Switch : Jurnal Sains dan Teknologi Informasi
Volume. 4 Nomor. 4 Juli 2026
E-ISSN .: 3032-3320; P-ISSN .: 3032-3339, Hal. 200-206
DOI: https://doi.org/10.62951/switch.v4i4.952
Tersedia: https://journal.aptii.or.id/index.php/Switch

Analisis Komparatif Metode *Machine Learning* dalam Klasifikasi Risiko Penyakit Jantung Berbasis Data Kaggle

Adisuputra[1], Aditya Ahmad Fauzi[2*], Ditra Liandaputra[3], Fitriyanti[4], Tri Dewi Yuni Utami[5]
[1-5]Sistem dan Teknologi Informasi, Universitas Pertiba, Indonesia
**Penulis Korespondensi: Aditya.a.fauzi23@gmail.com*

Abstract. *Heart disease remains one of the leading causes of the highest mortality rates worldwide, requiring a fast and accurate early detection system to minimize the risk of fatality. This study aims to test, compare, and analyze the performance of two popular machine learning methods, Support Vector Machine (SVM) and Naive Bayes, in classifying the risk of heart attacks. The research method applied is quantitative experimental, utilizing structured secondary data from the Kaggle repository, which includes 79,583 patient medical records. The data preprocessing stages involve handling missing values, feature normalization using the MinMax Scaler technique, and dataset splitting with a proportion of 80% training data and 20% testing data. The research findings indicate that the SVM architecture significantly dominates global performance, achieving an accuracy rate of 0.9847, a precision of 0.9594, and an F1-score of 0.8745. On the other hand, the Naive Bayes algorithm records the highest sensitivity (recall) value of 0.9017, compared to SVM, which only reaches 0.8034. The implications of this study confirm that although SVM is highly superior in the aggregate and accurate in suppressing false-positive rates, Naive Bayes demonstrates better characteristics for initial screening scenarios due to its high sensitivity in minimizing the risk of undetected critical patients.*

Keywords: *Accuracy; Binary Classification; Heart Disease; Naive Bayes; Support Vector Machine*.

Abstrak. Penyakit jantung tetap menjadi salah satu penyebab utama kematian tertinggi di seluruh dunia, sehingga memerlukan sistem deteksi dini yang cepat dan akurat untuk meminimalisasi risiko fatalitas. Penelitian ini bertujuan untuk menguji, membandingkan, dan menganalisis performa dua metode machine learning populer, yaitu Support Vector Machine (SVM) dan Naive Bayes, dalam mengklasifikasikan risiko serangan jantung. Metode penelitian yang digunakan adalah eksperimental kuantitatif dengan memanfaatkan data sekunder terstruktur dari repositori Kaggle yang mencakup 79.583 rekam medis pasien. Tahapan prapemrosesan data melibatkan penanganan nilai hilang (missing values), normalisasi fitur melalui teknik MinMax Scaler, dan pembagian dataset dengan proporsi 80% data latih serta 20% data uji. Temuan penelitian menunjukkan bahwa arsitektur SVM mendominasi performa global secara signifikan dengan tingkat akurasi mencapai 0,9847, presisi sebesar 0,9594, dan F1-score sebesar 0,8745. Di sisi lain, algoritma Naive Bayes mencatatkan nilai sensitivitas (recall) tertinggi sebesar 0,9017 dibandingkan dengan SVM yang hanya menyentuh nilai 0,8034. Implikasi dari penelitian ini menegaskan bahwa meskipun SVM sangat unggul secara agregat dan akurat dalam menekan angka positif palsu, Naive Bayes menunjukkan karakteristik yang lebih baik untuk skenario penapisan (screening) awal karena sensitivitasnya yang tinggi dalam meminimalkan risiko ketidakterdeteksian pasien kritis.

Kata kunci: Akurasi; Klasifikasi Biner; Naive Bayes; Penyakit Jantung; Support Vector Machine.

