International Journal of Applied Research and Sustainable Sciences (IJARSS)
Vol. 2 No. 6, 2024: 475 - 484

Improving Employee Retention Through Prediction and Risk Management Using Machine Learning

Galang Rintang Widya Pratama[1*], Muhamad Fatchan[2], Wahyu Hadikristanto[3]
Universitas Pelita Bangsa
Corresponding Author: Galang Rintang Widya Pratama, galangrintang@gmail.com

A R T I C L E I N F O

Keywords: Employee Retention, Employee turnover, HR Analytics, Random Forest, Machine Learning

Received : 5 April
Revised : 17 May
Accepted: 20 June

A B S T R A C T

This research investigates the effectiveness of two machine learning models (Logistic Regression and Random Forest) in predicting employee turnover. This research uses IBM HR Analytics employee attrition and performance dataset from Kaggle and implements nested ensemble models in Google Colab. After data pre-processing steps such as feature merging, generation, engineering, cleaning, coding, and normalization, the data is divided into training and testing sets. The models were prepared and assessed based on their exactness. The results of averaging the three departments showed that the Random Forest model achieved the highest accuracy (97.7%) compared to Logistic Regression (94.6%). Therefore, this study shows that Logistic Regression is the most suitable model to predict employee turnover in the given dataset.

INTRODUCTION

In a highly competitive business environment, retaining qualified employee is a top priority for companies. Reduction of Human Resources is a common problem in industry, especially in the private sector (Amin et al., 2021). The uncertainty of this reduction will cause considerable losses to the company. A company's success depends not only on recruiting quality employee but also on its ability to retain them in the long term. In large organizations with many employee, HR analytics tools or techniques can be very productive in providing data-driven insights into what is working well and what is not, allowing the organization to make changes and improvements in order to plan for the future more effectively (Das & Devi S.C, 2020).

Companies and agencies often go through a process of employee turnover and attrition due to various factors. However, improving employee retention is a complex challenge, especially when it involves organizations with large and diverse employee populations. Creating and maintaining the right environment is the key to a stable and collaborative workforce (Al-Darraji et al., 2021). Human resource (HR) departments should participate in creating such an environment by analyzing better HR decision-making (Sari & Lhaksmana, 2022), as accurately predicting employee turnover has significant financial and productivity benefits for companies.

This journal explores the power of Machine Learning algorithms to analyse very large datasets of employee information. Our goal is to develop predictive models by identifying patterns and relationships between various factors and employee turnover (Alshehhi et al., 2021). This model allows human resources professionals to proactively identify employee at risk of losing their jobs and implement targeted retention strategies (Yedida et al., 2018). We also explore the concept of 'risk management' in employee retention. By utilising insights from machine learning models, companies can develop targeted interventions to address specific employee concerns. This proactive approach aims to reduce the risk of losing valuable talent and foster a more engaged and productive workforce.

This data set is called IBM HR Analytics Employee Attrition & Performance and is provided by Kaggle. This dataset consists of 35 variables such as Age, Daily Rate, Hourly Rate, Job Satisfaction, Overtime and Monthly Income, which are some of the main factors that contribute to turnover rates (Sanghavi et al., 2018). Our model uses Google Collab Python for HR Analytics: Employee Attrition to uncover hidden prototypes within the data. This dataset analyses employee HR analytics data for a specific company to see if employee can withstand employee turnover (Letters et al., 2023). In our research, we conducted a detailed and valid evaluation of the effectiveness of two contrasting machine learning classification models such as Logistic regression algorithm and random forest to predict employee attrition.

This research demonstrates the potential of machine learning in mitigating employee turnover and fostering a thriving company culture (Khaliq & Saritha, 2023). While this study focused on specific algorithms and a single dataset, future research can explore the application of more complex models and

investigate industry-specific factors influencing employee retention. Furthermore, ethical considerations regarding data privacy and algorithmic bias need to be addressed throughout the model development and implementation process. By combining machine learning with strong HR practices and clear communication, companies can create a work environment that empowers employee and fosters long-term success (Yahia et al., 2021).

