ĐẠI HỌC HUẾ
KHOA KỸ THUẬT VÀ CÔNG NGHỆ
BỘ MÔN KHOA HỌC DỮ LIỆU VÀ TRÍ TUỆ NHÂN TẠO

VÕ HOÀNG PHÚC THỊNH
LÊ THANH LINH

ĐỒ ÁN 2

ĐỒ ÁN
PHÂN TÍCH DỮ LIỆU THỜI TIẾT

THÀNH PHỐ HUẾ, NĂM 2025

ĐẠI HỌC HUẾ
KHOA KỸ THUẬT VÀ CÔNG NGHỆ
BỘ MÔN KHOA HỌC DỮ LIỆU VÀ TRÍ TUỆ NHÂN TẠO

VÕ HOÀNG PHÚC THỊNH - 22E1020019
LÊ THANH LINH - 22E1010007

PHÂN TÍCH DỮ LIỆU THỜI TIẾT

ĐỒ ÁN
ĐỒ ÁN 2

Giảng viên hướng dẫn:
TS. Nguyễn Đăng Trị

THÀNH PHỐ HUẾ, NĂM 2025

LỜI CAM ĐOAN

Chúng tôi xin cam đoan đây là công trình nghiên cứu của riêng chúng tôi và được sự hướng dẫn đồ án khoa học của thầy Nguyễn Đăng Trị. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và được chính chúng tôi thực hiện. Những bảng biểu, dữ liệu phục vụ cho việc phân tích, biểu đồ, nhận xét, được chính chúng tôi thực hiện và các tài liệu tham khảo có ghi rõ trong phần tài liệu tham khảo.

Nếu phát hiện có bất kỳ sự gian lận nào chúng tôi xin hoàn toàn chịu trách nhiệm về nội dung Đồ án của mình. Khoa Kỹ thuật và Công nghệ - Đại học Huế không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

Thành phố Huế, ngày 13 tháng 6 năm 2025

Võ Hoàng Phúc Thịnh

Lê Thanh Linh

(Tác giả ký và ghi rõ họ tên)

MỤC LỤC

Danh mục hình ảnh

Danh mục bảng, biểu

Danh mục viết tắt

KNN	K-Nearest Neighbors
SVM	Support Vector Machine
ES	Expert Systems
XGBoost	Extreme Gradient Boosting
MAE	Mean Absolute Error
RMSE	Root Mean Squared Error
$R^2 Score$	Coefficient of determination

Phần I

MỞ ĐẦU

Chương 1

Tổng quan lý thuyết

1.1 Giới thiệu về học máy

Học máy là một lĩnh vực thuộc trí tuệ nhân tạo, tập trung vào việc nghiên cứu và phát triển các kỹ thuật cho phép hệ thống "học" tự động từ dữ liệu để giải quyết các vấn đề cụ thể.

Hiểu đơn giản, thuật ngữ này nói tới việc con người dạy máy tính nâng cao khả năng thực hiện các tác vụ cụ thể. Cụ thể là cung cấp các dữ liệu và thuật toán có sẵn để máy tính đưa ra dự đoán hoặc tự ra các quyết định. Thông thường, con người chỉ cần lập trình phần mềm với các dòng lệnh cụ thể để máy tính hiểu và thực hiện. Với học máy, máy tính sẽ tự "học" cách giải quyết công việc thông qua những dữ liệu đã được thu thập và cung cấp [1].

1.2 Phân loại học máy

Học máy được phân loại ra 3 thành phần chính:

- Học không giám sát (Unsupervised learning): là một phần của học máy, trong đó mô hình tự tìm kiếm các mẫu và cấu trúc trong dữ liệu mà không cần nhãn có sẵn. Thay vì dựa vào dữ liệu được gán nhãn trước, thuật toán này phân tích và nhóm các điểm dữ liệu dựa trên sự tương đồng và khác biệt dựa trên tệp dữ liệu đầu vào.

- Học có giám sát (Supervised Learning): là một phương pháp trong học máy, trong đó mô hình được huấn luyện bằng cách sử dụng các tập dữ liệu đã được gán nhãn. Thuật toán sẽ học cách nhận diện các mẫu và mối quan hệ giữa dữ liệu đầu vào và đầu ra, từ đó có thể dự đoán chính xác kết quả khi gặp dữ liệu mới trong thực tế.

- Học tăng cường (Reinforcement Learning) là một kỹ thuật trong học máy đào tạo phần mềm đưa ra quyết định nhằm thu về kết quả tối ưu nhất. Kỹ thuật này bắt chước quy trình học thử và sai mà con người sử dụng để đạt được mục tiêu đã đặt ra.

1.3 Lựa chọn mô hình

Trong đồ án lần này, chúng tôi sẽ sử dụng cả 2 phương pháp học đó là học có giám sát và học không giám sát để tiến hành phân tích dữ liệu.

Các mô hình mà chúng tôi lựa chọn để phân tích dữ liệu bao gồm như sau.

1.3.1 K-means

K-means là một thuật toán học không giám sát dùng để phân cụm dữ liệu (clustering). Mục tiêu của K-means là chia tập dữ liệu thành K cụm (clusters) sao cho các điểm trong cùng một cụm thì giống nhau và gần nhau, còn các cụm khác nhau thì xa nhau [2].

Với bài toán phân cụm dữ liệu, cách thức hoạt động được mô tả như sau.

- Chọn số cụm K cần phân chia (do người dùng xác định).
- Khởi tạo ngẫu nhiên K tâm cụm (centroids).
- Gán mỗi điểm dữ liệu vào cụm có tâm gần nhất theo khoảng cách Euclidean.
- Tính lại tâm cụm mới bằng trung bình tất cả điểm trong cụm đó.
- Lặp lại bước 3 và 4 cho đến khi không còn thay đổi về cụm hoặc đạt số vòng lặp tối đa.

Khoảng cách Euclidean trong bài toán phân cụm được mô tả với công thức như sau:

$$d(x, \mu) = \sqrt{\sum_{i=1}^{n} (x_i - \mu_i)^2}. \tag{1.1}$$

Trong đó:

- x: điểm dữ liệu.
- μ: là tâm cụm (centroid).

1.3.2 Linear Regression

Hồi quy tuyến tính (Linear Regression) là một phương pháp trong thống kê và máy học, được sử dụng để mô hình hóa mối quan hệ tuyến tính giữa một hoặc nhiều biến độc lập và một biến phụ thuộc [3].

Trong hồi quy tuyến tính, giả sử rằng mối quan hệ giữa các biến có thể được biểu diễn bằng một đường thẳng tuyến tính. Công thức toán học tổng quát của mô hình hồi quy tuyến tính có dạng như sau:

$$\hat{y} = \boldsymbol{w}^T x + b. \tag{1.2}$$

Trong đó,

- y: giá trị dự đoán (biến phụ thuộc);
- x: biến đầu vào (biến độc lập);
- b: hệ số chênh lệch (bias/intercept);
- $\boldsymbol{w}$: trọng số của mô hình (slope).

1.3.3 Random Forest Regression

Random Forest Regression là một mô hình học máy thuộc nhóm ensemble methods là một tập hợp các mô hình làm việc cùng nhau để đưa ra dự đoán hoặc phân loại, thường cho kết quả tốt hơn so với một mô hình duy nhất, được xây dựng dựa trên tập hợp của nhiều mô hình cây hồi quy nhị phân. Cụ thể, thuật toán sử dụng kỹ thuật bagging để tạo ra nhiều tập huấn luyện ngẫu nhiên từ tập dữ liệu gốc, sau đó huấn luyện từng cây hồi quy độc lập trên các tập con. Dự đoán đầu ra cuối cùng của mô hình được tính bằng trung bình dự đoán của tất cả các cây trong rừng [4].

$$\hat{y}(x) = \frac{1}{M} \sum_{m=1}^{M} T_m(\mathbf{x}). \tag{1.3}$$

- $\hat{y}$: giá trị dự đoán của mô hình;
- $\mathbf{x} \in \mathbb{R}^p$: vector đặc trưng đầu vào, với p đặc trưng;

- M: số lượng cây trong rừng;
- $T_m(\mathbf{x})$: giá trị dự đoán đầu ra của cây hồi quy thứ m tại đầu vào x.

1.3.4 XGBoost

XGBoost (Extreme Gradient Boosting) là một mô hình học máy thuộc họ Gradient Boosting Decision Tree, được thiết kế nhằm tối ưu hóa cả hiệu năng tính toán và độ chính xác dự đoán. XGBoost xây dựng một chuỗi các cây quyết định theo hướng nâng cao, trong đó mỗi cây mới học từ phần sai số còn lại của các cây trước. Điểm mạnh của XGBoost nằm ở khả năng xử lý tối ưu hóa đạo hàm bậc hai, chính quy hóa hàm mục tiêu và song song hóa dữ liệu.

Mô hình hoạt động bằng cách xây dựng một chuỗi các cây hồi quy liên tiếp, trong đó mỗi cây mới được huấn luyện để khắc phục sai số dự đoán của tổng các cây trước đó. So với các biến thể boosting truyền thống, XGBoost bổ sung thêm chính quy hóa hàm mục tiêu nhằm kiểm soát độ phức tạp mô hình, từ đó giảm quá khớp và tăng khả năng tổng quát hóa, giúp XGBoost trở thành một trong những thuật toán phổ biến nhất trong các ứng dụng học máy thực tiễn và nghiên cứu học thuật [5].

Công thức toán học tổng quát của mô hình có dạng:

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i), \quad f_k \in \mathcal{F}. \tag{1.4}$$

Trong đó,

- $\hat{y}_i^{(t)}$: giá trị dự đoán cho mẫu thứ i tại vòng lặp thứ t;
- $f_k(x_i)$: cây hồi quy thứ k áp dụng lên đầu vào x_i;
- $\sum_{k=1}^{t}$: biểu diễn tổng các cây từ 1 đến t, nhằm cộng gộp dần các mô hình yếu;
- $f_k \in \mathcal{F}$: một phần tử trong không gian hàm $\mathcal{F}$, tập hợp tất cả các cây quyết định có thể huấn luyện được;
- $\mathcal{F}$: tập các cây có dạng $f(x) = w_{q(x)}$, với q là hàm ánh xạ dữ liệu đầu vào x đến một nút lá, và w là trọng số tại lá đó.

Phần II

NỘI DUNG

Chương 2

Tổng quan dữ liệu

2.1 Mục tiêu

Mục tiêu của đồ án là tiến hành phân tích bộ dữ liệu về thời tiết nhằm phục vụ cho các hoạt động dự báo, phân tích ảnh hưởng, theo dõi thời tiết một cách chủ động nhằm dự báo và phòng ngừa trước các tình huống thiên tai bão lũ, cụ thể, nhóm sẽ thực hiện các mô hình phân tích, dự đoán như sau:

- Ứng dụng các kỹ thuật phân tích trong học máy để thực hiện phân tích các yếu tố môi trường nhằm phân cụm dự đoán, dự báo biến động thời tiết trước tương lai trong một đến bảy ngày tới nhằm có những biện pháp đề phòng, phòng tránh.
- Trực quan hóa dữ liệu bằng các loại biểu đồ trực quan, cho người đọc có cái nhìn tổng quan về các yếu tố môi trường.

2.2 Mô tả dữ liệu ban đầu

Bộ dữ liệu mà chúng tôi thực hiện các phép phân tích, các thuật toán, mô hình nhằm khai thác chiều sâu, mô phỏng, trực quan hóa của dữ liệu, áp dụng cho các ứng dụng trong thực tế lần này là bộ dữ liệu về thời tiết tại thành phố Hà Nội vào năm 2024.

Bộ dữ liệu về thời tiết Hà Nội là một bộ dữ liệu khá tốt mô tả chi tiết các đặc điểm của thời tiết như bức xạ khuếch tán, điểm sương, cường độ tia UV, hướng gió, gió giật, tốc độ gió, lượng mưa, ... từ ngày cuối năm 2023 đến hết năm 2024. Ngoài ra, còn có các chỉ số ngoài đất liền như áp suất mặt biển.

