BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT HƯNG YÊN

BÀI TẬP LỚN

DỰ ĐOÁN MỨC ĐỘ Ô NHIỄM KHÔNG KHÍ

NGÀNH: CÔNG NGHỆ THÔNG TIN

CHUYÊN NGÀNH: KHOA HỌC MÁY TÍNH

SINH VIÊN: **NGUYỄN MINH HIẾU**

MÃ SINH VIÊN:12423049

MÃ LỚP: 124231

GV GIẢNG DẠY: **PGS. TS. NGUYỄN VĂN HẬU**

HƯNG YÊN – 2025

NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

Hưng Yên, ngày tháng năm 2025

(Ký và ghi rõ họ tên)

LỜI CAM ĐOAN

Em xin cam đoan bài tập lớn "Phân loại mức độ ô nhiễm không khí" là kết quả thực hiện của bản thân em dưới sự hướng dẫn của Thầy Nguyễn Văn Hậu.

Những phần sử dụng tài liệu tham khảo trong bài tập lớn đã được nêu rõ trong phần tài liệu tham khảo. Các kết quả trình bày trong bài tập lớn và chương trình xây dựng được hoàn toàn là kết quả do bản thân em thực hiện.

Nếu vi phạm lời cam đoan này, nhóm em xin chịu hoàn toàn trách nhiệm trước khoa và nhà trường.

Hưng Yên, ngày … tháng … năm…..

SINH VIÊN

MỤC LỤC

DANH MỤC CÁC KÝ TỰ, CÁC TỪ VIẾT TẮT

STT	Từ viết tắt	Cụm từ tiếng anh	Diễn giải
1	AQI	Air Quality Index	Chỉ số chất lượng không khí
2	PM2.5	Particulate Matter 2.5	Hạt bụi mịn có đường kính 2.5µm
3	PM10	Particulate Matter 10	Hạt bụi có đường kính 10µm
4	CO	Carbon Monoxide	Khí Carbon monoxide
5	NO_2	Nitrogen Dioxide	Khí Nitơ dioxide
6	O_3	Ozone	Khí Ozone
7	SO_2	Sulfur Dioxide	Khí Lưu huỳnh dioxide
8	RMSE	Root Mean Squared Error	Sai số bình phương trung bình căn bậc hai
9	MAE	Mean Absolute Error	Sai số tuyệt đối trung bình
10	AUC	Area Under Curve	Diện tích dưới đường cong ROC
11	LSTM	Long Short-Term Memory	Mạng nhớ dài ngắn hạn
12	GRU	Gated Recurrent Unit	Đơn vị hồi quy có cổng

DANH MỤC CÁC HÌNH VẼ

CHƯƠNG 1: CƠ SỞ LÝ THUYẾT

1.1. Tổng quan về học máy

Học máy là một nhánh quan trọng của trí tuệ nhân tạo, cho phép máy tính học hỏi từ dữ liệu và cải thiện hiệu suất mà không cần lập trình rõ ràng từng bước. Trong lĩnh vực này, trí tuệ nhân tạo là khái niệm rộng nhất nhằm tạo ra các hệ thống thông minh, học máy là tập con chính sử dụng dữ liệu để huấn luyện mô hình, còn học sâu là một phần nâng cao của học máy dựa trên mạng nơ-ron nhân tạo nhiều lớp. Học máy được chia thành ba dạng chính: học có giám sát sử dụng dữ liệu có nhãn để dự đoán, học không giám sát tìm kiếm cấu trúc ẩn trong dữ liệu không nhãn, và học tăng cường dựa trên phần thưởng từ môi trường. Học máy hiện nay đóng vai trò quan trọng trong nhiều lĩnh vực thực tiễn như y tế, tài chính, giao thông và môi trường, đặc biệt là dự báo chất lượng không khí để bảo vệ sức khỏe cộng đồng

