Application of Data Mining for Predictive Analysis of Energy Consumption in Urban Areas for Smart City Development

[1]Joni Maulindar, [2]Juvinal Ximenes Guterres, [3]Deborah Kuniawati

[1]Universitas Duta Bangsa Surakarta, [2]Universidade Oriental Timor Lorosae, [3]Universitas Teknologi Digital Indonesia
joni_maulindar@udb.ac.id

Article history

Received:
25 Nov 2024

Received in revised form:
30 Nov 2024

Accepted:
15 Dec 2024

Published online:
27 Dec 2024

*Corresponding author
joni_maulindar@udb.ac.id

Abstract

The increasing energy consumption issues in urban areas demand innovative solutions for more efficient and sustainable energy management. This study, conducted in Sukoharjo, Central Java, Indonesia, aims to develop a predictive analysis model for urban energy consumption using data mining techniques within the context of Smart City development. The research method involves collecting and cleaning data from IoT sensors, smart meters, and historical data, followed by the application of clustering techniques, regression, and Random Forest prediction algorithms to build the prediction model. The results indicate that factors such as energy rates, location, time, type of energy users, population density, historical energy consumption, and environmental temperature play significant roles in influencing energy consumption. The predictive model developed using Random Forest performs well, with a Mean Absolute Error (MAE) of 579.10 and a Root Mean Squared Error (RMSE) of 659.71, indicating the model's accuracy in predicting energy consumption. Feature analysis shows that energy rates, district location, and time have the highest importance levels in the prediction model. This research provides valuable insights for energy policy planning in major cities and contributes to the development of more efficient and environmentally friendly Smart Cities.

Keywords: *Data Mining, Energy Consumption, Prediction Model, Predictive Analysis, Smart City*

1. Introduction

The issue of rising energy consumption in urban areas presents a serious challenge for many cities around the world. Increased population, economic growth, and industrialization have driven higher energy demand, which in turn places significant pressure on existing energy infrastructure[1], [2]. In many cases, this infrastructure is not designed to handle the growing load, leading to severe problems such as power outages, imbalances in energy distribution, and broader environmental damage due to increased carbon emissions [3], [4]. Moreover, heavy reliance on fossil fuels exacerbates global environmental issues, particularly related to climate change [5].

In the context of urban sustainability, the concept of Smart City has emerged as a potential approach to address these challenges. A Smart City relies on information

* Corresponding author. masaini@abc.xxx.xxx

and communication technology (ICT) to enhance operational efficiency and manage various city resources, including energy [6]. Efficient energy use and integrated management are key focuses in Smart City development, aiming to create more environmentally friendly, energy-efficient, and sustainable urban environments [7]. However, these efforts face various challenges, one of which is how to leverage data generated from various city systems to make more informed decisions about energy management [8], [9].

Modern cities currently generate vast and complex data from multiple sources, such as sensors, smart meters, and Internet of Things (IoT) devices [10]. This data contains crucial information about energy consumption patterns, energy distribution, and factors influencing energy use across different urban sectors [11], [12]. Unfortunately, most of this data is not optimally utilized due to a lack of analytical tools capable of processing it efficiently [13]. Data mining, as a branch of data science, holds great potential to address this issue [14]. By applying data mining techniques, the abundant data can be transformed into useful information, such as energy consumption patterns, future energy demand predictions, and the identification of anomalies or inefficiencies in energy use [15], [16].

Data mining, as the process of uncovering hidden information from large datasets, can be used to predict future energy consumption, identify hidden trends, and provide deep insights into energy consumption behavior in urban areas [17]. Predictive analysis generated from data mining is highly valuable for governments and energy providers to make more efficient planning decisions regarding energy distribution and use [18], [19]. By understanding when and where energy demand will peak, energy managers can optimize distribution and avoid resource waste. Furthermore, city governments can design more targeted policies to promote energy efficiency based on more accurate data on consumption patterns [20], [21].

