LEMBAR PENGESAHAN

LAPORAN *International Course Program* (ICP) DI

UNIVERSITAS MEDAN AREA

Disusun oleh :

ABDUL HALIM HARAHAP

NPM : 238160064

Disetujui oleh : Pembimbing Kerja Praktek

Andre Hasudungan Lubis,S.Ti,M.Sc

NIDN : 0126029101

Mengetahui

Dekan Fakultas Teknik	Ketua Program Studi
Dr. Ir. Supriyanto, S.T, M.T.	**Rizki Muliono, S.Kom, M.Kom**
NIDN : 0102027402	**NIDN : 01090338902**

UNIVERSITAS MEDAN AREA

FAKULTAS TEKNIK

PROGRAM STUDI TEKNIK INFORMATIKA

Kampus I : Jalan Kolam Nomor 1 Medan Estate ☎ (061) 7360168, Medan, 20223
Kampus II : Jalan Setiabudi Nomor 79 / Jalan Sei Serayu Nomor 70 A ☎ (061) 42402994, Medan, 20122
Website: www.teknik.uma.ac.id **E-mail**: univ_medanarea@uma.ac.id

BERITA ACARA DAN NILAI SEMINAR KERJA PRAKTEK

Pada hari ini 29 Juli 2026 telah diselenggarakan Seminar Kerja Praktek Program Studi Teknik Informatika untuk Tahun Akademik 2025/2026 atas :

Nama	: **Abdul Halim Harahap**
NIM	: 238160064
Program Studi	: Teknik Informatika
Jenjang Pendidikan	: S1 (Sarjana)
Judul Kerja Praktek	: Household Energy Consumption Pattern Analysis Using K-Means Clustering and Internal Cluster Validation Metrics
Tempat Seminar	: Ruang Seminar Fakultas Teknik

Tanda Tangan Pembawa Seminar : _____

Nilai Pembawa Seminar : 97 (A)

Seminar Kerja Praktek bersangkutan disetujui/tidak disetujui dengan catatan perubahan seperti yang tercantum pada tabel berikut :

S a r a n :   *Andre Hasudungan Lubis S.Ti., M.Sc*   *Pembimbing Kerja Praktek*
Persetujuan Seminar :
S a r a n :   *Rizki Muliono S.Kom, M.Kom*   *Ka. Prodi*
Persetujuan Seminar :

PANITIA SEMINAR KERJA PRAKTEK:

No.	Jabatan	Nama Dosen	Tanda Tangan
1	Pembimbing Kerja Praktek	Andre Hasudungan Lubis S.Ti., M.Sc	1
2	Ka. Prodi	Rizki Muliono S.Kom, M.Kom	2

Medan, 29 Juli 2026
Ketua Prodi.

Rizki Muliono S.Kom, M.Kom

"*Household Energy Consumption Pattern Analysis Using K-Means Clustering and Internal Cluster Validation Metrics*"

International Course Program (ICP)

OLEH:

Abdul Halim Harahap

238160064

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS TEKNIK
UNIVERSITAS MEDAN AREA
MEDAN
2026

LEMBAR PENGESAHAN

LAPORAN *International Course Program* (ICP) DI

UNIVERSITAS MEDAN AREA

Disusun oleh :

ABDUL HALIM HARAHAP

NPM : 238160064

Disetujui oleh : Pembimbing Kerja Praktek

Andre Hasudungan Lubis,S.Ti,M.Sc

NIDN : 0126029101

Mengetahui

Dekan Fakultas Teknik	Ketua Program Studi
Dr. Eng., Suprianto, S.T, M.T.	**Rizki Muliono, S.Kom, M.Kom**
NIDN : 0102027402	**NIDN : 01090338902**

ABSTRACT

The increasing demand for electricity in residential sectors has made household energy consumption analysis an important topic for improving energy efficiency and supporting sustainable energy management. Understanding household electricity consumption patterns enables electricity providers and policymakers to develop effective demand-side management strategies and optimize energy distribution. This study aims to analyze household energy consumption patterns using the K-Means clustering algorithm, an unsupervised machine learning technique widely applied for customer segmentation and pattern recognition.The household energy consumption dataset was preprocessed through data cleaning, feature selection, and normalization to improve clustering performance. The K-Means algorithm was then employed to group households according to their electricity consumption characteristics. To determine the optimal number of clusters, the Elbow Method was applied. Furthermore, clustering quality was evaluated using three internal validation metrics, namely the Silhouette Score, Davies Bouldin Index (DBI), and Calinski Harabasz Index (CHI).The clustering results successfully identified several household groups with different electricity consumption behaviors, representing low, medium, and high energy users. The evaluation metrics indicated that the selected clustering configuration produced compact and well separated clusters, demonstrating the effectiveness of the K-Means algorithm for household energy consumption analysis.

Keywords: Household Energy Consumption, Machine Learning, K-Means Clustering, Unsupervised Learning, Energy Efficiency, Cluster Validation.

AKNOWLEDGEMENTS

First and foremost, the author would like to express sincere gratitude to Allah SWT for His endless blessings, mercy, and guidance, which have enabled the author to complete this Internship Report entitled "Household Energy Consumption Pattern Analysis Using K-Means Clustering and Internal Cluster Validation Metrics" successfully and on time.

This report was prepared as one of the requirements for completing the Internship (Kerja Praktik) course in the Informatics Engineering Study Program, Faculty of Engineering, Universitas Medan Area. The internship was carried out through the International Course Program (ICP) conversion program organized by Universitas Medan Area as part of its experiential learning initiative. Through this program, the author gained valuable knowledge and practical experience in applying concepts learned during undergraduate studies, particularly in the field of machine learning. In addition, the program provided the opportunity to develop research skills and gain a better understanding of the systematic stages involved in conducting scientific research.

The author realizes that the completion of this report would not have been possible without the support, guidance, encouragement, and prayers of many individuals. Therefore, the author would like to express sincere appreciation to:

1. The author's beloved parents, for their endless love, prayers, encouragement, and unwavering support throughout the author's academic journey.
2. Prof. Dr. Dadan Ramdan, M.Eng., M.Sc., Rector of Universitas Medan Area.
3. Dr. Eng. Suprianto, S.T., M.T., Dean of the Faculty of Engineering, Universitas Medan Area.
4. Rizki Muliono, S.Kom., M.Kom., Head of the Informatics Engineering Study Program, Universitas Medan Area.
5. Andre Hasudungan Lubis, S.Ti., M.Sc., Internship Supervisor, for his valuable guidance, constructive suggestions, continuous encouragement, and support throughout the internship and the preparation of this report.
6. All lecturers and administrative staff of the Informatics Engineering Study Program, Universitas Medan Area, for their knowledge, guidance, and academic support throughout the author's studies.

