📘 Dataset Documentation

---

1. Dataset Summary

- **Source**: Individual Household Electric Power Consumption (UCI ML Repository).
- **Frequency**: Hourly (aggregated from raw 1-minute data).
- **Period**: December 2006 – November 2010.
- **Size**: 34,589 rows × 35 columns.
- **Goal**: Model electricity consumption patterns by combining **technical power measures, behavioral time patterns, environmental weather data, holiday effects, economic tariffs, and macroeconomic energy prices**.

---

2. Feature Documentation

🔷 A. Original Features (from raw dataset)

Feature	Unit	Description	Why it matters
Global_active_power	kW	Total household active power consumption.	Core target for forecasting demand.
Global_reactive_power	kVAR	Reactive power (energy stored/released by appliances).	Helps measure inefficiencies.
Voltage	Volt	Average voltage.	Indicator of appliance performance and grid stability.
Global_intensity	Ampere	Current intensity drawn.	Direct measure of load.
Sub_metering_1	Wh	Kitchen appliances (dishwasher, oven, microwave).	Detect appliance-level demand.
Sub_metering_2	Wh	Laundry appliances (washing machine, dryer, fridge, light).	Appliance-specific demand.

Feature	Unit	Description	Why it matters
Sub_metering_3	Wh	Water heater + air conditioning.	Captures heavy load appliances.

🔷 B. Derived Energy Features

Feature	Formula / Source	Why it matters
Total_sub_metering	Sub_metering_1 + Sub_metering_2 + Sub_metering_3	Captures metered consumption across appliances.
Non_Submetered_Energy	(Global_active_power × 1000 / 60) – Total_sub_metering	Hidden loads (TVs, computers, lighting not covered).
Power_Factor	Active / sqrt($Active^2$ + $Reactive^2$)	Efficiency ratio → closer to 1 = efficient use of electricity.

🔷 C. Temporal Features (from datetime)

Feature	Source	Why it matters
hour	Extracted from timestamp	Captures daily cycles (e.g., evening peaks).
day_of_week	Extracted (Mon–Sun)	Behavior differs between weekdays & weekends.
month	Extracted (1–12)	Seasonal variability.
Season	Derived (Winter, Spring, Summer, Autumn)	Models heating/cooling patterns.
hour_sin, hour_cos	sin/cos(2π·hour/24)	Encodes daily cycle without 23→0 discontinuity.
day_sin, day_cos	sin/cos(2π·day_of_week/7)	Captures weekly cycles smoothly.

🔷 D. Weather Features (external data – Meteostat API)

Feature	Source	Why it matters
Temperature_C	Meteostat hourly weather	Heating/cooling demand driver.
Humidity (%)	Meteostat	Impacts AC usage and comfort levels.
Precipitation (mm)	Meteostat	Affects occupancy (rain → more home activity).
Temp_Regime	Derived (Cold / Mild / Hot)	Simplifies modeling of climate effect.

🔷 E. Holiday & Calendar Features

Feature	Source	Why it matters
is_holiday	Python holidays package	Occupancy spikes on holidays.
holiday_name	Holidays package	Enables holiday-specific analysis (e.g., Christmas vs Easter).
day_type	Derived: {Workday, Weekend, Holiday}	General calendar effect.
Day_Before_Holiday	Shifted holiday flag	Captures pre-holiday preparation peaks.
Day_After_Holiday	Shifted holiday flag	Captures post-holiday recovery dips.

🔷 F. Tariff Features (pricing schedules)

Feature	Source	Why it matters
Heures_Creuses	French off-peak tariff schedule	Encourages load shifting during cheaper hours.
Tariff_Zone	Derived: {Off_Peak, Peak_Price}	Economic factor directly influencing energy use.

🔷 G. Macroeconomic Feature

Feature	Source	Why it matters
HICP_Energy_Index	Eurostat → FRED API → pandas_datareader (Series ID: CP0450FRM086NEST)	Captures long-term trend of **real energy prices** (electricity, gas, other fuels) in France. Higher index → higher prices → consumer conservation.

How it was built:

1. **Tool**: pandas_datareader in Python.
2. **API**: Connects to FRED (Federal Reserve Economic Data).
3. **Source**: Eurostat's *Harmonized Index of Consumer Prices (HICP)* for France.
4. **Series ID**: CP0450FRM086NEST.
5. **Result**: Monthly index interpolated to align with hourly household data.

📊 This feature adds a **macroeconomic dimension**, showing how long-term price trends affect household consumption.

---

3. Why These Features Together?

- **Technical (power measures)** → Explain baseline energy demand.
- **Behavioral (time + holidays)** → Capture when and why households consume electricity.
- **Environmental (weather)** → Energy use is highly climate-sensitive.
- **Economic (tariffs + energy prices)** → Prices and incentives directly affect behavior.
- **Efficiency (derived metrics)** → Power factor + hidden loads reveal inefficiencies.
- **Macroeconomic (HICP index)** → Accounts for broader economic context, not just household-level behavior.

Together, this dataset becomes **multi-dimensional**, suitable for:

- Short & long-term **load forecasting**.
- **Demand response studies** (peak shaving, tariff impact).
- **Smart home automation** (appliance-level efficiency).
- **Policy analysis** (how tariffs & prices affect demand).