**WATER QUALITY ANALYSIS**

**Harishma Sabu Ranjana S Ranjini S Roshni N Vithya S**

**Project Definition**

The project aims to analyze water quality data to determine the suitability of water for specific purposes, particularly for drinking. It involves identifying potential issues or deviations from regulatory standards and predicting water potability based on various water quality parameters. The project will encompass the following key components:

1. Analysis Objectives: The primary objectives are to assess water potability, identify deviations from regulatory standards, and understand relationships among water quality parameters.
2. Data Collection: Gather the provided water quality data, which includes parameters such as pH, Hardness, Solids, etc.
3. Visualization Strategy: Plan how to visualize parameter distributions, correlations, and potability status using appropriate visualization tools.
4. Predictive Modeling: Determine the machine learning algorithms and features to use for predicting water potability.

**Design Thinking**

**1. Analysis Objectives**

Objective 1: Assess Water Potability

* Description: This objective involves classifying water samples as either potable or non-potable (binary classification).
* Steps:
  + Establish criteria for water potability based on regulatory standards (e.g., WHO guidelines).
  + Develop a classification model to predict potability using supervised learning techniques.
  + Evaluate the model's performance using metrics such as accuracy, precision, recall, and F1-score.

Objective 2: Identify Deviations from Standards

* Description: Identify water quality parameters that deviate significantly from established regulatory standards.
* Steps:
  + Calculate summary statistics for each parameter.
  + Define thresholds or ranges for acceptable values based on regulations.
  + Generate reports or alerts for samples with out-of-spec parameters.

Objective 3: Understand Parameter Relationships

* Description: Explore relationships between different water quality parameters to gain insights.
* Steps:
  + Perform correlation analysis to assess the degree of association between parameters.
  + Visualize these relationships using scatter plots, heatmaps, or network graphs.
  + Interpret the visualizations to identify potential patterns and dependencies.

**2. Data Collection**

Data Source

* Dataset: The dataset for this project can be accessed from the following link: [Water Potability Dataset](https://www.kaggle.com/datasets/adityakadiwal/water-potability).
* Data Integrity: Verify the dataset's integrity and completeness, checking for missing values, duplicates, and data types.

Data Preprocessing

* Missing Values: Handle missing values through techniques such as imputation or removal.
* Outliers: Identify and address outliers in the data.
* Data Types: Ensure consistent data types for all variables.

Data Split

* Divide the dataset into training and testing subsets for model training and evaluation, typically using an 80-20 or 70-30 split ratio.

**3. Visualization Strategy**

Parameter Distributions

* Utilize data visualization tools (e.g., Matplotlib, Seaborn) to create histograms, box plots, or density plots to visualize the distributions of water quality parameters (e.g., pH, Hardness).
* Explore central tendencies, spread, and skewness of the distributions to understand their characteristics.

Correlation Analysis

* Calculate and visualize correlation coefficients (e.g., Pearson, Spearman) between pairs of parameters.
* Create correlation matrices and heatmaps to identify strong positive or negative correlations.
* Use scatter plots to visually represent relationships between selected pairs of parameters.

Potability Visualization

* Create visualizations to represent the potability status of water samples.
* Bar charts, pie charts, or stacked bar plots can be used to display the proportion of potable and non-potable samples.

**4. Predictive Modeling**

Algorithm Selection

* Choose suitable machine learning algorithms for binary classification tasks, such as Logistic Regression, Random Forest, Support Vector Machine (SVM), or Gradient Boosting.

Feature Engineering

* Engineer relevant features that may enhance the model's predictive performance, e.g., creating interaction terms or polynomial features.

Model Development

* Split the data into training and testing sets to facilitate model development and evaluation.
* Train the chosen classification algorithm(s) on the training data.
* Implement hyperparameter tuning to optimize model performance.

Model Evaluation

* Assess the model's performance using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.
* Use cross-validation techniques to ensure model robustness and avoid overfitting.

Recommendations

* Provide recommendations and insights based on the model's predictions.
* Highlight key features or parameters that strongly influence water potability.

**Conclusion**

In this initial phase, we have defined the project's objectives and outlined the design thinking process for analyzing water quality data to assess potability. The subsequent phases will involve data preprocessing, exploratory data analysis, model development, and evaluation. This systematic approach will enable us to effectively address the problem and provide valuable insights for water quality assessment and decision-making.