Journal of Information System, Applied, Management, Accounting and Research.
http://journal.stmikjayakarta.ac.id/index.php/jisamar ,
jisamar@stmikjayakarta.ac.id , jisamar2017@gmail.com
e-ISSN: 2598-8719 (Online), p-ISSN: 2598-8700 (Printed) , Vol. 10 No.3 (August 2026)

HETEROGENITAS SPASIAL PADA PREDIKSI HARGA RUMAH MULTI-KOTA: EVALUASI STACKING ENSEMBLE DAN INTERPRETASI SHAP

Spatial Heterogeneity in Multi-City House Price Prediction: Stacking Ensemble Evaluation and SHAP Interpretation

Gede Aby Sanjiwa Kurniawan [1],
I Nyoman Tri Anindia Putra[2]

Program Studi Sistem Informasi[1,2]
Fakultas Teknik dan Kejuruan[1,2]
Universitas Pendidikan Ganesha[1,2]

*Correspondent Author: aby@undiksha.ac.id

Authors Email: aby@undiksha.ac.id[1], tri.anindia@undiksha.ac.id[2]

Received: June 30, 2026**. Revised:** August 23, 2026**. Accepted:** August 25, 2026**. Issue Period:** Vol.10 No.3 (2026), Pp. 1020-1032

Abstrak: Prediksi harga rumah menggunakan model stacking ensemble umumnya dievaluasi secara agregat pada keseluruhan data uji, sehingga berpotensi menyembunyikan variasi performa yang sesungguhnya terjadi antar wilayah geografis. Penelitian ini menganalisis performa model Stacking Ensemble yang menggabungkan XGBoost, Random Forest, dan CatBoost dengan meta learner Ridge Regression pada dataset 100.000 rumah di tujuh kota Indonesia, dengan penekanan pada evaluasi yang terdekomposisi per kota dan analisis faktor penyebab disparitasnya. Hasil evaluasi agregat menunjukkan Stacking Ensemble mencapai performa terbaik dengan MAE sebesar Rp 1.106.456.898 dan R2 sebesar 0,8458, namun uji Wilcoxon signed rank menunjukkan keunggulan tersebut hanya signifikan secara statistik dibandingkan Random Forest, tidak terhadap CatBoost maupun XGBoost. Ketika performa didekomposisi per kota, ditemukan disparitas MAE lebih dari 2,5 kali lipat antara kota dengan performa terbaik dan terburuk, disertai variasi signifikansi statistik yang tidak seragam antar wilayah. Analisis lebih lanjut menunjukkan disparitas ini lebih berkaitan dengan variasi harga yang lebih besar pada kota tertentu dibandingkan keterbatasan jumlah data pelatihan. Analisis SHAP turut mengindikasikan kontribusi fitur lokasi yang cenderung lebih besar pada kota dengan performa lebih rendah, meskipun pola ini tidak berlaku seragam di seluruh kota yang diuji. Temuan ini menegaskan pentingnya evaluasi granular per wilayah dalam penelitian prediksi harga rumah multi kota.

Kata kunci: stacking ensemble, prediksi harga rumah, disparitas performa, SHAP, heterogenitas spasial.

DOI: 10.52362/jisamar.v10i3.2660

Journal of Information System, Applied, Management, Accounting and Research.
http://journal.stmikjayakarta.ac.id/index.php/jisamar ,
jisamar@stmikjayakarta.ac.id , jisamar2017@gmail.com
e-ISSN: 2598-8719 (Online), p-ISSN: 2598-8700 (Printed) , Vol. 10 No.3 (August 2026)

Abstract: *House price prediction using stacking ensemble models is commonly evaluated on an aggregate basis across the entire test set, which risks concealing performance variation that actually occurs across geographic regions. This study analyzes the performance of a Stacking Ensemble combining XGBoost, Random Forest, and CatBoost with a Ridge Regression meta learner on a dataset of 100,000 houses across seven cities in Indonesia, with an emphasis on city level decomposed evaluation and analysis of the underlying causes of disparity. Aggregate evaluation shows the Stacking Ensemble achieves the best performance with an MAE of Rp 1,106,456,898 and R2 of 0.8458, yet the Wilcoxon signed rank test indicates this advantage is statistically significant only compared to Random Forest, not to CatBoost or XGBoost. When performance is decomposed by city, MAE disparity of more than 2.5 times is found between the best and worst performing cities, accompanied by inconsistent statistical significance across regions. Further analysis shows this disparity is more closely related to greater price variation in certain cities than to limitations in training data volume. SHAP analysis further indicates that location feature contributions tend to be larger in cities with lower performance, although this pattern does not hold uniformly across all cities examined. These findings underscore the importance of granular, city level evaluation in multi city house price prediction research.*

Keywords: *stacking ensemble, house price prediction, performance disparity, SHAP, spatial heterogeneity.*

I. PENDAHULUAN

Permasalahan perumahan merupakan isu penting yang dihadapi masyarakat modern, terutama di daerah dengan pertumbuhan ekonomi tinggi seperti kota kota besar di Indonesia. Harga properti yang sangat dinamis dan dipengaruhi oleh berbagai faktor menjadikan fluktuasi harga rumah sebagai fenomena yang tidak dapat dihindari [1]. Sebagai contoh, studi di Kota Surabaya menunjukkan bahwa lokasi dan tipologi hunian turut membentuk preferensi serta tren harga pada kalangan generasi muda [2], sementara studi di kawasan aglomerasi Yogyakarta mengidentifikasi sejumlah faktor yang memengaruhi variasi harga rumah di wilayah tersebut [3]. Kondisi ini menjadi perhatian tidak hanya bagi masyarakat umum, tetapi juga bagi investor, pengembang properti, dan pemerintah daerah dalam merancang kebijakan hunian jangka panjang. Sehingga diperlukan pendekatan prediksi harga yang akurat dan adaptif untuk membantu berbagai pihak dalam pengambilan keputusan properti.

Sejumlah penelitian telah menggunakan pendekatan machine learning untuk memprediksi harga rumah. Model regresi linear banyak dimanfaatkan karena kemudahan interpretasinya, namun beberapa studi menunjukkan bahwa model berbasis pohon keputusan seperti Random Forest dapat menghasilkan akurasi yang lebih tinggi dibandingkan regresi linear maupun Gradient Boosted Trees pada kasus prediksi harga rumah, karena kemampuannya menangani hubungan antar variabel yang lebih kompleks dan non linear [4]. Model berbasis pohon keputusan lain seperti XGBoost juga menunjukkan performa yang baik dalam kondisi tersebut. Qolbi dkk. [5] melaporkan Random Forest menghasilkan R2 sebesar 0,942 pada prediksi harga rumah di Jakarta Pusat, sementara penelitian lain menunjukkan XGBoost mampu mencapai performa tinggi dengan optimasi hyperparameter pada data harga rumah [6].

