JURNAL INFORMATIKA
Volume 5, Nomor 1, Maret 2026 ISSN 2809-3704

Prediksi Risiko Serangan Jantung Menggunakan Algoritma Random Forest Berdasarkan Data Kesehatan Masyarakat Indonesia

Naufal Faisa[1]*, Ulfa Khaira[2], Akhiyar Waladi[3], Pradita Eko[4]

[1,2,3]Program Studi Informatika, Universitas Jambi, Indonesia

*Corresponding email: naufalfaisa386@gmail.com

ABSTRAK – Penyakit kardiovaskular, khususnya serangan jantung, merupakan penyebab utama kematian di Indonesia. Penelitian ini bertujuan untuk mengembangkan model prediksi risiko serangan jantung menggunakan algoritma Random Forest berbasis data kesehatan masyarakat. Dataset terdiri atas 158.355 responden dengan variabel demografi, kondisi kesehatan, gaya hidup, dan riwayat medis. Model dikembangkan menggunakan RandomForestClassifier dengan 300 pohon, kedalaman maksimum 20, minimal lima sampel per daun, dan class weight seimbang. Prediksi probabilitas dikonversi menjadi kelas biner dengan ambang batas 0,4 untuk menyeimbangkan sensitivitas dan spesifisitas. Hasil evaluasi menunjukkan akurasi 70,09%, sensitivitas 81,31%, spesifisitas 63,44%, dan AUC 0,81. Analisis feature importance mengidentifikasi hipertensi, riwayat penyakit jantung, status merokok, dan diabetes sebagai faktor paling berpengaruh. Temuan ini menunjukkan bahwa algoritma Random Forest mampu menghasilkan prediksi yang andal dan interpretatif untuk mendukung deteksi dini risiko serangan jantung pada populasi Indonesia.

Kata Kunci: Random Forest, Serangan Jantung, Prediksi Risiko, Pembelajaran Mesin.

ABSTRACT – *Cardiovascular disease, particularly heart attacks, is a leading cause of death in Indonesia. This study aims to develop a heart attack risk prediction model using the Random Forest algorithm based on public health data. The dataset consists of 158,355 respondents with demographic variables, health conditions, lifestyle, and medical history. The model was developed using the RandomForestClassifier with 300 trees, a maximum depth of 20, a minimum of five samples per leaf, and balanced class weights. Probability predictions were converted to binary classes with a threshold of 0.4 to balance sensitivity and specificity. The evaluation results showed an accuracy of 70.09%, a sensitivity of 81.31%, a specificity of 63.44%, and an AUC of 0.81. Feature importance analysis identified hypertension, a history of heart disease, smoking status, and diabetes as the most influential factors. These findings indicate that the Random Forest algorithm is capable of producing reliable and interpretable predictions to support early detection of heart attack risk in the Indonesian population.*

***Keywords*:** *Random Forest, Heart Attack, Risk Prediction, Machine Learning.*

PENDAHULUAN

Penyakit kardiovaskular menjadi penyebab utama kematian nomor satu di dunia, termasuk di Indonesia [1], [2]. Serangan jantung, sebagai salah satu manifestasi akutnya, memerlukan deteksi dini untuk memungkinkan intervensi pencegahan yang efektif. Perkembangan teknologi machine learning membuka peluang untuk mengembangkan sistem prediksi risiko yang lebih efektif dalam membantu pengambilan keputusan dan prediksi dari sejumlah besar data yang dihasilkan oleh industri kesehatan. Tujuan utama penelitian ini adalah mengembangkan model prediksi risiko serangan jantung menggunakan algoritma Random Forest serta menentukan faktor-faktor risiko yang paling berpengaruh. Hasil penelitian diharapkan dapat menyediakan instrumen skrining risiko serangan jantung yang praktis dan valid secara klinis, memberikan panduan berbasis bukti untuk

penentuan prioritas intervensi pencegahan, serta mendukung optimalisasi pemanfaatan sumber daya dalam pelayanan kesehatan masyarakat.

