

**SCHOOL OF COMPUTER SCIENCE AND APPLICATIONS**

A Project Synopsis

on

**Water Quality Prediction**

Master of Science in Data Science

Submitted by

**Bibin Mathew, Oliver North Rogers III**

**R22DG010, R22DG037**

Under the guidance of

Internal Guide

**Dr. Ambili**

November 22, 2023

Rukmini Knowledge Park, Kattigenahalli, Yelahanka, Bengaluru-560064

[www.reva.edu.in](http://www.reva.edu.in)

**CERTIFICATE**

This is to certify that the Minor project work entitled “\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_” submitted to the School of Computer Science and Applications, REVA University in partial fulfillment of the requirements for the award of the Degree of MSc in Data Science in the academic year 2023-2024 is a record of the original work done by **(Name ,SRN)** under my supervision and guidance and that this Minor project work has not formed the basis for the award of any Degree / Diploma / Associate ship / Fellowship or similar title to any candidate of any University.

Place:

Date :

Internal Guide Signature:

Water Quality Prediction

# **Introduction**

Water pollution is a result of the contamination of water bodies (such as rivers, lakes, oceans, groundwater, and even drinking water sources) by various harmful substances. These contaminants can be caused by natural sources or human activities, leading to a decrease in water quality.

Water pollution significantly impacts human health by introducing harmful contaminants into drinking water sources, leading to widespread waterborne diseases such as cholera, typhoid fever, and gastrointestinal illnesses. These contaminants, ranging from pathogens to toxic chemicals and heavy metals, pose acute and chronic health risks, causing organ damage, developmental issues, and even cancer. Simultaneously, water pollution disrupts ecosystems by deteriorating water quality, reducing oxygen levels, and introducing toxins that harm aquatic life. This disruption affects biodiversity, hampers natural processes, and leads to the decline of fisheries, altering entire ecosystems and compromising their resilience to further environmental stressors.

When water pollution problems persist without effective resolution, the consequences intensify across multiple dimensions. Continual pollution leads to the destruction of aquatic ecosystems, resulting in the loss of biodiversity, the destruction of fisheries, and the creation of oxygen-depleted "dead zones" where marine life cannot survive. Lengthy exposure to contaminated water sources poses severe health risks to humans, causing outbreaks of waterborne diseases, long-term health issues like organ damage and cancer, and even fatalities, particularly in vulnerable communities lacking access to clean water. Economic repercussions follow, with the reduced availability of clean water affecting agriculture, industry, and overall social well-being. The environment suffers from lasting contamination, destroying not only aquatic life, but also terrestrial ecosystems linked to water sources. The failure to address water pollution perpetuates a cycle of environmental degradation, health crises, economic difficulties, and ecological imbalance with far-reaching and interconnected impacts.

Water quality prediction plays a pivotal role in safeguarding human health and preserving ecosystems. This study aims to develop a robust predictive model for assessing and forecasting water quality parameters using machine learning techniques. Leveraging historical water quality data, meteorological factors, land use characteristics, and pollution sources, a predictive model is constructed to anticipate changes in water quality metrics. Various machine learning algorithms, including but not limited to regression, decision trees, and ensemble methods, are employed and compared for their efficacy in accurately predicting water quality parameters. The model's performance is evaluated using relevant evaluation metrics, emphasizing its precision, recall, accuracy, and root mean square error. Additionally, the study explores the impact of different feature sets and temporal variations on the model's predictive capabilities. The findings demonstrate the potential of machine learning-based predictive models in providing early warnings, aiding decision-making processes, and fostering proactive measures to ensure and enhance water quality for environmental sustainability and public health.

## **Scope of the Problem**

In India, a significant portion of the population, particularly in rural areas, has had limited awareness regarding water quality issues. While there's an increasing focus on water-related problems, especially in urban areas facing acute water scarcity or pollution, rural regions and marginalized communities often have limited awareness of these issues.

Several factors contribute to this lack of awareness:

**Limited Access to Information**: In rural areas, access to information about water quality, its importance, and potential health risks can be scarce. Information spreading and education programs often face challenges in reaching remote or underserved communities.

**Educational Gaps**: Many people, particularly in rural settings, might not have adequate education or awareness about the potential dangers of contaminated water sources. Understanding the importance of testing water quality or the risks associated with using polluted water for drinking or sanitation purposes might be lacking.

**Reliance on Traditional Sources**: In certain areas, reliance on traditional water sources such as wells, ponds, or rivers without proper testing or understanding of water quality standards prevails. Communities may continue to use these sources due to historical practices or lack of alternatives, despite potential contamination risks.

**Rapid Urbanization and Industrial Growth**: Urban centers experiencing rapid urbanization and industrialization face heightened water pollution concerns. These areas often grapple with challenges related to industrial discharge, inadequate wastewater treatment, and contamination from various sources, necessitating greater attention and action to address water quality issues.

