AN ANALYSIS OF CAUVERY RIVER WATER QUALITY USING MACHINE LEARNING ALGORITHMS

Dr. K. Shyamala
PG & Research Department of Computer Science
Dr. Ambedkar Government Arts College, Chennai, India
Shyamalakannan2000@gmail.com

L. Karishma
PG & Research Department of Computer Science
Dr. Ambedkar Government Arts College, Chennai, India
Karishmal.research@gmail.com
https://doi.org/10.34293/9789361639715.shanlax.ch.001

Abstract

Water is essential for Earth's ecosystems and supports crucial biological activities. Water plays a critical role in residential, industrial, agricultural and environmental applications. However, water resources are increasingly threatened by pollution from various sources including sewage effluents, industrial processes, oil spills and agricultural practices. To effectively assess and predict water quality, machine learning technology has been employed. Water Pollution Assessment involves the analysis of different contaminants and pollutants such as chemicals, pathogens and organic materials that can endanger the safety of water for human use. Accuracy of water quality prediction in machine learning not only relies on models used but also on the quality of the dataset. In this work, an analysis of water quality prediction using Cauvery river data set was presented. The performance of classification algorithms has significant results, with 89% accuracy and 97% precision.

Keywords: Water Quality Prediction – Machine Learning models-Cauvery river water-Evaluation metrics.

Introduction

Life on Earth is sustained by the presence of water, as it is essential for various biological processes and ecosystems to thrive. Water quality plays a crucial role in supporting humans, plants, animals and for maintaining a healthy environment [1]. In Southern India, Cauvery river serves as the important water resource, which flows through Karnataka and Tamilnadu. It is extensively used for drinking, agriculture and for supporting ecosystems. However, the river is growing polluted due to various industrial discharge, agricultural runoff, urbanization and untreated urban sewage.

This work utilized the Cauvery River dataset for evaluating water quality. The primary goal of this research was to proficiently perform feature selection and classification tasks for predicting the pollution and contamination level in the Cauvery River. To enhance the data processing and expedite analysis, Machine learning techniques were used [2]. Parameters such as Dissolved oxygen (DO), pH, biological oxygen demand (BOD), chemical oxygen

demand (COD), Nitrate, Phosphate, Conductivity, Fecal coliform, Turbidity and Total Dissolved Solids (TDS) were used to calculate the Water Quality Index. These parameters showcase the pollution caused by agricultural runoff, industrial discharge and biological contaminants present in the Cauvery river.

According to WQI thresholds, water has been classified into four categories. This classification shows the varying levels of pollution, which helps the government to treat the water accordingly. To analyze water pollution effectively, machine learning algorithms such as Random Forest, KNN, Naïve bayes, Stochastic Gradient boosting, gradient boosting algorithms were used. These methods have effectively categorized the water quality, based on the water quality index values.

Literature Review

Jitha P Nair et al. [3] Forecasted the WQ using Machine Learning techniques and their objective is to construct an accurate and efficient prediction model and classify the index value as per standard water quality thresholds. The data was collected from eleven sampling stations across myriad locations in the Bhavani River. WQI was calculated based on 27 attributes. For model development, data cleaning, data normalization, and feature selection were performed to derive machine Learning Models. The modeling phase included training phase which utilized Regression Algorithm and classification algorithm and subsequently, prediction model and classification models were employed. In their research, the water quality was predicted using the water quality index, where the highest accuracy was achieved with the MLP regressor and MLP Classifier at 81%.

Vellangiri J et al. [4] have proposed new approach which integrates both Aquila optimization with support vector machine. Here SVM takes care of classification and Aquila algorithm optimizes SVM. The dataset was collected from the 33 stations monitoring the Cauvery River for the Tamil Nadu Pollution Control Board. The prediction time using the proposed AO-SVM model is 7.5 milliseconds and accuracy is 96%.

