JUSTIAN, Jurnal Sistem Informasi Akuntansi
Vol. 6, No. 2, September 2025, pp. 186~193
ISSN: 2721-7523

Analisis Kinerja *Logistic Regression* dan *Random Forest* pada Deteksi *Fraud E-Commerce* Menggunakan SMOTE dan PCA

Kartika Handayani[1], Erni[2], Ardiyansyah[3], Agung Sasongko[4]

Info Artikel

Diterima September 27, 2025
Revisi September 29, 2025
Terbit September 30, 2025

Keywords:

E-commerce *Fraud* Detection,
SMOTE,
PCA,
Logistic Regression,
Random Forest

ABSTRACT

The rapid growth of e-commerce platforms has increased the volume and complexity of digital transactions, which is accompanied by a rising risk of fraudulent activities. This study aims to apply and evaluate the performance of Logistic Regression and Random Forest algorithms for fraud detection in e-commerce transactions. To address the class imbalance problem, the Synthetic Minority Over-sampling Technique (SMOTE) is employed, while dimensionality reduction is performed using Principal Component Analysis (PCA). The dataset is divided into training and testing sets using an 80:20 ratio. Model evaluation is conducted under four scenarios: baseline without additional preprocessing, SMOTE only, PCA only, and a combination of SMOTE and PCA. The results indicate that Random Forest consistently outperforms Logistic Regression across most evaluation metrics, including Recall, F1-Score, and Area Under the Curve (AUC). The application of SMOTE significantly improves the model's ability to identify fraudulent transactions, achieving the highest Recall of 80.79% using Random Forest. In contrast, the use of PCA, either alone or combined with SMOTE, tends to degrade model performance. This study concludes that Random Forest combined with SMOTE provides the most effective approach for fraud detection in highly imbalanced e-commerce transaction data.

Identitas Penulis:

Kartika Handayani[1], Erni[2], Ardiyansyah[3], Agung Sasongko[4]
Universitas Bina Sarana Informatika, Program Studi Informatika Kampus Kota Pontianak[1, 3], Program Studi Sistem Informasi Kampus Kota Pontianak[3,4]
Jl. Abdul Rahman Saleh No.18 A Pontianak[1,2]
Email: kartika.kth@bsi.ac.id[1] , erni.erx@bsi.ac.id[2] , ardi.arq@bsi.ac.id[3] , agung.ako@bsi.ac.id[4]

1. PENDAHULUAN

Perkembangan pesat *platform e-commerce* selama beberapa tahun terakhir meningkatkan volume dan kompleksitas transaksi digital, sekaligus memicu peningkatan kasus penipuan (*fraud*) yang merugikan konsumen dan pelaku usaha. Studi-studi terakhir menunjukan bahwa ancaman *fraud* di ekosistem *e-commerce* bukan hanya pada skala transaksi kartu kredit, tetapi juga meliputi penipuan akun penjual/pembeli dan manipulasi pesanan — hal ini menuntut metode deteksi yang lebih adaptif dan otomatis [1]

Metode data mining dan *machine learning* (ML) menjadi pendekatan utama untuk mendeteksi pola-pola perilaku *fraud* dari data transaksi, log pengguna, dan metadata perangkat. Dua algoritma yang sering dipilih dalam literatur adalah *Logistic Regression* (LR)—yang menawarkan interpretabilitas dan baseline cepat—dan *Random Forest* (RF)—sebuah metode *ensemble* yang kuat untuk menangani non-linearitas dan interaksi fitur. Berbagai penelitian komparatif melaporkan bahwa RF umumnya memberikan performa (mis. AUC, recall, F1) lebih baik dibanding LR pada dataset transaksi, namun LR tetap penting ketika interpretabilitas dan implementasi ringan diperlukan [2]

Salah satu tantangan kritis pada deteksi *fraud* adalah ketidakseimbangan kelas (*class imbalance*): kasus *fraud* biasanya jauh lebih sedikit daripada transaksi normal, sehingga model cenderung bias ke kelas mayoritas. Teknik *resampling* seperti SMOTE (*Synthetic Minority Over-sampling Technique*) banyak digunakan untuk mengatasi masalah ini dengan membuat sampel sintetis pada kelas minoritas, yang terbukti meningkatkan