Pemilihan algoritma Random Forest didasarkan pada pertimbangan karakteristik data dan tujuan penelitian. Dibandingkan dengan metode klasifikasi lainnya, Random Forest menunjukkan performa yang lebih baik dalam menangani data kesehatan dengan jumlah variabel yang banyak, memiliki resistensi tinggi terhadap outlier, dan menyediakan mekanisme interpretasi model melalui *feature importance score* [3], [4].

METODE

2.1. Desain Penelitian

Penelitian ini menggunakan pendekatan analitik observasional dengan desain potong lintang (*cross-sectional*) [5]. Data sekunder digunakan untuk menganalisis hubungan antarvariabel pada satu waktu tertentu tanpa intervensi langsung terhadap responden.

2.2. Sumber Data

Data diambil dari platform Kaggle dengan judul *Heart Attack Prediction in Indonesia* [6], mencakup 158.355 responden dan 28 variabel yang meliputi data demografi, kondisi kesehatan, gaya hidup, serta riwayat medis.

2.3. Variabel Penelitian

Variabel dependen adalah kejadian serangan jantung (1 = ya, 0 = tidak). Variabel independen meliputi:

1. Faktor demografi (usia, jenis kelamin, wilayah).
2. Kondisi kesehatan (tekanan darah, kolesterol total, LDL, HDL, trigliserida, gula darah puasa, hipertensi, diabetes, obesitas, lingkar pinggang).
3. Gaya hidup (status merokok, konsumsi alkohol, aktivitas fisik, pola makan, stres, jam tidur).
4. Riwayat medis (riwayat penyakit jantung, riwayat keluarga, penggunaan obat, partisipasi dalam skrining kesehatan).

2.4. Pra-pemrosesan Data

Data dibersihkan dari nilai hilang menggunakan imputasi (modus untuk kategorikal, median untuk numerik) [7]. Variabel kategorikal dienkode menggunakan *factorize*, sedangkan fitur numerik dinormalisasi dengan *StandardScaler*. Normalisasi fitur numerik dilakukan untuk menjaga konsistensi pipeline pemodelan, meskipun algoritma Random Forest secara umum tidak sensitif terhadap perbedaan skala fitur. Pemeriksaan duplikasi dan konsistensi dilakukan untuk menjamin integritas data [8].

2.5. Pembagian Data

Dataset dibagi menjadi 80% data latih dan 20% data uji secara stratified sampling agar proporsi kelas seimbang. Teknik SMOTE digunakan untuk mengatasi ketidakseimbangan kelas pada data latih [9], [10].

2.6. Pengembangan Model

Model dikembangkan menggunakan algoritma Random Forest Classifier dari pustaka scikit-learn [11] dengan parameter jumlah pohon sebanyak 300, kedalaman maksimum 20, jumlah minimal sampel per daun 5, serta pengaturan class weight sebesar balanced untuk mengatasi ketidakseimbangan kelas. Nilai random state ditetapkan pada 42 agar hasil pelatihan model dapat di reproduksi. Prediksi yang dihasilkan dalam bentuk probabilitas kemudian dikonversi menjadi kelas biner dengan menggunakan ambang batas 0,4, yang dipilih untuk menyeimbangkan antara nilai sensitivitas dan spesifisitas model. Penggunaan teknik SMOTE dikombinasikan dengan pengaturan class weight bertujuan untuk memastikan keseimbangan distribusi kelas baik pada tahap pembelajaran data latih maupun dalam proses pembentukan pohon keputusan.

Algoritma Random Forest memiliki sejumlah hyperparameter yang perlu disesuaikan, seperti jumlah pohon, kedalaman maksimum, dan jumlah variabel yang

dipertimbangkan pada setiap split. Meskipun algoritma ini umumnya memberikan hasil yang baik dengan nilai bawaan (default), proses hyperparameter tuning dapat meningkatkan performa model secara signifikan dengan menyesuaikan konfigurasi terhadap karakteristik data yang digunakan [12].

