PT Jurnal Cendekia Indonesia
Journal of Information Systems and Business Technology
Homepage: https://journal.jci.co.id/jisbt
Vol. 02 No. 03 (2026) Page: 843-847
E-ISSN: 3109-8886

Analisis Data Mining Menggunakan Metode Regresi Linier untuk Estimasi Nilai Hunian pada Dataset California Housing

[1]Dinnar Rizky, [2]Muhamad Andri Rian Riyadi

[12]Program Studi Sistem Informasi, Fakutas Ilmu Komputer, Universitas Pamulang, Kota Tanggerang Selatan, Indonesia

[1]rizkydinar5@gmail.com, [2]andririan183@gmail.com

Abstract

The property sector has highly fluctuating price dynamics, influenced by various factors ranging from geographical conditions to building facilities. Accurate housing price estimation is essential for developers and prospective buyers to make informed decisions. This study aims to apply Data Mining techniques using the Linear Regression algorithm to estimate the median value of housing in the California Housing dataset. This research methodology follows the stages of knowledge discovery in databases, which include data cleaning, data integration, transformation, and modeling. The dataset is processed by conducting correlation analysis to determine the main predictor variables, such as median income, geographic location, and population density. A Linear Regression model is built to map the relationship between these independent variables and the value of housing as the dependent variable. Model performance is evaluated using the statistical metrics Mean Absolute Error (MAE) and Coefficient of Determination (R^2). The results of the study are expected to show that Linear Regression is able to provide significant estimates with a reliable level of accuracy, where the median income variable is predicted to be the most dominant factor in determining house prices. This research contributes to the use of machine learning for more objective and efficient real estate market analysis.

Keywords: *Data Mining, Linear Regression, Price Estimation, California Housing, Prediction.*

Abstrak

Sektor properti memiliki dinamika harga yang sangat fluktuatif, dipengaruhi oleh berbagai faktor mulai dari kondisi geografis hingga fasilitas bangunan. Estimasi harga hunian yang akurat sangat diperlukan oleh pengembang dan calon pembeli untuk pengambilan keputusan yang tepat. Penelitian ini bertujuan untuk menerapkan teknik Data Mining menggunakan algoritma Regresi Linier untuk melakukan estimasi nilai tengah hunian pada dataset California Housing.Metodologi penelitian ini mengikuti tahapan penemuan pengetahuan dalam basis data (Knowledge Discovery in Databases), yang meliputi data cleaning, integrasi data, transformasi, hingga pemodelan. Dataset diproses dengan melakukan analisis korelasi untuk menentukan variabel prediktor utama, seperti pendapatan median (Median Income), lokasi geografis, dan kepadatan penduduk. Model Regresi Linier dibangun untuk memetakan hubungan antara variabel-variabel independen tersebut dengan nilai hunian sebagai variabel dependen.Kinerja model dievaluasi menggunakan metrik statistik Mean Absolute Error (MAE) dan Coefficient of Determination (R^2). Hasil penelitian diharapkan menunjukkan bahwa regresi linier mampu memberikan estimasi yang signifikan dengan tingkat akurasi yang dapat diandalkan, di mana variabel pendapatan median diprediksi menjadi faktor paling dominan dalam menentukan harga rumah. Penelitian ini memberikan kontribusi dalam pemanfaatan machine learning untuk analisis pasar real estat yang lebih objektif dan efisien.

Kata Kunci: *Data Mining, Regresi Linier, Estimasi Harga, California Housing, Prediksi*

A. PENDAHULUAN

Perkembangan sektor properti merupakan salah satu indikator stabilitas ekonomi di suatu wilayah. Namun, penentuan nilai hunian sering kali menghadapi kendala kompleksitas karena dipengaruhi oleh berbagai variabel yang bersifat dinamis, seperti letak geografis, kondisi sosial ekonomi, hingga fasilitas fisik bangunan. Di wilayah dengan pertumbuhan populasi yang pesat seperti California, fluktuasi harga rumah menjadi tantangan tersendiri bagi investor, pengembang, maupun masyarakat umum dalam mengambil keputusan finansial yang tepat. Estimasi harga yang tidak akurat dapat berdampak pada kerugian modal atau ketidakefisienan alokasi sumber daya.

