Jurnal Nasional Teknologi Komputer
Vol.6; No: 1; Januari 2026
Website: publikasi.hawari.id/index.php/jnastek
E-ISSN: 2808-4845; P-ISSN: 2808-7801

Makalah Penelitian

Perbandingan Algoritma XGBoost dan CatBoost dalam Prediksi Harga Rumah Berdasarkan Data Perumahan di Jabodetabek

Muh. Agum Nur Efendi[1], Imam Suharjo[2]

[1,2]Informatika, Fakultas Teknologi Informasi, Universitas Mercu Buana Yogyakarta
[1]201110040@student.mercubuana-yogya.ac.id*, [2]imam@mercubuana-yogya.ac.id

Corresponding Author: write name of corresponding author

ABSTRACT

This study aims to build and compare house price prediction models in Jabodetabek area using XGBoost and CatBoost machine learning algorithms. The data used are secondary data on residential properties obtained from real estate websites and public datasets, totaling 9,991 items. The variables used include numerical features such as land area, building area, number of bedrooms, number of bathrooms, garage, and building age, as well as categorical features such as location, type of certificate, building condition, and facilities.The target variable of this study is house prices in rupiah. The data pre-processing stages include data cleaning, unit conversion and price format, handling missing values using medians, and feature engineering by extracting city information from property locations. The dataset was divided into training and test data with an 80:20 ratio. The XGBoost model used one-hot encoding for categorical features and standardized numerical features, while CatBoost utilized native capabilities for handling categorical features. The models used logarithmic transformation on the target variable. The results showed that CatBoost performed superior to XGBoost. The CatBoost model produced an MAE of IDR 126,169,305, an RMSE of IDR 212,418,771, and an R^2 of 0.9907 on the test data, while XGBoost produced an MAE of IDR 239,321,909, an RMSE of IDR 393,835,569, and an R^2 of 0.9682. CatBoost's error distribution and MAPE were more stable, with a 91.80% of predictions 1,835 items having errors below 10%. In contrast, XGBoost achieved only 72.59% of predictions 1,451 items with errors below 10%.These findings indicate that CatBoost is effective for house price prediction, especially on datasets with a predominance of categorical features such as residential property data.

Keywords: *House price prediction, XGBoost, CatBoost, Machine Learning, Jabodetabek Property*

ABSTRAK

Penelitian ini bertujuan untuk membangun dan membandingkan model prediksi harga rumah di wilayah Jabodetabek menggunakan algoritma machine learning XGBoost dan CatBoost. Data yang digunakan merupakan data sekunder properti perumahan yang diperoleh dari situs real estat dan dataset publik, dengan total 9.991 data. Variabel yang digunakan meliputi fitur numerik seperti luas tanah, luas bangunan, jumlah kamar, jumlah kamar mandi, garasi, dan usia bangunan, serta fitur kategorikal seperti lokasi, jenis sertifikat, kondisi bangunan, dan fasilitas. Variabel target dalam penelitian ini adalah harga rumah dalam satuan rupiah. Tahapan pra-pemrosesan data meliputi pembersihan data, konversi satuan dan format harga, penanganan nilai hilang menggunakan median, serta rekayasa fitur dengan mengekstraksi informasi kota dari lokasi properti. Dataset kemudian dibagi menjadi data latih dan data uji dengan rasio 80:20. Pada model XGBoost, dilakukan One-Hot Encoding untuk fitur kategorikal dan standardisasi fitur numerik, sedangkan CatBoost memanfaatkan kemampuan native dalam menangani fitur kategorikal. Kedua model menggunakan transformasi logaritmik pada variabel target. Hasil evaluasi menunjukkan bahwa CatBoost memiliki performa yang lebih unggul dibandingkan XGBoost. Model CatBoost menghasilkan nilai MAE sebesar Rp 126.169.305, RMSE sebesar Rp 212.418.771, dan R^2 sebesar 0,9907 pada data uji, sedangkan XGBoost menghasilkan MAE Rp 239.321.909, RMSE Rp 393.835.569, dan R^2 sebesar 0,9682. Distribusi error dan MAPE CatBoost juga lebih stabil, dengan 91,80% prediksi memiliki error di bawah 10%, yakni dengan jumlah 1,835 data, sedangkan MAPE XGBoost memiliki prediksi error dibawah 10% berjumlah 72,59% dari total data yakni berjumlah 1.451 data. Temuan ini menunjukkan bahwa CatBoost sangat efektif untuk prediksi harga rumah, khususnya pada dataset dengan dominasi fitur kategorikal seperti data properti perumahan.

