

BAB II

TINJAUAN PUSTAKA

2.1 Penelitian yang Relevan

Sebagai bagian dari penyusunan landasan teori dan penegasan kebaruan, peneliti melakukan kajian terhadap penelitian terdahulu yang relevan. Kajian ini membantu peneliti memetakan perkembangan studi terkait, mengidentifikasi perbedaan pendekatan dan hasil, serta merumuskan *research gap* yang menjadi dasar penelitian peneliti. Berikut daftar jurnal yang digunakan:

Tahun	Penulis	Keterkaitan Penelitian	Hasil Penelitian	Gap Penelitian
<i>Pengaruh Nilai Tukar, Harga Dan Produksi Terhadap Ekspor Kayu Lapis Indonesia Ke Jepang Periode 2014-2023</i>				
2025	Arrashdan Tihami, Nurfatwa Al Agami dan Tria Apriliana	Menguji pengaruh nilai tukar, harga, dan produksi terhadap volume ekspor kayu lapis Indonesia ke Jepang dengan analisis regresi linear berganda.	Nilai tukar dan harga berpengaruh signifikan, sementara produksi tidak. Model memiliki nilai <i>Adjusted R-Square</i> sebesar 0,843.	Penelitian ini hanya focus pada komoditas tertentu saja.
<i>Comparative Analysis Using Multiple Regression Models for Forecasting Photovoltaic Power Generation</i>				
2024	Burhan U Din Abdullah, Shahbaz Ahmad Khanday, Nair Ul Islam, Suman Lata, Hoor Fatima, Sarvar Hussain Nengroo)	Menggunakan <i>multiple linear regression</i> sebagai model prediksi dan mengevaluasi kinerja dengan R^2 , MAE, MSE, RMSE pada skema <i>train-test split</i> .	Model regresi menunjukkan performa baik; nilai metrik dilaporkan jelas dalam tabel, misalnya R^2 mencapai 0,945 pada salah satu skenario.	Fokus pada prediksi energi bukan ekspor kehutanan dan validasi menggunakan split proporsional, bukan <i>time-based split</i> seperti penelitian ini.

*Analisis Daya Saing dan Faktor yang Mempengaruhi Volume Ekspor Biji Kopi
Indonesia Ke Jepang*

2021	Cindi Novariani, Muchtolifah, dan Sishadiyati	Menganalisis pengaruh produksi biji kopi, kurs rupiah, dan harga internasional terhadap volume ekspor biji kopi ke Jepang menggunakan regresi linear berganda.	Produksi biji kopi tidak berpengaruh signifikan, sementara kurs dan harga internasional berpengaruh negatif signifikan. Model memiliki nilai <i>R-squared</i> sebesar 0,929.	Peningkatan produksi domestik tidak secara langsung meningkatkan volume ekspor berbeda dengan fokus penelitian ini.
------	---	--	--	---

Medical Costs Estimation Using Linear Regression Method

2023	Ni Made Dita Dwikasaria, Ni Putu Sutramiania, Komang Sri Yanisa Putria, Nyoman Tri Rahaditya Kusumaa, Made Dimas Aldi Dwi Pramanaa, I Wayan Agus Surya Darmab	Menganalisis estimasi biaya pengobatan menggunakan regresi linear berganda dan sederhana berdasarkan faktor-faktor seperti usia, jenis kelamin, dan kebiasaan merokok.	Metode regresi linear berganda (MLR) lebih baik dibandingkan dengan regresi linear sederhana (SLR), dengan nilai R^2 sebesar 80% dan MSE serta MAE yang lebih rendah pada MLR.	Penelitian ini belum mengkaji penerapan regresi linear berganda dalam prediksi nilai ekspor, yang melibatkan variabel produk industri kehutanan.
------	---	--	--	--

*An Analysis of Implicit Value of Property Characteristics in Residential Property
Prices using a Hedonic Value Approach*

2021	Zulkarnain, Tania Arvianti	Menganalisis nilai implisit dari karakteristik properti yang mempengaruhi harga properti di Jakarta dengan menggunakan	MLR memberikan akurasi terbaik ($R^2 = 0.798$), dengan kedekatan fasilitas seperti stasiun MRT dan sekolah	Objek dan variabel penelitian bukan ekspor kehutanan, sehingga rujukan utamanya lebih pada alur pemodelan dan evaluasi metrik, bukan pada konteks
------	----------------------------	--	--	---

		metode regresi linear berganda, regresi kuantil, dan regresi vektor dukung (SVR).	menengah meningkatkan harga property, dibandingkan dengan model yang lain.	maupun variabel penelitian.
--	--	---	--	-----------------------------

Application of Multiple Linear Regression on Sales Prediction

2024	Chen Dong	Memprediksi volume penjualan dengan analisis regresi linear berganda berdasarkan pengeluaran di dalam toko, pengeluaran iklan daring, dan harga produk.	Pengeluaran di toko memiliki pengaruh lebih besar terhadap penjualan dibandingkan iklan daring. Model memiliki nilai R^2 sebesar 0,82381.	Model tidak mempertimbangkan optimalisasi alokasi dana promosi atau penentuan harga jual yang tepat serta objek penelitian masih belum relevan dengan penelitian ini.
------	-----------	---	---	---

