Just IT : Jurnal Sistem Informasi, Teknologi Informasi dan Komputer
Vol. 15, No. 2, Januari 2025, pp. 350 – 355
P-ISSN: 2089 – 0256, e-ISSN: 2598 – 3016

PENERAPAN ALGORITMA C4.5 UNTUK MEMPREDIKSI HARGA RUMAH

Iklas Anang Subekti*[1], Chrisfian Beni Andriano[2], Dimas Nurdiansyah[3], Rahmat Hidayat[4]
[1,2,3,4] Fakultas Sains dan Teknologi, Universitas Putra Bangsa
iklas.anang@gmail.com

Abstrak

Pasar properti yang dinamis sering kali menghadirkan tantangan dalam menentukan harga rumah secara akurat akibat berbagai faktor yang kompleks. Ketidakakuratan harga dapat merugikan pihak penjual maupun pembeli. Penelitian ini mengimplementasikan algoritma C4.5, sebuah algoritma berbasis pohon keputusan yang menggunakan *information gain* untuk memilih atribut utama dalam membangun model prediksi. Algoritma ini unggul dalam menangani data numerik maupun kategorikal, menjadikannya cocok untuk menganalisis atribut rumah seperti luas tanah, luas bangunan, jumlah kamar tidur, jumlah kamar mandi, dan luas garasi. Dengan membagi dataset berdasarkan atribut yang memberikan *gain ratio* tertinggi, algoritma C4.5 menghasilkan model yang dapat memprediksi harga rumah dengan akurasi mencapai 85,70%. Kemampuan algoritma ini dalam memberikan struktur pohon keputusan yang mudah dipahami oleh manusia juga menjadi keunggulan dalam mendukung interpretasi hasil. Hasil penelitian ini menunjukkan bahwa algoritma C4.5 dapat diimplementasikan secara efektif untuk membantu meningkatkan efisiensi dan transparansi di pasar properti, khususnya di wilayah Jakarta Selatan.

Kata kunci: *harga rumah, algoritma C4.5, pohon keputusan, prediksi*

Abstract

The dynamic property market often poses challenges in accurately determining house prices due to various complex factors. Inaccurate pricing can adversely impact both sellers and buyers. This study implements the C4.5 algorithm, a decision tree-based method that employs *information gain* to select the most significant attributes for building a predictive model. The algorithm excels in handling both numerical and categorical data, making it suitable for analyzing house attributes such as land size, building area, number of bedrooms, number of bathrooms, and garage size. By partitioning the dataset based on the attributes with the highest *gain ratio*, the C4.5 algorithm produces a model capable of predicting house prices with an accuracy of 85.70%. Additionally, the algorithm's ability to create interpretable decision tree structures provides a distinct advantage in facilitating result interpretation. The findings of this study demonstrate that the C4.5 algorithm can be effectively implemented to enhance efficiency and transparency in the property market, particularly in South Jakarta.

Keywords: *House Price, C4.5 Algorithm, Decision Tree, Prediction*

.

1. Pendahuluan

Dalam beberapa dekade terakhir, sektor properti menjadi salah satu pilar penting dalam perekonomian global. Pasar properti, khususnya sektor perumahan, mengalami pertumbuhan yang pesat, baik dari segi permintaan maupun harga. Hal ini didorong oleh meningkatnya kebutuhan masyarakat terhadap hunian, serta faktor eksternal seperti urbanisasi dan kebijakan ekonomi makro. Namun, dinamika pasar properti ini memunculkan tantangan baru, yaitu ketidakmampuan para pelaku pasar untuk memprediksi harga rumah secara akurat. Ketidakakuratan dalam menentukan harga sering kali merugikan baik penjual maupun pembeli. Bagi penjual, menetapkan harga yang terlalu tinggi dapat menurunkan daya tarik properti di pasar, sementara bagi pembeli, sulitnya menentukan harga yang wajar sesuai anggaran menjadi penghalang dalam proses pengambilan keputusan.

Harga rumah dipengaruhi oleh banyak faktor, seperti luas tanah, luas bangunan, jumlah kamar tidur, jumlah kamar mandi, serta keberadaan garasi (Lestari & Astuti, 2022). Faktor-faktor ini memengaruhi nilai jual properti dan dapat bervariasi antara satu rumah dengan rumah lainnya. Misalnya, rumah dengan luas tanah lebih besar dan jumlah kamar tidur lebih banyak biasanya dihargai lebih tinggi. Begitu juga dengan lokasi rumah yang dapat mempengaruhi permintaan dan penawaran, serta akhirnya harga jual rumah.