Kata Kunci: *Prediksi harga rumah, XGBoost, CatBoost, Machine Learning, Properti Jabodetabek*

1. Pendahuluan

Dua algoritma yang banyak digunakan dalam pemodelan prediksi harga rumah adalah Extreme Gradient Boosting (XGBoost) dan Categorical Boosting (CatBoost). XGBoost memperkenalkan peningkatan dalam hal regularisasi, komputasi paralel, dan teknik optimasi pohon, sehingga menghasilkan kinerja dan skalabilitas yang lebih unggul, sementara CatBoost memperkenalkan teknik baru seperti ordered boosting dan feature combination trees [1].

Meskipun keduanya merupakan algoritma gradient boosting yang mengembangkan rangkaian pohon keputusan secara berurutan, di mana setiap pohon bertujuan mengoreksi kesalahan yang dihasilkan oleh pohon terdahulu [2], karakteristik internal dan cara mereka memproses data berbeda secara signifikan. XGBoost menggunakan pendekatan tree boosting berbasis struktur CART yang dioptimalkan dengan teknik regularisasi dan terbukti unggul dalam pengolahan data tabular [3], sedangkan CatBoost mengakomodasi variabel kategorikal secara langsung, bekerja secara komputasional efisien, dan cenderung tidak mudah mengalami overfitting [4]. Perbedaan ini berpotensi menghasilkan performa yang berbeda pada data perumahan dengan karakteristik yang heterogen.

Perkembangan teknologi data dan komputasi dalam dekade terakhir telah mendorong munculnya berbagai pendekatan berbasis machine learning yang mampu memproses data berukuran besar dan memiliki keragaman fitur yang tinggi serta memungkinkan komputer belajar secara otomatis dari pengalaman sebelumnya dan meningkatkan kinerjanya tanpa memerlukan pemrograman yang kompleks [5]. Beberapa penelitian di Indonesia dalam beberapa tahun terakhir, algoritma modern seperti Random Forest, Gradient Boosting, dan XGBoost semakin sering digunakan dalam penelitian prediksi harga rumah di seluruh dunia [6].

Berdasarkan kondisi tersebut, penelitian ini bertujuan membandingkan performa algoritma XGBoost dan CatBoost dalam memprediksi harga rumah di wilayah Jabodetabek guna memperoleh gambaran empiris efektivitas kedua metode dalam konteks data perumahan di Indonesia.

2. Tinjauan Pustaka

Pendekatan konvensional penilaian properti seperti regresi linier memiliki keterbatasan dalam menangkap hubungan non-linier, sehingga mendorong penggunaan machine learning. Penelitian [6] menunjukkan bahwa algoritma ensemble seperti Random Forest, Gradient Boosting, dan XGBoost lebih akurat dibandingkan regresi linier dalam memprediksi harga properti, membuka peluang pengembangan algoritma boosting yang lebih mutakhir dan presisi.

2.1 Machine Learning

Pada machine learning, metode supervised learning melatih model menggunakan data yang telah memiliki label agar mampu mempelajari pola dan menghasilkan prediksi yang sesuai [7].

2.2 XGBoost

XGBoost bekerja dengan mengintegrasikan sejumlah pohon keputusan sederhana (weak learners) secara iteratif, di mana setiap pohon yang dibangun secara khusus diarahkan untuk mengoreksi kesalahan prediksi yang dihasilkan oleh pohon pada tahap sebelumnya [8].

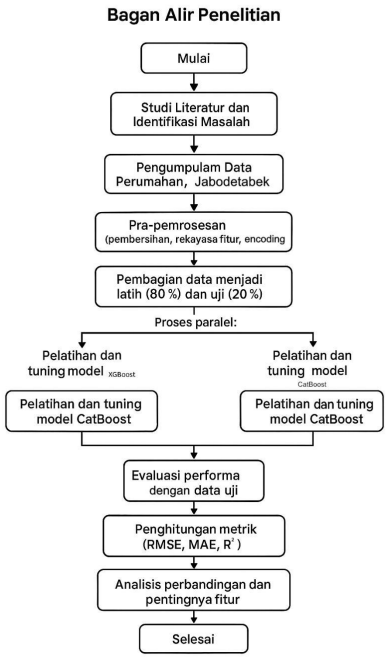
2.3 CatBoost

CatBoost menggunakan variasi dari target encoding yang disebut ordered encoding dengan menerapkan pendekatan ordered boosting untuk mencegah terjadinya kebocoran target (target leakage) [9].

Penelitian mengenai penerapan algoritma machine learning untuk prediksi harga properti telah banyak dilakukan secara global maupun di Indonesia. [10] menemukan bahwa XGBoost memberikan kinerja terbaik dalam prediksi harga rumah di wilayah urban dengan nilai R^2 sebesar 0,89, unggul dibandingkan Random Forest dan Support Vector Regression. [11] menunjukkan bahwa CatBoost memiliki RMSE 15% lebih rendah dibandingkan XGBoost pada dataset properti dengan banyak fitur kategorikal, menegaskan keunggulan CatBoost dalam menangani variabel kategorikal. [12] membuktikan bahwa penerapan XGBoost pada data properti Jakarta Selatan menghasilkan akurasi 87,3% dengan RMSE 128 juta, dengan luas bangunan, lokasi, dan jumlah kamar tidur sebagai fitur paling berpengaruh. [13] melaporkan bahwa XGBoost menghasilkan MAE terendah sebesar 97,5 juta dibandingkan LightGBM dan Random Forest, sehingga direkomendasikan untuk dataset properti Indonesia yang didominasi fitur numerik.