Seiring dengan kemajuan teknologi informasi, pendekatan tradisional dalam menaksir harga rumah mulai beralih ke arah otomatisasi berbasis data. Data mining hadir sebagai solusi untuk mengekstraksi pengetahuan tersembunyi dari dataset besar, seperti dataset California Housing. Dataset ini menyediakan informasi mendalam mengenai karakteristik perumahan yang mencakup pendapatan median penduduk, usia bangunan, hingga koordinat lokasi. Pengolahan data ini

Diterima Redaksi: 10-06-2026 | Selesai Revisi: 22-06-2026 | Diterbitkan Online: 23-06-2022

memerlukan metode statistik yang mampu memodelkan hubungan antarvariabel secara sistematis.

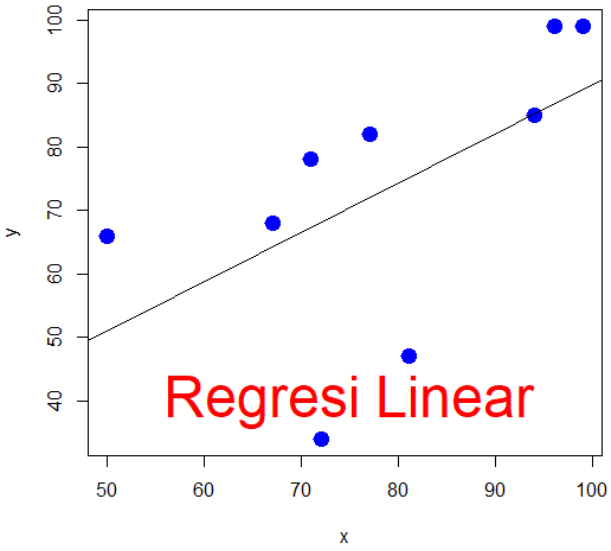
Metode Regresi Linier merupakan salah satu teknik dalam data mining yang sangat efektif untuk kasus estimasi nilai kontinu. Keunggulan metode ini terletak pada kemampuannya untuk mengidentifikasi sejauh mana variabel independen mempengaruhi variabel dependen secara linier, sehingga memudahkan peneliti dalam menginterpretasikan faktor-faktor dominan yang membentuk harga pasar. Meskipun sederhana, Regresi Linier memberikan fondasi yang kuat dalam analisis prediktif sebelum beralih ke model yang lebih kompleks.

Penelitian ini menggunakan Metode Prototype dalam proses pengembangannya. Penggunaan metode ini memungkinkan peneliti untuk melakukan pengembangan model secara iteratif, mulai dari perancangan model awal, pengujian akurasi, hingga perbaikan fitur berdasarkan hasil evaluasi secara berulang. Pendekatan ini memastikan bahwa model estimasi yang dihasilkan telah melalui tahap validasi yang ketat dan sesuai dengan karakteristik data yang ada.

Berdasarkan uraian tersebut, penelitian ini berfokus pada analisis data mining menggunakan metode regresi linier untuk estimasi nilai hunian pada dataset California Housing. Tujuan utama dari penelitian ini adalah untuk membangun model prediksi yang andal serta mengukur tingkat akurasinya menggunakan metrik evaluasi seperti R^2 Score dan Mean Absolute Error (MAE). Hasil dari penelitian ini diharapkan dapat memberikan kontribusi praktis sebagai alat bantu pengambilan keputusan di sektor real estat dan kontribusi teoretis dalam pengembangan model prediksi harga hunian.

B. METODOLOGI

Penelitian ini menggunakan pendekatan data mining dengan menerapkan metode Knowledge Discovery in Databases (KDD) untuk melakukan estimasi nilai hunian pada dataset California Housing. Metode ini dipilih karena memiliki tahapan yang sistematis dalam menggali informasi dari data, mulai dari pemilihan data hingga interpretasi hasil. Selain itu, penelitian ini juga menggunakan pendekatan Prototype dalam proses pengembangan model, sehingga model yang dihasilkan dapat diuji dan disempurnakan secara bertahap sesuai dengan kebutuhan penelitian.