JUSTIAN, Jurnal Sistem Informasi Akuntansi
Vol. 6, No. 2, September 2025, pp. 186~193
ISSN: 2721-7523

recall dan F1 pada model seperti RF dan LR ketika diaplikasikan pada dataset keuangan/*transactional*. Beberapa kajian juga mengombinasikan SMOTE dengan teknik pembersihan (mis. SMOTE-Tomek) atau optimasi *hyperparameter* untuk memaksimalkan hasil [3]

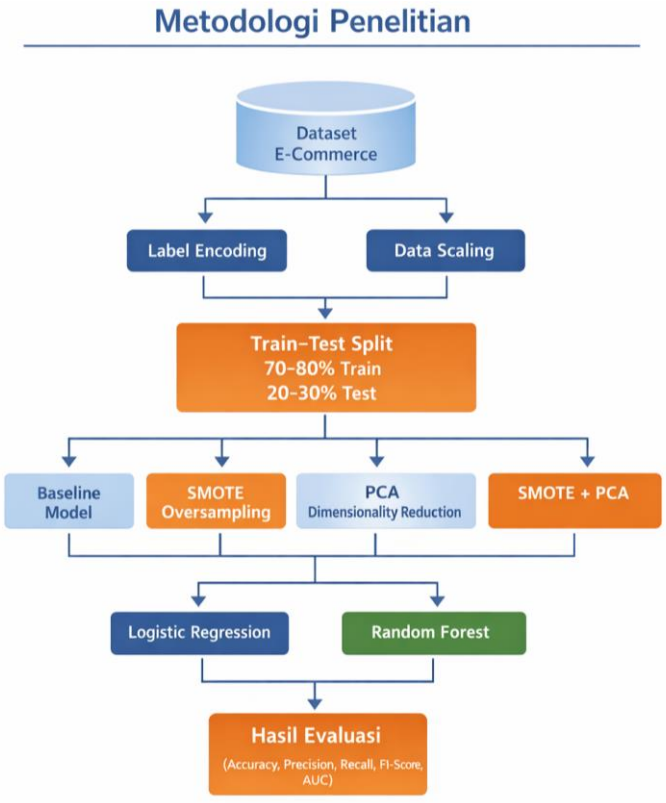
Selain itu, data *e-commerce* seringkali memiliki dimensi fitur yang tinggi (*behavioral features, one-hot categories, time features, device* metadata), yang dapat memperlambat pelatihan dan meningkatkan risiko *overfitting* [4]. *Principal Component Analysis* (PCA) merupakan teknik reduksi dimensi yang sering dipakai untuk mereduksi redundansi fitur sambil mempertahankan variansi terpenting—mempercepat pelatihan model dan kadang juga meningkatkan generalisasi model pada dataset transaksi. Beberapa studi mengaplikasikan PCA sebelum training model ML untuk menyeimbangkan trade-off antara kecepatan dan akurasi [5]

Penelitian ini hanya berfokus pada penerapan metode SMOTE dan PCA serta penggunaan algoritma *Logistic Regression* dan *Random Forest*, dengan tujuan mengetahui hasil evaluasi kinerja masing-masing model terhadap data *e-commerce*

.

2. METODE

Metode merupakan cara atau prosedur yang sistematis dan terstruktur yang digunakan untuk mencapai tujuan atau menyelesaikan suatu masalah. Berikut ini metode yang dilakukan dalam penelitian ini:



Gambar 1. Metode Penelitian

1. *Dataset*

Penelitian ini menggunakan dataset *E-Commerce Fraud Detection Dataset*, yang merupakan data publik yang diperoleh dari *Kaggle* dengan link : https://www.kaggle.com/datasets/umuttuygurr/e-commerce-*fraud*-detection-dataset. Dataset terdiri dari 299.695 data transaksi *e-commerce* dengan 17 atribut, yang mencerminkan karakteristik pengguna, transaksi, metode pembayaran, serta label kecurangan (*fraud*). Dataset ini digunakan untuk membangun dan mengevaluasi model deteksi *fraud* berbasis machine learning. Berikut atribut *E-Commerce Fraud Detection Dataset* :

