

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 GIÁNG 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

NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN	1
MỤC LỤC	3
DANH MỤC CÁC KÝ TỰ, CÁC TỪ VIẾT TẮT	4
DANH MỤC CÁC HÌNH VẼ	5
CHƯƠNG 1: CƠ SỞ LÝ THUYẾT	6
1.1. Tổng quan về học máy	6
1.2. Phát biểu bài toán	6
1.3. Mô hình học máy được lựa chọn	7
1.4. Nguyên lý hoạt động và giả thuyết	12
CHƯƠNG 2: DỮ LIỆU VÀ TIỀN XỬ LÝ	13
2.1 Giới thiệu bộ dữ liệu	13
2.2 Phân tích dữ liệu ban đầu	13
2.3 Tiền xử lý dữ liệu	14
CHƯƠNG 3: XÂY DỰNG VÀ HUẤN LUYỆN MÔ HÌNH	15
3.1. Quá trình xây dựng mô hình	15
3.2. Huấn luyện mô hình	15
3.3. Điều chỉnh tham số	15
2.4. Độ phức tạp và chi phí tính toán	16
CHƯƠNG 4: ĐÁNH GIÁ VÀ KẾT LUẬN	17
4.1. Đánh giá kết quả của mô hình Linear Regression	17
4.2. Đánh giá kết quả của mô hình Random Forest	17
4.3. So sánh các mô hình và thảo luận	18
KẾT LUẬN	20
TÀI LIỆU THAM KHẢO	21

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 $2.5\mu\text{m}$
3	PM10	Particulate Matter 10	Hạt bụi có đường kính $10\mu\text{m}$
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

Hình 1. 1 Mô hình hồi quy tuyến tính	8
Hình 1. 2 Bốn tập dữ liệu khác nhau với cùng xu hướng hồi quy.....	9
Hình 1. 3 Các cấu trúc cây được lấy mẫu và đánh nhãn	10
Hình 1. 4 Minh họa quá trình sampling và học cây quyết định	11
Hình 2. 1 Nguồn dữ liệu được sử dụng	13
Hình 2. 2 Xu hướng ô nhiễm theo thời gian.....	13
Hình 2. 3 Các yếu tố gây ô nhiễm	14
Hình 4. 1 Kết quả training Linear Regression.....	17
Hình 4. 2 Kết quả training Random Forest.....	18
Hình 4. 3 So sánh các mô hình được sử dụng	19

CHƯƠNG 1: CƠ SỞ LÝ THUYẾT

1.1. Tổng quan về học máy

Học máy (Machine Learning – ML) là một lĩnh vực nghiên cứu trong trí tuệ nhân tạo, tập trung vào việc phát triển và nghiên cứu các thuật toán thống kê có khả năng học từ dữ liệu và khai quật hóa sang dữ liệu chưa từng thấy, từ đó thực hiện các nhiệm vụ mà không cần được lập trình bằng các chỉ dẫn tường minh. Trong một phân ngành của học máy, những tiến bộ trong lĩnh vực học sâu (deep learning) đã cho phép mạng nơ-ron – một lớp các thuật toán thống kê – vượt trội hơn nhiều phương pháp học máy trước đây về hiệu năng.

ML được ứng dụng trong nhiều lĩnh vực, bao gồm xử lý ngôn ngữ tự nhiên, thị giác máy tính, nhận dạng giọng nói, lọc email, nông nghiệp và y học. Việc áp dụng ML vào các bài toán kinh doanh được gọi là phân tích dự đoán (predictive analytics).

Thống kê và các phương pháp tối ưu hóa toán học (lập trình toán học) tạo thành nền tảng của học máy. Khai phá dữ liệu (data mining) là một lĩnh vực nghiên cứu liên quan, tập trung vào phân tích dữ liệu thăm dò (EDA) thông qua học không giám sát.

