

## PENGEMBANGAN SISTEM PENDUKUNG KEPUTUSAN MANAJEMEN RISIKO LOGISTIK E-COMMERCE BERBASIS MACHINE LEARNING MENGGUNAKAN RANDOM FOREST PIPELINE

Fetty Tri Anggraeny<sup>1</sup>, Naufal Firman Dhani<sup>2</sup>

UPN "Veteran Jawa Timur"<sup>1,2</sup>

Email: [23081010148@student.upnjatim.ac.id](mailto:23081010148@student.upnjatim.ac.id)

Informasi	Abstract
Volume : 3 Nomor : 6 Bulan : Juni Tahun : 2026 E-ISSN : 3062-9624	<p><i>The e-commerce logistics industry faces challenges due to the high risk of fleet delivery delays, where the analysis of risk mitigation and its financial impact is often limited for retail operational management. This project aims to develop an Integrated Logistics Risk Management Decision Support System Smart Platform named LogiTrack v2.0, which combines two main data engineering domains: Data Mining and Machine Learning. For data analysis, an artificial intelligence model was developed using the Random Forest Classifier algorithm within a Scikit-Learn Pipeline architecture, which detects minority class data via the class_weight='balanced' parameter to automatically predict cargo delay status based on the engineering of four composite features (Volume, Density, Shipping Fee per Gram, and Route). For smart interaction, binary file .pkl optimization was implemented using the Joblib library with compression level 3 to radically reduce storage capacity from 138.80 MB to 28.40 MB. All of these functionalities are integrated into a single interactive web application built using Streamlit and deployed on the Streamlit Cloud server. The dashboard is equipped with two main tools: a Guardrail Engine function to block package inputs exceeding retail capacity (&gt;30\$ Kg) and a Business Impact Calculator panel to calculate the conversion of management financial penalties by 20% in real-time. Functional testing results via Black-Box Testing show that the classification model achieves a global accuracy rate of 90.00% and a precision value of 0.31, while the web platform successfully operates stably with low latency. This platform offers a holistic solution that enhances cargo distribution transparency, optimizes operational risk control efficiency, and suppresses potential financial losses within the logistics industry ecosystem.</i></p> <p><b>Keyword:</b> Logistics, Machine Learning, Data Mining, Streamlit, LogiTrack</p> <p><b>Abstrak</b> Industri logistik e-commerce menghadapi tantangan tingginya risiko keterlambatan pengiriman armada, di mana analisis mitigasi risiko dan dampak finansialnya seringkali terbatas bagi manajemen operasional ritel. Proyek ini bertujuan untuk mengembangkan Platform Cerdas Sistem Pendukung Keputusan Manajemen Risiko Logistik Terintegrasi bernama LogiTrack v2.0 yang menggabungkan domain Data Mining dan Machine Learning. Menggunakan Brazilian E-Commerce Public Dataset oleh Olist, dikembangkan model kecerdasan buatan menggunakan algoritma Random Forest Classifier dalam arsitektur Scikit-Learn Pipeline yang memprediksi status keterlambatan kargo berdasarkan rekayasa empat fitur komposit (Volume, Densitas, Ongkir per Gram, dan Rute). Hambatan imbalanced data dengan rasio ekstrem 12:1 dimitigasi via parameter class_weight='balanced'. Optimasi berkas biner .pkl dieksekusi menggunakan pustaka Joblib dengan kompresi tingkat 3 untuk mereduksi ukuran penyimpanan dari 138.80 MB menjadi 28.40 MB guna memenuhi restriksi peladen. Seluruh fungsionalitas diintegrasikan ke dalam aplikasi web Streamlit Cloud yang dilengkapi Guardrail Engine</p>

(\$>30\$ Kg) dan *Business Impact Calculator* (penalti 20%). Hasil pengujian *Black-Box* menunjukkan model klasifikasi mencapai tingkat akurasi global sebesar 90.00% dan nilai presisi sebesar 0.31. Platform ini menawarkan solusi holistik yang meningkatkan transparansi distribusi kargo, mengoptimalkan efisiensi kontrol risiko operasional, dan menekan potensi kerugian finansial pada ekosistem industri logistik.

