NGHIÊN CỨU ỨNG DỤNG CÁC MÔ HÌNH HỌC MÁY ĐỂ DỰ BÁO CHỈ SỐ CHẤT LƯỢNG NƯỚC MẶT VÙNG BÁN ĐẢO CÀ MAU

Nguyễn Đức Phong, Hà Hải Dương
Viện Nước, Tưới tiêu và Môi trường

Tóm tắt: *Chất lượng nước mặt vùng BĐCM đang bị ô nhiễm do ảnh hưởng của xả thải không đạt yêu cầu vào nguồn nước. Nguồn nước mặt trong vùng bị ô nhiễm phổ biến là hữu cơ và vi sinh với các thông số DO, BOD_5, COD, NH_4^+, tổng Coliform,... Trong vùng nghiên cứu, các địa phương thường dùng chỉ số chất lượng nước (WQI) để đánh giá chất lượng nước mặt và khả năng sử dụng của nguồn nước đối với từng mục đích khác nhau. Tuy nhiên, việc tính toán WQI từ các thông số quan trắc chất lượng nước còn gặp nhiều hạn chế do cần nhiều thông số quan trắc và tính toán còn tương đối phức tạp. Việc tìm phương pháp tính toán hiệu quả WQI là rất quan trọng và cần thiết nhằm phục vụ tốt hơn cho công tác đánh giá chất lượng nước mặt cho vùng nghiên cứu. Nghiên cứu này sẽ ứng dụng thuật toán (mô hình) học máy để tính toán WQI dựa vào số liệu đầu vào (thông số chất lượng nước tối thiểu) để giảm chi phí quan trắc chất lượng nước mặt. Nghiên cứu đã áp dụng phương pháp Bayes (BMA) để lựa chọn các thông số chất lượng nước tối ưu (pH, BOD_5, PO_4 và Coliform) để tính toán WQI. Kết quả cho thấy các mô hình học máy đã tính toán (dự báo) WQI dựa vào các thông số (tối tiểu) với độ chính xác cao. Theo đó mô hình Tăng cường độ dốc có kết quả dự báo chính xác nhất vì có hệ số xác định R^2 cao nhất (0,973), giá trị các sai số MAE, MSE và RMSE thấp nhất (3,24; 22,54; 4,75). Tiếp đến là mô hình Tăng cường độ dốc cực đại có R^2 là 0,966 và giá trị các sai số tương ứng (3,15; 28,95; 5,38). Mô hình Cây quyết định có R^2 là 0,944; giá trị các sai số là 4,46; 49,67; 7,04; Mô hình Tăng cường độ dốc nhẹ có R^2 là 0,928; giá trị các sai số là 5,95; 63,30; 7,95.*
Từ khóa: *Bán đảo Cà Mau, phương pháp BMA, mô hình học máy, chỉ số chất lượng nước mặt (WQI)*

Summary: *Surface water quality in the Ca Mau peninsula is being polluted due to the influence of unsatisfactory discharge into water sources. Surface water sources in polluted areas are organic and microbiological with parameters of DO, BOD_5, COD, NH_4^+, total Coliform, etc. In the study area, localities often use water quality index (WQI). to assess surface water quality and usability of water sources for different purposes. However, the calculation of WQI from water quality monitoring parameters still faces many limitations because it requires many monitoring parameters and is relatively complicated. Finding an effective WQI calculation method is very important and necessary to better serve the assessment of surface water quality for the study area. This study will apply a machine learning algorithm (model) to calculate the WQI based on the minimum input data (water quality parameters) to reduce the cost of surface water quality monitoring. The study applied Bayesian method (BMA) to select optimal water quality parameters (pH, BOD_5, PO_4 and Coliform) to calculate WQI. The results show that the machine learning models have calculated (predicted) WQI based on (minimum) parameters with high accuracy. Accordingly, Gradient Boosting model has the most accurate prediction results because it has the highest coefficient of determination R^2 (0.973), the lowest error values of MAE, MSE and RMSE (3.24; 22.54; 4.75). XGBoost model with R^2 of 0.966 and the corresponding error values (3.15; 28.95; 5.38). The Decision Tree model has an R^2 of 0.944; the lowest error values is 4.46; 49.67; 7.04); The LightGBM model has an R^2 of 0.928; false value value is 5.95; 63.30; 7.95.*
Keywords: *Ca Mau peninsula, Bayesian Model Averaging method, machine learning model, surface water quality index (WQI).*

1. ĐẶT VẤN ĐỀ

Vùng nghiên cứu (BĐCM) nằm ở phía Nam kênh Cái Sắn và hữu ngạn sông Hậu, có tổng diện tích tự nhiên khoảng 1.678.000 ha; bao gồm thành phố Cần Thơ, các tỉnh Hậu Giang, Sóc Trăng, Bạc Liêu, Cà Mau và phần phía Nam tỉnh Kiên Giang (gồm các huyện Giồng Riềng, An Biên, An Minh, Vĩnh Thuận, Gò Quao và các xã phía Nam các huyện Châu Thành, Tân Hiệp).

Ngày nhận bài: 16/01/2023
Ngày thông qua phản biện: 06/02/2023
Ngày duyệt đăng: 10/02/2023

Đối với vùng BĐCM, sông Hậu là con sông cấp nước chính cho vùng, tuy nhiên, một số đoạn sông của sông Hậu đã ghi nhận hiện tượng ô nhiễm cục bộ chất hữu cơ, với mức độ khác nhau do nước thải từ các khu công nghiệp và hoạt động khai thác cát, nuôi trồng thuỷ sản không qua xử lý, xả thẳng ra môi trường. Điển hình như đoạn sông Hậu qua khu vực Nhơn Mỹ và Cái Côn đã có dấu hiệu ô nhiễm hữu cơ và vi sinh. Bên cạnh đó, độ đục cũng ở mức khá cao do các sông ở khu vực này có phù sa lớn [2], [18]. Đối với chất lượng nước mặt hệ thống kênh nội vùng BĐCM, nước mặt vùng nghiên cứu chủ yếu bị ô nhiễm hữu cơ, chất dinh dưỡng, vi sinh vật và có xu hướng bị nhiễm phèn. Mức độ ô nhiễm hữu cơ, chất dinh dưỡng và vi sinh vật ngày càng tăng qua các năm [15]. Theo đánh giá chất lượng nước mặt của các địa phương trong vùng BĐCM, diễn biến chất lượng nước của hệ thống sông kênh của từng tỉnh trong vùng nghiên cứu được trình bày dưới đây:

- Chất lượng nước mặt tỉnh Hậu Giang: Chất lượng nước mặt tại một số sông, kênh địa bàn tỉnh đã bị ô nhiễm hữu cơ và ô nhiễm vi sinh. Nhiều thông số quan trắc môi trường nước mặt tại các khu vực này đã vượt tiêu chuẩn cho phép như: DO, TSS, $N\text{-}NO_2^-$, NH_4^+, BOD_5, COD và tổng Coliforms [5]. Tại một số tuyến sông chính trên địa bàn như khu vực kênh xáng Xà No, Cái Côn, Lái Hiếu, sông Hậu đoạn chảy qua địa bàn huyện Châu Thành, một số tuyến sông thuộc huyện Long Mỹ… chất lượng nước đều đã bị ô nhiễm;