Dalam era big data, metode berbasis algoritma data mining menjadi solusi yang semakin banyak digunakan untuk analisis pasar properti. Salah satu algoritma yang efektif adalah C4.5, dimana algoritma c4.5 dapat digunakan untuk membangun pohon keputusan berdasarkan pembagian data yang optimal. Algoritma ini memilih atribut yang paling relevan berdasarkan informasi gain, sehingga dapat menyederhanakan proses analisis tanpa kehilangan esensi dari data. menunjukkan bahwa algoritma C4.5 efektif dalam memprediksi variabel target dengan tingkat akurasi yang tinggi. Dengan kemampuan untuk mengolah data yang kompleks dan bervariasi, algoritma ini menawarkan pendekatan yang lebih andal dibandingkan metode konvensional.

Keberhasilan algoritma C4.5 dalam aplikasi prediksi harga properti juga telah dibuktikan oleh beberapa penelitian terdahulu. Misalnya, penelitian oleh (Farid & Fitrianah, 2021)menunjukkan bahwa algoritma pohon keputusan dapat mencapai tingkat akurasi hingga 86,24 % dalam memprediksi harga rumah, sementara algoritma regresi linear hanya mencapai akurasi 80%. Penelitian ini melibatkan iterasi sebanyak 10 kali, dengan ukuran sampel yang dihitung menggunakan G power Calculator, menetapkan cutoff 80% sebagai batas minimal untuk kekuatan analisis yang memadai. Meskipun analisis statistik menggunakan SPSS menunjukkan bahwa perbedaan akurasi antara kedua metode tidak signifikan, algoritma pohon keputusan inovatif terbukti lebih unggul dalam memperkirakan nilai properti di masa depan.

Penelitian ini melibatkan iterasi sebanyak 10 kali pada algoritma pohon keputusan baru, dengan ukuran sampel dihitung menggunakan G power Calculator dan menetapkan cutoff sebesar 80% sebagai batas minimal untuk kekuatan analisis yang memadai. Meskipun analisis statistik menggunakan SPSS menunjukkan bahwa perbedaan akurasi antara kedua metode tidak signifikan ($p=0.618$, $p>0.05$), algoritma pohon keputusan inovatif terbukti lebih unggul dalam memperkirakan nilai properti di masa depan. Sementara itu, penelitian oleh (Harman, 2024)menunjukkan bahwa algoritma C4.5 memiliki keunggulan dalam memprediksi penjualan produk makanan kuaci di PT Prima Niaga Indomas. Dengan memanfaatkan data historis penjualan, informasi produk, dan variabel eksternal seperti faktor pemasaran, penelitian ini berhasil mengembangkan model prediksi yang membantu perusahaan meramalkan penjualan secara lebih akurat.

Dalam konteks lokal, penelitian ini berfokus pada kawasan Jakarta Selatan, yang merupakan salah satu daerah strategis dengan dinamika pasar properti yang sangat kompetitif. Harga rumah di kawasan ini

cenderung dipengaruhi oleh kombinasi antara faktor intrinsik, seperti luas tanah dan kondisi fisik properti, serta faktor eksternal, seperti kedekatan dengan pusat bisnis. Namun, tantangan utama yang dihadapi adalah bagaimana membangun model prediksi harga rumah yang mampu mengakomodasi keragaman data dengan akurasi tinggi.

Penelitian ini bertujuan untuk membangun model prediksi harga rumah menggunakan algoritma C4.5 berdasarkan data sekunder yang meliputi atribut seperti luas tanah, luas bangunan, jumlah kamar tidur, jumlah kamar mandi, dan keberadaan garasi. Selain itu, penelitian ini juga bertujuan untuk mengidentifikasi faktor-faktor utama yang memengaruhi harga rumah di Jakarta Selatan. Dengan pendekatan berbasis algoritma data mining, penelitian ini diharapkan tidak hanya memberikan manfaat praktis bagi pelaku pasar properti, tetapi juga berkontribusi pada literatur ilmiah terkait penerapan algoritma data mining di sektor properti.

2. Metode Penelitian

Penelitian ini kami ambil dari salah satu agen properti dan juga kami ambil dari CV. Tunas Gemilang . Populasi yang digunakan mencakup seluruh harga rumah yang ada di Jakarta Selatan, sementara sampel yang diambil adalah harga rumah yang berada di wilayah tersebut. Data yang dianalisis terdiri dari 1010 transaksi penjualan rumah yang berada di Jakarta Selatan, yang diakses pada tanggal 1 November 2024.

