BAB I
PENDAHULUAN

1.1 Latar Belakang

Penyakit jantung adalah kondisi di mana bagian-bagian jantung seperti pembuluh darah, selaput, katup, dan otot jantung mengalami masalah. Penyakit ini bisa disebabkan oleh berbagai faktor, seperti penyumbatan pada pembuluh darah jantung, peradangan, infeksi, atau kelainan bawaan [1]. Sebelum pandemi *Covid-19*, penyakit jantung adalah salah satu penyakit penyebab kematian utama. Dalam data yang diterbitkan oleh WHO pada tahun 2021, kematian akibat penyakit jantung mencapai angka 17,8 juta kematian atau satu dari tiga kematian di dunia setiap tahun disebabkan oleh penyakit jantung [2]. Faktor risiko penyakit jantung meliputi gaya hidup tidak sehat seperti konsumsi makanan tinggi karbohidrat atau lemak, obesitas, kurang berolahraga, merokok, serta riwayat keluarga yang berperan signifikan dalam meningkatkan risiko penyakit jantung [3]. Diagnosis dini penyakit jantung sangat penting untuk meningkatkan kualitas hidup seseorang. Namun, proses diagnosis sering kali kompleks dan memerlukan pemeriksaan medis yang mendalam dan waktu yang tidak singkat. Oleh karena itu, diperlukan sebuah metode yang efisien untuk membantu dalam proses diagnosis penyakit jantung.

Salah satu metode yang dapat digunakan adalah klasifikasi berbasis *machine learning*. Klasifikasi adalah pengelompokkan data ke dalam kategori atau kelas tertentu berdasarkan fitur-fitur yang dimiliki oleh data tersebut [4]. Metode ini efektif karena mampu memproses data pasien dengan cepat dan menghasilkan prediksi yang jelas mengenai kemungkinan seseorang menderita penyakit jantung [5]. Di antara berbagai algoritma *machine learning* yang tersedia, *Random Forest* merupakan salah satu algoritma yang menunjukkan kinerja cukup baik dalam aplikasi medis, termasuk dalam klasifikasi penderita penyakit jantung [6]. Algoritma ini dipilih karena kemampuannya dalam menangani data yang kompleks dan bervariasi. *Random Forest* adalah algoritma *ensemble* yang menggabungkan

hasil dari banyak pohon keputusan untuk meningkatkan akurasi prediksi dan klasifikasi. Keunggulan utama dari *Random Forest* adalah kemampuannya dalam menangani *overfitting* serta memberikan hasil yang stabil meskipun data yang digunakan memiliki karakteristik yang beragam [7]. Algoritma ini dapat digunakan untuk mengklasifikasi kemungkinan seseorang menderita penyakit jantung atau tidak berdasarkan faktor-faktor risiko tertentu.

Pada penelitian sebelumnya yang dilakukan oleh (Fauzi dkk, 2023) berjudul "*Penerapan Algoritma K-Nearest Neighbor Dalam Klasifikasi Penyakit jantung*" didapatkan hasil akurasi sebesar 92,78% [8]. Metode dengan algoritma ini memiliki tingkat keakuratan yang signifikan untuk sebuah kasus klasifikasi penderita penyakit jantung menggunakan salah satu algoritma dari *machine learning*. Kemudian pada penelitian selanjutnya yang dilakukan oleh (Bowo dkk, 2023) dengan judul "*Implementasi Metode Naïve Bayes Untuk Klasifikasi Penderita Penyakit Jantung*" didapatkan hasil akurasi sebesar 86,84% [9]. Metode dengan algoritma ini memiliki tingkat keakuratan yang cukup baik tetapi lebih rendah dari algoritma pada penelitian sebelumnya. Ini menandakan bahwa hasil dari model klasifikasi untuk sebuah kasus penyakit jantung menggunakan metode *machine learning* dapat berbeda tergantung dari algoritma dan dataset yang digunakan.

