

BAB I PENDAHULUAN

1.1 Latar Belakang

Diabetes melitus atau yang lebih dikenal sebagai diabetes merupakan sebuah kondisi metabolisme yang serius dan memiliki dampak negatif yang signifikan terhadap kesehatan manusia [1]. Secara global, diperkirakan terdapat sekitar 422 juta orang dewasa yang hidup dengan diabetes, dan dari jumlah tersebut sekitar 1,5 juta orang meninggal dunia pada tahun 2014. Di Indonesia, diabetes menempati posisi sebagai penyebab kematian tertinggi ke 3, dengan persentase sebesar 6,7% setelah stroke yang mencatat 21,3% dan penyakit jantung koroner yang mencapai 12,9% pada tahun 2016, angka ini terus meningkat setiap tahunnya [2]. Hal ini menjadikan diabetes sebagai salah satu dari sepuluh penyakit yang paling umum menyebabkan kematian di seluruh dunia. Dengan demikian, sangat penting untuk mengembangkan metode diagnostik yang efektif dan efisien dalam mendeteksi penyakit diabetes secara cepat dan akurat.

Salah satu metode untuk mendeteksi penyakit diabetes adalah dengan menggunakan algoritma machine learning. Meskipun demikian, permasalahan yang sering terjadi adalah ketersediaan kualitas data dalam mendeteksi penyakit diabetes yang mempengaruhi kinerja machine learning. Secara umum, permasalahannya terletak pada ketidakseimbangan data yang menyebabkan kinerja machine learning menghasilkan tingkat akurasi prediksi yang rendah pada kelas minoritas dan tinggi pada kelas mayoritas [3]. Ketidakseimbangan kelas dalam dataset atau kesalahan dalam penyetelan parameter yang sering kali menjadi tantangan dalam analisis data. Dalam penelitian ini, metode optimasi Hyperparameter Tuning yang akan digunakan adalah Grid Search. Metode ini melakukan uji coba kombinasi dan validasi satu per satu, kemudian memilih kombinasi yang memberikan performa model terbaik untuk prediksi. Grid Search Cross Validation secara otomatis akan melakukan validasi dari setiap model kombinasi dan hyperparameter [4].

Bidang deteksi penyakit diabetes telah banyak memanfaatkan algoritma machine learning untuk menganalisis data. Salah satu algoritma yang populer dan mudah digunakan adalah K-Nearest Neighbor (KNN). Dalam konteks klasifikasi diabetes, KNN dapat memprediksi kemungkinan diabetes dengan menganalisis data individu dan membandingkannya dengan data pasien lain yang menunjukkan gejala atau faktor risiko yang serupa [5]. Beberapa penelitian juga telah dilakukan untuk membandingkan efektivitas berbagai model klasifikasi yang populer. Dalam penelitian untuk mendeteksi penyakit diabetes dengan membandingkan tiga model machine learning, yaitu Neural Network, Naive Bayes dan Logistic Regression menunjukkan bahwa algoritma Logistic Regression lebih baik dibandingkan dua algoritma lainnya [6]. Adapun algoritma lain seperti Support Vector Machine (SVM), dengan tujuan membangun sebuah model yang dapat memprediksi data tes yang diberikan. Salah satu keuntungan menggunakan metode ini untuk mendeteksi penyakit diabetes menggunakan sistem yaitu waktu yang diperlukan singkat [7].

Meskipun sudah banyak penelitian yang telah dilakukan mengenai penerapan algoritma machine learning, sebagian besar dataset yang berkaitan dengan penyakit masih tidak seimbang. Oleh karena itu, perlu untuk mengatasi masalah ketidakseimbangan ini dengan salah satu metode yang bisa digunakan yaitu sampling. Dalam penelitian ini, metode SMOTE akan diterapkan untuk menangani ketidakseimbangan data dengan cara menciptakan sampel sintetis bagi kelas minoritas dan mengurangi overfitting. Namun metode SMOTE memiliki kelemahan, yaitu memilih instance kelas minoritas secara acak untuk di-oversample dengan probabilitas uniform, sehingga menyebabkan data menjadi noise karena tidak membedakan area kelas yang overlap [8].