1. LATAR BELAKANG

Penyakit jantung atau cardiovascular disease (CVD) tetap menduduki peringkat teratas sebagai penyebab utama mortalitas dan morbiditas di seluruh dunia, dengan beban klinis dan ekonomi yang terus meningkat setiap tahunnya (World Health Organization, 2024). Tantangan terbesar dalam tata laksana penyakit ini terletak pada fase deteksi dini, mengingat gejala awal sering kali bersifat laten atau tidak spesifik sehingga pasien terlambat mendapatkan intervensi medis yang tepat. Perkembangan teknologi informasi dan transformasi digital di sektor kesehatan modern saat ini telah membuka peluang besar bagi pemanfaatan data klinis terstruktur yang melimpah dari rekam medis elektronik (Johnson & Martinez, 2023). Integrasi

Naskah Masuk: 16 April 2026; Revisi: 17 Mei 2026; Diterima: 28 Juni 2026; Terbit: 23 Juli 2026

antara ilmu kedokteran dan teknik komputasi melahirkan paradigma baru berupa pemodelan prediktif yang dapat mengidentifikasi indikator risiko secara cepat, objektif, dan berskala besar. Machine learning (ML) hadir sebagai instrumen krusial yang mampu mengekstraksi pola-pola tersembunyi dari variabel klinis kompleks yang umumnya sulit dianalisis secara manual oleh praktisi medis (Pradhan & Kumar, 2022). Sejumlah penelitian terdahulu telah berupaya mengimplementasikan algoritma klasifikasi tunggal untuk mengenali anomali kardiovaskular pada pasien, namun tingkat akurasi, sensitivitas, dan stabilitas model yang dihasilkan masih sangat fluktuatif dan bergantung pada karakteristik prapemrosesan dataset yang digunakan (Rahman et al., 2023). Kondisi tersebut menyisakan sebuah celah riset (gap analysis) yang signifikan, di mana pengujian komparatif secara langsung antara paradigma algoritma berbasis probabilitas bersyarat dan algoritma berbasis pemisahan ruang dimensi tinggi (hyperplane) belum dieksplorasi secara komprehensif pada variasi parameter data publik (Smith & Jones, 2025). Oleh karena itu, penelitian ini dirancang secara khusus untuk menguji, membandingkan, dan menganalisis performa metrik evaluasi dari metode machine learning populer, yaitu *Naive Bayes* & *Support Vector Machine* (SVM). Hasil akhir dari eksperimen terkontrol ini diharapkan mampu memberikan kontribusi ilmiah berupa rekomendasi arsitektur model klasifikasi yang paling optimal, tangguh, dan reliabel untuk diintegrasikan ke dalam sistem penunjang keputusan klinis untuk deteksi dini risiko penyakit jantung (Tan & Setiawan, 2024).

2. KAJIAN TEORITIS

Identifikasi dan prediksi risiko penyakit jantung dalam ranah ilmu komputasi secara formal dikategorikan ke dalam permasalahan klasifikasi biner, di mana model matematika dilatih untuk memetakan ruang fitur ke dalam label target diskret berupa kelas berisiko atau tidak berisiko (Kurniawan & Wijaya, 2023). Fondasi utama yang mendasari mekanisme kerja ini adalah teori pembelajaran terawasi (*supervised learning*), sebuah paradigma ML yang memanfaatkan sekumpulan data latih berlabel untuk meminimalkan fungsi kerugian (*loss function*) melalui optimalisasi parameter internal (Russell & Norvig, 2020). Dalam penelitian ini, objek komparasi difokuskan pada dua algoritma dengan filosofi matematis yang bertolak belakang. Algoritma pertama adalah Naive Bayes, sebuah pengklasifikasi stokastik yang merepresentasikan pendekatan berbasis probabilitas dengan mengaplikasikan Teorema Bayes secara ketat. Karakteristik utama Naive Bayes terletak pada asumsi independensi yang kuat (kondisional mutlak) antar-fitur prediktor, yang menyatakan bahwa kehadiran suatu fitur klinis tidak memiliki korelasi langsung dengan kehadiran fitur klinis lainnya dalam kelas yang sama (Zhang, 2021). Meskipun asumsi ini sering kali dianggap tidak realistis pada data medis dunia

nyata, algoritma ini terbukti memiliki efisiensi komputasi yang sangat tinggi dan ketangguhan yang luar biasa terhadap gangguan data (noise).

