Komparasi Metode SMOTE dan ADASYN Untuk Penanganan Data Tidak Seimbang MultiClass

by Fandi Yulian Pamuji

Submission date: 21-Sep-2023 02:17PM (UTC+0700)
Submission ID: 2172406322
File name: dan_ADASYN_Untuk_Penanganan_Data_Tidak_Seimbang_MultiClass.docx (104.38K)
Word count: 3789
Character count: 23593

Komparasi Metode SMOTE dan ADASYN Untuk Penanganan Data Tidak Seimbang MultiClass

Fandi Yulian Pamuji[1], Sephia Dwi Arma Putri[2]

[1,2]Fakultas Teknologi Informasi, Unmer Malang, Indonesia
[1]fandi.pamuji@unmer.ac.id. [2]sephia.putri@student.unmer.ac.id

Abstrak

Data Mining merupakan kegiatan yang menggabungkan berbagai cabang ilmu pengetahuan menjadi satu terdiri dari sistem basis data, statistika, machine learning dan visualization untuk menganalisis sebuah dataset yang besar guna mendapatkan karakteristik data yang bermanfaat. Untuk mengatasi permasalahan dataset tidak seimbang adalah dengan menyeimbangkan distribusi kelas tidak seragam di antara kelas-kelas dengan komparasi menggunakan metode SMOTE dan ADASYN supaya jumlahnya seimbang dari kelas mayoritas (negatif) maupun kelas minoritas (positif). Berdasarkan hasil eksperimen yang telah dilakukan dari 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 dibandingkan dengan menggunakan metode klasifikasi saja maupun menggunakan Metode ADASYN. Kemudian untuk dataset MultiClass nilai MCC dan Gmean yang paling tinggi menggunakan SMOTE + KNN dengan nilai tertinggi MCC = 0,64 dan nilai Gmean = 0,74 dari nilai MCC dan Gmean diatas akurasinya sudah bagus karena nilai sudah mencapai diatas 0,1 termasuk datanya sudah seimbang dengan menggunakan Metode SMOTE + KNN dapat mencapai kinerja prediksi yang lebih besar untuk menangani dataset tidak seimbang MultiClass. 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 + KNN.

Kata kunci : Data Mining, Data Tidak Seimbang, SMOTE, ADASYN, MultiClass

1. Pendahuluan

Data Mining merupakan kegiatan yang menggabungkan berbagai cabang ilmu pengetahuan menjadi satu terdiri dari sistem basis data, statistika, machine learning dan visualization untuk menganalisis sebuah dataset yang besar guna mendapatkan karakteristik data yang bermanfaat((SENASIF) & 2022, 2022). Machine Learning sangat beraneka ragam, seperti penyaringan spam email, computer vision, dan big data dimana program yang semacam ini akan sulit dikembangan dengan pemrograman eksplisit(Pamuji & Soeleman, 2020).

Banyak permasalahan data mining melibatkan imbalanced data. Dataset tidak seimbang kelas ini terjadi karena rasio yang tidak seimbang antara kasus yang satu dengan kasus yang lainnya(Pamuji & Ramadhan, 2021). Ketidakseimbangan kelas ini akan merugikan pada penelitian bidang data mining karena machine learning memiliki kesulitan dalam mengklasifikasikan kelas minoritas (jumlah instance yang kecil) dengan benar(V. P. Ramadhan & Pamuji, 2022). Beberapa algoritma mengasumsikan bahwa distribusi kelas yang diuji adalah seimbang sehingga dalam beberapa kasus menjadikan kesalahan dalam mengklasifikasikan hasil pada tiap kelas(Fernández et al., 2018).

Sedangkan over-sampling merupakan metode penyeimbangan distribusi kelas dengan mereplikasi instance pada kelas minoritas secara acak. Over-sampling meningkatkan kemungkinan munculnya overfitting karena menduplikasi instance secara sama persis mengajukan metode untuk pendekatan sampling pada pembelajaran dengan dataset tidak seimbang(Maldonado et al., 2019). Ide utama dari ADASYN merupakan menggunakan bobot distribusi untuk data pada kelas minoritas berdasarkan pada tingkat kesulitan belajar, dimana data sintesis dihasilkan dari kelas minoritas yang susah untuk belajar dibandingkan dengan data minoritas yang lebih mudah untuk belajar(Abdoh et al., 2018).