Pendekatan ensemble learning menawarkan solusi lebih lanjut dengan menggabungkan beberapa algoritma untuk meningkatkan akurasi dan stabilitas prediksi [7]. Dalam kerangka ensemble, stacking mengatasi keterbatasan bagging dan boosting dengan menggabungkan prediksi dari beberapa model heterogen melalui sebuah meta learner yang secara adaptif mempelajari kombinasi optimal dari setiap base learner [8]. Hibatulloh dkk. [9] melaporkan bahwa Stacking Ensemble menghasilkan kinerja terbaik dengan R2 sebesar 0,9076 dalam konteks prediksi harga rumah di Bandung, mengungguli Random Forest, Gradient Boosting, dan XGBoost secara individual, dan memperkuat temuan tersebut dengan analisis interpretabilitas berbasis SHAP yang mengidentifikasi luas tanah dan luas bangunan sebagai fitur paling berpengaruh terhadap harga.

Meskipun demikian, sebagian besar penelitian terdahulu memusatkan analisis pada satu kota, sehingga belum banyak terungkap apakah model yang dilatih pada data gabungan dari banyak kota mampu

DOI: 10.52362/jisamar.v10i3.2660

Journal of Information System, Applied, Management, Accounting and Research.
http://journal.stmikjayakarta.ac.id/index.php/jisamar ,
jisamar@stmikjayakarta.ac.id , jisamar2017@gmail.com
e-ISSN: 2598-8719 (Online), p-ISSN: 2598-8700 (Printed) , Vol. 10 No.3 (August 2026)

mempertahankan performa yang konsisten di seluruh wilayah tersebut. Isu ini dikenal secara luas dalam literatur internasional sebagai heterogenitas spasial (spatial heterogeneity), yaitu kondisi ketika sampel data spasial tidak mengikuti distribusi yang identik di seluruh area studi, sehingga model global yang dilatih dari seluruh area berpotensi menghasilkan prediksi yang kurang akurat pada wilayah lokal tertentu [10]. Beberapa studi valuasi properti berbasis machine learning turut mengindikasikan adanya variasi akurasi model valuasi otomatis (automated valuation model) pada konteks data yang berbeda [11], dan pendekatan seperti Geographically Neural Network Weighted Regression telah dikembangkan untuk mengakomodasi heterogenitas spasial pada prediksi harga rumah [12]. Namun, sejauh penelusuran penulis, penerapan konsep heterogenitas spasial pada kerangka stacking ensemble multi model, khususnya untuk konteks harga rumah di Indonesia, belum banyak dibahas.

Selain itu, penggunaan analisis interpretability seperti SHAP pada penelitian prediksi harga rumah umumnya baru sebatas menjelaskan kontribusi fitur secara global, dan belum banyak dimanfaatkan untuk membandingkan pola kontribusi fitur antar wilayah dengan tingkat akurasi yang berbeda. Padahal, jika suatu model menunjukkan akurasi yang lebih rendah pada wilayah tertentu, pemahaman mengenai fitur apa yang mendominasi prediksi pada wilayah tersebut dapat memberikan petunjuk mengenai penyebab disparitas performa itu sendiri.

Berdasarkan hal tersebut, penelitian ini bertujuan untuk: (1) mengukur perbandingan performa model Stacking Ensemble (XGBoost, Random Forest, dan CatBoost dengan meta learner Ridge) terhadap model tunggal dalam memprediksi harga rumah tinggal pada tujuh kota di Indonesia; (2) menganalisis apakah performa model tersebut konsisten di seluruh kota atau terdapat disparitas yang signifikan secara statistik; dan (3) menjelaskan kemungkinan penyebab disparitas performa antar kota tersebut melalui analisis interpretability berbasis SHAP.

Pemilihan XGBoost, Random Forest, dan CatBoost sebagai base learner dalam penelitian ini didasarkan pada perbedaan karakteristik pendekatan yang dimiliki masing masing model. Random Forest dipilih karena kemampuannya menangani hubungan non linear dan ketahanannya terhadap overfitting pada data tabular, sebagaimana ditunjukkan pada perbandingan performanya dengan regresi linear pada kasus prediksi harga rumah [4]. XGBoost dipilih karena efisiensi komputasi dan performanya yang tinggi pada data tabular dengan optimasi hyperparameter [5]. CatBoost dipilih karena kemampuannya menangani fitur kategorikal secara native serta karakteristik regularisasi yang berbeda dari XGBoost, sehingga berpotensi memberikan perspektif tambahan yang saling melengkapi dalam kerangka stacking [9]. Perbedaan karakteristik bias variance trade off dari ketiga model ini menjadikannya kandidat base learner yang saling melengkapi, sementara Ridge Regression dipilih sebagai meta learner karena kemampuan regularisasinya dalam menangani multikolinearitas antar prediksi base learner [9].

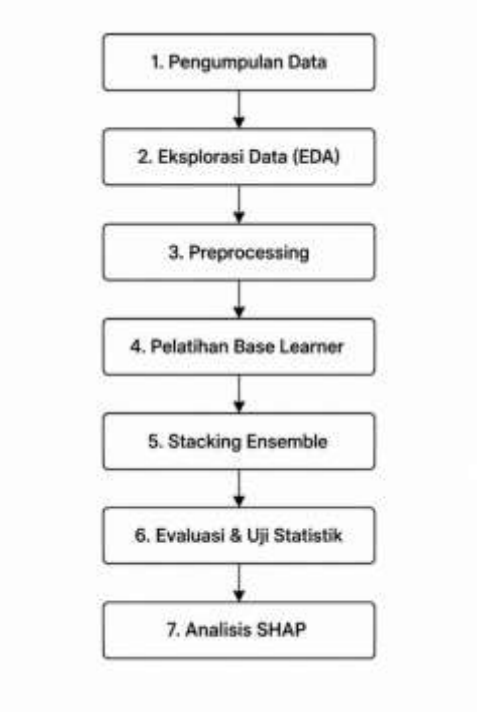
II. METODE DAN MATERI

2.1 Tahapan Penelitian

Penelitian ini dilakukan melalui tujuh tahapan utama secara berurutan, yaitu pengumpulan data, eksplorasi data, preprocessing, pelatihan base learner, pembentukan stacking ensemble, evaluasi model dan uji signifikansi statistik, serta analisis interpretabilitas SHAP. Alur tahapan tersebut divisualisasikan pada Gambar 1.

