JITET (Jurnal Informatika dan Teknik Elektro Terapan)
Vol. 13 No. 3S1, pISSN: 2303-0577 eISSN: 2830-7062 http://dx.doi.org/10.23960/jitet.v13i3S1.7737

PERBANDINGAN MODEL ENSEMBLE LARNING DALAM MEMPREDIKSI HARGA SEWA INDEKOS DI JAKARTA

Syawaludin[1*], Antonius Bagas Sunu W.A.[2]

[1,2]Politeknik Statistika STIS; Jl. Otto Iskandardinata No. 64C, Jakarta Timur, Indonesia 13330; Telp. (021) 8191437

Keywords:
XGBoost;
feature importance;
harga sewa kos;
prediksi harga.

Corespondent Email:
222212892@stis.ac.id

Abstrak. Ketersediaan kamar kos di Jakarta semakin dibutuhkan seiring dengan tingginya laju urbanisasi. Namun, lonjakan harga sewa menciptakan kebutuhan untuk memahami faktor-faktor yang mempengaruhinya secara akurat. Penelitian ini membandingkan empat model ensemble learning, yaitu XGBoost, CatBoost, Random Forest, dan LightGBM dalam memprediksi harga sewa kos berdasarkan data dari situs Mamikos. Data dikumpulkan melalui web scraping dan dilakukan pra-pemrosesan untuk menghapus nilai hilang, pencilan, dan mengubah variabel kategorikal menjadi numerik. Evaluasi model menggunakan metrik MAE, MSE, dan R^2, dengan hyperparameter tuning melalui Optuna. Hasil menunjukkan bahwa XGBoost memiliki performa terbaik dengan R^2 sebesar 0,6333. Analisis feature importance menunjukkan bahwa fasilitas seperti AC, kloset duduk, dan kamar mandi dalam memiliki pengaruh tertinggi terhadap harga sewa kos, lebih besar dibandingkan lokasi. Temuan ini mengindikasikan bahwa fasilitas kamar menjadi faktor utama dalam penentuan harga. Untuk peningkatan model di masa depan, disarankan penambahan fitur relevan dan penerapan feature engineering lanjutan.

Abstract. *The demand for rental rooms in Jakarta continues to increase in line with the rapid pace of urbanization. However, the surge in rental prices creates a need to accurately understand the influencing factors. This study compares four ensemble learning models XGBoost, CatBoost, Random Forest, and LightGBM in predicting room rental prices using data obtained from the Mamikos website. The data were collected through web scraping and preprocessed to remove missing values and outliers, as well as to convert categorical variables into numerical format. Model performance was evaluated using MAE, MSE, and R^2 metrics, with hyperparameter tuning conducted using Optuna. The results show that XGBoost achieved the best performance with an R^2 value of 0.6333. The feature importance analysis indicates that room facilities such as AC, sitting toilet, and private bathroom had the highest influence on rental price, surpassing the influence of location. These findings suggest that room amenities are the primary determinant in price formation. For future model improvements, it is recommended to include more relevant features and apply advanced feature engineering techniques.*

1. PENDAHULUAN

Tempat tinggal merupakan salah satu kebutuhan dasar manusia yang sangat penting untuk mendukung keberlangsungan hidup dan aktivitas sosial. Dalam hierarki kebutuhan, tempat tinggal dikategorikan sebagai kebutuhan primer yang harus dipenuhi bersama dengan pangan dan sandang [1]. Sebagai ruang untuk

berlindung, beristirahat, dan bersosialisasi, keberadaan tempat tinggal yang layak menjadi dambaan setiap individu [2] [3]. Bentuk tempat tinggal pun beragam, seperti rumah, apartemen, asrama, indekos, dan kontrakan [4] [5].

Jakarta merupakan ibu kota sekaligus pusat perekonomian Indonesia yang mengalami pertumbuhan penduduk sangat cepat akibat urbanisasi dan perpindahan penduduk dari daerah lain. Berdasarkan Urban Village Potential Statistics tahun 2024 yang dirilis oleh Badan Pusat Statistik (BPS), seluruh kelurahan di Provinsi DKI Jakarta telah tergolong sebagai kawasan urban dengan kepadatan penduduk yang tinggi [6]. Lonjakan jumlah penduduk ini memicu tingginya permintaan terhadap hunian sementara, seperti indekos, terutama di wilayah strategis yang dekat dengan pusat kegiatan bisnis, pendidikan, maupun transportasi.

Namun demikian, tingginya permintaan terhadap tempat tinggal di Jakarta tidak diimbangi dengan ketersediaan lahan, yang pada akhirnya menyebabkan lonjakan harga sewa. Laporan Jakarta Rental Apartment Market Overview 2023 dari Knight Frank menyebutkan bahwa harga sewa apartemen di Jakarta mengalami kenaikan sebesar 5% secara tahunan (year-on-year), yang 71 persen diantaranya berada di kawasan pusat bisnis (CBD) [7]. Sejumlah media nasional bahkan menyoroti bahwa harga sewa indekos di Jakarta sebanding dengan cicilan rumah di wilayah penyangga seperti Bogor, Depok, dan Bekasi, yang menandakan tingginya beban biaya hidup di ibu kota [8].

Di tengah kondisi tersebut, platform digital seperti Mamikos telah menjadi sumber utama informasi indekos bagi masyarakat. Aplikasi ini tercatat telah diunduh lebih dari lima juta kali dan mendapatkan penilaian yang cukup tinggi di Google Play Store [9]. Profil perusahaan menyebutkan bahwa Mamikos mengelola lebih dari 100.000 daftar indekos aktif di seluruh Indonesia dan melayani sekitar tujuh juta pengguna [10]. Volume data yang besar dan terus diperbarui ini menghadirkan peluang besar dalam penelitian berbasis data untuk memahami dinamika pasar sewa indekos secara lebih objektif dan ilmiah. Namun, data ini masih jarang dimanfaatkan dalam kajian akademik, terutama untuk membangun sistem prediksi harga kos secara otomatis. Permasalahan utama dalam penelitian ini adalah bagaimana membangun model prediksi harga sewa kos yang mampu menangani keberagaman harga sewa kos karena berbagai faktor.