Metode SMOTE dan ADASYN merupakan metode untuk mengatasi permasalahan dataset tidak seimbang Binary maupun MultiClass, SMOTE merupakan kombinasi dari mayoritas under-sampling dan mayoritas over-sampling. Bagian under-sampling hanyalah prosedur under-sampling umum(N. G. Ramadhan, 2021). Untuk bagian over-sampling dari SMOTE, sampel sintetis dibuat secara acak dengan menambahkan selisih bobot antara sampel ke-i dan k tetangga terdekatnya. ADASYN menghasilkan pengamatan sintetik sepanjang garis

lurus antara pengamatan kelas minoritas dan k-tetangga kelas minoritas terdekatnya(Gameng et al., 2019).

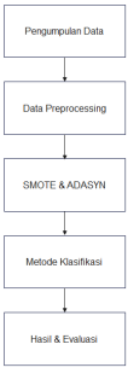
Untuk mengatasi permasalahan dataset tidak seimbang adalah dengan menyeimbangkan distribusi kelas tidak seragam di antara kelas-kelas dengan komparasi menggunakan metode SMOTE dan ADASYN supaya jumlahnya seimbang dari kelas mayoritas (negatif) maupun kelas minoritas (positif)(Feng et al., 2019). Metode klasifikasi yang digunakan untuk mencari nilai akurasi yang menggunakan machine learning, kemudian dataset tersebut diuji dan di evaluasi yang akan dipilih metode yang menghasilkan nilai akurasi paling tinggi(Jonathan et al., 2020).

Dari uraian permasalahan penelitian ini adalah bagaimana menangani data tidak seimbang MultiClass pada data public dari kumpulan data repository KEEL dengan komparasi metode SMOTE dan ADASYN dengan menggunakan metode klasifikasi. Penelitian ini bertujuan untuk mempertahankan jumlah kelas yang tidak seimbang agar tetap ideal dengan melakukan imputasi pada dataset yang tidak seimbang MultiClass dengan menguji metode SMOTE dan ADASYN(Mohammed et al., 2020).

2. Metode Penelitian

2.1 Alur Penelitian

Pada penelitian ini, data yang digunakan adalah data tidak seimbang MultiClass dari kumpulan repository KEEL. Data tidak seimbang MultiClass ini tersebut akan diolah menggunakan metode SMOTE dan ADASYN(Polat, 2019). Dalam penelitian ini akan dilakukan beberapa tahap seperti yang ada pada Gambar 1 dibawah ini.



Gambar 1. Alur Penelitian

2.2 Pengumpulan Data

Teknik pengumpulan data public yang dilakukan dengan mempersiapkan data tidak seimbang dari kumpulan repository KEEL berjumlah 5 dataset MultiClass yang terdiri dari 3 Class, 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(Gu et al., 2019). Data yang telah dikumpulkan dari data public pada Tabel 1 dibawah ini sebagai berikut:

Tabel 1. *Dataset MultiClass* Tidak Seimbang

Dataset	IR (Imbalance Ratio)	Instance	Atribute
balance	5.88	625	4
thyroid-new	5.0	215	5
wine	1.5	178	13
hayes-roth	1.7	132	4
cmc	12.33	443	7

Dari tabel 1 merupakan dataset yang terdiri dari 3 Class disetiap masing-masing dataset dan IR (Imbalance Ratio) berbeda-beda tergantung tingkat tidak seimbangan dari datasetnya. Untuk dataset balance merupakan dataset tingkat Imbalance Ratio rendah dengan nilai 5.88, dataset thyroid-new merupakan dataset tingkat Imbalance Ratio rendah dengan nilai 5.0, dataset wine merupakan dataset tingkat Imbalance Ratio rendah dengan nilai 1.5, dataset hayes-roth merupakan dataset tingkat Imbalance Ratio rendah dengan nilai 1.7, dan yang terakhir dataset cmc merupakan dataset tingkat Imbalance Ratio tinggi dengan nilai 12.33.

2.3 Data Preprocessing

Data transformation untuk mengubah dataset dalam bentuk yang sesuai dalam proses data mining. Normalization dilakukan untuk menskalakan nilai class dalam rentang class dalam rentang nilai 1, 2 dan 3 untuk dataset multi class. Selanjutnya dataset masing-masing di split 80% data training dan 20% data testing sebelum dieksekusi ke metode SMOTE dan ADASYN(Li et al., 2016).

