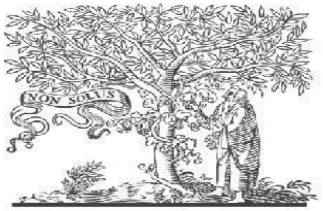
International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

# COPY RIGHT



ELSEVIER SSRN



IJIEMR Transactions, online available on 10th Apr 2023. Link :http://www.ijiemr.org/downloads.php?vol=Volume-12&issue=Issue 04

**10.48047/IJIEMR/V12/ISSUE 04/102**

Title **WATER QUALITY PREDICTION USING MACHINE LEARNING APPROACHES**

Volume 12, ISSUE 04, Pages: 820-827

Paper Authors

**Mr. Y. Venkata Narayana, Tangirala Meghana, Sunkara Monisha, Ravella Navya Sree, Vasireddy Kheerthhana**


USE THIS BARCODE TO ACCESS YOUR ONLINE PAPER

To Secure Your Paper As Per UGC Guidelines We Are Providing A Electronic Bar Code

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

# WATER QUALITY PREDICTION USING MACHINE LEARNING APPROACHES

**Mr. Y. Venkata Narayana**, Assistant Professor, Department of IT,
Vasireddy Venkatadri Institute of Technology, Nambur, Guntur Dt., Andhra Pradesh.

**Tangirala Meghana**, **Sunkara Monisha**, **Ravella Navya Sree**, **Vasireddy Kheerthhana**
UG Students, Department of IT,
Vasireddy Venkatadri Institute of Technology, Nambur, Guntur Dt., Andhra Pradesh.
meghanatangirala725@gmail.com, monisha.sunkara14@gmail.com,
navyasreeravella9@gmail.com, Vasireddykheerthhana@gmail.com

**Abstract**

The Water quality has deteriorated significantly over the last few decades owing to contamination and other factors. As a result, a model capable of making accurate estimates regarding water quality is required. Keeping track of the treated water outflow is critical for the stability and conservation of the environment. Moreover, inadequate sanitary facilities and a lack of knowledge contribute significantly to drinking water pollution. Water quality degradation has far-reaching consequences, including harming health, the environment, and infrastructure. Waterborne infections kill more than 1.5 million people (about the population of West Virginia) each year, according to the United Nations (UN), much more than accidents, crimes, and terrorism. As a result, it is extremely crucial to forecast the quality of water. Earlier, water quality was checked manually. Yet, systems taught with machine-learning methods such as Linear Regression and SVM (Support Vector Machines) classifiers are later employed independently. The present implementation predicts the quality of water utilizing different Supervised Machine Learning methods such as Linear Regression, K-Nearest Neighbour, Decision trees, XGBoost, and Random Forest trees. Finally, this study seeks the algorithm that provides the highest accuracy while still maintaining water quality. This study compares multiple Machine Learning algorithms in tracking the water purity through KNN, Decision tree, Random Forest trees, SVM, and Gradient Boosting Classifier. The Water Quality Index dataset from Kaggle was used to train this model.

**Keywords:** Water Quality prediction; Supervised Machine Learning; KNN; SVM; Decision Tree; XGBoost; Random Forest trees

## 1 Introduction

Water is the most significant supply, essential for all forms of life; unfortunately, it is constantly polluted. In terms of communication and reach, water is one of the most effective media. A result

of more development is a decline in water quality. Bad water quality has been identified as one of the key contributors to the spread of heinous illnesses. According to estimates, 80% of infections in underdeveloped nations are water-borne, resulting in 6 million deaths and 3 billion affected by chronic diseases. [1] The most common diseases in India are nausea, cholera, stomach flu, crypto-sporidium illnesses, and various types of hepatitis. [2]. Every year, water-borne illnesses cost India's GDP 0.6-1.44%. [3]. As a result, it is a crucial issue, especially in emerging countries like India. Nowadays, expensive and time-consuming lab and statistical analyses are used to evaluate water quality. These investigations require the collecting of samples, conveyance to labs, and a substantial amount of time and computation. If water is contaminated with disease-causing waste, speed is of the essence because it is a communicable medium. [4]. This study is based on a dataset obtained by the India Council of Research that specifies whether water is safe to be consumed by people, where 1 shows drinkable and 0 shows not fit for drinking.
{https://www.kaggle.com/dataset/shikadiwal/water-potability/). On the dataset, a representative collection of supervised Machine Learning algorithms was used to predict the water quality index (WQI).

