DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

BAB 1 PENDAHULUAN

1.1 Latar Belakang

Dalam era digital, industri *e-commerce* berkembang pesat dan semakin menjadi bagian penting dari kehidupan sehari-hari. Peningkatan volume transaksi online menuntut sistem logistik yang lebih efisien dan akurat dalam mengelola pengiriman barang. Salah satu tantangan utama dalam rantai pasok *e-commerce* adalah ketepatan waktu pengiriman. Ketepatan waktu ini tidak hanya memengaruhi kepuasan pelanggan tetapi juga berdampak pada biaya operasional dan efisiensi keseluruhan bisnis.

Berbagai faktor dapat mempengaruhi durasi pengiriman, termasuk metode pengiriman yang dipilih, lokasi pesanan, jumlah barang dalam pesanan, serta faktor eksternal seperti kondisi lalu lintas dan cuaca. Dalam beberapa kasus, keterlambatan pengiriman dapat menyebabkan ketidakpuasan pelanggan, penurunan reputasi *platform e-commerce*, dan bahkan hilangnya pelanggan potensial. Oleh karena itu, perusahaan e-commerce harus menerapkan strategi yang tepat untuk meningkatkan efisiensi pengiriman agar tetap kompetitif di pasar yang semakin ketat.

Seiring dengan perkembangan teknologi, pemanfaatan *Big Data* dan *Machine Learning* menjadi solusi yang menjanjikan untuk meningkatkan akurasi prediksi waktu pengiriman. Dengan analisis data historis, model *Machine Learning* dapat mengidentifikasi pola dalam pengiriman, mengoptimalkan proses logistik, dan memberikan estimasi yang lebih akurat kepada pelanggan. Penerapan model prediktif berbasis *Machine Learning* tidak hanya dapat meningkatkan efisiensi operasional perusahaan, tetapi juga memberikan pengalaman yang lebih baik kepada pelanggan dengan menyediakan estimasi pengiriman yang lebih presisi.

Selain itu, pemanfaatan data dalam skala besar juga memungkinkan perusahaan untuk mengidentifikasi hambatan dalam rantai pasok dan mengambil keputusan yang lebih tepat untuk mengatasinya. Dengan kombinasi strategi logistik yang lebih baik dan pemanfaatan teknologi canggih, perusahaan dapat meningkatkan efektivitas operasional dan mengurangi risiko keterlambatan pengiriman.

Analisis ini bertujuan untuk mengembangkan model prediksi waktu pengiriman berbasis *Machine Learning* dengan menganalisis berbagai faktor yang mempengaruhi durasi pengiriman. Dengan menggunakan pendekatan berbasis data, analisis ini diharapkan dapat memberikan kontribusi dalam peningkatan efisiensi rantai pasok di *industri e-commerce* serta membantu perusahaan dalam mengoptimalkan strategi logistik mereka.

1.2 Tujuan Analisis

Analisis ini memiliki beberapa tujuan utama sebagai berikut:

1. Mengembangkan model prediksi waktu pengiriman menggunakan algoritma *Machine Learning* berdasarkan faktor-faktor yang mempengaruhi durasi pengiriman, seperti metode pengiriman, lokasi pesanan, dan jumlah barang dalam pesanan.
2. Menganalisis fitur-fitur yang paling berpengaruh terhadap waktu pengiriman guna memahami faktor utama yang menentukan estimasi pengiriman.
3. Mengevaluasi performa berbagai model *Machine Learning* untuk menentukan model terbaik dalam memprediksi durasi pengiriman dengan akurasi tinggi.
4. Menyediakan solusi berbasis data yang dapat diimplementasikan dalam sistem manajemen rantai pasok *e-commerce* guna meningkatkan efisiensi operasional dan kepuasan pelanggan.
5. Mengidentifikasi potensi peningkatan dalam sistem logistik berdasarkan hasil analisis prediksi untuk memberikan rekomendasi strategis kepada perusahaan *e-commerce* dalam mengoptimalkan pengiriman.

