Interdisciplinary School of Management and Technology
Institut Teknologi Sepuluh Nopember

JURNAL TEKNOBISNIS
2022 Vol. 8(1), 74-85

Received 24 Mar, 2022; Revised 5 May, 2022; Accepted 26 May, 2022 | DOI: 10.12962/j24609463.v8i1.941

Employee Attrition Prediction using Machine Learning in Rolling Stock Manufacturing Company

Mu'ammar Itqon[1*], Jerry Dwi Trijoyo Purnomo[2]

ABSTRACT

Employee retention is crucial in human resource management, particularly in the rolling stock industry, characterized by a dynamic working environment and fierce competition. Elevated attrition rates can incur heightened recruitment expenses, productivity decline, and institutional knowledge loss. Addressing these concerns, this research endeavors to construct precise predictive models pinpointing employees at a heightened attrition risk. Employing the Team Data Science Process (TDSP) framework, three distinct machine learning algorithms were leveraged: Logistic Regression (LR), Naive Bayes (NB), and Random Forest (RF) to structure the employee dataset. TDSP procedure encompasses stages from data acquisition cleansing to descriptive analysis, dataset partitioning, algorithm deployment, and model appraisal. The evaluated variables include job designation, employment type, marital and educational status, gender, tenure, commute distance, and age. Model effectiveness was gauged via precision, recall, F1-score, and overall accuracy. The Random Forest algorithm surpassed its counterparts, boasting a remarkable accuracy of 93.1%. SHAP (SHapley Additive exPlanations) was incorporated for profound comprehension and model transparency. This analysis accentuated job role and employment type as pivotal in attrition forecasting. Such insights are instrumental for rolling stock firms to discern core determinants and craft potent, data-centric retention approaches.

KEYWORDS: Logistic Regression, Naive Bayes, Random Forest, Tdsp Framework, Shap Analysis.

[1]Information Technology Specialist, PT Industri Kereta Api (persero), Madiun, Indonesia
[2]Business Analytics, Interdisciplinary School of Management and Technology, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
[3]Department of Statistics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
*Corresponding author: itqon01@gmail.com

1. INTRODUCTION

Employees are invaluable assets to organizations, and their attrition can pose significant challenges, as highlighted by (Bhartiya et al., 2019). Reasons for attrition may vary, including low job satisfaction, inadequate wages, or an unsupportive work environment (Alduayj & Rajpoot, 2018). It is essential to grasp the myriad factors prompting employees to leave (Pratt et al., 2021). Addressing high-performing employee attrition requires understanding the departure's underlying contributors (Immaneni & Sailaja, 2019; Pratt et al., 2021). One sector confronting this attrition issue is the railway manufacturing industry in East Java. This company aims to be an all-encompassing solution provider for sustainable land transportation systems in Asia. The unpredictable high attrition could hamper this vision, thus highlighting the company's inability to fathom the factors inducing employee attrition and devising effective prediction strategies.

In this research, Random Forest (RF), Logistic Regression (LR), and Naive Bayes (NB) have been chosen as the machine learning algorithms. RF is favored due to its adeptness at handling diverse data types, both categorical and numerical, and its prowess in preventing overfitting, making it effective for employee attrition prediction (Fernández-Delgado et al., 2014; Probst et al., 2019). Despite its simplicity, LR effectively models target variable probabilities pertinent in the context of attrition predictions (Cramer, 2005). Meanwhile, NB's efficiency in handling extensive datasets makes it opt for HR datasets, generally characterized by a multitude of features(Kohavi & Edu, 1993; Zhao & Wang, 2018). With the inclusion of Explainable Machine Learning techniques (Ribeiro et al., 2016), this research not only aspires to predict potential employee departures but also to discern the significant contributing factors to such predictions. The insights gleaned are anticipated to guide the company's management in framing more effectual employee retention policies.

