



Penerapan Algoritma Linear Regression untuk Prediksi Harga Rumah Berdasarkan Faktor Lokasi dan Usia Bangunan

Application of Linear Regression Algorithm for House Price Prediction Based on Location and Building Age Factors

Muhamad Aris Purwanto^{1*}, Hasbi Firmansyah², Wahyu Asriyani³, Rizki Prasetyo Tulodo⁴

Universitas Pancasakti Tegal

Email : muhamadarisnurwanto3@gmail.com^{1*}, hasbifirmansyah@upstegal.ac.id², asriyani1409@gmail.com³,

Rizki.prasetyo.tulodo@gmail.com⁴

Article Info

Article history :

Received : 26-12-2025

Revised : 28-12-2025

Accepted : 30-12-2025

Pulished : 03-01-2026

Abstract

Property prices are fluctuating and influenced by many complex factors, making them difficult to predict manually with high accuracy. This study aims to build a house price prediction model per unit area using the Linear Regression algorithm. The dataset used is Real Estate Valuation data consisting of 414 transaction data with predictor variables including transaction date, building age, distance to the nearest MRT station, number of convenience stores, and geographic coordinates (latitude and longitude). The study was conducted using RapidMiner Studio software. The results of the correlation analysis show that distance to the MRT station has the strongest negative relationship with house prices. The resulting linear regression model is able to predict prices with a Root Mean Squared Error (RMSE) level of [RMSE value 8.290 +/- 0.000]. This study concludes that location factors (accessibility) have a more significant influence than building age in determining property prices in this dataset.

Keywords : Data Mining, Linear Regression, House Price Prediction

Abstrak

Harga properti merupakan nilai yang fluktuatif dan dipengaruhi oleh banyak faktor kompleks, sehingga sulit untuk diprediksi secara manual dengan akurasi tinggi. Penelitian ini bertujuan untuk membangun model prediksi harga rumah per satuan luas menggunakan algoritma *Linear Regression*. Dataset yang digunakan adalah data *Real Estate Valuation* yang terdiri dari 414 data transaksi dengan variabel prediktor meliputi tanggal transaksi, usia bangunan, jarak ke stasiun MRT terdekat, jumlah *convenience store*, serta koordinat geografis (latitude dan longitude). Penelitian dilakukan menggunakan perangkat lunak RapidMiner Studio. Hasil analisis korelasi menunjukkan bahwa jarak ke stasiun MRT memiliki hubungan negatif yang paling kuat terhadap harga rumah. Model regresi linear yang dihasilkan mampu memprediksi harga dengan tingkat *Root Mean Squared Error* (RMSE) sebesar [Nilai RMSE 8.290 +/- 0.000]. Penelitian ini menyimpulkan bahwa faktor lokasi (aksesibilitas) memiliki pengaruh lebih signifikan dibandingkan usia bangunan dalam penentuan harga properti pada dataset ini.

Kata Kunci : Data Mining, Linear Regression, Prediksi Harga Rumah

PENDAHULUAN

Properti, khususnya perumahan, merupakan salah satu kebutuhan primer manusia sekaligus instrumen investasi yang memiliki nilai ekonomi tinggi. Dalam satu dekade terakhir, dinamika



harga pasar (*market value*) properti cenderung mengalami fluktuasi yang signifikan. Bagi investor, pengembang, maupun calon pembeli, kemampuan untuk mengestimasi harga rumah secara akurat adalah hal yang krusial. Penilaian harga yang terlalu tinggi (*overpriced*) dapat menyebabkan properti sulit terjual, sedangkan penilaian yang terlalu rendah (*underpriced*) akan merugikan pemilik aset. (Produksi et al., 2019)

Secara tradisional, proses penilaian properti (*property appraisal*) seringkali dilakukan secara manual oleh penilai independen. Proses ini sangat bergantung pada subjektivitas penilai dan survei lapangan yang memakan waktu serta biaya. Selain itu, harga properti dipengaruhi oleh banyak variabel kompleks yang saling berkaitan, mulai dari atribut fisik bangunan (luas, usia bangunan) hingga faktor lokasi (aksesibilitas, kedekatan dengan fasilitas umum). Keterbatasan manusia dalam mengolah variabel-variabel tersebut secara simultan seringkali menyebabkan ketidakakuratan dalam estimasi harga. (Kamal, 2022)