Mahmoud Y. Shams et al. [5] developed techniques for optimization and tuning to enhance the accuracy of a few selected machine learning methods. Grid search is applied for optimizing and tuning the parameters of four classifications and four regression models. Data cleaning which includes the steps of data imputation and data normalization was done to prepare the data for processing. The worked Dataset has 7 attributes (features) and 1991 instances (records). This study concluded that the Gradient Boosting model performs well with 99.5% accuracy.

Md. Mehedi Hassan et al. [6] evaluated that WQI denotes one crucial measure for efficient water resources management. This dataset underwent a normalization technique with minimum value max normalization, missing values filled using Random Forest Approach, feature correlation, machine learning classification and importance of the model's features. The best performance achieved by this model is with Multinomial Logistic Regression having an accuracy greater than 99.83%.

Arun Josephraj [7] collects real time data of kaveri river through Internet of things device, in which the sensor can measure temperature, ph, total dissolved salts and turbidity all 4

parameters together. These data was collected and stored along with the historical data and then it is checked for any possible outlier and trends. This study was compared with different time series models which are best for predicting the efficiency of the model. The analysis concludes that to analyze and forecast future reading of Kaveri river, the Recurrent Neural Network model is by far the best for time series analysis. The aim of this research finding is to help government agencies, local populations and other individuals who feel concerned about the river's safety.

The existing studies on water quality prediction are limited in scope, as they classify and predict using only a small number of samples and stations. Most of these studies apply Z-score normalization, which, although effective, results in the loss of original measurement units and thereby reduces interpretability for stakeholders such as policymakers and environmental managers. Furthermore, the majority of studies have employed only a restricted set of classification algorithms, which limits the exploration of more advanced or diverse machine learning models. The inclusion of all 26 water quality parameters without effective feature selection may affect the model performance.

Study Area

The Cauvery River is also referred as 'The Ganga of the South'. It originates at Talakaveri at the Western Ghats in Karnataka and it drains into Bay of Bengal near Poompuhar in Mayiladuthurai District. The river traversing through Karnataka (40%), Tamilnadu (52%), Kerala (3.75%), Puducherry (4.25%) and totally it flows through around 800 Kilometers. Cauvery river is considered as Sacred river and has significant cultural and historical importance. This river plays a crucial role in irrigation, drinking water supply, power generation, industrial use and for environmental conservation [8].

Data Collection

The Cauvery river Stations include Mettur, Pallipalayam, MusiriFerrygate, BathiraKaliammanKoil, Sirumugai, Bhavanisagar, Bhavani, R.N.Pudur, Vairapalayam, P.Velur, Mohanur, Madathukulam, Thirumukkudal, Trichy U/S, Trichy D/S, Grand Anaicut, Coleroon, Pitchavaram, Karunthattankudi, Komarapalayam, Urrachikottai, Chirampalayam, Pugalur, Pettavaithalai, Kumbakonam, Sathiyamangalam, Kalingalray canal (B5), Kalingalray canal (B10), Thirumanimuthar, Vasista, Sarabanga. The data set consists of 1229 instances with 26 parameters has been collected from Tamilnadu Pollution Control Board from 2018 to 2023. For Efficient Water quality Index calculation parameters like Conductivity, pH, Turbidity, Dissolved Oxygen (DO), Biochemical Oxygen Demand (BOD), Chemical Oxygen Demand (COD), Nitrate (NO3), Fecal Coliforms, Total Dissolved Solids (TDS) and Phosphate (PO4) were used. This Study utilized Cauvery River dataset to train machine learning models.

Water Quality Index

The water quality index is measured using the WHO Standard values (Si), Unit weights (Wi) and Ideal values (V0) and Analyzed parameter (Vi). This index is used to find the

overall water quality status by simplifying the complex water quality data into a single numeric value by applying formula (1)

$WQI=\sum(Wi\times Vi)$ (1)

Where Wi represents the Unit Weights according to WHO and Vi can be calculated by applying formula (2)

$Vi = 100 - \left(\frac{v0-si}{si}\right) \times 50$ (Above max limit) (2)

If the ideal value is above the max limit of Standard value (Si), then the vi is calculated by multiplying 50 and if the ideal value is below the max limit of Standard value (Si), then the vi is calculated by multiplying 100 using formula (3)