Based on previous research, this research will analyze input data encompassing 2,546 employee records from 2017 to 2022. It will contain eight features about the determining factors for employee attrition in an Indonesian manufacturing company and the suitable machine learning method to forecast it. The study employs an experimental approach that begins with the collation of employee data, followed by data cleansing methods, descriptive data analysis, data elaboration for training and testing phases, experimenting with multiple classification algorithms, and eventually determining the most apt method for predicting employee attrition.

2. LITERATURE REVIEW

According to a study by (Mishra et al., 2016, industries or companies cannot sustain themselves in the long run without the capability for predictive analysis related to human resource management. Analyzing employees can aid organizations in reducing attrition issues by transforming how employees are treated and retained (Fallucchi et al., 2020; Yahia et al., 2021). To address attrition challenges, organizations ensure the right individuals are in the right place at the appropriate time through meticulous analysis

(Kakad et al., 2020; Srivastava & Eachempati, 2021). Several studies have been conducted to mitigate attrition using fictional or real data (Qutub et al., 2021).

The IBM dataset, comprising 1,470 samples with 34 features, is widely used for employee attrition prediction. Research using this IBM dataset (Pratt et al., 2021; Bhartiya et al., 2019; Srivastava & Eachempati, 2021) aims to provide an accurate model for predicting employee attrition by studying various models like Support Vector Machine (SVM), Decision Tree, Logistic Regression (Ponnuru, 2020), Random Forest (Gao et al., 2019; Pratt et al., 2021), XGBoost, and K Nearest Neighbors (KNN). Based on these studies, age and monthly income emerge as the predominant factors influencing attrition. Conversely, in research focused on the hospitality industry, the primary determinant for attrition was found to be the length of service (Immaneni & Sailaja, 2019)

Khera's research (Khera & Divya, 2019) in the IT sector reveals that gender and job experience do not significantly affect employee attrition. In India's manufacturing industry, employee turnover is primarily caused by low pay, limited career advancement opportunities, insufficient employee benefits, lack of recognition for work, strained employee relations, organizational culture, company policies, and unfair handling of complaints (Kaur & Padmanabhan, 2019). Meanwhile, in a study by Jennifer conducted on Business Process Outsourcing (BPO) firms or third-party employment agencies in the Philippines, based on data mining techniques, age was identified as the most significant factor affecting attrition. Specifically, the age group of 18-21 was found to be the most prone to attrition. In marital status, single individuals possess the highest likelihood of attrition following age. Moreover, the study indicated that employees with 0-2 years of tenure tend to resign more frequently.

3. METHODS

This research used several methodologies to analyze the dataset of rolling stock manufacturing employees from 2017 to 2022 in East Java. The chosen methods were based on the type of data available and the complexity of the relationships in the data.

Research Methods

The research procedure follows the TDSP methodology (see Figure 1), an agile and repetitive data science approach that efficiently produces predictive analytics solutions and smart applications.

- Data Acquisition and Understanding

  This stage involves acquiring the data necessary for the project and understanding its characteristics. It includes checking for missing values or outliers, which can affect the model's performance. Imputation techniques will be employed for missing data, while outliers will be further analyzed or removed depending on the context.

- Modelling

  At this point, the data is divided into a training set and a test set. The model is built using the training set, while its performance is evaluated using the test set. Three algorithms, namely Logistic Regression, Naïve Bayes, and Random Forest will be used on the training data to create predictive models.

- Model Evaluation

  Upon building the models, their performance will be tested on the test data using the evaluation metrics. This stage allows for assessing the models' performance and whether any adjustments need to be made.

- Deployment

  In this stage, the best-performing model is deployed for use. Leveraging the principles of Explainable Machine Learning, interpretations and explanations will be provided for each prediction made by the best-performing model. It allows for greater transparency and understanding of how the model is making its predictions.

