BAB I
PENDAHULUAN

1.1. Latar Belakang

Jantung adalah salah satu organ dalam tubuh manusia yang berperan untuk memompa darah ke seluruh tubuh. Sehingga fungsinya sangat krusial untuk kelangsungan hidup manusia. Namun, berdasarkan data dari Organisasi Kesehatan Dunia (WHO), penyakit jantung merupakan penyebab kematian teratas di dunia yang mengakibatkan sekitar 17,9 juta kematian setiap tahun dan mencakup sekitar 32% dari total kematian global [1]. Berdasarkan Kementerian Kesehatan Republik Indonesia, pada tahun 2022, penyakit yang paling umum dialami oleh masyarakat Indonesia adalah penyakit jantung dengan total kasus mencapai 15,5 juta, diikuti oleh kanker dengan jumlah 3,2 juta kasus dan stroke sebanyak 2,5 juta kasus. Dari hasil penelitian Kesehatan Dasar (Riskesdas) tahun 2013 dan 2018, menunjukkan ada peningkatan jumlah orang yang mengalami penyakit jantung, dari 0,5% di tahun 2013 menjadi 1,5% di tahun 2018. Selain itu, jumlah kasus serangan jantung pada kelompok usia di bawah 40 tahun meningkat sebesar 2% per tahun dari tahun 2000 hingga 2016. Secara keseluruhan, berdasarkan data dari Institute for Health Metrics and Evaluation (IHME) tahun 2019, jantung dan stroke merupakan penyebab kematian utama, menempati posisi pertama dan kedua. Pada tahun yang sama, jumlah kematian karena penyakit jantung mencapai 9,14 juta jiwa [2].

Penyakit jantung dipengaruhi oleh berbagai faktor risiko yang dibagi menjadi faktor yang tidak dapat diubah dan faktor yang dapat diubah. Faktor yang tidak dapat dimodifikasi termasuk usia, jenis kelamin, dan riwayat keluarga, sedangkan faktor yang bisa diubah meliputi hipertensi, kebiasaan merokok, pola makan, obesitas, kurangnya aktivitas fisik, serta konsumsi alkohol yang berlebihan [3]. Penyakit jantung merupakan kondisi ketika fungsi bagian-bagian jantung, seperti pembuluh darah, selaput jantung, katup, maupun otot jantung mengalami gangguan. Gangguan ini dapat dipicu oleh berbagai faktor, antara lain penyumbatan pada pembuluh darah, peradangan, infeksi, atau kelainan bawaan sejak lahir [4]. Penyakit jantung dapat ditangani

melalui berbagai upaya pencegahan dan pengelolaan faktor risiko. Pemanfaatan teknologi prediksi berbasis data dapat digunakan untuk membantu mengidentifikasi pola yang berkaitan dengan kondisi penyakit jantung berdasarkan sejumlah indikator. Mengingat hal tersebut, berbagai upaya terarah perlu diterapkan untuk mengurangi jumlah kematian yang diakibatkan oleh penyakit jantung, termasuk pemanfaatan teknologi sebagai alat prediksi untuk mendukung prediksi kondisi penyakit jantung sehingga dapat membantu proses pengambilan keputusan. Sebagian besar kasus penyakit jantung baru terdeteksi setelah serangan terjadi, padahal tindakan lebih awal bisa mencegah atau mengurangi kerusakan saraf yang berbahaya.

Kemajuan dalam digitalisasi layanan kesehatan dan teknologi *data mining* membuka peluang dalam pengembangan sistem prediksi penyakit jantung yang akurat [5]. *Data mining* sendiri mencakup beragam metode, di mana *machine learning* menempati posisi penting karena mampu mengembangkan model prediktif yang belajar langsung dari pola data [6]. Penelitian di Indonesia telah menunjukkan bahwa penerapan *machine learning* dalam memprediksi risiko penyakit jantung sangat bermanfaat untuk mendukung tindakan pencegahan yang lebih tepat [7]. *Machine learning* telah banyak diterapkan pada prediksi kesehatan, contoh penelitian pada penyakit *stroke* dengan tingkat akurasi tinggi dengan nilai AUC berkisar antara 0,64 hingga 0,99, dan secara konsisten menunjukkan performa yang lebih baik [8].

