UNP JOURNAL OF STATISTICS AND DATA SCIENCE
Volume 4 Nomor 3 Tahun 2026 Halaman 417-424.

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

Credit Card Fraud Detection under Extreme Class Imbalance: A Comparison of KNN and Logistic Regression

Felicia Sword dan Christopher Andreas*
Fakultas Teknologi Informasi, Universitas Ciputra, Surabaya, Indonesia
*Corresponding author: christopher.andreas@ciputra.ac.id

Submitted : 24 Juni 2026
Revised : 15 Juli 2026
Accepted : 31 Agustus 2026

ABSTRACT

Credit card fraud is a serious threat in the digital financial ecosystem and is characterised by extreme class imbalance, with fraudulent transactions typically below 1%. This study compares two standard classification algorithms, K-Nearest Neighbor (KNN) and Logistic Regression (LR), for detecting fraudulent transactions on the Sparkov dataset (1.85 million transactions; a stratified subsample of 100,000 rows; 0.52% fraud rate), and analyses the effect of the Synthetic Minority Over-sampling Technique (SMOTE). Preprocessing includes temporal feature engineering, haversine distance, leak-free per-card behavioural features, one-hot and label encoding, and z-score standardisation. Models are evaluated on a stratified 80:20 split using the confusion matrix, accuracy, precision, recall, F1-score, ROC-AUC, and PR-AUC, complemented by decision-threshold tuning, confidence intervals over five repetitions, and the McNemar test. No single model dominates across all metrics. At the default threshold, KNN baseline achieves the highest F1 (0.407) and precision (0.540), while LR baseline achieves the highest PR-AUC (0.246); LR+SMOTE leads on recall (0.712) and ROC-AUC (0.861) but with very low precision (0.025). Threshold tuning lets LR baseline reach the best F1 (0.422 at a 0.071 cut-off). McNemar shows the KNN–LR difference is not significant at baseline ($p = 0.282$) but significant under SMOTE ($p < 0.001$). The main finding is that under severe imbalance ROC-AUC can be misleading and PR-AUC is more informative; KNN baseline is a balanced detector without tuning, threshold-tuned LR baseline gives the best single operating point, and LR+SMOTE suits cases where recall is the priority.

Keywords: *Credit Card Fraud Detection, Class Imbalance, K-Nearest Neighbor, Logistic Regression, SMOTE.*

I. PENDAHULUAN

Penipuan kartu kredit merupakan salah satu ancaman paling signifikan dalam ekosistem keuangan digital global. Menurut Federal Trade Commission (FTC), kerugian individu akibat penipuan keuangan di Amerika Serikat melampaui USD 8,8 miliar pada tahun 2022, meningkat 30% dibandingkan tahun sebelumnya (Federal Trade Commission, 2023). Di Indonesia, Otoritas Jasa Keuangan (OJK) mencatat kerugian masyarakat akibat penipuan keuangan digital mencapai Rp 7,8 triliun selama November 2024, sementara kejahatan siber melonjak 550 persen pada periode yang sama (OJK, 2026). Seiring pesatnya pertumbuhan layanan keuangan digital, deteksi penipuan menjadi komponen penting dalam sistem keamanan keuangan modern. Sistem deteksi berbasis aturan (*rule-based systems*) yang secara tradisional digunakan oleh institusi keuangan sering kali menghadapi kesulitan untuk beradaptasi terhadap pola penipuan yang terus berkembang (Veldurthi, 2025). Oleh karena itu, pendekatan berbasis *machine learning* semakin banyak diterapkan karena kemampuannya mempelajari pola transaksi yang kompleks dari data historis. Namun, penerapan *machine learning* dalam deteksi penipuan menghadirkan tantangan yang tidak sederhana, terutama terkait karakteristik distribusi data yang sangat tidak seimbang.

Tantangan utama dalam deteksi penipuan adalah ketidakseimbangan kelas yang ekstrem: proporsi transaksi penipuan pada data nyata umumnya kurang dari 1%, sehingga model klasifikasi cenderung bias terhadap kelas mayoritas (transaksi sah). Akibatnya, akurasi menjadi metrik yang menyesatkan, dimana model yang selalu memprediksi "bukan penipuan" dapat mencapai akurasi di atas 99% namun gagal total mendeteksi penipuan (Thölke dkk., 2023). Untuk mengatasi permasalahan tersebut, berbagai teknik penanganan ketidakseimbangan telah dikembangkan, salah satunya adalah *Synthetic Minority Over-sampling Technique* (SMOTE) yang menghasilkan

DOI: https://doi.org/10.24036/ujsds/vol4-iss3/539

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

sampel sintetis pada kelas minoritas guna memperbaiki proses pembelajaran model (Chawla dkk., 2002). Selain memengaruhi proses pelatihan model, ketidakseimbangan kelas juga menimbulkan tantangan dalam proses evaluasi. Pada data yang sangat tidak seimbang, metrik yang umum digunakan dapat memberikan gambaran performa yang terlalu optimistis. Oleh karena itu, kurva *Precision–Recall* dan nilai *Area Under the Precision–Recall Curve* (*PR-AUC*) dianggap lebih informatif dibandingkan kurva *Receiver Operating Characteristic* (*ROC*) karena dapat tetap menunjukkan performa yang tampak baik meskipun presisi model rendah (Saito dan Rehmsmeier, 2015). Dengan demikian, evaluasi model pada kasus deteksi penipuan memerlukan perhatian tidak hanya terhadap algoritma yang digunakan, tetapi juga terhadap strategi penanganan ketidakseimbangan dan pemilihan metrik evaluasi yang tepat.

