

ANALISIS PENGARUH FITUR PROPERTI DAN PERBANDINGAN ALGORITMA MULTIPLE LINEAR REGRESSION, RANDOM FOREST REGRESSION DAN SUPPORT VECTOR REGRESSION DALAM PREDIKSI NILAI JUAL RUMAH DI KECAMATAN CIDAHU

Mohammad Samani Kamil^{1*} Iwan Rizal Setiawan² Prajoko³

¹Program Studi Teknik Informatika, Fakultas Sains dan Teknologi, Universitas Muhammadiyah Sukabumi, Jl. R. Syamsudin, S.H. No.50, Sukabumi, Jawa Barat, Telp. (0266) 218 345

Keywords: Machine Learning, Multiple Linear Regression, Prediksi Harga Rumah, Random Forest Regression, Seleksi Fitur, Streamlit.

Correspondent Email:
samanikamil2@gmail.com



Copyright © 2026 The Author(s), Published by Jurnal Informatika dan Teknik Elektro Terapan. This is an open-access article distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

Abstrak. Penelitian ini menganalisis pengaruh fitur properti terhadap nilai jual rumah di Kecamatan Cidahu, Kabupaten Sukabumi, menggunakan tiga algoritma Machine Learning Multiple Linear Regression (MLR), Random Forest Regression (RFR), dan Support Vector Regression (SVR) dengan kerangka kerja CRISP-DM. Data dikumpulkan melalui web scraping platform Pinhome (435 data mentah) dan diperkaya menggunakan SMOTE menjadi 2.104 data final. Seleksi fitur tiga tahap (korelasi Spearman, uji VIF, dan Feature Importance) menghasilkan dua fitur terpilih: luas tanah dan luas bangunan. Luas tanah terbukti sebagai fitur paling dominan ($\beta^* = 0,826260$, $t = 40,29$, $p < 0,001$). Evaluasi model menunjukkan RFR meraih kinerja terbaik dengan $R^2 = 0,9532$, MAE = Rp56.580.528, RMSE = Rp117.368.937, dan MAPE = 8,09%, mengungguli MLR ($R^2 = 0,5657$) dan SVR ($R^2 = 0,1994$). Pengujian kestabilan 5-Fold Cross-Validation memperkuat keunggulan RFR dengan rerata $R^2 = 0,9446$. Model RFR diimplementasikan dalam aplikasi web berbasis Streamlit sebagai prototipe estimasi harga yang objektif dan transparan.

Abstract. This study analyzes the impact of property features on house selling prices in Cidahu District, Sukabumi Regency, using three machine learning algorithms—Multiple Linear Regression (MLR), Random Forest Regression (RFR), and Support Vector Regression (SVR)—within the CRISP-DM framework. Data were collected via web scraping from the Pinhome platform (435 raw data points) and augmented using SMOTE to reach a final dataset of 2,104 entries. A three-stage feature selection process (Spearman correlation, VIF test, and Feature Importance) identified two key features: land area and building area. Land area proved to be the most dominant feature ($\beta^* = 0.826260$, $t = 40.29$, $p < 0.001$). Model evaluation revealed that RFR achieved the best performance—with $R^2 = 0.9532$, MAE = Rp56,580,528, RMSE = Rp117,368,937, and MAPE = 8.09%—outperforming MLR ($R^2 = 0.5657$) and SVR ($R^2 = 0.1994$). Stability testing via 5-fold cross-validation confirmed RFR's superiority, yielding an average R^2 of 0.9446. The RFR model was implemented in a Streamlit-based web application as a prototype for objective and transparent price estimation.

1. PENDAHULUAN

Sektor properti mengalami perkembangan signifikan seiring urbanisasi dan pertumbuhan ekonomi nasional. Rumah sebagai kebutuhan dasar sekaligus instrumen investasi memiliki posisi strategis dalam perekonomian. Fluktuasi harga yang tidak menentu menjadikan penentuan nilai jual rumah sebagai isu krusial bagi pembeli, penjual, dan pengembang [1].