Untuk mencapai tujuan ini, analisis dilakukan melalui eksplorasi dataset, pemrosesan data, pemilihan model terbaik, serta analisis hasil dan implementasi solusi yang diperoleh dari model *Machine Learning.*

BAB 2 EKSPLORASI DATA

2.1 Deskripsi *Dataset*

Dataset yang digunakan adalah DataCo Smart Suppy Chain, yang berisi informasi terkait proses rantai pasok dalam industri *e-commerce*, mencakup:

1. Informasi pesanan: waktu pemesanan, waktu pengiriman, dan metode pengiriman.
2. Detail produk: kategori produk, harga, dan jumlah item dalam pesanan.
3. Data geografis: lokasi pemesanan dan tujuan pengiriman.

Dataset ini terdiri dari 180.519 entri dan memiliki 53 fitur yang mencakup berbagai aspek operasional dalam rantai pasok *e-commerce*. Namun, tidak semua fitur relevan untuk analisis ini. Oleh karena itu, dilakukan proses seleksi fitur untuk memastikan bahwa hanya data yang memiliki pengaruh signifikan terhadap waktu pengiriman yang digunakan dalam analisis lebih lanjut.

Sebelum melakukan analisis lebih lanjut, data dalam *dataset* ini melalui beberapa tahap *preprocessing*:

1. Menghapus fitur yang tidak relevan

   Beberapa fitur seperti *email* pelanggan, kata sandi, dan ID transaksi tidak memiliki kontribusi terhadap prediksi waktu pengiriman sehingga dihapus dari *dataset*.
2. Menangani data yang hilang (*missing values*)

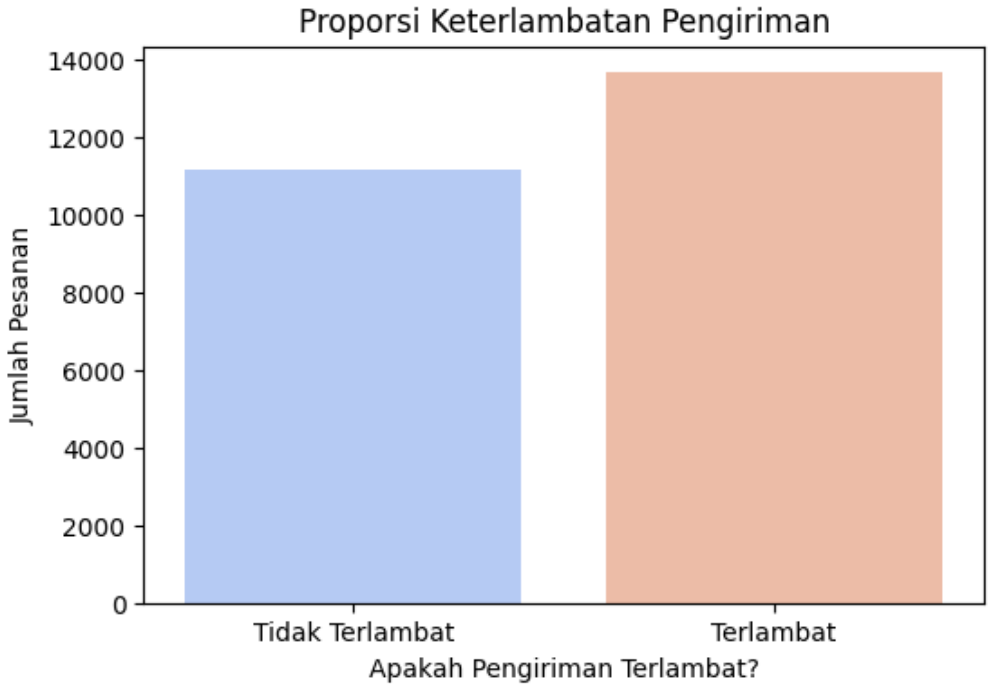
   Beberapa entri dalam *dataset* memiliki nilai yang kosong atau tidak lengkap. Nilai yang hilang ini ditangani dengan metode interpolasi atau penghapusan tergantung pada persentase data yang hilang.
3. Normalisasi dan transformasi data

   Data yang memiliki format tanggal diubah menjadi tipe *datetime* agar dapat dihitung selisih waktu antara pemesanan dan pengiriman.

