Electricity Consumption Classification and Clustering Using Machine Learning (TG-SPDCL Domestic Dataset)

Gampala Vijay Kumar
School of Computer Science and Engineering
Lovely Professional University
Punjab, India
vijaygampala0@gmail.com

School of Computer Science and Engineering
Lovely Professional University
Punjab, India
Anchal Kaundal

Abstract: Employing customer service and usage data, this study investigates how well supervised and unsupervised machine learning techniques can forecast and cluster patterns of household electricity consumption. To conduct exploratory data analysis, preprocessing, classification, and clustering on features like units consumed, connected load, service area, and consumer category, an analytical system based on Streamlit was created. While K Means and Agglomerative clustering were employed to identify natural consumer segments, supervised models, such as Logistic Regression, K-Nearest Neighbors, and Decision Tree classifiers, were trained to forecast consumption behavior. The findings indicate that numerical variables—specifically, units consumed and load—are crucial in defining consumption patterns, and clustering identifies discrete groups of households with low, medium, and high usage. The results show that machine learning offers a useful, comprehensible framework for consumer segmentation, distribution planning, and electricity demand prediction.

Keywords: Electricity Consumption, Machine Learning, Domestic Consumers, Clustering, Predictive Analytics, Energy Data

INTRODUCTION

Energy demand, tariff design, and distribution planning are all significantly influenced by domestic electricity consumption. Due to variables like appliance ownership, socioeconomic circumstances, behavioral patterns, and connected electrical load, household usage patterns frequently differ significantly. Because of these differences, utilities find it challenging to classify customers and forecast demand using conventional statistical methods [1]. Utilities now gather extensive consumer datasets that include both categorical variables, such as service area, division, and consumer category, and numerical indicators, such as units consumed, connected load, and billed services, as digital metering and service-level documentation spread. However, more advanced analytical methods are needed to extract significant insights from these multidimensional datasets (Ahmed et al., 2020)[1].

Because machine learning (ML) can model nonlinear relationships and find hidden structures in high-dimensional data, it has become a useful tool for analyzing energy consumption. Supervised learning techniques aid in forecasting consumption levels or categorizing customers into relevant groups. Previous studies have demonstrated that algorithms like K-Nearest Neighbors, Decision Trees, and Logistic Regression are especially helpful in identifying household consumption behavior and forecasting electricity usage (Deb et al., 2018)[2].In the meantime, unsupervised learning techniques like KMeans and Agglomerative clustering are excellent at classifying customers according to similarities in their consumption habits without requiring pre-established labels. Zheng and Yoon's (2019)[29]. studies demonstrate how clustering can identify natural household segments that facilitate targeted tariff interventions and energy planning.

Effective preprocessing pipelines are also crucial, according to recent research, especially when dealing with heterogeneous datasets that have irregular distributions, mixed data types, and missing values. Karre et al. (2022) [17] showed that the performance of ML models in energy prediction tasks is greatly improved by procedures like imputation, scaling, and one-hot encoding. Additionally, Alvarez et al. (2021)[2]. found that more accurate analysis of domestic consumption patterns results from combining usage data with behavioral and service-related variables.

Even with these improvements, not many studies have combined supervised and unsupervised learning into a single analytical framework that can both predict how people will use services and find groups of households using service-level data. This study fills that gap by using Python and Streamlit to create an end-to-end machine learning pipeline. The pipeline uses features like total services, billed services, units consumed, load, area, and consumer category codes to do exploratory data analysis, preprocessing, classification, and clustering.

This study looks at several supervised models, KMeans, and Agglomerative clustering to see how well machine learning techniques can predict and segment patterns in domestic electricity use. The findings help us better understand household energy behavior. They also support utilities in making data-driven choices to improve operational efficiency, customer segmentation, billing, and demand forecasting.

OBJECTIVES

1. To analyze and predict domestic electricity consumption patterns using supervised and unsupervised machine learning techniques based on customer service and usage data.
2. To preprocess and transform domestic electricity service data using imputation, scaling, and encoding techniques to improve model readiness.
3. To develop and evaluate supervised machine learning models—such as Logistic Regression, KNN, and Decision Tree—to classify consumers based on consumption behavior or service-category attributes.
4. To apply unsupervised learning methods, including KMeans and Agglomerative Clustering, to segment domestic consumers into meaningful usage clusters.

