TẠP CHÍ
KHÍ TƯỢNG THỦY VĂN

Bài báo khoa học

Xây dựng mô hình mô phỏng chỉ số chất lượng không khí cho thành phố Hồ Chí Minh bằng phương pháp học máy

Nguyễn Phúc Hiếu[1], Đỗ Dương Hoàng Vân[1], Đào Nguyên Khôi[1]*

[1] Khoa Môi trường, Trường ĐH Khoa học tự nhiên, ĐHQG-HCM;
phuchieu50@gmail.com; vanhoanggg25@gmail.com; dnkhoi@hcmus.edu.vn

*Tác giả liên hệ: dnkhoi@hcmus.edu.vn; Tel.: +84–088304379

Ban Biên tập nhận bài: 10/12/2023; Ngày phản biện xong: 11/1/2024; Ngày đăng bài: 25/5/2024

Tóm tắt: Nghiên cứu áp dụng các mô hình học máy bao gồm MLP (*Multi-layer Perceptron*), RF (*Random Forest*) và SVR (*Support Vector Regression*) để dự báo chỉ số chất lượng không khí tại Tp. Hồ Chí Minh. Dữ liệu đầu vào bao gồm chỉ số chất lượng không khí AQI (*Air Quality Index*) và 5 biến khí tượng (điểm sương, áp suất, nhiệt độ, độ ẩm, tốc độ gió) từ tháng 3/2019 đến tháng 6/2021, với 70% dữ liệu đầu vào được sử dụng cho giai đoạn huấn luyện và 30% dữ liệu còn lại sử dụng cho giai đoạn kiểm tra. Thông qua phân tích tương quan và phân tích tự tương quan một phần, 6 kịch bản với các thông số đầu vào khác nhau được xây dựng để mô phỏng chỉ số AQI. Kết quả cho thấy cả 3 mô hình đều có hiệu suất dự báo tốt ở cả 6 kịch bản. Trong đó, mô hình MLP với 5 thông số đầu vào (MLP-K5) cho hiệu quả dự báo tốt nhất với MSE = 0,0045, R^2 = 0,89, NSE = 0,886. Đối với mô hình SVR, mô hình SVR với 6 thông số đầu vào (SVR-K6) cho kết quả dự báo tốt nhất với MSE = 0,0048, R^2 = 0,88, NSE = 0,879. Đối với mô hình RF, mô hình RF với 6 thông số đầu vào (RF-K6) cho kết quả dự báo tốt nhất với MSE = 0,005, R^2 = 0,88, NSE = 0,875. Kết quả cho thấy, mô hình MLP có khả năng mô phỏng tốt chỉ số chất lượng không khí cho thành phố Hồ Chí Minh.

Từ khóa: Chỉ số chất lượng không khí; MLP; SVR; RF; Phương pháp học máy.

1. Đặt vấn đề

Hiện nay, thế giới ngày càng phát triển do quá trình công nghiệp hóa, đô thị hóa diễn ra mạnh mẽ. Đi đôi với sự phát triển là mối lo ngại lớn về vấn đề ô nhiễm không khí. Ô nhiễm không khí là chủ đề có tầm quan trọng cao, các vấn đề toàn cầu đã chứng minh rằng tác động gây hại của nó đến sức khỏe thể chất con người và hệ sinh thái [1]. Chất lượng không khí là tiêu chí quan trọng để đánh giá mức độ ô nhiễm của một khu vực. Chất lượng không khí xấu là 1 trong 5 nguy cơ lớn gây hại cho sức khỏe trên thế giới, ví dụ như tiếp xúc lâu dài với không khí ô nhiễm liên quan đến bệnh nhiễm trùng đường hô hấp, đau tim, đột quỵ và ung thư phổi [2]. Theo báo cáo của Ngân hàng Thế giới năm 2022 [3], nồng độ $PM_{2.5}$ trung bình hàng năm tại Việt Nam luôn cao hơn từ 4 đến 5 lần so với ngưỡng an toàn của Tổ chức Y tế thế giới là 10 ($\mu g/m^3$), và năm 2016 ước tính có hơn 60.000 ca tử vong có liên quan đến ô nhiễm không khí và theo Liên minh Toàn cầu về Sức khỏe và Ô nhiễm ước tính số ca tử vong lên hơn 50.000 vào năm 2019. Bênh cạnh đó, Hội đồng cố vấn kinh tế ước tính thiệt hại kinh tế của ô nhiễm không khí sẽ bằng 1% GDP vào năm 2020, dựa trên chi phí tiền tệ liên quan đến việc gia tăng tỷ lệ mắc bệnh và tổn thất lực lượng lao động.

Thành phố Hồ Chí Minh (TP.HCM) là trung tâm kinh tế của khu vực Nam bộ và cả nước, nằm ở ngã tư quốc tế giữa các con đường hàng hải từ Bắc xuống Nam, từ Tây sang Đông, là tâm điểm của khu vực Đông Nam Á. Đây là đầu mối giao thông nối liền các tỉnh

trong vùng, và nằm trong vùng chuyển tiếp giữa miền Đông Nam Bộ và đồng bằng sông Cửu Long (Hình 1) [4]. Theo Sở Giao thông vận tải TP.HCM, tính đến năm 2019 thành phố có khoảng 8,7 triệu phương tiện giao thông đang hoạt động, cùng với bụi thải từ hoạt động xây dựng trên địa bàn đã và đang làm trầm trọng thêm tình trạng ô nhiễm không khí trong thành phố. Theo đà phát triển của kinh tế, lượng dân cư ngày càng tăng. Tổng dân số của TP.HCM năm 2022 là hơn 9,389 triệu người, chiếm gần 9,44% dân số cả nước và 49,92% dân số vùng Đông Nam Bộ [5]. Dân số ngày càng đông cũng là một trong những nguyên nhân gây nên những vấn đề về môi trường mà trong đó phải kể đến là vấn đề về ô nhiễm không khí. Đây cũng là một trong những vấn đề bức thiết và được quan tâm hiện này tại khu vực, vì vậy TP.HCM được lựa chọn là khu vực nghiên cứu. Dự báo ô nhiễm không khí rất quan trọng đối với sự can thiệp sức khỏe cộng đồng và hoạch định chính sách kiểm soát ô nhiễm không khí [6]. Vì vậy, việc sử dụng mô hình hóa trong dự báo chất lượng không khí để kiểm soát ô nhiễm trở nên phổ biến tại nhiều quốc gia trên thế giới. Có nhiều phương pháp khác nhau để dự báo chất lượng không khí như mô hình thống kê, mô hình vật lý,…Trong đó, học máy đã được ứng dụng rộng rãi để dự báo chỉ số chất lượng không khí vì các ưu điểm như tính đơn giản, tính chính xác của kết quả dự báo, mô hình hóa được các mối quan hệ phức tạp và phi tuyến giữa một tập hợp dữ liệu đầu vào và mục tiêu, xử lý được bộ dữ liệu lớn. Do đó, sử dụng phương pháp học máy trong dự báo chất lượng không khí được các nhà nghiên cứu trong và ngoài nước quan tâm đến nhiều hơn, đặc biệt là trong sự bùng nổ phát triển của công nghệ 4.0.

