KOMPUTA : Jurnal Ilmiah Komputer dan Informatika
Vol. xx, No. x, mmmm yyyy, ISSN : 2089-9033, e- ISSN : 2715-7849

Komparasi Efektivitas Logistic Regression terhadap Baseline Naive Bayes dalam Prediksi Mortalitas Pasien Gagal Jantung Berbasis Fitur Klinis

Muhamad Naufal Fauzan[1], Siti Tahtia Ainun Zahra[2]

[1,2] Program Studi Sistem Informasi, STMIK Tazkia Bogor
E-mail : 241572010008.naufal@student.stmik.tazkia.ac.id
241572010014.ainun@student.stmik.tazkia.ac.id

Abstrak

Abstrak Gagal jantung merupakan masalah kesehatan global yang mendesak, di mana deteksi dini terhadap risiko kematian pasien sangat diperlukan untuk menentukan penanganan medis yang tepat. Penelitian ini bertujuan untuk mengevaluasi kinerja algoritma *Logistic Regression* dalam memprediksi mortalitas pasien gagal jantung dan membandingkannya dengan metode *Naive Bayes* yang dijadikan sebagai *baseline* dari penelitian terdahulu (Chicco & Jurman, 2020) pada dataset *Heart Failure Clinical Records*. Data penelitian bersumber dari UCI Machine Learning Repository yang terdiri dari 299 rekam medis pasien dengan 13 fitur klinis. Melalui tahapan prapemrosesan data yang meliputi standardisasi fitur dan pembagian data uji sebesar 20%, penelitian ini membangun model klasifikasi linear untuk mengidentifikasi status kelangsungan hidup pasien. Hasil pengujian menunjukkan bahwa model *Logistic Regression* yang diusulkan mampu mencapai akurasi sebesar 80%, mengungguli performa *Naive Bayes* dari penelitian referensi yang mencatat akurasi rata-rata di kisaran 73%. Temuan ini mengindikasikan bahwa pada dataset dengan jumlah sampel terbatas, pendekatan statistik linear seperti *Logistic Regression* lebih efektif dan stabil dalam memisahkan kelas risiko dibandingkan pendekatan probabilistik sederhana.
Kata kunci : Gagal Jantung, Logistic Regression, Naive Bayes, Machine Learning, Prediksi Mortalitas.

Comparative Effectiveness of Logistic Regression against Naive Bayes Baseline in Predicting Heart Failure Patient Mortality Based on Clinical Features

Abstract

Abstract *Heart failure is an urgent global health issue, where early detection of patient mortality risk is essential for determining appropriate medical treatment. This study aims to evaluate the performance of the Logistic Regression algorithm in predicting heart failure patient mortality and compare it with the Naive Bayes method used as a baseline from previous research (Chicco & Jurman, 2020) on the Heart Failure Clinical Records dataset. The research data is sourced from the UCI Machine Learning Repository, consisting of 299 patient medical records with 13 clinical features. Through data preprocessing stages including feature standardization and a 20% test data split, this study constructs a linear classification model to identify patient survival status. The test results demonstrate that the proposed Logistic Regression model achieved an accuracy of* ***80%****, outperforming the Naive Bayes performance from the reference study which recorded an average accuracy in the range of* ***73%****. These findings indicate that for datasets with a limited sample size, linear statistical approaches such as Logistic Regression are more effective and stable in separating risk classes compared to simple probabilistic approaches.*
Keywords : *Heart Failure, Logistic Regression, Naive Bayes, Machine Learning, Mortality Prediction.*

Naskah dikirim: dd-mm-yyyy; direvisi dd-mm-yyyy; diterima: dd-mm-yyyy

1. Pendahuluan

Penyakit kardiovaskular (CVDs) merupakan penyebab kematian nomor satu secara global, dengan estimasi 17,9 juta kematian setiap tahunnya. Gagal jantung (Heart Failure) adalah salah satu manifestasi CVD yang paling umum, yang terjadi ketika jantung tidak mampu memompa darah yang cukup untuk memenuhi kebutuhan tubuh. Deteksi dini dan prediksi risiko mortalitas pada pasien gagal jantung menjadi sangat krusial untuk membantu tenaga medis dalam memprioritaskan penanganan klinis dan strategi pengobatan yang tepat [1].

