Implementasi Regresi Linear Berganda dengan Metode Normal Equation untuk Prediksi Harga Rumah di Jakarta Selatan

Muhamad Dimas Saputra
Teknik Informatika/Sains dan Teknologi
UNIVERSITAS ISLAM NEGERI SYARIF HIDAYATULLAH JAKARTA
Tangerang Selatan, Indonesia
muhamad.dimas24@mhs.uinjkt.ac.id

***Abstract*—Sektor properti memiliki peran vital dalam ekonomi, namun penentuan harga rumah seringkali menjadi tantangan kompleks karena dipengaruhi oleh banyak faktor fisik. Penelitian ini bertujuan untuk mengimplementasikan metode Regresi Linear Berganda menggunakan pendekatan *Normal Equation* untuk memprediksi harga rumah. Berbeda dengan pendekatan iteratif, *Normal Equation* memberikan solusi analitik langsung untuk menemukan parameter model yang optimal dengan meminimalkan fungsi biaya secara matematis tanpa memerlukan *learning rate*. Data yang digunakan berasal dari dataset harga rumah di Jakarta Selatan yang diperoleh dari repositori GitHub, mencakup 1001 data transaksi dengan fitur luas tanah, luas bangunan, jumlah kamar tidur, jumlah kamar mandi, dan ketersediaan garasi. Hasil pengujian menunjukkan bahwa model mampu menghasilkan prediksi dengan nilai *R-squared* sebesar 0.836, yang mengindikasikan bahwa model dapat menjelaskan 83.6% variasi harga properti di wilayah tersebut, dengan tingkat kesalahan rata-rata (MAPE) sebesar 27.07%.**

***Index Terms*—Prediksi Harga Rumah, Regresi Linear, Normal Equation, Jakarta Selatan, Machine Learning, R-squared.**

I. Pendahuluan

A. Latar Belakang

Rumah merupakan kebutuhan primer yang sekaligus menjadi instrumen investasi strategis karena nilainya yang cenderung meningkat seiring waktu. Namun, harga rumah sangat fluktuatif dan dipengaruhi oleh berbagai variabel fisik seperti luas tanah, luas bangunan, serta fasilitas penunjang lainnya. Calon pembeli maupun penjual seringkali kesulitan menentukan harga wajar tanpa analisis data yang akurat [1].

Untuk mengatasi permasalahan tersebut, pendekatan *Machine Learning* khususnya Regresi Linear Berganda sering digunakan untuk memodelkan hubungan antara fitur properti dan harganya [2]. Dalam menyelesaikan masalah regresi linear, terdapat dua pendekatan utama: metode iteratif (seperti *Gradient Descent*) dan metode analitik (seperti *Normal Equation*). Meskipun *Gradient Descent* populer untuk data berskala sangat besar, *Normal Equation* menawarkan keunggulan berupa solusi eksak yang langsung meminimalkan fungsi biaya tanpa perlu menentukan parameter *learning rate* atau melakukan iterasi berulang, terutama untuk dataset dengan ukuran menengah [3].

Penelitian ini berfokus pada implementasi Regresi Linear Berganda menggunakan metode *Normal Equation* untuk memprediksi harga rumah di kawasan Jakarta Selatan. Dataset yang digunakan melalui tahapan pra-pemrosesan, termasuk pembersihan data (*cleaning*) dan normalisasi fitur menggunakan *Z-score*, sebelum diproses menggunakan operasi matriks untuk mendapatkan model prediksi terbaik.

B. Rumusan Masalah

Berdasarkan latar belakang di atas, rumusan masalah dalam penelitian ini adalah:

1) Bagaimana memodelkan hubungan antara variabel independen (luas tanah, luas bangunan, jumlah kamar tidur, jumlah kamar mandi, garasi) terhadap harga rumah di Jakarta Selatan?
2) Bagaimana implementasi solusi analitik *Normal Equation* ($\theta = (X^T X)^{-1} X^T y$) dalam meminimalkan galat prediksi?
3) Seberapa akurat kinerja model yang dihasilkan berdasarkan metrik *R-squared* (R^2), MSE, dan MAPE?

C. Tujuan Penelitian

Tujuan yang ingin dicapai dalam penelitian ini adalah:

1) Menganalisis karakteristik data historis properti untuk menentukan variabel-variabel yang signifikan dalam pembentukan harga.
2) Menerapkan pendekatan numerik untuk memodelkan data historis menjadi fungsi prediksi yang mampu menghasilkan estimasi harga properti yang wajar dan objektif.
3) Mengukur tingkat akurasi model yang dihasilkan dalam menaksir harga dibandingkan dengan data aktual.

