ỨNG DỤNG MÔ HÌNH HỌC MÁY DỰ BÁO CHẤT LƯỢNG NƯỚC DƯỚI ĐẤT: ĐIỂN HÌNH TẠI KHU VỰC THÀNH PHỐ HỘI AN, TỈNH QUẢNG NAM

APPLICATION OF MACHINE LEARNING MODELS IN UNDERGROUND WATER PREDICTION: A CASE STUDY IN HOIAN CITY, QUANGNAM PROVINCE

Lê Phước Cường*, Ngô Viết Thắng

Trường Đại học Bách khoa - Đại học Đà Nẵng[1]

*Tác giả liên hệ: lpcuong@dut.udn.vn

(Nhận bài: 10/02/2022; Chấp nhận đăng: 28/3/2022)

Tóm tắt - Bài báo nghiên cứu dự báo chất lượng nước dưới đất khu vực lân cận bãi rác Cẩm Hà, Tp. Hội An, Quảng Nam bằng các mô hình học máy. Nghiên cứu đã tiến hành phân tích bộ dữ liệu về chất lượng nước dưới đất trong mùa mưa và mùa khô. Bộ dữ liệu với 268 dòng, gồm 8 biến đầu vào (Fe, As, Mo, Co, Ni, Al, Zn, Pb) và 1 biến đầu ra (GWQI). Các tác giả đã nghiên cứu xác định mô hình dự báo tối ưu dựa vào các giá trị sai số tuyệt đối trung bình (MAE), sai số toàn phương trung bình (RMSE) và R^2. Ngôn ngữ R được dùng để tối ưu hoá các mô hình hồi quy tuyến tính (LR), rừng ngẫu nhiên (RF), máy hỗ trợ vec-tơ (SVM), K- điểm dữ liệu gần nhất (KNN), mạng lập thể (Cubist) với tỉ lệ "Huấn luyện":"Kiểm tra" từ 70:30 đến 85:15. Kết quả thu được cho thấy, mô hình Cubist ở tỷ lệ 70:30 là tối ưu nhất cho bộ dữ liệu tại khu vực lân cận bãi rác Cẩm Hà với độ tin cậy R^2 lần lượt là 98,8% và 96%.

Từ khóa - Học máy; nước dưới đất; Groundwater Quality Index (GWQI); bãi rác Cẩm Hà.

Abstract - This article studies to predict groundwater quality in the vicinity of Cam Ha landfill, Hoi An city, Quang Nam province by machine learning models. The study analyzed dataset on groundwater quality in rainy and dry seasons. Dataset with 268 lines, including 8 input variables (Fe, As, Mo, Co, Ni, Al, Zn, Pb) and 1 output oneis the groundwater quality index (GWQI). Authors determined the optimal forecasting model based on the mean absolute error (MAE), root mean square error (RMSE) and R^2. R language was used in order to optimize machine learning models, such as: linear regression (LR), random forest (RF), support vector machine (SVM), K-nearest neighbors (KNN), Cubist with Train:Test ratio from 70:30 to 85:15. The obtained results show that the Cubist model at the ratio 70:30 is the most optimal one for the dataset in the vicinity of Cam Ha landfill with the R^2 value of 98.8% and 96 %, respectively.

Key words - Machine learning; groundwater; Groundwater Quality Index (GWQI); Cam Ha landfill

1. Đặt vấn đề

Đất, nước dưới đất là nguồn tài nguyên vô cùng quý giá, đóng vai trò quyết định cho sự tồn tại và phát triển của con người. Hiện nay, vấn đề ô nhiễm nguồn nước đang là chủ đề nóng trên toàn cầu nói chung và Việt Nam nói riêng. Nguyên nhân gây nên tình trạng ô nhiễm môi trường nước như hiện nay chủ yếu từ ý thức của số ít người dân, các doanh nghiệp thiếu trách nhiệm và cả những bất cập, hạn chế trong công tác quản lý, bảo vệ môi trường [1].

Điển hình tại khu vực bãi rác Cẩm Hà nằm trên địa bàn thôn Bầu Ốc Thượng, xã Cẩm Hà, thành phố Hội An có diện tích khoảng 1,3 hectares (ha), sức chứa 100.000 m^3 là nơi tập trung rác thải của toàn thành phố. Hơn 40 năm tồn tại, giờ đây bãi rác đã trở nên quá tải, có thể hình dung bãi rác như một ngọn núi khổng lồ cao ngất chứa chất hàng nghìn thứ rác thải hỗn tạp chưa qua xử lý đổ về. Do đây là bãi rác tạm thời, không đảm bảo các điều kiện vệ sinh môi trường khiến cho cả khu vực nồng nặc một thứ mùi hôi thối bốc lên gây ảnh hưởng nghiêm trọng đến sức khỏe, đời sống của người dân, nguy cơ về ô nhiễm nguồn nước dưới đất là rất lớn. Chính vì vậy, việc dự báo chất lượng môi trường nước dưới đất có một ý nghĩa hết sức quan trọng trong quá trình phát triển chung của thành phố.

