Pengujian Metode SMOTE Untuk Penanganan Data Tidak Seimbang Pada Dataset Binary

Fandi Yulian Pamuji[1]

[1] *Universitas Merdeka Malang, Indonesia*
e-mail: fandi.pamuji@unmer.ac.id

Kata Kunci:

Pengujian
SMOTE
Data Tidak Seimbang
Dataset Binary

ABSTRAK

Dataset tidak seimbang jika berukuran besar maka tidak berpengaruh besar terhadap nilai akurasi, tetapi masalah akan muncul ketika dihadapkan pada data yang tidak normal seperti distribusi kelas tidak seragam di antara kelas-kelas dengan memiliki jumlah kelas mayoritas lebih banyak daripada kelas minoritas. Penelitian ini bertujuan untuk mengatasi jumlah kelas yang tidak seimbang agar tetap ideal dengan menggunakan metode SMOTE pada data tidak seimbang. Berdasarkan hasil eksperimen yang telah dilakukan, maka penulis menyimpulkan penelitian ini yaitu bahwa pengujian metode SMOTE dengan metode klasifikasi mampu menangani jumlah kelas mayoritas (negatif) dan kelas minoritas (positif) pada data tidak seimbang dengan menghasilkan nilai MCC dan Gmean mencapai kinerja prediksi yang lebih besar. Kinerja dari metode diukur dari nilai MCC dan Gmean dari masing-masing data tidak seimbang, namun untuk dataset Binary nilai MCC dan Gmean yang paling tinggi menggunakan metode SMOTE + KKN dengan nilai MCC = 0,90 dan nilai Gmean = 0,95 juga dapat mencapai kinerja prediksi yang lebih besar.

Keyword:

Testing
SMOTE
Unbalanced Data
Binary Dataset

ABSTRACT

If the dataset is not balanced, if it is large, it does not have a big effect on the accuracy value, but problems will arise when faced with abnormal data such as the class distribution is not uniform among classes by having more majority classes than minority classes. This study aims to overcome the number of unbalanced classes in order to remain ideal by using the SMOTE method on unbalanced data. Based on the results of the experiments that have been carried out, the authors conclude that this study is that the SMOTE method test with the classification method is able to handle the number of majority (negative) and minority (positive) classes in unbalanced data by producing MCC and Gmean values to achieve greater predictive performance. . The performance of the method is measured by the MCC and Gmean values of each unbalanced data, but for the Binary dataset the highest MCC and Gmean values using the SMOTE + KKN method with MCC values = 0.90 and Gmean values = 0.95 can also reach greater predictive performance.

ISSN: 2598-0076

PENDAHULUAN

Klasifikasi dataset tidak seimbang akan menimbulkan masalah pada data mining dan machine learning[1]. Himpunan data yang tidak seimbang adalah kasus khusus untuk masalah klasifikasi di mana distribusi kelas tidak seragam di antara kelas-kelas. Biasanya, mereka terdiri dari dua kelas: kelas mayoritas (negatif) dan kelas minoritas (positif)[2]. Dataset tidak seimbang jika berukuran besar maka tidak berpengaruh besar terhadap nilai akurasi, tetapi masalah akan muncul ketika dihadapkan pada data yang tidak normal seperti distribusi kelas tidak seragam di antara kelas-kelas dengan memiliki jumlah kelas mayoritas (negatif) lebih banyak daripada kelas minoritas (positif)[3]. Untuk mengatasi permasalahan dataset tidak seimbang adalah dengan menyeimbangkan distribusi kelas tidak seragam di antara kelas-kelas dengan menggunakan metode SMOTE over-sampling atau under-sampling supaya jumlahnya seimbang dari kelas mayoritas (negatif) maupun kelas minoritas (positif)[4]. Pada penelitian lain dijelaskan ketidakseimbangan data bukan satu-satunya yang mengakibatkan kurangnya kinerja klasifikasi, tetapi faktor data lain seperti disjungsi kecil, kebisingan, dan tumpang tindih[5].