## **Intent to address the Problem**

Water prediction models serve as powerful tools for individual awareness and understanding of water quality. These models utilize various data inputs, such as historical water quality data, weather patterns, land use, and pollution sources, to forecast changes in water quality. Leveraging these models can significantly enhance individual awareness. By predicting changes in water quality, these models can issue alerts or warnings to individuals in areas prone to contamination. This proactive approach enables people to take precautionary measures, such as using alternative water sources, boiling water for consumption, or employing water treatment methods when the quality is compromised. Understanding the predictive trends in water quality encourages behavioral changes among individuals. When people are aware of potential pollution sources or how their actions can impact water quality, they are more likely to adopt responsible practices. For instance, farmers might adjust their fertilizer use to minimize runoff, households might properly dispose of waste to prevent contamination, or industries might implement better wastewater treatment.

In conclusion, the use of water prediction models plays a crucial role in empowering individuals and communities with information crucial to making informed decisions regarding their water usage. By raising awareness, enabling proactive measures, encouraging responsible behavior, and fostering community engagement, these models become essential tools in ensuring access to clean and safe water for all.

# **Data Set and description of attributes**

The dataset that will be used for training our model is “Water Quality”. It contains a file with water quality metrics for 3276 instances of different water bodies and 10 attributes. This dataset was downloaded from Kaggle Datasets Repository. Link to the dataset <https://www.kaggle.com/datasets/adityakadiwal/water-potability/data>.

**Dataset Attributes**

**pH value**: PH is an important parameter in evaluating the acid–base balance of water. It is also an indicator of acidic or alkaline condition of water status. WHO has recommended maximum permissible limit of pH from 6.5 to 8.5. The current investigation ranges were 6.52–6.83 which are in the range of WHO standards.

**Hardness**: hardness is mainly caused by calcium and magnesium salts. These salts are dissolved from geologic deposits through which water travels. The length of time water is in contact with hardness producing material helps determine how much hardness there is in raw water. Hardness was originally defined as the capacity of water to precipitate soap caused by Calcium and Magnesium.

**Solids (Total dissolved solids - TDS)**: water can dissolve a wide range of inorganic and some organic minerals or salts such as potassium, calcium, sodium, bicarbonates, chlorides, magnesium, sulfates etc. These minerals produced un-wanted taste and diluted color in appearance of water. This is an important parameter for the use of water. The water with high TDS value indicates that water is highly mineralized. The desirable limit for TDS is 500 mg/l and maximum limit is 1000 mg/l which prescribed for drinking purposes.

**Chloramines**: chlorine and chloramine are the major disinfectants used in public water systems. Chloramines are mostly formed when ammonia is added to chlorine to treat drinking water. Chlorine levels up to 4 milligrams per liter (mg/L or 4 parts per million (ppm)) are considered safe in drinking water.

**Sulfate**: sulfates are naturally occurring substances that are found in minerals, soil, and rocks. They are present in ambient air, groundwater, plants, and food. The principal commercial use of sulfate is in the chemical industry. Sulfate concentration in seawater is about 2,700 milligrams per liter (mg/L). It ranges from 3 to 30 mg/L in most freshwater supplies, although much higher concentrations (1000 mg/L) are found in some geographic locations.

**Conductivity**: pure water is not a good conductor of electric current rather is a good insulator. Increase in ions concentration enhances the electrical conductivity of water. Generally, the number of dissolved solids in water determines the electrical conductivity. Electrical conductivity (EC) measures the ionic process of a solution that enables it to transmit current. According to WHO standards, EC value should not exceed 400 μS/cm.

**Organic Carbon**: Total Organic Carbon (TOC) in source waters comes from decaying natural organic matter (NOM) as well as synthetic sources. TOC is a measure of the total amount of carbon in organic compounds in pure water. According to US EPA < 2 mg/L as TOC in treated / drinking water, and < 4 mg/Lit in source water which is use for treatment.

**Trihalomethanes**: THMs are chemicals which may be found in water treated with chlorine. The concentration of THMs in drinking water varies according to the level of organic material in the water, the amount of chlorine required to treat the water, and the temperature of the water that is being treated. THM levels up to 80 ppm are considered safe in drinking water.

**Turbidity**: the turbidity of water depends on the quantity of solid matter present in the suspended state. It is a measure of light emitting properties of water and the test is used to indicate the quality of waste discharge with respect to colloidal matter. The mean turbidity value obtained for Wondo Genet Campus (0.98 NTU) is lower than the WHO recommended value of 5.00 NTU.