Tuy nhiên, trong bộ dữ liệu vẫn còn một số điểm dư thừa chưa được tốt, cần khắc phụ như:

- Lượng tuyết rơi: ở trường này dữ liệu toàn bộ đều bằng 0, dữ liệu năm 2024 cho thấy

không hề có tuyết rơi vào ngày cuối năm 2023 vàtoàn bộ năm 2024, nên khi đưa vào sẽ bị dư thừa dữ liệu

- Độ sâu tuyết tích tụ: Vì không có dữ liệu tuyết xuất hiện như đã giải thích ở biến trên nên trường dữ liệu độ dâu tuyết tích tụ cũng bị dư thừa
- Trạng thái cập nhật dữ liệu: toàn bộ dữ liệu đều được cập nhật thì mới có thể được ghi vào hoàn tất trong dataset, nên trường dữ liệu này cũng dư thừa

Hướng khắc phục: Loại bỏ những cột, hàng giá trị dư thừa, không phụ vụ quá nhiều cho quá trình phân tích và trực quan

Với nguồn nguyên liệu từ bộ dữ liệu thời tiết này, chúng tôi có thể đưa ra các nhận định và nêu ra các mô hình có thể áp dụng được đối với bộ dữ liệu này.

Bộ dữ liệu với các cột dữ liệu đa phần là dạng số, ít trường dữ liệu dạng chữ nên việc phân tích các giá trị liên tục, dự đoán các giá trị, chuẩn hóa, phân cụm, dự báo là khá tối ưu cho bộ dữ liệu. Nhận thấy điều đó, chúng tôi quyết định sử dụng các thuật toán của học máy, cụ thể là thuật toán học có giám sát và học không giám sát để thực hiện phân tích mô hình

Tuy nhiên, vì phạm vi bộ dữ liệu chỉ gói gọn trong 1 năm nên việc dự đoán, dự báo phạm vi rộng hơn như 2 năm, 5 năm có thể gặp khó khăn trong việc dự báo chuẩn xác.

Trước sự biến đổi khí hậu đang ngày càng gia tăng, Trái đất đang dần nóng lên, ô nhiễm khí quyển ngày càng lan rộng, đặc biệt là tại thành phố Hà Nội, nơi thường xuyên nằm trong top các nước có mức độ ô nhiễm nhất thế giới [6] thì đề tài “Thu thập và phân tích dữ liệu thời tiết” ra đời với mục đích mong muốn nghiên cứu, đề xuất những giải pháp nhằm giảm thiểu mức độ ô nhiễm tại thủ đô nước ta.

Dưới đây là hình ảnh minh họa bộ dữ liệu thời tiết 2.1

Mốc thời gian	Mức độ mây che phủ	Điểm sương	Bức xạ khuếch tán ngang(DHI)	Bức xạ trực tiếp bình thường(DNI)	Bức xạ mặt trời toàn phần ngang(GHI)
2023-12-31	79	18.9	40	329	207
2024-01-01	59	18.1	40	329	207
2024-01-02	77	19.6	40	329	208
2024-01-03	80	15.4	40	329	208
2024-01-04	100	14	40	330	209
2024-01-05	90	18	40	330	209
2024-01-06	89	18.4	40	331	210
2024-01-07	93	18.6	40	331	211
2024-01-08	92	19.4	40	332	211
2024-01-09	70	20	40	332	212
2024-01-10	99	19.2	41	333	213
2024-01-11	100	16.7	41	333	214
2024-01-12	100	16.9	41	334	214
2024-01-13	100	17.5	41	334	215
2024-01-14	100	18.1	41	335	216
2024-01-15	87	18.8	41	335	217
2024-01-16	97	18.2	41	336	218
2024-01-17	100	19.7	41	336	219
2024-01-18	59	19.6	41	337	220
2024-01-19	66	20.2	41	338	221
2024-01-20	83	20.1	41	338	222
2024-01-21	84	15.7	41	339	223
2024-01-22	88	8.9	42	340	224
2024-01-23	98	3.6	42	340	225

Hình 2.1: Hình ảnh minh họa bộ dữ liệu thời tiết

Bộ dữ liệu của chúng tôi bao gồm 35 cột dữ liệu và 366 dòng, dưới đây là mô tả chi tiết các trường dữ liệu có trong bộ dữ liệu của chúng tôi:

- Mốc thời gian: Dấu thời gian ghi nhận dữ liệu;
- Mức độ mây che phủ: Có thể là phần trăm hoặc chỉ số từ 0 đến 100;
- Điểm sương: Nhiệt độ mà tại đó không khí bão hòa và nước ngưng tụ, đơn vị °C;
- Bức xạ khuếch tán ngang (DHI): Đơn vị W/m²;
- Bức xạ trực tiếp bình thường (DNI): Đơn vị W/m²;
- Bức xạ mặt trời toàn phần ngang (GHI): Đơn vị W/m²;
- max dhi: Giá trị cực đại của DHI trong một khoảng thời gian nhất định;
- max dni: Giá trị cực đại của DNI;
- max ghi: Giá trị cực đại của GHI;
- Bức xạ mặt trời tổng: Đơn vị W/m²;
- Giá trị dự đoán bức xạ khuếch tán ngang(DHI): Đơn vị W/m²;
- Giá trị dự đoán bức xạ mặt trời toàn phần ngang(GHI): Đơn vị W/m²;

- Giá trị dự đoán bức xạ mặt trời tổng: Đơn vị W/m²;
- Nhiệt độ cao nhất: Nhiệt độ cao nhất ghi nhận trong ngày, đơn vị °C;
- Thời điểm nhiệt độ cao nhất: Dấu thời gian tại thời điểm ghi nhận nhiệt độ cao nhất trong ngày;
- Nhiệt độ thấp nhất: Nhiệt độ thấp nhất ghi nhận trong ngày, đơn vị °C;
- Thời điểm nhiệt độ thấp nhất: Dấu thời gian tại thời điểm ghi nhận nhiệt độ thấp nhất trong ngày;
- Nhiệt độ hiện tại: Nhiệt độ nghi nhận tại thời điểm ghi nhận dữ liệu, đơn vị °C;
- Nhiệt độ trung bình: Nhiệt độ trung bình của ngày ghi nhận dữ liệu, đơn vị °C;
- Tốc độ gió lớn nhất: Tốc độ gió lớn nhất ghi nhận trong ngày, đơn vị m/s;
- Thời điểm có hướng gió mạnh nhất: Dấu thời gian tại thời điểm nhiệt độ thấp nhất trong ngày;
- Hướng gió hiện tại: Hướng gió tại thời điểm ghi nhận dữ liệu;
- Tốc độ gió giật: Tốc độ gió giật tại thời điểm thu thập dữ liệu, đơn vị m/s;
- Tốc độ gió trung bình: Tốc độ gió trung bình của ngày ghi nhận dữ liệu, đơn vị m/s;
- Lượng mưa ghi nhận: Lượng mưa trung bình của ngày ghi nhận dữ liệu, đơn vị mm;
- Lượng mưa từ nguồn vệ tinh GPM: Lượng mưa trung bình của ngày ghi nhận dữ liệu được lấy từ vệ tinh;
- Áp suất không khí: Áp suất của không khí tại thời điểm ghi nhận dữ liệu;
- Độ ẩm tương đối: Độ ẩm tương đối trong không khí tại thời điểm ghi nhận dữ liệu;
- Áp suất mặt biển chuẩn: Áp suất mức nước biển tại thời điểm ghi nhận dữ liệu;
- Trạng thái cập nhật dữ liệu: Trạng thái hoàn tất ghi nhận dữ liệu.

Chương 3

Trực quan hóa dữ liệu

3.1 Tiền xử lý dữ liệu

Dữ liệu đầu vào là bộ dữ liệu về thời tiết về các yếu tố môi trường được ghi nhận tại thành phố Hà Nội. Trong bộ dữ liệu này, chúng tôi xin được viết lại bằng tiếng việt nhằm cho người đọc hình dung nhanh chóng về dữ liệu.

Bước đầu tiên trước khi tiền xử lý dữ liệu thì chúng tôi sẽ nhập các thư viện cần thiết cho các bước xử lý và phân tích dữ liệu.

```
import seaborn as sns
from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
from sklearn.decomposition import PCA
import pandas as pd
import matplotlib.pyplot as plt
```

Đoạn mã 3.1: Nhập các thư viện

Sau đó chúng tôi sẽ tiến hành nhập dữ liệu

```
data = pd.read_excel('D:/Doan2 ThayTri/DoAn2/weatherbit_hanoi_2024.xlsx')
```

Đoạn mã 3.2: Nhập dữ liệu

Để phục vụ cho công việc phân tích và trực quan hóa dữ liệu về sau, chúng tôi phải thực hiện quá trình xử lý dữ liệu nhằm lọc, xóa, sửa các cột, hàng dữ liệu dư thừa, lỗi và trống trong bộ dữ liệu.

3.1.1 Loại bỏ dữ liệu dư thừa

Chúng tôi tạm thời xóa bỏ các cột dữ liệu không sử dụng bằng hàm `data.drop()` trong Python vừa không ảnh hưởng đến dữ liệu gốc vừa không ảnh hưởng đến quá trình phân tích.

Các cột dữ liệu như "Lượng tuyết rơi", "Độ sâu tuyết tích tụ", "Trạng thái cập nhật dữ liệu", là các cột mà chúng tôi loại bỏ.

Đoạn mã mà chúng tôi thực hiện loại bỏ dữ liệu dư thừa như sau:

```
xoa_dulieu = data.drop(columns=["Luong tuyet roi",
                                "Do sau tuyet tich tu",
                                "Trang thai cap nhat du lieu"])
data = xoa_dulieu
```

Đoạn mã 3.3: Xóa bỏ dữ liệu thừa

Dưới đây là hình ảnh minh họa 4.2 sau khi đã loại bỏ ba biến "Lượng tuyết rơi", "Độ sâu tuyết tích tụ", "Trạng thái cập nhật dữ liệu".

	Mốc thời gian	Mức độ mây che phủ	Điểm sương	Bức xạ khuếch tán ngang(DHI)	Bức xạ trực tiếp bình thường(DNI)	Bức xạ mặt trời toàn phần ngang(GHI)	max_dhi	max_dni	max_ghi	Bức xạ mặt trời tổng	...	Hướng gió tại thời điểm gió mạnh nhất	Thời điểm có hướng gió mạnh nhất	Hướng gió hiện tại	Tốc độ gió giật	Tốc độ gió trung bình	Lượng mưa ghi nhận	Lượng mưa từ nguồn vệ tinh GPM	Áp suất không khí	Độ ẩm tương đối	Áp suất mặt biển chuẩn
0	2023-12-31	79	18.9	40	329	207	116	911	764	78	...	155	1704006000	155	3.6	0.9	0.0	0.0	1016	82	1017
1	2024-01-01	59	18.1	40	329	207	116	912	764	189	...	135	1704042000	135	7.6	1.6	0.0	0.0	1015	79	1017
2	2024-01-02	77	19.6	40	329	208	116	912	765	149	...	135	1704196800	135	7.4	1.4	0.0	0.0	1015	81	1016
3	2024-01-03	80	15.4	40	329	208	116	912	767	50	...	186	1704261600	186	10.4	1.8	0.0	0.0	1018	79	1019
4	2024-01-04	100	14.0	40	330	209	116	913	768	69	...	197	1704337200	197	6.0	1.4	0.0	0.0	1018	71	1019
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Hình 3.1: Hình ảnh minh họa sau khi loại bỏ dữ liệu thừa

3.1.2 Kiểm tra dữ liệu thiếu

Bước tiếp theo chúng tôi thực hiện xử lý dữ liệu là kiểm tra các dữ liệu thiếu và thay thế chúng, ở đây chúng tôi sử dụng hàm data.isnull() để kiểm tra.

Đoạn mã mà chúng tôi thực hiện kiểm tra dữ liệu thiếu như sau:

```
missing_value = data.isnull().sum()
missing_value
```

Đoạn mã 3.4: Kiểm tra dữ liệu thiếu

Với tổng giá trị thiếu trong dữ liệu là 0.

