Techno.COM, Vol. 24, No. 4, November 2025: 1359-1367

Penanganan Ketidakseimbangan Data Ekstrim pada Sistem Prediksi

Handling Extreme Data Imbalance in Prediction Systems

Ari Nugroho Putro[1], **Much Aziz Muslim**[2]
Department of Computer Science, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Semarang
E-mail : arinugrohoputro.students.unnes.ac.id [1] , *a212muslim@mail.unnes.ac.id*[2]
Received 30 October 2025; Revised 20 November 2025; Accepted 25 November 2025

Abstrak - Salah satu masalah utama dalam sistem prediksi adalah ketidakseimbangan data, di mana kelas tertentu sangat kurang terwakili dibandingkan dengan kelas lainnya. Ketidakseimbangan data dapat menyebabkan bias model, di mana model lebih mudah mendeteksi kelas mayoritas tetapi lemah dalam mendeteksi kelas minoritas. Terutama pada data dengan ketidakseimbangan ekstrem dengan IR >9, model memiliki akurasi tinggi tetapi performa recall rendah. Hal ini merugikan sistem prediksi yang memprioritaskan deteksi kelas minoritas. Penelitian ini bertujuan untuk meningkatkan recall pada dataset yang sangat tidak seimbang dengan menggunakan empat teknik penanganan ketidakseimbangan data, yaitu SMOTE dan OHIT pada level data, serta CSL dan CW pada level model. Teknik pada level data menyeimbangkan distribusi kelas dengan menambahkan data sintetis, sedangkan teknik pada level model meningkatkan sensitivitas terhadap kelas minoritas. Model yang digunakan sebagai baseline adalah LR untuk mengamati peningkatan recall dari keempat teknik penanganan ketidakseimbangan data. Dari hasil pengujian semua teknik penanganan ketidakseimbangan data, semuanya meningkatkan recall dengan margin sebesar 0,3243. Peningkatan recall tertinggi dicapai oleh LR-SMOTE dengan margin sebesar 0,3256. Penelitian ini menunjukkan bahwa recall model dapat ditingkatkan dengan menggunakan teknik penanganan ketidakseimbangan data.

Kata kunci – ketidakseimbangan data ekstrem, sistem prediksi, recall, penanganan ketidakseimbangan data

Abstract - *One major problem in prediction systems is data imbalance, where certain classes are significantly underrepresented compared to other classes. Data imbalance can cause model bias, where models easily detect majority classes but are weak in detecting minority classes. Especially in data with extreme imbalance with an IR >9, models have high accuracy but low recall performance. This is detrimental to prediction systems that prioritize the detection of minority classes. This study aims to improve recall on extremely imbalanced datasets by using four data imbalance handling techniques, namely SMOTE and OHIT at the data level, and CSL and CW at the model level. Data-level techniques balance class distribution by adding synthetic data, while model-level techniques increase sensitivity to minority classes. The model used as a baseline is LR to observe the increase in recall from the four data imbalance handling techniques. From the results of testing all data imbalance handling techniques, all of them increased recall by a margin of 0.3243. The highest increase in recall was achieved by LR-SMOTE with a margin of 0.3256. This study shows that model recall can be improved using data imbalance handling techniques.*

Keywords - *extreme data imbalance, prediction system, recall, data imbalance handling*

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v24i4.15005

1. PENDAHULUAN

Masalah ketidakseimbangan data menjadi masalah penting yang sering dihadapi dalam memodelkan sistem prediksi. Subset data kelas mayoritas mendominasi subset data kelas minoritas. Ketidakseimbangan data seperti ini, sering ditemui pada data-data medis [1], [2], [3], penipuan [4], [5], [6], prediksi churn [7], [8], [9], dan pendidikan [10], [11], [12]. dimana kelas minoritas menjadi fokus utama untuk dideteksi. Ketidakseimbangan data dapat membuat performa model menjadi bias dan berkinerja rendah dalam mendeteksi kelas minoritas [13]. Bias ini disebabkan karena model cenderung lebih mengenal pola mayoritas yang memiliki subset data dominan dan tidak sensitif terhadap kelas minoritas yang memiliki data kurang representatif.

Ketidakseimbangan data dapat dikategorikan berdasarkan *Imbalance Ratio* (IR) yang merupakan representasi rasio kelas mayoritas terhadap kelas minoritas [3]. Semakin tinggi IR, maka semakin tinggi juga ketidakseimbangan datanya. Data yang memiliki IR 1.9 sampai <9 dapat dikategorikan sebagai ketidakseimbangan sedang dan IR ≥ 9 dapat dikategorikan sebagai ketidakseimbangan ekstrim [14]. Sebagai contoh, apabila ketidakseimbangan data 0.20 : 0.80 maka IR adalah 4, jika ketidakseimbangannya adalah 0.03 : 0.90 maka IR adalah 90 [15].