3. Bahan & Metode

3.1. Alur Penelitian



Gambar 1. Bagan Alir Penelitian

Pada Gambar 3.1, penelitian ini diawali dengan studi literatur dan identifikasi masalah, kemudian dilanjutkan dengan pengumpulan data perumahan dari JABODETABEK. Data yang diperoleh dipra-pemrosesan untuk memastikan kualitas data sebelum digunakan dalam pemodelan.

Dataset selanjutnya dibagi menjadi data latih (80%) dan data uji (20%). Proses pemodelan dilakukan secara paralel menggunakan algoritma XGBoost dan CatBoost, disertai dengan penyesuaian parameter untuk memperoleh performa optimal.

Evaluasi model dilakukan menggunakan data uji dengan metrik RMSE, MAE, dan R^2. Tahap akhir penelitian meliputi analisis perbandingan performa kedua model serta analisis feature importance untuk mengidentifikasi variabel yang paling berpengaruh terhadap prediksi.

3.2. Kajian Literatur dan Perumusan Masalah

Pada tahap awal penelitian, dilakukan pengumpulan teori dan kajian studi terdahulu yang berkaitan dengan prediksi harga properti, baik dari riset dalam negeri maupun luar negeri. Selain itu, dilakukan studi mendalam mengenai struktur, kelebihan, dan kekurangan algoritma XGBoost serta kajian khusus terhadap algoritma CatBoost, terutama dalam mekanismenya mengelola fitur kategorikal. Berdasarkan kajian tersebut, dirumuskan pertanyaan penelitian untuk menentukan algoritma yang lebih akurat dan efisien dalam memprediksi harga rumah di wilayah Jabodetabek, yaitu antara XGBoost dan CatBoost.

3.3. Pengumpulan Data

Data yang digunakan dalam penelitian ini merupakan data sekunder properti di wilayah Jabodetabek yang diperoleh dari situs real estat populer maupun dataset publik. Data tersebut mencakup fitur numerik berupa luas tanah, luas bangunan, jumlah kamar tidur, kamar mandi, garasi, serta usia bangunan. Selain itu, digunakan pula fitur kategorikal yang meliputi lokasi, jenis sertifikat, kondisi properti, dan fasilitas. Variabel target dalam penelitian ini adalah harga rumah yang dinyatakan dalam satuan rupiah.

3.4. Pra-Pemrosesan Data

Data dipersiapkan melalui beberapa tahapan, yaitu pembersihan data dari nilai hilang, data duplikat, dan pencilan. Selanjutnya, dilakukan rekayasa fitur dengan menambahkan fitur baru, seperti harga per meter persegi dan rasio luas bangunan terhadap luas tanah. Pada tahap berikutnya, fitur kategorikal diproses menggunakan metode encoding, di mana algoritma XGBoost menerapkan One-Hot Encoding atau Label Encoding, sedangkan CatBoost memanfaatkan mekanisme internalnya untuk menangani fitur kategorikal tanpa memerlukan encoding manual. Setelah seluruh proses prapemrosesan selesai, data dibagi menjadi data latih sebesar 80% dan data uji sebesar 20%.

3.5. Pelatihan Model

Pada tahap pemodelan, kedua algoritma dilatih menggunakan data latih dan dilakukan proses pencarian parameter terbaik (*hyperparameter tuning*). Pada algoritma XGBoost, pelatihan model dilakukan dengan menyesuaikan parameter seperti jumlah estimator, kedalaman pohon, dan *learning rate*. Sementara itu, algoritma CatBoost dilatih dengan memanfaatkan deteksi fitur kategorikal secara otomatis serta penyesuaian parameter yang meliputi jumlah iterasi, kedalaman pohon, dan parameter regularisasi.

