BAB I
PENDAHULUAN

1.1 Latar Belakang

Diabetes melitus merupakan salah satu penyakit kronis yang menjadi perhatian utama kesehatan global. Menurut International Diabetes Federation (IDF), pada tahun 2021 terdapat sekitar 537 juta orang dewasa di seluruh dunia yang menderita diabetes, dan angka ini diproyeksikan akan meningkat menjadi 783 juta pada tahun 2045 [1]. Diabetes tidak hanya berdampak pada kualitas hidup penderita, tetapi juga meningkatkan risiko komplikasi serius seperti penyakit jantung, gagal ginjal, kerusakan saraf, dan gangguan penglihatan [2]. Deteksi dini diabetes sangat penting untuk mencegah komplikasi lebih lanjut dan memungkinkan intervensi medis yang tepat waktu.

Perkembangan teknologi machine learning telah memberikan peluang besar dalam bidang kesehatan, khususnya untuk deteksi dan prediksi penyakit [3]. Berbagai algoritma machine learning seperti Logistic Regression, K-Nearest Neighbors (KNN), Random Forest, dan XGBoost telah digunakan untuk mengklasifikasikan risiko diabetes berdasarkan data kesehatan pasien[4], [5], [6]. Namun, dalam implementasinya, peneliti sering menghadapi tantangan berupa ketidakseimbangan kelas (class imbalance) dalam dataset, di mana jumlah individu yang tidak menderita diabetes jauh lebih banyak dibandingkan dengan yang menderita diabetes. Ketidakseimbangan dapat menyebabkan model cenderung memprediksi kelas mayoritas dan mengabaikan kelas minoritas, yang justru merupakan kasus yang paling penting untuk dideteksi [7].

Deteksi dini diabetes sangat penting untuk mencegah komplikasi lebih lanjut dan memberikan kesempatan untuk melakukan intervensi medis yang lebih efektif.[8] Namun, tantangan besar yang dihadapi dalam mendeteksi diabetes adalah ketidakseimbangan kelas dalam dataset, terutama dalam konteks pengklasifikasian risiko diabetes. Ketidakseimbangan kelas ini terjadi ketika jumlah data untuk individu yang tidak menderita diabetes jauh lebih banyak

dibandingkan dengan individu yang terdiagnosis diabetes, sehingga model machine learning yang dilatih dengan data seperti ini cenderung lebih sering memprediksi kelas mayoritas, yaitu "tidak diabetes", dan mengabaikan kelas minoritas, yaitu "diabetes" . Hal ini mengurangi kemampuan model untuk mendeteksi individu yang benar-benar berisiko diabetes, yang justru menjadi sasaran utama dari deteksi dini.

Dataset Behavioral Risk Factor Surveillance System (BRFSS) 2015 yang digunakan dalam penelitian memiliki karakteristik ketidakseimbangan kelas yang signifikan [9]. Dari 253.680 data, hanya sekitar 13,8% yang terdiagnosis diabetes, sementara 86,2% sisanya adalah individu yang tidak menderita diabetes. Kondisi ini dapat mengakibatkan model machine learning menghasilkan akurasi yang tinggi secara keseluruhan, namun gagal dalam mengidentifikasi kasus diabetes yang sebenarnya [7]. Oleh karena itu, diperlukan teknik khusus untuk menangani masalah ketidakseimbangan kelas agar model dapat bekerja secara optimal dan memberikan prediksi yang adil untuk kedua kelas.

SMOTE-ENN (Synthetic Minority Over-sampling Technique with Edited Nearest Neighbors) merupakan salah satu teknik yang efektif untuk menangani ketidakseimbangan kelas [10]. SMOTE bekerja dengan cara membuat sampel sintetis baru untuk kelas minoritas berdasarkan interpolasi dari sampel yang sudah ada, sehingga meningkatkan jumlah data kelas minoritas. Sementara itu, ENN (Edited Nearest Neighbors) melakukan pembersihan data dengan menghapus sampel yang berada terlalu dekat dengan kelas mayoritas atau yang berpotensi menyebabkan misklasifikasi [10]. Kombinasi kedua teknik tidak hanya menyeimbangkan distribusi kelas, tetapi juga meningkatkan kualitas data sehingga model dapat belajar dengan lebih baik. Penelitian ini bertujuan untuk menganalisis performa berbagai algoritma machine learning dalam mendeteksi diabetes setelah penerapan teknik SMOTE-ENN, guna mengidentifikasi model yang paling efektif dan akurat untuk deteksi dini penyakit diabetes.

