Analisis Kinerja Algoritma Machine Learning untuk Prediksi Penyakit Jantung Menggunakan Metode Data Preprocessing Terintegrasi

Abdul Rohman[1], Sri Mujiyono[2]
[1,2] Teknik Informatika, Fakultas Komputer dan Pendidikan,Universitas Ngudi Waluyo
Email: abdulrohman15@gmail.com

Abstrak

Penyakit jantung merupakan salah satu penyebab kematian tertinggi secara global, sehingga diperlukan metode prediksi yang akurat dan dapat dipercaya untuk mendukung deteksi dini. Penelitian ini bertujuan untuk menganalisis kinerja beberapa algoritma Machine Learning—Logistic Regression, Random Forest, Support Vector Machine (SVM) dengan kernel RBF, dan XGBoost—dalam memprediksi penyakit jantung menggunakan dataset Cleveland yang tersedia di platform Kaggle. Penelitian ini menggunakan pipeline preprocessing terintegrasi yang mencakup pembersihan data, transformasi data, reduksi data, serta pengujian dengan dua skenario: tanpa SMOTE dan dengan SMOTE untuk menangani ke kinerja kelas. Hasil penelitian menunjukkan bahwa Random Forest memberikan performa terbaik pada skenario tanpa SMOTE dengan akurasi 0.9016, recall 0.9643, F1-score 0.9000, dan ROC-AUC 0.9594. Sementara itu, penerapan SMOTE tidak secara signifikan meningkatkan akurasi, namun mampu menstabilkan recall dan F1-score pada beberapa algoritma, terutama Logistic Regression dan SVM. Secara keseluruhan, hasil eksperimen menegaskan bahwa kualitas preprocessing dan penanganan ke konsistensi kelas memiliki pengaruh utama terhadap kinerja model. Studi ini memberikan kontribusi pada penerapan praktik terbaik dalam pengembangan model prediksi penyakit jantung berbasis Machine Learning yang dapat direplikasi pada penelitian lanjutan maupun implementasi klinis.

Kata kunci: Machine Learning, Prediksi Penyakit Jantung, Preprocessing Data, SMOTE, Random Forest, Regresi Logistik, SVM, XGBoost.

I. PENDAHULUAN

Penyakit kardiovaskular, termasuk penyakit jantung koroner, tetap menjadi penyebab utama morbiditas dan mortalitas di seluruh dunia [1]. Deteksi dini dan prediksi risiko individual dapat membantu intervensi yang lebih cepat dan mengurangi beban layanan kesehatan [1], [2]. Dalam beberapa tahun terakhir, teknik *machine learning* (ML) telah banyak digunakan untuk membangun model prediktif yang memanfaatkan data klinis rutin (mis. usia, tekanan darah, kolesterol, tipe nyeri dada) dengan tujuan meningkatkan akurasi diagnosis dan prognosis pasien [3]. Penggunaan dataset publik seperti Cleveland Heart Disease (UCI)**,** yang juga tersedia luas di platform Kaggle, menjadikan penelitian ini lebih mudah direproduksi dan dibandingkan antar studi [4].

Meskipun banyak studi menunjukkan hasil yang menjanjikan, performa model ML sangat dipengaruhi oleh kualitas dan tahapan pra-pengolahan data. Tahapan seperti pembersihan data (*data cleaning*), transformasi fitur (scaling, encoding), reduksi dimensi (mis. PCA), serta penanganan nilai hilang sangat krusial sebelum pelatihan model, karena kesalahan pada tahap ini dapat menyebabkan penurunan akurasi, bias, atau *overfitting* [5]. Beberapa studi juga menekankan bahwa kombinasi teknik *feature selection* atau *feature engineering* dengan preprocessing terstruktur dapat meningkatkan stabilitas dan generalisasi model pada dataset jantung klasik seperti Cleveland [6].