Một số nghiên cứu cho thấy được tính ứng dụng của học máy với khả năng dự báo chất lượng không khí [7–10]. Điển hình như nghiên cứu của Mehdi và cộng sự năm 2019 sử dụng mô hình RF (*Random Forest*), XGB (*XGBoost*) và DL (*Deep Learning*) dự báo nồng độ $PM_{2.5}$ tại thủ đô Tehran, Iran; nhìn chung kết quả cho thấy các mô hình có hiệu suất khá tốt với chỉ số R^2 trên 0,63, MAE trong khoảng 10-11,15 và RMSE trong khoảng 13,62-15,89 [11]. Nghiên cứu của Mauro Castelli và cộng sự năm 2020 đã sử dụng mô hình SVR (*Support Vector Regression*) kết hợp RBF (*Radial Basis Function*) để dự báo CO, O_3, SO_2, NO_2, $PM_{2.5}$, nhìn chung, kết quả dự báo đa số đều tốt. Chỉ số R^2 dao động trong khoảng 0,77-0,99, MAE dao động trong khoảng 0,08-0,46 [12]. Nghiên cứu của Doreswamy và cộng sự năm 2020 sử dụng mô hình MLP (*Multi-layer Perceptron*), RF, DT (*Decision Tree*), GB (*Gradient Boosting*) để dự báo nồng độ $PM_{2.5}$ tại huyện Bình Đông, Đài Loan với dữ liệu đầu vào gồm nồng độ CO, NO_2, SO_2, CO_2, PM_{10} và tốc độ gió, nhiệt độ. Kết quả dự báo tốt với $R^2 > 0,79$, MSE < 6,28, MAE < 0,04, RMSE < 0,17 [13]. Nghiên cứu của Mạc Duy Hưng và cộng sự năm 2017 đã ứng dụng mạng nơ-ron nhân tạo để xây dựng mô hình dự báo nồng độ SO_2 cực đại ngày cho thành phố Hà Nội. Kết quả cho thấy, ANN là có triển vọng để xây dựng mô hình dự báo thống kê chất lượng không khí, với giá trị của RMSE, RMSE và MAE lần lượt là 11,7%, 3,28 và 2,58 [14]. Một nghiên cứu khác của Nguyễn Thị Thu Phương và cộng sự năm 2020 sử dụng sử dụng mô hình MLP và SVM (*Support Vector Machine*) dự báo nồng độ O_3 đối lưu hàng giờ tại tỉnh Quảng Ninh. Nhìn chung, mô hình SVM có hiệu suất tốt hơn so với MLP, đặc biệt là trong các tình huống dao động lớn và nồng độ ozone cao [15].

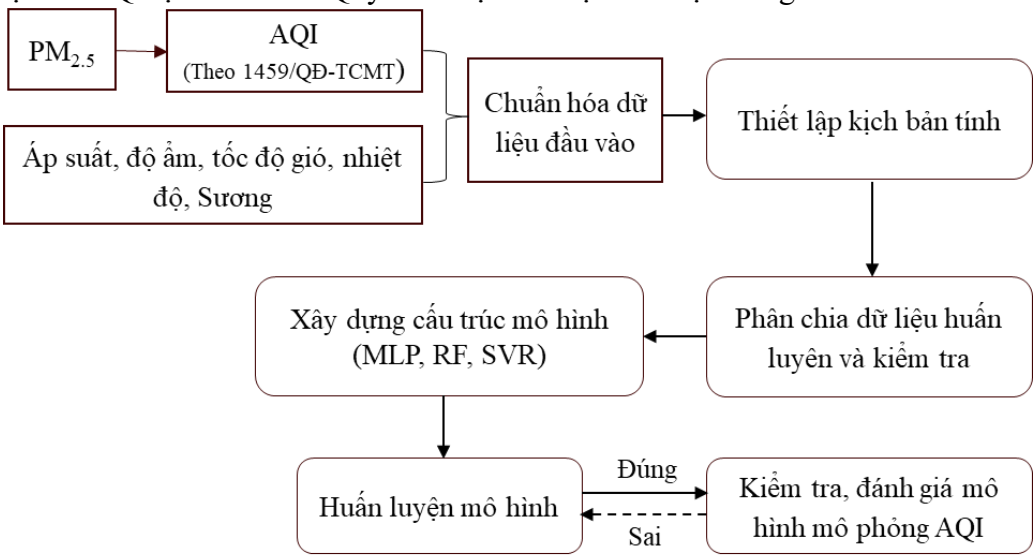
Có thể thấy, các mô hình học máy được áp dụng phổ biến và có hiệu quả trong việc mô phỏng và dự báo chất lượng không khí ở các khu vực ngoài nước. Tuy nhiên, các nghiên cứu về ứng dụng các mô hình học máy trong mô phỏng chất lượng không khí ở nước ta vẫn còn hạn chế, đặc biệt là tại TP.HCM vẫn chưa có nghiên cứu nào thực hiện. Vì vậy, nghiên cứu sẽ áp dụng các mô hình học máy để mô phỏng chỉ số chất lượng không khí tại khu vực TP.HCM.

Mục tiêu chính của nghiên cứu là dự báo AQI tại TP.HCM dựa trên các mô hình học máy với các mục tiêu cụ thể bao gồm: (1) Xây dựng các thuật toán học máy MLP, RF và SVR bằng ngôn ngữ python; (2) Dự báo AQI tại TP.HCM dựa trên các mô hình học máy đã xây dựng.

Hình 1. Khu vực nghiên cứu.

