



ĐẠI HỌC KHOA HỌC TỰ NHIÊN

Ứng dụng học máy dự báo mức độ ô nhiễm không khí tại các thành phố

Sinh viên :

Đỗ Quốc An - 22000067

Phạm Thị Duyên - 22000079

Trần Kiều Hạnh - 22000091

Giảng viên :

Cao Văn Chung

Ngày 20 tháng 4 năm 2025

Mục lục

Danh mục ký hiệu	5
Danh sách bảng	6
Danh sách hình vẽ	7
0 Mở đầu	9
0.1 Giới thiệu chung	9
0.2 Bài toán	9
0.3 Mục tiêu	10
0.4 Phạm vi nghiên cứu	10
0.5 Phương pháp nghiên cứu	10
1 Cơ sở lý thuyết	12
1.1 PCA (Principal Component Analysis)	12
1.1.1 Khái niệm	12
1.1.2 Mục tiêu	12
1.1.3 Nguyên lý hoạt động	13
1.1.4 Bài toán tối ưu	14
1.1.5 Ưu điểm và Nhược điểm	15
1.1.6 Một số ứng dụng	15
1.2 t-SNE (t-Distributed Stochastic Neighbor Embedding)	16
1.2.1 Khái niệm	16
1.2.2 Cấu trúc và nguyên lý hoạt động	16
1.2.3 Tham số quan trọng	17
1.3 MLP (Multi-layer Perceptron)	17
1.3.1 Khái niệm	17
1.3.2 Kiến trúc mô hình	17
1.3.3 Nguyên lý hoạt động	17
1.4 Random Forest	18
1.4.1 Khái niệm	18
1.4.2 Nguyên lý hoạt động	19
1.4.3 Thuật toán	19
1.4.4 Ưu điểm và Nhược điểm	20

1.4.5	Ứng dụng	20
1.5	K-means	21
1.6	GMM (Gaussian Mixture Model)	21
1.6.1	Khái niệm	21
1.6.2	Quy trình học GMM	22
1.6.3	Kiến thức lý thuyết liên quan đến GMM	22
1.6.4	Ứng dụng	23
2	Dữ liệu	24
2.1	Nguồn dữ liệu	24
2.2	Tiền xử lý dữ liệu	24
2.2.1	Đọc và khám phá dữ liệu ban đầu	24
2.2.2	Xử lý dữ liệu thiếu	25
2.2.3	Xử lý giá trị ngoại lai (Outliers)	25
2.2.4	Chuyển đổi dữ liệu	26
2.3	Phân tích thống kê mô tả dữ liệu	26
2.3.1	Tổng quan dữ liệu	26
2.3.2	Thống kê mô tả với df.describe()	26
3	Giảm chiều dữ liệu	27
3.1	PCA	27
3.1.1	Chuẩn hóa dữ liệu và phân tích thành phần chính	27
3.1.2	Trực quan hóa dữ liệu theo các cặp thành phần chính	30
3.1.3	Phân tích phương sai giải thích	31
3.1.4	Trực quan hóa mối quan hệ giữa đặc trưng và đầu ra	33
3.2	t-SNE	34
3.2.1	Phân tích thành phần chính	34
3.2.2	Trực quan hóa dữ liệu theo các thành phần chính	36
3.2.3	Trực quan hóa mối quan hệ giữa đặc trưng và đầu ra	37
3.3	So sánh PCA và t-SNE	38
4	Phân cụm dữ liệu	41
4.1	K-means	41
4.1.1	Phân cụm K-Means và lựa chọn số cụm K tối ưu	41
4.1.2	Đánh giá mối quan hệ giữa các mẫu dữ liệu đầu vào và đầu ra trong từng cụm	42
4.1.3	Trực quan hóa kết quả phân cụm	44
4.2	GMM (Gaussian Mixture Model)	46
4.2.1	Phân cụm GMM và lựa chọn số cụm phù hợp	46
4.2.2	Đánh giá mối quan hệ giữa các mẫu dữ liệu đầu vào và đầu ra trong từng cụm	48
4.2.3	Trực quan hóa kết quả phân cụm	50

4.3	So sánh hai phương pháp	52
4.3.1	So sánh chất lượng cụm theo độ đo định lượng	52
4.3.2	So sánh phân phối AQI trong các cụm	53
4.3.3	So sánh đặc trưng trong các cụm	53
4.3.4	So sánh độ ổn định AQI theo hệ số biến thiên (CV)	54
4.3.5	So sánh trực quan hóa cụm	54
4.3.6	Kết luận	54
5	Thực nghiệm và kết quả	55
5.1	Mô hình Random Forest	55
5.1.1	Thiết lập và cấu hình mô hình	55
5.1.2	Quá trình huấn luyện mô hình	56
5.1.3	Thực nghiệm mô hình Random Forest với dữ liệu khác nhau	57
5.1.4	Phân tích hiện tượng Overfitting	60
5.1.5	Trực quan hóa và đánh giá tương quan phần dư	64
5.2	Mô hình MLP (Multilayer Perceptron)	66
5.2.1	Thiết lập và cấu hình mô hình	66
5.2.2	Quá trình huấn luyện mô hình	66
5.2.3	Thực nghiệm mô hình MLP với dữ liệu khác nhau	67
5.2.4	Phân tích hiện tượng Overfitting	69
5.2.5	Trực quan hóa và đánh giá tương quan phần dư	73
5.3	So sánh hiệu suất giữa Random Forest và MLP	75
5.3.1	So sánh hiệu suất trên các chỉ số đánh giá	75
5.3.2	So sánh độ ổn định và phân tích phần dư	75
5.3.3	So sánh khả năng tổng quát hóa và hiện tượng Overfitting	76
5.3.4	Kết luận chung	76
6	Mô hình phân loại	77
6.1	Bài toán phân loại	77
6.1.1	Phân chia đầu ra thành 4 khoảng và xác định ngưỡng	77
6.1.2	Mô hình hóa bài toán phân loại AQI	77
6.2	Mô hình phân loại Naive Bayes	78
6.3	Mô hình phân loại Random Forest	81
6.4	So sánh Naive Bayes và Random Forest	83
6.4.1	Đánh giá Hiệu suất Tổng thể (Macro Metrics)	83
6.4.2	Ảnh hưởng của PCA (Principal Component Analysis)	83
6.4.3	Ảnh hưởng của Tỷ lệ Chia Train:Validation	84
6.4.4	Phân tích Chi tiết Kết quả Phân loại theo Nhãn	84
6.4.5	Giải thích Sự Khác biệt về Hiệu suất	85

6.4.6 Kết luận và Đề xuất	85
7 Kết luận	87
7.1 Tóm tắt kết quả	87
7.2 Đánh giá	88
7.3 Hướng phát triển	89

Danh mục ký hiệu

PCA	Phân tích thành phần chính.
t-SNE	T-Distributed Stochastic Neighbor Embedding, phương pháp giảm chiều dữ liệu và trực quan hóa dữ liệu.
MLP	Mạng Perceptron nhiều tầng.
GMM	Mô hình hỗn hợp Gaussian.
ReLU	Hàm kích hoạt Rectified Linear Unit.
Sigmoid	Hàm kích hoạt Sigmoid.
Tanh	Hàm kích hoạt Hyperbolic Tangent.
Softmax	Hàm kích hoạt Softmax, dùng trong phân loại nhiều lớp.
MSE	Sai số bình phương trung bình, dùng trong hồi quy.
Cross-entropy	Hàm mất mát Cross-entropy, dùng trong phân loại.
Gradient Descent	Thuật toán tối ưu hóa theo phương pháp giảm dần gradient.
SGD	Stochastic Gradient Descent, phương pháp giảm dần gradient ngẫu nhiên.
Adam	Thuật toán tối ưu hóa Adam, một biến thể của SGD.
n_components	Tham số chỉ định số lượng thành phần chính (PCA) hoặc số chiều đích (t-SNE).
perplexity	Tham số trong t-SNE, kiểm soát sự cân bằng giữa cấu trúc cục bộ và toàn cục.
random_state	Tham số đảm bảo tính tái lập của các quá trình ngẫu nhiên trong mô hình.
n_estimators	Tham số trong Random Forest, chỉ định số lượng cây quyết định trong rừng.
n_jobs	Tham số chỉ định số lượng lõi CPU sử dụng để tăng tốc độ huấn luyện.
learning_rate	Tốc độ học, tham số điều chỉnh bước cập nhật trọng số trong quá trình tối ưu hóa.
epochs	Số lượt duyệt toàn bộ tập dữ liệu huấn luyện trong quá trình huấn luyện mạng nơ-ron.

Danh sách bảng

1	Thống kê mô tả cho các trường dữ liệu định lượng	26
2	Mô tả dữ liệu đã chuẩn hóa	28
3	Số lượng thành phần chính cần thiết tương ứng với các ngưỡng phương sai và tỉ lệ giảm chiều.	32
4	Hệ số tương quan giữa AQI và các biến ô nhiễm không khí	34
5	Đánh giá chất lượng phân cụm theo các phương pháp giảm chiều	42
6	Giá trị BIC và số cụm tối ưu cho mô hình GMM áp dụng trên ba dạng dữ liệu đầu vào.	47
7	So sánh chất lượng phân cụm GMM trên ba loại dữ liệu dựa theo các chỉ số DBI và CHI	48
8	So sánh DBI và CHI giữa K-Means và GMM	53
9	Hiệu suất của Random Forest với ba phương pháp tiền xử lý (gốc, PCA, t-SNE) ở các tỉ lệ train:validation khác nhau	57
10	So sánh MSE và Overfit Ratio cho ba bộ dữ liệu (Original, PCA, t-SNE) với tỷ lệ Train:Test = 7:3	60
11	Kết quả huấn luyện và kiểm định (MSE) trên các loại tiền xử lý dữ liệu (bảng viền kín).	62
12	Bảng kết quả tổng quan theo tỉ lệ train:test	64
13	Hiệu suất của MLP trên 3 loại dữ liệu (Gốc, PCA, t-SNE) với các tỉ lệ train:validation	67
14	So sánh MSE và Overfit Ratio trên ba bộ dữ liệu (Original, PCA, t-SNE) với tỷ lệ Train:Test = 7:3	69
15	Ảnh hưởng của hệ số regularization alpha đến kết quả huấn luyện và kiểm định (MSE), với tỉ lệ chia dữ liệu 7:3 trên các phương pháp tiền xử lý khác nhau.	72
16	Bảng kết quả tổng quan đánh giá mô hình MLP theo các tỉ lệ train:test và loại dữ liệu khác nhau .	73
17	Kết quả đánh giá mô hình với các tỉ lệ chia train-validation và dữ liệu gốc vs PCA	78
18	So sánh tất cả các kết quả chi tiết (gốc vs PCA) ở các tỉ lệ Train:Val	79
19	Kết quả đánh giá mô hình với các tỉ lệ chia train-validation và dữ liệu gốc vs PCA	81
20	Chi tiết kết quả phân loại theo nhãn cho các tỉ lệ Train:Val và loại dữ liệu (Gốc vs PCA)	81

Danh sách hình vẽ

1	Biểu đồ phương sai giải thích của các thành phần chính	29
2	Ma trận biểu đồ phân tán cho 4 thành phần chính đầu tiên	30
3	Trực quan hóa dữ liệu theo các cặp thành phần chính	31
4	Phân tích phương sai giải thích và phương sai tích lũy	32
5	Ma trận tương quan giữa các biến PCA - AQI	33
6	Biểu đồ mật độ sau khi giảm chiều bằng t-SNE	35
7	Phân bố các mức chất lượng không khí (AQI_Bucket) sau khi giảm chiều bằng t-SNE	35
8	Phân bố chỉ số AQI trong không gian 2D sử dụng t-SNE	36
9	Hệ số tương quan Pearson giữa AQI và các thành phần t-SNE	37
10	Trực quan hóa dữ liệu ô nhiễm không khí bằng t-SNE (2D)	38
11	So sánh PCA và t-SNE cho AQI_Bucket	39
12	Biểu đồ Silhouette Score cho ba phương pháp biểu diễn dữ liệu: Gốc, PCA và t-SNE	42
13	Phân phối chỉ số AQI trong từng cụm	43
14	Giá trị trung bình các đặc trưng theo cụm với các phương pháp giảm chiều khác nhau	43
15	Phân phối và biến động AQI theo cụm	44
16	So sánh phân cụm K-Means trên ba không gian dữ liệu: gốc, PCA và t-SNE	45
17	Biểu diễn chỉ số AQI theo kích thước điểm trong không gian t-SNE	46
18	Biểu đồ BIC theo số cụm và loại ma trận hiệp phương sai cho từng kiểu dữ liệu: Original, PCA và t-SNE	47
19	Biểu đồ violin thể hiện phân phối chỉ số AQI theo các cụm được phân loại bằng GMM trên ba loại dữ liệu (Original, PCA, t-SNE). Giá trị trung bình AQI của từng cụm được đánh dấu phía trên bằng chữ đỏ.	48
20	Heatmap thể hiện giá trị trung bình của các đặc trưng ô nhiễm trong từng cụm GMM trên ba loại dữ liệu. Các ô được khoanh đỏ biểu thị đặc trưng có giá trị trung bình cao nhất trong cụm tương ứng.	49
21	Biểu đồ hộp (boxplot) minh họa sự phân bố AQI trong các cụm được phân cụm bằng GMM trên dữ liệu Original, PCA và t-SNE. Các hệ số biến động (CV) được chú thích bằng chữ đỏ trên các boxplot.	50
22	So sánh kết quả phân cụm bằng GMM trên ba dạng biểu diễn dữ liệu: Original, PCA và t-SNE. Các điểm dữ liệu được hiển thị trong không gian t-SNE 2D để thuận tiện cho trực quan hóa. Các tâm cụm được đánh dấu bằng dấu sao đỏ.	51

23	Trực quan phân cụm GMM trên không gian t-SNE 2D với kích thước điểm biểu thị giá trị AQI tương ứng. Các tâm cụm được đánh dấu bằng dấu sao đỏ.	52
24	Tương quan giữa phần dư và các đặc trưng đầu vào	65
25	Tương quan giữa phần dư và các đặc trưng đầu vào	74

Chương 0

Mở đầu

0.1 Giới thiệu chung

Ô nhiễm không khí là thách thức môi trường toàn cầu, phát sinh từ sự gia tăng dân số đô thị, hoạt động công nghiệp, giao thông và đốt nhiên liệu hóa thạch, dẫn tới nồng độ cao của bụi mịn ($PM_{2.5}$, PM_{10}) và các khí CO , NO_2 , SO_2 , O_3 , gây nguy hại nghiêm trọng đến sức khỏe con người. Chỉ số chất lượng không khí (AQI) ra đời nhằm tổng hợp các thông số ô nhiễm thành thang điểm trực quan, hỗ trợ cơ quan quản lý và người dân nắm bắt nhanh tình hình và khuyến nghị sức khỏe. Tuy nhiên, việc lắp đặt và vận hành cảm biến AQI khắp nơi gặp nhiều trở ngại về chi phí, thiết bị và xử lý dữ liệu thời gian thực. Do đó, nghiên cứu xây dựng mô hình dự báo AQI dựa trên dữ liệu ô nhiễm đã đo được là rất cần thiết, giúp ước lượng ở các khu vực thiếu trạm quan trắc, phát hiện sớm nguy cơ ô nhiễm và phục vụ công tác hoạch định chính sách hiệu quả.

0.2 Bài toán

Dự đoán giá trị chỉ số chất lượng không khí (AQI) dựa trên các thông số ô nhiễm.

Lý do chọn đề tài

- AQI là chỉ số tổng hợp phản ánh chất lượng không khí.
- Có tương quan cao với nhiều thông số đo lường (PM_{10} , CO , $PM_{2.5}$).
- Chỉ thiêu 15.85% giá trị AQI trong tập dữ liệu.

Biến mục tiêu

AQI (giá trị số liên tục)

Biến đầu vào

- Các thông số ô nhiễm: $PM_{2.5}$, PM_{10} , NO , NO_2 , NOx , NH_3 , CO , SO_2 , O_3

- Thông tin địa điểm: City (mã hóa thành biến phân loại)
- Thông tin thời gian: Date (trích xuất các thuộc tính như tháng, mùa, ngày trong tuần)

0.3 Mục tiêu

Báo cáo này nhằm nghiên cứu và ứng dụng các phương pháp học máy để dự báo mức độ ô nhiễm không khí tại các thành phố. Các mục tiêu cụ thể bao gồm:

- Hiểu rõ cấu trúc và đặc trưng tiềm ẩn trong dữ liệu AQI thông qua kỹ thuật giảm chiều và phân cụm.
- Xây dựng các mô hình học máy để dự báo chỉ số AQI dựa trên dữ liệu lịch sử.
- Đánh giá, so sánh hiệu quả mô hình trên các không gian dữ liệu khác nhau (gốc và đã giảm chiều).
- Phân tích sai số và các yếu tố ảnh hưởng nhằm đề xuất hướng cải thiện chất lượng dự báo trong thực tế.

0.4 Phạm vi nghiên cứu

Tổng quan về tập dữ liệu

- Số lượng mẫu: 29,531 bản ghi
- Số lượng thành phố: 26 thành phố
- Phạm vi thời gian: Từ 01/01/2015 đến 01/07/2020
- Các thông số ô nhiễm: PM2.5, PM10, NO, NO2, NOx, NH3, CO, SO2, O3, Benzene, Toluene, Xylene
- Chỉ số đo lường: AQI (Chỉ số chất lượng không khí) và AQI_Bucket (Phân loại mức độ ô nhiễm)

Tình trạng dữ liệu thiếu

- Tất cả các thành phố và ngày đều có dữ liệu (0% thiếu)
- Xylene có tỷ lệ thiếu cao nhất (61.32%)
- PM10 và NH3 có tỷ lệ thiếu tương đối cao (37.72% và 34.97%)
- AQI và AQI_Bucket thiếu 15.85% giá trị

0.5 Phương pháp nghiên cứu

- **Tiền xử lý dữ liệu:** Xử lý giá trị thiếu, chuẩn hóa dữ liệu, mã hóa one-hot cho các biến danh mục (City, AQI_Bucket), xử lý ngoại lai, và thêm đặc trưng thời gian (Year, Month, Day).
- **Giảm chiều dữ liệu:**

- + PCA: Giảm số chiều để giữ 95% phương sai, loại bỏ nhiễu.
 - + t-SNE: Giảm xuống 2 chiều để trực quan hóa cấu trúc dữ liệu.
- **Phân cụm dữ liệu:**
- + K-means: số cụm K được chọn bằng phương pháp silhouette score trong khoảng 2-10, với random_state=42 để đảm bảo tính tái lập.
 - + DBScan: eps được xác định bằng phương pháp elbow (khoảng 0.5-1.0), min_samples thử nghiệm [3, 5, 10, 15, 20, 25, 30]
- **Mô hình học máy:**
- + Random Forest Regressor: Sử dụng 100 cây, tối ưu bằng bagging và chọn đặc trưng ngẫu nhiên.
 - + Multi-Layer Perceptron (MLP): Mạng nơ-ron với 3 lớp ẩn (100, 50, 25 nơ-ron), tối ưu bằng Adam, hàm mất mát MSE.
- **Đánh giá:** So sánh hiệu suất trên dữ liệu gốc, PCA và t-SNE, sử dụng RMSE, MAE, R², với các tỷ lệ train:test là 8:2, 7:3, 6:4.
- **Phân tích thống kê:** Phân tích phân phối, tương quan, phần dư và tầm quan trọng đặc trưng để đánh giá mô hình và xác định các yếu tố chính ảnh hưởng đến AQI.

Chương 1

Cơ sở lý thuyết

1.1 PCA (Principal Component Analysis)

1.1.1 Khái niệm

PCA (Phân tích Thành phần Chính) là một kỹ thuật dùng để giảm số lượng đặc trưng (chiều) trong dữ liệu, nhưng vẫn giữ lại được phần lớn thông tin quan trọng. Thay vì làm việc với tất cả các đặc trưng ban đầu, PCA sẽ tạo ra một tập mới gồm các đặc trưng tổng hợp gọi là “thành phần chính”. Những thành phần này được sắp xếp theo mức độ thể hiện thông tin – tức là các đặc trưng mới đầu tiên sẽ mang nhiều thông tin nhất từ dữ liệu gốc.

1.1.2 Mục tiêu

Mục tiêu cốt lõi của PCA là tìm ra một hệ trực tọa độ mới (các thành phần chính) sao cho:

- **Phương sai của dữ liệu** trên mỗi trục mới là lớn nhất có thể, tức là các trục này giữ lại nhiều nhất sự biến thiên của dữ liệu.
- **Các trục tọa độ mới** là **trục giao** với nhau, đảm bảo tính không tương quan giữa các thành phần chính.

Bằng cách này, PCA cho phép:

- **Giảm số chiều dữ liệu:** Loại bỏ các chiều có phương sai thấp, thường chứa ít thông tin hoặc nhiễu.
- **Tăng cường khả năng phân tích và trực quan hóa:** Dữ liệu sau khi giảm chiều dễ dàng được phân tích và trực quan hóa hơn, đặc biệt trong không gian 2D hoặc 3D.
- **Cải thiện hiệu suất của các thuật toán học máy:** Giảm chiều dữ liệu giúp giảm độ phức tạp tính toán và nguy cơ quá khớp (overfitting).

1.1.3 Nguyên lý hoạt động

Chuẩn hóa dữ liệu

Bước đầu tiên trong PCA là chuẩn hóa dữ liệu. Ta thực hiện trừ trung bình để mỗi biến có trung bình bằng 0, đồng thời thường chia cho độ lệch chuẩn để mỗi biến có phương sai bằng 1. Việc chuẩn hóa giúp cân bằng các biến có đơn vị đo lường khác nhau, tránh cho những biến có giá trị lớn chi phối kết quả phân tích.

Tính toán ma trận hiệp phương sai

Sau khi chuẩn hóa, ta tính ma trận hiệp phương sai \mathbf{S} của tập dữ liệu \mathbf{X} (đã chuẩn hóa):

$$\mathbf{S} = \frac{1}{n-1} \mathbf{X}^T \mathbf{X}$$

Trong đó, n là số mẫu. Ma trận \mathbf{S} thể hiện mức độ tương quan giữa các biến.

Tìm giá trị riêng và vector riêng

Tiếp theo, ta giải bài toán trị riêng (eigen-decomposition) cho ma trận \mathbf{S} :

$$\mathbf{S}\mathbf{u}_i = \lambda_i \mathbf{u}_i$$

Trong đó:

- λ_i là giá trị riêng (eigenvalue),
- \mathbf{u}_i là vector riêng (eigenvector) tương ứng.

Các vector riêng xác định các hướng chính (principal directions) của dữ liệu, và giá trị riêng cho biết lượng phương sai khi chiếu dữ liệu lên các hướng đó.

Chọn các thành phần chính

Sắp xếp các giá trị riêng theo thứ tự giảm dần $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d$, và chọn k giá trị riêng lớn nhất cùng các vector riêng tương ứng $\{\mathbf{u}_1, \dots, \mathbf{u}_k\}$.

Các vector này tạo thành không gian mới có chiều k , và dữ liệu được chiếu lên không gian này như sau:

$$\mathbf{Z} = \mathbf{X}\mathbf{W}_k$$

Trong đó $\mathbf{W}_k = [\mathbf{u}_1, \dots, \mathbf{u}_k]$ là ma trận gồm các vector riêng chính, và \mathbf{Z} là dữ liệu sau khi giảm chiều.

1.1.4 Bài toán tối ưu

Định nghĩa bài toán tối ưu

Xét một tập dữ liệu gồm n điểm trong không gian d chiều: $X = \{x_1, x_2, \dots, x_n\}$, với mỗi $x_i \in \mathbb{R}^d$. Giả sử dữ liệu đã được chuẩn hóa (có trung bình bằng 0).

PCA giải quyết bài toán tối ưu sau:

1. Tìm vector đơn vị $w_1 \in \mathbb{R}^d$ sao cho phương sai của dữ liệu khi chiếu lên w_1 là lớn nhất:

$$\max_{\|w_1\|=1} \text{Var}(Xw_1)$$

2. Tiếp theo, tìm w_2 vuông góc với w_1 sao cho phương sai của dữ liệu khi chiếu lên w_2 là lớn nhất:

$$\max_{\|w_2\|=1} \text{Var}(Xw_2) \text{ với điều kiện } w_1^T w_2 = 0$$

3. Tổng quát, tìm w_k vuông góc với tất cả các vector trước đó sao cho phương sai chiểu là lớn nhất:

$$\max_{\|w_k\|=1} \text{Var}(Xw_k) \text{ với điều kiện } w_i^T w_k = 0, \forall i < k$$

Giải bài toán tối ưu

Phương sai của dữ liệu khi chiếu lên vector w có thể được biểu diễn thông qua ma trận hiệp phương sai C :

$$\text{Var}(Xw) = w^T C w$$

trong đó $C = (1/n)X^T X$ là ma trận hiệp phương sai của dữ liệu.

Bài toán tối ưu trở thành:

$$\max_{\|w\|=1} w^T C w$$

Đây là một bài toán tối ưu ràng buộc, có thể giải bằng phương pháp nhân tử Lagrange:

$$L(w, \lambda) = w^T C w - \lambda (w^T w - 1)$$

Đạo hàm của L theo w và cho bằng 0:

$$\frac{\partial L}{\partial w} = 2Cw - 2\lambda w = 0$$

$$\Rightarrow Cw = \lambda w$$

Điều này có nghĩa là w là một vector riêng của ma trận hiệp phương sai C và λ chính là giá trị riêng tương ứng.

Do $w^T C w = \lambda (w^T w = 1)$, phương sai của dữ liệu khi chiếu lên w chính là λ .

Vì vậy, để tối đa hóa phương sai, w_1 chính là vector riêng tương ứng với giá trị riêng lớn nhất của C . Tương tự, w_2 là vector riêng tương ứng với giá trị riêng lớn thứ hai, và cứ tiếp tục như vậy.

1.1.5 Ưu điểm và Nhược điểm

Ưu điểm

- **Giảm chiều dữ liệu hiệu quả:** PCA giúp giảm số lượng biến đầu vào mà vẫn giữ lại phần lớn thông tin quan trọng, từ đó giảm độ phức tạp của mô hình và tăng tốc độ xử lý.
- **Loại bỏ đa cộng tuyến:** PCA tạo ra các biến mới không tương quan với nhau, giúp giải quyết vấn đề khi các đặc trưng gốc có sự tương quan cao.
- **Giảm nhiễu:** Bằng cách loại bỏ các thành phần có phương sai thấp (thường được coi là nhiễu), PCA giúp làm rõ dữ liệu.
- **Phát hiện điểm ngoại lai:** PCA có thể xác định các điểm dữ liệu bất thường bằng cách chỉ ra những điểm lệch lạc đáng kể trong không gian chiều.
- **Trực quan hóa dữ liệu:** Bằng cách giảm dữ liệu xuống 2 hoặc 3 chiều, PCA hỗ trợ trực quan hóa dữ liệu phức tạp một cách dễ dàng hơn.

Nhược điểm

- **Khó diễn giải các thành phần chính:** Các thành phần chính là các tổ hợp tuyến tính của các biến gốc, do đó việc hiểu ý nghĩa thực tế của chúng có thể khó khăn.
- **Yêu cầu chuẩn hóa dữ liệu:** Trước khi áp dụng PCA, cần chuẩn hóa dữ liệu để tránh việc các biến có đơn vị đo lường lớn hơn chi phối kết quả.
- **Mất mát thông tin:** Việc giảm chiều dữ liệu có thể dẫn đến mất mát một phần thông tin, đặc biệt nếu không chọn đủ số lượng thành phần chính cần thiết.
- **Nhạy cảm với ngoại lệ:** PCA có thể bị ảnh hưởng bởi các giá trị ngoại lệ, dẫn đến kết quả không chính xác.
- **Chỉ xử lý mối quan hệ tuyến tính:** PCA chỉ nắm bắt được các mối quan hệ tuyến tính giữa các biến; nếu dữ liệu có cấu trúc phi tuyến, PCA có thể không hiệu quả.

1.1.6 Một số ứng dụng

Phân tích dữ liệu và học máy

- **Giảm chiều dữ liệu:** PCA giúp giảm số lượng biến đầu vào, từ đó giảm độ phức tạp của mô hình và tăng hiệu suất tính toán.
- **Tiền xử lý dữ liệu:** Trước khi áp dụng các thuật toán học máy, PCA có thể được sử dụng để loại bỏ nhiễu và tập trung vào các đặc trưng quan trọng nhất.

- **Trực quan hóa dữ liệu:** Bằng cách giảm dữ liệu xuống 2 hoặc 3 chiều, PCA hỗ trợ trực quan hóa dữ liệu phức tạp một cách dễ dàng hơn.

Di truyền học và sinh học

- **Phân tích cấu trúc di truyền:** PCA được sử dụng để xác định các mẫu phân bố di truyền trong các quần thể, giúp hiểu rõ hơn về sự đa dạng và mối quan hệ giữa các nhóm dân cư.
- **Phân tích dữ liệu biểu hiện gen:** Trong nghiên cứu sinh học, PCA giúp giảm chiều dữ liệu biểu hiện gen, từ đó phát hiện các mẫu và mối quan hệ giữa các gen.

1.2 t-SNE (t-Distributed Stochastic Neighbor Embedding)

1.2.1 Khái niệm

t-SNE là một thuật toán giảm chiều phi tuyến được phát triển để trực quan hóa dữ liệu chiều cao trong không gian hai hoặc ba chiều. Phương pháp này chuyển đổi khoảng cách giữa các điểm trong không gian gốc thành phân phối xác suất và tìm cách bảo toàn các quan hệ gần nhau trong không gian mới. Nó đặc biệt hiệu quả trong việc phát hiện cấu trúc cục bộ và phân cụm dữ liệu.

1.2.2 Cấu trúc và nguyên lý hoạt động

Thuật toán t-SNE hoạt động qua hai giai đoạn chính:

- **Giai đoạn 1 – Xây dựng phân phối xác suất ở không gian cao chiều:**

- + Với mỗi cặp điểm dữ liệu \mathbf{x}_i và \mathbf{x}_j , tính xác suất tương tự $p_{j|i}$ dựa trên khoảng cách Euclidean và tham số σ_i :

$$p_{j|i} = \frac{\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|\mathbf{x}_i - \mathbf{x}_k\|^2/2\sigma_i^2)}$$

- + Điều chỉnh σ_i sao cho entropy của phân phối đạt giá trị xác định bởi perplexity:

$$Perp(P_i) = 2^{H(P_i)}, \quad \text{với} \quad H(P_i) = -\sum_j p_{j|i} \log_2 p_{j|i}$$

- + Tính xác suất đối xứng:

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}, \quad \text{với} \quad \sum_{i,j} p_{ij} = 1$$

- **Giai đoạn 2 – Tối ưu hóa trong không gian chiều thấp:**

- + Ánh xạ các điểm vào không gian chiều thấp \mathbf{y}_i , sử dụng phân phối t-Student để tính xác suất q_{ij} :

$$q_{ij} = \frac{(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}}{\sum_k \sum_{l \neq k} (1 + \|\mathbf{y}_k - \mathbf{y}_l\|^2)^{-1}}$$

- + Tối thiểu hóa độ phân kỳ KL giữa hai phân phối P và Q bằng gradient descent:

$$\text{KL}(P \parallel Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

1.2.3 Tham số quan trọng

- **Perplexity:** Độ rỗng của phân phối xác suất trong không gian gốc, ảnh hưởng đến cấu trúc cục bộ/lớn được bảo toàn.
- **Learning rate:** Ảnh hưởng đến tốc độ hội tụ trong quá trình tối ưu hóa.
- **Số chiều đích:** Thường chọn là 2 hoặc 3 để thuận tiện cho việc trực quan hóa.

