

Performance Analysis of Random Forest Algorithm For Health Index Classification Of PLN Distribution Transformer

Meilly Sumarni¹, Erna Budhiarti Nababan^{2*}, and Opim Salim Sitompul³

^{1,2,3}Data Science And Artificial Intelligence, Universitas Sumatera Utara, Medan, Indonesia

*Corresponding author: ernabrn@usu.ac.id

ABSTRACT

The rapid growth in Indonesia's electricity demand has increased the urgency for effective asset management in PLN's distribution network, particularly for distribution transformers. As critical infrastructure, these transformers require accurate condition monitoring to reduce operational risks and optimize maintenance planning. Although PLN has adopted a Health Index (HI) system using parameters such as oil condition and load current, its maintenance strategy remains largely schedule-based and lacks strong data-driven support. This study addresses the gap by applying the Random Forest (RF) algorithm to classify transformer health conditions based on historical inspection data. Random Forest is an ensemble machine learning algorithm that builds multiple decision trees and merges their results to improve predictive accuracy and control overfitting. It is especially suitable for classification tasks involving complex and imbalanced datasets, making it an ideal choice for utility asset diagnostics. The objective is to develop an accurate, interpretable, and scalable model that supports the transition to condition-based maintenance. Performance evaluation was conducted using two data partitioning schemes: 70% training, 15% validation, and 15% testing; and 80% training, 10% validation, and 10% testing. The RF model consistently outperformed Logistic Regression across all metrics, achieving up to 99.54% accuracy, 99.53% precision, 99.54% recall, and 99.52% F1-score. It also demonstrated lower Mean Squared Error (MSE)—0.0194 and 0.0135—compared to Logistic Regression's 0.0575 and 0.0613. Feature importance analysis identified oil condition and load current as key predictors, while spatial visualization enabled intuitive geo-mapping of transformer health status. These results demonstrate that Random Forest is robust in handling imbalanced utility data and effective for practical deployment in asset health classification. The findings offer valuable insights to support PLN's digital transformation toward predictive maintenance and intelligent asset management.

Keywords: Transformer Health Index, Random Forest Classification, Asset health classification, predictive asset management.

1. INTRODUCTION

The rapid increase in electricity demand in Indonesia has significantly expanded the customer base and intensified the operational complexity of the distribution network managed by PLN, the national electricity company. These developments underscore the critical need for efficient asset management, especially for distribution transformers, which are critical in ensuring stable electricity delivery to end users. Due to their strategic importance and high capital value, optimising transformer maintenance is critical to reducing operational risks, extending asset life, and improving system reliability.

To support asset condition evaluation, PLN has adopted a Health Index (HI) framework that integrates various condition monitoring parameters, such as oil leakage, load current, grounding integrity, and physical condition, into a composite score for each transformer. The Health Index is a numerical representation of asset condition, typically scaled from 0 (Bad)

to 4 (Good), based on predefined parameters established by PLN. This scoring system enables standardised comparison across assets and serves as a reference for prioritising maintenance actions, scheduling replacements, and guiding risk-based financial planning. However, PLN's current asset management practices are still at the schedule-based level (maturity level 2), where routine maintenance is based on asset type and age, not actual condition data. Transitioning to condition-based maintenance (maturity level 3) demands accurate, scalable, data-driven HI prediction models.

Recent studies have demonstrated the effectiveness of machine learning (ML) techniques in improving transformer diagnostics and HI estimation. For example, (Selva et al., 2021) applied statistical distribution models to estimate transformer health index, which improved the objectivity of condition assessment. (Alqudsi & El-Hag, 2019) further highlighted the potential of ML in transformer HI prediction by combining feature selection and classification strategies that resulted in strong predictive performance across datasets from various utilities. Similarly, (Xiao et al., 2020) utilised Random Matrix Theory (RMT) to extract health-relevant features from high-dimensional data, demonstrating the benefits of data-driven models in real-time status assessment. In this study, all machine learning models were developed using Python, a widely adopted programming language in the data science community due to its rich ecosystem of ML libraries such as Scikit-learn, Pandas, and Matplotlib, which support efficient model development, evaluation, and visualization.

Among ML algorithms, Random Forest (RF) has emerged as a powerful tool due to its resistance to overfitting, ability to manage imbalanced data, and superior performance in classification tasks. Prasojo et al. (2023) showed that RF, when combined with data balancing techniques such as SMOTE (Synthetic Minority Over-sampling Technique), significantly improved the accuracy of fault classification in transformers, demonstrating its potential for practical application in utility diagnostics. SMOTE is a popular data augmentation method that synthetically generates new samples for underrepresented classes by interpolating between existing minority class instances, thus helping to reduce class imbalance during model training. However, despite these advances, the application of RF specifically for transformer HI classification in the context of PLN operations remains unexplored.