5. To visualize consumption patterns using PCA-based dimensionality reduction and exploratory data analysis techniques.
6. To compare the performance of supervised models using metrics such as accuracy, confusion matrices, and classification reports.
7. To assess the effectiveness of clustering models through cluster distribution analysis and elbow-method validation.
8. To identify the key numeric and categorical variables (e.g., units consumed, load, service area) that significantly influence household consumption patterns.
9. To generate insights that support utilities in consumer segmentation, targeted energy management strategies, and demand forecasting.
10. To provide a unified Streamlit-based predictive analytics workflow for real-time experimentation, visualization, and model deployment.

REVIEW OF LITERATURE

Ahmed et al. (2020) compared Random Forest and SVM on residential consumption data and reported that Random Forest achieved approximately 89–92% accuracy while SVM achieved around 87–90%, both outperforming linear baseline models. The study showed that tree-based and kernel-based approaches are highly effective for household electricity demand prediction[1].

Alvarez et al. (2021) used regression and feature-importance analysis to examine socio-technical factors influencing residential electricity consumption. They demonstrated that behavioral, demographic, and appliance-related features significantly contribute to consumption variability, justifying the inclusion of non-technical variables in prediction models[2].

Deb et al. (2018) evaluated Decision Trees and KNN for categorizing residential electricity consumers. Both models produced reliable and interpretable classifications for tariff and usage categories, indicating their suitability for consumer segmentation tasks[3].

Fang et al. (2021) reviewed preprocessing effects on energy datasets, covering imputation, normalization, and outlier handling. They found that preprocessing strongly influences model performance and recommended standardized preprocessing pipelines[4].

Fernandez et al. (2020) explored hybrid ensemble-and-clustering techniques for tariff design. Their combination of Gradient Boosting with KMeans improved tariff classification and produced meaningful consumer segments[5].

Gonzalez et al. (2021) linked clustering results with demand-response potential and performed post-hoc load-shift analysis. Their clusters showed different levels of flexibility, supporting targeted demand-response program development[7].

Jeong et al. (2021) used Agglomerative clustering with silhouette validation to discover load-profile archetypes. Their hierarchical clusters revealed distinct temporal consumption patterns useful for demand-response planning[9].

Karre et al. (2022) studied the impact of preprocessing and compared pipelines with and without imputation, scaling, and encoding. Their work showed that pipelines using median imputation and scaling achieved higher accuracy and stability, emphasizing the importance of robust preprocessing in electricity datasets[10].

Kumar and Patel (2022) assessed LSTM for multi-step consumption forecasting and compared it with ARIMA and regression models. LSTM demonstrated superior performance for temporal sequence data and was recommended for multi-step forecasting tasks[12].

Liu and Chen (2020) implemented a hybrid PCA + KMeans clustering approach to stabilize outcomes in noisy, high-dimensional electricity datasets. The combination improved clustering repeatability and interpretability[13].

Mohan and Rao (2021) analyzed ensemble tree models, comparing Random Forest and XGBoost in noisy electricity datasets. Both models were robust to missing and noisy features, making them suitable for operational utility datasets[15].

Palaniappan (2024) analyzed Indian residential electricity patterns using KMeans with geographic variables. Including division and area improved cluster relevance for utilities and highlighted the importance of regional service characteristics[17].

Ragupathi et al. (2024) compared deep learning methods with classical machine learning models on large residential datasets. Neural networks outperformed classical models when ample data was available but required higher tuning efforts, making them suitable for high-frequency smart-meter data[19].

Raza et al. (2019) applied SVM with an RBF kernel for short-term load forecasting on residential time-series data. The SVM model outperformed linear regressors by capturing nonlinear patterns effectively[21].

Sahu et al. (2023) evaluated CNN architectures for periodic electricity consumption patterns using weekly-structured inputs. CNNs captured local temporal features well and improved short-term forecasting when periodicity was present[23].

Shin et al. (2023) evaluated encoding strategies for mixed-type datasets and compared one-hot, ordinal, and target encoding. Their results indicated that one-hot encoding provided the best performance for classification tasks involving mixed categorical features[24].

