KINGSTON ENGINEERING COLLEGE

**COLLEGE CODE: 5113**

***DATA ANALYTICS WITH COGNOS***

# Project No.10- water quality analysis

**BATCH MEMBERS:**

1. INIYAN.P -au511321104032

2.KAMALESH.L-au511321104039

3.MUKESH.C-au511321104057

4.KUGAN.K-au511321104044

5.DHANUSH.M -au511321104021

## PHASE2:

## INNOVATION

## INTRODUCTION:

Water quality analysis is a critical aspect of environmental science and public health that seeks to assess and monitor the physical, chemical, and biological characteristics of water to ensure its suitability for various purposes. The quality of water has a profound impact on human health, ecosystems, and industrial processes, making it essential to understand and manage. This multifaceted field involves the collection, testing, and interpretation of data to determine the presence and concentration of various substances in water, such as pollutants, nutrients, minerals, and microorganisms.

Water quality analysis serves a variety of purposes, including the assessment of drinking water safety, the monitoring of aquatic ecosystems, the evaluation of wastewater treatment efficacy, and the identification of potential health hazards. It also plays a critical role in ensuring compliance with environmental regulations and standards.

This introduction sets the stage for a comprehensive exploration of the methods, significance, and applications of water quality analysis, highlighting its pivotal role in safeguarding our natural resources and protecting public health. In this context, it becomes evident that a thorough understanding of water quality is crucial for sustainable environmental management and responsible resource utilization.

### ABOUT THE DATA:

Where did we get the dataset?

### Kaggle:

The dataset provided on Kaggle, https://[www.kaggle.com/datasets/adityakadiwal/water-potability,](http://www.kaggle.com/datasets/adityakadiwal/water-potability) offers a valuable resource for our project aimed at forecasting water quality analysis

### Dataset Details:

The water\_potability.csv file contains water quality metrics for 3276 different water bodies.

* PH is an important parameter in evaluating the acid–base balance of water. It is also the indicator of acidic or alkaline condition of water status. WHO has recommended maximum permissible limit of pH from 6.5 to 8.5. The current investigation ranges were 6.52–

6.83 which are in the range of WHO standards.

* Hardness is mainly caused by calcium and magnesium salts. These salts are dissolved from geologic deposits through which water travels. The length of time water is in contact with hardness producing material helps determine how much hardness there is in raw water. Hardness was originally defined as the capacity of water to precipitate soap caused by Calcium and Magnesium.
* Water has the ability to dissolve a wide range of inorganic and some organic minerals or salts such as potassium, calcium, sodium, bicarbonates, chlorides, magnesium, sulfates etc. These minerals produced un-wanted taste and diluted color in appearance of water. This is the important parameter for the use of water. The water with high TDS value indicates that water is highly mineralized. Desirable limit for TDS is 500 mg/l and maximum limit is 1000 mg/l which prescribed for drinking purpose.



* Chlorine and chloramine are the major disinfectants used in public water systems. Chloramines are most commonly formed when ammonia is added to chlorine to treat drinking water. Chlorine

levels up to 4 milligrams per liter (mg/L or 4 parts per million (ppm)) are considered safe in drinking water.

## COLUMNS THAT WE USED:

⦁ The water\_potability.csv file contains water quality metrics for 3276 different water bodies.

### pH value:

PH is an important parameter in evaluating the acid–base balance of water. It is also the indicator of acidic or alkaline condition of water status. WHO has recommended maximum permissible limit of pH from

6.5 to 8.5. The current investigation ranges were 6.52–6.83 which are in the range of WHO standards.

### Hardness:

Hardness is mainly caused by calcium and magnesium salts. These salts are dissolved from geologic deposits through which water travels. The length of time water is in contact with hardness producing material helps determine how much hardness there is in raw water. Hardness was originally defined as the capacity of water to precipitate soap caused by Calcium and Magnesium.

### Solids (Total dissolved solids - TDS):

Water has the ability to dissolve a wide range of inorganic and some organic minerals or salts such as potassium, calcium, sodium, bicarbonates, chlorides, magnesium, sulfates etc. These minerals

produced un-wanted taste and diluted color in appearance of water. This is the important parameter for the use of water. The water with high TDS value indicates that water is highly mineralized. Desirable limit for TDS is 500 mg/l and maximum limit is 1000 mg/l which prescribed for drinking purpose.

### Chloramines:

Chlorine and chloramine are the major disinfectants used in public water systems. Chloramines are most commonly formed when ammonia is added to chlorine to treat drinking water. Chlorine levels up to 4 milligrams per liter (mg/L or 4 parts per million (ppm)) are considered safe in drinking water.

