JURNAL MEDIA INFORMATIKA [JUMIN]
Volume 7, Nomor 4, Juli – Agustus 2026, Page 1141-1150
ISSN 2808-005X (media online)
Available Online at http://ejournal.sisfokomtek.org/index.php/jumin

Prediksi Harga Rumah di Jakarta Menggunakan Pendekatan Hybrid K-Means Clustering dan Decision Tree

Kasih Nursaidah Husnina Br Siregar[1], Andini Nurjannah[2], Citra Ayuningtyas[3], Bagas Try Atmaja[4], Dedy Hartama[5]

[1,2,3,4,5)]Sistem Informasi, STIKOM Tunas Bangsa, Pematangsiantar, Indonesia

Email: [1,*]kasihsiregar009@email.com, [2]nurjannahandini6@email.com, [3]citraayuningtyas59@email.com, [4]bagastryatmaja1234@email.com, [5]dedyhartama@amiktunasbangsa.ac.id
Email Penulis Korespondensi: [1]kasihsiregar009@email.com

Abstrak− Harga rumah merupakan salah satu aspek penting dalam sektor properti yang dipengaruhi oleh berbagai karakteristik, seperti lokasi, luas tanah, luas bangunan, jumlah kamar, dan fasilitas pendukung. Penelitian ini bertujuan untuk menganalisis pengaruh penambahan informasi hasil clustering terhadap performa prediksi harga rumah di Jakarta menggunakan pendekatan Hybrid K-Means Clustering dan Decision Tree. Dataset yang digunakan adalah *Jakarta House Price Dataset* yang diperoleh dari Kaggle dengan jumlah 10.000 data rumah. Tahapan penelitian meliputi pembersihan data, penanganan *outlier* menggunakan metode *Interquartile Range* (IQR), transformasi data, pembentukan cluster menggunakan algoritma K-Means, serta pembangunan model prediksi menggunakan Decision Tree. Hasil penelitian menunjukkan bahwa model hybrid memperoleh nilai RMSE sebesar 2.097.559.711 dan R^2 sebesar 0,6044, yang lebih baik dibandingkan model *Baseline Decision Tree*. Sementara itu, proses clustering menghasilkan tiga cluster dengan nilai Silhouette Score sebesar 0,1436, yang menunjukkan bahwa kualitas pemisahan antar cluster masih tergolong rendah, namun tetap mampu memberikan informasi tambahan sebagai fitur pada model prediksi. Hasil analisis juga menunjukkan bahwa luas tanah dan luas bangunan merupakan faktor yang paling berpengaruh terhadap harga rumah. Dengan demikian, pendekatan Hybrid K-Means Clustering dan Decision Tree dapat digunakan sebagai alternatif dalam prediksi harga rumah, meskipun kualitas clustering masih perlu ditingkatkan pada penelitian selanjutnya.

Kata Kunci: Harga Rumah, K-Means Clustering, Decision Tree, Prediksi Harga Rumah, Machine Learning

Abstract− House prices are a critical aspect of the property sector, influenced by various characteristics such as location, land area, building area, number of bedrooms, and supporting facilities. This study aims to analyze the effect of incorporating clustering information on the performance of house price prediction in Jakarta using a hybrid approach of K-Means Clustering and Decision Tree. The dataset used is the Jakarta House Price Dataset obtained from Kaggle, consisting of 10,000 house records. The research stages include data cleaning, outlier handling using the Interquartile Range (IQR) method, data transformation, cluster formation using the K-Means algorithm, and building the prediction model using the Decision Tree. The results show that the hybrid model achieves an RMSE value of 2,097,559,711 and an R^2 of 0.6044, outperforming the baseline Decision Tree model. Meanwhile, the clustering process yields three clusters with a Silhouette Score of 0.1436, indicating that the separation quality between clusters is still relatively low, yet it remains capable of providing additional feature information to the prediction model. The analysis also reveals that land area and building area are the most influential factors driving house prices. Thus, the Hybrid K-Means Clustering and Decision Tree approach can be utilized as an alternative for house price prediction, although the clustering quality needs to be further improved in future research.

Keywords: House Price Prediction, K-Means Clustering, Decision Tree, House Price Prediction, Machine Learning

1. PENDAHULUAN

Harga rumah merupakan salah satu komponen penting dalam sektor properti yang nilainya dipengaruhi oleh berbagai karakteristik, seperti lokasi, luas tanah, luas bangunan, jumlah kamar tidur, jumlah kamar mandi, serta fasilitas pendukung lainnya. Variasi karakteristik tersebut menyebabkan perbedaan harga rumah yang cukup signifikan sehingga proses penentuan harga yang tepat menjadi tantangan bagi masyarakat, investor, maupun pengembang properti. Penetapan harga yang kurang akurat dapat berdampak pada kerugian finansial, mengurangi efisiensi transaksi, serta menurunkan daya saing properti di pasar. Oleh karena itu, diperlukan suatu pendekatan yang mampu mengolah data historis secara efektif untuk menghasilkan prediksi harga rumah yang lebih objektif, konsisten, dan akurat [1], [2], [3].

Perkembangan teknologi informasi yang semakin pesat serta meningkatnya ketersediaan data telah mendorong penerapan teknik *data mining* dan *machine learning* pada berbagai bidang, termasuk sektor properti. Berbagai penelitian menunjukkan bahwa algoritma *machine learning* mampu mengidentifikasi hubungan antara karakteristik rumah dengan nilai jualnya sehingga dapat dimanfaatkan sebagai alat bantu dalam proses prediksi harga rumah. Beragam metode telah digunakan, seperti Decision Tree, Random Forest, Support Vector Machine, hingga berbagai pendekatan berbasis *clustering*, yang terbukti mampu meningkatkan kualitas analisis serta akurasi prediksi harga rumah pada berbagai wilayah dan karakteristik data yang berbeda [4], [5], [6], [7], [8], [9].

