// mn nhớ ghi chú thích dưới từng hình ảnh, biểu đồ nha

Abstract

Dự báo và phân loại điều kiện thời tiết là một bài toán quan trọng trong khoa học dữ liệu do tính phức tạp, phi tuyến và chứa nhiều yếu tố bất định của dữ liệu khí tượng. Trong báo cáo này, chúng tôi xây dựng và so sánh hiệu năng của ba mô hình học máy có giám sát, bao gồm Random Forest, Softmax Regression và AdaBoost, nhằm phân loại điều kiện thời tiết dựa trên dữ liệu chuỗi thời gian thu thập từ các nguồn quan trắc và dự báo.

Quy trình nghiên cứu bao gồm các bước tiền xử lý dữ liệu, trích xuất và lựa chọn đặc trưng, chia tập dữ liệu theo cấu trúc thời gian nhằm tránh rò rỉ thông tin, huấn luyện mô hình và đánh giá hiệu năng. Các mô hình được so sánh thông qua các thước đo Precision, Recall và F1-score trên tập kiểm định và tập kiểm tra. Kết quả thực nghiệm cho thấy Random Forest đạt hiệu năng cao nhất và ổn định nhất so với hai mô hình còn lại.

Hiệu quả vượt trội của Random Forest cho thấy mô hình này đặc biệt phù hợp với dữ liệu thời tiết có tính phi tuyến, nhiều đặc trưng và chứa nhiễu, nhờ khả năng học các mối quan hệ phức tạp và giảm hiện tượng quá khớp thông qua cơ chế tổ hợp. Ngược lại, Softmax Regression cho thấy hạn chế trong việc mô hình hóa các quan hệ phi tuyến, trong khi AdaBoost nhạy cảm hơn với nhiễu trong dữ liệu. Kết quả nghiên cứu khẳng định vai trò của các mô hình tổ hợp, đặc biệt là Random Forest, trong bài toán phân loại điều kiện thời tiết với dữ liệu chuỗi thời gian.

Từ khóa: Random Forest, Softmax Regression, AdaBoost, time series, mối tương quan, tuyến tính, mất cân bằng

1. Giới thiệu

Phân loại điều kiện thời tiết đóng vai trò quan trọng trong nhiều lĩnh vực thực tiễn như quản lý đô thị, giao thông, du lịch, nông nghiệp và các hệ thống cảnh báo sớm thiên tai. Việc xác định chính xác trạng thái thời tiết không chỉ hỗ trợ ra quyết định mà còn góp phần giảm thiểu rủi ro và thiệt hại do các hiện tượng thời tiết bất thường gây ra. Với sự phát triển của khoa học dữ liệu và học máy, các phương pháp dựa trên dữ liệu đã trở thành hướng tiếp cận hiệu quả, bổ trợ cho các mô hình dự báo truyền thống dựa trên vật lý.

Tuy nhiên, bài toán phân loại thời tiết đặt ra nhiều thách thức đáng kể. Dữ liệu khí tượng mang tính phi tuyến, chịu ảnh hưởng đồng thời của nhiều yếu tố như nhiệt độ, độ ẩm, áp suất, tốc độ và hướng gió. Bên cạnh đó, dữ liệu thường tồn tại dưới dạng chuỗi thời gian (time series), có sự phụ thuộc theo thời gian và biến động mạnh theo mùa, gây khó khăn cho việc xây dựng các mô hình học máy ổn định. Ngoài ra, sự nhiễu và bất định trong các chỉ số khí tượng cũng ảnh hưởng trực tiếp đến độ chính xác của mô hình.

Xuất phát từ những thách thức trên, mục tiêu của nghiên cứu này là đánh giá khả năng phân loại điều kiện thời tiết của các thuật toán học máy có giám sát, từ đó xác định mô hình phù hợp nhất với đặc điểm dữ liệu khí tượng thực tế. Ba mô hình được lựa chọn để nghiên cứu và so sánh bao gồm Random Forest, AdaBoost và Softmax Regression. Đây là các mô hình đại diện cho những hướng tiếp cận khác nhau: mô hình tuyến tính đa lớp, mô hình tổ hợp dựa trên tăng cường, và mô hình tổ hợp dựa trên rừng cây quyết định.

2. Thu thập và tiền xử lý dữ liệu

a. Nguồn dữ liệu & Crawl dữ liệu

Dữ liệu thời tiết sử dụng trong nghiên cứu được thu thập từ Open-Meteo Archive API, một nền tảng cung cấp dữ liệu khí tượng lịch sử miễn phí với độ phân giải theo giờ. Open-Meteo tổng hợp dữ liệu từ các mô hình khí tượng và nguồn quan trắc đáng tin cậy, cho phép truy xuất các biến thời tiết đa dạng thông qua giao diện lập trình ứng dụng (API).

Trong nghiên cứu này, dữ liệu được thu thập cho khu vực thành phố Đà Nẵng. Phạm vi thời gian được lựa chọn từ ngày 01/01/2024 đến ngày 01/01/2026, đảm bảo số lượng mẫu đủ lớn để phục vụ cho quá trình huấn luyện và đánh giá các mô hình học máy.

Quá trình thu thập dữ liệu được thực hiện tự động thông qua việc gửi yêu cầu HTTP đến endpoint `archive-api.open-meteo.com/v1/archive` với các tham số truy vấn xác định trước. Các đặc trưng khí tượng được thu thập bao gồm: thời gian của dữ liệu (time), nhiệt độ không khí tại độ cao 2m (temperature_2m), độ ẩm tương đối (relative_humidity_2m), điểm sương (dew_point_2m), áp suất bề mặt (surface_pressure), lượng mưa (rain), mưa rào (showers), độ sâu của tuyết (snow_depth), mã điều kiện thời tiết (weather code), độ che phủ mây (cloud_cover), và tốc độ gió tại độ cao 10m (wind_speed_10m). Việc lựa chọn các biến này nhằm phản ánh đầy đủ các yếu tố vật lý ảnh hưởng trực tiếp đến trạng thái thời tiết.