2. Phương pháp nghiên cứu

Để đạt được các mục tiêu đề ra, nghiên cứu tiến hành thực hiện các bước sau: (1) Thu thập dữ liệu từ trạm Tổng lãnh sự quán Hoa Kỳ tại TP.HCM bao gồm áp suất, độ ẩm, tốc độ gió, nhiệt độ, sương, $PM_{2.5}$; (2) Tính toán AQI theo Quyết định số 1459/QĐ-TCMT của Tổng cục môi trường năm 2019; (3) Chuẩn hóa dữ liệu đầu vào; (4) Tính tương quan giữa các biến và thiết lập kịch bản tính toán; (5) Phân chia dữ liệu phục vụ quá trình huấn luyện và kiểm tra mô hình; (6) Xây dựng 03 mô hình học máy MLP, RF, SVR; (7) Huấn luyện, kiểm tra mô hình dự báo AQI tại TP.HCM. Quy trình cụ thể được thể hiện trong Hình 2.

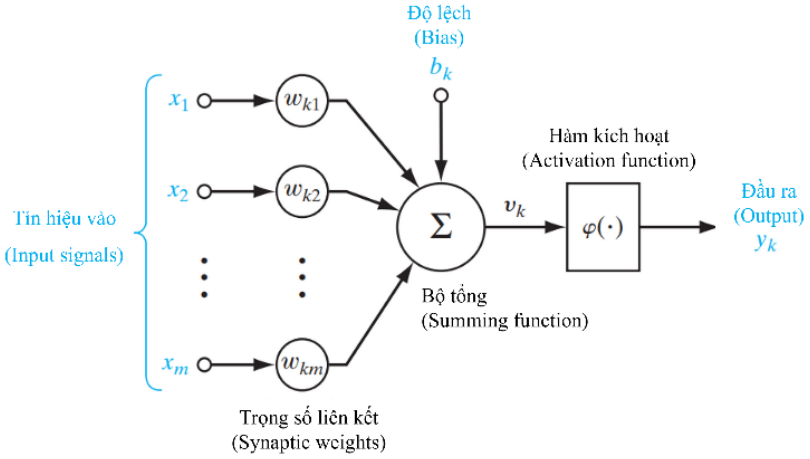


Hình 2. Khung nghiên cứu.

2.1. Mô hình học máy

2.2.1. Mô hình Multilayer Perceptron (MLP)

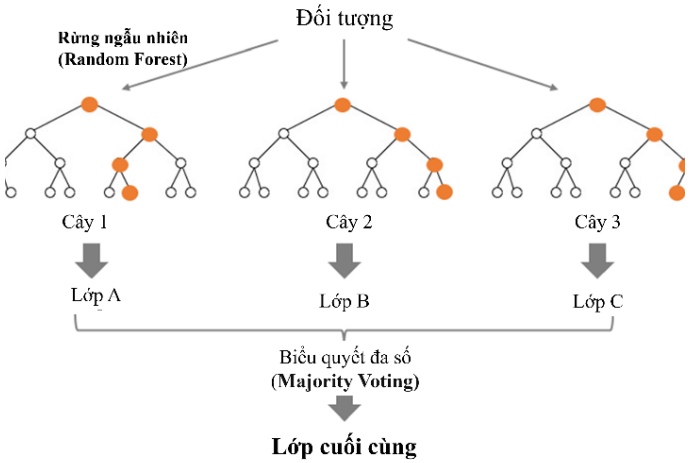
Mạng nơ-ron nhân tạo là một trong những thuật toán học máy phổ biến hiện nay. Nó được mô phỏng dựa trên sự hoạt động của các tế bào thần kinh của con người. MLP thuộc một trong những mạng nơ-ron phổ biến hiện nay. MLP gồm hệ thống các nơ-ron đơn giản được kết nối với nhau, là một mô hình đại diện cho một ánh xạ phi tuyến giữa một vectơ đầu vào và một vectơ đầu ra. Các node kết nối với nhau bằng trọng số và tín hiệu đầu ra, là một hàm của tổng các đầu vào cho node được sửa đổi bằng một hàm truyền phi tuyến đơn giản hay còn gọi là hàm kích hoạt. Kiến trúc của MLP có thể thay đổi nhưng nhìn chung sẽ bao gồm một số lớp nơ-ron nhân tạo. Lớp đầu vào không đóng vai trò tính toán mà chỉ đóng vai trò truyền vector đầu vào cho mạng. Một mô hình MLP có thể có một hoặc nhiều lớp ẩn và cuối cùng là một lớp đầu ra [16]. Hình 3 mô tả cấu trúc mô hình MLP.



Hình 3. Cấu trúc mô hình MLP.

2.2.2. Mô hình Random Forest (RF)

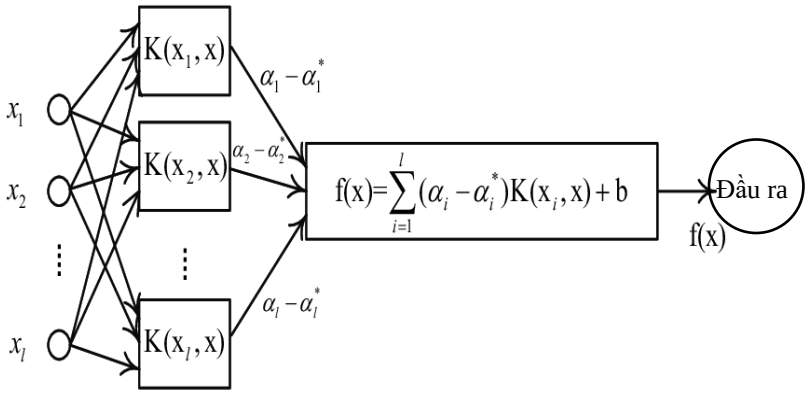
RF là thuật toán học học máy thuộc dạng học có giám sát. RF xây dựng nhiều cây quyết định trên các tập con khác nhau của tập dữ liệu đã cho và lấy giá trị trung bình để cải thiện độ chính xác dự đoán của tập dữ liệu đó. Thay vì dựa vào một cây quyết định, RF lấy dự đoán từ mỗi cây và dựa trên đa số dự đoán từ đó đưa ra kết quả dự đoán cuối cùng. Số lượng cây lớn hơn dẫn đến độ chính xác cao hơn và ngăn ngừa vấn đề overfitting. RF còn bổ sung tính ngẫu nhiên cho mô hình, thay vì tìm kiếm đặc tính quan trọng nhất trong khi tách node, nó sẽ tìm kiếm đặc tính tốt nhất trong số tập hợp con ngẫu nhiên. Hình 4 thể hiện cấu trúc của mô hình.



Hình 4. Mô hình RF.

