Machine Learning Approaches to Water Quality Forecasting

Submitted By:
BOYA DARSHAN REDDY(AM.EN.U4CSE22113)
MODIUM VEERA SAI NITHEESH REDDY(AM.EN.U4CSE22134)
PALLURU MOULI(AM.EN.U4CE22140)

ABSTRACT

Formal Description of the Problem:

The goal of this project is to develop a machine learning model that predicts the quality of water based on various physicochemical parameters such as pH, turbidity, dissolved oxygen levels, etc., to determine if the water is safe for consumption or use in agriculture, livestock, or human activities. This is critical to ensuring public health and environmental safety.

Well-Posed Problem:

- **Task (T):**
  - The task is to predict the quality of water (A,B,C,D,E) based on physicochemical properties (pH, dissolved oxygen, conductivity, etc.). The output is typically a classification (e.g., safe or unsafe) or a water quality index score on a numerical scale.
  - Our major classifications are:
    - A - Drinking Water Source without conventional treatment but after disinfection.
    - B - Outdoor bathing (Organised).
    - C - Drinking water source after conventional treatment and disinfection.
    - D - Propagation of Wild life and Fisheries.
    - E - Irrigation, Industrial Cooling, Controlled Waste disposal
- **Experience (E):**
  - The model gains experience by training on a labeled dataset that consists of water samples with their corresponding features (e.g., pH, turbidity, etc.) and labels (e.g., safe or unsafe). Experience grows as the model processes more water samples in the training set and uses the patterns it identifies to make predictions on unseen data.
- **Performance Measure (P):**
  - The model's performance will be evaluated based on standard classification metrics such as accuracy, precision, recall, F1-score, and AUC-ROC. These metrics will indicate how well the model can classify water samples as safe or unsafe, improving as it learns from more data.

Problem Statement:

"Given a dataset of water samples with various physicochemical properties (pH, turbidity, dissolved oxygen, etc.), design a machine learning model that can accurately predict whether the water is safe for consumption. The model's performance will be measured using accuracy and other relevant metrics, and its ability to predict will improve with more data and better training, ensuring reliable water quality assessment."

1. INTRODUCTION

Motivation :

As a student, the motivation for solving this problem is in two ways. First, it serves as an excellent learning opportunity to apply machine learning techniques in real-world scenarios, especially in the domain of environmental science and public health. By working on this project, we can gain hands-on experience with predictive modeling, data analysis, and classification tasks, which are essential skills.

Second, successfully predicting water quality can contribute to addressing global issues like access to clean drinking water and environmental conservation. This project allows us to apply our academic knowledge to solve meaningful, impactful problems while honing your technical expertise for future projects or professional development.

Benefits of solution :

Improved Water Quality Assessment: The machine learning model developed will enable automatic and accurate predictions of water quality based on physicochemical properties. This provides a quick and reliable way to determine whether water is safe for consumption or agricultural use, reducing the need for costly and time-consuming laboratory testing.*?*

Environmental and Public Health Impact: The solution has the potential to positively impact public health and environmental safety by providing a scalable, automated tool to monitor water quality. This could help local communities, governments, and organizations ensure that water sources meet health and safety standards.

Solution Use

The solution can play a vital role in real-life water quality problems.

Operationalization:

- The solution could be deployed as part of an automated system that continually assesses water quality in real-time using sensor data, or periodically when new water samples are collected.

Expected Lifetime:.

- **Long-Term (Operational)**: The model may need to run over the long term, possibly years. This requires the model to be regularly updated with new data and retrained periodically to ensure that it adapts to changing environmental conditions and new water quality parameters. We are expecting it to be used in real-life water quality prediction system(as water plays an important role in our life). Though,many technologies may come but we expect this should lay the foundation of all that future products. We are expecting long-life by these Maintenance Considerations:

    1. Data Updates
    2. Model Performance Monitoring
    3. Software Requirements

4. Adaptation to New Technologies

Functional and Non-Functional Requirements:

Functional:

- The solution must accurately predict water quality based on the given features.
- It must handle real-time data inputs (in an operational setting) or batch processing for periodic reporting.

Non-Functional:

- **Scalability**: The solution should be able to scale efficiently if applied to monitor water quality in multiple locations.
- **Reliability**: The system must be robust and capable of running without frequent interruptions.
- **Maintainability**: The codebase and model must be easy to maintain, update, and retrain as new data becomes available.

ML Algorithms

To solve the task of predicting water quality based on physicochemical properties, various machine learning algorithms can be employed. These algorithms are designed to handle classification problems and can be compared based on performance using specific evaluation metrics.

