

# BAB 1

## PENDAHULUAN

### 1.1 Latar Belakang Masalah

Penyakit jantung merupakan salah satu penyebab utama kematian di dunia. Menurut World Health Organization (WHO), sekitar 17,9 juta orang meninggal setiap tahunnya akibat penyakit kardiovaskular, yang mencakup serangan jantung dan stroke [1]. Sebagian besar kematian ini terjadi secara mendadak, tanpa gejala yang terlihat sebelumnya. Hal ini menunjukkan pentingnya deteksi dini agar risiko penyakit jantung dapat diketahui sejak awal dan ditindaklanjuti secara tepat.

Seiring berkembangnya teknologi komputasi dan tersedianya data medis dalam jumlah besar, pendekatan berbasis *machine learning* (ML) mulai banyak diterapkan dalam pengembangan sistem pendukung keputusan klinis. ML memiliki kemampuan untuk menganalisis data kompleks secara otomatis dan menemukan pola tersembunyi yang sulit dikenali oleh metode konvensional [2]. Dalam konteks kesehatan, ML telah digunakan dalam berbagai aplikasi, termasuk prediksi diabetes, kanker, penyakit ginjal, dan penyakit jantung [3, 4].

Algoritma *Random Forest* (RF) menjadi salah satu metode yang paling banyak digunakan dalam klasifikasi medis karena kemampuannya menangani fitur heterogen, mengurangi risiko *overfitting*, dan menghasilkan hasil prediksi yang stabil [5]. Selain itu, RF juga menyediakan metrik *feature importance*, yang dapat digunakan untuk mengetahui fitur mana yang paling berpengaruh terhadap prediksi. Hal ini memungkinkan dilakukannya proses seleksi fitur (*feature selection*) untuk menyaring fitur yang benar-benar relevan.

Seleksi fitur menjadi langkah penting dalam pipeline ML, terutama dalam domain medis, karena dapat meningkatkan akurasi, mempercepat proses pelatihan, serta meningkatkan interpretabilitas model [6, 7]. Dalam *Random Forest*, dua pendekatan populer untuk menghitung kepentingan fitur adalah *Mean Decrease in Impurity* (MDI) dan *Mean Decrease in Accuracy* (MDA) [8]. Alternatif lainnya adalah metode seperti *Recursive Feature Elimination* (RFE) atau teknik penalti seperti *L1-regularization* (Lasso) yang juga sering digunakan dalam penelitian medis [9].

Dalam penelitian-penelitian sebelumnya, terbukti bahwa menggabungkan seleksi fitur dengan algoritma RF mampu menghasilkan model klasifikasi penyakit

jantung yang lebih efisien dan interpretatif [10]. Namun, untuk mencapai performa terbaik, model ML juga memerlukan pengaturan parameter yang optimal (hyperparameter tuning). Parameter seperti jumlah pohon (n\_estimators), kedalaman maksimum pohon (max\_depth), dan minimum sampel per node dapat memengaruhi kinerja model secara signifikan [11]. Teknik pencarian seperti GridSearchCV atau RandomizedSearchCV telah terbukti efektif dalam menentukan kombinasi parameter terbaik dalam waktu yang efisien [12].

Penelitian ini bertujuan untuk membangun model prediksi penyakit jantung berbasis algoritma Random Forest yang dioptimalkan melalui proses seleksi fitur. Dataset yang digunakan merupakan dataset penyakit jantung dari UCI Machine Learning Repository [13]. Dengan pendekatan ini, diharapkan dapat diperoleh model klasifikasi yang akurat, efisien, dan memiliki interpretabilitas tinggi sehingga dapat dimanfaatkan sebagai sistem pendukung keputusan di lingkungan medis.

## 1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan sebelumnya, dapat disimpulkan bahwa terdapat kebutuhan akan sistem yang mampu membantu dalam mendeteksi penyakit jantung secara lebih dini dan akurat dengan memanfaatkan metode pembelajaran mesin. Oleh karena itu, penelitian ini dirancang untuk menjawab pertanyaan-pertanyaan utama yang menjadi fokus kajian, sebagaimana dirumuskan dalam poin-poin berikut.

