

BAB II

LANDASAN TEORI

2.1 Tinjauan Pustaka

2.2.1 Biaya Pengiriman Ekspor

Dalam konteks perdagangan internasional, aktivitas ekspor-impor tidak hanya melibatkan pertukaran barang antarnegara, tetapi juga mencakup serangkaian proses kompleks yang memerlukan perencanaan dan pengelolaan yang cermat. Salah satu aspek krusial dalam proses ini adalah biaya pengiriman ekspor, yang terdiri dari berbagai komponen seperti biaya transportasi, asuransi, pengurusan dokumen, bea cukai, dan jasa forwarder. Biaya transportasi, misalnya, dapat diminimalkan melalui penerapan metode transportasi yang efisien seperti metode Least Cost, yang telah terbukti efektif dalam mengurangi pengeluaran perusahaan pengiriman barang (Yertas et al., 2023). Selain itu, biaya pengurusan dokumen ekspor juga memerlukan perhatian khusus, karena ketidaktepatan dalam pengelolaannya dapat menyebabkan keterlambatan pengiriman dan peningkatan biaya operasional (Jaelani et al., n.d.). Penggunaan jasa forwarder dapat membantu dalam mengkoordinasikan proses pengiriman barang dari gudang hingga ke negara tujuan, namun juga menambah komponen biaya yang harus diperhitungkan secara cermat (Hansopaheluwakan et al., 2020). Oleh karena itu, pemahaman mendalam mengenai struktur biaya pengiriman

ekspor dan penerapan strategi efisiensi yang tepat sangat penting bagi perusahaan untuk meningkatkan daya saing dan keberlanjutan dalam pasar global.

2.2.1.1 Definisi dan Karakteristik Biaya Pengiriman Ekspor

Biaya pengiriman ekspor merupakan keseluruhan biaya logistik yang dikeluarkan dalam rangka mengirimkan barang dari dalam negeri menuju pasar internasional. Biaya ini dapat meliputi biaya angkut (freight), biaya asuransi, bea keluar, bea cukai, dan tambahan biaya lain yang bergantung pada jenis barang dan negara tujuan. Karakteristik biaya pengiriman ekspor bersifat dinamis, bergantung pada volume muatan, jarak tempuh, jenis komoditas, kondisi pasar logistik global, serta kebijakan perdagangan yang berlaku (Shiela et al., 2024).

Menurut (Nur Salam, 2022), variabilitas biaya ekspor juga dipengaruhi oleh aspek internal seperti jenis kapal yang digunakan (misalnya reefer untuk barang beku atau tanker untuk cairan), serta faktor eksternal seperti iklim dan permintaan pasar. Hal ini menjadikan proses estimasi biaya ekspor sebagai tantangan tersendiri yang membutuhkan pendekatan prediktif berbasis data historis.

2.2.1.2 Definisi dan Karakteristik Komoditas Ekspor

Komoditas ekspor adalah barang dagangan yang dikirim dari dalam negeri ke luar negeri sebagai bagian dari aktivitas perdagangan internasional. Komoditas ekspor memiliki karakteristik fisik dan penanganan yang berbeda-beda, sehingga membutuhkan jenis perlakuan logistik yang spesifik.

Menurut (Pambudi et al., 2020), barang seperti hasil laut beku (frozen seafood) harus dikirim dengan suhu rendah dan memerlukan kapal berpendingin (reefer), sedangkan bahan kimia atau cairan curah harus menggunakan kapal tanker. Komoditas elektronik atau barang bernilai tinggi membutuhkan perlindungan lebih dan biasanya memiliki biaya asuransi lebih besar. Perbedaan-perbedaan ini menyebabkan biaya pengiriman antar jenis komoditas menjadi sangat bervariasi, meskipun berat dan volumenya serupa.

2.2.1.3 Definisi dan Karakteristik Jenis Kapal

Jenis kapal pengangkut adalah salah satu faktor penting dalam logistik ekspor karena menentukan kapasitas, kecepatan, dan perlakuan khusus terhadap barang yang dikirim. Kapal pengangkut dalam ekspor dapat dibedakan menjadi beberapa jenis, seperti:

- a. **Container vessel:** digunakan untuk barang umum dalam kontainer.
- b. **Reefer vessel:** dilengkapi sistem pendingin untuk barang beku atau mudah rusak.
- c. **Tanker:** digunakan untuk cairan curah seperti minyak, bahan kimia, atau gas.

Menurut (Zebua et al., 2022), pemilihan jenis kapal yang tepat sangat memengaruhi struktur biaya pengiriman, karena perbedaan fasilitas dan spesifikasi kapal menimbulkan perbedaan tarif. Reefer vessel misalnya, membutuhkan energi tambahan untuk sistem pendingin sehingga biaya operasional dan pengiriman menjadi lebih tinggi dibanding kapal kontainer biasa.

2.2.2 Machine Learning untuk Prediksi

Machine learning merupakan cabang dari kecerdasan buatan yang berfokus pada pengembangan sistem yang dapat belajar dari data dan meningkatkan kinerjanya secara otomatis tanpa pemrograman eksplisit. Dalam konteks logistik dan ekspor, machine learning digunakan untuk memodelkan hubungan antara fitur logistik seperti berat, volume, dan jenis barang terhadap biaya pengiriman, serta memberikan estimasi biaya secara cepat dan akurat.

Menurut (Nur Salam, 2022), penerapan machine learning seperti regresi linier dalam analisis data ekspor dan impor terbukti mampu memberikan model prediksi yang kuat dan dapat diinterpretasikan. Selain itu, metode regresi sangat cocok untuk prediksi nilai kontinu seperti harga, biaya, atau jumlah.