### Sulfate:

Sulfates are naturally occurring substances that are found in minerals, soil, and rocks. They are present in ambient air, groundwater, plants, and food. The principal commercial use of sulfate is in the chemical industry. Sulfate concentration in seawater is about 2,700 milligrams per liter (mg/L). It ranges from 3 to 30 mg/L in most freshwater supplies, although much higher concentrations (1000 mg/L) are found in some geographic locations.

### Conductivity:

Pure water is not a good conductor of electric current rather’s a good insulator. Increase in ions concentration enhances the electrical conductivity of water. Generally, the amount of dissolved solids in water determines the electrical conductivity. Electrical conductivity (EC) actually measures the ionic process of a solution that enables it to

transmit current. According to WHO standards, EC value should not

exceeded 400 μS/cm.

### Organic\_carbon:

Total Organic Carbon (TOC) in source waters comes from decaying natural organic matter (NOM) as well as synthetic sources. TOC is a measure of the total amount of carbon in organic compounds in pure water. According to US EPA < 2 mg/L as TOC in treated / drinking water, and < 4 mg/Lit in source water which is use for treatment.

### Trihalomethanes:

THMs are chemicals which may be found in water treated with chlorine. The concentration of THMs in drinking water varies according to the level of organic material in the water, the amount of chlorine required to treat the water, and the temperature of the water that is being treated. THM levels up to 80 ppm is considered safe in drinking water.

### Turbidity:

The turbidity of water depends on the quantity of solid matter present in the suspended state. It is a measure of light emitting properties of water and the test is used to indicate the quality of waste discharge with respect to colloidal matter. The mean turbidity value obtained for Wondo Genet Campus (0.98 NTU) is lower than the WHO recommended value of 5.00 NTU.

## DETAILS OF LIBRARIES USED AND WAY TO DOWNLOAD:

In our water analysis forecasting project, we leverage a range of powerful libraries to facilitate data manipulation, model development, and evaluation. These libraries play a crucial role in transforming our design into an innovative solution. Here are the key libraries and a detailed explanation of how to download and install them:

### Pandas:

⦁ Download: Pandas can be installed using Python's package manager, pip. Open a command prompt or terminal and run pip install pandas.

⦁ Installation: After downloading, Pandas can be imported in your Python script using import pandas as pd.

### NumPy:

⦁ Download: NumPy can also be installed using pip. Run pip install numpy in your command prompt or terminal.

⦁ Installation: Import NumPy in your Python script using import numpy as np.

### Matplotlib and Seaborn:

⦁ Download: These visualization libraries can be installed with pip. Use pip install matplotlib seaborn in your command prompt or terminal.

⦁ Installation: Import Matplotlib and Seaborn in your Python script using import matplotlib.pyplot as plt and import seaborn as sns, respectively.

### Scikit-Learn:

⦁ Download: Scikit-Learn can be installed with pip. Run pip install scikit-learn in your command prompt or terminal.

⦁ Installation: Import Scikit-Learn in your Python script using import sklearn.

### Keras with TensorFlow Backend:

⦁ Download: Install Keras with TensorFlow using pip install tensorflow keras in your command prompt or terminal.

⦁ Installation: Import Keras in your Python script using import keras.

### Jupyter Notebooks:

⦁ Download: You can install Jupyter Notebooks by running pip install jupyter in your command prompt or terminal.

⦁ Installation: After installation, start a Jupyter Notebook server by running jupyter notebook in your command prompt or terminal. This will open a web-based interface for creating and running notebooks.

Once these libraries are downloaded and installed, you can import them into your Python scripts to access their functionality. These libraries provide a robust ecosystem for data analysis, machine learning, and data visualization, essential for transforming your design into an innovative solution for water analysis forecasting.

## HOW TO TRAIN AND TEST:

Training and testing a water quality analysis dataset involves using machine learning or statistical methods to build a model that can predict or classify water quality parameters based on :

### Data Collection:

* Gather a comprehensive and representative dataset containing information on water quality parameters. This dataset should include features such as temperature, pH, turbidity, dissolved oxygen, chemical concentrations (e.g., nitrate, phosphate, heavy metals), and biological indicators (e.g., E. coli counts).
* Ensure the dataset is labeled, meaning it includes the target variable you want to predict or classify (e.g., water quality categories like 'clean,' 'polluted,' or specific parameter concentrations).

### Data Preprocessing:

* Clean the dataset by handling missing values and outliers. This may involve imputation or removal of problematic data points.
* Normalize or standardize numerical features to bring them to a common scale.
* Encode categorical variables if needed (e.g., water source type) into numerical values.

### Data Splitting:

- Divide the dataset into two subsets: a training set and a testing set. The typical split is around 70-80% for training and 20-30% for testing.

### Feature Selection/Engineering:

- Analyze the importance of features. You may choose to select the most relevant features or engineer new features if needed.

