

Turnitin LLCC

2213030125_Muh. Junaidi

 SI-12

Document Details

Submission ID

trn:oid:::2945:392447655

Submission Date

Jun 10, 2026, 1:38 PM GMT+7

Download Date

Jun 10, 2026, 1:40 PM GMT+7

File Name

2213030125_Muh. Junaidi.pdf

File Size

690.4 KB

42 Pages

8,245 Words

54,404 Characters

28% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text
- Cited Text
- Small Matches (less than 8 words)

Top Sources

- 17%  Internet sources
- 14%  Publications
- 24%  Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Top Sources

- 17% Internet sources
- 14% Publications
- 24% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Internet	kc.umn.ac.id	1%
2	Internet	repository.itpln.ac.id	<1%
3	Internet	ejurnal.seminar-id.com	<1%
4	Student papers	Universitas Muhammadiyah Sukabumi on 2026-01-13	<1%
5	Internet	pustakagalerimandiri.co.id	<1%
6	Student papers	Universitas Islam Sumatera Utara on 2026-04-20	<1%
7	Student papers	Telkom University on 2025-07-04	<1%
8	Student papers	UPN Veteran Yogyakarta on 2026-05-04	<1%
9	Student papers	Binus University International on 2026-03-17	<1%
10	Internet	ejurnal.stmik-budidarma.ac.id	<1%
11	Student papers	Virginia Polytechnic Institute and State University on 2025-04-10	<1%

12	Publication	Laila Salsabilla Hanifa, Herlawati Herlawati, Rakhmat Purnomo. "Perbandingan K...	<1%
13	Internet	repository.umsu.ac.id	<1%
14	Internet	repository.iainpalopo.ac.id	<1%
15	Publication	Hasan Ibrohim, Harminto Mulyo, Gentur Wahyu Nyipto Wibowo. "Pemodelan Hyb...	<1%
16	Student papers	Sekolah Teknik Elektro & Informatika on 2025-07-18	<1%
17	Student papers	UPN Veteran Yogyakarta on 2026-05-29	<1%
18	Publication	Yusuf Deardo Purba, Daniel Simamora, Kristianos Zamili, Finky Notavianus. "Pene...	<1%
19	Internet	adoc.pub	<1%
20	Internet	digilibadmin.unismuh.ac.id	<1%
21	Student papers	Sriwijaya University on 2026-04-28	<1%
22	Student papers	UPN Veteran Yogyakarta on 2026-05-05	<1%
23	Student papers	Universitas Budi Luhur on 2026-01-31	<1%
24	Student papers	Universitas Indonesia on 2025-03-20	<1%
25	Student papers	Universitas Muhammadiyah Sukabumi on 2026-01-20	<1%

26	Student papers	Universitas Muhammadiyah Sukabumi on 2026-01-23	<1%
27	Internet	onepd.rokania.ac.id	<1%
28	Internet	journal.universitasmulia.ac.id	<1%
29	Internet	repository.ipb.ac.id	<1%
30	Publication	Alraifful Rahman, Valentinus Paramarta, Agnes Novita Ida, Mohammad Harits Ak...	<1%
31	Student papers	Telkom University on 2026-04-15	<1%
32	Student papers	Universitas Bengkulu on 2025-10-27	<1%
33	Publication	Hasriadi Hasriadi, Iqbal Iqbal. "Prediksi Harga Bitcoin Multivariat OHLC Berbasis ...	<1%
34	Publication	Rizky Dwi Prasetyo, Nadia Anissa Maori, Akhmad Khanif Zyen. "Pengaruh Optima...	<1%
35	Internet	etheses.uin-malang.ac.id	<1%
36	Student papers	School of Business and Management ITB on 2025-02-28	<1%
37	Student papers	UM Surabaya on 2025-09-12	<1%
38	Student papers	Universitas Budi Luhur on 2025-08-02	<1%
39	Student papers	Universitas Pamulang on 2026-05-04	<1%

40	Internet	docobook.com	<1%
41	Internet	repository.its.ac.id	<1%
42	Student papers	Sekolah Teknik Elektro & Informatika on 2025-08-20	<1%
43	Student papers	Universitas Khairun on 2026-04-27	<1%
44	Student papers	Universitas Muhammadiyah Jember on 2026-03-31	<1%
45	Student papers	Universitas Pancasila on 2026-03-04	<1%
46	Internet	ahmaddahlan.ac.id	<1%
47	Publication	Harliana Harliana, Tito Prabowo, Ady Alzhava Nuary. "Perbandingan Logistic Reg...	<1%
48	Internet	repository.unp.ac.id	<1%
49	Publication	Nor Riduan, Alva Hendi Muhammad. "Optimasi Model XGBoost dengan Genetic A...	<1%
50	Student papers	Telkom University on 2026-01-23	<1%
51	Student papers	Universitas Jember on 2025-10-06	<1%
52	Internet	ipm2kpe.or.id	<1%
53	Internet	jurnal.unprimdn.ac.id	<1%

54	Publication	Chaerul Hidayat, Agung Nugroho, Asep Suprianto. "Prediksi Risiko Stroke Berdas...	<1%
55	Student papers	LL DIKTI IX Turnitin Consortium Part III on 2026-05-25	<1%
56	Student papers	Universitas Dian Nuswantoro on 2025-12-05	<1%
57	Student papers	Universitas Islam Riau on 2025-07-18	<1%
58	Student papers	Universitas Muhammadiyah Palembang on 2026-05-09	<1%
59	Student papers	Universitas Negeri Jakarta on 2016-10-08	<1%
60	Student papers	Universitas Khairun on 2026-05-25	<1%
61	Internet	ioinformatic.org	<1%
62	Student papers	itera on 2024-11-07	<1%
63	Publication	Muhammad Luthfan Hawali, Walid Walid. "Pemilihan Neuron LSTM dan LSTM Bay...	<1%
64	Student papers	Perpustakaan on 2025-10-17	<1%
65	Publication	Syawaludin Author, Antonius Bagas Sunu W.A Author. "PERBANDINGAN MODEL E...	<1%
66	Student papers	Universitas Katolik Musi Charitas on 2026-02-09	<1%
67	Student papers	Universitas Katolik Widya Mandala Surabaya on 2025-12-15	<1%

68	Student papers	Universitas Khairun on 2026-05-20	<1%
69	Student papers	Universitas Pamulang on 2026-04-21	<1%
70	Internet	eprints.uny.ac.id	<1%
71	Internet	ojs.feb.uajm.ac.id	<1%
72	Internet	prima.uinbanten.ac.id	<1%
73	Internet	repository.uinjkt.ac.id	<1%
74	Internet	senter.ee.uinsgd.ac.id	<1%
75	Publication	Anis Athifah, Herlawati Herlawati, Fata Nidaul Khasanah. "Evaluasi Kinerja Cluste...	<1%
76	Publication	Evelyn Agustin, Ferdy Setiawan, Andri Wijaya. "Pengembangan Model Prediksi Pe...	<1%
77	Student papers	Fakultas Teknologi Industri on 2026-05-15	<1%
78	Student papers	UPN Veteran Yogyakarta on 2026-02-25	<1%
79	Student papers	UPN Veteran Yogyakarta on 2026-05-04	<1%
80	Student papers	Universitas Budi Luhur on 2026-04-01	<1%
81	Student papers	Universitas Khairun on 2026-05-18	<1%

82	Student papers	Universitas Muhammadiyah Sukabumi on 2026-01-26	<1%
83	Student papers	Universitas Pamulang on 2026-05-20	<1%
84	Student papers	Universitas Pancasila on 2026-02-21	<1%
85	Student papers	Universitas Putera Batam on 2026-02-03	<1%
86	Student papers	Universitas Sebelas Maret on 2018-12-19	<1%
87	Student papers	Universitas Sumatera Utara on 2024-11-19	<1%
88	Student papers	Universitas Sumatera Utara on 2026-04-15	<1%
89	Publication	Vivin Mahat Putri, M Syafrizal Zain, Satria Agus Darma. "Prediksi Harga Saham M...	<1%
90	Internet	dokumen.pub	<1%
91	Internet	eprints.ukmc.ac.id	<1%
92	Internet	ijircce.com	<1%
93	Internet	media.neliti.com	<1%
94	Internet	repository.ubt.ac.id	<1%
95	Internet	repository.upnjatim.ac.id	<1%

96	Publication	Mitra Unik, M. Eddo Aldo Fahrurrozi, Harmaini Harmaini, Chandra Kusuma. "Impl...	<1%
97	Publication	Muhamad Rizki Fauzi, Muhamad Handika, Alfian Awinanto, Ahmad Jurnaidi Wahi...	<1%
98	Publication	Rahmi Rahmi Burhanuddin. "KLASIFIKASI PENYAKIT PADI MELALUI CITRA DAUN ...	<1%
99	Student papers	Telkom University on 2026-01-21	<1%
100	Student papers	UIN Raden Intan Lampung on 2026-02-12	<1%
101	Student papers	UNIVERSITAS BUDI LUHUR on 2026-01-27	<1%
102	Student papers	Universitas Gadjah Mada on 2026-06-02	<1%
103	Student papers	Universitas Islam Negeri Sumatera Utara on 2026-06-06	<1%
104	Student papers	Universitas Islam Riau on 2025-10-23	<1%
105	Student papers	Universitas Muhammadiyah Palembang on 2026-04-20	<1%
106	Student papers	Universitas Muhammadiyah Purwokerto on 2025-11-13	<1%
107	Student papers	Universitas Pamulang on 2025-07-10	<1%
108	Student papers	Universitas Sangga Buana YPKP on 2026-02-23	<1%
109	Student papers	Universitas Tarumanagara on 2025-12-22	<1%

110	Student papers	Universitas Tarumanagara on 2025-12-23	<1%
111	Student papers	Universiteit Hasselt on 2024-05-14	<1%
112	Internet	e-jurnal.lppmunsera.org	<1%
113	Internet	ejournal.publine.or.id	<1%
114	Internet	id.nnetswitch.com	<1%
115	Student papers	itera on 2025-08-22	<1%
116	Internet	jurnal.stkippgritulungagung.ac.id	<1%
117	Internet	nanopdf.com	<1%
118	Internet	repository.dinamika.ac.id	<1%
119	Internet	www.j-cup.org	<1%
120	Internet	123dok.com	<1%
121	Publication	Elsyabeth Trivosa Casandra, Davin Nayaka, Hendrata Wibisana. "Analisa Kadar ...	<1%
122	Student papers	Fakultas Ekonomi Universitas Indonesia on 2015-11-23	<1%
123	Student papers	Fakultas Teknik on 2025-08-05	<1%

124	Student papers	Institut Teknologi PLN on 2025-09-02	<1%
125	Student papers	Islington College,Nepal on 2026-05-11	<1%
126	Publication	Melati Sinta Nurdanita, Azizah Azizah. "Estimasi Aggregate Loss Menggunakan P...	<1%
127	Publication	Muhammad Zhafir, Novi Rukhiviyanti. "Analisis Perbandingan Algoritma Machine...	<1%
128	Publication	Roni Habibi, Darfial Guslan. "RISIKO MAGANG MAHASISWA DENGAN PENDEKATA...	<1%
129	Publication	Roni Saputra. "A Comparative Analysis of Univariate and Multivariate LSTM Mode...	<1%
130	Student papers	STIE Mahardhika on 2025-07-24	<1%
131	Student papers	STT PLN on 2023-08-12	<1%
132	Student papers	Sekolah Teknik Elektro & Informatika on 2025-09-10	<1%
133	Publication	Siti Andini Utiahman, Andi Mulawati Mas Pratama. "Penerapan Support Vector ...	<1%
134	Publication	Suheli Suheli. "Manajemen Peserta Didik Berbasis Pesantren dalam Pembentuka...	<1%
135	Student papers	Telkom University on 2026-03-27	<1%
136	Student papers	UNIVERSITAS BUDI LUHUR on 2025-07-24	<1%
137	Student papers	UPN Veteran Yogyakarta on 2026-05-04	<1%