Kecamatan Cidahu, Kabupaten Sukabumi, merupakan wilayah periurban yang mengalami perkembangan properti pesat, namun belum terdapat kajian empiris yang menganalisis faktor penentu nilai jual rumah secara kuantitatif di wilayah tersebut. Transaksi properti masih bergantung pada penilaian subjektif penilai independen yang bekerja secara manual, dengan kelemahan mendasar berupa ketergantungan pada pertimbangan subjektif dan survei lapangan yang memakan waktu dan biaya [2].

Machine Learning muncul sebagai solusi yang menjanjikan karena kemampuannya mengekstraksi pola dari data historis dalam jumlah besar untuk menghasilkan prediksi harga yang lebih akurat dan efisien dibandingkan penilaian manual [3]. Landasan teoritis yang relevan adalah Hedonic Pricing Model (HPM), yang menyatakan bahwa harga suatu properti ditentukan oleh karakteristik-karakteristik pembentuknya, meliputi atribut fisik (luas bangunan, jumlah kamar), atribut lokasi, dan atribut lingkungan [4].

Penelitian terdahulu menunjukkan bahwa algoritma regresi seperti Multiple Linear Regression (MLR) mampu menghasilkan akurasi prediksi hingga 85% pada dataset properti [5], namun masih terdapat gap berupa: (1) belum adanya penerapan HPM di skala mikro wilayah Jawa Barat khususnya Kecamatan Cidahu, (2) seleksi variabel yang umumnya dilakukan secara manual bukan melalui pipeline Machine Learning, dan (3) perbandingan algoritma yang komprehensif antara metode regresi linier dan metode ensemble/kernel masih jarang dilakukan pada dataset properti lokal berskala kecil.

Penelitian ini bertujuan untuk: (1) mengidentifikasi fitur properti yang paling dominan memengaruhi nilai jual rumah di Kecamatan Cidahu menggunakan pipeline seleksi fitur otomatis, dan (2) membandingkan kinerja algoritma MLR, Random Forest

Regression (RFR), dan Support Vector Regression (SVR) untuk menentukan algoritma terbaik pada konteks data properti lokal.

2. TINJAUAN PUSTAKA

Hedonic Pricing Model (HPM) pertama kali dikembangkan oleh Rosen (1974) sebagai pendekatan ekonometrika untuk menganalisis harga barang heterogen berdasarkan fitur-fitur pembentuknya. Dalam konteks properti, harga rumah (P) merupakan fungsi dari atribut fisik, lokasi, dan lingkungan. HPM menjadi landasan teoritis penelitian ini karena secara eksplisit menghubungkan karakteristik fisik properti dengan nilai pasarnya.

Machine Learning adalah cabang kecerdasan buatan yang memungkinkan sistem komputer belajar dan meningkatkan kinerjanya melalui pemrosesan data tanpa harus diprogram ulang secara eksplisit [6]. Dalam konteks prediksi harga properti, pendekatan Supervised Learning digunakan di mana model dilatih menggunakan data berlabel untuk memetakan hubungan antara fitur input dan harga output.

Multiple Linear Regression (MLR) adalah teknik statistik yang mengkaji hubungan linier antara satu variabel dependen (Y) dan dua atau lebih variabel independen (X) secara bersamaan [7]. Model dinyatakan sebagai $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \varepsilon$. Keunggulan MLR adalah interpretabilitas koefisien yang tinggi, namun terbatas pada hubungan linier.

Random Forest Regression (RFR) adalah algoritma ensemble yang menciptakan banyak decision tree secara paralel menggunakan teknik bagging dan random feature selection [8]. Hasil regresinya diperoleh dari rata-rata prediksi semua pohon. Keunggulan RFR meliputi kemampuan mengelola hubungan non-linier, tidak sensitif terhadap outlier, dan menghasilkan Feature Importance secara alami.

Support Vector Regression (SVR) adalah adaptasi Support Vector Machine untuk masalah regresi yang menggunakan fungsi kernel (RBF, polinomial, atau linier) untuk menggambarkan hubungan non-linier yang kompleks antara fitur dan target [9]. SVR berupaya menemukan fungsi prediksi dengan deviasi maksimum epsilon (ϵ) dari nilai target.