JUSTIAN, Jurnal Sistem Informasi Akuntansi
Vol. 6, No. 2, September 2025, pp. 186~193
ISSN: 2721-7523

Tabel 1. Atribut *E-Commerce Fraud Detection Dataset*

No	Nama Atribut	Tipe Data	Deskripsi
1	*transaction_id*	*Integer*	Identitas unik setiap transaksi. Digunakan sebagai penanda data dan tidak berperan langsung sebagai fitur prediktor.
2	*user_id*	*Integer*	Identitas unik pengguna yang melakukan transaksi. Digunakan untuk merepresentasikan perilaku pengguna.
3	*account_age_days*	*Integer*	Usia akun pengguna dalam satuan hari sejak pertama kali terdaftar. Akun baru cenderung memiliki risiko *fraud* lebih tinggi.
4	*total_transactions_user*	*Integer*	Jumlah total transaksi yang pernah dilakukan oleh pengguna. Mencerminkan intensitas aktivitas pengguna.
5	*avg_amount_user*	*Float*	Rata-rata nilai transaksi historis pengguna. Digunakan untuk mendeteksi penyimpangan nilai transaksi.
6	*amount*	*Float*	Nilai nominal transaksi saat ini. Transaksi bernilai besar berpotensi meningkatkan risiko *fraud*.
7	*country*	Kategorikal (*Object*)	Negara asal pengguna atau lokasi transaksi. Digunakan untuk analisis risiko berbasis geografis.
8	*bin_country*	Kategorikal (*Object*)	Negara asal kartu pembayaran berdasarkan *Bank Identification Number* (BIN). Ketidaksesuaian dengan country dapat mengindikasikan *fraud*.
9	*channel*	Kategorikal (*Object*)	Saluran transaksi yang digunakan, seperti web atau aplikasi mobile.
10	*merchant_category*	Kategorikal (*Object*)	Kategori merchant tempat transaksi dilakukan. Setiap kategori memiliki tingkat risiko *fraud* yang berbeda.
11	*promo_used*	*Integer* (Biner)	Indikator penggunaan promo atau diskon (1 = ya, 0 = tidak).
12	*avs_match*	*Integer* (Biner)	Hasil verifikasi *Address Verification System* (AVS), menunjukkan kecocokan alamat penagihan.
13	*cvv_result*	*Integer* (Biner)	Hasil verifikasi kode keamanan kartu (CVV).
14	*three_ds_flag*	*Integer* (Biner)	Indikator penggunaan autentikasi 3D Secure (1 = digunakan, 0 = tidak).
15	*transaction_time*	Kategorikal / Waktu (*Object*)	Waktu terjadinya transaksi, dapat diekstraksi menjadi fitur temporal (jam, hari).
16	*shipping_distance_km*	*Float*	Jarak antara alamat penagihan dan alamat pengiriman dalam satuan kilometer.
17	*is_fraud*	*Integer* (Target)	Label kelas transaksi (1 = *fraud*, 0 = normal). Digunakan sebagai variabel target.

2. *Label Encoding*

Label Encoding Pengkodean label dilakukan yang bertujuan untuk melakukan pengkodean pada label kelas. Pengkodean label berfungsi untuk mengubah format data angka 0 hingga n_kelas-1 [6].

JUSTIAN, Jurnal Sistem Informasi Akuntansi
Vol. 6, No. 2, September 2025, pp. 186~193
ISSN: 2721-7523

3. Data Scaling

Data scaling menggunakan *standard scaler*. *Standard Scaler* adalah metode normalisasi yang paling umum digunakan. Metode ini bekerja dengan cara menghitung rata-rata dan standar deviasi dari data, kemudian membagi setiap nilai data dengan standar deviasi tersebut [7].

