Journal of Economic Sciences
Volume 5, Issue 1
https://www.jesciences.com

Research and Development Employees Attrition: A Machine Learning Based Exploration

Muhammad Atif Sattar[1,*], Muhammad Waseem[2], Khuram Shafi[2], Samina Nawab[2]

Affiliations
1. Kunming University of Science and Technology, Faculty of Management and Economics, China
2. Department of Management Sciences, COMSATS University Islamabad, Wah Campus, Pakistan

*Corresponding Author Email: 20242052@kust.edu.cn

Timeline
Received: Feb 26, 2026
Revised: May 29, 2026
Accepted: Jun 06, 2026
Published: Jun 29, 2026

DOI
https://doi.org/10.55603/jes.v5i1.a9

Abstract
Technical employees are the most vital asset of a company's R&D department to remain sustainable, their attrition is a significant concern from the company's perspective. Despite that technical employee attrition transpires for numerous causes, making it challenging for the predictive high authorities of the R&D department to anticipate these indicators in advance. The attrition of R&D employees affects knowledge retention, project continuity, and the pace of innovation. This research proposes a methodology for predicting attrition among R&D employees, enabling proactive workforce management rather than relying on retrospective metrics. This study developed a predictive Machine learning classification model utilizing 30 features that influence R&D employees attrition, derived from the IBM dataset, comprising 961 records. Consequently, five Machine learning classification predictive models- Decision Tree, Random Forest, Support Vector Machine, Logistic Regression, and K-Nearest Neighbour were developed and their performance assessed. Moreover, the analysis of variables affecting R&D employees' attrition revealed that stock option level, job involvement, job level, and job satisfaction were the most significant contributors. This research helps the organizations and managers to identify the attrition factors by using the machine learning models and develop sustainable strategies on retention.

Keywords: *R&D Employees, Attrition, Machine Learning*

JEL Classification: *F1, F12, F60*

1. Introduction

Organizational competitiveness and innovation depend on its essential Research and Development (R&D) department. Its primary role is to promote the development of new technologies, services, processes, and products while also enhancing current ones. Although the size and structure of R&D departments differ greatly across industries (Chou et al., 2023) . They all aim to improve an organization's knowledge database (Luo & Zhang, 2021), foster innovation (Tikas, 2023), and assurance long-term sustainability (Kumari et al., 2022) by staying abreast of or spearheading industry innovation.

Usually, spearheading innovation is unattainable without the R&D department's experienced employees like engineers, scientists, and analysts (Tikas, 2023) . These professionals drive innovation, enhance product quality, prevent knowledge, rapidly adapt market trends, mitigate potential risks (Bai et al., 2023), and exceed customer satisfaction to keep the organization competitive in the market (Wu et al., 2019) . R&D employees also acts as internal technical expert resources, providing advice on intricate technical matters and contributing product insights to other departments such as marketing and operations (Gupta & Wilemon, 1990; Szopik-Depczyńska et al., 2024) . This shows that the organization's performance and innovation hinge on the R&D department's highly specialized skill set employees.

Organizations face a laborious, complex, and challenging task in recruiting and retaining skilled R&D employees (Schmitz et al., 2021). To recruit these professionals, different HR activities such as job

advertisements, assessments, interviews, and onboarding are necessary; effective recruitment techniques are essential to ensure that personnel align with company culture and objectives (Chandler, 2023; Gadekar & Hiwarkar, 2023) . Conversely, organizations employ several effective retention strategies, such as competitive pay, flexible schedules, remote work opportunities, mentorship programs, training and development initiatives, and rewards to retain current R&D employees. Recruitment and retention strategies are crucial; if an organization fails to establish successful recruitment and retention strategies, it will result in the attrition of personnel, adversely impacting knowledge retention, project continuity, and the pace of innovation.

A market analysis reveals that the demand for qualified R&D employees has surged over the past decade, resulting in an attrition rate of 18% (Potocnik et al., 2021) . Consequently, businesses have become more concerned about the turnover of R&D personnel and are striving to fill these jobs with suitable candidates. The rationale is that technical employees possess specialized knowledge and abilities that are challenging to find in the labor market. The loss of such a person adversely affects the organization at various levels, including competitive advantage, project continuity, product quality deterioration, and disruption of the innovation pace.

Consequently, organizations emphasize on identifying factors influencing the attrition of R&D employees and prioritize the development of effective retention strategies to reduce the attrition rate. A number of studies of studies suggesting various attrition factors by traditional approaches (Ahn et al., 2020). Such as, Kiazad et al. (2024) mentioned that personal circumstances, career advancement opportunities, and the workplace environment are the dominant factors by using thematic analysis. Pirrolas & Correia (2022) adopted empirical study review and identified responsibility, leadership, schedule flexibility, recognition, and wage are the key factors of attrition. Singh et al. (2022) used hierarchical regression analysis and mentioned that ineffective top management are the vital factor in attrition. Machine learning (ML) prediction techniques have significantly advanced in the last decade. These developments can be useful for practitioners and organizational researchers.

ML is the subset of AI, has emerged a popular technique for prediction in all fields particularly in business field (Qutub et al., 2021) . ML helps businesses to analyze real-time business data ascertain hidden insights, increase predicting skills, and expedite informed decision-making (Gadekar & Hiwarkar, 2023; Saini & Saini, 2024) . Therefore, ML can be suitable in this research to analyze the R&D skilled employees data that can help to depict hidden pattern of attrition and to identify the R&D employees attrition factors. It can also help organizational managers to predict such employees who are at risk of attrition and to develop preemptive retention strategy.

Several studies used the ML Algorithms to predict the employee attrition and was applied on across various department employees of organization. However, R&D departments include distinct attributes that may result in varying attrition trends. Research explicitly targeting the R&D sector is scarce. Numerous research pointed out that the attrition rate in R&D department is higher than others department (Chang et al., 2008; Cordero et al., 1994; Grabner et al., 2024) . This offers a distinctive opportunity to utilize advanced predictive ML Algorithms on R&D department employees attrition. This study seeks to address that gap by offering data driven insights about R&D employees attrition, thereby assisting organizations in enhancing their skilled employees management strategies.

This research based on three main objectives. First, identify the key attrition factors in the R&D department. Second, predicting the high-risk employees attrition model. The third objective is to find out the best factors contributing to ML classification algorithm to achieve a high accuracy score. This research will help the managers and organization to enhance the company's sustainable management and productivity by proactively identifying employee attrition and fostering a positive organizational culture.

2. Literature Review

This section glances at the literature that is relevant to our research study. The relevant literature is founded on an overview of previous research findings and applied methodologies for employee attrition prediction. The advanced approaches were preferred for literature review.

