Feature Selection for Predictive Asset Health Management in Power Plants Using Random Forest

Muhammad Ghafur[1], Mochamad Ikbal Arifyanto[2]

[1,2]Faculty of Mathematics and Natural Sciences
Computational Science Study Program
Bandung Institute of Technology
Bandung, Indonesia
Email: 20922308@mahasiswa.itb.ac.id

In the power generation industry, effective asset health management is essential for optimizing maintenance strategies and preventing costly failures. This study focuses on feature selection for predictive asset health management using the Random Forest algorithm, applied to PLN Indonesia Power's power plant assets. A dataset of 64,999 entries containing asset criticality metrics such as Maintenance Priority Index (MPI), Asset Criticality Ranking (ACR), and other relevant features was analyzed. Using Recursive Feature Elimination (RFE), MPI, ACR, and ACRRANK were identified as the most significant predictors of asset health. The model achieved an accuracy of 93.64%, demonstrating the importance of feature selection in improving prediction performance. This research provides valuable insights into optimizing maintenance efforts through targeted data-driven predictions.

Keywords: *feature selection, predictive maintenance, asset health management, Random Forest, power plants, criticality ranking, maintenance priority index*.

1 Introduction

Effective asset health management plays a crucial role in ensuring operational reliability and minimizing maintenance costs in power plants. With high-value assets and the need for uninterrupted energy supply, power generation companies like PLN Indonesia Power must prioritize optimal maintenance strategies. Traditionally, maintenance in power plants has been performed based on time-based schedules or manual assessments of asset conditions. However, these approaches often lead to inefficiencies, including unnecessary maintenance or failure to address critical issues in time, which can result in unplanned downtimes.

Predictive maintenance, supported by data-driven methods, offers an advanced solution by forecasting asset health conditions based on historical and real-time data. Machine learning algorithms, particularly ensemble methods like Random Forest, have gained prominence in this field due to their ability to handle complex

Received ________, Revised _________, Accepted for publication __________
 ISSN: xxxx-xxxx, DOI: 10.5614/xxxx

data and provide robust predictions [1][2]. Random Forest, an ensemble of decision trees, excels at capturing non-linear relationships and dealing with multivariate datasets [3][4], making it well-suited for predicting asset health in large-scale operations such as power plants.

A critical step in predictive modeling is feature selection, which involves identifying the most relevant variables that contribute to accurate predictions. Reducing the number of features can improve model performance, reduce computational costs, and offer clearer insights into the factors that significantly impact asset health. In this context, Recursive Feature Elimination (RFE) is a widely used technique [5][6][7][8] to rank the importance of features by iteratively removing the least significant ones and training the model multiple times to select the best-performing features.

This study aims to explore the feature selection process for predictive asset health management using the Random Forest algorithm, applied to PLN Indonesia Power's dataset. The dataset contains asset-related information such as Maintenance Priority Index (MPI), Asset Criticality Ranking (ACR), and other critical variables that reflect the condition of various power plant assets. By identifying the most influential features in predicting asset health, this research aims to improve predictive model accuracy and offer actionable insights for enhancing asset maintenance strategies in power plants.

The objectives of this study are twofold: (1) to identify the most critical features contributing to asset health prediction using Recursive Feature Elimination (RFE) and Random Forest, and (2) to develop a highly accurate predictive model to assist PLN Indonesia Power in optimizing their maintenance efforts and reducing the risk of unexpected failures.

2 Literature Review

2.1 Predictive Maintenance and Asset Management

Predictive maintenance is a key approach in modern asset management, offering significant advantages over traditional time-based or reactive maintenance strategies. It utilizes data-driven techniques to predict equipment failures and optimize maintenance schedules, thereby reducing downtime and maintenance costs. Jardine et al. In [9] demonstrated the superiority of condition-based maintenance over other methods, showing that predictive maintenance can enhance operational efficiency by targeting interventions only when needed.

