Journal of Innovative and Creativity
Journal Homepage: https://joecy.org/index.php/joecy
e-mail: journaljoecy@gmail.com

Faculty of Education
University of Pahlawan Tuanku Tambusai
ISSN 2776-771X (Online)
ISSN 2962-570X (Printed)
6 (2) 2026, Page: 32865-32873

Pemodelan Harga Properti Residensial di Jabodetabek Berbasis XGBoost-SHAP dan Risiko Banjir

Amara Thuridha[1], Dedi Trisnawarman[2]
[1-2]Fakultas Teknologi Informasi, Universitas Tarumanagara, Indonesia
Corresponding Author: Amara Thuridha, e-mail: amara.825220153@stu.untar.ac.id

Published: June, 2026

ABSTRACT
The Greater Jakarta (Jabodetabek) property market is characterized by profound heterogeneity, with prices driven by physical building attributes, land values, accessibility, neighborhood amenities, and environmental risks. This study develops an explainable property price prediction framework utilizing Extreme Gradient Boosting (XGBoost) and SHapley Additive exPlanations (SHAP). The dataset comprises 31,730 residential property listings enriched with sub-district-level assessed land values (NJOP), accessibility to train stations and toll gates, public amenities, and a 2025 news-based flood risk index. The flood risk index was constructed from an initial corpus of 3,810 flood-related news records, which were subsequently filtered to yield 1,470 unique news articles published in 2025 concerning the Greater Jakarta region. Following data preprocessing, 31,729 observations were retained for modeling. The XGBoost regression model yielded a cross-validated R^2 (CV-R^2) of 0.902 ± 0.004, a hold-out R^2 of 0.908, and a Mean Absolute Percentage Error (MAPE) of 25.33%. Furthermore, the price segment classification model achieved an accuracy of 87.30% with a weighted F1-score of 0.872. The SHAP interpretability analysis revealed that building area, assessed land value (NJOP), land area, number of bathrooms, and proximity to the city center serve as the dominant predictors. Although the flood risk index did not emerge as the strongest global predictor, it proved critical as a contextual indicator, given that flood exposure is highly concentrated in specific sub-districts. This study demonstrates that integrating multi-source data with explainable machine learning can substantially enhance risk-based property valuation in metropolitan areas.

Keywords: Property Price Prediction; XGBoost; SHAP; Flood Risk; Greater Jakarta.

ABSTRAK
Pasar properti Jabodetabek memiliki karakteristik heterogen karena harga dipengaruhi oleh atribut fisik bangunan, nilai tanah, aksesibilitas, fasilitas lingkungan, dan risiko lingkungan. Penelitian ini mengembangkan kerangka prediksi harga properti yang dapat dijelaskan dengan menggunakan Extreme Gradient Boosting (XGBoost) dan SHapley Additive exPlanations (SHAP). Data yang digunakan terdiri dari 31.730 listing properti residensial yang diperkaya dengan nilai NJOP tingkat kecamatan, akses stasiun kereta, akses gerbang tol, fasilitas publik, serta indeks risiko banjir berbasis berita tahun 2025. Indeks risiko banjir dibangun dari 3.810 rekaman berita banjir yang difilter menjadi 1.470 berita unik tahun 2025 pada wilayah Jabodetabek. Setelah pembersihan data, sebanyak 31.729 observasi digunakan untuk pemodelan. Model regresi XGBoost menghasilkan CV-R^2 sebesar 0,902±0,004, hold-out R^2 sebesar 0,908, dan MAPE sebesar 25,33%. Model klasifikasi segmen harga mencapai akurasi 87,30% dengan weighted F1-score sebesar 0,872. Hasil interpretasi menunjukkan bahwa luas bangunan, nilai NJOP tanah, luas tanah, jumlah kamar mandi, dan jarak ke pusat kota merupakan faktor dominan. Indeks risiko banjir tidak menjadi prediktor global terkuat, tetapi penting sebagai indikator kontekstual karena paparan banjir terkonsentrasi pada beberapa kecamatan. Penelitian ini menunjukkan bahwa integrasi data multi-sumber dan machine learning yang dapat dijelaskan dapat mendukung penilaian properti berbasis risiko di kawasan metropolitan.

Kata Kunci: Prediksi Harga Properti, XGBoost, SHAP, Risiko Banjir, Jabodetabek.

