

**Predicting the Quality of Water using Machine Learning**

G Charan Kumar Naidu

Computer Science and Engineering

Kalasalingam Academy of Research and Education

Aditya Prakash

Computer Science and Engineering

Kalasalingam Academy of Research and Education

Ravi Bhushan

Computer Science and Engineering

Kalasalingam Academy of Research and Education

D Lakshmi Nivas

Computer Science and Engineering

Kalasalingam Academy of Research and Education

G Pavan Kumar

Computer Science and Engineering

Kalasalingam Academy of Research and Education

# Predicting the Quality of Water using Machine Learning

## *Abstract*

*The issue of poor water quality is getting worse every day due to a variety of factors, including urbanisation, industrial waste, population expansion that is out of control, sewage disposal, radioactive waste, plastics, and other dangerous compounds. Frequently, the pollution is the result of our own acts. It is urgent that this issue be resolved so that good water may be obtained for diverse uses. Building a machine learning model for analysing the water quality of Indian rivers is the main goal of our effort. Using the Indian Central Pollution Control Board's water quality guidelines, we divide the dataset into four categories. After feature engineering, a total of eight key features are taken into account while creating the classification model, including temperature, biochemical oxygen dissolve, pH value, and dissolved oxygen.* *Our machine learning model's accuracy of 96.1% is higher than all of the prior research on water quality prediction.*

**Keywords**: Classification model, Machine learning, Water Quality, Water Quality Index (WQI).

## I. INTRODUCTION

Humanity benefits much from the lakes, rivers, marshes, and other sources of water. Water resources can be a useful barometer for assessing a nation's capacity for sustainable growth. These water resources' quantity and quality will have a significant impact on the entire nation. The results of water bodies like lakes and rivers in metropolitan areas are commercial as well as noncommercial jobs like nature viewing, boating, recreation, etc., and so they become a good source of cash. Water bodies are getting worse every day, which is having a severe impact on people's life. Water bodies that have been contaminated can seriously harm aquatic life as well as humans and other creatures. These contaminated water bodies will negatively impact our life by spreading waterborne infections. Humans are suffering negative effects as a result of the declining water quality. Nearly 1.6 million people die each year as a result of waterborne infections, according to a United Nations Organization (UNO) report [14]. Poor water quality causes about 1.8 million deaths annually in poorer countries [10]. Therefore, a variety of methodologies are required to evaluate the water's quality and identify its likely future trends in order to minimise any negative effects on humanity as a whole.

## II. RELATED WORK

An effective machine learning-based model for the purpose of prediction and quality analysis of the water sample is the subject of numerous studies.In [13], the researchers suggested combining the particle swarm optimization (PSO) method with least squares SVM to predict the quality of the Liuxi River.Through simulation testing, the model estimates the quality of water with high accuracy.In the year 12 they carried out research on two lakes, Chini and Bera.The eleven features in the data set were used to predict the water's dissolved oxygen concentration.There were three levels for the concentrations of dissolved oxygen:low, moderate, or highThe Anova kernel was used to apply a support vector machine model to the data, and the result was an accuracy of 74%.

A tree-based decision-making model for determining the amount of chlorophyll in common water samples has been proposed in [8].Using the principle of algae transformation, this current information on chlorophyll helps to predict the future quality of water.In estimating the quality of the water, they achieved an accuracy of 80%.

An ANN-NAR, a time-series analysis, and an artificial feedforward neural network were used to create a model in [6] that could predict the quality of water.They discovered that the ANN-NAR time series algorithm produced the highest accuracy based on a variety of performance metrics like Regression, Root Mean, and Mean Squared Error.

At Kinta River in Malaysia's Perak region, the feed-forward MLP model for predicting the water quality index is described in [4].A single input layer, one hidden layer, and one layer of output neural nodes make up their three-layer model.This model outperforms other models, demonstrating that the MLP approach is a highly reliable method in this area.When it came to predicting the WQI, the MLP model performed the best.

In [5], a ANN-based model for predicting Ganga's quality was proposed.The efficiency of the model was evaluated using a variety of metrics, including precision and accuracy.The findings demonstrate that MLP with LM outperforms the GDA algorithm when it comes to predicting water quality.

Using the Bayes Model, LI Chaunqi monitors the water quality data from three George reservoirs in [7].In [15], they used a fuzzy neural model to predict Suzhou's water quality.Since river water is consumed by the general public, the pH value is an important criterion in [11] for determining the health of the water.In [1], they predicted a free-built wetland's water quality index. With a Mean Absolute Error (MAE) of 0.9984 and a coefficient of correlation (R2) of 0.0052, their study compares the overall AVM.

**III. METHODOLOGY**

The Ganga, Beas, and Sutlej rivers, all of which originate in the Himalayas, were the subjects of three years' worth of data collection for the purpose of this experiment, from 2012 to 2014.A machine learning-based model was then developed using this data to predict the water quality of these rivers.

