**Báo cáo Phân tích Khám phá Dữ liệu (EDA) - Dự báo Lượng mưa tại TP.HCM**

* **Mục tiêu dự án:** Xây dựng mô hình dự báo lượng mưa hàng ngày tại TP.HCM dựa trên dữ liệu khí tượng lịch sử.

**Phần 1: Tổng quan và Chất lượng Dữ liệu**

**1.1. Giới thiệu Bộ dữ liệu**  
**Dữ liệu NASA POWER API cho Thành phố Hồ Chí Minh**

**a) Chức năng (nội dung) của bộ dữ liệu là gì?**

Bộ dữ liệu này chứa thông tin khí tượng và bề mặt đất hàng ngày cho Thành phố Hồ Chí Minh, được thu thập tự động thông qua API NASA Prediction Of Worldwide Energy Resources (POWER). Dữ liệu bao gồm nhiều tham số quan trọng như:

* **Nhiệt độ**: trung bình, tối thiểu, tối đa, điểm sương, bầu ướt, nhiệt độ bề mặt đất
* **Độ ẩm**: tương đối, tuyệt đối
* **Lượng mưa**: mục tiêu dự đoán chính
* **Áp suất bề mặt**
* **Gió**: tốc độ trung bình, tối đa, hướng gió ở các độ cao khác nhau
* **Bức xạ mặt trời**: bức xạ sóng dài xuống bề mặt
* **Độ ẩm đất**: bề mặt, vùng rễ, toàn bộ mặt cắt

Dữ liệu được cung cấp cho mỗi ngày trong khoảng thời gian được chỉ định từ 01/01/2000 đến 30/04/2025, phục vụ mục tiêu dự đoán thời tiết cho toàn thành phố.

**b) Nguồn gốc của bộ dữ liệu?**

Dữ liệu gốc được cung cấp bởi dự án NASA POWER. Dữ liệu cụ thể trong bộ dữ liệu này được thu thập bằng cách gọi đến API công khai của NASA POWER:

* **Nguồn dữ liệu:** NASA POWER (Prediction Of Worldwide Energy Resources)
* **API Endpoint:** https://power.larc.nasa.gov/api/temporal/daily/point
* **Phương pháp:** Single request crawl với 20 parameters (trong giới hạn API)
* **Vị trí:** Tọa độ trung tâm TP.HCM (lat: 10.78, lon: 106.7)

**c) Cấu trúc các tập tin của bộ dữ liệu? Ý nghĩa từng tập tin.**

Dữ liệu sau khi thu thập được lưu trữ trong thư mục nasa\_power\_hcmc\_data với cấu trúc đơn giản:

1. **Tệp dữ liệu chính:**
   * **Tên tệp:** hcmc\_weather\_data\_{Ngày\_Bắt\_Đầu}\_{Ngày\_Kết\_Thúc}.csv
   * **Nội dung:** Chứa dữ liệu hàng ngày của tất cả 20 tham số khí tượng cho TP.HCM trong khoảng thời gian xác định. Mỗi hàng đại diện cho một ngày.

**d) Các thông tin nào mà các record (bản ghi) lưu trữ?**

Mỗi record (hàng trong tệp CSV) tương ứng với dữ liệu của một ngày cho TP.HCM và chứa các cột sau:

**Thông tin địa lý và thời gian:**

* **DATE:** Ngày ghi nhận dữ liệu (định dạng datetime)
* **LATITUDE:** Vĩ độ điểm đại diện TP.HCM (10.78)
* **LONGITUDE:** Kinh độ điểm đại diện TP.HCM (106.7)

**Nhóm 1 - Nhiệt độ, độ ẩm, mưa, áp suất (11 chỉ số):**

* **PRECTOTCORR:** Lượng mưa đã hiệu chỉnh (mm/ngày) - *Biến mục tiêu*
* **QV2M:** Độ ẩm tuyệt đối ở 2m (g/kg)
* **RH2M:** Độ ẩm tương đối ở 2m (%)
* **T2M:** Nhiệt độ trung bình ở 2m (°C)
* **T2MDEW:** Điểm sương/sương giá ở 2m (°C)
* **T2MWET:** Nhiệt độ bầu ướt ở 2m (°C)
* **T2M\_MAX:** Nhiệt độ tối đa ở 2m (°C)
* **TS:** Nhiệt độ bề mặt đất (°C)
* **T2M\_MIN:** Nhiệt độ tối thiểu ở 2m (°C)
* **ALLSKY\_SFC\_LW\_DWN:** Bức xạ sóng dài xuống bề mặt (MJ/m²/ngày)
* **PS:** Áp suất bề mặt (kPa)

**Nhóm 2 - Gió và độ ẩm đất (9 chỉ số):**

* **WS10M:** Tốc độ gió ở 10m (m/s)
* **WD10M:** Hướng gió ở 10m (Độ)
* **WS10M\_MAX:** Tốc độ gió tối đa ở 10m (m/s)
* **WS2M\_MAX:** Tốc độ gió tối đa ở 2m (m/s)
* **WS2M:** Tốc độ gió ở 2m (m/s)
* **WD2M:** Hướng gió ở 2m (Độ)
* **GWETPROF:** Độ ẩm đất theo mặt cắt (tỷ lệ 0-1)
* **GWETTOP:** Độ ẩm bề mặt đất (tỷ lệ 0-1)
* **GWETROOT:** Độ ẩm vùng rễ (tỷ lệ 0-1)

**Ghi chú quan trọng:**

* Giá trị -999 biểu thị dữ liệu thiếu hoặc không thể tính toán
* Dữ liệu được crawl trong 1 lần duy nhất với 20 parameters (tuân thủ giới hạn API)
* Tổng cộng có 20 weather parameters + 3 metadata columns cho việc dự đoán thời tiết TP.HCM

Tham khảo

* **NASA POWER Data Access Viewer:** [https://power.larc.nasa.gov/data-access-viewer/](https://www.google.com/url?q=https%3A%2F%2Fpower.larc.nasa.gov%2Fdata-access-viewer%2F)
* **NASA POWER Documentation:** [https://power.larc.nasa.gov/docs/](https://www.google.com/url?q=https%3A%2F%2Fpower.larc.nasa.gov%2Fdocs%2F)
* **NASA POWER API Documentation:** [https://power.larc.nasa.gov/docs/services/api/](https://www.google.com/url?q=https%3A%2F%2Fpower.larc.nasa.gov%2Fdocs%2Fservices%2Fapi%2F)

