CENTIVE
VOL. 3, NO. 1, PP.001-010, DEC 2023
ISBN : 978-602-53004-0-0

Conference on Electrical Engineering, Informatics, Industrial Technology, and Creative Media 2023

ANALISIS KETIDAKSEIMBANGAN KELAS DALAM PENGEMBANGAN MODEL KLASIFIKASI

Muhammad Reza Redo[1], Andreas Perdana[*2]

[1,2] *STMIK Dharma Wacana, Metro, Indonesia*

[1] reza.redo@hotmail.com
[2] andreas.perdana@dharmawacana.ac.id

Received on 30-10-2023, revised on 07-11-2023, accepted on 15-11-2023

Absrak

Penelitian ini bertujuan untuk mengkaji dan mengatasi masalah ketidakseimbangan kelas dalam pembelajaran mesin, sebuah tantangan umum yang sering menghambat efektivitas model. Dengan menggunakan dataset yang memiliki distribusi kelas yang tidak seimbang, penelitian ini mengusulkan pendekatan unik dengan menggabungkan kelas-kelas minoritas. Ketidakseimbangan kelas atau imbalance class merupakan masalah yang sering terjadi dalam proses pembelajaran mesin, ada banyak pendekatan atau metode yang digunakan untuk menyelesaikan permasalahan tersebut, pada penelitian ini terdapat sebuah dataset memiliki kelas yang tidak seimbang, dimana untuk mengatasi ketidakseimbangan kelas tersebut penelitian ini melakukan pendekatan dengan cara menggabungkan kelas-kelas minoritas. Nantinya kelas-kelas minoritas tersebut akan di evaluasi dengan model pembelajaran mesin yang menggunakan delapan algoritma berbeda yaitu Algoritma AdaBoost, Algoritma Gradient Boosting, Algoritma kNN, Algoritma Naïve Bayes, Algoritma Neural Network, Algoritma Random Forest, Algoritma SVM, dan Algoritma Decision Tree. Selain di evaluasi dengan delapan algoritma, data tersebut juga akan diterapkan pendekatan oversampling dan undersampling. Dari eksperimen tersebut diharapkan kita dapat melihat hasil evaluasi dari model pembelajaran mesin. Eksperimen ini diharapkan dapat memberikan wawasan tentang efektivitas metode oversampling dan undersampling dalam meningkatkan kinerja model pada dataset dengan ketidakseimbangan kelas. Hasil dari eksperimen ini akan memberikan gambaran yang lebih jelas tentang bagaimana mengatasi ketidakseimbangan kelas dalam konteks tertentu, serta memberikan pemahaman yang lebih mendalam tentang performa model pembelajaran mesin pada dataset tersebut.

Keywords: Imbalance Data, Learning Model, Minority Class, SMOTE

Corresponding Author:
Muhammad Reza Redo
STMIK Dharma Wacana
Jl. Kenanga No:10, Metro, Indonesia
Email: reza.redo@hotmail.com

I. PENDAHULUAN

Terdapat sebuah data dari hasil percakapan keluhan pelanggan e-materai, dari data tersebut didapatkan informasi bentuk kesalahan dan solusi yang harus dilakukan. Dari data tersebut terbentuklah delapan kelas jenis keluhan pelanggan. Dari delapan kelas tersebut didapatkan informasi bahwa terdapat ketidak seimbangan dalam jumlah record/rekaman data, awalnya penelitian ini ditujukan untuk melakukan prediksi keluhan pelanggan dengan melakukan *text minning*, akan tetapi dikarenakan terdapat kelas yang tidak seimbang maka langgkah atau pendekatan yang harus dilakukan adalah menyeimbangkan kelas-kelas tersebut, secara umum menyeimbangkan sebuah kelas itu dapat dilakukan dengan berberapa tehnik seperti oversampling dan undersampling, akan tetapi pada penelitian ini akan dilakukan penggabungan kelas-kelas minoritas terlebih dahulu sebelum dilakukan oversampling. Penggabungan kelas (*class*) pada data yang

