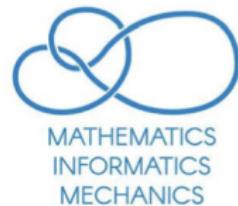


Ứng dụng học máy dự báo mức độ ô nhiễm không khí tại các thành phố

Đỗ Quốc An¹ Phạm Thị Duyên² Trần Kiều Hạnh³



Mục Lục

- ① Giới thiệu (Giữa kì)
- ② Kiến thức lý thuyết (Giữa kì)
- ③ Dữ liệu (Giữa kì)
- ④ Phân cụm dữ liệu
- ⑤ Thực nghiệm và kết quả (Giữa kì)
- ⑥ Mô hình phân loại
- ⑦ Kết luận và hướng phát triển

Mục lục

- 1 Giới thiệu (Giữa kì)
- 2 Kiến thức lý thuyết (Giữa kì)
- 3 Dữ liệu (Giữa kì)
- 4 Phân cụm dữ liệu
- 5 Thực nghiệm và kết quả (Giữa kì)
- 6 Mô hình phân loại
- 7 Kết luận và hướng phát triển

Giới thiệu

Lý do chọn đề tài

- Ô nhiễm không khí là vấn đề môi trường nghiêm trọng toàn cầu.
- Ảnh hưởng đến sức khỏe, gây bệnh hô hấp, tim mạch và ung thư.
- AQI giúp đánh giá mức độ ô nhiễm không khí.
- Việc đo AQI liên tục ở mọi nơi vẫn còn nhiều hạn chế.
- Cần xây dựng mô hình dự báo AQI từ dữ liệu quan trắc để cải thiện theo dõi chất lượng không khí.

Mục tiêu

- Xây dựng mô hình Random Forest, MLP dự báo AQI.
- So sánh hiệu suất trên dữ liệu gốc, PCA, t-SNE qua RMSE, MAE, R².
- Phân tích yếu tố ảnh hưởng: tầm quan trọng, tương quan, phần dư.
- Đề xuất cải tiến mô hình, đặc biệt ở ngưỡng ô nhiễm cao.

Giới thiệu

Phạm vi nghiên cứu

- **Dữ liệu:** city_day.csv, gồm 26 thành phố Ấn Độ (Ahmedabad, Delhi, Mumbai...), giai đoạn 2015–2020.
- **Phương pháp:** PCA, t-SNE, Random Forest, MLP.
- **Phân tích theo:** không gian (thành phố), thời gian (ngày).

Phương pháp nghiên cứu

- **Tiền xử lý:** xử lý thiếu, ngoại lai, chuẩn hóa, one-hot, đặc trưng thời gian.
- **Giảm chiều:** PCA (giữ phương sai chính), t-SNE (trực quan hóa 2D).
- **Mô hình:** RF (100 cây), MLP (1 lớp ẩn: 100, Adam, MSE).
- **Đánh giá:** RMSE, MAE, R^2 với tỉ lệ train:test = 8:2, 7:3, 6:4.
- **Phân tích:** tương quan, phân phối, phần dư, tầm quan trọng đặc trưng.

Mục lục

- 1 Giới thiệu (Giữa kì)
- 2 Kiến thức lý thuyết (Giữa kì)
- 3 Dữ liệu (Giữa kì)
- 4 Phân cụm dữ liệu
- 5 Thực nghiệm và kết quả (Giữa kì)
- 6 Mô hình phân loại
- 7 Kết luận và hướng phát triển

t-SNE (t-Distributed Stochastic Neighbor Embedding)

Tổng quan về t-SNE

- Phương pháp giảm chiều phi tuyến, thường dùng để trực quan hóa dữ liệu cao chiều (2D/3D).
- Bảo toàn cấu trúc cục bộ và quan hệ lân cận giữa các điểm dữ liệu.
- Hiệu quả trong việc phát hiện và quan sát cấu trúc cụm (clusters).

Nguyên lý hoạt động

- Mô hình hóa quan hệ điểm ở không gian cao chiều bằng phân phối Gaussian.
- Mô hình hóa ở không gian thấp chiều bằng phân phối t-Student.

Tham số quan trọng: Perplexity, Learning rate, Số chiều đích.

Random Forest (RF)

Random Forest là gì?

- Mô hình học máy tổ hợp, dùng nhiều cây quyết định độc lập.
- Dự đoán: Trung bình (hồi quy) hoặc đa số phiếu (phân loại).
- Ưu điểm: Xử lý dữ liệu nhiễu, nhiều đặc trưng; chống quá khớp.
- Ứng dụng: Phân loại, hồi quy, phát hiện bất thường.

Nguyên lý hoạt động

- **Bagging:** Lấy mẫu ngẫu nhiên có hoàn lại tạo tập con dữ liệu.
- **Ngẫu nhiên đặc trưng:** Chọn ngẫu nhiên tập con đặc trưng ở mỗi nút.
- **Tổng hợp:** Trung bình (hồi quy) hoặc biểu quyết (phân loại).

Mục lục

- 1 Giới thiệu (Giữa kì)
- 2 Kiến thức lý thuyết (Giữa kì)
- 3 Dữ liệu (Giữa kì)
- 4 Phân cụm dữ liệu
- 5 Thực nghiệm và kết quả (Giữa kì)
- 6 Mô hình phân loại
- 7 Kết luận và hướng phát triển

Tổng quan về tập dữ liệu

Nguồn: Quan trắc chất lượng không khí Ấn Độ (2015-2020).

- Tệp: city_day.csv (29,532 dòng x 16 cột)
- Cột chính: Date, City, PM2.5, PM10, NO, NO2, NOx, NH3, CO, SO2, O3, Benzene, Toluene, Xylene, AQI, AQI_Bucket.

Kiểu dữ liệu: Chuỗi, số nguyên, số thực.

Tiền xử lý dữ liệu

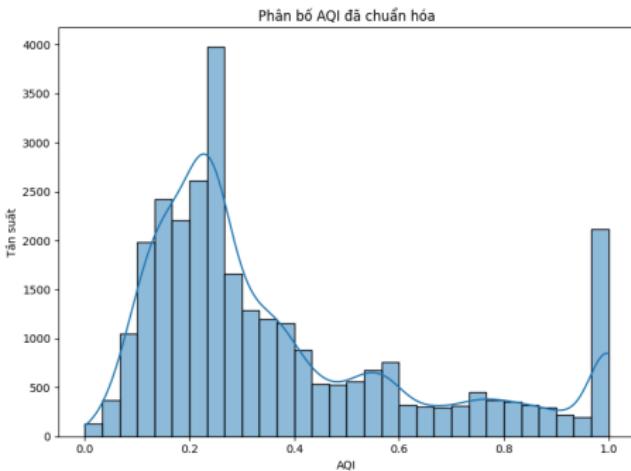
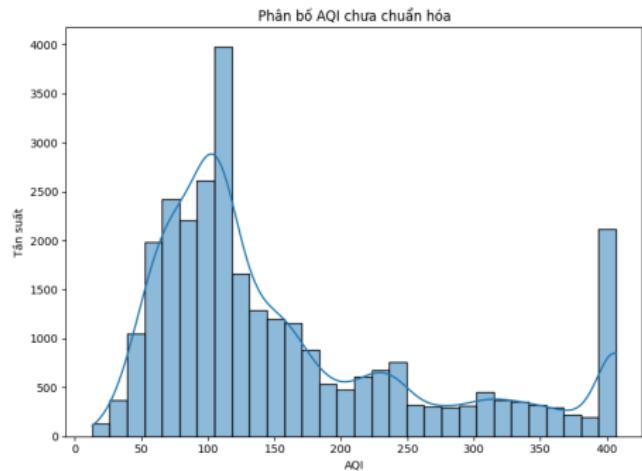
Mục tiêu: Làm sạch dữ liệu, xử lý giá trị thiếu và ngoại lai, chuẩn hóa để dữ liệu sẵn sàng cho mô hình học máy.

- **Đọc và Hiểu Dữ Liệu:** Kiểm tra kích thước, kiểu dữ liệu và các giá trị thiếu.
- **Xử Lý Giá Trị Thiếu:** Điền giá trị thiếu bằng trung bình theo nhóm hoặc toàn bộ dữ liệu.
- **Xử Lý Giá Trị Ngoại Lai:** Phát hiện và thay thế giá trị ngoại lai bằng IQR (Interquartile Range).
- **Chuyển Đổi Dữ Liệu:** Sử dụng One-Hot Encoding cho biến phân loại và chuẩn hóa dữ liệu bằng MinMaxScaler.

Kết quả:

- Dữ liệu không còn giá trị thiếu.
- Dữ liệu đã được chuẩn hóa trong phạm vi $[0, 1]$.
- Dữ liệu sẵn sàng cho mô hình học máy.

Tiền xử lý dữ liệu



Chuẩn hóa dữ liệu và đánh giá thành phần chính - PCA

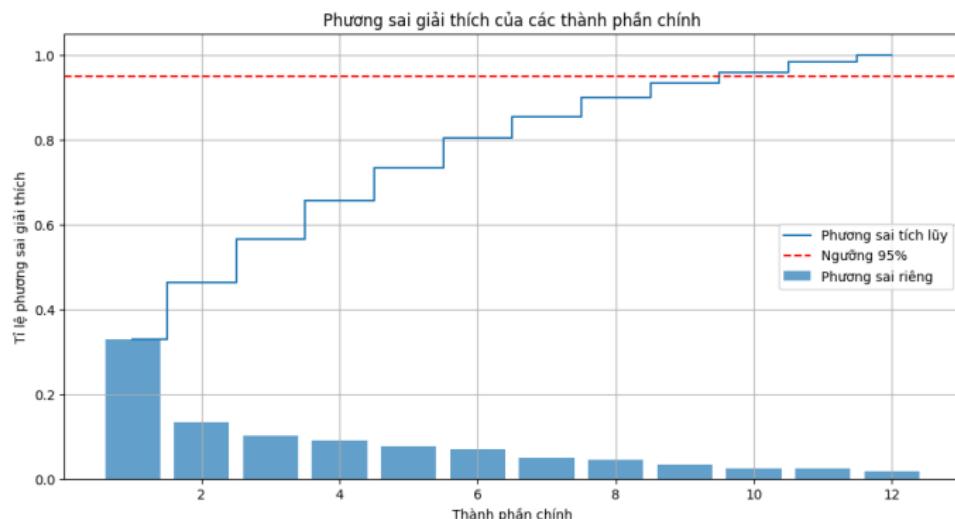
Mục tiêu

- Chuẩn hóa các biến đầu vào để các đặc trưng có cùng đơn vị đo.
- Giảm chiều dữ liệu mà vẫn giữ lại phần lớn thông tin quan trọng.

Các bước thực hiện

- ① **Chuẩn hóa dữ liệu:** Dữ liệu chuẩn hóa với trung bình 0 và độ lệch chuẩn 1. Ví dụ: $PM2.5$, $PM10$, NO , $NO2$, ...
- ② **Thực hiện PCA:** Dùng PCA để giảm số lượng thành phần mà vẫn giữ được thông tin quan trọng.
- ③ **Đánh giá phương sai giải thích:** Phương sai của từng thành phần:

Chuẩn hóa dữ liệu và đánh giá thành phần chính - PCA



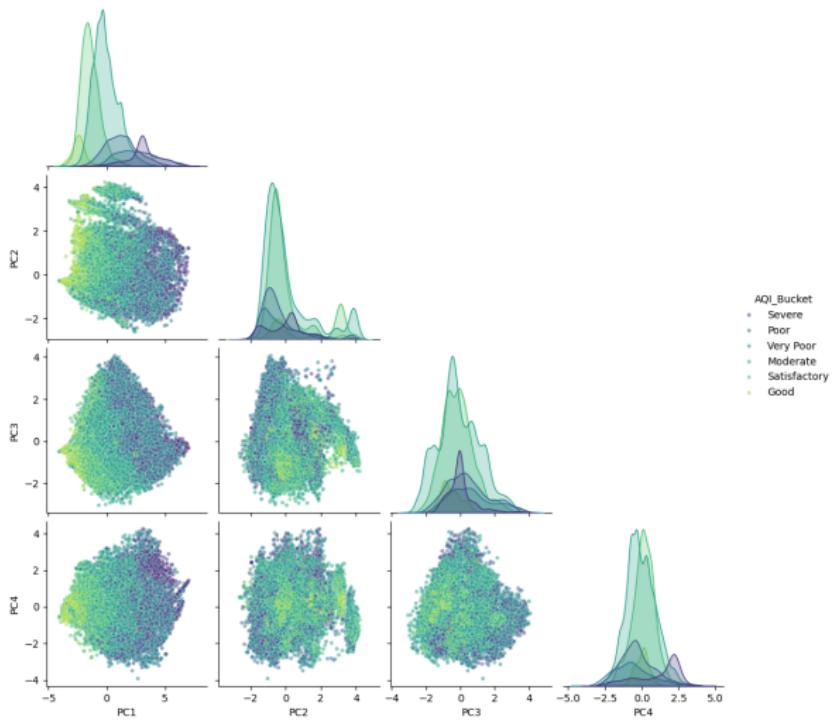
- ④ **Xác định số thành phần tối ưu:** Sử dụng 10 thành phần chính để giữ $\geq 95\%$ phương sai tích lũy.

Trực quan hóa dữ liệu theo các cặp thành phần chính - PCA

- Trích xuất 4 thành phần chính đầu tiên từ dữ liệu đã chuẩn hóa, cho phép giữ lại phần lớn thông tin quan trọng.
- Gắn nhãn mức độ ô nhiễm (AQI_Bucket) vào dữ liệu PCA để quan sát sự phân bố theo các mức ô nhiễm.
- Xây dựng các biểu đồ phân tán giữa các cặp thành phần chính, giúp quan sát cấu trúc dữ liệu và khả năng phân tách nhóm AQI trong không gian mới.

Trực quan hóa dữ liệu theo các cặp thành phần chính - PCA

Ma trận biểu đồ phân tán cho 6 thành phần chính đầu tiên



Trực quan hóa dữ liệu bằng *t-SNE* (2 thành phần)

Mục tiêu:

- Giảm chiều dữ liệu phi tuyến để trực quan hóa dữ liệu đa chiều trên mặt phẳng 2D.
- Quan sát phân bố dữ liệu theo mức độ ô nhiễm không khí (AQI_Bucket).

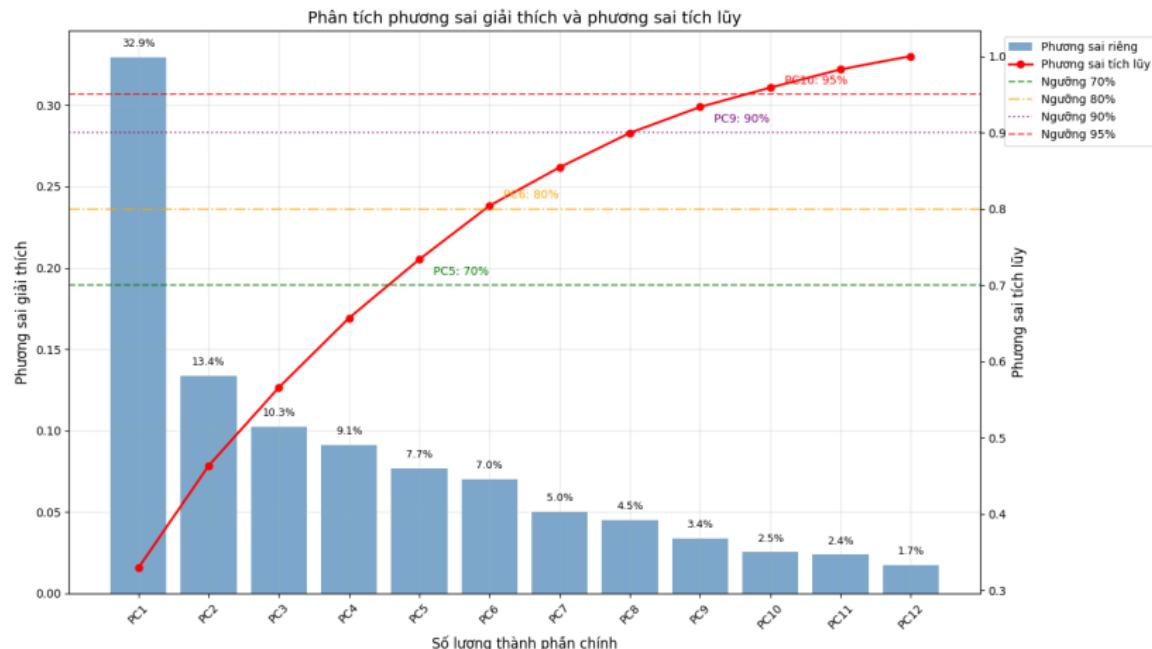
Cấu hình t-SNE:

- `n_components = 2, perplexity = 30, learning_rate = 200`
- `n_iter = 3000, random_state = 42`

Trực quan hóa dữ liệu bằng t -SNE (2 thành phần)



Lượng thông tin được bảo tồn - PCA

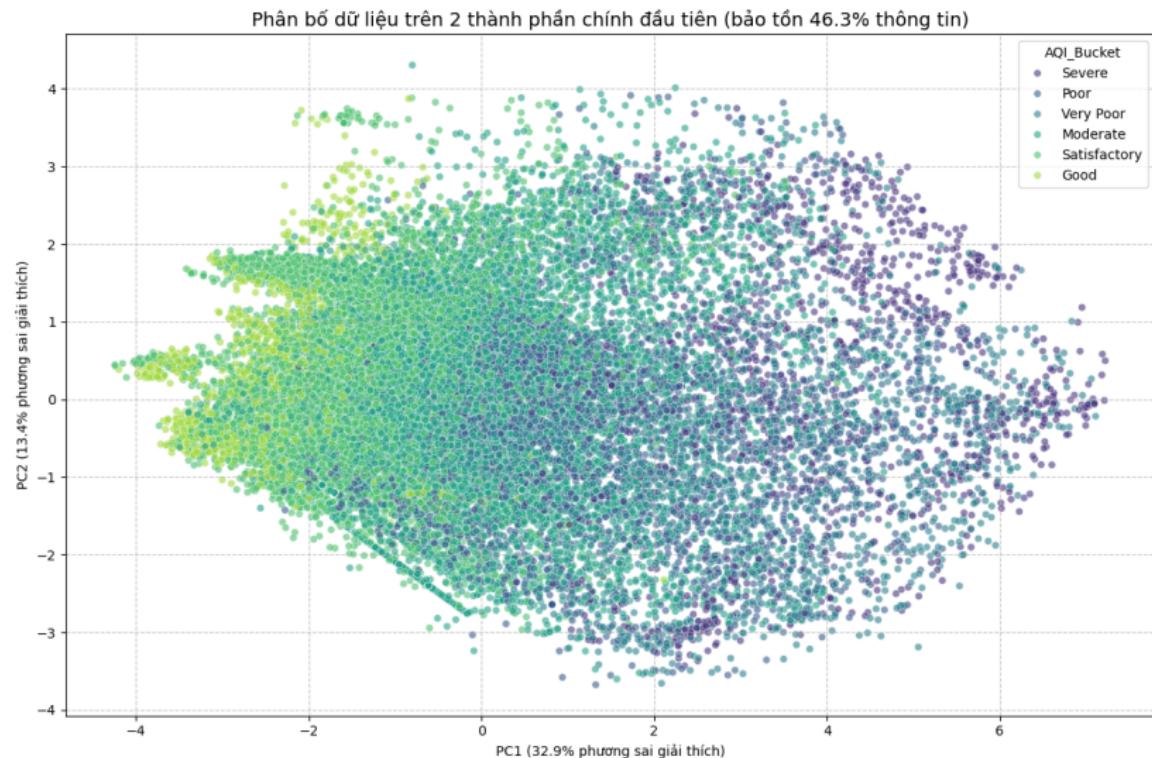


Lượng thông tin được bảo tồn - PCA

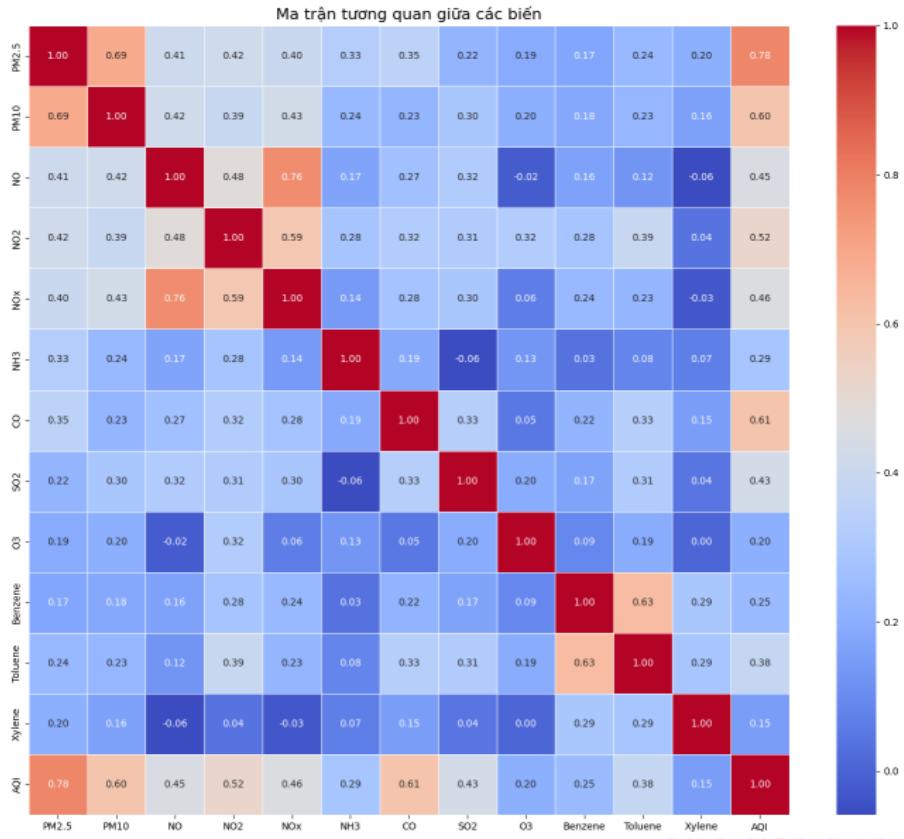
Số lượng thành phần chính cần thiết để đạt các ngưỡng phương sai khác nhau:

Ngưỡng phương sai	Số lượng thành phần chính cần thiết	Tỉ lệ giảm chiều (%)
0	50%	3
1	60%	4
2	70%	5
3	80%	6
4	90%	9
5	95%	10
6	99%	12

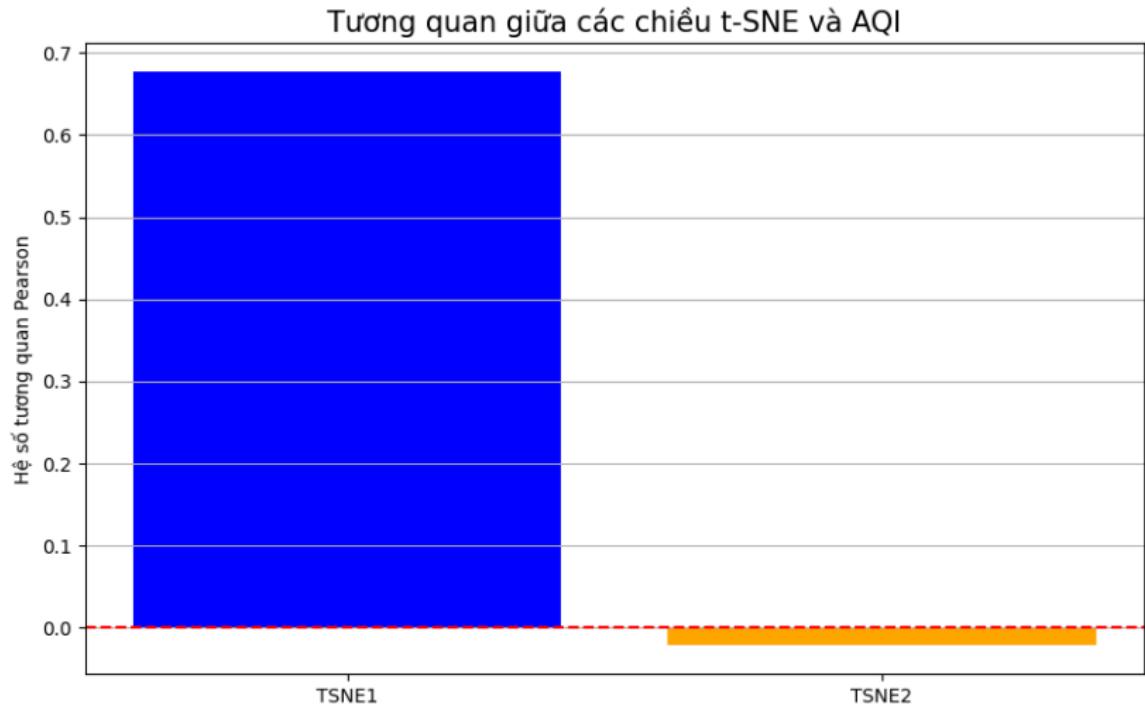
Lượng thông tin được bảo tồn - PCA



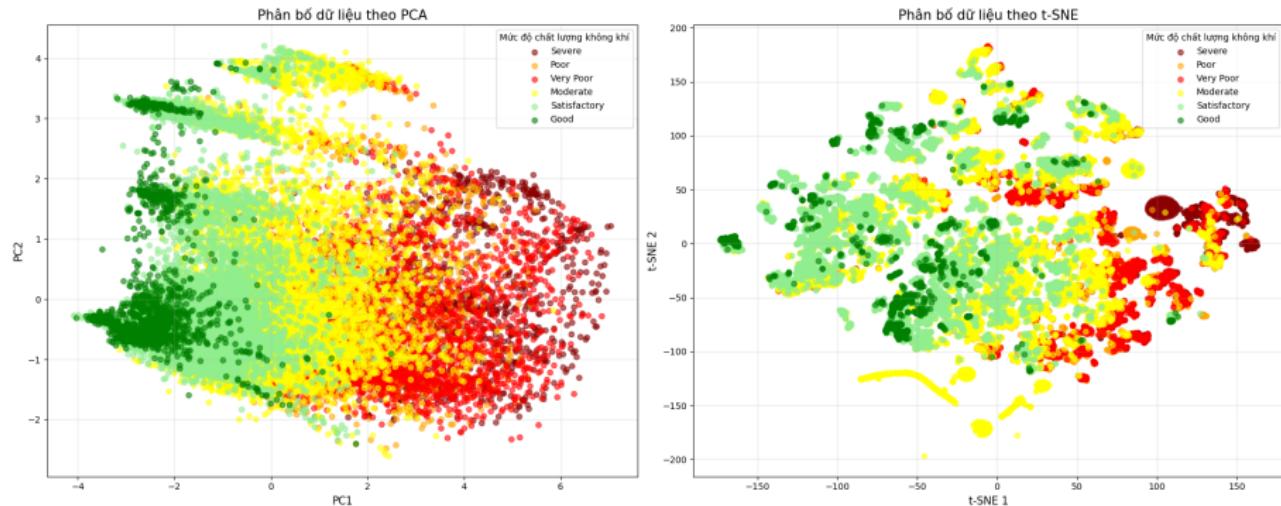
Trực quan hóa mối quan hệ các biến với AQI - PCA



Trực quan hóa mối quan hệ các biến với AQI - t-SNE



So sánh PCA và t-SNE



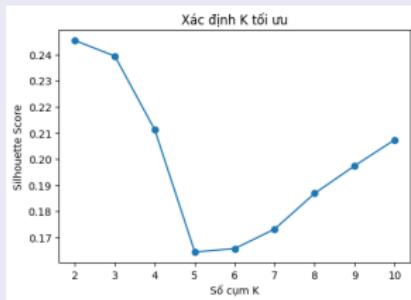
Mục lục

- 1 Giới thiệu (Giữa kì)
- 2 Kiến thức lý thuyết (Giữa kì)
- 3 Dữ liệu (Giữa kì)
- 4 Phân cụm dữ liệu
- 5 Thực nghiệm và kết quả (Giữa kì)
- 6 Mô hình phân loại
- 7 Kết luận và hướng phát triển

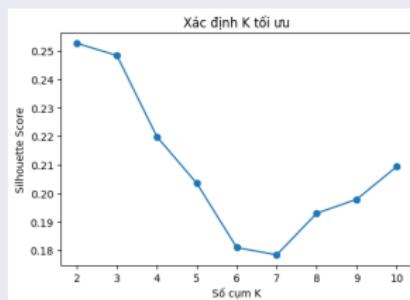
K_Means

Phân cụm K-Means & Lựa chọn K

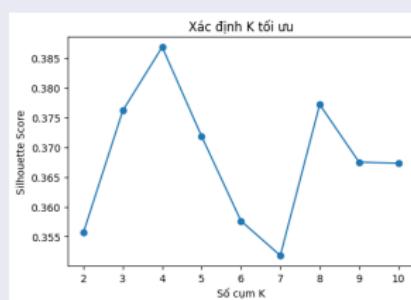
- Dữ liệu đầu vào đã được chuẩn hóa, bỏ trường đầu ra AQI.
- K-Means chạy với $K = 2 \dots 10$ trên 3 dạng dữ liệu:
 - Dữ liệu gốc (chuẩn hóa)
 - PCA (10 thành phần chính)
 - t-SNE (2 chiều)
- Lựa chọn số cụm dựa trên **Silhouette Score**.



(a) Dữ liệu gốc



(b) PCA



(c) t-SNE

Phân cụm K-Means & Lựa chọn K

Bảng: Số cụm K tối ưu và Silhouette Score

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

- t-SNE với $K = 4$ cho kết quả phân cụm tốt nhất.

Dánh giá mối quan hệ giữa mẫu đầu vào và đầu ra

Dánh giá chất lượng phân cụm

- Sử dụng 2 chỉ số:
 - Davies-Bouldin Index (DBI): càng nhỏ càng tốt.
 - Calinski-Harabasz Index (CHI): càng lớn càng tốt.

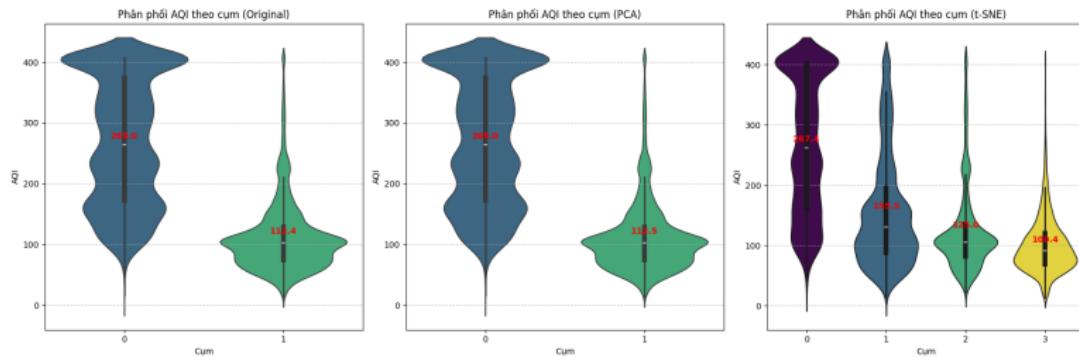
Bảng: Chất lượng phân cụm

Phương pháp	DBI ↓	CHI ↑
Dữ liệu gốc	1.8565	7511.38
PCA	1.8004	7949.71
t-SNE	0.8239	26069.27

- t-SNE phân cụm rõ nhất, đồng nhất nhất.

Dánh giá mối quan hệ giữa mẫu đầu vào và đầu ra

Phân phối chỉ số AQI trong các cụm

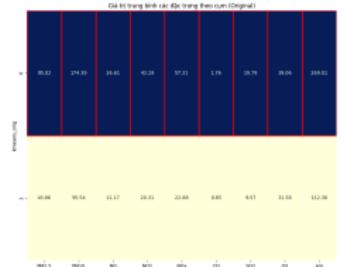


Hình: Phân phối chỉ số AQI trong từng cụm

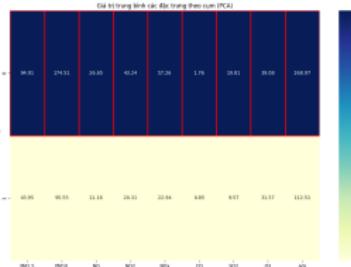
- Violin plot cho thấy phân phối AQI:
 - Original, PCA: 2 cụm, AQI trung bình cụm 0 cao (269), cụm 1 thấp (112).
 - t-SNE: 4 cụm, AQI giảm dần từ cụm 0 đến cụm 3 (267.4 → 100.4).
- t-SNE giúp phân biệt các mức độ ô nhiễm chi tiết hơn.

Dánh giá mối quan hệ giữa mẫu đầu vào và đầu ra

Mối quan hệ đặc trưng trong từng cụm



(a) Original



(b) PCA



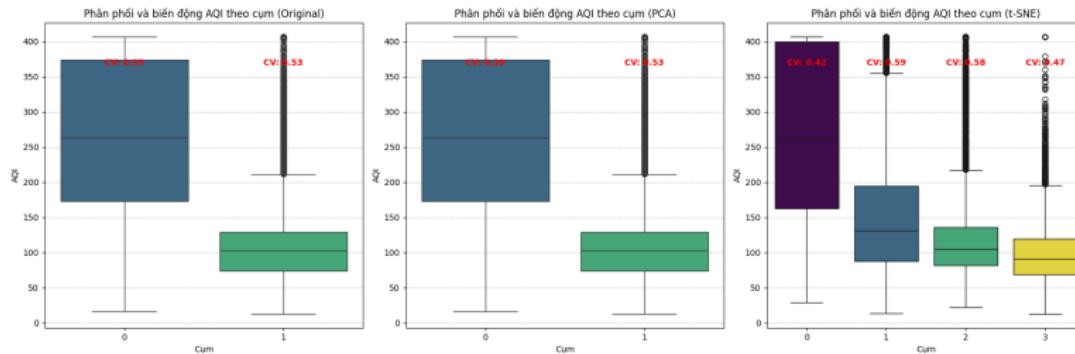
(c) t-SNE

Hình: Giá trị trung bình các đặc trưng theo cụm

- Heatmap giá trị trung bình các đặc trưng (chuẩn hóa):
 - Original, PCA: Các cụm tương đồng, phân tách kém.
 - t-SNE: Các cụm khác biệt rõ rệt, đa dạng màu sắc.
- K-Means hoạt động hiệu quả nhất trên không gian t-SNE.

Đánh giá mối quan hệ giữa mẫu đầu vào và đầu ra

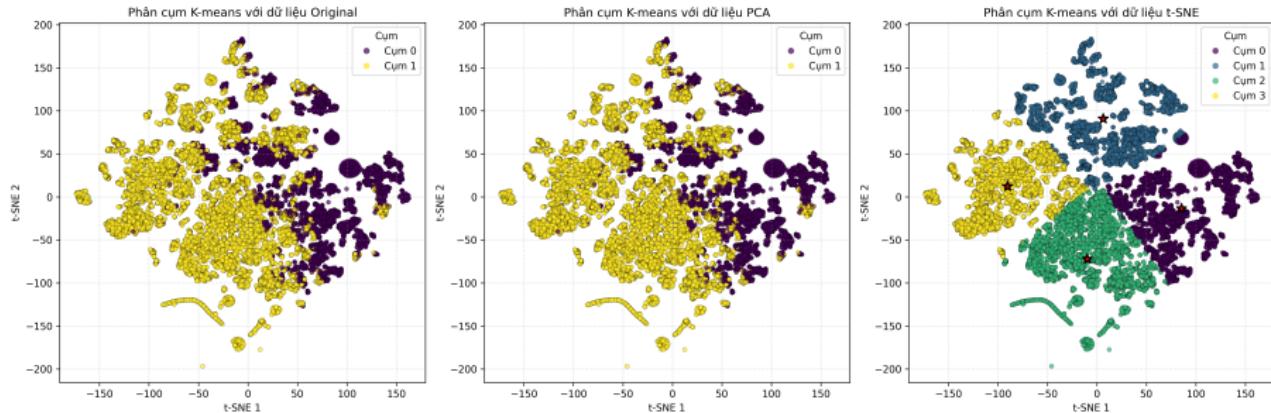
Biến động AQI trong từng cụm



Hình: Phân phối và biến động AQI theo cụm

- Đánh giá bằng hệ số biến thiên ($CV = \text{độ lệch chuẩn} / \text{trung bình}$).
- CV thấp: cụm đồng nhất về AQI.
- CV cao: ô nhiễm không đồng đều, có thể do nhiễu.
- Phân cụm giúp nhận diện các mức độ ô nhiễm khác nhau.

Trực quan hóa kết quả phân cụm

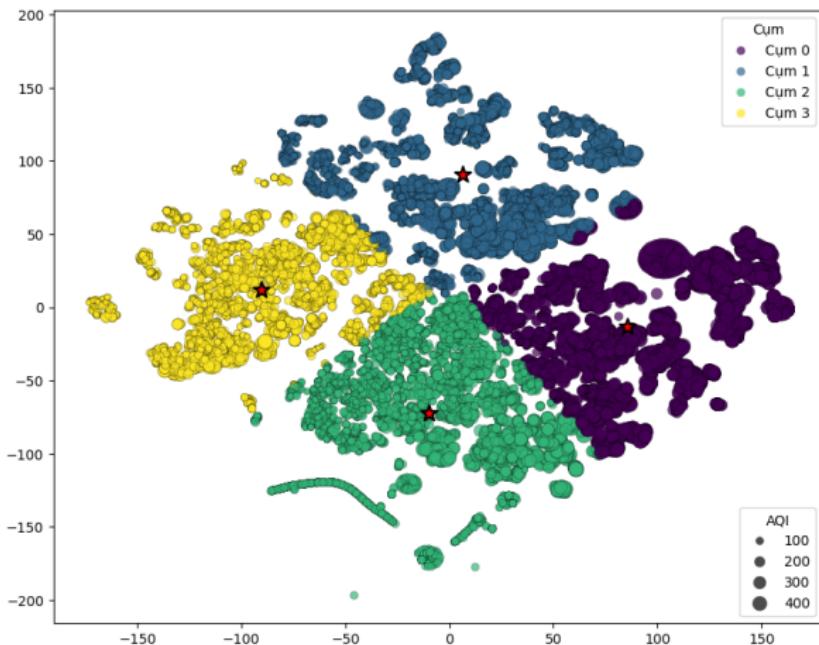


Hình: So sánh phân cụm K-Means trên ba không gian dữ liệu: gốc, PCA và t-SNE

- Original: phân cụm kém rõ ràng.
- PCA: phân cụm tốt hơn Original.
- t-SNE: ranh giới cụm rõ nét nhất, phân tách phi tuyến tốt.

Trực quan hóa kết quả phân cụm

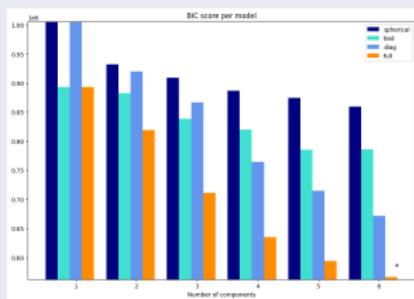
Trực quan hóa AQI trên không gian t-SNE



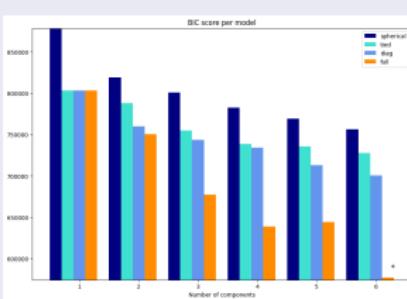
Hình: Biểu diễn chỉ số AQI theo kích thước điểm trong không gian t-SNE

Phân cụm GMM và lựa chọn số cụm phù hợp

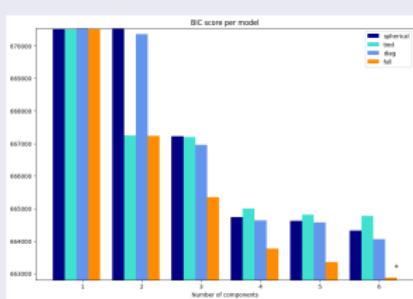
- Dữ liệu đầu vào đã được chuẩn hóa, loại bỏ trường AQI và huấn luyện GMM trên ba dạng: dữ liệu gốc, PCA (10 thành phần chính), và t-SNE (2 chiều).
- Kết hợp số cụm K từ 1 đến 6 với 4 loại hiệp phương sai: spherical, tied, diag, full, mỗi tổ hợp tạo ra một mô hình. Mô hình có **BIC** nhỏ nhất được chọn.



(a) Original



(b) PCA



(c) t-SNE

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

Kết quả số cụm tối ưu

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

Dữ liệu	K tối ưu	BIC
Chuẩn hóa	6	567002.8191
PCA (10 thành phần)	6	577498.0997
t-SNE (2 chiều)	6	662883.9523

Nhận xét: 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.

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

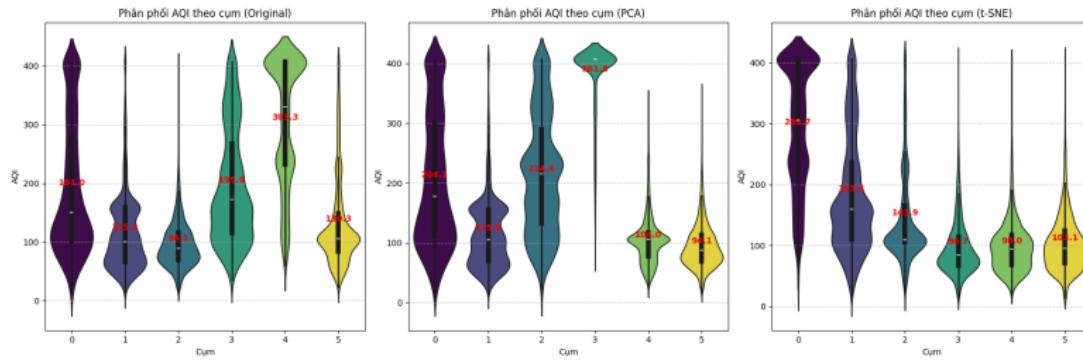
Sử dụng 2 chỉ số đánh giá phân cụm : DBI (càng nhỏ càng tốt) và CHI (càng lớn càng tốt).

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

Dữ liệu	DBI ↓	CHI ↑
Chuẩn hóa	2.6908	2591.6972
PCA	2.2503	3188.8180
t-SNE	0.9094	22287.9004

Nhận xét: t-SNE đạt chỉ số DBI thấp nhất và chỉ số CHI 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

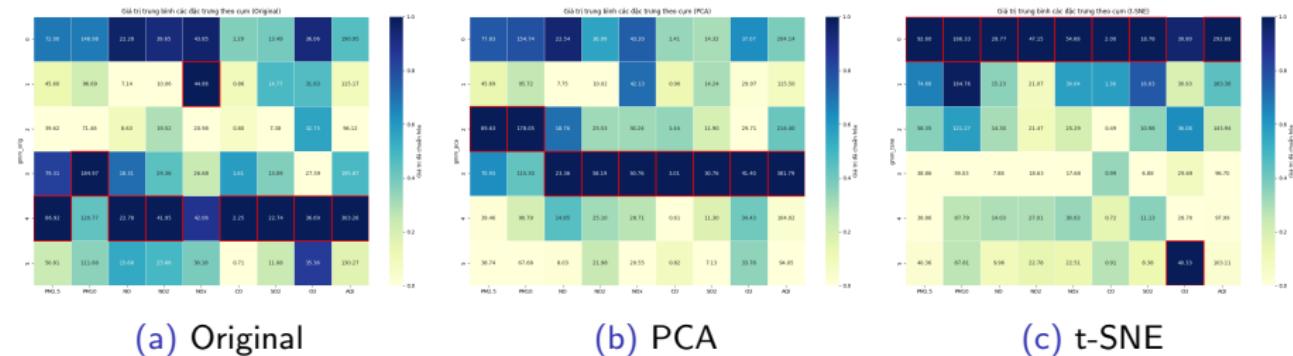


Hình: Phân phối chỉ số AQI theo các cụm trên ba loại dữ liệu

Nhận xét: PCA cân bằng và mở rộng dữ liệu; t-SNE phân tách tốt cụm AQI thấp-trung bình.

- Original: AQI 96.1–303.3, có giá trị cao bất thường.
- PCA: AQI 94.1–381.8, phân bố đều, cụm 3 cao nhất → dữ liệu mượt và mở rộng.
- t-SNE: AQI 96.7–292.7, tập trung thấp-trung bình, cụm 2 cao nhất → phân tách rõ.

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

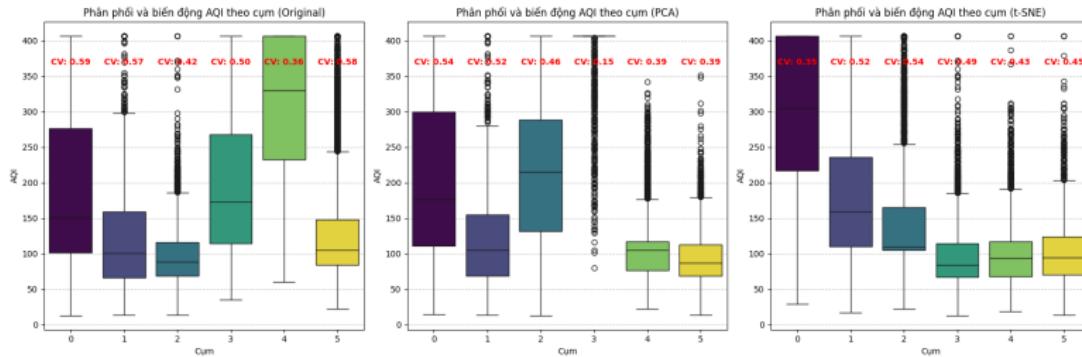


Hình: Heatmap thể hiện giá trị trung bình của các đặc trưng ô nhiễm trong từng cụm GMM trên ba loại dữ liệu.

Nhận xét:

- Tương quan: PM2.5, PM10, AQI thường tăng/giảm cùng nhau; NO, NO₂, NOx cũng đi kèm.
- Đặc trưng cụm: AQI cao → PM2.5, PM10, SO₂ cao; AQI thấp → đa số chỉ số thấp.
- Giảm chiều: Original, PCA phân cụm rõ; t-SNE tập trung cấu trúc cục bộ → cụm ít phân biệt.

Dánh giá mức độ biến động AQI trong từng cụm

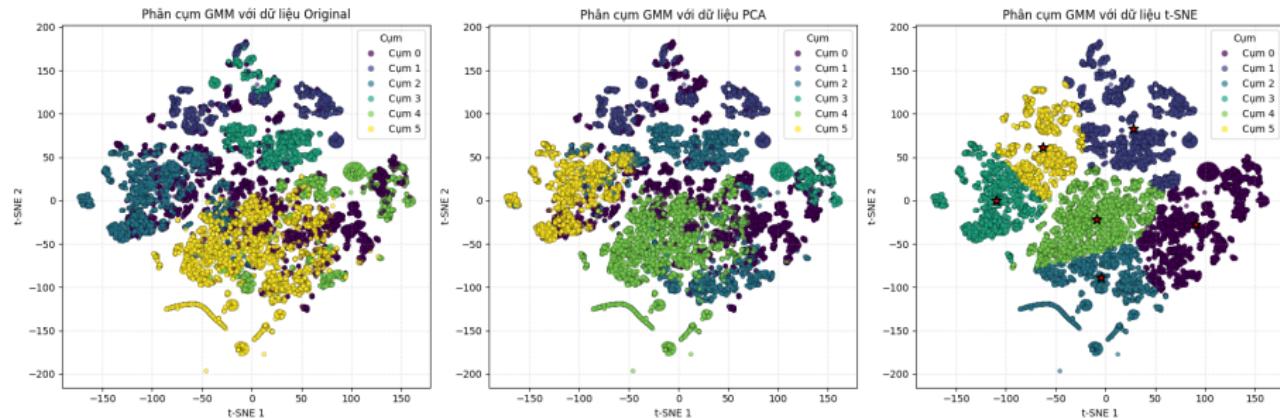


Hình: Phân bố AQI trong các cụm được phân cụm bằng GMM trên ba loại dữ liệu

Nhận xét: PCA tốt hơn trong tách cụm ổn định; t-SNE tạo cụm đồng đều về CV.

- Original: CV dao động 0.36–0.59; cụm AQI cao có CV thấp → ô nhiễm ổn định.
- PCA: CV thấp hơn; cụm 3 có CV = 0.15 → AQI ổn định nhất.
- t-SNE: CV 0.35–0.54 đều, không có cụm nổi bật về biến động.

Trực quan hóa kết quả phân cụm

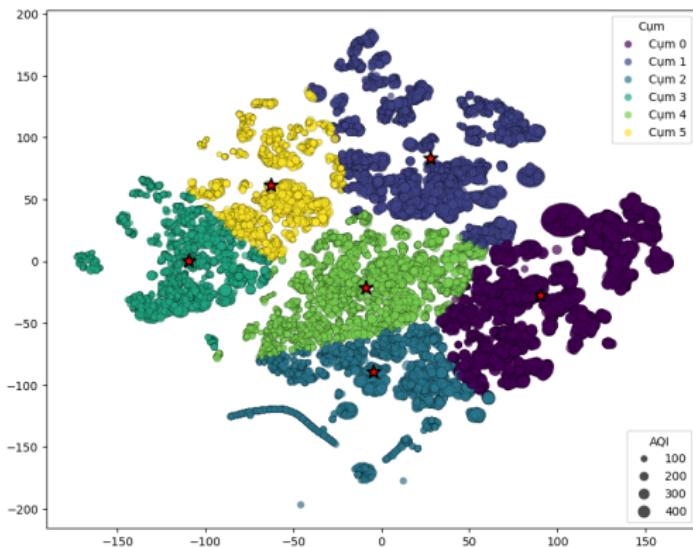


Hình: So sánh phân cụm GMM trên ba không gian dữ liệu: gốc, PCA và t-SNE

Nhận xét:

- Original: Cụm chồng chéo, phân tách kém.
- PCA: Cụm rõ hơn, vẫn có chồng lấn nhẹ.
- t-SNE: Cụm tách biệt rõ, phù hợp trực quan hóa.

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



Hình: Trực quan phân cụm GMM trên không gian t-SNE 2D với kích thước điểm biểu thị giá trị AQI tương ứng

Nhận xét:

- Cụm 0 có AQI cao, thể hiện ô nhiễm nặng.
- Cụm 3, 4 có AQI thấp, chất lượng không khí tốt.
- Cụm 1, 5 phân bố AQI trung bình, đồng đều.

Kết luận: t-SNE làm rõ phân vùng AQI và mối liên hệ giữa cụm và ô nhiễm.

So sánh K-Means và GMM

- **Chất lượng cụm:**

- K-Means + t-SNE: DBI thấp, CHI cao \Rightarrow cụm rõ ràng, tách biệt.
- GMM + PCA: ổn định hơn, phù hợp dữ liệu phức tạp.

- **Phân phối AQI:**

- K-Means + t-SNE: AQI giảm dần theo cụm ($267 \rightarrow 100$), tách biệt tốt.
- GMM + PCA: AQI cao, dao động lớn nhưng nội cụm ổn định.

- **Đặc trưng và độ ổn định (CV):**

- K-Means: rõ sự khác biệt (PM2.5, PM10), ổn định hơn ở AQI thấp.
- GMM: cụm ô nhiễm cao nổi bật, PCA giúp làm mượt AQI.

- **Trực quan hóa:**

- K-Means + t-SNE: ranh giới sắc nét, phân tích trực quan tốt.
- GMM: hiệu quả hơn trên PCA và t-SNE, rõ nhất ở t-SNE..

Kết luận:

K-Means + t-SNE \Rightarrow phân tích chi tiết AQI thấp-trung bình.

GMM + PCA \Rightarrow phù hợp dữ liệu ô nhiễm cao, phức tạp.

Mục lục

- 1 Giới thiệu (Giữa kì)
- 2 Kiến thức lý thuyết (Giữa kì)
- 3 Dữ liệu (Giữa kì)
- 4 Phân cụm dữ liệu
- 5 Thực nghiệm và kết quả (Giữa kì)
- 6 Mô hình phân loại
- 7 Kết luận và hướng phát triển

Thiết lập mô hình hồi quy

Mục tiêu: Dự đoán AQI từ các chỉ số ô nhiễm.

Mô hình:

- **MLP:** hidden_layer = 100, activation = 'relu', solver = 'adam'
- **Random Forest:** n_estimators=50, 100, 200

Dữ liệu: 12 đặc trưng (dữ liệu gốc), 10 đặc trưng (PCA – 95% phương sai), 2 đặc trưng (t-SNE – trực quan hóa).

Tỷ lệ Train:Validation: 8:2, 7:3, 6:4.

Chỉ số đánh giá: MSE, RMSE, MAE, R².

Hiệu suất mô hình MLP trên các loại dữ liệu

Bảng: So sánh hiệu suất MLP: Dữ liệu gốc, PCA và t-SNE

Loại dữ liệu	Tỷ lệ Train:Val	MSE	RMSE	MAE	R ²
Dữ liệu gốc	8:2	1239.90	35.21	22.85	0.8858
Dữ liệu gốc	7:3	1272.75	35.68	23.57	0.8836
Dữ liệu gốc	6:4	1272.77	35.68	23.31	0.8837
Dữ liệu PCA	8:2	1159.63	34.05	21.57	0.8932
Dữ liệu PCA	7:3	1124.81	33.54	21.38	0.8971
Dữ liệu PCA	6:4	1175.40	34.28	21.87	0.8926
Dữ liệu t-SNE	8:2	5029.95	70.92	53.83	0.5367
Dữ liệu t-SNE	7:3	4605.10	67.86	51.32	0.5788
Dữ liệu t-SNE	6:4	4589.90	67.75	51.27	0.5807

- MLP trên PCA có hiệu suất tốt nhất (MSE thấp nhất, R^2 cao nhất 0.8971), giúp MLP học hiệu quả hơn.
- Dữ liệu gốc ổn định, t-SNE có hiệu suất kém nhất, MSE cao và R^2 thấp, không phù hợp cho dự đoán với MLP.

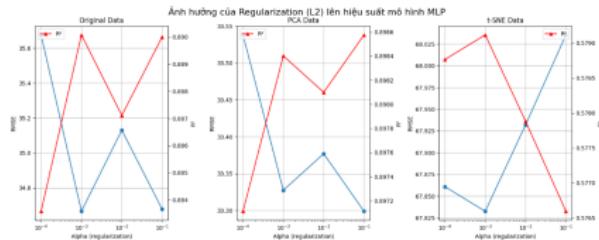
Đánh giá Overfitting của MLP



Hình: MSE giữa tập huấn luyện và validation trên MLP.

- Dữ liệu gốc và t-SNE có khoảng cách MSE lớn giữa train và validation \Rightarrow mô hình **overfitting**.
- PCA giúp giảm chênh lệch MSE \Rightarrow giảm overfitting nhưng chưa khắc phục hoàn toàn.
- Mô hình cần thêm regularization, giảm độ phức tạp hoặc bổ sung dữ liệu để cải thiện khả năng tổng quát.

Áp dụng Regularization để giảm Overfitting

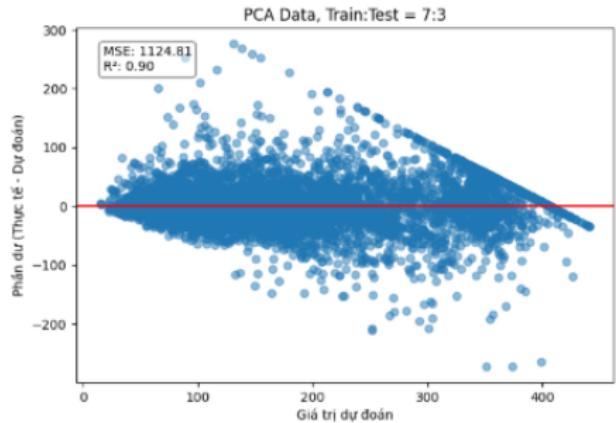


Hình: Ảnh hưởng của regularization (L2) lên hiệu suất MLP

- L2 regularization cải thiện hiệu suất trên dữ liệu gốc và PCA, với RMSE giảm nhẹ và R² tăng tại alpha tối ưu, nhưng không hiệu quả với t-SNE do underfitting.
- Kết quả: hiệu suất tốt trên PCA (RMSE ~ 33.3 , R² ~ 0.8986), nhưng kém trên t-SNE (RMSE ~ 68 , R² ~ 0.5790).

* Đã trình bày ở phần giữa kỳ.

Tương quan phần dư – đặc trưng đầu vào (MLP)



Hình: Scatter plot dự đoán vs. phản dư.

- Phản dư ngẫu nhiên quanh 0, phương sai tăng ở giá trị dự đoán cao (có outlier ± 300).
- Mô hình không có bias hệ thống, nhưng phản ứng tăng cho thấy phương sai không đồng nhất.

Hình: Tương quan phản dư và đặc trưng..

- Tương quan phản dư – đặc trưng rất thấp (<0.09) → không có mối quan hệ tuyến tính.
- Mô hình đã học tốt quan hệ tuyến tính nhưng có thể bỏ sót quan hệ phi tuyến.

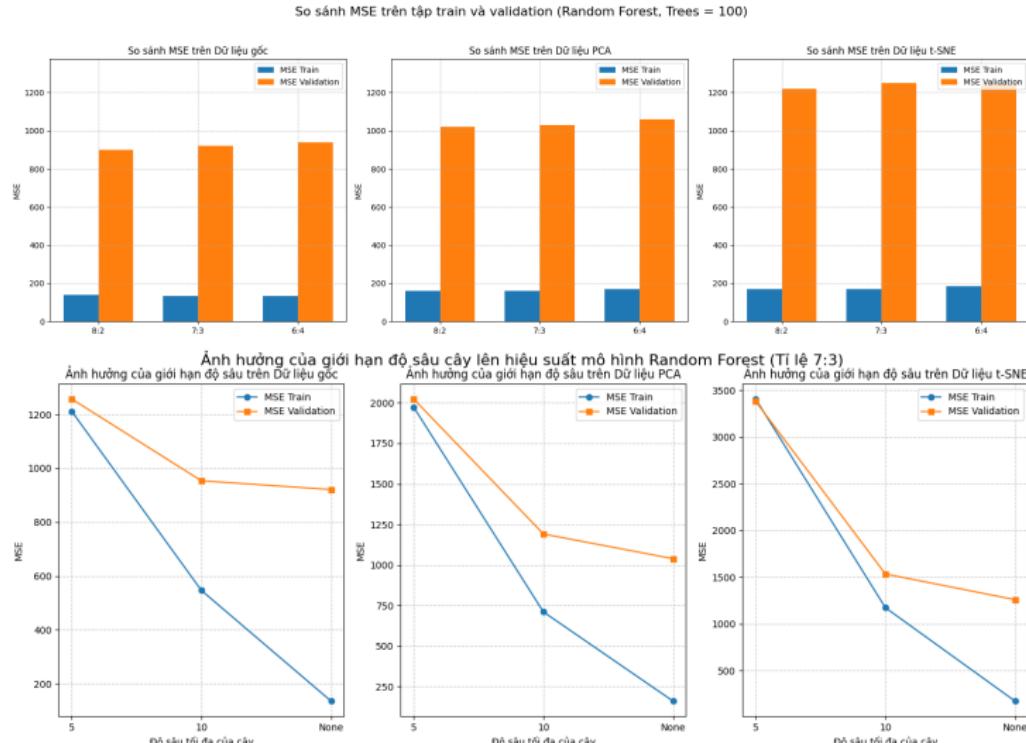
Hiệu suất Random Forest

Dữ liệu	Split	MSE	RMSE	R2	MAE
3*Gốc	8:2	892.3672	29.8725	0.9178	17.4586
	7:3	893.1509	29.8856	0.9183	17.5516
	6:4	903.9877	30.0664	0.9174	17.6793
3*PCA	8:2	1023.5769	31.9934	0.9057	19.1365
	7:3	988.4272	31.4393	0.9096	19.0530
	6:4	1016.8471	31.8880	0.9071	19.4024
3*t-SNE	8:2	1207.6390	34.7511	0.8888	20.6511
	7:3	1179.4381	34.3429	0.8921	20.4195
	6:4	1181.3736	34.3711	0.8921	20.5664

Bảng: Hiệu suất của Random Forest (số cây = 100) với ba phương pháp tiền xử lý (gốc, PCA, t-SNE) ở các tỉ lệ train:validation khác nhau

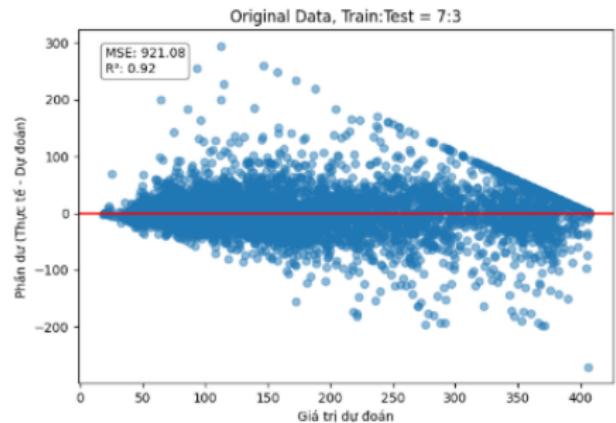
- Dữ liệu gốc cho hiệu suất tốt nhất (R^2 : 0.91-0.92, MSE: 921-926).
- Dữ liệu PCA và t-SNE có hiệu suất thấp hơn, đặc biệt là t-SNE
- Giảm chiều dữ liệu làm giảm hiệu suất, Random Forest tốt nhất trên dữ liệu gốc.

Đánh giá Overfitting của RF



Hình: Giải pháp Overfitting

Tương quan phần dư – đặc trưng đầu vào (RF)



Hình: Scatter plot dự đoán vs. phần dư.

- Phản ứng ngẫu nhiên quanh 0, phương sai không đồng nhất.
- $MSE \approx 921$, $R^2 \approx 0.92$, xuất hiện vài outlier ± 300 .



Hình: Tương quan phần dư và đặc trưng đầu vào.

- Phản ứng không có mối quan hệ tuyến tính mạnh với các đặc trưng (tương quan < 0.06), phù hợp với mô hình.

Kết quả

- **Tiền xử lý dữ liệu:** Dữ liệu đã được làm sạch, xử lý giá trị thiêu và chuẩn hóa để phục vụ cho mô hình.
- **Giảm chiều dữ liệu:**
 - **PCA:** Giảm chiều dữ liệu hiệu quả, giữ lại khoảng 95% phương sai và giúp cải thiện hiệu suất mô hình MLP
 - **t-SNE:** Mặc dù t-SNE hữu ích cho trực quan hóa, nhưng không phù hợp với mô hình dự đoán vì không giữ lại đủ thông tin quan trọng. Đặc biệt với mô hình MLP
- **So sánh mô hình:**
 - Cả **Random Forest** (RF) và **MLP** đều cho hiệu suất tốt trong dự đoán AQI.
 - RF phù hợp hơn với dữ liệu gốc, ổn định và cho độ chính xác cao.
- **Tỉ lệ phân chia dữ liệu:** Tỉ lệ phân chia 7:3 (train:test) mang lại kết quả ổn định và tốt nhất cho cả hai mô hình.
- **Thời gian huấn luyện:** MLP cần thời gian huấn luyện lâu hơn so với Random Forest.

* Đã trình bày ở phần giữa kỳ.

Mục lục

- 1 Giới thiệu (Giữa kì)
- 2 Kiến thức lý thuyết (Giữa kì)
- 3 Dữ liệu (Giữa kì)
- 4 Phân cụm dữ liệu
- 5 Thực nghiệm và kết quả (Giữa kì)
- 6 Mô hình phân loại
- 7 Kết luận và hướng phát triển

Phân chia đều ra thành 4 khoảng và xác định ngưỡng

Dữ liệu AQI gồm **29.531** mẫu, chia thành 4 nhóm gần đều nhau (mỗi nhóm khoảng 7382 mẫu):

- **Thấp:** 7.346 mẫu (24,9 %)
- **Trung bình:** 7.407 mẫu (25,1 %)
- **Cao:** 7.362 mẫu (24,9 %)
- **Rất cao:** 7.416 mẫu (25,1 %)

Các lớp AQI hiện tại đã đảm bảo mỗi khoảng có số lượng mẫu gần bằng nhau. Giả sử thứ tự mức độ AQI là: *Thấp < Trung bình < Cao < Rất cao.*

- **Nhãn 1 (AQI_Thap):** Lớp “Thấp” – 7.346 mẫu.
- **Nhãn 2 (AQI_TrungBinh):** Lớp “Trung bình” – 7.407 mẫu.
- **Nhãn 3 (AQI_Cao):** Lớp “Cao” – 7.362 mẫu.
- **Nhãn 4 (AQI_RatCao):** Lớp “Rất cao” – 7.416 mẫu.

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: 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: 4 lớp AQI đã xác định: "AQI_Thap", "AQI_TrungBinh", "AQI_Cao", "AQI_RatCao".

Mô hình phân loại Naive Bayes

Gaussian Naive Bayes: là thuật toán phân loại, phù hợp khi đặc trưng là biến liên tục và mỗi lớp có phân phối đặc trưng tuân theo phân phối Gaussian. GaussianNB ước tính xác suất mỗi điểm dữ liệu thuộc một lớp AQI, giả sử các biến đầu vào tuân theo phân phối chuẩn trong từng lớp ("Thấp", "Trung bình", "Cao", "Rất cao").

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).

Mô hình phân loại Naive Bayes

Bảng: 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

Nhận xét:

- Mô hình trên dữ liệu gốc luôn cho các chỉ số (70%) cao hơn khoảng 10% so với dữ liệu PCA (60%).
- Thay đổi tỷ lệ Train:Validation ($8:2 \rightarrow 7:3 \rightarrow 6:4$) hầu như không ảnh hưởng đến hiệu suất, cho thấy mô hình khá ổn định.

Mô hình phân loại Naive Bayes

Bảng: Chi tiết kết quả phân loại theo nhãn cho tỉ lệ Train:Val = 0.7:0.3

Loại dữ liệu	Nhãn	Precision	Recall	F1-score	Support
4*Gốc	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
4*PCA	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

Với PCA, hiệu suất giảm trên tất cả các nhãn, nhất là nhãn 1 (F1 từ 0.58 xuống 0.47) và nhãn 2 (từ 0.63 xuống 0.48), nhãn 3 cũng giảm từ 0.86 xuống 0.79.

Mô hình phân loại Random Forest

Bảng: 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

Nhận xét:

- Mô hình trên dữ liệu gốc (Accuracy 82–83%) luôn vượt trội hơn dữ liệu PCA (Accuracy 72–73%), chênh khoảng 10% trên mọi chỉ số.
- Thay đổi tỷ lệ Train:Validation ($8:2 \rightarrow 7:3 \rightarrow 6:4$) hầu như không ảnh hưởng đến kết quả

Mô hình phân loại Random Forest

Bảng: Chi tiết kết quả phân loại theo nhãn cho tỉ lệ Train:Val = 0.7:0.3

Loại dữ liệu	Nhãn	Precision	Recall	F1-score	Support
4*Gốc	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
4*PCA	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

Với PCA, hiệu suất giảm trên tất cả các nhãn, nhất là nhãn 1 và 2 (từ F1 0.75/0.78 xuống 0.65/0.64), nhãn 3 cũng giảm từ 0.92 xuống 0.85, nhãn 0 từ 0.85 xuống 0.78.

So sánh Naive Bayes và Random Forest

Accuracy: RF đạt khoảng 0.82–0.83 (gốc) và 0.72–0.73 (PCA), còn NB chỉ khoảng 0.70 (gốc) và 0.60 (PCA).

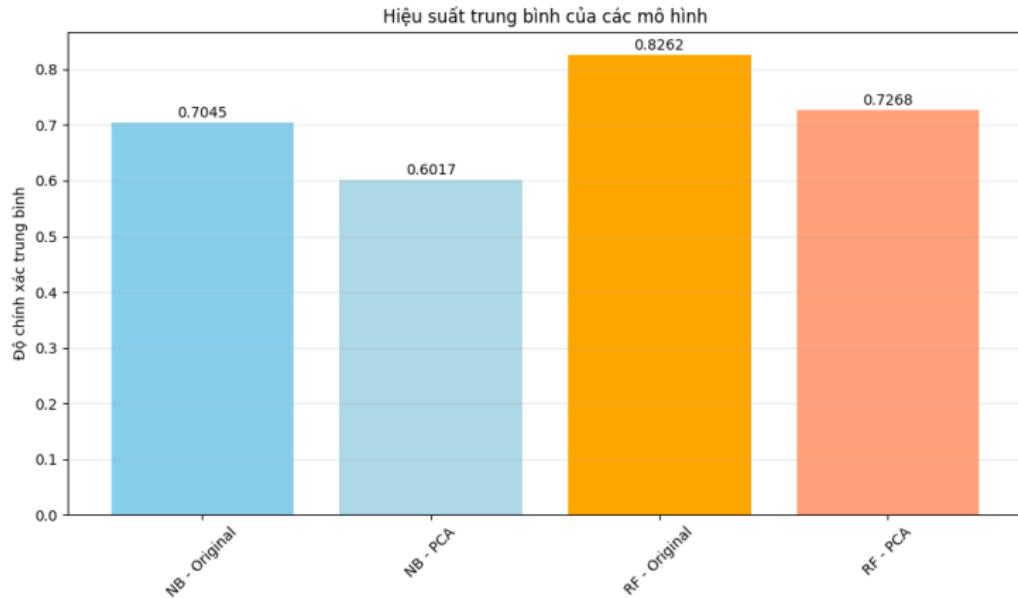
Precision/Recall/F1 (macro): Tương tự accuracy, RF luôn cao hơn NB (ví dụ, F1 macro gốc: RF 0.82, NB 0.70).

Ảnh hưởng của PCA: Cả hai giảm hiệu suất khoảng 0.10, nhưng NB bị suy giảm nghiêm trọng hơn (đặc biệt với các nhãn 1 và 2).

Tỷ lệ Train:Validation: Thay đổi từ 0.8:0.2 đến 0.6:0.4 không làm thay đổi đáng kể hiệu suất của cả hai mô hình.

Phân loại theo nhãn (gốc): RF mạnh hơn NB ở mọi nhãn, nhất là nhãn khó (1 và 2): NB (F1 0.56–0.62), RF (F1 0.74–0.78).

So sánh Naive Bayes và Random Forest



Kết quả chung: Random Forest (RF) luôn vượt trội so với Naive Bayes (NB) ở mọi kịch bản.

Mục lục

- 1 Giới thiệu (Giữa kì)
- 2 Kiến thức lý thuyết (Giữa kì)
- 3 Dữ liệu (Giữa kì)
- 4 Phân cụm dữ liệu
- 5 Thực nghiệm và kết quả (Giữa kì)
- 6 Mô hình phân loại
- 7 Kết luận và hướng phát triển

Kết luận và hướng phát triển

- Các kỹ thuật học máy bước đầu cho kết quả khả quan, nhưng vẫn còn nhiều khía cạnh có thể tối ưu thêm.
- Overfitting là vấn đề nổi bật ở các mô hình hồi quy – cần được xử lý bằng regularization và chọn mô hình phù hợp hơn.
- Giảm chiều chưa mang lại lợi ích rõ rệt cho mô hình học máy – cần cân nhắc mục tiêu khi áp dụng.
- Kết quả phân cụm gợi ý tiềm năng sử dụng như đặc trưng bổ sung để tăng hiệu quả dự báo và phân loại.
- Cần mở rộng đánh giá trên dữ liệu phong phú hơn và dùng cross-validation để tăng độ tin cậy mô hình.

**Cảm ơn thầy và các bạn đã
lắng nghe!**