Oleh karena itu, berdasarkan latar belakang yang sudah disebutkan sebelumnya, peneliti bertujuan untuk membuat dan menganalisis model klasifikasi penderita penyakit jantung menggunakan algoritma *Random Forest*. Model ini diharapkan dapat memberikan hasil klasifikasi yang akurat mengenai kemungkinan seseorang menderita penyakit jantung sehingga dapat membantu petugas medis dalam membuat keputusan yang lebih cepat dan tepat dalam mendiagnosis.

1.2 Maksud dan Tujuan

Maksud dari penelitian ini yaitu membuat model klasifikasi penderita penyakit jantung menggunakan algoritma *Random Forest*. Adapun tujuan dari penelitian ini yaitu menganalisis kinerja algoritma *Random Forest* dan menampilkan hasil akurasi yang diperoleh dari model klasifikasi yang telah dibuat.

1.3 Rumusan Masalah

Berdasarkan latar belakang masalah yang telah dijelaskan sebelumnya maka rumusan masalah dalam penelitian ini yaitu berapa tingkat akurasi model algoritma *Random Forest* yang diterapkan untuk klasifikasi penderita penyakit jantung?

1.4 Batasan Masalah

Dalam penelitian ini, terdapat beberapa batasan yang perlu diperhatikan guna mengarahkan fokus penelitian dan membatasi lingkup penelitian. Batasan-batasan tersebut adalah sebagai berikut:

1. Penelitian ini akan menggunakan tiga dataset yang sudah tersedia dan memiliki karakteristik yang berbeda-beda. Pemilihan dataset yang berbeda ini dilakukan untuk menguji kemampuan model dalam menangani variasi data dan fitur yang beragam pada sebuah data medis.
2. Penelitian ini hanya berfokus pada analisis dalam pembuatan model klasifikasi penderita penyakit jantung menggunakan algoritma *Random Forest*. Penelitian tidak akan mencakup penanganan atau pengobatan penyakit jantung.
3. Pengujian model akan dilakukan dengan berbagai proporsi data latih dan uji, mulai dari 90:10 hingga 50:50. Ini diterapkan untuk melihat bagaimana variasi proporsi data mempengaruhi kinerja model dan untuk menemukan proporsi terbaik yang memberikan hasil yang paling akurat pada setiap dataset.
4. Penelitian ini menggunakan dua metode pengujian yaitu pengujian menggunakan dataset asli dan pengujian menggunakan dataset yang telah di-*oversampling* dengan teknik SMOTE yang dilakukan pada setiap dataset. Hal ini bertujuan untuk mengatasi masalah ketidakseimbangan kelas dalam dataset asli dan untuk melihat seberapa pengaruhkah hasilnya pada model *Random Forest* yang dibuat.
5. Penelitian ini hanya akan melakukan optimasi pada parameter `random_state` dari algoritma *Random Forest*. Nilai `random_state` yang akan digunakan

adalah 2 dan 5, yang merupakan jumlah nilai dari label yang mewakili dari setiap dataset yang digunakan. Optimasi ini bertujuan untuk memastikan model memiliki kinerja yang stabil dan konsisten.

6. Evaluasi kinerja model dalam penelitian ini akan dilakukan dengan mengukur nilai akurasi saja. Metrik akurasi dipilih untuk menilai sejauh mana model dapat mengklasifikasikan data dengan benar, tanpa mempertimbangkan metrik evaluasi lainnya seperti presisi, sensitivitas, spesifisitas, atau *f1-score*.

1.5 Metode Penelitian

Adapun beberapa metode dan langkah-langkah yang digunakan dalam melaksanakan penelitian ini diantaranya:

1. Studi Literatur

   Studi literatur bertujuan untuk mempelajari teori dasar mengenai penyakit jantung, teori pendukung mengenai pengolahan data, metode *machine learning*, algoritma *Random Forest*, teknik klasifikasi dan bahasa pemrograman *Python*.

