JMI Jurnal Multidisiplin Indonesia

Volume 5 Number 4, 2026
E-ISSN: 2964-9048 | P-ISSN: 2963-2900
https://jmi.rivierapublishing.id/index.php/rp

Klasifikasi Tingkat Ketidakhadiran Pegawai Menggunakan Random Forest, K-Nearest Neighbor, dan Logistic Regression

Julius Warih Angkasa*, Wahyu Ananda Konik Purbaningrum
Universitas BPD, Indonesia
Email: julius.wa@ubpd.ac.id*, w.ananda.k@ubpd.ac.id

Abstract

Employee absenteeism affects organizational productivity, service quality, and the efficiency of human resource management. Attendance data are often used only for administrative reporting even though they can support predictive analytics. This study develops a machine learning-based classification model for employee absenteeism level using the Absenteeism at Work dataset. The target variable is transformed into two classes, with absenteeism greater than 8 hours treated as the positive class. The workflow includes data cleaning, label transformation, categorical encoding, feature scaling for selected models, class balancing using SMOTE, and model development using Random Forest, K-Nearest Neighbor, and Logistic Regression. Evaluation is conducted using stratified 10-fold cross-validation with accuracy, precision, recall, F1-score, and AUC. The results show that Random Forest achieves the highest accuracy of 89.9%, while Logistic Regression provides the most balanced performance with 28.8% precision, 65.5% recall, 39.4% F1-score, and 81.4% AUC. These findings indicate that model selection for employee absenteeism should not rely on accuracy alone, especially when the positive class is relatively small

Kata kunci: machine learning; ketidakhadiran pegawai; random forest; klasifikasi; manajemen sumber daya manusia

Abstrak

Ketidakhadiran pegawai merupakan salah satu faktor yang memengaruhi produktivitas organisasi, kualitas layanan, dan efisiensi pengelolaan sumber daya manusia. Data absensi yang tersedia pada banyak organisasi umumnya masih dimanfaatkan untuk pelaporan administratif, padahal data tersebut berpotensi digunakan untuk analisis prediktif. Penelitian ini bertujuan membangun model klasifikasi tingkat ketidakhadiran pegawai menggunakan pendekatan machine learning. Dataset penelitian menggunakan data Absenteeism at Work yang ditransformasi menjadi dua kelas, yaitu ketidakhadiran rendah dan ketidakhadiran tinggi dengan ambang lebih dari 8 jam sebagai kelas positif. Tahapan penelitian meliputi pembersihan data, transformasi label, encoding variabel kategorikal, standardisasi data pada model tertentu, penyeimbangan kelas menggunakan SMOTE, serta pengujian model menggunakan Random Forest, K-Nearest Neighbor, dan Logistic Regression. Evaluasi dilakukan menggunakan stratified 10-fold cross-validation dengan metrik accuracy, precision, recall, F1-score, dan AUC. Hasil pengujian menunjukkan bahwa Random Forest memperoleh accuracy tertinggi sebesar 89,9%, sedangkan Logistic Regression memberikan performa paling seimbang dengan precision 28,8%, recall 65,5%, F1-score 39,4%, dan AUC 81,4%. Temuan ini menunjukkan bahwa pemilihan model pada data ketidakhadiran pegawai tidak cukup didasarkan pada accuracy saja, tetapi perlu mempertimbangkan kemampuan model dalam mendeteksi kelas minoritas

Keywords : *absenteeism; classification; human resource management; machine learning; random forest*

PENDAHULUAN

Transformasi digital dalam pengelolaan sumber daya manusia mendorong organisasi untuk tidak hanya menyimpan data kepegawaian, tetapi juga memanfaatkannya sebagai dasar pengambilan keputusan. Salah satu data yang paling sering tersedia adalah data absensi pegawai (Chawla et al., 2002; Pedregosa & al., 2011).

Data ini umumnya berisi informasi mengenai waktu kehadiran, alasan ketidakhadiran, karakteristik pegawai, serta indikator lain yang berpotensi menjelaskan pola absensi (Kohavi, 1995; Nath et al., 2022; Repository, n.d.).

Ketidakhadiran pegawai yang tinggi dapat menimbulkan berbagai dampak, seperti menurunnya efektivitas kerja, bertambahnya beban kerja pegawai lain, menurunnya kualitas layanan, serta meningkatnya biaya operasional organisasi (Blázquez & al., 2025; Zupančič et al., 2024). Meskipun demikian, pada banyak instansi, analisis terhadap data absensi masih bersifat deskriptif dan belum diarahkan pada analisis prediktif untuk mendeteksi risiko ketidakhadiran secara lebih dini (Nawata, 2024; Rajath & Pandita, 2022; Skorikov & al., 2020; Sumeisey et al., 2025).

