**Khảo sát chi tiết về tập dữ liệu "Electricity Load Forecasting"**

**I. Giới thiệu**

Tập dữ liệu "Electricity Load Forecasting" là một nguồn tài nguyên quan trọng dành cho các nhà khoa học dữ liệu, kỹ sư năng lượng và nhà nghiên cứu quan tâm đến việc dự đoán nhu cầu điện năng. Được cung cấp bởi Saurabh Shahane trên nền tảng Kaggle ([Electricity Load Forecasting](https://www.kaggle.com/datasets/saurabhshahane/electricity-load-forecasting/data))tập dữ liệu này chứa dữ liệu lịch sử về tải điện năng, cho phép phân tích xu hướng tiêu thụ và xây dựng các mô hình dự báo.

Dự báo tải điện đóng vai trò thiết yếu trong việc quản lý lưới điện hiện đại, đặc biệt khi thế giới đang chuyển đổi sang các nguồn năng lượng tái tạo như năng lượng mặt trời và gió, vốn phụ thuộc nhiều vào khả năng dự đoán chính xác nhu cầu để cân bằng cung cầu.

**II. Tổng quan về tập dữ liệu**

Tập dữ liệu được thiết kế để giải quyết bài toán dự báo tải điện năng – tức là dự đoán lượng điện tiêu thụ tại các thời điểm cụ thể trong tương lai. Đây là nhiệm vụ quan trọng trong ngành năng lượng, giúp các nhà quản lý điều chỉnh nguồn cung, tránh tình trạng thiếu điện hoặc dư thừa không cần thiết.

**Mục tiêu cụ thể:**

* **Dự báo ngắn hạn:** Dự đoán nhu cầu trong vài giờ hoặc vài ngày tới, hữu ích cho lập kế hoạch vận hành hàng ngày.
* **Dự báo dài hạn:** Dự đoán xu hướng trong vài tháng hoặc vài năm, hỗ trợ lập chiến lược đầu tư cơ sở hạ tầng năng lượng.

**III. Các thuộc tính quan trọng trong tập dữ liệu**

Tập dữ liệu "Electricity Load Forecasting" bao gồm **17 thuộc tính**, chia thành các nhóm: 1 cột thời gian, 1 cột tải điện, 12 cột thời tiết (4 yếu tố × 3 thành phố), và 3 cột liên quan đến ngày lễ/trường học. Dưới đây là mô tả các thuộc tính chính:

**1. Cột thời gian**

* **datetime:** Thời gian của mỗi phép đo (ví dụ: 2023-01-01 00:00:00), kiểu dữ liệu Datetime.

**2. Cột tải điện**

* **nat\_demand:** Tải điện quốc gia (National electricity load), đơn vị có thể là kW hoặc MW, kiểu dữ liệu Float.

**3. Cột thời tiết (12 thuộc tính)**

* **T2M\_[city]:** Nhiệt độ tại độ cao 2 mét (°C), ví dụ: T2M\_toc (Tocumen, Panama City).
* **QV2M\_[city]:** Độ ẩm tương đối tại 2 mét (%), ví dụ: QV2M\_san (Santiago City).
* **TQL\_[city]:** Lượng mưa lỏng (mm), ví dụ: TQL\_dav (David City).
* **W2M\_[city]:** Tốc độ gió tại 2 mét (m/s), ví dụ: W2M\_toc.
  + **Thành phố:** toc (Tocumen), san (Santiago), dav (David).

**4. Cột liên quan đến ngày lễ và trường học (3 thuộc tính)**

* **Holiday\_ID:** Số định danh duy nhất cho ngày lễ, kiểu Integer.
* **holiday:** Chỉ số nhị phân (1 = ngày lễ, 0 = ngày thường), kiểu Integer.
* **school:** Chỉ số nhị phân (1 = trong kỳ học, 0 = kỳ nghỉ), kiểu Integer

**IV. Một số bài toán liên quan**

Dưới đây là các bài toán tiêu biểu có thể giải quyết với tập dữ liệu:

**a. Dự báo tải điện ngắn hạn**

* **Mô tả:** Dự đoán tải điện trong vài giờ hoặc vài ngày.
* **Ứng dụng:** Quản lý lưới điện hàng ngày, tối ưu hóa vận hành nhà máy điện.
* **Phương pháp:** ARIMA, Random Forest, LSTM.