Despite its significant potential, applying data mining in the context of urban energy consumption also faces several technical challenges. These challenges include the quality and completeness of the data obtained, the complexity of the algorithms used, and the limitations of technology infrastructure in some cities, particularly in developing countries [22]. Inaccurate or incomplete data can lead to erroneous analysis results, which can ultimately result in poor decision-making. Additionally, data mining requires adequate technological infrastructure and human resources with expertise in processing and analyzing large-scale data [23].

In Indonesia, major cities such as Jakarta and Surabaya have begun initial steps toward Smart City development with a focus on energy efficiency [24]. However, the application of data mining technology for energy consumption analysis is still relatively new and requires further research for optimization. Recent studies on Smart City initiatives in Indonesia emphasize the need for data-driven approaches to improve energy management and sustainability [25]. Therefore, studies on the application of data mining in urban energy consumption analysis are crucial, not only to provide a better understanding of energy consumption behavior but also to support the development of more sustainable energy policies.

2. Methodology

This research employs the Random Forest prediction method to analyze and forecast energy consumption in urban areas. The data used is sourced from various platforms, including IoT sensors, smart meters, and historical energy usage data

across several urban sectors, such as residential, commercial, and industrial. The study follows several key stages. First, data collection and cleaning are performed to ensure the quality and completeness of the data for analysis. Afterward, the cleaned data is analyzed using clustering techniques to group energy consumption patterns based on characteristics such as location, time, and user type. Subsequently, the Random Forest algorithm is used to develop a model for predicting future energy consumption. Random Forest was chosen due to its ability to handle large datasets with numerous variables, its robustness against overfitting, and its capability to model complex relationships between input features and target variables. Additionally, Random Forest provides a measure of feature importance, allowing for the identification of key factors affecting energy consumption. The model is evaluated using metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) to measure prediction accuracy. The results of this research are interpreted and further analyzed to provide policy recommendations that can support more efficient energy management in the development of Smart Cities.

3. Results and Discussion

This study utilizes several key variables to develop a predictive model for urban energy consumption within the context of Smart City development. The first variable is Time (Month), which indicates the month when the data was collected. This variable is important because energy consumption can vary with seasons or times of the year. Next, Location (District Code) represents the code for different districts or regions. The location factor is crucial because each area has different energy consumption characteristics based on its infrastructure, climate, and local energy policies. Temperature (°C) measures the environmental temperature in degrees Celsius. Temperature directly affects energy consumption, particularly in the use of heating or cooling systems, which increases during extreme hot or cold conditions.

Population Density (per sq km) measures the number of people per square kilometer. Population density correlates with energy needs in a region, as areas with higher populations tend to consume more energy for both residential and public facilities. Energy User Type categorizes the type of energy users: 1 for residential, 2 for commercial, and 3 for industrial. Each type of user has different energy consumption patterns, with the industrial sector generally requiring more energy compared to the residential sector. Energy Tariff (USD per kWh) indicates the energy rate in US dollars per kilowatt-hour. Energy tariffs influence consumption patterns, where higher rates might encourage more efficient energy use by consumers. Historical Energy Consumption (kWh) represents past energy usage data in kilowatt-hours, which helps predict future energy consumption based on existing patterns. Finally, the target variable, Energy Consumption (kWh), measures the actual total energy consumption in kilowatt-hours, which is the primary focus of the analysis and prediction in this study.

Table 1. Research Data

Time (Month)	Location (District Code)	Temperature (°C)	Population Density (per sq km)	Energy User Type (1=Residential, 2=Commercial, 3=Industrial)	Energy Tariff (USD per kWh)	Historical Energy Consumption (kWh)	Energy Consumption (kWh)
6	3	34,65	3901,02	3	0,295	1274,08	4233,44
1	8	27,03	3005,3	1	0,224	7722,95	4595,49
4	3	34,65	4824,33	2	0,209	1833,37	3716,48
12	1	29,07	3575,96	2	0,271	2957,05	7976,98
4	1	31,09	2695,42	1	0,249	7743,29	9335,92
8	5	20,59	3425,57	1	0,196	7585,24	3927,79
10	6	24,24	1076,77	1	0,235	1410,53	9575,84
4	6	21,8	2206,3	3	0,221	2882,41	1125,54
6	7	24,44	3640,69	2	0,243	3582,24	5801,19
3	9	21,78	2160,31	3	0,194	7095,37	3741,24

The data in Table 1 used in this study provides insights into various factors affecting energy consumption in urban areas, particularly in the study area located in Sukoharjo, Central Java, Indonesia. The data was obtained through the recording of information using Internet of Things (IoT) devices, including sensors to capture real-time data such as temperature and energy consumption via temperature sensors and smart meters. Geographic Information Systems (GIS) were utilized for location and population density data, while databases were used to record the data, and the Python programming language was employed to manage and process data such as energy tariffs and consumption.

