Jurnal RISTER : Riset Sistem Cerdas DOI: 10.25126/Riscer
Vol. 3, No. 2, November 2026, hlm. 49-61 e-ISSN: 3047-468X

Penerapan *Apache Spark* dan *Random Forest Regression* untuk Analisis dan Prediksi Harga Rumah

Herdayanti[*1], Rika[2], Sukriadi[3]
[1,2,3]Program Studi Teknik Informatika, Universitas Lamappapoleonro
[1,2,3]Jl.Salotungo No.62 Watansoppeng, Soppeng Sulawesi Selatan
Email: [1]herdhaerda08@gmail.com, [2]rikariska042@gmail.com, [3]sukriadi@unipol.ac.id

(Naskah masuk: 14-09-2026, diterima untuk diterbitkan: 22-11-2026)

Abstrak

Prediksi harga rumah pada data iklan properti sulit dilakukan karena sebaran harga sangat timpang dan memuat nilai ekstrem, sementara penelitian terdahulu sering melaporkan Random Forest tanpa menguji pengaruh pencilan dan validasi parameter. Penelitian ini membangun alur Apache Spark MLlib untuk data rumah di Bandung dengan kontribusi metodologis: (1) karakterisasi pengaruh penyaringan outlier terhadap data dan kinerja model, (2) pemilihan parameter Random Forest dengan 3-fold cross-validation tanpa melibatkan data uji, dan (3) perbandingan dengan Linear Regression pada data uji yang sama, dengan jumlah data yang dapat ditelusuri pada setiap tahap. Dataset berisi 7.609 baris dan delapan atribut. Setelah pembersihan (617 duplikat dan satu baris dengan luas bangunan tidak valid) diperoleh 6.991 baris, dan penyaringan batas operasional menghasilkan 6.845 baris (5.466 data latih dan 1.379 data uji). Konfigurasi terbaik (100 pohon, kedalaman maksimum 10) memiliki R^2 cross-validation 0,7434. Pada data uji, Random Forest menghasilkan MAE Rp1.401.144.747,20, RMSE Rp3.470.936.218,34, dan R^2 0,7454, lebih baik daripada Linear Regression (MAE Rp1.825.491.330,61; RMSE Rp3.870.302.942,97; R^2 0,6834). Penyaringan outlier menurunkan simpangan baku harga dari Rp15,26 miliar menjadi Rp7,02 miliar; eksperimen sebelum-sesudah tambahan hanya bersifat indikatif karena pembagian datanya tidak tetap. Hasil hanya berlaku untuk rumah dalam batas penyaringan (harga maksimal Rp100 miliar), satu pembagian data, dan eksekusi Spark lokal.

Kata kunci: Prediksi harga rumah, *Random Forest Regression, Apache Spark, cross-validation, Linear Regression.*

Abstract

House price prediction on property-listing data is difficult because prices are highly skewed and contain extreme values, while earlier studies often report Random Forest without testing the effect of outliers and parameter validation. This study builds an Apache Spark MLlib workflow for housing data from Bandung with methodological contributions: (1) characterizing the effect of outlier filtering on the data and on model performance, (2) selecting Random Forest parameters with 3-fold cross-validation without using the test data, and (3) comparing against Linear Regression on the same test set, with traceable record counts at every stage. The dataset contains 7,609 rows and eight attributes. After cleaning (617 duplicates and one row with an invalid building area), 6,991 rows remained, and operational outlier filtering produced 6,845 rows (5,466 training and 1,379 testing). The best configuration (100 trees, maximum depth 10) had a cross-validation R^2 of 0.7434. On the test set, Random Forest achieved an MAE of IDR 1,401,144,747.20, an RMSE of IDR 3,470,936,218.34, and an R^2 of 0.7454, outperforming Linear Regression (MAE IDR 1,825,491,330.61; RMSE IDR 3,870,302,942.97; R^2 0.6834). Outlier filtering reduced the standard deviation of price from IDR 15.26 billion to IDR 7.02 billion; an additional before-after experiment is indicative only because its data split was not fixed. The results apply only to houses within the filtering limits (price up to IDR 100 billion), a single data split, and local Spark execution.

Keywords: *house price prediction, Random Forest Regression, Apache Spark, cross-validation, Linear Regression.*

1. PENDAHULUAN

Rumah merupakan salah satu kebutuhan pokok sekaligus aset yang memiliki nilai ekonomi tinggi. Dalam kegiatan jual beli properti, penentuan harga rumah dipengaruhi oleh berbagai karakteristik, seperti lokasi, luas tanah, luas bangunan, jumlah kamar tidur, jumlah kamar mandi, serta fasilitas pendukung. Perbedaan karakteristik tersebut menyebabkan harga rumah memiliki rentang yang beragam,

bahkan pada wilayah geografis yang sama. Kondisi ini menjadikan analisis data penting untuk memahami hubungan antara karakteristik properti dan harga rumah. Selain itu, estimasi harga berdasarkan data dapat menjadi informasi pendukung bagi masyarakat dalam memahami nilai suatu properti sebelum melakukan transaksi jual beli.

Perkembangan teknologi *machine learning* memberikan alternatif untuk mempelajari hubungan antara karakteristik rumah dan harga melalui data historis. Sejumlah penelitian telah menerapkan algoritma *machine learning* untuk memprediksi harga rumah. Qolbi et al., (2025) membahas prediksi harga rumah di Jakarta Pusat menggunakan algoritma *machine learning*, sedangkan Marhananda & Fabiandika, (2026) menerapkan *Random Forest Regressor* untuk memprediksi harga rumah di Kota Malang. Warjiyono et al., (2024) juga membahas penggunaan algoritma *Random Forest* untuk memprediksi harga jual rumah. Selain itu, Balqis & Aini, (2026) mengkaji penerapan *Random Forest Regression* dalam memprediksi harga rumah. Penelitian-penelitian tersebut menunjukkan bahwa karakteristik properti dapat dimanfaatkan sebagai atribut dalam membangun model prediksi harga rumah.

Penelitian lainnya dilakukan oleh Fausi et al., (2025) yang membandingkan kinerja *Linear Regression*, *Random Forest*, dan *XGBoost*, dan melaporkan bahwa *Linear Regression* justru menghasilkan RMSE terendah pada datanya. Akhmadi & Budiman, (2025) membandingkan Regresi Linear, *Random Forest*, dan *Gradient Boosting*, Salsabila et al., (2025) membandingkan *Decision Tree Regressor* dengan *Random Forest Regressor*, dan Sobirin et al., (2025) menggunakan *Multiple Linear Regression* dan *Random Forest Regressor* dengan teknik *web scraping*. Berbagai penelitian tersebut menunjukkan bahwa *Random Forest* banyak digunakan dalam estimasi harga rumah, tetapi kinerjanya bergantung pada data dan praproses.