Sebagian besar penelitian di Indonesia lebih berfokus pada prediksi harga jual rumah. Tanamal et al., misalnya, menggunakan Random Forest untuk memprediksi harga rumah di Surabaya dan memperoleh akurasi tinggi [11]. Namun, pendekatan ini belum banyak diterapkan pada pasar sewa indekos, yang justru lebih fluktuatif dan dipengaruhi oleh variabel non-linier seperti lokasi mikro dan fasilitas sekitar. Selain itu, Random Forest memiliki keterbatasan, seperti ukuran model yang besar dan sensitivitas terhadap fitur yang berkorelasi tinggi [12]. Maula et al. membandingkan MLR, SVR, LightGBM, dan Random Forest untuk harga rumah di Jakarta dan Tangerang Selatan, dan menemukan bahwa model boosting seperti LightGBM lebih stabil [13]. Meskipun terdapat banyak kajian mengenai hunian, kajian terkait prediksi harga indekos, khususnya di Jakarta, masih sangat terbatas.

Metodologi ensemble learning menjadi pendekatan yang kian populer dalam menyelesaikan masalah prediksi, karena mampu menggabungkan keunggulan beberapa model dasar sekaligus untuk meningkatkan akurasi dan stabilitas hasil. Studi oleh Mienye dan Sun menunjukkan bahwa metode bagging seperti Random Forest serta boosting seperti Gradient Boosting dan XGBoost menunjukkan performa unggul di berbagai bidang prediktif [14]. Pastukh dan Khomyshyn juga membuktikan bahwa kombinasi Gradient Boosting dan Extra Trees menghasilkan prediksi harga properti yang sangat akurat berdasarkan metrik R^2, RMSE, dan MAE [15]. Selain itu, Zhao et al. menambahkan bahwa XGBoost tak hanya unggul dalam performa, tetapi juga mendukung analisis feature importance [16].

Berdasarkan latar belakang tersebut, penelitian ini bertujuan untuk membandingkan performa lima model ensemble learning, yaitu XGBoost, CatBoost, Random Forest, LightGBM, dan Optuna, dalam memprediksi harga sewa kos di Jakarta dengan menggunakan metrik evaluasi seperti Mean Absolute Error (MAE), Mean Squared Error (MSE), dan koefisien determinasi (R^2). Selain itu, penelitian ini juga bertujuan untuk mengidentifikasi

variabel-variabel yang paling berpengaruh terhadap harga sewa kos di Jakarta melalui analisis feature importance berdasarkan model dengan performa terbaik.

2. TINJAUAN PUSTAKA

Ensemble Learning merupakan pendekatan dalam pembelajaran mesin yang menggabungkan beberapa model dasar (base learners) untuk menghasilkan prediksi yang lebih akurat dan stabil dibandingkan penggunaan satu model saja. Tujuan utama metode ini adalah mengurangi variansi, mengatasi bias, dan meningkatkan generalisasi model pada data baru. Dua teknik utama dalam ensemble learning adalah bagging dan boosting. Bagging (Bootstrap Aggregating) bekerja dengan membuat beberapa subset data secara acak dari dataset asli untuk melatih model yang sama, sedangkan boosting melatih model secara berurutan di mana setiap model baru difokuskan untuk memperbaiki kesalahan model sebelumnya [17].

Random Forest adalah salah satu algoritma bagging yang memanfaatkan pohon keputusan sebagai model dasar. Setiap pohon dibangun menggunakan subset data dan subset fitur yang dipilih secara acak (random feature selection), sehingga tercipta variasi antar pohon. Proses prediksi pada klasifikasi dilakukan dengan majority voting, sementara pada regresi dilakukan dengan averaging hasil prediksi seluruh pohon. Keunggulan utama Random Forest adalah kemampuannya menangani data berdimensi tinggi, bersifat robust terhadap outlier, dan dapat mengukur tingkat kepentingan fitur (feature importance) [18]

Sementara itu, algoritma boosting seperti XGBoost, LightGBM, dan CatBoost bekerja dengan prinsip memperbaiki kesalahan prediksi secara iteratif. XGBoost menggunakan optimasi berbasis gradient boosting dengan regularisasi L1/L2 untuk mengendalikan kompleksitas model [19]. LightGBM mengimplementasikan teknik seperti Gradient-based One-Side Sampling (GOSS) dan Exclusive Feature Bundling (EFB) untuk mempercepat proses pelatihan [20], sedangkan CatBoost dirancang khusus untuk menangani variabel kategorikal dengan teknik ordered boosting dan target statistics [21].

3. METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif dan bersifat eksperimen komputasional dalam membandingkan performa beberapa model ensemble learning dalam memprediksi harga sewa kamar kos.

3.1. Sumber Data dan Variabel Penelitian

Data yang digunakan dalam penelitian ini diperoleh melalui proses web scraping dari situs pencarian indekos Mamikos (https://mamikos.com). Situs ini menyediakan informasi kos yang dapat diakses publik oleh pengguna umum, namun tidak menyediakan dataset dalam format open data yang siap unduh.

Proses scraping dilakukan pada bulan Juni 2025 dengan strategi pencarian berdasarkan nama kecamatan yang ada di Provinsi DKI Jakarta, tidak termasuk wilayah Kepulauan Seribu. Proses scraping menghasilkan sebanyak 8.142 entri kos yang berhasil dikumpulkan, mencakup seluruh kecamatan di DKI Jakarta.