Penelitian deteksi penipuan kerap terkendala oleh keterbatasan akses terhadap data transaksi nyata karena alasan kerahasiaan dan regulasi. Untuk mengatasi kendala tersebut, sejumlah simulator dan dataset sintetis telah dikembangkan, antara lain *PaySim* untuk transaksi uang elektronik dan *Sparkov Data Generation Tool* yang menghasilkan dataset transaksi kartu kredit yang digunakan pada penelitian ini (Lopez-Rojas dkk., 2016; Harris, 2016; Kaggle/Shenoy, 2020). Upaya lain dilakukan oleh Grover dkk. (2022) melalui *Fraud Dataset Benchmark* yang menstandarkan evaluasi lintas dataset penipuan dan menekankan pentingnya protokol evaluasi yang konsisten. Meskipun demikian, sebagian besar penelitian berfokus pada pengembangan model yang semakin kompleks, sementara studi yang melakukan perbandingan sistematis terhadap algoritma klasifikasi klasik menggunakan protokol evaluasi yang konsisten, termasuk analisis dampak penanganan ketidakseimbangan melalui SMOTE, masih relatif terbatas. Padahal, pemahaman yang jelas mengenai kinerja model-model dasar tetap penting karena model tersebut sering kali lebih mudah diinterpretasikan, lebih ringan secara komputasi, dan lebih realistis untuk diimplementasikan dalam lingkungan operasional (Assis dkk, 2024).

Lebih lanjut, masih terdapat kebutuhan akan evaluasi yang sistematis terhadap algoritma klasifikasi klasik pada skenario deteksi penipuan dengan ketidakseimbangan kelas yang ekstrem. Secara khusus, belum banyak penelitian yang secara bersamaan mengevaluasi kinerja relatif *K-Nearest Neighbor* (KNN) dan *Logistic Regression* (LR), serta menelaah bagaimana penerapan SMOTE memengaruhi performa kedua algoritma tersebut di bawah berbagai metrik evaluasi. Berdasarkan kesenjangan tersebut, penelitian ini bertujuan untuk membandingkan kinerja KNN dan LR dalam mendeteksi transaksi penipuan kartu kredit pada data yang sangat tidak seimbang, serta menganalisis pengaruh penerapan SMOTE terhadap kinerja relatif kedua algoritma. Untuk mencapai tujuan tersebut, penelitian ini menerapkan protokol evaluasi yang konsisten pada dataset Sparkov dengan mempertimbangkan berbagai metrik performa, penyetelan ambang keputusan, dan pengujian signifikansi statistik. Penelitian ini berkontribusi pada literatur deteksi penipuan kartu kredit melalui evaluasi yang sistematis dan transparan terhadap pengaruh pilihan algoritma, penerapan SMOTE, pemilihan metrik evaluasi, dan penyetelan ambang keputusan pada data yang sangat tidak seimbang. Secara khusus, penelitian ini: (1) menyajikan perbandingan empiris antara KNN dan LR di bawah protokol evaluasi yang konsisten; (2) menganalisis pengaruh penerapan SMOTE terhadap masing-masing algoritma; (3) mengevaluasi kinerja model, penyetelan ambang keputusan, dan pengujian signifikansi statistik untuk menghasilkan interpretasi yang lebih andal; serta (4) menunjukkan bahwa keunggulan model bergantung pada metrik dan titik operasi yang digunakan, sehingga pemilihan strategi pengambilan keputusan menjadi sama pentingnya dengan pemilihan algoritma itu sendiri.

II. METODE PENELITIAN

Bagian ini menguraikan teori yang relevan, sumber data dan variabel penelitian, praproses dan rekayasa fitur, penanganan ketidakseimbangan kelas, model beserta pemilihan hiperparameter, dan skema evaluasi.

A. Dataset, Praproses, dan Rekayasa Fitur

Penelitian menggunakan *Credit Card Transactions Fraud Detection Dataset* yang dibangkitkan dengan *Sparkov* (Kaggle/Shenoy, 2020; Harris, 2016), terdiri atas 1.852.394 transaksi periode 2019–2020 dengan tingkat penipuan yang sangat rendah. Demi efisiensi komputasi, diambil subsampel terstratifikasi sebanyak 100.000 transaksi yang mempertahankan tingkat penipuan asli (0,52%; 521 transaksi penipuan), dengan variabel dirangkum pada Tabel 1. Fitur temporal (*hour*, *day of week, is weekend*) diekstraksi dari waktu transaksi, dan *age* dihitung dari tanggal lahir. Jarak geografis antara pemegang kartu dan *merchant* dihitung menggunakan rumus *haversine*. Selain itu, tiga fitur perilaku dibentuk untuk menangkap pola penyalahgunaan: *amt zscore card* (seberapa tidak biasa nilai transaksi relatif terhadap riwayat kartu), *hours since previous*, dan *transaction_24h*. Fitur-fitur ini dihitung pada seluruh data (1,85 juta transaksi) sebelum penarikan *subsample*, karena memerlukan riwayat lengkap tiap kartu dan bersifat *leak-free* yaitu setiap transaksi hanya menggunakan riwayatnya sendiri. Variabel kategorik *category* dan *day*

DOI: https://doi.org/10.24036/ujsds/vol4-iss3/539

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

of week disandikan dengan *one-hot encoding* (*drop-first* untuk menghindari multikolinearitas), *gender* dengan label *encoding*, dan seluruh fitur numerik distandardisasi menggunakan *z-score* (*StandardScaler*). Standardisasi sangat penting bagi KNN yang berbasis jarak (Yusran dkk, 2025).