Dưới đây là bảng kiểm tra dữ liệu thiếu 3.2 được chúng tôi thực hiện với đoạn mã trên.

```
Mốc thời gian                                    0
Mức độ mây che phủ                               0
Điểm sương                                       0
Bức xạ khuếch tán ngang(DHI)                     0
Bức xạ trực tiếp bình thường(DNI)                0
Bức xạ mặt trời toàn phần ngang(GHI)             0
max_dhi                                          0
max_dni                                          0
max_ghi                                          0
Bức xạ mặt trời tổng                             0
Giá trị dự đoán bức xạ khuếch tán ngang(DHI)     0
Giá trị dự đoán bức xạ trực tiếp bình thường(DNI)  0
Giá trị dự đoán bức xạ mặt trời toàn phần ngang(GHI)  0
Giá trị dự đoán bức xạ mặt trời tổng             0
Nhiệt độ cao nhất                                0
Thời điểm nhiệt độ cao nhất                      0
Nhiệt độ thấp nhất                               0
Thời điểm nhiệt độ thấp nhất                     0
Nhiệt độ hiện tại                                0
Thời điểm ghi nhận nhiệt độ hiện tại             0
Nhiệt độ trung bình                              0
Cường độ tia UV max                              0
Tốc độ gió lớn nhất                              0
Hướng gió tại thời điểm gió mạnh nhất            0
Thời điểm có hướng gió mạnh nhất                 0
...
Độ ẩm tương đối                                  0
Áp suất mặt biển chuẩn                           0
dtype: int64
Tổng số giá trị thiếu trong dataframe là 0
```

Hình 3.2: Hình ảnh minh họa sau khi kiểm tra dữ liệu thiếu

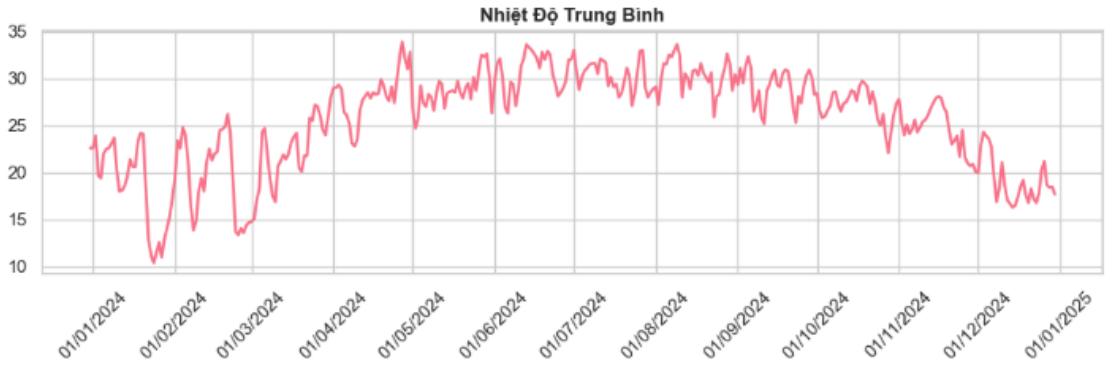
Đối với các bước chuẩn hóa dữ liệu, giảm chiều dữ liệu sẽ được thực hiện riêng ở từng bài toán vì mỗi bài toán sẽ thực hiện mỗi cột, trường dữ liệu khác nhau nên khi thực hiện bài toán nào, cần dùng trường dữ liệu nào thì chúng tôi sẽ tiến hành chuẩn hóa và giảm chiều dữ liệu với trường, cột dữ liệu đó.

3.2 Trực quan hóa dữ liệu

3.2.1 Biểu đồ xu hướng

Biểu đồ xu hướng (hay còn gọi là biểu đồ đường hồi quy, hoặc trend chart/trend line chart) là một loại biểu đồ được sử dụng để minh họa sự thay đổi của một hoặc nhiều biến số theo thời gian hoặc theo một thứ tự nhất định, đồng thời chỉ ra hướng phát triển tổng thể (xu hướng) của dữ liệu đó.

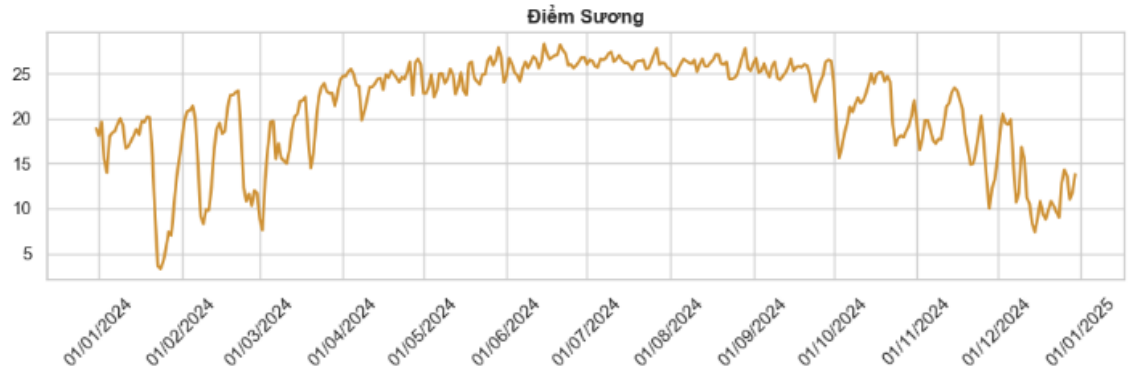
Dưới đây chúng tôi sẽ vẽ các biểu đồ xu hướng của các yếu tố thời tiết trong năm 2024.



Hình 3.3: Biểu đồ xu hướng nhiệt độ trung bình

Nhận xét

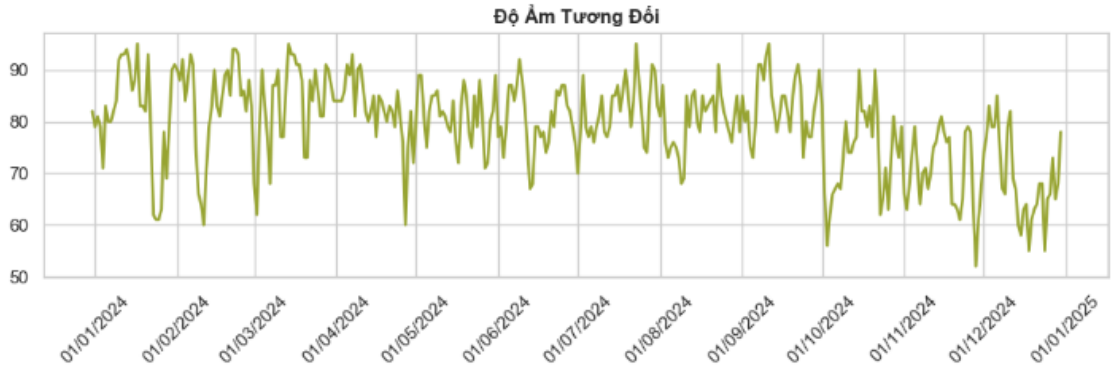
Biểu đô xu hướng nhiệt độ trung bình 3.3 cho ta cái nhìn tổng quát về nhiệt độ trung bình của năm 2024, với xu hướng nhiệt độ trung bình đầu năm khá thấp trong khoảng từ 10-25°C, tăng dần từ tháng 4 đến tháng 10, sau đó giảm dần về cuối năm. Điều này phản ánh đặc trưng khí hậu theo mùa, với mùa hè nóng và mùa đông mát mẻ.



Hình 3.4: Biểu đồ xu hướng điểm sương

Nhận xét

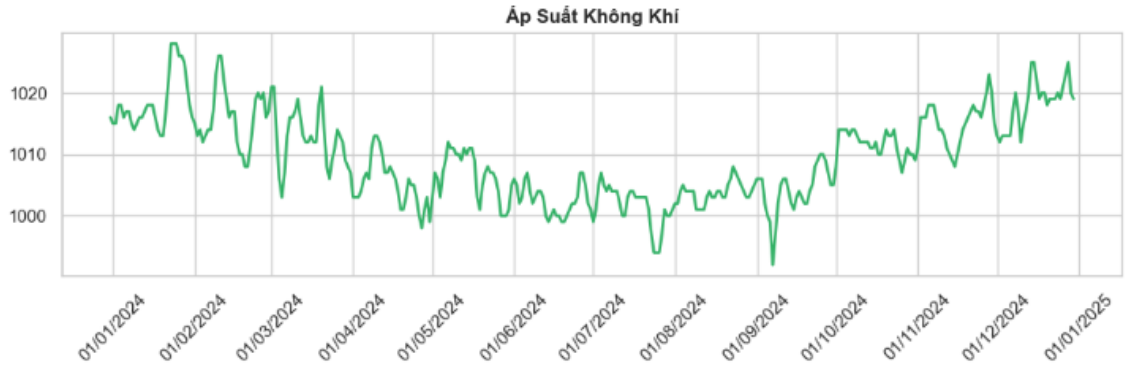
Biểu đồ điểm sương 3.4 có xu hướng biến động tương tự nhiệt độ, đầu năm điểm sương thấp, tăng vào mùa hè và giảm vào mùa đông. Điều này hợp lý vì điểm sương thường tăng cùng với độ ẩm và nhiệt độ cao.



Hình 3.5: Biểu đồ xu hướng độ ẩm tương đối

Nhận xét

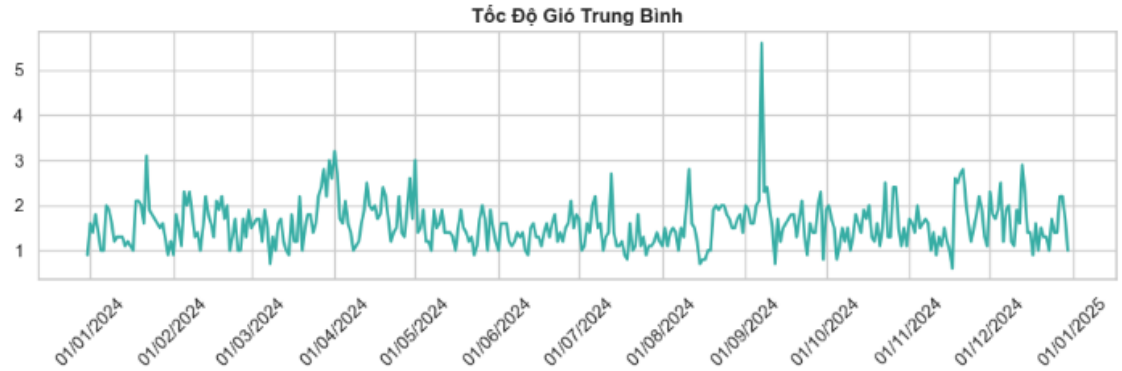
Biểu đồ độ ẩm 3.5 dao động quanh mức 70% – 90% trong suốt năm, với một số thời điểm giảm mạnh. Điều này cho thấy độ ẩm duy trì ở mức khá cao, phù hợp với khí hậu nhiệt đới gió mùa như thành phố Hà Nội, nhưng đôi khi có những giai đoạn khô hạn ngắn.



Hình 3.6: Biểu đồ xu hướng áp suất không khí

Nhận xét

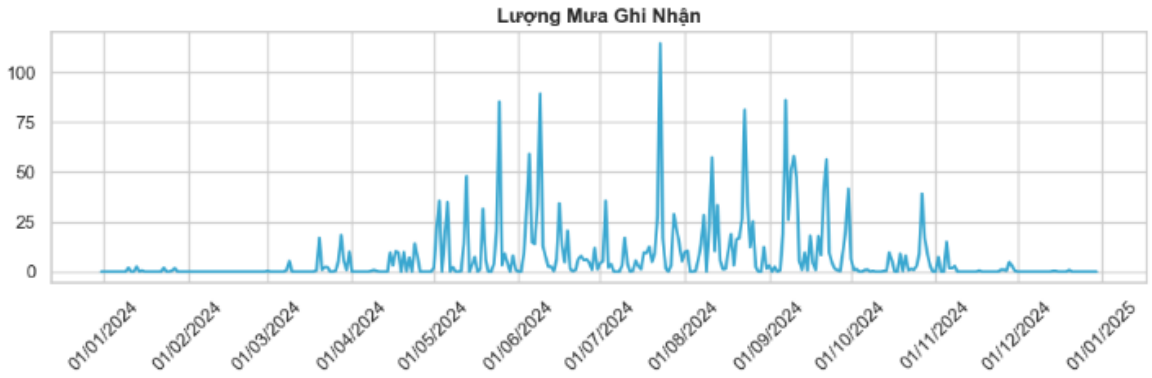
Biểu đồ áp suất không khí 3.6 cho thấy áp suất không khí có xu hướng mạnh vào đầu năm, sau đó giảm nhẹ đến giữa năm và tăng trở lại vào cuối năm. Các biến động nhẹ phản ánh sự thay đổi do thời tiết, đặc biệt là khi thành phố Hà Nội có hệ thống thời tiết nhiệt đới gió mùa.



