

Implementasi Feature Engineering dan Random Forest dalam Klasifikasi Kegagalan Mesin Menggunakan Dataset AI4I 2020

Bryan Mandagi

INFORMASI ARTIKEL	ABSTRAK
Sejarah Artikel: Diterima: Agustus 2026 Revisi: Agustus 2026 Diterima: Agustus 2026 Dipublikasi: Agustus 2026 Kata Kunci: Random Forest; Predictive Maintenance; Klasifikasi; Kegagalan Mesin; AI4I 2020 *Penulis Korespondensi: bryanmandagi24@gmail.com	Penelitian ini bertujuan untuk mengklasifikasikan kondisi kegagalan mesin pada sistem <i>predictive maintenance</i> menggunakan metode Random Forest dengan memanfaatkan AI4I 2020 Predictive Maintenance Dataset. Data diperoleh dari <i>UCI Machine Learning Repository</i> dan melalui tahapan <i>preprocessing</i> berupa penghapusan atribut yang tidak relevan, penanganan <i>data leakage</i>, serta <i>one-hot encoding</i> pada data kategorikal. Selanjutnya dilakukan <i>feature engineering</i> dengan membentuk fitur baru, yaitu <i>Power_W</i>, <i>Temp_diff_K</i>, dan <i>Wear_Torque</i>, untuk meningkatkan representasi karakteristik mesin. Model Random Forest kemudian dilatih menggunakan data yang telah distandardisasi dan dievaluasi menggunakan Confusion Matrix, ROC Curve, Histogram Distribusi Fitur, Scatter Plot, serta metrik Accuracy, Precision, Recall, dan F1-Score. Hasil penelitian menunjukkan bahwa model memperoleh akurasi sebesar 99,08%, precision sebesar 93,06%, recall sebesar 78,82%, F1-score sebesar 85,35%, serta AUC sebesar 0,982. Hasil tersebut menunjukkan bahwa metode Random Forest mampu mengklasifikasikan kondisi mesin dengan tingkat akurasi yang tinggi sehingga berpotensi diterapkan sebagai pendukung sistem <i>predictive maintenance</i> pada lingkungan otomasi industri. ABSTRACT <i>This study aims to classify machine failure conditions in a predictive maintenance system using the Random Forest algorithm based on the AI4I 2020 Predictive Maintenance Dataset. The dataset was obtained from the UCI Machine Learning Repository and processed through several preprocessing stages, including the removal of irrelevant attributes, elimination of data leakage, and one-hot encoding for categorical variables. Furthermore, feature engineering was performed by generating three additional features, namely Power_W, Temp_diff_K, and Wear_Torque, to improve machine condition representation. The Random Forest model was trained using standardized data and evaluated through Confusion Matrix, ROC Curve, Feature Distribution Histogram, Scatter Plot, and performance metrics including Accuracy, Precision, Recall, and F1-Score. The experimental results achieved an accuracy of 99.08%, precision of 93.06%, recall of 78.82%, F1-score of 85.35%, and an AUC value of 0.982. These findings demonstrate that the Random Forest algorithm provides excellent classification performance and has strong potential for supporting predictive maintenance applications in industrial automation systems.</i>

PENDAHULUAN

Perkembangan teknologi pada era Industri 4.0 mendorong penerapan kecerdasan buatan (*Artificial Intelligence*) dalam sistem otomasi industri, salah satunya melalui *predictive maintenance*. *Predictive maintenance* memanfaatkan data operasional mesin untuk

memprediksi potensi kegagalan sehingga proses perawatan dapat dilakukan sebelum terjadi kerusakan. Pendekatan ini mampu mengurangi *downtime*, meningkatkan efisiensi produksi, dan menekan biaya pemeliharaan.