For power plants, where asset reliability is crucial, predictive maintenance plays a critical role. As Lee and Ni in [10] highlighted, modern industrial systems increasingly rely on predictive maintenance to ensure continuous operation, particularly in power generation, where unscheduled outages can result in significant financial losses.

2.2 Machine Learning in Predictive Maintenance

Machine learning (ML) methods, such as decision trees, support vector machines (SVM), and neural networks, have been widely employed in predictive maintenance. These techniques analyze operational data to identify patterns and predict failures more accurately than traditional methods. Among these techniques, Random Forest stands out due to its robustness and accuracy in handling large, multivariate datasets.

Introduced by Breiman in [11], Random Forest is an ensemble learning method that aggregates the predictions of multiple decision trees to improve overall performance. In predictive maintenance, Random Forest has been particularly effective in dealing with noisy and imbalanced datasets as highlighted by Singh & Gupta in [12]. Zhao et al. in [13] applied Random Forest for failure prediction in power plants, demonstrating its ability to provide accurate, early warnings of potential failures.

2.3 Feature Selection in Predictive Models

Feature selection is an essential process in building efficient and accurate predictive models. It reduces model complexity, enhances interpretability, and improves predictive performance by identifying the most relevant features. Guyon et al. in [14] introduced Recursive Feature Elimination (RFE) as a technique for feature selection, particularly useful in models such as Random Forest. RFE iteratively removes the least important features based on their impact on model accuracy, streamlining the dataset and enhancing model performance.

In predictive maintenance, effective feature selection focuses the model on the key factors influencing asset health, such as maintenance priority, equipment condition, and operational history. Liu et al. in [15] applied feature selection techniques to improve fault detection models, and their findings indicated that reducing irrelevant features enhances prediction accuracy.

2.4 Feature Importance Analysis in Random Forest

Feature importance analysis is a critical aspect of Random Forest models, as it provides insight into which features contribute the most to predictive accuracy. The Random Forest algorithm computes feature importance by measuring how much each feature reduces uncertainty (measured via Gini impurity or entropy) across decision trees. This allows model developers to prioritize the most impactful variables and disregard less significant ones.

In predictive maintenance, understanding the importance of variables such as Maintenance Priority Index (MPI), Asset Criticality Ranking (ACR), and Equipment Criticality Ranking (ECR) is crucial for decision-making. Wang et al. in [16] emphasized the importance of feature analysis in predictive maintenance, showing how it can lead to more effective resource allocation and maintenance scheduling. Similarly, Zhang and Chen in [17] demonstrated that Random Forest-based feature importance analysis can enhance the prediction of transformer failures by identifying critical features that influence equipment degradation.

This study employs feature importance analysis using Random Forest to identify the most influential factors for asset health in power plants. By focusing on key features such as MPI and ACR, this research aims to improve the predictive maintenance model's performance and provide actionable insights for asset management at PLN Indonesia Power.

3 Methodology

This section outlines the approach used in building and evaluating the predictive model for asset health management in power plants, with a particular focus on the feature selection process. The methodology can be divided into several key steps: data collection, preprocessing, feature selection, model development, and model evaluation.

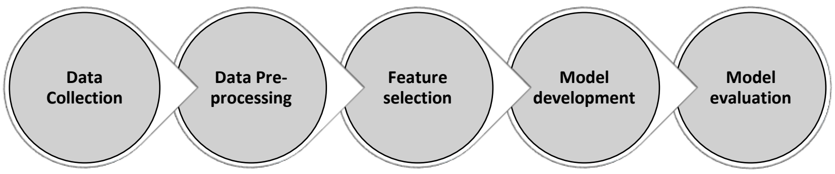

Figure 1 Research methodology

3.1 Data Collection

The dataset used in this study was collected from PLN Indonesia Power's data warehouse in January 2024. It contains 64,999 records of asset health data from various power plants, with the following features:

A. SiteID: Identifier for the power plant site

B. Status Mesin: Operational status of the machine (Healthy, Under Maintenance, etc.)

C. ACR (Asset Criticality Ranking): A ranking metric indicating the criticality of assets

D. ACRRANK: The numerical rank associated with the ACR

E. ECR (Equipment Criticality Ranking): Ranking of individual equipment within a system

F. SCR (System Criticality Ranking): Criticality of the entire system the equipment belongs to

G. MPI (Maintenance Priority Index): A score that determines the maintenance priority of the asset

H. Wellness: The last recorded health status of the asset

	STATUS	SCR	ECR	ACR	MPI	ACRRANK	WELLNESS	SITEID
0	OPERATING	7.18	4.33	31.09	31.09	69.70	HIJAU	MRC
1	OPERATING	7.18	5.55	39.85	39.85	27.27	HIJAU	MRC
2	OPERATING	7.18	4.56	32.74	32.74	60.61	HIJAU	MRC
3	OPERATING	7.18	4.42	31.74	31.74	66.67	HIJAU	MRC
4	OPERATING	7.18	4.33	31.09	31.09	69.70	HIJAU	MRC

	STATUS	SCR	ECR	ACR	MPI	ACRRANK	WELLNESS	SITEID
64994	INACTIVE	8.41	1.79	15.05	15.05	93.75	KUNING	JMB
64995	INACTIVE	8.41	1.79	15.05	15.05	93.75	KUNING	JMB
64996	INACTIVE	8.41	1.79	15.05	15.05	93.75	KUNING	JMB
64997	INACTIVE	8.41	1.79	15.05	15.05	93.75	KUNING	JMB
64998	OPERATING	6.78	4.45	30.17	30.17	16.55	HIJAU	CLG

Figure 2 Research dataset

This dataset is utilized to build a predictive model aimed at optimizing maintenance schedules and improving asset health management.

3.2 Data Pre-processing

Before building the predictive model, the data was cleaned and preprocessed to ensure the quality of input for machine learning:

A. Handling Missing Values: Any missing or incomplete records were either imputed using statistical techniques or removed if they were found to be uninformative.

B. Normalization: Continuous variables, such as ACR, ECR, SCR, and MPI, were normalized to ensure comparability and to improve model performance.

C. Categorical Encoding: Categorical features like Status Mesin were converted into numerical format using one-hot encoding to make them usable by machine learning algorithms.

D. Data Splitting: The dataset is divided into training data (90%) and testing data (10%) to ensure that the model can be tested with unseen data during the training process.

3.3 Feature Selection

The primary focus of this research is on selecting the most important features for the predictive model. Feature selection not only improves model accuracy but also enhances interpretability and reduces computational cost. In this study, Recursive Feature Elimination (RFE) is employed to select the most relevant features.

3.3.1 Recursive Feature Elimination (RFE)

RFE is a backward feature elimination method that works by recursively fitting the model and removing the least important features based on the model's performance. The process involves:

A. Initial Model Training: A Random Forest model is initially trained using all the features in the dataset.

B. Feature Ranking: The model provides an importance score for each feature based on the Gini impurity index. The feature that contributes the least to model performance is ranked the lowest.

C. Recursive Elimination: The least important feature is removed, and the model is retrained with the remaining features. This process is repeated until the desired number of features is selected or performance starts to degrade.

D. Cross-Validation: Throughout the feature elimination process, cross-validation is used to assess the impact of feature removal on model accuracy. The optimal subset of features is selected based on cross-validation results.

3.3.2 Feature Importance Analysis

To gain insight into which features are most critical for predicting asset health, feature importance scores are calculated based on the Random Forest algorithm. Random Forest computes feature importance by measuring the decrease in Gini impurity caused by each feature in the model's decision trees. Features with a higher importance score have a more significant impact on the model's predictive ability.