138	Student papers	UPN Veteran Yogyakarta on 2026-05-05	<1%
139	Student papers	Universitas Islam Indonesia on 2025-07-30	<1%
140	Student papers	Universitas Khairun on 2026-06-08	<1%
141	Student papers	Universitas Kristen Duta Wacana on 2026-05-07	<1%
142	Student papers	Universitas Kristen Satya Wacana on 2025-12-08	<1%
143	Student papers	Universitas Mohammad Husni Thamrin Jakarta on 2025-08-25	<1%
144	Student papers	Universitas Muhammadiyah Sukabumi on 2025-05-19	<1%
145	Student papers	Universitas Pendidikan Ganesha on 2024-11-21	<1%
146	Student papers	Universitas Pendidikan Indonesia on 2025-08-05	<1%
147	Student papers	Universitas Putera Batam on 2026-01-29	<1%
148	Student papers	Universitas Putera Batam on 2026-02-03	<1%
149	Student papers	Universitas Sumatera Utara on 2024-11-19	<1%
150	Student papers	Universitas Tarumanagara on 2025-12-19	<1%
151	Publication	Zaehol Fatah, Moch. Hegal Muktasim Billah. "Prediksi Nilai Akhir Siswa SMK Nuru..."	<1%

152	Student papers	iGroup on 2019-07-30	<1%
153	Internet	jurnal.larisma.or.id	<1%
154	Internet	lib.unnes.ac.id	<1%
155	Internet	text-id.123dok.com	<1%
156	Internet	wogoheso.tumblr.com	<1%
157	Internet	www.mahkotaproperty.com	<1%
158	Internet	www.majalahpangan.com	<1%
159	Internet	www.researchgate.net	<1%

14

BAB I

PENDAHULUAN

A. Latar Belakang

158
89
49
Harga rumah merupakan salah satu indikator penting dalam sektor properti yang memengaruhi berbagai keputusan, seperti investasi, pembelian, penjualan, hingga perencanaan pembangunan. Penentuan harga rumah merupakan permasalahan yang kompleks karena melibatkan berbagai faktor non-linear seperti kualitas bangunan, luas area, dan lokasi, sehingga metode konvensional sering kali kurang akurat (Balqis et al., 2026). Dengan adanya prediksi harga yang akurat, berbagai pihak seperti pembeli, penjual, pengembang, dan investor dapat membuat keputusan secara lebih rasional dan terukur. Oleh karena itu, pengembangan sistem prediksi harga rumah yang akurat menjadi kebutuhan yang sangat penting di era data saat ini (Li, 2023). Penerapan *data mining* menjadi fokus utama dalam analisis perbandingan algoritma untuk memprediksi suatu nilai atau kategori berdasarkan fitur-fitur yang tersedia (Kurnia et al., 2024).

2
127
90
26
156
Perkembangan teknologi di bidang *machine learning* telah membuka peluang untuk melakukan prediksi harga rumah secara otomatis dengan memanfaatkan data historis dan berbagai atribut properti. Dalam perkembangannya, algoritma berbasis *ensemble learning* seperti *Random Forest* dan *Extreme Gradient Boosting (XGBoost)* telah menjadi standar komputasi yang unggul dalam memprediksi nilai maupun harga sewa aset properti karena fleksibilitasnya dalam memproses fitur non-linear (Yantono et al., 2026). Beberapa algoritma yang umum digunakan dalam penelitian prediksi harga rumah meliputi *Linear Regression*, *K-Nearest Neighbor (KNN)*, *Artificial Neural Network (ANN)*, *Random Forest*, hingga metode *ensemble* lanjutan seperti *XGBoost*. Algoritma tersebut mampu mempelajari hubungan linear maupun *non-linear* yang kompleks antara fitur-fitur rumah, seperti lokasi, luas bangunan, jumlah kamar, dan variabel lainnya terhadap harga jual rumah (Hajiyianni & Stylianou, 2025). Oleh karena itu, sangat penting untuk mengembangkan strategi

yang tepat untuk melakukan prediksi atau peramalan yang akurat (Fitriyono et al., 2024).

Pada penelitian terdahulu yang menjadi jurnal acuan, telah dilakukan perbandingan beberapa algoritma *machine learning* dan *deep learning* untuk memprediksi harga rumah, di antaranya *ANN*, *K-Nearest Neighbor Regression*, *Linear Regression*, dan *Random Forest Regressor*. Selanjutnya, model *Random Forest* dioptimalkan menggunakan metode *Optuna Hyperparameter Tuning* dengan pendekatan *Tree-structured Parzen Estimator* (TPE) untuk meningkatkan akurasi. Hasil penelitian tersebut menunjukkan bahwa *Random Forest* yang telah dioptimalkan menghasilkan performa yang lebih baik dibandingkan model lainnya, dengan peningkatan nilai R^2 dan akurasi yang cukup signifikan (Fadzryan et al., 2025).

Meskipun demikian, hasil prediksi yang dihasilkan model tersebut belum sepenuhnya sempurna. Dalam jurnal tersebut masih ditemukan adanya selisih prediksi yang cukup besar pada beberapa data tertentu. Selisih yang cukup besar ini menunjukkan bahwa, meskipun model telah memiliki tingkat akurasi yang tinggi secara keseluruhan, masih terdapat kelemahan dalam memprediksi beberapa data, khususnya pada rumah dengan karakteristik yang lebih kompleks atau berada pada rentang harga tertentu (Hajiyianni & Stylianou, 2025).

Kondisi tersebut menunjukkan adanya kesenjangan antara harapan dan kenyataan. Secara ideal, model prediksi diharapkan mampu memberikan hasil yang mendekati harga aktual untuk seluruh data. Namun pada praktiknya, masih terdapat kasus *underprediction* maupun *overprediction* yang cukup signifikan. Hal ini dapat disebabkan oleh beberapa faktor, seperti distribusi data yang tidak merata (*skewed distribution*), keberadaan fitur yang kurang relevan atau belum optimal, serta penggunaan *hyperparameter* model yang belum berada pada konfigurasi terbaik (Sharma et al., 2024). Data yang tidak seimbang (*imbalanced data*) dapat terjadi dan memengaruhi hasil prediksi secara signifikan (Nugroho & Harini, 2024). Oleh karena itu, penggunaan *Grid Search CV* untuk *hyperparameter optimization* merupakan

langkah penting yang terbukti dapat meningkatkan akurasi model (Azzaria et al., 2025).

Selain itu, data harga rumah biasanya memiliki kombinasi fitur numerik dan kategorikal. Jika proses pra-pemrosesan tidak dilakukan dengan tepat, seperti penghapusan kolom yang tidak relevan, kesalahan dalam pengolahan data kategorikal, maupun tidak dilakukannya transformasi data yang diperlukan, maka model akan mengalami kesulitan dalam mempelajari pola sebenarnya yang terdapat di dalam data. Oleh karena itu, penelitian harus melakukan beberapa eksperimen pada pengaturan parameter algoritma untuk mencapai hasil yang lebih optimal (Amarya et al., 2025) dan dibutuhkan suatu pendekatan yang lebih sistematis dalam menyiapkan data dan membangun model prediksi (Hajjiyanni & Stylianou, 2025).

Berdasarkan permasalahan tersebut, dalam penelitian ini digunakan pendekatan yang sedikit berbeda dari jurnal acuan. Jika jurnal sebelumnya menggunakan Random Forest yang dioptimalkan dengan *Optuna*, maka dalam penelitian ini digunakan *XGBoost Regressor* yang dikenal memiliki performa tinggi pada tugas regresi dan mampu menangani hubungan non-linear secara lebih efektif. Selain itu, dilakukan transformasi logaritmik pada variabel target (harga) untuk mengatasi kemiringan data, serta diterapkan *ColumnTransformer* dengan *OneHotEncoder* untuk mengolah fitur kategorikal. Selanjutnya, dilakukan proses *hyperparameter tuning* menggunakan *GridSearchCV* untuk mendapatkan kombinasi parameter terbaik.

Dengan pendekatan tersebut, penelitian ini diharapkan mampu meningkatkan akurasi prediksi harga rumah dan mengurangi selisih yang signifikan antara harga aktual dan harga prediksi, sebagaimana yang masih ditemukan pada penelitian sebelumnya. Penelitian ini pada akhirnya diarahkan untuk menghasilkan model prediksi harga rumah yang lebih akurat, stabil, dan dapat diandalkan sebagai alat bantu pengambilan keputusan di bidang properti.

Selain itu, untuk meningkatkan validitas dan relevansi penelitian ini tidak hanya mengadopsi metode yang digunakan, tetapi juga mengimplementasikannya pada dataset “*Jakarta House Price Dataset*” sebagai representasi kondisi pasar properti di Indonesia. Langkah ini bertujuan untuk membandingkan performa metode yang

diusulkan dengan metode sebelumnya, sekaligus menguji kemampuan generalisasi model dalam konteks data properti di Indonesia. Dengan demikian, penelitian ini tidak hanya berfokus pada peningkatan performa model, tetapi juga pada evaluasi efektivitas metode dalam menghadapi karakteristik data properti yang berbeda dari penelitian sebelumnya.

B. Batasan Masalah

Agar penelitian ini lebih terarah dan tidak meluas, maka diberikan batasan masalah sebagai berikut:

1. Penelitian ini difokuskan pada upaya meningkatkan akurasi prediksi harga rumah dan meminimalkan selisih antara harga aktual dan harga hasil prediksi melalui optimasi model pembelajaran mesin.
2. Objek dalam penelitian ini adalah data harga rumah beserta variabel-variabel yang memengaruhinya, yang digunakan sebagai dasar dalam proses pelatihan dan pengujian model prediksi.
3. Model yang digunakan dalam penelitian ini dibatasi pada algoritma *XGBoost Regressor* yang diintegrasikan dalam *pipeline* dengan tahap *preprocessing* berupa transformasi logaritmik pada variabel target dan *One Hot Encoding* menggunakan *ColumnTransformer*, serta dilakukan optimasi menggunakan *GridSearchCV*.
4. Dataset yang digunakan dalam penelitian ini adalah “*Jakarta House Price Dataset*”. Dataset “*House Price 2023 Dataset*” hanya digunakan sebagai referensi dalam studi literatur dan bukan sebagai data eksperimen..

C. Rumusan Masalah

Berdasarkan identifikasi dan batasan masalah, maka rumusan masalah pada penelitian ini adalah:

1. Seberapa besar peningkatan akurasi model prediksi harga rumah setelah dilakukan *hyperparameter tuning* menggunakan *GridSearchCV* pada algoritma *XGBoost Regressor*?

2. Seberapa besar penurunan selisih antara harga aktual dan harga hasil prediksi setelah model *XGBoost Regressor* dilakukan optimasi menggunakan *GridSearchCV*?

D. Tujuan Penelitian

Tujuan dari penelitian ini adalah:

1. Mengetahui dan menganalisis secara kuantitatif tingkat peningkatan akurasi model prediksi harga rumah setelah dilakukan proses hyperparameter tuning menggunakan metode *GridSearchCV* pada algoritma *XGBoost Regressor*, yang diukur melalui nilai metrik evaluasi seperti R^2 , *MAE*, *MSE*, dan *RMSE*.
2. Mengetahui dan mengevaluasi besarnya penurunan selisih antara nilai harga rumah aktual dan nilai harga hasil prediksi setelah model *XGBoost Regressor* dilakukan proses optimasi menggunakan *GridSearchCV*, sebagai indikator keberhasilan model dalam menghasilkan prediksi yang lebih mendekati nilai sebenarnya.

E. Manfaat Penelitian

Berdasarkan uraian diatas berikut adalah manfaat dari penelitian ini:

1. Memberikan model prediksi harga rumah yang lebih akurat, dengan selisih antara harga aktual dan hasil prediksi yang lebih kecil sehingga penentuan harga jual/beli menjadi lebih tepat dan rasional bagi pembeli, penjual, pengembang, agen, dan investor.
2. Mendukung pengambilan keputusan berbasis data di bidang properti dan investasi, karena model dapat digunakan sebagai alat bantu untuk menilai potensi nilai properti dan menyusun strategi pengembangan atau penjualan yang lebih efektif.
3. Berkontribusi pada pengembangan keilmuan machine learning, khususnya penerapan *XGBoost Regressor* dan *GridSearchCV*, serta menegaskan pentingnya optimasi *hyperparameter* dan preprocessing data dalam membangun model regresi yang optimal dan andal.

