Social Science and Humanities Journal
Everant Journal
ISSN: 2456-2653
Volume 09 Issue 01 January 2025
https://sshjournal.com/
Impact Factor: 2024: 6.576
2023: 5.731
DOI: https://doi.org/10.18535/sshj.v9i01.1576

Tree-Based Algorithms performance in Predicting Household Energy Consumption

Ferdinand Nathanael[1,a*] | Jimmy Tjen[2,b] | Genrawan Hoendarto[3,c]

[1,2,3]Widya Dharma Pontianak University, Pontianak, West Kalimantan, Indonesia
[a]21421435_ferdinand_n@widyadharma.ac.id, [b]jimmy.tjen@mathmods.eu, [c]genrawan@widyadharma.ac.id

Received 03-12-2024
Revised 04-12-2024
Accepted 30-12-2024
Published 02-01-2025

Abstract:

Predicting household energy consumption is becoming increasingly important as we strive to manage energy costs and support environmental sustainability. This study takes a close look at how the random forest machine learning method can be used to forecast household energy usage. We used a detailed dataset from the UCI Machine Learning Repository, covering 47 months of minute-by-minute energy consumption data. By comparing Random Forest with other popular machine learning techniques like Gradient Boosting, Regression Tree, Support Vector Machine, and Naïve Bayes, we found that Random Forest stood out for its predictive accuracy, achieving 77.05%. While it does take longer to train, the benefits of accuracy make it a strong candidate for practical energy management solutions. Our findings suggest that Random Forest is particularly well-suited for forecasting household energy needs, providing reliable data that could help optimize energy use and craft effective energy-saving strategies. Looking ahead, future research should aim to improve dataset quality and explore advanced optimization techniques to push prediction accuracy even further.

Keywords: Random Forest, Prediction, Energy Consumption, Comparison, Machine Learning, Accuracy.

Introduction:

According to the International Energy Agency, residential energy consumption accounts for approximately 29% of global energy use [1]. This significant portion underscores the profound impact of household energy habits on overall energy demand and the environment. Moreover, with global energy consumption expected to rise by nearly 50% by 2050, driven largely by population growth and economic development, understanding and optimizing household energy consumption has become more critical than ever [2].

Household energy consumption significantly influences global energy demand, directly affecting individual expenses and contributing to broader environmental and economic impacts [3]. As the world faces escalating concerns over energy efficiency and sustainability, the need for accurate and reliable predictions of household energy usage has never been more critical. These predictions are essential for enabling smart consumption, reducing energy waste, and supporting the transition to green energy [4]. The challenge is compounded by the variability in household energy consumption patterns, influenced by climate, socio-economic status, and technological advancements [5].

Accurate energy consumption forecasting is crucial for designing efficient energy management

strategies, which help reduce energy costs and mitigate environmental impacts [6]. On a larger scale, these improvements can lead to significant reductions in greenhouse gas emissions, contributing to global efforts to combat climate change [7]. Furthermore, understanding consumption patterns can empower consumers to make informed decisions about their energy use, ultimately fostering a more sustainable and cost-efficient lifestyle [8]. Accurate prediction also facilitates the integration of renewable energy sources, helping to stabilize the grid and ensure a reliable energy supply [9].

Machine learning is widely used for predicting energy consumption [10]. Additionally, among all techniques, Random Forests are among the top-performing methods for energy consumption forecasting, offering significant improvements in predictive accuracy compared to traditional statistical models [11]. This paper investigates the application of machine learning with the Random Forest method for predicting household energy consumption. A Random Forests consists of an ensemble of decision trees, where each tree is constructed using a randomly sampled subset of the data [12] .The study aims to evaluate the method's effectiveness and accuracy in providing reliable forecasts, which are crucial for designing energy-saving strategies, optimizing household energy use, and informing policy decisions.