Hiện nay, việc ứng dụng học máy (machine learning) để đưa ra mô hình dự báo về chất lượng nước dưới đất đã được triển khai bởi một số nghiên cứu [2], [3], [4], nhưng đây là vấn đề còn khá mới mẻ ở nước ta. Bên cạnh đó, việc sử dụng chỉ số chất lượng nước dưới đất (GWQI) như một giá trị có khả năng cung cấp sự ảnh hưởng tổng hợp của từng thông số chất lượng trên toàn bộ chất lượng nước đã hỗ trợ tích cực cho việc dự báo chất lượng nước dưới đất trong thời gian qua [3], [4], [5], [6], [7], [8], [9], [10], [11]. Trong nghiên cứu này, tác giả đã phân tích một số mô hình có khả năng dự báo chất lượng nước dưới đất, bao gồm: Hồi quy tuyến tính (LR-Linear Regression), rừng ngẫu nhiên (RF - Random Forest), máy hỗ trợ véc-tơ (SVM - Support vector machine), K điểm dữ liệu gần nhất (KNN - K nearest neighbor), mạng lập thể (Cubist). Từ các giá trị sai số tuyệt đối trung bình (MAE), sai số toàn phương trung bình (RMSE) và R^2 tác giả đã đề xuất mô hình dự báo chất lượng nước dưới đất có độ chính xác cao nhất phù hợp với khu vực nghiên cứu tại thành phố Hội An, tỉnh Quảng Nam.

2. Đối tượng và phương pháp nghiên cứu

2.1. Đối tượng

Nghiên cứu này đã sử dụng bộ cơ sở dữ liệu chất lượng nước dưới đất gồm 268 dòng với 9 biến, trong đó có 8 biến đầu vào (Fe, As, Mo, Co, Ni, Al, Zn, Pb) và 1 biến đầu ra (GWQI), nước dưới đất được lấy mẫu tại khu vực lân cận bãi rác Cẩm Hà, xã Cẩm Hà, thành phố Hội An, tỉnh Quảng Nam (Hình 1). Tất cả các mẫu nước dưới đất của khu vực nghiên cứu được lấy trong cả hai mùa mưa (07/2021-12/2021) và mùa khô (1/2021-06/2021).

[1] The University of Danang - University of Science and Technology (Phuoc-Cuong Le, Viet-Thang Ngo)

Hình 1. *Các vị trí lấy mẫu nước dưới đất tại khu vực lân cận bãi rác Cẩm Hà, thành phố Hội An, tỉnh Quảng Nam*

2.2. Phương pháp phân tích hoá địa

Tổng cộng 268 mẫu nước dưới đất của tầng chứa nước Holocene được đem đi thực hiện phân tích các thông số hoá lý. Các mẫu nước dưới đất được lấy bằng phương pháp khoan sâu dưới lòng đất tại các khu vực nghiên cứu ở các độ sâu thích hợp, tuỳ từng địa điểm mà có độ sâu lấy mẫu dao động từ 10m đến 20m. Quá trình lấy mẫu được thực hiện tuân theo quy trình và các khuyến nghị của quy chuẩn kỹ thuật quốc gia về chất lượng nước ngầm do Bộ tài nguyên môi trường Việt Nam ban hành QCVN 09-MT:2015/BTNMT. Theo đó, quy trình lấy mẫu nước dưới đất đảm bảo các tiêu chuẩn yêu cầu của TCVN 6663-1:2011, ISO 5667-1:2006; TCVN 6663-3:2008, ISO 5667-3:2003; TCVN 6663-11:2011, ISO 5667-11:2009. 268 mẫu nước dưới đất tại các khu vực nghiên cứu được lấy trong suốt mùa mưa (07/2021-12/2021) và mùa khô (1/2021-06/2021).

Độ dẫn điện, pH và nhiệt độ các mẫu nước dưới đất được đo bằng thiết bị đo di động HANA EC-HI8733 và thiết bị AZ pH-8601. Các mẫu nước được thu thập và bảo quản trong các lọ polyetylen (đã được rửa qua bằng chính các mẫu nước ngầm đó) trước khi được phân tích các nguyên tố vi lượng và đa lượng. Các mẫu nước dưới đất trước khi phân tích được axit hoá bằng axit HNO_3 đạt chuẩn phân tích, 65% (Merck, Đức) đến độ pH trong khoảng 1-2. Các mẫu được giữ ổn định ở nhiệt độ phòng cho đến khi được đem đi phân tích các nguyên tố vi lượng, đa lượng.