Seleksi fitur adalah proses memilih subset fitur paling informatif untuk meningkatkan performa model dan mengurangi kompleksitas komputasi [10]. Metode yang digunakan

mencakup analisis korelasi Spearman untuk mengukur kekuatan hubungan monoton, uji Variance Inflation Factor (VIF) untuk mendeteksi multikolinieritas, dan Feature Importance dari Random Forest untuk mengukur kontribusi setiap fitur.

CRISP-DM (Cross Industry Standard Process for Data Mining) adalah kerangka kerja metodologis terstandarisasi yang menyediakan alur kerja terstruktur untuk proyek analisis data, meliputi Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, dan Deployment[11]. SMOTE (Synthetic Minority Over-sampling Technique) digunakan untuk mengatasi ketidakseimbangan dataset dengan menghasilkan data sintetis melalui interpolasi linier antar sampel menggunakan k-Nearest Neighbors [12].

Beberapa penelitian terdahulu telah mengkaji prediksi harga properti menggunakan Machine Learning. Maulana & Saputra (2025) menerapkan Linear Regression untuk memprediksi harga rumah di Tebet dengan $R^2 = 0,7713$ [13]. Labib et al. (2021) mengaplikasikan MLR untuk prediksi harga rumah di Jakarta Selatan dengan akurasi 66% [14]. Rianti et al. (2023) membandingkan Linear Regression, Random Forest, dan XGBoost pada dataset Kaggle [15]. Perbedaan penelitian ini adalah penerapan HPM di wilayah periurban Kecamatan Cidahu dengan pipeline seleksi fitur otomatis dan perbandingan tiga algoritma secara komprehensif.

3. METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif dengan kerangka kerja metodologis CRISP-DM. Objek penelitian adalah rumah hunian yang dijual di wilayah Kecamatan Cidahu, Kabupaten Sukabumi. Variabel dependen adalah harga jual rumah (Rupiah), sedangkan variabel independen awal terdiri dari luas bangunan (m^2), luas tanah (m^2), jumlah kamar tidur, dan jumlah kamar mandi.

A. Data Collection. Data dikumpulkan melalui web scraping menggunakan Python dengan library Requests dan BeautifulSoup pada platform Pinhome untuk listing rumah di Kecamatan Cidahu. Total data mentah yang diperoleh adalah 435 baris dengan 5 variabel. Proses scraping dilakukan dengan pendekatan berbasis HTTP request yang mengekstraksi konten HTML secara otomatis.

B. Data Preparation. Tahap persiapan data meliputi: (1) deduplikasi menghasilkan 50 data unik, (2) augmentasi data menggunakan SMOTE dengan target 750 sampel per kelas harga (4 kelas) menggunakan konfigurasi $k_neighbors = 3$ dan $random_state = 42$, menghasilkan 3.000 data sintetis, (3) post-processing untuk menjaga karakteristik riil variabel, (4) deteksi outlier menggunakan metode IQR dengan menghapus baris yang memiliki outlier pada ≥ 2 kolom secara bersamaan, menghasilkan dataset final 2.104 baris, (5) transformasi $\log_{10}$ pada variabel target untuk mengatasi distribusi right-skewed, (6) pembagian data latih (80% = 1.683 baris) dan data uji (20% = 421 baris).

C. Feature Selection. Seleksi fitur dilakukan melalui tiga tahap pada data latih: (1) Analisis Korelasi Spearman — fitur dipertahankan jika $|\rho| \geq 0,10$, (2) Uji VIF — fitur dipertahankan jika $VIF < 10$, (3) Feature Importance Random Forest — fitur dipertahankan jika skor $\geq 0,05$ (5%). Fitur final adalah fitur yang lolos ketiga tahap secara bersamaan (irisan/interseksi), memastikan pemilihan berbasis bukti statistik dari data.

D. Modeling. Tiga algoritma dibangun dan dibandingkan: (1) MLR diestimasi menggunakan OLS via statsmodels untuk inferensi statistik dan scikit-learn untuk prediksi, dengan Robust Standard Error HC3 untuk mengatasi heteroskedastisitas, (2) RFR dengan hyperparameter tuning menggunakan Grid Search 5-fold CV pada parameter $n_estimators \{100, 300, 500\}$, $max_depth \{None, 5, 10\}$, dan $min_samples_leaf \{1, 2, 4\}$, (3) SVR dengan kernel RBF, StandardScaler pada fitur dan target, serta Grid Search 5-fold CV pada parameter $C \{1, 10, 100, 1000\}$, $epsilon \{0,01, 0,05, 0,1, 0,5\}$, dan $gamma \{scale, auto\}$.