DOI: 10.52362/jisamar.v10i3.2660

Journal of Information System, Applied, Management, Accounting and Research.
http://journal.stmikjayakarta.ac.id/index.php/jisamar ,
jisamar@stmikjayakarta.ac.id , jisamar2017@gmail.com
e-ISSN: 2598-8719 (Online), p-ISSN: 2598-8700 (Printed) , Vol. 10 No.3 (August 2026)



Gambar 1. Diagram Alur Penelitian

2.2 Pengumpulan Data

Dataset yang digunakan terdiri atas 100.000 baris data rumah tinggal dari tujuh kota di Indonesia, yaitu Bekasi, Bogor, Surabaya, Bandung, Tangerang, Jakarta, dan Depok. Dataset memuat fitur numerik dan kategorikal, antara lain luas tanah, luas bangunan, jumlah kamar tidur, jumlah kamar mandi, jumlah lantai, tahun bangun, sertifikat, kondisi bangunan, akses jalan, keberadaan taman, keamanan, dan garasi, dengan variabel target berupa harga rumah dalam Rupiah. Dataset ini bersifat sintetis sehingga hasil penelitian perlu diinterpretasikan dengan mempertimbangkan batasan tersebut.

2.3 Eksplorasi Data (EDA)

Eksplorasi data dilakukan untuk memeriksa nilai hilang, duplikasi data, serta distribusi tiap kolom numerik melalui analisis skewness dan proporsi outlier menggunakan metode IQR. Hasil eksplorasi menunjukkan bahwa kolom Harga_Rumah memiliki skewness sebesar 1,871 (severe skew), sementara kolom numerik lain memiliki skewness mendekati nol. Eksplorasi juga menemukan bahwa kolom Kategori_Rumah merupakan label yang dibentuk langsung dari Harga_Rumah, ditunjukkan oleh rentang harga pada tiap kategorinya yang tidak saling tumpang tindih.

2.4 Preprocessing Data

Berdasarkan hasil eksplorasi, dilakukan feature selection dengan membuang kolom ID_Rumah karena tidak memiliki nilai prediktif, serta kolom Kategori_Rumah karena teridentifikasi sebagai sumber data leakage. Transformasi logaritma natural diterapkan pada Harga_Rumah untuk mengurangi skewness, dilakukan sebelum proses pembersihan outlier. Outlier kemudian dibersihkan menggunakan metode IQR yang diterapkan khusus pada Harga_Rumah dalam skala log, karena kolom numerik lain tidak menunjukkan indikasi outlier maupun skewness signifikan. Data selanjutnya dibagi menjadi data latih dan data uji dengan proporsi 80:20. Fitur kategorikal berkardinalitas rendah dikodekan menggunakan One Hot Encoding, sedangkan fitur Lokasi_Kota dan Lokasi_Kecamatan yang berkardinalitas tinggi dikodekan menggunakan target encoding [13] yang dihitung dari rata rata target pada skala log, dengan proses fitting hanya dilakukan pada data latih untuk mencegah kebocoran informasi ke data uji.

DOI: 10.52362/jisamar.v10i3.2660

Journal of Information System, Applied, Management, Accounting and Research.
http://journal.stmikjayakarta.ac.id/index.php/jisamar ,
jisamar@stmikjayakarta.ac.id , jisamar2017@gmail.com
e-ISSN: 2598-8719 (Online), p-ISSN: 2598-8700 (Printed) , Vol. 10 No.3 (August 2026)

2.5 Pelatihan Base Learner

Tiga *base learner* digunakan dalam penelitian ini, yaitu *XGBoost, Random Forest,* dan *CatBoost. CatBoost* dipilih sebagai salah satu *base learner* karena kemampuannya menangani fitur kategorikal secara native melalui pendekatan *ordered boosting* yang mengurangi bias prediksi akibat target leakage pada proses boosting konvensional [14]. Setiap base learner dituning menggunakan *RandomizedSearchCV* dengan skema 5 *fold cross validation* dan metrik MAE pada skala log sebagai fungsi objektif pencarian *hyperparameter*, mencakup parameter seperti jumlah estimator, kedalaman pohon, *learning rate*, serta parameter regularisasi masing masing algoritma.

2.6 Pelatihan Stacking Ensemble

Meta-features untuk melatih *meta learner* dibentuk melalui skema *out of fold (OOF)* 5 *fold* secara manual, yaitu setiap *base learner* dilatih pada empat *fold* data latih dan menghasilkan prediksi pada fold kelima yang tidak dilihat selama pelatihan, sehingga meta learner belajar dari prediksi yang belum pernah diketahui base learner sebelumnya. Prediksi akhir stacking ensemble dirumuskan sebagai kombinasi linear dari prediksi tiap base learner melalui meta learner *Ridge Regression*, sebagaimana ditunjukkan pada persamaan (1).

$$\hat{y}_{stacking} = \omega_0 + \omega_1 \hat{y}_{XGBoost} + \omega_2 \hat{y}_{Random\ Forest} + \omega_3 \hat{u}_{CatBoost} \quad (1)$$

dengan ŷ_stacking adalah prediksi akhir pada skala log, w0 adalah intercept, serta w1, w2, dan w3 adalah koefisien bobot hasil pelatihan Ridge Regression terhadap prediksi masing masing base learner.

2.7 Evaluasi Model dan Uji Statistik

Performa model dievaluasi menggunakan tiga metrik, yaitu Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), dan koefisien determinasi (R^2), yang dihitung setelah prediksi dikonversi kembali ke skala Rupiah menggunakan fungsi eksponensial. Ketiga metrik dirumuskan pada persamaan (2), (3), dan (4).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (2)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (3)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (4)$$

Keterangan untuk (2), (3), (4):

y_i = nilai aktual (harga rumah sebenarnya)

$\hat{y}_i$ = nilai prediksi model

$\bar{y}$ = rata-rata nilai aktual

n = jumlah data

Evaluasi dilakukan pada tingkat agregat dan tingkat per kota untuk melihat konsistensi performa model di seluruh wilayah. Signifikansi statistik selisih absolute error antara stacking ensemble dan tiap base learner individual diuji menggunakan uji Wilcoxon signed rank, sebuah uji non parametrik yang umum digunakan sebagai alternatif paired t test dalam membandingkan performa model machine learning secara berpasangan tanpa mengasumsikan distribusi normal pada data [15], dengan taraf signifikansi alpha 0,05, baik pada tingkat agregat maupun pada tingkat per kota.

2.8 Analisis Interpretabilitas SHAP

Analisis SHAP (SHapley Additive exPlanations) [16] digunakan untuk menjelaskan kontribusi tiap fitur terhadap prediksi model dengan performa terbaik. Nilai SHAP dihitung pada sampel sebanyak 2000 baris data uji. Untuk mengidentifikasi kemungkinan penyebab disparitas performa antar kota, nilai SHAP dibandingkan pada empat kota terpilih, yaitu dua kota dengan MAE tertinggi dan dua kota dengan MAE terendah.