## 2. Literature Survey

The methods that have been employed to alleviate problems with water quality are examined in this paper. Yet, other studies use machine learning techniques to find the optimum answer to the water quality problem. Classical laboratory testing and statistical methods are typically used in research to assist determine water quality. We learned more about India's water quality issue through analysis that made use of laboratory evaluations. Daud et al. [5] collected samples taken from multiple locations in India and used a manual lab analysis to assess them against various criteria, discovering a significant presence of Escherichia bacteria due to commercial and drainage waste. Alamgir et al. [6] looked at 45 variable samples. We investigated studies using machine learning approaches in the field of water quality after becoming acquainted with water quality research in India. Shafi et al. [7] employed 16 metrics for measuring water quality and an ANN with Bayesian regularisation to calculate the WQI. In order to forecast the WQI, Gazzaz et al. [8] used ANN. They discovered that their model accounted for 97.5% of the data's variation. They used 23 features to anticipate the WQI based on the cost of the sensors, which turned out to be fairly expensive when applied to an IoT system. The greatest number of studies either relied on physical laboratory analysis to compute the WQI, did not assess the WQI standard, or used too many factors to be effective. These ideas are enhanced by the proposed methodology.

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

## 3. Problem Identification

Degradation of water quality has far-reaching consequences, including harm to health, the environment, and infrastructure. Waterborne infections kill over 1.5 million people, far outnumbering accidents, crimes, and terrorism. As a result, forecasting water quality is critical. Previously, water quality was manually checked. Even though systems taught using Machine Learning methods such as Linear Regression and SVM classifiers are used individually. The current implementation predicts the quality of water using various supervised machine learning methods.

## 4. Proposed Methodology

The bulk of research used conventional lab analysis, could not forecast the criteria for water quality, or had far too many variables. Figure 1 shows the methodology being used and the suggested technique that expands on these concepts.

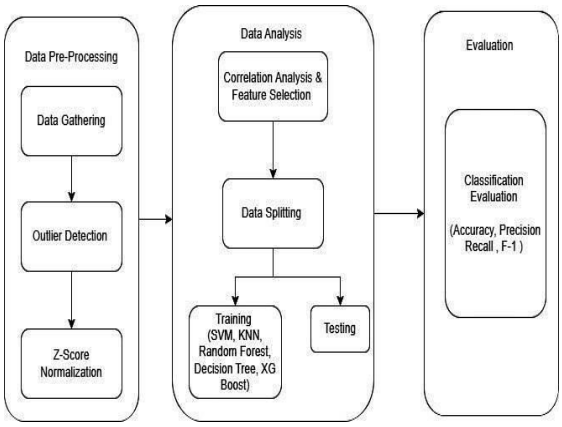



**Figure 1. Methodology flow**

## 5. Implementation:

### 5.1 Data pre-processing

Utilizing a box plot analysis, the data for this study were cleaned and collected from Kaggle https://www.kaggle.com/datasets/adityakadiwal/water-potability. To enable the WQI to be computed using the ten applicable parameters listed in Table 1, first data were cleaned up and then standardized using z-value normalization to adjust their range to 0-100 to ensure that they were all on the same scale.

| S.no | pH | Hardness | Solids | Chloramine | Sulphate | Conductivity | Organic_carbon | Trihalomethane | Turbidity | Portability |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 8.316766 | 214.3734 | 22018.42 | 8.059332 | 356.8861 | 363.2665 | 18.43652 | 100.3417 | 4.628771 | 0 |
| 2 | 9.092223 | 181.1015 | 17978.99 | 6.5466 | 310.1357 | 398.4108 | 11.55828 | 31.99799 | 4.075075 | 0 |
| 3 | 5.584087 | 188.3133 | 28748.69 | 7.544869 | 326.6784 | 280.4679 | 8.399735 | 54.91786 | 2.559708 | 0 |
| 4 | 10.22386 | 248.0717 | 28749.72 | 7.513408 | 393.6634 | 283.6516 | 13.7897 | 84.60356 | 2.672989 | 1 |
| 5 | 8.635849 | 203.3615 | 13672.09 | 4.563009 | 303.3098 | 474.6076 | 12.36382 | 62.79831 | 4.401425 | 0 |
| 6 | 11.18028 | 227.2315 | 25484.51 | 9.0772 | 404.0416 | 563.8855 | 17.92781 | 71.9766 | 4.370562 | 0 |
| 7 | 7.36064 | 165.5208 | 32452.61 | 7.550701 | 326.6244 | 425.3834 | 15.58681 | 78.74002 | 3.662292 | 0 |
| 8 | 7.119824 | 156.705 | 18730.81 | 3.606036 | 282.3441 | 347.715 | 15.92954 | 79.50078 | 3.445756 | 1 |
| 9 | 6.347270 | 186.7329 | 41065.23 | 9.629596 | 364.4877 | 516.7433 | 11.53978 | 75.07162 | 4.376348 | 0 |
| 10 | 9.18156 | 273.8138 | 24041.33 | 6.90499 | 398.3505 |  | 13.38734 | 71.45736 | 4.503661 | 0 |

**Table 1. Dataset**

### 5.2 Data Analysis

After achieving data pre-processing, a wide range of machine learning techniques were employed for analysis of data to evaluate the WQI with the least amount parameters possible. Before using a machine learning method, several preparatory processes, such as correlation analysis and data splitting, must be completed by using a machine learning method.