$Vi = 100 - \left(\frac{v0}{si}\right) \times 100$ (Below max limit) (3)

Table 2: Parameters with Standard value and Unit Weights

Parameters	Standard Value	Weights
pH	6.5 – 8.5	0.218
Conductivity	300	0.006
Turbidity	5	0.1
Dissolved Oxygen (DO)	6	0.37
Biochemical Oxygen Demand	5	0.1
Chemical Oxygen Demand (COD)	120	0.1
Nitrate (NO3)	45	0.412
Fecal Coliforms	<10	0.1
Total Dissolved Solids	<500	0.0037
Phosphate (PO4)	<0.1	0.1

Table 2 shows the standard value of each parameter and their unit weights. The Standard values of these attributes are taken from World Health Organization (WHO) [9], Bureau of Indian Standards (BIS), Central Pollution Control Board (CPCB) [10] and Environmental Protection Agencies Globally (EPA) [11].

Table 3: Water Quality Index Standards for Water Quality

Water Quality Index	Water Quality
0-100	Excellent
100-200	Good
200-300	Poor
>400	Very Poor

After Calculating Water Quality Index using formula (1) with respect to parameter limits and unit weights in Table 2, it was then classified according to the standards of water quality as shown in Table 3. In this table index value 0-100 indicates the water is natural without pollution and is safe for drinking and irrigation. 100-200 range indicates the water is slightly polluted and can be consumed after filtration process. 200-300 range indicates the water is

moderately polluted and can be used for irrigation of non-edible crops and industrial use. >400 indicates water is extremely polluted and cannot be used without treatment.

Data Pre-Processing

Raw data consists of errors, inconsistencies, missing values and irrelevant information which will lead to inaccurate predictions during the training of machine learning algorithms [12]. To avoid these issues Data processing, cleaning, reduction and normalization are required. Data pre-processing makes the data more usable for analysis and ensures its quality and accuracy by addressing issues like missing values, inconsistencies and error values. Complete case analysis (CCA) method is used for treating the missing values [13].

Normalization Method

Normalization is a technique used to standardize the parameter values in the dataset [14]. In Fecal Coliform some value lies in 17000 - 34000 but the range is 10 – 2500. Here, for standardizing the parameter values, Z-score normalization is applied using mean and standard deviation.

Methodology

This Study is a Quantitative approach using machine learning algorithms. The data set was obtained from Tamilnadu Pollution Control Board for Cauvery river from the year 2018 to 2023. The data was collected from water quality monitoring stations. The attributes include COD, BOD, conductivity, nitrate, fecal coliform, pH, Dissolved Oxygen, Turbidity, TDS, phosphate were used to calculate Water Quality Index. Machine learning algorithms were used for WQI Classification and Performance evaluation metrics such as accuracy, precision, recall, f1 score were used for comparing the machine learning algorithms [15]. The workflow model of this study was shown in Fig. 2. The analysis was conducted using Orange tool (version 3.39.0) with Python coding implementation facilitated through the Python widget. Model evaluation employed the common hold-out validation technique for training and testing with 70-30 ratio.

