PREDICTING PH IN WATER RESERVOIRS USING MACHINE LEARNING TECHNIQUES: A COMPARATIVE STUDY ON WATER QUALITY INDICES

Kozlovska P., Krzemińska A., Miller T.
University of Szczecin. tymoteusz.miller@usz.edu.pl

Introduction. Water quality monitoring plays a crucial role in maintaining healthy ecosystems and ensuring the safety of water resources for various purposes, such as drinking, agriculture, and industry [1, 2]. One essential aspect of water quality assessment is the measurement of pH, which can provide valuable insights into the overall health of aquatic ecosystems. Accurate pH prediction is necessary for timely intervention and effective management of water resources [3].

In recent years, machine learning techniques have emerged as powerful tools for modeling complex relationships between variables and making accurate predictions [4, 5,6]. The aim of this work was to develop an accurate predictive model for pH using various water quality indices by employing and comparing multiple machine learning techniques, including Generalized Linear Model, Deep Learning, Decision Tree, Random Forest, Gradient Boosted Trees, and Support Vector Machine. The models' performance is assessed using several evaluation metrics, and the most suitable model is identified for predicting pH based on the given water quality indices.

By developing an accurate pH prediction model, this study contributes to the ongoing efforts in environmental monitoring and management, helping stakeholders assess water quality more effectively and identify potential issues in water reservoirs. Furthermore, the findings from this comparative study can inform researchers and practitioners about the performance of different machine learning techniques in predicting pH and guide the selection of appropriate models for various applications in water quality assessment.

Material and methods. Area. Rusalka Lake, a small artificial water reservoir (2.9 ha) located in Kasprowicz Park, Szczecin, is formed by damming the waters of the Osowka and Warszowiec streams. The lake is elongated in shape, with a maximum depth of 1.7 m before sediment dredging and 3.0 m after. Average depth increased from 1.4 m to 2.0 m following dredging. Despite its lake-like appearance, Rusalka Lake's water retention time is approximately 30 days, making it more similar to a stream flood area. The lake is part of Szczecin's municipal storm drainage system, functioning as a storage reservoir, with water drained to the Western Odra River through an underground canal.

Methods. Water samples were collected monthly from January 2015 to December 2022 to analyze various water quality indices. These indices included Chlorophyll a (Chl a), Redox Potential (Eh), Temperature (Temp.), Chemical Oxygen Demand-Manganese (COD-Mn), Chemical Oxygen Demand-Chromium (COD-Cr), Biological Oxygen Demand (BOD), Dissolved Oxygen (DO), Water Saturation (WS), Nitrate (NO3), Nitrite (NO2), Ammonium (NH4), Total Nitrogen (TN), Soluble Reactive Phosphorus (SRP), Total Phosphorus (TP), Total Hardness (TH), Calcium (Ca), Magnesium (Mg), Chloride (Cl), Sulfate (SO4), Iron (Fe), Lead (Pb), Zinc (Zn), Cadmium (Cd), and Copper (Cu). The analysis followed the guidelines established by the American Public Health Association (APHA).

We employed various machine learning techniques, including Generalized Linear Model, Deep Learning, Decision Tree, Random Forest, Gradient Boosted Trees, and Support Vector Machine, using Rapid Miner Studio 10.1 with an academic license. To compare the performance of these models, we evaluated them based on several metrics such as Relative Error, Standard Deviation, Gains, Total Time, Training Time (1,000 Rows), Scoring Time, Root Mean Squared Error, Absolute Error, Lenient Relative Error, and Squared Error.

Results and discussion. The performance of different models was evaluated using various metrics. The results were as follows:

Model Performance:

1. Generalized Linear Model: Relative Error = 0.4, Total Time = 6047.0, Training Time (1,000 Rows) = 32.7, Scoring Time = 104.6
2. Deep Learning: Relative Error = 0.2, Total Time = 12145.0, Training Time (1,000 Rows) = 349.8, Scoring Time = 69.9
3. Decision Tree: Relative Error = 0.1, Total Time = 4943.0, Training Time (1,000 Rows) = 12.5, Scoring Time = 39.5
4. Random Forest: Relative Error = 0.3, Total Time = 42637.0, Training Time (1,000 Rows) = 75.2, Scoring Time = 65.7
5. Gradient Boosted Trees: Relative Error = 0.1, Total Time = 130547.0, Training Time (1,000 Rows) = 545.3, Scoring Time = 151.3
6. Support Vector Machine (SVM): Relative Error = 0.1, Total Time = 740183.0, Training Time (1,000 Rows) = 5459.6, Scoring Time = 708.0

Support Vector Machine (Kernel Model) details:

- Total number of Support Vectors: 6007
- Bias (offset): 6.494
- Weight coefficients (w) for each water quality index were provided (table 1).