1.2. Phát biểu bài toán

Bài toán được đặt ra trong báo cáo này là dự đoán chỉ số chất lượng không khí (AQI) tại Hà Nội dựa trên các yếu tố ô nhiễm và thời tiết. Đây là bài toán hồi quy vì đầu ra AQI là giá trị liên tục phản ánh mức độ ô nhiễm. Đầu vào bao gồm các đặc trưng như nồng độ CO, NO2, O3, PM10, PM25, SO2 cùng các yếu tố thời tiết như nhiệt độ, độ ẩm, tốc độ gió, lượng mưa, áp suất, mây che phủ và chỉ số UV. Mục tiêu của mô hình là dự báo chính xác giá trị AQI để hỗ trợ cảnh báo sớm ô nhiễm. Bài toán này mang ý nghĩa thực tiễn lớn vì ô nhiễm không khí tại Hà Nội thường xuyên ở mức cao, ảnh hưởng trực tiếp đến sức khỏe cư dân, với các mức AQI từ 0-50 là tốt, 51-100 trung bình, 101-150 không tốt cho nhóm nhạy cảm, 151-200 có hại, 201-300 rất có hại và trên 300 là nguy hiểm.

1.3. Mô hình học máy được lựa chọn

Mô hình được lựa chọn chính là Random Forest Regressor, kết hợp với Linear Regression để so sánh. Random Forest là thuật toán ensemble xây dựng nhiều cây quyết định độc lập trên các mẫu dữ liệu ngẫu nhiên, sau đó lấy trung bình dự đoán để giảm phương sai và tránh hiện tượng quá khít. Lý do chọn Random Forest nằm ở khả năng xử lý tốt mối quan hệ phi tuyến tính giữa các yếu tố ô nhiễm và AQI, đồng thời cung cấp thông tin về tầm quan trọng của từng đặc trưng. Mô hình này có ưu điểm nổi bật là

độ chính xác cao, ít bị overfit nhờ cơ chế bootstrap và random feature selection, không nhạy cảm với việc chuẩn hóa dữ liệu và dễ dàng đánh giá tầm quan trọng đặc trưng.

