Project Document: Water Quality Analysis - Phase 1

Problem Statement:
The project aims to analyze water quality data to evaluate its suitability for various purposes, with a focus on drinking water. Key objectives include identifying potential issues or deviations from regulatory standards and determining water potability based on multiple parameters. We must be defining analysis objectives, collecting water quality data, designing relevant visualizations, and constructing a predictive model.

1. Analysis Objectives:

This step involves defining specific objectives for the water quality analysis. These objectives guide our entire analysis process and help us stay focused on the key goals of the project. Here's a breakdown of the objectives:

a. Assessing Potability:

- The primary objective is to determine whether the water is safe for human consumption (whether it is potable or non-potable).
- To achieve this, we will compare the water quality metrics, such as pH, Hardness, Solids, etc., against established regulatory standards.

b. Identifying Deviations:

- To ensure water quality meets safety standards, we need to identify any deviations from these standards.
- This objective involves pinpointing areas where water quality falls below acceptable levels. For example, if pH or TDS levels are too high or too low, it could indicate potential issues.

c. Parameter Relationships:

- Understanding the relationships between different water quality parameters is crucial for a comprehensive analysis.
- For instance, we will explore how pH correlates with other parameters like Hardness or Solids, and how these relationships can impact water potability.

2. Data Collection

pH value: Measures water acidity/alkalinity. WHO recommends a pH range of 6.5 to 8.5 for safe drinking water.
Hardness: Mainly caused by calcium and magnesium salts, affecting soap precipitation.
Solids (Total Dissolved Solids - TDS): High TDS indicates highly mineralized water; desirable limit for drinking is 500 mg/L.
Chloramines: Common disinfectants in public water systems.
Sulfate: Naturally occurring substances; concentrations vary but are generally within safe limits.

Conductivity: Measures water's ability to conduct electric current; WHO recommends EC value not exceeding 400 µS/cm.
Organic Carbon: Total Organic Carbon (TOC) measures the total amount of carbon in organic compounds.
Trihalomethanes (THMs): Vary based on water characteristics; safe levels are up to 80 ppm in drinking water.
Turbidity: Measures solid matter in suspended form; WHO recommends values below 5.00 NTU.
Potability: Indicates if water is safe for human consumption (1 means Potable, 0 means Not potable).

3. Visualization Strategy:

The visualization strategy is essential for gaining insights from the data. It involves creating visual representations of the data to better understand its characteristics and relationships. Here's a detailed explanation of this step:

a. Parameter Distributions:

- To understand the distribution of each parameter, we will create visualizations such as histograms, density plots, or box plots.
- These visualizations will help us identify patterns and outliers in the data.

b. Exploring Correlations:

- We will generate correlation matrices and heatmaps to visualize how different parameters are related to each other.
- This will help us uncover significant relationships and dependencies between parameters, which can provide valuable insights into water quality.

c. Potability Visualization:

- We will create visual cues or visualizations that represent water potability.
- This could include scatterplots or color-coded charts to distinguish between potable and non-potable water samples based on the analyzed parameters.

4. Predictive Modeling:

Predictive modeling involves building a machine learning model to predict water potability based on the selected parameters.

a. Algorithm Selection:

- We will choose appropriate machine learning algorithms for classification tasks.
- Common algorithms include logistic regression, decision trees, random forests, or support vector machines.

b. Feature Selection:

- Identifying the most relevant features (parameters) from the dataset is crucial.
- We will determine which parameters have the most significant impact on predicting water potability and include them in the model.

c. Data Splitting:

- We will divide the dataset into two parts: a training set and a testing set.
- The training set will be used to train the predictive model, while the testing set will be used to evaluate its performance.

d. Model Evaluation:

- We will evaluate the model's performance using various metrics, such as accuracy, precision, recall, and F1-score.
- Cross-validation techniques may be employed to ensure the model's robustness and generalizability.