

BAB I

PENDAHULUAN

1.1 Latar Belakang

Klasifikasi adalah salah satu teknik dalam data mining yang berguna untuk menentukan kategori atau kelompok suatu data. Model klasifikasi berfungsi untuk memprediksi suatu data berdasarkan pola dari data yang sudah ada. Hasil dari model ini telah diuji pada berbagai jenis data dan terbukti efektif. Tujuan utama klasifikasi yaitu memprediksi nilai dari data yang belum diketahui, dengan memanfaatkan pola dan informasi yang diperoleh dari data yang sudah ada sebelumnya, sehingga dapat membantu dalam pengambilan keputusan awal, klasifikasi tidak hanya berguna untuk mengelompokkan data tetapi juga membantu dalam mengambil keputusan yang lebih baik berdasarkan analisis data yang mendalam[1].

Proses klasifikasi penting dalam sebuah penelitian mulai dari pengolahan data akademik, kesehatan, hingga industri. Klasifikasi dapat dilaksanakan dengan menggunakan berbagai algoritma, seperti *Support Vector Machine* (SVM), *Naïve Bayes*, *Decision Tree*, dan *K-Nearest Neighbor* (KNN), yang masing-masingnya mempunyai keunggulan dan kelemahannya tersendiri. Dalam penerapan klasifikasi dibantu oleh berbagai algoritma. Salah satu metode yang dipakai yaitu algoritma *Support Vector Machine* (SVM)[2].

Support Vector Machine (SVM) adalah metode yang seringkali digunakan untuk klasifikasi pada pembelajaran mesin. SVM beroperasi dengan mencari garis pemisah terbaik (hyperplane) yang dapat dibagi menjadi dua kelompok data dengan jarak terbesar. Keunggulan utama SVM adalah mampu menangani data yang memiliki banyak fitur, dalam jumlah yang besar dan tetap memberikan hasil yang akurat meskipun jumlah data pelatihan terbatas[3]. Namun, SVM memiliki beberapa kekurangan. Salah satunya adalah sensitivitas terhadap pemilihan parameter kernel yang optimal menentukan kombinasi parameter yang optimal cukup sulit karena proses pencariannya kompleks, cara untuk mengatasi kekurangan tersebut yaitu dengan menggunakan pencarian parameter otomatis seperti *Grid Search*[4]. Selain itu SVM mengalami kendala saat menghadapi data

yang tidak seimbang. Ketidakseimbangan ini bisa membuat hasil klasifikasi kurang akurat karena model cenderung lebih fokus pada kelas yang jumlah datanya lebih banyak (kelas mayoritas) dan mengabaikan kelas yang datanya lebih sedikit (kelas minoritas). Cara untuk mengatasi masalah ini adalah dengan menggunakan metode oversampling, seperti *Synthetic Minority Over-sampling Technique* (SMOTE)[5].

Penyakit jantung adalah salah satu faktor utama yang menyebabkan tingginya angka kematian secara global, penyakit jantung menyumbang penyakit yang diderita banyak orang. Berdasarkan WHO sekitar 17,9 juta orang meninggal setiap tahunnya dikarenakan penyakit jantung atau kardiovaskular[6]. Jantung berperan krusial dalam sistem sirkulasi dengan tugas utama memompa darah yang kaya oksigen ke seluruh bagian tubuh. Sehingga setiap gangguan pada organ jantung dapat berdampak serius pada kesehatan secara keseluruhan[7]. Penyakit jantung muncul akibat gangguan pada berbagai bagian jantung, termasuk pembuluh darah, selaput jantung, katup, atau otot jantung. Penyakit ini disebabkan oleh beberapa faktor seperti kolesterol tinggi, diabetes, tekanan darah tinggi, merokok, begadang, dan gaya hidup yang tidak sehat berkontribusi terhadap peningkatan risiko penyakit jantung[8]. Penyakit jantung salah satu penyakit tidak menular (PTM) yang berisiko menyebabkan kematian. Penyakit ini tidak disebabkan oleh infeksi kuman dan sering kali tidak menimbulkan gejala pada tahap awal, sehingga banyak orang sulit mendeteksinya sejak dini. Mendeteksi penyakit jantung sejak dini sangat penting untuk meningkatkan peluang kesembuhan dan mengurangi dampak penyakit[9]. Penelitian sebelumnya tentang klasifikasi gagal jantung telah diterapkan metode yang digunakan adalah metode SVM di mana hasil uji coba menunjukkan bahwa akurasi paling tinggi diperoleh pada kernel linear dengan akurasi 85,96%, RBF akurasi 85,84%, dan polynomial dengan akurasi 84,50%[2].

Penelitian ini mengusulkan metode SVM dan teknik SMOTE dalam mengatasi data tidak seimbang serta menerapkan optimasi *hyperparameter* dengan *grid search* pada klasifikasi penyakit jantung. Harapannya, penelitian ini dapat memberikan hasil akurasi yang optimal dalam klasifikasi data penyakit jantung.

