ISSN: 2598-8565 (media cetak)
ISSN: 2620-4339 (media online)

PENGEMBANGAN MODEL PREDIKSI HARGA KOS DI WILAYAH JIMBARAN MENGGUNAKAN METODE DECISION TREE REGRESSION

I Putu Krishna Paradipta✉, Putu Manik Prihatini, Ni Nyoman Harini Puspita
Jurusan Teknologi Informasi, Politeknik Negeri Bali, Badung, Indonesia
Email: krishnaparadipta@gmail.com

DOI: https://doi.org/10.46880/jmika.Vol10No2.pp597-606

ABSTRACT

The demand for temporary housing in Jimbaran continues to increase, making it more complex to determine boarding house prices due to various physical characteristics, location, and facilities. Previous studies have limitations in modeling nonlinear relationships among variables and have not integrated feature engineering and hyperparameter tuning to improve prediction performance. This study aims to develop a model for predicting boarding house prices using the decision tree regression algorithm. The model was developed through data preprocessing, feature engineering, and hyperparameter tuning using GridSearchCV with stratified K-fold cross-validation. The optimal hyperparameter combination was max_depth = 8, min_samples_split = 5, min_samples_leaf = 2, and max_leaf_nodes = 80. The model achieved R^2 values of 0.8871 for the training data and 0.8326 for the testing data, with an MAE of IDR 225,438 and an RMSE of IDR 303,080. The R^2 difference of 0.0546 indicates satisfactory generalization without significant overfitting. In addition to producing an accurate prediction model, this study contributes through the implementation of data preprocessing, feature engineering, hyperparameter tuning, and feature importance analysis to identify the factors that most significantly affect boarding house prices.

Keyword: Boarding House Price Prediction, Data Mining, Decision Tree Regression, Machine Learning.

ABSTRAK

Kebutuhan hunian sementara di wilayah Jimbaran terus meningkat sehingga penentuan harga kos menjadi semakin kompleks karena dipengaruhi oleh berbagai karakteristik fisik, lokasi, dan fasilitas. Penelitian sebelumnya masih memiliki keterbatasan dalam memodelkan hubungan nonlinier antarvariabel serta belum mengintegrasikan feature engineering dan hyperparameter tuning untuk meningkatkan performa prediksi. Penelitian ini bertujuan mengembangkan model prediksi harga kos di wilayah Jimbaran menggunakan algoritma Decision Tree Regression. Model dikembangkan melalui tahapan preprocessing data, feature engineering, hyperparameter tuning menggunakan GridSearchCV dengan Stratified K-Fold Cross Validation. Kombinasi hyperparameter terbaik yang diperoleh adalah max_depth = 8, min_samples_split = 5, min_samples_leaf = 2, dan max_leaf_nodes = 80. Model menghasilkan nilai R^2 sebesar 0,8871 pada data training dan 0,8326 pada data testing, dengan nilai MAE sebesar Rp225.438 dan RMSE sebesar Rp303.080. Selisih nilai R^2 sebesar 0,0546 menunjukkan bahwa model memiliki kemampuan generalisasi yang baik tanpa mengalami overfitting yang signifikan. Selain menghasilkan model prediksi yang baik, penelitian ini juga memberikan kontribusi melalui penerapan preprocessing data, feature engineering, hyperparameter tuning, serta analisis feature importance untuk mengidentifikasi faktor-faktor yang paling memengaruhi harga kos.

Kata Kunci: Prediksi Harga Kos, Data Mining, Decision Tree Regression, Machine Learning.

PENDAHULUAN

Kecamatan Jimbaran adalah salah satu wilayah strategis yang berada di Kabupaten Badung, Provinsi Bali. Wilayah ini mengalami perkembangan yang cukup pesat setiap tahunnya, baik pada sektor infrastruktur, ekonomi, pendidikan, maupun pariwisata. Keberadaan berbagai fasilitas umum, pusat aktivitas wisata, dan institusi pendidikan menjadikan Jimbaran sebagai kawasan dengan tingkat mobilitas penduduk yang tinggi (Athiyah et al., 2021). Di samping itu, Jimbaran merupakan salah satu kawasan yang menarik minat banyak mahasiswa, tenaga kerja, dan pendatang yang berasal dari berbagai daerah. Keberadaan sejumlah institusi pendidikan tinggi, seperti Politeknik Negeri Bali, didukung oleh kawasan industri pariwisata serta lokasinya yang berdekatan dengan pusat Kota Denpasar dan Bandara Internasional I Gusti Ngurah Rai, menjadikan wilayah ini kerap dipilih sebagai tempat menetap baik untuk sementara waktu maupun dalam jangka panjang. Situasi tersebut

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

berdampak pada meningkatnya permintaan akan hunian, terutama rumah kos, yang terus menunjukkan pertumbuhan dari waktu ke waktu.

Meningkatnya jumlah pendatang, terutama mahasiswa dan pekerja, berdampak pada bertambahnya jumlah rumah kos yang tersedia di wilayah Jimbaran. Rumah kos yang tersedia memiliki beragam fasilitas, lokasi, serta tingkat kenyamanan yang berbeda-beda (Labib Mu'tashim et al., 2021). Meskipun demikian, beragamnya pilihan tersebut kerap menimbulkan kesulitan bagi calon penghuni dalam memilih kos yang sesuai dengan kebutuhan serta kondisi keuangan mereka. Berbagai fasilitas, seperti luas kamar, keberadaan kamar mandi pribadi, akses internet, area parkir, tingkat keamanan, hingga kedekatan lokasi dengan pusat kegiatan, turut memengaruhi besaran tarif sewa yang ditentukan oleh pemilik kos (Hadi & Benedict, 2024). Akan tetapi, harga yang ditawarkan tidak selalu sebanding dengan fasilitas yang tersedia. Selain itu, proses pencarian dan pemilihan kos yang masih dijalankan dengan cara manual memerlukan waktu yang cukup lama serta berpotensi memberikan keputusan yang kurang tepat, khususnya bagi calon penyewa dengan keterbatasan anggaran. Maka dari itu, diperlukan suatu model prediksi yang mampu memperkirakan harga kos berdasarkan fasilitas yang tersedia, sehingga dapat membantu calon penyewa dalam mengambil keputusan yang lebih tepat dan efisien.