2.4 SMOTE

Metode 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(Blagus & Lusa, 2012). Contoh sintetis ini akan memungkinkan `pengklasifikasi yang digunakan untuk membangun batas keputusan yang lebih umum dan dengan demikian mengurangi efek overfitting(Lin et al.,

2021). 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 dataset(Jonathan et al., 2020).

$$X_{syn} = X_i + (X_{knn} - X_i) \times t \quad (1)$$

Pertama, mengidentifikasi vektor X_i dan mengidentifikasi K-nearest neighbors X_{knn}, kemudian menghitung perbedaan antara vektor fitur dan K-nearest neighbors, selanjutnya mengalikan perbedaan dengan angka acak antara 0 dan 1, kemudian menambahkan nomor keluaran ke vektor fitur untuk mengidentifikasi titik baru pada ruas garis dan terakhir mengulangi proses dari 1 hingga 4 untuk mengidentifikasi vektor fitur.

2.5 ADASYN

Metode ADASYN merupakan pendekatan sampling pada pembelajaran dengan dataset yang tidak seimbang. Ide utama dari ADASYN menggunakan bobot distribusi untuk data pada kelas minoritas berdasarkan pada tingkat kesulitan belajar, sehingga data sintesis dihasilkan dari kelas minoritas yang susah untuk belajar dibandingkan dengan data minoritas yang lebih mudah untuk belajar(Wang et al., 2019). ADASYN meningkatkan pembelajaran dengan dua cara. Pertama, mengurangi bias yang diakibatkan oleh ketidakseimbangan kelas dan yang kedua secara adaptif menggeser batas keputusan klasifikasi terhadap kesulitan data(Skryjomski & Krawczyk, 2017).

$$X_{new} = X_i + rand(0,1) * (X_i - X_1) \quad (2)$$

2.6 Metode Klasifikasi

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(Barus & Sanjaya, 2020). 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:

$$p_i = \frac{1}{1+e^{-(w^T x_i + b)}} \quad (3)$$

k-Nearest Neighbor (k-NN) adalah algoritma supervised learning dimana hasil dari instance yang baru diklasifikasikan berdasarkan mayoritas dari kategori k-tetangga terdekat. Metode k-NN menggunakan prinsip ketetanggaan (neighbor) untuk memprediksi kelas yang baru(Cahyanti et al., 2021). Jumlah tetangga yang dipakai adalah sebanyak k-tetanggaan. Setelah mengambil k tetangga terdekat pertama kemudian dihitung jumlah data yang mengikuti kelas yang ada dari k tetangga tersebut. Kelas dengan data terbanyak yang mengikutinya menjadi kelas pemenang yang diberikan sebagai label kelas pada data X. Pada k-NN, nilai k dapat memberikan pengaruh terhadap performa klasifikasi yang dihasilkan jika nilai k terlalu kecil(Arifin & Syalwah, 2020). Rumus k-NN menggunakan persamaan sebagai berikut:

$$dis(x_1, x_2) = \sqrt{\sum_{i=0}^{n}(x_{1i} - x_{2i})^2} \quad (4)$$

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. Teorema tersebut dikombinasikan dengan Naive dimana diasumsikan kondisi antar atribut saling bebas. Klasifikasi Naive Bayes diasumsikan bahwa ada atau tidak ciri tertentu dari sebuah kelas tidak ada hubungannya dengan ciri dari kelas lainnya(Supriyatna & Pustika, 2018). Rumus Naïve Bayes menggunakan persamaan sebagai berikut:

$$P(C|X) = \frac{P(X|C)P(c)}{P(x)} \quad (5)$$

2.7 Hasil dan Evaluasi

Pada tahapan ini dilakukan evaluasi terhadap tingkat akurasi dari masing-masing metode untuk melihat kinerja setiap metode yang digunakan. Pada penelitian ini metode SMOTE dan ADASYN akan dievaluasi dan divalidasi menggunakan alat ukur Confusion Matrix, MCC dan Gmean(Gameng et al., 2019).

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

$$MCC = \frac{TN \times TP - FP \times FN}{\sqrt{(TN+FN)(FP+TP)(TN+FP)(FN+TP)}} \quad (6)$$

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

$$G_{mean} = \sqrt{\frac{TP}{TP+FN} \times \frac{TN}{TN+FP}} \quad (7)$$

3. Hasil dan Pembahasan

4.

