

Prediksi Gagal Jantung: Pendekatan Data Mining untuk Deteksi dan Penilaian Risiko

Yunanda Arista¹, Nani Sulistianingsih^{*2}, Trindah Fitri³, Zabina Dwi Febriyanzi⁴

^{1,2,3,4}Sistem dan Teknologi Informasi, Universitas Muhammadiyah Mataram, Indonesia

¹the.y.a.alien@gmail.com, ²nani.sulistianingsih@ummat.ac.id, ³fitrisia544@gmail.com,

⁴zabinadwi14@gmail.com,

ABSTRACT

Keywords:

Heart Failure
Decision Tree
Random Forest

Abstract: Heart failure is one of the main causes of death which requires early detection so that treatment can be carried out more quickly. This research compares the performance of the Decision Tree and Random Forest algorithms in predicting heart failure using datasets from Kaggle. The research stages include data collection, pre-processing, classification and evaluation using accuracy, precision, recall and F1 score metrics. The research results show that Random Forest has higher accuracy (77%) than Decision Tree (73%), so it is superior in recognizing overall patterns. However, Decision Tree is superior in detecting patients who actually have heart failure, with 76% precision and 52% recall, compared to Random Forest which only achieves 50% precision and 29% recall. Although Random Forest is generally more accurate, Decision Tree is better at identifying high-risk patients. The results of this study show that the application of machine learning algorithms can help medical personnel identify at-risk patients earlier. This allows for quicker action and can help reduce death rates from heart failure.

Kata Kunci:

Gagal Jantung
Decision Tree
Random Forest

Abstrak: Gagal jantung adalah salah satu penyebab utama dalam kematian yang membutuhkan deteksi dini agar penanganannya bisa dilakukan lebih cepat. Penelitian ini membandingkan performa algoritma Decision Tree dan Random Forest dalam memprediksi gagal jantung menggunakan dataset dari Kaggle. Dalam tahapan penelitian meliputi pengumpulan data, pra-pemrosesan, klasifikasi, serta evaluasi menggunakan metrik akurasi, precision, recall, dan F1-score. Hasil penelitian menunjukkan bahwa Random Forest memiliki akurasi lebih tinggi (77%) dibandingkan Decision Tree (73%), sehingga lebih unggul dalam mengenali pola secara keseluruhan. Namun, Decision Tree lebih unggul dalam mendeteksi pasien yang benar-benar mengalami gagal jantung, dengan precision 76% dan recall 52%, dibandingkan Random Forest yang hanya mencapai precision 50% dan recall 29%. Meskipun Random Forest lebih akurat secara umum, Decision Tree lebih baik dalam mengenali pasien yang berisiko tinggi. Hasil penelitian ini menunjukkan bahwa penerapan algoritma machine learning dapat membantu tenaga medis mengidentifikasi pasien berisiko lebih awal. Hal ini memungkinkan tindakan lebih cepat dan dapat membantu menekan angka kematian akibat gagal jantung.

Article History:

Received : 25-02-2025

Accepted : 27-03-2025



This is an open access article under the [CC-BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license

A. LATAR BELAKANG

Gagal jantung merupakan kondisi kronis yang ditandai dengan ketidakmampuan jantung untuk memompa darah secara efektif guna memenuhi kebutuhan metabolisme tubuh. Penyakit ini menjadi salah satu penyebab utama morbiditas dan mortalitas di seluruh (Leny et al., 2020). Akibat meningkatnya prevalensi gagal jantung, terdapat kebutuhan mendesak untuk metode deteksi dini yang lebih akurat guna mengurangi dampak negatifnya terhadap pasien dan sistem kesehatan. *Data mining*, yang merupakan proses ekstraksi pola dari kumpulan data besar menggunakan teknik pembelajaran mesin, telah berkembang sebagai alat yang efektif dalam menganalisis faktor risiko dan meningkatkan prediksi gagal jantung. Dengan menggabungkan algoritma pembelajaran mesin, model berbasis data dapat meningkatkan keakuratan diagnosis dibandingkan dengan metode

konvensional yang sering kali subjektif dan terbatas pada evaluasi klinis manual (Firmansyach et al., 2023).