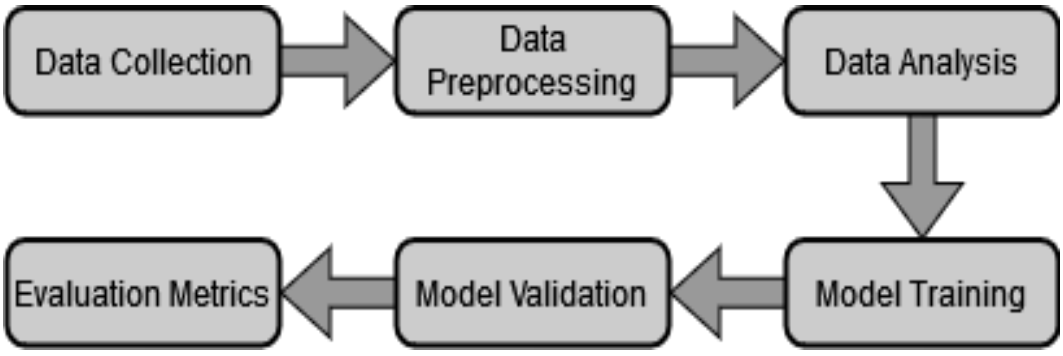
Several previous studies have used various methods to predict energy consumption. For instance, in reference [13] discussed how the Gradient Boosting Regression Tree (GBRT) and Autoregressive Integrated Moving Average Gradient Boosting Regression Tree (ARIMA-GBRT) are better than the other commonly used algorithm models in the electrical energy consumption prediction. In reference [14], the Support Vector Machine (SVM) method shows the most promising result, with RMSE valued at 4,75 and 3,59. In reference [15], Artificial Neural Networks (ANN) have proven to be highly effective in predicting and optimizing energy consumption across various applications, particularly in building energy management systems (BEM). In reference [16] highlights the effectiveness of regression trees in dealing with time-series data and the specific challenges posed by energy forecasting.

The paper is structured as follows: the first section is the introduction, which explains the objectives, problem, and solutions addressed in this research. The second section details the methodology and data used in this study, including data collection, preprocessing, and the implementations of the Random Forest algorithm. The third section presents the results and analysis, demonstrating the model's performance and predictive accuracy. The final section provides the conclusion of the research and provides recommendations for future research.

Method:

This Section details the methodology used to predict household energy consumption using the Random Forest algorithm. The goal of this research is to enhance the accuracy of energy consumption prediction models by developing a predictive model to estimate power consumption and comparing its performance with other methods. The steps followed in this research to assess the efficiency of our proposed strategy are outlined in Fig. 1.

Fig. 1. Research Steps



A. Data Collection

The dataset used in this study was obtained from the UCI Machine Learning Repository, specifically the "Individual household electric consumption" dataset by Georges Hebrail and Alice Berard [17]. This dataset contains comprehensive data gathered in a house in Sceaux (7km from Paris, France). The data spans from December 2006 to November 2010 (47 months), containing measurements of electric power consumption in a one-minute sampling rate. The dataset includes parameters such as date and time, global active power (in kilowatts), global reactive power (in kilowatts), voltage (in volts), global intensity (in amps), sub-metering from 1, 2, and 3 (in watt-hours).

B. Data Preprocessing

The dataset contains no missing data, so no imputation or missing data handling was required during preprocessing. Feature engineering involves creating time-based features such as hour of the day, day of the week, and month of the year. Additionally, lag features, such as previous hour consumption and previous hour intensity, were generated.

C. Data Analysis

The predictive model in this research was built and trained using a dataset that included 2,075,258 data samples.

D. Model Training and Validation

Random Forest is selected for its robustness and capability to manage complex relationships within a dataset. The model was trained on the training set using a computer equipped with an 8-core, 16 threats processor, 16GB of RAM, and a Graphics Card with 4GB of VRAM. To conduct the analysis, this dataset was divided into two subnets: training data and test data. 80% of the data was used as the training set to build the predictive model, while the remaining 20% was used as the test set to validate the model's accuracy.

E. Evaluation Metrics

After developing an energy consumption prediction model, it is critical to assess its performance to determine its accuracy in predicting energy consumption. The evaluation of the model is conducted to assess the performance of the Random Forest. The Model performance was evaluated using Root Mean Square Error (RMSE), Normalized Root Mean Square Error (NRMSE), and accuracy.

