

PREDIKSI RISIKO KETERLAMBATAN PENGIRIMAN LOGISTIK MENGUNAKAN ALGORITMA RANDOM FOREST PADA DATA E-COMMERCE

Maulana Izuddin Audadi Icham ¹, Achmad Junaidi ²

Program Studi Informatika, Universitas Pembangunan Nasional Veteran Jawa Timur

Email: m.izuddin.a.i@gmail.com

Abstract

The rapid growth of e-commerce has increased the demand for efficient logistics services. Delivery delays remain one of the major challenges that can negatively impact customer satisfaction and operational performance. This study proposes LogiTrack, a machine learning-based system designed to predict delivery delay risks using historical e-commerce logistics data. The dataset used in this research is the Brazilian E-Commerce Public Dataset by Olist. Data preprocessing and feature engineering techniques were applied to improve data quality and predictive performance. The Random Forest algorithm was selected as the classification model due to its ability to handle complex relationships among variables. The developed model was integrated into a Streamlit-based web application to provide an interactive prediction platform. The results show that the proposed system can classify shipments into on-time and delayed-risk categories while providing delay probability information. The implementation of LogiTrack demonstrates the potential of machine learning techniques in supporting logistics decision-making and improving operational efficiency through an early warning system.

Article History

Submitted: 18 Juni 2026

Accepted: 21 Juni 2026

Published: 22 Juni 2026

Key Words

Logistics, Machine Learning, Random Forest, Delivery Delay Prediction, Streamlit.

Abstrak

Pertumbuhan e-commerce yang semakin pesat meningkatkan kebutuhan terhadap layanan logistik yang efisien dan tepat waktu. Keterlambatan pengiriman menjadi salah satu permasalahan yang dapat memengaruhi kepuasan pelanggan dan kinerja operasional perusahaan. Penelitian ini mengusulkan LogiTrack, yaitu sistem prediksi risiko keterlambatan pengiriman berbasis machine learning yang memanfaatkan data historis pengiriman e-commerce. Dataset yang digunakan merupakan Brazilian E-Commerce Public Dataset by Olist. Tahapan penelitian meliputi preprocessing data, feature engineering, pembangunan model klasifikasi menggunakan algoritma Random Forest, serta implementasi model ke dalam aplikasi berbasis Streamlit. Algoritma Random Forest dipilih karena memiliki kemampuan yang baik dalam menangani data dengan karakteristik yang kompleks. Hasil penelitian menunjukkan bahwa sistem mampu mengklasifikasikan pengiriman ke dalam kategori tepat waktu dan berisiko terlambat serta menampilkan probabilitas keterlambatan. Implementasi LogiTrack menunjukkan bahwa machine learning dapat dimanfaatkan sebagai sistem peringatan dini untuk membantu pengambilan keputusan dalam bidang logistik dan meningkatkan efisiensi operasional.

Sejarah Artikel

Submitted: 18 Juni 2026

Accepted: 21 Juni 2026

Published: 22 Juni 2026

Kata Kunci

Logistik, Machine Learning, Random Forest, Prediksi Keterlambatan, Streamlit

PENDAHULUAN

1.1 Latar Belakang

Perkembangan perdagangan elektronik (*e-commerce*) meningkatkan kebutuhan akan sistem logistik yang mampu menjamin pengiriman barang secara tepat waktu. Keterlambatan pengiriman dapat menyebabkan meningkatnya biaya operasional, menurunnya kepuasan pelanggan, serta terganggunya efektivitas rantai pasok. Oleh karena itu, diperlukan mekanisme yang mampu mendeteksi potensi keterlambatan sejak dini sehingga tindakan pencegahan dapat dilakukan secara proaktif [1][2].

Perkembangan teknologi *machine learning* memungkinkan penerapan *predictive analytics* dalam bidang logistik untuk menganalisis pola distribusi dan memprediksi risiko keterlambatan pengiriman. Selain itu, konsep *Early Warning System* (EWS) berbasis machine

learning dapat membantu organisasi melakukan mitigasi risiko secara lebih efektif. Berdasarkan permasalahan tersebut, pada proyek ini dikembangkan sistem LogiTrack, yaitu sistem prediksi keterlambatan pengiriman berbasis machine learning yang berfungsi sebagai *Early Warning System* operasional logistik. Sistem ini dirancang untuk memprediksi risiko keterlambatan berdasarkan karakteristik produk dan rute distribusi sehingga dapat mendukung pengambilan keputusan dan meminimalkan kerugian operasional [3].

