Ứng dụng học máy với mô hình Random Forest để dự báo thời tiết tại Thành phố Hà Nội

Đồng Mạnh Hùng[1], Đoàn Quang Huy[1], Nguyễn Trần Huy[1,*]

Ngày 04 tháng 12 năm 2024

[1]Viện Trí tuệ nhân tạo, Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội, Việt Nam

Tóm tắt : Bài báo này nghiên cứu việc áp dụng phương pháp học máy Random Forest (RF) để dự báo nhiệt độ, độ ẩm và lượng mưa tại thành phố Hà Nội, Việt Nam, dựa trên 13 đặc trưng thời tiết. Quy trình nghiên cứu bao gồm các bước thu thập và tiền xử lý dữ liệu, phân chia dữ liệu thành tập huấn luyện và kiểm tra, huấn luyện mô hình, và đánh giá hiệu suất của mô hình qua các chỉ số như hệ số tương quan, sai số trung bình tuyệt đối (MAE) và căn bậc hai sai số trung bình (RMSE). Kết quả nghiên cứu cho thấy phương pháp RF đạt được hệ số tương quan trung bình (R) vượt ngưỡng 0,8361, với giá trị MAE dưới 1,1738 và RMSE dưới 1,6641. Những kết quả này khẳng định tiềm năng của phương pháp RF trong việc phát triển hệ thống dự báo thời tiết chính xác cao, đặc biệt phù hợp với điều kiện khí hậu tại Hà Nội.

Từ khóa: Random Forest, học máy, lượng mưa, Hà Nội, dự báo khí tượng.

1. Giới thiệu

Với sự phát triển nhanh chóng của công nghệ, nhiều lĩnh vực trong cuộc sống con người đã và đang thay đổi sâu sắc, đặc biệt nhờ vào sự tiến bộ của Trí tuệ Nhân tạo (AI). AI, với khả năng xử lý và phân tích dữ liệu vượt trội, đã trở thành trụ cột của kỷ nguyên 4.0, hỗ trợ và thay thế con người trong nhiều hoạt động, bao gồm cả dự báo thời tiết. Việc dự báo các yếu tố thời tiết như nhiệt độ, độ ẩm và lượng mưa đã được hưởng lợi đáng kể từ các mô hình AI, nhờ khả năng xử lý các tập dữ liệu phức tạp và tích hợp nhiều nguồn thông tin khác nhau.

Trong bối cảnh biến đổi khí hậu toàn cầu, các hiện tượng thời tiết cực đoan như mưa lớn, bão và hạn hán ngày càng khó dự đoán và xuất hiện với tần suất cao hơn, gây ra thiệt hại nghiêm trọng về con người và tài sản. Điều này đã làm gia tăng nhu cầu phát triển các hệ thống dự báo thời tiết chính xác, nhằm cung cấp thông tin kịp thời cho công tác phòng chống và giảm thiểu tác động của thiên tai.

Giữa các phương pháp học máy hiện đại, thuật toán Random Forest đã chứng tỏ hiệu quả vượt trội trong dự báo các yếu tố thời tiết như nhiệt độ, độ ẩm và lượng mưa. Thuật toán này không chỉ mang lại độ chính xác cao mà còn có khả năng khai thác tối đa giá trị từ các đặc trưng thời tiết, qua đó tăng cường năng lực dự báo và giảm thiểu rủi ro từ thiên tai, góp phần bảo vệ cộng đồng và thúc đẩy sự phát triển bền vững.

