**PREPROCESSING AND EXPLORE HOUSEHOLE ELECTRICITY**

***1.Mô tả tập dữ liệu***

-Tập dữ liệu Mức tiêu thụ điện gia đình là tập dữ liệu chuỗi thời gian đa biến mô tả mức tiêu thụ điện cho một hộ gia đinh trong vòng bốn năm.

-Kho dữ liệu này chứa 2075259 phép đo được tập hợp tại một ngôi nhà nằm ở Sceaux (cách Paris, Pháp 7 km) trong khoảng thời gian từ tháng 12 năm 2006 đến tháng 11 năm 2010 (47 tháng).

- sub\_meteing\_4 = (global\_active\_power \* 1000/60 – sub\_metering\_1 – sub\_metering\_2 – sub\_metering\_3) đại diện cho năng lượng hoạt động tiêu thụ mỗi phút (tính bằng watt-giờ) trong gia đình bởi các thiết bị điện không được đo bằng sub\_metering\_1, 2, 3.

-Trong tập dữ liệu chứa một số giá trị bị thiếu trong các phép đo (gần 1.25% số hàng). Tất cả các dấu thời gian lịch đều có trong tập dữ liệu nhưng đối với một số dấu thời gian, các giá trị đo lường bị thiếu: giá trị bị thiếu được biểu bị bằng việc không có giá trị giữa hai dấu phân tách thuộc tính dấu chấm phẩy liên tiếp. Ví dụ: tập dữ liệu hiển thị các giá trị bị thiếu vào ngày 28 tháng 4 năm 2007.

***2.Thông tin thuộc tính***

- date: ngày ở định dạng dd / mm / yyyy.

- time: thời gian ở định dạng hh : mm : ss.

- global\_active\_power: tổng công suất hoạt động mà hộ gia đình tiêu thụ (kilowatt).

- global\_reactive\_power: tổng công suất phản kháng hộ gia đình tiêu thụ (kilowatt).

- voltage: điện áp trung bình theo phút (tính bằng vol).

- global\_intensity: cường độ dòng điện trung bình theo phút toàn bộ của hộ gia đình (tính bằng ampe).

- sub\_metering\_1: năng lượng hoạt động số 1 (tính bằng watt-giờ năng lượng hoạt động). Nó tương ứng với nhà bếp, chứa chủ yếu là máy rửa bát, lò nướng và lò vi sóng.

- sub\_metering\_2: năng lượng hoạt động số 2 (tính bằng watt-giờ năng lượng hoạt động). Nó tương ứng với phòng giặt, chứa máy giặt, máy sấy quần áo, tủ lạnh và đèn chiếu sáng.

- sub\_metering\_3: năng lượng hoạt động số 3 (tính bằng watt-giờ năng lượng hoạt động). Nó tương ứng với một máy nước nóng điện và một máy lạnh.

- sub\_metering\_4: năng lượng hoạt động số 4 (tính bằng watt-giờ năng lượng hoạt động). Nó tương ứng với các thiết bị còn lại trong nhà (ngoài sub\_metering\_1, 2, 3).

3.Tiền xử lý dữ liệu

a.Xử lý dữ liệu thiếu trong tập dữ liệu

Graphical user interface, application

Description automatically generated with medium confidence

Figure 1: Dữ liệu thiếu trong tập dữ liệu

-Trong tập dữ liệu tồn tại các thời điểm thiếu dữ liệu được biểu thị bằng dấu “?”

Table

Description automatically generated with low confidence

Figure 2: Minh họa dữ liệu bị thiếu

Chart, pie chart

Description automatically generated

Figure 3: % dữ liệu bị thiếu

- Quan sát biểu đồ Pie có thể thấy số lượng dữ liệu bị thiếu chiếm một phần nhỏ so với tổng số lượng dữ liệu.

b.Thay thế dữ liệu bị thiếu

-Thay thế dữ liệu bị thiếu hay thay dấu “?” bằng giá trị “nan”.

-Thêm feature “sub\_metering\_4” – năng lượng hoạt động của các thiết bị còn lại trong nhà (ngoại trừ sub\_metering\_1, 2, 3).

-Công thức tính sub\_metering\_4:

sub\_meteing\_4 = (global\_active\_power \* 1000/60 – sub\_metering\_1 - sub\_metering\_2 – sub\_metering\_3)

3.Explore the dataset

a.

