

BAB 2

LANDASAN TEORI

Bab ini membahas tinjauan pustaka dan landasan teori yang menjadi dasar dalam penelitian mengenai prediksi harga mobil listrik menggunakan metode regresi linear berbasis machine learning. Pembahasan pada bab ini mencakup perkembangan kendaraan listrik di Eropa beserta karakteristik, faktor-faktor yang mempengaruhi harga mobil listrik, serta teori dan metode yang digunakan dalam membangun model prediksi. Landasan teori disusun untuk memberikan pemahaman yang komprehensif mengenai konteks penelitian, memperkuat dasar pemilihan metode, serta menjelaskan tahapan-tahapan yang dilakukan dalam proses pengolahan data dan pemodelan prediksi harga mobil listrik.

2.1 Konteks Pasar Kendaraan Listrik di Eropa

Dalam beberapa tahun terakhir, adopsi kendaraan listrik di kawasan Eropa mengalami peningkatan yang signifikan. Pertumbuhan ini didorong oleh kebijakan pemerintah Uni Eropa yang berfokus pada transisi energi, pengurangan emisi karbon, serta komitmen terhadap target netralitas karbon. Regulasi ketat terkait emisi kendaraan bermotor mendorong produsen otomotif untuk beralih dari kendaraan berbahan bakar fosil menuju kendaraan berbasis listrik [16].

Selain kebijakan lingkungan, pengembangan infrastruktur pengisian daya yang semakin luas dan merata turut meningkatkan tingkat adopsi kendaraan listrik. Investasi pada stasiun pengisian daya publik, termasuk pengisian cepat, meningkatkan kepercayaan konsumen dalam penggunaan kendaraan listrik secara jangka panjang [17]. Kondisi ini menjadikan Eropa sebagai salah satu pasar kendaraan listrik terbesar di dunia.

Seiring dengan meningkatnya jumlah kendaraan listrik baru, pasar kendaraan listrik bekas di Eropa juga mengalami pertumbuhan. Kendaraan listrik bekas menjadi alternatif bagi konsumen yang ingin beralih ke kendaraan ramah lingkungan dengan harga yang lebih terjangkau dibandingkan kendaraan baru [17]. Namun, pasar kendaraan listrik bekas memiliki karakteristik yang berbeda dibandingkan kendaraan konvensional karena harga jualnya dipengaruhi oleh faktor teknis yang lebih spesifik, terutama yang berkaitan dengan baterai [18].

Harga kendaraan listrik bekas dipengaruhi oleh berbagai faktor teknis,

seperti kapasitas baterai, jarak tempuh maksimum, efisiensi energi, kemampuan pengisian daya, performa kendaraan, serta usia pemakaian. Faktor-faktor tersebut berkaitan langsung dengan kondisi dan degradasi baterai yang berpengaruh terhadap performa kendaraan [19]. Semakin lama usia kendaraan, potensi penurunan kapasitas baterai akan semakin besar, sehingga berdampak pada penurunan nilai jual kendaraan.

Depresiasi harga pada kendaraan listrik juga memiliki karakteristik yang berbeda dibandingkan kendaraan berbahan bakar konvensional. Kondisi dan degradasi baterai menjadi faktor utama yang memengaruhi laju depresiasi kendaraan listrik [20]. Studi menunjukkan bahwa tingkat depresiasi kendaraan listrik dapat mencapai sekitar 1,16 persen per bulan, lebih tinggi dibandingkan kendaraan konvensional yang berada pada kisaran 0,87 persen per bulan [21]. Mengingat baterai merupakan komponen dengan biaya terbesar pada kendaraan listrik, perubahan kondisi baterai secara langsung memengaruhi nilai jual kembali kendaraan [22]. Oleh karena itu, pemodelan harga kendaraan listrik bekas berbasis data menjadi penting untuk merepresentasikan hubungan antara spesifikasi teknis kendaraan dan harga pasar secara objektif.