Tabel 1. Variabel penelitian

Tipe	Variabel	Deskripsi
Dependen (Y)	*Fraud*	Biner: 1 = penipuan, 0 = sah
Independen (X1)	*Category*	Kategori *merchant* (14 kategori, *one-hot*)
Independen (X2)	*Gender*	Jenis kelamin pemegang kartu (label *encoding)*
Independen (X3)	*Age*	Usia pemegang kartu (diturunkan dari tanggal lahir)
Independen (X4)	*City Population*	Populasi kota pemegang kartu
Independen (X5)	*Hour*	Jam transaksi (0–23)
Independen (X6)	*Day of Week*	Hari transaksi (*one-hot*)
Independen (X7)	*Is Weekend*	Indikator akhir pekan (1 = *Weekend;* 0 = Bukan *Weekend*)
Independen (X8)	*Distance*	Jarak *haversine* pemegang kartu dengan *merchant* (km)
Independen (X9)	*Amt Z-score Card*	Nilai transaksi relatif terhadap riwayat kartu (*z-score*)
Independen (X10)	*Hours since Previous*	Jam sejak transaksi sebelumnya pada kartu
Independen (X11)	*Transaction_24h*	Jumlah transaksi kartu dalam 24 jam terakhir
Independen (X12)	Amt	Nilai transaksi (USD) — berdasarkan hasil analisis data eksploratif, variabel ini dikecualikan dari model (menghindari multikolinearitas)

B. Penanganan Ketidakseimbangan Kelas

Dalam hal ini, terdapat dua kondisi yang dibandingkan: (a) *baseline* tanpa *resampling*, dan (b) penanganan ketidakseimbangan kelas melalui SMOTE. SMOTE hanya diterapkan pada data latih untuk membangkitkan sampel kelas minoritas sintetis; data uji dibiarkan tidak dimodifikasi agar evaluasi tetap mencerminkan distribusi nyata (Chawla dkk., 2002). SMOTE dipilih karena merupakan teknik *oversampling* yang paling banyak digunakan dan menjadi standar untuk klasifikasi dengan variabel target yang distribusinya tidak seimbang.

C. Model dan Pemilihan Hiperparameter

Penelitian ini membandingkan dua algoritma klasifikasi, yaitu *K-Nearest Neighbor* (KNN) dan *Logistic Regression* (LR), yang masing-masing dievaluasi pada kondisi *baseline* dan setelah penerapan SMOTE. Dengan demikian, diperoleh empat varian model, yaitu KNN *Baseline*, KNN+SMOTE, LR *Baseline*, dan LR+SMOTE. Kedua algoritma dipilih karena merepresentasikan dua paradigma pembelajaran yang berbeda dan saling melengkapi. KNN merupakan metode non-parametrik yang melakukan klasifikasi berdasarkan kedekatan lokal dalam ruang fitur, sedangkan LR merupakan metode parametrik yang membentuk batas keputusan global dan menghasilkan estimasi probabilitas kelas. Perbandingan antara pendekatan berbasis jarak dan pendekatan berbasis probabilitas memungkinkan analisis yang lebih jelas mengenai bagaimana karakteristik algoritma memengaruhi kinerja deteksi penipuan pada data dengan ketidakseimbangan kelas yang ekstrem, baik sebelum maupun sesudah penerapan SMOTE. Selain itu, kedua algoritma merupakan metode dasar (*baseline*) yang banyak digunakan dalam literatur sehingga dapat berfungsi sebagai titik rujukan yang kuat untuk evaluasi model yang lebih kompleks.

Nilai parameter k pada KNN ditentukan melalui validasi silang yang terstratifikasi lima lipatan (*5-fold stratified cross-validation*) pada data latih dengan menggunakan F1-score kelas penipuan sebagai kriteria pemilihan. Pencarian parameter dilakukan pada himpunan kandidat $k \in \{1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21\}$. Sementara itu, model LR dilatih menggunakan pengaturan bawaan pustaka *scikit-learn*, yaitu regularisasi L2 (*ridge*), parameter kekuatan regularisasi invers C = 1,0, solver *lbfgs*, dan tanpa pembobotan kelas (*class weighting*) (Pedregosa dkk., 2011). Jumlah iterasi maksimum ditingkatkan menjadi 1.000 untuk memastikan proses optimasi mencapai konvergensi. Seluruh eksperimen menggunakan nilai *random seed* sebesar 42 guna menjamin reproduktibilitas hasil penelitian.