Estimation of Future Number of Electric Vehicles and Charging Stations: Analysis of Sakarya Province with LSTM, GRU and Multiple Linear Regression Approaches

2025	Ayşe Tuğba Yapıcı, Nurettin Abut, Ahmet Yıldırım	Menggunakan MLR sebagai model baseline untuk prediksi berbasis deret waktu dan dibandingkan dengan LSTM/GRU. Evaluasi kinerja memakai R^2 , MAE, MSE, RMSE.	MLR cukup baik sebagai baseline R^2 sebesar 0,75, namun LSTM/GRU lebih akurat karena lebih mampu menangkap pola temporal.	Objek dan variabel penelitian bukan ekspor kehutanan, sehingga rujukan utamanya lebih pada alur pemodelan dan evaluasi metrik, bukan pada konteks maupun variabel penelitian.
------	--	---	---	---

Prediksi Hasil Produksi Cengkeh Di Indonesia Menggunakan Multiple Linear Regression

2025	Siti Nadiya Fauzani, Samsul Alam,	Memprediksi produksi cengkeh di Indonesia	Model regresi efektif dengan nilai R^2 sebesar 0,802. Luas lahan	Penelitian hanya berfokus pada prediksi produksi domestik tanpa
------	-----------------------------------	---	--	---

	Sunoko, Erip Suratno	menggunakan regresi linear berganda berdasarkan luas lahan dan produktivitas.	memiliki pengaruh yang lebih kuat terhadap produksi.	mengaitkannya dengan kinerja ekspor.
<i>Prediksi Harga Tiket Pesawat Domestik Rute Perjalanan Surabaya-Jakarta Menggunakan Metode Regresi Linear Berganda</i>				
2025	Enggi Sabrilla Assara, Hamzah Setiawan, Suprianto	Membangun model prediksi harga tiket Surabaya-Jakarta dengan regresi linear berganda berdasarkan maskapai, waktu keberangkatan, durasi, kapasitas bagasi, dan layanan.	Model regresi linear menunjukkan performa yang sangat baik dengan R^2 sebesar 96.4%, mampu mengidentifikasi faktor yang mempengaruhi harga tiket dengan akurasi tinggi.	Objek dan variabel penelitian bukan ekspor kehutanan, sehingga rujukan utamanya lebih pada alur pemodelan dan evaluasi metrik, bukan pada konteks maupun variabel penelitian.
<i>Pemodelan Prediksi Harga Kelapa Sawit menggunakan Regresi Linear Berganda Di Kecamatan Rantau Pulung Kabupaten Kutai Timur</i>				
2025	Talu Bella Lestiawati, Hanifah Ekawati, dan Eka Arriyanti	Memodelkan prediksi harga kelapa sawit berdasarkan luas kebun, tahun tanam, dan tonase menggunakan regresi linear berganda.	Tonase dan luas kebun berpengaruh signifikan terhadap harga, sementara tahun tanam tidak. Model memiliki nilai R-squared sebesar 0,946.	Secara teori, umur tanaman memengaruhi harga. Namun penelitian ini dibatasi pada tingkat koperasi lokal dan tidak memasukkan variabel eksternal, serta variabel tidak relevan dengan penelitian ini.

Berdasarkan tinjauan sepuluh penelitian yang relevan, sebagian besar studi yang menggunakan regresi linear berganda (MLR) fokus pada prediksi nilai atau volume ekspor untuk komoditas tertentu, seperti kayu lapis atau kopi, tanpa mempertimbangkan

produk kehutanan secara keseluruhan. Selain itu, banyak penelitian hanya menyoroti aspek domestik atau produksi tanpa mengaitkannya langsung dengan kinerja ekspor. Penggunaan pemisahan data berbasis waktu dalam evaluasi model prediksi yang relevan untuk perencanaan kebijakan jangka panjang masih terbatas dalam bentuk bulanan. Gap penelitian ini terletak pada penerapan MLR untuk memprediksi nilai ekspor produk kehutanan Indonesia secara agregat, dengan memasukkan variabel kunci seperti kayu bulat, kayu olahan, dan hasil hutan bukan kayu (HHBK). Penelitian ini juga akan menilai akurasi model dengan metrik yang jelas dan menggunakan pemisahan data berbasis waktu.

Pemilihan *multiple linear regression* (MLR) didasarkan pada kemampuannya untuk memodelkan hubungan linier antar variabel yang mempengaruhi ekspor produk industri kehutanan, seperti kayu bulat, kayu olahan, dan HHBK. MLR terbukti efektif dalam memprediksi ekspor dengan banyak faktor penentu, menghasilkan nilai R^2 yang tinggi dan akurasi yang baik. Koefisien MLR juga mempermudah pemahaman kontribusi masing-masing variabel terhadap nilai ekspor. Dengan demikian, MLR sangat cocok untuk penelitian ini, mengingat kemampuannya menjelaskan hubungan antara variabel yang mempengaruhi ekspor produk kehutanan Indonesia untuk dilakukan prediksi.