- Chất lượng nước mặt thành phố Cần Thơ: Chất lượng nước mặt tại các kênh rạch trên địa bàn thành phố Cần Thơ đã bị ô nhiễm hữu cơ và vi sinh [14]. Theo Sở Tài nguyên Môi trường Cần Thơ, các vị trí quan trắc năm 2020 trên 19 sông, kênh, rạch trên thì có 14 sông, kênh, rạch có chỉ số WQI nằm trong khoảng từ 51 đến 75 (chỉ sử dụng cho mục đích tưới tiêu hoặc tương đương); chỉ có 04/19 kênh, rạch có chỉ số WQI nằm trong khoảng từ 76 đến 90 (sử dụng cấp nước sinh hoạt, nhưng cần xử lý) đó là rạch Cái Sơn - Hàng Bàng quận Ninh Kiều; rạch Cam quận Bình Thủy; rạch Ba Láng quận Cái Răng và rạch Bò Ót quận Thốt Nốt [1];

- Chất lượng nước mặt tỉnh Sóc Trăng: Nước mặt thuộc tỉnh Sóc Trăng cũng bị ô nhiễm hữu cơ và vi sinh, giá trị COD, BOD5 tại hầu hết các điểm quan trắc đều vượt giới hạn cho phép [17]. Theo Sở Tài nguyên Môi trường Sóc Trăng, các vị trí quan trắc năm 2020 tại 19 sông, kênh được quan trắc chỉ có 04/19 kênh, rạch có chỉ số WQI nằm trong khoảng từ 76 đến 90 (sử dụng cấp nước sinh hoạt, nhưng cần xử lý). Có 10/19 sông, kênh có chỉ số WQI nằm trong khoảng từ 25 đến 75. Trong đó có 2 vị trí nước bị ô nhiễm nặng (WQI =25), cần các biện pháp xử lý là kênh Thạnh Lợi, kênh thị trấn Long Phú;

- Chất lượng nước mặt tỉnh Bạc Liêu: Nước mặt tỉnh Bạc Liêu cũng bị ô nhiễm hữu cơ và vi sinh, giá trị COD, BOD5 tại hầu hết các điểm quan trắc đều vượt giới hạn cho phép từ 1,2 – 10,1 lần. Theo Sở Tài nguyên Môi trường Bạc Liêu, các vị trí quan trắc năm 2020 tại 8 sông, kênh được quan trắc chỉ có 1/8 kênh, sông có chỉ số WQI nằm trong khoảng từ 76 đến 90 (sử dụng cấp nước sinh hoạt, nhưng cần xử lý). Có 6/8 sông, kênh có chỉ số WQI nằm trong khoảng từ 25 đến 75. Trong đó có 2 vị trí nước bị ô nhiễm nặng (WQI =25), cần các biện pháp xử lý là cống Cái Cùng (huyện Hoà Bình) và cống Đầu Bằng (TX xã Giá Rai). Ở một số vị trí như Cửa Gành Hào (huyện Đông Hải), Ngã tư Chủ Chí (huyện Phước Long), Cửa Nhà Mát (TP. Bạc Liêu), Cống Hưng Thành (huyện Vĩnh Lợi), Vĩnh Lộc - Ba Đình (huyện Hồng Dân) giá trị WQI nằm trong khoảng từ 49-70, nước chỉ sử dụng cho giao thông thủy và các mục đích tương đương khác [9, 6-8];

- Chất lượng nước mặt tỉnh Cà Mau: Nước mặt tỉnh Cà Mau cũng bị ô nhiễm hữu cơ và vi sinh, giá trị COD, BOD_5 tại hầu hết các điểm quan trắc đều vượt giới hạn cho phép từ 1,4 – 11,5 lần. Theo Sở Tài nguyên Môi trường Cà Mau, các vị trí quan trắc năm 2020 tại 20 vị trí được quan trắc chỉ có 1/20 vị trí (Cửa sông Cửa lớn) có chỉ số WQI nằm trong khoảng từ 76 đến 90. Có 10/20 vị trí có chỉ số WQI nằm trong khoảng từ 25 đến 50. Đây cũng là những vị trí nước bị ô nhiễm nặng, cần các biện pháp xử lý (WQI < 25) [12, 13, 10, 11].

Như vậy, hiện trạng môi trường nước mặt vùng BĐCM vẫn đang diễn biến khá phức tạp (với nhiều nguồn xả thải không được xử lý trước khi xả vào nguồn nước), đặc biệt là tại các đô thị, trung tâm và khu dân cư đông đúc sống 2 bên sông chất lượng nước mặt bị ô nhiễm chủ yếu do nước thải sinh hoạt, một số nơi khác lại do hoạt động sản xuất công nghiệp, nuôi trồng thủy sản... [3, 4].

Có thể thấy, các địa phương trong vùng nghiên cứu thường dùng chỉ số chất lượng nước (WQI) để đánh giá chất lượng nước mặt và khả năng sử dụng của nguồn nước đối với từng mục đích khác nhau và phải dựa vào rất nhiều thông số để tính toán WQI và quá trình tính toán tương đối phức tạp. Theo Quyết định số 1460/QĐ - TCMT của Tổng cục Môi trường ban hành về việc Hướng dẫn kỹ thuật tính toán và công bố chỉ số chất lượng nước của Việt Nam (VN WQI), số liệu để tính toán VN_WQI phải bao gồm tối thiểu 3/5 nhóm thông số, trong đó bắt buộc phải có nhóm IV (nhóm thông số hữu cơ và dinh dưỡng) phải có tối thiểu 3 thông số. Thực tế, các địa phương thường dùng 3 nhóm thông số Nhóm I (pH); Nhóm IV (DO, BOD_5, COD, TOC, N-NH_4, N-NO_3, N-NO_2, P-PO_4) và Nhóm V (Coliform).

Trên thế giới và Việt Nam, các kỹ thuật học máy đã được sử dụng rộng rãi để tính toán (dự báo) chất lượng nước mặt cũng như tính toán WQI [33]. Phương pháp này đã được chứng minh là có nhiều ưu điểm vượt trội (so với phương pháp truyền thống) để mô hình hóa các phương trình phi tuyến tính phức tạp trong nghiên cứu tài nguyên nước [29]. Mỗi thuật toán học máy đều có ưu nhược điểm khác nhau và phụ thuộc vào các biến đầu vào. Đối với dự báo chất lượng nước, các thuật toán học máy (Machine Learning) được sử dụng phổ biến là Adaboost [19], GBM [28], XGBoost [22], cây quyết định (DT) [32], [20], cây tăng cường (ExT) [21], rừng ngẫu nhiên (RF) [24], [27]. Các thuật toán học sâu (Deep Learning) được ứng dụng là tri giác đa lớp (MLP) [25], hàm cơ sở xuyên tâm (RBF) [26], mạng thần kinh chuyển tiếp nguồn cấp dữ liệu sâu (DFNN) [23], và mạng thần kinh tích chập (CNN) [31] đã được ứng dụng. Một số nghiên cứu còn ứng dụng rất nhiều thuật toán (cả học máy và học sâu) để tính toán [30]. Mặc dù có nhiều thuật toán được áp dụng và cho các kết quả khả quan, tuy nhiên còn gặp một số là có quá nhiều số liệu đầu vào phục vụ tính toán, điều này cần một lượng lớn số liệu quan trắc và kết quả mô hình có thể bị quá khớp với dữ liệu (overfitting).