© The Author(s). 2021 Open Access This article is licensed under a Creative Commons Atibution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0

PENDAHULUAN
Jabodetabek merupakan kawasan metropolitan terintegrasi yang berfungsi sebagai satu pasar hunian dan komuter. Pembentukan harga properti residensial di kawasan ini tidak hanya dipengaruhi oleh karakteristik bangunan, tetapi juga oleh aksesibilitas, fasilitas lingkungan, nilai fiskal tanah, serta

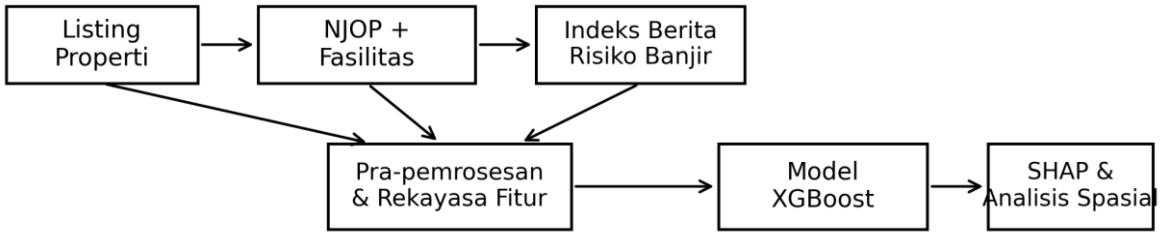
paparan risiko lingkungan. Kondisi ini membuat penilaian harga properti menjadi masalah komputasional yang kompleks karena hubungan antarfaktor bersifat tidak linear dan saling berinteraksi.

Penilaian properti berbasis data semakin penting karena proses appraisal manual membutuhkan waktu, biaya, dan sering kali sulit diterapkan pada pasar yang sangat heterogen. Model regresi linear masih bermanfaat untuk interpretasi awal, tetapi model tersebut memiliki keterbatasan dalam menangkap hubungan non-linear antara ukuran bangunan, lokasi, fasilitas publik, dan variabel risiko. Oleh karena itu, pendekatan machine learning seperti gradient boosting menjadi relevan untuk meningkatkan akurasi prediksi pada data tabular yang kompleks (Lorenz et al., 2023;Baur et al., 2023;Geerts et al., 2023;Deng, 2025;Sharma et al., 2024;Hasan et al., 2024;Wulan Sari & Prabowo, 2025). Selain faktor fisik dan lokasi, banjir merupakan salah satu risiko perkotaan utama di Jabodetabek. Banjir dapat memengaruhi akses mobilitas, biaya pemeliharaan bangunan, keputusan pembeli, dan persepsi nilai lingkungan. Namun, risiko banjir sering tidak tercatat secara langsung dalam listing properti. Penelitian ini membangun indeks risiko banjir berbasis berita tahun 2025 sebagai proxy paparan banjir yang dilaporkan media pada tingkat kecamatan (Tellman et al., 2021;Wing et al., 2022;Mutlu et al., 2023;Miller & Pinter, 2022;Islam et al., 2025;Seydi et al., 2022;Wang & Vivoni, 2022).

Tujuan penelitian ini adalah mengembangkan model prediksi harga properti berbasis XGBoost-SHAP dengan mengintegrasikan data listing properti, NJOP, aksesibilitas transportasi, fasilitas publik, dan indeks risiko banjir berbasis berita. Kontribusi utama penelitian ini adalah penyusunan pipeline data multi-sumber yang dapat dijelaskan serta analisis peran risiko banjir sebagai variabel kontekstual dalam prediksi harga properti (Lorenz et al., 2023;Sahakyan et al., 2021;Minh et al., 2022;Campbell et al., 2021;Boeing, 2025;Rey et al., 2023;Hadi & Benedict, 2024).

METODOLOGI PENELITIAN

Penelitian ini dilakukan melalui beberapa tahapan, yaitu pengumpulan data properti, integrasi data multi-sumber, rekayasa fitur, pemodelan XGBoost, evaluasi, serta interpretasi menggunakan SHAP. Alur penelitian disajikan pada Gambar 1 (Geerts et al., 2023;Boeing, 2025;Rey et al., 2023).



Gambar 1. Diagram alur penelitian

Pengumpulan dan Integrasi Data

Dataset utama terdiri dari 31.730 listing properti residensial di wilayah Jabodetabek. Setiap listing memuat atribut harga, luas bangunan, luas tanah, jumlah kamar tidur, jumlah kamar mandi, lokasi, dan tanggal publikasi. Setelah dilakukan pembersihan data dan penghapusan satu data dengan informasi kecamatan yang tidak lengkap, jumlah observasi yang digunakan dalam pemodelan adalah 31.729 observasi pada 191 kecamatan.