2. Pengumpulan Data

   Penelitian ini menggunakan tiga dataset berbeda yang tersedia secara publik untuk klasifikasi penderita penyakit jantung. Ketiga dataset ini dipilih karena memiliki karakteristik yang berbeda, yang akan membantu menguji keefektifan model dalam berbagai kondisi data. Data tersebut mencakup parameter medis seperti tekanan darah, kadar kolesterol, usia, dan faktor risiko lainnya.

3. Pra-pemrosesan Data

   Pada tahap ini data akan dibersihkan dari nilai-nilai yang hilang, duplikat, dan juga yang tidak sesuai. Proses ini meliputi penghapusan baris data yang tidak lengkap dan mengubah tipe data yang berupa objek menjadi numerik supaya bisa diproses. Kemudian data pada setiap dataset akan dilakukan *oversampling* dengan teknik SMOTE untuk menyeimbangkan kelas. Selanjutnya data akan dibagi menjadi dua set data latih dan data uji dengan

proporsi yang bervariasi dari 90:10 hingga 50:50. Proporsi ini akan diterapkan pada ketiga dataset untuk mengidentifikasi pengaruh ukuran data latih terhadap kinerja model.

4. Pembuatan Model

   Model klasifikasi akan dibuat menggunakan algoritma *Random Forest*. Algoritma ini dipilih karena kehandalannya dalam proses klasifikasi data yang kompleks dan memberikan hasil yang cukup akurat dan stabil.

5. Optimasi Parameter

   Penelitian ini akan mengoptimasi parameter `random_state` dengan nilai 2 dan 5, yang merupakan jumlah nilai dari label pada dataset yang digunakan. Parameter ini akan diuji untuk menentukan pengaruhnya terhadap kinerja model.

6. Evaluasi Model

   Kinerja model akan dievaluasi menggunakan metrik akurasi. Evaluasi akan dilakukan untuk kedua kondisi yaitu menggunakan dataset asli dan dataset yang sudah di-*oversampling*.

7. Analisis dan Kesimpulan

   Pada tahap ini dilakukan analis terhadap pengujian yang sudah dilakukan untuk memahami kekuatan serta kelemahan dari model yang sudah dibuat. Kemudian berdasarkan hasil evaluasi dan analisis, akan ditarik kesimpulan mengenai efektivitas algoritma *Random Forest* dalam klasifikasi penderita penyakit jantung.

1.6 Sistematika Penulisan

Sistematika penyusunan laporan skripsi ini dilakukan dengan sebagai berikut:

BAB I PENDAHULUAN

Pada bab ini materi yang dibahas mencakup topik yang berkaitan dengan dasar-dasar penulisan skripsi, seperti latar belakang penelitian, maksud dan tujuan penelitian, rumusan masalah, batasan masalah, metode penelitian, dan sistematika penulisan skripsi.

BAB II TINJAUAN PUSTAKA

Pada bab ini materi yang dibahas mencakup penelitian terkait yang sebelumnya pernah dilakukan dan teori-teori yang digunakan untuk mendukung dan mendasari penulisan skripsi ini.

BAB III ANALISIS DAN PERANCANGAN SISTEM

Pada bab ini materi yang dibahas mencakup analisis dan rancangan sistem yang akan dibuat, termasuk perancangan alur sistem, analisis data, analisis dari metode yang akan digunakan, serta spesifikasi kebutuhan sistem.

BAB IV IMPLEMENTASI DAN PENGUJIAN

Pada bab ini materi yang dibahas mencakup hasil implementasi, pengujian sistem dan evaluasi hasil keseluruhan.

BAB V KESIMPULAN DAN SARAN

Pada bab ini materi yang dibahas mencakup kesimpulan dari hasil penelitian serta saran untuk mengoptimalkan sistem dengan harapan sistem tersebut dapat menjadi lebih baik.