PREDIKSI PENYAKIT JANTUNG DENGAN *RANDOM FOREST* MELALUI TEKNIK IMPUTASI DAN *FEATURE SCALING* PADA *PREPROCESSING* DATA

Ainus Sya'adah[1], Aprillia Ayu Fadhilah[2], Ishaq Daulay[3]
Program Studi Teknologi, Universitas Bina Sarana Informatika[1], [2], [3]
Fakultas Teknik dan Informatika
Jakarta, Indonesia
17230850@bsi.ac.id[1], 17230876@bsi.ac.id[2], 17231056@bsi.ac.id[3]

***Abstract*— Heart disease is one of the leading causes of mortality worldwide, highlighting the need for reliable prediction methods to support early detection. In medical data analysis, challenges such as missing values and variations in feature scales are commonly encountered, making data *preprocessing* a crucial step prior to model construction. This study aims to analyze the impact of imputation techniques and *feature scaling* on the performance of the *Random Forest* algorithm in predicting heart disease. The dataset used contains missing values and varying feature scales; therefore, several *preprocessing* scenarios were implemented, including no *preprocessing*, data imputation, and a combination of imputation and *feature scaling*. *Random Forest* was employed as the classification method, and model performance was assessed using accuracy, precision, recall, and F1-score through a cross-validation approach. The results indicate that applying imputation techniques and *feature scaling* does not significantly improve model performance compared to using raw data. This finding can be attributed to the nature of *Random Forest*, which is relatively insensitive to differences in feature scales. Nevertheless, *preprocessing* remains an important step in enhancing data quality and ensuring readiness for the modeling process.**

Keywords*— *heart disease, Random Forest, data imputation, feature scaling, preprocessing.

***Abstrak*— Penyakit jantung merupakan salah satu penyebab utama tingginya angka kematian di dunia, sehingga diperlukan metode prediksi yang andal untuk mendukung upaya deteksi dini. Dalam pengolahan data medis, sering dijumpai permasalahan berupa nilai hilang serta perbedaan skala antar fitur, yang menjadikan tahap *preprocessing* sebagai bagian penting sebelum pemodelan dilakukan. Penelitian ini bertujuan untuk menganalisis pengaruh teknik imputasi dan *feature scaling* terhadap kinerja algoritma *Random Forest* dalam memprediksi penyakit jantung. Dataset yang digunakan mengandung sejumlah missing values dan memiliki variasi skala fitur, sehingga diterapkan beberapa skenario *preprocessing*, yaitu tanpa *preprocessing*, dengan imputasi data, serta kombinasi imputasi dan *feature scaling*. Algoritma *Random Forest* digunakan sebagai metode klasifikasi, sementara evaluasi kinerja model dilakukan menggunakan metrik akurasi, presisi, recall, dan F1-score melalui validasi silang. Hasil penelitian menunjukkan bahwa penerapan imputasi maupun kombinasi imputasi dan *feature scaling* tidak menghasilkan peningkatan kinerja yang signifikan dibandingkan dengan penggunaan data mentah. Hal ini berkaitan dengan karakteristik *Random Forest* yang relatif tidak sensitif terhadap perbedaan skala fitur. Meskipun demikian, *preprocessing* tetap berperan penting dalam meningkatkan kualitas dan kesiapan data sebelum tahap pemodelan.**

Kata Kunci*— *penyakit jantung, Random Forest, imputasi data, feature scaling, preprocessing.

I. PENDAHULUAN

Penyakit jantung terus menjadi penyebab kematian paling umum diseluruh dunia dan merupakan masalah kesehatan global yang membutuhkan perhatian serius. Berdasarkan laporan (WHO, 2018), penyakit kardiovaskular menyumbang sekitar 32% dari total kematian global setaip tahunnya. Angka ini menunjukkan bahwa deteksi dini risiko penyakit jantung sangat penting untuk menurunkan angka kematian. Di Indonesia, prevalensi penyakit jantung terus meningkat setiap tahunnya, sering kali dikaitkan dengan gaya hidup tidak sehat seperti merokok, kurang berolahraga, dan konsumsi makanan tinggi lemak jenuh (Wijaya, 2025). Upaya identifikasi dini dengan menggunakan teknologi kecerdasan buatan memberikan peluang besar bagi dunia medis untuk meningkatkan akurasi diagnosis serta pengambilan keputusan klinis.

Dalam bidang ilmu komputer dan kesehatan, penerapan algoritma *Machine Learning* (ML) untuk diagnosis penyakit jantung terus berkembang karena kemampuannya dalam memproses data yang besar dan kompleks. *Random Forest* (RF) adalah salah satu algoritma yang paling umum digunakan karena sangat efektif dalam mengelola data yang tidak memiliki hubungan linear dan membantu mengurangi risiko overfitting (Tamba & -, 2022). Metode *Random Forest* dapat mengklasifikasikan penyakit jantung dengan akurasi lebih dari 85% (Alfajr & Defiyanti, 2024). Model *Explainable Boosting Machine* (EBM) dengan langkah *preprocessing* yang baik memiliki akurasi yang lebih tinggi dibandingkan dengan model tanpa *preprocessing*. Hasil menunjukkan bahwa dalam industri medis, peningkatan kinerja algoritma klasifikasi dibantu oleh proses pengolahan data yang lebih baik.

Meskipun demikian, salah satu tantangan utama dalam penerapan algoritma *Machine Learning* (ML) pada data medis adalah kualitas data yang digunkan. Data medis sering kali menghadapi masalah seperti nilai yang hilang, perbedaan skala atribut, serta ketidakseimbangan jumlah data antar kelas yang dapat menurunkan performa model klasifikasi (Gullam Almuzadid & Egia Rosi Subhiyakto, 2025). Masalah ini sangat penting karena algoritma *Machine Learning* sangat bergantung pada struktur dan kualitas data yang diberikan. Karena itu, diperlukan beberapa langkah *preprocessing* seperti imputasi data, mengubah data ke bentuk standar, dan *feature scaling* agar data menjadi lebih representatif dan memenuhi persyaratan sebelum model dilatih (Faradeya & Subhiyakto, 2025).

Algortima pembelajaran mesin seperti *Support Vector Machine, Decision Tree*, dan *Random Forest* sangat efektif dalam diagnosis penyakit jantung menurut penelitian sebelumnya. Pengaruh tahapan *preprocessing* terhadap keberhasilan model belum banyak diteliti (Gullam Almuzadid & Egia Rosi Subhiyakto, 2025). Sebagai contoh, penelitian (Prakash et al., 2025) menemukan bahwa kombinasi imputasi MICE dan StandardScaler dapat meningkatkan akurasi klasifikasi risiko stroke hingga 98%. Di sisi lain, penelitian (Alsyar et al., 2025) menemukan bahwa penyesuaian fitur dapat meningkatkan kemampuan model *Random Forest* untuk memprediksi diabetes mellitus. Hasil menunjukkan bahwa *preprocessing* adalah komponen penting yang dapat meningkatkan keandalan model prediktif dan bukan hanya langkah teknis awal. Penelitian serupa yang secara komprehensif mempelajari kombinasi metode imputasi dan *feature scaling* untuk memprediksi penyakit jantung masih jarang ditemukan (Setiawan & Efendi, 2025).

Sebuah pendekatan eksploratif dalam teknik *preprocessing* diperlukan untuk kualitas data medis yang beragam dan kompleks. Untuk menangani nilai yang hilang, beberapa metode imputasi seperti rata-rata (*imputation mean*), median, dan *K-Nearest Neighbors* (KNN) dapat digunakan. Dalam proses ini, teknik *feature scaling* seperti *Min-Max* dan *StandardScaler* juga membantu memperbaiki rentang nilai setiap atribut, sehingga model dapat mencapai hasil yang stabil lebih cepat (Arman et al., 2025). Menurut (Wijaya, 2025), melakukan tahap *preprocessing* dengan benar dapat meningkatkan akurasi model klasifikasi hingga 10% dibandingkan dengan model tanpa *preprocessing*. Oleh karena itu, untuk mendapatkan model prediktif yang optimal ekplorasi dan evaluasi berbagai kombinasi teknik *preprocessing* menjadi lebih penting.

Dengan demikian, fokus penelitian ini adalah untuk mengevaluasi bagaimana kombinasi metode imputasi data dan *feature scaling* berdampak pada hasil klasifikasi penyakit jantung menggunakan algoritma *Random Forest*. Diharapkan penelitian ini akan meningkatkan pemahamin kami tentang pentingnya tahapan dalam meningkatkan akurasi prediksi penyakit jantung. Selain itu, hasil penelitian diharapkan dapat berguna bagi pengembang sistem pendukung keputusan berbasis AI, yang dapat membantu tenaga medis melakukan diagnosis dengan cepat dan akurat (Arman et al., 2025).