86

BAB II

KAJIAN TEORI

A. Kajian Teori

1. Prediksi Harga Rumah dan Machine Learning

a. Prediksi Harga Rumah (*House Price Prediction*)

Prediksi harga properti merupakan aspek krusial dalam pengambilan keputusan investasi di sektor *real estate* (Fadzryan et al., 2025). Sektor *real estate* berfungsi sebagai barometer penting kinerja ekonomi suatu bangsa, namun harga properti sering kali dipengaruhi oleh mekanisme penetapan harga yang rumit dan volatilitas yang melekat. Prediksi harga rumah juga melibatkan penemuan pola rumit di berbagai variabel data (Utomo et al., 2024).

b. Peran Machine Learning dalam Regresi

Machine Learning (Pembelajaran Mesin) telah muncul sebagai alternatif yang menjanjikan untuk penilaian properti, menawarkan analisis otomatis yang mampu mengenali pola rumit di berbagai variable (Utomo et al., 2024). Metode ini telah banyak digunakan untuk prediksi di berbagai sektor, termasuk property (Wijaya et al., 2024). Dalam konteks prediksi harga rumah, yang merupakan masalah regresi, model *Machine Learning* digunakan untuk memprediksi nilai numerik kontinu (harga) berdasarkan fitur-fitur property (Fadzryan et al., 2025). Berbagai algoritma machine learning seperti XGBoost, LightGBM, dan CatBoost telah banyak digunakan dalam prediksi harga rumah karena mampu menghasilkan performa regresi yang baik pada data properti multivariat (Wang & Zhao, 2022).

Pada ranah medis, penerapan *XGBoost* yang dioptimasi dengan *GridSearchCV* untuk penentuan kombinasi parameter seperti `n_estimators`, `learning_rate`, dan `max_depth` berhasil mencapai akurasi sekitar 93% dan nilai AUC tinggi pada klasifikasi tingkat hipertensi, menunjukkan kemampuan model dalam menangani pola data yang kompleks (Rais et al., 2025).

139

99

96

2. Pra-pemrosesan Data (Data Preprocessing)

Tahap pra-pemrosesan data adalah langkah penting yang harus dilakukan sebelum menerapkan data ke model *Machine Learning* (Syafei & Efrilianda, 2023). Tahap *preprocessing* memiliki pengaruh penting terhadap performa model machine learning karena kualitas data yang baik dapat membantu meningkatkan akurasi dan stabilitas model prediksi (Ristyan et al., 2025).

a. Eksplorasi dan Pembersihan Data

Tahap ini meliputi pemuatan data, eksplorasi awal, dan pembersihan data seperti penghapusan kolom yang tidak diperlukan atau tidak informatif (seperti kolom indeks yang tidak memiliki nilai prediktif). Pembersihan data memastikan kualitas data yang akan digunakan untuk pelatihan model sehingga dapat mengurangi noise dan meningkatkan performa machine learning (Amarya et al., 2025).

b. Transformasi Logaritmik pada Variabel Target

Dalam banyak masalah regresi harga, variabel target (seperti *price*) sering memiliki distribusi miring (*skewed*) ke kanan. Transformasi logaritmik ($y_{log} = \log(y)$) diterapkan pada variabel target untuk menstabilkan varians dan membuat distribusi data mendekati normal, yang dapat meningkatkan kinerja dan asumsi model regresi (Tr et al., 2025).

c. One-Hot Encoding

Fitur kategorikal (non-numerik) perlu diubah menjadi format numerik agar dapat diproses oleh model *Machine Learning*. One-Hot Encoding adalah teknik yang mengubah variabel kategorikal menjadi serangkaian variabel biner (0 atau 1), di mana setiap variabel baru merepresentasikan satu kategori asli (Utomo et al., 2024)(Bhakti et al., 2024).

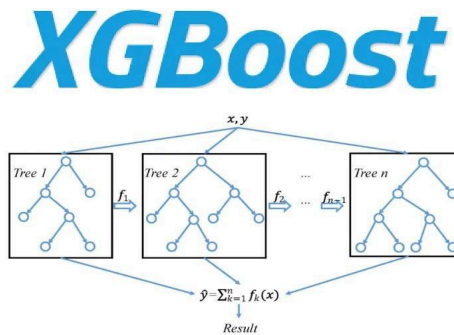
d. Pembagian Data

Data harus dibagi menjadi dua subset, yaitu data pelatihan (*training set*) dan data pengujian (*testing set*). Pembagian ini dilakukan untuk melatih model pada satu set data dan kemudian menguji kinerjanya pada data yang belum pernah dilihat sebelumnya (Utomo et al. 2024),(Islami & Nugroho, 2025). Pembagian dataset menjadi data pelatihan dan data pengujian sangat penting untuk memastikan model

mampu melakukan generalisasi dengan baik terhadap data baru yang belum pernah dipelajari sebelumnya (Joseph et al., 2022).

3. Algoritma Extreme Gradient Boosting (*XGBoost*)

a. Definisi Extreme Gradient Boosting (*XGBoost*)



Gambar 2. 1 Arsitektur *XGBoost*

XGBoost adalah kerangka *gradient-boosting* yang populer dan dikenal memiliki keandalan, presisi, dan portabilitas yang sangat baik (Wijaya et al., 2024). Gambar 2.1 mengilustrasikan mekanisme kerja algoritma ini, yang terdiri dari serangkaian *decision tree* yang saling terhubung secara sekuensial untuk memperbaiki kesalahan prediksi dari pohon sebelumnya. Algoritma ini merupakan implementasi dari *Gradient Boosting Decision Tree (GBDT)* yang dioptimalkan, menggunakan pemrosesan paralel dan teknik *regularization* untuk mengontrol kompleksitas *decision tree* dan mencegah *overfitting* (Syafei & Efrilianda, 2023).

b. Penerapan *XGBoost* dalam Regresi

Dalam konteks prediksi, *XGBoost* efektif digunakan untuk tugas regresi, seperti prediksi harga. Kemampuannya untuk mengelola data kompleks dan menangkap pola yang rumit menjadikannya pilihan yang andal untuk memprediksi tren. Penelitian menunjukkan bahwa *XGBoost* dapat mengungguli model regresi tradisional dan sering digunakan dalam tugas prediksi harga (Wijaya et al., 2024). Algoritma *XGBoost Regressor* telah terbukti menjadi instrumen *machine learning* yang sangat efektif dalam menangani problem prediksi berbasis regresi tabular dengan menghasilkan tingkat akurasi yang tinggi serta nilai error yang rendah (Kurniawan &

Indahyanti, 2024). Karakteristik arsitektur penguat gradiennya mampu memetakan pola variabel dependen secara optimal melalui minimalisasi fungsi kerugian secara bertahap.

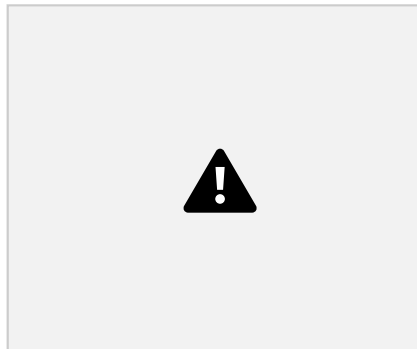
Pada domain finansial, *XGBoost* yang dioptimasi menggunakan *Particle Swarm Optimization (PSO)* terbukti mampu menurunkan kesalahan prediksi harga saham secara signifikan dibandingkan *XGBoost* tanpa optimasi maupun model lain, sehingga menegaskan pentingnya proses *hyperparameter* tuning pada algoritma ini (pebrianti et al., 2024)

4. Optimasi Hyperparameter Menggunakan *Grid Search*

a. Pentingnya Optimasi Hyperparameter

Performa model *Machine Learning*, termasuk *XGBoost*, sangat bergantung pada pemilihan nilai *hyperparameter* yang tepat (Wijaya et al., 2024). Optimasi Hyperparameter bertujuan untuk mencari kombinasi nilai terbaik untuk meningkatkan akurasi dan kinerja model (Fadzryan et al., 2025).

b. Metode Grid Search (*GridSearchCV*)



Gambar 2. 2 *Gridsearch CV*

Grid Search (GridSearchCV) adalah salah satu teknik optimasi *hyperparameter* yang secara sistematis menguji semua kombinasi parameter yang ditentukan dalam sebuah kisi (*grid*). Gambar 2.2, memperlihatkan visualisasi metode ini dalam bentuk kisi (*grid*) yang memetakan parameter-parameter tersebut secara sistematis. Optimasi *hyperparameter* menggunakan *GridSearchCV* dilakukan untuk memperoleh kombinasi parameter terbaik sehingga performa model dapat meningkat secara

optimal (Islami & Nugroho, 2025). Dalam studi kasus optimasi *XGBoost*, hasil evaluasi menunjukkan bahwa teknik Grid Search dapat memberikan hasil yang optimal dan secara substansial meningkatkan akurasi model dibandingkan metode dasar (Wijaya et al., 2024).

Penggunaan *GridSearchCV* juga diterapkan pada studi peramalan kedatangan wisatawan di Bali, di mana Random Forest, *XGBoost*, LSTM, dan model hibrida *RF-XGBoost* semuanya dioptimasi dengan pencarian grid untuk memperoleh performa terbaik dalam hal *RMSE* dan *MAPE* (Waciko et al., 2025).

c. Parameter Kunci *XGBoost* yang Dioptimalkan

Beberapa *hyperparameter* kunci pada *XGBoost* yang dioptimalkan melalui *tuning* meliputi:

- a) *n_estimators* : Menentukan jumlah pohon (*tree*) yang digunakan dalam model (Wijaya et al., 2024).
- b) *learning_rate (eta)* : Mengontrol ukuran langkah selama setiap iterasi. Nilai yang lebih kecil cenderung menghasilkan model yang lebih akurat tetapi memerlukan lebih banyak iterasi (Wijaya et al., 2024).
- c) *max_depth* : Menentukan kedalaman maksimum setiap pohon (Wijaya et al., 2024).
- d) *Subsample* : Rasio *instance* yang diambil secara acak untuk melatih setiap pohon (Wijaya et al., 2024).
- e) *colsample_bytree* : Rasio *instance* yang diambil secara acak untuk melatih setiap pohon (Wijaya et al., 2024).

5. Metrik Evaluasi Model Regresi

Evaluasi model bertujuan untuk mengukur kualitas dan kinerja model dalam memprediksi nilai actual (Wijaya et al., 2024). Metrik yang digunakan dalam penelitian ini adalah:

a. *Mean Absolute Error (MAE)*

MAE mengukur rata-rata magnitudo kesalahan dalam satu set prediksi. *MAE* memberikan informasi tentang seberapa dekat prediksi model dengan nilai actual (Fadzryan et al., 2025).

$$MAE = \frac{1}{n} + \sum_{i=1}^n |y_i - \hat{y}_i| \quad (2.1)$$

b. *Mean Squared Error (MSE)*

MSE mengukur rata-rata kuadrat dari kesalahan. Karena kesalahan dikuadratkan, metrik ini memberikan bobot yang lebih besar pada *outlier* (nilai ekstrem) (Fadzryan et al., 2025).

$$MSE = \frac{1}{n} + \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (2.2)$$

c. *Root Mean Squared Error (RMSE)*

RMSE adalah akar kuadrat dari *MSE*. Metrik ini sering disukai karena memiliki unit yang sama dengan variabel target aslinya, sehingga lebih mudah diinterpretasikan (Fadzryan et al., 2025)(Tr et al., 2025).

$$RMSE = \sqrt{MSE} \quad (2.3)$$

d. *Koefisien Determinasi (R^2 Score)*

R^2 mengukur proporsi varians pada variabel dependen yang dapat dijelaskan oleh model regresi. R^2 mendekati 1 menunjukkan model yang sangat baik dalam menjelaskan data (Fadzryan et al., 2025).