Nồng độ các nguyên tố vết như As, Mo, Co, Ni, Al, Zn và Pb được xác định bằng phương pháp quang phổ khối kết hợp cao tần cảm ứng (ICP-MS). Nguyên tố Fe được xác định nồng độ bằng phương pháp quang phổ hấp thụ nguyên tử (AAS). Các thí nghiệm phân tích hàm lượng kim loại được thực hiện tại Phân viện Bảo hộ và An toàn lao động miền Trung, Đà Nẵng và tại Trung tâm Nghiên cứu Bảo vệ Môi trường, Trường Đại học Bách khoa - Đại học Đà Nẵng. Tất cả các phương pháp phân tích các thông số hoá lý đều tuân theo các quy chuẩn của QCVN 09-MT:2015/BTNMT. Việc đảm bảo chất lượng/kiểm soát chất lượng (QA/QC) được thực hiện bởi các chuyên gia có chuyên môn sâu về phân tích hoá học của phòng thí nghiệm, bao gồm việc phân tích các mẫu trắng, phân tích lặp lại/mẫu và kiểm soát các chứng nhận chất lượng hoá chất phân tích.

2.3. Phương pháp học máy

Trong bài báo này, tác giả trình bày nghiên cứu trên 5 mô hình học máy (LR, RF, SVM, KNN và Cubist) để dự báo chất lượng nước dưới đất thông qua chỉ số GWQI. Dữ liệu để thực hiện được mô hình học máy cần đủ lớn, có độ tin cậy cao về các thông số nước dưới đất.

LR - Linear Regression (Hồi quy tuyến tính)

Trong thống kê, hồi quy tuyến tính là một cách tiếp cận tuyến tính để mô hình hóa mối quan hệ giữa một phản ứng vô hướng và một hoặc nhiều biến giải thích (còn được gọi là các biến phụ thuộc và độc lập). Giống như tất cả các hình thức phân tích hồi quy, hồi quy tuyến tính tập trung vào phân phối xác suất có điều kiện của phản hồi cho các giá trị của các yếu tố dự đoán, thay vì phân phối xác suất chung của tất cả các biến này, là lĩnh vực của phân tích đa biến.

RF - Random Forest (Rừng ngẫu nhiên)

RF là sự kết hợp của cây dự đoán, mỗi cây phụ thuộc vào giá trị của vector ngẫu nhiên được lấy mẫu độc lập (Independently) và với sự phân bố như nhau cho tất cả các cây có trong rừng. RF là một loại của thuật toán tổng hợp (Ensemble) được gọi là tổng hợp (aggregation) bootstrap và là một trong những phương pháp học máy phổ biến nhất.

SVM - Support vector machine (Máy hỗ trợ vector)

SVM là một trong những thuật toán phân lớp phổ biến và hiệu quả. SVM là một khái niệm trong thống kê và khoa học máy tính cho một tập hợp các phương pháp học có giám sát liên quan đến nhau để phân loại và phân tích hồi quy. SVM dạng chuẩn nhận dữ liệu vào và phân loại chúng vào hai lớp khác nhau. Do đó, SVM là một thuật toán phân loại nhị phân.

KNN - K Nearest neighbors (K- Điểm dữ liệu gần nhất)

KNN là thuật toán phân cụm (Clustering), là kỹ thuật học có giám sát sử dụng để phân loại (Classify) các điểm dữ liệu mới dựa trên vị trí (Position) của chúng trên điểm dữ liệu gần nhất. KNN dự đoán 1 mẫu mới sử dụng mẫu K- điểm dữ liệu gần nhất từ tập huấn luyện.

