Prediksi Gagal Jantung Berbasis *Machine learning* Menggunakan *Support Vector Machine* dan Regresi Logistik

Kadek Adi Hendrawan[1], Awaldi Rizki[2], Dody Kristianto Sinaga[3], Herman RuswanSuwarman[4]
Departemen Teknik Industri/Fakultas Industri Kreatif/Program Studi Teknik Industri
Universitas Teknologi Bandung
Kota Bandung, Indonesia
e-mail: [1]adilog09@gmail.com, [2]awaldirizki440@gmail.com, [3]dodykristiantosinaga@gmail.com, [4]herman@utb-univ.ac.id

Correspondence : *e-mail*: herman@utb-univ.ac.id

Diajukan: 22 Agustus 2024; Direvisi: 22 Agustus 2024; Diterima: 29 Agustus 2024

Abstrak

Penyakit kardiovaskular (CVD) merupakan penyebab utama kematian global, dan gagal jantung menjadi masalah kesehatan masyarakat yang krusial. Deteksi dini sangat penting untuk mengurangi risiko kematian dini dan meningkatkan kualitas hidup pasien. Penelitian ini mengembangkan model prediksi gagal jantung menggunakan algoritma machine learning*, yaitu Regresi Logistik dan Support Vector Machine (SVM), untuk meningkatkan akurasi diagnosis. Dataset yang digunakan mencakup 918 observasi dengan 11 fitur klinis. Hasil penelitian menunjukkan bahwa Regresi Logistik memiliki akurasi 82% dengan precision 0.80, recall 0.85, dan F1-Score 0.82, dengan rata-rata probabilitas prediksi 50%. Sebaliknya, SVM mencapai akurasi 85%, precision 0.82, recall 0.88, dan F1-Score 0.85. Hasil menunjukan bahwa kedua model ini mampu digunakan dalam proses prediksi. Penelitian ini memberikan wawasan tentang penerapan* machine learning *dalam deteksi dini gagal jantung dan meningkatkan kualitas perawatan kesehatan.*

Kata kunci: *Gagal Jantung, Machine learning, Regresi Logistik, dan SVM.*

Abstract

Cardiovascular diseases (CVD) are the leading cause of global mortality, with heart failure representing a critical public health issue. Early detection is crucial to reduce the risk of premature death and improve patient quality of life. This study developed heart failure prediction models using machine learning algorithms, specifically Logistic Regression and Support Vector Machine (SVM), to enhance diagnostic accuracy. The dataset used includes 918 observations with 11 clinical features. Results indicate that Logistic Regression achieved an accuracy of 82% with precision 0.80, recall 0.85, and F1-Score 0.82, with an average prediction probability of 50%. In contrast, SVM reached an accuracy of 85%, precision 0.82, recall 0.88, and F1-Score 0.85. The findings demonstrate that both models are viable for prediction processes. This research provides insights into the application of machine learning in early heart failure detection and contributes to improving healthcare quality.

Keywords*: Heart Failure,* Machine learning*, Logistic Regression, and SVM.*

1. Pendahuluan

Penyakit kardiovaskular (*Cardiovascular Diseases*/CVDs) adalah penyebab utama kematian di seluruh dunia, menyumbang sekitar 31% dari semua kematian global dengan sekitar 17,9 juta kematian setiap tahun. Sebagian besar kematian ini disebabkan oleh serangan jantung dan stroke, dengan sepertiga dari kematian tersebut terjadi pada individu di bawah usia 70 tahun [1]. Gagal jantung, sebagai salah satu konsekuensi umum dari CVD, menjadi masalah kesehatan masyarakat yang kritis [2]. Deteksi dini dan manajemen yang tepat bagi individu dengan risiko tinggi terhadap penyakit kardiovaskular sangat penting untuk mencegah kematian dini dan meningkatkan kualitas hidup pasien [3]. Namun, gejala penyakit ini sering kali tidak tampak jelas, menyebabkan keterlambatan dalam intervensi medis yang diperlukan [3].