Metode klasifikasi yang digunakan untuk mencari nilai akurasi menggunakan machine learning, kemudian dataset tersebut diuji dan di evaluasi yang akan dipilih metode yang menghasilkan nilai akurasi paling baik[6]. Selain itu ada metode data tidak seimbang, di uji dan di evaluasi dengan menggunakan metode SMOTE[7]. Pada penelitian ini menangani ketidakseimbangan pada kumpulan data repository KEEL dari kelas mayoritas (negatif) dan kelas minoritas (positif). Penelitian ini bertujuan untuk mempertahankan jumlah kelas yang tidak seimbang agar tetap ideal dengan melakukan imputasi pada dataset yang tidak seimbang[8].

METODE

SMOTE

SMOTE sebagai kombinasi dari mayoritas under-sampling dan mayoritas over-sampling. Bagian under-sampling hanyalah prosedur under-sampling umum. Untuk bagian over-sampling dari SMOTE, sampel sintetis dibuat secara acak dengan menambahkan selisih bobot antara sampel ke-i dan k tetangga terdekatnya[9]. Dengan metode over-sampling/under-sampling dengan mudah dapat membuat data-set menjadi seimbang tetapi metode ini mempunyai kelemahan, over-sampling pada data-set minority akan menuju model yang overfitting, karena over-sampling dilakukan dengan duplikasi data yang sudah mempunyai nilai yang sudah kecil, under-sampling pada majority juga dapat mengakibat data yang penting pembeda dua kelas menjadi diluar dari data-set[10].

ISSN: 2598-0076

O adalah kumpulan data asli
P adalah himpunan instance positif (instance kelas minoritas)
  Untuk setiap contoh *x* dalam *P*
  Temukan k-tetangga terdekat (instance kelas minoritas) ke *x* dalam *P*
  Dapatkan *y* dengan mengacak satu dari k instance
  selisih = *x* – *y*
  gap = bilangan acak antara *0* dan *1*
  n = *x* + *selisih* * *gap*
  Tambahkan *n* ke *O*
Akhir

Gambar 1. *The Synthetic Minority Oversampling Technique* (SMOTE)

Naïve Bayes

Naïve Bayes adalah salah satu metode machine learning yang memanfaatkan perhitungan probabilitas dan statistik, yaitu memprediksi peluang di masa depan berdasarkan pengalaman dimasa sebelumnya sehingga dikenal sebagai Teorema Bayes[11]. Rumus Naïve Bayes menggunakan persamaan sebagai berikut:

$$P(\mathrm{C}|\mathrm{X}) = \frac{P(x|c)P(c)}{P(x)} \quad (1)$$

Logistic Regression

Logistic Regression adalah kasus khusus dari model linier umum yang menyangkut analisis data biner. Logistic Regression merupakan algoritma klasifikasi machine learning yang digunakan untuk memprediksi probabilitas variabel dependen kategoris[12]. Dalam fungsi logistic, variabel dependen adalah variabel biner yang berisi data berkode 1 (berhasil) atau 0 (gagal), di mana fungsi tautannya adalah fungsi logistic menggunakan persamaan sebagai berikut:

$$p_i = \frac{1}{1 + e^{-(w^T x_i + b)}} \quad (2)$$

KNN

k-Nearest Neighbor (k-NN) adalah algoritma supervised learning dimana hasil dari instance yang baru diklasifikasikan berdasarkan mayoritas dari kategori k-tetangga terdekat[13]. Pada k-NN, nilai k dapat memberikan pengaruh terhadap performa klasifikasi yang dihasilkan jika nilai k terlalu kecil[14]. Rumus k-NN menggunakan persamaan sebagai berikut:

$$dis(x_1, x_2) = \sqrt{\sum_{i=0}^{n} (x_{1i} - x_{2i})^2} \quad (3)$$

ISSN: 2598-0076

MCC

MCC sebagai memperhitungkan positif maupun negatif dan umumnya dianggap sebagai ukuran seimbang yang dapat digunakan bahkan jika kelas memiliki ukuran yang sangat berbeda[15]. Persamaannya adalah sebagai berikut:

$$MCC = \frac{TN \times TP - FP \times FN}{\sqrt{(TN + FN)(FP + TP)(TN + FP)(FN + TP)}} \quad (4)$$