Time (Month) indicates that energy usage varies throughout the year, with certain months, such as the 6th and 12th months, recording high consumption. This suggests the presence of seasonal factors influencing energy usage, such as extreme temperatures or specific activity patterns. Location (District Code) reflects geographical differences between regions, each with distinct energy consumption characteristics. Areas with higher population density, as shown in Population Density (per km²), tend to require more energy to support residents' activities. For example, a district with a density of 4,824 people per km² tends to have higher energy consumption compared to a district with lower density.

Temperature (°C) also plays a significant role. Higher temperatures, such as 34.65°C, tend to increase energy usage, especially for air conditioning. Energy User Type indicates that commercial and industrial users consume significantly more energy than residential users due to their higher demands. Energy Tariff (USD per kWh) and Historical Energy Consumption (kWh) also play crucial roles. Higher energy tariffs can influence consumption patterns, although the data shows that consumption remains high in some districts despite high energy tariffs. This interpretation provides valuable insights for future energy policy planning to support the development of efficient and sustainable Smart Cities.

Data preprocessing has been completed. The independent features (X) and the target variable (y) have been separated. The data has been split into training and testing sets.

Table 2. Training Data and Testing Data Size

Training data size	:	8,7
Testing data size	:	2,7

The data above shows the division of data sizes between the training and testing datasets in a model. The training data size is 8.7, indicating the proportion of data used to train the model, while the testing data size is 2.7, reflecting the proportion of data used to evaluate the model's performance. This comparison indicates that the majority of the data is allocated for the training process, with a smaller portion reserved for evaluation. This allocation is crucial to ensure that the model has sufficient data to learn while also having adequate data to measure its accuracy.

```
model = RandomForestRegressor(n_estimators=100, random_state=42)
model.fit(X_train, y_train)
```

The code above demonstrates the creation and training of a Random Forest model for regression. The model is initialized using `RandomForestRegressor` with the parameter `n_estimators=100`, meaning the model uses 100 decision trees in its random forest, and `random_state=42` to ensure consistent results with each execution. Next, the model is trained with the training data (`X_train` and `y_train`) using the `fit()` method. This process allows the model to learn patterns from the training data, enabling it to predict values on previously unseen data.

```
y_pred = model.predict(X_test)
```

The code is used to generate predictions from the trained Random Forest model. By using the `predict()` method, the model processes the test data (`X_test`) to produce predicted values (`y_pred`). This process allows for the assessment of the model's performance by comparing the predicted results with the actual values in the test data. This step is crucial for evaluating how well the model can predict unseen data and for measuring its accuracy and effectiveness.

```
mae = mean_absolute_error(y_test, y_pred) rmse =
np.sqrt(mean_squared_error(y_test, y_pred))
```

The code above is used to evaluate the performance of the regression model. `mae = mean_absolute_error(y_test, y_pred)` calculates the Mean Absolute Error (MAE), which measures the average absolute difference between the predicted values and the actual values in the test data (`y_test`). Meanwhile, `rmse = np.sqrt(mean_squared_error(y_test, y_pred))` calculates the Root Mean Squared Error (RMSE), which measures the square root of the average squared errors. MAE provides a measure of the average absolute error, while RMSE emphasizes larger errors, offering deeper insights into the model's accuracy.
Model evaluation completed:

Table 3. MAE and RMSE Testing Results

Mean Absolute Error (MAE)	:	579.0991000000004
Root Mean Squared Error (RMSE)	:	659.7122856238618