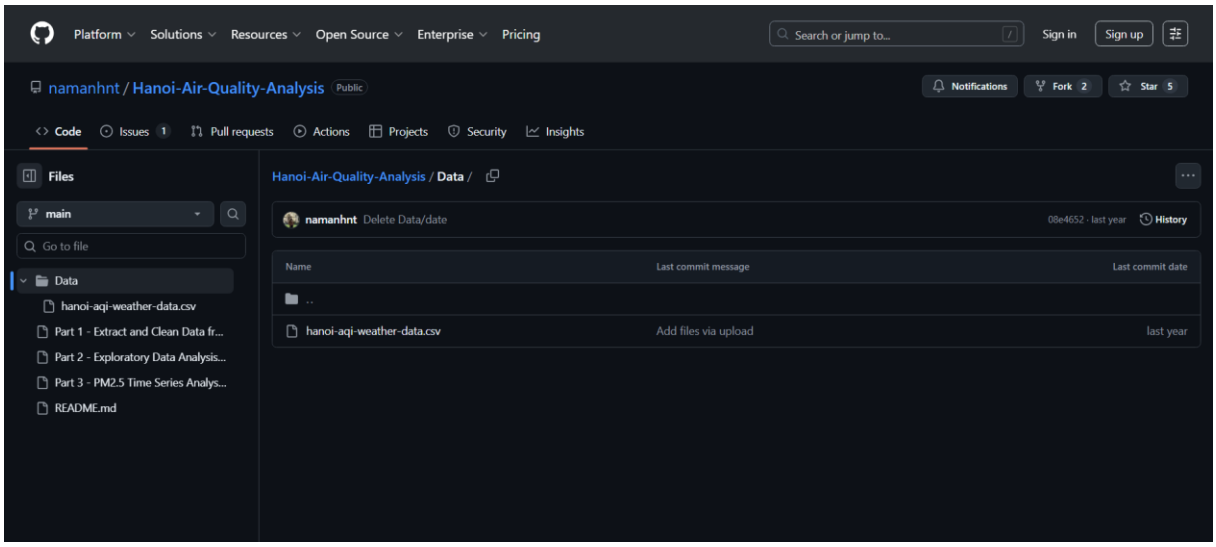
1.4. Nguyên lý hoạt động và giả thuyết

Về nguyên lý toán học, Random Forest sử dụng Gini impurity hoặc mean squared error để đo độ không tinh khiết của nút và chọn chia nhánh tối ưu theo thuật toán CART. Hàm mất mát thường là mean squared error trong bài toán hồi quy. Mô hình không yêu cầu giả thiết nghiêm ngặt về phân bố dữ liệu hay tuyến tính, chỉ cần các cây quyết định đa dạng và độc lập. Hạn chế chính là với dữ liệu lớn, thời gian huấn luyện có thể kéo dài dù có thể song song hóa. Linear Regression đơn giản hơn, giả định mối quan hệ tuyến tính và độc lập giữa các biến, nhưng kém hiệu quả khi dữ liệu có tính phi tuyến cao như trường hợp này.

CHƯƠNG 2: DỮ LIỆU VÀ TIỀN XỬ LÝ

2.1 Giới thiệu bộ dữ liệu

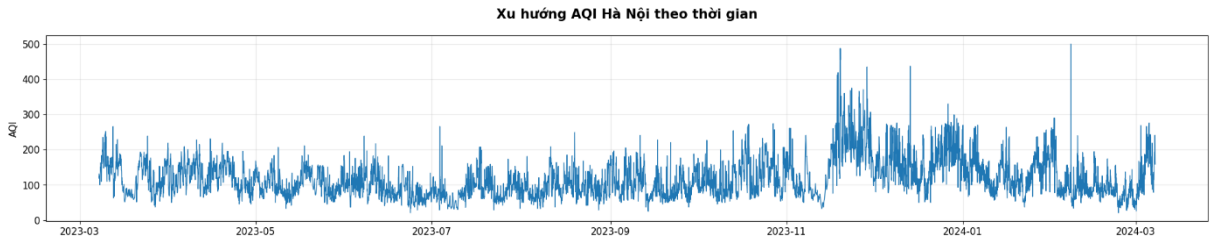
Bộ dữ liệu được sử dụng lấy từ nguồn mở trên GitHub của tác giả namanhnt, chứa thông tin chất lượng không khí và thời tiết tại Hà Nội theo giờ từ năm 2023 đến đầu 2024. Bộ dữ liệu gồm 8785 mẫu với 18 cột, bao gồm thời gian UTC, thành phố, mã quốc gia, múi giờ, chỉ số AQI cùng nồng độ các chất ô nhiễm chính CO, NO2, O3, PM10, PM25, SO2 và các yếu tố thời tiết như mây che phủ, lượng mưa, áp suất, độ ẩm tương đối, nhiệt độ, chỉ số UV và tốc độ gió.



Hình 2. 1 Nguồn dữ liệu được sử dụng

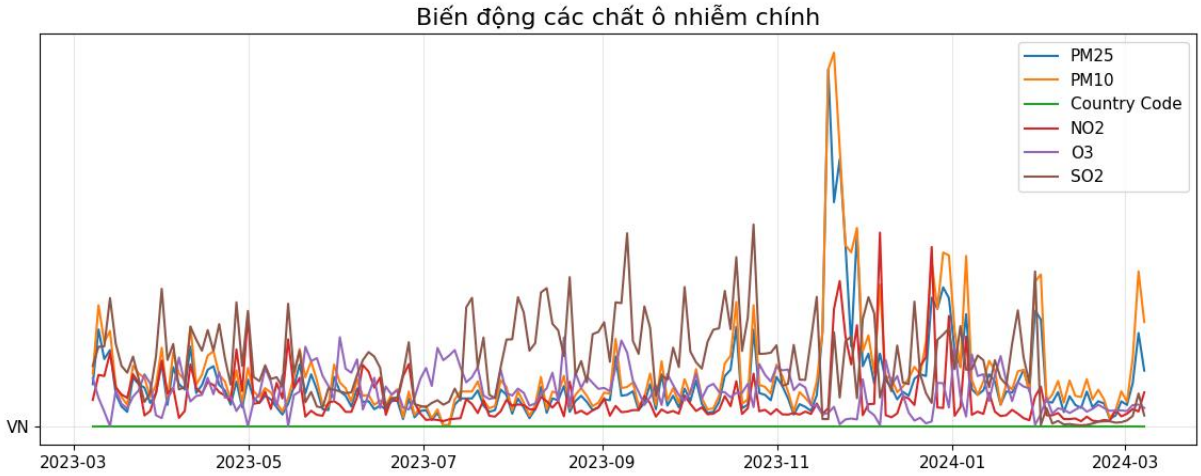
2.2 Phân tích dữ liệu ban đầu.

