Journal of Information Technology, Software Engineering and Computer Science
Website : https://ejournal.techcart-press.com/index.php/itsecs
E-ISSN 2962-0635
P-ISSN 2962-0538
Volume 5, Number 1, January 2027
DOI : **https://doi.org/10.58602/itsecs.v5i1.498**
Page 135-158

Prediksi Perilaku Konsumen UMKM E-Commerce Indonesia Menggunakan Random Forest, XGBoost, dan TreeSHAP

Mayadi[1]✉, Prio Kustanto[1*]✉, Dismita Citra Dewi[2]✉, Lenita Waty[3]✉, Hendarman Lubis[1]✉
[1]Informatika, Universitas Bhayangkara Jakarta Raya, Indonesia
[2]Ekonomi Syariah, Institut Agama Islam Pagaralam, Indonesia
[3]Bisnis Digital, Universitas Advent Indonesia, Indonesia

Abstrak: Transformasi digital usaha mikro, kecil, dan menengah (UMKM) Indonesia melalui lokapasar menghadapi tantangan sensitivitas biaya logistik antarpulau serta rendahnya pembelian ulang (*repeat order*). Penelitian ini merancang kerangka kerja prediktif berbasis *Random Forest*, *Extreme Gradient Boosting* (XGBoost), dan *TreeSHAP* untuk memprediksi perilaku konsumen sekaligus merumuskan strategi retensi pelanggan UMKM *e-commerce* Indonesia. Metodologi studi mencakup audit skema dua puluh sembilan atribut transaksi, isolasi kebocoran target (*target leakage*), penyeimbangan kelas sintetis, serta rancangan evaluasi klasifikasi (*accuracy*, *precision*, *recall*, *F1-score*, dan *ROC-AUC*) beserta atribusi nilai Shapley. Meskipun demikian, pelatihan model empiris, metrik evaluasi, serta nilai *TreeSHAP* belum dapat dieksekusi karena himpunan data hanya tersedia secara terfragmentasi dan modul komputasi khusus algoritma *boosting* serta komputasi *SHAP* belum terpasang pada lingkungan analitik. Audit kuantitatif terhadap tiga puluh empat catatan terverifikasi mengonfirmasi kesesuaian identitas aritmetika nilai transaksi bersih tanpa deviasi, anomali ketidaksesuaian kategori produk terhadap teks ulasan, serta hubungan deterministik tanpa varians antara peringkat produk, keterlambatan pengiriman, dan pembelian ulang terhadap skor loyalitas pelanggan. Data terverifikasi juga memperlihatkan kesenjangan ongkos kirim wilayah tertinggal sebesar Rp 78.571 dibandingkan Rp 21.615 antarkota sepulau, sementara keterlambatan pengiriman maupun pembayaran tunai di tempat mencatat pembelian ulang nol persen. Temuan audit dan sintesis literatur mendasari strategi retensi tersegmentasi berbasis subsidi ongkos kirim terarah, pengendalian layanan kurir per koridor, konversi pembayaran digital, serta pemulihan layanan empat tingkat risiko *churn* bagi UMKM Indonesia.
Kata Kunci: Perilaku Konsumen UMKM; Retensi Pelanggan; Random Forest; Extreme Gradient Boosting (XGBoost); TreeSHAP

Abstract: The digital transformation of Indonesian micro, small, and medium enterprises (MSMEs) through online marketplaces faces challenges related to inter-island logistics cost sensitivity and low repeat purchase rates. This study develops a predictive framework based on Random Forest, Extreme Gradient Boosting (XGBoost), and TreeSHAP to predict consumer behaviour while formulating customer retention strategies for Indonesian e-commerce MSMEs. The research methodology encompasses an audit of the 29 transaction attributes, target leakage isolation, synthetic class balancing, and a classification evaluation design incorporating accuracy, precision, recall, F1-score, and ROC-AUC, alongside Shapley value attribution. Nevertheless, empirical model training,

Prio Kustanto: *Penulis Korespondensi
Copyright © 2027, Mayadi, Prio Kustanto, Dismita Citra Dewi, Lenita Waty, Hendarman Lubis.

Journal of Information Technology, Software Engineering and Computer Science
Website : https://ejournal.techcart-press.com/index.php/itsecs
E-ISSN 2962-0635
P-ISSN 2962-0538
Volume 5, Number 1, January 2027
DOI : **https://doi.org/10.58602/itsecs.v5i1.498**
Page 135-158

evaluation metrics, and TreeSHAP values could not yet be executed because the dataset was only available in fragmented form, and the specialized computational modules for boosting algorithms and SHAP computation were not installed in the analytical environment. A quantitative audit of 34 verified records confirmed the arithmetic consistency of net transaction values without any deviations, identified anomalies in the mismatch between product categories and review text, and revealed deterministic relationships with zero variance among product ratings, delivery delays, repeat purchases, and customer loyalty scores. The verified data also revealed a shipping cost disparity between underserved regions (Rp 78,571) and intercity deliveries within the same island (Rp 21,615), while both delivery delays and cash-on-delivery payments recorded zero percent repeat purchases. The audit findings and literature synthesis provide the foundation for segmented customer retention strategies comprising targeted shipping subsidies, courier service quality control across delivery corridors, digital payment conversion, and a four-tier churn risk-based service recovery framework for Indonesian MSMEs.

Keywords: MSME Consumer Behavior; Customer Retention; Random Forest; Extreme Gradient Boosting (XGBoost); TreeSHAP

1. PENDAHULUAN

Ekosistem perdagangan elektronik (*e-commerce*) di Indonesia telah berkembang menjadi salah satu pasar dengan tingkat penetrasi pengguna tertinggi di dunia sekaligus bertindak sebagai instrumen strategis inklusi ekonomi bagi usaha mikro, kecil, dan menengah (UMKM)[1]. Adopsi kanal lokapasar nasional seperti Shopee Indonesia, Tokopedia yang terintegrasi dengan TikTok Shop, Lazada Indonesia, Blibli, serta kanal perdagangan percakapan seperti WhatsApp Business dan Instagram Shop memungkinkan pelaku UMKM memasarkan produk kriya dan kerajinan tangan, kuliner dan makanan ringan khas daerah, fesyen muslim dan batik, kopi olahan, kosmetik dan herbal lokal, aksesoris daerah, hingga peralatan rumah tangga melampaui batas wilayah provinsi asal penjual. Bukti empiris pada ratusan pelaku UMKM di Indonesia menunjukkan bahwa pemanfaatan perdagangan elektronik berasosiasi positif dengan peningkatan kinerja penjualan, perluasan pangsa pasar, serta penguatan daya saing usaha di tengah pergeseran preferensi konsumen menuju metode pembayaran digital seperti dompet elektronik, QRIS, transfer rekening virtual, hingga fasilitas cicilan digital (*paylater*). Sejalan dengan perluasan kanal akuisisi digital tersebut, tantangan manajerial yang dihadapi pelaku UMKM kini bergeser dari sekadar mendorong transaksi perdana menjadi menjaga kesinambungan pembelian ulang (*repeat order*) dan mencegah perpindahan konsumen ke penjual kompetitor (*customer churn*). Dalam lanskap persaingan lokapasar pihak ketiga di mana kuota promo voucher bersifat terbatas dan preferensi pelanggan mudah beralih akibat biaya tambahan pengiriman, kemampuan memprediksi perilaku pascatransaksi melalui pendekatan pembelajaran mesin yang dapat dijelaskan (*explainable machine learning*) menjadi faktor penentu ketahanan usaha UMKM jangka panjang.

Walaupun digitalisasi ritel membuka akses pasar nasional, UMKM Indonesia menghadapi kerentanan operasional, logistik, dan analitis yang berpotensi mengikis loyalitas pelanggan secara sistemik. Karakteristik geografis kepulauan Indonesia menciptakan disparitas tarif ekspedisi dan durasi pengiriman yang tajam antara rute intra-kota atau antarkota dalam satu pulau dibandingkan koridor antarpulau serta Wilayah Tertinggal, Terdepan, dan Terluar (3T). Pada transaksi produk UMKM bernilai rendah, porsi ongkos kirim riil kerap membengkak melampaui separuh hingga melebihi nilai bruto barang, sehingga pembeli sangat bergantung pada subsidi gratis ongkos kirim maupun kupon potongan harga dari platform; ketiadaan insentif yang memadai berisiko memicu sensitivitas biaya dan

Prio Kustanto: *Penulis Korespondensi
Copyright © 2027, Mayadi, Prio Kustanto, Dismita Citra Dewi, Lenita Waty, Hendarman Lubis.

Journal of Information Technology, Software Engineering and Computer Science
Website : https://ejournal.techcart-press.com/index.php/itsecs
E-ISSN 2962-0635
P-ISSN 2962-0538
Volume 5, Number 1, January 2027
DOI : **https://doi.org/10.58602/itsecs.v5i1.498**
Page 135-158

keengganan berbelanja ulang. Kerentanan tersebut diperparah oleh variasi ketepatan waktu layanan ekspedisi antarkoridor yang rentan mengalami keterlambatan moderat hingga keterlambatan kritis lebih dari tiga hari, yang berdampak langsung pada penurunan peringkat produk dan naiknya kerentanan perpindahan pelanggan. Di sisi pemodelan data, banyak sistem prediksi perilaku berbasis pembelajaran ansambel bekerja sebagai kotak hitam (*black-box*) yang tidak menyertakan penjelasan kausal bagi pengambil keputusan. Lebih krusial lagi, korpus data transaksi yang diperkaya dengan skor loyalitas turunan, narasi penalaran otomatis, serta ulasan konsumen rentan menyimpan kebocoran target (*target leakage*) dan ketidaksesuaian lintas-modal (*cross-modal inconsistency*), yang apabila tidak diaudit secara disiplin akan menghasilkan model dengan performa laboratorium yang bias namun mengalami penurunan generalisasi saat diterapkan pada strategi retensi nyata.

Sejumlah studi metodologis dan terapan telah memajukan pemodelan prediksi perilaku konsumen, pembatalan pesanan, dan perpindahan pelanggan menggunakan algoritma pohon keputusan ansambel serta kerangka kerja interpretabilitas. Secara teoretis, Breiman merumuskan *Random Forest* melalui agregasi *bootstrap* dan pemilihan fitur acak untuk menekan varians pohon keputusan tunggal serta menjaga ketahanan terhadap *noise* berdimensi tinggi, sementara Friedman [2] meletakkan fondasi *gradient boosting* aditif yang kemudian dioptimalkan skalabilitas, penanganan data jarang (*sparsity-aware*), dan regularisasi orde keduanya oleh Chen dan Guestrin [3] dalam arsitektur *Extreme Gradient Boosting* (*XGBoost*). Pada konteks data perilaku konsumen yang memiliki rasio kelas minoritas rendah , Chawla dkk. mengembangkan teknik *Synthetic Minority Over-sampling Technique* (*SMOTE*) untuk menginterpolasi sampel sintetis kelas minoritas secara geometris [4], yang selanjutnya diterapkan oleh Dhini dan Fauzan dalam kerangka pembelajaran ansambel guna memprediksi *customer churn* pada studi kasus perusahaan layanan *fixed broadband* [5]. Dalam ranah interpretabilitas model, Lundberg dan Lee [6] menyatukan enam metode atribusi aditif melalui kerangka *SHapley Additive exPlanations* (*SHAP*) berbasis nilai Shapley teori permainan kooperatif. Di sektor niaga-el Indonesia dan global, Samudra [7] mengomparasikan *Random Forest* dan *XGBoost* dengan interpretasi *SHAP* untuk memprediksi pembatalan pesanan di Shopee Indonesia, sedangkan Bernardes menunjukkan bahwa kombinasi pemodelan pembelajaran mesin berbasis pohon dan penjelasan *SHAP* mampu menguraikan pendorong keputusan pembelian daring secara transparan. Temuan serupa diperkuat oleh studi komparatif pengklasifikasi berbasis *SHAP* oleh Ardhani dan Tania serta studi empiris pemasaran digital UMKM Indonesia yang menegaskan peran strategis gratis ongkos kirim, diskon voucher, kualitas layanan, dan kepuasan pelanggan dalam membentuk niat beli ulang.

Untuk mengatasi keempat kesenjangan tersebut, penelitian ini mengusulkan sebuah kerangka kerja analitis tiga lapis yang menggabungkan audit integritas lintas-modal pra-pemodelan, arsitektur klasifikasi komparatif *Random Forest* dan *XGBoost* yang terisolasi dari kebocoran target, serta interpretabilitas *TreeSHAP* yang dipetakan langsung ke dalam strategi retensi pelanggan UMKM Indonesia. Kebaruan pendekatan yang diusulkan berpusat pada protokol verifikasi identitas aritmetika dan pengujian determinisme fitur sebelum proses pelatihan dijalankan, sehingga atribut turunan pembawa label masa depan maupun teks ulasan yang mengandung inkonsistensi kategori diisolasi secara sistematis dari ruang prediktor. Pada lapisan pemodelan, kerangka kerja ini menyandingkan mekanisme pengurangan varians melalui agregasi pohon acak [8] dengan pengurangan bias melalui optimasi gradien orde kedua teregularisasi, disertai penerapan penyeimbangan *SMOTE* [1] secara ketat di dalam lipatan pelatihan validasi silang terstratifikasi agar tidak terjadi kontaminasi sampel sintetis terhadap himpunan evaluasi. Lebih jauh lagi, penelitian ini menerapkan standar pelaporan ilmiah transparan terhadap keterbatasan lingkungan komputasi dan ketidaktersediaan korpus penuh, dengan

Prio Kustanto: *Penulis Korespondensi
Copyright © 2027, Mayadi, Prio Kustanto, Dismita Citra Dewi, Lenita Waty, Hendarman Lubis.