Atribut yang dikumpulkan meliputi informasi harga sewa per bulan, jenis kos (putra, putri, atau campur), nama kecamatan, rating konsumen, serta keberadaan fasilitas seperti AC, kasur, Wifi, kamar mandi dalam, akses 24 jam, dan kloset duduk. Variabel-variabel ini dipilih berdasarkan fitur yang paling umum dan relevan dalam praktik pencarian kos online oleh pengguna aplikasi Mamikos, serta merujuk pada penelitian-penelitian terdahulu yang menggunakan variabel serupa dalam memodelkan harga properti atau tempat tinggal sewa [22] [23]. Adapun atribut atau variabel yang digunakan disajikan dalam tabel berikut:

Tabel 1. Variabel Penelitian

Nama Variabel	Deskripsi	Deskripsi
nama	Nama Kos	Kategorik
harga	Harga sewa per bulan dalam rupiah	Numerik
jenis	Jenis kos (putra, putri, campur)	Kategorik
kecamatan	Lokasi kos berdasarkan kecamatan	Kategorik

Nama Variabel	Deskripsi	Deskripsi
rating	Nilai penilaian dari konsumen (skala 1–5)	Numerik
ac	Tersedia AC (1 = ya, 0 = tidak)	Kategorik
akses 24 jam	Boleh keluar/masuk 24 jam (1 = ya, 0 = tidak)	Kategorik
kasur	Tersedia kasur (1 = ya, 0 = tidak)	Kategorik
kloset duduk	Tersedia kloset duduk (1 = ya, 0 = tidak)	Kategorik
kamar mandi dalam	Tersedia kamar mandi dalam (1 = ya, 0 = tidak)	Kategorik
wifi	Tersedia Wi-Fi (1 = ya, 0 = tidak)	Kategorik

3.2. Pra-Pemrosesan Data

Sebelum data digunakan dalam proses pemodelan, dilakukan tahap pra-pemrosesan untuk memastikan kualitas dan konsistensi data. Proses ini mencakup pengecekan terhadap data yang hilang, penghapusan atau imputasi data yang tidak lengkap, serta deteksi dan penanganan pencilan (outlier) agar tidak memengaruhi hasil prediksi secara ekstrem. Selain itu, variabel kategorikal diubah ke dalam bentuk numerik agar dapat dikenali oleh algoritma Machine Learning, dan dilakukan normalisasi atau standardisasi pada variabel numerik untuk menyamakan skala antar fitur. Tahapan ini bertujuan untuk meningkatkan akurasi model dan memastikan bahwa data siap digunakan dalam analisis lebih lanjut.

3.3. Analisis Data Eksploratif

Setelah proses pra-pemrosesan selesai, dilakukan analisis deskriptif terhadap dataset untuk memahami karakteristik umum dari masing-masing variabel. Analisis ini meliputi pemeriksaan distribusi data serta analisis proporsi dalam setiap atribut kategorikal.

3.4. Membagi Data Menjadi Training dan Testing

Langkah ini penting untuk memastikan struktur data sesuai dengan asumsi pemodelan dan memberikan arahan awal dalam proses eksplorasi data. Setelah itu, data dibagi menjadi dua subset utama, yaitu data latih (training set) dan data uji (testing set) dengan rasio umum 80:20 [24]. Pembagian ini dilakukan secara acak dan bertujuan untuk mengevaluasi kemampuan generalisasi model terhadap data baru.

3.4. Optimalisasi Model

Untuk menyeimbangkan antara akurasi dan beban komputasi, dilakukan optimalisasi hyperparameter menggunakan Optuna. Optuna adalah sebuah framework open-source yang dirancang untuk mempermudah dan mempercepat proses pencarian kombinasi hyperparameter terbaik pada model machine learning. Optuna menerapkan pendekatan define-by-run, yang memungkinkan ruang pencarian dibentuk secara dinamis selama proses optimasi berlangsung, sehingga memberikan fleksibilitas yang lebih tinggi dibandingkan metode tradisional. Dalam proses pencarian, Optuna menggunakan algoritma Tree-structured Parzen Estimator (TPE) sebagai metode sampling, serta dilengkapi dengan mekanisme pruning seperti Asynchronous Successive Halving Algorithm (ASHA) yang berfungsi untuk menghentikan evaluasi terhadap kombinasi hyperparameter yang tidak menjanjikan. Pendekatan ini tidak hanya meningkatkan efisiensi proses optimasi, tetapi juga menghemat sumber daya komputasi secara signifikan [25].

3.5. Tahap Pengembangan Model

3.5.1. XGBoost

XGBoost (Extreme Gradient Boosting) adalah algoritma machine learning yang efisien dan berbasis pada teknik ensemble pohon keputusan [19]. Algoritma ini banyak digunakan dalam tugas-tugas yang melibatkan pemrosesan data skala besar karena kemampuannya dalam meningkatkan akurasi dan stabilitas prediksi dengan membangun banyak pohon keputusan dan menggabungkan hasilnya. Selain itu, XGBoost mendukung berbagai tujuan optimasi dan kriteria evaluasi yang dapat disesuaikan, serta memiliki kecepatan komputasi yang tinggi dan performa prediksi yang sangat baik, menjadikannya sangat cocok untuk masalah nonlinier yang kompleks [26].

3.5.2. Random Forest

Random Forest (RF) merupakan metode pembelajaran terintegrasi (ensemble learning) yang pertama kali diperkenalkan oleh Breiman dan hingga kini tetap menjadi salah satu algoritma paling populer. Model ini menggunakan pohon keputusan (decision tree) sebagai model dasar dan membangun sekumpulan pohon yang bervariasi melalui kombinasi subset data dan fitur yang berbeda. Prediksi akhir diperoleh melalui proses agregasi atau voting dari seluruh pohon, yang menghasilkan peningkatan akurasi dan stabilitas model dibandingkan metode pembelajaran tunggal. Tidak seperti regresi linear klasik, RF tidak terpengaruh oleh masalah multikolinearitas antar variabel input, sehingga cocok untuk data yang kompleks dan memiliki banyak fitur saling berkorelasi [27].