E. Evaluation. Evaluasi model menggunakan metrik R^2 , MAE, RMSE, dan MAPE pada skala Rupiah asli (setelah invers transformasi $\exp(m)$). Uji asumsi klasik MLR mencakup uji normalitas (Shapiro-Wilk), multikolinieritas (VIF), dan heteroskedastisitas (Breusch-Pagan). Stabilitas model diuji menggunakan 5-Fold Cross-Validation, dengan perbedaan kinerja antar algoritma diuji menggunakan Wilcoxon Signed-Rank Test ($\alpha = 0,05$).

F. Deployment. Model terbaik (RFR) diimplementasikan dalam aplikasi web interaktif berbasis Streamlit yang dapat diakses melalui browser tanpa instalasi khusus. Aplikasi menyediakan formulir input untuk keempat variabel fisik properti dengan validasi otomatis dan analisis komparasi posisi pasar.

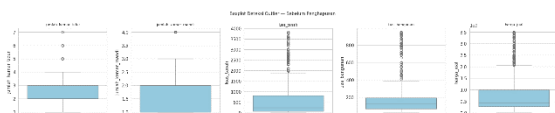
4. HASIL DAN PEMBAHASAN

A. Statistik Deskriptif dan Persiapan Data

Tabel 1. Statistik Deskriptif Dataset

Variabel	Min	Max	Mean	Median	Std Dev
Luas Bangunan (m ²)	30	950	150,44	121,24	127,56
Luas Tanah (m ²)	36	3.830	531,82	252,09	618,22
Kamar Tidur	1	7	2,52	2,00	0,94
Kamar Mandi	1	4	1,71	2,00	0,74
Harga Jual (Rp)	35 jt	3,5 M	776 jt	450 jt	640 jt

Dataset final terdiri dari 2.104 baris data setelah proses augmentasi SMOTE dan pembersihan outlier. Berdasarkan statistik deskriptif, harga jual rumah di Kecamatan Cidahu memiliki rentang dari Rp35.000.000 hingga Rp3.500.000.000, dengan mean Rp776.766.200 dan median Rp450.000.000. Selisih mean-median yang signifikan mengonfirmasi distribusi right-skewed, yang menjustifikasi penggunaan transformasi log1p pada variabel target. Luas tanah menunjukkan variasi tertinggi dengan standar deviasi 618,22 m² dan nilai maksimum 3.830 m².



Deteksi outlier menggunakan IQR mengidentifikasi persentase outlier per kolom berkisar 0% hingga 4,86%. Kebijakan penghapusan baris dengan outlier pada ≥ 2 kolom secara bersamaan berhasil menjaga kualitas data tanpa mengurangi sampel secara berlebihan, menghasilkan dataset bersih 2.104 baris yang dibagi menjadi 1.683 data latih dan 421 data uji.

B. Hasil Seleksi Fitur

Tabel 2. Hasil Seleksi Fitur

Fitur	Spearman ρ	VIF	Importance	Status
Luas Tanah (m ²)	0,9120	1,8315	90,09%	Lolos $\checkmark$
Luas Bangunan (m ²)	0,7475	1,7282	6,87%	Lolos $\checkmark$
Kamar Tidur	0,5820	1,3882	1,95%	Tidak Lolos $\times$
Kamar Mandi	0,7142	1,6661	1,09%	Tidak Lolos $\times$

Pada tahap korelasi Spearman, keempat fitur awal lolos dengan nilai korelasi: luas tanah ($\rho = 0,9120$), luas bangunan ($\rho = 0,7475$), jumlah kamar mandi ($\rho = 0,7142$), dan jumlah kamar tidur ($\rho = 0,5820$). Seluruh nilai berada jauh di atas ambang batas 0,10, menunjukkan hubungan monoton yang kuat dengan harga jual.