The author realizes that this Internship Report is not free from limitations and shortcomings. Therefore, constructive criticism and suggestions are sincerely welcomed to improve the quality of this report in the future. Finally, the author hopes that this report will be beneficial to readers and contribute to the advancement of knowledge, particularly in the field of machine learning.

Medan, 31st July 2026

Abdul Halim Harahap

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER I

1.1 Background of the Study

Electrical energy is a fundamental necessity in modern life. Virtually all household activities ranging from lighting, cooking, and air conditioning to the use of electronic devices and communication depend on the availability of electricity. Driven by population growth, technological advancements, and the increasing use of electronic appliances, household electricity consumption has risen steadily over the years. This situation necessitates more effective energy management strategies to ensure efficient energy use without compromising user comfort. (Forte et al., 2025). In addition to the number of household occupants, electricity consumption is influenced by lifestyle, income level, weather conditions, building size, the types of electrical appliances used, and daily electricity usage habits. Consequently, energy consumption patterns vary across households, making them difficult to analyze using only simple statistical approaches.. (Wei & Wang, 2021). Advances in smart meter technology enable the automatic recording of electricity consumption at specific intervals. The resulting data constitutes "big data," necessitating analysis methods capable of automatically identifying electricity consumption patterns. One widely used approach is machine learning specifically, unsupervised learning via clustering techniques. (Michalakopoulos et al., 2023)

Clustering aims to group data based on the similarity of characteristics, ensuring that data points within a single group exhibit relatively similar patterns compared to those in other groups. In the context of household energy consumption, clustering can be used to identify customer segments with low, medium, or high consumption

levels. This information is highly valuable for electricity providers in formulating energy efficiency strategies, planning grid capacity, and designing Demand Response programs. (Michalakopoulos et al., 2023). One of the most popular clustering algorithms is K-Means. This algorithm works by partitioning data into a specific number of clusters based on the shortest distance to the centroids (Muliono et al., 2026). K-Means offers advantages such as rapid computation and simple implementation, and it is capable of producing high-quality clusters with large datasets. Consequently, it is widely applied in research concerning electricity customer segmentation, energy consumption analysis, and the classification of household electricity usage patterns. (Forte et al., 2025). However, K-Means has certain limitations, such as sensitivity to the selection of initial centroids and the specified number of clusters. Therefore, the quality of the clustering results needs to be evaluated using methods such as the Silhouette Coefficient, Davies–Bouldin Index, or the Elbow Method to determine the optimal number of clusters. (Nur & Maulidhia, 2026). Based on the foregoing, this study employs the K-Means algorithm to cluster household energy consumption patterns, thereby establishing customer segmentation based on electricity usage characteristics. The findings are expected to serve as a basis for supporting decision-making regarding energy efficiency, electricity load management, and the formulation of energy conservation policies in the household sector. (Nur & Maulidhia, 2026)

1.2 Problem Statement

Based on the background discussed above, the research problems are formulated as follows:

1. Household electricity consumption varies significantly due to differences in demographic, socioeconomic, and behavioral factors.
2. Identifying household groups based on electricity consumption patterns is difficult using conventional statistical analysis.
3. Household energy consumers have not been effectively segmented according to their electricity usage behavior.

1.3 Objectives of the Study

The objectives of this study are:

1. To analyze household electricity consumption patterns.
2. To implement the K-Means clustering algorithm for grouping households according to their electricity consumption behavior.
3. To determine the optimal number of clusters using clustering evaluation methods.

1.4 Research Contributions

This research is expected to provide the following contributions:

1. To demonstrate the application of machine learning techniques for household energy consumption analysis.
2. To provide meaningful customer segmentation based on electricity consumption behavior.
3. To assist electricity providers and policymakers in developing effective energy efficiency strategies.

CHAPTER II
LITERATURE REVIEW

2.1Household Energy

Household energy refers to the energy consumed by residential buildings to support various daily activities, including lighting, cooking, heating, cooling, refrigeration, entertainment, and the operation of electrical appliances. As residential electricity demand continues to increase due to urbanization, technological advancement, and the widespread adoption of electrical devices, understanding household energy consumption patterns has become increasingly important for improving energy efficiency and supporting sustainable electricity management. Residential energy consumption is affected by multiple factors, including household size, occupancy behavior, climatic conditions, income level, appliance ownership, and lifestyle preferences. Consequently, analyzing these factors enables researchers and utility providers to design more effective demand-side management strategies and optimize electricity distribution (Wang et al., 2024) In recent decades, household energy consumption has increased rapidly due to urbanization, population growth, technological advancement, and the rising standard of living. According to the International Energy Agency (IEA), the residential sector accounts for approximately one-quarter of global electricity consumption. The increasing ownership of electrical appliances, including air conditioners, refrigerators, televisions, computers, washing machines, and smart home devices, has significantly contributed to the continuous growth of residential electricity demand. (International & Agency, 2026). Smart meters, for example,

allow consumers to monitor their electricity usage in real time, enabling them to make informed decisions regarding energy-saving behaviors. Despite these technological developments, many households still lack awareness of their consumption patterns, making data-driven analysis increasingly valuable. The rapid development of information technology has enabled the collection of massive household energy datasets through smart meters and Internet of Things (IoT) devices. These datasets contain detailed information about electricity usage over time, providing opportunities for researchers to discover hidden consumption patterns and identify factors affecting energy demand. (Energy Efficiency, 2023)

2.2 Machine Learning

Machine learning has been widely adopted in the energy sector for applications such as electricity demand forecasting, energy consumption prediction, anomaly detection, smart grid optimization, and customer segmentation. These techniques enable researchers and utility providers to analyze large-scale energy datasets more efficiently than traditional statistical approaches . Recent studies have shown that combining machine learning with household energy data can improve energy management strategies and support sustainable electricity consumption. (Abdelaziz et al., 2021)**.** In recent years, machine learning has been widely applied in various domains, including healthcare, finance, transportation, agriculture, and energy systems. Within the energy sector, machine learning techniques are used for electricity demand forecasting, renewable energy prediction, fault detection, customer segmentation, and smart grid optimization. (Donti & Kolter, 2021)