1.3 MLP (Multi-layer Perceptron)

1.3.1 Khái niệm

MLP (Mạng Perceptron nhiều tầng) là một dạng mạng nơ-ron nhân tạo, thường được sử dụng trong các bài toán như phân loại và hồi quy. Khác với mạng Perceptron đơn giản chỉ có một tầng, MLP có thêm nhiều tầng ẩn, cho phép mô hình học được các mối quan hệ phức tạp và phi tuyến tính trong dữ liệu.

1.3.2 Kiến trúc mô hình

Một mạng MLP bao gồm:

- **Tầng đầu vào:** Là nơi nhận dữ liệu đầu vào. Số lượng nơ-ron ở tầng này bằng với số lượng đặc trưng của dữ liệu.
- **Các tầng ẩn:** Gồm một hoặc nhiều tầng, mỗi tầng có nhiều nơ-ron. Các tầng này giúp mô hình học được các mẫu phức tạp trong dữ liệu. Mỗi nơ-ron tại đây đều sử dụng một hàm kích hoạt để tạo tính phi tuyến.
- **Tầng đầu ra:**
 - + Với bài toán phân loại, tầng này thường có một hoặc nhiều nơ-ron (tuỳ vào số lớp) và dùng hàm kích hoạt phù hợp như Softmax.
 - + Với bài toán hồi quy, tầng đầu ra thường có 1 nơ-ron và có thể không sử dụng hàm kích hoạt.
- **Trọng số và độ lệch (bias):** Là những giá trị cần học trong quá trình huấn luyện mô hình.

1.3.3 Nguyên lý hoạt động

Quá trình hoạt động của MLP có thể tóm tắt thành hai bước chính:

- **Lan truyền xuôi (Forward propagation):** Dữ liệu được truyền qua từng tầng của mạng. Mỗi nơ-ron tính toán đầu ra dựa trên đầu vào nhận được và hàm kích hoạt.

- **Lan truyền ngược (Backpropagation):** Sau khi có dự đoán, mô hình tính toán độ sai lệch so với kết quả thực tế và điều chỉnh trọng số để giảm sai số này. Việc cập nhật được thực hiện lặp đi lặp lại qua nhiều vòng huấn luyện.

Hàm kích hoạt

Hàm kích hoạt giúp mô hình học được các mối quan hệ phi tuyến giữa các đặc trưng. Một số hàm kích hoạt phổ biến:

- **ReLU:** Được dùng nhiều ở tầng ẩn do đơn giản và hiệu quả.
- **Sigmoid:** Phù hợp với bài toán phân loại nhị phân.
- **Tanh:** Tương tự Sigmoid nhưng cho giá trị đầu ra nằm trong khoảng -1 đến 1.
- **Softmax:** Dùng cho phân loại nhiều lớp.
- **Tuyến tính:** Thường dùng trong bài toán hồi quy.

Hàm mất mát (Loss function)

Hàm mất mát đo sự khác biệt giữa giá trị dự đoán và giá trị thực tế. Tùy thuộc vào loại bài toán, hàm mất mát sẽ khác nhau:

- **Hồi quy:** Thường dùng sai số bình phương trung bình (MSE).
- **Phân loại:** Thường dùng hàm mất mát entropy chéo (Cross-entropy).

Quá trình huấn luyện MLP

- **Khởi tạo trọng số:** Các trọng số ban đầu được chọn ngẫu nhiên.
- **Tối ưu hóa:** Sử dụng các thuật toán như Gradient Descent, SGD hoặc Adam để cập nhật trọng số nhằm giảm sai số.
- **Điều chỉnh siêu tham số:** Bao gồm các yếu tố như số tầng ẩn, số nơ-ron mỗi tầng, learning rate, số vòng lặp huấn luyện (epoch), hàm kích hoạt và hàm mất mát.

1.4 Random Forest

1.4.1 Khái niệm

Random Forest là một thuật toán học máy thuộc nhóm học có giám sát (supervised learning) và được sử dụng phổ biến trong các bài toán phân loại (classification) và hồi quy (regression). Thuật toán này là một dạng của tập hợp học (ensemble learning), nơi mà nhiều mô hình yếu (weak learners), cụ thể là các cây quyết định (decision trees), được kết hợp lại để tạo thành một mô hình mạnh mẽ hơn.

1.4.2 Nguyên lý hoạt động

Random Forest kết hợp hai kỹ thuật chính: Bagging và lựa chọn đặc trưng ngẫu nhiên:

- **Bagging (Bootstrap Aggregating):** Từ tập dữ liệu huấn luyện D gồm N mẫu, tạo nhiều tập con bằng cách lấy mẫu ngẫu nhiên có hoán lại. Mỗi tập con được dùng để huấn luyện một cây quyết định riêng biệt. Khoảng $\sim 1/3$ số mẫu không được chọn (gọi là *out-of-bag*) được dùng để đánh giá mô hình.
- **Lựa chọn đặc trưng ngẫu nhiên:** Tại mỗi nút phân chia của cây, thay vì xét toàn bộ M đặc trưng, chỉ một tập con nhỏ F được chọn ngẫu nhiên (thường $F = \sqrt{M}$ hoặc $F = \log_2(M) + 1$). Điều này giúp tăng tính đa dạng giữa các cây.
- **Xây dựng cây quyết định:** Mỗi cây được xây dựng bằng cách phân tách không gian đặc trưng thành các vùng con dựa trên các điều kiện phân tách tại các nút, sử dụng tập con dữ liệu bootstrap và tập con đặc trưng ngẫu nhiên đã chọn. Quá trình này tiếp tục cho đến khi cây đạt đến độ sâu tối đa hoặc các điều kiện dừng khác (ví dụ: số lượng mẫu trong nút nhỏ hơn một ngưỡng)
- **Dự đoán:** Sau khi xây dựng xong tất cả các cây trong rừng, quá trình dự đoán cho một mẫu dữ liệu mới được thực hiện bằng cách cho mẫu đó đi qua từng cây để thu nhận dự đoán riêng biệt của từng cây.
 - + **Bài toán phân loại:** Kết quả cuối cùng được xác định bằng phương pháp biểu quyết đa số (majority voting): lớp được dự đoán là lớp mà đa số các cây quyết định chọn
 - + **Bài toán hồi quy:** Kết quả dự đoán là giá trị trung bình của các dự đoán từ tất cả các cây

1.4.3 Thuật toán

Đầu vào:

- **Tập huấn luyện:** $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$
- **Số lượng cây:** B
- **Số lượng đặc trưng được chọn ngẫu nhiên tại mỗi nút:** m

Mã giả

Hàm `RandomForest(D, B, m)`:

```
H ← ∅ // Khi то tip hp cc cy
```

Cho i từ 1 đến B :

$D_i \leftarrow$ Lấy mẫu bootstrap từ D (lấy mẫu với thay thế)

$T_i \leftarrow$ Xây dựng cây quyết định từ D_i với m đặc trưng được chọn ngẫu nhiên tại mỗi nút
 $H \leftarrow H \cup -T_i$

Trả về H // Tập hợp các cây quyết định tạo thành rừng

Hàm Dự đoán(H , x):

Cho mỗi cây T_i trong H :

$y_i \leftarrow$ Dự đoán của T_i với đầu vào x

$y \leftarrow$ Kết hợp các y_i :

- Nếu là bài toán phân loại: y là nhãn được đa số y_i bỏ phiếu
- Nếu là bài toán hồi quy: y là trung bình của các y_i

Trả về y

1.4.4 Ưu điểm và Nhược điểm

Ưu điểm

- **Khả năng xử lý dữ liệu lớn và phức tạp:** Random Forest có thể xử lý lượng lớn dữ liệu với độ chính xác cao, ngay cả khi dữ liệu chứa nhiễu hoặc phân bố không đồng đều.
- **Giảm thiểu hiện tượng overfitting:** Nhờ cơ chế kết hợp nhiều cây quyết định được huấn luyện trên các mẫu ngẫu nhiên khác nhau, Random Forest giúp giảm nguy cơ quá khớp so với cây quyết định đơn lẻ.
- **Xử lý tốt dữ liệu thiếu:** Thuật toán vẫn hoạt động hiệu quả ngay cả khi dữ liệu bị mất mát hoặc thiếu, nhờ việc mỗi cây chỉ sử dụng một phần dữ liệu và đặc trưng khác nhau.

Nhược điểm

- **Yêu cầu tài nguyên tính toán lớn:** Do phải xây dựng và kết hợp nhiều cây quyết định, thuật toán tiêu tốn nhiều bộ nhớ và thời gian xử lý hơn so với các mô hình đơn giản như cây quyết định đơn lẻ.
- **Mô hình phức tạp, khó giải thích:** Mặc dù mỗi cây riêng lẻ dễ hiểu, nhưng khi kết hợp nhiều cây lại tạo thành “hộp đen”, gây khó khăn trong việc phân tích và giải thích kết quả.
- **Không tối ưu cho ứng dụng thời gian thực:** Thời gian dự đoán có thể chậm nếu cần phản hồi nhanh trong các hệ thống thời gian thực.

1.4.5 Ứng dụng

Ứng dụng của mô hình phân loại Random Forest:

- Phân loại email thành thư rác (spam) hoặc không phải thư rác (non-spam) dựa trên các đặc điểm như từ khóa trong tiêu đề, nội dung, người gửi.
- Phân loại hình ảnh, nhận dạng khuôn mặt trong các ứng dụng thị giác máy tính.

- Phân loại cảm xúc trong phân tích mạng xã hội hoặc phản hồi khách hàng (positive, negative, neutral).
- Xác định hoạt động gian lận trong tài chính, phân loại các ứng viên cho vay trung thành trong ngân hàng.

Ứng dụng của mô hình hồi quy Random Forest:

- Dự đoán giá bất động sản dựa trên các yếu tố như diện tích, số phòng ngủ, vị trí.
- Dự báo giá cổ phiếu và phân tích thị trường tài chính.
- Dự báo thời tiết, khí hậu dựa trên dữ liệu phức tạp từ nhiều nguồn khác nhau như vệ tinh và cảm biến khí tượng.
- Dự đoán các chỉ số sức khỏe dựa trên dữ liệu sinh học và xét nghiệm.

1.5 K-means

Ý tưởng chính: phân cụm dữ liệu thành K nhóm sao cho điểm trong cụm thì gần nhau (theo khoảng cách Euclid), còn khác cụm thì xa nhau.

1. Chọn ngẫu nhiên K tâm cụm.
2. Gán điểm vào cụm gần nhất.
3. Cập nhật tâm cụm = trung bình các điểm trong cụm.
4. Lặp lại đến khi ổn định.

K-Means tốt khi:

- Dữ liệu có cấu trúc rõ ràng, không quá nhiễu.
- Số lượng cụm biết trước hoặc có thể chọn nhờ kỹ thuật như Elbow, Silhouette Score.

Ưu điểm: Đơn giản, nhanh, dễ dùng.

Nhược điểm: Phải chọn K , nhạy cảm với nhiễu và tâm khởi tạo.

Ứng dụng: Nhận diện mẫu, xử lý ảnh, phân tích dữ liệu.

1.6 GMM (Gaussian Mixture Model)

1.6.1 Khái niệm

GMM là một mô hình xác suất thuộc nhóm mô hình biến ẩn, được sử dụng để biểu diễn dữ liệu như một tổ hợp của nhiều phân bố Gaussian. GMM giả định rằng dữ liệu được tạo ra từ một hỗn hợp của K phân bố Gaussian đa chiều. Mỗi cụm được biểu diễn bằng một phân bố Gaussian với các tham số riêng (trung bình μ_k , ma trận hiệp phương sai Σ_k , và trọng số π_k). GMM được sử dụng để phân cụm dữ liệu không nhãn, trong đó mỗi điểm dữ liệu được gán xác suất thuộc về một cụm cụ thể dựa trên phân bố Gaussian tương ứng.

1.6.2 Quy trình học GMM

Ước lượng tham số: Các tham số của GMM (μ_k, Σ_k, π_k) được ước lượng bằng phương pháp ước lượng hợp lý cực đại (Maximum Likelihood Estimation - MLE).

Thuật toán EM: Vì việc ước lượng trực tiếp các tham số của GMM là phức tạp, thuật toán Expectation-Maximization (EM) được sử dụng để tối ưu hóa các tham số này qua các bước lặp:

- **E-step (Expectation):** Tính xác suất mỗi điểm dữ liệu thuộc về từng cụm (xác suất hậu nghiệm).
- **M-step (Maximization):** Cập nhật các tham số (μ_k, Σ_k, π_k) để tối đa hóa hàm hợp lý.

1.6.3 Kiến thức lý thuyết liên quan đến GMM

Phân bố Gaussian đa chiều (Multi-Dimensional Gaussian Distribution)

Định nghĩa: Một phân bố Gaussian đa chiều được sử dụng để mô hình hóa dữ liệu trong không gian \mathbb{R}^d . Mỗi điểm dữ liệu $x_n \in \mathbb{R}^d$ được giả định tuân theo phân bố chuẩn với hàm mật độ xác suất:

$$p(x_n|\mu, \Sigma) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x_n - \mu)^T \Sigma^{-1} (x_n - \mu)\right)$$

trong đó:

- μ : vector trung bình (mean), biểu thị tâm của phân bố.
- Σ : ma trận hiệp phương sai (covariance matrix), mô tả hình dạng và độ phân tán của phân bố.
- $|\Sigma|$: định thức của ma trận hiệp phương sai.

Giả định: Dữ liệu trong bài toán được lấy mẫu độc lập từ một hoặc nhiều phân bố Gaussian đa chiều. GMM tổng quát hóa bằng cách kết hợp nhiều phân bố Gaussian này.

Ước lượng hợp lý cực đại (MLE) cho GMM

Mục tiêu: Tìm các tham số μ_k, Σ_k, π_k sao cho hàm hợp lý (likelihood) của dữ liệu được tối đa hóa:

$$p(X|\theta) = \prod_{n=1}^N \sum_{k=1}^K \pi_k p(x_n|\mu_k, \Sigma_k)$$

trong đó $\theta = \{\pi_k, \mu_k, \Sigma_k\}_{k=1}^K$ là tập hợp các tham số, và

$$\sum_{k=1}^K \pi_k = 1.$$

Thách thức: Việc tối ưu hóa trực tiếp hàm hợp lý là khó do có tổng bên trong logarit (log-sum), dẫn đến việc sử dụng thuật toán EM để giải quyết.

Thuật toán Expectation-Maximization (EM)

Mục đích: Tối ưu hóa các tham số của GMM một cách lặp đi lặp lại.

Các bước:

- **Khởi tạo:** Khởi tạo ngẫu nhiên các tham số μ_k, Σ_k, π_k .
- **E-step:** Tính xác suất hậu nghiệm (responsibility) rằng mỗi điểm dữ liệu x_n thuộc về cụm k :

$$\gamma_{nk} = \frac{\pi_k p(x_n | \mu_k, \Sigma_k)}{\sum_{j=1}^K \pi_j p(x_n | \mu_j, \Sigma_j)}.$$

- **M-step:** Cập nhật các tham số dựa trên γ_{nk} :

$$\begin{aligned}\mu_k &= \frac{\sum_{n=1}^N \gamma_{nk} x_n}{\sum_{n=1}^N \gamma_{nk}}, \\ \Sigma_k &= \frac{\sum_{n=1}^N \gamma_{nk} (x_n - \mu_k)(x_n - \mu_k)^T}{\sum_{n=1}^N \gamma_{nk}}, \\ \pi_k &= \frac{1}{N} \sum_{n=1}^N \gamma_{nk}.\end{aligned}$$

Quá trình E-step và M-step được lặp lại cho đến khi hội tụ hoặc đạt số vòng lặp tối đa.

1.6.4 Ứng dụng

GMM được ứng dụng rộng rãi trong nhiều lĩnh vực như phân cụm, nhận dạng mẫu, xử lý ảnh và học không giám sát. Một số lý do khiến GMM trở nên phổ biến:

- **Tính chất toán học thuận tiện:** Hàm mật độ Gaussian có dạng đóng, dễ tính toán đạo hàm và tích phân.
- **Phù hợp với dữ liệu thực tế:** Nhiều tập dữ liệu tự nhiên có xu hướng phân bố gần giống Gaussian.
- **Linh hoạt:** Phân bố Gaussian đa chiều có thể mô hình hóa các cụm với hình dạng elip, phù hợp với nhiều loại dữ liệu.

Chương 2

Dữ liệu

2.1 Nguồn dữ liệu

Bộ dữ liệu `city_day.csv` sử dụng trong đề tài là bộ dữ liệu quan trắc chất lượng không khí tại Ấn Độ trong giai đoạn từ năm 2015 đến 2020, được thu thập từ nhiều trạm đo tại các thành phố khác nhau.

Mỗi bản ghi đại diện cho một ngày đo tại một địa điểm cụ thể, bao gồm thông tin về thời gian, vị trí và các chỉ số ô nhiễm không khí như: AQI, O₃, CO, SO₂, NO₂, v.v.

Bộ dữ liệu có thể được tải xuống ở đây [1]

2.2 Tiền xử lý dữ liệu

2.2.1 Đọc và khám phá dữ liệu ban đầu

- **Đọc dữ liệu:** Dữ liệu được đọc từ file CSV `city_day.csv` sử dụng thư viện pandas

- **Thông tin cơ bản:**

- Kích thước dữ liệu: 29,531 bản ghi × 16 trường
 - Loại dữ liệu:
 - + Định lượng (13 trường): *PM2.5, PM10, ...*
 - + Định tính (3 trường): *City, Date, AQI_Bucket*.
 - Thời gian: Từ 01/01/2015 đến 01/07/2020
 - Thành phố: 26 thành phố khác nhau (Ahmedabad, Aizawl, Amaravati, ...)

- **Các trường dữ liệu:**

STT	Tên trường	Kiểu dữ liệu	Mô tả
1	City	object	Tên thành phố
2	Date	object	Ngày tháng (dd/mm/yyyy)
3	PM2.5	float64	Nồng độ bụi mịn PM2.5 ($\mu\text{g}/\text{m}^3$)
4	PM10	float64	Nồng độ bụi mịn PM10 ($\mu\text{g}/\text{m}^3$)
5	NO	float64	Nồng độ Nitơ Oxit
6	NO2	float64	Nồng độ Nitơ Dioxit
7	NOx	float64	Tổng Nitơ Oxit
8	NH3	float64	Nồng độ Amoniac
9	CO	float64	Nồng độ Carbon Monoxide
10	SO2	float64	Nồng độ Sulfur Dioxide
11	O3	float64	Nồng độ Ozone
12	Benzene	float64	Nồng độ Benzen
13	Toluene	float64	Nồng độ Toluen
14	Xylene	float64	Nồng độ Xylen
15	AQI	float64	Chỉ số chất lượng không khí
16	AQI_Bucket	object	Phân loại chất lượng không khí

2.2.2 Xử lý dữ liệu thiêu

- Phân tích dữ liệu thiêu:

- PM10 và NH3 có tỷ lệ thiêu trên 30%.
- Xylene có tỷ lệ thiêu cao nhất (61.32%).

- Xử lý:

1. Chuyển cột Date sang định dạng datetime và trích xuất thêm các cột Year, Month, Day.
2. Điền giá trị thiêu của các cột số bằng:
 - Giá trị trung bình theo từng thành phố.
 - Nếu vẫn thiêu, dùng trung bình toàn cục.
3. Cột AQI_Bucket được ánh xạ từ giá trị AQI.

2.2.3 Xử lý giá trị ngoại lai (Outliers)

- Phương pháp: Sử dụng khoảng tứ phân vị (IQR). Các giá trị nằm ngoài khoảng $[Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR]$ được coi là ngoại lai.
- Hành động: Thay thế các giá trị ngoại lai bằng giới hạn trên/dưới thay vì loại bỏ.

2.2.4 Chuyển đổi dữ liệu

- One-hot encoding:

 - Cột AQI_Bucket và City được chuyển sang dạng nhị phân bằng pd.get_dummies.

- Chuẩn hóa dữ liệu số:

 - Sử dụng MinMaxScaler đưa dữ liệu số về khoảng [0, 1].

2.3 Phân tích thống kê mô tả dữ liệu

2.3.1 Tổng quan dữ liệu

2.3.2 Thông kê mô tả với df.describe()

Tiến hành thống kê mô tả cho các trường dữ liệu định lượng. Kết quả được trình bày trong Bảng 1.

Bảng 1: Thông kê mô tả cho các trường dữ liệu định lượng

Thông kê	PM2.5	PM10	NO	NO2	NOx	NH3	CO	SO2	O3	Benzene	Toluene	Xylene	AQI
Count	29531	29531	29531	29531	29531	29531	29531	29531	29531	29531	29531	29531	29531
Mean	59.80	109.90	15.98	27.44	30.49	20.97	1.13	12.75	33.63	2.20	6.12	2.68	161.16
Std	38.12	54.51	13.20	19.12	21.59	13.32	0.87	8.78	18.03	2.24	6.29	1.76	105.20
Min	0.04	0.01	0.02	0.01	0.00	0.01	0.00	0.01	0.01	0.00	0.00	0.00	13.00
25%	31.83	63.78	6.11	12.38	14.60	10.98	0.54	6.07	19.96	0.22	0.69	1.17	87.00
50%	50.06	108.52	11.02	23.24	26.33	18.37	0.91	9.95	32.87	1.62	4.34	3.11	119.00
75%	76.35	129.36	22.60	37.42	41.26	27.07	1.53	15.95	43.03	3.26	8.28	3.11	215.00
Max	143.12	227.74	47.32	74.98	81.25	51.21	3.02	30.76	77.62	7.83	19.66	6.03	407.00

Chương 3

Giảm chiều dữ liệu

3.1 PCA

3.1.1 Chuẩn hóa dữ liệu và phân tích thành phần chính

Chuẩn hóa dữ liệu

Trước khi tiến hành phân tích thành phần chính (PCA), việc chuẩn hóa dữ liệu là bước tiền xử lý quan trọng để đảm bảo kết quả phân tích chính xác và đáng tin cậy. Quá trình này giúp giải quyết các vấn đề về tỷ lệ khác nhau giữa các biến, từ đó tạo điều kiện cho PCA phát huy hiệu quả tối đa.

Trong nghiên cứu này, chúng tôi thực hiện chuẩn hóa dữ liệu bằng phương pháp StandardScaler, một kỹ thuật phổ biến chuyển đổi dữ liệu sao cho mỗi đặc trưng có giá trị trung bình bằng 0 và độ lệch chuẩn bằng 1. Công thức chuẩn hóa được áp dụng như sau:

$$z = \frac{x - \mu}{\sigma}$$

Trong đó:

- z là giá trị sau khi chuẩn hóa
- x là giá trị gốc
- μ là giá trị trung bình của đặc trưng
- σ là độ lệch chuẩn của đặc trưng

Sau khi chuẩn hóa, ma trận dữ liệu kết quả X_scaled_df duy trì cấu trúc cột giống với dữ liệu ban đầu, nhưng với các giá trị đã được chuẩn hóa. Điều này đảm bảo rằng mỗi đặc trưng trong tập dữ liệu sẽ có giá trị trung bình bằng 0 và độ lệch chuẩn bằng 1, tạo điều kiện lý tưởng cho việc thực hiện PCA trong các bước tiếp theo của phân tích.

Bảng 2 trình bày thống kê mô tả cho dữ liệu đã chuẩn hóa:

1. Đánh giá tính hiệu quả của quá trình chuẩn hóa

Bảng 2: Mô tả dữ liệu đã chuẩn hóa

	PM2.5	PM10	NO	NO2	NOx	NH3	CO
count	29531.00	29531.00	29531.00	29531.00	29531.00	29531.00	29531.00
mean	-2.5793e-16	-3.9652e-16	-1.6169e-16	9.6244e-17	1.4629e-16	-2.1559e-16	1.0779e-16
std	1.0000e+00	1.0000e+00	1.0000e+00	1.0000e+00	1.0000e+00	1.0000e+00	1.0000e+00
min	-1.5676e+00	-2.0160e+00	-1.2088e+00	-1.4342e+00	-1.4120e+00	-1.5735e+00	-1.3030e+00
25%	-7.3364e-01	-8.4626e-01	-7.4743e-01	-7.8741e-01	-7.3581e-01	-7.5030e-01	-6.8122e-01
50%	-2.5543e-01	-2.5398e-02	-3.7579e-01	-2.1954e-01	-1.9256e-01	-1.9536e-01	-2.5520e-01
75%	4.3411e-01	3.5692e-01	5.0155e-01	5.2193e-01	4.9889e-01	4.5749e-01	4.5869e-01
max	2.1857e+00	2.1617e+00	2.3750e+00	2.4859e+00	2.3509e+00	2.2692e+00	2.1686e+00

- Giá trị trung bình (mean):** Các giá trị trung bình của tất cả chất ô nhiễm đều xấp xỉ 0. Điều này xác nhận rằng quá trình chuẩn hóa đã thành công trong việc đưa giá trị trung bình của mỗi biến về 0, phù hợp với yêu cầu của phương pháp StandardScaler. Các giá trị trung bình rất nhỏ gần 0 chỉ là sai số số học trong quá trình tính toán máy tính, không ảnh hưởng đến tính chất của dữ liệu đã chuẩn hóa.
- Độ lệch chuẩn (standard deviation):** Độ lệch chuẩn của tất cả các biến đều chính xác bằng 1, điều này khẳng định quy trình chuẩn hóa đã thành công trong việc biến đổi dữ liệu đưa các biến về cùng một thang đo với phương sai bằng 1. Đặc điểm này đảm bảo rằng mỗi biến sẽ có mức độ ảnh hưởng tương đương trong các phân tích thống kê tiếp theo, đặc biệt là trong phân tích thành phần chính (PCA).

2. Tính đối xứng và phân phối: Phân tích các giá trị phần tư (25%, 50%, 75%) cùng với giá trị cực tiểu và cực đại cho thấy có sự khác biệt trong phân phối của các chất ô nhiễm:

- PM2.5 và PM10: Giá trị trung vị (50%) lần lượt là -0.255 và -0.025, thấp hơn giá trị trung bình (xấp xỉ 0), cho thấy phân phối có độ lệch dương.
- Xylene thể hiện đặc điểm phân phối khác biệt khi giá trị trung vị (0.245) lớn hơn đáng kể so với trung bình, cho thấy phân phối có độ lệch âm.
- Ozone (O3) có giá trị trung vị (-0.042) gần với giá trị trung bình, cho thấy phân phối tương đối đối xứng.

3. Khoảng biến thiên: Khoảng biến thiên (range) của các chất ô nhiễm cũng thể hiện sự khác biệt đáng chú ý:

- NO2 có khoảng biến thiên lớn nhất với giá trị từ -1.434 đến 2.486, phản ánh sự biến động lớn của NO2 trong không khí.
- Benzene và Toluene có giá trị cực tiểu tương đối cao (lần lượt là -0.980 và -0.973) so với các chất ô nhiễm khác, cho thấy mức độ giảm thấp của các chất này ít hơn so với các chất ô nhiễm khác.
- PM10 và PM2.5 có giá trị cực đại tương đương nhau (khoảng 2.16-2.19), phản ánh mối tương quan trong sự tăng cao của hai loại bụi mịn này.

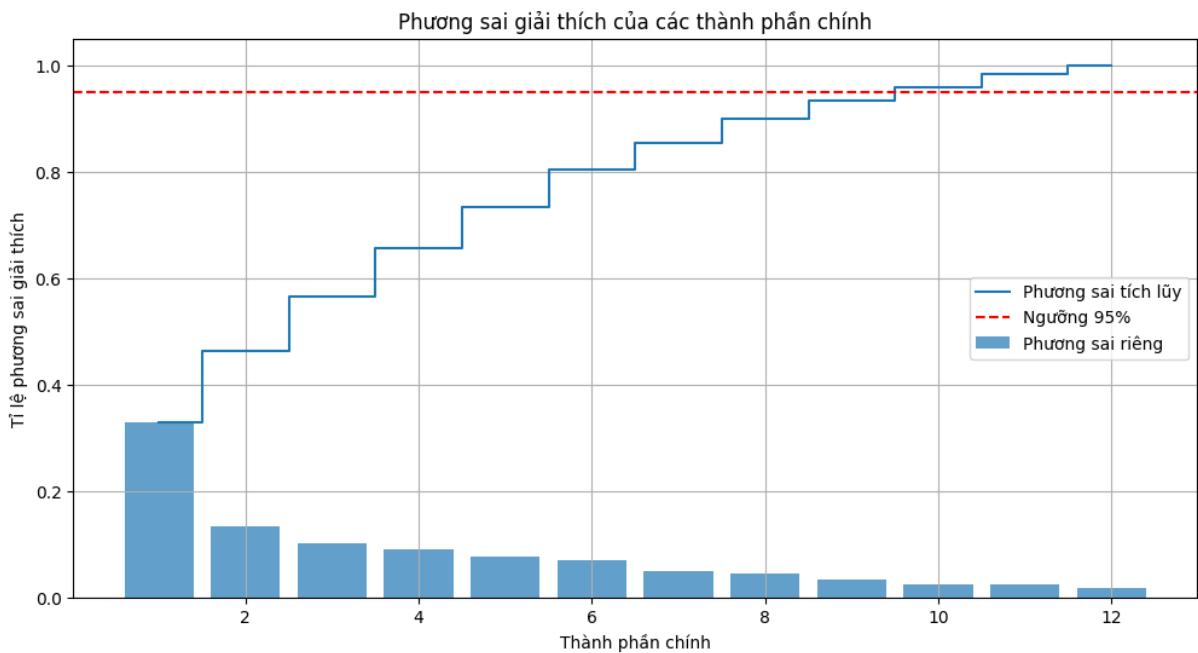
4. Đánh giá ngoại lai:

- Phân tích khoảng cách giữa các thông kê từ phân vị cho thấy không có giá trị ngoại lai cực đoan sau khi chuẩn hóa. Khoảng cách từ giá trị cực tiểu đến phần tư thứ nhất và từ phần tư thứ ba đến giá trị cực đại

tương đối cân đối với khoảng cách giữa các phần tử, phù hợp với tính chất của dữ liệu đã được chuẩn hóa.