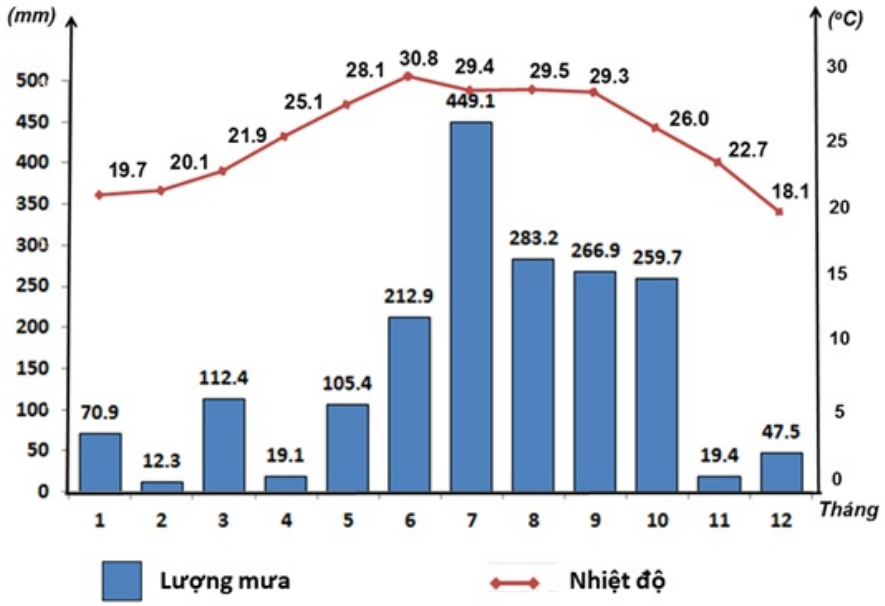
2. Khu vực nghiên cứu

Hình 1: Vị trí địa lý và đặc điểm địa hình của Hà Nội trên bản đồ lãnh thổ Việt Nam. (Nguồn: *worldatlas.com*)

Thành phố Hà Nội, thủ đô của Việt Nam, nằm tại khu vực đồng bằng Bắc Bộ, có địa hình chủ yếu là đồng bằng bằng phẳng và không giáp biển. Đây là vùng đất được hình thành và bồi đắp bởi phù sa từ Sông Hồng, con sông có lưu vực lớn thứ hai tại Việt Nam chỉ sau sông Mekong (Hình 1). Sự kết hợp giữa địa hình và khí hậu nóng ẩm mưa nhiều đã tạo điều kiện thuận lợi cho việc phát triển nền nông nghiệp, đặc biệt là trồng cây lúa nước, đóng góp vào sản lượng lương thực của vùng Đồng bằng Sông Hồng nói riêng và của Việt Nam nói chung.

Vị trí địa lý của Hà Nội nằm trong vùng khí hậu nhiệt đới gió mùa, khiến cho thời tiết tại đây chủ yếu là nóng ẩm quanh năm. Nhiệt độ trung bình trong phần lớn thời gian thường xuyên vượt 20^oC, và biên độ nhiệt độ giữa tháng cao nhất và tháng thấp nhất khá lớn, khoảng 10^oC. Độ ẩm trung bình luôn ở mức trên 70

Dữ liệu về thời tiết và khí tượng tại Hà Nội được cung cấp bởi Tổng cục Khí tượng Thủy văn thuộc Bộ Tài nguyên và Môi trường Việt Nam. Các đặc trưng thời tiết được thu thập bao gồm các yếu tố như nhiệt độ, gió, độ ẩm, áp suất, và lượng mưa. Bảng 1 liệt kê chi tiết các đặc trưng này được sử dụng trong tập dữ liệu nghiên cứu. Lưu ý rằng các thông số liên quan đến nhiệt độ, gió, độ ẩm, và áp suất được ghi nhận tại thời điểm thu thập số liệu, trong khi thông tin về lượng mưa được tính toán dựa trên khoảng thời gian từ một giờ trước đến thời điểm ghi nhận (xem thêm chú thích trong tài liệu).



Hình 2: Biểu đồ nhiệt độ, lượng mưa ở Hà Nội trong năm 2017.

