Journal of Information System Exploration and Research

https://shmpublisher.com/index.php/joiser

p-ISSN 2964-1160 | e-ISSN 2963-6361

Optimization of Energy Consumption Prediction with Random Forest Regressor and XGBoost Feature Importance

Risma Moulidya Syafei[1*], Tiara Lailatul Nikmah[2], Devi Nurul Anisa[3], Sidiq Noor Kharisma[4]

[1,2,3,4] *Department of Computer Science, Universitas Negeri Semarang, Indonesia*

DOI: https://doi.org/10.52465/joiser.v3i2.593

Received 28 January 2026; Accepted 08 February 2026; Available online 10 February 2026

Article Info

Keywords:

Energy consumption;
Southern California dataset;
Random forest regressor;
XGBoost feature importance

Abstract

Energy consumption is increasing as industry and technology advance. However, it will have a bad impact if its use is not properly controlled. Therefore, predicting energy consumption is needed to prevent energy waste and to streamline its use across several influencing factors. Predictions are made using the Random Forest Regressor method. Where regression and Random Forest techniques can produce accurate results for continuous values such as total energy consumption. The feature importance method is also used to select the most influential features. Where of the 40 features in the energy consumption dataset in Southern California, only 24 features were selected based on the average threshold of the gain value. The results showed that the use of XGBoost feature importance lowered the Mean Absolute Error (MAE) value of the Random Forest Regressor, which was 16.56 to 16.55. This value is the difference between the actual data and the predicted data. This proves that the model successfully predicts with a small error value. The application of feature importance in energy consumption prediction using Random Forest Regressor is expected to be more efficient in energy consumption, especially in the sectors that most affect the increase in energy consumption.

1. Introduction

Energy consumption is becoming higher along with the increasing demand for energy due to industrialization and rapid technological development [1], [2], [3], [4]. Especially in developed countries such as the United States and China which are always in the top 5 as countries with the largest energy consumption in the world [5], [6], [7]. Energy consumption continues to increase due to population growth, infrastructure development and rapidly increasing use of electronics [8], [9]. This makes it an important condition to optimize energy consumption patterns so that energy consumption can be controlled properly, efficiently and sustainably.

* ***Corresponding Author:***

Risma Moulidya Syafei,
Department of Computer Science,
Universitas Negeri Semarang,
Indonesia.
Email: syafeirisma@students.unnes.ac.id

In this study, a dataset of energy consumption in the Southern California region from January 2018 to January 2024 was used. Total energy consumption in California in 2023 is higher than most U.S. states, being just below Texas and Florida (492, 250, and 239 TWh), respectively) [10], [11], [12], [13].

The influence of energy consumption can also be caused by climate change [14], [15]. Climate change that causes extreme weather in every season, a relationship that has been increasingy studied using artificial intelligence approaches by Camps-Valls et al, 2025 [16], [17], [18]. Especially in the region of California which has a Mediterranean climate that has long and dry summers and short, humid winters [19].

The results of this prediction can determine the pattern of energy increase that can help design more efficient energy management policies [20]. Previous research has been cucted by Barua et al, 2025 [21] using Southern California regional energy consumption data. The study compared various algorithms such as Linear Regression, Random Forest Regressor and XGBoost. The results show that the Random Forest Regressor is the best model with the lowest error rate.

This research will combine the Random Forest Regressor algorithm and the feature importance of XGBoost. This feature importance is used to improve the accuracy of predictions, where only selected features are used, it can improve model performance by removing irrelevant features. The model will be more efficient in working by reducing computational complexity.

2. Method

This research uses 2 stages of data processing, namely data encoding and feature importance. The stage of this study can be seen in Figure 1.