Hình 3.7: Biểu đồ xu hướng tốc độ gió trung bình

Nhận xét

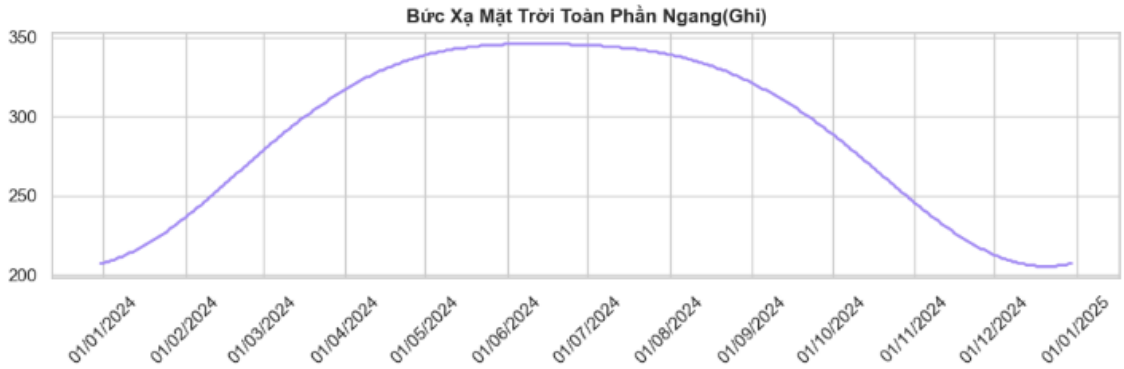
Tốc độ gió trung bình 3.7 khá ổn định quanh mức 1–2 m/s, thỉnh thoảng có các đỉnh nhọn phản ánh những cơn gió mạnh bất thường. Điều này cho thấy khí hậu nói chung ít gió mạnh, nhưng vẫn có những ngày gió mạnh đột ngột.



Hình 3.8: Biểu đồ xu hướng lượng mưa ghi nhận

Nhận xét

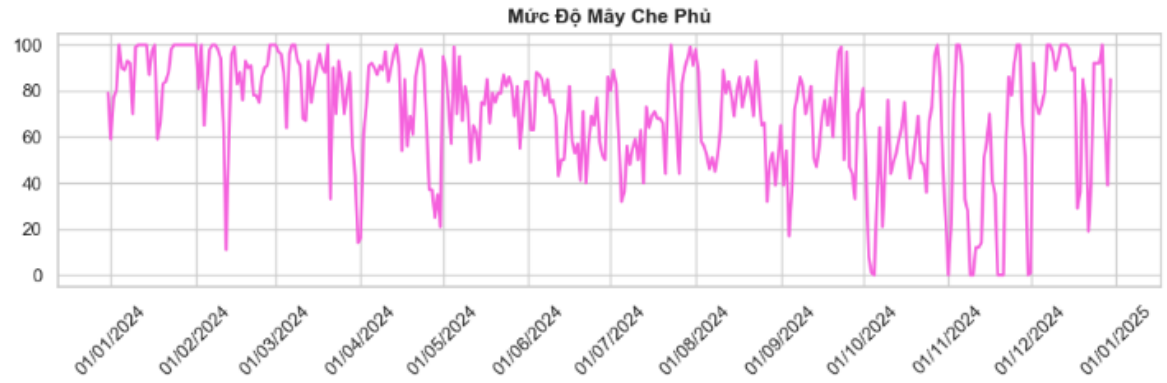
Biểu đồ xu hướng lượng mưa 3.8 có xu hướng tăng mạnh vào giữa năm, đặc biệt từ tháng 5 đến tháng 10, thời điểm mà mùa hè mưa bất chợt và bước sang thu với kiểu thời tiết ẩm lạnh. Biểu đồ này phản ánh một mô hình mưa theo mùa rõ rệt, đặc trưng của vùng khí hậu nhiệt đới.



Hình 3.9: Biểu đồ đường xu hướng bức xạ mặt trời toàn phần ngang

Nhận xét

Biểu đồ dạng parabol thể hiện mức bức xạ mặt trời 3.9 tăng cao khi bước vào giữa năm, và giảm dần về hai đầu năm. Điều này phù hợp với số giờ nắng trong ngày dài hơn vào mùa hè và giờ nắng ngắn hơn vào mùa đông.



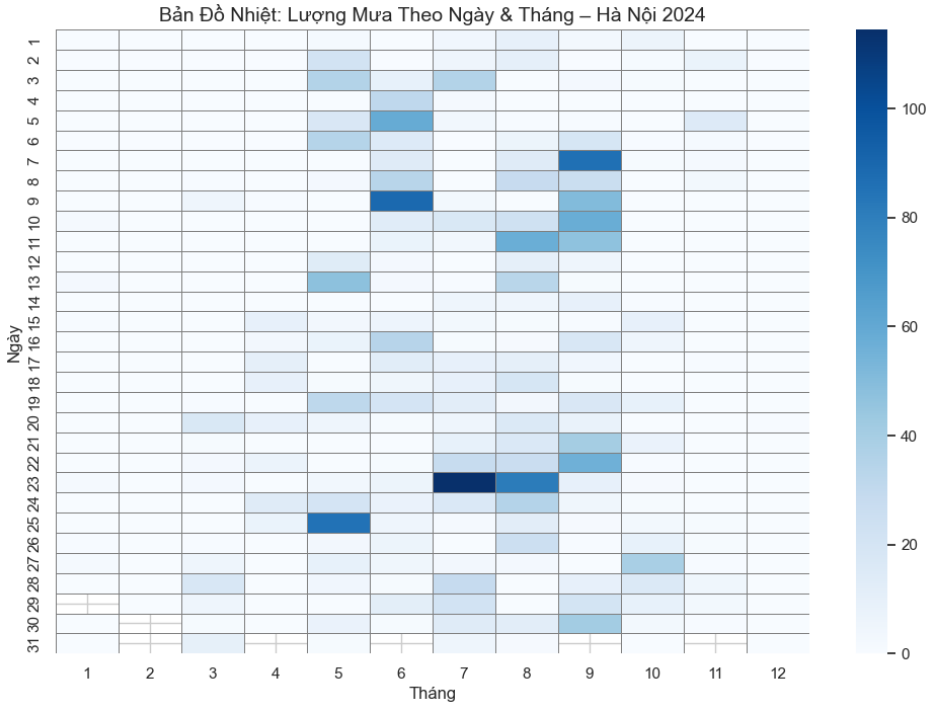
Hình 3.10: Biểu đồ đường xu hướng mức độ mây che phủ

Nhận xét

Mức độ mây che phủ 3.10 duy trì ở mức cao (>70%) trong phần lớn thời gian trong năm, nhưng có nhiều dao động mạnh, có thể phản ánh ảnh hưởng của các hệ thống mưa, áp thấp hoặc gió mùa làm tăng độ mây.

3.2.2 Biểu đồ nhiệt

Biểu đồ nhiệt (Heatmap) là một loại biểu đồ trực quan hóa dữ liệu mà trong đó cường độ của một giá trị được thể hiện bằng màu sắc. Dưới đây là biểu đồ nhiệt về lượng mưa trong năm 2024 với dữ liệu theo ngày và tháng.



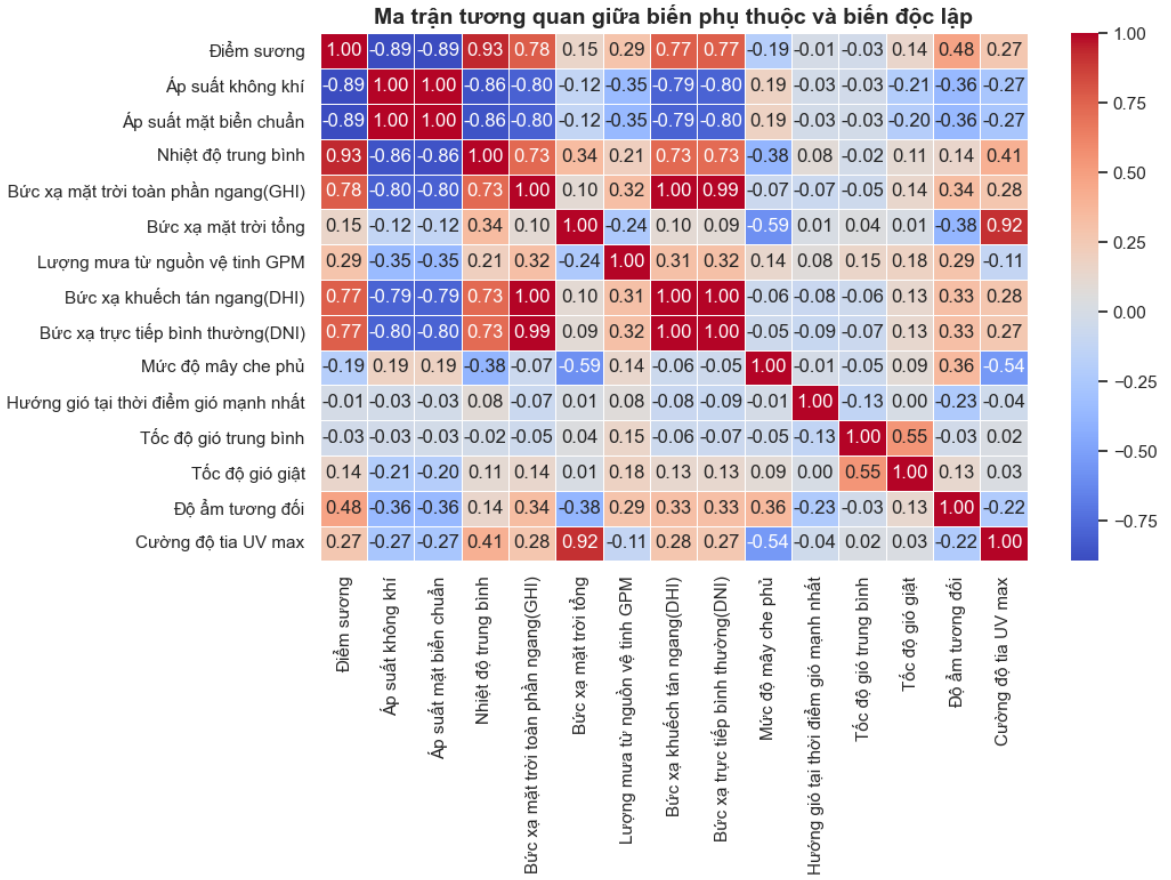
Hình 3.11: Biểu đồ nhiệt lượng mưa trong năm 2024

Biểu đồ nhiệt về lượng mưa theo ngày và tháng 3.11 tại Hà Nội năm 2024 trực quan hóa rõ ràng sự phân bố không đồng đều của lượng mưa trong năm. Mùa khô kéo dài từ Quý 1 đến Quý 4 (đầu và cuối năm) với lượng mưa rất thấp, trong khi mùa mưa tập trung mạnh mẽ vào Quý 3 (đặc biệt là tháng 7, 8, 9) với nhiều ngày ghi nhận lượng mưa lớn (biểu thị bằng màu xanh đậm). Biểu đồ cung cấp cái nhìn tổng quan hiệu quả về chu kỳ mưa theo mùa và các đợt mưa lớn trong năm, rất hữu ích cho việc phân tích khí hậu và quản lý tài nguyên nước.

3.2.3 Ma trận tương quan

Trong bất kỳ bài toán dự báo nào dựa trên dữ liệu thực tế, đặc biệt là những bài toán có độ phức tạp cao như dự báo thời tiết, nơi mà các yếu tố khí tượng luôn đan xen và tương tác một cách phi tuyến và khó kiểm soát, thì việc phân tích mối quan hệ giữa các biến đầu vào là bước không thể thiếu, đóng vai trò như nền tảng để hiểu rõ cấu trúc của dữ liệu, phát hiện ra các mối quan hệ ẩn giữa các yếu tố trong tự nhiên, đồng thời giúp định hướng thiết kế cấu trúc mô hình dự báo một cách tối ưu và hiệu quả.