Ketidakseimbangan data yang ekstrim berdampak pada model yang memiliki akurasi tinggi tetapi hanya karena keberhasilannya dalam memprediksi kelas mayoritas, namun model kesulitan untuk memprediksi kelas minoritas [16]. Hal ini menciptakan bias performa pada model, padahal model gagal dalam mendeteksi kejadian penting yang jarang muncul [17]. Oleh karena itu diperlukan pendekatan model yang memiliki *recall* tinggi tanpa mengorbankan penurunan signifikan terhadap akurasi dalam mendeteksi keseluruhan data [18].

Untuk mengatasi ketidakseimbangan data model perlu diberikan teknik penanganan ketidakseimbangan data, agar model memiliki *recall* tinggi dalam mendeteksi kelas minoritas [3], [19], [20]. Teknik ini membuat data menjadi seimbang melalui dua pendekatan berbeda, yaitu melalui pendekatan pembuatan data sintetis dan pendekatan model dengan memberikan bobot lebih sensitif pada kelas minoritas. Pendekatan pembuatan data sintetis yang bisa digunakan adalah teknik *Synthetic Minority Over Sampling Technique* (SMOTE) dan *Oversampling Hybrid Integrated Technique* (OHIT). SMOTE dan OHIT membuat data sintetis pada kelas minoritas, sehingga data menjadi seimbang dan memiliki IR 1 [15], [21]. Sedangkan untuk pendekatan model, teknik yang bisa diterapkan adalah *Cost Sensitive Learning* (CSL) dan *Class Weighting* (CW). CSL dan CW tidak membuat data sintetis seperti SMOTE dan OHIT, teknik ini membuat model memberikan sensitivitas dan perhatian yang lebih tinggi terhadap kelas minoritas [22], [23].

Penelitian ini berfokus untuk meningkatkan recall model pada sistem prediksi yang memiliki ketidakseimbangan data ekstrim dengan menerapkan 4 teknik penanganan ketidakseimbangan data yaitu SMOTE, OHIT, CSL, dan CW yang diterapkan pada Logistic Regression (LR) sebagai *baseline* model. 4 pendekatan ini dipilih karena masing-masing teknik memiliki pola kerja yang berbeda. SMOTE membuat data sintetis melalui kedekatan antar data, OHIT melakukan *clustering* data, CSL memberikan penalti pada model, dan CW memberikan bobot tinggi pada data minoritas. *Recall* merupakan metrik evaluasi yang menunjukkan sejauh mana model mampu mendeteksi kelas minoritas pada ketidakseimbangan data [24]. Recall menjadi penting pada sistem prediksi yang mengutamakan deteksi pada kelas-kelas minoritas yang jarang muncul [18].

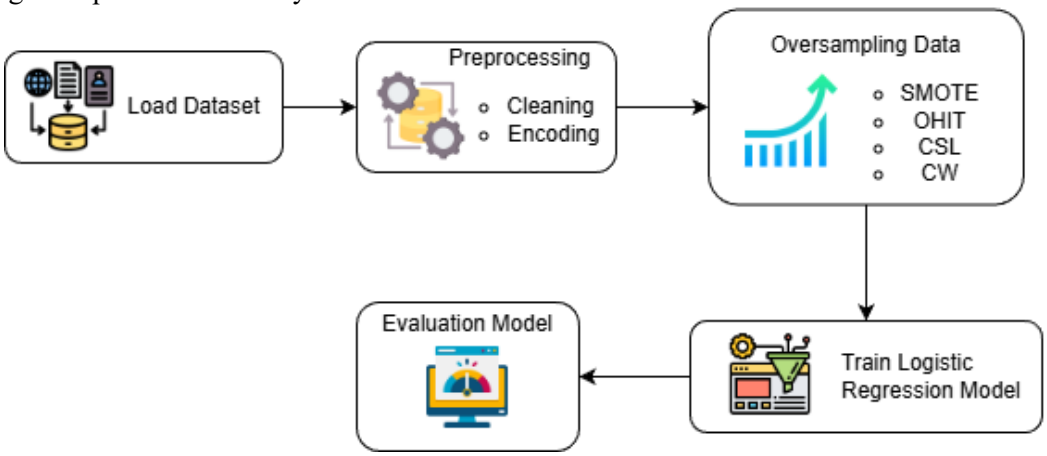
2. METODE PENELITIAN

2.1 Deskripsi Data

Penelitian ini menggunakan dataset *online course user engagement data* (OCUED) yang diunduh dari repositori kaggle pada tautan berikut: https://www.kaggle.com/datasets/thedevastator/online-course-user-engagement-data. Dataset ini merupakan catatan-catatan komperhensif tentang interaksi keterlibatan siswa dalam pembelajaran daring. Data ini