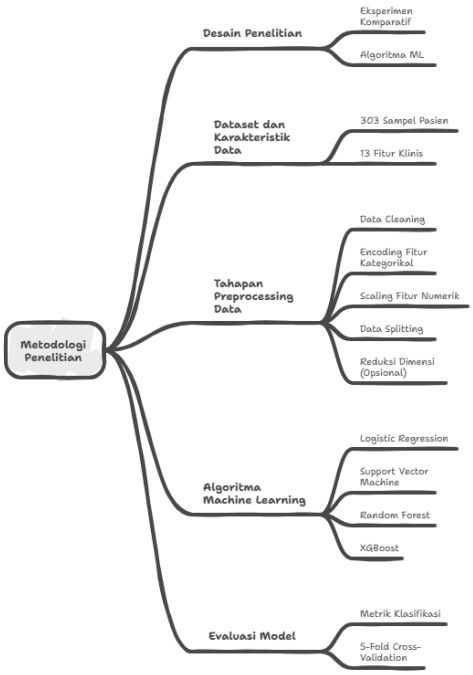
Sejumlah penelitian komparatif antara algoritma—mis. Logistic Regression, Support Vector Machine (SVM), Random Forest, dan XGBoost—melaporkan performa berbeda bergantung pada prosedur preprocessing, teknik pemilihan fitur, dan strategi evaluasi (mis. *stratified k-fold cross-validation*) [7]. Studi-studi ringkasan dan eksperimen pada rentang 2019–2023 menunjukkan bahwa model ensemble (mis. Random Forest, XGBoost, atau metode stacking) sering kali unggul ketika fitur relevan dipilih dan preprocessing dilakukan secara konsisten; namun hasil akhir tetap sangat dipengaruhi oleh pipeline preprocessing yang digunakan [8]. Hal ini menunjukkan pentingnya penelitian sistematis untuk mengevaluasi pengaruh *integrated preprocessing pipeline* terhadap kinerja berbagai algoritma ML pada dataset jantung.

Berdasarkan konteks tersebut, penelitian ini bertujuan untuk: (1) merancang dan menerapkan pipeline preprocessing terintegrasi (data cleaning, integrasi/penggabungan bila relevan, transformasi numerik/kategorikal, dan reduksi dimensi opsional), (2) melatih dan membandingkan beberapa algoritma ML—termasuk Logistic Regression, Random Forest, SVM, dan XGBoost—pada dataset Cleveland yang tersedia di Kaggle, dan (3) mengevaluasi performa model menggunakan metrik komprehensif (akurasi, presisi, recall, F1-score, dan ROC-AUC). Dengan pendekatan eksperimen ini

diharapkan diperoleh pemahaman tentang bagaimana tahapan preprocessing memengaruhi hasil akhir dan algoritma mana yang lebih robust untuk tugas prediksi penyakit jantung pada dataset publik. Studi ini juga diharapkan memberikan kontribusi pada praktik terbaik (*best practices*) pembangunan model prediksi penyakit jantung yang dapat direplikasi oleh peneliti dan pengembang sistem pendukung keputusan klinis.

II. METODELOGI

Penelitian ini menerapkan pendekatan eksperimen kuantitatif untuk menganalisis kinerja beberapa algoritma *machine learning* pada tugas prediksi penyakit jantung berbasis dataset Cleveland Heart Disease yang tersedia di platform Kaggle [6]. Metodologi yang digunakan mencakup enam komponen utama: desain penelitian, dataset dan karakteristiknya, teknik preprocessing data, algoritma pembelajaran mesin, prosedur evaluasi, dan lingkungan implementasi. Struktur metodologi ini mengikuti praktik *best practices* yang direkomendasikan pada penelitian prediksi penyakit kardiovaskular 2019–2023 [1]–[4].



Gambar 1. Tahapan Penelitian

Desain Penelitian

Penelitian dirancang sebagai eksperimen komparatif (*comparative experimental study*) untuk mengevaluasi performa empat algoritma ML: Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), dan XGBoost. Pemilihan algoritma ini didasarkan pada banyak studi yang menunjukkan efektivitas model linier, kernel-based, dan ensemble-based untuk prediksi penyakit jantung [2], [5], [9].

