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 21 tháng 4 năm 2025

Mục lục

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.
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

Danh sách hình vẽ

1 Giới thiệu đề tài

1.1 Lý do chọn đề tài

Ô nhiễm không khí hiện đang là một trong những thách thức môi trường nghiêm trọng nhất trên phạm vi toàn cầu. Sự gia tăng dân số đô thị, hoạt động công nghiệp, phương tiện giao thông và đốt nhiên liệu hóa thạch đã góp phần làm gia tăng nồng độ các chất ô nhiễm độc hại trong không khí như bụi mịn (PM2.5, PM10), khí CO, NO, SO và O. Những tác nhân này không chỉ ảnh hưởng tiêu cực đến môi trường tự nhiên mà còn gây ra nhiều vấn đề sức khỏe nghiêm trọng như bệnh hô hấp, tim mạch, và ung thư phổi.

Để giám sát và đánh giá mức độ ô nhiễm, chỉ số chất lượng không khí (AQI - Air Quality Index) đã được xây dựng như một công cụ tổng hợp giúp chuyển đổi các giá trị nồng độ ô nhiễm phức tạp thành một thang đo dễ hiểu cho cộng đồng. AQI được sử dụng rộng rãi bởi các tổ chức môi trường, chính phủ, và các hệ thống theo dõi chất lượng không khí để phản ánh nhanh tình trạng ô nhiễm và đưa ra khuyến nghị sức khỏe cho người dân.

Tuy nhiên, trong thực tế, việc đo lường và cập nhật AQI liên tục tại tất cả các địa điểm là một thách thức do hạn chế về thiết bị cảm biến, tài nguyên và thời gian xử lý. Chính vì vậy, việc phát triển các mô hình dự báo AQI dựa trên dữ liệu các chất ô nhiễm đã được đo là cần thiết. Các mô hình này không chỉ giúp ước lượng nhanh AQI tại những nơi chưa có thiết bị giám sát mà còn hỗ trợ đưa ra các cảnh báo sớm và hoạch định chính sách phòng ngừa ô nhiễm hiệu quả.

1.2 Mục tiêu

- Xây dựng mô hình Random Forest để dự báo AQI tại các thành phố dựa trên dữ liệu lịch sử về các chất ô nhiễm.
- So sánh hiệu suất của Random Forest trên dữ liệu gốc, dữ liệu PCA và dữ liệu t-SNE, sử dụng các chỉ số RMSE, MAE và R^2.
- Phân tích các yếu tố ô nhiễm chính ảnh hưởng đến AQI thông qua tầm quan trọng đặc trưng và tương quan, đồng thời đánh giá phần dư để hiểu hiệu suất mô hình theo các mức AQI (Good, Satisfactory, Moderate, Poor, Very Poor, Severe).
- Đề xuất cải tiến để nâng cao độ chính xác, đặc biệt ở các mức ô nhiễm cao (Poor, Very Poor/Severe), dựa trên phân tích phần dư.

1.3 Phạm vi nghiên cứu

- **Dữ liệu**: Tập dữ liệu chất lượng không khí hàng ngày (`city_day.csv`) từ 26 thành phố, từ 1/1/2015 đến 1/7/2020, bao gồm các chỉ số ô nhiễm và AQI.
- **Phương pháp**: Áp dụng PCA và t-SNE để phân tích và trực quan hóa dữ liệu, sử dụng Random Forest và MLP để phân tích hồi quy, dự báo AQI.
- **Phạm vi địa lý**: Các thành phố trong tập dữ liệu, bao gồm Ahmedabad, Delhi, Mumbai, Kolkata, v.v.

- **Phạm vi thời gian**: Phân tích dữ liệu từ 2015 đến 2020, tập trung vào xu hướng ô nhiễm theo ngày.

1.4 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.
- **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, R^2, 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.

2 Kiến thức lý thuyết

2.1 PCA (Principal Component Analysis)

2.1.1 PCA là gì?

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.

Một số đặc điểm nổi bật của PCA:

- **Giữ lại thông tin quan trọng nhất:** PCA tìm ra hướng có mức độ biến động (thay đổi) lớn nhất trong dữ liệu và ưu tiên giữ lại hướng đó.
- **Các đặc trưng mới không bị trùng lặp:** Các thành phần chính tạo ra không bị phụ thuộc vào nhau, giúp mô hình học máy hoạt động hiệu quả hơn.
- **Giảm chiều nhưng không làm mất cấu trúc dữ liệu:** Việc giảm số chiều giúp đơn giản hóa dữ liệu mà vẫn bảo toàn những mối quan hệ quan trọng.

2.1.2 PCA hoạt động như thế nào?

PCA bắt đầu bằng việc chuẩn bị dữ liệu (thường là chuẩn hóa về cùng thang đo), sau đó tìm ra các hướng chính trong không gian dữ liệu – đây là những hướng mà dữ liệu thay đổi mạnh nhất. Những hướng này chính là các thành phần chính.

Sau đó, dữ liệu được "chiếu" lên những hướng đó để tạo ra một phiên bản mới đơn giản hơn của dữ liệu, nhưng vẫn chứa phần lớn thông tin ban đầu. Quá trình này không những giúp mô hình học tốt hơn mà còn giúp giảm thời gian tính toán và tránh hiện tượng quá khớp.

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

2.2.1 t-SNE là gì?

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.

2.2.2 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:

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

2.2.3 Tham số quan trọng

- **Perplexity**: Điều khiển độ rộng 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.

2.3 MLP (Multi-layer Perceptron)

2.3.1 MLP là gì?

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.

2.3.2 Cấu trúc của MLP

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.

2.3.3 Cách MLP 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.

2.3.4 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.

2.3.5 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).

2.3.6 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.

2.4 Random Forest

2.4.1 Random Forest là gì?

Random Forest là một phương pháp học máy theo mô hình tổ hợp (ensemble learning) được đề xuất bởi Leo Breiman năm 2001 [1]. Thuật toán xây dựng một tập hợp các cây quyết định và đưa ra dự đoán bằng cách tổng hợp kết quả của từng cây. Phương pháp này được sử dụng phổ biến trong các bài toán phân loại và hồi quy, đặc biệt hiệu quả khi dữ liệu có nhiều đặc trưng và nhiễu.

2.4.2 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 hoàn 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.
- **Dự đoán**:
  - **Phân loại**: Mỗi cây đưa ra một dự đoán, kết quả cuối cùng là lớp được dự đoán nhiều nhất (đa số phiếu).
  - **Hồi quy**: Dự đoán được tính bằng trung bình dự đoán của tất cả các cây.

2.4.3 Cơ sở lý thuyết

Random Forest có một số tính chất lý thuyết quan trọng đảm bảo tính ổn định và tránh quá khớp:

- **Không quá khớp khi số cây lớn**: Khi số lượng cây đủ lớn, sai số tổng quát hóa PE^* hội tụ:

$$PE^* = P_{\mathbf{X},Y}\left(P_\Theta(h(\mathbf{X},\Theta) = Y) - \max_{j \neq Y} P_\Theta(h(\mathbf{X},\Theta) = j) < 0\right)$$

  Luật số lớn đảm bảo tính hội tụ, làm giảm nguy cơ quá khớp [1].

- **Độ mạnh và tương quan giữa các cây**:
  - *Độ mạnh* s: Là độ chính xác kỳ vọng của mỗi cây thành phần.
  - *Tương quan* $\bar{\rho}$: Là mức độ giống nhau giữa các cây trong rừng.

  Sai số tổng quát hóa được chặn bởi:

$$PE^* \leq \frac{\bar{\rho}(1 - s^2)}{s^2}$$

  Cho thấy mô hình hiệu quả khi mỗi cây mạnh (s cao) và không quá giống nhau ($\bar{\rho}$ thấp).

- **Hồi quy với Random Forest**: Với bài toán hồi quy, lỗi bình phương trung bình của rừng được giới hạn bởi:

$$PE^*(\text{forest}) \leq \bar{\rho} \cdot PE^*(\text{tree})$$

  Cho thấy lợi ích của việc kết hợp nhiều cây có tương quan thấp để giảm sai số tổng thể.

3 Dữ liệu

3.1 Nguồn dữ liệu

Tập dữ liệu 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, O_3, CO, SO_2, NO_2, v.v.

3.2 Tiền xử lý dữ liệu

3.2.1 Đọc và mô tả dữ liệu ban đầu

Đọc dữ liệu: Dữ liệu được tải từ tệp `city_day.csv` bằng phương thức `pandas.read_csv()`. Tập dữ liệu gồm **29,532 dòng** và **16 cột**.

Cấu trúc dữ liệu:

- **Thời gian - địa điểm:** `Date`, `City`.
- **Chỉ số ô nhiễm:** Bao gồm các chỉ số như PM2.5, O_3, CO, SO_2, NO_2, cùng các giá trị trung bình, cực đại và giờ cực đại.
- **Kiểu dữ liệu:** Dữ liệu bao gồm 5 cột dạng chuỗi, 7 cột dạng số nguyên, và 10 cột dạng số thực.
- **Dung lượng bộ nhớ:** Khoảng **111.7 MB**, không có giá trị thiếu trong các cột bắt buộc như `City` và `Date`.
- **Thống kê mẫu:**
  - `O3 Mean`: Trung bình 0.0286, giá trị min -0.0007, max 0.1074.
  - `State`: Có 48 giá trị duy nhất.
  - `County`: Có 137 giá trị duy nhất.