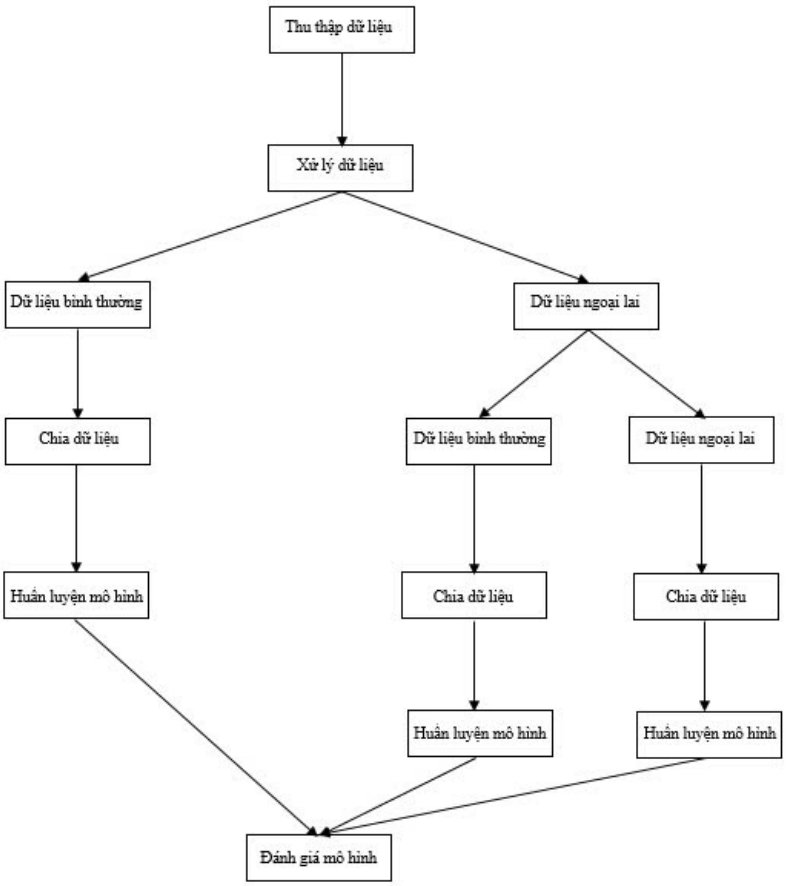
Đặc trưng	Mô tả	Đơn vị	Kiểu dữ liệu
temperature	Nhiệt độ	^oC	float
pressure	Áp suất khí quyển	hPa	int
wind-speed	Tốc độ gió	m/s	float
cloud	Độ bao phủ mây trên bầu trời	Phần trăm	int
humidity	Độ ẩm	Phần trăm	int
dew-point	Nhiệt độ điểm sương	^oC	float
sea-level-pressure	Áp suất tại mực nước biển	hPa	float
wind-direction	Hướng mà gió thổi từ	Tọa độ la bàn (^{o})	int
solar-rad	Năng lượng từ bức xạ mặt trời	W/m^2	int
precipitation	Lượng mưa	mm	float

Bảng 1: Các đặc trưng xuất hiện trong tập dữ liệu khí tượng tại Hà Nội.

3. Phương pháp nghiên cứu

Random Forest là một thuật toán học máy thuộc nhóm mô hình cây quyết định, được sử dụng chủ yếu cho các bài toán phân loại và hồi quy. Nó hoạt động bằng cách tạo ra nhiều cây quyết định (gọi là "cây con") từ các mẫu dữ liệu khác nhau và kết hợp kết quả của các cây này để đưa ra dự đoán cuối cùng. Quy trình này giúp giảm thiểu hiện tượng overfitting và tăng tính chính xác so với một cây quyết định đơn lẻ. Mỗi cây trong rừng được huấn luyện trên một tập dữ liệu con ngẫu nhiên và chỉ sử dụng một tập con các đặc trưng để chia tách, từ đó tạo ra sự đa dạng giữa các cây. Random Forest có ưu điểm là dễ sử dụng, hiệu quả trên các tập dữ liệu lớn và có khả năng xử lý dữ liệu thiếu hoặc bị nhiễu tốt. Dưới đây là hình ảnh về các giai đoạn của phương pháp nghiên cứu như trong Hình 3.

Trình tự các giai đoạn của phương pháp nghiên cứu được thực hiện theo các bước sau:



Hình 3: Sơ đồ mô hình nghiên cứu dự báo nhiệt độ, độ ẩm, lượng mưa tại Hà Nội

Bước 1: Sau khi thực hiện rà soát tài liệu và xác định vấn đề nghiên cứu, tiến hành thu thập dữ liệu chuỗi thời gian theo giờ[1] từ ngày 01/01/2020 đến 31/12/2023. Sau đó, tiến hành tiền xử lý dữ liệu[2]. Bộ dữ liệu ban đầu chưa bao gồm các biến mục tiêu cần dự đoán, vì vậy cần xử lý dữ liệu bằng cách trích xuất thông tin về nhiệt độ, độ ẩm và lượng mưa từ các dòng tiếp theo trong tập dữ liệu. Tiến hành trích xuất các đặc trưng về giờ, ngày và tháng từ cột 'datetime' của dữ liệu nhằm làm phong phú bộ đặc trưng. Việc này không chỉ giúp nâng cao khả năng phân tích dữ liệu mà còn tận dụng tính chu kỳ của thời gian để cải thiện độ chính xác của mô hình dự đoán.