Journal of Information Technology, Software Engineering and Computer Science
Website : https://ejournal.techcart-press.com/index.php/itsecs
E-ISSN 2962-0635
P-ISSN 2962-0538
Volume 5, Number 1, January 2027
DOI : **https://doi.org/10.58602/itsecs.v5i1.498**
Page 135-158

memfokuskan verifikasi kuantitatif pada catatan transaksi yang utuh teraudit serta menyintesiskan hasilnya bersama tolok ukur literatur ber-DOI untuk melahirkan protokol retensi adaptif.

Berdasarkan uraian latar belakang, tantangan empiris, dan kesenjangan tersebut, penelitian ini bertujuan menjawab empat pertanyaan utama: (1) bagaimana integritas numerik, kelengkapan skema, dan konsistensi lintas-modal pada korpus data transaksi perilaku konsumen UMKM *e-commerce* Indonesia; (2) bagaimana arsitektur protokol pemodelan *Random Forest* dan *XGBoost* bebas kebocoran target dirancang beserta metrik evaluasi *accuracy*, *precision*, *recall*, *F1-score*, dan *ROC-AUC* di bawah kondisi ketidakseimbangan kelas; (3) bagaimana mekanisme eksplanasi *TreeSHAP* menguraikan kontribusi fitur logistik, struktur harga, subsidi promosi, kanal, dan pembayaran secara aditif serta konsisten; dan (4) strategi retensi pelanggan berbasis bukti apa yang dapat diturunkan secara operasional bagi UMKM Indonesia dari pola empiris terverifikasi dan literatur terkait. Secara teoretis dan metodologis, studi ini berkontribusi pada perumusan protokol audit kebocoran target dan tata kelola data pengayaan sintetik pada analitik perilaku konsumen ritel. Secara praktis, artikel ini menghadirkan panduan segmentasi intervensi logistik, subsidi ongkos kirim, serta mitigasi risiko pembayaran bagi pelaku UMKM lintas platform dan koridor pengiriman Nusantara. Bagian selanjutnya menguraikan secara sistematis rancangan metodologi penelitian dari tahap audit korpus hingga formulasi matematis pemodelan prediktif dan dekomposisi nilai Shapley.

2. METODE PENELITIAN

2.1. Desain Penelitian, Protokol Audit Korpus, dan Keterbatasan Lingkungan Eksekusi

Penelitian ini mengadopsi desain kuantitatif-komputasional yang memadukan audit forensik kualitas data (*data integrity auditing*), perancangan arsitektur pemodelan klasifikasi berbasis ansambel pohon keputusan (*Random Forest* dan *XGBoost*), serta kerangka interpretabilitas *TreeSHAP* untuk merumuskan strategi retensi UMKM *e-commerce* Indonesia. Objek kajian berupa korpus transaksi perdagangan elektronik UMKM lintas provinsi di Indonesia yang terekam pada rentang waktu 6 Februari 2024 pukul 08:59:38 hingga 14 Oktober 2026 pukul 10:42:06, dengan penomoran identifikasi transaksi terpantau membentang dari TRX_UMKM_2026_00001 sampai TRX_UMKM_2026_01071. Struktur skema korpus terdiri atas dua puluh sembilan kolom yang mencakup dimensi kategori produk, kanal platform lokapasar, provinsi asal penjual dan tujuan pembeli, klasifikasi wilayah pengiriman, komponen finansial bruto hingga bersih, insentif voucher dan subsidi ongkos kirim, metode pembayaran, mitra ekspedisi, durasi dan status ketepatan waktu pengiriman, peringkat produk, status pembelian ulang, demografi pembeli, teks ulasan beserta metrik panjangnya, stempel waktu transaksi, serta atribut pengayaan seperti indeks sensitivitas ongkos kirim, skor loyalitas pelanggan, kategori risiko *churn*, rasio efektivitas diskon, narasi penalaran konsumen, dan tingkat kepercayaan posterior.

Sebelum pelatihan algoritma klasifikasi dijalankan, protokol audit penelitian melakukan pemeriksaan menyeluruh terhadap ketersediaan berkas korpus utuh serta kesiapan modul eksekusi pada lingkungan komputasi. Hasil audit teknis menyingkap dua kendala objektif yang membatasi eksekusi empiris secara penuh. Pertama, berkas tabular lengkap tidak terpasang pada sistem direktori lokal lingkungan analitik, sehingga akses terhadap korpus hanya tersedia dalam bentuk cuplikan fragmen yang terputus di beberapa bagian. Melalui penarikan *verbatim* dan verifikasi silang antarfragmen, berhasil dihimpun tepat tiga puluh empat catatan transaksi yang utuh seluruh kolom numerik serta kategorikalnya ($n = 34$, bertanggal 6 Maret 2024 hingga 10 September 2026), di samping sejumlah catatan parsial

Prio Kustanto: *Penulis Korespondensi
Copyright © 2027, Mayadi, Prio Kustanto, Dismita Citra Dewi, Lenita Waty, Hendarman Lubis.

Journal of Information Technology, Software Engineering and Computer Science
Website : https://ejournal.techcart-press.com/index.php/itsecs
E-ISSN 2962-0635
P-ISSN 2962-0538
Volume 5, Number 1, January 2027
DOI : https://doi.org/10.58602/itsecs.v5i1.498
Page 135-158

yang terpotong pada kolom naratif atau kehilangan kolom penutup. Kedua, pemeriksaan perangkat komputasi mengonfirmasi bahwa modul khusus untuk mengeksekusi pengklasifikasi *XGBoost* serta komputasi nilai *SHAP* belum terpasang di lingkungan eksekusi yang tersedia. Sesuai prinsip integritas akademik dan transparansi ilmiah, studi ini secara tegas tidak memfabrikasi angka keluaran *accuracy*, *precision*, *recall*, *F1-score*, *ROC-AUC*, maupun koefisien numerik *TreeSHAP* dari model yang belum dapat dijalankan pada populasi penuh. Sebaliknya, penelitian ini melaksanakan verifikasi matematis eksak dan analisis deskriptif-komparatif pada seluruh 34 catatan transaksi terverifikasi, merancang arsitektur alur pemodelan prediktif dan isolasi kebocoran secara lengkap untuk eksekusi skala penuh, serta menyintesis temuan tersebut dengan tolok ukur empiris dari publikasi jurnal bereputasi.

2.2. Taksonomi Skema Atribut dan Isolasi Kebocoran Target

Pada korpus perilaku konsumen yang telah diperkaya dengan skor komposit serta narasi penalaran otomatis, kesalahan paling fatal dalam pemodelan prediktif adalah membiarkan fitur turunan pasca-kejadian (*post-outcome attributes*) masuk ke dalam matriks predictor [9]. Apabila fitur pembentuk label atau atribut yang diturunkan dari label tetap dipertahankan sebagai masukan model, pengklasifikasi berbasis pohon keputusan akan membangun aturan pemecahan semu dengan performa evaluasi tinggi namun tidak memiliki kegunaan praktis bagi sistem deteksi dini UMKM [10]. Berdasarkan audit terhadap seluruh dua puluh sembilan kolom pada skema korpus, Tabel 1 mengklasifikasikan fungsi setiap atribut serta status isolasinya dalam alur pemodelan.

Tabel 1. Taksonomi Atribut dan Isolasi Kebocoran Target

Kelompok Atribut	Nama Kolom pada Skema	Tipe Data	Peran Analitis dalam Arsitektur Pemodelan	Status Isolasi Kebocoran Target
Identitas dan Waktu	transaksi_id, timestamp_transaksi	Teks, Tanggal-Waktu	Pengindeksan unik catatan; ekstraksi fitur temporal (jam pembelian, hari pekan, indikator akhir pekan, kuartal) untuk menangkap pola belanja musiman	transaksi_id dieksklusi; fitur kalender diekstraksi dari stempel waktu
Produk dan Kanal	kategori_produk, platform_ecommerce	Kategorikal	Prediktor utama segmentasi produk UMKM (7 kategori terpantau) dan karakteristik trafik lokapasar serta perdagangan sosial (5 platform)	Prediktor sah (Dipertahankan)
Geografi dan Rute	provinsi_penjual, provinsi_pembeli, wilayah_pengiriman	Kategorikal	Prediktor ketimpangan logistik kepulauan lintas provinsi (Intra-Kota, Antar-Kota Satu Pulau, Antar-Pulau Jawa-Luar Jawa, Wilayah Tertinggal / 3T)	Prediktor sah (Dipertahankan)
Struktur Finansial	nilai_transaksi_bruto_idr, diskon_voucher_platform_idr, ongkos_kirim_riil_idr, subsidi_gratis_o	Numerik Bulat dan Desimal	Mengukur skala nilai keranjang belanja, intensitas potongan voucher, tarif ongkos kirim riil, kompensasi subsidi ongkir, serta rasio efektivitas insentif	Prediktor sah dengan pengendalian multikolinearitas komposisi aritmetika

Prio Kustanto: *Penulis Korespondensi
Copyright © 2027, Mayadi, Prio Kustanto, Dismita Citra Dewi, Lenita Waty, Hendarman Lubis.

Journal of Information Technology, Software Engineering and Computer Science
Website : https://ejournal.techcart-press.com/index.php/itsecs
E-ISSN 2962-0635
P-ISSN 2962-0538
Volume 5, Number 1, January 2027
DOI : **https://doi.org/10.58602/itsecs.v5i1.498**
Page 135-158

Pembayaran dan Logistik	ngkir_idr, nilai_transaksi_netto_idr, rasio_efektivitas_diskon metode_pembayaran, layanan_ekspedisi, durasi_pengiriman_hari, status_keterlambatan_pengiriman	Kategorikal dan Numerik Bulat	Prediktor friksi transaksi pembayaran (COD, QRIS, dompet digital, Paylater, Virtual Account) dan reliabilitas pemenuhan SLA pengiriman	Prediktor sah; pada pemodelan risiko *churn* wajib diaudit terhadap ketergantungan aturan loyalitas
Umpan Balik dan Demografi	rating_produk, usia_pembeli, gender_pembeli, panjang_karakter_ulasan, jumlah_kata_ulasan	Numerik dan Kategorikal	Prediktor evaluasi kepuasan barang pascaterima, profil demografis pembeli, serta proksi panjang elaborasi ulasan konsumen	Demografi sah sejak awal; rating_produk dan metrik ulasan diisolasi pada mode *pre-rating*
Variabel Target Kandidat	status_repeat_order, kategori_risiko_churn	Biner dan Multikelas Ordinal	Target Skenario A: prediksi pembelian ulang (Ya / Tidak). Target Skenario B: klasifikasi kategori risiko *churn* (Rendah, Sedang, Tinggi, Kritis)	Variabel Target (Dilarang saling masuk sebagai prediktor)
Atribut Pengayaan Terisolasi	indeks_sensitivitas_ongkir, skor_loyalitas_pelanggan, penalaran_konsumen_cot, posterior_confidence, teks_ulasan_konsumen	Desimal dan Teks Bebas	skor_loyalitas_pelanggan, penalaran_konsumen_cot, dan posterior_confidence membawa bocoran label; teks_ulasan_konsumen mengandung anomali lintas-modal	Dieksklusi dari fitur prediktor utama (kecuali evaluasi terpisah pada indeks_sensitivitas_ongkir)

Melalui taksonomi pada Tabel 1, kerangka eksperimen prediktif dipetakan menjadi dua formulasi skenario yang bersih secara kausal dan bernilai manajerial bagi UMKM. Pada **Skenario A (Prediksi Retensi Pembelian Ulang)**, variabel target biner adalah status_repeat_order (Ya sebagai kelas positif retensi dan Tidak sebagai kelas negatif), dengan mengecualikan skor_loyalitas_pelanggan, kategori_risiko_churn, penalaran_konsumen_cot, posterior_confidence, serta teks_ulasan_konsumen dari himpunan masukan model. Pada **Skenario B (Prediksi Risiko Churn Dini)**, sasaran klasifikasi berupa kategori risiko kategori_risiko_churn (baik dalam skema empat kelas ordinal maupun skema biner berisiko tinggi yang menggabungkan Tinggi dan Kritis melawan Rendah dan Sedang), yang dijalankan secara ketat dalam mode *pre-rating* (sebelum rating bintang dan status *repeat order* terbentuk). Alasan pengisolasian rating_produk, status_repeat_order, dan skor_loyalitas_pelanggan pada Skenario B didasarkan pada temuan audit bahwa ketiga variabel tersebut terikat secara deterministik membentuk label kategori_risiko_churn, sehingga mode *pre-rating* memaksa *Random*

Prio Kustanto: *Penulis Korespondensi
Copyright © 2027, Mayadi, Prio Kustanto, Dismita Citra Dewi, Lenita Waty, Hendarman Lubis.

