# BAB I
# PENDAHULUAN

## 1.1. Latar Belakang

Penyakit jantung merupakan istilah umum yang digunakan pada seseorang yang mengalami disfungsi pada kerja fungsi jantung. Terdapat banyak jenis nama dari penyakit jantung seperti jantung koroner, kardiovaskuler, bahkan serangan jantung itu sendiri [1]. Penyakit jantung mengambil peran sebagai mengambil peran utama menjadi penyebab kematian di Indonesia. Berdasarkan data dari *World Health Organization* (WHO) pada tahun 2019, penyakit jantung dapat mencapai 31% dari total penyebab kematian di seluruh belahan dunia, angka kematiannya tersendiri diperkirakan hingga 17,7 juta jiwa pada tiap tahunnya. Angka kematian akibat penyakit kardiovaskular diprediksikan mengalami pertumbuhan tiap tahunnya hingga diperkirakan dapat mencapai 23,3 juta kematian pada 10 tahun mendatang [2]. Sedangkan laporan untuk kematian yang diakibatkan oleh penyakit *karidovaskular* di Indonesia menduduki presentase sebesar 12,9% berdasarkan data Kemenkes RI pada tahun 2017 [3]. Dengan memperhatikan hal ini, maka diperlukannya langkah dini untuk deteksi pada penyakit jantung yang bertujuan membantu mengurangi angka kematian disebabkan oleh penyakit jantung khususnya di Indonesia.

Pada dasarnya, penyakit jantung dapat dihindari dengan menjalani gaya hidup sehat dan menjaga pola makan yang teratur. Selain itu, deteksi dini juga memegang peranan penting dalam mencegah risiko kematian pada penderita penyakit jantung. Cara yang dapat dikatakan efektif untuk saat ini adalah dengan memanfaatkan teknologi informasi sebagai diagnosa dini penyakit jantung, hal ini merupakan situasi yang menantang disebabkan saling ketergantungan dari beberapa faktor [4]. Salah satu metode yang terbukti efektif adalah *data mining*, yang berfungsi sebagai alat yang kuat untuk mengidentifikasi pola dan hubungan tersembunyi dalam kumpulan data berukuran besar [5]. *Data mining* merupakan suatu proses analitik yang memanfaatkan metode-metode dari bidang matematika, statistika, kecerdasan buatan, dan pembelajaran mesin guna memperoleh informasi penting dan pengetahuan tersembunyi dari data dalam jumlah besar [6]. Teknik dalam *data mining* sangat bervariasi, diantaranya yang paling sering digunakan yakni klustering, asosiasi, estimasi, dan klasifikasi [7]. Dalam bidang medis, teknik klasifikasi banyak digunakan

dalam diagnosis dan analisis guna membantu dalam pengambilan keputusan.

Algoritma klasifikasi yang sering digunakan terkait *machine learning*, beberapa diantaranya adalah *Neural Network* (NN), *K-Nearest Neighbor* (KNN), *Naïve Bayes* (NB), *Decision Tree* (DT), dan *Support Vector Machine* (SVM) [7]. Untuk penelitian ini dipilih algoritma *Extreme Gradient Boosting* (XGBoost) sebagai metode klasifikasi untuk penyakit jantung. Algoritma XGBoost adalah salah satu model ensemble dari banyaknya algoritma *machine learning classifier* yang didasari dengan pohon keputusan dengan *Gadient Boosting* sebagai inti [8]. Pemilihan model XGBoost ini dikarenakan memiliki ketahanan terhadap *outlier* yang lebih besar, dengan waktu komputasi yang cepat untuk mendapatkan hasil yang akurat [9].

Untuk mendukung proses optimasi model klasifikasi, dilakukan juga penyetelan hyperparameter guna menemukan kombinasi parameter terbaik yang mampu meningkatkan kinerja model. Beberapa metode *hyperparameter tuning* diantaranya adalah *Random Search, Grid Search, Bayesian Optimization, Particle Swarm Optimization,* dan *Genetic Algorithm* [10]. Mengoptimalkan *hyperparameter* untuk pengklasifikasian adalah kunci untuk menemukan kombinasi terbaik dengan membandingkan hasil dari tiap uji coba [11]. Penelitian ini menggunakan *Grid Search* dengan *Cross-Validation* (GridSearchCV), yaitu sebuah pendekatan sistematis dalam mengeksplorasi berbagai kombinasi parameter melalui validasi silang.