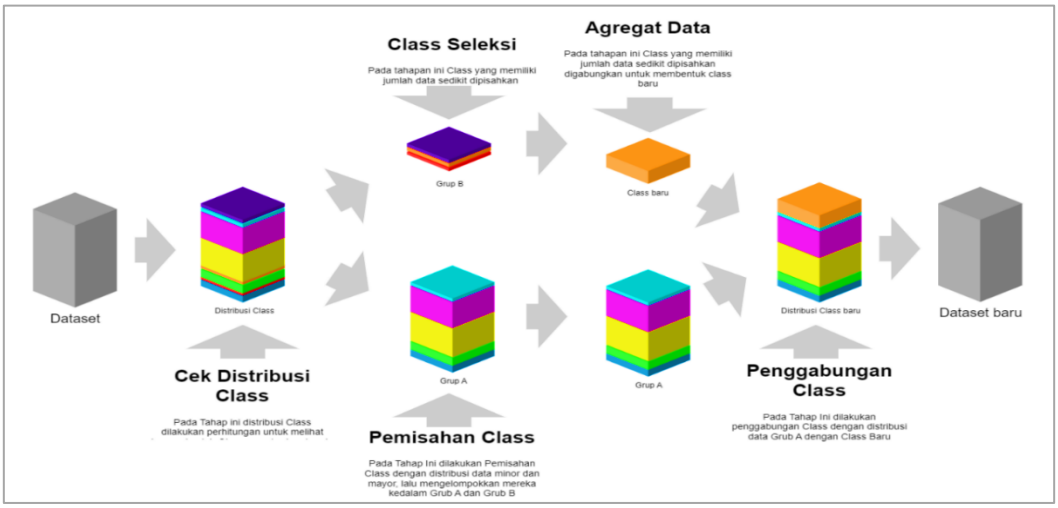
tidak seimbang adalah salah satu pendekatan yang digunakan dalam pemrosesan data[1], [1]–[4] untuk menangani ketidakseimbangan kelas. Ketidakseimbangan kelas terjadi ketika jumlah sampel dalam setiap kelas pada dataset tidak seimbang[5], [6], yaitu salah satu kelas memiliki jumlah yang jauh lebih sedikit daripada kelas lainnya. Ini bisa menjadi masalah dalam pembelajaran mesin karena model yang dibuat mungkin cenderung lebih baik dalam memprediksi kelas mayoritas, sementara kelas minoritas diabaikan. Penelitian sebelumnya tentang masalah ketidak seimbangan data dilakukan terhadap data yang tidak seimbang, mengikuti tren penelitian dalam normalisasi data. Penelitian tersebut menggambarkan sumber data yang digunakannya sebagai 7 database perpustakaan: *Springer*, *Elsevier*, *IEEExplore*, *Sage*, *ACM*, *Cambridge* dan *Wiley* metode berbasis Dekomposisi dan metode berbasis jarak *Hellinger*, untuk menyelesaikan masalah ketidakseimbangan data, kumpulan data terkait pemilihan fitur, penelitian tersebut melakukan pencarian manual untuk mengumpulkan makalah relevan yang diterbitkan dalam 2 dekade terakhir. dari penelitian tersebut di perkenalkan sebuah metode baru yang disebut "*Inverse Random Under Sampling* (*IRUS*)" yang meningkatkan akurasi klasifikasi multi label dan bermanfaat untuk pembelajaran ukuran dataset yang tidak teratur[1]. Sementara peneltian lain mengusulkan prosedur analisis baru berdasarkan ensemble learning untuk data yang tidak seimbang. Prosedur yang diusulkan melibatkan pembagian data menjadi beberapa legiun dengan mengelompokkan dan menggunakan teknik pengambilan sampel berlebih untuk memudahkan data dalam keadaan tidak seimbang dan mempelajari aturan klasifikasi berdasarkan metode hutan acak. Prosedur yang diusulkan memungkinkan *SMOTE* menghasilkan sampel yang mungkin berkontribusi pada batas klasifikasi[2], [7], [8]. Pada peneltian lain Algoritma baru untuk undersampling berdasarkan strukturalisasi *Compact Sets* (*CS*), yang di anggap mampu menangani data campuran dan tidak lengkap. algoritma tersebut digunakan untuk pengelompokan berbasis CS untuk memilih kumpulan kelas mayoritas yang sangat representatif namun tetap mampu mempertahankan objek kelas minoritas. *Compact set based Data Balancing* (*CDB*) bekerja untuk menyeimbangkan data campuran melalui pemilihan instance berbasis kumpulan kompak[9]–[11].

Pada penelitian ini kelas-kelas yang tidak seimbang akan pisahkan kemudian akan digabungkan menjadi satu kelas, setelah kelas tersebut memenuhi kriteria untuk diterapkan beberapa tehnik maka kelas kelas tersebut Kembali digabungkan menjadi satu dataset[12] untuk kemudian di proses lebih lanjut kami menyebutnya dengan kombinasi kelas minor dengan agregat data atau dengan istilah *Combination of Minor classes with Aggregate data* (*CMA*). Penggunaan agregasi data senidiri merupakan proses kunci dalam menganalisis data, dimana data yang digunakan penelitian ini adalah data sebuah klasifikasi keluhan pelanggan yang berupa *text chat*. Aggregasi sendiri banyak digunakan untuk menyediakan ringkasan statistik dan informasi berharga dalam analisis bisnis[13]–[15]. Proses ini melibatkan pengumpulan data dari berbagai sumber dan penggunaan *software agregator* data untuk mengelompokkan, memproses, dan menyajikan data dalam bentuk ringkasan yang lebih mudah dipahami. Adapun tujuan daripada penggabungan kelas ini adalah untuk melihat hasil evaluasi, sehingga didapatkan sebuah informasi apakah penggabungan kelas tersebut dapat mengatasi ketidak seimbangan kelas. Agregasi data sering digunakan untuk mengatasi ketidaklangsungan data, memahami tren, dan mengambil keputusan berdasarkan informasi yang relevan[16]. Hasil dari agregasi data membantu seorang analis statistik, untuk melakukan perhitungan mean, median, dan deviasi standar, serta memungkinkan analis tersebut untuk memahami kinerja keseluruhan dan mengidentifikasi peluang serta masalah dalam bisnis.