Do vậy, việc nghiên cứu ứng dụng các mô hình học máy để dự báo chỉ số chất lượng nước mặt vùng BĐCM là quan trọng và cần thiết. Nghiên cứu sẽ góp phần cung cấp thêm phương pháp tính toán chỉ số chất lượng nước mặt khoa học, hiệu quả, tốn ít chi phí nhằm thích hợp với điều kiện thực tế của các địa phương trong vùng Bán đảo Cà Mau.

2. PHƯƠNG PHÁP THỰC HIỆN

2.1. Mục tiêu

- Xây dựng được cơ sở khoa học tính toán chỉ số chất lượng nước mặt bằng phương pháp học máy;

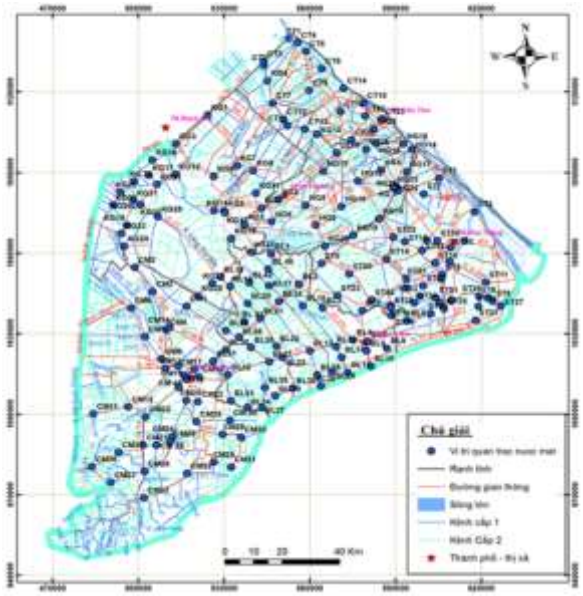
- Đề xuất được phương pháp tính toán chỉ số chất lượng nước mặt bằng phương pháp học máy thích hợp với điều kiện thực tế của các địa phương trong vùng Bán đảo Cà Mau.

2.2. Phương pháp thực hiện

Để đạt được các mục tiêu đề ra, các phương pháp nghiên cứu được sử dụng như sau:

2.2.1. Phương pháp đo đạc hiện trường, lấy mẫu phân tích

Khảo sát đo đạc, lấy mẫu phân tích các chỉ tiêu đánh giá ô nhiễm nước và chất lượng nước mặt, nước thải. Việc lấy mẫu nước mặt để đánh giá được xu thế chung biến đổi chất lượng nước [61], [68]. Các vị trí được quan trắc có tính điển hình, đại diện cho vùng nghiên cứu theo các tiêu chí đảm bảo phân bố theo các trục kênh chính của BĐCM (xem Hình 2.1).



Hình 2.1: Vị trí lấy mẫu phân tích chất lượng nước mặt vùng BĐCM

2.2.2. Phương pháp Bayes (BMA)

Phương pháp Bayes (BMA) khai thác nhân tố Bayes (BF) và chỉ số đo lường mức độ quân bình "compromise" giữa độ phức tạp và khả năng tiên lượng của mô hình (BIC) để chọn mô hình tối ưu. Đây là phương pháp mới khắc phục được vấn đề thừa biến (biến không có tác động thực tế) trong mô hình hồi quy tuyến tính đa biến [16].

Theo kết quả quan trắc chất lượng nước có rất nhiều thông số chất lượng nước là lý, hóa và vi sinh (pH, TSS, DO, BOD_5, COD, NH_4, PO_4, Coliform) quyết định đến ô nhiễm, tức là đến chất lượng nước (ở đây là giá trị WQI). Để xác định được các thông số đặc trưng phục vụ mô hình học máy trong vùng BĐCM, nghiên cứu đã dùng phương pháp Bayes để xác định được những biến (thông số chất lượng nước) có ảnh hưởng lớn đến WQI. Kết quả phân tích thống kê bằng phương pháp Bayes (BMA) sẽ xác định được các thông số chất lượng nước có ảnh hưởng lớn đến giá trị WQI, từ đó xác định được các thông số chính ảnh hưởng đến WQI.

2.2.3. Phương pháp học máy

Nghiên cứu sử dụng các mô hình học máy để tính toán (dự báo) WQI với 2 nhóm chính: thuật toán tăng cường, thuật toán cây quyết định vì đây là 2 thuật toán cho kết quả tính toán có độ chính xác cao, các thuật toán dễ hiểu và dễ triển khai.

2.2.3.1. Thuật toán tăng cường

Thuật toán tăng cường là một phương pháp được sử dụng trong máy học để giảm lỗi trong quá trình phân tích dữ liệu dự đoán. Các nhà khoa học dữ liệu đào tạo phần mềm máy học, hay còn gọi là các mô hình máy học, trên dữ liệu được gắn nhãn để dự đoán về dữ liệu chưa được gắn nhãn. Một mô hình máy học có thể dự đoán lỗi dựa trên độ chính xác của tập dữ liệu đào tạo. Để đào tạo mô hình thuật toán tăng cường, một thuật toán cần trải qua những bước tổng quát sau:

Bước 1: Thuật toán tăng cường chỉ định trọng số như nhau cho mỗi mẫu dữ liệu. Quá trình này cung cấp dữ liệu cho mô hình máy đầu tiên, được gọi là thuật toán cơ sở. Thuật toán cơ sở đưa ra dự đoán cho mỗi mẫu dữ liệu.

Bước 2: Thuật toán tăng cường đánh giá các dự đoán mô hình và tăng trọng số của các mẫu với một lỗi nghiêm trọng hơn. Quá trình này cũng chỉ định một trọng số dựa trên hiệu suất của mô hình. Mô hình cho ra các dự đoán xuất sắc sẽ có ảnh hưởng lớn đến quyết định cuối cùng.

Bước 3: Thuật toán chuyển dữ liệu được chỉ định trọng số sang cây quyết định tiếp theo.

Bước 4: Thuật toán lặp lại bước 2 và 3 đến khi các trường hợp lỗi đào tạo xảy ra thấp hơn ngưỡng nhất định.

Thuật toán tăng cường có những ưu điểm như sau:

- Dễ triển khai: Thuật toán tăng cường có các thuật toán dễ hiểu và dễ diễn giải, được đúc kết từ sai lầm. Các thuật toán này không yêu cầu bất cứ quá trình tiền xử lý dữ liệu nào, đồng thời còn có các quy trình tích hợp sẵn để xử lý dữ liệu còn thiếu.

- Giảm thiên kiến: Thiên kiến là sự tồn tại của tính không chắc chắn hoặc không chính xác trong kết quả của máy học. Các thuật toán tăng cường kết hợp nhiều máy học yếu theo phương pháp có trình tự liên tục cải thiện các dự đoán.

Hướng tiếp cận này giúp giảm mức độ thiên kiến cao thường gặp ở các mô hình máy học.