4. Pembagian Train & Test

Tahapan selanjutnya adalah melakukan validas pemisahan yang bertujuan untuk membagi data menjadi dua bagian, yaitu data pelatihan dan data pengujian. Data pelatihan dimanfaatkan untuk melatih model, sementara data pengujian digunakan untuk menguji keakuratan atau evaluasi model tersebut. Dalam penelitian ini, pemisahan dilakukan dengan menggunakan persentase pembagian sebesar 80:20, yang berarti 80% dari data untuk pelatihan dan 20% untuk pengujian [8].

5. SMOTE

Synthetic Minority Over Sampling Technique (SMOTE) menghasilkan sampel buatan dari kelas minoritas dengan menginterpolasi instanceyang ada yang sangat dekat satu sama lain [8].

6. PCA

PCA (*Principal Component Analysis*) diaplikasikan untuk menyederhanakan data tanpa menghilangkan informasi krusial [9].

7. *Logistic Regresstion*

Teknik statistik dan machine learning untuk memodelkan hubungan antara variabel dependen (target) dan satu atau lebih variabel independen (prediktor) sebagai sebuah garis lurus, bertujuan memprediksi nilai yang tidak diketahui berdasarkan data yang ada [10].

8. *Random Forest*

Random Forest adalah salah satu algoritma dalam machine learning yang sangat terkenal. Secara umum, *random forest* terdiri dari sekumpulan decision tree yang digabungkan untuk menciptakan model yang lebih tepat. Oleh karena itu, istilah *'forest'* digunakan untuk merujuk pada kumpulan decision tree tersebut. Random forest akan mengembangkan beberapa tree dengan menggunakan data contoh, di mana tree yang dibentuk selama proses pelatihan tidak akan tergantung pada tree yang telah ada sebelumnya. Kemudian, keputusan akan diambil dengan cara voting dari hasil yang paling banyak [11].

9. Evaluasi Hasil

Untuk membandingkan kinerja keseluruhan skema penelitian yang diusulkan, dilakukan evaluasi menggunakan: *accuracy, recall, precision, F1-Score dan AUC.*

3. HASIL

Berikut hasil dari hasil penerapan sesuai metode yang telah dirancang :

1. *Label Encoding*

✅ Data Setelah Label Encoding (5 baris pertama):

	country	bin_country	channel	merchant_category	transaction_time
0	2	2	1	4	5053
1	2	2	1	4	8612
2	2	2	0	4	10984
3	2	2	1	1	14409
4	2	9	1	0	15739

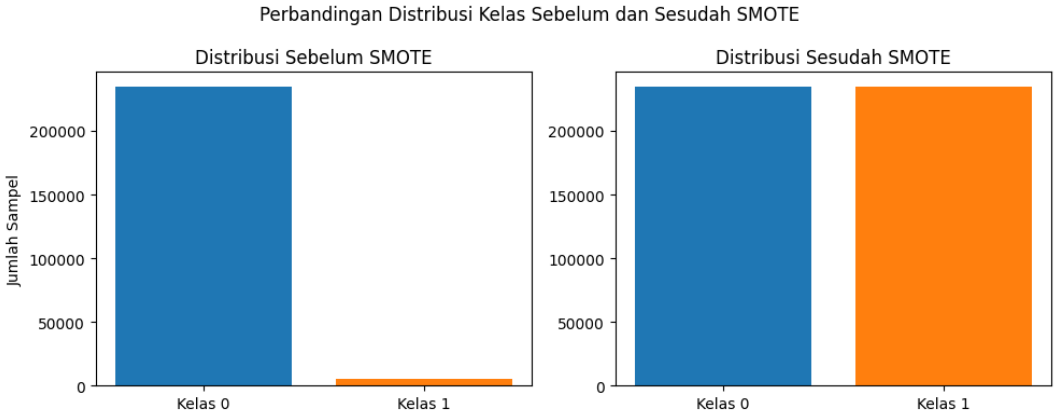
Gambar 2. Hasil *Label Encoding*

Gambar 2 menunjukkan contoh 5 baris pertama data setelah dilakukan proses label encoding, di mana atribut kategorikal seperti country, bin_country, channel, merchant_category, dan transaction_time telah dikonversi menjadi nilai numerik. Transformasi ini bertujuan agar data kategorikal dapat diproses oleh algoritma machine learning yang memerlukan input numerik, seperti Logistic Regression dan Random Forest. Nilai angka yang dihasilkan hanya berfungsi sebagai