Ma trận tương quan là công cụ hiệu quả để đánh giá mức độ tuyến tính giữa các biến.



Hình 3.12: Ma trận tương quan

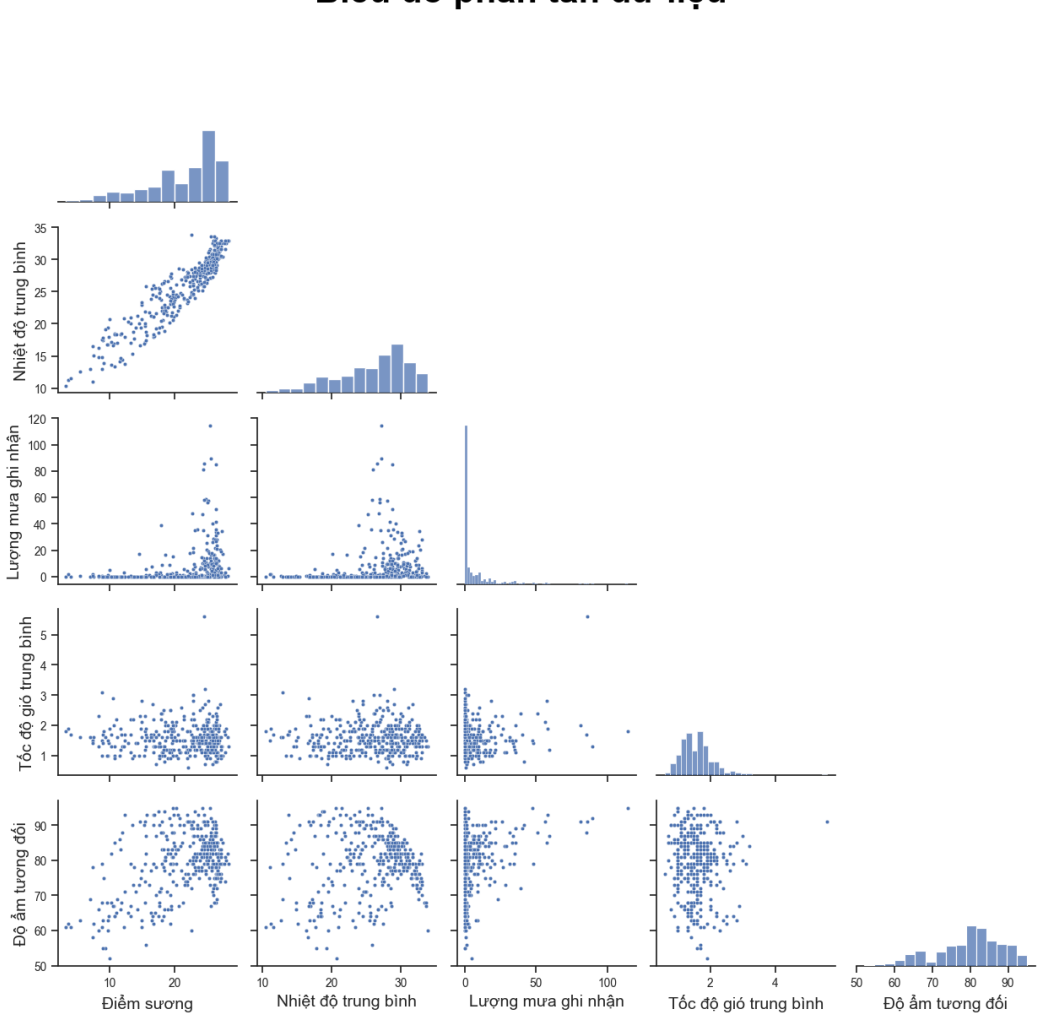
Thông qua việc phân tích ma trận tương quan giữa các biến đặc trưng trong 3.12, chúng tôi rút ra một số nhận định quan trọng cho quá trình chọn đặc trưng đầu vào khi xây dựng mô hình dự báo. Nhiệt độ trung bình thể hiện mối tương quan thuận rất mạnh với điểm sương (0.93), đồng thời có mối liên hệ nghịch mạnh với áp suất không khí và áp suất mặt biển (cùng mức -0.86), cho thấy vai trò chi phối của các yếu tố này đến trạng thái nhiệt trong khí quyển. Bức xạ mặt trời (GHI, DHI, DNI) cũng có liên hệ chặt với nhiệt độ và điểm sương, đặc biệt GHI có hệ số tương quan cao đến 0.91 với DNI và 0.99 với DHI, phản ánh tính đồng biến giữa các đại lượng bức xạ. Một số biến ít tương quan hơn với các biến còn lại như "Hướng gió tại thời điểm gió mạnh nhất", "tốc độ gió giật", hay "mức độ mây che phủ", đóng vai trò bổ trợ cho các đặc trưng chính.

Những phát hiện trên là cơ sở giúp nhóm lựa chọn biến đầu vào, loại bỏ dữ liệu dư thừa, từ đó tăng tính chính xác của các mô hình học máy trong giai đoạn huấn luyện.

3.2.4 Biểu đồ phân tán

Trước khi tiến hành huấn luyện các mô hình học máy nhằm dự báo các chỉ số thời tiết tại quan trọng tại Hà Nội, một bước trực quan hóa có tính chất then chốt cần được thực hiện là phân tích mối quan hệ giữa các biến thông qua biểu đồ phân tán. Phương pháp này không chỉ cho phép chúng tôi quan sát nhanh các xu hướng tuyến tính hay phi tuyến giữa biến độc lập và biến phụ thuộc, mà còn hỗ trợ việc kiểm tra giả định nền tảng về mối tương quan có ý nghĩa — một yếu tố ảnh hưởng trực tiếp đến độ chính xác và tính ổn định của các mô hình sau này.

Ngoài ra, việc nhận diện các mối quan hệ bất thường, phân bố chéo hoặc chồng lấp giữa các biến còn giúp làm rõ khả năng xảy ra hiện tượng đa cộng tuyến hoặc dữ liệu ngoại lai.



Hình 3.13: Biểu đồ phân tán tổng thể các biến khí tượng tại Hà Nội

Biểu đồ phân tán dữ liệu cho thấy mối quan hệ nổi bật giữa các yếu tố. Một trong những điểm nổi bật nhất là mối tương quan rất mạnh giữa "Điểm sương" và "Nhiệt độ trung bình". Các điểm dữ liệu trong biểu đồ phân tán tạo thành một đường chéo rõ ràng phản ánh mối quan hệ gần như đồng biến. Điều này có thể hiểu là khi nhiệt độ càng cao thì không khí càng giữ được nhiều hơi nước, làm điểm sương tăng lên. Bên cạnh đó, những ngày có điểm sương thấp thường trùng với các ngày có nhiệt độ thấp – phản ánh thời điểm mùa đông, không khí khô. Còn các cụm điểm ở vùng nhiệt độ cao và điểm sương cao cho thấy giai đoạn mùa hè ẩm nóng đặc trưng.

Biến "Lượng mưa ghi nhận" cho thấy một đặc điểm cực kỳ quan trọng của dữ liệu đó là sự phân mảnh rõ nét. Khoảng 85–90% dữ liệu rơi vào vùng mưa thấp dưới 10 mm, trong khi phần còn lại ở các mức cao đột biến, có thể lên tới trên 120 mm/ngày. Điều này khẳng định hiện tượng mưa lớn xảy ra không thường xuyên nhưng rất dữ dội, phản ánh tính khí hậu nhiệt đới gió mùa, nơi mà lượng mưa thường dồn vào một vài đợt mưa lớn kéo dài 1–2 ngày.

Ở khía cạnh gió, "Tốc độ gió trung bình" lại cho thấy phân bố tương đối chênh lệch khi hầu hết các biến giá trị gió thấp và trung nình chênh lệch với các giá trị gió cao. Điều này có thể được giải thích là "Tốc độ gió trung bình" có thể chịu ảnh hưởng bởi các yếu tố không có trong tập dữ liệu như vị trí địa lý. Tuy nhiên, sự hiện diện của các giá trị gió giật cao bất thường lên tới 7–8 m/s, có nguy cơ gây ra bão.

Một điểm thú vị khác là "Độ ẩm tương đối", chúng tôi phát hiện ra một mối tương quan âm nhẹ, độ ẩm cao thường xuất hiện ở tốc độ gió thấp và ngược lại. Điều này có thể lý giải rằng gió mạnh có xu hướng làm không khí "khô" hơn, là dấu hiệu của các đợt gió mùa – khi độ ẩm thấp và gió mạnh cùng lúc xuất hiện. Mối quan hệ này là một chỉ báo tiềm năng để phát hiện sớm các đợt không khí lạnh tràn về. Biến "Độ ẩm tương đối" có phân bố rất đều, tập trung chủ yếu trong vùng từ 65% đến 90%.

Đáng chú ý, sự phân tán của các biến còn lại như "Nhiệt độ trung bình" hay "Điểm sương" cho thấy một số vùng hình elip nghiêng, dấu hiệu của mối quan hệ đồng thời phi tuyến. Điều này gợi ý rằng nếu đưa "Độ ẩm tương đối" vào mô hình học máy, nên cân nhắc các thuật toán phi tuyến như Random Forest hoặc XGBoost thay vì hồi quy tuyến tính đơn thuần.

Chương 4

Các mô hình phân tích dữ liệu

4.1 Bài toán đặt ra

4.1.1 Bài toán số 1

Bài toán thứ nhất: Với tình hình khí hậu ngày càng biến đổi gay gắt, thảm họa thiên thai đang ngày càng phổ biến, thì việc biết được kiểu thời tiết dựa trên các đặc điểm, yếu tố môi trường xung quanh và đặc điểm của các yếu tố thời tiết trong lịch sử là rất quan trọng để dự đoán được kiểu thời tiết trong tương lai, vì vậy bài toán phân cụm dữ liệu thời tiết dựa trên các yếu tố môi trường được ra đời nhằm giải quyết vấn đề bài toán và mở ra hướng phát triển.

4.1.2 Bài toán số 2

Bài toán thứ hai: Trong bối cảnh biến đổi khí hậu toàn cầu và sự gia tăng các hiện tượng thời tiết cực đoan, nhu cầu dự báo thời tiết chính xác, ngắn hạn và cục bộ đang trở thành một yêu cầu cấp thiết phục vụ công tác quản lý đô thị, nông nghiệp, y tế cộng đồng và ứng phó thiên tai tại các đô thị lớn như Hà Nội. Mục tiêu cụ thể của bài toán là ứng dụng các mô hình học máy nhằm dự báo các thông số thời tiết quan trọng tại Hà Nội trong 7 ngày tiếp theo.

4.2 Mô hình K-Means phân loại thời tiết

Với lượng lớn thông tin của môi trường của các yếu tố như nhiệt độ, độ ẩm, lượng mưa, bức xạ, thì việc xác định kiểu thời tiết trở nên khó khăn khi nhìn vào các chỉ số về môi trường. Vì vậy, mô hình phân cụm thời tiết dựa trên các yếu tố môi trường được ra đời nhằm mục đích phục vụ cho các nghiên cứu, phân tích về thời tiết

4.2.1 Giới thiệu thuật toán K-Means

K-means là thuật toán phân cụm dựa trên trọng tâm, trong đó chúng ta tính toán khoảng cách giữa mỗi điểm dữ liệu và một trọng tâm để gán nó vào một cụm. Mục tiêu là xác định số lượng K nhóm trong tập dữ liệu.

Đây là một quá trình lặp đi lặp lại khi gán từng điểm dữ liệu vào các nhóm và các điểm dữ liệu dần dần được nhóm lại dựa trên các đặc điểm tương tự. Mục tiêu là giảm thiểu tổng khoảng cách giữa các điểm dữ liệu và tâm cụm, để xác định đúng nhóm mà mỗi điểm dữ liệu nên thuộc về.

Cụ thể, các bước mà thuật toán K-Means thực hiện bao gồm:

- Chọn số cụm K cần phân chia .
- Khởi tạo ngẫu nhiên K tâm cụm (centroids).
- Gán mỗi điểm dữ liệu vào cụm có tâm gần nhất (theo khoảng cách Euclidean).
- Tính lại tâm cụm mới bằng trung bình tất cả điểm trong cụm đó.
- Lặp lại bước 3 và 4 cho đến khi không còn thay đổi về cụm hoặc đạt số vòng lặp tối đa.