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (2.4)$$

B. Kajian Penelitian Terdahulu

Penelitian mengenai prediksi harga rumah telah banyak dilakukan dengan memanfaatkan berbagai algoritma *machine learning*, khususnya model berbasis *tree* (Dwi et al., 2024). (Dhiwa Aqsha, 2025) melakukan penelitian dengan membandingkan kinerja algoritma *Extreme Gradient Boosting (XGBoost)* dan *Random Forest* dalam memprediksi harga rumah di wilayah Jabodetabek. Hasil penelitian menunjukkan bahwa *Random Forest* memberikan performa yang lebih baik dengan nilai R^2 sebesar 0,77 dibandingkan *XGBoost* yang hanya mencapai R^2 sebesar 0,72. Meskipun demikian, penelitian ini menunjukkan bahwa *XGBoost* memiliki potensi yang besar sebagai model prediksi harga rumah, namun performanya belum optimal karena belum dilakukan proses optimasi *hyperparameter*.

Penelitian lain dilakukan oleh (Utomo et al., 2024) yang menggunakan algoritma *Decision Tree Regression (DTR)* untuk meningkatkan akurasi prediksi harga rumah dengan memanfaatkan atribut properti multivariat. Penelitian tersebut menghasilkan nilai *MAE* sebesar 0,1691 dan *RMSE* sebesar 0,2227, serta menekankan keunggulan model berbasis *tree* dalam hal interpretabilitas. Hasil penelitian ini menjadi dasar bahwa model *tree-based* sangat relevan untuk permasalahan prediksi harga rumah, namun masih memiliki keterbatasan dalam menangkap pola non-linear yang kompleks dibandingkan model ensemble lanjutan seperti *XGBoost*.

Penelitian mengenai estimasi nilai real estate juga dilakukan oleh (Utami et al., 2025) menggunakan algoritma *Random Forest Regression*. Melalui skema pembagian data sebesar 90:10, model mereka mampu mencapai skor R^2 sebesar 81,88%. Meskipun model tersebut menunjukkan performa yang cukup kuat, penggunaan parameter bawaan (*default*) masih membuka ruang pengembangan melalui algoritma *boosting* dan pencarian parameter yang lebih optimal.

Selanjutnya, (Fadzryan et al., 2025) mengusulkan kombinasi *Random Forest* dengan *Optuna Hyperparameter Tuning* menggunakan pendekatan *Bayesian Optimization*. Hasil penelitian menunjukkan peningkatan performa yang signifikan setelah proses optimasi, dengan nilai R^2 mencapai 0,92. Penelitian ini menegaskan

23 bahwa proses *hyperparameter tuning* memiliki peran yang sangat penting dalam meningkatkan akurasi model prediksi harga rumah, khususnya pada algoritma berbasis *ensemble tree*. Namun, penelitian ini masih terbatas pada penggunaan *Random Forest* dan belum mengeksplorasi potensi algoritma *XGBoost* dengan teknik optimasi lain.

2 Penelitian oleh (Wijaya et al., 2024) secara khusus membahas optimasi algoritma *XGBoost* menggunakan beberapa teknik *hyperparameter tuning*, yaitu *Random Search*, *Grid Search*, dan *Bayesian Optimization* pada kasus prediksi penjualan ritel. Hasil penelitian menunjukkan bahwa *Grid Search* memberikan hasil terbaik, dengan peningkatan nilai R^2 dari 97,31% menjadi 98,41%. Meskipun konteks penelitian ini bukan prediksi harga rumah, metodologi yang digunakan sangat relevan karena membuktikan bahwa kombinasi *XGBoost* dan *GridSearchCV* mampu meningkatkan performa model secara signifikan dan stabil.

100 Selain itu, (Syafei & Efrilianda, 2023) menerapkan *XGBoost* dan *LightGBM* pada kasus prediksi kebangkrutan perusahaan. Hasil penelitian menunjukkan bahwa algoritma *gradient boosting*, khususnya *XGBoost*, sangat efektif dalam menangani data yang kompleks dan berdimensi tinggi. Meskipun penelitian ini berfokus pada klasifikasi, temuan tersebut memperkuat argumen bahwa *XGBoost* merupakan algoritma yang andal dan fleksibel untuk berbagai permasalahan *machine learning*, termasuk regresi harga rumah.

64 Berdasarkan kajian terhadap penelitian-penelitian terdahulu tersebut, dapat disimpulkan bahwa algoritma berbasis *tree ensemble* seperti *Random Forest* dan *XGBoost* memiliki performa yang unggul dalam prediksi harga rumah. Namun, masih terdapat kesenjangan penelitian, khususnya pada penerapan *XGBoost* yang belum dioptimalkan secara maksimal menggunakan *GridSearchCV* dalam satu *pipeline* yang terintegrasi dengan pra-pemrosesan data regresi, seperti transformasi logaritmik dan pengolahan fitur kategorikal. Oleh karena itu, penelitian ini difokuskan pada optimasi *XGBoost* menggunakan *Grid Search Cross-Validation* untuk meningkatkan akurasi prediksi harga rumah dan meminimalkan selisih antara harga aktual dan harga hasil prediksi.

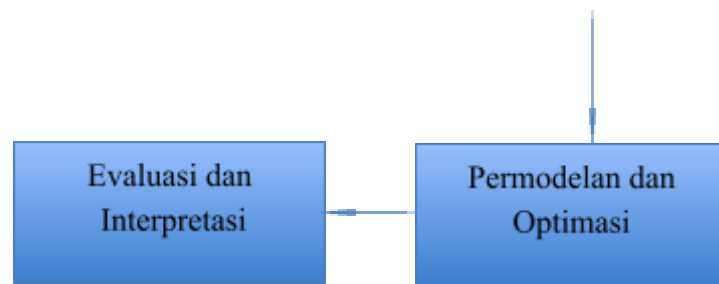
Berdasarkan beberapa penelitian terdahulu yang telah dibahas, pembaruan dalam penelitian ini terletak pada penerapan algoritma *XGBoost Regressor* yang dioptimasi menggunakan *GridSearchCV* dalam suatu pipeline yang terintegrasi dengan tahap pra-pemrosesan data, meliputi transformasi logaritmik pada variabel target serta pengolahan fitur kategorikal menggunakan *ColumnTransformer* dengan *One-Hot Encoding*.

Selain itu, penelitian ini tidak hanya berfokus pada peningkatan akurasi model, tetapi juga pada upaya meminimalkan selisih antara nilai prediksi dan nilai aktual. Lebih lanjut, penelitian ini mengadopsi metode dari penelitian acuan sebagai baseline pembanding, yang kemudian diimplementasikan dan dievaluasi pada dataset *Jakarta House Price Dataset* untuk menguji kemampuan generalisasi model dalam konteks data properti di Indonesia. Dengan pendekatan tersebut, diharapkan model yang dihasilkan menjadi lebih optimal, stabil, dan relevan dengan kondisi pasar properti di Indonesia.

C. Kerangka Berfikir

Kerangka berpikir dalam penelitian ini menggambarkan alur proses yang sistematis dalam membangun model prediksi harga rumah berbasis *machine learning*. Proses diawali dari pengumpulan dataset harga rumah yang kemudian dianalisis pada tahap *data understanding* untuk mengetahui karakteristik data. Selanjutnya dilakukan *data preparation* yang meliputi pembersihan data, penanganan *missing value*, serta transformasi dan pengolahan fitur agar data siap digunakan. Setelah itu, data yang telah diproses digunakan dalam tahap pemodelan menggunakan algoritma *XGBoost Regressor*, yang kemudian dioptimasi melalui proses *hyperparameter tuning* untuk mendapatkan performa terbaik. Tahap akhir adalah evaluasi model menggunakan metrik kinerja untuk menilai tingkat akurasi prediksi. Alur tersebut menunjukkan hubungan yang terstruktur dari data hingga menghasilkan model prediksi yang optimal, sebagaimana ditampilkan pada gambar kerangka berpikir penelitian.





Gambar 2. 3 Kerangka Berfikir

Gambar 2.3 menyajikan kerangka berpikir penelitian yang disusun untuk menggambarkan hubungan logis dan sistematis antara masalah yang teridentifikasi, landasan teoretis melalui studi literatur, pemilihan metode, hingga tahapan operasional yang ditempuh. Kerangka ini berfungsi sebagai cetak biru konseptual yang mengintegrasikan seluruh elemen penelitian guna mencapai tujuan yang telah ditetapkan secara efektif.

1. Studi Literatur

Studi literatur pada penelitian ini dilakukan untuk menentukan algoritma, skema pra-pemrosesan data, serta metode optimasi *hyperparameter* yang digunakan. Peneliti menelaah jurnal nasional dan internasional yang membahas prediksi harga rumah menggunakan *machine learning*, khususnya model regresi berbasis *tree ensemble*.

Berdasarkan hasil kajian literatur, algoritma *Extreme Gradient Boosting (XGBoost)* dipilih karena memiliki performa yang baik dalam menangani data tabular dengan hubungan non-linear. Selain itu, studi literatur digunakan untuk menetapkan tahapan pra-pemrosesan data, meliputi pembersihan data, pembagian data latih dan data uji, pengolahan fitur kategorikal menggunakan *One-Hot Encoding*, serta transformasi logaritmik pada variabel target untuk mengatasi distribusi data yang tidak seimbang.

Selanjutnya, peneliti mengkaji berbagai metode optimasi *hyperparameter* dan menetapkan *Grid Search Cross-Validation (GridSearchCV)* sebagai teknik optimasi yang digunakan, karena mampu mencari kombinasi parameter terbaik secara sistematis dan tervalidasi. Hasil studi literatur ini menjadi dasar penyusunan alur

metodologi penelitian dan perancangan *pipeline XGBoost Regressor* yang dioptimasi untuk meningkatkan akurasi prediksi harga rumah.

2. Identifikasi Kesenjangan dan Arah Solusi

Meskipun fondasi konseptual telah mantap, studi literatur menunjukkan adanya kesenjangan (gap) pada penelitian regresi harga rumah terdahulu, yaitu:

- a. Banyak model *ensemble* yang masih menghasilkan selisih prediksi ekstrem pada data tertentu, meskipun akurasi totalnya tinggi.
- b. Belum adanya integrasi *pipeline* yang eksplisit dan sistematis yang menggabungkan pra-pemrosesan data regresi spesifik (termasuk Transformasi Logaritmik untuk *skewness*) dengan *XGBoost* dan *GridSearchCV* pada dataset harga rumah yang sama.

Kesenjangan inilah yang menjadi dasar rumusan masalah penelitian, Dengan demikian, penelitian ini berupaya menyusun alur pra-pemrosesan dan prosedur *tuning* yang sistematis untuk memaksimalkan potensi model, yang hasilnya akan diukur melalui perbandingan metrik akurasi (R^2) dan kesalahan (*MAE*, *RMSE*) terhadap model *baseline*.

3. Tahap Pra-pemrosesan Data (Data Preprocessing)

Fase ini merupakan implementasi langsung dari kebutuhan *preprocessing* yang diidentifikasi untuk mendukung optimasi model regresi. Proses dimulai dengan Pembersihan Data dari Dataset Harga Rumah (*Jakarta House Prices dataset*), termasuk penanganan nilai yang hilang. Data dibagi menjadi Fitur (X) dan Target (y). Untuk menstabilkan varians dan meningkatkan asumsi model regresi, variabel Target (y) menjalani Transformasi Logaritmik ($\ln(y)$) untuk mengatasi distribusi yang miring (Pebrianti et al., 2024). Setelah Pembagian Data menjadi Data Pelatihan dan Data Pengujian, fase ini ditutup dengan *Preprocessing* Fitur menggunakan *ColumnTransformer* dengan *One-Hot Encoding* untuk mengkonversi fitur kategorikal menjadi format numerik.