Perkembangan teknologi informasi memberikan peluang dalam pemanfaatan metode *data mining* sebagai solusi untuk membantu proses penentuan harga kos sesuai dengan fasilitas yang tersedia (Muharrom, 2023). Metode *data mining* memungkinkan pengolahan data kos dalam jumlah besar untuk menemukan pola dan hubungan antara fasilitas dengan harga sewa yang sebelumnya sulit dianalisis secara manual. Salah satu algoritma yang mampu dipakai pada proses prediksi adalah *Decision Tree Regression*. Algoritma ini sanggup memberikan model keputusan yang mudah dimengerti, mempunyai struktur yang jelas, serta dapat mengolah data numerik maupun kategorikal (Ramadhan et al., 2023). Melalui penerapan algoritma *Decision Tree Regression*, data fasilitas rumah kos dapat dianalisis untuk menghasilkan prediksi harga sewa yang lebih akurat sehingga dapat membantu calon penyewa dalam menentukan pilihan kos sesuai kebutuhan fasilitas yang diinginkan.

Beberapa penelitian terdahulu telah membahas penerapan *machine learning* dan *data mining* dalam proses prediksi maupun pengambilan keputusan. Algoritma *Decision Tree Regression* mampu digunakan untuk memprediksi data numerik karena memiliki struktur model yang mudah dipahami serta performa yang cukup baik dalam menghasilkan prediksi (Saputra & Pratama, 2025). Metode *data mining* dapat membantu proses analisis data secara lebih efektif karena mampu mengolah dan mengekstraksi informasi yang berguna dari kumpulan data yang besar (Wijaya & Triayudi, 2023). Perbandingan analisis data antara algoritma *Support Vector Regression* dan *Decisiom Tree Regression* dapat dilakukan untuk mengetahui algoritma yang memberikan hasil prediksi yang lebih tepat (Wardana & Juanita, 2025). Selain itu, metode *machine learning* dapat membantu proses pengambilan keputusan secara lebih efektif dengan memanfaatkan pola dan informasi yang diperoleh dari data yang tersedia (Nurpadhilah & Saputera, 2026). Meskipun demikian, penelitian terdahulu masih memiliki beberapa keterbatasan, seperti belum optimalnya penerapan *preprocessing* data dan *feature engineering* dalam meningkatkan performa prediksi. Selain itu, belum ditemukannya penelitian yang secara khusus mengembangkan model prediksi harga kos yang menggunakan data rumah kos di wilayah Jimbaran (Karim et al., 2021).

Berbeda dengan penelitian sebelumnya, penelitian ini mengembangkan model prediksi harga kos di wilayah Jimbaran menggunakan algoritma *Decision Tree Regression* yang dikombinasikan dengan tahapan *preprocessing* data, *feature engineering*, *hyperparameter tuning* menggunakan *GridSearchCV*, serta validasi menggunakan *Stratified K-Fold Cross Validation*. Selain itu, penelitian ini juga menganalisis *feature importance* untuk mengidentifikasi atribut yang paling berpengaruh terhadap harga kos (Rayssa Buntoro, 2025). Dengan demikian, penelitian ini tidak hanya menghasilkan model prediksi harga kos dengan performa baik, tetapi juga memberikan pemahaman mengenai faktor-faktor yang memengaruhi penentuan harga kos di wilayah Jimbaran.

TINJAUAN PUSTAKA

Data Mining

Data Mining adalah bidang ilmu yang mempelajari proses penggalian informasi atau pengetahuan yang bernilai dari kumpulan data berukuran besar. Proses ini dilakukan dengan menerapkan berbagai metode, seperti statistik, matematika dan kecerdasan buatan untuk menemukan pola, hubungan, tren, maupun struktur tersembunyi yang sulit dikenali secara langsung melalui pengamanatan biasa. Pemanfaatan *data mining* telah banyak diterapkan dalam berbagai bidang karena mampu membantu proses pengambilan keputusan

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

melalui analisis data histori yang tersedia (Rahmadini, 2023). Selain itu, *data mining* juga dipakai untuk membantu proses prediksi dan analisis pola data sehingga informasi yang dihasilkan dapat dimanfaatkan sebagai dasar pengambilan keputusan yang lebih baik (Nuruzzaman et al., 2023).

Machine Learning

Machine learning adalah salah satu cabang dari kecerdasan buatan (*Artificial Intelligence*) yang memiliki fokus untuk mengembangkan algoritma serta metode agar komputer mampu mempelajari pola dari data dengan cara otomatis tanpa perlu diprogram dengan cara rinci (Ağbulut et al., 2021). Dalam prosesnya, *machine learning* memanfaatkan data latih untuk membentuk model matematis yang dapat merepresentasikan hubungan antarvariabel dalam data. Model tersebut kemudian dipakai untuk menjalankan prediksi, klasifikasi, maupun pengambilan keputusan terhadap data baru yang sebelumnya belum pernah diproses (Haryanto et al., 2023).

Preprocessing

Preprocessing atau pengolahan data merupakan tahapan awal yang memiliki peranan penting dalam proses *machine learning*. Tahapan ini dilakukan untuk mempersiapkan data mentah agar berada dalam kondisi yang baik sebelum digunakan dalam proses pelatihan model. Data yang didapatkan melalui beragam sumber sering kali masih mempunyai permasalahan, seperti masih memiliki nilai kosong (*missing value*), inkonsistensi data, kesalahan format, maupun ketidaksesuaian tipe data. Jika data tersebut langsung digunakan tanpa proses *preprocessing*, maka kualitas model yang dihasilkan dapat menurun (Çetin & Yıldız, 2022).

Feature Engineering

Feature Engineering merupakan proses mengubah, memilih, atau membentuk atribut baru dari data asli yang bertujuan untuk meningkatkan kualitas data dan kemampuan model *machine learning* dalam mengenali pola yang relevan (Qadri et al., 2023).

Train-Test Split

Train-Test Split merupakan metode validasi dalam *machine learning* yang digunakan untuk melakukan pembagian dataset menjadi dua bagian, yakni data latih (*training data*) serta data uji (*testing data*). Data latih dipakai dalam proses pembentukan model, sementara data uji dipakai untuk mengevaluasi kemampuan model ketika menjalankan prediksi terhadap data baru yang belum pernah dipelajari sebelumnya (Salazar et al., 2022). Tujuan utama penggunaan metode *Train-Test Split* adalah untuk mengukur kemampuan generalisasi model sehingga hasil evaluasi yang diperoleh lebih objektif dan dapat merepresentasikan kinerja model pada data baru. Pembagian dataset biasanya dilakukan menggunakan proporsi tertentu, seperti halnya 80% data latih serta 20% data uji, meskipun proporsi tersebut dapat disesuaikan dengan jumlah serta karakteristik data yang digunakan (Bichri et al., 2024).