Pada tahap uji VIF, semua fitur memiliki nilai VIF di bawah 10 (luas bangunan: 1,7282, luas tanah: 1,8315, kamar tidur: 1,3882, kamar mandi: 1,6661), sehingga tidak terdapat masalah multikolinieritas yang signifikan dan keempat fitur lolos tahap kedua.

Pada tahap Feature Importance Random Forest, hanya dua fitur yang memiliki skor $\geq 5\%$: luas tanah (90,09%) dan luas bangunan (6,87%). Jumlah kamar tidur (1,95%) dan kamar mandi (1,09%) tidak lolos ambang batas ini. Skor Feature Importance luas tanah yang mencapai 90,09% mengkonfirmasi dominasi absolut fitur ini dalam membentuk nilai properti di Kecamatan Cidahu. Dengan demikian, hanya luas tanah dan luas bangunan yang digunakan sebagai fitur final dalam pemodelan.

C. Hasil Pemodelan dan Uji Asumsi Klasik MLR

Model MLR menghasilkan $R^2 = 0,706$ (Adjusted $R^2 = 0,706$) pada data latih. Uji F dengan Robust HC3 menghasilkan $F = 1365,6237$ dengan $p\text{-value} = 0,000000$ ($p < 0,05$), sehingga H_{01} diterima: fitur terpilih secara simultan berpengaruh signifikan terhadap harga jual rumah.

Uji normalitas Shapiro-Wilk menghasilkan $W = 0,9587$ dengan $p\text{-value} = 0,000000$ ($p < 0,05$), menunjukkan residual masih lemah normalitasnya meskipun telah ditransformasi log1p — kondisi umum pada data harga properti right-skewed. Oleh karena itu,

inferensi statistik menggunakan Robust SE HC3. Uji VIF pada dua fitur terpilih menunjukkan VIF = 1,678 untuk keduanya, jauh di bawah batas 10. Uji Breusch-Pagan menghasilkan LM statistic = 41,4772 dengan p-value = 0,000000, mengindikasikan heteroskedastisitas yang menguatkan keputusan penggunaan HC3.

Hasil uji t parsial menunjukkan: luas tanah ($t = 40,2933$, $p < 0,001$) berpengaruh signifikan secara parsial, sedangkan luas bangunan ($t = 1,3901$, $p = 0,1647$) tidak berpengaruh signifikan secara parsial. Pengujian fitur paling dominan menggunakan Standardized Coefficients Beta menunjukkan luas tanah memiliki $\beta^* = 0,826260$, jauh melampaui luas bangunan ($\beta^* = 0,021786$). Temuan ini mengkonfirmasi H_{13} bahwa luas tanah adalah fitur paling dominan, konsisten dengan karakteristik wilayah periurban Cidahu di mana ketersediaan lahan luas menjadi aset utama penentu nilai properti.

D. Hyperparameter Tuning

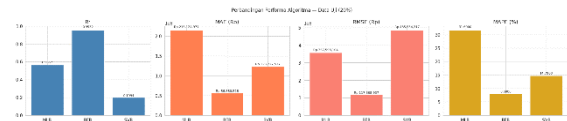
RFR: Grid Search 5-fold CV menghasilkan parameter terbaik $n_estimators = 500$, $max_depth = None$, $min_samples_leaf = 2$ dengan R^2 rata-rata CV = 0,9398. Parameter $max_depth = None$ memungkinkan pohon tumbuh optimal tanpa pembatasan kedalaman, yang efektif pada data dengan hubungan non-linier kompleks.

SVR: Grid Search 5-fold CV menghasilkan parameter terbaik $C = 1000$, $epsilon = 0,1$, $gamma = scale$ dengan R^2 rata-rata CV = 0,9107 pada data terstandarisasi. Nilai C yang tinggi (1000) menunjukkan model SVR memprioritaskan fitting data latih dengan margin yang lebih ketat.

