

BAB 1

PENDAHULUAN

1.1 Latar Belakang Masalah

Penyakit jantung merupakan salah satu penyebab utama kematian di dunia. Organisasi Kesehatan Dunia melaporkan bahwa penyakit jantung dan pembuluh darah menyumbang lebih dari 17,9 juta kematian setiap tahun secara global. [1] Kondisi ini menunjukkan bahwa penyakit jantung masih menjadi tantangan besar dalam bidang kesehatan masyarakat. Deteksi dini terhadap risiko penyakit jantung menjadi sangat penting karena dapat membantu menurunkan tingkat keparahan penyakit, mengurangi biaya pengobatan, serta meningkatkan kualitas hidup pasien.

Di Indonesia, penyakit jantung juga menjadi permasalahan kesehatan yang serius. Berdasarkan Riset Kesehatan Dasar tahun 2018, prevalensi penyakit jantung berdasarkan diagnosis dokter mencapai 1,5% atau sekitar 15 dari 1.000 penduduk [2]. Beberapa provinsi dengan prevalensi tertinggi antara lain Kalimantan Utara sebesar 2,2%, Daerah Istimewa Yogyakarta sebesar 2,0%, dan Gorontalo sebesar 2,0%. Data dari Kementerian Kesehatan Republik Indonesia menunjukkan bahwa jumlah kematian akibat penyakit kardiovaskular mencapai sekitar 651.481 orang per tahun, sehingga menjadikan penyakit ini sebagai penyebab kematian tertinggi di Indonesia. Selain itu, penyakit kardiovaskular menyumbang sekitar 38% dari total kematian pada tahun 2019 [2]. Studi epidemiologi juga memperkirakan bahwa sekitar 29,2% populasi dewasa berusia 40 tahun ke atas memiliki risiko tinggi terkena penyakit jantung atau stroke dalam kurun waktu sepuluh tahun mendatang.

Uraian tersebut menunjukkan adanya kesenjangan antara tingginya beban penyakit jantung dengan keterbatasan pendekatan deteksi dini yang tersedia saat ini. Permasalahan inti yang mendasari penelitian ini adalah bahwa model prediksi berbasis machine learning yang telah ada umumnya belum menangani ketidakseimbangan kelas secara memadai, sehingga akurasi tinggi yang dilaporkan dapat menyembunyikan kemampuan deteksi (recall) yang rendah pada kasus penyakit jantung; selain itu, model prediksi yang akurat sekalipun jarang disertai penjelasan yang dapat dipercaya secara klinis, sehingga sulit diadopsi oleh tenaga medis dalam praktik sehari-hari. Kombinasi kedua celah tersebut menjadi alasan utama mengapa prediksi penyakit jantung dipilih sebagai objek penelitian, dengan fokus tidak hanya pada akurasi model, tetapi juga pada penanganan

ketidakseimbangan kelas dan transparansi interpretasi.

Perkembangan teknologi informasi, khususnya di bidang *machine learning*, membuka peluang baru dalam pengolahan dan analisis data kesehatan secara lebih efisien. *Machine learning* memungkinkan penggalian pola tersembunyi dari data dalam jumlah besar untuk menghasilkan model prediksi yang andal. Salah satu algoritma *machine learning* yang banyak digunakan dan terbukti efektif dalam menangani permasalahan klasifikasi adalah *Random Forest* [3]. Algoritma ini merupakan metode *ensemble learning* yang membangun banyak pohon keputusan untuk menghasilkan model prediksi yang lebih stabil dan akurat, serta relatif tahan terhadap overfitting dibandingkan model pohon tunggal.

Salah satu tantangan utama dalam membangun model prediksi penyakit jantung adalah masalah ketidakseimbangan kelas pada dataset. *Dataset UCI Heart Disease* yang menjadi standar benchmark dalam penelitian ini memiliki distribusi kelas yang tidak seimbang antara pasien yang terdiagnosis penyakit jantung dan yang tidak. Kondisi ini dapat menyebabkan model *machine learning* menjadi bias, yaitu cenderung memprediksi kelas mayoritas sehingga kemampuan model dalam mendeteksi kasus penyakit jantung (kelas minoritas) menjadi rendah [4]. Padahal dalam konteks medis, kegagalan mendeteksi kasus positif (false negative) memiliki konsekuensi yang sangat serius bagi keselamatan pasien.