2.2.3 Data Mining

Data mining adalah proses mengekstraksi informasi atau pola yang berguna dari kumpulan data besar dengan menggunakan teknik statistik, matematika, dan kecerdasan buatan. Tujuan utamanya adalah menemukan pengetahuan tersembunyi yang dapat digunakan untuk pengambilan keputusan yang lebih baik.

Menurut (Sitti Rahmah & Aldi Bastiatul Fawait, 2024), data mining memungkinkan analisis data historis untuk memprediksi kejadian di masa depan, seperti dalam kasus prediksi penyakit kanker paru menggunakan algoritma regresi linear. Teknik ini

membantu dalam mengidentifikasi pola dan hubungan dalam data yang tidak terlihat secara langsung.

Dalam konteks logistik dan ekspor, data mining dapat digunakan untuk menganalisis dan memprediksi biaya pengiriman berdasarkan variabel-variabel seperti berat, volume, jenis barang, dan tujuan pengiriman. (Priyadi et al., 2019), menunjukkan bahwa model prediktif berbasis data mining dapat digunakan untuk memprediksi harga emas berdasarkan nilai tukar dolar, suku bunga BI, dan harga minyak mentah dunia.

2.2.4 Linear Regression

Linear Regression merupakan metode statistik yang digunakan untuk memodelkan hubungan linier antara satu atau lebih variabel independen (predictor) dengan variabel dependen (target). Dalam konteks prediksi biaya ekspor, Linear Regression digunakan untuk mengukur pengaruh faktor-faktor seperti berat, volume, dan biaya lainnya terhadap total biaya pengiriman.

a. Regresi Linear Sederhana (1 variabel independen):

$$\hat{y} = \beta_0 + \beta_1 x \quad (1)$$

dimana \hat{y} = prediksi, β_0 = intercept, β_1 = kemiringan garis.

b. Regresi Linear Berganda (lebih dari 1 variabel independen):

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n \quad (2)$$

Algoritma:

- a. Menggunakan metode *Ordinary Least Squares* (OLS) untuk menemukan koefisien regresi yang meminimalkan selisih kuadrat antara nilai prediksi dan nilai aktual.
- b. Proses iteratif meliputi inisialisasi koefisien, prediksi nilai \hat{y} , evaluasi error, dan optimasi menggunakan algoritma seperti *Gradient Descent* hingga hasil stabil.

Fungsi Biaya:

Mean Squared Error (MSE) digunakan untuk mengukur rata-rata kuadrat perbedaan prediksi dan nilai nyata.

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2 \quad (3)$$

Hasil:

Model regresi linear dapat digunakan untuk:

- a. Memprediksi nilai variabel dependen.
- b. Menginterpretasi pengaruh variabel independen melalui koefisien regresi.
- c. Mengevaluasi kualitas model dengan metrik seperti R-squared (R^2).

Menurut (Shiela et al., 2024), Linear Regression dapat memberikan hasil yang cukup akurat dan mudah diinterpretasikan, serta cocok digunakan sebagai baseline

model untuk prediksi logistik karena memiliki kompleksitas yang rendah dan kecepatan pelatihan yang tinggi.

Selain itu dalam penelitian (Fitri, 2023), membandingkan metode Linear Regression dengan Random Forest Regression dan Gradient Boosted Trees Regression dalam memprediksi harga rumah. Hasilnya menunjukkan bahwa meskipun Linear Regression memiliki akurasi yang lebih rendah dibandingkan metode lainnya, namun tetap memberikan hasil yang cukup baik dan mudah diinterpretasikan.

2.2.5 Random Forest Regression

Random Forest Regression adalah algoritma machine learning berbasis ensemble yang menggabungkan banyak model regresi secara paralel untuk meningkatkan akurasi prediksi. Metode ini efektif digunakan pada data yang bersifat kompleks dan non-linier, serta mampu menangani variabel kategorikal maupun numerik secara optimal. Selain itu, Random Forest dapat memberikan informasi mengenai tingkat kepentingan setiap fitur (feature importance), sehingga membantu mengidentifikasi faktor-faktor yang paling berpengaruh terhadap biaya pengiriman.

Algoritma:

- a. Menggunakan *bootstrap sampling* untuk membuat banyak subset data.
- b. Membangun model pohon regresi dari tiap subset.
- c. Prediksi akhir dihitung dari rata-rata seluruh prediksi pohon:

$$\hat{y} = \frac{1}{B} \sum_{b=1}^{B} f_{b(x)} \quad (4)$$

Di mana B = jumlah pohon, $f_b(x)$ = prediksi dari pohon ke-b.

Kelebihan:

- a. Akurat dan tahan terhadap overfitting.
- b. Menangani variabel numerik dan kategorikal.
- c. Memberikan *feature importance* untuk mengetahui variabel yang paling berpengaruh.

Hasil:

- a. Memberikan prediksi nilai kontinu.
- b. Dapat mengevaluasi pengaruh setiap fitur.
- c. Cocok untuk data dinamis seperti biaya pengiriman.

Penelitian oleh (Zebua et al., 2022), menunjukkan bahwa Random Forest menghasilkan nilai R^2 sebesar 85,87% dalam memprediksi tarif penerbangan, yang mencerminkan kemampuan algoritma ini dalam menangani data dinamis dan kompleks.

2.2.6 Evaluasi Akurasi Model Regresi

Evaluasi akurasi dalam pemodelan regresi bertujuan untuk mengukur seberapa dekat nilai hasil prediksi terhadap nilai aktual. Akurasi dalam konteks regresi tidak dinilai berdasarkan jumlah prediksi yang benar seperti pada klasifikasi, melainkan berdasarkan besarnya deviasi antara nilai hasil prediksi dengan nilai aktual. Oleh karena itu, evaluasi regresi memanfaatkan berbagai metrik berbasis kesalahan (*error-based metrics*), yang masing-masing memiliki karakteristik dan fungsi tersendiri.