Gmean

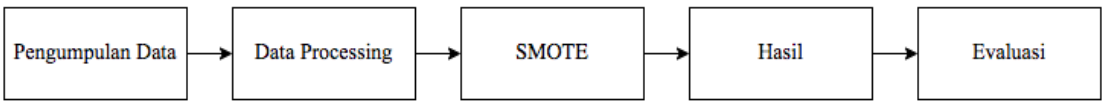
Gmean sebagai nilai rata-rata yang diperoleh dengan mengalikan semua data dalam suatu kelompok sampel[16]. Persamaannya adalah sebagai berikut:

$$G_{mean} = \sqrt{\frac{TP}{TP + FN} \times \frac{TN}{TN + FP}} \quad (5)$$

HASIL DAN PEMBAHASAN

Desain Penelitian

Pada penelitian ini, data yang digunakan adalah data tidak seimbang dari kumpulan repository KEEL. Data tidak seimbang ini tersebut akan diolah menggunakan metode SMOTE. Dalam penelitian ini akan dilakukan beberapa tahap seperti yang digambarkan pada Gambar 2.



Gambar 2. Desain Penelitian

Pengumpulan Data

Teknik pengumpulan data public yang dilakukan dengan mempersiapkan data tidak seimbang dari kumpulan repository KEEL berjumlah 5 dataset Binary, kemudian IR merupakan jumlah dari imbalance ratio tiap masing-masing data tidak seimbang, instance merupakan jumlah keseluruhan data tidak seimbang dan attribute merupakan jumlah atribut dari data tidak seimbang. Data yang telah dikumpulkan dari data public pada Tabel 1 dibawah ini sebagai berikut:

Tabel 1. Dataset Tidak Seimbang

Dataset	IR (Imbalance Ratio)	Instance	Attribute
ecoli4	15.8	336	7
pima	1.87	768	8
new-thyroid1	5.14	215	5
vehicle0	3.25	846	18
wisconsin	1.86	220	7

ISSN: 2598-0076

Hasil Eksperimen Dataset Binary

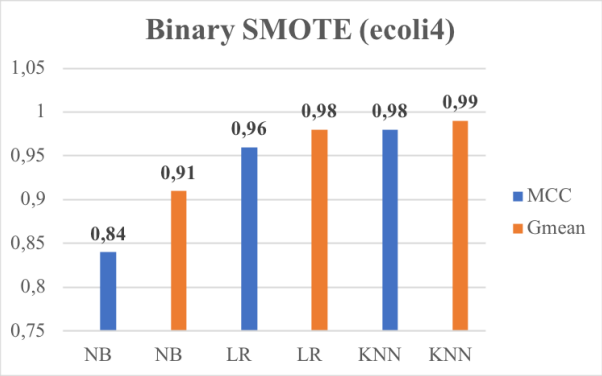
Pengujian dilakukan dengan 5 dataset Binary menggunakan SMOTE dan Metode Klasifikasi kemudian menggunakan perangkat lunak yang digunakan dalam bahasa Python. Hasil pengujian menggunakan dataset Binary dapat dilihat pada Tabel 2 sebagai berikut:

Tabel 2. Hasil Eksperimen Data Tidak Seimbang Binary

Dataset	Naïve Bayes		Logistic Regression		KNN	
	MCC	Gmean	MCC	Gmean	MCC	Gmean
ecoli4	0,84	0,91	0,96	0,98	0,98	0,99
pima	0,41	0,70	0,42	0,71	0,69	0,84
new-thyroid1	0,95	0,97	0,99	0,99	0,96	0,98
vehicle0	0,57	0,94	0,94	0,97	0,92	0,96
wisconsin	0,93	0,96	0,94	0,97	0,97	0,98

Binary SMOTE (ecoli4)

Pengujian dilakukan dengan dataset ecoli4 menggunakan SMOTE dan Metode Klasifikasi kemudian menggunakan perangkat lunak yang digunakan dalam bahasa Python. Hasil pengujian menggunakan dataset ecoli4 dapat dilihat pada Tabel 3 sebagai berikut:

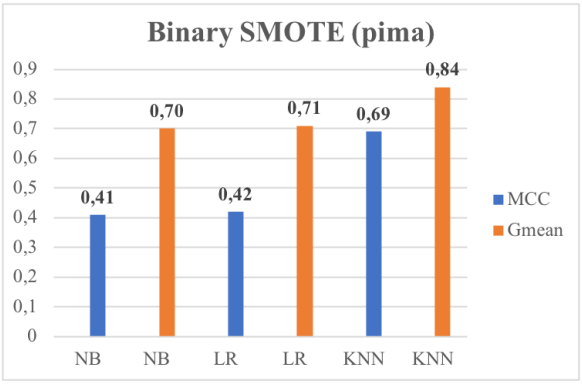


Gambar 3. Binary SMOTE (ecoli4)

Binary SMOTE (pima)

Pengujian dilakukan dengan dataset pima menggunakan SMOTE dan Metode Klasifikasi kemudian menggunakan perangkat lunak yang digunakan dalam bahasa Python. Hasil pengujian menggunakan dataset pima dapat dilihat pada Tabel 4 sebagai berikut:

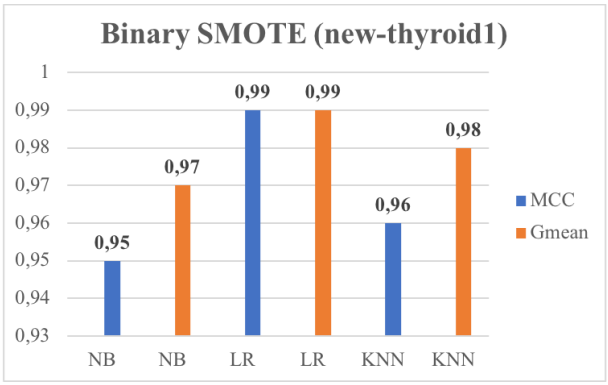
ISSN: 2598-0076



Gambar 4. Binary SMOTE (pima)

Binary SMOTE (new-thyroid1)

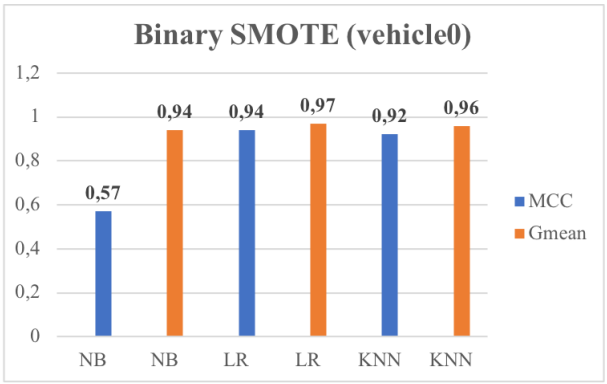
Pengujian dilakukan dengan dataset new-thyroid1 menggunakan SMOTE dan Metode Klasifikasi kemudian menggunakan perangkat lunak yang digunakan dalam bahasa Python. Hasil pengujian menggunakan dataset new-thyroid1 dapat dilihat pada Tabel 5 sebagai berikut:



Gambar 5. Binary SMOTE (new-thyroid1)

Binary SMOTE (vehicle0)

Pengujian dilakukan dengan dataset vehicle0 menggunakan SMOTE dan Metode Klasifikasi kemudian menggunakan perangkat lunak yang digunakan dalam bahasa Python. Hasil pengujian menggunakan dataset vehicle0 dapat dilihat pada Tabel 6 sebagai berikut:

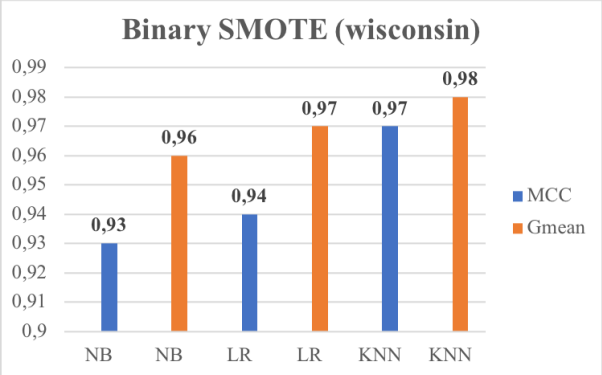


Gambar 6. Binary SMOTE (vehicle0)