2.2 Machine Learning

Machine learning merupakan pendekatan komputasional yang memungkinkan sistem untuk mempelajari pola dari data historis dan menghasilkan prediksi terhadap data baru. Dalam konteks prediksi nilai numerik, machine learning digunakan untuk memodelkan hubungan antara variabel input dan variabel target secara matematis. Penelitian ini menggunakan pendekatan *supervised learning*, di mana model dilatih menggunakan data yang telah memiliki label berupa harga kendaraan listrik. Pendekatan ini sesuai digunakan pada permasalahan prediksi harga karena tujuan utama model adalah memetakan hubungan antara spesifikasi teknis kendaraan dan harga pasar berdasarkan data historis.

Beberapa proses umum dalam machine learning meliputi pengumpulan data, pembersihan data, pemilihan fitur, pelatihan model, evaluasi, dan penerapan model. Jenis machine learning yang digunakan dalam penelitian ini termasuk dalam kategori supervised learning, yaitu model yang dilatih menggunakan data berlabel dimana input dan target output sudah diketahui dari awal [23].

2.3 Supervised Learning

Supervised learning adalah pendekatan pembelajaran mesin yang menggunakan dataset dengan pasangan input–output. Tujuan utama metode ini adalah membangun sebuah fungsi yang mampu memetakan input menjadi output secara akurat berdasarkan contoh data yang diberikan.

Supervised learning terbagi menjadi dua jenis utama:

1. Klasifikasi, digunakan ketika output berupa kategori.
2. Regresi, digunakan ketika output berupa nilai numerik atau kontinu.

Penelitian ini menggunakan pendekatan regresi karena variabel yang diprediksi adalah harga mobil bekas dalam bentuk angka [24].

2.4 Regresi

Regresi adalah metode statistik yang bertujuan menganalisis hubungan antara variabel independen (fitur) dengan variabel dependen (target). Dengan regresi, peneliti dapat mengetahui seberapa besar pengaruh suatu variabel terhadap variabel lain, serta dapat membangun model untuk memprediksi nilai baru berdasarkan pola yang sudah dipelajari dari data historis. Model regresi umumnya digunakan dalam bidang ekonomi, bisnis, dan teknik, termasuk dalam prediksi harga barang seperti rumah, properti, dan kendaraan [5].

2.5 Linear Regression

Regresi linear merupakan salah satu metode supervised learning yang digunakan untuk memodelkan hubungan linier antara satu atau lebih variabel independen terhadap variabel dependen yang bersifat kontinu. Dengan kata lain, perubahan kecil pada salah satu variabel input akan memengaruhi output secara proporsional [25]. Dalam penelitian ini, regresi linear digunakan untuk memprediksi harga kendaraan listrik berdasarkan beberapa spesifikasi teknis kendaraan. Oleh karena itu, model yang digunakan adalah *Multiple Linear Regression*, di mana harga kendaraan dipengaruhi oleh lebih dari satu variabel independen secara simultan.

2.6 Multiple Linear Regression

Multiple Linear Regression (MLR) merupakan pengembangan dari regresi linear sederhana yang melibatkan lebih dari satu variabel independen. Model ini sesuai digunakan ketika variabel target dipengaruhi oleh berbagai faktor, seperti pada kasus harga kendaraan yang dipengaruhi oleh kapasitas baterai, jarak tempuh, usia kendaraan, dan spesifikasi teknis lainnya.

Pembentukan model *Multiple Linear Regression* (MLR) dalam penelitian ini dimulai dari representasi hubungan dasar antara variabel dependen dan variabel independen. Secara matematis, hubungan tersebut diasumsikan bersifat linear dan dapat dinyatakan sebagai berikut.

$$y = f(x_1, x_2, \dots, x_n) \quad (2.1)$$

Untuk memodelkan hubungan tersebut secara kuantitatif, fungsi $f(\cdot)$ diasumsikan berbentuk kombinasi linear dari setiap variabel independen, sehingga diperoleh persamaan regresi linear sebagai berikut.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n \quad (2.2)$$

Karena data observasi pada dunia nyata mengandung ketidakpastian dan noise, maka ke dalam model ditambahkan komponen error yang merepresentasikan selisih antara nilai prediksi dan nilai aktual. Dengan demikian, persamaan regresi linear berganda dinyatakan sebagai berikut.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + \epsilon \quad (2.3)$$

Untuk memudahkan proses estimasi parameter pada banyak observasi, model regresi linear berganda tersebut direpresentasikan dalam bentuk matriks sebagai berikut.

