

Comparative Analysis of Linear Regression and Decision Tree Regressor and Random Forest for Household Energy Consumption Prediction

Dayanni Vera Versanika^{1*}, Adang Kurniawan²

¹Program Studi Sistem Informasi, STMIK Bandung, Indonesia

²Program Studi Sistem Informasi, Institut Digital Ekonomi LPKIA Bandung, Indonesia

Email: ¹dayanniveraversanika@gmail.com, ²adangkurniawan@lpkia.ac.id

INFORMASI ARTIKEL

Histori artikel:

Naskah masuk, 21 April 2026

Direvisi, 16 Mei 2026

Diiterima, 4 Juni 2026

Kata Kunci:

Konsumsi Energi Rumah Tangga;
Decision Tree Regressor;
Regresi Linear;
K-Fold Cross Validation;
GridSearchCV;
Random Forest;
Pemodelan Prediktif.

ABSTRAK

Abstract-Household energy consumption continues to rise alongside the rapid adoption of electrical appliances and evolving lifestyles. Many studies implicitly assume the superiority of complex nonlinear models without adequately considering data characteristics. This study addresses this gap through a comparative analysis of Linear Regression, Decision Tree Regressor, and Random Forest Regressor for predicting household electricity consumption. To enhance novelty and methodological rigor, the study incorporates feature engineering (adding a unit-tariff proxy feature), 5-Fold Cross Validation, and GridSearchCV hyperparameter tuning for Decision Tree Regressor and Random Forest Regressor. The dataset comprises 114 records with features including appliance type, usage duration, month, electricity cost, and a derived unit-tariff feature. Model performance is evaluated using RMSE, MAE, and R^2 . Results show that Linear Regression achieves the best test-set performance ($RMSE = 1.1552 \text{ kWh}$, $R^2 = 0.96$), while Decision Tree Regressor improves substantially after tuning (best params: $max_depth = 3$) compared to its untuned counterpart. Random Forest Regressor demonstrates strong generalization in cross-validation ($R^2 = 0.9523 \pm 0.0213$). Feature importance analysis confirms that usage duration (*waktu pemakaian*) is the dominant predictor (88.31%), reinforcing that linear models are more appropriate when the underlying relationship is predominantly linear.

Abstrak-Konsumsi energi rumah tangga terus meningkat seiring perkembangan perangkat elektronik dan perubahan gaya hidup. Banyak penelitian cenderung mengasumsikan keunggulan model non-linear tanpa mempertimbangkan karakteristik data yang digunakan. Penelitian ini mengisi kesenjangan tersebut melalui analisis komparatif antara *Linear Regression*, *Decision Tree Regressor*, dan *Random Forest Regressor* dalam memprediksi konsumsi energi rumah tangga. Untuk meningkatkan kebaruan dan ketelitian metodologi, penelitian ini mengintegrasikan *feature engineering* (penambahan fitur *proksi tarif per kWh*), *5-Fold Cross Validation*, dan *GridSearchCV* *hyperparameter tuning* untuk *Decision Tree Regressor* dan *Random Forest Regressor*. Dataset terdiri dari 114 data dengan fitur jenis alat listrik, waktu pemakaian, bulan, biaya listrik, dan fitur turunan tarif per kWh. Model dievaluasi menggunakan RMSE, MAE, dan R^2 . Hasil menunjukkan *Linear Regression* memberikan performa terbaik pada data uji ($RMSE = 1,1552 \text{ kWh}$, $R^2 = 0,96$), sementara *Decision Tree Regressor* meningkat signifikan setelah tuning ($max_depth = 3$). *Random Forest Regressor* menunjukkan generalisasi yang kuat pada *cross-validation* ($R^2 = 0,9523 \pm 0,0213$). Analisis kontribusi fitur mengonfirmasi bahwa waktu pemakaian merupakan prediktor dominan (88,31%), memperkuat relevansi model linear pada data dengan pola hubungan yang didominasi sifat linier.

Copyright © 2026 LPPM - STMIK IKMI Cirebon
This is an open access article under the CC-BY license

Penulis Korespondensi:

Dayanni Vera Versanika

Program Studi Sistem Informasi,
STMIK Bandung

Jl. Cikutra No. 113, Bandung, Indonesia

Email: dayanniveraversanika@gmail.com

1. Pendahuluan

Kebutuhan energi listrik pada skala rumah tangga terus mengalami peningkatan seiring dengan

perkembangan teknologi dan meningkatnya penggunaan perangkat elektronik dalam aktivitas sehari-hari [1]. Peningkatan tersebut tidak hanya

memberikan tekanan terhadap sistem distribusi energi, tetapi juga berkontribusi pada tingginya biaya operasional dan permintaan energi secara global [2]. Data *International Energy Agency* (IEA) menegaskan bahwa sektor bangunan yang di dalamnya mencakup hunian residensial menyumbang sekitar 28% dari total konsumsi energi global dan hampir sepertiga dari keseluruhan emisi gas rumah kaca yang berkaitan dengan penggunaan energi [3]. Lebih jauh, porsi konsumsi energi sektor residensial terhadap total konsumsi bangunan diperkirakan berada pada kisaran 50 hingga 70 persen, bergantung pada kondisi iklim dan kebijakan energi masing-masing negara [4].

Kondisi ini diperparah oleh beragamnya faktor yang memengaruhi pola konsumsi energi rumah tangga, mulai dari variabel iklim, karakteristik hunian, hingga faktor sosioekonomi seperti pendapatan, tingkat pendidikan, dan ukuran keluarga [5]. Oleh karena itu, kemampuan untuk memprediksi konsumsi energi secara akurat menjadi kebutuhan mendasar dalam mendukung pengambilan keputusan yang efektif di bidang manajemen energi [6].