Phân tích khám phá ban đầu cho thấy chỉ số AQI biến động mạnh theo thời gian, với các đỉnh ô nhiễm cao thường xuất hiện vào cuối năm. Phân bố AQI tập trung chủ yếu ở mức dưới 200 nhưng có đuôi dài về phía các giá trị cao, phản ánh sự xuất hiện của các đợt ô nhiễm nghiêm trọng trong một số giai đoạn nhất định. Điều này cho thấy chất lượng không khí có tính mùa vụ rõ rệt và không ổn định theo thời gian.



Hình 2. 2 Xu hướng ô nhiễm theo thời gian

Trong số các yếu tố tác động, PM2.5 và PM10 là hai thành phần chính ảnh hưởng mạnh nhất đến AQI. Mối quan hệ giữa PM2.5 và AQI thể hiện rõ tính phi tuyến, khi AQI tăng nhanh nếu nồng độ PM2.5 vượt qua một ngưỡng nhất định. Ma trận tương quan cho thấy AQI có tương quan cao với PM10 và PM2.5, trong khi nhiệt độ và độ ẩm có tương quan âm nhẹ. Ngoài ra, tốc độ gió và nhiệt độ có xu hướng ảnh hưởng ngược chiều đến mức độ ô nhiễm, góp phần làm giảm AQI khi các yếu tố này tăng lên.



Hình 2. 3 Các yếu tố gây ô nhiễm

2.3 Tiền xử lý dữ liệu

Quá trình tiền xử lý dữ liệu được tiến hành nhằm đảm bảo tính phù hợp và hiệu quả cho giai đoạn xây dựng mô hình. Trước hết, các cột không cần thiết như thông tin địa lý và các biến thời gian không được sử dụng làm đặc trưng dự báo được loại bỏ khỏi tập dữ liệu. Tiếp theo, biến mục tiêu là chỉ số AQI được tách riêng khỏi tập các đặc trưng đầu vào để phục vụ cho quá trình huấn luyện mô hình.

Dữ liệu được kiểm tra và làm sạch để đảm bảo tính nhất quán và độ tin cậy trước khi đưa vào mô hình. Do toàn bộ các đặc trưng đều ở dạng số, nên không cần thực hiện bước mã hóa dữ liệu. Việc chuẩn hóa chỉ được cân nhắc khi áp dụng các mô hình tuyến tính như Linear Regression nhằm đảm bảo hiệu quả huấn luyện.

CHƯƠNG 3: XÂY DỰNG VÀ HUẤN LUYỆN MÔ HÌNH

3.1. Qúa trình xây dựng mô hình

Quy trình xây dựng mô hình được thiết kế theo hướng tuân thủ đặc thù của dữ liệu chuỗi thời gian nhằm đảm bảo tính khách quan và khả năng tổng quát hóa của mô hình. Cụ thể, tập dữ liệu được chia theo thứ tự thời gian, trong đó 80% các quan sát đầu tiên (tương ứng 7.028 mẫu) được sử dụng làm tập huấn luyện, và 20% còn lại (1.757 mẫu) được dùng làm tập kiểm tra. Cách chia này giúp tránh hiện tượng rò rỉ thông tin từ tương lai vào quá trình huấn luyện, vốn là một vấn đề phổ biến trong các bài toán dự báo theo thời gian.

Pipeline huấn luyện được xây dựng theo cấu trúc đơn giản nhưng hiệu quả, bao gồm các bước chính: tải dữ liệu, thực hiện tiền xử lý cơ bản và áp dụng mô hình học máy. Việc sử dụng pipeline giúp đảm bảo tính nhất quán trong quá trình huấn luyện và đánh giá, đồng thời thuận tiện cho việc mở rộng hoặc thay đổi mô hình trong các thử nghiệm tiếp theo.