Dalam menghadapi tantangan ini, teknologi machine learning menawarkan solusi yang menjanjikan untuk meningkatkan akurasi prediksi dan diagnosis gagal jantung. Dengan kemampuannya untuk menganalisis dataset besar dan kompleks, *machine learning* dapat mengidentifikasi pola yang tidak terlihat

oleh metode konvensional dan membantu dalam penilaian risiko penyakit jantung dengan lebih tepat [4]. Penelitian ini berfokus pada pengembangan model prediksi penyakit jantung menggunakan dataset yang mencakup 918 observasi dari lima dataset jantung yang berbeda, dengan 11 fitur penting seperti usia, jenis kelamin, tipe nyeri dada, tekanan darah saat istirahat, dan kadar kolesterol [5].

Tujuan dari penelitian ini adalah untuk menganalisis kemungkinan terjadinya penyakit jantung berdasarkan atribut klinis yang tersedia dan memodelkan algoritma machine learning untuk memprediksi risiko penyakit jantung dengan akurasi tinggi. Dengan menggunakan algoritma seperti Regresi Logistik dan *Support Vector Machine* (SVM), penelitian ini bertujuan untuk memberikan solusi yang dapat diandalkan dalam deteksi dini dan manajemen gagal jantung, serta menawarkan wawasan berharga mengenai penerapan machine learning dalam praktik medis. Hasil dari penelitian ini diharapkan dapat berkontribusi pada peningkatan kualitas perawatan kesehatan dan pengambilan keputusan medis yang lebih informatif.

2. Metode Penelitian

Penelitian ini bertujuan untuk mengembangkan sebuah model prediksi gagal jantung menggunakan algoritma *machine learning*, khususnya Regresi Logistik dan *Support Vector Machine* (SVM). Proses penelitian dimulai dengan pengumpulan data medis dari sumber yang dapat dipercaya, meliputi atribut klinis seperti usia, jenis kelamin, tekanan darah, dan tingkat kolesterol, yang penting untuk menilai risiko gagal jantung.

Pada tahap pra-pemrosesan, berbagai langkah diambil untuk meningkatkan kualitas data yang akan digunakan dalam model. Data yang tidak lengkap ditangani dengan metode yang sesuai, data numerik dinormalisasi, dan variabel kategori dikonversi menjadi format numerik. Selain itu, dilakukan analisis eksploratif untuk memahami karakteristik data dan mengidentifikasi variabel kunci yang akan dimasukkan dalam model prediksi.

Setelah data siap, model prediksi gagal jantung dikembangkan dengan menggunakan algoritma Regresi Logistik dan SVM. Data dibagi menjadi dua set, yaitu set pelatihan dan set pengujian, untuk melatih dan menguji model. Regresi Logistik digunakan untuk memprediksi kemungkinan terjadinya gagal jantung, sementara SVM digunakan untuk klasifikasi yang lebih kompleks. Proses tuning *hyperparameter* dilakukan untuk mengoptimalkan performa model [6][7].

Setelah model selesai dibangun, diuji dengan data pengujian untuk menilai performa dalam kondisi nyata. Validasi akhir dilakukan dengan membandingkan hasil prediksi model dengan diagnosa medis aktual untuk menilai efektivitas model dalam praktik klinis.

Dataset yang digunakan bersumber dari Kaggle dan berisi data medis terkait gagal jantung, termasuk 918 observasi dengan 11 fitur klinis seperti usia, jenis kelamin, tipe nyeri dada, tekanan darah saat istirahat, dan kadar kolesterol. Dataset ini digunakan untuk mengembangkan model prediksi gagal jantung menggunakan algoritma machine learning seperti Regresi Logistik dan *Support Vector Machine* (SVM). Data ini membantu dalam analisis dan peningkatan akurasi prediksi serta diagnosis gagal jantung.