1. Logistic Regression

- **Description**: A basic linear model that can handle binary or multi-class classification problems. It is often used as a baseline model for comparison.
- **Suitability**: Works well if the relationship between the features and the target variable is line

2. Support Vector Machines (SVM)

- **Description**: A powerful algorithm for classification tasks that attempts to find a hyperplane that best separates the different classes.
- **Suitability**: Effective for high-dimensional data, especially when classes are separable. However, it's computationally expensive on larger datasets.

3. Random Forest

- **Description**: An ensemble method that builds multiple decision trees and merges them to improve accuracy and avoid overfitting.
- **Suitability**: Handles complex data and can capture non-linear relationships between features. Works well with large datasets.

4. K-Nearest Neighbors (KNN)

- **Description**: A simple algorithm that classifies a sample based on the majority class of its k nearest neighbors.
- **Suitability**: Works best with smaller datasets where the decision boundary is simple. Struggles with large datasets and complex feature spaces.

To evaluate the performance of the machine learning models, the following metrics will be used:

1. **Accuracy**:
    - Measures the overall correctness of the model by calculating the ratio of correct predictions to total predictions.
    - Formula:
      Accuracy = (True Positives + True Negatives) / Total Samples
2. **Precision**:
    - The ratio of correctly predicted positive observations to the total predicted positives. This metric is useful when the cost of false positives is high.
    - Formula:
      Precision = True Positives / (True Positives + False Positives)
3. **Recall** (Sensitivity or True Positive Rate):
    - The ratio of correctly predicted positive observations to all actual positives.
    - Formula:
      Recall = True Positives / (True Positives + False Negatives)
4. **F1-Score**:
    - The harmonic mean of precision and recall, providing a balance between the two. This metric is useful when dealing with imbalanced classes.
    - Formula:
      F1-Score = 2 * (Precision * Recall) / (Precision + Recall)
5. **AUC-ROC** (Area Under the Receiver Operating Characteristic Curve):
    - Measures the model's ability to distinguish between classes. It evaluates the performance across all classification thresholds.
    - AUC values range from 0 to 1, with 1 representing perfect classification.

2. Dataset finalization

https://ueppcb.uk.gov.in/pages/display/96-water-quality-data
This is where we have took information, we are analysing information from past three years.

The data is a monthly record of water quality parameters for a rivers in **India**, specifically monitoring various physicochemical and biological characteristics of the water. It includes temperature, pH, total dissolved solids (TDS), dissolved oxygen (DO), hardness, and concentrations of pollutants like nitrates and sulphates. Additionally, it measures biological contamination such as total coliforms (TC) and fecal coliforms (FC), which are important indicators of water safety.

Each month's data is assessed against the **Designated Best Use (DBU)** water quality criteria, which helps determine whether the water is suitable for different purposes such as drinking (with or without treatment), irrigation, industrial use, or recreational activities. In this case, The data includes water quality classification labels ranging from **A** to **E**, each representing different levels of water quality and suitability for specific uses. Here's a general breakdown:

- **A**: Drinking Water Source without conventional treatment but after disinfection.
- **B**: Water suitable for outdoor bathing (organized bathing areas).
- **C**: Drinking water source after conventional treatment and disinfection.
- **D**: Water suitable for the propagation of wildlife and fisheries.
- **E**: Water suitable for irrigation, industrial cooling, and controlled waste disposal

The dataset provided appears to have **25 features** (columns), excluding the month and temperature columns. Here's a breakdown of each feature and its importance in assessing water quality:

1. **Month**:
    - Describes the month when the data was collected.
    - Importance: Helps track seasonal variations in water quality.
2. **Temp (°C)**:
    - Water temperature in degrees Celsius.
    - Importance: Temperature affects the solubility of gases (like oxygen) and the rate of chemical and biological reactions. It's crucial for the aquatic life ecosystem.
3. **pH**:
    - The measure of the acidity or alkalinity of water.
    - Importance: pH affects the chemical solubility and biological availability of nutrients and heavy metals. The ideal pH for most aquatic life ranges between 6.5 and 8.5.
4. **TDS (Total Dissolved Solids)**:
    - The amount of dissolved substances in water, usually expressed in mg/l.
    - Importance: High TDS can indicate poor water quality, which affects drinking water taste and can harm aquatic life.
5. **EC (Electrical Conductivity, µS/cm)**:
    - Measures the water's ability to conduct electrical current, which increases with dissolved salts.
    - Importance: High conductivity often indicates a high level of dissolved salts, which can be harmful to agriculture, wildlife, and human consumption.