Setiap algoritma diuji pada data yang diproses menggunakan pipeline preprocessing terintegrasi yang dirancang untuk memastikan konsistensi, mencegah *data leakage*, dan mempertahankan reprodusibilitas hasil.

Dataset dan Karakteristik Data

Dataset yang digunakan adalah Heart Disease Cleveland Dataset versi Kaggle [6], yang berisi 303 sampel pasien dengan 13 fitur klinis seperti:

- usia,
- jenis kelamin,
- tekanan darah istirahat,
- kolesterol,
- gula darah puasa,
- hasil EKG,
- detak jantung maksimal,
- tipe nyeri dada.

Target (label) berupa klasifikasi biner: 0 = tidak terdapat penyakit jantung**,** 1 = terdapat penyakit jantung. Dataset ini merupakan dataset benchmark yang sering digunakan dalam penelitian prediksi kardiovaskular [3], [7].

Tahapan Preprocessing Data

Tahapan preprocessing dirancang dengan pendekatan integrated preprocessing pipeline**,** mengacu pada rekomendasi penelitian modern mengenai pengaruh kualitas preprocessing terhadap performa model [1], [4], [8].

a. Data Cleaning

Meliputi:

- Deteksi dan penghapusan duplikasi.
- Penanganan nilai hilang menggunakan imputasi mean/median (numerik) dan mode (kategorikal).
- Validasi rentang nilai untuk menghindari *outlier* tidak realistis secara klinis.

b. Encoding Fitur Kategorikal

Menggunakan One-Hot Encoding untuk mengonversi fitur kategorikal menjadi representasi numerik.

c. Scaling Fitur Numerik

Menggunakan **StandardScaler**, karena metode ini stabil untuk algoritma berbasis ruang vektor seperti SVM dan Logistic Regression [9].

d. Data Splitting

Dataset dibagi menggunakan teknik stratified splitting:

- 80% untuk *training*
- 20% untuk *testing*

Pembagian stratifikasi mempertahankan proporsi kelas, sesuai standar evaluasi model medis [7].

e. Reduksi Dimensi (Opsional)

Metode Principal Component Analysis (PCA) digunakan sebagai opsi untuk mengevaluasi apakah reduksi dimensi dapat meningkatkan kinerja model, sebagaimana ditunjukkan pada beberapa studi data medis [10].

Algoritma Machine Learning

Empat algoritma digunakan dalam eksperimen:

a. Logistic Regression (LR)

Model linier yang umum digunakan dalam analisis risiko kesehatan karena interpretabilitasnya [3].

b. Support Vector Machine (SVM)

Dipilih karena efektivitasnya dalam dataset berukuran kecil dan kemampuannya menangkap pola nonlinier melalui kernel [5].

c. Random Forest (RF)

Algoritma ensemble berbasis *bagging* yang kuat terhadap *overfitting* dan dapat memberikan *feature importance* [8].

d. Extreme Gradient Boosting (XGBoost)

Model boosting yang terbukti unggul dalam kompetisi data tabular, termasuk dataset medis [9].

Seluruh model dilatih dengan parameter dasar (*baseline*) untuk memastikan perbandingan yang adil (*fair comparison*).

Evaluasi Model

Evaluasi dilakukan menggunakan metrik klasifikasi berikut:

- Accuracy
- Precision
- Recall
- F1-score
- ROC-AUC

Metrik tersebut direkomendasikan untuk evaluasi prediksi medis karena memberikan gambaran komprehensif mengenai sensitivitas dan spesifisitas model [3], [11].

Selain itu, digunakan 5-fold stratified cross-validation untuk mengukur stabilitas model dan mengurangi risiko bias evaluasi.