1.2 Rumusan Masalah

Berdasarkan penjelasan latar belakang yang telah disampaikan, dapat dirumuskan beberapa pernyataan masalah sebagai berikut.

1.2.1 Permasalahan

Dalam klasifikasi penyakit jantung terdapat ketidakseimbangan data dimana jumlah data lebih banyak kelas mayoritas dibandingkan kelas minoritas. Selain itu *Support Vector Machine* (SVM) memiliki kelemahan dalam sensitivitas terhadap pemilihan parameter kernel dalam menentukan kombinasi parameter yang optimal.

1.2.2 Solusi Permasalahan

Dalam penelitian ini mengusulkan penerapan metode SVM pada klasifikasi penyakit jantung. Penelitian ini menerapkan teknik SMOTE sebagai penyeimbang jumlah kelas dalam dataset. Dalam memilih parameter dapat dilakukan dengan pencarian parameter otomatis dengan *grid search*.

1.2.3 Pertanyaan Penelitian

1. Berapa nilai akurasi dengan menggunakan metode SVM dengan *Grid Search* dalam klasifikasi pada data penyakit jantung?
2. Bagaimana hasil perbandingan kinerja metode SVM dengan penerapan SMOTE dan tanpa menggunakan SMOTE dalam melakukan klasifikasi penyakit jantung?

1.3 Tujuan dan Manfaat

1.3.1 Tujuan

1. Untuk mengetahui berapa hasil akurasi SVM dengan *Grid Search* untuk klasifikasi penyakit jantung.
2. Untuk mengetahui seberapa efektif metode SMOTE dalam menyeimbangkan jumlah kelas dataset penyakit jantung.

1.3.2 Manfaat

Dapat mengetahui berapa besar akurasi klasifikasi penyakit jantung yang menerapkan metode SVM dan teknik SMOTE yang pada akhirnya dapat digunakan sebagai rujukan atau perbandingan dengan metode lainnya.

1.4 Batasan Masalah

1. Dataset yang digunakan pada penelitian ini sebanyak 4.238 terdiri dari 16 fitur dan 1 label berisi 2 kelas yaitu “No” untuk kelas tidak terdeteksi penyakit jantung dengan jumlah 3594 dan “Yes” untuk kelas terdeteksi penyakit jantung dengan jumlah 644.

2. Data yang digunakan pada penelitian ini bersumber dari website kaggle pada link berikut: <https://www.kaggle.com/datasets/mirzahasnine/heart-disease-dataset> dari *American Heart Association National Heart, Lung, and Blood Institute*
3. Fitur-fitur yang digunakan untuk klasifikasi penyakit jantung yaitu *Gender, Age, Education, Currentsmoker, Cigsperday, Bpmeds, Prevalentstroke, Prevalenthyp, Diabetes, Totchol, Sysbp, Diabp, Bmi, Heartrate, Glucose, Heart_stroke*.
4. Metode yang digunakan untuk tahap klasifikasi penyakit jantung adalah SVM dengan kernel Linear dan kernel RBF, teknik SMOTE untuk mengatasi data tidak seimbang dan optimasi hyperparamter *Grid search*.
5. Evaluasi kinerja yang digunakan menggunakan confusion matrix yaitu *accuracy, precision, recall, f1-score*, dan AUC.

1.5 Sistematika Penulisan

Sistematika penulisan laporan ini disusun dengan tujuan untuk memberikan gambaran secara menyeluruh mengenai isi laporan

BAB I PENDAHULUAN

Bab ini membahas latar belakang masalah. Ini mencakup rumusan masalah, latar belakang, tujuan dan manfaat, batasan-batasan, dan sistematika laporan.

BAB II KAJIAN PUSTAKA

Berisi ide-ide utama yang diperlukan untuk menyelesaikan penelitian, serta penelitian sebelumnya yang telah dilakukan menggunakan pendekatan atau masalah yang sama untuk menyelesaikan penelitian.

BAB III METODE USULAN

Bagian ini memberikan penjelasan mengenai arsitektur sistem, *preprocessing* data, tahap-tahap atau sistematika dari setiap metode yang diajukan serta skenario pengujian.

BAB IV HASIL DAN PEMBAHASAN

Bab ini berisi hasil dan pembahasan tentang implementasi sistem dengan menggunakan metode *Support Vector Machine*, Teknik SMOTE, Mencari parameter terbaik dengan *Grid Search*. Hasil ini juga mencakup lingkungan serta hasil uji coba.

BAB V PENUTUP

Bagian ini mencakup kesimpulan dari penelitian, yang disesuaikan dengan rumusan yang telah dibuat, dan saran untuk penelitian selanjutnya.