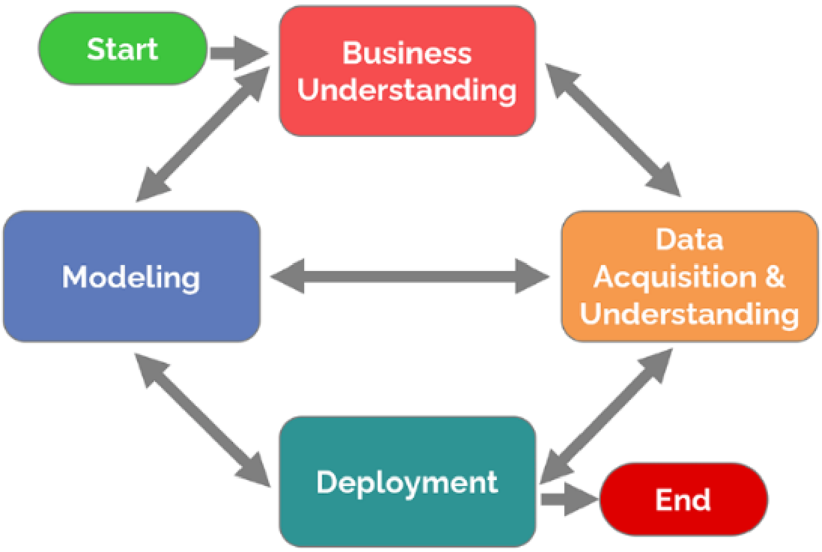

FIGURE 1. TDSP Framework

Data Collection & Research Variable

The data was extracted from the administrative records of the railway manufacturing company. The dataset encompasses information from 2017 to 2022 about 2,546 employees residing in East Java. The employee data attributes used to determine the employee attrition prediction model are shown in Table 1.

TABLE 1. Research Variable

Variable	Category	Scale
Attrition (Y)	0 = Employee still works at INKA	Nominal
	1 = Employee no longer works at INKA	
Permanent (X1)	0 = Contract Employee	Nominal
	1 = Permanent Employee	
Distance (X2)	-	Ratio

Variable	Category	Scale
Duration of Work (X3)	-	Ratio
Age (X4)	-	Ratio
Marital Status (X5)	1 = Unmarried Employee	Ordinal
	2 = Married Employee without children	
	3 = Married Employee with children	
Education (X6)	1 for High School Degree	Ordinal
	2 for Diploma Degree	
	3 for Bachelor's Degree	
	4 for Master's Degree	
	5 for Doctoral Degree	
Position (X7)	0 for Contract Employee Staff	Ordinal
	1 for Staff position grade 1	
	2 for Junior Specialist/Supervisor/Staff grade 2	
	3 for Manager/Junior Specialist/Staff grade 3	
	4 for Senior Manager/Middle Specialist/Staff grade 4	
	5 for General Manager/Senior Specialist/Staff grade 5	
Gender (X8)	0 = Female Employee	Nominal
	1 = Male Employee	

Analysis Methods

In the quest to predict employee attrition using machine learning, it is pivotal to select analytical methods that cater to the data's nature and distribution and uphold the prediction process's rigor and integrity. A range of algorithms were employed to obtain a holistic, yet precise, understanding of the determinants of attrition, each bringing forth distinct advantages in capturing the underlying patterns in the dataset.

- Logistic Regression
  Logistic regression, in essence, is a statistical technique that models and analyzes several predictor variables to predict a definite outcome. Due to its ability to transform data, Logistic Regression has the distinct advantage of predicting the likelihood of an event by fitting the data to a logistic curve. This is particularly useful when working with binary response variables, where the outcome is limited to two categories. The technique is valuable because the method does not require a linear relationship between the predictor and response variables, allowing for a non-linear boundary.

- Naïve Bayes
  The Naïve Bayes algorithm, grounded in Bayes' theorem, computes conditional probabilities in a simplified, 'naïve' assumption of independence between predictor variables. This classifier's merit lies in its ability to rapidly predict classes for new data instances without demanding extensive computational resources. Moreover, its probabilistic foundation makes it exceptionally adept at handling real-time

predictions, making it particularly relevant for a dataset that contains both categorical and numerical variables.