II. STUDI PUSTAKA

Machine Learning (ML) telah menjadi fondasi utama dalam pengembangan sistem pendukung keputusan klinis karena kemampuannya mengekstraksi pola tersembunyi dari data kesehatan yang kompleks dan non-liniear. Dalam konteks diagnosis penyakit, *Machine Learning* digunakan untuk memprediksi risiko klinis, mendeteksi kondisi secara dini, serta meningkatkan akurasi pengambilan keputusan medis. Tantangan utama dalam pemanfaatan *Machine Learning* pada data kesehatan meliputi variasi data antar pasien, ketidakseimbangan kelas, serta adanya noise missing values. Oleh karena itu, tahapan *preprocessing* dapat menyebabkan kesalahan interpretasi dan menurunkan performa model secara signifikan (Ren et al., 2024).

Random Forest (RF) merupakan algoritma klasifikasi berbasis *ensemble learing* yang menggabungkan banyak decisiom tree melalui bagging untuk menghasilkan prediksi akhir melalui votting mayoritas. *Random Forest* banyak digunakan dalam bidang medis karena ketahanannya terhadap overfitting, kemampuannya menangani data numerik maupun kategorikal, serta stabilitas performanya pada dataset klinis yang berskala menengah. Meskipun relatif kuat terhadap variasi skala data, interpretabilitas *Random Forest* lebih rendah dibandingkan model linear sehingga sering dilengkapi dengan teknik interpretasi tambahan seperti feature importance dan SHAP (Shapley Additive exPlanations). Selain itu, kinerja model *Random Forest* sangat bergantung pada pengaturan parameter tambahan seperti n_estimators, max depth, dan min samples leaf (Faradeya & Subhiyakto, 2025).

Preprocessing data adalah langkah fundamental sebelum model dilatih dan mencakup pembersihan data, standarisasi format, penghapusan duplikasi, normalisasi, serta identifikasi *outlier*. Pada domain rekam medis elektronik (EHR) *Electronic Heart Record*, *preprocessing* juga mencakup konsolidasi antar sumber data karena perbedaan pencatatan antar fasilitas kesehatan sering menhasilkan ketidakkonsistenan. Studi komputasional menunjukkan bahwa kesalahan minor dalam *preprocessing* dapat menurunkan akurasi prediksi hingga 25%, terutama oada data medis dengan tingkat heterogenitas tinggi (Torthi et al., 2024).

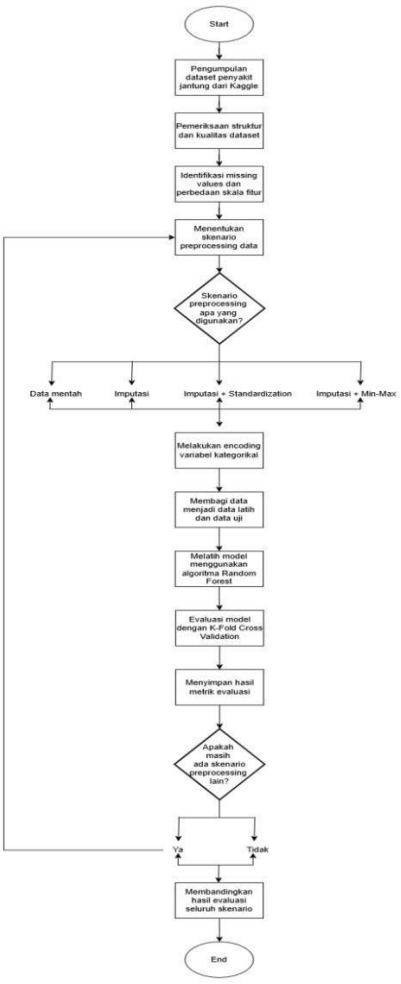
Penanganan missing values merupakan komponen penting dalam *preprocessing*. Teknik-teknik imputasi dasar seperti median dan mode sering digunakan karena sederhana dan stabil, namun metode ini dapat menghilangkan korelasi fitur. Sebaliknya, metode imputasi berbasis model seperti KNN-Imputer dan *Multiple Imputation by Chained Equations* (MICE) mampu mempertahankan hubungan antar variabel sehingga menghasilkan estimasi yang lebih representatif. Dalam data klinis dengan struktur kompleks, teknik seperti MICE lebih stabil dibandingkan imputasi sederhana, meskipun membutuhkan waktu komputasi lebih besar (Aracri et al., 2025).

Feature scaling merupakan proses penting untuk memastikan rentang nilai antar fitur tetap seimbang sehingga tidak ada fitur tertentu yang mendominasi proses pembelajaran model. Dua teknik yang paling umum adalah Standarizatiom (Z-score) dan Min-Maz Normalization. Walaupun algoritma berbasis decision tree seperti *Random Forest* tidak sensitif terhadap perubahan skala, beberapa penelitian melaporkan bahwa scaling tetap memberikan peningkatan performa terutama ketika digunakan bersamaan dengan teknik imputasi berbasis jarak (KNN). Penelitian empisis menunjukkan bahwa normalisasi mampu meningkatkan stabilitas dan akurasi model pada domain kesehatan (Pinheiro et al., 2025).

Pengaruh *preprocessing* terhadap performa model klasifikasi medis telah menjadi fokus berbagai penelitian. Meta-analisis menunjukkan bahwa gabungan antara imputasi yang tepat dan *feature scaling* dapat meningkatkan akurasi, recall, F1-score, serta kemampuan generalisasi model. Efektivitas peningkatan ini sangat bergantung pada karakteristik dataset seperti proporsi missing values dan kompleksitas fitur. Selain itu, menguji model dalam berbagai skenario *preprocessing* daripada hanya melakukan uji satu kali adalah cara yang lebih baik untuk mengevaluasi pengaruh setiap teknik (Alfajr & Defiyanti, 2024).

Penelitian terdahulu juga menunjukkan bagaimana metode tertentu memberikan kontribusi spesifik dalam *preprocessing*. Menurut (Setiawan & Efendi, 2025), metode *Random Forest* bisa memberikan hasil diagnosis yang cukup akurat, tetapi tingkat keberhasilannya sangat tergantung pada kualitas data yang digunakan. Studi lain mengenai data medis dengan tingkat missing tinggi menegaskan bahwa pemilihan teknik imputasi yang tepat sangat berpengaruh pada peningkatan reliabilitas dataset (Rasheed et al., 2024). Penelitian komparatif dalam (Seu et al., 2022) menemukan bahwa metode imputasi MICE dan KNN lebih unggul dibandingkan imputasi sederhana. Meskipun berbagai penelitian telah mengevaluasi *preprocessing*, sebagian besar hanya meninjau satu teknik secara terpisah. Oleh karena itu, masih terdapat celah penelitian karena belum banyak studi yang mengeksplorasi kombinasi teknik imputasi dan *feature scaling* secara terintegrasi dalam konteks prediksi penyakit jantung menggunakan *Random Forest*.

III. METODE PENELITIAN



A. Metode dan Alur Penelitian

Metode kuantitatif yang digunakan dalam penelitian ini adalah eksperimen komputasional. Metode ini diterapkan melalui serangkaian tahapan sistematis untuk mengevaluasi pengaruh Teknik Imputasi dan *Feature scaling* pada tahap *preprocessing* data terhadap kemampuan algoritma *Random Forest* dalam memprediksi penyakit jantung. Penelitian ini berfokus pada pengujian berbagai skenario perlakuan *preprocessing* terhadap data yang sama. Oleh karena itu, pendekatan eksperimen komputasional dipilih untuk memungkinkan pengamatan yang objektif dan terukur dari perbedaan kinerja model. Proses yang benar-benar dilakukan dalam penelitian ini membentuk jalur penelitian: pengumpulan data, *preprocessing*, pemodelan menggunakan *Random Forest*, dan evaluasi kinerja model. Analisis perbandingan performa model sebelum dan sesudah penerapan teknik imputasi dan *feature scaling* adalah tahapan kerja penelitian ini, bukan hasil dari penelitian sebelumnya. Untuk memulai penelitian, dataset penyakit kardiovaskular dikumpulkan dari repositori Kaggle dalam format CSV. Kemudian dataset ini diperiksa untuk memastikan konsistensi, yang mencakup pengecekan duplikasi, kesalahan format, dan nilai yang tidak valid. Selanjutnya proses *preprocessing* data dilakukan: fokus utama penelitian ini adalah imputasi nilai missing values dan penerapan *feature scaling* dengan berbagai teknik. Untuk menghindari bias dalam proses pembelajaran model, data yang telah diproses dibagi menjadi data latih dan data uji. Setelah itu, beberapa skenario *preprocessing* yang telah ditentukan digunakan untuk menguji model *Random Forest*. Evaluasi performa model dengan beberapa metrik evaluasi adalah tahap akhir penelitian. Tujuannya adalah untuk membandingkan hasil prediksi sebelum dan sesudah *preprocessing*.

B. Dataset Penelitian

Dataset *Cardiovascular Disease* yang digunakan dalam penelitian ini diperoleh dari repositori Kaggle dan terdiri dari sekitar 70.000 data observasi yang mewakili kondisi demografis, fisiologis, dan gaya hidup individu. Data disajikan dalam format CSV untuk mendapatkan pengolahan dan analisis.