2.1. Analisis Korelasi (Koefisien Korelasi Pearson)

Koefisien Korelasi Pearson mengukur kekuatan dan arah hubungan linear antara dua variabel [8]. Rumusnya adalah:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

Di mana:
- (r) adalah koefisien korelasi *Pearson*.
- (n) adalah jumlah pasangan data.
- $(\sum xy)$ adalah jumlah hasil perkalian antara (x) dan (y).
- $(\sum x)$ dan $(\sum y)$ adalah jumlah dari (x) dan (y) masing-masing.
- $(\sum x^2)$ dan $(\sum y^2)$ adalah jumlah kuadrat dari (x) dan (y).

2.2. Linearitas (Uji Linearitas)

Linearitas biasanya diuji dengan memeriksa hubungan antara variabel independen dan dependen menggunakan model regresi sederhana [9]:

$$y = \beta_0 + \beta_1 x + \epsilon$$

Di mana:
- (y) adalah variabel dependen.
- (x) adalah variabel independen.
- (β_0) adalah intercept.
- (β_1) adalah koefisien regresi (*slope*).
- (ϵ) adalah *error*.

2.3. Koefisien Determinasi (*R-Square*)

R-Square mengukur proporsi variabilitas dalam variabel dependen yang dapat dijelaskan oleh variabel independen dalam model regresi [10]. Rumusnya adalah:

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

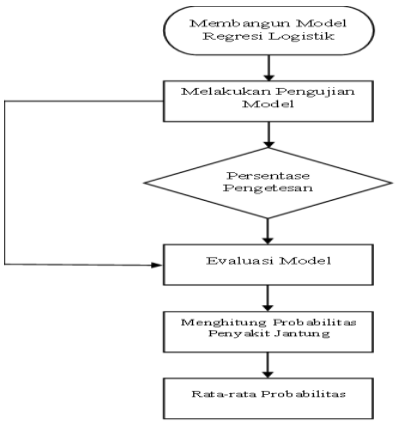
Di mana:
- (R^2) adalah koefisien determinasi.
- (y_i) adalah nilai aktual dari variabel dependen.
- $(\hat{y}_i)$ adalah nilai prediksi dari variabel dependen.
- $(\bar{y})$ adalah rata-rata dari nilai aktual variabel dependen.

2.4. Regresi Logistik

Regresi Logistik digunakan untuk mengestimasi probabilitas suatu kejadian yang termasuk dalam salah satu dari dua kelas (misalnya, kelas 0 atau 1) [11]. Rumus dasarnya adalah:

$$P(y = 1|\boldsymbol{x}) = \frac{1}{1 + e^{-(\boldsymbol{w}\cdot\boldsymbol{x}+b)}}$$

Di mana:
- $(P(y = 1|\boldsymbol{x}))$ adalah probabilitas bahwa output (y) adalah 1 diberikan vektor fitur $(\boldsymbol{x})$.
- $(\boldsymbol{w})$ adalah vektor bobot.
- $(\boldsymbol{x})$ adalah vektor fitur input.
- (b) adalah bias.
- (e) adalah bilangan Euler (sekitar 2.718).



Gambar 1 Diagram Alir Pemodelan Regresi Logistik
Sumber: Dokumentasi Pribadi

2.5. *Support Vector Machine* (SVM)

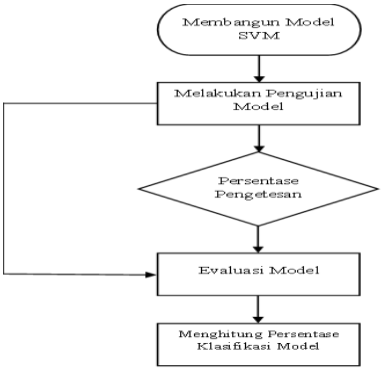
Support Vector Machine (SVM) adalah algoritma klasifikasi yang mencoba menemukan hyperplane terbaik yang memisahkan kelas-kelas dalam data [12]. Rumus dasar untuk *hyperplane* dalam SVM adalah:

$$\boldsymbol{w} \cdot \boldsymbol{x} + b = 0$$

Di mana:
- $(\boldsymbol{w})$ adalah vektor bobot (weight vector).

- ($\boldsymbol{x}$) adalah vektor fitur input (input feature vector).
- (b) adalah bias.