- Random Forest
  The Random Forest algorithm uses an ensemble learning approach, which involves training multiple decision trees and outputting the class mode for classification or the mean prediction for regression. One of its notable strengths is its ability to handle large datasets with higher dimensionality. It can also maintain accuracy when a large proportion of the data is missing. By leveraging bagging techniques and feature randomness, the Random Forest algorithm reduces overfitting and augments generalization, making predictions more stable and consistent over diverse data sets.

- Explainable Machine Learning (XML)
  In the burgeoning field of machine learning, explainability has emerged as a pivotal aspect, especially when the outcomes of machine learning models influence critical decisions. Techniques such as SHAP (SHapley Additive exPlanations) (Goldstein et al., 2015; Lundberg & Lee, 2017) are examples of XML methods to decipher the intricate decision-making of traditionally 'opaque' models. SHAP, is rooted in cooperative game theory and allocates contribution values for each feature, delineating how much each feature contributes to a particular prediction. Both techniques shed light on model behavior, instilling trust and enhancing model transparency(Doshi-Velez & Kim, 2017).

Evaluation Metrics

A suite of evaluation metrics was relied upon to appraise the performance of machine learning models rigorously, each furnishing distinct insights into various facets of model efficacy. The selected metrics encompass accuracy, precision, recall, and F1-score. An elucidation of these metrics is as follows:

- Accuracy
  This metric provides an overarching evaluation of the model by measuring the ratio of correct predictions to the total number of predictions made. Whereas accuracy is a straightforward metric, it may not be particularly informative in imbalanced datasets where one class significantly outnumbers the other.

- Precision
  Precision is important when the consequences of false positive predictions are significant. It measures the accuracy of positive predictions by calculating the ratio of true positives to the total number of true positives and false positives.

- Recall
  Recall becomes paramount when the implications of false negatives are substantial. Recall measures the ratio of true positives to the sum of true positives and false negatives. It focuses on the model's capability to identify all relevant instances correctly.

- F1-score
  The F1-score is the harmonic mean of precision and recall, balancing the two metrics to provide a more comprehensive view of model performance. It is especially useful when dealing with imbalanced class distributions, as it considers both precision and recall.

4. RESULTS

Descriptive Analysis

A descriptive analysis was conducted to gain an understanding of the characteristics of the data. Figure 2, which depicts employee attrition, shows that 900 employees did not experience attrition, while the other 1646 employees experienced attrition. It indicates that the number of employees who experienced attrition is higher than those who did not. This finding highlights the importance of understanding the factors contributing to employee attrition to develop effective retention strategies.

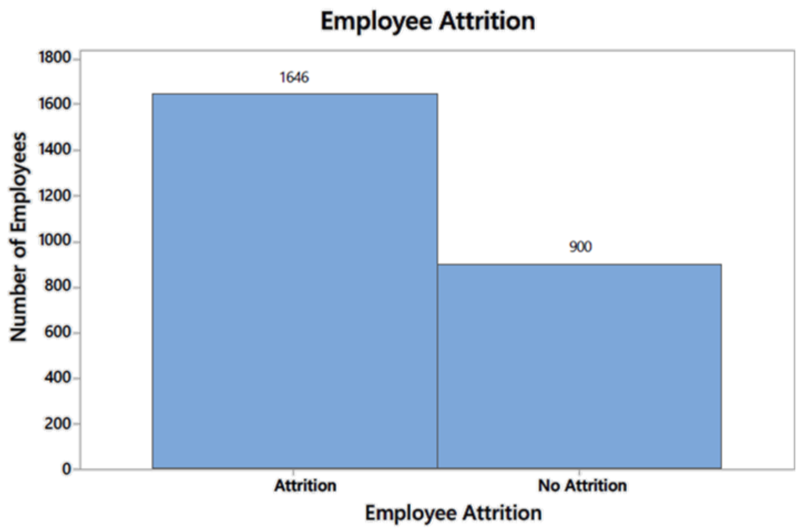

FIGURE 2. Employee Attrition Bar Chart

Descriptive details on variables with ratio scale are shown in Table 2. These variables include distance, tenure, and age. The table provides summary statistics, such as each variable's mean, standard deviation, and quartiles.