Bước 2: Thực hiện phân chia dữ liệu thành hai phần: dữ liệu ngoại lai và dữ liệu bình thường. Dữ liệu bình thường nằm trong đoạn $[Q_1 - 1.5 * IQR, Q_3 + 1.5 * IQR]$, còn dữ liệu ngoại lai sẽ nằm ngoài đoạn trên. Trong đó Q_1 là giá trị phần tư đầu tiên trong tập dữ liệu, Q_3 là giá trị phần tư thứ ba trong tập dữ liệu và IQR là khoảng tứ phân vị. Tập hợp các giá trị ngoại lai duy nhất trong cột lượng mưa được sử dụng để xác định dữ liệu ngoại lai. Sau khi tách dữ liệu đầu vào thành hai phần (dữ liệu bình thường và dữ liệu ngoại lai), tập dữ liệu ngoại lai tiếp tục được chia nhỏ thành hai phần: dữ liệu bình thường trong ngoại lai và dữ liệu ngoại lai đặc biệt.

[1]Các khung giờ trong tập dữ liệu được ghi lại theo múi giờ GMT+0, chậm hơn múi giờ địa phương nơi nghiên cứu 7h.

[2]Thuật ngữ 'tiền xử lý dữ liệu' thường được sử dụng trong lĩnh vực học máy để chỉ các quy trình liên quan đến việc tổ chức, làm sạch và chuẩn bị dữ liệu thô trước khi sử dụng để xây dựng các mô hình học máy.

Bước 3: Sau khi dữ liệu được tiền xử lý, quá trình chia dữ liệu được thực hiện để phân tách thành hai tập con: dữ liệu huấn luyện và dữ liệu kiểm tra. Dữ liệu huấn luyện được sử dụng để huấn luyện các mô hình học máy, trong khi dữ liệu kiểm tra được sử dụng để đánh giá hiệu suất của mô hình. Trong nghiên cứu này, tỷ lệ chia dữ liệu là 20 phần trăm cho dữ liệu kiểm tra (từ ngày 01/02/2023 trở về sau) và 80 phần trăm cho dữ liệu huấn luyện (trước ngày 01/02/2023). Đây là dự đoán chuỗi thời gian do đó dữ liệu huấn luyện và kiểm tra không được chia một cách ngẫu nhiên, mà chúng được chia theo trình tự thời gian để đảm bảo tính đúng đắn và chính xác của mô hình.

Bước 4: Bước tiếp theo là quá trình huấn luyện mô hình. Huấn luyện mô hình (Model Training) là quá trình dạy một mô hình học máy học cách đưa ra dự đoán hoặc quyết định dựa trên dữ liệu đầu vào. Quá trình này bao gồm việc cung cấp cho mô hình một tập dữ liệu huấn luyện chứa các đầu vào và kết quả tương ứng. Mục tiêu của huấn luyện là nhận diện các mẫu hoặc mối quan hệ trong dữ liệu đầu vào để dự đoán chính xác kết quả cho dữ liệu chưa được thấy trước. Quá trình huấn luyện mô hình được thực hiện thông qua một quy trình tối ưu hóa, được gọi là tinh chỉnh siêu tham số. Tinh chỉnh siêu tham số liên quan đến việc xác định sự kết hợp tối ưu của các siêu tham số để đạt được hiệu suất tốt nhất trên tập dữ liệu xác thực. Trong nghiên cứu này, phương pháp tối ưu hóa Optuna đã được áp dụng trong mô hình Random Forest để tìm ra các tham số tốt nhất.

Bước 5: Khi quá trình huấn luyện hoàn tất, mô hình đã được huấn luyện có thể được sử dụng để thực hiện dự đoán trên dữ liệu mới mà chưa được huấn luyện trong bước triển khai mô hình.

4. Phương pháp đánh giá

Để đánh giá hiệu quả của phương pháp Random Forest trong việc ước tính lượng mưa của Hà Nội, nghiên cứu đã áp dụng các chỉ số thống kê để đánh giá hiệu suất của mô hình.

1. ME (Mean Error): Thể hiện xu hướng của mô hình.

$$ME = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \tag{1}$$

2. MAE (Mean Absolute Error): Trung bình của sai số tuyệt đối, phản ánh độ lệch trung bình giữa giá trị dự đoán và thực tế.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2}$$

3. $RMSE$ (Root Mean Squared Error): Đo lường độ lệch giữa giá trị dự đoán và thực tế, nhấn mạnh hơn vào các sai số lớn do tính bình phương.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3}$$

4. R: Đo mức độ tương quan tuyến tính giữa giá trị thực tế và giá trị dự đoán.

$$R = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}} \tag{4}$$

5. R^2: Đo mức độ mô hình giải thích được sự biến thiên của dữ liệu.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{5}$$

Trong đó:

- y_i: giá trị thực tế.
- $\hat{y}_i$: giá trị dự đoán.
- n: số lượng mẫu.
- $\bar{y}$: giá trị trung bình của y.
- $\bar{\hat{y}}$: giá trị trung bình của $\hat{y}$.