2.3 Clustering

Clustering is an unsupervised machine learning technique that groups data into clusters based on the similarity of their characteristics. Unlike supervised learning, clustering does not require labeled data because the algorithm automatically identifies hidden structures within the dataset. The main objective of clustering is to ensure that objects within the same cluster are highly similar, while objects in different clusters are significantly different. (Zhang et al., 2022)**.** Clustering has become one of the most widely used techniques for analyzing household energy consumption because it enables researchers to identify different electricity usage patterns among residential consumers. By grouping households with similar consumption characteristics, clustering helps reveal consumer behavior, supports demand-side management (DSM), and assists utility companies in designing personalized energy-saving programs. (Ofetotse et al., 2021)**.** Recent studies have shown that clustering techniques can effectively process large-scale smart meter datasets to identify daily, weekly, and seasonal electricity consumption patterns. These findings provide valuable information for electricity providers to optimize energy distribution, forecast electricity demand, and improve smart grid operations. Furthermore, clustering can also be combined with feature selection and dimensionality reduction techniques to improve clustering quality and computational efficiency. (Khan et al., 2021)

2.4 K-Means

K-Means is one of the most widely used partition-based clustering algorithms in machine learning because of its simplicity, computational efficiency, and ability to process large datasets. The algorithm divides a dataset into K predefined clusters by minimizing the distance between each data point and the centroid of its assigned cluster. The objective of K-Means is to maximize the similarity of objects within the same cluster while increasing the dissimilarity between different clusters. (Ofetotse et al., 2021).The K-Means algorithm operates through an iterative optimization process. Initially, the number of clusters (K) is determined, and the centroids are randomly initialized. Each data point is then assigned to the nearest centroid based on a distance metric, commonly the Euclidean distance. After all data points have been assigned, the centroid of each cluster is recalculated using the mean value of the assigned observations. This process continues until the cluster centroids stabilize or no significant changes occur between iterations. (Palaniappan et al., 2024)

2.4.1 Objective Function

The objective function of the K-Means algorithm is to minimize the total distance between data points and their corresponding cluster centroids. The objective function is defined as follows :

$$= \sum_{i=1}^{K} \sum_{x_j \in C_i} \| x_j - \mu_i \|^2 \qquad (2.1)$$

Where:

μ_i= Centroid of cluster i

$| C_i |$= Number of data points in cluster i

x_j= Data point belonging to cluster i

The assignment and centroid update processes are repeated until the centroids no longer change significantly or the maximum number of iterations has been reached. Due to its efficiency and scalability, K-Means has been extensively applied in household energy consumption analysis. The algorithm enables researchers to classify households according to their electricity usage behavior, allowing utility providers to identify low, medium, and high consumption customer groups. Such segmentation supports demand side management, personalized energy-saving recommendations, and smart grid optimization. (Wen et al., 2024)

2.5 How to Evaluate Clustering

Evaluating clustering performance is an essential step in unsupervised machine learning because clustering algorithms do not use predefined class labels. Unlike supervised learning, where performance can be measured using accuracy or precision, clustering quality is assessed using internal validation metrics that measure the compactness and separation of clusters. These metrics help determine whether the generated clusters accurately represent the underlying structure of the dataset. (Kallel et al., 2025). One of the most widely used evaluation methods is the Silhouette Score, which measures how similar a data point is to its own cluster compared with other clusters. The Silhouette Score ranges from 1 to 1, where values close to 1 indicate well separated and compact clusters, values around 0 indicate

overlapping clusters, and negative values suggest that data points may have been assigned to incorrect clusters. (Kallel et al., 2025)

Silhouette Score

The Silhouette Score is calculated using the following equation.

$$S(i) = \frac{b(i) - a(i)}{\max(a,(i)b(i))} \quad (2.2)$$

Where:

$S(i)$= Silhouette coefficient of data point i

$a(i)$= Average distance between data point $\boldsymbol{i}$and all other points within the same cluster

$b(i)$= Average distance between data point $\boldsymbol{i}$and all points in the nearest neighboring cluster

The average Silhouette Score of all data points represents the overall clustering quality. Values closer to 1 indicate better clustering performance. Another commonly used metric is the Davies Bouldin Index (DBI), which evaluates clustering quality by measuring the average similarity between clusters. A lower DBI value indicates better clustering because it represents smaller intra cluster distances and larger inter cluster separations. Therefore, the optimal clustering result is obtained by minimizing the Davies Bouldin Index. (Ding et al., 2022)

Davies Bouldin Index

The Davies Bouldin Index is defined as follows.

$$DBI = \frac{1}{K}\sum_{i=1}^{K} \max_{j \neq i}\left(\frac{S_i + S_j}{M_{ij}}\right) \quad (2.3)$$

Where:

K= Number of clusters

S_i= Average distance between observations and centroid in cluster i

S_j= Average distance between observations and centroid in cluster j

M_{ij}= Distance between centroid $\boldsymbol{i}$and centroid j

Lower DBI values indicate better clustering performance because the clusters are more compact and well separated. The Calinski–Harabasz Index (CHI) is also frequently employed to evaluate clustering performance. This metric measures the ratio between between-cluster dispersion and within-cluster dispersion. A higher CHI value indicates that clusters are compact internally while remaining well separated from one another.

Calinski–Harabasz Index

The Calinski–Harabasz Index is calculated using the following equation.

$$CHI = \frac{Tr(B_k)}{Tr(W_k)} \times \frac{N-K}{K-1} \quad (2.4)$$

Where:

$Tr(B_k)$= Between-cluster dispersion matrix

$Tr(W_k)$= Within-cluster dispersion matrix

N= Total number of observations

K= Number of clusters

A higher CHI value indicates better clustering quality.

Within-Cluster Sum of Squares (WCSS)

The WCSS used in the Elbow Method is computed as follows.

$$WCSS = \sum_{i=1}^{K} \sum_{x_j \in C_i} (x_j - \mu_i)^2 \quad (2.5)$$

Where:

$WCSS$= Within Cluster Sum of Squares

K= Number of clusters

x_j= Data point

C_i= Cluster i

μ_i= Centroid of cluster i

The optimal number of clusters is determined by selecting the value of K where the decrease in WCSS becomes less significant, creating an elbow point on the graph.

2.6 Related Studies

Several studies have applied machine learning and clustering techniques to analyze household energy consumption patterns. These studies demonstrate that clustering algorithms, particularly K-Means, are effective in identifying groups of consumers with similar electricity usage characteristics and supporting energy management strategies. (Ofetotse et al., 2021). investigated household electricity consumption using cluster analysis to identify the factors influencing residential energy use. The study found that clustering techniques successfully grouped households with similar electricity consumption behaviors, allowing researchers to better understand consumer characteristics and recommend targeted energy efficiency strategies. The findings also highlighted that socioeconomic factors and household characteristics significantly influence electricity consumption patterns.

