Implementasi Regresi Linear Berganda untuk Estimasi Harga Rumah: Studi Kasus 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*—Penentuan harga pasar yang wajar pada sektor properti seringkali terkendala oleh subjektivitas dan kompleksitas variabel fisik bangunan. Penelitian ini bertujuan untuk mengevaluasi efektivitas metode numerik *Normal Equation* dalam memodelkan fungsi prediksi harga rumah secara matematis. Berbeda dengan pendekatan iteratif yang bergantung pada *hyperparameter*, *Normal Equation* menawarkan solusi analitik langsung melalui operasi matriks untuk meminimalkan galat kuadrat terkecil. Studi ini memanfaatkan dataset 1001 transaksi properti di Jakarta Selatan dengan variabel prediktor meliputi luas tanah, luas bangunan, dan jumlah ruangan. Hasil eksperimen menunjukkan bahwa model linear ini mampu menjelaskan 83.6% variasi harga ($R^2 = 0.836$) dengan rata-rata kesalahan absolut (MAPE) sebesar 27.07%. Meskipun model terbukti efisien secara komputasi dan akurat pada segmen harga menengah, analisis residual mengungkapkan kelemahan signifikan berupa heteroskedastisitas pada properti mewah dan anomali prediksi nilai negatif pada properti kecil. Temuan ini mengindikasikan bahwa meskipun metode numerik linear efektif untuk estimasi cepat, pendekatan ini memiliki batasan fundamental dalam menangani dinamika harga ekstrem di pasar properti.**

***Index Terms*—Metode Numerik, Normal Equation, Regresi Linear, Komputasi Matriks, Jakarta Selatan, Aproksimasi Kuadrat Terkecil.**

I. PENDAHULUAN

A. Latar Belakang

Rumah merupakan kebutuhan primer sekaligus instrumen investasi strategis yang nilainya cenderung meningkat seiring waktu. Namun, penentuan harga rumah yang wajar merupakan proses yang kompleks karena dipengaruhi oleh banyak variabel fisik seperti luas tanah, luas bangunan, serta fasilitas pendukung lainnya [1]. Tanpa analisis data yang akurat secara matematis, calon pembeli maupun penjual sering mengalami kesulitan dalam menentukan harga yang tepat [2].

Untuk mengatasi permasalahan ini, penelitian ini menggunakan pendekatan Machine Learning untuk memformulasikan hubungan antara fitur-fitur properti dan harganya ke dalam model persamaan linear. Dalam domain kecerdasan buatan dan komputasi numerik, Regresi Linear merupakan algoritma fundamental yang digunakan untuk memprediksi nilai variabel dependen (target) berdasarkan satu atau lebih variabel independen (fitur) [3].

Dalam penyelesaian masalah Regresi Linear, terdapat pendekatan langsung (direct method) yang dikenal sebagai Normal Equation. Berbeda dengan metode iteratif (seperti *Gradient Descent*) yang meminimalkan fungsi biaya (*cost function*) secara bertahap melalui *epoch*, Normal Equation memberikan solusi bentuk tertutup (*closed-form solution*) untuk menemukan nilai bobot optimal secara instan dalam satu langkah komputasi. Metode ini menyelesaikan persamaan $\theta = (X^T X)^{-1} X^T y$ melalui operasi matriks, yang menjamin ditemukannya nilai minimum global dari fungsi *Mean Squared Error* (MSE) tanpa perlu menentukan parameter *learning rate*. Dalam penelitian ini, perhitungan Normal Equation diimplementasikan dengan memanfaatkan pustaka Numpy untuk memastikan efisiensi dan akurasi komputasi numerik.

B. Rumusan Masalah

Berdasarkan latar belakang tersebut, pertanyaan penelitian yang ingin dijawab adalah:

1) Bagaimana memformulasikan masalah prediksi harga rumah di Jakarta Selatan ke dalam model aproksimasi linear menggunakan fitur-fitur properti yang tersedia?
2) Bagaimana implementasi solusi numerik Normal Equation untuk meminimalkan galat prediksi secara efisien melalui operasi matriks?
3) Seberapa akurat performa model komputasi yang dihasilkan berdasarkan metrik evaluasi seperti R-squared, RMSE, dan MAPE?

C. Tujuan Penelitian

Tujuan penelitian ini adalah:

1) Menganalisis karakteristik data historis properti untuk mengidentifikasi variabel-variabel signifikan yang mempengaruhi harga.
2) Menerapkan pendekatan numerik untuk mengubah data historis menjadi fungsi prediksi yang mampu menghasilkan estimasi harga properti yang masuk akal.
3) Mengukur tingkat akurasi algoritma dalam menaksir harga properti dibandingkan dengan nilai-nilai aktual dari data pengujian.