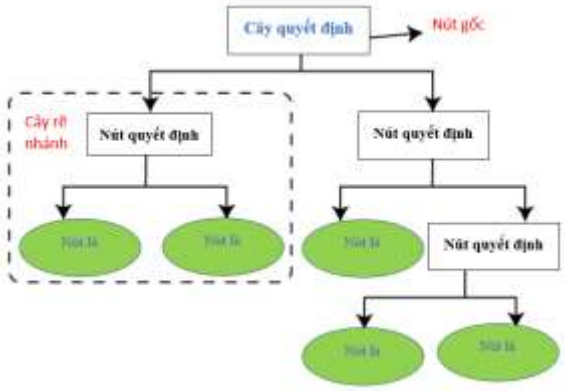
- Hiệu quả điện toán: Các thuật toán tăng cường ưu tiên những tính năng làm tăng độ chính xác của dự đoán trong quá trình đào tạo. Các thuật toán này giúp giảm thuộc tính dữ liệu và xử lý tập dữ liệu lớn một cách hiệu quả.

Tuy nhiên thuật toán tăng cường còn có những nhược điểm như dễ bị ảnh hưởng bởi dữ liệu ngoại lai. Các mô hình thuật toán tăng cường dễ bị ảnh hưởng bởi dữ liệu ngoại lai hoặc giá trị dữ liệu khác với phần còn lại của tập dữ liệu. Vì mỗi mô hình đều cố gắng khắc phục các lỗi của phiên bản tiền nhiệm, nên dữ liệu ngoại lai có thể làm kết quả bị sai lệch đáng kể.

2.2.3.2. Thuật toán cây quyết định

Cây quyết định là một thuật toán học tập có giám sát không tham số, được sử dụng cho cả nhiệm vụ phân loại và hồi quy. Nó có cấu trúc dạng cây, phân cấp, bao gồm nút gốc (root node), các nhánh, các nút bên trong (internal node) và các nút lá (leaf nodes). Cây quyết định bắt đầu bằng một nút gốc, không có bất kỳ nhánh nào đến. Các nhánh đi từ nút gốc sau đó đưa vào các nút bên trong, còn được gọi là nút quyết định. Dựa trên các đặc điểm sẵn có, cả hai loại nút đều tiến hành đánh giá để tạo thành các tập con đồng nhất, được ký hiệu bằng các nút lá, hoặc các nút đầu cuối. Các nút lá đại diện cho tất cả các kết quả có thể có trong tập dữ liệu.

Trong cây quyết định, để dự đoán lớp của tập dữ liệu đã cho, thuật toán bắt đầu từ nút gốc của cây. Thuật toán này so sánh các giá trị của thuộc tính gốc với thuộc tính bản ghi (tập dữ liệu thực) và dựa trên sự so sánh, đi theo nhánh và nhảy đến nút tiếp theo. Đối với nút tiếp theo, thuật toán lại so sánh giá trị thuộc tính với các nút con khác và di chuyển xa hơn. Nó tiếp tục quá trình cho đến khi nó đạt đến nút lá của cây (Hình 2.2).



Hình 2.2: Sơ đồ thuật toán cây quyết định

Quy trình hoàn chỉnh có thể được hiểu rõ hơn bằng cách sử dụng thuật toán dưới đây:

Bước 1: Bắt đầu cây với nút gốc (Đặt tên: S), nút này chứa tập dữ liệu hoàn chỉnh.

Bước 2: Tìm thuộc tính tốt nhất trong tập dữ liệu bằng cách sử dụng Phép đo lựa chọn thuộc tính (ASM).

Bước 3: Chia S thành các tập con chứa các giá trị có thể có cho các thuộc tính tốt nhất.

Bước 4: Tạo nút cây quyết định chứa thuộc tính tốt nhất.

Bước 5: Tạo một cách đệ quy cây quyết định mới bằng cách sử dụng các tập con của tập dữ liệu đã tạo ở bước -3. Tiếp tục quá trình này cho đến khi đạt đến một giai đoạn mà bạn không thể phân loại thêm các nút và được gọi là nút cuối cùng là nút lá.

Một số ưu điểm của thuật toán cây quyết định như sau:

- Dễ hiểu: các biểu diễn trực quan của cây quyết định giúp chúng dễ hiểu và dễ hiểu hơn. Bản chất phân cấp của cây quyết định cũng giúp bạn dễ dàng thấy thuộc tính nào là quan trọng nhất, điều này không phải lúc nào cũng rõ ràng với các thuật toán khác, như mạng nơ-ron.

- Ít hoặc không cần chuẩn bị dữ liệu: Cây quyết định có một số đặc điểm, làm cho nó linh hoạt hơn các bộ phân loại khác. Nó có thể xử lý các kiểu dữ liệu khác nhau, tức là các giá trị rời rạc hoặc liên tục và các giá trị liên tục có thể được chuyển đổi thành các giá trị phân loại thông qua việc sử dụng các ngưỡng.

- Linh hoạt hơn: Cây quyết định có thể được tận dụng cho cả nhiệm vụ phân loại và hồi quy, làm cho nó linh hoạt hơn so với một số thuật toán khác. Nó cũng không nhạy cảm với các mối quan hệ cơ bản giữa các thuộc tính; điều này có nghĩa là nếu hai biến có tương quan cao, thuật toán sẽ chỉ chọn một trong các đặc điểm để tách.

Tuy nhiên, thuật toán cây quyết định còn một số nhược điểm:

- Dễ bị hiện tượng mô hình tìm được quá khớp với dữ liệu (overfitting): Cây quyết định phức tạp có xu hướng quá mức và không tổng quát hóa tốt cho dữ liệu mới.

- Các công cụ ước tính phương sai cao: Các biến thể nhỏ trong dữ liệu có thể tạo ra một cây quyết định rất khác. Tính tổng hợp, hoặc tính trung bình của các ước tính, có thể là một phương pháp giảm phương sai của cây quyết định. Tuy nhiên, cách tiếp cận này bị hạn chế vì nó có thể dẫn đến các yếu tố dự báo có tương quan cao.

3. KẾT QUẢ VÀ THẢO LUẬN

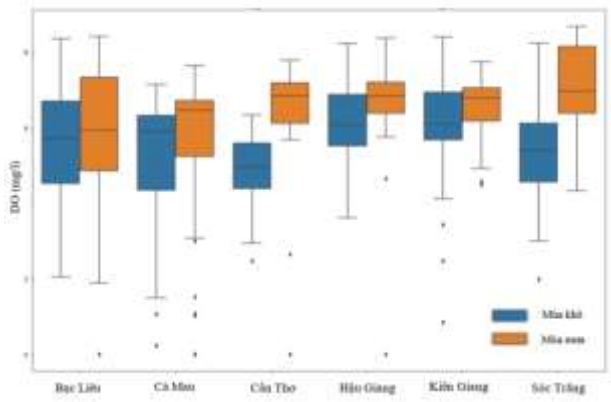
3.1. Kết quả quan trắc chất lượng nước mặt

Theo kết quả quan trắc chất lượng nước mặt năm 2016 tại các vị trí lấy mẫu vùng nghiên cứu [2] vào mùa khô và mùa mưa, kết quả quan trắc một số thông số chất lượng nước trong vùng nghiên cứu được tổng hợp trong Bảng 3.1. và các biểu đồ một số thông số chất lượng nước chính từ Hình 3.1 – Hình 3.4.