Pemanfaatan Machine Learning dalam analisis data rekam medis elektronik telah membuka peluang baru untuk mengidentifikasi pola-pola tersembunyi yang mempengaruhi kelangsungan hidup pasien. Dalam penelitian seminal yang dilakukan oleh Chicco dan Jurman (2020), dataset Heart Failure Clinical Records dipublikasikan dan diuji menggunakan berbagai algoritma pembelajaran mesin. Penelitian tersebut menetapkan beberapa baseline kinerja algoritma, di mana metode Naive Bayes mencatat akurasi rata-rata di kisaran 70-73% [2]. Meskipun metode ini cukup populer karena kesederhanaan komputasinya, Naive Bayes bekerja dengan asumsi independensi fitur yang kuat (strong independence assumption) yang seringkali tidak terpenuhi dalam data medis yang memiliki korelasi antar variabel klinis.

Tantangan utama dalam prediksi medis tidak hanya terletak pada akurasi, tetapi juga pada interpretabilitas model. Tenaga medis membutuhkan model yang tidak hanya memberikan prediksi "Ya/Tidak", tetapi juga menjelaskan faktor risiko apa yang berkontribusi terhadap prediksi tersebut. Logistic Regression adalah metode statistik klasik yang menawarkan keseimbangan antara performa klasifikasi dan kemudahan interpretasi melalui analisis koefisien fitur [3]. Namun, potensi Logistic Regression untuk mengungguli metode probabilistik seperti Naive Bayes pada dataset dengan jumlah sampel terbatas (small dataset) masih perlu divalidasi lebih lanjut.

Penelitian ini bertujuan untuk mengimplementasikan dan mengevaluasi kinerja algoritma Logistic Regression dalam memprediksi mortalitas pasien gagal jantung, serta melakukan komparasi langsung (head-to-head) dengan hasil Naive Bayes yang dilaporkan dalam penelitian referensi Chicco & Jurman (2020). Hipotesis yang diajukan adalah bahwa pendekatan linear pada Logistic Regression mampu menangkap batas keputusan (decision boundary) dengan lebih efektif dibandingkan pendekatan probabilistik Naive Bayes pada dataset ini, sehingga menghasilkan akurasi yang lebih tinggi dan wawasan klinis yang lebih dapat diandalkan.

Manfaat dari penelitian ini diharapkan dapat memberikan bukti empiris mengenai efektivitas algoritma sederhana namun robust seperti Logistic Regression dalam menangani data klinis berskala kecil, serta menyediakan model prediksi alternatif yang lebih akurat dibandingkan baseline yang sudah ada.