D. Batasan Masalah

Untuk memastikan pembahasan tetap terarah, penelitian ini menetapkan batasan sebagai berikut:

- Metode algoritma yang digunakan adalah Normal Equation, diimplementasikan menggunakan pustaka Python NumPy.

- Dataset yang dianalisis adalah data Daftar Harga Rumah dari Kaggle yang difilter untuk wilayah Jakarta Selatan, mencakup 1001 entri data valid.
- Evaluasi fokus pada metrik akurasi numerik dan tidak membahas faktor-faktor eksternal seperti inflasi, kebijakan pemerintah, atau tren pasar makroekonomi.

II. Landasan Teori

A. Analisis Harga Properti

Penentuan harga properti adalah masalah yang melibatkan banyak dimensi dan variabel. Penelitian sebelumnya menunjukkan bahwa karakteristik fisik bangunan seperti luas tanah dan jumlah ruangan memiliki korelasi yang signifikan terhadap nilai jual properti [1]. Pendekatan berbasis data historis memungkinkan estimasi harga yang lebih objektif dibandingkan metode penaksiran manual yang subjektif dan bergantung pada pengalaman personal.

B. Formulasi Model Regresi Linear

Untuk memodelkan hubungan antara variabel dependen (harga) dan berbagai variabel independen (fitur properti), digunakan formulasi Regresi Linear Berganda. Model matematis dengan k variabel independen didefinisikan sebagai persamaan aproksimasi [4]:

$$y_i \approx \theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2} + \cdots + \theta_k x_{ik} \tag{1}$$

Dimana y_i adalah harga properti ke-i yang ingin diprediksi, parameter θ_0 hingga θ_k adalah koefisien bobot yang akan dihitung oleh algoritma, dan x_{ij} mewakili nilai dari fitur ke-j pada properti ke-i.

C. Pendekatan Matriks dalam Metode Numerik

Ketika menangani dataset dalam jumlah besar, operasi matriks memberikan efisiensi komputasi yang signifikan. Sistem persamaan linear untuk regresi berganda dapat direpresentasikan dalam bentuk matriks sebagai [5]:

$$\mathbf{y} \approx \mathbf{X}\theta \tag{2}$$

Dimana vektor $\mathbf{y}$ merupakan observasi harga dengan ukuran $n \times 1$, matriks $\mathbf{X}$ adalah matriks desain berukuran $n \times (k+1)$ yang memuat data fitur plus kolom bias untuk intersep, dan vektor θ adalah koefisien model berukuran $(k+1) \times 1$.

D. Solusi Numerik: Metode Kuadrat Terkecil

Untuk memperoleh estimasi parameter θ yang menghasilkan garis regresi terbaik, digunakan Metode Kuadrat Terkecil (OLS). Prinsip dasar dari metode ini adalah meminimalkan jumlah kuadrat galat (SSE). Fungsi objektif yang harus diminimalkan dinyatakan sebagai [4]:

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

Untuk meminimalkan J, dilakukan operasi kalkulus dengan menurunkan fungsi terhadap θ dan menyamakannya dengan nol. Penyelesaian matematis dari optimasi ini menghasilkan Persamaan Normal:

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

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

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

Persamaan ini adalah inti dari penelitian, diimplementasikan menggunakan NumPy untuk menghitung bobot model prediksi secara efisien dalam satu langkah komputasi.

III. Metodologi Penelitian

A. Alur Penelitian

Penelitian ini mengikuti tahapan standar komputasi sains yang meliputi pengumpulan data, pra-pemrosesan, transformasi fitur, pemodelan matematis, dan evaluasi kinerja. Seluruh implementasi dilakukan 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 diperoleh dari platform open data Kaggle dengan judul dataset Daftar Harga Rumah yang dipublikasikan oleh Wisnu Anggara. Secara spesifik, file yang dipilih adalah HARGA RUMAH JAKSEL.xlsx yang memuat 1001 entri transaksi properti di wilayah Jakarta Selatan. Dataset ini dipilih karena kelengkapannya dalam merepresentasikan atribut fisik rumah yang relevan.