JUSTIAN, Jurnal Sistem Informasi Akuntansi
Vol. 6, No. 2, September 2025, pp. 186~193
ISSN: 2721-7523

penanda kategori dan tidak merepresentasikan urutan atau tingkat tertentu, sehingga data siap digunakan pada tahap preprocessing dan pemodelan selanjutnya dalam penelitian deteksi *fraud* e-commerce.

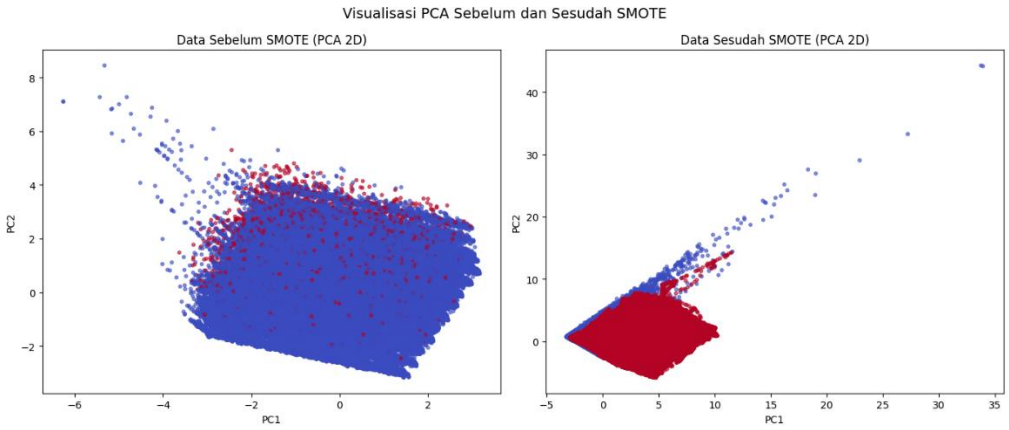
2. SMOTE



Gambar 3. Transformasi Data Menggunakan SMOTE

Gambar 3 memperlihatkan perbandingan distribusi kelas sebelum dan sesudah penerapan SMOTE. Sebelum SMOTE, dataset menunjukkan ketidakseimbangan kelas yang signifikan, dengan 234.466 data kelas non-*fraud* (Kelas 0) dan hanya 5.290 data kelas *fraud* (Kelas 1), sehingga model berpotensi bias terhadap kelas mayoritas. Setelah diterapkan SMOTE, jumlah data pada kelas minoritas ditingkatkan secara sintetis hingga seimbang dengan kelas mayoritas (234.466 : 234.466). Penyeimbangan ini bertujuan untuk meningkatkan kemampuan model dalam mengenali pola transaksi *fraud*, khususnya dalam meningkatkan nilai recall dan F1-score pada tahap pelatihan model.

3. PCA



Gambar 4. Visualisasi PCA Sebelum dan Sesudah SMOTE

Gambar tersebut menampilkan visualisasi data hasil reduksi dimensi menggunakan PCA dua dimensi (PC1 dan PC2) untuk membandingkan kondisi sebelum dan sesudah penerapan SMOTE. Pada visualisasi sebelum SMOTE, terlihat bahwa data kelas mayoritas mendominasi ruang fitur dan kelas minoritas (*fraud*) tersebar tipis serta saling tumpang tindih, sehingga batas pemisahan antar kelas sulit dibedakan. Setelah SMOTE diterapkan, jumlah data kelas minoritas meningkat secara signifikan