Data yang dapat digunakan faktanya tidak selalu dapat digunakan secara langsung, terutama pada bidang kesehatan sering kali ditemukan kasus data yang tidak seimbang (*imbalance*) dikarenakan data yang memiliki hasil negatif biasanya memiliki kelas mayoritas dibandingkan data dengan hasil positif sebagai kelas minoritas sehingga dapat menyebabkan kesalahan dalam klasifikasi [12]. Untuk mengatasi ketidakseimbangan kelas pada data, dapat dilakukan dengan pendekatan pada algoritma dan pendekatan pada *data processing*. Pendekatan pada algoritma berarti dengan fokus pada perbaikan algoritma yang digunakan tanpa mengubah data. Sedangkan untuk penelitian ini menggunakan pendekatan keduanya yakni pada *data processing* dengan menggunakan teknik *resampling* untuk menyeimbangkan distribusi pada level data, serta pada algoritma dengan menerapkan *hyperparameter tuning*. Teknik *resampling* dikategorikan kedalam tiga jenis, yaitu *oversampling, undersampling,* dan *hybridsampling* atau kombinasi dari keduanya [13].

Berdasakan penelitian terdahulu oleh Yang dan Guan pada tahun 2022 yang mengangkat judul penelitian "A Heart Disease Prediction Model Based on Feature Optimization and Smote-Xgboost Algorithm". Penelitian ini mengusulkan penggunaan prediksi penyakit jantung menggunakan algoritma Smote-XGBoost dengan pemilihan fitur untuk mencegah model overfitting, penggunaan Smote-Enn dalam memproses data yang tidak seimbang dengan data diperoleh dari data pasien rumah sakit secara langsung yang memiliki total 4232 sampel dengan 37 fitur termasuk numerik dan kategorik, terakhir penelitian ini menguji algoritma XGBoost yang dibandingkan dengan algoritma lainnya Random Forest, K-Nearest Neighbor, Logistic Regression, Decision Tree, dan Naïve Bayes. Hasil dari penelitian ini menjadikan cara yang diusulkan yakni Smote-XGBoost mendapatkan hasil paling tinggi diantara algoritma lainnya dengan nilai *accuracy* 93.4%, *precision* 92.6%, *recall* 98.1%, dan *F1-Score* 94.8% [14].

Penelitian lain yang dilakukan oleh Mengyin Lin dkk, tahun 2023 mengangkat judul penelitian "Detection of Ionospheric Scintillation Based on XGBoost Model Improved by SMOTE-ENN Technique", dilakukan kombinasi dengan mengimprovisasi algoritma *Extreme Gradient Boosting* (XGBoost) dengan menggunakan teknik *Synthetic Minority Oversampling Technique and Edited Nearest Neighbor* (SMOTE-ENN) untuk mendeteksi kejadian yang tidak seimbang terkait sintilasi ionosfer lemah, sedang, dan kuat. Hasil pengujian untuk metode yang ditingkatkan menunjukkan peningkatan yang signifikan dalam indikator evaluasi, dengan nilai recall relatif stabil di atas 90%, dan F1 skor lebih dari 92% [15].

Berdasarkan kajian latar belakang dan tinjauan pustaka yang telah dilakukan, penelitian ini mengambil judul "Optimasi Model Klasifikasi Penyakit Jantung Menggunakan *Extreme Gradient Boosting* dengan *Hyperparameter Tuning GridSearchCV* dan *Balancing Data SMOTE-ENN*". Data yang digunakan pada penelitian ini adalah data sekunder berasal dari situs Kaggle dengan total data bejumlah 445.132 data. Adapun tahap – tahap pada penelitian ini meliputi *preprocessing* data, selanjutnya dilakukan teknik *hybridsampling Synthetic Minority Oversampling Technique - Edited Nearest Neighbor* (SMOTE-ENN) yang merupakan kombinasi dari teknik *oversampling* dengan *Synthetic Minority Oversampling Technique* (SMOTE) dan teknik *undersampling* dengan *Edited Nearest Neighbor* (ENN), uji skenario dibandingkan dengan tanpa teknik *resampling* apapun pada model

algoritma *Extreme Gradient Boosting* (XGBOOST) yang selanjutnya dilakukan *hyperparameter tuning* GridSearchCV untuk mendapatkan parameter terbaik dalam pengujian model klasifikasi tersebut. Setelah itu diakhiri dengan perhitungan metrik evaluasi menggunakan *accuracy*, *sensitivity, specifity,* dan *g-mean* untuk mengukur kinerja model terhadap klasifikasi data. Penelitian ini bertujuan untuk menganalisis sejauh mana penggunaan teknik *resampling* untuk *balancing* data dan *hyperparameter tuning* dapat meningkatkan perfoma model XGBoost dalam mengatasi ketidakseimbangan pada data kelas penyakit jantung.