Gambar 1 *Metode Regresi Linier*

Data Selection
Tahap awal dalam penelitian ini adalah proses pemilihan data yang akan digunakan. Dataset yang digunakan adalah California Housing Dataset yang berisi berbagai atribut yang menggambarkan kondisi perumahan, seperti median income, jumlah populasi, jumlah rumah tangga, serta koordinat lokasi geografis. Selain itu, dataset ini juga memiliki variabel target berupa nilai median hunian yang digunakan sebagai acuan dalam proses prediksi. Pada tahap ini dilakukan identifikasi terhadap atribut-atribut yang relevan dengan tujuan penelitian. Tidak semua atribut dalam dataset digunakan, melainkan hanya atribut yang dianggap memiliki pengaruh terhadap nilai hunian. Proses seleksi ini bertujuan untuk menyederhanakan data serta meningkatkan efektivitas dalam proses analisis selanjutnya.

Data Cleaning
Setelah data dipilih, tahap selanjutnya adalah proses pembersihan data. Tahap ini bertujuan untuk memastikan bahwa data yang digunakan memiliki kualitas yang baik dan bebas dari kesalahan yang dapat memengaruhi hasil analisis. Dalam proses ini dilakukan penanganan terhadap data yang hilang (missing value), baik dengan cara menghapus data tersebut maupun menggantinya dengan nilai tertentu yang dianggap representatif.Selain itu, dilakukan juga pengecekan terhadap data yang terduplikasi serta data yang tidak konsisten. Data yang tidak sesuai atau tidak valid akan diperbaiki atau dihapus agar tidak mengganggu proses pemodelan. Dengan dilakukannya proses data cleaning, diharapkan data yang digunakan menjadi lebih akurat dan dapat menghasilkan model yang lebih optimal.

C) Data Transformation
Tahap berikutnya adalah transformasi data, yaitu proses mengubah data ke dalam bentuk yang sesuai untuk digunakan dalam pemodelan. Pada tahap ini dilakukan penyesuaian terhadap format data agar lebih mudah diproses oleh algoritma yang digunakan.Transformasi yang

dilakukan meliputi normalisasi atau standarisasi data numerik agar memiliki skala yang seimbang. Selain itu, dilakukan juga pemilihan atribut yang paling relevan serta analisis korelasi untuk mengetahui hubungan antar variabel. Proses ini bertujuan untuk memastikan bahwa data yang digunakan telah siap dan optimal untuk digunakan dalam tahap pemodelan.

Data Mining (Modeling)
Tahap data mining merupakan tahap inti dalam penelitian ini, yaitu proses pembangunan model menggunakan algoritma Regresi Linier. Metode ini digunakan untuk mengetahui hubungan antara variabel independen dengan variabel dependen, yaitu nilai hunian.Dalam proses ini, model akan dilatih menggunakan data yang telah dipersiapkan sebelumnya. Data biasanya dibagi menjadi dua bagian, yaitu data pelatihan dan data pengujian, agar model dapat diuji kemampuannya dalam melakukan prediksi terhadap data baru. Regresi linier bekerja dengan membentuk hubungan matematis yang menggambarkan pengaruh variabel-variabel independen terhadap nilai hunian, sehingga dapat digunakan untuk melakukan estimasi secara kuantitatif.

Evaluation dan Interpretation
Setelah model dibangun, tahap selanjutnya adalah evaluasi untuk mengukur kinerja model yang dihasilkan. Evaluasi dilakukan dengan menggunakan metrik seperti Mean Absolute Error (MAE) dan koefisien determinasi (R^2). Nilai MAE digunakan untuk mengetahui rata-rata kesalahan prediksi yang dihasilkan oleh model, sedangkan R^2 digunakan untuk mengetahui seberapa besar variabel independen mampu menjelaskan variabel dependen.Hasil evaluasi ini kemudian diinterpretasikan untuk mengetahui tingkat akurasi model serta untuk mengidentifikasi variabel yang memiliki pengaruh paling besar terhadap nilai hunian. Tahap ini sangat penting karena menentukan apakah model yang dibangun sudah layak digunakan atau masih perlu dilakukan perbaikan.