Tahapan proses data tidak seimbang menggunakan metode klasifikasi dataset MultiClass yang pertama proses data preprocessing dengan normalization dilakukan untuk menskalakan nilai class dalam rentang nilai 1, 2 dan 3 untuk dataset MultiClass, kemudian dataset MultiClass akan di proses menggunakan metode klasifikasi yang digunakan antara lain Logistic Regression, KNN dan Naïve Bayes. Dari hasil metode klasifikasi dataset MultiClass tersebut akan diambil nilai MCC dan Gmean untuk di analisa nantinya dengan hasil nilai MCC dan Gmean dataset MultiClass menggunakan SMOTE dan ADASYN.

MultiClass SMOTE dan ADASYN (*balance*)

Evaluasi pada penelitian ini dari hasil nilai MCC dan Gmean dari dataset MultiClass (balance) menggunakan SMOTE dan ADASYN pada Tabel 2 sebagai berikut:

Tabel 2. *MultiClass* SMOTE & ADASYN (*balance*)

Metode	Klasifikasi		ADASYN		SMOTE	
	MCC	Gmean	MCC	Gmean	MCC	Gmean
Logistic Regression	0.69	0.00	0.55	0.69	0.57	0.71
KNN (10)	0.76	0.00	0.50	0.58	0.50	0.58
Naive Bayes	0.83	0.00	0.77	0.82	0.73	0.80

Dari Tabel 2 untuk hasil tertinggi dari nilai MCC dan Gmean pada dataset MultiClass (balance) menggunakan SMOTE dan ADASYN adalah metode Klasifikasi Naïve Bayes dengan nilai tertinggi MCC = 0,83 dan nilai Gmean = 0,00 dari metode yang lain seperti Logistic Regression dan KNN. Kemudian menggunakan metode ADASYN dengan metode Klasifikasi Naïve Bayes nilai tertinggi MCC = 0,77 dan nilai Gmean = 0,82 dari metode yang lain seperti Logistic Regression dan KNN dan menggunakan metode SMOTE dengan metode Klasifikasi Naïve Bayes nilai tertinggi MCC = 0,73 dan nilai Gmean = 0,80 dari metode yang lain seperti Logistic Regression dan KNN. Penanganan distribusi kelas yang tidak seimbang pada dataset MultiClass (balance) menggunakan ADASYN dapat meningkatkan nilai akurasi MCC maupun Gmean pada metode Naïve Bayes.

MultiClass SMOTE dan ADASYN (*thyroid-new*)

Evaluasi pada penelitian ini dari hasil nilai MCC dan Gmean dari dataset MultiClass (thyroid-new) menggunakan SMOTE dan ADASYN pada Tabel 3 sebagai berikut:

Tabel 3. *MultiClass* SMOTE & ADASYN (*thyroid-new*)

Metode	Klasifikasi		ADASYN		SMOTE	
	MCC	Gmean	MCC	Gmean	MCC	Gmean
Logistic Regression	0.76	0.74	0.90	0.93	0.92	0.94
KNN (10)	0.57	0.56	0.70	0.81	0.72	0.81
Naive Bayes	0.92	0.93	0.95	0.96	0.95	0.97

Dari Tabel 3 untuk hasil tertinggi dari nilai MCC dan Gmean pada dataset MultiClass (thyroid-new) menggunakan SMOTE dan ADASYN adalah metode Klasifikasi Naïve Bayes dengan nilai tertinggi MCC = 0,92 dan nilai Gmean = 0,93 dari metode yang lain seperti Logistic Regression dan KNN. Kemudian menggunakan metode ADASYN dengan metode Klasifikasi Naïve Bayes nilai tertinggi MCC = 0,95 dan nilai Gmean = 0,96 dari metode yang lain seperti Logistic Regression dan KNN dan menggunakan metode SMOTE dengan metode Klasifikasi Naïve Bayes nilai tertinggi MCC = 0,95 dan nilai Gmean = 0,97 dari metode yang lain seperti Logistic Regression dan KNN. Penanganan distribusi kelas yang tidak seimbang pada dataset MultiClass (thyroid-new) menggunakan SMOTE dapat meningkatkan nilai akurasi MCC maupun Gmean pada metode Naïve Bayes.