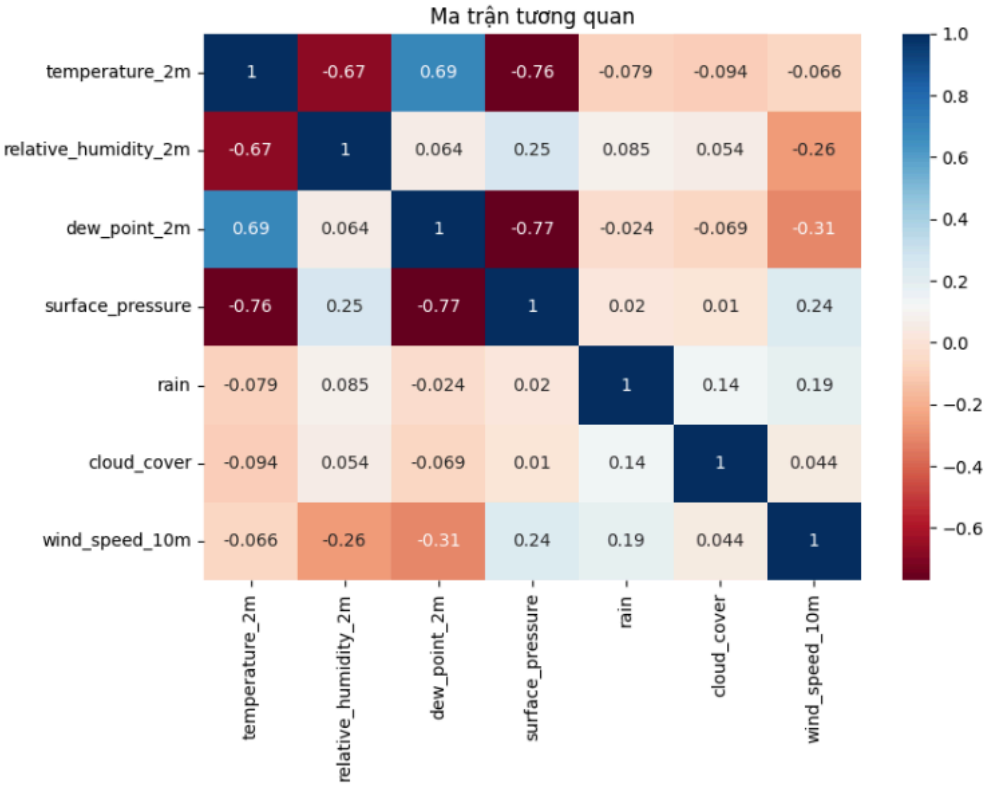
Sau khi nhận phản hồi từ API, dữ liệu JSON được chuyển đổi sang dạng bảng (DataFrame) để thuận tiện cho việc xử lý và phân tích. Tổng cộng 17.568 bản ghi theo giờ đã được thu thập thành công. Dữ liệu sau đó được lưu trữ dưới dạng tệp CSV nhằm phục vụ cho các bước tiền xử lý, phân tích khám phá dữ liệu và huấn luyện mô hình trong các giai đoạn tiếp theo của nghiên cứu.

Việc sử dụng API để crawl dữ liệu giúp đảm bảo tính tái lập của nghiên cứu, đồng thời cho phép mở rộng phạm vi không gian và thời gian trong các nghiên cứu tiếp theo một cách linh hoạt.

b. Khám phá và trực quan hóa dữ liệu

- Thống kê mô tả các đặc trưng (mean, std, min,...)
- Visualize tương quan (Sử dụng Heatmap để xem mối liên hệ giữa các yếu tố)
- Phân bổ nhãn (Kiểm tra nhãn có bị mất cân bằng ko)

Trước khi tiến hành phân tích thống kê chi tiết, nghiên cứu thực hiện **trực quan hóa mối quan hệ giữa các đặc trưng khí tượng** nhằm có cái nhìn tổng quan về cấu trúc dữ liệu. Ma trận tương quan Pearson được tính toán và biểu diễn dưới dạng **heatmap**, cho phép đánh giá nhanh mức độ phụ thuộc tuyến tính giữa các biến đầu vào. Từ đó hỗ trợ quá trình lựa chọn đặc trưng và diễn giải kết quả mô hình.

