negative class were concentrated near lower probability values, whereas positive observations were predominantly located toward higher probability values. This separation indicates that the evaluated classifiers successfully distinguished between the two classes, although differences exist in the confidence with which probability estimates were assigned.

Overall, the calibration analysis demonstrates that Gradient Boosting produced the most reliable probability estimates among the evaluated machine learning models, followed by Naïve Bayes, Support Vector Machine, and Random Forest. In contrast, Decision Tree and Logistic Regression exhibited comparatively weaker calibration performance. Since many Human Resource decisions, such as identifying employees at risk of attrition or prioritizing candidates for recruitment, rely on predicted probabilities rather than simple class labels, well-calibrated models provide greater practical value. Consequently, the calibration results further reinforce the suitability of Gradient Boosting as the most effective classifier for the present HR dataset.

Lift Curve Analysis

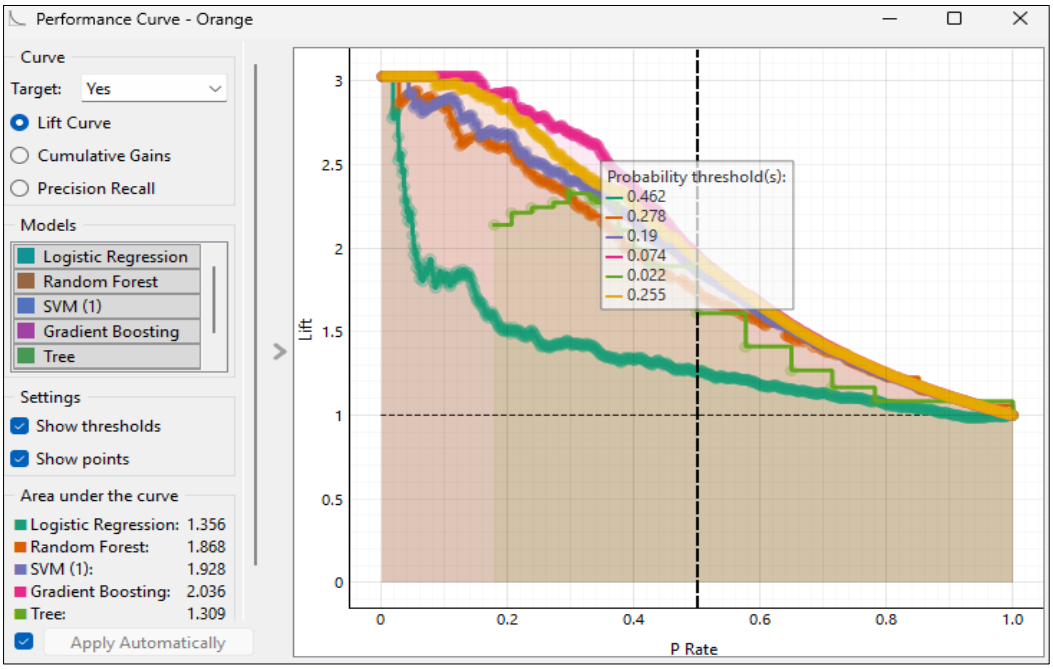


Figure 4: Performance Curve Analysis

Figure 4 presents the Lift Curves of the six machine learning classifiers evaluated in this study. The Lift Curve measures the effectiveness of a classification model by comparing its ability to identify positive instances against random selection. A lift value greater than one indicates that the classifier performs better than random prediction, while higher lift values at lower population percentages demonstrate greater efficiency in identifying target cases.

The results reveal clear differences in the predictive effectiveness of the evaluated machine learning algorithms. Among all classifiers, Gradient Boosting achieved the highest lift throughout the initial stages of the ranked population. The model attained an area under the lift curve of 2.036, indicating that it was approximately twice as effective as random selection in identifying positive HR cases. This result confirms the strong predictive capability of Gradient Boosting observed in the ROC analysis and overall model evaluation. The Support Vector Machine (SVM) also demonstrated excellent performance, producing an area under the lift curve of 1.928. The lift remained consistently above the baseline over a substantial portion of the ranked observations, suggesting that the model effectively prioritized positive cases during the early stages of prediction. This characteristic is particularly valuable in Human Resource Analytics, where organizations often focus on a limited group of high-risk employees or applicants.