Therefore, this study investigates the performance of the Random Forest algorithm in classifying the Health Index of distribution transformers based on PLN's historical inspection data. The goal is to develop an accurate, interpretable, and scalable classification model that facilitates the shift to condition-based maintenance. Thus, this research not only contributes to the academic understanding of predictive maintenance strategies but also provides actionable insights to improve transformer management in real-world electric utility environments.

2. LITERATURE REVIEW

Recent advancements in the classification of transformer Health Index (HI) have highlighted the critical role of data-driven approaches for predictive maintenance and operational reliability in power systems. Early studies employed hybrid intelligent systems such as the Neuro-Fuzzy model proposed by (Kadim et al., 2018) which demonstrated classification performance comparable to manual assessment schemes. However, the dependence on expert-defined rules limited scalability for large-scale utility datasets.

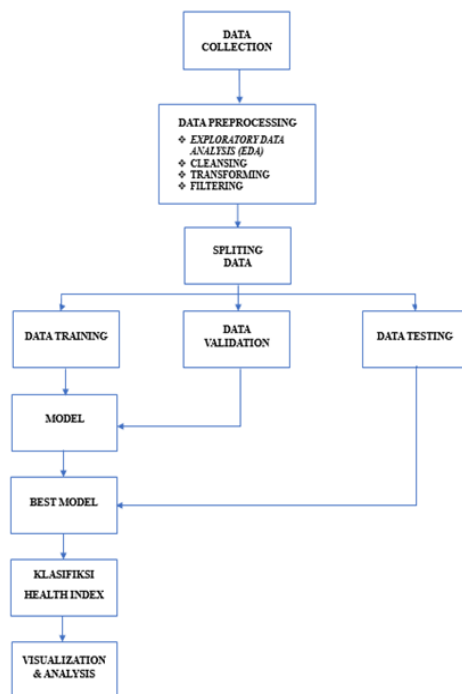
Addressing high-dimensionality issues, (Xiao et al., 2020) applied Random Matrix Theory (RMT) to extract statistical patterns from monitoring data, though the approach lacked adaptability for real-time classification and provided limited insight into feature contribution.

Subsequent efforts have incorporated machine learning models with embedded or wrapper-based feature selection. (Kadim et al., 2018) utilized Artificial Neural Networks (ANN) with regression-based selection, yielding generalizable input features, yet interpretability remained constrained by the ANN's black-box nature. Similarly, (Benhmed et al., 2018) evaluated multiple feature selection techniques—Information Gain, ReliefF, and CFS—combined with classifiers such as SVM and Random Forest, identifying physicochemical parameters as key predictors. Nevertheless, computational complexity and overfitting risks persisted, particularly in high-dimensional settings. (Liu et al., 2019) further explored tree-based embedded selection; however, model instability across varying datasets limited the robustness of feature relevance.

In parallel, Pirizadeh et al. (2023) proposed a cost-sensitive SVM framework tailored to imbalanced and high-dimensional data. SVM, or Support Vector Machine, is a supervised learning algorithm commonly used for classification tasks. It works by finding the optimal hyperplane that best separates different classes in the feature space, and is particularly effective in high-dimensional and complex datasets. Despite its theoretical strength, validation in this study was limited to synthetic environments, raising concerns about real-world generalization. Complementary findings from other domains reinforce the efficacy of ensemble methods. Surbakti et al. (2023) demonstrated Random Forest's superior performance in economic classification tasks involving GDP prediction, while Surbakti et al. (2023) successfully applied it for health-related outbreak forecasting, emphasizing its resilience to noisy, imbalanced data. While these studies affirm the potential of ensemble models and selective feature strategies, few have comprehensively assessed Random Forest's performance using actual transformer monitoring data from utility networks. The present study addresses this gap by evaluating Random Forest in conjunction with feature importance analysis, offering a scalable and interpretable approach for transformer health classification. This contributes not only to improved classification accuracy but also to actionable insights for asset management within digitalized utility infrastructures.

3. METHODS

3.1. Research Framework



The figure above shows the research framework consisting of several key stages. The process starts with the collection of field data obtained through monitoring and inspection systems. Next, a data preprocessing stage is performed which includes data exploration, cleaning, transformation, and filtering. After that, the data is divided into three parts: training data, validation data, and testing data. Training data is used to build the model, while validation data is used to select the best model. The best model is then used to classify the Health Index of the transformer. The final results of this process are visualised and analysed to generate insights that are useful for decision-making.