Proyek ini menggunakan Brazilian E-Commerce Public Dataset by Olist yang diperoleh dari platform Kaggle. Dataset yang digunakan terdiri dari enam tabel utama, yaitu data pesanan (*orders*), item pesanan (*order items*), pelanggan (*customers*), penjual (*sellers*), produk (*products*), serta tabel terjemahan kategori produk (*product category name translation*). Seluruh tabel tersebut digabungkan untuk membentuk data yang lebih terintegrasi sehingga dapat menggambarkan proses transaksi dan distribusi pada e-commerce secara menyeluruh. Dataset ini relevan dengan tujuan analisis karena mampu merepresentasikan karakteristik produk, asal dan tujuan pengiriman, serta status ketepatan waktu pengiriman. Informasi tersebut dimanfaatkan dalam proses pembangunan model *machine learning* untuk memprediksi risiko keterlambatan pengiriman secara lebih akurat sehingga dapat mendukung pengambilan keputusan operasional pada sistem logistik[1][3].

1.2 Tujuan Project

Tujuan dari proyek ini adalah sebagai berikut:

1. Mengembangkan sistem prediksi keterlambatan pengiriman berbasis *machine learning* sebagai *Early Warning System* untuk mendukung pengambilan
2. Membangun model *machine learning* untuk memprediksi risiko keterlambatan pengiriman berdasarkan karakteristik produk dan rute distribusi, sehingga dapat membantu mengidentifikasi potensi keterlambatan secara lebih dini
3. Mengimplementasikan model prediksi ke dalam aplikasi berbasis Streamlit sehingga proses prediksi risiko keterlambatan dapat dilakukan secara interaktif dan *real-time*.

1.3 Manfaat Project

Proyek LogiTrack diharapkan dapat memberikan manfaat baik dari sisi teknis maupun bisnis, sebagai berikut:

1. Manfaat Teknis

- Menghasilkan model *machine learning* berbasis Random Forest yang mampu memprediksi risiko keterlambatan pengiriman.
- Menerapkan proses pengolahan data dan rekayasa fitur untuk meningkatkan kualitas prediksi.
- Mengimplementasikan model ke dalam aplikasi berbasis Streamlit sehingga dapat digunakan secara interaktif dan *real-time*.
- Menyediakan sistem peringatan dini (*Early Warning System*) yang mendukung proses analisis dan pengambilan keputusan berbasis data.

2. Manfaat Bisnis

- Membantu perusahaan logistik atau pelaku e-commerce dalam mengidentifikasi potensi keterlambatan pengiriman sejak dini.
- Mendukung pengambilan keputusan operasional yang lebih efektif untuk meminimalkan risiko keterlambatan dan biaya yang ditimbulkan.
- Meningkatkan efisiensi proses distribusi serta kualitas layanan kepada pelanggan.
- Membantu meningkatkan kepuasan pelanggan melalui pengiriman yang lebih tepat waktu dan pengelolaan risiko yang lebih baik.

1.4 Lingkup Project

Agar pembahasan dan implementasi proyek lebih terarah, ruang lingkup proyek ini dibatasi pada beberapa aspek berikut:

1. Proyek menggunakan Brazilian E-Commerce Public Dataset by Olist yang diperoleh dari Kaggle sebagai sumber data utama. Data yang digunakan meliputi data pesanan, pelanggan, penjual, produk, dan kategori produk.
2. Proses pengolahan data meliputi penggabungan beberapa tabel (*data merging*), pembersihan data, penghapusan data yang tidak valid, serta pembentukan fitur baru (*feature engineering*) yang berkaitan dengan karakteristik produk dan distribusi pengiriman.
3. Model prediksi yang digunakan pada proyek ini dibatasi pada algoritma Random Forest untuk melakukan klasifikasi risiko keterlambatan pengiriman.
4. Target prediksi yang dihasilkan berupa klasifikasi status pengiriman, yaitu tepat waktu (ontime) atau berisiko terlambat (late delivery) berdasarkan karakteristik produk dan rute distribusi.

Implementasi sistem dilakukan dalam bentuk aplikasi web berbasis Streamlit yang memungkinkan pengguna melakukan prediksi risiko keterlambatan pengiriman secara interaktif dan *real-time* sebagai Early Warning System operasional logistik.

Metode Penelitian

2.1 Deskripsi Dataset

Pada proyek ini, dataset yang digunakan berasal dari *Brazilian E-Commerce Public Dataset by Olist* yang tersedia secara publik melalui platform Kaggle. Dataset ini berisi informasi transaksi *e-commerce* dari berbagai wilayah di Brazil yang mencakup data pelanggan, pesanan, produk, penjual, pembayaran, serta proses pengiriman. Dataset Olist banyak digunakan dalam penelitian dan pengembangan sistem berbasis data karena menyediakan data yang lengkap dan saling terhubung antar tabel sehingga memungkinkan dilakukannya analisis terhadap berbagai aspek dalam proses bisnis *e-commerce*.