D. Batasan Masalah

Agar pembahasan lebih terarah, penulis menetapkan batasan masalah sebagai berikut:

- **Metode Algoritma:** Penelitian menggunakan *Normal Equation* sebagai penyelesaian matematis untuk Regresi Linear, diimplementasikan menggunakan pustaka NumPy.

- **Dataset:** Data yang digunakan adalah dataset "HARGA RUMAH JAKSEL" yang terdiri dari 1001 baris data valid setelah pembersihan, bersumber dari repositori GitHub publik.
- **Evaluasi:** Fokus evaluasi pada metrik akurasi numerik dan tidak membahas faktor eksternal seperti inflasi atau kebijakan pemerintah.

II. Landasan Teori

A. Teori Harga Hedonik

Dasar pemodelan harga properti dalam penelitian ini mengacu pada Teori Harga Hedonik. Menurut Rosen [4], barang-barang heterogen seperti properti dinilai berdasarkan sekumpulan karakteristik atau atribut yang melekat padanya. Harga pasar dari sebuah properti (P) dipandang sebagai penjumlahan dari harga implisit masing-masing atributnya.

Secara konseptual, jika sebuah properti dideskripsikan oleh vektor karakteristik $Z = (z_1, z_2, \ldots, z_n)$, maka fungsi harganya adalah $P(Z) = P(z_1, z_2, \ldots, z_n)$. Teori ini menjadi landasan bahwa data historis yang memuat atribut fisik (seperti luas tanah dan jumlah kamar) dapat digunakan untuk menaksir harga wajar properti tersebut secara objektif.

B. Regresi Linear Berganda (Multiple Linear Regression)

Untuk memodelkan hubungan antara variabel dependen (harga) dan berbagai variabel independen (fitur properti), digunakan metode Regresi Linear Berganda. Sebagaimana dijelaskan oleh Walpole et al. [5], model regresi linear berganda dengan k variabel independen didefinisikan sebagai:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i, \quad i = 1, \ldots, n \quad (1)$$

Dimana:

- y_i adalah variabel respon (harga properti) ke-i.
- $\beta_0, \beta_1, \ldots, \beta_k$ adalah parameter regresi yang akan diestimasi.
- x_{ij} adalah nilai variabel prediktor ke-j pada observasi ke-i.
- ε_i adalah galat acak (*random error*) yang diasumsikan berdistribusi normal dengan rata-rata nol.

C. Pendekatan Matriks dalam Metode Numerik

Dalam komputasi numerik, penanganan data dalam jumlah besar dilakukan menggunakan operasi matriks untuk efisiensi komputasi. Menurut Chapra dan Canale [6], sistem persamaan linear untuk regresi berganda dapat direpresentasikan dalam bentuk matriks:

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon \quad (2)$$

Dimana $\mathbf{y}$ adalah vektor observasi harga berukuran $n \times 1$, $\mathbf{X}$ adalah matriks desain berukuran $n \times (k+1)$ yang memuat data fitur, β adalah vektor koefisien berukuran $(k+1) \times 1$, dan ε adalah vektor galat.

D. Metode Kuadrat Terkecil (Ordinary Least Squares)

Untuk mendapatkan estimasi parameter β yang menghasilkan garis regresi terbaik (*best fit*), digunakan Metode Kuadrat Terkecil (OLS). Prinsip dari metode ini adalah meminimalkan Jumlah Kuadrat Galat (*Sum of Squared Errors* - SSE). Fungsi objektif S yang harus diminimalkan adalah:

$$S = \sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon^T \varepsilon = (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta) \quad (3)$$

Untuk meminimalkan S, turunan parsial terhadap β disamakan dengan nol. Chapra dan Canale [6] menjabarkan bahwa solusi analitik dari optimasi ini menghasilkan **Persamaan Normal** (*Normal Equations*):

$$(\mathbf{X}^T \mathbf{X})\hat{\beta} = \mathbf{X}^T \mathbf{y} \quad (4)$$

Dengan asumsi bahwa matriks $(\mathbf{X}^T \mathbf{X})$ bersifat non-singular (memiliki invers), maka estimasi koefisien $\hat{\beta}$ dapat dihitung secara langsung menggunakan operasi aljabar linear:

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \quad (5)$$

Persamaan (5) inilah yang diimplementasikan menggunakan pustaka NumPy dalam penelitian ini untuk menghitung bobot model prediksi harga secara efisien tanpa proses iterasi.