**1.2. Kiểm tra Chất lượng Dữ liệu**  
🚀 STARTING COMPREHENSIVE DATA QUALITY ASSESSMENT

================================================================================

🔍 DATA QUALITY ASSESSMENT INITIALIZED

============================================================

📊 Dataset Overview:

- Shape: (9252, 23)

- Target Variable: Lượng mưa

- Date Column: Ngày

- Memory Usage: 2.20 MB

============================================================

📋 1. BASIC DATASET INFORMATION

============================================================

📏 Kích thước Dataset:

- Số dòng (observations): 9,252

- Số cột (variables): 23

- Tổng ô dữ liệu: 212,796

📊 Phân tích kiểu dữ liệu:

- float64: 22 columns (95.7%)

- datetime64[ns]: 1 columns (4.3%)

📂 Phân loại cột:

- Numerical columns: 22

- Categorical columns: 0

- Datetime columns: 1

============================================================

❓ 2. MISSING VALUES ANALYSIS

============================================================

📊 Tổng quan Missing Data:

- Tổng số ô thiếu dữ liệu: 0

- Tỷ lệ thiếu dữ liệu tổng thể: 0.00%

✅ Không có cột nào thiếu dữ liệu!

✅ Biến mục tiêu 'Lượng mưa' không thiếu dữ liệu

============================================================

🔄 3. DUPLICATE DATA ANALYSIS

============================================================

📊 Dòng trùng lặp hoàn toàn:

- Số dòng trùng lặp: 0

- Tỷ lệ trùng lặp: 0.00%

📊 Dòng trùng lặp (loại trừ tọa độ):

- Số dòng trùng lặp: 0

- Tỷ lệ trùng lặp: 0.00%

🎯 Phân tích giá trị trùng lặp của biến mục tiêu:

- Giá trị phổ biến nhất: 0.0

- Số lần xuất hiện: 2,381 (25.73%)

📊 Tỷ lệ ngày không mưa: 25.73%

============================================================

🔍 4. DATA CONSISTENCY CHECK

============================================================

🔍 Kiểm tra giá trị âm cho các biến phải dương:

✅ Lượng mưa: Không có giá trị âm

✅ Độ ẩm tương đối 2m: Không có giá trị âm

✅ Tốc độ gió 10m: Không có giá trị âm

✅ Tốc độ gió 2m: Không có giá trị âm

✅ Tốc độ gió tối đa 10m: Không có giá trị âm

✅ Tốc độ gió tối đa 2m: Không có giá trị âm

🔍 Kiểm tra phạm vi độ ẩm (0-100%):

✅ Độ ẩm tương đối 2m: Tất cả giá trị trong phạm vi [0,100]

🔍 Kiểm tra phạm vi hướng gió (0-360 độ):

✅ Hướng gió 10m: Tất cả giá trị trong phạm vi [0,360]

✅ Hướng gió 2m: Tất cả giá trị trong phạm vi [0,360]

🔍 Kiểm tra tính nhất quán nhiệt độ (min ≤ avg ≤ max):

✅ Tất cả dữ liệu nhiệt độ nhất quán

🔍 Kiểm tra giá trị cực đoan không thực tế:

✅ Lượng mưa: Không có giá trị cực đoan

✅ Nhiệt độ: Không có giá trị cực đoan

============================================================

📅 5. TEMPORAL DATA QUALITY CHECK

============================================================

✅ Tất cả ngày đều hợp lệ

📊 Phạm vi thời gian:

- Ngày bắt đầu: 2000-01-01

- Ngày kết thúc: 2025-04-30

- Tổng khoảng: 9,251 ngày (25.3 năm)

✅ Chuỗi thời gian liên tục, không thiếu ngày nào

✅ Không có ngày trùng lặp

🏆 OVERALL DATA QUALITY SCORE: 100.0/100

📊 Quality Level: 🟢 EXCELLENT

💡 Recommendation: Dữ liệu có chất lượng rất tốt, có thể tiến hành phân tích ngay.

📋 Chi tiết các vấn đề:

- Missing Data: 0.00%

- Duplicate Rows: 0.00%

- Consistency Issues: 0

- Invalid Dates: 0

- Missing Dates: 0

✅ DATA QUALITY ASSESSMENT COMPLETED

**1.3. Phân tích Phân phối Biến (Variable Distribution Analysis)**  
🚀 STARTING COMPREHENSIVE DISTRIBUTION ANALYSIS

================================================================================

📊 DISTRIBUTION ANALYSIS INITIALIZED

============================================================

- Total Features: 22

- Analysis Features: 20

- Target Variable: Lượng mưa

- Sample Size: 9,252

================================================================================

📋 5. DISTRIBUTION ANALYSIS SUMMARY REPORT

================================================================================

================================================================================

📈 1. DESCRIPTIVE STATISTICS SUMMARY

================================================================================

📊 Comprehensive Descriptive Statistics:

========================================================================================================================

🔢 Basic Statistics:

Feature Count Mean Median Mode Std Min Max

Lượng mưa 9252 3.9293 0.65 0.00 7.6473 0.00 143.31

Độ ẩm tuyệt đối 2m 9252 17.2692 18.15 19.11 2.3491 8.14 21.47

Độ ẩm tương đối 2m 9252 77.6260 82.49 88.46 12.4275 40.84 95.49

Nhiệt độ 2m 9252 27.4593 27.01 26.91 1.9548 19.89 34.30

Điểm sương 2m 9252 22.5162 23.50 24.35 2.4122 10.82 26.26

Nhiệt độ bầu ướt 2m 9252 24.9877 25.30 25.27 1.4066 17.23 28.29

Nhiệt độ tối đa 2m 9252 32.4488 31.23 30.24 3.5078 23.01 42.93

Nhiệt độ bề mặt đất 9252 28.0657 27.18 27.10 2.6741 19.89 37.19

Nhiệt độ tối thiểu 2m 9252 23.5858 23.84 24.12 1.8488 14.68 29.20

Bức xạ sóng dài xuống 9252 10.0590 10.13 10.16 0.2900 8.15 10.92

Áp suất bề mặt 9252 100.3785 100.37 100.29 0.1964 99.64 101.22

Tốc độ gió 10m 9252 2.9932 2.83 2.42 1.0663 0.68 7.08

Hướng gió 10m 9252 168.1869 147.05 125.20 76.0101 0.10 360.00

Tốc độ gió tối đa 10m 9252 4.8657 4.64 3.48 1.7382 1.22 10.96

Tốc độ gió tối đa 2m 9252 3.4204 3.28 2.37 1.2160 0.74 7.54

Tốc độ gió 2m 9252 2.0195 1.89 1.33 0.7876 0.45 4.82

Hướng gió 2m 9252 168.4265 146.90 246.10 76.8419 0.20 359.80

Độ ẩm đất mặt cắt 9252 0.6555 0.64 0.48 0.1342 0.48 0.99

Độ ẩm đất bề mặt 9252 0.6233 0.67 0.78 0.1956 0.16 0.97

Độ ẩm đất vùng rễ 9252 0.6395 0.63 0.47 0.1307 0.47 0.99

📐 Shape & Variability Statistics:

Feature Skewness Kurtosis CV(%) Range IQR

Lượng mưa 4.3368 34.0353 194.6239 143.31 4.5100

Độ ẩm tuyệt đối 2m -0.8629 -0.0783 13.6027 13.33 3.5400

Độ ẩm tương đối 2m -0.7455 -0.7241 16.0095 54.65 20.2625

Nhiệt độ 2m 0.6669 0.3792 7.1189 14.41 2.2500

Điểm sương 2m -1.1423 0.8762 10.7131 15.44 3.3700

Nhiệt độ bầu ướt 2m -1.3134 2.3580 5.6292 11.06 1.4500

Nhiệt độ tối đa 2m 0.8089 -0.3818 10.8103 19.92 4.8900

Nhiệt độ bề mặt đất 0.9042 0.1533 9.5282 17.30 3.2700

Nhiệt độ tối thiểu 2m -0.7688 1.4112 7.8385 14.52 1.8600

Bức xạ sóng dài xuống -1.5752 3.5724 2.8825 2.77 0.2600

Áp suất bề mặt 0.3026 0.2664 0.1957 1.58 0.2500

Tốc độ gió 10m 0.5451 -0.2745 35.6239 6.40 1.5725

Hướng gió 10m 0.0564 -0.7457 45.1938 359.90 122.5000

Tốc độ gió tối đa 10m 0.3725 -0.7643 35.7237 9.74 2.7600

Tốc độ gió tối đa 2m 0.3358 -0.7456 35.5521 6.80 1.9100

Tốc độ gió 2m 0.5471 -0.3738 38.9988 4.37 1.1800

Hướng gió 2m 0.0719 -0.7949 45.6234 359.60 124.6000

Độ ẩm đất mặt cắt 0.3658 -0.9798 20.4717 0.51 0.2300

Độ ẩm đất bề mặt -0.5643 -0.7362 31.3783 0.81 0.2900

Độ ẩm đất vùng rễ 0.3584 -0.9543 20.4419 0.52 0.2300

📊 Percentile Statistics:

Feature P5 Q1 Median Q3 P95

Lượng mưa 0.0000 0.0000 0.65 4.5100 18.4535

Độ ẩm tuyệt đối 2m 12.7655 15.5800 18.15 19.1200 19.9400

Độ ẩm tương đối 2m 54.7455 67.4375 82.49 87.7000 91.4200

Nhiệt độ 2m 24.9000 26.2300 27.01 28.4800 31.2900

Điểm sương 2m 17.7300 20.9800 23.50 24.3500 25.0400

Nhiệt độ bầu ướt 2m 22.1155 24.4300 25.30 25.8800 26.7200

Nhiệt độ tối đa 2m 28.3100 29.9100 31.23 34.8000 39.4000

Nhiệt độ bề mặt đất 24.9500 26.3000 27.18 29.5700 33.5300

Nhiệt độ tối thiểu 2m 19.9655 22.7800 23.84 24.6400 26.3600

Bức xạ sóng dài xuống 9.4500 9.9700 10.13 10.2300 10.3900

Áp suất bề mặt 100.0700 100.2500 100.37 100.5000 100.7245

Tốc độ gió 10m 1.5300 2.1500 2.83 3.7225 4.9600

Hướng gió 10m 41.7550 116.8000 147.05 239.3000 272.4450

Tốc độ gió tối đa 10m 2.4300 3.4400 4.64 6.2000 7.9100

Tốc độ gió tối đa 2m 1.6800 2.4400 3.28 4.3500 5.5400

Tốc độ gió 2m 0.9500 1.3900 1.89 2.5700 3.4700

Hướng gió 2m 42.5000 115.7000 146.90 240.3000 274.1450

Độ ẩm đất mặt cắt 0.4800 0.5300 0.64 0.7600 0.8900

Độ ẩm đất bề mặt 0.2600 0.4900 0.67 0.7800 0.8700

Độ ẩm đất vùng rễ 0.4700 0.5100 0.63 0.7400 0.8700

================================================================================

🎯 2. TARGET VARIABLE DEEP ANALYSIS: Lượng mưa

================================================================================

📈 Basic Descriptive Statistics:

- Count: 9,252

- Mean: 3.9293 mm

- Median: 0.6500 mm

- Mode: 0.0000 mm

- Standard Deviation: 7.6473 mm

- Variance: 58.4818

- Coefficient of Variation: 194.62%

📊 Range and Percentiles:

- Minimum: 0.0000 mm

- 5th Percentile: 0.0000 mm

- 25th Percentile (Q1): 0.0000 mm

- 50th Percentile (Median): 0.6500 mm

- 75th Percentile (Q3): 4.5100 mm

- 95th Percentile: 18.4535 mm

- Maximum: 143.3100 mm

- Range: 143.3100 mm

- IQR: 4.5100 mm

📐 Shape Statistics:

- Skewness: 4.3368

→ Interpretation: Highly right-skewed (lệch phải mạnh)

- Kurtosis: 34.0353

→ Interpretation: Leptokurtic (nhọn hơn normal)

🌧️ Vietnamese Meteorological Standards Classification (24h):

Based on Vietnamese National Weather Service Standards

- Không mưa (No Rain): 2,381 days (25.73%)

- Mưa lượng không đáng kể (0-0.6mm): 2,186 days (23.63%)

- Mưa nhỏ (0.6-6.0mm): 2,783 days (30.08%)

- Mưa (6.0-16.0mm): 1,304 days (14.09%)

- Mưa vừa (16.0-50.0mm): 568 days (6.14%)

- Mưa to (50.0-100.0mm): 27 days (0.29%)

- Mưa rất to (>100.0mm): 3 days (0.03%)

================================================================================

🔬 3. STATISTICAL DISTRIBUTION SHAPE TESTS

================================================================================

📊 Normality Test Results Summary:

====================================================================================================

Feature Sample\_Size Shapiro\_Normal DAgostino\_Normal JB\_Normal Skew\_Normal Kurt\_Normal AD\_Normal