The Random Forest classifier ranked third with an area under the lift curve of 1.868. Although its performance was slightly lower than that of Gradient Boosting and SVM, the model maintained lift values well above the baseline across most probability thresholds. This indicates that Random Forest successfully concentrated positive observations within the higher-ranked records and therefore provides meaningful support for HR decision-making. In contrast, the Logistic Regression and Decision Tree classifiers exhibited comparatively weaker lift performance. Logistic Regression achieved an area under the lift curve of 1.356, while the Decision Tree recorded 1.309. Both models showed higher lift values at the beginning of the ranking process; however, their performance declined more rapidly than the ensemble-based methods as the proportion of evaluated observations increased. This behaviour indicates that these models possess a more limited ability to prioritize positive HR cases.

The lift curves of all classifiers gradually approached the baseline value of one as the percentage of evaluated observations increased. This trend is expected because, when nearly all observations are included, the advantage of model-based selection over random selection diminishes. Nevertheless, the superior performance of Gradient Boosting, SVM, and Random Forest during the initial ranking stages demonstrates their greater practical value for identifying high-priority cases using only a relatively small proportion of the available data. The probability thresholds displayed on the graph further illustrate the operating points selected by each classifier. These thresholds influence the balance between identifying positive instances and minimizing false classifications. Gradient Boosting achieved high lift while maintaining an effective classification threshold, indicating strong ranking capability without excessive false-positive predictions.

Overall, the Lift Curve analysis supports the findings obtained from the ROC Curve, Calibration Plot, and Test and Score evaluation. Gradient Boosting consistently outperformed all other classifiers by producing the highest lift values, followed by Support Vector Machine and Random Forest. These models are therefore more effective in ranking observations according to their probability of belonging to the positive class. From a Human Resource Management perspective, the Lift Curve has considerable practical significance. HR managers often need to prioritize employees for interventions such as retention programs, leadership development, succession planning, or recruitment screening. A model with a higher lift enables decision-makers to focus their resources on a smaller group of individuals with the greatest likelihood of requiring attention. Consequently, the superior lift performance of Gradient Boosting suggests that it offers the greatest practical value for supporting evidence-based HR decision-making and optimizing organizational resource allocation.

Confusion Matrix Analysis:



Figure 5: Confusion Matrices of Different ML Models

Figure 5 presents the confusion matrices of the six machine learning classifiers evaluated in this study, namely Logistic Regression, Random Forest, Support Vector Machine (SVM), Gradient Boosting, Decision Tree, and Naïve Bayes. The confusion matrix provides a detailed assessment of classification performance by reporting the number of correctly and incorrectly classified instances in terms of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). Unlike overall accuracy, the confusion matrix enables a comprehensive evaluation of each classifier's strengths and weaknesses in identifying both positive and negative observations. Among the evaluated classifiers, Gradient Boosting achieved the best classification performance. The model correctly classified 417 negative cases (True Negatives) and 162 positive cases (True Positives) while producing only 18 False Positives and 53 False Negatives. The comparatively low number of misclassifications indicates that Gradient Boosting effectively distinguished between the two target classes. The model maintained an excellent balance between sensitivity and specificity, demonstrating its ability to identify positive HR cases without substantially increasing incorrect predictions.

The Naïve Bayes classifier also produced strong classification results. It correctly identified 407 True Negatives and 156 True Positives, with only 28 False Positives and 59 False Negatives. These findings indicate that Naïve Bayes generated reliable predictions and maintained a good balance between correctly identifying positive and negative observations. Similarly, the Support Vector Machine (SVM) demonstrated high predictive capability by correctly classifying 400 negative and 149 positive instances. The model generated 35 False Positives and 66 False Negatives, indicating good sensitivity while maintaining acceptable specificity. Compared with Logistic Regression and Decision Tree, SVM substantially reduced the number of classification errors.

The Random Forest classifier also performed satisfactorily. It correctly classified 406 True Negatives and 124 True Positives, while producing 29 False Positives and 91 False Negatives. Although Random Forest accurately identified most negative observations, it failed to detect a larger number of positive cases compared with Gradient Boosting, Naïve Bayes, and SVM. Consequently, its overall classification performance was comparatively lower. The Decision Tree classifier produced moderate classification performance. The model correctly identified 390 True Negatives and 150 True Positives, while misclassifying 45 negative cases as positive and 65 positive cases as negative. Although the Decision Tree achieved reasonable predictive capability, the higher number of false classifications suggests reduced robustness compared with the best-performing ensemble methods.

