WATER QUALITY ANALYSIS

Project summary:

Water quality describes the condition of water, including chemical, physical, and biological characteristics, usually with respect to its suitable for a particular purpose such as drinking.

Introduction:

The objective of water quality monitoring is to object quantitative information on the physical, chemical and biological characteristics of water via statistical sampling.

Definition:

Water quality analysis is also called hydro chemical analysis. That is to use chemical and physical methods to determine the content of various chemical components in water.

Abstract:

Water is perhaps the most precious natural resource after air. Though the surface of the earth ismostly consists of water, only a small part of it is usable, which makes this resource limited. Poor condition of water bodies are not only the indictor of environmental degradation, it is also a threat to the ecosystem.

K E Y W O R D S:

• Water Quality Assessment

• Water Quality Analysis

• Chain of Custody

Water Quality

Water Quality can be defined as the chemical, physical and biological characteristics of water,usually in respect to its suitability for a designated use. Water can be used for recreation, drinking,fisheries, agriculture or industry. Each of these designated uses has different defined chemical, physical and biological standards necessary to support that use.

Example :

There are stringent standards for water to be used for drinking or swimming compared to that used in agriculture or industry.

Water Quality Analysis:

After many years of research, water quality standards are put in place to ensure the suitability of efficient use of water for a designated purpose. Water quality analysis is to measure the requiredparameters of water, following standard methods, to check whether they are in accordance with theStandard.

Chart : Parameters for Water Quality Analysis

Selection of Methods:

The methods of water quality analysis are selected according to the requirement. The factors playingkey role for the selection of methods are:

(i) Volume and number of sample to be analysed

(ii) Cost of analysis

(iii) Precision required

Precision and Accuracy of Method Selected as Per RequirementWhat precision and accuracy to be maintained against a particular method is selected according to the need.

(i) Cost

(ii) Parameter

(iii) Use

Chain–of–Custody Procedures:

Properly designed and executed chain-of-custody forms will ensure sample integrity from collectionto data reporting. This includes the ability to trace possession and handling of the sample from the time of collection through analysis and final disposition. This process is referred to as “chain of- custody” and is required to demonstrate sample control when the data are to be used for regulation or litigation. A sample is considered to be under a person’s custody, if it is in the individual’s physical possession, in the individual’s sight, secured and tamper-proofed by that individual, or secured in an area restricted to authorized personnel. The following procedures summarize the major aspects of chain-of-custody::

Anomaly detection techniques:

Some common approaches for anomaly detection in water quality analysis include:

1**. **Statistical Methods****:

Techniques like Z-scores, Grubbs' test, or other statistical tests can be applied to detect outliers in the data.

2. **Machine Learning Algorithms****:

Algorithms like Isolation Forest, One-Class SVM, or Auto encoders can be trained on historical data to learn normal patterns and flag anomalies.

3. **Time Series Analysis****:

Methods like Exponential Smoothing, ARIMA, or Seasonal-Trend decomposition can be used to model and forecast expected values, then identify deviations.

4. **Cluster Analysis:**

Grouping similar water quality data points together and identifying points that do not belong to any cluster can help detect anomalies.

Here are some advanced techniques and considerations you might explore in phase 2:

1. **Ensemble Methods:** Combining multiple anomaly detection algorithms or models can often lead to more robust results.

2. **Deep Learning Approaches:**

Consider using deep learning models like convolutional neural networks (CNNs) or recurrent neural networks (RNNs) to capture complex patterns in the data.

3. **Transfer Learning:**

If applicable, you might leverage pre-trained models on related tasks and fine-tune them for water quality anomaly detection.

4. **Temporal and Spatial Analysis:**

Incorporating the temporal and spatial dimensions of the data can provide a more comprehensive understanding of water quality trends and anomalies.

5. **Data Fusion:**

Integrating data from multiple sources (e.g., sensors, satellite imagery, weather data) can enhance the accuracy of anomaly detection.

6. **Real-Time Monitoring:**

Implementing a system for real-time monitoring can allow for immediate response to detected anomalies, especially in critical situations.

7. **Model Interpretability:**

Ensuring that the anomaly detection models are interpretable is crucial for understanding the factors contributing to identified anomalies.

8. **Data Augmentation:**

Augmenting the dataset with synthetic data or additional features can help improve model generalization.

9. **Dynamic Thresholding**:

Using adaptive thresholds that change with time or environmental conditions can be more effective in certain scenarios.