### 5.2.1 Correlation Analysis

The diagnosis of dependent variables and forecasting the difficult-to-estimate variables by utilizing abundant factors were accomplished by employing correlation analysis to discover probable correlations between parameters.

### 5.2.2 Data Splitting

The Pre-processed dataset must be broken down into a training and testing set as a final step before employing the machine learning model.

### 5.2.3. Training:

The dataset was trained using five different Machine Learning techniques

**K-Nearest Neighbour**

By locating the N nearest neighbours of the given points and assigning the class to the majority of those neighbours, the K-Nearest Neighbour algorithm seeks to categorise. Both classification and regression barriers can be overcome with it.

**Decision Tree**

Despite being a supervised learning technique, Decision Tree is most typically employed to address categorization problems. It can also be used to address issues with regression. To determine the category of a given dataset, the approach begins at the decision tree's root node. This approach makes advantage of the branches and advances to the next node depending on comparison.

**Random Forest**

The findings of a model known as random forest are based on each of the multiple alternative base models that are used on different subsets of the input data. The cornerstone of the random forest model is a decision tree, which combines its benefits with the extra effectiveness of integrating several models.

**Support Vector Machine**

Although they can also be used for regression, support vector machines, or SVMs, are typically utilized for classification. For categorization purposes, input and desired output data are both provided to supervised learning systems in machine learning.

**Gradient Boosting Algorithm**

In most competitions, this is the most recent algorithm employed. A differentiable loss function can be optimised using the additive model that is used. It was employed with a loss function of "ls," a minimum sample size of 2, and a learning rate of 0.01.

### 5.3 Evaluation

These mentioned methodologies predict the algorithm with the finest probability using a voting classifier.

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

## 6. Results

Several metrics are employed to gauge how efficient machine learning algorithms are

### 6.1 Accuracy

The proportion of the model's correct predictions over all observed values is known as accuracy. The proposed methodology accuracy is shown in Figure2

Accuracy = (TP + TN) /(TP + FP + TN+FN)

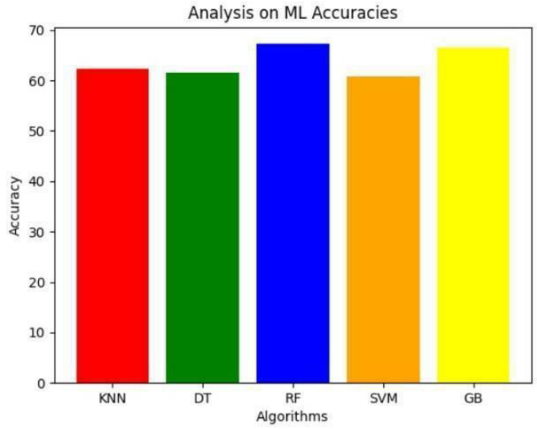



**Figure 2. Analysis on ML accuracies**

### 6.2 Precision

Any positive samples that are correctly or wrongly categorised as positive should be taken into account in Precision. The proposed methodology precision is shown in Figure 3.

Precision = (TP)/(TP + FP)

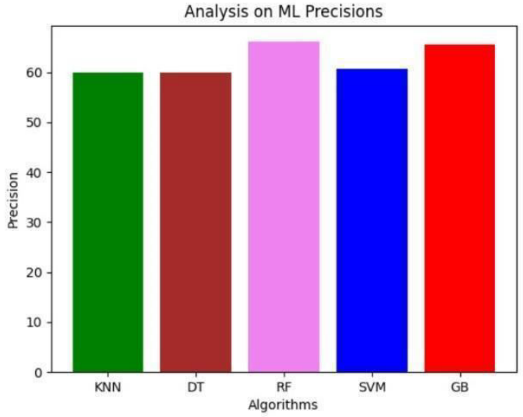


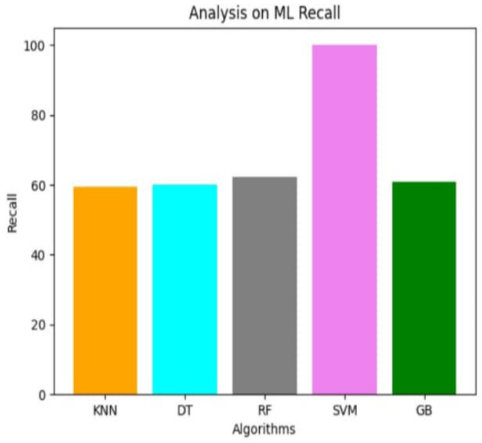

**Figure 3. Analysis on ML Precision**

### 6.3 Recall

Correctly identifying all positive samples is important to recall. The proposed methodology recall is shown in figure 4

