



Klasifikasi Tingkat Risiko Kardiovaskular Berbasis Pola Aktivitas Harian Menggunakan Random Forest

Abdul Hadi Al Hakim^{1*}, Al-Khowarizmi²

^{1*,2}Program Studi Sistem Informasi, Fakultas Ilmu Komputer dan Teknologi Informasi, Universitas Muhammadiyah Sumatera Utara, Medan, Indonesia

Email: ^{1*}abdulhadihakim03@gmail.com, ²alkhowarizmi@umsu.ac.id

Abstrak

Penyakit kardiovaskular merupakan penyebab kematian utama di dunia, sementara pemantauan mandiri yang non-invasif masih terbatas. Penelitian ini bertujuan membangun model klasifikasi tingkat risiko kardiovaskular (rendah, sedang, tinggi) berbasis pola aktivitas harian dan indikator kesehatan menggunakan algoritma Random Forest, serta mengimplementasikannya pada purwarupa sistem berbasis web. Data sekunder bersumber dari dataset Sleep Health and Lifestyle di Kaggle sebanyak 940 data mentah, dengan enam variabel masukan: tekanan darah, kategori indeks massa tubuh, detak jantung, tingkat stres, durasi dan gangguan tidur, serta jumlah langkah harian. Label risiko dibentuk dengan aturan ambang indikator kesehatan karena dataset tidak memiliki label risiko. Pra-pemrosesan meliputi pembersihan data, transformasi, standardisasi, dan penyeimbangan kelas dengan SMOTE pada data latih, menghasilkan 507 data bersih yang dibagi dengan rasio 80:20. Model dengan 100 pohon keputusan dievaluasi menggunakan confusion matrix. Hasil menunjukkan akurasi 99,77% dengan satu kesalahan klasifikasi pada data uji, serta F1-score rata-rata makro 99,80%. Karena label diturunkan dari variabel yang sama dengan fitur model, capaian ini mencerminkan kemampuan model mereplikasi aturan pelabelan dan belum merupakan bukti kemampuan prediksi klinis. Sistem web yang dikembangkan menyediakan formulir masukan, visualisasi hasil, grafik kepentingan fitur, dan riwayat pemeriksaan. Penelitian lanjutan memerlukan data berlabel diagnosis klinis dan validasi eksternal.

Kata Kunci: Random Forest, klasifikasi risiko kardiovaskular, SMOTE, pola aktivitas harian, sistem informasi web.

PENDAHULUAN

Penyakit kardiovaskular (*cardiovascular disease/CVD*) merupakan penyebab kematian nomor satu di dunia. Menurut World Health Organization (WHO, 2025), sekitar 19,8 juta orang meninggal akibat CVD pada tahun 2022, setara dengan kira-kira 32% seluruh kematian global, dan 85% di antaranya disebabkan oleh serangan jantung dan stroke. Upaya pencegahan bertumpu pada pengelolaan faktor risiko yang dapat dimodifikasi. American Heart Association merumuskan *Life's Essential 8*, yaitu delapan komponen kesehatan kardiovaskular yang mencakup pola makan, aktivitas fisik, paparan nikotin, kesehatan tidur, indeks massa tubuh, lipid darah, glukosa darah, dan tekanan darah, serta menempatkan kesehatan psikologis sebagai konteks yang penting (Lloyd-Jones dkk., 2022). Di Indonesia, pengaruh gaya hidup dan pengetahuan masyarakat terhadap risiko penyakit jantung koroner juga telah dikaji secara sistematis (Nasution dkk., 2025).

Sebagian indikator gaya hidup kini dapat diukur secara mudah dan non-invasif, misalnya jumlah langkah harian. Meta-analisis terhadap 17 studi kohort dengan 226.889 partisipan menunjukkan bahwa setiap penambahan 500 langkah per hari berkaitan dengan risiko mortalitas kardiovaskular yang lebih rendah (rasio hazard 0,93), sedangkan setiap penambahan 1.000 langkah berkaitan dengan risiko mortalitas semua penyebab yang lebih rendah (rasio hazard 0,85) (Banach dkk., 2023). Temuan ini mendukung pemanfaatan indikator aktivitas yang mudah dipantau sebagai dasar skrining mandiri yang bersifat pelengkap, bukan pengganti, pemeriksaan klinis.