MultiClass SMOTE dan ADASYN (*wine*)

Evaluasi pada penelitian ini dari hasil nilai MCC dan Gmean dari dataset MultiClass (wine) menggunakan SMOTE dan ADASYN pada Tabel 4 sebagai berikut:

Tabel 4. *MultiClass* SMOTE & ADASYN (*wine*)

Metode	Klasifikasi		ADASYN		SMOTE	
	MCC	Gmean	MCC	Gmean	MCC	Gmean
Logistic Regression	0.71	0.80	0.73	0.82	0.71	0.80
KNN (10)	0.65	0.75	0.69	0.77	0.65	0.75
Naive Bayes	0.57	0.26	0.53	0.23	0.54	0.29

Dari Tabel 4 untuk hasil tertinggi dari nilai MCC dan Gmean pada dataset MultiClass (wine) menggunakan SMOTE dan ADASYN adalah metode Klasifikasi Logistic Regression dengan nilai tertinggi MCC = 0,71 dan nilai Gmean = 0,80 dari metode yang lain seperti Naive Bayes dan KNN. Kemudian menggunakan metode ADASYN dengan metode Klasifikasi Logistic Regression nilai tertinggi MCC = 0,73 dan nilai Gmean = 0,82 dari metode yang lain seperti Naive Bayes dan KNN dan menggunakan metode SMOTE dengan metode Klasifikasi Logistic Regression nilai tertinggi MCC = 0,71 dan nilai Gmean = 0,80 dari metode yang lain seperti Naive Bayes dan KNN. Penanganan distribusi kelas yang tidak seimbang pada dataset MultiClass (wine) menggunakan ADASYN dapat

meningkatkan nilai akurasi MCC maupun Gmean pada metode Logistic Regression.

***MultiClass* SMOTE dan ADASYN (*hayes-roth*)**

Evaluasi pada penelitian ini dari hasil nilai MCC dan Gmean dari dataset MultiClass (hayes-roth) menggunakan SMOTE dan ADASYN pada Tabel 5 sebagai berikut:

Tabel 5. *MultiClass* SMOTE & ADASYN (*hayes-roth*)

Metode	Klasifikasi		ADASYN		SMOTE	
	MCC	Gmean	MCC	Gmean	MCC	Gmean
Logistic Regression	0.28	0.55	0.34	0.51	0.37	0.53
KNN (10)	0.48	0.65	0.73	0.81	0.64	0.75
Naive Bayes	0.69	0.82	0.71	0.78	0.70	0.78

Dari Tabel 5 untuk hasil tertinggi dari nilai MCC dan Gmean pada dataset MultiClass (balance) menggunakan SMOTE dan ADASYN adalah metode Klasifikasi Naïve Bayes dengan nilai tertinggi MCC = 0,69 dan nilai Gmean = 0,82 dari metode yang lain seperti Logistic Regression dan KNN. Kemudian menggunakan metode ADASYN dengan metode Klasifikasi KNN nilai tertinggi MCC = 0,73 dan nilai Gmean = 0,81 dari metode yang lain seperti Logistic Regression dan Naive Bayes kemudian menggunakan metode SMOTE dengan metode Klasifikasi Naive Bayes nilai tertinggi MCC = 0,70 dan nilai Gmean = 0,78 dari metode yang lain seperti Logistic Regression dan KNN. Penanganan distribusi kelas yang tidak seimbang pada dataset MultiClass (balance) menggunakan ADASYN dapat meningkatkan nilai akurasi MCC maupun Gmean pada metode KNN.

MultiClass SMOTE dan ADASYN (cmc)

Evaluasi pada penelitian ini dari hasil nilai MCC dan Gmean dari dataset MultiClass (cmc) menggunakan SMOTE dan ADASYN pada Tabel 6 sebagai berikut:

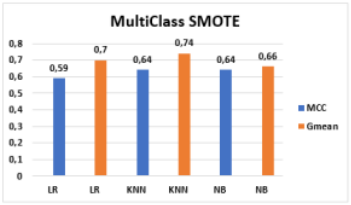
Tabel 6. *MultiClass* SMOTE & ADASYN (*cmc*)

Metode	Klasifikasi		ADASYN		SMOTE	
	MCC	Gmean	MCC	Gmean	MCC	Gmean
Logistic Regression	0.24	0.47	0.27	0.51	0.29	0.52
KNN (10)	0.42	0.59	0.53	0.68	0.52	0.67
Naive Bayes	0.23	0.48	0.24	0.46	0.28	0.48