3.2. Dataset Description

Table 1. Transformer Dataset

UNIT	WO NUM	LOC ATIO N	LON GITU DEX	LATI TUD EY	ASS ETN UM	MEAS URE DATE	DESCR IPTION	MEAS UREMENT VALUE	OBSER VATIO N	METER NAME	METER DESC	MEA SURE UNIT ID	C O N D I T I O N	HICON D VALUE
UP3. 6112	3746 89	GD0 1. 0344 2266	9554 3966 5900	5378 7777 610	2440 620	2019- 08-26- 12.11.1 2.97400 0	244062 0 TRAFO 18 TANO H ABEE	10000	NaN	BESAR ARUSN ETRAL TR	Besar Arus Netral Tr	%	2	Cukup

UNIT	WO NUM	LOC ATION	LON GITUDE	LATI TUD EY	ASS ETN UM	MEAS URE DATE	DESCR IPTION	MEAS UREMENT VALUE	OBSER VATION	METER NAME	METER DESC	MEA SURE UNIT ID	CON DIT ION	HICON D VALUE
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
UP3. 6114	9350 16	GD0 1. 0881 6112	9613 0220 0000	4152 1640 000	6090 331	2019- 11-08- 16.21.4 5.00000 0	609033 1 TRAFO GD JP 103	0.5	NaN	PENTA NAHA N	Pentana han Trafo	OHM	3	Baik

Table 1 presents a subset of the raw inspection data collected from PLN's distribution transformers. Each row represents an individual measurement record from a specific transformer unit. The attributes include metadata such as unit code, location coordinates (longitude and latitude), asset number, and inspection timestamp, as well as technical parameters like measured values, observation types, meter names, and descriptive notes from field personnel. The 'CONDITION' and 'HICON' columns represent qualitative assessments of transformer health, which are later mapped to numerical Health Index classes (0–4). This raw data forms the foundation for the preprocessing, feature engineering, and classification steps conducted in this study. The datasets used in this study were sourced from internal inspection reports and real-time monitoring systems of PLN distribution transformers. There are 15 attributes and 4 classes in the data. Data collection involved two main methods: (1) manual input by field personnel based on routine inspections, and (2) sensor-assisted measurements, specifically for temperature and voltage parameters, which are recorded and uploaded into the system.

Each data instance corresponds to an individual transformer unit and includes the following attributes: unit identifier, location (latitude and longitude), inspection date, measured value, and observation record. These features are key indicators used to assess transformer health and support the subsequent classification process.

3.3. Data Preprocessing

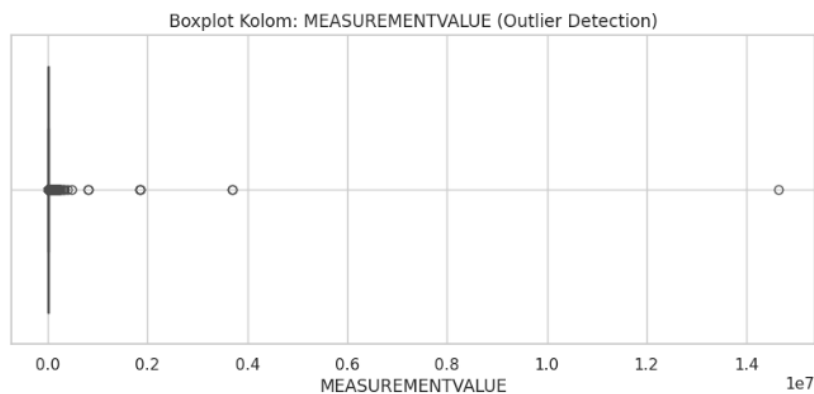
Data preprocessing is a crucial stage to ensure data quality and consistency before machine learning modelling. This process includes data cleaning to deal with missing values and outliers, data transformation through encoding and normalisation, and feature selection to retain the most relevant variables. An important part of preprocessing is Exploratory Data Analysis (EDA), which is an exploratory process to understand the data structure, identify hidden patterns, and detect anomalies with the help of descriptive statistics and visualisation. According to (Fernández et al., 2021), EDA enables early analysis of data quality and supports verification of initial hypotheses through statistical methods and graphical representations. This stage plays a role in improving the understanding of the data characteristics as well as being the basis for selecting appropriate algorithms. Overall, data preprocessing aims to improve the accuracy and effectiveness of the model by presenting the data in an optimal form.

for the training and testing process. The Exploratory Data Analysis (EDA) carried out is as follows:

- **Descriptive Statistics**

Descriptive statistics are used to understand the basic elements of the dataset before starting the further analysis process. From the analysis, it is known that the column 'Observation' has 6,951 rows, while 'Matername' has 11,683 rows, which indicates the possibility of missing values. In addition, categorical columns such as 'Observation', 'Matername', and 'Meterdesc' have mean, standard deviation, min, and max values that are NaN because they are not numerical, so they cannot be measured using a direct numerical statistical approach.