Pembelajaran mesin banyak digunakan untuk klasifikasi risiko penyakit jantung. *Random Forest*, yaitu ansambel pohon keputusan yang dilatih dengan sampel *bootstrap* dan pemilihan fitur acak (Breiman, 2001), menunjukkan kinerja tinggi pada berbagai studi. Agusyl dan Firmansyah (2023) melaporkan akurasi 91%. Hidayat dkk. (2023) memperoleh akurasi 94% pada dataset Kaggle berisi 1.025 data. Ridwan dkk. (2025) melaporkan akurasi 99,5% untuk Random Forest dibandingkan 98,5% untuk Gradient Boosting

pada 1.000 baris data. Sahelvi dkk. (2025) memperoleh akurasi 99% (skema 80:20) dan 98% (skema 70:30), dibandingkan 83% dan 86% pada K-Nearest Neighbors. Adapun Arafat dkk. (2025) memperoleh akurasi 90,16% untuk Random Forest dan 85,25% untuk XGBoost pada 303 data UCI. Studi-studi tersebut pada umumnya menggunakan target biner (ada atau tidaknya penyakit jantung), dan sebagian menggunakan variabel klinis seperti hasil elektrokardiogram atau nyeri dada (Sahelvi dkk., 2025; Arafat dkk., 2025).

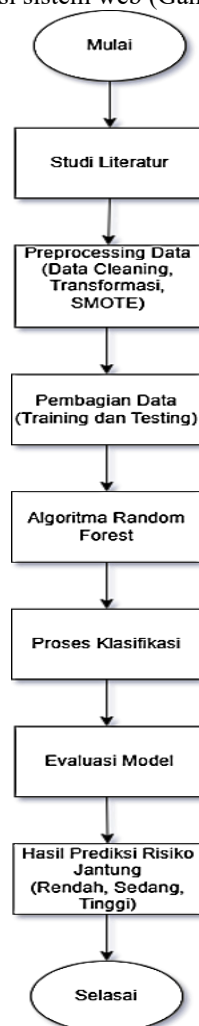
Pendekatan yang bertumpu pada target diagnosis klinis tidak selalu dapat diterapkan untuk pemantauan mandiri, karena menuntut data pemeriksaan yang tidak dimiliki pengguna awam. Kebaruan penelitian ini terletak pada (1) pemodelan tiga kategori risiko (rendah, sedang, tinggi) yang hanya menggunakan enam indikator non-invasif, yakni tekanan darah, kategori indeks massa tubuh, detak jantung, tingkat stres, durasi dan gangguan tidur, serta jumlah langkah harian; (2) implementasi model ke dalam sistem web yang dilengkapi visualisasi kepentingan fitur dan riwayat pemeriksaan; dan (3) pembahasan eksplisit mengenai validitas label berbasis aturan, yang berisiko menimbulkan kebocoran data (*data leakage*) dan hasil yang terlalu optimistis apabila fitur model sekaligus menentukan label (Kapoor dan Narayanan, 2023).

Permasalahan penelitian ini adalah sejauh mana algoritma Random Forest dapat mengklasifikasikan tiga tingkat risiko kardiovaskular dari enam indikator tersebut, dan bagaimana hasilnya diinterpretasikan secara tepat. Tujuan penelitian ini adalah (1) membangun model klasifikasi tingkat risiko kardiovaskular berbasis Random Forest; (2) mengevaluasi kinerja model menggunakan *confusion matrix*; (3) mengimplementasikan model ke dalam purwarupa sistem web; dan (4) menelaah keterbatasan validitas hasil akibat pelabelan berbasis aturan..

METODE

Desain penelitian

Penelitian ini merupakan penelitian kuantitatif eksperimental terapan dengan data sekunder. Tahapannya meliputi pengumpulan data, pelabelan risiko, pra-pemrosesan, pembagian data, pemodelan Random Forest, evaluasi, dan implementasi sistem web (Gambar 1).