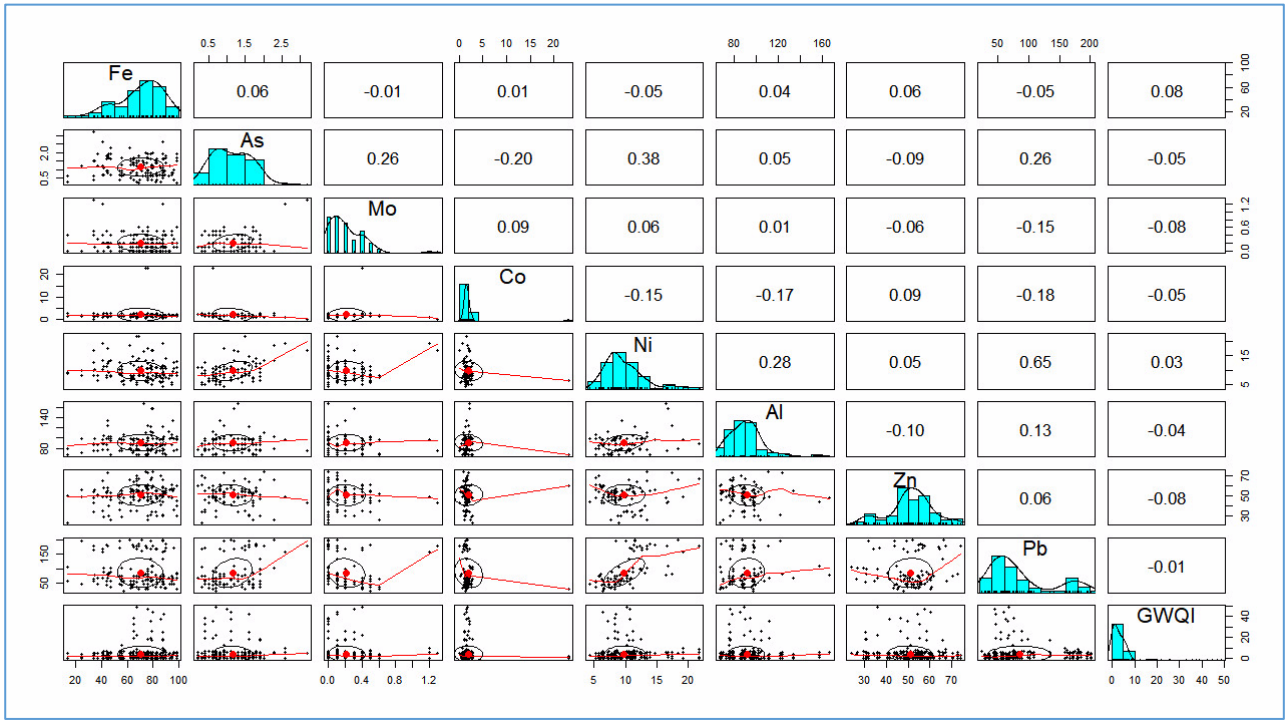
Cubist (Mạng lập thể)

Cubist là một thuật toán dựa trên các nguyên tắc (rules) được sử dụng để xây dựng các mô hình dự báo dựa trên việc phân tích dữ liệu đầu vào. Nó được phát triển dựa trên sự mở rộng của mô hình cây quyết định với khả năng xử lý lên tới hàng nghìn biến đầu vào.

Tác giả đã sử dụng ngôn ngữ R để thao tác các thuật toán của 5 mô hình học máy trên nhằm chọn ra mô hình học máy tối ưu trong việc dự báo chất lượng nước dưới đất dựa vào chỉ số GWQI.

Bộ dữ liệu được thực hiện trên phần mềm thống kê R- Studio. Số liệu đầu vào được tính toán và hiệu chỉnh phù hợp nhằm loại bỏ những số liệu không đáng tin cậy, phù hợp cho việc đọc dữ liệu đầu vào của máy tính.

Để thực hiện được mô hình học máy, tác giả đã phân tích các dữ liệu đầu vào của các biến thông qua biểu đồ mối tương quan các biến (Hình 2). Sau khi hiểu rõ được các dữ liệu, tác giả tiến hành phân chia bộ dữ liệu ban đầu thành các phần ngẫu nhiên gồm Huấn luyện (Training), Kiểm tra (Test) và Kiểm chứng (Validation) theo tỷ lệ train:test trong khoảng (70:30) - (85:15) (Bảng 1 và Bảng 2). Tiến hành thao tác các thuật toán từ 5 mô hình học máy, thu được kết quả mô hình tối ưu dựa vào chỉ số MAE, RMSE và R^2.



Hình 2. *Biểu đồ tương quan giữa các biến trong bộ dữ liệu*

3. Kết quả nghiên cứu và khảo sát

3.1. Phân tích điển hình mẫu nước dưới đất trong mùa mưa tháng 11/2021 tại khu vực nghiên cứu

Dựa vào đặc điểm, tính chất khu vực và các nguồn có khả năng gây ô nhiễm, tác giả đã chọn ra được 5 nhóm vị trí lấy mẫu nước dưới đất bao gồm: N_0, N_1, N_2, N_3, N_4 như Hình 1.

Trong đó:

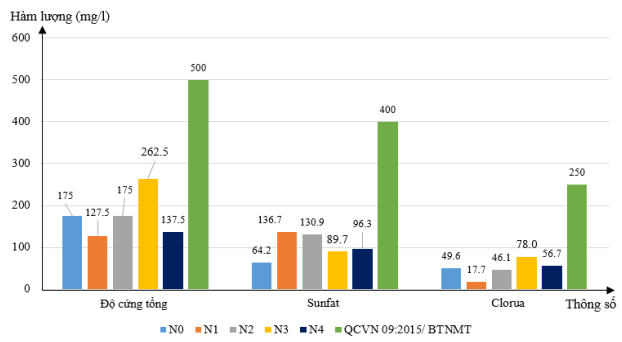
N_0 là điểm đại diện cho khu vực ít hoặc không chịu tác động bởi các nguồn ô nhiễm;