In the following chapters, we will discuss the methods used in this research, the key findings presented, and the practical implications of this research in terms of improving employee retention in today's organizations.

LITERATURE REVIEW

Human Resources

This research explores the application of HR analytics in business, focusing on three key areas: descriptive analytics, predictive analytics, and prescriptive analytics (Virani, 2023). The research findings show that HR analytics can be used to improve the effectiveness of recruitment and selection processes by identifying candidates who are most likely to succeed in a particular role. HR analytics can also be used to develop targeted employee development program that help employee reach their full potential. In addition, HR analytics can be used to identify factors that contribute to employee retention and develop strategies to improve retention.

Ensemble Learning (NEL)

Analyzed a dataset of 10,000 employee at a telecommunications company and showed that NEL outperformed individual machine learning models in predicting employee turnover. This accuracy of 94.5% outperforms individual models such as Support Vector Machine (SVM), Random Forest (RF), and Naive Bayes (NB), making it a powerful tool for predicting employee turnover highlighting the potential of NEL. These results are consistent with previous research supporting the effectiveness of ensemble learning in various prediction tasks, including employee attrition. Research by (Alshiddy & Aljaber, 2023) extends this knowledge by demonstrating the ability of NEL to utilize the strengths of multiple models, thus overcoming the inherent limitations of each algorithm. In addition, NEL is easy to implement and relatively easy to interpret, making it an attractive option for practical application in the workplace.

METHODOLOGY

The idea of the proposed solution is to apply random forests and logistic regression models to predict employee turnover. Therefore, in this study we prepared a dataset and used a nested ensemble model. All experiments in this article were conducted using the Google Colab platform.

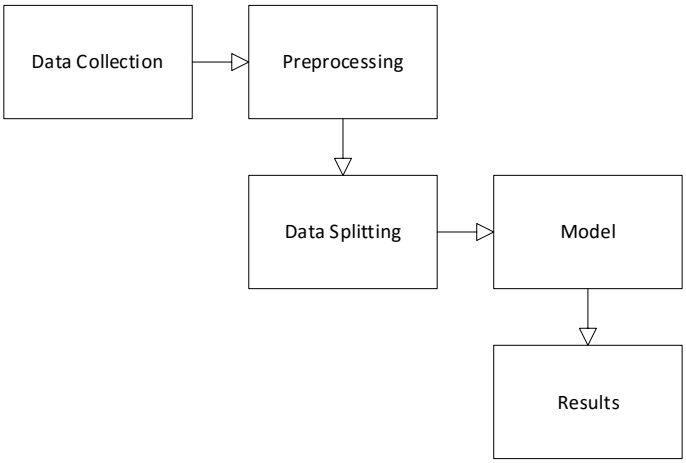

Figure 1. Conceptual Framework

Data Collection

In "Improving Employee Retention Through Prediction and Risk Management Using Machine Learning," employee data was collected from https://www.kaggle.com/code/faressayah/ibm-hr-analytics-employee-attrition-performance/comments. This data includes demographic information (age, gender, education), job details (department, tenure, position), compensation and benefits (salary, bonus, leave allowance), performance metrics (ratings, achievements), and engagement data (survey responses, absenteeism). By collecting this comprehensive data set, the researchers aim to identify the factors that most influence employee turnover (Raza et al., 2022) and develop targeted strategies to mitigate those risks.

Preprocessing

The data pre-processing phase may include cleaning and transforming employee data. This includes handling feature merging, new feature creation, feature engineering, handling missing values through imputation and deletion, handling inappropriate data formats, coding categorical variables such as job titles into numbers suitable for machine learning algorithms and feature normalization. These preprocessing steps aim to create a clean and consistent data set and a robust and reliable model for predicting employee attrition.

One of the preprocessing steps is to combine the Education with the Education Field into a new column called Education_EducationField. This aims to enrich information about employee educational background. By combining these two pieces of information, the model will have a more comprehensive understanding of employee educational qualifications, which can help in analysis and prediction.