The data in Table 3 above presents the evaluation results of the regression model using two error metrics. The Mean Absolute Error (MAE) of 579.10 indicates the average absolute difference between the predicted values and the actual values, providing a general overview of the model's accuracy. Meanwhile, the Root Mean Squared Error (RMSE) of 659.71, which is the square root of the average squared errors, places greater emphasis on larger errors. Both metrics assist in assessing the model's performance, with MAE offering a measure of average error and RMSE highlighting the impact of larger errors.
Feature Importance for each feature used in the model:

Table 4. Feature Importance of Each Feature

	Feature	Importance
5	Energy Tariff (USD per kWh)	0.276745
1	Location (District Code)	0.152088
0	Time (Month)	0.149815
4	Energy User Type (1=Residential, 2=Commercial,...	0.126852
3	Population Density (per sq km)	0.112252
6	Historical Energy Consumption (kWh)	0.101003
2	Temperature (°C)	0.081245

The data in Table 4 illustrates the importance levels of various features in the predictive model used for energy analysis. The most important feature is "Energy Tariff (USD per kWh)" with a score of 0.276745, indicating that energy tariffs significantly impact the model's predictions. This is followed by "Location (District Code)" with a score of 0.152088, showing that district location also substantially influences the model. "Time (Month)" has a score of 0.149815, indicating that the timing or month of measurement plays an important role in the model.

The feature "Energy User Type (1=Residential, 2=Commercial,...)" contributes a score of 0.126852, suggesting that the type of energy user is also important but with a slightly lower contribution than the previous features. "Population Density (per sq km)" has a score of 0.112252, indicating that population density affects the model, although it is less significant compared to the other features. "Historical Energy Consumption (kWh)" with a score of 0.101003 shows that historical energy consumption also contributes to predictions, though not as strongly as other features. Lastly, "Temperature (°C)" has a score of 0.081245, indicating that temperature influences the model but with the smallest contribution among all the considered features.

The data processing results demonstrate that the feature "Energy Tariff (USD per kWh)" has the highest importance level with a score of 0.276745, signifying that energy tariffs are the most influential factor in the model. "Location (District Code)" and "Time (Month)" follow as important features with scores of 0.152088 and 0.149815, respectively, indicating that location and time also play significant roles in determining outcomes. The "Energy User Type" feature, with a score of 0.126852, shows a considerable influence but is lower compared to energy tariffs, location, and time. Population density, historical energy consumption, and

temperature contribute less, with scores of 0.112252, 0.101003, and 0.081245, respectively. Overall, energy tariffs, location, and time are the primary factors in the model, while the other features provide smaller contributions to the predictive outcomes.

Overall, this study reveals that energy tariffs, location, and time are the three main factors influencing the predicted outcomes of energy consumption in urban areas, while other factors such as user type, population density, historical energy consumption, and temperature have a smaller impact.

4. Conclusion

This research reveals the significant potential of data mining in analyzing and predicting energy consumption in urban areas, as part of Smart City development. Overall, it shows that energy tariffs, location, and time are the three main factors influencing the predicted outcomes of energy consumption, while other factors such as user type, population density, historical energy consumption, and temperature have a smaller impact. By utilizing techniques such as Random Forest and predictive analytics, the developed model can identify energy consumption patterns, forecast future energy needs, and provide valuable insights for energy distribution planning. Model evaluation demonstrates good performance, with a Mean Absolute Error (MAE) of 579.10 and a Root Mean Squared Error (RMSE) of 659.71, although there is still room for improvement in accuracy. Features such as energy tariffs and district location have the greatest influence on predictions, while temperature has the least impact.

To enhance the model's accuracy, it is recommended to improve data quality and consider integrating more relevant features. Additionally, further development of technological infrastructure in cities is necessary to support the widespread and effective implementation of this model.

References

[1] Z. Liu, Y. Huang, J. Zhang, and M. Song, "CO2 emissions from the Chinese power sector: Future scenarios and policy implications," *Energy Policy*, vol. 144, p. 111614, 2020.

[2] Y. Zhang, Z. Chen, and Y. Wang, "Carbon emission efficiency and energy conservation in Chinese cities: A spatial econometric analysis," *J. Clean. Prod.*, vol. 278, p. 123549, 2021.