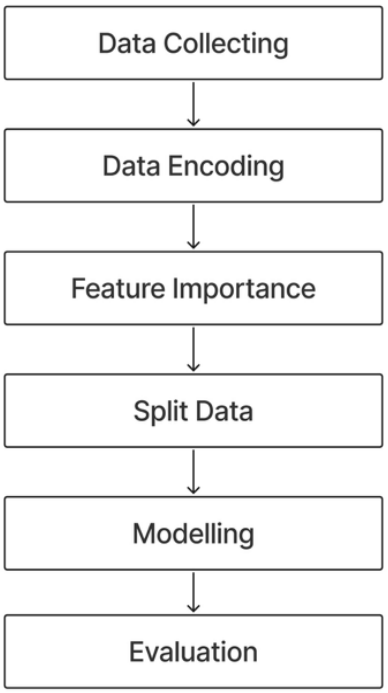

Figure 1. Stage of study

2.1. Data Collecting

The dataset was obtained from Kaggle with a total of 52 thousand data collected from January 2018 to January 2024. This data was collected from 100 facilities in Southern California such as smart meters, IoT sensors, building management systems, and regional utility companies. In addition, this data includes various natural factors such as season, temperature and humidity. There are 40 features, of which 34 are numerical features and 6 categorical features. The list of features can be seen in Table 1.

Table 1. Dataset Features List

No	Feature Name	Data Type	Short Description (Indonesian)
1	Timestamp	Object	Data logging time (per hour, from 2018–2024).
2	Building Type	Object	Building type (Residential, Commercial, Industrial).
3	Energy Consumption (kWh)	Float64	Total electrical energy consumption in kilowatt-hours.
4	Temperature (°C)	Float64	The outdoor temperature at energy consumption is recorded.
5	Humidity (%)	Float64	Relative humidity (%) at the time of energy recording.
6	Occupancy Rate (%)	Float64	The percentage of occupancy of the space in the building.
7	Lighting Consumption (kWh)	Float64	Electricity consumption for lighting.
8	HVAC Consumption (kWh)	Float64	Energy consumption for heating, ventilation, and air conditioning systems.
9	Energy Price ($/kWh)	Float64	The price of electricity per kilowatt-hour at that time.
10	Carbon Emission Rate (g CO2/kWh)	Float64	Carbon emissions per unit of energy used.
11	Power Factor	Float64	Real power to pseudo-power ratio (electrical efficiency).
12	Voltage Levels (V)	Float64	The level of electrical voltage at the time of recording.
13	Reactive Power (kVARh)	Float64	Reactive power consumed (in reactive kilovolt-amperes hours).
14	Power Outage Indicator	Int64	Power outage indicator (1 = out, 0 = normal).
15	Indoor Temperature (°C)	Float64	The temperature inside the building.
16	Building Age (years)	Float64	The age of the building in years.
17	Equipment Age (years)	Float64	The age of the main equipment in the building.
18	Energy Efficiency Rating	Float64	Building energy efficiency score.
19	Building Size (m^2)	Float64	The area of the building in square meters.
20	Window-to-Wall Ratio (%)	Float64	The ratio of the window area to the wall.
21	Insulation Quality Score	Float64	A building's thermal insulation quality score.
22	Historical Energy Consumption (kWh)	Float64	Energy consumption in the past.
23	Maintenance Status	Int64	Building maintenance status (0 = good, 1 = maintenance needed).
24	Demand Response Participation	Int64	Participation in the load reduction program (0 = no, 1 = yes).
25	Occupancy Schedule	Object	Schedule of activities/occupancy in the building.
26	Local Energy Production (kWh)	Float64	Locally generated energy (e.g. solar panels).
27	Grid Stability Score	Float64	Power grid stability score.

