LAPORAN TUGAS BESAR MATA KULIAH
KECERDASAN BUATAN

PREDIKSI PENYAKIT JANTUNG MENGGUNAKAN RANDOM FOREST PADA DATASET UCI CLEVELAND

Disusun oleh:

Kelompok dari Kelas A

Muhamad Saepul Rizal – 2306142

Fathir Miftah Nursalim – 2306135

Dosen Pengampu Mata Kuliah:

Leni Fitriani, S.Kom, M.Kom

INSTITUT TEKNOLOGI GARUT
JURUSAN ILMU KOMPUTER
PROGRAM STUDI TEKNIK INFORMATIKA

TAHUN AKADEMIK 2024/2025

1. Business Understanding

1.1 Permasalahan Dunia Nyata

Penyakit jantung merupakan salah satu penyakit tidak menular yang paling mematikan dan menjadi penyebab utama kematian di seluruh dunia. Menurut laporan World Health Organization (WHO), lebih dari 17 juta orang meninggal setiap tahunnya akibat penyakit jantung dan pembuluh darah. Di Indonesia sendiri, beban penyakit jantung semakin meningkat, menjadi salah satu penyumbang utama kematian dan pembiayaan di sektor kesehatan(Elektronik et al. 2023).

Salah satu tantangan besar dalam penanganan penyakit jantung adalah keterlambatan diagnosis. Penyakit jantung koroner seringkali tidak menunjukkan gejala hingga mencapai tahap lanjut, sehingga mempersulit proses deteksi dini. Kondisi ini membutuhkan solusi teknologi yang dapat melakukan klasifikasi risiko secara cepat dan akurat berdasarkan data kesehatan pasien seperti tekanan darah, kolesterol, detak jantung, dan hasil tes lainnya(Depari, Widiastiwi, and Santoni 2022).

Dengan meningkatnya jumlah data medis dan kemajuan teknologi kecerdasan buatan, metode klasifikasi berbasis algoritma seperti Random Forest menjadi solusi yang menjanjikan. Algoritma ini telah terbukti mampu memproses data kesehatan pasien dan memprediksi risiko penyakit jantung secara efisien, yang dapat membantu tenaga medis dalam membuat keputusan yang lebih tepat waktu dan berdasarkan data(Hidayat, Sunyoto, and Al Fatta 2023).

1.2 Tujuan Proyek

Tujuan utama dari proyek ini adalah untuk membangun sebuah model klasifikasi yang mampu memprediksi apakah seorang pasien berisiko menderita penyakit jantung berdasarkan data medis seperti tekanan darah, kadar kolesterol, detak jantung maksimum, dan fitur kesehatan lainnya. Model ini diharapkan dapat membantu dalam pengambilan keputusan medis yang lebih cepat dan akurat, terutama dalam proses skrining awal. Dengan pendekatan berbasis data, proses prediksi diharapkan menjadi lebih objektif dan tidak hanya mengandalkan intuisi klinis(Alfajr and Defiyanti 2024).

Dalam implementasinya, algoritma Random Forest dipilih karena kemampuannya dalam menangani data dengan kombinasi fitur numerik dan kategorikal, serta memberikan hasil klasifikasi yang akurat dan stabil. Selain itu, Random Forest juga dapat memberikan informasi penting mengenai fitur-fitur mana yang paling berkontribusi dalam proses prediksi. Tujuan jangka panjangnya adalah mengembangkan sistem yang dapat digunakan oleh instansi kesehatan sebagai alat bantu diagnosis berbasis kecerdasan buatan(Rahmada et al. 2024).

1.3 Pengguna Sistem

- Dokter dan tenaga medis
- Peneliti bidang kesehatan
- Sistem pendukung keputusan medis

1.4 Manfaat Implementasi AI

- Otomatisasi proses prediksi penyakit
- Membantu pengambilan keputusan berbasis data
- Meningkatkan efisiensi diagnosis awal

2. Data Understanding

2.1 Sumber Data

Dataset berasal dari Kaggle, menggunakan subset UCI Cleveland yaitu UCI Heart Disease Data, berikut linknya (https://www.kaggle.com/datasets/thisishusseinali/uci-heart-disease-data)

2.2 Deskripsi Fitur

Kolom	Deskripsi
age	Usia pasien
sex	Jenis kelamin (1=pria, 0=wanita)
cp	Jenis nyeri dada
trestbps	Tekanan darah saat istirahat
chol	Kolesterol serum
fbs	Gula darah puasa > 120 mg/dl
restecg	Hasil EKG istirahat
thalach	Detak jantung maksimal
exang	Angina yang diinduksi olahraga
oldpeak	Depresi ST akibat latihan
slope	Kemiringan segmen ST
ca	Jumlah pembuluh darah utama (0–3)
thal	Hasil tes thalassemia
target	0 = sehat, 1 = penyakit jantung

