# Prediction of Water Quality Classification using Machine Learning

Subjects: Computer Science, Artificial Intelligence

Contributor: Nur Hanisah Abdul Malek, Wan Fairos Wan Yaacob, Syerina Md Nasir

Machine Learning (ML) has been used for a long time and has gained wide attention over the last several years. It can handle a large amount of data and allow non-linear structures by using complex mathematical computations. However, traditional ML models do suffer some problems, such as high bias and overfitting. Therefore, this has resulted in the advancement and improvement of ML techniques, such as the bagging and boosting approach, to address these problems.

Keywords: water quality class ; water quality index ; supervised machine learning ; random forest ; gradient boosting ; decision tree

## 1. Introduction

Water pollution is a critical issue in Malaysia with a negative impact on water resources sustainability, which can cause an inadequate water supply to all people even though a large number of water resources are available [1]. The most important natural resource issue that humanity will have to address in the 21st century is water [2]. The combined impacts of human activities and climate change have resulted in significant changes in the run-off from many rivers and increasing water scarcity [2]. Water scarcity not only poses a threat to human life and social development, but also has a significant impact on the Gross Domestic Product [3]. To reduce the impact of water pollution, the monitoring and assessment of river water quality is crucial.

Water Quality Index (WQI) is an index that can represent the overall water quality status with a single score of the subindex based on six parameters, which are Dissolved Oxygen (DO) in percentage of saturation, Biochemical Oxygen Demand, Ammoniacal Nitrogen, pH, Total Suspended Solid (TSS) and Chemical Oxygen Demand [4]. It ranges from 0 to 100 and indicates the class of the water, whether it is clean, slightly polluted or polluted. If the WQI falls within the range of 81 to 100%, the river water status is classified as ‘clean’, a range between 60 to 80% as ‘slightly polluted’ and a range 0 to 59% as ‘polluted’ [4].

## 2. Machine Learning in Prediction of Water Quality Classification

Many studies have been conducted to address water quality problems. Most works employ manual laboratory analysis and statistical analysis to assist in regulating water quality [5][6][7], while other studies use Machine Learning methods to help to obtain optimized solutions to water quality problems [8][9][10][11][12]. A local researcher that used laboratory analysis has contributed to the understanding on the issue of water quality in Malaysia. Alias [5] collected water samples from 11 stations along the Pengkalan Chepa river basin, Kelantan, and analyzed them using Multi-Probe System for in situ tests and manual laboratory analysis for ex situ tests. It was found that the river was slightly polluted due to anthropogenic activities. Al-Badaii et al. [6] collected water samples from eight stations along the Semenyih river, Selangor, and analyzed them using manual laboratory analysis. They found that the Semenyih river was slightly polluted by suspended solids, nitrogen, ammoniacal nitrogen ($NH_3N$) and chemical oxygen demand (COD). Moreover, the river was extremely polluted with fecal coliform and phosphorus. This encouraged the further exploration of Machine Learning methodologies in the field of water quality.

Many works had been conducted to predict water quality using Machine Learning (ML) approaches. Some researchers used the traditional Machine Learning models, such as Decision Tree [13][14], Artificial Neural Network [12][15][16][17], Support Vector Machine [18][19][20], K-Nearest Neighbors [21] and Naïve Bayes [18][22][23]. However, in recent years, some researchers are moving towards more advanced ML ensemble models, such as Gradient Boosting and Random Forest [10][24][25][26][27].

Traditional Machine Learning models, such as the Decision Tree model, are frequently found in the literature and performed well on water quality data. However, decision-tree-based ensemble models, including Random Forest (RF) and Gradient Boosting (GB), always outperform the single decision tree [24]. Among the reasons for this are its ability to manage both regular attributes and data, not being sensitive to missing values and being highly efficient. Compared to other ML models, decision-tree-based models are more favorable to short-term prediction and may have a quicker calculation speed [26]. Gakii and Jepkoech [13] compared five different decision tree classifiers, which are Logistic Model Tree (LMT), J48, Hoeffding tree, Random Forest and Decision Stump. They found that J48 showed the highest accuracy of 94%, while Decision Stump showed the lowest accuracy. Another study by Jeihouni et al. [14] also compared five decision-tree-based models, which are Random Tree, Random Forest, Ordinary Decision Tree (ODT), Chi-square Automatic Interaction Detector and Iterative Dichotomiser 3 (ID3), to determine high water quality zones. They found that ODT and Random Forest produce higher accuracy compared to the other algorithms and the methods are more suitable for continuous datasets.