Data listing diperkaya dengan data NJOP per kecamatan, jarak ke stasiun kereta, jarak ke gerbang tol, jumlah fasilitas publik, skor fasilitas, indeks kriminalitas, dan indeks risiko banjir. Integrasi dilakukan pada tingkat kecamatan dan jarak spasial dihitung menggunakan koordinat kecamatan serta titik fasilitas terkait (Geerts et al., 2023;Boeing, 2025;Rey et al., 2023;Hadi & Benedict, 2024).

Tabel 1. Sumber data dan kunci integrasi

Dataset	Jumlah	Kunci Integrasi
Listing properti	31.730	Kecamatan
NJOP	191 area	Kecamatan
Berita banjir	1.470 berita	Kecamatan
Transportasi	152 titik	Jarak
Fasilitas publik	209 titik	Jarak

Rekayasa Fitur

Rekayasa fitur menghasilkan variabel struktural, spasial, aksesibilitas, fasilitas, risiko, dan fiskal. Variabel struktural meliputi luas bangunan, luas tanah, jumlah kamar, dan rasio bangunan terhadap tanah. Variabel spasial meliputi kota, kecamatan, koordinat, dan jarak ke pusat kota. Variabel aksesibilitas meliputi jarak ke stasiun kereta dan gerbang tol terdekat.

Variabel yang diturunkan langsung dari target harga, seperti harga per meter persegi, tidak digunakan dalam model regresi untuk menghindari kebocoran informasi. Variabel numerik yang memiliki distribusi miring ditransformasikan bila diperlukan, sedangkan variabel kategori dikodekan agar dapat diproses oleh model berbasis pohon keputusan.

Tabel 2. Kelompok fitur yang digunakan

Kelompok	Contoh	Tujuan
Struktural	Luas, kamar	Ukuran
Spasial	Kota, koordinat	Lokasi
Akses	Kereta, tol	Mobilitas
Fasilitas	Mall, pendidikan	Layanan
Risiko	Banjir, kriminalitas	Paparan
Fiskal	NJOP	Nilai tanah

Indeks Risiko Banjir Berbasis Berita

Indeks risiko banjir dibangun dari korpus berita banjir di wilayah Jabodetabek. Dataset awal berisi 3.810 rekaman berita. Data kemudian difilter hanya untuk tahun 2025 dan dideduplikasi berdasarkan kombinasi judul dan tanggal, sehingga diperoleh 1.470 berita unik. Untuk mengurangi kesalahan pemetaan lokasi, penyebutan level kota dipisahkan dari kecocokan level kecamatan (Tellman et al., 2021;Islam et al., 2025;Seydi et al., 2022;Wang & Vivoni, 2022).

Setiap kecocokan berita-lokasi diberi bobot berdasarkan kekuatan bukti. Lokasi eksplisit pada field lokasi diberi bobot 5, nama kecamatan yang muncul pada judul berita diberi bobot 3, sedangkan nama tempat yang hanya muncul pada query pencarian diberi bobot 2. Skor tiap kecamatan dijumlahkan dan dinormalisasi dengan transformasi log sehingga diperoleh indeks 0 sampai 1. Indeks ini tidak menggantikan peta hidrologis, tetapi digunakan sebagai proxy paparan banjir yang dilaporkan media.

Pemodelan dan Evaluasi

Penelitian ini membangun dua model XGBoost, yaitu model regresi untuk memprediksi harga properti dalam skala logaritmik dan model klasifikasi untuk memprediksi segmen harga. Data dibagi menggunakan skema hold-out 80:20 dan stabilitas model diuji menggunakan 5-fold cross-validation (Sharma et al., 2024;Sahakyan et al., 2021;Minh et al., 2022;Campbell et al., 2021).

Parameter utama yang digunakan adalah n_estimators = 500, max_depth = 8, learning_rate = 0,05, subsample = 0,85, colsample_bytree = 0,80, reg_alpha = 0,05, dan reg_lambda = 0,50. Kinerja regresi dievaluasi menggunakan R^2 dan MAPE, sedangkan kinerja klasifikasi dievaluasi menggunakan akurasi dan weighted F1-score. Interpretasi model dilakukan menggunakan SHAP.

