**TÓM TẮT**

Mục tiêu của nghiên cứu nhằm ứng dụng một số thuật toán học máy, bao gồm Extreme Gradient Boosting (XGB), Gradient Boosting Model (GBM), Support Vector Regression (SVR), và Radial Basis Function (RBF), để dự đoán WQI tại ba vị trí quan trắc trên sông Sài Gòn trong giai đoạn từ năm 2015 đến 2019. Chỉ số WQI được tính toán dựa trên Hướng dẫn kỹ thuật tính toán và công bố Chỉ số chất lượng nước Việt Nam theo Quyết định số 1460/QĐ-TCMT ngày 12 tháng 11 năm 2019 của Tổng cục Môi trường, sử dụng tám thông số: DO, BOD₅, COD, N-NH₄⁺, P-PO₄³⁻, pH, nhiệt độ và tổng coliform. Kết quả so sánh hiệu quả giữa các mô hình cho thấy, mô hình XGB vượt trội hơn các mô hình còn lại khi sử dụng cả tám thông số làm dữ liệu đầu vào. Cụ thể, mô hình XGB đạt sai số thấp nhất (PBIAS = 0,399**,** RMSE = 1,630 và MAE = 0,782) và độ tương quan cao nhất (R² = 0,960 và NSE = 0,953), tiếp theo là các mô hình GBM, SVR, MLP và RBF. Nghiên cứu cũng chứng minh rằng hiệu suất mô hình giảm đáng kể khi loại bỏ các thông số N-NH₄⁺ và P-PO₄³⁻, trong khi việc loại bỏ COD hoặc BOD₅ làm giảm nhẹ khả năng dự báo. Đồng thời, kết quả cũng cho thấy việc loại bỏ cùng lúc hai thông số chất lượng nước trong quá trình dự báo WQI bằng mô hình học máy không phải là phương án tối ưu, do làm giảm đáng kể độ chính xác dự báo. Những phát hiện này cho thấy rằng các mô hình ML có thể mô phỏng tốt chỉ số WQI tại khu vực nghiên cứu, điều này giúp giảm thiểu sai sót trong quá trình tính toán chỉ số phụ theo phương pháp truyển thống. Bên cạnh đó, mô hình XGB còn có thể giảm thiểu số lượng thông số đầu vào cần thiết để dự đoán WQI mà vẫn đảm bảo khả năng mô phỏng đạt yêu cầu và nắm bắt hiệu quả mối quan hệ tiềm năng giữa các thông số đầu vào với WQI. Nhìn chung, nghiên cứu này cung cấp khung phương pháp cho việc mô phỏng WQI dựa trên các thuật toán ML chính xác và tối ưu, góp phần hỗ trợ công tác đánh giá và giám sát chất lượng nước tại các khu vực thiếu hệ thống giám sát tự động.