II. METODE

A. Pemisahan Kelas

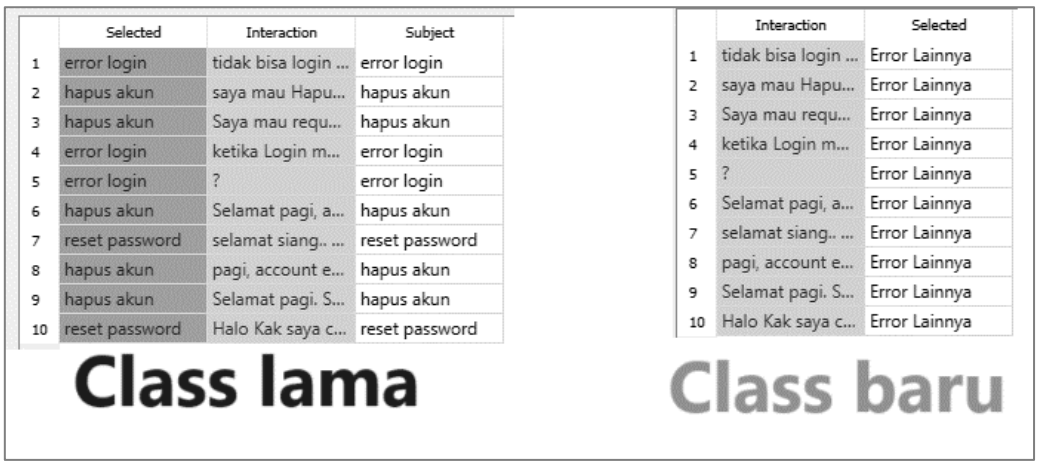
Pada tahapan pertama, Dataset dilakukan pengecekan distribusi class, hasil pengecekan tersebut memberikan informasi bahwa terdapat 8 Class yang berbeda pada dataset, dan ditemukan beberapa class minor yang artinya class minor tersebut adalah class yang memiliki data yang lebih sedikit dari class yang lain. selanjutnya *class-class* yang telah di observasi distribusinya dilakukan pengecekan jumlah data untuk kemudian dilakukan clasterisasi menjadi 2 klaster yaitu claster class *mayor* dan claster class *minor*, setelah di tetapkan Klaster *mayor* dan claster *minor* maka class class tersebut di pisahkan. Selanjutnya dilakukan pemisahan kelas menjadi 2 buah klaster yaitu klaster kelas yang memiliki data minor dan klaster yang memiliki data mayor, hal ini bertujuan untuk menjadikan kelas yang memiliki data minor untuk di gabungkan datanya dan dibentuk menjadi kelas baru dapat dilihat pada, Gambar 1. Preprosesing Data



Gambar 1. Preprosesing Data[17]

B. Penggabungan Class Minor (Aggregation Minor Class)

Pada tahapan ini kelas kelas dengan data minoritas akan di gabungkan menjadi 1 kelas tunggal dengan cara menggabungkan sejumlah besar data menjadi satu kesatuan yang lebih besar atau lebih tinggi. Proses penggabungan ini dapat melibatkan operasi matematika seperti penjumlahan, pengurangan, perkalian, atau pembagian tergantung pada tujuan agregasi dapat dilihat pada Gambar 2.



	Selected	Interaction	Subject
1	error login	tidak bisa login ...	error login
2	hapus akun	saya mau Hapu...	hapus akun
3	hapus akun	Saya mau requ...	hapus akun
4	error login	ketika Login m...	error login
5	error login	?	error login
6	hapus akun	Selamat pagi, a...	hapus akun
7	reset password	selamat siang.. ...	reset password
8	hapus akun	pagi, account e...	hapus akun
9	hapus akun	Selamat pagi. S...	hapus akun
10	reset password	Halo Kak saya c...	reset password

