VNUHCM - Trường Đại học Khoa học Tự nhiên
Khoa Công nghệ Thông tin

Báo cáo đồ án
Ứng dụng Công cụ AI
trong quy trình Phân tích Dữ liệu thực tế

Môn học: Phân tích dữ liệu thông minh

Nhóm QuQuBeTa

Sinh viên thực hiện:

Nguyễn Hữu Bền - 22120029
Nguyễn Tiến Quốc - 22120300
Nguyễn Trung Quốc - 22120301
Võ Thành Tâm - 22120324

Giảng viên hướng dẫn:

TS. Nguyễn Tiến Huy

Thành phố Hồ Chí Minh - 2025

Nội dung

1 Giới thiệu

1.1 Giới thiệu nhóm

MSSV	Họ và tên
22120029	Nguyễn Hữu Bền
22120300	Nguyễn Tiến Quốc
22120301	Nguyễn Trung Quốc
22120324	Võ Thành Tâm

Bảng 1.1: Danh sách thành viên

1.2 Mục tiêu đồ án

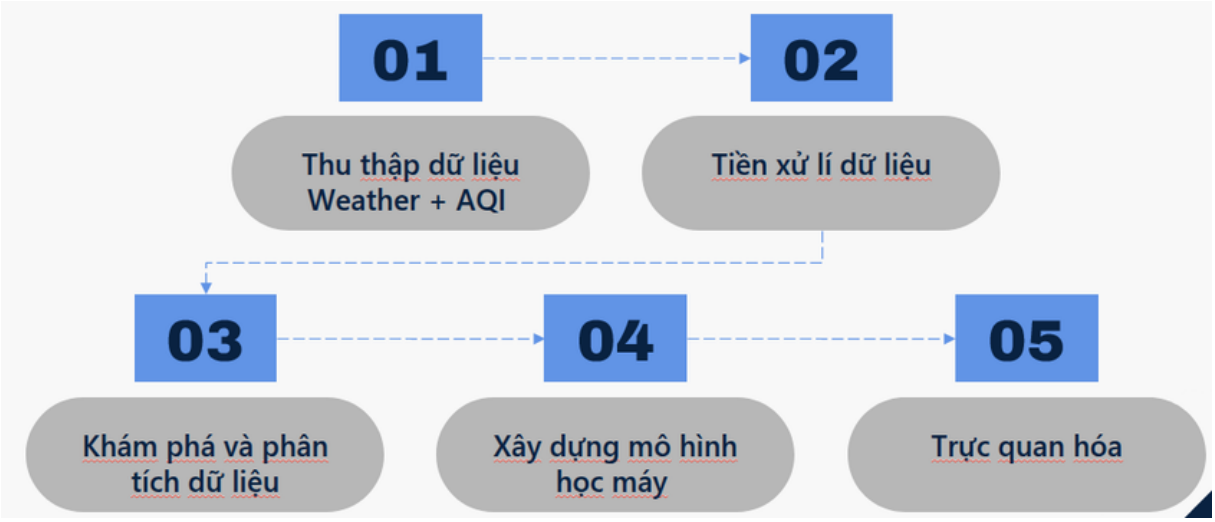
Mục tiêu chính của đồ án là:

- **Ứng dụng công cụ AI** để phân tích dữ liệu và dự đoán chất lượng không khí, thời tiết tại TP.HCM, giúp nhận diện mức độ ô nhiễm và dự báo chất lượng không khí. Ngoài ra, đồ án còn nhằm mục tiêu khám phá tiềm năng của các công cụ AI hiện đại như ChatGPT Plus và Grok trong việc hỗ trợ các nhà phân tích dữ liệu không chuyên, giúp tiết kiệm thời gian và tăng tính trực quan trong toàn bộ quy trình.

- **Xây dựng một pipeline phân tích dữ liệu hoàn chỉnh** từ thu thập, tiền xử lý, phân tích, xây dựng, đánh giá mô hình học máy và trực quan.

- **Phát triển kỹ năng thực tế**: Làm việc nhóm, rèn luyện kỹ năng xử lý dữ liệu thực tế, phân tích, trực quan hóa dữ liệu và áp dụng các mô hình học máy.

2 Quy trình Phân tích dữ liệu



Hình 2.1: Pipeline Quy trình phân tích dữ liệu

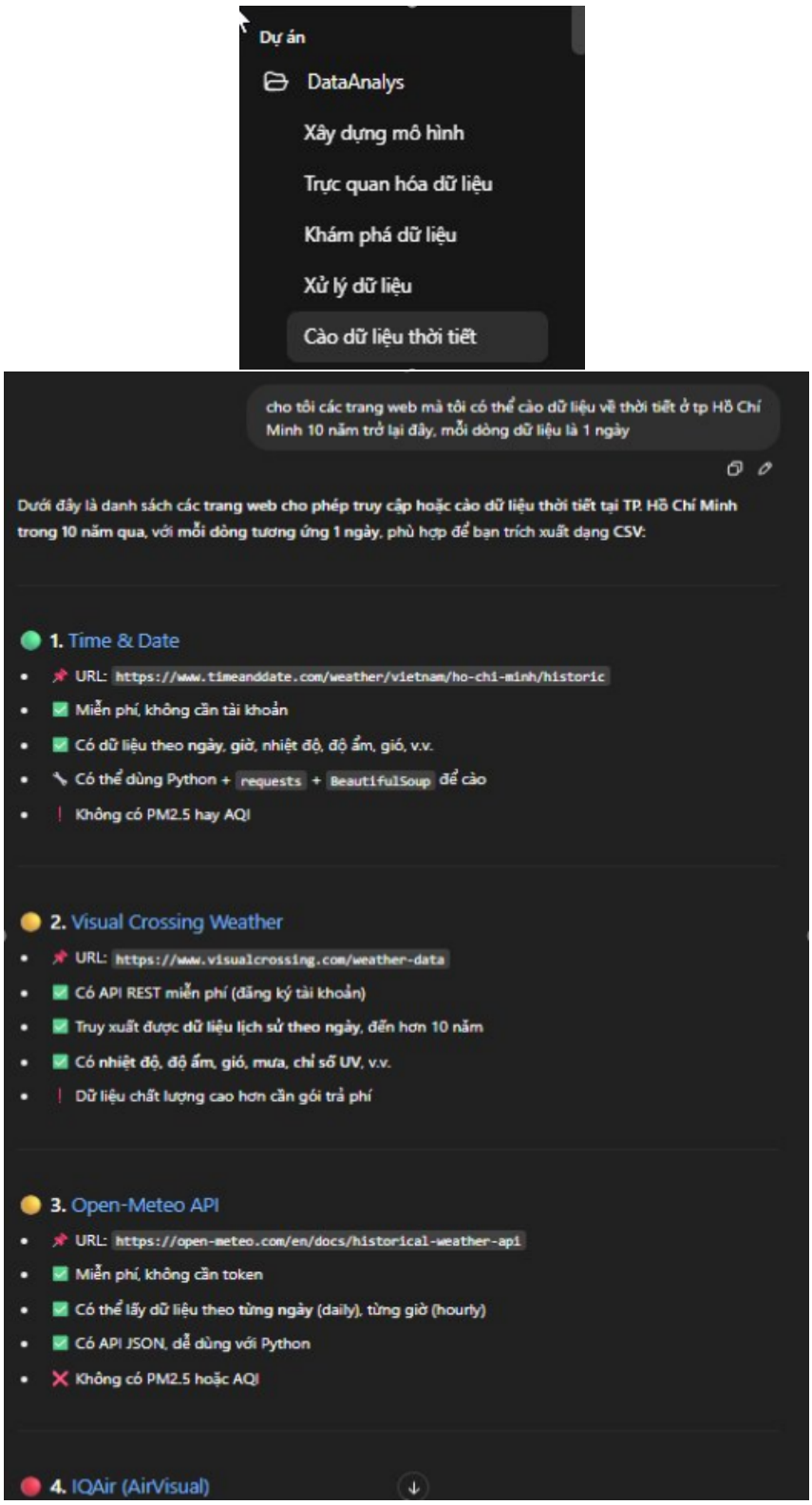
2.1 Ứng dụng AI trong quá trình phân tích

Việc sử dụng các mô hình ngôn ngữ lớn (LLMs) giúp nhóm rút ngắn đáng kể thời gian phân tích thông qua các chức năng như:

- Tăng tốc phân tích dữ liệu bằng cách sinh mã tự động, giải thích biểu đồ, gợi ý mô hình.
- Hỗ trợ sinh insight từ dữ liệu bằng cách đọc, tóm tắt hoặc đặt câu hỏi từ dữ liệu
- Hỗ trợ viết báo cáo, đặc biệt, nhóm đánh giá cao khả năng "tương tác đa vòng lặp" của các LLM, giúp đào sâu các insight tiềm ẩn thay vì chỉ dừng ở thống kê mô tả đơn thuần.

. Nhóm quyết định sử dụng các mô hình ngôn ngữ lớn (bản miễn phí và có trả phí): Grok và ChatGPTPlus

Đối với ChatGPT Plus(tương tự với Grok), nhóm thực hiện đưa ra các promt cho 5 khung chat trong 1 project (ứng với 5 công đoạn chính trong quá trình phân tích dữ liệu). Ở từng khung chat, lần lượt thực hiện prompt theo các công thức: T-A-G (Task, Action, Goal: mục tiêu)



Hình 2.2: Ví dụ về cách prompt và kết quả

2.2 Thu thập dữ liệu

Dữ liệu được thu thập từ các **trạm cảm biến môi trường** tại TP.HCM, bao gồm các yếu tố môi trường quan trọng như **Nhiệt độ, Độ ẩm, Tốc độ gió, Áp suất khí quyển** và các chỉ số ô nhiễm như **AQI, PM2.5, PM10, CO, NO2, SO2, NH3**

Các nguồn thu thập dữ liệu bao gồm:

- **Dữ liệu AQI** được thu thập qua **API** của OpenWeatherMap
- **Dữ liệu về các yếu tố thời tiết** được thu thập từ TimeandDate

.

Dữ liệu được thu thập **hàng giờ**, từ các trạm cảm biến và các nguồn API trực tuyến.

2.3 Tiền xử lí dữ liệu

Sau khi thu thập hai bộ được dữ liệu gồm thời tiết và AQI, tiến hành làm sạch và xử lí dữ liệu để có thể gộp 2 bộ dữ liệu theo thời gian.

1. **Làm sạch dữ liệu**: Tiến hành loại bỏ các đơn vị, xử lí lại thời gian.
2. Cột **Visiblity** nhóm tiến hành loại bỏ vì có quá nhiều giá trị lỗi.
3. **Chuyển đổi định dạng thời gian**: Cột **datetime** đã được chuẩn hóa về định dạng *datetime* để thuận tiện cho việc phân tích.
4. **Gộp hai bộ dữ liệu theo thời gian**: Vì dữ liệu AQI chỉ được lưu trữ từ 25/11/2020 nên nhóm sẽ tiến hành gộp và lấy thời gian từ 01/01/2021.

