Water Potability Using Machine Learning

# Objective

Design a model for predicting water potability.

# Dataset

Got the dataset for the model from Kaggle-- **Water Quality**(link: https://www.kaggle.com/datasets/sonialikhan/water-quality/data)

## Metadata

1. **pH value:**

pH evaluates the acid-base balance of the water. According to WHO pH should lie between 6.5 to 8.5. the current dataset ranges between 6.52 to 6.83.

1. **Hardness:**

Hardness in water is primarily caused by dissolved calcium and magnesium salts, which come from geological deposits. The longer water remains in contact with these materials, the higher the hardness level in the water.

1. **Solids:**

Water can dissolve various minerals like potassium, calcium, and magnesium, which can affect its taste and appearance. High Total Dissolved Solids (TDS) indicate mineral-rich water. For drinking, the desirable TDS limit is 500 mg/L, with a maximum of 1000 mg/L.

1. **Chloramines:**

Chlorine and chloramine are common disinfectants in public water systems. Chloramines form when ammonia is added to chlorine. Chlorine levels up to 4 mg/L (or 4 ppm) are safe for drinking water.

1. **Sulphate** (sulfate modified to Sulphate in Dataset)**:**

Sulphates are natural substances found in minerals, soil, and water. Seawater sulphate levels are about 2,700 mg/L, while freshwater usually has 3 to 30 mg/L, though some areas can have much higher concentrations.

1. **Conductivity:**

Pure water is a poor conductor of electricity, but adding dissolved solids (ions) increases its conductivity. Electrical conductivity (EC) measures how well water can transmit electricity, and according to WHO standards, it should not exceed 400 μS/cm.

1. **Organic\_carbon:**

Total Organic Carbon (TOC) in water, originating from natural decay and synthetic sources, measures organic carbon content. The US EPA recommends TOC levels under 2 mg/L in drinking water and 4 mg/L in source water for treatment.

1. **Trihalomethanes:**

THMs are chemicals in chlorinated water, with levels depending on organic content, chlorine used, and water temperature. Safe THM levels in drinking water are up to 80 ppm.

1. **Turbidity:**

Turbidity measures the cloudiness of water due to suspended solids. WHO recommended a limit of 5.00 NTU, indicating good water quality. WHO recommended a limit of 5.00 NTU, indicating good water quality.

1. **Potability:**

Indicates if water is safe for human consumption where 1 means Potable and 0 means Not potable.

1. **Y – variable:**

Potebility is the y variable.

# Data processing

Imported the dataset in RStudio.

Dataset contains a lot of missing values. The dataset has 3276 observations and ‘Na’ values are present in 1434 observations(“sum(is.na(data))”).   
Using “colSums(is.na(data))” I got missing values are present in pH(491), Sulphate(781), Trihalomethanes(162).

* **Deletion:**

(Deletion causes data loss.)

data.clean = na.omit(data)- deleted the rows that contain any missing values and stored cleaned data in data.clean.  
we are now left with 2011 observations(i.e: 1434 observations are deleted).

* **Imputation:**

1. Mean method:

Replacing ‘Na’ values with mean of data  
data$ph[which(is.na(data$ph))] = mean(data$ph, na.rm = TRUE)

data$Sulphate[which(is.na(data$Sulphate))] = mean(data$Sulphate, na.rm = TRUE)

data$Trihelomethain[which(is.na(data$Trihelomethain))] = mean(data$Trihelomethain, na.rm = TRUE)

1. Median method:

Replacing ‘Na’ values with median of data  
data$ph[which(is.na(data$ph))] = median(data$ph, na.rm = TRUE)

data$Sulphate[which(is.na(data$Sulphate))] = median(data$Sulphate, na.rm = TRUE)

data$Trihelomethain[which(is.na(data$Trihelomethain))] = median(data$Trihelomethain, na.rm = TRUE)

(I’ll work on the median imputated dataset as it is more robust against outliers and would retain the entire dataset)

* **Class Imbalance:**This is a classification problem statement, and it has two classes, ‘1’ and ‘0’. There are 1278 observations in class ‘1’ and 1998 observations in class ‘0’. So, the classes are partially balanced.
* **Normalization:**   
  The dataset contains a vast range so normalization is necessary.

Created a function of the above equation and applied it to all the columns excluding index and potability using the apply() function.

* **Data Split:**

The processed data is split into two parts train\_data & test\_data (90% and 10% respectively).  
It will help in further evaluating the data.