2.2 Landasan Teori

2.2.1 Ekspor Kehutanan

Ekspor kehutanan merupakan kegiatan perdagangan internasional yang memasarkan produk hasil hutan Indonesia ke pasar luar negeri, sejalan dengan definisi ekspor dalam Kamus Besar Bahasa Indonesia (KBBI). Produk yang diekspor mencakup kayu bulat (*log*), kayu olahan seperti kayu gergajian (*sawn wood*) dan kayu lapis (*plywood*), hasil hutan bukan kayu (HHBK) seperti rotan dan getah, serta produk hilir bernilai tambah seperti kertas dan furnitur (BPS, 2024). Dalam perekonomian nasional, ekspor kehutanan berperan sebagai sumber devisa non-migas yang mendorong industri domestik dan memperluas akses pasar internasional (Direktorat Statistik Distribusi, 2024)

Peran ekspor kehutanan diperkuat melalui kebijakan pemerintah. Peraturan Menteri LHK Nomor 13 Tahun 2020 menetapkan nilai ekspor hasil hutan sebagai indikator kontribusi sektor kehutanan terhadap perekonomian nasional. Selanjutnya, Peraturan Menteri LHK Nomor 8 Tahun 2021 menekankan fasilitasi ekspor dengan

memastikan legalitas dan keberlanjutan sumber bahan baku melalui instrumen Dokumen V-Legal dan Lisensi FLEGT, terutama untuk tujuan ekspor ke Uni Eropa.

Dalam teori perdagangan internasional, nilai ekspor dapat dijelaskan melalui pendekatan permintaan (*demand-side*) dan penawaran (*supply-side*). Pendekatan *supply-side* menekankan bahwa kemampuan ekspor ditentukan oleh kapasitas produksi yang memadai dan stabil (Indartik, 2008). Pada industri kehutanan, produksi kayu bulat, kayu olahan, dan HHBK mencerminkan kapasitas riil sektor dalam menyediakan komoditas ekspor. Peningkatan kapasitas produksi umumnya diikuti oleh peluang peningkatan nilai ekspor.

Di Indonesia, ekspor kehutanan dijalankan dalam kerangka regulasi untuk menjamin aspek legalitas dan keberlanjutan. Penerapan Sistem Verifikasi Legalitas dan Kelestarian (SVLK) ditujukan untuk memastikan produk kayu yang diekspor memiliki asal-usul yang jelas, berasal dari sumber yang sah, serta dikelola secara berkelanjutan. Pemenuhan ketentuan ini dibuktikan melalui dokumen seperti *V-Legal* atau Lisensi FLEGT, khususnya untuk tujuan ekspor ke Uni Eropa.

2.2.2 Produk Industri Kehutanan Indonesia

Produk industri kehutanan merujuk pada barang yang dihasilkan melalui pengolahan bahan baku dari sumber daya hutan. Proses ini bertujuan untuk meningkatkan nilai produk hutan mentah, seperti kayu, dengan mengubahnya menjadi barang setengah jadi atau produk jadi yang siap dipasarkan, baik di pasar domestik maupun internasional (Direktorat Jenderal Pengelolaan Hutan Lestari, 2024). Menurut Laporan Kinerja Direktorat Jenderal Pengelolaan Hutan Lestari, produk ekspor utama Indonesia meliputi kertas, panel kayu, *pulp*, furnitur kayu, dan bahan bangunan yang dirakit.

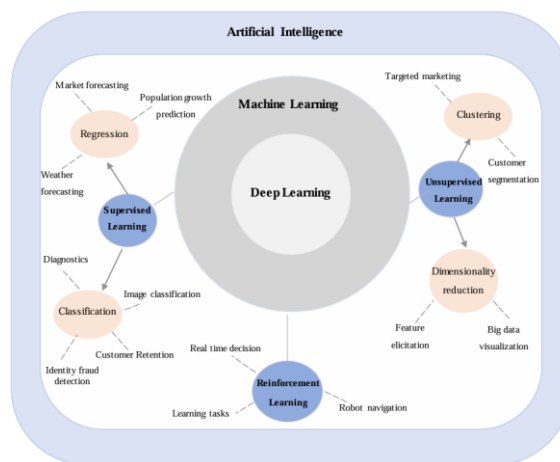
Dalam industri kehutanan, produk industri kehutanan terbagi menjadi dua sektor utama, yaitu sektor hulu dan hilir. Sektor hulu menghasilkan bahan baku seperti kayu bulat, yang kemudian diproses di sektor hilir menjadi produk bernilai tambah seperti kayu olahan dan produk non-kayu (HHBK). Ketiga sektor ini menjadi langkah strategis pemerintah untuk meningkatkan daya saing di pasar global.

