Water Quality Analysis

PHASE 5: PROJECT DOCUMENTATION & SUBMISSION

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

Water quality analysis serves several essential objectives that are crucial for maintaining human health, preserving ecosystems, and sustaining various water-dependent activities. The main objectives of water quality analysis include

1. **Ensuring Safe Drinking Water:**

One of the primary objectives of water quality analysis is to determine the safety of drinking water. This involves testing for contaminants such as bacteria, viruses, heavy metals, pesticides, and other harmful substances that could pose health risks to humans.

2. **Protecting Aquatic Ecosystems:**

Water quality analysis helps in understanding the health of aquatic ecosystems. By monitoring parameters like dissolved oxygen, pH, and nutrient levels, scientists can assess the impact of pollutants on aquatic life. This information is vital for preserving biodiversity and the overall balance of aquatic ecosystems.

3. **Supporting Sustainable Agriculture:**

Water quality analysis is essential for agricultural practices. Testing water used for irrigation ensures that it does not contain harmful substances that could damage crops or affect the soil. Monitoring water quality helps farmers make informed decisions about water usage and conservation.

4. **Preserving Industrial Processes:**

Many industries rely on water for their operations. Analyzing water quality is critical to prevent corrosion, scaling, and fouling in industrial equipment. It ensures that water used in manufacturing processes meets specific standards, enhancing the efficiency and longevity of machinery and products.

5. **Monitoring and Mitigating Pollution:**

Regular analysis of water quality helps in identifying pollution sources. By understanding the type and source of pollutants, appropriate measures can be taken to mitigate contamination and prevent further pollution, safeguarding both surface and groundwater resources.

6. **Compliance with Regulations:**

Governments and environmental agencies set regulations and standards for water quality. Regular analysis is necessary for industries, municipalities, and other entities to ensure compliance with these standards. Failure to meet these regulations can result in penalties and legal consequences.

7. **Research and Scientific Understanding:**

Water quality analysis is fundamental for scientific research. Researchers use water quality data to study pollution patterns, climate change effects, and the impact of human activities on water bodies. This research contributes to the development of new technologies and policies aimed at improving water quality.

8. **Public Awareness:**

Water quality analysis results can be used to raise public awareness about the importance of clean water and the consequences of pollution. Informed communities are more likely to actively participate in conservation efforts and advocate for policies that protect water resources.

Data Collection

Collecting accurate and reliable data is fundamental in water quality analysis. Here's a systematic guide to collecting data for water quality analysis:

1. **Define Objectives:**

Clearly define the purpose of your study. Determine whether you're assessing water for drinking, industrial use, aquatic life or general environmental health. This guides the selection of parameters and methods.

2. **Select Sampling Sites:**

Choose representative sites based on geography, human activity, and potential sources of pollution. Ensure that the sites cover the entire area of interest.

3. **Determine Sampling Frequency:**

Decide on the frequency of sampling. Regular sampling provides a comprehensive view of water quality changes over time.

4. **Select Parameters:**

Choose parameters relevant to your objectives. Common parameters include pH, temperature, dissolved oxygen, biochemical oxygen demand (BOD), chemical oxygen demand (COD), nutrients, heavy metals, and specific pollutants based on local concerns.

5. **Gather Equipment and Supplies:**

- **Sampling Containers:**

Clean, sterile containers for collecting water samples.

- **Field Testing Kits:**

Portable devices for immediate assessment of basic parameters.

- **Laboratory Equipment:**

If lab analysis is necessary, ensure you have the appropriate tools.

- **Preservatives:**

Chemicals like acid or sodium metabisulfite to preserve samples for specific tests.

6. Sampling Procedure:

Follow these guidelines when collecting water samples:

- **Properly label containers:**

Include site name, date, and time of collection.

- **Depth considerations:**

Take samples from different depths if assessing vertical stratification.

- **Avoid contamination:**

Use clean gloves, and avoid touching the inside of the sample container or cap.

- **Sample quantity:**

Collect an adequate volume, typically a liter, for thorough analysis.

7. Transportation and Preservation:

- **Transport samples:**

Keep samples cool and in the dark during transit to the lab.

- **Preservation:**

If immediate analysis isn't possible, preserve samples according to test requirements. Some samples need to be kept cold or in the dark to prevent chemical changes.

8. Data Recording:

- **Maintain a log:**

Record all relevant information, including sampling location, date, time, weather conditions, and any deviations from the standard procedure.

- **Note any observations:**

Document any unusual color, odor, or other qualitative features.

9. **Data Analysis:**

- **Quality Control:**

Regularly calibrate instruments, check the precision of measurements, and verify results.

- **Statistical Analysis:**

Use statistical methods to identify trends and correlations in your data.

10. **Report Findings:**

Compile your data into a comprehensive report. Include methodology, results, interpretations, and, if applicable, recommendations for corrective actions.

11. **Continuous Monitoring:**

For a more holistic understanding, consider implementing continuous monitoring systems using sensors and automated equipment, especially in critical or high-risk areas.

Visualization Strategy

Visualizing water quality analysis data is essential for understanding complex datasets, identifying patterns, and communicating findings effectively. Here's a strategy for visualizing water quality analysis data:

1. **Choose the Right Visualization Tools:**

- **Graphs and Charts:**

Use line charts to show trends over time for parameters like pH, dissolved oxygen, and pollutant levels. Bar charts can compare different locations or sources.

- **Maps:**

Geographic Information System (GIS) maps can display water quality variations across regions or specific points of interest. Color-coded maps can represent different water quality levels.

- **Heatmaps:**

Useful for displaying multiple parameters across different locations. Colors represent the intensity of the parameter, allowing for quick comparisons.

- **Scatter Plots:**

Useful for showing relationships between two variables, such as dissolved oxygen levels against temperature.

2. **Key Parameters Focus:**

- **Identify Key Parameters:**

Focus on crucial parameters like pH, dissolved oxygen, and pollutant levels. Visualize these parameters prominently for easy interpretation.

- **Thresholds and Standards:**

Overlay visualizations with regulatory standards. This provides a clear indication of whether the water quality meets the required standards.

3. **Temporal Analysis:**

- **Time-Series Charts:**

Display changes in water quality parameters over time. This helps in identifying seasonal trends, pollution events, or the impact of specific interventions.

4. **Spatial Analysis:**

- **GIS Mapping:**

Create maps that show water quality variations geographically. GIS tools allow you to layer different parameters, making it easier to spot spatial patterns and hotspots of pollution.

5. **Comparative Analysis:**

- **Side-by-Side Comparisons:**

Use bar charts or box plots to compare water quality parameters between different locations, sources, or time periods.

6. **Interactive Visualizations:**

- **Dashboard Tools:**

Utilize dashboard tools like Tableau or Power BI to create interactive visualizations. Stakeholders can interact with the data, filtering information based on their specific interests or queries.

7. **Annotations and Context:**

- **Annotations:**

Add annotations to highlight specific events or findings, providing context to the visualized data.

8. **Data Integrity:**

- **Data Transparency:**

Ensure that the data sources and methodology are transparently presented alongside the visualizations. This builds trust in the visualized results.

9. **User-Centric Approach:**

- **Stakeholder Feedback:**

Gather feedback from stakeholders to understand what visualizations are most helpful for their decision-making process. Adapt the visualizations based on their needs.

10. **Storytelling:**

- **Narrative Visualizations:**

Tell a story with your visualizations. Explain the context, present the problem, show the analysis, and conclude with actionable insights. This approach makes the data more engaging and understandable.