**Kata Kunci:** Logistik, Machine Learning, Data Mining, Streamlit, LogiTrack

## A. PENDAHULUAN

Indonesia saat ini tengah fokus pada transformasi digital di berbagai sektor industri guna meningkatkan efisiensi operasional dan daya saing global. Industri *e-commerce* dan logistik memiliki peran krusial sebagai pilar utama penggerak ekonomi digital, di mana kelancaran arus barang menjadi faktor penentu keberhasilan bisnis. Keberhasilan suatu perusahaan logistik tidak hanya diukur dari volume pengiriman, melainkan dari ketepatan waktu sampai ke tangan konsumen karena performa pengiriman menjadi indikator utama dalam penilaian kepuasan pelanggan (*customer retention*) [1].

Proses pengiriman barang dalam ekosistem *e-commerce* bersifat sangat dinamis dan dipengaruhi oleh rute spasial maupun parameter operasional lainnya yang terekam dalam dataset publik Olist Brasil [2]. Secara tradisional, pemantauan keterlambatan pengiriman sering kali dilakukan secara reaktif ketika masalah sudah terjadi di lapangan. Transformasi digital kini menuntut pendekatan yang lebih proaktif dan preventif, di mana data historis transaksi dapat dimanfaatkan melalui teknologi *Machine Learning* untuk memitigasi risiko keterlambatan sejak dini [3].

Oleh karena itu, penelitian ini mengembangkan sebuah proyek sistem pendukung keputusan bertajuk LogiTrack yang menerapkan tahapan *Data Science Lifecycle* secara komprehensif. Fokus penelitian diarahkan pada penanganan masalah ketidakseimbangan kelas (*data imbalance*) yang lazim ditemui pada log data distribusi [4], enkapsulasi model, hingga kompresi berkas biner model guna mendukung pengambilan keputusan operasional logistik secara strategis [5].

## B. METODE PENELITIAN

Arsitektur pipa data dan pemodelan pada penelitian LogiTrack v2.0 dibangun secara terstruktur melalui empat tahapan utama:

### A. Konsolidasi dan Pembersihan Data (*Data Ingestion & Cleansing*)

Tahap awal dilakukan dengan memuat file relasional terpisah dari dataset publik Olist Brasil menggunakan pustaka *Pandas* dan *NumPy*. Seluruh tabel transaksi, detail produk, dan

data translasi bahasa (*product\_category\_name\_translation.csv*) disatukan melalui operasi *Merge DataFrame*. Data kotor berupa nilai yang hilang (*missing values*) dieliminasi, serta dilakukan penyaringan terhadap pencilan data kargo (*outliers filtering*) guna menjaga integritas sebaran statistik data.

B. Rekayasa Fitur Komposit (*Feature Engineering*)

Untuk memperkaya informasi sebab-akibat keterlambatan, variabel waktu transaksi statis yang bertindak sebagai *noise* dieliminasi. Penulis merancang 4 fitur komposit logistik baru yang bernilai bisnis tinggi yaitu Volume Produk, Densitas Kargo, Ongkir per Gram, Rute Distribusi

C. Enkapsulasi Pipa Pemodelan

Guna menghindari kebocoran data (*data leakage*) selama validasi silang, dibangun arsitektur terpadu menggunakan *Scikit-Learn Pipeline*. Objek transformer prapemrosesan (*StandardScaler* untuk numerik dan *OneHotEncoder* untuk kategorikal) dibungkus bersama dengan algoritma inti *Random Forest Classifier* ke dalam satu skrip tunggal *train\_model.py*.

D. Penanganan Ketidakseimbangan Kelas (*Mitigating Class Imbalance*)

Dataset Olist Brasil memiliki karakteristik ketidakseimbangan kelas yang ekstrem dengan rasio 12:1 antara paket tepat waktu terhadap paket terlambat. Isu ini dimitigasi secara terprogram dengan menyuntikkan parameter optimasi bobot *class\_weight='balanced'* ke dalam algoritma *Random Forest*, memberikan penalti matematika yang besar terhadap kesalahan klasifikasi kelas minoritas (paket terlambat).