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon \quad (2.4)$$

Parameter β kemudian diestimasi menggunakan metode *Ordinary Least Squares* (OLS), yang bertujuan untuk meminimalkan jumlah kuadrat residual. Fungsi objektif OLS dirumuskan sebagai berikut.

$$S(\beta) = (\mathbf{Y} - \mathbf{X}\beta)^\top (\mathbf{Y} - \mathbf{X}\beta) \quad (2.5)$$

Dengan melakukan turunan parsial terhadap β dan menyamakannya dengan nol, diperoleh persamaan normal sebagai dasar perhitungan parameter regresi.

$$\mathbf{X}^T \mathbf{X} \beta = \mathbf{X}^T \mathbf{Y} \quad (2.6)$$

Apabila matriks $\mathbf{X}^T \mathbf{X}$ bersifat invertible, maka estimasi parameter regresi dapat diperoleh sebagai berikut.

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} \quad (2.7)$$

Nilai estimasi parameter tersebut selanjutnya disubstitusikan ke dalam model regresi sehingga menghasilkan persamaan akhir *Multiple Linear Regression* yang digunakan dalam penelitian ini.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + \varepsilon \quad (2.8)$$

Keterangan:

- Y : variabel dependen yang diprediksi, yaitu harga kendaraan listrik bekas.
- X_1, X_2, \dots, X_n : variabel independen yang digunakan dalam model regresi, meliputi karakteristik teknis kendaraan listrik.
- β_0 : konstanta atau *intercept* yang menyatakan nilai Y ketika seluruh variabel independen bernilai nol.
- $\beta_1, \beta_2, \dots, \beta_n$: koefisien regresi yang menunjukkan besar dan arah pengaruh masing-masing variabel independen terhadap variabel dependen.
- ε : error atau residual yang merepresentasikan selisih antara nilai aktual dan nilai prediksi yang dihasilkan oleh model.

2.7 Estimasi Parameter Regresi Linear

Estimasi parameter pada multiple linear regression dilakukan menggunakan metode *Ordinary Least Squares* (OLS). Metode ini bertujuan untuk menentukan nilai koefisien regresi yang meminimalkan jumlah kuadrat selisih antara nilai aktual dan nilai prediksi model. Metode OLS dipilih karena kesederhanaannya dan efektivitasnya dalam permasalahan regresi linier, khususnya ketika hubungan antar variabel diasumsikan bersifat linier [26].

Fungsi objektif metode *Ordinary Least Squares* (OLS) dirumuskan sebagai berikut.

$$\min \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (2.9)$$

di mana y_i merupakan nilai aktual, \hat{y}_i merupakan nilai prediksi, dan n merupakan jumlah data. Dengan meminimalkan fungsi tersebut, metode OLS menghasilkan estimasi koefisien regresi yang paling optimal dalam merepresentasikan hubungan linier antara variabel independen dan variabel dependen.

2.8 Keterkaitan Regresi Linear dengan Penelitian

Penggunaan multiple linear regression dalam penelitian ini didasarkan pada asumsi bahwa harga kendaraan listrik dapat direpresentasikan sebagai kombinasi linier dari beberapa spesifikasi teknis kendaraan. Pendekatan ini memungkinkan analisis hubungan antara variabel input dan harga kendaraan tanpa kompleksitas model non-linear. Model regresi linear juga memberikan keunggulan dalam interpretasi hasil, di mana koefisien regresi dapat digunakan untuk memahami arah hubungan antara spesifikasi kendaraan dan harga, sehingga sesuai dengan tujuan penelitian yang berfokus pada representasi hubungan antar variabel.