Bộ dữ liệu thu được 18 cột và 36624 mẫu dữ liệu

2.4 Tổng quan dữ liệu

Tên Cột	**Kiểu Dữ Liệu**	**Giải thích ý nghĩa**
`datetime`	datetime	Thời gian đầy đủ (ngày + giờ)
`date`	date	Ngày quan sát
`time`	time	Giờ quan sát
`temp`	numerical	Nhiệt độ (°C)
`weather`	categorical	Tình trạng thời tiết
`wind`	numerical	Tốc độ gió km/h
`wind_direction`	numerical	Hướng gió từ 0°–360°
`humidity`	numerical	Độ ẩm tương đối (%)
`pressure`	numerical	Áp suất khí quyển (mBar)
`aqi`	numerical	Chỉ số chất lượng không khí (Air Quality Index)
`co`	numerical	Nồng độ khí CO ($\mu g/m^3$)
`no`	numerical	Nồng độ NO ($\mu g/m^3$)
`no2`	numerical	Nồng độ NO_2 ($\mu g/m^3$)
`o3`	numerical	Nồng độ O_3 ($\mu g/m^3$)
`so2`	numerical	Nồng độ SO_2 ($\mu g/m^3$)
`pm2_5`	numerical	Nồng độ bụi mịn PM2.5 ($\mu g/m^3$)
`pm10`	numerical	Nồng độ bụi PM10 ($\mu g/m^3$)
`nh3`	numerical	Nồng độ NH_3 ($\mu g/m^3$)

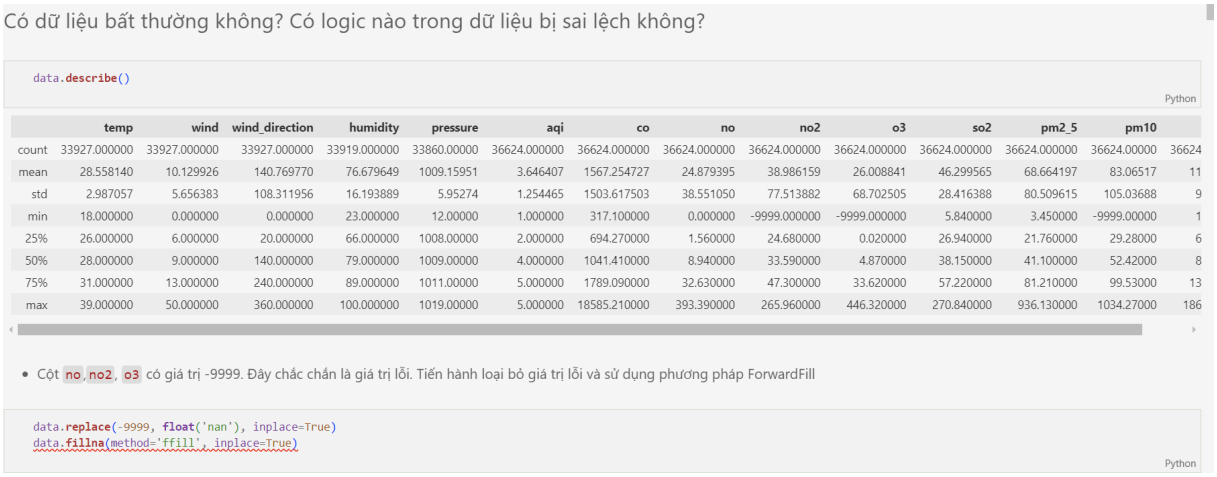
2.5 Khám Phá và Phân Tích Dữ Liệu

1. **Khám phá dữ liệu**:

   – Dữ liệu có giá trị thiếu trong một số cột. Nhóm đã sử dụng phương pháp ForwardFill để đảm bảo dữ liệu đầy đủ cho phân tích.

   – Đối với dữ liệu không khí, có một một vài giá trị lỗi -9999, nhóm tiến hành cũng loại bỏ và điền vào bằng ForwardFill.

   – Kiểm tra sự phân bố chuẩn của các yếu tố môi trường.



Có dữ liệu bất thường không? Có logic nào trong dữ liệu bị sai lệch không?

```
data.describe()
```

Python

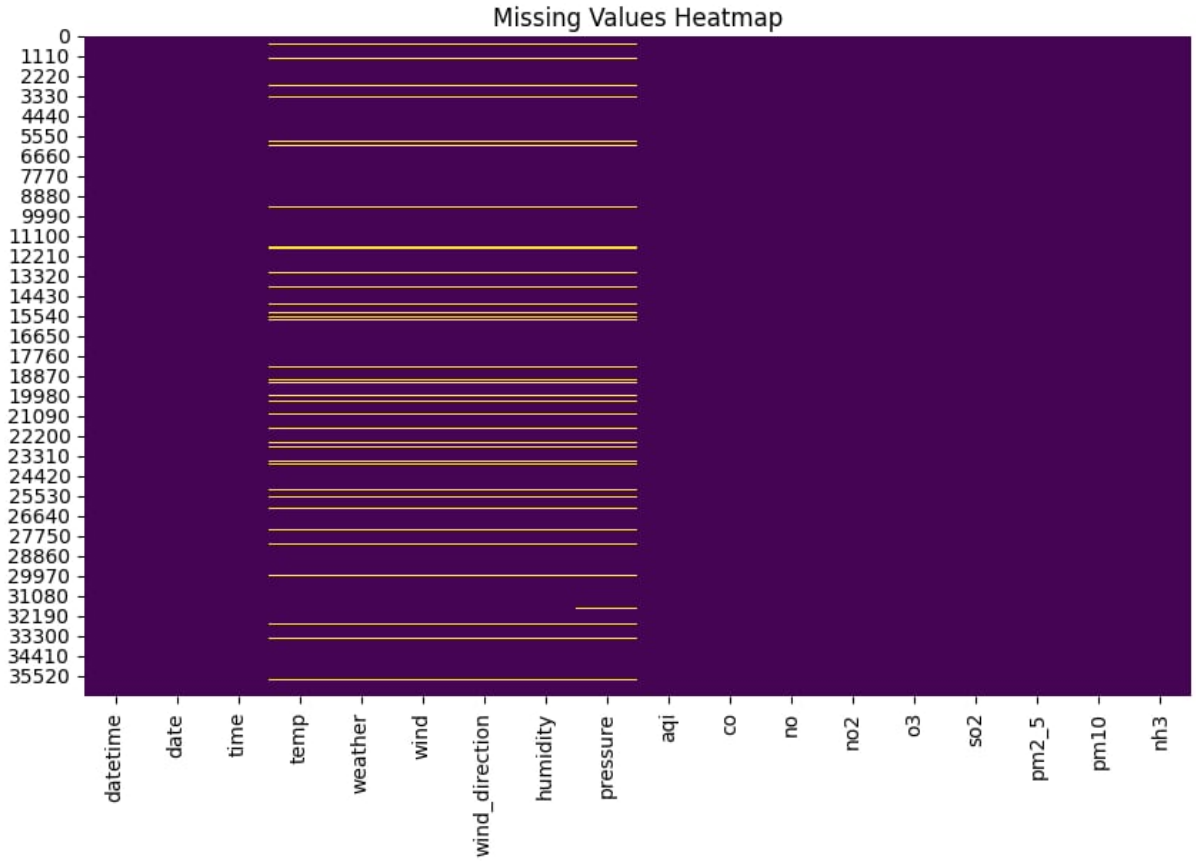
	temp	wind	wind_direction	humidity	pressure	aqi	co	no	no2	o3	so2	pm2_5	pm10	
count	33927.000000	33927.000000	33927.000000	33919.000000	33860.00000	36624.000000	36624.000000	36624.000000	36624.000000	36624.000000	36624.000000	36624.000000	36624.00000	36624
mean	28.558140	10.129926	140.769770	76.679649	1009.15951	3.646407	1567.254727	24.879395	38.986159	26.008841	46.299565	68.664197	83.06517	11
std	2.987057	5.656383	108.311956	16.193889	5.95274	1.254465	1503.617503	38.551050	77.513882	68.702505	28.416388	80.509615	105.03688	9
min	18.000000	0.000000	0.000000	23.000000	12.00000	1.000000	317.100000	0.000000	-9999.000000	-9999.000000	5.840000	3.450000	-9999.00000	1
25%	26.000000	6.000000	20.000000	66.000000	1008.00000	2.000000	694.270000	1.560000	24.680000	0.020000	26.940000	21.760000	29.28000	6
50%	28.000000	9.000000	140.000000	79.000000	1009.00000	4.000000	1041.410000	8.940000	33.590000	4.870000	38.150000	41.100000	52.42000	8
75%	31.000000	13.000000	240.000000	89.000000	1011.00000	5.000000	1789.090000	32.630000	47.300000	33.620000	57.220000	81.210000	99.53000	13
max	39.000000	50.000000	360.000000	100.000000	1019.00000	5.000000	18585.210000	393.390000	265.960000	446.320000	270.840000	936.130000	1034.27000	186

- Cột `no`, `no2`, `o3` có giá trị -9999. Đây chắc chắn là giá trị lỗi. Tiến hành loại bỏ giá trị lỗi và sử dụng phương pháp ForwardFill

```
data.replace(-9999, float('nan'), inplace=True)
data.fillna(method='ffill', inplace=True)
```

Python

Hình 2.3: Bảng mô tả các thông số thống kê của dữ liệu.



Hình 2.4: Biểu đồ Missing Values Heatmap.

2. **Phân Tích Dữ Liệu**:

   – Tính toán các chỉ số thống kê như mean (trung bình), std (độ lệch chuẩn), min (giá trị nhỏ nhất), và max (giá trị lớn nhất) cho các cột dữ liệu để hiểu rõ hơn về đặc điểm của dữ liệu.

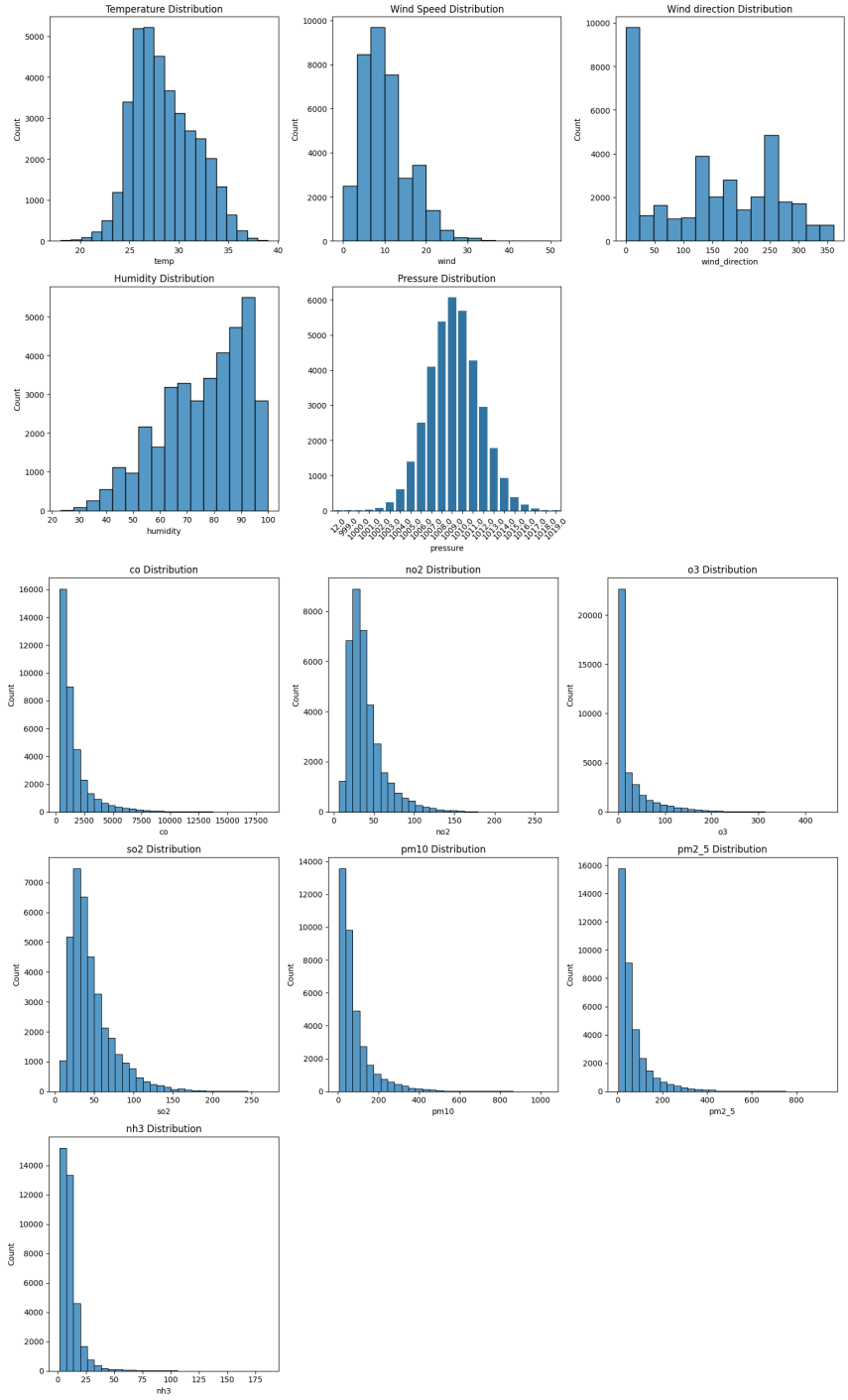
   – **Biểu đồ phân phối (Histogram)**: Được sử dụng để kiểm tra sự phân bố của các cột dữ liệu.