28	Solar Irradiance (W/m²)	Float64	The intensity of solar radiation at the site of the building.
29	Smart Plug Usage (kWh)	Float64	Energy consumption through smart plugs.
30	Water Usage (liters)	Float64	Water consumption in liters.
31	Energy Savings Target (%)	Float64	Energy savings target (%).
32	Room-Level Energy Consumption (kWh)	Float64	Energy consumption per room.
33	Zonal Heating/Cooling Data (kWh)	Float64	Energy usage data per zone (heating/cooling).
34	Electric Vehicle Charging Status	Int64	Electric vehicle charging status (0 = no, 1 = active).
35	Building Orientation	Object	The direction facing the building (e.g. North, South).
36	IoT Sensor Count	Float64	The number of active IoT sensors in the building.
37	Thermal Comfort Index	Float64	Thermal comfort index based on temperature & humidity.
38	Energy Savings Potential (%)	Float64	Potential energy savings based on usage patterns.
39	Peak Demand Reduction Indicator	Int64	Peak demand reduction indicator (1 = occurs, 0 = no).
40	Carbon Emission Reduction Category	Object	Carbon emission reduction categories (Low, Medium, High, Very High).

2.2. Data Encoding

Data encoding is done on a categorical feature because it will be difficult to process in an algorithmic model. Encoding was carried out on the Building Type, Carbon Emission Reduction Category, Building Orientation and Occupancy Shehule features. The method used is the label encoder, which works by converting each unique category of data into numbers in alphabetical order [22].

2.3. Feature Importance

Feature Importance uses the XGBoost algorithm which is famous for its performance. The advantage of this algorithm is its high performance and is very effective for datasets with many features without the need for many feature engineering [23]. This method uses the calculation of the gain value to determine the order of feature importance. The calculation of the gain value is obtained from how much the error value decreases in each split or fork of the XGBoost algorithm. The gain value shows how much a feature contributes to improving capital performance.

2.4. Split Data

Data sharing uses an 80:20 ratio for data train and data test. With 80% training data, the model has enough information to learn complex patterns. About 20% of the test data is used for testing so that the model is more representative of the new data. This comparison is based on the majority of dataset segments in previous studies [24]. This ratio is considered appropriate because it has a good balance between model accuracy and generalization capabilities.

2.5. Modelling

Modelling was done using the Random Forest regression algorithm. The use of regression algorithms is used because the target value is in the form of a continuous value, namely total energy consumption in kWH. In addition, the Random Forest algorithm was chosen because it has good performance in handling non-linear data. Random Forest works by combining predictions from multiple decision trees to get a more accurate final result. Each tree is constructed from a random subset of features so that it is able to reduce the variance and overfitting that often occur in the decision tree model [25].

The regression technique was chosen because it is able to provide direct quantitative predictions and allows analysis of the relationship patterns between input and output features that are non-

discrete. By using Random Forest-based regression techniques, the model can capture energy consumption dynamics more comprehensively and provide more stable and reliable prediction results.

2.6. Evaluation

The evaluation uses the Mean Absolute Error (MAE) metric that is suitable for regression models. This metric calculates the average absolute value of the predicted value and the actual value. If the value is smaller, the better the model will be.

3. Results and Discussion

Predictions are made using Random Forest Regressor to predict total energy consumption in the Southern California region. The dataset totals 52k data consisting of 34 categorical features and 6 numerical features. The data has been checked and there are no null or duplicate values. So that the data is declared clean and can go directly to the data encoding stage.

The next stage is data encoding, which is necessary because most features are categorical. Encoding is done to convert categorical data into numerical representations that can be processed by the Random Forest algorithm. After that, the processing of the importance feature was carried out using XGboost and 24 features were obtained that had the most influence on the model's performance.