4. Tahap Pemodelan dan Optimasi

Fase ini berfokus pada pembangunan dan optimalisasi model sesuai dengan rumusan masalah. Proses dimulai dengan Pembentukan *Pipeline Model Awal*, yang menggabungkan *preprocessing* fitur dengan *XGBoost Regressor (XGBR)*. Selanjutnya, model ini menjalani Optimasi *Hyperparameter* melalui teknik *Grid Search Cross-Validation (GridSearchCV)*. Proses *Grid Search* secara sistematis menguji kombinasi *hyperparameter* untuk mencari konfigurasi yang menghasilkan kinerja terbaik dan stabil (Sharma et al., 2024; Waciko et al., 2025). Secara paralel, Model *Baseline (XGBR tanpa tuning)* juga dilatih untuk dijadikan tolok ukur perbandingan. Hasil dari optimasi ini adalah Model *XGBoost Terbaik* yang telah di-*tuning*.

5. Tahap Evaluasi dan Interpretasi

Fase akhir bertujuan untuk menjawab secara kuantitatif kedua rumusan masalah penelitian dengan membandingkan kinerja Model Optimasi dengan Model *Baseline*. Model Terbaik digunakan untuk melakukan Prediksi Data Uji. Hasil prediksi logaritmik ($\ln(\text{harga})$) dikembalikan ke skala harga asli (harga riil) melalui Invers Transformasi (eksponensial). Terakhir, Evaluasi Kinerja Model dilakukan dengan menghitung metrik regresi standar:

- a. R^2 Score: Untuk mengukur peningkatan akurasi model (Menjawab Rumusan Masalah 1).
- b. MAE , MSE dan $RMSE$: Untuk mengukur penurunan selisih atau kesalahan prediksi menjawab rumusan masalah 2.

Hasil akhir dari evaluasi ini memberikan kesimpulan tentang tingkat Akurasi Prediksi Harga Rumah yang dicapai dan memvalidasi keberhasilan prosedur optimasi (Sharma et al., 2024).

82

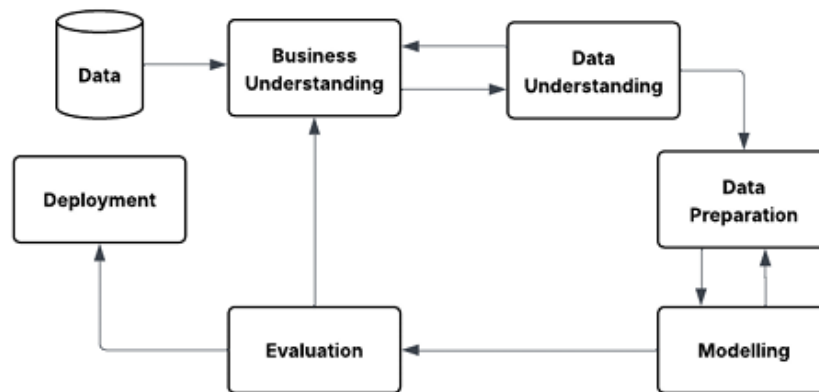
BAB III

METODE PENELITIAN

A. Alur Penelitian

12

Penelitian ini dilaksanakan dengan mengadopsi kerangka kerja CRISP-DM melalui serangkaian tahapan sistematis, mulai dari *business understanding*, *data understanding*, *data preparation*, *modelling*, *evaluation*, hingga tahap *deployment* model (Kurnia et al., 2024). Alur penelitian secara menyeluruh disajikan pada Gambar 3.1 yang menggambarkan keterkaitan setiap fase CRISP-DM secara terperinci, mulai dari prapemrosesan data mentah hingga evaluasi kinerja dan rencana pemanfaatan model prediksi harga rumah yang telah dioptimalkan.



Gambar 3. 1 Alur Peneletian

1

1. *Business Understanding*

65

Tahap *Business Understanding* bertujuan untuk mendefinisikan permasalahan utama dalam penelitian, yaitu bagaimana membangun model prediksi harga rumah yang akurat dan mampu meminimalkan kesalahan prediksi. Permasalahan ini muncul karena harga properti dipengaruhi oleh berbagai faktor, seperti lokasi, luas bangunan, luas tanah, serta jumlah fasilitas, sehingga sulit diprediksi secara manual tanpa bantuan pendekatan berbasis data.

Selain itu, penelitian ini juga merupakan pengembangan dari penelitian terdahulu yang menggunakan dataset luar negeri. Untuk menyesuaikan dengan kondisi di Indonesia, khususnya wilayah Jakarta, dilakukan proses *retraining* model menggunakan dataset lokal. Penelitian ini juga melakukan perbandingan antara model yang diadaptasi dari penelitian terdahulu dengan model yang dikembangkan dalam penelitian ini melalui proses optimasi parameter.

Adapun tujuan yang ingin dicapai dalam penelitian ini adalah sebagai berikut:

- Mengembangkan model prediksi harga rumah berbasis *machine learning* menggunakan algoritma XGBoost
- Mengoptimasi performa model menggunakan metode *GridSearchCV*
- Melakukan *retraining* model menggunakan dataset lokal untuk menyesuaikan dengan karakteristik pasar properti di Jakarta
- Membandingkan performa model hasil adaptasi penelitian terdahulu dengan model yang dikembangkan dalam penelitian ini
- Menghasilkan estimasi harga rumah yang akurat dan dapat diimplementasikan dalam aplikasi

Keberhasilan model diukur menggunakan beberapa metrik evaluasi, yaitu:

- R^2 Score untuk mengukur tingkat akurasi model
- Mean Absolute Error (MAE), Mean Squared Error (MSE), dan Root Mean Squared Error (RMSE) untuk mengukur tingkat kesalahan prediksi

2. Data Understanding

Tahap *Data Understanding* dilakukan untuk memahami karakteristik dataset yang digunakan. Dataset yang digunakan merupakan data harga rumah di Jakarta (*Jakarta House Price Dataset*) yang diperoleh dari Kaggle dan terdiri dari sekitar 10.000 data dengan beberapa atribut numerik dan kategorikal.

Fitur yang digunakan dalam penelitian ini meliputi:

- Variabel target (y): price
- Variabel fitur (X):
 - district
 - city

- 3) bed_rooms
- 4) bath_rooms
- 5) carport
- 6) land_area
- 7) building_area

Berdasarkan hasil eksplorasi data, diperoleh beberapa temuan penting:

- a. Terdapat missing values pada *land_area* dan *building_area*
- b. Terdapat kolom tidak relevan seperti *index* dan *Unnamed: 0*
- c. Distribusi data *price* bersifat skewed (tidak normal)

Temuan ini menjadi dasar dalam proses preprocessing pada tahap berikutnya.

3. Data Preparation

Tahap *Data Preparation* bertujuan untuk menyiapkan data agar siap digunakan dalam proses pemodelan. Proses ini melibatkan beberapa langkah penting untuk meningkatkan kualitas data.

Langkah-langkah yang dilakukan meliputi:

- a. Mengisi missing value pada *land_area* dan *building_area* menggunakan median
- b. Menghapus kolom yang tidak relevan seperti *index* dan *Unnamed: 0*
- c. Memisahkan data menjadi fitur (X) dan target (y)
- d. Melakukan transformasi logaritmik pada variabel target untuk mengurangi skewness

Selain itu, dilakukan proses preprocessing fitur sebagai berikut:

- a. Fitur kategorikal (*district*, *city*) diubah menggunakan One-Hot Encoding
- b. Fitur numerik digunakan secara langsung (*passthrough*)

Selanjutnya, data dibagi menjadi:

- a. Data training (80%)
- b. Data testing (20%)

Dengan menggunakan *random_state = 42* untuk menjaga konsistensi hasil.

Tahap *Data Preparation* ini dilakukan untuk memastikan data yang digunakan

memiliki kualitas yang baik sehingga model dapat mempelajari pola secara optimal dan menghasilkan prediksi yang lebih akurat.

4. Modelling

Tahap *Modeling* dilakukan dengan membangun model prediksi menggunakan algoritma XGBoost Regressor yang diintegrasikan dalam sebuah pipeline bersama preprocessing.

Untuk mendapatkan performa terbaik, dilakukan optimasi model menggunakan GridSearchCV dengan parameter sebagai berikut:

Tabel 3. 1 Tabel Parameter *GridSearch CV*

<i>Hyperparameter</i>	<i>Candidate Value</i>
<i>n_estimators</i>	[200, 300, 500]
<i>max_depth</i>	[3, 5, 7]
<i>learning_rate</i>	[0.01, 0.05, 0.1]
<i>subsample</i>	[0.8, 1.0]
<i>colsample_bytree</i>	[0.8, 1.0]
<i>gamma</i>	[0, 0.1, 0.2]
<i>min_child_weight</i>	[1, 3, 5]

Proses tuning dilakukan menggunakan:

- 3-fold cross-validation*
- Metrik evaluasi: *negative mean squared error*

Hasil dari proses ini adalah model terbaik yang memiliki kombinasi parameter optimal.

5. Evaluation

Tahap *Evaluation* dilakukan untuk menilai kinerja model dalam memprediksi harga rumah menggunakan data uji (*testing set*) yang tidak digunakan selama proses pelatihan. Hal ini bertujuan untuk memastikan kemampuan generalisasi model terhadap data baru.

Model menghasilkan prediksi dalam skala logaritmik karena sebelumnya dilakukan transformasi log pada variabel target. Oleh karena itu, hasil prediksi dikonversi kembali ke skala harga asli menggunakan fungsi eksponensial agar lebih mudah diinterpretasikan.

Evaluasi model dilakukan menggunakan beberapa metrik regresi, yaitu:

- a. *Mean Absolute Error (MAE)* untuk mengukur rata-rata kesalahan absolut antara nilai aktual dan prediksi.
- b. *Mean Squared Error (MSE)* untuk mengukur rata-rata kesalahan kuadrat.
- c. *Root Mean Squared Error (RMSE)* untuk memberikan interpretasi kesalahan dalam satuan yang sama dengan variabel target.
- d. *R² Score* untuk mengukur kemampuan model dalam menjelaskan variasi data.

Selain evaluasi secara numerik, dilakukan juga analisis visual untuk memahami performa model secara lebih mendalam. Analisis ini dilakukan melalui grafik *Actual vs Predicted* untuk melihat kesesuaian antara nilai aktual dan hasil prediksi, serta *Residual Plot* untuk mengetahui distribusi kesalahan prediksi dan mendeteksi adanya pola tertentu.

Selain itu, penelitian ini juga melakukan evaluasi komparatif antara dua model, yaitu model *XGBoost* yang dioptimasi menggunakan *GridSearchCV* sebagai model utama dan model *Random Forest* yang dioptimasi menggunakan metode *Optuna* sebagai model pembandingan. Perbandingan dilakukan menggunakan metrik evaluasi yang sama, sehingga dapat diketahui model mana yang memiliki kinerja lebih baik dalam memprediksi harga rumah pada dataset local (Aulia et al., 2023).

Hasil perbandingan tersebut selanjutnya dianalisis pada BAB IV untuk melihat peningkatan performa model setelah dilakukan proses *retraining* dan optimasi parameter.

6. Deployment

Tahap *Deployment* dilakukan dengan mengimplementasikan model ke dalam aplikasi berbasis web menggunakan Streamlit. Model terbaik yang telah disimpan dalam format *.pkl* menggunakan joblib dimuat kembali untuk digunakan dalam proses prediksi.

Aplikasi memungkinkan pengguna untuk memasukkan data properti, antara lain:

- a. wilayah (district)
- b. kota (city)
- c. jumlah kamar tidur (bedroom)

d. jumlah kamar mandi (bathroom)

e. garasi (carport)

f. luas tanah (land area)

g. luas bangunan (building area)

Sistem kemudian akan melakukan preprocessing secara otomatis dan menghasilkan prediksi harga rumah dalam skala asli. Hasil prediksi ditampilkan dalam bentuk estimasi harga serta rentang nilai untuk memberikan gambaran yang lebih realistis.

Penggunaan Streamlit dipilih karena kemudahan implementasi serta kemampuannya dalam membangun aplikasi interaktif berbasis Python dengan cepat.