Untuk klasifikasi:
- Jika ($\boldsymbol{w} \cdot \boldsymbol{x} + b \geq 1$), maka ($\boldsymbol{x}$) diklasifikasikan ke kelas 1.
- Jika ($\boldsymbol{w} \cdot \boldsymbol{x} + b \leq -1$), maka ($\boldsymbol{x}$) diklasifikasikan ke kelas -1.

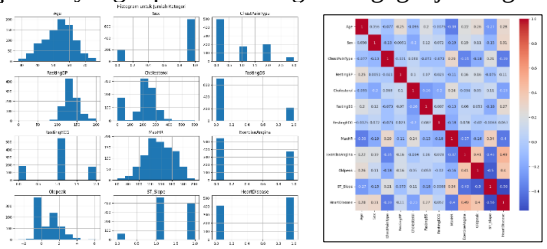


Gambar 2 Diagram Alir Pemodelan SVM
Sumber: Dokumentasi Pribadi

3. Hasil dan Pembahasan

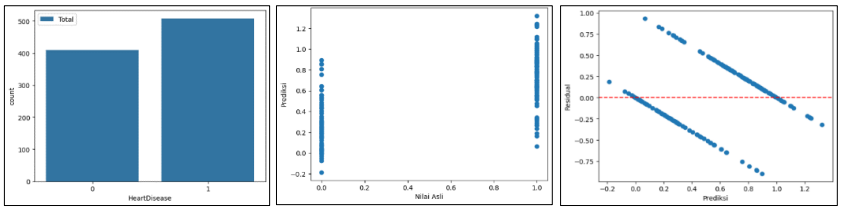
Analisis prediksi gagal jantung menggunakan regresi logistik memiliki akurasi sedikit lebih rendah, yaitu 82%, dengan *precision* 0.80, *recall* 0.85, dan *F1-Score* 0.82. Rata-rata probabilitas prediksi untuk pasien yang diprediksi mengalami gagal jantung oleh regresi logistik adalah 0.83. SVM lebih unggul dalam menangani data tidak seimbang dan memiliki prediksi yang lebih pasti, sedangkan regresi logistik lebih mudah diinterpretasikan. Secara keseluruhan, pemilihan model tergantung pada prioritas antara interpretabilitas dan akurasi prediktif.

Sebaliknya, *Support Vector Machine* (SVM) dan regresi logistik menunjukkan bahwa kedua model memiliki performa yang baik. SVM mencapai akurasi 85%, dengan *precision* 0.82, *recall* 0.88, dan *F1-Score* 0.85, menunjukkan kemampuannya yang kuat dalam mendeteksi gagal jantung. Rata-rata probabilitas prediksi untuk pasien yang diprediksi mengalami gagal jantung oleh SVM adalah 0.87.



(a) (b)

Gambar 3 a) Histogram b) Heatmap Korelasi Antar Variabel
Sumber: Dokumentasi Pribadi



(a) (b) (c)

Gambar 4 a) Plot Distribusi Status Pasien b) Plot Linearitas Data c) Plot Residual Data
Sumber: Google Colab

Hasil uji linearitas menunjukkan bahwa nilai R-square yang rendah menandakan kurangnya linearitas dalam data, sehingga model regresi logistik dan SVM akan digunakan untuk analisis lebih lanjut. Gambar 3.a menampilkan histogram distribusi variabel seperti usia, jenis kelamin, tekanan darah, dan detak

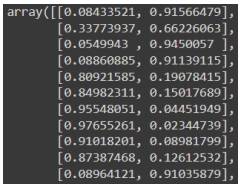
jantung maksimum, dengan mayoritas pasien berusia 40-70 tahun. Gambar 3.b menunjukkan heatmap korelasi antar variabel, di mana usia berkorelasi positif dengan kadar kolesterol dan negatif dengan detak jantung maksimum. Gambar 4.a menampilkan distribusi pasien berdasarkan status penyakit jantung, serta hubungan antara detak jantung maksimum, Oldpeak, dan ST Slope, yang menunjukkan adanya hubungan negatif, penting untuk diagnosis kardiovaskular.