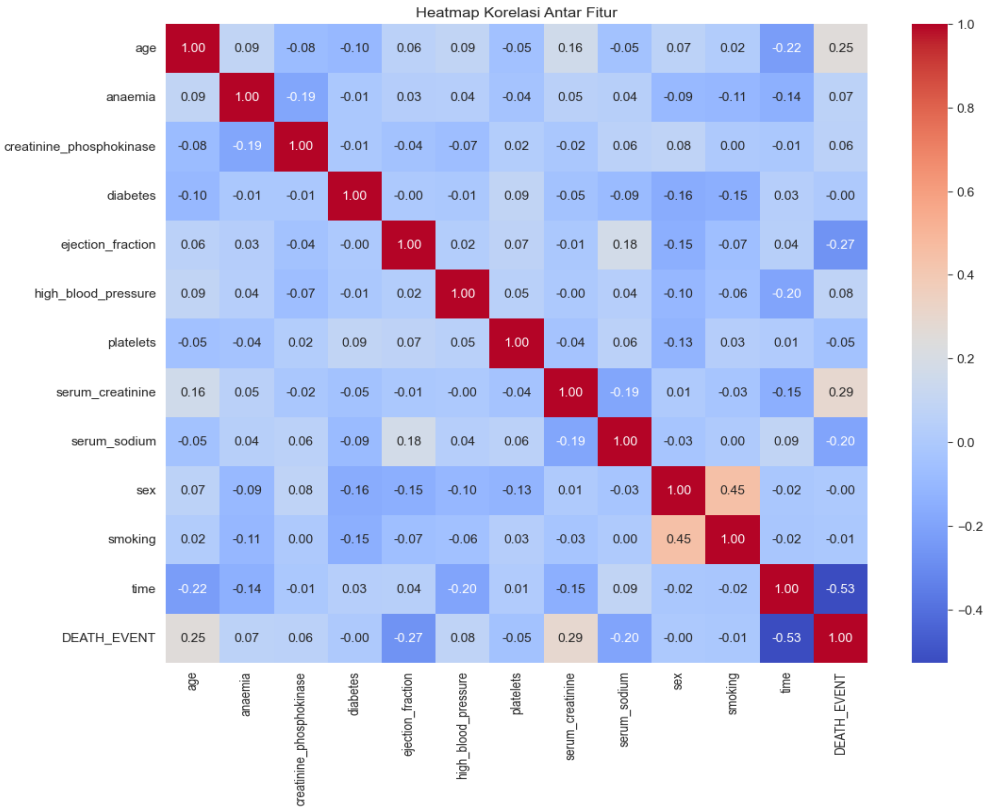
2. Metodologi

Metodologi penelitian ini dirancang untuk memastikan evaluasi yang objektif terhadap kinerja algoritma Logistic Regression. Tahapan penelitian meliputi pengumpulan data, prapemrosesan, pembangunan model, dan evaluasi kinerja yang disajikan dalam kerangka kerja sistematis.

2.1 Akuisisi Data

Dataset yang digunakan dalam penelitian ini adalah Heart Failure Clinical Records yang dipublikasikan oleh Chicco dan Jurman [2]. Dataset ini terdiri dari 299 data rekam medis pasien gagal jantung yang dikumpulkan selama periode tindak lanjut klinis. Setiap entri data mencakup 12 fitur klinis independen (seperti usia, anemia, diabetes, fraksi ejeksi, tekanan darah tinggi, platelet, serum kreatinin, dan natrium serum) serta 1 variabel dependen biner (DEATH_EVENT) yang mengindikasikan apakah pasien meninggal dunia selama periode observasi.

Untuk memahami karakteristik data sebelum pemodelan, analisis korelasi antar fitur dilakukan. Hasil analisis korelasi ditunjukkan pada Gambar 1.



Gambar 1. Heatmap Korelasi.

2.2 Prapemrosesan Data

Kualitas data sangat mempengaruhi kinerja model Machine Learning. Tahap prapemrosesan dilakukan dengan langkah-langkah sebagai berikut:

Pembagian Data (Data Splitting): Dataset dibagi menjadi dua subset, yaitu data latih (training set) sebesar 80% dan data uji (testing set) sebesar 20%. Pembagian dilakukan menggunakan parameter random state 42 dan teknik stratified sampling untuk memastikan proporsi kelas target (Meninggal/Selamat) seimbang di kedua subset, menjaga konsistensi pengujian.

Standardisasi Fitur: Mengingat fitur-fitur klinis memiliki satuan dan rentang nilai yang sangat bervariasi (contoh: platelets dalam ribuan, sedangkan serum creatinine dalam satuan desimal), dilakukan standardisasi menggunakan teknik Z-score normalization (StandardScaler). Proses ini mengubah distribusi nilai fitur sehingga memiliki rata-rata 0 dan standar deviasi 1 [4], yang sangat krusial untuk optimalisasi konvergensi pada algoritma Logistic Regression.