2.2.3. Mô hình Support Vector Regression (SVR)

Mô hình SVR được xây dựng dựa trên mô hình SVM với lợi thế: số lượng tham số tự do ít hơn, khả năng dự báo tốt hơn và huấn luyện nhanh hơn. Trong SVR, dữ liệu sẽ được ánh xạ vào không gian đặc trưng k-chiều, thông qua ánh xạ phi tuyến đưa mô hình hồi quy tuyến tính phù hợp với các điểm dữ liệu trong không gian này. Sau đó, dữ liệu tuyến tính thu được được sử dụng để dự báo trong không gian đặc trưng mới. Lúc này, ánh xạ từ không gian đầu vào vào không gian đặc trưng mới được xác định bởi hàm kernel. Hình 5 thể hiện mô hình SVR.



Hình 5. Cấu trúc mô hình SVR.

2.3. Thu thập xử lý và phân chia dữ liệu

2.3.1. Thu thập và xử lý số liệu

Dữ liệu được thu thập là các thông số được quan trắc liên tục từ trạm Lãnh sự quán Mỹ từ tháng 3/2019 đến tháng 6/2021, bao gồm 06 thông số như sau: nồng độ $PM_{2.5}$, nhiệt độ, tốc độ gió, độ ẩm không khí, điểm sương, áp suất. Sau khi thu thập dữ liệu nồng độ $PM_{2.5}$ tại điểm quan trắc, giá trị chỉ số chất lượng không khí (AQI) của $PM_{2.5}$ được tính theo hướng dẫn của Quyết định số 1459/QĐ-TCMT năm 2019. Bảng 1 thể hiện đặc trưng thống kê của các dữ liệu thu thập.

Bảng 1. Đặc trưng thống kê của dữ liệu

	Biến	**Min**	**Max**	**Trung bình**	**Độ lệch chuẩn**	**Đơn vị**
Dữ liệu chất lượng không khí	AQI	2,5	213,6	66,8	42,2	
Dữ liệu khí tượng	Nhiệt độ	23	31	27,9	1,5	°C
	Tốc độ gió	0,5	5,9	2,54	1,0	m/s
	Độ ẩm	47	100	77,4	10,7	%
	Điểm sương	14,5	26,5	23,4	2,4	°C
	Áp suất	1003	1014	1009	1,89	mb

Vì dữ liệu thu thập gồm nhiều thuộc tính với các tỷ lệ, đơn vị khác nhau và giá trị của các dữ liệu ban đầu có khoảng chênh lệch lớn nên thuật toán học máy có thể bị ảnh hưởng từ việc này. Do đó, thay đổi tỷ lệ các thuộc tính để tất cả các thuộc tính có chung một tỷ lệ là việc cần thiết. Điều này hữu ích cho các thuật toán có trọng số đầu vào như hồi quy và mạng nơ-ron nhân tạo. Nghiên cứu này sử dụng phương pháp chuẩn hóa MinMaxScaler để chuẩn hóa dữ liệu về khoảng [0;1] theo công thức:

$$x'_i = \frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)} \quad (1)$$

Trong đó: x'_i là giá trị x_i sau khi chuyển hóa; x_i là giá trị ban đầu; $\min(x_i)$ là giá trị nhỏ nhất của biến i; $\max(x_i)$ là giá trị lớn nhất của biến i.

2.3.2. Phân chia dữ liệu

Phân chia dữ liệu cũng là một trong những phần quan trọng trong việc xây dựng mô hình, việc phân chia dữ liệu kém có thể dẫn đến hiệu suất mô hình không chính xác và có nhiều thay đổi. Tập huấn luyện bao gồm các điểm dữ liệu được sử dụng trực tiếp trong việc xây dựng mô hình. Tập kiểm tra gồm các dữ liệu được dùng để đánh giá hiệu quả của mô hình. Để đảm bảo tính phổ quát, dữ liệu kiểm tra không được sử dụng trong quá trình xây dựng mô hình. Điều kiện cần để một mô hình hiệu quả là kết quả đánh giá trên cả tập huấn luyện và tập kiểm tra đều cao [17]. Các phương pháp phân chia dữ liệu có thể được phân loại thành phương pháp có giám sát và phương pháp không có giám sát [18]. Không có quy tắc thống nhất về cách phân chia tập huấn luyện và kiểm tra [19]. Tùy vào dữ liệu hiện có mà trong quá trình xây dựng mô hình sẽ đưa ra được tỷ lệ thích hợp nhất. Qua quá trình thử và sai, nghiên cứu đưa ra tỷ lệ tối ưu cho mô hình là 70:30, 70% dữ liệu sử dụng cho quá trình huấn luyện và 30% dữ liệu sử dụng cho quá trình kiểm tra.

2.4. Xây dựng kịch bản dữ liệu đầu vào cho mô hình

Lựa chọn thông số đầu vào là một trong những bước rất quan trọng trong việc xây dựng mô hình học máy bởi nó ảnh hưởng đến kết quả đầu ra cũng như tính chính xác của mô hình. Trong nghiên cứu này, phân tích tự tương quan từng phần và phân tích tương quan được áp dụng để xây dựng kịch bản dự báo.

2.4.1. Tự tương quan từng phần (PACF)

Biểu đồ tương quan là một cách trực quan để hiển thị mối tương quan nối tiếp trong dữ liệu thay đổi theo thời gian (tức là dữ liệu chuỗi thời gian). Tương quan nối tiếp (còn được gọi là tự tương quan) là trường hợp sai số tại một thời điểm trong thời gian di chuyển đến một điểm tiếp theo trong thời gian. PACF thu được mối tương quan tuyến tính của mỗi giá trị x_t của chuỗi với các giá trị khác ở các độ trễ khác nhau, như x_{t-1}, x_{t-2},… nhưng loại bỏ sự can thiệp của các giá trị khác. Ví dụ, mối tương quan giữa x_t và x_{t-2} có sự giao thoa của x_{t-1}, PACF sẽ loại bỏ sự can thiệp đó [20].

Bảng 2. Kết quả phân tích tự tương quan từng phần của AQI đối với các mức thời gian trễ.

Độ trễ (ngày)	t–1	t–2	t–3	t–4	t–5	t–6	t–7	t–8	t–9	t–10
Hệ số tương quan (r)	0,92	0,19	0,08	0,15	0,09	0,03	0,01	0,11	0,07	0,04

Bảng 2 thể hiện kết quả tự tương quan từng phần của dữ liệu AQI với các mức thời gian trễ khác nhau. Trong bảng kết quả, AQI có mức tương quan cao nhất với AQI (t - 1) với giá trị tương quan là 0,92, các giá trị AQI tại các mức thời gian trễ khác có giá trị tương quan nhỏ hơn 0,19. Do đó, nghiên cứu này sử dụng biến AQI (t - 1) làm biến đầu vào.