ISSN: 2598-0076

Binary SMOTE (wisconsin)

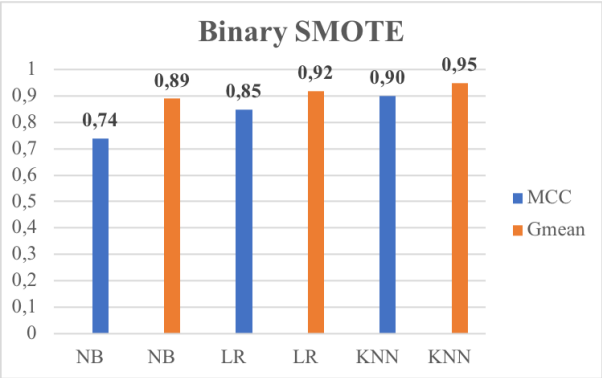
Pengujian dilakukan dengan dataset wisconsin menggunakan SMOTE dan Metode Klasifikasi kemudian menggunakan perangkat lunak yang digunakan dalam bahasa Python. Hasil pengujian menggunakan dataset wisconsin dapat dilihat pada Tabel 7 sebagai berikut:



Gambar 7. Binary SMOTE (wisconsin)

Average Binary SMOTE

Pengujian average dilakukan dengan 5 dataset Binary menggunakan SMOTE dan Metode Klasifikasi kemudian menggunakan perangkat lunak yang digunakan dalam bahasa Python. Hasil pengujian menggunakan 5 dataset Binary dapat dilihat pada Tabel 8 sebagai berikut:



Gambar 8. Average Binary SMOTE

SIMPULAN

Berdasarkan hasil eksperimen yang telah dilakukan, maka penulis menyimpulkan penelitian ini yaitu bahwa pengujian metode SMOTE dengan metode klasifikasi mampu menangani jumlah kelas mayoritas (negatif) dan kelas minoritas (positif) pada data tidak seimbang dengan menghasilkan nilai MCC dan Gmean mencapai kinerja prediksi yang lebih besar. Kinerja dari metode diukur dari nilai MCC dan Gmean dari masing-masing data tidak seimbang, namun untuk dataset Binary nilai MCC dan Gmean yang paling tinggi menggunakan metode SMOTE + KKN dengan nilai MCC = 0,90 dan nilai Gmean = 0,95 juga dapat mencapai kinerja prediksi yang lebih besar. Penanganan distribusi kelas yang tidak seimbang pada dataset Binary menggunakan metode

ISSN: 2598-0076

SMOTE + KKN dapat meningkatkan nilai akurasi MCC maupun Gmean. Hal tersebut menunjukkan bahwa proses penanganan terhadap distribusi kelas yang tidak seimbang pada tahap preprocessing data memberikan pengaruh terhadap nilai akurasi MCC maupun Gmean metode SMOTE + KKN.

DAFTAR RUJUKAN

[1] T. M. Alam *et al.*, "An investigation of credit card default prediction in the imbalanced datasets," *IEEE Access*, vol. 8, pp. 201173–201198, 2020, doi: 10.1109/ACCESS.2020.3033784.

[2] A. S. Tarawneh, A. B. A. Hassanat, K. Almohammadi, D. Chetverikov, and C. Bellinger, "SMOTEFUNA: Synthetic Minority Over-Sampling Technique Based on Furthest Neighbour Algorithm," *IEEE Access*, vol. 8, pp. 59069–59082, 2020, doi: 10.1109/ACCESS.2020.2983003.

[3] N. Liu, X. Li, E. Qi, M. Xu, L. Li, and B. Gao, "A novel ensemble learning paradigm for medical diagnosis with imbalanced data," *IEEE Access*, vol. 8, pp. 171263–171280, 2020, doi: 10.1109/ACCESS.2020.3014362.