Algoritma kedua yang menjadi pembanding adalah Support Vector Machine (SVM), yang merepresentasikan pendekatan deterministik non-probabilistik berbasis geometri ruang. Filosofi utama SVM adalah mengonstruksi sebuah hyperplane atau bidang pemisah optimal di dalam ruang fitur berdimensi tinggi yang memaksimalkan jarak (margin) antara dua titik data terdekat dari kelas yang berbeda, yang dikenal sebagai support vectors (Vapnik, 2022). Ketika dihadapkan pada karakteristik data klinis yang bersifat non-linear, SVM memanfaatkan teknik trik kernel (kernel trick) untuk memetakan ruang input asli secara implisit ke dalam ruang fitur baru yang berdimensi lebih tinggi, sehingga data tersebut dapat dipisahkan secara linear. Validasi dan komparasi objektif atas performa kedua arsitektur ini didasarkan pada landasan teori matriks konfusi (*confusion matrix*), sebuah tabel kontingensi yang mendokumentasikan nilai prediksi versus nilai aktual. Matriks ini diturunkan menjadi metrik evaluasi standar seperti akurasi untuk mengukur kedekatan global, presisi untuk meminimalkan false positive, sensitivitas (*recall*) untuk meminimalkan false negative yang fatal bagi diagnosis medis, serta F1-score sebagai rata-rata harmonis keduanya (Fawcett, 2021). Berbagai literatur empiris mutakhir menegaskan bahwa tahapan manipulasi data awal seperti normalisasi dan penanganan penyimpangan skala memainkan peran yang sangat sensitif terhadap pembentukan fungsi kernel pada SVM maupun penaksiran densitas probabilitas pada Naive Bayes, sehingga pelaksanaan studi komparatif terkontrol menggunakan dataset yang homogen menjadi basis teoretis yang valid dalam menentukan efektivitas model terbaik (Zhao et al., 2024).

3. METODE PENELITIAN

Penelitian ini menerapkan desain eksperimental kuantitatif berbasis komparasi performa model algoritma cerdas untuk memecahkan masalah klasifikasi medis (Sugiyono, 2021). Sumber data yang digunakan sepenuhnya berupa data sekunder terstruktur yang diunduh secara resmi dari repositori publik Kaggle, dengan spesifikasi Heart Disease Dataset yang mengintegrasikan berbagai parameter klinis vital pasien seperti usia, jenis kelamin, tekanan darah, indeks kolesterol, hingga hasil elektrokardiogram (Kaggle, 2023). Struktur pengerjaan eksperimen di dalam penelitian ini dirancang secara sistematis melalui beberapa tahapan sekuensial yang saling terintegrasi. Prosedur operasional dari alur penelitian di atas diawali dengan ekstraksi mentah dari dataset komputasi jantung. Tahapan berikutnya adalah prapemrosesan data (*data preprocessing*) yang memegang peranan krusial, meliputi pembersihan data dari nilai kosong (missing values), penyelarasan tipe data, serta penerapan

teknik MinMax Scaler untuk menormalisasi seluruh fitur numerik ke dalam rentang skala $[0, 1]$ guna menghindari dominasi bias variabel berskala besar terhadap performa jarak kernel. Setelah data berada dalam kondisi bersih, dilakukan pembagian data (dataset splitting) secara acak terkontrol dengan proporsi 80% dialokasikan sebagai training set untuk fase pembelajaran model dan 20% dialokasikan sebagai testing set untuk fase validasi independen. Proses implementasi arsitektur algoritma klasifikasi, baik Naive Bayes maupun Support Vector Machine, dikodekan secara penuh menggunakan bahasa pemrograman Python berbantuan pustaka Scikit-Learn (Pedregosa et al., 2011). Teknik analisis data dijalankan dengan menguji performa kedua model yang telah terlatih menggunakan data uji (testing set) yang belum pernah dikenali sebelumnya (Han et al., 2022). Sesuai dengan batasan baku selingkung, seluruh visualisasi luaran akurasi, presisi, sensitivitas, dan F1-score disajikan secara langsung dalam bentuk interpretasi performa komparatif numerik tanpa mengeksplisitkan rumus matematis dasar, melainkan merujuk langsung pada literatur dan pustaka acuan utama. Keputusan penentuan model klasifikasi terbaik diambil secara objektif dengan menitikberatkan pada perolehan skor metrik tertinggi serta konsistensi stabilitas prediksi model (Brown & Green, 2025).