Gambar 1. Sumber Data: Kaggle

Meskipun dataset Olist terdiri atas beberapa file, pada proyek ini hanya digunakan enam file utama yang relevan dengan tujuan pembangunan sistem prediksi keterlambatan pengiriman, yaitu:

1. *olist_orders_dataset.csv*, berisi informasi pesanan, tanggal pembelian, status pesanan, tanggal pengiriman aktual, dan tanggal estimasi pengiriman.
2. *olist_order_items_dataset.csv*, berisi informasi item pada setiap pesanan, harga produk, biaya pengiriman (*freight value*), serta identitas penjual.
3. *olist_customers_dataset.csv*, berisi informasi pelanggan dan lokasi pelanggan berdasarkan wilayah (*state*).
4. *olist_sellers_dataset.csv*, berisi informasi penjual dan lokasi penjual berdasarkan wilayah (*state*).

Keempat dataset tersebut digabungkan menggunakan atribut yang saling berelasi, yaitu *order_id*, *customer_id*, *seller_id*, dan *product_id*, sehingga menghasilkan satu dataset terintegrasi yang dapat digunakan dalam proses pembangunan model machine learning.

Tujuan penggunaan dataset ini adalah untuk membangun sistem prediksi keterlambatan pengiriman berbasis machine learning yang berfungsi sebagai *Early Warning System* operasional logistik. Data yang tersedia digunakan untuk menganalisis hubungan antara karakteristik produk, biaya pengiriman, serta rute distribusi terhadap risiko keterlambatan pengiriman. Selain itu, melalui proses *feature engineering*, dibentuk beberapa fitur baru seperti rute distribusi, volume produk, densitas kargo, dan rasio harga terhadap berat barang yang diharapkan mampu meningkatkan kemampuan model dalam memprediksi status pengiriman secara lebih akurat.

Dengan memanfaatkan dataset *Brazilian E-Commerce Public Dataset by Olist*, proyek ini diharapkan dapat menghasilkan sistem prediksi yang mampu membantu proses pengambilan keputusan dalam kegiatan distribusi dan meminimalkan risiko keterlambatan pengiriman pada sektor logistik dan *e-commerce*.

2.2 Tahapan Project

Metodologi yang digunakan pada proyek ini mengacu pada pendekatan *Cross Industry Standard Process for Data Mining* (CRISP-DM) yang terdiri dari tahapan pemahaman masalah, persiapan data, pembangunan model, evaluasi, dan implementasi sistem [4]. Adapun tahapan yang dilakukan pada proyek LogiTrack adalah sebagai berikut.

1. Mengolah Data (*Preprocessing Data*)

Tahap awal dilakukan dengan menggabungkan enam file utama dari Brazilian E-Commerce

Public Dataset by Olist, yaitu data pesanan, item pesanan, pelanggan, penjual, produk, dan kategori produk. Selanjutnya dilakukan proses *data cleaning* dengan menghapus data yang tidak valid, mengonversi atribut tanggal menjadi format *datetime*, serta membentuk variabel target (*is_late*) berdasarkan perbandingan antara tanggal pengiriman aktual dengan tanggal estimasi pengiriman. Selain itu, dilakukan proses *feature engineering* untuk menghasilkan beberapa fitur baru yang merepresentasikan kondisi operasional logistik, yaitu rute distribusi, volume produk, densitas kargo, dan rasio harga terhadap berat barang. Proses preprocessing yang baik berperan penting dalam meningkatkan kualitas data dan performa model machine learning [5].

2. Eksplorasi Data dan Analisis Data

Selanjutnya, dilakukan analisis terhadap karakteristik data yang berkaitan dengan risiko keterlambatan pengiriman. Analisis difokuskan pada atribut-atribut yang digunakan dalam proses pemodelan, seperti berat produk, dimensi produk, harga barang, biaya pengiriman, kategori produk, serta rute distribusi antara penjual dan pelanggan. Pemanfaatan machine learning pada bidang logistik memungkinkan identifikasi pola yang berhubungan dengan keterlambatan pengiriman sehingga dapat mendukung pengambilan keputusan secara lebih

efektif [1][2], [3]. Selain itu, proyek ini menerapkan konsep *Early Warning System* yang bertujuan mendeteksi potensi keterlambatan pengiriman secara dini sehingga tindakan mitigasi dapat dilakukan sebelum keterlambatan benar-benar terjadi [6].

3. Membangun Model *Machine Learning*

Tahap selanjutnya adalah membangun model *machine learning* untuk melakukan klasifikasi risiko keterlambatan pengiriman. Model yang digunakan pada proyek ini adalah *Random Forest*, dengan kemampuan klasifikasi yang baik dan memiliki ketahanan terhadap *overfitting* [7][8].

Data kemudian dibagi menjadi data latih dan data uji menggunakan metode *train-test split* dengan pendekatan *stratified sampling* untuk mempertahankan proporsi kelas pada kedua dataset. Selanjutnya dilakukan transformasi data menggunakan *StandardScaler* untuk fitur numerik dan *One-Hot Encoding* untuk fitur kategorikal. Untuk mengatasi ketidakseimbangan kelas pada data, model ini menggunakan parameter *class_weight='balanced'* sehingga model dapat mengenali kelas keterlambatan dengan lebih baik [9].