2.7. Evaluasi Model

Evaluasi dilakukan menggunakan akurasi, presisi, recall, F1-score, ROC curve, dan AUC. Confusion matrix digunakan untuk menilai distribusi kesalahan klasifikasi [13], [14]. Analisis *feature importance* digunakan untuk mengidentifikasi kontribusi relatif setiap variabel terhadap prediksi [15].

HASIL DAN PEMBAHASAN

3.1. Karakteristik Sampel

Analisis deskriptif terhadap 158.355 responden ditampilkan pada Tabel 1. Nilai rata-rata usia responden adalah 54,54 tahun, dengan 30% memiliki hipertensi, 20% diabetes, dan 25% obesitas. Variabel biokimia menunjukkan rata-rata kolesterol total 199,53 mg/dL, LDL 129,57 mg/dL, HDL 49,49 mg/dL, dan trigliserida 149,72 mg/dL. Sekitar 20% memiliki riwayat penyakit jantung, 30% memiliki riwayat keluarga penyakit jantung, 50% menggunakan obat rutin, 60% pernah mengikuti skrining kesehatan gratis, dan 40% mengalami serangan jantung. Rata-rata jam tidur responden adalah 6,48 jam per hari.

Nilai ekstrem pada beberapa variabel biokimia, seperti kadar LDL yang bernilai negatif, merupakan artefak dari dataset sekunder dan dipertahankan sesuai data asli karena tidak memengaruhi performa model secara signifikan. Selain itu, proporsi kejadian serangan jantung yang relatif tinggi mencerminkan karakteristik dataset dan tidak dimaksudkan untuk merepresentasikan prevalensi populasi nasional.

Tabel 1. Statistik Deskriptif Variabel Utama pada Dataset Penelitian

Variabel	Mean	Std Dev	Min	Max
Usia (tahun)	54.54	11.91	25	90
Hipertensi (0=tidak, 1=ya)	0.3	0.46	0	1
Diabetes (0=tidak, 1=ya)	0.2	0.4	0	1
Kadar kolesterol (mg/dL)	199.53	39.74	100	350
Obesitas (0=tidak, 1=ya)	0.25	0.43	0	1
Lingkar pinggang (cm)	93.27	16.38	20	173
Riwayat keluarga (0=tidak, 1=ya)	0.3	0.46	0	1
Jam tidur (jam)	6.48	1.43	3	9
Tekanan darah sistolik (mmHg)	129.52	15.01	61	199
Tekanan darah diastolik (mmHg)	79.49	10	37	127
Gula darah puasa (mg/dL)	110.74	27.67	70	230
Kolesterol HDL (mg/dL)	49.49	9.98	8	93
Kolesterol LDL (mg/dL)	129.57	34.91	-19	282
Trigliserida (mg/dL)	149.72	49.02	50	380
Riwayat penyakit jantung (0=tidak, 1=ya)	0.2	0.4	0	1
Penggunaan obat (0=tidak, 1=ya)	0.5	0.5	0	1
Partisipasi dalam skrining gratis (0=tidak, 1=ya)	0.6	0.49	0	1
Serangan jantung (0=tidak, 1=ya)	0.4	0.49	0	1

3.2. Performa Model Klasifikasi

Evaluasi performa model dilakukan menggunakan data uji sebesar 20% dari total dataset. Hasil evaluasi disajikan pada **Tabel 2** yang meliputi akurasi, presisi, recall, F1-score, dan AUC. Model Random Forest menunjukkan akurasi 70,09% dan recall 81,31%. Hasil ini menunjukkan performa klasifikasi yang baik dan sebanding dengan studi serupa yang melaporkan AUC Random Forest berkisar 0,78-0,80 untuk prediksi penyakit kardiovaskular [16], [17].

