**Water Quality Analysis**

**PROBLEM DEFINITION:**

The Water Quality predictor problem is to develop a machine learning model that can identify potential issues or deviations from regulatory standards and determine water potability based on various parameters. Access to safe drinking water is essential to health, a basic human right, and a component of effective policy for health protection. This is important in terms of health and development issues at a national, regional, and local level. This model could be used to identify and address water quality issues in different parts of the world, and to ensure that people have access to safe drinking water.

The input to the model would be a set of water quality metrics, such as pH, hardness, conductivity, and turbidity. The output of the model would be a prediction of whether the water is potable or not.

**Project Understanding**

The primary goal of this project is to innovate in the field of water quality analysis by leveraging advanced techniques and technologies to provide more accurate and actionable insights. Specifically, we will analyse water quality data collected from monitoring stations in Tamil Nadu and develop novel strategies for pollution monitoring and management.

**DESIGN THINKING:**

**1. Data Collection:**

The dataset used in this project comes from Kaggle which is a high-fidelity database with a five-star research rating in Google Scholar. There are nine training variables, such as pH, hardness, solids, chloramines etc. The dataset link is [Water-Potability-Dataset](https://www.kaggle.com/datasets/adityakadiwal/water-potability).

**2. Data Preprocessing:**

Clean the data by addressing missing values and outliers. We can use techniques like imputation, removal, or interpolation. Normalize or scale data if necessary.

**3. Exploratory Data Analysis (EDA):**

Visualize the data using histograms, box plots, scatter plots, or heatmaps to understand the distribution and relationships between variables. Identify patterns and trends in water quality parameters.

**4. Feature Engineering:**

Create new features or derive relevant features from existing ones if needed

**5. Machine Learning or Statistical Modelling:**

* Split your dataset into training and testing sets for model evaluation.
* The model of machine learning can be divided into supervised learning, semi-supervised learning, and unsupervised learning. There are many kinds of algorithms for each learning mode, such as decision trees, naive Bayes, random forests, etc. Each algorithm model has its own advantages and disadvantages.
* Among the many machine learning methods, artificial neural networks and support vector machine algorithms became popular in machine learning due to their large processing data and fast calculation speed. Therefore, selecting the above two algorithms to judge the drinking ability of water resources is expected to better achieve the desired purpose.

**6. Model Evaluation:**

Evaluate the model's performance using relevant metrics such as accuracy, precision, recall, F1-score, and ROC AUC.

**7. Model Tuning:**

Adjust the hyperparameters of a model to improve its performance.

**8. Interpretation and Insights:**

Interpret the model results to understand which water quality parameters are most influential in determining potability and compare accuracy of each model.

**Algorithms Used**

**Logistic Regression**

Logistic Regression is named for the function used at the core of the method, the logistic function.

The logistic function, also called the sigmoid function, was developed by statisticians to describe properties of population growth in ecology, rising quickly and maxing out at the carrying capacity of the environment. It’s an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1, but never exactly at those limits.

**Support Vector Classifier**

The objective of a Linear SVC (Support Vector Classifier) is to fit the data you provide, returning a "best fit" hyperplane that divides, or categorizes your data.

**XGBoost**

XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements machine learning algorithms under the Gradient Boosting framework. XGBoost provides a parallel tree boosting (also known as GBDT, GBM).

**Random Forest Classifier**

A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

**Model Training**

The RFC model was trained using a dataset that included water quality parameters as features and Potability values as the target variable. By feeding the model with historical data, it learned the relationships between these parameters and potability values. The trained model was then saved as a pickle file for later use, enabling real-time predictions of potability value.

The project leverages this model to provide timely and actionable insights into water quality conditions, helping to address water quality challenges in Tamil Nadu.