Phân tích thành phần chính

Sau khi chuẩn hóa, ta sẽ chỉ chọn các cột số (numeric) để áp dụng phương pháp PCA. Kết quả cho ta thấy:



Hình 1: Biểu đồ phương sai giải thích của các thành phần chính

1. Phân bố phương sai giữa các thành phần chính

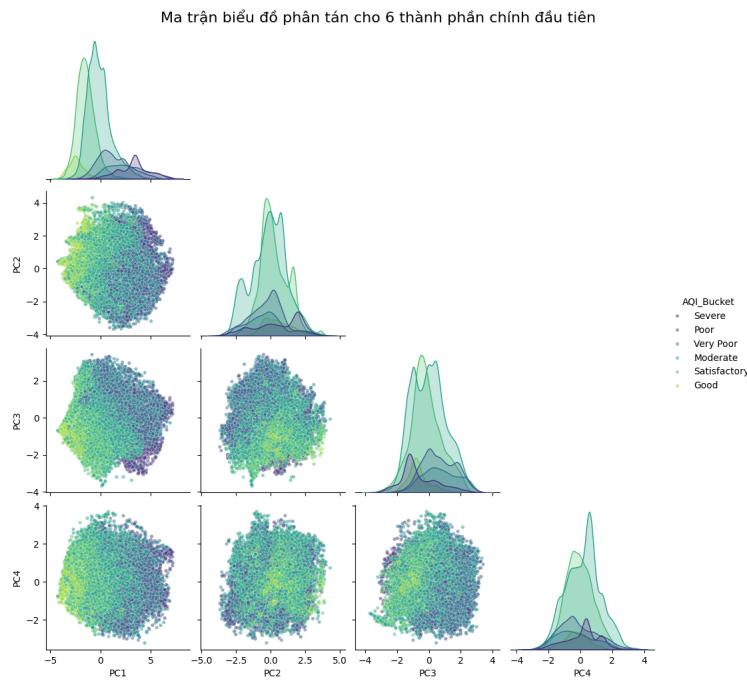
- Thành phần chính thứ nhất (PC1) đóng góp khoảng 33% tổng phương sai, chiếm tỷ lệ lớn nhất trong số các thành phần. Điều này cho thấy PC1 nắm giữ thông tin quan trọng nhất về biến động trong dữ liệu chất ô nhiễm và có khả năng phản ánh một yếu tố chung ảnh hưởng đến nhiều chất ô nhiễm cùng lúc.
- Thành phần chính thứ hai (PC2) giải thích khoảng 13% phương sai, giảm đáng kể so với PC1 nhưng vẫn đóng góp một phần quan trọng vào mô hình. PC2 có thể đại diện cho một nguồn ô nhiễm thứ cấp hoặc một quá trình môi trường khác biệt với yếu tố chính được phản ánh trong PC1.
- Từ PC3 đến PC6, mỗi thành phần giải thích khoảng 5-10% phương sai, cho thấy chúng vẫn nắm giữ thông tin có ý nghĩa về biến động trong dữ liệu.
- Các thành phần từ PC7 trở đi đóng góp rất nhỏ vào tổng phương sai (dưới 5% mỗi thành phần), cho thấy chúng có thể chỉ nắm giữ thông tin về nhiễu hoặc biến động ngẫu nhiên trong dữ liệu.

2. Phương sai tích lũy và ngưỡng lựa chọn

- Sáu thành phần chính đầu tiên giải thích khoảng 80% tổng phương sai trong dữ liệu.

- Để đạt được ngưỡng 95% phương sai (được thể hiện bằng đường nét đứt màu đỏ), cần sử dụng khoảng 10-11 thành phần chính. Cho phép giảm nhẹ số chiều của dữ liệu mà vẫn giữ được hầu hết thông tin.

Thực hiện hiển thị trực quan đối với dữ liệu theo từng cặp thành phần chính



Hình 2: Ma trận biểu đồ phân tán cho 4 thành phần chính đầu tiên

Ma trận biểu đồ phân tán (Scatter Plot Matrix) trên thể hiện mối quan hệ giữa sáu thành phần chính đầu tiên (PC1 đến PC4 được hiển thị rõ) từ phân tích PCA của dữ liệu chất ô nhiễm không khí. Các điểm dữ liệu được phân loại theo chỉ số chất lượng không khí (AQI) với sáu mức độ từ tốt (Good) đến nghiêm trọng (Severe).

- PC1 có mối liên hệ mạnh nhất với chỉ số AQI, cho thấy thành phần này có thể đại diện cho yếu tố chính ảnh hưởng đến chất lượng không khí, có thể là tổng nồng độ các chất ô nhiễm chính hoặc một nguồn phát thải chính.
- Mặc dù PC2 không thể hiện mối liên hệ tuyến tính rõ ràng với AQI, nhưng khi kết hợp với PC1, nó cung cấp thêm thông tin giúp phân biệt các mức độ ô nhiễm, đặc biệt là giữa mức "Good" và "Satisfactory".
- PC3 và PC4 thể hiện ít tương quan với AQI hơn, nhưng vẫn có thể đại diện cho các nguồn ô nhiễm cụ thể hoặc các yếu tố khí tượng ảnh hưởng đến sự phân tán của các chất ô nhiễm.

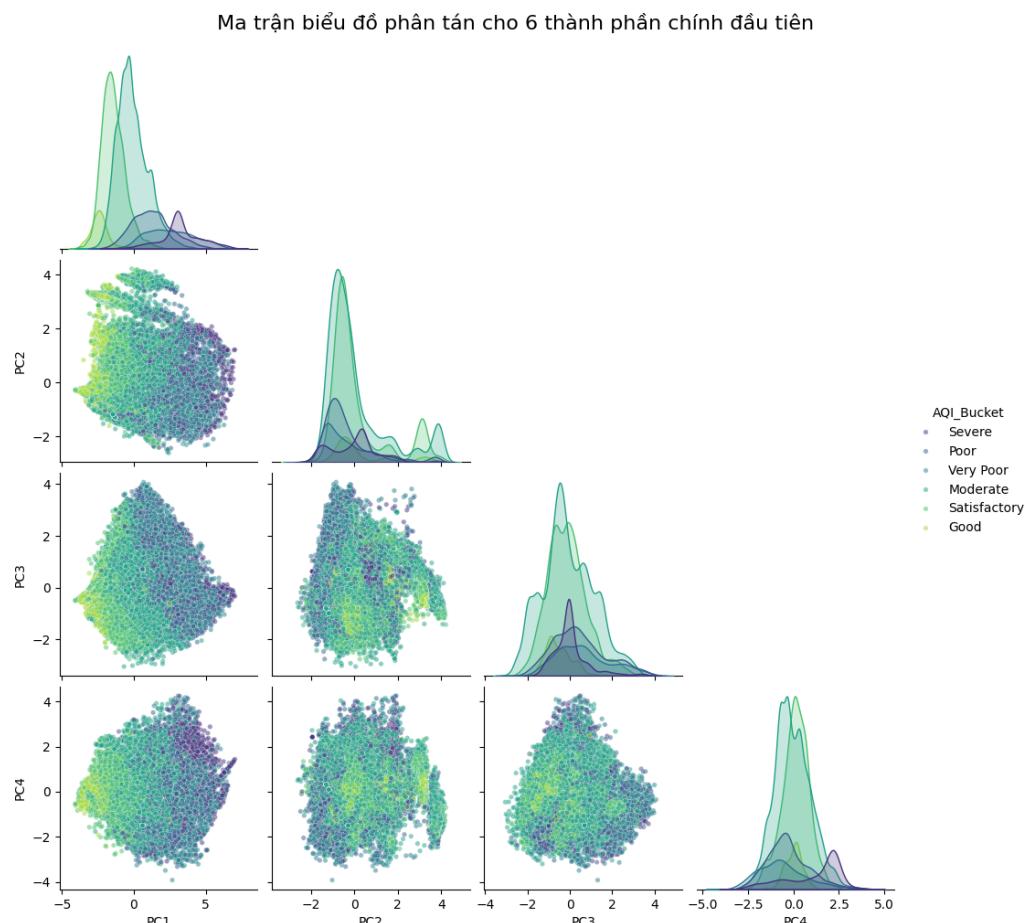
3.1.2 Trực quan hóa dữ liệu theo các cặp thành phần chính

Các cặp thành phần chính được vẽ trong không gian 2 chiều, cụ thể là giữa các thành phần chính PC1, PC2, PC1, PC3, PC2, PC3,... cho đến tối đa 4 cặp. Dưới đây là các bước thực hiện và kết quả trực quan hóa:

- Dữ liệu được phân nhóm theo mức AQI_Bucket và hiển thị dưới dạng các điểm trong đồ thị scatter.

- Các thành phần chính được đánh dấu bằng các trục PC_i, với tỉ lệ phương sai giải thích của mỗi thành phần được ghi rõ trong các trục.
- Từng cặp thành phần chính sẽ được vẽ trong từng đồ thị con.

Hình 5 dưới đây thể hiện kết quả trực quan hóa cho các cặp thành phần chính.

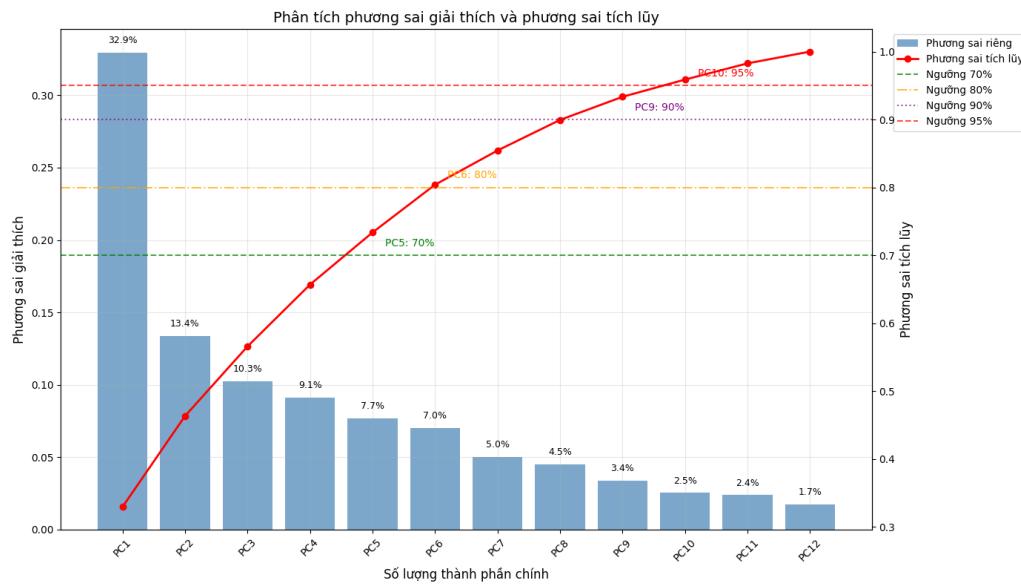


Hình 3: Trực quan hóa dữ liệu theo các cặp thành phần chính

Như có thể thấy, các cặp thành phần chính giúp phân biệt các mức AQI với nhau, ví dụ, các điểm màu đỏ (Severe) có xu hướng nằm ở một khu vực khác biệt so với các mức còn lại. Điều này chứng tỏ rằng các thành phần chính có thể hỗ trợ trong việc phân nhóm và phân tích dữ liệu chất lượng không khí.

3.1.3 Phân tích phương sai giải thích

Số lượng thành phần chính cần thiết để đạt các ngưỡng phương sai khác nhau:



Hình 4: Phân tích phương sai giải thích và phương sai tích lũy

Biểu đồ trên minh họa quá trình phân tích thành phần chính (PCA), trong đó:

- Phương sai giải thích riêng: được biểu diễn bằng các cột màu xanh, thể hiện phần trăm phương sai mà mỗi thành phần chính (PC) giải thích. Ví dụ: PC1 chiếm 32.9%, PC2 chiếm 13.4%, PC3 chiếm 10.3%, v.v.
- Phương sai tích lũy: được biểu diễn bằng đường màu đỏ, thể hiện tổng phương sai được giữ lại khi lần lượt cộng dồn các thành phần chính từ PC1 đến PCn.
- Các đường ngang (đứt nét) biểu thị các ngưỡng phương sai tiêu chuẩn như 70%, 80%, 90%, 95%, từ đó giúp xác định số lượng thành phần chính cần thiết để đạt mức độ giải thích tương ứng.

Ngưỡng phương sai	Số lượng thành phần chính cần thiết	Tỉ lệ giảm chiều (%)
50%	3	75.0%
60%	4	66.7%
70%	5	58.3%
80%	6	50.0%
90%	9	25.0%
95%	10	16.7%
99%	12	0.0%

Bảng 3: Số lượng thành phần chính cần thiết tương ứng với các ngưỡng phương sai và tỉ lệ giảm chiều.

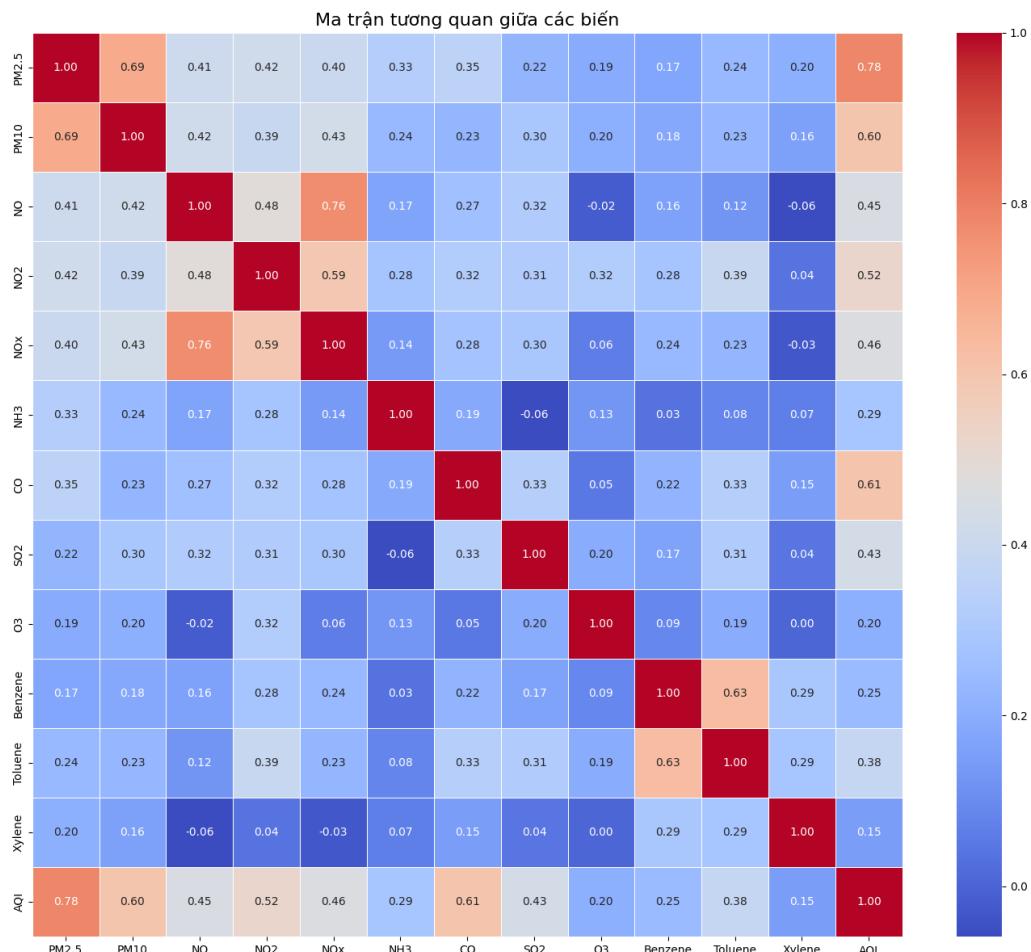
Nhận xét

- PCA giúp giảm số chiều của dữ liệu mà vẫn giữ lại phần lớn thông tin ban đầu.
- Nếu chỉ cần giữ lại khoảng 70–80% phương sai, ta chỉ cần 5–6 thành phần chính, tương đương với việc giảm từ 12 chiều xuống còn 5 hoặc 6 chiều, tiết kiệm 50–58% số chiều.
- Nếu yêu cầu giữ lại đến 95% phương sai, cần dùng tới 10 thành phần chính, chỉ giảm được 16.7% số chiều.

- Tùy theo mục tiêu ứng dụng (chính xác mô hình hay tốc độ tính toán), người dùng có thể lựa chọn những phù hợp để cân bằng giữa hiệu suất và độ phức tạp dữ liệu.

3.1.4 Trực quan hóa mối quan hệ giữa đặc trưng và đầu ra

Dưới đây là ma trận hệ số tương quan Pearson giữa một số thành phần chính (PCA) với đầu ra AQI.



Hình 5: Ma trận tương quan giữa các biến PCA - AQI

Ý nghĩa của ma trận tương quan

Ma trận tương quan hiển thị hệ số tương quan Pearson (r) giữa từng cặp biến. Hệ số này cho biết mức độ và chiều hướng mối liên hệ tuyến tính giữa hai biến:

- $r = 1$: Tương quan hoàn hảo cùng chiều.
- $r = -1$: Tương quan hoàn hảo ngược chiều.
- $r \approx 0$: Không có tương quan tuyến tính rõ ràng.

Ý nghĩa màu sắc trong biểu đồ tương quan:

- Đỏ đậm: Tương quan dương mạnh.

- Xanh đậm: Tương quan âm mạnh.
- Trắng hoặc màu nhạt: Ít hoặc không có tương quan.

Dựa trên ma trận tương quan, chỉ số AQI có mối tương quan chặt chẽ nhất với các chất ô nhiễm như PM2.5, PM10 và CO. Đây là những thành phần chính đóng vai trò quyết định trong việc xác định mức độ ô nhiễm không khí.

Ngoài ra, một số cặp biến khác như NO – NOx và Benzene – Toluene cũng thể hiện mối tương quan mạnh, phản ánh mối liên hệ về nguồn gốc hoặc cùng phát thải từ các hoạt động giao thông, công nghiệp hoặc nhiên liệu hóa thạch.

Bảng 4: Hệ số tương quan giữa AQI và các biến ô nhiễm không khí

Biến	Hệ số tương quan với AQI
AQI	1.000000
PM2.5	0.782712
CO	0.605764
PM10	0.601231
NO2	0.522026
NOx	0.461203
NO	0.453728
SO2	0.432119
Toluene	0.382466
NH3	0.286790
Benzene	0.246407
O3	0.203583
Xylene	0.148387

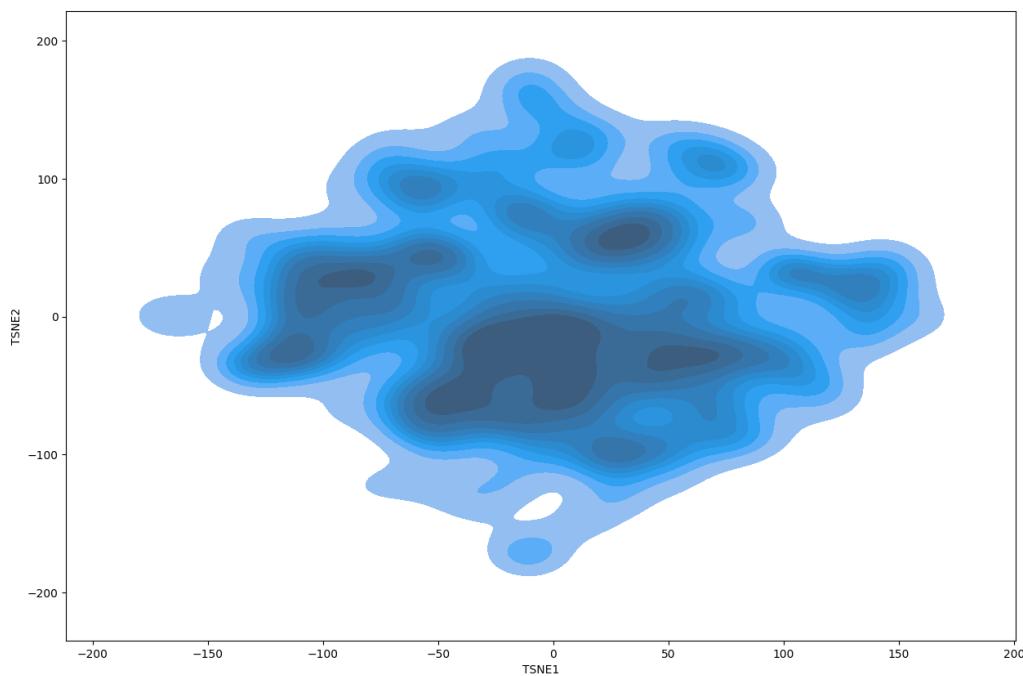
3.2 t-SNE

3.2.1 Phân tích thành phần chính

Sau chuẩn hóa, phương pháp t-SNE được áp dụng để giảm chiều dữ liệu và phân tích mức độ đóng góp của từng thành phần chính. Không giống như PCA tập trung vào việc bảo toàn phương sai toàn cục, t-SNE cố gắng bảo toàn cấu trúc cục bộ của dữ liệu, giúp phát hiện các cụm và mối quan hệ phức tạp hơn

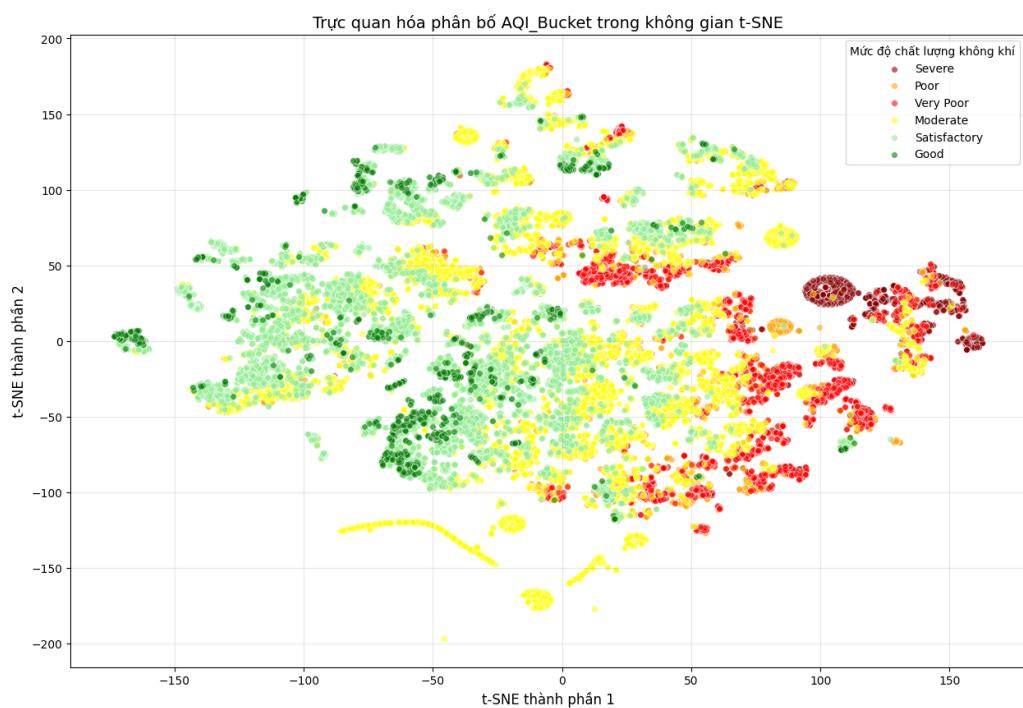
Ta thực hiện giảm chiều với t-SNE với các thông số:

- n components: Dữ liệu nhiều chiều ban đầu 12 được ánh xạ xuống 2 chiều để dễ dàng vẽ biểu đồ 2D.
- perplexity = 30: Là giá trị phổ biến, cân bằng giữa việc nắm được cấu trúc cục bộ và cấu trúc tổng thể.
- n iter = 1000: Cho chạy 1000 vòng lặp để thuật toán đủ thời gian hội tụ, cho ra kết quả tốt.
- random state = 42: Đặt hạt giống cho bộ sinh số ngẫu nhiên. Giúp việc chạy thuật toán t-SNE có thể tái lập (reproducible). Tức là nếu ta chạy lại thì sẽ thu được kết quả giống nhau. 42 là một con số phổ biến trong khoa học máy tính



Hình 6: Biểu đồ mật độ sau khi giảm chiều bằng t-SNE

Hình 9 thể hiện mật độ phân bố của các điểm dữ liệu trong không gian 2 chiều sau khi giảm chiều bằng t-SNE. Vùng màu đậm biểu thị khu vực tập trung nhiều điểm dữ liệu, cho thấy sự xuất hiện của các cụm dữ liệu tiềm năng. Dù không phân tách rõ ràng, dữ liệu vẫn có xu hướng hình thành một số vùng đậm đặc, phản ánh cấu trúc cục bộ của dữ liệu ban đầu.



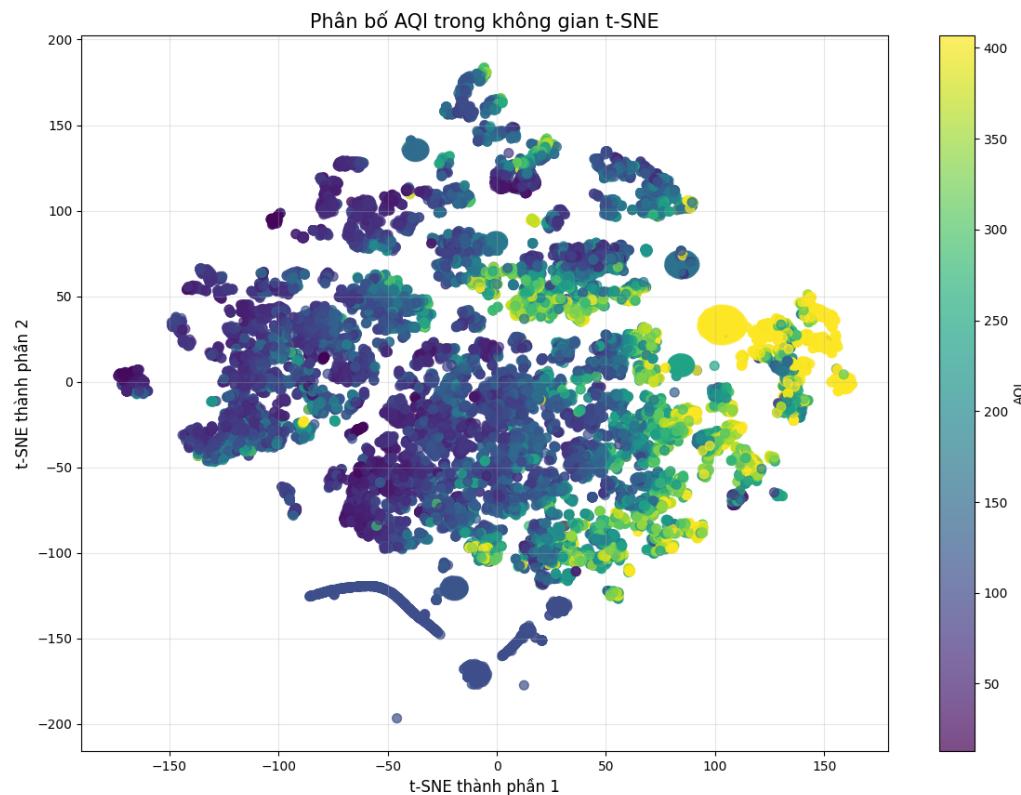
Hình 7: Phân bố các mức chất lượng không khí (AQI_Bucket) sau khi giảm chiều bằng t-SNE

Hình 10 minh họa cách các mức độ chất lượng không khí (*AQI_Bucket*) được phân bố trong không gian 2 chiều sau khi giảm chiều bằng t-SNE. Mỗi màu đại diện cho một mức AQI khác nhau (từ "Good" đến "Severe"). Ta có thể thấy một số mức AQI có xu hướng gom cụm lại (đặc biệt là "Severe" và "Good"), cho thấy t-SNE đã phần nào giữ được sự phân biệt giữa các nhóm chất lượng không khí.

3.2.2 Trực quan hóa dữ liệu theo các thành phần chính

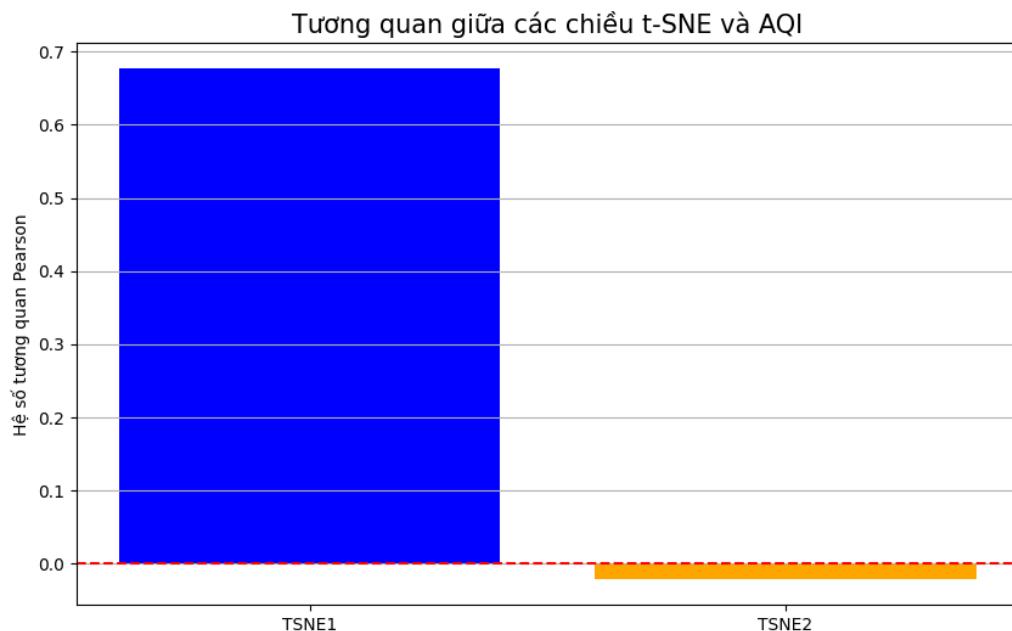
Sau khi thực hiện giảm chiều bằng t-SNE, dữ liệu được chuyển đổi thành một DataFrame với hai cột chính là TSNE1 và TSNE2, tương ứng với hai thành phần chính sau khi giảm chiều. Đồng thời, hai cột thông tin từ bộ dữ liệu ban đầu là AQI và AQI_Bucket cũng được thêm vào để phục vụ việc phân tích và biểu diễn trực quan.

Biểu đồ scatter bên dưới thể hiện phân bố các điểm dữ liệu trong không gian t-SNE với màu sắc biểu thị theo giá trị AQI. Các điểm có AQI cao hơn được biểu diễn với màu sắc sáng hơn (gần vàng), trong khi các điểm có AQI thấp hơn được biểu diễn bằng màu tối hơn (gần tím).



Hình 8: Phân bố chỉ số AQI trong không gian 2D sử dụng t-SNE

Để đánh giá mức độ liên hệ giữa từng chiều t-SNE với giá trị AQI, hệ số tương quan Pearson đã được tính toán giữa TSNE1, TSNE2 và AQI. Kết quả được thể hiện qua biểu đồ cột bên dưới.