### 1.2. Rumusan Masalah

Bedasarkan latar belakang yang telah didapat sebelumnya, penelitian ini menguraikan beberapa rumusan masalah berikut:

1. Bagaimana performa algoritma XGBoost dengan *hyperparameter tuning* sebagai model untuk mengatasi ketidakseimbangan data pada tiap kelas tanpa menggunakan teknik *resampling* pada dataset penyakit jantung?
2. Apakah kombinasi teknik *resampling* SMOTE-ENN untuk mengatasi ketidakseimbangan pada dataset penyakit jantung dapat meningkatkan performa dari model klasifikasi XGBoost dengan *hyperparameter tuning*?
3. Konsekuensi apa yang perlu untuk dipertimbangkan terkait penerapan teknik *resampling* SMOTE-ENN dalam model XGBoost dengan *hyperparameter tuning* yang digunakan untuk mengatasi rasio ketidakseimbangan kelas dalam klasifikasi pada data penyakit jantung?

### 1.3. Tujuan Penelitian

Berdasarkan perumusan masalah yang dijelaskan sebelumnya, tujuan akhir pada penelitian ini adalah sebagai berikut:

1. Menguji performa algoritma XGBoost dengan *hyperparameter tuning* sebagai model untuk mengatasi ketidakseimbangan tanpa menggunakan teknik *resampling* pada data penyakit jantung.
2. Mengevaluasi apakah kombinasi dari teknik *resampling* SMOTE-ENN pada dataset penyakit jantung dapat meningkatkan performa dari model klasifikasi XGBoost dengan *hyperparameter tuning*.

3. Menganalisis apakah terdapat konsekuensi yang perlu dipertimbangkan terkait penerapan teknik *resampling* dengan SMOTE-ENN dalam model XGBoost dengan *hyperparameter tuning* untuk mengatasi rasio ketidakseimbangan kelas dalam klasifikasi pada data penyakit jantung.

## 1.4. Batasan Masalah

Penelitian ini memiliki beberapa keterbarasan masalah meliputi:

1. Dataset yang digunakan pada penelitian ini adalah dataset sekunder dengan format CSV yang memiliki data cukup besar, yakni berjumlah 445.132 data.
2. Algoritma yang digunakan klasifikasi penyakit jantung pada penelitian ini adalah *Extreme Gradient Boosting* (XGBoost) dengan *hyperparameter tuning* GridSearchCV.
3. Penelitian ini menggunakan teknik SMOTE-ENN dengan kombinasi antara teknik *oversampling* SMOTE dan *undersampling* ENN, dalam mengatasi kelas yang tidak seimbang.
4. Pengukuran kinerja model menggunakan matriks evaluasi dalam mengukur seberapa baik performa model yang terdiri dari *accuracy* untuk mengukur prediksi benar, *sensitivity* sebagai *recall* kelas positif, *specifity* sebagai *recall* kelas negatif, dan *g-mean* untuk mengevaluasi dataset tidak seimbang.

## 1.5. Manfaat Penelitian

Penelitian yang dilakukan diharapkan dapat menghasilkan beberapa manfaat, yaitu:

1. Hasil dari uji coba dengan mengkombinasikan antara XGBoost dengan *hyperparameter tuning* dan *balancing data* pada penelitian ini diharapkan mampu menambahkan wawasan baru terkait seberapa efektif pendekatan yang telah digunakan dalam meningkatkan akurasi pada data yang memiliki kelas yang tidak seimbang.
2. Penelitian ini juga diharapkan memberikan referensi untuk peneliti maupun pengembang selanjutnya dengan bidang terkait. Dengan tujuan dan pemilihan metode yang baik, penelitian ini dapat menjadi acuan untuk penggunaan klasifikasi terhadap penyakit jantung.

*Halaman ini sengaja dikosongkan*