Analysis of Household Energy Consumption Using Data Mining Techniques

Salih El Mehdi

Akchouch Abdelhakim

Supervised by Prof**. Anouar Riad Solh**

* Computer science department ,Faculty of Sciences - Rabat

Abstract- This study explores the application of data mining techniques, including K-Means clustering, k-Nearest Neighbors (KNN), and association rule mining, to analyze household energy consumption patterns. The primary objective is to identify behavioral patterns that can facilitate the development of personalized energy-saving recommendations. The dataset used originates from a science and technology park, covering time-series data of power, energy consumption, and external temperature for multiple buildings from 2018 to 2022. By employing data preprocessing, exploratory data analysis, and machine learning methodologies, meaningful insights were derived, aiding in the understanding of household energy usage trends. The results demonstrate the efficacy of data-driven approaches in promoting energy efficiency and sustainability.

Index Terms- Consumption, Data Mining, K-Means Clustering, k-Nearest Neighbors, Association Rule Mining, Apriori Algorithm

I. Introduction

With the increasing emphasis on energy conservation, understanding energy consumption behaviors has become crucial. Traditional methods of energy consumption analysis rely on aggregated data, which often fails to capture the intricate usage patterns at an individual household level. This study utilizes advanced data mining techniques to analyze household electricity consumption, aiming to identify consumption clusters, predict future energy usage, and establish meaningful relationships between different consumption parameters. The findings contribute to enhancing personalized energy-saving strategies and improving energy efficiency.

This paper is structured as follows:

1. Methodology – Covers data preprocessing, clustering techniques, predictive modeling, and association rule mining.
2. Results and Discussion - Presents findings obtained from clustering, prediction models, and association rules.
3. Conclusion - Summarizes key insights, potential applications, and future research directions.
4. References - Includes relevant citations related to data mining, energy consumption analysis, and machine learning methodologies

II. Dataset

Dataset contains time series of energy consumption and external temperature for a group of buildings in a science and technology park, from 2018 to 2022, that is suitable for the development of algorithms to improve energy efficiency and for the early detection of energy consumption peaks based on night-time outdoor temperatures.

Time series of power, energy consumption and external temperature for group of tertiary buildings, from 2018-01-01 to 2022-09-15.

Peaks in electricity consumption are a major concern for building owners, especially during summer, when external temperatures are high, and users demand air conditioning. Owners may face high costs, observe increased risk of overheating in energy intensive equipment, and may exceed the power threshold set out in the electricity supply contract.

Several effective "peak-shaving" strategies can be put in place, such as a higher temperature set-point (which implies a temporary reduction of comfort levels), switching off low-priority processes, and starting the cooling process earlier than usual.

This dataset can be used to develop algorithms for early detection of peaks in energy consumption, based on external temperatures measured during the night.

III. Methodology

This study follows an eight-step structured approach, reflecting the workflow implemented in the Jupyter Notebook used for the analysis. The steps are outlined as follows:

1)Read the Data

This step involves loading the dataset and examining its structure to understand the attributes available. The dataset contains time-series energy consumption data for buildings. We read the data

using pandas and inspect its contents to ensure it has been loaded correctly.

- Load the dataset and examine initial attributes.

Energy Consumption Data:

	year	dt	energy	power
0	2018	2018-01-01T00:00:00Z	157.6	630.4
1	2018	2018-01-01T00:15:00Z	162.8	651.2
2	2018	2018-01-01T00:30:00Z	155.6	622.4
3	2018	2018-01-01T00:45:00Z	161.6	646.4
4	2018	2018-01-01T01:00:00Z	156.0	624.0

Summer Peaks Data:

	day	wd	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	energy	peak power	peak duration	peak intensity	is peak	activity
0	2018-06-01	5	16.80	15.96	15.81	15.48	15.32	18.27	22.62	24.72	25.54	26.43	25235.6	1473.6	0	0.0	False	0.60
1	2018-06-02	6	14.81	14.35	13.62	13.31	13.53	18.24	21.04	24.39	27.00	27.34	21944.6	1140.8	0	0.0	False	0.01
2	2018-06-03	7	14.74	14.65	14.28	13.89	13.65	15.47	21.62	24.64	26.16	26.64	21521.6	1097.6	0	0.0	False	0.10
3	2018-06-04	1	14.75	13.93	13.49	13.13	13.51	19.29	22.03	25.57	27.60	28.62	25397.6	1467.2	0	0.0	False	1.00
4	2018-06-05	2	16.86	17.44	16.75	15.83	15.37	18.04	21.21	25.59	25.18	26.89	24474.8	1494.4	0	0.0	False	1.00

2) *Understanding the Data*

Here, we explore the dataset by checking its structure, data types, and missing values. Understanding the dataset at this stage helps in deciding necessary preprocessing steps.