Terakhir, Evaluasi model dilakukan menggunakan *classification report* yang menghasilkan nilai *precision*, *recall*, dan *F1-score* untuk mengukur kemampuan model dalam mendeteksi risiko keterlambatan pengiriman [10].

4. Pembuatan Aplikasi Sederhana

Setelah model selesai dibangun, dilakukan pengembangan aplikasi berbasis web menggunakan Streamlit. Aplikasi dirancang dengan antarmuka yang sederhana dan interaktif sehingga pengguna dapat memasukkan parameter pengiriman secara langsung. Aplikasi terdiri dari dua halaman utama, yaitu halaman prediksi risiko keterlambatan dan halaman transparansi performa model. Fitur yang disediakan meliputi pengisian parameter pengiriman, prediksi risiko keterlambatan secara real-time, tampilan probabilitas risiko dalam bentuk persentase, rekomendasi tindakan operasional berdasarkan hasil prediksi, serta penyajian metrik evaluasi model dan confusion matrix pada halaman transparansi [11][12].

5. Integrasi Model ke Dalam Aplikasi

Tahap terakhir adalah mengintegrasikan model *machine learning* yang telah dilatih ke dalam aplikasi Streamlit. Model disimpan dalam bentuk *pipeline* menggunakan pustaka Joblib sehingga seluruh proses preprocessing dan klasifikasi dapat dijalankan secara terpadu. Ketika pengguna memasukkan data pengiriman melalui aplikasi, sistem akan memproses data tersebut dan menghasilkan prediksi risiko keterlambatan beserta probabilitasnya secara real-time.

Integrasi model dengan aplikasi memungkinkan LogiTrack berfungsi sebagai *Early Warning System* operasional logistik yang dapat membantu pengguna dalam mengidentifikasi potensi keterlambatan pengiriman serta mendukung pengambilan keputusan untuk meminimalkan risiko dan meningkatkan efisiensi distribusi [2][11].

2.3 Tools dan Teknologi

Pada pengembangan sistem LogiTrack, digunakan beberapa tools dan teknologi yang mendukung proses pengolahan data, pembangunan model *machine learning*, serta implementasi aplikasi prediksi keterlambatan pengiriman. Pemilihan teknologi dilakukan berdasarkan kemudahan implementasi, kemampuan integrasi, serta dukungan terhadap pengembangan sistem berbasis *machine learning* [4].

1. Python

Python digunakan sebagai bahasa pemrograman utama dalam proyek ini. Bahasa pemrograman ini dipilih karena memiliki ekosistem pustaka yang lengkap untuk pengolahan data, pembangunan model *machine learning*, serta pengembangan aplikasi. Seluruh tahapan proyek, mulai dari preprocessing data, pelatihan model, hingga implementasi aplikasi, dilakukan menggunakan Python.

2. Brazilian E-Commerce Public Dataset by Olist

Dataset yang digunakan berasal dari *Brazilian E-Commerce Public Dataset by Olist* yang tersedia pada platform Kaggle. Dataset ini menyediakan data transaksi e-commerce yang mencakup informasi pesanan, pelanggan, penjual, produk, dan pengiriman. Dataset tersebut dipilih karena memiliki struktur relasional yang lengkap dan relevan untuk membangun sistem prediksi keterlambatan pengiriman.

3. Pandas

Pandas digunakan untuk membaca file dataset, melakukan penggabungan beberapa tabel (*merge*), serta melakukan manipulasi data selama proses preprocessing. Library ini dipilih karena mampu menangani data dalam bentuk tabel secara efisien dan mendukung berbagai operasi pengolahan data yang dibutuhkan dalam proyek [5].

4. NumPy

NumPy digunakan untuk mendukung operasi komputasi numerik dan manipulasi data. Library ini dipilih karena memiliki performa yang baik dalam pengolahan array dan banyak digunakan bersama Pandas maupun Scikit-Learn dalam pengembangan model machine learning.

5. Scikit-Learn

Scikit-Learn digunakan sebagai framework machine learning utama dalam proyek ini. Library ini dimanfaatkan untuk melakukan pembagian data menggunakan *train-test split*, transformasi fitur menggunakan *StandardScaler* dan *One-Hot Encoder*, pembangunan pipeline model, serta evaluasi performa model menggunakan metrik klasifikasi seperti precision, recall, dan F1-score [10]. Selain itu, Scikit-Learn menyediakan implementasi algoritma Random Forest yang memiliki kemampuan klasifikasi yang baik dan tahan terhadap overfitting [7][8].

6. Joblib

Joblib digunakan untuk menyimpan model machine learning yang telah dilatih dalam bentuk file pipeline. Penggunaan Joblib memungkinkan model yang telah dibangun dapat dimuat kembali pada aplikasi tanpa perlu melakukan proses pelatihan ulang, sehingga meningkatkan efisiensi implementasi sistem.