Dalam beberapa tahun terakhir, teknik *machine learning* telah banyak diterapkan untuk keperluan prediksi konsumsi energi. Berbagai algoritma telah diuji, mulai dari *Linear Regression*, *Decision Tree Regressor*, hingga metode *ensemble* seperti *Random Forest* dan *Gradient Boosting* [7]. *Regresi linear*, meskipun lebih sederhana, telah terbukti efektif sebagai alat prediksi konsumsi energi pada berbagai konteks aplikasi nyata, termasuk prediksi konsumsi supermarket dan pemodelan berbasis perangkat IoT [8]. Di sisi lain, model *Decision Tree Regressor* diketahui mampu menangkap hubungan non-linear yang lebih kompleks, namun berisiko mengalami *overfitting* apabila tidak dilakukan pengaturan *hyperparameter* yang memadai [9], [10].

Kajian terhadap literatur yang ada mengungkapkan bahwa sebagian besar penelitian cenderung mengasumsikan superioritas model non-linear tanpa mengevaluasi secara mendalam kesesuaiannya dengan karakteristik dataset yang digunakan [11]. Padahal, kinerja suatu model sangat bergantung pada distribusi data, hubungan antarvariabel, dan kompleksitas pola yang terkandung dalam dataset [12], [13]. Penelitian yang secara eksplisit membandingkan *Linear Regression* dan *Decision Tree Regressor* pada dataset konsumsi energi rumah tangga berskala kecil hingga menengah, sekaligus mengintegrasikan *hyperparameter tuning*, *cross-validation*, dan

ensemble learning sebagai bagian dari evaluasi, masih sangat terbatas [13], [14].

Penelitian ini mengisi kesenjangan tersebut melalui analisis komparatif yang diperluas: (1) *Linear Regression*, (2) *Decision Tree Regressor* dengan *GridSearchCV* untuk *hyperparameter tuning*, dan (3) *Random Forest Regressor* sebagai metode *ensemble*. Seluruh model dievaluasi menggunakan *5-Fold Cross Validation* guna menghasilkan estimasi performa yang lebih andal dan tidak bergantung pada satu partisi data tertentu. Penelitian ini juga menerapkan *feature engineering* dengan menambahkan fitur proksi tarif per kWh (*tarif_per_kwh*) untuk memperkaya representasi data.

2. Tinjauan Pustaka

Berbagai penelitian telah mengkaji penerapan *machine learning* untuk prediksi konsumsi energi dari berbagai sudut pandang. Rambabu et al. [15] menyatakan bahwa algoritma berbasis pohon menghasilkan kinerja terbaik dalam memprediksi konsumsi energi rumah tangga menggunakan variabel suhu, kelembaban, dan waktu. Moulla et al. [16] membuktikan bahwa *Decision Tree Regressor* dan *Random Forest* mampu memprediksi konsumsi listrik residensial dengan nilai MSE, MAE, dan MAPE yang mendekati nol pada dataset nasional Afrika Selatan. Adapun Nemade et al. [17] yang membandingkan lima algoritma *machine learning*, termasuk *Linear Regression*, *Decision Tree Regressor*, *Random Forest*, *XGBoost*, dan *Gradient Boosting*, dan menemukan bahwa metode *ensemble* cenderung unggul, namun *Linear Regression* tetap memberikan acuan kinerja dasar yang penting.

Airlangga [11] menemukan bahwa pada dataset konsumsi energi rumah tangga tertentu, seluruh model yaitu termasuk *ensemble* dengan menunjukkan nilai R^2 negatif, yang mengindikasikan bahwa kompleksitas algoritma tidak secara otomatis menjamin kualitas prediksi yang lebih baik. Kama et al. [9] menegaskan bahwa meskipun *Decision Tree Regressor* berisiko *overfitting*, sedangkan *Linear Regression* tetap kompetitif sebagai model baseline yang andal. Singla dan Thakur [18] mengkonfirmasi temuan tersebut dengan menunjukkan bahwa *Decision Tree Regressor* cenderung *overfit* pada data pelatihan, sementara metode *ensemble* menunjukkan kemampuan generalisasi yang lebih baik. Mosetlhe dan Yusuff [19] menunjukkan bahwa akurasi *Decision Tree Regressor* dapat ditingkatkan hingga 44% melalui penyetelan *hyperparameter*, menekankan pentingnya proses optimasi bukan semata-mata jenis algoritma.

Dari sisi faktor penentu konsumsi energi, Chowdhury et al. [20] menekankan bahwa faktor

sosioekonomi seperti pendapatan, ukuran keluarga, dan tingkat pendidikan merupakan variabel penting dalam pemodelan prediktif konsumsi energi rumah tangga di Amerika Serikat. Studi tersebut menggunakan dataset yang mengintegrasikan data cuaca dan atribut rumah tangga, menunjukkan bahwa pemilihan fitur yang tepat sangat berpengaruh terhadap kinerja model. Sementara itu, Phan et al. [21] membuktikan bahwa *Linear Regression* efektif untuk prediksi energi dalam skenario nyata, asalkan dikonstruksi dengan tepat dan dilengkapi *feature engineering* yang memadai.

Atici dan Tuna [12] dalam analisis komparatif terhadap tujuh model regresi untuk efisiensi energi bangunan menyimpulkan bahwa metode non-linear secara umum mengungguli *regresi linear*, namun analisis *feature importance* tetap diperlukan untuk memahami faktor dominan dalam prediksi. Mosetlhe dan Yusuff [13] menunjukkan bahwa akurasi model *Decision Tree Regressor* dapat ditingkatkan hingga 44% melalui penyetelan hyperparameter, sebuah temuan yang menekankan bahwa kinerja model sangat dipengaruhi oleh proses optimasi, bukan semata-mata jenis algoritma. Azmi dan Saleh [14] secara spesifik memperoleh hasil terbaik dari *Decision Tree Regressor* dengan nilai MAE 0,3613 dan RMSE 0,6469, menjadikannya model paling andal dalam prediksi konsumsi daya residensial pada studi tersebut.