Berbagai teknik resampling telah dikembangkan untuk mengatasi permasalahan ketidakseimbangan kelas ini, di antaranya *SMOTE* (*Synthetic Minority Over-sampling Technique*), *SMOTE-ENN* (*SMOTE combined with Edited Nearest Neighbor*), dan *SMOTE-Tomek* (*SMOTE combined with Tomek Links*). Ketiga teknik ini memiliki pendekatan yang berbeda dalam menyeimbangkan distribusi kelas: *SMOTE* hanya melakukan oversampling pada kelas minoritas, *SMOTE-ENN* menggabungkan oversampling dengan pembersihan data ambigu menggunakan *ENN*, sedangkan *SMOTE-Tomek* membersihkan batas antar kelas menggunakan *Tomek Links* [5]. Meskipun demikian, penelitian yang secara khusus membandingkan efektivitas ketiga teknik resampling tersebut pada dataset *UCI Heart Disease* secara komprehensif masih sangat terbatas. Sebagian besar penelitian terdahulu hanya menerapkan satu teknik resampling tanpa membandingkannya dengan alternatif lain, sehingga belum dapat disimpulkan teknik mana yang paling efektif untuk konteks prediksi penyakit jantung.

Selain permasalahan ketidakseimbangan kelas, aspek interpretabilitas model merupakan isu krusial yang sering diabaikan dalam penelitian *machine learning* di bidang kesehatan. Tenaga medis tidak hanya membutuhkan angka akurasi yang

tinggi, tetapi juga penjelasan yang logis dan dapat dipahami mengenai faktor klinis apa saja yang mendorong suatu prediksi. Model yang bersifat *black-box* sulit diterima dalam praktik klinis karena dokter memerlukan transparansi untuk memvalidasi keputusan medis secara profesional dan etis [6]. Penelitian terdahulu berhenti pada evaluasi metrik kuantitatif semata tanpa memberikan interpretasi yang bagi penggunaanya.

Pendekatan Explainable AI (XAI), khususnya metode SHAP (SHapley Additive Explanations), hadir sebagai solusi untuk menjembatani kesenjangan antara performa model dan interpretabilitas klinis. SHAP mampu memberikan penjelasan berbasis kontribusi setiap fitur terhadap output prediksi model secara lokal (per individu pasien) maupun global (keseluruhan dataset), sehingga hasil prediksi model dapat dipahami dan diverifikasi oleh tenaga medis [6]. Integrasi SHAP ke dalam pipeline prediksi penyakit jantung dapat meningkatkan kepercayaan dan penerimaan model dalam lingkungan klinis yang sesungguhnya.

Pemilihan algoritma Random Forest pada penelitian ini didasarkan pada karakteristiknya sebagai metode *ensemble learning* yang relatif tahan terhadap *overfitting* dibandingkan model pohon keputusan tunggal [7], mampu menangani kombinasi fitur numerik dan kategorikal tanpa memerlukan asumsi distribusi data tertentu, serta memiliki dukungan algoritma interpretasi TreeSHAP yang efisien secara komputasi untuk model berbasis pohon [8]. Perbandingan tiga teknik *resampling* (SMOTE, SMOTE-ENN, dan SMOTE-Tomek) dilakukan karena, sebagaimana telah diuraikan pada paragraf sebelumnya, belum terdapat konsensus mengenai teknik *resampling* yang paling sesuai untuk dataset UCI Heart Disease multi-sumber; ketiga teknik tersebut mewakili strategi yang berbeda, yaitu *oversampling* murni (SMOTE) [5] dan *oversampling* yang dikombinasikan dengan pembersihan data pada batas antar kelas (SMOTE-ENN dan SMOTE-Tomek), sehingga perbandingan sistematis diperlukan untuk mengetahui *trade-off* masing-masing teknik terhadap *Recall*, yang merupakan metrik paling kritis dalam konteks medis karena berkaitan langsung dengan risiko *false negative*. Sementara itu, interpretasi berbasis SHAP dipilih karena model Random Forest bersifat *black-box* yang keputusannya sulit dipercaya dalam praktik klinis tanpa penjelasan yang transparan mengenai kontribusi tiap fitur terhadap prediksi [9]; SHAP dipilih dibandingkan metode XAI lain seperti LIME karena memiliki dasar matematis yang konsisten, yaitu nilai Shapley dari teori permainan kooperatif, serta didukung algoritma TreeSHAP yang secara khusus dioptimalkan untuk model berbasis pohon seperti Random Forest [8].