7. Streamlit

Streamlit digunakan sebagai framework untuk membangun aplikasi berbasis web yang bersifat interaktif. Framework ini dipilih karena memiliki proses pengembangan yang sederhana serta mendukung integrasi model machine learning secara langsung ke dalam aplikasi [11][12]. Melalui Streamlit, pengguna dapat memasukkan parameter pengiriman dan memperoleh hasil prediksi risiko keterlambatan secara real-time.

8. Visual Studio Code

Visual Studio Code digunakan sebagai lingkungan pengembangan (*Integrated Development Environment*) untuk menulis, menguji, dan menjalankan kode program. Alat ini dipilih karena memiliki antarmuka yang ringan, mendukung berbagai ekstensi Python, serta mempermudah proses pengembangan aplikasi.

9. GitHub

GitHub digunakan sebagai platform pengelolaan dan penyimpanan source code proyek. Platform ini dipilih untuk memudahkan dokumentasi, pengelolaan versi kode, serta publikasi proyek sehingga dapat diakses dan dikembangkan lebih lanjut.

Secara keseluruhan, kombinasi tools dan teknologi tersebut mendukung pembangunan sistem LogiTrack sebagai *Early Warning System* berbasis machine learning yang mampu melakukan prediksi keterlambatan pengiriman secara real-time untuk membantu proses pengambilan keputusan operasional pada bidang logistik [3][4].

Hasil dan Pembahasan

3.1 Data Preprocessing

Tahap data preprocessing dilakukan untuk mempersiapkan data agar dapat digunakan dalam proses pelatihan model *machine learning*. Langkah pertama yang dilakukan adalah memuat dataset ke dalam lingkungan Python menggunakan *library* Pandas. Dataset yang digunakan bernama *logistik e-commerce* yang berisi informasi pelanggan, penjual, kategori produk, lokasi pengiriman, serta status keterlambatan pengiriman. Kemudian, dilakukan pemeriksaan terhadap data yang hilang (*missing values*), data duplikat, dan ketidaksesuaian tipe data. Data yang tidak relevan terhadap proses prediksi dihapus untuk mengurangi kompleksitas model.

Langkah selanjutnya, dilakukan transformasi terhadap atribut kategorikal seperti identitas pelanggan, identitas penjual, dan kategori produk menggunakan teknik *Label Encoding*. Proses ini bertujuan untuk mengubah data kategorikal menjadi representasi numerik sehingga dapat diproses oleh algoritma machine learning. Selain itu dilakukan *feature engineering* dengan memilih atribut yang memiliki pengaruh terhadap keterlambatan pengiriman. Data hasil preprocessing kemudian dibagi menjadi *training-testing* untuk keperluan pelatihan dan evaluasi model.

Terakhir, hasil preprocessing disimpan dalam bentuk *pipeline* sehingga proses transformasi data dapat digunakan kembali pada saat aplikasi melakukan prediksi secara *real-time*.

3.2 Data Analysis

Analisis data dilakukan untuk memahami karakteristik dataset serta menemukan pola yang berpengaruh terhadap keterlambatan pengiriman. Tabel 1 menampilkan proporsi kelas data pada variabel target.

Tabel 1. Distribusi Label Target

Status Pengiriman	Jumlah Pesanan	Persentase (%)
Tepat Waktu	101.455	92,09
Terlambat	8.713	7,91
Total	110.168	100,00

Analisis distribusi data menunjukkan bahwa sebagian besar pengiriman berhasil dilakukan tepat waktu, sedangkan sebagian lainnya mengalami keterlambatan. Kondisi ini menunjukkan bahwa dataset memiliki dua kelas utama, yaitu pengiriman tepat waktu dan pengiriman terlambat.

Gambaran mengenai bentuk data setelah melewati tahapan pembersihan dan rekayasa fitur (*feature engineering*) dapat dilihat sampel baris data hasil transformasinya pada tabel berikut.