3.2.2 Chuẩn hóa và làm sạch dữ liệu

- **Xử lý thời gian:** Cột `Date` được chuyển sang định dạng `datetime` để dễ dàng thao tác với dữ liệu thời gian.
- **Xử lý giá trị không hợp lệ:** Các giá trị âm trong các chỉ số ô nhiễm như O_3, CO, SO_2, NO_2 được thay thế bằng giá trị `0` nhằm đảm bảo tính hợp lệ của dữ liệu.

3.2.3 Chuyển đổi và mã hóa dữ liệu

- **Mã hóa One-hot:** Các cột phân loại như `State`, `County`, và `City` được mã hóa one-hot, làm tăng số lượng cột từ 22 lên khoảng **350**.
- **Kiểu dữ liệu:** Dữ liệu số được giữ nguyên sau khi mã hóa, trong khi cột `Date` đã được chuyển đổi sang kiểu `datetime`.

3.2.4 Chuẩn hóa giá trị

- Các giá trị âm trong các chỉ số ô nhiễm được thay bằng 0, đảm bảo tính hợp lệ và ổn định cho các mô hình học máy.

3.2.5 Mô tả dữ liệu sau xử lý

- **Số dòng: 29,531** dòng sau khi loại bỏ giá trị thiếu.
- **Số cột:** Khoảng **350** cột sau khi thực hiện mã hóa one-hot.
- **Thống kê dữ liệu sau xử lý:**
  - `O3 Mean`: Trung bình 0.0286, giá trị min 0.0000, max 0.1074.
  - `CO Mean`: Trung bình 0.3298, giá trị max 7.5083.
  - SO_2 `Mean`: Trung bình 1.4333, giá trị max 321.6250.
  - NO_2 `Mean`: Trung bình 11.5116, giá trị max 140.6500.

3.3 Phân tích và trực quan hóa dữ liệu

3.3.1 Phân tích thống kê dữ liệu

Tổng quan dữ liệu Dữ liệu bao gồm thông tin về chất lượng không khí tại 26 thành phố của Ấn Độ từ ngày 01/01/2015 đến 01/07/2020:

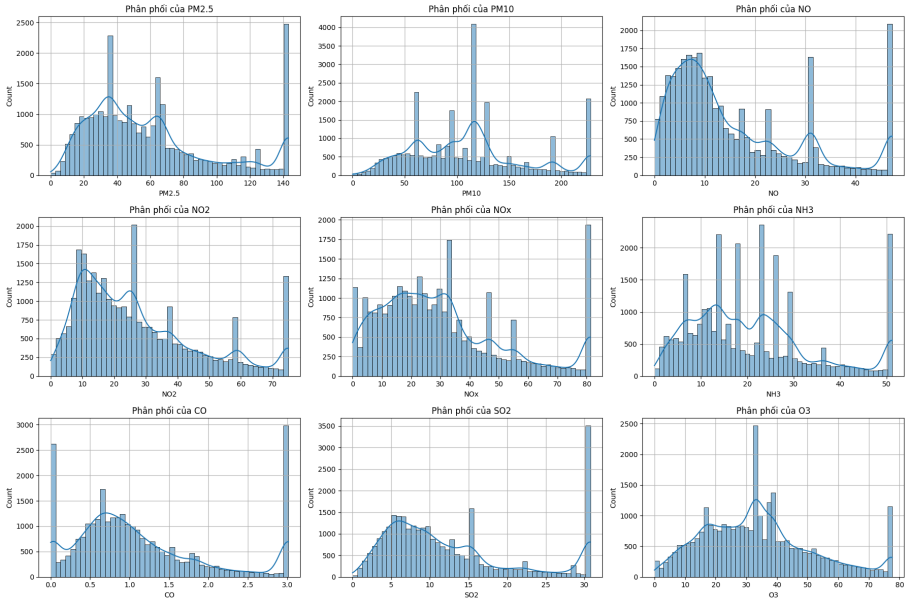
- **Kích thước**: 29,531 hàng, 16 cột.
- **Loại dữ liệu**:
  - **Định lượng** (13 trường): *PM2.5, PM10, ...*
  - **Định tính** (3 trường): *City, Date, AQI_Bucket*.
- **Số lượng thành phố**: 26 (ví dụ: Delhi, Mumbai, ...).

Thống 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: Thống kê mô tả cho các trường dữ liệu định lượng

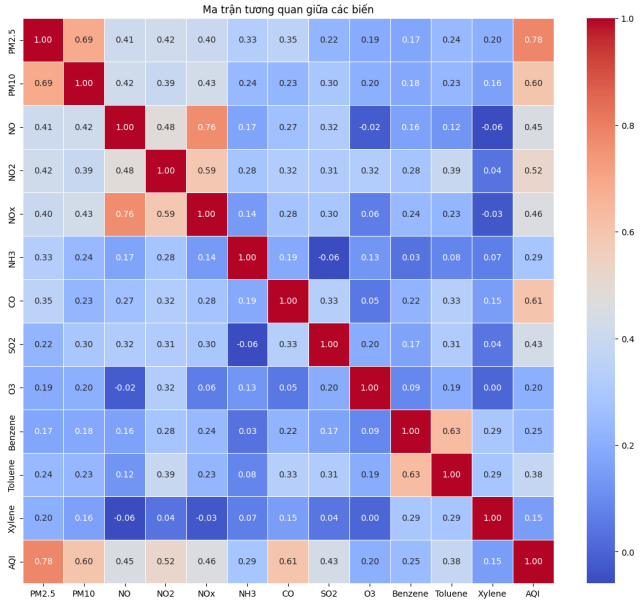
Thống kê	PM2.5	PM10	NO	NO2	NOx	NH3	CO	SO2	O3	Benzene	Toluene	Xylene	AQI
Count	29531	29531	29531	29531	29531	29531	29531	29531	29531	29531	29531	29531	29531
Mean	59.80	109.90	15.98	27.44	30.49	20.97	1.13	12.75	33.63	2.20	6.12	2.68	161.16
Std	38.12	54.51	13.20	19.12	21.59	13.32	0.87	8.78	18.03	2.24	6.29	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

Biểu đồ phân phối dữ liệu Để trực quan hóa phân phối của các biến định lượng, ta sử dụng biểu đồ histogram. Trong hình 1, thể hiện ví dụ histogram cho một số trường.



Hình 1: Biểu đồ histogram thể hiện phân phối của một số biến định lượng

Ma trận tương quan Để kiểm tra mối liên hệ tuyến tính giữa các biến định lượng, ta sử dụng ma trận tương quan Pearson như thể hiện trong Hình 2.



Các cặp biến tương quan mạnh nhất:

- **PM2.5 – AQI**: 0.7827
- **NO – NOx**: 0.7649
- **PM2.5 – PM10**: 0.6879
- **Benzene – Toluene**: 0.6337
- **CO – AQI**: 0.6058

Các biến như PM2.5, NO, CO, Benzene là các yếu tố cực kỳ ảnh hưởng tới chất lượng không khí (AQI).

Hình 2: Ma trận tương quan giữa các biến số

3.3.2 Chuẩn hóa dữ liệu và đánh giá thành phần chính

Chuẩn hóa dữ liệu Trước khi thực hiện phân tích thành phần chính (PCA), các trường dữ liệu định lượng được chuẩn hóa bằng phương pháp *StandardScaler* để đưa các biến về cùng thang đo (trung bình bằng 0, độ lệch chuẩn

bằng 1). Biến `AQI` được loại khỏi tập đặc trưng vì là biến mục tiêu trong các bài toán dự đoán sau này.

- Dữ liệu đầu vào có 12 biến: *PM2.5, PM10, NO, NO2, NOx, NH3, CO, SO2, O3, Benzene, Toluene, Xylene*.
- Sau chuẩn hóa, phân phối các biến được đưa về chuẩn, giúp PCA hoạt động hiệu quả hơn.

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

Bảng 2: Thống kê mô tả cho dữ liệu đã chuẩn hóa

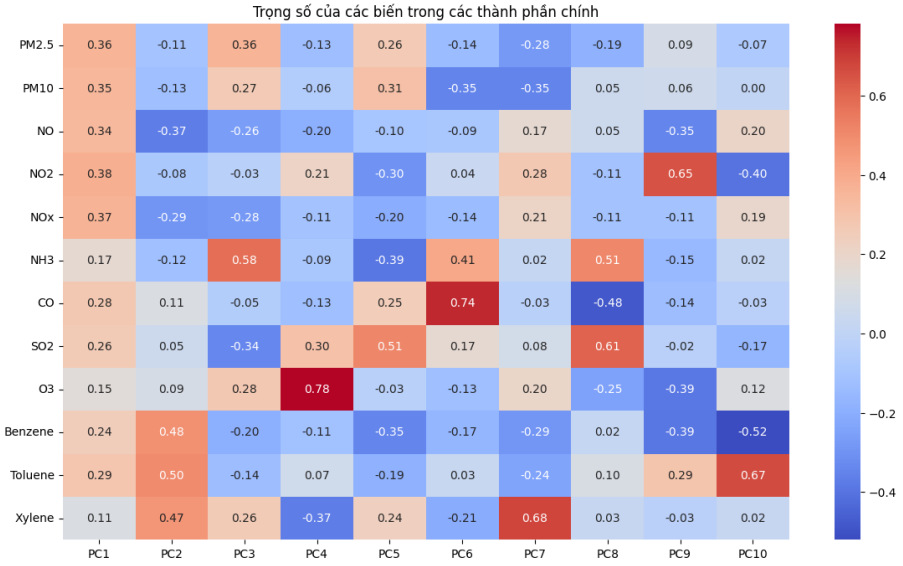
Thống kê	PM2.5	PM10	NO	NO2	NOx	NH3	CO	SO2	O3	Benzene	Toluene	Xylene
Count	...	...	...	...	...	...	...	...	...	...	...	...
Mean	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Std	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Min	...	...	...	...	...	...	...	...	...	...	...	...
Max	...	...	...	...	...	...	...	...	...	...	...	...

