## WATER QUALITY ANALYSIS

## TEAM MEMBER: S.ANANTHI

## (812121106003)

After many years of research, water quality standards are put in place to ensure the suitability of efficient use of water for a designated purpose. Water quality analysis is to measure the required parameters of water, following standard methods, to check whether they are in accordance with the standard.

## SYSTEM DESIGN

|  |  |
| --- | --- |
| **SELECTION OF PARAMETERS**  **SELECTION OF PARAMETERS**  **SELECTION OF METHODS**  **PRECISION AND ACCURACY OF METHOD SELECTED AS PER REQUIREMENT**  **PROPER SAMPLING**  **PROPER LABELING**   |  | | --- | | CHAIN-OF-CUSTODY PROCEDURES |   **PRESERVATION**    **ANALYSIS**  **REPORTING** Figure -1: Steps for Water Quality Analysis |

## MODULES

## Selection of Parameters

The parameters of water quality are selected entirely according to the need for a specific use of that water. Some examples are:

**Drinking:** As per WHO/CPCB Standards

#### Irrigation:

pH Conductivity

Sodium & Potassium Nutrients

Specific compounds

**Industries:** As per specific requirement **Domestic Consumption:** As per BIS Standards **Water Bodies:** As per CPCB guidelines

1. **Selection of Methods**

The methods of water quality analysis are selected according to the requirement. The factors playing key role for the selection of methods are:

* + 1. Volume and number of sample to be analyzed
    2. Cost of analysis
    3. Precision required
    4. Promptness of the analysis as required

## Precision and Accuracy of Method Selected as per Requirement

What precision and accuracy to be maintained against a particular method is decided according to the objective of the monitoring. The factors influencing this decision includes:

* Budget of Monitoring System
* Parameters to be Monitored
* Use of the Water

## Chain–of–Custody Procedures

Properly designed and executed chain-of-custody forms will ensure sample integrity from collection to data reporting. This includes the ability to trace possession and handling of the sample from the time of collection through analysis and final disposition. This process is referred to as “chain-of- custody” and is required to demonstrate sample control

when the data are to be used for regulation or litigation. Where litigation is not involved, chain-of-custody procedures are useful for routine control of samples.

A sample is considered to be under a person’s custody if it is in the individual’s physical possession, in the individual’s sight, secured and tamper-proofed by that individual, or secured in an area restricted to authorized personnel. The following procedures summarize the major aspects of chain- of-custody:

1. **Sample Labels:** Labels are used to prevent sample misidentification as well as to identify the collector, if required. In other words, labeling ensures the responsibility and accountability of the collector.
2. **Sample Seals:** Sample seals are used to detect unauthorized tampering with samples up to the time of analysis. So, it is essential to seal a sample before leaving the custody of the collector. Sealing must be done in such a way as one have to break the seal to access the sample.
3. **Field Log Book:** All the useful information related to a field survey or sampling should be recorded in a Log Book. At least the following data should be in the log book:
   1. Purpose of sampling
   2. Location of sampling point
   3. Name and address of field contact
   4. Producer of material being sampled and address, if different from location
   5. Type of sample
   6. Method, date, and time of preservation.
4. **Sample Analysis Request Sheet:** The sample analysis request sheet accompanies samples to the laboratory. The collector completes the field portion of such a form that includes most of the pertinent information noted in the log book. The laboratory portion of such a form is to be completed by laboratory personnel and includes: name of

person receiving the sample, laboratory sample number, date of sample receipt, condition of each sample (i.e., if it is cold or warm, whether the container is full or not, color, if more than one phase is present, etc.) and determinations to be performed.

1. **Sample Delivery to the Laboratory:** Sample(s) should be delivered to laboratory as soon as possible after collection, typically within 2 days. Where shorter sample holding times are required, special arrangements must be made to insure timely delivery to the laboratory. Where samples are shipped by a commercial carrier, the waybill number to be included in the sample custody documentation. Samples must be accompanied by a complete chain-of-custody record and a sample analysis request sheet.
2. **Receipt and Logging of Sample:** In the laboratory, the sample custodian inspects the condition and seal of the sample and reconciles label information and seal against the chain-of-custody record before the sample is accepted for analysis. After acceptance, the custodian assigns a laboratory number, logs sample in the laboratory log book and/or computerized laboratory information management system, and stores it in a secured storage room or cabinet or refrigerator at the specified temperature until it is assigned to an analyst.
3. **Assignment of Sample for Analysis:** The laboratory supervisor usually assigns the sample for analysis. Once the sample is in the laboratory, the supervisor or analyst is responsible for its care and custody.
4. **Disposal:** Samples are held for the prescribed amount and duration for the project or until the data have been reviewed and accepted. Samples are disposed usually after documentation. However, disposal must be in accordance with approved methods.

## Proper Sampling

Proper sampling is a vital condition for correct measurement of water quality parameters. Even if advanced techniques and sophisticated tools are used, the parameters can give an incorrect image of the actual scenario due to improper sampling. The proper sampling should fulfill the following criteria:

1. **Representative:** The data must represent the wastewater or water body being sampled. So, the following factors must be well planned for proper sampling:
   1. Process of Sampling
   2. Sampling size/volume
   3. Number of Sampling Locations
   4. Number of Samples
   5. Type of Samples
   6. Time Intervals

During sampling, these factors must also be taken care of:

* + - Choosing of proper sampling container
    - Avoiding contamination
    - Ensure the personal safety of the collector

1. **Reproducible:** The data obtained must be reproducible by others following the same sampling and analytical protocols.
2. **Defensible:** Documentation must be available to validate the sampling procedures. The data must have a known degree of accuracy and precision.
3. **Useful:** The data can be used to meet the objectives of the monitoring plan.

## Proper Labeling

Proper labeling prevents sample misidentification and ensures the responsibility and accountability of the collector. The sample container should be labeled properly, preferably by attaching an appropriately inscribed tag or label. Alternatively, the bottle can be labeled directly with a water- proof marker. Barcode labels are also available nowadays.

Information on the sample container or the tag should include at least:

* + 1. Sample code number (identifying location)
    2. Date and time of sampling
    3. Source and type of sample
    4. Pre-treatment or preservation carried out on the sample
    5. Any special notes for the analyst
    6. Sampler’s name

## Preservation

Usually a delay occurs between the collection and analysis of a sample. The characteristics of the sample can be changed

during this period. Therefore proper preservation is required in the way to laboratory after collection, and in the laboratory upto when analysis starts.