2.1. R&D employees attrition

Organizations need technical employees in the R&D department to sustain technological competitiveness. Due to the high demand for technical employees in the competitive business environment, organizations face recruitment and retention problems with technical employees (Potocnik et al., 2021) . Therefore, recruiting and retaining skilled employees has become vital to maintaining a sustainable business. Organizations must focus on effective strategies for recruiting and retaining skilled employees in developing performance. Failure to do so will result in a slow pace of innovation and ongoing project continuity.

In particular, skilled employee attrition in R&D department affects organizational performance as a major concern in all organizations. Whenever a skilled employee leaves, the ongoing projects become delayed and hire a new skilled employee is too costly and laborious to trained (Qutub et al., 2021) . Additionally there also exist a risk of core knowledge leaks resulting from employee attrition (Chung et al., 2023) . Consequently, organizations should minimize the attrition rates to continue ongoing projects and remain competitive. Therefore, managers and leaders should continuously identify the employees attrition factors and promptly take action to secure project discontinuity and innovation pace. Beforehand identifying high-risk employees leads to establish earlier retention strategy and prevent to hire new employees cost.

In last decade, various researchers have adopted conventional methods to predict R&D department employee attrition such as regression, thematic and correlation analysis. Kiazad et al. (2024) used the thematic analysis and identified personal circumstances, career advancement opportunities, and the workplace environment attrition factors of the science and technology employees. Pirrolas & Correia (2022) followed empirical study review technique and pinpointed responsibility, leadership, schedule flexibility, recognition, and wage attrition factors of skilled employees. Singh et al. (2022) followed hierarchical regression analysis and revealed that ineffective top management are the vital factor in attrition.

2.2. Approaches of employees attrition prediction

Prediction methodologies for employees attrition have been in a constant state of flux. The current statistical analysis methods has the drawback of performing very poorly when analyzing data with complicated or unusual patterns, and it frequently fails to have a high enough model prediction accuracy. To get over this restriction, additional research is now being done to improve the accuracy of employees attrition predictions by exploring more complex data science techniques, such ML.

El-Rayes et al. (2020) presented predicting employee attrition model using three ML classification algorithms models such as, Random Forest (RF), Decision Tree (DT) and Gradient Boosting Tree (GB) on Glassdoor's online data. RF found the best ML classification algorithm to predict the probability of an employee churn with 75% accuracy score. Company culture, senior management and compensation are the vital factors of employee's attrition.

Qutub et al. (2021) used IBM dataset from Kaggle website that contains 35 features and collected from 1470 employees. Adaboost (AB) Model, DT, GB, RF Regressor and Logistic Regressor (LR) classifier models was applied on the data to predict an employee's attrition probability. For evaluation of classifier algorithms accuracy, AUC and recall metrices was used and found Logistic model the best among others with 88.43% accuracy score.

Raza et al. (2022) presented Extra trees (ET) classifier algorithm performed well with 92% highest accuracy score. While, the others classifier algorithms like Support vector machine (SVM), DT and LR showed low accuracy score to predict the high-risk employee attrition. The total four classification was performed on IBM data set with 35 features of 1470 employees. On the same IBM dataset, Pratt et al. (2021) utilized six different ML algorithms like Gaussian NB, KNN, DT, SVM, LR and RF. RF performed best with 85.12% accuracy score.

The deep learning (DL) algorithms was applied to get the best accuracy score on the IBM dataset in the study of Al-Darraji et al. (2021), that is also a type of AI. The DL algorithm evidenced the 91% accuracy score. In another study, Chung et al. (2023) utilized eight ML classification algorithms, including XG Boost, RF, LR, Artificial Neural Network (ANN), SVM and Ensemble models on the IBM dataset downloaded from Kaggle website. Of the eight algorithms, Ensemble algorithm was prominent to predict the high-risk employee attrition with a 97% accuracy score. In the Ensemble algorithm RF and ANN was the base algorithm and LR was meta algorithm.

Using machine learning techniques on the employees attrition prediction previous research have reported significant results, numerous significant limitations remain. First, most studies develop the prediction models on the overall organization without examining R&D personnel, despite the pivotal role of retaining skilled research employees. Second, the studies rely on IBM dataset, which may limit the generalizability of the outcomes attributable to limited representation of organization environment and its simulated nature. Third, the research used advanced models to achieve the high predictive accuracy, while few studies emphasizes on model explainability and interpretability, which are vital for practical application of HR decision-making. Furthermore, methodological discrepancies related to evaluation, imbalancing dataset and feature engineering make comparison across studies challenging. Therefore, this study fulfill the gap for department-specific, practically applicable and more explainable attrition prediction models.

3. Methodology

This section articulates the methodology of this study for predicting the R&D employees attrition. The first subsection explains the data preprocessing techniques. The ML algorithms to predict the R&D employees attrition is deliberated in subsection two. The given below figure 1 is the methodology of employees prediction.

3.1 Data Preprocessing

Data processing technique is used to convert the raw data into useful and clear format that plays vital role in prediction of R&D employees attrition. In this subsection, data cleaning, handling missing information, data duplications, scaling, data balancing, encoding, feature engineering, and standardization data are the crucial techniques performed. These techniques enhance the quality of data and leads to accurate prediction of R&D employees attrition. The details of data preprocessing steps will be presented in this subsection.

3.1.1 Dataset

This research utilized the IBM dataset download from the Kaggle website. This dataset was selected following thorough evaluations of multiple relevant datasets, with the choice closely aligning with aims of this study. The dataset contains three categories of variables such as personal attributes, job related variables and career related variables, which are required to detect the R&D employees attrition. The dataset contains 961employees' information and each illustrated by 33 distinct independent variables. From 961 employees' information, 828 are categorized as non-attrition, while 133 are identified as attrition. Table 1 presents a comprehensive summary of the dataset's independent variables.