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v24i4.15005

menyediakan informasi berupa data demografis siswa, metrik aktivitas dan indikator kinerja siswa dalam pembelajaran daring di berbagai kelas kursus. Pada metrik demografis yang disediakan mencakup jenis kelamin siswa, tahun kelahiran, tingkat pendidikan sebelum mengikuti kursus, serta negara asal siswa. Pada metrik aktivitas, dataset memuat berbagai parameter keterlibatan siswa dengan platform seperti jumlah total interaksi dengan sistem, jumlah hari aktif selama mengikuti kursus, jumlah bab materi yang diakses, serta frekuensi aktivitas dalam forum diskusi. Dan pada metrik indikator kinerja mencakup nilai akhir yang diperoleh siswa, status apakah mereka melihat materi kursus, sejauh mana mereka mengeksplorasi materi secara mendalam, serta status sertifikasi sebagai penanda penyelesaian kursus. Dataset juga menyediakan informasi mengenai apakah siswa menyelesaikan kursus atau tidak.



Gambar 1. Metode penanganan ketidakseimbangan data ekstrim pada sistem prediksi

2.2 Prapemrosesan

Dataset OCUED memiliki 9 kelas kursus yang kemudian dipilih 2 kelas kursus yang memiliki nilai IR tertinggi. 2 kelas kursus yang memiliki nilai IR tertinggi adalah kelas kursus HarvardX/CS50x/2012 dengan nilai IR 130.80 dan kelas kursus HarvardX/CB22x/2013_Spring dengan nilai IR 77.13. Dataset OCUED memiliki 19 fitur yaitu *(index, Random, nplay_video, registered, nforum_posts, final_cc_cname_DI, LoE_DI, YoB, gender, start_time_DI, last_event_DI, roles, grade, incomplete_flag, nevents, viewed, explored, ndays_act, nchapters) dan setelah melewati proses cleaning data berupa drop duplicate, drop missing value, dan drop fitur-fitur yang tidak informatif dataset memiliki 5 fitur yaitu (nevents, viewed, explored, ndays_act, nchapters).*

Setelah proses pembersihan data selesai dilakukan, selanjutnya adalah melakukan normalisasi data menggunakan *minmax scaler* agar data memiliki karakteristik rentang nilai yang sama. Pada penelitian ini model validation yang digunakan adalah *holdout validation* dengan *split* 80:20 pada dataset, 80% subset data untuk *train* dan 20% subset data untuk *test*. Dataset memiliki 2 label kelas yaitu *certified* 1 sebagai kelas minoritas dan kelas *certified* 0 sebagai kelas mayoritas.

2.3 Teknik Penanganan Ketidakseimbangan Data

Dalam penelitian ini menerapkan 4 teknik penanganan ketidakseimbangan data. 2 teknik dengan pendekatan menambah data sintetis yaitu SMOTE dan OHIT dan 2 teknik dengan pendekatan menekankan model lebih sensitif terhadap kelas minoritas yaitu CSL dan CW. Penerapan 4 teknik penanganan ketidakseimbangan data ini dilakukan untuk memberi gambaran peningkatan *recall* yang dihasilkan oleh setiap teknik. SMOTE bekerja dengan cara membuat data sintetik baru untuk kelas minoritas berdasarkan kedekatan spasial dan temporal antar sampel [25]. OHIT mengelompokkan data kelas minoritas ke dalam beberapa *cluster* berdasarkan kemiripan temporal antar sampel, OHIT melakukan *oversampling* secara selektif di dalam setiap

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v24i4.15005

cluster untuk menghasilkan data baru yang lebih bervariasi namun tetap representatif terhadap karakteristik asli kelas minoritas [21]. CSL merupakan pendekatan yang mengintegrasikan bobot atau biaya yang berbeda untuk kesalahan klasifikasi pada setiap kelas [23]. Kesalahan dalam mengklasifikasikan sampel dari kelas minoritas dikenai penalti yang lebih besar dibandingkan kesalahan pada kelas mayoritas. CW memberikan bobot yang lebih besar pada kelas minoritas saat melatih model [22].