Khoảng cách Euclidean (Euclidean distance) là khoảng cách dùng để đo lường độ dài của đoạn thẳng nối hai điểm trong không gian. Nó còn được gọi là khoảng cách Pythagoras. Khoảng cách Euclidean là độ dài đường thẳng ngắn nhất giữa hai điểm trong không gian Euclid hoặc không gian nhiều chiều. Nó được tính bằng cách lấy căn bậc hai của tổng bình phương hiệu các tọa độ của hai điểm đó.

$$d(x, \mu) = \sqrt{\sum_{i=1}^{n} (x_i - \mu_i)^2} \tag{4.1}$$

Trong đó:

- x là điểm dữ liệu;
- μ là tâm cụm (centroid).

4.2.2 Ứng dụng thuật toán K-Means

Trong thuật toán phân cụm K-Means áp dụng cho các yếu tố thời tiết, chúng tôi sẽ sử dụng các biến đặc trưng, có thể làm ảnh hưởng đến mô hình phân cụm thời tiết.

Bước 1: Đầu tiên chúng tôi sẽ nhập các biến đặc trưng dùng để phân cụm vào và sau đó loại bỏ đi các giá trị trống

```
f_name = ['Muc do may che phu','Diem suong', 'Buc xa mat troi tong',
'Nhiet do cao nhat', 'Nhiet do thap nhat','Nhiet do trung binh',
'Cuong do tia UV max', 'toc do gio lon nhat', 'Toc do gio giat',
'Toc do gio trung binh','Luong mua ghi nhan',
'Luong mua tu nguon ve tinh GPM', 'Ap suat khong khi',
'Do am tuong doi']
# trich du lieu tu cac cot va loai cac dong thieu
f_data = data[f_name].dropna()
```

Đoạn mã 4.1: Nhập trường dữ liệu

Bước 2: Sau khi đã nhập các trường dữ liệu cần thiết cho việc phân cụm dữ liệu thời tiết, bước tiếp theo là thực hiện chuẩn hóa các cột dữ liệu nhằm đưa các chỉ số về dạng trung bình = 0 và phương sai = 1 nhằm tránh có chỉ số quá lớn ảnh hưởng đến mô hình.

```
Scaled = StandardScaler().fit_transform(f_data)
```

Đoạn mã 4.2: Chuẩn hóa dữ liệu

Bước 3: Sau khi đã chuẩn hóa, việc tiếp theo cần phải làm là giảm chiều dữ liệu, sở dĩ phải giảm chiều dữ liệu là vì với 14 chiều, gần như không thể trực quan hóa dữ liệu để hiểu được cấu trúc, mối quan hệ giữa các điểm. ngoài ra, giảm chiều còn làm giảm phức tạp tính toán giúp thuật toán tập trung vào việc tìm ra các thành phần chính nắm giữ phần lớn phương sai trong dữ liệu.

Nhóm chúng tôi sẽ thực hiện giảm chiều từ 14 chiều dữ liệu xuống còn 3 chiều dữ liệu tương ứng với 3 nhãn dữ liệu mới nhằm giảm vừa đủ và vừa giữ lại phần lớn phương sai trong dữ liệu.

```
pca = PCA(n_components=3)
pca_data = pca.fit_transform(Scaled)
dataframe_pca = pd.DataFrame(pca_data, columns=['Nhan 1','Nhan 2',
```

```
'Nhan 3'], index=f_data.index)
```

Đoạn mã 4.3: Giảm chiều dữ liệu

Bước 4: Tiến hành phân cụm dữ liệu sử dụng Kmeans từ thư viện sklearn với phân cụm từng nhãn lần lượt là 0:"Nắng nóng khô", 1: "Mưa ẩm", 2: "Trời mát"

```
kmeans = KMeans(n_clusters=3, random_state=42, n_init=20)
clusters_pca = kmeans.fit_predict(pca_data)
dataframe_pca['cluster'] = clusters_pca
clusters_labels = { 0: 'Nang nong kho',
                    1: 'Mua am',
                    2: 'Troi mat'}
dataframe_pca['label'] = dataframe_pca['cluster'].map(clusters_labels)
```

Đoạn mã 4.4: Phân cụm dữ liệu

Sau khi dữ liệu được phân thành 3 cụm nắng nóng khô, mưa ẩm, trời mát thì dưới đây chúng tôi sẽ cho xuất ra bảng tổng các giá trị của từng cụm.

	Nhãn	Biến đếm
0	Nắng nóng khô	130
1	Mưa ẩm	123
2	Trời mát	112

Bảng 4.1: Biến đếm số lượng mỗi cụm

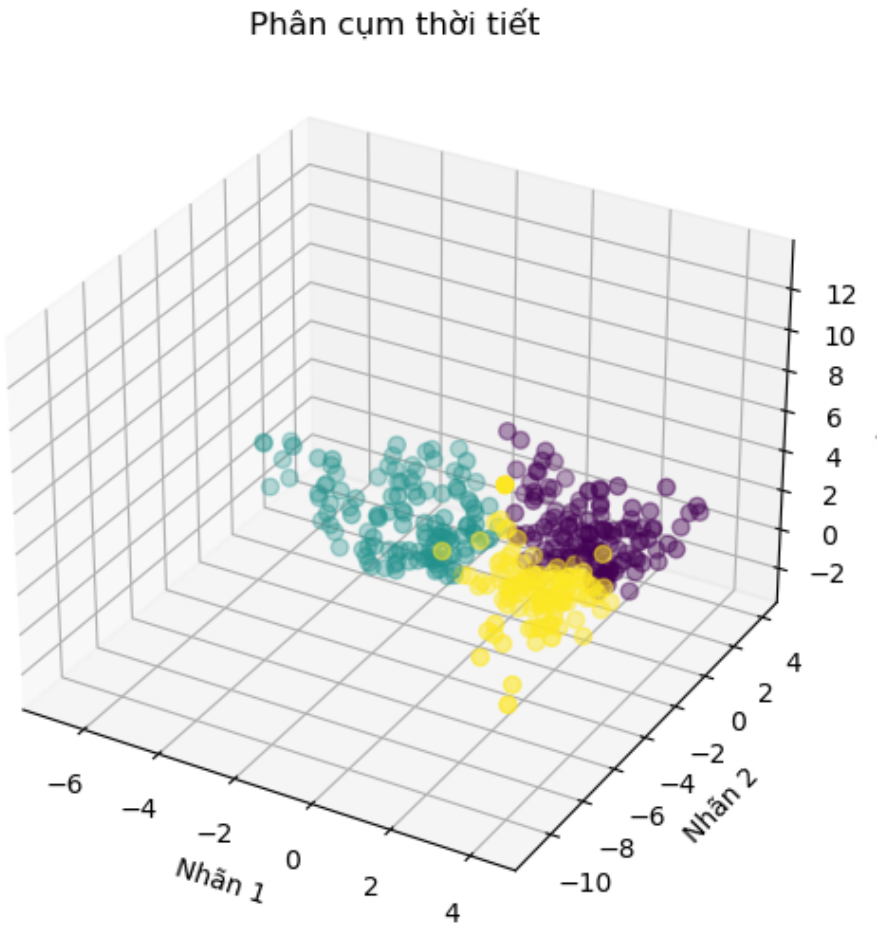
Sau khi đã phân cụm dữ liệu kiểu thời tiết, chúng tôi tiến hành vẽ biểu đồ nhằm mục đích trực quan hóa dữ liệu của mô hình sau khi phân cụm, cho người đọc cái nhìn dễ hình dung về các cụm dữ liệu thời tiết.

Dưới đây là đoạn mã sử dụng để vẽ biểu đồ phân cụm thời tiết sử dụng biểu đồ 3D

```
ax = fig.add_subplot(111, projection='3d')
ax.scatter(
    dataframe_pca['Nhan 1'],
    dataframe_pca['Nhan 2'],
    dataframe_pca['Nhan 3'],
    c=dataframe_pca['cluster'],  cmap='viridis', s=40)
```

```
ax.set_xlabel('Nhan 1')
ax.set_ylabel('Nhan 2')
ax.set_zlabel('Nhan 3')
ax.set_title('Phan cum thoi tiet')
plt.show()
```

Đoạn mã 4.5: Mã trực quan hóa phân cụm 3D



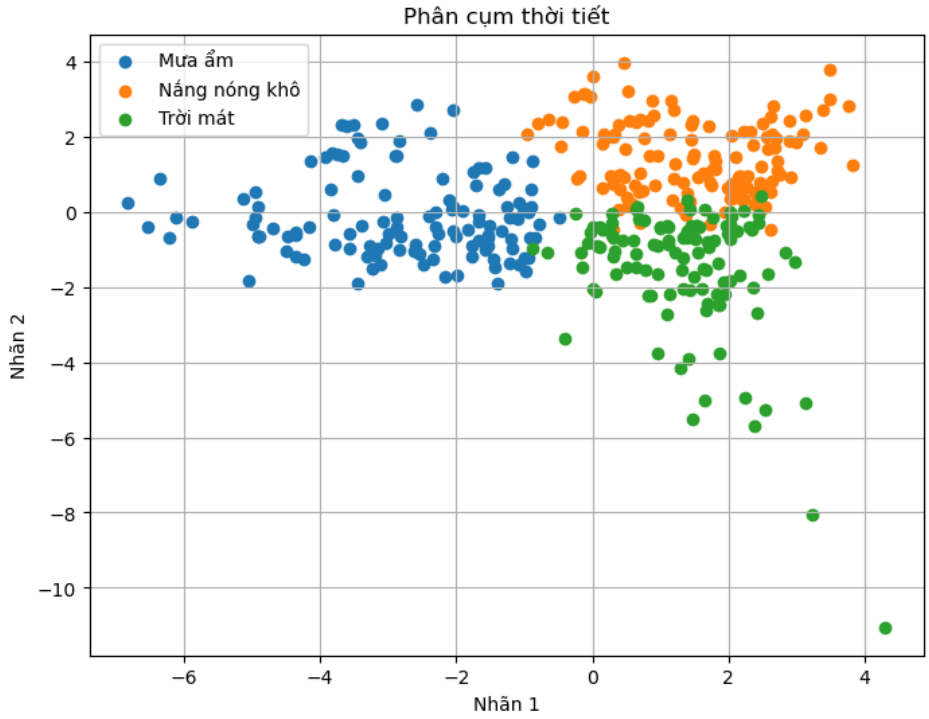
Hình 4.1: Biểu đồ 3D phân cụm dữ liệu thời tiết

Dưới đây là đoạn mã sử dụng để vẽ biểu đồ phân cụm thời tiết sử dụng biểu đồ 2D

```
plt.figure(figsize=(8,6))
for label in dataframe_pca['label'].unique():
    subset = dataframe_pca[dataframe_pca['label'] == label]
    plt.scatter(subset['Nhan 1'], subset['Nhan 2'], label=label)
plt.xlabel('Nhan 1')
```

```
plt.ylabel('Nhan 2')
plt.title('Phan cum thoi tiet')
plt.show()
```

Đoạn mã 4.6: Kỹ thuật phân cụm 2D



Hình 4.2: Biểu đồ 2D phân cụm dữ liệu thời tiết

Sau khi trực quan hóa thuật toán, ta có thể thấy dữ liệu phân thành 3 cụm riêng biệt.

- Cụm "Mưa ẩm"(xanh dương) và cụm "Nắng nóng khô" (cam) được phân tách khá tốt. Hai cụm này nằm ở hai vùng không gian khác nhau, chỉ số ít sự chồng lấn ở vùng giữa.
- Cụm "Trời mát"(xanh lá cây) có sự chồng lấn đáng kể với cụm "Nắng nóng khô"(cam), đặc biệt là ở phần trung tâm. Điều này cho thấy ranh giới giữa "Trời mát" và "Nắng nóng khô" không thực sự rõ ràng bằng cụm "Mưa ẩm" (xanh dương), hoặc có thể có một số điểm dữ liệu có đặc trưng lai giữa hai loại thời tiết này.
- Cụm "Trời mát" cũng có một chút chồng lấn với cụm "Mưa ẩm" ở phía bên trái, nhưng ít hơn so với "Nắng nóng khô".