Lượng mưa 5000 No No No No No No

Độ ẩm tuyệt đối 2m 5000 No No No No Yes No

Độ ẩm tương đối 2m 5000 No No No No No No

Nhiệt độ 2m 5000 No No No No No No

Điểm sương 2m 5000 No No No No No No

Nhiệt độ bầu ướt 2m 5000 No No No No No No

Nhiệt độ tối đa 2m 5000 No No No No No No

Nhiệt độ bề mặt đất 5000 No No No No No No

Nhiệt độ tối thiểu 2m 5000 No No No No No No

Bức xạ sóng dài xuống 5000 No No No No No No

Áp suất bề mặt 5000 No No No No No No

Tốc độ gió 10m 5000 No No No No No No

Hướng gió 10m 5000 No No No No No No

Tốc độ gió tối đa 10m 5000 No No No No No No

Tốc độ gió tối đa 2m 5000 No No No No No No

Tốc độ gió 2m 5000 No No No No No No

Hướng gió 2m 5000 No No No No No No

Độ ẩm đất mặt cắt 5000 No No No No No No

Độ ẩm đất bề mặt 5000 No No No No No No

Độ ẩm đất vùng rễ 5000 No No No No No No

🎯 Detailed Results for Lượng mưa:

- Shapiro-Wilk: statistic=0.561715, p-value=2.51e-77

- D'Agostino: statistic=4287.883628, p-value=0.00e+00

- Jarque-Bera: statistic=147454.458430, p-value=0.00e+00

- Skewness test: statistic=76.154761, p-value=0.00e+00

- Kurtosis test: statistic=52.413337, p-value=0.00e+00

- Anderson-Darling: statistic=682.307129, critical\_5%=0.786000

📈 Summary:

- Features possibly normal: 0 / 20

- No features follow normal distribution

📊 Non-parametric Test Results:

========================================================================================================================

🔍 Mann-Whitney U Test (Rain vs No Rain):

Feature MW\_Statistic MW\_P\_Value MW\_Significant

Độ ẩm tuyệt đối 2m 1432603.0 0.000000 Yes

Độ ẩm tương đối 2m 2215404.5 0.000000 Yes

Nhiệt độ 2m 9566249.5 0.000000 Yes

Điểm sương 2m 1435513.0 0.000000 Yes

Nhiệt độ bầu ướt 2m 3662549.5 0.000000 Yes

Nhiệt độ tối đa 2m 12665357.5 0.000000 Yes

Nhiệt độ bề mặt đất 10593322.0 0.000000 Yes

Nhiệt độ tối thiểu 2m 4852259.0 0.000000 Yes

Bức xạ sóng dài xuống 3555480.0 0.000000 Yes

Áp suất bề mặt 12486352.5 0.000000 Yes

Tốc độ gió 10m 8660874.0 0.000019 Yes

Hướng gió 10m 4675024.5 0.000000 Yes

Tốc độ gió tối đa 10m 10149620.5 0.000000 Yes

Tốc độ gió tối đa 2m 9879548.0 0.000000 Yes

Tốc độ gió 2m 8818993.0 0.000000 Yes

Hướng gió 2m 4742420.0 0.000000 Yes

Độ ẩm đất mặt cắt 4069286.5 0.000000 Yes

Độ ẩm đất bề mặt 3076642.0 0.000000 Yes

Độ ẩm đất vùng rễ 3544243.5 0.000000 Yes

🔍 Kruskal-Wallis Test (Multiple Rainfall Groups):

Feature KW\_Statistic KW\_P\_Value KW\_Significant

Độ ẩm tuyệt đối 2m 4715.242252 0.0 Yes

Độ ẩm tương đối 2m 4500.553097 0.0 Yes

Nhiệt độ 2m 638.735882 0.0 Yes

Điểm sương 2m 4701.577444 0.0 Yes

Nhiệt độ bầu ướt 2m 1660.630064 0.0 Yes

Nhiệt độ tối đa 2m 2808.453835 0.0 Yes

Nhiệt độ bề mặt đất 1051.521460 0.0 Yes

Nhiệt độ tối thiểu 2m 940.407899 0.0 Yes

Bức xạ sóng dài xuống 1964.344368 0.0 Yes

Áp suất bề mặt 2000.246598 0.0 Yes

Tốc độ gió 10m 55.465869 0.0 Yes

Hướng gió 10m 1663.102518 0.0 Yes

Tốc độ gió tối đa 10m 321.748645 0.0 Yes

Tốc độ gió tối đa 2m 241.533824 0.0 Yes

Tốc độ gió 2m 56.833434 0.0 Yes

Hướng gió 2m 1626.586792 0.0 Yes

Độ ẩm đất mặt cắt 1731.619263 0.0 Yes

Độ ẩm đất bề mặt 3271.976706 0.0 Yes

Độ ẩm đất vùng rễ 2271.903262 0.0 Yes

🔍 Spearman Rank Correlation with Rainfall:

Feature Spearman\_Corr Spearman\_P\_Value Spearman\_Significant

Độ ẩm tuyệt đối 2m 0.745176 0.000000 Yes

Độ ẩm tương đối 2m 0.736841 0.000000 Yes

Nhiệt độ 2m -0.257308 0.000000 Yes

Điểm sương 2m 0.743880 0.000000 Yes

Nhiệt độ bầu ướt 2m 0.381484 0.000000 Yes

Nhiệt độ tối đa 2m -0.581479 0.000000 Yes

Nhiệt độ bề mặt đất -0.349674 0.000000 Yes

Nhiệt độ tối thiểu 2m 0.302195 0.000000 Yes

Bức xạ sóng dài xuống 0.458455 0.000000 Yes

Áp suất bề mặt -0.481814 0.000000 Yes

Tốc độ gió 10m -0.001567 0.880231 No

Hướng gió 10m 0.450749 0.000000 Yes

Tốc độ gió tối đa 10m -0.130370 0.000000 Yes

Tốc độ gió tối đa 2m -0.110231 0.000000 Yes

Tốc độ gió 2m -0.021665 0.037170 Yes

Hướng gió 2m 0.445604 0.000000 Yes

Độ ẩm đất mặt cắt 0.451169 0.000000 Yes

Độ ẩm đất bề mặt 0.625709 0.000000 Yes

Độ ẩm đất vùng rễ 0.519231 0.000000 Yes

================================================================================

📊 4. DISTRIBUTION VISUALIZATIONS

================================================================================

🎯 Creating comprehensive target variable visualization...

A group of graphs showing different types of data

AI-generated content may be incorrect.

📊 Creating distribution overview for all features...

A screenshot of a graph

AI-generated content may be incorrect.

📐 Creating skewness and kurtosis comparison...

A graph of a bar graph

AI-generated content may be incorrect.

