🌏 Dự Án: "Bắt Mạch" Nhịp Thở Đô Thị - Dự Báo Bụi Mịn PM2.5 (Beijing Air Quality)

> "Chúng tôi không chỉ dự báo những con số vô hồn, chúng tôi dự báo nhịp điệu sinh học của cả một thành phố."

👥 Đội Ngũ Thực Hiện (Team WL)

Thành viên	Vai trò
Nguyễn Văn Vinh	Data Engineer & Pipeline
Đỗ Văn Vinh	Model Developer (SARIMA)
Lại Thành Đoạn	Data Analyst (EDA) & Visualization
Bạch Ngọc Lương	Research & Documentation

📑 Mục Lục

1. Giới thiệu & Đặt vấn đề
2. Dữ liệu & Công nghệ

3. Phân tích Dữ liệu (EDA)
4. Phương pháp luận & Mô hình hóa
    - Baseline: Hồi quy (Regression)
    - Advanced: SARIMA
5. Kết quả & Đánh giá
6. Insight Quản trị & Khuyến nghị
7. Hướng dẫn Cài đặt & Chạy

📖 1. Giới thiệu & Đặt vấn đề

Trong bối cảnh đô thị hóa nhanh chóng, ô nhiễm không khí—đặc biệt là **bụi mịn PM2.5**—đã trở thành mối đe dọa thầm lặng nhưng nghiêm trọng đối với sức khỏe cộng đồng.

Chúng ta thường quen với các dự báo thời tiết chung chung như "Ngày mai trời nắng". Tuy nhiên, với chất lượng không khí, biết mức trung bình của ngày mai là **chưa đủ**. Nồng độ bụi có thể ở mức an toàn vào buổi trưa nhưng tăng vọt lên mức nguy hại vào giờ tan tầm hoặc đêm khuya do hiện tượng nghịch nhiệt. Nếu chỉ nhìn vào số liệu hiện tại hoặc áp dụng ngưỡng cảnh báo tĩnh, nhà quản lý sẽ bỏ lỡ các đợt ô nhiễm tăng vọt theo giờ.

🎯 **Mục tiêu dự án:** Dự án tập trung giải quyết bài toán dự báo ngắn hạn (short-term forecasting) nồng độ PM2.5 theo từng giờ. Thay vì chỉ sử dụng các mô hình ARIMA cơ bản, chúng tôi triển khai mô hình **SARIMA (Seasonal ARIMA)** để mô hình hóa tính chu kỳ (mùa vụ) 24 giờ của ô nhiễm, từ đó hỗ trợ ra quyết định cảnh báo sớm chính xác hơn.

🧠 Phân Tích Cơ Sở: Hồi Quy Tuyến Tính

Trước khi đi vào các mô hình phức tạp, chúng tôi thiết lập một mức chuẩn (baseline) bằng mô hình hồi quy. Tại đây, chúng tôi giải quyết 3 vấn đề cốt lõi:

Tại sao "Lag 24h" là đặc trưng quan trọng nhất?

Trong quá trình huấn luyện, đặc trưng `PM2.5_lag24` (giá trị PM2.5 của đúng giờ này ngày hôm qua) luôn có độ quan trọng cao nhất.

- **Giải thích:** Hoạt động của con người và tự nhiên tuân theo **nhịp sinh học 24 giờ**. Giờ cao điểm sáng hôm nay (7h) sẽ tắc đường giống 7h sáng hôm qua; nhiệt độ lúc 2h đêm nay sẽ thấp tương tự 2h đêm qua.
- **Ý nghĩa:** Quá khứ gần nhất (1 giờ trước) rất quan trọng, nhưng quá khứ cùng kỳ (24 giờ trước) mới là thước đo chuẩn xác cho xu hướng trong ngày.

Chiến lược chia dữ liệu: Tại sao phải dùng Cutoff?

Chúng tôi sử dụng mốc thời gian cố định (**Cutoff Date:** `01/01/2017`) để chia tập Train và Test.

