Article

Not peer-reviewed version

Water Quality Class Modeling Using Machine Learning Algorithms at Roodeplaat Dam, South Africa

Kassahun Birhanu Tadesse * and Megersa Olumana Dinka

Posted Date: 14 June 2023

doi: 10.20944/preprints202306.1016.v1

Keywords: Decision Tree; linear regression; Naïve Bayes; Python; Support Vector Machine

Article

Water Quality Class Modeling Using Machine Learning Algorithms at Roodeplaat Dam, South Africa

Kassahun Birhanu Tadesse *[1] and Megerssa Olumana Dinka [1]

[1] Department of Civil Engineering Science, University of Johannesburg, Johannesburg, South Africa
* Correspondence: Author: kbirhan@gmail.com

Abstract: Water pollution is a common problem for dams situated within an urban or agricultural catchment. This can negatively affect the hydro ecosystem, drinking, recreational and other uses of water. In this study, the drinking water quality class of the Roodeplaat Dam, South Africa which faces pollution problems was modelled using machine learning algorisms in Python Jupyter Notebook 6.0.0. Eleven monthly water quality parameters recorded at five sampling stations from January 1981 to September 2017 were used for training and testing the model. Five machine learning classifiers: Gaussian Naïve Bayes (GNB), K-nearest neighbors (KNN), Decision Tree (DT), and Support Vector Machines (SVM), and Linear Regression (LR) at a test size of 20%, 25%, 30% and 40% were used to classify water into five classes (Excellent to Very bad). It was investigated that the dam water has only three classes of good, medium and bad. The prediction accuracies of machine learning algorithms from the highest to the lowest were 96.39%, 96.17%, 92.25%, 90.20, and 54.19% for KNN, DT, SVM, GNB, and LR, respectively. Therefore, KNN at test size of 30% was recommended to classify the water quality of Roodeplat Dam accurately. Hence, machine learning algorithms can be used to identify the class of water quality before the water is treated and distributed for drinking use.

Keywords: Decision Tree; linear regression; Naïve Bayes; Python; Support Vector Machine

1. Introduction

So many dams have been built in arid and semi-arid areas to maintain a continuous water supply for different uses such as drinking, recreation, ecology, and irrigation. To ensure sustainable use of the dams, water quality is a major concern that needs to be dealt with. Poor water quality in the dam can affect the human health, biodiversity, recreational use, and cost of water treatment. Due to intensive agriculture and urbanization around the dam, considerable amounts of wastewater, sewage and agricultural pollutants are released into reservoir systems, and are deteriorating water quality and limiting possible uses of water [1].

Hence, to monitor water quality status and to devise management strategies, many physicochemical parameters are sampled at a spatiotemporal basis. However, the measurement of large data sets involves huge inputs (money, time and energy), data analysis is complex and often hide important information. Scarce funding for water quality monitoring is initiating cost-efficient and thoughtful approaches for managing water quality. Hence, modeling water quality is of utmost importance for precise prediction of future water quality phenomenon. The traditional process-based modeling such as stochastic, deterministic or statistical techniques have limitations as they are costly, time-consuming and reliant on big datasets that entail unknown or unspecified data inputs [2].

The artificial intelligence (AI) models have been recognized as valuable options for modeling of nonlinear and complex water systems. AI are developed from input and output correlation without taking account of the internal process [2]. Different soft computing techniques such as AI and machine learning have been effectively used for modelling water quality [3–11].

The Roodeplaat Dam in South Africa is one of the typical examples for highly polluted dams situated in the urbanized areas. Many physicochemical parameters are sampled every day at different sites of the reservoir as part of regular water quality monitoring strategy. Water users are informed about potential health risks traditionally based on the concentration of individual parameters. This method is not easily understandable for all water users. In this study, the drinking water quality class of the Roodeplaat Dam, South Africa which faces pollution problems was modelled using different machine learning algorithms. Hence, the models prediction, for example, whether the water is excellent, good, medium, bad or very bad would be easily understandable for the public.

2. Methods

2.1. Study Area and Data

The study area is the Roodeplaat Dam situated at about 24 km northeast of Pretoria, South Africa. The dam was initially designed for irrigation, but later used for clean water supply and recreational sites around the dam. The water quality data recorded from 1981 to 2017 (instances of 2220) at five sampling stations, were collected in the format of the spreadsheet at the website of department of water affairs, South Africa. Eleven water quality parameters (Table 1): Calcium (Ca), Chloride (Cl), Total Dissolved Solids (TDS), Electrical Conductivity (EC), Fluoride (F), Potassium (K), Magnesium (Mg), Sodium (Na), Nitrate +Nitrite Nitrogen (NO3_N_NO2_N), pH, and Sulphate (SO4) were used for modelling. All the water quality parameters are expressed in milligram/liter (mg/L), except for pH (pH units) and EC (μSm^{-1}).