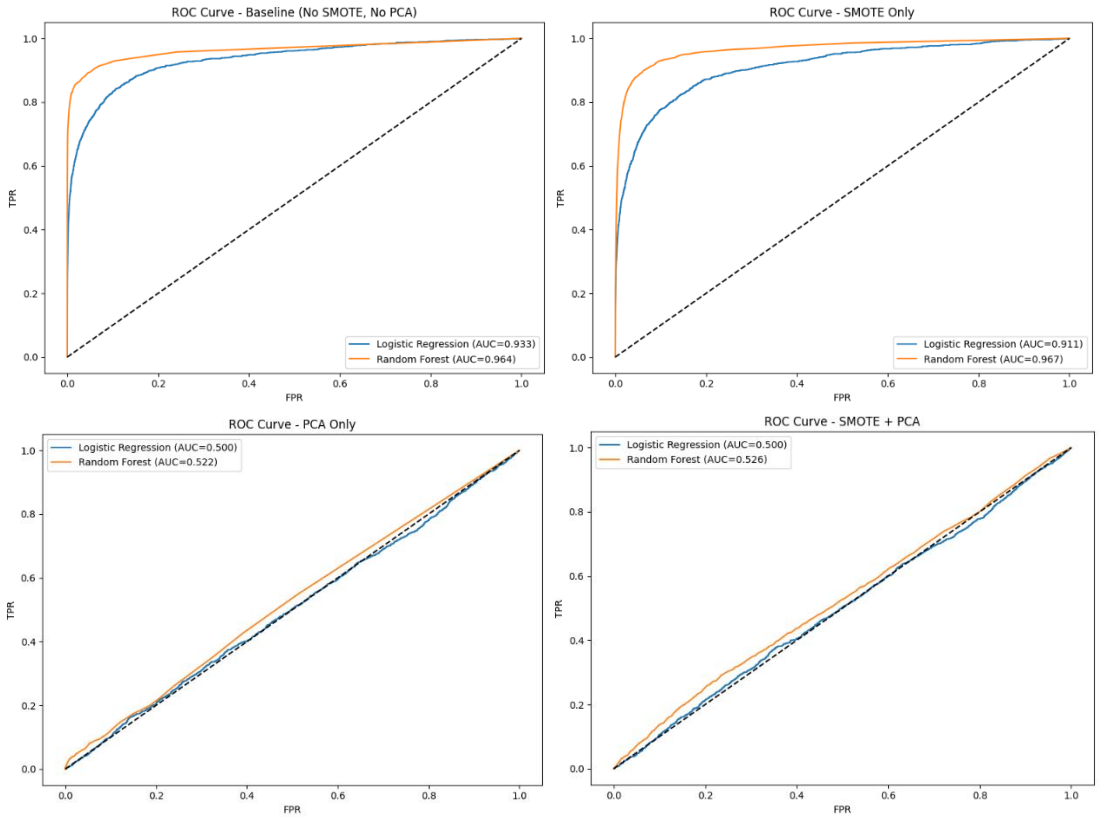
JUSTIAN, Jurnal Sistem Informasi Akuntansi
Vol. 6, No. 2, September 2025, pp. 186~193
ISSN: 2721-7523

dan distribusinya menjadi lebih merata di ruang PCA, meskipun masih terdapat tumpang tindih antar kelas. Visualisasi ini menunjukkan bahwa SMOTE berhasil menyeimbangkan distribusi kelas, namun sekaligus memperlihatkan bahwa struktur data menjadi lebih padat dan kompleks, yang dapat memengaruhi kinerja model pada tahap klasifikasi.

