

BAB I

PENDAHULUAN

1.1 Latar Belakang

Machine learning telah mengalami kemajuan yang signifikan dalam bidang medis, terutama dalam mendukung proses diagnostik. Salah satu studi yang sering dilakukan adalah klasifikasi penyakit jantung berdasarkan data pasien. Namun demikian [1], tantangan besar yang sering ditemui dalam analisis data medis adalah ketidakseimbangan distribusi kelas. Kondisi ini terjadi ketika kelas mayoritas memiliki jumlah sampel yang jauh lebih besar dibandingkan dengan kelas minoritas [2], [3]. Sebagai contoh, dalam dataset penyakit jantung, individu yang sehat biasanya jauh lebih banyak dibandingkan dengan mereka yang terdiagnosis menderita penyakit tersebut [4].

Data yang tidak seimbang dapat menyebabkan penurunan kinerja klasifikasi [5], terutama dalam mengidentifikasi kelas minoritas, yang justru sering kali lebih penting untuk dikenali secara akurat [6]. Model klasifikasi seperti K-Nearest Neighbors (KNN) dan Support Vector Machine (SVM) cenderung bias terhadap kelas mayoritas [7]. KNN menentukan keanggotaan kelas berdasarkan tetangga terdekat [8], sehingga menyulitkan data dari kelas minoritas untuk direpresentasikan dengan baik dalam prediksi karena jumlahnya yang terbatas [9]. Di sisi lain, SVM berusaha menemukan hyperplane pemisah antara kelas, namun pada data yang tidak seimbang, hyperplane yang dihasilkan cenderung terlalu dekat dengan kelas minoritas, menyebabkan banyak sampel minoritas salah diklasifikasikan [10].

Untuk mengatasi masalah ini, diperlukan teknik penyeimbangan data sebelum proses klasifikasi dilakukan. Salah satu teknik yang umum digunakan adalah SMOTE (Synthetic Minority Over-sampling Technique) [11]. Metode ini menghasilkan data sintesis baru dari kelas minoritas untuk mencapai distribusi kelas yang lebih seimbang. Namun, SMOTE memiliki beberapa kelemahan, termasuk kecenderungan menciptakan instance yang berada pada area bising atau borderline [12]. Oleh karena itu, beberapa pengembangan dari SMOTE telah diusulkan, seperti SMOTE yang dikombinasikan dengan Tomek Links dan SMOTE yang diintegrasikan dengan ENN (Edited Nearest Neighbor). SMOTE-Tomek bekerja dengan menghapus sampel yang saling tumpang tindih dari kelas mayoritas dan minoritas yang letaknya terlalu berdekatan (kasus borderline). Sementara itu, SMOTE-ENN meningkatkan kualitas dataset dengan menghapus instance yang salah diklasifikasikan berdasarkan tetangga terdekatnya (k-nearest neighbors), sehingga dapat mengurangi noise dan meningkatkan kualitas data [11].

Penelitian ini bertujuan untuk meningkatkan kinerja algoritma KNN dan SVM pada data yang tidak seimbang dengan menerapkan metode SMOTE, SMOTE-Tomek, dan SMOTE-ENN. Efektivitas model diukur menggunakan metrik seperti precision, recall, F1-score [12], serta akurasi, untuk mengetahui sejauh mana peningkatan performa klasifikasi setelah data diseimbangkan. Dengan pendekatan ini, diharapkan model dapat mengenali data dari kelas minoritas secara lebih efektif tanpa mengorbankan performa keseluruhan [13].

1.2 Rumusan Masalah

Berdasarkan latar belakang di atas, rumusan masalah dalam penelitian ini adalah:

1. Bagaimana pengaruh teknik SMOTE, SMOTE-Tomek, dan SMOTE-ENN terhadap keseimbangan distribusi kelas pada dataset kesehatan?
2. Sejauh mana peningkatan performa klasifikasi K-NN dan SVM setelah penerapan metode tersebut?
3. Bagaimana perbandingan akurasi, presisi, recall, dan F1-score sebelum dan sesudah penerapan teknik oversampling?

1.3 Tujuan dan Manfaat Penelitian

Tujuan Penelitian:

- Menerapkan teknik SMOTE dan hybrid SMOTE untuk menyeimbangkan distribusi kelas pada dataset kesehatan.
- Menganalisis dampak metode tersebut terhadap performa klasifikasi K-NN dan SVM.
- Membandingkan metode-metode tersebut dengan teknik oversampling lainnya berdasarkan metrik akurasi, presisi, recall, dan F1-score.

Manfaat Penelitian:

- Memberikan solusi baru dalam penanganan data tidak seimbang di bidang kesehatan.
- Meningkatkan akurasi model klasifikasi dalam mendeteksi pasien dengan kondisi kesehatan yang kritis.
- Menawarkan rekomendasi metode terbaik untuk menangani ketidakseimbangan data pada penelitian medis dan bidang lainnya.

1.4 Batasan Masalah

Untuk menjaga fokus dan kesesuaian penelitian, maka batasan masalah dalam penelitian ini adalah:

- Penelitian ini hanya berfokus pada dua algoritma machine learning: K-NN dan SVM.
- Dataset yang digunakan merupakan dataset klasifikasi biner yang tidak seimbang.
- Metode oversampling yang dibandingkan hanya terbatas pada SMOTE dan Hybrid SMOTE.
- Evaluasi performa model hanya berdasarkan metrik akurasi, presisi, recall, dan F1-score.
- Penelitian ini tidak membahas dataset klasifikasi multi-kelas.

1.5 Kebaruan Penelitian (Novelty)

Berikut adalah kebaruan penelitian ini berdasarkan studi literatur terhadap penelitian terdahulu terkait metode SVM, KNN, dan SMOTE:

Penelitian oleh Mohamed Hanafy Kotb dan Ruixing Ming (2021) berjudul *"Comparing SMOTE Family Techniques in Predicting Insurance Premium Defaulting using Machine Learning Models"* membandingkan teknik-teknik SMOTE untuk menangani ketidakseimbangan data pada sektor asuransi. SMOTE-Tomek menghasilkan performa terbaik ketika digunakan bersama SVM.

Penelitian oleh Arjun Puri dan Manoj Kumar Gupta (2022) menggunakan model hybrid ensemble yang menggabungkan K-Means SMOTE untuk oversampling dan ENN untuk undersampling. Pendekatan ini menunjukkan hasil yang lebih baik, terutama saat tingkat noise dalam dataset tinggi.

Adli Abdillah Nababan (2023) dalam penelitiannya menyatakan bahwa metode KNN tanpa oversampling menghasilkan akurasi yang lebih rendah dibandingkan setelah diterapkan LHS, yang menunjukkan pentingnya teknik pra-pemrosesan.

Nayema Khatun et al. (2024) menggunakan SMOTE, SMOTE-ENN, dan SMOTE-Tomek dalam prediksi diabetes tipe-2. Kombinasi teknik ini berhasil meningkatkan akurasi hingga 99.9%, menunjukkan potensi besar metode oversampling dalam meningkatkan deteksi awal penyakit.