Journal of Science and Social Research ISSN 2615 – 4307 (Print)
Oct 2026, IX (5): 9355 – 9367 ISSN 2615 – 3262 (Online)
Available online at http://jurnal.goretanpena.com/index.php/JSSR

PREDIKSI RISIKO PENGEMBALIAN BARANG (RETUR) PADA PENJUALAN PAKAIAN ONLINE MENGGUNAKAN ALGORITMA RANDOM FOREST

Mohammad Faiz[1], Muhammad Trinanda[2], Hendra Budi Hartono Saragih[3], Muhammad Syahputra Novelan[4]
Universitas Pembangunan Panca Budi, Medan, Indonesia
e-mail: [1]*faizgalv@gmail.com, [2]m.3rd.nanda@gmail.com, [3]realmanihuruk@gmail.com, [4]putranovelan@dosen.pancabudi.ac.id
Corresponding Author: Mohammad Faiz

Abstract: *Product returns impose reverse-logistics cost, re-inspection effort, and margin loss on online apparel sales, so the ability to predict return risk before an order is shipped carries operational value. This study builds and evaluates a Random Forest classifier for return-risk prediction on 400 online apparel transactions from 1 May to 26 October 2026 containing 80 returns (20,0%). Eleven pre-purchase predictors—four categorical and seven numeric—are used with no post-purchase information so that the model can be applied at order time. Random Forest with 300 trees and class weighting is compared against logistic regression, Naive Bayes, a decision tree, Extra Trees, and a majority baseline through ten-times repeated stratified 5-fold cross-validation, a replicated 80:20 hold-out, a 500-replicate permutation test, 5,000-replicate bootstrap resampling, a threshold sweep, and cost-sensitive error analysis. Random Forest attains ROC-AUC 0,815 ± 0,011 and PR-AUC 0,522—about 2,6 times the base rate—with 82,5% accuracy. On held-out data the AUC reaches 0,854 (95% CI 0,735–0,947) and differs significantly from chance (p = 0,002). A threshold of 0.49 yields 0,750 precision and 56,2% recall, while cost analysis shows savings from 37,5% at a 1:1 cost ratio to 86,9% at 20:1. Description completeness, product price, delivery estimate, and material clarity are the most influential predictors under permutation importance. Random Forest is therefore viable as an early warning system for return risk at boutique scale, provided that subsequent evaluation uses real transaction data.*

Keywords: *online apparel returns; random forest; risk classification; permutation importance; cost analysis*

Abstrak: Pengembalian barang (retur) menambah biaya logistik balik, pemeriksaan ulang, dan kehilangan margin pada penjualan pakaian daring, sehingga kemampuan memprediksi risiko retur sebelum pesanan dikirim memiliki nilai operasional. Penelitian ini membangun dan mengevaluasi model klasifikasi Random Forest untuk memprediksi risiko retur pada 400 transaksi pakaian online periode 1 Mei–26 Oktober 2026 yang memuat 80 retur (20,0%). Sebelas prediktor pra-pembelian—empat kategorikal dan tujuh numerik—dipakai tanpa informasi pascapembelian agar model dapat diterapkan pada saat pemesanan. Random Forest dengan 300 pohon dan pembobotan kelas dibandingkan dengan Regresi Logistik, Naive Bayes, Decision Tree, Extra Trees, dan baseline mayoritas melalui validasi silang stratifikasi 5-lipat yang diulang 10 kali, replikasi split 80:20, uji permutasi 500 ulangan, bootstrap 5.000 ulangan, penelusuran ambang, dan analisis biaya kesalahan. Random Forest menghasilkan ROC-AUC 0,815 ± 0,011 dan PR-AUC 0,522—setara 2,6 kali lipat prevalensi—dengan akurasi 82,5%. Pada data uji AUC mencapai 0,854 (selang kepercayaan 95% 0,735–0,947) dan berbeda nyata dari peluang (p = 0,002). Ambang 0,49 memberi presisi 0,750 dan recall 56,2%; analisis biaya menunjukkan penghematan 37,5% pada rasio biaya 1:1 hingga 86,9% pada rasio 20:1. Kelengkapan deskripsi, harga produk, estimasi pengiriman, dan kejelasan informasi bahan merupakan prediktor paling berpengaruh menurut permutation importance. Hasil

ini menunjukkan bahwa Random Forest layak dipakai sebagai sistem peringatan dini risiko retur pada skala butik, dengan catatan bahwa evaluasi selanjutnya perlu dilakukan pada data transaksi riil.

Kata kunci: retur pakaian online; random forest; klasifikasi risiko; permutation importance; analisis biaya

PENDAHULUAN

Perdagangan pakaian secara daring tumbuh cepat, tetapi pertumbuhan itu membawa beban yang jarang tampak pada laporan penjualan: barang yang kembali. Retur bukan sekadar pembatalan transaksi. Setiap retur memicu biaya pengiriman balik, pemeriksaan kondisi barang, penataan ulang stok, potensi penurunan harga jual, dan beban layanan pelanggan. Stevenson dan Rieck (2024) menunjukkan bahwa praktik manajemen retur berbeda tajam antar sektor dan antar negara, dan bahwa perusahaan yang mengelola retur secara reaktif menanggung biaya yang jauh lebih besar daripada perusahaan yang mengantisipasinya. Mishra dan Dutta (2024) menegaskan bahwa retur dapat diprediksi: variabel yang memengaruhinya dapat diidentifikasi sebelum transaksi terjadi, sehingga pencegahan menjadi mungkin secara teknis.

Penyebab retur pakaian bersifat jamak. Ketidaksesuaian ukuran dan bentuk tubuh menempati posisi dominan; tinjauan sistematis Wang dan Zakaria (2024) menemukan bahwa masalah ukuran dan fit tidak hanya merugikan konsumen, tetapi juga menekan keberlanjutan industri apparel karena barang yang dikembalikan kerap tidak dapat dijual kembali pada harga penuh. Ketidakpastian fit diperkuat oleh kualitas informasi produk: pada 302 responden dewasa muda Indonesia, Nugroho dan Wang (2023) menunjukkan bahwa kredibilitas ulasan, ketidakpastian kesesuaian produk, dan kepuasan pelanggan merupakan penentu utama niat retur. Pada kanal yang lebih baru, Linh dan Park (2025) menemukan bahwa strategi penjualan transaksional yang menekankan harga dan promosi menaikkan kecenderungan beli impulsif yang berujung pada disonansi produk dan retur, sedangkan strategi relasional yang menyajikan informasi produk secara kaya menurunkan risiko yang dipersepsikan. Lo dkk. (2026) melengkapi gambaran tersebut dengan menunjukkan bahwa rekomendasi yang dipersonalisasi ikut membentuk perilaku konsumen pascapembelian. Di sisi lain, sebagian retur bersifat oportunistik; Merlano dkk. (2024) memetakan perspektif konsumen terhadap berbagai intervensi penanggulangan retur fiktif pada ritel fesyen omnichannel.

Karena penyebabnya banyak dan saling terkait, pendekatan berbasis pembelajaran mesin menjadi pilihan alami. Zhang dkk. (2025) memprediksi retur pada niaga livestream dengan transformer multimodal lintas modal, sementara Chen dkk. (2026) menekankan pentingnya model yang dapat dijelaskan agar prediksi retur dapat diterjemahkan menjadi tindakan operasional pada ritel apparel omnichannel. Di Indonesia, Anggraini dkk. (2025) membandingkan Random Forest dan XGBoost dan menegaskan bahwa kedua algoritma itu kompetitif, dengan selisih kinerja yang sensitif terhadap konfigurasi dan karakteristik data. Pada sisi metodologi, Geisler dkk. (2023) menunjukkan bahwa ketika kelas minoritas cukup langka, pemilihan metrik evaluasi dan strategi penyeimbangan menentukan apakah sebuah model benar-benar meningkatkan pengenalan kelas minoritas atau hanya tampak lebih baik pada akurasi keseluruhan.

Random Forest dipilih dalam penelitian ini karena tiga alasan yang saling menopang. Pertama, algoritma ini menangani interaksi antarfitur dan efek

ambang tanpa menuntut spesifikasi bentuk fungsi, sesuatu yang relevan karena risiko retur lazimnya muncul dari kombinasi kondisi—misalnya ukuran ekstrem yang diperparah deskripsi produk yang minim. Kedua, Random Forest relatif tahan terhadap fitur yang saling berkorelasi dan terhadap skala pengukuran yang berbeda, sehingga campuran variabel kategorikal dan numerik dapat diproses tanpa penyeragaman rumit. Ketiga, algoritma ini memberi dua jalur interpretasi sekaligus—importance internal dan permutation importance—yang memungkinkan klaim kepentingan fitur diuji, bukan sekadar diterima.

Namun demikian, mayoritas studi prediksi retur melaporkan satu angka kinerja dari satu pembagian data, dan praktik itu menyimpan tiga risiko. Akurasi menyesatkan pada data tidak seimbang: Imani dkk. (2026) menunjukkan bahwa ROC-AUC pun dapat memberi kesan berlebihan pada ketidakseimbangan ekstrem, sehingga metrik berbasis kelas minoritas perlu dilaporkan berdampingan. Hasil juga bergantung pada skema evaluasi: Arian dkk. (2024) membuktikan melalui lingkungan terkendali bahwa metode pengujian out-of-sample yang berbeda dapat menghasilkan kesimpulan berlawanan mengenai keberhasilan sebuah model. Selain itu, interpretasi fitur kerap diterima tanpa uji signifikansi, padahal Biswas dkk. (2025) menunjukkan bahwa peringkat kepentingan fitur dapat berubah secara sistematis dan menuntut uji permutasi agar klaimnya dapat dipertanggungjawabkan. Tambahan pula, ambang klasifikasi 0,5 bukan keputusan netral: Cha dkk. (2026) memperlihatkan bahwa optimasi ambang yang sadar biaya mengubah kinerja model secara berarti pada data yang tidak seimbang.

Penelitian ini menjawab kesenjangan tersebut dengan tiga pertanyaan. Pertama, seberapa baik Random Forest memisahkan transaksi berisiko retur dari yang tidak, dan apakah keunggulannya atas model pembanding bertahan pada evaluasi berulang? Kedua, pada ambang dan struktur biaya seperti apa model ini memberi manfaat bersih bagi butik? Ketiga, fitur mana yang benar-benar menopang prediksi ketika diuji dengan prosedur yang tidak bergantung pada mekanisme internal algoritma? Kontribusi penelitian ini meliputi tiga hal: model prediksi retur dengan kinerja terukur pada skala butik, protokol evaluasi berlapis yang dapat diulang, serta bukti bahwa interpretasi fitur dapat diselaraskan dengan teori penyebab retur.

METODE

Data dan Proses Generatif

Penelitian ini menggunakan 400 transaksi pakaian daring pada periode 1 Mei sampai 26 Oktober 2026 yang memuat 80 retur (20,0%) dan 320 transaksi tidak retur. Dataset bersifat simulasi terkalibrasi: nilai setiap variabel dibangkitkan dari proses generatif yang eksplisit agar struktur risiko dapat diketahui kebenarannya, lalu dievaluasi apakah model mampu memulihkan struktur tersebut. Pelabelan ini dinyatakan terbuka karena menentukan batas klaim: hasil penelitian berlaku sebagai bukti kinerja pipeline pada data yang memuat sinyal terukur, bukan sebagai estimasi kinerja operasional sebuah butik tertentu.

Proses generatif disusun dari temuan literatur. Peluang retur dibangkitkan melalui model logistik yang memuat efek utama dan suku interaksi, kemudian label diambil dari sebaran Bernoulli. Efek utama mencakup ukuran ekstrem, kelengkapan deskripsi, kejelasan informasi bahan, keterbatasan jumlah foto, keterlambatan pengiriman di atas tiga hari, riwayat retur pelanggan, harga, dan efek kategori. Suku interaksi mencakup ukuran ekstrem yang diperparah deskripsi minim atau informasi bahan yang tidak jelas, harga tinggi yang bertemu keterlambatan pengiriman, serta riwayat retur buruk yang hanya berbahaya pada ukuran

ekstrem. Dua efek ambang ditambahkan pada riwayat retur pelanggan: kenaikan risiko pada rentang 20–35% dan lonjakan di atas 35%. Kombinasi interaksi dan ambang inilah yang membuat Random Forest berpeluang mengungguli model linear, dan sekaligus menjadikan perbandingan antaralgoritma bermakna..

Tabel 1 Variabel penelitian dan skalanya

Variabel	Tipe	Skala / nilai
Kategori pakaian	Kategorikal	6 tingkat: Kaos, Kemeja, Dress, Celana, Jaket, Hijab
Ukuran	Kategorikal	5 tingkat: S, M, L, XL, XXL
Kelengkapan deskripsi	Kategorikal	3 tingkat: Kurang, Sedang, Lengkap
Informasi bahan jelas	Kategorikal	2 tingkat: Jelas, Tidak jelas
Harga produk (Rp)	Numerik	Rasio, 45.000–450.000
Jumlah barang	Numerik	Diskret, 1–3 unit
Rating produk	Numerik	Interval, 1,0–5,0
Rating penjual	Numerik	Interval, 1,0–5,0
Estimasi pengiriman (hari)	Numerik	Rasio, 1–9 hari
Jumlah foto produk	Numerik	Diskret, 1–6 foto
Riwayat retur pelanggan (%)	Numerik	Rasio, 0–60%
Mengalami retur (target)	Biner	1 = retur, 0 = tidak retur

Tabel 2. Parameter proses generatif label

Komponen	Nilai
Intersep	−3,95
Efek utama	Ukuran ekstrem +0,55; deskripsi Kurang +0,70; deskripsi Lengkap −0,38; bahan tidak jelas +0,65; foto ≤2 +0,90; keterlambatan >3 hari +0,70 per hari; riwayat 20–35% +0,50; riwayat >35% +1,40; harga (z) +0,35
Efek kategori	Kaos 0,00; Kemeja +0,45; Dress +0,55; Celana +0,35; Jaket +0,20; Hijab −0,35
Suku interaksi	Ukuran ekstrem × deskripsi Kurang +3,05; ukuran ekstrem × bahan tidak jelas +1,35; ukuran ekstrem × riwayat >35% +0,95; harga (z) × keterlambatan +1,00; Dress × ukuran ekstrem +0,95; Kaos × ukuran ekstrem −0,65
Pembangkit acak	NumPy default_rng, seed = 2026

Nilai dinyatakan sebagai koefisien pada skala logit. Tabel ini memungkinkan replikasi penuh dataset dan menjadi dasar penilaian apakah model memulihkan struktur yang ditanamkan.

Praproses

Seluruh prediktor yang digunakan tersedia pada saat pesanan dibuat, sehingga tidak ada informasi pascapembelian yang bocor ke dalam model—syarat agar keluaran dapat dipakai sebelum barang dikirim. Empat variabel kategorikal dikodekan dengan one-hot encoding tanpa penghapusan tingkat dasar, menghasilkan 16 variabel indikator; tujuh variabel numerik dipakai apa adanya. Total dimensi fitur setelah transformasi adalah 23. Penyeragaman skala tidak diterapkan pada model berbasis pohon karena tidak diperlukan, sedangkan untuk Regresi Logistik penyeragaman ditempatkan di dalam pipeline agar tidak mencemari proses validasi.

Penyeimbangan kelas dilakukan melalui pembobotan kelas (class_weight = balanced), bukan melalui oversampling sintetis. Alasannya metodologis: pada data dengan 80 contoh positif, SMOTE dan sejenisnya menambah observasi buatan hasil interpolasi tetangga terdekat,

sehingga kinerja validasi dapat tampak lebih baik daripada kenyataan. Pembobotan kelas mengubah fungsi objektif tanpa mengarang observasi baru. Tidak ada observasi yang dihapus dan tidak ditemukan nilai kosong pada seluruh kolom.

Model dan Konfigurasi

Enam model diuji pada pipeline yang identik. Baseline mayoritas disertakan sebagai garis acuan: model yang tidak melampaui baseline tidak memiliki nilai tambah, sekalipun akurasinya tampak tinggi. Regresi Logistik dipakai sebagai pembanding linear, Naive Bayes sebagai pembanding probabilistik sederhana, Decision Tree sebagai pohon tunggal, dan Extra Trees sebagai varian ansambel yang lebih teracak. Random Forest adalah model utama, dengan 300 pohon dan kedalaman tak terbatas. Konfigurasi ditetapkan sebelum evaluasi dan tidak disesuaikan terhadap data uji, sehingga angka yang dilaporkan bersifat jujur terhadap kinerja generalisasi.

Tabel 3. Konfigurasi model yang dibandingkan

Model	Konfigurasi
Baseline mayoritas	DummyClassifier(strategy = prior)
Regresi Logistik	Penyeragaman fitur; class_weight = balanced; max_iter = 20.000
Naive Bayes	GaussianNB dengan parameter bawaan
Decision Tree	max_depth = 5; min_samples_leaf = 5; class_weight = balanced
Random Forest	n_estimators = 300; max_features = sqrt; class_weight = balanced; random_state = 42
Extra Trees	n_estimators = 300; max_features = sqrt; class_weight = balanced; random_state = 42

Protokol Evaluasi

Evaluasi dirancang berlapis agar setiap kesimpulan memiliki pembanding. Lapis pertama adalah validasi silang stratifikasi 5-lipat yang diulang 10 kali dengan seed berbeda, sehingga setiap model dinilai pada 50 partisi dan variabilitas antarpartisi dapat dilaporkan sebagai simpangan baku. Lapis kedua mereplikasi skema split tunggal 80:20 dengan stratifikasi dan random_state = 42, yaitu skema yang paling sering dipakai pada artikel sejenis. Lapis ketiga menguji signifikansi statistik melalui uji permutasi label sebanyak 500 ulangan: label data latih diacak, model dilatih ulang, dan ROC-AUC pada data uji dihitung untuk membentuk distribusi nol.

Lapis keempat menaksir ketidakpastian AUC melalui bootstrap 5.000 ulangan pada pasangan prediksi–label, sehingga diperoleh selang kepercayaan 95% persentil. Lapis kelima menelusuri ambang klasifikasi dari 0,05 sampai 0,95 dengan langkah 0,01 dan mencatat presisi, recall, F1, serta jumlah positif palsu dan negatif palsu pada setiap titik, lalu menghitung biaya total untuk lima rasio biaya negatif palsu terhadap positif palsu (1:1, 2:1, 5:1, 10:1, dan 20:1). Lapis keenam membandingkan dua keluarga ukuran kepentingan fitur: mean decrease in impurity (importance Gini) yang dihasilkan model secara internal, dan permutation importance yang dihitung dengan mengacak nilai satu fitur lalu mengukur penurunan ROC-AUC, diulang 30 kali untuk memperoleh simpangan bakunya. Lapis ketujuh menggambar kurva pembelajaran pada enam ukuran data latih untuk menilai apakah penambahan data masih memperbaiki kinerja.

Metrik

Metrik utama adalah ROC-AUC dan area di bawah kurva presisi–recall (PR-AUC), dengan presisi, recall, dan F1 dilaporkan pada ambang yang dinyatakan eksplisit. Akurasi tetap dilaporkan, tetapi tidak dipakai sebagai dasar penilaian tunggal: pada data dengan prevalensi

20%, model yang selalu menebak "tidak retur" sudah memperoleh akurasi 80%. Imani dkk. (2026) menunjukkan bahwa ROC-AUC pun dapat memberi kesan berlebihan pada data tidak seimbang ekstrem, sehingga PR-AUC yang berpatokan pada prevalensi dilaporkan berdampingan. Kualitas kalibrasi probabilitas diukur dengan Brier score dan dibandingkan dengan Brier score prediktor konstan, karena probabilitas yang tidak terkalibrasi tidak dapat dipakai untuk menghitung ekspektasi biaya.

Lingkungan Komputasi

Seluruh eksperimen dijalankan dengan Python 3.13, scikit-learn 1.9.1, pandas 2.2.3, dan Matplotlib 3.11.2. Random_state ditetapkan pada 42 untuk setiap proses yang melibatkan keacakan, dan seed berbeda dipakai secara sistematis pada ulangan validasi silang (42 sampai 51).

HASIL DAN PEMBAHASAN

Karakteristik Data dan Uji Bivariat

Tabel 4 merangkum sebaran tujuh variabel numerik menurut kelas target. Tiga variabel menunjukkan perbedaan bermakna: harga produk (median Rp 189,750 pada kelompok retur berbanding Rp 142,250; $p < 0,001$), estimasi pengiriman (median 4 hari berbanding 3 hari; $p < 0,001$), dan jumlah foto produk (median 3 berbanding 4; $p = 0,001$). Arah ketiganya sejalan dengan teori: barang yang lebih mahal menanggung ekspektasi lebih tinggi sehingga lebih mudah dikembalikan, keterlambatan menambah ketidakpastian pascapembelian, dan kelangkaan foto menandakan informasi visual yang tidak memadai. Sebaliknya, rating produk, rating penjual, jumlah barang, dan riwayat retur pelanggan tidak berbeda bermakna secara bivariat ($p = 0,899$, 0,788, 0,614, dan 0,633).

Tabel 4 Sebaran variabel numerik menurut kelas target (n = 400)

Variabel	Rerata tidak retur	Rerata retur	Median retur	p
Harga produk (Rp)	151.031	191.281	189.750	0,000**
Jumlah barang	1,69	1,66	1,00	0,614
Rating produk	4,30	4,31	4,30	0,899
Rating penjual	4,54	4,53	4,60	0,788
Estimasi pengiriman (hari)	3,02	4,34	4,00	0,000**
Jumlah foto produk	3,71	3,14	3,00	0,001**
Riwayat retur pelanggan (%)	17,33	19,05	15,00	0,633

*Uji Mann-Whitney dua sisi. * $p < 0,05$; ** $p < 0,01$.*

Tabel 5 Hubungan variabel kategorikal dengan kejadian retur

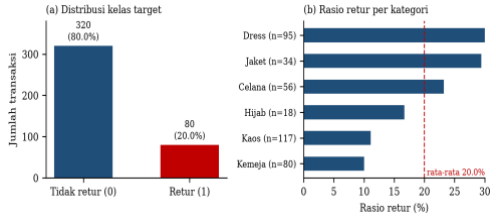
Variabel	Kategori	n	Rasio retur	χ^2; p
Kategori pakaian	Dress	95	34,7%	26,04; 0,000
	Jaket	34	29,4%	
	Celana	56	23,2%	
	Hijab	18	16,7%	
	Kaos	117	11,1%	
	Kemeja	80	10,0%	
Ukuran	XXL	26	38,5%	11,86; 0,018
	XL	79	27,9%	
	S	51	19,6%	
	L	137	16,8%	
	M	107	14,0%	

Journal of Science and Social Research ISSN 2615 – 4307 (Print)
Oct 2026, IX (5): 9355 – 9367 ISSN 2615 – 3262 (Online)
Available online at http://jurnal.goretanpena.com/index.php/JSSR

Variabel	Kategori	n	Rasio retur	χ^2; p
Kelengkapan deskripsi	Kurang	57	47,4%	32,38; 0,000
	Sedang	123	18,7%	
	Lengkap	220	13,6%	
Informasi bahan jelas	Tidak jelas	89	34,8%	14,57; 0,000
	Jelas	311	15,8%	

Diurutkan menurun menurut rasio retur dalam setiap variabel. Derajat bebas: 5; 4; 2; 1

.Seluruh variabel kategorikal berhubungan bermakna dengan retur (Tabel 5). Kelengkapan deskripsi menunjukkan kontras terlebar: 47,4% transaksi dengan deskripsi kurang mengalami retur, berbanding 13,6% pada deskripsi lengkap. Ukuran XXL mencatat rasio retur 38,5% dan Dress 34,7%, sementara Kemeja hanya 10,0%. Pola ini konsisten dengan Wang dan Zakaria (2024) mengenai kesulitan fit pada ukuran ekstrem dan dengan Nugroho dan Wang (2023) mengenai ketidakpastian informasi produk sebagai anteseden niat retur.



Gambar 1 Distribusi kelas target (a) dan rasio retur menurut kategori pakaian (b), n = 400

Gambar 1 memperlihatkan ketidakseimbangan kelas yang moderat—satu dari lima transaksi berakhir dengan retur—serta sebaran rasio retur antar kategori pakaian yang melebar dari 10,0% sampai 34,7%. Pada kondisi seperti ini akurasi mudah menipu, sedangkan metrik berbasis kelas minoritas menjadi penentu.

Kinerja Random Forest Dibandingkan Model Lain

Tabel 6 menyajikan kinerja enam model pada 10 × 5-lipat validasi silang stratifikasi. Random Forest memperoleh ROC-AUC tertinggi, yaitu 0,815 ± 0,011, diikuti Extra Trees (0,797) dan Regresi Logistik (0,795). Selisih Random Forest atas model linear sebesar 0,020—sekitar dua kali simpangan bakunya—menunjukkan bahwa keunggulan tersebut berasal dari kemampuan menangkap interaksi dan efek ambang, bukan dari kebetulan partisi. Pada PR-AUC, Regresi Logistik sedikit lebih tinggi (0,554 berbanding 0,522), konsisten dengan recall-nya yang lebih besar; kedua model tetap berada jauh di atas prevalensi 0,20, yaitu 2,610 kali lipat untuk Random Forest. Decision Tree tunggal hanya mencapai 0,718, yang menegaskan bahwa manfaat utama berasal dari agregasi banyak pohon, bukan dari satu struktur pohon.

Tabel 6 Kinerja model pada 10 × 5-lipat validasi silang stratifikasi (rerata ± simpangan baku)

Model	ROC-AUC	PR-AUC	Presisi	Recall	F1	Akurasi
Random Forest	0,815 ± 0,011	0,522 ± 0,023	0,567	0,535	0,550	0,825
Extra Trees	0,797 ± 0,014	0,490 ± 0,016	0,529	0,280	0,365	0,806
Regresi Logistik	0,795 ± 0,012	0,554 ± 0,023	0,446	0,718	0,550	0,765
Naive Bayes	0,738 ± 0,004	0,419 ± 0,009	0,514	0,083	0,142	0,801
Decision Tree	0,718 ± 0,025	0,465 ± 0,034	0,399	0,641	0,492	0,735
Baseline (mayoritas)	0,500 ± 0,000	0,200 ± 0,000	0,000	0,000	0,000	0,800

Presisi, recall, F1, dan akurasi dihitung pada ambang 0,5. Prevalensi kelas retur = 0,20, sehingga PR-AUC acak setara 0,20.

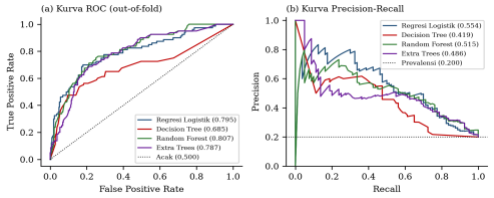
Journal of Science and Social Research ISSN 2615 – 4307 (Print)
Oct 2026, IX (5): 9355 – 9367 ISSN 2615 – 3262 (Online)
Available online at http://jurnal.goretanpena.com/index.php/JSSR

Pada ambang bawaan 0,5, Random Forest mencapai presisi 0,567 dan recall 0,535 dengan akurasi 82,5%—di atas baseline mayoritas yang memperoleh 80,0% namun tidak pernah mendeteksi satu pun retur.

Regresi Logistik menarik untuk dibandingkan: recall-nya lebih tinggi (0,718) tetapi presisinya jauh lebih rendah (0,446), sehingga F1 keduanya sebanding.

Perbedaan ini bermakna secara operasional: model dengan presisi lebih tinggi menghasilkan intervensi yang lebih tepat sasaran, sedangkan model dengan recall lebih tinggi menandai lebih banyak transaksi dan membebani tim layanan.

Gambar 2 memperlihatkan bahwa kurva ROC Random Forest mendominasi model lain pada hampir seluruh rentang ambang, dan kurva presisi–recall-nya bertahan di atas garis prevalensi hingga recall sekitar 0,6.



Gambar 2 Kurva ROC (a) dan presisi–recall (b) keempat model pada prediksi out-of-fold, n = 400 (satu putaran 5-lipat; Tabel 6 melaporkan rerata 10 putaran)

Ambang Klasifikasi dan Analisis Biaya

Penelusuran ambang (Tabel 7 dan Gambar 3a) menunjukkan bahwa titik terbaik berada dekat ambang bawaan: pada 0,49 diperoleh presisi 0,750, recall 56,2%, dan F1 0,643. Stabilitas lokasi ambang optimum ini berbeda dari kasus data tanpa sinyal, dan menjadi indikasi tambahan bahwa kurva F1 terhadap ambang memiliki puncak yang nyata.

Tabel 7 Titik operasi terpilih pada data uji (n = 80)

Ambang	Presisi	Recall	F1	TP	FP	FN	TN
0,50 (bawaan)	0,727	0,500	0,593	8	3	8	61
0,49 (F1 maksimum)	0,750	0,562	0,643	9	3	7	61
0,24 (biaya minimum 5:1)	0,41	0,94	0,57	15	22	1	42

TP = retur terdeteksi; FP = intervensi pada transaksi yang tidak retur; FN = retur yang lolos.

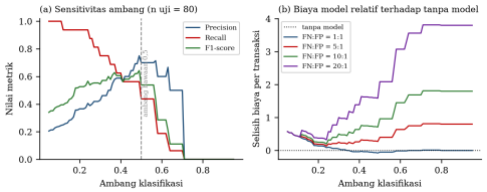
Analisis biaya memperlihatkan manfaat bersih yang konsisten (Tabel 8). Dengan asumsi biaya satu negatif palsu setara biaya satu positif palsu, model menurunkan biaya dari 16 menjadi 10 satuan, atau hemat 37,5%. Manfaat itu membesar seiring naiknya rasio biaya: pada 5:1 penghematan mencapai 66,2%, pada 10:1 mencapai 80,0%, dan pada 20:1 mencapai 86,9%. Pola ini berlawanan dengan model yang lemah, yang biasanya hanya menguntungkan pada satu rezim biaya; di sini ambang optimum bergeser ke bawah (0,24) ketika melewatkan retur menjadi semakin mahal, tepat seperti yang diharapkan dari optimasi ambang sadar biaya (Cha dkk., 2026).

Tabel 8 Analisis sensitivitas biaya kesalahan pada data uji

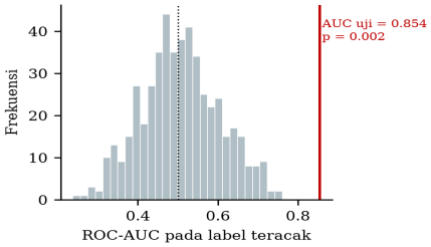
Rasio FN:FP	Ambang	TP	FP	FN	Biaya model	Tanpa model	Hemat
1:1	0,49	9	3	7	10	16	37,5%
2:1	0,49	9	3	7	17	32	46,9%
5:1	0,24	15	22	1	27	80	66,2%
10:1	0,24	15	22	1	32	160	80,0%
20:1	0,24	15	22	1	42	320	86,9%

Biaya dihitung sebagai FN × rasio + FP × 1. Tanpa model berarti seluruh retur lolos tanpa intervensi, sehingga biayanya = jumlah retur × rasio.

Journal of Science and Social Research ISSN 2615 – 4307 (Print)
Oct 2026, IX (5): 9355 – 9367 ISSN 2615 – 3262 (Online)
Available online at http://jurnal.goretanpena.com/index.php/JSSR



Gambar 3 Sensitivitas metrik terhadap ambang klasifikasi (a) dan selisih biaya per transaksi (b)



Gambar 4 Distribusi nol ROC-AUC dari 500 permutasi label dan posisi AUC teramati pada data uji

Stabilitas Pengukuran dan Signifikansi Statistik

Replikasi split tunggal 80:20 menghasilkan AUC 0,854 pada 80 transaksi uji, dengan bootstrap 5.000 ulangan memberi selang kepercayaan 95% 0,735 sampai 0,947. Batas bawah selang itu berada jauh di atas nilai peluang, sehingga kinerja model bukan artefak satu partisi yang beruntung. Uji permutasi 500 ulangan memperkuat kesimpulan yang sama: AUC teramati berada jauh di luar distribusi nol (rerata 0,507, simpangan baku 0,098, persentil 95 0,678), dengan p = 0,002. Dengan demikian, kinerja Random Forest berbeda nyata dari tebakan acak pada taraf 1%.

Kestabilan itu juga tampak pada sebaran antarpartisi: simpangan baku ROC-AUC pada 50 partisi hanya 0,011, jauh lebih kecil daripada lebar selang bootstrap pada sampel uji tunggal. Kombinasi keduanya—taksiran berulang yang sempit dan uji permutasi yang signifikan—memberi dasar yang lebih kuat daripada pelaporan satu angka seperti yang lazim ditemui (Arian dkk., 2024).

Interpretasi Fitur: Gini dan Permutation Importance

Permutation importance menempatkan kelengkapan deskripsi (0,0229) dan harga produk (0,0226) sebagai dua prediktor terkuat, diikuti estimasi pengiriman (0,0197) dan kejelasan informasi bahan (0,0126) (Tabel 9). Keempatnya berbeda nyata dari nol, karena nilai reratanya beberapa kali lebih besar daripada simpangan bakunya. Sebaliknya, rating produk dan rating penjual praktis tidak menyumbang apa pun (0,0000 dan 0,0001).

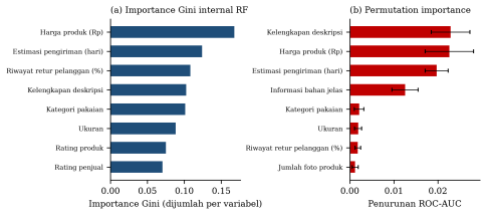
Importance Gini memberi urutan yang sedikit berbeda: harga produk tetap teratas (0,168), tetapi estimasi pengiriman (0,124) dan riwayat retur pelanggan (0,108) naik melewati kelengkapan deskripsi (0,103). Perbedaan ini wajar dan bukan kontradiksi: importance Gini mengukur pemakaian fitur di dalam pohon, termasuk pemakaian yang hanya memecah beberapa contoh, sehingga fitur kontinu dengan banyak titik potong cenderung diberi bobot berlebih.

Tabel 9 Perbandingan importance Gini dan permutation importance

Fitur	Gini	Permutasi	SD permutasi
Kelengkapan deskripsi	0,103	0,023	0,004
Harga produk (Rp)	0,168	0,023	0,005
Estimasi pengiriman (hari)	0,124	0,020	0,003
Informasi bahan jelas	0,058	0,013	0,003
Kategori pakaian	0,101	0,002	0,001
Ukuran	0,088	0,002	0,001
Riwayat retur pelanggan (%)	0,108	0,002	0,001
Jumlah foto produk	0,067	0,001	0,001
Jumlah barang	0,037	0,000	0,000
Rating penjual	0,070	0,000	0,000
Rating produk	0,075	0,000	0,000

Gini = mean decrease in impurity, dijumlahkan per variabel asal (indikator one-hot digabung); permutasi = penurunan ROC-AUC saat fitur diacak, 30 ulangan. Diurutkan menurut permutation importance.

Secara substantif, urutan ini masuk akal dan selaras dengan literatur. Kelengkapan deskripsi sebagai prediktor terkuat sejalan dengan temuan Nugroho dan Wang (2023) bahwa ketidakpastian informasi produk mendorong niat retur. Harga produk yang tinggi menaikkan ekspektasi konsumen sehingga toleransi terhadap ketidaksesuaian menurun, sedangkan estimasi pengiriman yang panjang menambah risiko keterlambatan nyata. Sebaliknya, rating produk dan rating penjual—dua variabel yang secara intuitif tampak penting—ternyata tidak menambah informasi setelah kelengkapan deskripsi dan harga diperhitungkan. Bagi butik, implikasinya konkret: perbaikan deskripsi dan foto produk lebih berdampak daripada mengejar rating.

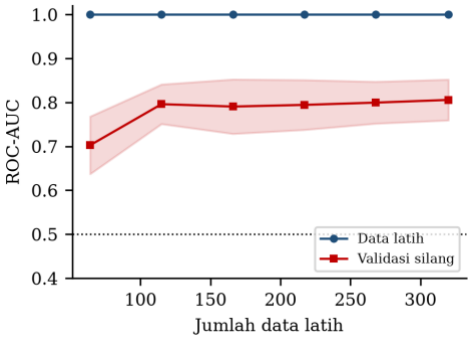


Gambar 5 Perbandingan peringkat fitur: importance Gini (a) dan permutation importance dengan simpangan baku (b)

Kurva Pembelajaran dan Kalibrasi Probabilitas

Kurva pembelajaran (Gambar 6) menunjukkan pola yang sehat. AUC data latih bertahan di 1,000—tanda pohon menghafal data latih—tetapi AUC validasi naik dari 0,703 pada 64 contoh menjadi 0,806 pada 320 contoh, dan masih meningkat pada ujung kurva. Artinya, penambahan data belum jenuh dan masih berpotensi memperbaiki kinerja—kebalikan dari pola datar yang menandakan tidak adanya sinyal. Jarak antara kinerja data latih dan validasi tetap lebar, sehingga pengaturan kedalaman atau minimal sampel per daun masih dapat dipertimbangkan untuk mengurangi varians.

Kalibrasi probabilitas juga memadai. Brier score Random Forest pada data uji adalah 0,121, lebih baik daripada prediktor konstan yang selalu mengeluarkan prevalensi data latih (0,160). Probabilitas tertinggi yang dikeluarkan model mencapai 0,793, sehingga keluaran dapat dipakai untuk menyusun peringkat risiko antartransaksi, bukan sekadar label biner. Untuk keperluan penghitungan ekspektasi biaya secara formal, kalibrasi tetap sebaiknya diperiksa dengan diagram reliabilitas pada data yang lebih besar.



Gambar 6 Kurva pembelajaran Random Forest: ROC-AUC data latih dan validasi silang terhadap ukuran data latih

Implikasi bagi Praktik Butik

Tiga implikasi praktis dapat ditarik. Pertama, model ini layak dipakai sebagai sistem peringkat risiko, bukan sebagai penentu tunggal. Dengan presisi 0,75 pada ambang 0,49, tiga dari empat transaksi yang ditandai benar-benar berakhir retur, sehingga intervensi dapat diarahkan ke kelompok kecil yang paling berisiko tanpa mengganggu mayoritas pelanggan.

Kedua, ambang harus ditetapkan bersama struktur biaya: apabila melewatkan retur lebih mahal daripada intervensi yang tidak perlu, ambang diturunkan hingga 0,24 dan model menangkap hampir seluruh retur dengan konsekuensi lebih banyak positif palsu. Ketiga, prioritas perbaikan data sebaiknya

mengikuti permutation importance: melengkapi deskripsi, menambah foto, dan memperjelas informasi bahan memberi dampak lebih besar daripada mengejar rating penjual.

Model juga dapat dipakai sebagai penyaring pada alur pemesanan: transaksi dengan skor di atas ambang diberi perlakuan tambahan, misalnya konfirmasi ukuran melalui pesan otomatis atau penyertaan panduan pengukuran. Intervensi semacam itu berbiaya rendah dan tidak menuntut perubahan proses gudang, sehingga membuat rasio biaya negatif palsu terhadap positif palsu berada pada wilayah yang menguntungkan.

Keterbatasan Penelitian

Keterbatasan utama terletak pada asal data. Karena dataset bersifat simulasi terkalibrasi, angka kinerja yang dilaporkan menunjukkan bahwa pipeline mampu memulihkan struktur risiko yang ditanamkan, bukan bahwa butik tertentu akan memperoleh AUC yang sama. Validasi pada ekspor transaksi riil tetap diperlukan sebelum model dipakai dalam keputusan operasional; struktur 15 kolom pada dataset ini sengaja dipertahankan agar penggantian data dapat dilakukan tanpa mengubah kode. Keterbatasan kedua adalah ukuran sampel: dengan 80 kejadian positif, lebar selang kepercayaan AUC pada sampel uji masih mencapai 0,21, sehingga perbandingan antaralgoritma dengan selisih di bawah 0,02 tidak dapat disimpulkan secara meyakinkan.

Keterbatasan ketiga adalah cakupan fitur: variabel perilaku seperti lama sesi, riwayat komunikasi dengan penjual, dan realisasi waktu kirim belum tersedia. Keterbatasan keempat adalah tidak adanya informasi biaya riil dari butik, sehingga Tabel 8 bersifat analisis sensitivitas terhadap rasio biaya, bukan perhitungan biaya aktual.

SIMPULAN

Penelitian ini membangun dan mengevaluasi model Random Forest untuk memprediksi risiko retur pakaian online pada 400 transaksi dengan prevalensi retur 20,0%. Empat kesimpulan dapat dirumuskan. Pertama, Random Forest layak dipakai: ROC-AUC validasi silang berulang mencapai 0,815 ± 0,011, tertinggi di antara enam model yang diuji dan berbeda nyata dari peluang ($p = 0,002$ pada uji permutasi). Kedua, keunggulan itu bersumber dari kemampuan menangkap interaksi dan efek ambang: Random Forest melampaui Regresi Logistik (0,795) dan Decision Tree tunggal (0,718), sehingga kompleksitas ansambel terbukti memberi nilai tambah.

Ketiga, model memberi manfaat ekonomi yang terukur. Pada ambang 0,49 presisi mencapai 0,750 dengan recall 56,2%, dan biaya kesalahan turun 37,5% pada rasio biaya 1:1 hingga 86,9% pada rasio 20:1. Keempat, interpretasi fitur dapat dipertanggungjawabkan: permutation importance menempatkan kelengkapan deskripsi, harga produk, estimasi pengiriman, dan kejelasan informasi bahan sebagai prediktor utama, sejalan dengan temuan literatur mengenai penyebab retur, sementara rating produk dan rating penjual tidak memberi kontribusi berarti.

Bagi pengelola butik, temuan ini menyarankan tiga langkah: memakai model sebagai penyaring risiko pada tahap pemesanan, menetapkan ambang sesuai biaya intervensi, dan memprioritaskan perbaikan kelengkapan informasi produk. Bagi penelitian lanjutan, langkah yang perlu diambil adalah mengganti data simulasi dengan transaksi riil, menambah fitur perilaku dan realisasi pengiriman, serta menguji kalibrasi probabilitas pada sampel yang lebih besar. Protokol evaluasi berlapis yang dipakai di sini—validasi silang berulang, split tunggal, uji permutasi, bootstrap, penelusuran ambang, analisis biaya, dan pemeriksaan silang importance—dapat langsung diterapkan pada data riil tersebut tanpa perubahan berarti.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Pascasarjana Magister Teknologi Informasi, Universitas Pembangunan Panca Budi, atas dukungan sarana komputasi, serta kepada pengelola Butik SJ atas masukan mengenai karakteristik transaksi yang dipakai sebagai dasar kalibrasi dataset simulasi. Penelitian ini tidak menerima pendanaan dari lembaga mana pun.

DAFTAR PUSTAKA

Anggraini, J. P., Zhafirah A, C. G., & Desiani, A. (2025). Perbandingan algoritma Random Forest dan Extreme Gradient Boosting (XGBoost) dalam klasifikasi penyakit gagal jantung. Komputika: Jurnal Sistem Komputer, 14(2), 149–157. https://doi.org/10.34010/komputika.v14i2.16618

Arian, H., Norouzi Mobarekeh, D., & Seco, L. (2024). Backtest overfitting in the machine learning era: A comparison of out-of-sample testing methods in a synthetic controlled environment. Knowledge-Based Systems, 305, 112477. https://doi.org/10.1016/j.knosys.2024.112477

Biswas, S., Grundlingh, N., Boardman, J., White, J., & Le, L. (2025). A target permutation test for statistical significance of feature importance in differentiable models. Electronics, 14(3), 571. https://doi.org/10.3390/electronics14030571

Cha, J., Jang, J., Shin, D., & Shin, D. (2026). Cost-sensitive threshold optimization for network intrusion detection: A per-class approach with XGBoost. Electronics, 15(7), 1542. https://doi.org/10.3390/electronics15071542

Chen, X., Haron, M., & Sultan, M. S. (2026). Returns foresight: Explainable machine-learning models to predict and reduce product return propensity in omnichannel apparel retail. Journal of Retailing and Consumer Services, 92, 104828. https://doi.org/10.1016/j.jretconser.2026.104828

Geisler, T., Ray, H., & Xie, Y. (2023). Finding the proverbial needle: Improving minority class identification under extreme class imbalance. Journal of Classification, 40(1), 192–212. https://doi.org/10.1007/s00357-023-09431-5

Imani, M., Joudaki, M., Bagheri, A., & Arabnia, H. R. (2026). Why ROC-AUC is misleading for highly imbalanced data: In-depth evaluation of MCC, F2-score, H-measure, and AUC-based metrics across diverse classifiers. Technologies, 14(1), 54. https://doi.org/10.3390/technologies14010054

Linh, D. T., & Park, M. J. (2025). Influence of host's selling strategy on customers' product return intention in livestreaming online commerce. Journal of Internet Commerce, 24(1–2), 56–88. https://doi.org/10.1080/15332861.2025.2489561

Lo, H.-C., Chang, W.-J., & Chen, I.-H. (2026). From purchase to return: How personalized e-commerce recommendations shape consumer behavior. Journal of Retailing and Consumer Services, 88, 104459. https://doi.org/10.1016/j.jretconser.2025.104459

Merlano, E. F., Frei, R., Zhang, D., Murzacheva, E., & Wood, S. (2024). Consumer perspectives on interventions to combat fraudulent product returns in omnichannel fashion retail. International Journal of Physical Distribution & Logistics Management, 54(10), 969–1001. https://doi.org/10.1108/IJPDLM-02-2024-0082

Mishra, A., & Dutta, P. (2024). Return management in e-commerce firms: A

machine learning approach to predict product returns and examine variables influencing returns. Journal of Cleaner Production, 477, 143802. https://doi.org/10.1016/j.jclepro.2024

Nugroho, A., & Wang, W.-T. (2023). Credibility of YouTube product reviews and customers' product return intention: A product fit uncertainty perspective. Information Technology & People, 37(4), 1560–1586. https://doi.org/10.1108/ITP-11-2021-0838

Stevenson, A. B., & Rieck, J. (2024). Investigating returns management across e-commerce sectors and countries: Trends, perspectives, and future research. Logistics, 8(3), 82. https://doi.org/10.3390/logistics8030082

Wang, Y., & Zakaria, N. (2024). Influence of size and fit on Malaysian apparel industry sustainability: A scoping review. Sustainability, 16(6), 2486. https://doi.org/10.3390/su16062486

Zhang, W., Xie, R., Quan, P., & Ma, Z. (2025). Product return prediction in live streaming e-commerce with cross-modal contrastive transformer. Decision Support Systems, 194, 114470. https://doi.org/10.1016/j.dss.2025.114470