Terakreditasi SINTA 5 SK :72/E/KPT/2024 Submitted: **03/06/2026**; Accepted: **28/06/2026**; Published: **30/07/2026**

JURNAL MEDIA INFORMATIKA [JUMIN]
Volume 7, Nomor 4, Juli – Agustus 2026, Page 1141-1150
ISSN 2808-005X (media online)
Available Online at http://ejournal.sisfokomtek.org/index.php/jumin

Meskipun demikian, setiap algoritma memiliki karakteristik, kelebihan, dan keterbatasan yang berbeda sehingga pemilihan metode sangat bergantung pada tujuan penelitian dan karakteristik data yang digunakan.

Selain proses prediksi, segmentasi data juga memiliki peranan penting dalam memahami karakteristik pasar properti. Salah satu algoritma yang banyak digunakan untuk melakukan segmentasi adalah K-Means Clustering, yaitu metode yang mengelompokkan data berdasarkan tingkat kemiripan karakteristik antar objek. Penelitian sebelumnya menunjukkan bahwa K-Means mampu mengelompokkan rumah berdasarkan lokasi, harga, luas tanah, luas bangunan, maupun atribut properti lainnya sehingga menghasilkan segmentasi pasar yang lebih terstruktur dan mudah dianalisis [1], [2], [3], [10]. Informasi hasil segmentasi tersebut dapat dimanfaatkan untuk mengetahui karakteristik masing-masing kelompok rumah sehingga membantu proses analisis pasar, penyusunan strategi pemasaran, maupun pengambilan keputusan dalam sektor properti.

Di sisi lain, Decision Tree merupakan salah satu algoritma *machine learning* yang banyak digunakan dalam permasalahan prediksi karena mampu memodelkan hubungan nonlinier antar variabel dan menghasilkan struktur keputusan yang mudah dipahami. Dibandingkan beberapa algoritma lain yang lebih kompleks, Decision Tree memiliki tingkat interpretabilitas yang tinggi sehingga mampu menunjukkan atribut-atribut yang paling berpengaruh terhadap hasil prediksi. Keunggulan tersebut menjadikan Decision Tree tidak hanya digunakan untuk memperoleh nilai prediksi, tetapi juga sebagai alat analisis dalam mengidentifikasi faktor-faktor yang memengaruhi harga rumah [4], [9], [11], [12]. Oleh karena itu, algoritma ini dipilih dalam penelitian ini karena mampu menghasilkan model prediksi yang mudah diinterpretasikan sekaligus memberikan informasi mengenai karakteristik properti yang paling berpengaruh terhadap harga rumah.

Berdasarkan penelitian-penelitian terdahulu, dapat diketahui bahwa sebagian besar penelitian masih mengimplementasikan proses segmentasi dan prediksi sebagai dua tahapan yang berdiri sendiri. Penelitian yang menggunakan K-Means umumnya hanya menghasilkan kelompok-kelompok data rumah berdasarkan karakteristik tertentu tanpa memanfaatkan informasi tersebut pada proses prediksi [1], [2], [3], [7]. Sebaliknya, penelitian yang menggunakan Decision Tree maupun algoritma prediksi lainnya lebih berfokus pada peningkatan akurasi model menggunakan atribut asli dataset tanpa mempertimbangkan informasi hasil segmentasi [4], [5], [10], [13], [6], [8], [11]. Kondisi tersebut menunjukkan bahwa pemanfaatan informasi cluster sebagai fitur tambahan pada model prediksi masih relatif terbatas, khususnya pada prediksi harga rumah di Jakarta. Selain itu, belum banyak penelitian yang mengevaluasi sejauh mana informasi hasil clustering mampu meningkatkan performa model Decision Tree dalam memprediksi harga rumah.

Berdasarkan kesenjangan penelitian tersebut, penelitian ini mengusulkan pendekatan *Hybrid K-Means Clustering dan Decision Tree* dengan memanfaatkan hasil clustering sebagai fitur tambahan pada proses prediksi harga rumah. Berbeda dengan penelitian sebelumnya yang hanya menggunakan atribut asli dataset, penelitian ini mengintegrasikan informasi cluster ke dalam model Decision Tree sehingga model tidak hanya memanfaatkan karakteristik individual setiap rumah, tetapi juga karakteristik kelompok rumah yang memiliki kemiripan. Pendekatan ini diharapkan mampu meningkatkan kemampuan model dalam mengenali pola data yang lebih kompleks sekaligus memberikan informasi mengenai karakteristik pasar properti di Jakarta.

Penelitian ini menggunakan Jakarta House Price Dataset yang diperoleh dari platform Kaggle dengan total 10.000 data rumah. Tahapan penelitian meliputi pembersihan data (*data cleaning*), penanganan *outlier* menggunakan metode Interquartile Range (IQR), transformasi data kategorikal, standardisasi data, pembentukan cluster menggunakan K-Means Clustering, serta pembangunan model prediksi menggunakan Decision Tree Regression. Evaluasi dilakukan menggunakan metrik Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), R-Squared (R^2), dan Silhouette Score untuk mengukur kualitas model prediksi dan hasil clustering. Hasil penelitian diharapkan dapat memberikan kontribusi dalam pengembangan metode prediksi harga rumah berbasis *machine learning*, sekaligus menjadi referensi bagi masyarakat, investor, maupun pengembang properti dalam menentukan estimasi harga rumah secara lebih objektif dan akurat.

2. METODOLOGI PENELITIAN

2.1 Dataset Penelitian

Penelitian ini menggunakan **Jakarta House Price Dataset** yang diperoleh dari platform Kaggle dan dapat diakses melalui https://www.kaggle.com/datasets/abiyyurasyiq/jakarta-house-price-dataset. Dataset terdiri atas 10.000 data rumah yang berada di wilayah Jakarta dengan sembilan atribut yang menggambarkan karakteristik setiap properti, yaitu index, price, district, city, bed_rooms, bath_rooms, carport, land_area, dan building_area. Atribut price digunakan sebagai variabel target yang akan diprediksi, sedangkan atribut lainnya digunakan sebagai variabel input dalam proses clustering dan prediksi harga rumah. Untuk memberikan gambaran mengenai data yang digunakan, beberapa baris pertama dataset ditampilkan pada Tabel 1, sedangkan deskripsi masing-masing atribut disajikan pada Tabel 2.