Class lama

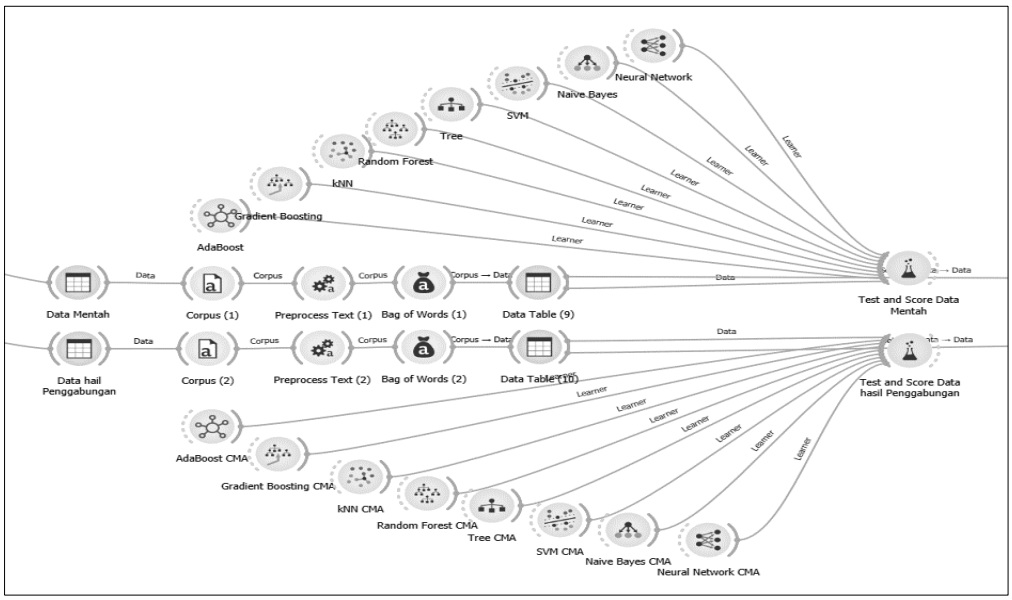
	Interaction	Selected
1	tidak bisa login ...	Error Lainnya
2	saya mau Hapu...	Error Lainnya
3	Saya mau requ...	Error Lainnya
4	ketika Login m...	Error Lainnya
5	?	Error Lainnya
6	Selamat pagi, a...	Error Lainnya
7	selamat siang.. ...	Error Lainnya
8	pagi, account e...	Error Lainnya
9	Selamat pagi. S...	Error Lainnya
10	Halo Kak saya c...	Error Lainnya

Class baru

Gambar 2. Class Data[17]

C. Uji score dengan algoritma

Pada tahapan ini hasil dari kelas baru akan di uji dengan delapan algoritma berbeda, yaitu : AdaBoost, Gradient Boosting, kNN, Naïve Bayes, Neural Network, Random Forest, SVM, dan Algoritna Decicion Tree. Pengujian ini bersifat uji komparasi untuk melihat nilai dari AUC, CA, F1, Preccicion dan Recall . Dengan uji komparasi ini, kita dapat mengevaluasi dan dapat melihat hasil yang optimal dalam aplikasi data tersebut. selanjutnya Data mentah dan Data yang telah di olah akan sama sama diberikan penerapan oversampling algoritma yang kemudian dari hasil oversampling tersebut data akan Kembali di uji dengan Delapan algortima menggunakan *learning model* yang sama, selanjutnya dilakukan uji terakhir dengan diberikan penerapan Undersampling algoritma yang kemudian dari hasil Undersampling tersebut data akan Kembali di uji dengan Delapan algortima, hal ini bertujuan untuk melihat rata-rata score yang dapat dihasilkan dari learning model tersebut