5. Kết quả và thảo luận

Thời gian	ME	MAE	$RMSE$	R^2	R
01:00:00	0.0141	0.5765	1.1888	0.7837	0.8752
04:00:00	-0.067	0.648	1.1302	0.7348	0.8859
07:00:00	-0.3245	1.1738	1.6641	0.826	0.912
10:00:00	-0.1118	1.0238	1.4074	0.885	0.9395
13:00:00	0.1057	0.9572	1.541	0.6827	0.8361
16:00:00	0.0936	0.8293	1.2098	0.7508	0.8429
19:00:00	-0.1085	0.6587	1.0744	0.8392	0.9104
22:00:00	-0.1264	0.6286	1.0905	0.86	0.9249

Bảng 2: Kết quả đánh giá ước tính lượng mưa theo giờ của 8 thời điểm khác nhau trong ngày tại Hà Nội.

Mô hình dự báo thời tiết về nhiệt độ, độ ẩm và lượng mưa tại Thành phố Hà Nội, Việt Nam được huấn luyện bởi mô hình Random Forest bằng cách sử dụng tập dữ liệu huấn luyện gồm 13 yếu tố ảnh hưởng. Kết quả đánh giá của phương pháp Random Forest xét tại các thời điểm 01:00:00, 04:00:00, 07:00:00, 10:00:00, 13:00:00, 16:00:00, 19:00:00, 22:00:00 được thể hiện trong bảng 2.

Kết quả từ bảng đánh giá ước tính lượng mưa theo giờ tại tám thời điểm khác nhau trong ngày cho thấy chỉ số ME (Mean Error) tại các thời 04:00:00, 07:00:00, 10:00:00, 19:00:00, 22:00:00 đều có giá trị âm, chỉ ra rằng lượng mưa thực tế đo được thấp hơn so với giá trị dự đoán. Ngược lại, tại các thời điểm 01:00:00, 13:00:00, 16:00:00, chỉ số ME có giá trị dương, cho thấy lượng mưa dự đoán từ mô hình thấp hơn so với lượng mưa thực tế quan sát được.

Chỉ số MAE (Mean Absolute Error) dao động trong khoảng từ 0,6765 đến 1,1738, trong khi chỉ số $RMSE$ (Root Mean Squared Error) có giá trị từ 1,0744 đến 1,6641. Chỉ số R^2 thể hiện tỷ lệ biến thiên của biến phụ thuộc (biến mục tiêu) mà mô hình có thể giải thích được dựa trên các biến độc lập, cho thấy mức độ phù hợp cao của mô hình. Cụ thể, giá trị R^2 đều lớn hơn 0,6872 và cao nhất đạt 0,885, đồng nghĩa với việc từ 68,72%

đến 88,5% biến thiên của dữ liệu dự báo có thể được giải thích bởi các giá trị quan trắc thực tế. Điều này chứng tỏ mô hình có khả năng dự đoán đáng tin cậy.

Mối quan hệ tương quan giữa lượng mưa ước tính từ dữ liệu vệ tinh và lượng mưa quan trắc đạt kết quả rất tốt, với hệ số tương quan R đều lớn hơn 0,8361, cho thấy sự tương đồng cao và độ tin cậy tốt giữa hai nguồn dữ liệu này.