3.5.3. Categorical Boosting

CatBoost adalah algoritma machine learning berbasis pohon keputusan yang dirancang secara khusus untuk menangani variabel kategorikal secara efisien melalui pendekatan boosting. Model ini diperkenalkan oleh Dorogush et al. sebagai bagian dari pengembangan algoritma gradient boosting yang lebih stabil dan akurat dalam membangun model prediktif, terutama pada masalah regresi [28]. Salah satu keunggulan utama CatBoost adalah kemampuannya dalam mengurangi risiko overfitting, yang sering menjadi kendala pada metode boosting tradisional. Berbeda dari pendekatan boosting lainnya, CatBoost secara internal mengadopsi teknik transformasi canggih yang disebut ordered boosting dan target statistics untuk mengonversi fitur kategorikal menjadi format numerik tanpa kehilangan informasi penting. Prokhorenkova et al. menyebut bahwa pendekatan ini sangat efektif terutama ketika jumlah data terbatas dan atribut kategorikal mendominasi struktur dataset. Selain itu, CatBoost mampu memanfaatkan keterkaitan antar fitur kategorikal secara eksplisit untuk meningkatkan kualitas prediksi serta memperkuat hubungan antar parameter dalam model [29].

3.5.4. Light Gradient Boosting Machine

LightGBM (LGBM) merupakan algoritma machine learning berbasis Gradient Boosting Decision Tree (GBDT) yang dikembangkan dengan fokus pada efisiensi pelatihan, penghematan memori, dan peningkatan akurasi dalam skenario skala besar. Untuk mencapai efisiensi tersebut, LGBM mengimplementasikan beberapa teknik utama seperti gradient-based one-sided sampling (GOSS), exclusive feature bundling (EFB), histogram-based decision tree algorithm, dan strategi pertumbuhan pohon secara leaf-wise seperti pada metode tradisional [30].

3.6. Evaluasi Model

Tahap evaluasi berfokus pada penilaian kualitas model prediktif yang telah dikembangkan menggunakan algoritma ensemble learning, yaitu Random Forest, XGBoost, CatBoost, dan LightGBM. Pada tahap ini, model dievaluasi untuk mengukur performanya dan menentukan seberapa baik model tersebut mampu menghasilkan prediksi yang akurat berdasarkan metrik evaluasi seperti MAE, MSE, dan R^2.

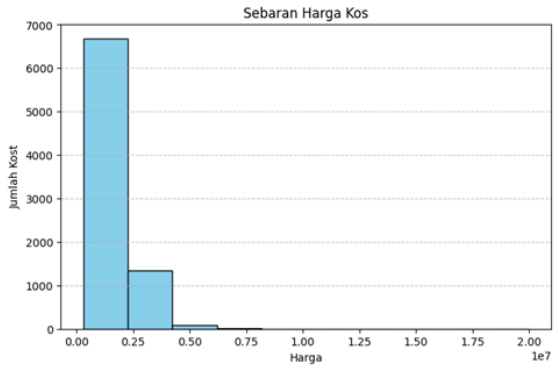
3.7. Feature Importance

Metode feature importance digunakan untuk mengetahui seberapa besar pengaruh masing-masing fitur terhadap hasil prediksi model. Metode ini bekerja dengan cara melihat perubahan kinerja model setelah nilai suatu fitur diacak. Jika pengacakan fitur tersebut menyebabkan penurunan akurasi model yang signifikan, maka fitur tersebut dianggap penting. Pendekatan ini bersifat independen terhadap jenis model (model-agnostic), sehingga dapat diterapkan pada berbagai algoritma machine learning. Melalui metode ini, diperoleh gambaran yang lebih jelas tentang fitur mana yang paling berperan dalam proses prediksi, sehingga dapat meningkatkan transparansi dan interpretabilitas model yang digunakan [31].

4. HASIL DAN PEMBAHASAN

4.1. Eksplorasi data

Eksplorasi data bertujuan untuk memperoleh pemahaman yang mendalam mengenai komponen dan isi data yang dianalisis, sehingga dapat mengidentifikasi informasi penting yang relevan dan bermanfaat dalam proses pengolahan data lebih lanjut [32].



Gambar 1. Sebaran Harga Sewa Kos

4.2. Pengecekan kualitas data

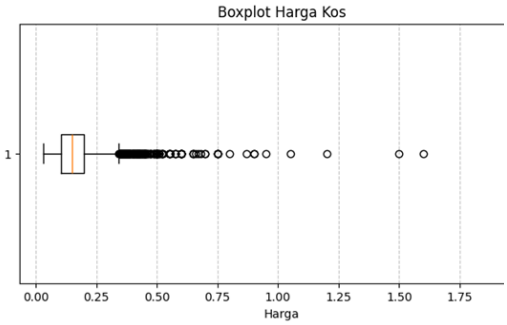
1) Data di Luar Jangkauan

Terdapat beberapa kos yang berada di kecamatan di luar Jakarta yang masuk ke dalam dataset. Penghapusan data dilakukan dengan fitur filter di Excel.

2) Missing Values

Terdapat beberapa missing value pada kolom kecamatan.

3) Outliers



Gambar 2. Boxplot Sebaran Harga Sewa Kos di Jakarta

Gambar 2 menunjukkan bahwa terlihat adanya sejumlah outlier dengan nilai yang sangat tinggi, yang dapat terjadi karena kesalahan penulisan digit, atau salah memasukkan nominal yang seharusnya tarif tahunan. Dari boxplot tersebut, terlihat juga bahwa tidak terdapat pencilan bawah.

4.3. Data Preparation

1) Penanganan Missing Values

Terdapat beberapa fitur yang memiliki nilai hilang (missing values) dengan proporsi yang bervariasi. Kolom yang terdapat missing value adalah kolom kecamatan.

2) Penanganan Outlier

Penanganan outlier dilakukan dengan memfilter data yang nilainya lebih dari kuartil atas ditambah 1,5 kali jarak antar kuartil [33]. Data yang berada di area tersebut akan langsung dihapus.

```
Q1 (25%)        : 1050000.0
Q3 (75%)        : 2000000.0
IQR             : 950000.0
Batas Bawah     : -375000.0
Batas Atas      : 3425000.0
```

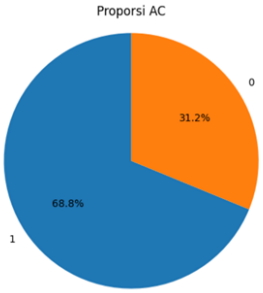
Gambar 3. Hasil Deteksi Outlier Berdasarkan IQR

Terlihat bahwa batas untuk pencilan atas adalah Rp3.425.000, sehingga terdapat 312 baris data yang dihapus. Penghapusan menyisakan 7797 baris data.