   – **Ma trận tương quan**: thể hiện mối quan hệ giữa các yếu tố môi trường (nhiệt độ, độ ẩm, gió) và chỉ số AQI.

   – Phân tích chuỗi thời gian cho các yếu tố như **Nhiệt độ, Độ ẩm, Áp suất, Tốc độ gió** giúp đánh giá xu hướng và mùa vụ.

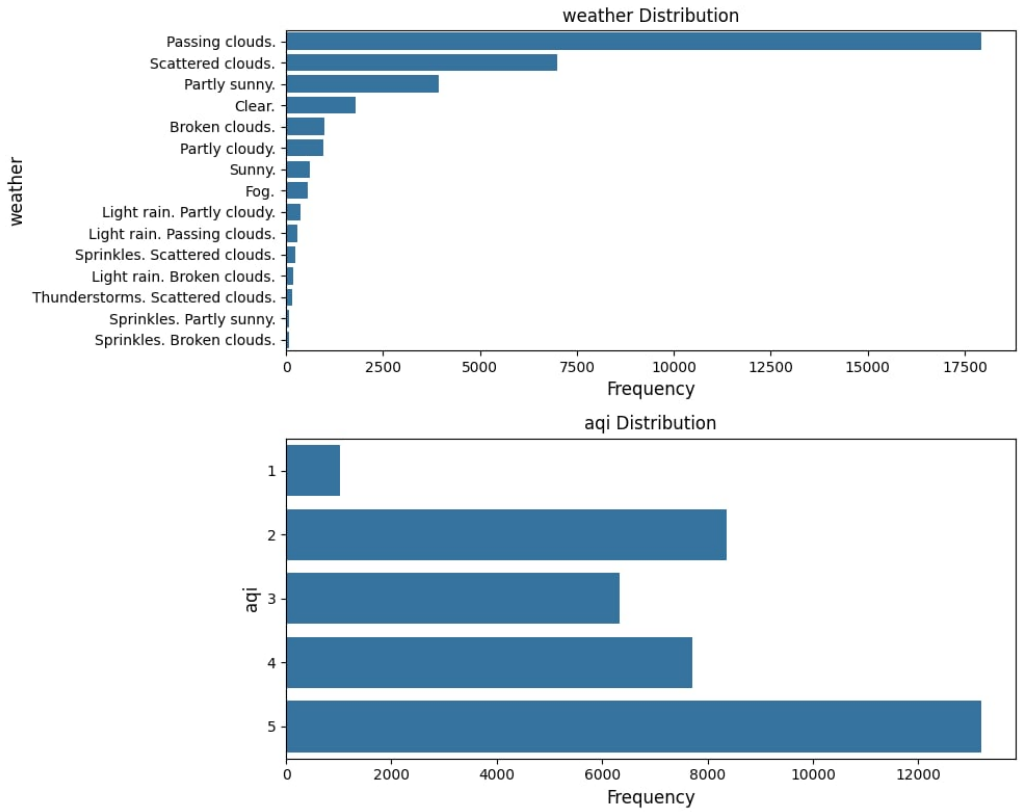
Các phân tích cho thấy rằng hầu hết các yếu tố đều có phân phối chuẩn, giúp cải thiện độ chính xác của mô hình học máy.

Sự phân phối của các đặc trưng trong dữ liệu và sự tương quan giữa chúng



Hình 2.5: Biểu đồ phân phối các chỉ số (các biến numeric).

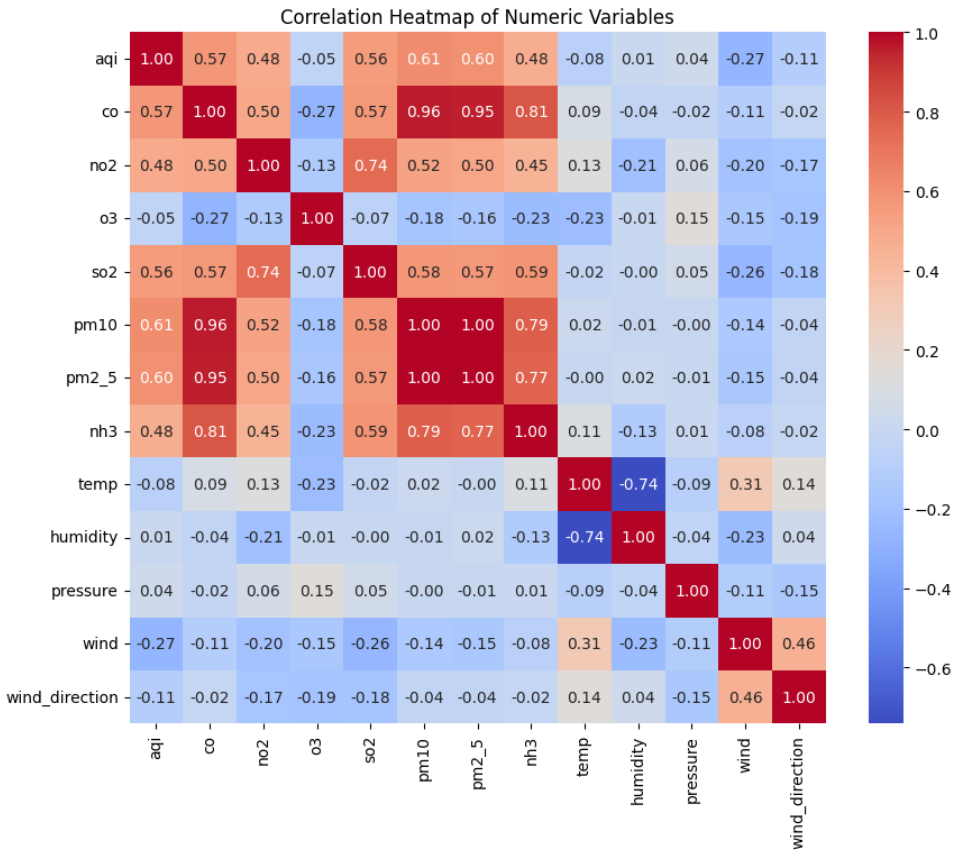
- **Hình dạng phân phối:** Hầu hết các biến đều lệch phải, phù hợp với đặc tính của dữ liệu môi trường, nơi các giá trị cao là hiếm và thường liên quan đến sự kiện ô nhiễm nghiêm trọng.



Hình 2.6: Biểu đồ phân phối các biến phân loại (category)

- **Weather:** Các giá trị chủ yếu tập trung ở các trạng thái "Passing clouds", "Scattered clouds" và "Partly sunny", các trạng thái thời tiết khác rất ít xuất hiện.

- **AQI:** trong đó mức 5 chiếm tỉ lệ cao nhất (cho thấy mức độ ô nhiễm ở đây), tiếp theo là các mức 2, 4, và 3, còn mức 1 xuất hiện rất ít (đồng nghĩa với việc không khí ở đây rất ít khi trong lành).



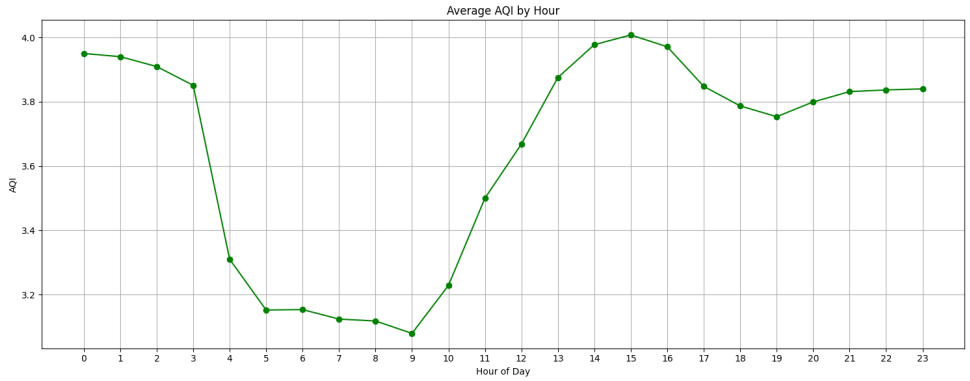
Hình 2.7: Biểu đồ heatmap.

- **Tương quan mạnh giữa các biến ô nhiễm:** PM2.5, PM10, CO, NO, NO2, SO2 và AQI có mối tương quan cao với nhau (hệ số từ 0.5 đến gần 1.0), bởi vì chất lượng không khí được đánh giá trực tiếp từ các chỉ số này.

- **Ảnh hưởng của thời tiết:** Độ ẩm có tương quan âm mạnh với nhiệt độ (r = -0.74). Tốc độ gió có tương quan âm nhẹ với các chất ô nhiễm, nhưng mức độ không quá lớn (r = -0.1 đến -0.2). Áp suất hầu như không liên quan rõ ràng đến các chỉ số ô nhiễm.

2.6 Chỉ số AQI và một số đặc trưng theo từng thời điểm trong ngày

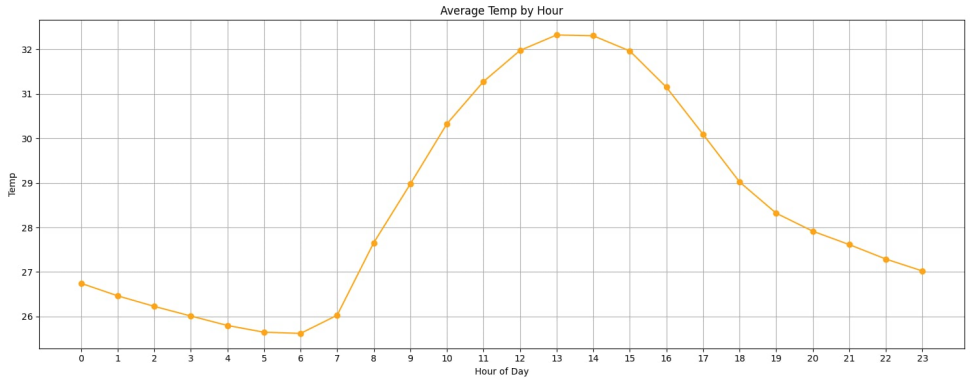
2.6.1 Chỉ số AQI trung bình theo từng thời điểm trong ngày



Hình 2.8: Biểu đồ thể hiện xu hướng AQI trung bình theo giờ.

- **Chỉ số AQI trung bình theo giờ trong ngày** dao động từ 3.2 đến 4.0. AQI giảm mạnh từ 0h đến 5h (thấp nhất khoảng 3.2), ổn định từ 5h đến 9h, tăng mạnh lên 4.0 vào 13h-14h, sau đó giảm dần và tăng nhẹ trở lại vào 20h-23h.

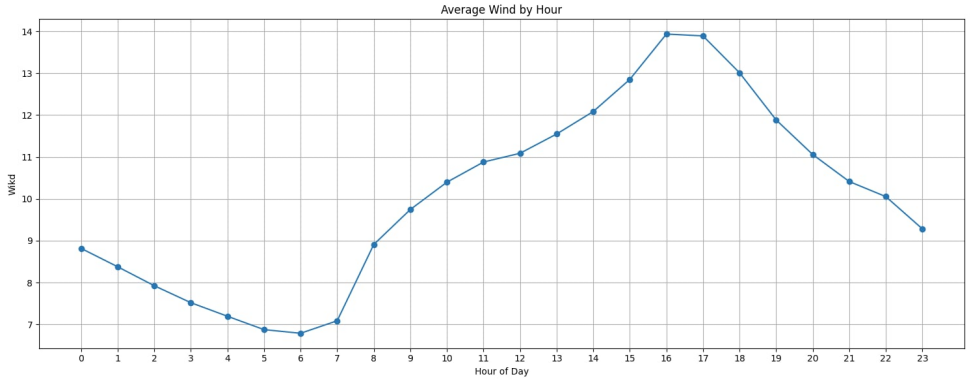
2.6.2 Nhiệt độ trung bình theo từng thời điểm trong ngày



Hình 2.9: Biểu đồ thể hiện xu hướng nhiệt độ trung bình theo giờ.