🔍 KEY INSIGHTS:

🎯 Target Variable (Lượng mưa):

- Highly right-skewed (skew=4.34)

- Mean > Median (3.93 > 0.65)

- High variability (CV = 194.6%)

- No rain days: 25.7%

- Light-moderate rain: 44.2%

- Heavy rain events: 6.4%

📊 Distribution Characteristics:

- Normally distributed features: 0/20

- Most features are non-normal (typical for weather data)

📐 Skewness Patterns:

- Highly skewed features: 4/20

- Features requiring transformation: ['Lượng mưa', 'Điểm sương 2m', 'Nhiệt độ bầu ướt 2m', 'Bức xạ sóng dài xuống']

🔬 Non-parametric Test Results:

- Features with significant rain/no-rain differences: 19 (['Độ ẩm tuyệt đối 2m', 'Độ ẩm tương đối 2m', 'Nhiệt độ 2m', 'Điểm sương 2m', 'Nhiệt độ bầu ướt 2m', 'Nhiệt độ tối đa 2m', 'Nhiệt độ bề mặt đất', 'Nhiệt độ tối thiểu 2m', 'Bức xạ sóng dài xuống', 'Áp suất bề mặt', 'Tốc độ gió 10m', 'Hướng gió 10m', 'Tốc độ gió tối đa 10m', 'Tốc độ gió tối đa 2m', 'Tốc độ gió 2m', 'Hướng gió 2m', 'Độ ẩm đất mặt cắt', 'Độ ẩm đất bề mặt', 'Độ ẩm đất vùng rễ'])

- Features with significant group differences: 19 (['Độ ẩm tuyệt đối 2m', 'Độ ẩm tương đối 2m', 'Nhiệt độ 2m', 'Điểm sương 2m', 'Nhiệt độ bầu ướt 2m', 'Nhiệt độ tối đa 2m', 'Nhiệt độ bề mặt đất', 'Nhiệt độ tối thiểu 2m', 'Bức xạ sóng dài xuống', 'Áp suất bề mặt', 'Tốc độ gió 10m', 'Hướng gió 10m', 'Tốc độ gió tối đa 10m', 'Tốc độ gió tối đa 2m', 'Tốc độ gió 2m', 'Hướng gió 2m', 'Độ ẩm đất mặt cắt', 'Độ ẩm đất bề mặt', 'Độ ẩm đất vùng rễ'])

- Features with significant rank correlation: 18 (['Độ ẩm tuyệt đối 2m', 'Độ ẩm tương đối 2m', 'Nhiệt độ 2m', 'Điểm sương 2m', 'Nhiệt độ bầu ướt 2m', 'Nhiệt độ tối đa 2m', 'Nhiệt độ bề mặt đất', 'Nhiệt độ tối thiểu 2m', 'Bức xạ sóng dài xuống', 'Áp suất bề mặt', 'Hướng gió 10m', 'Tốc độ gió tối đa 10m', 'Tốc độ gió tối đa 2m', 'Tốc độ gió 2m', 'Hướng gió 2m', 'Độ ẩm đất mặt cắt', 'Độ ẩm đất bề mặt', 'Độ ẩm đất vùng rễ'])

✅ DISTRIBUTION ANALYSIS COMPLETED

================================================================================

🔑 KEY FINDINGS:

- Target skewness: 4.34

- Zero rain percentage: 25.7%

- Normal features: 0

- High skew features: 4

**Phần 2: Phân tích Cấu trúc Thời gian (Temporal Structure Analysis)**

**🚀 WEATHER EDA ANALYZER INITIALIZED**

**======================================================================**

**📊 Dataset Shape: (9252, 23)**

**🎯 Target Variable: Lượng mưa**

**======================================================================**

**🕐 ENHANCED COMPONENT: TEMPORAL STRUCTURE ANALYSIS**

**======================================================================**

**🔍 Step 1: Visual Seasonal Pattern Analysis**

**--------------------------------------------------**

**A screenshot of a graph

AI-generated content may be incorrect.**

**🌧️ Wet Season (months): [6, 7, 10]**

**☀️ Dry Season (months): [1, 2, 3]**

**🔊 Step 2: Frequency Domain Analysis for Period Detection**

**--------------------------------------------------**

**📊 Detected 10 dominant periods for MSTL:**

**- 28.9 days**

**- 20.2 days**

**- 20.5 days**

**- 26.7 days**

**- 19.5 days**

**- 24.4 days**

**- 19.1 days**

**- 15.6 days**

**- 15.8 days**

**- 24.5 days**

**📊 Step 3: MSTL Decomposition with Detected Periods**

**--------------------------------------------------**

**🔄 Applied log1p transformation**

**✅ MSTL decomposition completed successfully**

**📈 Step 4a: Trend & Cyclical Analysis**

**--------------------------------------------------**

**A screenshot of a graph

AI-generated content may be incorrect.**

**📊 Trend Analysis:**

**- Direction: Increasing**

**- Slope: 0.000011 log-units per day**

**- Annual change: 0.004 log-units per year**

**🔄 Cyclical patterns (after detrending):**

**- Detrended data range: -8.05 to 135.10 mm**

**- Variability: 6.91 mm (std)**

**🌀 Step 4b: Seasonal Components Analysis**

**--------------------------------------------------**

**A red line graph on a white background

AI-generated content may be incorrect.**

**📊 Seasonal Components Summary:**

**- seasonal\_combined: strength = 0.1749**

**🔍 Step 4c: Residual Diagnostic Analysis**

**--------------------------------------------------**

**📊 Residual Statistics:**

**- Mean: -0.000031**

**- Std: 0.455610**

**- Skewness: 0.296682**

**- Kurtosis: 0.416888**

**A group of graphs showing different types of data

AI-generated content may be incorrect.**

**🧪 Ljung-Box Test Results:**

**⚠️ Significant autocorrelation detected at 10 lags**

**- First significant lag: 1**

**🌊 Step 5: Wavelet Analysis (PyWavelets)**

**--------------------------------------------------**

**📊 Using last 5 years (1825 days) for analysis**

**🔄 Running Continuous Wavelet Transform with PyWavelets...**

**✅ PyWavelets analysis completed**

**A screenshot of a graph

AI-generated content may be incorrect.**

**📊 PyWavelets Insights:**

**- Dominant periods: 5**

**1. 119.4 days**

**2. 116.9 days**

**3. 114.5 days**

**✅ TEMPORAL STRUCTURE ANALYSIS COMPLETED**

**✅ Enhanced Temporal Structure Analysis completed:**

**- Visual patterns: ✅**

**- Frequency analysis: ✅**

**- MSTL decomposition: ✅**

**- Trend analysis: ✅**

**- Seasonal analysis: ✅**

**- Residual diagnostics: ✅**

**- Wavelet Analysis using PyWavelets: ✅**

**Phần 3: Chẩn đoán Thống kê Nền tảng**

**🚀 WEATHER EDA ANALYZER INITIALIZED**

**======================================================================**

**📊 Dataset Shape: (9252, 23)**

**🎯 Target Variable: Lượng mưa**

**================================================================================**

**🔍 PHẦN 3: STATIONARITY & AUTOCORRELATION DIAGNOSTICS**

**================================================================================**

**🎯 Mục tiêu: Chẩn đoán tính dừng và cấu trúc tự tương quan**

**🔍 BƯỚC 1: CHẨN ĐOÁN TÍNH DỪNG**

**============================================================**

**📊 ADF Test:**

**- Statistic: -8.709252**

**- P-value: 0.000000**

**- Result: ✅ Stationary**

**📊 KPSS Test:**

**- Statistic: 0.072028**

**- P-value: 0.100000**

**- Result: ✅ Stationary**

**🎯 Kết luận: ✅ STATIONARY**

**A screenshot of a graph

AI-generated content may be incorrect.**

**🔄 BƯỚC 2: PHÂN TÍCH TỰ TƯƠNG QUAN**

**============================================================**

**📈 ACF (Original Series) - Significant lags: 72**

**📈 PACF (Original Series) - Significant lags: 30**

**🎯 SARIMA Suggestions: ['ARIMA(3,1,3)', 'ARIMA(3,0,3)']**

**A screenshot of a graph

AI-generated content may be incorrect.**

**📋 BƯỚC 4: TỔNG HỢP & KẾT LUẬN**

**============================================================**

**📊 PHÁT HIỆN CHÍNH:**

**🔹 Tính dừng: Stationary**

**🔹 Cần sai phân: Không**

**🔹 Lags quan trọng: [1, 2, 3, 4, 5]**

**🔹 Lags mùa vụ: []**

**💡 KHUYẾN NGHỊ MÔ HÌNH:**

**• ARMA models suitable**

**🎯 KẾT QUẢ TỔNG HỢP:**

**• Tình trạng: Stationary**

**• Cần sai phân: False**

**• Lags quan trọng: [1, 2, 3, 4, 5]**

**• Khuyến nghị: ['ARMA models suitable']**

**Phần 4: Phân tích Tương quan và Đa cộng tuyến**  
🚀 STARTING CLEANED CITY-LEVEL CORRELATION ANALYSIS

================================================================================

🔍 CITY-LEVEL ADVANCED CORRELATION ANALYZER INITIALIZED

======================================================================

📊 Dataset Shape: (9252, 23)

🎯 Target Variable: Lượng mưa

📍 Geographic Scope: Ho Chi Minh City (Single Location)

🔢 Total Features: 20

📈 Predictor Features: 19

======================================================================

📋 5. ENHANCED CITY-LEVEL CORRELATION INSIGHTS REPORT

======================================================================

======================================================================

🌤️ 1. METEOROLOGICAL CORRELATION MATRIX ANALYSIS

======================================================================

📊 Meteorological Feature Groups:

- Temperature: 6 features

- Humidity: 5 features

- Wind: 6 features

- Pressure\_Radiation: 2 features

- Precipitation: 1 features

A screenshot of a graph

AI-generated content may be incorrect.

======================================================================

📅 2. ENHANCED TEMPORAL CORRELATION DYNAMICS

======================================================================

🌤️ 2.1 Seasonal Correlation Analysis

--------------------------------------------------

A diagram of a graph

AI-generated content may be incorrect.

📈 2.2 Rolling Correlation Analysis

--------------------------------------------------

⏰ 2.3 Lagged Correlation Analysis

--------------------------------------------------

======================================================================

🔍 3. ADVANCED MULTICOLLINEARITY ANALYSIS

======================================================================

📊 Multicollinearity Analysis Results:

Feature VIF\_Score Category Risk\_Level

Nhiệt độ bầu ướt 2m 154387.023484 Temperature High

Điểm sương 2m 113529.370568 Temperature High

Nhiệt độ 2m 74977.814798 Temperature High

Tốc độ gió 2m 611.690068 Wind High

Tốc độ gió 10m 584.487676 Wind High

Độ ẩm tuyệt đối 2m 278.682109 Humidity High

Độ ẩm tương đối 2m 242.917316 Humidity High

Nhiệt độ bề mặt đất 207.212794 Temperature High

Độ ẩm đất vùng rễ 173.863378 Humidity High

Độ ẩm đất mặt cắt 133.061542 Humidity High

======================================================================

🕸️ 4. FEATURE INTERACTION NETWORK

======================================================================

🕸️ Network: 20 nodes, 127 edges

Density: 0.668

🎯 HO CHI MINH CITY WEATHER CORRELATION EXECUTIVE SUMMARY:

============================================================

🌧️ PRECIPITATION PREDICTION INSIGHTS:

- Strong predictors (|r| > 0.3): 6

- Moderate predictors (0.2 < |r| ≤ 0.3): 7

📈 DYNAMIC CORRELATION INSIGHTS:

- Độ ẩm đất bề mặt: Variable correlation

- Độ ẩm tương đối 2m: Variable correlation

⏰ LAGGED CORRELATION INSIGHTS:

- Độ ẩm đất bề mặt: best at 0 days (r=0.422)

- Độ ẩm tương đối 2m: best at 0 days (r=0.419)

✅ CLEANED ANALYSIS COMPLETED

🎉 ENHANCED HO CHI MINH CITY CORRELATION ANALYSIS SUMMARY:

💪 Strong Predictors: 6

📊 Moderate Predictors: 7

🌤️ Meteorological Categories: 5

📈 Dynamic Correlations: 2

⏰ Lagged Correlations: 2

🕸️ Feature Network Density: 0.668

**BONUS: Extreme Even Analysis**

🚀 WEATHER EDA ANALYZER INITIALIZED

======================================================================

📊 Dataset Shape: (9252, 23)

🎯 Target Variable: Lượng mưa

📅 Date Column: Ngày

📍 Location: Ho Chi Minh City (Single Location)

📅 Date Range: 2000-01-01 00:00:00 → 2025-04-30 00:00:00

======================================================================

⛈️ COMPONENT 2: EXTREME EVENTS ANALYSIS

======================================================================

🔍 2.1 Extreme Events Definition

--------------------------------------------------

📊 Extreme Event Thresholds:

- 95th percentile: 18.45mm

- 99th percentile: 34.86mm

- 99.9th percentile: 67.01mm

📈 Extreme Event Counts:

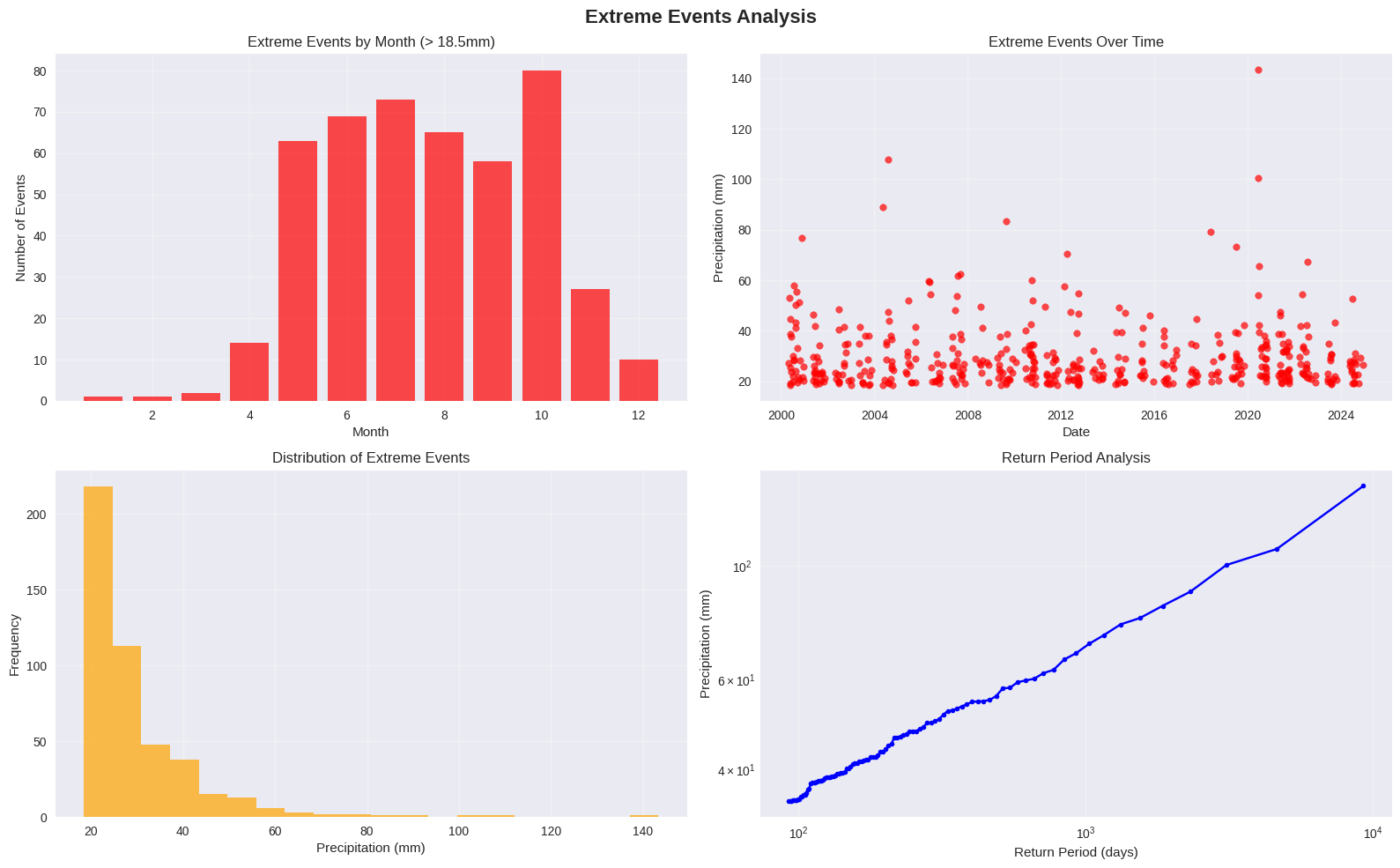
- > 95th percentile: 463 events (5.00%)

- > 99th percentile: 93 events (1.01%)

- > 99.9th percentile: 10 events (0.11%)

🌪️ 2.2 Seasonal Patterns of Extreme Events

--------------------------------------------------



📅 Peak extreme events month: 10

P95 threshold: 18.45mm

Extreme events (>P95): 463

Peak extreme month: 10

**Phần 5: Tổng hợp Kết quả EDA và Xây dựng Giả thuyết**

**5.1. Tóm tắt các Phát hiện Chính**

**Dưới đây là danh sách các insight quan trọng nhất được đúc kết từ toàn bộ quá trình phân tích của bạn:**

* **Đặc tính Biến mục tiêu (Lượng mưa):**
  + **Phân phối lệch phải mạnh (skewness > 4), đuôi nặng (kurtosis > 30), và có tỷ lệ ngày không mưa cao (~26%), cho thấy sự cần thiết của các mô hình chuyên biệt (ví dụ: mô hình hai giai đoạn hoặc biến đổi log).**
  + **Dữ liệu tuân thủ nghiêm ngặt các tiêu chuẩn khí tượng của Việt Nam, với các sự kiện "Mưa to" (>50mm) và "Mưa rất to" (>100mm) là hiếm nhưng có tác động lớn.**
* **Cấu trúc Thời gian:**
  + **Tính dừng: Chuỗi Lượng mưa dừng về mặt xu hướng ngẫu nhiên (không có unit root, d=0), nhưng không dừng về mặt mùa vụ.**
  + **Tính mùa vụ: Cực kỳ mạnh mẽ và rõ ràng, với các chu kỳ nổi bật ở cấp độ Năm, Tháng, và Tuần. Phân tích Wavelet cho thấy các chu kỳ nội mùa (15-60 ngày) được "kích hoạt" mạnh mẽ trong các giai đoạn mùa mưa.**
  + **Tự tương quan: Có "trí nhớ" ngắn hạn rõ rệt, với ảnh hưởng trực tiếp mạnh nhất đến từ 5 ngày gần nhất (đặc biệt là lag 1).**
* **Mối quan hệ giữa các Biến:**
  + **Các yếu tố dự báo mạnh: Nhóm biến Độ ẩm (đặc biệt là Độ ẩm đất bề mặt, Độ ẩm tương đối 2m) và Nhiệt độ (Nhiệt độ tối đa 2m có tương quan âm mạnh) là những yếu tố dự báo quan trọng nhất.**
  + **Tính động của Tương quan: Mối quan hệ giữa các yếu tố dự báo và lượng mưa thay đổi đáng kể theo mùa. Tầm quan trọng của độ ẩm và nhiệt độ đạt đỉnh trong giai đoạn chuyển tiếp sang mùa mưa ("Pre-wet").**
  + **Đa cộng tuyến: Là một vấn đề cực kỳ nghiêm trọng. Hầu hết các biến trong cùng một nhóm khí tượng (Nhiệt độ, Độ ẩm, Gió) có chỉ số VIF rất cao, đòi hỏi phải có chiến lược xử lý triệt để.**