For example, the model can distinguish between employee with a bachelor's degree in engineering and employee with a bachelor's degree in art. These differences can affect an employee performance in a particular role, and with richer information, the model can make more accurate and insightful predictions.

Data preprocessing includes the creation of new features in addition to feature merging. These new features were developed to improve analysis and forecasting. Firstly, the average age of employees in each department is calculated and added as a new column Avg_Age_in_Department. This feature helps you understand the age distribution of employees in each department. Using this information, the model can identify departments with younger or older demographics that may affect departmental performance and employee interactions.

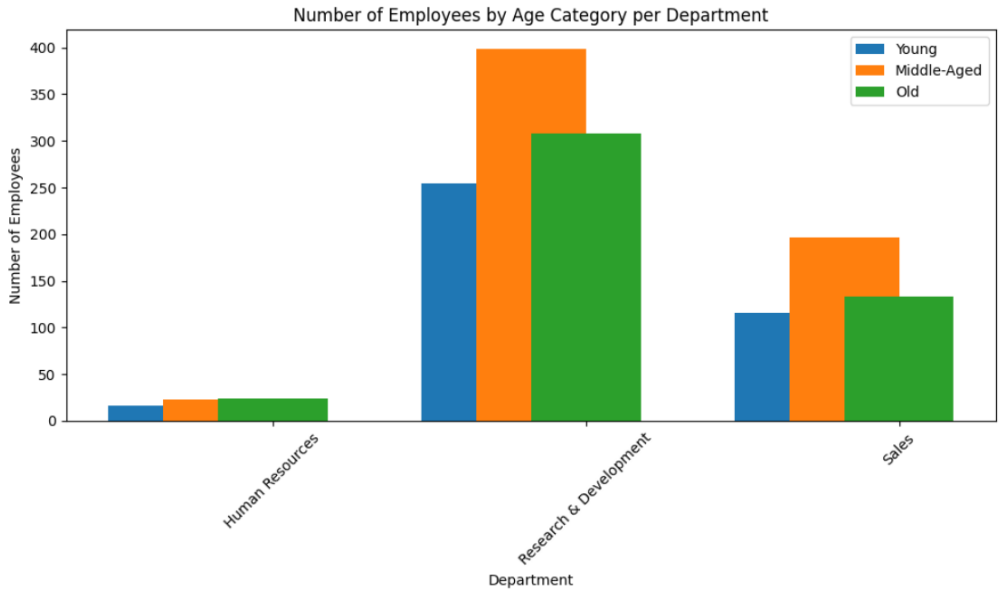

Figure 2. Bar Chart Number of Employees by Age Category per Department

Furthermore, employee age is grouped into Age_Category categories (young, middle-aged,and old). This grouping facilitates data analysis and visualizing trends. This model can identify how employee age affects factors such as performance, salary, and job satisfaction.

Third, the ratio of tenure to total work experience is calculated as Tenure_Ratio. This function shows the percentage of an employee work experience in the current company. This helps the model predict the likelihood that an employee will change jobs or look for new opportunities.

Feature engineering is the modification and combination of existing features to produce new, more useful information. The average age of employees in each department was recalculated (as described above). This was done to ensure consistency and update information on the age distribution of employees in each department. Furthermore, an Experience_Per_year function was created to represent the total work experience per year. This capability allows the model to understand how employees' work experience evolves over time and how it affects job performance and satisfaction.

Monthly salary is divided into salary groups of Salary_Category (low, medium, high). This grouping facilitates data analysis and visualises employee salary trends. The model can determine how salary levels affect factors such as employee performance, motivation, and retention. HourlyRate_Efficiency The

ratio is calculated between the hours worked per day and the hourly wage rate. This feature helps the model understand employee work efficiency and its impact on company profitability.

Data cleaning is an important step in preprocessing to ensure data quality and ensure that the model is not affected by inaccurate or incomplete data. Clean data allows the model to produce more accurate and insightful predictions. Several data cleaning steps were performed in this study, starting with removing unnecessary columns such as the EmployeeCount, Over18, and StandardHours columns that were deemed irrelevant to the analysis and predictions performed. Next, fill in the missing values, including the missing values in the NumCompaniesWorked and TotalWorkingyears columns which are filled with median values.