3.6. Evaluasi Model

Kinerja model prediksi harga rumah dievaluasi menggunakan metrik regresi berupa Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), dan koefisien determinasi (R^2) guna memperoleh gambaran performa model secara komprehensif.

a. Mean Absolute Error (MAE) mengukur rata-rata selisih absolut antara nilai aktual dan hasil prediksi model dalam satuan yang sama dengan variabel target, yaitu rupiah [14]. Metrik ini dirumuskan sebagai berikut:

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

Pada rumus (1), dengan:
- y_i = nilai harga rumah aktual
- $\hat{y}_i$ =.nilai harga rumah hasil prediksi
- n = jumlah data

b. Root Mean Squared Error (RMSE) mengukur akar dari rata-rata kuadrat selisih antara nilai aktual dan prediksi, serta lebih sensitif terhadap kesalahan besar atau *outlier* [15]. Metrik ini dirumuskan sebagai berikut:

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

Pada rumus (2), Nilai RMSE yang lebih kecil menunjukkan bahwa model memiliki tingkat kesalahan prediksi yang lebih rendah.

c. Koefisien determinasi (R^2) mengukur kemampuan model dalam menjelaskan variasi data harga rumah, dengan nilai mendekati 1 menunjukkan performa yang semakin baik [16]. Metrik ini dirumuskan sebagai berikut:

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

Pada rumus (3), dengan:
- $\bar{y}$ = nilai harga rumah aktual

Nilai R^2 yang tinggi menunjukkan bahwa model memiliki kemampuan prediksi yang baik dan tingkat generalisasi yang tinggi terhadap data uji.

3.7. Analisis dan Kesimpulan

Tahap analisis dilakukan dengan membandingkan hasil metrik evaluasi dari kedua model, serta menilai fitur-fitur yang paling berpengaruh terhadap hasil prediksi pada masing-

masing algoritma. Selanjutnya, dibahas faktor-faktor yang menyebabkan perbedaan performa antar model, termasuk keunggulan dalam penanganan fitur kategorikal dan efisiensi waktu pelatihan. Berdasarkan hasil analisis tersebut, ditarik kesimpulan penelitian dan disusun rekomendasi untuk pengembangan serta penelitian selanjutnya.

4. Hasil

4.1. Pengumpulan Data

Penelitian ini menggunakan data sekunder properti perumahan di wilayah Jabodetabek, diperoleh dari situs real estat dan dataset publik. Dataset awal berisi 9.991 baris dengan fitur numerik seperti luas tanah, luas bangunan, jumlah kamar tidur, kamar mandi, garasi, dan usia bangunan, serta fitur kategorikal seperti lokasi, jenis sertifikat, kondisi bangunan, dan fasilitas pendukung. Variabel targetnya adalah harga rumah (dalam rupiah), yang digunakan untuk pemodelan prediksi menggunakan XGBoost dan CatBoost.

4.2. Pra-Pemrosesan Data

Tahap pra-pemrosesan data dilakukan untuk memastikan kualitas data sebelum digunakan dalam pembangunan model. Proses pra-pemrosesan meliputi pembersihan data, transformasi data, transformasi variabel, serta rekayasa fitur.

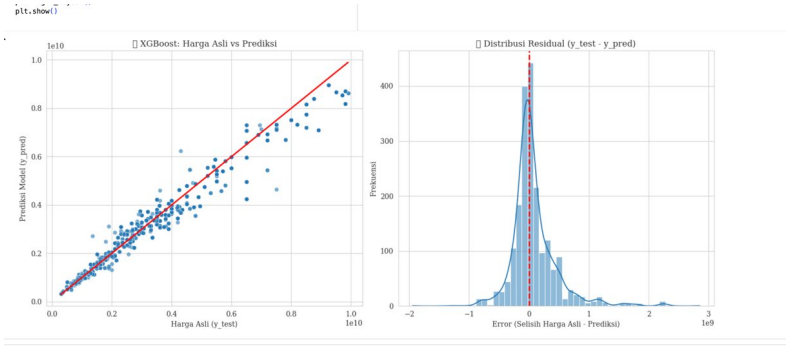
a. Data Cleaning
   Pembersihan data dilakukan dengan menghapus kolom yang tidak relevan seperti No, Judul, Agen, dan Kantor Agen. Data harga properti yang awalnya berbentuk teks (misalnya "Rp. 2,3 Milyar" atau "Rp. 800 Juta") dikonversi menjadi nilai numerik dalam rupiah. Selain itu, semua satuan ukuran, seperti Luas Tanah dan Luas Bangunan, diseragamkan ke meter persegi (m^2), dan nilai yang hilang diisi menggunakan median untuk mengurangi pengaruh outlier.
b. Rekayasa Fitur
   Dilakukan dengan mengekstraksi informasi kota dari kolom lokasi untuk membuat variabel kategorikal baru, Lokasi_Kota. Distribusi data menunjukkan sebagian besar properti berada di Tangerang (2.306 data), diikuti Jakarta Barat (2.067), Jakarta Timur (1.584), dan Jakarta Utara (1.249).
c. Pembagian Data
   Dataset dibagi menjadi 80% data latih, dan 20% data uji untuk keperluan pelatihan serta evaluasi model.

4.3. Pelatihan Model

Tahap pelatihan model dilakukan dengan melatih dua algoritma machine learning, yaitu XGBoost dan CatBoost, menggunakan data latih yang telah dipra-pemrosesan.