- **Outlier Detection**, is performed on the "MEASUREMENTVALUE" column using boxplot and Interquartile Range (IQR) approach

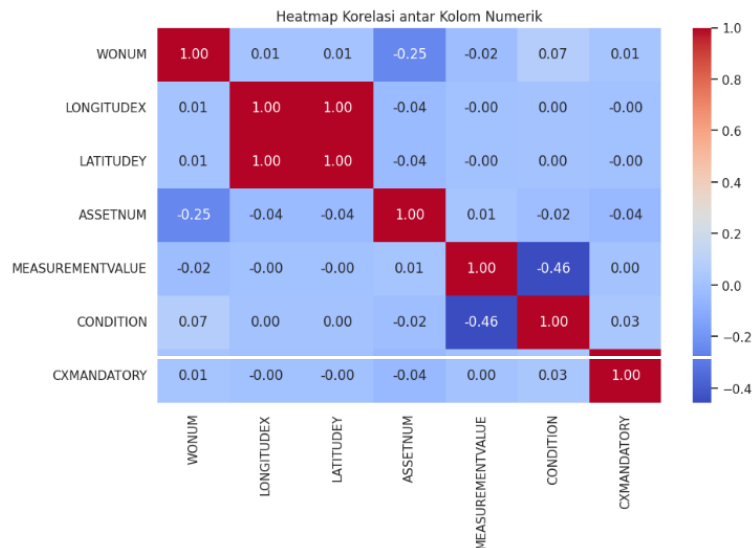


The results show that there are 803 data identified as outliers. These extreme values deviate significantly from the majority distribution, with some values approaching 14 million. This indicates the possibility of anomalies or measurement improprieties in some of the data that could affect model performance if not properly addressed.

- **Correlation Visualisation (Heatmap)**

Correlation analysis between numerical features was performed using heatmap visualisation to identify linear relationships between variables.

The heatmap shows that most feature pairs have weak correlations. The perfect correlation (1.00) between 'LONGITUDEX' and 'LATITUDEY' indicates redundancy and potential multicollinearity. 'MEASUREMENTVALUE' shows a moderate negative correlation with 'CONDITION' (-0.46), which highlights its relevance in predicting the health index. Meanwhile, features such as 'WONUM', 'ASSETNUM', and 'CXMANDATORY' showed negligible correlation, indicating that they are identifiers and may require transformation before modelling.



After EDA, data preparation involved cleansing, transformation, and filtering. Missing values in key columns ('MeasurementValue', 'Observation', etc.) were handled using simple imputation ('-' placeholder). Feature engineering was applied through rule-based mapping and iteration to generate two derived columns: 'Inspection' (from 'MeterName') and 'Condition' (from numerical or observation data). The transformed data was then merged and refined through filtering, retaining only 12 key transformer parameters. This process reduced the dataset from 11,683 to 9,627 rows, ensuring only relevant and clean data was used for model training.

3.4. Data Splitting

The dataset was partitioned using two splitting schemes to evaluate the model's generalization performance: Scheme 1 (70% train, 15% validation, 15% test) and Scheme 2 (80% train, 10% validation, 10% test). This approach allows assessment of model sensitivity to training set size and supports robust performance analysis.

- **Training Phase:**
The training set was used to build predictive models based on input features (*MeasurmenValue*, *Observation*, *Matername*) and the target variable (*HEALTH_INDEX*). Two models were trained: a baseline logistic regression for reference, and a Random Forest classifier selected for its capability to handle complex, nonlinear patterns and reduce overfitting.
- **Validation Phase**
The validation set was used for hyperparameter tuning and model selection. A hold-out validation technique was employed to ensure unbiased performance estimation. Model predictions on validation data were evaluated using Mean Squared Error (MSE), and the results were visualized to compare model effectiveness.
- **Testing Phase**
Final evaluation was conducted on the tet set, which remained unseen throughout training and validation. The model's performance was assessed using classification metrics including accuracy, precision, recall, and F1-score, derived from the confusion matrix.

This step ensures that the model generalizes well and performs reliably on new, real-world data.