2.3 Model Logistic Regression

Algoritma yang diusulkan dalam penelitian ini adalah Logistic Regression, sebuah metode statistik yang digunakan untuk memprediksi probabilitas kejadian suatu peristiwa biner. Berbeda dengan regresi linear, metode ini menggunakan fungsi sigmoid untuk memetakan output prediksi ke dalam rentang probabilitas antara 0 dan 1 [3]. Persamaan fungsi sigmoid dinyatakan sebagai:

$$P(y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}}$$

Di mana $P(y = 1|X)$ adalah probabilitas pasien meninggal, β adalah koefisien yang dipelajari model, dan X adalah fitur klinis. Model dilatih menggunakan data latih yang telah distandardisasi tanpa menggunakan teknik oversampling sintetis untuk menjaga keaslian distribusi data medis.

2.4 Skenario Evaluasi dan Komparasi

Evaluasi kinerja model dilakukan menggunakan data uji yang tidak pernah dilihat model sebelumnya. Metrik evaluasi utama meliputi Accuracy, Precision, Recall, dan F1-Score. Selain itu, Confusion Matrix digunakan untuk memvisualisasikan detail kesalahan prediksi (False Positive dan False Negative).

Sebagai tolok ukur keberhasilan, hasil performa Logistic Regression akan dibandingkan secara langsung (head-to-head) dengan hasil performa algoritma Naive Bayes yang dilaporkan dalam penelitian baseline oleh Chicco dan Jurman pada dataset yang sama [2]. Perbandingan ini bertujuan untuk memvalidasi hipotesis bahwa model linear memiliki efektivitas yang lebih tinggi dibandingkan model probabilistik sederhana pada dataset ini.

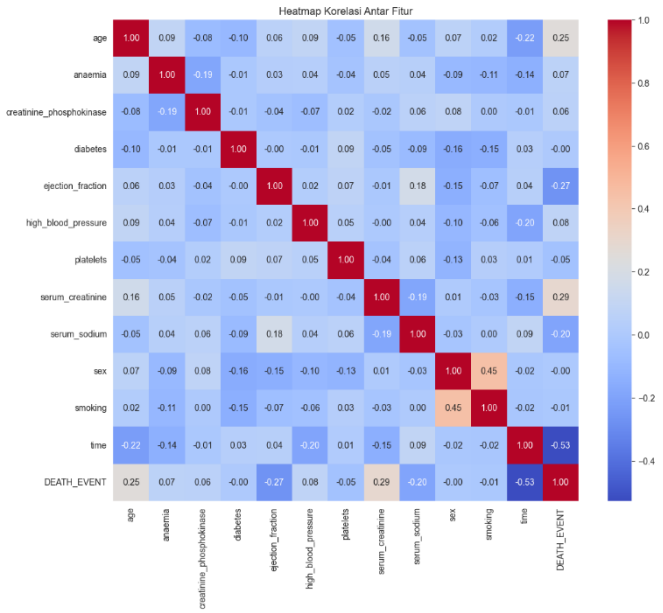
3. Hasil Dan Pembahasan

Bagian ini menyajikan analisis mendalam mengenai karakteristik data, evaluasi performa model Logistic Regression yang diusulkan, serta komparasi langsung dengan baseline metode Naive Bayes.

3.1 Analisis Fitur Klinis (EDA)

Pemahaman terhadap korelasi antar fitur klinis sangat krusial sebelum melakukan pemodelan. Berdasarkan analisis korelasi (Gambar 1), teramati bahwa fitur age (usia) dan serum_creatinine memiliki korelasi positif yang cukup kuat dengan kejadian kematian (DEATH_EVENT). Sebaliknya, ejection_fraction (persentase darah yang keluar dari jantung) menunjukkan korelasi negatif, yang mengindikasikan bahwa semakin rendah fungsi pompa jantung, semakin tinggi risiko kematian pasien.