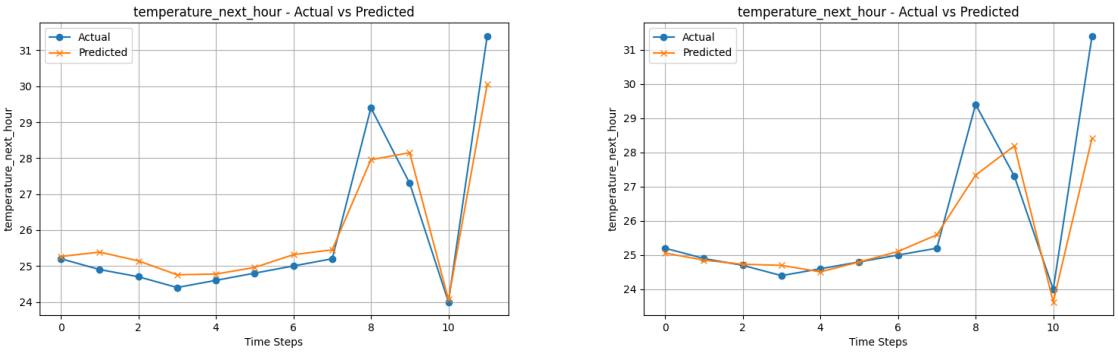
Bảng 3 trình bày các hàm đánh giá cho từng giá trị đầu ra, với giá trị trung bình được tính trong các giờ 01:00:00, 04:00:00, 07:00:00, 10:00:00, 13:00:00, 16:00:00, 19:00:00, 22:00:00.

Mục tiêu	ME	MAE	$RMSE$	R^2	R
Temperature	0.0859	0.3986	0.6441	0.9833	0.9918
Humidity	-0.2457	1.7666	2.5559	0.9728	0.9865
Precipitation	-0.037	0.2707	0.7977	0.4468	0.676

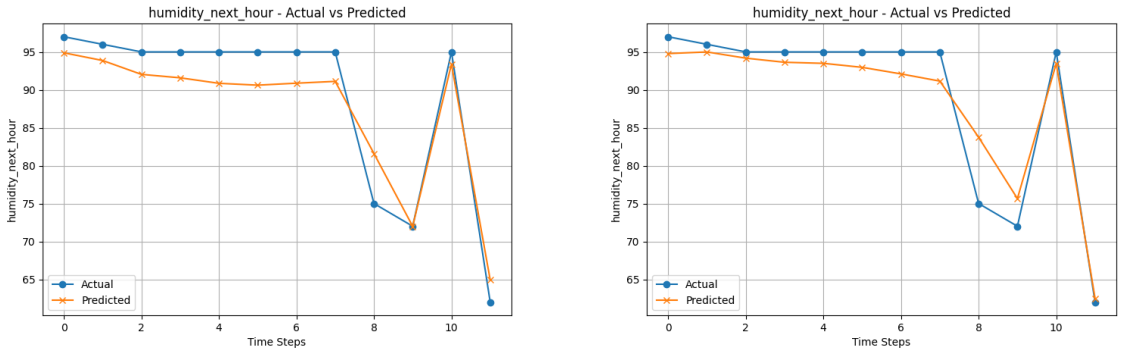
Bảng 3: Kết quả đánh giá cho việc ước tính từng giá trị đầu ra.

Mô hình dự báo cho thấy hiệu suất rất cao đối với nhiệt độ, với chỉ số $R^2 = 0.9833$ và $R = 0.9918$, chứng tỏ khả năng dự đoán chính xác. Các chỉ số về mặt sai số ở mức thấp, như $MAE = 0.3986$ và $RMSE = 0.6441$, cùng với giá trị $ME = 0.3986$ nhỏ, cho thấy sự ổn định và ít lệch trong dự đoán. Đối với độ ẩm, mô hình cũng hoạt động khá tốt với $R^2 = 0.9728$ và $R = 0.9865$, mặc dù sai số lớn hơn ($MAE = 1.7666$, $RMSE = 2.55559$) và ME = -0.2457 cho thấy có xu hướng dự đoán thấp hơn một chút. Tuy nhiên, hiệu suất dự đoán lượng mưa còn nhiều hạn chế, với $R^2 = 0.4468$ và $R = 0.676$ khá thấp, phản ánh mô hình chưa mô tả tốt sự biến động phức tạp của lượng mưa, mặc dù MAE và $RMSE$ vẫn ở mức tương đối thấp (lần lượt là 0.2707 và 0.7977). Điều này cho thấy dữ liệu đầu vào chưa đủ chi tiết hoặc chưa cân bằng để dự báo chính xác lượng mưa, đòi hỏi cải thiện cả về chất lượng lẫn đặc trưng dữ liệu.