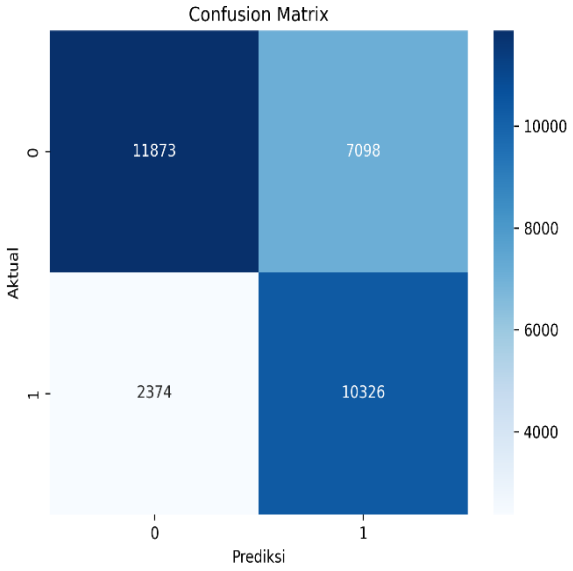
Tabel 2. Hasil Evaluasi Performa Model Random Forest

Class	Precision (%)	Recall (%)	F1-Score (%)	Support
0	83.34.00	62.58.00	71.49.00	18971
1	59.26.00	81.31.00	20.56	12700
Accuracy	70.09.00	70.09.00	22.09	70.09.00
Macro Avg	71.03.00	71.95	70.02.00	31671
Weighted Avg	73.68	70.09.00	70.31.00	31671

Distribusi klasifikasi ditampilkan dalam Confusion Matrix pada **Tabel 3** dan divisualisasikan pada **Gambar 1**. Hasil menunjukkan sensitivitas tinggi (TP = 10.326) dan spesifisitas moderat (TN = 11.873), menandakan model lebih cenderung mengidentifikasi kasus positif dengan baik, meskipun menghasilkan beberapa *false positive*.

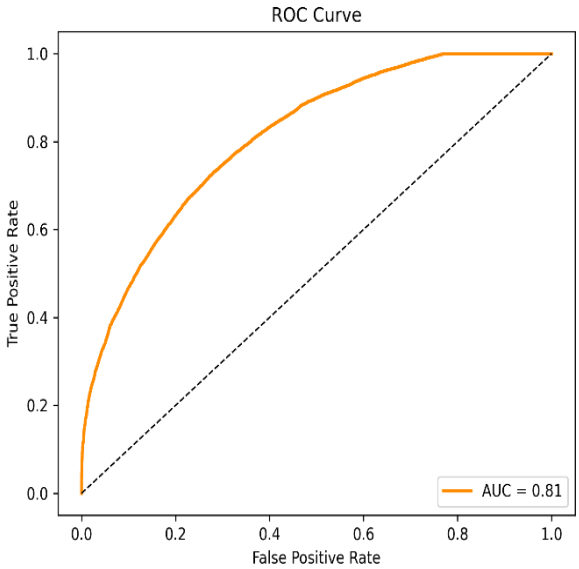
Tabel 3. Confusion Matrix Model Random Forest

	Prediksi 0	Prediksi 1
Aktual 0 (Negatif sebenarnya)	11873 (True Negative)	7098 (False Positive)
Aktual 1 (Positif sebenarnya)	2374 (False Negative)	10326 (True Positive)



Gambar 1. Visualisasi Confusion Matrix Model Random Forest

Kurva ROC yang ditunjukkan pada **Gambar 2** menggambarkan kemampuan diskriminatif model dalam membedakan individu berisiko dan tidak berisiko. Kurva ROC berada di atas garis diagonal dengan luas area (AUC) sebesar 0,81, menunjukkan performa klasifikasi yang baik.