### Model Selection:

- Choose an appropriate machine learning algorithm based on your problem. Common algorithms for water quality analysis include decision trees, random forests, support vector machines, and neural networks.

### Model Training:

- Use the training dataset to train the chosen model. The model will learn patterns in the data to make predictions or classifications.

### Model Evaluation:

* Evaluate the model's performance using the testing dataset.

Common evaluation metrics for classification tasks include accuracy, precision, recall, F1 score, and for regression tasks, metrics like mean squared error (MSE) or R-squared.

* Perform cross-validation if you have a limited dataset to ensure the model's generalization.

### Tuning Hyperparameters:

- Optimize the model's hyperparameters through techniques like grid search or random search to improve its performance.

### Interpret Results:

- Analyze the model's predictions to gain insights into water quality trends, identify influential factors, and understand the model's limitations.

### Deployment:

- Once you are satisfied with the model's performance, you can deploy it to make predictions or classifications on new, unseen data. This might involve integrating the model into a water quality monitoring system.

### Continuous Monitoring and Maintenance:

- Regularly update the model as more data becomes available or as water quality standards change.

Remember that the choice of modeling technique and the quality of your data are critical factors in the success of water quality analysis. Furthermore, interpretability and domain knowledge are crucial for understanding the implications of the model's predictions in the context of water quality management.

## METRICS USED FOR ACCURACY CHECK:

In water quality analysis, the choice of evaluation metrics depends on the specific task you are trying to accomplish. Here are some common metrics used to check the accuracy of water quality analysis models:

1. **Accuracy:** Accuracy is a general metric used for classification tasks. It measures the proportion of correctly classified samples out of the total samples in the dataset. While it provides a simple and easy-to- understand measure of overall model performance, it may not be the best choice if the dataset is imbalanced (e.g., if there are many more "clean" water samples than "polluted" ones), as it can be misleading in such cases.
2. **Precision:** Precision is the ratio of true positive predictions to the total number of positive predictions (true positives + false positives). It is particularly important when you want to minimize false positives, such as in cases where classifying water as polluted when it's not could lead to unnecessary actions or costs.
3. **Recall (Sensitivity):** Recall measures the ratio of true positive predictions to the total number of actual positive samples (true positives + false negatives). It is crucial when you want to avoid false negatives, such as when failing to detect actual pollution in water could have severe consequences.
4. **F1 Score:** The F1 score is the harmonic mean of precision and recall, and it provides a balance between these two metrics. It's particularly useful when you want to strike a balance between minimizing false positives and false negatives.
5. **Specificity:** Specificity is the ratio of true negative predictions to the total number of actual negative samples (true negatives + false positives). It is essential when you want to minimize false positives, similar to precision, but for the negative class.
6. **Mean Absolute Error (MAE):** MAE is a common metric for regression tasks in water quality analysis. It measures the average absolute difference between the predicted values and the actual values of the water quality parameter. It is easy to interpret, with lower values indicating better accuracy.
7. **Root Mean Squared Error (RMSE):**RMSE is another metric for regression tasks, and it measures the square root of the average of the squared differences between the predicted and actual values. RMSE gives more weight to larger errors and is sensitive to outliers.
8. **Coefficient of Determination (R-squared or R²):** R-squared quantifies the proportion of the variance in the water quality parameter that is explained by the model. An R-squared value close to 1 indicates a good fit, while a value close to 0 suggests the model does not explain much of the variance.
9. **Cohen's Kappa:** Cohen's Kappa is a metric that takes into account the agreement between the model's predictions and what would be expected by chance. It is often used when dealing with imbalanced datasets.

The choice of metrics should align with the specific goals of your water quality analysis project. It's important to consider the context, potential consequences of false positives and false negatives, and the nature of the dataset when selecting the most appropriate evaluation metrics.

### CONCLUSION:

⦁ In conclusion, water quality analysis plays a pivotal role in safeguarding our natural resources, protecting public health, and ensuring the sustainability of our environment. This multifaceted field, which encompasses the assessment and monitoring of physical, chemical, and biological characteristics of water, has far-reaching implications for various sectors, including public utilities, industry, agriculture, and conservation.

Through the systematic collection, testing, and interpretation of data, water quality analysis empowers us to make informed decisions and take proactive measures to address water-related challenges. Here are some key takeaways:

Water quality analysis is a cornerstone of environmental protection efforts. It enables us to identify and mitigate pollution sources, manage ecosystems, and maintain the delicate balance of aquatic life.

Access to clean and safe drinking water is a fundamental human right. Water quality analysis ensures that water treatment processes meet rigorous standards, minimizing health risks associated with contaminants and pathogens.

# THANK YOU.