D. Skema Evaluasi

Data dibagi secara terstratifikasi dengan rasio 80:20 menjadi data latih dan data uji. Kinerja model dievaluasi menggunakan *confusion matrix*, *accuracy, precision, recall, F1-score, ROC-AUC*, dan *PR-AUC*. Mengingat tujuan

DOI: https://doi.org/10.24036/ujsds/vol4-iss3/539

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

utama penelitian adalah mendeteksi transaksi penipuan yang merupakan kelas minoritas, analisis perbandingan terutama difokuskan pada *recall, F1-score*, dan *PR-AUC*. Untuk meningkatkan keandalan evaluasi pada data dengan ketidakseimbangan kelas yang ekstrem, dilakukan tiga prosedur tambahan. Pertama, dilakukan penyetelan ambang keputusan (*decision threshold tuning*). Ambang yang memaksimalkan *F1-score* ditentukan menggunakan *5-fold stratified cross-validation* pada data latih melalui prediksi *out-of-fold*, kemudian diterapkan satu kali pada data uji untuk memperoleh estimasi yang tidak bias. Kedua, kestabilan kinerja model dievaluasi melalui lima pembagian data acak yang berbeda, dan hasilnya dilaporkan dalam bentuk rata-rata dan simpangan baku melalui selang kepercayaan. Ketiga, uji signifikansi *McNemar* digunakan untuk mengevaluasi apakah perbedaan kinerja klasifikasi antara dua model bersifat signifikan secara statistik. Pengujian dilakukan baik untuk membandingkan algoritma yang berbeda (KNN *versus* LR) maupun untuk menilai dampak penerapan SMOTE pada algoritma yang sama (*baseline versus* SMOTE).

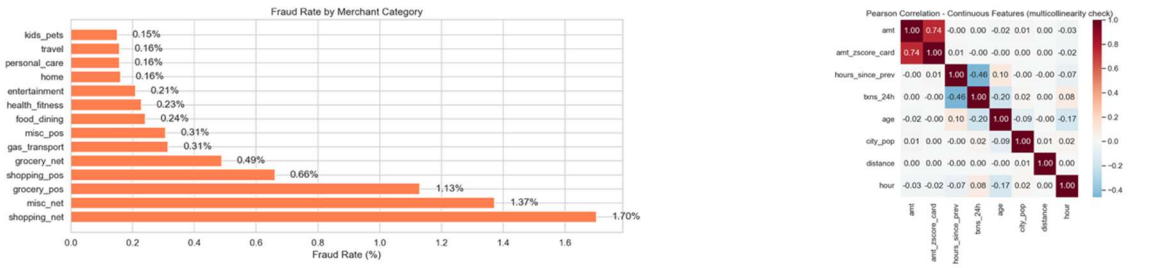
III. HASIL DAN PEMBAHASAN

Bagian ini menyajikan hasil analisis data eksploratif, tahap pemodelan dan evaluasi, penyetelan ambang keputusan, hingga pembahasan temuan utama dari penelitian ini.

A. Analisis Data Eksploratif

Dataset penelitian memiliki ketidakseimbangan kelas yang sangat ekstrem, dengan hanya 521 dari 100.000 transaksi (0,52%) yang berlabel sebagai penipuan (Gambar 1). Kondisi ini mengindikasikan bahwa metrik akurasi berpotensi memberikan gambaran kinerja yang menyesatkan karena dapat didominasi oleh kelas mayoritas. Oleh karena itu, evaluasi model dalam penelitian ini lebih difokuskan pada metrik yang sensitif terhadap kinerja pada kelas minoritas, yaitu *recall, F1-score*, dan *PR-AUC*. Selain itu, tingkat penipuan menunjukkan variasi yang cukup besar antar kategori *merchant* (lihat Gambar 2a), dimana transaksi pada kategori belanja daring dan beberapa kategori lainnya memiliki proporsi penipuan yang relatif lebih tinggi dibandingkan kategori lain. Berdasarkan analisis diskriminasi menggunakan AUC per fitur, yang relatif *robust* terhadap ketidakseimbangan kelas, fitur *amt* memiliki kemampuan pemisahan tertinggi (AUC = 0,845), diikuti oleh *amt zscore card* (AUC = 0,802). Fitur *hours since previous* juga menunjukkan kemampuan diskriminatif yang moderat (AUC = 0,649), yang mengindikasikan bahwa transaksi penipuan cenderung terjadi dalam interval waktu yang lebih dekat dengan transaksi sebelumnya. Sebaliknya, fitur *transaction_24h* hanya memiliki kemampuan pemisahan yang terbatas (AUC = 0,549). Selanjutnya, dilakukan pemeriksaan multikolinieritas antar fitur melalui analisis korelasi pada Gambar 2b. Berdasarkan Gambar 2b, diperoleh bahwa terdapat korelasi yang relatif tinggi antara *amt* dan *amt zscore card* (r = 0,74) mengindikasikan adanya hubungan linear yang kuat di antara kedua fitur tersebut. Oleh sebab itu, variabel *amt* dikeluarkan dari model.

Gambar 1. Distribusi kelas transaksi.