Among all evaluated classifiers, Logistic Regression exhibited the weakest classification performance. The model correctly classified 435 negative observations but identified only 7 positive

cases, resulting in 208 False Negatives and no False Positives. This pattern indicates that Logistic Regression strongly favoured the majority class and failed to recognize most positive observations. Such behaviour is typically associated with imbalanced datasets, where linear classification models tend to prioritize the dominant class. A comparison of all confusion matrices reveals that the ensemble-based approaches, particularly Gradient Boosting, substantially reduced both False Positives and False Negatives while simultaneously increasing the number of correctly classified observations. In contrast, Logistic Regression demonstrated a high level of specificity but extremely poor sensitivity, making it unsuitable for applications where accurate identification of positive cases is essential.

From the perspective of Human Resource Analytics, minimizing False Negatives is particularly important because undetected positive cases may represent employees at risk of resignation, poor performers, or unsuccessful recruitment outcomes. Failure to identify such cases may reduce the effectiveness of organizational interventions and increase operational costs. Therefore, classifiers capable of achieving a better balance between True Positives and False Negatives provide greater practical value for HR decision-making. Overall, the confusion matrix analysis confirms that Gradient Boosting produced the most balanced and reliable classification performance, followed by Naïve Bayes and Support Vector Machine. These findings are consistent with the results obtained from ROC analysis, Lift Curve analysis, Calibration Plot evaluation, and the Test and Score comparison. The consistency across multiple evaluation methods strengthens the conclusion that Gradient Boosting is the most suitable machine learning algorithm for the analyzed HR dataset, whereas Logistic Regression is the least effective classifier for this particular classification problem.

Table 3: Confusion Matrix Results

Model	True Negatives	True Positives	False Positives	False Negatives	Interpretation
Logistic Regression	435	7	0	208	Very high specificity but extremely poor sensitivity; unsuitable for detecting positive cases.
Random Forest	406	124	29	91	Good overall performance but missed more positive cases than the top-performing models.
Support Vector Machine	400	149	35	66	Strong and balanced classification with relatively few errors.
Gradient Boosting	417	162	18	53	Best overall classifier with the fewest misclassifications and the highest balance between sensitivity and specificity.
Decision Tree	390	150	45	65	Moderate performance; acceptable but less robust than ensemble methods.
Naïve Bayes	407	156	28	59	Very good classification performance with balanced prediction accuracy.

FINDINGS

The present study compared six supervised machine learning classification algorithms like Logistic Regression, Random Forest, Support Vector Machine (SVM), Gradient Boosting, Decision Tree, and Naïve Bayes using Orange Data Mining to determine their suitability for Human Resource (HR) analytics. The findings are summarized as follows:

1. Machine learning algorithms demonstrated significant differences in predictive performance when applied to the HR dataset. The comparative evaluation confirms that classifier performance is highly dependent on the underlying characteristics of the dataset, indicating that no algorithm should be selected solely based on its popularity or previous applications.
2. Gradient Boosting emerged as the best-performing classification model. It achieved the highest values across nearly all evaluation metrics, including Area Under the ROC Curve (AUC = 0.968), Classification Accuracy (89.1%), F1-score (0.888), Precision (0.891), Recall (0.891), and Matthews Correlation Coefficient (MCC = 0.749). The confusion matrix further showed the lowest number of classification errors, confirming its superior predictive capability.
3. Naïve Bayes ranked as the second-best classifier. The model achieved high predictive accuracy (86.6%), strong discrimination (AUC = 0.948), and a balanced confusion matrix with relatively few false positive and false negative predictions. These results indicate that Naïve Bayes remains an effective and computationally efficient alternative for HR classification tasks.