Timeline

Description automatically generated

Figure 4:Dữ liệu theo từng phút trong 4 năm của từng thuộc tính

-Hình ảnh trên mang lại mức độ quan sát trong bốn năm theo từng phút của 8 thuộc tính trong tập dữ liệu.

-Chúng ta có thể thấy trong ‘Sub\_metering\_3’, mức độ tiêu thụ trong cùng một năm ở từng thời điểm khá tương đương nhau và tổng thể thì ‘Sub\_metering\_3’ có xu hướng tăng dần.

-Hơn nữa, ‘sub\_metering\_4’ dường như giảm dần theo thời gian hoặc cho thấy xu hướng giảm.

-Chúng ta thấy làn sóng của hiệu ứng theo mùa trong ‘Global\_active\_power’.

b.

A picture containing timeline

Description automatically generated

Figure 5: Dữ liệu của thuộc tính 'Global\_active\_power' theo từng phút trong 4 năm

-Quan sát chung cho 4 năm ta thấy, khoảng tháng 2-3 và khoảng tháng 8-9, nơi ta thấy mức tiêu thụ giảm rõ rệt.

-Ngoài ra, ta thấy xu hướng giảm trong những tháng mùa hè và có lẽ lượng tiêu thụ nhiều hơn trong những tháng mùa đông.

c.

Text

Description automatically generated with low confidence

Figure 6: 'Global\_active\_power' của năm 2007 trong từng tháng

-Quan sát vào lượng tiêu thụ điện trong 12 tháng của năm 2007, ta có thể thấy tháng 4 và tháng 8 có lượng tiêu thụ điện khá ít, điều này có thể do các kỳ nghỉ làm cho việc sử dụng thiết bị điện giảm xuống.

d.

A picture containing table

Description automatically generated

Figure 7: 'Global\_active\_power' trong 12 tháng của năm 2008

-Quan sát thấy trong tháng 8, lượng tiêu thụ điện khá thấp và khá giống nhau trong các thời điểm, có thể là do gia đình đi nghỉ mát, du lịch.

e.

A picture containing text

Description automatically generated

Figure 8: 'Global\_active\_power' trong 12 tháng của năm 2009

-Quan sát thấy rõ việc thiếu dữ liệu của tháng 6 và tháng 8, đồng thời trong một vài ngày đầu tháng 8, lượng điện tiêu thụ điện cũng khá thấp và không có sự chênh lệch nhiều.

f.

Table

Description automatically generated with medium confidence

Figure 9: 'Global\_active\_power' trong 12 tháng của năm 2010

-Quan sát thấy việc thiếu dữ liệu trong các tháng 1, 3, 8 và 9, đồng thời cuối tháng 7 và đầu tháng 8 việc tiêu thụ điện giảm rõ rệt và có sự chênh lệch không đáng kể giữa các ngày.

g.Dữ liệu tiêu thụ điện trong 31 ngày của tháng 1 năm 2007.

A picture containing diagram

Description automatically generated

Letter

Description automatically generated

Letter

Description automatically generated

4.Data distribution

a.

Chart

Description automatically generated

Figure 10: Phân phối dữ liệu của 8 thuộc tính

-Quan sát thấy công suất hoạt động và phản kháng, cường độ cũng như công suất đo đếm phụ đều có sự phân bố lệch xuống theo các giá trị watt-giờ hoặc kilowatt nhỏ.

-Đồng thời, có thể thấy rằng sự phân bố dữ liệu điện áp rất giống Gaussian.

b.

Chart

Description automatically generated

Figure 11: Phân bố công suất điện từng năm

-Sự phân bố điện năng tiêu thụ hoạt động trong những năm đó rất giống nhau. Sự phân bố thực sự là hai phương thức với một đỉnh khoảng 0,3 KW và đỉnh khác khoảng 1,3 KW.

-Có một đuôi dài trên phân phối đến các giá trị KW cao hơn.

c.

Table

Description automatically generated

Figure 12: Phân phối dữ liệu tiêu thụ điện trong 12 tháng của năm 2007

-Nhìn chung, sự phân bố dữ liệu giống nhau mỗi tháng. Các trục của các biểu đồ dường như thẳng hàng (với các tỷ lệ tương tự) và có thể thấy rằng các đỉnh được dịch chuyển xuống trong các tháng ấm hơn và dịch chuyển lên trong các tháng lạnh hơn.

-Chúng ta có thể thấy phần đuôi dày hơn hoặc nổi bật hơn về phía giá trị KW lớn hơn trong những tháng mát mẻ hơn thừ tháng 12 đến tháng 3.