2.4 Model Prediksi

Model prediksi yang digunakan dalam penelitian ini adalah RL. Model RL merupakan model regresi linier yang menggunakan fungsi sigmoid yang memiliki keluaran pada rentang 0 hingga 1 sebagai interpretasi probabilitas [26]. Model LR sensitif terhadap distribusi kelas pada dataset [27], sehingga model ini efektif untuk menjadi *baseline* model untuk mengetahui peningkatan recall yang dihasilkan oleh setiap teknik penanganan ketidakseimbangan data. Model LR mengestimasi probabilitas p bahwa suatu himpunan fitur x termasuk dalam kelas positif menggunakan fungsi logistik

$$p\,(\,y = 1 \mid x\,) = \sigma(w^T x + b) \text{ [28]}$$

Dengan fungsi sigmoid yang didefinisikan dengan $\sigma(z) = \frac{1}{1+ e^{-z}}$ [28]. Selama proses pelatihan, model berusaha meminimalkan perbedaan antara probabilitas prediksi dan label sebenarnya menggunakan *maximum likelihood estimation*. Dalam praktiknya, hal ini dilakukan dengan meminimalkan fungsi *cross entropy loss*

$$J(w,b) = -\frac{1}{m}\sum_{i=1}^{m}[y^i \log(\sigma(w^T x^{(i)} + b)) + (1 - y^{(i)}) log(1 - \sigma(w^T x^{(i)} + b))] \text{ [28]}$$

Bobot w dan b bias diperbarui menggunakan algoritma optimisasi seperti *gradient descent*. Pembaruan dilakukan secara iteratif dengan menggerakkan parameter ke arah gradien penurunan tercepat sehingga nilai fungsi biaya $J(w,b)$ terus mengecil selama proses pelatihan berlangsung.

Model prediksi LR akan diterapkan pada 5 tahap. Tahap 1 menerapkan model LR tanpa teknik penanganan ketidakseimbangan data sebagai acuan *baseline* model. Tahap 2 menerapkan model LR menggunakan SMOTE untuk mengetahui peningkatan *recall* yang dihasilkan SMOTE. Tahap 3 menerapkan model LR menggunakan OHIT untuk mengetahui peningkatan *recall* yang dihasilkan OHIT. Tahap 4 menerapkan model LR menggunakan CSL untuk mengetahui peningkatan *recall* yang dihasilkan CSL. Tahap 5 menerapkan model LR menggunakan CW untuk mengetahui peningkatan *recall* yang dihasilkan CW.

2.5 Metrik Evaluasi

Akurasi adalah metrik evaluasi yang digunakan untuk mengukur seberapa banyak prediksi model yang benar dibandingkan dengan seluruh jumlah prediksi yang dilakukan [29]. Metrik ini digunakan dalam masalah klasifikasi untuk menilai kinerja model secara keseluruhan.

$$Akurasi = \ (TP + TN)/(TP + TN + FP + FN)$$

Recall mengukur seberapa banyak data dari kelas positif (kelas minoritas) yang berhasil dikenali oleh model [30]. Ini sangat penting dalam konteks *class imbalance*, karena menunjukkan kemampuan model dalam menangkap kelas minoritas.

$$Recall = \ TP/(TP + FN)$$

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v24i4.15005

Precision [30] mengukur seberapa banyak prediksi positif yang benar-benar positif . Metrik ini penting saat kesalahan positif palsu (*False Positives*) harus diminimalkan, seperti pada kasus *class imbalance*.

$$Precision = \ TP/(TP + FP)$$

F1-score [28] adalah rata-rata harmonik antara *precision* dan *recall*, dan digunakan untuk menyeimbangkan *trade off* antara keduanya [28]. Metrik ini sangat berguna ketika diperlukan keseimbangan antara deteksi kelas minoritas dan menghindari prediksi salah.

$$F1\ Score = 2. \ (Precision\ .\ Recall)/(Precision + Recall)$$

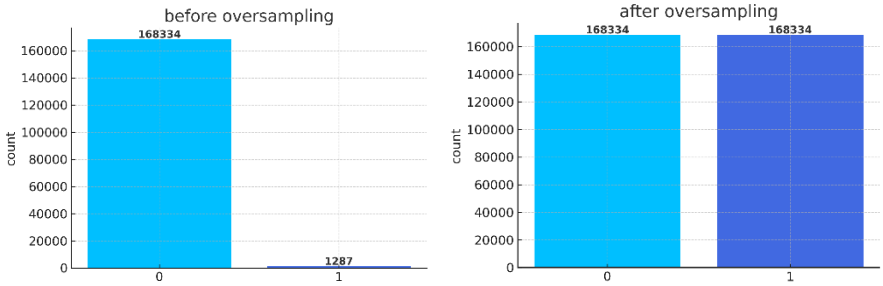
AUC mengukur kemampuan model dalam membedakan antara kelas positif dan negatif [28]. ROC (*Receiver Operating Characteristic*) sendiri adalah plot antara *True Positive Rate* dan *False Positive Rate*. Nilai AUC mendekati 1 menunjukkan bahwa model mampu membedakan kedua kelas dengan sangat baik.