Regresi

Regresi merupakan salah satu metode dalam *machine learning* yang dipakai untuk melakukan analisis korelasi antara variabel independen (X) dan variabel dependen (Y) yang mempunyai nilai numerik atau kontinu. Metode ini banyak digunakan dalam berbagai bidang untuk analisis data dan prediksi (Hasanah, 2023).

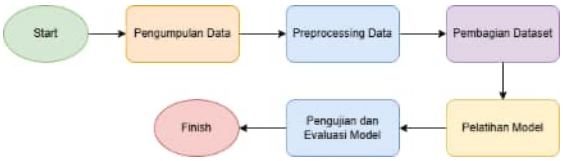
Decision Tree Regression

Decision Tree Regression merupakan salah satu algoritma regresi dalam *machine learning* yang menggunakan struktur pohon keputusan untuk memodelkan hubungan antara variabel independen dengan variabel dependen yang bersifat numerik. Algoritma ini bekerja dengan membangun model prediksi berbentuk pohon yang terdiri atas *root node*, *internal node*, dan *leaf node*. *Root node* merupakan titik awal pembentukan pohon, internal node berfungsi sebagai pemisah data berdasarkan atribut tertentu, sedangkan *leaf node* menunjukkan hasil prediksi atau nilai keluaran (Karim et al., 2021). Metode *Decision Tree Regression* melakukan proses pembagian data secara bertahap berdasarkan atribut yang dianggap paling optimal dalam mengurangi kesalahan prediksi (Nurani et al., 2023). Setiap percabangan pada pohon merepresentasikan aturan keputusan tertentu hingga data terbagi ke dalam kelompok-kelompok yang memiliki karakteristik serupa (Jumin et al., 2021).

METODE PENELITIAN

Riset ini termasuk ke dalam penelitian kuantitatif karena menggunakan data berbentuk angka serta pengolahan statistik dalam tahapan pengembangan model prediksi harga kos. Data yang dimanfaatkan mencakup atribut kos seperti ukuran kamar, fasilitas yang tersedia di kamar, fasilitas kamar mandi, fasilitas umum, serta jarak kos terhadap kampus Politeknik Negeri Bali yang diproses secara terstruktur guna menghasilkan informasi yang bersifat prediktif. Tahap-tahap penelitian yang dilakukan dalam riset ini ditampilkan pada Gambar 1 berikut ini.

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.



Gambar 1. Tahapan Penelitian

Gambar 1 memperlihatkan tahapan penelitian yang digunakan untuk membangun model prediksi harga kos menggunakan algoritma *Decision Tree Regression*. Algoritma ini dipilih karena mampu menghasilkan model yang mudah dipahami dan divisualisasikan sehingga mempermudah interpretasi hasil prediksi. Selain itu, *Decision Tree Regression* mampu menangani data numerik maupun kategorikal dengan baik serta tidak memerlukan asumsi hubungan linear antarvariabel. Karakteristik tersebut sesuai dengan dataset penelitian yang terdiri atas kombinasi atribut numerik dan kategorikal, seperti luas kamar, jarak ke kampus, lokasi, tipe kos, serta berbagai fasilitas yang digunakan untuk memprediksi harga kos. Penelitian diawali dengan proses pengumpulan data rumah kos di wilayah Jimbaran, kemudian dilanjutkan dengan *preprocessing* data untuk membersihkan dan mempersiapkan data sebelum digunakan. Selanjutnya, dataset dibagi menjadi data latih serta data uji memakai metode *Train-Test Split*. Data latih digunakan untuk proses *hyperparameter tuning* menggunakan *GridSearchCV* dengan metode *Stratified K-Fold Cross Validation* sebanyak 5 *fold*. Proses ini bertujuan memperoleh kombinasi parameter terbaik sebelum model akhir diuji menggunakan data uji. Selanjutnya, data uji digunakan untuk mengevaluasi performa model menggunakan metrik MAE, RMSE, dan R^2. Hasil evaluasi digunakan untuk menilai performa model dalam memprediksi harga kos.

Pengumpulan Data

Data yang dipakai pada penelitian ini merupakan data primer berupa data rumah kos di wilayah Jimbaran. Data diperoleh melalui proses pengumpulan manual dari berbagai kos yang berada di sekitar kawasan Jimbaran. Dataset yang digunakan terdiri dari sekitar 1100 data kos dengan atribut numerik dan kategorikal. Setelah dilakukan proses pembersihan data untuk mengatasi data yang tidak lengkap serta tidak konsisten, diperoleh sebanyak 1081 data yang siap digunakan pada tahap penelitian selanjutnya. Selanjutnya, data tersebut digunakan untuk proses *preprocessing* dan pembangunan model prediksi harga kos. Atribut yang dipakai pada penelitian ini mencakup lokasi, luas kamar, jenis kamar mandi, fasilitas kamar, fasilitas kamar mandi, fasilitas umum, jarak ke kampus Politeknik Negeri Bali, serta harga kos per bulan sebagai variabel target. Atribut-atribut tersebut dipakai selaku pedoman untuk mengembangkan model prediksi harga kos menggunakan algoritma *Decision Tree Regression*.

***Preprocessing* Data**

Tahap *preprocessing* dilakukan untuk memastikan data siap dipakai pada proses pelatihan model *machine learning*. Proses dimulai dengan pembersihan data melalui penghapusan data duplikat, penyeragaman format teks, serta penghapusan data harga kos yang kosong atau tidak valid. Selanjutnya dilakukan penyaringan harga kos pada rentang Rp1.000.000 hingga Rp6.000.000 untuk mengurangi pengaruh *outlier*. Atribut ukuran kamar diproses menjadi luas kamar dengan memisahkan nilai panjang dan lebar, sedangkan atribut jarak ke kampus dibersihkan dari satuan teks dan dikonversi menjadi numerik. Nilai kosong dalam atribut numerik kemudian diisi memakai nilai median. Dalam atribut fasilitas kamar, fasilitas kamar mandi, dan fasilitas umum dilakukan proses *multilabel feature extraction* dengan mengubah setiap fasilitas menjadi fitur biner bernilai 0 dan 1. Penelitian ini juga menerapkan *feature engineering* untuk menghasilkan fitur baru seperti total fasilitas, premium score, kategori jarak, rasio fasilitas terhadap luas kamar, dan transformasi logaritmik jarak. Setelah itu, variabel target harga kos ditransformasikan menggunakan fungsi logaritma natural (`log1p`) untuk meningkatkan stabilitas model. Pada tahap akhir, fitur kategorikal seperti lokasi, tipe kos, dan kategori jarak diproses menggunakan metode *One Hot Encoding*, sedangkan nilai kosong pada data kategorikal ditangani menggunakan `SimpleImputer`.