Keunggulan produk kehutanan Indonesia di pasar internasional terletak pada jaminan legalitas dan keberlanjutan. Setiap produk ekspor harus memenuhi standar Sistem Verifikasi Legalitas dan Kelestarian (SVLK), yang memastikan produk berasal

dari sumber yang sah dan dikelola secara berkelanjutan, dengan validasi melalui Dokumen V-Legal atau Lisensi FLEGT sebagai persyaratan eksportir.

2.2.3 Artificial Intelligence

Artificial Intelligence (AI) atau kecerdasan buatan merupakan bidang yang berfokus pada pengembangan sistem yang mampu meniru kemampuan kognitif manusia, seperti belajar, bernalar, memecahkan masalah, dan memahami bahasa (Darwish, 2024). Dalam penerapan, AI memanfaatkan data serta metode komputasi untuk membantu mesin menjalankan tugas yang umumnya memerlukan kecerdasan manusia



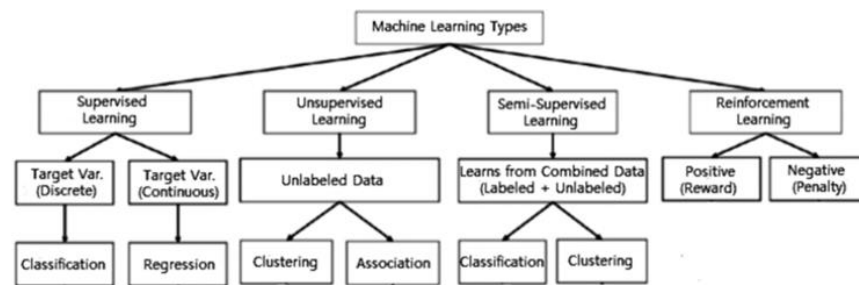
Gambar 2. 1 Hubungan Hierarkis Antara AI, ML, DL
Sumber: Darwish, 2024

Hubungan hierarkis antara *Artificial Intelligence* (AI), *Machine Learning* (ML), dan *Deep Learning* (DL) ditunjukkan pada Gambar 2.1. Secara hierarki, *artificial intelligence* menjadi payung yang mencakup ML serta DL. *Machine learning* merupakan cabang dari AI yang berfokus pada kemampuan sistem untuk belajar dari data dan meningkatkan kinerja seiring waktu tanpa harus diprogram aturan secara langsung. *Deep learning* merupakan bagian dari *machine learning* yang menggunakan jaringan saraf tiruan berlapis untuk mengenali pola yang lebih kompleks, misalnya pada data gambar dan teks (Paraijun et al., 2022).

2.2.4 Machine Learning

Machine Learning (ML) merupakan cabang dari *Artificial Intelligence* (AI) yang memungkinkan sistem belajar dari data sehingga performanya meningkat tanpa perlu

diprogram secara eksplisit untuk setiap aturan. ML penting dalam pendekatan *data-driven*, yaitu pendekatan yang menjadikan data sebagai dasar utama untuk analisis dan pengambilan keputusan, sehingga sistem dapat menemukan pola dari data untuk membangun solusi cerdas di berbagai bidang(Tulasi, 2024). Kategori utama dalam *machine learning* ditunjukkan pada Gambar 2.2.



Gambar 2. 2 Kategori *Machine Learning*

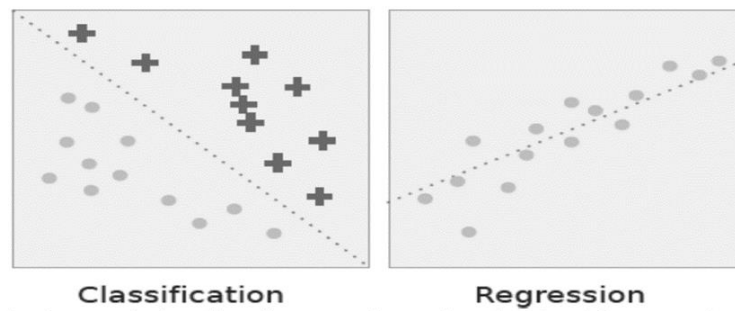
Sumber: Tulasi, 2024

Secara umum, *machine learning* memiliki beberapa pendekatan. Namun, penelitian ini memfokuskan pada *supervised learning* karena menggunakan data berlabel dan menargetkan keluaran yang jelas, sehingga lebih sesuai untuk kebutuhan prediksi yang terarah.

2.2.5 *Supervised Learning*

Supervised Learning adalah metode dalam *machine learning* yang melatih model dengan data berlabel. Setiap data yang diinput sudah memiliki label keluaran sebagai acuan, sehingga model belajar membentuk fungsi pemetaan dari input ke output. Targetnya adalah agar model dapat memberikan prediksi yang tepat ketika menghadapi data baru yang belum pernah diproses sebelumnya (Tulasi, 2024).

Pada *supervised learning* dikategorikan menjadi dua, yaitu klasifikasi dan regresi, sebagaimana ditunjukkan pada Gambar 2.3. Klasifikasi (*Classification*) digunakan untuk memprediksi target berupa kategori atau label. Regresi (*Regression*) digunakan untuk memprediksi target berupa nilai numerik yang bersifat kontinu dengan melihat hubungan antara variabel input dan output.