DOI: 10.52362/jisamar.v10i3.2660

Journal of Information System, Applied, Management, Accounting and Research.
http://journal.stmikjayakarta.ac.id/index.php/jisamar ,
jisamar@stmikjayakarta.ac.id , jisamar2017@gmail.com
e-ISSN: 2598-8719 (Online), p-ISSN: 2598-8700 (Printed) , Vol. 10 No.3 (August 2026)

III. PEMBAHASAN DAN HASIL

3.1 Hasil Pengumpulan dan Deskripsi Data

Dataset yang digunakan terdiri atas 100.000 baris data rumah tinggal yang tersebar di tujuh kota, yaitu Bekasi (14.372 baris), Bogor (14.359 baris), Surabaya (14.321 baris), Bandung (14.319 baris), Tangerang (14.307 baris), Jakarta (14.283 baris), dan Depok (14.039 baris), dengan proporsi yang relatif seimbang antar kota (masing masing sekitar 14 persen dari total data). Variabel target Harga_Rumah memiliki rentang antara Rp 323.572.967 hingga Rp 40.178.481.053, dengan rata rata sebesar Rp 5.047.165.000.

Hasil eksplorasi data menunjukkan skewness Harga_Rumah sebesar 1,871 yang tergolong severe skew, sementara kolom numerik lain menunjukkan variasi skewness dari mendekati nol (Luas_Tanah, Jumlah_Kamar_Tidur, Tahun_Bangun) hingga moderate (Luas_Bangunan 0,969, Jumlah_Kamar_Mandi 0,818, Jumlah_Lantai 1,179). Proporsi outlier menggunakan metode IQR pada kolom kolom tersebut relatif kecil (0 hingga 2,02 persen), kecuali pada Harga_Rumah yang mencapai 4,50 persen sebelum transformasi logaritma diterapkan.

Eksplorasi juga menemukan bahwa kolom Kategori_Rumah merupakan label yang dibentuk langsung dari Harga_Rumah, ditunjukkan oleh rentang harga pada tiap kategori yang tidak saling tumpang tindih (Murah: Rp 323.572.967 sampai Rp 799.973.139; Menengah: Rp 800.018.361 sampai Rp 1.999.985.089; Mewah: Rp 2.000.059.166 sampai Rp 4.999.958.723; Super Mewah: Rp 5.000.030.236 sampai Rp 40.178.481.053). Temuan ini mengonfirmasi bahwa Kategori_Rumah berpotensi menjadi sumber data leakage apabila tetap digunakan sebagai fitur prediktor, sehingga kolom tersebut dibuang pada tahap feature selection bersama dengan ID_Rumah yang tidak memiliki nilai prediktif.

3.2 Hasil Preprocessing

Transformasi logaritma natural pada Harga_Rumah berhasil menurunkan skewness dari 1,871 menjadi -0,041, mendekati distribusi normal. Penerapan metode IQR pada Harga_Rumah dalam skala log tidak menghasilkan penghapusan baris data (0 dari 100.000 baris berada di luar batas [19,472, 24,548]), yang menunjukkan bahwa transformasi logaritma sudah cukup efektif menangani ekor distribusi yang panjang tanpa perlu pembuangan data tambahan.

Dataset kemudian dibagi menjadi 80.000 baris data latih dan 20.000 baris data uji. Fitur kategorikal berkardinalitas rendah (delapan kolom dengan 4 hingga 7 kategori) dikodekan menggunakan One Hot Encoding dengan proses alignment antara data latih dan data uji untuk memastikan konsistensi kolom. Fitur Lokasi_Kota (7 kategori) dan Lokasi_Kecamatan (38 kategori) yang berkardinalitas tinggi dikodekan menggunakan target encoding [15], dengan nilai encoding dihitung dari rata rata Harga_Rumah pada skala log, dan proses fitting hanya dilakukan pada data latih untuk mencegah kebocoran informasi ke data uji.

3.3 Hasil Hyperparameter Tuning

Tuning hyperparameter dilakukan menggunakan RandomizedSearchCV dengan skema 5 fold cross validation dan metrik MAE pada skala log sebagai fungsi objektif. XGBoost dituning dengan 30 kandidat parameter (150 fit total, selesai dalam 2,2 menit), menghasilkan parameter terbaik n_estimators=500, max_depth=3, learning_rate=0,03, subsample=0,7, colsample_bytree=1,0, dengan CV MAE skala log sebesar 0,2296. Random Forest dituning dengan 30 kandidat (150 fit total, selesai dalam 27,7 menit), menghasilkan n_estimators=400, max_depth=10, min_samples_split=5, min_samples_leaf=2, max_features=None, dengan CV MAE skala log sebesar 0,2303. CatBoost dituning dengan 20 kandidat (100 fit total, selesai dalam 3,0 menit) mengikuti pendekatan ordered boosting [14], menghasilkan iterations=500, depth=4, learning_rate=0,03, l2_leaf_reg=9, dengan CV MAE skala log terendah di antara ketiga model, yaitu 0,2295.

Nilai max_depth yang relatif dangkal pada XGBoost (3) dan CatBoost (4), dibandingkan dengan Random Forest (10), mengindikasikan bahwa kedua model boosting tersebut mencapai performa optimal dengan pohon yang lebih sederhana namun berjumlah banyak (500 iterasi/estimator), sementara Random Forest memerlukan pohon yang lebih dalam untuk menangkap pola yang setara.

3.4 Hasil Evaluasi Model (Agregat)

Evaluasi pada 20.000 data uji yang tidak pernah dilihat selama pelatihan maupun tuning menunjukkan Stacking Ensemble mencapai performa terbaik pada ketiga metrik, dengan MAE sebesar Rp 1.106.456.898, RMSE sebesar Rp 1.672.641.652, dan R2 sebesar 0,8458. Perbandingan lengkap ditunjukkan pada Tabel 1.

DOI: 10.52362/jisamar.v10i3.2660

Journal of Information System, Applied, Management, Accounting and Research.
http://journal.stmikjayakarta.ac.id/index.php/jisamar ,
jisamar@stmikjayakarta.ac.id , jisamar2017@gmail.com
e-ISSN: 2598-8719 (Online), p-ISSN: 2598-8700 (Printed) , Vol. 10 No.3 (August 2026)

Tabel 1. Perbandingan Performa Model pada Data Uji (Skala Agregat)

Model	MAE (Rp)	RMSE (Rp)	R2
Stacking (XGB+RF+CatBoost, meta=Ridge)	1.106.456.898	1.672.641.652	0,8458
CatBoost	1.106.590.146	1.673.437.725	0,8456
XGBoost	1.107.047.971	1.674.064.558	0,8455
Random Forest	1.111.610.636	1.683.618.311	0,8438

Selisih performa antara Stacking Ensemble dan CatBoost sebagai base learner individual terbaik tergolong sangat kecil, yaitu MAE lebih rendah Rp 133.248 (setara 0,012 persen) dan R2 lebih tinggi 0,0002. Pola ini konsisten dengan koefisien bobot yang dipelajari meta learner Ridge terhadap prediksi OOF ketiga base learner, yaitu CatBoost sebesar 0,5770, XGBoost sebesar 0,3242, dan Random Forest sebesar 0,0995. Dominasi bobot CatBoost mengindikasikan bahwa meta learner menilai prediksi CatBoost sebagai yang paling andal di antara ketiga base learner, sehingga kontribusi tambahan dari kombinasi dengan XGBoost dan Random Forest terhadap prediksi akhir relatif terbatas.