Salah satu algoritma klasifikasi yang banyak digunakan dalam predictive maintenance adalah Random Forest karena memiliki akurasi yang tinggi dan mampu menangani data berdimensi besar. Penelitian ini menggunakan AI4I 2020 Predictive Maintenance Dataset untuk membangun model klasifikasi kondisi mesin. Sebelum proses klasifikasi dilakukan, data melalui tahapan *preprocessing* dan *feature engineering* untuk meningkatkan kualitas data. Selanjutnya model dievaluasi menggunakan Confusion Matrix, ROC Curve, Histogram, Scatter Plot, serta metrik Accuracy, Precision, Recall, dan F1-Score. Penelitian ini bertujuan menganalisis performa algoritma Random Forest dalam mengklasifikasikan kondisi kegagalan mesin pada sistem predictive maintenance.

METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif dengan memanfaatkan AI4I 2020 Predictive Maintenance Dataset sebagai sumber data. Tahapan penelitian dimulai dengan *preprocessing* data melalui penghapusan atribut yang tidak relevan, penanganan *data leakage*, dan *one-hot encoding* pada data kategorikal. Selanjutnya dilakukan *feature engineering* dengan membentuk fitur Power_W, Temp_diff_K, dan Wear_Torque untuk meningkatkan representasi karakteristik mesin.

Data kemudian dibagi menjadi data latih (75%) dan data uji (25%), dilanjutkan dengan proses standardisasi menggunakan StandardScaler. Model klasifikasi dibangun menggunakan algoritma Random Forest, kemudian dievaluasi menggunakan Confusion Matrix, ROC Curve, Histogram Distribusi Fitur, Scatter Plot, Feature Importance, serta metrik Accuracy, Precision, Recall, dan F1-Score untuk mengukur kinerja model.

HASIL DAN PEMBAHASAN

Penelitian ini menggunakan algoritma Random Forest untuk mengklasifikasikan kondisi mesin berdasarkan AI4I 2020 Predictive Maintenance Dataset. Sebelum proses klasifikasi dilakukan, data melalui tahap *preprocessing* berupa penghapusan atribut yang tidak relevan, penanganan *data leakage*, *one-hot encoding*, serta pembentukan fitur baru melalui *feature engineering*. Setelah model dilatih menggunakan data latih, dilakukan pengujian terhadap data uji untuk mengetahui performa model.

1. Hasil Evaluasi Model

Berdasarkan hasil pengujian, model Random Forest menghasilkan performa klasifikasi yang sangat baik. Nilai evaluasi yang diperoleh ditunjukkan pada Tabel 1.

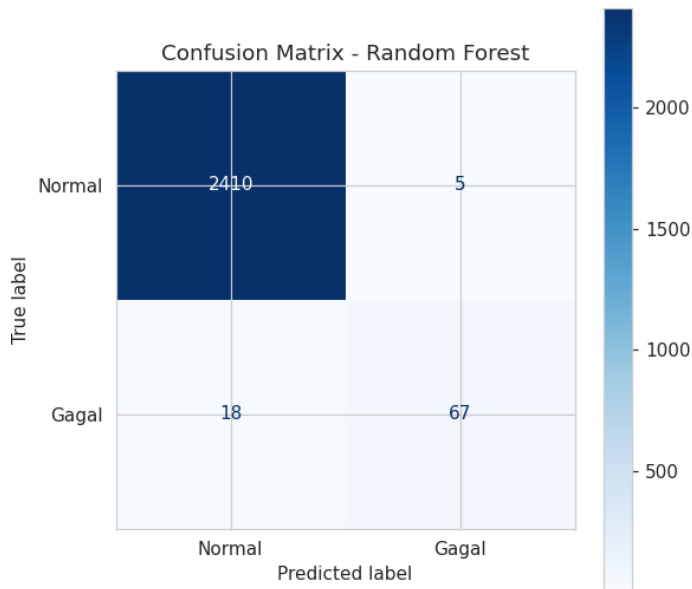
Table 1. Hasil Evaluasi Model Random Forest