3.5. Health Index Classification

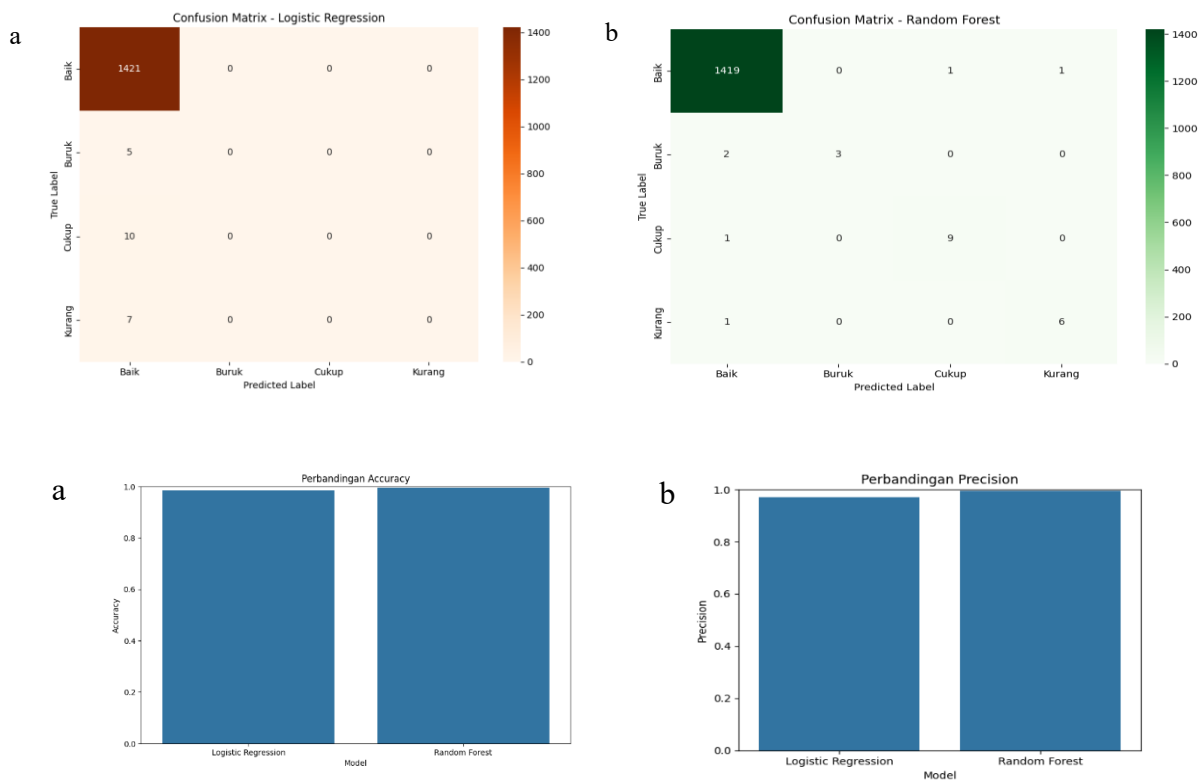
After the training, validation, and testing stages are completed and the best model (Random Forest) is obtained, the next process is the classification of the Health Index (HI) value. This process is carried out based on the prediction results of the test data, with the aim of determining the transformer condition class. Classification is done through a grouping method based on the majority value as follows:

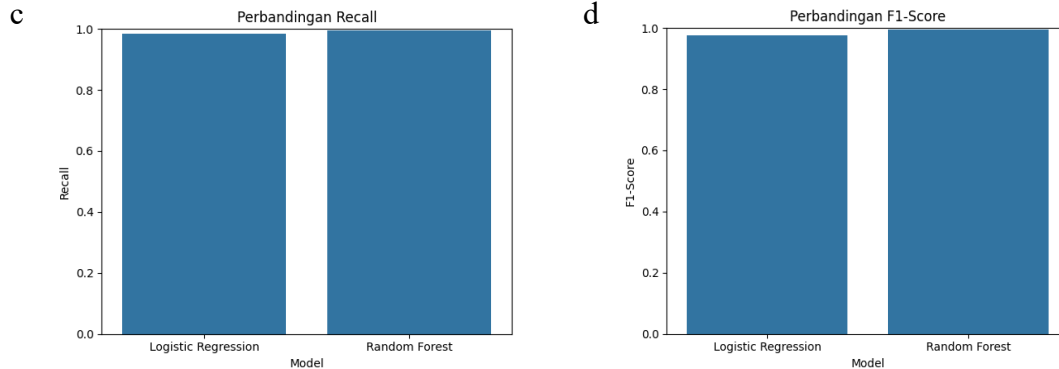
- Majority 0: HI = Bad
- Majority 1: HI = Poor
- Majority 2: HI = Fair
- Majority 3: HI = Good

4. RESULTS

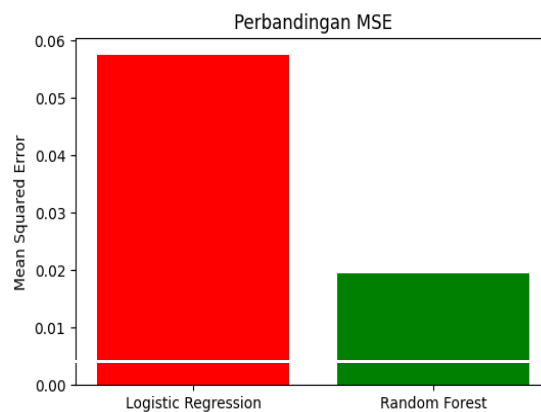
4.1. Modelling the Health Index with 70% Training Data, 15% Validation Data, and 15% Testing Data

The performance evaluation of the classification model was conducted using confusion matrix and key evaluation metrics (accuracy, precision, recall, and F1-score).