(a) (b)

Gambar 2. (a) Tingkat penipuan menurut kategori *merchant* dan (b) Matriks korelasi Pearson antarfitur numerik

DOI: https://doi.org/10.24036/ujsds/vol4-iss3/539

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

B. Hasil Pemodelan dan Perbandingan Kinerja Model

Pada pemodelan KNN, hasil validasi silang menunjukkan bahwa nilai optimal parameter (k) adalah 1, baik pada kondisi *baseline* maupun setelah penerapan SMOTE. Nilai *F1-score* cenderung menurun seiring meningkatnya nilai (k), sehingga k = 1 memberikan kinerja terbaik pada data latih. Fenomena ini terkait dengan tingkat ketidakseimbangan kelas yang sangat tinggi. Ketika nilai k meningkat, lingkungan tetangga suatu observasi semakin didominasi oleh transaksi sah sebagai kelas mayoritas, sehingga peluang kelas penipuan untuk memperoleh mayoritas suara menjadi semakin kecil. Akibatnya, sensitivitas model terhadap kelas minoritas cenderung menurun. Pemilihan k = 1 juga menyebabkan skor probabilitas yang dihasilkan KNN menjadi sangat diskret, dengan nilai yang umumnya mendekati 0 atau 1. Karakteristik ini membatasi fleksibilitas penyetelan ambang keputusan (*threshold tuning*), sebagaimana dibahas lebih lanjut pada Subbagian C. Sebaliknya, untuk pemodelan LR dapat dilakukan dengan parameter yang sudah ditetapkan. Tabel 2 menyajikan kinerja keempat model pada ambang keputusan default sebesar 0,5. Pada kondisi ini, KNN baseline memperoleh *F1-score* tertinggi, yaitu 0,407 pada pembagian utama (*seed* 42), serta mencapai *precision* tertinggi sebesar 0,540. Sementara itu, LR *baseline* menghasilkan nilai PR-AUC tertinggi, yaitu 0,246. Di sisi lain, LR+SMOTE mencapai recall tertinggi (0,712) dan ROC-AUC tertinggi (0,861), tetapi *precision* yang dihasilkan sangat rendah (0,025), sehingga nilai PR-AUC-nya juga relatif rendah (0,192). Selain itu, analisis *confusion matrix* pada menunjukkan bahwa kinerja *recall* yang tinggi pada LR+SMOTE diperoleh dengan konsekuensi meningkatnya jumlah positif palsu secara substansial. Pada data uji, model ini menghasilkan 2.852 positif palsu dibandingkan dengan 74 penipuan yang berhasil terdeteksi dengan benar, yang setara dengan sekitar 39 *alarm* palsu untuk setiap satu penipuan yang berhasil ditangkap. Temuan ini menunjukkan adanya *trade-off* yang kuat antara *recall* dan *precision* pada kondisi tersebut. Sebaliknya, LR *baseline* hanya mencapai *recall* sebesar 0,010 pada ambang *default,* yang mengindikasikan bahwa sebagian besar transaksi penipuan tidak berhasil diidentifikasi. Hasil ini menggambarkan keterbatasan penggunaan ambang keputusan standar pada data dengan ketidakseimbangan kelas yang ekstrem dan menjadi dasar dilakukannya penyetelan ambang keputusan yang dibahas pada Subbagian C.

Tabel 2. Perbandingan kinerja keempat model (ambang default 0,5)

Model	Accuracy	Precision	Recall	F1	ROC-AUC	PR-AUC
KNN Baseline	0,9950	0,5397	0,3269	0,4072	0,6627	0,1799
KNN + SMOTE	0,9930	0,3622	0,4423	0,3983	0,7191	0,1631
LR Baseline	0,9946	0,1429	0,0096	0,0180	0,8050	0,2457
LR + SMOTE	0,8559	0,0253	0,7115	0,0488	0,8607	0,1918

Kurva *ROC* pada Gambar 3a menunjukkan bahwa model LR memiliki kemampuan pemeringkatan (*ranking*) yang lebih baik dibandingkan model KNN. Temuan ini juga tercermin pada kurva *Precision–Recall* (Gambar 3b), yang lebih informatif untuk data dengan ketidakseimbangan kelas yang ekstrem. Berdasarkan nilai *Average Precision* (AP), LR baseline memperoleh kinerja terbaik (0,246), diikuti oleh LR+SMOTE (0,192), sedangkan KNN *baseline* dan KNN+SMOTE masing-masing mencapai AP sebesar 0,180 dan 0,163. Meskipun demikian, KNN *baseline* tetap menghasilkan *F1-score* tertinggi pada ambang keputusan *default*. Hasil ini menunjukkan bahwa kinerja suatu model pada metrik berbasis ambang tidak selalu sejalan dengan kualitas pemeringkatannya. Dalam kasus ini, titik operasi yang dihasilkan oleh KNN *baseline* pada ambang default memberikan keseimbangan *precision* dan *recall* yang lebih baik dibandingkan model lainnya.

(a) (b)

Gambar 3. (a) Kurva *ROC* dan (b) Kurva *Precision-Recall* dari keempat model.