Recall = (TP)/ (TP + FN)

**Figure 4. Analysis on ML Recall**




**6.4 F1 Score:** The F1 Score takes into account both factors and more accurately depicts overall accuracy. It has a scale of 0 to 1. The accuracy of the outcome increases with score. The proposed methodology F1-Score is shown in Figure 5.

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

F1 Score=(2 Precision*Recall) /(Precision + Recall).

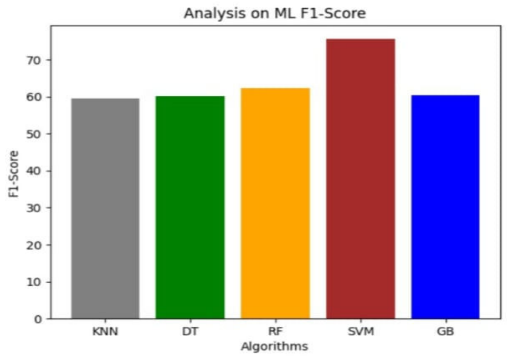



**Figure 5. Analysis on F1-Score**

### 6.5 ML Evaluations

How to evaluate the effectiveness of a machine learning model, as well as its benefits and drawbacks, is known as ML evaluation. It is shown in Table2

| Techniques | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| KNN | 62.28287841191067 | 59.923988804127106 | 59.32183730281473 | 59.4223635400106 |
| DT | 61.53846153846154 | 59.99059529256265 | 60.13119909268997 | 60.04158004158005 |
| RF | 67.24565756823822 | 66.10908968398356 | 62.32472419837096 | 62.218750000000014 |
| SVM | 60.7940446650124 | 60.69651741293532 | 100.0 | 75.54179566563467 |
| GB | 66.50124069478908 | 65.57295869708753 | 60.94313846788329 | 60.41689403033942 |

**Table 2. ML Evaluation**

### 6.6 Confusion Matrix

How effectively a categorization system performs is shown in a table called a confusion matrix. Confusion Matrix of KNN is shown in Table 3, Confusion Matrix of RF is shown in Table 4, Confusion Matrix of SVM is shown in Table 5, Confusion Matrix of GB is shown in Table 6