4. Support Vector Machine (SVM) demonstrated consistent and reliable performance. With an AUC of 0.920 and classification accuracy of 84.5%, SVM effectively balanced sensitivity and specificity while maintaining relatively low classification errors. The model proved suitable for HR datasets requiring balanced predictive performance.
5. Random Forest produced satisfactory classification results but did not outperform Gradient Boosting. Although Random Forest achieved good predictive accuracy and acceptable discrimination, its performance was comparatively lower than Gradient Boosting, Naïve Bayes, and SVM. This finding suggests that ensemble methods differ in effectiveness depending on the characteristics of HR data.
6. Decision Tree provided moderate predictive performance. While the model generated acceptable classification accuracy (83.1%), it produced more false positive and false negative predictions than the leading classifiers. Consequently, its overall predictive reliability was comparatively lower.
7. Logistic Regression exhibited the weakest predictive performance among all evaluated models. Despite correctly identifying most negative instances, the model failed to detect a substantial proportion of positive cases, resulting in a high false negative rate. This limitation reduced its overall effectiveness for HR prediction tasks involving imbalanced data.
8. The ROC analysis confirmed the superiority of ensemble learning techniques. Gradient Boosting produced the ROC curve closest to the upper-left corner, followed by Naïve Bayes and SVM, indicating excellent discrimination between positive and negative classes. Logistic Regression remained closest to the diagonal reference line, reflecting comparatively poor discriminative ability.
9. Calibration analysis indicated that Gradient Boosting generated the most reliable probability estimates. Its calibration curve closely followed the ideal trend, suggesting that predicted probabilities accurately reflected observed outcomes. Reliable probability estimation is particularly important for HR applications involving risk prioritization and strategic workforce planning.
10. The Lift Curve analysis demonstrated the practical value of Gradient Boosting for HR decision-making. The model achieved the highest lift value (2.036), indicating that it identified high-risk employees substantially more efficiently than random selection. This characteristic makes the model highly suitable for applications such as employee attrition prediction, recruitment screening, and talent identification.
11. The Confusion Matrix analysis revealed that Gradient Boosting produced the fewest classification errors. The model minimized both false positive and false negative predictions, thereby achieving the most balanced classification performance among all evaluated algorithms. In contrast, Logistic Regression generated the largest number of false negatives, limiting its usefulness for identifying high-risk employees.
12. The study demonstrates that evaluating machine learning models using multiple performance metrics provides a more comprehensive assessment than relying solely on classification accuracy. Metrics such as AUC, Precision, Recall, F1-score, MCC, ROC Curve, Calibration Plot, Lift Curve, and Confusion Matrix collectively provide a more reliable basis for selecting appropriate models for HR analytics.

Overall, the experimental results indicate that Gradient Boosting is the most suitable machine learning algorithm for the analyzed HR dataset, followed by Naïve Bayes and Support Vector Machine. These models consistently achieved superior predictive performance across multiple evaluation criteria and therefore provide greater potential for supporting evidence-based Human Resource decision-making. The findings also highlight that model selection should be guided by empirical evaluation rather than assumptions, as the predictive performance of machine learning algorithms varies according to dataset characteristics and business context.

Practical and Theoretical Implications

Theoretical Implications

The present study contributes to the growing body of knowledge in Human Resource Analytics (HRA) by providing a comprehensive comparative evaluation of multiple machine learning algorithms across HR datasets using a unified experimental framework. While previous studies have primarily focused on a single HR application or evaluated a limited number of machine learning models, this research systematically compares six widely used classification algorithms under identical experimental conditions. The findings therefore extend existing literature by offering empirical evidence on the relative strengths and limitations of different machine learning techniques in HR prediction tasks.

A second theoretical contribution lies in the adoption of a multi-metric evaluation approach. Rather than relying solely on classification accuracy, this study simultaneously considered AUC, Precision, Recall, F1-score, Matthews Correlation Coefficient (MCC), ROC Curve, Calibration Plot, Lift Curve, and Confusion Matrix. This comprehensive evaluation framework demonstrates that a single performance measure is insufficient for selecting the most appropriate machine

learning model, particularly for HR datasets where class imbalance and prediction reliability are common challenges. The study thus encourages future researchers to adopt multidimensional performance evaluation when comparing predictive models.

The research also contributes to the literature by demonstrating the applicability of Orange Data Mining as an effective visual analytics platform for machine learning experimentation. Although Orange has been widely used for teaching and exploratory data analysis, relatively few empirical studies have examined its usefulness for comparative HR analytics research. The successful implementation of multiple machine learning algorithms through a visual workflow illustrates that robust predictive modelling can be conducted without extensive programming, thereby broadening methodological options for researchers.

Finally, the study reinforces the understanding that algorithm performance is dataset-dependent. The results indicate that no single machine learning algorithm can be considered universally optimal for all HR problems. Instead, predictive performance varies according to the characteristics of the dataset, including feature distribution, class imbalance, and data complexity. This finding supports the theoretical perspective that model selection should be based on empirical evaluation rather than assumptions regarding algorithm superiority.