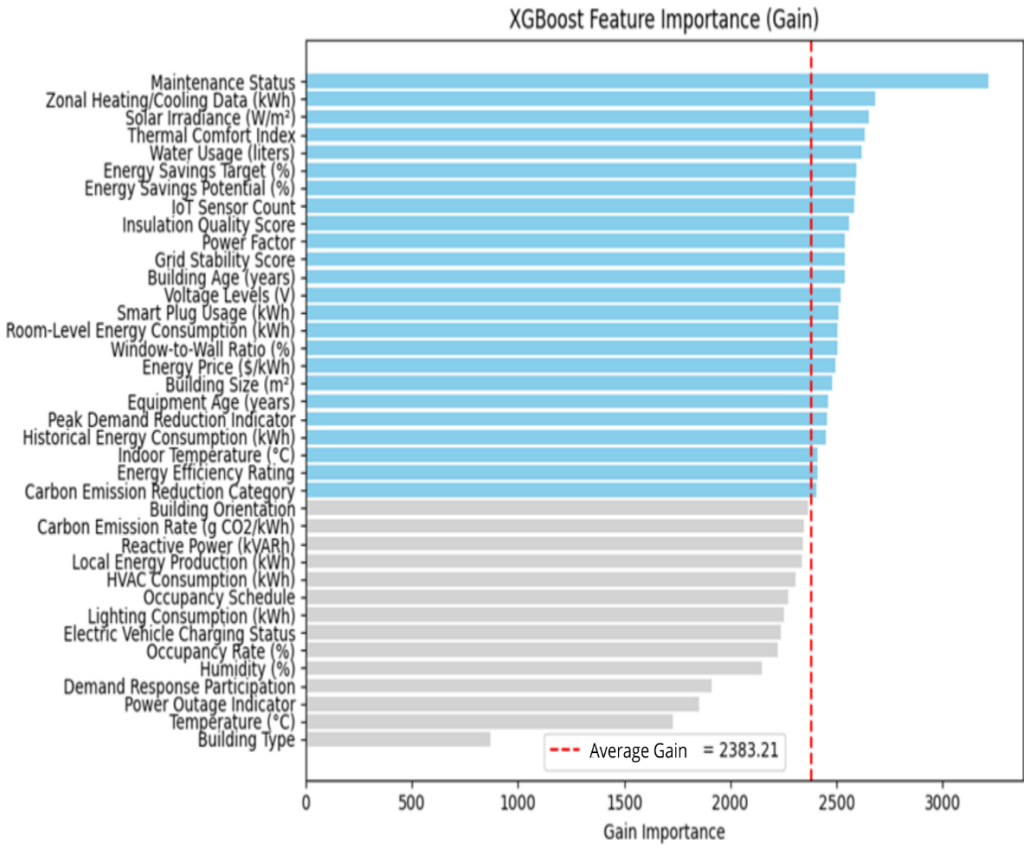

Figure 2. Feature importance results

After feature selection, the dataset is then divided into two parts, with an 80:20 ratio. As much as 80% of the data is used for training to build and train models, while the remaining 20% is used for testing in order to evaluate the model's performance against data that has never been seen before.

The regression model is built using the Random Forest Regressor algorithm. When compared to the previous implementation of feature importance, the MAE value has clearly decreased. The results of the comparison can be seen in Table 2.

Table 2. Comparison of regression models based on MAE values

Models	MAE
Random Forest Regressor	16.56
Random Forest Regressor + XGBoost Feature Importance	16.55

It can be seen in Table 2 that there was a decrease of 0.01 in the MAE value obtained by the Random Forest Regressor and XGBoost Feature Importance models. This proves that the use of feature importance can affect model performance by lowering the error value.

In addition, the model was also evaluated with three other metrics, namely Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). The results of the evaluation of the testing data can be seen in Table 3.

Table 3. Evaluation Results

Evaluation Metrics	Value
MAE	16.55
MSE	431.45
RMSE	20.77

The MAE value of 16.55 shows that on average, the model's prediction of a difference of ±16.55 kWh from the actual energy consumption value. This value is considered quite low due to the complexity of the data and the number of features analyzed. MAE is not very sensitive to outliers, thus providing a consistent picture of model performance.