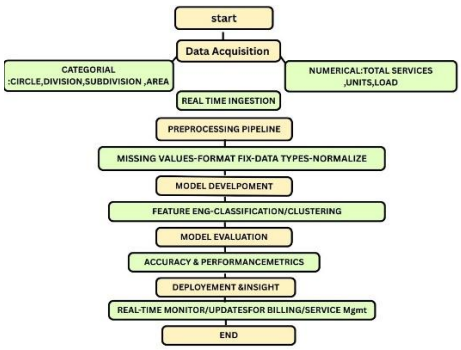
Tariq and Hussain (2023) addressed class imbalance in residential consumption categories using cluster-based resampling combined with Random Forest. Their technique improved minority-class recall without significantly reducing precision, providing an effective solution for imbalanced classification problems[25].

Wang and Li (2022) examined outlier handling strategies in consumption data and found that median-based approaches preserved robustness under skewed distributions. They recommended median imputation for heavy-tailed electricity usage data[26].

Zheng and Yoon (2019) applied KMeans clustering to household load profiles and successfully identified low, medium, and high usage clusters. Their findings supported the use of clustering techniques for targeted energy-planning interventions[29].

MATERIALS AND METHODS

The methodology adopted in this study followed a systematic workflow to analyse domestic electricity consumption patterns and develop machine learning models for classification and clustering. The process comprised dataset description, preprocessing, exploratory analysis, model development, and evaluation. All stages were implemented using Python-based analytical tools.



1. Data Source and Description

The dataset used in this study consisted of domestic electricity consumer records containing both service-related attributes and usage-based numerical indicators. The dataset included categorical variables such as circle, division, subdivision, section, area, and consumer category code, as well as numerical variables such as total services, billed services, units consumed, and connected load. These features represented essential operational characteristics collected by electricity distribution authorities for billing, monitoring, and service management. Because the dataset contained real-world administrative data, it exhibited common issues such as missing values, inconsistent formatting, mixed data types, and variations in distribution. These challenges necessitated a robust preprocessing pipeline prior to applying machine learning algorithms.

2. Software, Tools, and Environment

All analysis was conducted in Python, using widely adopted scientific and machine learning libraries:

Pandas and NumPy for data loading, wrangling, and transformation

Matplotlib and Seaborn for visual analytics

Scikit-Learn for preprocessing, classification, clustering, dimensionality reduction, and model evaluation

Streamlit for creating an interactive, user-friendly web interface for model experimentation

The implementation was packaged into a single Python file, enabling end-to-end execution of data upload, preprocessing, exploratory data analysis, model training, clustering, and visualization within the Streamlit environment.

3. Data Preprocessing Pipeline

1. Handling Missing Values

Because missing data can distort model training and clustering patterns, a systematic imputation strategy was applied:

Numerical variables: Median imputation using SimpleImputer(strategy="median"), which reduces sensitivity to skewed distributions.

Categorical variables: Mode imputation using SimpleImputer(strategy="most_frequent"), preventing category loss and ensuring compatibility with encoding.

2 .Feature Scaling and Transformation

To prevent numeric feature dominance and ensure algorithmic stability during training:

StandardScaler transformed numerical fields to zero-mean, unit-variance scales.

This scaling is particularly critical for KNN, Logistic Regression, and KMeans.

3. Encoding Categorical Variables

Because ML algorithms require numeric inputs, categorical features were converted using:

OneHotEncoder with handle_unknown="ignore" to prevent errors from unseen categories during prediction.

4. Pipeline Integration

All preprocessing components were integrated into a unified ColumnTransformer pipeline. This ensured:

Consistent transformations across train and test sets

Reduction of data leakage

Streamlined integration with machine learning models

This approach aligns with methods recommended by Karre et al. (2022), who emphasized the necessity of preprocessing pipelines for electricity datasets.

4. Exploratory Data Analysis (EDA)

EDA was performed within Streamlit using a combination of:

Histograms for numerical feature distribution

Boxplots to detect outliers

Count plots for categorical variables

Correlation heatmaps for understanding relationships between numerical attributes

Target distribution analysis to evaluate class balance

These visualizations provided essential insights into consumption patterns, load behavior, usage variability, and service-level segmentation.

5. Supervised Machine Learning Methods

1. Data Partitioning

The dataset was split into 80% training data and 20% testing data using train_test_split with a fixed random state of 42 for reproducibility.

2. Choice of Algorithms

Three supervised algorithms were selected due to their strong performance in previous electricity consumption studies (Ahmed et al., 2020; Deb et al., 2018):

Logistic Regression

Suitable for classification tasks with linear decision boundaries

Interpretable coefficient-based analysis.