Parameter	Nilai
Accuracy	99,08%
Precision	93,06%
Recall	78,82%
F1-Score	85,35%
AUC	0,982

Nilai akurasi sebesar 99,08% menunjukkan bahwa model mampu mengklasifikasikan kondisi mesin dengan tingkat ketepatan yang sangat tinggi. Selain itu, nilai precision sebesar 93,06%

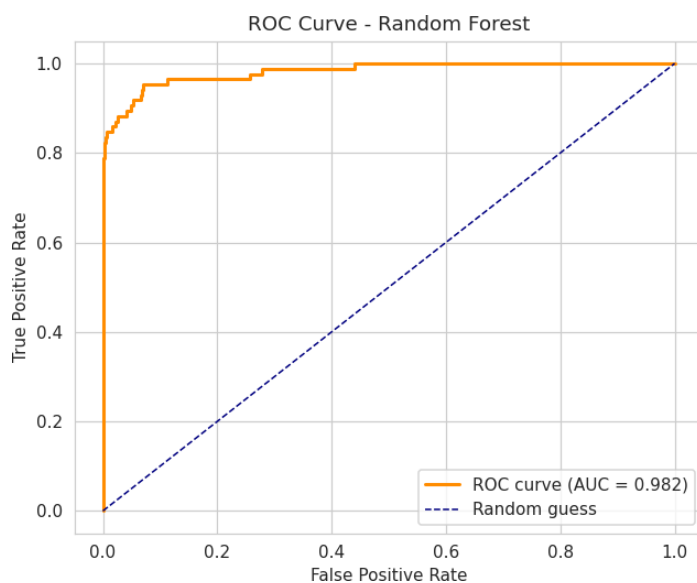
menunjukkan bahwa sebagian besar data yang diprediksi sebagai kegagalan mesin benar-benar merupakan kondisi gagal. Nilai recall sebesar 78,82% mengindikasikan bahwa model masih berhasil mendeteksi sebagian besar kasus kegagalan mesin, sedangkan nilai F1-score sebesar 85,35% menunjukkan keseimbangan yang baik antara precision dan recall.

2. Analisis Confusion Matrix



Berdasarkan Confusion Matrix, model berhasil mengklasifikasikan 2410 data normal dengan benar (*True Negative*) dan 67 data gagal dengan benar (*True Positive*). Sementara itu terdapat 5 data normal yang diprediksi sebagai gagal (*False Positive*) serta 18 data gagal yang diprediksi sebagai normal (*False Negative*). Hasil tersebut menunjukkan bahwa model memiliki tingkat kesalahan klasifikasi yang rendah sehingga mampu membedakan kondisi mesin normal dan gagal dengan baik