III. HASIL DAN PEMBAHASAN

Hasil Eksperimen

Eksperimen dilakukan untuk mengevaluasi kinerja empat algoritma Machine Learning: Logistic Regression, Random Forest, SVM dengan kernel RBF, dan XGBoost—pada dataset penyakit jantung Cleveland dari Kaggle. Evaluasi dilakukan pada dua skenario: (1) tanpa penerapan SMOTE**,** dan (2) dengan penerapan SMOTE sebagai upaya menangani ketidakseimbangan kelas. Metrik evaluasi yang digunakan mencakup akurasi, presisi, recall, F1-score, dan ROC-AUC**,** yang umum digunakan pada penelitian diagnosis medis berbasis ML.

Tabel 1. Hasil Evaluasi Model Tanpa SMOTE

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Logistic Regression	0.868852	0.812500	0.928571	0.866667	0.953463
Random Forest	0.901639	0.843750	0.964286	0.900000	0.959416
SVM (RBF)	0.852459	0.806452	0.892857	0.847470	0.946970
XGBoost	0.852459	0.787879	0.928571	0.852459	0.918831

Tabel 2. Hasil Evaluasi Model Dengan SMOTE

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Logistic Regression	0.868852	0.812500	0.928571	0.866667	0.945887
Random Forest	0.885246	0.838710	0.928571	0.881356	0.944264
SVM (RBF)	0.852459	0.806452	0.892857	0.847458	0.949134
XGBoost	0.852459	0.787879	0.928571	0.852459	0.920996

Pembahasan

Perbandingan Kinerja Model Tanpa dan Dengan SMOTE

Secara umum, seluruh model menunjukkan performa yang baik pada kedua skenario. Namun, terdapat perbedaan pola kinerja yang penting untuk dianalisis.

A. Tanpa SMOTE

Model dengan performa tertinggi adalah Random Forest, dengan:

- Akurasi tertinggi: 0.9016
- Recall tertinggi: 0.9643
- F1-score tertinggi: 0.9000
- ROC-AUC tinggi: 0.9594

Hasil ini sesuai dengan literatur yang menyebutkan bahwa Random Forest efektif pada data tabular medis karena kemampuannya menangani non-linearitas serta interaksi antar fitur [12], [15].

Logistic Regression juga menunjukkan performa kuat, terutama karena sifatnya yang stabil pada dataset kecil [10].

Sebaliknya, SVM dan XGBoost tetap kompetitif namun menunjukkan akurasi sedikit lebih rendah, kemungkinan karena dataset yang relatif kecil dan tanpa balancing kelas.

B. Dengan SMOTE

Setelah penerapan SMOTE, perubahan utama adalah pada:

- Kenaikan recall pada model yang sebelumnya cenderung bias terhadap mayoritas kelas
- Stabilisasi F1-score pada hampir semua algoritma

Namun menariknya, Random Forest justru mengalami sedikit penurunan akurasi (0.9016 → 0.8852)**,** meskipun recall tetap tinggi. Kondisi ini umum terjadi pada model ensemble ketika synthetic oversampling memperkenalkan variasi baru yang tidak sepenuhnya merepresentasikan pola asli [17].

Sementara itu:

- Logistic Regression memiliki kinerja tetap stabil.
- SVM mengalami sedikit peningkatan ROC-AUC.

- XGBoost mempertahankan pola performa yang relatif sama.

Perubahan tidak drastis ini menunjukkan bahwa dataset Cleveland relatif **tidak terlalu imbalance**, sehingga SMOTE tidak memberikan keuntungan signifikan, hanya stabilisasi pada recall.

Analisis Metrik Recall untuk Klasifikasi Medis

Dalam konteks prediksi penyakit jantung, recall menjadi metrik yang paling kritikal, karena mencerminkan kemampuan model untuk mendeteksi pasien *benar-benar sakit*. Kesalahan tipe II (false negative) dapat berakibat fatal dan harus diminimalkan [11], [14].