Phân tích thành phần chính (PCA) Sau khi chuẩn hóa, ta sẽ chỉ chọn các cột số (numeric) để áp dụng phương pháp PCA. Phương pháp PCA được áp dụng để giảm chiều dữ liệu với tham số `n_components=0.95`, nghĩa là ta muốn giữ lại **95%** lượng thông tin. Kết quả cho thấy...

- Có 10 thành phần chính tương ứng với 12 đặc trưng đầu vào.
- Thành phần chính đầu tiên (PC1) chiếm khoảng **30.87%**, ba thành phần đầu tiên cộng lại chiếm hơn **64%**.
- Để giữ lại **95% phương sai**, cần chọn **10 thành phần chính**.

Ý nghĩa của trọng số các biến trong các thành phần chính (PCA) Trong phân tích thành phần chính PCA, trọng số (loadings) thể hiện mức độ đóng góp của từng biến gốc vào các thành phần chính (PCs - Principal Components).

- Giá trị trọng số cao (gần 1 hoặc -1): Biến đó đóng góp mạnh vào thành phần chính.
- Giá trị trọng số gần 0: Biến đó không ảnh hưởng nhiều đến thành phần chính đó.
- Dấu dương biến tăng sẽ làm tăng giá trị thành phần chính.
- Dấu âm biến tăng sẽ làm giảm giá trị thành phần chính.



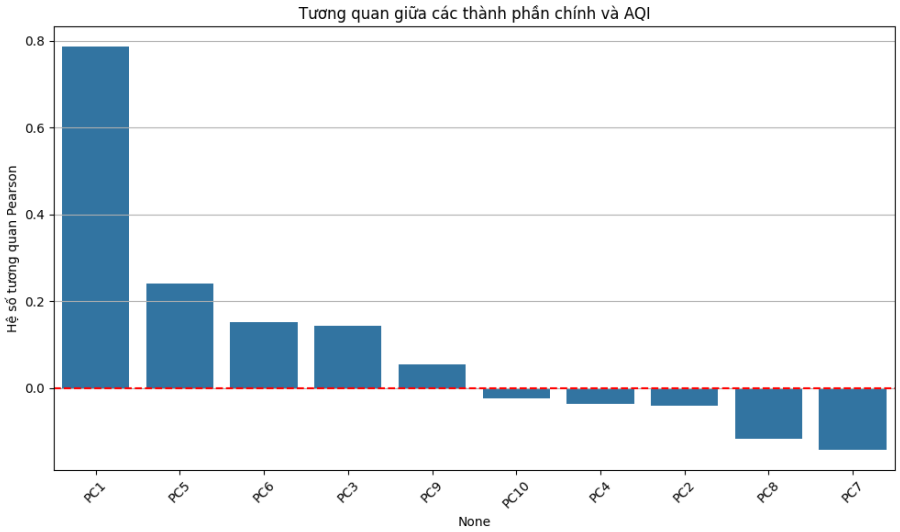
Hình 3: Trọng số của các biến trong các thành phần chính

Ví dụ:

- O_3 ở PC4 có trọng số 0.78 → O đóng góp rất mạnh vào PC4.
- CO ở PC6 có trọng số 0.74 → CO đóng góp chủ yếu vào PC6.

Phân tích mối quan hệ giữa các thành phần chính (PCA) với AQI Trong phương pháp PCA khi áp dụng cho dữ liệu ô nhiễm không khí, hệ số tương quan Pearson giữa các thành phần chính (PC1, PC2, ..., PC10) và AQI (Air Quality Index) hiểu đơn giản:

- Hệ số tương quan Pearson đo lường mức độ tuyến tính giữa 2 đại lượng.
- Giá trị gần 1 → Mối tương quan dương mạnh (cùng tăng/cùng giảm).
- Giá trị gần -1 → Mối tương quan âm mạnh (một tăng, một giảm).
- Giá trị gần 0 → Không có mối tương quan rõ rệt.



Hình 4: Trọng số của các biến trong các thành phần chính

Cụ thể trên biểu đồ này:

- PC1 có hệ số tương quan với AQI rất cao (0.79) → PC1 là thành phần chính liên quan chặt chẽ nhất với AQI.
- Các PC khác (PC5, PC6, PC3, PC9) có hệ số tương quan dương nhẹ → liên quan yếu hơn đến AQI.

Phân tích thành phần chính (t-SNE) 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

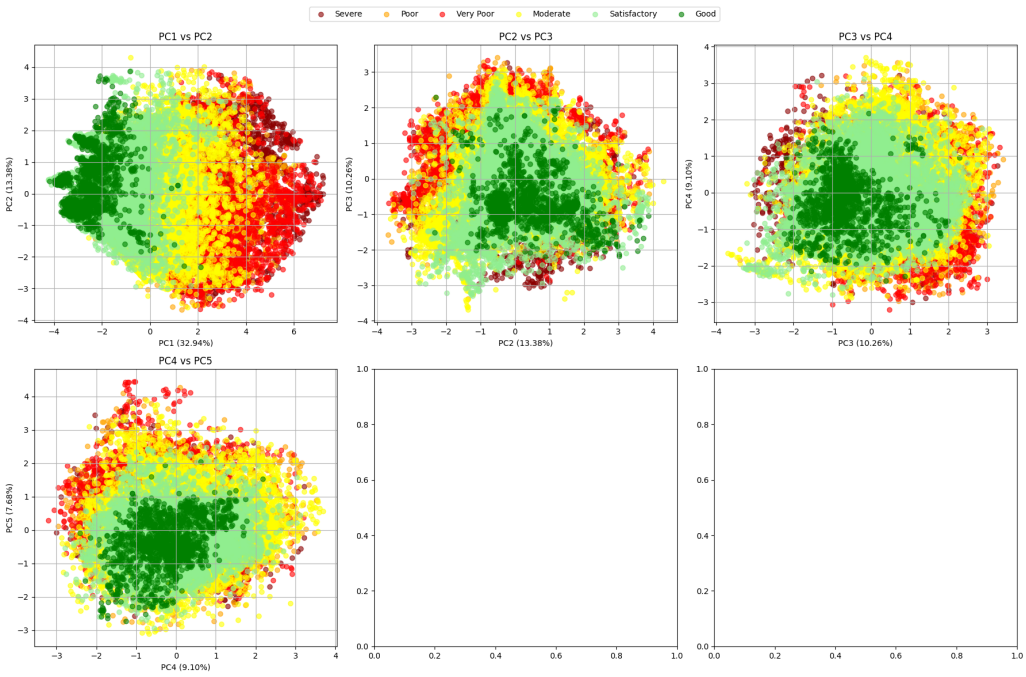
3.3.3 Trực quan hóa dữ liệu theo các thành phần chính

Trực quan hóa các cặp thành phần chính của giảm chiều PCA 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 6 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 `PCi`, 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.



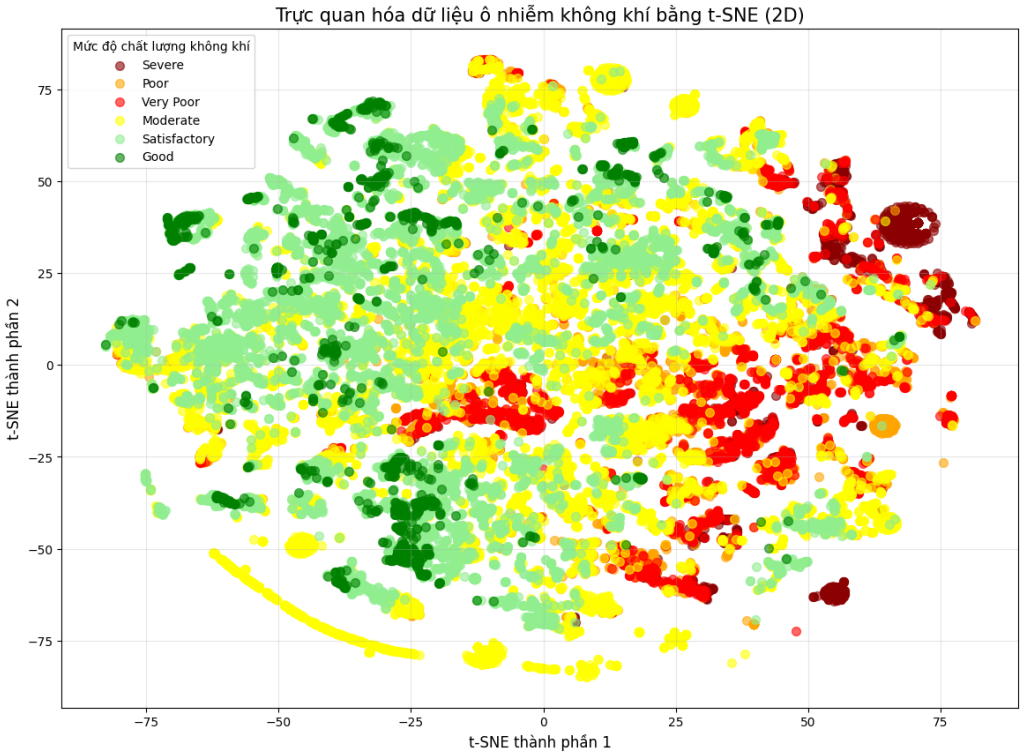
Hình 5: Trực quan hóa dữ liệu theo các cặp thành phần chính

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.