Pada dataset, atribut terdiri dari kedua metrik numerik dan kategorik. Usia, tinggi, berat, ap_hi, ap_lo, BMI, dan denyut jantung adalah metrik numerik, sedangkan metrik kategorik mencakup gender, cholesterol, gluc, smoke, alco, active, dan obesitas. Dalam penelitian ini, variabel target adalah kardiovaskular yang diberi nilai nol untuk individu tanpa indikasi penyakit. Dataset ini cocol untuk menguji efektivitas teknik imputasi dan pengukuran fitur dalam meningkatkan kinerja model klasifikasi karena memiliki karakteristik nmissing values dan perbedaan rentang skala antara fitur numerik.

C. Variabel Penelitian

Variabel independen adalah semua atribut dataset yang digunakan sebagai input model, variabel dependen adalah semua atribut numerik dan kategorikal. Variabel-variabel ini menunjukkan variabel yang dapat mempengaruhi kondisi kesehatan kardiovaskular seseorang. Variabel kardio yang berfungsi sebagai label kelas dalam proses klasifikasi, adalah variabel dependen penelitian ini. Hasil prediksi model *Random Forest*, seperti diagnosis penyakit jantung, ditentukan oleh variabel ini.

D. Tahap Preprocessing Data

Preprocessing data diperlukan untuk memastikan kualitas data yang digunakan dalam proses pemodelan sehingga model *Random Forest* dapat bekerja dengan baik. Ini diperlukan karena dataset yang digunakan memiliki fitur seperti missing values dan perbedaan rentang skala antar fitur numerik, yang jika tidak ditangani dengan tepat dapat memengaruhi kinerja model.

Fokus *preprocessing* penelitian adalah dua proses utama: imputasi dan pengembangan fitur. Imputasi missing values dilakukan untuk mengatasi ketidaklengkapan data sehingga seluruh atribut dapat digunakan seacara utuh selama proses pelatihan model. Pengembangan fitur

dilakukan untuk menyeragamkan skala fitur numerik sehingga tidak ada dominasi fitur tertentu karena perbedaan rentang nilai. Sebelum tahap pemodelan, kedua proses ini dilakukan untuk membandingkan kinerja model sebelum dan sesudah *preprocessing*, skenario eksperimen disusun.

1) Imputasi Data

Pada tahap imputasi, kumpulan data diperiksa untuk menemukan atribut yang memiliki nilai hilang (*missing values*). Missing values dapat menyebabkan lebih sedikit data yang dapat digunakan dalam proses pelatihan model, dan apabila tidak ditangani dengan benar, ini dapat menyebabkan prediksi menjadi kurang. Oleh karena itu, sebelum tahap pemodelan, proses imputasi dilakukan menggunakan algoritma *Random Forest*. Metode imputasi dalam penelitian ini disesuaikan dengan jenis data pada masing-masing karakteristik.

Imputasi data digunakan untuk mengatasi *missing values* pada fitur numerik, sedangkan imputasi median digunakan untuk mengatasi nilai hilang pada fitur kategorikal karena median lebih tahan terhadap pengaruh nilai ekstrem (*outlier*) dibandingkan nilai rata-rata, sehingga media dianggap lebih sesuai untuk data medis yang memiliki variasi nilai yang cukup besar. Dengan menggunakan proses imputasi ini, semua atribut kumpulan data dapat digunakan secara menyeluruh selama tahap pemodelan, tanpa menghilangkan data yang mungkin penting dan telah diamati. Selain itu, karakteristik distribusi data asli tetap dipertahankan.

2) *Feature scaling*

Feature scaling digunakan untuk menyeragamkan skalar antar fitur numerik dengan rentang nilai yang berbeda. Dalan dataset penyakit jantung, fitur numerik seperti usia, tekanan darah, dan indeks massa tubuh memiliki skala dan satuan yang berbeda, perbedaan skala ini dapat menyebabkan model memprioritaskan fitur dengan nilai numerik yang lebih besar. Namun, fitur-fitur ini tidak selalu lebih penting. Dalam penelitian ini, dua metode p*feature scaling* digunakan: Standarization (Z-score) dan Min-Max Normalization. Kedua metode ini diterapkan pada berbagai skenario eksperimen untuk mengevaluasi pengaruh mereka terhadap kinerja model *Random Forest*.

Tujuan Standarization adalah untuk mengubah fitur numerik dengan nilai rata-rata sebesar 0 dan standar deviasi sebesar 1. Teknik ini membantu menormalkan distribusi data dan mengurangi perbedaan skala antar fitur. Berikut ini adalah rumus Transformasi Standarization:

$$x_i' = \frac{x_i - \mu}{\sigma}$$

Di mana x_i merupakan nilai asli fitur ke i, μ nilai rata-rata fitur, σ adalah standar deviasi fitur, dan x'_i merupakan nilai hasil standarisasi.

Selain itu, Min-Max Normalization digunakan untuk mentransformasikan nilai fitur ke dalam rentang 0 hingga 1. Teknik ini mempertahankan bentuk distribusi data asli, namun menyamakan rentang nilai antar fitur sehingga lebih mudah diproses oleh model. Rumus Min-Max Normalization ditunjukkan sebagai berikut:

$$x_i' = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

Di mana x_{min} dan x_{max} masing-masing adalah nilai minimum dan maksimum dari fitur, serta x'_i merupakan hasil normalisasi.

Penerapam kedua teknik *feature scaling* ini memungkinkan dilakukan perbandingan performa model *Random Forest* sebelum dan sesudah penyeragaman skala data, sehingga pengaruh masing-masing teknik *preprocessing* dapat dianalisis secara sistematis.

E. Pemodelan Menggunakan Random Forest

Dalam penelitian ini, algoritma *Random Forest* digunakan sebagai metode klasifikasi untuk memprediksi penyakit jantung. Metode ensemble ini membangun sejumlah pohon keputusan secara acak dan menggunakan mekanisme suara mayoritas untuk menentukan kelas akhir untuk menggabungkan hasil prediksi dari setiap pohon. Metode ensemble ini memungkinkan model menghasilkan prediksi yang lebih stabil dibandingkan dengan metode satu pohon keputusan tunggal.

Selama proses pemodelan, *Random Forest* dilatih dengan data hasil *preprocessing* sesuai dengan skenario eksperimen yang telah ditetapkan. Dalam *Random Forest*, setiap pohon keputusan dibangun dengan menggunakan subset data dan fitur yang dipilih secara acak. Ini memungkinkan variasi antar pohon dan mengurangi risiko overfitting. Mekanisme ini sangat relevan untuk dataset medis yang memiliki banyak variabel individu.

Untuk menjaga kestabilan model dan memastikan bahwa hasil eksperimen dapat diulang, parameter *Random Forest* diatur pada nilai tertentu. Jumlah pohon keputusan yang dibangun dalam model dapat dihitung dengan menggunakan parameter n_estimators, lebih banyak pohon cenderung meningkatkan kestabilan prediksi. Untuk mencegah model menjadi terlalu kompleks, parameter max_depth digunakan untuk membatasi kedalaman maksimum setiap pohon. Selain itu, parameter min_samples_split dan min_samples_leaf digunakan untuk mengontrol jumlah data minimal pada proses pemisahan node dan node daun, sehingga struktur pohon tidak terbentuk secara berlebihan. Untuk memastikan bahwa hasil pelatihan model konsisten pada setiap skenario eksperimen, parameter random_state juga digunakan.

Tujuan konfigurasi parameter ini adalah untuk menghasilkan model *Random Forest* yang seimbang antara kemampuan prediksi dan kompleksitas model. Dengan cara ini, perbedaan performa yang dihasilkan pada setiap skenario eksperimen dapat lebih menunjukkan pengaruh teknik *preprocessing* daripada variasi parameter model.

F. Skenario Eksperimen

Dalam penelitian ini, empat skenario eksperimen disusun secara bertahap untuk melihat bagaimana metode *preprocessing* berpengaruh terhadap kinerja model *Random Forest* dalam prediksi penyakit jantung. Setiap skenario disusun dengan cara ini sehingga kontribusi masing-masing teknik *preprocessing*, yaitu imputasi data dan *feature scaling* dapat diamati secara terpisah dan secara kombinatif.

Dalam skenario pertama, data digunakan dalam kondisi awal sebagaimana diperoleh dari sumber dataset tanpa menggunakan imputasi atau *feature scaling*. Skenario ini berfungsi sebagai dasar, sehingga kinerja model *Random Forest* pada kondisi tanpa *preprocessing* dapat digunakan sebagai acuan utama dalam evaluasi dan perbandingan hasil eksperimen berikutnya.

Tujuan skenario kedua adalah untuk melihat pengaruh penanganan missing values terhadap performa model secara langsung. Skenario ini menggunakan teknik imputasi median untuk fitur numerik dan imputasi modus untuk fitur kategorikal, tetapi tidak melakukan *feature scaling*. Hasil skenario kedua dibandingkan dengan baseline untuk menunjukkan seberapa besar proses imputasi membantu meningkatkan kualitas data dan prediksi model.

Dalam skenario ketiga, imputasi median-modus diterapkan yang diikuti dengan *feature scaling* berupa Standarization (Z-score). Pada skenario ini, data dibersihkan dari missing values sebelum Standarization dilakukan untuk menyeragamkan skala antar fitur numerik. Tujuan dari skenario ini adalah untuk mengevaluasi bagaimana kombinasi Imputasi dan Standarization berdampak pada kinerja model *Random Forest*.