Tabel 3. Ringkasan kinerja model

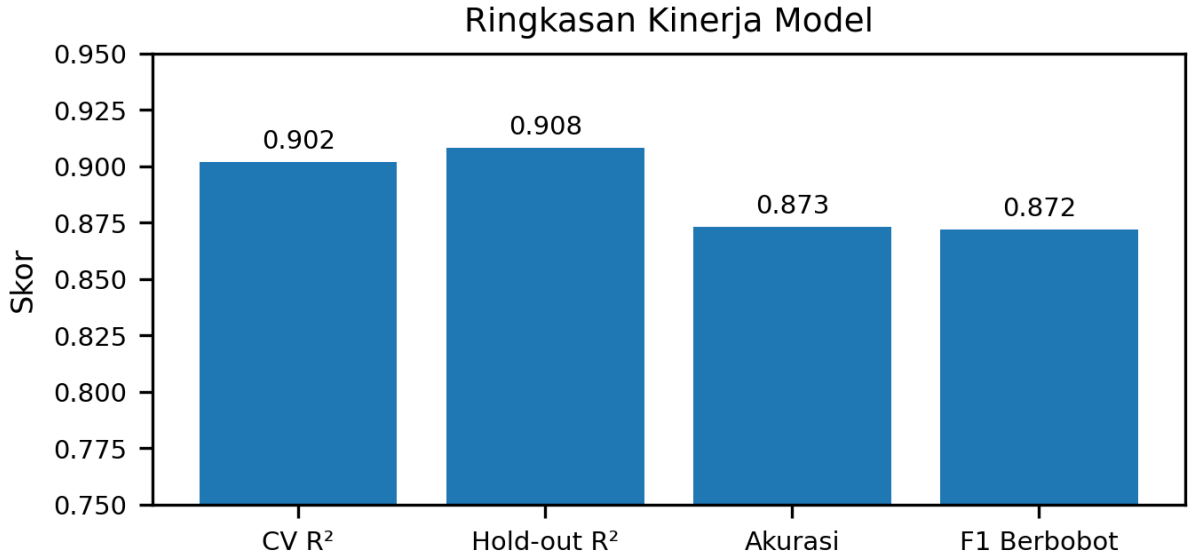
Tugas	Metrik	Nilai
Regresi	CV R^2	0,902±0,004
Regresi	Hold-out R^2	0,908
Regresi	MAPE	25,33%
Klasifikasi	Akurasi	87,30%
Klasifikasi	Weighted F1	0,872

HASIL PENELITIAN DAN PEMBAHASAN

Kinerja Model

Model regresi XGBoost menunjukkan kinerja yang stabil. Nilai CV-R^2 sebesar 0,902±0,004 menunjukkan bahwa model memiliki kemampuan generalisasi yang konsisten pada beberapa lipatan validasi. Nilai hold-out R^2 sebesar 0,908 menunjukkan bahwa model mampu menjelaskan sebagian besar variasi harga pada skala logaritmik. Nilai MAPE sebesar 25,33% masih dapat diterima pada pasar properti metropolitan yang memiliki variasi harga tinggi.

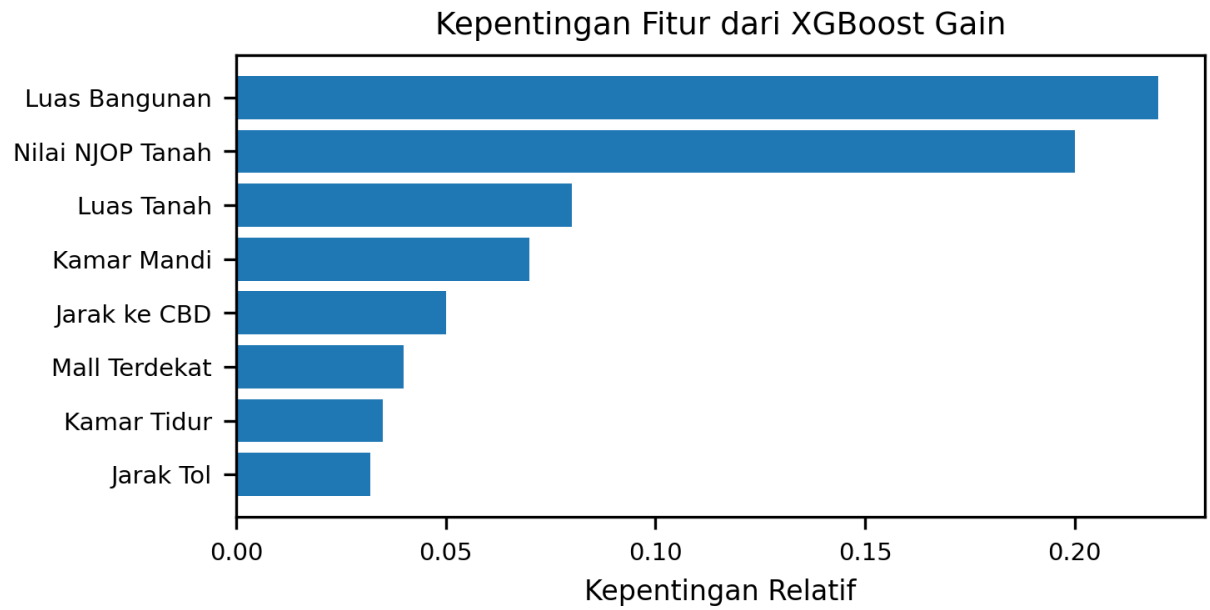
Model klasifikasi segmen harga juga menunjukkan performa yang baik dengan akurasi 87,30% dan weighted F1-score 0,872. Kesalahan klasifikasi paling mungkin terjadi pada batas antarsegmen harga karena kategori harga bersifat kontinu pada pasar nyata meskipun dibagi menjadi kelas diskrit untuk pemodelan.



