Data Mining for Smart Home Energy Consumption Analysis and Prediction

1. Introduction

In the context of increasing global energy demand and smart grid development, analyzing household electricity usage has become crucial. This project applies **data mining techniques** to the **Individual Household Electric Power Consumption Dataset** (UCI Machine Learning Repository) to uncover consumption patterns, predict future usage, and classify household behaviors.

The dataset contains over **2 million measurements** collected over nearly 4 years (2006–2010) from a single household, including sub-metered appliance usage, voltage, and current intensity readings. The data is **messy**, with **missing values**, **timestamp irregularities**, and **mixed numerical and categorical features**, making it ideal for realistic data mining practice.

2. Objectives

The project aims to:

1. **Preprocess and clean** real-world energy consumption data with noise and missing values.
2. **Extract meaningful features** (daily/weekly aggregations, temporal and weather-based features).
3. Apply multiple **data mining tasks**:
    - **Classification**: Detect abnormal or high-consumption days.
    - **Regression**: Predict next-day total energy consumption.
    - **Clustering**: Identify typical daily consumption profiles.
    - **Association Rules**: Discover patterns among appliance usage and time-of-day.
4. Optionally include **time-series forecasting** using advanced models.

3. Dataset Description

Dataset: Individual household electric power consumption

Attributes:

Attribute	Description
`Date`, `Time`	Timestamp of measurement (1-min sampling rate)
`Global_active_power`	Household global minute-averaged active power (kW)
`Global_reactive_power`	Reactive power (kW)
`Voltage`	Average voltage (V)
`Global_intensity`	Current intensity (A)
`Sub_metering_1`	Kitchen appliances consumption (Wh)
`Sub_metering_2`	Laundry appliances consumption (Wh)
`Sub_metering_3`	Water heater and AC consumption (Wh)

Issues:

- Missing and invalid readings (`?` or zero values)
- Irregular timestamps
- Strong seasonal patterns (daily/weekly)
- Requires aggregation and feature engineering

4. Data Preprocessing and Feature Engineering

Steps:

1. **Data Cleaning**
    - Remove or impute missing values.
    - Convert timestamps to proper `datetime` objects and resample.
2. **Feature Construction**
    - Aggregate features:
        - `Daily_total_power = sum(Global_active_power)`
        - `Peak_hour_power`, `Nighttime_usage`, `Weekend_usage`.
    - Create categorical features: day of week, season, working/non-working day.
    - Optional: integrate external **weather data** (temperature, humidity).
3. **Normalization**
    - Normalize continuous features using scaling when needed.

5. Methodology

5.1 Classification

Goal: Predict whether a given day's energy consumption is *"High"* or *"Low"* compared to the household's average.

- **Target variable:** Binary (1 = High consumption, 0 = Normal).
- **Evaluation:** Accuracy, F1-score, ROC-AUC.

5.2 Regression

Goal: Predict next-day total energy consumption (in kWh).

- **Evaluation:** MAE, RMSE, R^2.

5.3 Clustering

Goal: Identify groups of days with similar consumption profiles.

- **Evaluation:** Silhouette Score, Davies–Bouldin Index.
- **Interpretation:** e.g.
    - Cluster 1 → Weekdays, typical pattern.
    - Cluster 2 → Weekends, morning/evening peaks.
    - Cluster 3 → Abnormal/high consumption days.

5.4 Association Rule Mining

Goal: Discover frequent patterns between appliances and time periods.

- **Preprocessing:**
    - Discretize continuous features (e.g., Low/Medium/High sub-meter usage).
    - Encode as transactions (e.g., "High Sub_metering_1 + Nighttime" → "High Global_active_power").
- **Algorithm:** Apriori or FP-Growth (using `mlxtend`).
- **Metrics:** Support, Confidence, Lift.
- **Example Rule:**
  `IF (High Sub_metering_3) AND (Weekend) → High Total Consumption (Support=0.18, Confidence=0.74)`.

5.5 (Optional) Time-Series Forecasting

- **Goal:** Forecast future hourly/daily consumption.
- **Methods:** ARIMA, Prophet, or LSTM.
- **Evaluation:** RMSE, MAPE.

6. Evaluation & Results Interpretation

Task	Metric	Expected Outcome
Classification	Accuracy, F1	Identify high-consumption days with >85% accuracy
Regression	RMSE	Predict next-day usage within ±0.3 kWh
Clustering	Silhouette	Distinct weekday/weekend clusters
Association	Lift, Confidence	Discover meaningful usage correlations

Interpret patterns in relation to:

- Seasonal/weekday effects
- Appliance-level efficiency
- Energy-saving potential

7. Expected Deliverables

The submission date is January 15, 2026, on eClass. There will be a short 10-minute presentation in class.

1. **Data preprocessing notebook** (cleaning + feature extraction).
2. **Modeling notebook** (classification, regression, clustering, association).
3. **Final report** including:
    - Introduction
    - Methodology and experimental setup
    - Results and discussion
    - Conclusions and limitations

Grading criteria:
i) **Preparation:** description of preprocessing actions (10%)
ii) **Classification:** description of steps and evaluation of algorithms (20%)
iii) **Regression:** performance achieved on the training data (20%)
iv) **Clustering:** methodology followed and interpretation of clusters (20%)
v) **Association rule mining:** description of the methodology and interpretation of the generated rules (20%)

vi) **Final presentation of the project** (10%)
vii) **Time-series forecasting:** prediction results and analysis (10%)

The project can be carried out by **up to 3 students**.