Seiring dengan perkembangan teknologi informasi, teknik *Data Mining* atau Penambangan Data telah menjadi solusi efektif untuk mengatasi masalah tersebut. Dengan memanfaatkan data historis transaksi, komputer dapat dilatih untuk mengenali pola hubungan antara fitur-fitur properti dengan harganya. Salah satu metode statistika yang paling umum dan fundamental dalam prediksi nilai numerik adalah *Linear Regression* (Regresi Linear). Metode ini mampu memodelkan hubungan sebab-akibat antara variabel independen (seperti jarak ke transportasi umum) dan variabel dependen (harga rumah). (Conference, 2019)

Penelitian ini menggunakan dataset publik *Real Estate Valuation* yang mencakup data transaksi properti di distrik Xindian, New Taipei City. Dataset ini unik karena memuat variabel geospasial yang spesifik, yaitu jarak ke stasiun MRT (*Mass Rapid Transit*) dan jumlah *convenience store* di sekitar lokasi, yang diduga memiliki korelasi kuat terhadap harga. Penelitian ini akan menerapkan algoritma *Linear Regression* menggunakan perangkat lunak RapidMiner Studio untuk membangun model prediksi harga rumah serta menganalisis faktor mana yang paling dominan mempengaruhi harga. (Naradiasari & Wahyudi, 2022)

METODE PENELITIAN

1. Alur Penelitian

Penelitian ini dilakukan dengan mengikuti tahapan standar dalam proses *Knowledge Discovery in Database* (KDD) atau *Data Mining*. Tahapan penelitian dirancang secara sistematis mulai dari pengumpulan data hingga evaluasi model. (Han, 2007).

Secara garis besar, alur penelitian terdiri dari langkah-langkah berikut:

- Pengumpulan Data: Mengambil dataset dari sumber data sekunder.
- Pemahaman Data (*Data Understanding*): Menganalisis statistik deskriptif dan korelasi antar atribut.
- Pra-pemrosesan Data (*Data Preprocessing*): Membersihkan dan menyiapkan data agar siap dimodelkan.
- Pemodelan (*Modeling*): Menerapkan algoritma *Linear Regression*.



- e. Evaluasi (*Evaluation*): Mengukur akurasi model menggunakan *Root Mean Squared Error* (RMSE).

2. Sumber Data

Dataset yang digunakan dalam penelitian ini adalah "Real Estate Valuation Data Set" yang diperoleh dari repositori publik (seperti UCI Machine Learning Repository). Data ini berisi data historis pasar *real estate* di distrik Xindian, Kota New Taipei, Taiwan. (Virginia & Alimuddin, 2024)

Dataset terdiri dari 414 data transaksi (instances) dengan 7 atribut utama, yaitu:

Kode	Nama Atribut	Tipe Data	Keterangan (Peran)
X1	Transaction Date	Real	Tanggal transaksi 20.12.917
X2	House Age	Real	Usia bangunan dalam tahun
X3	Distance to MRT	Real	Jarak ke stasiun MRT terdekat (meter)
X4	Convenience Stores	Integer	Jumlah toko di sekitar lokasi (walking distance)
X5	Latitude	Real	Koordinat Lintang
X6	Longitude	Real	Koordinat Bujur
Y	House Price	Real	Label/Target (Harga per unit area)

3. Pra-pemrosesan Data (*Data Preprocessing*)

Sebelum data diolah, dilakukan beberapa penyesuaian menggunakan operator di RapidMiner:

- Seleksi Atribut (*Select Attributes*): Kolom No dihapus dari dataset karena hanya berfungsi sebagai penomoran indeks dan tidak memiliki korelasi logis terhadap harga rumah. Atribut yang dipertahankan adalah X1 hingga X6 dan Y.
- Penetapan Peran (*Set Role*): Atribut Y house price of unit area ditetapkan perannya sebagai Label (variabel dependen yang akan diprediksi). Atribut lainnya (X1-X6) secara otomatis menjadi fitur (variabel independen).
- Pemisahan Data (*Split Data*): Untuk menguji kehandalan model, data dibagi menjadi dua bagian dengan rasio 70:30.
 - 70% (290 data): Digunakan sebagai *Training Data* untuk melatih model.
 - 30% (124 data): Digunakan sebagai *Testing Data* untuk menguji akurasi prediksi.

4. Metode Pemodelan (*Linear Regression*)

Metode yang digunakan adalah Regresi Linear Berganda (*Multiple Linear Regression*). Metode ini dipilih karena variabel target (harga rumah) bersifat numerik kontinu (bukan kategori). Dimana model akan menghitung nilai koefisien (\$b\$) untuk setiap fitur guna meminimalkan selisih antara nilai prediksi dan nilai aktual. Operator *Linear Regression* pada



RapidMiner digunakan dengan pengaturan *feature selection* standar (M5 Prime atau *none* sesuai kebutuhan). (Details, 2025)

5. Pembagian Data (Data Splitting)

Data akan dibagi menjadi dua bagian untuk menguji keandalan model:

- Training Set (80%): Digunakan untuk melatih algoritma dalam mengenali pola antara lokasi, usia, dan harga.
- Testing Set (20%): Digunakan untuk menguji performa model dengan data yang belum pernah dilihat sebelumnya.

6. Evaluasi Model

Untuk mengukur sejauh mana akurasi prediksi, penelitian ini menggunakan tiga metrik evaluasi utama:

Metrik	Definisi
Mean Absolute Error (MAE)	Rata-rata selisih absolut antara harga asli dan prediksi.
Mean Squared Error (MSE)	Rata-rata selisih kuadrat, memberikan penalti lebih besar pada error yang besar.
R-Squared (R^2)	Menunjukkan seberapa besar persentase variasi harga yang dapat dijelaskan oleh lokasi dan usia bangunan.

HASIL DAN PEMBAHASAN

1. Statistik Deskriptif Data

Sebelum dilakukan pemodelan, analisis statistik deskriptif dilakukan untuk memahami karakteristik data. Dataset *Real Estate Valuation* terdiri dari 414 entri data. Berdasarkan eksplorasi data, diketahui bahwa rentang harga rumah (*Y house price of unit area*) sangat bervariasi, mulai dari nilai terendah 7.6 hingga tertinggi 117.5.

Variabel independen juga menunjukkan variasi yang signifikan, terutama pada variabel X3 (Jarak ke MRT) yang memiliki rentang dari lokasi yang sangat dekat dengan stasiun hingga lokasi pinggir yang jauh. Hal ini mengindikasikan bahwa dataset mencakup berbagai kondisi properti, mulai dari pusat kota hingga area sub-urban. (Itsnaini et al., 2025)

2. Analisis Korelasi (*Correlation Matrix*)

Untuk menentukan faktor mana yang paling mempengaruhi harga, dilakukan analisis matriks korelasi. Hasil visualisasi *heatmap* korelasi dari RapidMiner

Berdasarkan matriks di atas, diperoleh temuan sebagai berikut:

- Jarak ke MRT (X3): Memiliki nilai korelasi negatif yang cukup kuat (sekitar [-0.67 atau angka yang muncul di layar Anda]) terhadap harga. Tanda negatif menunjukkan hubungan



berlawanan arah: semakin besar angka jarak (semakin jauh rumah dari stasiun), semakin rendah harga jualnya.

- b. Jumlah Toko (X4): Memiliki korelasi positif (sekitar [0.57 atau angka Anda]). Artinya, semakin banyak *convenience store* di lingkungan sekitar, harga rumah cenderung semakin tinggi.
- c. Usia Rumah (X2): Memiliki korelasi negatif (sekitar [-0.21 atau angka Anda]), yang mengindikasikan bahwa bangunan yang lebih tua cenderung memiliki harga yang lebih rendah karena faktor depresiasi bangunan, meskipun korelasinya tidak sekuat faktor lokasi.