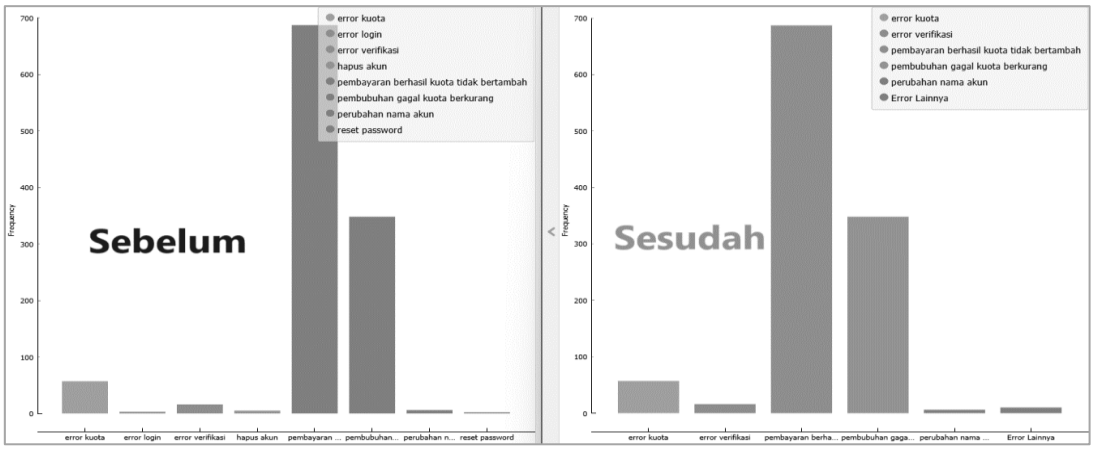


Gambar 3. Learning Model

III. HASIL DAN PEMBAHASAN

D. Hasil

Hasil penggabungan kelas ini melahirkan sebuah kelas baru dari pengabungan beberapa kelas sebelumnya. Dari hasil penggabungan tersebut maka diharapkan kelas baru dapat dilakukan sebuah analisis data. Setelah itu dilakukan pengecekan distribusi kelas Kembali, hal ini bertujuan untuk memastikan bahwa kelas kelas tersebut telah cukup sesuai dan layak untuk dilakukan tidakan selanjutnya. Dari hasil pengecekan distribusi kelas maka dapat di temukan informasi jumlah kelas baru yang sebelum nya berjumlah 8 kelas kini telah menjadi 6 kelas dapat dilihat pada Gambar 4 .



Gambar 4. Class Data Preprosesing[17]

Hasil Evaluasi Model dengan menggunakan Data lama (RAW data)

Table 1. Hasil Evaluasi Model Raw data

Model	*AUC*	*CA*	*F1*	*Prec*	*Recall*	*MCC*
AdaBoost	0.727	0.688	0.674	0.671	0.688	0.372
Gradient Boosting	0.797	0.723	0.709	0.707	0.723	0.445

kNN	0.718	0.671	0.638	0.627	0.671	0.309
Naive Bayes	0.567	0.03	0.05	0.618	0.03	0.013
Neural Network	0.767	0.707	0.696	0.692	0.707	0.418
Random Forest	0.785	0.711	0.682	0.693	0.711	0.4
SVM	0.669	0.464	0.433	0.592	0.464	0.182
Tree	0.696	0.673	0.667	0.665	0.673	0.364

Hasil Evaluasi Model dengan Data kombinasi kelas minor

Table 2. Hasil Evaluasi Model Raw data

Model	***AUC***	***CA***	***F1***	***Prec***	***Recall***	***MCC***
AdaBoost CMA	0.761	0.69	0.675	0.669	0.69	0.376
Gradient Boosting CMA	0.813	0.717	0.709	0.71	0.717	0.441
kNN CMA	0.699	0.646	0.605	0.6	0.646	0.242
Naive Bayes CMA	0.589	0.114	0.18	0.638	0.114	0.063
Neural Network CMA	0.766	0.695	0.685	0.682	0.695	0.399
Random Forest CMA	0.802	0.732	0.711	0.721	0.732	0.451
SVM CMA	0.628	0.433	0.392	0.577	0.433	0.145
Tree CMA	0.684	0.66	0.649	0.641	0.66	0.329

Hasil Evaluasi Model dengan Data lama yang di Oversampling

Table 3. Hasil Evaluasi Model Raw data yang di Oversampling

Model	***AUC***	***CA***	***F1***	***Prec***	***Recall***	***MCC***
AdaBoost	0.989	0.915	0.914	0.917	0.915	0.904
Gradient Boosting	0.991	0.912	0.911	0.912	0.912	0.9
kNN	0.974	0.877	0.868	0.869	0.877	0.86
Naive Bayes	0.978	0.826	0.819	0.824	0.826	0.803
Neural Network	0.987	0.935	0.934	0.934	0.935	0.926
Random Forest	0.994	0.928	0.927	0.927	0.928	0.918
SVM	0.858	0.455	0.43	0.607	0.455	0.402
Tree	0.963	0.896	0.893	0.891	0.896	0.882

Hasil Evaluasi Model dengan Data kombinasi kelas minor yang di Oversampling

Table 4. Hasil Evaluasi Model kelas minor yang di Oversampling

Model	***AUC***	***CA***	***F1***	***Prec***	***Recall***	***MCC***
AdaBoost CMA	0.985	0.893	0.891	0.892	0.893	0.872
Gradient Boosting CMA	0.983	0.881	0.88	0.882	0.881	0.858
kNN CMA	0.961	0.821	0.81	0.813	0.821	0.789
Naive Bayes CMA	0.959	0.754	0.743	0.751	0.754	0.708
Neural Network CMA	0.98	0.911	0.91	0.91	0.911	0.894
Random Forest CMA	0.991	0.914	0.913	0.913	0.914	0.897
SVM CMA	0.674	0.337	0.326	0.49	0.337	0.221
Tree CMA	0.952	0.874	0.87	0.869	0.874	0.849

Hasil Evaluasi Model dengan Data lama yang di Undersampling

Table 5. Hasil Evaluasi Model Raw data yang di Undersampling

Model	AUC	CA	F1	Prec	Recall	MCC
AdaBoost	0.344	0	0	0	0	-0.157
Gradient Boosting	0.283	0	0	0	0	-0.144
kNN	0.362	0	0	0	0	-0.179
Naive Bayes	0.304	0.125	0.071	0.05	0.125	0
Neural Network	0.295	0.125	0.062	0.042	0.125	0
Random Forest	0.346	0	0	0	0	-0.158
SVM	0.424	0	0	0	0	-0.179
Tree	0.415	0.062	0.042	0.031	0.062	-0.076