3.5 Hasil Uji Wilcoxon (Agregat)

Uji Wilcoxon signed rank dilakukan untuk menguji signifikansi statistik selisih absolute error antara Stacking Ensemble dan tiap base learner secara berpasangan pada seluruh data uji [16]. Hasil pengujian ditunjukkan pada Tabel 2.

Tabel 2. Hasil Uji Wilcoxon Signed Rank Test Stacking vs Base Learner (Agregat)

Perbandingan	Statistik	P-value	Signifikan (alpha=0,05)
Stacking vs XGBoost	98.836.116	0,1523	Tidak
Stacking vs Random Forest	95.911.255	0,0000	Ya
Stacking vs CatBoost	99.118.255	0,2775	Tidak

Hasil uji menunjukkan bahwa Stacking Ensemble secara statistik berbeda signifikan dari Random Forest (p kurang dari 0,05), namun tidak berbeda signifikan dari XGBoost (p=0,1523) maupun CatBoost (p=0,2775). Temuan ini konsisten dengan selisih metrik MAE dan R2 yang sangat kecil antara Stacking dan kedua model tersebut pada Tabel 1, dan memperkuat interpretasi bahwa keunggulan numerik Stacking Ensemble terhadap CatBoost dan XGBoost, meskipun konsisten arahnya, tidak cukup besar untuk dianggap signifikan secara statistik.

3.6 Hasil Evaluasi Per Kota

Untuk menguji konsistensi performa model di seluruh wilayah cakupan penelitian, evaluasi turut dilakukan pada tingkat per kota menggunakan label kota asli yang disimpan sebelum proses encoding. Hasil evaluasi MAE Stacking Ensemble dan base learner pada tiap kota ditunjukkan pada Tabel 3, diurutkan dari performa terbaik ke terburuk berdasarkan MAE Stacking.

Tabel 3. MAE per Kota untuk Stacking Ensemble dan Base Learner (Rp)

Kota	XGBoost	Random Forest	CatBoost	Stacking

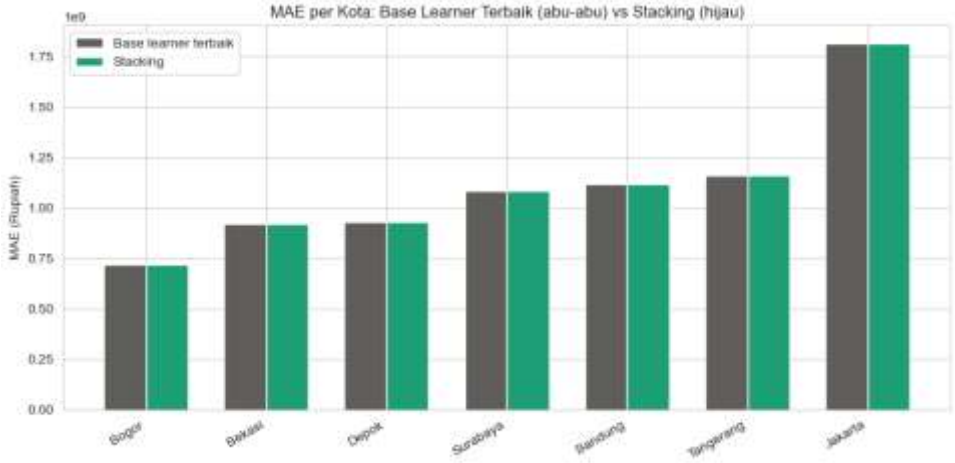
DOI: 10.52362/jisamar.v10i3.2660

Journal of Information System, Applied, Management, Accounting and Research.
http://journal.stmikjayakarta.ac.id/index.php/jisamar ,
jisamar@stmikjayakarta.ac.id , jisamar2017@gmail.com
e-ISSN: 2598-8719 (Online), p-ISSN: 2598-8700 (Printed) , Vol. 10 No.3 (August 2026)

Bogor	720.306.100	726.649.500	721.286.300	721.060.700
Bekasi	923.642.800	926.949.500	922.876.300	923.176.900
Depok	932.669.100	935.902.100	931.586.400	931.956.300
Surabaya	1.084.427.000	1.082.790.000	1.083.660.000	1.083.159.000
Bandung	1.122.435.000	1.125.094.000	1.120.683.000	1.120.868.000
Tangerang	1.161.961.000	1.171.518.000	1.163.956.000	1.162.660.000
Jakarta	1.814.841.000	1.823.583.000	1.813.070.000	1.813.293.000

Sebelum interpretasi lebih lanjut, dilakukan pemeriksaan konsistensi dengan menghitung rata rata tertimbang MAE Stacking lintas kota (berdasarkan proporsi jumlah sampel tiap kota) dan membandingkannya dengan MAE agregat pada Tabel 1. Kedua nilai tersebut identik hingga satuan Rupiah (selisih Rp 0), yang mengonfirmasi bahwa hasil evaluasi per kota konsisten dan tidak mengandung kesalahan perhitungan.

Tabel 3 menunjukkan pola disparitas performa yang jelas antar kota untuk seluruh model, bukan hanya pada Stacking Ensemble. Jakarta secara konsisten menunjukkan MAE tertinggi di antara ketujuh kota pada seluruh model (Rp 1,81 miliar untuk Stacking), sementara Bogor secara konsisten menunjukkan MAE terendah (Rp 721 juta untuk Stacking). Selisih MAE antara kota dengan performa terbaik dan terburuk mencapai lebih dari 2,5 kali lipat, jauh melampaui selisih performa antar model pada tingkat agregat yang dibahas pada bagian D. Perbandingan visual MAE Stacking Ensemble dan base learner terbaik pada tiap kota ditunjukkan pada Gambar 2.



Gambar 2. Perbandingan MAE Stacking Ensemble dan Base Learner Terbaik per Kota

3.7 Hasil Uji Wilcoxon Per Kota

Uji Wilcoxon signed rank turut dilakukan secara terpisah pada tiap kota untuk menguji apakah signifikansi statistik perbedaan performa Stacking terhadap tiap base learner konsisten di seluruh wilayah. Sebelum pengujian, dilakukan pemeriksaan jumlah sampel tiap kota pada data uji (berkisar 2.739 hingga 2.942 baris), yang dinilai cukup memadai untuk pelaksanaan uji ini. Ringkasan hasil ditunjukkan pada Tabel 4.