Table 1. Weight coefficients

w[Chl a] = 95.326	w[SRP] = -1319.678
w[Eh] = -98.492	w[TP] = 260.709
w[Temp.] = -348.134	w[TH] = 1363.217
w[COD-Mn] = -97.776	w[Ca] = -780.814
w[COD-Cr] = -331.585	w[Mg] = 1183.195
w[BOD] = -2.578	w[Cl] = -402.718
w[DO] = 39.341	w[SO4] = 147.365
w[WS] = -234.226	w[Fe] = -440.799
w[NO3] = -812.154	w[Pb] = 41.708
w[NO2] = 244.132	w[Zn] = 191.438
w[NH4] = -753.425	w[Cd] = -685.026
w[TN] = -0.489	w[Cu] = -51.695

Evaluation Metrics:

- Root Mean Squared Error: 0.995 +/- 0.117 (micro average: 1.001 +/- 0.000)
- Absolute Error: 0.635 +/- 0.037 (micro average: 0.635 +/- 0.773)
- Lenient Relative Error: 10.01% +/- 0.29% (micro average: 10.01% +/- 10.38%)
- Squared Error: 1.001 +/- 0.242 (micro average: 1.001 +/- 4.203)
- Correlation: 0.984 +/- 0.004 (micro average: 0.984)

Optimal parameters for SVM (Gamma and C values):

- Best performance was obtained with GAMMA (RBF) = 0.5 and C = 10.0, with a performance score of 0.11497013552915694.

Based on the results, the Decision Tree, Gradient Boosted Trees, and Support Vector Machine models had the lowest relative error (0.1) among the tested models. However, the Support Vector Machine demonstrated the best performance considering the optimal parameters and performance score

However, the Support Vector Machine model, with optimal GAMMA (RBF) = 0.5 and C = 10.0, achieved the best performance score of 0.11497013552915694. This indicates that the Support Vector Machine model is the most accurate and suitable for predicting pH based on the given water quality indices.

The other models, such as the Generalized Linear Model and Deep Learning, had higher relative errors, which implies less accurate predictions. The Random Forest model performed better than the Generalized Linear Model and Deep Learning, but it was not as accurate as the Support Vector Machine.

It is worth noting that the Support Vector Machine model had the longest total time and training time (1,000 rows) among all models. This could be a potential drawback when considering computational efficiency and scalability, especially for large datasets. However, the model's superior predictive performance may justify the trade-off in computational resources.

In conclusion, the Support Vector Machine model, with optimal parameters, demonstrated the best performance in predicting pH based on the water quality indices. This model can potentially be applied in various environmental monitoring and management applications, helping to assess water quality and identify potential issues more effectively. Further research could explore other machine learning techniques or feature selection methods to improve model performance and computational efficiency.

Conclusion. In conclusion, this study aimed to develop an accurate predictive model for pH using various water quality indices. Several machine learning techniques were employed and compared, including Generalized Linear Model, Deep Learning, Decision Tree, Random Forest, Gradient Boosted Trees, and Support Vector Machine. Based on the evaluation metrics, the Support Vector Machine model with optimal parameters (GAMMA (RBF) = 0.5 and C = 10.0) demonstrated the best performance in predicting pH.

Future research could explore other machine learning techniques or feature selection methods to further enhance model performance and computational efficiency. Additionally, the developed models can be tested on other water bodies to validate their generalizability and applicability in different environmental contexts.

References:

1. Gómez-Gutiérrez A, Miralles MJ, Corbella I, García S, Navarro S, Llebaria X. La calidad sanitaria del agua de consumo [Drinking water quality and safety]. Gac Sanit. 2016 Nov;30 Suppl 1:63-68

2. William Schaduw JN. Seawater Quality Analysis in Mantehage Island for Integrated and Sustainable Marine Tourism Development. Pak J Biol Sci. 2021 Jan;24(12):1333-1339.

3. Bundschuh M, Weyers A, Ebeling M, Elsaesser D, Schulz R. Narrow pH Range of Surface Water Bodies Receiving Pesticide Input in Europe. Bull Environ Contam Toxicol. 2016 Jan;96(1):3-8.

4. Zampieri G, Vijayakumar S, Yaneske E, Angione C. Machine and deep learning meet genome-scale metabolic modeling. PLoS Comput Biol. 2019 Jul 11;15(7):e1007084.

5. Manco L, Maffei N, Strolin S, Vichi S, Bottazzi L, Strigari L. Basic of machine learning and deep learning in imaging for medical physicists. Phys Med. 2021 Mar;83:194-205.

6. Puthongkham P, Wirojsaengthong S, Suea-Ngam A. Machine learning and chemometrics for electrochemical sensors: moving forward to the future of analytical chemistry. Analyst. 2021 Oct 25;146(21):6351-6364.