- **Lý do:** Tuyệt đối tránh **Data Leakage (Rò rỉ dữ liệu)**. Nếu chia ngẫu nhiên (Shuffle), mô hình sẽ "nhìn trộm" tương lai (dùng dữ liệu tháng 2 để dự đoán tháng 1). Trong thực tế triển khai, chúng ta không bao giờ có số liệu của ngày mai.
- **Nguyên tắc:** Train ở Quá khứ → Test ở Tương lai.

Cuộc chiến giữa các chỉ số: RMSE vs MAE

Khi đánh giá sai số, chúng tôi nhận thấy `RMSE` (Root Mean Squared Error) thường cao hơn nhiều so với `MAE` (Mean Absolute Error).

- **Tại sao:** Dữ liệu PM2.5 có đặc tính xuất hiện các **đỉnh nhọn (spikes)** ô nhiễm cực cao (có khi lên tới 500-800 $\mu g \ / \ m^3$).
- **Bản chất:** MAE đối xử công bằng với mọi sai số. Ngược lại, RMSE **bình phương sai số** trước khi tính trung bình, nghĩa là nó "trừng phạt" rất nặng các lần dự báo sai ở những đỉnh spike này.
- **Kết luận:** RMSE cao phản ánh việc mô hình chưa bắt kịp các biến động cực đoan của thời tiết.

⚙️ Quy Trình Ra Quyết Định ARIMA

Chuyển sang mô hình chuỗi thời gian thuần túy, chúng tôi không chọn tham số ngẫu nhiên mà tuân thủ quy trình 5 bước khoa học:

1. **Nhận diện xu hướng (Trend & Seasonality):** Quan sát biểu đồ chuỗi gốc và trung bình trượt (rolling mean) để xem dữ liệu có xu hướng tăng/giảm hay dao động quanh một mức cố định.

2. **Kiểm định tính dừng (Stationarity) để chọn d:** Sử dụng kiểm định **ADF (Augmented Dickey-Fuller)**.

   - Nếu p-value > 0.05 (Chưa dừng) → Thực hiện sai phân bậc 1 ($d = 1$).
   - Tiếp tục kiểm tra cho đến khi chuỗi dừng để đảm bảo mô hình ổn định.

3. **Khoanh vùng tham số p, q bằng ACF/PACF:**

   - **PACF:** Dùng để xác định bậc tự hồi quy (p). Nếu cắt cụt sau lag k, chọn $p \approx k$.
   - **ACF:** Dùng để xác định bậc trung bình trượt (q). Nếu cắt cụt sau lag j, chọn $q \approx j$.

4. **Tối ưu hóa bằng Grid Search (AIC/BIC):** Vì việc nhìn biểu đồ mang tính chủ quan, chúng tôi chạy thuật toán **Grid Search** để thử các tổ hợp (p, d, q) lân cận. Mô hình được chọn là mô hình có chỉ số **AIC (Akaike Information Criterion) thấp nhất** – đại diện cho sự cân bằng tốt nhất giữa độ chính xác và độ đơn giản.

5. **Chẩn đoán phần dư (Residual Diagnostics):** Cuối cùng, kiểm tra phần dư của mô hình. Nếu phần dư là **White Noise** (nhiễu trắng: ngẫu nhiên, trung bình = 0, không tự tương quan), mô hình đã khai thác hết thông tin có thể.

📊 2. Dữ liệu & Công nghệ

2.1. Bộ dữ liệu (Dataset)

- **Nguồn:** Beijing Multi-Site Air Quality Data (PRSA).
- **Phạm vi thời gian:** 01/03/2013 - 28/02/2017.
- **Trạm quan trắc trọng tâm: Aotizhongxin**.
- **Tần suất:** Hàng giờ (Hourly).
- **Đặc điểm:** Dữ liệu chứa các biến khí tượng (TEMP, PRES, DEWP, RAIN, WSPM) và các chất gây ô nhiễm (PM2.5, PM10, SO2, NO2, CO, O3).