Gambar 1. Tahapan Penelitian

Gambar 1. merupakan tahapan penelitian yang dimulai dengan mengidentifikasi masalah, dilanjutkan dengan kajian literatur yang mengacu pada jurnal internasional terkemuka serta beberapa jurnal nasional. Selanjutnya, data publik mengenai harga rumah diambil dari jurnal sebelumnya dan diproses melalui tahap pre-processing dan data cleaning untuk menghapus data yang tidak relevan atau tidak lengkap, sehingga menghasilkan data yang lebih bersih dan siap digunakan. Setelah itu, dilakukan analisis deskriptif terhadap data, dan selanjutnya dilakukan pemodelan menggunakan algoritma C4.5. Tahapan terakhir adalah pengujian model dan penyimpulan hasil penelitian.

Algoritma C4.5 adalah metode pembelajaran mesin yang digunakan untuk membuat pohon keputusan berdasarkan data input. Algoritma ini membangun model pohon dengan memilih atribut yang memberikan informasi maksimum berdasarkan indeks gain rasio.

Tahapan dalam algoritma C4.5 meliputi:

1. Perhitungan *Entropy*

   Entropy mengukur tingkat ketidakpastian data dan didefinisikan dengan rumus:

$$E(S) = -\sum_{i=0}^{n} p_i \, log_2(p_i)$$

   dimana:
   $E(S)$: *entropy dataset S*
   p_i : proporsi data dalam kategori *i*

2. Perhitungan *Gain*

   Gain digunakan untuk mengukur seberapa besar pengurangan *entropy* setelah membagi data berdasarkan atribut tertentu. Rumusnya:

$$Gain(A) = E(S) - \sum_{v \in values(A)} \frac{|S_v|}{|S|} E(S_v)$$

   dimana:
   A : atribut yang dianalisis
   $|S_v|$: jumlah data dalam subset S_v setelah pembagian atribut *A*
   $|S|$: jumlah total data dalam *dataset S*
   $E(S_v)$: *entropy* dari subset S_v

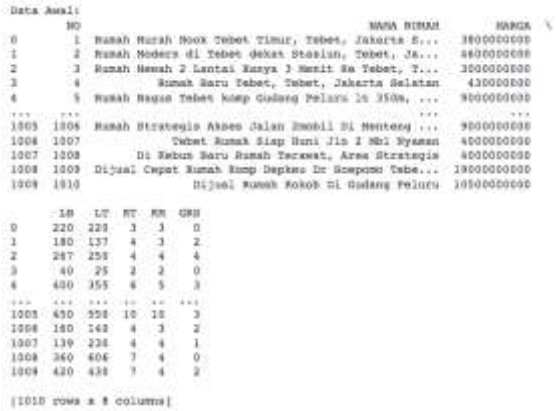
3. Hasil dan Pembahasan

Penelitian ini menggunakan aplikasi berbasis web, yaitu Google Colab, untuk

melakukan proses pembersihan data, transformasi data, visualisasi data, serta penerapan *machine learning* dan simulasi.

Pre-Processing

Data yang digunakan untuk pelatihan dan pengujian model perlu diproses dengan cermat agar model dapat mempelajari pola dengan lebih efisien (Adetunji et al., 2021). Tahap *pre-processing* ini bertujuan untuk menghasilkan dataset final yang akan dimasukkan ke dalam model yang akan dibangun (Kurniawan et al., 2023). Dalam penelitian ini, data yang digunakan berjumlah 1010 sampel dengan 6 atribut, yaitu Harga, Luas Bangunan (LB), Luas Tanah (LT), Jumlah Kamar Tidur (KT), Jumlah Kamar Mandi (KM), dan Luas Garasi (GRS).