DOI: https://doi.org/10.24036/ujsds/vol4-iss3/539

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

C. Penyetelan Ambang Keputusan dan KNN dengan Pembobotan Jarak (*Distance-Weighted* KNN)

Ambang keputusan *default* sebesar 0,5 tidak selalu optimal pada data dengan ketidakseimbangan kelas yang tinggi karena secara implisit mengasumsikan distribusi kelas dan biaya kesalahan yang relatif seimbang. Dalam konteks deteksi penipuan, kondisi tersebut dapat menyebabkan model menjadi terlalu konservatif dalam memberikan label penipuan. Oleh karena itu, dilakukan penyetelan ambang keputusan dengan memilih nilai yang memaksimalkan *F1-score*. Untuk menghindari bias evaluasi, pemilihan ambang dilakukan menggunakan data latih melalui validasi silang, kemudian ambang terpilih diterapkan satu kali pada data uji. Tabel 3 menunjukkan kinerja model setelah penyetelan ambang keputusan. Pada LR *baseline*, ambang optimal yang diperoleh adalah sekitar 0,071, yang meningkatkan *F1-score* dari 0,018 menjadi 0,422. Peningkatan yang substansial ini menunjukkan bahwa rendahnya kinerja LR pada ambang *default* terutama disebabkan oleh pemilihan titik operasi yang kurang sesuai untuk data dengan ketidakseimbangan kelas yang ekstrem, bukan karena keterbatasan kemampuan diskriminatif model. Setelah penyetelan ambang, LR *baseline* menghasilkan *F1-score* tertinggi di antara seluruh model yang dievaluasi, sehingga dapat dianggap sebagai titik operasi terbaik dalam konteks penelitian ini. Sebaliknya, kinerja KNN tidak mengalami perubahan setelah penyetelan ambang keputusan. Hal ini disebabkan oleh karakteristik skor probabilitas yang dihasilkan pada k = 1, yang umumnya hanya mengambil nilai 0 atau 1 sehingga tidak menyediakan fleksibilitas yang cukup untuk menggeser titik operasi melalui perubahan ambang keputusan.

Tabel 3. Hasil penyetelan ambang.

Model	Ambang	*Precision*	*Recall*	*F1*
LR *Baseline*	0,071	0,500	0,365	0,422
LR + SMOTE	0,913	0,356	0,404	0,378

Catatan: Model KNN tidak disertakan karena tidak mengalami perubahan.

Untuk mengatasi keterbatasan skor probabilitas yang sangat diskret pada KNN dengan k = 1, dilakukan eksperimen tambahan menggunakan pembobotan jarak (*weights* = *distance*). Ketika nilai k dipilih kembali melalui validasi silang, hasil yang diperoleh tetap k = 1, sehingga kinerja model identik dengan KNN berbobot seragam. Temuan ini dapat dijelaskan karena pada k = 1 hanya terdapat satu tetangga yang berkontribusi pada proses klasifikasi, sehingga skema pembobotan tidak memengaruhi hasil prediksi. Selanjutnya, dilakukan pengujian pada nilai k yang lebih besar (k = 21) untuk mengevaluasi pengaruh pembobotan jarak terhadap karakteristik skor probabilitas. Pada konfigurasi ini, pembobotan jarak menghasilkan probabilitas yang jauh lebih kontinu, dengan 1.296 nilai probabilitas unik dibandingkan hanya 22 nilai unik pada pembobotan seragam. Karakteristik tersebut memungkinkan dilakukannya penyetelan ambang keputusan secara lebih efektif. Setelah penyetelan ambang, KNN berbobot jarak mencapai *F1-score* sebesar 0,440, sedikit lebih tinggi dibandingkan KNN berbobot seragam (0,434), dan sebanding dengan kinerja terbaik yang diperoleh model-model lain dalam penelitian ini. Meskipun demikian, keunggulan utama pembobotan jarak tidak terletak pada peningkatan *F1-score* yang relatif kecil, melainkan pada kemampuannya menghasilkan skor probabilitas yang lebih halus (*continuous probability scores*). Dengan demikian, model KNN menjadi lebih fleksibel untuk disesuaikan dengan berbagai titik operasi melalui penyetelan ambang keputusan. Sebaliknya, KNN dengan k = 1 menghasilkan probabilitas yang hampir sepenuhnya biner sehingga ruang untuk penyesuaian ambang menjadi sangat terbatas. Fakta bahwa validasi silang tetap memilih k = 1 menunjukkan bahwa konfigurasi tersebut memang memberikan *F1-score* terbaik pada ambang *default*, meskipun tidak menghasilkan probabilitas yang ideal untuk proses kalibrasi dan penyetelan ambang berikutnya.

D. Analisis Kestabilan dan Pengujian Signifikansi

Selang kepercayaan yang dihitung dari lima pembagian data acak (Tabel 4) menunjukkan bahwa hasil evaluasi relatif stabil. Sebagai contoh, KNN+SMOTE memperoleh *F1-score* sebesar 0,337 ± 0,036, sedangkan LR+SMOTE mencapai *recall* sebesar 0,692 ± 0,018. Nilai simpangan baku yang relatif kecil mengindikasikan bahwa pola perbedaan kinerja antar model konsisten pada berbagai pembagian data.