4. Hasil Penerapan

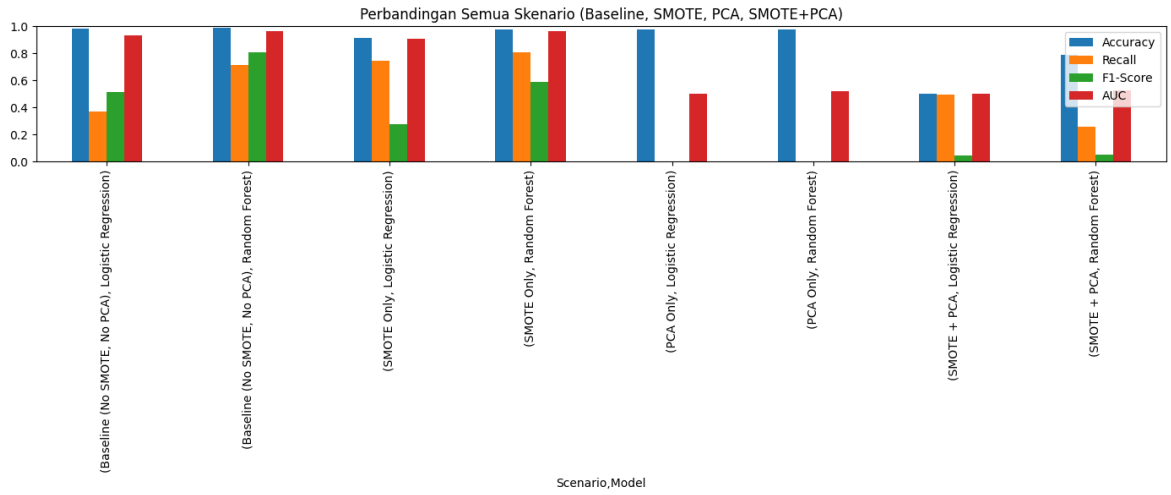
Tabel 1. Atribut *E-Commerce Fraud Detection Dataset*

No	Skenario	Model	*Accuracy*	*Precision*	*Recall*	*F1-Score*	*AUC*
1	*Baseline (No SMOTE, No PCA)*	*Logistic Regression*	0.9846	0.8446	0.3699	0.5145	0.9326
2	*Baseline (No SMOTE, No PCA)*	***Random Forest***	**0.9925**	**0.9266**	0.7163	**0.8080**	0.9641
3	*SMOTE Only*	*Logistic Regression*	0.9131	0.1687	0.7481	0.2753	0.9109
4	*SMOTE Only*	***Random Forest***	0.9753	0.4656	**0.8079**	0.5907	**0.9667**
5	*PCA Only*	*Logistic Regression*	0.9779	0.0000	0.0000	0.0000	0.4997
6	*PCA Only*	*Random Forest*	0.9778	0.2143	0.0023	0.0045	0.5221
7	*SMOTE + PCA*	*Logistic Regression*	0.5040	0.0222	0.4985	0.0425	0.5004
8	*SMOTE + PCA*	*Random Forest*	0.7865	0.0277	0.2549	0.0500	0.5256



Gambar 5. Gambar ROC

JUSTIAN, Jurnal Sistem Informasi Akuntansi
Vol. 6, No. 2, September 2025, pp. 186~193
ISSN: 2721-7523



Gambar 5. Perandingan Semua Skenario Pengujian

Perbandingan kinerja algoritma Logistic Regression dan Random Forest pada empat skenario, yaitu tanpa preprocessing (baseline), penerapan SMOTE, penerapan PCA, serta kombinasi SMOTE dan PCA. Hasil menunjukkan bahwa Random Forest secara konsisten menghasilkan performa yang lebih baik dibanding Logistic Regression pada seluruh skenario, khususnya pada metrik Recall, F1-Score, dan AUC yang lebih relevan untuk deteksi *fraud*. Pada skenario baseline, Random Forest mencapai AUC sebesar 0,964 dan F1-Score sebesar 0,808, jauh lebih tinggi dibanding Logistic Regression yang memiliki Recall rendah akibat ketidakseimbangan kelas. Penerapan SMOTE terbukti meningkatkan kemampuan deteksi *fraud* pada kedua model, terutama meningkatkan Recall hingga 80,79% pada Random Forest. Sebaliknya, penerapan PCA tanpa SMOTE maupun kombinasi SMOTE dan PCA tidak memberikan peningkatan kinerja yang signifikan dan bahkan menurunkan performa model, yang mengindikasikan hilangnya informasi penting terkait pola *fraud* pada proses reduksi dimensi.