Terakreditasi SINTA 5 SK :72/E/KPT/2024 Submitted: **03/06/2026**; Accepted: **28/06/2026**; Published: **30/07/2026**

JURNAL MEDIA INFORMATIKA [JUMIN]
Volume 7, Nomor 4, Juli – Agustus 2026, Page 1141-1150
ISSN 2808-005X (media online)
Available Online at http://ejournal.sisfokomtek.org/index.php/jumin

Tabel 1. Cuplikan Dataset

Index	Price (Miliar)	District	City	Bed Rooms	Bath Rooms	Carport	Lad Area	Builg Area
1	5,9	Citra Garden	Jakarta Barat	2	4	2	250	350
2	2,7	Jelambar	Jakarta Barat	4	2	0	100	225
3	2,2	Jelambar	Jakarta Barat	3	3	0	60	140
...	...	...	...	...	...	...	...	...
10000	48,0	Menteng	Jakarta Pusat	5	5	0	716	400

Berdasarkan Tabel 1, setiap baris pada dataset merepresentasikan satu unit rumah beserta karakteristiknya. Dataset memuat informasi mengenai lokasi rumah, jumlah kamar tidur, jumlah kamar mandi, kapasitas kendaraan, luas tanah, luas bangunan, serta harga rumah yang digunakan sebagai dasar dalam proses segmentasi dan prediksi.

Tabel 2**.** Deskripsi Atribut Dataset

No	Atribut	Keterangan
1	index	Identitas data
2	price	Harga rumah
3	district	Kecamatan lokasi rumah
4	city	Kota administrasi
5	bed_rooms	Jumlah kamar tidur
6	bath_rooms	Jumlah kamar mandi
7	carport	Kapasitas kendaraan
8	land_area	Luas tanah (m^2)
9	building_area	Luas bangunan (m^2)

Sebelum dilakukan proses pemodelan, dataset terlebih dahulu dieksplorasi untuk mengetahui karakteristik data yang digunakan. Berdasarkan hasil eksplorasi diketahui bahwa dataset terdiri atas atribut numerik dan kategorikal. Setelah melalui proses pembersihan data (data cleaning) dan penanganan outlier, jumlah data yang digunakan pada penelitian ini menjadi 7.523 data. Ringkasan statistik deskriptif dari atribut numerik yang digunakan ditunjukkan pada Tabel 3.

Tabel 3. Statistik Deskriptif Dataset

Atribut	Minimum	Maksimum	Rata-rata	Standar Deviasi
Price (Rp)	1.850.000	20.500.000.000	4.344.571.000	3.329.495.000
Bed Rooms	0	8	3,63	1,22
Bath Rooms	0	7	2,97	1,11
Carport	0	3	1,17	0,82
Land Area (m^2)	0	480	160,13	95,10
Building Area (m^2)	1	570	205,47	114,82

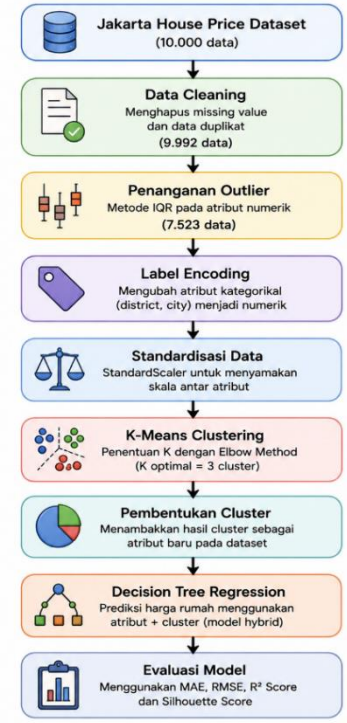
Berdasarkan Tabel 3, dataset hasil *data preparation* memiliki variasi karakteristik yang cukup tinggi. Nilai rata-rata harga rumah sebesar Rp4,34 miliar dengan rentang harga mulai dari Rp1,85 juta hingga Rp20,5 miliar, menunjukkan bahwa harga rumah di Jakarta memiliki variasi yang besar. Selain itu, atribut land area dan building area juga memiliki rentang nilai yang cukup lebar sehingga diperlukan proses penanganan *outlier* sebelum dilakukan proses clustering dan pemodelan agar kualitas data menjadi lebih representati.

2.2 Tahapan Penelitian

Penelitian Penelitian ini dilakukan melalui beberapa tahapan yang meliputi pengumpulan data, *data preparation*, pembentukan cluster menggunakan algoritma K-Means Clustering, pembangunan model prediksi menggunakan Decision Tree Regression, serta evaluasi model. Alur penelitian secara keseluruhan ditunjukkan pada Gambar 1.

Terakreditasi SINTA 5 SK :72/E/KPT/2024 Submitted: **03/06/2026**; Accepted: **28/06/2026**; Published: **30/07/2026**

JURNAL MEDIA INFORMATIKA [JUMIN]
Volume 7, Nomor 4, Juli – Agustus 2026, Page 1141-1150
ISSN 2808-005X (media online)
Available Online at http://ejournal.sisfokomtek.org/index.php/jumin



Gambar 1. Alur Penelitian

Berdasarkan Gambar 1, penelitian diawali dengan penggunaan Jakarta House Price Dataset yang terdiri atas 10.000 data rumah beserta atribut karakteristik properti. Tahap pertama yang dilakukan adalah data cleaning, yaitu menghapus data yang memiliki nilai kosong (*missing value*) dan data duplikat sehingga diperoleh 9.992 data yang siap diproses lebih lanjut.