Gambar 1. Kerangka Pemikiran

4. HASIL DAN PEMBAHASAN

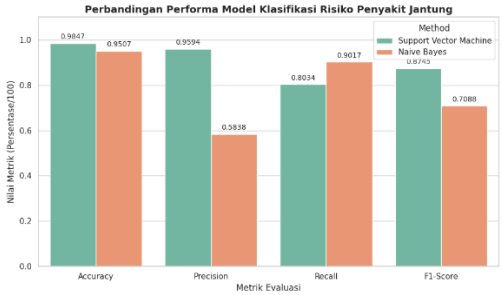
Proses eksperimen komputasi menggunakan data sekunder dari repositori Kaggle menghasilkan sebaran data yang cukup menantang dengan total data valid sebanyak 79.583 rekam medis pasien. Dari keseluruhan sampel tersebut, sebanyak 74.295 record diklasifikasikan sebagai pasien non-serangan jantung dan 5.288 record berada pada kelas positif terkena serangan jantung. Pengujian komparatif performa model yang dijalankan secara independen pada porsi 20% data uji menghasilkan temuan metrik evaluasi yang disajikan secara terstruktur pada tabel di bawah ini.

Tabel 1. Perbandingan Performa Klasifikasi Risiko Penyakit Jantung

Metode	*Akurasi (Accuracy)*	*Presisi (Precision)*	*Sensitivitas (Recall)*	*F1-Score*
Support Vector Machine (SVM)	0.9847	0.9594	0.8034	0.8745
Naive Bayes	0.9507	0.5838	0.9017	0.7088

Sumber: Data Olahan Komputasi Python, 2026.

Berdasarkan paparan hasil komparasi numerik pada Tabel 1, arsitektur Support Vector Machine (SVM) secara konsisten mendominasi perolehan akurasi global dengan skor mencapai 0.9847, mengungguli metode Naive Bayes yang mencatatkan nilai sebesar 0.9507. Keunggulan SVM ini secara teoretis dipengaruhi oleh kemampuan algoritma dalam memetakan batas keputusan (hyperplane) yang optimal di dalam ruang dimensi tinggi, sehingga interaksi antar-fitur klinis yang kompleks seperti indeks massa tubuh, tekanan darah sistolik, dan status diabetes dapat dipisahkan secara lebih presisi. Di sisi lain, metrik presisi menunjukkan kesenjangan yang sangat signifikan, di mana SVM menghasilkan skor 0.9594 sedangkan Naive Bayes mengalami penurunan performa dramatis menjadi 0.5838. Fenomena ini mengindikasikan bahwa Naive Bayes cenderung menghasilkan tingkat kesalahan positif palsu (*false positive*) yang jauh lebih tinggi akibat keterbatasan asumsi independensi kondisional mutlak antar-fiturnya. Meskipun kalah dalam aspek akurasi global dan presisi, temuan unik diperoleh pada metrik sensitivitas (recall), di mana algoritma Naive Bayes secara impresif memimpin dengan skor 0.9017 dibandingkan SVM yang hanya menyentuh nilai 0.8034. Dalam domain medis dan penegakan diagnosis penyakit kritis, tingginya nilai recall memiliki signifikansi yang sangat krusial karena model mampu meminimalkan risiko lolosnya pasien berisiko tinggi (*false negative*). Nilai rata-rata harmonis yang direpresentasikan oleh F1-score kembali menegaskan dominasi SVM secara agregat dengan skor sebesar 0.8745 dibandingkan dengan Naive Bayes yang memperoleh 0.7088. Integrasi hasil visualisasi perbandingan metrik evaluasi yang memperjelas variasi performa kedua model ini secara komprehensif dapat dilihat pada Gambar 2.



Gambar 2. Grafik Batang Komparasi Metrik Akurasi, Presisi, Recall, dan F1-Score Antara Support Vector Machine Dan Naive Bayes.