Table 1. Water Quality Parameters, weights, and thresholds.

Parameters	TDS	pH	EC	SO4	NO3	TAL	Na	Ca	Mg	K	Cl
Wi	5	4	3	3	3	3	3	2	2	2	2
Threshold*	450	9	400mS/cm	100	6	200	100	50	30	100	100

*Source: [12].

2.2. Modelling

2.2.1. Water Quality Index

A water quality index (WQI) is a means by which water quality data is summarized for consistently reporting to the public. Eleven parameters (Table 1) were considered to calculate WQI. Four successive steps were followed to calculate WQI (Equation 1-4): assigning weight (wi), relative weight calculation (RWi) (Equation 1), and calculation of quality rating scale (Equation 2). The least and the most significant parameters weights of 2 and 5, respectively were assigned to TDS.

$$RWi = \frac{wi}{\sum_i^n wi} \tag{1}$$

$$QRi = \frac{Ci}{DWSi} * 100 \tag{2}$$

$$WSi = RWi * QRi \tag{3}$$

$$WQI = \sum_i^n WSi \tag{4}$$

where QRi is the quality rating for each chemical parameter i, Ci is the concentration of each chemical parameter i in each water sample (mg/L), n is the total number of parameters, and DWSi is South African maximum limit for drinking for each chemical parameter i [12]. The water suitability class was determined based on [13,14] classification: excellent water (WQI <50), good water (50 <WQI <100), poor water (100 <WQI <200), very poor water (200 <WQI <300), and water unsuitable for drinking (WQI > 300).

2.2. Models

Machine learning Algorithms: Gaussian Naïve Bayes (GNB), KNNeighbors (KNN), Decision Tree (DT), Support Vector Machines (SVM), and Linear Regression (LR) were used for water quality classification. Scikit machine learning library in Anaconda (Python) Jupyter Notebook 6.0.0. software was used for water quality prediction and accuracy testing. Modelling was performed at a test size of 20%, 25%, 30% and 40% to classify water into five quality classes. A machine learning algorithm having the highest prediction accuracy was recommended for future water quality modelling.

3. Results and Discussion

3.1. Water Quality Class

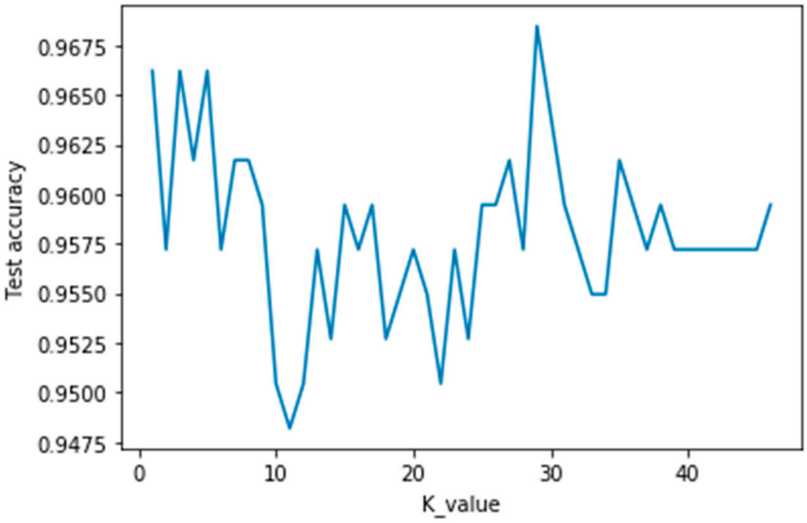
The Roodeplaat water classification classes for all instances were only three: Good, Medium, and Bad. No water quality class of "Excellent" and "Very bad "were observed. "Medium" class tells that the water quality is protected with only a slight level of harm, and frequently threatened when water quality class is "Bad".

3.2. Model Prediction Accuracy

The comparison of machine learning algorithms prediction accuracy at different test sizes was shown in Table 2. Maximum prediction accuracy values are indicated in bold. The KNNeighbor machine learning algorithm was the best predictor at all test sizes. The optimal number of nearest neighbours (K) required to predict water quality class was 30 as shown in Figure 1. Then Decision tree classifier was the second-best model especially at test size of 40%. The linear regression model was the least predictor and not good for predicting water quality class of the Roodeplaat dam.

Table 2. Model prediction accuracies at different test sizes.

