## WATER QUALITY ANAYSIS

**PROJECT DEFINITION:**

The project involves analyzing water quality data to assess the suitability of water for specific purposes, such as drinking. The objective is to identify potential issues or deviations from regulatory standards and determine water potability based on various parameters. This project includes defining analysis objectives, collecting water quality data, designing relevant visualizations, and building a predictive model.

DESIGN THINKING:

1. **ANALYSIS OBJECTIVES:**

Define specific objectives for analyzing water quality data, including assessing potability, identifying deviations from standards, and understanding parameter relationships.

* **Evaluation of Potability of Water:**

Identify whether the water complies with the requirements and specifications set forth for safe drinking water.

Methods:

Calculate the parameters of water quality (such as pH, turbidity, and pollutants) and compare them to legal requirements (such as EPA, WHO) for drinking water.

To show the proportion of samples passing or failing potability requirements, create visuals like pie charts or bar graphs.

Determine which substances or parameters routinely exceed safe limits and call for cleanup.

* **Detecting Standards Deviations:**

Finding and recording occasions where water quality measurements depart from established environmental or regulatory norms is the goal.

Methods:

Data on the quality of the water tested is compared to specified standards and thresholds.

When parameters go above predetermined limits, create alerts or notifications to allow for quick response and correction.

Follow trends and potential reasons, such as pollution sources or seasonal variations, by keeping track of variances over time.

* **Understanding Parameter Relationships:**

Explore the connections and interactions between various water quality measures to learn more about the effects and processes on the environment.

Methods:

For a visual representation of the interactions between parameters, make scatter plots, heatmaps, or correlation matrices.

To measure the intensity and direction of relationships between parameters, use statistical analysis (such as correlation coefficients).

Examine cause-and-effect connections to comprehend how modifications to one parameter may have an influence on others (for example, nitrogen levels affecting algal blooms).

1. **Data Collection:**

Gather the provided water quality data containing parameters like pH, Hardness, Solids, etc. Here's a step-by-step guide on how to collect water quality data for parameters like pH, hardness, solids, and others:

* **Define Data Requirements:**

- Clearly define the specific parameters you need to collect data for. In this case, you mentioned pH, hardness, and solids, so ensure you have a complete list of desired parameters.

* **Identify Data Sources:**

- Determine where you can obtain the data. Potential sources include:

- Government agencies: Environmental or water quality agencies at the federal, state, or local levels often collect and maintain water quality data.

- Research institutions: Universities, research organizations, and environmental groups may have collected relevant data.

- Monitoring stations: Automated water quality monitoring stations at various water bodies.

- Field sampling: Conduct your own water sampling at specific locations if necessary.

* **Access Existing Data**:

- Contact relevant organizations or agencies to inquire about the availability of historical water quality data. They may provide datasets or access to their databases.

* **Collect Real-time Data:**

- If you require real-time or up-to-date data, consider using online databases or APIs provided by government agencies or research institutions. These sources often offer downloadable datasets or access to streaming data.

* **Design a Sampling Plan (if applicable):**

- If you need to collect new data through field sampling, develop a sampling plan. This plan should outline:

- Sampling locations: Identify the specific sites where you will collect water samples. These sites should be representative of the area of interest.

- Sampling frequency: Determine how often you will collect samples, whether daily, weekly, monthly, or seasonally.

- Sample collection methods: Specify the methods and equipment needed for collecting samples (e.g., water bottles, samplers, probes).

- Sample preservation: Describe how you will handle and store samples to maintain their integrity until analysis.

* **Data Collection:**

- Collect water samples according to your sampling plan. Follow standard procedures to ensure consistency and accuracy in data collection.

- Measure the parameters of interest using appropriate instruments and techniques. For example:

- pH: Use a pH meter or pH indicator solutions.

- Hardness: Conduct hardness titration tests.

- Solids: Determine total dissolved solids (TDS) or total suspended solids (TSS) using suitable methods.

* **Quality Control:**

- Implement quality control measures to ensure the accuracy and reliability of data. This may include:

- Calibrating instruments regularly.

- Running duplicate samples for comparison.

- Documenting field conditions and any deviations from the sampling plan.

* **Data Recording:**

- Record all data accurately, including sample location, date, time, and parameter values. Use a well-organized data sheet or electronic data logging system.

* **Data Storage**:

- Store collected data securely in a central database or file system. Ensure data is backed up regularly to prevent loss.

* **Data Analysis:**

- After data collection, analyze the collected data to derive insights, identify trends, and make assessments based on your specific objectives.

* **Data Sharing:**

- Share your findings and data with relevant stakeholders, including government agencies, research partners, or the public, as appropriate.

* **Compliance and Ethics:**

- Ensure that all data collection activities comply with relevant regulations, permits, and ethical guidelines, especially if working in sensitive environments.

By following these steps, you can gather water quality data effectively and ensure that the data you collect is reliable and suitable for your analysis and research purposes.

1. **Visualization Strategy:**

Plan how to visualize parameter distributions, correlations, and potability using suitable tools. Here's a strategy for visualizing parameter distributions, correlations, and potability using suitable tools:

* **Identify Key Parameters**:

- Begin by identifying the key water quality parameters relevant to your analysis. These may include pH, turbidity, dissolved oxygen, various contaminants, temperature, and more.