Predictive Modeling

Predictive modeling in water quality analysis involves using statistical, mathematical, or computational techniques to predict future water quality parameters based on historical data and other relevant factors. This modeling approach is valuable for understanding how different variables influence water quality and for making informed decisions about water resource management and pollution control. Here's how predictive modeling can be applied in water quality analysis:

1. Data Collection and Preprocessing:

- **Data Collection:**

Gather historical data on water quality parameters (e.g., pH, dissolved oxygen, pollutants) from various sources such as sensors, laboratories, and remote sensing technologies.

- **Data Preprocessing:**

Cleanse the data by handling missing values, outliers, and inconsistencies. Properly formatted and clean data is crucial for accurate modeling.

2. Feature Selection and Engineering:

- **Feature Selection:**

Identify relevant features (variables) that influence water quality. Correlation analysis and domain knowledge can help select the most important features.

- **Feature Engineering:**

Create new features that might provide valuable information, such as seasonal patterns or pollutant ratios.

3. Model Selection:

- **Regression Models:**

Linear regression, multiple regression, and nonlinear regression models can predict continuous water quality parameters.

- **Classification Models:**

Logistic regression and decision trees can predict binary outcomes like water quality being above or below a certain standard.

- **Machine Learning Models:**

Algorithms like Random Forest, Support Vector Machines, and Neural Networks can capture complex patterns in the data.

4. Model Training and Validation:

- **Training:**

Use historical data to train the selected model. The model learns the patterns in the data.

- **Validation:**

Validate the model using a separate dataset not used in training. Techniques like cross-validation ensure the model's generalizability.

5. Model Evaluation and Optimization:

- **Evaluation Metrics:**

Use metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), or R-squared to evaluate the model's accuracy.

- **Optimization:**

Fine-tune the model parameters and features to improve its performance. This process might involve techniques like grid search or random search.

6. Deployment and Monitoring:

- **Deployment:**

Implement the model in a real-world scenario to make predictions. It could be integrated into a monitoring system for continuous assessment.

- **Monitoring:**

Continuously monitor the model's performance and retrain it periodically with new data to maintain accuracy, especially if the environmental conditions change.

7. Decision Support and Policy Making:

- **Utilize Predictions:**

Use the predictions for proactive decision-making, such as issuing warnings when water quality is predicted to deteriorate.

- **Policy Formulation:**

Base environmental policies and regulations on predictive models to prevent and control water pollution effectively.

Challenges of Water Quality Analysis:

Water quality analysis faces several challenges, ranging from technological and methodological limitations to issues related to data interpretation and policy implementation. Here are some of the key challenges associated with water quality analysis:

1. Complexity of Contaminants:

Water sources can be contaminated by a wide range of pollutants, including chemicals, heavy metals, pathogens, and emerging contaminants like pharmaceuticals and microplastics. Developing methods to detect and quantify these diverse pollutants is a significant challenge**.**

2. Detection Limits:

Some contaminants require extremely sensitive detection methods, and detecting them at low concentrations (parts per billion or even parts per trillion) is a technical challenge. Improving the sensitivity of analytical techniques is crucial.

3.Sampling Variability:

Water quality can vary spatially and temporally. Obtaining representative samples that accurately reflect the quality of the entire water body is challenging. Inadequate sampling can lead to inaccurate assessments of water quality.

4. Data Accuracy and Precision:

Analytical methods must be accurate and precise to ensure reliable results. Calibration, equipment maintenance, and proper quality control procedures are essential but can be difficult to maintain consistently.

5.Integration:

Integrating data from various sources and formats (field measurements, lab analyses, remote sensing) into a comprehensive dataset can be complex. Developing standardized formats and protocols for data integration is a challenge.

Annovation And Future Plans:

1. Sensor Technology:

Advancements in sensor technology allow the development of compact, affordable, and portable sensors for real-time monitoring. These sensors can measure various parameters such as pH, dissolved oxygen, and specific pollutants, providing instant data without the need for extensive lab work.

2. Nanotechnology:

Nanomaterials are being utilized to create highly sensitive and selective sensors. Nanoscale sensors can detect trace amounts of contaminants and are especially useful for monitoring emerging pollutants and heavy metals in water.

3.Remote Sensing and IoT:

Integration of remote sensing technologies, satellite imagery, and the Internet of Things (IoT) enables continuous monitoring of large water bodies. IoT devices can transmit real-time data to central databases, allowing for timely analysis and response to water quality changes.

4. Data Analytics and Artificial Intelligence:

Advanced data analytics, machine learning, and AI algorithms are being applied to process vast amounts of water quality data efficiently. These techniques can identify patterns, predict trends, and provide valuable insights for decision-making and early warning systems.

5.Biosensors:

Biological sensors or biosensors use living organisms or biological molecules to detect pollutants. These sensors are highly specific, allowing for the detection of specific contaminants with high accuracy. Biosensors are being developed for various applications, including detecting pathogens and chemical pollutants.

Solving The Challenges for Innovation

Innovations in water quality analysis are instrumental in overcoming the challenges associated with monitoring, detecting contaminants, ensuring data accuracy, and implementing effective policies. Here's how innovation can address these challenges:

1.Improved Sensing Technologies:

- **Challenge:**
  Traditional lab-based analysis is time-consuming and costly.
- **Innovation:**
  Real-time sensors and microfluidic devices enable rapid on-site analysis, reducing costs and providing instant results. These sensors can be deployed in remote or resource-limited areas.

2.Enhanced Sensitivity and Selectivity:

- **Challenge:**
  Detection of trace amounts of contaminants is difficult.
- **Innovation:**
  Nanotechnology and advanced materials enable the development of highly sensitive and selective sensors. Nanomaterial-based sensors can detect even low concentrations of pollutants, ensuring accurate analysis.

3. Data Accuracy and Interpretation:

- **Challenge:**

Ensuring data accuracy and interpreting vast datasets are complex tasks.

- **Innovation:**
  Artificial intelligence (AI) and machine learning algorithms process large datasets, identify patterns, and provide real-time analysis. Predictive analytics help in anticipating water quality changes, enabling proactive measures.

4. Remote Sensing and IoT Integration:

- **Challenge:**
  Monitoring large water bodies in real-time is challenging.
- **Innovation:**

Remote sensing technologies, coupled with IoT devices, provide continuous monitoring and transmit data to central databases. This integration ensures timely analysis and immediate responses to water quality fluctuations.

5. Citizen Science Initiatives:

- **Challenge:**
  Limited resources for extensive data collection.
- **Innovation:**
  Engaging citizens through mobile apps and affordable testing kits increases data collection points. Crowdsourced data enhance the quantity and diversity of information available for analysis, empowering communities to actively participate in water quality monitoring**.**

Mechine learning Algorithms for predictive Analysis

1. Linear Regression:

- **Use:**
  Predicting a numerical value (e.g., pollutant concentration) based on one or more input variables (e.g., temperature, pH).

Advantage:

- **Simple and easy to interpret.**
- **Useful for understanding linear relationships between variables.**

2.Decision Trees:

- **Use:**
  Classification and regression tasks. Decision trees split the data based on features to make predictions**.**

Advantage:

- **Easy to understand**
- **handle categorical and numerical data**
- **can reveal important variables in the prediction process.**

3. Random Forest:

- **Use:**
  Ensemble method using multiple decision trees for prediction.

Advantage:

- **Improved accuracy compared to individual decision trees.**
- **Handles overfitting and works well with large and diverse datasets.**

4. Gradient Boosting:

- **Use:**
  Sequentially adds models (typically decision trees) to correct errors of previous models**.**

Advantage:

- **Often provides higher accuracy than random forests, especially when tuned properly.**
- **It combines the predictions of several base estimators in a way that minimizes errors.**

5. Support Vector Machines (SVM):

- **Use:**
  Classification and regression tasks. SVM finds the optimal hyperplane that best divides the data into classes or predicts a continuous outcome.