III. Metodologi Penelitian

A. Alur Penelitian

Penelitian ini dilakukan mengikuti tahapan standar *Data Science* yang meliputi pengumpulan data, pra-pemrosesan, transformasi fitur, pemodelan matematis, dan evaluasi. Alur kerja secara keseluruhan diimplementasikan menggunakan bahasa pemrograman Python dengan pustaka Pandas untuk manajemen data dan NumPy untuk komputasi aljabar linear.

B. Pengumpulan dan Pemahaman Data

Dataset yang digunakan dalam penelitian ini adalah data sekunder "Harga Rumah Jakarta Selatan" yang bersumber dari repositori publik GitHub. Data mentah diunduh dalam format *Comma Separated Values* (CSV). Dataset ini terdiri dari 1001 entri data transaksi properti dengan 7 atribut (kolom) awal, yaitu:

- **NV**: Nomor urut atau indeks data (tidak digunakan dalam pemodelan).
- **LT**: Luas Tanah (m^2).
- **LB**: Luas Bangunan (m^2).
- **JKT**: Jumlah Kamar Tidur.
- **JKM**: Jumlah Kamar Mandi.
- **GRS**: Kapasitas Garasi (Kategorikal: "ADA"/"TIDAK ADA" atau numerik).
- **HARGA**: Harga transaksi rumah (Variabel Target/Dependen).

C. Pra-pemrosesan Data (Data Preprocessing)

Sebelum data dapat diproses oleh algoritma *Normal Equation*, dilakukan serangkaian tahapan pembersihan dan transformasi:

1) **Pembersihan Data**: Memastikan tidak ada nilai yang hilang (*missing values*) atau duplikasi baris yang dapat membiaskan hasil prediksi.
2) **Konversi Tipe Data**: Memastikan seluruh kolom fitur (LT, LB, JKT, JKM) memiliki tipe data numerik (*integer*) agar dapat dihitung secara matematis.
3) **Encoding Variabel Kategorikal**: Fitur 'GRS' (Garasi) yang semula berisi data teks ditransformasi menjadi biner numerik. Nilai "ADA" dikonversi menjadi 1, dan nilai lainnya (atau "TIDAK ADA") dikonversi menjadi 0.
4) **Seleksi Fitur**: Kolom 'NV' dan 'KOTA' (jika ada) dihapus karena tidak relevan dengan kalkulasi numerik harga. Variabel independen ($\mathbf{X}$) yang terpilih adalah: [LT, LB, JKT, JKM, GRS].

D. Pembagian Data Latih dan Uji

Untuk menguji generalisasi model, dataset dibagi menjadi dua bagian:

- **Data Latih (*Training Set*)**: Sebesar 90% dari total data (900 data), digunakan untuk menghitung parameter model (θ).
- **Data Uji (*Testing Set*)**: Sebesar 10% dari total data (101 data), digunakan untuk validasi kinerja model.

Pembagian dilakukan dengan pengacakan (*shuffling*) menggunakan *random state* 831 untuk memastikan konsistensi hasil eksperimen setiap kali kode dijalankan.

E. Implementasi Normal Equation dengan NumPy

Inti dari penelitian ini adalah implementasi algoritma *Normal Equation* tanpa menggunakan pustaka *machine learning* tingkat tinggi (seperti Scikit-Learn), melainkan menggunakan operasi matriks murni dengan NumPy. Langkah-langkah komputasinya adalah sebagai berikut:

1) Penambahan Bias Term: Dalam persamaan garis linear $y = mx + c$, terdapat konstanta c (intersep). Dalam bentuk matriks, hal ini direpresentasikan dengan menambahkan kolom yang berisi angka 1 (satu) di awal matriks fitur $\mathbf{X}$.

$$\mathbf{X}_{bias} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{bmatrix} \tag{6}$$

Penambahan ini dilakukan menggunakan fungsi `np.concatenate` atau `np.c_` pada pustaka NumPy.