Hình 9: Hệ số tương quan Pearson giữa AQI và các thành phần t-SNE

Từ biểu đồ ta thấy:

Chiều TSNE1 có hệ số tương quan dương khá cao với AQI (0.68), cho thấy chiều này có liên hệ đáng kể với biến mục tiêu AQI.

Ngược lại, chiều TSNE2 gần như không có tương quan với AQI (hệ số gần 0), cho thấy chiều này không mang nhiều thông tin liên quan đến sự thay đổi của AQI.

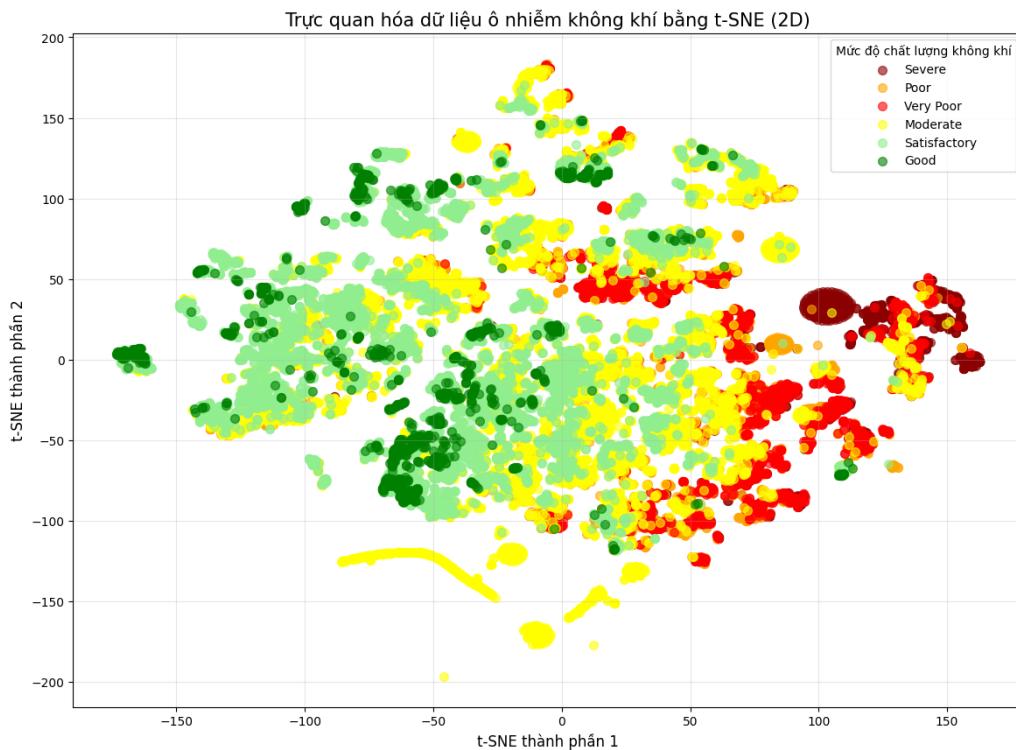
→ Nhận xét:

Tổng thể, việc sử dụng t-SNE giúp trực quan hóa rõ ràng sự phân bố của AQI trong không gian hai chiều. Chiều TSNE1 cho thấy tiềm năng cao trong việc đại diện cho sự biến thiên của AQI, và có thể đóng vai trò quan trọng nếu áp dụng các mô hình học máy sau này. Việc phát hiện các vùng tập trung AQI cao/thấp cũng hỗ trợ tốt cho việc xác định các cụm ô nhiễm không khí.

3.2.3 Trực quan hóa mối quan hệ giữa đặc trưng và đầu ra

Để kiểm tra mối quan hệ giữa các đặc trưng đầu vào và đầu ra AQI_Bucket, kỹ thuật t-SNE đã được sử dụng để giảm chiều dữ liệu xuống không gian 2D. Việc giảm chiều này cho phép trực quan hóa cấu trúc phân cụm trong dữ liệu và đánh giá xem các mức AQI có được phân biệt rõ ràng hay không.

Hình dưới đây mô tả phân bố của các mẫu dữ liệu trên mặt phẳng hai chiều sau khi áp dụng t-SNE, với mỗi điểm đại diện cho một quan sát và được tô màu theo mức chất lượng không khí (AQI_Bucket):



Hình 10: Trực quan hóa dữ liệu ô nhiễm không khí bằng t-SNE (2D)

Biểu đồ cho thấy các nhóm AQI_Bucket có xu hướng tập trung thành từng cụm riêng biệt, đặc biệt rõ ràng với các mức cực đoan như Good và Severe.

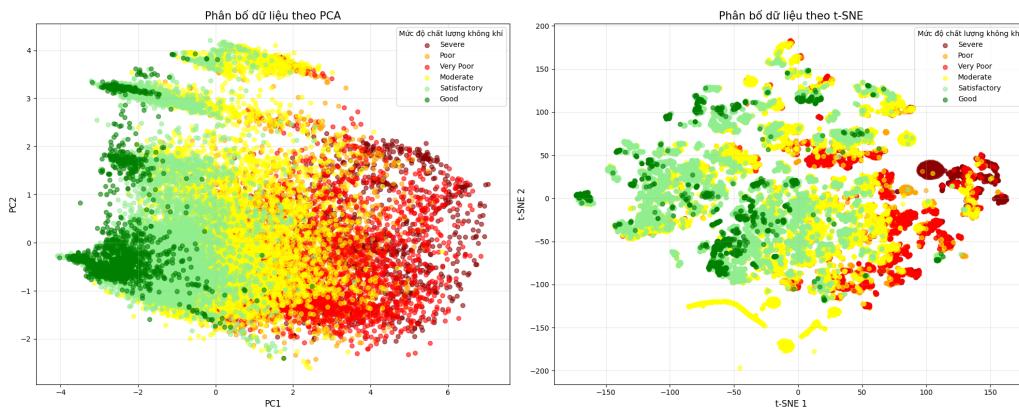
Các điểm dữ liệu có mức độ ô nhiễm cao (Severe, Very Poor, Poor) phân bố rõ ràng tại các vùng riêng biệt so với nhóm dữ liệu có mức AQI thấp (Good, Satisfactory).

Điều này phản ánh rằng các đặc trưng đầu vào (các chỉ số môi trường như PM2.5, PM10, NO2, CO, SO2, v.v.) có ảnh hưởng đáng kể và phân biệt rõ ràng giữa các mức AQI, cho thấy mối liên hệ mạnh giữa đặc trưng và đầu ra.

Từ trực quan hóa này, có thể khẳng định rằng việc sử dụng các đặc trưng hiện có để dự báo mức chất lượng không khí là hoàn toàn khả thi và có tiềm năng áp dụng trong các mô hình học máy.

3.3 So sánh PCA và t-SNE

Biểu đồ trực quan hóa dữ liệu bằng hai kỹ thuật giảm chiều phổ biến: PCA (Principal Component Analysis) và t-SNE (t-distributed Stochastic Neighbor Embedding)



Hình 11: So sánh PCA và t-SNE cho AQI_Bucket

PCA - Principal Component Analysis

- **Ý nghĩa:** Biểu diễn dữ liệu theo 2 thành phần chính (PC1 và PC2), giữ lại phần lớn phương sai của dữ liệu gốc.
- **Nhận xét:** Dữ liệu có xu hướng phân tách tuyến tính:
 - Phía bên trái (PC1 thấp) chủ yếu là các mẫu có chất lượng không khí tốt: *Good, Satisfactory*.
 - Phía bên phải (PC1 cao) dần chuyển sang các mức AQI xấu hơn: *Moderate, Very Poor, Poor, Severe*.
- Điều này cho thấy thành phần PC1 có tương quan mạnh với mức độ chất lượng không khí, như đã thể hiện trong biểu đồ tương quan trước đó.

t-SNE - t-distributed Stochastic Neighbor Embedding

- **Ý nghĩa:** Kỹ thuật phi tuyến để trực quan hóa dữ liệu nhiều chiều trong không gian 2D. Mặc dù không bảo toàn phương sai, t-SNE giữ được cấu trúc cụm và khoảng cách cục bộ.
- **Nhận xét:**
 - Không phân bố theo trực tuyến rõ ràng, nhưng t-SNE vẫn nhóm được các mức AQI gần nhau về mặt bản chất.
 - Các vùng có chất lượng không khí xấu như *Severe, Poor* có xu hướng nằm gần nhau.
 - Các mức tốt hơn như *Good, Satisfactory* cũng tập trung ở vùng khác.
- So với PCA, t-SNE cho thấy một số cụm rõ rệt hơn, phản ánh khả năng phân tách cục bộ tốt hơn, dù việc giải thích các trực là khó hơn.

So sánh giữa PCA và t-SNE

- **PCA (Principal Component Analysis)** là phương pháp tuyến tính có thể lượng hóa bằng phương sai, giữ lại các thành phần chính giúp bảo toàn cấu trúc toàn cục. Phù hợp cho các bài toán tiền xử lý, giảm nhiễu, và dễ diễn giải các trực tọa độ. Tuy nhiên, nếu dữ liệu có cấu trúc phi tuyến hoặc cụm ẩn sâu, PCA khó phát hiện được rõ ràng.

- **t-SNE (t-distributed Stochastic Neighbor Embedding)** là kỹ thuật phi tuyến mạnh mẽ trong trực quan hóa dữ liệu nhiều chiều, giúp giữ lại cấu trúc cục bộ và phát hiện các cụm tự nhiên một cách rõ ràng. Dù vậy, t-SNE không bao toàn phuong sai và rất khó diễn giải ý nghĩa các chiều tọa độ.
- Trong bối cảnh bài toán dự báo ô nhiễm không khí, nơi dữ liệu thường mang tính phi tuyến cao và có phân cụm tự nhiên (theo mức độ AQI), t-SNE thể hiện ưu thế vượt trội về mặt trực quan hóa và khám phá cấu trúc dữ liệu.
- Tuy nhiên, PCA vẫn có vai trò quan trọng khi cần chuẩn bị dữ liệu cho các thuật toán học máy truyền thống, do khả năng giảm chiều hiệu quả và giữ lại thông tin toàn cục.

Tóm lại, việc lựa chọn giữa PCA và t-SNE phụ thuộc vào mục đích sử dụng: nếu để trực quan hóa và khám phá cụm dữ liệu thì t-SNE là lựa chọn tốt; còn nếu để giảm chiều cho học máy hoặc diễn giải dữ liệu thì PCA sẽ hiệu quả hơn.

Chương 4

Phân cụm dữ liệu

Trong phần này, chúng em thực hiện phân cụm dữ liệu đầu vào (sau khi loại bỏ trường đầu ra) bằng hai phương pháp: K-Means và Gaussian Mixture Model (GMM). Sau khi phân cụm, chúng em đánh giá mối quan hệ giữa các mẫu trong từng cụm và phân tích mối liên hệ giữa các đầu ra tương ứng của các cụm đã tạo. Các đánh giá được thực hiện dựa trên các độ đo định lượng, đồng thời cung cấp trực quan hóa minh họa rõ ràng.

4.1 K-means

4.1.1 Phân cụm K-Means và lựa chọn số cụm K tối ưu

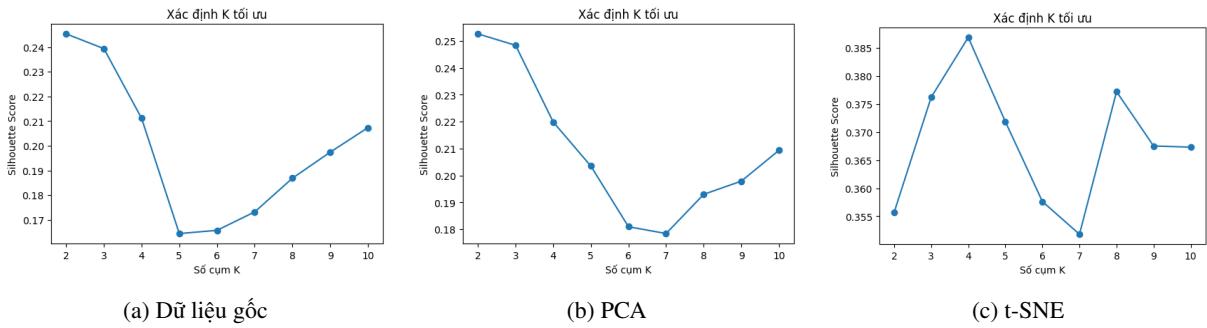
Sau khi loại bỏ trường đầu ra AQI, dữ liệu đầu vào đã được chuẩn hóa để đảm bảo các thuộc tính có cùng tỷ lệ ảnh hưởng trong quá trình phân cụm.

Lựa chọn số cụm K phù hợp

Thuật toán K-Means được áp dụng với các giá trị K khác nhau từ 2 đến 10. Để lựa chọn số cụm phù hợp, chỉ số *Silhouette Score* được sử dụng nhằm đánh giá chất lượng phân cụm. Quá trình này được thực hiện trên ba dạng dữ liệu:

- Dữ liệu gốc đã được chuẩn hóa.
- Dữ liệu sau khi giảm chiều bằng PCA (10 thành phần chính).
- Dữ liệu sau khi giảm chiều bằng t-SNE (2 chiều).

Biểu đồ Silhouette Score theo số cụm K cho từng trường hợp:



Hình 12: Biểu đồ Silhouette Score cho ba phương pháp biểu diễn dữ liệu: Gốc, PCA và t-SNE

Kết quả số cụm tối ưu và đánh giá

Số cụm K tối ưu được chọn dựa trên điểm silhouette cao nhất, kết quả như sau:

Loại dữ liệu	Số cụm K tối ưu	Silhouette Score
Chuẩn hóa	2	0.2454
PCA (10 thành phần)	2	0.2526
t-SNE (2 chiều)	4	0.3868

Kết quả cho thấy dữ liệu sau khi giảm chiều bằng t-SNE cho phân cụm tốt nhất với điểm silhouette cao nhất (0.3868), tương ứng số cụm $K = 4$. Do đó, trong các bước tiếp theo, phương án sử dụng cụm này được ưu tiên để phân tích và trực quan hóa dữ liệu.

4.1.2 Đánh giá mối quan hệ giữa các mẫu dữ liệu đầu vào và đầu ra trong từng cụm

Đánh giá chất lượng phân cụm qua các chỉ số định lượng

Hai chỉ số được sử dụng để đánh giá mức độ phân tách và độ đồng nhất của các cụm như sau:

- **Davies-Bouldin Index (DBI):** Giá trị càng thấp cho thấy các cụm càng tách biệt và đồng nhất hơn.
- **Calinski-Harabasz Index (CHI):** Giá trị càng cao cho thấy cấu trúc cụm càng rõ ràng.

Kết quả đánh giá:

Bảng 5: Đánh giá chất lượng phân cụm theo các phương pháp giảm chiều

Phương pháp giảm chiều	Davies-Bouldin Index ↓	Calinski-Harabasz Index ↑
ORIG (Dữ liệu gốc)	1.8565	7511.3845
PCA	1.8004	7949.7083
t-SNE	0.8239	26069.2734

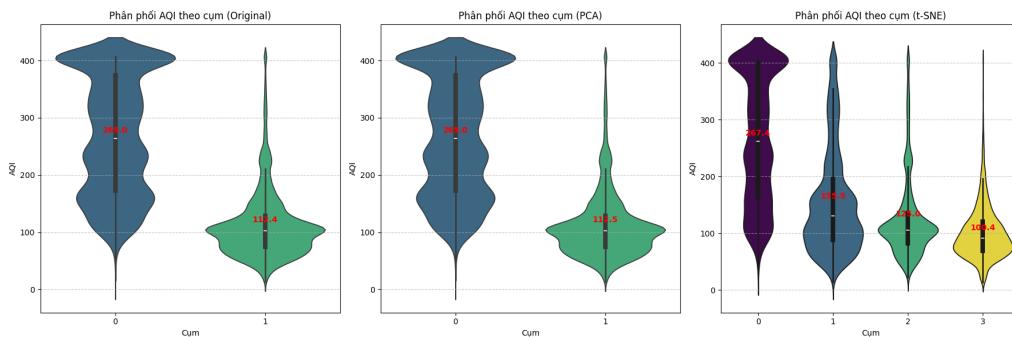
Chú thích:

- Chỉ số Davies-Bouldin (DB) càng nhỏ \Rightarrow chất lượng phân cụm càng tốt.
- Chỉ số Calinski-Harabasz (CH) càng lớn \Rightarrow cụm càng tách biệt rõ ràng.

Có thể thấy, phương pháp t-SNE cho kết quả phân cụm tốt nhất với DBI thấp nhất (0.8239) và CHI cao nhất (26069.2734), cho thấy sự tách biệt và đồng nhất của các cụm rõ ràng hơn.

Phân phối chỉ số AQI trong các cụm

Biểu đồ violin plot thể hiện phân phối chỉ số AQI trong từng cụm giúp đánh giá sự khác biệt về đầu ra giữa các nhóm dữ liệu. Giá trị trung bình AQI trong từng cụm cũng được thể hiện rõ.



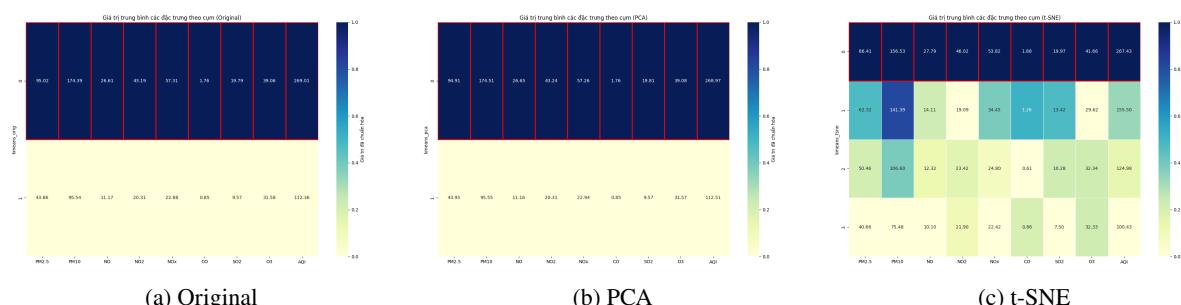
Hình 13: Phân phối chỉ số AQI trong từng cụm

Nhận xét: Biểu đồ violin thể hiện sự phân phối chỉ số AQI theo từng cụm khi áp dụng thuật toán KMeans trên ba không gian đặc trưng khác nhau.

- **Original và PCA:** Chỉ tạo ra hai cụm, trong đó cụm 0 có giá trị AQI trung bình cao (khoảng 269), còn cụm 1 có giá trị thấp hơn đáng kể (khoảng 112). Điều này cho thấy sự phân tách cụm còn đơn giản, chưa phản ánh được đầy đủ các mức độ ô nhiễm khác nhau.
- **t-SNE:** Thu được bốn cụm với giá trị trung bình AQI giảm dần từ cụm 0 đến cụm 3 (267.4 → 100.4). Điều này cho thấy không gian t-SNE giúp KMeans phân biệt rõ hơn các mức độ ô nhiễm, hỗ trợ phân tích AQI một cách chi tiết và hiệu quả hơn.

Mối quan hệ giữa các đặc trưng đầu vào trong từng cụm

Biểu đồ heatmap biểu diễn giá trị trung bình (đã chuẩn hóa) của các đặc trưng trong từng cụm giúp phân tích đặc trưng điển hình của mỗi cụm.



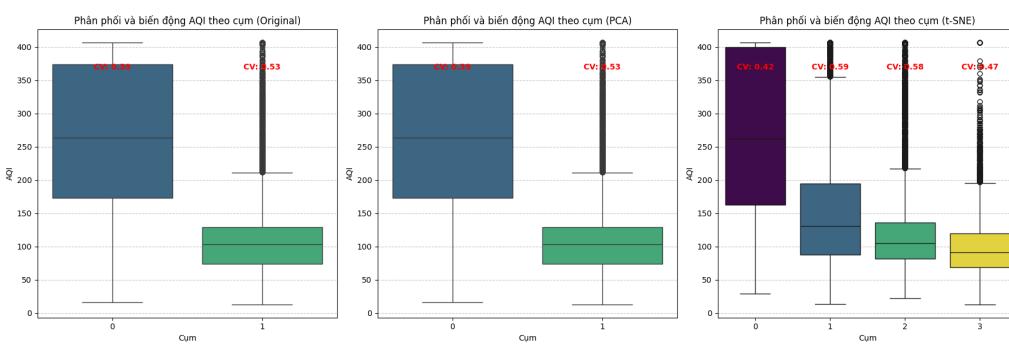
Hình 14: Giá trị trung bình các đặc trưng theo cụm với các phương pháp giảm chiều khác nhau

→ **Nhận xét:**

- (a) Original và (b) PCA: Các cụm phân tách chưa rõ ràng, thể hiện qua sự đồng nhất về màu sắc giữa các cụm. Điều này cho thấy KMeans chưa khai thác hiệu quả đặc trưng trong không gian gốc và không gian PCA để tạo ra các cụm đặc trưng.
- (c) t-SNE: Các cụm có sự khác biệt rõ ràng hơn về giá trị trung bình của các đặc trưng. Sự đa dạng về màu sắc chứng tỏ KMeans hoạt động hiệu quả hơn trong không gian t-SNE, hỗ trợ tốt cho việc phân tách cụm dựa trên đặc trưng.

Đánh giá mức độ biến động AQI trong từng cụm

Độ ổn định của chỉ số AQI trong mỗi cụm được đánh giá bằng hệ số biến thiên (CV = độ lệch chuẩn / giá trị trung bình).



Hình 15: Phân phối và biến động AQI theo cụm

Nhận xét:

- Các cụm có CV thấp cho thấy AQI ổn định, phản ánh tính đồng nhất của các mẫu.
- Một số cụm có CV cao, biểu thị mức độ không đồng đều trong ô nhiễm, có thể do nhiều hoặc ảnh hưởng từ nhiều yếu tố khác nhau.

→ **Kết luận:**

Phân tích cho thấy phân cụm không chỉ nhóm các mẫu dữ liệu đầu vào có tính chất tương đồng mà còn giúp nhận diện các mức độ ô nhiễm khác nhau thông qua chỉ số AQI. Cả kết quả định lượng và trực quan đều ủng hộ việc sử dụng t-SNE kết hợp K-Means với $K = 4$ là phương án phân cụm tốt nhất trong trường hợp này.

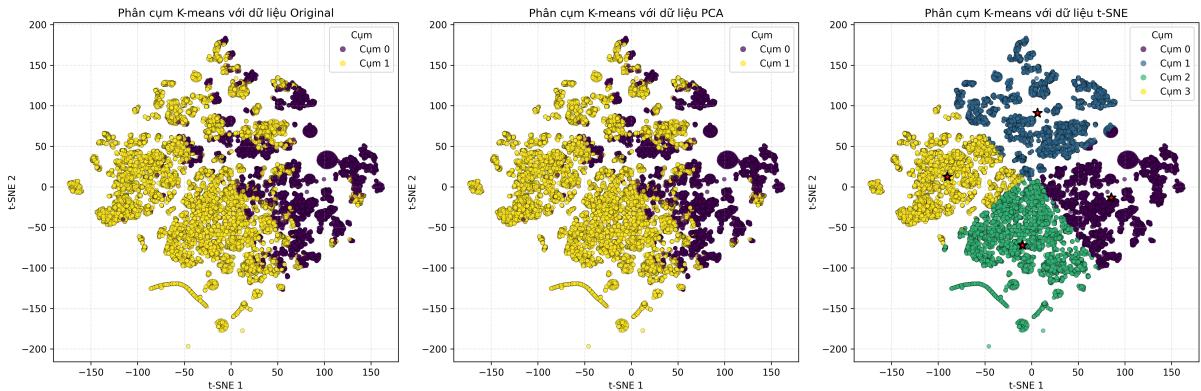
4.1.3 Trực quan hóa kết quả phân cụm

So sánh phân cụm trên không gian t-SNE

Để trực quan hóa kết quả phân cụm, dữ liệu được ánh xạ xuống không gian 2 chiều bằng phương pháp t-SNE (t-distributed Stochastic Neighbor Embedding), một kỹ thuật giảm chiều phi tuyến giúp bảo toàn cấu trúc cục bộ của dữ liệu.

Hình 16 minh họa kết quả phân cụm K-Means trên ba không gian dữ liệu:

- **Original:** Dữ liệu gốc chưa qua biến đổi.
- **PCA:** Dữ liệu đã giảm chiều tuyến tính bằng PCA.
- **t-SNE:** Dữ liệu đã giảm chiều phi tuyến bằng t-SNE.



Hình 16: So sánh phân cụm K-Means trên ba không gian dữ liệu: gốc, PCA và t-SNE

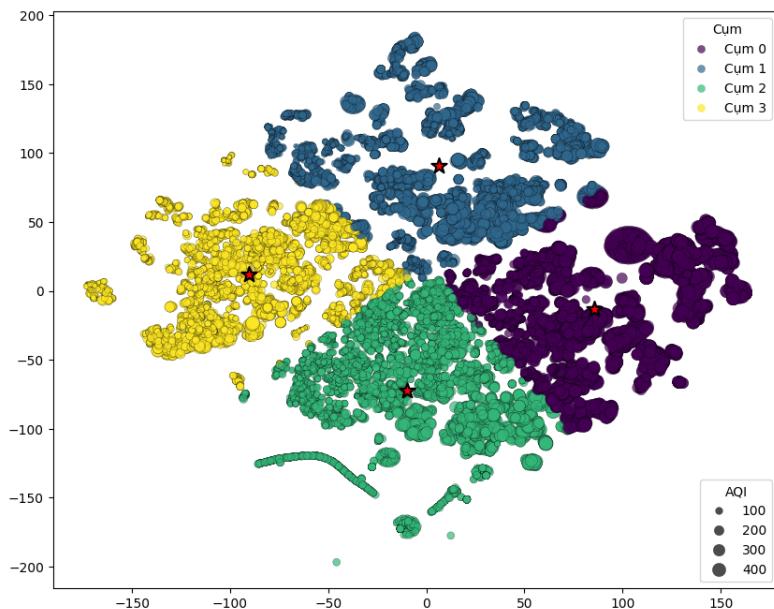
Nhận xét:

- Ở cả ba không gian, các điểm dữ liệu được phân cụm rõ ràng với màu sắc phân biệt.
- Phân cụm trên không gian PCA giúp gom cụm tốt hơn so với dữ liệu gốc.
- Phân cụm trên không gian t-SNE có ranh giới cụm rõ nét nhất, nhờ khả năng phân tách phi tuyến mạnh mẽ.

Biểu diễn AQI theo kích thước điểm trong không gian t-SNE

Biểu đồ dưới đây thể hiện kết quả phân cụm trên không gian t-SNE, trong đó **kích thước điểm tương ứng với chỉ số AQI**, nhằm giúp hiểu rõ hơn về mức độ ô nhiễm trong từng cụm.

- Kích thước điểm lớn biểu thị AQI cao, tức mức độ ô nhiễm không khí cao.
- Kích thước điểm nhỏ biểu thị AQI thấp, tức chất lượng không khí tốt.



Hình 17: Biểu diễn chỉ số AQI theo kích thước điểm trong không gian t-SNE

Nhận xét:

- Một số cụm chủ yếu gồm các điểm có kích thước nhỏ, phản ánh khu vực có chất lượng không khí tốt.
- Một vài cụm có các điểm với kích thước lớn rõ rệt, thể hiện nhóm điểm có AQI cao, cảnh báo mức độ ô nhiễm không khí.
- Sự khác biệt về kích thước trong các cụm thể hiện khả năng K-Means nhóm các mẫu dữ liệu có đặc trưng AQI tương đồng.

→ Kết luận:

- Trực quan hóa bằng t-SNE giúp làm nổi bật cấu trúc cụm ngay cả trong trường hợp dữ liệu nhiều chiều.
- Kết hợp t-SNE và K-Means cho phép phân chia dữ liệu thành các nhóm đặc trưng, giúp thuận tiện cho việc phân tích thêm về mức độ ô nhiễm không khí.
- Các biểu đồ phân phối và heatmap đặc trưng giúp hiểu sâu hơn về đặc tính từng cụm, đặc biệt là mối liên hệ giữa các chỉ số khí thải và AQI.

4.2 GMM (Gaussian Mixture Model)

4.2.1 Phân cụm GMM và lựa chọn số cụm phù hợp

Tương tự như K-Means, thuật toán GMM được áp dụng sau khi loại bỏ trường đầu ra AQI và chuẩn hóa dữ liệu đầu vào.

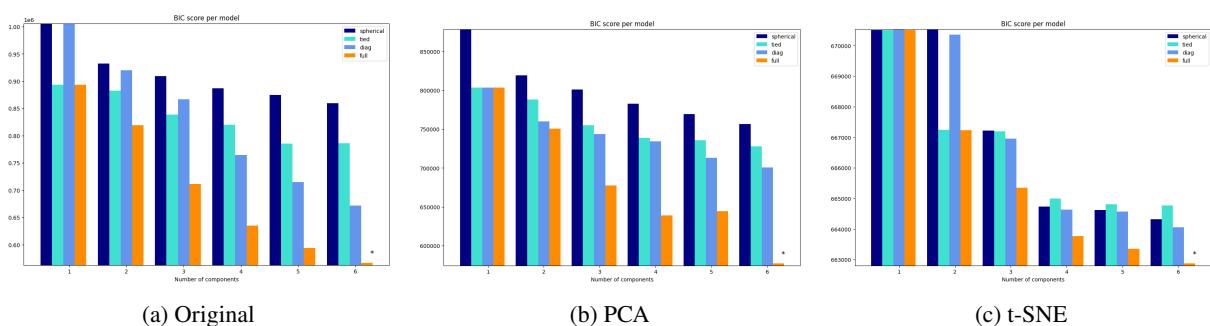
Lựa chọn số cum phù hợp

Để lựa chọn số cụm phù hợp, tiêu chí thông tin Bayes (Bayesian Information Criterion – BIC) được sử dụng nhằm đánh giá chất lượng phân cụm. Với mỗi dạng dữ liệu, các mô hình GMM được huấn luyện trên các số cụm từ 1 đến 6, kết hợp với bốn loại ma trận hiệp phương sai: spherical, tied, diag, và full. Mỗi tổ hợp (số cụm, loại hiệp phương sai) cho ra một mô hình, và mô hình có giá trị BIC nhỏ nhất sẽ được chọn. Quá trình phân cụm được thực hiện trên ba dạng dữ liệu: dữ liệu gốc đã chuẩn hóa (Original), dữ liệu sau khi giảm chiều bằng PCA (10 thành phần chính) và t-SNE (2 chiều).

Kết quả được minh họa bằng biểu đồ BIC thể hiện sự biến thiên theo số cụm và loại hiệp phương sai:

- Trục hoành biểu diễn số cụm.
 - Trục tung biểu diễn giá trị BIC (càng thấp càng tốt).
 - Mỗi màu sắc đại diện cho một loại cấu trúc hiệp phương sai.
 - Đầu “*” đánh dấu mô hình có BIC thấp nhất – chính là mô hình được lựa chọn.