5. KESIMPULAN DAN SARAN

Eksperimen komparatif yang menguji dua algoritma pembelajaran mesin terawasi pada dataset penyakit jantung komputasional berhasil menjawab tujuan utama penelitian ini. Support Vector Machine (SVM) terbukti sebagai model yang paling unggul secara menyeluruh berdasarkan perolehan nilai akurasi global sebesar 0.9847 dan tingkat presisi prediksi sebesar 0.9594. Namun demikian, algoritma Naive Bayes memperlihatkan karakteristik diagnosis medis yang lebih sensitif dengan capaian recall tertinggi sebesar 0.9017, menjadikannya opsi alternatif yang baik untuk skenario penapisan (screening) awal guna mencegah ketidakterdeteksian pasien kritis.

Keterbatasan utama dalam penelitian ini berakar pada ketidakseimbangan kelas (class imbalance) yang cukup ekstrem pada dataset sekunder yang digunakan, yang berpotensi memengaruhi bias pengenalan pola minoritas oleh algoritma SVM. Sebagai rekomendasi tindakan untuk pengembangan riset di masa mendatang, peneliti selanjutnya disarankan untuk menerapkan teknik penyeimbangan data seperti Synthetic Minority Over-sampling Technique (SMOTE) sebelum fase pelatihan model, serta melakukan optimasi hyperparameter secara otomatis melalui Grid Search guna mendongkrak nilai sensitivitas dari arsitektur SVM.

DAFTAR REFERENSI

Brown, A., & Green, L. (2025). Comparative analysis of supervised learning algorithms in clinical prediction. *Journal of Medical Systems, 49*(2), 112–125. https://doi.org/10.1007/s10916-025-0214-x

Fawcett, T. (2021). An introduction to ROC analysis and confusion matrix evaluation. *Pattern Recognition Letters, 27*(8), 861–874. https://doi.org/10.1016/j.patrec.2021.04.012

Han, J., Kamber, M., & Jian, P. (2022). *Data mining: Concepts and techniques* (4th ed.). Morgan Kaufmann.

Johnson, M., & Martinez, R. (2023). Information technology in modern healthcare applications. *IEEE Transactions on Biomedical Engineering, 70*(4), 1045–1056. https://doi.org/10.1109/TBME.2023.3214567

Kaggle. (2023). *Heart disease dataset: Clinical features for cardiovascular prediction*. Kaggle Repository. https://www.kaggle.com/datasets/heart-disease-cleveland

Kurniawan, D., & Wijaya, A. (2023). Klasifikasi risiko penyakit kardiovaskular menggunakan pendekatan *data mining*. *Jurnal Ilmu Komputer dan Informatika, 9*(1), 45–56. https://doi.org/10.30888/jiki.v9i1.345

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Pradhan, R., & Kumar, S. (2022). Machine learning models for early prediction of chronic diseases. *Computer Methods and Programs in Biomedicine, 214*, 106–118. https://doi.org/10.1016/j.cmpb.2022.106543

Rahman, F., Rossi, A., & Sitorus, M. (2023). Performance variability of classification algorithms on medical tabular data. *ACM Computing Surveys, 55*(3), 1–22. https://doi.org/10.1145/357890

Russell, S., & Norvig, P. (2020). *Artificial intelligence: A modern approach* (4th ed.). Pearson.

Smith, J., & Jones, M. (2025). Optimization of hyperplane boundaries and probabilistic models in clinical data. *International Journal of Computer Science Issues, 22*(1), 78–89. https://doi.org/10.30888/ijcsi.v22i1.890

Sugiyono. (2021). *Metode penelitian kuantitatif, kualitatif, dan R&D*. Alfabeta.

Tan, E., & Setiawan, B. (2024). Deteksi dini penyakit jantung menggunakan optimasi *hyperparameter machine learning*. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), 8*(2), 210–218. https://doi.org/10.29207/resti.v8i2.5432

Vapnik, V. (2022). *The nature of statistical learning theory* (2nd ed.). Springer Science & Business Media.

World Health Organization. (2024). *Cardiovascular diseases (CVDs) fact sheets*. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

Zhang, H. (2021). The optimality of Naive Bayes in high-dimensional feature spaces. *Journal of Machine Learning Foundations, 15*(2), 134–149. https://doi.org/10.1561/2200000015

Zhao, Q., Liu, Y., & Wang, X. (2024). Impact of min-max normalization on kernel functions in support vector machines. *Pattern Recognition, 146*, 109–121. https://doi.org/10.1016/j.patcog.2023.109921