Gambar 2. 3 Ilustrasi dari Kategori Klasifikasi dan Regresi
Sumber: Tulasi, 2024

2.2.6 Regresi

Regresi (*Regression*) adalah teknik pada *supervised learning* yang digunakan untuk memodelkan hubungan antara variabel dependen (target) dan satu atau lebih variabel independen (bebas) (Damodar N. Gujarati & Dawn C. Porter, 2009).. Fokus utamanya adalah menghasilkan prediksi nilai rata-rata variabel target berdasarkan variabel bebas yang ditentukan, sehingga membantu membaca pola hubungan secara statistic.

Berdasarkan jumlah variabel independen, regresi dibedakan menjadi regresi linier sederhana (*linear regression*) untuk satu variabel independen dan regresi linier berganda (*multiple linear regression*) untuk dua atau lebih variabel independen. Regresi linier berganda banyak digunakan dalam analisis ekonomi karena dapat melihat pengaruh beberapa faktor terhadap satu hasil yang diamati.

2.2.7 Multiple Linear Regression

Multiple Linear Regression adalah pengembangan dari analisis regresi sederhana, yang hanya melibatkan dua variabel (Abdurrasyid et al., 2021; Damodar N. Gujarati & Dawn C. Porter, 2009b). Jika regresi sederhana mengkaji hubungan antara satu variabel dependen (Y) dan satu variabel independen (X), regresi berganda mempelajari pengaruh dua atau lebih variabel independen terhadap variabel dependen secara bersamaan. Tujuan utama dari model ini adalah untuk memperkirakan nilai rata-rata variabel dependen (Y) dengan mempertimbangkan nilai-nilai variabel independen (X) yang telah diketahui (Jatnika et al., 2025).

Model *multiple linear regression* dengan k -variabel dapat direpresentasikan dengan persamaan berikut:

$$\hat{Y} = b_0 + b_1X_1 + \dots + b_kX_k \quad (2.1)$$

Dalam persamaan tersebut, b_0 menyatakan *intercept* atau konstanta, sedangkan b_1 hingga b_k merupakan koefisien regresi menunjukkan perubahan rata-rata nilai $\hat{Y}$ akibat perubahan satu unit pada suatu variabel independen, dengan asumsi variabel lainnya konstan. Dengan demikian, analisis pengaruh tiap variabel independen terhadap variabel dependen secara terpisah.

Metode *multiple linear regression* didasarkan pada asumsi dalam kerangka *Classical Linear Regression Model* (CLRM), antara lain hubungan linier, *error* dengan nilai rata-rata nol, serta ragam error yang konstan atau homoskedastis (Damodar N. Gujarati & Dawn C. Porter, 2009). Dalam kaitannya dengan multikolinearitas, CLRM menekankan bahwa permasalahan utama muncul pada kondisi multikolinearitas tidak menyebabkan estimator menjadi bias, namun dapat meningkatkan varians koefisien regresi sehingga interpretasi pengaruh parsial perlu dilakukan secara lebih hati-hati (Mason & Perreault, 1991; Upendra et al., 2023).

2.2.8 Ordinary Least Squares (OLS)

Ordinary Least Squares (OLS) adalah metode estimasi parameter untuk mencari nilai parameter (*intercept* dan koefisien) yang membuat selisih antara nilai aktual dan nilai prediksi (*residual*) sekecil mungkin, dengan kriteria meminimalkan jumlah kuadrat residual. Secara konseptual, jika y_i adalah nilai aktual dan $\hat{y}_i$ adalah nilai prediksi, maka OLS memilih parameter β sehingga fungsi berikut minimum (Hansen, 2021):

$$\min_{\beta} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (2.2)$$

Dengan keterangan:

- a) n : jumlah observasi dalam dataset.
- b) y_i : nilai aktual variabel dependen pada observasi ke- i .
- c) $\hat{y}_i$: nilai prediksi variabel dependen pada observasi ke- i .
- d) $(y_i - \hat{y}_i)$: residual/galat prediksi observasi ke- i .
- e) β : vektor parameter model yang mencakup *intercept* dan koefisien regresi.

Fungsi tersebut menunjukkan bahwa OLS memilih nilai β yang meminimalkan total kuadrat residual, sehingga perbedaan antara nilai aktual dan nilai prediksi secara keseluruhan menjadi sekecil mungkin.

Dalam bentuk matriks, model regresi linier dapat ditulis sebagai:

$$y = X\beta + \varepsilon \quad (2.3)$$

Dengan keterangan:

- a) y : vektor nilai aktual
- b) X : matriks desain yang berisi variabel independen
- c) β : vektor parameter
- d) ε : vektor error.

Estimator OLS untuk menghitung parameter β adalah:

$$\hat{\beta} = (X^T X)^{-1} X^T y \quad (2.4)$$

Dengan keterangan:

- a) $\hat{\beta}$: vektor estimasi parameter OLS.
- b) X^T : transpose dari matriks X .
- c) $(X^T X)^{-1}$: invers dari matriks $X^T X$
- d) $X^T y$: hasil perkalian transpose matriks fitur dengan vektor target.