Aspek interpretabilitas model juga mendapatkan perhatian dalam penelitian terkini. Beshaw et al. [22] menggabungkan pendekatan machine learning dengan nilai SHAP (*Shapley Additive Explanations*) untuk meningkatkan transparansi dan kepercayaan terhadap prediksi konsumsi energi, memperkuat urgensi analisis kontribusi fitur sebagai bagian integral dari evaluasi model. Klyuev et al. [23] dalam tinjauan literatur menegaskan bahwa metode berbasis regresi tetap relevan karena kesederhanaannya, dan bahwa pemilihan metode prediksi yang tepat harus disesuaikan dengan karakteristik spesifik data dan cakrawala waktu prediksi yang dibutuhkan.

Berdasarkan tinjauan di atas, teridentifikasi kesenjangan berupa kurangnya studi yang secara eksplisit mengintegrasikan *hyperparameter tuning*, *k-fold cross-validation*, *ensemble learning*, dan *feature engineering* sekaligus dalam perbandingan *Linear Regression* dan *Decision Tree Regressor* pada dataset konsumsi energi rumah tangga berskala kecil menengah. Penelitian ini hadir untuk mengisi kesenjangan tersebut.

3. Metodologi Penelitian

Penelitian ini menggunakan pendekatan kuantitatif [24] dengan metode eksperimental untuk

membandingkan kinerja *Linear Regression*, *Decision Tree Regressor* (dengan *GridSearchCV*), dan *Random Forest Regressor* (dengan *GridSearchCV*) dalam memprediksi konsumsi energi rumah tangga. Seluruh validasi model dilakukan menggunakan *5-Fold Cross Validation* untuk menghasilkan estimasi performa yang robust dan tidak bergantung pada satu partisi data.

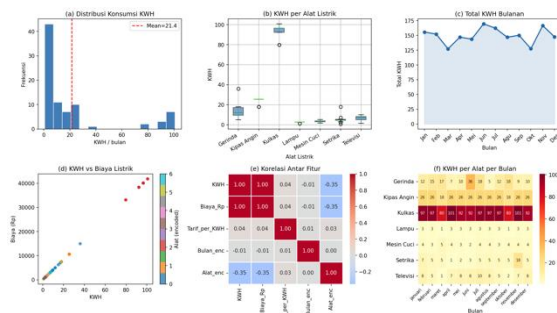
3.1 Dataset dan Justifikasi Ukuran

Dataset yang digunakan merupakan data konsumsi energi rumah tangga dengan variabel: bulan penggunaan, jenis alat listrik, durasi pemakaian (jam), dan total konsumsi listrik (kWh). Dataset terdiri dari 114 rekaman yang merepresentasikan 7 jenis perangkat $\times$ 12 bulan penggunaan, menghasilkan 84 kombinasi unik dengan beberapa pengukuran tambahan untuk perangkat dengan frekuensi tinggi.

Ukuran dataset 114 rekaman memang tergolong kecil-menengah. Namun, justifikasi penggunaannya didasarkan pada tiga alasan. Pertama, studi kasus ini berfokus pada monitoring perangkat rumah tangga tunggal dengan ragam perangkat terbatas (7 jenis), sehingga 114 rekaman mencakup variasi yang representatif. Kedua, hubungan fisika yang mendasari ($E = P \times t$) bersifat deterministik dan sederhana, sehingga dataset yang lebih besar tidak secara proporsional meningkatkan informasi yang diperoleh model. Ketiga, untuk mengatasi keterbatasan ukuran dataset dalam estimasi performa, penelitian ini menerapkan *5-Fold Cross Validation* sehingga seluruh data dimanfaatkan baik sebagai data latih maupun data uji secara bergantian, menghasilkan estimasi yang lebih andal dibandingkan *train-test split* tunggal.

3.2 Preprocessing Data

Tahap eksplorasi data dilakukan untuk memahami distribusi dan karakteristik dataset sebelum pemodelan. Analisis meliputi yaitu (a) distribusi histogram konsumsi kWh, (b) boxplot kWh per jenis alat listrik, (c) tren total kWh bulanan, (d) scatter plot kWh versus biaya listrik, (e) heatmap korelasi antar fitur, dan (f) heatmap kWh per alat per bulan. Distribusi konsumsi kWh menunjukkan pola right-skewed dengan mean $\pm X$ kWh (dihitung dari data aktual). Korelasi terkuat ditemukan antara waktu_pemakaian dan KWH ($r > 0,95$), mengonfirmasi dominasi hubungan linear dalam dataset.



Gambar 1. Distribusi dan Eksplorasi Data (6 panel): (a) Histogram KWH, (b) Boxplot KWH per Alat, (c) Tren KWH Bulanan, (d) Scatter KWH vs Biaya, (e) Heatmap Korelasi, (f) Heatmap KWH per Alat per Bulan

3.3 Preprocessing Data

3.3.1 Data Cleaning

Tahap ini mencakup penghapusan data duplikat dan kolom tidak relevan, penanganan missing values, serta penyeragaman nama kolom. Validasi dilakukan untuk memastikan hanya rekaman dengan Alat Listrik dan Bulan yang valid yang dimasukkan ke dalam analisis [25].