$$AUC = \int \_0\text{^}1\ \ [\![TPR\ (FPR)\ \ d\ (FPR)]\!]$$

3. HASIL DAN PEMBAHASAN

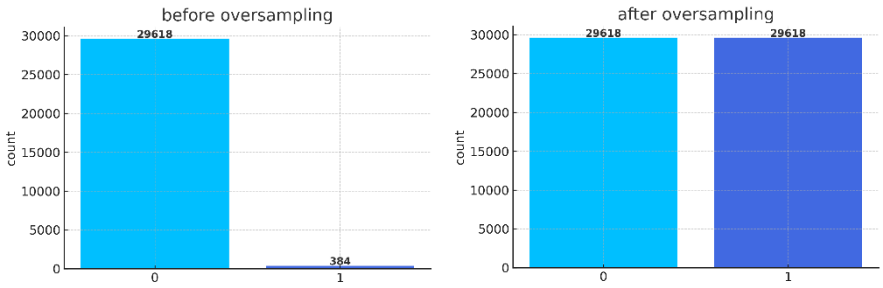
Penelitian ini menggunakan 2 kelas kursus yang memiliki IR tertinggi yaitu HarvardX/CS50x/2012 dengan IR 130.80 dan HarvardX/CB22x/2013_Spring dengan IR 77.13. Setelah prapemrosesan data bersih Kelas HarvardX/CS50x/2012 memiliki jumlah fitur 5 dan jumlah data 169621. Sedangkan untuk kelas HarvardX/CB22x/2013_Spring memiliki jumlah fitur 5 dan jumlah data 24001. Dataset memiliki 2 label kelas yaitu *certified* 1 sebagai kelas minoritas dan kelas *certified* 0 sebagai kelas mayoritas. Data kemudian di split menjadi 80 :20 yang selanjutnya dilakukan modeling. Pada modeling akan diterapkan 4 teknik penanganan ketidakseimbangan data pada model LR.

Teknik penanganan ketidakseimbangan data SMOTE dan OHIT bekerja dengan cara membuat data sintetis atau melakukan *oversampling* untuk menyeimbangkan data pada kelas minoritas seperti yang terlihat pada gambar 1 dan 2. Perbedaan cara kerja SMOTE dan OHIT terletak pada pembuatan data sintetisnya. SMOTE membuat data sintetis berdasar interpolasi linier antara satu sampel minoritas dengan tetangganya sedangkan OHIT melakukan *clustering* pada data minoritas terlebih dahulu, lalu melakukan *oversampling* secara selektif di dalam setiap *cluster*. Selanjutnya untuk teknik CSL dan CW tidak melakukan penambahan data sintetis ke dalam kelas minoritas. CSL dan CW mempertahankan asli namun memberikan sensitivitas lebih tinggi terhadap kelas minoritas. CSL akan memberikan denda lebih besar pada model apabila salah memprediksi kelas minoritas sedangkan CW memberikan bobot yang lebih besar kepada kelas minoritas saat melatih model.



Gambar 2. Hasil *over sampling* kelas kursus HarvardX/CS50x/2012

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v24i4.15005



Gambar 3. Hasil *over sampling* kelas kursus HarvardX/CB22x/2013_Spring

Pada setiap teknik penanganan ketidakseimbangan data yang diterapkan metrik *recall* dapat dilihat pada tabel 1, metrik akurasi dapat dilihat pada tabel 2, dan *margin* peningkatan *recall* dapat dilihat pada tabel 3. Hasil yang diperoleh menunjukkan bahwa dari seluruh teknik penanganan ketidakseimbangan data yang diterapkan, semuanya mampu meningkatkan angka *recall* dari *baseline* model secara signifikan di rata-rata 0.3416 dengan mengorbankan sedikit akurasi di rata-rata -0.0173. Penurunan akurasi ini adalah penurunan yang kecil atau tidak signifikan dibandingkan kenaikan yang dialami pada *recall*. Seluruh teknik penanganan ketidakseimbangan data memiliki *margin* peningkatan *recall* sebesar 0.3243.

Tabel 1. Evaluasi hasil *recall* setiap model

Model	HarvardX /CS50x/2012	HarvardX /CB22x/2013 Spring	Rata-Rata Peningkatan
LR (Baseline)	0.3658	0.9481	Baseline
LR-SMOTE	0.9961	1.0000	0.3411
LR-OHIT	0.9961	1.0000	0.3411
LR-CSL	1.0000	1.0000	0.3430
LR- CW	0.9961	1.0000	0.3411
Seluruh Model	0.3416		

Tabel 2. Evaluasi hasil akurasi setiap model

Model	HarvardX /CS50x/2012	HarvardX /CB22x/2013 Spring	Rata-Rata Penurunan
LR (Baseline)	0.9941	0.9967	Baseline
LR-SMOTE	0.9667	0.9928	-0.0155
LR-OHIT	0.9634	0.9930	-0.0171
LR-CSL	0.9594	0.9910	-0.0201
LR- CW	0.9648	0.9928	-0.0165
Seluruh Model	-0.0173		