Pembagian Dataset

Dataset dibagi menggunakan metode *Train-Test Split* dengan komposisi 80% atau 864 data latih dan 20% atau 217 data uji. Data latih digunakan untuk proses pelatihan model serta pencarian kombinasi *hyperparameter* terbaik menggunakan *GridSearchCV*. Pada proses tersebut diterapkan metode *Stratified K-Fold Cross Validation* sebanyak 5 fold sehingga setiap kombinasi parameter dapat dievaluasi secara berulang. Setelah parameter terbaik diperoleh, model akhir diuji menggunakan data uji yang tidak pernah digunakan selama proses pelatihan maupun validasi sehingga hasil evaluasi dapat menggambarkan kemampuan generalisasi model.

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Pelatihan Model

Model prediksi dibangun memakai algoritma *Decision Tree Regression* melalui bantuan library *Scikit-learn*. Sebelum proses pelatihan, data diproses menggunakan *ColumnTransformer* dan *Pipeline* untuk menggabungkan tahapan *preprocessing* dan pelatihan model secara terstruktur. Fitur kategorikal diproses memakai *One Hot Encoding*, sedangkan fitur numerik digunakan secara langsung (*passthrough*). Variabel target harga kos ditransformasikan menggunakan fungsi logaritma natural (log1p) untuk meningkatkan stabilitas model regresi. Proses pelatihan model dilakukan memakai metode *GridSearchCV* untuk mencari kombinasi *hyperparameter* terbaik berdasarkan atas nilai R^2. *Hyperparameter* yang diuji terdiri dari max_depth, min_samples_leaf, min_samples_split, dan max_leaf_nodes. Selain itu, penelitian ini menggunakan metode *Stratified K-Fold Cross Validation* sebanyak 5 fold untuk mengevaluasi setiap kombinasi *hyperparameter* secara konsisten pada data pelatihan sehingga diperoleh *hyperparameter* dengan performa terbaik.

Tabel 1. Pelatihan Model

Hyperparameter	Nilai
max_depth	8, 10, 12, 15
min_samples_split	5, 10, 15
min_samples_leaf	2, 4, 6
max_leaf_nodes	40, 60, 80

Tabel 1 menunjukkan daftar kombinasi *hyperparameter* yang diuji pada proses *hyperparameter tuning* menggunakan *GridSearchCV* untuk memperoleh model *Decision Tree Regression* terbaik. Parameter `max_depth` diuji pada nilai 8, 10, 12, dan 15 untuk menentukan kedalaman maksimal pohon keputusan agar model tidak terlalu kompleks maupun terlalu sederhana. Parameter `min_samples_split` dilakukan pengujian terhadap nilai 5, 10, dan 15 untuk mengatur total minimal data yang dibutuhkan sebelum suatu node dibagi menjadi cabang baru. Selanjutnya, `min_samples_leaf` dilakukan pengujian terhadap nilai 2, 4, dan 6 untuk menentukan total minimal data pada setiap node akhir (*leaf node*) sehingga model lebih stabil dan tidak mudah mengalami *overfitting*. Parameter `max_leaf_nodes` diuji pada nilai 40, 60, dan 80 untuk membatasi jumlah node akhir pada pohon keputusan agar model tetap optimal. Pengujian beberapa kombinasi parameter tersebut dilakukan menggunakan *GridSearchCV* yang menerapkan metode *Stratified K-Fold Cross Validation* sebanyak 5 *fold* pada data pelatihan. Nilai R^2 dari setiap *fold* dihitung dan dirata-ratakan untuk menentukan kombinasi *hyperparameter* dengan performa terbaik. Selanjutnya, kombinasi *hyperparameter* terbaik digunakan untuk membangun model akhir yang kemudian diuji menggunakan data pengujian yang tidak digunakan selama proses pelatihan maupun validasi sehingga kemampuan generalisasi model dapat dievaluasi.

Pengujian dan Evaluasi Model

Model yang sudah dilatih kemudian diuji memakai data uji untuk mengetahui kemampuan prediksi terhadap data baru. Evaluasi model dilakukan pada data latih dan data uji memakai metrik regresi berupa *Mean Absolute Error* (MAE), *Root Mean Squared Error* (RMSE), dan *R-Squared* (R^2). Nilai prediksi yang sebelumnya ditransformasikan menggunakan logaritma dikembalikan ke bentuk aslinya menggunakan fungsi expm1 agar hasil evaluasi sesuai dengan harga kos sebenarnya. Selain itu, dilakukan analisis selisih nilai R^2 antara data latih serta data uji untuk mengetahui indikasi *overfitting* pada model.

MAE (*Mean Absolute Error*) mengukur rata-rata selisih absolut antara nilai prediksi (ŷi) dan nilai aktual (yi) pada seluruh data pengamatan (n). Semakin kecilnya nilainya, maka semakin baik performa model.

$$MAE = \frac{1}{n}\sum_{i}^{n} = 1 \mid yi - \hat{y}i \mid$$

RMSE (*Root Mean Squared Error*) mengukur rata-rata kesalahan prediksi melalui akar kuadrat dari rata-rata kuadrat selisih antara nilai aktual (yi) dan nilai prediksi (ŷi) pada seluruh data pengamatan (n). Semakin kecil nilai RMSE, semakin baik performa model.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i}^{n} = 1\,(yi - \hat{y}i)^2}$$

R^2 (*Koefisien Determinasi atau R-Squared*) digunakan untuk kemampuan model dalam menjelaskan variasi data target berdasarkan perbandingan antara jumlah kuadrat error $\sum_{i}^{n} = 1\,(yi - \hat{y}i)^2$ dengan jumlah kuadrat total $\sum_{i}^{n} = 1\,(yi - \bar{y}i)^2$ di mana (yi) merupakan nilai aktual, (ŷi) merupakan nilai hasil prediksi, dan (ȳ) merupakan rata-rata nilai aktual. Nilai R^2 berada pada rentang 0 sampai 1. Semakin mendekati 1, semakin baik kemampuan model dalam menjelaskan hubungan anatara variabel input dengan output.