Journal of Information Technology, Software Engineering and Computer Science
Website : https://ejournal.techcart-press.com/index.php/itsecs
E-ISSN 2962-0635
P-ISSN 2962-0538
Volume 5, Number 1, January 2027
DOI : **https://doi.org/10.58602/itsecs.v5i1.498**
Page 135-158

Forest dan *XGBoost* untuk menggali sinyal prediktif asli dari harga, ongkos kirim riil, subsidi platform, kanal, metode pembayaran, koridor pengiriman, dan mitra ekspedisi.

2.3. Rekayasa Fitur Rasio, Pengkodean Atribut, dan Penyeimbangan Kelas Berbasis SMOTE

Tahap pra-pemrosesan dirancang untuk mentransformasikan himpunan transaksi mentah menjadi matriks fitur terstruktur yang menjaga invariansi skala pohon keputusan sekaligus memperjelas daya interpretasi *TreeSHAP*. Oleh karena pemisahan simpul pada *Random Forest* dan *XGBoost* bergantung pada pengurutan peringkat nilai fitur, variabel numerik kontinu dan diskrit tidak memerlukan standardisasi maupun normalisasi min-maks [4], [11]. Namun demikian, untuk menyingkap interaksi nonlinier antara skala belanja produk dan friksi logistik yang tidak tampak dari nominal rupiah mentah secara terpisah, dirancang empat fitur rasio komplementer yang diturunkan langsung dari kolom transaksi terverifikasi [12]:

$$r\_"ongkir" = "ongkos\_kirim\_riil\_idr" / "nilai\_transaksi\_bruto\_idr", \quad r\_"beban" = \max(0, "ongkos\_kirim\_riil\_idr" - "subsidi\_gratis\_ongkir\_idr") / "nilai\_transaksi\_bruto\_idr"$$

$$r\_"ongkir" = "ongkos\_kirim\_riil\_idr" / "nilai\_transaksi\_bruto\_idr", \quad r\_"beban" = \max(0, "ongkos\_kirim\_riil\_idr" - "subsidi\_gratis\_ongkir\_idr") / "nilai\_transaksi\_bruto\_idr"$$

Keempat fitur turunan tersebut memiliki makna ekonomi langsung bagi pengambil keputusan UMKM: $r\_"ongkir"$ mengukur persentase tarif kurir mentah terhadap harga barang; $r\_"beban"$ mengukur sisa porsi ongkos kirim yang benar-benar membebani kantong pembeli setelah dikurangi subsidi; $r\_"cakupan"$ mengukur persentase tarif ekspedisi yang berhasil ditutup oleh subsidi gratis ongkos kirim; dan $r\_"hemat"$ membedakan transaksi di mana total subsidi melampaui ongkos kirim ($r\_"hemat" > 0$) dari transaksi di mana sisa ongkos kirim menyebabkan tagihan bersih melampaui harga barang bruto ($r\_"hemat" < 0$).

Untuk variabel kategorikal dengan jumlah nilai unik terbatas—yaitu kategori_produk, platform_ecommerce, wilayah_pengiriman, metode_pembayaran, dan layanan_ekspedisi—diterapkan pengkodean biner indikator (*one-hot encoding*) guna menjaga keterpisahan atribusi masing-masing kategori saat dipetakan oleh *TreeSHAP*, sehingga koefisien Shapley tidak terdistorsi oleh pengurutan angka arbitrer. Sementara itu, untuk kolom provinsi_penjual dan provinsi_pembeli yang memiliki kardinalitas tinggi hingga puluhan provinsi di Indonesia, digunakan pengkodean target beraturan di luar lipatan (*out-of-fold regularized target encoding*) atau pengelompokan klaster koridor geografis guna mencegah ledakan dimensi dan pemecahan pohon yang terlalu terfragmentasi. Dari stempel waktu timestamp_transaksi, diekstraksi komponen siklikal jam transaksi, indikator akhir pekan, serta kuartal waktu agar efek musiman dapat dimodelkan tanpa mengikutsertakan nomor identifikasi berurut[13].

Pada masalah prediksi perilaku konsumen ritel, distribusi kelas target lazim mengalami ketidakseimbangan struktural—sebagai gambaran, pada himpunan terverifikasi hanya 5,88% transaksi berada pada kategori risiko Kritis, 17,65% pada kategori Tinggi, dan 32,35% mencatat status status_repeat_order = 'Ya'. Ketidakseimbangan tersebut berpotensi mengarahkan pengklasifikasi untuk memprioritaskan akurasi kelas mayoritas dengan mengorbankan sensitivitas deteksi kelas minoritas berisiko tinggi. Untuk mengatasinya, kerangka kerja ini mengintegrasikan teknik *SMOTE* yang dirumuskan oleh Chawla dkk [14]. guna membangkitkan sampel sintetis kelas minoritas melalui interpolasi linier acak di sepanjang ruas penghubung pasangan tetangga terdekat kelas yang sama:

$$\mathbf{x}\_"sintetis" = \mathbf{x}\_i + \lambda \cdot (\mathbf{x}\_zi - \mathbf{x}\_i), \quad \lambda \sim \mathcal{U}(0,1)$$

Prio Kustanto: *Penulis Korespondensi
Copyright © 2027, Mayadi, Prio Kustanto, Dismita Citra Dewi, Lenita Waty, Hendarman Lubis.

Journal of Information Technology, Software Engineering and Computer Science
Website : https://ejournal.techcart-press.com/index.php/itsecs
E-ISSN 2962-0635
P-ISSN 2962-0538
Volume 5, Number 1, January 2027
DOI : **https://doi.org/10.58602/itsecs.v5i1.498**
Page 135-158

di mana x_i merupakan vektor fitur observasi kelas minoritas dan x_zi adalah salah satu dari k tetangga terdekatnya pada kelas tersebut. Untuk fitur gabungan numerik dan kategorikal, diterapkan varian *SMOTE-NC* (*Synthetic Minority Over-sampling Technique for Nominal and Continuous*) yang menghitung jarak fitur nominal melalui nilai median simpangan baku fitur kontinu. Secara metodologis, proses *oversampling* diisolasi secara eksklusif di dalam setiap lipatan pelatihan validasi silang terstratifikasi dan tidak pernah dijalankan pada keseluruhan himpunan sebelum pemisahan uji, guna menjamin ketiadaan kebocoran sampel sintetis ke himpunan validasi serta mencegah estimasi metrik *ROC-AUC* yang terinflasi.

2.4. Formulasi Matematis Random Forest dan XGBoost

Kerangka pemodelan komparatif memadukan dua arsitektur ansambel pohon keputusan yang bertumpu pada prinsip pengurangan galat yang saling melengkapi: paradigma pengurangan varians secara paralel melalui *Random Forest* dan paradigma pengurangan bias secara sekuensial melalui *XGBoost*.

Algoritma *Random Forest* yang diperkenalkan oleh Breiman mengonstruksi himpunan sebanyak T pohon klasifikasi independen $\{h(\mathbf{x}, \Theta\_t)\}\_(t = 1)\^T$, di mana vektor parameter acak Θ_t mengendalikan dua sumber stokastisitas: pertama, pengambilan sampel baris *bootstrap* $\mathcal{D}\_t$ dengan pengembalian dari dataset pelatihan $\mathcal{D}$; dan kedua, pemilihan subset acak fitur berjumlah $m\_"try" < p$ dari total p fitur prediktor pada setiap simpul pemecahan internal [1]. Penentuan fitur terbaik dan titik potong optimal pada setiap simpul memaksimalkan penurunan ketidakmurnian Gini (*Gini impurity*):

$$G(S) = 1 - \sum\_(c = 1)\^C \blacksquare p\_c\^2\,, \quad \Delta G(S, a) = G(S) - (|S\_L|/|S|\ G(S\_L) + |S\_R|/|S|\ G(S\_R))$$

di mana p_c adalah proporsi sampel kelas c di dalam simpul S dengan C kelas target, sedangkan S_L dan S_R merupakan simpul anak kiri dan kanan hasil pemisahan atribut a. Prediksi probabilitas kelas pada sebuah transaksi uji $\mathbf{x}$ diperoleh melalui rata-rata proporsi daun di seluruh pohon ansambel:

$$\hat{P}\_"RF"\ (y = c \mid \mathbf{x}) = 1/T \sum\_(t = 1)\^T \blacksquare \hat{p}\_t\ (y = c \mid \mathbf{x})$$

Dengan membatasi jumlah fitur kandidat pada setiap pemecahan dan merata-ratakan prediksi ratusan pohon yang terdekorrelasi, *Random Forest* menekan varians prediktif secara substansial tanpa menaikkan bias, sehingga relatif tahan terhadap pencilan nilai belanja ataupun *noise* pencatatan.

Berbeda dengan mekanisme paralel *Random Forest*, algoritma *XGBoost* yang dikembangkan oleh Chen dan Guestrin memperluas metode *gradient boosting* dari Friedman dengan membangun barisan pohon regresi CART $\{f\_k\}\_(k = 1)\^t$ secara aditif, di mana setiap pohon baru mengoreksi kesalahan prediksi kumulatif dari seluruh pohon sebelumnya:

$$\hat{y}\_i\^((t)) = \sum\_(k = 1)\^t \blacksquare f\_k\ (\mathbf{x}\_i) = \hat{y}\_i\^((t-1)) + \eta f\_t\ (\mathbf{x}\_i), \quad f\_k \in \mathcal{F}$$

dengan $\eta \in (0,1]$ melambangkan laju penyusutan pembelajaran (*learning rate*) untuk mengerem langkah koreksi dan $\mathcal{F}$ merepresentasikan ruang fungsi seluruh pohon keputusan. Guna mengendalikan *overfitting* pada dataset berfitur majemuk, *XGBoost* mengoptimalkan fungsi objektif teregularisasi pada setiap langkah iterasi t:

$$\mathcal{L}\^((t)) = \sum\_(i = 1)\^n \blacksquare l\,(y\_i, \hat{y}\_i\^((t-1)) + f\_t\,(\mathbf{x}\_i)) + \Omega(f\_t), \quad \Omega(f\_t) = \gamma K + 1/2\ \lambda \sum\_(j = 1)\^K \blacksquare w\_j\^2 + \alpha \sum\_(j = 1)\^K \blacksquare |w\_j|$$

di mana $l(\cdot)$ menyatakan fungsi kerugian *log-loss* biner atau entropi silang multikelas, K adalah jumlah simpul daun pohon ke-t, w_j merupakan bobot prediksi pada daun ke-j, γ adalah penalti kompleksitas penambahan simpul daun, serta λ dan α berturut-turut adalah koefisien regularisasi L_2 dan L_1 atas bobot daun [3]. Melalui ekspansi deret Taylor orde

Prio Kustanto: *Penulis Korespondensi
Copyright © 2027, Mayadi, Prio Kustanto, Dismita Citra Dewi, Lenita Waty, Hendarman Lubis.

Journal of Information Technology, Software Engineering and Computer Science
Website : https://ejournal.techcart-press.com/index.php/itsecs
E-ISSN 2962-0635
P-ISSN 2962-0538
Volume 5, Number 1, January 2027
DOI : **https://doi.org/10.58602/itsecs.v5i1.498**
Page 135-158

kedua atas fungsi kerugian dengan gradien pertama $g\_i = \partial\_(y\hat{}((t-1)))\, l(y\_i, y\hat{}((t-1)))$ dan turunan kedua (Hessian) $h\_i = \partial\_(y\hat{}((t-1)))\hat{}2\, l(y\_i, y\hat{}((t-1)))$, skor perolehan struktural (*gain*) untuk mengevaluasi pemecahan simpul I menjadi partisi kiri I_L dan kanan I_R dirumuskan secara analitis sebagai:

$$"Gain" = 1/2\,[(\sum\_(i \in I\_L) g\_i)\hat{}2/(\sum\_(i \in I\_L) h\_i + \lambda) + (\sum\_(i \in I\_R) g\_i)\hat{}2/(\sum\_(i \in I\_R) h\_i + \lambda) - (\sum\_(i \in I) g\_i)\hat{}2/(\sum\_(i \in I) h\_i + \lambda)] - \gamma$$

Pemecahan simpul hanya dipertahankan apabila nilai perolehan melampaui penalti γ, sementara mekanisme penyampelan acak baris dan kolom per pohon—yang menurut Chen dan Guestrin terinspirasi dari *Random Forest*—semakin memperkuat kemampuan generalisasi algoritma. Tabel 2 menyajikan ruang pencarian hiperparameter yang dirancang untuk penyetelan sistematis kedua model pada korpus perilaku konsumen UMKM.