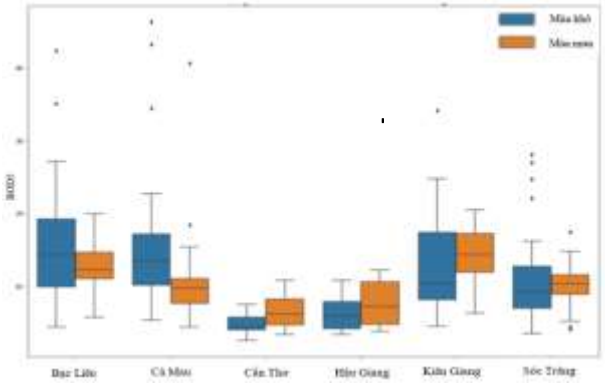
Qua phân tích ở trên, thấy rằng nước mặt vùng nghiên cứu chủ yếu bị ô nhiễm hữu cơ, chất dinh dưỡng, vi sinh vật. Các thông số vượt nhiều lần TCCP là DO, BOD_5, NH_4^+ và tổng Coliform (đây cũng là những thông số ô nhiễm điển hình của vùng nghiên cứu). Mức độ ô nhiễm hữu cơ, chất dinh dưỡng và vi sinh vật ngày càng tăng qua các năm vượt TCCP từ 1,1 – 3,7 lần). Các kênh/rạch bị ô nhiễm là Cái Sơn Hàng Bàng; rạch Cam; Ba Láng và Bò Ót... (Cần Thơ); kênh Xà No, Cái Côn, Lái Hiếu, một số tuyến sông thuộc huyện Long Mỹ (Hậu Giang); Kênh 8 mét, kênh 16 mét, kênh 30/4, sông Cổ Cò, kênh chợ Thị xã Vĩnh Châu, kênh TT Huỳnh Hữu Nghĩa, kênh thị trấn Châu Thành (Sóc Trăng); kênh Quản Lộ - Phụng Hiệp, Phó sinh – Phước Long; kênh Bạc Liêu - Cà Mau (Bạc Liêu); và các kênh rạch thuộc thành phố Cà Mau.

Bảng 3.1: Tổng hợp kết quả phân tích chất lượng nước vùng BĐCM

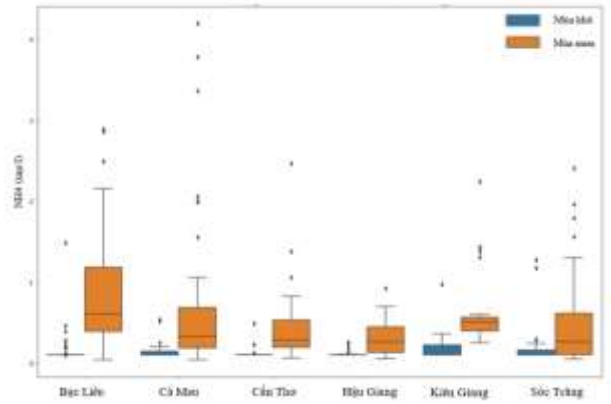
Thông số	Đặc trưng	Mùa khô (N=239)	Mùa mưa (N=239)	Tổng (N=478)
pH	Trung bình (SD)	7.61 (0.335)	7.25 (0.252)	7.43 (0.346)
	Trung vị [Min, Max]	7.59 [5.57, 8.75]	7.24 [6.52, 8.07]	7.40 [5.57, 8.75]
DO	Trung bình (SD)	5.48 (1.48)	6.25 (1.68)	5.87 (1.63)
	Trung vị [Min, Max]	5.68 [0.210, 8.40]	6.62 [0, 8.68]	6.16 [0, 8.68]
BOD_5	Trung bình (SD)	11.5 (7.44)	13.6 (30.0)	12.6 (21.9)
	Trung vị [Min, Max]	9.40 [2.20, 46.3]	10.5 [3.40, 420]	10.2 [2.20, 420]
COD	Trung bình (SD)	20.5 (13.5)	27.1 (61.4)	23.8 (44.5)
	Trung vị [Min, Max]	16.9 [4.00, 96.8]	20.5 [6.70, 880]	19.0 [4.00, 880]
NH_4	Trung bình (SD)	0.238 (0.464)	0.967 (2.15)	0.603 (1.59)
	Trung vị [Min, Max]	0.100 [0.100, 3.99]	0.430 [0.0100, 20.2]	0.200 [0.0100, 20.2]
PO_4	Trung bình (SD)	0.422 (3.68)	0.752 (5.24)	0.587 (4.53)
	Trung vị [Min, Max]	0.100 [0.0100, 55.9]	0.110 [0.0100, 65.5]	0.100 [0.0100, 65.5]
TSS	Trung bình (SD)	113 (119)	225 (284)	169 (224)
	Trung vị [Min, Max]	68.0 [7.40, 732]	110 [4.00, 1530]	95.4 [4.00, 1530]
Coliform	Trung bình (SD)	22600 (38600)	38800 (63100)	30700 (52900)
	Trung vị [Min, Max]	5200 [180, 320000]	12000 [180, 540000]	7900 [180, 540000]



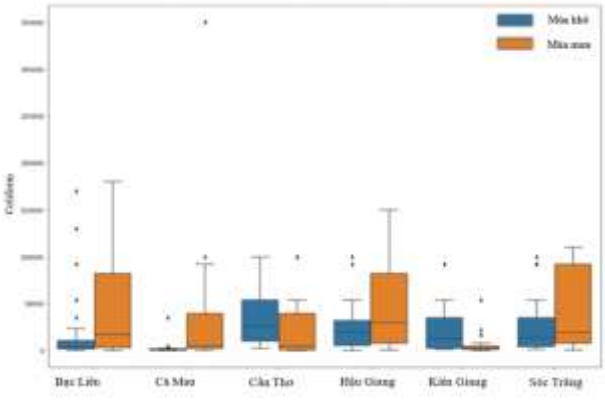
Hình 3.1: Biểu đồ kết quả quan trắc DO vùng BĐCM (Mùa khô và mùa mưa 2016)



Hình 3.2: Biểu đồ kết quả quan trắc BOD_5 vùng BĐCM (Mùa khô và mùa mưa 2016)

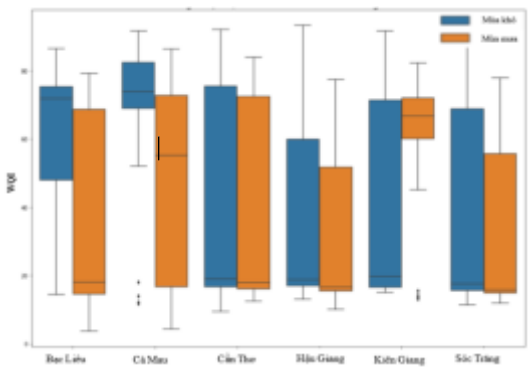


Hình 3.3: Biểu đồ kết quả quan trắc NH_4^+ vùng BĐCM (Mùa khô và mùa mưa 2016)



Hình 3.4: Biểu đồ tổng Coliform vùng BĐCM (Mùa khô và mùa mưa 2016)