$$R^2 = \frac{\sum_{i}^{n}=1\,(yi-\hat{y}i)^2}{\sum_{i}^{n}=(yi-\bar{y}i)^2}$$

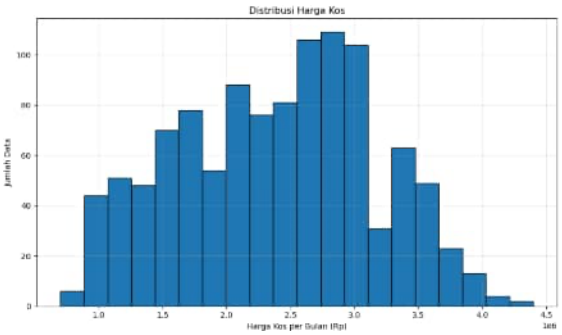
This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

ISSN: 2598-8565 (media cetak)
ISSN: 2620-4339 (media online)

HASIL DAN PEMBAHASAN

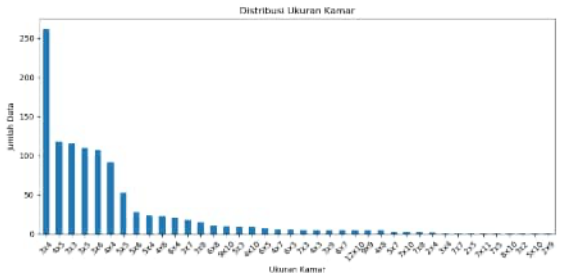
Visualisasi Dataset

Distribusi data pada variabel harga kos, luas kamar, serta jarak menuju kampus Politeknik Negeri Bali divisualisasikan untuk mengidentifikasi sebaran jumlah data pada masing-masing interval nilai. Visualisasi tersebut memungkinkan pengamatan terhadap frekuensi kemunculan data, seperti banyaknya kos yang berada pada kisaran harga tertentu, memiliki ukuran kamar tertentu, maupun berlokasi pada jarak tertentu dari kampus. Adapun hasil visualisasi distribusi data tersebut disajikan pada Gambar 2 hingga Gambar 4 berikut.



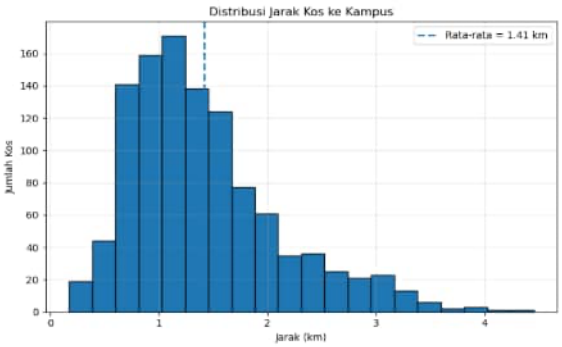
Gambar 2. Distribusi Harga Kos

Berdasarkan Gambar 2, sebagian besar data kos berada pada rentang harga Rp1.500.000 sampai Rp3.000.000 per bulan. Hal ini menunjukkan bahwa mayoritas kos di wilayah Jimbaran termasuk dalam kategori harga menengah. Selain itu, jumlah kos dengan harga di atas Rp5.000.000 relatif lebih sedikit sehingga distribusi data tidak merata pada seluruh rentang harga.



Gambar 3. Ukuran Kamar

Berdasarkan Gambar 3, sebagian besar kos memiliki luas kamar antara 9 m² hingga 16 m². Kondisi ini menunjukkan bahwa ukuran kamar sedang merupakan tipe kamar yang paling banyak tersedia pada dataset penelitian.



Gambar 4. Jarak Ke Kampus Politeknik Negeri Bali

Berdasarkan Gambar 4, sebagian besar kos ditemukan dalam jarak yang tidak mencapai 3 km dari kampus. Gambar tersebut memperlihatkan bahwa lokasi yang dekat dengan kampus menjadi karakteristik dominan pada data kos yang digunakan dalam penelitian.

Hasil analisis distribusi data menunjukkan bahwa dataset memiliki variasi karakteristik yang cukup beragam pada aspek harga, luas kamar, dan jarak ke kampus. Variasi tersebut memberikan informasi yang cukup bagi model *Decision Tree Regression* untuk mempelajari pola hubungan antara atribut kos dan harga sewa sehingga mendukung proses pembangunan model prediksi.

Hasil Preprocessing Data

Hasil *preprocessing* menunjukkan bahwa seluruh data berhasil dibersihkan dan ditransformasikan ke dalam format yang sesuai untuk proses *machine learning*. Proses yang dilakukan meliputi pembersihan nama kolom, penghapusan data duplikat, validasi tipe data, filtering harga, pengolahan ukuran kamar dan jarak kampus, pengisian nilai kosong, *encoding* fasilitas, *feature engineering*, transformasi logaritma, stratifikasi data, dan *One Hot Encoding*. Hasil *preprocessing* dapat dilihat dalam Tabel 2 hingga Tabel 4 dibawah ini.

Tabel 2. Hasil Filtering Harga

Kriteria	Nilai
Harga Minimum	Rp1.000.000
Harga Maksimum	Rp6.000.000

Tabel 2 menampilkan proses filtering harga, menunjukkan bahwa data harga kos yang digunakan dalam penelitian berada pada rentang Rp1.000.000 hingga Rp6.000.000 per bulan. Pembatasan rentang harga ini dilakukan untuk mengurangi pengaruh data ekstrem (*outlier*) yang dapat menyebabkan model menghasilkan pola prediksi yang kurang stabil. Dengan mempertahankan data pada rentang harga yang umum

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

ditemukan di wilayah Jimbaran, sehingga model mampu lebih fokus mempelajari karakteristik kos yang memiliki keterkaitan sehingga menambahkan kualitas proses pelatihan dan akurasi prediksi.

Tabel 3. Hasil Encoding Fasilitas

Data Awal	AC	Wifi	Lokasi
AC, Wifi, Lokasi	1	1	1
Wifi, Lokasi	0	1	1
AC	1	0	0

Tabel 3 menampilkan proses *encoding* menunjukkan bahwa data fasilitas yang sebelumnya berbentuk teks berhasil diubah menjadi data numerik biner bernilai 0 dan 1. Proses ini memungkinkan algoritma *Decision Tree Regression* mengenali keberadaan setiap fasilitas secara individual. Selain itu, teknik ini mampu mempertahankan informasi fasilitas secara lebih detail dibandingkan jika seluruh fasilitas digabungkan menjadi satu atribut.