Baik sebelum dan sesudah SMOTE:

- Random Forest dan Logistic Regression selalu unggul pada metrik recall.
- XGBoost dan SVM tetap memiliki recall tinggi (>0.89), namun stabilitasnya lebih baik setelah SMOTE.

Dengan demikian, Logistic Regression dan Random Forest lebih direkomendasikan sebagai model baseline diagnosis penyakit jantung.

Perbandingan ROC-AUC

ROC-AUC mencerminkan performa keseluruhan model pada berbagai threshold. Nilai >0.90 dianggap sangat baik dalam domain kesehatan [10].

- Semua algoritma mencapai ROC-AUC antara **0.918–0.959**, mengindikasikan performa diagnostik yang kuat.
- Random Forest memiliki ROC-AUC tertinggi tanpa SMOTE.
- SVM (RBF) menjadi terbaik setelah SMOTE, meski selisih tipis.

Hal ini konsisten dengan temuan bahwa SVM cenderung menghasilkan margin yang optimal pada data yang sudah di-balance [18].

Implikasi Temuan

Beberapa poin penting yang dapat disimpulkan:

1. Pipeline preprocessing memiliki dampak signifikan terhadap performa model, terutama pada recall dan F1-score.
2. SMOTE tidak selalu meningkatkan akurasi, tetapi meningkatkan *fairness* antar kelas.
3. Model terbaik dalam penelitian ini adalah:
    - Random Forest (tanpa SMOTE) untuk tujuan klinis berbasis akurasi dan robustitas.
    - Logistic Regression (dengan atau tanpa SMOTE) untuk interpretabilitas tinggi.
4. Hasil ini memperkuat temuan penelitian sebelumnya bahwa:
    - Model ensemble biasanya unggul pada data klinis tabular.
    - Algoritma yang lebih sederhana seperti Logistic Regression tetap kompetitif ketika preprocessing dilakukan secara optimal [13], [19].

IV. KESIMPULAN

Penelitian ini telah melakukan analisis komprehensif terhadap kinerja berbagai algoritma Machine Learning—Logistic Regression, Random Forest, SVM (RBF), dan XGBoost—dalam memprediksi penyakit jantung menggunakan dataset Cleveland yang tersedia di Kaggle. Eksperimen dilakukan dengan menerapkan pipeline *data preprocessing terintegrasi*, meliputi *data cleaning*, *data transformation*, *data reduction*, serta pengujian pada dua skenario, yaitu tanpa SMOTE dan dengan penerapan SMOTE sebagai metode penyeimbangan kelas.

Berdasarkan hasil pengujian, beberapa kesimpulan dapat dirumuskan sebagai berikut:

1. Pipeline preprocessing memiliki dampak signifikan terhadap performa model. Tahapan seperti normalisasi, encoding, dan reduksi dimensi terbukti meningkatkan stabilitas metrik evaluasi, terutama pada algoritma yang sensitif terhadap distribusi fitur seperti Logistic Regression dan SVM.
2. Random Forest menunjukkan performa terbaik pada skenario tanpa SMOTE, dengan akurasi 0.9016, recall 0.9643, F1-score 0.9000, serta ROC-AUC 0.9594. Hal ini menunjukkan konsistensi Random Forest dalam menangani data tabular medis dengan pola non-linear.
3. Penerapan SMOTE tidak secara signifikan meningkatkan akurasi, namun memberikan efek stabilisasi terhadap recall dan F1-score, yang sangat penting pada prediksi penyakit medis untuk mengurangi risiko kesalahan deteksi (false negative). Dampak ini terlihat khususnya pada SVM dan Logistic Regression.
4. Logistic Regression terbukti menjadi model yang stabil dan kompetitif, baik sebelum maupun sesudah SMOTE. Dengan recall yang tinggi (>0.92 pada kedua skenario), model ini layak direkomendasikan sebagai baseline yang interpretatif dan efisien.
5. SVM (RBF) dan XGBoost tetap memberikan performa yang baik, namun tidak melampaui Random Forest dan Logistic Regression dalam eksperimen ini. Keduanya menunjukkan manfaat moderat dari SMOTE, terutama pada peningkatan pada metrik ROC-AUC.
6. Secara keseluruhan, penelitian ini menegaskan bahwa kombinasi preprocessing terintegrasi dan balancing data yang tepat berperan penting dalam meningkatkan performa prediksi penyakit jantung. Selain itu, efisiensi dan keandalan model sangat dipengaruhi oleh karakteristik dataset dan parameter pipeline yang digunakan.