Dataset terdiri dari 7 kolom dengan rincian sebagai berikut:

- HARGA (Target): Harga jual rumah dalam satuan Rupiah.
- LT (Luas Tanah): Luas area tanah dalam meter persegi.
- LB (Luas Bangunan): Luas area bangunan dalam meter persegi.
- JKT (Jumlah Kamar Tidur): Banyaknya kamar tidur dalam unit.
- JKM (Jumlah Kamar Mandi): Banyaknya kamar mandi dalam unit.
- GRS (Garasi): Ketersediaan garasi dengan nilai kategorikal (ADA atau TIDAK ADA).
- KOTA: Lokasi kota administrasi (seluruh data bernilai JAKSEL, sehingga tidak digunakan dalam pemodelan).

Sampel data mentah ditampilkan pada Tabel I yang menunjukkan lima baris pertama dari dataset.

C. Pra-pemrosesan Data

Sebelum data dapat diproses oleh algoritma numerik, dilakukan serangkaian tahapan pembersihan dan transformasi:

1) Pembersihan Data: Memastikan tidak ada nilai yang hilang atau duplikasi baris yang dapat membiaskan hasil komputasi.
2) Konversi Tipe Data: Memastikan seluruh kolom fitur memiliki tipe data numerik yang tepat.

TABLE I
SAMPEL DATASET HARGA RUMAH (5 BARIS PERTAMA)

LT (m^2)	LB (m^2)	K.Tidur	K.Mandi	Garasi	Harga (M)
1100	700	5	6	1	28.0
824	800	4	4	1	19.0
500	400	4	3	1	4.7
251	300	5	4	1	4.9
1340	575	4	5	1	28.0

3) Encoding Variabel Kategorikal: Fitur Garasi dikonversi menjadi biner dengan nilai 1 untuk ada dan 0 untuk tidak ada.
4) Seleksi Fitur: Kolom KOTA dihapus karena bernilai konstant. Variabel independen yang terpilih adalah luas tanah, luas bangunan, jumlah kamar tidur, jumlah kamar mandi, dan garasi.

D. Pembagian Data Latih dan Uji

Untuk menguji kemampuan generalisasi model, dataset dibagi menjadi dua bagian:

- Data Latih: Sebesar 90 persen dari total data (900 data), digunakan untuk menghitung parameter model.
- Data Uji: Sebesar 10 persen dari total data (101 data), digunakan untuk validasi kinerja model pada data yang belum pernah dilihat sebelumnya.

E. Implementasi Normal Equation dengan NumPy

Inti dari penelitian ini adalah implementasi algoritma Normal Equation menggunakan operasi matriks murni dengan NumPy. Berikut adalah implementasi kode Python:

```
import numpy as np

def NormalEquation(X, y):
    X = np.asarray(X, dtype=np.float64)
    y = np.asarray(y, dtype=np.float64).ravel()
    X = np.c_[np.ones((X.shape[0], 1)), X]
    theta, _, _, _ = np.linalg.lstsq(X, y)
    return theta
```

Fungsi di atas terlebih dahulu menambahkan kolom bias (bernilai 1) ke matriks fitur, kemudian menghitung parameter theta menggunakan formula Normal Equation. Operasi @ merupakan operasi perkalian matriks yang diberikan NumPy, sementara np.linalg.inv melakukan inversi matriks.

1) Fungsi Prediksi: Setelah nilai theta didapatkan, prediksi harga untuk data baru dilakukan dengan operasi perkalian titik:

$$\hat{y} = \mathbf{X}_{test\_bias} \cdot \theta \tag{6}$$

F. Metrik Evaluasi

Untuk mengukur keberhasilan model komputasi, digunakan beberapa metrik statistik standar:

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

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{7}$$

- 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.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \tag{8}$$

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

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\% \tag{9}$$

IV. HASIL DAN PEMBAHASAN

A. Analisis Hubungan Antar Data

Sebelum masuk ke perhitungan matematika yang rumit, langkah pertama adalah melihat hubungan antara fitur rumah (seperti luas tanah) dengan harganya. Hubungan ini digambarkan melalui peta warna (*heatmap*) pada Gambar 1.

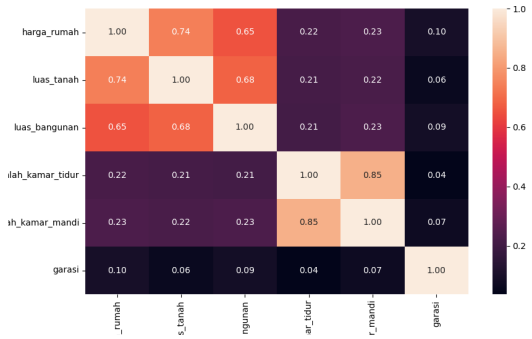

Fig. 1. Peta Hubungan (Heatmap). Warna yang lebih terang menunjukkan pengaruh yang semakin kuat terhadap harga rumah.