Jika asumsi dasar regresi linier terpenuhi, seperti error dengan rata-rata nol, varians konstan, dan tidak berkorelasi, maka dalam kerangka Gauss-Markov, estimator OLS bersifat linear, tidak bias, dan memiliki varians terkecil di antara estimator lain. Meskipun asumsi ini jarang sempurna dalam praktik, OLS tetap banyak digunakan karena sederhana dan hasil estimasinya mudah diinterpretasikan (Upendra et al., 2023).

2.2.9 Variabel Penelitian

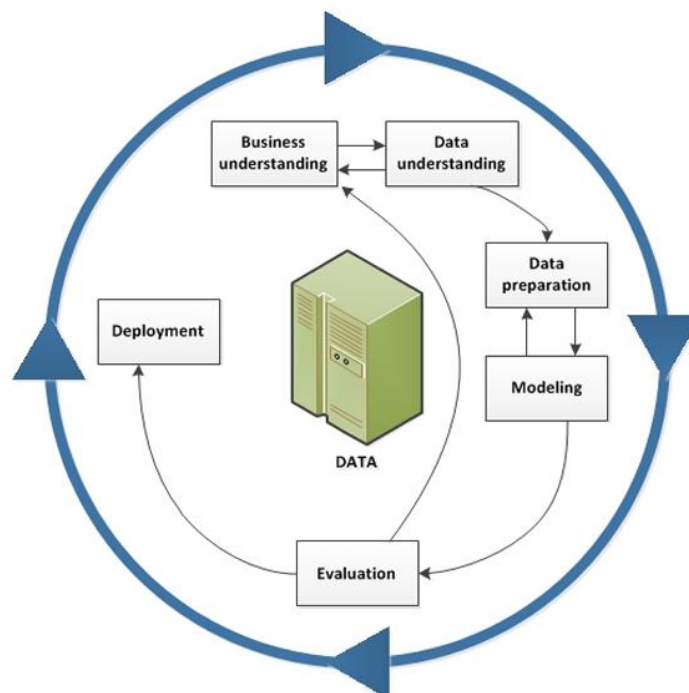
Variabel penelitian adalah atribut atau karakteristik yang nilainya dapat berubah dan dapat diukur maupun diamati selama penelitian (Damodar N. Gujarati & Dawn C. Porter, 2009). Dalam konteks analisis regresi, variabel dikelompokkan berdasarkan hubungan ketergantungan, yakni variabel dependen yang nilainya dipengaruhi oleh satu atau lebih variabel independen. Klasifikasi ini penting karena memperjelas posisi setiap variabel dalam model sekaligus menunjukkan arah hubungan yang diuji.

1. Variabel dependen (Y) adalah variabel utama yang nilainya dijelaskan atau diprediksi berdasarkan pengaruh variabel lain dalam model.
2. Variabel independen (X) adalah variabel penjelas yang digunakan sebagai prediktor untuk menerangkan perubahan atau variasi pada variabel dependen (Y).

2.2.10 Crisp-DM (*Cross-Industry Standard Process for Data Mining*)

Penelitian ini menerapkan kerangka kerja CRISP-DM (*Cross-Industry Standard Process for Data Mining*) untuk memastikan seluruh proses penelitian berlangsung secara sistematis dan terstruktur. CRISP-DM merupakan metodologi yang telah digunakan secara luas dalam proyek *data mining* karena mencakup seluruh tahapan analisis, mulai dari perumusan masalah hingga penerapan hasil analisis (Azzura et al., 2022; IBM, 2018a; Ummah et al., 2021).

Metodologi CRISP-DM terdiri dari enam fase utama yang saling terhubung dan membentuk alur kerja yang berkelanjutan. Alur kerja dan hubungan antar fase akan diilustrasikan pada Gambar 2.4.



Gambar 2. 4 Alur Kerja CRISP-DM

Sumber: IBM, 2018

Siklus CRISP-DM terdiri dari enam fase utama, yang dijelaskan sebagai berikut

1. *Business Understanding*: Fase ini berfokus pada pemahaman tujuan bisnis, penerjemahannya ke dalam tujuan proyek, serta penetapan kriteria keberhasilan.
2. *Data Understanding*: Tahap ini melibatkan pengumpulan dan eksplorasi data untuk memahami karakteristik, pola, dan kualitasnya.
3. *Data Preparation*: Fase ini mencakup pembersihan, pemilihan, integrasi, dan transformasi data menjadi dataset yang siap digunakan dalam proses pemodelan.
4. *Modeling*: Pada tahap ini, berbagai algoritma diterapkan dan disesuaikan untuk menghasilkan model dengan performa terbaik.
5. *Evaluation*: Model yang telah dibangun dievaluasi untuk memastikan hasilnya sesuai dengan tujuan penelitian dan kebutuhan analisis.
6. *Deployment*: Fase ini mengimplementasikan model dan wawasan ke dalam operasional bisnis melalui laporan atau integrasi sistem.