Penelitian sebelumnya oleh sebagian penulis juga menerapkan pembelajaran mesin pada objek lain. Sukriadi et al., (2026) menerapkan *Random Forest* untuk memprediksi kepatuhan pendaftaran tanah, sedangkan Sukriadi et al., (2023) menerapkan *text mining* dan *Naïve Bayes* dengan RapidMiner untuk mengklasifikasikan judul skripsi berdasarkan konsentrasi, dengan 49 dari 75 judul (65%) sesuai. Kedua penelitian tersebut tidak membahas harga rumah, tetapi menunjukkan alur pengolahan data hingga pemodelan yang menjadi dasar penelitian ini, yang kemudian dikembangkan pada regresi harga rumah dengan Apache Spark.

Dari rujukan tersebut terdapat tiga celah. Pertama, sebagian studi memakai data berukuran kecil, misalnya 1.460 properti pada Balqis & Aini, (2026), dan belum menguraikan pengaruh nilai ekstrem yang lazim pada data iklan properti; pada data penelitian ini luas bangunan mencapai 3.006.000 m² dan harga mencapai ratusan miliar rupiah. Kedua, hasil perbandingan algoritma bergantung pada data: *Linear Regression* terbaik pada Fausi et al., (2025), sedangkan *XGBoost* terbaik pada data Ames, Iowa (Sharma et al., 2024). Keunggulan *Random Forest* karena itu perlu diuji pada setiap dataset dengan validasi parameter. Ketiga, evaluasi regresi pada Spark yang ditinjau memakai data non-properti, yaitu durasi perjalanan taksi dan Netflix Prize (Dritsas & Trigka, 2025), sehingga manfaat Spark untuk regresi harga rumah pada data berukuran ribuan baris belum dilaporkan. Uraian celah penelitian tersebut dibatasi pada informasi yang telah dijelaskan dalam rujukan dan tidak dimaksudkan untuk menyatakan bahwa seluruh penelitian terdahulu memiliki keterbatasan yang sama. Kebaruan penelitian ini bersifat metodologis dan terukur: (1) karakterisasi pengaruh penyaringan outlier terhadap distribusi data dan, secara indikatif, terhadap MAE, RMSE, dan R^2; (2) pemilihan parameter *Random Forest* dengan 3-fold *cross-validation* tanpa melibatkan data uji, dibandingkan dengan *Linear Regression*; dan (3) satu alur PySpark dengan jumlah data yang dapat ditelusuri pada setiap tahap (7.609, 6.991, dan

6.845 baris). Penelitian ini bukan pengembangan algoritma baru.

Apache Spark dipilih dengan tiga pertimbangan. Pertama, API *Pipeline* pada MLlib merangkai *StringIndexer*, *OneHotEncoder*, *VectorAssembler*, dan model dalam satu objek sehingga transformasi pada data latih dan data uji identik dan eksperimen dapat diulang. Kedua, Spark mendukung skalabilitas pada lingkungan terdistribusi; Dritsas & Trigka, (2025) melaporkan bahwa waktu pemrosesan model regresi pada Spark meningkat secara linear terhadap ukuran data, yang relevan bagi data iklan properti yang terus bertambah. Ketiga, atribut lokasi yang kategorikal diproses dengan *StringIndexer* dan *OneHotEncoder*, lalu digabung dengan atribut numerik oleh *VectorAssembler* dalam satu alur. Penelitian ini dijalankan pada konfigurasi lokal (*local mode*) dengan data ribuan baris, sehingga keunggulan kecepatan Spark tidak diasumsikan dan tidak diklaim. Waktu komputasi pada konfigurasi ini dilaporkan pada Subbab Evaluasi Waktu Komputasi, tetapi tanpa pembanding sehingga tidak dipakai untuk mengklaim keunggulan kecepatan.

Urgensi penelitian ini bersifat spesifik: harga pada data iklan properti sangat condong ke kanan dan memuat nilai ekstrem, sehingga model yang dilaporkan tanpa penanganan pencilan dan tanpa validasi parameter dapat menyesatkan pengguna estimasi harga. Penelitian ini mengkaji enam prediktor, yaitu lokasi, jumlah kamar tidur, jumlah kamar mandi, jumlah *carport*, luas tanah, dan luas bangunan. Tujuannya adalah (1) mengukur pengaruh penyaringan outlier terhadap kinerja model, (2) memilih parameter *Random Forest* melalui 3-fold *cross-validation*, dan (3) membandingkannya dengan *Linear Regression* pada data uji yang sama.

Penelitian dibatasi pada dataset data_rumah.csv dan eksekusi Apache Spark dalam konfigurasi local[*]. Penyaringan nilai ekstrem diterapkan sebelum pembagian data sehingga data uji juga hanya memuat rumah dalam batas penyaringan; hasil tidak berlaku untuk rumah di luar batas tersebut. Pengujian menggunakan satu pembagian data 80:20 dengan seed 42; tuning dilakukan melalui 3-fold *cross-validation* pada data pelatihan dan evaluasi akhir pada data pengujian. Perbandingan hanya mencakup *Random Forest* dan *Linear Regression*, dan penelitian tidak mengklaim pengujian pada klaster terdistribusi atau keunggulan kecepatan Spark.

2. METODE PENELITIAN

2.1 Jenis dan Sumber Data

Penelitian ini merupakan penelitian kuantitatif dengan pendekatan eksperimen pemodelan regresi. Data yang digunakan merupakan data sekunder yang tersimpan dalam berkas data_rumah.csv yang diambil dari platform Kaggle pada bulan Oktober yang berisi informasi iklan properti rumah di wilayah Bandung. Dataset terdiri atas 7.609 baris data dan delapan atribut, yaitu house_name, location, bedroom_count, bathroom_count, carport_count, price, land_area, dan building_area (m2). Atribut house_name berisi nama atau deskripsi properti, location menunjukkan lokasi rumah, sedangkan atribut lainnya memuat jumlah kamar tidur, jumlah kamar mandi, jumlah carport, harga rumah, luas tanah, dan luas bangunan.

Atribut house_name berisi nama atau deskripsi properti, location menunjukkan lokasi rumah, sedangkan atribut lainnya berisi informasi jumlah kamar tidur, kamar mandi, carport, harga rumah, luas tanah, dan luas bangunan. Variabel target dalam penelitian ini adalah price, sedangkan variabel prediktor yang digunakan meliputi location, bedroom_count, bathroom_count, carport_count, land_area, dan building_area (m2). Atribut house_name tidak digunakan dalam proses pemodelan.

Pada pembacaan Pandas tercatat 7.609 baris dan pada Spark 7.610 baris, dengan 8 kolom. Satu baris tambahan pada Spark tersingkirkan pada tahap pembersihan sehingga kedua jalur menghasilkan 6.991 baris. Data tersebut selanjutnya digunakan untuk

pembersihan, pengolahan, dan pemodelan menggunakan algoritma Random Forest Regression.

2.2 Tahapan Penelitian

Tahapan penelitian dilakukan secara bertahap, dimulai dari pembacaan dataset, pemeriksaan kualitas data, pembersihan data, analisis eksploratif, penanganan nilai ekstrem, pembentukan fitur, pembagian data, pemodelan menggunakan algoritma Random Forest Regression, pengujian kombinasi parameter, hingga evaluasi hasil prediksi.