Another popular Machine Learning model to predict water quality is Artificial Neural Network (ANN). ANN is a remarkable data-driven model that can cater both linear and non-linear associations among output and input data. It is used to treat the non-linearity of water quality data and the uncertainty of contaminant source. However, the performance of ANN can be obstructed if the training data are imbalanced and when all initial weights of the parameter have the same value. In India, Aradhana and Singh [8] used ANN algorithms to predict water quality. They found that Lavenberg Marquardt (LM) algorithm has a better performance than the Gradient Descent Adaptive (GDA) algorithm. Abyaneh [5] used ANN and multivariate linear regression models in his research and found that the ANN model outperforms the MLR model. However, the research only assessed the performance of the ANN model using root-mean-square error (RMSE), coefficient of correlation (r) and bias values. Although ANN models are the most broadly used, they have a drawback as the prediction power becomes weak if they are used with a small dataset and the testing data are outside the range of the training data [28].

Support Vector Machine has also been extensively used in water quality studies. Some studies proved that SVM is the best model in predicting water quality compared to other models. A study by Babbar and Babbar [11] found that Support Vector Machine and Decision Tree are the best classifiers because they have the lowest error rate, which is 0%, in classifying water quality class compared to ANN, Naive Bayes and K-NN classifiers. It also revealed that ML models can quickly determine the water quality class if the data provided represent an accurate representation of domain knowledge. In China, Liu and Lu [12] developed the SVM and ANN model to predict phosphorus and nitrogen. They found that SVM model achieves a better forecasting accuracy compared to the ANN model. This is because the SVM model optimizes a smaller number of parameters acquired from the principle of structural risk minimization, hence avoiding the occurrence of overtraining data to have a better generalization ability [12]. This is supported by another study in Eastern Azerbaijan, Iran [16]. They found that SVM has a better performance compared to the K-Nearest Neighbor algorithm in estimating two water quality parameters, which are total dissolved solid and conductivity. The results showed smaller error and higher $R^2$ than the results attained in Abbasi et al.'s report [4]. Naïve Bayes has also been widely used for predicting water quality. A study by Vijay and Kamaraj [22] found that Random Forest and Naïve Bayes produce better accuracy and low classification error compared to the C5.0 classifier. However, traditional ML models, for example, Decision Tree, ANN, Naïve Bayes and SVM, do not perform well. They have some weaknesses, such as a high tendency to be biased and a high variance [22]. For example, SVM uses the structural risk minimization principle to address overfitting problem in Machine Learning by reducing the model's complexity and fitting the training data successfully [29]. Meanwhile, the Bayes model uses prior and posterior probabilities in order to prevent overfitting problems and bias from using only sample information. In ANN, the training process takes a longer time and overfitting problems may occur if there are too many layers, while the prediction error may be affected if there are not enough layers [30]. Overfitting is a fundamental issue in supervised Machine Learning that prevents the perfect generalization of the model to fit the data observed on the training data, as well as unseen data on the testing set. Hence, overfitting occurs due to the presence of noise, a limited training set size, and classifier complexity [30]. One of the strategies considered by many previous works to reduce the effects of overfitting is to adopt more advanced methods, such as the ensemble method.