Figure 1: Methodology

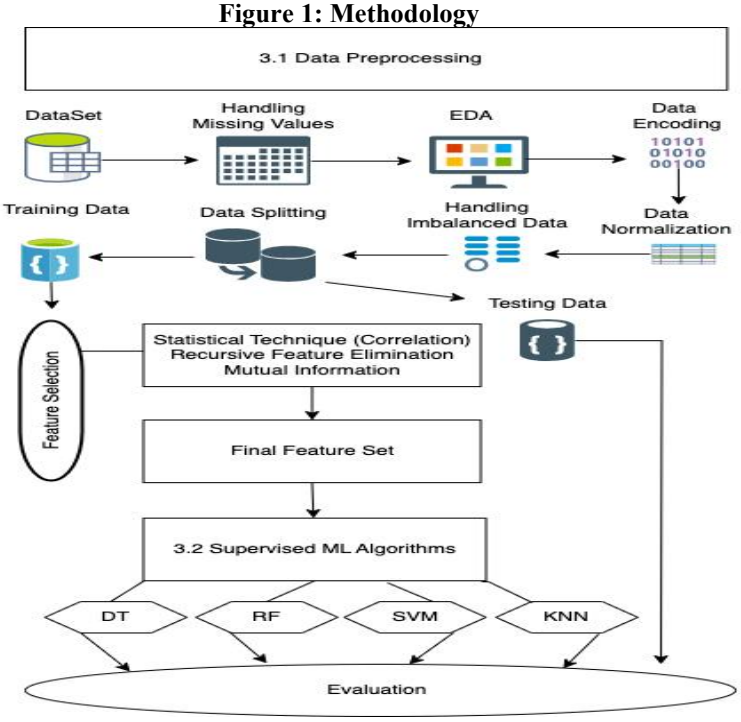


Table 1: list of Independent variables

Category	Variable Name	Type
Personal Attributes	Age	Integer
	Education	Object
	Education Field	Object
	Gender	Object
	Marital Status	Object
Job Related	Business Travel	Object
	Daily Rate	Integer
	Distance From Home	Integer
	Environmental Satisfaction	Integer
	Hourly Rate	Integer
	Job Involvement	Integer
	Job Level	Integer
	Job Role	Object
	Job Satisfaction	Integer
	Monthly Income	Integer
	Monthly Rate	Integer
	Over Time	Object
	Salary Hike	Integer
	Performance Rating	Integer
	Relationship Satisfaction	Integer
	Stock Option Level	Integer
	Training Time Last Year	Integer
	Work Life Balance	Integer

Career Related	Num Companies Worked	Integer
	Total Working Years	Integer
	Years at Company	Integer
	Years in Current Role	Integer
	Years Since Last Promotion	Integer
	Years with Current Manager	Integer

3.1.2 Handling Missing Information

Analysis is performed on the correct data format. Handling missing values and duplication data are vital steps of data analysis. Properly dealing with missing and duplication leads to model accuracy. If not done carefully, bias and misinterpretation problems can occur in the end results. No missing values and duplications were found in the IBM dataset. While, some variable contained only one information such as Employee count, Over 18, and Standard Hours. These variables were removed to keep the correct format of dataset.

3.1.3 Exploratory Data Analysis (EDA)

EDA was performed on the dataset to discover the hidden patterns and other characteristics. In EDA three techniques like, univariate, bivariate and multivariate was used.

3.1.3.1 Univariate Analysis

Figure 2 is about the description of categorical variables. Which showed that male employees are greater in comparison to female employees. Maximum employees are married and work as laboratory technicians or research scientists. Majority of employees have a background in the medical or life sciences and they travel rarely. Since plenty employees are classified as "NO" for overtime, that shows overtime is not a common practice.

Figure 2: Univariate Analysis for Categorical variables

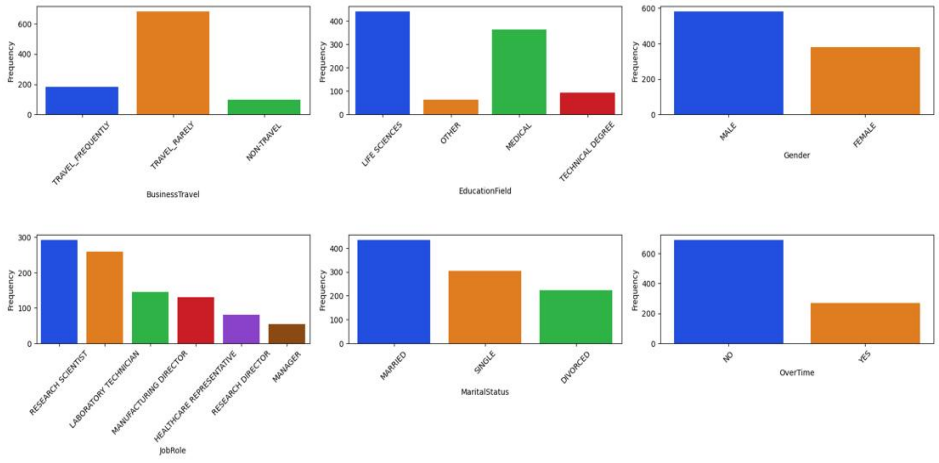
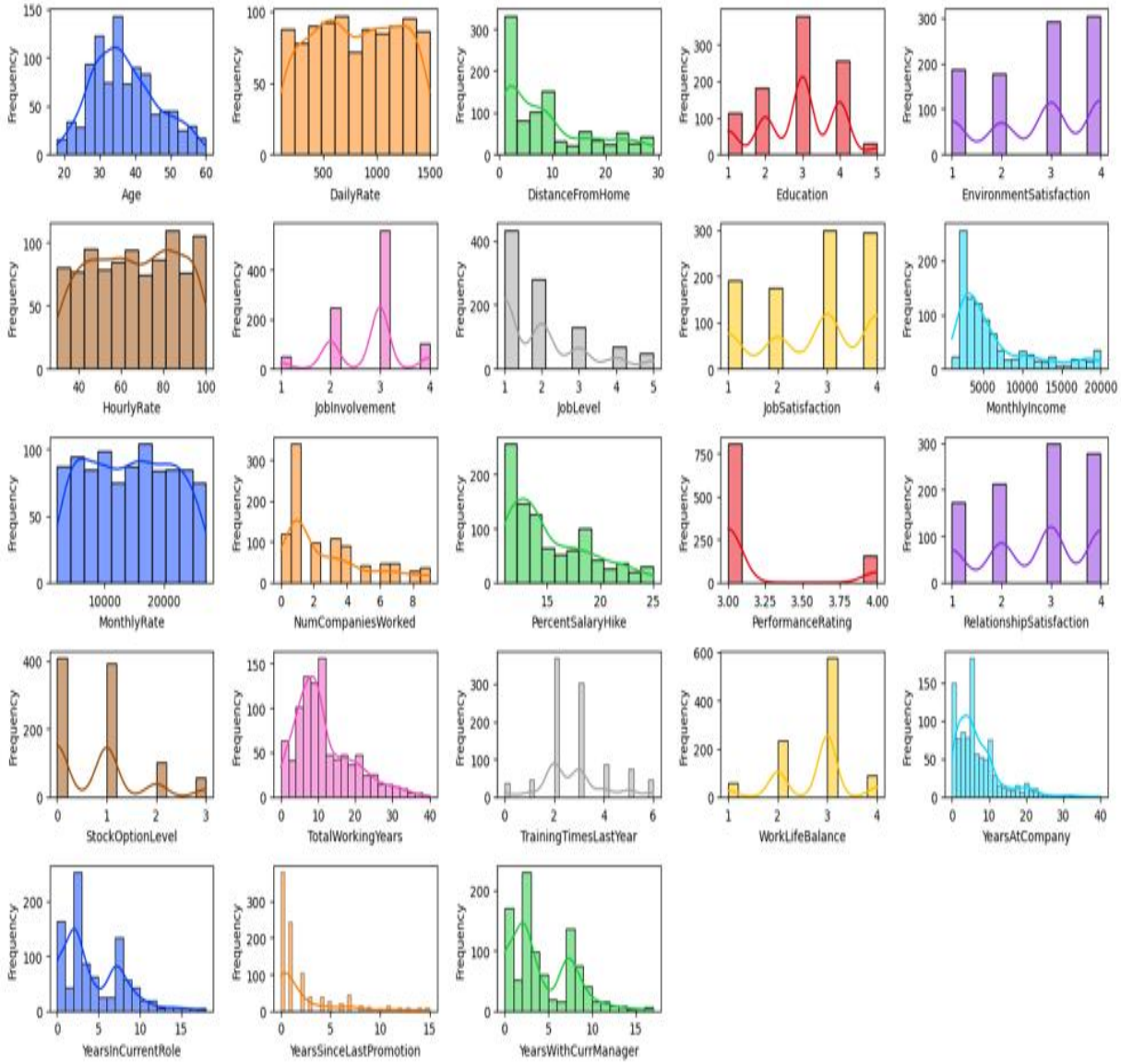