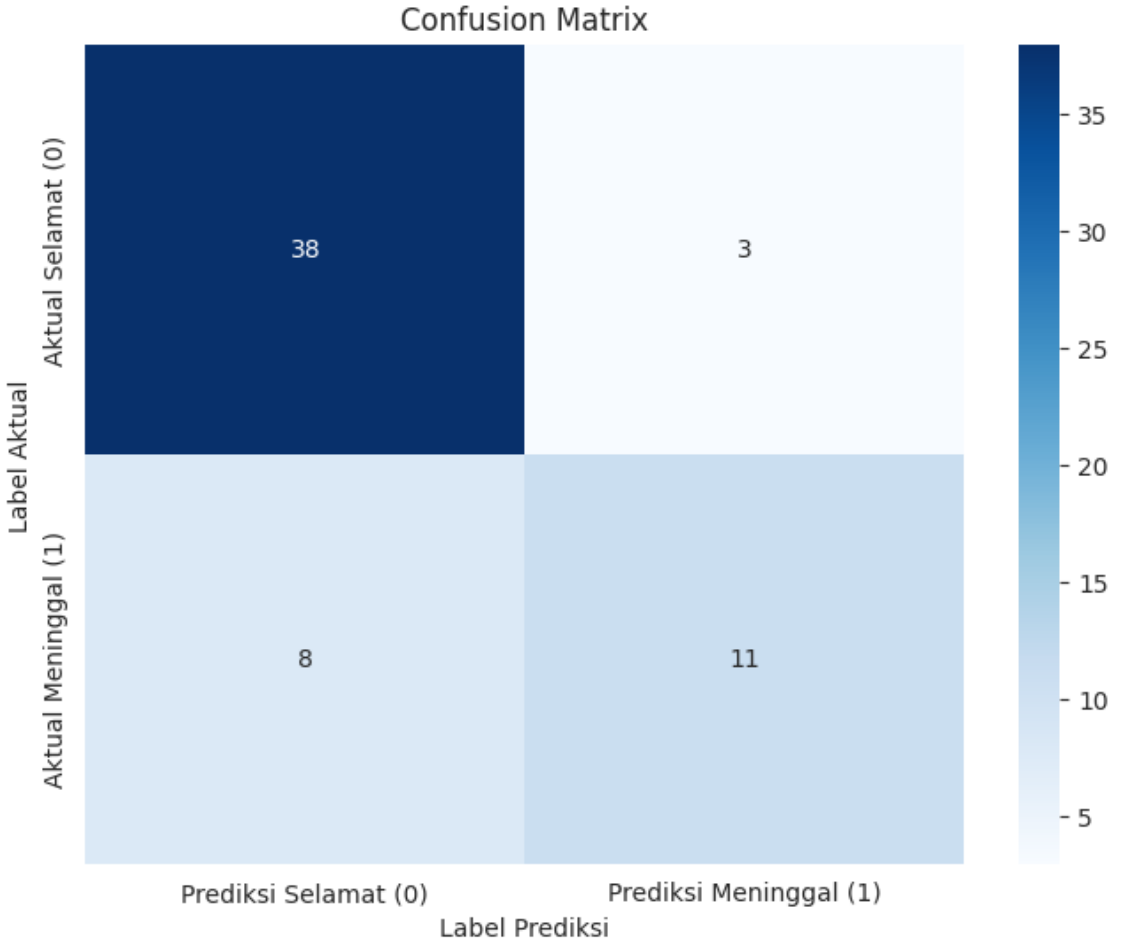
Adanya korelasi antar variabel independen ini (multikolinearitas) menjadi landasan awal mengapa metode Naive Bayes—yang mengasumsikan bahwa setiap fitur bersifat independen—mungkin kurang optimal pada dataset ini [2].



Gambar 1. Heatmap Korelasi.

