Water quality analysis

Phase 1

Annai college of engineering and technology Kumbakonam

Submitted by:

Kalaimani.k

Prabhaharan.R

Gokul.R

Mohamed niyas.S

Daniel.B

Problem Definition and Design Thinking

Agenta:

Introduction:

Water quality analysis is a critical process that involves assessing the safety and purity of water for consumption and other purposes. By examining various factors like pH levels, mineral content, and potential contaminants, we can ensure that the water we rely on meets established standards for health and environmental safety. This analysis plays a vital role in safeguarding public health and the well-being of our communities.

Project definition:

The project involves analyzing water quality data to assess the suitability of water for specific purposes, such as drinking. The objective is to identify potential issues or deviations from regulatory standards and determine water potability based on various parameters. This project includes defining analysis objectives, collecting water quality data, designing relevant visualizations, and building a predictive model.

Design Thinking:

Analysis Objectives:

The primary objective of this water quality analysis is to rigorously evaluate the composition and characteristics of the provided water samples. This assessment aims to ascertain the suitability of the water for various purposes, including human consumption, agricultural use, and environmental impact. Specifically, we aim to:

1. Assess Potability: Determine if the water meets established safety standards for human consumption, ensuring it is free from harmful contaminants and within acceptable pH ranges.

2. Identify Deviations from Standards: Detect and quantify any deviations from regulatory guidelines or recommended thresholds for key parameters, such as pH levels, mineral content, and microbial load.

3. Understand Parameter Relationships: Analyze the interplay between different water quality parameters to gain insights into their potential synergies or correlations, which may impact the overall quality and usability of the water.

By achieving these objectives, we aim to provide a comprehensive evaluation of the water's quality, enabling informed decision-making for its appropriate use and ensuring the well-being of communities and the environment.

Data Collection:

About Dataset

Admittance to safe drinking-water is fundamental for wellbeing, an essential common freedom and a part of viable strategy for wellbeing insurance. This is significant as a wellbeing and improvement issue at a public, provincial and nearby level. In certain locales, it has been demonstrated the way that interests in water supply and sterilization can yield a net financial advantage, since the decreases in unfavorable wellbeing impacts and medical care costs offset the expenses of undertaking the mediations.

1. pH esteem:

PH is a significant boundary in assessing the corrosive base equilibrium of water. It is likewise the sign of acidic or antacid state of water status. WHO has suggested greatest reasonable restriction of pH from 6.5 to 8.5. The ongoing examination ranges were 6.52-6.83 which are in the scope of WHO principles.

2. Hardness:

Hardness is primarily brought about by calcium and magnesium salts. These salts are disintegrated from geologic stores through which water voyages. The period of time water is in touch with hardness creating material decides how much hardness there is in crude water. Hardness was initially characterized as the limit of water to encourage cleanser brought about by Calcium and Magnesium.

3. Solids (All out broke down solids - TDS):

Water can disintegrate many inorganic and a few natural minerals or salts, for example, potassium, calcium, sodium, bicarbonates, chlorides, magnesium, sulfates and so on. These minerals created un-needed taste and weakened variety in appearance of water. This is the significant boundary for the utilization of water. The water with high TDS esteem shows that water is profoundly mineralized. Helpful breaking point for TDS is 500 mg/l and most extreme cutoff is 1000 mg/l which endorsed for drinking reason.

4. Chloramines:

Chlorine and chloramine are the significant sanitizers utilized openly water frameworks. Chloramines are most generally shaped when alkali is added to chlorine to treat drinking water. Chlorine levels up to 4 milligrams for each liter (mg/L or 4 sections for every million (ppm)) are viewed as protected in drinking water.

5. Sulfate:

Sulfates are normally happening substances that are tracked down in minerals, soil, and shakes. They are available in encompassing air, groundwater, plants, and food. The key business utilization of sulfate is in the substance business. Sulfate fixation in seawater is around 2,700 milligrams for every liter (mg/L). It goes from 3 to 30 mg/L in most freshwater supplies, albeit a lot higher focuses (1000 mg/L) are tracked down in a few geographic areas.