1. Mean Absolute Error (MAE)

Mean Absolute Error (MAE) mengukur rata-rata dari nilai absolut selisih antara hasil prediksi dan nilai aktual. Metrik ini bersifat linear, di mana setiap kesalahan dihitung secara setara tanpa memperbesar pengaruh kesalahan besar. Menurut (Botchkarev, 2021), MAE memiliki interpretasi yang sederhana dan tidak sensitif terhadap outlier, sehingga cocok digunakan ketika distribusi kesalahan bersifat seragam.

2. Mean Squared Error (MSE)

Mean Squared Error (MSE) merupakan rata-rata dari kuadrat selisih antara nilai aktual dan hasil prediksi. Berbeda dengan MAE, MSE memberikan penalti lebih besar terhadap kesalahan yang lebih besar, karena nilai kesalahan dikalikan dengan dirinya sendiri. (Lee et al., 2021), menyatakan bahwa MSE sering digunakan sebagai fungsi

objektif dalam pelatihan model karena bersifat kontinu dan dapat diturunkan secara matematis, sehingga memudahkan proses optimasi.

3. Root Mean Square Error (RMSE)

Root Mean Square Error (RMSE) adalah metrik evaluasi regresi yang digunakan untuk mengukur deviasi antara nilai prediksi dan nilai aktual. RMSE mengkuadratkan selisih antara nilai aktual dan prediksi, lalu dihitung rata-ratanya, dan diambil akarnya. Rumusnya adalah sebagai berikut:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

Keterangan:

1. y_i : nilai aktual ke- i
2. \hat{y}_i : nilai prediksi ke- i
3. n : jumlah data

Menurut (Botchkarev, 2021), RMSE memberikan penalti lebih besar terhadap kesalahan yang besar, sehingga lebih sensitif terhadap outlier dibanding metrik lain seperti MAE. Oleh karena itu, RMSE cocok digunakan saat kesalahan besar perlu dihindari, misalnya dalam prediksi biaya logistik.

Dan menurut (Zhu, 2020) menambahkan bahwa nilai RMSE juga dipengaruhi oleh ukuran sampel. Pada data yang besar, RMSE lebih stabil dan representatif. Namun, jika distribusi error tidak normal atau banyak outlier, RMSE sebaiknya digunakan bersama metrik lain untuk validasi yang lebih objektif.

4 R-Squared (R^2)

R^2 (koefisien determinasi) menunjukkan proporsi variansi yang dijelaskan oleh model. Nilainya berkisar antara 0 hingga 1—semakin mendekati 1, semakin kuat model menjelaskan data. Nilai tinggi tidak selalu berarti model sempurna jika tidak didukung metrik error lainnya (Hodson, 2022).

Penggunaan kombinasi MAE, MSE, RMSE, dan R^2 secara bersamaan memberikan gambaran menyeluruh terhadap akurasi dan kualitas model regresi. Keempat metrik ini saling melengkapi dalam menilai performa prediksi, baik dari sisi kestabilan, sensitivitas terhadap outlier, maupun tingkat variasi yang berhasil dijelaskan oleh model.

2.2.7 CRISP-DM (Cross Industry Standard Process for Data Mining)

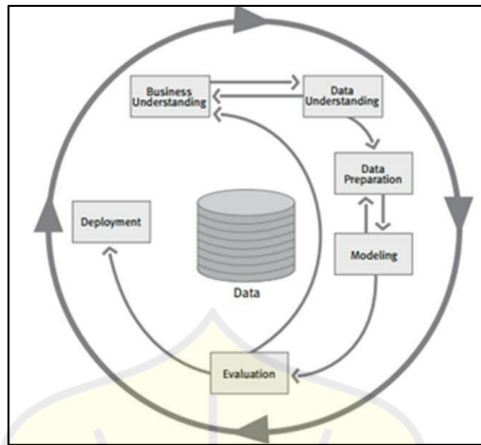
Setelah pemilihan algoritma yang tepat seperti Linear Regression dan Random Forest Regression, diperlukan suatu pendekatan metodologis yang sistematis untuk menerapkan model prediktif tersebut secara efisien dan sesuai dengan tujuan bisnis. Salah satu pendekatan yang umum digunakan dalam proyek data mining adalah CRISP-DM (Cross Industry Standard Process for Data Mining).

CRISP-DM adalah kerangka kerja metodologis yang dirancang untuk memandu proses data mining secara sistematis dan terstruktur. Dikembangkan pada akhir 1990-an oleh konsorsium yang terdiri dari Daimler-Benz, SPSS, dan NCR, CRISP-DM telah menjadi standar de facto dalam industri dan akademisi karena fleksibilitas dan pendekatannya yang terstruktur.

CRISP-DM membagi proses data mining menjadi enam fase utama: pemahaman bisnis, pemahaman data, persiapan data, pemodelan, evaluasi, dan implementasi. Setiap fase dirancang untuk memastikan bahwa proyek data mining berjalan sesuai dengan tujuan bisnis dan menghasilkan model yang dapat diandalkan.

Menurut penelitian oleh (Schröer et al., 2021), CRISP-DM tetap menjadi standar industri yang paling banyak digunakan dalam proyek data mining. Namun, studi tersebut juga mencatat bahwa sebagian besar penelitian tidak melanjutkan hingga fase implementasi, yang menunjukkan adanya tantangan dalam menerapkan model ke dalam proses bisnis secara nyata.

Dalam konteks prediksi biaya pengiriman ekspor, penerapan CRISP-DM memungkinkan analisis yang sistematis terhadap data historis, identifikasi pola, dan pengembangan model prediktif yang dapat membantu perusahaan dalam pengambilan keputusan strategis.