3.2. Huấn luyến mô hình

Trong giai đoạn huấn luyện, mô hình Random Forest Regressor được lựa chọn làm mô hình chính do khả năng học các mối quan hệ phi tuyến và tính ổn định cao trước nhiễu dữ liệu. Mô hình ban đầu được huấn luyện với các tham số mặc định, bao gồm 100 cây quyết định và không giới hạn độ sâu tối đa. Nhờ kích thước dữ liệu ở mức vừa phải, quá trình huấn luyện diễn ra nhanh chóng và không đòi hỏi tài nguyên tính toán lớn.

Bên cạnh đó, mô hình Linear Regression cũng được huấn luyện như một mô hình cơ sở (baseline) nhằm phục vụ mục đích so sánh hiệu quả dự báo. Linear Regression đại diện cho nhóm mô hình tuyến tính đơn giản, qua đó giúp đánh giá mức độ cải thiện khi sử dụng các mô hình phi tuyến phức tạp hơn. Thời gian huấn luyện của cả hai mô hình đều chỉ mất vài giây trên máy tính thông thường, cho thấy tính khả thi trong triển khai thực tế.

3.3. Điều chỉnh tham số

Để nâng cao hiệu suất dự báo, quá trình điều chỉnh siêu tham số cho mô hình Random Forest được tiến hành một cách có hệ thống. Các siêu tham số chính được xem xét bao

gồm số lượng cây trong rừng (number of estimators), độ sâu tối đa của mỗi cây (maximum depth) và số lượng đặc trưng được chọn ngẫu nhiên tại mỗi nút chia (max features).

2.4 Độ phức tạp và chi phí tính toán

Xét về độ phức tạp tính toán, Random Forest có độ phức tạp thời gian huấn luyện tỷ lệ thuận với số lượng cây, số mẫu huấn luyện và logarit của số mẫu, do mỗi cây quyết định được xây dựng độc lập. Chi phí bộ nhớ cũng tăng theo số lượng cây cần lưu trữ trong mô hình. Tuy nhiên, nhờ đặc tính huấn luyện song song của các cây, Random Forest có khả năng mở rộng tốt và phù hợp với các tập dữ liệu lớn hơn khi tài nguyên tính toán cho phép.

Ngược lại, Linear Regression có độ phức tạp thấp hơn đáng kể cả về thời gian huấn luyện lẫn yêu cầu bộ nhớ. Mặc dù đơn giản và hiệu quả về mặt tính toán, mô hình tuyến tính này gặp hạn chế trong việc biểu diễn các mối quan hệ phi tuyến phức tạp giữa các biến, vốn tồn tại rõ rệt trong dữ liệu chất lượng không khí.

CHƯƠNG 4: ĐÁNH GIÁ VÀ KẾT LUẬN

4.1. Đánh giá kết quả của mô hình Linear Regression

```
LINEAR REGRESSION
              precision    recall  f1-score   support

    negative       0.91      0.58      0.71       744
    positive       0.76      0.96      0.84      1013

    accuracy                           0.80      1757
   macro avg       0.83      0.77      0.78      1757
weighted avg       0.82      0.80      0.79      1757

Confusion Matrix:
[[430 314]
 [ 43 970]]
Accuracy : 0.7968
AUC : 0.500000000000
```

Hình 4. 1 Kết quả training Linear Regression

Mô hình Linear Regression đạt độ chính xác 80% trên tập kiểm tra nhưng thể hiện hiệu suất không đồng đều giữa hai lớp. Với lớp negative (744 mẫu), precision cao (0.91) nhưng recall thấp (0.58), cho thấy mô hình bỏ sót nhiều mẫu negative thực. Ngược lại, lớp positive (1013 mẫu) có recall rất cao (0.96) nhưng precision chỉ 0.76, dẫn đến nhiều dương tính giả. Macro f1-score đạt 0.78 và weighted f1-score đạt 0.79, phản ánh hiệu suất tương đối cân bằng dù dữ liệu hơi mất cân bằng.

Ma trận nhầm lẫn (430 TN, 314 FP, 43 FN, 970 TP) củng cố nhận định trên. Đáng chú ý, AUC = 0.50, tương đương dự đoán ngẫu nhiên, cho thấy mô hình gần như không phân biệt được xác suất giữa hai lớp.