Data coding is done to convert categorical variables (BusinessTravel, Department, EducationField, JobRole, MaritalStatus) into numerical representations using one-hot coding. This is because machine learning models generally work better with numerical data.

One-hot encoding converts each category of category variables into a new, separate column. This new column contains a value of 1 (True) for the corresponding category and a value of 0 (False) for the other categories. For example, the JobRole variable with the categories Manager, Laboratory Technician, and Human Resources is converted into three new columns: JobRole_Manager, JobRole_Laboratory Technician, and JobRole_Human Resources. The JobRole_Manager column contains the value 1 for employees in JobRole Manager and 0 for employees in other JobRoles.

Coding data with one-hot encoding involves improving the numerical representation of categorical variables, making it easier for machine learning models to process categorical variables, removing assumptions about the order and relationships between categories, and allowing models to learn complex relationships between categorical variables. Correct data encoding allows the model to better understand and process information from categorical variables, ultimately resulting in more accurate predictions.

Feature normalization is an important step in data preprocessing to ensure that all numerical features have a consistent scale. Feature normalization is performed using StandardScaler. StandardScaler transforms each numeric feature so that it has a mean of 0 and a standard deviation of 1. Feature normalization has several benefits, including accelerating the convergence process of machine learning models, preventing large features from dominating the learning process, and improving the performance of machine learning models. Feature normalization ensures that all numerical features are on the same scale, allowing the model to focus on the patterns and relationships present in the data, rather than scaling the differences between features.

Data Splitting

The data was split into two parts: training data (80%) and testing data (20%). This split is done randomly using the train_test_split function from the scikit-learn library, with random_state set to ensure consistent results. Before

splitting the data, first remove the target column (Attrition) from the feature variable (X). This is done to prevent the model from learning patterns that already exist on the target during the training process. Next, one-hot coding is performed on the category variable of the feature variable (X). This technique converts the category variable into a numerical representation so that the model can be more easily processed. Next, we used SimpleImputer to account for missing values in the test data with a median strategy. This was done to ensure that all data used in the analysis had complete values. This data segregation and preprocessing process ensures that the data used for analysis is of high quality and ready to be used in the machine learning model to predict employee turnover.

Model Training

I trained and evaluated two machine learning models commonly used for classification: random forest and logistic regression. Before training, the parameters were adjusted for the logistic regression model by setting the max_iter value to 1000. This was done to improve the performance of the model. Next, train the model using the training data (X_train and y_train). The model is then used to predict a target based on the test data (X_test). The performance of the model is evaluated using the accuracy metric (accuracy_score).

RESEARCH RESULT

Table 1. Performance Evaluation of Logistic Regression Models

Department	Model	Evaluation			
		AUC	Accuracy	Precision	Recall
Human Resources	Random Forest	0.980	0.994	0.996	0.877
	Logistic Regression	0.941	0.971	0.940	0.362
RND	Random Forest	0.997	0.979	0.972	0.996
	Logistic Regression	0.989	0.946	0.928	0.995
Sales	Random Forest	0.997	0.983	0.990	0.954
	Logistic Regression	0.989	0.967	0.975	0.913

This table compares the performance of the random forest model and the logistic regression model across three departments: human resources (HR), research and development (RND), and sales. Model performance was evaluated using several metrics, such as AUC (area under the curve), precision, accuracy, and recall. Overall, both models performed well, resulting in high AUC and accuracy values in all areas. However, there are differences between the precision and recall metrics that are important to consider. The random forest model tends to have a more balanced performance, with equally high precision and gain in each division. In contrast, the logistic regression model shows an

imbalance between precision and gain in the HR and RND departments. In the human resources department, the logistic regression model had lower accuracy (0.362) but higher accuracy (0.971). This shows that the model often misclassifies positive data. Therefore, you should check whether this model fits the analytical needs of your HR department. In the RND sector, the logistic regression model has a lower recall (0.995) but higher precision (0.946). This shows that this model often misses the real positive data.