Gambar 2. Ringkasan kinerja regresi dan klasifikasi

Analisis kepentingan fitur menunjukkan bahwa harga properti terutama dipengaruhi oleh atribut struktural dan nilai tanah. Luas bangunan dan nilai NJOP tanah menjadi dua prediktor paling kuat. Hasil ini konsisten dengan karakteristik pasar properti, yaitu kapasitas sfisik bangunan dan nilai lahan menjadi dasar utama pembentukan harga (Lorenz et al., 2023;Baur et al., 2023;Geerts et al., 2023;Deng, 2025;Wulan Sari & Prabowo, 2025;Hadi & Benedict, 2024).

Faktor aksesibilitas dan fasilitas berperan sebagai sinyal tambahan. Jusmlah mall terdekat, fasilitas pendidikan, dan jarak menuju tol memberikan informasi mengenai kenyamanan dan kualitas lingkungan. Dengan demikian, pendekatan multi-sumber lebih informatif dibandingkan model yang hanya menggunakan atribut listing properti (Boeing, 2025;Rey et al., 2023).

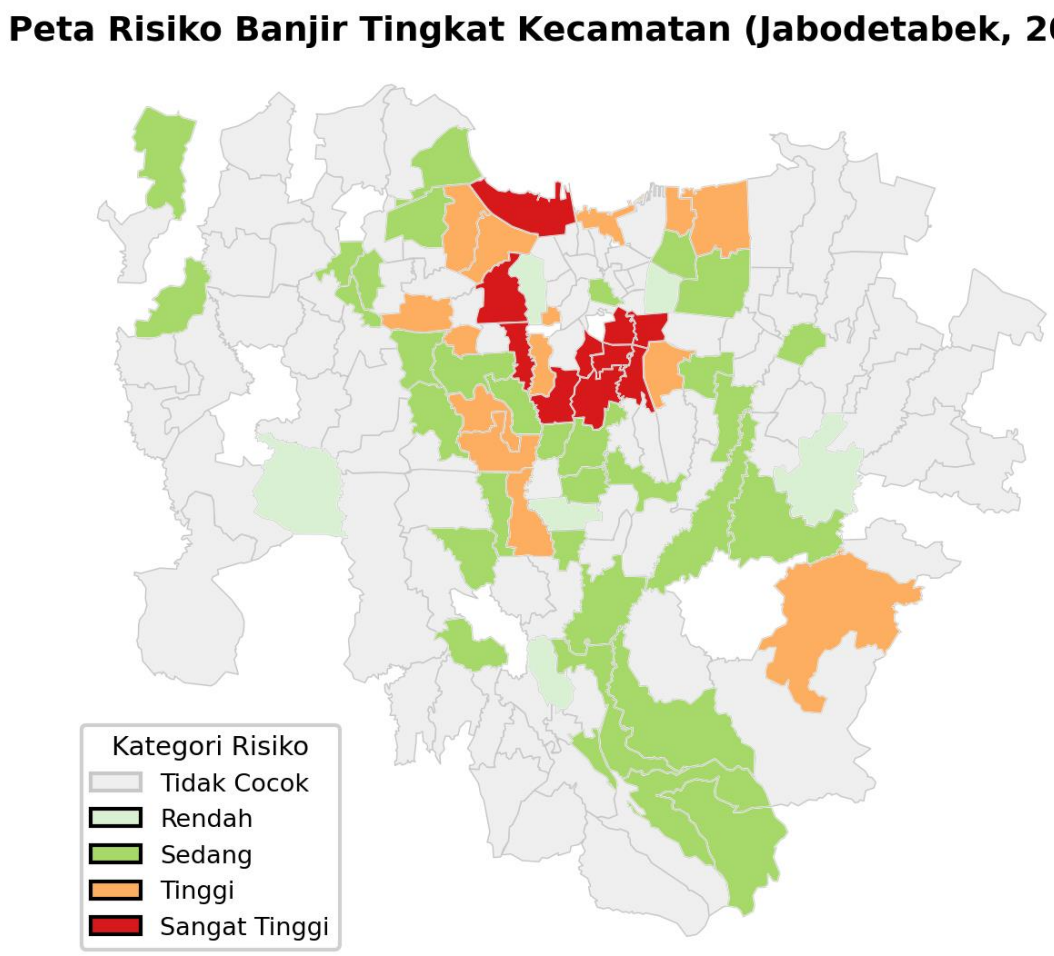


Gambar 3. Kepentingan fitur berdasarkan XGBoost gain