6. Conductivity:

Unadulterated water is certainly not a decent conveyor of electric flow Rather's a decent encasing. Expansion in particles fixation upgrades the electrical conductivity of water. By and large, how much disintegrated solids in water decides the electrical conductivity. Electrical conductivity (EC) really gauges the ionic course of an answer that empowers it to send flow. As indicated by WHO norms, EC worth shouldn't surpassed 400 μS/cm.

7. Organic carbon:

Absolute Natural Carbon (TOC) in source waters comes from rotting normal natural matter (NOM) as well as engineered sources. TOC is a proportion of the aggregate sum of carbon in natural mixtures in unadulterated water. As per 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 synthetic substances which might be found in water treated with chlorine. The convergence of THMs in drinking water shifts as per the degree of natural material in the water, how much chlorine expected to treat the water, and the temperature of the water that is being dealt with. THM levels up to 80 ppm is viewed as protected in drinking water.

9. Turbidity:

The turbidity of water relies upon the amount of strong matter present in the suspended state. It is a proportion of light transmitting properties of water and the test is utilized to show the nature of waste release as for colloidal matter. The mean turbidity esteem acquired for Wondo Genet Grounds (0.98 NTU) is lower than the WHO suggested worth of 5.00 NTU.

10. Potability:

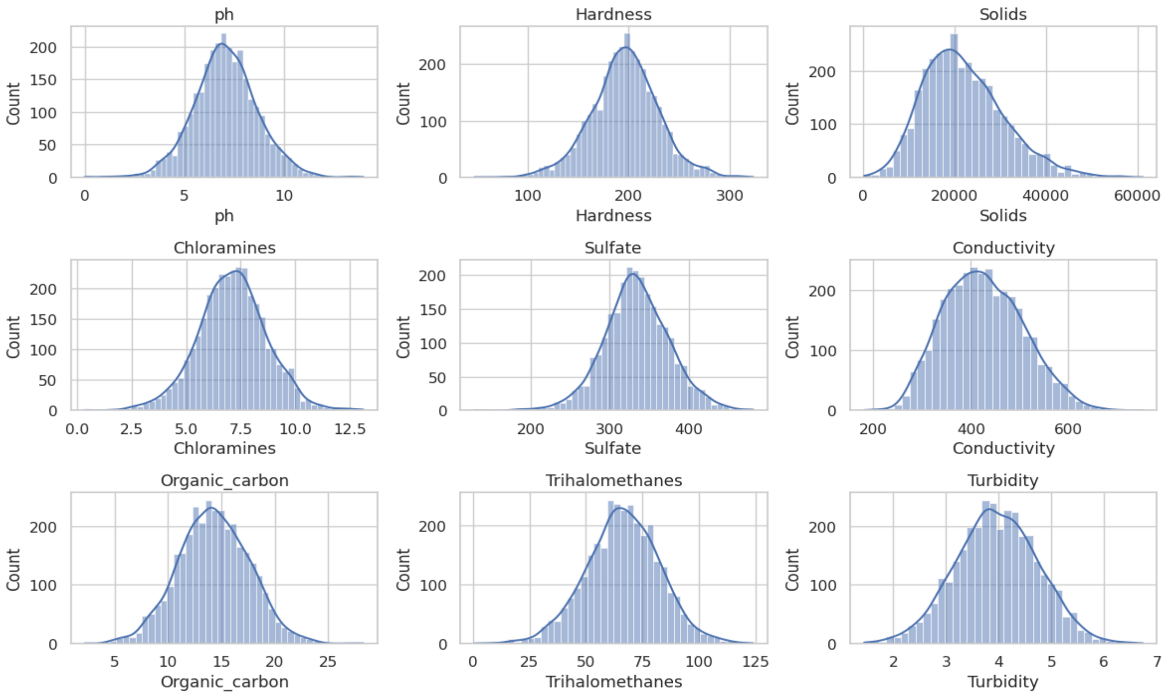
Shows in the event that water is alright for human utilization where 1 method Consumable and 0 methods Not consumable.