4. KESIMPULAN

Berdasarkan hasil evaluasi, dapat disimpulkan bahwa Random Forest dengan penerapan SMOTE memberikan kinerja terbaik dalam mendeteksi *fraud* pada dataset e-commerce, ditunjukkan oleh nilai Recall, F1-Score, dan AUC yang paling tinggi dibandingkan skenario lainnya. Logistic Regression tetap menunjukkan performa yang memadai sebagai model baseline, namun kurang optimal dalam menangani ketidakseimbangan kelas. Sementara itu, penggunaan PCA tidak direkomendasikan pada penelitian ini karena cenderung menurunkan kemampuan model dalam mengenali transaksi *fraud*. Dengan demikian, penelitian ini menegaskan bahwa pemilihan algoritma dan teknik preprocessing yang tepat sangat berpengaruh terhadap efektivitas sistem deteksi *fraud* e-commerce.

REFERENSI

[1] W. Priatna, S. Yulianto, J. Prasetyo, S. Wijono, E. Maria, and D. Manongga, "Deteksi Anomali dalam Penipuan E-commerce Menggunakan Hybrid Autoencoder-Transformer Frameworks," *JEPIN (Jurnal Edukasi dan Penelit. Inform.*, vol. 11, no. 1, pp. 33–40, 2025.

[2] J. K. Afriyie *et al.*, "A supervised machine learning algorithm for detecting and predicting *fraud* in credit card transactions," *Decis. Anal. J.*, vol. 6, no. November 2022, p. 100163, 2023, doi: 10.1016/j.dajour.2023.100163.

[3] P. Sundaravadivel, R. A. Isaac, D. Elangovan, D. KrishnaRaj, V. V. L. Rahul, and R. Raja, "Optimizing credit card *fraud* detection with random forests and SMOTE," *Sci. Rep.*, vol. 15, no. 1, 2025, doi: 10.1038/s41598-025-00873-y.

[4] K. S. Roy, P. B. Udas, B. Alam, and K. Paul, "Unveiling Hidden Patterns: A Deep Learning Framework Utilizing PCA for *Fraud*ulent Scheme Detection in Supply Chain Analytics," *Int. J. Intell. Syst. Appl.*, vol. 17, no. 2, pp. 14–30, 2025, doi: 10.5815/ijisa.2025.02.02.

[5] Z. Salekshahrezaee, J. L. Leevy, and T. M. Khoshgoftaar, "The effect of feature extraction and data sampling on credit card *fraud* detection," *J. Big Data*, vol. 10, no. 1, 2023, doi: 10.1186/s40537-023-00684-w.

JUSTIAN, Jurnal Sistem Informasi Akuntansi
Vol. 6, No. 2, September 2025, pp. 186~193
ISSN: 2721-7523

[6] F. Ernawan, K. Handayani, M. Fakhreldin, and Y. Abbker, "Light Gradient Boosting with Hyper Parameter Tuning Optimization for COVID-19 Prediction," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 8, pp. 514–523, 2022, doi: 10.14569/IJACSA.2022.0130859.
[7] A. C. Purba and T. Handhayani, "Perbandingan Algoritma K-Means, Affinity Clustering, Dan Minibatch K-Means Untuk Analisis Segmentasi Pasar," vol. 13, no. 1, pp. 54–63, 2024.
[8] K. Handayani and E. Erni, "Penerapan Light Gradient Boosting Dalam Prediksi Rasio Klik Tayang," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 1, pp. 13–18, 2023, doi: 10.36040/jati.v7i1.6010.
[9] R. Rianti *et al.*, "Penerapan PCA dan Algoritma Clustering untuk Analisis Mutu Perguruan Tinggi di LLDIKTI Wilayah IV," vol. 18, 2024.
[10] S. Khoiriyah and Z. Fatah, "Penerapan Algoritma Linear Regression dalam Memprediksi Harga Rumah Menggunakan RapidMiner," vol. 3, no. 2, pp. 107–115, 2024.
[11] M. M. Alvanof and R. K. Dinata, "Penerapan Algoritma Random Forest dalam Deteksi dan Klasifikasi Ransomware," vol. 5, no. 2, 2024.