Sr.No	**Machin learning Algorithms**	**Quality prediction accuracy (%) at test size of**			
		20%	**25%**	**30%**	**40%**
1	Naïve Bayes Classifier	89.94%	89.94%	90.09%	90.20
2	KNNeighbor Classifier	**96.17%**	**96.21%**	**96.39%**	**96.17%**
3	Decision Tree	95.27%	95.32%	95.95%	**96.17%**
4	Support Vector Machine	91.44%	92.25%	92.25%	92.25%
5	Linear Regression	55.26%	54.19%	55.55%	54.89%



4. Conclusion

Roodeplaat Dam water quality class was modelled using five machine learning algoithms run using Anaconda Jupyter Notebook/ Python software. The water quality of the dam was found to have good, medium, and bad classes. The prediction accuracy of all the machine learning algorithms except linear regression were very good. However, the accuracy of K-nearest neighbors model was the best of all at all test sizes. The prediction accuracy (96.39%) of KNN at test size of 30% was the highest of all test sizes and hence, recommended to classify the water quality of Roodeplaat Dam accurately. Hence, machine learning algorithms can be used to identify the quality class of the water sample and are useful to inform the water managers and the users. Hence, the water quality must be monitored frequently before the water is treated and distributed for drinking use.

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Acknowledgments: The corresponding author acknowledges the University of Johannesburg for providing a postdoctoral research fellowship.

Conflict of interest: The authors declare no conflict of interest

References

1. Chen, S.K., Jang, C.S. and Chou, C.Y., 2019. Assessment of spatiotemporal variations in river water quality for sustainable environmental and recreational management in the highly urbanized Danshui River basin. Environmental monitoring and assessment, 191(2), p.100.
2. Ahmed, A.N., Othman, F.B., Afan, H.A., Ibrahim, R.K., Fai, C.M., Hossain, M.S., Ehteram, M. and Elshafie, A., 2019. Machine learning methods for better water quality prediction. *Journal of Hydrology*, *578*, p.124084.
3. Azimi, S., Moghaddam, M.A. and Monfared, S.H., 2019. Prediction of annual drinking water quality reduction based on Groundwater Resource Index using the artificial neural network and fuzzy clustering. *Journal of contaminant hydrology*, *220*, pp.6-17.
4. Camejo, J., Pacheco, O. and Guevara, M., 2013, January. Classifier for drinking water quality in real time. In *2013 International Conference on Computer Applications Technology (ICCAT)* (pp. 1-5). IEEE.
5. Mohammadpour, R., Shaharuddin, S., Zakaria, N.A., Ghani, A.A., Vakili, M. and Chan, N.W., 2016. Prediction of water quality index in free surface constructed wetlands. *Environmental Earth Sciences*, *75*(2), p.139.
6. Babbar, R. and Babbar, S., 2017. Predicting river water quality index using data mining techniques. *Environmental Earth Sciences*, *76*(14), p.504.
7. Chou, J.S., Ho, C.C. and Hoang, H.S., 2018. Determining quality of water in reservoir using machine learning. *Ecological informatics*, *44*, pp.57-75.
8. Kamyab-Talesh, F., Mousavi, S.F., Khaledian, M., Yousefi-Falakdehi, O. and Norouzi-Masir, M., 2019. Prediction of Water Quality Index by Support Vector Machine: a Case Study in the Sefidrud Basin, Northern Iran. *Water Resources*, *46*(1), pp.112-116.
9. Najafzadeh, M. and Ghaemi, A., 2019. Prediction of the five-day biochemical oxygen demand and chemical oxygen demand in natural streams using machine learning methods. *Environmental monitoring and assessment*, *191*(6), p.380.
10. Ribeiro, V.H.A. and Reynoso-Meza, G., 2018, July. Multi-objective Support Vector Machines Ensemble Generation for Water Quality Monitoring. In *2018 IEEE Congress on Evolutionary Computation (CEC)* (pp. 1-6). IEEE.
11. Ross, A.C. and Stock, C.A., 2019. An assessment of the predictability of column minimum dissolved oxygen concentrations in Chesapeake Bay using a machine learning model. *Estuarine, Coastal and Shelf Science*, *221*, pp.53-65.
12. Holmes, S., 1996. South African Water Quality Guidelines. Volume 1: Domestic Use. *Department of Water Affairs and Forestry*, Second Edition.
13. Sahu, P. and Sikdar, P.K., 2008. Hydrochemical framework of the aquifer in and around East Kolkata Wetlands, West Bengal, India. *Environmental Geology*, *55*(4), pp.823-835.

doi:10.20944/preprints202306.1016.v1

14. Prasad, M., Sunitha, V., Reddy, Y.S., Suvarna, B., Reddy, B.M. and Reddy, M.R., 2019. Data on water quality index development for groundwater quality assessment from Obulavaripalli Mandal, YSR district, AP India. *Data in brief*, *24*, pp.103846.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.