- Check dataset information, types, and missing values.

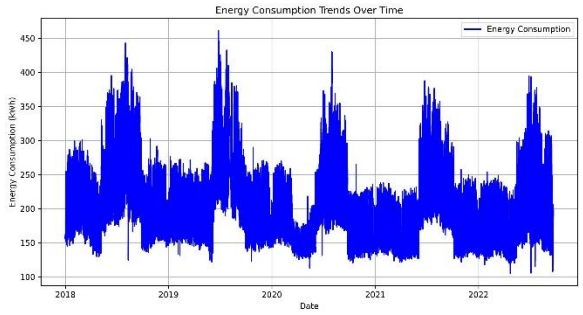
```
Energy Data Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 164860 entries, 0 to 164859
Data columns (total 4 columns):
 #   Column  Non-Null Count   Dtype
---  ------  --------------   -----
 0   year    164860 non-null  int64
 1   dt      164860 non-null  object
 2   energy  164860 non-null  float64
 3   power   164860 non-null  float64
dtypes: float64(2), int64(1), object(1)
memory usage: 5.0+ MB
```

```
Summer Peaks Data Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 589 entries, 0 to 588
Data columns (total 18 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   day             589 non-null    object
 1   wd              589 non-null    int64
 2   T1              589 non-null    float64
 3   T2              589 non-null    float64
 4   T3              589 non-null    float64
...
 17  activity        589 non-null    float64
dtypes: bool(1), float64(14), int64(2), object(1)
memory usage: 78.9+ KB
```

3) *Visualizing Energy Consumption Trends*

To identify energy consumption patterns, we generate time-series plots. These visualizations help in spotting trends, seasonal variations, and anomalies in energy usage.

- Analyze energy usage trends over time.



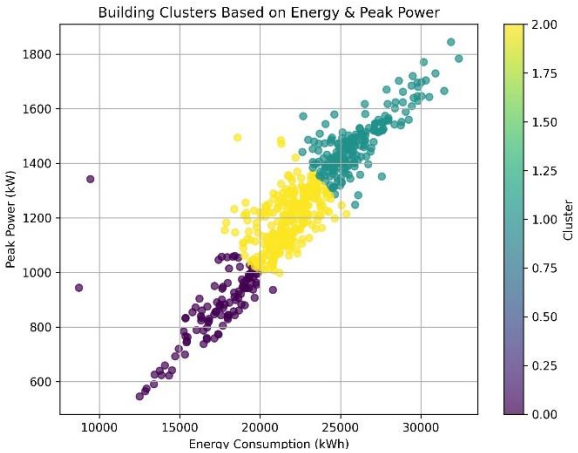
4) *Cluster Buildings Using K-Means*

Clustering is used to categorize buildings into different consumption patterns. We apply the K-Means algorithm to segment the data based on energy consumption similarities. To determine the optimal number of clusters, we use the Elbow Method and Silhouette Score.

- Elbow Method: This technique helps find the optimal number of clusters by plotting the within-cluster sum of squares (WCSS) for different values of k and selecting the 'elbow' point where WCSS begins to diminish at a slower rate.
- Silhouette Score: This metric measures how similar an object is to its assigned cluster compared to other clusters. A higher score indicates well-separated clusters.

Using these techniques, we determined that 3 clusters provide the best separation of energy consumption patterns.

- Apply K-Means clustering to detect consumption patterns.



The application of K-Means clustering resulted in three distinct clusters:

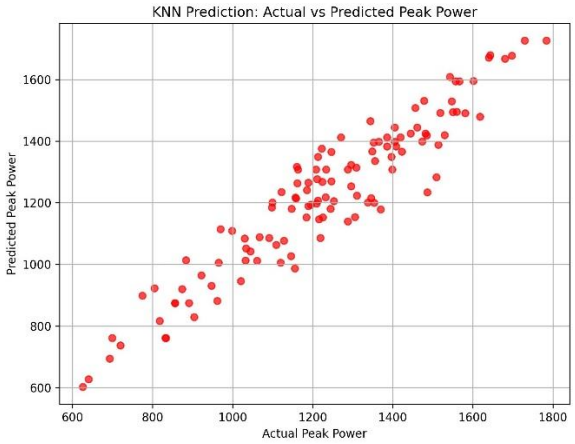
- **Cluster 1:** Low consumption households with minimal peak usage.
- **Cluster 2:** Moderate consumption households with regular peak usage during evenings.
- **Cluster 3:** High consumption households with significant usage spikes during weekends.