N_1 là điểm đại diện cho khu vực chịu tác động từ khu chăn nuôi gia súc, gia cầm và bãi rác;

N_2 là điểm đại diện cho khu vực chịu tác động từ nhà máy đốt rác và bãi rác;

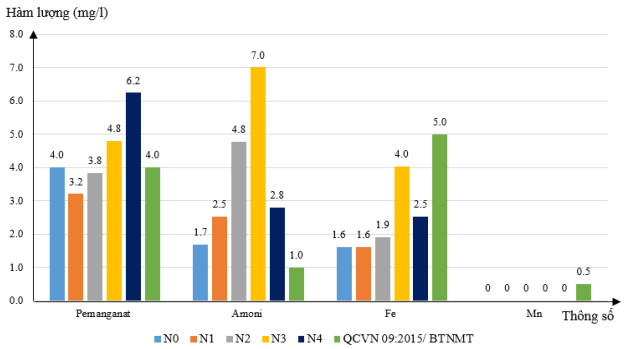
N_3 là điểm đại diện cho khu vực chịu tác động trực tiếp từ bãi rác, nghĩa trang;

N_4 là điểm đại diện cho khu vực chịu tác động từ bãi rác.



Hình 3. *Biểu đồ phân tích kết quả mẫu nước dưới đất tại khu vực bãi rác Cẩm Hà*

Kết quả phân tích các chỉ số nước dưới đất tại khu vực lân cận bãi rác Cẩm Hà được biểu diễn trên các biểu đồ Hình 3 và Hình 4.



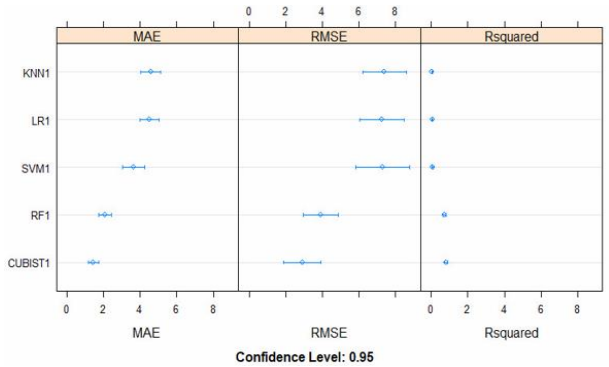
Hình 4. *Biểu đồ phân tích kết quả mẫu nước dưới đất tại khu vực bãi rác Cẩm Hà*

3.2. Kết quả thực hiện mô hình học máy

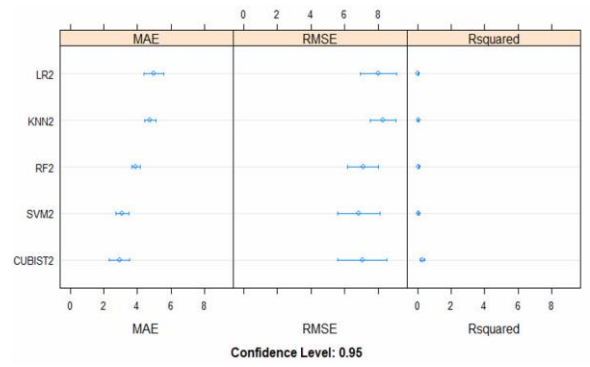
Tác giả thực hiện phân tích mối tương quan giữa các biến đầu vào và biến đầu ra GWQI (Hình 2) và nhận thấy rằng có sự tương quan thấp. Do vậy, cần mô hình học máy để giải quyết vấn đề dự báo chất lượng nước dưới đất tại khu vực nghiên cứu. Thông thường, nếu có sự tương quan lớn thì chỉ cần sử dụng đường tuyến tính giản đơn (linear regression) là có thể giải quyết được vấn đề dự báo, điều này chứng minh sự cần thiết việc áp dụng mô hình học máy trong dự báo chất lượng nước dưới dưới đất tại khu vực nghiên cứu.

Sau khi thực hiện mô hình từ dữ liệu Training và Validation lần lượt theo các tỷ lệ (70:30) - (85:15), thu được bảng kết quả về các chỉ số MAE, RMSE và R^2 như Bảng 1 và Bảng 2.

Kết quả mô hình tối ưu thu được khi thực hiện mô hình học máy với tỷ lệ 70:30 được thể hiện qua Hình 5 và Hình 6.