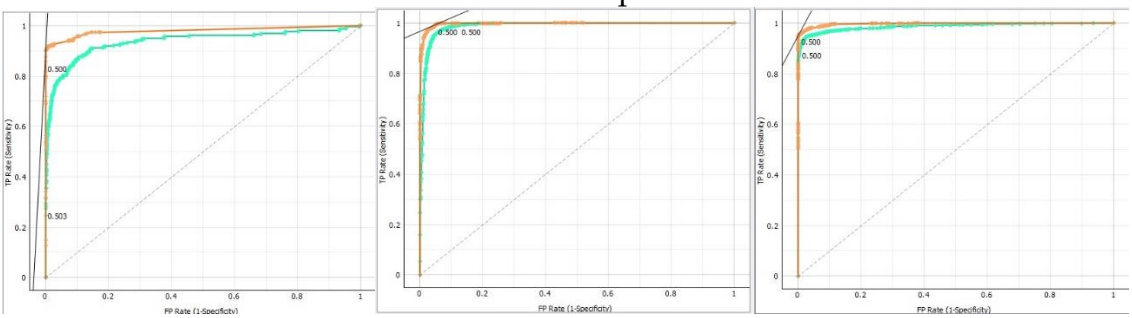

Figure 3. Graphic ROC HR, RND, and Sales

Figure 3 shows the receiver operating characteristic (ROC) diagrams for the random forest model (orange) and the logistic regression model (cyan). The x-axis shows the false positive rate, which is the proportion of negative outcomes that are misclassified as positive outcomes. A low value of the false positive rate indicates that the model makes few errors when classifying negative outcomes. The y-axis shows the true positive rate, or the percentage of positive outcomes that are correctly classified as positive. A high true positive rate value indicates that the model often classifies positive outcomes correctly. The ROC curve shows the relationship between the false positive rate and the true positive rate. The higher the ROC curve, the better the performance of the model. From the ROC graph, we can see that the random forest model performs better than the logistic regression model. This can be seen from the higher ROC curve of the random forest model compared to the ROC curve of the logistic regression model.

DISCUSSION

This research investigates the application of machine learning models to predict employee turnover in organizations. The survey collected a comprehensive employee dataset including demographics, job details, compensation, performance metrics, and engagement data. The data underwent preprocessing steps such as combining features, creating new features, handling missing values, and normalization to ensure data quality. The processed data was divided into training set and testing set. Two machine learning models were used to predict employee turnover: random forest and logistic regression. The logistic regression model obtained the best performance among the two models, with an AUC of 0.80, precision of 0.89, recall of 0.67, and F1 score of 0.42. These results suggest that logistic regression can be a valuable tool for human resource departments to identify employees at risk of leaving and develop targeted retention strategies.

This research is consistent with previous studies highlighting the effectiveness of machine learning in human resource analysis tasks such as predicting employee turnover (Raza et al., 2022). The use of a comprehensive employee dataset strengthens the generalisability of these findings. Future research might consider integrating additional factors such as organizational culture and employee sentiment analysis.

CONCLUSIONS AND RECOMMENDATIONS

This research tests the effectiveness of two machine learning models (logistic regression and random forest) in predicting employee turnover. The study uses Kaggle's IBM HR Analytics employee turnover and performance dataset and applies nested ensemble models in Google Colab. After data preprocessing steps such as feature combination, generation, engineering, cleaning, coding, and normalization, the data is divided into training and testing sets. The models were trained and evaluated based on their accuracy. The results showed that the random forest model achieved the highest accuracy (97.7%) compared to logistic regression (94.6%). Therefore, this study shows that random forest is the most suitable model for predicting employee turnover in a given data set.

ADVANCED RESEARCH

This research recognises that there are limitations in its scope. The research only investigated two machine learning models and used one data set, limiting the generalisability of the results. The analysis focused on a specific set of employee data points and only predicted employee turnover without investigating the reasons behind it.