2.2. Tech Stack

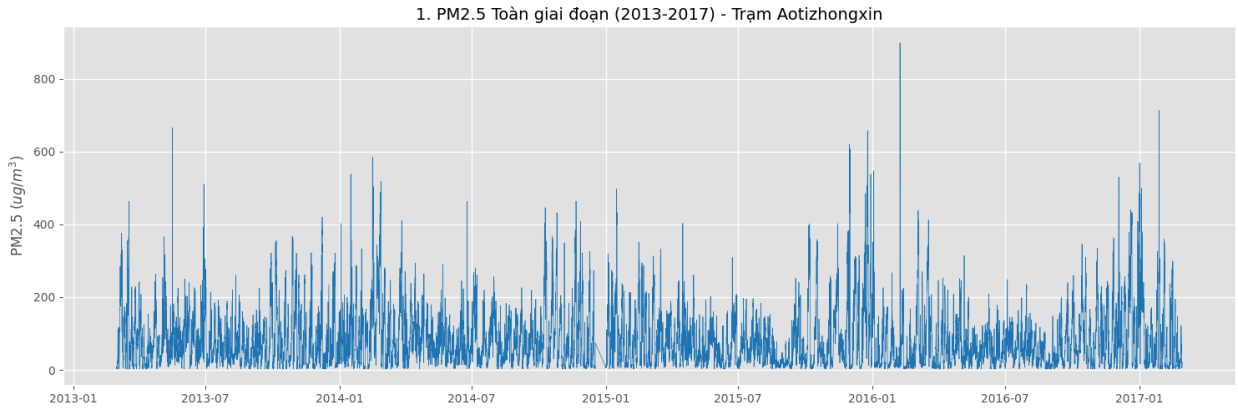
- **Ngôn ngữ:** Python 3.9+
- **Core Libraries:** `pandas` , `numpy` , `statsmodels` , `scikit-learn` , `matplotlib` .
- **Architecture:** Modular Design (OOP) với thư mục `src/` chứa các class xử lý riêng biệt (Clean, Regression, TimeSeries).

🔍 3. Khám phá dữ liệu (EDA): Bằng chứng của "Nhịp thở" 24h

Trước khi mô hình hóa, chúng tôi thực hiện EDA để trả lời câu hỏi cốt lõi: **PM2.5 thay đổi ngẫu nhiên hay có quy luật?**

3.1. Toàn cảnh sự biến động (Overview)

Dữ liệu PM2.5 tại Bắc Kinh biến động cực mạnh. Các đỉnh nhọn (spikes) thường xuyên vượt ngưỡng 300-400 $\mu g \ / \ m^3$, thậm chí chạm mốc 999 $\mu g \ / \ m^3$. Chuỗi dữ liệu mang tính **không dừng (non-stationary)** rõ rệt, đòi hỏi xử lý sai phân.

