Vol. 17, No. 1, April 2025

IMPROVE

ISSN(e): - / ISSN(p) : 1979-8342

MENERAPKAN SMOTE PADA KLASIFIKASI DATA PENYAKIT STROKE

Muhammad Ibnu Choldun Rachmatullah[1], Sari Armiati[2], Mubassiran[3]

[1] Universitas Telkom, [2,3] Universitas Logistik dan Bisnis Internasional

[1]muhammadibnucholdun@telkomuniversity.ac.id,[2]sari@ulbi.ac.id,[3]mubassiran@ulbi.ac.id

Abstrak— Masalah ketidakseimbangan kelas (imbalanced dataset) merupakan tantangan utama dalam pengolahan data, terutama pada sistem klasifikasi biner seperti prediksi penyakit stroke. Model klasifikasi cenderung bias terhadap kelas mayoritas, yang menyebabkan performa rendah dalam mendeteksi kelas minoritas. Penelitian ini menerapkan metode SMOTE (Synthetic Minority Over-sampling Technique) untuk menyeimbangkan dataset stroke dari Kaggle yang terdiri dari 5110 data pasien. Model klasifikasi yang digunakan adalah Random Forest, dengan pembagian data 80% untuk pelatihan dan 20% untuk pengujian. Hasil eksperimen menunjukkan bahwa sebelum penerapan SMOTE, model memiliki akurasi tinggi sebesar 93,93% namun gagal mendeteksi kasus stroke (precision, recall, dan F1-score = 0%). Setelah penerapan SMOTE, recall meningkat menjadi 14,52%, precision menjadi 15,52%, dan F1-score menjadi 15,00%, meskipun akurasi menurun menjadi 90,02%. Hal ini menunjukkan bahwa SMOTE berhasil meningkatkan sensitivitas model terhadap kelas minoritas, menjadikannya lebih efektif untuk deteksi kondisi medis yang jarang terjadi.

Kata kunci— Imbalanced, Dataset, SMOTE, Random Forest, Penyakit Stroke

Abstract— The problem of class imbalance (imbalanced dataset) is a major challenge in data processing, especially in binary classification systems such as stroke prediction. Classification models tend to be biased towards the majority class, resulting in low performance in detecting minority classes. This study applies the SMOTE (Synthetic Minority Over-sampling Technique) method to balance a stroke dataset from Kaggle consisting of 5,110 patient data. The classification model used is Random Forest, with 80% data for training and 20% for testing. Experimental results show that before the application of SMOTE, the model had a high accuracy of 93.93% but failed to detect stroke cases (precision, recall, and F1-score = 0%). After the application of SMOTE, recall increased to 14.52%, precision to 15.52%, and F1-score to 15.00%, although accuracy decreased to 90.02%. This indicates that SMOTE successfully increases the model's sensitivity to the minority class, making it more effective for detecting rare medical conditions.

***Keywords*— Imbalanced, Dataset, SMOTE, Random Forest, Stroke Disease**

I. PENDAHULUAN

Dalam era data besar (big data), peran data mining dan pembelajaran mesin (machine learning) semakin dominan dalam membantu pengambilan keputusan berbasis data. Salah satu tantangan terbesar yang sering dihadapi dalam pengolahan data adalah masalah ketidakseimbangan kelas (imbalanced dataset). Ketidakseimbangan ini terjadi ketika distribusi kelas dalam dataset sangat tidak berimbang, sehingga satu atau beberapa kelas memiliki jumlah sampel yang jauh lebih sedikit dibandingkan kelas lainnya. Masalah ini sering ditemukan dalam berbagai bidang seperti deteksi penipuan (fraud detection), diagnosis medis, prediksi kegagalan mesin, dan lainnya.

Model klasifikasi cenderung bias terhadap kelas mayoritas, sehingga performa pada kelas minoritas sangat rendah. Hal ini menyebabkan ketidakakuratan sistem dalam mengidentifikasi kejadian penting namun jarang terjadi, seperti kasus penyakit langka atau transaksi penipuan. Untuk mengatasi permasalahan ini, beberapa teknik telah dikembangkan, salah satunya adalah metode SMOTE (Synthetic Minority Over-sampling Technique). SMOTE adalah metode over-sampling yang bekerja dengan cara menciptakan sampel sintetis berdasarkan interpolasi linier dari sampel minoritas yang ada. Teknik ini dianggap lebih unggul dibanding over-sampling konvensional seperti duplikasi, karena tidak menyebabkan overfitting seburuk metode duplikasi [1].