Machine learning memberikan peluang untuk mengolah data absensi menjadi model prediksi yang mampu mengklasifikasikan tingkat risiko ketidakhadiran pegawai (Araujo et al., 2019; Han et al., 2005; He & Garcia, 2009; Lawrance et al., 2021). Dengan model klasifikasi yang baik, manajemen dapat melakukan intervensi lebih cepat, misalnya melalui monitoring khusus, penyesuaian jadwal kerja, atau evaluasi faktor-faktor penyebab ketidakhadiran (Davis & Goadrich, 2006; Saito & Rehmsmeier, 2015).

Urgensi penelitian ini didasarkan pada kebutuhan organisasi untuk beralih dari pendekatan reaktif ke proaktif dalam manajemen ketidakhadiran. Dengan model klasifikasi yang baik, manajemen dapat melakukan intervensi lebih cepat, seperti monitoring khusus, penyesuaian jadwal kerja, atau evaluasi faktor-faktor penyebab ketidakhadiran (Lawrance et al., 2021). Tanpa model prediktif, organisasi akan terus mengalami kerugian operasional akibat ketidakhadiran yang tidak terdeteksi sejak dini.

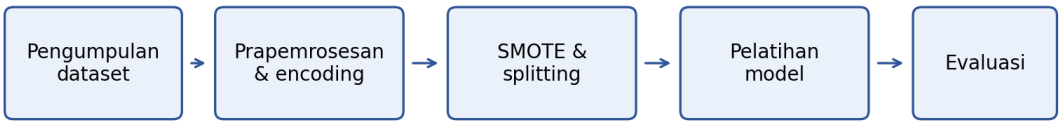
Beberapa penelitian sebelumnya menunjukkan bahwa pendekatan machine learning dapat diterapkan untuk memprediksi absenteeism, baik untuk kebutuhan analisis klasifikasi maupun dukungan pengambilan keputusan berbasis sistem administrasi (Breiman, 2001; Hosmer Jr. et al., 2013; Lemaître et al., 2017). Namun, masih diperlukan naskah penelitian yang sederhana, replikatif, dan mudah diadaptasi untuk konteks organisasi di Indonesia.

Berdasarkan latar belakang tersebut, penelitian ini bertujuan membandingkan performa tiga algoritma machine learning, yaitu Random Forest, K-Nearest Neighbor, dan Logistic Regression, dalam mengklasifikasikan tingkat ketidakhadiran pegawai. Kontribusi penelitian ini terletak pada penerapan model klasifikasi yang aplikatif, penggunaan penyeimbangan data untuk mengatasi ketidakseimbangan kelas, serta interpretasi fitur penting sebagai dasar rekomendasi manajerial.

METODE PENELITIAN

Desain Penelitian

Penelitian ini menggunakan pendekatan kuantitatif eksperimental (Repository, n.d.). Tahapan penelitian meliputi pengumpulan dataset, prapemrosesan, pembentukan label kelas, pembagian data latih dan data uji secara tervalidasi, pelatihan model, evaluasi performa, dan interpretasi hasil (Kohavi, 1995; Repository, n.d.). Model utama yang digunakan adalah *Random Forest*, sedangkan *K-Nearest Neighbor* dan *Logistic Regression* digunakan sebagai model pembanding (Breiman, 2001).



Gambar 1. Alur umum penelitian machine learning untuk klasifikasi tingkat ketidakhadiran pegawai

Dataset Penelitian

Dataset penelitian disarankan menggunakan dataset publik "Absenteeism at Work" dari UCI Machine Learning Repository atau dataset lokal organisasi yang memiliki struktur serupa. Dalam penelitian ini, variabel target dibentuk menjadi dua kelas, yaitu ketidakhadiran rendah dan ketidakhadiran tinggi. Salah satu skenario yang sering dipakai adalah menggunakan nilai median atau ambang kebijakan tertentu pada variabel absenteeism time in hours untuk membentuk label biner.