Collecting water quality data :

Importing libraries:

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

Load the Data: Use a data analysis tool such as Python with libraries like Pandas to load the dataset into your environment.

main_dat = pd.read_csv("/kaggle/input/water-potability/water_potability.csv")

ks = main_dat.copy() #copy of original data set

Explore the Data: Get familiar with the dataset by examining its structure, columns, and basic statistics.

ks.head()

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
0	NaN	204.890456	20791.31898	7.300212	368.516441	564.308654	10.379783	86.990970	2.963135	0
1	3.716080	129.422921	18630.05786	6.635246	NaN	592.885359	15.180013	56.329076	4.500656	0
2	8.099124	224.236259	19909.54173	9.275884	NaN	418.606213	16.868637	66.420093	3.055934	0
3	8.316766	214.373394	22018.41744	8.059332	356.886136	363.266516	18.436525	100.341674	4.628771	0
4	9.092223	181.101509	17978.98634	6.546600	310.135738	398.410813	11.558279	31.997993	4.075075	0

ks.sample()

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
659	5.555353	154.300684	20503.430050	9.644997	313.470297	355.206969	18.468690	75.140362	4.536146	0
2272	8.384296	223.328185	27463.654790	6.476753	352.952803	318.042648	10.645164	64.209337	3.460998	0
2006	6.538207	214.992866	12330.406570	7.300092	389.817036	465.352665	22.089402	24.532773	3.426266	1
801	8.900865	211.306812	9592.151333	8.863272	348.437820	333.775327	18.267951	68.333170	4.518751	1
2886	NaN	206.036295	8667.720239	6.329952	353.529381	599.546019	21.118938	55.932324	4.128746	0

Ks. Shape()

```
(3276, 10)
```

Ks.columns()

```
Index(['ph', 'Hardness', 'Solids', 'Chloramines', 'Sulfate', 'Conductivity',
       'Organic_carbon', 'Trihalomethanes', 'Turbidity', 'Potability'],
      dtype='object')
```

Checking Null values:

pd.isnull(ks).sum()

```
ph                 491
Hardness             0
Solids               0
Chloramines          0
Sulfate            781
Conductivity         0
Organic_carbon       0
Trihalomethanes    162
Turbidity            0
Potability           0
dtype: int64
```

ks.dropna(inplace=True)

pd.isnull(ks).sum()

ks.describe()

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
count	2011.000000	2011.000000	2011.000000	2011.000000	2011.000000	2011.000000	2011.000000	2011.000000	2011.000000	2011.000000
mean	7.085990	195.968072	21917.441375	7.134338	333.224672	426.526409	14.357709	66.400859	3.969729	0.403282
std	1.573337	32.635085	8642.239815	1.584820	41.205172	80.712572	3.324959	16.077109	0.780346	0.490678
min	0.227499	73.492234	320.942611	1.390871	129.000000	201.619737	2.200000	8.577013	1.450000	0.000000
25%	6.089723	176.744938	15615.665390	6.138895	307.632511	366.680307	12.124105	55.952664	3.442915	0.000000
50%	7.027297	197.191839	20933.512750	7.143907	332.232177	423.455906	14.322019	66.542198	3.968177	0.000000
75%	8.052969	216.441070	27182.587065	8.109726	359.330555	482.373169	16.683049	77.291925	4.514175	1.000000
max	14.000000	317.338124	56488.672410	13.127000	481.030642	753.342620	27.006707	124.000000	6.494749	1.000000

ks.nunique()

```
ph                2011
Hardness          2011
Solids            2011
Chloramines       2011
Sulfate           2011
Conductivity      2011
Organic_carbon    2011
Trihalomethanes   2011
Turbidity         2011
Potability           2
dtype: int64
```

ks.info()

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2011 entries, 3 to 3271
Data columns (total 10 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   ph               2011 non-null   float64
 1   Hardness         2011 non-null   float64
 2   Solids           2011 non-null   float64
 3   Chloramines      2011 non-null   float64
 4   Sulfate          2011 non-null   float64
 5   Conductivity     2011 non-null   float64
 6   Organic_carbon   2011 non-null   float64
 7   Trihalomethanes  2011 non-null   float64
 8   Turbidity        2011 non-null   float64
 9   Potability       2011 non-null   int64
dtypes: float64(9), int64(1)
memory usage: 172.8 KB
```

Handle Missing Values:

Identify missing values in the dataset. Decide how to handle missing data, which may include:

Imputation: Filling missing values with a mean, median, or mode of the column.

Visualization parameter distribution:

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:

Sulfate 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 litre (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. Portability:

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

Predictive model:

LINEAR ALGEBRA:

```
Import numpy as np
```

DATA PROCESSING:

```
Import pandas as pd
```

INPUT DATA FILES read_ only:

```
Import os
for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))
```

IMPORT THE DATA:

```
Dataframed=pd.read_csv("input/waterpotability/water_potability.csv")
```

INPUT THE HEAD:

```
dataframed.head()
```

ADDING COLUMNS AND VALUES:

```
dataframed.columns.values.tolist()
['ph',
'Hardness',
 'Solids',
 'Chloramines',
 'Sulfate',
 'Conductivity',
 'Organic_carbon',
 'Trihalomethanes',
 'Turbidity',
'Potability']
```

GET INFORMATION:

```
dataframed.info()
<class'pandas.core.frame.DataFrame'> RangeIndex:3276entries,0to3275
Datacolumns(total10columns):
 #              Column                          Non-NullCount                  Dtype
 ---             ------                          --------------                  ----
0                ph                              2785non-null
```

float64

1 Hardness 3276non-null

float64

2 Solids 3276non-null

float64

3 Chloramines 3276non-null

float64

4 Sulfate 2495non-null

float64

5 Conductivity 3276non-null

float64

6 Organic_ carbon 3276non-null

float64

7 Trihalomethanes 3114non-null

float64

8 Turbidity 3276non-null

float64

9 Potability 3276non-null int64

dtypes:float64(9),int64(1)

memoryusage:256.1KB

RETURNS DESCRIPTION:

dataframed.describe()

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
count	2785.000000	3276.000000	3276.000000	3276.000000	2495.000000	3276.000000	3276.000000	3114.000000	3276.000000	3276.000(
mean	7.080795	196.369496	22014.092526	7.122277	333.775777	426.205111	14.284970	66.396293	3.966786	0.390110
std	1.594320	32.879761	8768.570828	1.583085	41.416840	80.824064	3.308162	16.175008	0.780382	0.487849
min	0.000000	47.432000	320.942611	0.352000	129.000000	181.483754	2.200000	0.738000	1.450000	0.000000
25%	6.093092	176.850538	15666.690297	6.127421	307.699498	365.734414	12.065801	55.844536	3.439711	0.000000
50%	7.036752	196.967627	20927.833607	7.130299	333.073546	421.884968	14.218338	66.622485	3.955028	0.000000
75%	8.062066	216.667456	27332.762127	8.114887	359.950170	481.792304	16.557652	77.337473	4.500320	1.000000
max	14.000000	323.124000	61227.196008	13.127000	481.030642	753.342620	28.300000	124.000000	6.739000	1.000000

MISSING VALUES:

dataframed.isnull().sum()

DATA CLEANING:

In this phase, the team focused on the visualization and treatment of missing values to prepare the data for the exploratory data analysis and model building phase.

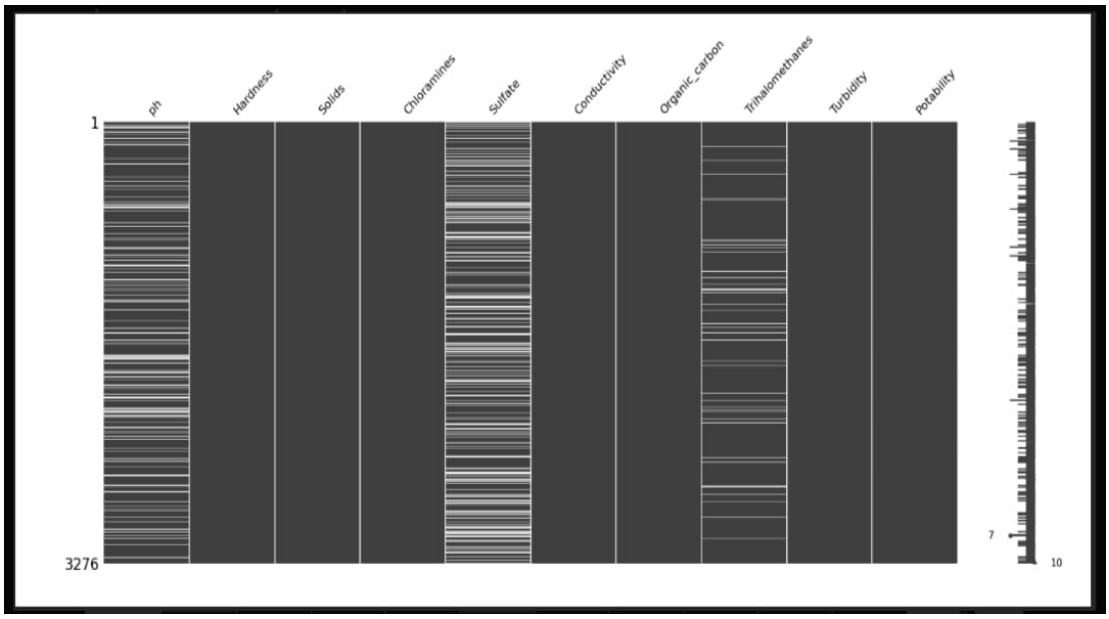
Dealing with missing values is very important for creating a powerful prediction model, therefore, this phase is vital for the study's success.

MISSING VALUE SVISUALIZATION:

Import missing noasmsno

msno.matrix(dataframed)

plt.show()



CONCLUSION:

Maintaining and improving water quality is a shared responsibility that requires the concerted effort of all stakeholders. This analysis serves as a foundation for informed decision-making and action towards ensuring clean and safe water for the community and the environment. It is imperative to act promptly and diligently to safe guard our water resources for current and future generations.