Meanwhile, MSE measures the square average of the difference between the predicted value and the actual value. The MSE value is higher than MAE because the weighting process is large error squared, so it is very sensitive to outliers. The value of 431.45 kWh^2 is still within reasonable limits, due to the large amount of data used by the dataset with 52,000 rows.

The results of the evaluation show that the Random Forest Regressor model has a good performance in predicting the total energy consumption of the Southern California region. The three evaluation metrics provide an indication that the difference between prediction and reality is quite small and practically acceptable. The conformity of the MAE and RMSE values also indicates that the model is not only accurate overall, but also not significantly affected by extreme values or noise in the data.

This model has proven to be effective in processing data with many features and high complexity, including a combination of categorical and numerical features. These results support the use of Random Forest in energy prediction scenarios, especially on large, heterogeneous real-world data.

4. Conclusion

The Random Forest Regressor model applied to predict total energy consumption in the Southern California region showed good performance with an MAE of 16.55, an MSE of 431.45 and an RMSE of 20.77, indicating an accurate and consistent prediction. Using a dataset with 52,000 data and 40 features, the model was able to effectively see the complex patterns of the relationship between input variables and energy consumption. In addition, feature importance shows the main factors as the main determinants of energy consumption that can help in future energy efficiency planning. Suggestions for future research are that the model be tested on datasets from different regions or time periods to test the generalization capabilities of the model. In addition, additional features such as real-time power equipment usage data and integration with IoT technology can further improve prediction accuracy.

References

[1] S. A. Reza *et al.*, "Predicting energy consumption patterns with advanced machine learning techniques for sustainable urban development," *Journal of Computer Science and Technology Studies*, vol. 7, no. 1, pp. 265–282, 2025.

[2] R. Chandrasekaran and S. K. Paramasivan, “Advances in deep learning techniques for short-term energy load forecasting applications: A review,” *Archives of Computational Methods in Engineering*, vol. 32, no. 2, pp. 663–692, 2025.

[3] J. H. Hefni, L. R. Ozzora, F. V. Ferdinand, K. Julita, and A. Ng, “ANALISIS KORELASI ANTARA KONSUMSI ENERGI DAN URBANISASI TERHADAP PERTUMBUHAN MANUFAKTUR,” *JMBI UNSRAT (Jurnal Ilmiah Manajemen Bisnis dan Inovasi Universitas Sam Ratulangi).*, vol. 12, no. 1, pp. 378–390, Apr. 2025, doi: 10.35794/jmbi.v12i1.61312.

[4] J. Yang, Y. Yu, T. Ma, C. Zhang, and Q. Wang, “Evolution of energy and metal demand driven by industrial revolutions and its trend analysis,” *Chinese Journal of Population, Resources and Environment*, vol. 19, no. 3, pp. 256–264, Sep. 2021, doi: 10.1016/j.cjpre.2021.12.028.

[5] F. Tarumanegara, H. Sidik, and F. J. Sanjaya, “Performa Produksi Dan Konsumsi Sumber Energi Negara-Negara Dunia Pada Tahun 2001-2022,” *Jurnal Terekam Jejak*, vol. 2, no. 1, pp. 1–27, 2024.

[6] International Energy Agency, “Global Energy Review 2025,” 2025. [Online]. Available: www.iea.org

[7] McKinsey & Company, “Global Energy Perspective 2025,” 2025. [Online]. Available: https://www.mckinsey.com/

[8] R. N. Sari and Y. P. Sari, “Pengaruh Urbanisasi Terhadap Konsumsi Energi di Indonesia,” *Media Riset Ekonomi Pembangunan (MedREP)*, vol. 2, no. 1, 2025.

[9] S. A. Reza *et al.*, “Predicting energy consumption patterns with advanced machine learning techniques for sustainable urban development,” *Journal of Computer Science and Technology Studies*, vol. 7, no. 1, pp. 265–282, 2025.

[10] SEDS, “U.S. Energy Information Administration, State energy data system (seds),” https://www.eia.gov/state/seds/sep_fuel/html/pdf/fuel_use_es.