3.4 Model Development

After selecting the optimal features, the final Random Forest model was developed. Random Forest was chosen due to its ability to handle large datasets and its robustness in dealing with noise and feature redundancy. The hyperparameters of the Random Forest model, such as the number of trees and maximum depth, were optimized using grid search with cross-validation.

The final model was trained using the selected features, and its performance was evaluated based on key metrics such as accuracy, precision, recall, and F1-score.

3.5 Model Evaluation

The performance of the predictive model was evaluated using the following metrics:

A. Accuracy: The proportion of correct predictions made by the model out of all predictions.

B. Precision: The ability of the model to correctly identify positive instances.

C. Recall: The ability of the model to capture all positive instances.

D. F1-Score: The harmonic mean of precision and recall, providing a balanced measure of model performance.

By using these various evaluation methods, the performance of the model can be measured thoroughly, both in terms of the predictions generated and a deeper understanding of the features that contribute to the predictions

4 Results and Discussion

This section presents the results of the predictive model developed for asset health management in power plants and discusses the significance of feature selection in improving model performance and interpretability. The focus is on the classification accuracy and the importance of selected features after applying Recursive Feature Elimination (RFE).

4.1 Model Performance

The Random Forest classifier achieved an overall accuracy of 93.64%, demonstrating strong predictive capabilities for asset health management. The classification report, as shown in Figure 3, provides detailed performance metrics for each class of asset health (1 is green, 2 is yellow, and 3 is red).

```
Accuracy: 0.9364615384615385

Classification Report:
              precision    recall  f1-score   support

           1       0.96      0.97      0.96      5320
           2       0.83      0.83      0.83      1089
           3       0.80      0.52      0.63        91

    accuracy                           0.94      6500
   macro avg       0.86      0.77      0.81      6500
weighted avg       0.94      0.94      0.94      6500
```

Figure 3 Classification report

Figure 4 shows the accuracy of the Random Forest model when using features selected by the Recursive Feature Elimination (RFE) method compared to the Random Forest model without RFE. The model with RFE has a higher accuracy, which is around 93.64%, compared to the model without RFE with an accuracy close to 93.50%.

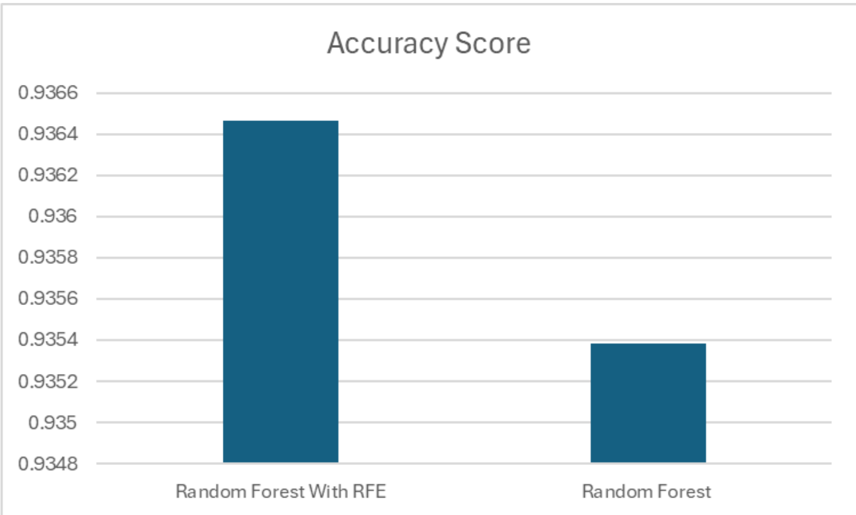

Figure 4 Comparison RFE implementation in Random Forest

Figure 5 shows the performance of the various models: Random Forest, XGBoost, LightGBM, and Neural Network. Random Forest emerges as the model with the highest accuracy, at around 93.6%, followed by XGBoost with an accuracy of around 92.5% and LightGBM at around 91.8%. Neural Network recorded the lowest performance with an accuracy of about 88.9%.