Contoh atribut yang dapat digunakan pada penelitian ini meliputi:

1. reason for absence
2. month of absence
3. day of the week
4. seasons
5. transportation expense
6. distance from residence to work
7. service time
8. age
9. workload average/day
10. hit target
11. disciplinary failure
12. education
13. son, pet, weight, height, dan body mass index

Prapemrosesan Data

Tahap prapemrosesan dilakukan untuk memastikan data dapat digunakan secara optimal oleh model machine learning. Pertama, data diperiksa untuk memastikan tidak terdapat duplikasi, inkonsistensi nilai, dan atribut yang tidak relevan. Kedua, variabel target ditransformasi menjadi label biner sesuai ambang yang telah ditentukan. Ketiga, atribut kategorikal diubah menjadi representasi numerik menggunakan label encoding

atau one-hot encoding. Keempat, standardisasi diterapkan khusus pada model yang sensitif terhadap skala data, terutama K-Nearest Neighbor dan Logistic Regression. Kelima, jika distribusi kelas tidak seimbang, diterapkan Synthetic Minority Over-sampling Technique (SMOTE) pada data latih untuk mengurangi bias model terhadap kelas mayoritas.

Tabel 1. Tahapan prapemrosesan data

No	Tahap	Deskripsi
1	Cleaning data	Memeriksa duplikasi, nilai anomali, dan atribut yang tidak relevan.
2	Transformasi label	Mengubah variabel target menjadi dua kelas: rendah dan tinggi.
3	Encoding	Mengubah atribut kategorikal menjadi numerik agar dapat diproses model.
4	Scaling	Melakukan standardisasi pada fitur numerik untuk KNN dan Logistic Regression.
5	Balancing	Menerapkan SMOTE pada data latih bila distribusi kelas tidak seimbang.

Pemodelan

Tiga algoritma yang digunakan dalam penelitian ini adalah Random Forest, K-Nearest Neighbor, dan Logistic Regression. Random Forest dipilih sebagai model utama karena mampu menangani hubungan nonlinier dan interaksi antarfitur melalui mekanisme ensemble dari banyak decision tree. K-Nearest Neighbor dipilih sebagai pembanding karena sederhana dan mudah diinterpretasikan berdasarkan kedekatan jarak antarinstance. Logistic Regression digunakan sebagai baseline model karena umum dipakai dalam klasifikasi biner dan memiliki interpretabilitas koefisien yang baik.

Rancangan parameter yang dapat digunakan pada penelitian ini disajikan pada Tabel 2.

Tabel 2. Rancangan parameter model

Model	Parameter utama	Rentang pengujian
Random Forest	n_estimators, max_depth, min_samples_split	100–300; 5–15; 2–10
K-Nearest Neighbor	n_neighbors, metric, weights	3–9; euclidean/manhattan; uniform/distance
Logistic Regression	C, solver, max_iter	0.1–10; lbfgs/liblinear; 100–500

Evaluasi Model

Evaluasi model dilakukan menggunakan stratified 10-fold cross-validation untuk memperoleh estimasi performa yang lebih stabil pada data klasifikasi. Metrik yang digunakan adalah accuracy, precision, recall, F1-score, dan Area Under the Curve (AUC). Accuracy digunakan untuk melihat proporsi prediksi benar secara umum, precision dan recall untuk menilai ketepatan klasifikasi pada kelas positif, F1-score untuk menyeimbangkan precision dan recall, sedangkan AUC dipakai untuk menilai kemampuan model membedakan dua kelas secara keseluruhan.

$$Accuracy = (TP + TN) / (TP + TN + FP + FN)$$
$$Precision = TP / (TP + FP)$$
$$Recall = TP / (TP + FN)$$
$$F1\text{-}score = 2 \times Precision \times Recall / (Precision + Recall)$$

HASIL DAN PEMBAHASAN

Hasil Prapemrosesan

Dataset yang digunakan terdiri atas 740 record. Variabel target dibentuk menjadi dua kelas dengan ketentuan ketidakhadiran lebih dari 8 jam sebagai kategori tinggi. Berdasarkan transformasi tersebut, diperoleh 677 record (91,49%) pada kelas ketidakhadiran rendah dan 63 record (8,51%) pada kelas ketidakhadiran tinggi. Distribusi ini menunjukkan adanya ketidakseimbangan kelas yang cukup besar, sehingga SMOTE diterapkan pada data latih di setiap fold untuk menyeimbangkan jumlah sampel kelas minoritas terhadap kelas mayoritas sebelum proses pelatihan model dilakukan.