2) Perhitungan Parameter Theta: Vektor parameter optimal θ (yang berisi bobot $\beta_0, \beta_1, \dots, \beta_k$) dihitung menggunakan rumus:

$$\theta = (\mathbf{X}_{bias}^T \mathbf{X}_{bias})^{-1} \mathbf{X}_{bias}^T \mathbf{y}_{train} \tag{7}$$

Dalam kode Python, operasi ini diterjemahkan menjadi:

```
theta = np.linalg.inv(X.T.dot(X)).dot(X.T).dot(y)
```

3) Fungsi Prediksi: Setelah nilai θ didapatkan, prediksi harga ($\hat{y}$) untuk data baru dilakukan dengan operasi perkalian titik (*dot product*):

$$\hat{y} = \mathbf{X}_{test\_bias} \cdot \theta \tag{8}$$

F. Metrik Evaluasi

Untuk mengukur keberhasilan model, digunakan beberapa metrik statistik standar:

- **Mean Squared Error (MSE)**: Mengukur rata-rata kuadrat kesalahan.

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{9}$$

- **Root Mean Squared Error (RMSE)**: Akar dari MSE, memberikan gambaran kesalahan dalam satuan Rupiah.
- **R-squared (R^2)**: Koefisien determinasi yang menunjukkan seberapa baik variabel independen menjelaskan variasi variabel dependen.

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \tag{10}$$

- **Mean Absolute Percentage Error (MAPE)**: Persentase rata-rata kesalahan prediksi terhadap nilai aktual.

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100\% \tag{11}$$

IV. Hasil dan Pembahasan

A. Analisis Statistik Deskriptif dan Korelasi

Sebelum dilakukan pemodelan, analisis korelasi dilakukan untuk melihat hubungan linear antara variabel independen dan variabel dependen (Harga). Berdasarkan matriks korelasi (*Correlation Matrix*) pada tahap eksplorasi data, ditemukan pola hubungan sebagai berikut:

- **Luas Tanah (LT) dan Luas Bangunan (LB)**: Memiliki korelasi positif yang paling kuat terhadap harga. Hal ini wajar mengingat di wilayah Jakarta Selatan, komponen nilai tanah (*land value*) seringkali lebih dominan dibandingkan nilai bangunan itu sendiri.
- **Jumlah Kamar (JKT & JKM)**: Memiliki korelasi positif moderat. Semakin banyak kamar, harga cenderung naik, namun tidak sesignifikan luas tanah.
- **Garasi**: Menunjukkan korelasi positif lemah namun relevan, mengindikasikan bahwa properti dengan fasilitas garasi memiliki nilai tambah tersendiri.

B. Evaluasi Kinerja Model

Model Regresi Linear yang dibangun menggunakan metode *Normal Equation* diuji menggunakan 101 data uji (X_{test}) yang tidak pernah dilihat model selama proses pelatihan. Ringkasan kinerja model disajikan pada Tabel I.

TABLE I
METRIK EVALUASI KINERJA MODEL

Metrik Evaluasi	Nilai Hasil Pengujian
Mean Squared Error (MSE)	1.577×10^{19}
Root Mean Squared Error (RMSE)	Rp 3.971.650.938
R-squared (R^2)	0.836
MAPE	27.07%

1) Interpretasi R-squared (R^2): Nilai R^2 sebesar **0.836** mengindikasikan bahwa **83.6%** variabilitas atau keragaman harga rumah di Jakarta Selatan dapat dijelaskan oleh model ini menggunakan lima fitur fisik (LT, LB, JKT, JKM, GRS). Sisanya sebesar 16.4% dipengaruhi oleh faktor-faktor lain yang tidak tercakup dalam dataset, seperti:

- Lokasi spesifik (misal: jarak ke jalan raya utama atau MRT).
- Kondisi lingkungan (misal: bebas banjir atau keamanan).
- Kualitas interior (misal: lantai marmer vs keramik biasa).

Angka 0.836 tergolong *strong goodness of fit* untuk data properti yang sifatnya sangat heterogen.

2) Analisis Tingkat Kesalahan (RMSE dan MAPE): Nilai RMSE sebesar Rp 3.97 Miliar menunjukkan standar deviasi dari sisaan (*residuals*). Meskipun angka ini terlihat besar secara nominal, hal ini harus dilihat dalam konteks harga properti Jakarta Selatan yang rentang harganya sangat lebar (mulai dari Rp 1 Miliar hingga di atas Rp 100 Miliar).

Metrik MAPE sebesar **27.07%** memberikan gambaran yang lebih proporsional. Artinya, jika sebuah rumah memiliki harga asli Rp 10 Miliar, model cenderung memprediksi harga di kisaran Rp 7.3 Miliar hingga Rp 12.7 Miliar. Tingkat kesalahan ini masih dapat diterima untuk tujuan estimasi awal (*preliminary valuation*), namun memerlukan penyesuaian manual (*appraisal*) untuk penentuan harga transaksi final.