Tabel 2. Ruang Hiperparameter Random Forest dan XGBoost

Algoritma	Parameter Matematis Model	Ruang Kandidat Pencarian	Peran dalam Pengendalian Bias-Varians dan Regularisasi
Random Forest	Jumlah pohon ansambel (T)	200; 500; 1.000	Menstabilkan konvergensi estimasi probabilitas kelas melalui agregasi *bagging*
Random Forest	Kedalaman maksimum pohon ($d\_\max$)	6; 10; 15; Tanpa batas	Membatasi orde interaksi antarfitur agar pohon tidak menghafal *noise* sampel mikro
Random Forest	Jumlah fitur per pemecahan ($m\_"try"$)	$\lfloor\sqrt{p}\rfloor$; $\lfloor\log\_2 p\rfloor$; $0{,}5p$	Mendekorelasikan struktur antar-pohon sehingga fitur dominan tidak menutupi sinyal sekunder
Random Forest	Sampel minimum simpul daun ($n\_\min$)	1; 3; 5; 10	Menghaluskan probabilitas di daun terminal, krusial saat menghadapi kategori berfrekuensi rendah
Random Forest	Skema pembobotan kelas	Seimbang terbalik terhadap frekuensi; Seragam	Mengompensasi ketimpangan rasio *churn* tinggi tanpa harus selalu melakukan duplikasi sampel
XGBoost	Jumlah putaran boosting (B)	100; 300; 600 (dengan penghentian dini)	Mengatur banyaknya koreksi residu berurutan sebelum terjadinya *overfitting*
XGBoost	Laju pembelajaran (η)	0,01; 0,05; 0,10	Memperkecil langkah pembaruan setiap pohon agar konvergensi gradien berlangsung lebih halus
XGBoost	Kedalaman maksimum pohon ($d\_\max$)	3; 5; 7	Menentukan derajat interaksi fitur tertinggi yang diizinkan (misal rute × ongkir × subsidi)
XGBoost	Rasio subsampling baris dan kolom (ρ_r, ρ_c)	0,70; 0,85; 1,00	Mengintroduksi keragaman stokastik pada baris dan fitur guna menekan varians model
XGBoost	Regularisasi bobot daun dan pemangkasan (λ, α, γ)	$\lambda \in \{1{,}5{,}10\}$; $\alpha \in \{0; 0{,}5\}$; $\gamma \in \{0; 0{,}1; 0{,}5\}$	Menahan pertumbuhan cabang berperolehan gradien kecil serta menstabilkan bobot daun ekstrem

Prio Kustanto: *Penulis Korespondensi
Copyright © 2027, Mayadi, Prio Kustanto, Dismita Citra Dewi, Lenita Waty, Hendarman Lubis.

Journal of Information Technology, Software Engineering and Computer Science
Website : https://ejournal.techcart-press.com/index.php/itsecs
E-ISSN 2962-0635
P-ISSN 2962-0538
Volume 5, Number 1, January 2027
DOI : https://doi.org/10.58602/itsecs.v5i1.498
Page 135-158

2.5. Protokol Validasi Terstratifikasi dan Metrik Evaluasi Klasifikasi

Untuk mengukur daya generalisasi di luar sampel pelatihan tanpa mengorbankan proporsi kelas minoritas, rancangan validasi menerapkan pembagian himpunan latih dan uji terstratifikasi dengan proporsi 80:20, yang dikombinasikan dengan validasi silang terstratifikasi 5-lipatan atau 10-lipatan (*stratified k-fold cross-validation*) pada bagian pelatihan guna memilih konfigurasi hiperparameter terbaik. Teknik stratifikasi menjaga agar persentase kasus risiko Kritis dan Tinggi serta kelas status_repeat_order = 'Ya' tersebar proporsional pada seluruh lipatan[15]. Kinerja prediktif kedua algoritma dievaluasi secara komprehensif menggunakan lima metrik berbasis matriks konfusi (*True Positive* TP, *True Negative* TN, *False Positive* FP, dan *False Negative* FN).
Metrik pertama adalah *Accuracy* yang menghitung rasio keseluruhan prediksi tepat terhadap total observasi uji:

$$"Accuracy" = (TP + TN)/(TP + TN + FP + FN)$$

Mengandalkan *Accuracy* secara tunggal pada dataset tidak seimbang dapat menghasilkan kesimpulan menyesatkan karena pengklasifikasi yang hanya memprediksi kelas mayoritas tetap tampak berkinerja tinggi. Oleh karena itu, analisis wajib menyertakan *Precision* (mengukur proporsi prediksi positif yang benar-benar positif, berperan menjaga efisiensi anggaran promosi UMKM agar subsidi tidak salah sasaran ke pelanggan yang sudah loyal) dan *Recall* (mengukur proporsi kasus positif aktual yang berhasil ditangkap model, berperan krusial menekan risiko lolosnya pelanggan berpotensi *churn*), bersama *F1-score* sebagai rata-rata harmonik keduanya[16]:

$$"Precision" = TP/(TP + FP), \quad "Recall" = TP/(TP + FN), \quad "F1 - score" = (2 \times "Precision" \times "Recall")/("Precision" + "Recall")$$

Pada skema multikelas empat tingkat risiko *churn*, pelaporan diarahkan pada nilai *macro-averaged Precision*, *Recall*, dan *F1-score* agar kelas Kritis dan Tinggi memperoleh bobot evaluasi setara dengan kelas mayoritas Rendah. Selanjutnya, metrik kelima yaitu *Area Under the Receiver Operating Characteristic Curve* (*ROC-AUC*) mengukur probabilitas bahwa pengklasifikasi memberikan skor risiko lebih tinggi kepada satu sampel positif acak dibandingkan satu sampel negatif acak di seluruh rentang ambang batas keputusan $\tau \in [0,1]$.

$$"ROC - AUC" = \int \_0\text{^}1▒"TPR" \ (〖"FPR"〗 \text{^}(-1)\ (u))du, \quad "TPR" = TP/(TP + FN), \quad "FPR" = FP/(FP + TN)$$

Keunggulan *ROC-AUC* terletak pada sifatnya yang invarian terhadap pemilihan satu ambang batas diskrit tertentu, sehingga pemilik UMKM dapat menyetel batas probabilitas pemicu kampanye retensi secara fleksibel sesuai kapasitas anggaran operasional bulanan[17].

2.6. Kerangka Kerja Explainable AI Menggunakan TreeSHAP

Untuk mengubah keluaran probabilistik *Random Forest* dan *XGBoost* menjadi panduan strategi retensi yang dapat dipertanggungjawabkan, penelitian ini memanfaatkan kerangka *SHAP* (*SHapley Additive exPlanations*) dari Lundberg dan Lee melalui implementasi eksak *TreeSHAP* yang diperkenalkan oleh Lundberg dkk. Pendekatan ini mengatasi kelemahan pengukuran kepentingan fitur bawaan berbasis reduksi ketidakmurnian Gini atau frekuensi pemecahan pohon yang kerap tidak konsisten dan bias ke arah variabel kontinu atau kategorikal berkardinalitas tinggi [18].
Secara formal, *SHAP* merepresentasikan penjelasan atas keluaran model $f(\mathbf{x})$ pada sebuah transaksi spesifik $\mathbf{x}$ melalui model penjelasan aditif linier $g(\mathbf{z}')$ atas vektor biner fitur sederhana $\mathbf{z}' \in \{0,1\}\text{^}M$:

Prio Kustanto: *Penulis Korespondensi
Copyright © 2027, Mayadi, Prio Kustanto, Dismita Citra Dewi, Lenita Waty, Hendarman Lubis.

Journal of Information Technology, Software Engineering and Computer Science
Website : https://ejournal.techcart-press.com/index.php/itsecs
E-ISSN 2962-0635
P-ISSN 2962-0538
Volume 5, Number 1, January 2027
DOI : **https://doi.org/10.58602/itsecs.v5i1.498**
Page 135-158

$$g(\mathbf{z}') = \phi\_0 + \sum\_(i = 1)\^M \phi\_i \; z\_i'$$

di mana $\phi\_0 = \mathbb{E}[f(\mathbf{X})]$ menyatakan nilai ekspektasi dasar (*base value*) pada populasi latar belakang dan $\phi\_i \in \mathbb{R}$ mengukur besaran serta arah kontribusi fitur ke-i terhadap selisih prediksi transaksi tersebut dari nilai dasar. Lundberg dan Lee [6] membuktikan bahwa solusi unik yang secara serentak memenuhi aksioma akurasi lokal (*local accuracy*), ketiadaan atribusi fitur absen (*missingness*), dan konsistensi (*consistency*) diberikan oleh nilai Shapley dari teori permainan kooperatif:

$$\phi\_i\,(f, \mathbf{x}) = \sum\_(S \subseteq F \setminus \{i\}) (|S|!\;(|F| - |S| - 1)!)/|F|!\;[f\_(S \cup \{i\})\,(\mathbf{x}\_(S \cup \{i\})) - f\_S\,(\mathbf{x}\_S)]$$

dengan F menyatakan himpunan seluruh fitur, S melambangkan setiap subhimpunan koalisi fitur yang tidak memuat fitur ke-i, serta $f\_S\,(\mathbf{x}\_S) = \mathbb{E}[f(\mathbf{X}) \mid \mathbf{X}\_S = \mathbf{x}\_S]$ menyatakan ekspektasi prediksi model ketika hanya himpunan fitur dalam S yang diketahui [6].
Berhubung perhitungan eksak nilai Shapley pada model umum menuntut waktu eksponensial $\mathcal{O}(2\^|F|)$, algoritma *TreeSHAP* [10] memanfaatkan struktur rekursif pohon keputusan dengan melacak proporsi seluruh koalisi fitur yang melewati cabang kiri dan kanan di sepanjang jalur pemecahan, sehingga mereduksi kompleksitas komputasi menjadi polinomial $\mathcal{O}(TLD\^2)$ dengan T jumlah pohon, L jumlah daun maksimum, dan D kedalaman maksimum pohon [10]. Selain atribusi individu ϕ_i, *TreeSHAP* memungkinkan dekomposisi interaksi berpasangan (*SHAP interaction values*) $\Phi\_(i,j)$ untuk mengkuantifikasi efek sinergis antarvariabel—seperti interaksi antara ongkos_kirim_riil_idr dan subsidi_gratis_ongkir_idr, atau interaksi antara wilayah_pengiriman dan durasi_pengiriman_hari [10]. Melalui tiga bentuk keluaran visual—peringkat dampak global (*global summary* dari $1/n \sum\_(k = 1)\^n |\phi\_i\^((k))|$), kurva ketergantungan nonlinier (*dependence plot*) untuk menemukan titik balik ambang batas biaya, serta dekomposisi air terjun lokal (*local waterfall*) untuk memicu intervensi per transaksi—*TreeSHAP* menyediakan jembatan analitis antara pembelajaran mesin dan strategi bisnis UMKM [7, 10].

3. HASIL DAN PEMBAHASAN

3.1. Audit Integritas Aritmetika, Kelengkapan Catatan, dan Anomali Lintas-Modal pada Korpus Terverifikasi

Pelaksanaan audit data forensik atas fragmen korpus yang berhasil diekstraksi menyingkap gambaran ganda mengenai kualitas data: di satu sisi, variabel akuntansi dan promosi memiliki integritas matematis yang sangat presisi; namun di sisi lain, terdapat cacat kelengkapan baris serta ketidaksinkronan semantik antara atribut terstruktur dengan teks naratif. Pada pengujian terhadap seluruh tiga puluh empat catatan transaksi yang terverifikasi utuh (n = 34, mencakup transaksi dari TRX_UMKM_2026_00001 sampai TRX_UMKM_2026_01071), ditemukan bahwa kolom nilai_transaksi_netto_idr mematuhi persamaan identitas aritmetika eksak berikut:

$$\text{"}nilai\_transaksi\_netto\_idr\text{"} = \text{"}nilai\_transaksi\_bruto\_idr\text{"} - \text{"}diskon\_voucher\_platform\_idr\text{"} + (\text{"}ongkos\_kirim\_riil\_idr\text{"} - \text{"}subsidi\_gratis\_ongkir\_idr\text{"})$$

dengan galat absolut maksimum bernilai tepat nol rupiah di seluruh 34 catatan. Sejalan dengan hal tersebut, kolom rasio_efektivitas_diskon terbukti mengikuti formula rasio insentif gabungan terhadap nilai bruto produk:

$$\text{rasio\_efektivitas\_diskon} = \frac{\text{diskon\_voucher\_platform\_idr} + \text{subsidi\_gratis\_ongkir\_idr}}{\text{nilai\_transaksi\_bruto\_idr}}$$

Prio Kustanto: *Penulis Korespondensi
Copyright © 2027, Mayadi, Prio Kustanto, Dismita Citra Dewi, Lenita Waty, Hendarman Lubis.

Journal of Information Technology, Software Engineering and Computer Science
Website : https://ejournal.techcart-press.com/index.php/itsecs
E-ISSN 2962-0635
P-ISSN 2962-0538
Volume 5, Number 1, January 2027
DOI : **https://doi.org/10.58602/itsecs.v5i1.498**
Page 135-158

dengan deviasi absolut maksimum hanya sebesar 4,68 × 10^{-5} yang murni berasal dari pembulatan empat angka di belakang koma. Hasil ini mengonfirmasi bahwa kalkulasi moneter inti dalam dataset bersih dari kesalahan pencatatan akuntansi, sekaligus mengingatkan bahwa penyertaan simultan seluruh komponen moneter bruto, voucher, ongkos kirim, subsidi, dan nilai bersih dalam satu matriks fitur memerlukan perhatian terhadap pembagian bobot fitur antarvariabel yang berkorelasi erat.
Sebaliknya, inspeksi terhadap aspek struktural dan konten tekstual memperlihatkan lima kelompok anomali krusial yang dirangkum pada Tabel 3. Anomali tersebut meliputi keterputusan rekaman baris transaksi, hilangnya nilai kepercayaan posterior pada beberapa entri, diskrepansi usia antar-ekstraksi fragmen, kontradiksi antara kolom kategorikal terstruktur dengan isi teks ulasan konsumen (teks_ulasan_konsumen), serta inkonsistensi logika pada kolom narasi penalaran (penalaran_konsumen_cot).