Proses pengolahan data dilakukan menggunakan bahasa pemrograman Python dengan bantuan pustaka Pandas dan Apache Spark. Pandas digunakan untuk analisis dan visualisasi data, sedangkan Apache Spark digunakan dalam proses pengolahan data dan pemodelan. Algoritma Random Forest Regression diterapkan menggunakan pustaka PySpark MLlib.

2.3 Pembersihan Data

Pemeriksaan awal menunjukkan bahwa dataset memiliki 7.609 baris (7.610 pada pembacaan Spark) dan delapan atribut. Pemeriksaan menemukan 617 baris duplikat dan tidak ditemukan nilai kosong pada atribut dataset. Setelah baris duplikat dihapus, jumlah data menjadi 6.992 baris pada Pandas dan 6.993 baris pada Spark (selisih satu baris yang sama dengan selisih pembacaan awal). Selain itu, ditemukan satu baris dengan luas bangunan sebesar -250 m², yang tidak valid secara logis sehingga baris tersebut dikeluarkan.

Pembersihan data dilakukan menggunakan fungsi dropDuplicates() untuk menghapus baris duplikat, dan dropna() dijalankan sebagai pemeriksaan nilai kosong pada atribut yang diperlukan. Selanjutnya, atribut numerik dikonversi ke tipe data yang sesuai. Data kemudian disaring dengan ketentuan harga rumah, luas tanah, dan luas bangunan harus lebih dari nol. Jumlah kamar tidur, kamar mandi, dan carport harus bernilai minimal nol, serta atribut lokasi tidak boleh kosong.

Setelah proses pembersihan tersebut, jumlah data menjadi 6.991 baris. Selanjutnya, dilakukan penanganan nilai ekstrem untuk mengurangi pengaruh data dengan nilai yang terlalu tinggi terhadap pemodelan. Batas penyaringan yang digunakan adalah harga rumah maksimal Rp100.000.000.000, luas tanah maksimal 5.000 m², luas bangunan maksimal 3.000 m², jumlah kamar tidur maksimal 20, jumlah kamar mandi maksimal 20, dan jumlah carport maksimal 10. Karena penyaringan dilakukan sebelum pembagian data, data uji juga hanya memuat rumah yang berada dalam batas tersebut.

Berdasarkan batas tersebut, jumlah data yang digunakan untuk pemodelan menjadi 6.845 baris. Sebanyak 146 baris dikeluarkan karena berada di luar batas penyaringan yang telah ditentukan. Batas tersebut merupakan kriteria penyaringan dalam eksperimen ini dan bukan batas umum untuk seluruh properti rumah.

Tahapan preprocessing merupakan bagian penting dalam mempersiapkan data sebelum digunakan untuk pemodelan. Pangestu & Iswahyudi, (2025) membahas pengaruh preprocessing data terhadap performa model regresi. Oleh karena itu, pemeriksaan data, penanganan nilai kosong, penghapusan duplikasi, dan penyaringan nilai ekstrem dilakukan untuk meningkatkan kualitas data yang digunakan dalam penelitian.

2.4 Analisis Eksploratif

Analisis eksploratif dilakukan untuk mengetahui karakteristik dan pola dalam dataset. Proses ini meliputi pemeriksaan statistik deskriptif, distribusi harga rumah, luas tanah, luas bangunan, jumlah rumah berdasarkan lokasi, serta hubungan antaratribut numerik.

Jumlah data pada setiap tahap dihitung dalam satu rangkaian pemeriksaan sehingga dapat ditelusuri: 7.609 baris awal (7.610 pada Spark), 6.992 baris setelah penghapusan 617 duplikat (6.993 pada Spark), 6.991 baris setelah satu baris tidak valid dikeluarkan, dan 6.845 baris setelah penyaringan nilai ekstrem. Pembacaan menggunakan Pandas dan Spark menghasilkan jumlah data bersih yang sama, yaitu 6.991 baris.

Analisis korelasi Pearson dilakukan pada 6.991 baris data bersih (sebelum

penyaringan nilai ekstrem) untuk mengetahui hubungan linear antara atribut numerik dan harga rumah. Korelasi luas tanah (land_area) dengan harga (price) sebesar 0,3499, yang menunjukkan hubungan positif dengan kekuatan lemah hingga sedang.

Selain itu, dilakukan standardisasi menggunakan StandardScaler terhadap lima atribut numerik prediktor, yaitu jumlah kamar tidur, jumlah kamar mandi, jumlah carport, luas tanah, dan luas bangunan. Analisis Principal Component Analysis (PCA) juga dilakukan terhadap lima atribut numerik tersebut. Hasil PCA menghasilkan dua komponen utama dengan total variasi yang dapat dijelaskan sebesar 63,50%. PCA digunakan untuk membantu memahami struktur data dan visualisasi, bukan sebagai bagian dari pipeline prediksi Random Forest.

2.5 Pembentukan Fitur dan Model

Pemodelan dilakukan menggunakan algoritma *Random Forest Regression* dari PySpark MLlib. Sebelum proses pelatihan, fitur kategorikal dan numerik disiapkan agar dapat digunakan oleh algoritma.

Atribut location merupakan fitur kategorikal yang diubah menjadi indeks numerik menggunakan StringIndexer dengan pengaturan handleInvalid='keep'. Selanjutnya, OneHotEncoder digunakan untuk mengubah indeks lokasi menjadi representasi vektor kategorikal.

Lima atribut numerik, yaitu bedroom_count, bathroom_count, carport_count, land_area, dan building_area (m2), digabungkan dengan vektor lokasi menggunakan VectorAssembler. Seluruh fitur tersebut disimpan dalam kolom features, sedangkan kolom price digunakan sebagai target prediksi.

Proses transformasi fitur dan pemodelan disusun menggunakan Pipeline agar setiap tahapan dapat dijalankan secara berurutan. Model Random Forest Regression kemudian dilatih menggunakan data pelatihan.

Pengujian hyperparameter dilakukan menggunakan 3-fold *cross-validation* pada data pelatihan. Parameter numTrees diuji pada nilai 50 dan 100, sedangkan maxDepth diuji pada nilai 5 dan 10, sehingga terdapat empat kombinasi. Pemilihan konfigurasi terbaik didasarkan pada rata-rata R^2 hasil *cross-validation*. Data pengujian tidak digunakan dalam proses pemilihan parameter. Sebelum cross-validation, data pelatihan dikonsolidasikan menjadi 8 partisi (coalesce) dan seed Random Forest ditetapkan 42.

Hasil cross-validation menunjukkan konfigurasi numTrees=100 dan maxDepth=10 sebagai konfigurasi terbaik dengan R^2 rata-rata 0,743419. Kombinasi numTrees=50 dan maxDepth=10 memperoleh 0,737499; numTrees=100 dan maxDepth=5 memperoleh 0,698503; sedangkan numTrees=50 dan maxDepth=5 memperoleh 0,697875.

2.6 Pembagian Data dan Evaluasi

Dataset setelah penyaringan nilai ekstrem sebanyak 6.845 baris dibagi menggunakan randomSplit([0.8, 0.2], seed=42), yaitu 5.466 baris data pelatihan dan 1.379 baris data pengujian. Data pelatihan digunakan untuk cross-validation, tuning, dan pelatihan model, sedangkan data pengujian disimpan untuk evaluasi akhir.

