BAB I
PENDAHULUAN

1.1 Latar Belakang

Penyakit jantung merupakan istilah umum untuk seseorang yang memiliki gangguan pada fungsi kerja jantung. Berdasarkan data *World Health Organization* (WHO) tahun 2017, angka kematian akibat penyakit jantung mencapai 17,7 juta jiwa setiap tahunnya, dan penyakit jantung menyumbang 31% dari total kematian di seluruh dunia. Angka kematian ini diprediksi akan mengalami peningkatan setiap tahunnya dan diperkirakan mencapai 23,3 juta jiwa pada tahun 2030 (Medyati dkk., 2018). Di Indonesia, penyakit jantung juga menjadi salah satu penyebab kematian tertinggi. Laporan Pengelolaan Program dan Laporan Keuangan Badan Penyelenggara Jaminan Sosial (BPJS) Kesehatan tahun 2022 menunjukkan bahwa penyakit jantung menghabiskan biaya sebesar Rp 12,144 triliun dengan total 15 juta pasien. Angka ini cenderung mengalami peningkatan dibandingkan tahun sebelumnya dengan biaya pengeluran sebesar Rp 8,671 triliun (Khadijah Nur Azizah, 2023). Oleh karena itu, deteksi dini penyakit jantung sangat diperlukan untuk menurunkan jumlah angka kematian di dunia dan Indonesia.

Pada dasarnya, banyak hal yang dapat dilakukan untuk mencegah penyakit jantung, salah satunya adalah menerapkan gaya hidup sehat. Selain itu, deteksi dini penyakit jantung sangat penting untuk mencegah kematian pada penderitanya. Salah satu cara untuk melakukan deteksi dini adalah dengan memanfaatkan teknologi informasi. *Data mining* dapat diartikan sebagai gabungan dari berbagai cabang ilmu pengetahuan, seperti basis data, statisika, *machine learning*, visualisasi, dan pengetahuan informasi. Data mining telah berhasil diterapkan dalam berbagai bidang ilmu, seperti ekonomi, bioinformatika, genetika, kedokteran, Pendidikan, dan sebagainya. *Data mining* terbagi menjadi beberapa teknik, di antaranya klastering, asosiasi, prediksi, dan klasifikasi (Pristyanto, 2019). Pada permasalahan penyakit jantung, teknik yang digunakan adalah klasifikasi.

Terdapat beberapa algoritma klasifikasi yang umum digunakan untuk memprediksi penyakit jantung seperti, *Support Vector Machine* (SVM), *K-Nearest Neighbor*, *Neural Netwrok*, *Decision Tree*, *Logistic Regression*, *AdaBoost*, *Naïve*

Bayes, dan *Fuzzy Logic* (Haq dkk., 2018). Pada penelitian ini akan menggunakan algoritma *Extreme Gradient Boosting* (XGBoost). XGboost merupakan salah satu algoritma *ensemble learning* yang memiliki kinerja yang sangat baik dalam klasifikasi maupun regresi. keunggulan utama pada algoritma ini terletak pada kemampuan dalam mencegah *overfitting* dan hasil prediksi yang relatif tinggi terhadap data yang hilang ataupun data yang tidak seimbang (Ayu dkk., 2023).

Pada kasus nyata, permasalahan umum yang sering ditemukan dalam data medis adalah *imbalance data*. Kondisi ini terjadi ketika terdapat ketidakseimbangan jumlah antara kelas mayoritas dan minoritas. Akibatnya, model klasifikasi cenderung memprediksi kelas mayoritas, sehingga dapat menghasilkan hasil klasifikasi yang salah. Hal ini berpotensi menyebabkan kesalahan dalam penanganan pasien (Mutmainah, 2021).

Terdapat dua pendekatan untuk mengatasi ketidakseimbangan kelas, yaitu pendekatan pada level algoritmia dan pendekatan pada level data. Pendekatan level algoritma berfokus pada perbaikan algoritma klasifikasi tanpa mengubah disribusi kelas pada data. Sementara itu, pendekatan level data berfokus pada penggunaan berbagai macam teknik *resampling* untuk membuat distribusi kelas menjadi seimbang. Teknik *resampling* sendiri dibagi menjadi tiga jenis, yaitu *oversampling*, *undersampling*, dan *hybrid* (Alberth dkk., 2022).

Berdasarkan penelitian yang dilakukan oleh Agili Lopo dan Dwi Hartomo pada tahun 2023 dengan judul penelitian "Evaluating Sampling Techniques for Healthcare Insurance Fraud Detection in Imbalanced Dataset", dilakukan perbandingan berbagai teknik *resampling* pada model XGBoost yang telah di optimasi dengan *hyperparameter tuning*. Algoritma *resampling* yang digunakan adalah SMOTE, ROS, RUS, dan IHT. Selain itu, penelitian ini juga mencari berbagai rasio sampling untuk setiap teknik *resampling*. Hasil penelitian menunjukkan bahwa algortima SMOTE dengan rasio 70:30 memberikan performa terbaik, dengan nilai AUC sebesar 0,76 dan *g-mean* sebesar 0,74 (Agili Lopo & Dwi Hartomo, 2023).