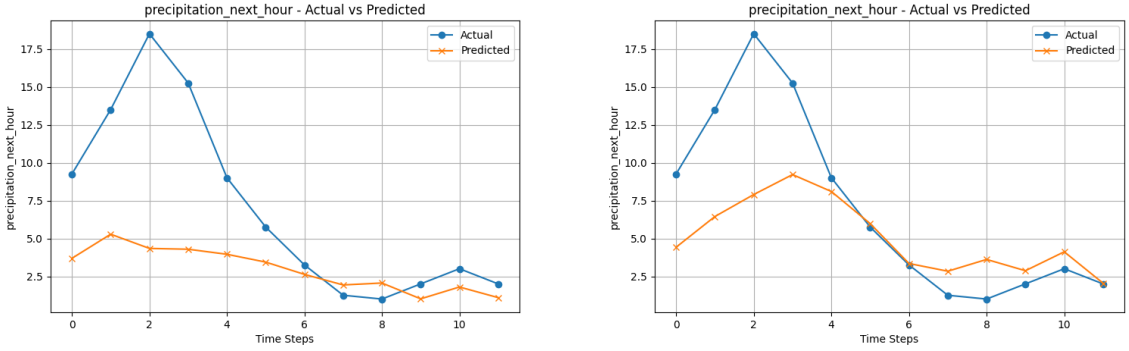
Đánh giá khả năng dự báo nhiệt độ, độ ẩm và lượng mưa của mô hình sau khi cải tiến thông qua các biểu đồ được trình bày dưới đây. Cụ thể, biểu đồ bên trái thể hiện kết quả dự báo khi áp dụng thuật toán Random Forest trên toàn bộ tập dữ liệu ban đầu, trong khi biểu đồ bên phải thể hiện kết quả dự báo sau khi áp dụng thuật toán Random Forest trên tập dữ liệu được chia.



Hình 4: So sánh giữa nhiệt độ



Hình 5: So sánh giữa độ ẩm.



Hình 6: So sánh giữa lượng mưa.

Mô hình đã thể hiện sự cải thiện rõ rệt trong khả năng dự báo, như được thể hiện qua các biểu đồ trên. Cụ thể, các chỉ số đánh giá mô hình cho hai phương pháp được so sánh như sau: Với hình bên trái, áp dụng thuật toán Random Forest trên toàn bộ tập dữ liệu ban đầu, các chỉ số đánh giá lần lượt là: Mean Squared Error (MSE) = 17.29, Mean Absolute Error (MAE) = 2.71 và R^2 Score = 0.56. Ngược lại, hình bên phải, khi áp dụng thuật toán Random Forest trên tập dữ liệu đã được chia, cho thấy các chỉ số được cải thiện đáng kể với MSE = 8.52, MAE = 1.79 và R^2 Score = 0.75. Sự giảm sút của MSE và MAE, cũng như sự gia tăng của R^2 Score cho thấy mô hình đã hoạt động tốt hơn, với khả năng dự báo gần đúng hơn và mức độ giải thích biến thiên của dữ liệu cao hơn. Điều này chứng tỏ rằng việc chia dữ liệu có thể góp phần nâng cao độ chính xác và tính ổn định của mô hình dự báo.

6. Kết luận

Trong nghiên cứu này, nhóm tác giả đã áp dụng thuật toán Random Forest để dự báo ba yếu tố thời tiết chính: nhiệt độ, độ ẩm, và lượng mưa, với trọng tâm đặc biệt vào các hiện tượng mưa cực đoan. Kết quả thực nghiệm cho thấy thuật toán Random Forest đạt hiệu suất dự báo đáng tin cậy, với hệ số tương quan trung bình (R) vượt ngưỡng 0,8361, sai số trung bình tuyệt đối (MAE) dưới 1,1738 và căn bậc hai sai số trung bình (RMSE) dưới 1,6641. Những kết quả này không chỉ khẳng định tính hiệu quả và khả năng ứng dụng cao của Random Forest trong lĩnh vực dự báo thời tiết, mà còn tạo tiền đề cho việc phát triển các hệ thống dự báo thời tiết có độ chính xác cao hơn trong tương lai.