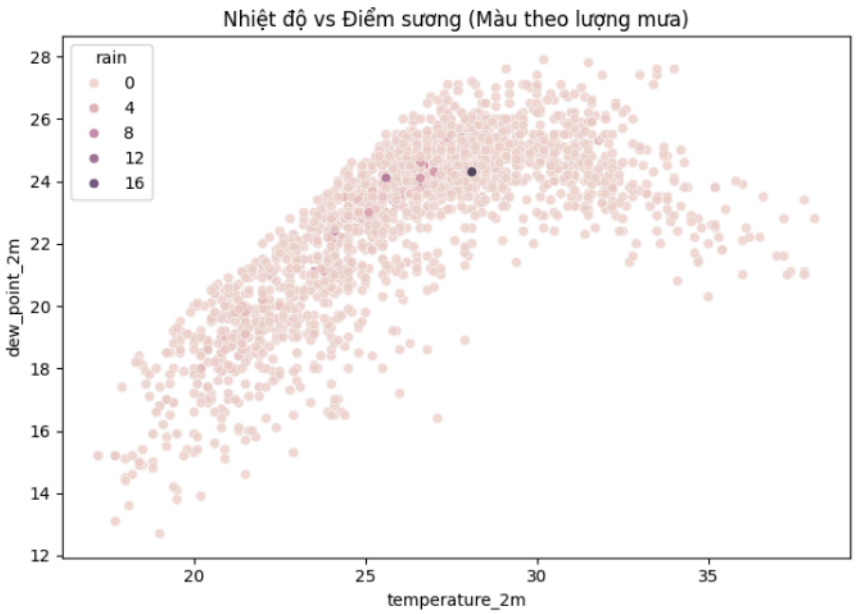


Hình : **Ma trận tương quan Pearson** giữa các đặc trưng khí tượng

Kết quả cho thấy tồn tại một số mối tương quan mạnh và có ý nghĩa về mặt vật lý. Nhiệt độ tại độ cao 2m có tương quan âm mạnh với áp suất bề mặt (−0.76) và độ ẩm tương đối (−0.67), đồng thời tương quan dương mạnh với điểm sương (0.69). Điều này phản ánh đúng bản chất khí tượng học, khi nhiệt độ tăng thường đi kèm với sự giảm áp suất và thay đổi độ ẩm không khí, trong khi điểm sương có mối liên hệ chặt chẽ với nhiệt độ.

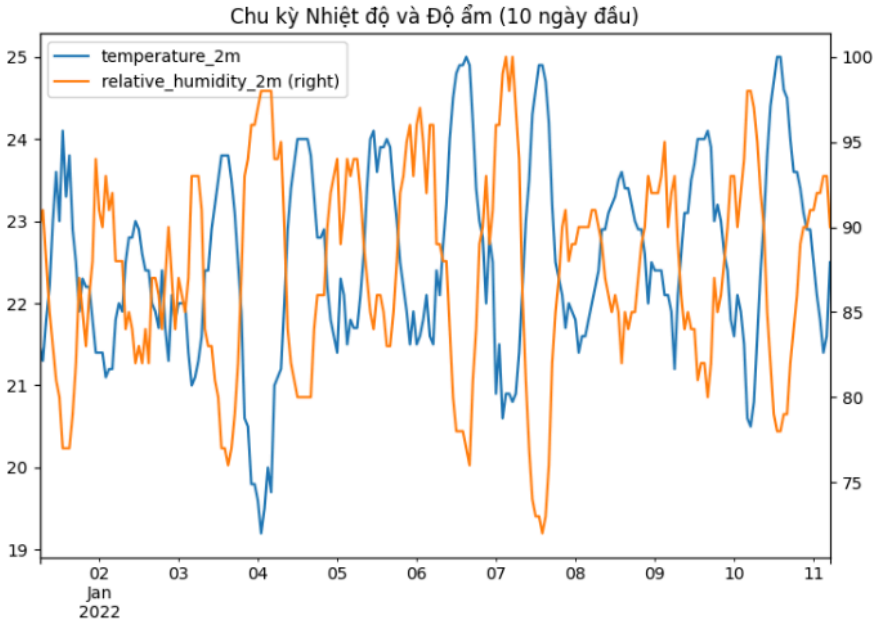
Tương tự, điểm sương và áp suất bề mặt có tương quan âm rất mạnh (−0.77), cho thấy khi không khí ẩm hơn (điểm sương cao), áp suất bề mặt có xu hướng giảm. Mối quan hệ này đóng vai trò quan trọng trong việc hình thành các trạng thái thời tiết như mưa hoặc nhiều mây.

Ngược lại, các đặc trưng liên quan đến **lượng mưa, độ che phủ mây và tốc độ gió** có mức tương quan tuyến tính thấp với các biến còn lại. Ví dụ, lượng mưa có tương quan rất nhỏ với nhiệt độ (−0.079), điểm sương (−0.024) và áp suất (0.02), trong khi độ che phủ mây gần như không tương quan tuyến tính đáng kể với các đặc trưng khác. Điều này cho thấy các hiện tượng này có thể phụ thuộc vào các mối quan hệ phi tuyến hoặc tương tác phức tạp giữa nhiều yếu tố khí tượng.



Hình: Minh họa mối quan hệ giữa **nhiệt độ tại độ cao 2 m và điểm sương**

Biểu đồ phân tán cho thấy tồn tại mối quan hệ gần tuyến tính giữa nhiệt độ và điểm sương, phù hợp với bản chất vật lý của các đại lượng khí tượng này. Khi nhiệt độ tăng, điểm sương cũng có xu hướng tăng theo, phản ánh sự gia tăng hàm lượng hơi nước trong không khí. Đáng chú ý, các giá trị mưa lớn chủ yếu xuất hiện trong vùng có điểm sương cao, cho thấy trạng thái ẩm của không khí đóng vai trò quan trọng trong việc hình thành mưa. Kết quả này gợi ý rằng điểm sương là một đặc trưng quan trọng giúp mô hình phân biệt các trạng thái thời tiết có mưa và không mưa.



Hình : Chu kỳ biến thiên của nhiệt độ và độ ẩm tương đối trong 10 ngày đầu của tập dữ liệu

Có thể quan sát rõ tính chu kỳ theo ngày của nhiệt độ, trong khi độ ẩm tương đối có xu hướng biến thiên ngược pha với nhiệt độ. Khi nhiệt độ đạt giá trị cao, độ ẩm tương đối

thường giảm và ngược lại. Mối quan hệ này phản ánh quy luật khí tượng học cơ bản và cho thấy dữ liệu chuỗi thời gian chứa thông tin động học quan trọng, có thể được khai thác thông qua các đặc trưng trễ (lag features) trong quá trình xây dựng mô hình.