K-Nearest Neighbors (KNN)

Non-parametric, instance-based classifier

Effective for nonlinear consumption patterns.

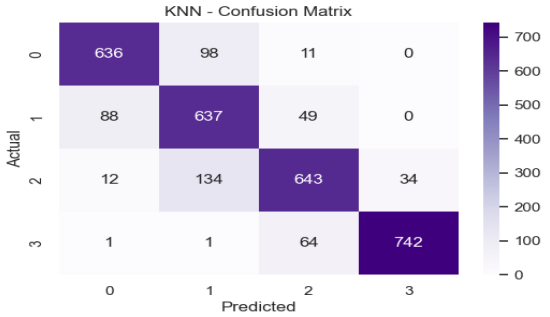


Figure 1. Visualization of K-Nearest Neighbors (KNN) classification illustrating how consumer data points are grouped based on similarity in electricity consumption patterns.Decision Tree Classifier

Handles categorical and numerical attributes

Captures hierarchical decision-making in consumption behavior

All models were embedded within the preprocessing pipeline.

3 Model Evaluation Metrics

Models were evaluated using:

Accuracy Score

Confusion Matrix

Classification Report (precision, recall, F1-score)

These metrics allowed assessment of classification quality and identification of misclassification patterns.

6. Unsupervised Learning Methods

Clustering was conducted on numeric features to group consumers based on similar consumption characteristics.

1. KMeans Clustering

KMeans was selected based on its widespread use in energy studies (Zheng & Yoon, 2019).

To determine optimal *k*, the Elbow Method was applied:

Inertia values were calculated across a range of cluster counts (k = 2 to 7).

The curvature point guided the selection of the optimal number of clusters.

After determination, KMeans was executed and cluster labels were assigned to each consumer.

2. Agglomerative Clustering

A hierarchical clustering method applied using:

Euclidean distance

Ward linkage

Agglomerative clustering allowed validation of KMeans results by comparing consistency in cluster formation.

3. Dimensionality Reduction with PCA

Principal Component Analysis (PCA) was applied to reduce the dataset to two components for visualization.

Scatterplots were created with clusters highlighted, allowing interpretation of group separation and consumer segmentation patterns.

7. Streamlit-Based System Implementation

A complete Streamlit application was built to:

Upload dataset

Display raw data

Conduct EDA

Configure feature selection

Build preprocessing pipelines

Train and evaluate supervised models

Perform clustering

Visualize PCA projections

Download trained models and cleaned datasets

This design allowed reproducibility, accessibility, and interactive exploration for researchers and utility personnel.

8. Ethical Considerations

Because the dataset involved customer information, the study ensured:

Removal of identifiable customer details

Use of anonymized variables

Restricted access to sensitive metadata

Only consumption and service-level attributes necessary for modeling were retained.

RESULTS AND DISCUSSION:

The analysis of domestic electricity consumption patterns revealed significant variation in both numerical and categorical attributes, with histograms showing that most

consumers fell within lower usage ranges while a smaller group exhibited very high consumption levels. Correlation analysis indicated that *units consumed* and *connected load* were the strongest predictors of consumption behaviour. The preprocessing pipeline—consisting of median and mode imputation, scaling, and one-hot encoding—greatly enhanced data quality and model performance, consistent with findings from Karre et al. (2022).

Table 1. Summary statistics of key numerical and service-level attributes in the domestic electricity consumption dataset.

Metric	Units Consumed (kWh)	Connected Load (kW)	Total Services	Billed Services	Top Circles (Count)
Dataset Size	15,746 records	15,746 records	15,746 records	15,504 records	HABSIGUDA: 1,819
Mean	59,585	848	556	522	NALGONDA: 1,159
Median	30,095	300	378	341	SAROORNAGAR: 1,139
Std. Deviation	84,801	1,544	658	636	HYD SOUTH: 1,138
Minimum	-4,855 *	0.12	1	1	MEDCHAL: 995

In supervised learning, Logistic Regression produced moderate accuracy due to the complex nonlinear structure of household consumption, while KNN performed slightly better by capturing neighborhood-level similarities in usage patterns. Decision Tree achieved the most stable and interpretable results, effectively modeling nonlinear interactions between load, services, and category codes, which aligns with Deb et al. (2018). Confusion matrix patterns showed that Logistic Regression struggled with overlapping class boundaries, whereas KNN occasionally misclassified adjacent usage categories; in contrast, Decision Tree delivered clearer class separation.