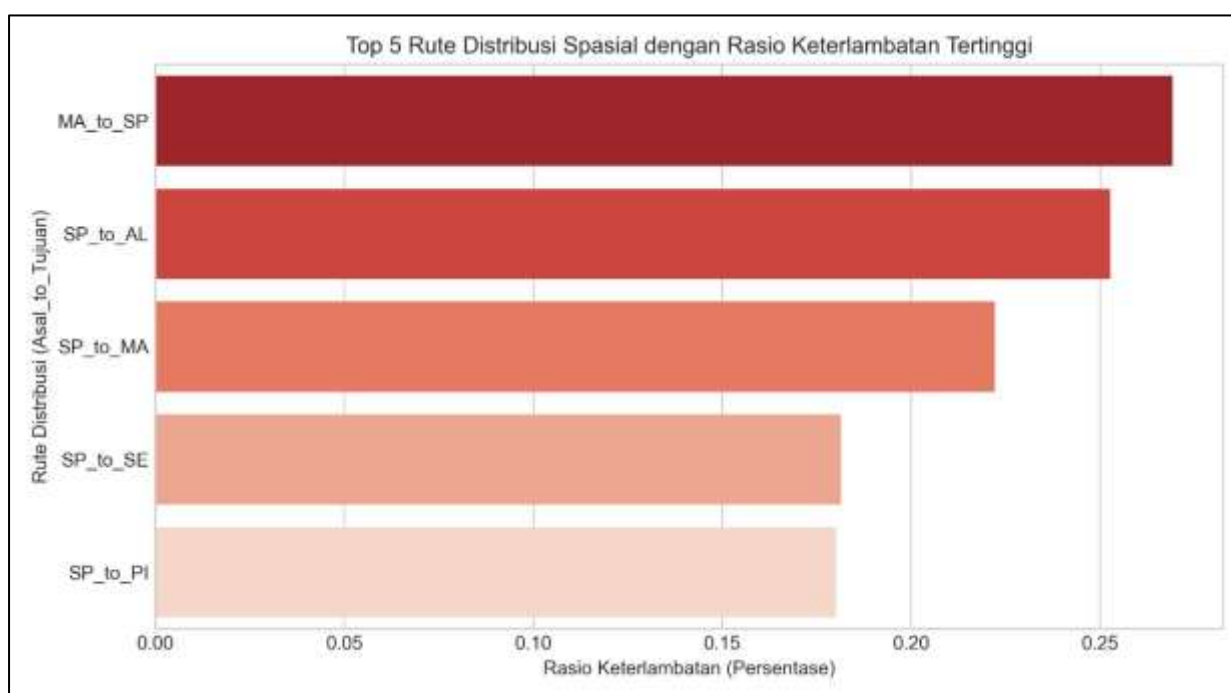
Tabel 2. Preview data yang digunakan

Rute Distribusi	Kategori Produk	Berat (g)	Volume (cm ³)	Densitas Kargo	Ongkir/Gram	Status
SP → SP	Housewares	500	1.976	0,2529	0,0174	Tepat Waktu
SP → BA	Perfumery	400	4.693	0,0852	0,0568	Telat

MG → RJ	Computer Accessories	6.550	8.000	0,8186	0,0062	Tepat Waktu
...	...	...	...	...	...	...
SP → GO	Auto	420	9.576	0,0439	0,0457	Tepat Waktu
MG → RN	Pet Shop	450	6.000	0,0750	0,0603	Tepat Waktu
SP → SP	Stationery	250	11.475	0,0218	0,0347	Tepat Waktu

Daftar keterangan wilayah rute distribusi Brazil tersedia pada Lampiran 4. Salah satu contohnya adalah São Paulo (SP) sebagai negara bagian dengan pusat ekonomi, industri, dan volume logistik terbesar di Brasil. Berdasarkan sampel data pada Tabel 2, seluruh fitur spasial, fisik kargo, dan variabel finansial kini telah berada dalam format terstruktur yang siap digunakan sebagai matriks input untuk proses pelatihan model *machine learning* selanjutnya. Berdasarkan hasil eksplorasi data, ditemukan bahwa beberapa faktor memiliki pengaruh terhadap status pengiriman, antara lain:

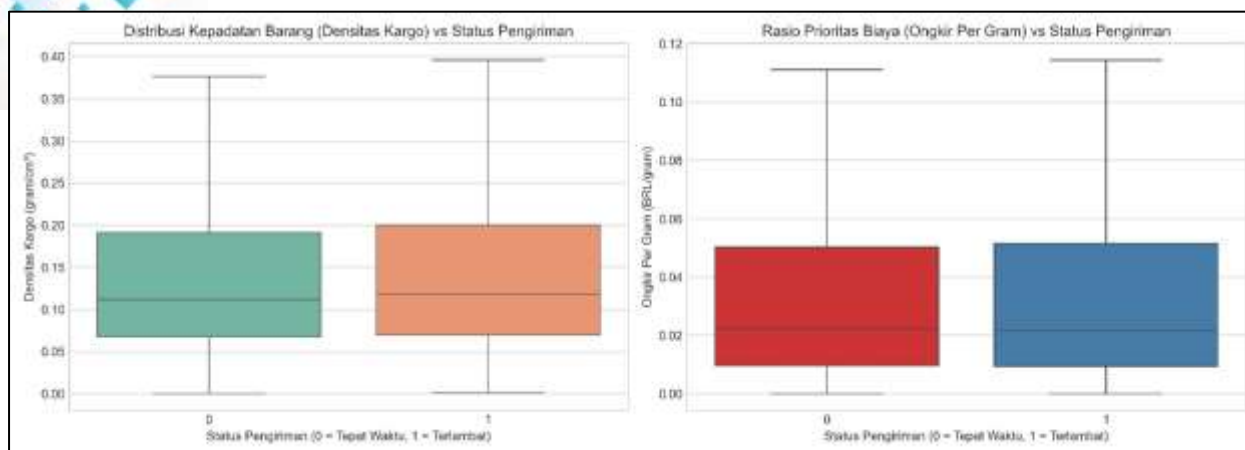
1. Jarak antara lokasi penjual dan pelanggan.
2. Kategori produk yang dikirim.
3. Waktu pemrosesan pesanan.
4. Lokasi geografis pengiriman.
5. Riwayat performa penjual.