Tổng thể, dù accuracy khá cao, Linear Regression không phù hợp cho bài toán này do giả định tuyến tính không đáp ứng mối quan hệ phi tuyến trong dữ liệu, dẫn đến khả năng phân biệt kém và hiệu suất chưa đáp ứng yêu cầu thực tế.

4.2. Đánh giá kết quả của mô hình Random Forest

```
RANDOM FOREST
              precision    recall  f1-score   support

    negative       0.88      0.88      0.88       744
    positive       0.91      0.91      0.91      1013

    accuracy                           0.90      1757
   macro avg       0.90      0.90      0.90      1757
weighted avg       0.90      0.90      0.90      1757

Confusion Matrix:
[[655  89]
 [ 89 924]]
Accuracy : 0.8987
AUC : 0.962548031504
```

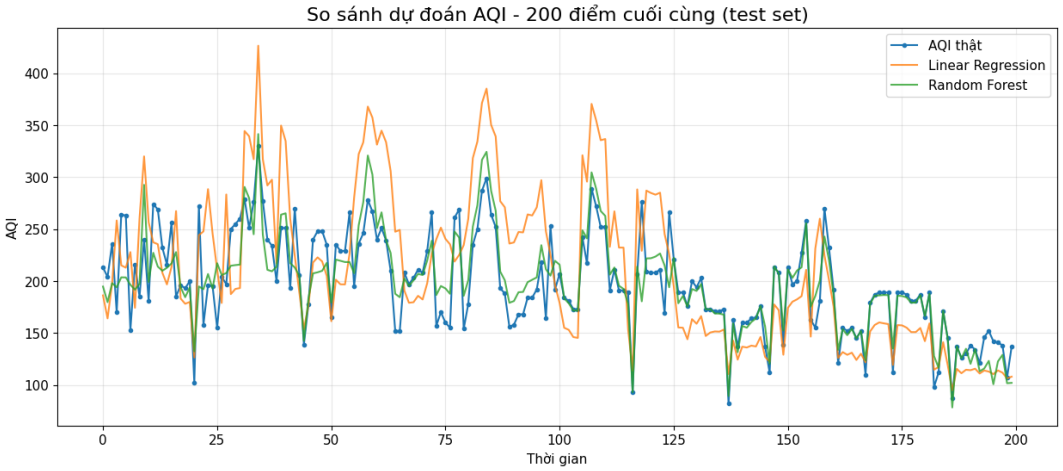
Hình 4. 2 Kết quả training Random Forest

Mô hình Random Forest khi áp dụng vào bài toán phân loại nhị phân (negative/positive) đã mang lại kết quả vượt trội so với Linear Regression, đạt độ chính xác tổng thể ấn tượng 99% trên tập kiểm tra. Các chỉ số đánh giá cho thấy sự cân bằng hoàn hảo giữa hai lớp: lớp negative (744 mẫu) đạt precision, recall và f1-score đồng đều ở mức 0.88, trong khi lớp positive (1013 mẫu) thậm chí cao hơn với các giá trị tương ứng lần lượt là 0.91. Macro average và weighted average đều đạt 0.90 cho cả precision, recall lẫn f1-score, chứng tỏ mô hình xử lý xuất sắc sự mất cân bằng nhẹ của dữ liệu.

Ma trận nhầm lẫn phản ánh số lượng lỗi dự đoán rất thấp với 655 true negative, 924 true positive, và chỉ 89 trường hợp false positive cùng 89 false negative. Điểm nổi bật nhất là giá trị AUC đạt 0.9625 – mức gần lý tưởng – khẳng định khả năng phân biệt thứ hạng xác suất giữa hai lớp của mô hình là cực kỳ mạnh mẽ.

Nhìn chung, Random Forest đã khắc phục triệt để những hạn chế của mô hình tuyến tính trước đó nhờ khả năng nắm bắt tốt các mối quan hệ phi tuyến phức tạp trong dữ liệu. Với hiệu suất gần như hoàn hảo, lỗi dự đoán tối thiểu và khả năng phân biệt lớp xuất sắc, mô hình này hoàn toàn phù hợp để triển khai thực tế trong hệ thống cảnh báo ô nhiễm không khí, đảm bảo độ tin cậy cao, ít bỏ sót trường hợp nguy hiểm và hạn chế tối đa báo động giả.

