BAB I
PENDAHULUAN

1.1. Latar Belakang

Di seluruh dunia, penyakit jantung masih menduduki peringkat teratas sebagai penyebab kematian [1]. Pada tahun 2019, World Health Organization (WHO) memperkirakan kematian yang diakibatkan oleh penyakit jantung mencapai angka 17.9 juta jiwa [2]. Penyakit jantung membawa dampak yang luas, tidak hanya pada kesehatan individu tetapi juga pada kondisi ekonomi dan sosial masyarakat, terutama karena penyakit ini cenderung menyerang kelompok usia produktif [3]. Gejala penyakit jantung pada fase awal seringkali tidak disadari oleh pengidapnya karena mirip dengan gejala penyakit ringan yang dianggap tidak membahayakan, terutama pada pasien yang masih berusia muda dan produktif [4]. Akibatnya, banyak pasien baru melakukan pemeriksaan kondisi jantung ketika gejala sudah cukup parah.

Pada tahap ini, deteksi dini sangat penting dilakukan agar keterlambatan dalam penanganan penyakit jantung dapat dicegah. Salah satunya adalah dengan memanfaatkan teknologi sebagai model prediksi untuk diagnosis dini sehingga dapat mempercepat penanganan dan pengobatan penyakit jantung serta meminimalkan risiko komplikasi. Teknologi yang digunakan salah satunya adalah data mining [5]. Data mining telah diterapkan dalam berbagai bidang ilmu, seperti ekonomi, kedokteran, pendidikan, dan sebagainya. Salah satu pendekatan dalam data mining adalah pemanfaatan algoritma machine learning, yaitu cabang dari kecerdasan buatan yang berfokus pada pembuatan algoritma dan model statistik. Algoritma ini memungkinkan komputer untuk membuat prediksi atau pengambila keputusan berdasarkan data yang dipelajari tanpa perlu diberikan instruksi secara langsung [6].

Namun, penerapan algoritma machine learning juga menghadapi sejumlah tantangan, terutama terkait dengan ketidakseimbangan data pada jumlah kelas mayoritas dan minoritas. Ketidakseimbangan tersebut menyebabkan model menjadi bias terhadap kelas mayoritas, sehingga performa model mengalami

penurunan. Selain itu, tingginya jumlah fitur yang digunakan dalam dataset dapat menyebabkan model menjadi kurang efisien, terutama jika sebagian fitur tersebut tidak relevan atau redundan. Oleh karena itu, penyeimbangan data dan pemilihan fitur sangat penting untuk peningkatan performa model sehingga diperlukan metode atau teknik tambahan. Beberapa peneliti telah melakukan studi terkait penerapan model algoritma berbasis supervided learning untuk deteksi penyakit jantung. Dari berbagai penelitian yang dilakukan, para peneliti mencatat adanya variasi dalam tingkat akurasi yang diperoleh.

Random Forest, sebagai salah satu dari algoritma *machine learning,* terbukti efektif dalam berbagai aplikasi klasifikasi, termasuk di bidang medis [7]. Random Forest merupakan algoritma ensemble learning yang bekerja dengan cara menggabungkan banyak pohon keputusan untuk membuat prediksi. Setiap pohon keputusan dilatih pada subset data yang berbeda dan menggunakan subset fitur yang berbeda, yang membantu mengurangi overfitting dan meningkatkan generalisasi model [8]. Namun, seperti algoritma machine learning lainnya, Random Forest dapat mengalami penurunan kinerja klasifikasi secara signifikan ketika dihadapkan pada dataset yang tidak seimbang [9] dan fitur-fitur yang kurang relevan [10] .

Untuk mengatasi masalah ketidakseimbangan data, metode Synthetic Minority Over-sampling Technique - Edited Nearest Neighbors (SMOTE-ENN) dapat digunakan. SMOTE menghasilkan sampel sintetis baru dari kelas minoritas untuk menyeimbangkan distribusi kelas, sedangkan ENN membersihkan noise dan tumpang tindih antara kelas [11]. Metode ini telah terbukti efektif dalam meningkatkan kinerja model klasifikasi pada dataset medis yang tidak seimbang, termasuk dalam prediksi penyakit [12]. Dengan menerapkan SMOTE-ENN, diharapkan model Random Forest dapat belajar dari data yang lebih seimbang, sehingga akurasi sistem prediksi untuk pasien dengan penyakit jantung dapat meningkat.