Tabel 4. Hasil *Feature Engineering*

Nama Fitur	Fungsi
total_fasilitas	Menghitung jumlah fasilitas kos
premium_score	Mengukur kualitas fasilitas premium
kamar_besar	Menandai kamar berukuran besar
super_dekat	Menandai kos yang sangat dekat dengan kampus
fasilitas_per_m^2	Rasio fasilitas terhadap luas kamar
premium_per_m^2	Rasio fasilitas premium terhadap luas kamar

Tabel 4 menampilkan hasil proses *feature engineering* menghasilkan beberapa fitur baru yang tidak tersedia pada data asli. Fitur-fitur tersebut dirancang untuk merepresentasikan kualitas kos secara lebih baik dibandingkan atribut dasar. Sebagai contoh, fitur premium_score mampu menggambarkan tingkat eksklusivitas fasilitas yang dimiliki kos, sedangkan fitur fasilitas_per_m2 dan premium_per_m2 menggambarkan kualitas fasilitas relatif terhadap ukuran kamar. Berdasarkan hasil pengujian model, kedua fitur tersebut memberikan kontribusi yang signifikan terhadap peningkatan performa prediksi karena mampu memiliki hubungan yang lebih kompleks antara fasilitas dan harga kos.

Hasil Pelatihan Model

Model dibangun menggunakan algoritma *Decision Tree Regression* dengan bantuan library *Scikit-learn*. Penelitian ini menggunakan *Pipeline* dan *ColumnTransformer* agar *preprocessing* dan pelatihan model berjalan secara terintegrasi. Selain itu, dilakukan proses hyperparameter *tuning* menggunakan *GridSearchCV* dan *Stratified K-Fold Cross Validation* sebanyak 5 fold.

Pada proses *tuning*, beberapa kombinasi nilai hyperparameter diuji, yaitu max_depth sebesar 8, 10, 12, 15 ; min_samples_split sebesar 5, 10, 15 ; min_samples_leaf sebesar 2, 4, 6 ; serta max_leaf_nodes sebesar 40, 60, 80. Setiap kombinasi parameter dievaluasi menggunakan metrik R^2 pada proses *cross validation* untuk menentukan konfigurasi yang menghasilkan model dengan kinerja terbaik. Evaluasi pada tahap ini dilakukan menggunakan data pelatihan (*training data*) melalui metode *Cross Validation*. Tujuan evaluasi ini adalah untuk menentukan kombinasi hyperparameter terbaik sebelum model diuji menggunakan data pengujian (*testing data*). Hasil pengujian dengan beberapa parameter ditunjukkan pada Tabel 5 di bawah ini.

Tabel 5. Hasil Pengujian Kombinasi Parameter

No	max_ depth	min_ samples_ split	min_ samples_ leaf	max_ leaf_ nodes	mean R^2 CV
1	8	5	2	80	0.7825
2	15	5	2	80	0.7824
3	8	5	2	80	0.7824
4	10	5	2	80	0.7807
5	15	10	2	80	0.7783
6	12	10	2	80	0.7783
7	10	10	2	80	0.7779
8	10	10	4	80	0.7765
9	12	10	4	80	0.7755
10	15	10	4	80	0.7755

Berdasarkan Tabel 5, kombinasi max_depth = 8, min_samples_split = 5, min_samples_leaf = 2, dan max_leaf_nodes = 80 menghasilkan nilai Mean R^2 *cross validation* tertinggi sebesar 0,7825 sehingga dipilih sebagai kombinasi parameter terbaik. Nilai tersebut menunjukkan bahwa model mampu menjelaskan variasi data dengan lebih baik dibandingkan kombinasi parameter lainnya. Nilai Mean R^2 yang diperoleh digunakan sebagai dasar pemilihan parameter terbaik sebelum dilakukan evaluasi menggunakan data pengujian (*testing data*).

Kombinasi ini menghasilkan performa terbaik karena mampu membentuk pohon keputusan yang cukup kompleks untuk menangkap pola data, namun tetap terkontrol sehingga model tidak mudah

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

mengalami *overfitting*. Sementara itu, kombinasi dengan nilai max_depth yang lebih besar atau min_samples_split dan min_samples_leaf yang lebih tinggi menghasilkan nilai Mean R^2 yang lebih rendah. Hal ini menunjukkan bahwa penambahan kompleksitas pohon maupun pembatasan pemecahan node yang terlalu besar tidak mampu meningkatkan kemampuan model dalam memprediksi harga kos. Hasil pengujian dengan parameter terbaik ditunjukkan pada Tabel 6.

Tabel 6. Hasil *Hyperparameter Tuning*

Hyperparameter	Nilai
max_depth	8
min_samples_split	5
min_samples_leaf	2
max_leaf_nodes	80

Berdasarkan Tabel 6, diperoleh parameter terbaik yaitu max_depth = 8, min_samples_split = 5, min_samples_leaf = 2 dan max_leaf_nodes = 80. Nilai max_depth = 8 menghasilkan kompleksitas pohon yang cukup untuk menangkap pola data tanpa menyebabkan *overfitting*, sedangkan parameter lainnya membantu meningkatkan kemampuan generalisasi model. Oleh karena itu, kombinasi parameter tersebut dipilih sebagai model akhir pada penelitian ini.

Hasil Pengujian Model

Hasil pengujian model menunjukkan tingkat performa algoritma *Decision Tree Regression* dalam memprediksi harga kos berdasarkan *data testing*. Evaluasi dilakukan menggunakan metrik *R-Squared* (R^2), *Mean Absolute Error* (MAE), dan *Root Mean Squared Error* (RMSE) untuk mengukur akurasi dan tingkat kesalahan prediksi model. Hasil evaluasi model dapat dilihat pada Tabel 7 di bawah ini.

Tabel 7. Hasil Pengujian Model

Data	R^2	MAE	RMSE
Training	0.8871	Rp189.898	Rp251.789
Testing	0.8326	Rp225.438	Rp303.080

Berdasarkan Tabel 7, model memperoleh nilai R^2 sebesar 0,8871 pada *data training* dan 0,8326 pada *data testing*. Hasil tersebut menunjukkan bahwa model mampu menjelaskan sekitar 83,26% variasi harga kos berdasarkan karakteristik fisik kamar, fasilitas yang disediakan dan lokasi kos. Hasil ini menunjukkan bahwa atribut-atribut yang digunakan memiliki hubungan yang kuat terhadap harga sewa kos di wilayah Jimbaran sehingga model mampu menghasilkan prediksi yang cukup baik.

