

BAB I

PENDAHULUAN

1.1 Latar Belakang

Penyakit kardiovaskular (CVD) merupakan krisis kesehatan global dan penyebab utama kematian di seluruh dunia. Berdasarkan data terbaru Survei Kesehatan Indonesia (SKI) tahun 2023, prevalensi penyakit jantung masih menjadi perhatian serius dengan angka nasional 0,85%, bahkan mencapai 1,67% di wilayah tertentu seperti Yogyakarta [1]. Menurut data dari WHO kematian pada tahun 2022 mencapai 19,8 juta kematian akibat penyakit kardiovaskular (CVD), mewakili sekitar 32% angka kematian global [2].

Seiring perkembangan teknologi, *machine learning* menawarkan solusi potensial untuk membantu mengidentifikasi individu dengan profil berisiko tinggi berdasarkan analisis data historis pasien [3]. Untuk mendalami analisis faktor risiko ini, penelitian ini akan menggunakan dataset publik penyakit jantung yang diperoleh dari platform Kaggle. Dataset ini terdiri dari 1.190 baris data pasien dengan 12 atribut di dalamnya, termasuk usia, jenis kelamin, tekanan darah, kolesterol, dan status penyakit jantung.

Terdapat beberapa metode yang dapat digunakan untuk klasifikasi, di antaranya adalah metode *Logistic regression*, *Decision tree*, dan *Random forest*. *Logistic regression* digunakan untuk menggambarkan data dan untuk menjelaskan hubungan antara satu variabel biner dependen dan satu atau lebih variabel independen nominal, ordinal, interval atau rasio tingkat [4]. *Decision tree* juga seringkali digunakan untuk membuat keputusan klasifikasi. Cara kerja algoritma ini dimulai dengan menilai bobot setiap atribut dengan perhitungan entropi (Persamaan 1), kemudian menghitung keterkaitan antar atribut menggunakan *information gain* [5]. *Random forest* juga bisa diartikan terbentuk dari sekumpulan *Decision tree* atau pohon Keputusan [5].

Penelitian ini menggunakan metrik presisi, recall, dan f1-score untuk mengukur performa kedua algoritma. Hasil penelitian menunjukkan bahwa Algoritma Random Forest memiliki akurasi yang lebih tinggi dibandingkan dengan

Logistic Regression, dengan nilai akurasi 95%, presisi 95%, recall 95% dan f1-score 95%. Di sisi lain, algoritma Logistic Regression memiliki nilai akurasi sebesar 85%, presisi 85%, recall 85% dan f1-score 85% [6]. Berdasarkan latar belakang di atas peneliti yang di ajukan sebuah penelitian yang berfokus untuk membandingkan kinerja *Logistic regression*, *Decision tree*, dan *Random forest* dalam klasifikasi risiko penyakit jantung.

1.2 Rumusan Masalah

Berdasarkan konteks tersebut, penelitian ini berfokus pada perbandingan Algoritma Logistic Regression, Decision Tree dan Random Forest untuk mengukur seberapa akurat Algoritma tersebut dalam klasifikasi risiko penyakit jantung. Berdasarkan uraian tersebut, rumusan masalah dalam penelitian ini adalah sebagai berikut "Bagaimana perbandingan kinerja (*Accuracy*, *Precision*, *Recall*, *F1-Score*) antara algoritma Logistic Regression, Decision Tree, dan Random Forest dalam melakukan klasifikasi risiko penyakit jantung pada dataset publik UCI Heart Disease"

1.3 Batasan Masalah

Pada skripsi ini, untuk menghindari ruang lingkup yang terlalu jauh, adapun batasan masalah yang digunakan adalah sebagai berikut :

1. Penelitian ini menggunakan dataset publik UCI Heart Disease yang diperoleh dari platform Kaggle, yang terdiri dari 1.190 baris data dan 12 atribut. Penelitian ini tidak menggunakan data primer atau dataset lain di luar sumber tersebut.
2. Algoritma *machine learning* yang dibandingkan terbatas pada *Logistic regression*, *Decision tree*, dan *Random forest*. Algoritma klasifikasi lainnya tidak diuji dalam penelitian ini.
3. Kinerja model dievaluasi dan dibandingkan hanya berdasarkan metrik *accuracy*, *precision*, *recall*, *F1-score*. Metrik evaluasi lainnya seperti AUC-ROC, *specificity*, atau *log-loss* tidak menjadi fokus utama perbandingan.

1.4 Tujuan Penelitian

Berikut adalah beberapa tujuan dalam penelitian ini, antara lain :

“Menentukan algoritma yang memberikan performa terbaik dalam klasifikasi risiko penyakit jantung di antara *Logistic regression*, *Decision tree*, dan *Random forest*, dengan menggunakan metrik evaluasi *accuracy*, *precision*, *recall*, *F1-score* pada dataset yang sama”.

1.5 Manfaat Penelitian

Manfaat dari penelitian ini dapat diharapkan :

1. Meningkatkan pemahaman dalam menggunakan machine learning dalam mengklasifikasi risiko penyakit jantung.
2. Membandingkan Algoritma terbaik dalam mengklasifikasi.
3. Diharapkan memberikan referensi dan pemahaman dalam bahayanya risiko penyakit jantung jika tidak di tangani secara dini.

1.6 Sistematika Penulisan

Untuk memperoleh gambaran yang lebih mudah dipahami mengenai isi penulisan skripsi ini , sistematika penulisan ditulis dengan menguraikan bab demi bab, meliputi :

BAB I PENDAHULUAN

Berisikan tentang latar belakang, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian dan sistematika penulisan.

BAB II TINJAUAN PUSTAKA

Berisi tentang studi literatur, dasar teori yang digunakan dalam skripsi yang berasal dari jurnal, buku, dan internet yang menjadi dasar dari penelitian ini.

BAB III METODE PENELITIAN

Bab ini akan membahas lebih detail tahapan dalam membandingkan kinerja algoritma Logistic Regression, Decesion Tree dan Random Forest yang dihasilkan.

BAB IV HASIL DAN PEMBAHASAN

Bab ini akan membahas mengenai proses dan hasil dari analisis kinerja algoritma Logistic Regression, Decesion Tree dan Random Forest yang telah dilakukan, menentukan algoritma mana yang terbaik dalam mengklasifikasi risiko penyakit jantung.

BAB V PENUTUP

Berisi tentang kesimpulan yang diambil dari proses penelitian dan saran dari penulis

terkait pengembangan dan progres lanjutan untuk mendapatkan data yang lebih spesifik.