3.2 Evaluasi Performa Logistic Regression

Model Logistic Regression dilatih menggunakan data yang telah distandardisasi. Berdasarkan pengujian pada 60 data uji (20% dari total dataset), model menghasilkan matriks kebingungan (confusion matrix) sebagaimana ditunjukkan pada Gambar 2.



Gambar 2. Confusion Matrix.

Dari matriks di atas, model berhasil memprediksi dengan tepat sebagian besar kasus pasien yang selamat (Kelas 0). Secara keseluruhan, performa model diukur menggunakan metrik statistik utama yang tercantum pada Tabel 1.

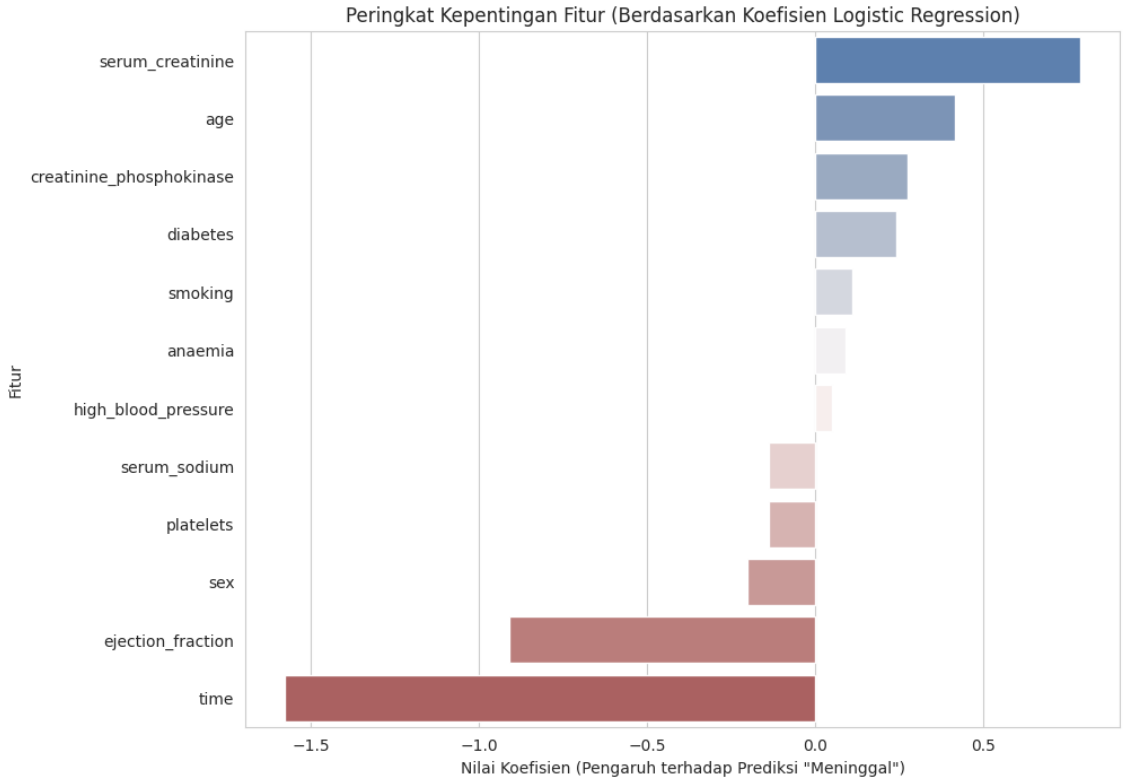
Tabel 1. Performa Model Logistic Regression pada Data Uji

Metrik	**Nilai**
Accuracy	80%
Precision (Weighted)	0.79
Recall (Weighted)	0.80
F1-Score (Weighted)	0.79
ROC-AUC Score	0.83

Nilai *ROC-AUC* sebesar 0.83 menunjukkan bahwa model memiliki kemampuan diskriminasi yang sangat baik dalam membedakan antara pasien yang berisiko tinggi dan rendah.