C. Komparasi Nilai Aktual vs Prediksi

Untuk memvisualisasikan akurasi model, grafik sebar (*scatter plot*) antara nilai aktual (y) dan nilai prediksi ($\hat{y}$) menunjukkan pola linear yang kuat. Titik-titik data berkumpul di sekitar garis diagonal 45°, yang menandakan bahwa prediksi model cukup konsisten mengikuti tren harga asli.

Namun, terdapat sedikit bias pada data dengan harga ekstrem tinggi (properti *luxury* di atas Rp 50 Miliar). Model cenderung melakukan *under-prediction* (prediksi lebih rendah dari harga asli) pada segmen ini. Hal ini disebabkan oleh keterbatasan Regresi Linear dalam menangkap faktor "kemewahan" non-linear yang biasanya mendongkrak harga properti kelas atas secara eksponensial.

D. Diskusi Koefisien Model (Nilai Theta)

Berdasarkan hasil perhitungan *Normal Equation*, diperoleh vektor bobot θ. Secara matematis, model ini memberitahu kita "harga wajar" per unit fitur.

- Koefisien untuk **Luas Tanah** memiliki bobot positif terbesar. Ini mengonfirmasi bahwa di Jakarta Selatan, tanah adalah aset yang paling berharga. Setiap penambahan $1m^2$ tanah akan meningkatkan harga properti secara signifikan.
- Koefisien **Garasi** bernilai positif, yang berarti keberadaan garasi secara statistik menaikkan harga jual dibandingkan rumah tanpa garasi, dengan asumsi luas tanah dan bangunan sama.

E. Kelebihan dan Kelemahan Pendekatan Normal Equation

Dalam eksperimen ini, penggunaan *Normal Equation* terbukti lebih efisien dibandingkan *Gradient Descent* karena:

1) **Tanpa Iterasi**: Komputasi langsung selesai dalam satu langkah eksekusi matriks, sangat cepat untuk 1000 data (kurang dari 1 detik).
2) **Tanpa Tuning**: Tidak perlu mencari *learning rate* (α) yang tepat, menghilangkan risiko divergensi.

Namun, kelemahannya adalah kompleksitas komputasi matriks inversi $O(n^3)$. Jika data bertambah menjadi jutaan baris di masa depan, metode ini akan menjadi sangat lambat dan memakan memori, sehingga *Gradient Descent* akan menjadi pilihan yang lebih baik nantinya.

V. KESIMPULAN

Berdasarkan hasil eksperimen, dapat disimpulkan bahwa:

1) Metode *Normal Equation* terbukti efektif dan efisien untuk memprediksi harga rumah pada dataset berukuran menengah (1000 data) karena tidak memerlukan tuning parameter *learning rate*.
2) Fitur fisik (Luas Tanah, Luas Bangunan, Jumlah Kamar) memiliki korelasi yang sangat kuat dengan harga rumah di Jakarta Selatan, dibuktikan dengan skor R^2 mencapai 0.836.
3) Model ini dapat digunakan sebagai alat estimasi awal (*first-pass estimation*) bagi agen properti atau pembeli, dengan catatan adanya margin error rata-rata sebesar 27%.

REFERENCES

[1] R. N. T. Siregar, V. Sitorus, and W. P. Ananta, "Analisis prediksi harga rumah di bandung menggunakan regresi linear berganda," *Journal of Creative Student Research (JCSR)*, vol. 1, no. 6, pp. 395–404, 2023.

[2] R. R. Hallan and I. N. Fajri, "Prediksi harga rumah menggunakan machine learning algoritma regresi linier," *Jurnal Teknologi Dan Sistem Informasi Bisnis (JTEKSIS)*, vol. 7, no. 1, pp. 57–62, 2025.

[3] A. Ng and T. Ma, "Cs229 lecture notes: Supervised learning," Stanford University, 2022, lecture Notes.

[4] S. Rosen, "Hedonic prices and implicit markets: product differentiation in pure competition," *Journal of political economy*, vol. 82, no. 1, pp. 34–55, 1974.

[5] R. E. Walpole, R. H. Myers, S. L. Myers, and K. Ye, *Probability and Statistics for Engineers and Scientists*, 9th ed. Boston, MA: Pearson Education, 2012.

[6] S. C. Chapra and R. P. Canale, *Numerical Methods for Engineers*, 7th ed. New York, NY: McGraw-Hill Education, 2015.