Sample dataset:

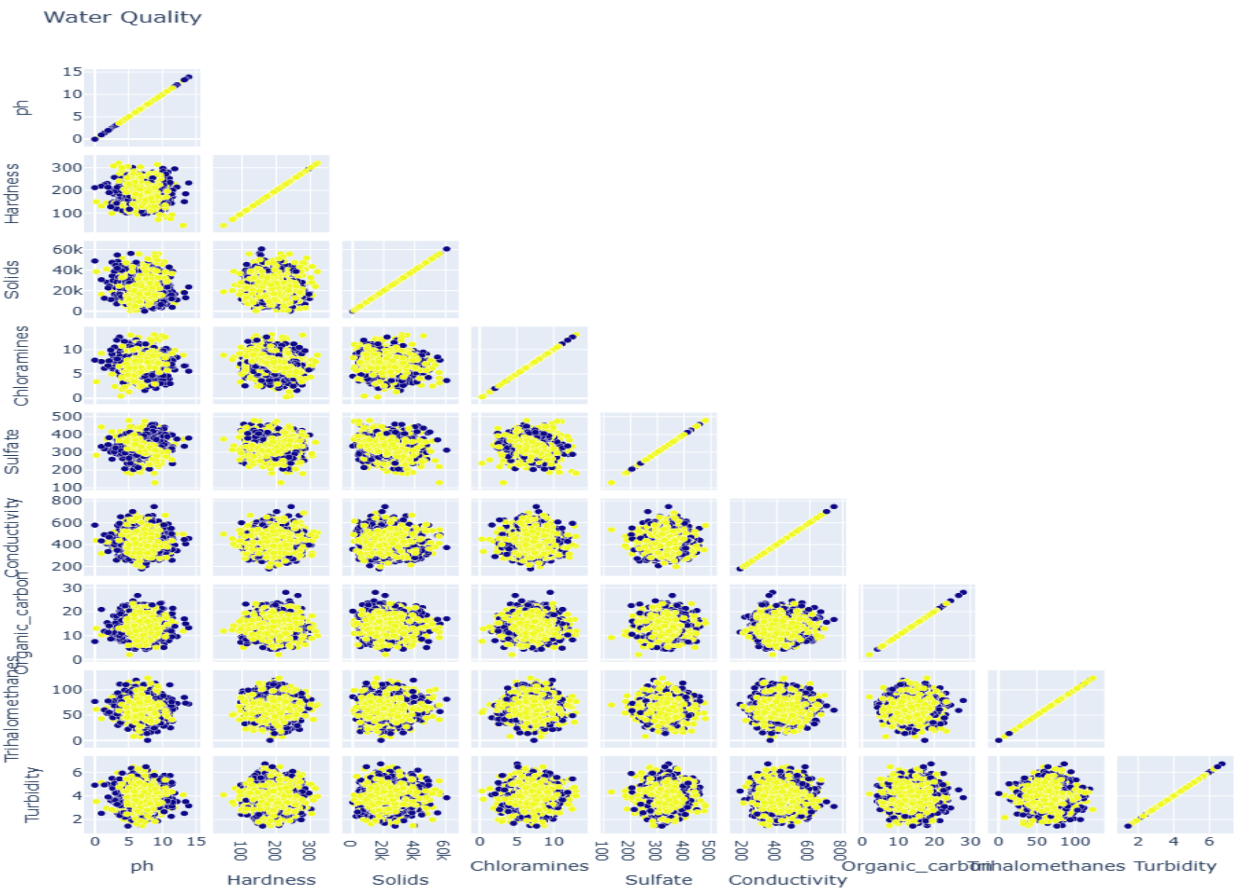
Our sample dataset : water_potability.csv

Visualization:

Univariate analysis



Bivariate analysis



Predictive modeling:

The best machine learning algorithm for water quality analysis depends on several factors, including the specific goals of the analysis, the nature of the data, and the available computational resources. Here are some commonly used algorithms for water quality analysis, along with their strengths:

1. Random Forests:

Strengths:

Can handle a large number of features without overfitting.