Từ góc độ lý thuyết, mô hình học “gần đúng có xác suất đúng” (Probably Approximately Correct – PAC learning) cung cấp một khuôn khổ toán học và thống kê để mô tả học máy. Phần lớn các thuật toán học máy truyền thống và học sâu đều có thể được mô tả như là quá trình tối thiểu hóa rủi ro thực nghiệm (empirical risk minimisation) trong khuôn khổ này.

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-

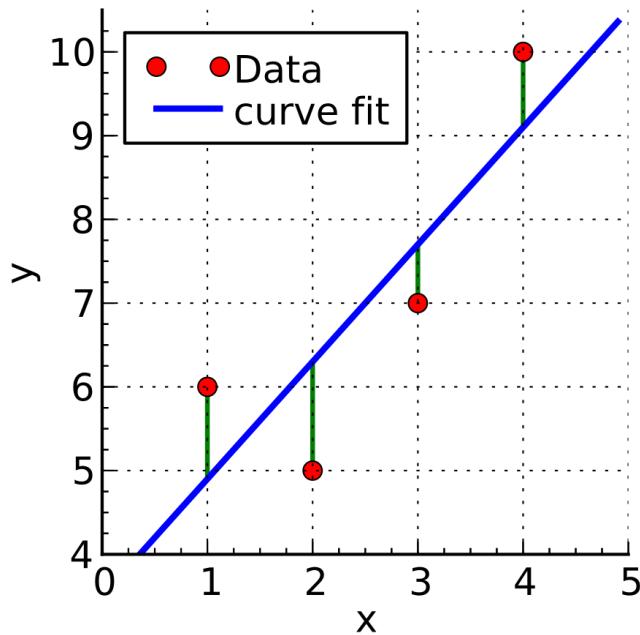
150 không tốt cho nhóm nhạy cảm, 151-200 có hại, 201-300 rất có hại và trên 300 là nguy hiểm.

1.3. Mô hình học máy được lựa chọn

a) Linear Regression

Trong thống kê, hồi quy tuyến tính là một mô hình điều đó ước tính mối quan hệ giữa a và hướng phản hồi (biến phụ thuộc) và một hoặc nhiều biến giải thích (hồi quy hoặc biến independent). Một mô hình có chính xác một biến giải thích là a hồi quy tuyến tính đơn giản; một mô hình có hai hoặc nhiều biến giải thích là a hồi quy tuyến tính bội.^[1] Thuật ngữ này khác biệt với hồi quy tuyến tính đa biến, dự đoán nhiều tương quan biến phụ thuộc chứ không phải là một biến phụ thuộc duy nhất.

Trong hồi quy tuyến tính, các mối quan hệ được mô hình hóa bằng cách sử dụng các hàm dự đoán tuyến tính người mẫu không rõ thông số đang là ước tính từ các dữ liệu. Thông thường nhất, các có điều kiện (conditional mean) trong số các phản hồi đưa ra giá trị của các biến giải thích (hoặc các yếu tố dự đoán) được giả định là một hàm affine trong số những giá trị đó; ít phổ biến hơn, có điều kiện trung bình hoặc một số khác lượng tử được sử dụng. Giống như mọi hình thức của phân tích hồi quy, hồi quy tuyến tính tập trung vào phân bố xác suất có điều kiện của phản hồi đưa ra các giá trị của các yếu tố dự đoán, thay vì trên phân bố xác suất chung trong số tất cả các biến này, đó là miền của phân tích đa biến.