Figure 3 is about the univariate analysis of numerical variables that depicts the distribution shape of each variable. Age, Daily Rate, Education, Environment Satisfaction, Hourly Rate, Job Involvement, Job Satisfaction, Monthly Rate, and Relationship Satisfaction, are skewed. Moderately positive skewed are Distance from Home, Percent Salary Hike, Stock Option Level, Training Time Last Year, Years in Current Role, and Year with Current Manager, while Work Life Balance is moderately negative skewed. Job Level, Monthly Income, Number of Companies Worked, Performance Rating, Total Working Years, Years at Company, and Years since Last Promotion are highly positive skewed.

Figure 3: Univariate Analysis for Numerical variables



3.1.3.2 Bivariate Analysis

Analyzing two variables at a time is considered as bivariate analysis. Three bivariate analyses were performed in this subsection.

Bivariate analysis between numerical and categorical variables is shown in Figure 4. A high attrition was found among bachelor's and master's degree holders and high participation employees. The employees who got training 2 or 3 times last year and have already worked in one company are higher. These employees was new in the company and their job and environment satisfaction level was low and high level. The performance rating 3, work balance life rating 3 and none stock option level were greater in numbers.

Figure 5 is also a bivariate analysis between remaining numeric and categorical variables. This revealed that the higher probability of employees attrition was found in fewer promotions, younger employees (age 18 to 30), shorter in tenure, lower monthly income (6000 to 1500), stagnant role, and office distance from home greater than 10 km.