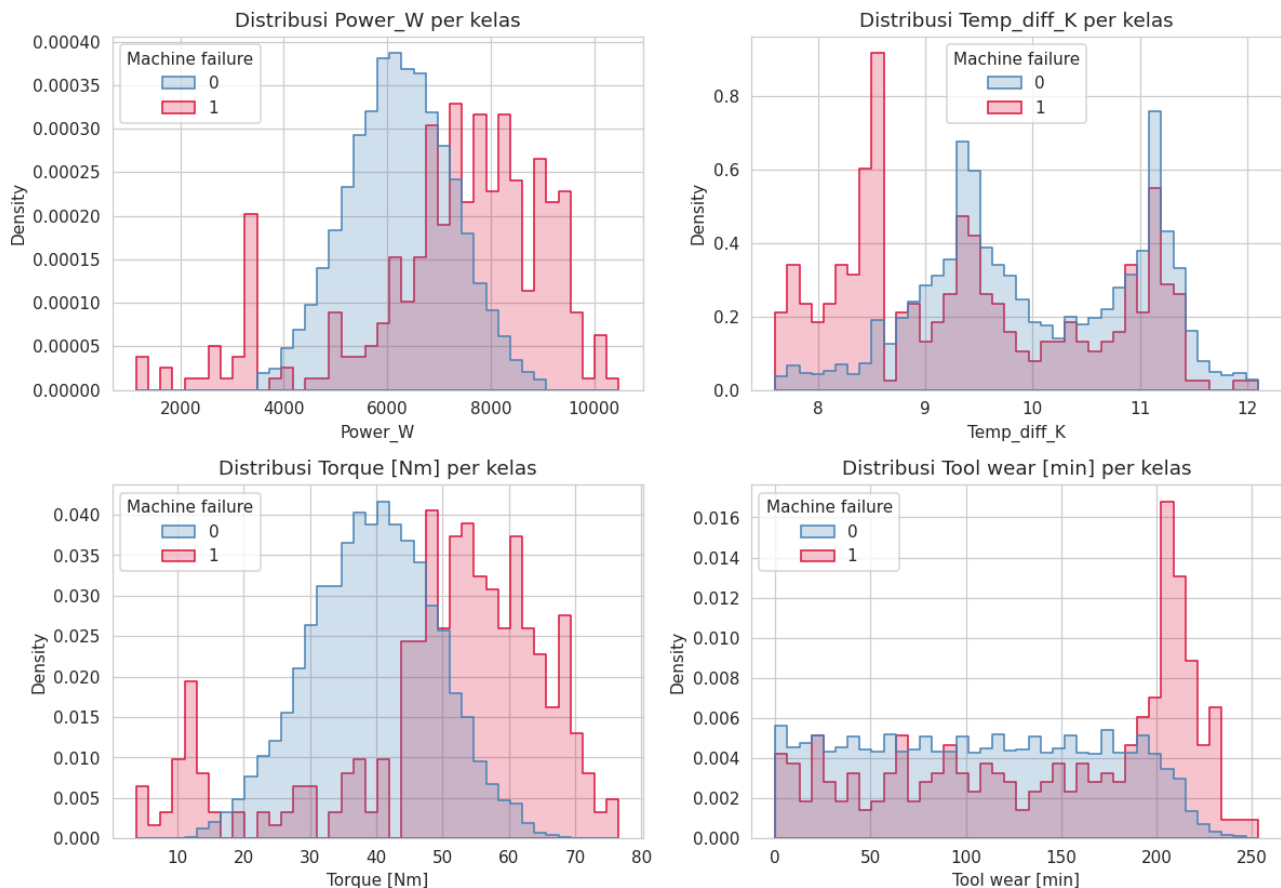
3. Analisis ROC Curve



Kurva ROC menunjukkan hubungan antara True Positive Rate dan False Positive Rate pada berbagai nilai ambang klasifikasi. Berdasarkan hasil pengujian diperoleh nilai Area Under Curve (AUC) sebesar 0,982. Nilai tersebut termasuk dalam kategori excellent classification, sehingga menunjukkan bahwa Random Forest memiliki kemampuan yang sangat baik dalam membedakan mesin yang mengalami kegagalan dan mesin yang berada pada kondisi normal.