DOI: 10.52362/jisamar.v10i3.2660

Journal of Information System, Applied, Management, Accounting and Research.
http://journal.stmikjayakarta.ac.id/index.php/jisamar ,
jisamar@stmikjayakarta.ac.id , jisamar2017@gmail.com
e-ISSN: 2598-8719 (Online), p-ISSN: 2598-8700 (Printed) , Vol. 10 No.3 (August 2026)

Tabel 4. Ringkasan Hasil Uji Wilcoxon Signed Rank Test per Kota (Stacking vs Base Learner)

Kota	vs XGBoost	vs Random Forest	vs CatBoost
Bogor	Tidak (p=0,6789)	Ya (p=0,0009)	Tidak (p=0,9142)
Bekasi	Tidak (p=0,8271)	Ya (p=0,0281)	Tidak (p=0,6111)
Depok	Tidak (p=0,6584)	Tidak (p=0,1617)	Tidak (p=0,9766)
Surabaya	Tidak (p=0,1234)	Tidak (p=0,8242)	Tidak (p=0,5590)
Bandung	Ya (p=0,0494)	Ya (p=0,0070)	Tidak (p=0,6988)
Tangerang	Tidak (p=0,4879)	Tidak (p=0,0590)	Tidak (p=0,1997)
Jakarta	Tidak (p=0,8098)	Ya (p=0,0255)	Tidak (p=0,4847)

Pola signifikansi Stacking terhadap Random Forest pada evaluasi agregat (signifikan) ditemukan konsisten pada empat dari tujuh kota (Bogor, Bekasi, Bandung, Jakarta), namun tidak signifikan pada tiga kota lainnya (Depok, Surabaya, Tangerang). Sementara itu, Stacking tidak pernah berbeda signifikan dari CatBoost pada satu kota pun, konsisten dengan hasil agregat. Terhadap XGBoost, hanya Bandung yang menunjukkan perbedaan signifikan, sementara enam kota lainnya tidak. Variasi hasil signifikansi antar kota ini perlu diinterpretasikan dengan hati hati mengingat pengujian dilakukan berulang pada tujuh kota dan tiga pasangan model sekaligus (21 pengujian), sehingga terdapat risiko kesalahan tipe I akibat multiple testing yang tidak dikoreksi pada penelitian ini.

3.8 Analisis Penyebab Disparitas Antar Kota

Untuk menelusuri kemungkinan penyebab disparitas performa yang ditemukan pada Tabel 3, dilakukan analisis tambahan yang menggabungkan MAE Stacking per kota dengan jumlah sampel data latih dan karakteristik sebaran Harga_Rumah pada skala Rupiah asli per kota. Hasil analisis ditunjukkan pada Tabel 5.

Tabel 5. Perbandingan MAE, Jumlah Sampel Latih, dan Sebaran Harga per Kota

Kota	Jumlah Sampel Latih	Std Harga (Rp)	Rentang Harga (Rp)	MAE Stacking (Rp)
Bogor	11.484	2.533.773.000	16.103.280.000	721.060.700
Bekasi	11.430	3.171.500.000	20.066.970.000	923.176.900
Depok	11.300	3.234.959.000	19.400.350.000	931.956.300
Surabaya	11.420	3.837.566.000	23.960.040.000	1.083.159.000
Bandung	11.396	3.887.112.000	23.356.460.000	1.120.868.000
Tangerang	11.507	4.495.205.000	28.744.870.000	1.162.660.000
Jakarta	11.463	6.363.446.000	38.648.180.000	1.813.293.000

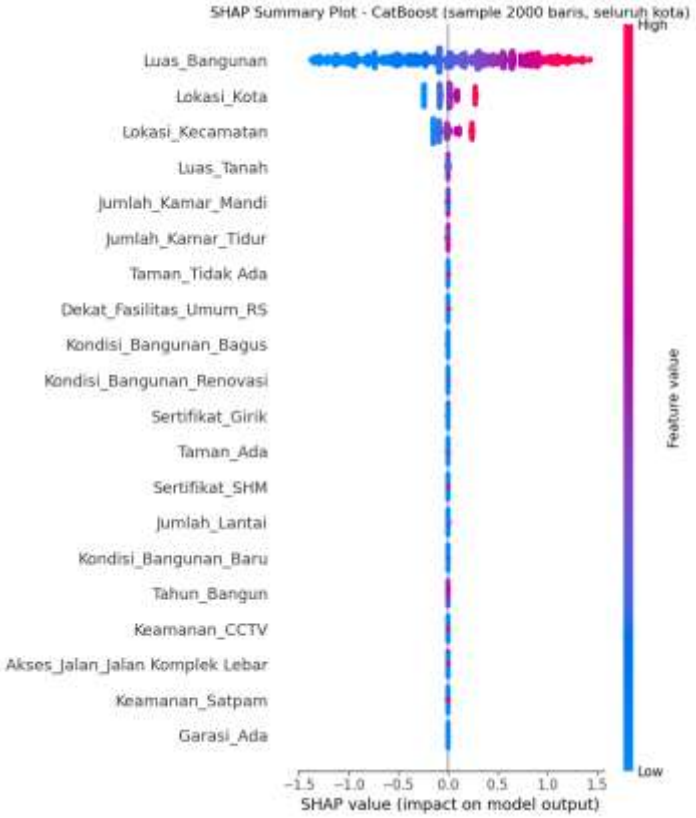
DOI: 10.52362/jisamar.v10i3.2660

Journal of Information System, Applied, Management, Accounting and Research.
http://journal.stmikjayakarta.ac.id/index.php/jisamar ,
jisamar@stmikjayakarta.ac.id , jisamar2017@gmail.com
e-ISSN: 2598-8719 (Online), p-ISSN: 2598-8700 (Printed) , Vol. 10 No.3 (August 2026)

Tabel 5 menunjukkan bahwa jumlah sampel data latih antar kota relatif seragam, berkisar antara 11.300 hingga 11.507 baris (selisih kurang dari 2 persen), sehingga disparitas performa yang ditemukan pada Tabel 3 tidak dapat dijelaskan oleh perbedaan kuantitas data latih antar kota. Sebaliknya, standar deviasi dan rentang Harga_Rumah menunjukkan pola yang sejalan dengan MAE: Jakarta memiliki standar deviasi harga tertinggi (Rp 6,36 miliar) sekaligus MAE tertinggi, sementara Bogor memiliki standar deviasi harga terendah (Rp 2,53 miliar) sekaligus MAE terendah. Kota kota lain menunjukkan pola serupa secara umum, meskipun urutan Surabaya dan Bandung pada std_harga sedikit berbeda urutan dengan MAE keduanya yang berdekatan. Pola ini mengindikasikan bahwa disparitas performa model antar kota lebih berkaitan dengan variasi harga yang lebih besar pada kota tertentu (yang secara inheren lebih sulit diprediksi dengan presisi tinggi), bukan karena keterbatasan jumlah data pelatihan.