TABLE 2. Descriptive Analysis of Ratio Scale Variables

Variable	N	Mean	SE Mean	StDev	Min	Q1	Median	Q3	Max
Distance	2546	39,659	0,844	42,589	5	14	24	34	190
Length of Working	2546	8,621	0,192	9,676	0	2	3	11	37
age	2546	32,749	0,226	11,394	19	24	28	38	73

Evaluation of Machine Learning Models

Table 3 summarizes the performance metrics for Logistic Regression, Naive Bayes, and Random Forest to provide a clearer picture of the performance of each method tested.

TABLE 3. Machine Learning Algorithm Performance Metrics

Method	Presisi Kelas 0	Presisi Kelas 1	Recall Kelas 0	Recall Kelas 1	F1-score Kelas 0	F1-score Kelas 1	Accuration
Regresi Logistik	0.84	0.94	0.90	0.91	0.87	0.92	0.904
Naive Bayes	0.84	0.93	0.88	0.91	0.86	0.92	0.900
Random Forest	0.90	0.95	0.91	0.94	0.90	0.95	0.931

Based on the summary of results in Table 2, there is variation in the performance of the three models. Specifically, the Random Forest is the most efficient model regarding precision, recall, F1 score, and accuracy. High precision indicates that the model can accurately classify positive classes, while high recall indicates the model's ability to identify most or all positive cases. The F1 score provides an overview of the balance between precision and recall, which is also important to ensure that the model does not prioritize one over the other. Accuracy is a general indicator of model performance that shows the proportion of correct predictions in the overall data.

SHAP Analysis

In the era of increasingly massive data, data-driven decision-making has become the norm in various fields. However, decisions generated by machine learning models are often difficult for humans to understand because these models tend to be "black boxes" (Samek et al., 2017). In this context, the importance of explainable machine learning emerges as a means to bridge the gap between prediction accuracy and model interpretation. Explainable machine learning aims to create models that are capable of making accurate predictions and can be easily interpreted by humans. Thus, stakeholders can understand the reasoning behind the predictions generated by the model and feel more confident in acting based on these insights.

One technique in explainable machine learning is the SHAP (SHapley Additive exPlanations) method. The SHAP method allows measuring each feature's influence on the prediction made by the model. SHAP uses the concept of Shapley values from cooperative game theory to allocate each feature's fair "contribution" to the prediction. The SHAP method will be used to analyze the influence of attributes on attrition predictions generated by three machine learning models: Logistic Regression, Naive Bayes, and Random Forest. Through this analysis, it is hoped that an understanding of how each attribute contributes to attrition prediction to make informed and data-driven decisions can be gained. The random data used to calculate the force plot is shown in Table 4.

TABLE 4. Random data used to calculate the force plot

Attribute	Value
Attrition	1
permanent	0
distance	19
length of working	2

Attribute	Value
age	64
marital status	1
gender	1
position	0
education	1

In this context, focus will be given to the Random Forest model in the SHAP analysis, as it was found from previous evaluations that this model had the best performance in prediction with an accuracy of 93.1%. In the Random Forest model, it was found that temporary staff position (position_0), permanent status (permanent_1), and non-permanent employee status (permanent_0) were features with the highest average absolute SHAP values, indicating that these features were most influential in predicting employee attrition. The plot results in Figures 3 and 4 show that the Random Forest model offers a stronger and more varied understanding due to its non-linear nature. By combining the predictive power of Random Forest and the interpretation of SHAP values, a model can be created that performs well in terms of accuracy and provides insights into features that influence employee attrition.