Gambar 2. Kurva ROC Model Random Forest

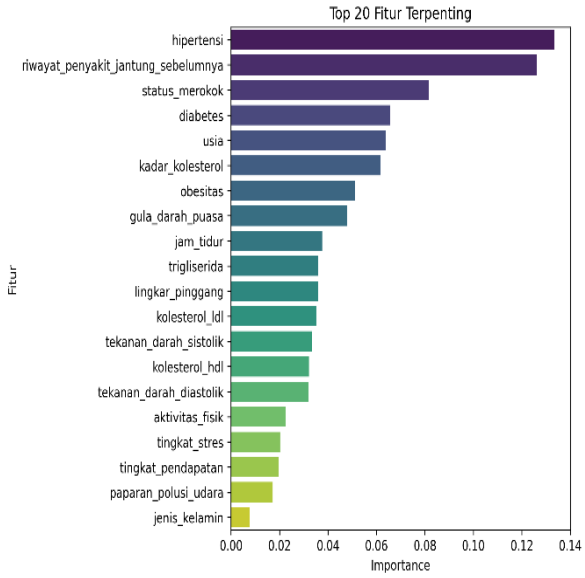
3.3. Faktor Risiko Utama

Hasil analisis *feature importance* disajikan pada Tabel 4 dan divisualisasikan pada Gambar 3. Sepuluh variabel dengan kontribusi tertinggi terhadap prediksi adalah hipertensi,

riwayat penyakit jantung, status merokok, diabetes, usia, kadar kolesterol total, obesitas, gula darah puasa, jam tidur, dan trigliserida.

Tabel 4. Nilai *Feature Importance* dari Variabel Prediktor (10 Fitur Teratas)

Peringkat	Fitur	Importance Score
1	hipertensi	0.13.00
2	riwayat_penyakit_jantung	0.13.00
3	status_merokok	0.08.00
4	diabetes	00.07
5	usia	0.06.00
6	kadar_kolesterol	00.06
7	obesitas	00.05
8	gula_darah_puasa	00.05
9	jam_tidur	00.04
10	trigliserida	00.04



Gambar 3. Feature Importance dari 20 Fitur Teratas

Faktor tersebut merepresentasikan kombinasi faktor metabolik, perilaku, dan kardiovaskular yang telah lama diidentifikasi sebagai penyebab utama peningkatan risiko serangan jantung. Temuan ini konsisten dengan penelitian epidemiologis sebelumnya yang menunjukkan peran dominan hipertensi, riwayat penyakit jantung, status merokok, diabetes, usia, kadar kolesterol, obesitas, gula darah puasa, jam tidur, dan trigliserida dalam kejadian penyakit jantung koroner. Analisis feature importance digunakan sebagai indikator kontribusi relatif variabel terhadap prediksi model dan tidak diinterpretasikan sebagai hubungan kausal [18].

KESIMPULAN

Penelitian ini berhasil mengembangkan model prediksi risiko serangan jantung menggunakan algoritma Random Forest dengan performa baik (akurasi 70,09%, AUC 0,81). Faktor risiko utama yang memengaruhi prediksi meliputi hipertensi, riwayat penyakit jantung, status merokok, diabetes, usia, kadar kolesterol, obesitas, gula darah puasa, jam tidur, dan trigliserida. Model ini berpotensi digunakan sebagai alat skrining awal berbasis data untuk mendukung pencegahan penyakit kardiovaskular di Indonesia. Pengembangan lebih lanjut disarankan dengan validasi eksternal dan integrasi faktor genetik serta sosial-ekonomi untuk meningkatkan generalisasi model.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Program Studi Informatika Universitas Jambi atas dukungan dan bimbingan selama penelitian berlangsung, serta kepada pengelola dataset Heart Attack Prediction in Indonesia di platform Kaggle atas ketersediaan data untuk analisis.

REFERENSI

[1] "Cardiovascular diseases (CVDs)." Accessed: Nov. 08, 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] "Penyakit Jantung Penyebab Utama Kematian, Kemenkes Perkuat Layanan Primer." Accessed: Nov. 08, 2025. [Online]. Available: https://kemkes.go.id/eng/penyakit-

jantung-penyebab-utama-kematian-kemenkes-perkuat-layanan-primer

[3] O. R. Olaniran, S. F. Olaniran, A. R. R. Alzahrani, N. M. S. Alharbi, and A. A. Alzahrani, "Random Forest Adaptation for High-Dimensional Count Regression," *Math. 2025, Vol. 13, Page 3041*, vol. 13, no. 18, p. 3041, Sep. 2025, doi: 10.3390/MATH13183041.