E. Perbandingan Performa Algoritma pada Data Uji

Tabel 3. Perbandingan Performa Algoritma pada Data Uji

Algoritma	R^2	MAE (Rp)	RMSE (Rp)	MAPE (%)
MLR	0,5657	215.124.977	357.598.104	31,69%
SVR	0,1994	123.227.527	485.534.317	14,79%
RFR	0,9532	56.580.528	117.368.937	8,09%



Evaluasi pada data uji hold-out 20% menunjukkan RFR mencapai performa terbaik pada semua metrik: $R^2 = 0,9532$, MAE = Rp56.580.528, RMSE = Rp117.368.937, dan MAPE = 8,09%. RFR mampu menjelaskan 95,32% variasi harga jual dengan rata-rata kesalahan prediksi hanya 8,09% dari harga aktual. MLR berada pada posisi kedua dengan $R^2 = 0,5657$, MAE = Rp215.124.977, RMSE = Rp357.598.104, dan MAPE = 31,69%. SVR menghasilkan performa terendah dengan $R^2 = 0,1994$, MAE = Rp123.227.527, RMSE = Rp485.534.317, dan MAPE = 14,79%.

Rendahnya performa SVR ($R^2 = 0,1994$) kemungkinan disebabkan oleh kompleksitas hubungan non-linier antara luas tanah dan harga properti yang tidak sepenuhnya tertangkap oleh kernel RBF dengan hanya dua fitur, serta sensitivitas SVR terhadap outlier yang tetap ada meskipun telah dilakukan pembersihan. MLR yang terbatas pada hubungan linier menghasilkan R^2 moderat (0,5657), mengkonfirmasi bahwa hubungan antara fitur fisik properti dan harga jual bersifat non-linier.

F. Evaluasi Kestabilan Model — 5-Fold Cross-Validation

Tabel 4. Hasil 5-Fold Cross-Validation

Algoritma	R^2 Mean	R^2 Std Dev	Ranking
MLR	0,7089	0,0261	3 (Terburuk)
SVR	0,9136	0,0074	2
RFR	0,9446	0,0076	1 (Terbaik)

Evaluasi 5-Fold Cross-Validation pada seluruh dataset menghasilkan: RFR (R^2 mean = 0,9446, std = 0,007581), SVR (R^2 mean = 0,9136, std = 0,007418), dan MLR (R^2 mean = 0,7089, std = 0,026052). RFR menunjukkan R^2 mean tertinggi sekaligus standar deviasi terkecil, mengindikasikan stabilitas dan konsistensi terbaik pada berbagai pembagian data. MLR memiliki standar deviasi tertinggi (0,026052), menunjukkan sensitivitas yang lebih besar terhadap pembagian data.

Uji Wilcoxon Signed-Rank Test atas skor 5-fold CV menghasilkan p-value = 0,0625 untuk semua pasangan perbandingan (MLR vs RFR, MLR vs SVR, RFR vs SVR). Karena $p > 0,05$, secara statistik formal H_{04} tidak dapat ditolak. Ketidaksignifikanan ini mencerminkan keterbatasan statistical power uji Wilcoxon

dengan hanya 5 pasang data ($n = 5$ fold), bukan berarti ketiga algoritma identik secara performa. Berdasarkan perbandingan nominal, perbedaan praktis antara RFR dan algoritma lainnya sangat signifikan secara substansial.

G. Implementasi Sistem (Deployment)

Model RFR terbaik diintegrasikan ke dalam aplikasi web interaktif berbasis Streamlit yang dapat diakses melalui browser. Aplikasi menyediakan: (1) formulir input untuk keempat variabel fisik properti dengan validasi otomatis yang mencegah nilai non-positif dan kesalahan logis (luas bangunan > luas tanah), (2) analisis komparasi posisi pasar berupa badge status harga dan progress bar interaktif yang membandingkan estimasi pengguna terhadap distribusi statistik harga Kecamatan Cidahu, (3) halaman informasi transparansi cara kerja model dan batasan penggunaan hasil estimasi.

Tabel 5. Sampel Prediksi Harga Jual pada Data Uji

Bangunan (m ²)	Tanah (m ²)	Aktual (Rp)	Pred. MLR (Rp)	Pred. RFR (Rp)	Pred. SVR (Rp)
237,19	837,11	750.000.000	883.713.926	1.021.558.013	782.662.963
71,92	71,92	223.895.034	325.806.750	227.440.039	241.609.966
146,84	1.609,27	2.004.932.956	2.343.500.566	2.538.634.087	2.783.933.267
64,74	71,05	226.579.438	325.157.354	223.938.334	245.823.976

Berdasarkan Tabel 5, terlihat bahwa prediksi RFR konsisten lebih mendekati nilai aktual dibandingkan MLR dan SVR pada sebagian besar sampel. Pada rumah dengan luas tanah 71,92 m² dan harga aktual Rp223.895.034, RFR memprediksi Rp227.440.039 (selisih hanya 1,58%), sementara MLR memprediksi Rp325.806.750 (selisih 45,52%). Hal ini mengkonfirmasi kemampuan RFR dalam menangkap pola non-linier yang kompleks pada data properti lokal berskala kecil, terutama untuk properti dengan karakteristik yang tidak umum dalam dataset.