Tabel 3. *Margin* peningkatan *recall* setiap model

Model	Margin Peningkatan Recall (Recall + Akurasi)
LR-SMOTE	0.3256
LR-OHIT	0.3240
LR-CSL	0.3229
LR- CW	0.3246
Seluruh Model	0.3243
LR-SMOTE	0.3256

Seluruh teknik penanganan ketidakseimbangan data selalu konsisten dalam meningkatkan nilai *recall* baik pada subset data HarvardX/CS50x/2012 dan HarvardX /CB22x/

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v24i4.15005

2013_Spring. Pada subset data HarvardX/CS50x/2012 *baseline* model hanya mencapai *recall* sebesar 0.3658 dengan akurasi 0.9941. sedangkan setelah menerapkan teknik penanganan ketidakseimbangan data *recall* naik dengan signifikan. Model LR-SMOTE menaikkan *recall* diangka 0.9961 dengan penurunan akurasi diangka 0.9667, LR-OHIT menaikkan *recall* diangka 0.9961 dengan penurunan akurasi diangka 0.9634, LR-CSL menaikkan *recall* diangka 1.0000 dengan penurunan akurasi diangka 0.9594, LR- CW menaikkan *recall* sebesar 0.9961 dengan penurunan akurasi diangka 0.9648. Sedangkan pada subset data HarvardX /CB22x/ 2013_Spring *baseline* model mencapai *recall* sebesar 0.9481 dengan akurasi 0.9967 sedangkan setelah menerapkan teknik penanganan ketidakseimbangan data *recall* naik signifikan dengan angka sempurna sebesar 1.0000 pada semua model. Meski terjadi sedikit penurunan akurasi yang bervariasi pada setiap model. LR-SMOTE terjadi penurunan akurasi diangka 0.9928, LR-OHIT terjadi penurunan akurasi diangka 0.9930, LR-CSL terjadi penurunan akurasi diangka 0.9910, LR- CW terjadi penurunan akurasi diangka 0.9648.

Meski terdapat sedikit penurunan akurasi di seluruh model dan subset data, namun secara keseluruhan teknik penanganan ketidakseimbangan data yang memiliki kinerja paling tinggi adalah LR-SMOTE dengan *margin* peningkatan recall 0.3256. Selanjutnya disusul oleh teknik LR- CW dengan *margin* peningkatan recall 0.3246. Urutan ketiga ada LR-OHIT dengan *margin* peningkatan *recall* 0.3240. Lalu untuk urutan terakhir adalah teknik LR-CSL dengan *margin* peningkatan *recall* 0.3229.

Tabel 4. Evaluasi hasil *precision, f1 score, dan AUC* setiap model

Dataset \ Model		LR (Baseline)	LR-SMOTE	LR-OHIT	LR-CSL	LR- CW
HarvardX /CS50x/2012	Precision	0.7121	0.1849	0.1713	0.1573	0.1766
	F1 Score	0.4833	0.3120	0.2924	0.2719	0.3001
	AUC	0.9952	0.9965	0.9959	0.9963	0.9963
HarvardX /CB22x/2013 _Spring	Precision	0.8202	0.6416	0.6470	0.5877	0.6416
	F1 Score	0.8795	0.7817	0.7857	0.7403	0.7817
	AUC	0.9990	0.9989	0.9989	0.9988	0.9989

Pada table 4 teknik penanganan ketidakseimbangan data seperti LR-SMOTE, LR-OHIT, LR-CSL, dan LR-CW menunjukkan penurunan *precision* dan *f1 score* dibanding model *baseline*. Penurunan ini terjadi karena oversampling menambah lebih banyak contoh kelas minoritas, sehingga model menjadi lebih agresif dalam memprediksi kelas positif. Akibatnya, jumlah *false positive* meningkat dan *precision* menurun. Namun penurunan ini merupakan konsekuensi umum pada dataset dengan ketidakseimbangan ekstrim, di mana prioritas utama adalah meningkatkan kemampuan mendeteksi kelas minoritas. Sementara itu, nilai AUC pada seluruh teknik tetap sangat tinggi dan relatif stabil, menunjukkan bahwa kemampuan model dalam membedakan kelas sebenarnya tidak menurun secara signifikan meskipun precision menurun.