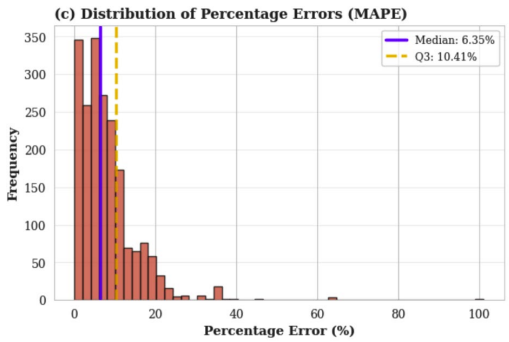
a. Pelatihan Model XGBoost
   Model pertama yang digunakan adalah XGBoost. Sebelum pelatihan, fitur kategorikal diubah menggunakan One-Hot Encoding, fitur numerik distandarisasi dengan Standard Scaler, dan variabel target harga rumah ditransformasi secara logaritmik untuk

menstabilkan varians. Performa model dievaluasi menggunakan MAE, RMSE, dan R^2 pada data latih dan uji. Hasil evaluasi menunjukkan MAE sebesar Rp 228.851.603 (latih) dan Rp 239.321.909 (uji), RMSE Rp 381.492.634 (latih) dan Rp 393.835.569 (uji), serta R^2 0,9683 (latih) dan 0,9682 (uji). Nilai MAE dan RMSE yang rendah serta R^2 mendekati 1 menunjukkan bahwa XGBoost mampu memprediksi harga rumah dengan akurat dan menjelaskan sebagian besar variasi data.



Gambar 2. Scatter Plot Harga Asli vs Prediksi XGBoost

Gambar 2 menunjukkan dua visualisasi evaluasi model XGBoost. Scatter plot di sisi kiri memperlihatkan hubungan antara harga rumah aktual (sumbu x) dan prediksi model (sumbu y), dengan garis merah diagonal mewakili prediksi sempurna. Sebagian besar titik berada di sekitar garis tersebut, menandakan model menangkap pola harga dengan baik, meski error sedikit meningkat pada properti berharga tinggi. Sisi kanan menampilkan distribusi residual yang menyerupai normal dan terpusat di nol, menunjukkan model tidak bias, meski ekor kanan yang lebih panjang mengindikasikan beberapa kasus overestimation pada harga tinggi.



Gambar 3. Distribusi MAPE Model XGBoost

Table 1. Statistik Error Model XGBoost

Statistik	Error Absolut (Miliar Rp)	Error Persentase (MAPE, %)
Minimum	0,001	0,06
Kuartil 1 (Q1, 25%)	0,052	3,24

Median (50%)	0,125	6,35
Rata-rata (Mean)	0,239	7,94
Kuartil 3 (Q3, 75%)	0,306	10,41
Persentil 90 (P90)	0,530	–
Maksimum	2,852	101,09

Table 2. Distribusi Akurasi Prediksi Berdasarkan MAPE XGBoost

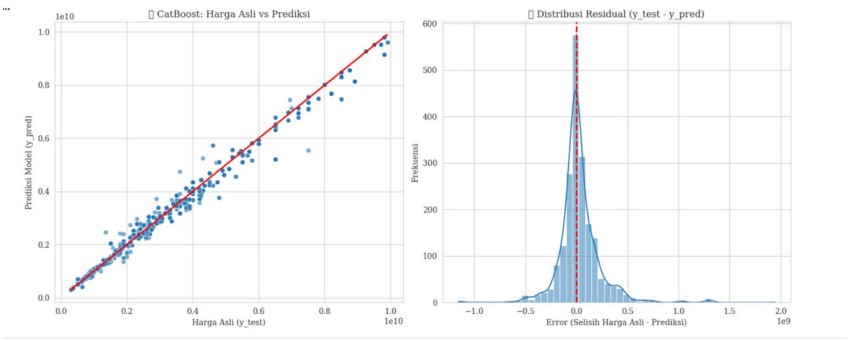
Kategori	Rentang Error (%)	Jumlah Prediksi	Persentase
Excellent	0 – 10	1.451	72,59%
Good	10 – 50	543	27,16%
Fair	50 – 100	4	0,20%
Poor	100 – 200	1	0,05%
Very Poor	> 200	0	0,00%

Gambar 3, Tabel 1 menunjukkan bahwa model XGBoost memiliki distribusi error yang terkonsentrasi pada nilai rendah. Berdasarkan **Tabel 1**, nilai median MAPE sebesar **6,35%** dan kuartil ketiga sebesar **10,41%** menunjukkan bahwa **75% prediksi memiliki error di bawah 10,41%**. Selain itu, nilai persentil ke-90 (P90) pada error absolut sebesar **Rp 0,530 miliar** mengindikasikan bahwa **90% prediksi memiliki selisih harga di bawah Rp 0,5 miliar**.