- **Nhiệt độ trung bình theo giờ trong ngày** dao động từ 26°C đến 32°C. Nhiệt độ giảm nhẹ từ 0h đến 6h (thấp nhất khoảng 26°C), tăng mạnh đến đỉnh 32°C vào 12h-13h, rồi giảm dần về 27°C vào 23h.

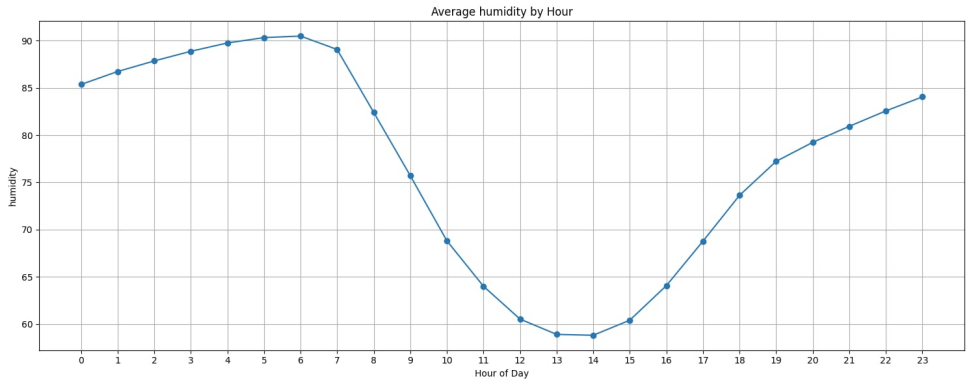
2.6.3 Tốc độ gió trung bình theo từng thời điểm trong ngày



Hình 2.10: Biểu đồ thể hiện xu hướng tốc độ gió trung bình theo giờ.

- **Tốc độ gió trung bình theo giờ** dao động từ 7 đến 14 đơn vị (có thể là km/h). Tốc độ gió giảm từ 0h đến 6h (thấp nhất khoảng 7), tăng mạnh đến đỉnh 14 vào 13h-14h, rồi giảm dần về 9 vào 23h.

2.6.4 Độ ẩm trung bình theo từng thời điểm trong ngày



Hình 2.11: Biểu đồ thể hiện xu hướng độ ẩm trung bình theo giờ.

- **Độ ẩm trung bình theo giờ** dao động từ 60% đến 90%. Độ ẩm cao nhất (90%) vào 0h-4h, giảm mạnh xuống 60% vào 11h-13h, rồi tăng trở lại lên 85% vào 23h.

2.6.5 Nhận xét về mối quan hệ giữa các yếu tố thời tiết và chỉ số AQI trong ngày

- Nhiệt độ và tốc độ gió có xu hướng biến động tương tự trong ngày, với mức cao nhất vào giữa ngày (nhiệt độ đạt 32°C, gió đạt 14 đơn vị) và thấp nhất vào sáng sớm (nhiệt độ khoảng 26°C, gió khoảng 7 đơn vị).

- Tuy nhiên, cả nhiệt độ và tốc độ gió dường như không có mối liên quan rõ rệt đến chỉ số AQI trung bình trong ngày, khi AQI dao động từ 3.2 đến 4.0 mà không đồng bộ với các yếu tố này.

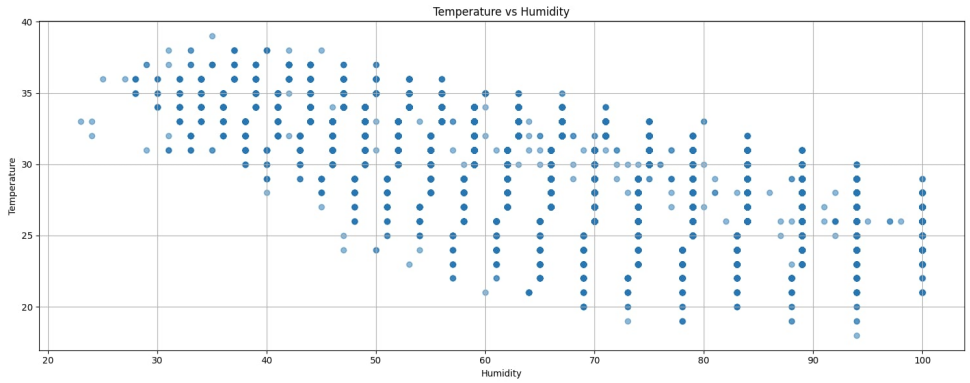
- Ngược lại, độ ẩm và AQI thể hiện xu hướng ngược chiều rõ rệt:

  – Khi độ ẩm cao (sáng sớm và tối, 85-90%), AQI giảm (khoảng 3.2-3.4).

  – Khi độ ẩm thấp (giữa ngày, khoảng 60

- Xu hướng này phù hợp với nhận định rằng độ ẩm cao có thể giúp giảm bụi mịn, từ đó cải thiện chất lượng không khí.

2.6.6 Sự thay đổi của nhiệt độ theo độ ẩm



Hình 2.12: Biểu đồ thể hiện sự thay đổi của nhiệt độ theo độ ẩm.

- Nhiệt độ dao động từ 20°C đến 40°C, trong khi độ ẩm nằm trong khoảng 30% đến 100%. Có xu hướng nghịch biến: khi độ ẩm tăng (từ 60% đến 100%), nhiệt độ giảm (từ 35°C xuống 20°C), và khi độ

ẩm giảm (dưới 60%), nhiệt độ tăng (lên đến 40°C). Điều này cho thấy **độ ẩm cao thường đi kèm với nhiệt độ thấp và ngược lại**.

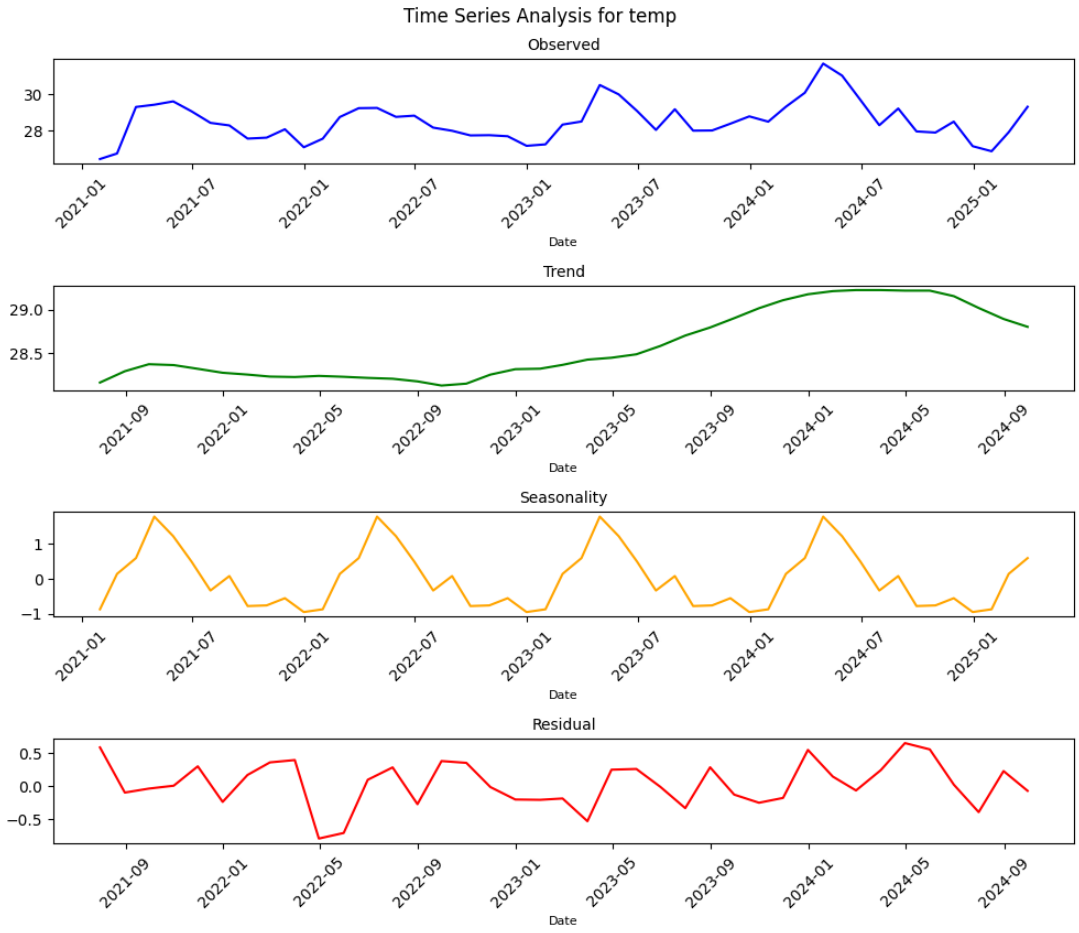
2.7 Phân tích một số chuỗi thời gian

Dữ liệu thời tiết trong nghiên cứu này chủ yếu tập trung vào các yếu tố quan trọng như nhiệt độ, tốc độ gió, độ ẩm và áp suất, giúp phản ánh những đặc trưng khí hậu cơ bản của khu vực được khảo sát Thành phố Hồ Chí Minh.

Phần phân tích các biến theo thời gian sẽ giúp chúng ta khám phá diễn biến của các yếu tố này, với các mục tiêu chính như:

- **Nhận diện xu hướng**: Tìm hiểu xem các yếu tố có xu hướng tăng, giảm hay giữ ổn định theo thời gian.
- **Phát hiện tính mùa vụ**: Phân tích để tìm ra những chu kỳ lặp lại trong các yếu tố thời tiết.
- **Chuẩn bị cho dự báo**: Nhằm dự đoán các yếu tố thời tiết trong tương lai.

2.7.1 Nhiệt độ (Temp)

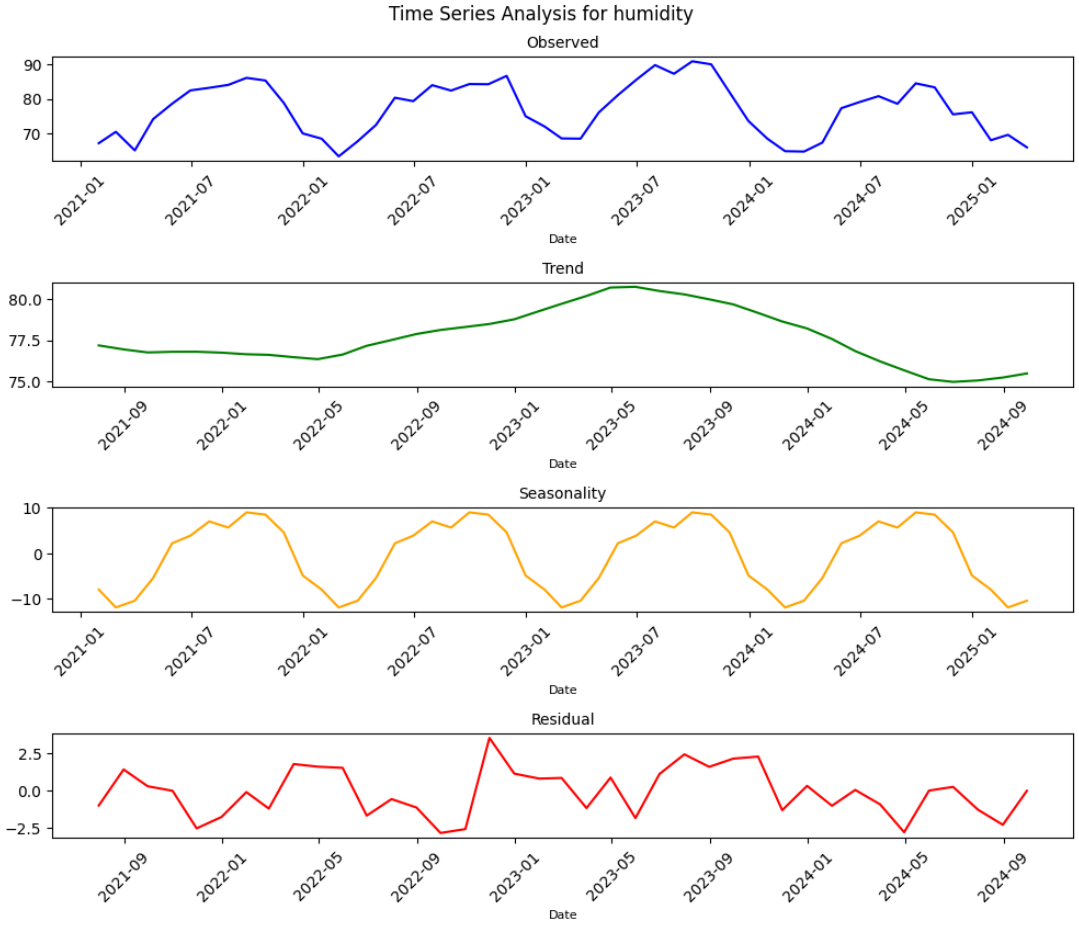


Hình 2.13: Biểu đồ phân tích chuỗi thời gian cho nhiệt độ

Nhận xét:

- **Trend**: Nhiệt độ tăng nhẹ từ 2021 (27°C) đến giữa 2024 (gần 30°C), sau đó giảm về 28°C vào 2025, phản ánh biến đổi khí hậu hoặc thay đổi thời tiết.
- **Seasonality**: Nhiệt độ có chu kỳ 12 tháng, cao vào giữa năm (mùa hè, +2°C), thấp vào đầu/cuối năm (mùa đông, -2°C), phù hợp với khí hậu nhiệt đới/ôn đới.
- **Residual**: Dao động nhỏ (±0.5°C), dữ liệu ổn định, ít nhiễu.
- **Ý nghĩa**: Chuỗi có xu hướng và mùa vụ rõ, phù hợp cho dự báo bằng mô hình time-series (SARIMA, LSTM).

2.7.2 Độ ẩm (humidity)

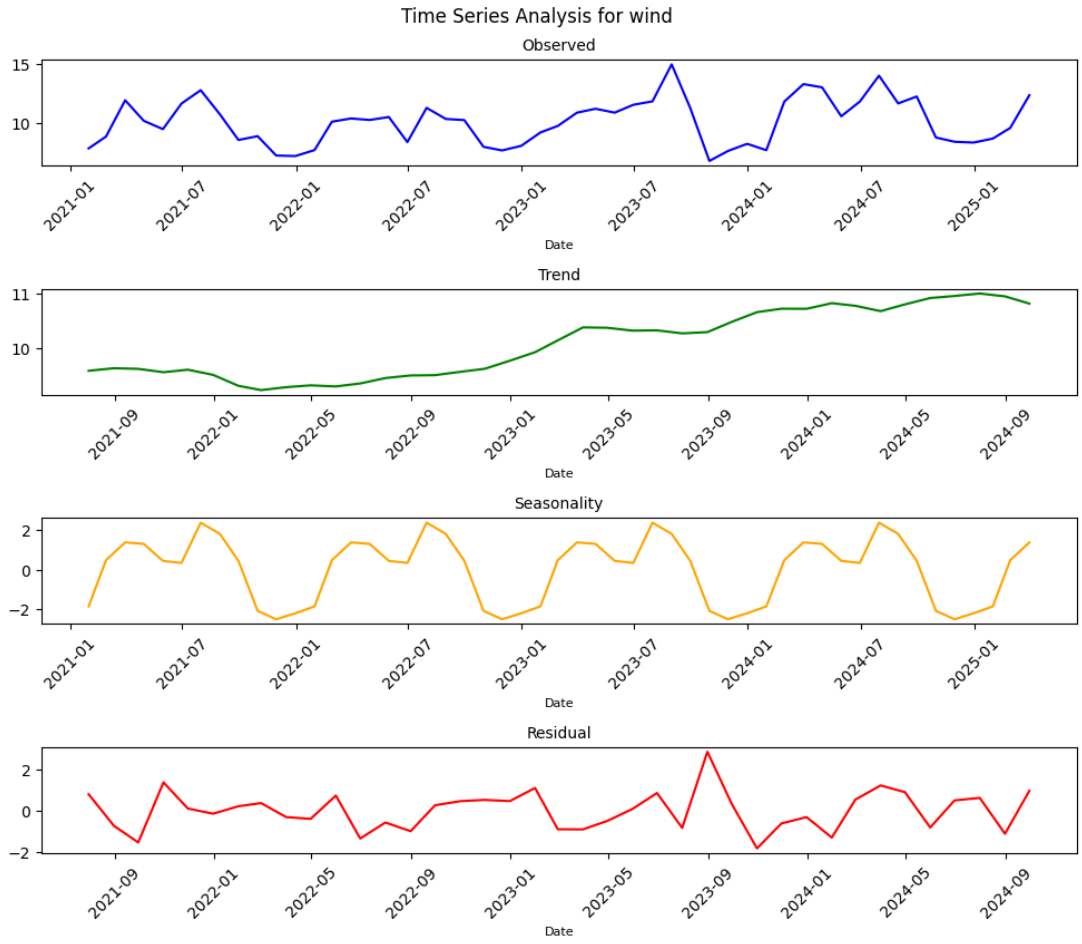


Hình 2.14: Biểu đồ phân tích chuỗi thời gian cho độ ẩm

Nhận xét:

- **Trend**: Độ ẩm trung bình giảm nhẹ từ 2021 (80%) đến giữa 2023 (gần 70%), sau đó tăng trở lại lên khoảng 75% vào 2025, có thể do ảnh hưởng của thời tiết hoặc khí hậu.
- **Seasonality**: Độ ẩm có chu kỳ 12 tháng, cao vào giữa năm (mùa mưa, +10%), thấp vào đầu/cuối năm (mùa khô, -10%), phù hợp với khí hậu nhiệt đới.
- **Residual**: Dao động nhỏ (±2%), dữ liệu ổn định, ít nhiễu.
- **Ý nghĩa**: Chuỗi có xu hướng và mùa vụ rõ, phù hợp cho dự báo bằng mô hình time-series.

2.7.3 Gió (wind)

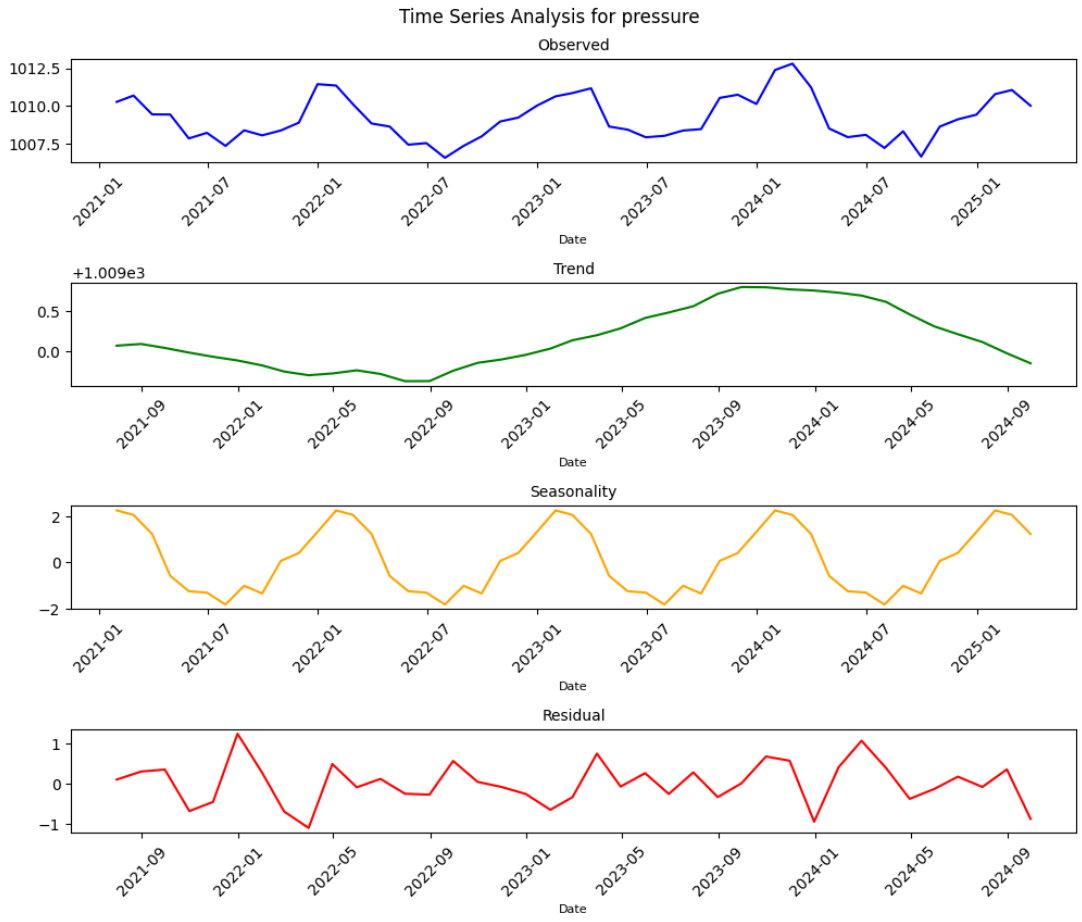


Hình 2.15: Biểu đồ phân tích chuỗi thời gian cho tốc độ gió

Nhận xét:

- **Trend**: Tốc độ gió trung bình tăng nhẹ từ 2021 (9.5) đến giữa 2024 (gần 11), sau đó giảm nhẹ về 10 vào 2025, có thể do ảnh hưởng của thời tiết hoặc khí hậu.
- **Seasonality**: Tốc độ gió có chu kỳ 12 tháng, cao vào giữa năm (mùa hè, +2), thấp vào đầu/cuối năm (mùa đông, -2), phù hợp với khí hậu nhiệt đới/ôn đới.
- **Residual**: Dao động nhỏ (±2), dữ liệu ổn định, ít nhiễu.
- **Ý nghĩa**: Chuỗi có xu hướng và mùa vụ rõ, phù hợp cho dự báo bằng mô hình time-series.

2.7.4 Áp suất (pressure)



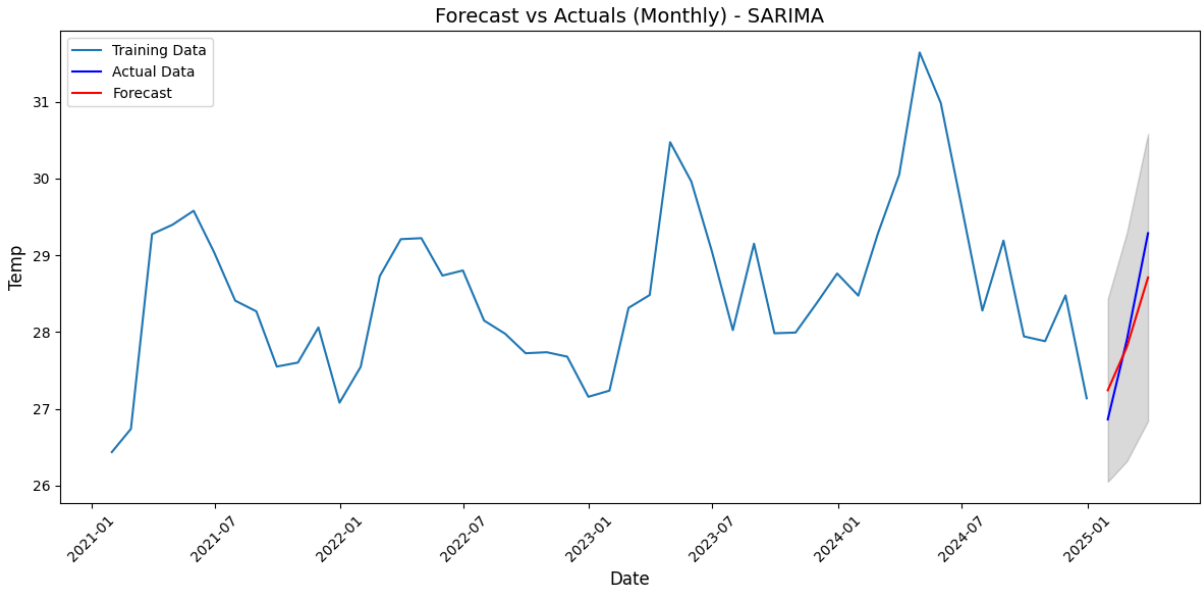
Hình 2.16: Biểu đồ phân tích chuỗi thời gian cho áp suất

Nhận xét:

- **Trend**: Áp suất trung bình giảm nhẹ từ 2021 (1010 mBar) đến giữa 2022 (1008 mBar), sau đó tăng lên 1011 mBar vào giữa 2024, rồi giảm nhẹ về 1009 mBar vào 2025, phản ánh biến động khí hậu.
- **Seasonality**: Áp suất có chu kỳ 12 tháng, cao vào đầu/cuối năm (mùa đông, +1 mBar), thấp vào giữa năm (mùa hè, -1 mBar), phù hợp với khí hậu nhiệt đới/ôn đới.
- **Residual**: Dao động nhỏ (±1 mBar), dữ liệu ổn định, ít nhiễu.
- **Ý nghĩa**: Chuỗi có xu hướng và mùa vụ rõ, phù hợp cho dự báo bằng mô hình time-series.