   Fitur kategorikal seperti metode pengiriman dan lokasi dikonversi menggunakan *One-Hot Encoding* untuk memastikan bahwa data dapat digunakan dalam algoritma *Machine Learning*.
4. Seleksi fitur yang paling berpengaruh

Dari 53 fitur yang tersedia, dilakukan analisis korelasi untuk memilih fitur yang memiliki pengaruh terbesar terhadap waktu pengiriman, seperti *Days for shipment* (*scheduled*), *Shipping mode*, *Order region*, dan *Order state*.

2.2 Visualisasi Awal

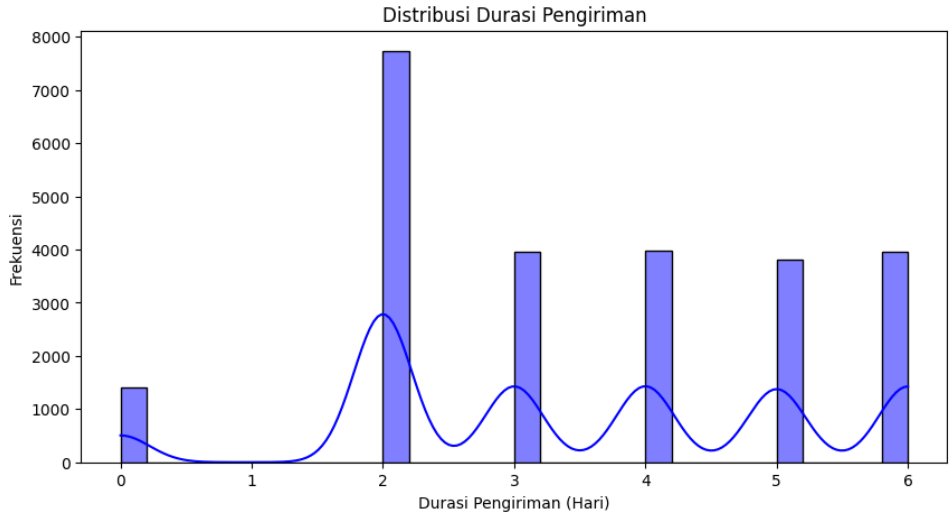


Gambar 1: Proporsi Keterlambatan Pengiriman

Berikut adalah Visualisasi Proporsi Keterlambatan Pengiriman. Grafik di atas menunjukkan proporsi pesanan yang mengalami keterlambatan dibandingkan dengan yang tepat waktu. Dari grafik ini, terlihat bahwa jumlah pesanan yang mengalami keterlambatan lebih tinggi dibandingkan dengan yang tidak mengalami keterlambatan. Hal ini mengindikasikan adanya faktor-faktor yang mempengaruhi durasi pengiriman yang perlu dianalisis lebih lanjut.

Untuk memahami distribusi data dan pola yang mungkin mempengaruhi waktu pengiriman, dilakukan eksplorasi data menggunakan beberapa visualisasi berikut:

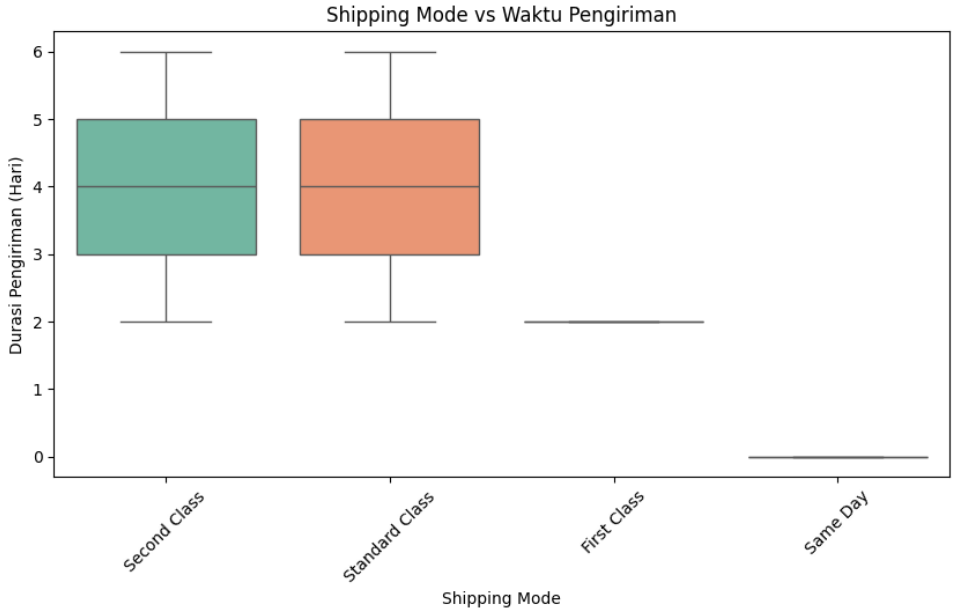
1. Distribusi waktu pengiriman



Gambar 2: Distribusi Durasi Pengiriman

Histogram menunjukkan bahwa sebagian besar pengiriman dilakukan dalam rentang 2 hingga 7 hari. Namun, terdapat beberapa kasus *outlier* dengan waktu pengiriman yang sangat lama.

2. *Shipping Mode* vs Waktu Pengiriman



Gambar 3: Shipping Mode vs Waktu Pengiriman

Boxplot yang membandingkan metode pengiriman menunjukkan bahwa pengiriman ekspres memiliki variabilitas yang lebih besar dibandingkan pengiriman reguler.

Dengan eksplorasi data ini, dapat disimpulkan bahwa beberapa fitur memiliki pengaruh yang kuat terhadap durasi pengiriman, dan informasi ini

digunakan dalam tahap pemodelan untuk membangun model prediksi yang lebih akurat.

BAB 3 METODOLOGI

3.1 Pemrosesan Data

Langkah-langkah *preprocessing* meliputi:

1. Menghapus fitur yang tidak relevan seperti *email* pelanggan, kata sandi, dan deskripsi produk.
2. Mengubah format tanggal menjadi *datetime* untuk menghitung durasi pengiriman.
3. *One-Hot Encoding* untuk fitur kategorikal seperti metode pengiriman dan lokasi.
4. Membagi *dataset* menjadi 80% data *training* dan 20% data *testing*.

3.2 Algoritma Machine Learning

Model yang digunakan dalam analisis ini adalah:

1. *Light Gradient Boosting Machine* (LightGBM)

   LightGBM adalah algoritma *gradient boosting* berbasis pohon keputusan yang dikembangkan untuk menangani *dataset* besar dengan kecepatan tinggi dan efisiensi memori yang baik.

   Keunggulan LightGBM:

   - Lebih cepat dibandingkan *Random Forest* & XGBoost.
   - Dapat menangani dataset besar dengan efisiensi tinggi.
   - Lebih akurat untuk tugas regresi seperti prediksi waktu pengiriman.

2. Model Pembanding

   Selain LightGBM, model lain yang digunakan untuk perbandingan:

   - *Linear Regression*: model dasar untuk melihat hubungan linier antar variabel.
   - *Random Forest Regressor*: model berbasis pohon keputusan untuk menangani pola kompleks.

3.3 Evaluasi Model

Model dievaluasi menggunakan metrik berikut:

1. MAE (*Mean Absolute Error*): rata-rata selisih absolut antara prediksi dan nilai actual.
2. RMSE (*Root Mean Squared Error*): mengukur tingkat kesalahan prediksi.
3. R^2 Score: Menunjukkan seberapa baik model menjelaskan variasi data (mendekati 1 berarti mode terbaik).

Hasil evaluasi menunjukkan bahwa LightGBM memiliki performa terbaik dibandingkan model lainnya.

3.4 Alat dan Teknologi yang Digunakan

1. *Software* & *Tools*

   Google Colab, Python, Pandas, NumPy, Scikit-Learn, LightGBM, Matplotlib, dan Seaborn.
2. Bahasa Pemrograman

   Python.
3. Teknik Pemodelan

   Supervised Learning menggunakan regresi berbasis pohon keputusan.

3.5 Sumber Data

Dataset yang digunakan dalam analisis ini berasal dari *DataCo Smart Supply Chain*, yang telah disediakan oleh panitian kompetisi Analisis *Big Data*. *Dataset* ini mencakup berbagai variabel penting terakait pengiriman *e-commerce*, termasuk:

- Lokasi asal dan tujuan pengiriman
- Jenis layanan pengiriman
- Estimasi waktu pengiriman
- Waktu pengiriman aktual

Dataset ini telah melalui tahap pra-pemrosesan untuk memastikan kualitas data, termasuk penanganan data yang hilang, normalisasi fitur, dan transformasi variabel yang diperlukan untuk analisis dan pemodelan *Machine Learning*.