Biểu đồ BIC theo số cụm cho từng trường hợp:



Hình 18: Biểu đồ BIC theo số cụm và loại ma trận hiệp phương sai cho từng kiểu dữ liệu: Original, PCA và t-SNE

Kết quả số cum tối ưu và đánh giá

Số cụm phù hợp được chọn dựa trên tiêu chí BIC thấp nhất, kết quả như sau:

Bảng 6: Giá trị BIC và số cụm tối ưu cho mô hình GMM áp dụng trên ba dạng dữ liệu đầu vào.

Loại dữ liệu	Số cụm tối ưu	Giá trị BIC
Dữ liệu gốc (chuẩn hóa)	6	567002.8191
PCA (10 thành phần)	6	577498.0997
t-SNE (2 chiều)	6	662883.9523

Kết quả cho thấy dữ liệu gốc (đã chuẩn hóa) mang lại hiệu quả phân cụm tốt nhất với giá trị BIC thấp nhất (567002.8191), tương ứng số cụm là 6.

4.2.2 Đánh giá mối quan hệ giữa các mẫu dữ liệu đầu vào và đầu ra trong từng cụm

Đánh giá chất lượng phân cụm qua các chỉ số định lượng

Hai chỉ số **Davies-Bouldin Index (DBI)** và **Calinski-Harabasz Index (CHI)** được sử dụng để đánh giá mức độ phân tách và độ đồng nhất của các cụm tương tự phương pháp KMeans, trong đó DBI càng thấp và CHI càng cao chứng tỏ cụm càng tách biệt và rõ ràng.

Kết quả đánh giá:

Bảng 7: So sánh chất lượng phân cụm GMM trên ba loại dữ liệu dựa theo các chỉ số DBI và CHI

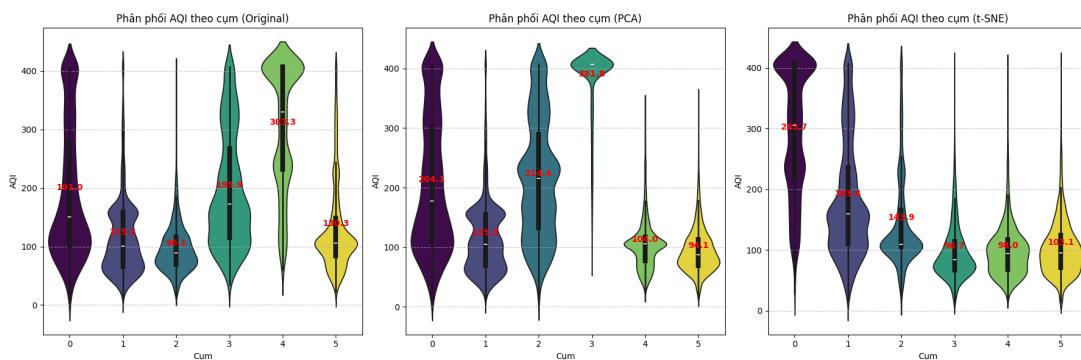
Phương pháp	Davies-Bouldin Index	Calinski-Harabasz Index
Chuẩn hóa (ORIG)	2.6908	2591.6972
PCA (10 thành phần)	2.2503	3188.8180
t-SNE (2 chiều)	0.9094	22287.9004

Chú thích:

- Chỉ số Davies-Bouldin (DB) càng nhỏ \Rightarrow chất lượng phân cụm càng tốt.
- Chỉ số Calinski-Harabasz (CH) càng lớn \Rightarrow cụm càng tách biệt rõ ràng.

Kết quả cho thấy phân cụm GMM trên dữ liệu t-SNE đạt chỉ số Davies-Bouldin thấp nhất và chỉ số Calinski-Harabasz cao nhất, chứng tỏ các cụm được phân tách rõ ràng và đồng nhất hơn so với dữ liệu Original và PCA.

Phân phối chỉ số AQI trong các cụm



Hình 19: Biểu đồ violin thể hiện phân phối chỉ số AQI theo các cụm được phân loại bằng GMM trên ba loại dữ liệu (Original, PCA, t-SNE). Giá trị trung bình AQI của từng cụm được đánh dấu phía trên bằng chữ đỏ.

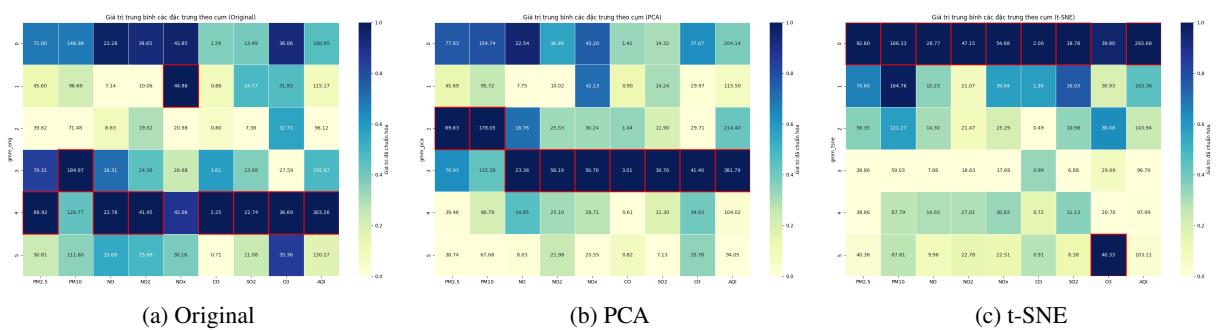
Nhận xét:

- Với dữ liệu *Original*, AQI trung bình dao động từ 96.1 đến 303.3, cho thấy sự phân bố khá đồng đều nhưng nổi bật với một số giá trị cao bất thường, phản ánh tính biến động lớn của dữ liệu gốc.
- Với *PCA*, AQI trung bình nằm trong khoảng 94.1 đến 381.8, phân bố đồng đều hơn so với dữ liệu gốc, trong đó cụm 3 ghi nhận giá trị cao nhất (381.8), cho thấy PCA giúp làm mượt dữ liệu và mở rộng phạm vi giá trị hiệu quả.

- Với t-SNE, AQI trung bình dao động từ 96.7 đến 292.7, tập trung nhiều ở mức thấp và trung bình, với cụm 2 đạt giá trị cao nhất (292.7), thể hiện sự phân tách rõ ràng hơn ở các mức AQI không quá cao.

Tóm lại: PCA hỗ trợ phân bố AQI cân bằng và mở rộng phạm vi giá trị từ 94.1 đến 381.8, làm mượt dữ liệu hiệu quả hơn, trong khi t-SNE nổi bật ở khả năng phân tách các cụm có AQI thấp và trung bình (lên đến 292.7), phù hợp cho phân tích chi tiết ở các mức giá trị này.

Mối quan hệ giữa các đặc trưng đầu vào trong từng cụm



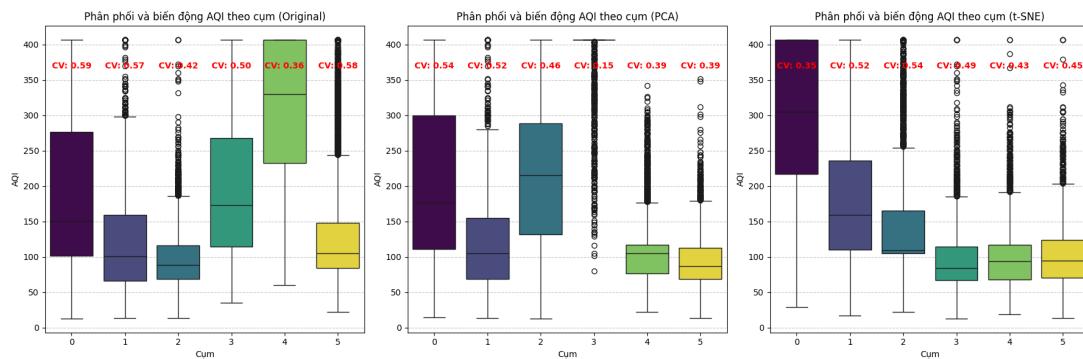
Hình 20: Heatmap thể hiện giá trị trung bình của các đặc trưng ô nhiễm trong từng cụm GMM trên ba loại dữ liệu. Các ô được khoanh đỏ biểu thị đặc trưng có giá trị trung bình cao nhất trong cụm tương ứng.

Nhận xét:

- Sự tương quan giữa các đặc trưng:** Các đặc trưng như PM2.5, PM10 và AQI thường có xu hướng tương quan thuận với nhau, thể hiện qua việc chúng cùng đạt giá trị cao hoặc thấp trong cùng một cụm. Điều này phản ánh mối liên hệ chặt chẽ giữa nồng độ bụi mịn và chất lượng không khí tổng thể. Tương tự, các chỉ số NO, NO₂ và NO_x cũng thường đi cùng nhau trong các cụm, do có nguồn phát thải tương đồng (giao thông, công nghiệp).
- Đặc trưng nổi bật trong từng cụm:** Cụm có AQI cao (ví dụ: cụm 4 trong dữ liệu *Original* với AQI = 303.26) thường có PM2.5, PM10 và SO₂ cao, cho thấy ảnh hưởng của bụi mịn và khí thải công nghiệp. Ngược lại, cụm có AQI thấp (ví dụ: cụm 3 trong dữ liệu *t-SNE* với AOI = 96.70) đi kèm với giá trị thấp của hầu hết các đặc trưng, phản ánh chất lượng không khí tốt.
- Khác biệt giữa các phương pháp giảm chiều:** Dữ liệu *Original* và *PCA* cho thấy sự phân cụm rõ ràng hơn so với *t-SNE*, đặc biệt trong việc làm nổi bật các cụm có đặc trưng ô nhiễm cao. Điều này có thể do PCA bảo toàn phương sai tổng thể tốt hơn, trong khi *t-SNE* tập trung vào cấu trúc cục bộ. Một số cụm trong *t-SNE* có xu hướng phân tán hơn, dẫn đến sự khác biệt về giá trị trung bình không rõ rệt.

Đánh giá mức độ biến động AQI trong từng cụm

Độ ổn định của chỉ số AQI trong mỗi cụm được đánh giá bằng hệ số biến thiên (CV = độ lệch chuẩn / giá trị trung bình).



Hình 21: Biểu đồ hộp (boxplot) minh họa sự phân bố AQI trong các cụm được phân cụm bằng GMM trên dữ liệu Original, PCA và t-SNE. Các hệ số biến động (CV) được chú thích bằng chữ đỏ trên các boxplot.

Nhận xét:

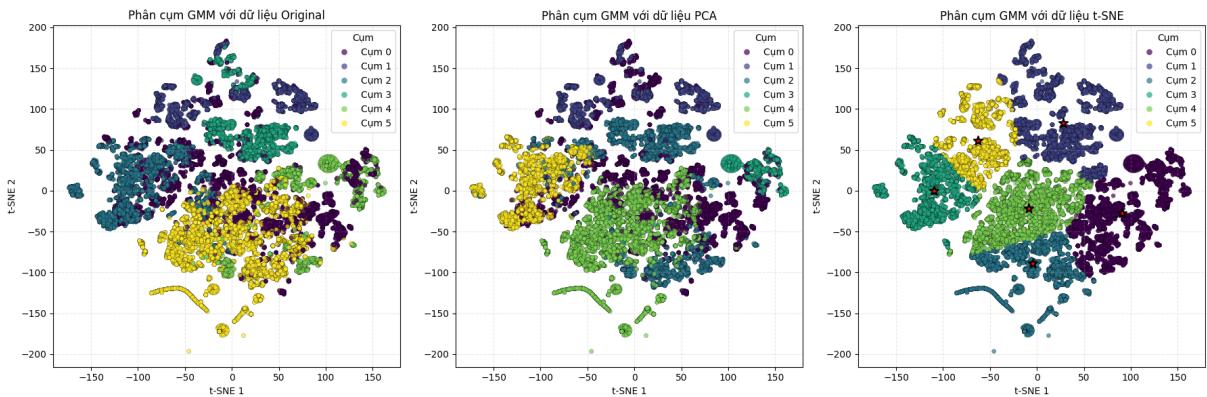
- Dữ liệu Original:** Các cụm có CV dao động từ 0.36 đến 0.59, cho thấy mức độ biến động tương đối cao. Cụm 4 có độ biến động thấp nhất ($CV = 0.36$) – là cụm có AQI cao trung bình nhất, phản ánh tình trạng ô nhiễm ổn định ở mức cao. Ngược lại, các cụm như 0, 1 và 5 có CV cao hơn (khoảng 0.57–0.59), thể hiện sự chênh lệch lớn trong chất lượng không khí nội tại mỗi cụm.
- Dữ liệu PCA:** Biến động AQI được thu hẹp hơn, đặc biệt nổi bật là cụm 3 với CV rất thấp (0.15), cho thấy AQI trong cụm này gần như ổn định. Các cụm còn lại có CV từ 0.39 đến 0.54, thấp hơn so với dữ liệu Original, phản ánh hiệu quả của PCA trong việc gom cụm theo phương sai tổng thể.
- Dữ liệu t-SNE:** Mức độ biến động phân bố khá đồng đều, CV dao động trong khoảng 0.35 đến 0.54. Không có cụm nào có CV quá thấp hoặc quá cao, cho thấy t-SNE tạo ra các cụm có phân bố AQI tương đối ổn định. Tuy nhiên, so với PCA, t-SNE không tạo ra sự khác biệt rõ rệt giữa các cụm về mức độ biến động AQI.

Tổng thể, việc đánh giá CV giúp xác định cụm nào đại diện cho vùng ô nhiễm ổn định và cụm nào có chất lượng không khí biến động nhiều. PCA thể hiện ưu thế trong việc cô lập cụm có mức biến động thấp nhất, trong khi t-SNE tạo ra các cụm đồng đều hơn về phương sai.

4.2.3 Trực quan hóa kết quả phân cụm

So sánh phân cụm trên không gian t-SNE

Hình 22 minh họa kết quả phân cụm của từng phương pháp. Các điểm dữ liệu được tô màu theo cụm tương ứng, trong khi tâm cụm (cluster centers) được đánh dấu bằng dấu sao đỏ. So sánh này giúp đánh giá mức độ phân tách giữa các cụm khi áp dụng các phương pháp biến đổi dữ liệu khác nhau.



Hình 22: So sánh kết quả phân cụm bằng GMM trên ba dạng biểu diễn dữ liệu: Original, PCA và t-SNE. Các điểm dữ liệu được hiển thị trong không gian t-SNE 2D để thuận tiện cho trực quan hóa. Các tâm cụm được đánh dấu bằng dấu sao đỏ.

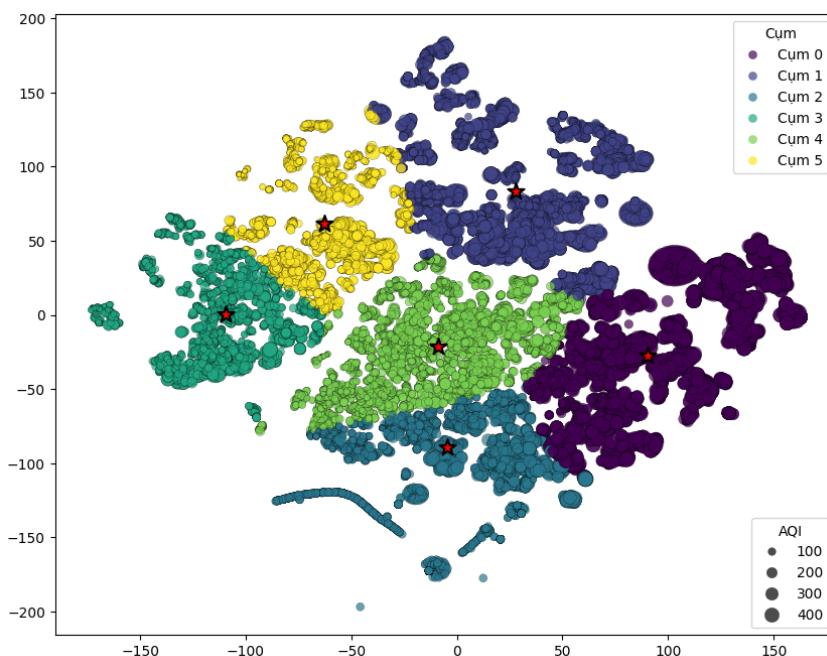
Nhận xét:

- **Original:** Các cụm phân bố khá chồng chéo và rải rác. Dễ thấy các cụm như 0, 1 và 5 có sự giao thoa lớn, thể hiện khả năng phân tách không rõ ràng của không gian gốc. Các cụm không thực sự tách biệt, điều này có thể ảnh hưởng đến hiệu quả của mô hình phân loại hoặc dự đoán sau này.
- **PCA:** Các cụm bắt đầu được phân biệt tốt hơn. Một số cụm như cụm 3 và cụm 4 đã có xu hướng tách ra tương đối rõ ràng, tuy nhiên vẫn tồn tại hiện tượng chồng lấn giữa một số cụm như 0 và 1. Việc áp dụng PCA giúp giảm nhiễu và giữ lại phương sai chính, từ đó cải thiện khả năng phân cụm của GMM so với dữ liệu gốc.
- **t-SNE:** Các cụm được phân tách rõ ràng và có ranh giới tương đối tách biệt. Không gian t-SNE giúp các cụm có cấu trúc gần nhau trong không gian cao chiều được biểu diễn gần nhau trong không gian 2 chiều, đồng thời đẩy các cụm khác nhau ra xa. Đây là biểu diễn trực quan hỗ trợ mạnh cho việc đánh giá chất lượng phân cụm.

Nhìn chung, phân cụm trong không gian t-SNE cho thấy sự tách biệt rõ ràng hơn cả, hỗ trợ tốt cho việc trực quan hóa và đánh giá kết quả. Tuy nhiên, vì t-SNE là phương pháp phi tuyến không bảo toàn cấu trúc toàn cục nên chỉ nên dùng cho mục đích minh họa trực quan, không nên áp dụng trực tiếp để huấn luyện mô hình học máy.

Biểu diễn AQI theo kích thước điểm trong không gian t-SNE

Hình 23 trực quan hóa kết quả phân cụm GMM trên không gian t-SNE 2D, trong đó kích thước điểm (marker size) biểu thị giá trị AQI của từng mẫu. Cách thể hiện này giúp quan sát sự phân bố AQI theo từng cụm, từ đó đánh giá mức độ đồng nhất hoặc khác biệt giữa các nhóm trong các lớp AQI. Các điểm lớn tương ứng với giá trị AQI cao, phản ánh mức độ ô nhiễm nặng, trong khi các điểm nhỏ thể hiện chất lượng không khí tốt với AQI thấp. Tâm cụm được đánh dấu bằng dấu sao đỏ để dễ nhận diện vị trí trung tâm của từng cụm.



Hình 23: Trực quan phân cụm GMM trên không gian t-SNE 2D với kích thước điểm biểu thị giá trị AQI tương ứng. Các tâm cụm được đánh dấu bằng dấu sao đỏ.

Nhận xét:

- Các cụm thể hiện sự khác biệt đáng kể về mức AQI trung bình. Cụm 0 (màu tím đậm bên phải) có nhiều điểm có kích thước lớn, cho thấy giá trị AQI cao, biểu hiện mức độ ô nhiễm không khí nghiêm trọng hơn.
- Ngược lại, cụm 3 (màu xanh ngọc, phía dưới bên trái) và cụm 4 (màu xanh lá cây, trung tâm) chứa nhiều điểm nhỏ hơn, cho thấy chất lượng không khí tại các điểm này tốt hơn (AQI thấp).
- Một số cụm như cụm 1 và cụm 5 có phân bố AQI tương đối đồng đều, với kích thước điểm dao động trong khoảng trung bình.

t-SNE giúp làm nổi bật các vùng dữ liệu có mức AQI tương đồng, hỗ trợ đánh giá mối liên hệ giữa cụm và mức độ ô nhiễm. Việc mã hóa AQI bằng kích thước điểm cho phép nhận diện rõ các vùng ô nhiễm cao. Cụm 0 và 1 thường có AQI cao, trong khi cụm 3 và 4 chủ yếu gồm các điểm có AQI thấp, cho thấy sự phân hóa rõ rệt.

4.3 So sánh hai phương pháp

4.3.1 So sánh chất lượng cụm theo độ đo định lượng

Chất lượng phân cụm của K-Means và GMM được đánh giá qua hai chỉ số: Davies-Bouldin Index (DBI – càng thấp càng tốt) và Calinski-Harabasz Index (CHI – càng cao càng tốt). Bảng dưới đây trình bày kết quả trên ba loại dữ liệu: chuẩn hóa (ORIG), PCA (10 thành phần) và t-SNE (2 chiều).

Bảng 8: So sánh DBI và CHI giữa K-Means và GMM

Phương pháp	Dữ liệu	Số cụm	DBI ↓	CHI ↑
K-Means	ORIG	2	1.8565	7511.38
	PCA (10)	2	1.8004	7949.71
	t-SNE (2D)	4	0.8239	26069.27
GMM	ORIG	6	2.6908	2591.70
	PCA (10)	6	2.2503	3188.82
	t-SNE (2D)	6	0.9094	22287.90

Nhận xét:

- **K-Means** đạt chất lượng tốt nhất trên t-SNE (DBI thấp, CHI cao), thể hiện cụm tách biệt và đồng nhất.
- **GMM** hoạt động tốt hơn trên PCA nhưng kém hơn K-Means trên t-SNE, dù có lợi thế mô hình hóa phân phối phức tạp.

4.3.2 So sánh phân phối AQI trong các cụm

Biểu đồ violin plot phản ánh phân phối chỉ số AQI trong từng cụm:

- **K-Means:**

- Với ORIG và PCA: chỉ tạo 2 cụm, AQI dao động lớn.
- Với t-SNE: tạo 4 cụm, AQI trung bình giảm dần (267.4 → 100.4), phân biệt rõ các mức độ ô nhiễm.

- **GMM:**

- Với ORIG và PCA: phân phối AQI rộng (96.1–381.8), thể hiện tính biến động lớn.
- Với t-SNE: AQI phân bố tập trung ở mức thấp–trung bình, nhưng kém sắc nét hơn K-Means.

Kết luận: K-Means trên t-SNE tách biệt tốt các mức AQI; GMM trên PCA phù hợp để nhận diện vùng ô nhiễm cao.

4.3.3 So sánh đặc trưng trong các cụm

Thông qua heatmap giá trị trung bình các đặc trưng trong cụm:

- **K-Means:**

- Với ORIG, PCA: cụm tương đối đồng nhất.
- Với t-SNE: rõ sự khác biệt về PM2.5, PM10, SO₂, v.v.

- **GMM:**

- Với ORIG, PCA: làm nổi bật cụm có nồng độ ô nhiễm cao.
- Với t-SNE: các cụm phân tán, khác biệt không rõ như K-Means.

Kết luận: K-Means trên t-SNE tạo cụm đặc trưng rõ; GMM trên PCA giúp phân tích các cụm ô nhiễm cao.

4.3.4 So sánh độ ổn định AQI theo hệ số biến thiên (CV)

- K-Means:

- Trên t-SNE: CV thấp, AQI ổn định.
- Trên ORIG, PCA: CV cao hơn, chỉ số AQI không đồng đều.

- GMM:

- ORIG: CV dao động 0.36–0.59, cụm ô nhiễm cao ổn định hơn.
- PCA: CV rất thấp ở cụm 3 (0.15), thể hiện sự mượt hoá bởi PCA.
- t-SNE: CV đồng đều (0.35–0.54), nhưng không vượt trội như PCA.

Kết luận: GMM trên PCA cho AQI ổn định nhất; K-Means trên t-SNE ổn định hơn ở vùng ô nhiễm thấp.

4.3.5 So sánh trực quan hóa cụm

- K-Means:

- Trên t-SNE: cụm ranh giới sắc nét, AQI phân hoá tốt.
- Trên ORIG, PCA: cụm chồng lấn nhiều.

- GMM:

- Trên t-SNE: cụm tách biệt nhưng một số vùng chồng chéo.
- Trên PCA: cải thiện so với ORIG, nhưng kém rõ hơn t-SNE.

Kết luận: K-Means trực quan hóa tốt hơn trên t-SNE, hỗ trợ phân tích trực quan hiệu quả hơn.

4.3.6 Kết luận

K-Means:

- **Ưu điểm:** Tối ưu trên t-SNE, cụm rõ ràng, hiệu quả với AQI thấp/trung bình.
- **Nhược điểm:** Ít cụm hơn trên ORIG và PCA, kém phân biệt.

GMM:

- **Ưu điểm:** Linh hoạt với số cụm lớn, mô hình Gaussian phù hợp dữ liệu phức tạp.
- **Nhược điểm:** Trên t-SNE kém hơn K-Means, cụm chồng lấn.

Tóm lại:

- **K-Means + t-SNE:** phù hợp cho phân tích chi tiết, trực quan mức AQI thấp–trung bình.
- **GMM + PCA:** phù hợp hơn với dữ liệu ô nhiễm cao, có tính phức tạp.

Chương 5

Thực nghiệm và kết quả

5.1 Mô hình Random Forest

5.1.1 Thiết lập và cấu hình mô hình

Lựa chọn và giải thích các siêu tham số chính

Việc lựa chọn các siêu tham số cho Random Forest là một bước quan trọng để tối ưu hóa hiệu suất của mô hình.

- **n_estimators:** Số lượng cây quyết định trong rừng. Giá trị lớn hơn thường dẫn đến hiệu suất tốt hơn nhưng cũng làm tăng thời gian tính toán. Giá trị này được thiết lập là 50, 100 và 200 để xem xét nhiều trường hợp cây
- **max_depth:** Độ sâu tối đa của mỗi cây quyết định. Nếu không được chỉ định (hoặc đặt là None), các nút sẽ được mở rộng cho đến khi tất cả các lá đều “thuần” (pure) hoặc cho đến khi tất cả các lá chứa ít hơn **min_samples_split** mẫu. Việc giới hạn độ sâu giúp kiểm soát độ phức tạp của mô hình và giảm nguy cơ quá khớp. Ở đây ta thử với 3 trường hợp là none, 5 và 10.
- **min_samples_split:** Số lượng mẫu tối thiểu cần thiết để tách một nút nội bộ. Nếu một nút có số lượng mẫu nhỏ hơn giá trị này, nó sẽ không được tách nữa và trở thành một nút lá. Giá trị mặc định thường là 2.
- **min_samples_leaf:** Số lượng mẫu tối thiểu cần thiết ở một nút lá. Điều này đảm bảo rằng mỗi lá có một số lượng mẫu đáng kể, giúp làm mượt mô hình, đặc biệt là trong hồi quy. Giá trị mặc định là 1.
- **random_state:** Được sử dụng để kiểm soát tính ngẫu nhiên trong quá trình xây dựng cây (ví dụ, chọn mẫu **bootstrap**, chọn đặc trưng tại mỗi lần tách). Việc đặt một giá trị cụ thể (ví dụ: 42 đảm bảo rằng kết quả có thể được tái tạo).

Khởi tạo đối tượng RandomForestRegressor trong thư viện (scikit-learn)

Đối tượng **RandomForestRegressor** được khởi tạo từ mô-đun `sklearn.ensemble` của thư viện scikit-learn. Quá trình khởi tạo bao gồm việc truyền các giá trị siêu tham số đã chọn vào hàm khởi tạo của lớp.

5.1.2 Quá trình huấn luyện mô hình

Sau khi mô hình đã được thiết lập và cấu hình, bước tiếp theo là huấn luyện mô hình bằng cách sử dụng dữ liệu đã chuẩn bị.

Chia tập huấn luyện và tập xác thực (train/validation split)

- Trước khi huấn luyện, dữ liệu thường được chia thành hai tập con: **tập huấn luyện** (training set) và **tập xác thực** (validation set). Mô hình sẽ học từ tập huấn luyện và sau đó được đánh giá trên tập xác thực để đo lường khả năng tổng quát hóa của nó trên dữ liệu mới, từ đó điều chỉnh các siêu tham số một cách phù hợp.
- Thư viện `scikit-learn` cung cấp hàm `train_test_split` từ module `sklearn.model_selection` để thực hiện việc này. Trong quá trình thử nghiệm, có thể sử dụng các tỉ lệ chia như sau:
 - 80% huấn luyện và 20% xác thực (tỉ lệ 8:2)
 - + Kích thước tập huấn luyện: (23624, 12)
 - + Kích thước tập validation: (5907, 12)
 - 70% huấn luyện và 30% xác thực (tỉ lệ 7:3)
 - + Kích thước tập huấn luyện: (20671, 12)
 - + Kích thước tập validation: (8860, 12)
 - 60% huấn luyện và 40% xác thực (tỉ lệ 6:4)
 - + Kích thước tập huấn luyện: (17718, 12)
 - + Kích thước tập validation: (11813, 12)

Gọi phương thức `.fit()` để huấn luyện

Sau khi có tập huấn luyện (`X_train, y_train`), mô hình Random Forest Regressor được huấn luyện bằng cách gọi phương thức `.fit()`. Phương thức này nhận đầu vào là ma trận đặc trưng của tập huấn luyện và vector mục tiêu tương ứng.

5.1.3 Thực nghiệm mô hình Random Forest với dữ liệu khác nhau

Dữ liệu	Split	Cây	MSE	RMSE	R2	MAE
Gốc	8:2	50	896.7948	29.9465	0.9174	17.5425
		100	892.3672	29.8725	0.9178	17.4586
		200	890.1362	29.8351	0.9180	17.4293
	7:3	50	900.3161	30.0053	0.9176	17.6563
		100	893.1509	29.8856	0.9183	17.5516
		200	887.3809	29.7889	0.9188	17.4954
	6:4	50	910.3369	30.1718	0.9168	17.7527
		100	903.9877	30.0664	0.9174	17.6793
		200	898.0534	29.9675	0.9180	17.6019
PCA	8:2	50	1034.1755	32.1586	0.9047	19.2811
		100	1023.5769	31.9934	0.9057	19.1365
		200	1016.5509	31.8834	0.9064	19.0404
	7:3	50	1002.0973	31.6559	0.9083	19.1856
		100	988.4272	31.4393	0.9096	19.0530
		200	982.3533	31.3425	0.9101	19.0027
	6:4	50	1028.2536	32.0664	0.9061	19.5285
		100	1016.8471	31.8880	0.9071	19.4024
		200	1009.0498	31.7655	0.9078	19.3373
t-SNE	8:2	50	1222.0471	34.9578	0.8874	20.7382
		100	1207.6390	34.7511	0.8888	20.6511
		200	1207.1450	34.7440	0.8888	20.6579
	7:3	50	1183.5152	34.4023	0.8917	20.4775
		100	1179.4381	34.3429	0.8921	20.4195
		200	1177.6691	34.3172	0.8923	20.3995
	6:4	50	1192.7864	34.5367	0.8910	20.6712
		100	1181.3736	34.3711	0.8921	20.5664
		200	1182.5431	34.3881	0.8920	20.5479

Bảng 9: Hiệu suất của Random Forest với ba phương pháp tiền xử lý (gốc, PCA, t-SNE) ở các tỉ lệ train:validation khác nhau

Bảng kết quả cho thấy sự ảnh hưởng của ba yếu tố chính đến hiệu suất mô hình Random Forest: **phương pháp tiền xử lý dữ liệu** (Gốc, PCA, t-SNE), **tỷ lệ chia dữ liệu** (Split: 8:2, 7:3, 6:4), và **số lượng cây** trong rừng (50, 100, 200). Hiệu suất được đo lường bằng các chỉ số: MSE (Mean Squared Error), RMSE (Root Mean Squared Error), R2 (R-squared), và MAE (Mean Absolute Error). Nhìn chung, giá trị MSE, RMSE, MAE càng thấp và R2 càng gần 1 thì mô hình càng tốt.