Figure 4:Bivariate Analysis for Numerical vs Categorical Variables

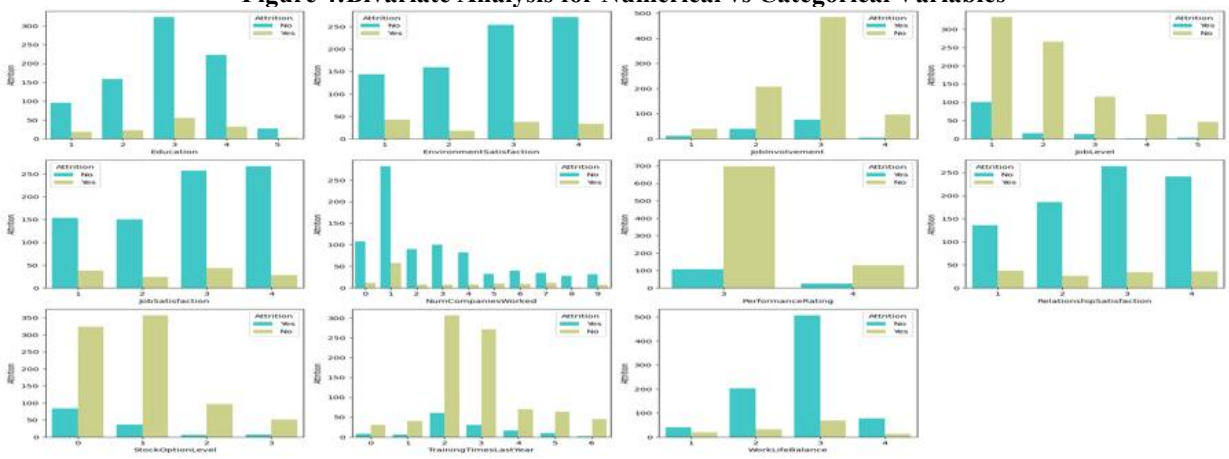


Figure 5: Bivariate Analysis for Numeric vs Categorical variable

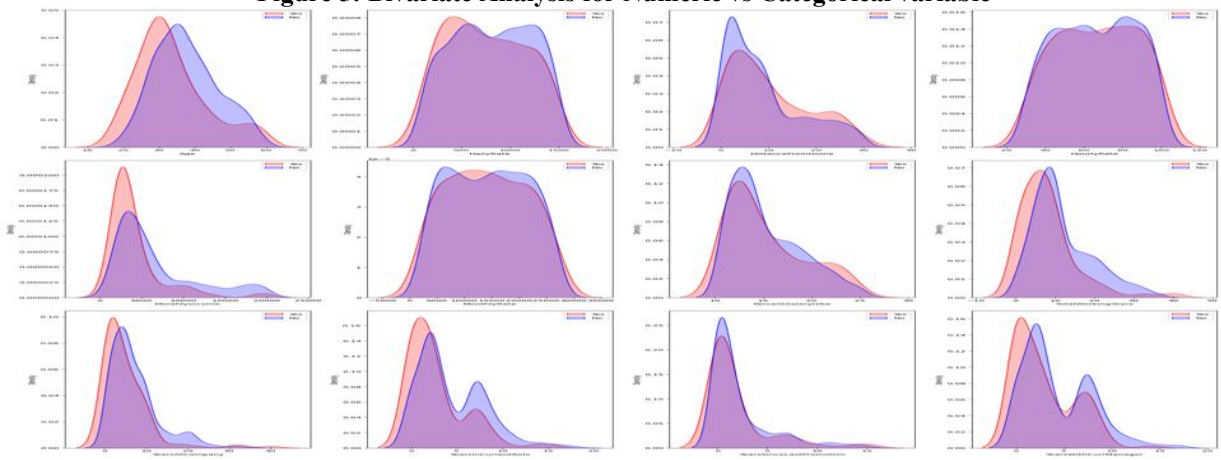
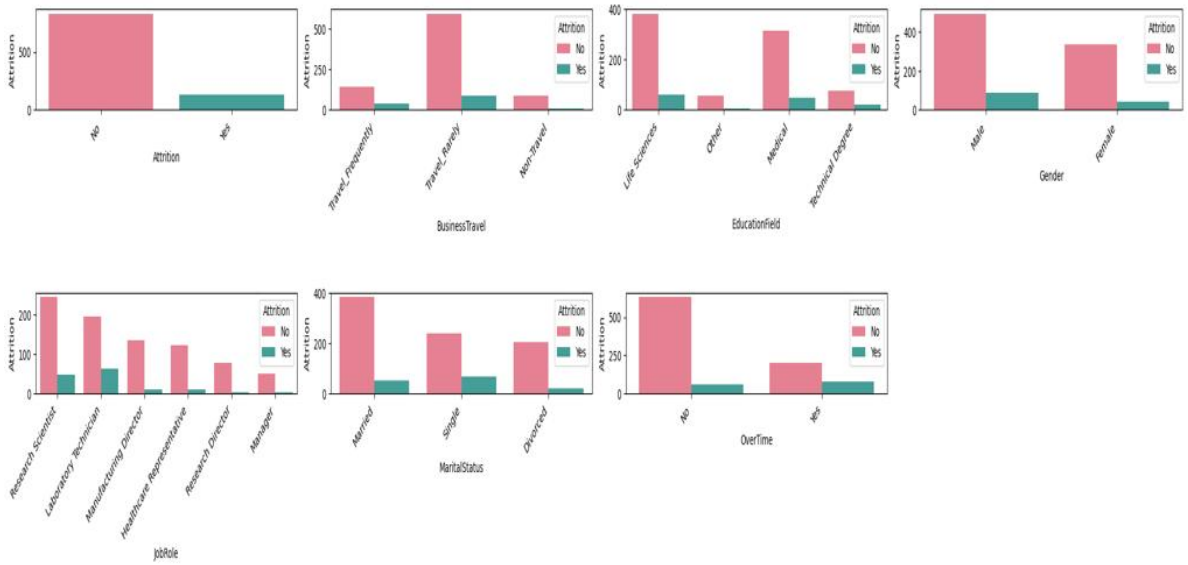
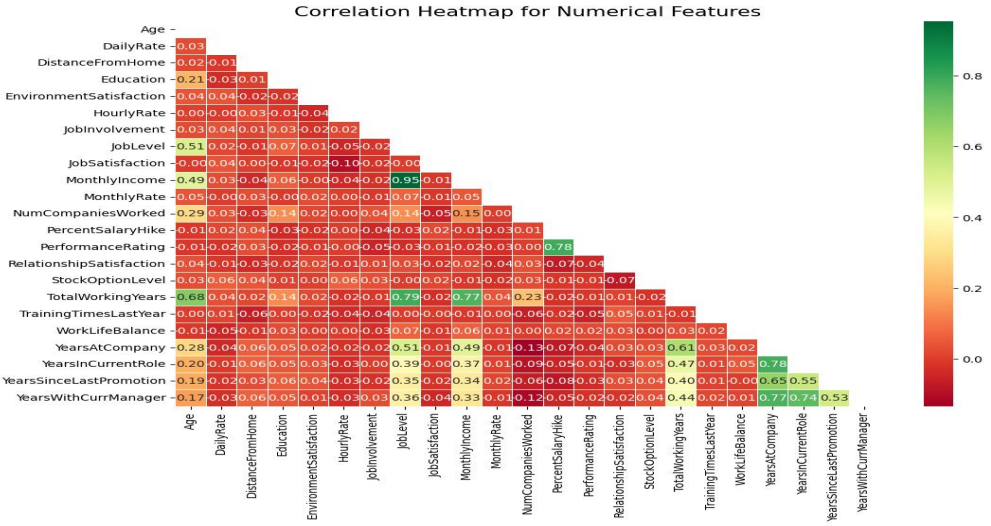


Figure 6:Bivariate Analysis for Categorical vs Categorical variables



The last bivariate analysis between numeric and numeric variables is shown in figure 7 called heatmap. That is used to find out the multicollinearity between two numeric independent variables. High correlation exist between, job level, monthly income, total working years, percent salary hike, years in current role and years at company. Removing the high correlated variables can increase the model accuracy.

Figure 7: Numeric vs Numeric variable



3.1.4 Data Encoding

Converting the categorical data into numerical form to run ML algorithms is called data encoding. Two encoding algorithms were run on the dataset. Label encoder was applied on the dependent variables that contains only two categories and one hot encoding (OHE) on the categorical independent variables.

3.1.5 Data Normalization

The standard scaler was utilized to convert numerical variables into a unit scale. Which is crucial in ML, especially for models sensitive to variations in variable values. Standardization guarantees that all variables are comparable, enables models to perform more efficiently, execute more rapidly, and provide more accurate forecasts. The equation 1 is the formula for standardizing:

$$z = \frac{x_i - \bar{x}_i}{s_i} \quad \ldots(1)$$

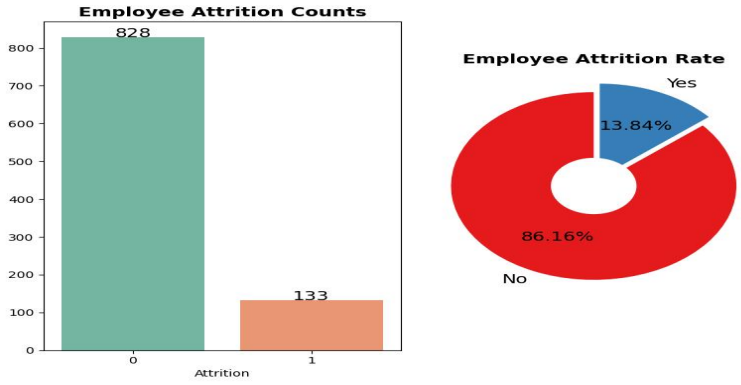
3.1.6 Handling Imbalanced Data

In employees' attrition detection, imbalanced classes in IBM data are a common problem that can affecting predictive model accuracy. This study tackles this issue using a dataset of 921 employee attrition. Figure 8 exhibited a considerable imbalance: 828 instances of typical non-attrition (label 0) and 133 instances of attrition (label 1). The disproportionate distribution of labels results in biased predictions that favor the dominant label, diminishes the model's capacity to generalize to unseen data, and requires intricate assessment criteria to measure and enhance model performance.

This research used the Adaptive Synthetic Sampling (ADASYN) technique to address the imbalanced dataset problem. This reliable oversampling technique produces synthetic samples for the minority label. ADASYN adaptability adjusts the weights of minority label based on their level of difficulty in learning (Balaram & Vasundra, 2022) . On minority labels, it enhances model performance without sacrificing

overall accuracy. The dataset had 828 for each label, i.e., non-attrition (label 0) and attrition (label 1) after the ADASYN algorithm was applied.

