Implementasi Metode Numerik Menggunakan Pendekatan 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 **algoritma numerik** berbasis *Normal Equation* untuk menyelesaikan masalah prediksi harga rumah. Berbeda dengan pendekatan iteratif heuristik, *Normal Equation* memberikan solusi analitik eksak untuk menemukan parameter model optimal (θ) melalui operasi matriks langsung, meminimalkan fungsi biaya secara matematis tanpa memerlukan penentuan *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 eksperimen komputasi menunjukkan bahwa metode ini 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*—Metode Numerik, Normal Equation, Regresi Linear, Komputasi Matriks, Jakarta Selatan, 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 secara matematis [1].

Untuk mengatasi permasalahan tersebut, pendekatan **Metode Numerik** digunakan untuk memformulasikan hubungan antara fitur properti dan harganya ke dalam model persamaan matematis [2]. Dalam domain komputasi numerik, penyelesaian masalah regresi ini dapat dilakukan melalui dua pendekatan utama: metode iteratif (seperti *Gradient Descent*) dan metode analitik langsung (seperti *Normal Equation*). Meskipun *Gradient Descent* populer untuk data masif, **Normal Equation** menawarkan keunggulan berupa solusi tertutup (*closed-form solution*) yang langsung meminimalkan galat dalam satu langkah komputasi matriks, menjadikannya sangat efisien untuk dataset berukuran menengah [3].

Penelitian ini berfokus pada implementasi algoritma **Normal Equation** menggunakan pustaka numerik (NumPy) untuk memprediksi harga rumah di kawasan Jakarta Selatan. Dataset melalui tahapan pra-pemrosesan, termasuk pembersihan data (*cleaning*) dan normalisasi, sebelum diproses menggunakan operasi aljabar linear untuk mendapatkan koefisien model terbaik.

B. Rumusan Masalah

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

1) Bagaimana memformulasikan masalah prediksi harga rumah di Jakarta Selatan ke dalam model persamaan linear menggunakan variabel fisik yang tersedia?
2) Bagaimana implementasi solusi numerik *Normal Equation* ($\theta = (X^T X)^{-1} X^T y$) untuk meminimalkan galat prediksi secara efisien?
3) Seberapa akurat kinerja model komputasi 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 signifikan dalam persamaan model.
2) Menerapkan pendekatan numerik untuk memodelkan data historis menjadi fungsi prediksi yang mampu menghasilkan estimasi harga properti yang wajar.
3) Mengukur tingkat akurasi algoritma 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, diimplementasikan menggunakan pustaka Python 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 dapat digunakan untuk menaksir harga wajar properti tersebut secara objektif.

B. Formulasi Model Linear

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

$$y_i = \theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2} + \cdots + \theta_k x_{ik} + \varepsilon_i, \quad i = 1, \ldots, n \quad (1)$$

Dimana:

- y_i adalah variabel respon (harga properti) ke-i.
- $\theta_0, \theta_1, \ldots, \theta_k$ adalah parameter bobot (koefisien) yang akan dihitung.
- 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:

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

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, θ adalah vektor koefisien berukuran $(k+1) \times 1$, dan ε adalah vektor galat.

D. Solusi Numerik: Metode Kuadrat Terkecil

Untuk mendapatkan estimasi parameter θ 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 $J(\theta)$ yang harus diminimalkan adalah:

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

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

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

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

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

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 komputasi sains 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 numerik, dilakukan serangkaian tahapan pembersihan dan transformasi:

1) **Pembersihan Data**: Memastikan tidak ada nilai yang hilang (*missing values*) atau duplikasi baris yang dapat membiaskan hasil komputasi.
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' 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 $\theta_0, \theta_1, \dots, \theta_k$) 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 @ X) @ X.T @ 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 komputasi, 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 Eksplorasi Data (Exploratory Data Analysis)

Sebelum dilakukan komputasi numerik, analisis korelasi antar variabel dilakukan untuk memahami faktor-faktor yang paling mempengaruhi harga rumah.

Fig. 1. Matriks Korelasi Fitur terhadap Harga Rumah. Warna yang lebih terang menunjukkan korelasi positif yang lebih kuat.

Berdasarkan Gambar 1, dapat dilihat bahwa variabel **Luas Tanah (LT)** memiliki koefisien korelasi tertinggi terhadap harga (mendekati 0.81). Hal ini mengindikasikan bahwa dalam penentuan harga properti di Jakarta Selatan, komponen tanah memegang peranan yang lebih dominan dibandingkan luas bangunan atau jumlah kamar.

B. Evaluasi Kinerja Model Numerik

Model yang dibangun menggunakan *Normal Equation* dievaluasi menggunakan data uji. Ringkasan metrik akurasi disajikan pada Tabel I.

TABLE I
METRIK EVALUASI KINERJA MODEL

Metrik	Nilai
R^2 Score	0.836
RMSE (Rupiah)	3.97 Miliar
MAPE (Persentase Error)	27.07%

1) Analisis Linearitas (Prediksi vs Aktual): Untuk memvalidasi akurasi model secara visual, dilakukan pemetaan antara harga aktual (sumbu X) dan harga prediksi model (sumbu Y).

Fig. 2. Grafik Sebar (*Scatter Plot*) antara Harga Aktual dan Prediksi. Garis putus-putus merah merepresentasikan prediksi sempurna ($y = x$).