Tabel 3. Ringkasan Anomali Data dan Implikasi Pemodelan

Kategori Anomali	Transaksi Teridentifikasi	Deskripsi Temuan Empiris	Implikasi Metodologis bagi Pemodelan Prediktif
Keterputusan Baris dan Nilai Hilang	TRX_UMKM_2026_00003, 00074, 00112, 00286, 00494, 00540, 00561, 00671, 01071	Sejumlah baris terpotong di tengah teks naratif sehingga struktur kolom terputus atau menyambung dengan baris lain; nilai posterior_confidence tidak tersedia pada fragmen 00540, 00671, dan 01071	Memerlukan penyaringan ketat atas integritas jumlah kolom per baris sebelum pemodelan, serta eksklusi posterior_confidence dari fitur prediksi
Diskrepansi Usia Lintas Fragmen	TRX_UMKM_2026_00100, TRX_UMKM_2026_01069	Ekstraksi fragmen berbeda mencatat nilai usia pembeli 52 versus 56 tahun pada transaksi 00100, serta 38 versus 39 tahun pada transaksi 01069	Menuntut validasi checksum berkas sumber asli sebelum melaporkan statistik titik tunggal demografi atau memakai usia sebagai fitur utama
Kontradiksi Kategori Produk vs. Ulasan Teks	TRX_UMKM_2026_00001, 00002, 00111, 00525, 00560, 00701, 00792, 01015, 01016, 01111	Kategori tercatat sebagai *Kriya & Kerajinan Tangan* namun ulasan memuji bahan batik (00001); tercatat *Kecantikan & Herbal* namun ulasan membahas bau pernis tas kerajinan (00002), kopi botolan (00560), atau keripik sambal (00792); tercatat *Fashion Muslim & Batik* namun ulasan memuji keripik sambal (00701, 01016, 01111)	Fitur teks maupun analisis sentimen berbasis pemrosesan bahasa alami atas kolom ulasan tidak boleh dipakai untuk mengklasifikasikan produk atau loyalitas karena tidak sinkron dengan SKU terstruktur
Kontradiksi Kurir/Durasi Sistem vs. Ulasan	TRX_UMKM_2026_00001, 00019, 00112	Sistem mencatat durasi 4 hari namun ulasan menulis kurir tiba dalam dua hari (00001); sistem mencatat status Tepat Waktu namun ulasan mengeluhkan pengiriman molor sehari (00019); teks naratif menyebut JNE 11 hari	Mengindikasikan bahwa kolom teks ulasan dihasilkan dari pengayaan templat sintetis secara terpisah dari mesin pembangkit variabel logistik tabular

Prio Kustanto: *Penulis Korespondensi
Copyright © 2027, Mayadi, Prio Kustanto, Dismita Citra Dewi, Lenita Waty, Hendarman Lubis.

Journal of Information Technology, Software Engineering and Computer Science
Website : https://ejournal.techcart-press.com/index.php/itsecs
E-ISSN 2962-0635
P-ISSN 2962-0538
Volume 5, Number 1, January 2027
DOI : **https://doi.org/10.58602/itsecs.v5i1.498**
Page 135-158

Kontradiksi Logika pada Narasi Penalaran	TRX_UMKM_2026_00018, 00701, 00702, 01015, 01016	versus kolom tabel Shopee Xpress 12 hari (00112) Kolom status_repeat_order bernilai Tidak namun teks penalaran_konsumen_cot menyertakan klausa baku *"dan retensi repeat order aktif"* karena kategori_risiko_churn berlabel Rendah	Mengindikasikan kuat bahwa penalaran_konsumen_cot dibangkitkan memakai templat berbasis label risiko *churn*, sehingga menjadi sumber kebocoran label langsung apabila dimasukkan ke model

Temuan pada Tabel 3 memberikan pelajaran metodologis yang sangat berharga dalam penggunaan korpus ritel yang diperkaya secara otomatis. Apabila peneliti secara naif mengekstraksi fitur teks dari teks_ulasan_konsumen atau penalaran_konsumen_cot ke dalam *Random Forest* dan *XGBoost*, algoritma akan mengeksploitasi frasa-frasa templat yang secara implisit menyandikan label kategori_risiko_churn—misalnya deskripsi mengenai solidnya loyalitas pada kasus Rendah versus kerentanan pindah ke kompetitor pada kasus Tinggi—sehingga menghasilkan kinerja prediktif yang tampak tinggi tetapi rapuh. Pengisolasian kedua kolom teks tersebut sebagaimana ditetapkan dalam Tabel 1 melindungi integritas eksperimen agar model benar-benar belajar dari sinyal operasional perdagangan elektronik yang autentik.

3.2. Penemuan Struktur Deterministik pada Skor Loyalitas dan Risiko Churn

Penyelidikan komputasional lanjutan terhadap tiga puluh empat transaksi terverifikasi mengungkap adanya struktur matematis deterministik yang mengatur pembentukan nilai skor_loyalitas_pelanggan dan kategori_risiko_churn. Pengelompokan data berdasarkan kombinasi tiga variabel—yakni rating_produk, status_keterlambatan_pengiriman, dan status_repeat_order—menghasilkan sebelas grup unik yang masing-masing memiliki simpangan baku nol (nilai minimum, maksimum, dan rata-rata identik) pada kolom skor_loyalitas_pelanggan, seperti disajikan pada Tabel 4.

Tabel 4. Pemetaan Deterministik Skor Loyalitas dan Risiko Churn

Peringkat Produk (rating_produk)	Status Keterlambatan Pengiriman	Status Pembelian Ulang (status_repeat_order)	Jumlah Sampel (n)	Skor Loyalitas Pelanggan (Min = Maks = Rerata)	Kategori Risiko Churn
1	Terlambat Kritis >3 Hari	Tidak	1	0,05	Kritis
2	Terlambat Kritis >3 Hari	Tidak	1	0,18	Kritis
2	Terlambat 1-2 Hari	Tidak	1	0,22	Tinggi
3	Terlambat Kritis >3 Hari	Tidak	1	0,30	Tinggi
3	Terlambat 1-2 Hari	Tidak	3	0,35	Tinggi
3	Tepat Waktu	Tidak	1	0,45	Tinggi
4	Terlambat 1-2 Hari	Tidak	3	0,48	Sedang
4	Tepat Waktu	Tidak	6	0,57	Sedang

Prio Kustanto: *Penulis Korespondensi
Copyright © 2027, Mayadi, Prio Kustanto, Dismita Citra Dewi, Lenita Waty, Hendarman Lubis.

Journal of Information Technology, Software Engineering and Computer Science
Website : https://ejournal.techcart-press.com/index.php/itsecs
E-ISSN 2962-0635
P-ISSN 2962-0538
Volume 5, Number 1, January 2027
DOI : **https://doi.org/10.58602/itsecs.v5i1.498**
Page 135-158

5	Tepat Waktu	Tidak	6	0,70	Rendah
4	Tepat Waktu	Ya	5	0,88	Rendah
5	Tepat Waktu	Ya	6	0,98	Rendah

Pemeriksaan numerik atas sebelas baris pada Tabel 4 memperlihatkan pola kontribusi aljabar yang konsisten. Pada kondisi pengiriman Tepat Waktu tanpa pembelian ulang (Tidak), setiap peningkatan satu bintang rating_produk dari 3 ke 4 dan 5 menaikkan skor_loyalitas_pelanggan dari 0,45 menjadi 0,57 dan 0,70. Ketika terjadi keterlambatan ringan Terlambat 1-2 Hari, skor tersebut berkurang sebesar 0,09 hingga 0,10 poin dibandingkan kondisi tepat waktu (turun dari 0,45 menjadi 0,35 pada peringkat 3, dan dari 0,57 menjadi 0,48 pada peringkat 4), sedangkan Terlambat Kritis >3 Hari menjatuhkan skor sekitar 0,15 poin (turun dari 0,45 menjadi 0,30 pada peringkat 3). Di pihak lain, terjadinya pembelian ulang (Ya) memberikan tambahan bobot sekitar +0,30 poin yang mendorong peringkat 4 tepat waktu dari 0,57 ke 0,88 serta mengangkat peringkat 5 tepat waktu menuju batas plafon 0,98. Secara analitis, sebelas kombinasi nilai tersebut selaras dengan fungsi komposit linier berbobot $0{,}50 \times (\text{rating\_produk} - 1)/4$ ditambah komponen ketepatan waktu (0,20 untuk tepat waktu; 0,10 untuk terlambat 1–2 hari; 0,05 untuk terlambat kritis) ditambah bonus 0,30 apabila status_repeat_order = 'Ya', dengan pembulatan aritmetika dua desimal dan batas atas pada angka 0,98. Selanjutnya, kategori_risiko_churn terpetakan langsung oleh ambang batas skor loyalitas: Kritis (< 0,20), Tinggi (0,22–0,45), Sedang (0,48–0,57), dan Rendah (≥ 0,70).

Penemuan hubungan deterministik tanpa varians ini membawa implikasi fundamental bagi pemodelan berbasis *Random Forest*, *XGBoost*, dan *TreeSHAP*. Pertama, oleh karena variabel seperti nilai_transaksi_bruto_idr, diskon_voucher_platform_idr, subsidi_gratis_ongkir_idr, indeks_sensitivitas_ongkir, metode_pembayaran, maupun wilayah_pengiriman tidak masuk ke dalam persamaan langsung pembentuk skor_loyalitas_pelanggan, maka pelatihan model untuk memprediksi kategori_risiko_churn sambil menyertakan rating_produk, status_keterlambatan_pengiriman, dan status_repeat_order akan membuat pohon keputusan mereplikasi aturan tangga Tabel 4 secara dangkal dan mendorong atribusi *TreeSHAP* bagi faktor harga serta promosi mendekati nol. Oleh karena itu, formulasi pemodelan yang bermakna secara ilmiah dan strategis bagi UMKM wajib dijalankan melalui Skenario A (memprediksi status_repeat_order) dan Skenario B mode *pre-rating* (memprediksi risiko *churn* tanpa rating_produk, status_repeat_order, dan skor_loyalitas_pelanggan), sehingga algoritma benar-benar mengidentifikasi faktor transaksi, subsidi, rute, kurir, dan metode pembayaran yang melatarbelakangi retensi konsumen.

3.3. Profil Deskriptif Perilaku Konsumen, Logistik Kepulauan, Kanal, dan Pembayaran

Analisis deskriptif dan tabulasi silang terhadap tiga puluh empat transaksi terverifikasi (n = 34) menyingkap dinamika perilaku konsumen UMKM *e-commerce* Indonesia di berbagai segmen demografi, produk, lokapasar, metode pembayaran, dan koridor logistik. Dari sisi demografis, pembeli perempuan mendominasi sebanyak 24 transaksi (70,59%) dibandingkan pembeli laki-laki sebanyak 10 transaksi (29,41%), dengan rentang usia pembeli membentang dari 18 tahun hingga 57 tahun. Distribusi tujuh kategori produk pada sampel terverifikasi menunjukkan bahwa Kuliner & Makanan Ringan Khas serta Fashion Muslim & Batik sama-sama menempati porsi terbesar dengan masing-masing 10 transaksi (29,41%), diikuti oleh Kriya & Kerajinan Tangan dan Peralatan Rumah Tangga Lokal masing-masing 4 transaksi (11,76%), Aksesoris & Souvenir Daerah sebanyak 3 transaksi (8,82%), Kopi & Minuman Olahan sebanyak 2 transaksi (5,88%), serta Kecantikan & Herbal Lokal sebanyak 1 transaksi (2,94%).

Prio Kustanto: *Penulis Korespondensi
Copyright © 2027, Mayadi, Prio Kustanto, Dismita Citra Dewi, Lenita Waty, Hendarman Lubis.

Journal of Information Technology, Software Engineering and Computer Science
Website : https://ejournal.techcart-press.com/index.php/itsecs
E-ISSN 2962-0635
P-ISSN 2962-0538
Volume 5, Number 1, January 2027
DOI : **https://doi.org/10.58602/itsecs.v5i1.498**
Page 135-158

Penelaahan struktur finansial pada 34 catatan tersebut menyoroti kesenjangan mencolok antara nilai belanja barang bruto dengan biaya pengiriman riil yang harus ditanggung oleh ekosistem niaga-el kepulauan. Agregat nilai_transaksi_bruto_idr mencapai Rp 10.323.000 dengan rerata Rp 303.618, namun median belanja bruto hanya berada di angka Rp 179.000 (rentang dari Rp 21.000 pada transaksi kuliner TRX_UMKM_2026_00014 hingga Rp 1.444.000 pada transaksi kriya TRX_UMKM_2026_00015). Kemencengan ke kanan ini mencerminkan adanya perbedaan segmen nilai transaksi antar-kategori: Kriya & Kerajinan Tangan mencatat rerata bruto tertinggi sebesar Rp 881.250 (pembelian ulang 50,00%; risiko tinggi/kritis 0,00%) dan Peralatan Rumah Tangga Lokal mencatat rerata Rp 512.750 (pembelian ulang 50,00%; risiko tinggi/kritis 25,00%), sedangkan Aksesoris & Souvenir Daerah hanya mencatat rerata Rp 106.000 (pembelian ulang 33,33%; risiko tinggi/kritis 33,33%) dan Kuliner & Makanan Ringan Khas mencatat rerata Rp 164.500 (pembelian ulang 20,00%; risiko tinggi/kritis 30,00%), sementara Fashion Muslim & Batik mencatat rerata Rp 223.100 (pembelian ulang 30,00%; risiko tinggi/kritis 10,00%), Kopi & Minuman Olahan rerata Rp 247.500 (pembelian ulang 50,00%; risiko tinggi/kritis 50,00%), serta Kecantikan & Herbal Lokal sebesar Rp 58.000 (1 transaksi berisiko tinggi).