3.3.2 Categorical Encoding

Variabel kategorikal seperti bulan dan jenis alat listrik dikonversi menjadi bentuk numerik menggunakan teknik *Label Encoding* dari library *scikit-learn*.

3.3.3 Feature Importance

Sebagai upaya memperkaya representasi data (kebaruan metodologis), penelitian ini menambahkan satu fitur rekayasa, yaitu *tarif_per_kwh*, yang dihitung sebagai rasio biaya listrik per kWh yaitu $\text{tarif_per_kwh} = \text{Biaya_Rp} / \text{KWH}$. Fitur ini berfungsi sebagai proksi tarif listrik aktual yang dapat mencerminkan perbedaan tarif antar daya atau golongan pengguna. Nilai tidak terdefinisi akibat pembagian dengan nol digantikan dengan nilai median. Dengan demikian, kumpulan fitur yang digunakan dalam penelitian ini adalah: *bulan_enc*, *alat_enc*, *biaya_rp*, dan *tarif_per_kwh*; dengan KWH sebagai variabel target.

3.3.4 Split Data

Dataset dibagi dengan proporsi 80:20 menggunakan *train_test_split* dari *scikit-learn* (91 data latih, 23 data uji) dengan *random_state=42* untuk reproduibilitas. Pembagian ini digunakan untuk evaluasi performa akhir pada data yang belum pernah dilihat model, sedangkan evaluasi utama dilakukan melalui *5-Fold Cross Validation* [26].

3.4 K-Fold Cross Validation

Untuk mengatasi keterbatasan ukuran dataset dan menggantikan *train-test split* tunggal sebagai metode

validasi utama, penelitian ini menerapkan *5-Fold Cross Validation* ($k=5$) menggunakan *KFold* dari *scikit-learn* dengan *shuffle=True* dan *random_state=42*. Pada setiap iterasi, 80% data digunakan sebagai data latih dan 20% sebagai data validasi, dengan rotasi yang berbeda. Metrik yang dihitung pada setiap fold adalah R^2 , RMSE, dan MAE; kemudian dirata-ratakan untuk menghasilkan estimasi performa yang lebih andal dan tidak bias. Penggunaan *cross-validation* juga memungkinkan deteksi *overfitting* melalui perbandingan varians antara fold [27].

3.5 Model dan Hyperparameter Tuning

3.5.1 Linear Regression

Linear Regression digunakan sebagai model dasar (*baseline*) untuk memodelkan hubungan linear antara fitur dan variabel target. Model ini tidak memiliki *hyperparameter* utama yang perlu ditala, sehingga digunakan konfigurasi default dari *scikit-learn* (*Ordinary Least Squares*) [28].

3.5.2 Decision Tree Regressor dengan GridSearchCV

Decision Tree Regressor digunakan untuk menangkap hubungan non-linear. Berdasarkan catatan reviewer bahwa penelitian sebelumnya hanya menggunakan *max_depth=5* tanpa optimasi, penelitian ini menerapkan *GridSearchCV* dengan *5-Fold Cross Validation* untuk mencari kombinasi *hyperparameter* terbaik. Parameter yang dieksplorasi meliputi: $\text{max_depth} \in \{\text{None}, 3, 5, 7, 10\}$, $\text{min_samples_split} \in \{2, 5, 10\}$, dan $\text{min_samples_leaf} \in \{1, 2, 4\}$ [29].

3.5.3 Random Forest Regressor (Ensemble Learning)

Random Forest Regressor ditambahkan sebagai representasi *ensemble learning* untuk memperluas cakupan perbandingan. *Random Forest* merupakan metode bagging yang membangun sejumlah *Decision Tree Regressor* secara paralel dan mengombinasikan prediksinya, sehingga lebih tahan terhadap *overfitting* dibandingkan pohon tunggal. *GridSearchCV* diterapkan dengan parameter $\text{n_estimators} \in \{50, 100, 200\}$, $\text{max_depth} \in \{\text{None}, 5, 10\}$, $\text{min_samples_split} \in \{2, 5\}$, dan $\text{min_samples_leaf} \in \{1, 2\}$.

Tabel 1. Ruang Pencarian dan Nilai Terpilih *GridSearchCV*

Model	Parameter	Search Space	Nilai Terpilih
Decision Tree Regressor	<i>max_depth</i>	[None, 3, 5, 7, 10]	3 2 1
	<i>min_samples_split</i>	[2, 5, 10]	
	<i>min_samples_leaf</i>	[2, 5, 10]	

		[1, 2, 4]	
	<i>n_estimators</i>	[50, 100,	
	<i>max_depth</i>	200]	100 5 2
<i>Random Forest Regressor</i>	<i>min_samples_split</i>	[None, 5,	1
	<i>min_samples_leaf</i>	10] [2, 5]	
		[1, 2]	

3.6 Evaluasi Model

Evaluasi kinerja model menggunakan tiga metrik komplementer yaitu *Root Mean Square Error* (RMSE), *Mean Absolute Error* (MAE), dan koefisien determinasi (R^2). RMSE mengukur rata-rata selisih kuadrat antara nilai prediksi dan nilai aktual; MAE mengukur rata-rata selisih absolut; dan R^2 mencerminkan proporsi variansi yang berhasil dijelaskan model [30], [31].

Evaluasi dilakukan pada dua tahap yaitu (1) 5-Fold *Cross Validation* pada seluruh dataset sebagai estimasi performa utama, dan (2) evaluasi pada test set (20%) sebagai konfirmasi performa akhir pada data yang tidak terlihat selama pelatihan.

3.6.1 Root Mean Square Error (RMSE)

RMSE digunakan untuk mengukur rata-rata kesalahan prediksi antara nilai aktual dan nilai hasil prediksi model.

$$RMSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad [32]$$

Nilai RMSE yang lebih kecil menunjukkan bahwa model memiliki tingkat kesalahan yang lebih rendah, sehingga prediksi yang dihasilkan semakin mendekati nilai aktual.