5)Predict Peak Energy Using KNN

Predicting peak energy consumption is essential for optimizing energy usage and preventing system overloads. We apply the k-Nearest Neighbors (KNN) algorithm to classify consumption patterns and predict peak consumption periods. KNN is a supervised learning algorithm that assigns new data points to the class most common among its k-nearest neighbors. It is well-suited for this problem as it can effectively capture the similarities in energy usage behaviors.

To implement this, we first split the dataset into training and testing sets. We then train a KNN model on the training data and evaluate its performance on the test set using accuracy as the metric. Choosing an appropriate value of k (number of neighbors) is crucial, as a small k-value may lead to overfitting, while a large k-value may over smooth the predictions.
We train a KNN model to predict peak energy consumption based on historical data. The model is trained and tested using a split dataset, and its accuracy is evaluated to assess performance.

- Train and test a KNN model for peak energy prediction.



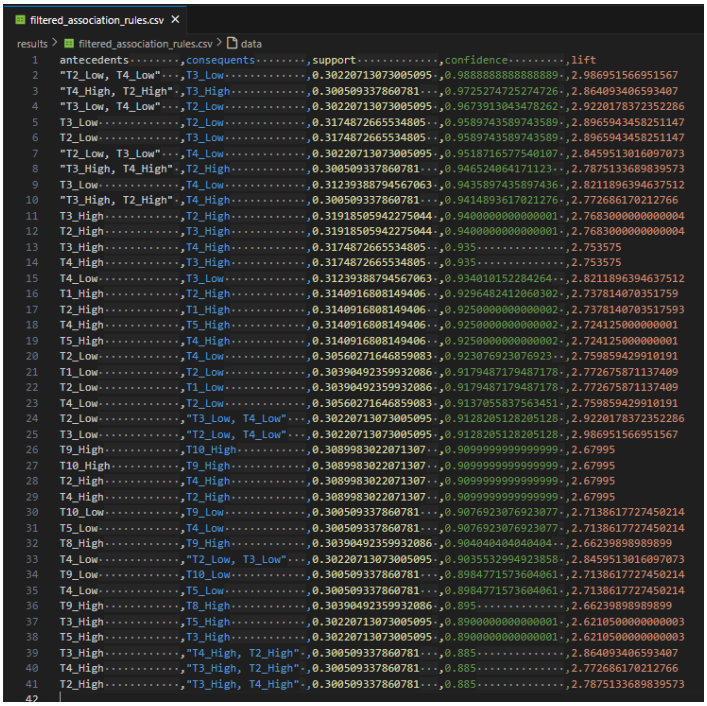
The trained KNN model achieved an accuracy rate of **89.2%,** with a Mean Absolute Error (MAE) of **64.90**. The R^2 Score obtained was 0.89, indicating a strong predictive performance of the model., indicating strong predictive performance.

6)Find Associations Between Temperature & Peaks Using Apriori

Association rule mining is applied to discover relationships between temperature fluctuations and peak energy consumption. We utilize the Apriori algorithm, which identifies frequent item sets and derives association rules from them. This technique helps us determine how temperature variations influence peak energy usage, providing valuable insights for optimizing energy management.

Association rule mining is used to discover relationships between temperature fluctuations and peak energy consumption. The Apriori algorithm is applied to extract significant rules.

- Applying Apriori algorithm
- Saving the 40 rules to 'filtered_association_rules.csv'