Secara keseluruhan, keunggulan RFR dalam prediksi harga properti di Kecamatan Cidahu dapat dijelaskan oleh tiga faktor utama: (1) kemampuan mengelola hubungan non-linier antara luas tanah dan harga yang tidak dapat ditangkap oleh MLR, (2) ketahanan terhadap outlier yang masih tersisa setelah proses

pembersihan data, dan (3) kemampuan mempertimbangkan interaksi antar fitur melalui mekanisme decision tree ensemble. Temuan ini sejalan dengan penelitian Rianti et al. (2023) yang menunjukkan Random Forest unggul dibandingkan Linear Regression pada dataset properti dengan karakteristik non-linier [15].

5. KESIMPULAN

Berdasarkan hasil analisis, pengujian hipotesis, dan evaluasi pemodelan Machine Learning, penelitian ini menghasilkan kesimpulan sebagai berikut:

1. Luas tanah merupakan fitur properti yang paling dominan memengaruhi nilai jual rumah di Kecamatan Cidahu, Kabupaten Sukabumi, dengan nilai Standardized Beta (β^*) = 0,826260 ($t = 40,29$, $p < 0,001$), jauh melampaui luas bangunan ($\beta^* = 0,021786$). Proses seleksi fitur tiga tahap (Spearman, VIF, Feature Importance) secara otomatis mengidentifikasi luas tanah dan luas bangunan sebagai dua fitur final dari empat fitur awal, dengan luas tanah memiliki Feature Importance sebesar 90,09%. Secara simultan, kedua fitur berpengaruh signifikan terhadap harga jual ($F = 1365,62$, $p < 0,001$, $R^2 = 0,706$).
2. Random Forest Regression (RFR) merupakan algoritma dengan kinerja terbaik dan paling akurat dalam memprediksi nilai jual rumah. Pada evaluasi data uji hold-out 20%, RFR mencapai $R^2 = 0,9532$, MAE = Rp56.580.528, RMSE = Rp117.368.937, dan MAPE = 8,09%, secara signifikan mengungguli MLR ($R^2 = 0,5657$, MAPE = 31,69%) dan SVR ($R^2 = 0,1994$, MAPE = 14,79%). Kestabilan RFR dikonfirmasi melalui 5-Fold Cross-Validation dengan rerata $R^2 = 0,9446$ (std = 0,008).
3. Pendekatan ensemble non-linier (RFR) terbukti jauh lebih efektif dibandingkan pendekatan linier (MLR) dan kernel (SVR) dalam memodelkan nilai properti yang bersifat kompleks dan non-linier, terutama untuk dataset properti lokal berskala kecil.
4. Model RFR berhasil diimplementasikan dalam aplikasi web berbasis Streamlit sebagai prototipe estimasi harga rumah yang objektif, transparan, dan dapat diakses publik, memberikan kontribusi praktis bagi masyarakat, penilai properti, dan pengembang di Kecamatan Cidahu.

Rekomendasi untuk penelitian selanjutnya mencakup integrasi variabel non-

fisik/lokasional (jarak ke fasilitas publik, lebar akses jalan, kondisi bangunan), perluasan cakupan wilayah analisis ke tingkat kabupaten/regional, serta eksplorasi algoritma Gradient Boosting (XGBoost, LightGBM, CatBoost) sebagai komparator tambahan.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Bapak Iwan Rizal Setiawan, M.T. selaku Pembimbing I dan Bapak Prajoko, S.Pd.I, M.Kom. selaku Pembimbing II atas bimbingan dan arahnya. Terima kasih kepada Fakultas Sains dan Teknologi Universitas Muhammadiyah Sukabumi atas fasilitas yang diberikan, serta kepada keluarga dan rekan-rekan atas dukungan moral selama penelitian berlangsung.