Sau khi phân tích mối quan hệ giữa các đặc trưng, nghiên cứu tiến hành xem xét **phân bố nhãn điều kiện thời tiết** nhằm đánh giá mức độ cân bằng của tập dữ liệu. Việc phân tích phân bổ nhãn giúp nhận diện sự chênh lệch giữa các lớp, đặc biệt trong bối cảnh dữ liệu thời tiết thường có xu hướng thiên lệch về các trạng thái phổ biến.

Bảng: Phân bố nhãn điều kiện thời tiết (Tổng cộng 17.568 bản ghi theo giờ)

Mã thời tiết	0	1	2	3	51	53	55	61	63	65
Tần suất	1935	1523	1185	6037	3149	740	172	302	225	44

Có thể quan sát thấy sự mất cân bằng rõ rệt giữa các nhãn, trong đó một số mã xuất hiện với tần suất rất cao, trong khi các mã khác chỉ chiếm tỷ lệ nhỏ. Cụ thể, các mã thời tiết phổ biến như mã 3 (6037 mẫu), mã 51 (3149 mẫu), mã 0 (1935 mẫu) chiếm phần lớn tổng số quan sát. Ngược lại, các hiện tượng thời tiết cường độ cao như mã 65 (44 mẫu), mã 55 (172 mẫu) và mã 63 (225 mẫu) xuất hiện với tần suất rất thấp. Sự chênh lệch này phản ánh đúng đặc điểm khí hậu thực tế của khu vực Đà Nẵng, nơi các trạng thái thời tiết cực đoan xảy ra không thường xuyên.

Tình trạng mất cân bằng nhãn như trên đặt ra thách thức lớn cho các mô hình học máy, đặc biệt là các mô hình tuyến tính như Softmax Regression, vốn có xu hướng thiên lệch về các lớp chiếm ưu thế và suy giảm khả năng nhận diện các lớp hiếm. Nếu giữ nguyên các mã thời tiết gốc, mô hình có nguy cơ đạt độ chính xác tổng thể cao nhưng hiệu năng kém đối với các trạng thái thời tiết quan trọng như mưa lớn hoặc mưa rất to.

Do đó, trong nghiên cứu này, các mã thời tiết gốc được gom nhóm lại thành các lớp thời tiết tổng quát hơn, nhằm vừa giảm mức độ mất cân bằng dữ liệu, vừa đảm bảo ý nghĩa khí tượng học của từng lớp. Việc gom nhóm này giúp cải thiện tính ổn định của quá trình huấn luyện, tăng khả năng khái quát hóa của mô hình và tạo điều kiện cho việc so sánh công bằng giữa các thuật toán phân loại khác nhau.

// thêm biểu đồ đặc điểm (temperature, humidity, wind speed ….) trước và sau khi weather group

c. Tiền xử lý dữ liệu chuyên sâu

- Gom nhóm weather code:

Dữ liệu gốc từ Open-Meteo sử dụng bảng mã WMO (World Meteorological Organization) với rất nhiều biến thể chi tiết (ví dụ: các mã 51, 53, 55 đại diện cho các mức độ mưa phùn khác nhau). Để tối ưu hóa cho mô hình phân loại và giảm bớt độ nhiễu không cần thiết, nhóm đã thực hiện kỹ thuật gom nhóm các mã tương đồng. Cụ thể, các mã phản ánh trạng thái ít mây hoặc mây rải rác (0, 1, 2, 3) được quy ước về nhóm "Nắng/Mây". Các

mã liên quan đến hiện tượng ngưng tụ và kết tủa bao gồm mưa phùn (51, 53, 55) và mưa (61, 63, 65) được hợp nhất thành nhóm "Mưa". Việc chuyển đổi này giúp tập dữ liệu trở nên cân bằng hơn, hỗ trợ mô hình tập trung vào các đặc trưng vật lý quan trọng thay vì bị phân tán bởi các định danh mã hóa quá chi tiết.

- Xử lý dữ liệu chuỗi thời gian, lag feature…

Do thời tiết là một hiện tượng có tính tuần tự và phụ thuộc chặt chẽ vào các trạng thái quá khứ, chúng tôi đã áp dụng kỹ thuật trích xuất đặc trưng trễ (Lag Features) để mô hình hóa mối quan hệ này. Các cột dữ liệu quan trọng như nhiệt độ, độ ẩm và áp suất khí quyển được dịch chuyển theo thời gian để tạo ra các biến mới như lag_1 (trạng thái cách đây 1 giờ), lag_3 (cách đây 3 giờ) và lag_6 (cách đây 6 giờ). Kỹ thuật này cho phép các mô hình tĩnh như Softmax hay Random Forest "nhìn" thấy được xu hướng biến động của thời tiết (ví dụ: áp suất giảm đột ngột hoặc nhiệt độ hạ nhanh trong vài giờ trước đó), từ đó cải thiện đáng kể khả năng dự báo cho các khung giờ tiếp theo.

- Lựa chọn đặc trưng, chuẩn hóa dữ liệu (quan trọng là với Softmax)

Để đảm bảo tính hiệu quả về mặt tính toán và tránh hiện tượng quá khớp (overfitting), một ma trận tương quan (Correlation Heatmap) đã được xây dựng. Dựa trên kết quả này, nhóm đã loại bỏ các đặc trưng có độ tương quan quá thấp với biến mục tiêu hoặc các biến có hiện tượng đa cộng tuyến mạnh (ví dụ như giữa nhiệt độ 2m và điểm sương).