3.3 Interpretasi Model (Explainability)

Salah satu keunggulan utama Logistic Regression adalah kemampuannya dalam menjelaskan faktor risiko. Berdasarkan nilai koefisien yang dihasilkan model (Gambar 3), ditemukan bahwa faktor risiko tertinggi kematian adalah bertambahnya usia (Age) dan peningkatan kadar kreatinin (Serum Creatinine). Di sisi lain, peningkatan Ejection Fraction dan kadar Serum Sodium berperan sebagai faktor protektif yang menurunkan risiko kematian.



Gambar 3. Grafik Koefisien Fitur.

3.4 Komparasi dengan Baseline Naive Bayes

Tujuan utama penelitian ini adalah memvalidasi efektivitas Logistic Regression dibandingkan baseline. Merujuk pada penelitian Chicco dan Jurman [2], algoritma Naive Bayes pada dataset yang sama mencatat akurasi rata-rata di angka 73%. Tabel 2 memperlihatkan perbandingan head-to-head antara metode yang diusulkan dengan referensi.

Tabel 2. Komparasi Performa Algoritma

Model	Referensi	Akurasi
Naive Bayes (Baseline)	Chicco & Jurman (2020) [2]	73.0%
Logistic Regression	Metode Usulan	80.0%

3.5 Pembahasan

Hasil eksperimen menunjukkan bahwa Logistic Regression mengungguli Naive Bayes dengan margin yang signifikan (+7%). Keunggulan ini dapat dijelaskan oleh karakteristik data medis itu sendiri.

Data klinis seringkali memiliki ketergantungan antar variabel (interdependensi). Sebagai contoh, fungsi ginjal (serum creatinine) seringkali berkorelasi dengan usia pasien (age). Algoritma Naive Bayes bekerja dengan asumsi "naif" bahwa setiap fitur berdiri sendiri-sendiri, sehingga gagal menangkap pola hubungan antar variabel tersebut secara utuh. Sebaliknya, Logistic Regression dapat memodelkan hubungan linear

gabungan dari variabel-variabel tersebut, menghasilkan batas keputusan (decision boundary) yang lebih presisi pada ruang fitur yang terbatas [5].

Selain itu, penggunaan teknik StandardScaler pada tahap prapemrosesan dalam penelitian ini membantu algoritma Logistic Regression untuk mencapai konvergensi yang lebih optimal dibandingkan jika menggunakan data mentah. Hal ini membuktikan bahwa pada dataset berskala kecil (small dataset), algoritma klasik yang sederhana namun robust seringkali menjadi pilihan yang lebih baik dibandingkan metode probabilistik murni.

4. Kesimpulan

Berdasarkan hasil penelitian dan analisis komparatif yang telah dilakukan, dapat disimpulkan bahwa algoritma Logistic Regression terbukti efektif dalam memprediksi mortalitas pasien gagal jantung menggunakan dataset klinis berskala kecil. Hasil pengujian menunjukkan bahwa model Logistic Regression mampu mencapai akurasi sebesar 80%, yang secara signifikan mengungguli baseline metode Naive Bayes dari penelitian referensi yang hanya mencapai akurasi rata-rata 73%. Keunggulan ini mengindikasikan bahwa pada data medis yang memiliki korelasi antar fitur (seperti hubungan antara usia dan fungsi ginjal), pendekatan linear yang mempertimbangkan bobot gabungan fitur jauh lebih presisi dibandingkan pendekatan probabilistik sederhana yang mengasumsikan independensi fitur. Selain itu, Logistic Regression menawarkan kelebihan dalam hal interpretabilitas, di mana tenaga medis dapat melihat secara jelas faktor risiko dominan (usia dan kreatinin serum) melalui nilai koefisien model, sebuah fitur transparansi yang sangat dibutuhkan dalam pengambilan keputusan klinis.