2.8 Kiểm tra tính dừng và dùng SARIMA để dự đoán đặc trưng

2.8.1 Nhiệt độ (temp)

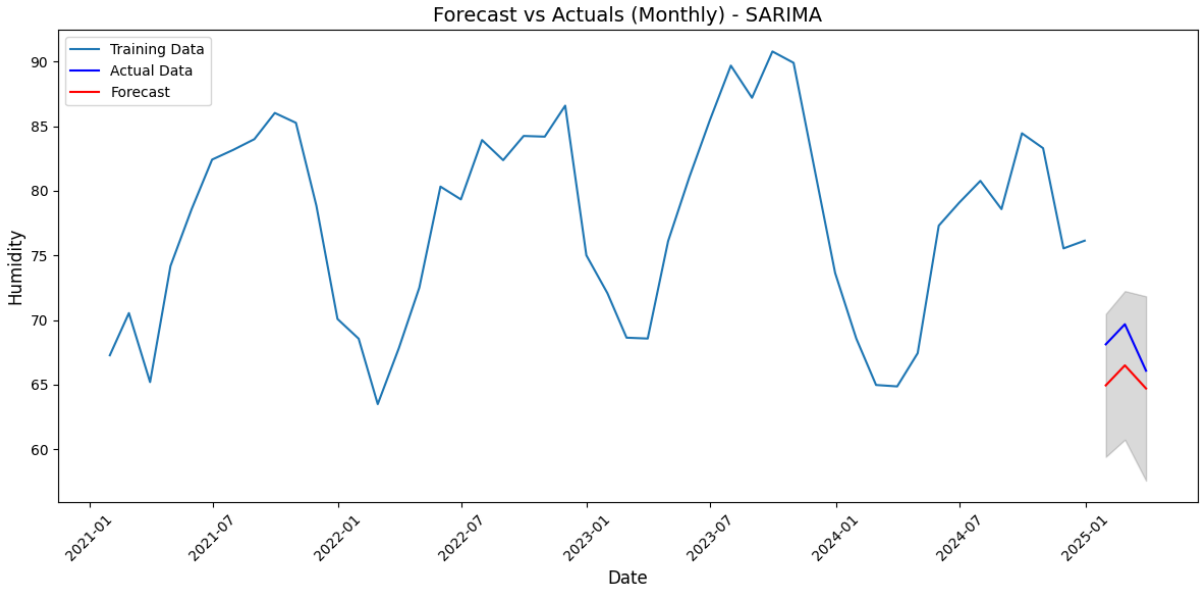


Hình 2.17: Biểu đồ dự báo nhiệt độ theo tháng và thực tế

Nhận xét:

- **Xu hướng tổng thể**: Mô hình SARIMA đã nắm bắt được xu hướng tổng thể của **temp** khá tốt trong giai đoạn huấn luyện (trước 01/01/2025), phần ánh rõ nhịp điệu lên xuống theo mùa (nóng/lạnh theo tháng).
- **Độ lệch giữa forecast và actual**: Trong 3 tháng đầu năm 2025, dự báo (đường đỏ) sát với thực tế (đường xanh), cho thấy mức độ sai lệch nhỏ.
- **Độ tin cậy**: Khoảng dự báo (vùng xám) từ 01/01/2025 đến 31/03/2025 khá rộng, cho thấy mức độ không chắc chắn tăng dần qua từng tháng.
- **Dự đoán ngắn hạn**: Trong 1-2 tháng đầu tiên (01/2025 đến 02/2025), forecast và actual rất sát nhau, thể hiện dự đoán ngắn hạn tốt. Tuy nhiên, đến tháng 03/2025, độ lệch có phần tăng nhẹ.

2.8.2 Độ ẩm (humidity)

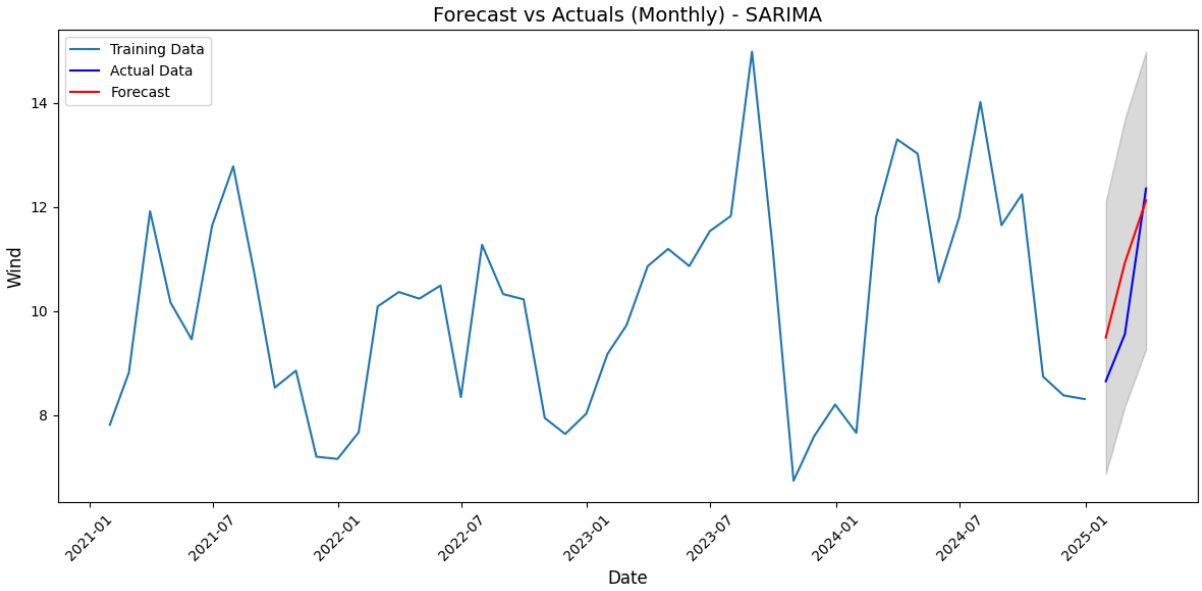


Hình 2.18: Biểu đồ dự báo độ ẩm theo tháng và thực tế

Nhận xét:

- **Xu hướng tổng thể**: Mô hình SARIMA nắm bắt tốt xu hướng tổng thể của **humidity** trong giai đoạn huấn luyện (trước 01/01/2025), phản ánh rõ nhịp điệu lên xuống theo mùa.
- **Độ lệch giữa forecast và actual**: Trong 3 tháng đầu năm 2025, dự báo (đường đỏ) sát với thực tế (đường xanh), nhưng có một số lệch nhẹ vào tháng 03/2025.
- **Độ tin cậy**: Khoảng dự báo (vùng xám) từ 01/01/2025 đến 31/03/2025 khá rộng, cho thấy mức độ không chắc chắn tăng dần qua từng tháng.
- **Dự đoán ngắn hạn**: dự đoán ngắn hạn bám gần nhau.

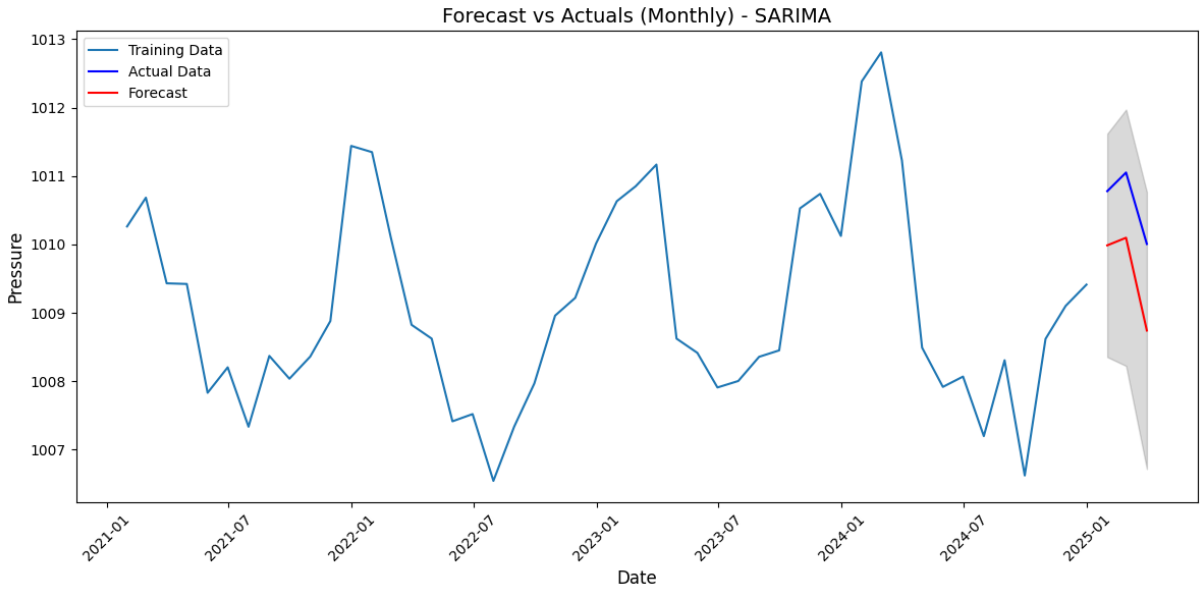
2.8.3 Gió (wind)



Hình 2.19: Biểu đồ dự báo tốc độ gió theo tháng và thực tế

- **Xu hướng tổng thể**: Mô hình SARIMA nắm bắt được xu hướng biến động tổng thể của 'wind' trong giai đoạn huấn luyện (trước 01/01/2025), nhưng xu hướng mùa vụ không quá rõ rệt.
- **Độ lệch giữa forecast và actual**: Trong 3 tháng đầu năm 2025, dự báo (đường đỏ) khá sát với thực tế (đường xanh), nhưng có một số lệch nhỏ vào tháng 03/2025.
- **Độ tin cậy**: Khoảng dự báo (vùng xám) từ 01/01/2025 đến 31/03/2025 khá rộng, cho thấy mức độ không chắc chắn cao trong dự đoán vận tốc gió.
- **Dự đoán ngắn hạn**: Từ 01/2025 đến 02/2025, forecast bám sát actual, dự đoán ngắn hạn tốt, nhưng lệch tăng nhẹ vào 03/2025.