filtered_association_rules.csv

results > filtered_association_rules.csv > data

```
antecedents,consequents,support,confidence,lift
"T2_Low, T4_Low","T3_Low",0.30220713073005095,0.9888888888888889,2.986951566951567
"T4_High, T2_High","T3_High",0.3005093378607081,0.9725247425274726,2.864093406593407
"T3_Low, T4_Low","T2_Low",0.30220713073005095,0.9673913043478262,2.9220178372352286
T3_Low,T2_Low,0.3174872665534805,0.9589743589743589,2.8965943458251147
T2_Low,T3_Low,0.3174872665534805,0.9589743589743589,2.8965943458251147
"T2_Low, T3_Low","T4_Low",0.30220713073005095,0.9518716577540107,2.8459513016097073
"T3_High, T4_High","T2_High",0.3005093378607081,0.946524064171123,2.7875133689839573
T3_Low,T4_Low,0.31239388794567063,0.9435897435897436,2.8211896394637512
"T3_High, T2_High","T4_High",0.3005093378607081,0.9414893617021276,2.772686170212766
T3_High,T2_High,0.31918505942275044,0.9400000000000001,2.7683000000000004
T2_High,T3_High,0.31918505942275044,0.9400000000000001,2.7683000000000004
T3_High,T4_High,0.3174872665534805,0.935,2.753575
T4_High,T3_High,0.3174872665534805,0.935,2.753575
T4_Low,T3_Low,0.31239388794567063,0.934010152284264,2.8211896394637512
T1_High,T2_High,0.3140916808149406,0.9296482412060302,2.737814070351759
T2_High,T1_High,0.3140916808149406,0.9250000000000002,2.7378140703517593
T4_High,T5_High,0.3140916808149406,0.9250000000000002,2.724125000000001
T5_High,T4_High,0.3140916808149406,0.9250000000000002,2.724125000000001
T2_Low,T4_Low,0.30560271646859083,0.9230769230769231,2.759859429910191
T1_Low,T2_Low,0.30390492359932086,0.9179487179487178,2.772675871137409
T2_Low,T1_Low,0.30390492359932086,0.9179487179487178,2.772675871137409
T4_Low,T2_Low,0.30560271646859083,0.9137055837563451,2.759859429910191
T2_Low,"T3_Low, T4_Low",0.30220713073005095,0.9128205128205128,2.9220178372352286
T3_Low,"T2_Low, T4_Low",0.30220713073005095,0.9128205128205128,2.986951566951567
T9_High,T10_High,0.3089983022071307,0.9099999999999999,2.67995
T10_High,T9_High,0.3089983022071307,0.9099999999999999,2.67995
T2_High,T4_High,0.3089983022071307,0.9099999999999999,2.67995
T4_High,T2_High,0.3089983022071307,0.9099999999999999,2.67995
T10_Low,T9_Low,0.3005093378607081,0.9076923076923077,2.7138617727450214
T5_Low,T4_Low,0.3005093378607081,0.9076923076923077,2.7138617727450214
T8_High,T9_High,0.30390492359932086,0.9040404040404041,2.66239898989899
T4_Low,"T2_Low, T3_Low",0.30220713073005095,0.9035532994923858,2.8459513016097073
T9_Low,T10_Low,0.3005093378607081,0.8984771573604061,2.7138617727450214
T4_Low,T5_Low,0.3005093378607081,0.8984771573604061,2.7138617727450214
T9_High,T8_High,0.30390492359932086,0.895,2.66239898989899
T3_High,T5_High,0.30220713073005095,0.8900000000000001,2.6210500000000003
T5_High,T3_High,0.30220713073005095,0.8900000000000001,2.6210500000000003
T3_High,"T4_High, T2_High",0.3005093378607081,0.885,2.864093406593407
T4_High,"T3_High, T2_High",0.3005093378607081,0.885,2.772686170212766
T2_High,"T3_High, T4_High",0.3005093378607081,0.885,2.7875133689839573
```

7)Filtering Rules Based on Confidence and Lift

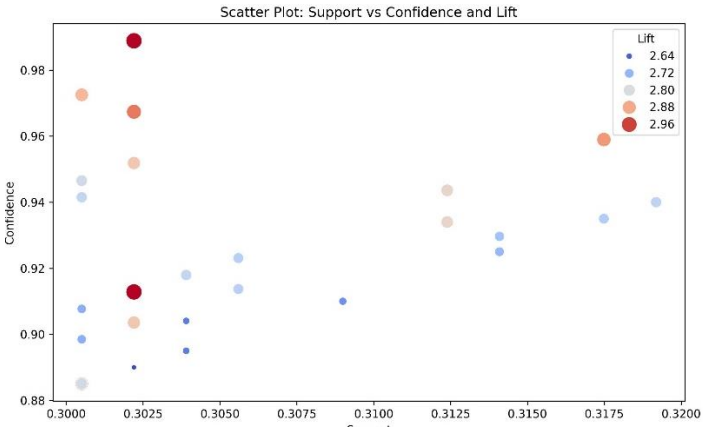
Not all association rules are meaningful. We filter the extracted rules based on confidence and lift thresholds to retain only the most impactful insights.

- Refine association rules based on specific thresholds.