Selanjutnya dilakukan penanganan *outlier* menggunakan metode Interquartile Range (IQR) pada atribut numerik price, bed_rooms, bath_rooms, carport, land_area, dan building_area. Tahap ini bertujuan menghilangkan data ekstrem yang berpotensi memengaruhi kualitas hasil clustering maupun prediksi sehingga diperoleh 7.523 data yang lebih representatif.

Tahap berikutnya adalah transformasi data menggunakan Label Encoding untuk mengubah atribut kategorikal district dan city menjadi bentuk numerik agar dapat diproses oleh algoritma *machine learning*. Setelah itu dilakukan standardisasi menggunakan StandardScaler sehingga seluruh atribut memiliki skala yang relatif sama sebelum proses clustering dilakukan.

Data hasil standardisasi kemudian diproses menggunakan algoritma K-Means Clustering untuk mengelompokkan rumah berdasarkan kemiripan karakteristiknya. Jumlah cluster optimal ditentukan menggunakan Elbow Method, kemudian label cluster yang dihasilkan ditambahkan sebagai atribut baru pada dataset.

Tahap selanjutnya adalah membangun model Decision Tree Regression menggunakan dua pendekatan, yaitu Baseline Decision Tree dan Hybrid K-Means Clustering–Decision Tree. Pada model hybrid, atribut cluster digunakan sebagai fitur tambahan sehingga model diharapkan mampu mengenali pola data yang lebih baik dibandingkan model baseline.

Tahap terakhir adalah evaluasi model menggunakan Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), R-Squared (R^2), dan Silhouette Score. Hasil evaluasi digunakan untuk mengukur kualitas clustering sekaligus membandingkan performa model baseline dan model hybrid dalam memprediksi harga rumah di Jakarta.

2.3 Data Preparation

Tahap *data preparation* bertujuan untuk menghasilkan dataset yang bersih, konsisten, dan siap digunakan dalam proses *clustering* maupun pemodelan. Pada tahap ini dilakukan beberapa proses pembersihan dan transformasi data agar kualitas dataset meningkat. Proses diawali dengan menghapus data yang memiliki *missing value* menggunakan fungsi dropna(), kemudian menghapus data yang terduplikasi menggunakan fungsi drop_duplicates(). Selanjutnya, penanganan *outlier* dilakukan dengan metode *Interquartile Range* (IQR) pada atribut *price*, *bed_rooms*, *bath_rooms*, *carport*, *land_area*, dan *building_area* untuk mengurangi pengaruh nilai-nilai ekstrem yang dapat memengaruhi hasil analisis.

Terakreditasi SINTA 5 SK :72/E/KPT/2024 Submitted: **03/06/2026**; Accepted: **28/06/2026**; Published: **30/07/2026**

JURNAL MEDIA INFORMATIKA [JUMIN]
Volume 7, Nomor 4, Juli – Agustus 2026, Page 1141-1150
ISSN 2808-005X (media online)
Available Online at http://ejournal.sisfokomtek.org/index.php/jumin

Setelah proses pembersihan data selesai, atribut kategorikal yaitu *district* dan *city* ditransformasikan ke dalam bentuk numerik menggunakan metode *Label Encoding* sehingga dapat diproses oleh algoritma machine learning. Tahap berikutnya adalah melakukan standardisasi data menggunakan *StandardScaler* agar seluruh atribut memiliki skala yang seragam. Standardisasi ini penting untuk mencegah atribut dengan rentang nilai yang lebih besar mendominasi proses perhitungan jarak pada algoritma *K-Means* maupun memengaruhi kinerja model prediksi. Dengan melalui seluruh tahapan *data preparation* tersebut, diperoleh dataset yang lebih representatif, berkualitas, dan mampu meningkatkan performa proses *clustering* serta akurasi model prediksi yang dibangun [14].

2.4 K-Means Clustering

Proses segmentasi rumah dilakukan menggunakan algoritma K-Means Clustering. Sebelum proses clustering dilakukan, seluruh atribut terlebih dahulu distandardisasi menggunakan StandardScaler agar setiap atribut memiliki kontribusi yang seimbang terhadap pembentukan cluster.

Jumlah cluster optimal ditentukan menggunakan Elbow Method dengan membandingkan nilai inertia pada beberapa jumlah cluster. Berdasarkan hasil evaluasi diperoleh tiga cluster sebagai jumlah cluster optimal [1], [2], [3], [10]. Model K-Means dibangun menggunakan parameter yang ditunjukkan pada Tabel 4.

Tabel 4. Parameter K-Means Clustering

Parameter	Nilai
Jumlah Cluster (*n_clusters*)	3
Inisialisasi (*init*)	k-means++
Maksimum Iterasi (*max_iter*)	300
Jumlah Inisialisasi (*n_init*)	auto
Random State	42

Berdasarkan Tabel 4, parameter tersebut digunakan agar proses clustering dapat direproduksi pada penelitian selanjutnya serta menghasilkan pembentukan cluster yang konsisten. Label cluster yang dihasilkan kemudian ditambahkan sebagai atribut baru pada dataset dan digunakan sebagai fitur tambahan pada model hybrid.

2.5 Decision Tree Regression

Prediksi harga rumah dilakukan menggunakan algoritma Decision Tree Regression. Penelitian ini membangun dua model, yaitu Baseline Decision Tree dan Hybrid K-Means Clustering–Decision Tree. Model baseline menggunakan atribut karakteristik rumah tanpa informasi cluster, sedangkan model hybrid memanfaatkan atribut cluster hasil K-Means sebagai fitur tambahan.

Sebelum proses pelatihan model dilakukan, dataset dibagi menjadi data latih dan data uji menggunakan metode train_test_split dengan random_state = 42. Proporsi pembagian data ditunjukkan pada Tabel 5.