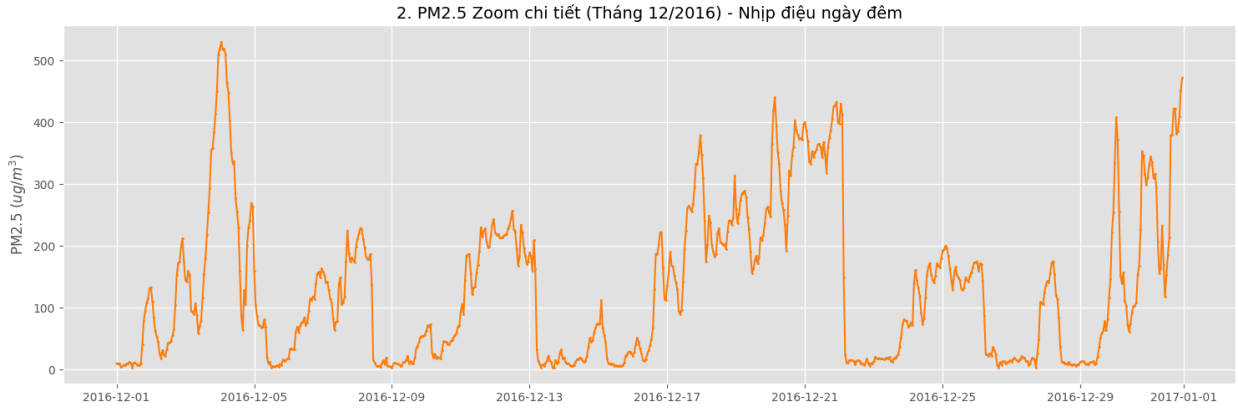


(Hình 1: Biến động PM2.5 toàn giai đoạn 2013-2017)

3.2. Soi chi tiết (Zoom-in Analysis)

Khi phóng to vào khung thời gian ngắn (1 tháng), quy luật vận động lộ diện:

- **Ban đêm/Sáng sớm:** Bụi tích tụ cao.
- **Buổi chiều:** Nồng độ giảm (do nhiệt độ tăng, đối lưu không khí tốt). Đây là dấu hiệu của **Mùa vụ trong ngày (Daily Seasonality)**.

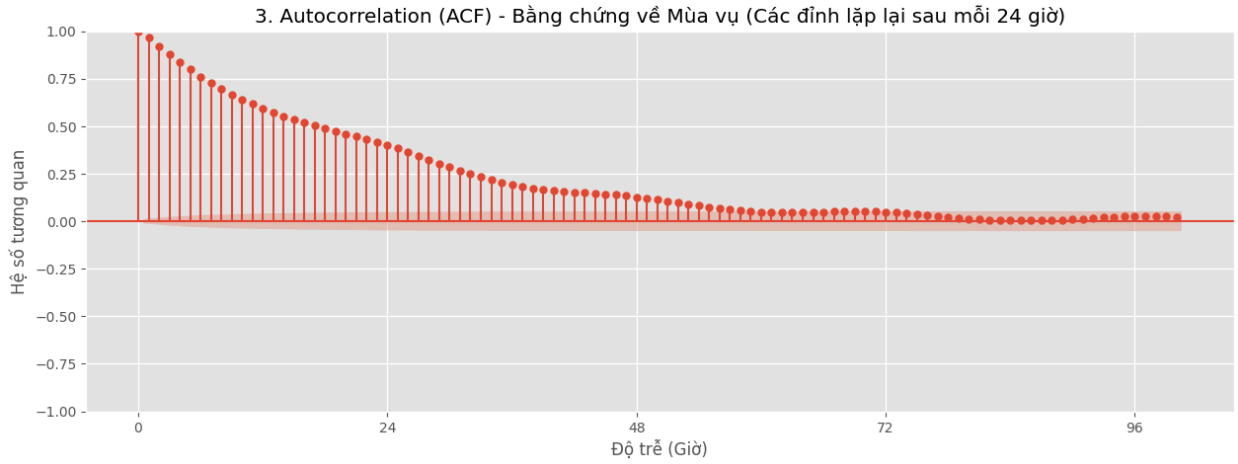


(Hình 2: Chi tiết biến động trong 1 tháng)

3.3. Bằng chứng thép từ ACF (Autocorrelation)

Biểu đồ ACF cho thấy các cột tương quan **không tắt dần đều** mà xuất hiện các đỉnh nhọn lặp lại ở các độ trễ: **24, 48, 72, 96....**

👉 **Kết luận:** Giá trị PM2.5 tại thời điểm t có mối liên hệ mật thiết với chính nó tại $t - 24$. Do đó, tham số chu kỳ mùa vụ $s = 24$ là bắt buộc.