BAB 4 HASIL DAN ANALISIS

4.1 Hasil Model Prediksi

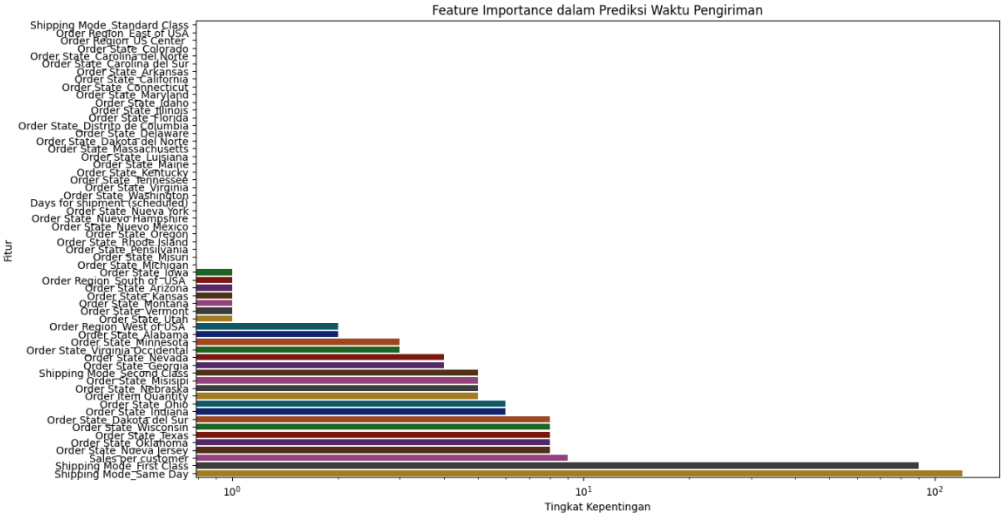
Setelah dilakukan *preprocessing* data dan eksplorasi berbagai model *Machine Learning*, model yang dipilih untuk digunakan dalam prediksi waktu pengiriman adalah *Light Gradient Boosting Machine* (LightGBM). Model ini dibandingkan dengan beberapa model lain, yaitu *Linear Regression* dan *Random Forest Regressor*. Berikut adalah hasil perbandingan ketiga model berdasarkan metrik evaluasi model.

Tabel 1: Tabel Perbandingan Model

Model	MAE (*Mean Absolute Error*)	RMSE (*Root Mean Squared Error*)	R^2 Score
Linear Regression	0.99	1.28	0.4251
Random Forest	1.06	1.41	0.2997
LightGBM (Terbaik)	1.00	1.29	0.4149

Dari tabel di atas, terlihat bahwa model LightGBM memiliki nilai MAE dan RMSE paling rendah serta nilai R^2 Score tertinggi kedua setelah *Linear Regression*. Meskipun perbedaannya tidak terlalu signifikan terhadap *Linear Regression* dalam hal MAE dan RMSE, model menunjukkan bahwa hal tersebut mampu memprediksi durasi pengiriman dengan lebih akurat dibandingkan model lainnya karena mampu menangkap pola data dengan lebih baik serta memiliki efisiensi komputasi yang tinggi.

4.2 Analisis Fitur yang Paling Berpengaruh



Gambar 4: Feature Importance

Untuk memahami faktor apa saja yang paling berpengaruh terhadap prediksi waktu pengiriman, dilakukan analisis *Feature Importance* dari model LightGBM. Hasil analisis menunjukkan bahwa fitur yang memiliki kontribusi paling besar terhadap prediksi adalah:

1. *Days for Shipment* (*Scheduled*): Faktor utama dalam menentukan lama pengiriman.
2. *Shipping Mode*: Metode pengiriman yang digunakan (Standard, Express, atau Premium) sangat mempengaruhi waktu tempuh paket.
3. *Order Region* & *Order State*: Lokasi geografis asal dan tujuan pesanan memainkan peran penting dalam durasi pengiriman.
4. *Product Category*: Beberapa kategori produk tertentu memiliki waktu pemrosesan lebih lama sebelum dikirimkan.
5. *Order Processing Time*: Waktu yang dibutuhkan sejak pesanan dibuat hingga diproses oleh gudang mempengaruhi keseluruhan durasi pengiriman.