Hình 1. 1 Mô hình hồi quy tuyến tính

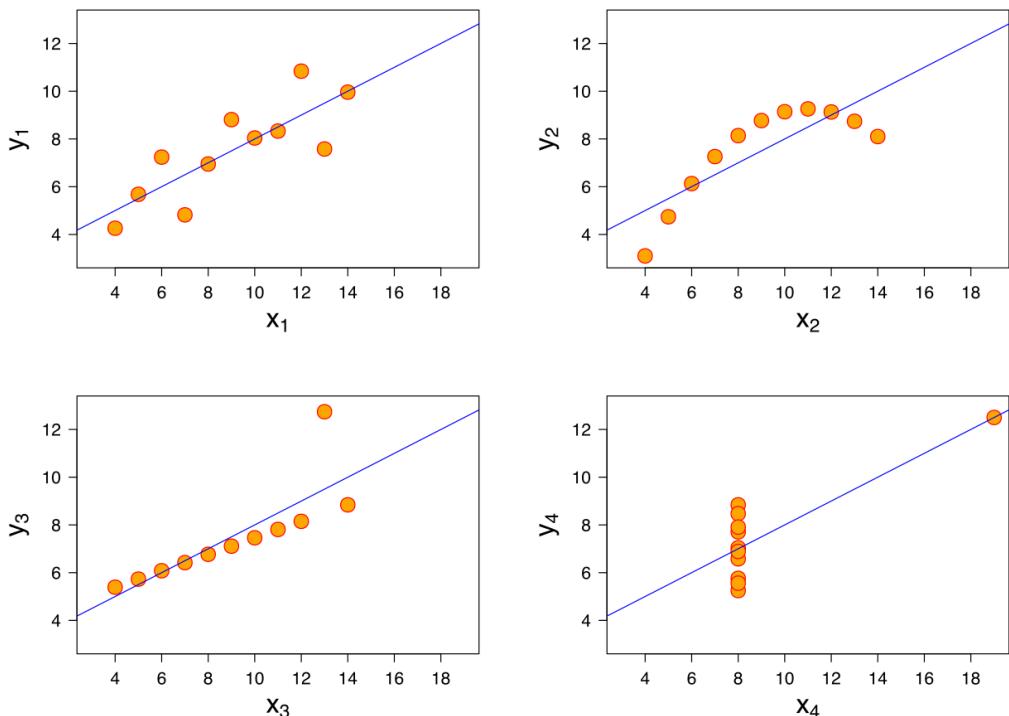
Hồi quy tuyến tính cũng là một loại học máy thuật toán, cụ thể hơn là a giám sát thuật toán học từ các bộ dữ liệu được gắn nhãn và ánh xạ các điểm dữ liệu đến các hàm tuyến tính được tối ưu hóa nhất có thể được sử dụng để dự đoán trên các bộ dữ liệu mới.^[3]

Hồi quy tuyến tính là loại phân tích hồi quy đầu tiên được nghiên cứu nghiêm ngặt và được sử dụng rộng rãi trong các ứng dụng thực tế.^[4] Điều này là do các mô hình phụ thuộc tuyến tính vào các tham số chưa biết của chúng dễ khớp hơn các mô hình có liên quan phi tuyến tính với các tham số của chúng và do các thuộc tính thống kê của các công cụ ước tính thu được dễ xác định hơn.

Hồi quy tuyến tính có nhiều ứng dụng thực tế. Hầu hết các ứng dụng thuộc một trong hai loại lớn sau:

- Nếu mục tiêu là để giảm lỗi, tức là. phương sai vào dự đoán hoặc dự báo, hồi quy tuyến tính có thể được sử dụng để khớp mô hình dự đoán với mô hình quan sát được tập dữ liệu giá trị của các biến phản hồi và giải thích. Sau khi phát triển mô hình như vậy, nếu các giá trị bổ sung của các biến giải thích được thu thập mà không có giá trị phản hồi đi kèm thì mô hình phù hợp có thể được sử dụng để đưa ra dự đoán về phản hồi.

- Nếu mục tiêu là giải thích sự biến thiên của biến phản hồi có thể được quy cho sự biến thiên của các biến giải thích, thì phân tích hồi quy tuyến tính có thể được áp dụng để định lượng mức độ của mối quan hệ giữa phản hồi và các biến giải thích, và đặc biệt là để xác định liệu một số biến giải thích có thể không có mối quan hệ tuyến tính nào với phản hồi cả hoặc để xác định tập hợp con nào của các biến giải thích có thể chứa thông tin dư thừa về phản hồi.