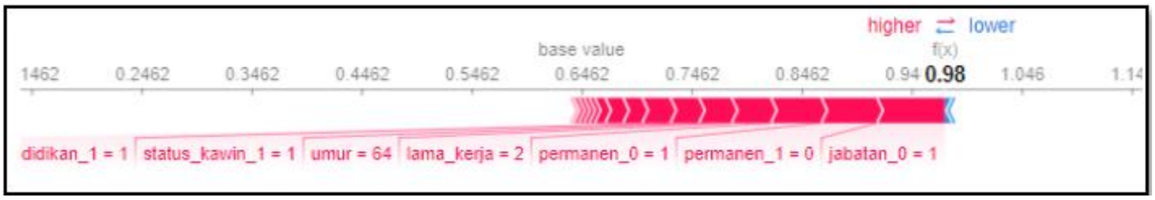

FIGURE 3. Force Plot Random Forest

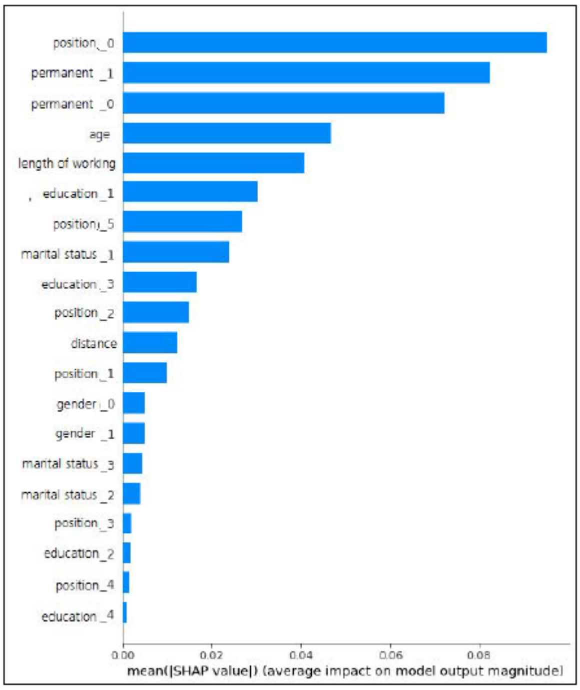

FIGURE 4. Summary Plot Random Forest

5. CONCLUSIONS

In the exploration of employee attrition in the rollingstock sector using advanced analytical methods, the following conclusions have been deduced:

- Among employed methodologies - Logistic Regression (LR), Naive Bayes (NB), and Random Forest (RF) - the latter emerged superior. Specifically, the Random Forest algorithm showcased an optimal performance, evidenced by an accuracy metric of 93.1% and consistently higher precision, recall, and F1-score scores. This result underscores the algorithm's heightened capability in detecting and predicting employee attrition nuances compared to its counterparts.
- Employing the SHAP framework, a subset of the Explainable Machine Learning domain, it was discerned that attributes such as job position, employment status (permanent versus contractual), and marital status were prominent influencers in determining attrition tendencies. It implies a complex interplay between organizational variables and external demographic factors influencing employees' decisions to remain or exit.
- Recognizing these influencers can empower organizations to devise data-driven, targeted strategies. For instance, delineating clear career pathways, considering transitions from contractual to permanent roles for deserving candidates, and sculpting policies that promote work-life balance emerge as pivotal initiatives.
- From a managerial perspective, these conclusions necessitate a balanced approach to employee retention. Beyond just professional incentives, there's an imperative need to address the holistic needs of employees, interlacing both their career aspirations and personal well-being. Integrating such a nuanced approach, anchored by the insights from the Random Forest algorithm, can pave the way for enhanced employee loyalty, culminating in sustained organizational success.

REFERENCES

Alduayj, S. S., & Rajpoot, K. (2018). Predicting Employee Attrition using Machine Learning. *2018 International Conference on Innovations in Information Technology (Iit)*, 93–98.

Bhartiya, N., Jannu, S., Shukla, P., & Chapaneri, R. (2019). Employee Attrition Prediction Using Classification Models. *2019 IEEE 5th International Conference for Convergence in Technology, I2CT 2019*. https://doi.org/10.1109/I2CT45611.2019.9033784

Cramer, J. S. (2005). The Origins of Logistic Regression. *SSRN Electronic Journal*. https://doi.org/10.2139/SSRN.360300

Doshi-Velez, F., & Kim, B. (2017). A Roadmap for a Rigorous Science of Interpretability. *ArXiv Preprint ArXiv:1702.08608v1*.