Tabel 4. Analisis Kestabilan Melalui Selang Kepercayaan

Model	*F1*	*Recall*	*ROC-AUC*	*PR-AUC*
KNN *Baseline*	0,385 ± 0,029	0,310 ± 0,027	0,654 ± 0,013	0,162 ± 0,022
KNN + SMOTE	0,337 ± 0,036	0,392 ± 0,042	0,694 ± 0,021	0,121 ± 0,025
LR *Baseline*	0,004 ± 0,007	0,002 ± 0,004	0,775 ± 0,021	0,228 ± 0,037

DOI: https://doi.org/10.24036/ujsds/vol4-iss3/539

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

Model	*F1*	*Recall*	*ROC-AUC*	*PR-AUC*
LR + SMOTE	0,047 ± 0,002	0,692 ± 0,018	0,836 ± 0,016	0,155 ± 0,026

Uji *McNemar* (Tabel 5) digunakan untuk mengevaluasi signifikansi statistik perbedaan prediksi antar model. Pada kondisi *baseline*, perbedaan antara KNN dan LR tidak signifikan secara statistik (p = 0,282). Sebaliknya, setelah penerapan SMOTE, perbedaan antara kedua algoritma menjadi signifikan (p < 0,0001). Namun, hasil tersebut perlu diinterpretasikan dengan mempertimbangkan karakteristik kesalahan yang dihasilkan masing-masing model. Pada LR+SMOTE, peningkatan kemampuan mendeteksi transaksi penipuan disertai peningkatan jumlah positif palsu yang sangat besar (sekitar 2.850 kasus), sehingga signifikansi statistik yang diperoleh tidak sepenuhnya mencerminkan peningkatan kualitas keputusan klasifikasi secara keseluruhan. Analisis dalam satu algoritma menunjukkan bahwa penerapan SMOTE mengubah pola prediksi secara signifikan baik pada KNN maupun LR (p < 0,0001). Pada KNN, perubahan tersebut relatif terbatas, dengan 52 prediksi yang sebelumnya benar berubah menjadi salah dan hanya 12 prediksi yang berubah dari salah menjadi benar. Sebaliknya, pada LR, penerapan SMOTE menghasilkan perubahan yang jauh lebih besar dan terutama ditandai oleh peningkatan jumlah positif palsu, dengan sekitar 2.846 alarm tambahan untuk memperoleh 73 kasus penipuan yang berhasil dideteksi. Secara keseluruhan, hasil analisis kestabilan dan pengujian statistik mendukung kesimpulan bahwa tidak terdapat satu model yang secara konsisten unggul pada seluruh kriteria evaluasi. Perbedaan *F1-score* antar model terdepan berada pada kisaran yang sebanding dengan variasi hasil antar pembagian data (sekitar ±0,03), sementara perbandingan langsung antara KNN dan LR pada kondisi *baseline* tidak menunjukkan perbedaan yang signifikan secara statistik.

Tabel 5. Hasil uji signifikansi McNemar

Perbandingan	χ^2	p-value	Kesimpulan
LR *Baseline* vs KNN *Baseline*	1,157	0,2821	Tidak signifikan
LR + SMOTE vs KNN + SMOTE	2593,503	< 0,0001	Signifikan (KNN lebih sedikit galat)
KNN *Baseline* vs KNN + SMOTE	23,766	< 0,0001	Signifikan (SMOTE mengubah prediksi)
LR *Baseline* vs LR + SMOTE	2632,403	< 0,0001	Signifikan (SMOTE mengubah prediksi)

E. Pembahasan

Temuan utama yang diperoleh adalah pemilihan metrik evaluasi dan ambang keputusan berpengaruh besar terhadap interpretasi kinerja model. KNN dengan k = 1 menghasilkan *F1-score* tertinggi pada ambang *default*, tetapi kemampuan pemeringkatannya relatif terbatas. Sebaliknya, LR menunjukkan kemampuan pemeringkatan yang lebih baik, tercermin dari nilai *PR-AUC* dan *ROC-AUC* yang lebih tinggi, meskipun kinerjanya pada ambang *default* kurang optimal sebelum dilakukan penyetelan ambang. Temuan ini mendukung argumen bahwa pada data yang sangat tidak seimbang, kemampuan pemeringkatan dan kinerja pada satu titik operasi merupakan dua aspek yang perlu dievaluasi secara terpisah (Saito dan Rehmsmeier, 2015). Lalu, dampak penerapan SMOTE bergantung pada algoritma yang digunakan. Pada KNN, SMOTE meningkatkan *recall* tetapi menurunkan *precision* dan *F1-score.* Sebaliknya, pada LR, SMOTE meningkatkan *recall* secara signifikan, namun dengan konsekuensi meningkatnya jumlah positif palsu. Dalam penelitian ini, LR *baseline* yang disertai penyetelan ambang menghasilkan keseimbangan yang lebih baik antara *precision* dan *recall* dibandingkan LR+SMOTE. Selain itu, fitur *amt zscore card* merupakan prediktor terkuat yang digunakan dalam pemodelan setelah pengecualian fitur *amt* untuk menghindari multikolinearitas. Dengan tetap mempertahankan sebagian besar kemampuan diskriminatif fitur asli, fitur tersebut memberikan representasi yang lebih stabil bagi kedua algoritma. Secara keseluruhan, hasil penelitian menunjukkan bahwa tidak terdapat model yang unggul pada seluruh skenario evaluasi. KNN *baseline* memberikan keseimbangan terbaik tanpa penyetelan tambahan, LR dengan ambang tersetel menghasilkan *F1-score* tertinggi, sedangkan LR+SMOTE lebih sesuai ketika prioritas utama adalah memaksimalkan *recall.*