Hình 1. 2 Bốn tập dữ liệu khác nhau với cùng xu hướng hồi quy

Các mô hình hồi quy tuyến tính thường được trang bị bằng cách sử dụng bình phương tối thiểu tiếp cận, nhưng chúng cũng có thể được trang bị theo những cách khác, chẳng hạn như bằng cách giảm thiểu các "thiểu phù hợp" ở một số nơi khác chuẩn mực (như với độ lệch tuyệt đối ít nhất hồi quy), hoặc bằng cách giảm thiểu một phiên bản bị phạt của bình phương tối thiểu hàm cost như trong hồi quy sườn núi (L^2 -hình phạt thông thường) và lasso (L^1 -hình phạt bình thường). Sử dụng các Lỗi bình phương trung bình (MSE) vì chi phí trên một tập dữ liệu có nhiều giá trị ngoại lệ lớn có thể dẫn đến một mô hình phù hợp với các giá trị ngoại lệ hơn dữ liệu thực do tầm quan trọng cao hơn mà MSE gán cho các lỗi lớn. Vì vậy, nên sử dụng các hàm chi phí mạnh đối với các giá trị ngoại lệ nếu tập dữ liệu có nhiều giá trị lớn ngoại lệ. Ngược lại, phương pháp bình phương tối thiểu có thể được sử dụng để khống các mô hình không phải là mô hình tuyến

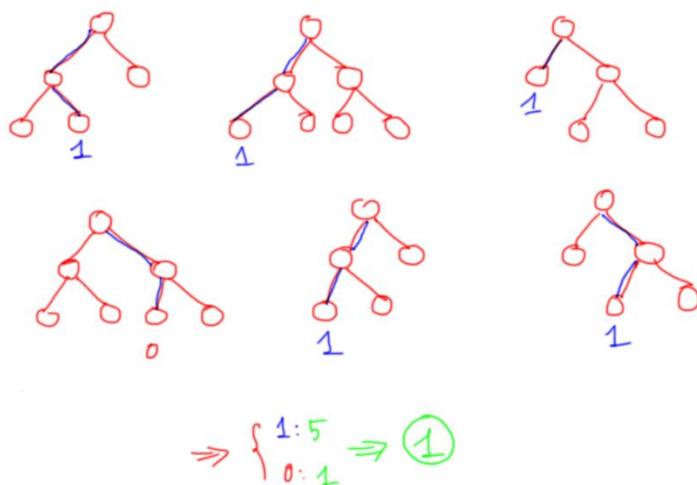
tính. Do đó, mặc dù các thuật ngữ "bình phương nhỏ nhất" và "mô hình tuyến tính" có mối liên hệ chặt chẽ với nhau nhưng chúng không đồng nghĩa với nhau.

b) Random Forest

Random là ngẫu nhiên, Forest là rừng, nên ở thuật toán Random Forest mình sẽ xây dựng nhiều cây quyết định bằng thuật toán Decision Tree, tuy nhiên mỗi cây quyết định sẽ khác nhau (có yếu tố random). Sau đó kết quả dự đoán được tổng hợp từ các cây quyết định.

Ở bước huấn luyện thì mình sẽ xây dựng nhiều cây quyết định, các cây quyết định có thể khác nhau (phần sau mình sẽ nói mỗi cây được xây dựng như thế nào).

Sau đó ở bước dự đoán, với một dữ liệu mới, thì ở mỗi cây quyết định mình sẽ đi từ trên xuống theo các node điều kiện để được các dự đoán, sau đó kết quả cuối cùng được tổng hợp từ kết quả của các cây quyết định.