Dari Tabel 6 untuk hasil tertinggi dari nilai MCC dan Gmean pada dataset MultiClass (balance) menggunakan SMOTE dan ADASYN adalah metode Klasifikasi KNN dengan nilai tertinggi MCC = 0,42 dan nilai Gmean = 0,59 dari metode yang lain seperti Logistic Regression dan Naive Bayes. Kemudian menggunakan metode ADASYN dengan metode Klasifikasi KNN nilai tertinggi MCC = 0,53 dan nilai Gmean = 0,68 dari metode yang lain seperti Logistic Regression dan Naive Bayes kemudian menggunakan metode SMOTE dengan metode Klasifikasi KNN nilai tertinggi MCC = 0,52 dan nilai Gmean = 0,67 dari metode yang lain seperti Logistic Regression dan Naive Bayes. Penanganan distribusi kelas yang tidak seimbang pada dataset MultiClass (cmc) menggunakan ADASYN dapat meningkatkan nilai akurasi MCC maupun Gmean pada metode KNN.

Tabel 7. *MultiClass* SMOTE

Dataset	Logistic Regression		KNN = 10		Naïve Bayes	
	MCC	Gmean	MCC	Gmean	MCC	Gmean
balance	0.57	0.71	0.50	0.58	0.73	0.80
thyroid-new	0.92	0.94	0.72	0.81	0.95	0.97
wine	0.71	0.80	0.72	0.81	0.54	0.29
hayes-roth	0.37	0.53	0.72	0.81	0.7	0.78
cmc	0.37	0.53	0.52	0.67	0.28	0.48
Average	0.59	0.70	0.64	0.74	0.64	0.66

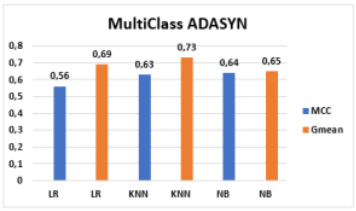


Gambar 2. *MultiClass* SMOTE

Dari Tabel 7 dan Gambar 2 untuk hasil tertinggi dari nilai MCC dan Gmean pada dataset MultiClass menggunakan SMOTE adalah metode Klasifikasi KNN dengan nilai tertinggi MCC = 0,64 dan nilai Gmean = 0,74 dari metode yang lain seperti Logistic Regression dan Naive Bayes. Penanganan distribusi kelas yang tidak seimbang pada dataset MultiClass menggunakan SMOTE dapat meningkatkan nilai akurasi MCC maupun Gmean pada metode KNN.

Tabel 8. *MultiClass* ADASYN

Dataset	Logistic Regression		KNN = 10		Naïve Bayes	
	MCC	Gmean	MCC	Gmean	MCC	Gmean
balance	0.55	0.69	0.50	0.58	0.77	0.82
thyroid-new	0.90	0.93	0.70	0.81	0.95	0.96
wine	0.73	0.82	0.69	0.77	0.53	0.23
hayes-roth	0.34	0.51	0.73	0.81	0.71	0.78
cmc	0.27	0.51	0.53	0.68	0.24	0.46
Average	0.56	0.69	0.63	0.73	0.64	0.65



Gambar 3. *MultiClass* ADASYN

Dari Tabel 8 dan Gambar 3 untuk hasil tertinggi dari nilai MCC dan Gmean pada dataset MultiClass menggunakan ADASYN adalah metode Klasifikasi KNN dengan nilai tertinggi MCC = 0,63 dan nilai Gmean = 0,73 dari metode yang lain seperti Logistic Regression dan Naive Bayes. Penanganan distribusi kelas yang tidak seimbang pada dataset MultiClass menggunakan ADASYN dapat meningkatkan nilai akurasi MCC maupun Gmean pada metode KNN.

5. Kesimpulan dan Saran

Berdasarkan hasil eksperimen yang telah dilakukan dari 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 dibandingkan dengan menggunakan metode klasifikasi saja maupun menggunakan Metode ADASYN. Kemudian untuk dataset MultiClass nilai MCC dan Gmean yang paling tinggi menggunakan SMOTE + KNN dengan nilai tertinggi MCC = 0,64 dan nilai Gmean = 0,74 dari nilai MCC dan Gmean diatas akurasinya sudah bagus karena nilai sudah mencapai diatas 0,1 termasuk datanya sudah seimbang dengan menggunakan Metode SMOTE + KNN dapat mencapai kinerja prediksi yang lebih besar untuk menangani dataset tidak seimbang MultiClass. 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 + KNN.