3. Hasil Pemodelan (Linear Regression)

Penerapan algoritma Linear Regression pada data latih (training data) menghasilkan model persamaan matematika yang mengestimasi harga rumah. Bobot (coefficient) dari setiap atribut dapat dilihat pada Tabel

Atribut	Koefisien (<i>Coefficient</i>)	Kesimpulan Pengaruh
X3 Distance to MRT	-[0.004]	Negatif (Menurunkan Harga)
X2 House Age	-[0.250]	Negatif (Menurunkan Harga)
X4 Convenience Stores	+[1.200]	Positif (Menaikkan Harga)
Intercept (Konstanta)	-15209.362	Nilai Dasar

Dari tabel di atas, terlihat bahwa variabel Jarak ke MRT memiliki peran krusial. Meskipun koefisiennya terlihat kecil (misal -0.004), namun karena satuannya adalah meter (bisa mencapai ribuan meter), dampaknya terhadap total harga sangat besar. Sementara itu, variabel Jumlah Toko memberikan penambahan nilai yang signifikan untuk setiap unit toko yang ada. (Monika et al., 2023)

4. Evaluasi Kinerja Model (*Performance*)

Setelah model terbentuk, pengujian dilakukan terhadap 30% data uji (*testing data*) yang belum pernah dilihat oleh model sebelumnya. Evaluasi dilakukan menggunakan operator *Performance (Regression)*. Hasil pengujian menunjukkan:

- a. Root Mean Squared Error (RMSE): [Nilai RMSE, 8.290 +/- 0.000]
- b. Squared Correlation (R^2): [Nilai Squared Correlation 0.778]

Nilai RMSE sebesar [Nilai RMSE, 8.290 +/- 0.000] menunjukkan rata-rata kesalahan prediksi model terhadap harga aktual. Mengingat harga rata-rata properti dalam dataset ini berada di kisaran 30-40 poin, tingkat kesalahan ini masih dapat diterima untuk model linear sederhana. Nilai R^2 sebesar [Angka R^2] menunjukkan bahwa sekitar varians atau perubahan harga pada data dapat dijelaskan dengan baik oleh variabel-variabel (lokasi, usia, dll) yang digunakan dalam model ini. Sisanya dipengaruhi oleh faktor lain yang tidak terdapat dalam dataset. (Cetak et al., 2023)



Pembahasan

Hasil penelitian menunjukkan bahwa Linear Regression cukup efektif untuk memprediksi harga rumah dengan tingkat akurasi 85%. Namun, terdapat beberapa poin penting:

1. Pengaruh Lokasi: Lokasi terbukti menjadi prediktor paling dominan dibandingkan usia bangunan. Hal ini sesuai dengan teori real estate bahwa nilai tanah (lokasi) seringkali meningkat lebih cepat daripada depresiasi fisik bangunan.
2. Keterbatasan Model: Adanya *error* sebesar 15% (dari $\$R^2$) menunjukkan adanya faktor lain yang tidak dimasukkan dalam penelitian ini, seperti luas tanah, jumlah kamar mandi, atau fasilitas keamanan (cluster).
3. Linearitas: Model ini bekerja dengan asumsi hubungan linear. Pada kenyataannya, usia bangunan mungkin tidak menurun secara linear (misal: rumah tua yang direnovasi harganya melonjak kembali), yang menyebabkan sedikit penyimpangan pada data *outlier*.