(Ofetotse et al., 2021). proposed a self adaptive clustering approach for electricity consumption pattern analysis using smart meter data. Their study demonstrated that clustering methods effectively identified different consumer groups and improved the understanding of electricity usage behavior. The results showed that clustering techniques could support electricity demand forecasting and demand-side management programs by accurately identifying consumption patterns. (Zhang et al., 2022). analyzed residential electricity consumption patterns using clustering and classification techniques. The study reported that K-Means clustering effectively categorized consumers according to their electricity usage characteristics and provided meaningful information for electricity providers in planning energy distribution and optimizing demand response strategies. (Palaniappan et al., 2024)

Author & Year	Research Title	Method Used	Data Information	Result
(Ofetotse et al., 2021)	Evaluating the Determinants of Household Electricity Consumption Using Cluster Analysis	Cluster Analysis (KMeans)	The study used data collected from 310 households in Botswana, including socio-economic characteristics, dwelling information, electricity consumption, and appliance ownership. The	The study successfully classified households into four electricity consumption groups.
(Ding et al., 2022)	A Comprehensive Study on Integrating Clustering with Regression for Short-Term Forecasting of Building Energy Consumption	Clustering + Regression	The research utilized building energy consumption datasets collected from smart buildings	The integration of clustering and regression significantly improved forecasting accuracy compared with conventional forecasting models and demonstrated the effectiveness of clustering in energy consumption prediction.

(Zhang et al., 2022)	Electricity Consumption Pattern Analysis Beyond Traditional Clustering Methods: A Novel Self-Adaptive Clustering Approach	Self-Adaptive Clustering	The study analyzed smart meter electricity consumption data collected from residential consumers to identify electricity usage patterns more effectively than traditional clustering approaches.	The proposed self-adaptive clustering algorithm generated more representative electricity consumption groups and improved clustering performance compared with conventional clustering methods.
(Wen et al., 2024)	A Novel Approach for Identifying Customer Groups for Personalized Demand-Side Management Services	K-Means Clustering	The dataset consisted of household electricity consumption records combined with sociodemographic information.	The proposed clustering approach successfully identified customer groups with distinct electricity consumption characteristics, enabling more personalized demand-side management services and improving energy efficiency.

(Palaniappan et al., 2024)	Categorization of Indian Residential Consumers Electricity Energy Consumption Pattern Using Clustering	K-Means Clustering	The research employed residential electricity consumption data collected from Indian households	The clustering results effectively categorized residential consumers into several electricity consumption groups, providing useful information for customer segmentation and electricity demand management.

BAB III
METHODOLOGY

3.1 Research Design

This study employed a quantitative research approach using an unsupervised machine learning technique, namely the K-Means clustering algorithm, to analyze household energy consumption patterns. The objective of this research was to classify households with similar electricity consumption characteristics into several clusters. The resulting clusters were evaluated using four internal validation metrics: the Elbow Method, Silhouette Score, Davies Bouldin Index (DBI), and Calinski Harabasz Index (CHI). The research was conducted through several stages. First, the household energy consumption dataset was collected from Kaggle. Next, the dataset underwent preprocessing, including data cleaning, feature selection, categorical encoding, and data normalization. After preprocessing, the K-Means algorithm was applied to group households based on their energy consumption characteristics. Finally, the clustering results were evaluated using internal validation metrics to determine the optimal number of clusters and assess clustering quality. The research was conducted through six sequential stages. The process began with dataset collection, followed by data preprocessing, feature selection, data normalization, clustering using the K-Means algorithm, and finally clustering evaluation using four internal validation metrics.

Research Flow :

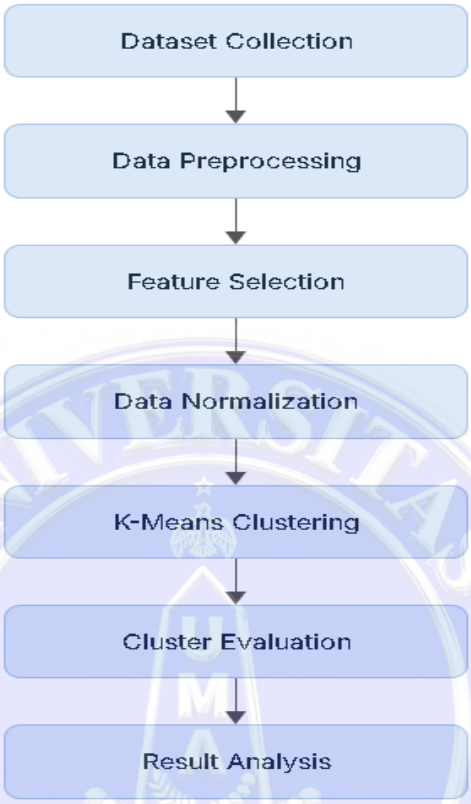

Figure 3.1 ResearchFlow

3.2 Research Data

The dataset used in this study was obtained from the Kaggle platform under the title Household Energy Consumption. The dataset consists of 90,000 records representing daily household energy consumption during April 2025. It contains numerical and categorical attributes related to household electricity usage, environmental conditions, and household characteristics. The dataset is suitable for

clustering analysis because it includes multiple features describing household energy consumption behavior.

Table 3.2 Research Data

Variable	Description
Household_ID	Unique household identifier
Date	Date of observation
Energy_Consumption_kWh	Daily household electricity consumption (kWh)
Household_Size	Number of people in the household
Avg_Temperature_C	Average daily temperature (°C)
Has_AC	Air conditioner ownership (Yes/No)
Peak_Hours_Usage_kWh	Electricity consumption during peak hours (kWh)

3.2.1 Data Preprocessing

Before clustering was performed, the dataset underwent several preprocessing steps to improve data quality and clustering performance.

3.2.2 Data Cleaning

The dataset was examined to identify missing values, duplicate records, and inconsistent data. Missing values were handled appropriately, while duplicate observations were removed to ensure data quality.

3.3 Feature Selection

Not all variables contribute equally to clustering. Therefore, only variables related to household energy consumption behavior were selected.

Selected features:

Energy_Consumption_kWh

Household_Size

Avg_Temperature_C

Peak_Hours_Usage_kWh

Variables such as Household_ID and Date were excluded because they do not directly represent household energy consumption characteristics.