2.9 Evaluasi Model Regresi

Evaluasi model dilakukan untuk mengukur sejauh mana hasil prediksi harga kendaraan listrik bekas mendekati nilai harga aktual pada data pengujian. Dalam konteks prediksi harga, evaluasi tidak hanya berfokus pada akurasi model, tetapi juga pada besarnya kesalahan prediksi dalam satuan moneter yang relevan bagi pengguna dan pelaku pasar. Oleh karena itu, penelitian ini menggunakan beberapa metrik evaluasi berbasis error, yaitu Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), serta Mean Absolute Percentage Error (MAPE).

Koefisien determinasi (R^2) tidak digunakan sebagai metrik kesalahan utama karena nilai R^2 hanya menunjukkan proporsi variasi variabel dependen yang dapat dijelaskan oleh model, bukan besarnya error prediksi secara langsung. Dalam prediksi harga kendaraan listrik bekas, metrik berbasis error lebih representatif untuk menggambarkan ketepatan estimasi harga yang dihasilkan oleh model.

2.9.1 Mean Absolute Error (MAE)

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (2.10)$$

MAE mengukur rata-rata nilai absolut selisih antara harga aktual (y_i) dan harga hasil prediksi (\hat{y}_i). Metrik ini memiliki keunggulan karena berada pada satuan yang sama dengan variabel harga, sehingga mudah diinterpretasikan secara langsung. Dalam penelitian ini, nilai MAE merepresentasikan rata-rata kesalahan prediksi harga kendaraan listrik bekas, sehingga dapat memberikan gambaran yang intuitif mengenai ketepatan model.

2.9.2 Mean Squared Error (MSE)

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (2.11)$$

MSE menghitung rata-rata kuadrat selisih antara nilai aktual dan nilai prediksi. Dengan melakukan pengkuadratan error, metrik ini memberikan penalti yang lebih besar terhadap kesalahan prediksi yang ekstrem. Hal ini penting dalam konteks prediksi harga kendaraan listrik bekas, karena kesalahan prediksi yang terlalu besar dapat menyebabkan estimasi harga yang tidak realistis dan berpotensi menyesatkan pengambilan keputusan.

2.9.3 Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{MSE} \quad (2.12)$$

RMSE merupakan akar kuadrat dari MSE sehingga nilai error kembali ke satuan yang sama dengan variabel harga. RMSE sering digunakan sebagai metrik evaluasi utama karena sensitif terhadap error besar, namun tetap mudah diinterpretasikan. Dalam penelitian ini, RMSE digunakan untuk menilai stabilitas model dalam memprediksi harga kendaraan listrik bekas, khususnya pada data dengan variasi harga yang tinggi.

2.9.4 Mean Absolute Percentage Error (MAPE)

$$MAPE = \frac{100\%}{n} \sum_{i=1}^n \frac{|y_i - \hat{y}_i|}{y_i} \quad (2.13)$$

MAPE mengukur rata-rata kesalahan prediksi dalam bentuk persentase terhadap nilai aktual. Metrik ini memberikan gambaran kesalahan relatif model dan memungkinkan perbandingan performa model pada skala harga yang berbeda. Dalam penelitian ini, MAPE digunakan untuk menilai tingkat kesalahan prediksi harga kendaraan listrik bekas secara proporsional, sehingga dapat memberikan indikasi tingkat keandalan model dalam konteks pasar kendaraan listrik.

2.9.5 Confidence Interval Prediksi

Selain metrik error seperti MAE, MSE, RMSE, dan MAPE, dalam evaluasi model regresi linear juga penting untuk mengetahui tingkat kepercayaan prediksi yang dihasilkan, terutama ketika model digunakan untuk memprediksi data baru. Confidence interval (CI) memberikan rentang nilai di sekitar prediksi di mana harga aktual kemungkinan besar berada dengan tingkat keyakinan tertentu, misalnya 95%.