4.2.3 Tiểu kết

Bài toán đầu tiên đặt ra là phân cụm dữ liệu kiểu thời tiết nhằm phân loại, gán nhãn kiểu thời tiết dựa trên các yếu tố môi trường như nhiệt độ, độ ẩm, lượng mưa,..., đã hoàn tất khi với mỗi chỉ số thu thập được từ môi trường của ngày hôm đó, tổng hợp lại đưa vào mô hình sẽ gán nhãn được kiểu thời tiết là gì và sẽ sử dụng các nhãn của mô hình để thực hiện tiếp các mô hình phân tích dự đoán tiếp theo

4.3 Mô hình hồi quy tuyến tính

4.3.1 Multivariate Linear Regression

Multivariate Linear Regression là một phần mở rộng của mô hình hồi quy tuyến tính bội , trong đó nhiều biến phụ thuộc, tức là nhiều đầu ra được mô hình hóa đồng thời như là các hàm tuyến tính của cùng một tập hợp hoặc các tập hợp riêng biệt của nhiều biến độc lập. Mô hình này cho phép khai thác cấu trúc liên kết giữa các biến phụ thuộc, giúp cải thiện độ chính xác dự đoán so với việc xây dựng các mô hình đơn lẻ một cách riêng biệt.

Công thức toán học tổng quát của mô hình có dạng:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}. \tag{4.2}$$

Trong đó,

- $\mathbf{Y}$: ma trận biến phụ thuộc;
- $\mathbf{X}$: ma trận các biến độc lập;
- $\mathbf{B}$: ma trận hệ số hồi quy;
- $\mathbf{E}$: ma trận sai số ngẫu nhiên.

4.3.2 Chuẩn bị dữ liệu đầu vào

Căn cứ vào ma trận tương quan về các mối liên hệ giữa các biến trong tập dữ liệu, chúng tôi trích xuất ra biến đặc trưng và biến mục tiêu nhằm phục vụ cho việc huấn luyện mô hình.

Biến đặc trưng features = [“Điểm sương”, “Áp suất không khí”, “Áp suất mặt biển chuẩn”, “Nhiệt độ trung bình”, “Bức xạ mặt trời toàn phần ngang(GHI)”, “Bức xạ mặt trời tổng”, “Lượng mưa từ nguồn vệ tinh GPM”, “Bức xạ khuếch tán ngang(DHI)”, “Bức xạ trực tiếp bình thường(DNI)”, “Mức độ mây che phủ”, “Hướng gió tại thời điểm gió mạnh nhất”, “Tốc độ gió trung bình”, “Tốc độ gió giật”, “Độ ẩm tương đối”, “Cường độ tia UV max”].

Biến mục tiêu targets = [“Nhiệt độ hiện tại”, “Lượng mưa ghi nhận”, “Độ ẩm tương đối”, “Hướng gió hiện tại”, “Tốc độ gió giật”, “Cường độ tia UV max”].

4.3.3 Chuẩn hóa dữ liệu

Chuẩn hóa dữ liệu là một bước tiền xử lý dữ liệu cực kỳ quan trọng trong quy trình xây dựng mô hình học máy, với mục tiêu đưa các đặc trưng về cùng một thang đo.

```
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
```

Đoạn mã 4.7: Kỹ thuật chuẩn hóa dữ liệu

4.3.4 Huấn luyện mô hình

Sau khi thực hiện quá trình chuẩn bị dữ liệu, chúng tôi đến với phần huấn luyện mô hình.

Chúng tôi chia tập dữ liệu ban đầu thành tập train/test với tỷ lệ 80/20.

```
X_train, X_test, y_train, y_test=train_test_split(X_scaled, y_target,
test_size=0.2, shuffle=False)
model_dict_MLR = {}
for col in targets:
    y_target = y[col]
    model = LinearRegression()
    model.fit(X_train, y_train)
    model_dict_MLR[col] = model
```

Đoạn mã 4.8: Mô hình Multivariate Linear Regression

4.3.5 Kết quả dự báo của mô hình

Sau khi hoàn tất quá trình huấn luyện, căn cứ vào sự phân bổ tổng thể của dữ liệu, mô hình đưa ra dự đoán như *Hình 4.3*.

Dự báo thời tiết ngày	Nhiệt độ (°C)	Lượng mưa (mm)	Độ ẩm (%)	Hướng gió (°)	Tốc độ gió (m/s)	Cường độ tia UV (μW/cm²)
31/12/2024	26.2	3.6	80.2	170.7	7.2	4.4
01/01/2025	26.2	3.6	80.2	170.7	7.2	4.4
02/01/2025	26.1	3.6	80.2	170.7	7.2	4.4
03/01/2025	26.1	3.6	80.2	170.7	7.2	4.4
04/01/2025	26.1	3.6	80.2	170.7	7.2	4.4
05/01/2025	26.1	3.6	80.2	170.7	7.2	4.4
06/01/2025	26.1	3.6	80.2	170.7	7.2	4.4

Hình 4.3: Kết quả dự báo của mô hình Multivariate Linear Regression

4.3.6 Tiểu kết

Như kết quả thể hiện trong *Hình 4.3*, mô hình Multivariate Linear Regression đã đưa ra dự báo thời tiết tại Hà Nội trong 7 ngày tiếp theo, tính từ ngày 31/12/2024 đến ngày 06/01/2025. Qua đó, phản ánh các thông số dự báo có xu hướng ổn định, phản ánh tính chất đều đặn của điều kiện khí hậu cận nhiệt đới ẩm trong khoảng thời gian ngắn hạn này. mức nhiệt độ dao động không đáng kể quanh mức 26.1 – 26.2 °C, độ ẩm ổn định ở mức 80.2%, và lượng mưa duy trì ở mức 3.6 mm/ngày, cho thấy rằng mô hình đã thành công trong việc nắm bắt được cấu trúc tuyến tính và ổn định của thời tiết.

4.4 Mô hình Random Forest Regression

Mô hình Random Forest Regression đóng vai trò về mặt hiệu suất mô hình, khả năng khái quát hoá, và diễn giải mối quan hệ phi tuyến giữa các biến trong tập dữ liệu.

Chúng tôi sử dụng chung dữ liệu đầu vào với mô hình Multivariate Linear Regression nhằm huấn luyện và đưa ra dự báo.

4.4.1 Huấn luyện mô hình

Sau khi thực hiện quá trình chuẩn bị dữ liệu, tiền xử lý dữ liệu, chuẩn hóa dữ liệu, chúng tôi đến với phần huấn luyện mô hình.

Chúng tôi chia tập dữ liệu ban đầu thành tập train/test với tỷ lệ 80/20.

```
X_train, X_test, y_train, y_test=train_test_split(X_scaled, y_target,
test_size=0.2, shuffle=False)
model_dict_RF = {}
for col in targets:
    y_target = y[col]
    test_size=0.2, shuffle=False)
    model = RandomForestRegressor(n_estimators=100, random_state=42)
    model.fit(X_train, y_train)
    model_dict_RF[col] = model
```

Đoạn mã 4.9: Mô hình Random Forest Regression

4.4.2 Kết quả dự báo của mô hình

Sau khi hoàn tất quá trình huấn luyện, căn cứ vào sự phân bổ tổng thể của dữ liệu, mô hình đưa ra dự đoán như *Hình 4.4*.

Dự báo thời tiết ngày	Nhiệt độ (°C)	Lượng mưa (mm)	Độ ẩm (%)	Hướng gió (°)	Tốc độ gió (m/s)	Cường độ tia UV (μW/cm²)
31/12/2024	26.2	3.7	80.0	169.9	7.2	4.4
01/01/2025	25.8	3.7	80.0	169.8	7.2	4.4
02/01/2025	25.6	3.7	80.0	169.8	7.2	4.4
03/01/2025	25.4	3.7	80.0	169.8	7.2	4.4
04/01/2025	25.4	3.7	80.0	169.8	7.2	4.4
05/01/2025	25.4	3.7	80.0	169.8	7.2	4.4
06/01/2025	25.4	3.7	80.0	169.8	7.2	4.4

Hình 4.4: Kết quả dự báo của mô hình Random Forest Regression

4.4.3 Tiểu kết

Như kết quả thể hiện trong *Hình 4.4*, mô hình Random Forest Regression đã đưa ra dự báo thời tiết tại Hà Nội trong 7 ngày tiếp theo, tính từ ngày 31/12/2024 đến ngày 06/01/2025.

Dữ liệu đầu ra cho thấy mô hình đã dự báo hợp lý xu hướng giảm nhiệt độ nhẹ và ổn định theo thời gian, với giá trị nhiệt độ giảm từ 26.2°C xuống còn 25.4°C trong vòng một tuần. Mức độ biến thiên này phản ánh khả năng nắm bắt các chuyển động vi mô trong chuỗi dữ liệu thời tiết, điều mà các mô hình tuyến tính trước đó có thể bỏ qua.

4.5 Mô hình XGBoost

Trong bước tiếp theo của đồ án, mô hình XGBoost được chúng tôi lựa chọn nhằm khai thác tối ưu tiềm năng dự báo ngắn hạn từ dữ liệu thời tiết của Hà Nội. Đây là một trong những thuật toán boosting hiệu quả nhất hiện nay, nổi bật với khả năng xử lý dữ liệu phi tuyến, tối ưu hóa gradient theo từng bước lặp và kiểm soát quá khớp một cách linh hoạt thông qua các siêu tham số.

So với các biến thể boosting truyền thống, XGBoost bổ sung thêm chính quy hóa hàm mục tiêu nhằm kiểm soát độ phức tạp mô hình, từ đó giảm quá khớp và tăng khả năng tổng quát hóa, giúp XGBoost trở thành một trong những thuật toán phổ biến nhất trong các ứng dụng học máy thực tiễn và nghiên cứu học thuật.

4.5.1 Huấn luyện mô hình

Chúng tôi sử dụng chung dữ liệu đầu vào với mô hình Multivariate Linear Regression và Random Forest Regression nhằm huấn luyện và đưa ra dự báo.

XGBoost được thiết kế dựa trên khái niệm boosting theo gradient, trong đó nhiều mô hình cây quyết định yếu được huấn luyện nối tiếp nhau, mỗi mô hình học từ sai số còn lại của mô hình trước đó. Cơ chế này cho phép mô hình cải thiện dần độ chính xác của dự báo, đặc biệt hiệu quả trong các bài toán có cấu trúc dữ liệu phức tạp như dự báo thời tiết, nơi các mối quan hệ giữa các thuộc tính thường phi tuyến tính và có nhiều tương tác chéo lẫn nhau.

```
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y_target,
test_size=0.2, shuffle=False)
model_dict_XGB = {}
for col in targets:
    y_target = y[col]
    model = XGBRegressor(
        n_estimators=100, learning_rate=0.05,
        max_depth=5, subsample=0.8,
        colsample_bytree=0.8,random_state=42, verbosity=0)
    model.fit(X_train, y_train)
    model_dict_XGB[col] = model
```

Đoạn mã 4.10: Mô hình XGBoost

4.5.2 Kết quả dự báo của mô hình

Sau khi hoàn tất quá trình huấn luyện, căn cứ vào việc điều chỉnh, cập nhật các tham số một cách phù hợp, mô hình đưa ra dự đoán như *Hình 4.5*.

Dự báo thời tiết ngày	Nhiệt độ (°C)	Lượng mưa (mm)	Độ ẩm (%)	Hướng gió (°)	Tốc độ gió (m/s)	Cường độ tia UV (μW/cm²)
31/12/2024	26.4	3.1	80.4	176.5	7.1	4.4
01/01/2025	26.3	2.9	80.4	181.9	7.0	4.4
02/01/2025	26.3	2.9	80.4	186.1	6.8	4.4
03/01/2025	26.3	2.9	80.4	193.1	6.3	4.4
04/01/2025	26.3	2.9	80.4	196.0	6.0	4.4
05/01/2025	26.3	2.9	80.4	196.7	5.6	4.4
06/01/2025	26.3	2.9	80.4	197.7	5.6	4.4

Hình 4.5: Kết quả dự báo thời tiết 7 ngày của mô hình XGBoost

4.5.3 Tiểu kết

Như kết quả thể hiện trong *Hình 4.5*, mô hình XGBoost thể hiện khả năng mô hình hóa tốt các mối quan hệ phi tuyến tính trong dữ liệu dữ liệu khí hậu, với độ ổn định cao và biến động nhỏ trong các chỉ số dự báo.