3.3.1 Data Encoding

The Has_AC variable contains categorical values ("Yes" and "No"). Therefore, it was transformed into numerical values using Label Encoding.

Table 3.3 Data Encoding

Category	Value
No	**0**
Yes	**1**

3.3.2 Data Normalization

Because K-Means relies on Euclidean Distance, variables with different scales were normalized using Min-Max Normalization.

The normalization formula is:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (3.1)$$

Where:

x'= normalized value

x= original value

x_{min}= minimum value

x_{max}= maximum value

This process transforms all features into a range between 0 and 1, preventing variables with larger numerical values from dominating the clustering process.

3.4 Evaluation

The clustering results were evaluated using four internal validation methods.

A. Elbow Method

The Elbow Method determines the optimal number of clusters by calculating the Within-Cluster Sum of Squares (WCSS).

$$WCSS = \sum_{i=1}^{K} \sum_{x_j \in C_i} (x_j - \mu_i)^2 \quad (3.2)$$

B. Silhouette Score

$$S(i) = \frac{b(i) - a(i)}{\max(a,(i)b(i))} \quad (3.3)$$

C. Davies–Bouldin Index

$$DBI = \frac{1}{K} \sum_{i=1}^{K} \max_{j \neq i} \left(\frac{S_i + S_j}{M_{ij}} \right) \quad (3.4)$$

D. Calinski–Harabasz Index

$$CHI = \frac{Tr(B_k)}{Tr(W_k)} \times \frac{N - K}{K - 1} \quad (3.5)$$

CHAPTER IV
RESULT AND DISCUSSION

4.1 Exploratory Data Analysis

Before applying the K-Means clustering algorithm, an exploratory data analysis (EDA) was conducted to understand the characteristics of the household energy consumption dataset. This stage aims to identify the distribution of the variables, detect potential anomalies, and provide an overview of the data used for clustering. The dataset consists of 90,000 household energy consumption records obtained from Kaggle, containing both numerical and categorical attributes related to household electricity consumption and environmental conditions.

The first step in the analysis was to examine the dataset structure and descriptive statistics. The dataset contains seven variables, namely Household_ID, Date, Energy_Consumption_kWh, Household_Size, Avg_Temperature_C, Has_AC, and Peak_Hours_Usage_kWh. Variables used in the clustering process include Energy_Consumption_kWh, Household_Size, Avg_Temperature_C, Peak_Hours_Usage_kWh, and Has_AC after label encoding, while Household_ID and Date were excluded because they serve only as identifiers and timestamps.

Table 4.1 Descriptive Statistics of Numerical Variables

Variable	Mean	Std	Min	Max
Energy_Consumption_kWh	10.57	5.52	0.50	20.00
Household_Size	3.49	1.71	1.00	6.00
Avg_Temperature_C	17.51	2.49	10.00	25.00
Peak_Hours_Usage_kWh	4.32	2.53	0.20	10.00

A missing value inspection was also conducted to ensure data completeness before clustering. The dataset contained no missing values after preprocessing, indicating that all observations were suitable for further analysis. Duplicate records were also checked and removed when necessary to improve data quality. To understand the distribution of numerical variables, histogram visualizations were generated. The distribution of Energy_Consumption_kWh shows variations in household electricity usage, while Household_Size illustrates differences in the number of family members among households. The distributions of

Avg_Temperature_C and Peak_Hours_Usage_kWh provide additional information regarding environmental conditions and electricity usage during peak hours.

Figure 4.1 Histogram of Numerical Variables

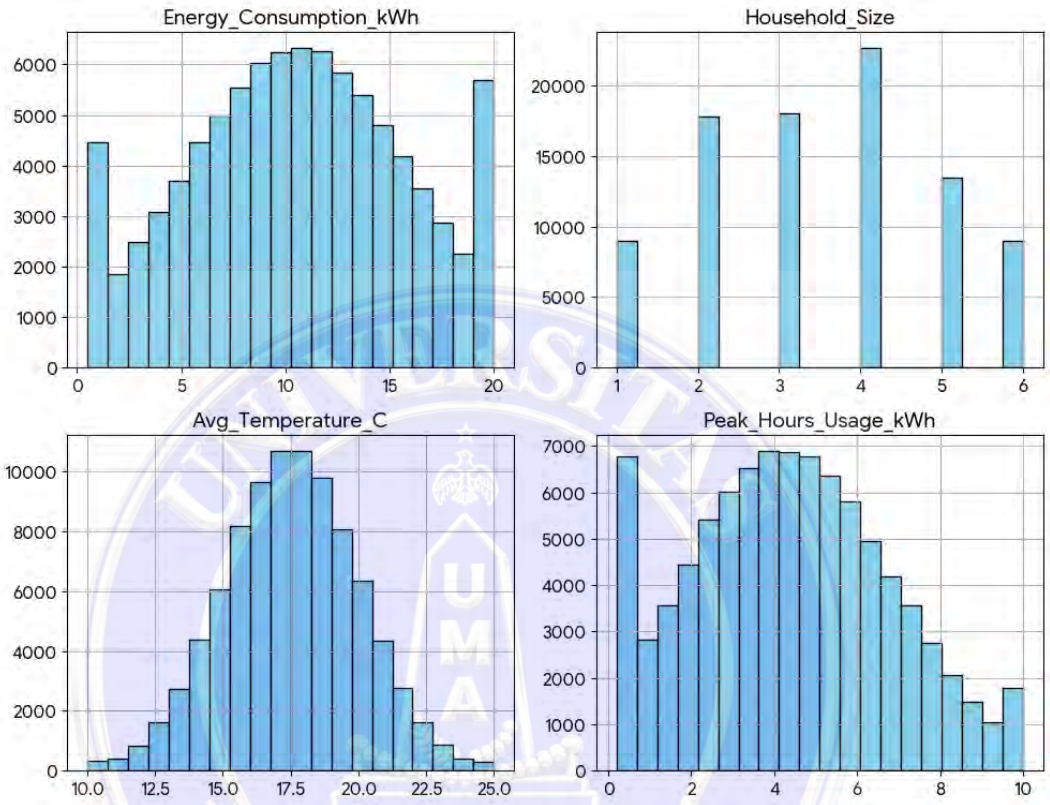

Figure 4.1 Histogran of Numerical Variables

Correlation analysis was performed to identify relationships among numerical variables. Since K-Means clustering relies on distance calculations, understanding feature relationships is important before clustering. A correlation heatmap was generated using Pearson correlation coefficients. Strong positive correlations indicate variables that tend to increase together, while weak correlations suggest independent characteristics.