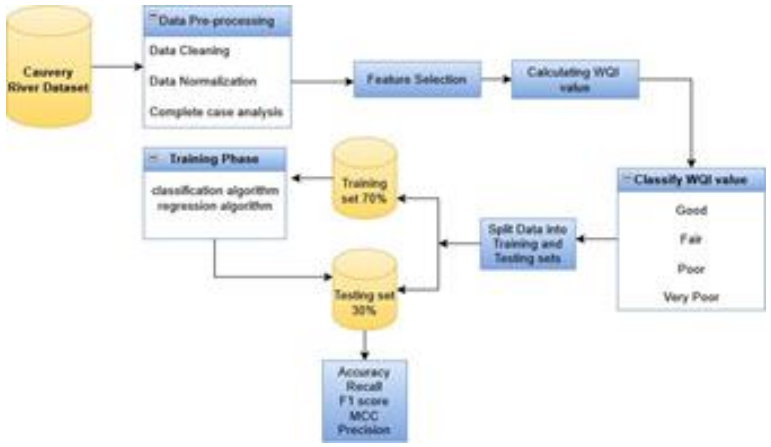

Fig. 2. Workflow Model of this Study

Feature Selection

To improve the predictive accuracy of the WQI, feature selection was performed by prioritizing relevant attributes. The dataset initially contained 26 attributes, but those strongly linked to the major sources of water pollution are agricultural runoff, industrial pollution and domestic sewage. Attributes such as COD, BOD, conductivity, nitrate, fecal coliform, pH, Dissolved Oxygen, Turbidity, TDS, phosphate were identified as key indicators of these pollution sources. Based on the analysis, these 10 attributes were selected for the model. This selection process reduced complexity while ensuring the inclusion of parameters most relevant to WQI prediction.

WQI Prediction Model

A Water Quality prediction model uses the machine learning algorithms to analyse the trends in the river water quality dataset using Logistic regression, Naive Bayes, Random Forest, KNN, Stochastic Gradient boosting, gradient boosting algorithms, Support vector machine and Decision tree method. In this study, WQI is the target variable and other parameters in the dataset are independent variables in this regression process.

Logistic regression is a supervised learning algorithm used primarily for binary classification tasks by modelling the probability of an outcome based on input variables. It is simple, interpretable, computationally efficient and works well with both numerical and categorical data.

Naïve Bayes was a probabilistic classifier based on Bayes theorem, assuming independence among predictors. It is highly efficient for large datasets and works well with categorical data and text based applications like spam detection and sentiment analysis. Despite its simplicity, it performs surprisingly well in real-world scenarios but may struggle when predictor independence is violated or with highly correlated features.

Random Forest is a supervised machine learning algorithm that combines several decision trees to improve prediction accuracy and reduce overfitting. It Handles missing data, ranks feature importance, works well with large datasets and is resistant to overfitting.

Gradient Boosting is an ensemble learning technique that builds models sequentially, with each model correcting the errors of its predecessor. It is highly effective for both classification and regression tasks and is renowned for its predictive accuracy.

Support Vector Machine is a supervised learning algorithm designed for classification and regression tasks. It identifies an optimal hyperplane to separate data points into distinct classes and can handle both linear and non-linear problems effectively using kernel functions. SVM is particularly suitable for high- dimensional datasets.

K-Nearest Neighbors is a non-parametric algorithm that classifies data points based on the majority class of their closet neighbors. It is simple to implement and performs well on small datasets, but its computational complexity increases with larger dataset due to distance calculations.

Stochastic Gradient Descent is an optimization algorithm commonly used to minimize loss functions in machine learning models, including classifiers and regressors. It is computationally efficient for large-scale datasets but requires careful tuning of hyper

parameters, such as the learning rate, to achieve optimal performance.

Decision Tree Model is a supervised learning model that splits data into branches based on feature values to make predictions. It is intuitive and interpretable, capable of handling both classification and regression tasks. Pruning must be used to prevent overfitting problems.

WQI Classification Model

A classification model for WQI was developed by analyzing patterns within a comprehensive river water quality dataset. This model utilizes the water quality index as a class label to categorize water quality into four distinct standards – Excellent, Good, Poor and very Poor based on different water quality parameters. more

Results and Discussion

This Study analyze the major parameters which affects the water quality of Cauvery River. Among the various parameters, the parameters which affects the water quality alone are selected for calculating the water quality index. Then the water quality index is categorized into 0 and 1 which depicts the water quality is good or bad. Based on this results the water quality is classified using various machine learning algorithms.