Gambar 2.1 Metode CRIPS-DM

2.2.8 Pemodelan UML

Di dunia pengembangan sistem berorientasi obyek, Unified Modeling Language (UML) adalah gabungan dari bahasa pemodelan yang dikembangkan oleh Booch, Objects Modeling Technique (OMT), dan Object Oriented Software Engineering (OOSE). Metode ini menggabungkan proses analisis dan desain ke dalam tahapan iteratif. Tahapan-tahap ini mencakup pengenalan objek dan kelas, pengenalan hubungan semantik antara objek dan kelas, perincian interface, dan implementasi (Voutama, 2022).





2.2.7.1 Use Case Diagram

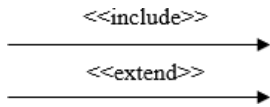
Use Case Diagram merupakan salah satu jenis diagram dalam Unified Modeling Language (UML) yang digunakan untuk memodelkan interaksi antara pengguna (aktor) dan sistem berdasarkan fungsi-fungsi yang akan dijalankan. Diagram ini bersifat fungsional dan bertujuan untuk menggambarkan apa saja yang dapat

dilakukan oleh aktor terhadap sistem dan bagaimana sistem merespons interaksi tersebut.

Menurut (Nugraha & Rosmeida, n.d.), dalam jurnalnya yang berjudul Analisis dan Perancangan Sistem Informasi Akademik Berbasis Web dengan UML (Studi Kasus: SMK Swasta Teladan Medan), penggunaan Use Case Diagram mampu memperjelas kebutuhan sistem karena menggambarkan secara visual fungsi-fungsi yang harus dimiliki oleh sistem berdasarkan interaksi pengguna. Hal ini mempermudah proses validasi dan verifikasi kebutuhan fungsional sistem sebelum tahap implementasi dimulai.

Table 1 Komponen Use Case Diagram

Simbol	Nama Komponen	Fungsi
	Aktor (<i>Actor</i>)	untuk memberikan input ke sistem, menerima informasi dari sistem, atau melakukan keduanya—mengirim dan menerima informasi.
	Dependency	Untuk menghubungkan antara dua elemen, di mana perubahan pada satu elemen memengaruhi elemen lainnya
	Association	hubungan antara elemen-elemen struktural yang saling terhubung dalam sebuah sistem.
	Generalization	Untuk menggambarkan relasi khusus antara objek anak (<i>child</i>) dan induk (<i>parent</i>).



Simbol	Nama Komponen	Fungsi
	Include dan extends	include menggambarkan perilaku yang wajib dipenuhi agar suatu <i>event</i> dapat terjadi extends menunjukkan perilaku tambahan yang hanya terjadi di bawah kondisi tertentu

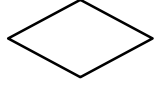

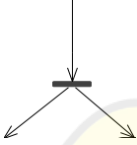

2.2.7.2 Diagram Activity

Menurut (Indriyani & Winarko, 2020), *Activity Diagram* adalah diagram yang digunakan untuk menggambarkan alur kerja (workflow) atau aktivitas dalam suatu sistem perangkat lunak. Diagram ini memvisualisasikan urutan aktivitas, kondisi keputusan, dan aksi, sehingga memberikan pemahaman yang jelas tentang proses yang sedang berlangsung dalam sistem.

Diagram ini berguna untuk menggambarkan proses bisnis atau logika sistem secara visual, memudahkan komunikasi antara tim pengembang dan pemangku kepentingan.

Table 2 Komponen activity Diagram

Simbol	Nama Komponen	Fungsi
	<ul style="list-style-type: none"> • <i>Start State</i> • <i>End State</i> 	<ul style="list-style-type: none"> • Sebagai tanda awal proses dari <i>activity diagram</i>. • menggambarkan akhir atau terminal dari pada sebuah <i>activity diagram</i>
	<i>Activity</i>	mengambarkan sebuah pekerjaan atau tugas dalam workflow

Simbol	Nama Komponen	Fungsi
	<i>Decision</i>	mengindikasikan suatu kondisi kemungkinan perbedaan transisi
	<i>State Transition</i>	menunjukkan kegiatan apa berikutnya setelah suatu kegiatan sebelumnya
	<i>Fork/percabangan</i>	menunjukkan kegiatan yang dilakukan secara paralel atau untuk menggabungkan dua kegiatan paralel menjadi satu
	<i>Join (penggabungan)</i>	menunjukkan adanya dekomposisi

2.2.9 Database dan DBMS

Menurut (Ramadhan & Mukhaiyar, 2020), Database merupakan suatu sistem yang dirancang untuk menyimpan, mengelola, dan mengakses data secara efisien. Data yang tersimpan di dalamnya biasanya terstruktur dan digunakan untuk satu atau lebih tujuan tertentu dalam format digital. Untuk mengelola data digital tersebut, digunakan Database Management System (DBMS), yaitu perangkat lunak yang berfungsi untuk menyimpan isi database, melakukan pengelolaan data, serta memfasilitasi proses pencarian dan akses informasi. Contoh DBMS yang banyak digunakan saat ini antara lain: MySQL, SQL Server, Microsoft Access, Oracle, dan PostgreSQL.

2.1.8.1 MySQL

Menurut (Lase & Alasi, 2024),MySQL adalah salah satu sistem manajemen basis data relasional (RDBMS) yang bersifat open source dan banyak digunakan dalam pengembangan aplikasi web. Popularitas MySQL didukung oleh kemampuannya yang ringan, cepat, serta mudah diintegrasikan dengan berbagai bahasa pemrograman seperti PHP, Python, dan Java.

Dan sangat cocok untuk pengembangan sistem informasi skala kecil hingga menengah karena mendukung struktur data yang kompleks, mudah digunakan, dan memiliki komunitas pengguna yang luas sebagai sumber bantuan teknis selama proses pengembangan.