Hasil Indeks Risiko Banjir Tahun 2025

Variabel banjir berbasis berita bukan prediktor global terkuat dalam model, tetapi tetap penting secara spasial dan kebijakan. Indeks ini menangkap wilayah yang berulang kali muncul dalam pemberitaan banjir tahun 2025, bukan mengukur kedalaman air secara langsung. Oleh karena itu, variabel ini lebih tepat digunakan sebagai indikator lokal yang mendukung interpretasi risiko (Tellman et al., 2021;Wing et al., 2022;Mutlu et al., 2023;Miller & Pinter, 2022;Islam et al., 2025;Seydi et al., 2022;Wang & Vivoni, 2022).

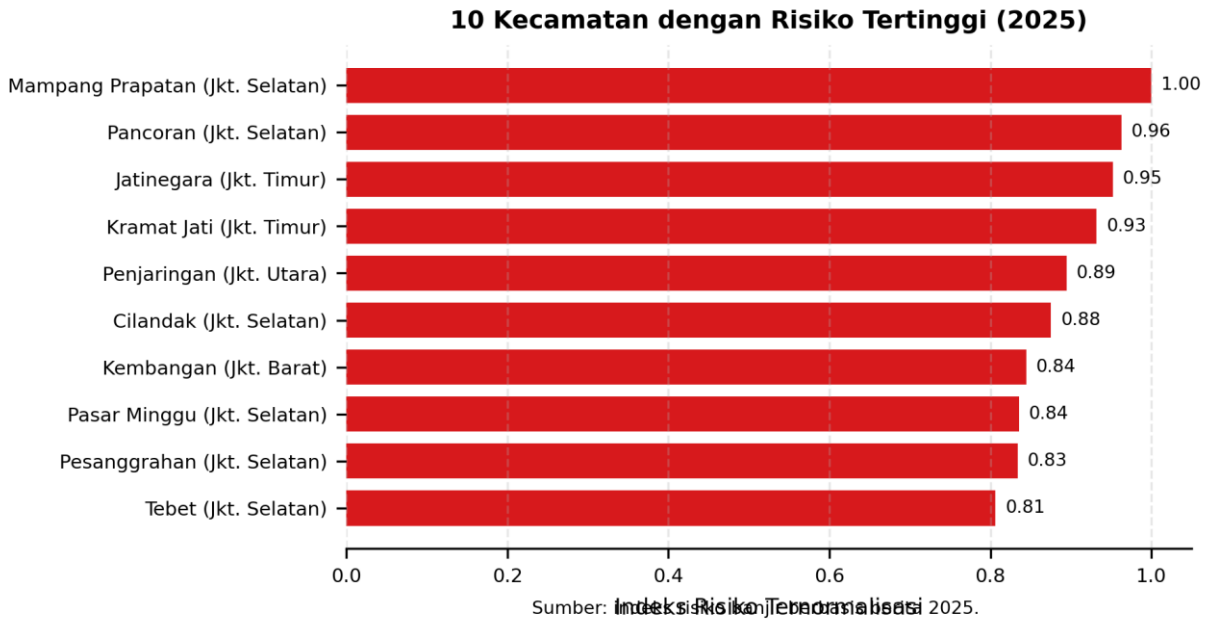
Peta pada **Gambar 4** menunjukkan bahwa paparan banjir tahun 2025 tidak tersebar merata. Konsentrasi tertinggi terlihat pada beberapa wilayah Jakarta Selatan dan Jakarta Timur, sedangkan konsentrasi sekunder muncul pada Jakarta Utara serta beberapa wilayah barat dan selatan pinggiran metropolitan. Gambar 5 melengkapi peta dengan menampilkan sepuluh kecamatan dengan indeks risiko tertinggi.