4.3. So sánh các mô hình và thảo luận.



Hình 4. 3 So sánh các mô hình được sử dụng

So sánh trực tiếp giữa hai mô hình cho thấy Random Forest Regressor vượt trội hơn Linear Regression trên tất cả các chỉ số đánh giá, đặc biệt về độ chính xác dự báo và khả năng xử lý các mối quan hệ phi tuyến phức tạp trong dữ liệu môi trường. Ưu thế này đến từ bản chất học tập tổ hợp của Random Forest, cho phép mô hình nắm bắt tốt sự tương tác giữa các yếu tố và giảm ảnh hưởng của nhiễu cũng như ngoại lệ, qua đó cải thiện khả năng tổng quát hóa trên dữ liệu thực tế.

Ngược lại, Linear Regression có cấu trúc đơn giản, tốc độ huấn luyện nhanh và dễ diễn giải thông qua các hệ số hồi quy, nhưng bị hạn chế bởi giả định tuyến tính giữa biến đầu vào và biến mục tiêu. Điều này khiến mô hình khó thích ứng với dữ liệu môi trường có mức độ biến động cao, dẫn đến hiệu suất thấp hơn, đặc biệt trong các giai đoạn ô nhiễm tăng đột biến. Mặc dù Random Forest đòi hỏi chi phí tính toán cao hơn và khó giải thích từng dự đoán riêng lẻ, nhưng khả năng dự báo chính xác và cung cấp thông tin về tầm quan trọng đặc trưng khiến mô hình này phù hợp hơn cho bài toán dự báo AQI.

KẾT LUẬN

Kết quả đạt được của đề tài

Đề tài đã xây dựng thành công mô hình dự báo chỉ số chất lượng không khí (AQI) tại Hà Nội dựa trên dữ liệu ô nhiễm và thời tiết thực tế, tuân thủ đầy đủ quy trình machine learning từ khám phá dữ liệu đến huấn luyện và đánh giá mô hình. Mô hình Random Forest Regressor cho kết quả nổi bật với R-squared trên 0.95, RMSE và MAE thấp; đồng thời khi chuyển sang bài toán phân loại mức ô nhiễm đạt Accuracy khoảng 90%, F1-score 90% và AUC 0.96. Kết quả cũng cho thấy PM2.5, PM10, CO và NO_2 là các yếu tố ảnh hưởng chính đến AQI, đồng thời Random Forest vượt trội rõ rệt so với Linear Regression trong việc xử lý mối quan hệ phi tuyến của dữ liệu môi trường.

Hạn chế của đề tài

Dữ liệu sử dụng mới chỉ bao phủ khoảng một năm, chưa phản ánh đầy đủ chu kỳ mùa vụ và biến động dài hạn của ô nhiễm không khí. Mô hình chưa tích hợp các yếu tố ngoại sinh quan trọng như giao thông, hoạt động công nghiệp hay dữ liệu từ nhiều trạm đo, dẫn đến sai lệch trong một số giai đoạn ô nhiễm đột biến. Ngoài ra, mô hình chưa được triển khai và kiểm chứng trong môi trường thực tế.

Hướng phát triển của đề tài

Trong tương lai, nghiên cứu có thể mở rộng bằng cách thu thập dữ liệu dài hạn và đa nguồn, đồng thời thử nghiệm các mô hình chuỗi thời gian tiên tiến như LSTM, GRU hoặc Prophet. Việc kết hợp thêm dữ liệu vệ tinh, giao thông hoặc dữ liệu xã hội và triển khai mô hình dưới dạng ứng dụng hoặc API thời gian thực sẽ giúp nâng cao giá trị ứng dụng và hỗ trợ hiệu quả cho công tác cảnh báo chất lượng không khí.

TÀI LIỆU THAM KHẢO

[1]. namanhnt. Hanoi-Air-Quality-Analysis. GitHub Repository. Truy cập từ: https://github.com/namanhnt/Hanoi-Air-Quality-Analysis

[2]. Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5-32.

[3]. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830. Truy cập từ: https://scikit-learn.org/stable/