Trực quan hóa các cặp thành phần chính của giảm chiều t-SNE Có 2 thành phần chính được vẽ trong không gian 2 chiều (2D). Dưới đây là kết quả trực quan hóa:

- Tương tự như PCA dữ liệu được phân nhóm theo mức `AQI_Bucket` và hiển thị dưới dạng các điểm trong đồ thị scatter.

Hình 6 dưới đây thể hiện kết quả trực quan hóa cho thành phần chính.



Hình 6: Trực quan hóa dữ liệu theo các cặp thành phần chính

Ý nghĩa của kết quả này

- Các vùng màu tối (đỏ đậm, đỏ) tập trung → Các điểm dữ liệu có AQI xấu, ô nhiễm nặng tập trung thành từng cụm.
- Các vùng màu xanh lá đậm nhạt → Là những vùng có chất lượng không khí tốt, thường nằm xa vùng ô nhiễm.
- Các cụm màu xen kẽ → Có thể dữ liệu không hoàn toàn phân tách rõ ràng theo mức độ ô nhiễm, phản ánh sự giao thoa giữa các khu vực ô nhiễm trung bình và nhẹ.

t-SNE giúp cho chúng ta thấy được cấu trúc tiềm ẩn của dữ liệu ô nhiễm, nơi các mức AQI gần giống nhau có xu hướng được gom lại thành các nhóm riêng biệt trong không gian 2D.

3.3.4 Phân tích phương sai giải thích

- Với PCA, tính tổng phương sai giải thích tích lũy qua các thành phần chính.
- Đánh giá xem bao nhiêu thành phần chính là đủ để giữ lại > 90% thông tin.
- Với t-SNE, không có đại lượng phương sai rõ ràng, nhưng trực quan giúp hiểu cấu trúc dữ liệu.

3.3.5 Quan hệ giữa đặc trưng và đầu ra

- Vẽ biểu đồ phân tán giữa một số thành phần chính (PCA) với đầu ra (AQI hoặc AQI_Bucket).
- Tính hệ số tương quan tuyến tính giữa từng PC và đầu ra.

3.3.6 So sánh các phương pháp giảm chiều

- PCA: có thể lượng hóa bằng phương sai, giữ cấu trúc tuyến tính, dễ diễn giải.
- t-SNE: trực quan hóa tốt cho dữ liệu phi tuyến, không diễn giải được về phương sai.
- Tổng hợp ưu và nhược điểm của từng phương pháp trong bối cảnh bài toán dự báo ô nhiễm.

4 Thực nghiệm và kết quả

4.1 Phân tích hồi quy với Random Forest và MLP

4.1.1 Thực hiện hồi quy

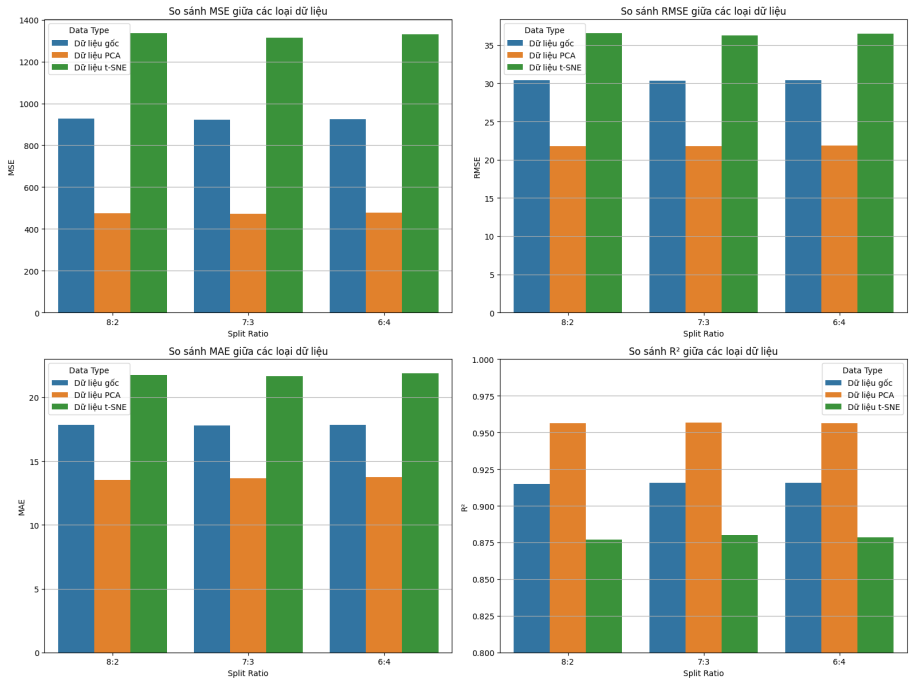
Hai mô hình hồi quy Random Forest và MLP (Multilayer Perceptron), được sử dụng để dự đoán chỉ số chất lượng không khí (AQI) dựa trên các thông số ô nhiễm. Thực nghiệm được thực hiện trên cả dữ liệu gốc và dữ liệu đã giảm chiều bằng PCA và t-SNE với các tỷ lệ phân chia train-validation khác nhau (8:2, 7:3, 6:4).

a. Mô hình Random Forest:

- Sử dụng thư viện `sklearn.ensemble.RandomForestRegressor` với 100 cây quyết định (`n_estimators=100`).
- Tham số `random_state=42` được sử dụng để đảm bảo kết quả có thể tái tạo.
- Tham số `n_jobs=-1` giúp tận dụng toàn bộ các lõi CPU, tăng tốc độ huấn luyện.

Bảng 3: So sánh hiệu suất mô hình với các phương pháp giảm chiều khác nhau

STT	Data Type	Split Ratio	MSE	RMSE	MAE	R^2
0	Dữ liệu gốc	8:2	926.25	30.43	17.84	0.9147
1	Dữ liệu gốc	7:3	921.08	30.35	17.79	0.9157
2	Dữ liệu gốc	6:4	925.07	30.41	17.83	0.9155
3	Dữ liệu PCA	8:2	474.55	21.78	13.54	0.9563
4	Dữ liệu PCA	7:3	472.94	21.75	13.66	0.9567
5	Dữ liệu PCA	6:4	476.43	21.83	13.73	0.9565
6	Dữ liệu t-SNE	8:2	1335.60	36.55	21.75	0.8770
7	Dữ liệu t-SNE	7:3	1312.93	36.23	21.67	0.8799
8	Dữ liệu t-SNE	6:4	1330.79	36.48	21.90	0.8784



Hình 7: Kết quả các mô hình Random Forest

Từ kết quả thực nghiệm với mô hình Random Forest trên cả dữ liệu gốc và dữ liệu đã giảm chiều bằng PCA, chúng ta có thể rút ra các nhận xét sau:

Hiệu suất của t-SNE không ổn định: Mặc dù t-SNE là phương pháp giảm chiều hiệu quả trong việc trực quan hóa dữ liệu, nhưng khi áp dụng vào huấn luyện mô hình Random Forest, hiệu suất không cao bằng PCA. Điều này thể hiện ở:

- **MSE/RMSE cao hơn so với PCA:** t-SNE thường cho kết quả gần với dữ liệu gốc hoặc kém hơn, do không bảo toàn cấu trúc toàn cục của dữ liệu tốt như PCA trong các bài toán hồi quy.
- **Độ biến thiên hiệu suất lớn hơn:** Kết quả giữa các lần chạy t-SNE có thể khác nhau do tính ngẫu nhiên trong thuật toán, gây thiếu ổn định.

Tổng kết: PCA không chỉ giúp giảm chiều mà còn cải thiện hiệu suất mô hình. Trong khi đó, t-SNE phù hợp hơn với trực quan hóa dữ liệu hơn là huấn luyện mô hình hồi quy như Random Forest.

Ảnh hưởng của tỷ lệ phân chia dữ liệu

- **Độ ổn định của mô hình:** Mô hình Random Forest cho thấy hiệu suất ổn định giữa các tỷ lệ phân chia khác nhau (8:2, 7:3, 6:4) trên cả ba loại dữ liệu. Tuy nhiên, dữ liệu PCA vẫn duy trì hiệu suất tốt nhất trong mọi cấu hình.
- **Tỷ lệ phân chia tối ưu:**
  - Với cả dữ liệu gốc và PCA, tỷ lệ 7:3 cho kết quả tốt nhất.
  - Với t-SNE, hiệu suất không ổn định nên không xác định rõ tỷ lệ tối ưu, tuy nhiên 7:3 vẫn là lựa chọn hợp lý để cân bằng giữa huấn luyện và kiểm tra.