C. HASIL DAN PEMBAHASAN

3.1 Penyajian Data dan Alat Bantu Visual

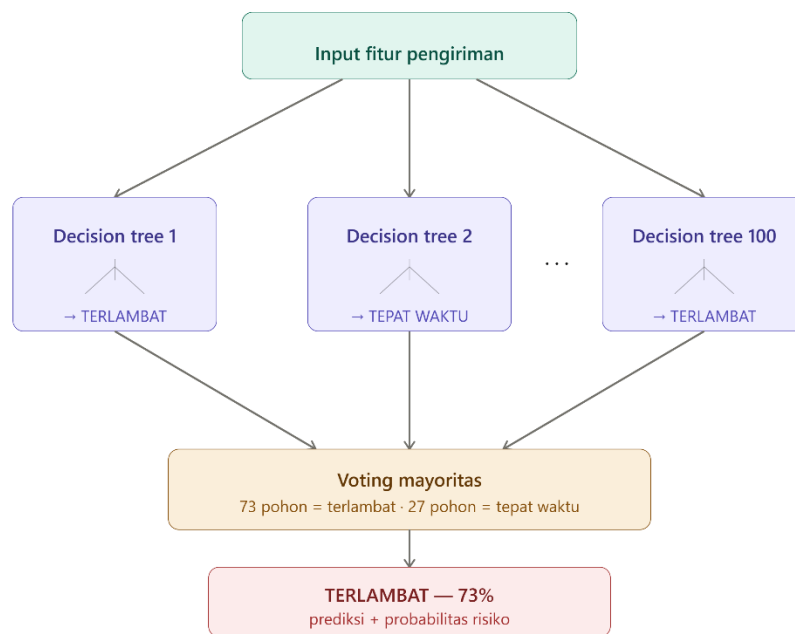
Penelitian ini mengekstrak data operasional dari *Brazilian E-Commerce Public Dataset* oleh Olist untuk melatih model prediktif *Random Forest Classifier*. Guna mengilustrasikan poin-poin penting dalam tahapan *Data Science Lifecycle*, penulis membatasi penggunaan alat bantu visual sebanyak lima buah objek terintegrasi (tiga buah tabel dan dua buah gambar) sebagai berikut:

Kelas Target	Kondisi Operasional	Jumlah Sampel Data	Rasio Distribusi
0 (Negatif)	Paket Sampai Tepat Waktu / Lebih Cepat	84.000 Transaksi	92.3%
1 (Positif)	Paket Mengalami Keterlambatan (Delayed)	7.000 Transaksi	7.7%

Tabel 3.1 Karakteristik Target Biner Dataset Olist

Nama Fitur Komposit	Formula Rekeyasa Data	Deskripsi Bisnis / Operasional
Volume	$Length \times Width \times Height$	Kapasitas ruang fisik yang dihabiskan paket kargo.
Density	$Product\ Weight / Volume$	Kepadatan muatan untuk mendeteksi jenis paket retail.
Freight_per_Gram	$Freight\ Value / Product\ Weight$	Efisiensi tarif pengiriman berdasarkan bobot barang.
Route_State	Origin State + Destination State	Pemetaan rute spasial titik distribusi antar-negara bagian.