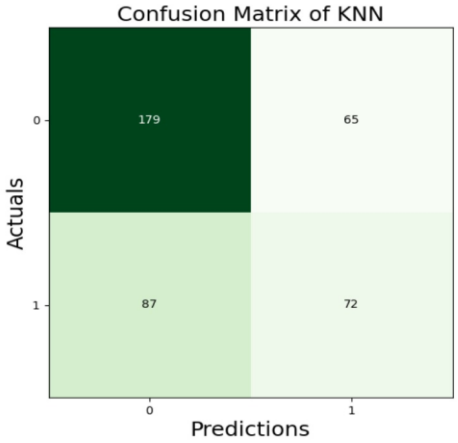



**Table 3. Confusion Matrix of KNN**

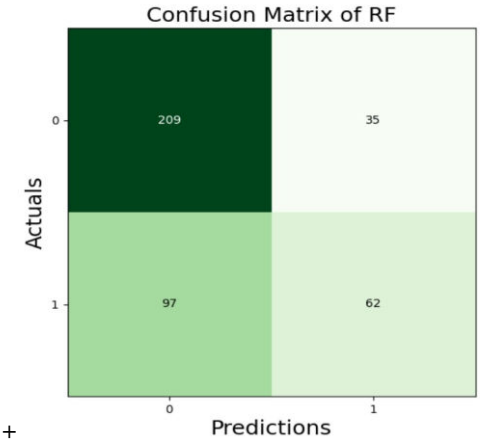



**Table 4. Confusion Matrix of RF**

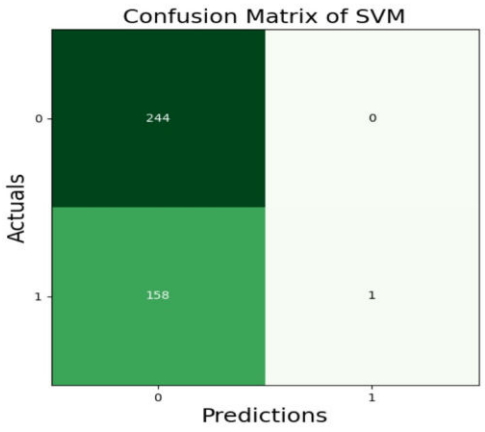



**Table 5. Confusion Matrix of SVM**

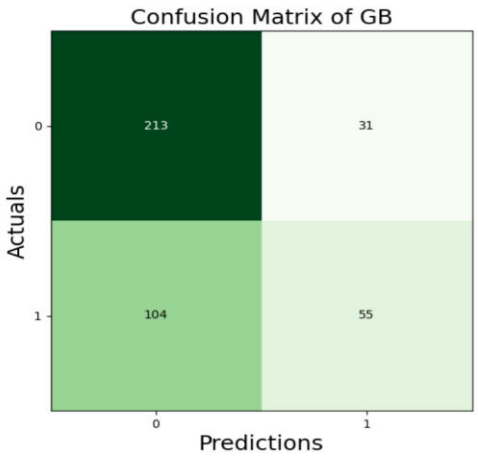



**Table 6. Confusion Matrix of GB**

## 7. Conclusion and Future Work

WQI controls the quality of the water, which is one of our most important resources for survival. It used to need a pricey and drawn-out lab procedure to test the water quality. The purpose of this work was to investigate an innovative machine learning method for predicting water quality.

Nevertheless, in this manner, we trained the model using data that was acquired from Kaggle rather than using the actual concentrations of contaminants in the water. So, in the future there is a chance to construct an IoT system in the future that can forecast water quality based on real-time data provided to the IoT system.

## 8. References

1.PCRWR. National Water Quality Monitoring Programme, Fifth Monitoring Report (2005-2006); Pakistan Council of Research in Water Resources Islamabad, Pakistan, 2007, Available in online: http://www.pcrwr.gov.pk/Publications/Water%20Quality%20Reports/Water%20Quality%20 Monitoring%20Report%202005-06.pdf (accessed on 23 August 2019).

2. Mehmood, S.; Ahmed, A.; Khalid, N.; Javed, T. Drinking Water Quality in Capital City of Pakistan. Open Access Sci. Rep. 2013, 2. [CrossRef]

3. PCRWR. Water Quality of Filtration Plants, Monitoring Report; PCRWR: Islamabad, Pakistan, 2010. Available online: http://www.pcrwr.gov.pk/Publications/water%20Quality%20Reports/FILTRTAION%20PLANTS% 20REPOT-CDA.pdf

4. Gazza, N.M.; Yusoff, M.K.; Aris, A.Z.; Jahir. H.; Ramli, M.F. Artificial neural network modelling of the water quality index for Kinta River (Malaysia) using water quality variables as predictors. Mar. Pollute. Bull. 2012, 64, 2409-2420. [CrossRef]

5. Daud, M.K.; Nafees, M.; Ali, S.; Rizwan, M.; Bajwa, R.A.; Shakoor, M.B.; Arshad, M.U.; Chatha, S.A.S.; Deeba, F.; Murad, W.; et al. Drinking water quality status and contamination in Pakistan. BioMed Res. Int. 2017, 2017, 7908183. [CrossRef]

6. Alamgir, A.; Khan, M.N.A.; Hany; Shaukat, S.S.; Mehmood, K.; Ahmed, A.; Ali, S.J.; Ahmed, S. Public health quality of drinking water supply in Orangi town, Karachi, Pakistan. Bull. Environ. Pharmacal. Life Sci. 2015, 4, 88–94.

7. Shafi, U.; Mumtaz, R.; Anwar, H.; Qamar, A.M.; Khurshid, H. Surface Water Pollution Detection using Internet of Things. In Proceedings of the 2018 15th International Conference on Smart Cities: Improving Quality of Life Using ICT & IoT (HONET-ICT), Islamabad, Pakistan, 8–10 October 2018; pp. 92–96.

8. Ahmad, Z.; Rahim, N.; 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. [CrossRef]

9. Saki Zadeh, M. Artificial intelligence for the prediction of water quality index in groundwater systems. Model. Earth Syst. Environ. 2016, 2, 8. [CrossRef]

10.Abyaneh, H.Z. Evaluation of the multivariate linear regression and artificial neural networks in prediction of water quality parameters. J. Environ. Health Sci. Eng. 2014, 12, 40. [CrossRef]