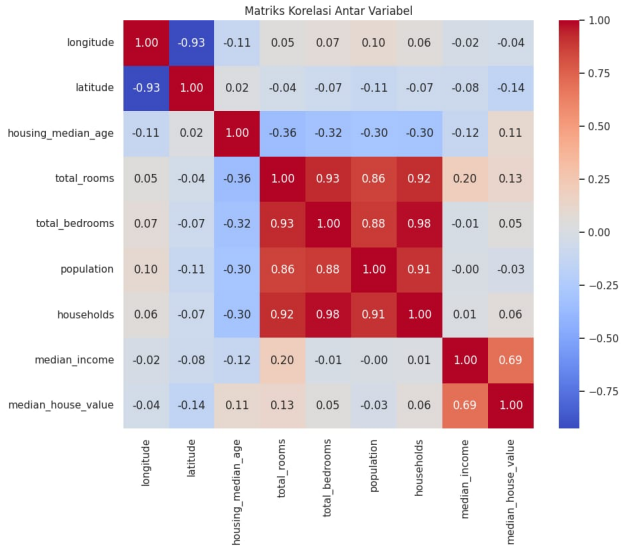
C. HASIL DAN PEMBAHASAN

Bagian ini menyajikan hasil dari proses pengolahan data serta pembahasan terhadap model yang telah dibangun menggunakan metode regresi linier. Seluruh proses analisis dilakukan secara sistematis dengan mengacu pada tahapan yang telah dijelaskan pada bagian metodologi, yaitu mulai dari tahap pemilihan data (data selection), pembersihan data (data cleaning), transformasi data (data transformation), hingga proses pemodelan dan evaluasi model. Setiap tahapan tersebut memiliki peran penting dalam menghasilkan model yang optimal dan dapat digunakan secara efektif dalam melakukan estimasi nilai hunian.Pada tahap awal, dilakukan pemahaman terhadap karakteristik dataset yang digunakan, termasuk distribusi data, tipe variabel, serta kualitas data yang tersedia. Selanjutnya, data yang telah melalui proses pembersihan dan transformasi digunakan dalam pembangunan model regresi linier untuk mengetahui hubungan antara variabel independen dan variabel dependen. Proses ini bertujuan untuk menghasilkan model matematis yang mampu merepresentasikan pola hubungan dalam data secara sistematis.

Hasil yang diperoleh dari proses pemodelan kemudian dianalisis lebih lanjut untuk mengetahui sejauh mana model mampu melakukan prediksi terhadap nilai hunian. Analisis dilakukan dengan melihat hubungan antar variabel, mengidentifikasi variabel yang paling berpengaruh, serta membandingkan hasil prediksi dengan nilai aktual. Selain itu, dilakukan juga evaluasi model menggunakan metrik tertentu untuk mengukur tingkat akurasi dan performa model secara keseluruhan.Melalui rangkaian proses tersebut, diharapkan dapat diperoleh pemahaman yang lebih mendalam mengenai pola hubungan antara faktor-faktor yang mempengaruhi nilai hunian, serta menilai kemampuan metode Regresi Linier dalam menghasilkan estimasi yang akurat dan dapat diandalkan. Hasil analisis ini juga menjadi dasar dalam menarik kesimpulan serta memberikan rekomendasi terkait pemanfaatan model dalam konteks yang lebih luas.

Analisis Deskriptif Dataset
Analisis deskriptif dilakukan untuk memahami karakteristik dasar dari dataset California Housing yang digunakan dalam penelitian. Dataset ini terdiri dari beberapa atribut numerik seperti median income, jumlah populasi, jumlah rumah tangga, serta koordinat lokasi geografis. Variabel target dalam penelitian ini adalah nilai median hunian.Berdasarkan hasil analisis awal, diketahui bahwa setiap atribut memiliki distribusi nilai yang berbeda. Variabel median income menunjukkan variasi yang cukup signifikan antarwilayah, yang mengindikasikan adanya perbedaan tingkat ekonomi masyarakat. Sementara itu, variabel populasi dan jumlah rumah tangga juga menunjukkan penyebaran data yang cukup luas. Hal ini menunjukkan bahwa dataset memiliki keragaman yang baik sehingga dapat digunakan untuk membangun model prediksi yang representatif.

