BAB I
PENDAHULUAN

1. Latar Belakang

Penyakit diabetes merupakan penyakit kronis dengan prevalensi terus meningkat secara global sehingga deteksi dini menjadi aspek penting dalam pengendalian penyakit ini. Teknik *machine learning* banyak digunakan dalam prediksi penyakit diabetes karena kemampuanya mengelola data klinis dalam jumlah besar serta mengenali pola yang kompleks. Namun, salah satu tantangan utama pada dataset medis, termasuk data diabetes, adalah ketidakseimbangan kelas (*imbalanced data*), dimana jumlah data pasien non-diabetes jauh lebih dominan dibandingkan pasien sebagai kelas minoritas. Kondisi ini berdampak pada menurunnya kemampuan model klasifikasi dalam mengenali pasien diabetes secara akurat [1][2].

Ketidakseimbangan data menyebabkan algoritma pembelajaran mesin seperti Random Forest dan XGBoost cenderung bias terhadap kelas mayoritas, sehingga nilai matrik evalausi seperti recall dan F1-score pada kelas minoritas menjadi kurang optimal. Oleh karena itu, diperlukan Teknik penanganan imbalanced data sebelum proses pelatihan model agar peforma klasifikasi menjadih lebih seimbang. Berbagai metode sampling telah diterapkan untuk memperbaiki distribusi kelas, seperti oversampling, undersampling, serta hybred sampling yang bertujuan untuk mengurangi kesalahan klasifikasi pada kelas minoritas [3].

Pada penelitian juga menunjukan bahwa penerapan Teknik pemyeimbangan data pada dataset diabetes mampu meningkatkan peforma model Random Forest dan XGBoost, khususnya dalam mendetekasi pasien diabetes sebagai kelas minoritas. Penelitian lain juga melaporkan bahwa pendekatan penanganan ketidakseimbangan kelas dapat membantu model mempelajari karakteristik kelas minoritas secara efektif. Temuan tersebut menegaskan bahwa penanganan ketidakseimbangan kelas merupakan Langkah krusial dalam

meningkatkan peforma klasifikasi pada kasus diabetes [3].

Berdasarkan permasalahan tersebut, penelitian ini akan menganalisis pengaruh berbagai metode sampling terhadap kinerja model klasifikasi Random Forest dan XGBoost pada dataset diabetes yang memiliki ketidakseimbangan kelas. Evaluasi kinerja model akan dilakukan menggunakan metrik F1-Score dan akurasi untuk menentukan metode sampling yang paling efektif dalam meningkatkan kemampuan model dalam memprediksi kelas minoritas.

2. Rumusan Masalah

Berdasar latar belakang masalah seperti di 1.1, maka rumusan masalah dalam penelitian ini adalah sebagai berikut:

1. Bagaimana pengaruh ketidakseimbangan kelas (imbalanced data) terhadap performa model Random Forest dan XGBoost dalam klasifikasi diabetes?
2. Bagaimana pengaruh penerapan metode sampling SMOTE, ROS, SMOTENN, ADASYN, RUS, ClusterCentroids, Tomek Links, and ENN terhadap peningkatan performa model klasifikasi diabetes?
3. Metode sampling manakah yang memberikan performa terbaik dalam menangani data imbalanced pada klasifikasi diabetes?

3. Batasan Masalah

Agar penelitian ini lebih terfokus dan tidak menyimpang dari tujuan yang telah ditetapkan, maka diberikan Batasan masalah sebagai berikut:

1. Penelitian ini menggunakan dataset diabetes yang diperoleh dari platform Kaggle sebagai sumber data sekunder, dengan karakteristik ketidakseimbangan kelas anatar pasien diabetes dan non-diabetes.
2. Algoritma klasifikasi yang digunakan dalam penelitian ini dibatasi pada Random Forest dan XGBoost.
3. Teknik penangan ketidakseimbangan kelas yang diterapkan dibatasi

pada metode sampling, yang meliputi oversampling dan undersampling

4. Proses evaluasi kinerja model dilakukan menggunakan metrik akurasi dan F1-score.

4. Tujuan Penelitian

Adapun dalam penelitian ini yang bertujuan untuk sebagai berikut:

1. Menganalisis kinerja model Random Forest dan XGBoost pada klasifikasi diabetes dengan kondisi imbalanced data.
2. Mengetahui pengaruh berbagai metode sampling SMOTE, ROS, SMOTENN, ADASYN, RUS, ClusterCentroids, Tomek Links, and ENN dalam meningkatkan kemampuan model memprediksi kelas minoritas diabetes.
3. Menentukan metode sampling paling efektif berdasarkan evaluasi F1-score dan akurasi. buatkan poin.

5. Manfaat Penelitian

Hasil penelitian ini diharapkan dapat memberikan manfaat sebagai berikut:

1. Memberikan pemahaman tentang kinerja Random Forest dan XGBoost pada klasifikasi diabetes dengan kondisi data tidak seimbang.
2. Mengetahui efektivitas berbagai metode sampling dalam meningkatkan kemampuan model mendeteksi kelas minoritas (pasien diabetes).
3. Memberikan rekomendasi metode sampling optimal berdasarkan evaluasi F1-score dan akurasi untuk menangani masalah imbalanced data.
4. Membantu peneliti atau praktisi dalam memilih algoritma dan metode sampling yang tepat untuk klasifikasi data medis.

6. Sistematika Penulisan

Skripsi ini disususn dengan sistematika penulisan yang terdiri dari lima bab untuk mempermudah pembaca memahami alur penelitian :

1) BAB I PENDAHULUAN, berisi latar belakang masalah, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, dan sistematika penulisan. Bab ini memberikan gambaran umum mengenai permasalahan yang diteliti dan arah penelitian.
2) BAB II TINJAUAN PUSTAKA memuat studi literatur terkait penelitian sebelumnya, dasar teori mengenai machine learning, ketidakseimbangan kelas pada dataset medis, serta metode sampling yang digunakan untuk menangani masalah tersebut. Bab ini menjadi landasan teoritis dan konsep yang mendukung penelitian.
3) BAB III METODE PENELITIAN menjelaskan objek penelitian, jenis dan sumber data, tahapan preprocessing, metode pemodelan Random Forest dan XGBoost, teknik penanganan data imbalance, serta metrik evaluasi kinerja model. Bab ini juga membahas desain penelitian dan alur eksperimen secara sistematis.
4) BAB IV HASIL DAN PEMBAHASAN memaparkan hasil pemodelan, analisis kinerja model pada berbagai metode sampling, pembahasan temuan utama, serta interpretasi hasil eksperimen untuk menjawab rumusan masalah penelitian.
5) BAB V PENUTUP berisi kesimpulan yang diperoleh dari penelitian, saran untuk pengembangan penelitian selanjutnya, serta kontribusi penelitian bagi pengembangan ilmu pengetahuan dan praktik di bidang kesehatan dan data mining.