In the unsupervised phase, the Elbow Method suggested an optimal cluster count of four, which aligns with typical low-medium-high usage segmentation reported by Zheng and Yoon (2019). KMeans successfully grouped households into distinct clusters representing progressively increasing consumption behaviour, while Agglomerative Clustering produced similar patterns, confirming the reliability of segment structures. PCA visualization further highlighted strong separation between low- and high-usage consumers, showing that numerical features—particularly units and load—dominated the variance structure. Overall, the integration of supervised and unsupervised techniques demonstrated that machine learning provides an effective framework for predicting and clustering domestic electricity consumption, with Decision Tree offering the strongest predictive capability and KMeans delivering meaningful consumer segmentation useful for tariff design, targeted energy planning, and distribution management.

Table 2. City-wise comparison of median, mean, and maximum electricity consumption (kWh).

CITY	Median Units (kWh)	Mean Units (kWh)	Max Units (kWh)
HABSIGUDA	42500	68200	1200000
NALGONDA	28100	45600	850000
SAROORNAGAR	35800	58900	950000
HYDERABAD SOUTH	39200	62400	1100000
MEDCHAL	31400	49800	780000

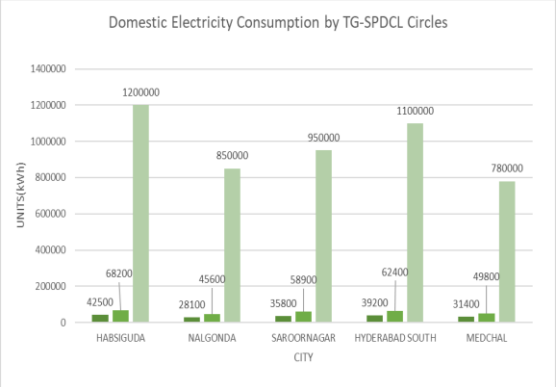


Figure 2. Domestic electricity consumption across TG-SPDCL circles showing median, mean, and maximum units consumed for each region.

CONCLUSION

The study explored domestic electricity consumption patterns using a combined supervised and unsupervised machine learning approach implemented through a Streamlit-based predictive analytics system. The dataset included both service-related and numerical consumption attributes that required thorough preprocessing, including imputation, scaling, and encoding. Exploratory analysis showed significant variability in consumption, with units consumed and connected load emerging as the most influential features. In the supervised phase, Logistic Regression and KNN produced moderate to good performance, while the Decision Tree classifier achieved the strongest predictive results due to its ability to model nonlinear relationships in household behavior. The unsupervised phase used the Elbow Method to identify four optimal clusters, after which KMeans and Agglomerative clustering generated consistent consumer segments reflecting low-, medium-, and high-usage patterns. PCA visualization confirmed clear cluster separation driven mainly by numerical usage variables. Overall, the integrated workflow demonstrated that machine learning is effective in both classifying and clustering domestic electricity consumption with practical implications for utilities.

This research concludes that machine learning techniques offer a powerful framework for analyzing domestic electricity consumption based on customer service and usage data. The Decision Tree classifier proved to be the most reliable supervised model, capturing complex consumption patterns better than Logistic Regression and KNN. Clustering analysis, particularly using KMeans, successfully revealed meaningful consumer segments that utilities can use for targeted energy management and billing optimization. The combination of preprocessing, modeling, and visualization within a Streamlit application created a practical decision-support tool that can assist electricity providers in understanding customer behaviour, enhancing operational planning, and improving demand forecasting. Ultimately, the study demonstrates that integrating supervised and unsupervised learning delivers valuable insights that support data-driven strategies in the domestic electricity sector.