Sumber: pengolahan 3.810 rekaman berita banjir menjadi 1.470 berita unik tahun 2025.

Gambar 4. Peta indeks risiko banjir berbasis berita tahun 2025

Berdasarkan **Gambar 4**, wilayah dengan risiko sangat tinggi tampak terkonsentrasi pada koridor perkotaan utama. Hal ini menunjukkan bahwa risiko banjir perlu dipahami sebagai faktor lokasi yang bersifat kontekstual dan tidak dapat diberlakukan sebagai diskon harga seragam untuk seluruh wilayah Jabodetabek.



Gambar 5. Sepuluh kecamatan dengan indeks risiko banjir tertinggi

Gambar 5 menunjukkan bahwa Mampang Prapatan, Pancoran, Jatinegara, Kramat Jati, dan Penjaringan menempati posisi tertinggi. Informasi ini dapat digunakan untuk membantu pembeli, analis properti, dan pemerintah daerah dalam mengidentifikasi wilayah yang memerlukan perhatian lebih pada aspek mitigasi banjir.

Pembahasan penelitian ini menunjukkan bahwa model XGBoost-SHAP tidak hanya memberikan nilai prediksi, tetapi juga membantu menjelaskan faktor yang memengaruhi harga properti. Nilai NJOP yang kuat sebagai prediktor menunjukkan bahwa informasi fiskal tanah masih memiliki hubungan erat dengan dinamika pasar. Namun, NJOP belum sepenuhnya merepresentasikan variasi harga aktual karena harga properti juga dipengaruhi oleh aksesibilitas, fasilitas, karakteristik bangunan, dan risiko lingkungan. Oleh karena itu, model yang dapat dijelaskan dapat membantu pemerintah daerah, analis properti, dan calon pembeli memahami faktor pembentuk harga secara lebih transparan (Lorenz et al., 2023;Baur et al., 2023;Geerts et al., 2023;Deng, 2025;Sahakyan et al., 2021;Minh et al., 2022;Campbell et al., 2021).

Dari sisi risiko banjir, indeks berbasis berita tahun 2025 perlu dipahami sebagai indikator paparan yang bersifat kontekstual, bukan sebagai pengganti peta hidrologi resmi. Hasil ini tetap bermanfaat karena menunjukkan wilayah yang berulang kali muncul dalam pemberitaan banjir dan dapat menjadi sinyal awal bagi proses evaluasi agunan, keputusan asuransi, dan perencanaan mitigasi. Untuk platform properti, penyajian informasi aksesibilitas, fasilitas, dan kategori risiko banjir dapat mengurangi asimetri informasi sehingga pembeli dapat membandingkan properti secara lebih bertanggung jawab (Tellman et al., 2021;Wing et al., 2022;Mutlu et al., 2023;Miller & Pinter, 2022;Islam et al., 2025;Seydi et al., 2022;Wang & Vivoni, 2022).

KESIMPULAN

Penelitian ini membangun kerangka prediksi harga properti berbasis XGBoost-SHAP dengan mengintegrasikan data listing properti, NJOP, aksesibilitas, fasilitas, kriminalitas, dan indeks risiko banjir berbasis berita tahun 2025. Model regresi mencapai hold-out R^2 sebesar 0,908 dan MAPE sebesar 25,33%, sedangkan model klasifikasi mencapai akurasi 87,30% dan weighted F1-score sebesar 0,872. Hasil analisis menunjukkan bahwa luas bangunan, nilai NJOP tanah, luas tanah, jumlah kamar mandi, dan jarak ke pusat kota merupakan faktor paling berpengaruh terhadap prediksi harga. Indeks risiko banjir tidak menjadi prediktor global terkuat, tetapi tetap relevan karena paparan banjir bersifat

terkonsentrasi pada wilayah tertentu. Penelitian selanjutnya dapat menambahkan data transaksi aktual, data genangan resmi, data hidrologis, serta validasi spasial untuk memperkuat generalisasi model.