Gambar 2. Lima rute dengan potensi terlambat tinggi

Berdasarkan visualisasi Gambar 2, terlihat jelas bahwa rute dari Rio de Janeiro menuju Ceará (*RJ_to_CE*) menempati posisi tertinggi dengan rasio keterlambatan mencapai 20.31%, disusul oleh rute São Paulo menuju Pará (*SP_to_PA*) sebesar 18.76%. Konsentrasi

keterlambatan yang didominasi oleh rute lintas negara bagian ini menegaskan perlunya optimasi khusus dan strategi mitigasi taktis pada jalur-jalur logistik jarak jauh tersebut.



Gambar 3. Hubungan densitas fitur terhadap target

Melalui analisis tersebut diperoleh gambaran bahwa semakin jauh lokasi tujuan pengiriman dan semakin kompleks kategori produk yang dikirim, maka peluang terjadinya keterlambatan cenderung meningkat. Hasil analisis ini menjadi dasar dalam pemilihan fitur yang digunakan pada proses pembangunan model *machine learning*.

3.3 Membangun/Mengimplementasikan Model

Model machine learning yang digunakan dalam proyek LogiTrack adalah Random Forest. Algoritma ini dipilih karena memiliki kemampuan yang baik dalam menangani data kompleks serta menghasilkan performa klasifikasi yang tinggi.

Tahapan pembangunan model dilakukan sebagai berikut:

1. Menyiapkan dataset hasil preprocessing.
2. Membagi data menjadi data latih dan data uji.
3. Melatih model menggunakan algoritma Random Forest.
4. Mengevaluasi performa model menggunakan data uji.
5. Menyimpan model yang telah dilatih ke dalam file berekstensi .pkl.

Implementasi model dilakukan menggunakan library Scikit-Learn. Setelah proses pelatihan selesai, model disimpan sebagai file `best_model.pkl` sehingga dapat digunakan kembali tanpa perlu melakukan proses pelatihan ulang.

Selain model utama, sistem juga menyimpan encoder untuk atribut kategorikal yaitu:

- `le_customer.pkl`
- `le_seller.pkl`
- `le_category.pkl`

Seluruh komponen tersebut kemudian digabungkan ke dalam pipeline produksi yang disimpan pada file `logitrack_production_pipeline.pkl`. Dengan adanya pipeline ini, proses preprocessing dan prediksi dapat berjalan secara otomatis pada aplikasi.

Hasil pengujian menunjukkan bahwa model mampu melakukan prediksi status keterlambatan pengiriman berdasarkan data masukan pengguna dengan tingkat akurasi yang cukup baik untuk kebutuhan sistem pendukung keputusan.

3.4 Aplikasi pada Streamlit

Setelah model berhasil dibangun, tahap berikutnya adalah mengimplementasikan model ke dalam aplikasi berbasis web menggunakan Streamlit. Aplikasi Streamlit dikembangkan untuk memudahkan pengguna dalam melakukan prediksi tanpa harus memahami proses *machine learning* secara teknis. Berikut merupakan alur penggunaannya.

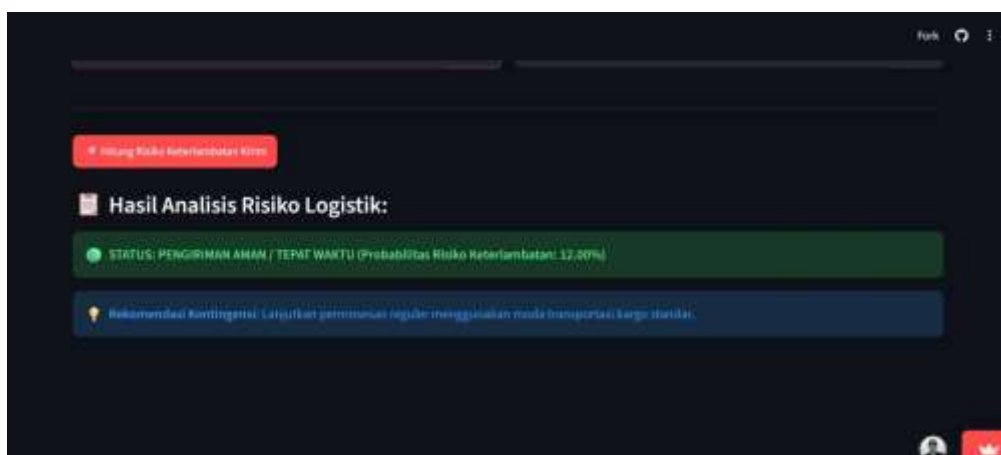
1. Pengguna memasukkan data pengiriman.