3.9 Hasil Analisis SHAP

Analisis SHAP dilakukan pada model CatBoost sebagai base learner dengan performa individual terbaik [14], menggunakan sampel 2.000 baris data uji. Hasil SHAP summary plot pada Gambar 3 menunjukkan bahwa Luas_Bangunan merupakan fitur dengan kontribusi paling dominan terhadap prediksi harga, jauh melampaui fitur lainnya, disusul oleh Lokasi_Kota dan Lokasi_Kecamatan pada posisi kedua dan ketiga. Fitur fitur lain seperti jumlah kamar, sertifikat, kondisi bangunan, dan fasilitas keamanan menunjukkan kontribusi yang jauh lebih kecil.



Gambar 3. SHAP Summary Plot pada Model CatBoost

Untuk menelusuri kemungkinan penyebab disparitas performa antar kota yang ditemukan pada bagian 3.6 dan 3.7, nilai SHAP dibandingkan pada empat kota terpilih berdasarkan MAE CatBoost, yaitu Tangerang dan Jakarta (dua kota dengan MAE tertinggi) serta Bogor dan Bekasi (dua kota dengan MAE terendah). Rata rata nilai absolut SHAP untuk tiga fitur teratas pada keempat kota tersebut ditunjukkan pada Tabel 6.

DOI: 10.52362/jisamar.v10i3.2660

Journal of Information System, Applied, Management, Accounting and Research.
http://journal.stmikjayakarta.ac.id/index.php/jisamar ,
jisamar@stmikjayakarta.ac.id , jisamar2017@gmail.com
e-ISSN: 2598-8719 (Online), p-ISSN: 2598-8700 (Printed) , Vol. 10 No.3 (August 2026)

Tabel 6. Rata-Rata SHAP Value Tiga Fitur Teratas pada Empat Kota Terpilih

Fitur	Tangerang (MAE tinggi)	Jakarta (MAE tinggi)	Bogor (MAE rendah)	Bekasi (MAE rendah)
Luas_Bangunan	0,6189	0,6098	0,6213	0,6250
Lokasi_Kota	0,0901	0,2757	0,2410	0,0842
Lokasi_Kecamatan	0,1000	0,2413	0,1479	0,0885

Kontribusi Luas_Bangunan relatif konsisten pada keempat kota (berkisar 0,610 hingga 0,625), menunjukkan bahwa fitur ini berperan setara di seluruh wilayah. Pola yang lebih menarik ditemukan pada kontribusi fitur lokasi: Jakarta, sebagai kota dengan MAE tertinggi, menunjukkan gabungan bobot SHAP Lokasi_Kota dan Lokasi_Kecamatan tertinggi di antara keempat kota (0,5170), sementara Bekasi, sebagai salah satu kota dengan MAE terendah, menunjukkan gabungan bobot terendah (0,1727). Namun demikian, pola ini tidak sepenuhnya konsisten pada Tangerang, yang meskipun tergolong kota dengan MAE tinggi, menunjukkan gabungan bobot fitur lokasi yang justru lebih rendah (0,1901) dibandingkan Bogor yang tergolong kota dengan MAE rendah (0,3889). Temuan ini mengindikasikan bahwa kontribusi fitur lokasi yang lebih besar berasosiasi dengan performa yang lebih rendah pada sebagian kota (khususnya Jakarta), namun tidak dapat digeneralisasi sebagai pola yang berlaku seragam pada seluruh kota dengan MAE tinggi.

3.10 Pembahasan

Penelitian ini menemukan bahwa Stacking Ensemble menghasilkan performa terbaik secara numerik pada evaluasi agregat, namun keunggulannya terhadap CatBoost dan XGBoost secara individual tidak signifikan secara statistik. Temuan ini konsisten dengan penelitian penilaian properti berbasis image fusion yang turut mengindikasikan variasi akurasi model pada konteks data berbeda [11], serta melengkapi studi Hibatulloh dkk. yang melaporkan performa Stacking Ensemble lebih unggul signifikan pada konteks satu kota [9], dengan kemungkinan perbedaan hasil dipengaruhi cakupan wilayah dan karakteristik dataset yang berbeda.

Temuan yang lebih substansial adalah teridentifikasinya disparitas performa yang konsisten di seluruh model berdasarkan wilayah kota, dengan selisih MAE lebih dari 2,5 kali lipat antara kota terbaik dan terburuk. Fenomena ini sejalan dengan konsep heterogenitas spasial yang telah didokumentasikan pada studi internasional [10], [12], yang menyatakan model global berpotensi menghasilkan prediksi kurang akurat pada wilayah lokal tertentu akibat distribusi data yang tidak identik antar wilayah. Analisis lanjutan menunjukkan disparitas tersebut lebih berkaitan dengan variasi harga yang lebih besar pada kota tertentu dibandingkan keterbatasan jumlah data pelatihan, dan analisis SHAP memberikan indikasi tambahan bahwa kontribusi fitur lokasi yang lebih besar berasosiasi dengan performa lebih rendah pada sebagian kota, meski pola ini bersifat indikatif dan tidak berlaku seragam di seluruh kota yang diuji.

Beberapa keterbatasan perlu dipertimbangkan dalam interpretasi hasil. Dataset yang digunakan bersifat sintetis, sehingga pola disparitas antar kota perlu divalidasi lebih lanjut pada data transaksi riil. Uji Wilcoxon yang dilakukan berulang pada 21 kombinasi kota dan model tidak menerapkan koreksi multiple testing, sehingga sebagian hasil signifikan pada tingkat kota berpotensi mengandung kesalahan tipe I. Penggunaan uji signifikansi frekuentis dalam penelitian ini juga memiliki keterbatasan metodologis yang telah didiskusikan pada literatur, yang mengusulkan pendekatan Bayesian sebagai alternatif untuk perbandingan performa model yang lebih informatif [16]. Analisis SHAP pun hanya dilakukan pada model CatBoost dan sampel 2.000 baris untuk efisiensi komputasi, sehingga pola yang ditemukan belum tentu sepenuhnya merepresentasikan mekanisme pada Stacking Ensemble secara keseluruhan. Penelitian lanjutan disarankan memvalidasi temuan pada data riil, menerapkan koreksi multiple testing yang sesuai, serta mengeksplorasi pendekatan yang secara eksplisit mengakomodasi heterogenitas spasial seperti Geographically Weighted Regression atau variannya.