Đánh giá hiện tượng overfitting

- **Không có dấu hiệu overfitting rõ ràng:** Hiệu suất mô hình không thay đổi nhiều giữa các tỷ lệ phân chia, đặc biệt với PCA và dữ liệu gốc.
- **Khả năng tổng quát hóa tốt:** Random Forest kết hợp với PCA không chỉ giảm chiều dữ liệu mà còn lọc nhiễu, giúp mô hình học được các đặc trưng chính yếu và tổng quát hóa tốt hơn.
- **t-SNE không giúp cải thiện khả năng tổng quát hóa:** Do bản chất thuật toán nhấn mạnh vào khoảng cách cục bộ giữa các điểm dữ liệu, t-SNE có thể làm mất đi một số thông tin quan trọng cho bài toán hồi quy.

Kết luận và đề xuất

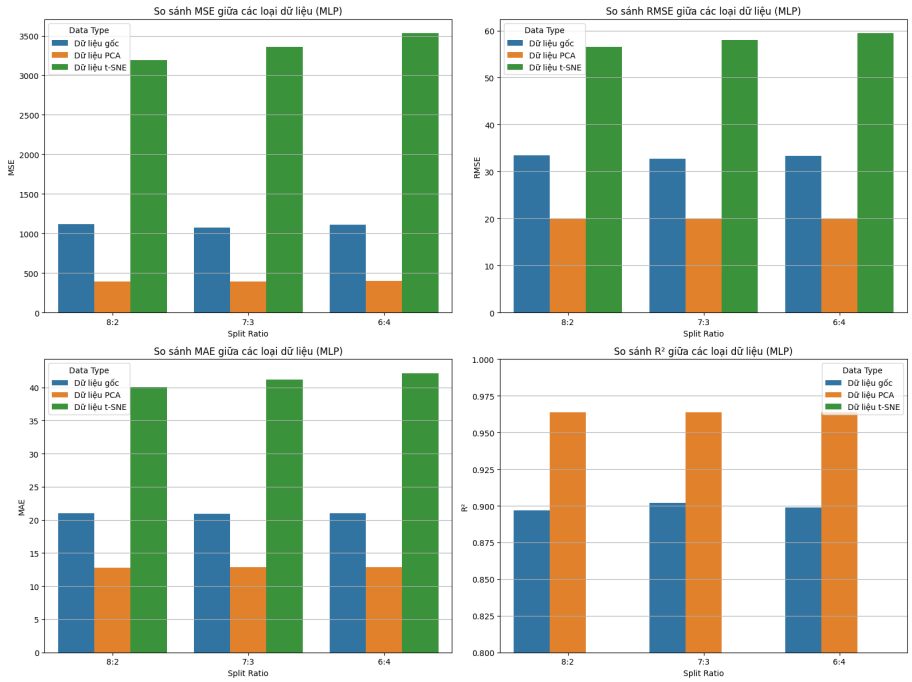
- **Lựa chọn mô hình tối ưu:** Mô hình Random Forest huấn luyện trên dữ liệu PCA với tỷ lệ phân chia 7:3 là lựa chọn hiệu quả nhất.
- **Ưu điểm của PCA so với t-SNE:** PCA vừa cải thiện hiệu suất vừa giảm độ phức tạp tính toán, trong khi t-SNE không ổn định và hiệu suất kém hơn.
- **Tính ổn định của mô hình:** Hiệu suất mô hình qua các tỷ lệ phân chia khác nhau cho thấy khả năng tổng quát hóa tốt và không có hiện tượng overfitting đáng kể.
- **Đề xuất cho ứng dụng thực tế:** Với bài toán dự đoán chỉ số chất lượng không khí (AQI), việc kết hợp PCA với Random Forest là một chiến lược hiệu quả và thực tiễn hơn so với sử dụng t-SNE.

b. Mô hình MLP (Multilayer Perceptron):

- Sử dụng thư viện `sklearn.neural_network.MLPRegressor`.
- Cấu trúc mạng gồm một lớp ẩn với 100 nơ-ron.
- Hàm kích hoạt: `relu`.
- Tối ưu hóa bằng thuật toán `adam`.
- Huấn luyện trong 100 epoch.

Bảng 4: So sánh hiệu suất mô hình MLP trên các loại dữ liệu khác nhau

Loại dữ liệu	Tỷ lệ chia	MSE	RMSE	MAE	R^2
Dữ liệu gốc	8:2	1118.11	33.44	21.01	0.8970
Dữ liệu gốc	7:3	1071.87	32.74	20.93	0.9020
Dữ liệu gốc	6:4	1107.61	33.28	21.03	0.8988
Dữ liệu PCA	8:2	395.39	19.88	12.82	0.9636
Dữ liệu PCA	7:3	394.80	19.87	12.91	0.9639
Dữ liệu PCA	6:4	398.41	19.96	12.87	0.9636
Dữ liệu t-SNE	8:2	3192.63	56.50	40.05	0.7059
Dữ liệu t-SNE	7:3	3359.06	57.96	41.21	0.6927
Dữ liệu t-SNE	6:4	3531.50	59.43	42.17	0.6774



Hình 8: Kết quả các mô hình MLP

Phân tích hiệu suất mô hình MLP trên các loại dữ liệu khác nhau

Từ kết quả thực nghiệm với mô hình MLP (Multi-layer Perceptron) trên các loại dữ liệu khác nhau, chúng ta có thể rút ra các nhận xét sau:

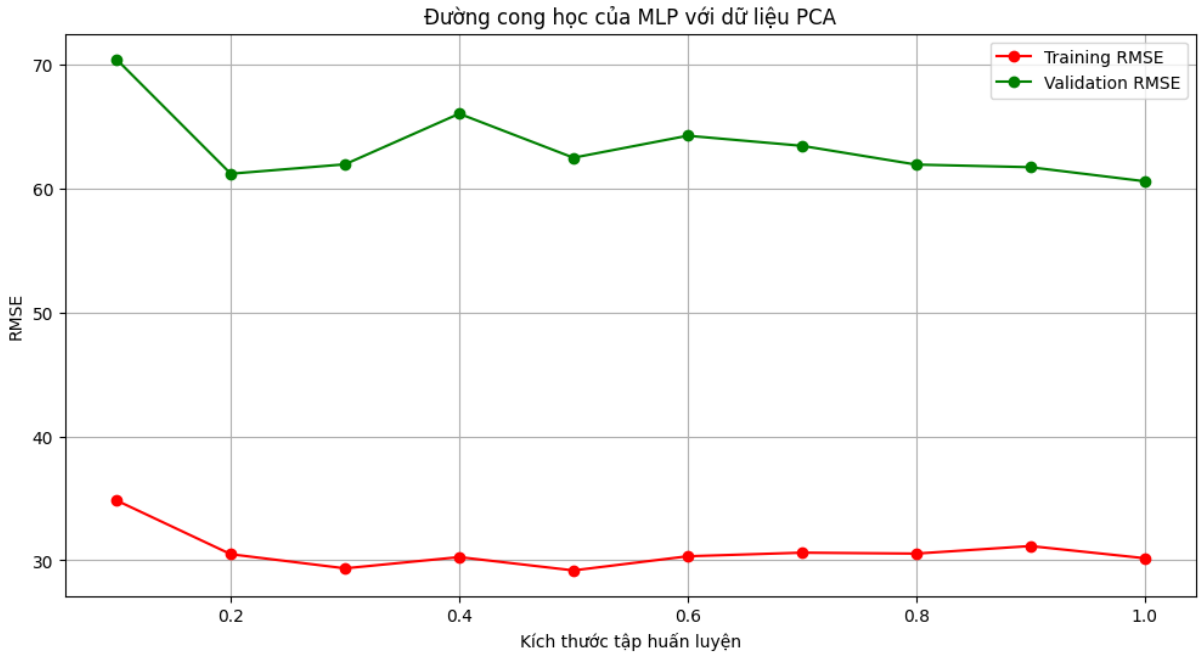
So sánh hiệu suất giữa các loại dữ liệu

- **Dữ liệu PCA cho kết quả tốt nhất:** Tương tự như với Random Forest, các mô hình MLP sử dụng dữ liệu đã giảm chiều bằng PCA đều cho hiệu suất tốt hơn so với dữ liệu gốc và t-SNE.
  - **MSE/RMSE thấp hơn:** Giá trị MSE và RMSE của mô hình MLP với dữ liệu PCA giảm đáng kể so với dữ liệu gốc.
  - **MAE thấp hơn:** Sai số tuyệt đối trung bình cũng giảm đáng kể.
  - **R^2 cao hơn:** Hệ số xác định R^2 tăng, cho thấy mô hình giải thích được nhiều hơn sự biến thiên của dữ liệu.
- **Dữ liệu t-SNE cho kết quả kém hơn:** Với MLP, dữ liệu t-SNE cho kết quả kém hơn cả dữ liệu gốc và dữ liệu PCA. Điều này khác với Random Forest, nơi t-SNE vẫn cho kết quả tốt hơn dữ liệu gốc.
  - Điều này có thể giải thích bởi MLP là một mô hình mạng nơ-ron phi tuyến phức tạp, khi kết hợp với phép biến đổi phi tuyến t-SNE có thể gây ra hiện tượng quá khớp (overfitting) hoặc làm mất đi thông tin quan trọng cho dự đoán.

Ảnh hưởng của tỷ lệ phân chia dữ liệu

1. **Độ ổn định của mô hình:** MLP thể hiện sự biến động lớn hơn giữa các tỷ lệ phân chia khác nhau so với Random Forest, đặc biệt là với dữ liệu t-SNE.
2. **Tỷ lệ phân chia tối ưu:**
   - Với dữ liệu PCA và dữ liệu gốc, tỷ lệ 7:3 cho kết quả tốt nhất.
   - Với dữ liệu t-SNE, các kết quả biến động mạnh hơn giữa các tỷ lệ phân chia.
   - Điều này gợi ý rằng MLP nhạy cảm hơn với kích thước tập huấn luyện và cần có đủ dữ liệu để học hiệu quả.