Data pelatihan digunakan untuk membangun model Random Forest Regression, sedangkan data pengujian digunakan untuk mengukur kemampuan model dalam memprediksi harga rumah yang belum digunakan pada proses pelatihan.

Evaluasi model dilakukan menggunakan tiga metrik, yaitu Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), dan R-squared (R^2). MAE digunakan untuk mengukur rata-rata selisih absolut antara harga aktual dan harga prediksi. RMSE digunakan untuk mengukur akar rata-rata kuadrat kesalahan sehingga lebih sensitif terhadap kesalahan prediksi yang besar. Sementara itu, R^2 digunakan untuk mengetahui seberapa besar variasi harga rumah yang dapat dijelaskan oleh model.

Model pembanding Linear Regression dilatih pada data pelatihan yang sama dengan pipeline fitur yang sama (maxIter=50, parameter lain bawaan PySpark MLlib) dan dievaluasi

pada data pengujian yang sama. Model Random Forest terbaik dengan numTrees=100 dan maxDepth=10 memperoleh MAE sebesar Rp1.401.144.747,20, RMSE sebesar Rp3.470.936.218,34, dan R^2 sebesar 0,7454 pada data pengujian. Nilai R^2 menunjukkan bahwa model menjelaskan sekitar 74,54% variasi harga rumah pada data pengujian.

3. HASIL DAN PEMBAHASAN

3.1 Hasil Pemeriksaan dan Pembersihan Data

Pemeriksaan awal terhadap dataset menunjukkan 7.609 baris data (7.610 pada pembacaan Spark) dengan delapan atribut. Pemeriksaan kualitas data menemukan 617 baris duplikat, sedangkan tidak ditemukan nilai kosong. Setelah baris duplikat dihapus, jumlah data menjadi 6.992 baris pada Pandas dan 6.993 baris pada Spark (selisih satu baris yang sama dengan selisih pembacaan awal). Selanjutnya, ditemukan satu baris dengan nilai building_area (m2) sebesar -250 m². Nilai tersebut tidak valid karena luas bangunan tidak mungkin negatif, sehingga baris terkait dikeluarkan.

Pembersihan dilakukan menggunakan fungsi dropDuplicates() untuk menghapus data duplikat dan dropna() untuk menangani baris yang memiliki nilai kosong pada atribut yang diperlukan. Selanjutnya, data disaring berdasarkan ketentuan harga, luas tanah, dan luas bangunan harus lebih dari nol, sedangkan jumlah kamar tidur, kamar mandi, dan carport harus bernilai minimal nol. Atribut lokasi juga harus memiliki nilai.

Proses pembersihan menghasilkan 6.991 baris data bersih. Jumlah ini sama pada pembacaan Pandas dan Spark dan menjadi dasar seluruh analisis eksploratif pada subbab ini, sedangkan penyaringan nilai ekstrem menghasilkan 6.845 baris untuk pemodelan.

Tabel 1. Hasil Pemeriksaan dan Pembersihan Data

Indikator pemeriksaan	Hasil
Jumlah baris awal (Pandas / Spark)	7.609 / 7.610
Jumlah atribut	8
Baris duplikat	617
Nilai kosong	0
Data setelah penghapusan duplikat (Pandas / Spark)	6.992 / 6.993
Data dikeluarkan pada validasi (Pandas / Spark)	1 / 2
Data setelah pembersihan	6.991
Data setelah penyaringan nilai ekstrem	6.845
Baris dikeluarkan pada penyaringan	146 (2,09%)

Setelah pembersihan, rata-rata harga rumah tercatat sekitar Rp5.003.054.284, rata-rata jumlah kamar tidur 4,51, kamar mandi 3,20, carport 1,40, luas tanah 275,27 m², dan luas bangunan 677,91 m². Nilai maksimum luas bangunan yang sangat tinggi dibandingkan rata-rata menunjukkan perlunya pemeriksaan nilai ekstrem sebelum pemodelan.

Penyaringan nilai ekstrem menggunakan batas operasional berupa harga maksimal Rp100 miliar, luas tanah maksimal 5.000 m², luas bangunan maksimal 3.000 m², jumlah kamar tidur dan kamar mandi maksimal 20, serta jumlah carport maksimal 10. Batas tersebut merupakan kriteria eksperimen berdasarkan pemeriksaan distribusi dan karakteristik atribut, bukan batas universal.

Setelah penyaringan, data berkurang dari 6.991 menjadi 6.845 baris. Sebanyak 146 baris atau 2,09% dikeluarkan karena berada di luar batas operasional. Pengeluaran tersebut tidak berarti seluruh baris pasti salah, tetapi merupakan keputusan penyaringan untuk membatasi nilai ekstrem dalam eksperimen.

3.2 Persebaran Lokasi Rumah

Berdasarkan hasil pengelompokan data Spark sebelum penyaringan tambahan nilai ekstrem, lokasi dengan jumlah rumah terbanyak pada hasil yang ditampilkan adalah Arcamanik, Bandung sebanyak 360 rumah, Mandalajati, Bandung sebanyak 350 rumah, dan Cicendo, Bandung sebanyak 339 rumah. Selanjutnya terdapat Andir sebanyak 333 rumah dan Lengkong sebanyak 324 rumah.

Tabel 2. Lima Lokasi dengan Jumlah Rumah Terbanyak

Perbedaan rata-rata harga antarwilayah menunjukkan bahwa lokasi merupakan salah satu atribut yang berpotensi memberikan

informasi dalam prediksi harga rumah. Sebagai contoh, Cicendo memiliki jumlah observasi lebih sedikit dibandingkan Arcamanik, tetapi rata-rata harga rumah yang tercatat lebih tinggi. Kondisi tersebut menunjukkan bahwa jumlah rumah dalam suatu wilayah tidak selalu sejalan dengan rata-rata harganya.

Oleh karena itu, atribut location dipertahankan sebagai salah satu fitur prediktor dalam model. Meskipun demikian, perbedaan rata-rata harga tersebut tidak secara langsung membuktikan bahwa lokasi menjadi penyebab perubahan harga, karena harga rumah juga dapat dipengaruhi oleh luas tanah, luas bangunan, jumlah kamar, dan karakteristik lainnya.

3.3 Hasil Analisis Harga dan Korelasi

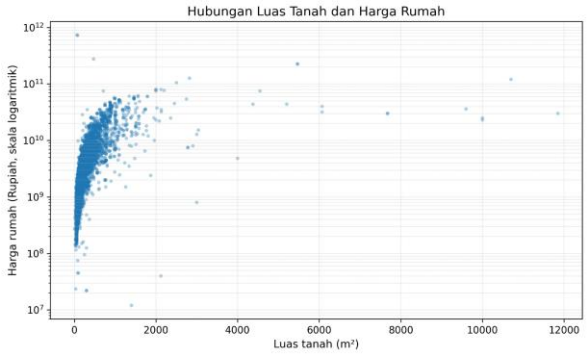
Analisis statistik deskriptif terhadap 6.991 baris data bersih menunjukkan bahwa harga rumah memiliki rata-rata Rp5.003.054.284 dan median Rp2.200.000.000. Median yang lebih rendah daripada rata-rata menunjukkan distribusi harga yang menceng ke kanan, yaitu terdapat sejumlah rumah dengan harga jauh lebih tinggi daripada sebagian besar data. Harga terendah tercatat Rp12.000.000 dan harga tertinggi mencapai Rp735.000.000.000.