Figure 8: Imbalanced Dataset



3.1.9 Data Splitting

The data was partitioned into two subsets (training and testing datasets) in an 70:30 ratio, which is essential for developing ML models. Dividing data into training and testing sets enables the model to learn from the training set, assess its performance on the testing set, and ultimately evaluate its capacity to manage new data. This prevents the model from being overly specialized to the training data and aids in its performance across diverse data type.

3.2 Feature Selection

Feature selection method is vital to improve model performance, reduce overfitting, improve model interpretability, reduce training times and build more robust model. Three feature selection methods was applied like, correlation, recursive feature selection and mutual information. Correlation feature selection method performed well among the others. Monthly income, total working years, years in current role and years with current manager independent variables was identified highly correlated by using correlation feature selection method and removed from the dataset to build robust model.

3.3 Machine Learning Algorithms

In this section, the mechanism and concept of ML classification models are examined such as Decision Tree, K-Nearest Neighbor, Support Vector Machines and Random Forest. These four ML classification algorithm was applied to predict the R&D employees' attrition.

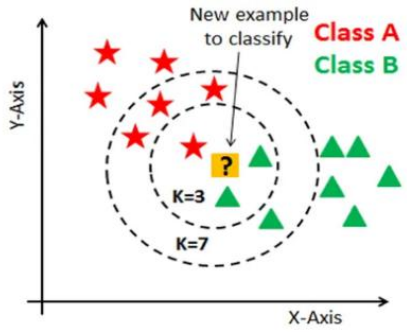
3.3.1 K-Nearest Neighbor (KNN):

K-nearest neighbor is also the poplar and simplest distance based ML model used for classification and regression problems. It designates a class label according to the predominant class among its nearest neighbors in the new dataset (Shokrzade et al., 2021) . It functions according to the principal of proximity and similarity within feature space as shown in figure 9.

3.3.2 Decision Tree (DT):

DT is also used for regression and classification problems and consider as a supervised ML algorithms. This model operates by recursively partitioning the whole dataset into subsets according to variable values, so forming a tree-like structure for prediction purposes (Klusowski & Tian, 2024). Easily interpretable, visually readable and non-parametric are the advantages of DT.

Figure 9: KNN Algorithm



3.3.3 Random Forest (RF):

Breiman (1996) proposed a distance based called RF model. This model calculate the output on the basis of DT by utilizing multiple random sample(Boateng et al., 2020). Random selection is used to choose the explanatory variables for every node in the decision tree fitting procedure (Hu & Szymczak, 2023). Randomness is enhanced by using this technique, and as randomness is enhanced, decision tree correlations reduce, resulting in minimizing prediction error. RF reduces overfitting and ensures greater predictive power and stability. Furthermore, it is particularly helpful when there is an unequal distribution of the data and more data from a certain class.

3.3.4 Support Vector Machine (SVM):

A non-stochastic binary linear model for ML classification is the SVM. The SVM model classifies data by identifying the border with the largest area to which the data belongs, and it generates a decision boundary based on the data that establishes the category to which the data belongs. Both non-linear and linear data can be classified by it (Roy & Chakraborty, 2023). When performing classification tasks, the SVM model aim to minimize the classification error while maximizing the margin classification that classifies both classes (Cervantes et al., 2020) . It may be used for prediction issues when the dependent variable is continuous, as well as classification analysis based on mathematical optimization approaches.

4. Results

This study used precision, recall, F1 score, accuracy, and AUC score metrics to evaluate the attrition of R&D employees. The following are the metrics specifics:

Accuracy: Accuracy provides the model's overall effectiveness. Calculated as the given formula in equation 2

$$Accuracy = \frac{TP + TN}{(TP + TN + FP + FN)} \quad \ldots(2)$$

Precision: Precision provides the model's proportion of true positives among predicted positives. Calculated as the given formula in equation 3

$$Precision = \frac{TP}{(TP + FP)} \quad \ldots(3)$$

Recall: Recall score provides proportion of true positives among actual positives of the model. Calculated as the given formula in equation 4

$$Recall = \frac{TP}{(TP + FN)} \quad \ldots(4)$$

F1 Score: F1 score is the Harmonic mean of recall and precision as mentioned equation 5.

$$F1 - Score = \frac{2(Precision * Recall)}{(Precision + Recall)} \quad \ldots(5)$$

AUC Score: The abbreviation of AUC is Area under the Precision-Recall curve. It the most strong evaluation metric when the dataset is imbalanced. Higher AUC scores (near to 1) indicates best model performance.

Table 2: Evaluation Metrics

Classification Metrics for Correlation-Based Feature Selection:						
	Model	Accuracy	Precision	Recall	F1 Score	AUC
2	Random Forest	0.935743	0.935876	0.935743	0.935749	0.975458
3	SVM	0.905622	0.905996	0.905622	0.905630	0.965971
4	K-Nearest Neighbors	0.879518	0.892571	0.879518	0.878251	0.959686
0	Logistic Regression	0.847390	0.847828	0.847390	0.847266	0.935330
1	Decision Tree	0.841365	0.841546	0.841365	0.841289	0.841019

Resulting from analyzing the performance of the classification model predicting R&D employees attrition, Random forest showed the leading performance in all evaluation metrics including accuracy (0.935743), precision (0.935876), recall (0.935743), F1 score (0.935749), and AUC score (0.975458).

Overall, the Random Forest ML model excelled compared to other single models like KNN, SVM, Logistic Regression and DT. This highlights that the single model performed well compared to hybrid ML models if the stratum is the main concern of research. It also indicates single model consume low computational resources, while hybrid model consume more computational resources. Single models provide a balance of interpretability, accuracy and computational efficiency.

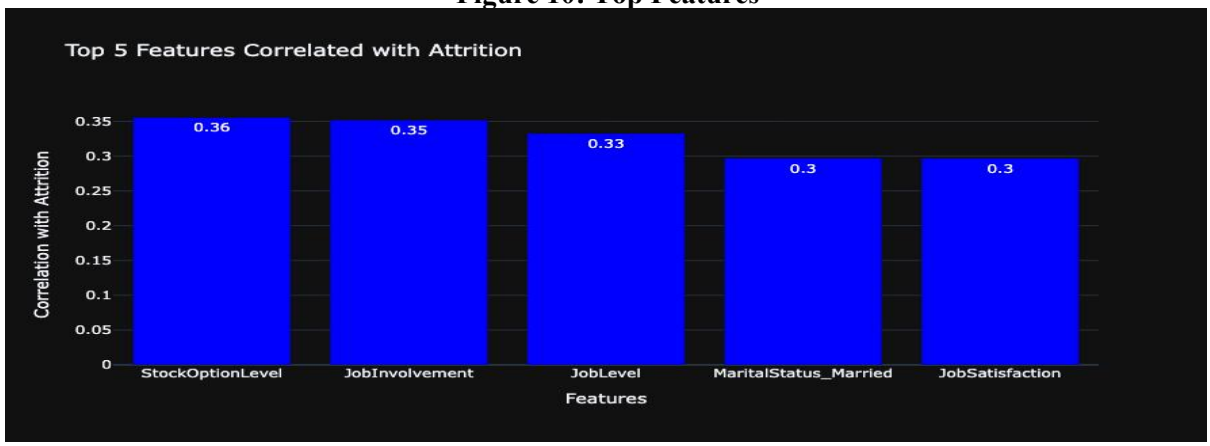
This study employs an RF model that generates an ideal algorithm using numerous decision trees, effectively analyzing intricate nonlinear interactions across several levels to appropriately assess the complicated interrelations of personnel- related variables. Table 2 presents the outcomes of computing the F1 score, accuracy and AUC in relation to the attrition prediction value and the employees prediction value using the test data.

