**Water Quality Analysis**



|  |  |
| --- | --- |
| Team Member 1 | Nandhini M 19886B0ED0C87919C114C9DDDC5C9560 |
| Team Member 2 | Hemasri R 785ED556BB0F30280674CEFE0CA02567 |

|  |  |
| --- | --- |
| Department | AI&DS – III |
| College Name | Adhi College Of Engineering And Technology |
| Domain Name | Data Analytics with Cognos |

**Problem Statement :**

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.

**Introduction:**

In this phase, we have prepared and analyzed the water quality dataset to ensure it is suitable for subsequent analysis and modeling. Our primary objectives were to handle missing values, address outliers, visualize parameter distributions, understand correlations between variables, and identify potential deviations from water quality standards.

**Design Thinking :**

In this phase, we will apply the principles of Design Thinking to solve the problem at hand. Design Thinking is an iterative process that focuses on understanding the needs of stakeholders and creating user-centered solutions. Here are the steps we will follow:

1. **Empathize:** Understand the perspectives and needs of stakeholders, including water quality regulators, public health officials, and the general public. Conduct interviews, surveys, or gather domain knowledge to gain insights into their concerns and requirements regarding water quality assessment.

2. **Define :** Clearly define the project objectives and scope. Establish the primary goal of assessing water potability and ensuring compliance with regulatory standards. Identify key success criteria and project milestones. Define the data sources and parameters that will be used for analysis.

3. **Ideate :** Brainstorm potential solutions for assessing water quality and determining potability. Explore various data sources and analytical methods that can be applied to this problem. Consider innovative approaches for data visualization and predictive modeling.

4. **Prototype :** Develop a prototype workflow that outlines the data collection, preprocessing, analysis, and visualization steps. Create a preliminary data visualization plan to present water quality data effectively. Outline a high-level plan for building a predictive model to assess water potability.

5. **Test :** Apply the prototype workflow to a small subset of the water quality data to test its functionality. Evaluate the initial data visualizations for clarity and effectiveness in conveying information. Assess the feasibility of the proposed predictive modeling approach.

6. **Iterate :** Gather feedback from stakeholders, data analysts, and project team members based on the prototype's performance. Refine the workflow, data visualization techniques, and modeling approach based on feedback and lessons learned during testing.

7. **Implementation :** Scale up the solution to analyze the entire water quality dataset. Implement the finalized data visualization and predictive modeling pipelines.

8. **Communication :** By Preparing comprehensive reports and presentations to communicate the project's findings, insights, and recommendations to stakeholders and the broader audience. Ensure that the analysis results are presented in an easily understandable and actionable format.

By following this Design Thinking approach, we aim to develop a robust solution for assessing water quality and determining water potability, with a strong focus on meeting the needs and expectations of stakeholders and regulatory standards.

**Research Anomaly Detection Techniques :**

* Conduct an extensive review of anomaly detection techniques and algorithms. Consider both traditional statistical methods and machine learning-based approaches.
* Evaluate the suitability of each technique for identifying unusual patterns in water quality parameters.

**Data preprocessing:**

Prepares the water quality dataset for anomaly Detection. This process involves handling the missing values , Outlier Detection and Treatment and data visualization in the dataset.

The code for the data preprocessing of the water quality dataset is given below ,

**Handling missing values:**

We are Identified and addressed missing values in the dataset.

Missing data are also observed in the below code:

import pandas as pd

import numpy as np

from google.colab import drive

drive.mount("/content/drive")

# Load the dataset

data = pd.read\_csv('/content/drive/MyDrive/water\_potability.csv')

# Check for missing values

missing\_values = data.isna().sum()

print(missing\_values)

# Remove rows with any missing values

data = data.dropna()

# Impute missing values with the mean of each column

data = data.fillna(data.mean())

# Interpolate missing values linearly

data = data.interpolate(method='linear')

# Check for missing values after handling

missing\_values = data.isna().sum()

print(missing\_values)

# Save the processed dataset

data.to\_csv('processed\_dataset.csv', index=False)

import pandas as pd

from google.colab import files

# Save the DataFrame with extracted features to a CSV file

data.to\_csv('processed\_dataset.csv', index=False)

files.download('processed\_dataset.csv')

**Outlier detection and treatment:**

We Detected outliers in the dataset, focusing on parameters that could significantly impact water quality assessment, such as pH, hardness.

The code for the outlier detection of the water quality dataset is given below,

from google.colab import drive

drive.mount('/content/drive')

Mounted at /content/drive

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

data = pd.read\_csv('/content/drive/MyDrive/processed\_dataset.csv')

# Create a box plot to visualize outliers

plt.figure(figsize=(8, 6))

sns.boxplot(x=data["ph"])

plt.title(f'Box Plot for {"ph"}')

plt.xlabel("ph")

plt.show()

plt.figure(figsize=(8, 6))

sns.boxplot(x=data["Hardness"])

plt.title(f'Box Plot for {"Hardness"}')

plt.xlabel("Hardness")

plt.show()

summary\_stats = data.describe()

print(summary\_stats)