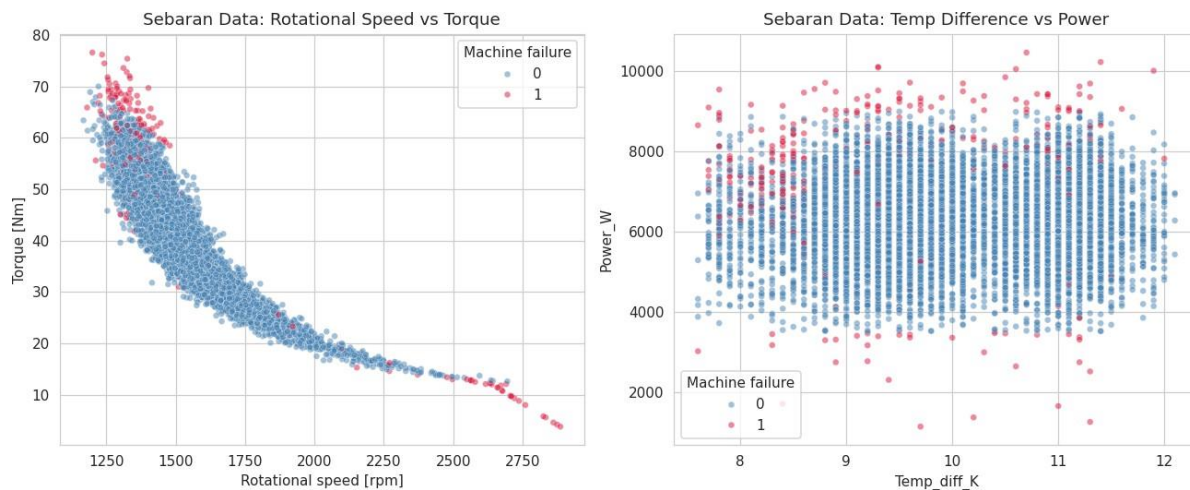
4. Analisis Histogram Distribusi Fitur

Histogram Distribusi Fitur (Normal vs Gagal)



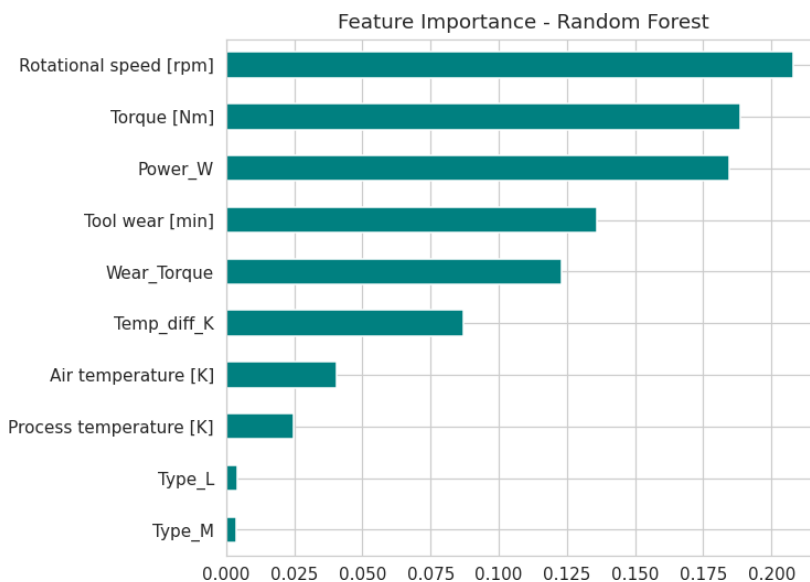
Histogram distribusi fitur menunjukkan perbedaan karakteristik antara data mesin normal dan data mesin gagal. Pada fitur Power_W, Torque, dan Tool Wear, terlihat bahwa data kegagalan mesin cenderung berada pada nilai yang lebih tinggi dibandingkan data normal. Sementara itu, pada fitur Temp_diff_K masih terdapat sedikit tumpang tindih antara kedua kelas, namun distribusi keseluruhan tetap menunjukkan pola yang berbeda. Hal ini mengindikasikan bahwa fitur hasil *feature engineering* mampu membantu model dalam membedakan kondisi mesin.

5. Analisis Sebaran Data (Scatter Plot)



Scatter plot pertama memperlihatkan hubungan antara Rotational Speed dan Torque, dimana data kegagalan mesin cenderung berada pada daerah dengan kombinasi torsi tinggi dan kecepatan putar yang lebih rendah. Scatter plot kedua menunjukkan hubungan antara Temperature Difference dan Power, yang memperlihatkan bahwa sebagian besar data kegagalan memiliki nilai daya yang relatif tinggi. Sebaran tersebut menunjukkan adanya pola yang dapat dipelajari oleh algoritma Random Forest untuk melakukan klasifikasi.