Dari sisi insentif platform dan beban pengapalan, total diskon_voucher_platform_idr berjumlah Rp 432.000 dan dinikmati oleh tepat separuh sampel (17 dari 34 transaksi atau 50,00%), sedangkan total subsidi_gratis_ongkir_idr mencapai Rp 531.000 yang menjangkau 27 dari 34 transaksi (79,41%). Kendati cakupan subsidi ongkos kirim cukup merata, total ongkos_kirim_riil_idr yang dibebankan mencapai Rp 1.378.000 (rerata Rp 40.529 per pengiriman; rentang Rp 10.000 hingga Rp 120.000), sehingga subsidi gratis ongkos kirim hanya menutupi 38,53% dari total biaya pengiriman aktual. Akibatnya, total nilai_transaksi_netto_idr yang dibayar pembeli membengkak menjadi Rp 10.738.000 (rerata Rp 315.824)—atau 4,02% lebih tinggi di atas nilai produk bruto—karena sisa beban ongkos kirim yang tidak tersubsidi (Rp 847.000) melampaui total voucher diskon (Rp 432.000). Secara proporsional, tepat separuh transaksi (17/34 atau 50,00%) mencatat nilai bersih pembayaran melebihi harga bruto barang (nilai_transaksi_netto_idr > nilai_transaksi_bruto_idr), sebanyak 5 transaksi (14,71%) membukukan nilai bersih tepat sama dengan harga bruto karena insentif menutupi penuh ongkos kirim, dan hanya 12 transaksi (35,29%) yang benar-benar memperoleh tagihan bersih di bawah harga produk bruto. Tabel 5 menguraikan tabulasi silang tingkat pembelian ulang dan insiden risiko *churn* di berbagai kanal lokapasar, metode pembayaran, serta status ketepatan waktu pengiriman.

Tabel 5. Tabulasi Retensi Lintas Kanal, Pembayaran, dan Pengiriman

Dimensi Variabel dan Subkelompok	Jumlah (n)	Proporsi (%)	Keterangan Karakteristik Terpantau	Tingkat Pembelian Ulang (Ya)	Insiden Risiko Churn (Tinggi + Kritis)
Total Sampel Terverifikasi	34	100,00%	Bruto rerata Rp 303.618 (Median Rp 179.000); Netto rerata Rp 315.824	11 (32,35%)	8 (23,53%)
Kanal Lokapasar (platform_ecommerce)					
TikTok Shop Tokopedia	14	41,18%	Kanal dengan volume tertinggi; menampung 5 dari 8 kasus risiko tinggi/kritis	4 (28,57%)	5 (35,71%)

Prio Kustanto: *Penulis Korespondensi
Copyright © 2027, Mayadi, Prio Kustanto, Dismita Citra Dewi, Lenita Waty, Hendarman Lubis.

Journal of Information Technology, Software Engineering and Computer Science
Website : https://ejournal.techcart-press.com/index.php/itsecs
E-ISSN 2962-0635
P-ISSN 2962-0538
Volume 5, Number 1, January 2027
DOI : **https://doi.org/10.58602/itsecs.v5i1.498**
Page 135-158

Shopee Indonesia	8	23,53%	Melayani transaksi kriya bernilai tertinggi serta nol kasus risiko tinggi/kritis	4 (50,00%)	0 (0,00%)
Blibli	7	20,59%	Frekuensi tinggi pada pengiriman rute antarpulau serta Wilayah Tertinggal / 3T	1 (14,29%)	2 (28,57%)
WhatsApp Business & Instagram Shop	3	8,82%	Kanal perdagangan percakapan langsung; belum mencatat pembelian ulang	0 (0,00%)	1 (33,33%)
Lazada Indonesia	2	5,88%	Transaksi antarkota sepulau dengan ketepatan waktu penuh	2 (100,00%)	0 (0,00%)
Metode Pembayaran (metode_pembayaran)					
QRIS Bank / E-Wallet	13	38,24%	Instrumen pembayaran paling dominan lintas kanal dan rute luar Jawa	4 (30,77%)	3 (23,08%)
ShopeePay / GoPay / OVO / DANA	8	23,53%	Dompet digital terintegrasi ekosistem aplikasi lokapasar	3 (37,50%)	2 (25,00%)
Paylater / Cicilan Digital	7	20,59%	Mencatat tingkat pembelian ulang tertinggi di antara metode bervolume $n \geq 4$	3 (42,86%)	1 (14,29%)
Cash on Delivery / COD	4	11,76%	Tingkat retensi terendah disertai insiden risiko tinggi/kritis tertinggi	0 (0,00%)	2 (50,00%)
Virtual Account Bank	2	5,88%	Transaksi transfer terencana; bebas dari keterlambatan dan risiko tinggi	1 (50,00%)	0 (0,00%)
Status Keterlambatan Pengiriman					
Tepat Waktu	24	70,59%	Seluruh kasus pembelian ulang pada korpus teraudit terjadi di kelompok ini	11 (45,83%)	1 (4,17%)
Terlambat 1–2 Hari	7	20,59%	Tidak ada pembelian ulang meski sebagian masih memperoleh rating bintang 4	0 (0,00%)	4 (57,14%)
Terlambat Kritis >3 Hari	3	8,82%	Seluruh transaksi jatuh ke kategori	0 (0,00%)	3 (100,00%)

Prio Kustanto: *Penulis Korespondensi
Copyright © 2027, Mayadi, Prio Kustanto, Dismita Citra Dewi, Lenita Waty, Hendarman Lubis.

Journal of Information Technology, Software Engineering and Computer Science
Website : https://ejournal.techcart-press.com/index.php/itsecs
E-ISSN 2962-0635
P-ISSN 2962-0538
Volume 5, Number 1, January 2027
DOI : **https://doi.org/10.58602/itsecs.v5i1.498**
Page 135-158

					risiko Tinggi atau Kritis	

Tiga pola perilaku signifikan tampak menonjol dari Tabel 5. Pertama, pada himpunan 34 catatan transaksi yang terverifikasi, ketepatan waktu pengiriman berasosiasi sangat kuat dengan perilaku pembelian ulang: seluruh sebelas transaksi berstatus status_repeat_order = 'Ya' terjadi ketika paket tiba berstatus Tepat Waktu dengan peringkat produk 4 atau 5, sementara di antara sepuluh transaksi yang mengalami keterlambatan (Terlambat 1-2 Hari ataupun Terlambat Kritis >3 Hari), tidak satu pun menghasilkan pembelian ulang (0,00%) dan tujuh di antaranya (70,00%) tergolong berisiko *churn* Tinggi atau Kritis. Kedua, di antara metode pembayaran, Cash on Delivery / COD (n = 4) menampakkan kerentanan terbesar dengan 0,00% pembelian ulang dan 50,00% insiden risiko Tinggi/Kritis (meliputi TRX_UMKM_2026_00858 dan 00860), berbanding terbalik dengan pengguna Paylater / Cicilan Digital yang membukukan 42,86% pembelian ulang dan hanya 14,29% risiko tinggi. Pola kerentanan *COD* ini konsisten dengan temuan Samudra [7] di lokapasar Shopee bahwa transaksi *COD* memiliki korelasi kuat terhadap pembatalan maupun ketidakstabilan penyelesaian pesanan dibandingkan metode pembayaran di muka (*prepaid*). Ketiga, seluruh tujuh transaksi berharga produk bruto di bawah Rp 100.000 mengalami rasio ongkos kirim terhadap harga barang antara 37,93% hingga 119,05% serta mencatat indeks_sensitivitas_ongkir tinggi pada kisaran 0,83 hingga 0,98 (enam di antaranya menyentuh plafon 0,98, jauh di atas rerata keseluruhan sampel sebesar 0,680), yang menandakan tekanan biaya logistik yang sangat sensitif pada keranjang belanja mikro.

Selanjutnya, Tabel 6 merinci ketimpangan biaya dan durasi antar-koridor geografis serta variasi kinerja mitra ekspedisi. Data menunjukkan bahwa pengiriman ke Wilayah Tertinggal / 3T (n = 7) mencatat rerata ongkos_kirim_riil_idr sebesar Rp 78.571 dan rerata durasi 9,00 hari (rentang 5 hingga 12 hari)—yaitu 3,64 kali lipat lebih mahal dan 2,66 kali lipat lebih lama dibandingkan koridor Antar-Kota Satu Pulau (n = 13, rerata ongkir Rp 21.615, durasi 3,38 hari), serta 6,29 kali lipat lebih mahal dibandingkan rute Intra-Kota (rerata Rp 12.500). Ketimpangan tarif ekstrem terlihat misalnya pada transaksi TRX_UMKM_2026_00803 dari Maluku ke Nusa Tenggara Timur yang menanggung ongkos kirim Rp 120.000 tanpa subsidi atas produk aksesoris senilai Rp 156.000 (76,92% dari nilai produk), serta transaksi kuliner TRX_UMKM_2026_00014 dari Jawa Timur ke DKI Jakarta seharga Rp 21.000 yang dibebani tarif kurir Rp 25.000 (119,05% dari harga produk).

Tabel 6. Kinerja Logistik Lintas Koridor dan Ekspedisi

Dimensi Logistik	Jumlah (n)	Rerata Ongkos Kirim Riil (Rp)	Rerata Durasi Pengiriman (Hari)	Insiden Terlambat (>0 Hari)	Tingkat Pembelian Ulang (Ya)	Insiden Risiko Tinggi + Kritis
Koridor Wilayah (wilayah_pengiriman)						
Intra-Kota	2	12.500	4,00	2 (100,00%)	0 (0,00%)	2 (100,00%)
Antar-Kota Satu Pulau	13	21.615	3,38	4 (30,77%)	5 (38,46%)	3 (23,08%)
Antar-Pulau Jawa–Luar Jawa	12	43.500	6,25	3 (25,00%)	3 (25,00%)	2 (16,67%)
Wilayah Tertinggal / 3T	7	78.571	9,00	1 (14,29%)	3 (42,86%)	1 (14,29%)
Mitra Layanan Ekspedisi (layanan_ekspedisi)						

Prio Kustanto: *Penulis Korespondensi
Copyright © 2027, Mayadi, Prio Kustanto, Dismita Citra Dewi, Lenita Waty, Hendarman Lubis.

Journal of Information Technology, Software Engineering and Computer Science
Website : https://ejournal.techcart-press.com/index.php/itsecs
E-ISSN 2962-0635
P-ISSN 2962-0538
Volume 5, Number 1, January 2027
DOI : **https://doi.org/10.58602/itsecs.v5i1.498**
Page 135-158

Shopee Xpress	14	51.286	6,50	3 (21,43%)	4 (28,57%)	2 (14,29%)
J&T Express	7	37.286	6,00	2 (28,57%)	1 (14,29%)	1 (14,29%)
JNE	4	27.000	3,50	2 (50,00%)	2 (50,00%)	2 (50,00%)
Pos Indonesia	4	29.000	4,25	1 (25,00%)	2 (50,00%)	0 (0,00%)
SiCepat	3	28.333	3,67	0 (0,00%)	2 (66,67%)	1 (33,33%)
Anteraja	1	15.000	6,00	1 (100,00%)	0 (0,00%)	1 (100,00%)
Paxel	1	75.000	9,00	1 (100,00%)	0 (0,00%)	1 (100,00%)

Temuan menarik yang tampak berlawanan dengan intuisi muncul pada koridor Wilayah Tertinggal / 3T di Tabel 6: meskipun menempuh jarak paling jauh dengan durasi terlama dan biaya tertinggi, koridor 3T dalam sampel terverifikasi membukukan tingkat pembelian ulang tertinggi sebesar 42,86% (3/7) serta insiden keterlambatan terendah sebesar 14,29% (hanya terjadi pada TRX_UMKM_2026_00987). Fenomena ini dapat dijelaskan melalui keselarasan antara ekspektasi waktu tiba yang realistis dan dukungan kompensasi subsidi ongkir: pengiriman 5 hingga 12 hari menuju Maluku Utara, Kepulauan Riau, atau Papua tetap berstatus Tepat Waktu sesuai estimasi koridor, dan ketika pelaku usaha atau platform memberikan subsidi ongkos kirim serta potongan voucher—seperti pada TRX_UMKM_2026_00500 menuju Maluku Utara (subsidi ongkir penuh Rp 55.000 plus voucher Rp 50.000) dan TRX_UMKM_2026_01069 menuju Papua (rasio efektivitas diskon 25,25%)—konsumen merespons dengan penilaian tinggi dan pembelian ulang. Sebaliknya, pada rute jarak dekat Intra-Kota (n = 2), keterlambatan kritis 6 hari pada transaksi TRX_UMKM_2026_00858 di Jawa Tengah serta keterlambatan pada TRX_UMKM_2026_00860 di Jawa Timur langsung memicu penilaian rendah (bintang 2 dan 3) dan risiko *churn* Kritis/Tinggi karena pembeli di koridor yang sama memiliki ekspektasi ketepatan pengiriman jauh lebih ketat. Dari sisi mitra ekspedisi, meskipun Shopee Xpress mencatat rerata tarif tertinggi (Rp 51.286) dan durasi 6,50 hari karena paling banyak menangani pengiriman jarak jauh dan 3T, rasio keterlambatannya termasuk rendah (21,43%) di bawah J&T Express (28,57%) dan JNE (50,00%), sedangkan Pos Indonesia tidak mencatat satu pun kasus risiko tinggi/kritis (0,00%) serta SiCepat membukukan ketepatan waktu sempurna (0,00% terlambat)—meski angka per ekspedisi dengan sampel kecil ($n \leq 4$) wajib ditafsirkan secara hati-hati sebagai indikator deskriptif awal.