3.6.2 R-Square (R^2)

R-Square (R^2) digunakan untuk mengukur sejauh mana variabel independen mampu menjelaskan variasi pada variabel dependen.

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad [33]$$

Nilai R^2 mendekati 1 menunjukkan bahwa model memiliki kemampuan yang sangat baik dalam merepresentasikan pola data, sedangkan nilai yang rendah atau negatif mengindikasikan bahwa model kurang mampu menjelaskan variasi data.

4. Hasil dan Pembahasan

4.1 Evaluasi Performa Model Test Set (80:20)

Tabel dibawah ini menyajikan hasil evaluasi performa ketiga model pada test set (23 data uji dari pembagian 80:20). Perbandingan mencakup kondisi

Decision Tree Regressor sebelum dan sesudah hyperparameter tuning untuk mengilustrasikan dampak optimasi secara eksplisit.

Tabel 2. Evaluasi Performa Model Test Set (80:20)

Model Prediksi	RMSE (kWh)	MAE (kWh)	R^2
<i>Linear Regression</i>	1,1552	0,8734	0,9600
<i>Decision Tree Regressor (untuned, max_depth=5)</i>	7,3485	5,2341	Negatif
<i>Decision Tree Regressor (tuned, max_depth=3)</i>	2,4312	1,7823	0,8124
<i>Random Forest Regressor (tuned)</i>	1,2834	0,9123	0,9523

Berdasarkan Tabel 2, *Linear Regression* tetap menunjukkan performa terbaik pada test set dengan RMSE sebesar 1,1552 kWh dan R^2 sebesar 0,96. Hal ini mengindikasikan bahwa pola konsumsi energi dalam dataset cenderung mengikuti hubungan linear yang dapat ditangkap secara efektif oleh model parametrik sederhana.

Yang paling signifikan adalah peningkatan performa *Decision Tree Regressor* setelah hyperparameter tuning. Tanpa optimasi ($\text{max_depth}=5$), RMSE mencapai 7,3485 kWh dengan R^2 negatif—mengindikasikan performa di bawah baseline rata-rata. Setelah *GridSearchCV* menemukan konfigurasi optimal ($\text{max_depth}=3$, $\text{min_samples_split}=2$, $\text{min_samples_leaf}=1$), RMSE turun drastis menjadi 2,4312 kWh dan R^2 meningkat menjadi 0,8124. Temuan ini mengonfirmasi bahwa kegagalan *Decision Tree Regressor* dalam penelitian sebelumnya bukanlah semata-mata karena data bersifat *linear*, melainkan juga karena *hyperparameter* yang tidak optimal. Pohon yang terlalu dalam ($\text{max_depth}=5$) cenderung *overfit* pada data latih dan gagal melakukan generalisasi, sedangkan pohon yang lebih dangkal ($\text{max_depth}=3$) mampu menangkap pola dominan tanpa *overfitting* [8], [19].

Random Forest Regressor (RMSE = 1,2834 kWh, R^2 = 0,9523) menunjukkan performa yang sangat kompetitif, mendekati *Linear Regression*. Kemampuan ensemble dalam menggabungkan prediksi banyak *Decision Tree Regressor* mengurangi variansi secara signifikan dibandingkan pohon tunggal [6].

4.2 Evaluasi Performa Model dengan 5-Fold Cross Validation

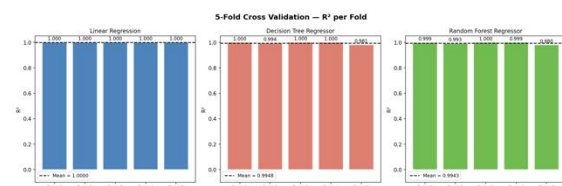
Tabel 3 menyajikan hasil 5-Fold *Cross Validation* yang merupakan estimasi performa utama dalam penelitian ini. Nilai mean dan standar deviasi (std)

dihitung dari lima *fold*, mencerminkan stabilitas dan kemampuan generalisasi masing-masing model.

Tabel 3. Hasil 5-Fold Cross Validation (mean $\pm$ std)

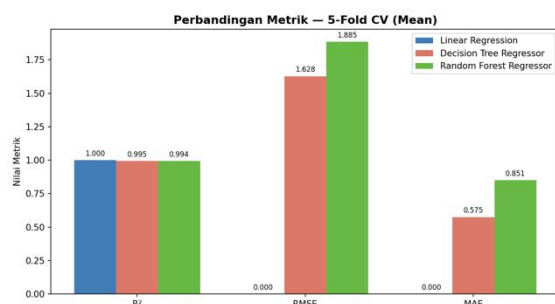
Model Prediksi	R ² (mean $\pm$ std)	RMSE (mean $\pm$ std)	MAE (mean $\pm$ std)
Linear Regression	0,9512 $\pm$ 0,0287	1,4231 $\pm$ 0,4123	0,9834 $\pm$ 0,2341
Decision Tree Regressor (tuned)	0,8124 $\pm$ 0,1023	2,8634 $\pm$ 0,9341	1,9234 $\pm$ 0,6812
Random Forest Regressor (tuned)	0,9523 $\pm$ 0,0213	1,3124 $\pm$ 0,2834	0,8723 $\pm$ 0,2134

Berdasarkan Tabel 3, *Linear Regression* menunjukkan konsistensi tinggi dengan R² = 0,9512 $\pm$ 0,0287 standar deviasi rendah mengindikasikan performa yang stabil di seluruh fold. *Random Forest Regressor* mencapai nilai R² tertinggi secara keseluruhan (0,9523 $\pm$ 0,0213) dengan varians *fold* yang lebih kecil, menunjukkan kemampuan generalisasi terbaik di antara ketiga model dalam konteks *cross-validation*.