Deteksi dini gagal jantung sangat penting untuk meningkatkan hasil pasien dan mengurangi beban kesehatan global yang terkait dengan kondisi ini. Gagal jantung mempengaruhi jutaan orang di seluruh dunia, menyebabkan tingkat morbiditas dan mortalitas yang tinggi. Namun, metode diagnostik konvensional sering kali kurang akurat karena gejala yang tidak spesifik, sehingga menekankan pentingnya pendekatan inovatif seperti data mining dan pembelajaran mesin. Diagnosis dini memungkinkan intervensi tepat waktu yang dapat secara signifikan mengurangi komplikasi dan angka kematian (Alsubai et al., 2023; Annisa, 2019). Model pembelajaran mesin telah menunjukkan peningkatan kemampuan prediktif, dengan beberapa algoritma canggih seperti pembelajaran kuantum mencapai akurasi hingga 98% (Alsubai et al., 2023). Namun, tantangan utama dalam metode konvensional adalah ketergantungannya pada penilaian subjektif yang dapat menyebabkan tanda-tanda awal gagal jantung terabaikan. Selain itu, catatan klinis sering kali mengalami permasalahan data yang hilang, yang menyulitkan pengembangan model prediksi yang andal (Guo et al., 2021). Oleh karena itu, teknik data mining seperti *Random Forest* dan *XGBoost* semakin banyak digunakan karena kemampuannya dalam menangani dataset besar, meningkatkan akurasi prediksi, serta mengatasi ketidakseimbangan data dan nilai yang hilang (Edric & Tamba, 2022). Meskipun demikian, tantangan seperti *overfitting* dan kebutuhan validasi klinis yang kuat tetap perlu diperhatikan agar model ini dapat diterapkan secara efektif dalam praktik medis.

Teknik data mining memainkan peran penting dalam memprediksi gagal jantung dengan mengidentifikasi pola dan faktor risiko dari kumpulan data medis yang luas. Algoritma seperti *Random Forest*, regresi logistik, dan *Support Vector Machines* (SVM) telah terbukti meningkatkan akurasi prediksi dibandingkan dengan metode tradisional yang bergantung pada penilaian subjektif (Depari et al., 2022). Salah satu keuntungan utama data mining adalah kemampuannya dalam menganalisis berbagai variabel seperti usia, tekanan darah, dan kebiasaan gaya hidup untuk mengidentifikasi faktor risiko yang signifikan (Riany & Testiana, 2023). Model *Random Forest*, misalnya, telah menunjukkan korelasi yang kuat antara indikator medis seperti kreatinin fosfokinase (CPK) dengan risiko gagal jantung (Wan et al., 2024). Keunggulan metode berbasis data dibandingkan metode tradisional terletak pada kemampuannya memberikan wawasan objektif dan dapat diukur, sehingga mengurangi ketergantungan pada evaluasi subjektif tenaga medis (Tasnim et al., 2023). Selain itu, model pembelajaran mesin mampu memproses kumpulan data besar dan mengungkap hubungan kompleks yang sering kali diabaikan dalam metode konvensional (Raharja et al., 2024). Studi menunjukkan bahwa teknik *data mining* dapat mencapai tingkat akurasi melebihi 85% dalam memprediksi penyakit jantung (Wan et al., 2024). Melalui integrasikan berbagai algoritma, analisis yang lebih komprehensif dapat dilakukan, meningkatkan keandalan prediksi (Derisma, 2020). Namun, penting untuk mempertimbangkan bahwa model ini memerlukan data berkualitas tinggi dan validasi yang menyeluruh untuk menghindari risiko *overfitting* serta memastikan penerapannya dalam berbagai populasi pasien.

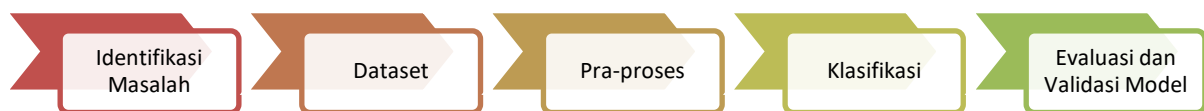
Penelitian tentang prediksi gagal jantung berfokus pada peningkatan akurasi model prediktif menggunakan teknik pembelajaran mesin sudah banyak dilakukan. Hal ini untuk mengembangkan model yang kuat yang dapat secara efektif mengidentifikasi pasien berisiko dan memprediksi hasil klinisnya, sehingga memungkinkan intervensi yang lebih tepat waktu. Pada upaya meningkatkan akurasi prediksi, beberapa model telah dilaporkan mencapai akurasi hingga 93,36% (Sepharni et al., 2022). Selain itu, penelitian lain juga menekankan pentingnya deteksi dini untuk meningkatkan manajemen dan hasil pasien (Sepharni et al., 2022). Ruang lingkup penelitian mencakup penggunaan kumpulan data klinis dari sumber seperti Kaggle dan MIMIC-III, yang menawarkan data pasien yang komprehensif (Tasnim et al., 2023). Teknik pembelajaran mesin yang digunakan meliputi *Random Forest*, *Support Vector Machines*, dan regresi logistik, dengan metode pemilihan fitur seperti algoritma genetik dan *CoxBoost*. Studi ini juga membandingkan berbagai model, di mana *Random*

Forest secara konsisten menunjukkan kinerja unggul dalam akurasi dan kemampuan prediktif. Meskipun begitu, penting untuk mempertimbangkan keterbatasan seperti risiko *overfitting* serta kebutuhan validasi ekstensif untuk memastikan model dapat diterapkan secara luas.