Ketidakseimbangan data memiliki pengaruh signifikan terhadap performa algoritma klasifikasi. Ketidakseimbangan kelas menyebabkan algoritma seperti Decision Tree dan SVM menghasilkan akurasi tinggi tetapi sebenarnya hanya berhasil memprediksi kelas mayoritas, sehingga metrik seperti precision, recall, dan F1-score menjadi penting untuk mengevaluasi kinerja secara menyeluruh. Dalam konteks sistem klasifikasi real-world, data yang tidak seimbang adalah hal umum. Misalnya, dalam bidang kesehatan, dataset untuk mendeteksi kanker payudara memiliki kasus positif (kanker) yang jauh lebih sedikit dibandingkan dengan kasus negatif (tidak kanker). Jika model hanya mengandalkan akurasi sebagai metrik utama, maka prediksi semua data sebagai kelas negatif tetap memberikan akurasi tinggi namun kurang memberikan makna dalam aplikasi nyata.

SMOTE diperkenalkan oleh Chawla dkk. (2002) sebagai metode untuk menciptakan data sintetis dari data minoritas. Algoritma ini bekerja dengan memilih sampel dari kelas minoritas dan kemudian menciptakan titik baru di sepanjang garis yang menghubungkan titik tersebut dengan tetangga terdekatnya. Teknik ini terbukti dapat meningkatkan performa klasifikasi dibanding hanya melakukan duplikasi data minoritas. Beberapa penelitian telah menunjukkan bahwa SMOTE mampu meningkatkan sensitivitas model terhadap kelas minoritas. Sebagai contoh, Yulianti dkk. (2024) menerapkan SMOTE pada sistem prediksi risiko kredit. Penelitian-penelitian juga menunjukkan efektivitas SMOTE dalam berbagai kasus. Handoko dan Aditya melakukan penelitian berkaitan dengan deteksi penyakit, menggunakan SMOTE digunakan untuk menyeimbangkan dataset diagnosis diabetes. Marzuqi dkk. melakukan penelitian berkaitan dengan prediksi dropout mahasiswa, menunjukkan bahwa penerapan SMOTE berhasil meningkatkan akurasi prediksi mahasiswa yang berpotensi drop out [4]. Klasifikasi Kurnia dan Saputra melakukan penelitian tentang analisis sentimen, membuktikan bahwa SMOTE mampu memperbaiki performa klasifikasi sentimen negatif yang sebelumnya terabaikan [5]. SMOTE sering digunakan bersamaan dengan algoritma seperti Decision Tree, K-Nearest Neighbor, Naive Bayes, Random Forest, dan SVM. Kombinasi ini telah terbukti meningkatkan performa model secara signifikan.

Pada penelitian ini dilakukan penerapan SMOTE pada dataset penyakit stroke dengan menggunakan pengklasifikasi random forest. Bagian awal berisi pendahuluan, dilanjutkan dengan metode penelitian. Selanjutnya dilakukan analisis dan pembahasan dari hasil eksperimen, kemudian diakhiri dengan kesimpulan.

II. METODOLOGI PENELITIAN

Metode penelitian terdiri dari tahapan-tahapan penelian sebagai berikut, yang masing-masing tahapan diuraikan pada subbab berikutnya:

a. Menyiapkan dataset
b. Menyeimbangkan data dengan SMOTE
c. Menentukan metode split data
d. Menentukan metode pengklasifikasi
e. Menghitung kinerja.

A. Menyiapkan Dataset

Dataset yang digunakan dalam penelitian ini adalah dataset healthcare-dataset-stroke-data.csv yang diambil dari Kaggle. Dataset Stroke Prediction ini terdiri dari 5110 data pasien dengan 12 atribut yang merepresentasikan informasi demografis, riwayat kesehatan, dan gaya hidup. Setiap baris mewakili satu individu, dengan kolom stroke sebagai variabel target yang menunjukkan apakah pasien pernah mengalami stroke (1) atau tidak (0). Atribut lainnya mencakup jenis kelamin (gender), usia (age), status hipertensi dan penyakit jantung (hypertension, heart_disease), status pernikahan (ever_married), jenis pekerjaan (work_type), tipe tempat tinggal (Residence_type), kadar rata-rata glukosa darah (avg_glucose_level), indeks massa tubuh (bmi), serta status merokok (smoking_status). Sebagian besar data lengkap, kecuali kolom bmi yang memiliki 201 nilai hilang. Dataset ini relevan untuk digunakan dalam penelitian klasifikasi biner, khususnya prediksi risiko stroke, dan sangat potensial untuk penerapan teknik penyeimbangan data seperti SMOTE karena kemungkinan besar distribusi kelas target tidak seimbang.