3.4. Strategi Retensi Pelanggan Berbasis Bukti untuk UMKM E-Commerce Indonesia

Integrasi antara temuan empiris pada catatan transaksi terverifikasi (n = 34), audit ketimpangan logistik antarpulau, serta kerangka literatur pemasaran digital UMKM Indonesia memungkinkan perumusan strategi retensi pelanggan yang terarah secara ekonomi dan dapat dieksekusi langsung di lapangan. Prinsip sentral strategi ini adalah menghentikan praktik pemberian diskon harga merata (*blanket discounting*) yang menggerus margin tipis pelaku UMKM, dan menggantinya dengan alokasi subsidi ongkos kirim, perlindungan *Service Level Agreement* (SLA) ekspedisi, serta pemulihan layanan yang dikalibrasi menurut profil keranjang belanja, koridor geografis, metode pembayaran, dan tingkat risiko *churn*.

Strategi pertama berfokus pada rebalansifikasi antara subsidi ongkos kirim bersyarat dan voucher loyalitas berdasarkan nilai keranjang belanja. Audit finansial menyingkap bahwa 50,00% transaksi tidak menerima potongan voucher platform, sementara pada

Prio Kustanto: *Penulis Korespondensi
Copyright © 2027, Mayadi, Prio Kustanto, Dismita Citra Dewi, Lenita Waty, Hendarman Lubis.

Journal of Information Technology, Software Engineering and Computer Science
Website : https://ejournal.techcart-press.com/index.php/itsecs
E-ISSN 2962-0635
P-ISSN 2962-0538
Volume 5, Number 1, January 2027
DOI : **https://doi.org/10.58602/itsecs.v5i1.498**
Page 135-158

seluruh tujuh transaksi berharga di bawah Rp 100.000, tarif ongkos kirim riil melonjak hingga 37,93% sampai 119,05% dari harga barang sehingga mendorong indeks sensitivitas ongkos kirim menuju kisaran 0,83–0,98. Studi empiris perilaku belanja daring di Indonesia menunjukkan bahwa program gratis ongkos kirim berpengaruh signifikan meningkatkan niat beli ulang melalui kepuasan pelanggan, sedangkan diskon harga langsung tidak selalu berdampak signifikan secara langsung terhadap niat beli ulang. Oleh karena itu, bagi produk UMKM bernilai di bawah median sampel (Rp 179.000)—yang didominasi kategori kuliner khas (rerata Rp 164.500), aksesoris daerah (rerata Rp 106.000), serta kecantikan herbal—pelaku UMKM direkomendasikan untuk mengalihkan anggaran promosi dari potongan persentase harga ke subsidi ongkos kirim bersyarat dengan nilai minimum keranjang atau merancang paket *bundling* berisi dua hingga tiga unit produk guna mendorong total belanja melampaui kisaran Rp 150.000 sebagai target operasional agar rasio tarif kurir terhadap harga barang tertekan di bawah 20%. Sebaliknya, pada kategori bernilai tinggi di atas Rp 500.000—seperti Kriya & Kerajinan Tangan (rerata Rp 881.250), Peralatan Rumah Tangga Lokal (rerata Rp 512.750), dan segmen premium Fashion Muslim & Batik (TRX_UMKM_2026_00342), di mana indeks sensitivitas ongkir pada rute non-3T berada di tingkat rendah antara 0,24 hingga 0,43—retensi lebih tepat diarahkan pada voucher loyalitas bertingkat untuk pembelian berikutnya dan optimalisasi opsi pembayaran *Paylater / Cicilan Digital* yang pada sampel teraudit membukukan tingkat pembelian ulang tertinggi (42,86% untuk kategori n ≥ 4).

Strategi kedua menyasar manajemen ekspedisi berbasis koridor geografis dan komunikasi proaktif ekspektasi waktu tiba. Kenyataan pada himpunan teraudit bahwa 100% dari sebelas kasus pembelian ulang hanya terjadi pada kondisi Tepat Waktu, sedangkan seluruh sepuluh pesanan terlambat mencatat 0,00% pembelian ulang, menegaskan bahwa kepastian pemenuhan janji pengiriman lebih menentukan retensi dibandingkan kecepatan absolut semata. Hal ini diperkuat oleh kontras kinerja antara koridor Wilayah Tertinggal / 3T—yang mampu mencatat pembelian ulang 42,86% walaupun rata-rata pengiriman memakan waktu 9,00 hari karena tetap berstatus Tepat Waktu—melawan koridor Intra-Kota di mana keterlambatan langsung berujung pada penilaian bintang 2 hingga 3 dan risiko *churn* Tinggi/Kritis. Secara operasional, UMKM perlu memetakan tingkat ketepatan waktu masing-masing mitra logistik per koridor tujuan (mengacu pada variasi insiden keterlambatan antar-kurir pada Tabel 6), memprioritaskan kurir dengan rekam jejak tepat waktu terbaik di rute Jawa–luar Jawa dan 3T, menetapkan estimasi waktu tiba (*Estimated Time of Arrival*) yang realistis sejak awal, serta mengaktifkan pesan proaktif pembaruan status pengiriman apabila resi menunjukkan potensi deviasi jadwal sebelum konsumen memberikan ulasan buruk. Selain itu, investasi pada kemasan pelindung berlapis untuk produk kuliner rapuh dan kerajinan tangan pada pengiriman lintas pulau menjaga peringkat produk tetap berada di rentang bintang 4–5.

Strategi ketiga adalah mitigasi risiko pada segmen pembayaran tunai di tempat (*COD*) dan migrasi transaksi perdagangan percakapan ke kanal bersubsidi. Dalam data teraudit, metode Cash on Delivery / COD (n = 4) mencatat 0,00% pembelian ulang serta 50,00% risiko *churn* tinggi atau kritis, sejalan dengan literatur yang mengaitkan *COD* dengan ketidakstabilan penyelesaian transaksi di lokapasar Indonesia. Di samping itu, kanal WhatsApp Business & Instagram Shop (n = 3) juga belum mencatat satu pun pembelian ulang (0,00%) karena pembeli tidak menikmati subsidi ongkos kirim otomatis dari penyedia lokapasar. Pelaku UMKM perlu menerapkan verifikasi konfirmasi pesanan sebelum pengiriman (*pre-dispatch confirmation*) khusus untuk pesanan *COD* guna memastikan kesiapan penerima dan kesesuaian alamat, sekaligus menawarkan insentif peralihan metode pembayaran—seperti prioritas pemrosesan paket di hari yang sama atau bonus potongan mikro—bagi konsumen *COD* agar berpindah ke QRIS Bank / E-Wallet atau dompet elektronik pada transaksi kedua. Untuk pelanggan akuisisi awal dari Instagram

Prio Kustanto: *Penulis Korespondensi
Copyright © 2027, Mayadi, Prio Kustanto, Dismita Citra Dewi, Lenita Waty, Hendarman Lubis.

Journal of Information Technology, Software Engineering and Computer Science
Website : https://ejournal.techcart-press.com/index.php/itsecs
E-ISSN 2962-0635
P-ISSN 2962-0538
Volume 5, Number 1, January 2027
DOI : **https://doi.org/10.58602/itsecs.v5i1.498**
Page 135-158

atau WhatsApp, pelaku usaha dapat menyediakan tautan *checkout* etalase resmi di Shopee, Tokopedia, TikTok Shop, atau Blibli guna mengoptimalkan pemanfaatan subsidi gratis ongkos kirim platform tanpa mengorbankan arus kas internal UMKM.

Strategi keempat mengorganisasikan seluruh hasil deteksi prediktif dan atribusi *TreeSHAP* ke dalam matriks intervensi empat tingkat risiko churn sebagai panduan aksi harian bagi pemilik usaha, sebagaimana dirinci pada Tabel 8. Parameter waktu respons dan besaran kompensasi pada matriks tersebut dirancang sebagai protokol operasional yang terinspirasi oleh struktur insentif teramati pada korpus (di mana rasio_efektivitas_diskon pada kasus retensi berhasil mencapai kisaran 14–32%) serta kerangka minimalisasi biaya pembatalan berbasis *SHAP*.

Tabel 7. Matriks Intervensi Retensi Empat Tingkat Risiko

Tingkat Risiko *Churn* (Rentang Skor Loyalitas)	Profil Empiris pada Sampel Terverifikasi (n = 34)	Faktor Pendorong Utama Teridentifikasi	Rekomendasi Protokol Intervensi Retensi UMKM	Rekomendasi Kanal dan Jendela Waktu Aksi
Rendah (0,70–0,98)	17 transaksi (50,00%); seluruhnya Tepat Waktu dengan rating 4–5; mencakup seluruh 11 kasus pembelian ulang aktif [1]	Kepuasan produk tinggi dan pengiriman tepat waktu; namun 6 dari 17 transaksi bernilai skor 0,70 belum melakukan pembelian ulang [1]	Fokus mengonversi pembeli bintang 5 non-ulang (skor = 0,70) menjadi pelanggan loyal melalui kartu ucapan terima kasih personal, sampel produk ukuran mini, keanggotaan pelanggan prioritas, serta voucher pembelian ulang berbatas waktu (misalnya masa berlaku 14 hari) [2]	Kartu sisipan fisik di dalam kemasan dan pesan tindak lanjut otomatis H+3 pascapenerimaan via obrolan lokapasar
Sedang (0,48–0,57)	9 transaksi (26,47%); didominasi rating 4 tanpa pembelian ulang; 3 terlambat 1–2 hari dan 6 tepat waktu [1]	Ketidakpuasan ringan terhadap kemasan/produk, tarif ongkos kirim tinggi tanpa subsidi penuh, atau keterlambatan 1–2 hari [3]	Alokasikan subsidi gratis ongkos kirim terarah untuk pesanan kedua (diprioritaskan bagi transaksi berindeks sensitivitas ongkir ≥ 0,70 atau koridor antarpulau/3T), disertai jajak pendapat singkat perbaikan kualitas produk agar skor bergerak ke bintang 5 [1]	Kampanye pengingat bervoucher ongkos kirim pada jendela retensi hari ke-10 hingga ke-21 pascatransaksi
Tinggi (0,22–0,45)	6 transaksi (17,65%); seluruhnya berating 2–3 tanpa pembelian ulang; 5 dari 6 mengalami	Keterlambatan logistik 1–2 hari atau kritis yang berinteraksi dengan keluhan kualitas produk atau biaya pengiriman [4]	Aktifkan protokol pemulihan layanan (*service recovery*): permohonan maaf personal dari pengelola toko, penggantian unit rusak/cacat tanpa	Kontak proaktif langsung oleh pengelola toko dalam kurun kurang dari 24 jam sejak status Terlambat atau

Prio Kustanto: *Penulis Korespondensi
Copyright © 2027, Mayadi, Prio Kustanto, Dismita Citra Dewi, Lenita Waty, Hendarman Lubis.