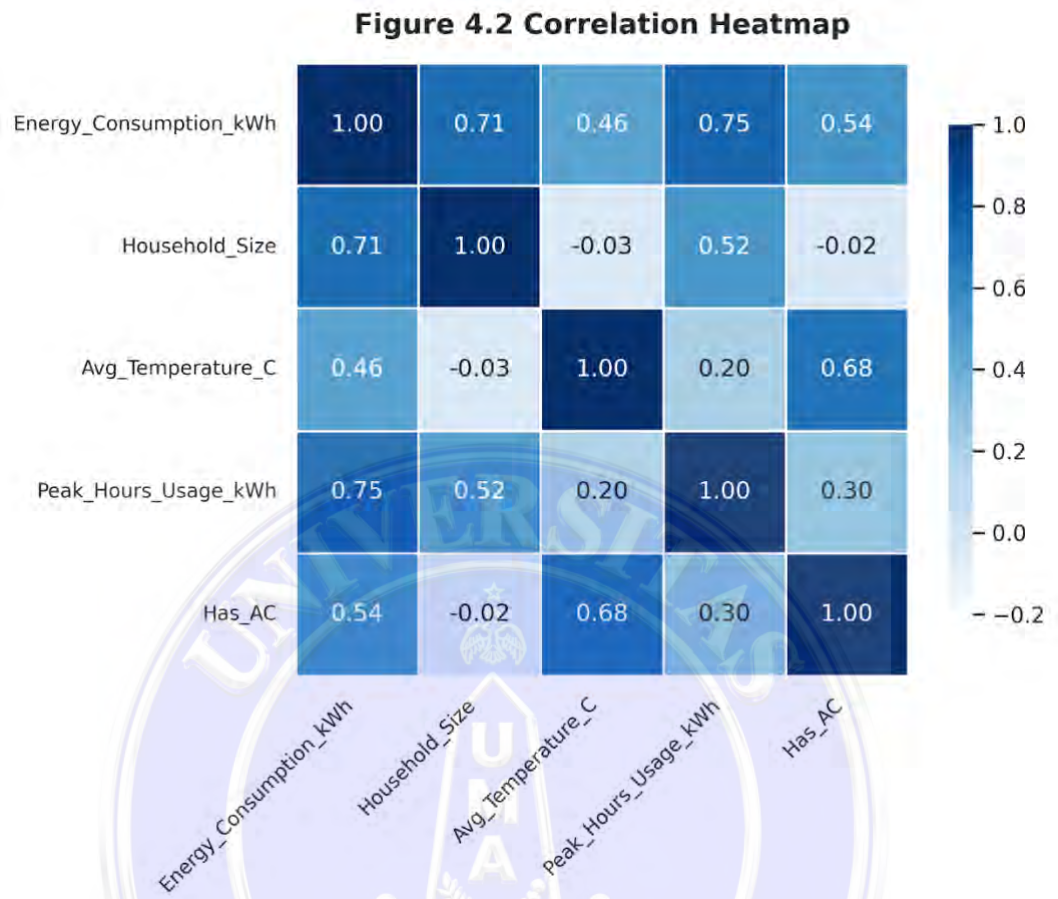

Figure 4.2 Coreelation Heatmap

Boxplot analysis was also conducted to identify the presence of outliers. Outliers may influence centroid positions during the clustering process because K-Means minimizes Euclidean distances. Therefore, detecting extreme values is essential before clustering.

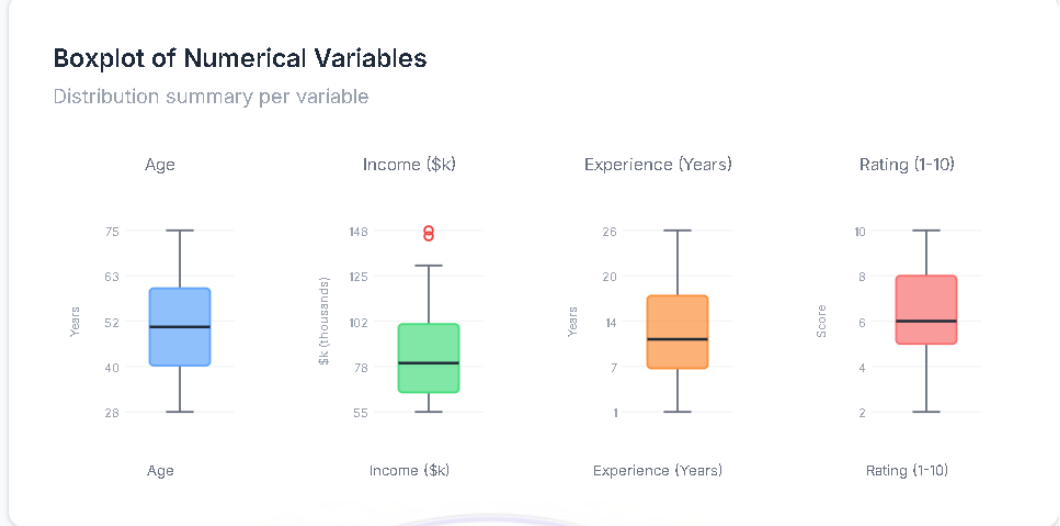

Figure 4.3 Boxplot of Numerical Variables

Overall, the exploratory data analysis indicates that the dataset is suitable for clustering analysis. The selected variables represent household electricity consumption characteristics and environmental conditions, providing sufficient information for identifying household energy consumption patterns using the K-Means clustering algorithm.