Tabel 5. Pembagian Dataset

Dataset	Jumlah Data	Persentase
Data Latih (*Training*)	6.018	79,99%
Data Uji (*Testing*)	1.505	20,01%

Model Decision Tree dibangun menggunakan parameter yang ditunjukkan pada Tabel 6.

Tabel 6. Parameter Decision Tree Regression

Parameter	Nilai
Max Depth	3
Min Samples Split	20
Min Samples Leaf	10
Random State	42

Pengaturan parameter tersebut bertujuan menghasilkan struktur pohon keputusan yang lebih sederhana sehingga mampu mengurangi risiko *overfitting* dan meningkatkan kemampuan generalisasi model. Pendekatan

Terakreditasi SINTA 5 SK :72/E/KPT/2024 Submitted: **03/06/2026**; Accepted: **28/06/2026**; Published: **30/07/2026**

JURNAL MEDIA INFORMATIKA [JUMIN]
Volume 7, Nomor 4, Juli – Agustus 2026, Page 1141-1150
ISSN 2808-005X (media online)
Available Online at http://ejournal.sisfokomtek.org/index.php/jumin

yang mengombinasikan K-Means Clustering dan Decision Tree telah digunakan dalam berbagai penelitian *data mining* untuk meningkatkan kualitas segmentasi maupun analisis data [13], [15]. Selain itu, kombinasi clustering dan model prediksi juga terbukti mampu membantu model mengenali pola data yang lebih kompleks [6].

2.6 Evaluasi Model

Evaluasi dilakukan untuk mengukur kualitas model prediksi maupun hasil clustering. Performa model prediksi dievaluasi menggunakan tiga metrik, yaitu Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), dan R-Squared (R^2).

Nilai MAE digunakan untuk mengukur rata-rata kesalahan absolut antara nilai aktual dan hasil prediksi. RMSE digunakan untuk mengukur besar kesalahan prediksi dengan memberikan penalti yang lebih besar terhadap kesalahan bernilai tinggi, sedangkan R^2 digunakan untuk mengukur kemampuan model dalam menjelaskan variasi data.

Selain itu, kualitas hasil clustering dievaluasi menggunakan Silhouette Score yang mengukur tingkat kemiripan data di dalam satu cluster dibandingkan dengan cluster lainnya. Semakin tinggi nilai Silhouette Score, semakin baik kualitas cluster yang dihasilkan.

Hasil evaluasi tersebut digunakan untuk membandingkan performa Baseline Decision Tree dan Hybrid K-Means Clustering–Decision Tree dalam memprediksi harga rumah di Jakarta serta mengevaluasi kualitas segmentasi yang dihasilkan oleh algoritma K-Means Clustering.

3. HASIL DAN PEMBAHASAN

3.1 Hasil Data Preparation

Tahap *data preparation* dilakukan untuk menghasilkan dataset yang bersih dan siap digunakan pada proses clustering maupun prediksi harga rumah. Dataset awal terdiri atas 10.000 data rumah yang diperoleh dari Jakarta House Price Dataset. Proses pembersihan diawali dengan menghapus data yang memiliki *missing value* dan data duplikat sehingga jumlah data berkurang menjadi 9.992 data. Selanjutnya dilakukan penanganan *outlier* menggunakan metode Interquartile Range (IQR) pada atribut numerik, yaitu *price*, *bed_rooms*, *bath_rooms*, *carport*, *land_area*, dan *building_area*. Setelah proses tersebut diperoleh dataset akhir sebanyak 7.523 data yang digunakan pada tahap pemodelan.

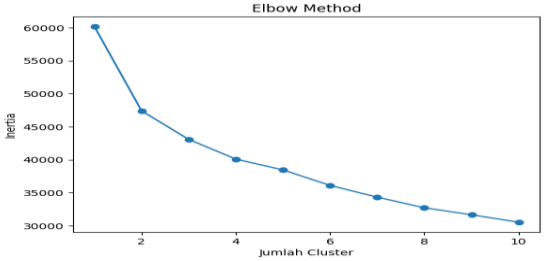
Tabel 7. Hasil Data Preparation

Tahap	Jumlah Data
Dataset Awal	10.000
Setelah Cleaning	9.992
Setelah Outlier Removal	7.523

Berdasarkan Tabel 7, proses *data preparation* berhasil menghilangkan data yang tidak lengkap serta data ekstrem yang berpotensi memengaruhi proses pembentukan cluster dan model prediksi. Penghapusan *outlier* menghasilkan data yang lebih homogen sehingga diharapkan mampu meningkatkan stabilitas model machine learning. Tahapan ini juga penting untuk mengurangi pengaruh nilai yang menyimpang terhadap proses pembelajaran model sehingga hasil prediksi menjadi lebih representatif.

3.2 Hasil Clustering Menggunakan K-Means

Proses clustering dilakukan menggunakan algoritma K-Means Clustering. Sebelum pembentukan cluster, jumlah cluster optimal ditentukan menggunakan Elbow Method dengan membandingkan nilai *inertia* pada beberapa jumlah cluster. Berdasarkan hasil visualisasi pada Gambar 2, titik siku (*elbow point*) terlihat pada nilai K = 3, sehingga penelitian ini menggunakan tiga cluster.