2.3 Ukuran dan Format

- Jumlah data: 606 baris × 14 kolom
- Format: CSV
- Tipe data: numerik dan kategorikal

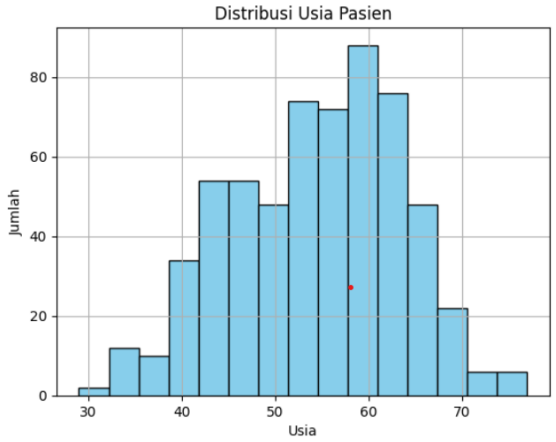
2.4 Tipe Data dan Target Klasifikasi

- Dataset terdiri dari 14 kolom (fitur) dengan 606 baris data pasien.
- Semua kolom bertipe numerik (integer atau float).
- Fitur terdiri dari kombinasi:
  - Data kategorikal numerik: sex, cp, fbs, restecg, exang, slope, ca, thal
  - Data numerik kontinu: age, trestbps, chol, thalach, oldpeak
- Tidak terdapat data dalam format teks atau string.
- Target klasifikasi berada di kolom target:
  - 0 = pasien sehat
  - 1 = pasien mengidap penyakit jantung
- Masalah ini termasuk dalam klasifikasi biner (binary classification).
- Tidak diperlukan proses encoding tambahan karena semua data sudah dalam format numerik.

3. Exploratory Data Analysis (EDA

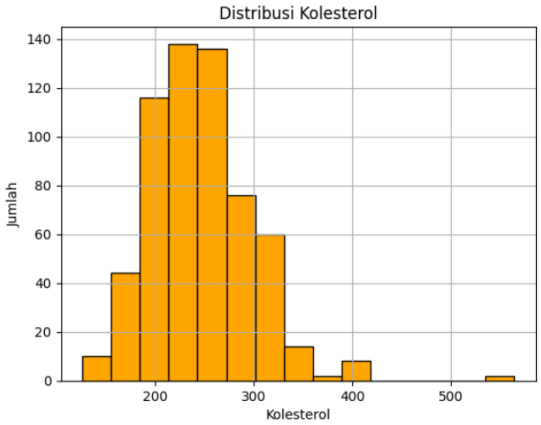
3.1 Visualisasi distribusi data

Gambar 1 Histogram distribusi usia pasien menunjukkan bahwa mayoritas pasien berusia antara 50 hingga 60 tahun. Distribusi ini berbentuk normal condong ke kanan.



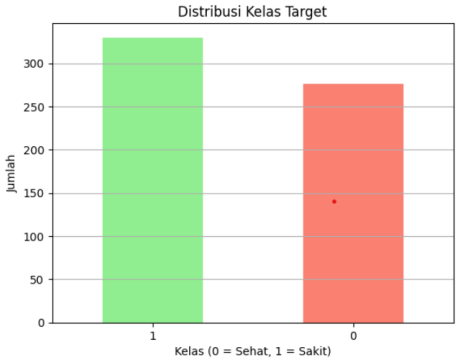
Gambar 1 Distribusi Usia Pasien

Gambar 2 Distribusi kolesterol pasien berkisar antara 150–300 mg/dL, dengan puncak di sekitar 230 mg/dL.