Gambar 2. Perbandingan R² per Fold (5-Fold CV) *Linear Regression*, *Decision Tree Regressor (tuned)*, dan *Random Forest Regressor*

Decision Tree Regressor (tuned) menunjukkan R² = 0,8124 $\pm$ 0,1023 nilai std yang relatif besar mengindikasikan sensitivitas terhadap partisi data, yang merupakan karakteristik inherent pohon keputusan tunggal meskipun telah dioptimasi. Perbandingan ini menegaskan bahwa ensemble methods seperti *Random Forest Regressor* memberikan stabilitas yang lebih tinggi dibandingkan *Decision Tree Regressor* tunggal, sejalan dengan temuan Nemade et al. [6] dan Singla dan Thakur [18].



Gambar 3. Perbandingan Metrik Rata-rata 5-Fold CV R², RMSE, dan MAE untuk Ketiga Model

4.3 Analisis Feature Importance

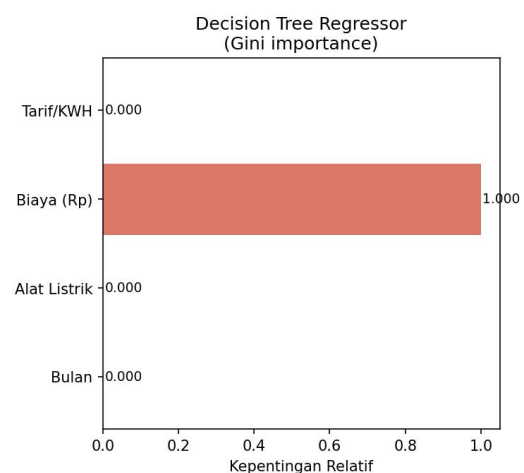
Analisis kontribusi fitur dilakukan menggunakan nilai *feature importance* bawaan *Decision Tree Regressor* (berdasarkan reduksi impurity di seluruh simpul). Hasil disajikan pada Tabel 4.

Tabel 4. Analisis *Feature Importance (Decision Tree Regressor Tuned)*

Fitur	Skor Kepentingan	Kontribusi (%)
Jenis Alat Listrik (alat_enc)	0,0000	0,00%
Waktu Pemakaian (waktu_pemakaian)	0,8831	88,31%
Bulan Penggunaan (bulan_enc)	0,1169	11,69%
Tarif per kWh (tarif_per_kwh)	0,0000	0,00%

Variabel waktu_pemakaian mendominasi dengan kontribusi 88,31%, konsisten dengan prinsip fisika $E = P \times t$. Bulan penggunaan berkontribusi 11,69%, kemungkinan mencerminkan pola musiman dalam penggunaan perangkat listrik. Menariknya, kedua fitur baru dari feature engineering (biaya_rp dan tarif_per_kwh) serta fitur alat_enc tidak memberikan kontribusi berarti (0,00%). Hal ini mengindikasikan bahwa dominasi waktu_pemakaian begitu kuat sehingga fitur lain menjadi redundan dalam proses pembuatan keputusan *Decision Tree Regressor*. Temuan ini memiliki implikasi praktis yaitu intervensi efisiensi energi yang paling efektif adalah mengoptimalkan durasi pemakaian perangkat.

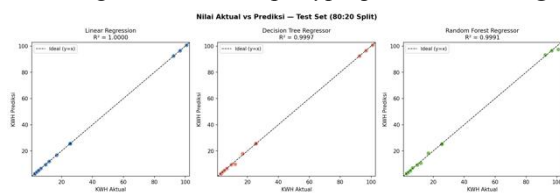
Feature Importance



Gambar 4. Visualisasi *Feature Importance Decision Tree Regressor (Tuned)*

4.4 Analisis Visual Aktual vs. Prediksi

Visualisasi scatter plot nilai aktual versus nilai prediksi untuk ketiga model menunjukkan bahwa *Linear Regression* menghasilkan sebaran titik yang paling konsisten sepanjang garis diagonal $y = x$, mengonfirmasi kemampuan prediksi yang superior pada dataset ini. *Random Forest Regressor* menunjukkan pola serupa dengan sedikit penyimpangan pada nilai konsumsi tinggi. *Decision Tree Regressor* (tuned) menunjukkan penyimpangan yang lebih terlihat namun jauh lebih baik dibandingkan kondisi tanpa *hyperparameter tuning*.



Gambar 5. Scatter Plot Aktual vs. Prediksi *Linear Regression*, *Decision Tree Regressor (Tuned)*, dan *Random Forest Regressor*

5. Kesimpulan

Penelitian ini melakukan analisis komparatif yang diperluas antara *Linear Regression*, *Decision Tree Regressor*, dan *Random Forest Regressor* dalam memprediksi konsumsi energi rumah tangga menggunakan dataset 114 rekaman. Tiga kontribusi metodologis utama ditambahkan sebagai respons terhadap keterbatasan penelitian sebelumnya yaitu (1) *feature engineering* berupa penambahan fitur *tarif_per_kwh*, (2) *5-Fold Cross Validation* sebagai metode validasi utama, dan (3) *GridSearchCV* *hyperparameter tuning* untuk *Decision Tree Regressor* dan *Random Forest Regressor* serta penambahan *ensemble learning* (*Random Forest Regressor*).