Advantage:

- **Effective in high-dimensional spaces**
- **Can handle both linear and non-linear relationships between variables.**

6. Neural Networks:

- **Use:**
  Deep learning techniques for complex, non-linear relationships within large datasets.

  Advantage:
  - **Can learn intricate patterns in data but require large amounts of data for training and can be computationally intensive.**

Data Loading:

In water quality analysis, loading data accurately and efficiently is crucial for obtaining meaningful results. Here is a step-by-step guide on how to load and handle data for water quality analysis

1. Define Your Objectives:

Clearly define the objectives of your water quality analysis. Understand what parameters you need to measure, and what kind of data (e.g., chemical, physical, biological) you will be dealing with.

2. Data Collection:

Collect water samples from various sources such as rivers, lakes, wells, or wastewater treatment plants. Ensure that the samples are collected using proper techniques and stored appropriately to maintain their integrity.

3. Data Types:

Water quality data can include various types such as numerical (pH levels, chemical concentrations), categorical (water type, pollution level), spatial (location data), and temporal (time and date of sample collection) data.

4. Data Sources:

Obtain data from reliable sources such as government agencies, research institutions, or your own field measurements. Ensure that the data is accurate, complete, and relevant to your analysis.

5. Data Storage:

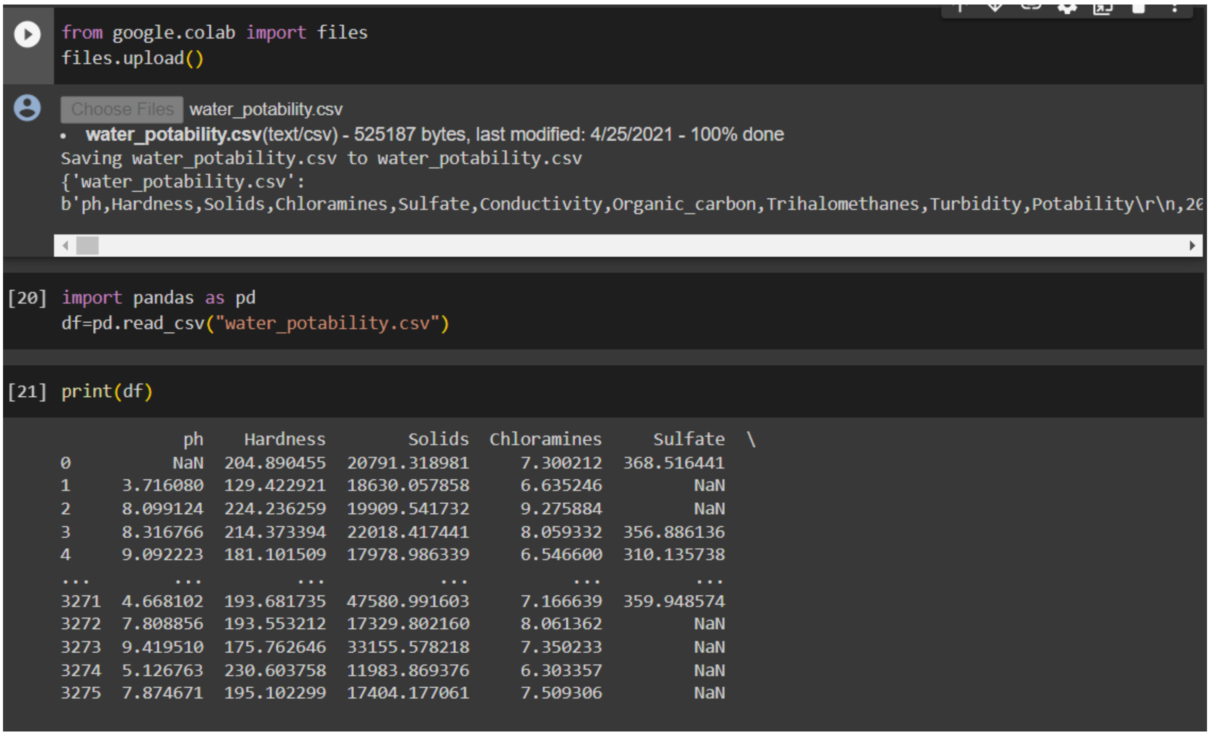
Organize your data in a structured format. Consider using spreadsheets, databases, or specialized software designed for environmental data management. Each parameter should ideally have its own column, and each row should represent a unique sample.

Code:

import pandas as pd

df=pd.read_csv(“water_probability.csv”)

Print(df)



```
from google.colab import files
files.upload()
```

Choose Files water_potability.csv

- **water_potability.csv**(text/csv) - 525187 bytes, last modified: 4/25/2021 - 100% done

```
Saving water_potability.csv to water_potability.csv
{'water_potability.csv':
b'ph,Hardness,Solids,Chloramines,Sulfate,Conductivity,Organic_carbon,Trihalomethanes,Turbidity,Potability\r\n,2
```

```
[20] import pandas as pd
     df=pd.read_csv("water_potability.csv")
```

```
[21] print(df)
```

```
            ph    Hardness        Solids  Chloramines     Sulfate  \
0          NaN  204.890455  20791.318981     7.300212  368.516441
1     3.716080  129.422921  18630.057858     6.635246         NaN
2     8.099124  224.236259  19909.541732     9.275884         NaN
3     8.316766  214.373394  22018.417441     8.059332  356.886136
4     9.092223  181.101509  17978.986339     6.546600  310.135738
...        ...         ...           ...          ...         ...
3271  4.668102  193.681735  47580.991603     7.166639  359.948574
3272  7.808856  193.553212  17329.802160     8.061362         NaN
3273  9.419510  175.762646  33155.578218     7.350233         NaN
3274  5.126763  230.603758  11983.869376     6.303357         NaN
3275  7.874671  195.102299  17404.177061     7.509306         NaN
```

Fiq – 1 : Data Loading

Data Preprocessing:

Data preprocessing is a crucial step in any data analysis or machine learning project. It involves cleaning, transforming, and organizing the raw data into a format suitable for analysis or modeling. Here's why data preprocessing is essential

1. Handling Missing Data:

Real-world datasets often have missing values. Preprocessing techniques such as imputation (filling in missing values) ensure that the analysis or model-building process is not compromised due to missing data.

2. Dealing with Noisy Data:

Noise in data can come from various sources, leading to inaccuracies. Preprocessing techniques such as smoothing can help reduce noise, ensuring that the data used for analysis is more reliable.

3. Handling Categorical Data:

Machine learning algorithms usually work with numerical data, so categorical variables need to be converted into numerical representations. Techniques like one-hot encoding or label encoding are used to convert categorical data into a format suitable for algorithms.