3. Phương pháp nghiên cứu

a. Các mô hình nghiên cứu

Giới thiệu về model (nếu có điểm mới: kỹ thuật mới, kỹ thuật mới)

+ Softmax: Cách hàm tính xác suất đa lớp, tại sao chọn nó làm mốc so sánh…
+ Random: Cơ chế bỏ phiếu từ cây quyết định, tại sao nó mạnh trong giảm overfitting…

Cơ chế tính toán xác suất đa lớp

Softmax Regression (hay còn gọi là Multinomial Logistic Regression) là một bản mở rộng của Logistic Regression dành cho các bài toán phân loại có nhiều hơn hai lớp. Trong bài toán này, mô hình được sử dụng để phân loại ba trạng thái thời tiết: "Nắng/Mây", "Mưa" và "Thời tiết cực đoan".

Về mặt kỹ thuật, mô hình tính toán một số điểm (score) cho mỗi lớp bằng cách sử dụng tổ hợp tuyến tính của các đặc trưng đầu vào (nhiệt độ, độ ẩm, áp suất, các biến trễ...). Sau đó, hàm Softmax được áp dụng để chuyển đổi các điểm số này thành xác suất:

$$P(y = j \mid \mathbf{x}) = \frac{e^{\mathbf{w}_j^T \mathbf{x}}}{\sum_{k=1}^{K} e^{\mathbf{w}_k^T \mathbf{x}}}$$

- $\mathbf{x}$: vector đặc trưng đầu vào (features)
- y: nhãn (class)
- j: lớp thứ j
- K: tổng số lớp
- $\mathbf{w}_j$: vector trọng số ứng với lớp j
- $\mathbf{w}_j^T \mathbf{x}$: **score (logit)** của lớp j

Trong đó, xác suất của một lớp tỷ lệ thuận với hàm mũ của điểm số mà nó đạt được, đảm bảo rằng tổng xác suất của tất cả các lớp luôn bằng 1. Lớp có xác suất cao nhất sẽ được chọn làm nhãn dự báo cuối cùng cho khung giờ đó.

Lý do lựa chọn làm mốc so sánh (Baseline Model)

Chúng tôi lựa chọn Softmax Regression làm mô hình cơ sở (baseline) vì những lý do sau:

- Tính đơn giản và tường minh: Là một mô hình tuyến tính, Softmax giúp thiết lập một mức hiệu năng tiêu chuẩn. Nếu các mô hình phức tạp hơn (như Random Forest hay AdaBoost) không đạt được kết quả vượt trội hơn rõ rệt, thì mô hình đơn giản như Softmax sẽ được ưu tiên để tiết kiệm chi phí tính toán.
- Kiểm chứng tính tuyến tính của dữ liệu: Kết quả của Softmax giúp nhóm nghiên cứu đánh giá liệu các đặc trưng thời tiết (như mối quan hệ giữa nhiệt độ và khả năng có mưa) có thể được phân tách một cách tuyến tính hay không. Nếu Softmax đạt kết quả thấp, điều đó minh chứng cho tính phi tuyến phức tạp của dữ liệu khí tượng, từ đó khẳng định sự cần thiết của các thuật toán tổ hợp (Ensemble Learning).
- Tốc độ huấn luyện: Softmax cực kỳ nhanh và hiệu quả với dữ liệu đã được chuẩn hóa, giúp quy trình thực nghiệm diễn ra nhanh chóng trước khi triển khai các mô hình nặng hơn.

Ứng dụng vào bài toán

Trong nghiên cứu này, Softmax Regression đóng vai trò là "thước đo" để so sánh khả năng học máy. Mô hình nhận đầu vào là các vector đặc trưng sau khi đã được chuẩn hóa bằng StandardScaler. Việc so sánh kết quả giữa Softmax và các mô hình cây quyết định sẽ làm rõ liệu việc kết hợp nhiều đặc trưng thời quá khứ (lag features) có tạo ra các ngưỡng quyết định phức tạp mà một hàm tuyến tính không thể bắt kịp hay không.

+ AdaBoost: Cách tập trung vào mẫu dữ liệu khó, cập nhật trọng số,....

Thứ nhất, AdaBoost được triển khai theo thuật toán SAMME, cho phép xử lý trực tiếp bài toán đa lớp trong một mô hình boosting thống nhất, thay vì sử dụng các chiến lược phân rã bài toán thủ công (One vs Rest).

Thứ hai, AdaBoost được huấn luyện trên tập đặc trưng đã qua xử lý lần hai, bao gồm các đặc trưng chuỗi thời gian được thiết kế bổ sung bên cạnh các biến khí tượng gốc. Cách tiếp cận này cho phép AdaBoost khai thác không chỉ thông tin tức thời, mà còn cả xu hướng và trạng thái tích lũy của hệ thống thời tiết.

Thứ ba, nghiên cứu không sử dụng decision stump truyền thống (cây độ sâu 1) làm bộ phân loại yếu, mà thay thế bằng cây quyết định nông có kiểm soát cấu trúc, được thể hiện thông qua các ràng buộc về độ sâu, số mẫu tối thiểu tại nút và số đặc trưng được xét khi chia nhánh. Cách thiết kế này cho phép các weak learner mô hình hóa các quan hệ phi tuyến giữa các yếu tố khí tượng, đồng thời hạn chế overfitting trong môi trường dữ liệu nhiễu.

Cách hoạt động (model áp dụng vào bài toán phân loại này như thế nào + lý do)

- **Random Forest**
- **Softmax Regression**
- AdaBoost

Trong AdaBoost, weak learner thường được chọn là decision stump (tạm dịch: gốc cây quyết định- cây quyết định độ sâu 1). Đây là một mô hình cực kỳ đơn giản- chỉ sử dụng một đặc trưng duy nhất và một ngưỡng chia để tách dữ liệu thành hai nhóm. Tuy nhiên, để phù hợp với bài toán phân loại điều kiện thời tiết sau khi gom nhóm, mô hình cần được điều chỉnh cả về chiến lượng phân loại đa lớp (cụ thể là 3 lớp) lẫn biểu diễn dữ liệu đầu vào.