Untuk model regresi linear, confidence interval untuk prediksi individual dihitung menggunakan rumus berikut:

$$\hat{y}_{\text{new}} \pm t_{\alpha/2, df} \cdot s_{\text{err}} \cdot \sqrt{1 + x_{\text{new}}^T (X^T X)^{-1} x_{\text{new}}} \quad (2.14)$$

Keterangan:

- \hat{y}_{new} : prediksi harga untuk data baru
- $t_{\alpha/2, df}$: nilai t-Student untuk tingkat kepercayaan $(1 - \alpha)$ dan derajat bebas $df = n - p - 1$
- s_{err} : residual standard error, dihitung dari data training
- x_{new} : vektor fitur termasuk intercept dari data baru
- X : matriks fitur training termasuk intercept

Residual standard error (s_{err}) dihitung dengan rumus:

$$s_{\text{err}} = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - p - 1}} \quad (2.15)$$

Dengan menggunakan confidence interval, peneliti dapat menilai stabilitas dan keandalan prediksi model, sehingga tidak hanya mengandalkan nilai prediksi tunggal (\hat{y}) tetapi juga mempertimbangkan kemungkinan variasi harga aktual. Hal ini penting dalam konteks estimasi harga mobil listrik bekas, karena harga aktual dapat dipengaruhi oleh faktor-faktor pasar yang tidak sepenuhnya tercakup dalam model.

2.10 Dataset Mobil Listrik

Dataset yang digunakan dalam penelitian ini merupakan dataset kendaraan listrik yang diperoleh dari platform publik seperti kaggle. Dataset tersebut berisi data mobil listrik yang dipasarkan di wilayah Eropa dan mencakup berbagai atribut teknis kendaraan yang relevan dengan analisis harga.

Atribut yang tersedia dalam dataset meliputi:

- merek dan model kendaraan,
- performa akselerasi,
- kecepatan maksimum,
- jarak tempuh maksimum (*driving range*),
- efisiensi energi (Wh/km),
- kemampuan pengisian cepat (*fast charging*),
- sistem penggerak (*powertrain*),
- jumlah kursi,
- harga kendaraan listrik dalam satuan Euro.

Perlu diperhatikan bahwa dataset tidak menyediakan informasi eksplisit mengenai tahun produksi kendaraan. Namun, berdasarkan karakteristik model yang tercantum, seluruh kendaraan dalam dataset merupakan model modern yang umumnya dirilis dan dipasarkan pada rentang tahun 2018 hingga 2024. Oleh karena itu, penelitian ini menetapkan asumsi tahun produksi 2021 sebagai tahun referensi untuk seluruh kendaraan guna menjaga konsistensi variabel.

Dengan asumsi tersebut, variabel usia kendaraan dianggap bernilai konstan sehingga tidak dimasukkan sebagai fitur dalam proses pemodelan. Pendekatan

ini dilakukan untuk menghindari bias akibat ketidaklengkapan data sekaligus memastikan bahwa model regresi linear hanya memanfaatkan variabel yang benar-benar tersedia dan relevan.

2.11 Data Preprocessing

Sebelum data digunakan dalam proses pelatihan model, diperlukan beberapa tahap preprocessing untuk memastikan kualitas data tetap terjaga.

- **Handling Missing Values**

Menangani data yang hilang dengan menghapus atau mengisi nilai yang kosong.

- **Encoding Data Kategorikal**

Mengubah data kategorikal menjadi format numerik menggunakan metode seperti label encoding atau one-hot encoding.

- **Normalisasi atau Standarisasi**

Menyamakan skala antar variabel agar model dapat belajar dengan lebih stabil.

- **Outlier Handling**

Menghapus atau menangani data ekstrem yang dapat menyebabkan bias pada model.

- **Train-Test Split**

Memisahkan data menjadi data latih dan data uji untuk mengevaluasi performa model secara objektif.

Tahapan preprocessing ini penting agar data yang digunakan memiliki kualitas baik dan menghasilkan model prediksi yang lebih akurat.