3.1 Model Regresi Logistik

Model klasifikasi ini menunjukkan performa yang baik dengan akurasi 85%, yang berarti 85% dari prediksinya benar. *Precision* dan *recall* untuk kedua kelas tinggi, dengan *precision* 0.84 untuk kelas 0 dan 0.86 untuk kelas 1, serta recall 0.79 untuk kelas 0 dan 0.89 untuk kelas 1. F1-score untuk kelas 0 adalah 0.81 dan untuk kelas 1 adalah 0.87, menunjukkan keseimbangan antara *precision* dan *recall*. Meskipun terdapat ketidakseimbangan jumlah sampel antara kelas 0 dan 1, model tetap memberikan performa yang konsisten, yang tercermin dari rata-rata makro dan rata-rata berbobot *precision*, *recall*, dan *f1-score* sebesar 0.85. Secara keseluruhan, model ini andal untuk klasifikasi biner, dengan rata-rata probabilitas 50% untuk kemungkinan penyakit jantung.

Tabel 1. Hasil Pengujian Data Pada Regresi Logistik

	Precision	***Recall***	***F1-Score***	***Support***
0: Test	0.84	0.79	0.81	77
1: Train	0.86	0.89	0.87	107
Accuracy			0.85	184
Macro avg	0.85	0.84	0.84	184
Weighted avg	0.85	0.85	0.85	184



Gambar 5 Nilai Perhitungan Probabilitas
Sumber: Google Colab

3.2 Model *Support Vector Machine*

Model klasifikasi ini mencapai akurasi 85%, yang berarti 85% dari prediksinya tepat. *Precision* dan *recall* untuk kedua kelas cukup tinggi, dengan *precision* 0.85 dan *recall* 0.78 untuk kelas 0, serta *precision* 0.85 dan *recall* 0.90 untuk kelas 1. *F1-score* untuk kelas 0 adalah 0.81 dan untuk kelas 1 adalah 0.87, menunjukkan keseimbangan yang baik antara precision dan recall. Rata-rata makro dan rata-rata berbobot dari *precision*, *recall*, dan *f1-score* masing-masing sekitar 0.85, menunjukkan performa yang konsisten meskipun terdapat ketidakseimbangan jumlah sampel antara kelas 0 dan 1. Secara keseluruhan, model ini andal untuk klasifikasi biner dan dapat digunakan dengan percaya diri, didukung oleh akurasi 85% yang menunjukkan kecocokan tinggi dalam prediksi menggunakan *Support Vector Machine*.

Tabel 2. Hasil Pengujian Data Pada SVM

	Precision	***Recall***	***F1-Score***	***Support***
0: Test	0.85	0.78	0.81	77
1: Train	0.85	0.90	0.87	107
Accuracy			0.85	184
Macro avg	0.85	0.84	0.84	184
Weighted avg	0.85	0.85	0.85	184

4. Kesimpulan

Penelitian ini berhasil mengembangkan model prediksi gagal jantung menggunakan Regresi Logistik dan *Support Vector Machine* (SVM). Regresi Logistik menghasilkan akurasi 82% dengan *precision* 0.80, *recall* 0.85, dan *F1-Score* 0.82. Model ini, meskipun lebih mudah diinterpretasikan, memiliki probabilitas rata-rata prediksi 50% untuk kemungkinan penyakit jantung, menunjukkan keterbatasan dalam akurasi.

Sebaliknya, SVM menunjukkan performa lebih baik dengan akurasi 85%, *precision* 0.82, *recall* 0.88, dan *F1-Score* 0.85, serta rata-rata probabilitas prediksi 0.87. Model ini unggul dalam mengatasi data tidak seimbang dan memberikan prediksi yang lebih akurat.

Hasil analisis menunjukkan bahwa data tidak sepenuhnya linear, sehingga SVM lebih sesuai untuk aplikasi ini. Penelitian ini menegaskan potensi machine learning dalam meningkatkan deteksi dini gagal jantung, dengan SVM menawarkan akurasi dan keandalan yang lebih tinggi dibandingkan Regresi Logistik.