Table 4. Evaluation Metrics

Metric	Gradient Boosting	Random Forest	Decision Tree	Logistic Regression	Stochastic Gradient Boosting	Naive Bayes	KNN	SVM
CA	89	88	88.8	88.5	82	81	74	70
F1	100	99.8	99.6	99	95.8	94.4	77.3	62.7
Precision	97.4	99.7	99.7	98.6	94.3	92.5	84.2	85.4
Recall	100	99.9	99.5	99.5	97.3	96.3	71.5	49.5
MCC	99.9	99.6	99.4	98.4	93	90.6	65	51.2

Table 4. shows the Evaluation metrics such as Classification Accuracy (CA), F1-score, precision, recall and Matthews correlation coefficient (MCC) of various machine learning algorithms. Fig.3. shows the Graphical representation of Evaluation metrics and it shows that the gradient boosting and random forest works best, while comparing with SVM. Logistic Regression has highest accuracy but it has a significantly greater training time than others.

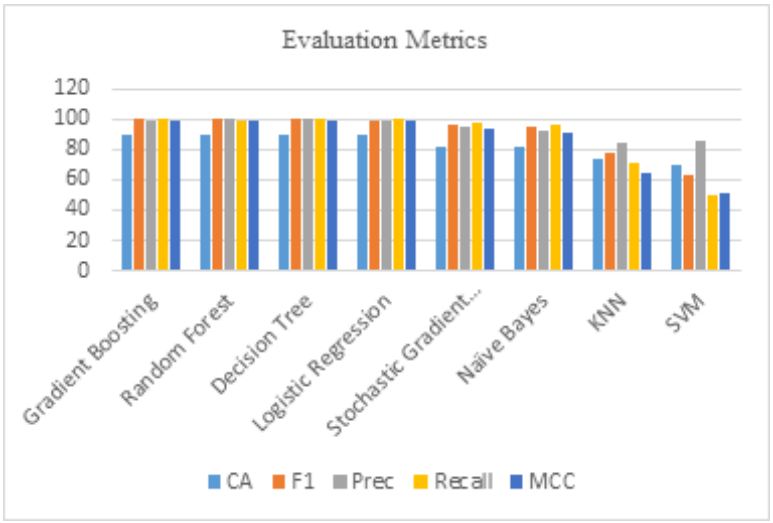

Fig.3. Comparison of Evaluation metrics of various ML models

Fig.4. shows the comparison of machine learning models based on its accuracy. The Accuracy (%) of these models were evaluated based on their performance. Gradient boosting works as the best- performing model with 89% accuracy when comparing with other models. Logistic regression and decision tree models also performs well with accuracies of 88.5% and 88.8%.

Random Forest shows a decent performance with an accuracy of 88%. Stochastic gradient boosting model's performance is slightly ahead of naive bayes model. Comparing with other algorithms, KNN & SVM algorithms have lowest accuracy and these algorithms does not exhibit optimal performance for this data set.

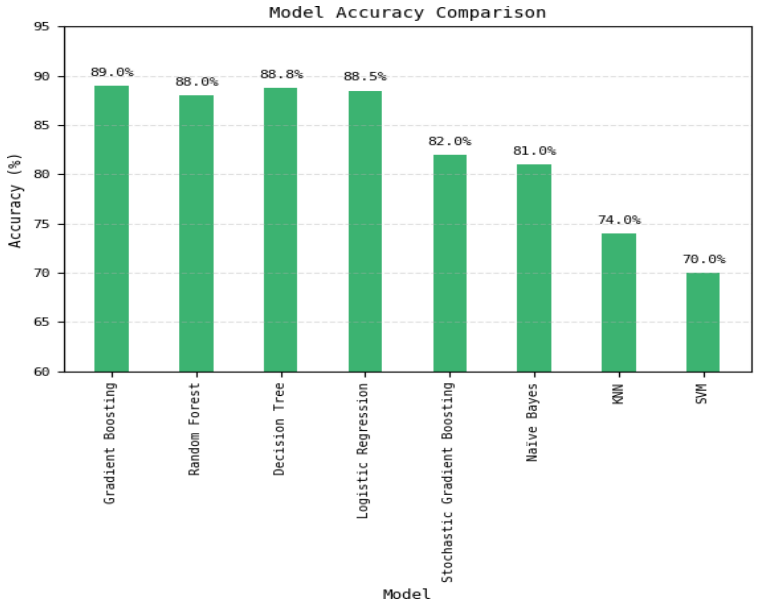

Fig.4. Comparison of Accuracy for various ML models

This study utilized only 10 water quality parameters of the 26 available, selected manually to emphasize indicators of industrial and agricultural runoff. The highest accuracy achieved was 89% with Gradient Boosting, and feature selection was not performed using systematic or automated methods.