Korelasi Pearson antara luas tanah (land_area) dan harga rumah (price) sebesar 0,3499, yang menunjukkan hubungan positif dengan tingkat korelasi lemah hingga sedang. Korelasi luas bangunan dengan harga sebesar 0,0010, sehingga hubungan linearnya hampir tidak terlihat pada dataset ini. Korelasi jumlah kamar tidur dengan harga sebesar 0,1311, jumlah kamar mandi sebesar 0,1267, dan jumlah carport sebesar 0,1831. Hubungan setiap atribut secara individual dengan harga relatif terbatas, sehingga penggunaan beberapa atribut secara bersama-sama diperlukan untuk mempelajari pola harga. Angka korelasi ini dihitung dari 6.991 baris sebelum penyaringan nilai ekstrem.

Dengan demikian, pemodelan menggunakan beberapa atribut secara bersama-sama diperlukan untuk memperoleh prediksi yang lebih baik. Atribut lokasi, jumlah kamar tidur, jumlah kamar mandi, jumlah carport, luas tanah, dan luas bangunan digunakan sebagai prediktor dalam penelitian ini.

Tabel 2. Hasil Analisis Harga dan Korelasi

Lokasi	Jumlah rumah	Rata-rata harga (sekitar)
Arcamanik, Bandung	360	Rp1,495 miliar
Mandalajati, Bandung	350	Rp1,762 miliar
Cicendo, Bandung	339	Rp8,830 miliar
Andir, Bandung	333	Rp3,594 miliar
Lengkong, Bandung	324	Rp5,599 miliar



Gambar 1. Hubungan luas tanah dan harga rumah (sumbu harga berskala logaritmik)

3.4 Hasil Principal Component Analysis

Principal Component Analysis (PCA) digunakan untuk mengeksplorasi struktur hubungan antar lima atribut numerik prediktor, yaitu jumlah kamar tidur, jumlah kamar mandi, jumlah carport, luas tanah, dan luas bangunan. PCA tidak digunakan sebagai fitur pada pipeline prediksi Random Forest.

Hasil PCA menunjukkan komponen pertama menjelaskan 42,85% variasi dan komponen kedua 20,65%, sehingga keduanya bersama-sama menjelaskan 63,50% variasi dari lima atribut numerik.

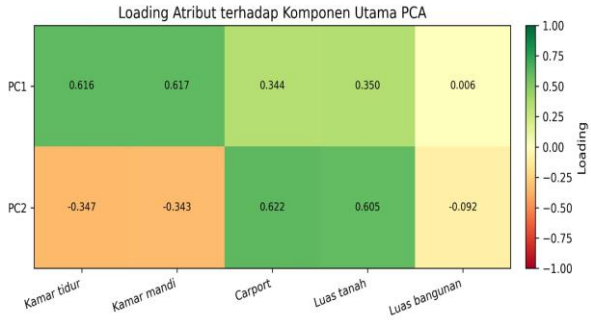
Hasil tersebut menunjukkan bahwa dua komponen utama dapat mempertahankan sebagian besar informasi variasi dari lima atribut numerik yang dianalisis. Meskipun demikian, masih terdapat sekitar 36,50% variasi yang tidak terwakili oleh dua komponen tersebut. Gambar 2

menampilkan proyeksi data pada dua komponen utama tersebut, sedangkan Gambar 3 menunjukkan bahwa PC1 didominasi jumlah kamar tidur (0,616) dan kamar mandi (0,617), diikuti luas tanah (0,350) dan carport (0,344), sedangkan luas bangunan hampir tidak berkontribusi (0,006). PC2 didominasi carport (0,622) dan luas tanah (0,605) dengan arah berlawanan terhadap kamar tidur (−0,347) dan kamar mandi (−0,343).

Tabel 3. Principal Component Analysis

Komponen	Konfigurasi
Jumlah data pemodelan	6.845
Data pelatihan	5.466
Data pengujian	1.379
Rasio pembagian	80:20
Seed	42
Cross-validation	3-fold
Jumlah kombinasi parameter	4

Penggunaan PCA dalam penelitian prediksi harga rumah juga dikaji oleh Sutanto et al., (2025) melalui kombinasi PCA dan Random Forest Regression. Penelitian tersebut menjadi referensi pendukung dalam penggunaan PCA untuk menganalisis struktur data sebelum pemodelan. Namun, pada penelitian ini PCA hanya digunakan untuk eksplorasi dan visualisasi, bukan sebagai bagian dari pipeline prediksi Random Forest.



Gambar 3. Loading atribut terhadap komponen utama PCA

3.5 Hasil Penanganan Nilai Ekstrem Pembagian Data

Pemeriksaan sebelum penyaringan menunjukkan nilai maksimum luas bangunan mencapai 3.006.000 m², jauh di atas rata-rata 677,91 m². Ketimpangan tersebut menjadi alasan untuk melakukan penyaringan operasional. Namun, penelitian tidak melakukan verifikasi satu per satu terhadap sumber observasi atau analisis sensitivitas dengan beberapa batas berbeda.

Nilai ekstrem dapat memengaruhi proses pembelajaran model karena algoritma regresi berusaha mempelajari pola dari seluruh observasi yang digunakan. Apabila terdapat sejumlah observasi dengan nilai yang sangat jauh dari sebagian besar data, model berpotensi menghasilkan prediksi yang kurang sesuai untuk karakteristik rumah pada umumnya.

Penyaringan operasional mengurangi data dari 6.991 menjadi 6.845 baris, sehingga 146 baris atau 2,09% dikeluarkan. Setelah itu, data dibagi menjadi 5.466 baris pelatihan dan 1.379 baris pengujian.

Pembagian data menggunakan randomSplit ([0.8, 0.2], seed=42). Seluruh proses cross-validation dan pemilihan parameter dilakukan pada data pelatihan, sedangkan data pengujian hanya digunakan untuk evaluasi akhir agar tidak terlibat dalam pemilihan model.

3.6 Hasil Cross-Validation dan Hyperparameter Tuning Random Forest Regression

Empat kombinasi parameter diuji menggunakan 3-fold cross-validation. Nilai R^2 rata-rata tertinggi diperoleh pada numTrees=100 dan maxDepth=10 sebesar 0,743419. Hasil ini digunakan untuk memilih model Random Forest yang kemudian dievaluasi pada data pengujian. Konfigurasi terbaik berada pada nilai maksimum yang diuji (numTrees=100 dan maxDepth=10), sehingga optimum yang sebenarnya dapat berada di luar ruang pencarian ini. Pengujian belum mencakup kombinasi parameter di luar ruang pencarian tersebut dan simpangan baku antar-lipatan belum dihitung. Dengan demikian, konfigurasi ini dinyatakan sebagai konfigurasi terbaik di antara empat kombinasi yang diuji, bukan sebagai optimum global.