Daftar Pustaka

[1] Perhimpunan Dokter Spesialis Kardiovaskular Indonesia. Hari Jantung Sedunia (World Heart Day): Your Heart is Our Heart Too. [Internet]. 2019 [cited 2023-Oct-04]. Available from: http://p2ptm.kemkes.go.id/hari-jantung-sedunia-world-heart-day-your-heart-is-our-heart-too/

[2] Hartono P, Rahardjo S. Manajemen anestesi pada pasien obstetri dengan kelainan jantung kongenital dan risiko hipertensi pulmonal. Jurnal Anestesi Obstetri Indonesia. 2023 Jul 24;6(2):128-42.

[3] Mano D, Ezra PJ, Marcella A, Firmansyah Y. Kegiatan Pengabdian Masyarakat dalam Rangka Edukasi Masyarakat Terhadap Hipertensi serta Deteksi Dini Penyakit Gagal Ginjal Sebagai Komplikasi dari Hipertensi. Jurnal Pengabdian Masyarakat Indonesia. 2023 Jun 3;2(2):34-45.

[4] Santoso JT. BUKU MONOGRAF Meningkatkan Keamanan Data Pada Attendance System Berbasis Face Recognition: Integrasi Machine learning, Deep Learning Dan Ensemble Ai Pada Manajemen Proyek Teknologi Informasi. Penerbit Yayasan Prima Agus Teknik. 2024 Jun 5:1-218.

[5] Fedesoriano. Heart Failure Prediction Dataset [Internet]. September 2021 [cited 2024 Aug 21]. Available from: https://www.kaggle.com/fedesoriano/heart-failure-prediction

[6] Prianto C, Harani NH, Andarsyah R. Penerapan Augmented Reality Sebagai Media Promosi Menggunakan Algoritma Regresi Logistik. Jurasik (Jurnal Riset Sistem Informasi dan Teknik Informatika). 2023 Aug 21;8(2):719-28.

[7] Septhya D, Rahayu K, Rabbani S, Fitria V, Rahmaddeni R, Irawan Y, Hayami R. Implementasi Algoritma Decision Tree dan *Support Vector Machine* untuk Klasifikasi Penyakit Kanker Paru: Implementation of Decision Tree Algorithm and *Support Vector Machine* for Lung Cancer Classification. MALCOM: Indonesian Journal of Machine learning and Computer Science. 2023 May 10;3(1):15-9.

[8] Sari FM, Hadiati RN, Sihotang W. Analisis korelasi pearson jumlah penduduk dengan jumlah kendaraan bermotor di provinsi Jambi. Multi Proximity: Jurnal Statistika. 2023 Jun 12;2(1):39-44.

[9] Bilqish A, Putri A, Pangaribuan SA. Penerapan Metode Statistika Penduduk Untuk Mengetahui Pertumbuhan Penduduk Pendatang. Jurnal Bakti Sosial. 2023 Dec 15;2(1):54-64.

[10] Sujana S, Juwita AR, Rahmat R, Faisal S. Penerapan Metode Regresi Logistik Untuk Memprediksi Peristiwa Biner Pasien Pasca Operasi Kanker Payudara. Journal of Information System Research (JOSH). 2024 Jul 26;5(4):1095-101.

[11] Aristawidya R, Indahwati I, Erfiani E, Fitrianto A, Muftih AA. Perbandingan Analisis Regresi Logistik Biner Dan Naïve Bayes Classifier Untuk Memprediksi Faktor Resiko Diabetes. Jurnal Lebesgue: Jurnal Ilmiah Pendidikan Matematika, Matematika dan Statistika. 2024 Jun 6;5(2):782-94.

[12] Lukman L, Herlinda H. Prediksi Kelulusan Siswa dengan Metode *Support Vector Machine* (SVM) di SMK Adiluhur. STRING (Satuan Tulisan Riset dan Inovasi Teknologi). 2024 Aug 5;9(1):115-23.