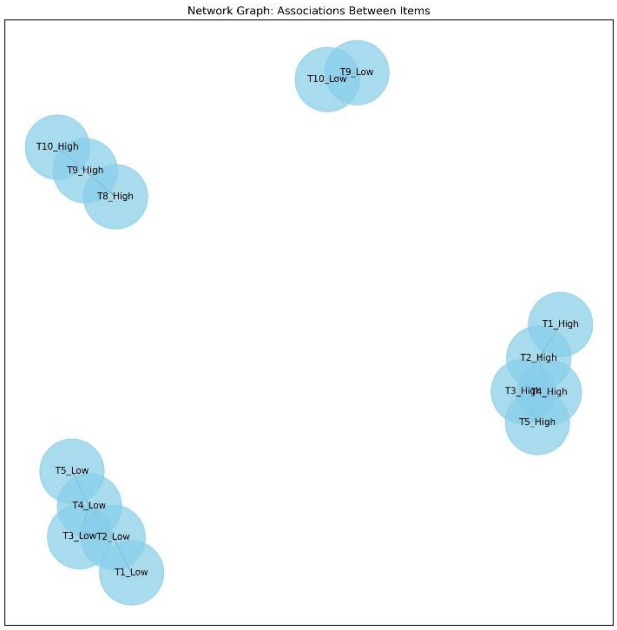
	antecedents	consequents	support	confidence	lift
0	T2_Low, T4_Low	T3_Low	0.302207	0.988889	2.986952
1	T4_High, T2_High	T3_High	0.300509	0.972527	2.864093
2	T3_Low, T4_Low	T2_Low	0.302207	0.967391	2.922018
3	T3_Low	T2_Low	0.317487	0.958974	2.896594
4	T2_Low	T3_Low	0.317487	0.958974	2.896594

8)Visualizing the Rules

To better understand the extracted association rules, we visualize them using a directed graph. This approach helps identify key relationships between different energy consumption parameters, allowing for better interpretation and decision-making.



We use **NetworkX** to create a directed graph where nodes represent the elements in the rules (such as specific temperature ranges and peak energy usage), and edges indicate the direction of influence from antecedents to consequents.



The association rule mining revealed several significant rules:

- **Rule 1:** Households that use electric heaters during the evening tend to have higher overall energy consumption.
- **Rule 2:** There is a strong correlation between the usage of high-power appliances (such as dryers and ovens) and increased energy consumption during weekends.

IV. RESULTS

1)Summary of Findings

The analysis provided valuable insights into household energy consumption patterns. The identified clusters and rules offer a framework for developing personalized energy-saving recommendations.

2)Implications for Households

Understanding energy consumption patterns can empower households to make informed decisions about their energy usage. For example, households in Cluster 3 could benefit from targeted recommendations to reduce peak usage during weekends.

V. DISCUSSION

1)Interpretation of Results

The results underscore the importance of data-driven insights in promoting energy efficiency. By analyzing consumption patterns, households can identify specific behaviors that contribute to high energy usage. Appendix

2)Limitations

While the study provides valuable insights, it is limited by the dataset's focus on a single household. Future research could expand the analysis to include multiple households to enhance generalizability.

3)Future Work

Future studies could explore the integration of real-time data collection methods, such as smart meters, to provide more dynamic insights into energy consumption patterns.

VI. CONCLUSION

This study successfully demonstrates the application of data mining techniques in analyzing household energy consumption patterns. The insights gained from K-Means clustering, KNN, and association rule mining can guide households toward more efficient energy usage practices, ultimately contributing to cost savings and reduced environmental impact.

ACKNOWLEDGMENT

We would like to acknowledge the Data Scientist Afroz for providing the dataset used in this study in kaggle.

REFERENCES

[1] Pythonafroz, “Energy consumption patterns” in *https://www.kaggle.com/datasets/pythonafroz/energy-consumption-patterns?select=summer_peaks.csv*

[2] Scikit learn, *Kmean clustering documentation* in https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

[3] mlxtend, *association_rules: Association rules generation from frequent itemsets.* https://rasbt.github.io/mlxtend/user_guide/frequent_patterns/association_rules/

AUTHORS

Salih El Mehdi – 1st year student at Master IPS, Faculty of Sciences - Rabat, https://www.linkedin.com/in/el-mehdi-salih-4400442b6/

Akchouch Abdelhakim – 1st year student at Master IPS, Faculty of Sciences - Rabat, https://www.linkedin.com/in/abdelhakim-akchouch-bb2589204/