Conclusion

This study analyzed the performance of various machine learning models for Water Quality prediction based on multiple evaluation metrics. Gradient boosting model works best for this data set among SVM, KNN, Logistic Regression, Random Forest, Stochastic Gradient Boosting, Decision tree and Naive bayes. The Cauvery river data with ten parameters were collected with 1229 instances for calculating Water Quality Index value. The Water quality index values are categorized and then WQI prediction was done. It is concluded that Gradient boosting model outperforms than the other models in forecasting water quality prediction with 89% accuracy and 97% precision.

Future Work

To enhance the accuracy of water quality prediction for the Cauvery River, dominant features identified through Exploratory Data Analysis (EDA) and Principal Component Analysis (PCA) can be utilized for more precise modelling. Additionally, applying standardization techniques may improve the performance and interpretability of machine learning models.

References

1. Haghiabi, Amir Hamzeh, Ali Heidar Nasrolahi, and Abbas Parsaie. "Water quality prediction using machine learning methods." *Water Quality Research Journal* 53.1 (2018): 3-13.
2. Gong, Youdi, et al. "A survey on dataset quality in machine learning." *Information and Software Technology* 162 (2023): 107268.
3. Nair, Jitha P., and M. S. Vijaya. "River water quality prediction and index classification using machine learning." *Journal of physics: Conference series*. Vol. 2325. No. 1. IOP Publishing, 2022.
4. Vellingiri, J., et al. "AO-SVM: a machine learning model for predicting water quality in the cauvery river." *Environmental Research Communications* 6.7 (2024): 075025.
5. Shams, Mahmoud Y., et al. "Water quality prediction using machine learning models based on grid search method." *Multimedia Tools and Applications* 83.12 (2024): 35307-35334.
6. Hassan, Md Mehedi, et al. "Efficient prediction of water quality index (WQI) using machine learning algorithms." *Human-Centric Intelligent Systems* 1.3 (2021): 86-97.
7. Indian Urban Data Exchange (IUDX). (n.d.). *Cities*. https://iudx.org.in/cities/ *(Accessed May 13, 2025)*
8. Sharma, Rakesh. "Assessment of Irrigation Water Quality of the River Cauvery and Its Distributaries in Tamil Nadu, India through Hydrochemical Analysis." (2025).

9. World Health Organization. (n.d.). Drinking-water quality guidelines. https://www.who.int/teams/environment-climate-change-and-health/water-sanitation-and-health/water-safety-and-quality/drinking-water-quality-guidelines
10. Central Pollution Control Board. (2019, October 11). WHO guidelines for drinking water quality. https://cpcb.nic.in/who-guidelines-for-drinking-water-quality/
11. United States Environmental Protection Agency. (2023, August 10). What are water quality standards? https://www.epa.gov/wqs-tech/what-are-water-quality-standards
12. Maharana, Kiran, Surajit Mondal, and Bhushankumar Nemade. "A review: Data pre-processing and data augmentation techniques." *Global Transitions Proceedings* 3.1 (2022): 91-99.
13. Martijn W. Heymans, Jos W.R. Twisk, Handling missing data in clinical research, Journal of Clinical Epidemiology, Volume 151, 2022, Pages 185-188, ISSN 0895-4356, https://doi.org/10.1016/j.jclinepi.2022.08.016
14. Vinay, Sachin. (2021). STANDARDIZATION IN MACHINE LEARNING.
15. Solangi, Ghulam Shabir, et al. "Machine learning, Water Quality Index, and GIS-based analysis of groundwater quality." *Water Practice & Technology* 19.2 (2024): 384-400.