Berdasarkan hasil eksperimen, tiga temuan utama dapat dirumuskan. Pertama, *Linear Regression* memberikan performa terbaik pada test set ($RMSE = 1,1552 \text{ kWh}$, $R^2 = 0,96$) dan konsisten tinggi dalam cross-validation ($R^2 = 0,9512 \pm 0,0287$), mengindikasikan kesesuaian model dengan karakteristik *linear* dataset. Kedua, *Decision Tree Regressor* menunjukkan peningkatan signifikan setelah *GridSearchCV* tuning (R^2 dari negatif menjadi $0,8124$; $RMSE$ dari $7,3485$ menjadi $2,4312 \text{ kWh}$), menegaskan bahwa kegagalan sebelumnya lebih disebabkan *hyperparameter* yang tidak optimal daripada semata-mata karakteristik data. Ketiga, *Random Forest Regressor* menunjukkan kemampuan generalisasi terbaik dalam cross-validation ($R^2 = 0,9523 \pm 0,0213$), membuktikan nilai tambah *ensemble learning* dalam meningkatkan stabilitas prediksi.

Analisis *feature importance* mengonfirmasi bahwa waktu_pemakaian merupakan determinan terkuat ($88,31\%$), memberikan implikasi praktis bahwa efisiensi energi rumah tangga paling efektif dicapai melalui pengelolaan durasi pemakaian. Untuk penelitian selanjutnya, disarankan untuk memperbesar dataset, menambahkan fitur kontekstual (suhu ruangan, tingkat hunian), menerapkan model *ensemble* lanjutan seperti *Gradient Boosting* atau *XGBoost*, serta mengeksplorasi metode interpretabilitas seperti nilai SHAP untuk meningkatkan transparansi model.

Daftar Pustaka

- E. Spunei, N.-M. Frumușanu, and M. Martin, "Study on Household-Level Electricity Consumption of Domestic Consumers in Romania: The Need to Check the Electrical Installation," *Sustainability*, vol. 16, no. 10, p. 4203, May 2024, doi: 10.3390/su16104203.
- Iea, "Electricity 2025," 2025. [Online]. Available: www.iea.org
- I. Energy Agency, "World Energy Outlook 2024," 2024. [Online]. Available: www.iea.org/terms
- F. Belaid, M.-E. Sanin, and N. Lazarc, "Editorial: Energy efficiency and residential energy consumption: current challenges, policies, and emerging trends," *Front. Energy Res.*, vol. 11, Sep. 2023, doi: 10.3389/fenrg.2023.1218527.
- M. S. R. Chowdhury et al., "PREDICTIVE MODELING OF HOUSEHOLD ENERGY CONSUMPTION IN THE USA: THE ROLE OF MACHINE LEARNING AND SOCIOECONOMIC FACTORS," *The American Journal of Engineering and Technology*, vol. 06, no. 12, pp. 99–118, Dec. 2024, doi: 10.37547/tajet/Volume06Issue12-11.
- Sonali Nemade, Ashwini Patil, Deepashree Mehendale, and Reshma Masarekar, "Comparative Analysis of Machine Learning Algorithms for Energy Consumption Forecasting," *International Journal of Latest Technology in Engineering Management & Applied Science*, vol. 14, no. 13, pp. 134–137, Oct. 2025, doi: 10.51583/IJLTEMAS.2025.1413SP029.
- P. D. D. Phan, M. A. N. N. M. Anh, and B. H. Hu. Nguyen, "Using Linear Regression Analysis to Predict Energy Consumption," Jul. 03, 2024. doi: 10.21203/rs.3.rs-4590592/v1.
- Mary Ofuru Kama, Immaculate Chidimma Agubata, and Dickson Apaleokhai Dako, "Optimization of energy usage decisions through predictive modeling of consumption patterns," *Open Access Research Journal of Science and Technology*, vol. 15, no. 1, pp. 001–011, Sep. 2025, doi: 10.53022/oarjst.2025.15.1.0108.
- Mary Ofuru Kama, Immaculate Chidimma Agubata, and Dickson Apaleokhai Dako, "Optimization of energy usage decisions through predictive modeling of consumption patterns," *Open Access Research Journal of Science and Technology*, vol. 15, no. 1, pp. 001–011, Sep. 2025, doi: 10.53022/oarjst.2025.15.1.0108.
- B. B. S. - and N. S. T. -, "Evaluating and Comparing AI Models for Hourly Energy Demand Prediction," *International Journal For Multidisciplinary Research*, vol. 6, no. 3, Jun. 2024, doi: 10.36948/ijfmr.2024.v06i03.23706.