Berdasarkan uraian permasalahan di atas, penelitian ini mengusulkan pengembangan model prediksi penyakit jantung menggunakan algoritma *Random Forest* dengan tiga kontribusi utama: pertama, membandingkan efektivitas tiga teknik resampling (*SMOTE*, *SMOTE-ENN*, dan *SMOTE-Tomek*) untuk menentukan pendekatan terbaik dalam menangani ketidakseimbangan kelas pada dataset UCI Heart Disease; kedua, menerapkan GridSearchCV untuk optimasi hyperparameter model secara sistematis; dan ketiga, mengintegrasikan metode SHAP sebagai pendekatan Explainable AI untuk menghasilkan interpretasi yang transparan dan dapat dipahami oleh tenaga medis. Penelitian ini menggunakan dataset gabungan dari empat sumber UCI Heart Disease (Cleveland, Hungarian, Switzerland, dan VA Long Beach) untuk meningkatkan generalisasi dan keragaman data. Diharapkan penelitian ini memberikan kontribusi yang lebih komprehensif dibandingkan penelitian-penelitian sebelumnya, baik dari segi performa model maupun transparansi hasil prediksi.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, ditemukan adanya kebutuhan akan model prediksi penyakit jantung yang tidak hanya akurat tetapi juga transparan dan dapat diinterpretasikan. Oleh karena itu, penelitian ini dilakukan untuk menjawab pertanyaan-pertanyaan riset sebagai berikut:

- A. Bagaimana pengaruh penggunaan teknik resampling *SMOTE*, *SMOTE-ENN*, dan *SMOTE-Tomek* terhadap performa model *Random Forest* dalam memprediksi penyakit jantung pada dataset UCI Heart Disease?
- B. Teknik resampling manakah yang menghasilkan performa terbaik berdasarkan metrik akurasi, precision, recall, F1-score, dan AUC-ROC pada model *Random Forest* yang telah dioptimasi menggunakan GridSearchCV?
- C. Bagaimana penerapan metode SHAP dapat menghasilkan interpretasi yang transparan terhadap keputusan model *Random Forest* dalam memprediksi penyakit jantung, serta fitur klinis apa yang paling berpengaruh terhadap hasil prediksi?

1.3 Batasan Permasalahan

Agar penelitian lebih terarah, batasan masalah dalam penelitian ini adalah:

- A. Dataset yang digunakan adalah UCI Heart Disease Dataset yang mencakup empat sumber data, yaitu Cleveland, Hungarian, Switzerland, dan VA Long Beach, dengan 14 atribut klinis yang telah digunakan secara luas dalam penelitian prediksi penyakit jantung.
- B. Algoritma klasifikasi yang digunakan adalah Random Forest dengan optimasi menggunakan metode GridSearchCV
- C. Teknik resampling yang dibandingkan dalam penelitian ini meliputi tiga metode, yaitu SMOTE, SMOTE-ENN, dan SMOTE-Tomek.
- D. Evaluasi performa model difokuskan pada metrik klasifikasi yang meliputi akurasi, precision, recall, F1-score, dan AUC-ROC.
- E. Interpretasi model menggunakan metode SHAP (*SHapley Additive Explanations*) yang mencakup analisis pengaruh fitur secara global (keseluruhan dataset) dan lokal (per individu pasien).
- F. Penelitian ini tidak mencakup pengembangan sistem atau aplikasi dan berfokus pada eksperimen pemodelan dan analisis hasil.