Berdasarkan latar belakang dan kajian literatur sebelumnya, penelitian ini mengangkat judul "Analisis Perbandingan Teknik Oversampling pada Extreme Gradient Boosting (XGBoost) untuk Mengatasi Ketidakseimbangan Kelas pada

Klasifikasi Penyakit Jantung". Data penelitian yang digunakan merupakan data sekunder yang berasal dari situs Kaggle dengan jumlah total data sebanyak 445.132 data. Tahap awal penelitian meliputi *preprocessing* data. Selanjutnya, dilakukan teknik *oversampling* menggunakan beberapa algoritma, seperti *Random Oversampling* (ROS), *Synthetic Minority Oversampling Technique* (SMOTE), dan *Adaptive Synthetic* (ADASYN). Kemudian, melakukan pelatihan model menggunakan algoritma XGBoost. Beberapa metrik evaluasi seperti akurasi, sensitivitas, spesifisitas, dan *g-mean* akan digunakan dalam penelitian untuk mengukur kinerja model. Tujuan penelitian ini adalah untuk mencari tahu apakah penggunaan teknik *oversampling* dapat meningkatkan performa model XGBoost dalam mengatasi ketidakseimbangan kelas pada klasifikasi penyakit jantung.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, penelitian ini merumuskan beberapa pertanyaan sebagai berikut:

1. Bagaimana kinerja algoritma XGBoost dalam mengatasi ketidakseimbangan kelas tanpa menggunakan teknik *oversampling* pada dataset penyakit jantung?
2. Apakah kombinasi antara Extreme Gradient Boosting (XGBoost) dengan teknik *oversampling* dapat meningkatkan kinerja model klasifikasi dalam mengatasi ketidakseimbangan kelas pada dataset penyakit jantung?
3. Apakah terdapat trade-off yang perlu dipertimbangkan dalam menerapkan teknik *oversampling* pada model XGBoost untuk mengatasi ketidakseimbangan kelas pada klasifikasi penyakit jantung?

1.3 Tujuan Penelitian

Berdasarkan permasalahan yang telah diuraikan, maka tujuan yang ingin dicapai pada penelitian ini adalah sebagai berikut:

1. Menguji kinerja algoritma *Extreme Gradient Boosting* (XGBoost) dalam mengatasi ketidakseimbangan kelas pada klasifikasi penyakit jantung tanpa menerapkan teknik *oversampling*.
2. Menilai apakah kombinasi antara teknik *oversampling* dan XGBoost dapat menghasilkan model klasifikasi yang lebih unggul dalam mengatasi

ketidakseimbangan kelas pada dataset penyakit jantung dibandingkan dengan penggunaan XGBoost tanpa *oversampling*.

3. Menganalisis *trade-off* yang mungkin terjadi dalam menerapkan teknik *oversampling* dan XGBoost untuk mengatasi ketidakseimbangan kelas pada klasifikasi penyakit jantung.

1.4 Batasan Masalah

Batasan masalah dalam penelitian ini adalah sebagai berikut:

1. Penelitian ini menggunakan dataset yang diperoleh dari Kaggle dalam format CSV dengan jumlah data yang cukup besar, yaitu 445.132 data.
2. Penelitian ini menggunakan algoritma Extreme Gradient Boosting (XGBoost) sebagai metode klasifikasi untuk memprediksi penyakit jantung.
3. Penelitian ini menggunakan beberapa teknik *oversampling*, yaitu ROS, SMOTE, dan ADASYN, untuk mengatasi ketidakseimbangan kelas.
4. Kinerja model dievaluasi menggunakan beberapa metrik, yaitu akurasi, sensitivitas, spesifisitas, dan g-mean untuk mengukur seberapa baik model dapat mengklasifikasikan kelas positif dan negatif.

1.5 Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan beberapa manfaat, yaitu:

1. Dengan mengevaluasi kombinasi antara XGBoost dan teknik *oversampling*, penelitian ini diharapkan dapat memberikan wawasan tentang efektivitas pendekatan tersebut dalam meningkatkan akurasi prediksi yang memiliki ketidakseimbangan kelas.
2. Penelitian ini diharapkan dapat menjadi referensi bagi para peneliti, praktisi, atau pengembang yang berkerja pada bidang terkait. Dengan memperhatikan tujuan bisnis, distribusi kelas, dan memilih metode yang terbaik, diharapkan dapat membantu pengambilan keputusan terhadap klasifikasi penyakit jantung