Daftar Pustaka:

(SENASIF), F. P.-S. N. S. I., & 2022, undefined. (2022). Pengujian Metode SMOTE Untuk Penanganan Data Tidak Seimbang Pada Dataset Binary. *Jurnalfti.Unmer.Ac.Id*, *2022*(September), 3200–3208. https://jurnalfti.unmer.ac.id/index.php/senasif/article/view/403

Abdoh, S. F., Abo Rizka, M., & Maghraby, F. A. (2018). Cervical cancer diagnosis using random forest classifier with SMOTE and feature reduction techniques. *IEEE Access*, *6*, 59475–59485. https://doi.org/10.1109/ACCESS.2018.2874063

Arifin, T., & Syalwah, S. (2020). Prediksi Keberhasilan Immunotherapy Pada Penyakit Kutil Dengan Menggunakan Algoritma Naïve Bayes. *Jurnal Responsif*, *2*(1), 38–43.

Barus, O. P., & Sanjaya, T. (2020). *Prediksi Tingkat Keberhasilan Pengobatan Kanker Menggunakan Imunoterapi Dengan Metode Naive Bayes*. *5*(1), 1–6.

Blagus, R., & Lusa, L. (2012). Evaluation of SMOTE for high-dimensional class-imbalanced microarray data. *Proceedings - 2012 11th International Conference on Machine Learning and Applications, ICMLA 2012*, *2*(1), 89–94. https://doi.org/10.1109/ICMLA.2012.183

Cahyanti, F. L. D., Gata, W., & Sarasati, F. (2021). Implementasi Algoritma Naïve Bayes dan K-Nearest Neighbor Dalam Menentukan Tingkat Keberhasilan Immunotherapy Untuk Pengobatan Penyakit Kanker Kulit. *Jurnal Ilmiah Universitas Batanghari Jambi*, *21*(1), 259. https://doi.org/10.33087/jiubj.v21i1.1189

Feng, W., Huang, W., & Bao, W. (2019). Imbalanced Hyperspectral Image Classification with an Adaptive Ensemble Method Based on SMOTE and Rotation Forest with Differentiated Sampling Rates. *IEEE Geoscience and Remote Sensing Letters*, *16*(12), 1879–1883. https://doi.org/10.1109/LGRS.2019.2913387

Fernández, A., García, 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. https://doi.org/10.1613/jair.1.11192

Gameng, H. A., Gerardo, B. B., & Medina, R. P. (2019). Modified Adaptive Synthetic SMOTE to Improve Classification Performance in Imbalanced Datasets. *ICETAS 2019 - 2019 6th IEEE International Conference on Engineering, Technologies and Applied Sciences*, 19–23. https://doi.org/10.1109/ICETAS48360.2019.9117287

Gu, Q., Wang, X. M., Wu, Z., Ning, B., & Xin, C. S. (2016). An improved SMOTE algorithm based on genetic algorithm for imbalanced data classification. *Journal of Digital Information Management*, *14*(2), 92–103.

Jonathan, B., Putra, P. H., Ruldeviyani, Y., Network, F. D., & Indonesia, U. (2020). *Observation Imbalanced Data Text to Predict Users Selling Products on Female Daily with SMOTE ,*. 81–85.

Li, J., Fong, S., & Zhuang, Y. (2016). Optimizing SMOTE by Metaheuristics with Neural Network and Decision Tree. *Proceedings -*

2015 3rd International Symposium on Computational and Business Intelligence, ISCBI 2015, 26–32. https://doi.org/10.1109/ISCBI.2015.12

Lin, M., Zhu, X., Hua, T., Tang, X., Tu, G., & Chen, X. (2021). *Detection of Ionospheric Scintillation Based on XGBoost Model Improved by SMOTE-ENN Technique*. 1–22.

Maldonado, S., López, J., & Vairetti, C. (2019). An alternative SMOTE oversampling strategy for high-dimensional datasets. *Applied Soft Computing Journal*, *76*, 380–389. https://doi.org/10.1016/j.asoc.2018.12.024

Mohammed, A. J., Hassan, M. M., & Kadir, D. H. (2020). Improving classification performance for a novel imbalanced medical dataset using smote method. *International Journal of Advanced Trends in Computer Science and Engineering*, *9*(3), 3161–3172. https://doi.org/10.30534/ijatcse/2020/104932020

Pamuji, F. Y., & Ramadhan, V. P. (2021). Komparasi Algoritma Random Forest dan Decision Tree untuk Memprediksi Keberhasilan Immunotheraphy. *Jurnal Teknologi Dan Manajemen Informatika*, *7*(1), 46–50. https://doi.org/10.26905/jtmi.v7i1.5982