B. Menyeimbangkan Data dengan SMOTE

Tujuan dari penyeimbangan dataset adalah agar supaya setiap kelas mempunyai mempunyai frekuensi/jumlah data yang sama. Pada penelitian ini untuk menyeimbangkan data menggunakan teknik yang disebut dengan Synthetic Minority Oversampling Technique (SMOTE) yang pertama kali diusulkan oleh Chawla dkk [1] [6]. Dengan proses SMOTE maka yang sebelumnya jumlah data antara kelas mayoritas dengan kelas minoritas tidak seimbang menjadi seimbang.

C. Menentukan Metode Split Data

Metode pemisahan data yang digunakan adalah dengan menggunakan proporsi sebagai berikut: 80 % untuk training dan 20 % untuk pengujian.

D. Menentukan metode pengklasifikasi

Decision Tree memberikan teknik mengklasifikasikan data dengan menghasilkan struktur seperti pohon. Node internal pohon mewakili pilihan (biner) untuk setiap atribut, sedangkan cabang pohon menandakan hasil dari pilihan yang diinginkan. Dalam beberapa tahun terakhir, banyak jenis pohon keputusan telah diperkenalkan oleh para peneliti, dan salah satu pengklasifikasi yang berbentuk pohon yang paling banyak digunakan adalah Random Forest (RF). RF mewakili kumpulan pohon yang diproduksi untuk menghindari risiko ketidakstabilan dan meminimalkan kemungkinan pelatihan sampel yang berlebihan. Pohon-pohon ini juga dibuat untuk mengurangi overfitting menggunakan teknik pemangkasan [7][8]. Random forest ini secara progresif mengurangi node tanpa mengganggu kinerja classifier secara keseluruhan.

E. Menghitung Kinerja

Untuk menghitung kinerja metode pengklasifikasi pada data yang seimbang kurang cocok jika menggunakan ukuran akurasi, karena nilai akurasi sangat tergantung pada kelas data mayoritas, padahal ternyata kelas minoritas sebenarnya yang seharusnya lebih mendapatkan perhatian. Walaupun bukan ukuran kinerja yang cocok untuk klasifikasi data tidak seimbang, nilai akurasi tetap akan dihitung untuk melihat perubahan nilai akurasi antara sebelum dan sesudah penerapan SMOTE. Untuk mengukur kinerja algoritma klasifikasi pada data tidak seimbang , empat metrik evaluasi digunakan. Metrik evaluasi ini didasarkan pada confusion matrix [9][10] seperti yang ditunjukkan pada tabel 1. Matriks ini terdiri dari dua baris

dan dua kolom yaitu: TP (True Positive), TN (True Negative), FP (False Positive), dan FN (False Negative).

TABEL I
CONFUSION MATRIX

Kelas Aktual	Kelas hasil prediksi	
	Positif	Negatif
Positif	TP	TN
Negatif	FP	FN

Rumus dari ukuran kinerja yang digunakan adalah sebagai berikut:

$$Recall = \frac{TP}{TP+FN} \quad (1)$$

$$Presisi = \frac{TP}{TP+FP} \quad (2)$$

$$F1 - Score = 2 * \frac{(Recall*Presisi)}{(Recall+Presisi)} \quad (3)$$

Area di bawah kurva ROC (Receiver Operating Characteristic) atau Area Under Curve (AUC) adalah metrik evaluasi lain yang digunakan untuk mengukur kinerja kumpulan data yang tidak seimbang [11]. ROC adalah kurva dua dimensi yang mewakili kompromi antara tingkat True Positive dan False Positive. Sedangkan area di bawah kurva ROC atau AUC digunakan untuk menilai keakuratan pengklasifikasi. Pengklasifikasi yang memberikan nilai AUC lebih tinggi artinya mempunyai kinerja yang lebih baik. Ukuran akurasi juga dihitung sebagai tambahan dengan rumus sebagai berikut:

$$Akurasi = \frac{TP+TN}{TP+FP+FN+TN} \quad (4)$$

III. HASIL DAN PEMBAHASAN

Dataset tersebut memiliki 5110 baris dan 12 kolom yang mencakup informasi demografi dan klinis pasien seperti usia, jenis kelamin, riwayat hipertensi dan penyakit jantung, status pernikahan, jenis pekerjaan, kadar glukosa darah rata-rata, indeks massa tubuh (BMI), status merokok, dan label target yaitu apakah pasien mengalami stroke. Langkah pertama dalam pra-pemrosesan adalah menangani nilai yang hilang. Kolom bmi teridentifikasi memiliki nilai kosong dan diimputasi menggunakan median untuk menjaga kestabilan distribusi data. Setelah itu, semua fitur kategorikal diidentifikasi dan dikonversi menjadi representasi numerik menggunakan teknik One-Hot Encoding. Model klasifikasi yang digunakan dalam eksperimen ini adalah Random Forest Classifier, sebuah ensemble model yang kuat dalam menangani data dengan fitur-fitur yang beragam dan non-linear. Model pertama dilatih menggunakan data pelatihan tanpa modifikasi apa pun terhadap distribusi kelas. Setelah pelatihan, model diuji pada data pengujian dan metrik evaluasi yang dihitung mencakup: akurasi, presisi, recall, dan F1-score. Berdasarkan output, model menunjukkan akurasi yang cukup tinggi, tetapi nilai recall dan F1-score cenderung rendah, terutama dalam mendeteksi kasus stroke (minority class). Hal ini adalah indikator klasik dari ketidakseimbangan kelas, di mana model cenderung bias terhadap mayoritas (tidak terkena stroke), sehingga gagal mengenali dengan baik kelas minoritas.

Untuk mengatasi masalah di atas maka diterapkan SMOTE (Synthetic Minority Over-sampling Technique) pada data pelatihan. SMOTE bekerja dengan menghasilkan sampel sintetis dari kelas minoritas dengan cara interpolasi di ruang fitur, sehingga distribusi kelas menjadi lebih seimbang tanpa menghapus data mayoritas. Setelah SMOTE diterapkan, model Random Forest kembali dilatih pada data pelatihan yang telah diseimbangkan. Ketika model diuji pada set pengujian yang sama, terjadi peningkatan signifikan pada metrik recall dan F1-score, menunjukkan bahwa model kini mampu mengenali lebih banyak kasus stroke dibandingkan sebelumnya. Walaupun dalam beberapa kasus presisi bisa menurun (karena bertambahnya false positive), peningkatan recall dan F1 menunjukkan bahwa model menjadi lebih sensitif terhadap kasus stroke yang penting dalam konteks medis.

Dari hasil eksperimen didapatkan hasil sebagai berikut:

TABEL III
HASIL EKSPERIMEN

Ukuran Kinerja	Tanpa SMOTE	Dengan SMOTE
Accuracy	93,93%	90,02%
Precision	0,00%	15,52%
Recall	0,00%	14,52%
F1-score	0,00%	15,00%

Hasil eksperimen menunjukkan perbandingan kinerja model klasifikasi dalam dua kondisi: tanpa menggunakan SMOTE dan dengan menggunakan SMOTE. Evaluasi dilakukan berdasarkan empat metrik utama: akurasi, presisi, recall, dan F1-score. Secara umum, terlihat bahwa penerapan SMOTE memberikan dampak signifikan terhadap kemampuan model dalam mengenali kelas minoritas (kasus stroke), meskipun menyebabkan sedikit penurunan pada akurasi keseluruhan.

Tanpa penggunaan SMOTE, model mencapai akurasi tinggi sebesar 93.9%, namun metrik lain seperti precision, recall, dan F1-score semuanya bernilai nol. Hal ini mengindikasikan bahwa meskipun model tampak sangat akurat secara keseluruhan, ia sama sekali tidak mampu mengklasifikasikan kasus stroke dengan benar. Ini merupakan gejala klasik dari ketidakseimbangan kelas yang ekstrem, di mana model lebih memilih untuk selalu memprediksi kelas mayoritas (tidak stroke), karena secara statistik hal tersebut menghasilkan akurasi tinggi. Namun, dari sudut pandang aplikasi nyata seperti prediksi penyakit, kegagalan dalam mendeteksi kasus positif sama artinya dengan kegagalan sistem secara keseluruhan.