Hasil Evaluasi Model dengan Data kombinasi kelas minor yang di Undersampling

Table 6. Hasil Evaluasi Model Raw data

Model	AUC	CA	F1	Prec	Recall	MCC
AdaBoost CMA	0.589	0.278	0.264	0.262	0.278	0.135
Gradient Boosting CMA	0.515	0.139	0.131	0.135	0.139	-0.034
kNN CMA	0.469	0.111	0.115	0.167	0.111	-0.075
Naive Bayes CMA	0.65	0.222	0.147	0.114	0.222	0.074
Neural Network CMA	0.59	0.194	0.192	0.222	0.194	0.035
Random Forest CMA	0.45	0.083	0.076	0.077	0.083	-0.102
SVM CMA	0.334	0.194	0.188	0.213	0.194	0.038
Tree CMA	0.53	0.111	0.102	0.101	0.111	-0.069

Hasil Evaluasi rata rata keseluruhan model terhadap 6 bentuk perlakukan data yang berbeda

Table 7. Hasil Evaluasi rata-rata

Data	Kelas	Data	Seimbang	Rasio Data/Class	AUC	CA	F1	Prec	Recall	MCC	Resiko Overfitting
RAW	8	1123	Tidak	1: 495	0.716	0.583	0.569	0.658	0.583	0.313	Menengah
CMA	6	1123	Tidak	1: 169	0.718	0.586	0.576	0.655	0.586	0.306	Menengah
RAW+Oversampling	8	5496	Ya	1: 8	0.967	0.843	0.837	0.860	0.843	0.824	Tinggi
CMA+Oversampling	6	4122	Ya	1: 6	0.936	0.798	0.793	0.815	0.798	0.761	Menengah
RAW+Undersampling	8	16	Ya	1: 8	0.347	0.039	0.022	0.015	0.039	-0.112	Rendah
CMA+Undersampling	6	36	Ya	1: 6	0.516	0.167	0.152	0.161	0.167	0.000	Rendah

E. Pembahasan

Pada proses preprosesing data kita telah melihat kelas baru terbentuk dari penggabungan 3 kelas berbeda yaitu kelas bernama "error lainnya" kelas baru ini terbentuk dari penggabungan kelas "error login, hapus akun dan reset password" berdasarkan [Gambar 2] dari hasil observasi pada kelas sebelumnya ditemukan fakta bahwa data pada kelas "error login dan reset password" merupakan kelas dengan jumlah data paling rendah yaitu hanya memiliki 2 buah data saja. dari observasi [Gambar 4] juga terdapat kelas dengan nama "perubahan akun" kelas ini tidak ikut digabungkan menjadi kelas minoritas karena tidak relevan terhadap kelas baru, sementara alasan 3 kelas baru yang di gabungkan adalah karena kelas tersebut memilki

kesamaan prilaku atau tindakan penanganan. dalam beberapa kasus penggabungan kelas ada pertimbangan untuk melakukan ekstrasi fitur yang kemudian mengelompokkan kedalam fitur-fitur yang memiliki korelasi dengan hubungan yang sama, namun pada kasus ini penggabungan kelas melalui ekstraksi fitur dan kesamaan hubungan tidak mungkin dilakukan karena bentuk data adalah data text.

Setelah kelas-kelas tersebut digabung, langkah selanjutnya adalah melakukan evaluasi model[Gambar 3] pembelajaran mesin dengan menggunakan dua bentuk dataset berbeda namun masing masing data diberi perlakuan yang sama, dari hasil evaluasi model tersebut maka dapat ditemukan fakta bahwa hasil dari evaluasi model data awal [Table 1] dengan evaluasi data yang sudah dilakukan penggabungan kelas tidak terlalu menunjukan perbedaan yang signifikan hal ini dikarenakan data train dan test berasal dari sumber data yang sama, jika merujuk pada masing masing tabel hasil [Table 7], memang terlihat hasil evaluasi pada data yang telah di lakukan penggabungan kelas sedikit memilki nilai lebih besar walaupun tidak terlalu signifikan, jika kita melihat pada nilai *Matthews Correlation Coefficient*(MCC) Algoritma *Nueral Network* dan *kNN* pada Data awal [Table 1] memiliki nilai yang sedikit lebih tinggi daripada menggunakan Data yang telah digabungkan kelasnya [Table 2], hasil evaluasi tersebut menunjukan bahwa penerapan tehnik ini masih memiliki kekurangan pada ke dua algoritma tersebut. MCC sendiri dapat memberikan gambaran tentang kinerja model klasifikasi pada kelas yang tidak seimbang[18]–[21]. dengan rendahnya nilai MCC tersebut bisa dipastikan bahwa data benar tidak seimbang.