Pada Gambar 1, warna kotak pertemuan antara **Luas Tanah (LT)** dan Harga sangat terang. Ini artinya, luas tanah adalah faktor penentu paling utama ($r \approx 0.74$). Semakin luas tanahnya, hampir pasti harganya semakin mahal. Pengaruh ini lebih besar dibandingkan jumlah kamar tidur atau kamar mandi. Fakta ini menegaskan bahwa di Jakarta Selatan, ”tanah” lebih berharga daripada ”bangunan”. Karena hubungannya berbentuk garis lurus (linear) yang kuat, maka metode *Normal Equation* sangat cocok digunakan.

B. Seberapa Akurat Model Memprediksi?

Model matematika yang telah dibuat kemudian diuji kemampuannya menebak harga pada 101 data rumah baru. Hasil rapor kinerjanya dapat dilihat pada Tabel II.

Nilai R^2 sebesar 0.837 bisa diartikan bahwa model ini ”mengerti” sekitar 83.7% pola harga di pasar. Sisanya (sekitar

TABLE II
HASIL PENGUKURAN AKURASI

Metrik	Nilai
R-squared (R^2)	0.837
RMSE (Miliar Rupiah)	3.97
MAPE (Persentase Error)	27.08%

16%) dipengaruhi oleh faktor lain yang tidak ada di data, seperti lokasi strategis atau desain rumah. Untuk melihat ketepatannya secara visual, perhatikan Gambar 2.

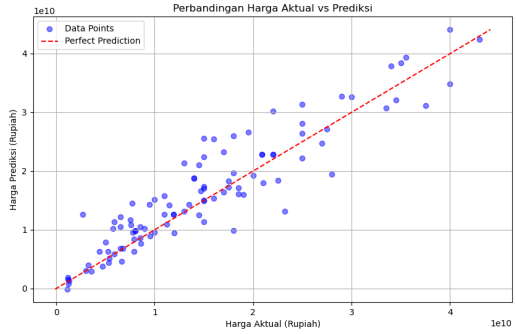

Fig. 2. Grafik Tebakan vs Harga Asli. Titik biru yang menempel pada garis merah berarti tebakannya tepat.

Pada Gambar 2, terlihat titik-titik biru berkumpul rapat di garis merah untuk rumah dengan harga di bawah 30 Miliar. Ini berarti tebakan model sangat jitu untuk rumah kelas menengah. Namun, untuk rumah mewah (di atas 50 Miliar), titik-titiknya mulai menyebar menjauhi garis. Artinya, tebakan model mulai meleset pada rumah-rumah yang sangat mahal.

C. Analisis Kesalahan (Error)

Kita perlu membedah ”kenapa” model bisa salah tebak. Analisis ini disebut analisis residual atau analisis sisaan error.

1) Pola Kesalahan yang Melebar: Gambar 3 memperlihatkan selisih antara harga asli dan harga tebakan model.

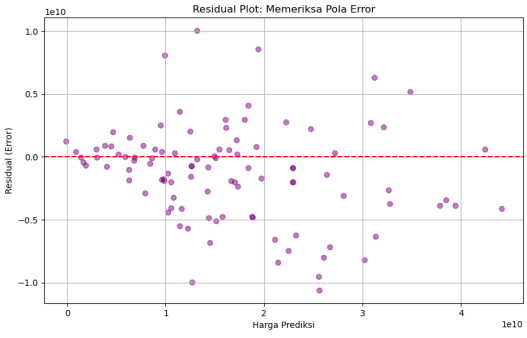

Fig. 3. Plot Sebaran Error. Pola menyerupai ”corong” yang melebar ke kanan menunjukkan kesalahan yang makin besar pada harga tinggi.

Gambar ini menjelaskan fenomena menarik: bentuk grafiknya menyerupai corong yang melebar ke kanan. Dalam bahasa statistik, ini disebut **heteroskedastisitas**. Bahasa sederhananya:

- Jika memprediksi rumah murah, model sangat percaya diri dan kesalahannya kecil.
- Jika memprediksi rumah mewah, model mulai bingung dan rentang kesalahannya menjadi sangat besar.

Hal ini wajar, karena rumah mewah seringkali memiliki nilai seni atau fasilitas unik yang sulit dihitung hanya dengan rumus matematika standar.