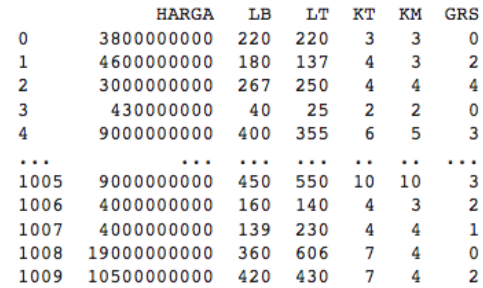


Gambar 2. *Pre-Processing*

Data *Cleaning*

Pada tahap awal proses data cleaning, dilakukan penghapusan kolom-kolom yang tidak relevan untuk prediksi. Kolom 'NO' yang kemungkinan berisi nomor urut atau ID dan kolom 'NAMA RUMAH' yang berisi nama rumah, keduanya dihapus karena tidak memberikan informasi yang signifikan dalam proses pemodelan harga rumah. Proses penghapusan kolom ini dilakukan dengan menggunakan perintah data.drop (columns=['NO', 'NAMA RUMAH']). Selanjutnya, data kemudian dipisahkan menjadi dua bagian, yaitu fitur (X) dan target (y). Fitur berisi seluruh kolom selain HARGA, sementara target hanya berisi kolom HARGA. Pemisahan ini memastikan bahwa hanya data yang relevan digunakan untuk pelatihan model, sementara data target digunakan untuk evaluasi hasil prediksi.

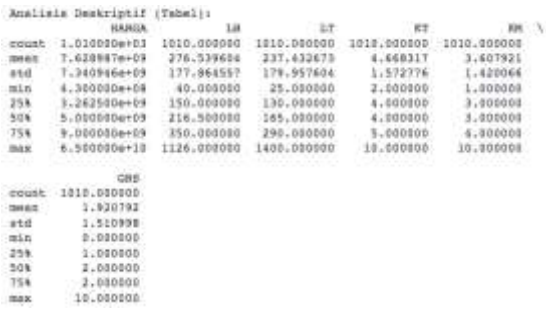
Selain itu, proses standarisasi juga diterapkan pada data fitur menggunakan StandardScaler. Standarisasi ini bertujuan untuk menyamakan skala antar fitur, sehingga tidak ada fitur yang dominan hanya karena skala atau rentang nilainya yang lebih besar, yang dapat mempengaruhi kinerja model secara negatif.



Gambar 4. Data Hasil *Cleaning*

Analisis Deskriptif

Penelitian ini melakukan analisis deskriptif terhadap data yang tersedia, dan hasil analisis disajikan pada Gambar 4. Berdasarkan Gambar 4, dari total 1010 data yang ada, diperoleh informasi bahwa harga rata-rata rumah di Jakarta Selatan adalah sebesar 7,63 milyar rupiah. Selain itu, rata-rata luas tanah adalah 237 m2 dan luas bangunan mencapai 276 m2. Rumah-rumah tersebut memiliki rata-rata 4 kamar tidur, 3 kamar mandi, dan luas garasi sebesar 1 m2.



Gambar 5. Analisis Deskriptif

Akurasi

Dari hasil pengujian yang ditunjukkan pada Gambar 6, terlihat bahwa akurasi yang diperoleh mencapai 0,8570 atau 85,70%, yang

cukup memadai untuk memprediksi harga rumah berdasarkan variabel independen yang digunakan.

Pengujian Model

Setelah model dilatih menggunakan dataset pelatihan, langkah selanjutnya dalam penelitian ini adalah menguji kinerja model prediktif. Proses ini dilakukan dengan menghapus data harga asli dan membiarkan model memprediksi harga rumah. Harga yang diprediksi kemudian dibandingkan dengan harga aktual, dan selisihnya dihitung. Hasilnya, seperti yang terlihat pada Gambar 6, menunjukkan adanya perbedaan antara harga yang sebenarnya dan harga yang diprediksi, dengan kolom selisih menunjukkan penurunan harga yang cukup signifikan, yakni rata-rata penurunan sebesar 27,14%.

Hasil Prediksi vs Actual (dengan Selisih dan Persentase Penurunan):

	Actual_HARGA	Predicted_HARGA	Price_Difference	Price_Drop
629	8900000000	8000000000	900000000	True
788	6500000000	2000000000	4500000000	True
684	6500000000	6500000000	0	False
516	37000000000	38000000000	-1000000000	False
529	18500000000	21000000000	-2500000000	False
657	8500000000	13000000000	-4500000000	False
552	2150000000	1100000000	1050000000	True
531	2250000000	2500000000	-250000000	False
321	4000000000	3000000000	1000000000	True
70	3799000000	3000000000	799000000	True

Gambar 6. Hasil Pengujian Aktual vs Prediksi

Hasil

Berdasarkan perhitungan diatas penelitian ini mendapat nilai *gain* tertinggi sebesar 0,8570 dan Setelah itu dilanjutkan memilih *node* hingga semua atribut memiliki kelas. Jika semua atribut sudah memiliki kelas.