2.2.10 Software dan Pemrograman Terkait

Pengembangan sistem prediksi berbasis machine learning tidak lepas dari dukungan perangkat lunak dan bahasa pemrograman yang tepat. Dalam penelitian ini, digunakan Python sebagai bahasa pemrograman utama karena fleksibilitasnya, komunitas yang luas, serta banyaknya pustaka (library) yang mendukung pengolahan data dan pembelajaran mesin.

Beberapa pustaka yang digunakan meliputi:

- a. Scikit-learn, yang menyediakan berbagai algoritma machine learning seperti regresi, klasifikasi, dan clustering.

- b. Pandas, yang sangat berguna untuk manipulasi dan analisis data dalam bentuk tabel (dataframe).
- c. Matplotlib, yang digunakan untuk visualisasi data dalam bentuk grafik atau plot.

Untuk pengembangan dan dokumentasi kode secara interaktif, digunakan Jupyter Notebook, yang memungkinkan integrasi penulisan kode, visualisasi, dan dokumentasi dalam satu antarmuka. Hasil prediksi kemudian disajikan melalui antarmuka web menggunakan bahasa pemrograman PHP, yang umum digunakan dalam pengembangan web dinamis.

2.2.9.1 Editor Jupyter

Jupyter Notebook merupakan platform open-source berbasis web yang memungkinkan integrasi kode, visualisasi, dan dokumentasi dalam satu dokumen interaktif. Fitur ini menjadikannya alat yang sangat efektif dalam bidang data science, pembelajaran mesin, dan penelitian ilmiah.

Menurut (Rule et al., 2019), Jupyter Notebook telah menjadi standar de facto dalam penelitian data science karena kemampuannya mengintegrasikan kode, output, dan narasi secara dinamis. Penelitian ini menggunakan Jupyter untuk pengembangan model Random Forest Regression dan eksplorasi data.

2.2.9.2 Bahasa Pemrograman Python dan Library Terkait

Python merupakan bahasa pemrograman tingkat tinggi yang bersifat open-source dan mendukung berbagai paradigma pemrograman, termasuk prosedural, berorientasi objek, dan fungsional. Sintaksisnya yang sederhana dan kemampuannya dalam mengintegrasikan berbagai pustaka (*library*) menjadikan Python sebagai pilihan utama dalam pengembangan aplikasi data science dan machine learning. Menurut (Joshi & Tiwari, 2023), Python menyediakan berbagai pustaka yang dapat dikategorikan ke dalam tiga kelompok utama: pengumpulan data, analisis dan pemrosesan data, serta visualisasi data.

Menurut (Sundaram et al., 2023) ,Dalam konteks penelitian ini, beberapa pustaka Python yang digunakan meliputi:

1. Pandas: Digunakan untuk manipulasi dan analisis data tabular. Pandas menyediakan struktur data *DataFrame* yang memungkinkan operasi seperti pembersihan data, transformasi, dan agregasi dilakukan dengan sintaks yang ringkas namun ekspresif. Kemampuan Pandas dalam menangani data yang memiliki nilai hilang (*missing values*), melakukan penggabungan (*merge*) dan penyatuan (*join*) dataset, serta operasi *time series* sangat mendukung tahap pra-pemrosesan data dalam penelitian ini.
2. NumPy: Digunakan untuk komputasi numerik, menyediakan array multidimensi yang efisien dan berbagai fungsi matematika tingkat tinggi. NumPy menjadi

fondasi bagi banyak pustaka ilmiah Python lainnya dan memungkinkan komputasi numerik berjalan dengan performa tinggi, yang sangat krusial dalam proses pelatihan model machine learning.

3. Matplotlib dan Seaborn: Digunakan untuk visualisasi data. Matplotlib adalah pustaka standar yang digunakan untuk membuat grafik dan plot dalam Python, mendukung berbagai jenis visualisasi seperti diagram batang, histogram, dan scatter plot. Seaborn dibangun di atas Matplotlib dan menyediakan antarmuka tingkat tinggi yang lebih intuitif dalam membuat visualisasi statistik yang menarik secara visual. Kombinasi kedua pustaka ini digunakan dalam penelitian untuk mengeksplorasi distribusi data, hubungan antar variabel, serta mempresentasikan hasil prediksi model.
4. Scikit-learn: Merupakan pustaka Python yang populer untuk implementasi algoritma machine learning, termasuk *Random Forest Regression*. Scikit-learn menyediakan antarmuka yang konsisten dan dokumentasi yang lengkap, sehingga memudahkan pengguna dalam membangun dan mengevaluasi model prediktif.

2.2.9.3 Library Scikit-learn

Scikit-learn merupakan pustaka open-source berbasis Python yang secara luas digunakan dalam pengembangan model machine learning. Pustaka ini menyediakan berbagai algoritma pembelajaran mesin, baik yang bersifat supervised maupun unsupervised. Salah satu keunggulan utamanya adalah kemudahan penggunaan, efisiensi dalam kinerja, dokumentasi yang lengkap, serta konsistensi pada antarmuka

pemrogramannya (API). Dengan ketergantungan yang minimal dan lisensi BSD yang terbuka, Scikit-learn banyak digunakan dalam lingkungan akademis maupun industry (Pedregosa et al., 2012).

Pustaka ini mendukung integrasi dengan pustaka lain seperti NumPy dan SciPy, serta menyediakan fitur-fitur penting seperti pemilihan model, preprocessing, dan validasi silang (*cross-validation*), yang membuatnya sangat fleksibel dan kuat untuk berbagai kebutuhan analisis data. Antarmuka yang konsisten juga memudahkan pengguna dalam menerapkan dan membandingkan berbagai metode algoritma secara efisien.

Dalam penelitian ini, Scikit-learn dimanfaatkan untuk membangun model Decision Tree Regression. Penggunaannya mempermudah proses pelatihan model dan evaluasi performa secara sistematis, sehingga mendukung analisis prediktif yang dibutuhkan dalam studi ini.