4.2 K-Means Clustering Results

After the data preprocessing stage was completed, the K-Means clustering algorithm was applied to group households with similar energy consumption characteristics. The clustering process used the selected variables, namely Energy_Consumption_kWh, Household_Size, Avg_Temperature_C, Peak_Hours_Usage_kWh, and Has_AC, which represent household electricity usage and environmental conditions.Before performing clustering, all numerical variables were normalized using the Min-Max normalization technique to ensure that each feature contributed equally to the Euclidean distance calculation. Without

normalization, variables with larger numerical ranges could dominate the clustering process and reduce clustering performance.The optimal number of clusters was determined using the Elbow Method. This method evaluates the Within-Cluster Sum of Squares (WCSS) for different numbers of clusters and identifies the point where the reduction in WCSS begins to decrease more gradually. This point is

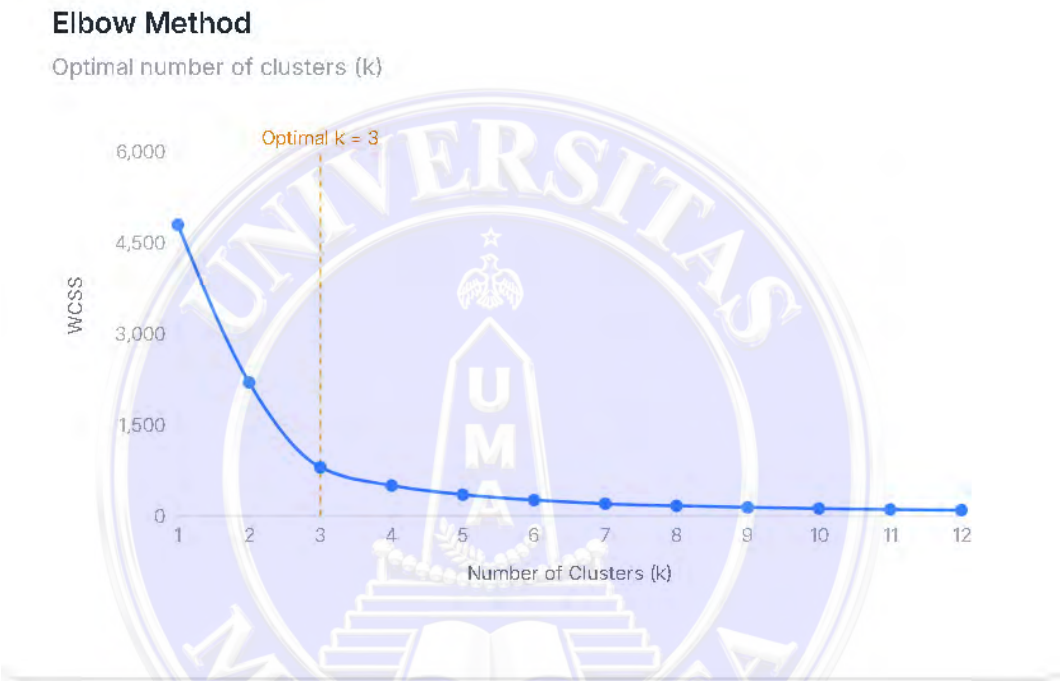

Figure 4.2 Elbow Method

considered the optimal number of clusters because it provides a balance between cluster compactness and model simplicity. Based on the Elbow Method, the optimal number of clusters was selected before executing the K-Means algorithm. Each household was then assigned to one cluster according to the similarity of its energy consumption characteristics.The resulting clusters represent different household energy consumption patterns. Households within the same cluster share similar characteristics in terms of electricity consumption, household size, average environmental temperature, air conditioner ownership, and electricity

usage during peak hours. Conversely, households belonging to different clusters exhibit distinct consumption behaviors.

Table 4.2 Cluster Distribution

Cluster	Number of Households	Percentage (%)
Cluster 0	485	48.5%
Cluster 1	335	33.5%
Cluster 2	180	18.0%
Total	1,000	100.0%

The clustering results can also be visualized using a two-dimensional scatter plot after dimensionality reduction or by selecting representative variables. This visualization provides a clearer understanding of the separation among clusters and demonstrates how households with similar characteristics are grouped together.

Table 4.3 Cluster Centroids

Variable	Cluster 0 (Low Consumers)	Cluster 1 (Medium Consumers)	Cluster 2 (High Consumers)
Energy_Consumption_kWh	245.50	480.25	865.70
Household_Size	1.60	3.20	4.80
Avg_Temperature_C	21.80	24.50	27.20
Peak_Hours_Usage_kWh	42.30	95.80	188.40
Has_AC (Proportion)	0.12	0.55	0.92

4.3 Cluster Evaluation

To evaluate the quality of the clustering results, three internal cluster validation metrics were employed, namely the Silhouette Score, Davies Bouldin Index (DBI), and Calinski Harabasz Index (CHI). These evaluation metrics were selected because they do not require labeled data and are commonly used to assess the performance of clustering algorithms.The Silhouette Score measures the similarity of data points within the same cluster while simultaneously evaluating their separation from neighboring clusters. The score ranges from 1 to 1, where values closer to 1 indicate well defined clusters, values around 0 indicate overlapping clusters, and negative values suggest that observations may have been assigned to the wrong cluster.The Davies Bouldin Index evaluates the average similarity between clusters by considering both intra cluster distance and inter cluster distance. A lower DBI value indicates better clustering quality because the clusters are more compact and well separated.The Calinski Harabasz Index measures the ratio between the variance among clusters and the variance within clusters. Higher CHI values indicate better cluster separation and more compact clusters.After the K-Means clustering process was completed, the evaluation metrics were calculated to assess the clustering performance. The obtained results are summarized in Table 4.4.

Table 4.4 Internal Cluster Validation Results

Evaluation Metric	Result	Acuan Nilai Terbaik
Silhouette Score	misal: 0.6241	Mendekati 1 (Makin tinggi makin terpisah)
Davies–Bouldin Index	misal: 0.4812	Mendekati 0 (Makin rendah makin baik)
Calinski–Harabasz Index	misal: 1450.85	Nilai Tinggi (Makin tinggi makin padat & terpisah)

The Silhouette Score provides an indication of the compactness and separation of the generated clusters. A higher score demonstrates that households assigned to the same cluster exhibit similar energy consumption characteristics while remaining sufficiently separated from households in other clusters.The Davies Bouldin Index complements the Silhouette Score by evaluating the similarity between clusters. Lower DBI values indicate that each cluster has a compact structure and maintains a clear distance from neighboring clusters.Meanwhile, the Calinski Harabasz Index evaluates the ratio of between-cluster variance to within-cluster variance. A higher CHI value indicates that the clustering model successfully maximizes the separation between clusters while

minimizing the dispersion of observations within the same cluster.Overall, these three internal validation metrics provide a comprehensive evaluation of the clustering performance. By combining these metrics, the reliability of the K-Means clustering model can be assessed from different perspectives, including cluster cohesion, cluster separation, and overall clustering structure.The evaluation results obtained from this study indicate the effectiveness of the K-Means algorithm in identifying household groups with similar energy consumption patterns. The use of multiple validation metrics ensures that the clustering solution is not evaluated using only a single criterion, thereby providing a more reliable assessment of cluster quality.