Gambar 1. Tahapan penelitian

Dataset dan variabel

Data diperoleh dari dataset publik *Sleep Health and Lifestyle* di Kaggle (Tharmalingam, t.t.). Penyusunnya menyatakan bahwa data tersebut bersifat sintetis dan dibuat untuk tujuan ilustrasi, sehingga tidak merepresentasikan pasien nyata. Dataset yang digunakan terdiri atas 940 data mentah dengan 15 atribut. Dari seluruh atribut dipilih enam variabel masukan, yaitu tekanan darah (mmHg), kategori indeks massa tubuh (IMT: normal, *overweight*, *obese*), detak jantung (bpm), tingkat stres (rendah, sedang, tinggi), durasi tidur (jam) beserta gangguan tidur (tidak ada, *insomnia*, *sleep apnea*), dan jumlah langkah per hari. Pemilihan variabel ini selaras dengan komponen aktivitas fisik, kesehatan tidur, indeks massa tubuh, dan tekanan darah pada *Life's Essential 8* (Lloyd-Jones dkk., 2022).

Pelabelan risiko

Dataset tidak memiliki label risiko kardiovaskular, sehingga label dibentuk otomatis menggunakan aturan berbasis ambang indikator kesehatan. Aturan ini disusun peneliti dengan mengacu pada pedoman pengendalian penyakit jantung dan pembuluh darah (Kementerian Kesehatan Republik Indonesia, 2009) (Tabel 1). Label dikodekan 0 (Rendah), 1 (Sedang), dan 2 (Tinggi).

Tabel 1. Aturan pelabelan tingkat risiko

Kategori	Kriteria
Rendah (0)	Seluruh parameter normal: tekanan darah < 120/80 mmHg, IMT normal, jumlah langkah harian tinggi, tidak ada stres berat, dan tidak ada gangguan tidur
Sedang (1)	Tidak memenuhi kriteria Tinggi, tetapi salah satu terpenuhi: tekanan darah 120–139 mmHg, IMT <i>overweight</i> , atau tingkat stres sedang
Tinggi (2)	Salah satu terpenuhi: tekanan darah $\geq$ 140/90 mmHg, IMT <i>obese</i> , atau gangguan tidur (<i>sleep apnea</i> /insomnia)

Pra-pemrosesan data

Pembersihan data dilakukan dengan memeriksa nilai kosong, data duplikat, dan nilai tidak valid, sehingga diperoleh 507 data bersih dari 940 data mentah. Variabel kategorikal diubah menjadi numerik dan seluruh fitur numerik distandardisasi dengan *Standard Scaling*. Data dibagi menjadi data latih dan data uji dengan rasio 80:20. Ketidakseimbangan kelas pada data latih ditangani dengan *Synthetic Minority Over-sampling Technique* (SMOTE), yang membangkitkan sampel sintetis kelas minoritas melalui interpolasi antartetangga terdekat (Chawla dkk., 2002). SMOTE diterapkan hanya pada data latih agar data uji tetap merepresentasikan distribusi asli dan terhindar dari kebocoran data.

Pemodelan Random Forest

Model dibangun dengan *RandomForestClassifier* pada pustaka scikit-learn (Python) dengan parameter jumlah pohon (*n_estimators*) 100, kedalaman maksimum (*max_depth*) 10, kriteria pemisahan *gini*, jumlah fitur per pemisahan akar kuadrat dari jumlah fitur (*max_features = sqrt*), dan *random_state* 42. Setiap pohon dilatih pada sampel *bootstrap* dan prediksi akhir ditentukan melalui pemungutan suara mayoritas (Breiman, 2001).

Evaluasi model

Kinerja model dievaluasi pada data uji menggunakan *confusion matrix*. Akurasi keseluruhan dihitung sebagai proporsi prediksi benar terhadap seluruh data uji. Presisi, *recall*, dan *F1-score* dihitung per kelas dengan pendekatan satu-lawan-lainnya (*one-vs-rest*) menurut Persamaan (1)–(3), kemudian dirata-ratakan secara makro.

$$\text{Presisi} = \text{TP} / (\text{TP} + \text{FP}) \quad (1)$$

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN}) \quad (2)$$

$$\text{F1-score} = 2 \times (\text{Presisi} \times \text{Recall}) / (\text{Presisi} + \text{Recall}) \quad (3)$$

dengan TP, FP, dan FN berturut-turut adalah *true positive*, *false positive*, dan *false negative*.