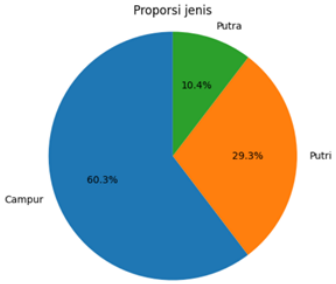
3) Konversi Variabel Kategorikal

Terdapat delapan variabel kategorik yang digunakan sebagai variabel prediktor. Enam variabel tersebut telah diatur untuk hanya bernilai 0 dan 1. Dua variabel lain akan dikodekan menggunakan metode one-hot-encoding, (cari artikel one hot encoding dalam analisis ensembel).

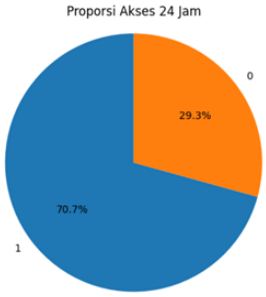
4.4. Analisis Deskriptif



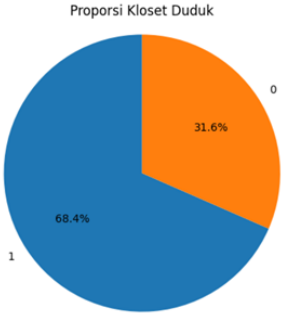
Gambar 4. Diagram Lingkaran Proporsi Ac



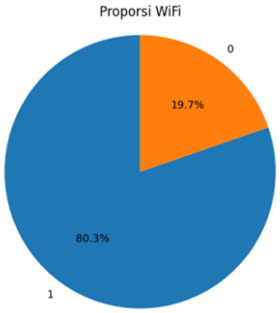
Gambar 5. Diagram Lingkaran Proporsi Jenis Kos



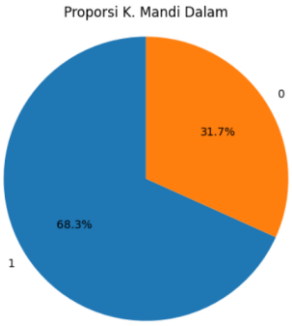
Gambar 6. Diagram Lingkaran Proporsi Akses 24 Jam



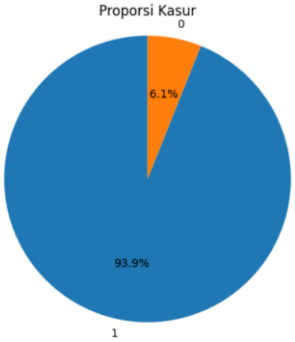
Gambar 7. Diagram Lingkaran Proporsi Kloset Duduk



Gambar 8. Diagram Lingkaran Proporsi Wifi



Gambar 9. Diagram Lingkaran Proporsi Kamar Mandi Dalam



Gambar 10. Diagram Lingkaran Proporsi Kasur

Berdasarkan hasil analisis deskriptif menggunakan diagram lingkaran, Sebagian besar indekos di Jakarta telah menyediakan fasilitas pendingin udara (AC), dengan persentase sebesar 68,8%. Ditinjau dari jenis kos, kos campur merupakan kategori yang paling dominan dengan proporsi 60,3%, diikuti oleh indekos khusus putri sebesar 29,3% dan indekos khusus putra sebesar 10,4%. Selain itu, mayoritas indekos di Jakarta dapat diakses selama 24 jam tanpa adanya pembatasan jam malam (70,7%). Fasilitas dasar seperti kasur, akses Wi-Fi, dan kloset duduk umumnya tersedia di sebagian besar indekos. Namun demikian, fasilitas kamar mandi dalam masih belum menjadi standar yang umum ditemui di seluruh indekos di wilayah Jakarta.

4.5. Evaluasi Model

Pada tahap pemodelan, data dibagi menjadi dua bagian, yaitu data latih dan data uji, dengan proporsi 80% untuk pelatihan dan 20% untuk pengujian. Selanjutnya, dilakukan proses standardisasi menggunakan metode StandardScaler. Langkah ini bertujuan untuk menyamakan skala antar fitur agar perbedaan rentang nilai tidak mempengaruhi kinerja model. Standarisasi membantu mengurangi variabilitas antar fitur dan mempermudah model dalam mempelajari hubungan antara fitur masukan dan target secara lebih efisien. Untuk memperoleh model terbaik, akan dilakukan hyperparametric tuning untuk setiap model dengan menggunakan Optuna. Tabel 2 menunjukkan parameter terbaik yang diperoleh dari output Optuna.

Tabel 2. Parameter Terbaik Setiap Model

Algoritma	Hyperparameter

Random Forest	{'n_estimators': 371, 'max_depth': 42, 'min_samples_split': 10, 'min_samples_l 'max_features': 'sqrt'}
XGBoost	{'n_estimators': 199, 'max_depth': 1 'learning_rate': 0.0513176093482410 'min_child_weight': 7, 'subsample': 0.6754546789505042, 'colsample_byt 0.6077048569680631}
CatBoost	{'iterations': 490, 'learning_rate': 0.22113355829268982, 'depth': 6, 'l2_lea 9.944216522211834}
LightGBM	{'n_estimators': 179, 'max_depth': 1 'learning_rate': 0.111343348454803 'num_leaves': 106, 'min_child_samples 'subsample': 0.7358762833181154, 'colsample_bytree': 0.5013875072863196}

Evaluasi model dilakukan untuk membandingkan performa dua algoritma, yaitu Random Forest, XGBoost, CatBoost, dan LightGBM dalam memprediksi harga kos. Tabel di bawah ini menyajikan tiga metrik evaluasi utama yang digunakan: Mean Absolute Error (MAE), Mean Squared Error (MSE), dan R-squared (R^2).