Penerapan machine learning dalam bidang kesehatan, khususnya dalam mendeteksi penyakit diabetes, telah menunjukkan hasil yang menjanjikan.[11] Berbagai algoritma machine learning, seperti Logistic Regression, K-Nearest

Neighbors (KNN), Random Forest, dan XGBoost, telah digunakan untuk mengklasifikasikan risiko diabetes berdasarkan data kesehatan individu

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, maka rumusan masalah dalam penelitian ini adalah:

1. Bagaimana pengaruh teknik SMOTE-ENN dalam menangani ketidakseimbangan kelas pada dataset deteksi diabetes?
2. Bagaimana perbandingan kinerja algoritma Logistic Regression, K-Nearest Neighbors (KNN), Random Forest, dan XGBoost dalam mendeteksi diabetes setelah penerapan teknik SMOTE-ENN?
3. Algoritma machine learning manakah yang menghasilkan akurasi, precision, recall, dan F1-score terbaik untuk deteksi penyakit diabetes dengan kelas yang tidak seimbang?
4. Bagaimana implementasi sistem deteksi diabetes berbasis web menggunakan model machine learning terbaik yang telah dilatih dengan teknik SMOTE-ENN?

1.3 Batasan Masalah

Agar penelitian ini lebih terarah dan fokus, maka ditetapkan batasan masalah sebagai berikut:

1. Dataset yang digunakan adalah Behavioral Risk Factor Surveillance System (BRFSS) 2015 dengan 253.680 data dan 22 variabel yang terkait dengan faktor risiko diabetes.
2. Teknik penanganan ketidakseimbangan kelas yang digunakan adalah SMOTE-ENN (Synthetic Minority Over-sampling Technique with Edited Nearest Neighbors).
3. Algoritma machine learning yang dibandingkan adalah Logistic Regression, K-Nearest Neighbors (KNN), Random Forest, dan XGBoost.

4. Metrik evaluasi yang digunakan meliputi accuracy, precision, recall, F1-score, dan Area Under Curve (AUC-ROC).
5. Implementasi sistem dilakukan dalam bentuk aplikasi web berbasis Flask yang memungkinkan pengguna melakukan prediksi risiko diabetes berdasarkan data kesehatan pribadi.

1.4 Tujuan Penelitian

Tujuan yang ingin dicapai dari penelitian ini adalah:

1. Menganalisis efektivitas teknik SMOTE-ENN dalam menangani ketidakseimbangan kelas pada dataset deteksi diabetes.
2. Membandingkan kinerja algoritma Logistic Regression, K-Nearest Neighbors (KNN), Random Forest, dan XGBoost dalam mendeteksi diabetes setelah penerapan SMOTE-ENN berdasarkan metrik evaluasi accuracy, precision, recall, F1-score, dan AUC-ROC.
3. Mengidentifikasi algoritma machine learning terbaik untuk deteksi penyakit diabetes dengan kelas yang tidak seimbang.
4. Mengimplementasikan sistem deteksi diabetes berbasis web yang dapat digunakan untuk prediksi risiko diabetes secara interaktif.

1.5 Manfaat Penelitian

Manfaat yang diharapkan dari penelitian ini dapat dibagi menjadi dua aspek, yaitu manfaat teoritis dan manfaat praktis:

Manfaat Teoritis

Penelitian ini diharapkan dapat memberikan kontribusi terhadap pengembangan ilmu pengetahuan dalam bidang machine learning dan kesehatan, khususnya terkait penanganan ketidakseimbangan kelas dalam klasifikasi penyakit. Hasil penelitian dapat menjadi referensi bagi peneliti lain yang ingin mengembangkan sistem deteksi penyakit dengan karakteristik dataset yang tidak seimbang, serta memberikan wawasan tentang efektivitas teknik SMOTE-ENN dalam meningkatkan performa model klasifikasi.