B. Teknik Pengumpulan Data

Teknik pengumpulan data yang diterapkan dalam penelitian ini adalah *Studi Dokumentasi* (Data Sekunder). Teknik ini dipilih karena data yang digunakan, yaitu dataset harga rumah, sudah tersedia secara publik dan terstruktur, sehingga penelitian ini berfokus pada analisis komputasi terhadap data yang telah ada.

a) Jenis dan Sumber Data

1. Jenis Data

Data yang digunakan merupakan data kuantitatif sekunder dalam bentuk tabular. Data ini meliputi variabel independen (X) yang merepresentasikan atribut properti, seperti luas area, jumlah kamar, dan lokasi, serta variabel dependen (Y) yaitu harga jual rumah (*price*).

2. Sumber Data

Data sekunder dalam penelitian ini diperoleh dari repositori data publik, yaitu Kaggle. Dataset yang digunakan adalah *Jakarta House Price Dataset* yang dapat diakses melalui tautan berikut: "[Jakarta House Price Dataset](#)".

Pemilihan dataset ini didasarkan pada kelengkapan atribut serta relevansinya dengan penelitian prediksi harga rumah menggunakan metode *machine learning*, khususnya algoritma XGBoost Regressor. Dengan karakteristik tersebut, dataset yang

digunakan dinilai mampu merepresentasikan kondisi pasar properti di Jakarta serta mendukung proses pelatihan dan evaluasi model prediksi harga rumah secara efektif.

b) Prosedur Studi Dokumentasi

Studi dokumentasi dilakukan melalui beberapa tahapan sistematis untuk memastikan dataset yang digunakan sesuai dengan kebutuhan penelitian, yaitu sebagai berikut:

1. Identifikasi Kebutuhan Data

Menentukan spesifikasi dataset yang diperlukan, yaitu data historis harga rumah yang mencakup variabel numerik dan kategorikal. Kriteria utama adalah ketersediaan fitur-fitur penting yang berpengaruh terhadap harga properti, sebagaimana dijelaskan pada Subbab Struktur Data.

2. Pencarian dan Seleksi *Dataset*

Melakukan pencarian dataset pada platform Kaggle dengan kata kunci yang relevan. Proses seleksi dilakukan berdasarkan jumlah data, kelengkapan atribut, serta kesesuaian dengan kebutuhan penelitian.

3. Pengunduhan Data

Dataset yang telah memenuhi kriteria kemudian diunduh untuk selanjutnya digunakan dalam tahap *data preparation*, yang meliputi proses pembersihan data, transformasi logaritmik, serta encoding fitur sebelum digunakan pada tahap pemodelan.

150

19

72

BAB IV

HASIL DAN PEMBAHASAN

60

A. Data Hasil Penelitian

1. Hasil Eksploitasi Data

Tahap eksplorasi data dilakukan untuk memahami karakteristik dataset yang digunakan dalam penelitian, meliputi struktur data, kualitas data, serta distribusi nilai pada variabel target. Dataset yang digunakan merupakan data harga rumah di wilayah Jakarta yang terdiri dari berbagai atribut properti seperti lokasi, jumlah kamar, serta luas bangunan dan tanah.

Berdasarkan hasil eksplorasi awal, ditampilkan beberapa baris pertama dataset untuk memberikan gambaran umum mengenai struktur dan isi data.

	index	price	district	city	bed_rooms	bath_rooms	carport	land_area	building_area
0	1	59000000000	Citra Garden	Jakarta Barat	2	4	2	250.0	350.0
1	2	27000000000	Jelambar	Jakarta Barat	4	2	0	100.0	225.0
2	3	22000000000	Jelambar	Jakarta Barat	3	3	0	60.0	140.0
3	4	19000000000	Jelambar	Jakarta Barat	3	2	0	60.0	120.0
4	5	21000000000	Tanjung Duren	Jakarta Barat	4	3	0	56.0	108.0

Gambar 4. 1 Jakarta House Price dataset

Berdasarkan Gambar 4.1, dataset memiliki beberapa atribut utama yaitu *price*, *district*, *city*, *bed_rooms*, *bath_rooms*, *carport*, *land_area*, dan *building_area*. Variasi nilai pada setiap atribut menunjukkan bahwa dataset memiliki keragaman karakteristik properti yang dapat digunakan dalam proses pemodelan.

Selanjutnya dilakukan analisis struktur dataset untuk mengetahui jumlah data, tipe data, serta keberadaan nilai yang hilang (*missing values*).

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 9 columns):
#   Column      Non-Null Count  Dtype
---  -
0   index       10000 non-null  int64
1   price       10000 non-null  int64
2   district    10000 non-null  object
3   city        10000 non-null  object
4   bed_rooms   10000 non-null  int64
5   bath_rooms  10000 non-null  int64
6   carport     10000 non-null  int64
7   land_area   9997 non-null   float64
8   building_area 9995 non-null   float64
dtypes: float64(2), int64(5), object(2)
memory usage: 703.3+ KB
```

Gambar 4. 2 Struktur Dataset

Berdasarkan Gambar 4.2, dataset terdiri dari 10.000 data dengan kombinasi tipe data numerik dan kategorikal. Pada tahap ini juga teridentifikasi adanya nilai yang hilang pada beberapa atribut.

Untuk mengetahui jumlah missing value secara lebih rinci, dilakukan pengecekan pada setiap kolom.

```
***
      0
index  0
price  0
district  0
city  0
bed_rooms  0
bath_rooms  0
carport  0
land_area  3
building_area  5
dtype: int64
```

Gambar 4. 3 Missing Value Dataset

Berdasarkan Gambar 4.3, diketahui bahwa terdapat nilai kosong pada atribut *land_area* sebanyak 3 data dan *building_area* sebanyak 5 data, sedangkan atribut lainnya tidak memiliki nilai kosong. Keberadaan missing value ini perlu ditangani pada tahap data preparation agar tidak memengaruhi kinerja model.

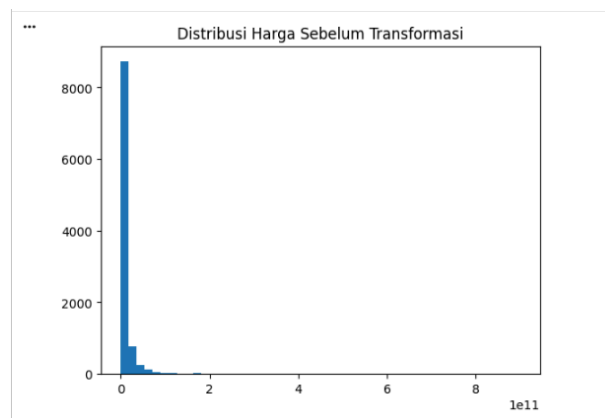
Selain itu, dilakukan analisis statistik deskriptif untuk memahami distribusi nilai pada masing-masing atribut numerik.

	index	price	bed_rooms	bath_rooms	carport	land_area	building_area
count	10000.00000	1.000000e+04	10000.000000	10000.000000	10000.000000	9997.000000	9995.000000
mean	5000.50000	1.004029e+10	4.253500	3.521300	1.578900	296.610183	319.962481
std	2886.89568	2.004519e+10	3.176609	2.900136	1.743467	502.921686	357.676034
min	1.00000	1.850000e+06	0.000000	0.000000	0.000000	1.000000	1.000000
25%	2500.75000	2.500000e+09	3.000000	2.000000	1.000000	100.000000	139.000000
50%	5000.50000	4.500000e+09	4.000000	3.000000	1.000000	177.000000	230.000000
75%	7500.25000	9.800000e+09	5.000000	4.000000	2.000000	335.000000	400.000000
max	10000.00000	9.000000e+11	63.000000	63.000000	31.000000	17714.000000	15331.000000

Gambar 4. 4 Statistik Deskriptif Dataset

Berdasarkan Gambar 4.4, terlihat bahwa variabel *price* memiliki rentang nilai yang sangat besar, mulai dari nilai minimum yang relatif kecil hingga nilai maksimum yang sangat tinggi. Nilai standar deviasi yang tinggi menunjukkan bahwa data memiliki penyebaran yang luas dan kemungkinan terdapat nilai ekstrem (*outlier*).

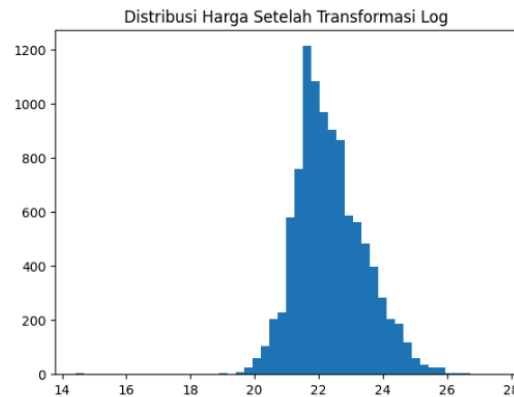
Untuk melihat bentuk distribusi data secara visual, dilakukan analisis menggunakan histogram pada variabel *price* sebelum dilakukan transformasi.



Gambar 4. 5 Distribusi Harga sebelum Transformasi

Berdasarkan Gambar 4.5, terlihat bahwa distribusi harga rumah bersifat tidak simetris dan cenderung condong ke kanan (*right-skewed*). Kondisi ini menunjukkan bahwa sebagian besar data berada pada rentang harga rendah, namun terdapat beberapa nilai yang sangat tinggi sehingga menyebabkan distribusi tidak normal.

Untuk mengatasi permasalahan tersebut, dilakukan transformasi logaritmik pada variabel target guna menstabilkan distribusi data.



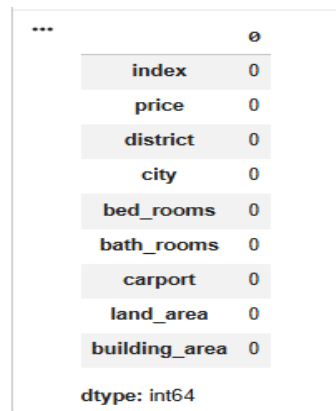
Gambar 4. 6 Distribusi Harga Setelah Transformasi

16 Berdasarkan Gambar 4.6, terlihat bahwa distribusi data menjadi lebih simetris dan mendekati distribusi normal setelah dilakukan transformasi logaritmik. Transformasi ini bertujuan untuk mengurangi pengaruh nilai ekstrem serta meningkatkan kemampuan model dalam mempelajari pola data.

2. Proses transformasi data

21 Tahap transformasi data dilakukan untuk mempersiapkan dataset agar siap digunakan dalam proses pemodelan. Proses ini meliputi penanganan missing value, pembersihan data, serta transformasi fitur agar sesuai dengan kebutuhan algoritma machine learning.

107 Berdasarkan hasil tahap sebelumnya, ditemukan adanya nilai kosong (*missing values*) pada beberapa atribut, sehingga dilakukan proses imputasi menggunakan nilai median. Setelah proses tersebut dilakukan, dilakukan pengecekan ulang untuk memastikan bahwa tidak terdapat lagi nilai yang hilang.



...	0
index	0
price	0
district	0
city	0
bed_rooms	0
bath_rooms	0
carport	0
land_area	0
building_area	0
dtype: int64	

Gambar 4. 7 Hasil Penanganan Missing Value

Berdasarkan Gambar 4.7, seluruh atribut telah memiliki nilai lengkap tanpa adanya missing value. Hal ini menunjukkan bahwa proses imputasi berhasil dilakukan dan dataset telah siap untuk tahap berikutnya.

Selanjutnya dilakukan proses pembersihan data dengan menghapus atribut yang tidak relevan, seperti kolom *index*, yang tidak memiliki pengaruh terhadap proses prediksi.

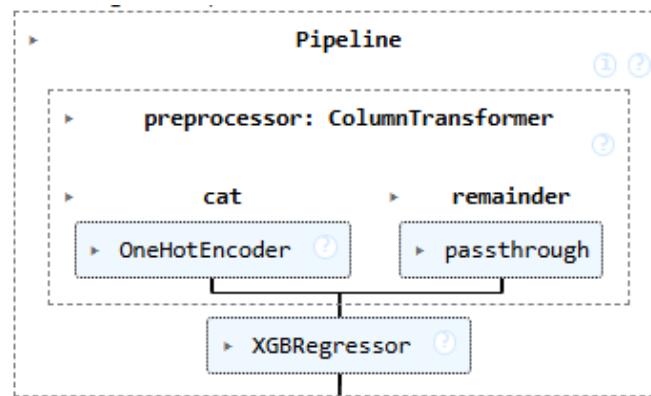
```
Index(['price', 'district', 'city', 'bed_rooms', 'bath_rooms',
      'carport', 'land_area', 'building_area'],
      dtype='object')
```

Gambar 4. 8 Penghapusan Kolom Tidak Relevan

Berdasarkan Gambar 4.8, struktur dataset telah disesuaikan dengan hanya mempertahankan atribut yang relevan untuk proses pemodelan.

Selain itu, dilakukan transformasi logaritmik pada variabel target (*price*) untuk mengatasi distribusi data yang tidak normal. Transformasi ini bertujuan untuk menstabilkan varians serta mengurangi pengaruh nilai ekstrem sehingga model dapat mempelajari pola data dengan lebih optimal.

Untuk menangani perbedaan tipe data, digunakan metode *ColumnTransformer* yang memungkinkan penerapan transformasi yang berbeda pada masing-masing jenis fitur. Fitur kategorikal seperti *district* dan *city* diubah menggunakan teknik *One-Hot Encoding*, sedangkan fitur numerik diteruskan tanpa perubahan (*passthrough*).



Gambar 4. 9 Pipeline

Berdasarkan Gambar 4.9, seluruh proses preprocessing dan pemodelan diintegrasikan dalam sebuah pipeline yang terdiri dari tahap transformasi fitur dan algoritma XGBoost Regressor. Penggunaan pipeline ini bertujuan untuk memastikan bahwa seluruh proses transformasi dilakukan secara konsisten baik pada data pelatihan maupun data pengujian.

Dengan dilakukannya proses transformasi data, dataset telah berada dalam kondisi bersih, terstruktur, dan siap digunakan pada tahap pemodelan.

3. Hasil modeling

Tahap modeling dilakukan untuk membangun model prediksi harga rumah berdasarkan data yang telah melalui proses pembersihan dan transformasi pada tahap sebelumnya. Model yang digunakan adalah algoritma XGBoost Regressor yang diintegrasikan dalam sebuah pipeline bersama proses preprocessing, sehingga seluruh tahapan transformasi dan pemodelan dapat dilakukan secara otomatis dan konsisten.

Untuk memperoleh performa model yang optimal, dilakukan proses *hyperparameter tuning* menggunakan metode *GridSearchCV*. Proses ini bertujuan untuk mencari kombinasi parameter terbaik melalui pengujian berbagai kemungkinan konfigurasi parameter yang telah ditentukan sebelumnya.