Performa model yang cukup baik dipengaruhi oleh tahapan *preprocessing* data, *feature engineering*, serta proses *hyperparameter tuning* menggunakan *GridSearchCV* dengan *Stratified K-Fold Cross Validation*. Tahapan tersebut membantu model memperoleh kombinasi *hyperparameter* yang optimal sehingga mampu menangkap hubungan antara karakteristik kos dan harga sewa secara lebih baik.

Selain itu, nilai MAE sebesar Rp225.438 dan RMSE sebesar Rp303.080 menunjukkan bahwa rata-rata kesalahan prediksi model relatif kecil sehingga model masih mampu memberikan hasil prediksi yang cukup baik. Nilai MAE mengindikasikan bahwa rata-rata selisih antara harga prediksi dan harga aktual relatif kecil dibandingkan rentang harga kos pada dataset. Sementara itu, nilai RMSE yang sedikit lebih besar menunjukkan adanya beberapa data dengan kesalahan prediksi yang lebih tinggi, namun secara keseluruhan model masih mampu menghasilkan prediksi yang stabil. Selisih nilai R^2 antara *data training* dan *testing* sebesar 0,0546 menunjukkan bahwa model tidak mengalami *overfitting* yang signifikan dan masih mampu melakukan generalisasi dengan baik terhadap data baru. Hasil penelitian ini memperoleh nilai R^2 sebesar 0,8326 pada *data testing*, yang menunjukkan bahwa model mampu menjelaskan sekitar 83,26% variasi harga kos. Hasil tersebut lebih tinggi dibandingkan penelitian (Saputra & Pratama, 2025) yang memperoleh nilai R^2 sebesar 0,5822 pada prediksi harga rumah di daerah Tebet, sehingga menunjukkan bahwa model pada penelitian ini memiliki kemampuan yang lebih baik dalam menjelaskan variasi data target. Perbedaan hasil tersebut dipengaruhi oleh penerapan tahapan *preprocessing* data, *feature engineer*ing, serta *hyperparameter tuning* menggunakan *GridSearchCV* dengan *Stratified K-Fold Cross Validation* yang membantu menghasilkan model dengan kemampuan generalisasi yang lebih baik. Hasil *feature importance* ditunjukkan pada Tabel 8 di bawah ini.

Tabel 8. Feature Importance

Fitur	Importance
luas_kamar	0.068457
jarak_ke_kampus	0.041765
fasilitas_per_m^2	0.036440
total_fasilitas	0.437743

Tabel 8 menampilkan empat fitur dengan nilai *feature importance* tertinggi yang paling berpengaruh terhadap prediksi harga kos. Hasil *feature importance* menunjukkan bahwa fitur total_fasilitas memberikan kontribusi terbesar terhadap proses prediksi harga kos dengan nilai *importance* sebesar 43,77%. Hal ini

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

mengindikasikan bahwa jumlah fasilitas yang tersedia merupakan faktor dominan yang memengaruhi harga sewa kos. Sementara itu, fitur luas_kamar (6,85%), jarak_ke_kampus (4,18%), dan fasilitas_per_m2 (3,64%) juga berkontribusi dalam pembentukan model, meskipun pengaruhnya relatif lebih kecil. Hasil ini menunjukkan bahwa kelengkapan fasilitas memiliki hubungan yang lebih kuat terhadap harga kos dibandingkan faktor ukuran kamar maupun jarak ke kampus.

Hasil penelitian ini menunjukkan bahwa kelengkapan fasilitas menjadi faktor utama dalam menentukan harga kos di wilayah Jimbaran. Semakin lengkap fasilitas yang disediakan, seperti ac, wifi, kamar mandi dalam, dan fasilitas pendukung lainnya, semakin tinggi juga harga sewa yang ditawarkan. Kondisi ini menunjukkan bahwa calon penyewa tidak hanya mempertimbangkan ukuran kamar maupun kedekatan lokasi dengan kampus, tetapi juga kualitas dan kelengkapan fasilitas yang diperoleh. Oleh karena itu, fitur total_fasilitas menjadi atribut yang paling berpengaruh dalam pembentukan model prediksi harga kos.

KESIMPULAN

Penelitian ini berhasil membangun model prediksi harga kos di wilayah Jimbaran menggunakan metode *Decision Tree Regression*. Model yang dikembangkan mampu memprediksi harga kos berdasarkan karakteristik fisik kamar, fasilitas yang tersedia, serta lokasi kos yang memiliki tingkat akurasi cukup baik.

Hasil penelitian menunjukkan bahwa karakteristik fisik kamar, fasilitas, dan lokasi memiliki hubungan yang cukup kuat terhadap harga sewa kos. Tahapan *preprocessing* dan *feature engineering* yang diterapkan mampu membantu model dalam mengenali pola hubungan antar atribut secara lebih efektif. Berdasarkan analisis *feature importance*, total_fasilitas merupakan faktor yang paling berpengaruh dalam menentukan harga kos, diikuti oleh luas_kamar, jarak_ke_kampus, dan fasilitas per _meter persegi.

Evaluasi model menunjukkan bahwa *Decision Tree Regression* menghasilkan nilai R^2 sebesar 0,8326, MAE sebesar Rp225.438, dan RMSE sebesar Rp303.080 pada data pengujian. Nilai tersebut menunjukkan bahwa model mampu menjelaskan sekitar 83,26% variasi harga kos dan mempunyai tingkatan kesalahan prediksi yang relatif rendah. Hasil tersebut mengindikasikan bahwa *Decision Tree Regression* mampu memodelkan hubungan *non-linear* antar atribut dengan baik sehingga dapat dimanfaatkan sebagai pendukung pengambilan keputusan dalam prediksi harga kos di wilayah Jimbaran.

DAFTAR PUSTAKA

Ağbulut, Ü., Gürel, A. E., & Biçen, Y. (2021). Prediction of daily global solar radiation using different machine learning algorithms: Evaluation and comparison. *Renewable and Sustainable Energy Reviews*, *135*. https://doi.org/10.1016/j.rser.2020.110114

Athiyah, U., Hananta, A., Maulidi, T., Putra, V. M. E., Purba, T. F. H., & Bakowatun, E. A. W. (2021). Sistem Pendukung Keputusan Prediksi Harga Rumah Kost untuk Mahasiswa IT Telkom Purwokerto Menggunakan Metode Fuzzy Tsukamoto Berbasis Web. Journal of Dinda: Data Science, Information Technology, and Data Analytics, 1(2), 77-81.