Meski terjadi penurunan di berbagai metrik evaluasi, hasil peningkatan *recall* terjadi di seluruh teknik penanganan ketidakseimbangan data. Hal ini menunjukkan bahwa penanganan ketidakseimbangan data efektif dilakukan pada data yang memiliki ketidakseimbangan ekstrim. Terutama apabila sistem prediksi mengutamakan deteksi kelas minoritas yang jarang terjadi. Tanpa melakukan penanganan ketidakseimbangan data, model akan cenderung bias dan kesulitan dalam mendeteksi kelas minoritas yang tidak memiliki data yang cukup representatif.

4. KESIMPULAN

Berdasarkan hasil penelitian, dapat disimpulkan bahwa penerapan teknik penanganan ketidakseimbangan data pada data yang memiliki ketidakseimbangan data ekstrim efektif dalam meningkatkan *recall* model. Total keseluruhan dampak peningkatan *recall* yang disebabkan oleh seluruh teknik penanganan ketidakseimbangan data adalah meningkatkan *recall* dengan *margin*

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v24i4.15005

peningkatan sebesar 0.3243. Teknik penanganan ketidakseimbangan data yang memiliki kinerja terbaik adalah LR-SMOTE dengan *margin* peningkatan *recall* sebesar 0.3256.

DAFTAR PUSTAKA

[1] Md. A. Talukder, A. S. Talaat, M. Kazi, and A. Khraisat, "XAI-HD: an explainable artificial intelligence framework for heart disease detection," *Artif Intell Rev*, vol. 58, no. 12, p. 385, Oct. 2025, doi: 10.1007/s10462-025-11385-6.

[2] N. F. de Groot, "A contextual integrity approach to genomic information: what bioethics can learn from big data ethics," *Med Health Care Philos*, vol. 27, no. 3, pp. 367–379, Sep. 2024, doi: 10.1007/s11019-024-10211-0.

[3] A. X. Wang, V. T. Le, H. N. Trung, and B. P. Nguyen, "Addressing imbalance in health data: Synthetic minority oversampling using deep learning," *Comput Biol Med*, vol. 188, Apr. 2025, doi: 10.1016/j.compbiomed.2025.109830.

[4] H. Ahmad, B. Kasasbeh, B. Aldabaybah, and E. Rawashdeh, "Class balancing framework for credit card fraud detection based on clustering and similarity-based selection (SBS)," *International Journal of Information Technology (Singapore)*, vol. 15, no. 1, pp. 325–333, Jan. 2023, doi: 10.1007/s41870-022-00987-w.

[5] R. K. L. Kennedy, Z. Salekshahrezaee, F. Villanustre, and T. M. Khoshgoftaar, "Iterative cleaning and learning of big highly-imbalanced fraud data using unsupervised learning," *J Big Data*, vol. 10, no. 1, Dec. 2023, doi: 10.1186/s40537-023-00750-3.

[6] D. Breskuvienė and G. Dzemyda, "Enhancing credit card fraud detection: highly imbalanced data case," *J Big Data*, vol. 11, no. 1, Dec. 2024, doi: 10.1186/s40537-024-01059-5.

[7] M. Sagming, R. Heymann, and M. V. Visaya, "Using topological data analysis and machine learning to predict customer churn," *J Big Data*, vol. 11, no. 1, Dec. 2024, doi: 10.1186/s40537-024-01020-6.

[8] Y. Suh, "Machine learning based customer churn prediction in home appliance rental business," *J Big Data*, vol. 10, no. 1, Dec. 2023, doi: 10.1186/s40537-023-00721-8.

[9] J. B. G. Brito *et al.*, "A framework to improve churn prediction performance in retail banking," *Financial Innovation*, vol. 10, no. 1, Dec. 2024, doi: 10.1186/s40854-023-00558-3.

[10] S. Sarker, M. K. Paul, S. T. H. Thasin, and M. A. M. Hasan, "Analyzing students' academic performance using educational data mining," *Computers and Education: Artificial Intelligence*, vol. 7, Dec. 2024, doi: 10.1016/j.caeai.2024.100263.

[11] A. López-García, O. Blasco-Blasco, M. Liern-García, and S. E. Parada-Rico, "Early detection of students' failure using Machine Learning techniques," *Operations Research Perspectives*, vol. 11, Dec. 2023, doi: 10.1016/j.orp.2023.100292.

[12] M. Yağcı, "Educational data mining: prediction of students' academic performance using machine learning algorithms," *Smart Learning Environments*, vol. 9, no. 1, Dec. 2022, doi: 10.1186/s40561-022-00192-z.

[13] T. Althaqafi, F. Saleem, and A. A. M. AL-Ghamdi, "Enhancing student performance prediction: the role of class imbalance handling in machine learning models," *Discover Computing*, vol. 28, no. 1, Dec. 2025, doi: 10.1007/s10791-025-09576-4.