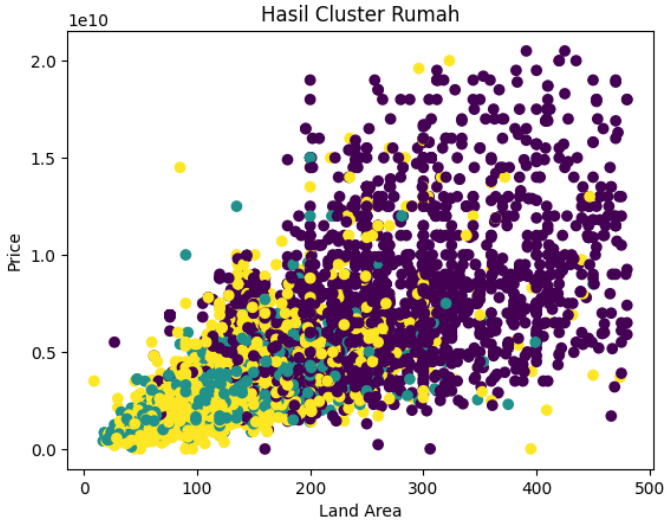


Gambar 2. Elbow Method

Terakreditasi SINTA 5 SK :72/E/KPT/2024 Submitted: **03/06/2026**; Accepted: **28/06/2026**; Published: **30/07/2026**

JURNAL MEDIA INFORMATIKA [JUMIN]
Volume 7, Nomor 4, Juli – Agustus 2026, Page 1141-1150
ISSN 2808-005X (media online)
Available Online at http://ejournal.sisfokomtek.org/index.php/jumin

Setelah jumlah cluster ditentukan, algoritma K-Means digunakan untuk mengelompokkan rumah berdasarkan karakteristik yang dimiliki. Hasil pengelompokan ditampilkan pada Gambar 3.



Gambar 3. Hasil Clustering Rumah

Kualitas hasil clustering dievaluasi menggunakan Silhouette Score dan menghasilkan nilai sebesar 0,1436.

Tabel 8. Hasil Evaluasi Clustering

Metrik	Nilai
Silhouette Score	0,1436

Berdasarkan Tabel 8, nilai **Silhouette Score sebesar 0,1436** menunjukkan bahwa proses clustering mampu membentuk tiga kelompok rumah berdasarkan karakteristik yang dimiliki, namun tingkat pemisahan antar cluster masih tergolong rendah. Nilai tersebut mengindikasikan bahwa masih terdapat beberapa data yang memiliki kemiripan dengan anggota pada cluster lain sehingga batas antar kelompok belum terbentuk secara jelas.

Kondisi tersebut diduga disebabkan oleh karakteristik harga rumah di Jakarta yang memiliki variasi tinggi serta adanya tumpang tindih (*overlap*) antar atribut, seperti luas tanah, luas bangunan, jumlah kamar, dan lokasi. Akibatnya, beberapa rumah memiliki karakteristik yang hampir serupa meskipun berada pada kelompok yang berbeda.

Meskipun demikian, informasi cluster tetap dimanfaatkan sebagai fitur tambahan pada model prediksi karena mampu memberikan informasi mengenai karakteristik kelompok rumah yang tidak diperoleh secara langsung dari atribut individual. Temuan ini sejalan dengan penelitian [6] yang menyatakan bahwa informasi hasil clustering dapat membantu model prediksi dalam mengenali pola data yang lebih kompleks meskipun kualitas cluster belum sepenuhnya optimal.

3.3 Hasil Prediksi Harga Rumah

Penelitian ini membandingkan dua model prediksi, yaitu Baseline Decision Tree dan Hybrid K-Means Clustering–Decision Tree. Perbandingan performa kedua model ditunjukkan pada Tabel 5.

Tabel 9. Perbandingan Hasil Prediksi

Model	MAE	RMSE	R^2
Baseline Decision Tree	1.284.156.810	2.321.111.672	0,5156
Hybrid K-Means + Decision Tree	1.388.181.177	2.097.559.711	0,6044

Berdasarkan Tabel 9, model hybrid menghasilkan nilai RMSE yang lebih rendah dibandingkan model baseline, yaitu 2.097.559.711 dibandingkan 2.321.111.672. Selain itu, nilai R^2 meningkat dari 0,5156 menjadi 0,6044, yang menunjukkan bahwa model hybrid mampu menjelaskan sekitar 60,44% variasi harga rumah pada dataset yang digunakan.

Terakreditasi SINTA 5 SK :72/E/KPT/2024
Submitted: **03/06/2026**; Accepted: **28/06/2026**; Published: **30/07/2026**

JURNAL MEDIA INFORMATIKA [JUMIN]
Volume 7, Nomor 4, Juli – Agustus 2026, Page 1141-1150
ISSN 2808-005X (media online)
Available Online at http://ejournal.sisfokomtek.org/index.php/jumin

Peningkatan nilai R^2 menunjukkan bahwa penambahan atribut cluster memberikan informasi tambahan mengenai karakteristik kelompok rumah sehingga model mampu mempelajari pola hubungan antar atribut dengan lebih baik. Namun demikian, nilai MAE pada model hybrid sedikit lebih besar dibandingkan model baseline. Hal ini menunjukkan bahwa meskipun secara keseluruhan model hybrid memiliki kemampuan prediksi yang lebih baik berdasarkan RMSE dan R^2, masih terdapat beberapa prediksi individual yang memiliki selisih absolut cukup besar terhadap nilai sebenarnya.

Hasil penelitian ini sejalan dengan penelitian [6] yang menunjukkan bahwa pemanfaatan hasil clustering sebagai fitur tambahan dapat membantu model prediksi dalam menangkap pola data yang lebih kompleks dibandingkan penggunaan model prediksi tanpa proses segmentasi data.

3.4 Analisis Faktor Yang Mempengaruhi Harga Rumah

Analisis feature importance dilakukan untuk mengetahui atribut yang paling berpengaruh terhadap prediksi harga rumah. Hasil analisis ditunjukkan pada Tabel 10.