Dalam skenario keempat, imputasi median-modus digunakan dan diikuti dengan Normalization Min-Max, sebuah teknik *feature scaling* yang mengubah nilai fitur ke dalam rentang 0-1. Tujuan dari skenario ini adalah untuk membandingkan efektivitas Normalization Min-Max terhadap Standarization dalam konteks pemodelan *Random Forest* pada dataset penyakit jantung.

Secara objektif perbandingan kinerja model *Random Forest*, keempat skenario eksperimen tersebut digunakan. Hasil yang berbeda dari masinh-masing skenario dianalisis untuk menentukan metode *preprocessing* yang paling efektif untuk meningkatkan kinerja prediski penyakit jantung.

G. Evaluasi Model

Untuk menilai kemampuan algoritma Random Fprest untuk memprediksi penyakit jantung pada setiap skenario eksperimen yang telah ditetapkan, evaluasi performa model dilakukan, Dalam penelitian ini, metode *K-Fold Cross Validation* dengan jumlah lipatan sebanyak lima (K=5). Metode ini membagi dataset menjadi lima bagian yang sama. Pada setiap iterasi, satu bagian digunakan sebagai data uji dan empay bagian lainnya digunakan sebagai data latih. Proses ini dilakukan secara bergantian hingga seluruh dataset digunakan sebagai data uji.

Tujuan penggunaan *K-Fold Cross Validation* adalah untuk menghasilkan hasil evaluasi yang lebih stabil dan tidak bias serta mengurangi ketergantungan hasil pada satu bagian data. Metode ini menunjukkan kemampuan model secara umum dengan menghasilkan perofrma model sebagai nilai trata-rata dari seluruh proses pengujian.

Dalam penelitian ini, metrik evaluasi yang digunakan termasuk akurasi, presisi, recall, *F1-Score*, dan *matrix confusion*. Akurasi mengukur proporsi prediksi yang benar terhadap seluruh data uji, sedangkan presisi mengukur tingkat ketepatan model dalam memprediksi kelas positif.

Recall menjadi metrik penting dalam prediksi penyakit jantung karena berkaitan dengan kemampuan model untuk mengidentifikasi pasien dengan indikasi penyakit jantung. F1-score memberikan gambaran yang lebih baik tentang performa model, terutama dalam kasus ketidakseimbangan kelas, karena digunakan sebagai ukuran keseimbangan antara presisi dan recall. Untuk melihat kesalahan prediksi lebih lanjut, *confusion matrix* digunakan untuk menganalisis distribusi hasil prediksi model berdasarkan nilai *true positive, true negative, false positive*, dan *valse negative*. Selanjutnya hasil evaluasi dari setiap skenario eksperimen dibandingkan untuk melihat bagaimana penerapan metode imputasi dan *feature scaling* berdampak pada kemmapuan model *Random Forest* untuk memprediksi penyakit jantung.

IV. PEMBAHASAN DAN HASIL

A. Eksplorasi Data Awal

Tahap eksplorasi data awal dilakukan menggunakan dataset Heart Disease UCI yang diperoleh dari Kaggle dalam format CSV. Dataset ini terdiri dari 920 baris data dan 16 kolom, sehingga cukup representatif untuk pengujian model *Random Forest*. Pada tahap ini, data masih berada dalam kondisi mentah dan belum mengalami proses *preprocessing* apapun. Pemeriksaan awal difokuskan pada struktur data untuk mengidentifikasi jenis atribut yang dimiliki, baik numerik maupun katgeorikal. Hasil eksplorasi menunjukkan bahwa dataset memiliki kombinasi tipe data integer, float, dan object yang perlu diperhatikan pada tahap pengolahan selanjutnya dengan analisis statistik deskriptif dilakukan pada fitur numerik untuk melihat rentang dan sebaran nilai dimana ditemukan perbedaan skala yang cukup signifikan antar fitur seperti tekanan darah dan kadar kolesterol. Selain itu, pemeriksaan *Missing value* menunjukkan bahwa beberapa atribut mengandung nilai kosong dalam jumlah yang cuku besar, khususnya pada fitur ca, thal, dan

slope. tersebut teridentifikasi dalam bentuk NaN, bukan simbol khusus sehingga dapat ditangani secara langsung melalui teknik imputasi. Seluruh proses pada tahap ini bertujuan untuk memperoleh pemahaman awal terhadap karakteristik data dan menjadi dasar penentuan strategi *preprocessing* pada tahap berikutnya.

1) *Gambaran Umum Data*

Dataset Penyakit Jantung digunakan dalam penelitian ini untuk menyimpan data klinis pasien tentang faktor risiko penyakit jantung. Dataset terdiri dari 920 baris data (sampel) dan 16 kolom (fitur), yang masing-masing terdiri dari satu variabel target dan fitur numerik dan kategorikal, menurut hasil eksplorasi awal yang dilakukan menggunakan Google Colab.

Distribusi tipe data pada dataset ditunjukkan sebagai berikut:

- Fitur numerik: *age, trestbps, chol, thalch, oldpeak, ca.*
- Fittur kategorikal: *sex, dataset, cp, fbs, restecg, exang, slope, thal.*
- Variabel target: *num.*

Variabel target *num* menunjukkan kondisi penyakit jantung: nilai 0 menunjukkan pasien tidak memiliki penyakit jantung, dan nilai 1-4 menunjukkan tingkat keparahan penyakit jantung. Untuk keperluan pemodelan, variabel target diubah menjadi klasifikasi biner.

2) *Identifikasi Masalah pada Data*

Pada tahap eksplorasi, pemeriksaan kualitas data dilakukan untuk mengidentifikasi masalah potensial yang dapat memengaruhi hasil pemodelan. Menurut pemeriksaan, dataset mengandung nilai hilang pada beberapa fitur numerik dan kategorikal.

Fitur	**Jumlah Missing**
sex	0
dataset	0
cp	0
Fbs	90
restecg	2
exang	55
slope	309
thal	486

Tabel 1 Jumlah Missing value pada Fitur Katgeorikal

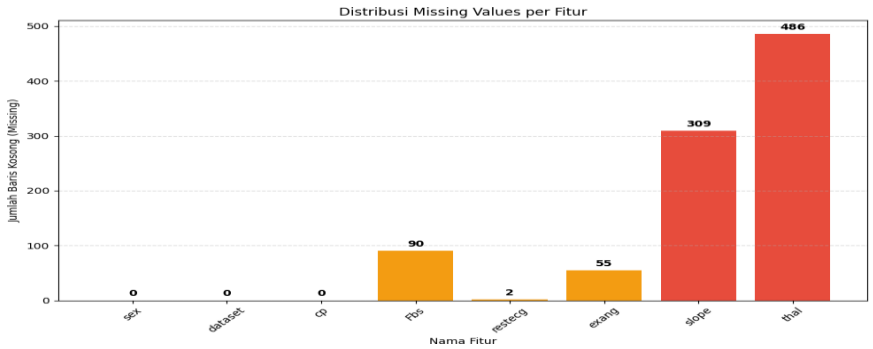


Diagram 1 Jumlah Missing Value pada Fitur Katgeorikal

Fitur	Jumlah Missing
age	0
trestbps	59
chol	30
thalch	55
oldpeak	62
ca	611

Tabel 2 Jumlah Missing value pada Fitur Numerik

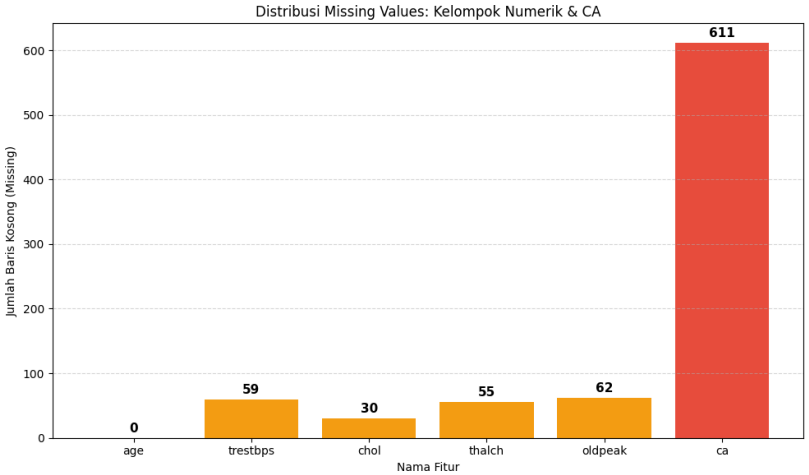


Diagram 2 Jumlah Missing Value pada Fitur Numerik

Hasil menunjukkan bahwa dataset belum sepenuhnya bersih dan proses imputasi data diperlukan sebelum digunakan dalam pemodelan. Apabila tidak ditangani dengan baik, *Missing value* yang cukup signifikan terutama pada atribut *ca* dan *thal* dapat mengurangi kinerja model.