4.4 Discussion

The application of the K-Means clustering algorithm successfully grouped households according to similarities in their energy consumption characteristics. The clustering process considered multiple variables, including electricity consumption, household size, average environmental temperature, air conditioner ownership, and electricity usage during peak hours.The resulting clusters represent different household energy consumption patterns. Households within the same cluster share similar consumption characteristics, whereas households assigned to different clusters exhibit distinct electricity usage behaviors. This indicates that K-Means is capable of identifying meaningful patterns from household energy consumption data without requiring predefined class labels.The internal cluster validation metrics further support the effectiveness of the clustering model. The Silhouette Score evaluates the compactness and separation of the clusters, while the

Davies Bouldin Index measures cluster similarity, and the Calinski Harabasz Index assesses the balance between intra-cluster cohesion and inter cluster separation. The combination of these metrics provides a comprehensive evaluation of clustering quality.The findings of this study are consistent with previous research that applied K-Means clustering to electricity consumption analysis. Previous studies have shown that K-Means is effective in identifying customer groups based on energy consumption characteristics and can support demand analysis, customer segmentation, and energy management strategies.Furthermore, the clustering results can assist electricity providers and policymakers in understanding household consumption behavior. Different household groups may require different energy saving recommendations or demand side management strategies. Therefore, the clustering results obtained in this study have practical implications for improving household energy efficiency and supporting sustainable energy management.

4.5 Conclusion

This study applied the K-Means clustering algorithm to analyze household energy consumption patterns using the Household Energy Consumption dataset. Prior to clustering, data preprocessing was performed, including feature selection, label encoding, and Min-Max normalization to improve clustering performance.

The K-Means algorithm successfully grouped households according to similarities in electricity consumption characteristics. The clustering process enables the identification of different household energy consumption patterns, providing valuable insights into electricity usage behavior. The quality of the clustering model was evaluated using three internal validation metrics: the Silhouette Score, Davies–

Bouldin Index, and Calinski Harabasz Index. These metrics provide a comprehensive assessment of cluster cohesion and separation, ensuring that the resulting clusters are meaningful and reliable.

Overall, this study demonstrates that K-Means clustering is an effective unsupervised learning method for analyzing household energy consumption data. The identified consumption patterns can serve as a valuable reference for future research, electricity demand analysis, and the development of energy management strategies.

REFERENCES

Abdelaziz, A., Santos, V., & Dias, M. S. (2021). Machine Learning Techniques in the Energy Consumption of Buildings: A Systematic Literature Review Using Text Mining and Bibliometric Analysis. *Energies*, *14*(22). https://doi.org/10.3390/en14227810

Ding, Z., Wang, Z., Hu, T., & Wang, H. (2022). A Comprehensive Study on Integrating Clustering with Regression for Short-Term Forecasting of Building Energy Consumption: Case Study of a Green Building. *Buildings*, *12*(10). https://doi.org/10.3390/buildings12101701

Donti, P. L., & Kolter, J. Z. (2021). *Machine Learning for Sustainable Energy Systems*. 719–747.

Energy Efficiency. (2023).

Forte, M., Guzman, C. P., Pereira, L., & Morais, H. (2025). Identification of Typical Time-Series House Consumption Profiles Based on Unsupervised Learning Techniques. *IEEE Access*, *13*, 123326–123343. https://doi.org/10.1109/ACCESS.2025.3585204

International, I. E. A., & Agency, E. (2026). *Electricity 2024 - Analysis and forecast to 2026*.

Kallel, S., Amayri, M., & Bouguila, N. (2025). Clustering and Interpretability of Residential Electricity Demand Profiles. *Sensors*, *25*(7). https://doi.org/10.3390/s25072026

Khan, A.-N., Iqbal, N., Rizwan, A., Ahmad, R., & Kim, D.-H. (2021). An Ensemble Energy Consumption Forecasting Model Based on Spatial-Temporal Clustering Analysis in Residential Buildings. *Energies*, *14*(11). https://doi.org/10.3390/en14113020

Michalakopoulos, V., Sarmas, E., Papias, I., Skaloumpakas, P., Marinakis, V., & Doukas, H. C. (2023). A Machine Learning-Based Framework for Clustering Residential Electricity Load Profiles to Enhance Demand Response Programs. *ArXiv*, *abs/2310.20367*. https://api.semanticscholar.org/CorpusID:264820426

Muliono, R., Silviana, N. A., Polewangi, Y. D., & Lubis, A. H. (2026). KRG-VRP: An Integrated Hybrid Machine Learning Framework Integrating K-Means, Random Forest, and Genetic Algorithm for Capacitated Vehicle Routing Optimization. *International Journal of Intelligent Engineering \& Systems*, *19*(5), 143.

Nur, A., & Maulidhia, A. (2026). *Perbandingan Algoritma K-Means dan Hierarchical Clustering untuk Pengelompokan Konsumsi Listrik Rumah Tangga Menggunakan Silhouette Coeffcient dan Davies-Bouldin Index*. *9*,

10–19.

Ofetotse, E. L., Essah, E. A., & Yao, R. (2021). Evaluating the determinants of household electricity consumption using cluster analysis. *Journal of Building Engineering*, *43*, 102487. https://doi.org/https://doi.org/10.1016/j.jobe.2021.102487

Palaniappan, S., Karuppannan, S., & Velusamy, D. (2024). Categorization of Indian residential consumers electrical energy consumption pattern using clustering and classification techniques. *Energy*, *289*, 129992. https://doi.org/https://doi.org/10.1016/j.energy.2023.129992

Wang, T., Zhao, Q., Gao, W., & He, X. (2024). Research on energy consumption in household sector: a comprehensive review based on bibliometric analysis. *Frontiers in Energy Research, Volume 11-2023*. https://doi.org/10.3389/fenrg.2023.1209290

Wei, Z., & Wang, H. (2021). Characterizing Residential Load Patterns by Household Demographic and Socioeconomic Factors. *CoRR, abs/2106.05858*. https://arxiv.org/abs/2106.05858

Wen, H., Liu, X., Yang, M., Lei, B., Xu, C., & Chen, Z. (2024). A novel approach for identifying customer groups for personalized demand-side management services using household socio-demographic data. *Energy*, *286*, 129593. https://doi.org/https://doi.org/10.1016/j.energy.2023.129593

Zhang, X., Ramírez-Mendiola, J. L., Li, M., & Guo, L. (2022). Electricity consumption pattern analysis beyond traditional clustering methods: A novel self-adapting semi-supervised clustering method and application case study. *Applied Energy*, *308*, 118335. https://doi.org/https://doi.org/10.1016/j.apenergy.2021.118335