Hình 5. *Biểu đồ kết quả từ dữ liệu Training với tỷ lệ 70:30*



Hình 6. *Biểu đồ kết quả từ dữ liệu Validation với tỷ lệ 70:30*

Bảng 1. *Kết quả thực hiện mô hình từ dữ liệu Huấn luyện (Training)*

Training															
Split Ratio	MAE (min)					RMSE (min)					R^2 (max)				
	LR	SVM	KNN	RF	Cubist	LR	SVM	KNN	RF	Cubist	LR	SVM	KNN	RF	Cubist
70:30	2,43	1,36	1,99	0,90	**0,74**	3,06	1,61	2,85	1,33	**0,87**	0,269	0,516	0,161	0,987	**0,988**
71:29	2,43	1,57	2,62	0,80	**0,68**	3,22	1,97	3,79	1,05	**0,95**	0,269	0,394	0,554	0,985	**0,995**
72:28	3,18	1,87	3,07	**0,67**	0,69	4,20	2,14	4,24	0,95	**0,83**	0,393	0,191	0,449	0,986	**0,992**
73:27	2,42	1,81	2,66	0,82	**0,74**	2,98	2,07	3,25	1,05	**1,02**	0,258	0,270	0,394	0,991	**0,992**
74:26	2,38	1,39	1,65	0,68	**0,61**	2,95	1,75	2,12	0,84	**0,76**	0,341	0,338	0,602	0,976	**0,990**
75:25	2,81	1,69	2,35	**0,67**	0,70	3,46	1,98	2,98	**0,84**	0,93	0,259	0,396	0,346	0,974	**0,987**
76:24	2,63	1,60	2,33	0,67	**0,61**	3,38	1,78	3,43	0,90	**0,82**	0,198	0,422	0,627	0,991	**0,994**
77:23	2,74	1,93	2,79	**0,71**	**0,71**	3,18	2,39	3,71	**0,87**	0,89	0,141	0,206	0,598	0,987	**0,992**
78:22	2,63	1,69	3,04	0,77	**0,71**	3,07	2,33	4,33	1,09	**0,84**	0,234	0,166	0,395	**0,987**	0,900
79:21	2,76	1,95	2,74	0,79	**0,73**	3,40	2,30	3,95	**0,94**	1,01	0,209	0,199	0,613	0,985	**0,989**
80:20	2,95	1,49	2,43	**0,53**	0,64	3,68	1,86	3,19	**0,64**	0,95	0,181	0,461	0,592	**0,994**	0,991
81:19	2,67	1,84	2,74	0,72	**0,71**	3,21	2,23	3,68	**0,85**	1,15	0,275	0,192	0,581	0,992	**0,994**
82:18	3,18	1,64	2,75	0,79	**0,77**	3,74	2,01	3,92	1,00	**0,94**	0,246	0,363	0,556	0,985	**0,988**
83:17	2,96	1,66	2,79	0,79	**0,59**	3,37	1,96	3,75	0,98	**0,76**	0,279	0,421	0,494	**0,991**	0,989
84:16	2,67	1,39	2,02	0,73	**0,64**	3,14	1,78	2,61	0,95	**0,80**	0,301	0,392	0,477	0,982	**0,990**
85:15	2,13	1,49	2,53	0,81	**0,47**	2,69	1,88	3,07	0,93	**0,66**	0,179	0,356	0,611	0,986	**0,994**

Bảng 2. *Kết quả thực hiện mô hình từ dữ liệu Kiểm chứng (Validation)*

Validation															
Split Ratio	MAE (min)					RMSE (min)					R2 (max)				
	LR	SVM	KNN	RF	Cubist	LR	SVM	KNN	RF	Cubist	LR	SVM	KNN	RF	Cubist
70:30	3,43	1,44	3,38	2,66	**0,93**	4,30	1,72	4,36	3,46	**1,26**	0,12	0,41	0,24	0,36	**0,96**
71:29	3,03	1,43	2,55	1,70	**0,91**	3,73	1,78	3,48	2,13	**1,25**	0,09	0,37	0,18	0,66	**0,92**
72:28	2,86	1,65	3,25	2,20	**1,04**	4,30	1,92	4,69	2,83	**1,28**	0,23	0,17	0,13	0,38	**0,92**
73:27	2,50	1,40	2,38	1,72	**0,97**	3,52	1,76	3,74	2,14	**1,32**	0,16	0,34	0,13	0,53	**0,80**
74:26	2,31	1,51	2,63	2,08	**1,23**	2,56	1,70	3,35	2,41	**1,48**	0,13	0,47	0,09	0,58	**0,88**
75:25	2,656	1,41	2,65	2,13	**0,88**	3,28	1,77	3,51	2,46	**1,39**	0,11	0,49	0,25	0,28	**0,84**
76:24	1,89	1,20	2,33	1,89	**1,09**	2,30	1,53	2,82	2,45	**1,37**	0,26	0,66	0,14	0,41	**0,90**
77:23	1,91	1,29	2,59	2,11	**1,40**	2,22	**1,62**	3,23	2,55	1,68	0,12	0,50	0,19	0,30	**0,59**
78:22	1,96	1,29	1,68	1,33	**1,20**	2,37	1,69	2,21	1,69	**1,58**	0,20	0,39	0,20	0,30	**0,42**
79:21	2,10	1,29	2,30	1,67	**1,05**	2,46	1,62	2,70	1,93	**1,33**	0,17	0,41	0,24	0,36	**0,78**
80:20	1,71	1,44	2,17	1,69	**1,15**	2,11	1,66	2,76	1,94	**1,51**	0,28	0,35	0,29	0,31	**0,62**
81:19	2,00	**1,12**	1,55	1,46	1,45	2,33	1,52	1,85	**1,64**	1,91	0,18	0,53	0,32	0,64	**0,43**
82:18	2,74	**1,34**	1,71	1,71	1,54	3,26	1,56	2,14	2,04	**2,02**	0,15	0,37	0,34	**0,51**	0,49
83:17	1,85	**1,24**	1,97	1,81	1,27	2,39	1,47	2,46	2,24	**1,71**	0,18	0,58	0,50	0,43	**0,60**
84:16	2,79	1,36	3,00	2,00	**1,13**	3,37	1,76	3,50	2,43	**1,39**	0,25	0,47	0,19	**0,65**	0,57
85:15	2,50	**1,37**	2,78	1,92	1,56	2,74	**1,78**	3,40	2,30	2,06	0,274	0,343	0,343	0,318	**0,392**