IV. KESIMPULAN

DOI: 10.52362/jisamar.v10i3.2660

Penelitian ini menganalisis performa model Stacking Ensemble yang menggabungkan XGBoost, Random Forest, dan CatBoost dengan meta learner Ridge untuk prediksi harga rumah pada tujuh kota di Indonesia, dengan penekanan pada evaluasi yang tidak hanya bersifat agregat tetapi juga terdekomposisi per kota. Hasil evaluasi agregat menunjukkan Stacking Ensemble mencapai performa terbaik dengan MAE sebesar Rp 1.106.456.898 dan R2 sebesar 0,8458, namun keunggulannya terhadap CatBoost dan XGBoost secara individual tidak signifikan secara statistik berdasarkan uji Wilcoxon signed rank, dan hanya berbeda signifikan dibandingkan Random Forest. Ketika dekomposisi per kota dilakukan, ditemukan disparitas performa yang konsisten di seluruh model, dengan MAE tertinggi pada Jakarta dan MAE terendah pada Bogor, selisih mencapai lebih dari 2,5 kali lipat. Signifikansi statistik keunggulan Stacking terhadap Random Forest juga tidak seragam, hanya konsisten pada empat dari tujuh kota. Analisis lanjutan menunjukkan disparitas ini lebih berkaitan dengan variasi harga yang lebih besar pada kota tertentu dibandingkan keterbatasan jumlah data pelatihan, yang relatif seragam antar kota. Analisis SHAP turut mengindikasikan bahwa kontribusi fitur lokasi cenderung lebih besar pada kota dengan performa lebih rendah, khususnya Jakarta, meskipun pola ini tidak berlaku seragam pada seluruh kota yang diuji. Temuan ini menegaskan bahwa evaluasi performa model prediksi harga rumah secara agregat berisiko menyembunyikan variasi penting yang terjadi pada tingkat wilayah, dan bahwa faktor penyebab variasi tersebut perlu ditelusuri melalui analisis yang lebih granular. Penelitian lanjutan disarankan memvalidasi temuan ini pada data transaksi riil serta mengeksplorasi pendekatan yang secara eksplisit mengakomodasi heterogenitas spasial.

REFERENASI

[1] A. Fauzi, N. Maulidah, R. Supriyadi, H. Nalatissifa, and S. Diantika, “Prediksi Harga Properti Di Indonesia Menggunakan Algoritma Random Forest,” *J. Artif. Intell. Digit. Bus.*, vol. 4, no. 1, pp. 43–49, 2025.

[2] A. N. Widari and E. B. Santoso, “Analisis Preferensi Generasi Milenial Terhadap Tren Harga Hunian di Kota Surabaya Berdasarkan lokasi dan Tipologi,” *J. PENATAAN RUANG*, vol. 20, pp. 172–180, 2025, doi: https://doi.org/10.12962/j2716179X.v20iII.5326 Analisis.

[3] D. Chandraderia, V. N. Siwi, and S. Fevriera, “ANALISIS FAKTOR-FAKTOR YANG MEMPENGARUHI HARGA RUMAH DI AREA AGLOMERASI YOGYAKARTA,” *J. Pembang. Wil. dan Kota*, vol. 18, no. 2, pp. 128–139, 2022, doi: 10.14710/pwk.v18i2.37603.

[4] D. N. M. Cahyani *et al.*, “Comparison Of Decision Tree , Linear Regression , and Random Forest Regressor Models for Predicting House Prices,” *J. Ilm. MERPATI*, vol. 12, no. 1, pp. 62–71, 2024.

[5] M. B. S. Qolbi, T. N. Puteh, R. Rivandi, and C. Rozikin, “Prediksi Harga Rumah Di Jakarta Pusat Menggunakan Algoritma Machine Learning,” *J. Ilmu Komput. dan Bisnis*, vol. 16, no. 1, pp. 16–24, 2025, doi: 10.47927/jikb.v16i1.840.

[6] H. Sharma, H. Harsora, and B. Ogunleye, “An Optimal House Price Prediction Algorithm : XGBoost,” *Analytics*, vol. 3, pp. 30–45, 2024, doi: doi.org/10.3390/analytics3010003.

[7] P. Mahajan, S. Uddin, and F. Hajati, “Ensemble Learning for Disease Prediction : A Review,” *Healthcare*, vol. 11, p. 21, 2023, doi: https://doi.org/10.3390/healthcare11121808.

[8] U. Park, Y. Kang, H. Lee, and S. Yun, “A Stacking Heterogeneous Ensemble Learning Method for the Prediction of Building Construction Project Costs,” *Appl. Sci*, vol. 12, no. 9729, p. 11, 2022.

[9] M. N. Hibatulloh, G. D. Prakoso, A. D. Putri Yunus, and T. D. Putra, “Prediksi Harga Rumah di Bandung 2024 Menggunakan Ensemble Learning: Analisis Komparatif dan Interpretabilitas,” *J. Inform. J. Pengemb. IT*, vol. 10, no. 2, pp. 484–493, 2025, doi: 10.30591/jpit.v10i2.8200.

[10] R. Cellmer, A. Cichulska, and M. Bełej, “Spatial Analysis of Housing Prices and Market Activity with the Geographically Weighted Regression,” *Int. J. Geo-Information Artic.*, vol. 9, p. 19, 2020, doi: doi:10.3390/ijgi9060380.

DOI: 10.52362/jisamar.v10i3.2660

[11] L. Deng, "Real estate valuation with multi-source image fusion and enhanced machine learning pipeline," *PLoS One*, vol. 20, no. 5, pp. 1–32, 2025, doi: 10.1371/journal.pone.0321951.

[12] Z. Wang, Y. Wang, and S. Wu, "House Price Valuation Model Based on Geographically Neural Network Weighted Regression : The Case Study of," *Int. J. Geo-Information Artic.*, vol. 11, p. 25, 2022, doi: https://doi.org/10.3390/ijgi11080450.

[13] F. Pargent, F. Pfisterer, J. Thomas, and B. Bischl, "Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features," *Comput. Stat.*, vol. 37, no. 5, pp. 2671–2692, 2022, doi: 10.1007/s00180-022-01207-6.

[14] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost : unbiased boosting with categorical features," *Adv. Neural Inf. Process. Syst.*, vol. 31, pp. 6639–6649, 2018.

[15] A. Benavoli, G. Corani, and M. Zaffalon, "Time for a Change : a Tutorial for Comparing Multiple Classifiers Through Bayesian Analysis," *J. Mach. Learn. Res.*, vol. 18, no. 77, pp. 1–36, 2017.

[16] S. M. Lundberg and S. Lee, "A Unified Approach to Interpreting Model Predictions," *Adv. Neural Inf. Process. Syst.*, vol. 30, pp. 4766–4777, 2017.

DOI: 10.52362/jisamar.v10i3.2660