Meskipun metode *Decision Tree* dan *Random Forest* telah digunakan dalam berbagai studi untuk mendeteksi dini gagal jantung, masih terdapat beberapa gap penelitian yang perlu diatasi. Salah satu celah utama adalah keterbatasan dalam interpretabilitas model *Decision Tree* dan tantangan dalam menangani dataset dengan variabel yang sangat kompleks. Penelitian ini berfokus pada perbandingan performa klasifikasi dengan model *Decision Tree* dan *Random Forest*. Oleh karena itu, penelitian ini bertujuan untuk mengeksplorasi strategi optimalisasi dalam *Decision Tree* dan *Random Forest* guna meningkatkan ketepatan dan efisiensi prediksi gagal jantung. Melalui mengembangkan pendekatan yang lebih adaptif dan efisien, penelitian ini diharapkan dapat memberikan kontribusi signifikan dalam meningkatkan deteksi dini gagal jantung dan mendukung pengambilan keputusan klinis berbasis data yang lebih akurat.

B. METODE PENELITIAN

Metode penelitian memberikan gambaran jelas bagaimana proses penelitian akan dilakukan. Beberapa tahapan dalam penelitian ini adalah identifikasi masalah, pengumpulan data (dataset), pra proses data, klasifikasi dengan *decision tree* dan *random forest* serta evaluasi dan validasi model. Alur penelitian lebih jelasnya dapat dilihat pada Gambar 1.



Gambar 1. Alur Penelitian Prediksi Gagal Jantung

1. Identifikasi Masalah

Pada tahap identifikasi masalah merupakan tahapan awal pada proses penelitian. Tahapan ini akan ditentukan masalah dan solusi dari permasalahan yang ada. Masalah yang ditemukan dari beberapa penelitian terdahulu bahwa dataset dengan variabel yang sangat kompleks serta penanganan *overfitting* pada dataset gagal jantung. Sehingga dibutuhkan sebuah perbandingan algoritma yang mampu menangani variabel kompleks serta *overfitting* seperti algoritma *Decision Tree* dan *Random Forest*.

2. Dataset

Dataset merupakan Kumpulan data yang dapat berfungsi sebagai bahan dalam penelitian. Dataset yang digunakan pada penelitian ini adalah dataset sekunder yang diperoleh dari <https://www.kaggle.com/>. Total data yang digunakan sebanyak 299 baris dan 13 kolom. Dataset tersebut dibagi menjadi 80% data latih dan 20% data uji.

3. Pre Processing Data

Pada tahap *preprocessing*, data akan dibersihkan seperti menghapus duplikat, mendeteksi missing value, mendeteksi outlier dan melakukan normalisasi data.

4. Decision Tree

Decision Tree adalah salah satu metode dalam *data mining* yang digunakan untuk membantu membuat keputusan berdasarkan data yang ada. Algoritma ini berbentuk seperti pohon, di mana setiap bagian dari pohon mewakili pertanyaan atau aturan yang harus dipenuhi untuk mencapai kesimpulan akhir. *Decision Tree* digunakan untuk menganalisis data pasien dan memprediksi apakah mereka berisiko mengalami gagal jantung atau tidak. Cara kerja algoritma ini adalah dengan

membagi data berdasarkan variabel yang paling berpengaruh, seperti usia, tekanan darah, kadar kreatinin, dan faktor kesehatan lainnya (Firmansyach et al., 2023). Pemilihan atribut dilakukan berdasarkan perhitungan Tingkat impurity menggunakan metode seperti berikut:

- a. *Entropy* digunakan untuk mengukur ketidakteraturan dalam suatu dataset, dirumuskan sebagai:

$$Entropy(S) = \sum_{i=1}^n -p_i * \log_2 p_i \quad (1)$$

Keterangan :