Gambar 6. Pohon keputusan

Pohon keputusan pada gambar menjelaskan cara memprediksi harga properti berdasarkan beberapa faktor, seperti luas bangunan (LB), luas tanah (LT), jumlah kamar tidur (KT), dan faktor tambahan lainnya (GRS). Pohon ini dimulai dengan membagi properti berdasarkan luas bangunan, apakah lebih kecil atau sama dengan 323,5 meter persegi. Jika luas bangunan lebih kecil, harga rata-rata properti adalah 481 miliar. Properti ini kemudian dikelompokkan lagi berdasarkan luas tanah. Misalnya, jika luas tanah kurang dari atau sama dengan 60,5 meter persegi, harga properti turun menjadi 321 miliar, dengan harga pasti pada beberapa properti, seperti 430 miliar.

Untuk properti dengan luas tanah lebih besar, jumlah kamar tidur menjadi faktor utama. Properti dengan jumlah kamar tidur lebih sedikit memiliki harga yang lebih rendah, berkisar antara 280 miliar hingga 380 miliar, tergantung pada faktor lainnya. Di sisi lain, jika luas bangunan properti lebih besar dari 323,5 meter persegi, harga rata-rata naik menjadi 579 miliar. Properti dengan luas bangunan yang sangat besar, lebih dari 375,9 meter persegi, memiliki harga jauh lebih tinggi, hingga 900 miliar.

Secara sederhana, pohon keputusan ini menunjukkan bahwa semakin besar luas bangunan, semakin tinggi harga properti. Luas tanah dan jumlah kamar tidur juga memengaruhi harga, terutama untuk properti dengan luas bangunan sedang. Properti besar hampir selalu memiliki harga tinggi, sedangkan properti kecil atau menengah memiliki harga yang lebih bervariasi tergantung pada kombinasi faktor-faktor lainnya.

4. Kesimpulan

Penelitian ini berhasil membangun model prediksi harga rumah menggunakan algoritma C4.5 dengan tingkat akurasi sebesar 85,70%. Model ini memanfaatkan atribut penting seperti luas tanah, luas bangunan, jumlah kamar tidur, jumlah kamar mandi, dan luas garasi untuk memprediksi harga rumah di Jakarta Selatan. Hasil penelitian menunjukkan bahwa algoritma C4.5 mampu memberikan prediksi yang cukup akurat dan relevan dalam membantu penjual dan pembeli menentukan harga rumah yang sesuai dengan kondisi pasar. Penggunaan teknik *pre-processing* dan standarisasi data juga berkontribusi signifikan terhadap performa model. Temuan ini dapat diterapkan sebagai solusi praktis dalam sektor properti untuk meningkatkan transparansi dan

.

efisiensi pasar rumah.

Daftar Pustaka

Adetunji, A. B., Akande, O. N., Ajala, F. A., Oyewo, O., Akande, Y. F., & Oluwadara, G. (2021). House Price Prediction using Random Forest Machine Learning Technique. *Procedia Computer Science*, *199*, 806–813. https://doi.org/10.1016/j.procs.2022.01.100

Farid, M., & Fitrianah, D. (2021). Rekomendasi Pemilihan Restoran Berdasarkan Rating Online Menggunakan Algoritma C4.5. *Jurnal Telekomunikasi Dan Komputer*, *11*(1), 9. https://doi.org/10.22441/incomtech.v11i1.9791

Harman, R. (2024). Computer Based Information System Journal PENERAPAN ALGORITMA C4.5 UNTUK MEMPREDIKSI PENJUALAN BARANG PADA PT PRIMA NIAGA INDOMAS. *CBIS JOURNAL*, *12*(01). http://ejournal.upbatam.ac.id/index.php/cbishttp://ejournal.upbatam.ac.id/index.php/cbis

Kurniawan, I., Cahya Putri Buani, D., Apriliah, W., Amegia Saputra, R., & Korespondensi, P. (2023). *IMPLEMENTASI ALGORITMA RANDOM FOREST UNTUK MENENTUKAN PENERIMA BANTUAN RASKIN IMPLEMENTATION OF RANDOM FOREST ALGORITHM FOR DETERMINING RECIPIENTS OF RASKIN*. *10*(2), 421–428. https://doi.org/10.25126/jtiik.202396225

Lestari, E. S., & Astuti, I. (2022). Penerapan Random Forest Regression Untuk Memprediksi Harga Jual Rumah Dan Cosine Similarity Untuk Rekomendasi Rumah Pada Provinsi Jawa Barat. *Jurnal Ilmiah FIFO*, *14*(2), 131. https://doi.org/10.22441/fifo.2022.v14i2.003