Tabel 4. Hasil 3-Fold Cross-Validation

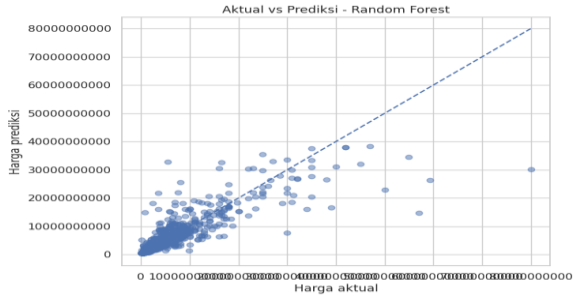
numTrees	maxDepth	R^2 cross-validation
100	10	0,743419
50	10	0,737499
100	5	0,698503
50	5	0,697875

Tabel 5. Perbandingan Kinerja Model pada Data Pengujian

Metrik	Random Forest (100,10)	Linear Regression
MAE	Rp1.401.144.747,20	Rp1.825.491.330,61
RMSE	Rp3.470.936.218,34	Rp3.870.302.942,97
R^2	0,7454	0,6834

Hasil perbandingan menunjukkan Random Forest memiliki MAE sekitar 23,25% lebih rendah dan RMSE sekitar 10,32% lebih rendah dibandingkan Linear Regression. Nilai R^2 Random Forest juga lebih tinggi sebesar 0,06197. Hasil ini menunjukkan Random Forest lebih baik pada tiga metrik evaluasi dalam eksperimen dan pembagian data yang sama. Meskipun demikian, hasil ini belum membuktikan bahwa Random Forest selalu lebih unggul pada dataset atau pembagian data lain.

Nilai R^2 sebesar 0,7454 menunjukkan bahwa model menjelaskan sekitar 74,54% variasi harga rumah pada data pengujian. Nilai tersebut menunjukkan kemampuan penjelasan model pada pembagian data ini, tetapi tidak berarti seluruh prediksi memiliki tingkat ketepatan yang sama untuk setiap rumah.



Gambar 4. Perbandingan harga aktual dan prediksi pada data pengujian (Random Forest terbaik).

Pada grafik aktual dan prediksi, garis diagonal menjadi acuan kondisi ketika harga prediksi sama dengan harga aktual. Titik yang dekat dengan garis menunjukkan prediksi yang lebih mendekati nilai aktual, sedangkan titik yang jauh menunjukkan selisih prediksi yang lebih besar. Pola ini perlu dibaca bersama MAE, RMSE, dan R^2, bukan sebagai satu-satunya dasar penilaian model.

3.7 Perbandingan dengan Penelitian Terdahulu

Hasil ini dibandingkan secara indikatif dengan penelitian lain karena dataset, wilayah, rentang harga, dan pembagian data berbeda. Noorihsan et al., (2026) melaporkan R^2 76,66% dan MAPE 25,27% untuk prediksi harga properti di Kota Malang dengan *Random Forest*; R^2 penelitian ini (0,7454) berada pada kisaran yang serupa. Sebaliknya, Fausi et al., (2025) melaporkan *Linear Regression* lebih baik daripada *Random Forest* pada datanya, sedangkan pada penelitian ini *Random Forest* lebih baik pada ketiga metrik. Perbedaan ini menunjukkan bahwa peringkat algoritma bergantung pada data dan penanganan pencilan. Perlu dicatat bahwa R^2 penelitian ini dihitung pada data uji yang telah disaring, sehingga tidak sebanding langsung dengan studi yang mengevaluasi pada data uji tanpa penyaringan. Perbandingan pada bagian ini dibatasi pada penelitian yang metrik dan informasinya telah tersedia dalam naskah. Perbandingan lintas penelitian bersifat indikatif karena perbedaan dataset, wilayah, proses penyaringan, dan skema pengujian.

3.8 Analisis kesalahan prediksi

Analisis residual pada model Random Forest menunjukkan rata-rata kesalahan absolut (mean absolute error) sebesar Rp1.401.144.747,20. Kesalahan absolut terbesar mencapai Rp49.815.734.018,50, dengan residual minimum sebesar -Rp27.266.672.823,56 dan residual maksimum sebesar Rp49.815.734.018,50. Residual positif menunjukkan bahwa harga aktual lebih tinggi daripada hasil prediksi (model memprediksi terlalu rendah), sedangkan residual negatif menunjukkan bahwa hasil prediksi lebih tinggi daripada harga aktual. Sebagai contoh, rumah

dengan harga aktual Rp80.000.000.000 diprediksi sebesar sekitar Rp30.184.265.981,50, sehingga menghasilkan kesalahan sekitar Rp49.815.734.018,50. Sebaliknya, rumah dengan harga aktual Rp5.500.000.000 diprediksi sekitar Rp32.766.672.823,56. Temuan ini memperlihatkan bahwa model masih mengalami kesulitan pada sebagian rumah dengan nilai harga yang sangat tinggi atau karakteristik yang kurang terwakili dalam data pelatihan. Dengan demikian, nilai MAE dan R^2 yang baik secara agregat tidak berarti seluruh prediksi memiliki tingkat kesalahan yang seragam.

Rincian kesalahan berdasarkan kuartil harga, lokasi, serta kontribusi 1% dan 5% observasi dengan kesalahan terbesar belum dihitung karena tabel prediksi per observasi tidak disertakan dalam bahan analisis. Oleh sebab itu, pembahasan residual pada penelitian ini dibatasi pada nilai kesalahan agregat dan contoh prediksi yang telah tersedia.

3.9 Pengaruh penyaringan pencilan (outlier)

Pengaruh penyaringan terhadap data terlihat pada statistik deskriptif. Setelah 146 baris (2,09%) dikeluarkan, rata-rata harga turun dari Rp5,003 miliar menjadi Rp4,467 miliar dan simpangan baku harga dari Rp15,26 miliar menjadi Rp7,02 miliar, sedangkan harga maksimum turun dari Rp735 miliar menjadi Rp80 miliar. Luas bangunan maksimum turun dari 3.006.000 m^2 menjadi 2.500 m^2 dan simpangan bakunya dari 35.949,75 menjadi 206,19. Jumlah kamar tidur dan kamar mandi maksimum turun dari 99 menjadi 20, jumlah carport maksimum dari 40 menjadi 10, dan luas tanah maksimum dari 11.850 m^2 menjadi 4.378 m^2.