Hình 1.3 Các cấu trúc cây được lấy mẫu và đánh nhãn

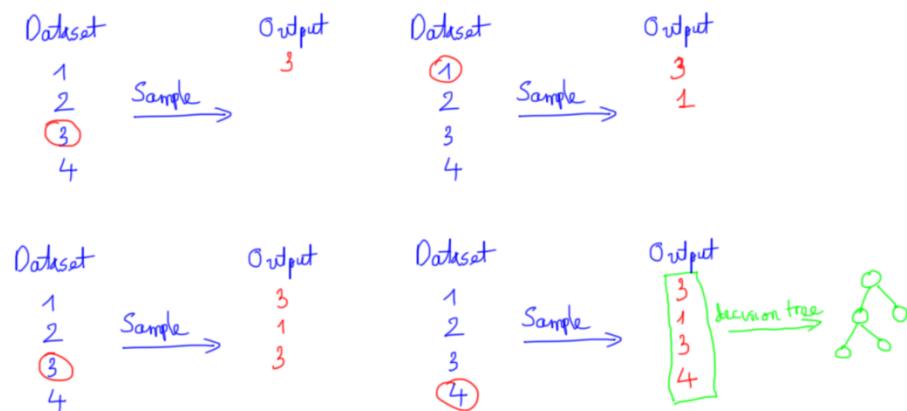
Ví dụ như trên, thuật toán Random Forest có 6 cây quyết định, 5 cây dự đoán 1 và 1 cây dự đoán 0, do đó mình sẽ vote là cho ra dự đoán cuối cùng là 1.

Xây dựng thuật toán Random Forest

Giả sử bộ dữ liệu của mình có n dữ liệu (sample) và mỗi dữ liệu có d thuộc tính (feature).

Để xây dựng mỗi cây quyết định mình sẽ làm như sau:

Lấy ngẫu nhiên n dữ liệu từ bộ dữ liệu với kỹ thuật Bootstrapping, hay còn gọi là random sampling with replacement. Tức khi mình sample được 1 dữ liệu thì mình không bỏ dữ liệu đó ra mà vẫn giữ lại trong tập dữ liệu ban đầu, rồi tiếp tục sample cho tới khi sample đủ n dữ liệu. Khi dùng kỹ thuật này thì tập n dữ liệu mới của mình có thể có những dữ liệu bị trùng nhau.



Hình 1.4 Minh họa quá trình sampling và học cây quyết định

Sau khi sample được n dữ liệu từ bước 1 thì mình chọn ngẫu nhiên ở k thuộc tính ($k < n$). Giờ mình được bộ dữ liệu mới gồm n dữ liệu và mỗi dữ liệu có k thuộc tính.

Dùng thuật toán Decision Tree để xây dựng cây quyết định với bộ dữ liệu ở bước trước đó.

Do quá trình xây dựng mỗi cây quyết định đều có yếu tố ngẫu nhiên (random) nên kết quả là các cây quyết định trong thuật toán Random Forest có thể khác nhau.

Thuật toán Random Forest sẽ bao gồm nhiều cây quyết định, mỗi cây được xây dựng dùng thuật toán Decision Tree trên tập dữ liệu khác nhau và dùng tập thuộc tính khác nhau. Sau đó kết quả dự đoán của thuật toán Random Forest sẽ được tổng hợp từ các cây quyết định.

Khi dùng thuật toán Random Forest, mình hay để ý các thuộc tính như: số lượng cây quyết định sẽ xây dựng, số lượng thuộc tính dùng để xây dựng cây. Ngoài ra, vẫn có các thuộc tính của thuật toán Decision Tree để xây dựng cây như độ sâu tối đa, số phần tử tối thiểu trong 1 node để có thể tách.