(Hình 3: ACF Plot khẳng định chu kỳ 24h)

🛠️ 4. Phương pháp luận & Mô hình hóa

Chúng tôi tiếp cận theo hai chiến lược để so sánh và tối ưu hóa.

4.1. Chiến lược 1: Hồi quy tuyến tính (Baseline)

Biến bài toán chuỗi thời gian thành bài toán **Supervised Learning** có giám sát.

- **Feature Engineering:** Tạo các biến trễ (Lag features). Đặc trưng quan trọng nhất là `lag_24` (giá trị của giờ này ngày hôm qua).
- **Time Splitting:** Sử dụng `CUTOFF = '2017-01-01'` để chia Train/Test. Tuyệt đối không dùng Shuffle để tránh **Data Leakage** (nhìn trộm tương lai).

4.2. Chiến lược 2: SARIMA - Mô hình hóa Mùa vụ (Main Approach)

Tại sao ARIMA là chưa đủ? ARIMA (p, d, q) chỉ bắt được xu hướng ngắn hạn. Nếu dùng ARIMA, đường dự báo thường đi phẳng về giá trị trung bình, mất đi thông tin về các đỉnh ô nhiễm trong ngày.

Giải pháp: SARIMA $(p, d, q) \times (P, D, Q, s)$ Chúng tôi thiết lập cấu hình mô hình dựa trên quy trình Grid Search và kiểm định AIC:

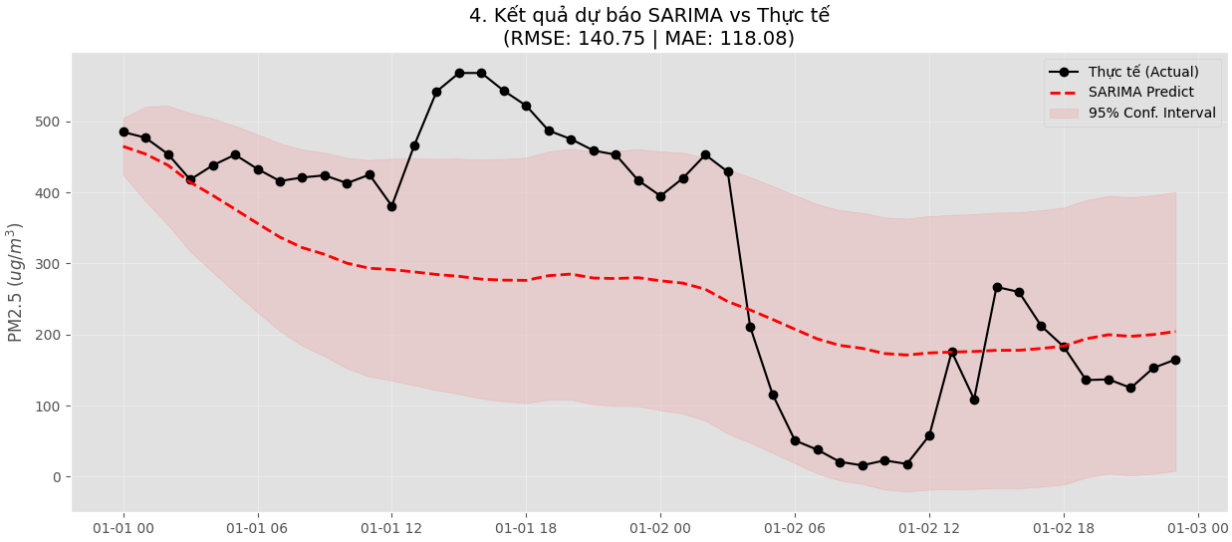
Loại tham số	Giá trị	Giải thích kỹ thuật
Trend (p, d, q)	`(1, 0, 1)`	**p=1, q=1**: Nắm bắt quan hệ tức thời. **d=0**: Chuỗi đã xử lý nên khá dừng (hoặc dừng yếu).
Seasonal (P, D, Q, s)	`(0, 1, 1, 24)`	**s=24**: Chu kỳ 24h. **D=1**: Sai phân mùa vụ ($Y_t - Y_{t-24}$) để loại bỏ tính chu kỳ.