2.4.2. Phân tích tương quan

Phân tích tương quan có thể xác định mối liên hệ tuyến tính giữa biến chất lượng không khí và các biến khí tượng. Bảng 3 thể hiện kết quả tương quan theo thứ tự từ cao đến thấp giữa biến AQI và các biến khí tượng trễ 1 ngày. Phân tích cho thấy, điểm sương có mối quan hệ chặt chẽ nhất với AQI (r = -0,403), tương quan thấp nhất là biến nhiệt độ với r = -0,229. Các biến áp suất, độ ẩm, tốc độ gió có mức tương quan lần lượt giảm dần.

Bảng 3. Tương quan giữa AQI (t) và các biến khí tượng (t-1).

	Điểm sương (t-1)	**Áp suất (t-1)**	**Độ ẩm (t-1)**	**Tốc độ gió (t-1)**	**Nhiệt độ (t-1)**
AQI (t)	-0,403	0,371	-0,278	-0,274	-0,229

2.4.3. Xây dựng kịch bản biến đầu vào cho mô hình

Thông qua phân tích PACF và phân tích tương quan giữa các biến khí tượng và biến AQI, 6 kịch bản được xây dựng tương ứng với số lượng biến đầu vào tăng dần. Cụ thể, các kịch bản dự báo được thể hiện trong Bảng 4.

Bảng 4. Kịch bản dự báo AQI của các mô hình MLP, RF và SVR.

STT	Kịch bản	Thông số đầu vào					
1	K1	AQI (t-1)					
2	K2	AQI (t-1)	Điểm sương (t-1)				
3	K3	AQI (t-1)	Điểm sương (t-1)	Áp suất (t-1)			
4	K4	AQI (t-1)	Điểm sương (t-1)	Áp suất (t-1)	Độ ẩm (t-1)		
5	K5	AQI (t-1)	Điểm sương (t-1)	Áp suất (t-1)	Độ ẩm (t-1)	Tốc độ gió (t-1)	
6	K6	AQI (t-1)	Điểm sương (t-1)	Áp suất (t-1)	Độ ẩm (t-1)	Tốc độ gió (t-1)	Nhiệt độ (t-1)

2.5. Đánh giá hiệu suất cho mô hình

Để đánh giá hiệu suất mô hình, nghiên cứu này sử dụng các chỉ số thống kê như hệ số tương quan R^2, hệ số hiệu quả Nash-Sutcliffe (NSE) và sai số bình phương trung bình (MSE). Giá trị của R^2 và NSE càng tiến về gần 1 cho thấy hiệu quả mô hình càng cao, giá trị MSE càng tiến về gần 0 cho thấy sai số càng nhỏ, mức độ dự báo càng tốt. Các chỉ số thống kê có công thức như sau:

Hệ số tương quan R^2

$$R^2 = \left[\frac{\sum_{i=1}^{n}\left(y_i^{obs}-\bar{y}^{obs}\right)\times\left(y_i^{sim}-\bar{y}^{sim}\right)}{\sqrt{\sum_{i=1}^{n}\left(y_i^{obs}-\bar{y}^{obs}\right)^2}\sqrt{\sum_{i=1}^{n}\left(y_i^{sim}-\bar{y}^{sim}\right)^2}}\right]^2 \quad (3)$$

Sai số bình phương trung bình (MSE)

$$MSE = \frac{1}{n}\sum_{i=1}^{N}(y_i^{obs} - y_i^{sim})^2 \quad (4)$$

Chỉ số Nash-Sulcliffe (NSE)

$$NSE = 1 - \frac{\sum_{i=1}^{n}\left(y_i^{obs}-y_i^{sim}\right)^2}{\sum_{i=1}^{n}\left(y_i^{obs}-\bar{y}^{obs}\right)^2} \quad (5)$$

Trong đó y_i^{obs} là giá trị quan trắc thứ i; y_i^{sim} là giá trị dự báo thứ i; $\bar{y}^{obs}$ là giá trị quan trắc trung bình; $\bar{y}^{sim}$ là giá trị dự báo trung bình.

3. Kết quả và thảo luận

3.1. Mô hình MLP

Chỉ số chất lượng không khí được dự báo bằng mô hình MLP theo 6 kịch bản. Kết quả dự báo được thể hiện trong Bảng 5. Qua đó cho thấy mô hình MLP có khả năng dự báo chất lượng không khí tốt. Cụ thể, kết quả giai đoạn kiểm tra đều cao hơn giai đoạn huấn luyện và các giá trị MSE < 0,0064, R^2 > 0,84 và NSE > 0,84. Ngoài ra, trong 6 kịch bản, kịch bản có hiệu quả dự báo cao nhất là kịch bản 5 thông số đầu vào (MLP-K5) với MSE = 0,0052, R^2 = 0,86, NSE = 0,869 cho quá trình huấn luyện và MSE = 0,0045, R^2 = 0,89, NSE = 0,89 cho quá trình kiểm tra. Ngược lại, kịch bản có hiệu quả thống kê thấp nhất là kịch bản 1

(MLP-K1) với MSE = 0,0064, R^2 = 0,84, NSE = 0,84 cho quá trình huấn luyện và MSE = 0,0048, R^2 = 0,88, NSE = 0,88 cho quá trình kiểm tra. Đồ thị so sánh diễn biến chất lượng không khí trong 2 giai đoạn huấn luyện và kiểm tra giữa kết quả từ MLP-K5 và quan trắc được thể hiện trong Hình 6.

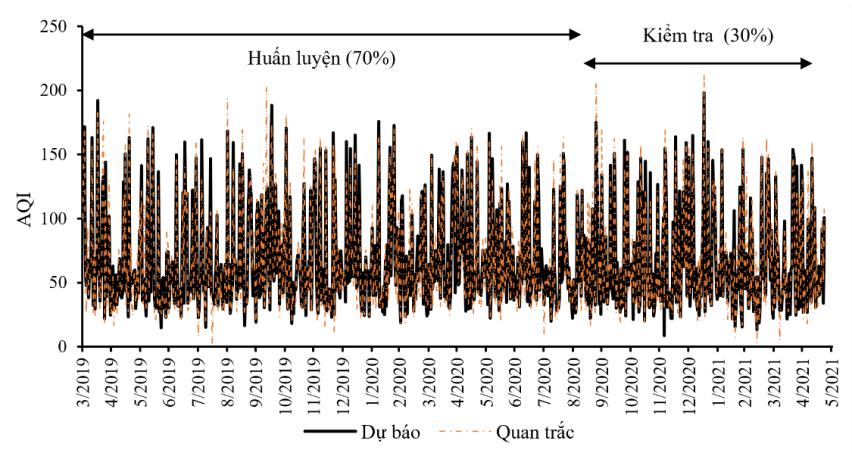
Cấu trúc tổng thể và các tham số của mô hình MLP với 5 thông số đầu vào (MLP-K5) được thể hiện trong Bảng 6. Mô hình MLP với 3 lớp ẩn có số lượng nơ–ron trong mỗi lớp khác nhau. Hàm kích hoạt được chọn sử dụng là hàm ReLU với tốc độ học là 0,001. Hàm tối ưu được sử dụng là Adam.