**b. Phân tích tác động của thời tiết đến tải điện**

* **Mô tả:** Nghiên cứu ảnh hưởng của thời tiết đến tiêu thụ điện.
* **Ứng dụng:** Cải thiện dự báo, đánh giá tác động biến đổi khí hậu.
* **Phương pháp:** Tương quan Pearson, hồi quy đa biến.

**c. Phát hiện bất thường trong tiêu thụ điện**

* **Mô tả:** Xác định các điểm bất thường trong tải điện.
* **Ứng dụng:** Phát hiện sự cố, quản lý rủi ro.
* **Phương pháp:** Isolation Forest, One-Class SVM.

**d. Dự báo tải điện dài hạn**

* **Mô tả:** Dự đoán xu hướng tiêu thụ điện trong vài tháng/năm.
* **Ứng dụng:** Lập kế hoạch cơ sở hạ tầng, chiến lược mua bán điện.
* **Phương pháp:** Hồi quy tuyến tính, Prophet.

**e. Phân tích ảnh hưởng của ngày lễ và kỳ học**

* **Mô tả:** Đánh giá tác động của ngày lễ/kỳ học đến tải điện.
* **Ứng dụng:** Cải thiện dự báo trong dịp đặc biệt.
* **Phương pháp:** t-test, hồi quy với biến giả.

**V. Phương pháp nghiên cứu**

Dự báo và phân tích tải điện có thể sử dụng các phương pháp sau:

* **Phân tích thống kê:** Tương quan (Pearson), t-test để khám phá mối quan hệ giữa các biến.
* **Hồi quy:**
  + Hồi quy tuyến tính: Dự đoán đơn giản.
  + Hồi quy đa biến: Kết hợp nhiều yếu tố.
  + Ridge/Lasso: Xử lý dữ liệu phức tạp.
* **Phân tích chuỗi thời gian:** ARIMA, SARIMA cho dự báo dựa trên thời gian.
* **Học máy:** Random Forest, XGBoost tối ưu hóa dự đoán với dữ liệu phức tạp.
* **Học sâu:** LSTM, GRU khai thác mẫu chuỗi thời gian dài hạn.
* **Phát hiện bất thường:** Isolation Forest, DBSCAN để tìm điểm bất thường.

**VI. Kết quả đạt được**

* **Dự báo ngắn hạn:** MAPE 2-5%, RMSE 50-200 MW (LSTM, XGBoost).
* **Dự báo dài hạn:** MAPE 5-10%, R² 0.8-0.9 (Prophet).
* **Phân tích thời tiết:** R² 0.6-0.8, nhiệt độ là yếu tố chính (hồi quy).
* **Phân tích ngày lễ/kỳ học:** Tải điện giảm 5-15% ngày lễ, R² 0.3-0.5 (t-test).
* **Phát hiện bất thường:** F1-Score 0.8-0.9 (Isolation Forest).  
  *Lưu ý: Kết quả phụ thuộc vào dữ liệu thực tế và mô hình cụ thể.*

**VII. Kết luận và hướng phát triển**

**1. Kết luận**

Tập dữ liệu "Electricity Load Forecasting" là nguồn tài nguyên giá trị cho nghiên cứu và ứng dụng dự báo nhu cầu điện năng. Với cấu trúc chuỗi thời gian và các biến thời tiết, ngày lễ, trường học, nó hỗ trợ xây dựng mô hình dự đoán chính xác và khám phá mẫu tiêu thụ năng lượng. Dù có thách thức như giá trị thiếu hoặc dữ liệu lớn, đây vẫn là công cụ đáng để khai thác.

**2. Hướng phát triển**

* Thu thập thêm dữ liệu (kinh tế, xã hội) để tăng độ chính xác.
* Phát triển mô hình lai (ARIMA-LSTM) để tối ưu hiệu suất.
* Triển khai dự báo thời gian thực tích hợp IoT.
* Nghiên cứu tác động của biến đổi khí hậu đến nhu cầu điện trong tương lai.

References

[Electricity Load Forecasting](https://www.kaggle.com/datasets/saurabhshahane/electricity-load-forecasting)