6. Analisis Feature Importance



Hasil analisis *feature importance* menunjukkan bahwa Rotational Speed [rpm] merupakan fitur yang memiliki pengaruh terbesar dalam proses klasifikasi, diikuti oleh Torque [Nm], Power_W, Tool Wear, dan Wear_Torque. Sementara itu, fitur Type_L dan Type_M memiliki kontribusi yang relatif kecil terhadap keputusan model. Hasil ini menunjukkan bahwa parameter operasional mesin memiliki pengaruh yang lebih besar dibandingkan jenis produk dalam menentukan kondisi kegagalan mesin.

Secara keseluruhan, hasil penelitian menunjukkan bahwa algoritma Random Forest mampu memberikan performa klasifikasi yang sangat baik pada AI4I 2020 Predictive Maintenance Dataset. Kombinasi proses *preprocessing* dan *feature engineering* berhasil meningkatkan kualitas data sehingga model mampu menghasilkan tingkat akurasi yang tinggi dan kesalahan klasifikasi yang rendah.

KESIMPULAN

Berdasarkan hasil penelitian, dapat disimpulkan bahwa algoritma Random Forest mampu mengklasifikasikan kondisi kegagalan mesin pada AI4I 2020 Predictive Maintenance Dataset dengan performa yang sangat baik. Model yang dibangun melalui tahapan *preprocessing*, *feature engineering*, dan standardisasi data menghasilkan nilai akurasi sebesar 99,08%, precision sebesar 93,06%, recall sebesar 78,82%, F1-score sebesar 85,35%, serta AUC sebesar 0,982.

Hasil analisis menggunakan Confusion Matrix, ROC Curve, Histogram, Scatter Plot, dan Feature Importance menunjukkan bahwa model mampu membedakan kondisi mesin normal dan gagal secara efektif. Fitur Rotational Speed, Torque, dan Power_W merupakan faktor yang paling berpengaruh dalam proses klasifikasi. Dengan demikian, algoritma Random Forest berpotensi diterapkan sebagai metode pendukung sistem predictive maintenance pada lingkungan otomasi industri untuk membantu proses pemantauan kondisi mesin dan mengurangi risiko terjadinya kegagalan operasional.

Daftar Pustaka

- [1] A. Matzka, "Explainable Artificial Intelligence for Predictive Maintenance Applications: A Survey," *IEEE Access*, vol. 9, pp. 118784–118806, 2021.
- [2] D. Chicco and G. Jurman, "The Advantages of the Matthews Correlation Coefficient over F1 Score and Accuracy in Binary Classification Evaluation," *BMC Genomics*, vol. 21, no. 6, pp. 1–13, 2020.
- [3] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [4] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [5] UCI Machine Learning Repository, "AI4I 2020 Predictive Maintenance Dataset," University of California, Irvine, 2020.
- [6] S. K. S. Fan, Y. C. Hsu, and C. C. Tsai, "Machine Learning Techniques for Predictive Maintenance in Smart Manufacturing," *IEEE Access*, vol. 10, pp. 55733–55747, 2022.
- [7] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. Cambridge, MA, USA: MIT Press, 2012.
- [8] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
- [9] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. New York, NY, USA: Springer, 2009.

- [10] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Burlington, MA, USA: Morgan Kaufmann, 2012.
- [11] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed. Pearson, 2021.
- [12] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.
- [13] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, 2nd ed. New York, NY, USA: Springer, 2021.
- [14] T. M. Mitchell, *Machine Learning*. New York, NY, USA: McGraw-Hill, 1997.
- [15] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, “A Survey on Concept Drift Adaptation,” *ACM Computing Surveys*, vol. 46, no. 4, pp. 1–37, 2014.