Root Mean Squared Error (RMSE) is a commonly used measure to evaluate the accuracy of a predictive model. It calculates the square root of the average squared differences between the actual value and the predicted value. It is expressed as Eq. (1):

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

y_i represents the actual value, $\hat{y}_i$ denotes the predicted values, and n is the number of samples. RMSE measures how well the model's predictions match the actual data. A lower RMSE value indicates better predictive accuracy.

Normalized Root Mean Squared Error (NRMSE) is the RMSE normalized by the mean of the observed data. This normalization allows for a more meaningful comparison between different datasets or models by expressing the error as a proportion of the mean value of the actual data. It is given by Eq. (2):

$$NRMSE = \frac{n}{\sum_{i=1}^{n} y_i} \times RMSE \quad (2)$$

Where $\frac{n}{\sum_{i=1}^{n} y_i}$ denotes the reciprocal of the mean of actual values, by normalizing RMSE, NRMSE provides a relative measure of the prediction error, making it easier to compare performance across different contexts.

Accuracy is derived from NRMSE and is expressed as Eq. (3):

$$accuracy = (1 - NRMSE) \times 100\% \quad (3)$$

The metric converts the NRMSE into a percentage format, where a higher value indicates better model performance. An accuracy of 100% would mean perfect predictions with no error, while lower values indicate increasing prediction errors.

Result and Discussion:

In this section, we will show the results of the Random Forest method. Before comparing it with another method, we will first demonstrate the variation in the performance of the Random Forest method based on the number of trees used in training the models.

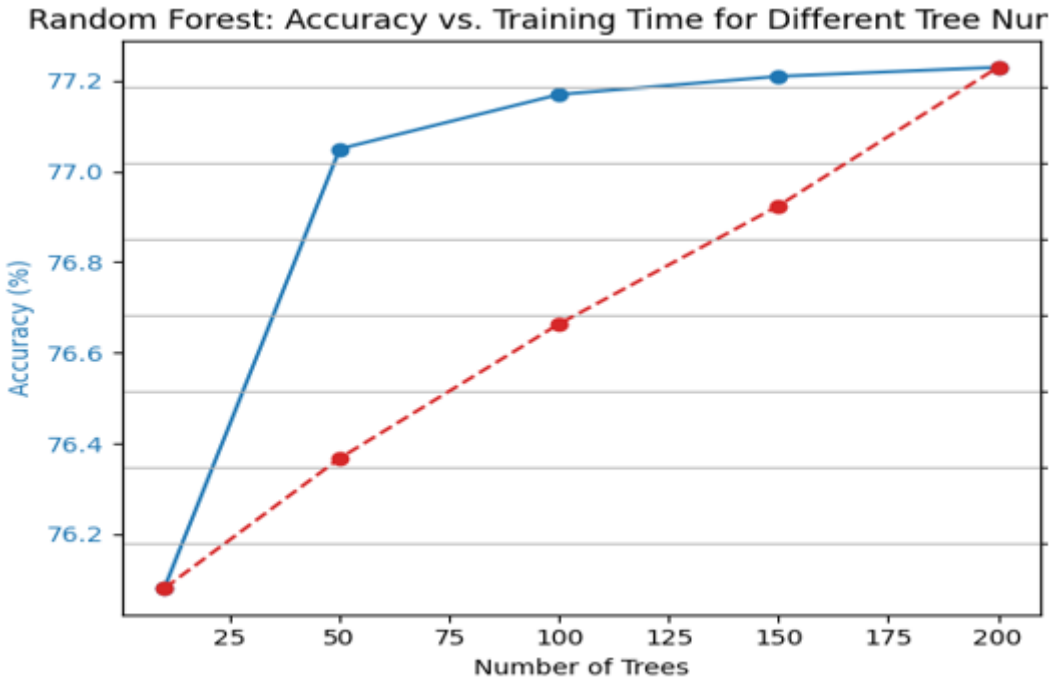


Fig. 2. Random Forest's tree comparison graph

Table 1. Random Forest's tree training time comparison

Numbers of Tree	Training Time
10	82.2s
50	423.9s
100	776.6s
150	1086.6s
200	1450s

Table 2. Random Forest's tree accuracy comparison

Numbers of Tree	RMSE	NRMSE	Accuracy
10	0.26	0.34	76.08%
50	0.25	0.23	77.05%
100	0.25	0.23	77.17%
150	0.25	0.23	77.21%
200	0.25	0.23	77.23%