Hasil Pengujian Model

Tabel 3 menyajikan ringkasan hasil eksperimen setelah proses prapemrosesan dan evaluasi model. Nilai yang ditampilkan merupakan rata-rata hasil stratified 10-fold cross-validation pada tiga algoritma yang dibandingkan. Hasil ini menunjukkan bahwa model yang unggul pada accuracy belum tentu menjadi model yang paling baik untuk mendeteksi kelas ketidakhadiran tinggi yang jumlahnya jauh lebih sedikit.

Tabel 3. Hasil Evaluasi model

Model	Accuracy	Precision	Recall	F1-score	AUC
Random Forest	0,899	0,315	0,174	0,215	0,806
K-Nearest Neighbor	0,730	0,183	0,619	0,281	0,695
Logistic Regression	0,822	0,288	0,655	0,394	0,814

Narasi hasil dapat dirumuskan sebagai berikut: Berdasarkan pengujian menggunakan stratified 10-fold cross-validation, Random Forest memperoleh accuracy tertinggi sebesar 89,9%. Namun, Logistic Regression menunjukkan performa yang lebih seimbang untuk mendeteksi kelas ketidakhadiran tinggi dengan precision 28,8%, recall 65,5%, F1-score 39,4%, dan AUC 81,4%. Sementara itu, K-Nearest Neighbor memiliki recall yang cukup tinggi, tetapi precision dan AUC-nya masih berada di bawah Logistic Regression.

Hasil pengujian menunjukkan bahwa Random Forest tidak menjadi model terbaik pada semua metrik, meskipun accuracy-nya paling tinggi. Temuan ini perlu dibaca secara hati-hati karena dataset memiliki distribusi kelas yang tidak seimbang. Dalam kondisi seperti ini, accuracy yang tinggi dapat muncul karena model lebih sering memprediksi kelas mayoritas, yaitu ketidakhadiran rendah. Hal tersebut terlihat pada Random Forest yang memperoleh accuracy 89,9%, tetapi recall untuk kelas ketidakhadiran tinggi hanya 17,4% dengan F1-score 21,5%. Dengan kata lain, model ini cukup baik dalam

mempertahankan prediksi benar secara umum, namun masih lemah dalam menangkap pegawai yang benar-benar berisiko mengalami ketidakhadiran tinggi.

K-Nearest Neighbor menunjukkan pola yang berbeda. Model ini memiliki recall 61,9%, artinya cukup banyak kasus ketidakhadiran tinggi berhasil terdeteksi. Akan tetapi, precision-nya hanya 18,3% dan AUC sebesar 69,5%, sehingga masih banyak prediksi positif yang ternyata tidak tepat. Kondisi ini mengindikasikan bahwa K-Nearest Neighbor relatif sensitif terhadap distribusi lokal data, pemilihan jumlah tetangga, serta karakteristik fitur yang berasal dari kombinasi variabel numerik dan kategorikal. Walaupun standardisasi telah diterapkan, model ini belum memberikan keseimbangan performa yang optimal.

Logistic Regression justru memberikan performa paling seimbang pada penelitian ini. Model ini memperoleh precision 28,8%, recall 65,5%, F1-score 39,4%, dan AUC 81,4%. Hasil tersebut menunjukkan bahwa meskipun Logistic Regression bersifat lebih sederhana dan linear dibanding Random Forest, model ini lebih efektif dalam membedakan pegawai dengan risiko ketidakhadiran tinggi pada konfigurasi data yang digunakan. Temuan ini juga menandakan bahwa hubungan antarfitur pada dataset tidak selalu menuntut model ensemble yang kompleks untuk menghasilkan prediksi yang berguna secara praktis.

Secara praktis, hasil penelitian ini menegaskan bahwa pemilihan model untuk data kepegawaian perlu mempertimbangkan tujuan analisis. Apabila organisasi hanya berfokus pada ketepatan klasifikasi umum, Random Forest dapat dipertimbangkan karena accuracy-nya paling tinggi. Namun, apabila tujuan utama adalah mendeteksi sedini mungkin pegawai dengan risiko ketidakhadiran tinggi, maka Logistic Regression lebih layak dipilih karena memberikan keseimbangan yang lebih baik antara precision, recall, F1-score, dan AUC. Dengan demikian, model prediktif yang dihasilkan dapat lebih relevan untuk mendukung monitoring, intervensi dini, dan kebijakan manajemen kepegawaian berbasis data.