1. Ảnh hưởng của Phương pháp Tiền xử lý Dữ liệu

Dữ liệu Gốc:

- **Luôn cho kết quả tốt nhất** trên tất cả các chỉ số (MSE, RMSE, MAE thấp nhất; R2 cao nhất) so với PCA và t-SNE ở mọi cấu hình số cây và tỷ lệ chia.
- Điều này cho thấy các đặc trưng ban đầu của dữ liệu đã chứa đựng thông tin hữu ích và phù hợp cho mô hình Random Forest. Việc áp dụng các kỹ thuật giảm chiều như PCA và t-SNE trong trường hợp này không những không cải thiện mà còn làm suy giảm hiệu suất.

Dữ liệu PCA:

- Hiệu suất **kém hơn đáng kể** so với dữ liệu gốc nhưng **tốt hơn nhiều** so với t-SNE.

- Ví dụ, với split 8:2 và 200 cây, MSE của PCA là 1016.55 so với 890.13 của dữ liệu gốc. R2 cũng giảm từ ~0.918 xuống ~0.906.
- Điều này ngụ ý rằng quá trình giảm chiều bằng PCA có thể đã loại bỏ một số thông tin quan trọng hoặc các thành phần chính được giữ lại chưa tối ưu cho việc dự đoán của Random Forest.

Dữ liệu t-SNE:

- Cho **kết quả kém nhất** trong ba phương pháp. Các giá trị lỗi (MSE, RMSE, MAE) đều cao nhất và R2 thấp nhất.
- Ví dụ, với split 8:2 và 200 cây, MSE của t-SNE là 1207.14, cao hơn nhiều so với PCA (1016.55) và dữ liệu gốc (890.13). R2 chỉ đạt khoảng ~0.888.
- t-SNE chủ yếu là một kỹ thuật giảm chiều để trực quan hóa, tập trung vào việc bảo toàn cấu trúc lân cận cục bộ. Điều này có thể không phù hợp cho các tác vụ hồi quy với Random Forest, nơi mà mối quan hệ tổng thể giữa các đặc trưng có thể quan trọng hơn.

2. Ảnh hưởng của Số lượng Cây

- **Xu hướng chung:** Trong hầu hết các trường hợp, việc **tăng số lượng cây** (từ 50 lên 100, rồi lên 200) **cải thiện nhẹ hiệu suất** của mô hình.
 - MSE, RMSE, MAE có xu hướng giảm.
 - R2 có xu hướng tăng.
- **Ví dụ (Dữ liệu gốc, Split 8:2):**
 - 50 cây: MSE 896.79, R2 0.9174
 - 100 cây: MSE 892.36, R2 0.9178
 - 200 cây: MSE 890.13, R2 0.9180
- **Nhận xét:** Điều này phù hợp với lý thuyết của Random Forest, nơi việc tăng số lượng cây giúp giảm phương sai của mô hình và làm cho dự đoán ổn định hơn. Tuy nhiên, mức độ cải thiện giảm dần khi số cây đã tương đối lớn (ví dụ, sự cải thiện từ 100 lên 200 cây thường ít hơn so với từ 50 lên 100 cây).

3. Ảnh hưởng của Tỷ lệ Chia Dữ liệu

Không có một tỷ lệ chia nào hoàn toàn vượt trội trong mọi trường hợp, nhưng có thể thấy một số xu hướng:

- Với **dữ liệu Gốc và PCA**, tỷ lệ chia **7:3** thường cho kết quả tốt nhất hoặc rất cạnh tranh, đặc biệt khi sử dụng 200 cây.
 - Gốc (200 cây, 7:3): MSE 887.38, R2 0.9188 (tốt nhất tổng thể)
 - PCA (200 cây, 7:3): MSE 982.35, R2 0.9101 (tốt nhất cho PCA)

- Với **dữ liệu t-SNE**, tỷ lệ chia **7:3** cũng thường cho kết quả tốt hơn so với 8:2 và 6:4.
 - t-SNE (200 cây, 7:3): MSE 1177.66, R2 0.8923 (tốt nhất cho t-SNE)
- Tỷ lệ **6:4** (ít dữ liệu huấn luyện hơn) đôi khi cho kết quả kém hơn một chút, đặc biệt với dữ liệu gốc. Điều này có thể là do mô hình không có đủ dữ liệu để học hiệu quả.

4. Kết luận

1. **Dữ liệu gốc là lựa chọn tốt nhất:** Với bộ dữ liệu này, mô hình Random Forest hoạt động hiệu quả nhất khi sử dụng trực tiếp dữ liệu gốc mà không qua các bước tiền xử lý giảm chiều như PCA hay t-SNE. Các kỹ thuật này có thể đã làm mất thông tin quan trọng.
2. **Số lượng cây nhiều hơn thường tốt hơn:** Tăng số lượng cây (ví dụ lên 200) giúp cải thiện hiệu suất. Có thể xem xét thử nghiệm với số cây lớn hơn nữa (ví dụ 300, 500) để xem có cải thiện đáng kể hay không, đồng thời cân nhắc thời gian huấn luyện.
3. **Tỷ lệ chia 7:3 tỏ ra hiệu quả:** Tỷ lệ chia 70% dữ liệu cho huấn luyện và 30% cho kiểm định (validation) thường như là một lựa chọn cân bằng và mang lại hiệu suất tốt cho nhiều trường hợp.
4. **Cấu hình tốt nhất được quan sát:** Dựa trên bảng, cấu hình cho kết quả tốt nhất là:
 - **Dữ liệu:** Gốc
 - **Split:** 7:3
 - **Số cây:** 200
 - Với các chỉ số: MSE = 887.3809, RMSE = 29.7889, R2 = 0.9188, MAE = 17.4954.

Lưu ý:

- Kết quả này đặc thù cho bộ dữ liệu và mô hình Random Forest cụ thể.
- Việc tinh chỉnh các siêu tham số khác của Random Forest (như `max_depth`, `min_samples_split`) có thể mang lại những cải thiện thêm.
- Đối với PCA, số lượng thành phần chính được giữ lại không được nêu rõ, điều này cũng ảnh hưởng lớn đến kết quả của PCA.

5.1.4 Phân tích hiện tượng Overfitting

Overfitting trên dữ liệu không phân cụm theo tỉ lệ 7:3

Bảng 10: So sánh MSE và Overfit Ratio cho ba bộ dữ liệu (Original, PCA, t-SNE) với tỷ lệ Train:Test = 7:3

Dataset	MSE (Train)			MSE (Val)			Overfit Ratio		
	50	100	200	50	100	200	50	100	200
Original	137.6162	130.5773	127.4320	900.3161	893.1509	887.3809	6.5422	6.8400	6.9636
PCA	159.7102	150.5923	146.3994	1002.0973	988.4272	982.3533	6.2745	6.5636	6.7101
t-SNE	167.8875	160.3246	157.1639	1183.5152	1179.4381	1177.6691	7.0495	7.3566	7.4933

1. Dấu hiệu Chính của Overfitting:

- Chênh lệch lớn giữa MSE (Train) và MSE (Val):

- Trên tất cả các bộ dữ liệu và với mọi số lượng cây (50, 100, 200), giá trị **MSE (Train)** (Mean Squared Error trên tập huấn luyện) đều **thấp hơn đáng kể** so với **MSE (Val)** (Mean Squared Error trên tập kiểm định).
- Ví dụ, với bộ dữ liệu **Original** và 200 cây: MSE (Train) là 127.4320 trong khi MSE (Val) lên tới 887.3809.
- Điều này cho thấy mô hình hoạt động rất tốt trên dữ liệu nó đã "học thuộc" (tập train) nhưng lại kém hiệu quả khi dự đoán trên dữ liệu mới mà nó chưa từng thấy (tập val). Đây là đặc điểm điển hình của overfitting.

- Chỉ số Overfit Ratio cao:

- Overfit Ratio** được tính bằng $\frac{\text{MSE}(\text{Val})}{\text{MSE}(\text{Train})}$. Tỉ lệ này càng cao thì mức độ overfitting càng nghiêm trọng.
- Tất cả các giá trị Overfit Ratio trong bảng đều lớn hơn 1 rất nhiều (từ 6.2745 đến 7.4933), khẳng định mạnh mẽ tình trạng overfitting.
- Ví dụ, với bộ dữ liệu **t-SNE** và 200 cây, Overfit Ratio là 7.4933, nghĩa là lỗi trên tập validation cao gấp gần 7.5 lần lỗi trên tập training.

2. Ảnh hưởng của Số lượng Cây:

- MSE (Train):** Khi số lượng cây tăng từ 50 lên 200, MSE (Train) có xu hướng **giảm nhẹ** trên cả ba bộ dữ liệu. Điều này hợp lý vì mô hình Random Forest phức tạp hơn (nhiều cây hơn) có khả năng học kỹ hơn dữ liệu huấn luyện.
- MSE (Val):** MSE (Val) cũng có xu hướng **giảm nhẹ** khi số lượng cây tăng, nhưng mức giảm này không đáng kể bằng mức giảm của MSE (Train).

- **Overfit Ratio:** Điều thú vị là Overfit Ratio lại có xu hướng **tăng nhẹ** khi số lượng cây tăng trên cả ba bộ dữ liệu.
 - Ví dụ, trên bộ dữ liệu **Original**, Overfit Ratio tăng từ 6.5422 (50 cây) lên 6.9636 (200 cây).
 - Điều này gợi ý rằng việc tăng số lượng cây, mặc dù có thể cải thiện một chút hiệu suất trên tập validation, nhưng lại làm cho mô hình càng "khớp" chặt hơn với tập train, dẫn đến tỷ lệ overfitting cao hơn. Tuy nhiên, Random Forest thường khá kháng với overfitting khi tăng số cây so với các thuật toán khác, nhưng ở đây vẫn thấy một sự gia tăng nhẹ.

3. So sánh giữa các Bộ dữ liệu:

- **Original:** Cho thấy mức độ overfitting đáng kể.
- **PCA:**
 - Có MSE (Train) và MSE (Val) cao hơn so với bộ dữ liệu Original.
 - Tuy nhiên, Overfit Ratio của PCA (ví dụ: 6.7101 với 200 cây) lại **thấp hơn một chút** so với Original (6.9636 với 200 cây). Điều này có thể gợi ý rằng PCA, dù làm tăng lỗi tổng thể, nhưng lại giúp giảm nhẹ mức độ overfitting so với không dùng PCA.
- **t-SNE:**
 - Bộ dữ liệu này cho thấy **MSE (Val) cao nhất** và cũng có **Overfit Ratio cao nhất** trong cả ba bộ dữ liệu (ví dụ: 7.4933 với 200 cây).
 - Điều này cho thấy việc áp dụng t-SNE trong trường hợp này không những không cải thiện hiệu suất tổng thể mà còn làm cho mô hình bị overfitting nghiêm trọng hơn. Có thể các đặc trưng được tạo ra bởi t-SNE quá đặc thù cho tập huấn luyện.

4. Kết luận: Mô hình Random Forest đang gặp phải tình trạng **overfitting rõ rệt** trên tất cả các cấu hình thử nghiệm. Mô hình học quá tốt dữ liệu huấn luyện nhưng không khái quát hóa được trên dữ liệu mới.

- Việc **tăng số lượng cây** có vẻ làm tăng nhẹ tỷ lệ overfitting.
- Trong số các phương pháp tiền xử lý dữ liệu, **t-SNE** dẫn đến overfitting nghiêm trọng nhất.
- **PCA** có thể làm giảm nhẹ mức độ overfitting so với dữ liệu gốc, nhưng đi kèm với việc tăng lỗi dự đoán chung.

Để cải thiện, có thể xem xét các kỹ thuật như điều chỉnh siêu tham số của Random Forest (ví dụ: `max_depth`, `min_samples_split`, `min_samples_leaf`), sử dụng các kỹ thuật regularization khác, hoặc thu thập thêm dữ liệu nếu có thể.

Bảng 11: Kết quả huấn luyện và kiểm định (MSE) trên các loại tiền xử lý dữ liệu (bảng viền kín).

Dataset	max_depth	MSE (Train)	MSE (Val)	Overfit Ratio
Original	5	1140.9298	1215.2233	1.0651
	10	556.8662	940.2108	1.6884
	None	130.5773	893.1509	6.8400
PCA	5	1768.0525	1791.0925	1.0130
	10	689.3788	1145.6438	1.6618
	None	150.5923	988.4272	6.5636
t-SNE	5	2907.7389	2909.7022	1.0007
	10	1133.3090	1468.8688	1.2961
	None	160.3246	1179.4381	7.3566

Áp dụng giới hạn độ sâu để cải thiện Overfitting

1. Hiệu quả giảm Overfitting (Overfit Ratio):

- Khi không giới hạn độ sâu (`max_depth = None`), Overfit Ratio rất cao trên cả ba bộ dữ liệu (Original: 6.8400, PCA: 6.5636, t-SNE: 7.3566), cho thấy mô hình học quá khớp với dữ liệu huấn luyện.
- Giới hạn `max_depth = 10` đã **giảm đáng kể** Overfit Ratio (Original: 1.6884, PCA: 1.6618, t-SNE: 1.2961).
- Khi `max_depth = 5`, Overfit Ratio **tiếp tục giảm mạnh**, xuống rất gần 1 (Original: 1.0651, PCA: 1.0130, t-SNE: 1.0007). Điều này có nghĩa là lỗi trên tập huấn luyện và tập kiểm định gần như tương đương, cho thấy overfitting đã được kiểm soát rất tốt.

2. Ảnh hưởng đến Lỗi Huấn luyện (MSE Train):

- Việc giảm `max_depth` làm **tăng MSE (Train)**. Điều này là dự kiến vì mô hình đơn giản hơn (cây nông hơn) sẽ khó học thuộc lòng dữ liệu huấn luyện hơn.
- Ví dụ, trên bộ Original, MSE (Train) tăng từ 130.5773 (`max_depth = None`) lên 1140.9298 (`max_depth = 5`).

3. Ảnh hưởng đến Lỗi Kiểm định (MSE Val) - Điểm then chốt:

- Mặc dù Overfit Ratio giảm mạnh, **MSE (Val) lại có xu hướng tăng lên** khi giới hạn `max_depth`.
 - Trên bộ dữ liệu **Original**:
 - `max_depth = None`: MSE (Val) = 893.1509
 - `max_depth = 10`: MSE (Val) = 940.2108 (cao hơn)
 - `max_depth = 5`: MSE (Val) = 1215.2233 (cao hơn đáng kể)
 - Xu hướng tương tự cũng xảy ra với bộ dữ liệu **PCA** và **t-SNE**. Ví dụ, với t-SNE, MSE (Val) tăng từ 1179.4381 (`max_depth = None`) lên 2909.7022 (`max_depth = 5`).

4. Phân tích và Kết luận:

- Việc giới hạn độ sâu của cây (`max_depth`) là một phương pháp **hiệu quả để giảm hiện tượng overfitting**, thể hiện qua chỉ số Overfit Ratio giảm xuống gần 1. Tuy nhiên, trong trường hợp này, việc giảm overfitting quá mạnh (đặc biệt với `max_depth = 5`) đã dẫn đến việc **tăng lỗi trên tập kiểm định (MSE Val)**. Điều này cho thấy mô hình có thể trở nên quá đơn giản và rơi vào tình trạng **underfitting** (không học đủ các đặc điểm quan trọng của dữ liệu).
- Mặc dù mô hình với `max_depth = None` có Overfit Ratio cao, nó lại cho kết quả MSE (Val) tốt nhất (thấp nhất) trên cả ba bộ dữ liệu. Điều này gợi ý rằng sự phức tạp của mô hình khi không giới hạn độ sâu có thể đã giúp nó nắm bắt được những quy luật hữu ích trong dữ liệu, dù phải trả giá bằng việc học cả nhiễu.
- Nếu mục tiêu là giảm Overfit Ratio mà vẫn giữ hiệu năng chấp nhận được, `max_depth = 10` có vẻ là một sự cân bằng tốt hơn so với `max_depth = 5`. Tuy nhiên, cần cân nhắc giữa việc giảm Overfit Ratio và việc duy trì MSE (Val) ở mức thấp. Đôi khi, một mô hình có một chút overfitting nhưng hiệu suất dự đoán tổng thể tốt hơn có thể được ưu tiên hơn một mô hình ít overfitting nhưng dự đoán kém chính xác hơn trên dữ liệu mới.

5.1.5 Trực quan hóa và đánh giá tương quan phần dư

1. Phân tích thống kê phần dư

Data Type	Train:Test	Residual Mean	Residual Std	Residual Min	Residual Max
Original	8:2	-0.1733	29.8720	-271.5510	285.3883
PCA	8:2	0.1210	31.9932	-271.5510	290.9600
t-SNE	8:2	0.5115	34.7473	-272.0000	290.1411
Original	7:3	-0.3332	29.8838	-271.4573	287.4168
PCA	7:3	-0.0547	31.4392	-277.1235	287.8800
t-SNE	7:3	0.1609	34.3426	-272.0000	294.5992
Original	6:4	-0.6754	30.0588	-282.1778	298.3900
PCA	6:4	-0.3779	31.8858	-283.3400	289.6250
t-SNE	6:4	-0.1391	34.3708	-291.0000	294.5708

Bảng 12: Bảng kết quả tổng quan theo tỉ lệ train:test

Bảng "kết quả tổng quan theo tỉ lệ train:test" cung cấp các chỉ số thống kê về phần dư cho các loại dữ liệu (Original, PCA, t-SNE) và các tỷ lệ chia tập huấn luyện:kiểm tra khác nhau.

- **Trung bình Phần dư (Residual Mean):**

- Các giá trị trung bình của phần dư trong hầu hết các trường hợp đều **khá gần 0** (ví dụ: -0.1733 cho Original 8:2; 0.1210 cho PCA 8:2; -0.0547 cho PCA 7:3).
- **Ý nghĩa:** Trung bình phần dư gần 0 cho thấy mô hình không có xu hướng dự đoán cao hơn (overestimate) hay thấp hơn (underestimate) giá trị thực tế một cách có hệ thống. Đây là một đặc điểm của một mô hình không bị chệch (unbiased).

- **Độ lệch chuẩn Phần dư (Residual Std), Phần dư Nhỏ nhất (Residual Min) và Lớn nhất (Residual Max):**

- **Độ lệch chuẩn (Residual Std):** Dao động trong khoảng từ 29.87 đến 34.75.
- **Khoảng dao động của phần dư (Min/Max):** Rất lớn. Ví dụ, với dữ liệu Original và tỷ lệ 8:2, phần dư dao động từ -271.5510 đến 285.3883. Các trường hợp khác cũng cho thấy khoảng dao động rộng tương tự.
- **Ý nghĩa:** Mặc dù trung bình phần dư gần 0, độ lệch chuẩn tương đối lớn và đặc biệt là khoảng cách rất rộng giữa giá trị lỗi nhỏ nhất và lớn nhất cho thấy **sai số của mô hình trên các dự đoán cụ thể có thể rất đáng kể**. Điều này ngụ ý rằng trong khi mô hình có thể tốt về mặt trung bình, hiệu suất trên từng điểm dữ liệu riêng lẻ có thể không ổn định và có những trường hợp mô hình dự đoán sai lệch lớn so với thực tế.

2. Phân tích Tương quan giữa Phần dư và Đặc trưng



Hình 24: Tương quan giữa phần dư và các đặc trưng đầu vào

Biểu đồ heatmap "Tương quan giữa phần dư và đặc trưng (toute bộ dữ liệu gốc)" cho thấy các hệ số tương quan giữa phần dư của mô hình và các đặc trưng đầu vào (NH₃, Benzene, O₃, CO, NO, Xylene, Toluene, PM_{2.5}, NO₂, NOx, SO₂, PM10) đều **rất gần 0**. Ví dụ, tương quan với NH₃ là 0.0064, với Benzene là 0.0023, và với PM10 là -0.035.

- **Ý nghĩa:** Điều này là một dấu hiệu tốt, cho thấy phần dư (sai số dự đoán) không có mối quan hệ tuyến tính rõ ràng với bất kỳ đặc trưng nào. Nói cách khác, mô hình đường như đã nắm bắt được thông tin từ các đặc trưng này và không còn lỗi hệ thống nào liên quan đến chúng mà mô hình chưa giải thích được.

3. Nhận định chung

Mô hình **phù hợp ở mức độ cơ bản** nhưng **chưa thực sự tối ưu** về độ chính xác và độ tin cậy cho các dự đoán riêng lẻ.

- **Ưu điểm:** Mô hình không cho thấy lỗi thiên lệch hệ thống (trung bình phần dư gần 0) và phần dư không có tương quan tuyến tính với các đặc trưng đầu vào, cho thấy mô hình đã học được các mối quan hệ cơ bản trong dữ liệu.
- **Hạn chế:** Độ biến động của sai số (độ lệch chuẩn) và biên độ sai số (min/max) rất lớn, chỉ ra rằng mô hình có thể tạo ra các dự đoán với sai số đáng kể ở các trường hợp cụ thể.

Để cải thiện, cần xem xét các phương pháp nhằm giảm phương sai của lỗi dự đoán, có thể bao gồm việc tinh chỉnh thêm các tham số mô hình, sử dụng các thuật toán khác, hoặc kỹ thuật xử lý đặc trưng nâng cao hơn.

5.2 Mô hình MLP (Multilayer Perceptron)

5.2.1 Thiết lập và cấu hình mô hình

Lựa chọn và giải thích các siêu tham số chính

Việc lựa chọn các siêu tham số cho mô hình MLP là rất quan trọng để đạt được hiệu suất tối ưu. Các siêu tham số chính được sử dụng trong thực nghiệm bao gồm:

- **hidden_layer_sizes**: Xác định số neuron trong lớp ẩn. Thử nghiệm với 100, 75 và 50 neuron để khảo sát hiệu quả mô hình.
- **activation**: Dùng hàm kích hoạt `relu` nhờ khả năng học tốt và khắc phục hiện tượng mất gradient.
- **solver**: Sử dụng `adam`, một thuật toán tối ưu thích nghi hiệu quả, phổ biến trong huấn luyện mạng nơ-ron.
- **alpha**: Hệ số regularization L2 (0.0001) giúp giảm overfitting.
- **learning_rate**: Dùng kiểu `adaptive` để tự điều chỉnh khi mô hình không còn cải thiện.
- **learning_rate_init**: Tốc độ học ban đầu là 0.001 – giá trị mặc định phù hợp với Adam.
- **max_iter**: Giới hạn 500 vòng lặp để cân bằng giữa thời gian và khả năng hội tụ.
- **early_stopping**: Bật dừng sớm nếu không có cải thiện trên tập validation, giúp tránh overfitting.
- **random_state**: Thiết lập hạt giống 42 để đảm bảo kết quả có thể tái lập.

Khởi tạo đối tượng MLPRegressor trong thư viện (scikit-learn)

Đối tượng **MLPRegressor** được khởi tạo từ mô-đun `sklearn.neural_network` của thư viện scikit-learn. Quá trình khởi tạo bao gồm việc truyền các giá trị siêu tham số đã chọn vào hàm khởi tạo của lớp.

5.2.2 Quá trình huấn luyện mô hình

Sau khi mô hình `MLPRegressor` đã được thiết lập và cấu hình, bước tiếp theo là huấn luyện mô hình bằng cách sử dụng dữ liệu đã chuẩn bị.

Chia tập huấn luyện và tập xác thực (train/validation split)

Việc chia tập huấn luyện và xác thực được thực hiện tương tự như trong mô hình Random Forest.

Gọi phương thức `.fit()` để huấn luyện

Sau khi phân chia dữ liệu, mô hình `MLPRegressor` được huấn luyện bằng cách gọi phương thức `.fit()` trên tập huấn luyện (`X_train, y_train`). Phương thức này sẽ huấn luyện mạng neural nhiều lớp với số lượng neuron trong lớp ẩn được xác định trước (100, 75, 50), cùng các siêu tham số như hàm kích hoạt `relu`, tốc độ học ban đầu 0.001, số vòng lặp tối đa 500 và kích hoạt chế độ dừng sớm `early_stopping=True`.

Mô hình sau đó được đánh giá trên tập xác thực (X_{val} , y_{val}) thông qua các chỉ số như MSE, RMSE, R^2 và MAE, nhằm so sánh hiệu quả giữa các cấu hình khác nhau của mô hình MLP.

5.2.3 Thực nghiệm mô hình MLP với dữ liệu khác nhau

Bảng 13: Hiệu suất của MLP trên 3 loại dữ liệu (Gốc, PCA, t-SNE) với các tỉ lệ train:validation

Dữ liệu	Split	Hidden size	MSE	RMSE	R^2	MAE
Gốc	8:2	50	1174.5383	34.2715	0.8918	22.6244
		75	1116.8762	33.4197	0.8971	21.6696
		100	1100.8416	33.1789	0.8986	21.5505
	7:3	50	1191.7850	34.5222	0.8910	23.3985
		75	1096.4259	33.1123	0.8997	21.8996
		100	1104.7961	33.2385	0.8989	21.9691
	6:4	50	1231.1262	35.0874	0.8875	23.6628
		75	1157.7174	34.0252	0.8942	22.6721
		100	1103.3577	33.2168	0.8992	21.8300
PCA	8:2	50	1116.9583	33.4209	0.8971	21.3847
		75	1084.4043	32.9303	0.9001	20.8577
		100	1079.5648	32.8567	0.9006	20.7728
	7:3	50	1065.6990	32.6450	0.9025	20.9966
		75	1060.1471	32.5599	0.9030	20.8943
		100	1060.4338	32.5643	0.9030	20.9114
	6:4	50	1088.0748	32.9860	0.9006	21.3286
		75	1075.7949	32.7993	0.9017	21.0405
		100	1073.4318	32.7633	0.9019	21.1382
t-SNE	8:2	50	4053.4055	63.6664	0.6266	46.3661
		75	3814.2513	61.7596	0.6487	44.6810
		100	3727.0438	61.0495	0.6567	44.2544
	7:3	50	4087.8641	63.9364	0.6261	46.7758
		75	3887.4279	62.3492	0.6444	45.3975
		100	3903.2121	62.4757	0.6430	45.2820
	6:4	50	4186.0233	64.6995	0.6176	47.4012
		75	3962.4543	62.9480	0.6380	45.9481
		100	4406.2571	66.3796	0.5975	48.7203

1. Ảnh hưởng của phương pháp tiền xử lý dữ liệu

Dữ liệu Gốc:

- Mô hình đạt hiệu suất ổn định và cao, với các chỉ số R^2 dao động quanh 0.89–0.90 và MSE dưới 1235 trong hầu hết các cấu hình.
- Điều này cho thấy dữ liệu gốc đã chứa đủ thông tin cần thiết cho mô hình học hiệu quả, không cần thiết phải giảm chiều.

Dữ liệu PCA:

- PCA cho kết quả khá tương đương dữ liệu gốc trong nhiều trường hợp. Với tỉ lệ 7:3 và 75 nơ-ron, PCA đạt $MSE = 1060.15$, $R^2 = 0.9030$ – thậm chí nhỉnh hơn một chút so với dữ liệu gốc.
- Tuy nhiên, sự cải thiện là không đáng kể và không nhất quán ở mọi cấu hình. Điều này gợi ý rằng PCA có thể tăng tốc huấn luyện mà không làm giảm nhiều hiệu suất.

Dữ liệu t-SNE:

- t-SNE cho kết quả thấp nhất trong cả ba phương pháp, với R^2 chỉ quanh 0.62–0.65 và MSE luôn trên 3700.
- Nguyên nhân là do t-SNE vốn tối ưu cho trực quan hóa, bảo toàn cấu trúc cục bộ thay vì toàn cục – điều này không phù hợp cho mô hình hồi quy vốn cần đặc trưng toàn cục để dự đoán chính xác.

2. Ảnh hưởng của số lượng Nơ-ron ẩn

- **Xu hướng chung:** Tăng số lượng nơ-ron ẩn (50 lên 75, 100) thường cải thiện nhẹ hiệu suất MLP trên cả 3 loại dữ liệu (Gốc, PCA, t-SNE).
 - MSE, RMSE, MAE giảm dần; R^2 tăng nhẹ hoặc ổn định.
- **Ví dụ (Dữ liệu gốc, Split 8:2):**
 - 50 node: MSE 1174.5383, R^2 0.8918
 - 75 node: MSE 1116.8762, R^2 0.8971
 - 100 node: MSE 1100.8416, R^2 0.8986
- **Nhận xét:** Tăng node ẩn cải thiện khả năng biểu diễn, nhưng hiệu quả giảm dần (như với 100 node ở dữ liệu Gốc và t-SNE). Cần chọn số node hợp lý để tránh overfitting hoặc tăng chi phí tính toán.

3. Ảnh hưởng của tỷ lệ chia dữ liệu

Không có một tỷ lệ chia nào hoàn toàn vượt trội trong mọi trường hợp, nhưng có thể thấy một số xu hướng:

- Với **dữ liệu Gốc**, tỷ lệ chia **7:3** thường cho kết quả tốt nhất hoặc rất cạnh tranh, đặc biệt khi sử dụng kích thước ẩn 75.
 - Gốc (75, 7:3): MSE 1096.4259, R^2 0.8997 (tốt nhất cho dữ liệu Gốc)
 - Gốc (100, 6:4): MSE 1103.3577, R^2 0.8992 (rất cạnh tranh - kết quả gần như tốt nhất)
- Với **dữ liệu PCA**, tỷ lệ chia **7:3** cũng cho kết quả tốt nhất, đặc biệt khi sử dụng kích thước ẩn 75 hoặc 100.
 - PCA (75, 7:3): MSE 1060.1471, R^2 0.9030 (tốt nhất cho PCA)
 - PCA (100, 7:3): MSE 1060.4338, R^2 0.9030 (tương đương, kết quả gần như tốt nhất)
- Với **dữ liệu t-SNE**, tỷ lệ chia **8:2** thường cho kết quả tốt hơn so với 7:3 và 6:4, đặc biệt khi sử dụng kích thước ẩn 100.
 - t-SNE (100, 8:2): MSE 3727.0438, R^2 0.6567 (tốt nhất cho t-SNE)
- **Tỷ lệ 6:4** (ít dữ liệu huấn luyện hơn) thường cho kết quả kém hơn, đặc biệt với dữ liệu t-SNE, có thể do mô hình không có đủ dữ liệu để học hiệu quả.