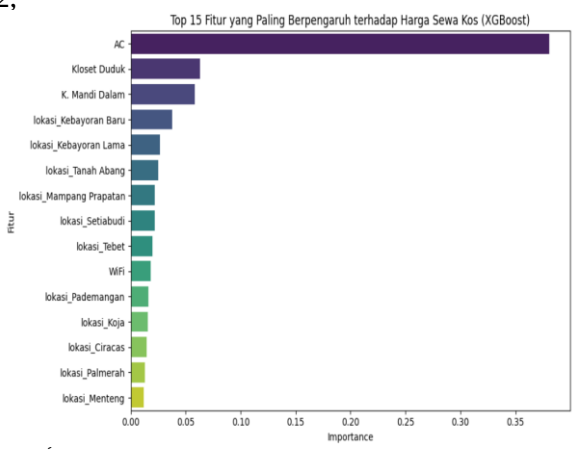
Tabel 3. Perbandingan Performa Model pada Parameter Optimal

Algoritma	MAE	MSE	R^2
Random Forest	0,4533	0,3817	0,6212
XGBoost	0,4449	0,3695	0,6333
CatBoost	0,4468	0,3721	0,6308
Light GBM	0,4501	0,3740	0,6289

Berdasarkan tabel 3, dapat dilihat bahwa XGBoost mendapatkan hasil terbaik, yang memiliki MSE sebesar 0,3695 dan R^2 tertinggi sebesar 0,6333. Hasil ini sesuai dengan penelitian yang dilakukan oleh Szczepanek (2022) menunjukkan bahwa XGBoost memiliki kinerja yang stabil dan dapat diandalkan, khususnya ketika jumlah variabel penelitian terbatas. Meskipun dalam penelitiannya LightGBM menunjukkan hasil terbaik setelah dilakukan optimasi, XGBoost tetap menjadi salah satu model yang unggul karena mampu memberikan hasil yang baik [32].

4.6. Feature Importance

Hasil dari analisis feature importance menunjukkan kontribusi masing-masing fitur terhadap prediksi harga kos.



Gambar 11. Feature Importance Variabel Yang Berpengaruh Terhadap Harga Sewa Kos

Berdasarkan hasil analisis feature importance menggunakan model XGBoost, terlihat bahwa fitur AC memiliki pengaruh paling signifikan dalam menentukan harga sewa kos, jauh melampaui fitur lainnya. Hal ini menunjukkan bahwa keberadaan fasilitas AC menjadi faktor utama yang dipertimbangkan dalam penentuan harga kos di dataset ini. Fitur lain yang juga cukup berpengaruh adalah Kloset Duduk dan Kamar Mandi Dalam, yang menempati posisi kedua dan ketiga. Artinya, selain pendingin ruangan, kenyamanan fasilitas sanitasi pribadi menjadi faktor penting yang mempengaruhi harga sewa kos.

Di sisi lain, variabel lokasi ternyata memiliki pengaruh yang lebih kecil secara individu, meskipun tetap masuk dalam 15 fitur teratas. Lokasi seperti Kebayoran Baru, Kebayoran Lama, dan Setiabudi tercatat sedikit berkontribusi dalam penentuan harga, namun tidak sebesar pengaruh fasilitas kamar. Hal ini dapat mengindikasikan bahwa dalam pasar kos yang dianalisis, harga lebih dipengaruhi oleh kelengkapan fasilitas kamar dibandingkan hanya sekadar lokasi. Meski demikian, lokasi tetap relevan sebagai pelengkap faktor harga, tetapi nilai tambah dari fasilitas kamar menjadi daya tarik utama dalam menentukan nilai sewa kos.

Prediksi harga sewa kos dengan menggunakan metode ensemble learning menunjukkan hasil yang belum terlalu tinggi. Bahkan pada model terbaik, yaitu XGBoost, nilai koefisien determinasi (R^2) yang diperoleh hanya sebesar 0,6333. Berdasarkan penelitian

oleh Ahn et al. (2023) model XGBoost cenderung menghasilkan nilai R^2 yang lebih baik ketika digunakan pada data berdimensi tinggi dan telah melalui proses pembersihan data seperti penghilangan pencilan (outlier) [34]. Hal ini menunjukkan bahwa kualitas data sangat berpengaruh terhadap kinerja model. Ketika pencilan tidak ditangani, kemungkinan besar model yang terbaik akan berbeda.

Menariknya, meskipun dalam penelitian ini terdapat banyak variabel kategorik, XGBoost tetap menunjukkan performa prediksi yang lebih baik dibandingkan CatBoost. Hasil ini bertolak belakang dengan temuan dari Kulkarni (2022), yang menyatakan bahwa CatBoost secara khusus dirancang untuk menangani data kategorikal secara efisien, tanpa perlu proses preprocessing seperti one-hot encoding. CatBoost memiliki keunggulan dalam mempertahankan informasi variabel kategorik selama proses pelatihan dan dilengkapi dengan algoritma seperti ordered boosting dan dynamic learning rate scheduling, yang pada dasarnya dirancang untuk meningkatkan akurasi dan efisiensi model [23]. Oleh karena itu, perbedaan hasil ini menunjukkan bahwa perlu dilakukan analisis lebih lanjut untuk memahami faktor-faktor yang membuat XGBoost lebih unggul dalam konteks prediksi harga sewa kos, meskipun secara teoritis CatBoost lebih unggul dalam menangani data kategorikal.

Selain itu, nilai R^2 sebesar 0,6333 ini sejalan dengan temuan pada penelitian sebelumnya oleh Faisal Ardiansyah (2020), yang menggunakan metode Random Forest untuk memprediksi harga sewa kost di Yogyakarta dan memperoleh nilai R^2 sebesar 65,91%. Kedua hasil ini menunjukkan bahwa terbatasnya jumlah dan jenis variabel prediktor turut mempengaruhi rendahnya performa model [35].

Untuk meningkatkan akurasi prediksi di masa mendatang, disarankan agar dilakukan penambahan variabel prediktor yang lebih relevan dan informatif, misalnya melalui teknik scraping yang lebih dalam dari berbagai sumber data. Selain itu, eksplorasi terhadap variabel lag, interaksi antar fitur, serta teknik feature engineering lainnya juga dapat membantu meningkatkan performa model. Tak kalah penting, optimasi hyperparameter secara sistematis juga dapat mendukung tercapainya nilai R^2 yang lebih tinggi dan model yang lebih andal.