The correlation feature selection method was used to measure the contribution of each variable in the RF prediction model. The following steps were used to figure out the contribution for a certain variable. First, to measure the relevance, the correlation was calculated between features and target variable. To measure redundancy, correlation was calculated between features. Second, A Merit score was calculated by using a heuristic that balances relevance and redundancy. Equation 6 is Merit score formula:

$$Merit\ score = \frac{k.\overline{r_{cf}}}{\sqrt{k + (k-1).\overline{r_{ff}}}} \quad \ldots(6)$$

Lastly, an empty set were created and iteratively features were added to the empty set based on the Merit score. The features contained in the empty set provides the best features by using the trade-off between relevance and redundancy. Figure 10 showed the top 5 features contributed to the prediction of R&D employees attrition. In particular, stock option level contributed the most to the prediction. Additionally, the two main features contributing attrition prediction were Job involvement and job level. Generally job related variables were the highest contributor to predicting the employees' attrition.

Figure 10: Top Features



5. Conclusion

The aim of this paper to find the main factors of R&D employees attrition and find out the machine learning model to predict the attrition of an employee. The evaluation table shows that Random Forest achieved the highest performance among Logistic regression, Support vector Machines, Decision Tree and K-nearest neighbor utilizing the IBM dataset and creating an R&D employees attrition prediction model. Several studies used the different single ML algorithms, feature selection and sampling methods to predict the employee attrition across various departments but didn't achieve high accuracy (Al-Darraji et al., 2021; El-Rayes et al., 2020; Pratt et al., 2021; Raza et al., 2022). This research study achieved high performance by using a stratum strategy, that is considered as highly effective for target analysis.

Furthermore, stock option level , job involvement, job level and job satisfaction are the top features contributed to predict the R&D employees attrition model by using the top feature algorithm. In deep analysis on the IBM data a high attrition rate was found in young male, high participation in work and low experienced R&D employees from the univariate and bivariate analysis. These employees environment satisfaction, performance rating, work balance life was high. Some more hidden factors like they are laboratory technicians or research scientists, shorter in tenure, lower monthly income (6000 to 1500), stagnant role, and office distance from home greater than 10 km.

From a pragmatic perspective, considering the attrition factors of R&D personnel, the machine learning model for predicting R&D employee attrition will enhance the efficiency and efficacy of the company's financial, temporal, and human resource management costs associated with attrition. predictive managers and high authorities utilizing this methodology will enhance the company's sustainable

management and productivity by proactively identifying employee attrition and fostering a positive organizational culture.

This research study has a few limitation regardless of its contributions. First, in this research study the IBM dataset of one company was used to find the attrition factors and machine learning model, which may leads to non-generalizability of the findings to organizational contexts and other cultural. Second, tree, distance and mathematical model was applied to predict the attrition. Future research could explore this framework in diverse industries and cultural settings to validate and extend the current findings.

This study presented a single model with high performance but further can be improved. Other than single model, hybrid model like, boosting, bagging, deep learning and ensemble techniques can be applied to improve further accuracy. Different ML algorithms can be performed to identify the hidden patterns like relationship between the variables and feature engineering. In future, the research will be conducted on the non-technical employees attrition and compare the attrition factors between technical and non-technical employees.

Acknowledgement:
The author acknowledges the useful comments from the Editor and anonymous reviewers. Moreover, all remaining errors are our own.

Data Availability Statement:
Data is self-collected from published (secondary sources) and self-regressed and will be provided on demand.

Funding, if any:
This research has received no external funding.

Conflict of Interest Disclosure Statement:
There is no conflict of interest among the authors of the study.

Ethical Approval:
Ethical approval has been obtained from the relevant forum if necessary.

References

Ahn, J., Hwang, J., Kim, D., Choi, H., & Kang, S. (2020). A Survey on Churn Analysis in Various Business Domains. *IEEE Access*, 220816–220839. https://doi.org/10.1109/ACCESS.2020.3042657

Al-Darraji, S., Honi, D. G., Fallucchi, F., Abdulsada, A. I., Giuliano, R., & Abdulmalik, H. A. (2021). Employee Attrition Prediction Using Deep Neural Networks. *Computers*, *10*(11), 141. https://doi.org/10.3390/computers10110141

Bai, L., Zhao, X., Kang, S., Ma, Y., & Zhang, B. (2023). Dynamic assessment of stakeholder conflict risk for R&D project portfolios. *Engineering, Construction and Architectural Management*. https://doi.org/10.1108/ECAM-03-2023-0301

Balaram, A., & Vasundra, S. (2022). Prediction of software fault-prone classes using ensemble random forest with adaptive synthetic sampling algorithm. *Automated Software Engineering*, *29*(1), 6. https://doi.org/10.1007/s10515-021-00311-z

Boateng, E. Y., Otoo, J., & Abaye, D. A. (2020). Basic Tenets of Classification Algorithms K-Nearest-Neighbor, Support Vector Machine, Random Forest and Neural Network: A Review. *Journal of*

Data Analysis and Information Processing, *08*(04), 341–357. https://doi.org/10.4236/jdaip.2020.84020

Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., & Lopez, A. (2020). A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*, *408*, 189–215. https://doi.org/10.1016/j.neucom.2019.10.118

Chandler, J. (2023). Participant Recruitment. In *The Cambridge Handbook of Research Methods and Statistics for the Social and Behavioral Sciences* (pp. 179–201). Cambridge University Press. https://doi.org/10.1017/9781009010054.010

Chang, J. Y., Choi, J. N., & Kim, M. U. (2008). Turnover of highly educated R&D professionals: The role of pre-entry cognitive style, work values and career orientation. *Journal of Occupational and Organizational Psychology*, *81*(2), 299–317. https://doi.org/10.1348/096317907X204453