Selain itu, dilakukan eksperimen tambahan yang membandingkan Random Forest (100 pohon, kedalaman 10) yang dilatih sebelum dan sesudah penyaringan. Data bersih 6.991 baris dibagi 80:20 (seed 42) sebelum penyaringan menjadi 5.656 data latih dan 1.335 data uji; penyaringan pada data latih menyisakan 5.537 baris, dan kedua model dievaluasi pada data uji yang disaring dengan batas yang sama. Pada Tabel 6, MAE turun dari Rp1.321.340.177,58 menjadi Rp1.179.593.767,54 (sekitar 10,73%), RMSE dari Rp3.171.086.850,67 menjadi Rp2.658.098.325,58 (sekitar 16,18%), dan R^2 naik dari 0,8249 menjadi 0,8770. Hasil ini hanya bersifat indikatif dan tidak dijadikan bukti utama, karena jumlah data uji tersaring pada eksekusi tersebut tercatat 1.386 baris, lebih besar daripada data uji sebelum penyaringan (1.335 baris). Hal ini menunjukkan bahwa pembagian data tidak tetap antar-aksi Spark pada eksekusi tersebut, sehingga data latih dan data uji dapat berbeda antar-perhitungan dan keterpisahannya tidak dapat dipastikan. R^2 sebelum penyaringan (0,8249) dan R^2 setelah penyaringan (0,8770) lebih tinggi daripada R^2 pada Tabel 5 (0,7454), yang kemungkinan mencerminkan perbedaan data uji dan pembagian data, sehingga tidak dibandingkan dengan Tabel 5. Hasil utama penelitian adalah Tabel 5, yang memakai pembagian data tetap.

Tabel 6. Hasil Eksperimen Tambahan (Indikatif) Sebelum dan Sesudah Penyaringan Outlier

Metrik	Sebelum penyaringan	Setelah penyaringan
MAE	Rp1.321.340.177,58	Rp1.179.593.767,54
RMSE	Rp3.171.086.850,67	Rp2.658.098.325,58
R^2	0,8249	0,8770

3.10 Evaluasi Waktu Komputasi Apache Spark

Pengukuran waktu dilakukan satu kali pada Apache Spark 4.0.5 dengan konfigurasi local[*] di Google Colab. Data yang dipakai adalah 5.459 baris latih dan 1.386 baris uji dari satu kali pembagian ulang pada eksperimen sejenis Tabel 6. Jumlah data latih ini berbeda dari 5.537 baris pada Tabel 6 karena pembagian data tidak tetap antar-aksi Spark, sehingga tidak identik dengan data pada Tabel 5 maupun Tabel 6. Metrik model pada eksekusi ini (R^2 0,7803) berasal dari pembagian yang berbeda dan tidak dibandingkan dengan Tabel 5. Total waktu 449,6421 detik (sekitar 7 menit 30 detik), dengan pelatihan satu model Random Forest (100 pohon, kedalaman 10) 421,6009 detik (93,8%), serta *preprocessing* 8,9636, prediksi 8,0316, dan evaluasi 11,0459 detik (Tabel 7). Pengukuran ini

tunggal, tanpa ulangan, tanpa pembanding platform lain, dan data latih tidak diatur partisinya, sehingga waktu dapat memuat biaya penjadwalan pada data berukuran ribuan baris. Karena itu angka ini hanya menggambarkan biaya komputasi alur pada mode lokal dan tidak digunakan untuk menyimpulkan efisiensi Spark.

Tabel 7. Hasil Evaluasi Waktu Komputasi Apache Spark

Tahapan	Waktu (detik)
Preprocessing	8,9636
Training	421,6009
Prediksi	8,0316
Evaluasi	11,0459
Total	449,6421

3.11 Keterbatasan dan Validitas

Pertama, penyaringan nilai ekstrem dilakukan sebelum pembagian data sehingga data uji juga disaring. Akibatnya R^2 0,7454 hanya berlaku untuk rumah dalam batas penyaringan (harga maksimal Rp100 miliar, luas tanah maksimal 5.000 m², luas bangunan maksimal 3.000 m², serta batas jumlah kamar dan carport). Batas tersebut ditetapkan setelah memeriksa distribusi data dan bukan batas universal, serta belum diuji sensitivitasnya terhadap alternatif ambang. Nilai ekstrem seperti luas bangunan 3.006.000 m² juga belum diverifikasi ke sumber data karena dapat berupa kesalahan input.

Kedua, validasi parameter memakai 3-fold *cross-validation* pada empat kombinasi parameter, dan konfigurasi terbaik berada di tepi ruang pencarian. Evaluasi akhir tetap memakai satu pembagian data 80:20 dengan seed 42, sehingga variasi kinerja antar pembagian data belum tergambar dan simpangan baku antar-lipatan belum dilaporkan. Eksperimen sebelum-sesudah penyaringan (Tabel 6) juga memakai pembagian data yang tidak tetap antar-aksi Spark, sehingga hanya bersifat indikatif.

Ketiga, model pembanding hanya *Linear Regression* tanpa pengujian statistik atas perbedaan kinerja; algoritma lain seperti *Gradient Boosted Trees* belum diuji. Keempat, waktu komputasi hanya diukur pada Spark local[*] tanpa pembanding platform lain dan tanpa ulangan, sehingga belum menunjukkan efisiensi relatif Spark, dan klaster terdistribusi tidak diuji. Kelima, data berupa harga tawaran pada satu wilayah tanpa atribut seperti koordinat, usia bangunan, atau kondisi rumah, sehingga hasil tidak dapat digeneralisasi ke kota lain.

4. KESIMPULAN DAN SARAN

Penelitian ini menerapkan alur analisis dan prediksi harga rumah menggunakan Apache Spark dan Random Forest Regression. Dari 7.609 baris data awal dengan delapan atribut, penghapusan 617 baris duplikat dan satu baris dengan luas bangunan tidak valid menghasilkan 6.991 baris data bersih. Penyaringan nilai ekstrem berdasarkan batas operasional mengurangi data menjadi 6.845 baris, dengan 146 baris atau 2,09% dikeluarkan. Data kemudian dibagi menjadi 5.466 baris pelatihan dan 1.379 baris pengujian.

Pengujian empat kombinasi parameter melalui 3-fold cross-validation menghasilkan konfigurasi terbaik numTrees=100 dan maxDepth=10 dengan R^2 rata-rata 0,7434. Pada data pengujian, Random Forest memperoleh MAE Rp1.401.144.747,20, RMSE Rp3.470.936.218,34, dan R^2 0,7454. Model pembanding Linear Regression memperoleh MAE Rp1.825.491.330,61, RMSE Rp3.870.302.942,97, dan R^2 0,6834. Dengan demikian, Random Forest menunjukkan kinerja lebih baik pada ketiga metrik dalam eksperimen ini.

Kontribusi penelitian bersifat metodologis dan empiris: (1) karakterisasi pengaruh penyaringan outlier, yaitu simpangan baku harga turun dari Rp15,26 miliar menjadi Rp7,02 miliar, sedangkan eksperimen sebelum-sesudah tambahan hanya bersifat indikatif karena pembagian datanya tidak tetap; (2) pemilihan parameter dengan 3-fold *cross-validation* tanpa melibatkan data uji; (3) perbandingan dengan *Linear Regression*, yang menunjukkan *Random Forest* lebih baik pada ketiga metrik; dan (4) alur PySpark dengan jumlah data yang dapat ditelusuri pada setiap tahap, serta pengukuran waktu komputasi pada Spark lokal yang dilaporkan tanpa pembanding

platform lain. Analisis kesalahan menunjukkan kesalahan terbesar terjadi pada rumah berharga sangat tinggi, dengan selisih hingga Rp49,8 miliar. Temuan berlaku untuk rumah dalam batas penyaringan, satu pembagian data, dan Spark mode lokal, dan bukan klaim algoritma baru atau keunggulan komputasi terdistribusi.