4. Feature Scaling:

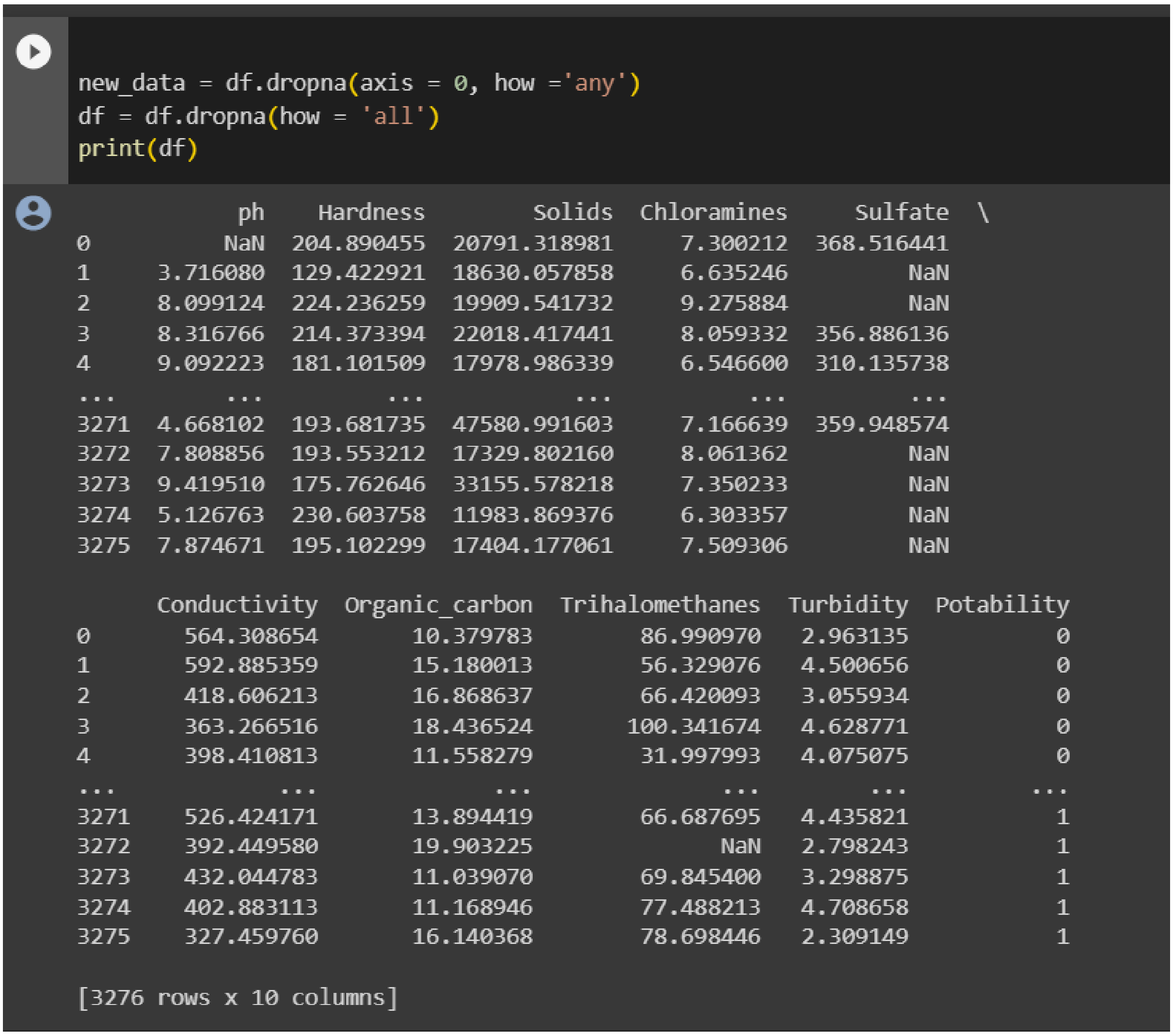
Features often have different scales. Algorithms like Support Vector Machines and k-Nearest Neighbors are sensitive to feature scales. Preprocessing techniques like min-max scaling or standardization (Z-score normalization) bring all features to a similar scale, preventing one feature from dominating due to its larger scale.

5. Feature Engineering:

Preprocessing involves creating new features from existing ones to enhance the model's performance. These new features might capture important patterns in the data, improving the model's accuracy and interpretability.

Code:

```
new_data = df.dropna(axis = 0, how = ’any’)
    Df = df.dropna(how = ’all’)
    Print(df)
```



```
new_data = df.dropna(axis = 0, how ='any')
df = df.dropna(how = 'all')
print(df)
```

```
            ph    Hardness        Solids  Chloramines     Sulfate  \
0          NaN  204.890455  20791.318981     7.300212  368.516441
1     3.716080  129.422921  18630.057858     6.635246         NaN
2     8.099124  224.236259  19909.541732     9.275884         NaN
3     8.316766  214.373394  22018.417441     8.059332  356.886136
4     9.092223  181.101509  17978.986339     6.546600  310.135738
...        ...         ...           ...          ...         ...
3271  4.668102  193.681735  47580.991603     7.166639  359.948574
3272  7.808856  193.553212  17329.802160     8.061362         NaN
3273  9.419510  175.762646  33155.578218     7.350233         NaN
3274  5.126763  230.603758  11983.869376     6.303357         NaN
3275  7.874671  195.102299  17404.177061     7.509306         NaN

      Conductivity  Organic_carbon  Trihalomethanes  Turbidity  Potability
0       564.308654       10.379783        86.990970   2.963135           0
1       592.885359       15.180013        56.329076   4.500656           0
2       418.606213       16.868637        66.420093   3.055934           0
3       363.266516       18.436524       100.341674   4.628771           0
4       398.410813       11.558279        31.997993   4.075075           0
...            ...             ...              ...        ...         ...
3271    526.424171       13.894419        66.687695   4.435821           1
3272    392.449580       19.903225              NaN   2.798243           1
3273    432.044783       11.039070        69.845400   3.298875           1
3274    402.883113       11.168946        77.488213   4.708658           1
3275    327.459760       16.140368        78.698446   2.309149           1

[3276 rows x 10 columns]
```

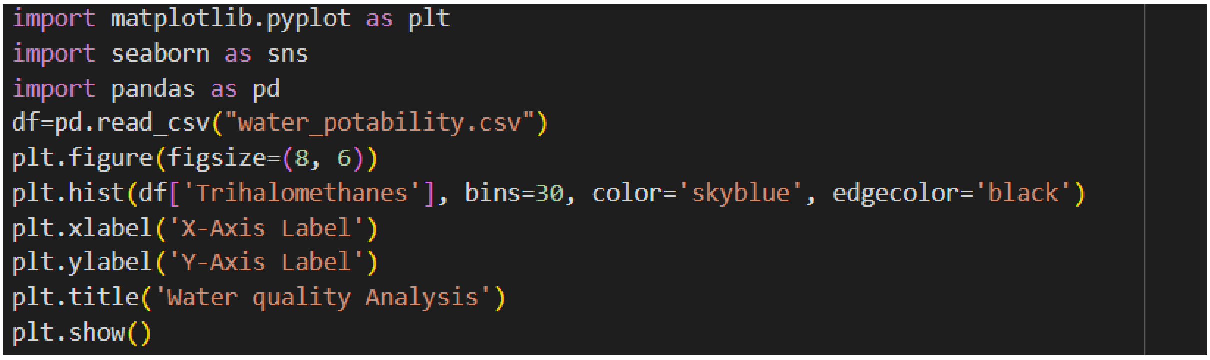
Fiq – 2 : DataPreprocessing

Visualization:

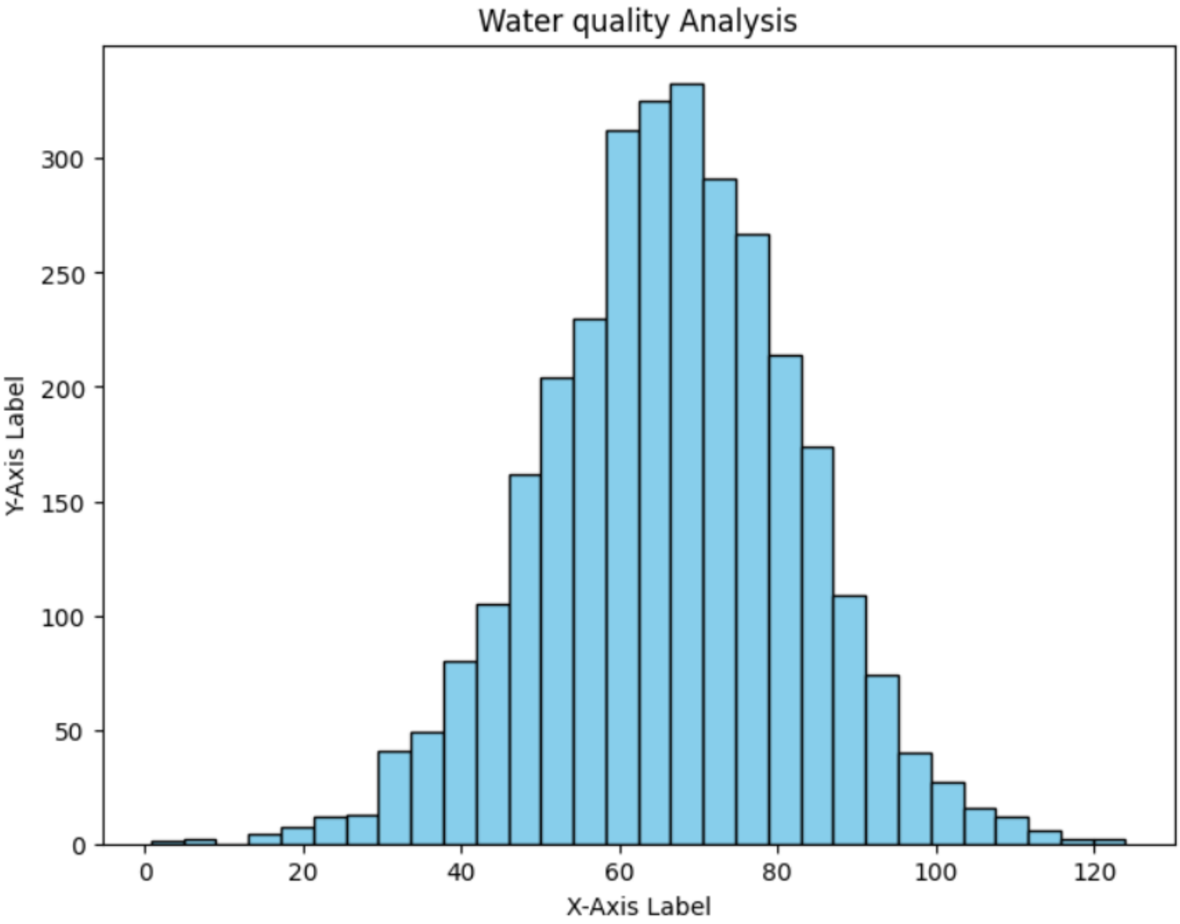
Loading the dataset allows you to create visualizations. Data visualizations (charts, graphs, maps, etc.) provide a clear and intuitive way to convey complex information and help stakeholders understand the insights derived from the data.

Code:

```
 import matplotlib.pyplot as plt
 import seaborn as sns
 import pandas as pd
 data = pd.read_csv('water_portability.csv')
plt.figure(figsize=(8, 6))
plt.hist(data['column_name'], bins=30, color='skyblue', edgecolor='black')
plt.xlabel('X-Axis Label')
plt.ylabel('Y-Axis Label')
plt.title('water Quality Analysis')
plt.show()
```



```
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd
df=pd.read_csv("water_potability.csv")
plt.figure(figsize=(8, 6))
plt.hist(df['Trihalomethanes'], bins=30, color='skyblue', edgecolor='black')
plt.xlabel('X-Axis Label')
plt.ylabel('Y-Axis Label')
plt.title('Water quality Analysis')
plt.show()
```



Fiq – 3: Virtualization for Histogram

Visualization and Algorithms

Visualization libraries are essential tools in data analysis and interpretation. They allow you to represent complex data sets in graphical or pictorial formats, making it easier to identify patterns, trends, and insights. Here are some advantages of using visualization libraries.

Advantages of Visualization Libraries:

1. Data Comprehension:

Visualizations provide a clear and concise way to understand large and complex datasets. Patterns and trends that might not be apparent in raw data can be easily identified through visual representations.

2. Communication:

Visualizations are powerful tools for communicating findings to non-technical stakeholders. A well-designed graph or chart can convey insights more effectively than raw data, making it easier for others to understand the information.

3. Decision Making:

Visualizations enable better decision-making by providing a visual summary of data. Decision-makers can quickly grasp the situation and make informed choices based on the visualized information.

4. Identification of Outliers:

Visualizations can highlight outliers or anomalies in the data, making it easier to detect errors or unusual patterns that might require further investigation.

5. Exploratory Data Analysis (EDA):

Visualization libraries are crucial for EDA, allowing data scientists to explore the data, generate hypotheses, and identify relationships between variables.

6. Storytelling:

Visualizations can be used to tell a compelling data-driven story. By creating a sequence of visualizations, you can guide your audience through a narrative, leading to a better understanding of the data and its implications.

Types of Visualization Libraries

Histogram

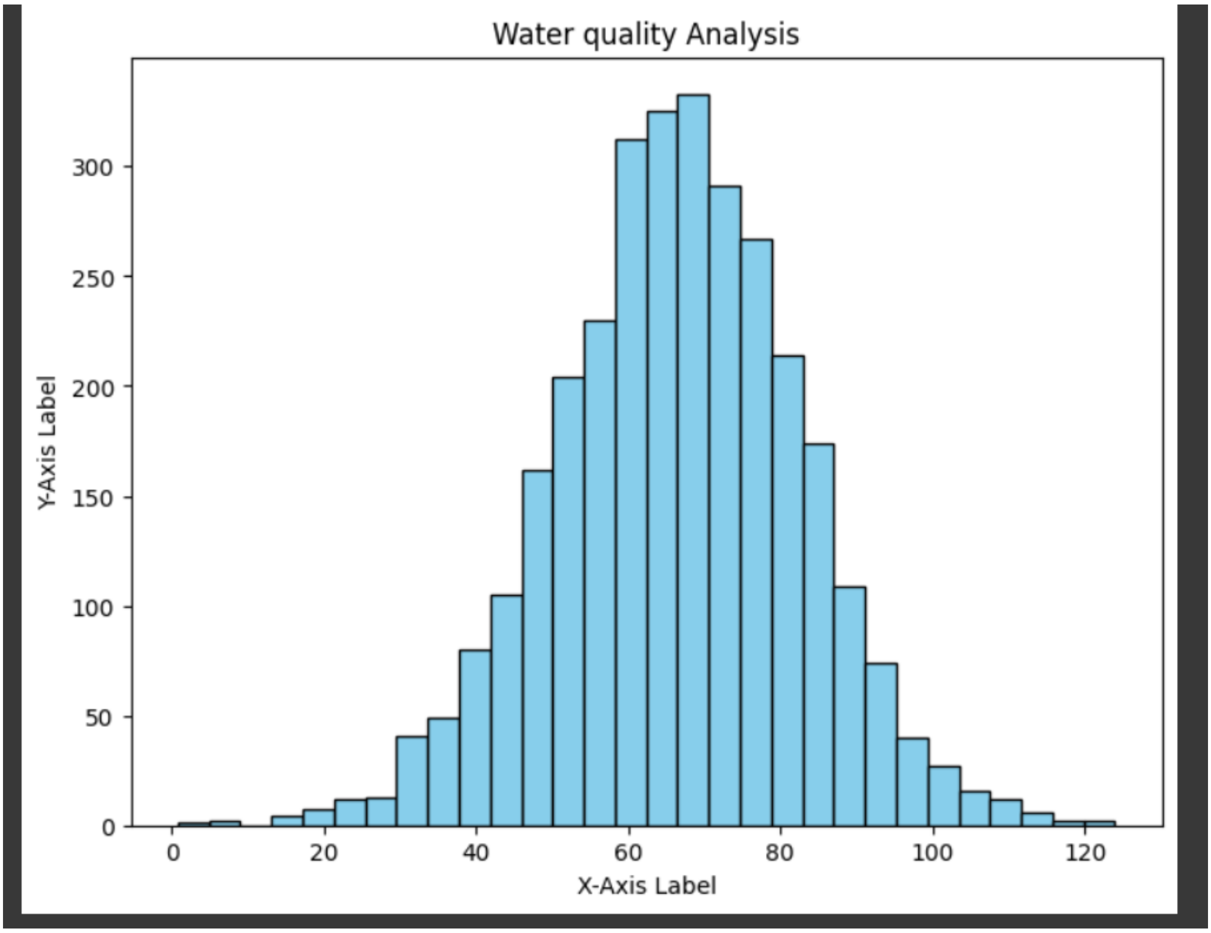
A histogram is a graphical representation of the distribution of a dataset. It is an estimate of the probability distribution of a continuous variable. To construct a histogram, the first step is to "bin" the range of values—that is, divide the entire range of values into a series of intervals—and then count how many values fall into each interval. The bins are usually specified as consecutive, non-overlapping intervals of a variable. The bins (intervals) must be adjacent and are often (but not necessarily) of equal size.