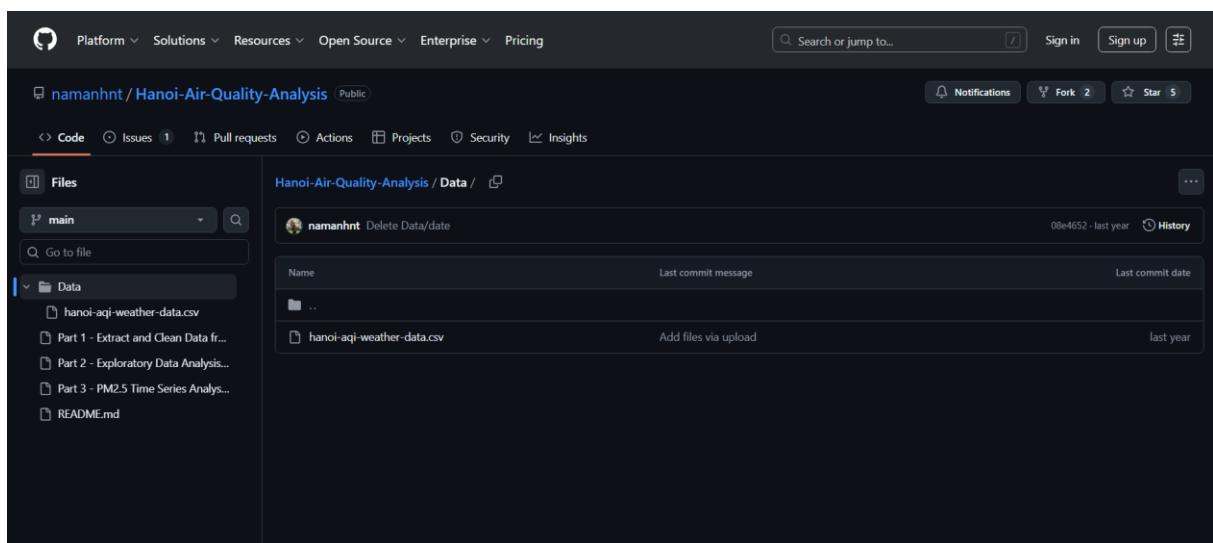
1.4. Nguyên lý hoạt động và giả thuyết

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

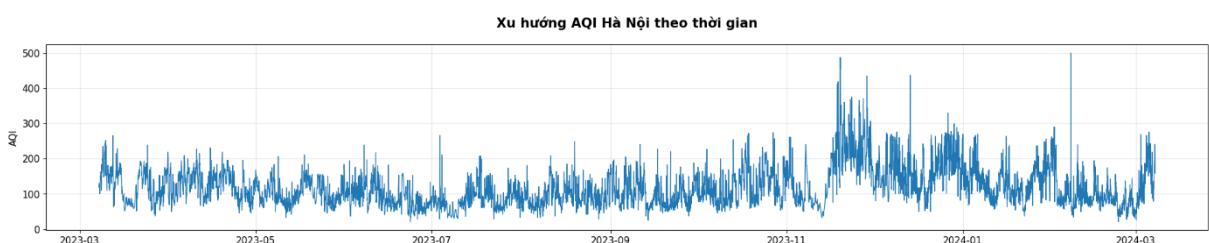
Bộ dữ liệu được sử dụng lấy từ nguồn mở trên GitHub của tác giả namanhnt, chứa thông tin chất lượng không khí và thời tiết tại Hà Nội theo giờ từ năm 2023 đến đầu 2024. Bộ dữ liệu gồm 8785 mẫu với 18 cột, bao gồm thời gian UTC, thành phố, mã quốc gia, múi giờ, chỉ số AQI cùng nồng độ các chất ô nhiễm 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