Trong nghiên cứu này, thuật toán SAMME (Stagewise Additive Modeling using a Multi-class Exponential loss) được sử dụng để mở rộng AdaBoost cho bài toán đa lớp một cách nhất quán về mặt toán học. Khác với chiến lược one-vs-rest, SAMME xây dựng một mô hình boosting duy nhất cho toàn bộ các lớp, trong đó mỗi bộ phân loại yếu đưa ra dự đoán trực tiếp trên không gian đa lớp. Cách tiếp cận này cho phép AdaBoost khai thác đồng thời mối quan hệ giữa các lớp thời tiết, thay vì huấn luyện các bộ phân loại nhị phân độc lập.

Trong SAMME, trọng số của mỗi bộ phân loại yếu được xác định dựa trên sai số phân loại đa lớp, với hệ số:

$$\alpha^{(m)} = \log \frac{1 - err^{(m)}}{err^{(m)}} + \log(K - 1).$$

trong đó K là số lớp thời tiết. Khi K = 2, SAMME thu về đúng AdaBoost chuẩn, tuy nhiên khi K > 2, số hạng log(K - 1) đóng vai trò then chốt:

- Tăng mức đóng góp của các bộ phân loại có hiệu năng tốt hơn ngẫu nhiên
- Đảm bảo sự cân bằng giữa các lớp trong quá trình boosting

b. Quy trình đánh giá khoa học

- Phân chia dữ liệu, tìm kiếm siêu tham số tốt nhất

Trong nghiên cứu này, các siêu tham số quan trọng ví dụ như số vòng boosting (n_estimators) và tốc độ học (learning_rate), không được lựa chọn thủ công mà được xác định thông qua Grid Search. Phương pháp này cho phép đánh giá có hệ thống nhiều cấu hình khác nhau, từ đó lựa chọn tập tham số mang lại hiệu năng phân loại tốt nhất.

Quá trình tìm kiếm siêu tham số được thực hiện kết hợp với TimeSeriesSplit, trong đó dữ liệu được chia theo đúng trình tự thời gian. Cách chia này đảm bảo mô hình chỉ học từ dữ liệu trong quá khứ để dự đoán dữ liệu ở các thời điểm sau, tránh hiện tượng rò rỉ thông tin và phản ánh đúng kịch bản triển khai thực tế của bài toán phân loại thời tiết.

- Metric đánh giá

Hiệu năng của các mô hình Random Forest, Softmax Regression và AdaBoost có thể được đánh giá thông qua các chỉ số: Accuracy, Precision, Recall và F1-score. Các chỉ số này được sử dụng để đo lường khả năng dự đoán của mô hình

Accuracy phản ánh độ chính xác tổng thể của mô hình thông qua tỷ lệ các mẫu được dự đoán đúng trên toàn bộ tập dữ liệu. Chỉ số này xét đồng thời các trường hợp dự đoán đúng dương và âm, tuy nhiên có thể gây hiểu lầm khi dữ liệu mất cân bằng.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision đo lường độ chính xác của các dự đoán dương, thể hiện tỷ lệ các mẫu được dự đoán đúng là dương trên tổng số các mẫu được phân loại là dương. Chỉ số này cho thấy mức độ tin cậy của các dự đoán dương và giúp đánh giá khả năng hạn chế các dự đoán sai dương.

$$\text{Precision} = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

Recall đánh giá khả năng mô hình nhận diện đúng các trường hợp dương thông qua tỷ lệ các mẫu dương được phát hiện trên tổng số các mẫu dương thực sự. Chỉ số này phản ánh mức độ nhạy của mô hình đối với lớp dương và khả năng giảm thiểu các trường hợp bỏ sót.

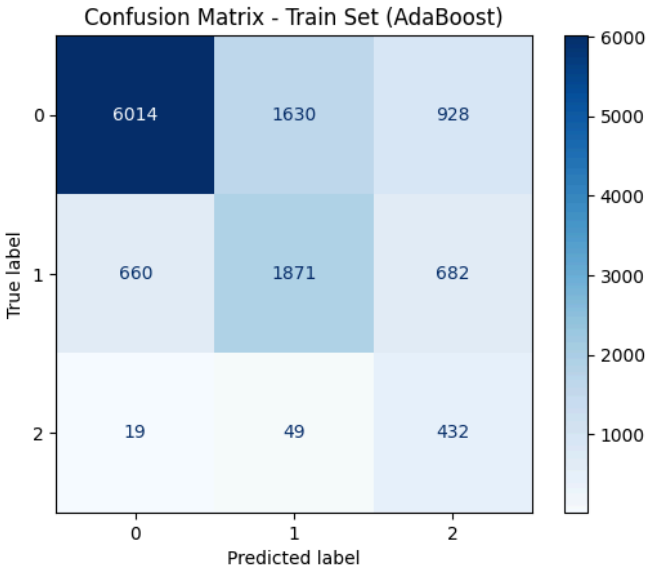
$$\text{Recall} = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

F1-score là chỉ số tổng hợp giữa Precision và Recall, nhằm cân bằng giữa độ chính xác của dự đoán dương và khả năng bao phủ các mẫu dương. Giá trị F1-score càng cao cho thấy hiệu năng phân loại tổng thể của mô hình càng tốt.