1. **Data Distribution:**
   Histograms provide a clear visual summary of the distribution of the data. You can easily see whether the data is symmetric, skewed, unimodal, bimodal, or multimodal.
2. **Central Tendency:**
   Histograms can help you identify the central tendency of the data, such as the mean, median, and mode. For example, in a symmetric distribution, these measures tend to be around the center of the histogram.
3. **Variability:**
   You can assess the variability or spread of the data. A wide histogram indicates high variability, while a narrow histogram indicates low variability.
4. **Shape:**
   Histograms can reveal the shape of the data distribution. Common shapes include normal (bell-shaped)

CODE:

```
import pandas as pd import matplotlib.pyplot as plt
import seaborn as sns
data = pd.read_csv("water.zip")
plt.figure(figsize=(8, 6))
plt.hist(data['Turbidity'], bins=30, color='green', edgecolor='blue')
 plt.title('Turbidity')
plt.xlabel('X-axis label')
 plt.ylabel('Y-axis label')
 plt.show()
```

OUTPUT:



Scatter Plots

It is a popular type of data visualization that uses Cartesian coordinates to display values for two variables. Each data point is represented as a dot, allowing you to observe the relationship between the two variables. Here are some advantages of using scatter plots in data analysis:

1. Visualizing Relationships:

Scatter plots help you understand the relationship between two variables. You can quickly identify patterns such as linear or non-linear correlations, clusters, or outliers.

2. Identifying Correlations:

Scatter plots are excellent for identifying correlations between variables. Positive correlations (variables increase together), negative correlations (one variable increases while the other decreases), or the absence of correlation can be easily spotted.

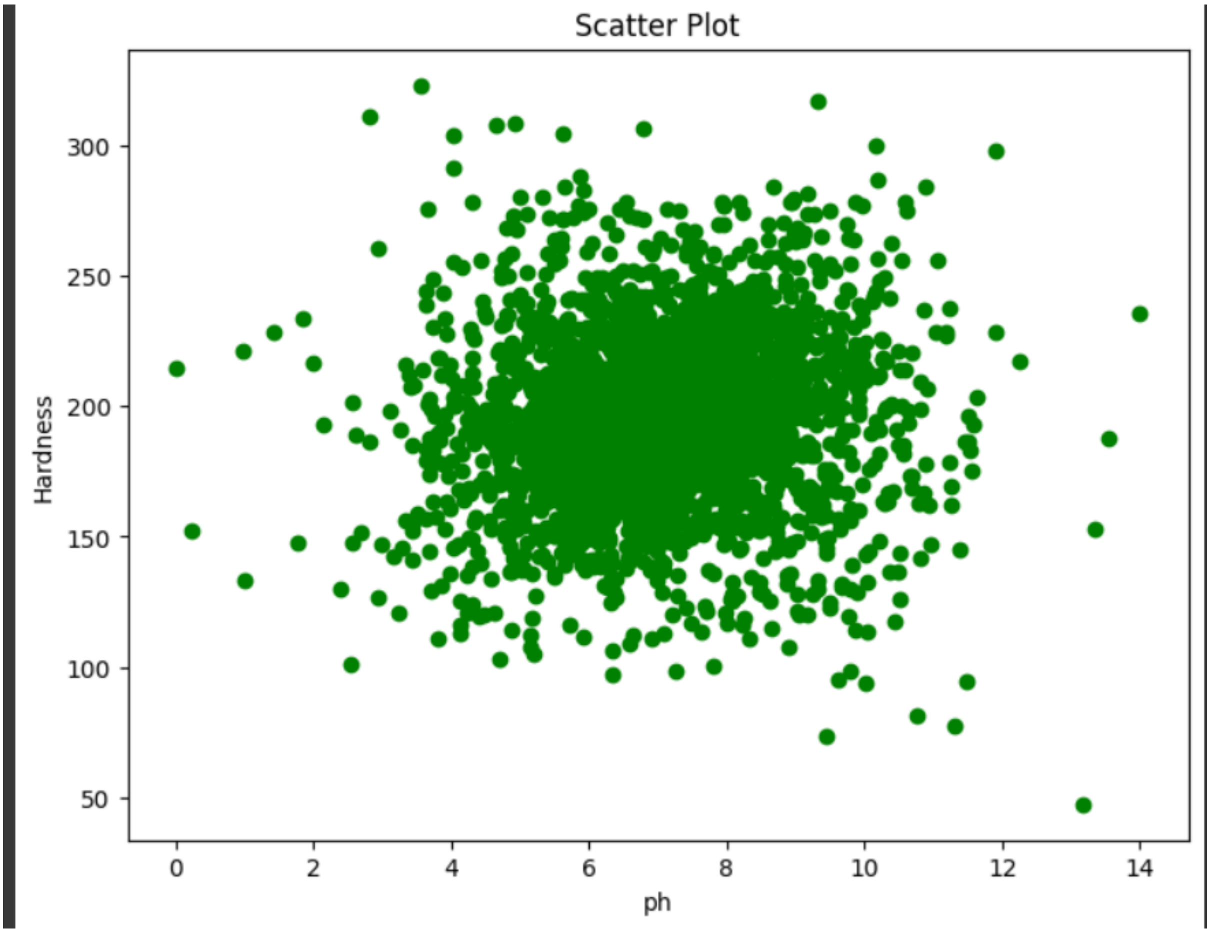
3. Outlier Detection:

Outliers, or data points that significantly differ from the rest of the data, can be visually identified on a scatter plot. Outliers might be errors in the data or represent significant events that need further investigation.

CODE:

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
data = pd.read_csv("water.zip")
plt.figure(figsize=(8, 6))
plt.scatter(data['ph'], data['Hardness'], color='green', marker='o')
plt.title('Scatter Plot')
 plt.xlabel('ph')
plt.ylabel('Hardness')
plt.show()
```

OUTPUT:



Correlation Matrices

Correlation matrices are used in data analysis and statistics for several important reasons

1. Understanding Relationships:

Correlation matrices show how strongly pairs of variables are related. A correlation close to 1 indicates a strong positive relationship, while a correlation close to -1 indicates a strong negative relationship. A correlation near 0 suggests a weak or no linear relationship between variables.

2. Feature Selection:

In machine learning and statistics, understanding correlations can help in feature selection. Highly correlated variables might provide redundant information. Identifying and removing highly correlated features can improve the performance of some machine learning algorithms and simplify the model interpretation.

3. Multicollinearity Detection:

Correlation matrices are used to identify multicollinearity in regression analysis. Multicollinearity occurs when two or more independent variables in a regression model are highly correlated, leading to unstable coefficient estimates. Identifying multicollinearity helps in making necessary adjustments to the regression model.

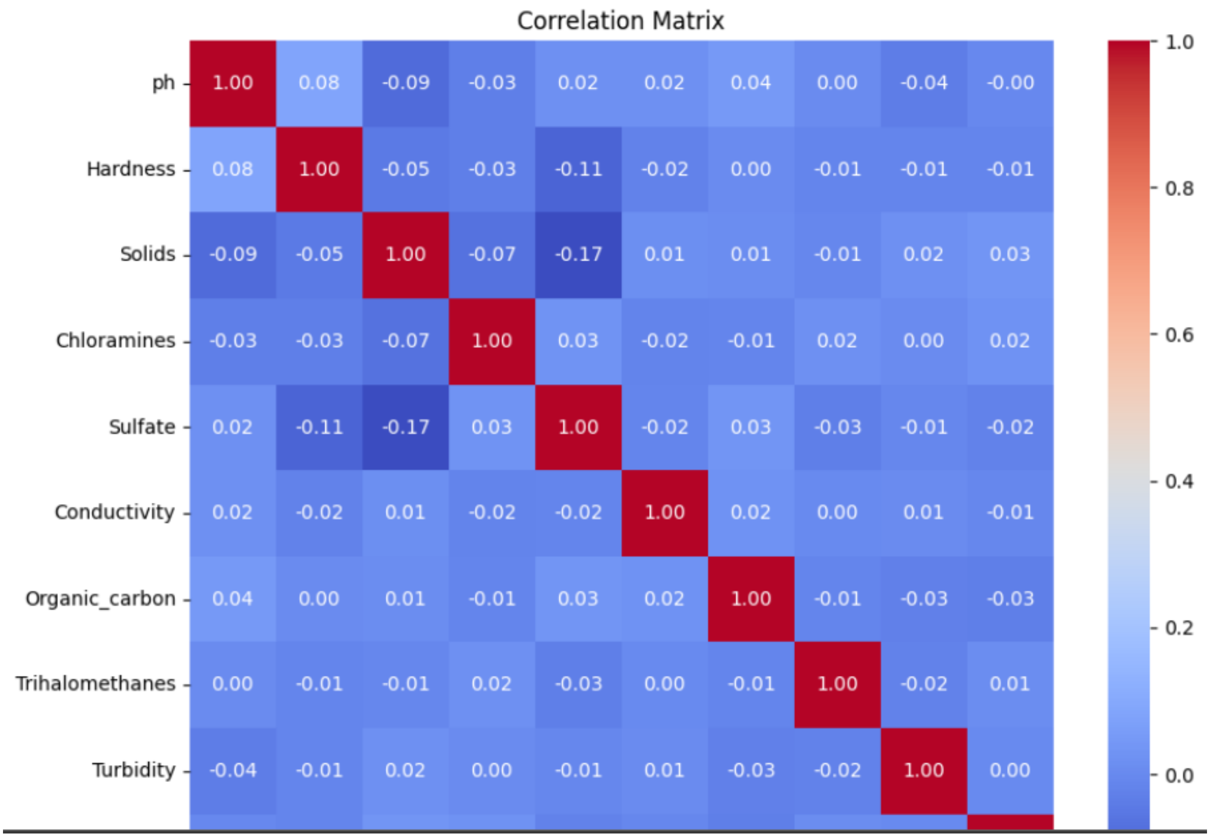
4.Data Cleaning:

Correlation matrices can highlight potential data errors.

CODE:

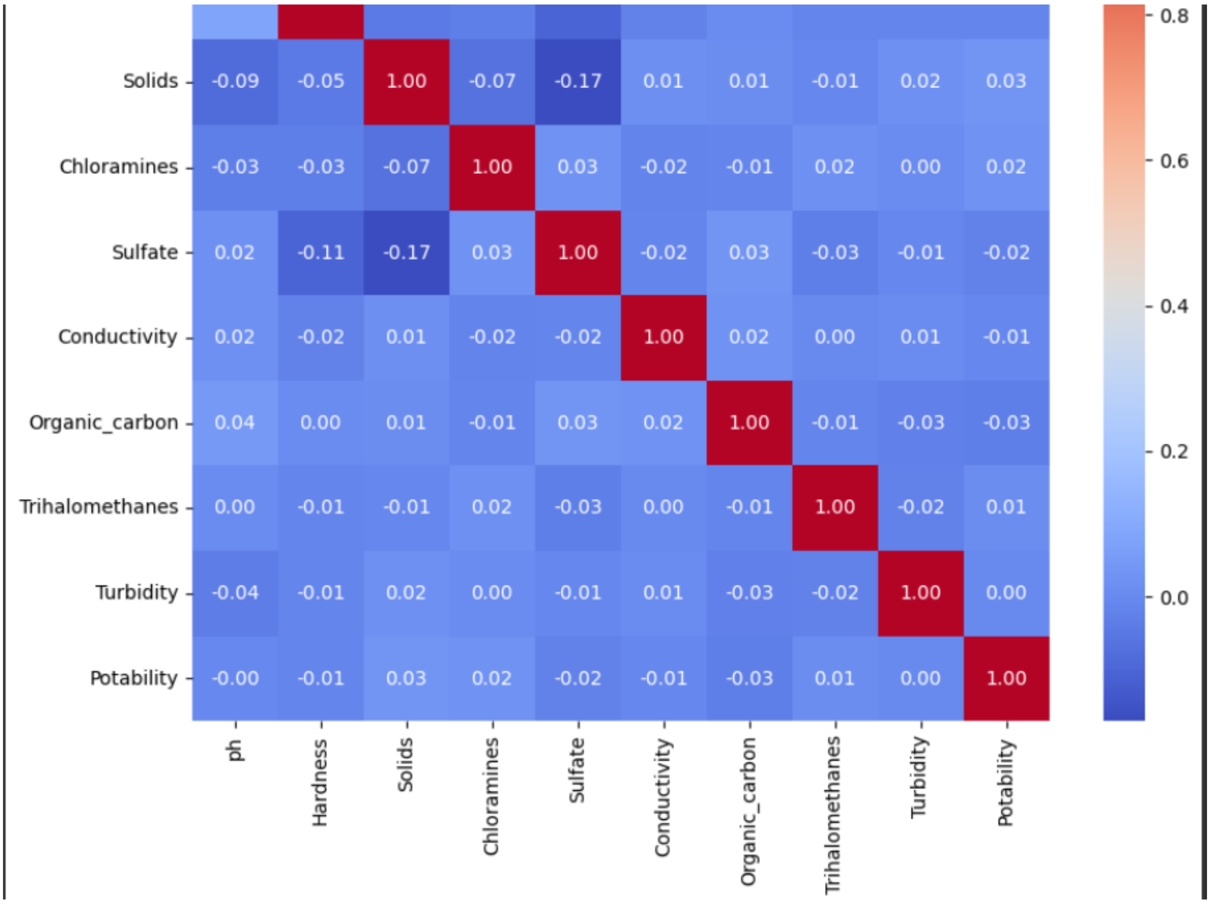
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
data = pd.read_csv("water.zip")
plt.figure(figsize=(10, 8))
correlation_matrix = data.corr() sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Correlation Matrix')
plt.show()
```

OUTPUT:



Correlation Matrix

ph	1.00	0.08	-0.09	-0.03	0.02	0.02	0.04	0.00	-0.04	-0.00
Hardness	0.08	1.00	-0.05	-0.03	-0.11	-0.02	0.00	-0.01	-0.01	-0.01
Solids	-0.09	-0.05	1.00	-0.07	-0.17	0.01	0.01	-0.01	0.02	0.03
Chloramines	-0.03	-0.03	-0.07	1.00	0.03	-0.02	-0.01	0.02	0.00	0.02
Sulfate	0.02	-0.11	-0.17	0.03	1.00	-0.02	0.03	-0.03	-0.01	-0.02
Conductivity	0.02	-0.02	0.01	-0.02	-0.02	1.00	0.02	0.00	0.01	-0.01
Organic_carbon	0.04	0.00	0.01	-0.01	0.03	0.02	1.00	-0.01	-0.03	-0.03
Trihalomethanes	0.00	-0.01	-0.01	0.02	-0.03	0.00	-0.01	1.00	-0.02	0.01
Turbidity	-0.04	-0.01	0.02	0.00	-0.01	0.01	-0.03	-0.02	1.00	0.00



Logistic Regression

Logistic Regression is a statistical method used for binary classification problems, where the outcome variable is categorical and has two classes. It predicts the probability of an instance belonging to a particular category. Unlike Linear Regression, which predicts a continuous outcome, Logistic Regression models the probability that the given input belongs to a specific category.