2.2.9.4 Bahasa Pemrograman PHP

PHP adalah bahasa pemrograman server-side untuk pengembangan web. Kode PHP dijalankan di server sebelum hasilnya dikirim ke browser. Bahasa ini bisa disisipkan dalam HTML dan berinteraksi dengan database seperti MySQL. PHP mendukung berbagai sistem operasi dan punya komunitas besar. Cocok untuk membuat website dinamis, dari yang sederhana sampai kompleks (Sinlae et al., n.d.).

2.2 Kajian Penelitian Terdahulu

Pada bagian ini, akan diberikan beberapa penelitian yang relevan dengan judul skripsi ini, yaitu “ judul skripsi”. Penelitian-penelitian terkait ini memberikan dasar teori serta pengetahuan tentang penggunaan algoritma Linear Regression dalam melakukan prediksi pada bidang logistik, pengiriman, dan keuangan.

2.2.1 Paper 1

Penelitian oleh Zebua et al. (2025) dalam jurnal "Prediksi Tarif Penerbangan Menggunakan Algoritma Random Forest dan Regresi Linier", mengungkapkan bahwa kombinasi regresi linier dan algoritma ensemble learning dapat mencapai nilai R^2 sebesar 85,87% dalam memprediksi tarif penerbangan yang sangat dinamis. Studi ini mendukung bahwa regresi linier tetap relevan sebagai baseline model yang baik dalam prediksi biaya.

Dalam penelitian ini, atribut yang digunakan untuk membentuk model prediksi mencakup tujuan penerbangan, jarak tempuh, musim atau periode liburan (sebagai indikator waktu yang memengaruhi lonjakan harga), waktu pembelian tiket (misalnya early booking atau last-minute), kelas penerbangan (ekonomi, bisnis, first class), serta jenis maskapai. Hubungan antar variabel ini sangat penting, karena masing-masing faktor secara signifikan memengaruhi fluktuasi harga tiket, seperti meningkatnya tarif saat musim liburan atau perbedaan harga berdasarkan maskapai dan kelas penerbangan yang dipilih. (Zebua et al., 2022).

2.2.1.1 Masalah Penelitian

Masalah utama dalam penelitian ini adalah sulitnya memprediksi tarif penerbangan yang sangat dinamis dan dipengaruhi oleh banyak faktor. Perubahan harga tiket secara cepat akibat musim, waktu pemesanan, jenis maskapai, dan kelas layanan menyebabkan kesulitan bagi konsumen dan penyedia layanan dalam mengestimasi biaya secara akurat. Oleh karena itu, dibutuhkan suatu model prediktif yang dapat menangkap kompleksitas tersebut.

2.2.1.2 Penyelesaian yang Dilakukan

Untuk mengatasi permasalahan tersebut, peneliti menggunakan pendekatan kombinasi antara algoritma Linear Regression dan Random Forest Regression. Pendekatan ini dipilih karena Linear Regression mampu memberikan interpretasi hubungan linier antar variabel, sementara Random Forest dapat menangani kompleksitas dan interaksi non-linier yang sulit ditangkap oleh metode linier biasa.

2.2.1.3 Metodologi yang Digunakan

Penelitian ini menerapkan dua algoritma machine learning, yaitu Linear Regression sebagai baseline model dan Random Forest Regression sebagai model utama. Data yang digunakan dalam penelitian mencakup sejumlah fitur yang relevan terhadap prediksi tarif penerbangan, seperti tujuan penerbangan, jarak tempuh, periode waktu pembelian (early booking atau last-minute), musim liburan, jenis maskapai, dan kelas penerbangan. Model yang dibangun kemudian dievaluasi menggunakan metrik

R-squared (R^2) untuk mengukur seberapa baik model mampu menjelaskan variabilitas atau perubahan dalam harga tiket.

2.2.1.4 Hasil Penelitian

Hasil penelitian menunjukkan bahwa algoritma Random Forest Regression memberikan performa terbaik dengan nilai R^2 sebesar 85,87%, yang berarti model mampu menjelaskan sebagian besar variabilitas dalam tarif penerbangan. Linear Regression tetap memberikan hasil yang baik dan digunakan untuk interpretasi awal. Kombinasi kedua metode terbukti efektif dalam menangani data harga yang fluktuatif dan kompleks.

2.2.2 Paper 2

Penelitian oleh Pambudi et al. (2020) dalam jurnal "Penerapan Algoritma Random Forest untuk Prediksi Status Pengiriman Barang pada Perusahaan Ekspedisi", menunjukkan bahwa metode machine learning seperti Random Forest dapat digunakan untuk memprediksi status pengiriman barang dengan akurasi sebesar 76,6%. Meskipun menggunakan algoritma yang berbeda dari Linear Regression, penelitian ini menegaskan pentingnya pemanfaatan machine learning dalam meningkatkan efisiensi pengelolaan logistik.

Dalam penelitian ini, meskipun variabel-variabel yang digunakan tidak dijelaskan secara eksplisit, namun mengingat fokusnya pada status pengiriman, besar kemungkinan model memanfaatkan atribut seperti lokasi pengirim dan penerima,

waktu pengiriman, jenis barang, estimasi waktu tempuh, serta status logistik terkini (misalnya: sedang di gudang, dalam perjalanan, atau telah diterima). Hubungan antar variabel ini penting untuk membentuk pola prediksi yang dapat mengidentifikasi apakah pengiriman akan tepat waktu, terlambat, atau mengalami kendala operasional lainnya (Pambudi et al., 2020).

2.2.2.1 Masalah Penelitian

Penelitian oleh Pambudi et al. (2020) berfokus pada permasalahan ketidakpastian dan ketidaktepatan dalam prediksi status pengiriman barang pada perusahaan ekspedisi. Masalah utama yang dihadapi adalah sulitnya menentukan apakah suatu pengiriman akan sampai tepat waktu, terlambat, atau mengalami kendala operasional lainnya, mengingat banyaknya variabel yang mempengaruhi proses pengiriman, seperti lokasi pengirim dan penerima, jenis barang, serta kondisi logistik yang dinamis.