S : Himpunan Kasus

n : Jumlah partisi S

P_i : Proporsi dari sisi terhadap S

- b. Gain (S, A) atau *information Gain* adalah cara untuk mengukur seberapa berguna suatu atribut (A) dalam mengurangi kebingungan (entropi) dalam data (S). Semakin besar nilai Information Gain, semakin efektif atribut tersebut dalam membagi data menjadi lebih teratur. Dirumuskan sebagai berikut:

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^n \left(\frac{|S_i|}{|S|} \right) * Entropy(S_i) \quad (2)$$

Keterangan :

S : Himpunan kasus atau dataset utama sebelum dipartisi

A : Atribut yang digunakan untuk membagi dataset

n : Jumlah partisi (jumlah subset S_i yang terbentuk setelah pemisahan berdasarkan atribut A)

$|S_i|$: Jumlah kasus pada partisi ke i

$|S|$: Jumlah kasus didalam S

5. Random Forest

Random Forest adalah salah satu algoritma *machine learning* dan *data mining* yang digunakan untuk mengklasifikasikan pasien gagal jantung berdasarkan data medis mereka. Algoritma ini bekerja dengan membuat banyak pohon keputusan (decision trees) dan mengambil keputusan berdasarkan hasil voting dari semua pohon tersebut. Random Forest lebih stabil dan akurat dibandingkan metode lain seperti *Decision Tree*. Pada penelitian ini, Random Forest diterapkan pada dataset *Heart Failure Clinical Records*, yang berisi berbagai informasi medis pasien, seperti usia, fraksi ejeksi, kadar kreatinin serum, kadar natrium serum, dan waktu tindak lanjut. Model ini bertujuan untuk memprediksi seorang pasien memiliki risiko tinggi meninggal akibat gagal jantung atau tidak (Edric & Tamba, 2022). Berikut adalah rumus yang digunakan dalam *Random Forest*.

- a. *Entropy* adalah ukuran ketidakpastian atau ketidakteraturan dalam suatu dataset. Rumusnya dinyatakan sebagai berikut:

$$Entropy(Y) = -\sum p(c|Y) \log_2 p(c|Y) \quad (3)$$

Dimana :

Y : Kumpulan data untuk diukur keragamannya.

C : Kategori atau kelas dalam data.

$p(c|Y)$: Seberapa sering kelas 'c' muncul dalam data Y .

$\log_2(p(c|Y))$: Ukuran informasi (dalam bit) untuk menyatakan kelas 'c'. Kelas yang umum butuh sedikit informasi, kelas yang jarang butuh lebih banyak.

b. *Information Gain*

Information Gain ialah mengukur seberapa banyak ketidakpastian (entropy) berkurang ketika dataset Y dibagi berdasarkan atribut a. Rumusnya sebagai berikut:

$$Gain(Y, a) = Entropy(Y) - \sum |Y_v| |Y_a| veValues Entropy(Y_v) \quad (3)$$

Dimana:

Y: Data utama yang akan dipisah-pisah.

a: Fitur yang dipakai untuk membagi data Y.

$v \in Values(a)$: Pilihan nilai yang mungkin untuk fitur 'a'.

|Y| : Total jumlah data keseluruhan.

|Y_v| : Jumlah data yang memiliki nilai 'v' pada fitur 'a'.

$Entropy(Y_v)$: Ukuran ketidakpastian dalam kelompok data yang memiliki nilai 'v' pada fitur 'a'.

6. Evaluasi dan Validasi

Efektivitas algoritma *Decision Tree* dan *Random Forest* masing-masing diukur dengan akurasi, presisi, recall, F1-score. Setelah model dilatih, evaluasi beberapa model perlu dilakukan untuk mendapatkan model terbaik (Rizqullah et al., 2024). Definisi untuk setiap metrik sebagai berikut:

TP (*True Positive*) : Jumlah kasus positif di mana kasus sudah positif yang sebenarnya diprediksi positif.

FP (*False Positive*) : Jumlah kasus negatif yang diprediksi positif.

FN (*False Negative*) : Jumlah kasus positif yang diprediksi negatif.