Berdasarkan konteks permasalahan di atas, peneliti memanfaatkan salah satu metode *machine learning* yaitu XGBoost dalam memprediksi penyakit jantung. *Extreme Gradient Boosting* (XGBoost) merupakan sebuah metode pengembangan dari teknik *boosting* pohon keputusan, yang termasuk dalam keluarga algoritma *ensemble* [9]. Akan tetapi, penerapan *machine learning* dalam prediksi medis masih menghadapi sejumlah tantangan teknis, terutama dalam memperoleh konfigurasi model yang optimal dan menghasilkan data latih yang representatif. Meskipun distribusi kelas pada dataset penelitian relatif seimbang, proses pembentukan sampel sintetis dan

pembersihan data tetap perlu diuji untuk mengetahui pengaruhnya terhadap proses pembelajaran model. Selain itu, kinerja model *machine learning* sangat tergantung pada penyesuaian *hyperparameter* yang tepat. Tanpa konfigurasi *hyperparameter* yang optimal, model dapat mengalami *underfitting* atau *overfitting* yang berdampak pada kemampuan prediksi. Oleh karena itu, penelitian ini menggunakan SMOTE-ENN untuk melaksanakan *oversampling* sekaligus pembersihan data melalui kombinasi pembentukan sampel sintetis dan *Edited Nearest Neighbors* (ENN) [10]. Selanjutnya, *hyperparameter tuning* menggunakan Optuna diterapkan untuk memperoleh konfigurasi parameter XGBoost yang sesuai melalui proses optimasi otomatis [11]. Kombinasi kedua pendekatan tersebut digunakan untuk menganalisis pengaruh SMOTE-ENN dan optimasi *hyperparameter* Optuna terhadap kinerja XGBoost dalam prediksi penyakit jantung.

Penelitian Pramesti, Anggraeny, dan Saputra pada tahun 2025 berjudul "Klasifikasi Data Tidak Seimbang Menggunakan SMOTE-ENN dengan Optimasi Bayesian" mengembangkan model pendeteksian gagal jantung berbasis XGBoost yang diperkuat dengan teknik penyeimbangan data SMOTE-ENN serta penyetelan *hyperparameter* menggunakan *Bayesian Optimization*. Menggunakan dataset lebih dari 250 ribu sampel dari Kaggle, penelitian ini mengevaluasi performa model melalui metrik akurasi, presisi, recall, dan F1-score. Hasil eksperimen menunjukkan bahwa penerapan SMOTE-ENN disertai *Bayesian Optimization* menghasilkan performa tertinggi, dengan akurasi mencapai 0,9375 dan F1-*score* 0,9425, jauh lebih baik dibandingkan model tanpa penyeimbangan data yang memiliki *recall* sangat rendah. Temuan ini menegaskan bahwa kombinasi teknik *balancing* dan optimasi *hyperparameter* efektif dalam meningkatkan kemampuan model mengidentifikasi kasus gagal jantung pada kelas minoritas [12].

Penelitian yang dilakukan oleh Srinivas pada tahun 2022 mengusulkan model "Hyperparameter Optuna XGBoost" (hyOPTXg), yaitu sistem prediksi penyakit jantung berbasis algoritma XGBoost yang dioptimalkan melalui teknik *hyperparameter tuning* menggunakan *framework* Optuna. Model ini

dievaluasi pada tiga dataset berbeda. Hasilnya menunjukkan bahwa hyOPTXg mampu memberikan kinerja lebih baik dibandingkan model lain yang telah diajukan sebelumnya, dengan akurasi sebesar 94,7% pada Cleveland dataset, 89,3% pada Heart Failure dataset, dan 88,5% pada Heart Disease UCI dataset, sehingga membuktikan bahwa optimasi *hyperparameter* memiliki peran penting dalam meningkatkan kinerja XGBoost pada klasifikasi penyakit kardiovaskular [13].

Berdasarkan studi literatur yang dilakukan oleh peneliti sebelumnya, studi ini akan mengintegrasikan teknik SMOTE-ENN serta metode *Hyperparameter Tuning* OPTUNA pada algoritma XGBoost. Studi ini akan memanfaatkan data sekunder yang berasal dari *open source* UCI Machine Learning Repository yang merupakan kumpalan data dari Cleveland, Hungarian, Long Beach VA, Switzerland dan gabungan dataset Statlog. Data tersebut terdiri dari 1190 sampel yang diperoleh dari data pasien dengan 14 fitur [14]. Langkah-langkah penelitian mencakup proses *preprocessing* data, seleksi fitur, penerapan SMOTE-ENN, pembagian data, serta pemodelan menggunakan XGBoost. Selanjutnya, Optuna digunakan untuk melakukan optimasi *hyperparameter* pada model XGBoost. Penerapan kedua metode tersebut dilakukan melalui beberapa skenario untuk mengetahui pengaruh masing-masing metode terhadap kinerja model.