Implementasi sistem

Model terlatih diintegrasikan ke dalam aplikasi web yang menerima enam variabel masukan dari pengguna terautentikasi, memprosesnya melalui tahap pra-pemrosesan, lalu menampilkan kategori risiko beserta probabilitasnya.

HASIL DAN PEMBAHASAN

Hasil evaluasi model

Setelah pembersihan diperoleh 507 data bersih dengan enam variabel masukan dan satu label risiko. *Confusion matrix* pada data uji (Tabel 2) yang memuat total 432 data dengan rincian: 234 Rendah, 130 Sedang, dan 68 Tinggi.

Tabel 2. *Confusion matrix* data uji

Aktual \ Prediksi	Rendah	Sedang	Tinggi
Rendah	234	0	0
Sedang	1	129	0
Tinggi	0	0	68

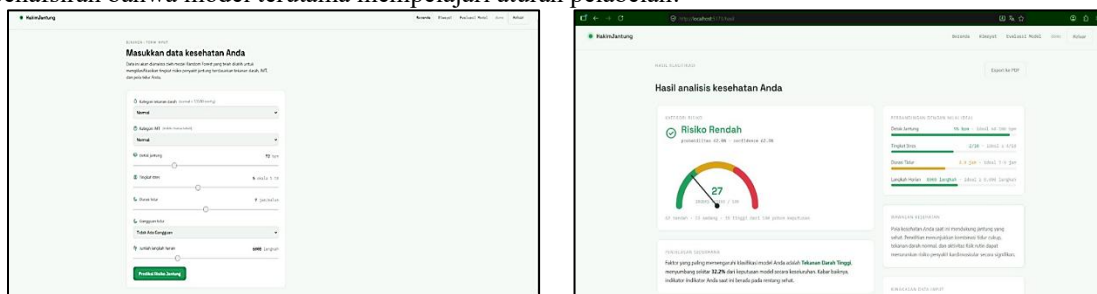
Model mengklasifikasikan 431 dari 432 data uji dengan benar, sehingga akurasi keseluruhan mencapai 99,77%. Satu-satunya kesalahan adalah satu data kelas Sedang yang diprediksi sebagai Rendah. Metrik per kelas disajikan pada Tabel 3.

Tabel 3. Presisi, *recall*, dan *F1-score* per kelas (%)

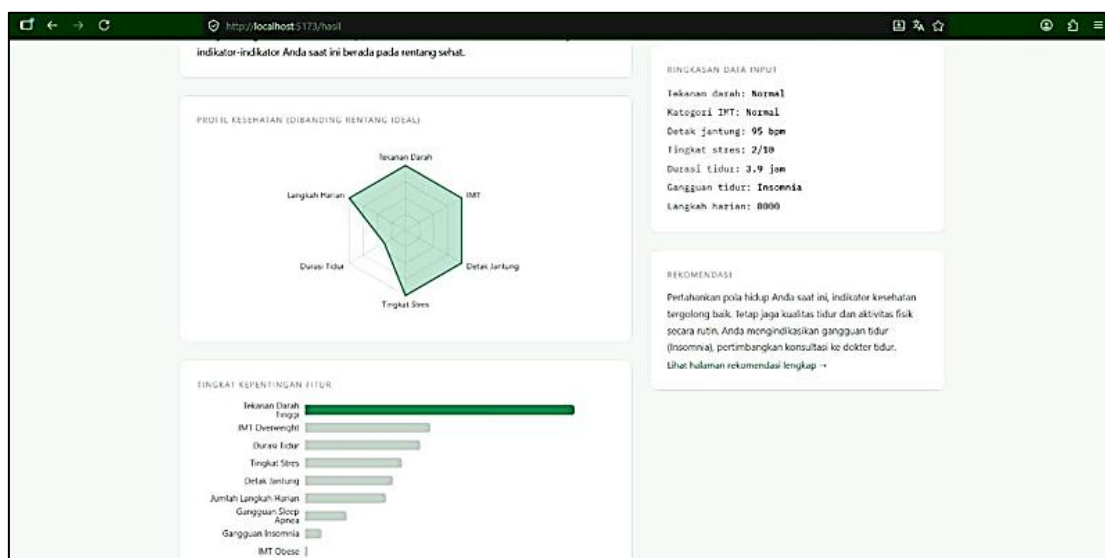
Kelas	Presisi	Recall	F1-score	Jumlah data
Rendah	99,57	100,00	99,79	234
Sedang	100,00	99,23	99,61	130
Tinggi	100,00	100,00	100,00	68
Rata-rata makro	99,86	99,74	99,80	432