3) *Ukuran Pemusatan Data*

Nilai tengah distribusi data untuk setiap fitur numerik dihitung dengan menggunakan ukuran pemusatan data. Nilai rata-rata (*mean*) dan median yang diperoleh dari statistik deskriptif dataset sebelum *preprocessing* dimasukkan dalam analisis ini.

Fitur	Mean	Median
age	53.51	54
trestbps	132.13	130
chol	199.13	223
thalch	137.55	140
oldpeak	0.88	0.0
ca	0.68	0

Tabel 3 Ukuran Pemusatan Data Fitur Numerik (sebelum preprocessing)

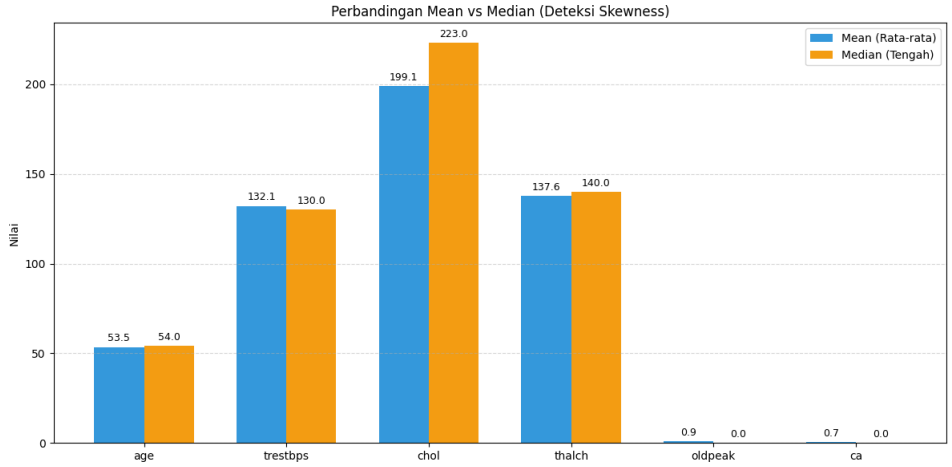


Diagram 3 Ukuran Pemusatan Data Fitur Numerik (sebelum preprocessing)

Adanya *Missing value* antara median dan mean untuk fitur tertentu seperti *chol* dan *oldpeak* menunjukkan distribusi data yang tidak simetris dan kemungkinan adanya pencilan (*outlier*).

4) *Ukuran Penyebaran Data*

Nilai standar deviasi, nilai minimum, dan nilai maksimum digunakan dalam analisis ini untuk mengukur variasi data pada setiap fitur.

Fitur	Std	Min	Max
age	9.42	28	77
trestbps	19.07	0	200

chol	110.78	0	603
thalch	25.93	60	202
oldpeak	1.09	-2.6	6.2
ca	0.94	0	3

Tabel 4 Ukuran Penyebaran Data Fitur Numerik

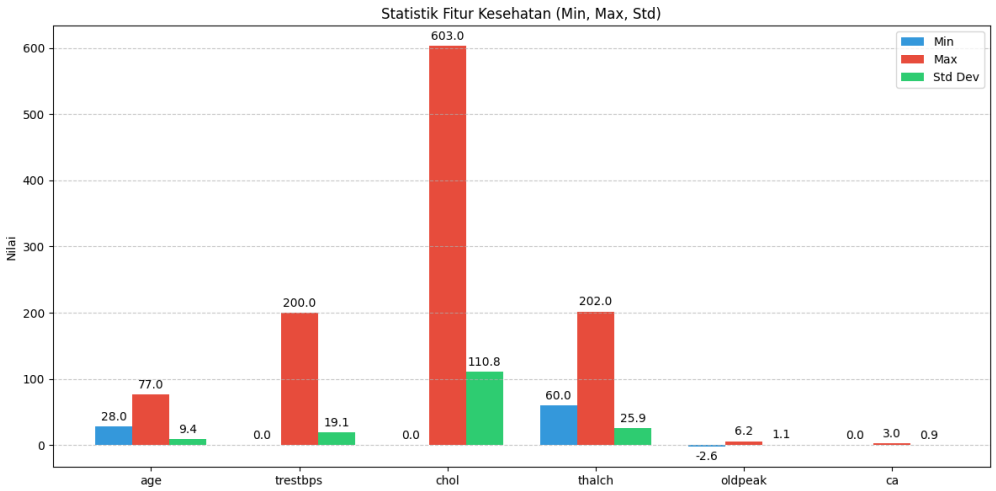


Diagram 4 Ukuran Penyebaran Data Fitur Numerik

Adanya variasi data yang tinggi serta kemungkinan pencilan (*outlier*) yang dapat memengaruhi proses pelatihan model, ditunjukkan oleh rentang yang cukup besar pada fitur *chol* dan *oldpeak.*

5) *Distribusi Data*

Analisis distribusi data dilakukan untuk mengetahui bentuk sebaran data apakah cenderung normal atau memiliki kemiringan (*skewness*). Hasil eksplorasi statistik serta pengamatan nilai minimun dan maksimum menunjukkan bahwa beberapa fitur numerik menunjukkan distribusi yang tidak normal dan cenderung skewed terutama pada fitur *chol*, *oldpeak*, dan *ca.*

Untuk meningkatkan skalabilitas dan konsistensi data sebelum proses pemodelan, kondisi distribusi data yang tidak seragam ini dipertimbangkan saat menggunakan teknik *preprocessing* seperti standarization dan imputasi missing value..

B. Hasil Preprocessing Data

Sebelum digunakan dalam proses pemodelan, tahapan *preprocessing* data yang bertujuan untuk meningkatkan kualitas dataset dibahas dalam bab ini. *Preprocessing* dilakukan untuk mengatasi masalah umum dengan data mentah. Ini termasuk variabel kategorikal yang algoritma pemodelan tidak dapat langsung memproses, perbedaan skala antara fitur numerik, dan missing values. Setiap tahapan proses *preprocessing* termasuk imputasi missing values, penskalaan fitur numerik, dan encoding variabel kategorikal berkontribusi pada peningkatan integritas data, konsistensi struktur dataset, dan kesiapan data untuk memenuhi kebutuhan algoritma yang digunakan. Hasil *preprocessing* dievaluasi untuk memastikan bahwa tidak ada informasi penting yang hilang dan untuk memastikan bahwa data yang dihasilkan lebih representatif dan stabil.

1) *Ringkasan Perubahan Data (Data Overview)*

Untuk memberikan gambaran umum tentang kondisi dataset sebelum dan sesudah *preprocessing*, ringkasan perubahan data diberikan. Ringkasan ini berfungsi sebagai tahap

awal evaluasi untuk melihat dampak *preprocessing* secara makro, sebelum analisis lebih lanjut dilakukan pada setiap tahap.

Dataset masih mengalami beberapa masalah sebelum *preprocessing*. Ini termasuk nilai kosong pada beberapa atribut, perbedaan rentang nilai yang cukup besar antara fitur numerik, dan variabel kategorikal yang masih berbentuk teks. Jika tidak ditangani dengan benar, kondisi ini dapat menurunkan kinerja model dan menyebabkan bias dalam proses analisis.

Dataset mengalami perubahan nilai kualitas dan struktur setelah *preprocessing*. Proses imputasi mengatasi nilai yang hilang sehingga data tidak lagi memiliki missing values. Penskalaan fitur numerik memungkinkan rentang nilai yang lebih seragam, yang membantu algoritma pemodelan bekerja dengan lebih baik. Selain itu, proses encoding telah digunakan untuk mengubah variabel kategorikal menjadi angka, yang memungkinkan seluruh atribut dataset diproses secara komputasional.

Dalam tabel ringkasan, kondisi dataset sebelum dan sesudah *preprocessing* dibandingkan, termasuk jumlah data, atribut, nilai kosong, dan perubahan tipe data. Ringkasan ini menunjukkan bahwa *preprocessing* berhasil meningkatkan kesiapan dataset untuk tahap pemodelan dan analisis.