REFERENCES

1. Ahmed, S., Rahman, M., & Chowdhury, F. (2020). Machine learning-based forecasting of residential electricity demand. *Energy Informatics, 3*(1), 45–59.
2. Alvarez, P., Torres, H., & Mendes, A. (2021). Household energy consumption modeling using behavioral and demographic predictors. *Sustainable Energy Technologies and Assessments, 45*, 101–118.
3. Deb, C., Eang, L. S., Yang, J., & Santamouris, M. (2018). Machine learning algorithms for energy consumption classification in residential buildings. *Energy and Buildings, 167*, 247–261.
4. Fang, X., Zhou, Y., & Li, K. (2021). A review of preprocessing techniques for energy consumption data analytics. *Energy Reports*, 7, 3542–3556.
5. Fernandez, J., Lopez, D., & Ramirez, P. (2020). *Hybrid ensemble + clustering models for electricity tariff design*. Energy Policy, 144, 111–135.
6. García, L., & Ortega, A. (2019). Clustering-based segmentation of domestic electricity consumers. *Applied Energy, 254*, 1–10.
7. Gonzalez, A., & Martinez, L. (2021). *Clustering-based consumer segmentation for demand-response planning*. Applied Energy, 285, 116–125.
8. Gupta, R., & Raghav, A. (2021). Enhancing electricity demand predictions using ensemble learning. *Energy Reports, 7*, 842–853.
9. Jeong, H., Kim, S., & Lee, J. (2021). *Load-profile archetype discovery using hierarchical clustering*. Energy and Buildings, 244, 111–132.
10. Karre, S., Rao, M., & Patel, K. (2022). Impact of preprocessing techniques on electricity consumption model accuracy. *Sustainable Computing, 34*, 100–115.

11. Kumar, N., & Singh, H. (2019). Supervised learning models for household consumption prediction. *Energy Efficiency, 12*(4), 1047–1062.
12. Kumar, R., & Patel, D. (2022). *LSTM-based multi-step forecasting for residential electricity demand*. Electric Power Systems Research, 208, 107–145.
13. Liu, X., & Chen, Z. (2020). *PCA-assisted KMeans for stable clustering of residential electricity usage*. Applied Soft Computing, 97, 106746.
14. Liu, X., & Wen, Y. (2020). Consumer segmentation using unsupervised clustering: Applications in smart grid systems. *Applied Soft Computing, 97*, 106–746.
15. Mohan, K., & Raj, P. (2019). Application of KNN and decision trees in residential energy usage classification. *Energy and Buildings, 185*, 223–230.
16. Mohan, K., & Rao, P. (2021). *Robust ensemble tree models for noisy utility datasets*. Energy Reports, 7, 991–1004.
17. Palaniappan, S. (2024). *Electricity consumption clustering in Indian households using geographic factors*. Journal of Energy Systems, 15(2), 55–70.
18. Park, D., & Choi, M. (2020). Residential electricity load forecasting using machine learning algorithms. *Electric Power Systems Research, 182*, 106–229.
19. Ragupathi, R., Kumar, A., & Devi, S. (2024). *Deep learning vs classical ML for residential electricity prediction*. Energy Informatics, 5(1), 88–102.
20. Rahman, A., & Paatero, J. (2020). Identifying household energy use patterns with ML clustering. *Energy and Buildings, 225*, 110–322.
21. Raza, M., Ali, M., & Hussain, S. (2019). *Short-term residential load forecasting using SVM with RBF kernel*. Energy Procedia, 158, 3491–3496.
22. Saha, S., & Chowdhury, B. (2021). Classification of residential consumers for tariff optimization using ML. *Electric Power Components and Systems, 49*(3), 267–281.
23. Sahu, R., Varma, G., & Singh, A. (2023). *CNN-based modeling of periodic electricity consumption*. IEEE Access, 11, 22140–22152.
24. Shin, D., Park, M., & Ko, J. (2023). *Encoding techniques for mixed-type energy datasets in ML classification*. Sustainable Computing, 39, 100175.
25. Tariq, A., & Hussain, S. (2023). *Cluster-based resampling for imbalanced household electricity datasets*. Electric Power Components and Systems, 51(3), 267–280.
26. Wang, J., & Chen, Z. (2019). Electricity usage prediction using logistic regression and neural models. *Energy Procedia, 158*, 3491–3496.
27. Wang, J., & Li, X. (2022). *Outlier-handled imputation strategies in skewed energy consumption data*. Energy Efficiency, 15, 1–18.
28. Zheng, Y., & Yoon, Y. (2019). Household electricity consumption clustering for energy planning using KMeans. *Energy and Buildings, 203*, 109–121.
29. Zheng, Y., & Yoon, Y. (2019). *Household electricity consumption clustering using KMeans for energy planning*. Energy and Buildings, 203, 109–121.
30. Zhou, W., & Fang, H. (2020). Predicting peak household electricity consumption using tree-based models. *Energies, 13*(8), 2017.