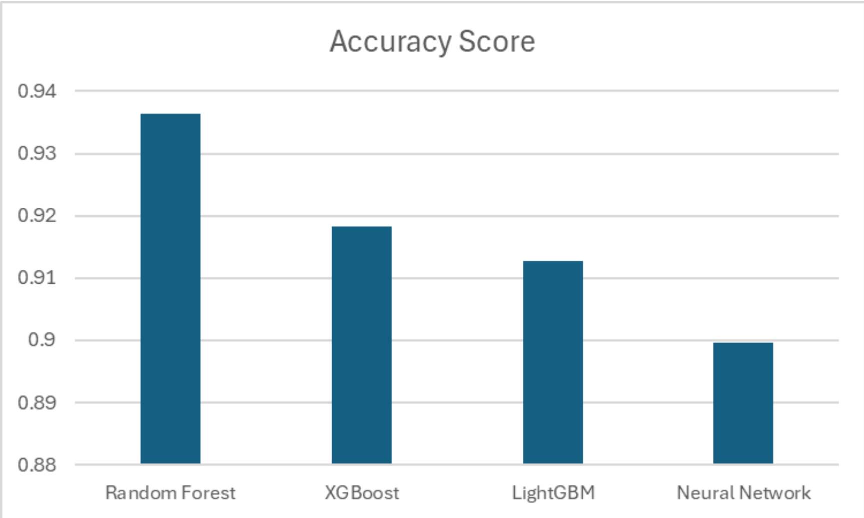

Figure 5 Comparison accuracy score between multiple algorithm

4.2 Feature Importance Analysis

The Recursive Feature Elimination (RFE) method was applied to identify the most important features influencing the model's predictions. Four features were selected, with their respective importance values as shown in Table 1.

Table 1 Fitur importance analysis

No	Feature	Importance score
1	MPI	0.34183961891228254
2	ACR	0.28896850314364525
3	ACRRANK	0.19782556140185004
4	SCR	0.1713663165422222

As seen from the importance scores, MPI contributes the most to the model's predictive capability, indicating that assets with a higher maintenance priority are more likely to experience issues that require intervention. This aligns with realworld maintenance practices where higher MPI values signify critical assets in need of urgent attention.

ACR and ACRRANK also significantly influence the model, representing the overall and ranking-based criticality of the assets. These two features help differentiate between assets that are crucial for plant operations and those with less immediate risk.

SCR, though the least important among the selected features, still plays a role in understanding system-level criticality, especially in the context of interconnected assets in power plants.

5 Conclusion

This research presents a robust framework for predictive asset health management in power plants, employing the Random Forest algorithm to enhance decision-making processes related to maintenance practices. The primary objective of the study was to identify and evaluate key features that significantly influence asset health predictions. Through the application of Recursive Feature Elimination (RFE), a large dataset was effectively distilled into four critical features: Maintenance Priority Index (MPI), Asset Criticality Ranking (ACR), ACRRANK, and System Criticality Ranking (SCR).

The Random Forest model achieved an accuracy of 93.64%, with commendable performance metrics for precision, recall, and F1 scores, indicating the robustness and reliability of the model in classifying asset health states. The highest feature importance was assigned to MPI, underscoring its role as a key determinant in maintenance prioritization. This finding aligns with existing literature that emphasizes the necessity of integrating predictive analytics into asset management frameworks to facilitate proactive maintenance strategies.

The results of this study not only confirm the utility of machine learning techniques in predictive maintenance but also highlight the significance of feature selection methods in enhancing model interpretability and performance. Utilizing RFE demonstrated that a reduced feature set can maintain high predictive accuracy while providing actionable insights to maintenance teams. This reduction is critical in operational settings where data complexity can hinder timely decision-making.