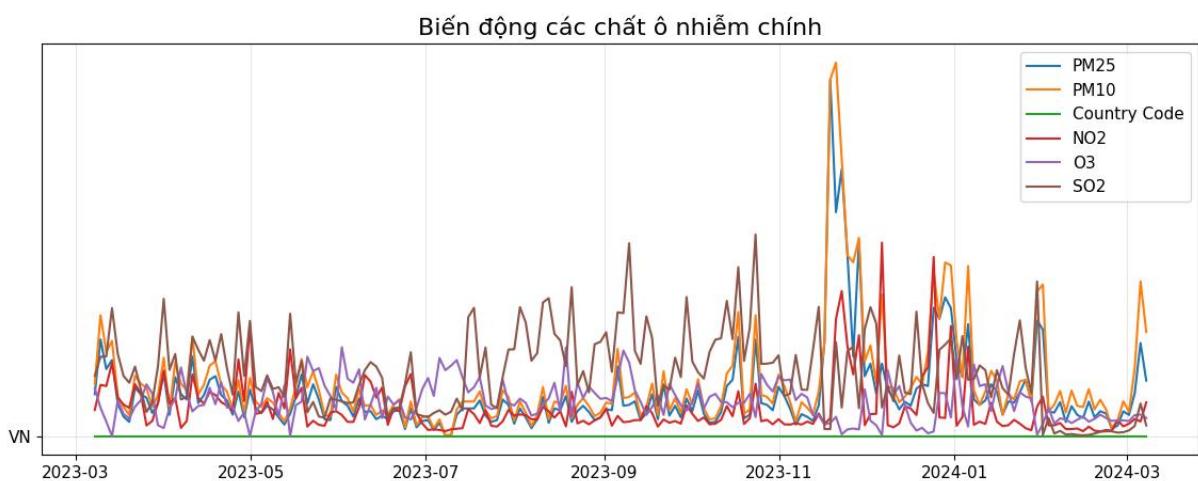
2.2 Phân tích dữ liệu ban đầu.

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. Mỗi 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. Quá 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 luyện 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 nhiều 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) cũng có 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					

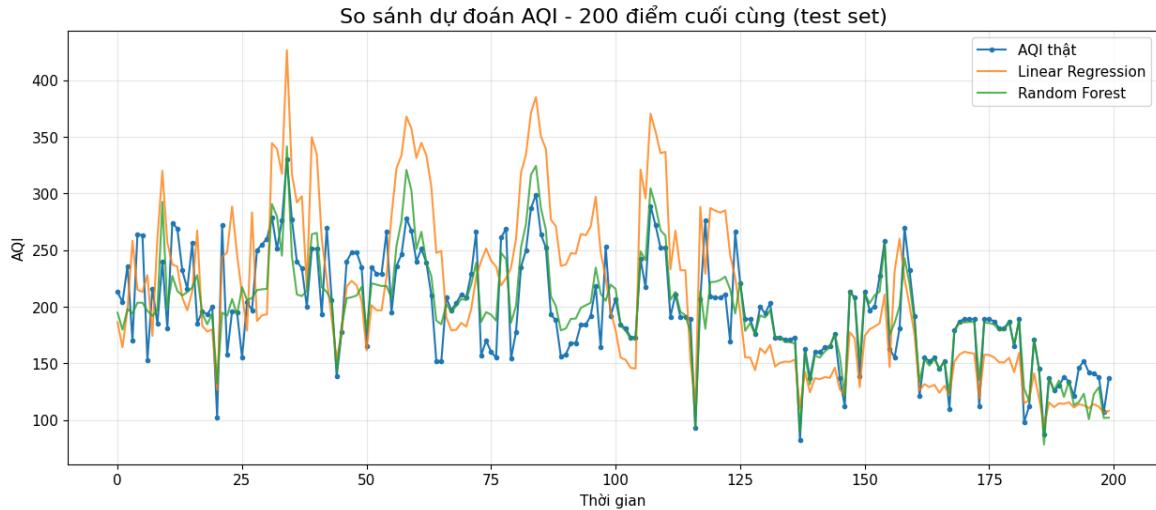
Hình 4. 2 Kết quả training Random Forest

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 án 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 nhiều cung 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 NO₂ 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/>