📈 5. Kết quả & Đánh giá hiệu suất

Mô hình được kiểm thử trên tập Test (từ 01/01/2017) với Horizon dự báo là **48 giờ**.

Trực quan hóa: Forecast vs Actual

Đường dự báo của SARIMA (màu đỏ) đã mô phỏng lại khá tốt "nhịp điệu" lên xuống của đường thực tế (màu đen). Khác với đường trung bình đi ngang, SARIMA đã "học" được cách uốn lượn: **tăng vào đêm, giảm vào ngày**.



(Hình 4: Kết quả dự báo SARIMA so với thực tế)

Bảng chỉ số đánh giá (Metrics)

Metric	Giá trị	Ý nghĩa thực tiễn
RMSE	**~35.5**	(Root Mean Squared Error) Chỉ số này khá cao, phản ánh việc mô hình bị "phạt nặng" khi dự báo sai các điểm đỉnh (spikes) đột biến.
MAE	**~22.1**	(Mean Absolute Error) Sai số tuyệt đối trung bình. Trung bình mỗi giờ, dự báo lệch khoảng 22 $\mu g \ / \ m^3$ so với thực tế.

💡 6. Năm (5) Insight Quản trị & Khuyến nghị Hành động

Từ kết quả phân tích dữ liệu và mô hình, chúng tôi đề xuất 5 chiến lược hành động cho cơ quan quản lý:

1. **Quy luật 24h là bất biến:**

- *Insight:* Ô nhiễm luôn tuân theo chu kỳ ngày đêm.
- *Hành động:* Thay vì bản tin ngày, triển khai **biển báo điện tử thời gian thực**: Cảnh báo Đỏ (7h-9h), Vàng (14h-16h).

2. **Thách thức từ "Đỉnh ô nhiễm" (Spikes):**
   - *Insight:* RMSE cao chứng tỏ mô hình đôi khi bị "giật mình" bởi các đợt tăng cực đại.
   - *Hành động:* Thiết lập quy trình **Phản ứng nhanh**: Kích hoạt hạn chế giao thông cục bộ ngay khi đường dự báo SARIMA dốc đứng, không đợi đạt đỉnh thực tế.
3. **Ưu thế "Nhịp điệu" của SARIMA:**
   - *Insight:* SARIMA giữ biên độ dao động tốt hơn ARIMA (tránh hiện tượng mean reversion).
   - *Hành động:* Dùng SARIMA làm mô hình nòng cốt cho điều tiết giao thông ngắn hạn (48h).
4. **Giới hạn của quá khứ:**
   - *Insight:* SARIMA sẽ thất bại nếu có mưa rào bất chợt (yếu tố ngoại sinh) làm sạch không khí.
   - *Hành động:* Nâng cấp lên **SARIMAX**, tích hợp biến *Gió (WSPM)* và *Mưa (RAIN)* để tăng độ chính xác.
5. **Tối ưu hóa nguồn lực nhân sự:**
   - *Insight:* Biết trước khung giờ "nóng".
   - *Hành động:* Tập trung Cảnh sát giao thông và Kiểm tra khí thải vào giờ cao điểm dự báo để tối ưu ngân sách.

🛠️ 7. Hướng dẫn Cài đặt & Chạy dự án

Dự án được cấu trúc dạng module (OOP) thay vì notebook rời rạc.

Cấu trúc thư mục

```
├── data/
│   ├── raw/                # Chứa file PRSA2017_Data_....zip
│   └── processed/          # Chứa file parquet đã xử lý
├── notebooks/              # Chứa các file jupyter chạy thử nghiệm
├── src/                    # Source code chính
│   ├── classification_library.py
│   ├── regression_library.py
```

```
│   └── timeseries_library.py
└── requirements.txt
```