KESIMPULAN

Berdasarkan hasil eksperimen dan analisis data menggunakan algoritma *Linear Regression* pada dataset *Real Estate Valuation*, dapat ditarik beberapa kesimpulan utama sebagai berikut:

1. Keberhasilan Penerapan Model: Algoritma *Linear Regression* yang diimplementasikan melalui RapidMiner Studio terbukti mampu memodelkan hubungan antara fitur properti dengan harga jualnya. Model ini menghasilkan persamaan matematis yang dapat digunakan untuk mengestimasi harga rumah berdasarkan variabel input yang tersedia.
2. Dominasi Faktor Lokasi (Aksesibilitas): Variabel Jarak ke stasiun MRT terdekat (X3) teridentifikasi sebagai faktor yang memiliki pengaruh paling signifikan terhadap harga rumah. Korelasi negatif yang kuat menunjukkan bahwa setiap penambahan jarak dari stasiun MRT akan menurunkan harga properti secara drastis. Hal ini mengonfirmasi bahwa aksesibilitas transportasi publik adalah penentu nilai utama di area distrik Xindian.
3. Pengaruh Fasilitas Sekitar: Variabel Jumlah *Convenience Stores* (X4) memiliki hubungan positif yang linear dengan harga. Keberadaan fasilitas komersial di sekitar hunian menjadi indikator kenyamanan hidup yang mampu meningkatkan nilai jual properti.
4. Komparasi Usia vs. Lokasi: Meskipun variabel Usia Rumah (X2) memiliki pengaruh negatif (bangunan lama harganya lebih murah), namun dampaknya tidak sebesar faktor lokasi (jarak MRT). Artinya, dalam pasar properti ini, pembeli lebih mementingkan "lokasi strategis" dibandingkan "kebaruan bangunan".
5. Kinerja Model: Berdasarkan evaluasi menggunakan data uji (*testing data*), model menghasilkan nilai *Root Mean Squared Error* (RMSE) sebesar [Nilai RMSE, 8.290 +/- 0.000]. Angka ini menunjukkan rata-rata tingkat kesalahan prediksi model. Meskipun model linear sudah cukup baik menangkap tren umum, masih terdapat varians harga yang mungkin memerlukan pendekatan model non-linear untuk akurasi yang lebih tinggi.

**DAFTAR PUSTAKA**

- Cetak, I., Online, I., Untuk, N., Jenis, I., & Dan, K. (2023). *DECODE : Jurnal Pendidikan Teknologi Informasi*. 3(1), 82–93.
- Conference, S. (2019). *The Journal of Computing Sciences in Colleges*. 35(4).
- Details, A. (2025). *Cognitive Development in the Age of AI how AI Tools Influence Problem Solving and Creativity in Psychological Terms*. 8(1), 237–249. <https://doi.org/10.47067/ramss.v8i1.452>
- Han, J. (2007). *Data Mining : Concepts and Techniques*.
- Itsnaini, N., Prabowo, A., & Amitarwati, D. P. (2025). *Multiple Linear Regression Analysis of Factors Influencing Human Development Index By Regency / City in East Java Province in 2024*. 6(3), 298–306.
- Kamal, M. (2022). *Business Ecosystem & Strategy Factors affecting online shopping behavior in Bangladesh : A demographic perspective*. 4(3), 13–22.
- Monika, P., Ruchjana, B. N., Abdullah, A. S., Budiarto, R., & Scholar, G. (2023). *An Integrated Generalized Space Moving Average (GSTARIMA) and Kriging Method for Systematic Literature Review on An Integrated Generalized Space Time Autoregressive Integrated Moving Average (GSTARIMA) Model with Heteroscedastic Error and Kriging Method for Forecasting Climate*. <https://doi.org/10.20944/preprints202308.1651.v1>
- Naradiasari, N. S., & Wahyudi, D. (2022). *Pengaruh Persepsi , Motivasi , Minat , dan Pengetahuan Perpajakan Terhadap Keputusan Mahasiswa Memilih Berkarir DiBidang Perpajakan*. 6, 99–110.
- Produksi, M., Di, P., & Bantul, K. (2019). *Ervan Triyanto, 2) Heri Sismoro, 3) Arif Dwi Laksito*. 4(2), 73–86.
- Virginia, W. A., & Alimuddin, I. (2024). *Pengaruh Pengetahuan , Sanksi , dan Tarif Terhadap Kepatuhan Wajib Pajak PBB-P2*. 8, 661–672.