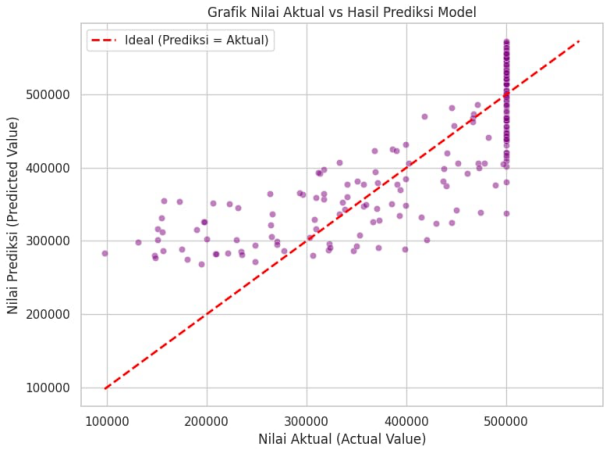
Analisis Korelasi Antar Variabel
Untuk mengetahui hubungan antarvariabel, dilakukan analisis korelasi menggunakan matriks korelasi. Hasil analisis menunjukkan bahwa beberapa variabel memiliki hubungan yang cukup kuat terhadap nilai hunian, terutama median income yang menunjukkan korelasi positif yang signifikan.Hal ini mengindikasikan bahwa semakin tinggi pendapatan rata-rata suatu wilayah, nilai hunian di wilayah tersebut cenderung semakin tinggi. Selain itu, variabel lain seperti jumlah rumah tangga dan populasi memiliki hubungan yang lebih lemah terhadap nilai hunian. Analisis ini menjadi dasar dalam menentukan variabel mana yang paling berpengaruh dalam proses pemodelan.



***Gambar 2**. Analisis Korelasi Antar Variabel*

C) Hasil Pemodelan Regresi Linier

Model prediksi dibangun menggunakan algoritma Regresi Linier untuk mempelajari hubungan antara variabel independen dan variabel dependen. Dataset yang telah melalui tahap preprocessing kemudian digunakan dalam proses pelatihan model.Hasil pemodelan menunjukkan bahwa model mampu membentuk hubungan linier antara variabel median income, populasi, dan lokasi terhadap nilai hunian. Koefisien regresi yang dihasilkan menunjukkan bahwa median income merupakan variabel yang paling dominan dalam mempengaruhi nilai hunian dibandingkan variabel lainnya Model yang dibangun kemudian digunakan untuk melakukan prediksi terhadap data pengujian. Hasil prediksi menunjukkan bahwa model mampu mengikuti pola data aktual meskipun terdapat beberapa perbedaan nilai pada beberapa data.



***Gambar 3** Grafik Prediksi vs Aktual*

D) Evaluasi Model

Evaluasi model dilakukan untuk mengukur tingkat akurasi dari model yang telah dibangun. Hasil evaluasi menunjukkan bahwa nilai Mean Absolute Error (MAE) berada pada tingkat yang relatif rendah, yang berarti rata-rata kesalahan prediksi model masih dalam batas yang dapat diterima.Selain itu, nilai koefisien determinasi (R^2) menunjukkan bahwa model mampu menjelaskan sebagian besar variasi dalam data. Hal ini menunjukkan bahwa model Regresi Linier yang digunakan memiliki kemampuan yang cukup baik dalam melakukan estimasi nilai hunian. Namun demikian, masih terdapat selisih antara nilai aktual dan nilai prediksi pada beberapa data, yang menunjukkan bahwa model belum sepenuhnya sempurna. Hal ini dapat disebabkan oleh adanya variabel lain yang belum dimasukkan dalam model.