## i. Dataset

The open government data (OGD) platform known as India-data.gov.in [9] was the source of the data.The Indian government provides this platform in support of the Open Data initiative.There are seven parameters in the data:temperature, BOD, dissolved oxygen, conductivity, pH, fecal and total coliform counts, and temperature.The data were divided into two sets: a training set with 300 samples and a validation set with 72 samples.

 **Data Collection**: To identify trends and patterns in the data, we used a third-party platform provided by the Indian government to collect water data. A machine learning system for analyzing water quality will be built on the basis of this data.

**Data Preprocessing:** We are aware that irregular and noisy raw data are always present.Preprocessing the data improves the quality of the model.The Mean Imputation technique was used to fill in a number of missing values in this data.

**Water Quality Standards:** The water quality classification of Indian rivers as determined by the CPCB of India's water quality criteria is presented in Table 1 [2][3].

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Parameter** | **Raw water for Drinking (A)** | **Bathing Water (B)** | **Agriculture Water (C)** | **Polluted**  **Water (D)** |
| pH | 6.5 < n < 8.5 | 6 < n < 9 | 6 < n < 8.5 | n <= 6 or 9 >= n |
| BOD | n < 3 | n < 4 | n < 5 | n >= 5 |
| DO | 6 < n | 5 < n | 2 < n | n <= 2 |
| COD | n < 15 | n < 20 | n < 50 | n >= 50 |
| NO3 | n < 5 | n < 5 | n < 5 | n >= 5 |

Table 1: Water Quality standards for different types of water

##  Prediction Model Used

Decision tree, support vector machine (SVM), K-Nearest Neighbor (KNN), Nave Bayes method, and multi-layer perceptron (MLP) are the five distinct machine learning models developed in this work.

## A. K-Nearest Neighbour

Both classification and regression problems can be solved with KNN.In fact, it is a very straightforward algorithm that uses distance functions to classify the new cases.KNN uses a variety of distance functions, including:Manhattan distance, Minkowski distance, and Euclidean distance.For continuous variables, the distance functions described above are utilized.Hamming distance is utilized for categorical variables.

## B. Naive Bayes Method

Naive Bayes is a straightforward but surprisingly effective model for predictive analysis.It is utilized for both multi-class and binary classification.P(x1, x2, x3|h) is not used to calculate the value of each attribute; instead, it is assumed that they are conditionally independent and is calculated as P(x1|h) \* P(x2|h), and so on.Since attributes rarely interact with one another, this is actually a very strong assumption that is unlikely to be true in real data.

## C. Support Vector Machine (SVM)

Based on the supervised machine learning paradigm, SVM is a well-known algorithm that can be used to analyze regression and classification issues.When transforming the data, SVM typically employs the Kernel trick, which aids in locating the ideal boundary between the various outputs.It is known that smaller datasets perform better with SVM.

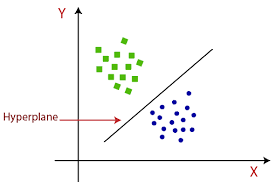


Fig. 1. Hyperplane division in SVM

## D. Decision Tree

## A decision tree model has a structure that looks like a tree and can be used for both classification and regression problems.It makes decisions based on a tree-like structure, as its name suggests.Decision trees' implicit feature selection is one of their advantages.The performance of the tree is not harmed by the data's nonlinearity.

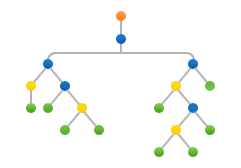


Fig. 2. Decision Tree

#### E. Multi-Layer Perceptron

An MLP artificial neural network is a feed forward model with an input layer, one or more hidden layers, and an output layer.Analyzing the classification, prediction, and recognition of patterns requires the use of MLPs, which are extremely beneficial.Additionally, it is well-known for resolving nonlinear data-related issues.MLP employs the backpropagation algorithm for training purposes.

## IV. PERFORMANCE METRICS

We need a test dataset to measure the efficiency of a machine learning model in order to determine its performance.Cross validation steps in to save the day by evaluating the machine learning model's performance on a separate dataset.In our study, we employed 10-fold cross validation, which divides the dataset into ten equal parts for the purposes of training and validation and measuring the model's performance.For performance evaluation, we calculate f1-sccore, accuracy, precision, and recall.

The most straightforward method for evaluating the model's performance is **confusion metrics**.Classification problems make use of it.