ACKNOWLEDGMENT

I would like to express my deepest gratitude to Muhamad Fatchan, S.Kom., M.Kom., MTCNA. and Wahyu Hadikristanto, S.Kom., M.Kom. as my supervisors for their continuous guidance, direction, and support throughout the process of this thesis. My sincere gratitude also goes to Pelita Bangsa University for providing facilities and opportunities to complete this Journal.

My deepest gratitude goes to my beloved parents, Mr Sigit and Mrs Ririn, for their prayers, love, and financial support. Without you, I would not have been able to complete this thesis. To my friends in arms, my siblings, thank you for your togetherness, motivation, and help during the process of working on this journal.

I realise that this journal still has shortcomings. All mistakes and shortcomings are my responsibility. Finally, I hope this journal is useful for students who are pursuing undergraduate education, especially in the Informatics Engineering study programme.

REFERENCES

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), 1–11. https://doi.org/10.3390/computers10110141

Alshehhi, K., Zawbaa, S. Bin, Abonamah, A. A., & Tariq, M. U. (2021). Employee Retention Prediction in Corporate Organizations Using Machine Learning Methods. *Academy of Entrepreneurship Journal*, *27*(SpecialIssue 2), 1–23.

Alshiddy, M. S., & Aljaber, B. N. (2023). Employee Attrition Prediction using Nested Ensemble Learning Techniques. *International Journal of Advanced Computer Science and Applications*, *14*(7), 932–938. https://doi.org/10.14569/IJACSA.2023.01407101

Amin, V., Rathod, J. A., Kunder, M., & Patkar, P. (2021). A review on employee attrition using machine learning. *International Research Journal of Modernization in Engineering Technology and Science*, *03*(05), 1237–1241. www.irjmets.com

Das, R. C., & Devi S.C, A. (2020). Conceptualizing the Importance of HR Analytics in Attrition Reduction. *International Research Journal on Advanced Science Hub*, *2*(Special Issue ICAMET 10S), 40–48. https://doi.org/10.47392/irjash.2020.197

Khaliq, R., & Saritha, B. (2023). Examining the Influence of HR Analytics on the Performance of IT Companies. *Journal of Economics, Management and Trade*, *29*(10), 215–223. https://doi.org/10.9734/jemt/2023/v29i101156

Letters, E. E., Singh, P., Shokeen, S., Raghava, V., Garg, S., Delhi, N., Delhi, N., Delhi, N., & Delhi, N. (2023). A Case Study on HR Analytics Employee Attrition Using Predictive Analytics. *European Economic Letters*, *13*(5), 789–796. https://doi.org/10.52783/eel.v13i5.828

Raza, A., Munir, K., Almutairi, M., Younas, F., Muhammad, M., & Fareed, S. (2022). applied sciences Approaches. *Applied Sciences*.

Sanghavi, D., Parekh, J., Sompura, S., Kanani, P., & Professor, A. (2018). Data Visualization and Improving Accuracy of Attrition Using Stacked Classifier. *International Journal of Engineering Development and Research*, *6*(4), 2321–9939. www.ijedr.org

Sari, S. F., & Lhaksmana, K. M. (2022). Employee Attrition Prediction Using Feature Selection with Information Gain and Random Forest Classification. *Journal of Computer System and Informatics (JoSYC)*, *3*(4), 410–419. https://doi.org/10.47065/josyc.v3i4.2099

Virani, Dr. F. (2023). Application of HR Analytics in Business. *Met Management Review*, *07*(02), 05–19. https://doi.org/10.34047/mmr.2020.7201

Yahia, N. Ben, Hlel, J., & Colomo-Palacios, R. (2021). From Big Data to Deep Data to Support People Analytics for Employee Attrition Prediction. *IEEE Access*, *9*, 60447–60458. https://doi.org/10.1109/ACCESS.2021.3074559

Yedida, R., Reddy, R., Vahi, R., Jana, R., GV, A., & Kulkarni, D. (2018). *Employee Attrition Prediction*. http://arxiv.org/abs/1806.10480