Journal of Information Technology, Software Engineering and Computer Science
Website : https://ejournal.techcart-press.com/index.php/itsecs
E-ISSN 2962-0635
P-ISSN 2962-0538
Volume 5, Number 1, January 2027
DOI : **https://doi.org/10.58602/itsecs.v5i1.498**
Page 135-158

		keterlambatan pengiriman [1]		prosedur retur rumit, serta penawaran kupon pemulihan bernilai signifikan (sekitar 15–25%, setara rasio insentif efektif yang terpantau pada segmen sensitif) disertai pilihan kurir paling handal di rute tersebut [1]	rating ≤ 3 terdeteksi
Kritis	(0,05–0,18)	2 transaksi (5,88%): TRX_UMKM_2026_00671 (antarpulau, 10 hari, rating 1) dan 00858 (intra-kota *COD*, 6 hari, rating 2) [1]	Keterlambatan kritis >3 hari yang merusak kepuasan secara ekstrem, diperparah kegagalan produk atau friksi *COD* [1]	Jalankan eskalasi penyelamatan krisis dan audit operasional: evaluasi penonaktifan sementara mitra ekspedisi pada rute gagal terkait, opsi pengembalian dana parsial atau pengiriman ulang prioritas bebas biaya, serta investigasi akar masalah pengemasan/stok sebelum mengirim penawaran pemulihan kepercayaan	Eskalasi manajerial dalam kurang dari 6 jam pascakejadian terlambat kritis dan evaluasi kinerja kurir mingguan

Strategi kelima berkaitan dengan tata kelola data dan audit kebersihan umpan balik pelanggan sebelum otomasi analitik. Temuan inkonsistensi lintas-modal pada Tabel 3—di mana teks ulasan membahas produk yang berbeda dari kategori terstruktur atau narasi penalaran menyisipkan klausa retensi yang bertentangan dengan fakta kolom status_repeat_order—memberikan peringatan penting bagi penyedia platform maupun konsultan digitalisasi UMKM. Sebelum mengintegrasikan model analitik sentimen berbasis teks atau dasbor prediksi *churn* ke dalam operasional toko, pelaku ekosistem wajib memastikan tautan integritas relasional antara identifikasi SKU produk dengan teks ulasan pelanggan serta menghindari penggunaan variabel komposit pasca-kejadian sebagai fitur peringatan dini. Hanya dengan tata kelola data yang bersih dari kebocoran dan tervalidasi lintas-modal, keluaran model *Random Forest*, *XGBoost*, dan *TreeSHAP* dapat diandalkan untuk memandu keputusan subsidi dan retensi UMKM secara berkelanjutan.

4. KESIMPULAN

Penelitian ini telah membangun rancangan metodologis lengkap serta menyajikan hasil audit kuantitatif yang komprehensif untuk memprediksi perilaku konsumen dan merumuskan strategi retensi pelanggan bagi UMKM *e-commerce* Indonesia dengan memadukan algoritma *Random Forest*, *XGBoost*, dan interpretabilitas *TreeSHAP*. Secara jujur dan transparan, penelitian ini menegaskan bahwa pelatihan empiris kedua algoritma, komputasi lima metrik evaluasi klasifikasi (*accuracy*, *precision*, *recall*, *F1-score*, dan *ROC-AUC*), serta penghitungan koefisien atribusi *TreeSHAP* pada keseluruhan populasi dataset

Prio Kustanto: *Penulis Korespondensi
Copyright © 2027, Mayadi, Prio Kustanto, Dismita Citra Dewi, Lenita Waty, Hendarman Lubis.

Journal of Information Technology, Software Engineering and Computer Science
Website : https://ejournal.techcart-press.com/index.php/itsecs
E-ISSN 2962-0635
P-ISSN 2962-0538
Volume 5, Number 1, January 2027
DOI : **https://doi.org/10.58602/itsecs.v5i1.498**
Page 135-158

belum dapat dieksekusi karena ketidaktersediaan himpunan data penuh pada lingkungan penyimpanan dan belum terpasangnya modul komputasi khusus algoritma *boosting* serta komputasi nilai Shapley. Alih-alih merekayasa metrik kinerja yang tidak memiliki pijakan komputasi, artikel ini membuktikan kelayakan arsitektur pemodelannya melalui audit forensik terhadap tiga puluh empat catatan transaksi yang utuh terverifikasi (n = 34) dan menyintesis proyeksi model dari tolok ukur literatur jurnal yang relevan.

Empat kesimpulan pokok dapat ditarik dari hasil audit empiris dan perancangan strategi dalam studi ini. Pertama, verifikasi matematis terhadap 34 catatan transaksi membuktikan konsistensi aritmetika mutlak tanpa deviasi pada kalkulasi nilai_transaksi_netto_idr serta galat pembulatan minimal $4{,}68 \times 10^{-5}$ pada rasio_efektivitas_diskon, namun audit struktural menemukan sejumlah keterputusan baris dan anomali lintas-modal antara kategori produk terstruktur dengan isi teks ulasan maupun kontradiksi templat pada narasi penalaran yang harus diisolasi sebelum pelatihan. Kedua, pengujian kelompok mengungkap struktur deterministik tanpa varians antara kombinasi rating_produk, status_keterlambatan_pengiriman, dan status_repeat_order terhadap skor_loyalitas_pelanggan dan kategori_risiko_churn, yang menegaskan bahwa pemodelan prediktif bebas kebocoran target wajib dijalankan pada target pembelian ulang atau dalam mode *pre-rating*. Ketiga, analisis empiris menyingkap ketimpangan tarif logistik kepulauan yang tajam—di mana rerata ongkos kirim menuju Wilayah Tertinggal / 3T mencapai Rp 78.571 dibandingkan Rp 21.615 pada rute Antar-Kota Satu Pulau—serta menunjukkan bahwa keterlambatan pengiriman dan metode pembayaran *COD* berasosiasi dengan tingkat pembelian ulang nol persen pada sampel teraudit, sementara pengiriman tepat waktu di rute 3T yang ditopang subsidi ongkos kirim terarah tetap mencatat pembelian ulang sebesar 42,86%. Keempat, penyatuan temuan audit dengan mekanisme penyeimbangan kelas *SMOTE* dan eksplanasi lokal-ke-global *TreeSHAP* menghasilkan strategi retensi UMKM empat tingkat (Rendah, Sedang, Tinggi, Kritis) yang berfokus pada rebalansifikasi anggaran diskon ke subsidi ongkos kirim bersyarat untuk produk berharga di bawah median, pengendalian SLA ekspedisi per koridor geografis, migrasi bertahap pembeli *COD* menuju pembayaran digital, serta pemulihan layanan proaktif dalam jendela 6–24 jam.

Sebagai arah penelitian selanjutnya, kerangka eksperimen bebas kebocoran target yang dirumuskan dalam studi ini perlu dieksekusi penuh pada infrastruktur komputasi yang memiliki akses terhadap keseluruhan baris korpus serta dilengkapi perangkat lunak pelatihan *Random Forest*, *XGBoost*, *SMOTE-NC*, dan *TreeSHAP*. Pengujian skala penuh tersebut hendaknya menyajikan kurva *ROC* dan *Precision-Recall*, mengaudit dampak teknik *resampling* terhadap kalibrasi probabilitas di luar lipatan validasi, serta mengeksplorasi nilai interaksi berpasangan *TreeSHAP* untuk menguji ambang batas rasio ongkos kirim optimal pada setiap kategori produk unggulan UMKM di Indonesia.

PERNYATAAN KONTRIBUSI PENULIS (CREDIT)

Mayadi: Penulisan – draf awal, Supervisi. **Prio Kustanto**: Penulisan – draf awal, Penulisan – tinjauan dan penyuntingan. **Dismita Citra Dewi**: Penulisan – draf awal, Kurasi Data. **Lenita Waty**: Penulisan – draf awal, Administrasi proyek. **Hendarman Lubis**: Penulisan – draf awal, Analisis.

PERNYATAAN KONFLIK KEPENTINGAN

Para penulis menyatakan bahwa tidak terdapat kepentingan finansial atau hubungan pribadi yang diketahui yang dapat memengaruhi hasil penelitian yang dilaporkan dalam artikel ini.

Prio Kustanto: *Penulis Korespondensi
Copyright © 2027, Mayadi, Prio Kustanto, Dismita Citra Dewi, Lenita Waty, Hendarman Lubis.

Journal of Information Technology, Software Engineering and Computer Science
Website : https://ejournal.techcart-press.com/index.php/itsecs
E-ISSN 2962-0635
P-ISSN 2962-0538
Volume 5, Number 1, January 2027
DOI : **https://doi.org/10.58602/itsecs.v5i1.498**
Page 135-158

PERNYATAAN PENDANAAN

Penulis tidak menerima dukungan finansial untuk penelitian, kepenulisan, dan/atau publikasi artikel ini.

PERNYATAAN PENGGUNAAN AI GENERATIF DAN TEKNOLOGI BERBANTUAN AI DALAM PROSES PENULISAN

Para penulis menggunakan Ollama Desktop dengan model Qwen 3.8 Flash next untuk meningkatkan kejelasan penulisan artikel ini. Seluruh konten yang dibantu oleh AI telah ditinjau dan disunting oleh penulis, dan para penulis bertanggung jawab sepenuhnya atas isi akhir yang dipublikasikan.

REFERENCES

[1] E. Aminullah *et al.*, "Interactive Components of Digital MSMEs Ecosystem for Inclusive Digital Economy in Indonesia," *J. Knowl. Econ.*, vol. 15, no. 1, pp. 487–517, Mar. 2024, doi: 10.1007/s13132-022-01086-8.

[2] A. Zhafira and F. Yunita, "Analisis Prediksi Pembatalan Pesanan E-Commerce Menggunakan Algoritma Random Forest dan XGBoost untuk Mitigasi Kerugian Finansial," *Indones. J. Multidiscip. Soc. Technol.*, vol. 4, no. 3, pp. 3143–3152, Aug. 2026, doi: 10.69693/ijmst.v4i3.12964.

[3] T. Chen and C. Guestrin, "XGBoost," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA: ACM, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[4] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.

[5] A. Dhini and M. Fauzan, "Predicting Customer Churn using ensemble learning: Case Study of a Fixed Broadband Company," *Int. J. Technol.*, vol. 12, no. 5, p. 1030, Dec. 2021, doi: 10.14716/ijtech.v12i5.5223.

[6] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, in NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, pp. 4768–4777.

[7] M. B. Samudra, M. Afrina, A. Meiriza, R. D. Kurnia, and A. Ariani, "Analisis Perbandingan Random Forest dan XGBoost dalam Prediksi Pembatalan Pesanan Shopee Menggunakan Interpretasi SHAP Comparison of Random Forest and XGBoost for Shopee Order Cancellation Prediction Using SHAP Interpretation," vol. 8, no. 2, pp. 820–832, 2026, doi: 10.47065/bits.v8i2.10960.

[8] A. Fannouch, J. Gharib, and Y. Gahi, "Enhancing DataOps practices through innovative collaborative models: A systematic review," *Int. J. Inf. Manag. Data Insights*, vol. 5, no. 1, p. 100321, Jun. 2025, doi: 10.1016/j.jjimei.2025.100321.

[9] L. N. Mishra, A. Rangari, S. Nagrare, and S. K. Nayak, "A taxonomy for detecting and preventing temporal data leakage in machine learning-based build prediction: A dual-platform empirical validation," *PLoS One*, vol. 21, no. 5, p. e0340167, May 2026, doi: 10.1371/journal.pone.0340167.

[10] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, vol. 4, no. 9, p. 100804, Sep. 2023, doi: 10.1016/j.patter.2023.100804.

[11] J. M. H. Pinheiro *et al.*, "The Impact of Feature Scaling in Machine Learning: Effects on Regression and Classification Tasks," *IEEE Access*, vol. 13, pp. 199903–199931,

Prio Kustanto: *Penulis Korespondensi
Copyright © 2027, Mayadi, Prio Kustanto, Dismita Citra Dewi, Lenita Waty, Hendarman Lubis.

Journal of Information Technology, Software Engineering and Computer Science
Website : https://ejournal.techcart-press.com/index.php/itsecs
E-ISSN 2962-0635
P-ISSN 2962-0538
Volume 5, Number 1, January 2027
DOI : **https://doi.org/10.58602/itsecs.v5i1.498**
Page 135-158

2025, doi: 10.1109/ACCESS.2025.3635541.
[12] W. Ruan and Z. Zhai, "Predicting User Purchase Behavior via Multidimensional Feature Engineering and Explainable Machine Learning," *J. Theor. Appl. Electron. Commer. Res.*, vol. 21, no. 10, p. 352, Oct. 2026, doi: 10.3390/jtaer21100352.
[13] Y. Yang, X. Li, Y. Guan, H. Wang, C. Kong, and J. Jiang, "LHP: Logical hypergraph link prediction," *Expert Syst. Appl.*, vol. 222, p. 119842, Jul. 2023, doi: 10.1016/j.eswa.2023.119842.
[14] N. V. Chawla, A. Lazarevic, L. O. Hall, and K. W. Bowyer, "SMOTEBoost: Improving Prediction of the Minority Class in Boosting," 2003, pp. 107–119. doi: 10.1007/978-3-540-39804-2_12.
[15] M. K. Mayangsari, I. Syarif, and A. Barakbah, "Evaluation of Stratified K-Fold Cross Validation for Predicting Bug Severity in Game Review Classification," *Kinet. Game Technol. Inf. Syst. Comput. Network, Comput. Electron. Control*, Jul. 2023, doi: 10.22219/kinetik.v8i3.1740.
[16] K. M. Sujon, R. Hassan, K. Choi, and M. A. Samad, "Accuracy, precision, recall, f1-score, or MCC? empirical evidence from advanced statistics, ML, and XAI for evaluating business predictive models," *J. Big Data*, vol. 12, no. 1, p. 268, Dec. 2025, doi: 10.1186/s40537-025-01313-4.
[17] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006, doi: 10.1016/j.patrec.2005.10.010.
[18] S. M. Lundberg *et al.*, "From local explanations to global understanding with explainable AI for trees," *Nat. Mach. Intell.*, vol. 2, no. 1, pp. 56–67, Jan. 2020, doi: 10.1038/s42256-019-0138-9.

Prio Kustanto: *Penulis Korespondensi
Copyright © 2027, Mayadi, Prio Kustanto, Dismita Citra Dewi, Lenita Waty, Hendarman Lubis.