Bichri, H., Chergui, A., & Hain, M. (2024). Investigating the Impact of Train / Test Split Ratio on the Performance of Pre-Trained Models with Custom Datasets. In *IJACSA) International Journal of Advanced Computer Science and Applications* (Vol. 15, Number 2). https://doi.org/10.14569/IJACSA.2024.0150235

Çetin, V., & Yıldız, O. (2022). A comprehensive review on data preprocessing techniques in data analysis. *Pamukkale University Journal of Engineering Sciences*, *28*(2), 299–312. https://doi.org/10.5505/pajes.2021.62687

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

Haryanto, C., Rahaningsih, N., & Muhammad Basysyar, F. (2023). Komparasi Algoritma Machine Learning Dalam Memprediksi Harga Rumah. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *7*(1), 533–539. https://doi.org/10.36040/jati.v7i1.6343

Hasanah, M., & Nurmalitasari, N. (2023, July). Prediksi Harga Indekost Menggunakan Analisis Regresi Linier Berganda. In *Prosiding Seminar Nasional Teknologi Informasi dan Bisnis* (pp. 313-318).

Jumin, E., Basaruddin, F. B., Yusoff, Y. B. M., Latif, S. D., & Ahmed, A. N. (2021). Solar radiation prediction using boosted decision tree regression model: A case study in Malaysia. *Environmental Science and Pollution Research*, *28*(21), 26571–26583. https://doi.org/10.1007/s11356-021-12435-6

Karim, R., Alam, M. K., & Hossain, M. R. (2021, August 10). Stock market analysis using linear regression and decision tree regression. *2021 1st International Conference on Emerging Smart Technologies and Applications, ESmarTA 2021*.

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

https://doi.org/10.1109/eSmarTA52612.2021.9515762

Labib Mu'tashim, M., Damayanti, S. A., Zaki, H. N., Muhayat, T., & Wirawan, R. (2021). *Analisis Prediksi Harga Rumah Sesuai Spesifikasi Menggunakan Multiple Linear Regression*. *3*. https://doi.org/10.52958/iftk.v17i3.3635

Muharrom, M. (2023). Analisis Penggunaan Orange Data Mining untuk Prediksi Harga USDT/BIDR Binance. *Bulletin of Information Technology (BIT)*, *4*(2), 178–184. https://doi.org/10.47065/bit.v3i1

Nurani, A. T., Setiawan, A., & Susanto, B. (2023). Perbandingan Kinerja Regresi Decision Tree dan Regresi Linear Berganda untuk Prediksi BMI pada Dataset Asthma. *Jurnal Sains Dan Edukasi Sains*, *6*(1), 34–43. https://doi.org/10.24246/juses.v6i1p34-43

Nurpadhilah, N. A., & Saputera, A. (2026). Sentiment Analysis and Characteristics of Youtube User Opinions Toward Samsung and Iphone Brands Using TF-IDF with Naive Bayes and KNN Comparison and Mcnemar Test. In *Jurnal Media Computer Science* (Vol. 5, Number 2). https://doi.org/10.37676/jmcs.v5i2.11311

Sholeh, M., Nurnawati, E. K., & Lestari, U. (2023). Penerapan Data Mining dengan Metode Regresi Linier untuk Memprediksi Data Nilai Hasil Ujian Menggunakan Rapid Miner. *Jurnal Informatika Sunan Kalijaga (JISKA), 8*(1), 10-21.

Qadri, A. M., Raza, A., Munir, K., & Almutairi, M. S. (2023). Effective Feature Engineering Technique for Heart Disease Prediction with Machine Learning. *IEEE Access*, *11*, 56214–56224. https://doi.org/10.1109/ACCESS.2023.3281484

Rahmadini, R., LorencisLubis, E. E., Priansyah, A., RWN, Y., & Meutia, T. (2023). Penerapan data mining untuk memprediksi harga bahan pangan di Indonesia menggunakan algoritma K-Nearest Neighbor. *Jurnal Mahasiswa Akuntansi Samudra, 4*(4), 223-235. https://doi.org/10.33059/jmas.v4i4.7074

Ramadhan, A. I., Almajid, N., & Ginting, D. Y. (2023). Penerapan Decision Tree Regression dalam Memprediksi Harga Rumah di Provinsi Jawa Barat. *Jejaring Penelitian dan Pengabdian Masyarakat (JPPM)*, *1*(3). https://doi.org/10.58776/jriti.v1i3.64

Rayssa Buntoro, D. (2025). Pergeseran Feature Importance pada Prediksi Pasar Saham Teknologi Menggunakan Machine Learning: Studi Komparatif Pra dan Pasca Pandemi COVID-19. *JAMI: Jurnal Ahli Muda Indonesia*, *6*(2), 114–119. https://doi.org/10.46510/jami.v6i2.380

Salazar, J. J., Garland, L., Ochoa, J., & Pyrcz, M. J. (2022). Fair train-test split in machine learning: Mitigating spatial autocorrelation for improved prediction accuracy. *Journal of Petroleum Science and Engineering*, *209*, 109885. https://doi.org/10.1016/j.petrol.2021.109885

Saputra, R. A., & Pratama, A. (2025). Implementasi Decision Tree Untuk Prediksi Harga Rumah di Daerah Tebet. *Journal of Information System Management (JOISM),* 6(2). https://doi.org/10.24076/joism.2025v6i2.1928

Wardana, F. A., & Juanita, S. (2025). Prediksi Jumlah Tenaga Kerja Asing di Jawa Barat Menggunakan Perbandingan Algoritma Support Vector Regression dan Decision Tree Regression. *JIPI (Jurnal Ilmiah Penelitian Dan Pembelajaran Informatika)*, *10*(2), 890–899. https://doi.org/10.29100/jipi.v10i2.6003

Wijaya, Y. F., & Triayudi, A. (2023). Penerapan Data Mining Pada Prediksi Harga Emas dengan Menggunakan Algoritma Regresi Linear Berganda dan ARIMA. *Journal of Computer System and Informatics (JoSYC)*, *5*(1), 73–81. https://doi.org/10.47065/josyc.v5i1.4615

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.