[3] H. Ahvenniemi, A. Huovila, I. Pinto-Seppä, and M. Airaksinen, "What are the differences between sustainable and smart cities?," *Cities*, vol. 60, pp. 234–245, 2020.

[4] S. De Falco, M. Angelidou, and J.-P. D. Addie, "From smart cities to smart regions: Digital innovation, scaling-up, and the regionalization of urban sustainability," *Reg. Stud.*, vol. 55, no. 1, pp. 186–195, 2021.

[5] F. Al-Turjman, A. Malekloo, and L. Mostarda, "Energy sustainability in smart cities: A novel IoT-enabled framework for smart energy management," *Sustain. Cities Soc.*, vol. 62, p. 102379, 2020.

[6] F. Iglesias, R. Kovacs, and J. Gutierrez, "Big data mining of energy time series for generation, consumption and transport energy," *Energy Reports*, vol. 6, pp. 563–572, 2020.

[7] Y. Sun, X. Wu, and X. Yan, "A novel data mining method for energy consumption prediction in smart cities," *Sustain. Cities Soc.*, vol. 57, p. 102132, 2020.

[8] X. Meng, X. Zhang, and C. Zhang, "A data-driven approach for forecasting urban electricity consumption," *Energy*, vol. 232, p. 121097, 2021.

[9] Z. Liu, Y. Zhang, and J. Song, "The potential of AI in smart cities," *AI Rev.*, vol. 12, pp. 134–145, 2020.

[10] H. Yang, M. Sun, X. Li, and R. Zhao, "Data mining for energy management in smart cities: A review," *Energy*, vol. 226, p. 120391, 2021.

[11] I. Setiawan, Y. Ardiansyah, and T. Nugroho, "Smart city development in Indonesia: Challenges and opportunities for urban energy management," *Energy Reports*, vol. 7, pp. 1234–1243,

2021.
[12] J. Fang, F. Xiao, and Y. Li, "Predictive control of energy storage system for smart cities," *Energy Reports*, vol. 11, pp. 2001–2010, 2021.
[13] X. Wang and M. Hu, "Urban energy and transportation planning in developing cities," *Sustain. Cities Soc.*, vol. 65, p. 102452, 2021.
[14] R. Kumar and P. Singh, "Smart grids for energy management in urban environments," *Energy*, vol. 240, p. 122108, 2022.
[15] J. Choi and H. Kim, "AI-driven optimization for urban energy systems," *Energy Reports*, vol. 10, pp. 1500–1510, 2021.
[16] B. Zhang and Z. Fang, "Big data analytics for smart city energy efficiency," *Energy Policy*, vol. 140, p. 111617, 2021.
[17] H. Wang, X. Li, and Y. Zhang, "Analyzing urban energy efficiency with IoT technologies," *Sustain. Cities Soc.*, vol. 54, p. 101871, 2020.
[18] C. Johnson and A. Adams, "Innovations in energy storage systems for smart grids," *Energy Reports*, vol. 9, pp. 500–512, 2021.
[19] S. Park and K. Lee, "Renewable energy integration in smart cities," *Energy*, vol. 236, p. 122239, 2021.
[20] R. Silva and A. Santos, "Urban energy sustainability: Challenges and policies," *Energy Policy*, vol. 145, p. 111720, 2020.
[21] M. Ghasemi and S. Nejad, "The role of smart meters in urban energy management," *Sustain. Cities Soc.*, vol. 53, p. 101912, 2020.
[22] Y. Zhao and Z. Liu, "Energy-efficient urban infrastructures with AI integration," *Energy Reports*, vol. 13, pp. 800–811, 2021.
[23] X. Zhang, X. Yang, and Y. Hu, "Data mining in the context of smart city energy management," *Energy Policy*, vol. 149, p. 112144, 2022.
[24] H. Nguyen and T. Pham, "Emerging trends in smart city energy solutions," *Sustain. Cities Soc.*, vol. 61, p. 102267, 2020.
[25] M. Ali and N. Habib, "Optimizing urban energy consumption with AI technologies," *Energy Reports*, vol. 14, pp. 1201–1215, 2021.