Interpretasi Hasil

Berdasarkan hasil analisis yang dilakukan, dapat disimpulkan bahwa variabel median income memiliki pengaruh paling besar terhadap nilai hunian. Hal ini menunjukkan bahwa faktor ekonomi merupakan faktor utama dalam menentukan harga rumah.Selain itu, variabel lokasi juga memberikan kontribusi terhadap nilai hunian, meskipun tidak sebesar median income. Sementara itu, variabel populasi memiliki pengaruh yang relatif lebih kecil terhadap nilai hunian.Hasil ini sejalan dengan konsep ekonomi bahwa harga properti sangat dipengaruhi oleh tingkat pendapatan masyarakat serta kondisi wilayah. Dengan demikian, model yang dibangun dapat memberikan gambaran yang cukup baik mengenai faktor-faktor yang memengaruhi nilai hunian.

Implementasi Model

Model yang telah dibangun dapat digunakan sebagai alat bantu dalam melakukan estimasi nilai hunian berdasarkan data yang tersedia. Dengan memasukkan variabel seperti median income, populasi, dan lokasi, model dapat memberikan prediksi nilai hunian secara cepat dan objektif.Selain itu, model ini juga dapat dikembangkan lebih lanjut menjadi sistem pendukung keputusan yang dapat membantu pengguna dalam menganalisis harga properti. Dengan adanya model ini, proses pengambilan keputusan dapat dilakukan secara lebih sistematis dan berbasis data.

D. PENUTUP

Simpulan

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa penerapan metode data mining menggunakan algoritma regresi linier pada dataset California Housing mampu digunakan untuk melakukan estimasi nilai hunian secara cukup baik. Melalui tahapan Knowledge Discovery in Databases (KDD), proses pengolahan data dilakukan secara sistematis mulai dari pemilihan data, pembersihan data, transformasi data, hingga proses pemodelan dan evaluasi. Hasil analisis menunjukkan bahwa beberapa variabel memiliki pengaruh yang signifikan terhadap nilai hunian, terutama variabel median income yang menjadi faktor dominan dalam menentukan harga rumah. Model Regresi Linier yang dibangun mampu merepresentasikan hubungan antara variabel independen dan variabel dependen dalam bentuk persamaan matematis yang dapat digunakan untuk

melakukan prediksi. Berdasarkan hasil evaluasi model, diperoleh bahwa tingkat akurasi model berada pada kategori cukup baik, yang ditunjukkan oleh nilai kesalahan prediksi yang relatif rendah serta nilai koefisien determinasi yang mampu menjelaskan sebagian besar variasi data. Hal ini menunjukkan bahwa model yang dibangun memiliki kemampuan yang cukup baik dalam melakukan estimasi nilai hunian, meskipun masih terdapat beberapa keterbatasan dalam menangkap seluruh kompleksitas data.Secara keseluruhan, penelitian ini menunjukkan bahwa metode Regresi Linier dapat digunakan sebagai pendekatan awal dalam analisis prediktif terhadap nilai hunian, serta dapat memberikan gambaran mengenai faktor-faktor yang mempengaruhi harga properti secara lebih objektif dan berbasis data.

E. DAFTAR PUSTAKA

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Computer Science, 7, e623.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI Magazine, 17(3), 37-54.

Han, J., Kamber, M., & Pei, J. (2011). Data mining: Concepts and techniques (3rd ed.). Morgan Kaufmann.

Larose, D. T., & Larose, C. D. (2014). Discovering knowledge in data: An introduction to data mining (2nd ed.). John Wiley & Sons.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). Introduction to linear regression analysis (6th ed.). John Wiley & Sons.

Pace, R. K., & Barry, R. P. (1997). Sparse spatial autoregressions. Statistics & Probability Letters, 33(3), 291-297.

Phan, T. D. (2018). Housing price prediction using machine learning algorithms. Journal of Computer Science, 14(2), 233-242.

Pressman, R. S., & Maxim, B. R. (2020). Software engineering: A practitioner's approach (9th ed.). McGraw-Hill Education