4. Kết luận

1. **PCA là lựa chọn tốt nhất:** Với bộ dữ liệu này, mô hình MLP hoạt động hiệu quả nhất so với dữ liệu gốc và t-SNE. Dữ liệu t-SNE cho thấy hiệu suất kém hơn đáng kể, có thể do mất thông tin quan trọng trong quá trình giảm chiều.
2. **Kích thước lớp ẩn lớn hơn cải thiện hiệu suất:** Tăng kích thước lớp ẩn (từ 50 lên 75, 100) thường mang lại kết quả tốt hơn, đặc biệt với dữ liệu PCA, với các chỉ số như R^2 cao hơn và MSE thấp hơn.
3. **Tỷ lệ chia 7:3 là tối ưu:** mang lại hiệu suất tốt nhất trong nhiều trường hợp, đặc biệt với dữ liệu PCA, đạt được sự cân bằng giữa huấn luyện và đánh giá.
4. **Cấu hình tốt nhất được quan sát:** Dựa trên bảng, cấu hình cho kết quả tốt nhất là:
 - **Dữ liệu:** PCA
 - **Split:** 7:3
 - **Kích thước lớp ẩn:** 75
 - VỚI CÁC CHỈ SỐ: MSE = 1060.1471, RMSE = 32.5599, R^2 = 0.9030, MAE = 20.8943.

5.2.4 Phân tích hiện tượng Overfitting

Overfitting trên dữ liệu không phân cụm theo tỉ lệ 7:3

Bảng 14: So sánh MSE và Overfit Ratio trên ba bộ dữ liệu (Original, PCA, t-SNE) với tỷ lệ Train:Test = 7:3

Dataset	MSE (Train)			MSE (Val)			Overfit Ratio		
	50	75	100	50	75	100	50	75	100
Original	1034.6920	934.2126	963.1497	1092.6157	1038.1074	1044.8168	1.0560	1.1112	1.0848
PCA	971.5205	935.3130	939.3451	1017.9475	999.9754	995.7916	1.0478	1.0691	1.0601
t-SNE	4072.3131	3760.1801	3648.7369	4003.1245	3701.3815	3627.3772	0.9830	0.9844	0.9941

1. Dấu hiệu Chính của Overfitting:

- **Chênh lệch giữa MSE (Train) và MSE (Val) không quá lớn:**
 - Trên cả ba bộ dữ liệu và với các kích thước mạng 50, 75, 100, giá trị **MSE (Train)** và **MSE (Val)** khá sát nhau, chỉ chênh lệch nhẹ.
 - Ví dụ, với bộ dữ liệu **Original** và kích thước 100, MSE (Train) là 963.1497 trong khi MSE (Val) là 1044.8168.
 - Điều này cho thấy mô hình có khả năng khai thác khá tốt, không bị quá khớp (overfitting) rõ rệt.
- **Chỉ số Overfit Ratio gần bằng 1:**

- **Overfit Ratio** được tính bằng $\frac{\text{MSE}(\text{Val})}{\text{MSE}(\text{Train})}$. Giá trị này ở mức khoảng từ 0.98 đến 1.09 cho cả ba bộ dữ liệu.
- Các giá trị Overfit Ratio trên đều rất gần 1, chứng tỏ sự cân bằng tốt giữa lỗi trên tập huấn luyện và tập kiểm định, không xuất hiện hiện tượng overfitting nghiêm trọng.
- Ví dụ, với bộ dữ liệu **Original** và kích thước 50, Overfit Ratio là 1.0560, nghĩa là lỗi trên tập validation chỉ cao hơn khoảng 5.6% so với tập training.
- **Trường hợp dữ liệu t-SNE khác biệt:**

- Với bộ dữ liệu **t-SNE**, MSE (Train) và MSE (Val) cũng khá gần nhau và Overfit Ratio thậm chí còn thấp hơn 1 (từ 0.98 đến 0.99), tức lỗi trên tập validation còn thấp hơn hoặc bằng lỗi trên tập training.
- Điều này có thể cho thấy mô hình chưa bị overfitting trên dữ liệu t-SNE hoặc có thể dữ liệu sau khi giảm chiều bằng t-SNE làm thay đổi đặc tính của dữ liệu.

2. Ảnh hưởng của số lượng nơ-ron:

- **MSE (Train):** Khi kích thước hidden layer tăng từ 50 lên 100, **MSE (Train) nhìn chung giảm nhẹ hoặc giữ ổn định** trên cả ba bộ dữ liệu. Ví dụ, với bộ **Original**, MSE (Train) giảm từ **1034.69** (50 nơ-ron) xuống **963.15** (100 nơ-ron). Điều này cho thấy mô hình học được nhiều đặc trưng hơn khi tăng số lượng nơ-ron.
- **MSE (Val):** Trên tập validation, MSE cũng có xu hướng **ổn định hoặc giảm nhẹ** theo chiều tăng kích thước hidden layer. Chẳng hạn, với bộ **PCA**, MSE (Val) giảm từ **1017.95** (50) xuống **995.79** (100). Điều này phản ánh rằng mô hình không bị overfitting nghiêm trọng khi tăng độ phức tạp.
- **Overfit Ratio:** Overfit Ratio trên cả ba bộ dữ liệu đều duy trì ở mức **thấp và ổn định**, không tăng đáng kể theo kích thước hidden layer. Đặc biệt, trên bộ **t-SNE**, Overfit Ratio luôn **nhỏ hơn 1** – cho thấy mô hình thậm chí hoạt động tốt hơn trên tập validation, có thể do hiệu ứng ngẫu nhiên hoặc lợi thế phân tách rõ ràng hơn sau t-SNE.
- **Nhận xét chung:** Việc tăng số lượng nơ-ron từ 50 lên 100 không làm tăng đáng kể lỗi kiểm tra hoặc hiện tượng overfitting. Điều này gợi ý rằng mô hình MLP có thể mở rộng về độ phức tạp (với hidden layer lớn hơn) mà vẫn duy trì được khả năng tổng quát hóa tốt trên các dạng dữ liệu đã chuẩn hóa hoặc giảm chiều. Tuy nhiên, mức cải thiện là **không lớn**, cho thấy cần thận trọng khi tăng thêm số lớp hoặc số nơ-ron.

3. So sánh giữa các Bộ dữ liệu:

- **Original:**
 - Cho thấy mức độ overfitting khá thấp với Overfit Ratio dao động khoảng 1.05 đến 1.11, cho thấy mô hình có sự khớp khá tốt giữa tập huấn luyện và tập kiểm tra.
 - MSE (Train) và MSE (Val) ở mức trung bình, không quá chênh lệch lớn.
- **PCA:**

- MSE (Train) và MSE (Val) đều thấp hơn so với dữ liệu gốc Original, đặc biệt MSE (Val) giảm đáng kể (ví dụ 995.79 so với 1044.82 với hidden size 100).
 - Overfit Ratio thấp hơn Original (khoảng 1.04 - 1.07), cho thấy PCA giúp giảm nhẹ hiện tượng overfitting đồng thời cải thiện hiệu suất mô hình trên tập kiểm tra.
- **t-SNE:**
- MSE (Train) và MSE (Val) rất cao, gần như bằng nhau, và Overfit Ratio nhỏ hơn 1 (khoảng 0.98 - 0.99), chứng tỏ mô hình gần như không bị overfitting nhưng lại thể hiện hiệu suất kém do lỗi rất lớn.
 - Điều này cho thấy việc áp dụng t-SNE trong trường hợp này không giúp cải thiện hiệu suất mô hình, có thể do đặc trưng t-SNE không phù hợp hoặc làm mất thông tin quan trọng, dẫn đến hiệu quả dự đoán kém.

4. Kết luận: Mô hình MLP huấn luyện trên ba bộ dữ liệu không phân cụm cho thấy tình trạng **overfitting nhẹ đến vừa phải**, với Overfit Ratio chủ yếu dao động từ **1.05 đến 1.15**. Điều này cho thấy mô hình học tốt trên tập huấn luyện và vẫn duy trì khả năng tổng quát hóa khá ổn trên tập kiểm tra.

- Việc **tăng số lượng nơ-ron trong hidden layer** từ 50 lên 100 nhìn chung **không làm tăng đáng kể mức độ overfitting**. Overfit Ratio chỉ tăng nhẹ ở một vài cấu hình và vẫn nằm trong ngưỡng chấp nhận được, phản ánh rằng mô hình có thể mở rộng về độ phức tạp mà không bị suy giảm hiệu suất quá nhiều.
- **Dữ liệu PCA giúp giảm tổng thể MSE trên tập validation so với dữ liệu gốc**, cho thấy việc giảm chiều bằng PCA giúp nổi bật các đặc trưng quan trọng. Tuy nhiên, độ chênh lệch giữa MSE huấn luyện và kiểm tra cũng tăng nhẹ, dẫn đến Overfit Ratio không cải thiện nhiều, thậm chí đôi khi còn cao hơn so với dữ liệu gốc.
- **Dữ liệu t-SNE cho Overfit Ratio thường nhỏ hơn 1**, nghĩa là mô hình có thể **underfit nhẹ** hoặc hoạt động cân bằng hơn trên tập validation. Mặc dù MSE train và val khá sát nhau, nhưng giá trị MSE tuyệt đối vẫn còn cao, cho thấy t-SNE có thể chưa phải cách giảm chiều tối ưu cho mô hình hồi quy như MLP.

Hướng cải thiện: Cần xem xét điều chỉnh các siêu tham số của MLP như số nơ-ron, learning rate, số epoch, kết hợp với các kỹ thuật regularization nhằm giảm overfitting. Ngoài ra, mở rộng kích thước tập huấn luyện hoặc kết hợp thêm các kỹ thuật tiền xử lý dữ liệu khác cũng có thể giúp cải thiện hiệu suất và khả năng tổng quát của mô hình.

Ảnh hưởng của hệ số regularization alpha đến Overfitting

1. Hiệu quả giảm Overfitting (Overfit Ratio):

- **Dữ liệu gốc (Original):** Khi tăng hệ số regularization alpha từ 0.0001 lên 0.01, tỉ số Overfit Ratio dao động nhẹ quanh mức 1.11. Điều này cho thấy regularization đã giúp duy trì sự ổn định của mô hình, tuy nhiên chưa có tác động đáng kể trong việc giảm overfitting.

Bảng 15: Ảnh hưởng của hệ số regularization alpha đến kết quả huấn luyện và kiểm định (MSE), với tỉ lệ chia dữ liệu 7:3 trên các phương pháp tiền xử lý khác nhau.

Dataset	Alpha	MSE (Train)	MSE (Val)	Overfit Ratio
Original	0.0001	934.2126	1038.1074	1.1112
	0.0010	929.6150	1029.3433	1.1073
	0.0100	935.8106	1037.4994	1.1087
PCA	0.0001	935.3130	999.9754	1.0691
	0.0010	937.5284	999.1341	1.0657
	0.0100	934.7500	999.2212	1.0690
t-SNE	0.0001	3760.1801	3701.3815	0.9844
	0.0010	3779.9609	3721.8199	0.9846
	0.0100	3780.4383	3722.9339	0.9848

- Dữ liệu sau PCA:** Overfit Ratio thấp hơn so với dữ liệu gốc, nằm trong khoảng từ 1.0691 đến 1.0657, và nhìn chung có xu hướng giảm nhẹ khi alpha tăng. Điều này cho thấy việc kết hợp PCA với regularization giúp kiểm soát overfitting tốt hơn, đồng thời duy trì độ khái quát của mô hình trên tập kiểm định.
- Dữ liệu sau t-SNE:** Overfit Ratio luôn nhỏ hơn 1 (từ 0.9844 đến 0.9848), phản ánh rằng mô hình học trên t-SNE có thể đã bị *underfitting* nhẹ hoặc rất khớp với tập kiểm định. Sự ổn định này khi thay đổi alpha cho thấy regularization gần như không ảnh hưởng nhiều đến mô hình, có thể do đặc tính giảm chiều phi tuyến mạnh mẽ của t-SNE.

Nhận xét chung: Regularization thông qua hệ số alpha có tác dụng ổn định mô hình trên dữ liệu gốc và PCA, với mức giảm nhẹ Overfit Ratio ở PCA. Trong khi đó, trên t-SNE, do dữ liệu đã được giảm chiều phi tuyến mạnh mẽ, tác động của regularization là không đáng kể. Việc lựa chọn phương pháp tiền xử lý phù hợp kết hợp với điều chỉnh tham số regularization là rất quan trọng trong việc kiểm soát overfitting và nâng cao hiệu quả tổng quát của mô hình.

4. Phân tích và kết luận:

- Bảng 15 cho thấy việc điều chỉnh hệ số regularization alpha ảnh hưởng đến hiệu năng mô hình MLP ở các mức độ khác nhau, tùy thuộc vào cách tiền xử lý dữ liệu. Nhìn chung, MSE trên tập huấn luyện (Train) biến động rất ít theo alpha, trong khi MSE trên tập kiểm định (Val) thay đổi nhẹ, phản ánh mức độ cải thiện khái quát hóa của mô hình là có nhưng không mạnh.
- Trên dữ liệu gốc (Original), Overfit Ratio duy trì ổn định quanh mức 1.11 khi tăng alpha, cho thấy regularization giúp kiểm soát mô hình nhưng chưa giảm rõ rệt hiện tượng overfitting.
- Trên dữ liệu PCA, Overfit Ratio luôn thấp hơn dữ liệu gốc (dao động từ 1.0691 đến 1.0657), và giảm nhẹ theo alpha, cho thấy regularization hoạt động hiệu quả hơn khi mô hình được huấn luyện trên dữ liệu đã giảm chiều. Đồng thời, MSE kiểm định trên PCA cũng thấp hơn đáng kể so với dữ liệu gốc, cho thấy PCA giúp cải thiện khả năng dự đoán của mô hình.
- Ngược lại, trên dữ liệu t-SNE, cả MSE huấn luyện và kiểm định đều rất cao (gấp gần 4 lần dữ liệu gốc), và Overfit Ratio < 1 (khoảng 0.9844), phản ánh mô hình có thể bị underfitting hoặc không học được đặc trưng

hiệu quả. Điều này là hợp lý do t-SNE không bảo toàn cấu trúc toàn cục và không phù hợp làm đầu vào cho mô hình hồi quy như MLP.

- Tóm lại, regularization giúp ổn định và phân nào giảm overfitting, nhưng hiệu quả phụ thuộc mạnh vào đặc trưng dữ liệu đầu vào. Kết quả tốt nhất đạt được khi kết hợp PCA với regularization – giúp giảm nhiễu, giảm chiều và giữ lại thông tin chính yếu. Trong khi đó, t-SNE tỏ ra không hiệu quả trong bối cảnh này. Việc chọn đúng kỹ thuật tiền xử lý là yếu tố then chốt để nâng cao hiệu quả dự đoán và khả năng tổng quát hóa của mô hình MLP.

5.2.5 Trực quan hóa và đánh giá tương quan phần dư

1. Phân tích thống kê phần dư

Bảng 16: Bảng kết quả tổng quan đánh giá mô hình MLP theo các tỉ lệ train:test và loại dữ liệu khác nhau

Data Type	Train:Test	Residual Mean	Residual Std	Residual Min	Residual Max
Original	8:2	-0.4806	33.1755	-274.9650	286.6950
PCA	8:2	0.8627	32.8454	-289.8600	292.0967
t-SNE	8:2	-0.4124	61.0481	-253.7982	291.4984
Original	7:3	0.3539	33.2366	-277.1871	282.0469
PCA	7:3	-0.0891	32.5642	-308.5134	292.8222
t-SNE	7:3	0.1447	62.4755	-268.2937	287.7785
Original	6:4	0.4157	33.2142	-341.3483	298.6320
PCA	6:4	-0.4360	32.7604	-316.2273	287.4688
t-SNE	6:4	-0.7372	66.3756	-285.5636	288.7181

- PCA cho phần dư ổn định nhất:** Dữ liệu PCA có độ lệch chuẩn phần dư (Residual Std) thấp nhất (32.5642–32.8454) trên cả ba tỷ lệ Train:Test, so với dữ liệu gốc và t-SNE. Điều này cho thấy dự đoán trên dữ liệu PCA có độ biến thiên thấp, phù hợp với MSE (Val) thấp nhất (1060.4338–1079.5648).
- Phần dư trung bình gần 0:** Giá trị trung bình phần dư (Residual Mean) trên các bộ dữ liệu và tỷ lệ Train:Test dao động trong khoảng nhỏ (-0.7372 đến 0.8627), cho thấy mô hình MLP không có thiên lệch rõ rệt. Dữ liệu PCA với tỷ lệ 7:3 có Residual Mean gần 0 nhất (-0.0891), thể hiện sự cân bằng tốt nhất.
- Phạm vi phần dư rộng:** Phần dư có giá trị tối thiểu (Residual Min) và tối đa (Residual Max) dao động lớn, đặc biệt trên dữ liệu PCA (-316.2273 đến 292.8222) và dữ liệu gốc (-341.3483 đến 298.6320). Điều này chỉ ra rằng mô hình vẫn gặp các trường hợp dự đoán sai lệch lớn, cần cải thiện qua tiền xử lý hoặc tối ưu hóa mô hình.
- t-SNE có hiệu suất kém nhất:** Dữ liệu t-SNE ghi nhận độ lệch chuẩn phần dư cao nhất (61.0481–66.3756) và phạm vi phần dư rộng (-285.5636 đến 291.4984), đi kèm R^2 thấp (0.5975–0.6567) và MSE cao (3727.0438–4406.2571). Điều này cho thấy t-SNE có thể làm mất thông tin quan trọng trong quá trình giảm chiều.
- Tỷ lệ 7:3 tối ưu cho PCA:** Với tỷ lệ Train:Test = 7:3, dữ liệu PCA đạt MSE (Val) thấp nhất (1060.4338), Residual Std thấp (32.5642) và Residual Mean gần 0 (-0.0891), cho thấy sự ổn định và chính xác cao trong dự đoán.

Lưu ý:

- Dữ liệu PCA là lựa chọn tốt nhất về độ ổn định và chính xác, nhưng phạm vi phần dư lớn (Residual Min/Max) cho thấy cần áp dụng thêm các kỹ thuật tiền xử lý (ví dụ: chuẩn hóa, xử lý ngoại lệ) hoặc tinh chỉnh siêu tham số để giảm sai lệch cực đại.
- Hiệu suất kém của t-SNE có thể do số chiều giảm quá thấp hoặc không phù hợp với bài toán. Cần thử nghiệm thêm với các cấu hình t-SNE khác (ví dụ: thay đổi perplexity hoặc số chiều).
- Kết quả phụ thuộc vào đặc điểm bộ dữ liệu và cấu hình mô hình MLP. Cần kiểm tra thêm trên các bộ dữ liệu khác hoặc với các mô hình khác (ví dụ: Random Forest, XGBoost) để đánh giá tính tổng quát.

2. Phân tích tương quan giữa phần dư và đặc trưng



Hình 25: Tương quan giữa phần dư và các đặc trưng đầu vào

Biểu đồ heatmap "Tương quan giữa phần dư và đặc trưng đầu vào" thể hiện các hệ số tương quan giữa phần dư của mô hình và các đặc trưng đầu vào (O3, NH3, NO, Benzene, CO, PM2.5, Xylene, NO2, SO2, Toluene, NOx, PM10) đều **rất gần 0**. Cụ thể, tương quan với O3 là 0.015, với NH3 là -0.0013, và với PM10 là -0.026.

- Ý nghĩa:** Điều này cho thấy phần dư (sai số dự đoán) không có mối quan hệ tuyến tính đáng kể với bất kỳ đặc trưng nào, phản ánh rằng mô hình MLP đã xử lý tốt thông tin từ các đặc trưng. Tuy nhiên, mức độ tương quan thấp cũng có thể ám chỉ rằng mô hình chưa khai thác hết các mối quan hệ phi tuyến tính hoặc cần thêm đặc trưng bổ sung để giải thích phần dư còn lại.

3. Nhận định chung

Mô hình MLP **đạt hiệu suất ở mức chấp nhận được** nhưng **chưa tối ưu** về độ chính xác và độ ổn định cho các dự đoán chi tiết.

- **Ưu điểm:** Mô hình không biểu hiện thiên lèch hệ thống (trung bình phần dư gần 0) và phần dư không có tương quan tuyến tính đáng kể với các đặc trưng đầu vào (tương quan từ -0.026 đến 0.015), cho thấy mô hình đã học được các mối quan hệ cơ bản trong dữ liệu. Đặc biệt, dữ liệu PCA với tỷ lệ Train:Test 7:3 cho kết quả tốt nhất với Residual Mean gần 0 (-0.0891) và MSE thấp (1060.4338).
- **Hạn chế:** Độ biến thiên của phần dư cao, với độ lệch chuẩn lớn (32.5642–66.3756) và phạm vi phần dư rộng (từ -341.3483 đến 298.6320 trên dữ liệu gốc), cho thấy mô hình vẫn gặp sai lệch lớn ở một số trường hợp. Hiệu suất của t-SNE đặc biệt kém (R^2 từ 0.5975–0.6567, MSE cao 3727.0438–4406.2571), có thể do mất thông tin trong quá trình giảm chiều.

Để nâng cao hiệu quả, cần tập trung vào việc giảm độ biến thiên của sai số, thông qua các phương pháp như tinh chỉnh siêu tham số.

5.3 So sánh hiệu suất giữa Random Forest và MLP

5.3.1 So sánh hiệu suất trên các chỉ số đánh giá

- **Random Forest (RF):** Đạt hiệu suất tốt nhất trên dữ liệu gốc với MSE = 887.38, RMSE = 29.79, R^2 = 0.9188, MAE = 17.50 (200 cây, tỷ lệ train:test = 7:3). Tuy nhiên, hiệu suất giảm khi áp dụng PCA (MSE = 982.35) và t-SNE (MSE = 1177.67).
- **MLP:** Hiệu quả nhất trên dữ liệu PCA với MSE = 1060.15, RMSE = 32.56, R^2 = 0.9030, MAE = 20.89 (75 nơ-ron ẩn, 7:3). Tuy nhiên, mô hình hoạt động kém trên t-SNE (MSE = 3727.04).

Nhận xét: Random Forest vượt trội trên dữ liệu gốc, trong khi MLP hiệu quả hơn trên dữ liệu đã giảm chiều bằng PCA nhờ khả năng học các mối quan hệ phi tuyến tính.

5.3.2 So sánh độ ổn định và phân tích phần dư

- **Residual Mean:** Cả hai mô hình có trung bình phần dư gần bằng 0 (RF: -0.6754 đến 0.5115; MLP: -0.7372 đến 0.8627). Trong đó, MLP trên PCA (7:3) đạt Residual Mean = -0.0891, ổn định hơn RF (-0.0547).
- **Residual Std:** RF có độ lệch chuẩn phần dư thấp hơn (29.87–34.75) so với MLP (32.56–66.38), đặc biệt là trên t-SNE. Điều này cho thấy RF có độ ổn định tốt hơn.
- **Residual Min/Max:** Cả hai mô hình đều có phần dư biến thiên rộng (RF: -291.00 đến 298.39; MLP: -341.35 đến 298.63). Tuy nhiên, MLP trên PCA có biên độ phần dư nhỏ hơn, cho thấy khả năng kiểm soát sai số tốt hơn trên dữ liệu này.

Nhận xét: Random Forest ổn định hơn trên dữ liệu gốc, trong khi MLP hoạt động tốt hơn khi được kết hợp với kỹ thuật giảm chiều như PCA.

5.3.3 So sánh khả năng tổng quát hóa và hiện tượng Overfitting

- **Overfit Ratio:** RF có Overfit Ratio cao (6.54–7.49) khi không giới hạn độ sâu cây (max_depth = None), giảm xuống 1.00–1.69 khi giới hạn max_depth, cho thấy hiện tượng overfitting khá nghiêm trọng. Ngược lại, MLP có Overfit Ratio thấp (0.98–1.11), đặc biệt trên t-SNE có giá trị nhỏ hơn 1, cho thấy khả năng tổng quát hóa tốt hơn.
- **Ảnh hưởng siêu tham số:** Tăng số lượng cây (RF) làm tăng nhẹ Overfit Ratio, trong khi tăng số lượng nơ-ron trong MLP không ảnh hưởng nhiều nhờ vào kỹ thuật regularization.

Nhận xét: MLP tổng quát hóa tốt hơn nhờ được hỗ trợ bởi regularization và tiền xử lý (PCA). Trong khi đó, RF cần được tinh chỉnh thêm, đặc biệt là giới hạn độ sâu cây để tránh overfitting.

5.3.4 Kết luận chung

- **Random Forest:** Phù hợp nhất với dữ liệu gốc, đạt hiệu suất cao ($MSE = 887.38$, $R^2 = 0.9188$), ổn định (Residual Std ≈ 30), nhưng dễ bị overfitting nếu không giới hạn độ sâu cây. Thích hợp cho các bài toán yêu cầu tốc độ huấn luyện nhanh và độ ổn định cao.
- **MLP:** Hoạt động tốt nhất trên dữ liệu PCA ($MSE = 1060.15$, $R^2 = 0.9030$), có khả năng học các mối quan hệ phi tuyến tính, ít bị overfitting nhờ regularization, nhưng hiệu suất kém trên dữ liệu t-SNE. Thích hợp cho dữ liệu đã qua tiền xử lý và chuẩn hóa tốt.
- **Khuyến nghị:**
 - Sử dụng **Random Forest** cho dữ liệu thô hoặc khi yêu cầu mô hình đơn giản, dễ giải thích.
 - Sử dụng **MLP** khi dữ liệu đã được xử lý tốt (như PCA) và cần khai thác các mối quan hệ phi tuyến.
 - Cần tinh chỉnh thêm các siêu tham số để cải thiện hiệu suất, như `max_depth` trong RF hoặc số lượng nơ-ron và hệ số regularization `alpha` trong MLP.

Chương 6

Mô hình phân loại

6.1 Bài toán phân loại

6.1.1 Phân chia đầu ra thành 4 khoảng và xác định ngưỡng

Dựa trên phân phối dữ liệu AQI đã cho:

- Rất cao: 7416 mẫu (25.1%)
- Trung bình: 7407 mẫu (25.1%)
- Cao: 7362 mẫu (24.9%)
- Thấp: 7346 mẫu (24.9%)

Tổng số mẫu là 29531. Mỗi khoảng trong số 4 khoảng sẽ có xấp xỉ $29531/4 \approx 7382.75$ mẫu. Các lớp AQI hiện tại đã đáp ứng tốt yêu cầu này.

Nhãn mới cho bài toán phân loại Giả sử thứ tự mức độ AQI là: Thấp < Trung bình < Cao < Rất cao.

- Nhãn 1 (hoặc “AQI_Thap”):** Tương ứng với lớp “Thấp” (7346 mẫu).
- Nhãn 2 (hoặc “AQI_TrungBinh”):** Tương ứng với lớp “Trung bình” (7407 mẫu).
- Nhãn 3 (hoặc “AQI_Cao”):** Tương ứng với lớp “Cao” (7362 mẫu).
- Nhãn 4 (hoặc “AQI_RatCao”):** Tương ứng với lớp “Rất cao” (7416 mẫu).

6.1.2 Mô hình hóa bài toán phân loại AQI

Mục tiêu bài toán Xây dựng mô hình dự đoán một trong 4 nhãn AQI (“AQI_Thap”, “AQI_TrungBinh”, “AQI_Cao”, “AQI_RatCao”) dựa trên các đặc trưng đầu vào.

Đặc trưng đầu vào (Input Features) Các yếu tố có thể ảnh hưởng đến AQI:

- Các thông số ô nhiễm: PM2.5, PM10, NO, NO2, NOx, NH3, CO, SO2, O3
- Thông tin địa điểm: City (mã hóa thành biến phân loại)
- Thông tin thời gian: Date (trích xuất các thuộc tính như tháng, mùa, ngày trong tuần)

Nhãn đầu ra (Output Labels) 4 lớp AQI đã xác định: “AQI_Thap”, “AQI_TrungBinh”, “AQI_Cao”, “AQI_RatCao”.

6.2 Mô hình phân loại Naive Bayes

Naive Bayes được khởi tạo và sử dụng là Gaussian Naive Bayes (qua lớp GaussianNB() của thư viện scikit-learn)

Gaussian Naive Bayes (GNB) thường được dùng cho bài toán phân loại, đặc biệt khi các đặc trưng (features) đầu vào là các biến liên tục và giả thiết rằng mỗi lớp (class) có phân phối xác suất của từng đặc trưng tuân theo phân phối Gaussian (hội tụ chuẩn).

Ưu điểm chính:

- Đơn giản, dễ triển khai, tốc độ huấn luyện và suy đoán (prediction) rất nhanh.
- Khi dữ liệu thật sự xấp xỉ phân phối chuẩn (hoặc ít nhất gần đúng chuẩn), hiệu quả phân loại khá tốt.
- Không cần quá nhiều tham số (chỉ cần tính trung bình và phương sai cho mỗi đặc trưng trên mỗi lớp).

Trong bài toán dự báo ô nhiễm không khí (AQI) này, các chỉ số ô nhiễm (PM2.5, PM10, NO2, CO,...) và biến thời gian (Year, Month, Day) đều là dữ liệu liên tục. GaussianNB sẽ tính xác suất của mỗi điểm dữ liệu thuộc một lớp AQI nhất định, dựa trên giả thiết rằng mỗi biến đầu vào (ví dụ PM2.5) theo phân phối chuẩn trong mỗi lớp (“Thấp”, “Trung bình”, “Cao”, “Rất cao”).

Đánh giá hiệu suất mô hình theo tỷ lệ Train:Validation và loại dữ liệu (Gốc vs PCA)

Bảng 17: Kết quả đánh giá mô hình với các tỷ lệ chia train-validation và dữ liệu gốc vs PCA

Tỷ lệ Train:Val	Dữ liệu	Accuracy	Precision (macro)	Recall (macro)	F1-score (macro)
0.8:0.2	Gốc	0.7036	0.7000	0.7036	0.7011
0.8:0.2	PCA	0.6027	0.6031	0.6027	0.6010
0.7:0.3	Gốc	0.7042	0.7004	0.7042	0.7017
0.7:0.3	PCA	0.6009	0.5997	0.6009	0.5981
0.6:0.4	Gốc	0.7057	0.7015	0.7056	0.7029
0.6:0.4	PCA	0.6015	0.6001	0.6015	0.5986

Chi tiết kết quả phân loại theo nhãn cho các tỷ lệ Train:Validation và loại dữ liệu (Gốc vs PCA)

Bảng 18: So sánh tất cả các kết quả chi tiết (gốc vs PCA) ở các tỉ lệ Train:Val

Loại dữ liệu	Train:Val	Nhãn	Metrics			
			Precision	Recall	F1-score	Support
Gốc	0.8:0.2	0	0.73	0.80	0.76	1469
		1	0.59	0.54	0.56	1482
		2	0.62	0.61	0.62	1473
		3	0.86	0.86	0.86	1483
PCA	0.8:0.2	0	0.66	0.76	0.71	1469
		1	0.47	0.50	0.49	1482
		2	0.49	0.43	0.46	1473
		3	0.79	0.72	0.75	1483
Gốc	0.7:0.3	0	0.73	0.80	0.76	2204
		1	0.58	0.55	0.57	2222
		2	0.63	0.59	0.61	2209
		3	0.86	0.87	0.87	2225
PCA	0.7:0.3	0	0.66	0.77	0.71	2204
		1	0.47	0.50	0.48	2222
		2	0.48	0.41	0.44	2209
		3	0.79	0.72	0.75	2225
Gốc	0.6:0.4	0	0.73	0.80	0.76	2938
		1	0.58	0.56	0.57	2963
		2	0.63	0.59	0.61	2945
		3	0.86	0.88	0.87	2967
PCA	0.6:0.4	0	0.66	0.77	0.71	2938
		1	0.47	0.49	0.48	2963
		2	0.49	0.42	0.45	2945
		3	0.79	0.73	0.76	2967

Nhìn chung, **mô hình trên dữ liệu gốc cho kết quả vượt trội hơn hẳn so với dữ liệu đã giảm chiều bằng PCA**. Việc thay đổi tỷ lệ Train:Validation không gây ra sự khác biệt đáng kể về hiệu suất trong cả hai trường hợp.

So sánh Dữ liệu Gốc và Dữ liệu PCA

- Hiệu suất tổng thể:** Trên cả ba tỷ lệ phân chia, mô hình sử dụng **dữ liệu gốc** luôn đạt được các chỉ số (Accuracy, Precision, Recall, F1-score) cao hơn hẳn so với mô hình sử dụng **dữ liệu PCA**. Cụ thể, các chỉ số trên dữ liệu gốc luôn dao động quanh mức ~70%, trong khi trên dữ liệu PCA chỉ ở mức ~60%, tức là **thấp hơn khoảng 10%**.
- Phân tích theo từng nhãn:**
 - Hiệu suất giảm trên tất cả các nhãn khi dùng dữ liệu PCA.
 - Sự sụt giảm này ảnh hưởng nặng nề nhất đến các nhãn vốn đã khó phân loại hơn. Ví dụ, **Nhãn 1** và **Nhãn 2** có F1-score trên dữ liệu gốc là ~0.57 và ~0.61, nhưng đã giảm mạnh xuống chỉ còn ~0.48 và ~0.45 trên dữ liệu PCA.
 - Ngay cả với nhãn dễ phân loại nhất là **Nhãn 3** (F1-score ~0.87 trên dữ liệu gốc), hiệu suất cũng giảm đáng kể xuống còn ~0.75 trên dữ liệu PCA.

So sánh các Tỷ lệ Train:Validation

- Việc thay đổi tỷ lệ từ **8:2** sang **7:3** và **6:4** (tức là giảm dữ liệu huấn luyện và tăng dữ liệu kiểm định) **không gây ra sự thay đổi đáng kể** nào về hiệu suất của mô hình.
- Trên dữ liệu gốc, Accuracy chỉ thay đổi rất nhỏ (từ 0.7036 lên 0.7057). Tương tự, trên dữ liệu PCA, các chỉ số gần như đứng yên.
- Điều này cho thấy mô hình khá ổn định và hiệu suất của nó trong trường hợp này không phụ thuộc nhiều vào sự thay đổi nhỏ trong tỷ lệ phân chia dữ liệu.

Giải thích

Nguyên nhân chính cho sự khác biệt hiệu suất giữa hai loại dữ liệu là do **bản chất của việc giảm chiều bằng PCA**.

Dữ liệu PCA trong thí nghiệm này chỉ giữ lại 65% phương sai của dữ liệu gốc. Điều này có nghĩa là 35% phương sai đã bị loại bỏ.

- Mất mát thông tin:** Phương sai trong dữ liệu thường chứa đựng thông tin và các đặc trưng quan trọng giúp mô hình phân biệt giữa các lớp (nhãn). Khi 35% thông tin này bị mất đi, mô hình có ít "mạnh mẽ" hơn để học cách phân loại chính xác.
- Tác động đến mô hình:** Đặc biệt, những thông tin giúp phân biệt các lớp khó (như **Nhãn 1** và **Nhãn 2**) có thể nằm trong phần phương sai đã bị loại bỏ. Điều này dẫn đến việc mô hình nhận dạng các lớp này kém đi rất nhiều, kéo theo hiệu suất tổng thể sụt giảm.

Kết luận

Việc sử dụng PCA để giảm chiều dữ liệu đã làm mất một lượng thông tin quan trọng, khiến cho nhiệm vụ phân loại của mô hình trở nên khó khăn hơn và dẫn đến hiệu suất sụt giảm trên mọi phương diện so với việc sử dụng toàn bộ dữ liệu gốc. Trong trường hợp này, lợi ích về tốc độ tính toán (nếu có) khi dùng PCA đã không thể bù đắp được sự mất mát về độ chính xác của mô hình.

6.3 Mô hình phân loại Random Forest

Đánh giá hiệu suất mô hình theo tỷ lệ Train:Validation và loại dữ liệu (Gốc vs PCA)

Bảng 19: Kết quả đánh giá mô hình với các tỷ lệ chia train-validation và dữ liệu gốc vs PCA

Tỷ lệ Train:Val	Dữ liệu	Accuracy	Precision (macro)	Recall (macro)	F1-score (macro)
0.8:0.2	Gốc	0.8224	0.8225	0.8225	0.8223
0.8:0.2	PCA	0.7266	0.7257	0.7266	0.7260
0.7:0.3	Gốc	0.8272	0.8269	0.8272	0.8270
0.7:0.3	PCA	0.7301	0.7288	0.7301	0.7293
0.6:0.4	Gốc	0.8291	0.8286	0.8291	0.8288
0.6:0.4	PCA	0.7235	0.7221	0.7235	0.7227

Chi tiết kết quả phân loại theo nhãn cho các tỷ lệ Train:Validation và loại dữ liệu (Gốc vs PCA)

Bảng 20: Chi tiết kết quả phân loại theo nhãn cho các tỷ lệ Train:Val và loại dữ liệu (Gốc vs PCA)

Loại dữ liệu	Train:Val	Nhãn	Metrics			
			Precision	Recall	F1-score	Support
Gốc	0.8:0.2	0	0.85	0.87	0.86	1469
		1	0.75	0.73	0.74	1482
		2	0.77	0.78	0.78	1473
		3	0.92	0.90	0.91	1483
PCA	0.8:0.2	0	0.76	0.79	0.78	1469
		1	0.65	0.62	0.63	1482
		2	0.64	0.64	0.64	1473
		3	0.86	0.85	0.86	1483
Gốc	0.7:0.3	0	0.85	0.88	0.87	2204
		1	0.75	0.75	0.75	2222
		2	0.78	0.77	0.77	2209
		3	0.92	0.91	0.92	2225
PCA	0.7:0.3	0	0.78	0.80	0.79	2204
		1	0.65	0.62	0.64	2222
		2	0.64	0.64	0.64	2209
		3	0.85	0.86	0.86	2225
Gốc	0.6:0.4	0	0.86	0.88	0.87	2938
		1	0.77	0.74	0.76	2963
		2	0.77	0.78	0.78	2945
		3	0.92	0.91	0.91	2967
PCA	0.6:0.4	0	0.77	0.79	0.78	2938
		1	0.63	0.62	0.62	2963
		2	0.63	0.63	0.63	2945
		3	0.85	0.86	0.86	2967

Nhìn chung, mô hình trên dữ liệu gốc cho kết quả vượt trội hơn hẳn so với dữ liệu đã giảm chiều bằng PCA (chỉ giữ lại 65% phương sai). Việc thay đổi tỷ lệ Train:Validation không gây ra sự khác biệt đáng kể về hiệu suất trong cả hai trường hợp.

So sánh Dữ liệu Gốc và Dữ liệu PCA

- **Hiệu suất tổng thể:** Trên cả ba tỷ lệ phân chia, mô hình sử dụng **dữ liệu gốc** luôn đạt được các chỉ số (Accuracy, Precision, Recall, F1-score) cao hơn hẳn so với mô hình sử dụng **dữ liệu PCA**. Cụ thể, các chỉ số trên dữ liệu gốc luôn dao động quanh mức $\sim 82\text{-}83\%$, trong khi trên dữ liệu PCA chỉ ở mức $\sim 72\text{-}73\%$, tức là **thấp hơn khoảng 10%**.
- **Phân tích theo từng nhãn:** (Sử dụng tỷ lệ 8:2 làm ví dụ đại diện, vì các tỷ lệ khác cho kết quả tương tự)
 - Hiệu suất giảm trên tất cả các nhãn khi dùng dữ liệu PCA.
 - Sự sụt giảm này ảnh hưởng nặng nề nhất đến các nhãn vốn đã khó phân loại hơn. Ví dụ, **Nhãn 1** và **Nhãn 2** có F1-score trên dữ liệu gốc là ~ 0.74 và ~ 0.78 , nhưng đã giảm mạnh xuống chỉ còn ~ 0.63 và ~ 0.64 trên dữ liệu PCA.
 - Ngay cả với nhãn dễ phân loại nhất là **Nhãn 3** (F1-score ~ 0.91 trên dữ liệu gốc), hiệu suất cũng giảm đáng kể xuống còn ~ 0.86 trên dữ liệu PCA. **Nhãn 0** cũng cho thấy sự sụt giảm tương tự từ ~ 0.86 xuống ~ 0.78 .

So sánh các Tỷ lệ Train:Validation

- Việc thay đổi tỷ lệ từ **8:2** sang **7:3** và **6:4** (tức là giảm dữ liệu huấn luyện và tăng dữ liệu kiểm định) **không gây ra sự thay đổi đáng kể** nào về hiệu suất của mô hình.
- Trên dữ liệu gốc, Accuracy chỉ thay đổi rất nhỏ (từ 0.8224 lên 0.8272 rồi 0.8291). Tương tự, trên dữ liệu PCA, Accuracy dao động nhẹ quanh mức 0.72-0.73 (0.7266, 0.7301, 0.7235).
- Điều này cho thấy mô hình khá ổn định và hiệu suất của nó trong trường hợp này không phụ thuộc nhiều vào sự thay đổi nhỏ trong tỷ lệ phân chia dữ liệu.

Giải thích

Nguyên nhân chính cho sự khác biệt hiệu suất giữa hai loại dữ liệu là do **bản chất của việc giảm chiều bằng PCA**.

Dữ liệu PCA trong thí nghiệm này chỉ giữ lại 65% phương sai của dữ liệu gốc. Điều này có nghĩa là 35% phương sai đã bị loại bỏ.

- **Mất mát thông tin:** Phương sai trong dữ liệu thường chứa đựng thông tin và các đặc trưng quan trọng giúp mô hình phân biệt giữa các lớp (nhãn). Khi 35% thông tin này bị mất đi, mô hình có ít "mạnh mẽ" hơn để học cách phân loại chính xác.
- **Tác động đến mô hình:** Đặc biệt, những thông tin giúp phân biệt các lớp khó hơn (như **Nhãn 1** và **Nhãn 2** với F1-score thấp hơn ban đầu) có thể nằm trong phần phương sai đã bị loại bỏ. Điều này dẫn đến việc mô hình nhận dạng các lớp này kém đi rất nhiều, kéo theo hiệu suất tổng thể sụt giảm.

Kết luận

Việc sử dụng PCA để giảm chiều dữ liệu (chỉ còn 65% phương sai) đã làm mất một lượng thông tin quan trọng, khiến cho nhiệm vụ phân loại của mô hình trở nên khó khăn hơn và dẫn đến hiệu suất giảm trên mọi phương diện so với việc sử dụng toàn bộ dữ liệu gốc. Trong trường hợp này, lợi ích về tốc độ tính toán hoặc giảm độ phức tạp (nếu có) khi dùng PCA đã không thể bù đắp được sự mất mát đáng kể về độ chính xác của mô hình.

6.4 So sánh Naive Bayes và Random Forest

Nhìn chung, mô hình **Random Forest** tỏ ra vượt trội hơn hẳn so với **Naive Bayes** trong tất cả các kịch bản thử nghiệm (cả với dữ liệu gốc và dữ liệu đã qua PCA, cũng như ở các tỷ lệ chia Train:Validation khác nhau).

6.4.1 Đánh giá Hiệu suất Tổng thể (Macro Metrics)

Accuracy (Độ chính xác tổng thể)

- **Naive Bayes (NB):** Đạt khoảng 0.70 trên dữ liệu gốc và giảm xuống khoảng 0.60 khi dùng PCA.
 - **Random Forest (RF):** Đạt khoảng 0.82 – 0.83 trên dữ liệu gốc và giảm xuống khoảng 0.72 – 0.73 khi dùng PCA.
- ⇒ **Nhận xét:** RF có accuracy cao hơn NB khoảng 0.12 – 0.13 trên cả dữ liệu gốc và PCA. Điều này cho thấy RF phân loại tổng thể tốt hơn đáng kể.

Precision, Recall, F1-score (macro)

Xu hướng tương tự như Accuracy. Các chỉ số của RF luôn cao hơn NB. Ví dụ, với tỷ lệ 0.8 : 0.2 trên dữ liệu gốc:

- NB: F1-score (macro) = 0.7011
- RF: F1-score (macro) = 0.8223

⇒ **Nhận xét:** RF cân bằng tốt hơn giữa Precision và Recall trên các lớp, dẫn đến F1-score cao hơn.

6.4.2 Ảnh hưởng của PCA (Principal Component Analysis)

Đối với Naive Bayes

PCA làm giảm đáng kể hiệu suất của NB. Accuracy giảm khoảng 0.10, và các chỉ số F1-score theo từng nhãn cũng giảm mạnh (ví dụ, nhãn 1 từ 0.56 xuống 0.49, nhãn 2 từ 0.62 xuống 0.46 ở tỷ lệ 0.8 : 0.2).

⇒ **Giải thích:** Naive Bayes giả định các đặc trưng là độc lập có điều kiện. PCA tạo ra các thành phần chính là tổ hợp tuyến tính của các đặc trưng gốc, nhằm tối đa hóa phương sai. Có thể các thành phần chính này không còn giữ được (hoặc làm yếu đi) mối quan hệ "ngây thơ" mà Naive Bayes dựa vào để tính toán xác suất, hoặc

các đặc trưng quan trọng cho việc phân loại của NB lại không nằm trong các thành phần chính có phuong sai lớn nhất. Việc giảm chiều có thể đã loại bỏ thông tin hữu ích cho mô hình đơn giản như NB.

Đối với Random Forest

PCA cũng làm giảm hiệu suất của RF, nhưng mức độ giảm ít nghiêm trọng hơn so với NB. Accuracy giảm khoảng 0.10. Mặc dù giảm, RF với PCA (Accuracy $\sim 0.72 - 0.73$) vẫn cho kết quả tốt hơn NB trên dữ liệu gốc (Accuracy ~ 0.70).

⇒ **Giải thích:** Random Forest có khả năng tự xử lý các đặc trưng tương quan và lựa chọn đặc trưng quan trọng thông qua cấu trúc cây quyết định. Mặc dù PCA có thể loại bỏ một số nhiễu hoặc giảm tính đa cộng tuyến, việc giảm chiều cũng có thể làm mất đi một số thông tin hoặc tương tác đặc trưng tinh vi mà RF có khả năng khai thác. Tuy nhiên, do bản chất mạnh mẽ hơn, RF vẫn duy trì được hiệu suất tương đối tốt.

6.4.3 Ảnh hưởng của Tỷ lệ Chia Train:Validation

Cả Naive Bayes và Random Forest đều cho thấy hiệu suất khá ổn định qua các tỷ lệ chia Train:Validation khác nhau (0.8 : 0.2, 0.7 : 0.3, 0.6 : 0.4). Sự thay đổi về Accuracy, Precision, Recall, F1-score (macro) là rất nhỏ (thường ở hàng phần nghìn).

⇒ **Nhận xét:** Điều này cho thấy với lượng dữ liệu trong tập validation (từ 20% đến 40% tổng dữ liệu), cả hai mô hình đều đưa ra đánh giá hiệu suất nhất quán. Có thể suy ra rằng lượng dữ liệu huấn luyện (từ 60% đến 80%) là đủ để các mô hình học được các mẫu cơ bản.

6.4.4 Phân tích Chi tiết Kết quả Phân loại theo Nhãn

Naive Bayes (Dữ liệu Gốc)

- Phân loại tốt nhất cho nhãn 3 (F1-score $\sim 0.86 - 0.87$) và nhãn 0 (F1-score ~ 0.76).
- Gặp khó khăn hơn với nhãn 1 (F1-score $\sim 0.56 - 0.57$) và nhãn 2 (F1-score $\sim 0.61 - 0.62$).

Random Forest (Dữ liệu Gốc)

- Phân loại rất tốt cho nhãn 3 (F1-score $\sim 0.91 - 0.92$) và nhãn 0 (F1-score $\sim 0.86 - 0.87$).
- Phân loại khá tốt cho nhãn 1 (F1-score $\sim 0.74 - 0.76$) và nhãn 2 (F1-score $\sim 0.77 - 0.78$).

⇒ **Nhận xét:**

- Cả hai mô hình đều dễ dàng phân biệt nhãn 3 và nhãn 0 hơn so với nhãn 1 và 2. Điều này có thể cho thấy các đặc trưng của nhãn 1 và 2 có sự chồng chéo nhiều hơn hoặc khó phân biệt hơn.
- Random Forest cải thiện đáng kể hiệu suất trên các nhãn khó (1 và 2) so với Naive Bayes. Ví dụ, với nhãn 1 (tỷ lệ 0.8 : 0.2, gốc), RF (0.74) tốt hơn NB (0.56) rất nhiều.

Với dữ liệu PCA

Hiệu suất trên tất cả các nhãn đều giảm đối với cả hai mô hình, nhưng mức giảm ở NB nghiêm trọng hơn. NB với PCA cho kết quả rất thấp ở nhãn 1 và 2 (F1-score chỉ còn $\sim 0.4x$).

6.4.5 Giải thích Sự Khác biệt về Hiệu suất

Naive Bayes

Là một thuật toán đơn giản, nhanh chóng, dựa trên định lý Bayes với giả định "ngây thơ" về tính độc lập của các đặc trưng.

- Khi giả định này bị vi phạm nhiều (thường xảy ra trong dữ liệu thực tế), hiệu suất có thể bị ảnh hưởng.
- Không có khả năng mô hình hóa các mối quan hệ phức tạp, phi tuyến giữa các đặc trưng.

Random Forest

Là một thuật toán học máy mạnh mẽ thuộc họ ensemble learning (học tập tổ hợp), cụ thể là bagging kết hợp với việc chọn ngẫu nhiên tập con đặc trưng khi xây dựng mỗi cây quyết định.

- Có khả năng nắm bắt các mối quan hệ phức tạp, phi tuyến và tương tác giữa các đặc trưng.
- Ít bị overfitting hơn so với một cây quyết định đơn lẻ.
- Robust với nhiễu và các đặc trưng không liên quan.

Sự vượt trội của RF cho thấy bộ dữ liệu này có thể chứa các mối quan hệ đặc trưng phức tạp mà NB không thể mô hình hóa hiệu quả.

6.4.6 Kết luận và Đề xuất

Lựa chọn mô hình

Random Forest là lựa chọn vượt trội hơn rõ rệt cho bài toán này dựa trên tất cả các chỉ số đánh giá. Nó mang lại độ chính xác và F1-score cao hơn đáng kể so với Naive Bayes.

Sử dụng PCA

Trong trường hợp này, PCA làm giảm hiệu suất của cả hai mô hình. Dữ liệu gốc cho kết quả tốt hơn. Có thể các thành phần chính được tạo ra không giữ được thông tin quan trọng cho việc phân loại bằng các mô hình này, hoặc bản chất của dữ liệu gốc phù hợp hơn. **Không nên sử dụng PCA** với bộ dữ liệu và các mô hình này trừ khi có lý do đặc biệt (ví dụ: yêu cầu giảm mạnh thời gian huấn luyện cho các mô hình phức tạp hơn nữa và chấp nhận sự suy giảm hiệu suất).

Tỷ lệ Train:Validation

Sự thay đổi tỷ lệ chia trong khoảng 0.8 : 0.2 đến 0.6 : 0.4 không ảnh hưởng lớn đến kết quả đánh giá tổng thể của các mô hình, cho thấy tính ổn định nhất định. Tỷ lệ 0.8 : 0.2 hoặc 0.7 : 0.3 có thể được ưu tiên để dành nhiều dữ liệu hơn cho việc huấn luyện.

Hướng cải thiện

- Đối với Random Forest, có thể thử tinh chỉnh thêm các siêu tham số (hyperparameters) như số lượng cây, độ sâu tối đa của cây, số lượng đặc trưng tối thiểu để chia nhánh, v.v., để có thể cải thiện thêm một chút hiệu suất.
- Cân xem xét kỹ hơn đặc điểm của nhãn 1 và nhãn 2, vì đây là các nhãn mà cả hai mô hình (đặc biệt là NB) gặp khó khăn hơn. Có thể cần thêm kỹ thuật tiền xử lý dữ liệu, feature engineering hoặc thu thập thêm dữ liệu/dặc trưng liên quan đến các nhãn này.
- Nếu mục tiêu là đạt hiệu suất cao nhất, nên tập trung vào tối ưu hóa Random Forest trên dữ liệu gốc.

Tóm lại, thực nghiệm cho thấy Random Forest là một mô hình mạnh mẽ và phù hợp hơn cho bộ dữ liệu này so với Naive Bayes, và việc áp dụng PCA đã gây tác động tiêu cực đến kết quả phân loại.

Chương 7

Kết luận

Báo cáo này đã trình bày quá trình ứng dụng các kỹ thuật học máy để dự báo và phân loại mức độ ô nhiễm không khí tại các thành phố, dựa trên một tập dữ liệu quan trắc chất lượng không khí tại Ấn Độ từ năm 2015 đến 2020. Quá trình này bao gồm các giai đoạn tiền xử lý dữ liệu, giảm chiều dữ liệu, phân cụm dữ liệu, xây dựng mô hình dự báo hồi quy (Random Forest, MLP) và mô hình phân loại (Naive Bayes, Random Forest Classifier).

7.1 Tóm tắt kết quả

- **Tiền xử lý dữ liệu:** Dữ liệu ban đầu đã trải qua các bước làm sạch cần thiết bao gồm xử lý giá trị thiếu (điền bằng giá trị trung bình theo thành phố hoặc trung bình toàn cục), chuẩn hóa dữ liệu số về khoảng [0, 1] bằng MinMaxScaler, và mã hóa one-hot cho các biến danh mục (City, AQI_Bucket). Các giá trị ngoại lai cũng được xử lý bằng phương pháp IQR, thay thế bằng giới hạn trên/dưới.
- **Giảm chiều dữ liệu:** Hai phương pháp PCA và t-SNE đã được áp dụng.
 - PCA cho thấy 6 thành phần đầu tiên giải thích khoảng 80% tổng phương sai, và cần 10-11 thành phần để giữ lại 95% phương sai.
 - t-SNE tỏ ra hiệu quả hơn trong việc trực quan hóa và phân tách các cụm dữ liệu so với PCA, đặc biệt khi dữ liệu có tính phi tuyến cao và mục tiêu là khám phá cấu trúc cục bộ.
- **Phân cụm dữ liệu:**
 - *K-Means:* Khi áp dụng trên dữ liệu đã giảm chiều bằng t-SNE với 4 cụm cho kết quả tốt nhất dựa trên Silhouette Score (0.3868) và các chỉ số DBI (0.8239), CHI (26069.2734). Phân cụm K-Means trên t-SNE giúp phân biệt rõ hơn các mức độ ô nhiễm.
 - *Gaussian Mixture Model (GMM):* Dữ liệu gốc (đã chuẩn hóa) với 6 cụm cho giá trị BIC thấp nhất. Tuy nhiên, khi đánh giá bằng DBI và CHI, GMM trên dữ liệu t-SNE (0.9094 và 22287.9004) cho thấy khả năng phân tách cụm tốt hơn so với dữ liệu gốc và PCA.

- *So sánh K-Means và GMM:* K-Means kết hợp với t-SNE cho thấy hiệu quả vượt trội trong việc phân tách các cụm dữ liệu một cách rõ ràng. GMM trên PCA lại cho thấy độ ổn định AQI trong các cụm tốt hơn, đặc biệt là ở các cụm ô nhiễm cao.

- **Mô hình dự báo hồi quy:**

- *Random Forest Regressor:*
 - * Hiệu suất tốt nhất đạt được khi sử dụng dữ liệu gốc (chưa qua giảm chiều), với tỷ lệ chia 7:3 và 200 cây, đạt MSE khoảng 887.38 và R^2 khoảng 0.9188.
 - * Việc áp dụng PCA trước khi huấn luyện làm giảm hiệu suất, t-SNE cho kết quả kém nhất.
 - * Hiện tượng overfitting được quan sát thấy ở hầu hết các cấu hình. Tăng số lượng cây có thể làm tăng nhẹ overfitting. Giới hạn độ sâu của cây (`max_depth`) là một phương pháp hiệu quả để giảm overfitting nhưng có thể dẫn đến underfitting.
- *Multi-Layer Perceptron (MLP):*
 - * Cấu hình với 3 lớp ẩn (100, 50, 25 nơ-ron), tối ưu bằng Adam và hàm mất mát MSE được sử dụng.
 - * MLP cũng cho kết quả tốt nhất trên dữ liệu gốc so với PCA và t-SNE. Cụ thể, với tỷ lệ chia 7:3, dữ liệu gốc cho MSE (Val) thấp hơn.
 - * Hiện tượng overfitting cũng xảy ra với MLP. Việc điều chỉnh hệ số regularization `alpha` cho thấy `alpha` nhỏ (ví dụ 0.001) giúp giảm overfitting mà không làm tăng đáng kể MSE trên tập validation.
 - * So với Random Forest, MLP thường cho MSE cao hơn một chút trên dữ liệu gốc với cùng điều kiện thử nghiệm.

- **Mô hình Phân loại:** Bài toán được đặt ra là phân loại AQI thành 4 khoảng giá trị.

- *Naive Bayes:* Cho kết quả tương đối, với dữ liệu gốc thường cho hiệu suất tốt hơn so với khi dùng PCA.
- *Random Forest Classifier:* Thường cho hiệu suất cao hơn Naive Bayes trên cả dữ liệu gốc và PCA.
 - * Ảnh hưởng của PCA: Việc áp dụng PCA trước khi huấn luyện Random Forest Classifier không nhất quán, đôi khi làm giảm nhẹ hiệu suất tổng thể (Macro F1-score).
 - * Ảnh hưởng của tỷ lệ chia Train:Validation: Tỷ lệ 8:2 và 7:3 thường cho kết quả tốt và ổn định.
- *So sánh chung:* Random Forest Classifier tỏ ra là mô hình mạnh mẽ hơn cho bài toán phân loại AQI trong nghiên cứu này.

7.2 Đánh giá

- **Ưu điểm:**

- Nghiên cứu đã thực hiện một cách có hệ thống các bước tiền xử lý, khám phá dữ liệu và xây dựng đa dạng các mô hình học máy.

- Việc so sánh chi tiết giữa các phương pháp giảm chiều, phân cụm, và các mô hình hồi quy, phân loại trên các dạng dữ liệu khác nhau mang lại cái nhìn sâu sắc về đặc điểm dữ liệu và hiệu quả của từng kỹ thuật.
- Các kết quả được trực quan hóa rõ ràng, giúp dễ dàng diễn giải và đưa ra nhận định.

- **Nhược điểm:**

- Các mô hình dự báo hồi quy (Random Forest, MLP) đều gặp phải vấn đề overfitting đáng kể, cho thấy mô hình học quá tốt trên tập huấn luyện nhưng khả năng tổng quát hóa trên dữ liệu mới còn hạn chế.
- Các kỹ thuật giảm chiều dữ liệu (PCA và t-SNE) chưa cho thấy hiệu quả trong việc cải thiện chất lượng các mô hình hồi quy và phân loại trong trường hợp này, thậm chí đôi khi còn làm giảm hiệu suất.

7.3 Hướng phát triển

Dựa trên các kết quả và đánh giá đã trình bày, một số hướng phát triển tiềm năng cho nghiên cứu này bao gồm:

- **Giải quyết vấn đề Overfitting cho mô hình hồi quy:**

- Tiếp tục tinh chỉnh sâu hơn các siêu tham số của Random Forest và MLP, có thể sử dụng các kỹ thuật tìm kiếm siêu tham số tự động (Grid Search, Random Search, Bayesian Optimization).
- Áp dụng các kỹ thuật regularization mạnh mẽ hơn (ví dụ: tăng giá trị alpha cho MLP một cách cẩn trọng, sử dụng Dropout cho mạng nơ-ron) hoặc các phương pháp ensemble phức tạp hơn.
- Thu thập thêm dữ liệu huấn luyện đa dạng hơn nếu có thể để cải thiện khả năng tổng quát hóa.

- **Khám phá các thuật toán dự báo và phân loại khác:**

- Đối với hồi quy: Thử nghiệm với Gradient Boosting (XGBoost, LightGBM, CatBoost), Support Vector Regression (SVR).
- Đối với phân loại: Thử nghiệm với Support Vector Machines (SVM), K-Nearest Neighbors (KNN), và các mô hình ensemble khác.
- Xem xét các kiến trúc mạng nơ-ron sâu hơn và chuyên biệt hơn cho dữ liệu chuỗi thời gian (nếu áp dụng) như LSTM, GRU.

- **Cải tiến Feature Engineering:** Nghiên cứu và xây dựng các đặc trưng mới từ dữ liệu thời gian (ví dụ: các đặc trưng trễ, trung bình trượt) và không gian (ví dụ: thông tin từ các trạm lân cận) có thể mang lại thông tin hữu ích hơn cho mô hình.

- **Ứng dụng kết quả phân cụm:**

- Sử dụng nhãn cụm được xác định bởi K-Means trên t-SNE hoặc GMM trên PCA như một đặc trưng đầu vào cho các mô hình dự báo và phân loại để xem xét liệu nó có cải thiện hiệu suất không.

- **Đánh giá chéo (Cross-Validation):** Sử dụng các kỹ thuật đánh giá chéo (ví dụ: K-Fold Cross-Validation) để có được ước lượng hiệu suất mô hình ổn định và đáng tin cậy hơn, đặc biệt khi tinh chỉnh siêu tham số.
- **Đánh giá trên tập dữ liệu lớn hơn và đa dạng hơn:** Kiểm tra tính tổng quát của các phương pháp và kết quả trên các bộ dữ liệu ô nhiễm không khí từ nhiều khu vực và quốc gia khác nhau, hoặc dữ liệu với tần suất lấy mẫu cao hơn.

Tài liệu tham khảo

- [1] Hiren Vora. *city_day - Air Pollution Dataset*. <https://www.kaggle.com/datasets/hirenvora/city-daycsv>. 2020.
- [2] Leo Breiman. ?Random Forests? **in***Machine Learning*: 45.1 (2001), **pages** 5–32.
- [3] James Lever, Martin Krzywinski **and** Naomi Altman. ?Principal Component Analysis? **in***Nature Methods*: 14.7 (2017), **pages** 641–642.
- [4] Jason Shlens. *A Tutorial on Principal Component Analysis*. <https://arxiv.org/abs/1404.1100>. 2014.