Sebagaimana ditunjukkan pada Gambar 2, titik-titik data (biru) terkonsentrasi di sekitar garis diagonal merah. Hal ini menandakan bahwa model mampu memprediksi harga dengan cukup presisi pada rentang harga Rp 1 Miliar hingga Rp 20 Miliar. Namun, terlihat adanya penyimpangan (*deviation*) yang semakin melebar pada properti *luxury* dengan harga di atas Rp 50 Miliar, di mana model cenderung memprediksi lebih rendah (*under-prediction*) dari harga aslinya.

C. Analisis Residual (Error Analysis)

Dalam metode numerik, analisis sisaan (*residual*) sangat krusial untuk memastikan bahwa model memenuhi asumsi statistik yang mendasarinya.

1) Uji Heteroskedastisitas: Grafik residual digunakan untuk melihat apakah varian error bersifat konstan atau berubah seiring besarnya nilai prediksi.

Pada Gambar 3, terlihat pola menyerupai "corong" yang melebar ke arah kanan. Ini menunjukkan gejala **Heteroskedastisitas**, artinya tingkat kesalahan prediksi model semakin besar (dalam nominal Rupiah) ketika memprediksi rumah yang harganya semakin mahal. Fenomena ini wajar terjadi pada data properti karena variansi fitur pada rumah mewah jauh lebih kompleks daripada rumah standar.

Fig. 3. Plot Residual. Sumbu Y menunjukkan selisih antara harga asli dan prediksi. Pola melebar ke kanan mengindikasikan adanya Heteroskedastisitas.

2) Uji Normalitas Error: Histogram distribusi error dibuat untuk melihat karakteristik statistik dari kesalahan prediksi.

Fig. 4. Histogram Distribusi Error. Kurva berbentuk lonceng menunjukkan error berdistribusi normal.

Gambar 4 memperlihatkan kurva berbentuk lonceng simetris (*bell-shaped curve*) yang terpusat di angka nol. Hal ini mengonfirmasi bahwa galat (*error*) model berdistribusi normal, sehingga metode *Least Squares* yang digunakan pada *Normal Equation* adalah estimator yang valid secara statistik.

D. Tinjauan Sampel Kasus

Untuk memberikan gambaran konkret mengenai performa model di dunia nyata, Tabel II menampilkan perbandingan harga asli dan prediksi pada 5 data acak dari data uji.

Dari tabel sampel di atas, terlihat bahwa untuk rumah kelas menengah (luas tanah $< 300m^2$), selisih prediksi relatif kecil. Namun, pada rumah besar (luas tanah $1000m^2$), selisih prediksi menjadi cukup signifikan, yang konsisten dengan temuan pada analisis residual sebelumnya.

TABLE II
SAMPEL PERBANDINGAN HARGA ASLI VS PREDIKSI

Luas Tanah	Harga Asli	Prediksi	Selisih
120 m^2	Rp 2.5 M	Rp 2.7 M	+ Rp 0.2 M
300 m^2	Rp 8.0 M	Rp 7.2 M	- Rp 0.8 M
60 m^2	Rp 1.2 M	Rp 1.3 M	+ Rp 0.1 M
500 m^2	Rp 15.0 M	Rp 13.5 M	- Rp 1.5 M
1000 m^2	Rp 45.0 M	Rp 38.0 M	- Rp 7.0 M

V. KESIMPULAN DAN SARAN

A. Kesimpulan

Berdasarkan hasil implementasi dan evaluasi algoritma numerik *Normal Equation* terhadap data harga rumah di Jakarta Selatan, dapat ditarik beberapa kesimpulan analitis sebagai berikut:

1) **Efektivitas Metode Analitik**: Pendekatan *Normal Equation* terbukti sangat efisien untuk dataset berukuran menengah (1001 data), di mana solusi optimal (θ) dapat diperoleh secara instan melalui komputasi matriks tanpa memerlukan proses iterasi (*looping*) atau penyesuaian *hyperparameter*.
2) **Dominasi Fitur Fisik**: Model berhasil menjelaskan **83.6%** variasi harga properti ($R^2 = 0.836$). Hal ini mengonfirmasi secara kuantitatif bahwa variabel fisik—terutama **Luas Tanah** dan **Luas Bangunan**—adalah determinan utama nilai properti di Jakarta Selatan. Sisa varians sebesar 16.4% kemungkinan besar dipengaruhi oleh faktor eksternal tak terukur dalam dataset ini, seperti prestise lokasi, aksesibilitas jalan, dan risiko banjir.
3) **Batasan Model Linear pada Data Heterogen**: Meskipun akurasi global cukup tinggi, analisis residual menunjukkan adanya gejala **Heteroskedastisitas**. Model cenderung memiliki tingkat kesalahan yang membesar (MAPE rata-rata 27.07%) pada segmen properti mewah (*luxury*). Ini menunjukkan bahwa hubungan antara fitur fisik dan harga pada level harga ekstrem tidak sepenuhnya linear.

B. Saran

Untuk pengembangan penelitian selanjutnya dan peningkatan akurasi model, penulis menyarankan:

- **Penambahan Fitur Lokasi**: Mengingat jargon properti "Lokasi, Lokasi, Lokasi", disarankan untuk menambahkan fitur geospasial yang lebih spesifik, seperti "Kecamatan", "Jarak ke MRT", atau "Lebar Jalan Depan".
- **Transformasi Data**: Untuk mengatasi masalah heteroskedastisitas pada rumah mewah, penelitian selanjutnya dapat menerapkan transformasi logaritma pada variabel target ($\log(y)$) atau menggunakan pendekatan model non-linear.
- **Regularisasi**: Penerapan teknik regularisasi (seperti *Ridge* atau *Lasso*) dapat dipertimbangkan dalam formulasi matriks untuk mengurangi potensi *overfitting* jika jumlah fitur diperbanyak.

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.