WATER QUALITY ANALYSIS

Phase 4: Development part 2

Submitted by:

Chakali Madhu Kumar

au723921243010

naturalmadhu1@gmail.com

Using visualization libraries (e.g., Matplotlib, Seaborn) to create histograms, scatter plots, and correlation matrices of water quality analysis.

Content:

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. In some regions, it has been shown that investments in water supply and sanitation can yield a net economic benefit, since the reductions in adverse health effects and health care costs outweigh the costs of undertaking the interventions.

The water\_potability.csv file contains water quality metrics for 3276 different water bodies.

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.

Visuliasation:

import pandas as pd *# data processing, CSV file I/O*

import numpy as np *# linear algebra*

import matplotlib.pyplot as plt

import seaborn as sns

import plotly.express as px

from sklearn.preprocessing import StandardScaler

data = pd.read\_csv('/kaggle/input/water-quality-and-potability/water\_potability.csv')

linkcode

data.shape *#rows,columns*

(3276, 10)

In [4]:

data.head()

Out[4]:

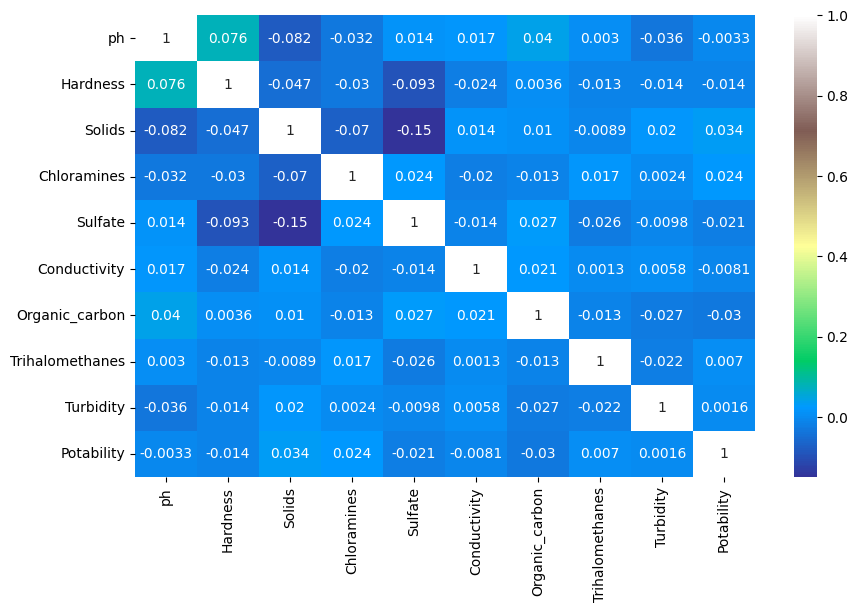
|  | ph | Hardness | Solids | Chloramines | Sulfate | Conductivity | Organic\_carbon | Trihalomethanes | Turbidity | Potability |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | NaN | 204.890455 | 20791.318981 | 7.300212 | 368.516441 | 564.308654 | 10.379783 | 86.990970 | 2.963135 | 0 |
| 1 | 3.716080 | 129.422921 | 18630.057858 | 6.635246 | NaN | 592.885359 | 15.180013 | 56.329076 | 4.500656 | 0 |
| 2 | 8.099124 | 224.236259 | 19909.541732 | 9.275884 | NaN | 418.606213 | 16.868637 | 66.420093 | 3.055934 | 0 |
| 3 | 8.316766 | 214.373394 | 22018.417441 | 8.059332 | 356.886136 | 363.266516 | 18.436524 | 100.341674 | 4.628771 | 0 |
| 4 | 9.092223 | 181.101509 | 17978.986339 | 6.546600 | 310.135738 | 398.410813 | 11.558279 | 31.997993 | 4.075075 |  |

sns.heatmap(data.corr(),annot= True,cmap='terrain')

fig= plt.gcf()

fig.set\_size\_inches(10,6)

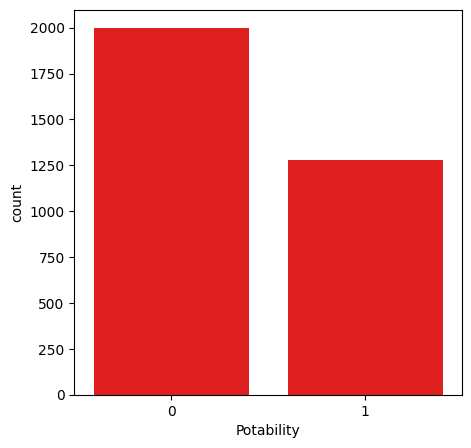
plt.show()



plt.figure(figsize=(5,5))