6. **DO (Dissolved Oxygen, mg/l)**:
    - The amount of oxygen dissolved in the water.
    - Importance: Essential for aquatic life. Low DO levels can indicate pollution, which may be harmful to all aquatic organisms.
7. **Chloride (mg/l)**:
    - Concentration of chloride ions in the water.
    - Importance: High chloride levels may indicate contamination from human or animal waste, industrial pollution, or deicing salt used on roads.
8. **Hardness (mg/l)**:
    - Measure of the concentration of calcium and magnesium salts.
    - Importance: Water hardness affects the taste of water and can also contribute to scaling in pipes. It influences the suitability of water for domestic use.
9. **Calcium (mg/l)**:
    - Concentration of calcium ions.
    - Importance: Calcium is important for bone health in humans but can contribute to water hardness and scaling issues in water pipes.
10. **Magnesium (mg/l)**:

- Concentration of magnesium ions.
- Importance: Like calcium, magnesium contributes to water hardness and affects the suitability of water for consumption and industrial use.

11. **Alkalinity (mg/l)**:

- The water's ability to neutralize acids.
- Importance: Alkalinity helps maintain a stable pH in water bodies and protects aquatic organisms from drastic pH changes.

12. **BOD (Biochemical Oxygen Demand, mg/l)**:

- The amount of dissolved oxygen needed by aerobic biological organisms to break down organic material.
- Importance: High BOD indicates high levels of organic pollution, which can lead to oxygen depletion in water bodies.

13. **COD (Chemical Oxygen Demand, mg/l)**:

- The amount of oxygen required to chemically oxidize organic compounds.
- Importance: High COD can indicate high levels of pollution from industrial waste or untreated sewage.

14. **TC (Total Coliform, MPN/100 ml)**:

- The count of total coliform bacteria present in water.
- Importance: Coliform bacteria indicate the presence of harmful pathogens and fecal contamination, which can pose health risks for human consumption.

15. **FC (Fecal Coliform, MPN/100 ml)**:

- The count of fecal coliform bacteria present in water.

- Importance: Indicates the level of contamination by human or animal feces, which can lead to serious waterborne diseases.

16. **Fecal Streptococcus (MPN/100 ml)**:

- Concentration of fecal streptococcus bacteria.
- Importance: Like fecal coliform, this is used as an indicator of fecal contamination and potential presence of pathogens.

17. **Nitrate-N (mg/l)**:

- Concentration of nitrogen in the form of nitrates.
- Importance: High nitrate levels can indicate agricultural runoff and are harmful if present in drinking water, especially for infants (can cause "blue baby syndrome").

18. **Nitrite-N (mg/l)**:

- Concentration of nitrogen in the form of nitrites.
- Importance: Similar to nitrate, nitrites are harmful at high levels and can indicate pollution from fertilizers or sewage.

19. **Sulphate (mg/l)**:

- Concentration of sulfate ions.
- Importance: High sulfate levels can cause gastrointestinal discomfort and can be indicative of industrial pollution.

20. **Phosphate (mg/l)**:

- Concentration of phosphate ions.
- Importance: Phosphates can lead to eutrophication, causing algal blooms and depletion of oxygen, which can harm aquatic life.

21. **Fluoride (mg/l)**:

- Concentration of fluoride ions.
- Importance: Fluoride is beneficial for dental health in small amounts, but excessive levels can cause fluorosis, affecting teeth and bones.

22. **Sodium (mg/l)**:

- Concentration of sodium ions.
- Importance: High sodium levels affect water taste and can be harmful to people with certain medical conditions (e.g., hypertension) and to crops sensitive to salt.

23. **Potassium (mg/l)**:

- Concentration of potassium ions.
- Importance: Essential nutrient for plants, but high levels in water can indicate pollution or agricultural runoff.

24. **SAR (Sodium Adsorption Ratio)**:

- A measure of the sodium content relative to calcium and magnesium.
- Importance: High SAR values can reduce soil permeability, affecting irrigation and plant growth.

25. **Designated Best Use (DBU) Water Quality Criteria**:

- Indicates the designated best use of water based on quality (A to E categories).
- Importance: This feature categorizes the water according to its suitability for human consumption, agriculture, industry, or wildlife preservation, which is a key outcome of the water quality assessment process.

These features collectively provide a comprehensive picture of water quality and its suitability for various uses, ranging from drinking water to agricultural irrigation and wildlife support.

As of our research, we have observed that this data is not used in any of research or any project until now. So, we expect it to be first and will be able to enhance our models based on these datasets.