Berdasarkan hasil dan keterbatasan, saran penelitian selanjutnya dirumuskan sebagai pengujian yang dapat dibuktikan:

1. Uji sensitivitas penyaringan outlier. Gunakan pembagian data yang tetap (data di-*cache* sebelum randomSplit), lalu bandingkan batas operasional dengan batas persentil (misalnya P1–P99 dari data latih) dan beberapa alternatif ambang, lalu laporkan perubahan MAE, RMSE, dan R^2 pada data uji yang sama serta pada data uji tanpa penyaringan.
2. Validasi yang lebih kuat. Gunakan *cross-validation* berulang untuk evaluasi akhir, perluas ruang pencarian parameter (numTrees dan maxDepth yang lebih besar), dan laporkan rata-rata serta simpangan baku antar-lipatan.
3. Perbandingan algoritma. Tambahkan *Gradient Boosted Trees* dan *Decision Tree* pada dataset, pembagian data, dan metrik yang sama, disertai pengujian statistik atas selisih kinerja antar-lipatan.
4. Atribut tambahan. Uji apakah koordinat lokasi, usia bangunan, dan kondisi rumah menurunkan RMSE uji dibandingkan model enam prediktor pada skema validasi yang sama.
5. Evaluasi Apache Spark. Lengkapi pengukuran waktu pada local[*] dengan pembanding pustaka lokal seperti scikit-learn pada data dan konfigurasi yang sama, beberapa ukuran data, dan klaster terdistribusi bila tersedia, dengan metrik waktu pelatihan dan rasio percepatan.
6. Verifikasi nilai ekstrem. Telusuri nilai ekstrem ke sumber data untuk memisahkan kesalahan input dari rumah mewah yang sah sebelum disaring.

DAFTAR PUSTAKA

Akhmadi, & Budiman, F. (2025). Prediksi Harga Rumah Menggunakan Algoritma Regresi Linier,Random Forest, Dan Gradient Boosting. *JURIKOM (Jurnal Riset Komputer)*, *12*(6), 1018–1025. https://doi.org/https://doi.org/10.30865/jurikom.v12i6.9369

Balqis, F. H., & Aini, Q. (2026). Prediksi Harga Rumah menggunakan Algoritma Random Forest Regression. *Sistemasi: Jurnal Sistem Informasi*, *15*(2), 433–445. https://doi.org/https://doi.org/10.32520/stmsi.v15i2.5726

Dritsas, E., & Trigka, M. (2025). Applying Machine Learning on Big Data With Apache Spark. *IEEE Access*, *13*, 53377–53393. https://doi.org/https://doi.org/10.1109/ACCESS.2025.3552042

Fausi, M. R., Handika, M., Awinanto, A., Wahidin, A. J., Rahmatullah, B., & Kurniawati, I. (2025). Analisis Perbandingan Kinerja Algoritma Linear Regression , Random Forest , dan XGBoost dalam Prediksi Harga Rumah. *RIGGS: Journal of Artificial Intelligence and Digital Business*, *4*(4), 1541–1548. https://doi.org/https://doi.org/10.31004/riggs.v4i4.3620

Marhananda, A. A., & Fabiandika, F. E. (2026). Sistem Prediksi Harga Rumah Di Kota Malang Menggunakan Metode Random Forest Regressor Berbasis Machine Learning. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *10*(1), 1686–1693. https://doi.org/https://doi.org/10.36040/jati.v10i1.16594

Noorihsan, S. Y. F., Widhianingsih, T. D. A., & Kuswanto, H. (2026). Predictive Analytics for Property Valuation Using Random Forest in Malang City. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *6*(1), 186–192. https://doi.org/https://doi.org/10.57152/malcom.v6i1.2411

Pangestu, A., & Iswahyudi, R. T. (2025). Pengaruh Data Preprocessing Terhadap Performa Regresi Linier Dalam Prediksi Saham. *IKRA-ITH Informatika: Jurnal Komputer Dan Informatika*, *9*(2), 204–210. https://doi.org/https://doi.org/10.37817/ikraith-informatika.v9i2

Qolbi, M. B. S., Puteh, T. N., Rivandi, & Rozikin, C. (2025). Prediksi Harga Rumah Di Jakarta Pusat Menggunakan Algoritma Machine Learning. *JIKB: Jurnal Ilmu Komputer Dan Bisnis*, *16*(1), 16–24. https://doi.org/https://doi.org/10.47927/jikb.v16i1.840

Salsabila, M., Adiwijaya, R. P., & Octaviana, L. N. (2025). Implementasi Machine Learning untuk Memprediksi Harga Rumah dengan Menggunakan Metode Decision Tree Regressor dan Random Forest Regressor. *SETRUM: Sistem Kendali-Tenaga-Elektronika-Telekomunikasi-Komputer*, *14*(1), 46–56. https://doi.org/https://doi.org/10.62870/setrum.v14i1.31025

Sharma, H., Harsora, H., & Ogunleye, B. (2024). An Optimal House Price Prediction Algorithm : XGBoost. *Analytics*, *3*(1), 30–45. https://doi.org/https://doi.org/10.3390/analytics3010003

Sobirin, R., Kusrini, & Wibowo, F. W. (2025). Estimasi harga properti rumah menggunakan algoritma multiple linear regression dan random forest regressor dengan teknik web scraping pada website penjualan rumah di yogyakarta. *JIPI (Jurnal Ilmiah Penelitian Dan Pembelajaran Informatika)*, *10*(4), 3301–3312. https://doi.org/https://doi.org/10.29100/jipi.v10i4.6463

Sukriadi, Achriaty, N., Nugraha, M. S., & Fitranda. (2026). Prediksi Kepatuhan Pendaftaran Tanah Berbasis Machine Learning Menggunakan Algoritma Random Forest. *JISTI: Jurnal Ilmiah Sistem Informasi Dan Teknik Informatika*, *9*(2), 299–308. https://doi.org/https://doi.org/10.57093/jisti.v9i2.434

Sukriadi, Ismail, & Andzar, A. . (2023). Penerapan Text Mining Dalam Klasifikasi Judul Skripsi Yang Diusulkan Mahasiswa Menggunakan Metode Naïve Bayes. *JISTI: Jurnal Ilmiah Sistem Informasi Dan Teknik Informatika*, *6*(2), 184–196. https://doi.org/https://doi.org/10.57093/jisti.v6i2.174

Sutanto, Y., Amin, B. Al, Setyadi, H. A., & Purnama, B. E. (2025). Prediksi Harga Perumahan Menggunakan Metode Principal Component Analysis Dan Random Forest Regresi. *JTIIK: Jurnal Teknologi Informasi Dan Ilmu Komputer*, *12*(6), 1243–1250. https://doi.org/https://doi.org/10.25126/jtiik.2025126

Warjiyono, Rais, A. N., Alfarobi, I., Hadi, S. W., & Kurniawan, W. (2024). Analisa Prediksi Harga Jual Rumah Menggunakan Algoritma Random Forest Machine Learning. *JURSISTEKNI (Jurnal Sistem Informasi Dan Teknologi Informasi)*, *6*(2), 416–423. https://doi.org/https://doi.org/10.52005/jursistekni.v6i2.323