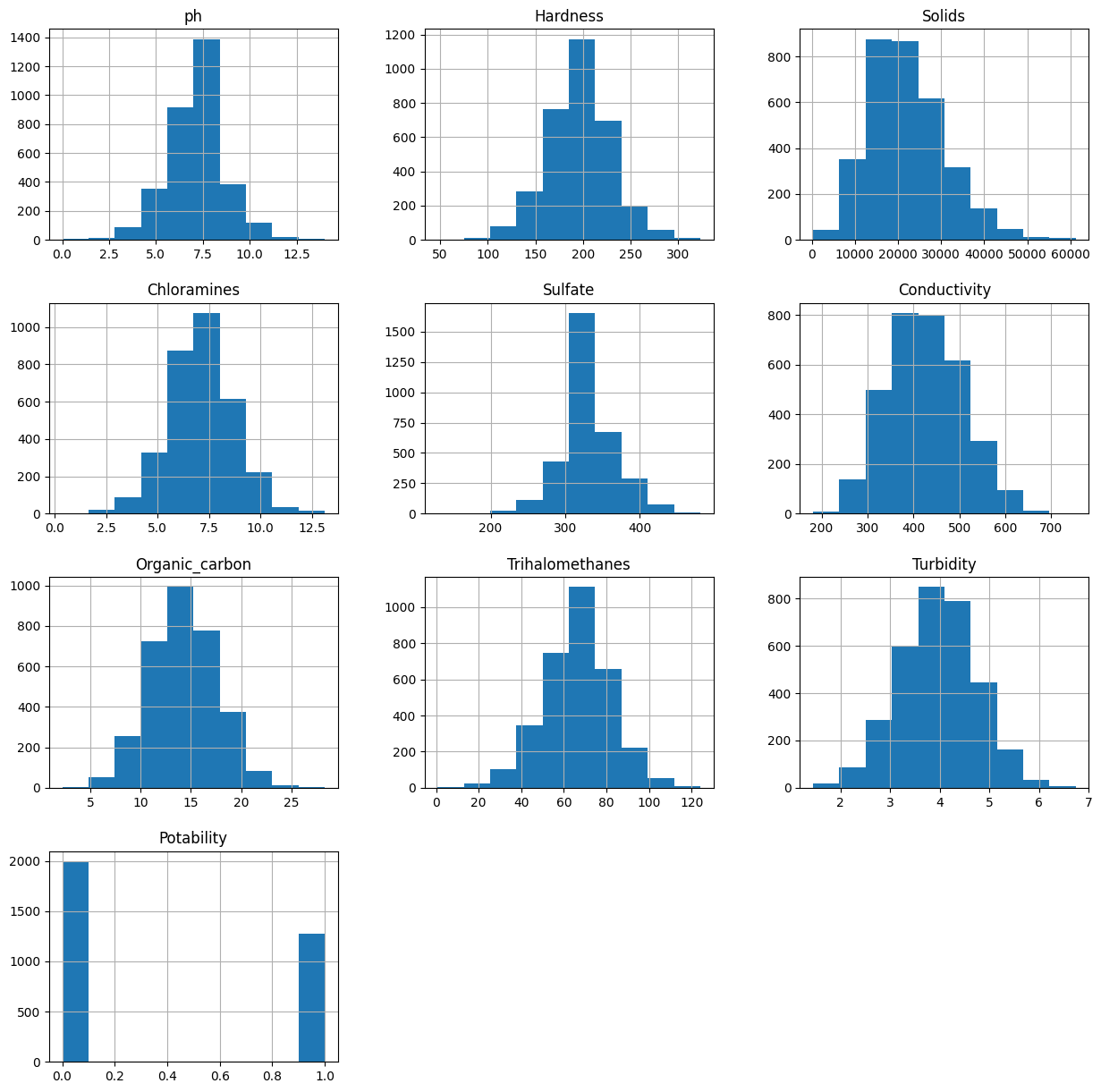
sns.countplot( x=data["Potability"], color="red")

plt.show()



data.hist(figsize=(15,15))

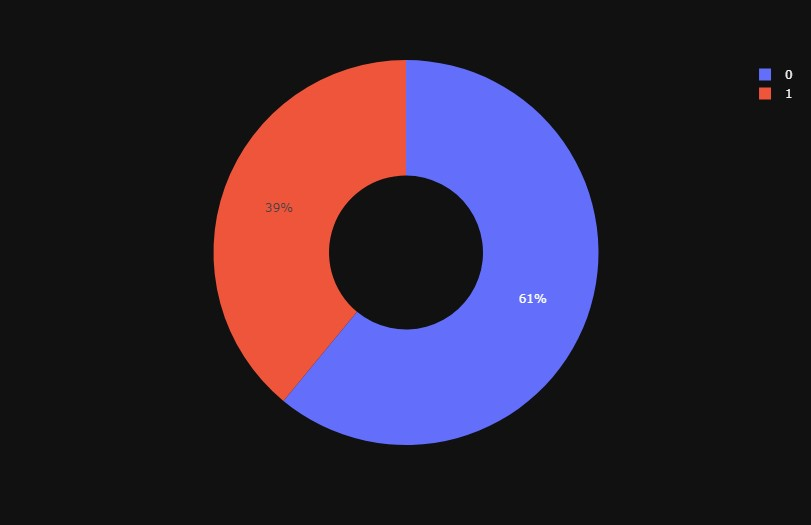
plt.show()



In [18]:

fig = px.pie(data,names ="Potability",hole = 0.4,template ="plotly\_dark")

fig.show()

predictive models to determine water potability based on water quality parameters.

Decision tree:

from sklearn.tree import DecisionTreeClassifier

dt = DecisionTreeClassifier(criterion= 'entropy', min\_samples\_split= 3,)

dt.fit(X\_train,Y\_train)

DecisionTreeClassifier

DecisionTreeClassifier(criterion='entropy', min\_samples\_split=3)

Y\_test

Output:

2541 0

2605 0

330 1

515 0

400 1

..

482 0

2970 0

50 0

839 0

374 1

Name: Potability, Length: 656, dtype: int64

Y\_prediction=dt.predict(X\_test) *#if in Y\_test 2541--> 0 is match with this data the model is correct but first is 1*

from sklearn.metrics import accuracy\_score, confusion\_matrix

confusion\_matrix(Y\_prediction,Y\_test)

*#here is TP FP True Positive is Y test and Y prediction is 1 False Positive is Y tesi is 0 Y prediction is 1 (1= good water 0 is bad water)*

*# FN TN False Negative is Y test is 1 Yprediction is 0 True Negative is Y test and Y prediction is 0*

*Output:*

array([[262, 124],

[140, 130]])

# Model Optimization /Hyper Parameter Tuning::

from sklearn.model\_selection import GridSearchCV

from sklearn.model\_selection import RepeatedStratifiedKFold

dt= DecisionTreeClassifier()

criterion = ["gini","entropy"]

splitter = ['best','random ']

min\_samples\_split=range (1,10)

parameters = dict(criterion=criterion,splitter= splitter, min\_samples\_split= min\_samples\_split)

cv= RepeatedStratifiedKFold(n\_splits = 5,random\_state=101)

grid\_search\_cv\_dt= GridSearchCV(estimator=dt, param\_grid=parameters,scoring='accuracy',cv=cv)

*#grid will use every parameter to decide the best*

*#cv is a cross validation which spliting the test to 5 pieces and compare the each other to 5 times*

*#and we using together those*

grid\_search\_cv\_dt.fit(X\_train,Y\_train)

/opt/conda/lib/python3.10/site-packages/sklearn/model\_selection/\_validation.py:378: FitFailedWarning:

1000 fits failed out of a total of 1800.

The score on these train-test partitions for these parameters will be set to nan.

If these failures are not expected, you can try to debug them by setting error\_score='raise'.

Below are more details about the failures:

--------------------------------------------------------------------------------

200 fits failed with the following error:

Traceback (most recent call last):

File "/opt/conda/lib/python3.10/site-packages/sklearn/model\_selection/\_validation.py", line 686, in \_fit\_and\_score

estimator.fit(X\_train, y\_train, \*\*fit\_params)

File "/opt/conda/lib/python3.10/site-packages/sklearn/tree/\_classes.py", line 889, in fit

super().fit(

File "/opt/conda/lib/python3.10/site-packages/sklearn/tree/\_classes.py", line 177, in fit

self.\_validate\_params()

File "/opt/conda/lib/python3.10/site-packages/sklearn/base.py", line 600, in \_validate\_params

validate\_parameter\_constraints(

File "/opt/conda/lib/python3.10/site-packages/sklearn/utils/\_param\_validation.py", line 97, in validate\_parameter\_constraints

raise InvalidParameterError(

sklearn.utils.\_param\_validation.InvalidParameterError: The 'min\_samples\_split' parameter of DecisionTreeClassifier must be an int in the range [2, inf) or a float in the range (0.0, 1.0]. Got 1 instead.

800 fits failed with the following error:

Traceback (most recent call last):

File "/opt/conda/lib/python3.10/site-packages/sklearn/model\_selection/\_validation.py", line 686, in \_fit\_and\_score

estimator.fit(X\_train, y\_train, \*\*fit\_params)

File "/opt/conda/lib/python3.10/site-packages/sklearn/tree/\_classes.py", line 889, in fit

super().fit(

File "/opt/conda/lib/python3.10/site-packages/sklearn/tree/\_classes.py", line 177, in fit

self.\_validate\_params()

File "/opt/conda/lib/python3.10/site-packages/sklearn/base.py", line 600, in \_validate\_params

validate\_parameter\_constraints(

File "/opt/conda/lib/python3.10/site-packages/sklearn/utils/\_param\_validation.py", line 97, in validate\_parameter\_constraints

raise InvalidParameterError(

sklearn.utils.\_param\_validation.InvalidParameterError: The 'splitter' parameter of DecisionTreeClassifier must be a str among {'random', 'best'}. Got 'random ' instead.

/opt/conda/lib/python3.10/site-packages/sklearn/model\_selection/\_search.py:952: UserWarning:

One or more of the test scores are non-finite: [ nan nan 0.57664122 nan 0.5798855 nan

0.58038168 nan 0.57916031 nan 0.58041985 nan

0.58118321 nan 0.58053435 nan 0.58049618 nan

nan nan 0.58519084 nan 0.58431298 nan

0.58473282 nan 0.58461832 nan 0.58278626 nan

0.58534351 nan 0.58305344 nan 0.58793893 nan]

Output:

GridSearchCV

estimator: DecisionTreeClassifier

DecisionTreeClassifier