Selanjutnya, **Tabel 2** menunjukkan bahwa sebanyak **1.451 prediksi (72,59%)** berada pada kategori *excellent* (MAPE ≤ 10%) dan **543 prediksi (27,16%)** pada kategori *good* (10–50%). Prediksi dengan error tinggi jumlahnya sangat terbatas, yaitu hanya **5 prediksi (0,25%)** dengan MAPE di atas **50%**, yang menunjukkan bahwa prediksi ekstrem hanya terjadi pada sebagian kecil data.

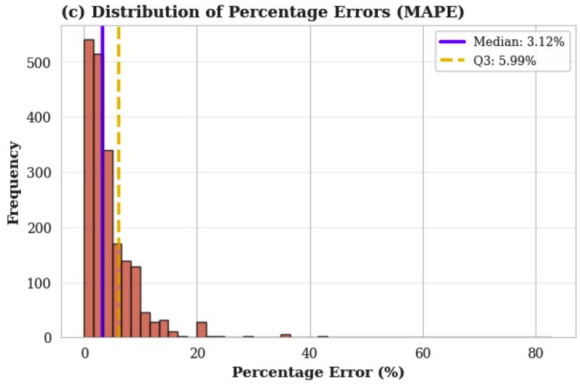
b. Pelatihan Model CatBoost

Model kedua yang digunakan adalah CatBoost, yang mampu menangani fitur kategorikal secara langsung tanpa perlu encoding manual. Model dilatih dengan 800 iterasi, depth 6, subsample 0,85, dan regularisasi L2 sebesar 3, serta variabel target harga rumah ditransformasi logaritmik untuk menstabilkan varians. Evaluasi menggunakan MAE, RMSE, dan R^2 menunjukkan performa yang sangat baik, dengan MAE Rp 117.976.192 (latih) dan Rp 126.169.305 (uji), RMSE Rp 195.530.045 (latih) dan Rp 212.418.771 (uji), serta R^2 0,9917 (latih) dan 0,9907 (uji). Nilai error yang lebih rendah dibanding XGBoost, CatBoost mampu memprediksi harga properti Jabodetabek dengan nilai 91,80% berada di rentang error 0 sampai 10 persen dari harga asli, 8,15% berada di rentang error 10 sampai 50 persen dari harga asli, dan 0,05 persen berada di rentang error 50 sampai 100 persen dari harga aslinya



Gambar 4. Scatter Plot Harga Asli vs Prediksi Catboost

Gambar 4 memperlihatkan evaluasi model CatBoost. Scatter plot di sisi kiri menunjukkan sebaran prediksi yang lebih rapat mengikuti garis diagonal dibanding XGBoost, menandakan akurasi dan konsistensi yang lebih tinggi, terutama pada harga menengah hingga tinggi. Distribusi residual di sisi kanan lebih simetris dan terpusat di nol, menunjukkan kesalahan prediksi lebih kecil, stabil, dan risiko overfitting lebih rendah.



Gambar 5. Distribusi MAPE Model Catboost

Tabel 3. Statistik Error Model Catboost

Statistik	Error Absolut (Miliar Rp)	Error Persentase (MAPE, %)
Minimum	0,000	0,00
Kuartil 1 (Q1, 25%)	0,026	1,44
Median (50%)	0,069	3,12
Rata-rata (Mean)	0,126	4,51
Kuartil 3 (Q3, 75%)	0,171	5,99
Persentil 90 (P90)	0,316	–
Maksimum	1,945	82,93

Tabel 4. Distribusi Akurasi Prediksi Berdasarkan MAPE CatBoost

Kategori	Rentang Error (%)	Jumlah Prediksi	Persentase
Excellent	0 – 10	1.835	91,80%
Good	10 – 50	163	8,15%

Fair	50 – 100	1	0,05%
Poor	100 – 200	0	0,00%
Very Poor	> 200	0	0,00%

Berdasarkan **Tabel 3**, model CatBoost menunjukkan distribusi error yang rendah dan terkonsentrasi pada nilai kecil, dengan median *Mean Absolute Percentage Error* (MAPE) sebesar **3,12%** dan kuartil ketiga sebesar **5,99%**, yang berarti **75% prediksi memiliki error di bawah 6%.** Selain itu, median error absolut sebesar **Rp 0,069 miliar** dan nilai persentil ke-90 (P90) sebesar **Rp 0,316 miliar** menunjukkan bahwa sebagian besar prediksi memiliki selisih harga yang relatif kecil.