[4] S. Maldonado, J. López, and C. Vairetti, "An alternative SMOTE oversampling strategy for high-dimensional datasets," *Appl. Soft Comput. J.*, vol. 76, pp. 380–389, 2019, doi: 10.1016/j.asoc.2018.12.024.

[5] W. Feng, W. Huang, and W. Bao, "Imbalanced Hyperspectral Image Classification with an Adaptive Ensemble Method Based on SMOTE and Rotation Forest with Differentiated Sampling Rates," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 12, pp. 1879–1883, 2019, doi: 10.1109/LGRS.2019.2913387.

[6] X. Xu, W. Chen, and Y. Sun, "Over-sampling algorithm for imbalanced data classification," *J. Syst. Eng. Electron.*, vol. 30, no. 6, pp. 1182–1191, 2019, doi: 10.21629/JSEE.2019.06.12.

[7] Y. Yan, R. Liu, Z. Ding, X. Du, J. Chen, and Y. Zhang, "A parameter-free cleaning method for SMOTE in imbalanced classification," *IEEE Access*, vol. 7, pp. 23537–23548, 2019, doi: 10.1109/ACCESS.2019.2899467.

[8] A. Fernández, S. García, F. Herrera, and N. V. Chawla, "SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary," *J. Artif. Intell. Res.*, vol. 61, pp. 863–905, 2018, doi: 10.1613/jair.1.11192.

[9] G. Douzas, F. Bacao, and F. Last, "Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE," *Inf. Sci. (Ny).*, vol. 465, pp. 1–20, 2018, doi: 10.1016/j.ins.2018.06.056.

[10] J. Mathew, C. K. Pang, M. Luo, and W. H. Leong, "Classification of Imbalanced Data by Oversampling in Kernel Space of Support Vector Machines," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 29, no. 9, pp. 4065–4076, 2018, doi: 10.1109/TNNLS.2017.2751612.

ISSN: 2598-0076

[11] V. P. Ramadhan and F. Y. Pamuji, "Jurnal Teknologi dan Manajemen Informatika Analisis Perbandingan Algoritma Forecasting dalam Prediksi Harga Saham LQ45 PT Bank Mandiri Sekuritas (BMRI)," vol. 8, no. 1, pp. 39–45, 2022.

[12] D. Dablain, B. Krawczyk, and N. V. Chawla, "DeepSMOTE: Fusing Deep Learning and SMOTE for Imbalanced Data," pp. 1–14, 2021, [Online]. Available: http://arxiv.org/abs/2105.02340

[13] F. Y. Pamuji and V. P. Ramadhan, "Komparasi Algoritma Random Forest dan Decision Tree untuk Memprediksi Keberhasilan Immunotheraphy," *J. Teknol. dan Manaj. Inform.*, vol. 7, no. 1, pp. 46–50, 2021, doi: 10.26905/jtmi.v7i1.5982.

[14] R. Das, S. K. Biswas, D. Devi, and B. Sarma, "An Oversampling Technique by Integrating Reverse Nearest Neighbor in SMOTE: Reverse-SMOTE," *Proc. - Int. Conf. Smart Electron. Commun. ICOSEC 2020*, no. Icosec, pp. 1239–1244, 2020, doi: 10.1109/ICOSEC49089.2020.9215387.

[15] F. Y. Pamuji and M. A. Soeleman, "Improved number detection for low resolution image using the canny algorithm," *Proc. - 2020 Int. Semin. Appl. Technol. Inf. Commun. IT Challenges Sustain. Scalability, Secur. Age Digit. Disruption, iSemantic 2020*, pp. 638–642, 2020, doi: 10.1109/iSemantic50169.2020.9234190.

[16] T. Lee, M. Kim, and S. P. Kim, "Data augmentation effects using borderline-SMOTE on classification of a P300-based BCI," *8th Int. Winter Conf. Brain-Computer Interface, BCI 2020*, pp. 9–12, 2020, doi: 10.1109/BCI48061.2020.9061656.

ISSN: 2598-0076