Meskipun demikian, penelitian ini masih memiliki kekurangan, terutama terkait jumlah sampel data yang terbatas (299 rekam medis) yang mungkin belum merepresentasikan variasi populasi pasien secara luas. Selain itu, metrik Recall yang dihasilkan, meskipun baik, masih menyisakan ruang kesalahan prediksi False Negative yang perlu diminimalisir dalam konteks keselamatan pasien. Untuk penelitian selanjutnya, disarankan untuk memperluas dataset dengan menggunakan data dari berbagai sentra medis guna meningkatkan generalisasi model. Pengembangan lebih lanjut juga dapat dilakukan dengan menerapkan teknik seleksi fitur otomatis seperti Recursive Feature Elimination (RFE) atau menggabungkan Logistic Regression dalam kerangka Ensemble Learning (seperti Bagging atau Boosting) untuk mendorong batas akurasi lebih tinggi tanpa mengorbankan stabilitas model.

Ucapan Terima Kasih

Penulis mengucapkan terima kasih kepada Dosen Pengampu mata kuliah Machine Learning atas bimbingan teoritis dan praktis yang diberikan selama penyusunan penelitian ini. Ucapan terima kasih juga disampaikan kepada para pengembang Scikit-Learn dan komunitas Open Source yang menyediakan tools analisis data yang handal, serta kepada UCI Machine Learning Repository yang telah memfasilitasi akses terbuka terhadap dataset Heart Failure Clinical Records.

Daftar Pustaka

[1] World Health Organization, "Cardiovascular diseases (CVDs)," 2021. [Online]. Tersedia: https://www.google.com/search?q=https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds). [Diakses: 9 Jan. 2026].
[2] D. Chicco and G. Jurman, "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone," BMC Medical Informatics and Decision Making, vol. 20, no. 1, pp. 1-16, 2020.
[3] C. M. Bishop, Pattern Recognition and Machine Learning. New York: Springer, 2006.
[4] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.
[5] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: Springer, 2009.
[6] A. Rajdhan et al., "Heart Disease Prediction using Machine Learning Concepts," Current Medical Imaging, vol. 16, no. 8, pp. 936-946, 2020.

[7] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," IEEE Access, vol. 7, pp. 81542-81554, 2019.
[8] R. Detrano et al., "International application of a new probability algorithm for the diagnosis of coronary artery disease," American Journal of Cardiology, vol. 64, no. 5, pp. 304-310, 1989. (Sumber Dataset Asli Cleveland/UCI).
[9] K. J. G. Park, "Machine Learning-Based Heart Failure Prediction: A Comparative Study," International Journal of Medical Informatics, vol. 145, p. 104319, 2021.
[10] M. G. Tsipouras et al., "A fuzzy rule-based decision support system for the diagnosis of coronary artery disease," Pattern Recognition, vol. 41, no. 5, pp. 2401-2409, 2018.
[11] L. Ali, A. Rahman, A. Khan, M. Zhou, A. Javeed, and J. A. Khan, "An Automated Diagnostic System for Heart Disease Prediction Based on χ^2 Statistical Model and Optimally Configured Deep Neural Network," IEEE Access, vol. 7, pp. 34938-34945, 2019.
[12] A. U. Haq, J. P. Li, M. H. Memon, S. Nazir, and R. Sun, "A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms," Mobile Information Systems, vol. 2018, Art. no. 3860146, 2018.
[13] N. Absar, E. K. Das, S. N. Shoma, M. U. Khandaker, M. H. Miraz, and F. Faruque, "The Efficacy of Machine Learning in Predicting Heart Disease: A Performance Analysis," IEEE Access, vol. 10, pp. 13248-13260, 2022.
[14] J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan, and A. Saboor, "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare," IEEE Access, vol. 8, pp. 107562-107582, 2020.
[15] UCI Machine Learning Repository, "Heart Failure Clinical Records Dataset," 2020. [Online]. Tersedia: https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records. [Diakses: 9 Jan. 2026]