Bảng 5. Hiệu quả dự báo của mô hình MLP theo 6 kịch bản thông số đầu vào.

Kịch bản	Huấn luyện			Kiểm tra		
	MSE	R^2	NSE	MSE	R^2	NSE
K1	0,0064	0,84	0,84	0,0048	0,88	0,88
K2	0,0061	0,85	0,85	0,0048	0,88	0,88
K3	0,0061	0,85	0,85	0,0048	0,88	0,88
K4	0,0055	0,86	0,86	0,0052	0,87	0,87
K5	0,0052	0,86	0,87	0,0045	0,89	0,89
K6	0,0057	0,86	0,86	0,0045	0,89	0,89

Bảng 6. Cấu trúc và tham số trong mô hình MLP-K5.

	Số lớp ẩn	Loại lớp ẩn	Số nơ–ron
Cấu trúc mô hình	1	Dense	8
	2	Dense	5
	3	Dense	3
	Hàm kích hoạt	ReLU	
	Tốc độ học	0,001	
Tham số mô hình	Hàm tối ưu	Adam	
	Epochs	8000	
	Batch_size	150	



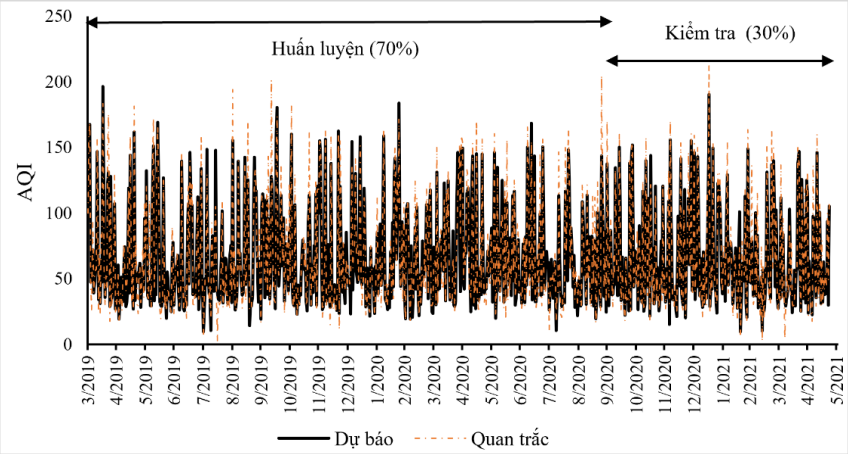
Hình 6. Diễn biến giá trị AQI dự báo và quan trắc của mô hình MLP-K5 trong giai đoạn 2019-2020.

3.2. Mô hình RF

Chỉ số chất lượng không khí được dự báo bằng mô hình RF theo 6 kịch bản. Kết quả dự báo được thể hiện trong Bảng 7. Qua đó cho thấy mô hình RF có khả năng dự báo chất lượng không khí tốt mặc dù kết quả của giai đoạn huấn luyện đều cao hơn giai đoạn kiểm tra. Nhìn chung, các giá trị MSE < 0,004, R^2 > 0,9 và NSE > 0,90 trong giai đoạn huấn luyện, và MSE < 0,006, R^2 > 0,84, và NSE > 0,84 trong giai đoạn kiểm tra. Trong 6 kịch bản, kịch bản có hiệu quả thống kê tốt nhất là kịch bản 6 (RF-K6) với MSE = 0,004, R^2 = 0,9, NSE = 0,90 cho quá trình huấn luyện và MSE = 0,005, R^2 = 0,88, NSE = 0,88 cho quá trình kiểm tra. Đồ thị so sánh diễn biến chất lượng không khí trong 2 giai đoạn huấn luyện và kiểm tra giữa kết quả từ RF-K6 và quan trắc được thể hiện trong Hình 7.

Bảng 7. Hiệu quả dự báo của mô hình RF theo 6 kịch bản thông số đầu vào.

Kịch bản	Huấn luyện			Kiểm tra		
	MSE	R^2	NSE	MSE	R^2	NSE
K1	0,004	0,91	0,91	0,006	0,86	0,86
K2	0,002	0,93	0,93	0,006	0,84	0,84
K3	0,002	0,95	0,95	0,005	0,85	0,85
K4	0,0009	0,98	0,95	0,005	0,85	0,85
K5	0,003	0,92	0,96	0,005	0,87	0,87
K6	0,004	0,90	0,90	0,005	0,88	0,88



Hình 7. Diễn biến giá trị AQI dự báo và quan trắc của mô hình RF-K6 trong giai đoạn 2019-2020.

Cấu trúc và các tham số của mô hình RF được sử dụng cho kịch bản 6 thể hiện trong Bảng 8. Mô hình được xây dựng với các tham số n_estimator là 8000, min_sample_split là 2, min_sample_leaf là 8, max_depth là 50.

Bảng 8. Tham số trong mô hình RF-K6.

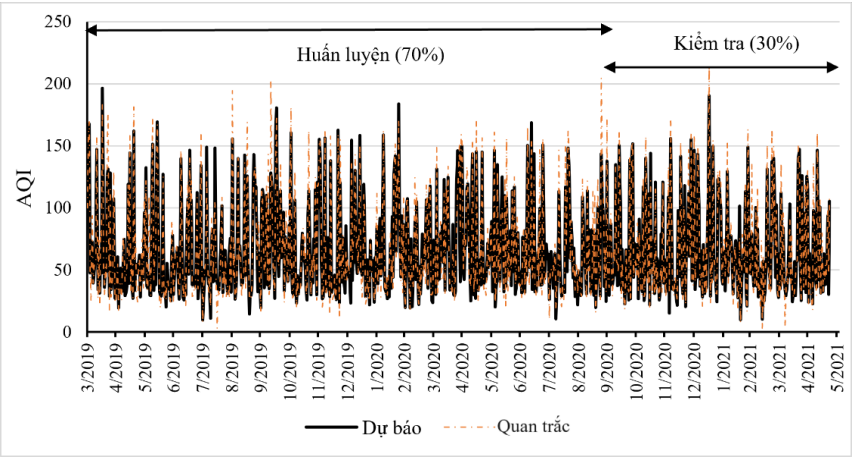
Tham số mô hình	**N_estimators**	**8000**
	Min_sample_split	2
	Min_sample_leaf	8
	Max_depth	50