Tabel 10. Feature Importance

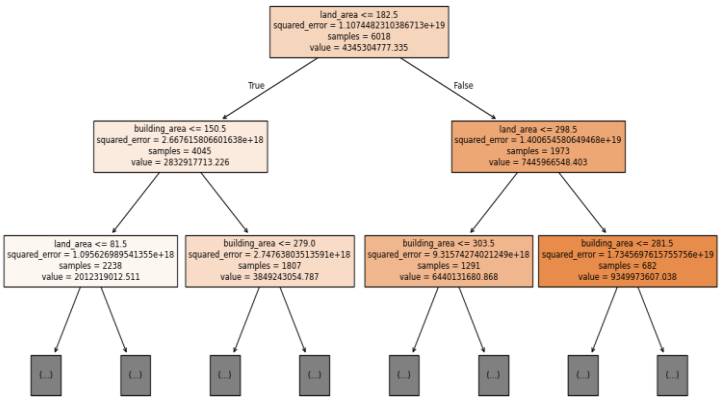
Fitur	Importance
land_area	0,829593
building_area	0,170407

Berdasarkan Tabel 6, atribut land_area memiliki nilai *feature importance* sebesar 0,829593, sedangkan building_area memiliki nilai 0,170407. Hasil tersebut menunjukkan bahwa luas tanah merupakan faktor yang paling dominan dalam menentukan harga rumah pada dataset penelitian.

Besarnya kontribusi atribut land_area menunjukkan bahwa nilai lahan masih menjadi faktor utama dalam pembentukan harga rumah di Jakarta. Kondisi ini sesuai dengan karakteristik pasar properti di wilayah perkotaan, di mana keterbatasan lahan menyebabkan harga tanah memiliki pengaruh yang lebih besar dibandingkan luas bangunan. Temuan ini juga sejalan dengan beberapa penelitian sebelumnya [6], [10], [11] yang menyatakan bahwa luas tanah merupakan salah satu variabel utama dalam prediksi harga properti.

3.5 Analisis Decision Tree

Visualisasi struktur Decision Tree pada Gambar 4 menunjukkan bahwa atribut land_area menjadi root node dalam proses prediksi harga rumah. Hal ini menunjukkan bahwa luas tanah merupakan atribut pertama yang digunakan model untuk membagi data karena memiliki kemampuan paling tinggi dalam mengurangi nilai kesalahan (*impurity*) selama proses pembentukan pohon keputusan.



Gambar 4. Struktur Decision Tree Hybrid

Pada percabangan berikutnya, model menggunakan atribut building_area sebagai variabel pembeda untuk menghasilkan prediksi harga rumah yang lebih spesifik. Rumah dengan luas tanah yang lebih besar cenderung memiliki nilai prediksi harga yang lebih tinggi dibandingkan rumah dengan luas tanah yang lebih kecil. Sementara itu, luas bangunan digunakan untuk membedakan harga rumah pada kelompok yang memiliki luas tanah relatif serupa.

Struktur pohon keputusan tersebut menunjukkan bahwa model membangun proses prediksi secara bertahap berdasarkan aturan-aturan keputusan yang mudah diinterpretasikan. Salah satu keunggulan Decision

Terakreditasi SINTA 5 SK :72/E/KPT/2024 Submitted: **03/06/2026**; Accepted: **28/06/2026**; Published: **30/07/2026**

JURNAL MEDIA INFORMATIKA [JUMIN]
Volume 7, Nomor 4, Juli – Agustus 2026, Page 1141-1150
ISSN 2808-005X (media online)
Available Online at http://ejournal.sisfokomtek.org/index.php/jumin

Tree dibandingkan beberapa algoritma lain adalah kemampuannya menghasilkan aturan prediksi yang dapat dipahami secara langsung oleh pengguna. Oleh karena itu, selain memberikan performa prediksi yang cukup baik, model yang dihasilkan juga mampu menjelaskan faktor-faktor utama yang memengaruhi harga rumah.

Hasil visualisasi ini konsisten dengan analisis *feature importance*, di mana land_area dan building_area menjadi dua atribut yang memberikan kontribusi terbesar terhadap proses prediksi harga rumah.

4. KESIMPULAN

Penelitian ini berhasil menerapkan pendekatan Hybrid K-Means Clustering dan Decision Tree untuk melakukan segmentasi serta prediksi harga rumah di Jakarta menggunakan *Jakarta House Price Dataset*. Hasil proses clustering menghasilkan tiga cluster dengan nilai Silhouette Score sebesar 0,1436. Meskipun nilai tersebut menunjukkan bahwa kualitas pemisahan antar cluster masih tergolong rendah, informasi cluster tetap dapat dimanfaatkan sebagai fitur tambahan pada proses prediksi.

Hasil evaluasi menunjukkan bahwa model Hybrid K-Means Clustering–Decision Tree memperoleh nilai RMSE sebesar 2.097.559.711 dan R^2 sebesar 0,6044, yang lebih baik dibandingkan model Baseline Decision Tree. Peningkatan nilai R^2 tersebut menunjukkan bahwa penambahan informasi hasil clustering mampu membantu model dalam menjelaskan variasi data harga rumah dengan lebih baik. Selain itu, hasil analisis *feature importance* dan struktur Decision Tree menunjukkan bahwa luas tanah (*land_area*) dan luas bangunan (*building_area*) merupakan faktor yang paling berpengaruh terhadap prediksi harga rumah di Jakarta.

Penelitian ini masih memiliki beberapa keterbatasan. Nilai Silhouette Score yang relatif rendah menunjukkan bahwa kualitas cluster yang terbentuk belum optimal sehingga karakteristik antar cluster masih memiliki tingkat kemiripan yang cukup tinggi. Selain itu, penelitian ini hanya membandingkan pendekatan hybrid dengan model Decision Tree sebagai model dasar sehingga belum dapat menunjukkan performanya terhadap algoritma lain yang lebih kompleks, seperti Random Forest, Gradient Boosting, atau XGBoost.

Oleh karena itu, penelitian selanjutnya disarankan untuk mengeksplorasi metode clustering yang lebih sesuai dengan karakteristik data, melakukan optimasi parameter model, serta membandingkan pendekatan hybrid dengan algoritma prediksi lainnya agar diperoleh model prediksi harga rumah yang memiliki tingkat akurasi dan kemampuan generalisasi yang lebih baik.