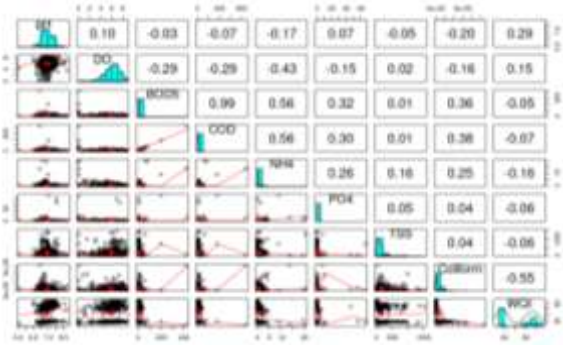
Qua kết quả tính toán VN_WQI có thể thấy các vị trí bị ô nhiễm nặng chiếm khoảng 50%: vùng phía Đông Bắc (Thành phố Cần Thơ); vùng Đông Nam (Sóc Trăng – Bạc Liêu); vùng phía Tây (huyện An Minh, An Biên tỉnh Kiên Giang) và vùng giữa Bán đảo (Vị Thanh, Cà Mau). Trong đó, các tuyến kênh bị ảnh hưởng bởi ô nhiễm là Kênh Cà Mau – Bạc Liêu, Quản Lộ - Phụng Hiệp; các kênh rạch trong đô thị của các đô thị lớn (Cần Thơ và Sóc Trăng). Đối với các tỉnh trong vùng nghiên cứu, biên độ dao động của VN_WQI cũng rất khác nhau, các tỉnh Hậu Giang, Sóc Trăng, Kiên Giang và Cần Thơ có giá trị WQI thấp và biên độ dao động lớn (đồng nghĩa với ô nhiễm nghiêm trọng hơn); 2 tỉnh Bạc Liêu và Cà Mau có giá trị VN_WQI lớn hơn 50, nên chất lượng nước mặt không bị ô nhiễm nghiêm trọng. Tuy nhiên, tỉnh Cà Mau có một vài điểm ô nhiễm cục bộ (chủ yếu ở thành phố Cà Mau do nước thải từ sinh hoạt và công nghiệp). Đối với các vị trí bị ảnh hưởng mặn (độ mặn > 3 g/l) cần phải có biện pháp xử lý hoặc không dùng cho sinh hoạt và sản xuất nông nghiệp (xem Hình 3.5).



Hình 3.5: Biểu đồ WQI của các tỉnh trong vùng BĐCM (tháng 4 và 10/2016)

3.2. Ứng dụng phương pháp Bayes (BMA) lựa chọn các thông số phục vụ xây dựng mô hình học máy

Theo kết quả quan trắc chất lượng nước ở mục 3.1 có rất nhiều thông số chất lượng nước là lý, hóa và vi sinh (pH, TSS, DO, BOD_5, COD, NH_4, PO_4, Coliform) quyết định đến ô nhiễm, tức là đến chất lượng nước (ở đây là giá trị WQI). Để xác định được các thông số đặc trưng phục vụ mô phỏng chất lượng nước trong vùng BĐCM, nghiên cứu đã dùng phương pháp Bayes để xác định được những biến (thông số chất lượng nước) có ảnh hưởng lớn đến WQI (Hình 3.6).



Hình 3.6: Biểu đồ tương quan của các thông số chất lượng nước và WQI

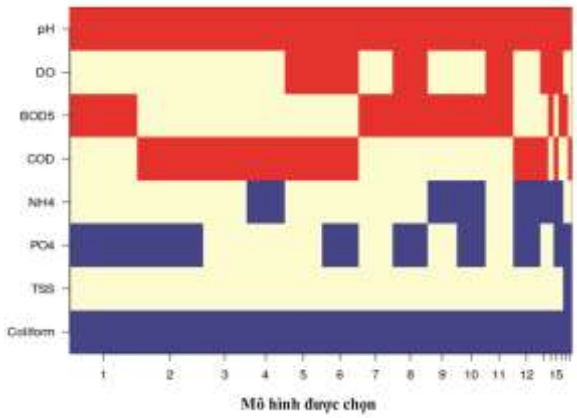
Theo Hình 3.6, mối tương quan giữa WQI và các thông số không cao, mức tương quan đáng kể với một số thông số như Coliform (0,55); tiếp đến là pH (0,29); NH_4 (0,16); DO (0,15) do vậy việc chọn các thông số (tối ưu) để tính toán WQI bằng mô hình học máy sẽ khó khăn. Để chọn được các thông số (tối ưu), nghiên cứu đã áp dụng phương pháp Bayes (BMA), kết quả phân tích thống kê bằng BMA đã xác định được các thông số chất lượng nước có ảnh hưởng lớn đến giá trị WQI là DO, COD, BOD_5, NH_4^+ và tổng Coliform. Từ đó đã xác định được các thông số chính ảnh hưởng đến WQI Hình 3.7).

- Theo kết quả phân tích theo phương pháp BMA, xác suất xuất hiện (theo mô hình được chọn) của từng thông số ảnh hưởng đến WQI như sau: pH, Coliform (100%); PO4 (55,3%); COD (52,8), BOD_5 (47,2%); DO (31,8%); NH_4^+ (29,1%) và TSS (1,6%);

- Có 5 mô hình tối ưu được chọn như sau (Hình 3.6).

  - Mô hình 1: pH, BOD_5, PO_4 và Coliform (tần suất hậu định là 13,4%);
  - Mô hình 2: pH, COD, PO_4 và Coliform (tần suất hậu định là 13,2%);
  - Mô hình 3: pH, COD và Coliform (xác suất hậu định là 8,7%);
  - Mô hình 4: pH, COD, NH_4 và Coliform (xác suất hậu định là 7,5%);
  - Mô hình 5: pH, DO, COD và Coliform (xác suất hậu định là 7,5%).

Qua phân tích ở trên cho thấy, mô hình 1 là mô hình tối ưu nhất vì có tần suất hậu định cao nhất. Do vậy chọn mô hình 1 để tính toán (dự báo) WQI bằng các thuật toán học máy (sẽ được thực hiện trong Mục 3.3).



Hình 3.7: Biểu đồ lựa chọn thông số chất lượng nước quan trọng theo BMA

3.3. Nghiên cứu tính toán chỉ số chất lượng nước mặt bằng phương pháp học máy cho vùng BĐCM

3.3.1. Tiêu chí đánh giá các mô hình học máy

Các tiêu chí đánh giá (hiệu chỉnh) các mô hình học máy được trình bày trong các công thức (1) đến (4):

- Sai số trung bình tuyệt đối (MAE): là 1 chỉ số phổ biến để tính sai số nhằm đánh giá (kiểm định) mô hình đối với các biến liên tục, được xác định theo công thức (1). Trong đó, Pi là giá trị dự báo và Mi là giá trị thực đo. Giá trị MAE càng thấp thì kết quả tính toán càng chính xác.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Pi - Mi|. \quad (1)$$

- Sai số toàn phương trung bình (MSE) của một phép ước lượng là trung bình của bình phương các sai số, tức là sự khác biệt giữa các giá trị dự đoán và giá trị thực đo và được tính toán theo theo công thức (2). Giá trị MSE càng thấp thì kết quả tính toán càng chính xác.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \quad (2)$$

- RMSE là căn bậc hai của trung bình của các sai số bình phương. RMSE là thước đo mức độ dàn trải của những phần dư này, nói cách khác, nó cho bạn biết mức độ tập trung của dữ liệu xung quanh đường phù hợp nhất. RMSE là độ lệch chuẩn của các phần dư (sai số dự đoán) và được tính toán theo theo công thức (3). Giá trị RMSE càng thấp thì kết quả tính toán càng chính xác.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Q_A^i - Q_p^i)^2}. \quad (3)$$