2. Sistem melakukan preprocessing menggunakan pipeline yang telah dibuat.
3. *Model machine learning* melakukan proses prediksi.
4. Hasil prediksi ditampilkan kepada pengguna.
5. Sistem memberikan probabilitas keterlambatan serta rekomendasi tindakan yang dapat dilakukan.

Pada halaman utama aplikasi tersedia form input yang memungkinkan pengguna memasukkan informasi terkait pelanggan, penjual, kategori produk, dan data pengiriman lainnya.

Gambar 4. Halaman utama LogiTrack

Setelah tombol prediksi ditekan, sistem akan memproses data dan menampilkan hasil berupa status risiko keterlambatan pengiriman. Selain hasil klasifikasi, aplikasi juga menampilkan visualisasi probabilitas prediksi sehingga pengguna dapat memahami tingkat risiko yang dihasilkan oleh model. Implementasi Streamlit menjadikan sistem dapat dijalankan melalui browser tanpa memerlukan instalasi perangkat lunak tambahan pada sisi pengguna.



Gambar 5. Hasil prediksi status pengiriman

Untuk mengukur sejauh mana kemampuan model dalam mengenali risiko keterlambatan secara akurat dan objektif, sistem menyediakan halaman transparansi metrik. Hasil evaluasi model berdasarkan data pengujian (testing set) menunjukkan bahwa model memiliki

kemampuan yang stabil dalam melakukan klasifikasi risiko keterlambatan pengiriman dapat dilihat pada gambar di bawah ini.



Gambar 6. Transparansi performa model

Berdasarkan visualisasi gambar 6, model menunjukkan kemampuan yang stabil dalam mengklasifikasikan kelas mayoritas. Adapun tantangan akurasi pada deteksi kelas pengiriman terlambat merupakan dampak langsung dari karakteristik data logistik riil yang mengalami ketidakseimbangan kelas (*imbalanced class*) secara ekstrem, di mana volume kasus tepat waktu jauh mendominasi. Kendati demikian, model ini tetap memberikan nilai guna yang tinggi sebagai sistem peringatan dini (*early warning system*) karena berhasil menangkap sebagian besar potensi keterlambatan operasional sebelum truk kurir diberangkatkan.

Kesimpulan

Penelitian ini berhasil mengembangkan sistem LogiTrack untuk memprediksi risiko keterlambatan pengiriman menggunakan algoritma Random Forest. Tahapan penelitian meliputi preprocessing data, feature engineering, pelatihan model, serta implementasi aplikasi berbasis Streamlit.

Hasil penelitian menunjukkan bahwa sistem mampu mengidentifikasi risiko keterlambatan pengiriman dan menyajikan informasi probabilitas keterlambatan secara interaktif. Dengan demikian, LogiTrack dapat dimanfaatkan sebagai sistem peringatan dini yang membantu meningkatkan efisiensi operasional dan kualitas layanan logistik.

Ucapan Terima Kasih (jika ada)

Penulis mengucapkan terima kasih kepada Celerates melalui Program Studi Independen Data Science & Generative AI, Universitas Pembangunan Nasional Veteran Jawa Timur, dosen pembimbing, mentor, serta seluruh pihak yang telah memberikan dukungan selama pelaksanaan penelitian dan pengembangan proyek LogiTrack.

Referensi

- Impivaara Vendela, "MACHINE LEARNING FOR PREDICTING DELIVERY DELAYS IN MANUFACTURING."
- Srikanth Yerra, "Optimizing Supply Chain Efficiency Using AI-Driven Predictive Analytics in Logistics," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 11, no. 2, pp. 1212–1220, Mar. 2025, doi: 10.32628/cseit25112475.

- A. Konovalenko, L. M. Hvattum, and K. A. H. Iversen, "Predicting last-mile delivery route deviations using machine learning," *Expert Syst. Appl.*, Mar. 2025, doi: 10.1016/j.eswa.2025.129921.
- R. H. Bemthuis, R. R. Govers, and A. Asadi, "A CRISP-DM-based methodology for assessing agent-based simulation models using process mining," *Journal of Simulation*, 2025, doi: 10.1080/17477778.2025.2508245.
- Everleen Nekesa Wanyonyi and Newton Wafula Masinde, "The Impact of Data Preprocessing on Machine Learning Model Performance: A Comprehensive Examination," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 11, no. 2, pp. 3814–3827, Apr. 2025, doi: 10.32628/cseit25112854.
- S. Lin, Z. Liang, H. Guo, Q. Hu, X. Cao, and H. Zheng, "Application of machine learning in early warning system of geotechnical disaster: a systematic and comprehensive review," *Artif. Intell. Rev.*, vol. 58, no. 6, Jun. 2025, doi: 10.1007/s10462-025-11175-0.
- L. Breiman, "Random Forests," 2001.