KESIMPULAN

Penelitian ini membandingkan tiga algoritma machine learning, yaitu Random Forest, K-Nearest Neighbor, dan Logistic Regression, untuk mengklasifikasikan tingkat ketidakhadiran pegawai. Hasil pengujian menunjukkan bahwa Random Forest memperoleh accuracy tertinggi sebesar 89,9%, sedangkan Logistic Regression memberikan performa paling seimbang dengan precision 28,8%, recall 65,5%, F1-score 39,4%, dan AUC 81,4%. Temuan ini menunjukkan bahwa evaluasi model pada data yang tidak seimbang tidak cukup hanya berfokus pada accuracy, tetapi juga perlu memperhatikan kemampuan model dalam mendeteksi kelas minoritas.

Secara metodologis, penelitian ini memberikan kerangka kerja yang mudah direplikasi untuk pengembangan sistem pendukung keputusan di bidang manajemen kepegawaian. Secara praktis, Logistic Regression lebih direkomendasikan ketika organisasi memerlukan deteksi dini terhadap pegawai dengan risiko ketidakhadiran

tinggi, sedangkan Random Forest dapat dipilih apabila prioritas utama terletak pada ketepatan klasifikasi secara umum.

Penelitian selanjutnya dapat menambahkan optimasi hyperparameter yang lebih mendalam, seleksi fitur, explainable AI, atau pengujian pada dataset lokal dari instansi pendidikan maupun organisasi sektor publik agar model semakin kontekstual dan aplikatif.

REFERENSI

Araujo, V. S., Rezende, T. S., Guimarães, A. J., Araujo, V. J. S., & de Campos Souza, P. V. (2019). A Hybrid Approach of Intelligent Systems to Help Predict Absenteeism at Work in Companies. *SN Applied Sciences*, *1*, 536. https://doi.org/10.1007/s42452-019-0547-x

Blázquez, P. L., & al., et. (2025). *Predicting Workplace Absenteeism Using Machine Learning: A Pilot Study in Occupational Health*.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Chawla, N. V, Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Davis, J., & Goadrich, M. (2006). *The Relationship Between Precision-Recall and ROC Curves BT - Proceedings of the 23rd International Conference on Machine Learning*. 233–240. https://doi.org/10.1145/1143844.1143874

Han, H., Wang, W.-Y., & Mao, B.-H. (2005). *Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning BT - Advances in Intelligent Computing* (Vol. 3644, pp. 878–887).

He, H., & Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Hosmer Jr., D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression* (3rd ed.). Wiley.

Kohavi, R. (1995). *A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection BT - Proceedings of the 14th International Joint Conference on Artificial Intelligence*. 1137–1145.

Lawrance, N., Petrides, G., & Guerry, M.-A. (2021). Predicting Employee Absenteeism for Cost Effective Interventions. *Decision Support Systems*, *147*, 113539. https://doi.org/10.1016/j.dss.2021.113539

Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research*, *18*(17), 1–5.

Nath, G., Wang, Y., Coursey, A., Saha, K. K., Prabhu, S., & Sengupta, S. (2022). Incorporating a Machine Learning Model into a Web-Based Administrative Decision Support Tool for Predicting Workplace Absenteeism. *Information*, *13*(7), 320. https://doi.org/10.3390/info13070320

Nawata, K. (2024). Evaluation of Physical and Mental Health Conditions Related to Employees' Absenteeism. *Frontiers in Public Health*.

Pedregosa, F., & al., et. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Rajath, J., & Pandita, D. (2022). *Machine Learning as a Data Science Tool to Predict*

Absenteeism for Factory Workers.
Repository, U. C. I. M. L. (n.d.). *Absenteeism at Work*. University of California, Irvine. https://archive.ics.uci.edu/dataset/445/absenteeism+at+work
Saito, T., & Rehmsmeier, M. (2015). The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLoS ONE*, *10*(3), e0118432. https://doi.org/10.1371/journal.pone.0118432
Skorikov, S., & al., et. (2020). *Prediction of Absenteeism at Work using Data Mining Techniques*.
Sumeisey, N. L., Misriati, T., & Nawawi, I. (2025). Klasifikasi Ketidakhadiran di Tempat Kerja Menggunakan Metode Support Vector Machine. *Jurnal Komputer, Informasi Dan Teknologi*, *5*(1).
Zupančič, P., Lavrač, N., & al., et. (2024). Predicting Employee Absence from Historical Absence Profiles with Machine Learning. *Applied Sciences*, *14*(16), 7037. https://doi.org/10.3390/app14167037