Parameter	**Sebelum *Preprocessing***	**Sesudah *Preprocessing***
Jumlah Baris	920	920
Jumlah Fitur	16	22
Missing Values	Ada, pada fitur numerik & kategorikal	0
Duplikasi Data	Tidak terdeteksi	Tidak ada
Tipe Data	Campuran numerik dan kategorikal	Seluruh fitur numerik

Tabel 5 Kondisi Data Sebelum dan Sesudah Preprocessing

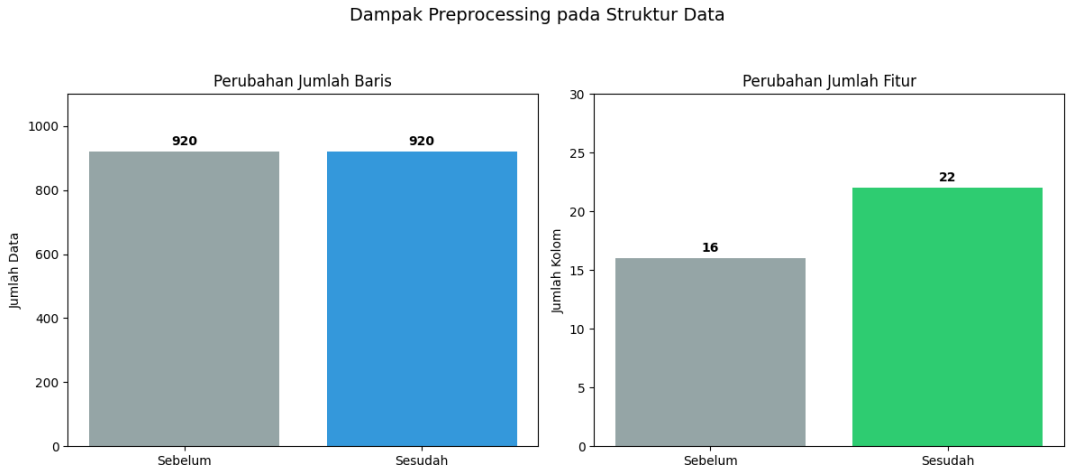


Diagram 5 Kondisi Data Sebelum dan Sesudah Preprocessing

Jumlah baris data tidak mengalami perubahan karena pada penelitian ini tidak dilakukan penghapusan observasi. Namun, jumlah fitur mengalami peningkatan dari 15 menjadi 22 sebagai dampak dari penerapan *One-Hot Encoding* pada variabel kategorikal. Selain itu, seluruh nilai yang hilang (*missing values*) telah berhasil ditangani melalui proses imputasi, sehingga dataset akhir tidak lagi mengandung nilai kosong dan siap digunakan pada tahap pemodelan.

2) Hasil Imputasi Data

Dalam penelitian ini, nilai median untuk fitur numerik dan nilai modus untuk fitur kategorikal digunakan, dataset awal mengandung nilai yang tidak ada pada fitur numerik dan kategorikal. Metode ini dipilih karena kemampuan untuk mengisi nilai kosong tanpa mengubah distribusi data secara signifikan. Ini terutama berlaku untuk data dengan nilai ekstrem atau tidak berdistribusi normal.

Setelah *preprocessing* selesai, perbandingan statistik deskriptif sebelum dan sesudah imputasi dilakukan untuk memastikan bahwa proses imputasi tidak mengubah karakteristik data secara substansial. Tujuan dari perbandingan ini adalah untuk menilai konsistensi nilai pusat dan sebaran data setelah *preprocessing* selesai.

Fitur	Mean Sebelum	Mean Sesudah	Std Dev Sebelum	Std Dev Sesudah
age	53.51	53.50	9.42	9.41
trestbps	132.13	132.11	19.07	19.02
chol	199.13	199.10	110.78	110.65
thalch	137.55	137.53	25.93	25.90
oldpeak	0.88	0.87	1.09	1.08
ca	0.68	0.67	0.94	0.93

Tabel 6 Perbandingan Statistik Fitur Numerik Sebelum dan Sesudah Imputasi

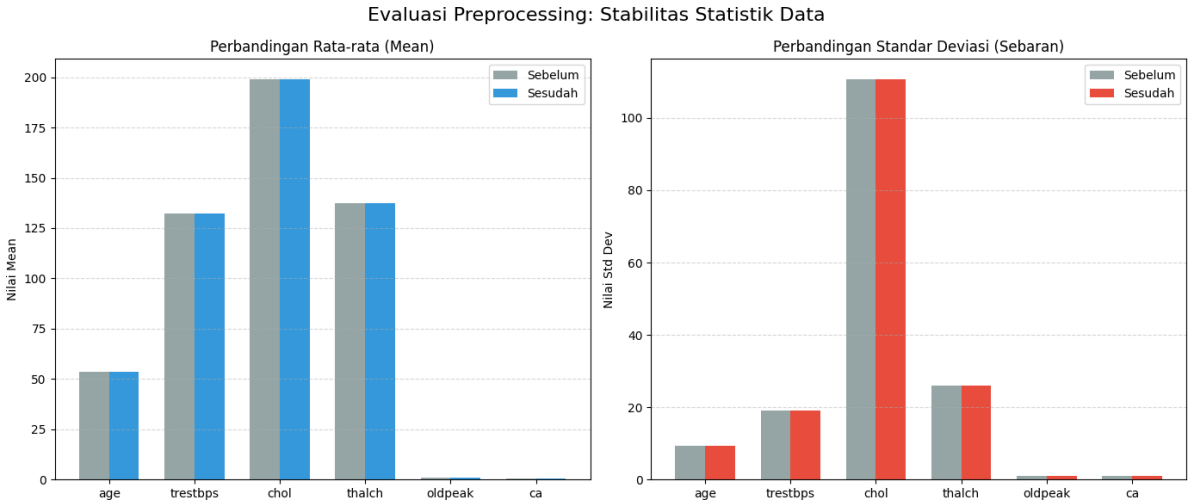


Diagram 6 Perbandingan Statistik Fitur Numerik Sebelum dan Sesudah Imputasi

Nilai rata-rata dan standar deviasi sebelum dan sesudah imputasi tidak menunjukkan pergeseran yang ekstrem. Hal ini mengindikasikan bahwa metode imputasi yang diterapkan mampu mengisi nilai kosong tanpa mengubah karakteristik utama data, khususnya dalam hal kecenderungan pusat dan tingkat sebaran.

3) Hasil Penskalaan Data (Feature scaling)

Setelah proses imputasi, dilakukan penskalaan terhadap fitur numerik menggunakan metode *StandardScaler*. Metode ini mentransformasikan data sehingga setiap fitur memiliki nilai rata-rata mendekati nol dan standar deviasi mendekati satu, tanpa menghilangkan pola distribusi data yang ada.

Fitur	Sebelum Scaling (Min-Max)	Sesudah Scaling (Min-Max)
age	28 – 77	-2.71 – 2.49
trestbps	0 – 200	-7.16 – 3.69
chol	0 – 603	-1.83 – 3.70
thalch	60 – 202	-3.09 – 2.56
oldpeak	-2.6 – 6.2	-3.27 – 5.06
ca	0 – 3	-0.36 – 4.41

Tabel 7 Perbandingan Rentang Nilai Sebelum dan Sesudah Scaling

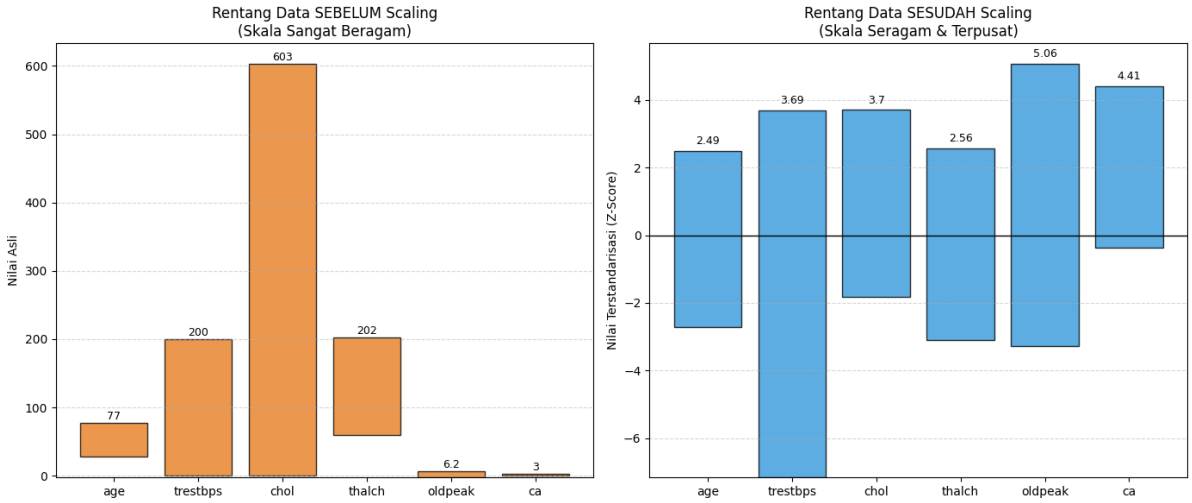


Diagram 7 Perbandingan Rentang Nilai Sebelum dan Sesudah Scaling

Hasil penskalaan menunjukkan bahwa seluruh fitur numerik berada pada skala yang relatif sebanding. Visualisasi menggunakan tabel sebelum dan sesudah scaling memperlihatkan bahwa fitur yang sebelumnya memiliki rentang nilai lebih besar tidak lagi mendominasi distribusi data. Kondisi ini mendukung proses pemodelan dengan memastikan bahwa setiap fitur memberikan kontribusi yang lebih proporsional dalam pembelajaran pola data oleh algoritma *Random Forest.*

4) Penangan Outlier

Tahap ini dimulai dengan eksplorasi visual dengan bloxpot untuk menemukan pencilan (*outlier*) data. Hasil pengamatan menunjukkan bahwa meskipun terdapat nilai esktrem pada beberapa fitur, nilai-nilai tersebut tetap berada di batas yang wajar dan tidak menyebabkan distorsi yang signifikan terhadap distribusi mayoritas data.