```
Fitting 3 folds for each of 972 candidates, totalling 2916 fits
Best model saved as 'house_price_pipeline_tuned.pkl'

...

Best Hyperparameters:

regressor__colsample_bytree
→ 0.8

regressor__gamma
→ 0

regressor__learning_rate
→ 0.1

regressor__max_depth
→ 3

regressor__min_child_weight
→ 3

regressor__n_estimators
→ 500

regressor__subsample
→ 1.0
```

Gambar 4. 10 Hasil Hyperparameter Tuning

Berdasarkan Gambar 4.10, diperoleh kombinasi parameter terbaik yang digunakan dalam model, yaitu *n_estimators* sebesar 500, *max_depth* sebesar 3, *learning_rate* sebesar 0.1, *subsample* sebesar 1.0, serta *colsample_bytree* sebesar 0.8. Selain itu, parameter *gamma* bernilai 0 dan *min_child_weight* sebesar 3, yang menunjukkan bahwa model tidak menggunakan penalti tambahan pada pembentukan split dan tetap menjaga kompleksitas model agar tidak terlalu tinggi. Proses pencarian parameter dilakukan dengan jumlah kombinasi yang besar, sehingga model yang dihasilkan telah berada pada kondisi optimal.

Selanjutnya, model terbaik yang diperoleh digunakan untuk melakukan prediksi terhadap data uji, dan hasilnya dievaluasi menggunakan beberapa metrik untuk mengukur kinerja model.

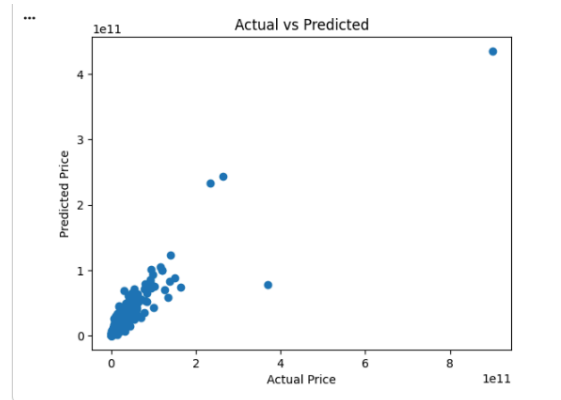
```
... Performance of the Tuned Model:
Mean Absolute Error: 0.24189826922988403
Mean Squared Error: 0.13007322338135857
Root Mean Squared Error: 0.3606566558118103
R2 Score: 0.8889843619954766
```

Gambar 4. 11 Hasil Evaluasi model

Berdasarkan Gambar 4.11, diperoleh nilai Mean Absolute Error (MAE) sebesar 0.2419, Mean Squared Error (MSE) sebesar 0.1301, Root Mean Squared Error

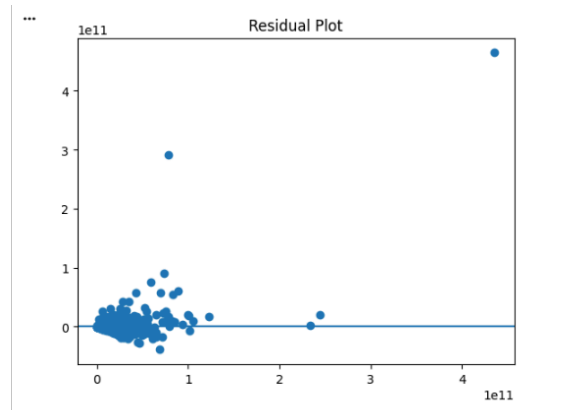
(RMSE) sebesar 0.3607, serta nilai R^2 Score sebesar 0.8889. Nilai error yang relatif kecil menunjukkan bahwa model memiliki tingkat kesalahan prediksi yang rendah, sedangkan nilai R^2 yang mendekati 1 menunjukkan bahwa model mampu menjelaskan sebagian besar variasi data harga rumah dengan baik.

Untuk memahami performa model secara lebih mendalam, dilakukan analisis visual menggunakan grafik perbandingan nilai aktual dan prediksi serta residual plot.



Gambar 4. 12 Grafik *Actual vs Predicted*

Berdasarkan Gambar 4.12, terlihat bahwa sebagian besar titik data berada di sekitar garis diagonal, yang menunjukkan bahwa nilai prediksi yang dihasilkan model cukup mendekati nilai aktual. Meskipun terdapat beberapa titik yang menyimpang, terutama pada nilai harga yang tinggi, secara umum model mampu mengikuti pola data dengan baik.



Gambar 4. 13 Grafik *Residual Plot*

46 Berdasarkan Gambar 4.13, terlihat bahwa residual tersebar di sekitar garis nol
tanpa membentuk pola tertentu. Hal ini menunjukkan bahwa model tidak mengalami
bias yang signifikan dan mampu melakukan prediksi dengan cukup baik. Namun,
39 terdapat beberapa residual yang cukup besar pada nilai tertentu, yang
mengindikasikan adanya data ekstrem atau outlier yang sulit diprediksi oleh model.
Keberadaan outlier dalam penelitian ini tidak dihilangkan karena data tersebut masih
merepresentasikan kondisi nyata dalam pasar properti, di mana harga rumah dapat
bervariasi secara signifikan akibat faktor lokasi, fasilitas, maupun karakteristik
khusus lainnya. Menghapus outlier justru berpotensi mengurangi kemampuan model
dalam memahami variasi data secara menyeluruh serta menurunkan kemampuan
generalisasi model terhadap data baru. Oleh karena itu, outlier tetap dipertahankan
106 agar model dapat belajar dari seluruh variasi data yang ada dan menghasilkan
prediksi yang lebih realistis.

31 Dengan demikian, hasil modeling menunjukkan bahwa model XGBoost yang
telah dioptimasi mampu menghasilkan performa yang baik dalam memprediksi harga
rumah, serta memiliki kemampuan generalisasi yang cukup baik terhadap data uji.
43

4. Hasil Perbandingan

6 Pada tahap ini dilakukan perbandingan kinerja antara dua model yang digunakan
dalam penelitian, yaitu model utama *XGBoost Regressor* yang telah dioptimasi
menggunakan *GridSearchCV* dan model pembanding *Random Forest Regressor*.
30 Tujuan dari perbandingan ini adalah untuk mengetahui model mana yang memiliki
performa terbaik dalam memprediksi harga rumah berdasarkan dataset yang
digunakan.

Model *XGBoost* dibangun menggunakan pipeline yang telah terintegrasi dengan
proses preprocessing dan dilakukan optimasi hyperparameter menggunakan metode
GridSearchCV. Sementara itu, model *Random Forest* dilatih ulang menggunakan
dataset yang sama sebagai pembanding untuk memastikan evaluasi dilakukan secara
adil dan konsisten.

Hasil kinerja kedua model dibandingkan menggunakan beberapa metrik evaluasi, yaitu *Mean Absolute Error (MAE)*, *Mean Squared Error (MSE)*, *Root Mean Squared Error (RMSE)*, dan *R² Score*.

Tabel 4. 1 Hasil Perbandingan

Metrik	XGBoost + GridSearchCV	Random Forest + Optuna
MAE	0.2418	0.2428
MSE	0.1300	0.1381
RMSE	0.3606	0.3716
R ² Score	0.8889	0.8822

Berdasarkan Tabel 4.1, dapat dilihat bahwa model *XGBoost* yang dioptimasi menggunakan *GridSearchCV* menunjukkan performa yang lebih baik dibandingkan model *Random Forest* pada seluruh metrik evaluasi. Nilai *R² Score* pada model *XGBoost* lebih tinggi, yaitu sebesar 0.8889 dibandingkan dengan 0.8822 pada model *Random Forest*, yang menunjukkan bahwa model *XGBoost* mampu menjelaskan variasi data dengan lebih baik.

Selain itu, nilai error pada model *XGBoost* juga lebih rendah dibandingkan *Random Forest*. Nilai *Mean Absolute Error (MAE)*, *Mean Squared Error (MSE)*, dan *Root Mean Squared Error (RMSE)* pada model *XGBoost* menunjukkan tingkat kesalahan prediksi yang lebih kecil. Hal ini menunjukkan bahwa hasil prediksi yang dihasilkan oleh model *XGBoost* lebih mendekati nilai aktual dibandingkan dengan model *Random Forest*.

Perbedaan performa ini menunjukkan bahwa algoritma *XGBoost* memiliki kemampuan yang lebih baik dalam menangkap pola hubungan antara variabel fitur dan target, terutama pada data dengan karakteristik yang kompleks. Selain itu, proses optimasi hyperparameter menggunakan *GridSearchCV* juga berkontribusi dalam meningkatkan performa model secara signifikan.

Dengan demikian, dapat disimpulkan bahwa model *XGBoost* yang telah dioptimasi merupakan model yang paling optimal dalam penelitian ini dan lebih unggul dibandingkan model *Random Forest* dalam memprediksi harga rumah. Model ini selanjutnya digunakan sebagai model utama dalam proses analisis dan implementasi pada tahap berikutnya.

5. Implementasi

Pada tahap ini, model yang telah dibangun dan dioptimasi diimplementasikan dalam bentuk aplikasi berbasis web menggunakan framework Streamlit. Implementasi ini bertujuan untuk mempermudah pengguna dalam melakukan prediksi harga rumah secara interaktif berdasarkan parameter yang dimasukkan.

Aplikasi yang dikembangkan memiliki antarmuka yang sederhana dan mudah digunakan, di mana pengguna dapat memasukkan berbagai informasi terkait properti seperti wilayah (*district*), kota (*city*), jumlah kamar tidur, jumlah kamar mandi, kapasitas mobil (*carport*), luas tanah, serta luas bangunan.

Gambar 4. 14 Tampilan Awal Aplikasi

Berdasarkan Gambar 4.14, aplikasi menyediakan form input yang memungkinkan pengguna memasukkan data properti secara lengkap. Setiap input telah disesuaikan dengan fitur yang digunakan dalam model, sehingga memastikan kesesuaian antara data input dengan data pelatihan.

Setelah pengguna memasukkan data dan menekan tombol prediksi, sistem akan memproses data tersebut menggunakan model XGBoost yang telah disimpan dalam bentuk file *.pkl*. Proses ini mencakup tahap preprocessing secara otomatis melalui pipeline yang telah dibangun sebelumnya.

Masukkan Data Rumah

Wilayah	Kota
Citra garden	Jakarta Barat
Kamar Tidur	Kamar Mandi
2	4
Kapasitas Mobil (Garasi)	
2	
Luas Tanah (m ²)	
250.00	
Luas Bangunan (m ²)	
350.00	

Prediksi Harga

Prediksi Harga Terbaik:

Rp 6,206,508,032

Estimasi Harga:

Rp 5,585,857,024 - Rp 6,827,159,040

Gambar 4. 15 Tampilan Hasil Prediksi

Berdasarkan Gambar 4.15, aplikasi menampilkan hasil prediksi dalam dua bentuk, yaitu nilai estimasi harga utama (*prediksi harga terbaik*) serta rentang estimasi harga. Rentang ini diberikan untuk memberikan gambaran variasi harga yang mungkin terjadi di pasar.

Hasil prediksi yang ditampilkan menunjukkan bahwa aplikasi mampu memberikan estimasi harga rumah secara cepat dan responsif berdasarkan input pengguna. Dengan demikian, implementasi sistem ini dapat digunakan sebagai alat bantu dalam pengambilan keputusan, baik bagi calon pembeli, penjual, maupun pihak terkait di bidang properti.

B. Pembahasan

1. Model Paling Optimal

Berdasarkan hasil pemodelan dan evaluasi yang telah dilakukan, model yang paling optimal dalam penelitian ini adalah algoritma XGBoost Regressor yang telah dioptimasi menggunakan metode GridSearchCV. Hal ini ditunjukkan oleh nilai kinerja model yang lebih baik dibandingkan dengan model pembanding Random Forest pada seluruh metrik evaluasi.