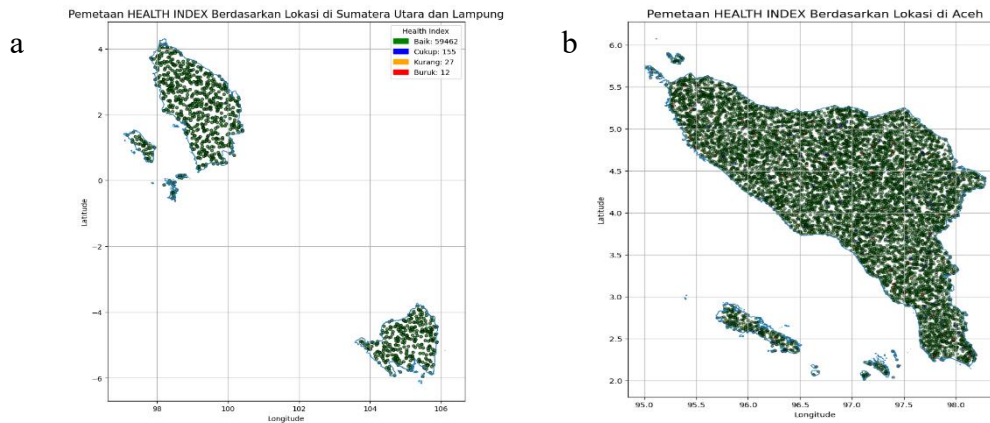




The confusion matrix results showed that Random Forest was able to classify most of the data very accurately, especially in the majority class ‘Good’ with 4,731 out of 4,811 data successfully predicted correctly. Although there were still misclassifications in the minority classes, the overall performance of the model remained high. Compared to Logistic Regression, Random Forest showed significant advantages in all evaluation metrics, with accuracy of 99.54%, precision of 99.53%, recall of 99.54%, and F1-score of 99.52%. These values indicate that the model is not only able to accurately identify the majority class, but also has a more balanced performance in detecting the minority class. Thus, Random Forest proved to be a more reliable algorithm in classifying the Health Index of distribution transformers on imbalanced data.



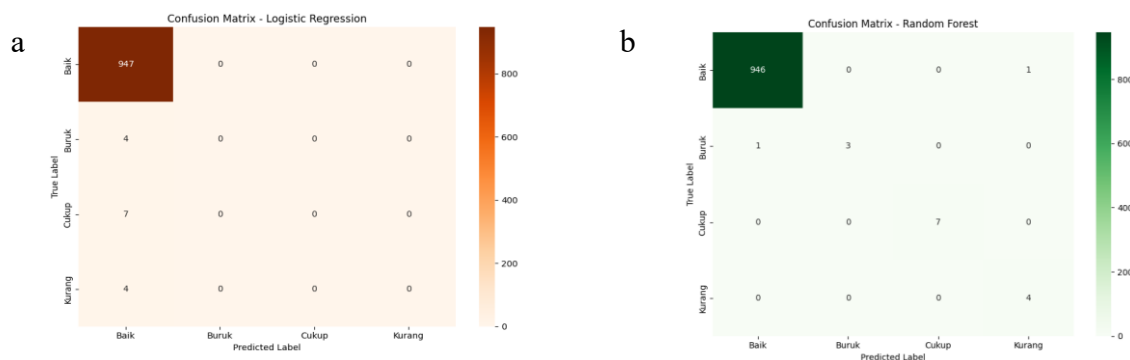
The Mean Square Error (MSE) results show that the Random Forest model (MSE = 0.0194) has a lower prediction error rate than Logistic Regression (MSE = 0.0575). With a smaller MSE value, Random Forest is proven to provide more accurate predictions and is closer to the actual data.