From the results of testing different numbers of trees in the Random Forest method, it can be concluded that the accuracy difference between using 10 and 50 trees is significant, with a 1% improvement. However, the accuracy increases only by 0.1% when the number of trees increases from 50 to 100, and similarly from 100 to 150. A further increase from 150 to 200 trees results in a 0.01% improvement. Based on these results, we select 50 trees as the optimal number for the Random Forest method, as it provides a 1% increase in accuracy compared to 10 trees, with only a 5-minute difference in training time.

We also compared the Random Forest method with the Gradient Boosting, Regression Tree Support Vector Machine, and Naïve Bayes methods. The parameters evaluated include training time and accuracy shown in Table 3.

Table 3. Training Time Comparison

Method	Training Time
Random Forest (50)	423.9s
Gradient Boosting	180.9s
Regression Tree	11.1s
Support Vector Machine	48.9s
Naïve Bayes	1.4s

Table 4. Accuracy Comparison

Method	RMSE	NRMSE	Accuracy
Random Forest (50)	0.25	0.23	77.05%
Gradient Boosting	0.26	0.24	76.10%
Regression Tree	0.36	0.33	67.48%
Support Vector Machine	0.26	0.24	75.80%
Naïve Bayes	0.44	0.44	56.42%

From Table 4, it is evident that while the Random Forest method requires a longer training time, it achieves higher accuracy (77.0.5%) than the other methods. Despite the longer training time, the Random Forest method shows a 1% improvement in accuracy over the Gradient Boosting method. Among the five methods tested, we chose the Random Forest method due to its superior accuracy, despite its substantially longer training time of 400 seconds compared to the other methods, which have training times of less than 190 seconds.

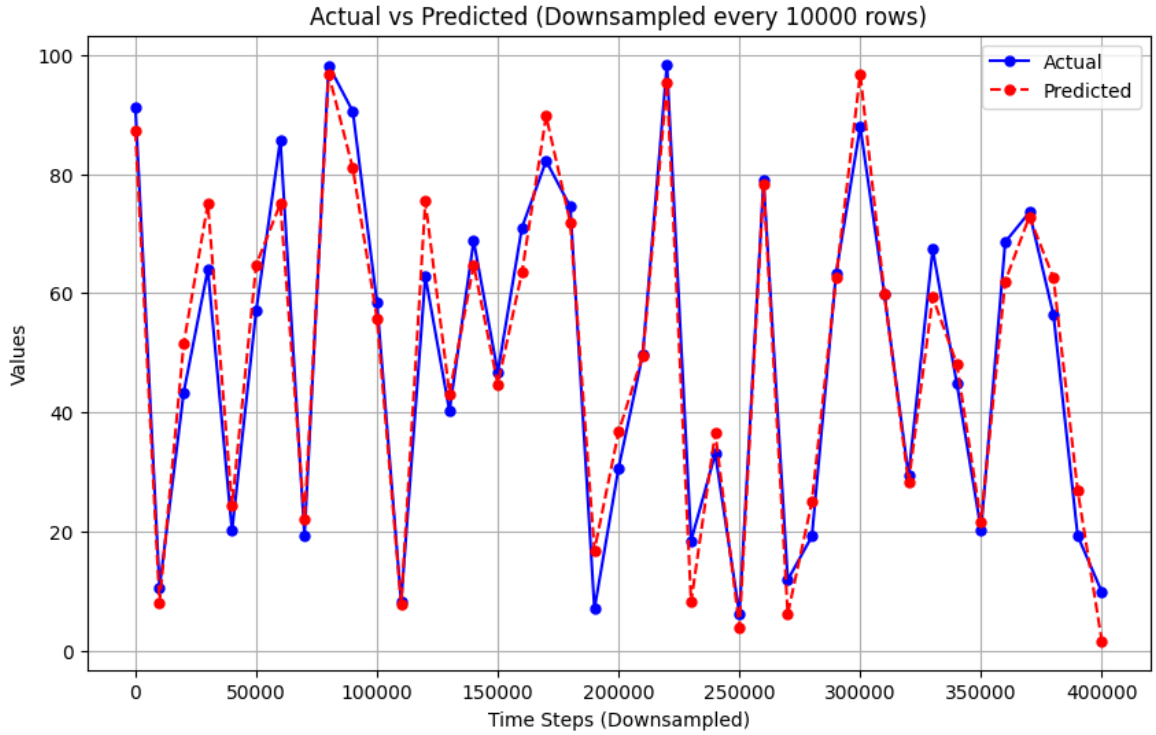


Fig. 3. Actual vs Predicted Values

Conclusion:

In this study, we employed Random Forest techniques to predict energy consumption using a dataset of 8 parameters. Among the various methods tested, Random Forest emerged as the most accurate, achieving an accuracy of 77.05%, despite having a longer training time than other approaches.

One significant limitation of this research is the dataset's quality, which may hinder the model's

ability to generalize effectively to larger or more varied datasets. Future research could focus on implementing advanced optimization techniques to improve prediction accuracy further. additionally, applying this methodology to other datasets or incorporating more variables could offer valuable insight and enhance the robustness of energy consumption prediction.

References:

[1] IEA, Global Energy Review 2021, Paris: https://www.iea.org/reports/global-energy-review-2021, 2021.

[2] IEA, "Global Energy Stauts Report 2019," https://www.iea.org/reports/global-energy-co2-status-report-2019, Paris, 2019.

[3] M. S. Bakare, A. Abdulkarim, M. Zeeshan and A. N. Shuaibu, "A comprehensive overview on demand side energy management towards smart grids: challenges, solutions, and future direction," *Energy Informatics,* vol. 6, pp. 1-59, 2023.

[4] H. Lund and B. V. Mathiesen, "Energy system analysis of 100% renewable energy systems—The case of Denmark in years 2030 and 2050," *Energy,* vol. 34, no. 5, pp. 524-531, 2009.

[5] A. Kavousian, R. Rajagopal and M. Fischer, "Determinants of residential electricity consumption: Using smart meter data to examine the effect of climate, building characteristics, appliance stock, and occupants' behavior," *Energy,* vol. 55, pp. 184-194, 2013.

[6] R. Mathumitha, P. Rathika and K. Manimala, "Intelligent deep learning techniques for energy consumption," *Artificial Intelligence Review,* vol. 57, p. 35, 2024.

[7] IEA, "World Energy Outlook 2021," https://www.iea.org/reports/world-energy-outlook-2021, Paris, 2021.

[8] DOE, "Energy Savers Guide," https://www.energy.gov/energysaver/energy-saver-guide-tips-saving-money-and-energy-home, 2020.

[9] IEA, "Renewable Energy Market Update - June2023,"https://www.iea.org/reports/renewable-energy-market-update-june-2023, Paris, 2023.

[10] S. Bourhnane, M. R. Abid,, R. Lghoul, K. Zine-Dine, N. Elkamoun and D. Benhaddou, "Machine learning for energy consumption prediction and scheduling in smart buildings," *SN Applied Sciences,* vol. 2, 2020.

[11 S. S. Aravind, P. Tanna and P. Vittaldas , "Modeling Energy Consumption Using Machine Learning," *Frontiers in Manufacturing Technology,* vol. 2, 2022.

[12 L. Breiman, "Random Forests," *Machine Learning,* vol. 45, no. 1, pp. 5-32, 2001.

[13 P. Nie, M. Roccotelli, M. P. Fanti, Z. Ming and Z. Li, "Prediction of home energy consumption based on gradient boosting regression tree," *Energy Report,* vol. 7, pp. 1246-1255, 2021.

[14] "Energy consumption prediction by using machine learning for smart building: Case study in Malaysia," *Developments in the Built Environment,* vol. 5, 2021.

[15] P. Michailidis, I. Michailidis, S. Gkelios and E. Kosmatopoulos, "Artificial Neural Network Applications for Energy Management in Buildings: Current Trends and Future Directions," *Energies,* vol. 17, no. 3, 2024.

[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, 2024.

[17] G. Hebrail and A. Berard, "Individual Household Electric Power Consumption," UCI Machine Learning Repository, 2012.