Oleh karena itu, tidak ada proses penghapusan (*trimming*) atau pembatasan (*capping*) terhadap *outlier* dalam penelitian ini. Keputusan ini didasarkan pada fakta bahwa algoritma *Random Forest* yang digunakan adalah metode berbasis pohon keputusan yang relatif tahan terhadap nilai ekstrem. Selain itu, metode ini dipilih untuk menjaga keutuhan jumlah data, yang memungkinkan tahapan *preprocessing* berikutnya untuk berkonsentrasi pada proses imputasi dan penskalaan fitur.

5) Encoding Variabel Kategorikal

Setiap variabel kategorikal diubah menjadi angka numerik menggunakan metode One-Hot Encoding dengan parameter drop_first = True agar algoritma *Machine Learning* dapat memproses data. Tujuan dari penerapan parameter ini adalah untuk mencegah multikolinearitas antar fitur hasil encoding sambil mempertahankan informasi kategori yang jelas dalam representasi numerik.

age	trestbps	chol	thalch	oldpeak	ca	sex_Male	cp_atypical	cp_nonanginal	thal_fixed
63	145	233	150	2.3	0	1	0	0	1
67	160	286	108	1.5	3	1	0	0	0
37	130	250	187	3.5	0	1	1	0	0
41	130	204	172	1.4	0	0	0	1	0

Tabel 8 Sampel Data Hasil Feature Engineering (Encoding)

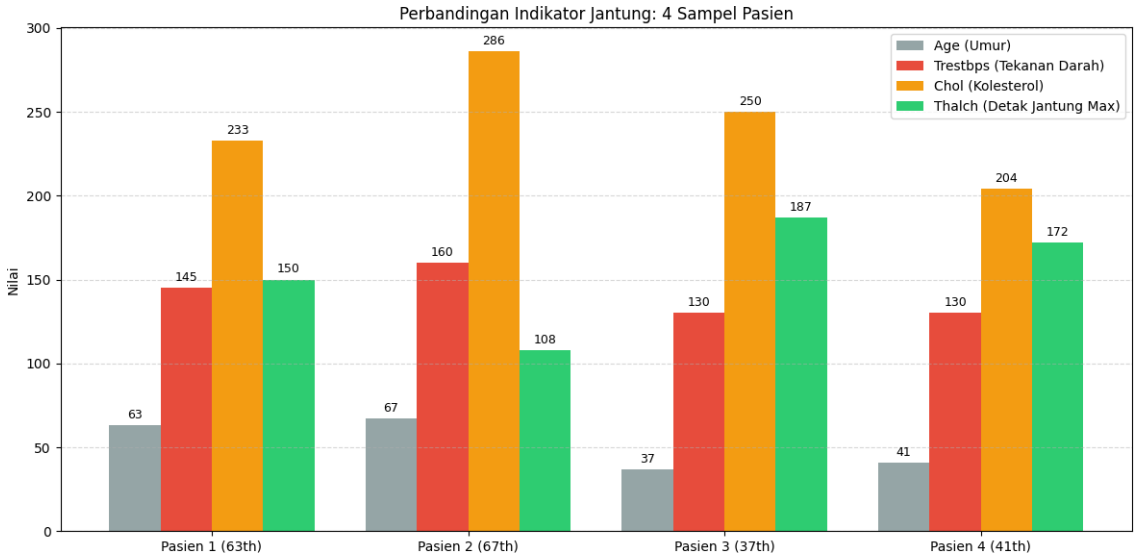


Diagram 8 Sampel Data Hasil Feature Engineering (Encoding)

Setelah seluruh tahapan *preprocessing* selesai, dataset yang dihasilkan sepenuhnya numerik, memiliki skala fitur yang seragam, dan tidak memiliki nilai yang hilang. Kondisi ini menunjukkan bahwa data telah siap untuk digunakan pada tahap pemodelan dan evaluasi kinerja model, yang akan dibahas pada subbab berikutnya.

C. Hasil Pemodelan

Tahap pemodelan dilakukan menggunakan algoritma *Random Forest* dengan pendekatan 5-*Fold Cross Validation*. Pendekatan ini digunakan untuk memperoleh evaluasi kinerja model yang lebih stabil dan representatif, dengan memanfaatkan seluruh data secara bergantian sebagai data latih dan data uji. Evaluasi kinerja model dilakukan pada tiga skenario yang berbeda untuk mengamati pengaruh tahapan *preprocessing* terhadap performa model.

- **Skenario A:** Menggunakan data mentah tanpa penerapan imputasi dan *feature scaling*.
- **Skenario B:** Menggunakan data yang telah melalui proses imputasi namun tanpa *feature scaling*.

- **Skenario C:** Menggunakan data yang telah melalui proses imputasi dan *feature scaling* secara lengkap.

Skenario	Akurasi	Presisi	Recall	F1-score
Data Mentah	0.4652	0.4456	0.4652	0.4011
Imputasi	0.4652	0.4456	0.4652	0.4011
Imputasi + Scaling	0.4641	0.4457	0.4641	0.3992

Tabel 9 Evaluasi Kinerja Model Berdasarkan Tahap Preprocessing

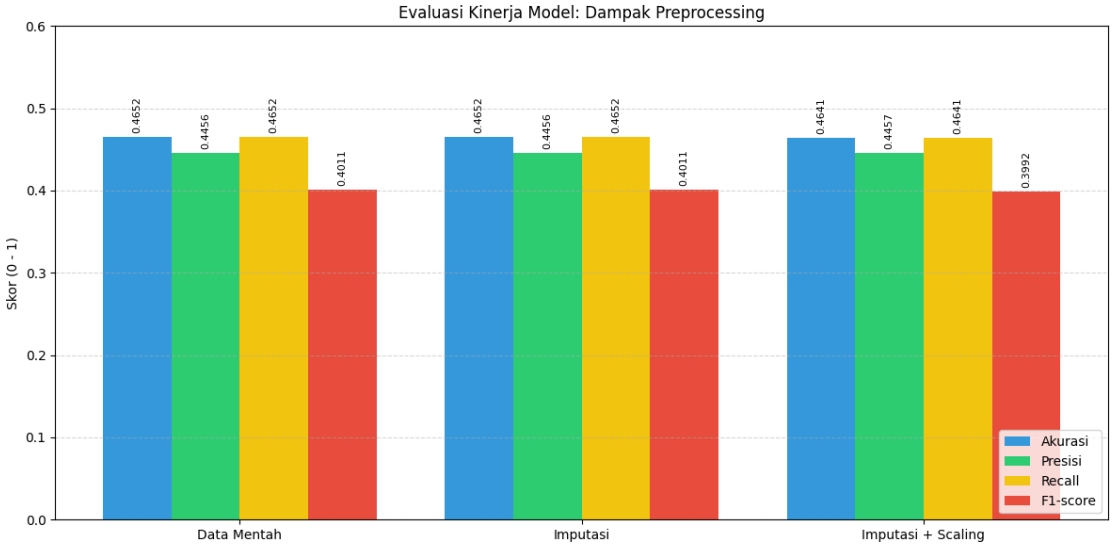


Diagram 9 Evaluasi Kinerja Model Berdasarkan Tahap Preprocessing

Berdasarkan hasil yang diperoleh, terlihat bahwa perbedaan nilai kinerja antar skenario relatif kecil. Nilai akurasi, presisi, recall, dan F1-score tidak mengalami perubahan yang signifikan setelah penerapan imputasi maupun kombinasi imputasi dan *feature scaling*. Meskipun demikian, variasi nilai metrik yang dihasilkan tetap memberikan gambaran mengenai dampak *preprocessing* terhadap stabilitas dan konsistensi kinerja model, khususnya dalam konteks penggunaan algoritma *Random Forest* yang relatif robust terhadap perbedaan kondisi data.

D. Hasil Pemodelan

Jika dibandingkan dengan penggunaan data mentah, penerapan imputasi dan pengembangan fitur tidak meningkatkan kinerja model secara signifikan, seperti yang ditunjukkan oleh hasil pemodelan. Hasil ini dapat dijelaskan oleh sifat algoritma *Random Forest* yang dikenal relatif tahan terhadap perbedaan skala fitur dan ketiadaan nilai yang tidak ada dalam jumlah terbatas.

Kinerja model dalam skenario dengan penerapan imputasi menghasilkan hasil yang sama dengan skenario data mentah. Kondisi ini menunjukkan bahwa *Random Forest* memiliki kemampuan untuk menangani ketidaklengkapan data secara efektif melalui mekanisme internalnya. Contoh mekanisme internal ini termasuk memilih subset data dan menambahkan fitur acak pada setiap pohon keputusan yang dibangun. Dengan menggunakan mekanisme ini, dampak kekurangan nilai terhadap proses pembelajaran model menjadi sangat kecil.

Selain itu, skenario ketiga tidak melihat peningkatan performa yang signifikan setelah menerapkan *feature scaling*. Hasil ini sejalan dengan teori yang menyatakan bahwa algoritma keputusan berbasis pohon tidak bergantung pada perhitungan jarak antar fitur. Oleh karena itu, proses penskalaan data tidak memengaruhi pembentukan struktur pohon atau hasil prediksi model secara signifikan.