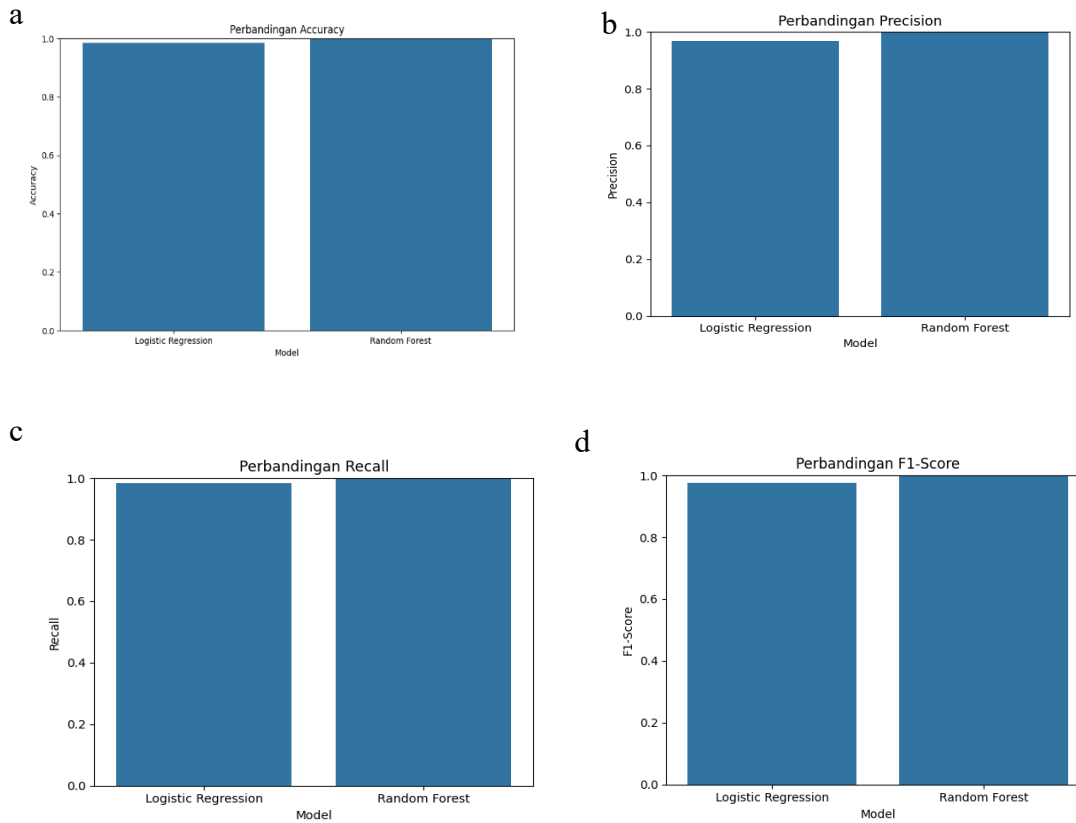


Spatial visualisation was used to map the transformer health index classification results. This final stage facilitates interpretation by displaying asset locations and their respective health index classes on a map.

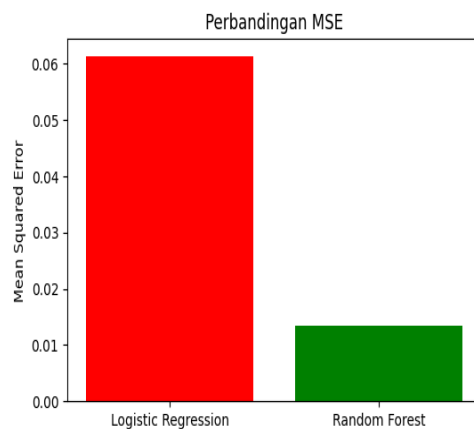
4.2. Modelling the Health Index with 80% Training Data, 10% Validation Data, and 10% Testing Data

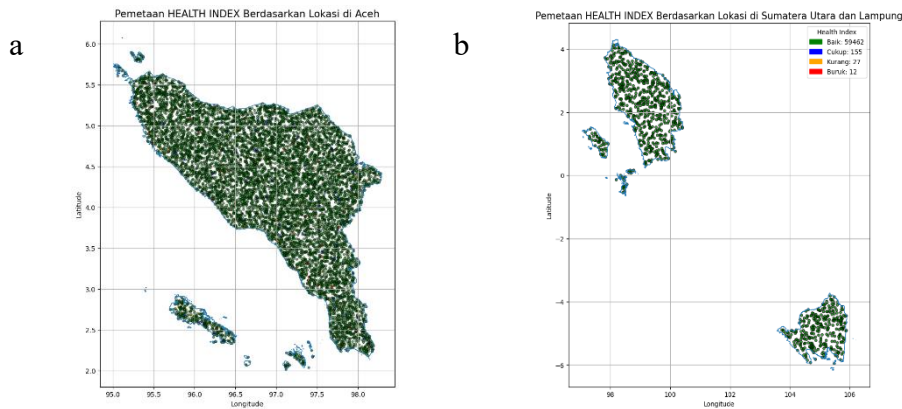
After the classification process is carried out with a data split of 80% for training, 10% for validation, and 10% for testing, model evaluation using confusion matrix is essential to assess classification performance. Confusion matrix compares the model's predicted results with the actual labels, identifying types of errors such as false positive and false negative. Based on the confusion matrix results, the model is able to predict very well in each class, for example for the Good class, 4731 out of 4811 data were classified correctly, although there were some data that were misclassified to other classes. Similarly, for the Bad, Fair, and Poor classes, the model showed a high level of accuracy although there were still a few prediction errors. Evaluation of the Random Forest model performance metrics also showed excellent results with an accuracy value of 99.54%, precision 99.53%, recall 99.54%, and F1-score 99.52%. These results show that the model is able to classify almost all data correctly, minimise prediction errors, and has a good balance between precision and recall. Thus, the 80-10-10 data split produced an effective and reliable Health Index classification model for various classes in the test data.





In the test results, the baseline model using Logistic Regression produces an MSE value of 0.0613, while the Random Forest model has a much lower MSE of 0.0135. This shows that the Random Forest model is able to minimise prediction errors significantly compared to the baseline model. As part of the interpretation of the results, a visualisation of the Health Index prediction mapping against the distribution transformer locations was also conducted.