Selain itu, fitur-fitur yang kurang relevan dapat ditangani dengan menggunakan metode seleksi fitur, salah satunya adalah RFECV yang telah terbukti dapat meningkatkan kinerja model klasifikasi [13]. Recursive Feature Elimination with Cross-Validation (RFECV) adalah metode seleksi fitur yang efektif yang secara

rekursif menghilangkan fitur-fitur yang kurang penting, memungkinkan model untuk fokus pada subset fitur yang paling relevan. Dengan mengurangi dimensi data dan berfokus pada fitur-fitur yang relevan, RFECV dapat meningkatkan efisiensi komputasi dan mengurangi risiko overfitting pada model.

Dengan memanfaatkan SMOTE-ENN untuk penyeimbangan data dan RFECV untuk seleksi fitur, penelitian ini bertujuan untuk mengoptimalkan kinerja algoritma Random Forest dalam prediksi penyakit jantung. Penelitian ini akan menggunakan lima dataset dari UCI Machine Learning Repository, yaitu Cleveland, Hungarian, Long Beach VA, Switzerland, dan Statlog untuk mengevaluasi generalisasi model. Evaluasi model akan dilakukan menggunakan beberapa metrik meliputi akurasi, sensitivitas, spesifisitas, dan F1-score.

1.2.Rumusan Masalah

Dari penjelasan latar belakang sebelumnya, dapat dirumuskan beberapa masalah utama yang akan dibahas dalam penelitian ini sebagai berikut:

1. Bagaimana pengaruh implementasi balancing data dengan metode *Synthetic Minority Over-sampling Technique - Edited Nearest Neighbors* pada kinerja model prediksi penderita penyakit jantung?
2. Bagaimana pengaruh implementasi seleksi fitur dengan metode *Recursive Feature Elimination with Cross-Validation* pada model prediksi penderita penyakit jantung?
3. Apakah kombinasi antara algoritma *Random Forest* dengan *SMOTE-ENN* dan *RFECV* dapat meningkatkan kinerja model prediksi penderita penyakit jantung?

1.3. Tujuan Penelitian

Berdasarkan permasalahan yang telah jelaskan, penelitian ini memiliki tujuan sebagai berikut:

1. Mengetahui pengaruh implementasi balancing data dengan *Synthetic Minority Over-sampling Technique - Edited Nearest Neighbors* pada kinerja model prediksi penderita penyakit jantung.

2. Mengetahui pengaruh seleksi fitur dengan *Recursive Feature Elimination with Cross-Validation* pada model prediksi penderita penyakit jantung.
3. Menguji kinerja kombinasi antara *Random Forest* dengan *SMOTE-ENN* dan *RFECV* dalam meningkatkan kinerja model klasifikasi penderita penyakit jantung.

1.4. Manfaat Penelitian

Dari penelitian ini diharapkan dapat memberikan manfaat sebagai berikut:

1. Diharapkan evaluasi kombinasi antara *Random Forest* dengan *SMOTE-ENN* dan *RFECV* yang dilakukan pada penelitian ini dapat memberikan wawasan tambahan mengenai efektivitas pendekatan tersebut dalam meningkatkan akurasi model klasifikasi.
2. Menyediakan sumber referensi untuk penelitian lanjutan di bidang klasifikasi penyakit jantung, pengolahan data medis, dan penerapan machine learning dalam kesehatan, terutama yang melibatkan algoritma Random Forest dan teknik balancing data SMOTE-ENN serta seleksi fitur RFECV.
3. Dapat dijadikan sebagai referensi penelitian bagi para praktisi di bidang terkait. Dengan memperhatikan tujuan bisnis, distribusi kelas, dan memilih metode yang terbaik, diharapkan dapat membantu pengambilan keputusan terhadap klasifikasi penyakit jantung.

1.5. Batasan Masalah

Dalam penelitian ini terdapat beberapa batasan yang digunakan untuk memastikan penelitian berfokus pada permasalahan yang diteliti. Batasan yang dimaksud sebagai berikut:

1. Dataset yang digunakan merupakan dataset sekunder yang diperoleh dari situs penyedia dataset open source, yaitu UCI Machine Learning Repository. Dataset ini terdiri dari total 1.190 sampel data pasien yang disimpan dalam format file DATA.

2. Penelitian ini menggunakan algoritma Random Forest sebagai metode klasifikasi utama, sehingga tidak mencakup penggunaan algoritma lain.
3. Data akan diproses menggunakan teknik balancing data SMOTE-ENN untuk mengatasi masalah ketidakseimbangan kelas, serta seleksi fitur menggunakan metode RFECV.
4. Terdapat 14 atribut yang digunakan, antara lain: Age, Sex, Cp, Trestbps, Chol, Fbs, Restecg, Thalach, Exang, Oldpeak, Slope, Ca, Thal, dan Target.

Halaman ini sengaja dikosongkan