Selanjutnya, **Tabel 4** menunjukkan bahwa mayoritas prediksi berada pada kategori *excellent*, yaitu sebanyak **1.835 prediksi (91,80%)** dengan MAPE ≤ **10%**, diikuti oleh **163 prediksi (8,15%)** pada kategori *good*. Prediksi dengan error tinggi hampir tidak ditemukan, dengan hanya **1 prediksi (0,05%)** yang memiliki MAPE antara **50%–100%,** dan **tidak terdapat prediksi dengan error di atas 100%.**

4.4. Analisis Perbandingan

Table 5. Perbandingan Kinerja Model XGBoost dan CatBoost

Metrik	XGBoost	CatBoost
Median Error Absolut (Miliar Rp)	0,125	**0,069**
Median MAPE (%)	6,35	**3,12**
Kuartil 3 MAPE (Q3, %)	10,41	**5,99**
P90 Error Absolut (Miliar Rp)	0,530	**0,316**
Prediksi Excellent (≤ 10%)	72,59% (1.451)	**91,80% (1.835)**
Prediksi Error > 50%	0,25% (5)	**0,05% (1)**
Prediksi Error > 100%	0,05% (1)	**0,00% (0)**

Berdasarkan **Tabel 5**, model **CatBoost menunjukkan performa yang lebih unggul dibanding XGBoost** dalam memprediksi harga rumah di wilayah Jabodetabek. Keunggulan tersebut terlihat dari **median MAPE yang lebih rendah (3,12% dibanding 6,35%),** error absolut yang lebih kecil, serta proporsi prediksi dengan akurasi tinggi yang lebih besar, di mana **91,80%** prediksi CatBoost berada pada kategori *excellent* dibanding **72,59%** pada XGBoost.

Selain itu, CatBoost juga menunjukkan **jumlah prediksi ekstrem yang lebih sedikit**, dengan hanya **0,05%** prediksi memiliki MAPE di atas **50%** dan **tidak terdapat prediksi dengan error di atas 100%**, sedangkan pada XGBoost masih ditemukan prediksi ekstrem meskipun dalam jumlah terbatas. Keunggulan ini berkaitan dengan kemampuan CatBoost dalam menangani fitur kategorikal secara langsung melalui *ordered boosting* dan *target encoding*, sehingga mampu mengurangi risiko *data leakage* dan *overfitting*. Distribusi error yang lebih rapat dan stabil ini menegaskan efektivitas CatBoost pada dataset dengan dominasi fitur kategorikal, sejalan dengan temuan pada penelitian sebelumnya.

5. Kesimpulan

Berdasarkan hasil analisis dan perbandingan kinerja model XGBoost dan CatBoost dalam memprediksi harga rumah di wilayah Jabodetabek, dapat disimpulkan bahwa kedua algoritma mampu memberikan performa prediksi yang baik. Namun, **CatBoost menunjukkan kinerja yang lebih unggul dan stabil dibandingkan XGBoost pada seluruh metrik evaluasi yang digunakan.**

Keunggulan CatBoost ditunjukkan oleh **median error absolut yang lebih rendah** sebesar **Rp 0,069 miliar** dibandingkan **Rp 0,125 miliar** pada XGBoost, serta **median MAPE sebesar 3,12%**, yang hampir setengah dari nilai XGBoost (**6,35%**). Selain itu, distribusi kesalahan CatBoost lebih terkonsentrasi pada nilai rendah, tercermin dari **kuartil ketiga MAPE (Q3) sebesar 5,99%**, sementara XGBoost mencapai **10,41%.** Dari sisi stabilitas, **90% prediksi CatBoost memiliki error absolut di bawah Rp 0,316 miliar**, lebih kecil dibandingkan XGBoost yang mencapai **Rp 0,530 miliar**.

Proporsi prediksi dengan tingkat akurasi tinggi (*excellent*, MAPE ≤ 10%) pada CatBoost mencapai **91,80%,** jauh melampaui XGBoost yang hanya mencapai **72,59%**. Selain itu, CatBoost hampir tidak menghasilkan prediksi ekstrem, dengan **tidak terdapat prediksi dengan error di atas 100%,** menunjukkan tingkat risiko kesalahan yang lebih rendah.

Secara keseluruhan, hasil penelitian ini menegaskan bahwa **CatBoost lebih efektif dan andal dalam memprediksi harga rumah pada dataset dengan dominasi fitur kategorikal,** terutama karena kemampuannya menangani fitur kategorikal secara langsung melalui *ordered boosting* yang mampu mengurangi *overfitting* dan *data leakage*. Dengan demikian, CatBoost direkomendasikan sebagai model yang lebih sesuai untuk permasalahan prediksi harga properti di wilayah Jabodetabek.

REFERENSI

[1] A. F. Limas Ptr, M. M. Siregar, dan I. Daniel, "Analysis of Gradient Boosting, XGBoost, and CatBoost on Mobile Phone Classification," Journal of Computer Networks, Architecture and High Performance Computing, vol. 6, no. 2, pp. 661–670, Apr. 2024.

[2] M. R. P. Putra, S. Juwariyah, M. Ridwan, dan R. Marco, "Optimasi Prediksi Kelayakan Pinjaman dengan Teknik Resampling dan Algoritma Boosting," Komputika: Jurnal Sistem Komputer, vol. 14, no. 1, pp. 41–51, Apr. 2025. doi: 10.34010/komputika.v14i2.15485.