Berdasarkan latar belakang masalah dari beberapa penelitian diatas, maka pada penelitian ini diusulkan menggunakan metode Hyperparameter Tuning dan SMOTE. Metode ini digunakan untuk identifikasi penyakit diabetes agar meningkatkan akurasi model machine learning dan mengurangi masalah ketidakseimbangan data sehingga memberikan informasi yang lebih akurat dan membantu dalam diagnosis dini yang lebih efektif. Sehingga pada penelitian ini

mengangkat judul yaitu “Perbandingan Performa Algoritma Machine Learning untuk Mendeteksi Penyakit Diabetes”.

1.2 Rumusan Masalah

Berdasarkan latar belakang diatas, dapat dirumuskan masalah sebagai berikut:

1. Apakah penggunaan metode SMOTE dalam mengatasi ketidakseimbangan data dapat meningkatkan kinerja klasifikasi machine learning dalam mendeteksi?
2. Bagaimana perbandingan performa algoritma machine learning yang memberikan hasil akurasi paling akurat dalam mendeteksi dan pengaruh penggunaan Stacking Classifier dalam meningkatkan akurasi model dibandingkan dengan algoritma individu?

1.3 Batasan Masalah

Untuk memastikan fokus yang jelas dan menghindari penyimpangan serta pelebaran pokok masalah dalam penyusunan penelitian ini, peneliti telah menetapkan batasan masalah sebagai berikut:

1. Penerapan data mining pada dataset publik yang tersedia di kaggle berfokus pada data medis penyakit diabetes. <https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database>
2. Metode balancing data yang digunakan dalam penelitian ini adalah metode SMOTE (Synthetic Minority Over-sampling Technique).
3. Proses normalisasi data menggunakan MinMaxScalar.
4. Evaluasi model dilakukan menggunakan metrik akurasi, precision, recall, F1-score dan AUC-ROC.
5. Teknik evaluasi model yang digunakan dalam penelitian ini adalah K-Fold Cross Validation.

6. Teknik evaluasi kinerja akhir yang digunakan dalam penelitian ini adalah Confusion Matrix

1.4 Tujuan Penelitian

Tujuan yang akan dicapai oleh peneliti dalam penelitian ini adalah:

1. Mengetahui pengaruh penggunaan SMOTE dalam meningkatkan kinerja klasifikasi machine learning dalam mendeteksi diabetes.
2. Mengetahui cara untuk mengoptimalkan algoritma yang digunakan agar mendapat hasil akurasi yang lebih baik.
3. Mengetahui pengaruh penggunaan Stacking Classifier dalam meningkatkan akurasi model dibandingkan dengan algoritma individu seperti Logistic Regression, Random Forest, SVM, dan KNN.

1.5 Manfaat Penelitian

Hal yang bermanfaat dari penelitian ini, sebagai berikut:

1. Bagi Pemerintah

Mendukung kebijakan kesehatan dalam prediksi dini penyakit diabetes dan meningkatkan efisiensi pelayanan diagnosa yang lebih cepat.

2. Bagi Warga

Mendapat informasi teknologi medis untuk mempercepat diagnosa dan memungkinkan deteksi dini serta tindakan pencegahan agar meningkatkan kualitas hidup.

3. Bagi Penulis

Mengembangkan ilmu dan keterampilan machine learning dalam analisis data serta memberikan kontribusi pada penelitian kesehatan.

1.6 Sistematika Penulisan

BAB I PENDAHULUAN

Bab ini berisi Latar belakang masalah, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, dan sistematika penulisan sebagai panduan awal dalam memahami keseluruhan penelitian.

BAB II TINJAUAN PUSTAKA

Bab ini berisi tinjauan pustaka, dasar-dasar teori yang digunakan terkait algoritma machine learning, metode yang digunakan dalam penelitian ini, serta penelitian-penelitian terdahulu yang mendukung penelitian.

BAB III METODE PENELITIAN

Bab ini terdapat metode yang digunakan dalam penelitian, tinjauan umum tentang objek penelitian, teknik pengumpulan data, analisis data, metode evaluasi performa algoritma, serta teknik optimasi yang diterapkan.

BAB IV HASIL DAN PEMBAHASAN

Bab ini merupakan tahapan yang penulis lakukan dalam menyajikan hasil eksperimen dan pengujian model machine learning pada dataset, analisis performa dari algoritma yang dibandingkan, serta pembahasan mengenai temuan dan interpretasi hasil penelitian.

BAB V PENUTUP

Bab ini berisi kesimpulan dan saran yang dapat peneliti rangkum selama proses penelitian serta pengembangan lebih lanjut, baik dalam penelitian maupun implementasi pada dunia nyata.