Setelah SMOTE diterapkan untuk menyeimbangkan jumlah data antara kelas positif dan negatif, performa model berubah secara signifikan. Akurasi menurun menjadi 90.0%, namun yang lebih penting adalah presisi meningkat menjadi 15.5%, recall menjadi 14.5%, dan F1-score menjadi 15.0%. Meskipun angka-angka ini masih

rendah, mereka jauh lebih representatif dibandingkan nol mutlak sebelumnya. Ini menunjukkan bahwa model mulai mampu mengidentifikasi sebagian kecil dari kasus stroke yang sebenarnya. Penurunan akurasi sebesar sekitar 3.9% adalah pengorbanan yang wajar dalam konteks klasifikasi ketidakseimbangan, karena tujuan utamanya bukan sekadar akurasi global, tetapi deteksi kelas minoritas yang krusial secara klinis.

Dari perspektif medis dan praktis, recall dan F1-score adalah metrik yang lebih penting dibanding akurasi ketika menyangkut deteksi penyakit. Dalam kasus ini, peningkatan recall dari 0 menjadi 14.5% berkat SMOTE adalah sangat berarti, karena recall mengukur kemampuan model dalam mendeteksi kasus yang benar-benar stroke. F1-score yang meningkat menunjukkan bahwa keseimbangan antara precision dan recall juga membaik, meskipun masih jauh dari ideal.

REFERENSI

[1] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). "SMOTE: Synthetic Minority Over-sampling Technique." Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953

[2] T. Yulianti, A. H. Cahyana, M. Komarudin, Y. Mulyani, dan H. D. Septama, "Penilaian Pembayaran Kredit dengan Logistic Regression dan Random Forest pada Home Credit," *Jurnal Pseudocode*, vol. 11, no. 2, pp. 79–88, Sep. 2024, doi: 10.33369/pseudocode.11.2.79-88.

[3] C. B. Handoko dan C. S. K. Aditya, "Penerapan Teknik SMOTE dalam Mengatasi Imbalance Data Penyakit Diabetes Menggunakan Algoritma ANN," *Smart Comp: Jurnalnya Orang Pintar Komputer*, vol. 14, no. 1, pp. 13–20, Jan. 2025, doi: 10.30591/smartcomp.v14i1.7045.

[4] T. A. Marzuqi, E. Kristiani, dan Marcel, "Prediksi Mahasiswa Drop-Out di Universitas XYZ," *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)*, vol. 11, no. 6, pp. 1345–1350, Des. 2024, doi: 10.25126/jtiik.2024118689.

[5] Kurnia, I. Purnamasari, dan D. D. Saputra, "Analisis Sentimen dengan Metode Naïve Bayes, SMOTE dan Adaboost pada Twitter Bank BTN," Jurnal JTIK (Jurnal Teknologi Informasi dan Komunikasi), vol. 7, no. 2, pp. 235–242, Apr. 2023, doi: 10.35870/jtik.v7i2.707.

[6] Fernandez, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for Learning from Imbalanced Data : Progress and Challenges , Marking the 15-year Anniversary. Journal of Artificial Intelligence Research, 61, 863–905.

[7] Erdiansyah, U., Lubis, A. I., & Erwansyah, K. (2022). Komparasi Metode K-Nearest Neighbor dan Random Forest Dalam Prediksi Akurasi Klasifikasi Pengobatan Penyakit Kutil. JURNAL MEDIA INFORMATIKA BUDIDARMA, 6, 208–214. https://doi.org/10.30865/mib.v6i1.3373

[8] Yuliati, I. F., & Sihombing, P. R. (2021). Penerapan Metode Machine Learning dalam Klasifikasi Risiko Kejadian Berat Badan Lahir Rendah di Indonesia Implementation of Machine Learning Method in Risk Classification on Low Birth weight in Indonesia. Matrik: Jurnal Manajemen, Teknik Informatika, Dan Rekayasa Komputer, 20(2), 417–426. https://doi.org/10.30812/matrik.v20i2.1174

[9] Indransyah, R., Chrisnanto, Y. H., Sabrina, P. N., Informatika, P. S., Jenderal, U., Yani, A., Abdillah, N., & Sentimen, K. (2022). KLASIFIKASI SENTIMEN PERGELARAN MOTOGP DI INDONESIA MENGGUNAKAN ALGORITMA CORRELATED NAÏVE BAYES CLASIFIER. 60–66.

[10] Nabillah, A., Alam, S., & Resmi, M. G. (2022). Twitter User Sentiment Analysis Of TIX ID Applications Using Support Vector Machine Algorithm. 3(1), 14–27.

[11] Anis, M., & Ali, M. (2017). Investigating the Performance of Smote for Class Imbalanced Learning : A Case Study of Credit Scoring Datasets. European Scientific Journal, 13(33), 340–353. https://doi.org/10.19044/esj.2017.v13n33p340,