[4] D. Priyanto, H. Hairani, K. Marzuki, and M. Innuddin, "Optimization of Random Forest for Health Data Classification Using PCA and K-Means SMOTE-ENN," *Eng. Technol. Appl. Sci. Res.*, vol. 15, no. 5, pp. 27646–27652, Oct. 2025, doi: 10.48084/ETASR.12976.

[5] X. Wang and Z. Cheng, "Cross-Sectional Studies: Strengths, Weaknesses, and Recommendations," *Chest*, vol. 158, no. 1, pp. S65–S71, Jul. 2020, doi: 10.1016/j.chest.2020.03.012.

[6] "Heart Attack Prediction in Indonesia." Accessed: Nov. 08, 2025. [Online]. Available: https://www.kaggle.com/datasets/ankushpanday2/heart-attack-prediction-in-indonesia

[7] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, and O. Tabona, "A survey on missing data in machine learning," *J. Big Data 2021 81*, vol. 8, no. 1, pp. 1–37, Oct. 2021, doi: 10.1186/S40537-021-00516-9.

[8] E. Alshdaifat, D. Alshdaifat, A. Alsarhan, F. Hussein, and S. M. F. S. El-Salhi, "The Effect of Preprocessing Techniques, Applied to Numeric Features, on Classification Algorithms' Performance," *Data 2021, Vol. 6, Page 11*, vol. 6, no. 2, p. 11, Jan. 2021, doi: 10.3390/DATA6020011.

[9] S. Matharaarachchi, M. Domaratzki, and S. Muthukumarana, "Enhancing SMOTE for imbalanced data with abnormal minority instances," *Mach. Learn. with Appl.*, vol. 18, p. 100597, Dec. 2024, doi: 10.1016/J.MLWA.2024.100597.

[10] Y. Zhang, L. Deng, and B. Wei, "Imbalanced Data Classification Based on Improved Random-SMOTE and Feature Standard Deviation," *Math. 2024, Vol. 12, Page 1709*, vol. 12, no. 11, p. 1709, May 2024, doi: 10.3390/MATH12111709.

[11] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324/METRICS.

[12] P. Probst, "Hyperparameters, tuning and meta-learning for random forest and other machine learning algorithms," Jul. 2019, doi: 10.5282/EDOC.24557.

[13] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006, doi: 10.1016/J.PATREC.2005.10.010.

[14] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," *Sci. Reports 2024 141*, vol. 14, no. 1, pp. 1–14, Mar. 2024, doi: 10.1038/s41598-024-56706-x.

[15] M. L. Wallace *et al.*, "Use and misuse of random forest variable importance metrics in medicine: demonstrations through incident stroke prediction," *BMC Med. Res. Methodol. 2023 231*, vol. 23, no. 1, pp. 1–12, Jun. 2023, doi: 10.1186/S12874-023-01965-X.

[16] L. Yang *et al.*, "Study of cardiovascular disease prediction model based on random forest in eastern China," *Sci. Reports 2020 101*, vol. 10, no. 1, pp. 1–8, Mar. 2020, doi: 10.1038/s41598-020-62133-5.

[17] C. Yan, E. Gao, Y. Xing, J. Wang, and S. Liu, "Machine Learning Models for Cardiovascular Disease Prediction: A Comparative Study", doi: 10.1101/2024.05.27.596092.

[18] H. E. Bays *et al.*, "Ten things to know about ten cardiovascular disease risk

factors – 2022," *Am. J. Prev. Cardiol.*, vol. 10, p. 100342, Jun. 2022, doi: 10.1016/J.AJPC.2022.100342.