Alur kerja CRISP-DM bersifat fleksibel dan iteratif, sehingga memungkinkan peneliti untuk kembali ke tahap sebelumnya apabila ditemukan kebutuhan atau temuan baru, seperti perlunya penyesuaian data pada tahap pemodelan (IBM, 2018).

2.2.11 One-Hot Encoding

One-hot encoding merupakan metode transformasi data yang bertujuan mengonversi variabel kategorikal menjadi bentuk numerik melalui representasi biner (Carrizosa et al., 2022). Setiap kategori dibuatkan satu kolom dummy, dengan nilai 1 ketika suatu observasi masuk ke dalam kategori tersebut dan 0 ketika tidak. Dengan skema ini, kategori diperlakukan sebagai label yang setara, sehingga tidak memunculkan asumsi adanya urutan atau jarak tertentu antar kategori.

Dalam regresi linier, teknik ini menjadi relevan karena model membutuhkan seluruh fitur input dalam format numerik, sementara variabel seperti wilayah atau periode waktu umumnya bersifat kategorikal. Jika kategorikal nominal seperti bulan langsung diubah menjadi angka (misalnya Januari = 1, Februari = 2), model bisa salah paham dan mengira angka tersebut menunjukkan urutan atau jarak tertentu. Padahal, pada data kategori, angka itu hanya label, bukan ukuran. Karena itu, *one-hot encoding* sering digunakan agar kategori tetap bisa dimasukkan ke model tanpa membuat model menafsirkan adanya urutan atau jarak, sehingga hasil analisis lebih tepat dan tidak menimbulkan bias.

2.2.12 Robust Scaler

Robust Scaler adalah metode penskalaan data numerik yang menggunakan median dan rentang interkuartil (IQR), yang membuatnya lebih tahan terhadap pengaruh outlier dibandingkan metode penskalaan lainnya. Metode ini sangat efektif untuk data yang mengandung nilai ekstrim, karena tidak dipengaruhi oleh nilai yang sangat tinggi atau rendah, sehingga memberikan hasil penskalaan yang lebih stabil (Munawir Lasiyono & Gomes Soares, 2025).

Normalisasi data menggunakan robust scaler dapat dihitung dengan rumus berikut:

$$x' = \frac{x - \text{median}(x)}{\text{IQR}(x)} \quad (2.5)$$

Dengan keterangan:

- a. x = nilai asli data
- b. $\text{median}(x)$ = nilai median (kuartil ke-2) dari data
- c. $\text{IQR}(x)$ = rentang interkuartil, yaitu selisih antara kuartil atas ($Q3$) dan kuartil bawah ($Q1$), atau $\text{IQR}(x) = Q3 - Q1$

Dengan menggunakan robust scaler, data akan dinormalisasi berdasarkan median dan IQR, yang menjadikannya lebih tahan terhadap outlier dan memungkinkan skala antar variabel menjadi lebih sebanding, sehingga proses pemodelan *machine learning* dapat berjalan dengan lebih optimal dan stabil.

2.2.13 Evaluasi Model Regresi

Evaluasi model regresi dilakukan untuk menilai sejauh mana model mampu menjelaskan pola data dan menghasilkan prediksi yang akurat. Penilaian ini menggunakan metrik statistik yang mengukur tingkat kesalahan prediksi dan kecocokan model terhadap data, sehingga kinerja model dapat dievaluasi secara objektif dan konsisten (Assara & Setiawan, 2025; Hartarti et al., 2025; Made et al., 2023,).

1. Koefisien Determinasi (R^2)

Koefisien Determinasi (R^2) adalah metrik ini mengukur seberapa besar persentase variasi pada variabel dependen (Y) yang dapat dijelaskan oleh variabel independen (X) dalam model. Nilai R^2 berkisar antara 0 hingga 1, semakin mendekati 1, semakin baik model dalam menjelaskan variasi data, sementara nilai mendekati 0 menunjukkan kemampuan model yang lemah (Damodar N. Gujarati & Dawn C. Porter, 2009a).

Berikut adalah rumusnya:

$$R^2 = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \quad (2.6)$$

Dimana,

- a) R^2 : Koefisien Determinasi, metrik yang mengukur seberapa besar variasi variabel dependen yang dapat dijelaskan oleh model.
- b) Y_i : Nilai aktual (nilai asli) dari observasi ke-i.
- c) $\hat{Y}_i$: Nilai yang diprediksi oleh model untuk observasi ke-i.

2. Mean Squared Error (MSE)

Mean Squared Error (MSE) adalah metrik evaluasi yang mengukur rata-rata kuadrat selisih antara nilai aktual dan nilai prediksi model regresi. Karena menggunakan pengkuadratan, MSE memberikan penalti lebih besar pada kesalahan yang besar. Nilai MSE yang lebih kecil menunjukkan kemampuan prediksi model yang lebih baik.