UCAPAN TERIMAKASIH

Penulis mengucapkan terima kasih kepada Tuhan Yang Maha Esa atas segala rahmat dan karunia-Nya sehingga penelitian ini dapat diselesaikan dengan baik. Penulis juga mengucapkan terima kasih kepada dosen pembimbing, keluarga, serta seluruh pihak yang telah memberikan dukungan dan bantuan selama proses penelitian. Selain itu, penulis mengucapkan terima kasih kepada Kaggle yang telah menyediakan dataset yang digunakan dalam penelitian ini.

REFERENCES

[1] N. Septiani and R. Herdiana, "Penerapan Algoritma K-Means Clustering Untuk Harga Rumah di Jakarta Selatan Nuraeni Septiani Sekolah Tinggi Manajemen Informatika dan Komputer (STMIK) IKMI Cirebon Saeful Anwar Sekolah Tinggi Manajemen Informatika dan Komputer (STMIK) IKMI Cirebon," *Trending J. Ekon. Akunt. dan Manaj.*, vol. 1, no. 2, pp. 35–47, 2023.

[2] A. Suderajat, A. Z. Kamalia, and D. Afandi, "Segmentasi Harga Pasar Properti Rumah Menggunakan Algoritma K-Means Berdasarkan Lokasi, Harga dan Luas Bangunan," *DINAMIK*, vol. 30, no. 11, pp. 1–10, 2025.

[3] B. G. Aji, D. C. A. Sondawa, M. R. Gifari, and S. Wijayanto, "Penerapan Algoritma K-Means Untuk Clustering Harga Rumah Di Bandung," *J. Ilm. Inform. Glob.*, vol. 14, no. 2, pp. 17–23, 2023, doi: 10.36982/jiig.v14i2.3189.

[4] R. A. Saputra and A. Pratama, "Implementasi Decision Tree Untuk Prediksi Harga Rumah Di Daerah Tebet," *J. Inf. Syst. Manag.*, vol. 6, no. 2, pp. 164–170, 2025, doi: 10.24076/joism.2025v6i2.1928.

[5] N. N. Sari, T. T. Anisah, and R. Fitriani, "Implementasi Machine Learning Untuk Prediksi Harga Laptop Menggunakan Algoritma Regresi Linear Berganda," *J. Manaj. Inform.*, vol. 14, no. 2, pp. 162–177, 2024, doi: 10.34010/jamika.v14i2.12923.

[6] Z. Huang and G. Lai, "A House Price Prediction Model Based on K-means Clustering and Random Forest in Guangzhou," *Front. Business, Econ. Manag.*, vol. 10, no. 2, pp. 377–381, 2023, doi: 10.54097/fbem.v10i2.11077.

[7] A. Hjort, I. Scheel, D. E. Sommervoll, and J. Pensar, "Locally interpretable tree boosting: An application

Terakreditasi SINTA 5 SK :72/E/KPT/2024 Submitted: **03/06/2026**; Accepted: **28/06/2026**; Published: **30/07/2026**

to house price prediction," *Decis. Support Syst.*, vol. 178, no. September 2023, p. 114106, 2024, doi: 10.1016/j.dss.2023.114106.

[8] S. Lahmiri, S. Bekiros, and C. Avdoulas, "A comparative assessment of machine learning methods for predicting housing prices using Bayesian optimization," *Decis. Anal. J.*, vol. 6, no. November 2022, p. 100166, 2023, doi: 10.1016/j.dajour.2023.100166.

[9] M. Salsabila, R. P. Adiwijaya, and L. N. Octavia, "Implementasi Machine Learning Untuk Memprediksi Harga Rumah Dengan Menggunakan Metode Decision Tree Regressor Dan Random Forest Regressor," *Setrum Sist. Kendali-Tenaga-elektronika-telekomunikasi-komputer*, vol. 14, no. 1, p. 46, 2025, doi: 10.62870/setrum.v14i1.31025.

[10] V. R. Maulani, M. A. Barata, and P. E. Yuwita, "A Improving House Price Clustering Results with K-means through the Implementation of One-hot Encoding Pre-processing Technique," *J. Appl. Informatics Comput.*, vol. 9, no. 3, pp. 741–748, 2025, doi: 10.30871/jaic.v9i3.9481.

[11] B. A. Ghaisani, G. A. Damayanti, R. M. A. Rani, N. Fitri, and W. Steven, "Pemodelan Klasifikasi Kategori Harga Rumah Menggunakan Algoritma Decision Tree dengan Pendekatan CRISP-DM," *J. Inform.*, vol. 11, no. 2, 2025.

[12] S. Muchammad Rosyid Aridho, A. Hasna Khaira Aswha, T. Dwi Wahyuni, and A. Arum Sari, "Analisis Faktor yang Mempengaruhi Harga Rumah Menggunakan Decision Tree," *Pros. Semin. Nas. Teknol. Inf. dan Bisnis*, pp. 386–392, 2025, doi: 10.47701/f1bwa603.

[13] H. A. Maulana, N. Fitriyah, D. El Baradei, F. Ardiyanto, and M. Arifin, "Analisis Data Pendidikan Menggunakan Metode Data Mining dengan Algoritma Decision Tree dan K-Means Clustering," *J. Tek. Inform.*, vol. 6, no. 1, 2026.

[14] L. Oluwaseye Joel, W. Doorsamy, and B. Sena Paul, "A Review of Missing Data Handling Techniques for Machine Learning," *Int. J. Innov. Technol. Interdiscip. Sci.*, vol. 5, no. 3, pp. 971–1005, 2022, [Online]. Available: https://doi.org/10.15157/IJITIS.2022.5.3.971-1005

[15] T. P. Kaloka, A. Marta, and N. D. Syofrin, "K-Means and Decision Tree Models for Analysis Students Perception of Digital Marketing," *Indones. J. Mach. Learn. Comput. Sci.*, vol. 6, no. April, pp. 453–461, 2026.