Hasil implementasi sistem web

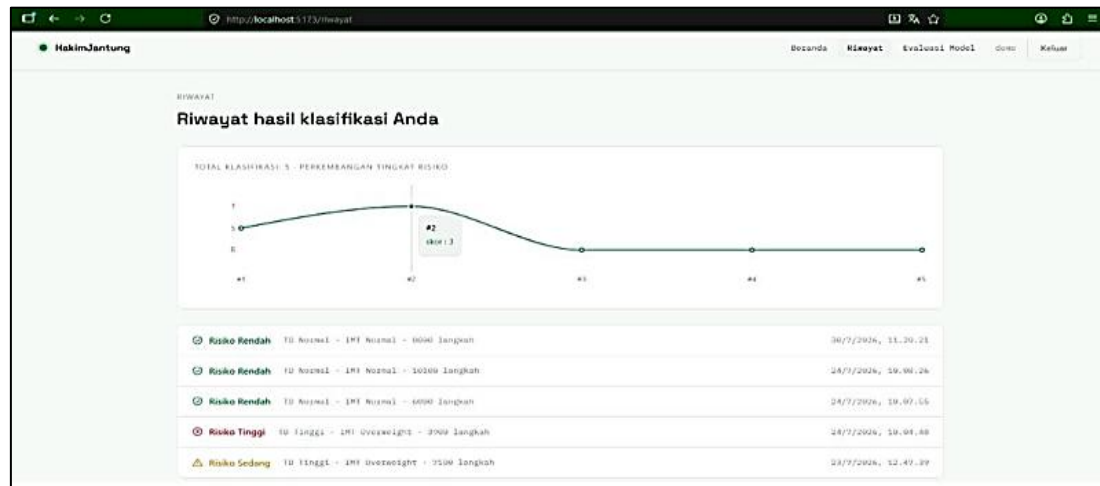
Sistem web terdiri atas halaman beranda, autentikasi (masuk dan daftar), formulir masukan, halaman hasil, dan halaman riwayat. Pada formulir masukan, pengguna memilih kategori tekanan darah dan IMT melalui menu pilihan, mengatur detak jantung, tingkat stres, durasi tidur, dan jumlah langkah harian melalui penggeser (*slider*), serta memilih jenis gangguan tidur (Gambar 2). Halaman hasil menampilkan kategori risiko beserta probabilitasnya dalam bentuk indikator speedometer, ringkasan data masukan, perbandingan nilai pengguna dengan nilai ideal, grafik radar profil kesehatan, grafik kepentingan fitur, penjelasan sederhana tentang model, dan fitur ekspor laporan ke PDF. Grafik kepentingan fitur (Gambar 3) menunjukkan bahwa tekanan darah tinggi merupakan fitur yang paling dominan, dengan batang sekitar dua kali lebih panjang daripada fitur peringkat kedua. Fitur berikutnya secara berurutan adalah IMT *overweight*, durasi tidur, tingkat stres, detak jantung, jumlah langkah harian, gangguan *sleep apnea*, gangguan insomnia, dan IMT *obese*. Dominasi tekanan darah tinggi sejalan dengan aturan pelabelan pada Tabel 1, karena kategori tekanan darah menentukan penetapan kelas risiko tinggi. Kondisi ini mendukung penafsiran bahwa model terutama mempelajari aturan pelabelan.



Gambar 2. Formulir masukan dan hasil klasifikasi risiko



Gambar 3. Kepentingan fitur pada model Random Forest



Gambar 4. Halaman riwayat hasil klasifikasi

Pembahasan

Akurasi 99,77% berada dalam kisaran 90,16% hingga 99,5% yang dilaporkan studi terdahulu (Agusyul dan Firmansyah, 2023; Hidayat dkk., 2023; Ridwan dkk., 2025; Sahelvi dkk., 2025; Arafat dkk., 2025). Namun, angka tersebut tidak dapat dibandingkan secara langsung. Studi terdahulu pada umumnya memprediksi target biner ada atau tidaknya penyakit jantung yang tersedia pada datasetnya, sedangkan penelitian ini memprediksi label tiga kelas yang dibentuk dari aturan ambang oleh peneliti.