Complete and unequivocal preservation of samples, whether domestic wastewater, industrial wastes, or natural waters, is a practical impossibility because complete stability for every constituent never can be achieved. At best, preservation techniques only retard chemical (especially, hydrolysis of constituents) and biological changes that inevitably continue after sample collection.

No single method of preservation is entirely satisfactory; the preservative is chosen with due regard to the determinations to be made. Preservation methods are limited to pH control, chemical addition, the use of amber and opaque bottles, refrigeration, filtration, and freezing.

## Analysis

The samples, after reaching laboratory, are analyzed, according to the requisite parameters, following standard methods and protocols.

## Reporting

The ultimate procedure of water analysis is to prepare a proper report against the submitted requisition. The report must be authenticated before handing over the authority. All data should be kept in the laboratory log and preferably in laboratory database.

An alternative way to present the overall quality of water is to express it in the form of Water Quality Index (WQI). WQI is a concise numerical representation of overall water quality of a water body, which is convenient to interpret and used widely. WQI expresses the overall quality of water with a single digit, instead of many digits for all the WQP. Thus, it is readily conceivable for common people.

After many years of research, water quality standards are put in place to ensure the suitability of efficient use of water for a designated purpose. Water quality analysis is to measure the required parameters of water, following standard methods, to check whether they are in accordance with the standard.

## SYSTEM DESIGN

|  |  |
| --- | --- |
| **SELECTION OF PARAMETERS**  **SELECTION OF PARAMETERS**  **SELECTION OF METHODS**  **PRECISION AND ACCURACY OF METHOD SELECTED AS PER REQUIREMENT**  **PROPER SAMPLING**  **PROPER LABELING**   |  | | --- | | CHAIN-OF-CUSTODY PROCEDURES |   **PRESERVATION**    **ANALYSIS**  **REPORTING** Figure -1: Steps for Water Quality Analysis |

## MODULES

## Selection of Parameters

The parameters of water quality are selected entirely according to the need for a specific use of that water. Some examples are:

**Drinking:** As per WHO/CPCB Standards

#### Irrigation:

pH Conductivity

Sodium & Potassium Nutrients

Specific compounds

**Industries:** As per specific requirement **Domestic Consumption:** As per BIS Standards **Water Bodies:** As per CPCB guidelines

1. **Selection of Methods**

The methods of water quality analysis are selected according to the requirement. The factors playing key role for the selection of methods are:

* + 1. Volume and number of sample to be analyzed
    2. Cost of analysis
    3. Precision required
    4. Promptness of the analysis as required

## Precision and Accuracy of Method Selected as per Requirement

What precision and accuracy to be maintained against a particular method is decided according to the objective of the monitoring. The factors influencing this decision includes:

* Budget of Monitoring System
* Parameters to be Monitored
* Use of the Water

## Chain–of–Custody Procedures

Properly designed and executed chain-of-custody forms will ensure sample integrity from collection to data reporting. This includes the ability to trace possession and handling of the sample from the time of collection through analysis and final disposition. This process is referred to as “chain-of- custody” and is required to demonstrate sample control

when the data are to be used for regulation or litigation. Where litigation is not involved, chain-of-custody procedures are useful for routine control of samples.

A sample is considered to be under a person’s custody if it is in the individual’s physical possession, in the individual’s sight, secured and tamper-proofed by that individual, or secured in an area restricted to authorized personnel. The following procedures summarize the major aspects of chain- of-custody:

1. **Sample Labels:** Labels are used to prevent sample misidentification as well as to identify the collector, if required. In other words, labeling ensures the responsibility and accountability of the collector.
2. **Sample Seals:** Sample seals are used to detect unauthorized tampering with samples up to the time of analysis. So, it is essential to seal a sample before leaving the custody of the collector. Sealing must be done in such a way as one have to break the seal to access the sample.
3. **Field Log Book:** All the useful information related to a field survey or sampling should be recorded in a Log Book. At least the following data should be in the log book:
   1. Purpose of sampling
   2. Location of sampling point
   3. Name and address of field contact
   4. Producer of material being sampled and address, if different from location
   5. Type of sample
   6. Method, date, and time of preservation.
4. **Sample Analysis Request Sheet:** The sample analysis request sheet accompanies samples to the laboratory. The collector completes the field portion of such a form that includes most of the pertinent information noted in the log book. The laboratory portion of such a form is to be completed by laboratory personnel and includes: name of

person receiving the sample, laboratory sample number, date of sample receipt, condition of each sample (i.e., if it is cold or warm, whether the container is full or not, color, if more than one phase is present, etc.) and determinations to be performed.

1. **Sample Delivery to the Laboratory:** Sample(s) should be delivered to laboratory as soon as possible after collection, typically within 2 days. Where shorter sample holding times are required, special arrangements must be made to insure timely delivery to the laboratory. Where samples are shipped by a commercial carrier, the waybill number to be included in the sample custody documentation. Samples must be accompanied by a complete chain-of-custody record and a sample analysis request sheet.
2. **Receipt and Logging of Sample:** In the laboratory, the sample custodian inspects the condition and seal of the sample and reconciles label information and seal against the chain-of-custody record before the sample is accepted for analysis. After acceptance, the custodian assigns a laboratory number, logs sample in the laboratory log book and/or computerized laboratory information management system, and stores it in a secured storage room or cabinet or refrigerator at the specified temperature until it is assigned to an analyst.
3. **Assignment of Sample for Analysis:** The laboratory supervisor usually assigns the sample for analysis. Once the sample is in the laboratory, the supervisor or analyst is responsible for its care and custody.
4. **Disposal:** Samples are held for the prescribed amount and duration for the project or until the data have been reviewed and accepted. Samples are disposed usually after documentation. However, disposal must be in accordance with approved methods.

## Proper Sampling

Proper sampling is a vital condition for correct measurement of water quality parameters. Even if advanced techniques and sophisticated tools are used, the parameters can give an incorrect image of the actual scenario due to improper sampling. The proper sampling should fulfill the following criteria:

1. **Representative:** The data must represent the wastewater or water body being sampled. So, the following factors must be well planned for proper sampling:
   1. Process of Sampling
   2. Sampling size/volume
   3. Number of Sampling Locations
   4. Number of Samples
   5. Type of Samples
   6. Time Intervals

During sampling, these factors must also be taken care of:

* + - Choosing of proper sampling container
    - Avoiding contamination
    - Ensure the personal safety of the collector

1. **Reproducible:** The data obtained must be reproducible by others following the same sampling and analytical protocols.
2. **Defensible:** Documentation must be available to validate the sampling procedures. The data must have a known degree of accuracy and precision.
3. **Useful:** The data can be used to meet the objectives of the monitoring plan.

## Proper Labeling

Proper labeling prevents sample misidentification and ensures the responsibility and accountability of the collector. The sample container should be labeled properly, preferably by attaching an appropriately inscribed tag or label. Alternatively, the bottle can be labeled directly with a water- proof marker. Barcode labels are also available nowadays.

Information on the sample container or the tag should include at least:

* + 1. Sample code number (identifying location)
    2. Date and time of sampling
    3. Source and type of sample
    4. Pre-treatment or preservation carried out on the sample
    5. Any special notes for the analyst
    6. Sampler’s name

## Preservation

Usually a delay occurs between the collection and analysis of a sample. The characteristics of the sample can be changed

during this period. Therefore proper preservation is required in the way to laboratory after collection, and in the laboratory upto when analysis starts.

Complete and unequivocal preservation of samples, whether domestic wastewater, industrial wastes, or natural waters, is a practical impossibility because complete stability for every constituent never can be achieved. At best, preservation techniques only retard chemical (especially, hydrolysis of constituents) and biological changes that inevitably continue after sample collection.

No single method of preservation is entirely satisfactory; the preservative is chosen with due regard to the determinations to be made. Preservation methods are limited to pH control, chemical addition, the use of amber and opaque bottles, refrigeration, filtration, and freezing.

## Analysis

The samples, after reaching laboratory, are analyzed, according to the requisite parameters, following standard methods and protocols.

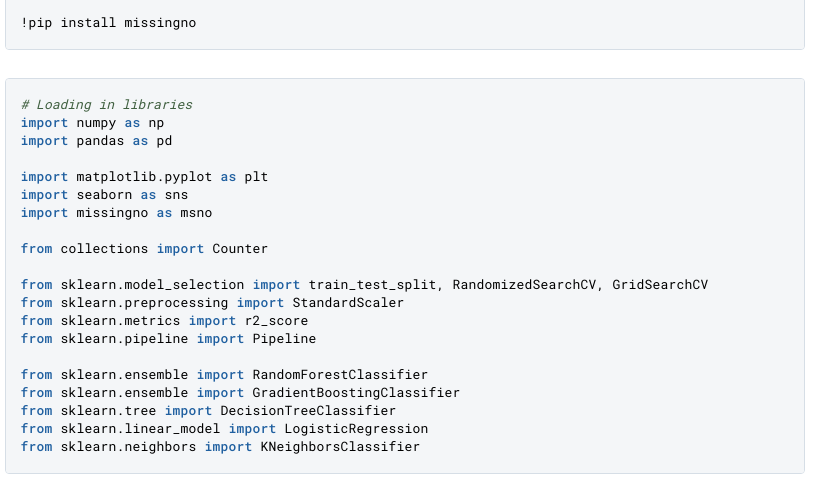
## Reporting

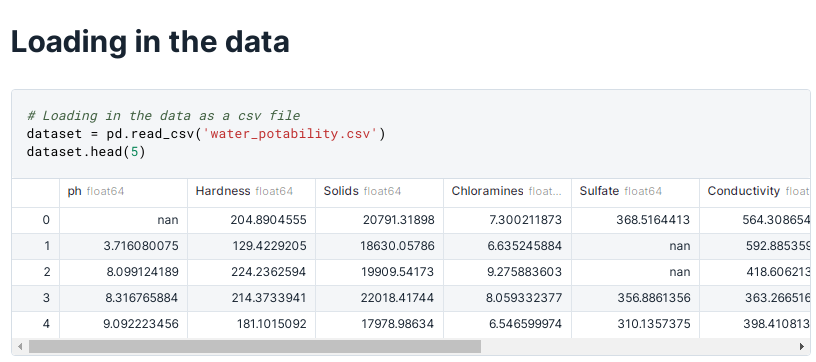
The ultimate procedure of water analysis is to prepare a proper report against the submitted requisition. The report must be authenticated before handing over the authority. All data should be kept in the laboratory log and preferably in laboratory database.

An alternative way to present the overall quality of water is to express it in the form of Water Quality Index (WQI). WQI is a concise numerical representation of overall water quality of a water body, which is convenient to interpret and used widely. WQI expresses the overall quality of water with a single digit, instead of many digits for all the WQP. Thus, it is readily conceivable for common people.

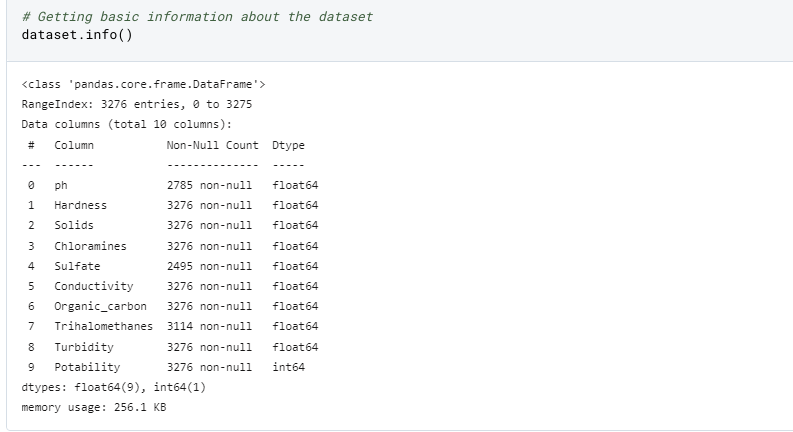
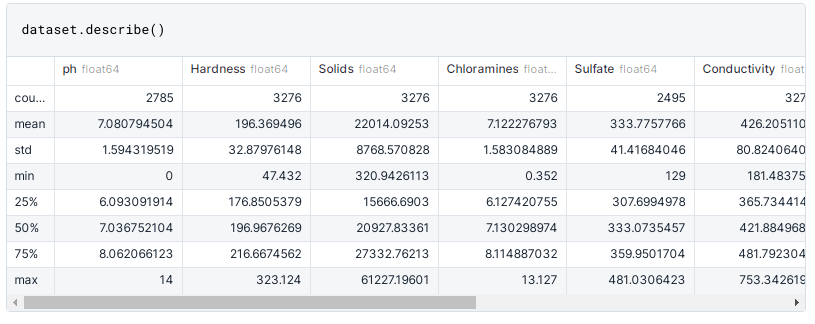
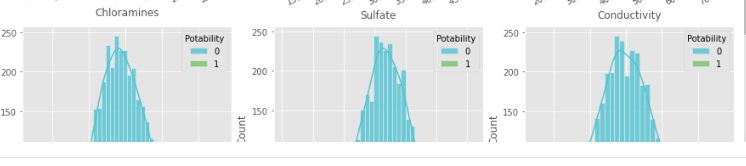
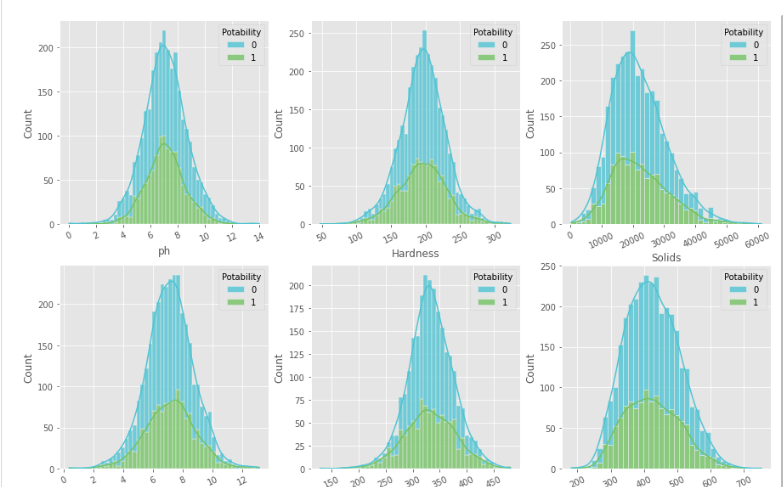
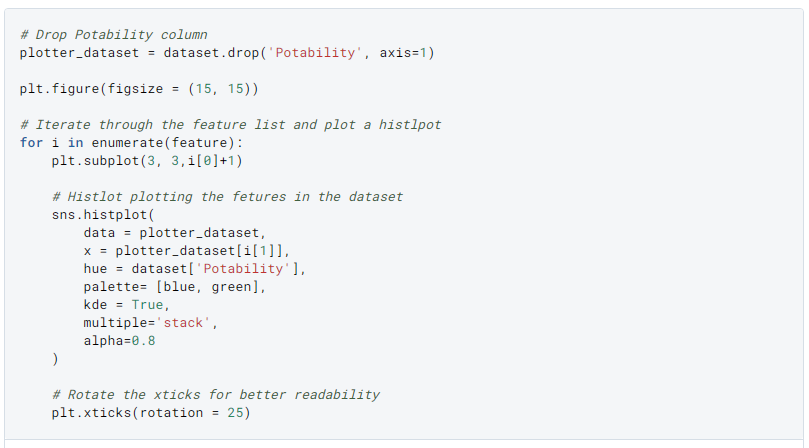
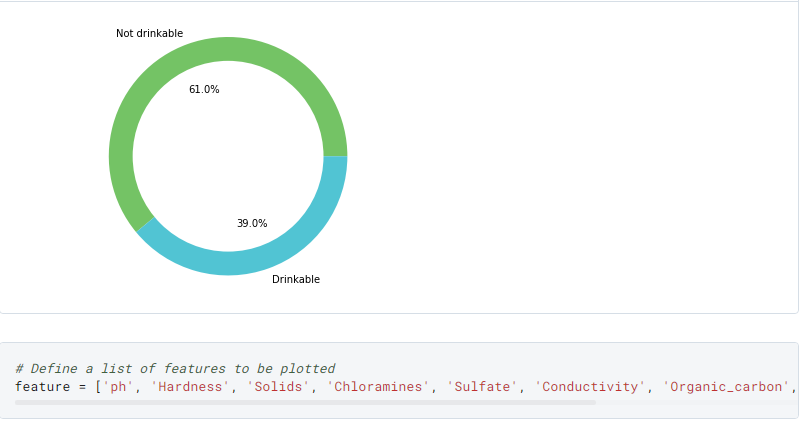
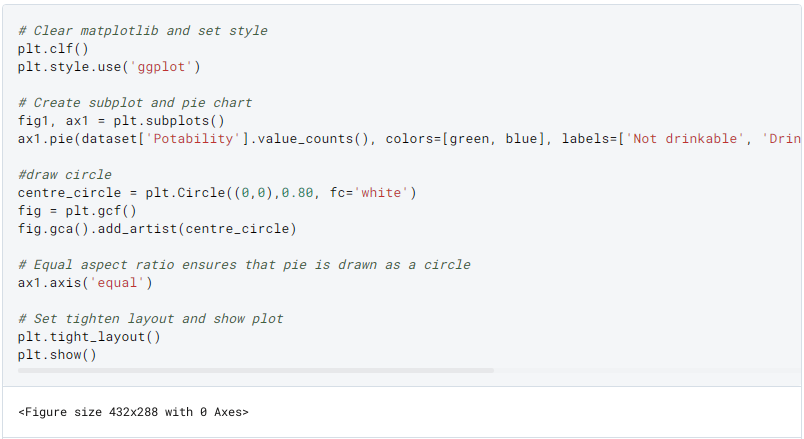
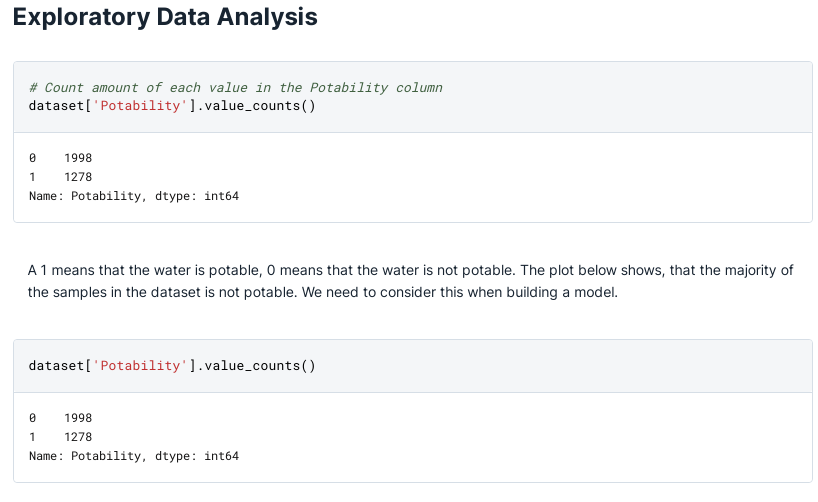
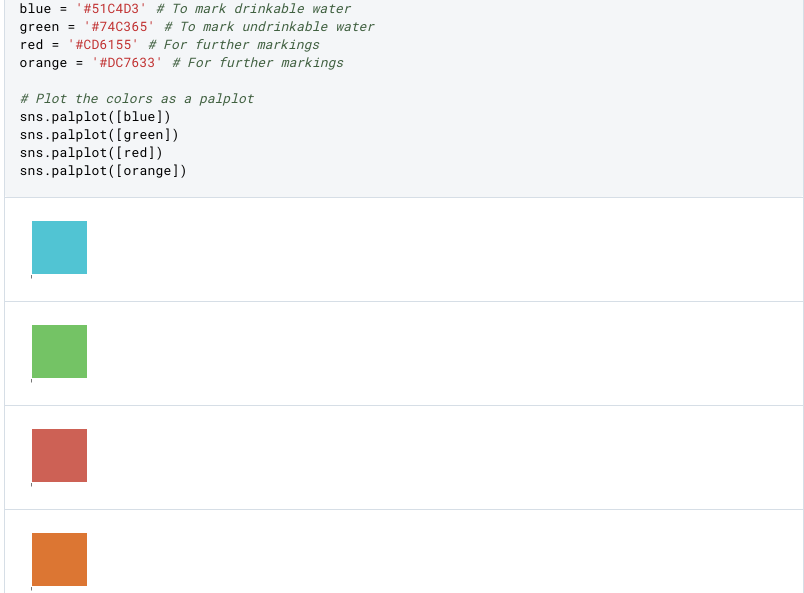
Water is the base for life as we know it. Every human needs water to survive. In many countries however, people do not have access to clean drinking water. The water quality dataset provides data on more than 3000 water samples. The data contains information about chemical components and wether or not the water istacutally drinkable.

This notebook will determine what makes clean, drinkable water. To accomplish this, supervised machine learning models are used.

The dataset contains 9 columns in total.



* **ph**: The ph value of the water, which describes the acid-base balance of the water. A ph-value of 6.5 - 8.5 is recommended by the who.
* **Sardness**: The more calcium and magnesium the water contains, the harder the water is. Though these minerals are not harmful to consume, they could have an impact on the potability of the water.
* **Solids (Total dissolved solids - TDS)**: Measurement of how many organic and inorgnaic materials are contained in the water. Desirable limit for TDS is 500 mg/l, maximum limit is 1000 mg/l.
* **Chloramines**: Chlorine and chloramine are a common disinfectant. Chlorine of 4 mg/L or 4 parts per million (ppm) are considered save.
* **Sulfate**: Naturally occuring mineral, that is much more higher in seawater than in feshwater.
* **Conductivity**: Measurement of how conductivce the water is, meaning how well energy flows through it. According to WHO standards, the electric conductivity (EC) should not be higher than 400 μS/cm.
* **Organic carbon**: Total Organic Carbon (TOC) is the result of decaying organic matter in water. According to US EPA < 2 mg/L of TOC is considered drinkable water.
* **Trihalomethanes**: A chemical that occurs in water treated with chlorine. Levels up to 80 ppm are considered safe.
* **Turbidity**: Depends on the amount of solid matter in the water. The WHO recommends a value of 5.00 NTU.
* **Potability**: States wether water is safe. 0 = not safe to drink, 1 = safe to drink.

The colors below are going to be used throughout the entire notebook. Blue will show dinkable, or potable, water. The muddy green color will be used in water that is not save for consumption.The WHO recommends ph-values of drinking water to be within the range of 6.5 and 9.5. My homecountry Germany also mandates that tap water should have a ph value between 6.5 and 9.5. With that in mind, let's see how many of our samples are within that range.

In a guideline for drinking water published by the WHO, the ph-value of water alone is not sufficient when it come to the potability of water. However, the ph-value is very important to ensure water disinfection and clarification. Therefore, pipes made out of copper or certain steels should not be used to transports water with very low ph-values, because the water is too acidic and would dissolve some of the metal in the pipe, making the water unsave due to the high amounts of metal in the water. The ph-value could therefore be correlated to other quality factors of the water.

## As the EU considers everything below 250 mg/L of Sulfate as save, the majority of the samples would not be considered Potable. The WHO however would see all of the samples as just fine, because they set the limit at 500 mg/L.

**DATA PREPAREATION: Dealing with missing data**

The ph, Sulfate and Trihalomethanescolumnes contain a lot of missing values. Dropping all of these values would mean that we lose a considerable amount of data. Instead, we will look into replacing missing values with the mean of median.

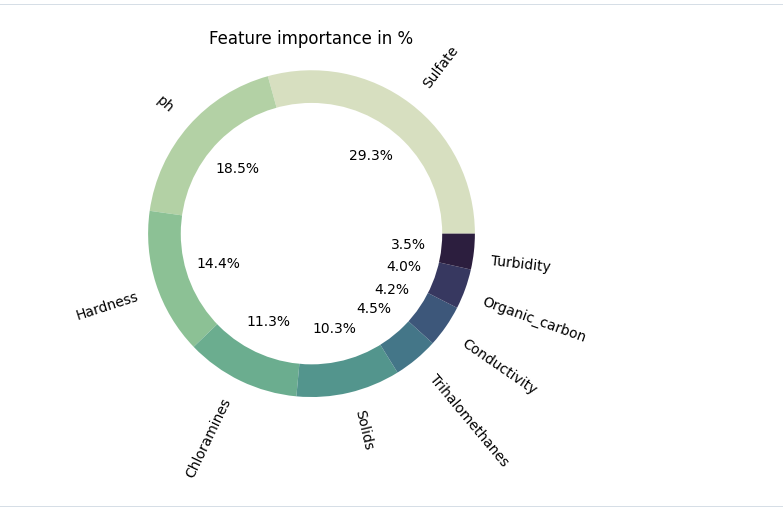
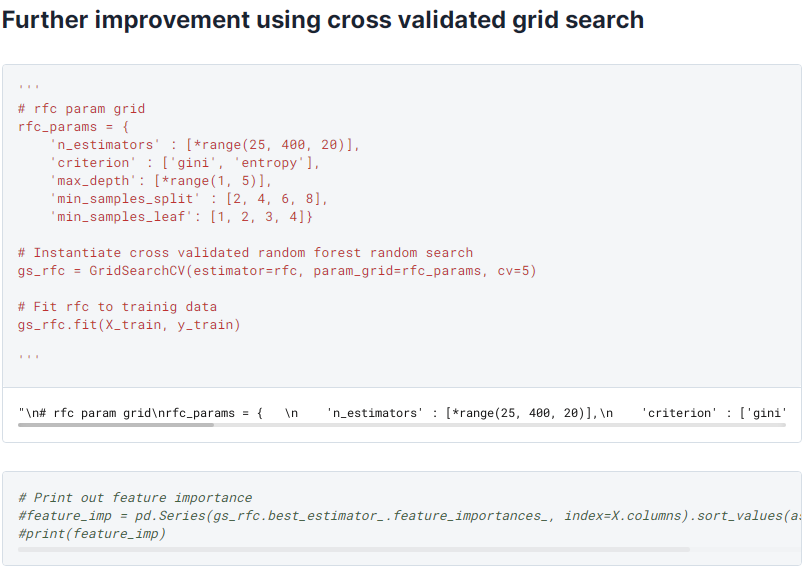
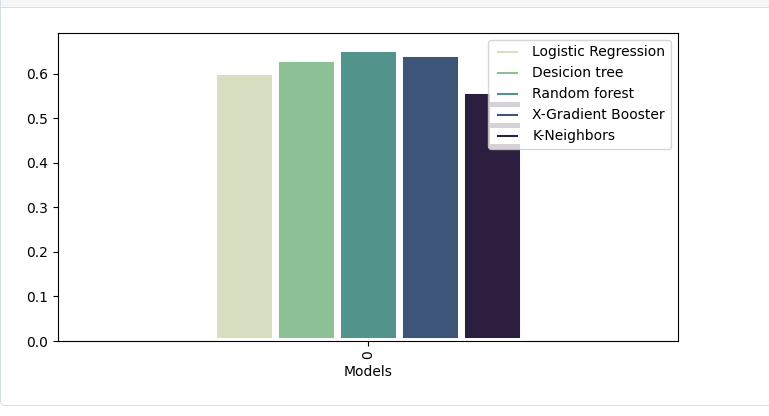
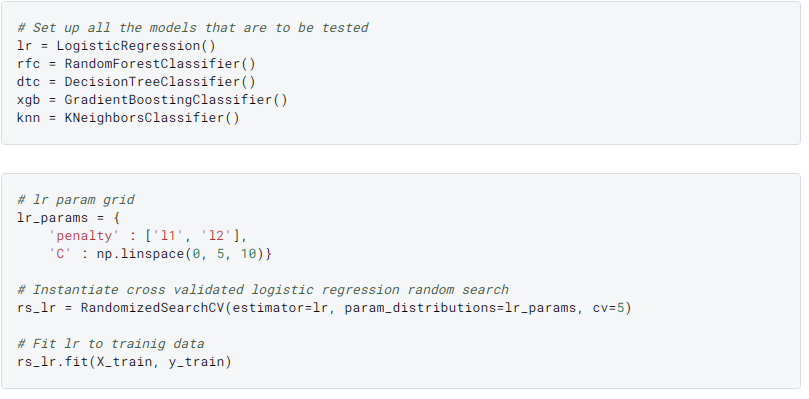
### Median and mean are very similar. Missing values will be replaces by the median value.Checking for correlations

## Building a model

**Randomized, cross validated search**

Time to build a machine learning model. To determine what makes, potable water, five different machine learning algortihms are going to be used. To get (relatively) good results, a cross validated random search is used on every model.

The random search randomly seelcts parameters from a predefined grid and to find the best results within these parameters. Computing the results of all the possible paramters would be computationally expensive. Doing a random search saves both time and computational power. After all models have gone through a randomized search, the best model will be selected. To get the absolute best results from that model, a grid search is performed, where all possible cominations of parmateres are tested to find out the best parameters.



➢ Analysing water quality is one of the key topics of machine learning research.

➢ In order to train a machine learning model that can determine if a certain water sample is safe or unsafe for eating, we must first understand all the parameters that impact water potability.

➢ This process is also known as water potability analysis.

➢ We'll be utilising a Kaggle dataset that includes information on all of the key elements that have an impact on the potability of water for the water quality analysis challenge.

➢ Before building a model using machine learning to predict whether the water specimen is acceptable or unsafe for eating.

➢ we must first quickly examine each characteristic of this dataset because all of the elements that determine water quality are crucial.

**Overview of the process:**

**1. \*\*Sample Collection\*\*:**

- Select sampling locations that are representative of the water source or system being studied.

- Use clean, non-contaminated containers and equipment for sample collection.

- Collect samples at different depths if applicable (e.g., for lakes and reservoirs).

**2. \*\*Sample Preservation\*\*:**

- Preserve the samples as necessary to maintain their integrity until analysis. Common preservation methods include:

- Refrigeration: For most samples, storing at 4°C is sufficient.

- Chemical preservatives: Adding specific reagents to prevent microbial growth or chemical changes.

**3. \*\*Physical Characteristics Analysis\*\*:**

- Measure physical parameters, including temperature, turbidity, and conductivity.

- Conduct a visual inspection for color, odor, and sediment.

**4. \*\*Chemical Characteristics Analysis\*\*:**

- Determine the chemical composition of the water. Common chemical parameters include:

- pH (acidity or alkalinity)

- Dissolved oxygen (DO)

- Chemical oxygen demand (COD)

- Biochemical oxygen demand (BOD)

- Nutrient concentrations (nitrate, phosphate, etc.)

- Metals (e.g., iron, lead, copper)

- Organic compounds (e.g., pesticides, volatile organic compounds)

**5. \*\*Biological Characteristics Analysis\*\*:**

- Assess the presence of microorganisms, such as coliform bacteria or fecal coliforms, as indicators of microbial contamination.

- Evaluate the diversity and abundance of aquatic organisms in the water, including algae, plankton, and macroinvertebrates.

**6. \*\*Toxicological Analysis\*\*:**

- Determine the presence of toxic substances or pollutants using specialized tests for specific contaminants, such as heavy metals, pesticides, or organic pollutants.

**7. \*\*Data Interpretation\*\*:**

- Compare the obtained data with established water quality standards, guidelines, or regulatory limits to assess the water's suitability for its intended use.

- Identify potential water quality issues or areas of concern.

**8. \*\*Reporting and Communication\*\*:**

- Compile the results into a comprehensive report.

- Communicate the findings to relevant authorities, stakeholders, and the public as necessary.

**9. \*\*Quality Control\*\*:**

- Implement quality control measures to ensure the accuracy and reliability of the data, including the use of certified reference materials and duplicate samples.

**10. \*\*Ongoing Monitoring\*\*:**

- Continuously monitor water quality over time to detect trends and assess the effectiveness of any remediation efforts.

**11. \*\*Remediation and Management\*\*:**

- Implement measures to improve water quality if issues are identified, such as treating water or addressing pollution sources.

Water quality analysis is an ongoing process that helps protect human health, ecosystems, and water resources. The specific parameters and methods used in analysis may vary depending on the purpose, regulatory requirements, and the characteristics of the water source being studied.

**Feature engineering for water quality analysis:**

**1. \*\*Temporal Features\*\*:**

- Time-based features can help capture seasonality and trends. Examples include:

- Day of the week, month, or year.

- Time of day.

- Moving averages or rolling statistics to smooth out data.

**2. \*\*Lagged Features\*\*:**

- Create lagged features by shifting measurements in time. This can capture delayed effects and autocorrelation. For instance, you might use data from the past few days or weeks to predict the current water quality.

**3. \*\*Statistical Aggregates\*\***:

- Calculate statistical aggregates such as mean, median, standard deviation, and percentiles for water quality parameters over different time intervals (e.g., hourly, daily, monthly). These can provide a summary of the data distribution.

**4. \*\*Interactions\*\*:**

- Create interaction features by combining two or more water quality parameters. For instance, you might calculate the ratio of nitrate to phosphate concentrations to assess nutrient balance.

**5. \*\*Derivative Features\*\*:**

- Compute the rate of change for specific water quality parameters over time. Derivatives can be useful in understanding trends and identifying sudden changes or anomalies.

**6. \*\*Geospatial Features\*\*:**

- If you have data from different locations, consider including geospatial features, such as latitude and longitude, distance to pollution sources, or proximity to specific geographical features.

**7. \*\*Weather Data Integration\*\*:**

- Incorporate weather-related features (e.g., precipitation, temperature, humidity) as they can influence water quality. Combining water quality data with weather data can help in understanding how environmental factors impact water parameters.

**8. \*\*Seasonal Decomposition\*\*:**

- Decompose the time series data into trend, seasonality, and residual components. These components can be used as separate features for analysis.

**9. \*\*Time Since Last Event\*\*:**

- Calculate the time since significant water quality events or occurrences (e.g., heavy rainfall, pollution incidents) to assess the persistence of impacts.

**10. \*\*Historical Averages and Trends\*\*:**

- Compute historical averages and trends for each water quality parameter, providing context for the current measurement.

**11. \*\*Frequency Domain Analysis\*\*:**

- Use techniques such as Fourier transforms to extract frequency domain features that may reveal periodic patterns in the data.

**12. \*\*Principal Component Analysis (PCA)\*\*:**

- Reduce the dimensionality of the data by using PCA to create new features that capture the most significant variations in water quality parameters.

**13. \*\*Cross-Correlation\*\*:**

- Calculate cross-correlation between different water quality parameters to identify relationships and dependencies between them.

**14. \*\*Categorical Encoding\*\***:

- Encode categorical variables like water source type, treatment methods, or regulatory zones into numerical features using techniques like one-hot encoding.

**15. \*\*Target Transformation\*\*:**

- Apply transformations to the target variable, such as log transformation, to make it more amenable to modeling and account for skewed distributions.

**16. \*\*Feature Scaling and Normalization\*\*:**

- Normalize or scale features to ensure they are on a common scale, especially when using machine learning algorithms sensitive to feature scales.

The choice of feature engineering techniques depends on the specific characteristics of your water quality dataset and the goals of your analysis. Experiment with different features and assess their impact on model performance to identify the most informative features for your water quality analysis.

Model evaluation:

**1. \*\*Data Splitting\*\*:**

- Split the dataset into training, validation, and test sets to train the model, tune hyperparameters, and evaluate its performance on unseen data.

**2. \*\*Cross-Validation\*\*:**

- Use techniques like k-fold cross-validation to assess the model's robustness and generalization ability by repeatedly splitting the data into training and validation subsets.

**3. \*\*Performance Metrics\*\*:**

a. \*\*Regression Metrics\*\*:

- Mean Absolute Error (MAE): Measures the average absolute difference between predicted and actual values.

- Root Mean Square Error (RMSE): Provides a measure of the model's prediction error by taking the square root of the mean of squared differences.

- R-squared (R2) or Coefficient of Determination: Indicates the proportion of variance explained by the model.

b. \*\*Classification Metrics\*\*:

- Accuracy: Measures the proportion of correctly classified instances.

- Precision: Calculates the ratio of true positive predictions to the total positive predictions.

- Recall (Sensitivity or True Positive Rate): Measures the ability of the model to identify positive instances.

- F1-Score: The harmonic mean of precision and recall, balancing precision and recall.

- Area Under the Receiver Operating Characteristic Curve (AUC-ROC): Evaluates the model's ability to distinguish between classes in binary classification problems.

**4. \*\*Confusion Matrix\*\*:**

- For classification tasks, create a confusion matrix to visualize the true positives, true negatives, false positives, and false negatives. This helps in understanding model performance, especially when dealing with imbalanced datasets.

**5. \*\*Residual Analysis\*\*:**

- For regression tasks, analyze the model residuals (the differences between predicted and actual values) to check for patterns or biases. A well-behaved residual plot should show random scattering around zero.

**6. \*\*Feature Importance\*\*:**

- Assess the importance of each feature in the model using techniques like feature importance scores or permutation importance to understand which variables are most influential in water quality predictions.

7. \*\*Model Selection\*\*:

- Compare the performance of different models (e.g., linear regression, decision trees, random forests, neural networks) to identify the most suitable model for the specific water quality analysis task.

8. \*\*Overfitting and Underfitting Analysis\*\*:

- Check for signs of overfitting (the model is too complex and fits noise) or underfitting (the model is too simple to capture the underlying patterns) by examining training and validation performance.

**9. \*\*Bias and Fairness Evaluation\*\*:**

- Assess the model for potential biases or unfairness, especially when water quality analysis has implications for environmental justice and social equity.

**10. \*\*External Validation\*\*:**

- If available, compare model predictions with independent data sources or conduct field tests to validate the model's performance in real-world conditions.

**11. \*\*Error Analysis\*\*:**

- Analyze the specific types of errors the model makes and investigate potential causes. This can guide improvements in data quality and model design.

**12. \*\*Model Interpretability\*\*:**

- Ensure that the model is interpretable, especially if the results need to be communicated to stakeholders. Techniques like SHAP values and feature importance can help in understanding model predictions.

**13. \*\*Cost-Benefit Analysis\*\*:**

- Consider the costs associated with model errors and the benefits of accurate predictions. This can help in selecting an appropriate model threshold or decision boundary.

The choice of evaluation techniques and metrics should be tailored to the specific objectives of the water quality analysis and the type of models being used. Regular model evaluation is essential to ensure that the model remainsaccurate and reliable over time, especially in dynamic environmental conditions.

**Visualization:**

Model evaluation is a crucial step in water quality analysis to assess the performance of predictive models and determine their reliability in estimating or classifying water quality parameters. Below are some common techniques and metrics for evaluating models in the context of water quality analysis:

1. \*\*Data Splitting\*\*:

- Split the dataset into training, validation, and test sets to train the model, tune hyperparameters, and evaluate its performance on unseen data.

2. \*\*Cross-Validation\*\*:

- Use techniques like k-fold cross-validation to assess the model's robustness and generalization ability by repeatedly splitting the data into training and validation subsets.

3. \*\*Performance Metrics\*\*:

a. \*\*Regression Metrics\*\*:

- Mean Absolute Error (MAE): Measures the average absolute difference between predicted and actual values.

- Root Mean Square Error (RMSE): Provides a measure of the model's prediction error by taking the square root of the mean of squared differences.

- R-squared (R2) or Coefficient of Determination: Indicates the proportion of variance explained by the model.

b. \*\*Classification Metrics\*\*:

- Accuracy: Measures the proportion of correctly classified instances.

- Precision: Calculates the ratio of true positive predictions to the total positive predictions.

- Recall (Sensitivity or True Positive Rate): Measures the ability of the model to identify positive instances.

- F1-Score: The harmonic mean of precision and recall, balancing precision and recall.

- Area Under the Receiver Operating Characteristic Curve (AUC-ROC): Evaluates the model's ability to distinguish between classes in binary classification problems.

4. \*\*Confusion Matrix\*\*:

- For classification tasks, create a confusion matrix to visualize the true positives, true negatives, false positives, and false negatives. This helps in understanding model performance, especially when dealing with imbalanced datasets.

5. \*\*Residual Analysis\*\*:

- For regression tasks, analyze the model residuals (the differences between predicted and actual values) to check for patterns or biases. A well-behaved residual plot should show random scattering around zero.

6. \*\*Feature Importance\*\*:

- Assess the importance of each feature in the model using techniques like feature importance scores or permutation importance to understand which variables are most influential in water quality predictions.

7. \*\*Model Selection\*\*:

- Compare the performance of different models (e.g., linear regression, decision trees, random forests, neural networks) to identify the most suitable model for the specific water quality analysis task.

8. \*\*Overfitting and Underfitting Analysis\*\*:

- Check for signs of overfitting (the model is too complex and fits noise) or underfitting (the model is too simple to capture the underlying patterns) by examining training and validation performance.

9. \*\*Bias and Fairness Evaluation\*\*:

- Assess the model for potential biases or unfairness, especially when water quality analysis has implications for environmental justice and social equity.

10. \*\*External Validation\*\*:

- If available, compare model predictions with independent data sources or conduct field tests to validate the model's performance in real-world conditions.

11. \*\*Error Analysis\*\*:

- Analyze the specific types of errors the model makes and investigate potential causes. This can guide improvements in data quality and model design.

12. \*\*Model Interpretability\*\*:

- Ensure that the model is interpretable, especially if the results need to be communicated to stakeholders. Techniques like SHAP values and feature importance can help in understanding model predictions.

13. \*\*Cost-Benefit Analysis\*\*:

- Consider the costs associated with model errors and the benefits of accurate predictions. This can help in selecting an appropriate model threshold or decision boundary.

PROGRAM:

import plotly.graph\_objs as go

index\_vals =

data['Potability'].astype('category').cat.codes

fig = go.Figure(data=go.Splom(

dimensions=[dict(label='ph',

values=data['ph']),

dict(label='Hardness',

values=data['Hardness']),

dict(label='Solids',

values=data['Solids']),

dict(label='Chloramines',

values=data['Chloramines'])

dict(label='Sulfate',

values=data['Sulfate']),

dict(label='Conductivity',

values=data['Conductivity']),

dict(label='Organic\_carbon',

values=data['Organic\_carbon']),

dict(label='Trihalomethanes',

values=data['Trihalomethanes']),

dict(label='Turbidity',values=data['Turbidity'])],

showupperhalf=False,

text=data['Potability'],

marker=dict(color=index\_vals,

showscale=False,

line\_color='white', line\_width=0.5)

))

fig.update\_layout(

title='Water Quality',

width=1000,

height=1000,

)

fig.show().

CONCLUSION:

In conclusion, water quality analysis is a vital component of environmental monitoring, public health, and the sustainable management of water resources. It encompasses a comprehensive process of assessing the physical, chemical, and biological characteristics of water to determine its suitability for various purposes.