DAFTAR PUSTAKA

- [1] R. N. Tuah Siregar, V. Sitorus, and W. Pramudia Ananta, "Analisis Prediksi Harga Rumah di Bandung Menggunakan Regresi Linear Berganda," vol. 1, no. 6, 2023.
- [2] M. Firman, T. Sendjaja, M. Program Studi, and I. Perbanas Institute, "Impression: Jurnal Teknologi dan Informasi Pemanfaatan Big Data Penawaran dan Transaksi Properti untuk," vol. 4, no. 2, 2025.
- [3] M. A. Purwanto, H. Firmansyah, W. Asriyani, and R. P. Tulodo, "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," pp. 162–168, 2026.
- [4] J. Parhusip, A. S. Julian, F. N. Hidayat, and J. T. Souk, "Pengaplikasian Algoritma Simple Linear Regression untuk Prediksi Harga Rumah di Jabodetabek Berdasarkan Fitur Lokasi dan Luas Bangunan," vol. 18, no. 2, pp. 110–117, 2025.
- [5] U. Lathifah *et al.*, "IMPLEMENTASI METODE LINEAR REGRESSION UNTUK PREDIKSI," vol. 8, no. 1, pp. 1129–1137, 2024.
- [6] S. Baharsyah, K. Pelayanan, P. Pratama, B. Satu, and D. J. Pajak, "FAKTOR-FAKTOR PENENTU NILAI PASAR KENDARAAN RODA EMPAT BEKAS: PENDEKATAN HEDONIC PRICING MODEL PADA," vol. 3, no. 3, pp. 1353–1362, 2025.
- [7] D. P. Egistin, M. Y. Rauza, R. H. Ramadhan, S. Ramadani, and K. Kunci, "Analisis regresi linier sederhana dan penerapannya," vol. 1, no. 2, pp. 69–78, 2025.
- [8] C. D. Putranto, K. Pelayanan, P. Pratama, M. Bungo, and D. J. Pajak, "PENGARUH KARAKTERISTIK FISIK DAN AKSESIBILITAS LOKASI TERHADAP NILAI PASAR RUMAH TAPAK: STUDI HEDONIC PRICING MODEL DI KOTA DEPOK, JAWA BARAT," vol. 3, no. 3, pp. 1344–1352, 2025.
- [9] B. Karyadi, "Jurnal teknologi pendidikan," vol. 8, no. 2, pp. 253–258, 2023, doi: 10.32832/educate.v8i02.14843.
- [10] E. Retnoningsih and R. Pramudita, "Mengenal Machine Learning Dengan Teknik Supervised dan Unsupervised Learning Menggunakan Python," vol. 7, no. 2, pp. 156–165, 2020.
- [11] A. Julian indraswara, "PERANCANGAN SISTEM MACHINE LEARNING UNTUK KLASIFIKASI KENDARAAN PADA PERSIMPANGAN TRAFFIC LIGHT," 2022.
- [12] R. S. Nurhalizah, R. Ardianto, and P. Purwono, "Analisis Supervised dan Unsupervised Learning pada Machine Learning: Systematic Literature Review," *J. Ilmu Komput. dan Inform.*, vol. 4, no. 1, pp. 61–72, 2024, doi: 10.54082/jiki.168.
- [13] A. R. Pratama, F. Wabula, H. Ilmandry, M. L. Isabela, and R. Sianipar, "Literature Review The Impact of Machine Learning in Modern Industries," no. 2021, 2025.
- [14] M. Labib, S. A. Damayanti, H. N. Zaki, T. Muhayat, and R. Wirawan, "Analisis Prediksi Harga Rumah Sesuai Spesifikasi Menggunakan Multiple Linear Regression," vol. 4221, pp. 238–245, 2021.
- [15] A. Rianti, N. W. A. Majid, and A. Fauzi, "CRISP-DM: Metodologi Proyek Data Science," in *Prosiding Seminar Nasional Teknologi Informasi Dan Bisnis (SENATIB) 2023*, 2023, pp. 104–107.