Irham Maulana Johani | Hermawan Sentyaki Sarjito | Jonathan

2025-11-13

Daftar Isi

Daftar Gambar

Sekolah Vokasi
School of Vocational Studies

Analisis Prediktif Penyakit Jantung Menggunakan Algoritma Klasifikasi (Random Forest) Berbasis Data Klinis

1 Latar Belakang

Gagal jantung merupakan salah satu penyakit kardiovaskular kronis dengan angka kesakitan dan kematian yang tinggi secara global, sehingga menjadi beban besar bagi sistem kesehatan [Savarese et al., 2022]. Studi epidemiologi terbaru menunjukkan bahwa jumlah kasus gagal jantung terus meningkat terus meningkat, terutama di kawasan Asia dengan jumlah data 33% di Asia Tengah hingga 186% dalam rentang tahun 1990 hingga 2019 dan negara berpenghasilan menengah [Feng et al., 2024]. Kondisi ini tidak hanya berdampak pada peningkatan angka hospitalisasi dan readmission, tetapi juga mengurangi kualitas hidup pasien dan menambah beban ekonomi masyarakat serta fasilitas pelayanan kesehatan [Yan et al., 2023]. Dalam konteks tersebut, identifikasi dini pasien yang berisiko tinggi menjadi sangat penting untuk menekan angka rawat inap berulang dan mortalitas.

Dalam praktik klinis, penilaian risiko pasien gagal jantung masih banyak bergantung pada kombinasi penilaian subjektif klinisi, pedoman klinis, dan pengamatan terhadap berbagai parameter klinis secara manual. Berbagai penelitian juga memperkuat bahwa pasien gagal jantung memiliki angka readmission dan mortalitas jangka pendek yang tinggi yang merefleksikan belum optimal dalam stratifikasi risiko di layanan kesehatan [Sabouri et al., 2023]. Hal ini menimbulkan risiko bahwa sebagian pasien dengan profik risiko tinggi tidak teridentifikasi tepat waktu sehingga tidak mendapatkan prioritas pemantauan dan intervensi yang lebih intensif.

Seiring meluasnya penggunaan electronic health records (EHR), data klinis seperti karakteristik pasien, hasil laboratorium, diagnosis, prosedur, dan terapi kini tersimpan dalam jumlah besar dan relatif terstruktur di fasilitas pelayanan kesehatan. Integrasi data EHR dengan metode analitik, seperti kecerdasan buatan dan machine learning telah dilaporkan mampu mentransformasi cara prediksi dan manajemen risiko penyakit kardiovaskular [Tsai et al., 2025]. Berdasarkan penelitian lain menunjukkan, bahwa teknik data mining dapat dimanfaatkan untuk mengekstraksi parameter klinis penting serta data laboratorium guna memperbaiki subclassificiation gagal jantung [Vuori et al., 2023].

Sejalan dengan perkembangan tersebut, berbagai algoritma klasifikasi seperti logistic regression, random forest, support vector machine, dan gradient boosting telah diterapkan untuk diagnosis, prediksi, dan prognosis gagal jantung, dengan performa yang umumnya lebih baik daripada skor risiko tradisional [Saqib et al., 2024]. Studi lain yang memanfaatkan Heart Failure Prediciton Dataset juga menunjukkan bahwa perbandingan beberapa model klasifikasi memungkinkan identifikasi algoritma yang optimal untuk klasifikasi gagal jantung [Chulde-Fernández et al., 2025]. Di Indonesia, terdapat penelitian yang telah mengevaluasi berbagai model machine learning untuk memprediksi keparahan dan readmission pasien gagal jantung serta menemukan model dengan performa terbaik untuk diintegrasikan ke dalam aplikasi pemantauan mandiri pasien [Indriany et al., 2024]. Meskipun demikian, diperlukan penerapan dan analisis model klasifikasi berbasis Random Forest untuk mengidentifikasi pasien berisiko tinggi gagal jantung sehingga dapat dibangun model prediksi dengan performa yang dapat menghasilkan pengetahui mengenai faktor-faktor klinis yang berkontribusi terhadap kategori risiko tinggi tersebut.

2 Rumusan Masalah

1. Bagaimana karakteristik data klinis pasien gagal jantung yang ditanjau dari distirbusi kelas risiko serta variabel klinisnya?
2. Bagaimana kinerja model klasifikasi Random Forest dalam mengidentifikasi pasien berisiko tinggi gagal jantung berdasarkan metrik evaluasi seperti akurasi, presisi, recall, dan F1-score?
3. Fitur-fitur klinis apa saja yang paling berkontribusi terhadap penentuan kategori pasien berisiko tinggi gagal jantung berdasarkan hasil analisis model Random Forest?