[11] A. Barua *et al.*, “Optimizing energy consumption patterns in southern california: An ai-driven approach to sustainable resource management,” *Journal of Ecohumanism*, vol. 4, no. 1, pp. 2920–2935, 2025.

[12] S. Hossain *et al.*, “Forecasting Energy Consumption Trends with Machine Learning Models for Improved Accuracy and Resource Management in the USA,” *Journal of Business and Management Studies*, vol. 7, no. 1, pp. 200–217, 2025.

[13] G. Tao, Q. M. Jiang, and C. Huang, “Guidelines for enhancing the energy performance regarding accessory dwelling units in Southern California,” *Journal of Building Engineering*, vol. 99, p. 111621, 2025.

[14] R. Qu, R. Kou, and T. Zhang, “The Impact of Weather Variability on Renewable Energy Consumption: Insights from Explainable Machine Learning Models,” *Sustainability*, vol. 17, no. 1, p. 87, Dec. 2024, doi: 10.3390/su17010087.

[15] A. C. R. Gonçalves, X. Costoya, R. Nieto, and M. L. R. Liberato, “Extreme weather events on energy systems: a comprehensive review on impacts, mitigation, and adaptation measures,” *Sustainable Energy Research*, vol. 11, no. 1, p. 4, Jan. 2024, doi: 10.1186/s40807-023-00097-6.

[16] D. N. Putri, A. P. Ramadhani, C. P. Manurung, and B. Purba, “Konsep Konsumsi Energi Di Indonesia Serta Menganalisis Keterkaitannya Dengan Emisi Gas Rumah Kaca,” *Jurnal Ilmiah Wahana Pendidikan*, vol. 10, no. 14, pp. 338–345, 2024.

[17] R. C. Tarumingkeng, “Pengaruh Perubahan Iklim,” 2024.

[18] G. Camps-Valls *et al.*, “Artificial intelligence for modeling and understanding extreme weather and climate events,” *Nat. Commun.*, vol. 16, no. 1, p. 1919, Feb. 2025, doi: 10.1038/s41467-025-56573-8.

[19] S. Stefanidis, K. Ioannou, N. Proutsos, I. Karmiris, and P. Stefanidis, “Comparative Analysis of Machine Learning Algorithms for Potential Evapotranspiration Estimation Using Limited Data at a High-Altitude Mediterranean Forest,” *Atmosphere (Basel).*, vol. 16, no. 7, p. 851, Jul. 2025, doi: 10.3390/atmos16070851.

[20] A. Prasyas and S. M. Ibrahim, “Integrasi Teknologi AI dalam Perancangan Smart Building: Studi Implementasi dan Efisiensi Energi,” *Jurnal Rekayasa Sipil dan Arsitektur*, vol. 1, no. 1, pp. 61–72, 2025.

[21] A. Barua *et al.*, “Optimizing energy consumption patterns in southern california: An ai-driven approach to sustainable resource management,” *Journal of Ecohumanism*, vol. 4, no. 1, pp. 2920–2935, 2025.

[22] N. Ulhasanah, Y. H. Chrisnanto, and J. E. Chrisnanto, “Klasifikasi Multi-Label Jenis dan Warna Buah Menggunakan Convolutional Neural Network (CNN) dengan Encoder Fitur,” *TEMATIK*, vol. 12, no. 1, pp. 72–80, 2025.

[23] S. Banitaan and H. Alquran, “Improving Suicide Ideation Detection Through Feature Engineering and Machine Learning,” *IEEE Access*, 2025.

[24] 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 & Applications*, vol. 15, no. 2, 2024.

[25] E. Halabaku and E. Bytyçi, “Overfitting in Machine Learning: A Comparative Analysis of Decision Trees and Random Forests.,” *Intelligent Automation & Soft Computing*, vol. 39, no. 6, 2024.