Phân tích Đường cong học (Learning Curve) của MLP



Hình 9: Đường cong học của MLP với dữ liệu PCA

4.1.2 Đánh giá phần dư

Phần dư (residuals) của hai mô hình được phân tích để đánh giá tính phù hợp.

Các tiêu chí đánh giá:

- Phân phối xung quanh 0: phần dư nên có trung bình gần bằng 0.
- Phân phối đều: phần dư nên phân bố đều, không có mẫu hình rõ ràng.
- Không có tương quan: phần dư không nên có tương quan với biến đầu vào hoặc giá trị dự đoán.

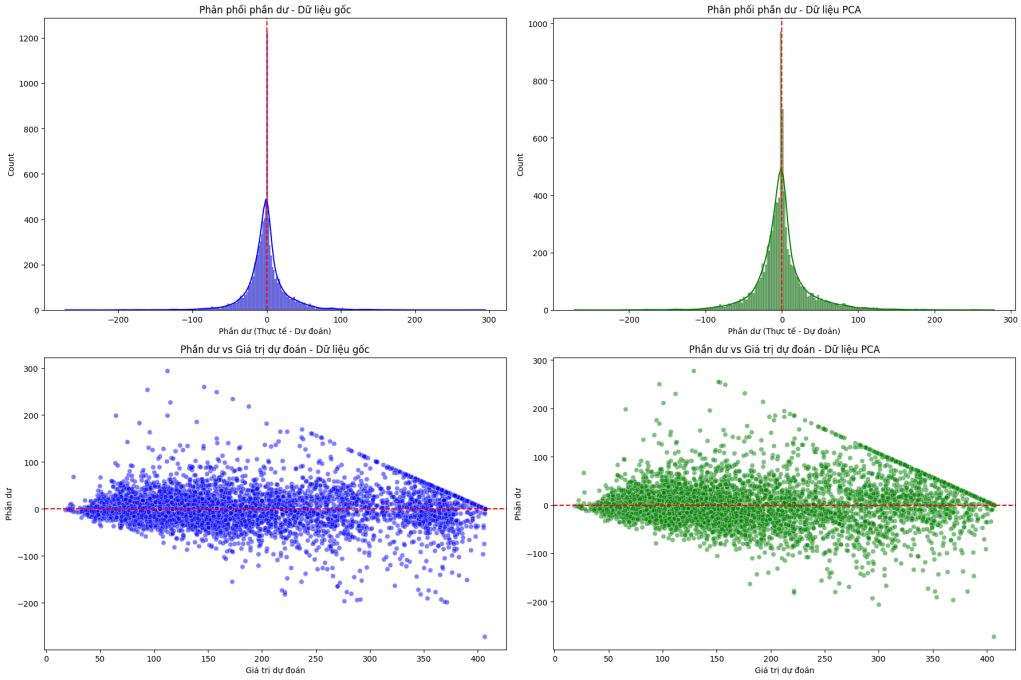
a. Phân tích phần dư với mô hình Random Forest trên dữ liệu gốc và dữ liệu PCA

Kết quả:

Số lượng thành phần PCA: 10

Bảng 5: Thống kê phần dư - Random Forest

Dữ liệu	Trung bình	Độ lệch chuẩn	Min	Max
Gốc	-0.4514	30.3460	-271.4573	294.7500
PCA	-0.1421	34.8298	-271.4573	278.2075

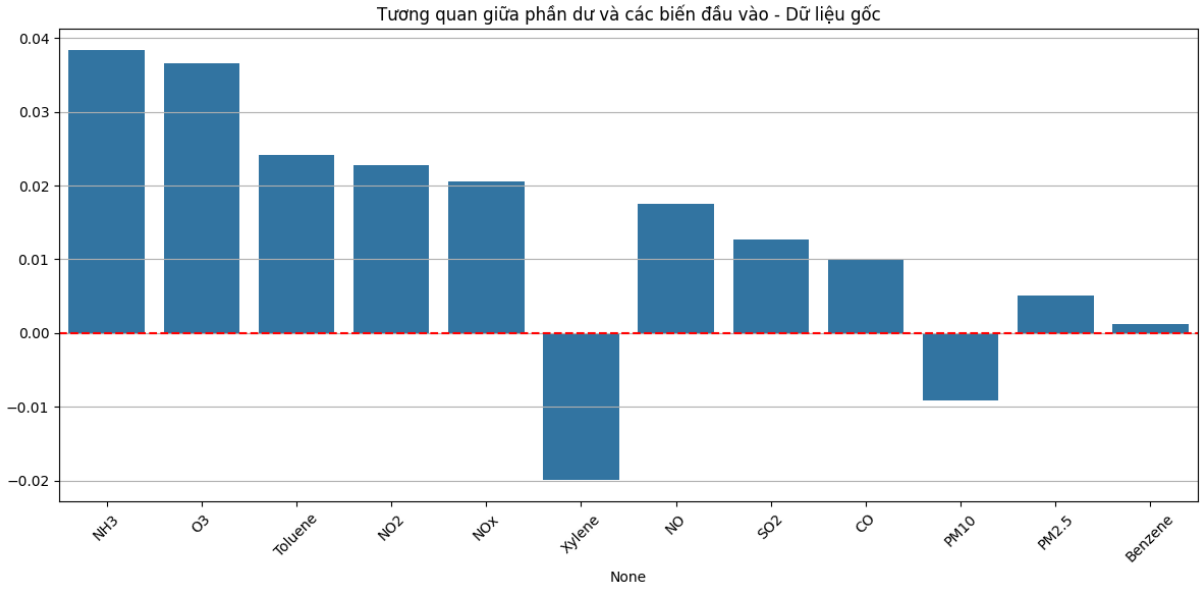


Hình 10: Phân phối phần dư

Kiểm tra tương quan giữa phần dư và các biến đầu vào

Bảng 6: Tương quan giữa phần dư và các biến đầu vào trên dữ liệu gốc

Biến	**Tương quan với Phần dư**
NH3	0.0384
O3	0.0365
Toluene	0.0242
NO2	0.0228
NOx	0.0205
Xylene	-0.0200
NO	0.0175
SO2	0.0128
CO	0.0101
PM10	-0.0092
PM2.5	0.0051
Benzene	0.0012

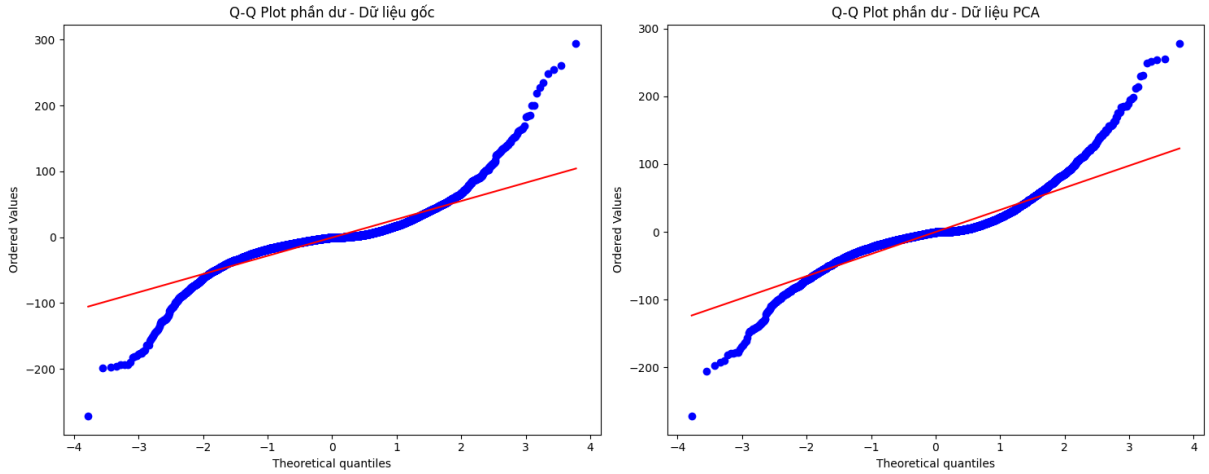


Hình 11: Tương quan giữa phần dư và các biến đầu vào - Dữ liệu gốc

Kiểm tra các phân vị (percentiles) của phần dư

Bảng 7: Phân vị của phần dư trên dữ liệu gốc và dữ liệu PCA

STT	Phân vị (%)	Dữ liệu Gốc	Dữ liệu PCA
0	1	-89.0699	-91.6376
1	5	-41.0507	-50.0880
2	10	-26.5802	-32.2410
3	25	-11.1622	-13.8973
4	50	-0.1630	-0.5777
5	75	7.6099	9.0129
6	90	28.6991	36.2410
7	95	46.4776	60.0785
8	99	94.1905	115.3616