Selanjutnya, model dievaluasi menggunakan metrik, yaitu *precision*, *recall*, *F1-score*, *accuracy*, dan *Area Under the Curve* (AUC). Pemanfaatan beberapa metrik bertujuan untuk memberikan ilustrasi mengenai kinerja model dalam melaksanakan klasifikasi penyakit jantung. Evaluasi dilakukan pada setiap skenario untuk mengetahui pengaruh penerapan SMOTE-ENN dan *hyperparameter tuning* Optuna terhadap kinerja XGBoost. Dengan demikian, hasil evaluasi diharapkan dapat menunjukkan skenario dengan performa terbaik serta kemampuan model dalam mengidentifikasi kelaspositif dan negatif dengan cara yang optimal. Hasil dari penelitian ini diharapkan dapat menjadi acuan dalam pengembangan *machine learning* yang lebih efisien.

1.2. Rumusan Masalah

Dengan merujuk pada latar belakang yang telah disampaikan sebelumnya, maka fokus penelitian ini dijelaskan dalam rumusan masalah berikut:

1. Bagaimana pengaruh penerapan SMOTE-ENN dalam menangani perbedaan distribusi kelas pada dataset penyakit jantung?
2. Bagaimana pengaruh *hyperparameter tuning* Optuna terhadap optimasi performa algoritma XGBoost dalam prediksi penyakit jantung?
3. Bagaimana kinerja model XGBoost setelah menerapkan teknik SMOTE-ENN dan optimasi *hyperparameter* menggunakan Optuna dalam prediksi penyakit jantung?

1.3. Tujuan Penelitian

Untuk menjawab permasalahan yang sudah diidentifikasi, penelitian ini akan diarahkan pada tujuan berikut:

1. Menganalisis pengaruh penerapan SMOTE-ENN terhadap distribusi kelas dan performa XGBoost pada data penyakit jantung.
2. Menganalisis pengaruh optimasi Optuna terhadap performa algoritma XGBoost dalam melakukan prediksi penyakit jantung.
3. Mengevaluasi kinerja model XGBoost pada setiap skenario pengujian serta menganalisis efektivitas kombinasi SMOTE-ENN dan Optuna dalam meningkatkan performa prediksi penyakit jantung.

1.4. Manfaat Penelitian

Oleh karena itu, penelitian ini diharapkan dapat memberikan manfaat baik dari segi teoretis maupun praktis, antara lain:

1. Penerapan kombinasi algoritma XGBoost dengan teknik SMOTE-ENN dan *hyperparameter tuning* Optuna diharapkan dapat memberikan kontribusi dalam pengembangan metode klasifikasi penyakit jantung berbasis *machine learning*.
2. Menjadi referensi bagi mahasiswa dan peneliti di Indonesia dalam mengembangkan model *machine learning* untuk menangani perbedaan distribusi kelas pada data kesehatan serta mendukung pengembangan penelitian lebih lanjut di bidang prediksi penyakit jantung.

1.5. Batasan Masalah

Dalam studi ini terdapat sejumlah batasan yang diterapkan untuk memastikan penelitian terfokus pada isu yang diteliti. Batasan yang yang dimaksud adalah sebagai berikut:

1. Penelitian berfokus pada klasifikasi penyakit jantung ke dalam dua kelas, yaitu tidak sakit dan sakit.
2. Optimasi *hyperparameter* XGBoost menggunakan Optuna dengan pendekatan *Tree-structured Parzen Estimator* (TPE) dan 150 *trial* pada setiap kombinasi *split* dan *fold*.
3. Penelitian tidak membandingkan XGBoost dengan algoritma klasifikasi lain di luar skenario eksperimen yang telah ditentukan.
4. Penelitian dibatasi pada klasifikasi dan evaluasi kinerja model, tanpa membahas implementasi model secara langsung dalam lingkungan medis.