Robust against outliers and noisy data.

Provides feature importance scores.

Use Case:

Suitable for situations where there are a large number of features or complex interactions between features.

2. Gradient Boosting (e.g., XGBoost, LightGBM):

Strengths:

Highly accurate and powerful for complex relationships in the data.

Can handle missing data and outliers well.

Less likely to overfit compared to individual decision trees.

Use Case:

When high prediction accuracy is a top priority and there are complex relationships in the data.

3. Support Vector Machines (SVM):

Strengths:

Effective for separating classes in high-dimensional spaces.

Works well when there's a clear margin of separation between classes.

Use Case:

Suitable when there's a distinct separation between different water quality categories.

4. Neural Networks:

Strengths:

Extremely flexible and capable of learning complex relationships.

Can handle large amounts of data.

Use Case:

When there are very complex and non-linear relationships in the data, and you have a substantial amount of data for training.

5. Logistic Regression:

Strengths:

Simple and easy to interpret.

Works well when the relationship between features and target variable is roughly linear.

Use Case:

Suitable for cases where you want a straightforward understanding of how each feature influences water quality.

6. Decision Trees:

Strengths:

Easy to interpret and explain.

Can handle both numerical and categorical data.

Use Case:

Useful for gaining insights into which features are most influential in determining water quality.

7. K-Nearest Neighbors (KNN):

Strengths:

Simple and intuitive concept.

Can handle non-linear relationships.

Use Case:

When the local characteristics of data points are important in determining water quality.

8. Naive Bayes:

Strengths:

Simple and computationally efficient.

Can work well even with a small amount of data.

Use Case:

Suitable for situations where independence assumptions among features hold reasonably well.

SAMPLE SOURCE CODE:

```
# Import necessary libraries
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score

# Load your water quality dataset
# Assume 'X' contains your feature columns (e.g., pH, hardness, etc.) and 'y' contains the target variable (potability)
# Make sure your data is appropriately preprocessed and cleaned

# Split data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Initialize the Logistic Regression model
model = LogisticRegression()

# Train the model
model.fit(X_train, y_train)

# Predict potability on the test set
y_pred = model.predict(X_test)
```

Calculate accuracy

accuracy = accuracy_score(y_test, y_pred)

print(accuracy: {accuracy*100:.2f}%")

Explanation:

- **Import Libraries**: This code snippet starts by importing the necessary libraries, including pandas for data handling, scikit-learn for machine learning functionalities, and specific modules like Logistic Regression for the algorithm.
- **Load Data**: Assuming you have a dataset with features (X) and the target variable (y), load it into your Python environment. Make sure your data is appropriately preprocessed and cleaned before using it in a machine learning model.
- **Split Data**: The dataset is split into training and testing sets. This allows us to train the model on one portion of the data and evaluate its performance on another, unseen portion.
- **Initialize the Model**: We create an instance of the Logistic Regression model.
- **Train the Model**: The model is trained using the **fit()** method on the training data.
- **Predict**: We use the trained model to make predictions on the test set.
- **Calculate Accuracy**: Finally, we calculate the accuracy of the model by comparing the predicted labels with the actual labels.

Conclusion:

In conclusion, this water quality analysis project has provided valuable insights into the safety and suitability of the examined water sources. By employing a combination of data collection, rigorous testing, and advanced analytical techniques, we have gained a comprehensive understanding of the key parameters influencing water quality. This information is essential for making informed decisions regarding water resource management and ensuring the well-being of the communities relying on these sources. Moving forward, continued monitoring and proactive measures will be instrumental in upholding and improving water quality standards, ultimately safeguarding the health and vitality of our environment and its inhabitants.