- Bagaimana membangun model deteksi penyakit jantung berbasis Random Forest yang akurat?
- Seberapa besar pengaruh seleksi fitur terhadap performa model?
- Fitur klinis apa saja yang paling berkontribusi terhadap prediksi menurut skor *feature importance*?

## 1.3 Batasan Permasalahan

Untuk menjaga fokus penelitian dan menghindari ruang lingkup yang terlalu luas, penelitian ini dibatasi pada aspek-aspek tertentu yang relevan dengan tujuan utama. Batasan ini ditetapkan agar proses pengumpulan data, pengolahan, serta analisis dapat dilakukan secara lebih terarah dan sistematis. Adapun batasan yang dimaksud adalah sebagai berikut.

- Data penelitian menggunakan dataset Heart Disease UCI (Cleveland, Hungarian, Switzerland, Long–Beach–VA).
- Variabel target dibinarisasi: 0 (tidak sakit) dan 1 (sakit).
- Algoritma yang diuji hanya Random Forest.
- Evaluasi menggunakan *train/test split* 80/20 dan metrik Accuracy, ROC–AUC, Confusion Matrix.
- Bahasa pemrograman Python 3.11 dengan scikit-learn  $\geq 0.24$ .

#### 1.4 Tujuan Penelitian

Penelitian ini bertujuan untuk memberikan solusi terhadap permasalahan yang telah dirumuskan sebelumnya. Tujuan-tujuan tersebut difokuskan pada pengembangan model prediktif menggunakan metode pembelajaran mesin dan langkah-langkah analisis yang menyertainya, seperti dijelaskan pada poin-poin berikut.

- Membangun model deteksi penyakit jantung menggunakan RF.
- Menguji pengaruh seleksi fitur berbasis *feature importance* terhadap performa model.
- Mengidentifikasi fitur klinis paling relevan.

#### 1.5 Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan manfaat baik secara teoritis maupun praktis. Manfaat yang diperoleh tidak hanya dirasakan oleh peneliti dan pengembang sistem, tetapi juga oleh institusi kesehatan dan masyarakat umum. Adapun manfaat yang dimaksud adalah sebagai berikut.

- **Manfaat Akademis:** Penelitian ini menambah literatur mengenai penerapan algoritma Random Forest dalam klasifikasi medis, khususnya untuk deteksi penyakit jantung. Selain itu, penelitian ini menunjukkan bagaimana metode seleksi fitur berbasis *feature importance* dapat digunakan untuk meningkatkan efisiensi dan akurasi model.

- **Manfaat Praktis:** Hasil dari penelitian ini dapat menjadi dasar pengembangan sistem pendukung keputusan (decision support system) untuk membantu tenaga kesehatan dalam memprediksi risiko penyakit jantung berdasarkan data klinis pasien.
- **Manfaat Klinis:** Dengan mengidentifikasi fitur-fitur klinis yang paling berpengaruh terhadap prediksi, penelitian ini dapat membantu dalam prioritas pemeriksaan, penanganan dini, dan edukasi risiko terhadap pasien yang rentan mengalami penyakit jantung koroner.

## 1.6 Sistematika Penulisan

Untuk mempermudah pemahaman terhadap isi skripsi ini, penulisan disusun secara sistematis dalam beberapa bab. Masing-masing bab disusun secara runtut dan saling berkaitan satu sama lain guna menjelaskan langkah-langkah penelitian dari awal hingga kesimpulan. Sistematika penulisan laporan ini adalah sebagai berikut.

- **Bab 1 PENDAHULUAN** berisi latar belakang masalah, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, dan sistematika penulisan.
- **Bab 2 LANDASAN TEORI** menjelaskan teori-teori yang digunakan, seperti penyakit jantung, pembelajaran mesin, *decision tree*, *random forest*, *randomized search*, dan *confusion matrix*.
- **Bab 3 METODOLOGI PENELITIAN** berisi penjelasan mengenai alur penelitian, pengumpulan data, pembangunan model, dan tahap perancangan serta pembangunan sistem.
- **Bab 4 HASIL DAN DISKUSI** berisi hasil implementasi dan analisis dari pembangunan model *machine learning* serta sistem yang dibangun.
- **Bab 5 SIMPULAN DAN SARAN** berisi simpulan dari penelitian yang dilakukan serta saran untuk pengembangan dan penelitian selanjutnya.