**Potability**: indicates if water is safe for human consumption where 1 means Potable and 0 means Not Potable.

# **Methodology Diagram**

Predicting water quality involves analyzing different parameters and patterns within water bodies. The algorithms used to create this model include:

* **Linear Regression**: Simple and effective for establishing relationships between water quality parameters and other variables. Predicts a continuous numeric value, making it suitable for basic predictions.
* **Support Vector Machines (SVM)**: SVM is used for classification and regression tasks. It works well with small to medium-sized datasets and can handle complex relationships between input variables. SVM seeks to find an optimal hyperplane that best separates data points into different classes or predicts continuous values.
* **Random Forest (RF)**: Random Forest is an ensemble learning method that builds multiple decision trees and merges their predictions. It's useful for handling large datasets and capturing complex interactions among various water quality parameters.
* **Gradient Boosting Machines (GBM)**: GBM builds multiple decision trees sequentially, where each tree corrects the errors of its predecessor. XGBoost and LightGBM are popular GBM implementations used for water quality prediction due to their efficiency and accuracy.
* **Decision Trees (DT):** Decision trees break down data into smaller subsets based on different attributes. They're easy to interpret and visualize. However, they might not capture complex relationships as effectively as ensemble methods like Random Forests.
* **Artificial Neural Networks (ANN)**: ANNs simulate the human brain's neural structure to process information. They're effective for complex, nonlinear relationships in water quality data. Multi-layer perceptron (MLP) is a popular ANN architecture used for water quality prediction.

Some decision metrics are used, including accuracy, precision, recall, and F1 score. Here's an explanation of each metric and how to calculate them:

* **Accuracy:** Accuracy measures the overall correctness of the system's predictions. It calculates the ratio of correctly predicted instances to the total number of instances.

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

Where:

* TP (True Positive) represents the number of instances that were correctly classified as positive (credit score clarified).
* TN (True Negative) represents the number of instances that were correctly classified as negative (credit score not clarified).
* FP (False Positive) represents the number of instances that were incorrectly classified as positive.
* FN (False Negative) represents the number of instances that were incorrectly classified as negative.
* **Precision:** Precision focuses on the accuracy of positive predictions. It calculates the ratio of true positive predictions to the total number of positive predictions made by the system.

**Precision = TP / (TP + FP)**

Precision provides an indication of how many of the positive predictions are correct.

* **Recall:** Recall measures the ability of the system to correctly identify positive instances. It calculates the ratio of true positive predictions to the total number of actual positive instances.

**Recall = TP / (TP + FN)**

Recall provides an indication of how well the system "recalls" or captures the positive instances.

* **F1 Score:** The F1 score combines precision and recall into a single metric to provide an overall evaluation of the system's performance. It is the harmonic means of precision and recall.

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

The F1 score gives equal weight to precision and recall, providing a balanced measure of the system's performance.

These metrics will allow us to evaluate the performance of our models by considering both the positive and negative instances.

The methodology can be described in a 5-step process.

* Load the respective dataset
* Preprocess the dataset.
* Apply various algorithms suitable for prediction.
* Evaluate the derived models using performance metrics and error measure.
* Select the best model with high accuracy and low error rate.

#### **FLOW CHART**

Load The Water Quality Dataset

Preprocess the Data

Apply various algorithms suitable for prediction

Evaluate the models using performance metrics

Select the Best Model

# **Software Requirement for Project**

To create this machine learning (ML) model, we’ll need a few software components and libraries:

**Integrated Development Environment (IDE)**: Install any development environment that supports Python programing language. Popular IDEs include Jupyter Notebook, Anaconda, PyCharm, VS Codes, etc. We used Jupyter Notebook to create our Machine Learning Model.

**Python**: We used Python as our programming language. Jupyter Notebook is typically used with Python, so you'll need a Python installation on your system.

**Data Manipulation Libraries**: NumPy and Pandas libraries are used for data manipulation and analysis.

**Machine Learning Libraries**: Libraries such as Scikit-Learn, and XGBoost are being used for Machine Learning.

**Visualization Libraries**: Matplotlib and Seaborn libraries are used for data visualization.

# **References**

* Devendra Dohare, Shriram Deshpande and Atul Kotiya. "Analysis of Ground Water Quality Parameters: A Review". Research Journal of Engineering Sciences, ISSN 2278 – 9472, Vol. 3(5), 26-31, May (2014).
* S. P. Gorde1, M. V. Jadhav2. "Assessment of Water Quality Parameters: A Review", S. P. Gorde et al Int. Journal of Engineering Research and Applications, ISSN: 2248-9622, Vol. 3, Issue 6, Nov-Dec 2013, pp.2029-2035.
* Gujrati, Ashwin, Jha, Vibhuti Bhushan, Nidamanuri, Rama Rao. "Satellite-based Optical Water Type Classification of Inland Waters Bodies of India", 2023 International Conference on Machine Intelligence for GeoAnalytics and Remote Sensing (MIGARS) | 979-8-3503-4542-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/MIGARS57353.2023.10064493.
* S. Babu, Banavath Baby Nagaleela, Cheekurimelli Ganesh Karthik and Lakshmi Narayana Yepuri. "Water Quality Prediction using Neural Networks", 2023 International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF) | 979-8-3503-3436-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICECONF57129.2023.10084120.
* Kishan Singh Rawat, K.K. Gupta. "Appraisal of Groundwater Quality using Artificial Neural Network (ANN): A case study of Penisular”. India, 2023 International Conference on Advancement in Computation & Computer Technologies (InCACCT) | 979-8-3503-9648-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/InCACCT57535.2023.10141707.