The ensemble method is a Machine Learning technique that combines several base learners' decisions to produce a more precise prediction than what can be achieved with having each base learner's decision [16]. This method has also gained wide attention among researchers recently. The diversity and accuracy of each base learner are two important features to make the ensemble learners work properly [17]. The ensemble method ensures the two features in several ways based on its working principle. There are two commonly used ensemble families in Machine Learning, which are bagging and boosting. Both the bagging and boosting methods provide a higher stability to the classifiers and are good in reducing variance. Boosting can reduce the bias, while bagging can solve the overfitting problem [31]. A famous ensemble model that uses the bagging algorithm is Random Forest. It is a classification model that uses multiple base models, typically

decision trees, on a given subset of data independently and makes decisions based on all models [25]. It uses feature randomness and bagging when building each individual decision tree to produce an independent forest of trees. Random Forest carries all the advantages of a decision tree with the added effectiveness of using several models [32]. Another popular ensemble model is Gradient Boosting. Gradient Boosting is a Machine Learning technique that trains multiple weak classifiers, typically decision trees, to create a robust classifier for regression and classification problems. It assembles the model in a stage-wise way similar to other boosting techniques and it generalizes them by optimizing a suitable cost function. In the GB algorithm, incorrectly classified cases for a step are given increased weight during the next step. The advantages of GB are that it has exceptional accuracy in predicting and fast process [33]. Therefore, advanced models, such as Random Forest and Gradient Boosting, should be employed to cater for the lack of basic ML models.

## References


1. Ling, J.K.B. Water Quality Study and Its Relationship with High Tide and Low Tide at Kuantan River. Bachelor's Thesis, Universiti Malaysia Pahang, Gambang, Malaysia, 2010. Available online: http://umpir.ump.edu.my/id/eprint/2449/1/JACKY_LING_KUO_BAO.PDF (accessed on 22 February 2022).
2. Xu, J.; Gao, X.; Yang, Z.; Xu, T. Trend and Attribution Analysis of Runoff Changes in the Weihe River Basin in the Last 50 Years. Water 2022, 14, 47.
3. Wahab, M.A.A.; Jamadon, N.K.; Mohmood, A.; Syahir, A. River Pollution Relationship to the National Health Indicated by Under-Five Child Mortality Rate: A Case Study in Malaysia. Bioremediat. Sci. Technol. Res. 2015, 3, 20–25.
4. Abbasi, T.; Abbasi, S.A. Water Quality Indices; Elsevier: Amsterdam, The Netherlands, 2012.
5. Abyaneh, H.Z. Evaluation of multivariate linear regression and artificial neural networks in prediction of water quality parameters. J. Environ. Health Sci. Eng. 2014, 12, 40.
6. Alias, S.W.A.N. Ecosystem Health Assessment of Sungai Pengkalan Chepa Basin: Water Quality and Heavy Metal Analysis. Sains Malays. 2020, 49, 1787–1798.
7. Al-Badaii, F.; Shuhaimi-Othman, M.; Gasim, M.B. Water quality assessment of the Semenyih river, Selangor, Malaysia. J. Chem. 2013, 2013, 871056.
8. Asadollah, S.B.H.S.; Sharafati, A.; Motta, D.; Yaseen, Z.M. River water quality index prediction and uncertainty analysis: A comparative study of machine learning models. J. Environ. Chem. Eng. 2021, 9, 104599.
9. Chen, K.; Chen, H.; Zhou, C.; Huang, Y.; Qi, X.; Shen, R.; Liu, F.; Zuo, M.; Zou, X.; Wang, J. Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data. Water Res. 2020, 171, 115454.
10. Lerios, J.L.; Villarica, M.V. Pattern Extraction of Water Quality Prediction Using Machine Learning Algorithms of Water Reservoir. Int. J. Mech. Eng. Robot. Res. 2019, 8, 992–997.
11. Sengorur, B.; Koklu, R.; Ates, A. Water quality assessment using artificial intelligence techniques: SOM and ANN—A case study of Melen River Turkey. Water Qual. Expo. Health 2015, 7, 469–490.
12. Aradhana, G.; Singh, N.B. Comparison of Artificial Neural Network algorithm for water quality prediction of River Ganga. Environ. Res. J. 2014, 8, 55–63.
13. Gakii, C.; Jepkoech, J. A Classification Model for Water Quality analysis Using Decision Tree. Eur. J. Comput. Sci. Inf. Technol. 2019, 7, 1–8.
14. Jeihouni, M.; Toomanian, A.; Mansourian, A. Decision tree-based data mining and rule induction for identifying high quality groundwater zones to water supply management: A novel hybrid use of data mining and GIS. Water Resour. Manag. 2020, 34, 139–154.
15. Ahmad, Z.; Rahim, N.A.; Bahadori, A.; Zhang, J. Improving water quality index prediction in Perak River basin Malaysia through a combination of multiple neural networks. Int. J. River Basin Manag. 2017, 15, 79–87.
16. Gazzaz, N.M.; Yusoff, M.K.; Aris, A.Z.; Juahir, H.; Ramli, M.F. Artificial neural network modeling of the water quality index for Kinta River (Malaysia) using water quality variables as predictors. Mar. Pollut. Bull. 2012, 64, 2409–2420.
17. Hameed, M.; Sharqi, S.S.; Yaseen, Z.M.; Afan, H.A.; Hussain, A.; Elshafie, A. Application of artificial intelligence (AI) techniques in water quality index prediction: A case study in tropical region, Malaysia. Neural Comput. Appl. 2017, 28, 893–905.
18. Babbar, R.; Babbar, S. Predicting river water quality index using data mining techniques. Environ. Earth Sci. 2017, 76, 1–15.