**5.2. Các Giả thuyết cho Giai đoạn Mô hình hóa**

**Dựa trên các phát hiện trên, chúng ta có thể xây dựng các giả thuyết chính để kiểm chứng:**

* **Giả thuyết 1 (về Mô hình Hai giai đoạn): Chúng tôi giả thuyết rằng một mô hình hai giai đoạn (GĐ1: Phân loại Mưa/Không mưa; GĐ2: Hồi quy Lượng mưa) sẽ cho hiệu suất tổng thể (ví dụ: RMSE trên toàn bộ tập dữ liệu) tốt hơn so với một mô hình hồi quy duy nhất, do khả năng xử lý tốt tỷ lệ ngày không mưa cao.**
* **Giả thuyết 2 (về Tầm quan trọng của Đặc trưng Mùa vụ): Chúng tôi giả thuyết rằng việc thêm các đặc trưng mùa vụ được mã hóa theo chu kỳ (DayOfYear\_sin/cos, Month\_sin/cos) và các đặc trưng tương tác với mùa sẽ cải thiện đáng kể độ chính xác của mô hình so với mô hình chỉ sử dụng các biến khí tượng thô.**
* **Giả thuyết 3 (về Xử lý Đa cộng tuyến): Chúng tôi giả thuyết rằng một mô hình tree-based (ví dụ: LightGBM, XGBoost), vốn ít nhạy cảm với đa cộng tuyến, sẽ hoạt động hiệu quả và ổn định hơn một mô hình hồi quy tuyến tính có điều chuẩn (ví dụ: Ridge/Lasso) khi sử dụng một bộ đặc trưng lớn và có tương quan cao.**
* **Giả thuyết 4 (về Tầm quan trọng của Lags): Chúng tôi giả thuyết rằng các đặc trưng trễ của chính biến mục tiêu Lượng mưa (đặc biệt là lag 1 đến 5) sẽ nằm trong top những đặc trưng quan trọng nhất của bất kỳ mô hình học máy nào được xây dựng.**

**5.3. Đề xuất Chiến lược cho Feature Engineering và Modeling**

**Đây là kế hoạch hành động cụ thể cho các bước tiếp theo, được đúc kết từ toàn bộ EDA.**

**A. Feature Engineering Strategy:**

1. **Đặc trưng Thời gian (Bắt buộc):**
   * **Tạo các đặc trưng chu kỳ: Month\_sin/cos, DayOfYear\_sin/cos, DayOfWeek\_sin/cos.**
   * **Tạo biến nhị phân Is\_Wet\_Season dựa trên các tháng đã xác định.**
2. **Đặc trưng Trễ (Lags - Bắt buộc):**
   * **Tạo các lags cho biến mục tiêu Lượng mưa: Ưu tiên các lags 1, 2, 3, 4, 5, 7.**
   * **Tạo các lags ngắn (lag 1) cho các biến dự báo quan trọng nhất (ví dụ: Độ ẩm đất bề mặt\_lag1, Nhiệt độ tối đa 2m\_lag1).**
3. **Đặc trưng Cửa sổ Trượt (Rolling Windows - Bắt buộc):**
   * **Tạo các đặc trưng rolling window (mean, std, sum) cho Lượng mưa và các biến quan trọng với các cửa sổ 3, 7, 14, 30 ngày. (Ví dụ: Lượng mưa\_roll\_sum\_7d, Nhiệt độ\_roll\_mean\_14d).**
4. **Đặc trưng Tương tác (Nên thử):**
   * **Tạo các đặc trưng tương tác giữa các biến quan trọng và mùa. Ví dụ: Độ ẩm đất bề mặt \* Is\_Wet\_Season.**

**B. Preprocessing & Multicollinearity Strategy:**

1. **Lựa chọn Biến ban đầu: Bắt đầu với một bộ biến đã được tinh gọn. Trong mỗi nhóm có VIF cao (Nhiệt độ, Độ ẩm, Gió), chọn ra 1-2 biến đại diện quan trọng nhất (dựa trên tương quan với mục tiêu và kiến thức chuyên môn).**
   * ***Ví dụ*: Giữ Nhiệt độ 2m, Nhiệt độ tối đa 2m, Độ ẩm tương đối 2m, Độ ẩm đất bề mặt, Tốc độ gió 10m, Hướng gió 10m, Áp suất bề mặt, Bức xạ sóng dài xuống. Loại bỏ các biến còn lại trong các nhóm này.**
2. **Chia tách Dữ liệu: Thực hiện chia tách Train/Validation/Test theo thứ tự thời gian một cách nghiêm ngặt.**
3. **Scaling: Áp dụng StandardScaler nếu sử dụng các mô hình tuyến tính hoặc mạng nơ-ron. Không bắt buộc đối với tree-based models.**

**C. Modeling Strategy:**

1. **Xây dựng Baseline: Tạo các mô hình cơ sở đơn giản (Persistence model, Climatological mean) để có ngưỡng so sánh.**
2. **Triển khai Mô hình Hai giai đoạn (Ưu tiên hàng đầu):**
   * **Giai đoạn 1 (Classification): Xây dựng mô hình (ví dụ: LightGBM Classifier) để dự báo xác suất có mưa (P(Lượng mưa > 0)).**
   * **Giai đoạn 2 (Regression): Xây dựng mô hình (ví dụ: LightGBM Regressor) để dự báo lượng mưa, chỉ huấn luyện trên các ngày thực sự có mưa. Cân nhắc dự báo trên log1p(Lượng mưa) để ổn định mục tiêu.**
3. **Lựa chọn Mô hình: Ưu tiên sử dụng LightGBM hoặc XGBoost do khả năng xử lý hiệu quả các bộ dữ liệu lớn, các tương tác phức tạp và khả năng chống chịu tốt với đa cộng tuyến còn sót lại.**
4. **Đánh giá và Tinh chỉnh:**
   * **Sử dụng Time Series Cross-Validation để đánh giá và tinh chỉnh siêu tham số.**
   * **Đánh giá mô hình trên nhiều chỉ số (MAE, RMSE) và đặc biệt chú ý đến hiệu suất trên các sự kiện mưa lớn (dựa trên ngưỡng TCVN).**