[14] B. Zhu, X. Jing, L. Qiu, and R. Li, "An Imbalanced Data Classification Method Based on Hybrid Resampling and Fine Cost Sensitive Support Vector Machine," *Computers, Materials and Continua*, vol. 79, no. 3, pp. 3977–3999, 2024, doi: 10.32604/cmc.2024.048062.

[15] S. Matharaarachchi, M. Domaratzki, and S. Muthukumarana, "Enhancing SMOTE for imbalanced data with abnormal minority instances," *Machine Learning with Applications*, vol. 18, p. 100597, Dec. 2024, doi: 10.1016/j.mlwa.2024.100597.

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v24i4.15005

[16] M. Mujahid *et al.*, "Data oversampling and imbalanced datasets: an investigation of performance for machine learning and feature engineering," *J Big Data*, vol. 11, no. 1, Dec. 2024, doi: 10.1186/s40537-024-00943-4.
[17] P. M. Vieira and F. Rodrigues, "An automated approach for binary classification on imbalanced data," *Knowl Inf Syst*, vol. 66, no. 5, pp. 2747–2767, May 2024, doi: 10.1007/s10115-023-02046-7.
[18] W. Chen, K. Yang, Z. Yu, Y. Shi, and C. L. P. Chen, "A survey on imbalanced learning: latest research, applications and future directions," *Artif Intell Rev*, vol. 57, no. 6, Jun. 2024, doi: 10.1007/s10462-024-10759-6.
[19] M. Büttner *et al.*, "Conquering class imbalances in deep learning-based segmentation of dental radiographs with different loss functions," *J Dent*, vol. 148, Sep. 2024, doi: 10.1016/j.jdent.2024.105063.
[20] B. Kim, Y. Ko, and J. Seo, "Novel regularization method for the class imbalance problem," *Expert Syst Appl*, vol. 188, Feb. 2022, doi: 10.1016/j.eswa.2021.115974.
[21] T. Zhu, C. Luo, Z. Zhang, J. Li, S. Ren, and Y. Zeng, "Minority oversampling for imbalanced time series classification," *Knowl Based Syst*, vol. 247, Jul. 2022, doi: 10.1016/j.knosys.2022.108764.
[22] O. Volk and G. Singer, "An adaptive cost-sensitive learning approach in neural networks to minimize local training–test class distributions mismatch," *Intelligent Systems with Applications*, vol. 21, Mar. 2024, doi: 10.1016/j.iswa.2023.200316.
[23] I. Araf, A. Idri, and I. Chairi, "Cost-sensitive learning for imbalanced medical data: a review," *Artif Intell Rev*, vol. 57, no. 4, Apr. 2024, doi: 10.1007/s10462-023-10652-8.
[24] M. Salmi, D. Atif, D. Oliva, A. Abraham, and S. Ventura, "Handling imbalanced medical datasets: review of a decade of research," *Artif Intell Rev*, vol. 57, no. 10, Oct. 2024, doi: 10.1007/s10462-024-10884-2.
[25] D. Elreedy, A. F. Atiya, and F. Kamalov, "A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning," *Mach Learn*, vol. 113, no. 7, pp. 4903–4923, Jul. 2024, doi: 10.1007/s10994-022-06296-4.
[26] A. R. Ghavamipour, F. Turkmen, and X. Jiang, "Privacy-preserving logistic regression with secret sharing," *BMC Med Inform Decis Mak*, vol. 22, no. 1, Dec. 2022, doi: 10.1186/s12911-022-01811-y.
[27] K. Hwang, W. Kang, and Y. Jung, "Application of the class-balancing strategies with bootstrapping for fitting logistic regression models for post-fire tree mortality in South Korea," *Environ Ecol Stat*, vol. 30, no. 3, pp. 575–598, Sep. 2023, doi: 10.1007/s10651-023-00573-8.
[28] M. Ismail, S. Alrabaee, K. K. R. Choo, L. Ali, and S. Harous, "A Comprehensive Evaluation of Machine Learning Algorithms for Web Application Attack Detection with Knowledge Graph Integration," *Mobile Networks and Applications*, vol. 29, no. 3, pp. 1008–1037, Jun. 2024, doi: 10.1007/s11036-024-02367-z.
[29] D. Ananda Agustina Pertiwi and M. Aziz Muslim, "Improvement accuracy of gradient boosting in app rating prediction on google playstore," *Journal of Numerical Optimization and Technology Management*, vol. 1, no. 2, p. 2023, doi: 10.00000/jnotm.0000.00.00.000.
[30] M. Jebbari, B. Cherradi, S. Hamida, and A. Raihani, "Identifying learning styles in MOOCs environment through machine learning predictive modeling," *Educ Inf Technol (Dordr)*, vol. 29, no. 16, pp. 20977–21014, Nov. 2024, doi: 10.1007/s10639-024-12637-8.

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v24i4.15005