2.2.2.2 Penyelesaian yang Dilakukan

Untuk mengatasi masalah tersebut, peneliti menggunakan pendekatan machine learning, khususnya algoritma Random Forest, guna membangun model prediksi yang dapat mengidentifikasi status pengiriman dengan lebih akurat. Model ini dirancang untuk memanfaatkan berbagai atribut pengiriman guna menangkap pola dan hubungan antar variabel yang mempengaruhi keberhasilan pengiriman tepat waktu.

2.2.2.3 Metodologi yang Digunakan

Metodologi penelitian meliputi pengumpulan data pengiriman yang mencakup atribut-atribut seperti lokasi pengirim dan penerima, waktu pengiriman, jenis barang, estimasi waktu tempuh, serta status logistik terkini. Data tersebut kemudian diproses dan digunakan untuk melatih model Random Forest. Algoritma ini dipilih karena kemampuannya dalam menangani data dengan variabel kompleks dan menghasilkan prediksi yang robust. Evaluasi model dilakukan dengan mengukur akurasi prediksi status pengiriman.

2.2.2.4 Hasil Penelitian

Hasil penelitian menunjukkan bahwa algoritma Random Forest berhasil memprediksi status pengiriman barang dengan akurasi sebesar 76,6%. Hal ini menegaskan bahwa metode machine learning dapat memberikan kontribusi signifikan dalam meningkatkan efisiensi dan keandalan pengelolaan logistik pada perusahaan ekspedisi. Meskipun model menggunakan algoritma berbeda dari Linear Regression, hasil ini mendukung pentingnya penerapan teknologi kecerdasan buatan dalam memecahkan masalah prediksi operasional.

2.2.3 Paper 3

Menurut penelitian yang dilakukan oleh Ratih (2019) dalam jurnal berjudul "Prediksi Penjualan Produk Elektronik Yang Terlaris Pada CV. Istana Komputer Palembang Menggunakan Algoritma Regresi Linear Sederhana", penerapan regresi linier sederhana dapat memprediksi penjualan produk elektronik dengan nilai RMSE

sebesar 1083 dan absolute error sebesar 1065. Hasil ini menunjukkan bahwa regresi linier sederhana efektif dalam memodelkan hubungan antara variabel penjualan dengan faktor-faktor seperti jenis produk elektronik, waktu penjualan, harga, diskon atau promosi, serta tren historis penjualan. Variabel-variabel tersebut saling berkontribusi terhadap prediksi, misalnya harga dan diskon dapat langsung memengaruhi minat beli, sementara tren historis mencerminkan pola pembelian di waktu sebelumnya (Ratih, 2019).

2.2.3.1 Masalah Penelitian

Penelitian yang dilakukan oleh Ratih (2019) mengangkat permasalahan ketidakpastian dalam memprediksi penjualan produk elektronik yang terlaris pada CV. Istana Komputer Palembang. Kesulitan muncul karena penjualan dipengaruhi oleh berbagai faktor seperti jenis produk, waktu penjualan, harga, diskon, serta tren historis yang kompleks dan berubah-ubah, sehingga dibutuhkan metode yang mampu memodelkan hubungan antar variabel ini secara efektif.

2.2.3.2 Penyelesaian yang Dilakukan

Untuk mengatasi permasalahan tersebut, peneliti menggunakan algoritma regresi linier sederhana sebagai metode prediksi. Metode ini diaplikasikan untuk mengidentifikasi hubungan linier antara variabel penjualan dan faktor-faktor pendukung seperti harga, diskon, dan tren penjualan sebelumnya. Dengan pendekatan ini, diharapkan dapat diperoleh model yang dapat memprediksi jumlah penjualan produk elektronik secara akurat.

2.2.3.3 Metodologi yang Digunakan

Metodologi penelitian meliputi pengumpulan data penjualan produk elektronik yang mencakup variabel-variabel seperti jenis produk, waktu penjualan, harga, diskon atau promosi, dan data historis penjualan. Data tersebut kemudian dianalisis menggunakan regresi linier sederhana untuk membangun model prediksi. Validasi model dilakukan dengan menggunakan metrik evaluasi seperti Root Mean Square Error (RMSE) dan absolute error untuk mengukur tingkat kesalahan prediksi.

2.2.3.4 Hasil Penelitian

Hasil penelitian menunjukkan bahwa regresi linier sederhana mampu memprediksi penjualan produk elektronik dengan nilai RMSE sebesar 1083 dan absolute error sebesar 1065. Nilai ini menunjukkan efektivitas metode dalam memodelkan hubungan antara variabel penjualan dengan faktor-faktor pendukung. Temuan ini menegaskan bahwa regresi linier sederhana dapat menjadi alat yang handal untuk membantu perusahaan dalam merencanakan strategi penjualan berdasarkan prediksi yang diperoleh.

2.2.4 Paper 4

Penelitian oleh Hakim dan Utari (2020) dalam jurnal "Prediksi Jumlah Pembelian Sepatu Dengan Penerapan Metode Regresi Linear" menunjukkan bahwa penggunaan regresi linier dalam sistem prediksi jumlah pembelian memudahkan pegawai toko dalam memprediksi jumlah pembelian, sehingga dapat meningkatkan efisiensi pengadaan barang. Regresi linier digunakan untuk menganalisis hubungan

antara beberapa variabel, seperti musim atau bulan penjualan, ukuran sepatu, jenis sepatu, jumlah stok sebelumnya, dan event promosi. Variabel-variabel ini saling berinteraksi dan mempengaruhi jumlah pembelian, dimana musim atau bulan penjualan dapat mempengaruhi tren permintaan, ukuran dan jenis sepatu menentukan preferensi konsumen, serta jumlah stok sebelumnya yang mencerminkan pola pembelian di masa lalu, sementara event promosi berpotensi meningkatkan jumlah pembelian dalam periode tertentu (Lukman Hakim & Utari, 2020).