Model XGBoost menghasilkan nilai Mean Absolute Error (MAE) sebesar 0.2418, Mean Squared Error (MSE) sebesar 0.1300, Root Mean Squared Error (RMSE) sebesar 0.3606, serta nilai R² Score sebesar 0.8889. Nilai-nilai tersebut menunjukkan

3 bahwa model memiliki tingkat kesalahan prediksi yang rendah serta kemampuan yang tinggi dalam menjelaskan variasi data harga rumah.

116 Jika dibandingkan dengan model Random Forest, XGBoost menunjukkan
83 performa yang lebih unggul dengan nilai error yang lebih kecil serta nilai R^2 yang
8 lebih tinggi. Hal ini mengindikasikan bahwa XGBoost lebih mampu menangkap pola hubungan antara variabel input dan target, terutama pada data dengan karakteristik yang kompleks dan memiliki variasi yang tinggi.

37 Selain itu, proses optimasi hyperparameter menggunakan GridSearchCV juga berperan penting dalam meningkatkan performa model. Pemilihan kombinasi parameter yang optimal memungkinkan model bekerja secara lebih efisien dan mengurangi risiko overfitting maupun underfitting.

153 Dengan demikian, dapat disimpulkan bahwa model XGBoost yang telah
108 dioptimasi merupakan model yang paling optimal dalam penelitian ini dan layak digunakan sebagai model utama dalam sistem prediksi harga rumah yang dikembangkan.

2. Evaluasi Hasil Modeling

24 Tujuan utama penelitian ini adalah untuk menghasilkan model prediksi harga rumah yang akurat dan dapat digunakan sebagai alat bantu dalam pengambilan keputusan di bidang properti. Berdasarkan hasil evaluasi model, algoritma XGBoost yang telah dioptimasi mampu menghasilkan nilai R^2 Score sebesar 0.8889 serta nilai error yang relatif rendah, yang menunjukkan bahwa model memiliki tingkat akurasi yang tinggi dalam memprediksi harga rumah.

28 Hasil ini menunjukkan bahwa model telah mampu memenuhi tujuan bisnis yang ditetapkan, yaitu memberikan estimasi harga yang mendekati nilai aktual. Dengan adanya model ini, pengguna seperti calon pembeli, penjual, maupun investor properti dapat memperoleh gambaran harga rumah secara lebih objektif dan berbasis data, sehingga dapat membantu dalam proses pengambilan keputusan.

146 Selain evaluasi numerik, analisis visual melalui grafik *actual vs predicted*
85 menunjukkan bahwa sebagian besar data berada di sekitar garis diagonal, yang menandakan bahwa hasil prediksi model memiliki kedekatan dengan nilai aktual.

36 Sementara itu, grafik *residual plot* menunjukkan bahwa residual tersebar secara acak di sekitar garis nol tanpa membentuk pola tertentu, yang mengindikasikan bahwa model tidak mengalami bias yang signifikan.

3. Kesesuaian Hasil

Pada tahap awal penelitian, tujuan yang ingin dicapai adalah membangun model prediksi harga rumah yang mampu memanfaatkan data historis untuk menghasilkan estimasi harga yang akurat. Selain itu, model diharapkan dapat mengatasi permasalahan distribusi data yang tidak merata serta meningkatkan performa prediksi melalui teknik optimasi.

138 Berdasarkan hasil yang diperoleh, model yang dibangun telah mampu memenuhi ekspektasi tersebut. Transformasi logaritmik yang diterapkan berhasil mengatasi permasalahan distribusi data yang tidak normal, sementara optimasi menggunakan *GridSearchCV* mampu meningkatkan performa model secara signifikan. Hal ini ditunjukkan dengan nilai evaluasi yang baik serta kemampuan model dalam mengikuti pola data.

40 Dengan demikian, hasil penelitian ini telah sesuai dengan tujuan yang dirumuskan pada tahap awal dan menunjukkan bahwa pendekatan yang digunakan efektif dalam menyelesaikan permasalahan prediksi harga rumah.

4. Validitas dan Generalisasi Model

63 Validitas model dalam penelitian ini ditunjukkan melalui penggunaan teknik pembagian data menjadi data pelatihan dan data pengujian dengan rasio 80:20. Model dilatih menggunakan data pelatihan dan diuji pada data yang tidak pernah dilihat sebelumnya, sehingga hasil evaluasi mencerminkan kemampuan model dalam melakukan prediksi pada data baru.

3 Nilai R^2 Score yang tinggi serta error yang relatif rendah menunjukkan bahwa model memiliki kemampuan generalisasi yang baik terhadap data uji. Selain itu, penggunaan teknik pipeline memastikan bahwa proses preprocessing dilakukan secara konsisten, sehingga mengurangi risiko kesalahan dalam penerapan model.

105 Namun demikian, hasil penelitian ini masih memiliki keterbatasan, yaitu dataset yang digunakan hanya mencakup wilayah Jakarta. Oleh karena itu, model mungkin belum sepenuhnya dapat digeneralisasikan untuk wilayah lain dengan karakteristik pasar yang berbeda. Untuk meningkatkan generalisasi, diperlukan penambahan dataset dari wilayah yang lebih luas pada penelitian selanjutnya.

3 Hasil analisis residual juga menunjukkan bahwa tidak terdapat pola tertentu pada distribusi kesalahan, sehingga memperkuat bahwa model memiliki kemampuan generalisasi yang baik terhadap data uji. Meskipun terdapat beberapa nilai residual yang cukup besar akibat data ekstrem, hal tersebut tetap mencerminkan kondisi nyata dalam dataset dan tidak mengurangi validitas model secara keseluruhan.

5. Rekomendasi Penerapan Model

28 Model yang telah dikembangkan dalam penelitian ini berpotensi untuk diterapkan dalam sistem prediksi harga rumah berbasis web, seperti aplikasi yang telah diimplementasikan menggunakan Streamlit. 81 Aplikasi ini memungkinkan pengguna untuk memasukkan data properti secara langsung dan memperoleh hasil prediksi harga secara real-time.

40 Penerapan model ini dapat dimanfaatkan oleh berbagai pihak, seperti agen properti, calon pembeli, maupun investor, untuk membantu dalam menentukan harga jual atau estimasi nilai properti secara lebih objektif. Selain itu, model ini juga dapat dikembangkan lebih lanjut dengan menambahkan fitur lain seperti lokasi geografis yang lebih detail, kondisi lingkungan, serta tren pasar properti.

13 Untuk pengembangan selanjutnya, disarankan untuk menggunakan dataset yang lebih besar dan beragam serta mencoba algoritma lain sebagai pembanding guna meningkatkan performa model. Selain itu, integrasi dengan sistem berbasis API atau platform digital lainnya dapat meningkatkan kemudahan akses dan penggunaan model dalam skala yang lebih luas. 14

BAB V

PENUTUP

A. Kesimpulan

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa penerapan *hyperparameter tuning* menggunakan metode *Grid Search Cross Validation (GridSearchCV)* pada algoritma *XGBoost Regressor* mampu meningkatkan akurasi model prediksi harga rumah. Hasil evaluasi menunjukkan bahwa model *XGBoost* dengan *GridSearchCV* memperoleh nilai R^2 Score sebesar 0.8889, lebih tinggi dibandingkan model *Random Forest* dengan *Optuna* sebesar 0.8822, sehingga terjadi peningkatan akurasi sebesar 0.0067.

Selain itu, optimasi menggunakan *GridSearchCV* juga mampu menurunkan selisih antara harga aktual dan harga hasil prediksi. Hal tersebut ditunjukkan melalui penurunan nilai *Mean Absolute Error (MAE)* sebesar 0.0010, *Mean Squared Error (MSE)* sebesar 0.0081, dan *Root Mean Squared Error (RMSE)* sebesar 0.0110 dibandingkan model *Random Forest* dengan *Optuna*. Penurunan nilai error tersebut menunjukkan bahwa hasil prediksi model menjadi lebih mendekati nilai harga rumah sebenarnya serta mampu meminimalkan kesalahan prediksi secara lebih optimal.

Dengan demikian, penelitian ini berhasil menjawab rumusan masalah mengenai seberapa besar peningkatan akurasi model dan seberapa besar penurunan selisih antara nilai aktual dan hasil prediksi setelah dilakukan optimasi hyperparameter menggunakan *GridSearchCV* pada algoritma *XGBoost Regressor*. Selain itu, model yang dihasilkan juga berhasil diimplementasikan ke dalam aplikasi berbasis web menggunakan *Streamlit* sehingga dapat dimanfaatkan sebagai alat bantu prediksi harga rumah yang lebih akurat, stabil, dan efektif dalam mendukung pengambilan keputusan di bidang properti.

B. Saran

Berdasarkan hasil penelitian yang telah dilakukan, disarankan agar pengembangan selanjutnya dapat memperluas penggunaan dataset dengan cakupan wilayah yang lebih beragam guna meningkatkan kemampuan generalisasi model, serta menambahkan fitur-fitur yang lebih kompleks seperti lokasi geografis, akses fasilitas, dan kondisi lingkungan agar akurasi prediksi semakin meningkat. Selain itu, sistem yang telah dibangun dapat dikembangkan lebih lanjut dengan peningkatan tampilan antarmuka (*UI/UX*), penambahan fitur visualisasi data, serta integrasi dengan API properti untuk memperoleh data secara real-time. Penelitian berikutnya juga dapat mengeksplorasi penggunaan algoritma lain atau metode optimasi tambahan untuk membandingkan performa model. Lebih lanjut, model dan aplikasi yang telah dikembangkan diharapkan dapat diterapkan secara nyata pada platform atau organisasi di bidang properti sebagai alat bantu pengambilan keputusan yang berbasis data.