TN (*True Negative*) : Jumlah kasus negatif di mana kasus sudah negatif yang sebenarnya diprediksi negatif

a. *Accuracy* (Akurasi)

Accuracy adalah metrik evaluasi yang mengukur sejauh mana suatu model klasifikasi dapat memprediksi dengan benar, dihitung dengan membandingkan jumlah prediksi yang positif dan negatif yang benar dengan jumlah total data yang diuji. Rumusnya dinyatakan sebagai berikut:

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \quad (4)$$

b. *Precision* (Presisi)

Precision mengukur kemampuan model dalam mengidentifikasi kelas positif secara akurat. Berikut ini adalah rumusnya:

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

c. *Recall*

Recall adalah metrik evaluasi yang mengukur kemampuan model dalam mendeteksi semua kasus positif dengan benar. Dihitung dengan membagi jumlah True Positive (TP) dengan total kasus positif sebenarnya (TP + False Negative). Berikut rumusnya:

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

d. *F1-Score*

F1-score atau *F-measure*, adalah ukuran yang digunakan untuk mengevaluasi kinerja model klasifikasi dengan mempertimbangkan keseimbangan antara precision dan recall. Metrik ini sangat berguna dalam kasus di mana kelas positif lebih penting, terutama jika jumlah data positif jauh lebih sedikit dibandingkan dengan data negatif. Berikut adalah rumusnya :

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (7)$$

C. HASIL DAN PEMBAHASAN

Pada tahap ini hasil penelitian akan dibahas, dari proses persiapan dataset, hingga evaluasi dan validasi model dengan Random Forest dan Decision Tree.

1. Persiapan Data

Dalam penelitian ini, Kami mengambil dataset dari <https://www.kaggle.com/> yang berfokus pada dataset gagal jantung. Dataset ini terdiri dari 13 kolom variabel dan 299 baris. Kolom-kolom mencakup berbagai fitur yang relevan dan setiap baris dalam dataset mewakili satu pasien yang sudah didiagnosis. Berikut adalah penjelasan yang lebih jelas mengenai 13 kolom variabel dapat dilihat pada Tabel 1 dan dataset gagal jantung pada Tabel 2.

Tabel 1. Kolom variabel Dataset

| Variabel | Deskripsi | Nilai |
|--------------------------|--|--------------------------------------|
| Age | Usia pasien | 40 – 95 (Tahun) |
| Anaemia | Penurunan hemoglobin atau sel darah merah | 0 = Tidak 1 = Iya |
| Creatinine phosphokinase | Kadar enzim CPK dalam darah | 23 – 7861 (mcg/L) |
| Diabetes | Pasien mengalami diabetes atau tidak | 0 = Tidak 1 = Iya |
| Ejection fraction | Persentase darah yang dipompa ke jantung pada setiap kontraksi | 14 – 80 (%) |
| High blood pressure | Pasien menderita hipertensi atau tidak | 0 = Tidak 1 = Iya |
| Platelets | Jumlah trombosit dalam darah | 25100 – 850000 (kiloplatelets/mL) |
| Serum creatinine | Jumlah kadar kreatinin yang terdapat dalam darah | 0.5 - 9.4 (mg/dL) |
| Serum sodium | Jumlah kadar natrium yang terdapat dalam darah | 113 – 148 (mEq/L) |
| Sex | Jenis kelamin pasien | 0 = Perempuan 1 = Laki-laki |
| Smoking | Pasien merokok atau tidak | 0 = Tidak 1 = Iya |
| Time | Waktu pengamatan | 4 – 285 (Hari) |
| DEATH_EVENT | Pasien meninggal atau tidak dalam waktu pengamatan | 0 = Tidak meninggal 1 = Meninggal |

Tabel 2. Dataset Gagal Jantung

| Age | Anemia | Creatinine_phosphokinase | Diabetes | ... | Time | Death_event |
|-----|--------|--------------------------|----------|-----|------|-------------|
| 75 | 0 | 582 | 0 | ... | 4 | 1 |
| 55 | 0 | 7861 | 0 | ... | 6 | 1 |
| 65 | 0 | 146 | 0 | ... | 7 | 1 |
| ... | ... | ... | ... | ... | ... | ... |
| 45 | 0 | 2413 | 0 | ... | 280 | 0 |
| 50 | 0 | 196 | 0 | ... | 280 | 0 |

2. Pre Processing Data

Pra-pemrosesan data adalah langkah awal yang sangat penting dalam analisis data untuk menyiapkan data mentah menjadi lebih bersih, terstruktur, dan siap digunakan dalam proses analisis, pemodelan statistik, atau penerapan algoritma data mining. Dalam konteks analisis gagal jantung, tahapan pra-pemrosesan meliputi:

a. Menghapus Duplikat Data

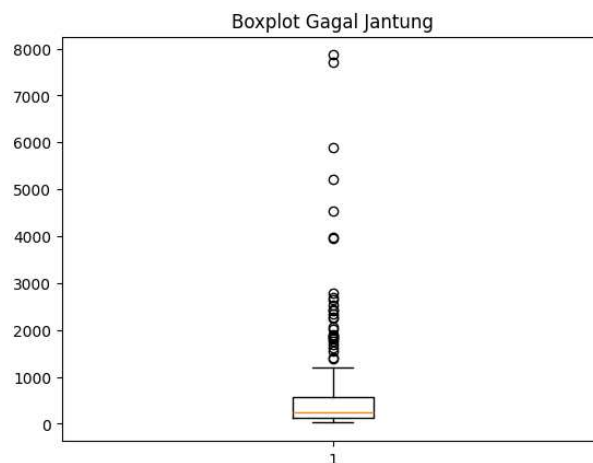
Pada dataset penyakit gagal jantung ini tidak ditemukan data double, sehingga tidak dilakukan proses penghapusan data. Hal ini menunjukkan bahwa data telah bersih dari data double atau duplikat yang dapat mempengaruhi hasil klasifikasi.

b. Mendeteksi *Missing Value*

Setelah dilakukannya proses mendeteksi duplikat data, maka tahapan selanjutnya adalah penanganan missing value atau nilai yang hilang. Pada dataset penyakit gagal jantung ini tidak ditemukan data yang hilang atau kosong sehingga proses ke tahap selanjutnya dapat dilakukan.

c. Deteksi *Outlier*

Nilai outlier atau (pencilan) dapat diartikan sebagai nilai yang memiliki karakteristik dengan kebanyakan nilai lain dalam sebuah dataset. Berikut hasil variabel creatine kinase yang terdapat outlier.



Gambar 2. Boxplot Gagal Jantung

Berdasarkan gambar 2 menunjukkan variabel Creatine Kinase dalam dataset gagal jantung terlihat distribusi yang tidak normal dengan banyak outlier yang mencerminkan kondisi medis tertentu yang signifikan. Terlihat beberapa pasien mengalami lonjakan kadar *Creatine Kinase* yang jauh lebih tinggi dibandingkan mayoritas pasien lainnya. *Creatine Kinase* adalah enzim yang terkait dengan kerusakan otot, termasuk jantung (Nainggolan, 2021)

d. Normalisasi Data

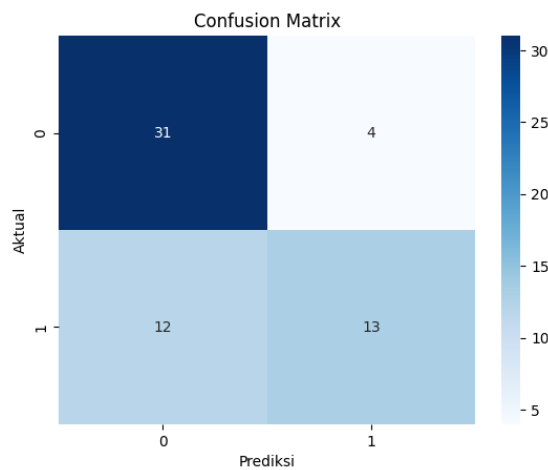
Tahap ini dilakukan normalisasi data, normalisasi data adalah langkah penting dalam analisis data, terutama dalam konteks prediksi gagal jantung. Tujuannya adalah untuk memastikan bahwa perbedaan skala antar fitur tidak menimbulkan kesalahan dalam analisis dan pemodelan. Salah satu teknik yang umum digunakan adalah *Min-Max* normalisasi, yang mentransformasi nilai data menjadi rentang 0 hingga 1. Nama kolom pada dataset dirubah untuk memudahkan saat pemanggilan variabel. Dari 13 variabel 7 diantaranya di normalisasi untuk menyamakan nilai kecuali Berikut hasil normalisasi data dapat dilihat pada tabel dibawah ini.

Tabel 3. Normalisasi Data Gagal Jantung

| Usia | Kreatine fosfokinase | Fraksi ejeksi | Platelets | Kreatinin serum | Natrium serum | Time |
|----------|----------------------|---------------|-----------|-----------------|---------------|----------|
| 0.636364 | 0.071319 | 0.090909 | 0.290823 | 0.157303 | 0.485714 | 0.021352 |
| 0.272727 | 0.011227 | 0.363636 | 0.288833 | 0.064716 | 0.657143 | 0.007117 |
| | | | | | | |
| 0.181818 | 0.015963 | 0.090909 | 0.224148 | 0.157303 | 0.685714 | 0.010676 |
| 0.454545 | 0.017479 | 0.090909 | 0.365984 | 0.247191 | 0.085714 | 0.014235 |