Tingkat kesesuaian yang hampir sempurna perlu dibaca dengan hati-hati. Label risiko diturunkan dari tekanan darah, IMT, tingkat stres, dan gangguan tidur, yang juga menjadi fitur masukan model. Pohon keputusan memisahkan data melalui ambang nilai, sehingga ansambel pohon sangat mudah merekonstruksi aturan berbasis ambang tersebut. Satu kesalahan yang terjadi, yaitu data Sedang yang diprediksi Rendah, kemungkinan berada di sekitar batas antarkategori. Dengan demikian, akurasi ini lebih tepat ditafsirkan sebagai tingkat kesesuaian model terhadap aturan pelabelan, bukan kemampuan memprediksi kejadian penyakit jantung. Kondisi ini sejalan dengan salah satu bentuk kebocoran data yang dibahas Kapoor dan Narayanan (2023), yaitu penggunaan fitur yang merupakan penentu variabel target sehingga estimasi kinerja menjadi terlalu optimistis. Penerapan SMOTE hanya pada data latih sudah mencegah kebocoran dari tahap penyeimbangan kelas, tetapi tidak menghilangkan kebocoran yang berasal dari proses pelabelan.

Keterbatasan lain perlu dicatat. Pertama, data bersifat sintetis dan berukuran kecil sehingga belum merepresentasikan populasi nyata. Kedua, evaluasi memakai satu pembagian data 80:20 tanpa validasi silang maupun validasi eksternal. Ketiga, penelitian hanya menguji satu algoritma sehingga belum ada pembandingan seperti pohon keputusan tunggal atau regresi logistik. Keempat, aturan pelabelan belum divalidasi oleh tenaga medis.

Meskipun demikian, penelitian ini menghasilkan purwarupa sistem yang mengubah enam indikator yang mudah diperoleh menjadi umpan balik risiko yang cepat, lengkap dengan visualisasi, penjelasan fitur, dan riwayat pemantauan. Dalam bentuk saat ini, sistem berfungsi sebagai alat bantu edukasi dan pemantauan mandiri, bukan alat diagnosis. Nilai tambah pembelajaran mesin baru dapat diuji apabila model dilatih dengan hasil klinis nyata, misalnya diagnosis penyakit kardiovaskular, dengan variabel penentu label dikeluarkan dari fitur. Pengguna dengan kategori Sedang atau Tinggi tetap disarankan berkonsultasi dengan tenaga Kesehatan.

KESIMPULAN

Penelitian ini membangun model Random Forest untuk mengklasifikasikan tiga tingkat risiko kardiovaskular berdasarkan enam indikator non-invasif dan mengimplementasikannya ke dalam purwarupa sistem web. Pada data uji, model mencapai akurasi 99,77% dan *F1-score* rata-rata makro 99,80% dengan satu kesalahan klasifikasi. Karena label dibentuk dari aturan yang memakai variabel yang sama dengan fitur model, hasil ini menunjukkan kesesuaian model terhadap aturan pelabelan dan belum membuktikan kemampuan prediksi klinis. Sistem web yang dihasilkan menyediakan formulir masukan, visualisasi hasil, grafik kepentingan fitur, dan riwayat pemeriksaan sebagai alat bantu pemantauan mandiri.

Saran untuk penelitian selanjutnya adalah (1) menggunakan data nyata berlabel diagnosis klinis dengan ukuran yang lebih besar; (2) mengeluarkan variabel penentu label dari fitur model; (3) menerapkan validasi silang dan validasi eksternal serta membandingkan dengan algoritma lain; (4) memvalidasi aturan pelabelan bersama tenaga medis; dan (5) mengintegrasikan data dari perangkat sensor atau *wearable* dengan notifikasi berbasis ambang, sebagaimana diterapkan pada sistem pemantauan berbasis sensor, disertai pengujian *usability* dan keamanan data kesehatan pengguna.