Gambar 2 Distribusi Kolesterol

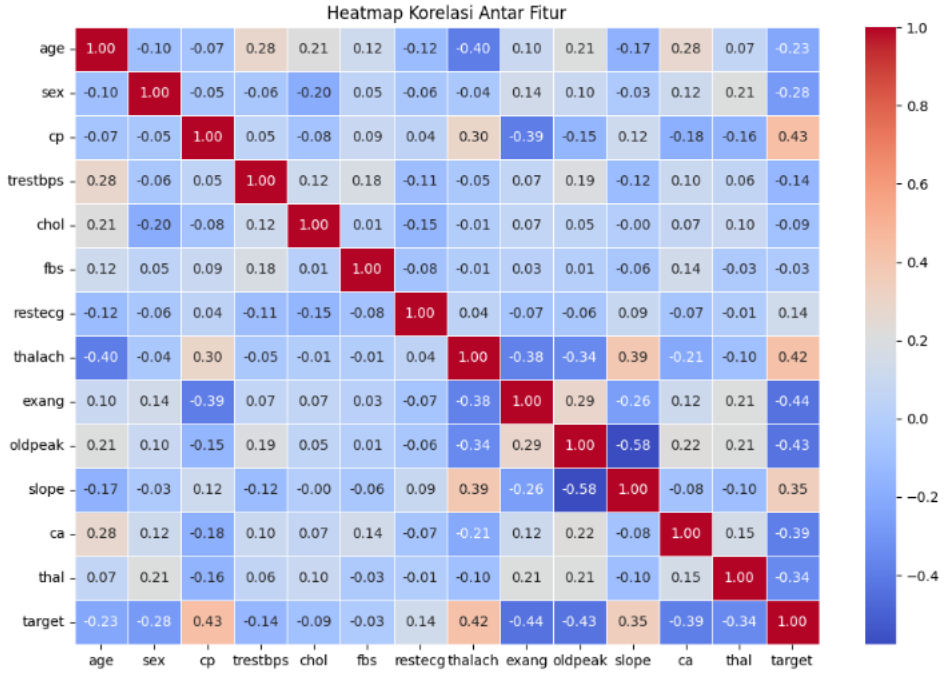
Gambar 3 Distribusi kelas target menunjukkan jumlah pasien berpenyakit jantung (kelas 1) sebanyak 330 dan pasien sehat (kelas 0) sebanyak 276, sehingga dataset ini relatif seimbang.



Gambar 3 Distribusi Kelas Target

3.2 Analisis korelasi antar fitur

Gambar 4 Heatmap korelasi menunjukkan fitur-fitur yang memiliki hubungan signifikan dengan target, di antaranya cp, thalach, oldpeak, dan thal.



Gambar 4 Korelasi Fitur

3.3 Deteksi data tidak seimbang

Distribusi target (Gambar 3) menunjukkan bahwa jumlah data antara kelas 0 dan 1 cukup seimbang, sehingga tidak diperlukan teknik penanganan imbalanced dataset seperti SMOTE atau undersampling.

3.4 Insight awal dari pola data

Fitur cp dan thalach berasosiasi kuat dengan kemungkinan penyakit jantung. Pasien dengan nilai cp tinggi (jenis nyeri dada tertentu) dan thalach tinggi cenderung memiliki hasil target positif (memiliki penyakit jantung). Ini didukung oleh hasil visualisasi feature importance pada modeling.

4. Data Preparation

Pembersihan data dilakukan dengan memeriksa missing value. Hasilnya, seluruh kolom tidak memiliki nilai kosong. Duplikasi data juga tidak ditemukan.

Proses encoding dilakukan menggunakan pd.get_dummies() meskipun semua fitur sudah numerik, sebagai langkah aman untuk memastikan model tidak menganggap fitur kategorikal sebagai ordinal.

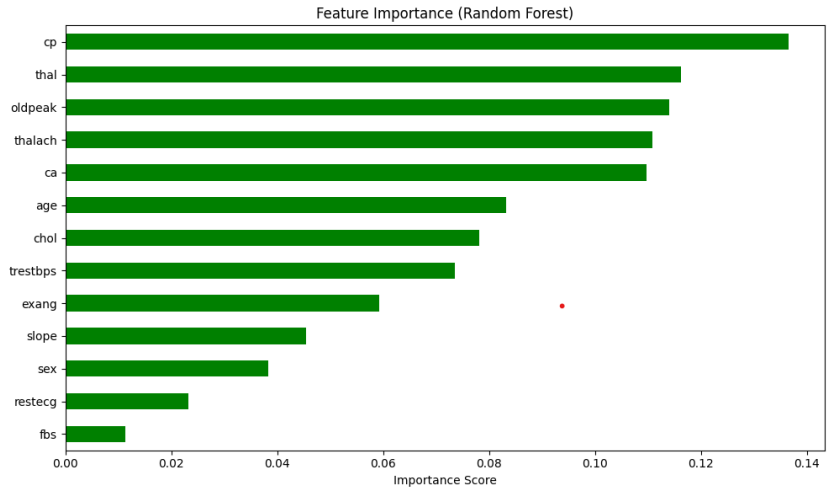
Semua fitur dinormalisasi menggunakan MinMaxScaler agar berada pada skala 0–1. Dataset dibagi menjadi data latih dan data uji dengan rasio 80:20 dan stratifikasi berdasarkan target, guna menjaga proporsi kelas.

5. MODELING

Algoritma yang digunakan adalah Random Forest Classifier karena mampu memberikan akurasi tinggi, bekerja baik untuk data tabular, dan menghasilkan interpretasi berupa feature importance.

Model dioptimalkan menggunakan GridSearchCV untuk memilih parameter terbaik: n_estimators, max_depth, dan min_samples_split.