1.4 Tujuan Penelitian

Penelitian ini bertujuan untuk:

- A. Membandingkan efektivitas teknik resampling SMOTE, SMOTE-ENN, dan SMOTE-Tomek dalam meningkatkan performa model Random Forest untuk prediksi penyakit jantung pada dataset UCI Heart Disease.
- B. Mengidentifikasi teknik resampling yang paling optimal untuk dikombinasikan dengan algoritma Random Forest yang telah dioptimasi menggunakan GridSearchCV berdasarkan metrik evaluasi yang ditetapkan.
- C. Menerapkan metode SHAP untuk menghasilkan interpretasi yang transparan terhadap hasil prediksi model Random Forest, sehingga dapat diidentifikasi faktor-faktor klinis yang paling berpengaruh terhadap risiko penyakit jantung.

1.5 Manfaat Penelitian

- A. Memberikan kontribusi terhadap pengembangan model prediksi penyakit jantung berbasis *machine learning*, khususnya terkait perbandingan teknik *resampling* dan penerapan *Explainable AI* pada model *Random Forest*.
- B. Menghasilkan perbandingan sistematis antara tiga teknik *resampling* yang dapat menjadi acuan metodologis bagi penelitian serupa pada bidang medis, sekaligus membuka arah baru dalam *feature engineering* berbasis interpretasi model.
- C. Menawarkan model prediksi yang akurat dan dapat diinterpretasi melalui pendekatan SHAP, sehingga berpotensi digunakan sebagai alat bantu pendukung keputusan klinis yang transparan dan berbasis data dalam skrining risiko penyakit jantung secara dini.
- D. Memberikan interpretasi faktor-faktor klinis yang paling berpengaruh terhadap prediksi penyakit jantung, yang dapat dimanfaatkan sebagai landasan pengembangan *Clinical Decision Support System* maupun perancangan program edukasi dan kebijakan pencegahan yang lebih terarah.

1.6 Sistematika Penulisan

Berisikan uraian singkat mengenai struktur isi penulisan laporan penelitian, dimulai dari Pendahuluan hingga Simpulan dan Saran.

Sistematika penulisan laporan adalah sebagai berikut:

A. Bab 1 PENDAHULUAN

Berisi latar belakang permasalahan mengenai tingginya angka kematian akibat penyakit jantung dan kebutuhan akan model prediksi yang akurat sekaligus dapat diinterpretasikan, rumusan masalah, batasan permasalahan, tujuan penelitian, manfaat penelitian, dan sistematika penulisan.

B. Bab 2 LANDASAN TEORI

Memuat kajian teori yang relevan meliputi penyakit jantung dan faktor risikonya, konsep data mining dan machine learning, algoritma Random Forest, ketidakseimbangan kelas pada dataset medis, teknik resampling (SMOTE, SMOTE-ENN, dan SMOTE-Tomek), optimasi hyperparameter

menggunakan GridSearchCV, Explainable AI, metode SHAP, metrik evaluasi model, serta penelitian-penelitian terdahulu yang relevan.

C. Bab 3 METODOLOGI PENELITIAN

Menjelaskan desain dan jenis penelitian, sumber data dan objek penelitian, tahapan penelitian secara keseluruhan, proses preprocessing data, penerapan dan perbandingan ketiga teknik resampling, pelatihan model Random Forest dengan GridSearchCV, evaluasi performa model, serta penerapan metode SHAP untuk interpretasi hasil prediksi.

D. Bab 4 HASIL DAN DISKUSI

Menyajikan dan menganalisis hasil penelitian yang mencakup perbandingan performa ketiga teknik resampling, evaluasi model Random Forest terbaik berdasarkan metrik yang ditetapkan, hasil visualisasi dan interpretasi SHAP baik secara global maupun lokal, serta diskusi mengenai perbedaan hasil penelitian ini dibandingkan dengan penelitian-penelitian terdahulu.

E. Bab 5 KESIMPULAN DAN SARAN

Berisi kesimpulan yang menjawab seluruh rumusan masalah penelitian, serta saran untuk pengembangan penelitian selanjutnya terkait prediksi penyakit jantung berbasis machine learning.