DAFTAR PUSTAKA

- Agusyul, A. Y., & Firmansyah. (2023). Prediksi penyakit jantung menggunakan algoritma Random Forest. *Jurnal Minfo Polgan*, 12(2). <https://doi.org/10.33395/jmp.v12i2.13214>
- Arafat, C., Ramadhan, M. C., Jafar, F. A., Pramudya, M. R. A., & Ismanto, E. (2025). Perbandingan algoritma Random Forest dan XGBoost untuk klasifikasi penyakit jantung berdasarkan data medis. *Jurnal Fasilkom*, 15(2), 430–435. <https://ejurnal.umri.ac.id/index.php/JIK/article/download/9927/3810/33590>
- Banach, M., Lewek, J., Surma, S., Penson, P. E., Sahebkar, A., Martin, S. S., Bajraktari, G., Henein, M. Y., Reiner, Z., Bielecka-Dabrowa, A., & Bytyci, I. (2023). The association between daily step count and all-cause and cardiovascular mortality: A meta-analysis. *European Journal of Preventive Cardiology*, 30(18), 1975–1985. <https://doi.org/10.1093/eurjpc/zwad229>
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- 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. <https://doi.org/10.1613/jair.953>
- Hidayat, Sunyoto, A., & Al Fatta, H. (2023). Klasifikasi penyakit jantung menggunakan Random Forest Clasifier. *Jurnal Sistem Komputer dan Kecerdasan Buatan*, 7(1), 31–40. <https://jurnal.tau.ac.id/index.php/siskom-kb/article/download/464/341>
- Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9), 100804. <https://doi.org/10.1016/j.patter.2023.100804>
- Kementerian Kesehatan Republik Indonesia. (2009). *Keputusan Menteri Kesehatan Republik Indonesia Nomor 854/MENKES/SK/IX/2009 tentang Pedoman Pengendalian Penyakit Jantung dan Pembuluh Darah*. Kementerian Kesehatan RI. <https://pasal.id/peraturan/permen/kepmenkes-854-menkes-sk-ix-2009>
- Lloyd-Jones, D. M., Allen, N. B., Anderson, C. A. M., Black, T., Brewer, L. C., Foraker, R. E., Grandner, M. A., Lavretsky, H., Perak, A. M., Sharma, G., & Rosamond, W. (2022). Life's Essential 8: Updating and enhancing the American Heart Association's construct of cardiovascular health. *Circulation*, 146(5), e18–e43. <https://doi.org/10.1161/CIR.0000000000001078>
- Maharani, D., Atmia, K., & Novita, S. (2025). Blynk-based air handling unit (AHU) air filter monitoring system prototype at Juwata Tarakan Airport Terminal. *Indonesian Journal of Multidisciplinary Scientific Studies*, 3(5), 392–397. <https://doi.org/10.33151/ijomss.v3i5.696>
- Nasution, I. S., Rahmadani, A. D., Audina, W., Sari, D. P., & Sari, N. D. (2025). Systematic review: Pengaruh gaya hidup dan pengetahuan masyarakat terhadap risiko penyakit jantung koroner. *Sehat Rakyat: Jurnal Kesehatan Masyarakat*, 4(2), 287–298. <https://journal.yp3a.org/index.php/sehatrakyat/article/view/4337>
- Ridwan, R., Handayani, H. H., Lestari, S. A. P., & Cahyana, Y. (2025). Evaluasi kinerja algoritma Random Forest dan Gradient Boosting untuk klasifikasi penyakit jantung. *Jurnal Komtika (Komputasi dan Informatika)*, 9(1). <https://journal.unimma.ac.id/komtika/article/view/13450>
- Sahelvi, E., Cikita, P., Sapitri, R. M., Rahmaddeni, R., & Efrizoni, L. (2025). Perbandingan algoritma K-Nearest Neighbors dan Random Forest untuk rekomendasi gaya hidup sehat dalam mencegah penyakit jantung. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, 5(3), 830–840. <https://doi.org/10.57152/malcom.v5i3.1972>
- Tharmalingam, L. (t.t.). *Sleep health and lifestyle dataset* [Data set]. Kaggle. Diakses 5 Oktober 2026, dari <https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset>
- World Health Organization. (2025). *Cardiovascular diseases (CVDs)*. Diakses 5 Oktober 2026, dari [https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-\(cvds\)](https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds))