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

Đỗ Quốc An [1] Phạm Thị Duyên [1] Trần Kiều Hạnh [1]

[1]Khoa Toán Cơ Tin Học
Đại học Khoa học Tự nhiên
Đại học quốc gia Hà Nội

Mục Lục

Giới thiệu

Lý do chọn đề tài

- Ô nhiễm không khí là vấn đề môi trường nghiêm trọng toàn cầu.
- Anh 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 dự báo AQI.
- So sánh hiệu suất trên dữ liệu gốc, PCA, t-SNE qua RMSE, MAE, R^2.
- 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 (3 lớp ẩn: 100–50–25, 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.

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.
- Tối thiểu hóa độ phân kỳ KL giữa hai phân phối qua gradient descent.

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).

Dữ liệu

Nguồn và mô tả 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.

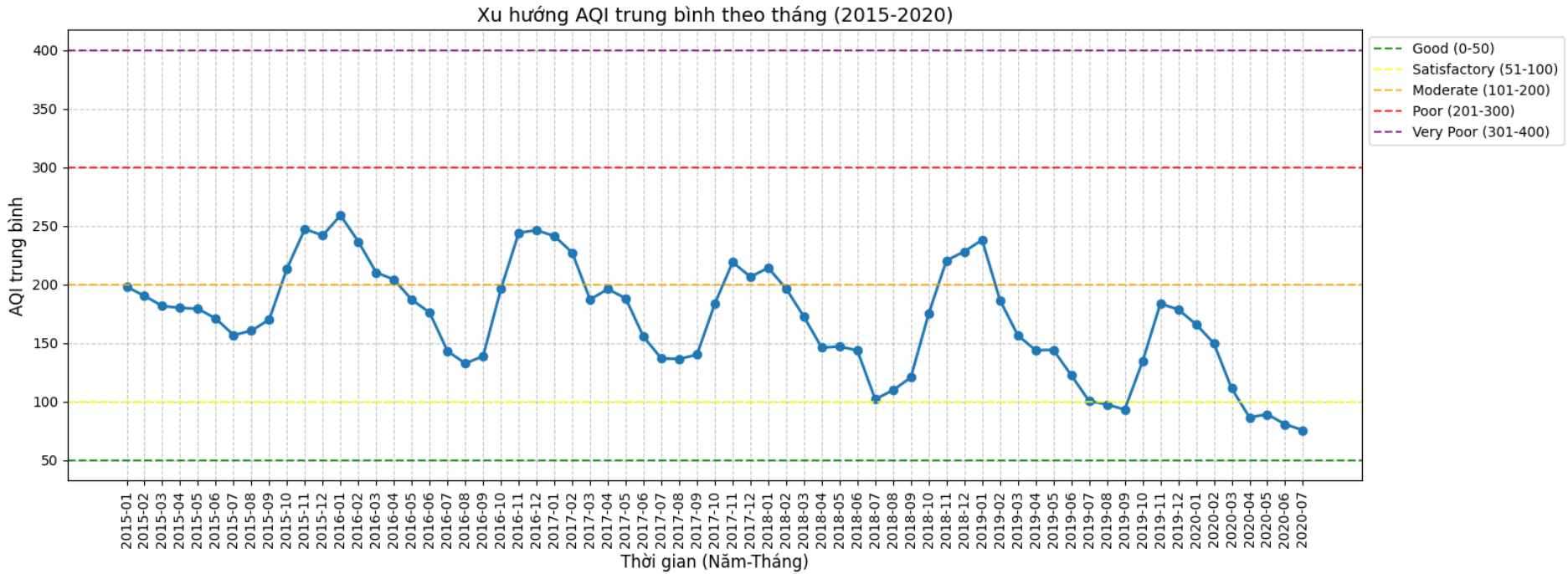
Tiền xử lý dữ liệu

- **Xử lý thời gian:** Chuyển `Date` sang `datetime`.
- **Giá trị không hợp lệ:** Thay thế giá trị âm bằng 0.
- **Giá trị thiếu:** Loại bỏ dòng thiếu AQI (còn 29,531 dòng).
- **Chuẩn hóa:** Dùng `StandardScaler` cho các chỉ số ô nhiễm.

Phân tích và trực quan hóa

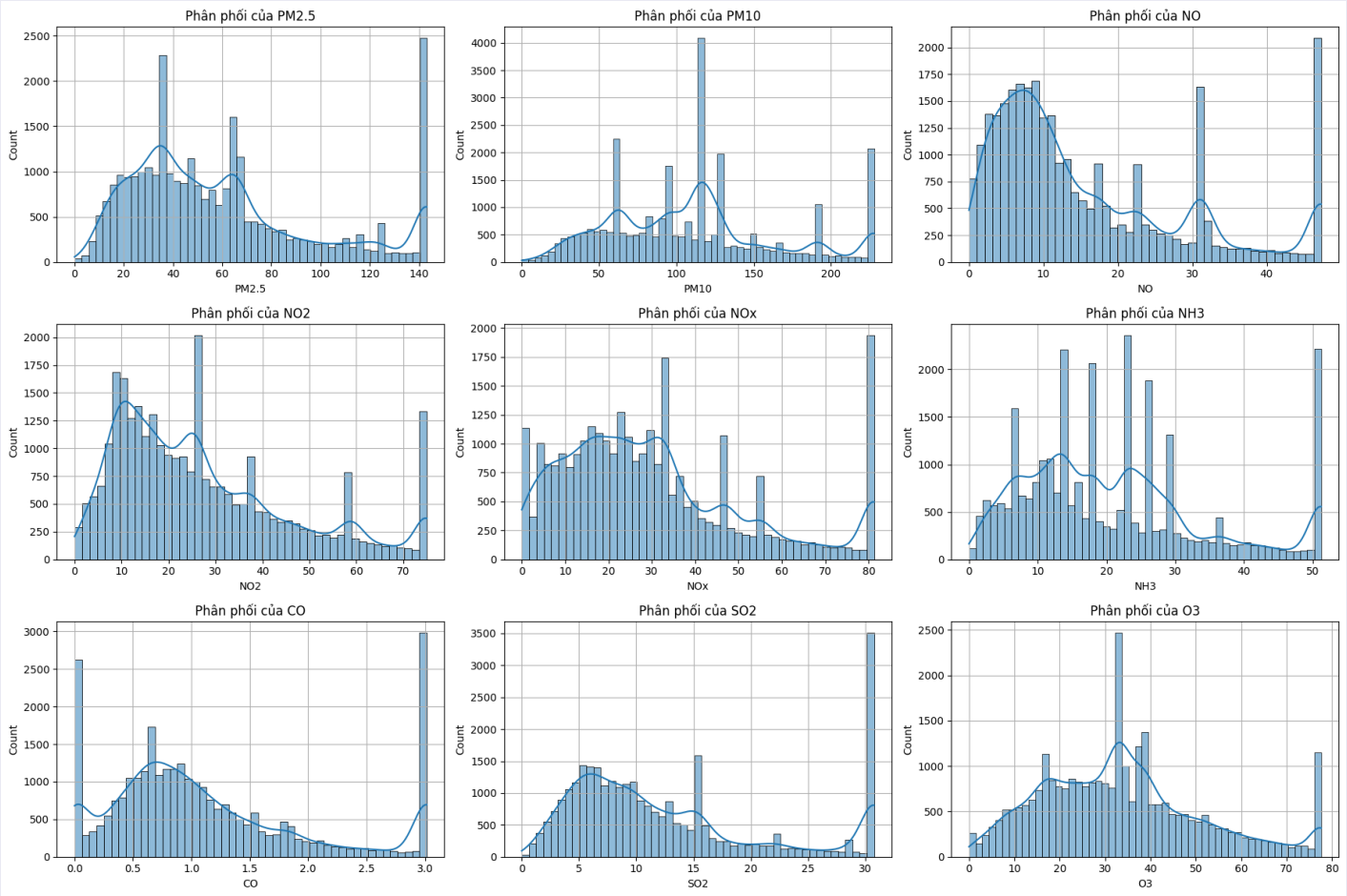
Thống kê mô tả

- Cung cấp cái nhìn tổng quan về dữ liệu: min, max, mean, std, quartiles.
- Ví dụ: AQI trung bình = 161.16, dao động từ 13 đến 407.

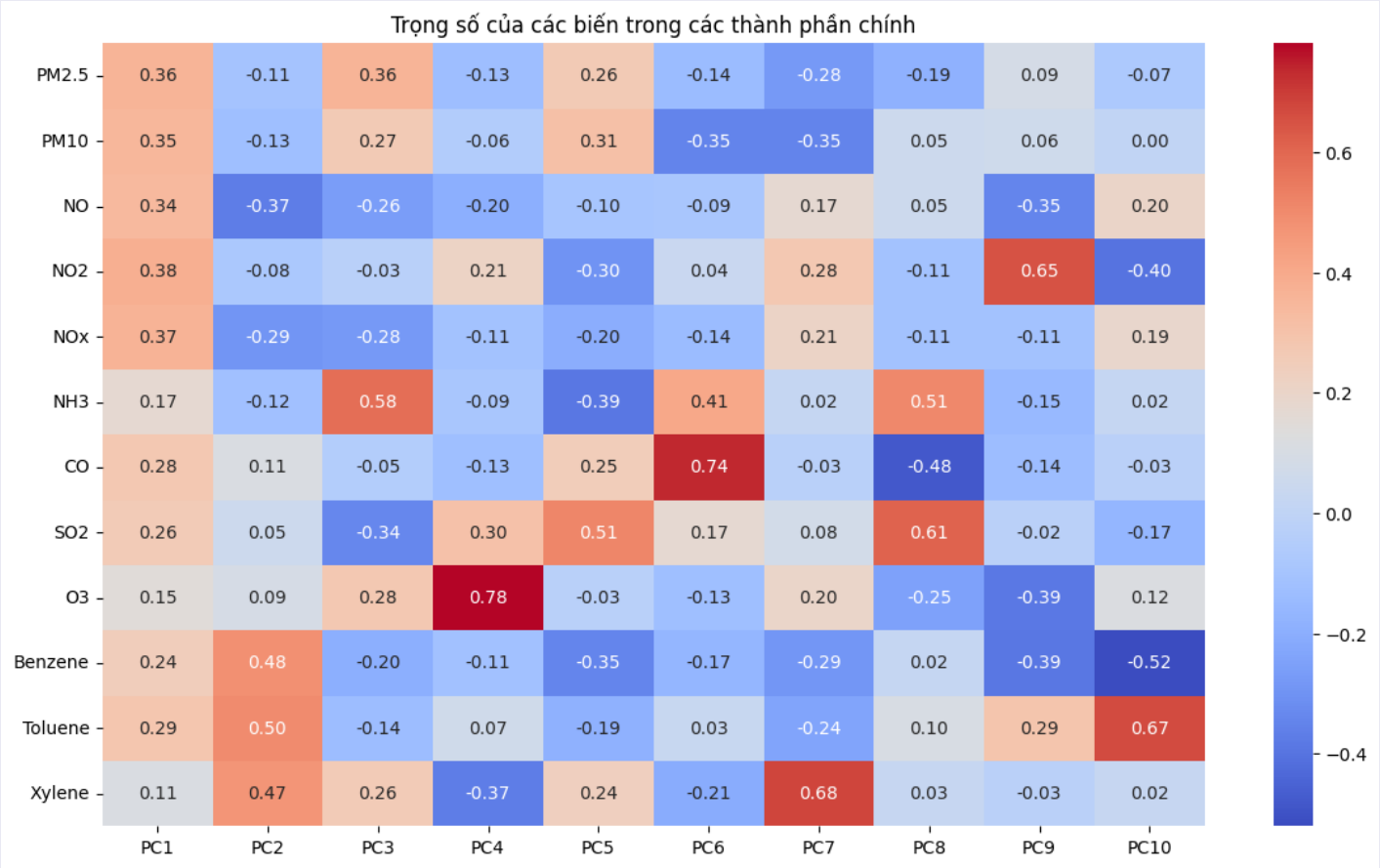


Phân tích và trực quan hóa

Biểu đồ phân phối (Histogram)



Ma trận tương quan



Phân tích và trực quan hóa - Giảm chiều

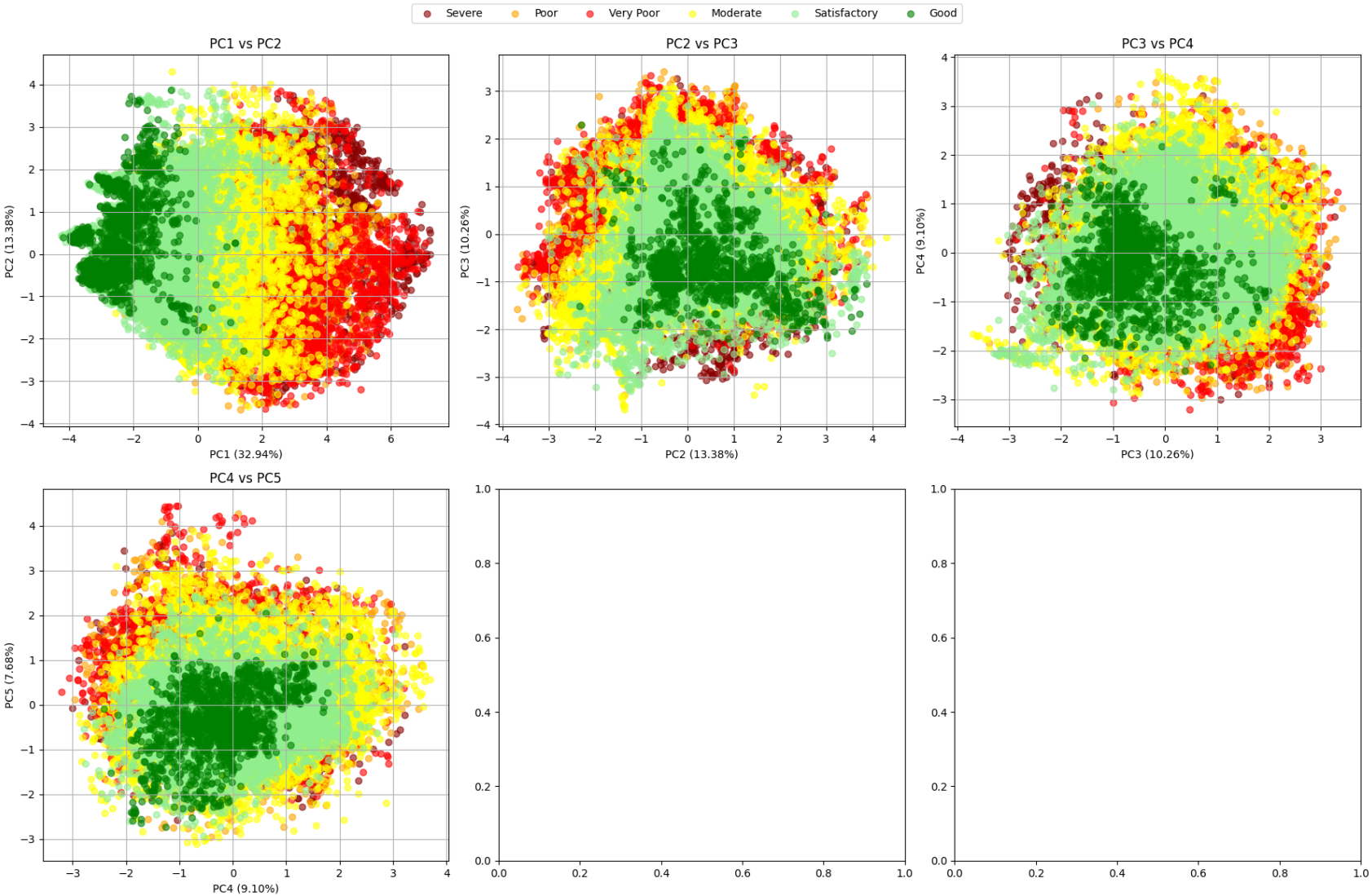
Phân tích thành phần chính (PCA)

- Giữ 95% phương sai $\rightarrow$ giảm từ 12 biến gốc xuống còn 10 thành phần chính.
- PC1 giải thích khoảng 30.87% phương sai trong dữ liệu.
- Trọng số: Các biến gốc đóng góp vào các thành phần chính (VD: O3 đóng góp vào PC4, CO vào PC6).
- Tương quan PCA với AQI: PC1 có tương quan mạnh nhất với AQI (0.79), cho thấy đây là thành phần quan trọng nhất.

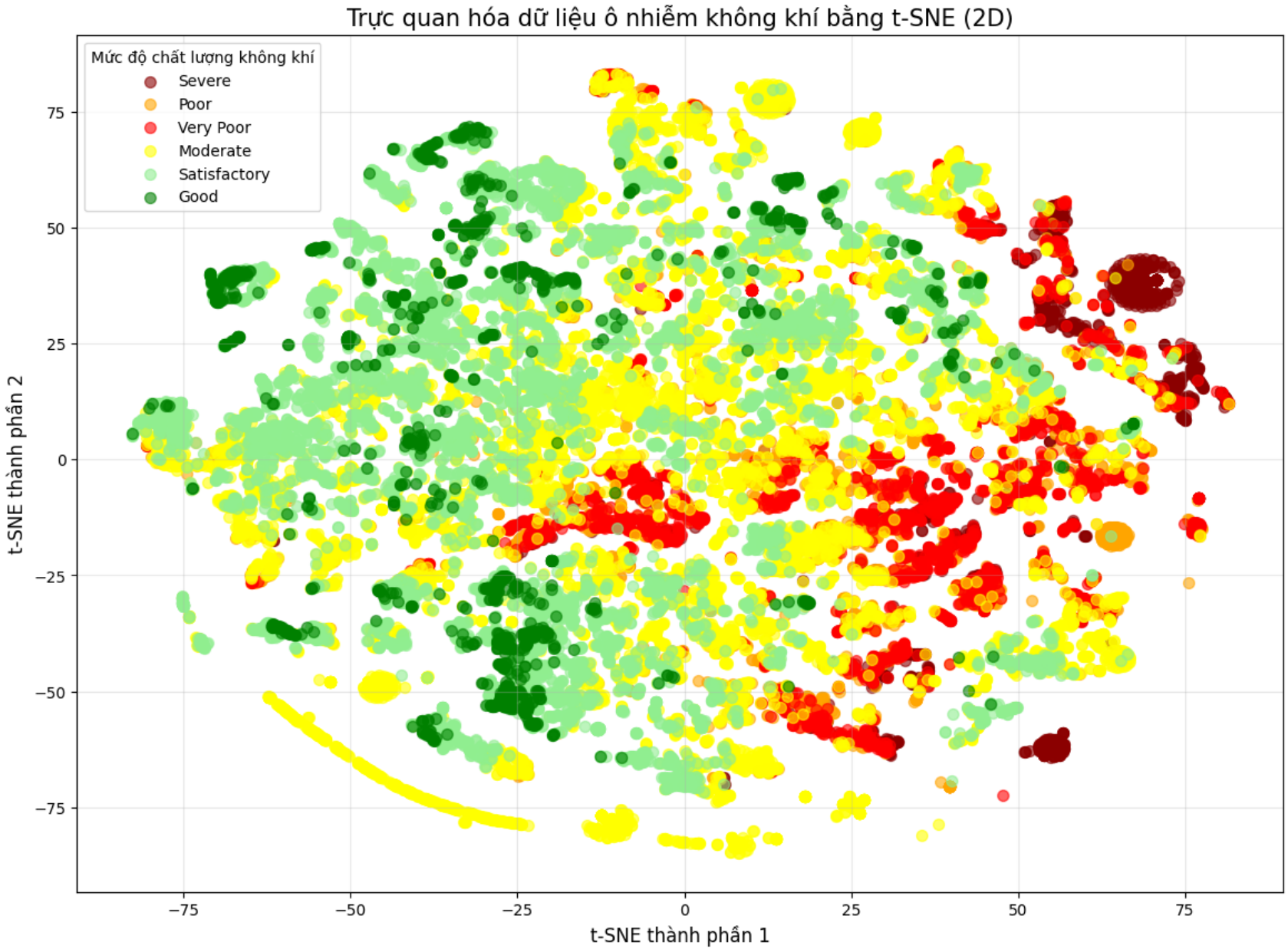
Phân tích t-SNE

- Sử dụng t-SNE để giảm chiều xuống 2D và trực quan hóa dữ liệu.
- Thông số: perplexity=30, n_iter=1000, giúp tối ưu hóa việc phân nhóm và tìm hiểu cấu trúc dữ liệu.

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



Trực quan hóa dữ liệu theo thành phần chính (t-SNE)



Thiết lập mô hình hồi quy

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

Mô hình:

- **Random Forest:** `n_estimators=100`, `random_state=42`, `n_jobs=-1`.
- **MLP:** `hidden_layer_sizes=(100,)`, `activation='relu'`, `solver='adam'`, `max_iter=100`, `random_state=42`.

Dữ liệu: 12 đặc trưng (dữ liệu gốc), 10 đặc trưng (PCA - 95% phương sai).

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

Chỉ số đánh giá: MSE, RMSE, MAE, R^2.

Hiệu suất Random Forest

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

Loại dữ liệu	**Tỷ lệ Train:Val**	**MSE**	**RMSE**	**MAE**	R^2
Dữ liệu gốc	8:2	926.25	30.43	17.84	0.9147
Dữ liệu gốc	7:3	921.08	30.35	17.79	0.9157
Dữ liệu gốc	6:4	925.07	30.41	17.83	0.9155
Dữ liệu PCA	8:2	474.55	21.78	13.54	0.9563
Dữ liệu PCA	7:3	472.94	21.75	13.66	0.9567
Dữ liệu PCA	6:4	476.43	21.83	13.73	0.9565
Dữ liệu t-SNE	8:2	1335.60	36.55	21.75	0.8770
Dữ liệu t-SNE	7:3	1312.93	36.23	21.67	0.8800
Dữ liệu t-SNE	6:4	1330.79	36.48	21.90	0.8780

Nhận xét:

- PCA vượt trội so với dữ liệu gốc ($R^2 \approx 0.957$ so với 0.916).
- PCA giảm nhiễu, cải thiện MSE, RMSE, MAE.
- Tỷ lệ 7:3 tối ưu cho PCA ($R^2 = 0.9567$).

Hiệu suất MLP

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	1077.63	32.83	20.49	0.901
Dữ liệu gốc	7:3	1017.84	31.90	20.03	0.907
Dữ liệu gốc	6:4	1058.48	32.53	20.39	0.903
Dữ liệu PCA	8:2	1254.53	35.42	22.31	0.884
Dữ liệu PCA	7:3	1242.55	35.25	22.53	0.886
Dữ liệu PCA	6:4	1244.80	35.28	22.49	0.886
Dữ liệu t-SNE	8:2	3192.63	56.50	40.05	0.706
Dữ liệu t-SNE	7:3	3359.06	57.96	41.21	0.693
Dữ liệu t-SNE	6:4	3531.50	59.43	42.17	0.677

Nhận xét:

- Dữ liệu gốc vượt trội so với PCA ($R^2 \approx 0.907$ vs. ≈ 0.886).
- PCA giảm hiệu suất MLP, có thể do mất thông tin phi tuyến.
- Tỷ lệ 7:3 tối ưu cho dữ liệu gốc ($R^2 = 0.907$).

So sánh hiệu suất RF vs. MLP (Tỷ lệ Train:Val 7:3)

Bảng: Hiệu suất MLP và RF trên các loại dữ liệu với tỷ lệ Train:Val 7:3

Mô hình	**Loại dữ liệu**	**MSE**	**RMSE**	**MAE**	R^2
MLP	Dữ liệu gốc	1017.84	31.90	20.03	0.907
MLP	Dữ liệu PCA	1242.55	35.25	22.53	0.886
MLP	Dữ liệu t-SNE	3359.06	57.96	41.21	0.693
Random Forest	Dữ liệu gốc	921.08	30.35	17.79	0.9157
Random Forest	Dữ liệu PCA	472.94	21.75	13.66	0.9567
Random Forest	Dữ liệu t-SNE	1312.93	36.23	21.67	0.8800

Kết luận:

- PCA giúp cải thiện hiệu suất cho cả MLP (tăng từ $R^2 = 0.907$ lên 0.886) và Random Forest (tăng từ $R^2 = 0.9157$ lên 0.9567).
- Random Forest hoạt động ổn định hơn trên mọi loại dữ liệu, đặc biệt hiệu suất cao trên dữ liệu PCA với $R^2 = 0.9567$.
- t-SNE làm giảm hiệu suất, đặc biệt là với MLP ($R^2 = 0.693$), và có sự giảm hiệu suất nhẹ với Random Forest ($R^2 = 0.8800$).

Đánh giá phần dư - RF + PCA (Tỷ lệ 7:3)

Thống kê phần dư

- Trung bình: ≈ -0.14 (gần 0, mô hình không chệch).
- Độ lệch chuẩn: 34.83.
- Min/Max: -271.46 / 278.21.

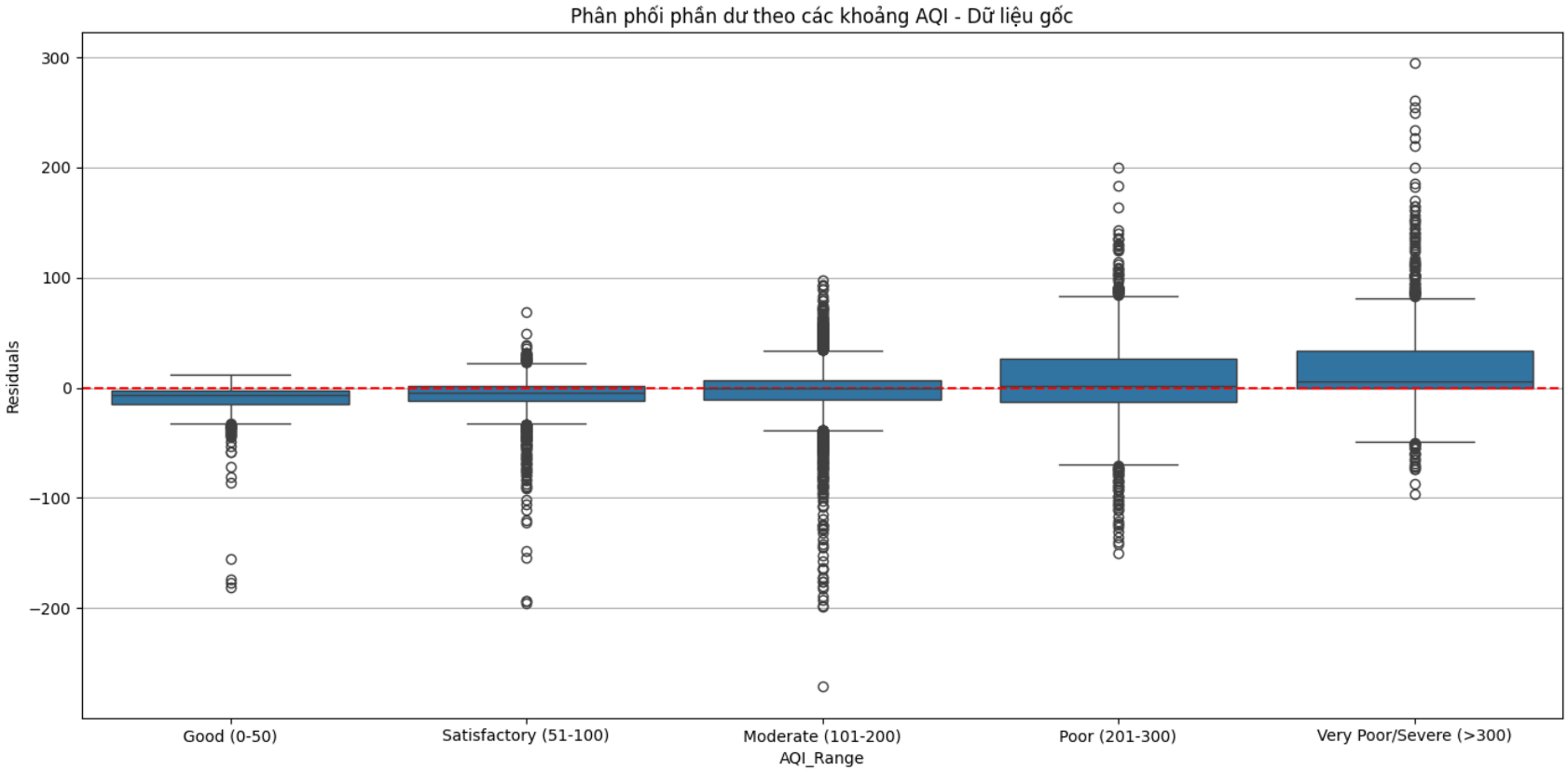
Phân phối phần dư

- Tập trung quanh 0, nhưng phân phối không hoàn toàn chuẩn (đuôi nặng).
- Q-Q plot cho thấy độ lệch nhẹ khỏi phân phối chuẩn.

Phân tích phần dư

- **Tương quan với biến đầu vào (PCA):** Gần 0, mô hình học tốt quan hệ dữ liệu.
- **Phần dư vs. Giá trị dự đoán:** Không có mẫu hình rõ ràng (tốt); sai

Phân phối phần dư theo khoảng AQI - RF + PCA (7:3)



Hình: Phân tích sai số theo từng mức độ ô nhiễm.

Phân tích phần dư theo khoảng AQI (RF + PCA)

Nhận xét:

- Mô hình dự đoán tốt hơn ở các mức AQI thấp và trung bình (Good, Satisfactory, Moderate).
- Sai số (độ biến thiên phần dư) tăng lên ở các mức AQI cao (Poor, Very Poor, Severe).
- Có xu hướng dự đoán thấp hơn giá trị thực tế (underestimation) ở các mức ô nhiễm rất cao.

Kết quả

Tổng quan

- Mô hình dự báo AQI được xây dựng thành công với **Random Forest** và **MLP**.
- Áp dụng **PCA** và **t-SNE** để giảm chiều và trực quan hóa dữ liệu.

Hiệu suất mô hình

- **Random Forest kết hợp với PCA** đạt hiệu suất tốt nhất:
  - $R^2 = 0.957$
  - RMSE $= 21.75$
- **PCA** giúp cải thiện hiệu suất của **Random Forest**, nhưng làm giảm hiệu suất của **MLP**.

Đánh giá

Tổng quan

- **Random Forest + PCA:** Hiệu quả cao trong dự đoán AQI tổng thể.
- **PCA:** Giảm chiều, tăng độ chính xác và giảm nhiễu cho mô hình Random Forest.

Ứng dụng và hạn chế

- **Ứng dụng thực tế:** Có thể sử dụng để ước tính AQI trong các môi trường khác nhau.
- **Hạn chế:** Mô hình giảm độ chính xác khi dự đoán các mức ô nhiễm cực đoan (Very Poor/Severe).

Hướng phát triển

Cải thiện mô hình

- Thử nghiệm XGBoost, LightGBM.
- Tinh chỉnh siêu tham số RF và MLP.
- Sử dụng LSTM, GRU cho dự báo chuỗi thời gian.
- Áp dụng kỹ thuật xử lý mất cân bằng dữ liệu.

Dữ liệu

- Bổ sung dữ liệu từ nhiều thành phố/quốc gia.
- Thêm đặc trưng thời tiết, giao thông.
- Sử dụng dữ liệu có độ phân giải thời gian cao hơn.

Phân tích sâu hơn

- Nghiên cứu sai số ở các mức AQI cao, đề xuất cải tiến.