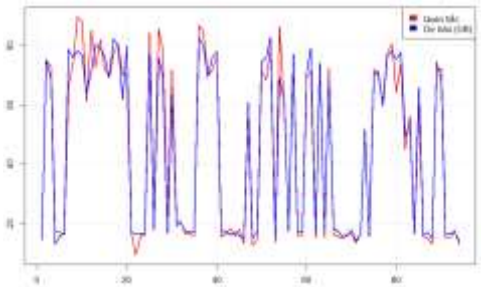
- Hệ số xác định (R^2): phản ánh phần trăm phương sai của y có thể giải thích bởi mô hình được xác định theo theo công thức (4). Trong đó, ESS là tổng các độ lệch bình phương của phần dư; TSS là tổng các độ lệch bình phương. Giá trị R^2 dao động từ 0 đến 1, giá trị R^2 càng gần 1 thì kết quả tính toán càng chính xác.

$$R^2 = 1 - (ESS/TSS) \quad (4)$$

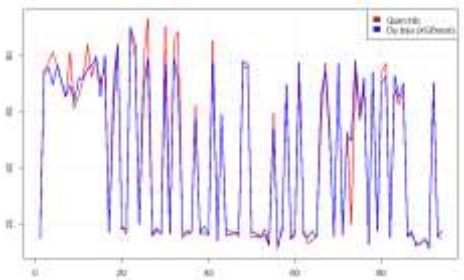
3.3.2. Kết quả đánh giá các mô hình học máy

Việc xây dựng mô hình học máy theo 5 bước chính như sau:

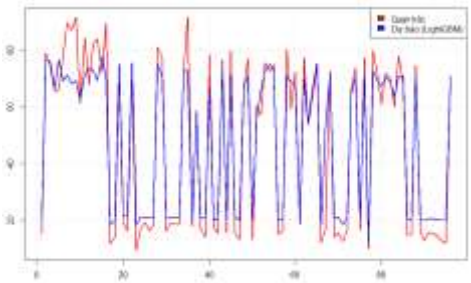
Căn cứ vào kết quả lựa chọn các thông số phục vụ xây dựng mô hình học máy bằng phương pháp Bayes (BMA), nghiên cứu đã lựa chọn Mô hình 1 với 4 thông số pH, BOD_5, PO_4 và Coliform để làm số liệu đầu vào dự báo WQI theo 4 thuật toán (mô hình) học máy là Tăng cường độ dốc, Tăng cường độ dốc cực đại, Tăng cường độ dốc nhẹ và Cây quyết định. Kết quả dự báo WQI và các biểu đồ so sánh giữa giá trị WQI dự báo và thực đo đối với tập số liệu thử nghiệm theo 4 mô hình học máy khác nhau được trình bày trong Hình 3.8.



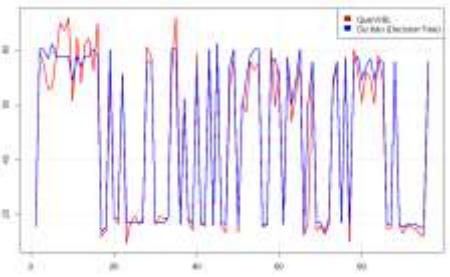
a) Mô hình hồi quy tăng cường độ dốc



b) Mô hình hồi quy tăng cường độ dốc cực đại



c) Mô hình hồi quy tăng cường độ dốc nhẹ



d) Mô hình hồi quy cây quyết định

Hình 3.8: Biểu đồ so sánh giữa giá trị WQI dự báo và thực đo đối với tập số liệu thử nghiệm

theo 4 mô hình học máy khác nhau

Kết quả đánh giá (dựa vào 4 tiêu chí) các mô hình học máy dự báo chỉ số chất lượng nước mặt vùng BĐCM được trình bày trong Bảng 3.2. Theo đó mô hình Tăng cường độ dốc có kết quả dự báo chính xác nhất vì có hệ số xác định R^2 cao nhất (0,973), giá trị các sai số MAE, MSE và RMSE thấp nhất (3,24; 22,54; 4,75). Tiếp đến là mô hình Tăng cường độ dốc cực đại có R^2 là 0,966 và giá trị các sai số tương ứng (3,15; 28,95; 5,38). Mô hình Cây quyết định có R^2 là 0,944; giá trị các sai số là 4,46; 49,67; 7,04; Mô hình Tăng cường độ dốc nhẹ có R^2 là 0,928; giá trị các sai số là 5,95; 63,30; 7,95). Có thể thấy, các mô hình học máy được áp dụng trong nghiên cứu này đều có thể dự đoán tốt WQI cho khu vực nghiên cứu (hệ số xác định rất cao, lớn hơn 0,9). Đây là cơ sở khoa học vững chắc và cũng là kết quả quan trọng để có thể ứng dụng các mô hình học máy trong tính toán WQI cho các vùng khác có điều kiện tương tự như vùng BĐCM, nhất là trong điều kiện khó khăn trong công tác quan trắc đầy đủ các thông số chất lượng nước để phục vụ tính toán WQI theo phương pháp truyền thống.

Bảng 3.2: Bảng thống kê kết quả đánh giá các mô hình học máy dự báo chỉ số chất lượng nước mặt vùng BĐCM

Mô hình	Các thông số đầu vào	Thông số đầu ra	Tiêu chí đánh giá			
			MAE	MSE	RMSE	R^2
Gradient Boosting	pH, BOD_5, PO_4, Coliform	WQI	3,24	22,54	4,75	0,973
XGBoost	pH, BOD_5, PO_4, Coliform	WQI	3,15	28,95	5,38	0,966
LightGBM	pH, BOD_5, PO_4, Coliform	WQI	5,95	63,30	7,95	0,928
Decision Tree	pH, BOD_5, PO_4, Coliform	WQI	4,46	49,67	7,04	0,944

4. KẾT LUẬN

Kết quả nghiên cứu đã xây dựng được cơ sở khoa học tính toán chỉ số chất lượng nước mặt bằng phương pháp học máy và đề xuất được phương pháp tính toán chỉ số chất lượng nước mặt bằng phương pháp học máy thích hợp với điều kiện thực tế của các địa phương trong vùng Bán đảo Cà Mau. Trong đó, nghiên cứu đã ứng dụng phương pháp Bayes (BMA) để lựa chọn các thông số (tối ưu) phục vụ xây dựng mô hình học máy tính toán WQI với 4 thông số chính là pH, BOD_5, PO_4, Coliform (ít và tối ưu hơn so với phương pháp truyền thống).

Theo kết quả tính toán (dự báo) WQI của các mô hình học máy cho thấy rằng mô hình tăng cường độ dốc (Gradient Boosting) có kết quả dự báo chính xác nhất, tiếp đến là mô hình tăng cường độ dốc cực đại (XGBoost), Mô hình Cây quyết định (Decision Tree) và mô hình tăng cường độ dốc nhẹ (LightGBM). Tất cả các mô hình này có độ chính xác cao, từ 92,8% đến 97,3%.

Như vậy, 2 loại mô hình học máy tăng cường và cây quyết định đều có thể tính toán (dự báo) WQI cho khu vực nghiên cứu với độ chính xác cao, và có thể áp dụng cho các khu vực khác với điều kiện tương tự. Điều này sẽ giúp các địa phương cải thiện hơn trong công tác đánh giá và quản lý chất lượng nước mặt.