2.2.4.1 Masalah Penelitian

Penelitian oleh Hakim dan Utari (2020) mengangkat permasalahan dalam memprediksi jumlah pembelian sepatu di toko yang seringkali sulit dilakukan secara akurat oleh pegawai toko. Kesulitan ini disebabkan oleh pengaruh berbagai faktor yang saling berinteraksi, seperti musim atau bulan penjualan, ukuran dan jenis sepatu, jumlah stok sebelumnya, serta event promosi, sehingga diperlukan metode yang dapat mengintegrasikan faktor-faktor tersebut untuk meningkatkan efisiensi pengadaan barang.

2.2.4.2 Penyelesaian yang Dilakukan

Untuk mengatasi permasalahan tersebut, peneliti menerapkan metode regresi linier dalam sistem prediksi jumlah pembelian. Metode ini digunakan untuk menganalisis hubungan antara variabel-variabel yang memengaruhi pembelian dan membangun model prediksi yang dapat membantu pegawai toko memperkirakan kebutuhan pengadaan barang dengan lebih tepat.

2.2.4.3 Metodologi yang Digunakan

Metodologi penelitian meliputi pengumpulan data penjualan sepatu yang mencakup variabel-variabel seperti musim atau bulan penjualan, ukuran dan jenis sepatu, jumlah stok sebelumnya, dan event promosi. Data tersebut dianalisis menggunakan regresi linier untuk mendapatkan model prediksi jumlah pembelian. Evaluasi model dilakukan dengan mengamati sejauh mana model dapat menangkap pola hubungan antar variabel serta kemampuannya dalam memperkirakan jumlah pembelian.

2.2.4.4 Hasil Penelitian

Hasil penelitian menunjukkan bahwa penerapan regresi linier dalam sistem prediksi jumlah pembelian sepatu dapat memudahkan pegawai toko dalam memperkirakan kebutuhan stok dengan akurat. Model ini membantu meningkatkan efisiensi dalam pengadaan barang dengan memperhitungkan pengaruh musim, preferensi konsumen berdasarkan ukuran dan jenis sepatu, pola pembelian masa lalu, serta dampak event promosi. Temuan ini menegaskan pentingnya metode regresi linier sebagai alat bantu dalam manajemen stok toko sepatu.

2.2.5 Paper 5

Menurut penelitian yang dilakukan oleh Nur Salam (2022) dalam jurnal berjudul "*Inferensi Model Regresi Linear untuk Ekspor dan Impor Provinsi Kalimantan Selatan Tahun 2020*", analisis regresi linier digunakan untuk menjelaskan hubungan antara variabel independen dan dependen dalam konteks

ekspor dan impor. Penelitian ini memberikan wawasan tentang bagaimana regresi linier dapat digunakan untuk menganalisis data perdagangan internasional. Dalam hal ini, variabel seperti nilai ekspor dan nilai impor dianalisis untuk melihat pengaruh dan korelasinya, dengan kemungkinan tambahan variabel seperti tahun dan jenis komoditas sebagai faktor pendukung. Nilai ekspor atau impor dapat dijadikan sebagai variabel dependen, sementara tahun dan jenis komoditas berperan sebagai variabel independen yang mempengaruhi fluktuasi perdagangan. Dengan model ini, dapat diperoleh pemahaman yang lebih baik tentang dinamika perdagangan antar waktu dan antar jenis produk (Nur Salam, 2022).

2.2.5.1 Masalah Penelitian

Penelitian yang dilakukan oleh Nur Salam (2022) mengangkat permasalahan mengenai bagaimana menganalisis hubungan antara nilai ekspor dan impor di Provinsi Kalimantan Selatan dalam tahun 2020. Permasalahan utama terletak pada kebutuhan untuk memahami faktor-faktor yang mempengaruhi fluktuasi perdagangan internasional, termasuk pengaruh tahun dan jenis komoditas terhadap nilai ekspor dan impor, sehingga diperlukan metode yang tepat untuk mengidentifikasi pola dan korelasi antar variabel tersebut.

2.2.5.2 Penyelesaian yang Dilakukan

Untuk mengatasi permasalahan tersebut, peneliti menggunakan model regresi linier sebagai alat analisis untuk menguji hubungan antara variabel independen seperti tahun dan jenis komoditas dengan variabel dependen berupa nilai ekspor dan impor.

Dengan pendekatan ini, diharapkan dapat diperoleh pemahaman yang lebih mendalam tentang dinamika perdagangan internasional di tingkat provinsi.

2.2.5.3 Metodologi yang Digunakan

Metodologi penelitian meliputi pengumpulan data perdagangan ekspor dan impor Provinsi Kalimantan Selatan pada tahun 2020, termasuk variabel nilai ekspor, nilai impor, tahun, dan jenis komoditas. Data tersebut kemudian dianalisis menggunakan teknik regresi linier untuk mengukur pengaruh variabel independen terhadap variabel dependen. Model regresi diuji untuk validitas dan keakuratannya dalam menjelaskan hubungan antar variabel.

2.2.5.4 Hasil Penelitian

Hasil penelitian menunjukkan bahwa model regresi linier mampu menjelaskan hubungan antara variabel tahun dan jenis komoditas dengan nilai ekspor dan impor di Provinsi Kalimantan Selatan. Temuan ini memberikan wawasan penting mengenai faktor-faktor yang mempengaruhi perdagangan internasional di tingkat regional serta memperkuat peran regresi linier sebagai metode yang efektif dalam analisis data ekonomi dan perdagangan.