3 Tujuan Penelitian

1. Mendeskripsikan karakteristik data klinis pasien gagal jantung termasuk distirbusi kelas risiko dan variabel-variabel klinis yang relevan.
2. Membangun dan mengevaluasi model klasifikasi Random Forest untuk mengidentifikasi pasien berisiko tinggi gagal jantung menggunakan metrik evaluasi seperti akurasi, presisi, recall, dan F1-score.
3. Mengidentifikasi fitur-fitur klinis yang paling berpengaruh terhadap penentuan kategori pasien berisiko tinggi gagal jantung berdasarkan hasil analisis model Random Forest.

4 Sumber dan Karakteristik Data

- **Sumber Data:** Dataset publik "Heart Failure Prediction Dataset" yang diperoleh dari platform Kaggle.
- **Karakteristik Data:** Dataset awal terdiri dari 918 observasi (baris) dan 12 variabel (kolom).
- **Target Variabel:** `HeartDisease` (Biner: 1 = Sakit Jantung, 0 = Normal).
- **Variabel Prediktor:** Mencakup data demografis (`Age`, `Sex`), data klinis (`RestingBP`, `Cholesterol`), dan hasil tes (`MaxHR`, `RestingECG`, dll.).
- **Kualitas Data:** Terdapat potensi masalah kualitas data, seperti nilai '0' yang tidak logis pada kolom `Cholesterol` dan `RestingBP`, yang harus ditangani pada tahap *Data Preparation*.

5 Metodologi Data Mining (CRISP-DM)

Proyek ini mengadopsi kerangka kerja *CRoss-Industry Standard Process for Data Mining* (CRISP-DM) untuk memastikan proses analisis berjalan sistematis, terdokumentasi, dan fokus pada tujuan. Berikut adalah penjabaran dari enam fase CRISP-DM yang diterapkan dalam konteks analisis *Heart Failure Prediction Dataset*.

5.1 Business Understanding (Pemahaman Konteks)

Fase ini berfokus pada pemahaman tujuan dan persyaratan proyek dari perspektif klinis.

- **Tujuan Bisnis/Klinis:** Tujuan utamanya adalah mengurangi angka mortalitas dan morbiditas akibat gagal jantung. Masalahnya adalah keterlambatan diagnosis atau penilaian risiko yang tidak optimal di fasilitas layanan kesehatan.
- **Asesmen Situasi:** Penilaian risiko klinis saat ini masih bersifat subjektif dan manual. Terdapat ketersediaan data klinis (EHR) yang melimpah namun belum dimanfaatkan secara optimal. Dataset "Heart Failure Prediction" dari Kaggle akan digunakan sebagai proksi untuk data klinis ini.

- **Tujuan Data Mining:** Menerjemahkan masalah klinis menjadi tugas data mining. Tujuannya adalah: **Membangun model klasifikasi** yang mampu memprediksi apakah seorang pasien memiliki penyakit jantung (`HeartDisease` = 1) atau tidak (`HeartDisease` = 0) berdasarkan 11 variabel prediktor klinis.
- **Rencana Proyek:** Menggunakan R dan RStudio, menerapkan algoritma **Random Forest**, dan mengevaluasi model menggunakan metrik standar (Akurasi, Presisi, Recall, F1-Score) serta *variable importance*.

5.2 Data Understanding (Pemahaman Data)

Fase ini melibatkan pengumpulan dan eksplorasi data awal untuk memahami karakteristik dan kualitasnya.

- **Pengumpulan Data:** Data `heart.csv` (Heart Failure Prediction Dataset) dimuat ke dalam R.
- **Deskripsi Data:** Dataset terdiri dari 918 observasi dan 12 variabel. Variabel target adalah `HeartDisease` (biner). Variabel prediktor mencakup data demografis (`Age`, `Sex`), klinis (`RestingBP`, `Cholesterol`, `FastingBS`), dan hasil tes EKG/latihan (`RestingECG`, `MaxHR`, `ExerciseAngina`, `Oldpeak`, `ST_Slope`).
- **Eksplorasi Data:** Dilakukan analisis statistik deskriptif (`summary()`) dan visualisasi awal (histogram, bar plot) untuk memahami distribusi setiap variabel dan hubungannya dengan variabel target `HeartDisease`.
- **Verifikasi Kualitas Data: Temuan Kunci:** Identifikasi masalah kualitas data yang kritis. Ditemukan nilai `0` yang tidak logis secara biologis pada variabel `RestingBP` (tekanan darah tidak mungkin 0) dan `Cholesterol` (kolesterol serum tidak mungkin 0). Nilai-nilai ini akan diperlakukan sebagai data anomali atau *missing values* yang harus ditangani.

5.3 Data Preparation (Persiapan Data)

Ini adalah fase terpenting di mana data mentah dibersihkan dan ditransformasi menjadi dataset yang siap untuk *modeling*.

- **Seleksi Data:** Semua 11 variabel prediktor akan digunakan, karena *variable importance* dari Random Forest akan membantu menyaringnya nanti.
- **Pembersihan Data (Cleansing):**
  - Baris data dengan `RestingBP = 0` akan **dihapus** (di-filter), karena jumlahnya sedikit dan nilainya jelas tidak valid.
  - Nilai `Cholesterol = 0` akan ditangani, namun tidak dengan penghapusan baris karena jumlahnya mungkin signifikan.
- **Pemisahan Data (Data Splitting):** Dataset yang telah dibersihkan (`df_clean`) akan segera dibagi menjadi **data latih (80%)** dan **data uji (20%)** menggunakan `createDataPartition`.
  - **Catatan Kritis:** Langkah ini *harus* dilakukan **sebelum** imputasi untuk mencegah *data leakage*, di mana informasi statistik dari data uji (misalnya, median) "bocor" ke dalam proses pelatihan.
- **Konstruksi & Imputasi Data:**
  - Nilai median untuk `Cholesterol` akan dihitung **hanya** dari data latih (`train_data`) yang memiliki nilai `Cholesterol > 0`.
  - Median ini kemudian akan digunakan untuk **mengimputasi** (mengganti) nilai `Cholesterol = 0` baik di `train_data` maupun di `test_data`.
- **Format Data:**

  - Variabel-variabel kategorikal (`Sex`, `ChestPainType`, `RestingECG`, `ExerciseAngina`, `ST_Slope`, `FastingBS`) akan dikonversi menjadi tipe data `factor` agar dapat diinterpretasikan dengan benar oleh model di R.
  - Variabel target `HeartDisease` juga akan dikonversi menjadi `factor`.

5.4 Modeling (Pemodelan)

Pada fase ini, model prediktif dibangun dan dilatih.

- **Pemilihan Teknik Modeling:** Sesuai tujuan penelitian, algoritma **Random Forest (`method = "rf"`)** dipilih. Alasan: performa tinggi, kemampuan menangani data campuran (numerik/kategorikal), dan menghasilkan metrik *variable importance* secara inheren.
- **Desain Pengujian:** Untuk mendapatkan estimasi kinerja model yang stabil dan *robust* (tidak *overfitting*), akan digunakan teknik validasi silang, yaitu **10-Fold Cross-Validation (`method = "cv", number = 10`)** pada data latih.
- **Pembangunan Model:** Model `model_rf` akan dilatih menggunakan fungsi `train()` dari paket `caret` pada `train_data`. Formula yang digunakan adalah `HeartDisease ~ .`, yang berarti semua variabel prediktor lain digunakan untuk memprediksi `HeartDisease`. Proses *pre-processing* (`center`, `scale`) juga diterapkan untuk standarisasi data numerik.

5.5 Evaluation (Evaluasi)

Fase ini fokus pada penilaian kinerja model, baik secara teknis maupun relevansinya terhadap tujuan.

- **Evaluasi Kinerja Teknis:** Model `model_rf` yang telah dilatih akan digunakan untuk membuat prediksi pada data yang belum pernah dilihat sebelumnya, yaitu `test_data`.
- **Metrik Evaluasi:** Kinerja model akan diukur menggunakan **Confusion Matrix**. Metrik utama yang akan dianalisis adalah:
  - **Akurasi:** Persentase prediksi yang benar secara keseluruhan.
  - **Presisi:** Seberapa akurat prediksi positif (dari semua yang diprediksi sakit, berapa yang benar-benar sakit).
  - **Recall (Sensitivity):** Seberapa baik model menemukan semua kasus positif (dari semua yang benar-benar sakit, berapa yang berhasil terdeteksi). **Metrik ini sangat penting** dalam konteks medis untuk meminimalkan *False Negative* (pasien sakit tapi diprediksi sehat).
  - **F1-Score:** Rata-rata harmonik dari Presisi dan Recall.
- **Evaluasi Tujuan Proyek:**
  - Kinerja model akan dievaluasi untuk menjawab Rumusan Masalah #2.
  - Analisis **Variable Importance (`varImp(model_rf)`)** akan diekstrak dari model untuk mengidentifikasi fitur klinis apa yang paling berpengaruh. Ini akan menjawab Rumusan Masalah #3.

5.6 Deployment (Penyebaran)

Fase ini menguraikan bagaimana hasil proyek akan digunakan.

- **Rencana Deployment:** Dalam konteks proposal ini, "deployment" berarti presentasi hasil dan model yang fungsional. Model R yang telah dilatih (`model_rf`) adalah *output* utama.
- **Rencana Monitoring (Implikasi):** Model ini dilatih pada dataset statis. Untuk penerapan nyata, model perlu dipantau kinerjanya terhadap data pasien baru dan dilatih ulang secara berkala.

- **Laporan Akhir & Presentasi:** Seluruh temuan, termasuk kinerja model, *confusion matrix*, dan plot *variable importance*, akan didokumentasikan dalam laporan akhir (makalah ini).
- **Potensi Penerapan:** Seperti yang dijelaskan dalam "Dampak", model ini dapat menjadi dasar untuk **Sistem Pendukung Keputusan (Decision Support System)** sederhana bagi praktisi medis untuk melakukan *screening* awal risiko penyakit jantung.

6 Teknik dan Algoritma yang Digunakan

Proyek ini menerapkan beberapa fungsi data mining, yang dibagi menjadi fungsi mayor dan minor:

- **Fungsi Mayor (Tujuan Utama):**
  - **Klasifikasi:** Ini adalah fungsi utama proyek. Tujuannya adalah untuk membangun model yang dapat memprediksi kelas target biner (`HeartDisease`: 0 atau 1).
  - **Algoritma: Random Forest** dipilih sebagai algoritma klasifikasi utama. Alasan pemilihannya adalah kemampuannya menangani data kategorikal dan numerik secara bersamaan, ketahanannya terhadap *overfitting* (dibandingkan *decision tree* tunggal), dan kemampuannya untuk menghasilkan metrik *variable importance*.
- **Fungsi Minor (Pendukung):**
  - **Pembersihan Data:** Menghapus baris dengan data yang tidak logis (misal `RestingBP = 0`).
  - **Transformasi Data (Imputasi):** Mengganti nilai `Cholesterol = 0` dengan nilai median yang dihitung dari data latih.
  - **Seleksi Fitur (Feature Importance):** Meskipun tidak dilakukan seleksi fitur secara manual di awal, model Random Forest secara inheren menyediakan output *variable importance*. Ini akan digunakan untuk menjawab rumusan masalah ketiga, yaitu mengidentifikasi prediktor klinis yang paling berpengaruh.

7 Rencana Implementasi dan Alat yang Digunakan

Proyek ini akan diimplementasikan menggunakan bahasa pemrograman **R** dalam lingkungan pengembangan terintegrasi (IDE) **RStudio**.

Paket (libraries) R utama yang akan digunakan adalah:

- `tidyverse` (termasuk `dplyr` dan `ggplot2`): Untuk keperluan manipulasi, pembersihan data, dan visualisasi data.
- `caret`: Sebagai framework utama untuk proses *machine learning*, termasuk pembagian data (*data splitting*), *pre-processing*, pelatihan model (*cross-validation*), dan evaluasi model.
- `randomForest`: Untuk implementasi algoritma Random Forest.

8 Hasil yang Diharapkan dan Dampak

8.1 Hasil yang Diharapkan (Output Proyek)

1. **Deskripsi Karakteristik Data:** Analisis dan visualisasi mengenai distribusi data pasien, termasuk perbandingan antar kelas (Normal vs. Sakit Jantung) dan korelasi antar variabel klinis.
2. **Model Prediktif:** Sebuah model klasifikasi R (`model_rf`) yang telah dilatih dan divalidasi, yang mampu memprediksi risiko penyakit jantung berdasarkan fitur-fitur klinis.