Manfaat Praktis

1. Bagi Tenaga Kesehatan: Sistem yang dikembangkan dapat digunakan sebagai alat bantu untuk melakukan skrining awal risiko diabetes pada pasien, sehingga memungkinkan deteksi dini dan intervensi medis yang lebih cepat.
2. Bagi Masyarakat: Aplikasi web yang dihasilkan memberikan akses mudah bagi masyarakat umum untuk melakukan pemeriksaan mandiri terkait risiko diabetes berdasarkan data kesehatan pribadi mereka, sehingga meningkatkan kesadaran akan pentingnya kesehatan dan pencegahan diabetes.
3. Bagi Peneliti Selanjutnya: Hasil penelitian dapat menjadi landasan untuk pengembangan penelitian lebih lanjut dalam bidang machine learning untuk kesehatan, khususnya dalam eksplorasi teknik-teknik lain untuk menangani ketidakseimbangan kelas atau penerapan algoritma yang lebih canggih seperti deep learning.
4. Bagi Institusi Pendidikan: Penelitian dapat digunakan sebagai bahan pembelajaran dan referensi dalam mata kuliah terkait machine learning, data mining, dan aplikasi teknologi informasi di bidang kesehatan.

1.6 Sistematika Penulisan

BAB I PENDAHULUAN

Bab berisi latar belakang masalah yang menjadi dasar dilakukannya penelitian, rumusan masalah yang ingin dijawab, batasan masalah agar penelitian lebih fokus dan terarah, tujuan penelitian yang ingin dicapai, manfaat penelitian baik secara teoritis maupun praktis, serta sistematika penulisan yang menjelaskan struktur keseluruhan skripsi.

BAB II TINJAUAN PUSTAKA

Bab memuat kajian literatur dan studi penelitian terdahulu yang relevan dengan topik penelitian, serta dasar-dasar teori yang digunakan sebagai landasan penelitian. Teori-teori yang dibahas meliputi supervised learning, machine learning, penyakit diabetes, preprocessing data (Label Encoder dan StandardScaler), teknik SMOTE-ENN, algoritma-algoritma machine learning yang digunakan (Logistic Regression,

K-Nearest Neighbors, Random Forest, dan XGBoost), serta metrik evaluasi model (precision, recall, F1-score, accuracy, dan ROC-AUC).

BAB III METODE PENELITIAN

Bab menjelaskan secara detail tentang objek penelitian, alur penelitian yang dilakukan, serta alat dan bahan yang digunakan dalam penelitian. Alur penelitian mencakup tahapan data collection, preprocessing data (data preparation, data cleaning, label encoding, standardization, data split, dan penerapan SMOTE-ENN), pemodelan menggunakan algoritma machine learning, evaluasi model, hingga implementasi aplikasi web berbasis Flask untuk deteksi diabetes.

BAB IV HASIL DAN PEMBAHASAN

Bab menyajikan hasil implementasi dari setiap tahapan penelitian, mulai dari pengumpulan data, preprocessing, penerapan SMOTE-ENN, pelatihan model dengan berbagai algoritma machine learning, hingga evaluasi performa masing-masing model. Pembahasan juga mencakup visualisasi data, confusion matrix, kurva ROC, dan perbandingan kinerja algoritma berdasarkan metrik evaluasi. Selain itu, bab ini menampilkan implementasi aplikasi web yang telah dikembangkan beserta cara penggunaannya.

BAB V PENUTUP

Bab berisi kesimpulan yang merupakan jawaban dari rumusan masalah dan pencapaian tujuan penelitian, serta saran-saran untuk pengembangan penelitian lebih lanjut di masa mendatang. Kesimpulan merangkum temuan utama penelitian terkait efektivitas SMOTE-ENN dan perbandingan kinerja algoritma machine learning dalam deteksi diabetes, sementara saran memberikan rekomendasi untuk perbaikan dan pengembangan sistem di masa depan.