IV. KESIMPULAN

Penelitian ini membandingkan kinerja KNN dan LR untuk deteksi penipuan kartu kredit pada dataset Sparkov yang sangat tidak seimbang, baik pada kondisi baseline maupun setelah penerapan SMOTE. Hasil penelitian menunjukkan bahwa tidak terdapat model yang secara konsisten unggul pada seluruh metrik evaluasi. KNN *baseline* memberikan keseimbangan terbaik antara *precision* dan *recall* pada ambang *default,* sedangkan LR menunjukkan kemampuan pemeringkatan yang lebih baik dan mencapai *F1-score* tertinggi setelah dilakukan penyetelan ambang

DOI: https://doi.org/10.24036/ujsds/vol4-iss3/539

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

keputusan. Sementara itu, LR+SMOTE menghasilkan *recall* tertinggi, tetapi dengan peningkatan positif palsu yang substansial. Temuan ini menunjukkan bahwa pada data yang sangat tidak seimbang, pemilihan metrik evaluasi dan titik operasi sama pentingnya dengan pemilihan algoritma. Secara khusus, penyetelan ambang keputusan terbukti menjadi strategi yang efektif untuk meningkatkan kinerja LR tanpa memerlukan *oversampling*. Keterbatasan penelitian ini terletak pada penggunaan dataset sintetis, sehingga penelitian selanjutnya dapat memanfaatkan dataset yang lebih besar atau lebih beragam serta mengevaluasi model *ensemble* dan pendekatan kalibrasi probabilitas.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada penyedia dataset dan Universitas Ciputra yang telah mendukung kegiatan penelitian ini.

DAFTAR PUSTAKA

Assis, A., Dantas, J., & De Andrade, E. C. (2024). The performance-interpretability trade-off: a comparative study of machine learning models. *Journal of Reliable Intelligent Environments*, 11. https://doi.org/10.1007/s40860-024-00240-0

Chawla, N. V., Bowyer, K. W., Hall, L. O., dan Kegelmeyer, W. P. (2002), "SMOTE: Synthetic Minority Over-sampling Technique", *Journal of Artificial Intelligence Research*, Vol. 16, hal. 321–357.

Federal Trade Commission (2023), New FTC Data Show Consumers Reported Losing Nearly $8.8 Billion to Scams in 2022, https://www.ftc.gov/news-events/news/press-releases/2023/02/new-ftc-data-show-consumers-reported-losing-nearly-88-billion-scams-2022.

Grover, P., Xu, J., Tittelfitz, J., Cheng, A., Li, Z., Zablocki, J., Liu, J., dan Zhou, H. (2022), "Fraud Dataset Benchmark and Applications", *arXiv*, https://doi.org/10.48550/arXiv.2208.14417.

Harris, B. (2016), Sparkov Data Generation Tool [Source code], GitHub, https://github.com/namebrandon/Sparkov_Data_Generation.

Kaggle/Shenoy, K. (2020), Credit Card Transactions Fraud Detection Dataset [Dataset], https://www.kaggle.com/datasets/kartik2112/fraud-detection.

Lopez-Rojas, E. A., Elmir, A., dan Axelsson, S. (2016), "PaySim: A Financial Mobile Money Simulator for Fraud Detection", *Proceedings of the 28th European Modeling and Simulation Symposium (EMSS 2016)*.

Otoritas Jasa Keuangan (2026), Kejahatan Siber Melonjak 550 Persen, OJK Ingatkan Pentingnya Keamanan Digital, ANTARA News, https://www.antaranews.com/berita/5363118.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., dan Duchesnay, É. (2011), "Scikit-learn: Machine Learning in Python", *Journal of Machine Learning Research*, Vol. 12, hal. 2825–2830.

Saito, T., dan Rehmsmeier, M. (2015), "The Precision-Recall Plot Is More Informative Than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets", *PLOS ONE*, Vol. 10, No. 3, e0118432.

Thölke, P., Mantilla-Ramos, Y.-J., Abdelhedi, H., Maschke, C., Dehgan, A., Harel, Y., Kemtur, A., Berrada, L. M., Sahraoui, M., Young, T., Pépin, A. B., Khantour, C. E., Landry, M., Pascarella, A., Hadid, V., Combrisson, E., O'Byrne, J., & Jerbi, K. (2023). Class imbalance should not throw you off balance: Choosing the right classifiers and performance metrics for brain decoding with imbalanced data. *bioRxiv*. https://doi.org/10.1101/2022.07.18.500262

Veldurthi, A. K. (2025). The Role of AI and Machine Learning in Fraud Detection for Financial Services. *Journal of Computer Science and Technology Studies*. https://doi.org/10.32996/jcsts.2025.7.4.88

Yusran, M., Sadik, K., Soleh, A. M., & Suhaeni, C. (2025). Effect of Feature Normalization and Distance Metrics on K-Nearest Neighbors Performance for Diabetes Disease Classification. *Journal of Mathematics, Computations and Statistics*. https://doi.org/10.35580/jmathcos.v8i2.8012