How Logistic Regression Works:

1. Sigmoid Function (Logistic Function):

Logistic Regression uses the sigmoid function to map any real-valued number to the range of 0 and 1. The sigmoid function is defined as:

[sigma(z) = \frac{1}{1 + e^{-z}\]

where

(z = b_0 + b_1x_1 + b_2x_2 + ldots + b_nx_n) (a linear combination of input features and their corresponding weights).

2. Probability Prediction:

The sigmoid function converts the linear combination of input features into a probability score between 0 and 1. If the probability is greater than or equal to 0.5, the instance is classified as the positive class; otherwise, it is classified as the negative class.

CODE:

```
 from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, classification_report
from sklearn.ensemble import RandomForestClassifier
import pandas as pd
data=pd.read_csv("water.zip")
data_cleaned = data.dropna(axis=1)
X = data_cleaned.drop(['Potability'], axis=1)
y = data_cleaned['Potability']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
logistic_regression_model = LogisticRegression(random_state=42)
logistic_regression_model.fit(X_train, y_train)
logistic_regression_predictions = logistic_regression_model.predict(X_test)
logistic_regression_accuracy = accuracy_score(y_test, logistic_regression_predictions)
print('Logistic Regression Accuracy:', logistic_regression_accuracy)
print('Logistic Regression Classification Report:')
print(classification_report(y_test, logistic_regression_predictions))# Initialize the Random Forest Classifier model
```

```
random_forest_model = RandomForestClassifier(random_state=42)
random_forest_model.fit(X_train, y_train)
random_forest_predictions = random_forest_model.predict(X_test)
random_forest_accuracy = accuracy_score(y_test,
random_forest_predictions)
print('Random Forest Classifier Accuracy:',
random_forest_accuracy)
print('Random Forest Classifier Classification Report:')
print(classification_report(y_test, random_forest_predictions))
```

OUTPUT:

```
 Logistic Regression Accuracy: 0.6280487804878049
Logistic Regression Classification Report:
              precision    recall  f1-score   support

           0       0.63      1.00      0.77       412
           1       0.00      0.00      0.00       244

    accuracy                           0.63       656
   macro avg       0.31      0.50      0.39       656
weighted avg       0.39      0.63      0.48       656


  _warn_prf(average, modifier, msg_start, len(result))
/usr/local/lib/python3.10/dist-
packages/sklearn/metrics/_classification.py:1344:
UndefinedMetricWarning: Precision and F-score are ill-defined and
being set to 0.0 in labels with no predicted samples. Use
`zero_division` parameter to control this behavior.
```

Random Forest Classifier

Random Forest is an ensemble learning method used for both classification and regression tasks. It creates a "forest" of decision trees, where each tree is trained on a random subset of the dataset and makes its own prediction. The final prediction in classification tasks is determined by a majority vote from all the individual trees.

How Random Forest Works

1. Bootstrapping:

Random Forest builds multiple decision trees through a process called bootstrapping. It creates random subsets of the original dataset with replacement. Each subset is used to train a decision tree.

2. Feature Selection:

For each split in the decision tree, a random subset of features is considered. This randomness ensures that the individual trees are diverse and not highly correlated.

3. Decision Trees:

Each subset of data is used to train a decision tree. Decision trees in a Random Forest can grow deep and capture complex patterns in the data.

CODE:

```
  from sklearn.ensemble importRandomForestClassifier
random_forest_model = RandomForestClassifier(random_state=42)
random_forest_model.fit(X_train, y_train)
random_forest_predictions = random_forest_model.predict(X_test)
random_forest_accuracy = accuracy_score(y_test,
random_forest_predictions)
print('Random Forest Classifier Accuracy:',
random_forest_accuracy)
print('Random Forest Classifier Classification Report:')
print(classification_report(y_test, random_forest_predictions))
```

OUTPUT:

```
Random Forest Classifier Accuracy: 0.6173780487804879
Random Forest Classifier Classification Report:
              precision    recall  f1-score   support

           0       0.65      0.83      0.73       412
           1       0.47      0.26      0.33       244

    accuracy                           0.62       656
   macro avg       0.56      0.54      0.53       656
weighted avg       0.59      0.62      0.58       656
```

Water quality analysis can help assess water quality and determine potability by providing insights into the following key parameters:

- **Physical parameters:**
  These include turbidity, color, odor, taste, and temperature. Turbidity indicates the presence of suspended particles in the water, which can interfere with disinfection and make it cloudy or hazy. Color, odor, and taste can be indicative of certain contaminants, such as algae blooms or industrial pollutants. Temperature can affect the growth of bacteria and other microorganisms.
- **Chemical parameters:**
  These include pH, dissolved oxygen, total dissolved solids, hardness, alkalinity, chloride, nitrates, sulfates, fluoride, and arsenic. pH affects the solubility of other chemicals in water and can also affect the growth of microorganisms. Dissolved oxygen is essential for aquatic life and for the proper functioning of water treatment systems. Total dissolved solids are a measure of the total amount of dissolved solids in water, including minerals, salts, and organic matter. Hardness is a measure of the amount of calcium and magnesium in water. Alkalinity is a measure of the

water's ability to neutralize acids. Chloride is a common contaminant in water that can come from agricultural runoff, industrial discharges, and saltwater intrusion. Nitrates are a byproduct of fertilizer use and can be harmful to infants. Sulfates are a common contaminant in water that can come from industrial discharges and septic systems. Fluoride is added to some drinking water supplies to help prevent tooth decay. Arsenic is a naturally occurring contaminant that can be harmful to human health.

- **Biological parameters:**
  These include coliform bacteria, E. coli, and other pathogenic microorganisms. Coliform bacteria are a group of bacteria that are present in the intestines of humans and animals. E. coli is a specific type of coliform bacteria that is often used as an indicator of fecal contamination. Pathogenic microorganisms can cause a variety of diseases, such as diarrhea, cholera, and typhoid fever.

By analyzing these parameters, water quality professionals can identify potential problems with water quality and take steps to improve it. For example, if the turbidity level is high, the water may need to be filtered to remove the suspended particles. If the pH level is too low or too high, it may need to be adjusted. If the nitrate level is too high, the water may need to be treated to remove the nitrates. And if coliform bacteria or other pathogenic microorganisms are present, the water may need to be disinfected.

Once water quality has been improved, it can be assessed for potability. Potability is defined as the suitability of water for human consumption. To be considered potable, water must meet certain standards for physical, chemical, and biological parameters. These standards are set by government agencies and vary from country to country.
If water meets all of the potability standards, it is considered safe to drink. However, even if water meets the potability standards, it may still have a taste or odor that some people find objectionable. In these cases, the water may need to be treated to remove the taste and odor.
Overall, water quality analysis is an essential tool for assessing water quality and determining potability. By understanding the insights from the analysis, water quality professionals can take steps to protect public health and ensure that everyone has access to safe drinking water.

GitHub repository link:

https://github.com/IT-Thamarai-demo/project-IBM.git