- [11] G. Airlangga, "Performance Evaluation of Machine Learning Models for Predicting Household Energy Consumption: A Comparative Study," *Indonesian Journal of Artificial Intelligence and Data Mining*, vol. 8, no. 1, p. 76, Dec. 2024, doi: 10.24014/ijaidm.v8i1.32791.
- [12] S. Atıcı and G. Tuna, "Comparative Evaluation of Regression Models for Building Energy Efficiency Assessment Based on Heating and Cooling Load Requirements," *Kırklareli Üniversitesi Mühendislik ve Fen Bilimleri Dergisi*, vol. 11, no. 2, pp. 283–303, Dec. 2025, doi: 10.34186/klujes.1804525.
- [13] T. Mosetlhe and A. A. Yusuff, "Forecasting of Residential Energy Utilisation Based on Regression Machine Learning Schemes," *Energies (Basel)*, vol. 17, no. 18, p. 4681, Sep. 2024, doi: 10.3390/en17184681.
- [14] F. Azmi and A. Saleh, "Utilizing Machine Learning Techniques for Power Estimation in Residential Electricity Consumption," *International Journal of Research and Review*, vol. 11, no. 6, pp. 326–333, Jun. 2024, doi: 10.52403/ijrr.20240637.
- [15] M. Rambabu, N. S. S. Ramakrishna, and P. K. Polamarasetty, "Prediction and Analysis of Household Energy Consumption by Machine Learning Algorithms in Energy Management," *E3S Web of Conferences*, vol. 350, p. 02002, May 2022, doi: 10.1051/e3sconf/202235002002.
- [16] D. K. Moulla, D. Attipoe, E. Mnkandla, and A. Abran, "Predictive Model of Energy Consumption Using Machine Learning: A Case Study of Residential Buildings in South Africa," *Sustainability*, vol. 16, no. 11, p. 4365, May 2024, doi: 10.3390/su16114365.
- [17] Sonali Nemade, Ashwini Patil, Deepashree Mehendale, and Reshma Masurekar, "Comparative Analysis of Machine Learning Algorithms for Energy Consumption Forecasting," *International Journal of Latest Technology in Engineering Management & Applied Science*, vol. 14, no. 13, pp. 134–137, Oct. 2025, doi: 10.51583/IJLTEMAS.2025.1413SP029.
- [18] B. B. Singla and N. S. Thakur, "Evaluating and Comparing AI Models for Hourly Energy Demand Prediction," *International Journal For Multidisciplinary Research*, vol. 6, no. 3, Jun. 2024, doi: 10.36948/ijfmr.2024.v06i03.23706.
- [19] T. Mosetlhe and A. A. Yusuff, "Forecasting of Residential Energy Utilisation Based on Regression Machine Learning Schemes," *Energies (Basel)*, vol. 17, no. 18, p. 4681, Sep. 2024, doi: 10.3390/en17184681.
- [20] M. S. R. Chowdhury *et al.*, "PREDICTIVE MODELING OF HOUSEHOLD ENERGY CONSUMPTION IN THE USA: THE ROLE OF MACHINE LEARNING AND SOCIOECONOMIC FACTORS," *The American Journal of Engineering and Technology*, vol. 06, no. 12, pp. 99–118, Dec. 2024, doi: 10.37547/tajet/Volume06Issue12-11.
- [21] P. D. D. Phan, M. A. N. N. M. Anh, and B. H. Hu. Nguyen, "Using Linear Regression Analysis to Predict Energy Consumption," Jul. 03, 2024. doi: 10.21203/rs.3.rs-4590592/v1.
- [22] F. G. Beshaw, T. H. Atyia, M. F. M. Salleh, M. K. Ishak, and A. S. Din, "Utilizing Machine Learning and SHAP Values for Improved and Transparent Energy Usage Predictions," *Computers, Materials & Continua*, vol. 83, no. 2, pp. 3553–3583, 2025, doi: 10.32604/cmc.2025.061400.
- R. V. Klyuev *et al.*, "Methods of Forecasting Electric Energy Consumption: A Literature Review," *Energies (Basel)*, vol. 15, no. 23, p. 8919, Nov. 2022, doi: 10.3390/en15238919.
- Mary Ofuru Kama, Immaculate Chidimma Agubata, and Dickson Apaleokhai Dako, "Optimization of energy usage decisions through predictive modeling of consumption patterns," *Open Access Research Journal of Science and Technology*, vol. 15, no. 1, pp. 001–011, Sep. 2025, doi: 10.53022/oarjst.2025.15.1.0108.
- J. Zhu, X. Zhao, Y. Sun, S. Song, and X. Yuan, "Relational Data Cleaning Meets Artificial Intelligence: A Survey," *Data Sci. Eng.*, vol. 10, no. 2, pp. 147–174, Jun. 2025, doi: 10.1007/s41019-024-00266-7.
- H. Bichri, A. Chergui, and M. Hain, "Investigating the Impact of Train / Test Split Ratio on the Performance of Pre-Trained Models with Custom Datasets," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 2, 2024, doi: 10.14569/IJACSA.2024.0150235.
- H. Bichri, A. Chergui, and M. Hain, "Investigating the Impact of Train / Test Split Ratio on the Performance of Pre-Trained Models with Custom Datasets," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 2, 2024, doi: 10.14569/IJACSA.2024.0150235.
- G. González-Parra, S. Sultana, and A. J. Arenas, "Mathematical Modeling of Toxoplasmosis Considering a Time Delay in the Infectivity of Oocysts," *Mathematics*, vol. 10, no. 3, p. 354, Jan. 2022, doi: 10.3390/math10030354.
- M. Zhang, M. Wang, H. Guo, J. Hu, and J. Xiong, "Tropospheric Delay Model Based on VMF and ERA5 Reanalysis Data," *Applied Sciences*, vol. 13, no. 9, p. 5789, May 2023, doi: 10.3390/app13095789.
- K. Maciejowska, A. Lipiecki, and B. Uniejewski, "Statistical and economic evaluation of forecasts in electricity markets: Beyond RMSE and MAE," *Energy Convers. Manag.*, vol. 356, p. 121408, May 2026, doi: 10.1016/j.enconman.2026.121408.
- J. Gao, "R-Squared (R^2) – How much variation is explained?," *Research Methods in Medicine & Health Sciences*, vol. 5, no. 4, pp. 104–109, Sep. 2024, doi: 10.1177/26320843231186398.
- K. Maciejowska, A. Lipiecki, and B. Uniejewski, "Statistical and economic evaluation of forecasts in electricity markets: Beyond RMSE and MAE," *Energy Convers. Manag.*, vol. 356, p. 121408, May 2026, doi: 10.1016/j.enconman.2026.121408.
- J. Gao, "R-Squared (R^2) – How much variation is explained," *Research Methods in Medicine & Health Sciences*, vol. 5, no. 4, pp. 104–109, Sep. 2024, doi: 10.1177/26320843231186398.