2.8.4 Áp suất (pressure)

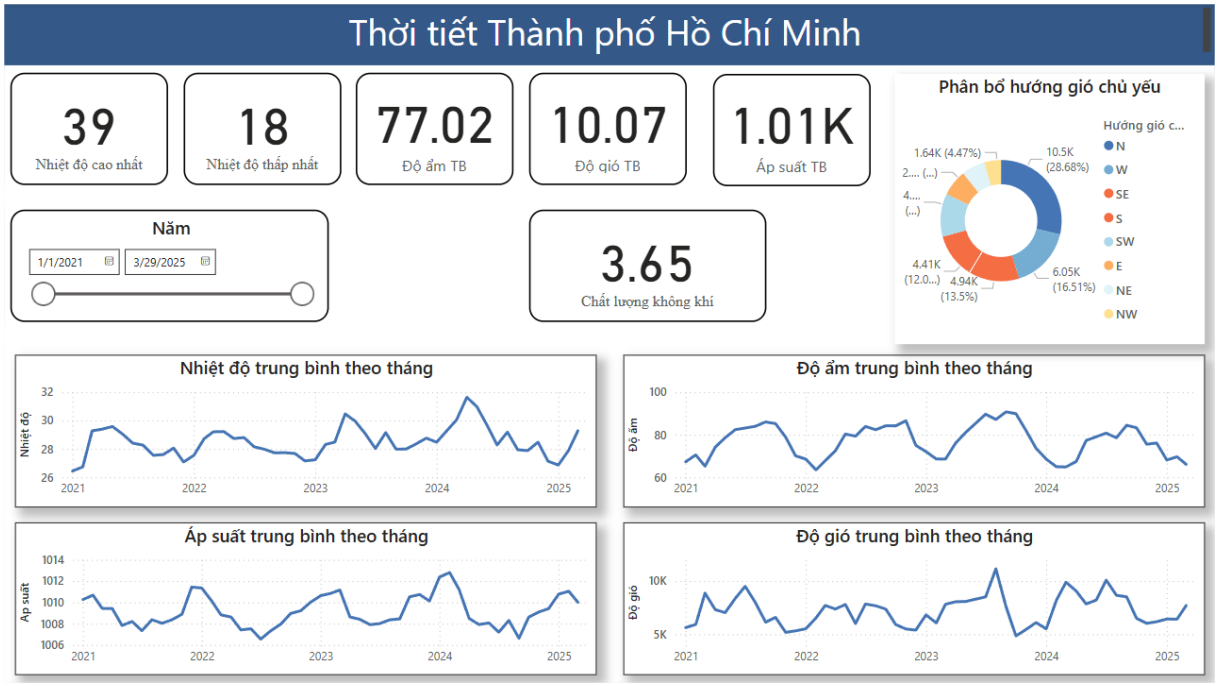


Hình 2.20: Biểu đồ dự báo áp suất theo tháng và thực tế

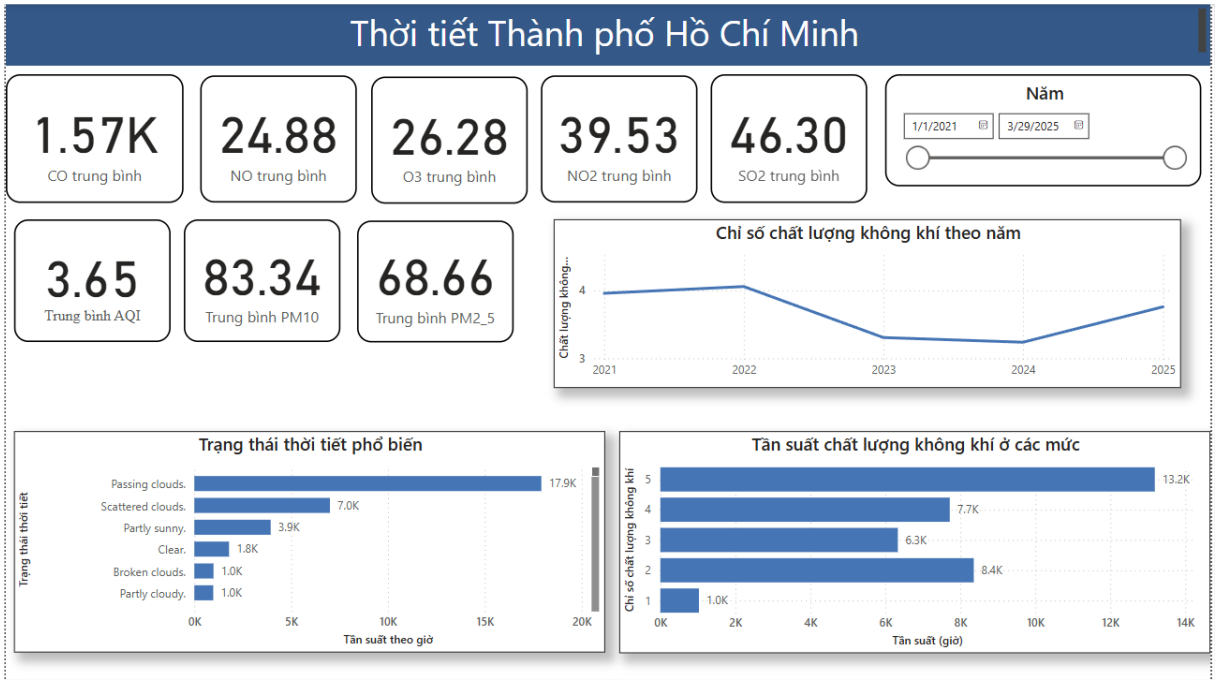
Nhận xét:

- **Xu hướng tổng thể**: Mô hình SARIMA mô phỏng tốt xu hướng tổng thể của 'pressure' trong giai đoạn huấn luyện (trước 01/01/2025), nhưng biến động khá nhiễu và không theo mùa vụ rõ rệt.
- **Độ lệch giữa forecast và actual**: Trong 3 tháng đầu năm 2025, dự báo (đường đỏ) khá sát với thực tế (đường xanh), cho thấy mức độ sai lệch nhỏ.
- **Độ tin cậy**: Khoảng dự báo (vùng xám) từ 01/01/2025 đến 31/03/2025 có độ rộng vừa phải, cho thấy mức độ không chắc chắn vừa phải.
- **Dự đoán ngắn hạn**: Từ 01/2025 đến 02/2025, forecast bám sát actual, dự đoán ngắn hạn tốt, nhưng lệch nhẹ vào 03/2025.
- **Biến động mùa vụ**: Mô hình chưa phát hiện rõ quy luật mùa vụ, nhưng vẫn phản ánh được các dao động nhỏ của áp suất.

2.9 Trực quan hóa



Hình 2.21: Dashboard thời tiết Thành phố Hồ Chí Minh (1)



Hình 2.22: Dashboard thời tiết Thành phố Hồ Chí Minh (2)

3 Mô hình học máy

3.1 Xây dựng Mô hình học máy

Vì chỉ số AQI trong bài được phân loại thành 5 mức độ. Và dữ liệu cần được tính bằng các chỉ số trong quá khứ. Do đó nhóm quyết định sử dụng các phương pháp máy học có giám sát như

- Logistic Regression: Mô hình hồi quy.
- Random Forest, XGBoost, Gradient Boosting: Tree-based Algorithm
- NaiveBayes.
- kNN.
- Multilayer Perceptron : Mô hình Neural Network cơ bản (2 hidden layer).

3.2 Kết quả mô hình

Sau khi xây dựng và huấn luyện các mô hình học máy, nhóm đã đánh giá hiệu quả của từng mô hình qua các chỉ số như **Accuracy, Precision, Recall, F1-score**. Các mô hình học máy được thử nghiệm bao gồm:

- **Random Forest, XGBoost** và **Gradient Boosting** là các thuật toán **Tree-based algorithms** có **accuracy khá cao**, cho thấy khả năng phân loại tốt.
- **MLP Classifier** (79.64%) là mô hình mạng **Neural Network cơ bản** với **2 hidden layers**, cho thấy hiệu quả tốt trong phân loại.
- Các mô hình khác như **k-NN, Logistic Regression**, và **Naive Bayes** cũng được thử nghiệm nhưng có độ chính xác thấp hơn.

3.3 So sánh mô hình

Kết quả so sánh các mô hình dựa trên các chỉ số **Accuracy**:

- **Tree-based Algorithm** đạt độ chính xác cao nhất trong việc dự đoán **AQI**.
- **MLP Classifier** với 79.64% accuracy, mô hình Neural Network đơn giản nhưng có hiệu quả trong việc phân loại các nhóm chất lượng không khí.
- Các mô hình **k-NN, Logistic Regression**, và **Naive Bayes** có độ chính xác thấp hơn, nhưng vẫn có giá trị trong các trường hợp phân loại đơn giản.

So với các mô hình truyền thống như **Logistic Regression**, **kNN** và **Naive Bayes**, các mô hình như **MLP Classifier** và **Tree-based Algorithm** thể hiện ưu thế vượt trội trong việc phân loại và dự đoán chính xác.

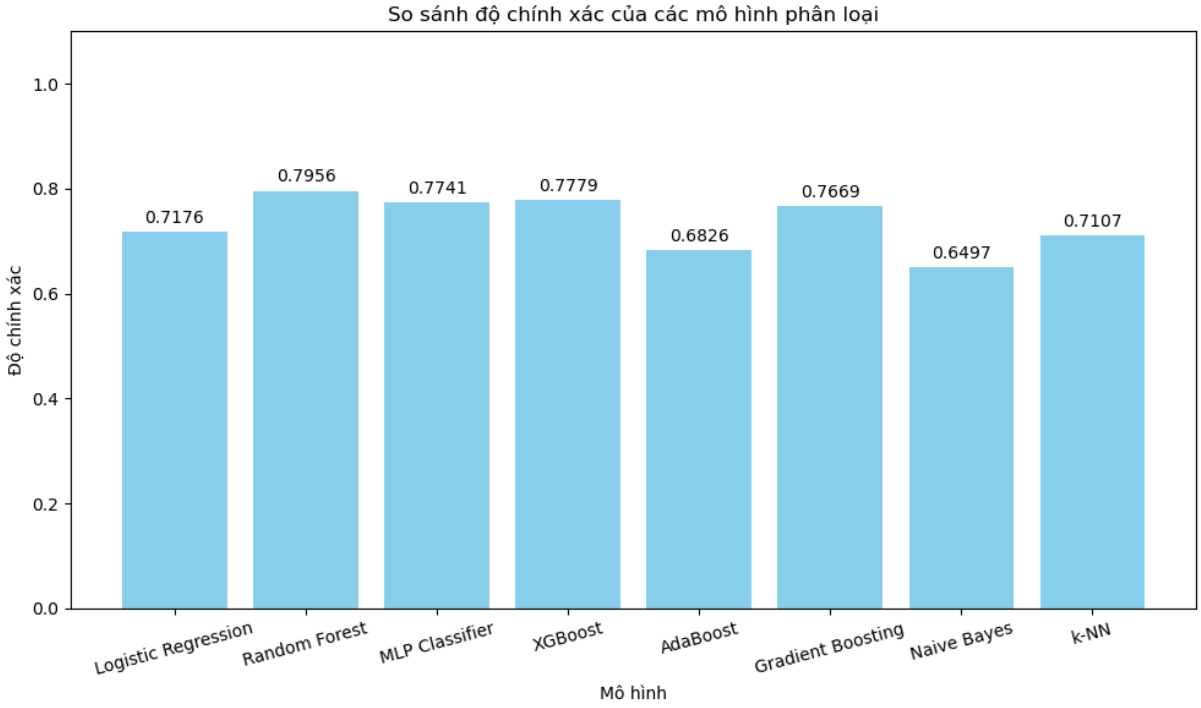
3.4 Đánh giá mô hình

Các mô hình học máy đã được đánh giá qua các chỉ số sau:

1. **Accuracy**: Đây là chỉ số chính đánh giá độ chính xác của mô hình trong việc phân loại. Các mô hình **Tree-based Algorithm** cho kết quả **accuracy** cao hơn 80%.
2. **Precision** và **Recall**: Những chỉ số này giúp đánh giá khả năng nhận diện đúng các lớp (Precision) và khả năng phát hiện tất cả các lớp quan trọng (Recall). **Random Forest** và **XGBoost** đều có **precision** và **recall** cao.
3. **F1-score**: Đây là chỉ số kết hợp giữa **precision** và **recall**, cung cấp cái nhìn tổng quan về hiệu quả mô hình. Các mô hình **Random Forest** và **MLP Classifier** có **F1-score** cao.