Furthermore, the implications of this research extend beyond predictive accuracy to encompass optimization of maintenance schedules and resource allocation. By focusing on assets identified as high priority by the selected features, organizations can mitigate the risk of unexpected failures, thereby improving overall operational efficiency.

References

[1] S. M. Hashemi, R. M. Botez, and G. Ghazi. (2024). *Robust Trajectory Prediction Using Random Forest Methodology Application to UAS-S4 Ehécatl*. Aerospace, vol. 11, no. 1, p. 49.

[2] M. Anisetti, C. A. Ardagna, A. Balestrucci, N. Bena, E. Damiani, and C. Y. Yeun. (2023). *On the Robustness of Random Forest Against Untargeted Data Poisoning: An Ensemble-Based Approach*. IEEE Trans. Sustain. Comput., vol. 8, no. 4, pp. 540–554.

[3] Y. Zhang, Z. Wang, and J. Yuan. (2021). *A Proximity Forest for Multivariate Time Series Classification*. pp. 766–778.

[4] C. Browne et al. (2021). *Multivariate random forest prediction of poverty and malnutrition prevalence*. PLoS One, vol. 16, no. 9, p. e0255519.

[5] V. Srinivasan, S. Hariprasath, and G. Thangavel. (2024). *Recursive feature elimination and multisupport vector machine in healthcare analytics*. Deep Learning Applications in Translational Bioinformatics, Elsevier, pp. 17–32.

[6] A. M. Priyatno, W. F. Ramadhan Sudirman, and R. J. Musridho. (2024). *Feature selection using non-parametric correlations and important features on recursive feature elimination for stock price prediction*. Int. J. Electr. Comput. Eng., vol. 14, no. 2, p. 1906.

[7] N. S. Yadav, V. P. Sharma, D. S. D. Reddy, and S. Mishra. (2023). *An Effective Network Intrusion Detection System Using Recursive Feature Elimination Technique*. RAiSE-2023, Basel Switzerland: MDPI, p. 99.

[8] F. O. Albasheer Mohamed and M. Agarwal. (2024). *Using Recursive Feature Elimination Feature Selection based Machine Learning Classifier for Attack Classification on UNSW-NB 15 dataset.* IEEE 9th International Conference for Convergence in Technology (I2CT), IEEE, pp. 1–7.

[9] Jardine, A. K. S., Lin, D., & Banjevic, D. (2006). *A review on machinery diagnostics and prognostics implementing condition-based maintenance*. Mechanical Systems and Signal Processing, 20(7), 1483–1510.

[10] Lee, J., & Ni, J. (2013). *Infotronics-based intelligent maintenance system and its impact on closed-loop maintenance operations in predictive manufacturing*. Journal of Intelligent Manufacturing, 24(1), 1–11.

[11] Breiman, L. (2001). *Random Forests*. Machine Learning, 45(1), 5–32.

[12] Singh, R., & Gupta, A. (2019). *Predictive maintenance techniques in industry 4.0*: *A review*. Journal of King Saud University - Computer and Information Sciences.

[13] Zhao, R., Yan, R., Chen, Z., Mao, K., Wang, P., & Gao, R. X. (2017). *Deep learning and its applications to machine health monitoring*. Mechanical Systems and Signal Processing, 115, 213–237.

[14] Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). *Gene selection for cancer classification using support vector machines.* Machine Learning, 46(1-3), 389–422.

[15] Liu, K., Wang, S., & Yao, X. (2018). *Feature selection using hybrid ReliefF and recursive feature elimination for fault diagnosis*. Journal of Manufacturing Systems, 47, 35–45.

[16] Wang, C., Wang, J., & Xu, H. (2020). *A comprehensive review of machine learning applied to predictive maintenance for wind turbines*. Applied Sciences, 10(21), 7820.

[17] Zhang, Y., & Chen, M. (2019). *Transformer fault diagnosis using Random Forest algorithm based on feature selection*. Energy Reports, 5, 775–783. –783.