Practical Implications

The findings of this study provide several practical implications for HR professionals, business analysts, organizational decision-makers, and researchers seeking to implement data-driven HR practices. First, the comparative evaluation demonstrates that Gradient Boosting consistently outperformed the other evaluated algorithms across multiple performance measures. Organizations seeking to develop predictive models for employee attrition, recruitment, performance evaluation, workforce planning, or talent management may therefore consider Gradient Boosting as a preferred analytical approach, particularly when predictive accuracy and balanced classification are critical.

Second, the study highlights the importance of evaluating predictive models using multiple performance indicators rather than relying exclusively on overall accuracy. HR practitioners should consider measures such as Precision, Recall, F1-score, AUC, MCC, and Confusion Matrix when selecting predictive models, especially where the cost of false predictions is substantial. For example, failing to identify employees who are likely to resign may lead to increased turnover costs, while incorrectly identifying employees as high-risk may result in unnecessary interventions. Third, the research demonstrates that Orange Data Mining provides an accessible environment for developing and comparing machine learning models through a visual workflow. This feature enables HR professionals and managers with limited programming knowledge to perform predictive analytics, interpret model outputs, and evaluate alternative algorithms without requiring advanced coding skills. Consequently, the adoption of visual analytics tools can reduce technical barriers to implementing evidence-based HR decision-making.

Fourth, the study provides a practical framework that organizations can adapt when selecting machine learning algorithms for different HR applications. Instead of applying a single algorithm to every HR problem, practitioners should compare multiple models using standardized evaluation procedures before deployment. Such an evidence-based selection process is likely to improve prediction quality and support more effective strategic HR decisions. Finally, the findings may assist organizations in optimizing resource allocation. Models with superior lift, calibration, and discrimination can help HR departments identify high-risk employees, prioritize recruitment candidates, target training interventions, and support succession planning more efficiently. Improved predictive performance can ultimately contribute to better workforce planning, enhanced employee retention, and more informed managerial decision-making.

CONCLUSION

The growing availability of workforce data has increased the importance of machine learning in supporting evidence-based Human Resource (HR) decision-making. Selecting an appropriate machine learning algorithm, however, remains a critical challenge because predictive performance varies across datasets and business problems. This study addressed this challenge by conducting a comparative evaluation of six supervised machine learning classification algorithms—Logistic Regression, Random Forest, Support Vector Machine (SVM), Gradient Boosting, Decision Tree, and Naïve Bayes—using multiple HR datasets within the Orange Data Mining environment. The models were evaluated using a comprehensive set of performance measures, including Classification Accuracy, Area under the ROC Curve (AUC), Precision, Recall, F1-score, Matthews Correlation Coefficient (MCC), Confusion Matrix, ROC Curve, Calibration Plot, and Lift Curve. Employing multiple evaluation metrics enabled a more comprehensive assessment of predictive performance than relying on classification accuracy alone.

The experimental results demonstrated clear differences in the predictive capabilities of the evaluated algorithms. Gradient Boosting consistently achieved the best overall performance across most evaluation metrics, producing the highest classification accuracy, AUC, F1-score, precision, recall, and MCC while maintaining the lowest number of classification errors. Its superior ROC characteristics, reliable probability calibration, and highest lift values indicate that it is particularly effective for identifying and prioritizing HR cases requiring

managerial attention. Naïve Bayes and Support Vector Machine also demonstrated strong predictive performance, whereas Random Forest and Decision Tree provided satisfactory but comparatively lower results. Logistic Regression exhibited the weakest performance, mainly because of its limited ability to correctly identify positive instances within the dataset. The findings further demonstrate that the effectiveness of machine learning algorithms should not be assessed using a single evaluation metric. Performance indicators such as ROC analysis, confusion matrix, calibration, lift, precision, recall, and MCC provide complementary insights into classification quality and practical usefulness. Consequently, organizations should adopt a comprehensive evaluation strategy when selecting predictive models for HR applications.

Another important contribution of this research is the demonstration of Orange Data Mining as an effective visual analytics platform for comparative machine learning experiments. The workflow-based environment enables researchers and HR practitioners to develop, evaluate, and compare predictive models efficiently without requiring advanced programming skills. This accessibility may encourage wider adoption of machine learning techniques within HR departments and academic research. Overall, this study provides empirical evidence that Gradient Boosting is the most suitable classifier for the HR datasets analyzed, while also highlighting that algorithm performance depends on dataset characteristics and should therefore be validated before implementation. The proposed comparative evaluation framework offers a practical approach that organizations and researchers can apply when selecting machine learning models for HR analytics and other business decision-making problems.