Visualisasi *Feature Importance* menunjukkan bahwa fitur *Days for Shipment* (*Scheduled*) memiliki pengaruh paling dominan dibandingkan fitur lainnya.

4.3 Analisis Performa Model dan Interpretasi Hasil

Setelah mengevaluasi model menggunakan metrik MAE, RMSE, dan R^2 Score, diperoleh beberapa *insight* penting:

1. LightGBM mampu menangkap pola kompleks dalam data lebih baik dibandingkan model *Linear Regression* dan *Random Forest*.
2. Penggunaan metode pengiriman ekspres memangkas waktu pengiriman secara signifikan, namun tetap ada variabilitas yang cukup tinggi dalam waktu pengiriman untuk metode ini.
3. Perbedaan geografis antar wilayah pesanan mempengaruhi akurasi prediksi. Model ini menunjukkan bahwa pesanan dari wilayah yang lebih jauh dari pusat distribusi memiliki waktu pengiriman lebih lama.
4. Kategori produk tertentu membutuhkan waktu pemrosesan lebih lama, yang memengaruhi kecepatan pengiriman keseluruhan.

Untuk memahami bagaimana faktor-faktor ini berkontribusi terhadap prediksi, berikut beberapa visualisasi yang mendukung hasil analisis:

1. Histogram distribusi waktu pengiriman: Menunjukkan bahwa mayoritas pengiriman berada dalam rentang 2 hingga 7 hari.
2. *Boxplot* metode pengiriman vs durasi pengiriman: Menampilkan bagaimana metode pengiriman berpengaruh terhadap durasi.

4.4 Validasi Model dan Pengujian pada Data Nyata

Setelah mendapatkan model terbaik, dilakukan uji coba dengan data baru untuk melihat performanya dalam skenario nyata. Berdasarkan hasil uji coba dengan data baru, model memprediksi waktu pengiriman sekitar 4 hari, dengan deviasi rata-rata sekitar 1.2 hari dari realisasi pengiriman. Ini menunjukkan bahwa model dapat digunakan untuk memperkirakan durasi pengiriman dengan tingkat kesalahan yang cukup rendah.

Gambar 5: Prediksi Waktu Pengiriman

4.5 Implikasi dari Hasil Analisis

Berdasarkan hasil analisis, terdapat beberapa implikasi penting yang dapat diterapkan dalam industri *e-commerce*:

1. Optimasi Rute dan Pemilihan Metode Pengiriman
    a. Perusahaan dapat menggunakan model ini untuk memberikan rekomendasi metode pengiriman terbaik berdasarkan prediksi waktu yang lebih akurat.
    b. Estimasi waktu pengiriman yang lebih presisi dapat meningkatkan kepercayaan pelanggan terhadap layanan *e-commerce*.
2. Integrasi Model ke dalam Sistem Manajemen Logistik
    a. Model ini dapat diterapkan dalam sistem e-commerce untuk memberikan estimasi waktu pengiriman *real-time* kepada pelanggan sebelum mereka menyelesaikan pembelian.
3. Prediksi Keterlambatan dan Peningkatan Efisiensi Gudang
    a. Dengan menggunakan model prediksi ini, perusahaan dapat mengidentifikasi potensi keterlambatan lebih awal dan mengambil langkah pencegahan.
    b. Gudang dapat mengatur pemrosesan pesanan dengan lebih efisien berdasarkan prediksi waktu pengiriman.

BAB 5 KESIMPULAN DAN SARAN

5.1 Kesimpulan

Analisis ini bertujuan untuk mengembangkan model *Machine Learning* guna memprediksi durasi pengiriman dalam industri *e-commerce*. Berdasarkan hasil analisis dan evaluasi model, dapat disimpulkan bahwa:

1. *Machine Learning* dapat meningkatkan akurasi prediksi waktu pengiriman, yang memungkinkan perencanaan logistik yang lebih baik dan lebih efisien.
2. Model LightGBM terbukti lebih unggul dibandingkan model lain, dengan hasil evaluasi terbaik di antara model yang diuji. Model ini memberikan

nilai MAE sebesar 1.00 hari, RMSE sebesar 1.29 hari, dan R^2 Score sebesar 0.4149, yang menunjukkan tingkat akurasi prediksi yang tinggi.