Preprocessing memiliki peran penting dalam menjamin kualitas data dan konsistensi proses analisis, meskipun tahapan tersebut tidak berkontribusi secara langsung terhadap peningkatan performa model pada penelitian ini. Proses imputasi memastikan bahwa data tidak kehilangan nilai kosong, dan skala fitur membuat kondisi data lebih terstandarisasi dan siap digunakan apabila algoritma lain yang sensitif terhadap skala fitur diterapkan.

Secara keseluruhan, hasil penelitian ini menunjukkan bahwa, dalam kasus prediksi penyakit jantung menggunakan algoritma *Random Forest*, *preprocessing* lebih berfungsi sebagai upaya untuk meningkatkan kualitas dan kesiapan data daripada sebagai komponen penting dalam meningkatkan kinerja model. Hasil ini menunjukkan bahwa memilih teknik *preprocessing* yang sesuai dengan fitur algoritma yang digunakan sangat penting.

V. KESIMPULAN DAN SARAN

A. Kesimpulan

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa algoritma *Random Forest* mampu menghasilkan prediksi penyakit jantung yang stabil pada berbagai kondisi data. Tujuan penelitian untuk mengkaji pengaruh teknik imputasi dan *feature scaling* terhadap kinerja model telah tercapai melalui pengujian beberapa skenario *preprocessing*. Hasil pengujian menunjukkan bahwa penerapan imputasi data, maupun kombinasi antara imputasi dan *feature scaling*, tidak memberikan peningkatan kinerja yang signifikan terhadap nilai akurasi, presisi, recall, dan F1-score jika dibandingkan dengan penggunaan data tanpa *preprocessing*. Perbedaan performa antar skenario relatif kecil, dengan nilai akurasi yang berada pada rentang yang hampir sama. Dengan demikian, secara empiris dapat disimpulkan bahwa proses *preprocessing* tidak berkontribusi secara langsung terhadap peningkatan performa prediksi penyakit jantung menggunakan algoritma *Random Forest* dalam penelitian ini.

Temuan tersebut dapat dijelaskan oleh karakteristik algoritma *Random Forest* yang berbasis pohon keputusan, sehingga relatif tidak terpengaruh oleh perbedaan skala fitur maupun keberadaan sejumlah missing values. Hal ini didukung oleh mekanisme internal *Random Forest*, seperti pemilihan subset data dan fitur secara acak. Meskipun tidak berdampak signifikan terhadap peningkatan kinerja model, tahapan *preprocessing* tetap berperan penting dalam meningkatkan kualitas, konsistensi, dan kesiapan data, terutama dalam menangani nilai kosong pada fitur ca dan thal yang sebelumnya cukup tinggi. Penelitian ini menegaskan bahwa pemilihan teknik *preprocessing* perlu disesuaikan dengan karakteristik algoritma *Machine Learning* yang digunakan agar proses pengolahan data dapat berjalan secara optimal. Hasil penelitian ini dapat menjadi landasan bagi pengembangan studi selanjutnya dalam menentukan metode *preprocessing* dan algoritma yang sesuai untuk prediksi penyakit jantung.

B. Saran

Berdasarkan temuan penelitian, disarankan agar studi selanjutnya menggunakan dataset yang lebih besar dan lebih bervariasi guna menghasilkan kesimpulan yang lebih representatif. Selain itu, ruang lingkup pengujian dapat diperluas dengan melakukan perbandingan antara algoritma *Random Forest* dan algoritma *Machine Learning* lain yang lebih peka terhadap perbedaan skala fitur, seperti Support Vector Machine, K-Nearest Neighbors, dan Logistic Regression. Dengan demikian, pengaruh teknik *preprocessing* dapat dianalisis secara lebih menyeluruh. Penelitian lanjutan juga dapat mengkaji penerapan metode imputasi yang lebih kompleks, teknik seleksi fitur, serta optimasi parameter model untuk mengidentifikasi potensi peningkatan kinerja

prediksi. Melalui pengembangan tersebut, diharapkan sistem prediksi penyakit jantung yang dihasilkan memiliki tingkat akurasi dan keandalan yang lebih tinggi.

REFERENCES

Alfajr, N. H., & Defiyanti, S. (2024). Prediksi Penyakit Jantung Menggunakan Metode *Random Forest* Dan Penerapan Principal Component Analysis (Pca). *Jurnal Informatika Dan Teknik Elektro Terapan*, *12*(3S1). https://doi.org/10.23960/jitet.v12i3s1.5055

Alsyar, A., Amin Putra, R., Ramadhani, W. A., Hidayatullah, F. R., & Ismanto, E. (2025). Pemodelan Prediktif Diabetes Menggunakan Pendekatan Multimodel *Machine Learning* dan Deep Learning Predictive Modeling of Diabetes Using Multimodel *Machine Learning* and Deep learning Approaches. *Jurnal Computer Science and Information Technology (CoSciTech)*, *6*(2), 158–165. http://ejurnal.umri.ac.id/index.php/coscitech/indexhttps://doi.org/10.37859/coscitech.v6i2.9812

Aracri, F., Bianco, M. G., Quattrone, A., & Sarica, A. (2025). Bridging the Gap: Missing Data Imputation Methods and Their Effect on Dementia Classification Performance. *Brain Sciences*, *15*(6), 1–16. https://doi.org/10.3390/brainsci15060639

Arman, D., Indayana, N. S., Okmayura, F., Anjani, S. P., Nur, F., Farhan, M., & Faturrahman, A. (2025). *Jurnal Software Engineering and Information System (SEIS) PEMODELAN MACHINE LEARNING DENGAN ALGORITMA RANDOM FOREST*. *5*(2).

Faradeya, M. A.-Z., & Subhiyakto, E. R. (2025). Klasifikasi Penyakit Gagal Jantung Menggunakan Algoritma Naive Bayes. *Jurnal Algoritma*, *22*(1), 115–127. https://doi.org/10.33364/algoritma/v.22-1.2178

Gullam Almuzadid, & Egia Rosi Subhiyakto. (2025). Stroke Risk Classification Using the Ensemble Learning Method of XGBoost and *Random Forest*. *Journal of Applied Informatics and Computing*, *9*(3), 828–837. https://doi.org/10.30871/jaic.v9i3.9528

Pinheiro, J. M. H., de Oliveira, S. V. B., Silva, T. H. S., Saraiva, P. A. R., de Souza, E. F., Godoy, R. V., Ambrosio, L. A., & Becker, M. (2025). *The Impact of Feature scaling In Machine Learning: Effects on Regression and Classification Tasks*.

Prakash, P., Street, K., Narayanan, S., Fernandez, B. A., Shen, Y., & Shu, C. (2025). Benchmarking *Machine Learning* missing data imputation methods in large-scale mental health survey databases. *Artificial Intelligence in Health*, *2*(1), 81–92. https://doi.org/10.36922/aih.4406

Rasheed, S., Kumar, G. K., Rani, D. M., Kantipudi, M. V. V. P., & Anila, M. (2024). Heart Disease Prediction Using GridSearchCV and *Random Forest*. *EAI Endorsed Transactions on Pervasive Health and Technology*, *10*, 1–8. https://doi.org/10.4108/eetpht.10.5523

Ren, W., Liu, Z., Wu, Y., Zhang, Z., Hong, S., & Liu, H. (2024). Moving Beyond Medical Statistics: A Systematic Review on Missing Data Handling in Electronic Health Records. *Health Data Science*, *4*, 1–18. https://doi.org/10.34133/hds.0176

Setiawan, M. A., & Efendi, M. H. (2025). *Pengaruh Teknik Preprocessing terhadap Kinerja Model Explainable Boosting Machine (EBM) untuk Prediksi Serangan Jantung*. *8*(2), 317–326.

Seu, K., Kang, M. S., & Lee, H. (2022). An Intelligent Missing Data Imputation

Techniques: A Review. *International Journal on Informatics Visualization*, *6*(1–2), 278–283. https://doi.org/10.30630/joiv.6.1-2.935

Tamba, S. P., & -, E. (2022). Prediksi Penyakit Gagal Jantung Dengan Menggunakan *Random Forest*. *Jurnal Sistem Informasi Dan Ilmu Komputer Prima(JUSIKOM PRIMA)*, *5*(2), 176–181. https://doi.org/10.34012/jurnalsisteminformasidanilmukomputer.v5i2.2445

Torthi, R., Marapatla, A. D. K., Mande, S., Gadiraju, H. K. V., & Kanumuri, C. (2024). Heart Disease Prediction Using *Random Forest* Based Hybrid Optimization Algorithms. *International Journal of Intelligent Engineering and Systems*, *17*(2), 134–144. https://doi.org/10.22266/ijies2024.0430.12

WHO. (2018). Cardiovascular diseases (CVDs) Key facts What are the risk factors for cardiovascular disease ? *World Health Organization*, *May 2017*, 1–8.

Wijaya, A. P. (2025). Perbandingan Algoritma Klasifikasi Random Foresst dengan Naïve Bayes Classifier pada Studi Penyakit Berdasarkan Pola Nutrisi. *Remik: Riset Dan E-Jurnal Manajemen Informatika Komputer*, *9*(1), 429–438.