5. KESIMPULAN

Berdasarkan hasil penelitian, berbagai model ensembel mampu memberikan performa yang cukup baik, untuk memprediksi harga kos di Jakarta, dengan terbatasnya data yang tersedia. Setiap metode tidak memiliki perbedaan performa yang signifikan. Penggunaan Optuna sebagai hyperparametric tuning, dapat membantu meningkatkan performa model, tetapi tidak signifikan. Untuk meningkatkan performa model, diperlukan adanya penambahan variabel prediktor yang dapat menjelaskan penyebab keberagaman harga kos.

Model XGBoost menunjukkan performa terbaik dalam memprediksi harga sewa kos di Jakarta dibandingkan dengan model ensemble lainnya seperti CatBoost, Random Forest, dan LightGBM. Model ini berhasil mencapai nilai koefisien determinasi (R^2) sebesar 0,6333 dan Mean Squared Error (MSE) sebesar 0,3695, yang menunjukkan kemampuannya dalam menjelaskan lebih dari 63% variasi harga sewa kos. Meskipun demikian, nilai R^2 tersebut masih tergolong sedang, sehingga masih ada ruang untuk peningkatan akurasi model. Berdasarkan hasil analisis feature importance, diketahui bahwa fasilitas kamar seperti AC, kloset duduk, dan kamar mandi dalam memiliki pengaruh yang paling besar terhadap harga sewa, mengungguli pengaruh faktor lokasi. Temuan ini mengindikasikan bahwa kenyamanan fasilitas lebih menentukan harga dibandingkan letak geografis kos dalam konteks pasar sewa di Jakarta.

Selain itu, hasil penelitian ini juga memperlihatkan bahwa meskipun CatBoost dirancang untuk menangani variabel kategorikal secara efisien, model XGBoost justru memberikan hasil yang lebih baik dalam konteks data yang digunakan. Hal ini menunjukkan bahwa performa model tidak hanya dipengaruhi oleh jenis algoritma, tetapi juga sangat bergantung pada kualitas data dan proses pra-pemrosesan, termasuk penanganan outlier. Untuk penelitian selanjutnya, disarankan untuk menambahkan variabel prediktor yang lebih relevan dan informatif, melakukan eksplorasi terhadap interaksi antar variabel dan variabel lag, serta menerapkan

teknik feature engineering dan optimasi hyperparameter secara sistematis, seperti menggunakan framework Optuna. Dengan langkah-langkah tersebut, diharapkan model prediksi dapat menghasilkan kinerja yang lebih tinggi dan memberikan hasil yang lebih akurat serta aplikatif.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada pihak-pihak terkait yang telah memberi dukungan terhadap penelitian ini, khususnya kepada Bu Rani Nooraeni selaku pembimbing, Serta teman-teman yang telah mendukung selama penelitian dan penulisan ini.

DAFTAR PUSTAKA

[1] N. I. Imansari, "Praktikum Mengenai Kebutuhan Atau Utilitas Dalam Khidupan Sehari-hari," J. Masharif al-Syariah J. Ekon. Perbank. Syariah, vol. 5, no. 2, p. 90, 2020.

[2] J. Wibowati, "Pengaruh Kualitas Pelayanan Terhadap Kepuasan Pelanggan Pada Pt Muarakati Baru Satu Palembang," J. Manaj., vol. 8, no. 2, pp. 15–31, 2021, doi: 10.36546/jm.v8i2.348.

[3] M. Kharisma and I. F. Susilowati, "Tinjauan Yuridis Terhadap Pengaturan Pemanfaatan Rumah Negara Selain Sebagai Tempat Tinggal Di Indonesia," NOVUM J. Huk., vol. 7, no. 3, pp. 12–26, 2020.

[4] M. Adam, A. Yassin, M. Adam, A. L. Yassin, S. R. I. H. Wahyuningrum, and S. W. Firmandhani, "Logo TA , DAFT , UNDIP APARTEMEN SEWA UNTUK MAHASISWA DENGAN PENDEKATAN ARSITEKTUR KONTEMPORER," vol. 04, no. 1, p. 1988, 1988.

[5] I. Fawzia and D. N. Andini, "Tipologi Pola Ruang Rumah Kost Mahasiswa Di Banjarbaru," J. Rivet, vol. 2, no. 01, pp. 62–68, 2022, doi: 10.47233/rivet.v2i01.543.

[6] Badan Pusat Statistik, Urban Village Potential Statistics of DKI Jakarta Province 2024, Jakarta, 2024. [Online]. Available: https://jakarta.bps.go.id.

[7] Knight Frank, Jakarta Rental Apartment Market Overview H2-2023, 2024. [Online]. https://content.knightfrank.com/research/2847/

[8] F. Risar, "Harga Kos di Jakarta Setara Cicilan Rumah? Ini 4 Alasannya," IDN Times, Mei 2025. [Online]. https://www.idntimes.com/

[9] Google Play, "Mamikos – Cari & Sewa Kos Mudah," Google Play Store, 2025. [Online]. Available: https://play.google.com/store/apps/details?id=com.git.mami.kos

[10] Jobstreet, "Company Profile: Mamikos.com," Jobstreet Indonesia, 2025. [Online]. https://id.jobstreet.com/companies/mamikos-168557133083389

[11] R. Tanamal, N. Minoque, T. Wiradinata, Y. Soekamto, and T. Ratih, "House Price Prediction Model Using Random Forest in Surabaya City," TEM J., vol. 12, no. 1, pp. 126–132, 2023, doi: 10.18421/TEM121-17.

[12] M. Maia, A. R. Azevedo, and A. Ara, "Predictive Comparison Between Random Machines and Random Forests," J. Data Sci., vol. 19, no. 4, pp. 593–614, 2021, doi: 10.6339/21-JDS1025.