Pamuji, F. Y., & Soeleman, M. A. (2020). Improved number detection for low resolution image using the canny algorithm. *Proceedings - 2020 International Seminar on Application for Technology of Information and Communication: IT Challenges for Sustainability, Scalability, and Security in the Age of Digital Disruption, ISemantic 2020*, 638–642. https://doi.org/10.1109/iSemantic50169.2020.9234190

Polat, K. (2019). A Hybrid Approach to Parkinson Disease Classification using speech signal : The combination of SMOTE and Random Forests. *2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT)*, 1–3. https://doi.org/10.1109/EBBT.2019.8741725

Ramadhan, N. G. (2021). Comparative Analysis of ADASYN-SVM and SMOTE-SVM Methods on the Detection of Type 2 Diabetes Mellitus. *Scientific Journal of Informatics*, *8*(2), 276–282. https://doi.org/10.15294/sji.v8i2.32484

Ramadhan, V. P., & Pamuji, F. Y. (2022). *Jurnal Teknologi dan Manajemen Informatika Analisis Perbandingan Algoritma Forecasting dalam Prediksi Harga Saham LQ45 PT Bank Mandiri Sekuritas (BMRI)*. *8*(1), 39–45.

Skryjomski, P., & Krawczyk, B. (2017). Influence of minority class instance types on SMOTE imbalanced data oversampling. *Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications*, *74*, 7–21. http://proceedings.mlr.press/v74/skryjomski17a.html

Supriyatna, A., & Mustika, W. P. (2018). Komparasi Algoritma Naive bayes dan SVM Untuk Memprediksi Keberhasilan Imunoterapi Pada Penyakit Kutil. *J-SAKTI (Jurnal Sains Komputer Dan Informatika)*, *2*(2), 152. https://doi.org/10.30645/j-sakti.v2i2.78

Wang, X., Xu, P., Yang, Q., Wu, G., & Wei, F. (2019). Fault Prediction Method of Access Control Terminal Based on Euclidean Distance Center SMOTE Method. *Proceedings of 2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems, CCIS 2018*, *6027*, 84–89. https://doi.org/10.1109/CCIS.2018.8691196

Komparasi Metode SMOTE dan ADASYN Untuk Penanganan Data Tidak Seimbang MultiClass

ORIGINALITY REPORT

15%	14%	6%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	ejournal.medan.uph.edu Internet Source	2%
2	ejournal.itats.ac.id Internet Source	2%
3	Submitted to Universitas Airlangga Student Paper	2%
4	Submitted to Universitas Brawijaya Student Paper	2%
5	janitra.org Internet Source	1%
6	lataviaroberson.com Internet Source	1%
7	www.dqlab.id Internet Source	1%
8	repository.bsi.ac.id Internet Source	1%
9	ichi.pro Internet Source	1%

10	www.scilit.net Internet Source	<1 %
11	edoc.pub Internet Source	<1 %
12	erepo.unud.ac.id Internet Source	<1 %
13	www.mdpi.com Internet Source	<1 %
14	Nurlaelatul Maulidah, Riki Supriyadi, Dwi Yuni Utami, Fuad Nur Hasan, Ahmad Fauzi, Ade Christian. "Prediksi Penyakit Diabetes Melitus Menggunakan Metode Support Vector Machine dan Naive Bayes", Indonesian Journal on Software Engineering (IJSE), 2021 Publication	<1 %
15	repo.itera.ac.id Internet Source	<1 %
16	journal.universitasbumigora.ac.id Internet Source	<1 %
17	kc.umn.ac.id Internet Source	<1 %
18	sisfotenika.stmikpontianak.ac.id Internet Source	<1 %
19	jurnalmahasiswa.stiesia.ac.id Internet Source	<1 %

20	www.scribd.com Internet Source	<1%
21	123dok.com Internet Source	<1%
22	ejurnal.stmik-budidarma.ac.id Internet Source	<1%
23	jurnal.univbinainsan.ac.id Internet Source	<1%
24	stt-pln.e-journal.id Internet Source	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		

Komparasi Metode SMOTE dan ADASYN Untuk Penanganan Data Tidak Seimbang MultiClass

GRADEMARK REPORT

FINAL GRADE

/0

GENERAL COMMENTS

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7