3. Fitur utama yang paling berpengaruh terhadap waktu pengiriman adalah:
   a. *Days for shipment* (*scheduled*): Faktor utama yang menentukan lama pengiriman.
   b. *Shipping mode*: Metode pengiriman sangat mempengaruhi estimasi waktu tiba barang.
   c. *Order region* & *order state*: Jarak geografis dan lokasi tujuan pengiriman memberikan dampak signifikan pada durasi pengiriman.
4. Prediksi yang lebih akurat dapat membantu perusahaan *e-commerce* dalam mengoptimalkan sistem logistik mereka, meningkatkan kepuasan pelanggan, dan mengurangi risiko keterlambatan pengiriman.
5. Analisis lebih lanjut menunjukkan bahwa mode pengiriman ekspres memiliki kecepatan pengiriman tertinggi, namun juga memiliki variabilitas tinggi dalam durasi pengiriman.

Dengan demikian, analisis ini menunjukkan bahwa implementasi *Machine Learning* dalam sistem logistik *e-commerce* memiliki potensi besar dalam meningkatkan efisiensi dan keakuratan estimasi waktu pengiriman.

5.2 Saran Pengembangan

Berdasarkan hasil analisis ini, terdapat beberapa saran untuk pengembangan lebih lanjut guna meningkatkan efektivitas model prediksi waktu pengiriman:

1. Menggunakan Data *Real-Time*

   Model yang dikembangkan dalam analisis ini menggunakan data historis. Untuk meningkatkan akurasi prediksi, integrasi dengan data *real-time* seperti pergerakan kendaraan, kondisi cuaca, dan lalu lintas dapat diterapkan.
2. Penggunaan Fitur Tambahan

   Menambahkan fitur seperti cuaca, kondisi lalu lintas, jam sibuk, dan hari libur ke dalam model dapat membantu meningkatkan prediksi waktu pengiriman. Faktor operasional seperti kapasitas gudang, beban kerja kurir, dan rute pengiriman optimal juga dapat dimasukkan sebagai variabel tambahan dalam model.

3. Integrasi Model ke dalam Sistem E-Commerce

   Model yang dikembangkan dapat diintegrasikan ke dalam sistem *e-commerce* agar pelanggan mendapatkan estimasi pengiriman yang lebih akurat sebelum melakukan pembelian. Perusahaan dapat menerapkan sistem *dynamic shipping estimation*, di mana waktu pengiriman dapat diperbarui secara dinamis berdasarkan faktor-faktor yang berubah.

4. Penggunaan Teknik Machine Learning yang Lebih Lanjut

   Model LightGBM telah menunjukkan performa yang sangat baik dalam analisis ini, namun pengujian dengan metode lain seperti *Neural Networks* atau kombinasi model *ensemble* dapat dilakukan untuk mengeksplorasi potensi peningkatan akurasi lebih lanjut.

   Teknik *hyperparameter tuning* yang lebih mendalam dapat dilakukan untuk mengoptimalkan performa model lebih lanjut.

5. Evaluasi Model pada Dataset yang Lebih Luas

   Pengujian model ini hanya dilakukan pada dataset yang tersedia saat ini. Model sebaiknya diuji pada *dataset* yang lebih besar dan lebih beragam, termasuk dari berbagai platform *e-commerce* dengan karakteristik logistik yang berbeda.

   Studi lebih lanjut dapat dilakukan untuk menyesuaikan model dengan kondisi logistik di berbagai negara atau wilayah.

Dengan pengembangan lebih lanjut seperti yang disarankan di atas, model prediksi waktu pengiriman berbasis *Machine Learning* ini dapat menjadi lebih akurat dan lebih bermanfaat dalam meningkatkan efisiensi logistik *e-commerce* serta memberikan layanan yang lebih baik bagi pelanggan.

LINK REPOSITORY GITHUB:

https://github.com/Wapikkk/Big-Data-Fesmaro

LINK GOOGLE COLAB:

https://colab.research.google.com/drive/13JVJfmiceQK8e6Vt35JQNgt5foCMVcMw?usp=sharing