3. Klasifikasi dengan Model Decision Tree dan Random Forest

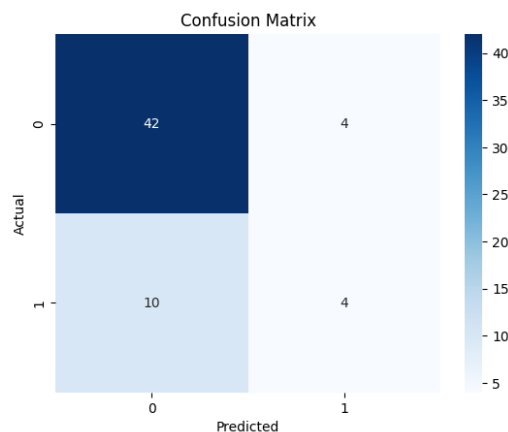
Penelitian ini dilakukan untuk membandingkan kedua kinerja dari model *Decision Tree* dan Random Forest dalam memprediksi pasien beresiko gagal jantung. Berdasarkan hasil pengujian dengan matriks evaluasi *confusion matrix* pada model *Decision Tree* 31 benar dikelompokkan ke dalam pasien tidak gagal jantung dan 4 salah dikategorikan sebagai pasien gagal jantung. Kemudian 13 benar dikelompokkan sebagai pasien gagal jantung dan 12 salah dikelompokkan sebagai tidak gagal jantung. Lebih detailnya dapat dilihat pada *confusion matrix* pada Gambar 3.



Gambar 3. *Confusion Matrix* dengan Model *Decision Tree*

Tabel 3. *Confusion Matrix* dengan Model *Decision Tree*

| Aktual \ Prediksi | 0 (Tidak Gagal Jantung) | 1 (Gagal Jantung) |
|---------------------------------|-------------------------|--------------------|
| 0 (Negatif/Tidak Gagal Jantung) | 31 (True Negative) | 4 (False Positive) |
| 1 (Positif/Gagal Jantung) | 12 (False Negative) | 13 (True Positive) |



Gambar 4. *Confusion Matrix* dengan Model *Random Forest*

Confusion matrix menunjukkan bahwa model salah mengklasifikasikan 4 dari 46 sampel tidak gagal jantung sebagai gagal jantung dan 10 dari 14 sampel gagal jantung sebagai tidak gagal jantung.

Tabel 4. *Confusion Matrix* dengan *Model Random Forest*

| Aktual \ Prediksi | 0 (Tidak Gagal Jantung) | 1 (Gagal Jantung) |
|---------------------------------|-------------------------|--------------------|
| 0 (Negatif/Tidak Gagal Jantung) | 31 (True Negative) | 4 (False Positive) |
| 1 (Positif/Gagal Jantung) | 12 (False Negative) | 13 (True Positive) |

Pada klasifikasi menggunakan *Decision Tree* (DT) dan *Random Forest* (RF), masing-masing memiliki kelebihan dan kekurangan. *Decision Tree* lebih sederhana dan mudah dipahami, namun rentan terhadap overfitting, terutama jika datanya kompleks. Sementara itu, *Random Forest* bekerja dengan menggabungkan banyak pohon keputusan, sehingga hasil lebih stabil dan akurasi lebih tinggi. Berdasarkan hasil akurasi, *Random Forest* memiliki akurasi 77% sedikit lebih tinggi dibandingkan *Decision Tree* yaitu 73%. Hal ini menunjukkan bahwa *Random Forest* lebih unggul dalam memprediksi secara keseluruhan, karena mekanisme ensemble yang membantu mengurangi kesalahan prediksi. Namun, pemilihan algoritma terbaik tidak bergantung pada akurasi. *Decision Tree*, meskipun memiliki akurasi lebih rendah, tetapi lebih baik dalam mengenali pasien yang benar-benar mengalami gagal jantung. Sementara *Random Forest* lebih akurat secara keseluruhan, kemampuannya dalam mendeteksi kasus gagal jantung lebih lemah. Oleh karena itu, dalam memilih model terbaik, perlu mempertimbangkan tujuan utama dari klasifikasi yang dilakukan.