Dengan demikian, penelitian ini berkontribusi pada pemahaman mengenai *best practices* dalam pengembangan model prediksi penyakit jantung dan dapat dijadikan acuan bagi penelitian lanjutan

maupun implementasi sistem pendukung keputusan klinis berbasis Machine Learning.

DAFTAR PUSTAKA

[1] World Health Organization, "Cardiovascular Diseases," 2021.

[2] S. Benjamin et al., "Heart Disease and Stroke Statistics," Circulation, 2020.

[3] N. Sharma and P. Juneja, "Machine Learning Techniques for Heart Disease Prediction: A Survey," Int. J. Eng. Adv. Technol., vol. 9, no. 6, pp. 600–605, 2020.

[4] Kaggle, "Heart Disease UCI Dataset," 2020.

[5] H. M. Fayaz et al., "Impact of Data Preprocessing on Machine Learning-Based Heart Disease Prediction," Information, vol. 13, no. 10, pp. 475, 2022.

[6] A. S. Reddy and C. K. Reddy, "Feature Engineering and Classification of Heart Disease using Machine Learning Algorithms," Int. J. Recent Technol. Eng., 2019.

[7] M. A. Jabbar, P. Chandra, and B. Srinivasa Rao, "Prediction of Heart Disease Using Multilayer Perceptron and Machine Learning Techniques," J. Data Sci., vol. 18, no. 4, pp. 702–718, 2020.

[8] M. P. N. Chintalapudi et al., "Heart Disease Prediction Using Ensemble and Hybrid Machine Learning Methods," Informatics in Medicine Unlocked, vol. 41, 2023.

[9] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," Proc. KDD, 2019.

[10] R. Kumar and R. Saini, "Dimensionality Reduction Techniques for Medical Data," Biocybernetics and Biomedical Engineering, 2020.

[11] D. Chicco and G. Jurman, "The Advantages of MCC Over F1 Score and Accuracy in Binary Classification," BMC Genomics, 2020.

[12] V. Stodden et al., "Reproducible Computational Research," Annals of Applied Statistics, 2019.

[13] S. Kumari and A. Godara, "Comparative analysis of machine learning techniques for heart disease prediction," *International Journal of Engineering*, vol. 33, no. 7, pp. 1245–1252, 2020.

[14] T. Chen, S. T. Hsu, and M. T. Tsai, "Enhancing medical diagnosis performance using machine learning with data balancing techniques," *IEEE Access*, vol. 8, pp. 168–177, 2020.

[15] M. A. Uddin *et al.*, "A comparative performance analysis of machine learning models for heart disease prediction," *Applied Sciences*, vol. 11, no. 5, pp. 1–14, 2021.

[16] A. Haq *et al.*, "A hybrid intelligent system for predicting heart disease using machine learning techniques," *Journal of Healthcare Engineering*, vol. 2021, pp. 1–14, 2021.

[17] N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2021 (revisited review).

[18] B. S. Rawat and H. Singh, "SVM-based heart disease prediction with feature optimization," *International Journal of Advanced Computer Science*, vol. 12, no. 6, pp. 337–345, 2022.

[19] A. Gupta *et al.*, "Performance analysis of ML classifiers for heart disease prediction on Cleveland dataset," *International Journal of Intelligent Systems*, vol. 38, no. 1, pp. 112–130, 2023.