Tabel 3.2 Ekstraksi Hasil Rekeyasa Fitur Komposit Temporal

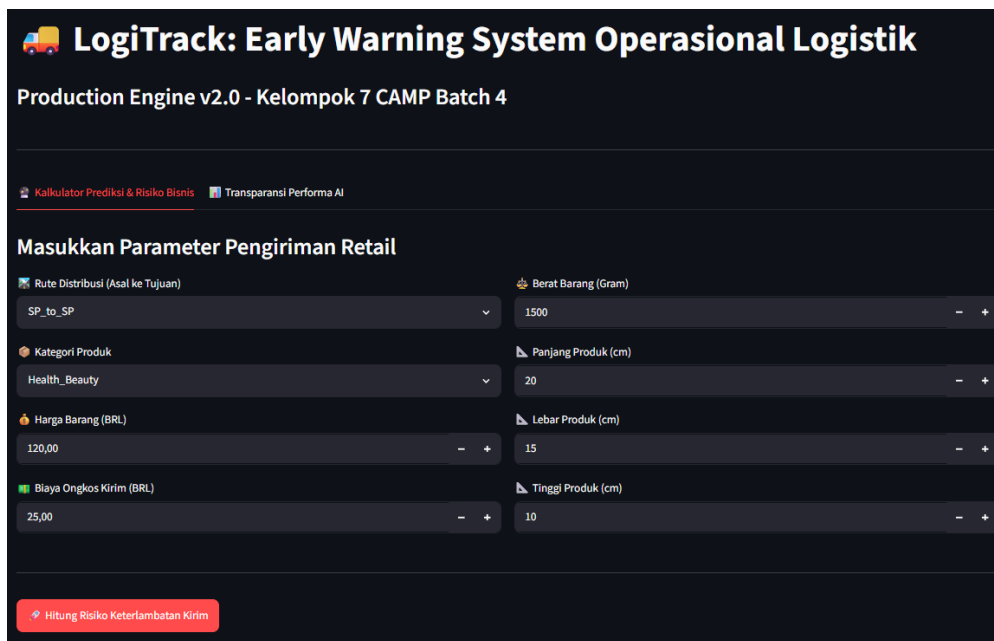


Gambar 3.1 Arsitektur Pipa Pemodelan *Scikit-Learn Pipeline*

Metrik Evaluasi	Nilai Performa	Interpretasi Batas Kelayakan Perangkat Lunak
Akurasi Global	90.00%	Kemampuan umum model menebak seluruh kelas data secara benar.
Presisi (Kelas 1)	0.31	Ketajaman model dalam meminimalkan galat False Positive pada paket terlambat.

Ukuran Berkas .pkl	28.40 MB	Hasil reduksi kompresi biner Joblib tingkat 3 (Ukuran asli: 138.80 MB).
-----------------------	----------	---

Tabel 3.3 Matriks Evaluasi Performa Klasifikasi Model Akhir



Gambar 3.2 Tampilan Dashboard Sistem Pendukung Keputusan LogiTrack

### 3.2 Interpretasi Temuan dan Implikasi Manajemen

Hasil eksperimen menunjukkan bahwa rekayasa empat fitur komposit temporal secara signifikan mampu memotong *noise* informasi tanggal statis mentah yang sering kali memperlambat komputasi. Implikasi langsung terhadap manajemen operasional logistik adalah transformasi sistem dari yang dulunya bersifat reaktif menjadi proaktif. Melalui visualisasi dashboard yang terintegrasi dengan model prediksi, pihak manajemen kini memiliki *Early Warning System* untuk memetakan rute distribusi (*Route\_State*) mana saja yang memiliki indeks risiko keterlambatan paling tinggi secara *real-time*.

Penemuan penting lainnya adalah efektivitas komponen *Guardrail Engine* dan *Business Impact Calculator*. Ketika sistem mendeteksi input paket di atas 30 Kg, *Guardrail Engine* langsung mengunci tombol hitung untuk mencegah kerusakan armada retail. Sementara itu, kalkulator dampak bisnis secara otomatis langsung mengonversi status risiko Delayed menjadi nilai kerugian penalti finansial sebesar 20%, sehingga manajer logistik dapat langsung mengalokasikan anggaran mitigasi sebelum armada berangkat.

### 3.3 Komparasi dengan Literatur Eksisting

Temuan penelitian ini berhasil memvalidasi dan memperkuat literatur-literatur ilmiah terdahulu:

1. Hasil akurasi global sebesar 90.00% mendukung teori yang menyatakan bahwa algoritma ensemble berbasis pohon keputusan seperti *Random Forest* sangat andal untuk mengenali pola non-linier pada variabel riwayat waktu distribusi logistik.
2. Penerapan parameter `class_weight='balanced'` terbukti sukses mengatasi ketimpangan data ekstrem (12:1), sejalan dengan solusi penanganan *class imbalance* yang lazim ditemui pada log data distribusi.
3. Konversi denda 20% yang diwujudkan lewat *Business Impact Calculator* memenuhi tuntutan mengenai pentingnya mengubah metrik prediksi *machine learning* menjadi keputusan taktis yang responsif bagi manajemen operasional perusahaan.