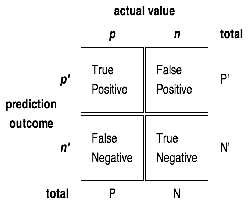


Fig. 3. Description of Confusion Matrix

Although confusion metrics aren't a performance measurement in and of themselves, they are helpful to nearly every performance measurement tool in calculating performance by utilizing its values.

|  |  |
| --- | --- |
| **Accuracy** (%) = (True Positive + True Negative × 100) / Total Sample Space | (i) |
| **Precision** = (True Positive ×100) / (True Positive + False Positive) | (ii) |
| **Recall** = (True Positive ×100) / True Positive + False Negative | (iii) |
| **F1-Score** = (Recall \* Precision \* 2 ×100) / (Precision + Recall) | (iv) |

## V. RESULTS AND OBSERVATIONS

Expectations can be estimated in different ways like exactness, review, accuracy, F1-score. We utilized this multitude of four boundaries to really take a look at the better model for our information.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Technique** | **Accuracy** | **Precision** | **F1-score** | **Recall** |
| MLP | 93.6% | 0.93 | 0.92 | 0.92 |
| KNN | 92.2% | 0.92 | 0.91 | 0.90 |
| NB | 88.6% | 0.87 | 0.86 | 0.86 |
| DT | 94.4% | 0.94 | 0.94 | 0.94 |
| ***SVM*** | ***96.1%*** | ***0.95*** | ***0.95*** | ***0.95*** |

Table 3: Summary of observations using best eight attributes

In this work, five different machine learning models namely: support vector Machine (SVM), Decision tree, K-Nearest Neighbor (KNN), Naïve Bayes method and Multi-layer perceptron (MLP) were used. It is evident from the results obtained that the SVM based model outperforms all other classifiers in terms of performance.

## VI. CONCLUSION

A machine learning classifier for predicting the quality of Indian river water was built successfully in this work.With 96.1 percent accuracy, we find that SVM performs the best when compared to other machine learning models.Due to the oxygen depletion, we also observed a strong correlation between temperature and Biochemical Oxygen Demand (BOD).Because dissolved oxygen (DO) has a negative correlation with temperature, DO varies slightly during the summer.As a result, this model can be used to estimate the quality of Indian river water.

## REFERENCES

1. Afifah, T., Ahmed, A. N., & El-Shafie, A. (2014). Dissolved oxygen prediction using Support Vector Machine in Terengganu River. Middle East Journal of Scientific Research, 21(11), 2182-2188.
2. Central Pollution Control Board, Available: http://www.cpcb.gov.in
3. Central Pollution Control Board, Available: http://cpcb.nic.in/water-quality-criteria/
4. Gazzaz, N. M., Yusoff, M. K., Aris, A. Z., Juahir, H., & Ramli, M. F. (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), 2409-2420.
5. Giri, A., & Singh, N. B. (2014). Comparison of Artificial Neural Network Algorithm for Water Quality Prediction of River Ganga. Environmental Research Journal, 8(2),
6. Khan, Y., & See, C. S. (2016, April). Predicting and analyzing water quality using Machine Learning: A comprehensive model. In 2016 IEEE Long Island Systems, Applications and Technology Conference (LISAT) (pp. 1-6). IEEE.
7. Li, C., & Wang, W. (2009, December). Assessment of the water quality near the dam area of Three Gorges Reservoir based on Bayes. In 2009 First International Conference on Information Science and Engineering (pp. 145-148). IEEE.
8. Lu, J., & Huang, T. (2009, August). Data mining on forecast raw water quality from online monitoring station based on decision-making tree. In 2009 Fifth International Joint Conference on INC, IMS and IDC (pp. 706-709). IEEE.
9. Open Government Data Platform (OGD), Available: India http://www.data.gov.in

[10] Payen, G. (2011). Worldwide needs for safe drinking water are underestimated: billions of people are impacted. Aquafed.

[11] Sharma, D., & Kansal, A. (2011). Water quality analysis of River Yamuna using water quality index in the national capital territory, India (2000–2009). Applied water science, 1(3-4), 147-157.

[12] Solanki, A., Agrawal, H., & Khare, K. (2015). Predictive Analysis of Water Quality Parameters using Deep Learning. International Journal of Computer Applications, 125(9), 0975-8887.

[13] Xiang, Y., & Jiang, L. (2009, January). Water quality prediction using LS-SVM and particle swarm optimization. In 2009 Second International Workshop on Knowledge Discovery and Data Mining (pp. 900-904). IEEE.

[14] Zeilhofer, P., Zeilhofer, L. V. A. C., Hardoim, E. L., Lima, Z. M. D., & Oliveira, C. S. (2007). GIS applications for mapping and spatial modeling of urban-use water quality: a case study in District of Cuiabá, Mato Grosso, Brazil. Cadernos de saude publica, 23, 875-884.

[15] Zhu C., Hao Z. Fuzzy Neural Network Model and Its Application in Water Quality Evaluation. International Conference on Environmental Science and Information Application Technology 2009 Jul 4 (Vol. 1, pp. 251-254). IEEE.