### WATER QUALITY ANALYSIS USING MACHINE LEARNING

**TEAM MEMBER**

**412621243025:JANANI.S**

# Phase-3 Document Submission

**Project:** *Water Quality Analysis*



### OBJECTIVE:

A "Water Quality Analysis Project" is a research or environmental monitoring initiative aimed at assessing and ensuring the safety and purity of water sources, such as rivers, lakes, reservoirs, groundwater, and drinking water supplies.

**Phase 3*: Development Part 1***

***1. Data Collection:***

* *Start by obtaining the water quality dataset. You may find such datasets on government websites, environmental organizations, or research institutions' data repositories.*

***2.Data Loading:***

* *Once you have the dataset, load it into your preferred data analysis environment. Popular tools for this purpose include Python with libraries like Pandas, or R.*

***3.Data Inspection:***

* *Begin by inspecting the dataset to understand its structure and the variables it contains. This typically includes checking the number of rows and columns and the data types of each column.*

***4. Handling Missing Values:***

*Identify and handle missing values. Missing data can significantly impact your analysis, so it's important to deal with them. Common strategies include:*

* *Removing rows with missing values.*
* *Imputing missing values with statistical measures (e.g., mean, median, mode).*
* *Using domain knowledge to estimate missing values.*

***5. Handling Outliers:***

*Identify and address outliers in the dataset. Outliers can skew your analysis and should be treated appropriately. Some methods for handling outliers include:*

* *Visualizing data using box plots and histograms to identify outliers.*
* *Applying techniques like the Z-score or IQR (Interquartile Range) to detect and deal with outliers.*

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

*EDA helps you gain insights into the dataset. You can perform the following steps as part of EDA:*

1. ***Univariate Analysis:***

* *Visualize the distribution of each parameter in the dataset. You can use , kernel density plots, or bar charts.*

1. ***Bivariate Analysis:***

* *Examine relationships between pairs of variables. You can use scatter plots to visualize correlations between water quality parameters.*

1. ***Multivariate Analysis:***

* *Consider how multiple variables interact with each other. This can be rough techniques like correlation matrices and heatmaps.*

***7. Visualization:***

*Create meaningful visualizations to help interpret the data. This could include:*

* *Scatter plots to visualize relationships between water quality parameters.*
* *Box plots to show the distribution of parameters by categories (e.g., location or time).*
* *Correlation matrices and heatmaps to display relationships between variables.*

***8. Identify Deviations from Standards:***

* *If you have established standards or guidelines for water quality, compare your data to these standards to identify any deviations. This can be done through summary statistics or visualizations.*

***9. Data Preprocessing for Analysis:***

* *Depending on your specific analysis goals, you may need to further preprocess the data. This could include feature scaling, encoding categorical variables, or creating new features.*

***10. Document and Report:***

* *Document your findings, insights, and the steps you've taken during data preprocessing and EDA. Visualizations and clear explanations will help communicate your results effectively.*

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

**PROGRAM:**

import pandas as pd

# Load the dataset from a CSV file

df = pd.read\_csv('water\_quality\_dataset.csv')

# Check for missing values in the dataset

df.isnull().sum()

import matplotlib.pyplot as plt

# Create a boxplot for each water quality parameter

for col in df.columns:

    plt.boxplot(df[col])

    plt.title(col)

    plt.show()

# Create histograms of all water quality parameters

df.hist(figsize=(10, 10))

plt.show()

# Create a correlation matrix to identify correlations between water quality parameters

corr\_matrix = df.corr()

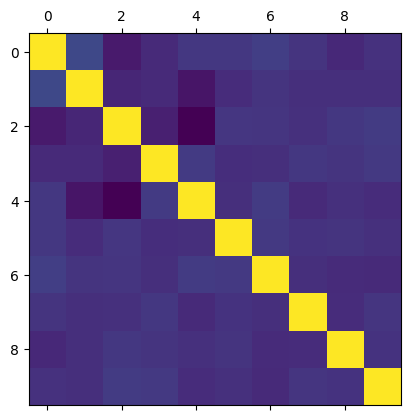
plt.matshow(corr\_matrix)

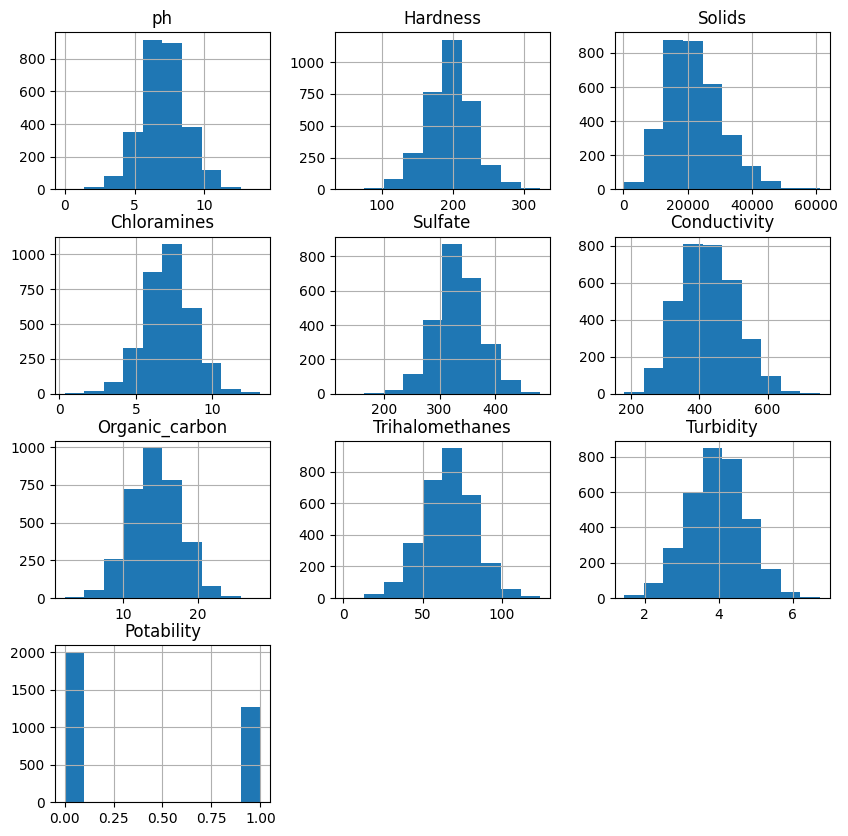
plt.show()

# Calculate the mean and standard deviation of each water quality parameter

df.describe()

***OUTPUT:***





ph 491

Hardness 0

Solids 0

Chloramines 0

Sulfate 781

Conductivity 0

Organic\_carbon 0

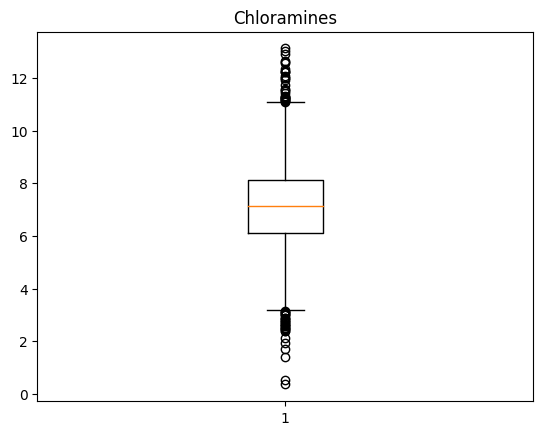
Trihalomethanes 162

Turbidity 0

Potability 0

dtype:int64



nt64