[13] S. F. Al Maula, N. A. D. Setiawan, E. Pusporani, and S. Z. Jannah, "Modeling House Selling Prices in Jakarta and South Tangerang Using Machine Learning Prediction Analysis," Barekeng, vol. 19, no. 1, pp. 107–118, 2025, doi: 10.30598/barekengvol19iss1pp107-118.

[14] I. D. Mienye and Y. Sun, "A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects," IEEE Access, vol. 10, no. August, pp. 99129–99149, 2022, doi: 10.1109/ACCESS.2022.3207287.

[15] O. Pastukh and V. Khomyshyn, "Using ensemble methods of machine learning to predict real estate prices," CEUR Workshop Proc., vol. 3896, pp. 438–447, 2024.

[16] Y. Zhao, R. Ravi, S. Shi, Z. Wang, E. Y. Lam, and J. Zhao, "PATE: Property, Amenities, Traffic and Emotions Coming Together for Real Estate Price Prediction," Proc. - 2022 IEEE 9th Int. Conf. Data Sci. Adv. Anal. DSAA 2022, 2022, doi: 10.1109/DSAA54385.2022.10032416.

[17] A. G. Karegowda and M. A. Jayaram, "Ensemble Learning in Machine Learning: Concepts, Algorithms, and Applications," Applied Sciences, vol. 13, no. 4, p. 2147, 2023, doi: 10.3390/app13042147.

[18] N. H. Alfajr and S. Defiyanti, "Prediksi Penyakit Jantung Menggunakan Metode Random Forest dan Penerapan Principal Component Analysis (PCA)," JITET, vol. 12, no. 3S1, pp. 3457–3464, Oct. 2024, doi: 10.23960/jitet.v12i3S1.5055.

[19] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[20] G. Ke et al., "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in Advances in Neural Information Processing Systems 30 (NIPS 2017), 2017.

[21] L. Prokhorenkova, G. Gusev, A. Vorobev, A. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," in Advances in Neural Information Processing Systems 31 (NeurIPS 2018), 2018.
[22] N. R. Mohede, B. Rahmat, and K. Kartini, "Prediction of Boarding House Rental Prices Using Multiple Linear Regression Method," Int. J. Educ. Inf. Technol. Others, vol. 7, no. 3, pp. 191–199, 2024.
[23] C. S. Kulkarni, "Advancing Gradient Boosting: A Comprehensive Evaluation of the CatBoost Algorithm for Predictive Modeling," J. Artif. Intell. Mach. Learn. Data Sci., vol. 1, no. 5, pp. 54–57, 2022, doi: 10.51219/jaimld/chinmay-shripad-kulkarni/29.
[24] M. N. Hibatulloh, G. D. Prakoso, A. D. Putri Yunus, and T. D. Putra, "Prediksi Harga Rumah di Bandung 2024 Menggunakan Ensemble Learning: Analisis Komparatif dan Interpretabilitas," J. Inform. J. Pengemb. IT, vol. 10, no. 2, pp. 484–493, 2025, doi: 10.30591/jpit.v10i2.8200.
[25] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A Next-generation Hyperparameter Optimization Framework," Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., pp. 2623–2631, 2019, doi: 10.1145/3292500.3330701.
[26] A. Kaya, "Conflicted Principals, Uncertain Agency: The International Monetary Fund and the Great Recession," Glob. Policy, vol. 3, no. 1, pp. 24–34, 2012, doi: 10.1111/j.1758-5899.2011.00096.x.
[27] Z. Jin, J. Shang, Q. Zhu, C. Ling, W. Xie, and B. Qiang, "RFRSF: Employee Turnover Prediction Based on Random Forests and Survival Analysis," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 12343 LNCS, pp. 503–515, 2020, doi: 10.1007/978-3-030-62008-0_35.
[28] A. V. Dorogush, V. Ershov, and A. Gulin, "CatBoost: gradient boosting with categorical features support," pp. 1–7, 2018, [Online]. Available: http://arxiv.org/abs/1810.11363
[29] P. Liudmila, G. Gleb, V. Aleksandr, D. Anna Veronika, and G. Andrey, "Catboost: unbiased boosting with categorical features," Adv. Neural Inf. Process. Syst., no. Section 4, pp. 6638–6648, 2018.
[30] M. R. Machado, S. Karray, and I. T. De Sousa, "LightGBM: An effective decision tree gradient boosting method to predict customer loyalty in the finance industry," 14th Int. Conf. Comput. Sci. Educ. ICCSE 2019, no. Nips, pp. 1111–1116, 2019, doi: 10.1109/ICCSE.2019.8845529.
[31] C. Molnar, *Interpretable Machine Learning*, "5.5 Feature Importance," [Online]. Available: https://christophm.github.io/interpretable-ml-book/feature-importance.html. [Accessed: Jul. 13, 2025].
[32] R. Szczepanek, "Daily Streamflow Forecasting in Mountainous Catchment Using XGBoost, LightGBM and CatBoost," Hydrology, vol. 9, no. 12, 2022, doi: 10.3390/hydrology9120226.
[33] Dhiwa Aqsha, "Perbandingan Kinerja Algoritma Extreme Gradient Boosting Dan Random Forest Untuk Prediksi Harga Rumah Di Jabodetabek," J. Ilmu Komput. dan Sist. Inf., vol. 13, no. 1, pp. 1–7, 2025, doi: 10.24912/jiksi.v13i1.32863.
[34] J. M. Ahn, J. Kim, and K. Kim, "Ensemble Machine Learning of Gradient Boosting (XGBoost, LightGBM, CatBoost) and Attention-Based CNN-LSTM for Harmful Algal Blooms Forecasting," Toxins (Basel)., vol. 15, no. 10, 2023, doi: 10.3390/toxins15100608.
[35] F. Ardiansyah, "SISTEM PREDIKSI HARGA SEWA KOST DENGAN MENGGUNAKAN RANDOM FOREST ANALYTICS (Studi Kasus : Kost Eksklusif di Daerah Istimewa Yogyakarta) TUGAS AKHIR," Tugas Akhir, 2020.