3. **Laporan Evaluasi Kinerja:** Penilaian kuantitatif kinerja model pada data uji, disajikan dalam bentuk *confusion matrix* beserta metrik turunannya (Akurasi, Presisi, Recall, F1-Score). Laporan ini akan menunjukkan seberapa akurat model dalam mengidentifikasi kasus *True Positive* (pasien sakit terdeteksi) dan *True Negative* (pasien sehat terdeteksi), serta tingkat kesalahannya (*False Positive* dan *False Negative*).
4. **Insight Kunci (Variable Importance):** Sebuah visualisasi (plot) yang mengurutkan fitur-fitur klinis dari yang paling berpengaruh hingga yang paling tidak berpengaruh dalam menentukan prediksi model. Diharapkan fitur-fitur yang terkait dengan hasil EKG dan tes fisik (seperti `ST_Slope`, `Oldpeak`, `MaxHR`, dan `ExerciseAngina`) akan teridentifikasi sebagai prediktor paling signifikan.

8.2 Dampak dan Potensi Penerapan

Hasil dari proyek ini memiliki dampak praktis yang signifikan. Model prediktif yang dihasilkan dapat memberikan *insight* berbasis data bagi praktisi medis mengenai faktor-faktor risiko utama penyakit jantung.

Secara konkret, model ini dapat digunakan sebagai **Sistem Pendukung Keputusan (Decision Support System)** untuk *screening* awal di fasilitas layanan kesehatan. Pasien yang diidentifikasi oleh model memiliki profil risiko tinggi (diprediksi sebagai '1') dapat segera dirujuk untuk pemeriksaan diagnostik lebih lanjut yang lebih intensif (seperti angiografi). Hal ini berpotensi mempercepat proses diagnosis, memungkinkan intervensi medis yang lebih dini, dan pada akhirnya membantu menekan angka mortalitas akibat gagal jantung.

Pustaka

Bryan Chulde-Fernández, Denisse Enríquez-Ortega, Cesar Guevara, Paulo Navas, Andrés Tirado-Espín, Paulina Vizcaíno-Imacaña, Fernando Villalba-Meneses, Carolina Cadena-Morejon, Diego Almeida-Galarraga, and Patricia Acosta-Vargas. Classification of heart failure using machine learning: A comparative study. *Life*, 15(3):496, 2025.

Jiayu Feng, Yuhui Zhang, and Jian Zhang. Epidemiology and burden of heart failure in asia. *JACC: Asia*, 4(4):249–264, 2024.

Finna E Indriany, Kemal N Siregar, Budhi Setianto Purwowiyoto, Bambang Budi Siswanto, Indrajani Sutedja, and Hendy R Wijaya. Predicting the risk of severity and readmission in patients with heart failure in indonesia: A machine learning approach. *Healthcare Informatics Research*, 30(3):253–265, 2024.

Maziar Sabouri, Ahmad Bitarafan Rajabi, Ghasem Hajianfar, Omid Gharibi, Mobin Mohebi, Atlas Haddadi Avval, Nasim Naderi, and Isaac Shiri. Machine learning based readmission and mortality prediction in heart failure patients. *Scientific Reports*, 13(1):18671, 2023.

Muhammad Saqib, Prinka Perswani, Abraar Muneem, Hassan Mumtaz, Fnu Neha, Saiyad Ali, and Shehroze Tabassum. Machine learning in heart failure diagnosis, prediction, and prognosis. *Annals of Medicine and Surgery*, 86(6):3615–3623, 2024.

Gianluigi Savarese, Peter Moritz Becher, Lars H Lund, Petar Seferovic, Giuseppe MC Rosano, and Andrew JS Coats. Global burden of heart failure: a comprehensive and updated review of epidemiology. *Cardiovascular research*, 118(17):3272–3287, 2022.

Ming-Lung Tsai, Kuan-Fu Chen, and Pei-Chun Chen. Harnessing electronic health records and artificial intelligence for enhanced cardiovascular risk prediction: A comprehensive review. *Journal of the American Heart Association*, 14(6):e036946, 2025.

Matti A Vuori, Tuomo Kiiskinen, Niina Pitkänen, Samu Kurki, Hannele Laivuori, Tarja Laitinen, Sampo Mäntylahti, Aarno Palotie, FinnGen, and Teemu J Niiranen. Use of electronic health record data mining for heart failure subtyping. *BMC Research Notes*, 16(1):208, 2023.

Tao Yan, Shijie Zhu, Xiujie Yin, Changming Xie, Junqiang Xue, Miao Zhu, Fan Weng, Shichao Zhu, Bitao Xiang, Xiaonan Zhou, et al. Burden, trends, and inequalities of heart failure globally, 1990 to 2019: a secondary analysis based on the global burden of disease 2019 study. *Journal of the American Heart Association*, 12(6):e027852, 2023.