4. Bàn luận

4.1. Bàn luận về kết quả phân tích mẫu nước dưới đất mùa mưa tháng 11/2021

Kết quả phân tích mẫu nước dưới đất tại khu vực bãi rác Cẩm Hà được so sánh với QCVN 09-MT:2015/BTNMT: Quy chuẩn kĩ thuật quốc gia về chất lượng nước dưới đất. Theo kết quả thu được từ Hình 3 và Hình 4, các thông số về độ cứng tổng (tính theo $CaCO_3$), Sunfat, Clorua, Mn, có hàm lượng (mg/l) thấp, đều nằm trong giới hạn cho phép QCVN 09-MT:2015/BTNMT.

Chỉ số Pemanganat tại vị trí N_3 và N_4 vượt lần lượt là 1,2 và 1,55 lần so với QCVN 09-MT:2015/BTNMT. Hàm

lượng amoni tại tất cả các vị trí đều vượt giới hạn cho phép QCVN 09-MT:2015/BTNMT, cụ thể vị trí N_0 vượt 1,7 lần, N_1 vượt 2,5 lần, N_2 vượt 4,8 lần, N_3 vượt 7,0 lần và vị trí N_4 vượt 2,8 lần. Hàm lượng amoni trong nước dưới đất cao không gây độc trực tiếp mà sản phẩm chuyển hóa từ amoni là nitrit và nitrat là yếu tố gây độc hại. Nguyên nhân dẫn đến hàm lượng amoni cao có thể một phần là do hoạt động sản suất nông nghiệp sử dụng quá nhiều phân bón và thuốc hóa học, hoặc do nguồn ô nhiễm từ bãi rác Cẩm Hà.

Với kết quả điển hình cụ thể tại khu vực này cho thấy, nguy cơ ô nhiễm là rất lớn, trong thời gian đến việc áp dụng các mô hình học máy để dự báo và thường xuyên cập nhật tình hình ô nhiễm thông qua việc ứng dụng mô hình tối ưu từ nghiên cứu này là hoàn toàn cấp thiết.

4.2. Bàn luận kết quả thực hiện mô hình học máy

Sau khi hoàn thiện thuật toán chạy các mô hình đã chọn, ta thu được bảng kết quả với cách chia tỷ lệ từ tập dữ liệu ban đầu trong khoảng (70:30) – (85:15). Mô hình được lựa chọn là mô hình tối ưu cho dự báo chất lượng nước dưới đất tại khu vực khảo sát dựa vào giá trị nhỏ nhất của các chỉ số như sai số tuyệt đối trung bình (MAE), sai số toàn phương trung bình (RMSE) và độ tin cậy R^2 cao nhất. Từ bảng kết quả dữ liệu Training Bảng 1 cho thấy, mô hình Cubist, RF là 2 mô hình có chỉ số MAE, RMSE thấp nhất và độ tin cậy R^2 cao nhất. Trong đó, mô hình Cubist là mô hình có chỉ số MAE, RMSE thấp nhất ở tỷ lệ (85:15) có giá trị lần lượt là 0,47 và 0,66 và độ tin cậy R^2 là 99,4%. Tiếp theo, để kiểm định lại độ chính xác của mô hình ta sử dụng tập dữ liệu Validation cho kết quả ở Bảng 2. Quan sát Bảng 2 thì ở tỷ lệ (85:15) cho kết quả R^2 không khả quan (39%). Kết hợp 2 bảng kết quả, nhìn chung mô hình Cubist có kết quả cao nhất khi Huấn luyện và Kiểm chứng lại ở tỷ lệ (70:30). Các mô hình khác không cho kết quả khả quan khi được Huấn luyện và Kiểm chứng lại, hoặc cho kết quả khả quan khi Huấn luyện nhưng khi kiểm tra lại thì không đạt được kết quả tốt (R^2 không cao).