[3] S. N. Ruscikasani, R. R. N. Oktalivia, F. R. Putra, A. J. Wahidin, B. Rahmatullah, dan I. Kurniawati, "Prediksi Pembelian E-Commerce Menggunakan XGBoost Berbasis Perilaku Sesi Pengguna," Journal of Artificial Intelligence and Digital Business (RIGGS), vol. 4, no. 4, pp. 5666–5672, 2025.

[4] S. Aurahmana dan U. Mahmudah, "Prediksi Perilaku Konsumtif Remaja Menggunakan Algoritma CatBoost Berbasis Machine Learning," Indonesian Journal on Data Science, vol. 3, no. 2, pp. 78–86, Nov. 2025.

[5] A. F. Istianto, A. I. Hadiana, dan F. R. Umbara, "Prediksi Curah Hujan Menggunakan Metode Categorical Boosting (CatBoost)," JATI (Jurnal Mahasiswa Teknik Informatika), vol. 7, no. 4, pp. 2930–2937, Aug. 2023.

[6] B. W. Sari dan D. Prabowo, "Analisis Perbandingan Prediksi Harga Rumah Dengan Random Forest, Gradient Boosting, dan XGBoost," Intellect: Indonesian Journal of Innovation Learning and Technology, vol. 4, no. 1, pp. 42–51, Jun. 2025. doi: 10.57255/intellect.v4i1.1385.

[7] C. P. Ananda, "Machine Learning untuk Prediksi Gaya Hidup Berdasarkan Socioeconomic Status (SES) Menggunakan Algoritma CatBoost: Studi Kasus Mahasiswa UIN Jakarta," Skripsi, Program Studi Sistem Informasi, Fakultas Sains dan Teknologi, Universitas Islam Negeri Syarif Hidayatullah Jakarta, 2023.

[8] F. H. Syahadah, R. T. Subagio, dan P. Rizqiyah, "Penerapan XGBoost dalam Prediksi Pendaftaran Siswa Baru Bimbingan Belajar QSC di Kota Cirebon," JITET (Jurnal Informatika dan Teknik Elektro Terapan), vol. 13, no. 3S1, pp. 1082–1089, 2025. doi: 10.23960/jitet.v13i3S1.7998.

[9] F. Madani dan A. H. Lubis, "CatBoost Algorithm Implementation for Classifying Women's Fashion Products," JITE (Journal of Informatics and Telecommunication Engineering), vol. 9, no. 1, pp. 249–260, Jul. 2025. doi: 10.31289/jite.v9i1.15604.

[10] A. Alaoui, et al., "A Comparative Study of Machine Learning Models for Real Estate Price Prediction," International Journal of Advanced Computer Science and Applications, vol. 12, no. 4, pp. 215–222, 2021. doi: 10.14569/IJACSA.2021.0120427.

[11] X. Chen dan Y. Zhang, "CatBoost for Real Estate Valuation: A Comparative Analysis," Journal of Property Research, vol. 40, no. 2, pp. 145–163, Jun. 2023. doi: 10.1080/09599916.2023.2185432.

[12] A. M. Siregar, et al., "Penerapan Machine Learning untuk Prediksi Harga Rumah di Wilayah Jakarta Selatan," JKBTI (Jurnal Komputer, Bisnis dan Teknologi Informasi), vol. 4, no. 1, pp. 45–53, 2022.

[13] R. Wijaya dan S. Pratama, "Comparative Analysis of Ensemble Methods for Housing Price Prediction in Surabaya," IJIST (International Journal of Information Systems and Technology), vol. 7, no. 2, pp. 88–97, 2023.

[14] M. B. S. Qolbi, T. N. Puteh, Rivandi, dan C. Rozikin, "Prediksi Harga Rumah di Jakarta Pusat Menggunakan Algoritma Machine Learning," Jurnal Ilmu Komputer dan Bisnis (JIKB), vol. 16, no. 1, pp. 16–24, Mei 2025. doi: 10.47927/jikb.v16i1.840.

[15] M. R. Fauzi, M. Handika, A. Awinanto, A. J. Wahidin, B. Rahmatullah, dan I. Kurniawati, "Analisis Perbandingan Kinerja Algoritma Linear Regression, Random Forest, dan XGBoost dalam Prediksi Harga Rumah," Journal of Artificial Intelligence and Digital Business (RIGGS), vol. 4, no. 4, pp. 1541–1548, 2025.

[16] M. G. A. Rianto, A. L. Prasasti, dan A. Novianty, "Implementasi Model XGBoost untuk Prediksi Jumlah Transaksi dan Total Pendapatan di Jaringan Restoran CV Balibul," Jurnal Nasional SAINS dan TEKNIK, vol. 2, no. 2, pp. 43–46, 2024. doi: 10.25124/jnst.v2i2.8750.