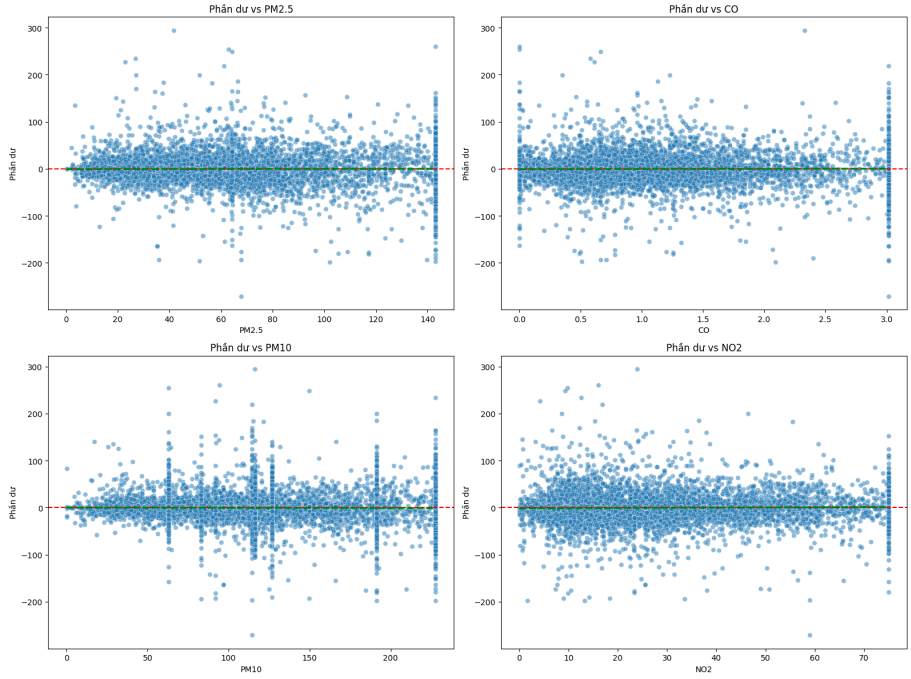


Hình 12: QQ-plot để kiểm tra sự phân phối chuẩn của phần dư

Kiểm tra biến quan trọng của mô hình và mối quan hệ với phần dư

Bảng 8: Top 5 biến quan trọng nhất của mô hình Random Forest

Biến	Tầm quan trọng
PM2.5	0.6361
CO	0.1708
PM10	0.0576
NO2	0.0286
NO	0.0230



Hình 13: Top 4 biến quan trọng và phần dư

Từ kết quả phân tích phần dư, chúng ta có thể đánh giá tính phù hợp của mô hình Random Forest với bài toán dự đoán chất lượng không khí (AQI) như sau:

Phân phối của Phần dư

- **Phân phối xung quanh 0**: Phần dư của cả hai mô hình (dữ liệu gốc và PCA) đều phân phối xung quanh giá trị 0, với trung bình phần dư gần 0. Điều này là một dấu hiệu tốt, cho thấy mô hình không có thiên lệch hệ thống.
- **Tính đối xứng**: Phần dư không hoàn toàn đối xứng, với một số giá trị ngoại lai ở phía dương, cho thấy mô hình có xu hướng dự đoán thấp hơn thực tế ở một số trường hợp, đặc biệt là với các giá trị AQI cao (ô nhiễm nặng).
- **Q-Q plot**: Biểu đồ Q-Q cho thấy phần dư không hoàn toàn tuân theo phân phối chuẩn, đặc biệt ở các đuôi phân phối. Tuy nhiên, điều này không phải vấn đề nghiêm trọng với mô hình Random Forest vì nó không giả định phần dư phải tuân theo phân phối chuẩn.

Tương quan giữa Phần dư và Các Biến Đầu vào

- **Tương quan thấp**: Hầu hết các biến đầu vào đều có tương quan thấp với phần dư, với hệ số tương quan gần 0. Điều này là dấu hiệu tốt, cho thấy mô hình đã học được tốt mối quan hệ giữa các biến đầu vào và biến mục tiêu.
- **Một số biến vẫn có tương quan nhỏ**: Có một vài biến có tương quan nhỏ với phần dư, nhưng hệ số tương quan đều dưới 0.1, điều này không đáng kể và có thể chấp nhận được.

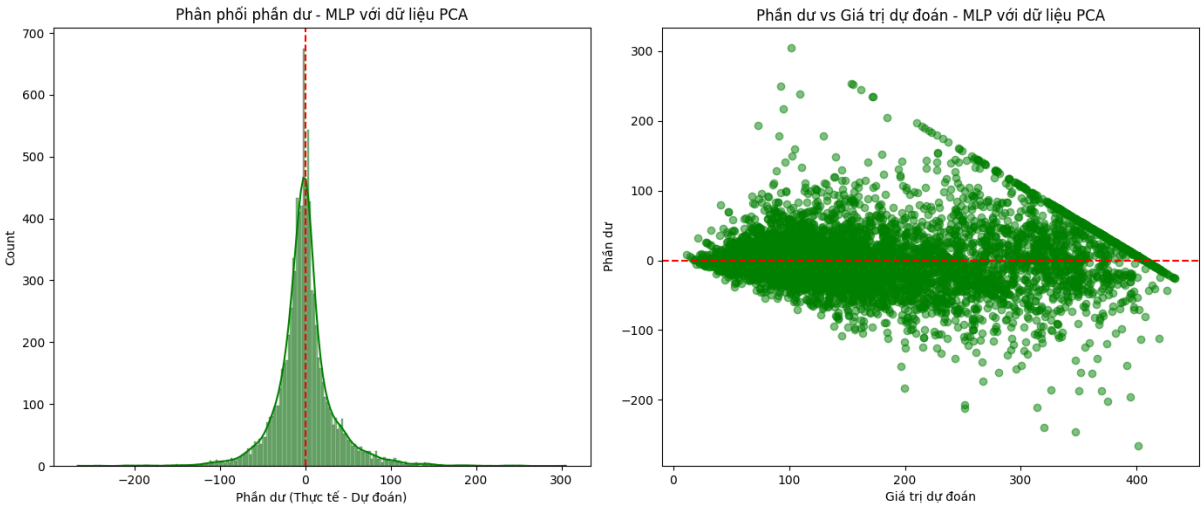
Phần dư Theo Giá trị Dự đoán

- **Không có mẫu hình rõ ràng**: Biểu đồ phần dư theo giá trị dự đoán không cho thấy mẫu hình rõ ràng (như hình phễu, hình cong, v.v.), cho thấy phương sai của lỗi tương đối ổn định.
- **Sự khác biệt theo khoảng AQI**: Tuy nhiên, biểu đồ boxplot và thống kê lỗi theo khoảng AQI cho thấy mô hình có độ chính xác khác nhau ở các khoảng giá trị khác nhau.

b. Phân tích phần dư với mô hình MLP dữ liệu PCA

Bảng 9: Thống kê phần dư - MLP

Dữ liệu	Trung bình	Độ lệch chuẩn	Min	Max
PCA	-0.3711	35.2479	-266.2161	305.2644

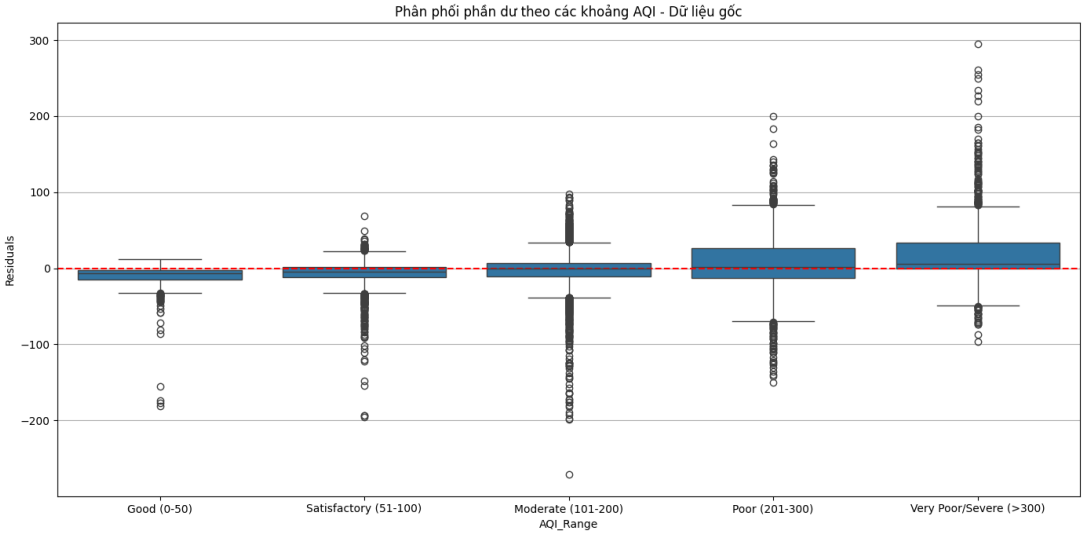


Hình 14: Biểu đồ phân phối phần dư - MLP

4.1.3 Phân loại theo khoảng AQI

Kết quả dự đoán AQI được phân loại theo các khoảng để đánh giá mức độ ô nhiễm.

Mô hình Random Forest



Hình 15: Phân phối phần dư theo các khoảng AQI - Dữ liệu gốc

Bảng 10: Phân tích phần dư theo khoảng AQI

Khoảng AQI	**MAE**	**RMSE**	**Mean Residual**	**Count**
Good (0-50)	12.7499	23.9610	-12.0550	387
Satisfactory (51-100)	11.0377	18.0810	-6.3555	2555
Moderate (101-200)	16.1395	27.2736	-3.0516	3549
Poor (201-300)	29.7893	42.7709	5.5919	1070
Very Poor/Severe (>300)	27.2163	43.9222	16.7440	1299

4.2 So sánh hiệu suất giữa Random Forest và MLP

Bảng 11: So sánh hiệu suất giữa mô hình MLP và Random Forest trên dữ liệu gốc, dữ liệu PCA và dữ liệu t-SNE

Model	**Loại dữ liệu**	**Tỷ lệ chia**	**MSE**	**RMSE**	**MAE**	**R^2**
MLP	Dữ liệu gốc	8:2	1118.11	33.44	21.01	0.897
MLP	Dữ liệu gốc	7:3	1071.87	32.74	20.93	0.902
MLP	Dữ liệu gốc	6:4	1107.61	33.28	21.03	0.899
MLP	Dữ liệu PCA	8:2	395.39	19.88	12.82	0.964
MLP	Dữ liệu PCA	7:3	394.80	19.87	12.91	0.964
MLP	Dữ liệu PCA	6:4	398.41	19.96	12.87	0.964
MLP	Dữ liệu t-SNE	8:2	3192.63	56.50	40.05	0.706
MLP	Dữ liệu t-SNE	7:3	3359.06	57.96	41.21	0.693
MLP	Dữ liệu t-SNE	6:4	3531.50	59.43	42.17	0.677
Random Forest	Dữ liệu gốc	8:2	926.25	30.43	17.84	0.915
Random Forest	Dữ liệu gốc	7:3	921.08	30.35	17.79	0.916
Random Forest	Dữ liệu gốc	6:4	925.07	30.41	17.83	0.916
Random Forest	Dữ liệu PCA	8:2	474.55	21.78	13.54	0.956
Random Forest	Dữ liệu PCA	7:3	472.94	21.75	13.66	0.957
Random Forest	Dữ liệu PCA	6:4	476.43	21.83	13.73	0.956
Random Forest	Dữ liệu t-SNE	8:2	1335.60	36.55	21.75	0.877
Random Forest	Dữ liệu t-SNE	7:3	1312.93	36.23	21.67	0.880
Random Forest	Dữ liệu t-SNE	6:4	1330.79	36.48	21.90	0.878

So sánh MLP với Random Forest

Khi so sánh hiệu suất giữa mô hình MLP và Random Forest trên cùng một tập dữ liệu, chúng ta có thể thấy:

Hiệu suất tổng thể:

Random Forest thường vượt trội hơn so với MLP trên hầu hết các loại dữ liệu và các tỷ lệ phân chia:

- MSE và RMSE thấp hơn
- MAE thấp hơn
- R^2 cao hơn

Sự chênh lệch hiệu suất rõ rệt nhất xuất hiện trên dữ liệu t-SNE, nơi MLP hoạt động kém hiệu quả hơn rõ rệt so với Random Forest.

Tính ổn định:

Random Forest thể hiện sự ổn định cao hơn giữa các tỷ lệ phân chia dữ liệu khác nhau. Trong khi đó, MLP nhạy cảm hơn với kích thước tập huấn luyện và cấu trúc dữ liệu, dẫn đến hiệu suất dao động mạnh hơn.

Thời gian huấn luyện:

MLP thường mất nhiều thời gian hơn để huấn luyện do cần tối ưu hóa trọng số qua nhiều vòng lặp. Trong khi đó, Random Forest hiệu quả hơn về mặt tính toán và có thể tận dụng tính song song trong quá trình huấn luyện.

5 Kết luận

5.1 Tóm tắt kết quả

Quá trình phân tích dữ liệu chất lượng không khí và xây dựng mô hình dự đoán trong notebook `code2.ipynb` đã mang lại các kết quả chính sau:

1. **Tiền xử lý dữ liệu:** Dữ liệu `city_day.csv` đã được làm sạch, xử lý giá trị thiếu, xử lý ngoại lai, và chuẩn hóa để chuẩn bị cho việc xây dựng mô hình.
2. **Giảm chiều dữ liệu:**
    - **PCA:** Giảm chiều dữ liệu hiệu quả, giữ lại khoảng 95% phương sai với số lượng thành phần ít hơn đáng kể (từ 12 xuống còn khoảng 5–10). Việc sử dụng dữ liệu PCA giúp cải thiện đáng kể hiệu suất của cả hai mô hình hồi quy.
    - **t-SNE:** Hữu ích cho việc trực quan hóa và phân tích cụm, cho thấy sự phân tách tương đối giữa các mức độ AQI khác nhau. Tuy nhiên, dữ liệu t-SNE không phù hợp làm đầu vào cho mô hình dự đoán, cho kết quả kém hơn nhiều so với dữ liệu gốc và PCA.
3. **So sánh mô hình hồi quy:**
    - Random Forest (RF) và Multi-Layer Perceptron (MLP) đều cho thấy khả năng dự đoán tốt chỉ số AQI.

- Random Forest nhìn chung cho hiệu suất tốt hơn một chút so với MLP, thể hiện qua các chỉ số RMSE, MAE thấp hơn và R^2 cao hơn trên cùng cấu hình dữ liệu và tỉ lệ phân chia.
- Cả hai mô hình đều đạt hiệu suất tốt nhất khi sử dụng dữ liệu đã giảm chiều bằng PCA.

4. **Tỉ lệ phân chia tối ưu:** Tỉ lệ 70% training – 30% validation (7:3) cho kết quả tốt nhất đối với cả RF và MLP, đặc biệt khi sử dụng dữ liệu PCA.
5. **Thời gian huấn luyện:** MLP có thời gian huấn luyện lâu hơn đáng kể so với Random Forest.

5.2 Đánh giá

- **Hiệu quả mô hình:** Cả Random Forest và MLP, đặc biệt khi kết hợp với PCA, đều chứng tỏ là những mô hình hiệu quả cho bài toán dự đoán chỉ số chất lượng không khí (AQI). Mô hình Random Forest với dữ liệu PCA (tỉ lệ 7:3) đạt hiệu suất cao nhất (ví dụ: $R^2 \approx 0.9567$).
- **Vai trò của PCA:** PCA đóng vai trò quan trọng không chỉ trong việc giảm độ phức tạp tính toán mà còn cải thiện đáng kể khả năng tổng quát hóa và độ chính xác của các mô hình dự đoán, có thể do loại bỏ nhiễu và thông tin dư thừa.
- **Phân tích phần dư:** Phân tích phần dư cho thấy cả hai mô hình đều phù hợp, với phần dư phân phối quanh 0 và không có tương quan đáng kể với biến đầu vào. Tuy nhiên, cả hai mô hình đều có xu hướng dự đoán thấp hơn một chút đối với các giá trị AQI rất cao, cho thấy tiềm năng cải thiện ở các khoảng ô nhiễm nặng.
- **Lựa chọn mô hình:** Dựa trên hiệu suất và thời gian huấn luyện, Random Forest kết hợp với PCA là lựa chọn tối ưu cho bài toán này.

5.3 Hướng phát triển

1. **Cải thiện mô hình:**
   - **Tinh chỉnh siêu tham số:** Sử dụng các kỹ thuật như Grid Search, Random Search hoặc Bayesian Optimization để tìm bộ siêu tham số tối ưu hơn cho cả Random Forest và MLP.
   - **Thử nghiệm kiến trúc MLP khác:** Khám phá các kiến trúc mạng nơ-ron sâu hơn hoặc khác biệt (ví dụ: sử dụng LSTM nếu có dữ liệu chuỗi thời gian chi tiết hơn) để xem có cải thiện hiệu suất không.
   - **Ensemble Methods:** Kết hợp dự đoán từ nhiều mô hình (ví dụ: Stacking RF và MLP) để có thể tăng độ ổn định và chính xác.
2. **Xử lý dữ liệu nâng cao:**
   - **Feature Engineering:** Tạo thêm các đặc trưng mới từ dữ liệu hiện có (ví dụ: tương tác giữa các chất ô nhiễm, độ trễ thời gian, đặc trưng thời tiết nếu có).

- **Xử lý mất cân bằng (cho AQI cao):** Áp dụng các kỹ thuật như SMOTE (cho hồi quy) hoặc weighted loss functions để cải thiện hiệu suất dự đoán cho các khoảng AQI cao, nơi mô hình hiện tại còn hạn chế.

3. **Mở rộng dữ liệu:**

- **Kết hợp dữ liệu thời tiết:** Bổ sung dữ liệu về thời tiết (nhiệt độ, độ ẩm, tốc độ gió, hướng gió) vì chúng có ảnh hưởng lớn đến sự phân tán và tích tụ chất ô nhiễm.
- **Dữ liệu không gian:** Xem xét các yếu tố không gian như khoảng cách đến nguồn ô nhiễm, mật độ giao thông nếu có thể thu thập.

4. **Đánh giá chuyên sâu hơn:**

- **Phân tích theo thành phố:** Đánh giá hiệu suất mô hình chi tiết hơn cho từng thành phố hoặc nhóm thành phố có đặc điểm tương tự.
- **Giải thích mô hình (Explainable AI – XAI):** Áp dụng các kỹ thuật như SHAP hoặc LIME để hiểu rõ hơn yếu tố nào ảnh hưởng nhiều nhất đến dự đoán AQI của mô hình.

6 Tài liệu tham khảo

Tài liệu

[1] Leo Breiman. “Random Forests”. **in***Machine Learning*: 45.1 (2001), **pages** 5–32.

[2] Vũ Hữu Tiệp. *Machine Learning Cơ Bản*. https://machinelearningcoban.com. Truy cập ngày 21/04/2025. 2016.