BAB I
PENDAHULUAN

1.1 Latar Belakang

Ketidakseimbangan kelas pada dataset menjadi tantangan umum dalam penerapan algoritma klasifikasi machine learning. Pada kasus nyata seperti diagnosis penyakit, deteksi penipuan, dan prediksi kerusakan mesin, jumlah data pada satu kelas (biasanya kelas mayoritas) jauh lebih dominan dibandingkan kelas lainnya. Ketidakseimbangan ini menyebabkan model cenderung bias terhadap kelas mayoritas dan gagal mendeteksi kelas minoritas secara akurat, padahal kelas minoritas sering kali lebih penting untuk diprediksi

Dua teknik yang umum digunakan untuk mengatasi masalah ini adalah SMOTE (Synthetic Minority Oversampling Technique) dan NearMiss. SMOTE merupakan metode oversampling yang menghasilkan sampel sintetis untuk kelas minoritas berdasarkan interpolasi [17], sementara NearMiss adalah metode undersampling yang memilih sampel dari kelas mayoritas berdasarkan jarak terdekat terhadap data minoritas [19]. Beberapa penelitian menunjukkan bahwa efektivitas kedua metode tersebut sangat tergantung pada karakteristik dataset yang digunakan [8][9].

Random Forest menjadi algoritma yang populer untuk tugas klasifikasi karena kemampuannya menangani data berskala besar dan mengurangi overfitting. Dalam penelitian Nugroho & Harini, kombinasi Random Forest dan SMOTE mampu meningkatkan akurasi hingga 97,5% dalam klasifikasi diabetes [6]. Di sisi lain, penelitian Alamsyah et al. justru menemukan bahwa NearMiss menghasilkan nilai F1-score dan akurasi yang lebih tinggi dalam kasus klasifikasi penyakit [7], menunjukkan bahwa masing-masing metode balancing memiliki keunggulan tergantung pada konteks dataset yang digunakan.

Berdasarkan permasalahan tersebut, penelitian ini bertujuan untuk membandingkan kinerja metode SMOTE dan NearMiss dalam menangani dataset tidak seimbang menggunakan algoritma Random Forest. Untuk itu, penulis mengembangkan aplikasi klasifikasi universal berbasis Streamlit yang memungkinkan pengguna memuat berbagai dataset klasifikasi, menerapkan balancing, melatih model, mengevaluasi performa, dan melakukan prediksi secara interaktif, tanpa perlu mengubah kode secara manual.

1.2 Rumusan Masalah

Berdasar latar belakang yang telah dijelaskan, penelitian ini berfokus pada tantangan dalam menangani ketidakseimbangan data pada berbagai dataset menggunakan metode SMOTE dan NearMiss. Oleh karena itu, rumusan masalah yang diajukan dalam penelitian ini adalah:

1. Bagaimana pengaruh ketidakseimbangan kelas terhadap performa model klasifikasi Random Forest pada berbagai dataset dari domain kesehatan, keuangan, dan industri?
2. Bagaimana perbandingan efektivitas metode SMOTE dan NearMiss dalam menangani ketidakseimbangan data berdasarkan metrik *precision, recall, dan F1-score*?

1.3 Batasan Masalah

Penelitian ini memiliki beberapa batasan ruang lingkup guna menjaga fokus analisis dan kejelasan tujuan.

1. Pada penelitian ini hanya menggunakan dataset dengan karakteristik ketidakseimbangan yang signifikan, dataset yang digunkan meliputi, *Predictive Maintencance, Diabetes, Spam SMS, Customer Churn, Breast Cancer, GiveMeSomeCredit, Spam Email, dan Employee Performance.*
2. Penelitian ini hanya berfokus membandingkan dua metode penanganan ketidakseimbangan data, yaitu SMOTE (*Synthetic Minority Oversampling Technique)* dan *NearMiss* Sebagai Teknik *undersampling*. Dan pada penelitian ini metode penanganan ketidakseimbangan lainnya seperti ADASYN, Borderline-SMOTE, atau Teknik hybrid tidak digunakan.

3. Model yang digunkan dalam penelitian ini adalah *Random Forest*. Tidak dilakukan pembandingan model dengan model Machine learning lainnya seperti SVM, Decision Tree, atau Deep Learning.
4. Kinerja model dievaluasi menggunakan precision, recall, accuracy F1-Score, ROC AUC Score dan Confusion Matrix
5. Tidak dilakukan analisis terhadap performa model dalam lingkungan produksi atau sistem real-time

1.4 Tujuan Penelitian

Tujuan penelitian ini adalah untuk menganalisis kinerja metode SMOTE dan NearMisss dalam menangani ketidakseimbangan kelas pada berbagai dataset menggunakan algoritma *Random Forest*. Evaluasi dilakukan dengan metrik *precision, recall, accuracy F1-Score, ROC AUC Score* dan *Confusion Matrix* untuk menentukan metode yang lebih optimal sesuai dengan karakteristik dataset. Hasil penelitian ini diharapkan dapat memberikan panduan dalam pemilihan metode ketidakseimbangan data agar dapat meningkatkan akurasi dan keadilan model klasifikasi di berbagai dataset.

1.5 Manfaat Penelitian

Penelitian ini diharapkan dapat menjadi panduan bagi peneliti dan praktisi dalam memilih metode yang sesuai untuk menangani ketidakseimbangan data guna meningkatkan akurasi model. Secara luas penelitian ini dapat berkontribusi pada pengembangan sistem berbasis *machine learning* yang lebih adil dan andal dalam berbagai bidang, seperti Kesehatan, keuangan, dan industri.

1.6 Sistematika Penulisan

Untuk mempermudah dalam memahami skripsi ini, maka penulis materi disusun dengan sistematika sebagai berikut:

Contoh :

BAB I PENDAHULUAN, berisi Latar belakang masalah, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian dan sistematika penulisan.

BAB II TINJAUAN PUSTAKA, berisi tinjauan pustaka, dasar-dasar teori yang

digunakan dan dijadikan dasar penelitian dalam skripsi ini.

BAB III METODE PENELITIAN, didalamnya terdapat tinjauan umum tentang objek penelitian, analisis masalah, tahap rancangan, serta alat dan bahan yang digunakan dalam penelitian

BAB IV HASIL DAN PEMBAHASAN, bab ini merupakan tahapan yang penulis lakukan dalam hasil penelitian yang dicapai, serta menjelaskan hasil uji coba rancangan yang telah dibuat

BAB V PENUTUP, berisi kesimpulan dan saran yang dapat peneliti rangkum selama proses penelitian yang penulis berikan untuk peneli ti selanjutnya.