PROBLEM DEFINITION:

The water potability predictor problem is to develop a machine learning model that can predict the potability of water based on its water quality metrics. Access to safe drinking-water is essential to health, a basic human right and a component of effective policy for health protection. This is important as a health and development issue at a national, regional and local level. This model could be used to help identify and address water quality issues in different parts of the world, and to ensure that people have access to safe drinking water.

The input to the model would be a set of water quality metrics, such as pH, hardness, conductivity, and turbidity. The output of the model would be a prediction of whether the water is potable or not.

DATASET CONTENT:

1. 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.

2. 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.

3. 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.

4. 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.

5. 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.

6. 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.

7. 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.

8. 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.

9. 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.

10. Potability:

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

DESIGN THINKING:

1. Data Collection:

The dataset used in this project comes from Kaggle which is a high-fidelity database with five-star research rating in google scholar. There are nine training variables, such as pH, hardness, solids, chloramines etc… The dataset link is Water-Potability-Dataset.

2. Data Preprocessing:

Clean the data by addressing missing values and outliers. We can use techniques like imputation, removal, or interpolation. Normalize or scale data if necessary.

3. Exploratory Data Analysis (EDA):

Visualize the data using histograms, box plots, scatter plots, or heatmaps to understand the distribution and relationships between variables. Identify patterns and trends in water quality parameters.

4. Feature Engineering:

Create new features or derive relevant features from existing ones if needed

5. Machine Learning or Statistical Modelling:

- Split your dataset into training and testing sets for model evaluation.
- The model of machine learning can be divided into supervised learning, semi-supervised learning, and unsupervised learning. There are many kinds of algorithms for each learning mode, such as decision trees, naive Bayes, random forests, etc. Each algorithm model has its own advantages and disadvantages.
- Among the many machine learning methods, artificial neural networks and support vector machine algorithms became popular in machine learning due to their large processing data and fast calculation speed. Therefore, selecting the above two algorithms to judge the drinking ability of water resources is expected to better achieve the desired purpose.

6. Model Evaluation:

Evaluate the model's performance using relevant metrics such as accuracy, precision, recall, F1-score, and ROC AUC.

7. Model Tuning:

Adjust the hyperparameters of a model to improve its performance.

8. Interpretation and Insights:

Interpret the model results to understand which water quality parameters are most influential in determining potability and compare accuracy of each model.