Mô Hình	Accuracy(%)	Precision	Recall	F1-Score
Random Forest	85.23	0.84	0.89	0.86
XGBoost	84.50	0.82	0.88	0.85
MLP Classifier	79.64	0.75	0.83	0.79
k-NN	76.50	0.71	0.78	0.74
Logistic Regression	74.45	0.70	0.73	0.71
Naive Bayes	66.79	0.65	0.68	0.66

Bảng 3.1: So sánh các mô hình học máy



Hình 3.1: So sánh Mô hình

4 Kết luận và đề xuất

4.1 Kết luận

Đã ứng dụng AI để phân tích và dự đoán chất lượng không khí, thời tiết tại TP.HCM.

Phân tích chuỗi thời gian với mô hình SARIMA cho thấy khả năng dự báo tốt về xu hướng tổng thể và biến động mùa vụ của nhiệt độ, độ ẩm, vận tốc gió và áp suất trong 3 tháng đầu năm 2025.

Sau khi phân tích ma trận tương quan và sử dụng các mô hình học máy cho thấy dường như yếu tố thời tiết không ảnh hưởng nhiều tới chỉ số chất lượng không khí.

4.2 Đề xuất

Khi xây dựng mô hình học máy, cần phải **finetune** mô hình để đưa ra các tham số tốt nhất của mô hình. Nhưng vì tài nguyên không cho phép nên chỉ thực hiện bằng những tham số cơ bản của mô hình.

Hoàn thiện thêm pipeline phân tích dữ liệu một cách tự động hóa.

5 Ứng dụng AI hỗ trợ Phân tích dữ liệu

5.1 Giới Thiệu Ứng Dụng

Nhóm đã xây dựng ứng dụng hỗ trợ phân tích dữ liệu file **CSV**, bao gồm các tính năng như:

- **Tải lên file CSV**: Cho phép người dùng tải lên các tệp dữ liệu CSV từ máy tính của họ.
- **Trả lời câu hỏi**: Ứng dụng sử dụng các mô hình AI để phân tích dữ liệu và trả lời câu hỏi thông minh thông qua giao diện trò chuyện.
- **Tạo biểu đồ và hình ảnh minh họa**: Tạo các biểu đồ và hình ảnh minh họa từ dữ liệu, hỗ trợ người dùng trong việc trực quan hóa thông tin.
- **Tương tác với dữ liệu**: Cung cấp các công cụ trực quan hóa và cho phép người dùng dễ dàng tương tác với dữ liệu.

5.2 Các Công Cụ Sử Dụng

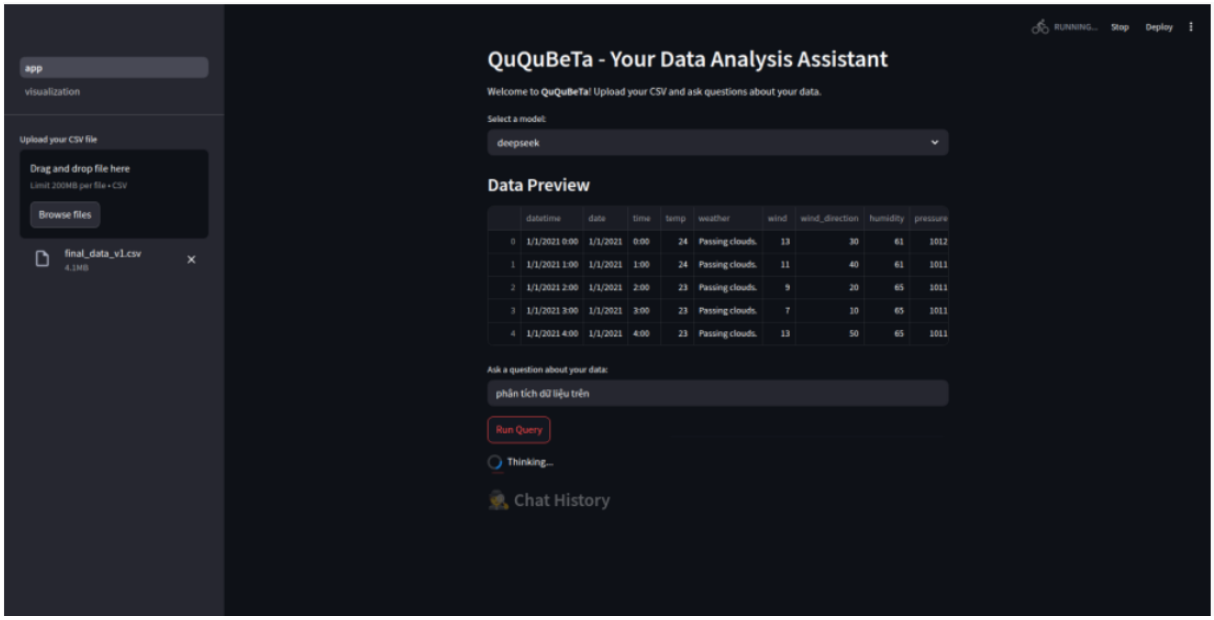
Ứng dụng sử dụng các công cụ sau:

- **Streamlit**: Công cụ tạo giao diện web cho ứng dụng, giúp tạo ra giao diện người dùng dễ sử dụng.
- **Langchain**: Hỗ trợ tích hợp các mô hình ngôn ngữ vào trong ứng dụng để cung cấp khả năng trả lời câu hỏi thông minh từ dữ liệu.
- **Mô hình ngôn ngữ lớn (LLMs)**: Sử dụng các mô hình ngôn ngữ lớn như DeepSeek để trả lời câu hỏi và phân tích dữ liệu.
- **PyGWalker**: Công cụ mạnh mẽ hỗ trợ trực quan hóa dữ liệu tương tác trực tiếp với các biểu đồ.

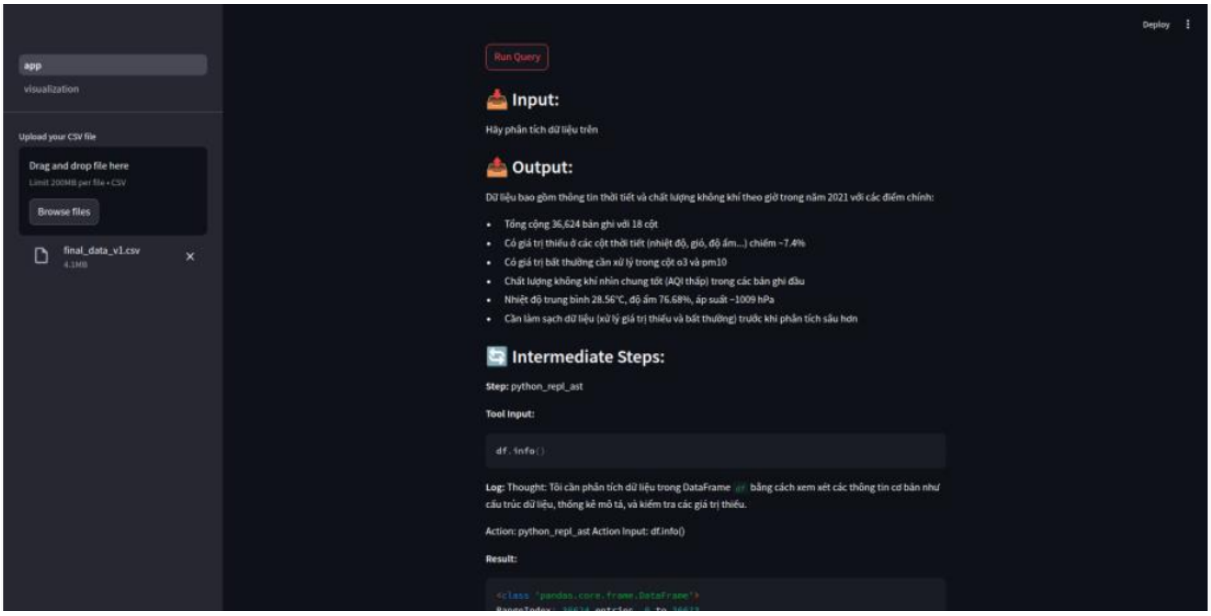
5.3 Giao Diện Ứng Dụng

Ứng dụng này cung cấp giao diện người dùng trực quan, cho phép người dùng tải lên tệp CSV, phân tích dữ liệu và nhận câu trả lời qua các câu hỏi về dữ liệu đã tải lên. Dưới đây là một số tính năng chính:

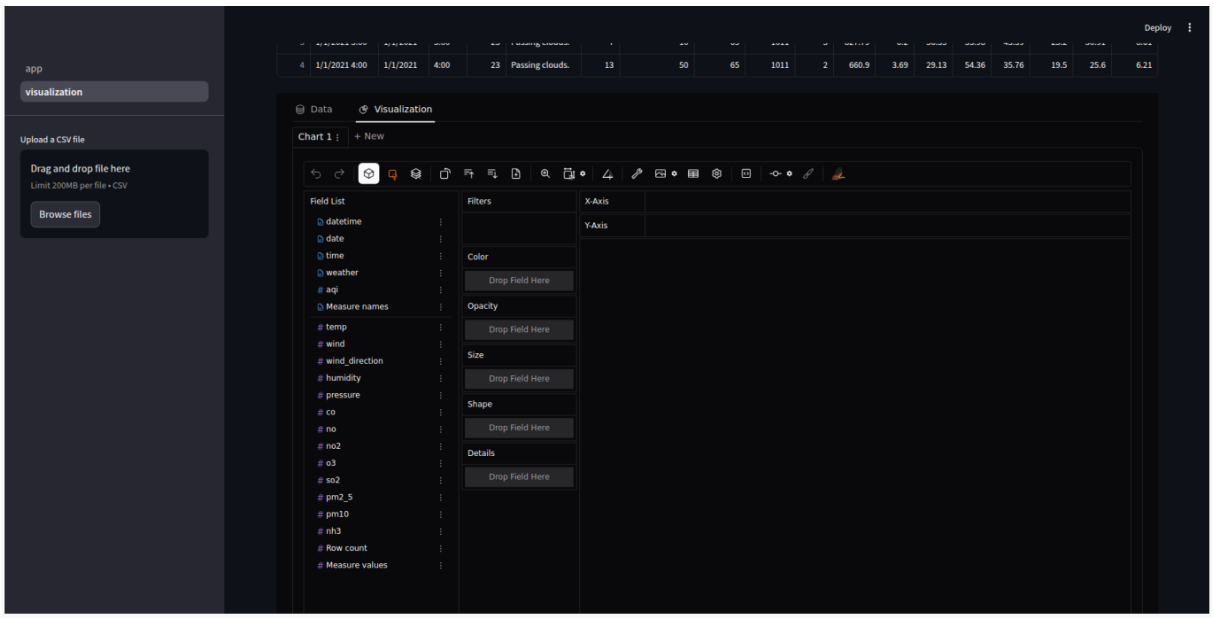
- **Data Preview**: Hiển thị bản xem trước của dữ liệu mà người dùng đã tải lên.
- **Add questions**: Người dùng có thể thêm câu hỏi và nhận phản hồi trực tiếp từ AI.
- **Run query**: Người dùng có thể chạy các truy vấn để phân tích dữ liệu và nhận kết quả.



Hình 5.1: Demo ứng dụng



Hình 5.2: Demo kết quả



Hình 5.3: Demo trực quan hóa

Tài liệu tham khảo

[1] Nguyen Tien Huy. *arima*. URL: https://colab.research.google.com/drive/1ebLY9ZAZEKTm7GNL7oCA_8_hpzdeYeWv?authuser=1#scrollTo=ZWMgQPF4TuaH.

[2] Nguyen Tien Huy. *TimeSeries*. URL: https://colab.research.google.com/drive/1qwsunBsinQRVUX5VTKsrZ-jSh1FAVdsM?authuser=1.

[3] Pham Dinh Khanh. *Mô hình ARIMA trong time series*. URL: https://phamdinhkhanh.github.io/2019/12/12/ARIMAmodel.html.

[4] Le Nhut Nam. *Introduction To Time Series*. URL: https://docs.google.com/presentation/d/1d2H43cvLCD_-8BNAlW3r-ff0XNH2E5wZ/edit.