Experimental selanjutnya adalah dengan menerapkan tehnik oversampling dan undersampling pada kedua data yang berbeda, dari hasil oversampling terlihat masalah ketidakseimbangan kelas pada algoritma sebelumnya dapat teratasi hanya saja pada algoritma SVM masih menunjukan nilai rendah[], perlu diketahui bahwa parameter untuk model SVM ini menggunakan kernel RBF dengan penalti cost =1 dan 100 iterasi, parameter ini sejauh ini adalah parameter yang paling optimal, kelemahan dari SVM sendiri karena kedua bentuk data bukanlah sebuah regresi hal ini menyadikan algoritma ini tergolong lemah jika digunakan dengan bentuk data tersebut. selanjutnya hasil evaluasi pada data yang dilakukan undersampling justru menunjukan bahwa hasil evaluasi model pembelajaran nilai F1, Precicion dan Recall terdapat nilai 0 pada hasil evaluasi model dengan menggunakan dataset awal yang tidak dilakukan penggabungan kelas, hal ini dikarenakan dataset yang tidak digabungkan kelasnya terdapat dua kelas minoritas dengan jumlah record data sebanyak dua buah saja, sehingga mengakibatkan algoritma undersampling mengambil record data terendah sebagai parameter undersampling data. Situasi di mana presisi dan F1-score keduanya 0 biasanya menunjukkan masalah serius dalam kualitas model. Dari hasil evaluasi tersebut mengindikasikan bahwa model mungkin saja terindikasi overfiftting karena dari informasi awal kita sudah tau bahwa dataset memiliki masalah ketidak seimbangan dalam kelas.

IV. KESIMPULAN

Dalam proses preprocessing data, terjadi pembentukan kelas baru yang disebut "error lainnya" dengan menggabungkan tiga kelas berbeda: "error login," "hapus akun," dan "reset password" karena kelas "error login dan reset password" memiliki data yang sangat sedikit. Meskipun penggabungan kelas tersebut tidak memiliki dampak signifikan pada model, karena data pelatihan dan pengujian berasal dari sumber yang sama, terdapat beberapa peningkatan sejumlah metrik evaluasi. Namun, perbedaan ini tidak signifikan, dan nilai Matthews Correlation Coefficient (MCC) yang rendah menunjukkan ketidakseimbangan data. Eksperimen penelitian yang melibatkan oversampling berhasil mengatasi masalah ketidakseimbangan kelas pada beberapa algoritma, kecuali SVM, yang tetap memiliki nilai rendah. Hasil evaluasi pada dataset yang mengalami undersampling menunjukkan nilai F1, Precision, dan Recall sebesar 0, mengindikasikan masalah serius dalam kualitas model, seperti overfitting apabila dilakukan oversampling. Pada dataset yang menggabungkan kelasnya, empat algoritma memiliki nilai MCC negatif, menunjukkan bahwa model sepenuhnya salah dalam memisahkan kedua kelas. Secara keseluruhan, penelitian ini menyoroti bahwa masalah ketidakseimbangan kelas menjadi masalah utama yang perlu diatasi dalam pengembangan model klasifikasi. Penggabungan kelas minoritas pada dataset yang tidak seimbang dianggap tidak berhasil mengatasi permasalahan tersebut

UCAPAN TERIMAKASIH

Penulis mengucapkan terimakasih kepada saudara Marchel Aji Alfero atas bantuannya yang bersedia membantu mengumpulkan dataset ini.

Terima Kasih atas bantuan Pembiayaan LPPM STMIK Dharma Wacana atas kegiatan Konfrensi ilmiah ini.

REFERENCES

[1] H. Ali, M. N. Mohd Salleh, R. Saedudin, K. Hussain, and M. F. Mushtaq, "Imbalance class problems in data mining: a review," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 14, no. 3, p. 1552, Jun. 2019, doi: 10.11591/ijeecs.v14.i3.pp1552-1563.

[2] M. Z. Abedin, C. Guotai, P. Hajek, and T. Zhang, "Combining weighted SMOTE with ensemble learning for the class-imbalanced prediction of small business credit risk," *Complex Intell. Syst.*, vol. 9, no. 4, pp. 3559–3579, 2023, doi: 10.1007/s40747-021-00614-4.

[3] . H., O. S. Sitompul, E. B. Nababan, . T., D. Abdullah, and A. S. Ahmar, "A New Diversity Technique for Imbalance Learning Ensembles," *Int. J. Eng. Amp Technol.*, vol. 7, no. 2.14, p. 478, Apr. 2018, doi: 10.14419/ijet.v7i2.11251.

[4] K. Raghavendar, I. Batra, and A. Malik, "Novel Framework for Resources Optimization to Solve Class Imbalance Problems," *Proceedings - 2021 International Conference on Computing Sciences, ICCS 2021*. Institute of Electrical and Electronics Engineers Inc., pp. 143–147, 2021. doi: 10.1109/ICCS54944.2021.00036.