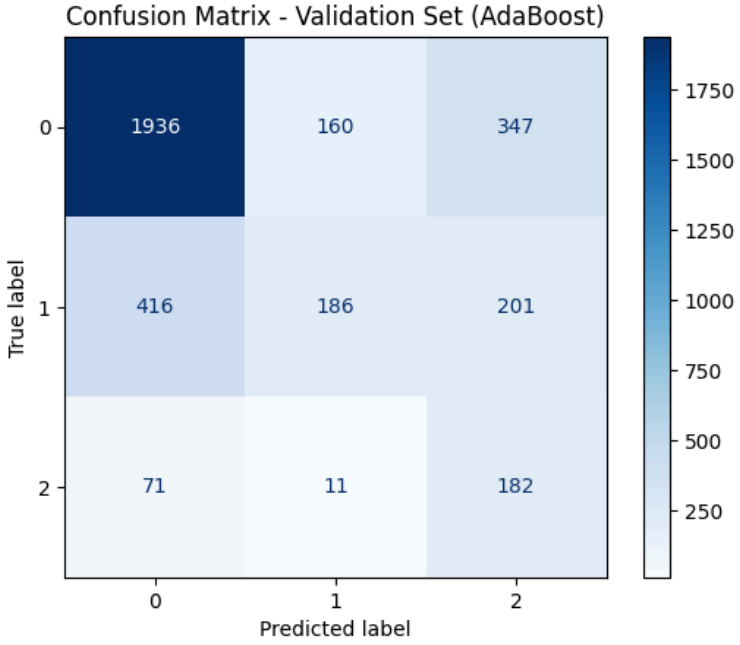
$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

4. **Thực nghiệm và kết quả**
    a. **Trên tập train**

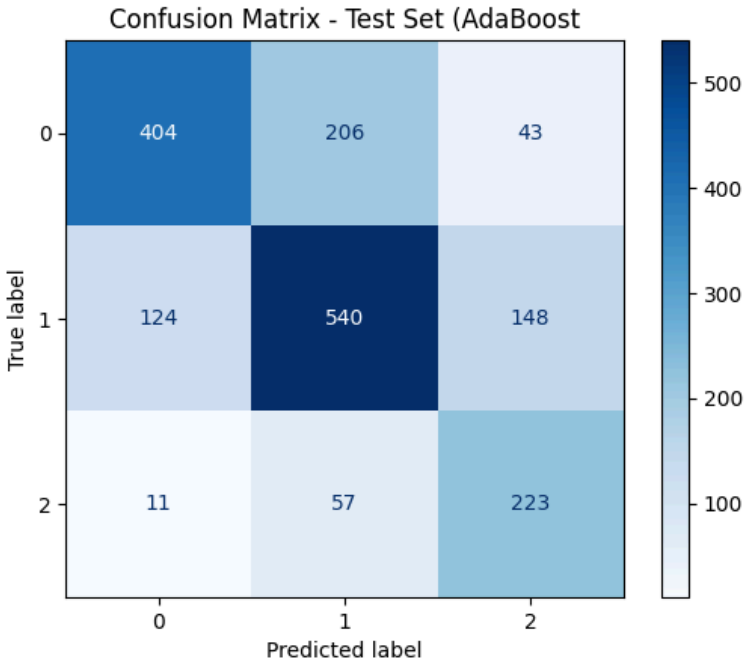
	Precision	**Recall**	**F1 - Score**	**Support**
Clo				



b. Trên tập validation



c. Trên tập test



Hình: Ma trận nhầm lẫn trong việc phân loại thời tiết sử dụng mô hình AdaBoost

Ma trận nhầm lẫn cho thấy mô hình AdaBoost phân loại tương đối hiệu quả ba trạng thái thời tiết *cloudy (lớp 0), drizzle (lớp 1) và rain (lớp 2),* với phần lớn dự đoán đúng tập trung trên đường chéo chính. Trong đó, lớp *drizzle* đạt số lượng dự đoán đúng cao nhất, phản ánh đây là trạng thái có đặc trưng khí tượng tương đối nhất quán sau khi được gom nhóm.

Các lỗi phân loại chủ yếu xuất hiện giữa *cloudy* và *drizzle*, cũng như giữa *drizzle* và *rain*. Đây là hiện tượng hợp lý về mặt vật lý, do *drizzle* nằm ở vùng chuyển tiếp giữa không mưa (*cloudy*) và mưa rõ rệt (*rain*). Trong nhiều thời điểm, lượng mưa rất nhỏ hoặc gián đoạn khiến các đặc trưng như độ ẩm, độ che phủ mây và áp suất của *drizzle* gần với *cloudy*, trong khi ở các thời điểm khác, chúng lại tiệm cận *rain*.

Ngược lại, mô hình ít nhầm lẫn trực tiếp giữa *cloudy* và *rain*, cho thấy hai trạng thái này có sự khác biệt rõ ràng hơn về lượng mưa và cấu trúc khí tượng. Nhìn chung, các sai lệch dự đoán tập trung tại vùng biên khí tượng, nơi ranh giới giữa các trạng thái thời tiết không sắc nét và các đặc trưng đầu vào có sự chồng lấn mạnh, làm tăng độ khó cho bài toán phân loại đa lớp.

- Độ quan trọng của đặc trưng (Trích từ Random Forest) để chứng minh yếu tố nào thực sự định hình điều kiện thời tiết đà nẵng

Hình: Mức độ quan trọng của các đặc trưng trong mô hình Random Forest

- So sánh 3 mô hình với nhau: độ chính xác, thời gian huấn luyện

5. Thảo luận

- Chỉ ra ưu/nhược điểm của từng mô hình

Random Forest

Softmax Regression

AdaBoost

AdaBoost (SAMME) thể hiện ưu điểm rõ rệt trong việc khai thác các mẫu dữ liệu khó phân loại thông qua cơ chế cập nhật trọng số mẫu sau mỗi vòng lặp. Các quan sát bị phân loại sai sẽ được tăng trọng số, buộc các bộ phân loại yếu ở các vòng sau tập trung nhiều hơn vào những vùng dữ liệu phức tạp. Nhờ đó, mô hình có khả năng cải thiện dần hiệu năng tổng thể, đặc biệt tại các vùng biên giữa các lớp thời tiết có đặc điểm khí tượng chồng lấn.