* **Data Collection and Preprocessing:**

- Collect and preprocess your water quality data, ensuring it is clean, consistent, and organized. You may need to clean outliers and missing values.

* **Choose Visualization Tools:**

- Select appropriate visualization tools based on the nature of your data and the insights you want to convey. Common tools include:

- **Box Plots**: Show the distribution of data, including median, quartiles, and potential outliers. Useful for understanding parameter ranges.

- **Histograms**: Visualize the frequency distribution of a single variable. Ideal for showing the distribution of continuous variables like pH levels or temperature.

- **Scatter Plots**: Display the relationship between two variables, making it suitable for identifying correlations between different water quality parameters.

- **Heatmaps**: Depict correlations between multiple variables by using color intensity to represent strength and direction. Helpful for identifying complex patterns.

- **Line Charts**: Show trends and variations in data over time. Useful for visualizing seasonal changes in water quality.

-**GIS Mapping**: If location data is available, create geographic maps to show spatial variations in water quality parameters.

-**Waterfall Charts**: Useful for visualizing changes in water quality over time or along a water treatment process.

-**Parallel Coordinate Plots:** Ideal for visualizing multidimensional data and identifying relationships between parameters.

* **Parameter Distributions**:

- Create histograms, box plots, or density plots to visualize the distributions of individual water quality parameters. Use appropriate bin sizes and labels for clarity.

* **Correlation Analysis:**

- Generate scatter plots or heatmaps to explore correlations between pairs of parameters. Calculate correlation coefficients (e.g., Pearson, Spearman) to quantify relationships.

* **Potability Assessment:**

- To assess water potability, you can use visualizations like stacked bar charts, pie charts, or donut charts to show the proportion of potable and non-potable water samples.

- Overlay histograms or box plots for key parameters with a distinction between potable and non-potable water to identify potential differences.

* **Documentation:**

- Document your visualization choices, data sources, and methodology to enable others to replicate and understand your analysis.

1. **Predictive Modelling:**

Decide on the machine learning algorithms and features to use for predicting water potability. Here's a guide to help you decide on the machine learning algorithms and features for this task:

* **Data Preparation:**

Before choosing algorithms and features, ensure your dataset is properly prepared:

- **Data Cleaning**: Handle missing values, outliers, and any data quality issues.

- **Data Split**: Divide the dataset into training, validation, and test sets to evaluate model performance.

* **Define the Target Variable:**

- In this case, the target variable is "water potability," which is binary (potable or non-potable). Ensure your dataset includes this variable.

* **Feature Selection and Engineering:**

- Conduct exploratory data analysis to identify the most relevant features that may impact water potability. Features to consider include:

- pH

- Hardness

- Solids (e.g., TDS or TSS)

- Chloramines

- Sulphate

- Conductivity

- Organic carbon

- Turbidity

- Trihalomethanes

- ...and any other relevant water quality parameters.

- Consider transforming or scaling features if necessary (e.g., normalization or standardization).

* **Feature Importance Analysis:**

- Use feature importance techniques such as tree-based models (e.g., Random Forest) or feature selection algorithms (e.g., Recursive Feature Elimination) to identify the most informative features for predicting water potability.

* **Machine Learning Algorithms:**

- Choose machine learning algorithms suitable for binary classification tasks like water potability prediction. Some options to consider include:

- **Logistic Regression**: A simple yet effective algorithm for binary classification.

- **Random Forest**: A powerful ensemble method that handles feature importance and non-linearity well.

- **Gradient Boosting**: Algorithms like XGBoost or LightGBM can perform well in classification tasks and handle imbalanced datasets.

- **Support Vector Machines (SVM):** Useful for binary classification tasks when there's a clear separation between classes.

- **Neural Networks**: Deep learning models can capture complex relationships but may require a larger dataset and more computational resources.

- Try multiple algorithms and evaluate their performance using appropriate metrics (e.g., accuracy, precision, recall, F1-score, ROC-AUC) on the validation set.

* **Imbalanced Data Handling:**

- If your dataset is imbalanced (i.e., one class significantly outnumbers the other), consider techniques like oversampling, under sampling, or using weighted loss functions to address this issue.

* **Hyperparameter Tuning:**

- Optimize hyperparameters for the selected algorithm(s) using techniques like grid search or random search.

* **Model Evaluation:**

- Assess model performance on the test dataset to ensure it generalizes well to unseen data.

- Consider using appropriate evaluation metrics, confusion matrices, and ROC curves to understand model behaviour.

* **Interpretability:**

- Depending on the context, consider using interpretable machine learning models or techniques to explain the predictions, especially if the model is intended for regulatory or public health purposes.

* **Model Deployment:**

- If the model performs satisfactorily, deploy it in a production environment to make real-time predictions about water potability.

* **Continuous Monitoring:**

- Regularly monitor and update the model as new data becomes available to ensure its accuracy and relevance.

**Dataset Link:**[**https://www.kaggle.com/datasets/adityakadiwal/water-potability**](https://www.kaggle.com/datasets/adityakadiwal/water-potability)**Y**

**TEAM MEMBERS**:

AAZAM KHAN 2021504002

SASMITHA A 2021504038

SWETHA K 2021504051

GOKUL R 2021504515