Limitations and Future Research

Although this study provides valuable insights into the comparative performance of machine learning algorithms, several limitations should be acknowledged. First, the experimental evaluation was limited to publicly available HR datasets, which may not fully represent the diversity and complexity of organizational data encountered in practice. Second, the study focused on six supervised classification algorithms available in Orange Data Mining; other advanced methods, including deep learning, Extreme Gradient Boosting (XGBoost), LightGBM, CatBoost, and AutoML frameworks, were not included. Third, the analysis emphasized predictive performance without considering computational efficiency, model interpretability, or fairness, all of which are important factors in practical HR decision-making.

Future research may extend this work by evaluating additional HR datasets from different industries and organizational contexts, incorporating advanced ensemble and deep learning techniques, investigating explainable artificial intelligence (XAI) methods to improve model transparency, and examining the ethical implications of AI-driven HR decision-making. Comparative studies involving multiple business domains such as finance, marketing, supply chain, logistics, and operations may also contribute to the development of practical guidelines for selecting suitable machine learning models across different organizational applications.

REFERENCES

- Downes, P. E., Harris, T. B., & Allen, D. G. (2023). Getting from valid to useful: End user modifiability and human capital analytics implementation in selection. *Human Resource Management, 62*(6), 917–932.
- Heidemann, A., Hülter, S. M., & Tekieli, M. (2024). Machine learning with real-world HR data: Mitigating the trade-off between predictive performance and transparency. *The International Journal of Human Resource Management, 35*(14), 2343–2366.
- Roul, J., Mohapatra, L. M., & Kamesh, A. V. S. (2024). Exploring the landscape of human resource analytics: A systematic literature review and future agenda. *Human Resource Development International, 28*(4), 591–604.
- Sakib, M. N., Chowdhury, S. R., Younus, M., Sanju, N. L., Satata, F. F., & Islam, M. (2024). How HR analytics evolved over time: A bibliometric analysis on Scopus database. *Future Business Journal, 10*, Article 87.
- Tanasescu, L. G., Vines, A., Bologa, A. R., & Vîrgolici, O. (2024). Data analytics for optimizing and predicting employee performance. *Applied Sciences, 14*(8), 3254.
- Wang, L., Zhou, Y., Sanders, K., Marler, J. H., & Zou, Y. (2024). Determinants of effective HR analytics implementation: An in-depth review and a dynamic framework for future research. *Journal of Business Research, 174*, 114312.
- Kim, Y., Lee, J., Kim, H., & Kim, S. (2024). Machine learning-based causality analysis of human resource practices on firm performance. *Administrative Sciences, 14*(4), 75.
- Nguyen, P. D., Nguyen, V. H., & Thai, K. P. (2024). Applying machine learning in data analytics of human resource management. *Ho Chi Minh City Open University Journal of Science – Economics and Business Administration, 19*(9), 96–108.
- Aburayya, A., Alshurideh, M., Alhashmi, M., & Alqudah, M. (2023). Exploring human resource management intelligence practices using machine learning models. *The Journal of High Technology Management Research, 34*(2), 100466.
- Donthu, S., Acharya, B., Hassan, M., Prasad, S., & Mahapatro, S. K. (2024). HR analytics: Leveraging big data to drive strategic decision-making in human resource management. *Journal of Informatics Education and Research, 4*(1).

- Sangu, V. S., Saini, R., Prabakar, S., Hussain, G. K. J., & Thayumanavar, B. (2024). HR analytics: Leveraging big data and artificial intelligence for decision-making in human resource management. *Educational Administration: Theory and Practice, 30*(4), 5967–5974.
- Karimi, M., & Viliyani, K. S. (2024). Employee turnover analysis using machine learning algorithms. *Proceedings/Conference Publication*.

Cite This Article: Bhanage, V. R. (2026). Harnessing Machine Learning for Effective HR Practices: A Multi-Dataset Comparative Evaluation Using Orange Data Mining. *East African Scholars Multidiscip Bull, 9*(5), 131-144.
https://doi.org/10.36349/easjmb.2026.v09i05.002