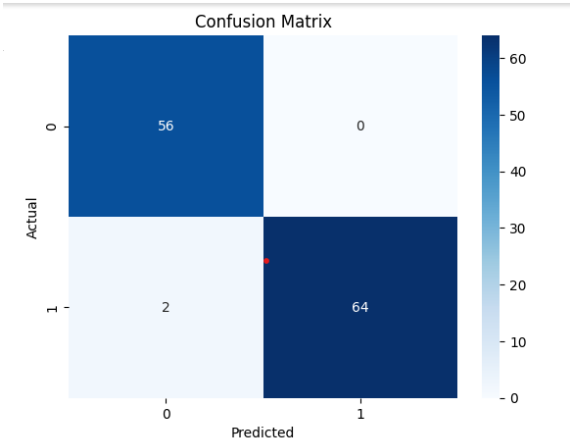
Visualisasi feature importance ditampilkan dalam grafik batang horizontal untuk menunjukkan kontribusi setiap fitur.



Gambar 5 Feature Important

6. EVALUATION

Confusion Matrix menunjukkan performa model yang sangat baik: 56 pasien sehat dan 64 pasien sakit diprediksi dengan benar, hanya 2 kesalahan (false negative).



Gambar 6 Confusion Matrix

Classification Report:

- Accuracy: 98.36%
- Precision: 0.97 (kelas 0), 1.00 (kelas 1)
- Recall: 1.00 (kelas 0), 0.97 (kelas 1)
- F1-Score: 0.98 untuk kedua kelas

Classification Report:

	precision	recall	f1-score	support
0	0.97	1.00	0.98	56
1	1.00	0.97	0.98	66
accuracy			0.98	122
macro avg	0.98	0.98	0.98	122
weighted avg	0.98	0.98	0.98	122

Gambar 7 Classification report

Model mampu mendeteksi pasien sakit dengan sangat baik (recall 0.97), dan tidak ada kesalahan pada prediksi pasien sehat (recall 1.00). Ini penting dalam konteks medis.

7. KESIMPULAN

Model prediksi penyakit jantung dengan Random Forest berhasil dibangun dan menunjukkan kinerja sangat tinggi (akurasi 98.36%).

Fitur-fitur penting yang memengaruhi klasifikasi antara lain cp, thal, oldpeak, dan thalach.

Tujuan proyek tercapai: memprediksi risiko penyakit jantung secara akurat berdasarkan data medis numerik.

Kelebihannya tidak memerlukan preprocessing kompleks, akurasi tinggi, dan hasil dapat diinterpretasikan.

Sedangkan keterbatasan dataset kecil, hanya berasal dari Cleveland, belum diuji pada populasi Indonesia.

Rekomendasi:

- Uji model pada data real dari rumah sakit lokal
- Bandingkan dengan algoritma lain (XGBoost, SVM)
- Bangun aplikasi berbasis web/Android sederhana

8. DAFTAR PUSTAKA

Alfajr, Nur Halizah, and Sofi Defiyanti. 2024. “METODE RANDOM FOREST DAN PENERAPAN PRINCIPAL COMPONENT ANALYSIS (PCA).” 12(3).

Depari, Deo Haganta, Yuni Widiastiwi, and Mayanda Mega Santoni. 2022. “Perbandingan Model Decision Tree, Naive Bayes Dan Random Forest Untuk Prediksi Klasifikasi Penyakit Jantung.” *Informatik : Jurnal Ilmu Komputer* 18(3): 239. doi:10.52958/iftk.v18i3.4694.

Elektronik, Jurnal, Ilmu Komputer Udayana, Stephania Getrudis, Inaconta Sadipun, Gusti Ngurah Anom, Cahyadi Putra, M Cs, Jalan Raya, and Kampus Unud. 2023. “Analisis Algoritma Random Forest Dalam Memprediksi Penyakit Jantung Koroner.” 11(4): 2654–5101. https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset.

Hidayat, Hidayat, Andi Sunyoto, and Hanif Al Fatta. 2023. “Klasifikasi Penyakit Jantung Menggunakan Random Forest Clasifier.” *Jurnal SISKOM-KB (Sistem Komputer dan Kecerdasan Buatan)* 7(1): 31–40. doi:10.47970/siskom-kb.v7i1.464.

Rahmada, Alfin, Erliyan Redy Susanto, Sistem Informasi, Fakultas Teknik, and Universitas Teknokrat Indonesia. 2024. “Peningkatan Akurasi Prediksi Penyakit Jantung Dengan Teknik SMOTEENN Pada Algoritma Random Forest Improving Heart Disease Prediction Accuracy with SMOTEENN Technique on Random Forest Algorithm.” 4(12): 795–803.