TÀI LIỆU THAM KHẢO

[1] Bộ Tài nguyên và Môi trường (2015), *Báo cáo môi trường Quốc gia 2015*, Hà Nội.

[2] Bộ Tài nguyên và Môi trường; (2018), *Báo cáo hiện trạng môi trường Quốc gia năm 2018*, Hà Nội.

[3] Lê Thị Siêng (2003), *Nghiên cứu diễn biến môi trường nước do hoạt động nuôi tôm ở các tỉnh Bạc Liêu, Cà Mau ảnh hưởng tới môi trường và đề xuất các biện pháp khắc phục*, Viện

Khoa học Thủy lợi Miền Nam, Hồ Chí Minh.

[4] Lê Thị Siêng (2006), *Nghiên cứu xây dựng loại hình nuôi tôm sú công nghiệp sử dụng các đối tượng sinh học để xử lý nguồn nước trong các ao nuôi và sau khi nuôi góp phần bảo vệ môi trường*, Viện Khoa học Thủy lợi Miền Nam, Hồ Chí Minh.

[5] Sở TN&MT Hậu Giang; (2016), *Báo cáo hiện trạng môi trường tỉnh Hậu Giang năm 2011 - 2015*, Hậu Giang.

[6] Sở TN&MT Bạc Liêu (2012), *Điều tra, khảo sát đánh giá tình hình ô nhiễm môi trường do hoạt động NTTS và xây dựng các giải pháp xử lý, giảm thiểu ô nhiễm trên địa bàn tỉnh Bạc Liêu*, Bạc Liêu.

[7] Sở TN&MT Bạc Liêu (2013), *Xây dựng chiến lược quản lý và bảo vệ môi trường vùng biển, ven biển tỉnh Bạc Liêu đến năm 2020 và định hướng đến năm 2030*, Bạc Liêu.

[8] Sở TN&MT Bạc Liêu (2016), *Báo cáo hiện trạng môi trường tỉnh Bạc Liêu năm 2011 - 2015*, Bạc Liêu.

[9] Sở TNMT Bạc Liêu (2012), *Điều tra, đánh giá hiện trạng và phân vùng xả nước thải vào các nguồn tiếp nhận trên địa bàn tỉnh Bạc Liêu*, Bạc Liêu.

[10] Sở TN&MT Cà Mau (2016), *Báo cáo hiện trạng môi trường tỉnh Cà Mau năm 2011 - 2015.*

[11] Sở TN&MT Cà Mau (2016), *Điều tra, khảo sát đánh giá hiện trạng ô nhiễm môi trường nuôi trồng thủy sản và đề xuất biện pháp xử lý*, Cà Mau.

[12] Sở TNMT Cà Mau (2016), *Điều tra, đánh giá và đề xuất quản lý tác nhân gây ô nhiễm môi trường nước vùng ven biển trên địa bàn tỉnh Cà Mau*, Cà Mau.

[13] Sở TNMT Cà Mau (2016), *Điều tra, khảo sát đánh giá hiện trạng ô nhiễm môi trường nuôi trồng thủy sản và đề xuất biện pháp xử lý*, Cà Mau.

[14] Sở TN&MT Cần Thơ; (2016), *Báo cáo hiện trạng môi trường tỉnh Cần Thơ năm 2011 - 2015*.

[15] Tăng Đức Thắng (2015), *Nghiên cứu các biện pháp khoa học công nghệ đánh giá và quản lý nguồn nước, giảm thiểu ô nhiễm trong các hệ thống thủy lợi ĐBSCL*, Viện khoa học Thủy lợi Miền Nam.

[16] Nguyễn Văn Tuấn (2020), *Mô hình hồi quy và khám phá khoa học*, Nhà xuất bản Tổng hợp, Thành phố Hồ Chí Minh.

[17] Viện Kỹ thuật Biển (2015), *Quy hoạch Tài nguyên nước tỉnh Sóc Trăng*, Tp. Hồ Chí Minh.

[18] Phạm Thế Vinh (2020), *Nghiên cứu đề xuất các giải pháp tổng thể cải thiện và bảo vệ môi trường nước phục vụ phát triển bền vững đồng bằng sông Cửu Long*, Viện Khoa học Thủy lợi miền Nam, Hồ Chí Minh.

[19] Anthony A. Adegoke và các cộng sự. (2018), "Epidemiological Evidence and Health Risks Associated With Agricultural Reuse of Partially Treated and Untreated Wastewater: A Review", *Frontiers in public health*. 6, tr. 337-337.

[20] Mahreen Ahmed, Rafia Mumtaz và Syed Mohammad (2021), "Analysis of water quality indices and machine learning techniques for rating water pollution: A case study of Rawal Dam, Pakistan", *Water Supply*. 21.

[21] Seyed Babak Haji Seyed Asadollah và các cộng sự. (2021), "River water quality index prediction and uncertainty analysis: A comparative study of machine learning models",

Journal of Environmental Chemical Engineering. 9(1), tr. 104599.

[22] Shine Bedi và các cộng sự. (2020), "Comparative evaluation of machine learning models for groundwater quality assessment", *Environmental Monitoring and Assessment*.

[23] Benjamin Bowes và các cộng sự. (2022), "Reinforcement learning-based real-time control of coastal urban stormwater systems to mitigate flooding and improve water quality", *Environmental Science: Water Research & Technology*. 8.

[24] Ali El Bilali, Abdeslam Taleb và Youssef Brouziyne (2021), "Groundwater quality forecasting using machine learning algorithms for irrigation purposes", *Agricultural Water Management*. 245, tr. 106625.

[25] Nabeel M. Gazzaz và các cộng sự. (2012), "Artificial neural network modeling of the water quality index for Kinta River (Malaysia) using water quality variables as predictors", *Marine Pollution Bulletin*. 64(11), tr. 2409-2420.

[26] Mohammed Hameed và các cộng sự. (2017), "Application of artificial intelligence (AI) techniques in water quality index prediction: a case study in tropical region, Malaysia", *Neural Computing and Applications*. 28.

[27] Manel Naloufi và các cộng sự. (2021), "Evaluating the Performance of Machine Learning Approaches to Predict the Microbial Quality of Surface Waters and to Optimize the Sampling Effort", *Water*. 13(18), tr. 2457.

[28] Al-Akhir Nayan và các cộng sự. (2020), *River Water Quality Analysis and Prediction Using GBM*, 219-224.

[29] Grey S. Nearing và các cộng sự. (2021), "What Role Does Hydrological Science Play in the Age of Machine Learning?", *Water Resources Research*. 57(3).

[30] Dao Nguyen Khoi và các cộng sự. (2022), "Using Machine Learning Models for Predicting the Water Quality Index in the La Buong River, Vietnam", *Water*. 14, tr. 1552.

[31] D. Venkata Vara Prasad và các cộng sự. (2022), "Analysis and prediction of water quality using deep learning and auto deep learning techniques", *Science of The Total Environment*. 821, tr. 153311.

[32] Neha Radhakrishnan và Anju Pillai (2020), *Comparison of Water Quality Classification Models using Machine Learning*, 1183-1188.

[33] Tiyasha, Tran Minh Tung và Zaher Mundher Yaseen (2020), "A survey on river water quality modelling using artificial intelligence models: 2000–2020", *Journal of Hydrology*. 585, tr. 124670.