5. CONCLUSION

This study investigates the Health Index classification of PLN distribution transformers using the Random Forest algorithm, with comparative evaluation against a basic Logistic Regression model. Two different data partitioning strategies were applied: (1) 70% training, 15% validation, and 15% testing; and (2) 80% training, 10% validation, and 10% testing. In both configurations, Random Forest consistently outperformed Logistic Regression in all key evaluation metrics. The model achieved outstanding classification results, with 99.54% accuracy, 99.53% precision, 99.54% recall, and 99.52% F1 score. These values indicate that Random Forest not only performs very well in identifying the dominant class ('Good'), but also shows resilience in detecting the minority class with minimal misclassification. In terms of regression-based performance, Random Forest recorded significantly lower Mean Square Error (MSE) values - 0.0194 in the 70-15-15 split and 0.0135 in the 80-10-10 split - compared to Logistic Regression, which produced MSE values of 0.0575 and 0.0613 respectively. This confirms the superior predictive accuracy and generalisation ability of Random Forest in handling complex and imbalanced data sets. The incorporation of spatial visualisation further enhances the interpretability of the results, allowing the health condition of distribution transformers to be intuitively mapped based on geographic location. This step is crucial in supporting asset maintenance planning and infrastructure risk management. In conclusion, the Random Forest algorithm shows strong potential as a reliable and accurate method for electrical asset health condition classification, particularly in operational environments where data imbalance and classification precision are critical. Future research could extend this approach by incorporating temporal data and real-time monitoring inputs, aiming at building adaptive models for predictive maintenance and smart grid applications.

REFERENCES

- Alqudsi, A., & El-Hag, A. (2019). Application of machine learning in transformer health index prediction. *Energies*, 12(14), 1–13. <https://doi.org/10.3390/en12142694>
- Benhmed, K., Mooman, A., Younes, A., Shaban, K., & El-Hag, A. (2018). Feature selection for effective health index diagnoses of power transformers. *IEEE Transactions on Power Delivery*, 33(6), 3223–3226. <https://doi.org/10.1109/TPWRD.2017.2762920>
- Fernández, D., Epifanio, I., & McMillan, L. F. (2021). Archetypal analysis for ordinal data. *Information Sciences*, 579, 281–292. <https://doi.org/10.1016/j.ins.2021.07.095>

- Kadim, E. J., Azis, N., Jasni, J., Ahmad, S. A., & Talib, M. A. (2018). Transformers health index assessment based on neural-fuzzy network. *Energies*, 11(4). <https://doi.org/10.3390/en11040710>
- Liu, H., Zhou, M., & Liu, Q. (2019). An embedded feature selection method for imbalanced data classification. *IEEE/CAA Journal of Automatica Sinica*, 6(3), 703–715. <https://doi.org/10.1109/JAS.2019.1911447>
- Pirizadeh, M., Farahani, H., & Kheradpisheh, S. R. (2023). Imbalance factor: a simple new scale for measuring inter-class imbalance extent in classification problems. *Knowledge and Information Systems*, 65(10), 4157–4183. <https://doi.org/10.1007/s10115-023-01881-y>
- Prasojo, R. A., Putra, M. A. A., Ekojono, Apriyani, M. E., Rahmanto, A. N., Ghoneim, S. S. M., Mahmoud, K., Lehtonen, M., & Darwish, M. M. F. (2023). Precise transformer fault diagnosis via random forest model enhanced by synthetic minority over-sampling technique. *Electric Power Systems Research*, 220. <https://doi.org/10.1016/j.epsr.2023.109361>
- Selva, A. M., Azis, N., Shariffudin, N. S., Kadir, M. Z. A. A., Jasni, J., Yahaya, M. S., & Talib, M. A. (2021). Application of statistical distribution models to predict health index for condition-based management of transformers. *Applied Sciences (Switzerland)*, 11(6). <https://doi.org/10.3390/app11062728>
- Surbakti, S., Hayatunnufus, & T. Henny Febriana. (2023). Prediction of Dengue Fever in Coastal Areas of North Sumatera (Kuala Namu and Belawan) With Random Forest and Support Vector Machine (SVM) Methods. *Data Science: Journal of Computing and Applied Informatics*, 7(2), 103–110. <https://doi.org/10.32734/jocai.v7.i2-14355>
- Xiao, F., Cui, Y., Wang, W., Zhang, C., He, X., Gu, J., & Cao, L. (2020). Online status assessment o distribution network transformer based o random matrix theory. *2020 IEEE 4th Conference on Energy Internet and Energy System Integration: Connecting the Grids Towards a Low-Carbon High-Efficiency Energy System, EI2 2020*, 4227–4231. <https://doi.org/10.1109/EI250167.2020.9346854>