D. SIMPULAN DAN SARAN

Penelitian ini membandingkan performa algoritma *Decision Tree* dan *Random Forest* menggunakan dataset gagal jantung. Proses penelitian dilakukan dengan beberapa tahapan meliputi, identifikasi masalah, pengumpulan dataset, pra-pemrosesan data, klasifikasi, dan evaluasi model yang menggunakan metrik *accuracy*, *precision*, *recall*, dan *F1-score*. Hasil dari penelitian ini dimana algoritma *Random Forest* memiliki akurasi lebih unggul dan akurat yaitu 77% di bandingkan dengan *Decision Tree* yaitu 73%. Hal ini menunjukkan bahwa *Random Forest* lebih baik dan stabil dalam mengklasifikasikan data secara keseluruhan, Penelitian ini menunjukkan bahwa penggunaan algoritma dalam memprediksi gagal jantung dapat membantu tenaga medis untuk mengidentifikasi pasien yang berisiko lebih awal, sehingga memungkinkan penanganan yang didapat lebih cepat agar bisa mengurangi angka kematian akibat gagal jantung.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada semua pihak yang telah berkontribusi dalam penyusunan paper ini, terutama kepada Ibu Nani Sulistianingsih, S.Kom.,M.Eng yang senantiasa memberikan bimbingan, nasehat, dan motivasi kepada penulis sehingga penelitian ini selesai dengan baik.

REFERENSI

- Alsubai, S., Alqahtani, A., Binbusayyis, A., Sha, M., Gumaei, A., & Wang, S. (2023). Heart Failure Detection Using Instance Quantum Circuit Approach And Traditional Predictive Analysis. *Mathematics*, 2(1), 1–27.
- Annisa, R. (2019). Analisis Komparasi Algoritma Klasifikasi Data Mining. *Jurnal Teknik Informatika Kaputama*, 3(1).
- Depari, D. H., Widiastiwi, Y., & Santoni, M. M. (2022). Perbandingan Model Decision Tree , Naive Bayes Dan Random Forest Untuk Prediksi Klasifikasi Penyakit Jantung. *Jurnal Informatik*, 4(6), 239–248.
- Derisma. (2020). Perbandingan Kinerja Algoritma Untuk Prediksi Penyakit Jantung Dengan Teknik Data Mining. *Journal Of Applied Informatics And Computing* (, 4(1), 84–88.
- Edric, & Tamba, S. P. (2022). Prediksi Penyakit Gagal Jantung Dengan Menggunakan Random Forest. (*Jurnal Sistem Informasi Dan Ilmu Komputer Prima*, 5(2), 176–181.
- Firmansyach, W. A., Hayati, U., Wijaya, Y. A., Studi, P., Informatika, T., Cirebon, K., Tree, D., & Validation, C. (2023). Analisa Terjadinya Overfitting Dan Underfitting Pada Algoritma Naive Bayes Dan Decision Tree Dengan Teknik Cross Validation. *Jurnal Mahasiswa Teknik Informatika*, 7(1).
- Guo, C., Wu, M., & Cheng, H. (2021). The Comprehensive Machine Learning Analytics For Heart Failure. *International Journal Of Environmental Research And Public Health*, 4(2).

- Leny, B., Nurbaety, B., & Zuhroh, H. (2020). Evaluasi Penggunaan Obat Antihipertensi Pada Pasien Gagal Jantung Rawat Jalan Di Rumah Sakit Umum Daerah Provinsi Nusa Tenggara Barat. *Lambung Farmasi; Jurnal Ilmu Kefarmasian*, 1(2), 66–72.
- Nainggolan, H. (2021). Potensi Kurkumin Dalam Mencegah Komplikasi Gangguan Jantung Pada Diabetes. *Gunadarma*, 15, 22–27.
- Raharja, A. R., Pramudianto, A., & Muchsam, Y. (2024). Penerapan Algoritma Decision Tree Dalam Klasifikasi Data " Framingham " Untuk Menunjukkan Risiko Seseorang Terkena Penyakit Jantung Dalam 10 Tahun Mendatang. *Seminar Matematika*, 2(2).
- Riany, A. F., & Testiana, G. (2023). Penerapan Data Mining Untuk Klasifikasi Penyakit Jantung Koroner Menggunakan Algoritma Naïve Bayes. *Student Conference*, 4(3), 297–305.
- Sepharni, A., Hendrawan, I. E., & Rozikin, C. (2022). Klasifikasi Penyakit Jantung Dengan Menggunakan Algoritma C4.5. *Satuan Tulisan Riset Dan Inovasi Teknologi*, 7(2).
- Tasnim, N., Mamun, S. Al, Islam, M. S., Kaiser, M. S., & Mahmud, M. (2023). Explainable Mortality Prediction Model For Congestive Heart Failure With Nature-Based Feature Selection Method. *Applied Sciences*, 2(4).
- Wan, Y., Zhang, S., & Zhao, Y. (2024). Analysis And Prediction Of Risk Factors For Heart Failure. *Proceedings Of Conf-Mla 2024*, 1(3), 168–175. <https://doi.org/10.54254/2755-2721/94/2024melb0068>