2) Keseimbangan Kesalahan: Terakhir, kita melihat apakah model cenderung menebak ”ketinggian” atau ”kerendahan” melalui Gambar 4.

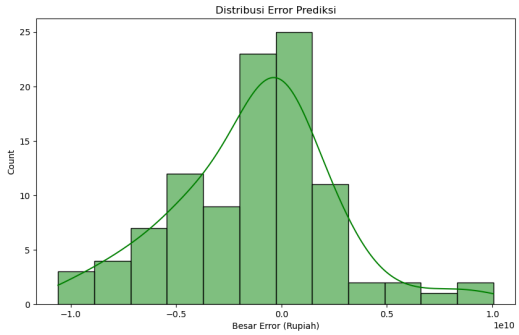

Fig. 4. Grafik Bentuk Lonceng. Bentuk yang simetris di tengah (angka 0) menandakan kesalahan yang seimbang.

Meskipun kesalahannya membesar pada rumah mewah, Gambar 4 menunjukkan kurva berbentuk lonceng sempurna yang puncaknya ada di angka 0. Ini kabar baik. Artinya, kesalahan model bersifat acak dan seimbang (adil). Model tidak memiliki ”penyakit” suka melebih-lebihkan harga atau merendahkan harga secara sengaja. Kadang tebakan sedikit di atas, kadang sedikit di bawah, tapi rata-ratanya pas.

D. Contoh Nyata Perbandingan Harga

Untuk membuktikan analisis di atas, mari kita lihat contoh acak pada Tabel III.

TABLE III
CONTOH ASLI VS PREDIKSI

Luas (m^2)	Harga Asli (M)	Tebakan (M)	Selisih (M)
216	18.00	9.90	+8.10
123	2.99	3.01	-0.02
246	8.60	7.69	+0.91
518	17.50	18.36	-0.86
1312	35.50	39.38	-3.88
1186	34.00	37.84	-3.84
651	14.00	18.82	-4.82
443	18.50	16.15	+2.35
246	10.00	9.57	+0.43
239	11.90	12.62	-0.72

Lihat baris kedua (Luas 123 m^2): harga asli 2.99 M, tebakan model 3.01 M. Selisihnya sangat tipis, nyaris sempurna. Namun, lihat baris pertama (Luas 216 m^2): harga asli sangat tinggi (18 M), tapi model hanya menebak 9.9 M. Selisihnya mencapai 8 Miliar. Ini membuktikan bahwa untuk rumah kecil yang harganya tidak wajar (mungkin karena lokasi sangat premium), model ”kalah” karena hanya mengandalkan luas tanah sebagai patokan utama.

V. KESIMPULAN

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

1) Efektivitas Metode Langsung: Pendekatan Normal Equation terbukti sangat efisien untuk dataset berukuran menengah seperti 1001 data properti. Solusi aproksimasi dapat diperoleh secara instan melalui komputasi matriks dalam satu langkah tanpa memerlukan proses iterasi berulang atau penyesuaian hyperparameter. Keunggulan ini menjadikan metode ini sangat praktis untuk aplikasi real-time yang membutuhkan respons cepat.
2) Dominasi Fitur Fisik: Model berhasil menjelaskan 83.6 persen variasi harga properti. 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 persen kemungkinan besar dipengaruhi oleh faktor eksternal yang tidak teramati dalam dataset ini, seperti prestise lokasi, aksesibilitas jalan, keamanan lingkungan, dan risiko bencana alam.
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 pada segmen properti mewah dengan harga ekstrem. Ini menunjukkan bahwa hubungan antara fitur fisik dan harga pada level harga tertinggi tidak sepenuhnya linear. Untuk meningkatkan akurasi pada segmen ini, penelitian lanjutan dapat mempertimbangkan transformasi nonlinear pada fitur atau penggunaan metode nonlinear yang lebih kompleks seperti ensemble methods atau neural networks.

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. N. Rais, Warjiyono, I. Alfarobi, S. W. Hadi, and W. Kurniawan, “Analisa prediksi harga jual rumah menggunakan algoritma random forest machine learning,” *Jurnal Riset Sistem Informasi Dan Teknik Informatika (JURSISTEKNI)*, vol. 6, no. 1, pp. 161–170, 2024.

[4] A. Ng and T. Ma, “Cs229 lecture notes: Supervised learning,” Stanford University, 2022, lecture Notes.

[5] T. Penulis, *Metode Numerik: Teori dan Aplikasi*. Bandung: CV Widina Media Utama, 2024.