Trong nghiên cứu này, AdaBoost không sử dụng decision stump truyền thống mà kết hợp với cây quyết định nông có ràng buộc cấu trúc. Cách tiếp cận này giúp các weak learner có đủ năng lực để học các quan hệ phi tuyến cục bộ giữa các yếu tố khí tượng như nhiệt độ, độ ẩm và áp suất, đồng thời vẫn kiểm soát được hiện tượng overfitting thông qua giới hạn độ sâu cây, số mẫu tối thiểu và số đặc trưng được xét tại mỗi nút chia.

Tuy nhiên, AdaBoost (SAMME) cũng bộc lộ một số hạn chế. Do cơ chế tăng trọng số mạnh cho các mẫu bị phân loại sai, mô hình trở nên nhạy cảm với nhiễu và các mẫu ngoại lai, vốn phổ biến trong dữ liệu khí tượng thực tế. Bên cạnh đó, trong bối cảnh dữ liệu có cấu trúc phức tạp và phân bố không đồng đều giữa các lớp thời tiết, hiệu năng của AdaBoost vẫn chưa vượt trội so với Random Forest, đặc biệt về tính ổn định và khả năng tổng quát hóa.

- Model nào phù hợp nhất + giải thích lý do

Dựa trên kết quả thực nghiệm, Random Forest được xác định là mô hình phù hợp nhất cho bài toán phân loại điều kiện thời tiết tại Đà Nẵng. Điều này xuất phát từ đặc thù dữ liệu khí tượng có tính phi tuyến cao và tồn tại nhiều mối quan hệ tương tác phức tạp giữa các đặc trưng, trong khi Random Forest có khả năng khai thác hiệu quả các tương tác này mà không cần giả định tuyến tính, đồng thời giảm phương sai thông qua cơ chế tổ hợp nhiều cây quyết định. Bên cạnh đó, mô hình cung cấp thông tin về độ quan trọng của đặc trưng, góp phần nâng cao tính ổn định và khả năng diễn giải của kết quả phân loại.

So với Softmax Regression, Random Forest vượt trội hơn trong việc mô hình hóa các ranh giới phân lớp phi tuyến, đặc biệt tại các vùng biên giữa các trạng thái thời tiết có đặc trưng chồng lấn như *cloudy* và *drizzle*. Đối với AdaBoost, mặc dù mô hình này chú trọng tốt vào các mẫu khó phân loại, hiệu năng lại nhạy cảm với nhiễu và phân bố lớp không cân bằng, khiến kết quả kém ổn định hơn so với Random Forest.

6. Kết luận và hướng phát triển

a. Kết luận

Trong quá trình thực hiện đề tài, việc thu thập, xử lý và tổ chức dữ liệu đóng vai trò nền tảng cho toàn bộ quy trình xây dựng mô hình phân loại thời tiết. Nhóm đã thu thập thành công dữ liệu khí tượng theo giờ tại thành phố Đà Nẵng với số lượng mẫu đủ lớn để phản ánh đặc trưng biến thiên theo thời gian của các yếu tố khí tượng. Dữ liệu được xử lý qua nhiều bước bao gồm làm sạch, xử lý giá trị thiếu và ngoại lai, chuẩn hóa theo thời gian, đồng thời đơn giản hóa và gom nhóm các mã thời tiết ban đầu thành các nhóm có ý nghĩa khí tượng rõ ràng.

Việc gom nhóm weather code giúp giảm độ phức tạp của không gian nhãn, hạn chế sự chồng lấn giữa các trạng thái thời tiết và cải thiện khả năng học của các mô hình phân loại. Bên cạnh đó, dữ liệu được tiếp cận dưới dạng chuỗi thời gian, cho phép khai thác tốt hơn tính liên tục và quy luật biến thiên theo thời gian của các đặc trưng khí tượng, đặc biệt tại các vùng chuyển tiếp giữa các trạng thái thời tiết.

Trên cơ sở dữ liệu đã xử lý, ba mô hình Softmax Regression, AdaBoost (SAMME) và Random Forest được đánh giá và so sánh. Kết quả cho thấy Random Forest đạt hiệu năng và độ ổn định cao nhất trong bối cảnh dữ liệu khí tượng có tính phi tuyến và phân bố lớp không đồng đều. Nhờ khả năng khai thác tương tác đặc trưng, giảm nhiễu và duy trì thời

gian suy luận hợp lý, Random Forest được xác định là mô hình phù hợp nhất cho hệ thống phân loại và cảnh báo thời tiết gần thời gian thực.

b. Hướng phát triển

Trước hết, việc tối ưu hóa bộ tham số bằng các phương pháp thông minh như Bayesian Optimization hoặc các thuật toán tiến hóa có thể giúp cải thiện thêm hiệu năng phân loại so với cách dò lưới truyền thống.

Thứ hai, mở rộng và cân bằng tập dữ liệu, đặc biệt đối với các lớp thời tiết hiếm, thông qua các kỹ thuật oversampling hoặc tổng hợp dữ liệu, sẽ góp phần nâng cao khả năng tổng quát hóa của mô hình.

Cuối cùng, việc kết hợp thêm các đặc trưng đa nguồn và nghiên cứu triển khai kèm cơ chế giám sát hiệu suất theo thời gian sẽ giúp hệ thống thích ứng tốt hơn với sự biến đổi của điều kiện khí tượng.