Mô hình đã đưa ra dự báo thời tiết tại Hà Nội trong 7 ngày tiếp theo, tính từ ngày 31/12/2024 đến ngày 06/01/2025. Cụ thể, nhiệt độ dao động nhẹ từ 26.3°C đến 26.4°C, lượng mưa giảm nhẹ từ 3.1 mm xuống còn 2.9 mm và độ ẩm duy trì ổn định ở mức 80.4%. Điều đáng chú ý là mô hình đã dự đoán xu hướng dịch chuyển rõ rệt về hướng gió tăng từ 176.5° lên đến 197.7° và tốc độ gió giảm dần qua từng ngày từ 7.1 m/s xuống còn 5.6 m/s. Những thay đổi này phản ánh được sự biến động khí quyển thực tế, điều mà các mô hình tuyến tính hoặc ensemble đơn giản khó có thể phát hiện.

XGBoost, với cơ chế học tăng cường boosting, đã cho thấy năng lực vượt trội trong việc nhận diện và khai thác các tín hiệu phức tạp ẩn trong dữ liệu đầu vào. Khả năng này giúp mô hình không chỉ đưa ra dự báo chính xác mà còn dự báo theo hướng linh hoạt, thích ứng với các chuỗi dữ liệu có tính dao động theo thời gian như dữ liệu thời tiết.

Phần III

KẾT LUẬN VÀ KIẾN NGHỊ

Chương 5

Kết Luận và hướng phát triển

5.1 Kết luận

5.1.1 Kết luận bài toán số 1

. Bài toán đặt ra là phân cụm dữ liệu thời tiết dựa trên các yếu tố môi trường nhằm dự báo thời tiết trong tương lai dựa vào nhãn dữ liệu từ quá khứ đã hoàn thành khi mô hình đã thực hiện phân cụm từ 366 dòng dữ liệu thành 3 cụm với nhãn lần lượt của 3 cụm là "Nắng nóng khô", "Mưa ẩm", "Trời mát", qua quá trình chuẩn hóa, giảm chiều dữ liệu sử dụng phương pháp PCA cho các biến đặc trưng từ 14 chiều dữ liệu xuống còn 3 chiều dữ liệu, thuật toán đã phân cụm dữ liệu thành công khi chỉ với các biến đầu vào đặc trưng như "Điểm sương", "Nhiệt độ trung bình, thấp nhất, cao nhất", "Độ ẩm tương đối", "Cường độ tia UV", "Tốc độ gió giật", "Tốc độ gió trung bình, lớn nhất", "Lượng mưa".

Dưới đây là hình ảnh minh họa kết quả sau khi thực hiện phân cụm thời tiết.

	Mức độ mây che phủ	Điểm sương	Bức xạ mặt trời tổng	Nhiệt độ cao nhất	Nhiệt độ thấp nhất	Nhiệt độ trung bình	Cường độ tia UV max	Tốc độ gió lớn nhất	Tốc độ gió giật	Tốc độ gió trung bình	Lượng mưa ghi nhận	Lượng mưa từ nguồn vệ tinh GPM	Áp suất không khí	Độ ẩm tương đối	label
100	97	22.1	71	25.0	21.9	23.45	3.6	2.0	6.4	1.2	0.3	0.3	1013	91	Mưa ẩm
101	84	23.5	101	29.7	23.5	26.60	4.0	2.4	8.8	1.6	0.0	0.0	1012	87	Trời mát
102	90	23.5	116	30.5	24.9	27.70	3.6	2.9	9.6	1.9	0.0	0.0	1010	82	Trời mát
103	96	23.9	121	30.3	25.8	28.05	3.6	3.9	10.0	2.5	0.0	0.0	1007	80	Trời mát
104	100	24.4	109	31.0	26.0	28.50	3.7	2.9	8.0	2.0	0.0	0.0	1007	82	Trời mát
105	91	24.5	98	29.7	26.1	27.90	3.7	3.0	10.0	1.9	9.5	9.5	1008	85	Trời mát
106	54	23.2	271	32.0	25.0	28.50	9.7	3.0	10.8	2.0	3.3	3.3	1007	77	Nắng nóng khô
107	85	24.9	141	30.3	26.3	28.30	4.5	2.8	8.0	1.7	10.3	10.3	1006	85	Trời mát
108	56	24.5	256	32.0	24.7	28.35	10.5	3.0	8.8	1.8	9.5	9.5	1004	84	Nắng nóng khô
109	69	25.3	244	33.6	26.1	29.85	9.1	3.0	8.0	2.4	0.0	0.0	1001	82	Nắng nóng khô

Hình 5.1: Minh họa kết quả phân cụm thời tiết

5.1.2 Kết luận bài toán số 2

Trong quá trình tìm ra lời giải cho bài toán số 2, chúng tôi đã tiến hành xây dựng và huấn luyện ba mô hình học máy khác nhau bao gồm: Multiple Linear Regression, Random Forest Regression và XGBoost. Mỗi mô hình được đào tạo để dự báo các biến mục tiêu thời tiết quan trọng tại Hà Nội lần lượt là "Nhiệt độ hiện tại", "Lượng mưa ghi nhận", "Độ ẩm tương đối", "Hướng gió hiện tại", "Tốc độ gió giật", "Cường độ tia UV max".

Để đảm bảo tính khách quan trong lựa chọn mô hình tối ưu, nhóm tiến hành đánh giá đồng thời hiệu suất của các mô hình trên cơ sở ba chỉ số thông dụng trong lĩnh vực học máy:

- MAE (Mean Absolute Error) $\rightarrow$ đo lường sai số tuyệt đối trung bình;
- RMSE (Root Mean Squared Error) $\rightarrow$ nhấn mạnh các sai số lớn;
- R^2 Score (Hệ số xác định) $\rightarrow$ phản ánh mức độ mô hình giải thích được phương sai của dữ liệu.

Công thức toán học của 3 phương pháp đánh giá các mô hình lần lượt như sau:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|. \tag{5.1}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}. \tag{5.2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}. \tag{5.3}$$

Trong đó,

- n: số lượng điểm dữ liệu;
- y_i: giá trị thực tế tại điểm thứ i;
- $\hat{y}_i$: giá trị dự báo tại điểm thứ i;
- $\bar{y}$: giá trị trung bình của tất cả giá trị thực tế.

Biến mục tiêu	Mô hình	MAE	RMSE	R^2 Score
Nhiệt độ hiện tại	Multiple Linear Regression	0.171	0.219	0.996
Lượng mưa ghi nhận	Multiple Linear Regression	0.000	0.000	1.000
Độ ẩm tương đối	Multiple Linear Regression	0.000	0.000	1.000
Hướng gió hiện tại	Multiple Linear Regression	0.000	0.000	1.000
Tốc độ gió giật	Multiple Linear Regression	0.000	0.000	1.000
Cường độ tia UV max	Multiple Linear Regression	0.000	0.000	1.000
Nhiệt độ hiện tại	Random Forest	0.327	0.431	0.986
Lượng mưa ghi nhận	Random Forest	0.058	0.266	0.998
Độ ẩm tương đối	Random Forest	0.586	1.286	0.974
Hướng gió hiện tại	Random Forest	1.445	2.593	0.999
Tốc độ gió giật	Random Forest	0.017	0.048	0.999
Cường độ tia UV max	Random Forest	0.044	0.064	0.999
Nhiệt độ hiện tại	XGBoost	0.412	0.520	0.980
Lượng mưa ghi nhận	XGBoost	0.475	1.051	0.963
Độ ẩm tương đối	XGBoost	1.871	2.488	0.902
Hướng gió hiện tại	XGBoost	7.737	10.018	0.979
Tốc độ gió giật	XGBoost	0.083	0.108	0.996
Cường độ tia UV max	XGBoost	0.133	0.171	0.992

Hình 5.2: Đánh giá mức độ hiệu quả của ba mô hình

Các kết quả đánh giá được tổng hợp trong 5.2, cho thấy sự khác biệt. Cụ thể:

- Mô hình Multiple Linear Regression đạt kết quả gần như tuyệt đối (≈ 1.000) trên mọi biến mục tiêu, đồng thời MAE và RMSE gần bằng 0. Tuy nhiên, điều này đặt ra nghi vấn về hiện tượng quá khớp (overfitting) hoặc chia tách dữ liệu chưa hợp lý.
- Random Forest Regression thể hiện hiệu năng ổn định, với R^2 từ khoảng 0.974 đến 0.999 và sai số thấp, đặc biệt trên các biến Tốc độ gió giật và Cường độ tia UV max.
- Trong khi đó, XGBoost Regression tuy có sai số lớn hơn một chút trên một số biến, như Độ ẩm tương đối ($R^2 \approx 0.902$) hay Hướng gió hiện tại (MAE ≈ 7.737), nhưng lại thể hiện khả năng dự báo tốt với các biến biến động mạnh, Tốc độ gió giật và Cường độ tia UV max. Điều này minh chứng cho tính hiệu quả của thuật toán tăng cường theo gradient khi xử lý các mối quan hệ phi tuyến phức tạp trong dữ liệu khí hậu.

Tổng hòa các yếu tố về độ chính xác, khả năng tổng quát hóa và tính ổn định, chúng tôi kết luận rằng mô hình Random Forest Regression là lựa chọn tối ưu nhất cho bộ dữ liệu thời tiết tại Hà Nội.

5.2 Kiến nghị hướng phát triển

5.2.1 Bài toán số 1

Đối với bài toán thứ nhất, tuy mô hình phân cụm của thuật toán K-Mean hoạt động rất tốt khi phân cụm dữ liệu thời tiết dựa trên các yếu tố đặc trưng môi trường thành 3 cụm, tuy nhiên việc dự đoán kiểu thời tiết trong tương lai lại phải thủ công tìm lại các đặc trưng thời tiết để biết kiểu dữ liệu đó là gì.

Vì vậy, chúng tôi đã đề ra hướng phát triển cho mô hình phân cụm đó là sử dụng tiếp các mô hình dự đoán, dự báo kiểu thời tiết trong tương lai dựa trên các đặc trưng yếu tố thời tiết, các nhãn đã được phân cụm và dữ liệu đã được chuẩn hóa và giảm chiều.

Mô hình dự đoán, dự báo kiểu thời tiết trong tương lai mà chúng tôi kiến nghị sử dụng bao gồm như sau:

- Decision Tree / Random Forest.
- KNN (K-Nearest Neighbors).
- SVM (Support Vector Machine).
- Gradient Boosting / XGBoost.

5.2.2 Bài toán số 2

Từ những kết quả đạt được, chúng tôi rút ra được là đồ án có thể tiếp tục được mở rộng theo các hướng sau:

- Mở rộng thời gian và không gian dự báo sử dụng các mô hình học sâu như LSTM, áp dụng mô hình cho toàn bộ tỉnh thành xây dựng hệ thống cảnh báo thời tiết theo thời gian thực.
- Triển khai hệ thống dự báo sử dụng API và tạo dashboard tích hợp vào bản đồ khí tượng số phục vụ cộng đồng, cơ quan khí tượng thủy văn quốc gia.

Đường dẫn mã nguồn:

Toàn bộ mã nguồn và biểu đồ chúng tôi xin được gói trong đường dẫnTại đây.

Tài liệu tham khảo

[1] E. Alpaydin, *Machine learning*. MIT press, 2021.

[2] J. Yadav and M. Sharma, "A review of k-mean algorithm," *Int. J. Eng. Trends Technol*, vol. 4, no. 7, pp. 2972–2976, 2013.

[3] X. Su, X. Yan, and C.-L. Tsai, "Linear Regression," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 4, no. 3, pp. 275–294, 2012.

[4] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[5] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. ACM, 2016, pp. 785–794.

[6] N. A. Thịnh, Đoàn Ngọc Đức, L. Đức Hải, N. B. Trung, and K. V. Quỳ, "Các yếu tố ảnh hưởng đến mức sẵn lòng chi trả của người dân để cải thiện chất lượng không khí tại quận Tây Hồ, thành phố Hà Nội," *Tạp chí Môi trường*, 2024.