3.3. Mô hình SVR

Mô hình SVR được sử dụng để dự báo chỉ số chất lượng không khí tại TP.HCM theo 6 kịch bản. Hiệu quả dự báo được thể hiện trong Bảng 9. Qua đó cho thấy hiệu quả dự báo của SVR ở mức tốt. Cũng như mô hình MLP, hiệu quả dự báo giai đoạn kiểm tra đều cao hơn giai đoạn huấn luyện. Cả 6 kịch bản đều cho $R^2 = 0,84$, MSE < 0,0065, NSE = 0,84 cho giai đoạn huấn luyện, và $R^2 = 0,88$, MSE < 0,0051, NSE > 0,87 đối với giai đoạn kiểm tra. Kịch bản cho dự báo tốt nhất là kịch bản 6 (SVR-K6) với MSE = 0,0064, $R^2 = 0,84$, NSE = 0,84 cho giai đoạn huấn luyện, và MSE = 0,0048, $R^2 = 0,88$, NSE = 0,88 cho giai đoạn kiểm tra. Đồ thị so sánh diễn biến chất lượng không khí trong 2 giai đoạn huấn luyện và kiểm tra giữa kết quả từ SVR-K6 và quan trắc được thể hiện trong Hình 8.

Bảng 9. Hiệu quả dự báo của mô hình SVR theo 6 kịch bản thông số đầu vào.

Kịch bản	Huấn luyện			Kiểm tra		
	MSE	R^2	NSE	MSE	R^2	NSE
K1	0,0065	0,84	0,835	0,005	0,88	0,875
K2	0,0065	0,84	0,837	0,0051	0,87	0,872
K3	0,0064	0,84	0,839	0,005	0,88	0,875
K4	0,0064	0,84	0,84	0,0049	0,88	0,877
K5	0,0063	0,84	0,841	0,005	0,88	0,875
K6	0,0064	0,84	0,84	0,0048	0,88	0,879



Hình 8. Diễn biến giá trị AQI dự báo và quan trắc của mô hình SVR-K6 trong giai đoạn.

Cấu trúc và các tham số của mô hình SVR được sử dụng cho kịch bản 6 thể hiện trong Bảng 10. Mô hình được xây dựng với hàm kernel là rbf, hệ số gamma là 0,0005, hệ số epsilon là 0,002, max_iter là 12000.

Bảng 10. Tham số trong mô hình SVR-K6.

	Kernel	**rbf**
Tham số mô hình	C	100
	Gamma	0,0005
	Epsilon	0,002
	Max_iter	12000

4. Kết luận

Nghiên cứu đã thực hiện dự báo chỉ số chất lượng không khí cho thông số $PM_{2.5}$ tại khu vực TP.HCM dựa trên 3 mô hình MLP, RF và SVR. Kết quả cho thấy cả 3 mô hình đều có khả năng dự báo tốt với chỉ số MSE < 0,0065, R^2 > 0,84, NSE > 0,835 trong cả 2 giai đoạn huấn luyện và kiểm tra. Đối với mô hình MLP, mô hình MLP-K5 (kịch bản với 5 thông số đầu vào) cho hiệu suất dự báo tốt nhất. Đối với RF và SVR, mô hình RF-K6 và SVR-K6 cho hiệu suất dự báo tốt nhất với 6 thông số đầu vào. Nhìn chung, mô hình MLP-K5 là mô hình tối ưu nhất vì cho kết quả huấn luyện, kiểm tra tốt nhất, đồng thời sử dụng ít dữ liệu đầu vào nhất tròn cả ba mô hình, cụ thể mô hình MLP với 5 thông số đầu vào là AQI(t-1) và các biến khí tượng trễ 1 ngày là sương, áp suất, độ ẩm, tốc độ gió cho hiệu quả dự báo tốt nhất với MSE = 0,0045, R^2 = 0,89 và NSE = 0,886. Kết quả đạt được trong nghiên cứu này khá tương đồng với kết quả từ nhóm nghiên cứu [14] thực hiện mô phỏng chất lượng không khí tại thành phố Hà Nội, các nghiên cứu đều cho thấy các mô hình dựa trên mạng nơ ron nhân tạo có hiệu quả mô phỏng tốt chất lượng không khí tại khu vực.

Bên cạnh các kết quả đạt được, nghiên cứu vẫn còn một số hạn chế, cụ thể nghiên cứu chỉ thực hiện dự báo tại 1 trạm Tổng lãnh sự quán Mỹ tại TP.HCM, cũng như chỉ tính AQI do $PM_{2.5}$ mà chưa xem xét đến các thông số ô nhiễm khắc như NO_2, SO_2,…Trong nghiên cứu tiếp theo sẽ thực hiện tại nhiều trạm quan trắc hơn cũng như tính toán AQI dựa trên các thông số khác cùng với thông số $PM_{2.5}$. Tuy nhiên, kết quả đạt được trong nghiên cứu đã cho thấy ưu điểm, tiềm năng ứng dụng của các thuật toán học máy trong dự báo chỉ số chất lượng không khí. Thông qua đó, các nghiên cứu tiếp theo sẽ ứng dụng các thuật toán này để dự báo những thông số chất lượng không khí khác hoặc sử dụng các biến đầu vào đa dạng hơn.

Đóng góp tác giả: N.P.H.: phương pháp, tính toán và phân tích kết quả, viết bản thảo; Đỗ D.H.V.: thu thập dữ liệu, tính toán và phân tích kết quả, viết bản thảo; Đ.N.K.: Lên ý tưởng, phương pháp, viết và chỉnh sửa bản thảo.

Lời cảm ơn: Nghiên cứu này được thực hiện dưới sự tài trợ của Sở Khoa Học và Công Nghệ Tp.HCM và được thực hiện bởi Viện Khoa học và Công nghệ Tính toán (ICST) thông qua Hợp đồng thực hiện nhiệm vụ khoa học và công nghệ số 11/2020/HĐ–QPTKHCN ngày 22 tháng 04 năm 2020.

Lời cam đoan: Tập thể tác giả cam đoan bài báo này là công trình nghiên cứu của tập thể tác giả, chưa được công bố ở đâu, không được sao chép từ những nghiên cứu trước đây; không có sự tranh chấp lợi ích trong nhóm tác giả.