Chou, H.-C., Lu, W.-M., Kweh, Q. L., & Tsai, C.-Y. (2023). Using hierarchical network data envelopment analysis to explore the performance of national research and development organizations. *Expert Systems with Applications*, *234*, 121109. https://doi.org/10.1016/j.eswa.2023.121109

Chung, D., Yun, J., Lee, J., & Jeon, Y. (2023). Predictive model of employee attrition based on stacking ensemble learning. *Expert Systems with Applications*, *215*, 119364. https://doi.org/10.1016/j.eswa.2022.119364

Cordero, R., DiTomaso, N., & Farris, G. F. (1994). Career development opportunities and likelihood of turnover among R&D professionals. *IEEE Transactions on Engineering Management*, 223–232. https://doi.org/10.1109/17.310137

El-Rayes, N., Fang, M., Smith, M., & Taylor, S. M. (2020). Predicting employee attrition using tree-based models. *International Journal of Organizational Analysis*, *28*(6), 1273–1291. https://doi.org/10.1108/IJOA-10-2019-1903

Gadekar, B., & Hiwarkar, T. (2023). A Critical Evaluation of Business Improvement through Machine Learning: Challenges, Opportunities, and Best Practices. *International Journal on Recent and Innovation Trends in Computing and Communication*, *11*(10s), 264–276. https://doi.org/10.17762/ijritcc.v11i10s.7627

Grabner, I., Seiter, M., Wabnegg, M., & Wirth, H. (2024). The role of target difficulty and career tournaments in retaining creative R&D employees. *Contemporary Accounting Research*, *41*(2), 1058–1088. https://doi.org/10.1111/1911-3846.12931

Gupta, A. K., & Wilemon, D. (1990). Improving R&D/Marketing relations: R&D's perspective. *R&D Management*, *20*(4), 277–290. https://doi.org/10.1111/j.1467-9310.1990.tb00718.x

Hu, J., & Szymczak, S. (2023). A review on longitudinal data analysis with random forest. *Briefings in Bioinformatics*, *24*(2). https://doi.org/10.1093/bib/bbad002

Kiazad, K., Restubog, S. L. D., Hom, P. W., Capezio, A., Holtom, B., & Lee, T. (2024). STEMming the tide: New perspectives on careers and turnover. *Journal of Organizational Behavior*, *45*(3), 335–343. https://doi.org/10.1002/job.2780

Klusowski, J. M., & Tian, P. M. (2024). Large Scale Prediction with Decision Trees. *Journal of the American Statistical Association*, *119*(545), 525–537. https://doi.org/10.1080/01621459.2022.2126782

Kumari, B., Sahney, S., & Madhukar, A. (2022). Aligning individual and organizational R&D goals for self-sustainability: investigating preferences of researchers in selected CSIR laboratories, India.

International Journal of Productivity and Performance Management, *71*(5), 1642–1673. https://doi.org/10.1108/IJPPM-12-2019-0556

Luo, T., & Zhang, Z. (2021). Multi-network embeddedness and innovation performance of R&D employees. *Scientometrics*, *126*(9), 8091–8107. https://doi.org/10.1007/s11192-021-04106-7

Pirrolas, O. A. C., & Correia, P. M. A. R. (2022). Literature Review on Human Resource Churning—Theoretical Framework, Costs and Proposed Solutions. *Social Sciences*, *11*(10), 489. https://doi.org/10.3390/socsci11100489

Potočnik, K., Anderson, N. R., Born, M., Kleinmann, M., & Nikolaou, I. (2021). Paving the way for research in recruitment and selection: recent developments, challenges and future opportunities. *European Journal of Work and Organizational Psychology*, *30*(2), 159–174. https://doi.org/10.1080/1359432X.2021.1904898

Pratt, M., Boudhane, M., & Cakula, S. (2021). Employee Attrition Estimation Using Random Forest Algorithm. *Baltic Journal of Modern Computing*, *9*(1). https://doi.org/10.22364/bjmc.2021.9.1.04

Qutub, A., Al-Mehmadi, A., Al-Hssan, M., Aljohani, R., & Alghamdi, H. S. (2021). Prediction of Employee Attrition Using Machine Learning and Ensemble Methods. *International Journal of Machine Learning and Computing*, *11*(2), 110–114. https://doi.org/10.18178/ijmlc.2021.11.2.1022

Raza, A., Munir, K., Almutairi, M., Younas, F., & Fareed, M. M. S. (2022). Predicting Employee Attrition Using Machine Learning Approaches. *Applied Sciences*, *12*(13), 6424. https://doi.org/10.3390/app12136424

Roy, A., & Chakraborty, S. (2023). Support vector machine in structural reliability analysis: A review. *Reliability Engineering & System Safety*, *233*, 109126. https://doi.org/10.1016/j.ress.2023.109126

Saini, A., & Saini, V. (2024). IMPACT OF AI ON BUSINESS MANAGEMENT. *International Journal of Engineering Science and Humanities*, *14*(Special Issue 1), 143–146. https://doi.org/10.62904/51j6ve95

Schmitz, P., Yockel, S., Mizumoto, C., Cheatham, T., & Brunson, D. (2021). Advancing the Workforce That Supports Computationally and Data Intensive Research. *Computing in Science & Engineering*, *23*(5), 19–27. https://doi.org/10.1109/MCSE.2021.3098421

Shokrzade, A., Ramezani, M., Akhlaghian Tab, F., & Abdulla Mohammad, M. (2021). A novel extreme learning machine based kNN classification method for dealing with big data. *Expert Systems with Applications*, *183*, 115293. https://doi.org/10.1016/j.eswa.2021.115293

Singh, R., Sharma, P., Foropon, C., & Belal, H. M. (2022). The role of big data and predictive analytics in the employee retention: a resource-based view. *International Journal of Manpower*, *43*(2), 411–447. https://doi.org/10.1108/IJM-03-2021-0197

Szopik-Depczyńska, K., Dembińska, I., Barczak, A., Szczepaniak, K., Secka, J., & Ioppolo, G. (2024). Does marketing orientation impact the innovation activity of R&D departments influenced by user-driven innovation? The correspondence analysis. *European Journal of Innovation Management*, *27*(5), 1793–1811. https://doi.org/10.1108/EJIM-03-2023-0196

Tikas, G. D. (2023). Measuring R&D team's focus towards innovation: Scale development and empirical validation. *Journal of Engineering and Technology Management*, *69*, 101766. https://doi.org/10.1016/j.jengtecman.2023.101766

Wu, Y., Tan, S., Luo, Q., & Yan, X. (2019). Research on Performance Evaluation Index System of R&D Personnel in Start-Up Software Enterprises. *Proceedings of the 5th Annual International Conference on Social Science and Contemporary Humanity Development (SSCHD 2019)*. https://doi.org/10.2991/sschd-19.2019.68