# Project Proposal: Machine Learning-Based Water Quality Indexing

**Team BD**

**Abdullah Al Mofael, Fahim Muntasir Rabbi**

## 1. Introduction

Water quality assessment is essential for monitoring environmental health and public safety. Traditional water quality indexing methods rely on standard calculations that may not capture complex relationships between parameters. This project aims to develop a machine learning (ML) application using Python to predict water quality classification based on multiple water quality parameters. The model will provide a classification indicating how poor or good the water quality is by leveraging artificial neural networks (ANN), support vector machines (SVM), decision trees (DT), and k-nearest neighbors (K-NN).

## 2. Problem Statement

Existing water quality indices (WQI) rely on predefined formulas that may not generalize well to varying environmental conditions. By using ML models, we aim to improve classification accuracy and adaptability by learning from historical river water quality data. The model will classify water quality into categories (e.g., Excellent, Good, Moderate, Poor, Very Poor) based on threshold values for pH, electrical conductivity (EC), dissolved oxygen (DO), biological oxygen demand (BOD), chemical oxygen demand (COD), total dissolved solids (TDS), suspended solids (SS), total solids (TS), chlorine, total alkalinity, NH3-N, NO2-N, NO3-N, Phosphate, and turbidity.

## 3. Data Description

The dataset consists of river water quality measurements stored in CSV format. Each row represents a recorded sample from a specific location and time, while each column represents a water quality parameter. The dataset contains numerical values representing the concentration or level of each parameter.

### 3.1 Data Structure

The CSV file follows a structured format:

**Water Quality Data Structure**

| Sample_ID | pH | EC (µS/cm) | DO (mg/L) | BOD (mg/L) | COD (mg/L) | TDS (mg/L) | SS (mg/L) | TS (mg/L) | Chlorine (mg/L) | Total Alkalinity (mg/L) | Turbidity (NTU) | NH3-N (mg/L) | NO2-N (mg/L) | NO3-N (mg/L) | Phosphate (mg/L) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|

### 3.2 Data Source

Water Quality data was collected from various sampling locations of the Dhaleshwari river and Bhairab river, Bangladesh, during 2023. Sampling and analysis were conducted by Department of Environment, Bangladesh and Department of Environmental Science and Technology at Jashore University of Science and Technology, Bangladesh.

## 3.2 Description of Parameters

- **Sample_ID**: Unique identifier for each recorded sample.
- **pH**: Measures the acidity or alkalinity of water, ranging from 0 (acidic) to 14 (alkaline).
- **EC (Electrical Conductivity in µS/cm)**: Indicates the amount of dissolved ions in water, reflecting salinity and contamination levels.
- **DO (Dissolved Oxygen in mg/L)**: Represents oxygen availability for aquatic life; lower levels indicate pollution.
- **BOD (Biological Oxygen Demand in mg/L)**: Measures the amount of oxygen consumed by microorganisms; higher values suggest organic pollution.
- **COD (Chemical Oxygen Demand in mg/L)**: Quantifies oxygen required to break down pollutants; used for detecting industrial contamination.
- **TDS (Total Dissolved Solids in mg/L)**: Sum of all dissolved substances, affecting water clarity and taste.
- **SS (Suspended Solids in mg/L)**: Particulate matter floating in water, impacting water transparency and aquatic life.
- **TS (Total Solids in mg/L)**: Combined measure of dissolved and suspended solids.
- **Chlorine (mg/L)**: Indicator of disinfectants or industrial pollution in water.
- **Total Alkalinity (mg/L)**: Measures the water's buffering capacity against pH changes.
- **Turbidity (NTU - Nephelometric Turbidity Unit)**: Indicates water clarity, where higher values suggest high levels of suspended particles.
- **NH3-N (Ammonia-Nitrogen in mg/L)**: Represents ammonia concentration in water, which can be toxic to aquatic life at high levels.
- **NO2-N (Nitrite-Nitrogen in mg/L)**: A product of nitrogen cycle; high concentrations indicate contamination from wastewater or fertilizers.
- **NO3-N (Nitrate-Nitrogen in mg/L)**: A crucial nutrient for aquatic life, but excessive amounts lead to eutrophication and water pollution.
- **Phosphate (mg/L)**: A key nutrient for plant growth; excess levels contribute to algal blooms and oxygen depletion in water bodies.

## 3.3 Data Considerations

- **Acceptable Limits**: Each parameter has a regulatory threshold indicating safe water conditions. These limits will be used as decision boundaries for water quality classification.
- **Data Cleaning**: Handling missing values, duplicates, and outliers to ensure data integrity.
- **Temporal & Spatial Variability**: The dataset may include seasonal or location-based variations that need to be accounted for during modeling.

This structured dataset will serve as input for machine learning models to classify water quality into predefined categories based on established thresholds.

## 1. Methodology

The project will follow these stages:

### 4.1 Data Preprocessing

- Handling missing values
- Standardization/normalization of numerical features
- Outlier detection and removal

### 4.2 Feature Extraction

- Identifying key parameters influencing water quality classification
- Generating new features (e.g., ratio-based indicators) if necessary

### 4.3 Feature Analysis

- Correlation analysis between water quality parameters
- Principal Component Analysis (PCA) for dimensionality reduction

### 4.4 Model Development

- Implementing ML models: ANN, SVM, DT, K-NN

### 4.5 Model Assessment

- Performance evaluation using accuracy, precision, recall, and F1-score
- Cross-validation to ensure model robustness

### 4.6 Analysis of Results

- Comparing ML models with traditional WQI calculations

## 2. Expected Outcomes

- A trained ML model that classifies water quality into different categories
- Identification of the most influential parameters in water quality classification
- Performance comparison between ML models and traditional WQI methods

## 3. Timeline and Task Allocation

| Task | Duration |
|---|---|
| Data Collection & Preprocessing | 1 Week |
| Feature Extraction and Analysis | 1 Week |
| Model Development (ANN, SVM, DT, K-NN) | 2 Weeks |
| Model Assessment and Optimization | 2 Weeks |
| Analysis of Result and Report Compilation | 1 Week |