19. Liu, M.; Lu, J. Support vector machine—An alternative to artificial neuron network for water quality forecasting in an agricultural nonpoint source polluted river? Environ. Sci. Pollut. Res. 2014, 21, 11036–11053.
20. Mohammadpour, R.; Shaharuddin, S.; Chang, C.K.; Zakaria, N.A.; Ab Ghani, A.; Chan, N.W. Prediction of water quality index in constructed wetlands using support vector machine. Environ. Sci. Pollut. Res. 2015, 22, 6208–6219.
21. Sattari, M.T.; Joudi, A.R.; Kusiak, A. Estimation of Water Quality Parameters with Data—Driven Model. J.-Am. Water Work. Assoc. 2016, 108, E232–E239.
22. Vijay, S.; Kamaraj, K. Ground Water Quality Prediction using Machine Learning Algorithms in R. Int. J. Res. Anal. Rev. 2019, 6, 743–749.
23. Muhammad, S.Y.; Makhtar, M.; Rozaimee, A.; Aziz, A.A.; Jamal, A.A. Classification model for water quality using machine learning techniques. Int. J. Softw. Eng. Its Appl. 2015, 9, 45–52.
24. Bui, D.T.; Khosravi, K.; Tiefenbacher, J.; Nguyen, H.; Kazakis, N. Improving prediction of water quality indices using novel hybrid machine-learning algorithms. Sci. Total Environ. 2020, 721, 137612.
25. 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, 2210.
26. Lu, H.; Ma, X. Hybrid decision tree-based machine learning models for short-term water quality prediction. Chemosphere 2020, 249, 126169.
27. Naghibi, S.A.; Hashemi, H.; Berndtsson, R.; Lee, S. Application of extreme gradient boosting and parallel random forest algorithms for assessing groundwater spring potential using DEM-derived factors. J. Hydrol. 2020, 589, 125197.
28. Khosravi, K.; Mao, L.; Kisi, O.; Yaseen, Z.M.; Shahid, S. Quantifying hourly suspended sediment load using data mining models: Case study of a glacierized Andean catchment in Chile. J. Hydrol. 2018, 567, 165–179.
29. Vapnik, V. The Nature of Statistical Learning Theory; Springer: New York, NY, USA, 1995.
30. Xu, T.; Coco, G.; Neale, M. A predictive model of recreational water quality based on adaptive synthetic sampling algorithms and machine learning. Water Res. 2020, 177, 115788.
31. Ahmed, S.; Mahbub, A.; Rayhan, F.; Jani, R.; Shatabda, S.; Farid, D.M. Hybrid methods for class imbalance learning employing bagging with sampling techniques. In Proceedings of the Computational Systems and Information Technology for Sustainable Solution (CSITSS), Bengaluru, India, 21–23 December 2017; pp. 1–5.
32. Liaw, A.; Wiener, M. Classification and regression by randomForest. R News 2002, 2, 18–22.
33. Prakash, R.; Tharun, V.P.; Devi, S.R. A Comparative Study of Various Classification Techniques to Determine Water Quality. In Proceedings of the Inventive Communication and Computational Technologies (ICICCT), Coimbatore, India, 20–21 April 2018; pp. 1501–1506.