DAFTAR PUSTAKA

Baur, K., Rosenfelder, M., & Lutz, B. (2023). Automated real estate valuation with machine learning models using property descriptions. *Expert Systems with Applications*, *213*, 119147. https://doi.org/10.1016/j.eswa.2022.119147

Boeing, G. (2025). Modeling and Analyzing Urban Networks and Amenities With OSMnx. *Geographical Analysis*, *57*(4), 567–577. https://doi.org/10.1111/gean.70009

Campbell, T. W., Roder, H., Georgantas, R. W., & Roder, J. (2021). *Exact Shapley Values for Local and Model-True Explanations of Decision Tree Ensembles*. http://arxiv.org/abs/2112.10592

Deng, L. (2025). Real estate valuation with multi-source image fusion and enhanced machine learning pipeline. *PLOS One*, *20*(5), e0321951. https://doi.org/10.1371/journal.pone.0321951

Geerts, M., vanden Broucke, S., & De Weerdt, J. (2023). A Survey of Methods and Input Data Types for House Price Prediction. *ISPRS International Journal of Geo-Information*, *12*(5), 200. https://doi.org/10.3390/ijgi12050200

Hadi, N., & Benedict, J. (2024). Implementasi Machine Learning Untuk Prediksi Harga Rumah Menggunakan Algoritma Random Forest. *Computatio : Journal of Computer Science and Information Systems*, *8*(1), 50–61. https://doi.org/10.24912/computatio.v8i1.15173

Hasan, M. H., Jahan, M. A., Ali, M. E., Li, Y.-F., & Sellis, T. (2024). *A Multi-Modal Deep Learning Based Approach for House Price Prediction*. http://arxiv.org/abs/2409.05335

Islam, T., Zeleke, E. B., Afroz, M., & Melesse, A. M. (2025). A Systematic Review of Urban Flood Susceptibility Mapping: Remote Sensing, Machine Learning, and Other Modeling Approaches. *Remote Sensing*, *17*(3), 524. https://doi.org/10.3390/rs17030524

Lorenz, F., Willwersch, J., Cajias, M., & Fuerst, F. (2023). Interpretable machine learning for real estate market analysis. *Real Estate Economics*, *51*(5), 1178–1208. https://doi.org/10.1111/1540-6229.12397

Miller, R. G., & Pinter, N. (2022). Flood risk and residential <scp>real-estate</scp> prices: Evidence from three <scp>US</scp> counties. *Journal of Flood Risk Management*, *15*(2), 1–12. https://doi.org/10.1111/jfr3.12774

Minh, D., Wang, H. X., Li, Y. F., & Nguyen, T. N. (2022). Explainable artificial intelligence: a comprehensive review. *Artificial Intelligence Review*, *55*(5), 3503–3568. https://doi.org/10.1007/s10462-021-10088-y

Mutlu, A., Roy, D., & Filatova, T. (2023). Capitalized value of evolving flood risks discount and nature-based solution premiums on property prices. *Ecological Economics*, *205*, 107682. https://doi.org/10.1016/j.ecolecon.2022.107682

Rey, S., Arribas-Bel, D., & Wolf, L. J. (2023). *Geographic Data Science with Python*. Chapman and Hall/CRC. https://doi.org/10.1201/9780429292507

Sahakyan, M., Aung, Z., & Rahwan, T. (2021). Explainable Artificial Intelligence for Tabular Data: A Survey. *IEEE Access*, *9*, 135392–135422. https://doi.org/10.1109/ACCESS.2021.3116481

Seydi, S. T., Kanani-Sadat, Y., Hasanlou, M., Sahraei, R., Chanussot, J., & Amani, M. (2022).

Comparison of Machine Learning Algorithms for Flood Susceptibility Mapping. *Remote Sensing*, *15*(1), 192. https://doi.org/10.3390/rs15010192

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

Tellman, B., Sullivan, J. A., Kuhn, C., Kettner, A. J., Doyle, C. S., Brakenridge, G. R., Erickson, T. A., & Slayback, D. A. (2021). Satellite imaging reveals increased proportion of population exposed to floods. *Nature*, *596*(7870), 80–86. https://doi.org/10.1038/s41586-021-03695-w

Wang, Z., & Vivoni, E. R. (2022). Mapping Flash Flood Hazards in Arid Regions Using CubeSats. *Remote Sensing*, *14*(17), 4218. https://doi.org/10.3390/rs14174218

Wing, O. E. J., Lehman, W., Bates, P. D., Sampson, C. C., Quinn, N., Smith, A. M., Neal, J. C., Porter, J. R., & Kousky, C. (2022). Inequitable patterns of US flood risk in the Anthropocene. *Nature Climate Change*, *12*(2), 156–162. https://doi.org/10.1038/s41558-021-01265-6

Wulan Sari, B., & Prabowo, D. (2025). Analisis Perbandingan Prediksi Harga Rumah Dengan Random Forest, Gradient Boosting, dan XGBoost. *Intellect : Indonesian Journal of Learning and Technological Innovation*, *4*(1), 42–51. https://doi.org/10.57255/intellect.v4i1.1385