Fallucchi, F., Coladangelo, M., Giuliano, R., & De Luca, E. W. (2020). Predicting employee attrition using machine learning techniques. *Computers*, *9*(4). https://doi.org/10.3390/computers9040086

Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D., & Fernández-Delgado, A. (2014). Do We Need Hundreds of Classifiers to Solve Real World Classification Problems? *Journal of Machine Learning Research*, *15*, 3133–3181. http://www.mathworks.es/products/neural-network.

Gao, X., Wen, J., & Zhang, C. (2019). An Improved Random Forest Algorithm for Predicting Employee Turnover. *Mathematical Problems in Engineering*, *2019*. https://doi.org/10.1155/2019/4140707

Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation. *Journal of Computational and Graphical Statistics*, *24*(1). https://doi.org/10.1080/10618600.2014.907095

Immaneni, K. M., & Sailaja, V. N. (2019). A study on factors effecting the employees attrition in hotel industry with reference Hyderabad. *International Journal of Management*, *10*(6). https://doi.org/10.34218/IJM.10.6.2019.017

Kakad, S., Kadam, R., Deshpande, P., Karde, S., & Lalwani, R. (2020). Employee Attrition Prediction System. *IJISET-International Journal of Innovative Science, Engineering & Technology*, *7*(9).

Kaur, J., & Padmanabhan, H. K. (2019). A Study of Employee Attrition Rate at Selected Manufacturing Industry. In *International Journal of Research and Innovation in Social Science*.

Khera, S. N., & Divya. (2019). Predictive Modelling of Employee Turnover in Indian IT Industry Using Machine Learning Techniques. *Vision*, *23*(1). https://doi.org/10.1177/0972262918821221

Kohavi, R., & Edu, S. (1993). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, *2*.

Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, *2017-December*.

Mishra, S. N., Lama, D. R., & Pal, Y. (2016). Human Resource Predictive Analytics (HRPA) For HR Management In Organizations. *International Journal of Scientific & Technology Research*, *5*(5).

Ponnuru, S. R. (2020). Employee Attrition Prediction using Logistic Regression. *International Journal for Research in Applied Science and Engineering Technology*, *8*(5). https://doi.org/10.22214/ijraset.2020.5481

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

Probst, P., Boulesteix, A.-L., & Bischl, B. (2019). Tunability: Importance of Hyperparameters of Machine Learning Algorithms. *Journal of Machine Learning Research*, *20*, 1–32. http://jmlr.org/papers/v20/18-444.html.

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

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "why should i trust you?" explaining the predictions of any classifier. *NAACL-HLT 2016 - 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Demonstrations Session*. https://doi.org/10.18653/v1/n16-3020

Samek, W., Wiegand, T., & Müller, K.-R. (2017). *Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models*. http://arxiv.org/abs/1708.08296

Srivastava, P. R., & Eachempati, P. (2021). Intelligent Employee Retention System for Attrition Rate Analysis and Churn Prediction: An Ensemble Machine Learning and Multi- Criteria Decision-Making Approach. *Journal of Global Information Management*, *29*(6). https://doi.org/10.4018/JGIM.20211101.oa23

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*. https://doi.org/10.1109/ACCESS.2021.3074559

Zhao, Z., & Wang, X. (2018). Multi-Segments Naïve Bayes Classifier in Likelihood Space. *IET Computer Vision*, *12*(6), 882–891. https://doi.org/10.1049/iet-cvi.2017.0546

How to cite this article:

Itqon, M. Purnomo, J. D. T. (2022). Employee Attrition Prediction Using Machine Learning in Rolling Stock Manufacturing Company. *Jurnal Teknobisnis*, 8(1): 74-85. DOI: 10.12962/j24609463.v8i1.941