Rumus MSE dinyatakan sebagai berikut:

$$MSE = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \quad (2.7)$$

Di mana:

- a) Y_i adalah nilai aktual observasi ke-i,
- b) $\hat{Y}_i$ adalah nilai prediksi observasi ke-i,
- c) n adalah jumlah observasi.

3. Root Mean Squared Error (RMSE)

Root Mean Squared Error (RMSE) mengukur besar kesalahan prediksi model melalui akar kuadrat rata-rata selisih kuadrat antara nilai prediksi dan nilai aktual. Karena menggunakan pengkuadratan, RMSE sensitif terhadap kesalahan besar. Nilai RMSE yang lebih kecil menunjukkan kemampuan prediksi model yang lebih baik (Chicco et al., 2021).

Berikut adalah rumusnya:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2} \quad (2.8)$$

Dimana,

- a) *RMSE: Root Mean Squared Error*, metrik yang mengukur standar deviasi dari galat (kesalahan) prediksi.
 - b) n : Jumlah total observasi atau titik data.
 - c) Σ : Simbol Sigma, yang berarti "jumlah total dari".
 - d) Y_i : Nilai aktual (nilai asli) dari observasi ke- i .
 - e) $\hat{Y}_i$: Nilai yang diprediksi oleh model untuk observasi ke- i .
4. *Mean Absolute Error (MAE)*

Mean Absolute Error (MAE) adalah metrik evaluasi yang mengukur rata-rata nilai absolut dari selisih antara nilai aktual dan nilai prediksi. Berbeda dengan MSE dan RMSE, MAE tidak menggunakan pengkuadratan sehingga setiap kesalahan diperlakukan secara proporsional dan lebih mudah diinterpretasikan karena berada pada satuan yang sama dengan variabel target. Nilai MAE yang lebih kecil menunjukkan kinerja prediksi model yang lebih baik (Van Pham & Nguyen-Ngoc, 2022).

Rumus MAE dinyatakan sebagai berikut:

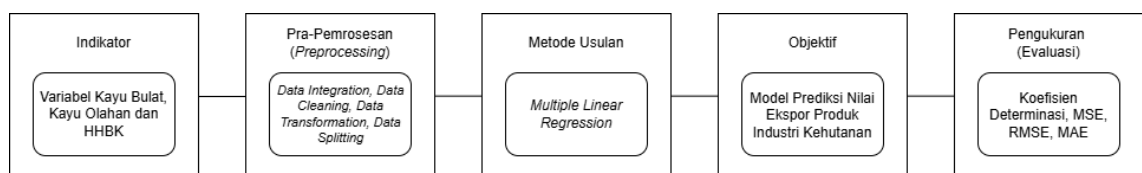
$$MAE = \frac{1}{n} \sum_{i=1}^n |Y_i - \hat{Y}_i| \quad (2.9)$$

Di mana:

- a) Y_i adalah nilai aktual observasi ke- i ,
- b) $\hat{Y}_i$ adalah nilai prediksi observasi ke- i ,
- c) n adalah jumlah observasi.

2.3 Kerangka Pemikiran

Bagian ini menjelaskan alur pemikiran penelitian yang dimulai dari tahap persiapan data hingga evaluasi model sebagai dasar penarikan kesimpulan.



Gambar 2. 5 Kerangka Pemikiran Penelitian

Kerangka pemikiran penelitian divisualisasikan dalam Gambar 2.5, yang menggambarkan keterkaitan antara setiap komponen utama penelitian.

1. Indikator

Data yang digunakan meliputi produk industri kehutanan berupa kayu bulat, kayu olahan, dan HHBK sebagai prediktor serta nilai ekspor sebagai variabel target yang disusun pada tingkat provinsi dan periode bulanan dari tahun 2020–2024.

2. Pra-Pemrosesan (*Preprocessing*)

Persiapan data dimulai dari menggabungkan data ekspor dan produksi berdasarkan provinsi, tahun, dan bulan, pembersihan data termasuk penanganan *missing value*, transformasi variabel kategorikal dengan *one-hot encoding* untuk provinsi dan bulan serta penskalaan variabel numerik, lalu pembagian data berbasis waktu dengan 2020–2023 sebagai data latih dan 2024 sebagai data uji.

3. Metode Usulan

Model yang digunakan adalah regresi linier berganda (*multiple linear regression*) karena dapat memprediksi nilai ekspor dari beberapa variabel produksi secara bersamaan, bersifat sederhana dan transparan, serta koefisiennya mudah diinterpretasikan.

4. Tujuan Penelitian (Objektif)

Penelitian bertujuan membangun model prediksi nilai ekspor produk industri kehutanan berbasis data historis produksi untuk mendukung analisis kinerja dan pengambilan keputusan ekspor bagi Kementerian Kehutanan.

5. Evaluasi Model (*Measurement/Evaluation*)

Kinerja model dievaluasi dengan membandingkan nilai aktual dan prediksi menggunakan metrik R^2 , MSE, RMSE, dan MAE untuk menilai kemampuan penjelasan model dan besar kesalahan prediksi.