7. Hướng nghiên cứu trong tương lai

Các nghiên cứu tiếp theo sẽ tập trung vào việc phát triển và đánh giá các mô hình phân loại hiệu quả, nhằm phân loại chính xác giữa dữ liệu ngoại lai và dữ liệu bình thường. Mục tiêu là áp dụng các phương pháp phân loại không chỉ cho dữ liệu đầu vào mà còn cho dữ liệu đầu ra, từ đó cải thiện khả năng phát hiện và xử lý các giá trị bất thường. Ngoài ra, việc thử nghiệm và áp dụng các mô hình khác như mạng nơron sâu (deep neural networks), XGBoost hoặc các mô hình chuỗi thời gian tiên tiến như ARIMA, SARIMA, sẽ giúp nâng cao độ chính xác của dự đoán. Những mô hình này có thể được điều chỉnh để dự đoán các yếu tố thời tiết trong vòng 24 giờ tiếp theo, cung cấp cái nhìn chi tiết hơn về xu hướng và biến động thời tiết. Đặc biệt, vấn đề dự đoán các giá trị ngoại lai vẫn cần được cải thiện hơn nữa để tăng cường khả năng nhận diện và xử lý những trường hợp bất thường trong dữ liệu. Việc này không chỉ giúp nâng cao độ chính xác của các mô hình dự đoán mà còn hỗ trợ việc ra quyết định trong các ứng dụng thực tiễn, chẳng hạn như cảnh báo sớm cho các hiện tượng thời tiết cực đoan. Do đó, nghiên cứu và ứng dụng các kỹ thuật mới để cải thiện mô hình dự đoán giá trị ngoại lai là rất cần thiết.

Tài liệu tham khảo

1. AI cách mạng hóa dự báo thời tiết, theo dõi khí hậu và dự đoán mang tính thập niên - Steven Dewitte, Jan P. Cornelis, Richard Müller, Adrian Munteanu - August 2021.

2. AI cho dự báo thời tiết - Silvia Conti - January 2024.

3. Phân tích so sánh các kĩ thuật khai phá dữ liệu cho việc dự báo lượng mưa ở Malaysia - Suhaila Zainudin, Dalia Sami Jasim, Azuraliza Abu Bakar - December 2016.

4. Cải thiện độ chính xác của tỉ lệ mưa từ cảm biến vệ tinh quang học cùng với học máy - một phương pháp với nền tảng là những Khu rừng ngẫu nhiên áp dụng vào MSG Seviri - Meike Kühnlein, Tim Appelhans, Boris Thies, Thomas Nauss - February 2014

5. So sánh Khu rừng ngẫu nhiên và máy Vector hỗ trợ (SVM) trong dự báo lượng mưa theo thời gian thực thông qua radar - Pao-Shan Yu, Tao-Chang Yang, Szu-Yin Chen, Chen-Min Kuo, Hung-Wei Tseng - September 2017.

6. Kết hợp Khu rừng ngẫu nhiên và Hồi quy Vector hỗ trợ bình phương tối thiểu trong việc cải thiện Chi tiết hóa dữ liệu Mưa cực đoan - Quoc Bao Pham, Tao-Chang Yang, Chen-Min Kuo, Hung-Wei Tseng, Pao-Shan Yu - March 2019.

7. Ước tính lượng mưa sử dụng dữ liệu vệ tinh Himawari-8 dựa trên mô hình học máy Random Forest - Nguyễn Vinh Thư, Bùi Thị Khánh Hòa, Nguyễn Minh Cường, Hoàng Thị Thanh Thuật, Nguyễn Thị Hoàng Anh - November 2023.

8. Phương pháp rừng ngẫu nhiên có trọng số dựa trên địa hình để ước lượng lượng mưa định lượng từ radar - Xuebing Yang, Qiuming Kuang, Wensheng Zhang, Guoping Zhang - May 2017.

9. Một thuật toán rừng ngẫu nhiên để dự báo ngắn hạn các sự kiện mưa lớn - Saurabh Das, Rohit Chakraborty, Animesh Maitra - March 2017

10. Mô hình dự báo ngắn hạn lượng mưa dựa trên thuật toán rừng ngẫu nhiên - Nita H. Shah, Anupam Priamvada, Bipasha Paul Shukla - July 2023

11. Rainfall Prediction Using Random Forest and Decision Tree Algorithms - Sevierda Raniprima, Nanang Cahyadi, Vivi Monita - JUNE 2024

12. Predicting Rainfall Using Random Forest and CatBoost Model - Sung-Chi Hsu, Alok Kumar Sharma, Radius Tanone and Yan-Tang Ye - April 2024

13. Nghiên cứu công nghệ dự báo mưa AI thí điểm tại TP. Hồ Chí Minh - Phạm Thanh Long, Lê Văn Phận, Nguyễn Phương Đông1, Lê Hồng Dương, TrầnTuấn Hoàng - 2022

14. Chollet, F. Deep Learning with Python, Public: Manning Publications, 2017

15. Nguyễn Đình Thúc, N.Đ. Trí Tuệ Nhân Tạo Mạng Nơron - Phương Pháp Và Ứng Dụng. NXB GiáoDục, 2000.