5. Kết luận

Kết quả nghiên cứu, phân tích và thực hiện mô hình học máy cho thấy, việc áp dụng khoa học công nghệ tiên tiến, thông minh, tự động trong công tác quản lý và giám sát chất lượng môi trường nước dưới đất tại khu vực khảo sát là hữu ích và cực kỳ quan trọng hiện nay. Thông qua đó, có thể biết được hiện trạng ô nhiễm môi trường nước dưới đất tại khu vực khảo sát, từ đó đưa ra các giải pháp quản lý và khắc phục tình trạng ô nhiễm. Hơn nữa trong thời đại hiện nay, với sự tham gia, hỗ trợ của các phần mềm giúp cho việc tính toán, đưa ra dự báo, các dữ liệu khó tính toán trên cơ sở từ các dữ liệu đã có hoặc dễ tính toán diễn ra thuận lợi hơn.

Bên cạnh đó, việc đánh giá chất lượng nguồn nước dưới đất cũng gặp một vài thách thức như mẫu thu thập ở quy mô lớn, xử lý số liệu mất nhiều thời gian, thiết bị, hoá chất và nguồn lực con người. Ngoài ra, việc tính toán các chỉ số chất lượng nước dưới đất là một quá trình lâu dài, cần nguồn lực kinh tế lớn. Vì vậy, để giải quyết những vấn đề này, học máy (machine learning) là cách tiếp cận tiềm năng và tiết kiệm chi phí, hiệu quả và đáng tin cậy trong đánh giá chất lượng nước dưới đất.

Lời cảm ơn: Nghiên cứu này được tài trợ bởi Bộ Giáo dục và Đào tạo Việt Nam trong đề tài mã số B2022-DNA-04.

TÀI LIỆU THAM KHẢO

[1] S. Varol, A. Davraz, "Evaluation of the groundwater quality with WQI (Water Quality Index) and multivariate analysis: a case study of the Tefenni plain (Burdur/Turkey)", *Environmental Earth Sciences.*, Vol. 73, No. 4, 2015, pp. 1725-1744,.

[2] Ahmed, U., Mumtaz, R., Anwar, H., Shah, A.A., Irfan, R., García-Nieto, J., "Efficient water quality prediction using supervised Machine Learning", *Water*, 2019, 11 (11), 2210.

[3] Bui, D.T., Khosravi, K., Tiefenbacher, J., Nguyen, H., Kazakis, N., "Improving prediction of water quality indices using novel hybrid machine-learning algorithms", *Science of The Total Environment*, 2020a, p. 137612.

[4] Bui, D.T., Hoang, N.D., Martínez-Alvarez, F., Ngo, P.T.T., Hoa, P.V., Pham, T.D., Samui, P., Costache, R. "A novel deep learning neural network approach for predicting flash flood susceptibility: a case study at a high frequency tropical storm area", *Sci. Total Environ.* 2020b, 701, 134413.

[5] Kazakis, N., Mattas, C., Pavlou, A., Patrikaki, O., Voudouris, K. "Multivariate statistical analysis for the assessment of groundwater quality under different hydrogeological regimes", *Environmental Earth Sciences*, 2017, 76 (9), 349.

[6] Kim, J., Han, H., Johnson, L.E., Lim, S., Cifelli, R. "Hybrid machine learning framework for hydrological assessment", *J. Hydrol.*, 2019, 577, 123913.

[7] Li, P.Y., Wu, J.H., Qian, H. "Groundwater quality assessment based on entropy weighted osculating value method", *Int. J. Environ. Sci.* 2010, 1 (4), 621-630.

[8] Li, Z., Yang, T., Huang, C.S., Xu, C.Y., Shao, Q., Shi, P., Wang, X., Cui, T. "An improved approach for water quality evaluation: TOPSIS-based informative weighting and ranking (TIWR) approach", *Ecol. Indicat.* 2018, 89, 356-364.

[9] Lu, H., Ma, X. "Hybrid decision tree-based machine learning models for short- term water quality prediction", *Chemosphere*, 2020, 249, 126169.

[10] Maier, H.R., Dandy, G.C. "Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications", *En- viron. Model. Software*, 2000, 15 (1), 101-124.

[11] Palani, S., Liong, S.Y., Tkalich, P. "An ANN application for water quality fore- casting", *Mar. Pollut. Bull.*, 2008, 56 (9), 1586-1597.