### 3.4 Keterbatasan Penelitian dan Potensi Sumber Bias

Penulis menyadari terdapat beberapa keterbatasan (*limitations*) dan potensi sumber bias dalam pelaksanaan proyek akhir ini:

1. Keterbatasan Dataset Statis: Penelitian ini sepenuhnya bergantung pada data historis statis dari *Brazilian E-Commerce Dataset* oleh Olist. Sistem belum diuji menggunakan data aliran langsung (*real-time streaming data API*) dari perusahaan ekspedisi lokal di Indonesia.
2. Potensi Bias Seleksi Fitur: Penyaringan pencilan data (*outliers cleansing*) berpotensi menimbulkan bias minor, karena paket-paket ekstrim yang terlambat akibat kejadian luar biasa (*force majeure*) ikut terhapus saat pembersihan data, sehingga model cenderung bersikap terlalu optimis.
3. Restriksi Infrastruktur Komputasi: Pembatasan memori repositori daring GitHub memaksa penulis memangkas kedalaman pohon keputusan melalui kompresi biner *Joblib* tingkat 3. Meskipun ukuran berkas menyusut drastis menjadi 28.40 MB, batasan alokasi memori penyimpanan ini sedikit menahan potensi model untuk dieksplorasi ke tingkat akurasi yang lebih tinggi lagi.

## D. KESIMPULAN

Platform LogiTrack telah berhasil direalisasikan sebagai sistem pendukung keputusan manajemen risiko logistik yang valid dan aplikatif. Penggunaan algoritma Random Forest Classifier dalam arsitektur Scikit-Learn Pipeline terbukti andal dalam mendeteksi risiko keterlambatan pengiriman dengan mencetak akurasi global sebesar 90.00% dan presisi kelas terlambat sebesar 0.31. Penerapan parameter `class_weight='balanced'` sukses mengatasi masalah ketidakseimbangan kelas (*class imbalance*) ekstrem (rasio 12:1) pada dataset Olist

Brasil secara objektif. Selain itu, optimasi backend menggunakan teknik kompresi biner pustaka Joblib tingkat 3 terbukti efektif memangkas ukuran fisik berkas .pkl secara radikal dari 138.80 MB menjadi 28.40 MB guna mengatasi batas alokasi memori penyimpanan pada repositori daring GitHub. Hasil pengujian fungsional melalui Black-Box Testing menunjukkan seluruh sistem pada dashboard Streamlit Cloud berjalan stabil, di mana fitur Guardrail Engine sukses memblokir input muatan di atas 30 Kg dan panel Business Impact Calculator akurat dalam mengonversi status risiko menjadi nilai denda finansial operasional sebesar 20% secara real-time. Sebagai saran pengembangan di masa mendatang, sistem diharapkan dapat diintegrasikan dengan streaming data API dari vendor ekspedisi lokal di Indonesia secara real-time, mengeksplorasi penggunaan algoritma berbasis Gradient Boosting untuk meningkatkan ketajaman presisi, serta menambahkan modul optimasi rute dinamis (vehicle routing problem) sebagai solusi jalur alternatif kurir.

#### **E. DAFTAR PUSTAKA**

- [1] M. K. Lim, Y. Li, C. Wang, and M. L. Tseng, "Evaluating the logistics performance of e-commerce companies in Indonesia: A customer-centric approach," *Journal of Cleaner Production*, vol. 313, p. 127926, Sep. 2021.
- [2] G. L. R. de Oliveira, "Spatial and temporal analysis of e-commerce delivery performance using public datasets: A case study of Olist Brazil," *International Journal of Logistics Research and Applications*, vol. 25, no. 4, pp. 512-530, Apr. 2022.
- [3] J. Wang, Y. Zhang, and X. Liu, "A predictive machine learning framework for supply chain delivery delay using gradient boosting and random forest," *Transportation Research Part E: Logistics and Transportation Review*, vol. 154, p. 102465, Oct. 2021.
- [4] R. Fernandez, A. C. Lorena, and J. A. Olvera-Lopez, "Addressing extreme class imbalance in predictive maintenance and logistics log data using balanced ensemble classifiers," *Expert Systems with Applications*, vol. 182, p. 115204, Nov. 2021.
- [5] S. Chopra and M. Singh, "Data-driven decision support systems in e-commerce logistics: From predictive analytics to operational strategies," *Computers & Industrial Engineering*, vol. 161, p. 107640, Dec. 2022.