[5] G. Idakwo, "Structure–activity relationship-based chemical classification of highly imbalanced Tox21 datasets," *Journal of Cheminformatics*, vol. 12, no. 1. 2020. doi: 10.1186/s13321-020-00468-x.

[6] "A Review on Class Imbalance Problem: Analysis and Potential Solutions," *Int. J. Comput. Sci. Issues*, vol. 14, no. 6, pp. 43–51, Nov. 2017, doi: 10.20943/01201706.4351.

[7] D. Gyoten, M. Ohkubo, and Y. Nagata, "Imbalanced data classification procedure based on SMOTE," *Total Qual. Sci.*, vol. 5, no. 2, pp. 64–71, Jan. 2020, doi: 10.17929/tqs.5.64.

[8] J. Grzyb, "SVM ensemble training for imbalanced data classification using multi-objective optimization techniques," *Applied Intelligence*, vol. 53, no. 12. pp. 15424–15441, 2023. doi: 10.1007/s10489-022-04291-9.

[9] Y. Villuendas-Rey and M. ia Matilde Garc\' ia-Lorenzo, "Mixed Data Balancing through Compact Sets Based Instance Selection," in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, Springer Berlin Heidelberg, 2013, pp. 254–261. doi: 10.1007/978-3-642-41822-8_32.

[10] J. Potthast, V. Grimm, and J. Rubart, "Immersive Experience of Multidimensional Data using Mixed Reality based Scatterplots," *ACM International Conference Proceeding Series*. Association for Computing Machinery, pp. 594–596, 2022. doi: 10.1145/3543758.3547515.

[11] L. Wang, Q. Zhang, X. Niu, Y. Ren, and J. Xia, "Outlier detection of mixed data based on neighborhood combinatorial entropy," *Computers, Materials and Continua*, vol. 69, no. 2. Tech Science Press, pp. 1765–1781, 2021. doi: 10.32604/cmc.2021.017516.

[12] "(PDF) Data Keluhan pelanggan." Accessed: Oct. 30, 2023. [Online]. Available: https://www.researchgate.net/publication/374169609_Data_Keluhan_pelanggan?channel=doi&linkId=651291102c6cfe2cc21013dd&showFulltext=true

[13] "What is Data Aggregation?," Data Management. Accessed: Oct. 29, 2023. [Online]. Available: https://www.techtarget.com/searchdatamanagement/definition/data-aggregation

[14] Y. Wang, Y. Yuan, G. Wang, and Y. Ma, "Graph cells: Top-k structural-textual aggregated query over information networks," *Information Sciences*, vol. 547. Elsevier Inc., pp. 354–366, 2021. doi: 10.1016/j.ins.2020.08.057.

[15] A. Cuzzocrea, "BigMDHealth: Supporting Multidimensional Big Data Management and Analytics over Big Healthcare Data via Effective and Efficient Multidimensional Aggregate Queries over Key-Value Stores," *Lecture Notes on Data Engineering and Communications Technologies*, vol. 165. Springer Science and Business Media Deutschland GmbH, pp. 187–194, 2023. doi: 10.1007/978-981-99-0741-0_13.

[16] R. Hans, "Pelajari Seluk Beluk Tugas Data Analyst & Fungsinya." Accessed: Oct. 29, 2023. [Online]. Available: https://dqlab.id/pelajari-seluk-beluk-tugas-data-analyst-and-fungsinya

[17] R. Redo and A. Perdana, "PENGGABUNGAN CLASS PADA DATA YANG TIDAK SEIMBANG." Oct. 2023. doi: 10.13140/RG.2.2.26131.45608.

[18] K. S. Nugroho, "Confusion Matrix untuk Evaluasi Model pada Supervised Learning," Medium. Accessed: Oct. 30, 2023. [Online]. Available: https://ksnugroho.medium.com/confusion-matrix-untuk-evaluasi-model-pada-unsupervised-machine-learning-bc4b1ae9ae3f

[19] K. Abhishek and G. Hamarneh, "Matthews correlation coefficient loss for deep convolutional networks: Application to skin lesion segmentation," *Proceedings - International Symposium on*

Biomedical Imaging, vol. 2021-April. IEEE Computer Society, pp. 225–229, 2021. doi: 10.1109/ISBI48211.2021.9433782.

[20] D. Chicco and G. Jurman, "The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification," *BioData Mining*, vol. 16, no. 1. BioMed Central Ltd, 2023. doi: 10.1186/s13040-023-00322-4.

[21] D. Chicco and G. Jurman, "A statistical comparison between Matthews correlation coefficient (MCC), prevalence threshold, and Fowlkes–Mallows index," *Journal of Biomedical Informatics*, vol. 144. Academic Press Inc., 2023. doi: 10.1016/j.jbi.2023.104426.