Tài liệu tham khảo

1. Nguyen, T.N.T.; Bui, Q.H.; Pham, V.H.; Luu, V.H.; Man, D.C.; Pham, N.H.; Le, T.H.; Nguyen, T.T. Particulate matter concentration mapping from MODIS satellite data: a Vietnamese case study. *Environ. Res. Lett.* **2015**, *10(9)*, 095016.
2. Lelieveld, J.; Poschl, U. Chemists can help to solve the air pollution health crisis. *Nature* **2017**, *551*, 291–293.
3. World Bank. Vietnam - Country climate and development report. Washington DC., Chapter 1: Vietnam's Development Model and Climate Challenges. 2022, pp.9.
4. Sở Tài Nguyên và Môi trường TP.HCM. Báo cáo hiện trạng môi trường TP.HCM năm 2021.
5. Tổng Cục Thống Kê. Kết quả toàn bộ tổng điều tra dân số và nhà ở năm 2022.
6. Gou, Q.; He, Z.; Li, S.; Li, X.; Meng, J.; Hou, Z.; Liu, J.; Chen, Y. Air pollution forecasting using artificial and wavelet neural networks with meteorological conditions. *Aerosol Air Qual. Res*. **2020**, *20*, 1429–1439.
7. Karimian, H.; Li, Q.; Wu, C.; Qi, Y.; Mo, Y.; Chen, G.; Zhang, X.; Sachdeva, S. Evaluation of different machine learning approaches to forecasting $PM_{2.5}$ mass concentrations. *Aerosol Air Qual. Res*. **2019**, *19*, 1400–1410.
8. Doreswamy.; Harishkumar, K.S.; Yogesh, K.M.; Gad, I. Forecasting air pollution particular matter ($PM_{2.5}$) using machine learning regression models. *Procedia Comput. Sci.* **2020**, *171*, 2057–2066.
9. Zaman, N.A.F.K.; Kanniah, K.D.; Kaskaoutis, D.G.; Latif, M.T. Evaluation of machine learning models for estimating $PM_{2.5}$ concentrations across Malaysia. *Appl. Sci.* **2021**, *11(16)*, 7326.
10. Preetham Vignesh, P.; Hiang, J.H.; Kishore, P. Predicting $PM_{2.5}$ concentration across USA using machine learning. *Earth Space Sci.* **2023**, *10(10)*, e2023EA002911.
11. Zamani Joharestani, M.; Cao, C.; Ni, X.; Bashir, B.; Talebiesfandarani, S. $PM_{2.5}$ prediction based on Random Forest, XGBoost and Deep Learning multisource remote sensing data. *Atmos.* **2019**, *10(7)*, 373.
12. Castelli, M.; Clemente, F.M.; Popovic, A.; Silva, S.; Vanneschi, L. A machine learning approach to predict air quality in California. *Complexity* **2020**, 8049504.
13. Doreswamy, Harishkumar, K.S.; Yogesh, K.M.; Ibrahim, G. Forecasting air pollution particulate matter ($PM_{2.5}$) using machine learning regression models. *Procedia Comput. Sci*. **2020**, *171*, 2057-2066.
14. Hưng, M.D.; Dũng, N.T.; Cơ, H.X. Nghiên cứu ứng dụng mạng nơron nhân tạo để xây dựng mô hình dự báo nồng độ SO_2 cực đại ngày. *Tạp chí Khoa Học & Công Nghệ Đại học Thái Nguyên* **2017**, *166(06)*, 127–132.
15. Phuong, N.T.T.; Hung, M.D.; Nam, D.T.; Dung, N.T. Forecast of hourly tropospheric ozone concentration in Quang Ninh using MLP and SVM. *J. Sci: Earth Env. Sci*. **2020**, *36(3)*, 46–54.
16. Gardner, M.W.; Dorling, S.R. Artificial neural networks (the multilayer perceptron) - A review of applications in the atmosphere sciences. *Atmos. Environ*. **1998**, *32(14)*, 2627–2636.
17. Tiệp, V.H. Machine Learning cơ bản. https://machinelearningcoban.com/.

18. Cabaneros, S.M.; Calautit, J.K.; Hughes, B.R. A review of artificial neural network models for ambient air pollution prediction. *Environ. Modell. Software* **2019**, *119*, 285–304.
19. Chen, Y.; Song, L.; Liu, Y.; Yang, L.; Li, D. A review of the artificial neural network models for water quality prediction. *Appl. Sci*. **2020**, *10(17)*, 5776.
20. Flores, J.H.F.; Engel, P.M.; Pinto, R.C. Autocorrelation and partial autocorrelation functions to improve neural networks models on univarite time series forecasting. Proceeding of the 2012 International Joint Conference on Neural Networks, 2012, pp. 1–8.

Developing model for simualtion of air quality index in Ho Chi Minh City using machine learning algorithms

Nguyen Phuc Hieu[1], Do Duong Hoang Van[1], Dao Nguyen Khoi[1]*

[1] Falcuty of Environment, University of Science, VNU-HCM; phuchieu50@gmail.com; vanhoanggg25@gmail.com; dnkhoi@hcmus.edu.vn

Abstract: The aim of this study is to predict the Air Quality Index (AQI) in Ho Chi Minh City using three machine learning algorithms: Multilayer Perceptron (MLP), Random Forest (RF), and Support Vector Regression (SVR). The input data comprise AQI and five meteorological variables (dew point, pressure, temperature, humidity, and wind speed) recorded from March 2019 to June 2021. Seventy percent of the input data is used for the training phase, while the remaining 30% is applied for the testing phase. Correlation analysis and partial autocorrelation analysis were employed to develop six scenarios of input variables. The results indicate that all three machine learning models exhibit robust performance in predicting AQI for Ho Chi Minh City. Notably, the MLP model demonstrates superior predictive capabilities, particularly in the scenario incorporating five input variables (MLP-K5), yielding the best results with MSE = 0.0045, R^2 = 0.89, and NSE = 0.886. The SVR model, under the scenario of six input variables (SVR-K6), achieves optimal performance in AQI prediction with MSE = 0.0048, R^2 = 0.88, and NSE = 0.879. Similarly, the RF model, in the scenario utilizing six input variables (RF-K6), yields the most accurate prediction results with MSE = 0.005, R^2 = 0.88, and NSE = 0.875. In conclusion, the findings evince the proficiency of the Multilayer Perceptron model (MLP) in accurately simulating the Air Quality Index for Ho Chi Minh City.

Keywords: Air Quality Index; Mutilayer Perceptron; Support Vector Regression; Random Forest; Machine learning.