[ ]

summary\_stats = data.describe()

print(summary\_stats)

account\_circle

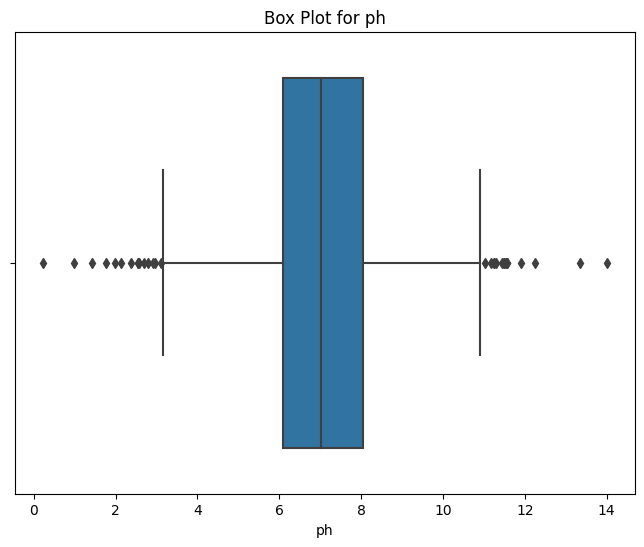
[ ]

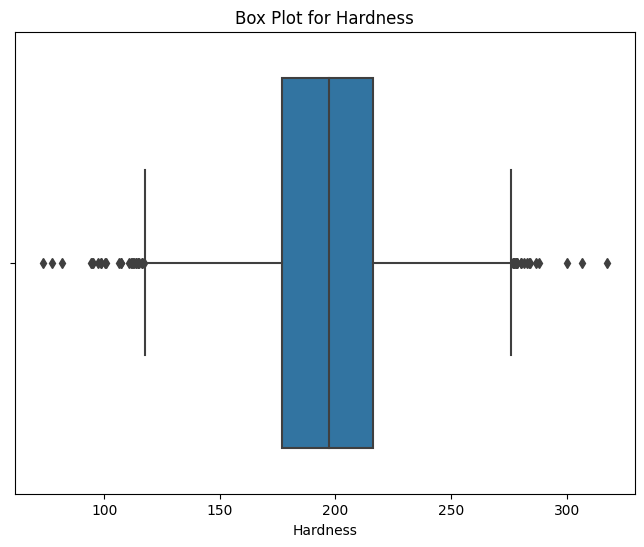
summary\_stats = data.describe()

print(summary\_stats)

account\_circle

**Output:**





**Data Visualization:**

We Visualized parameter distributions to understand the central tendencies and spread of the data. We have distribution using the box plot to find the ph and hardness.

**Code:**

plt.figure(figsize=(12, 6))

sns.set(style='whitegrid')

sns.histplot(data=data, x='p', bins=20, kde=True)

plt.title('pH Distribution')

plt.xlabel('pH')

plt.ylabel('Frequency')

plt.show()

# Explore correlations between parameters

correlation\_matrix = data.corr()

plt.figure(figsize=(10, 8))

sns.set(style='white')

sns.heatmap(correlation\_matrix, annot=True, cmap='coolwarm')

plt.title('Correlation Heatmap')

plt.show()

# Scatter plot to visualize relationships

plt.figure(figsize=(10, 6))

sns.scatterplot(data=data, x='Hardness', y='Chloramines', hue='Potability')

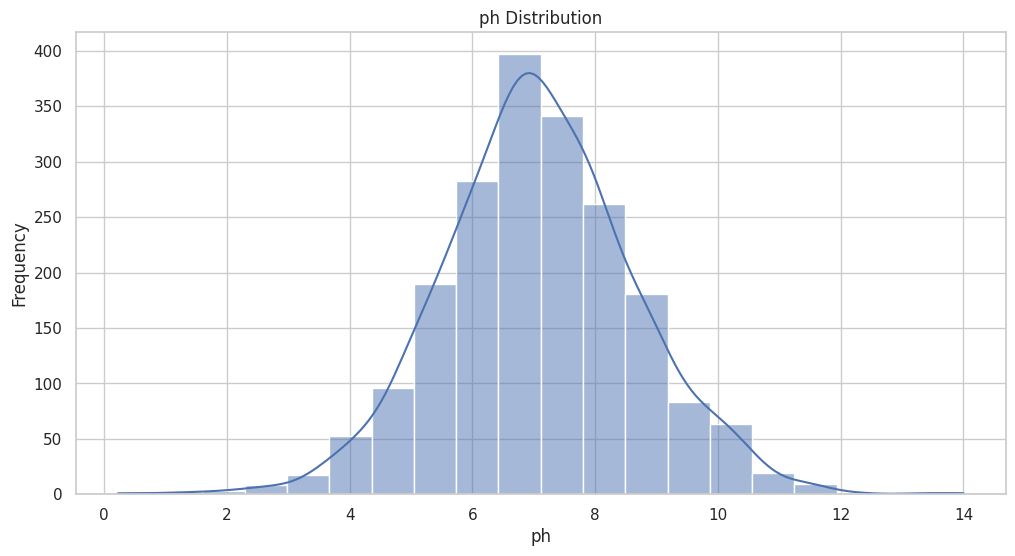
plt.title('Hardness vs. Chloramines')

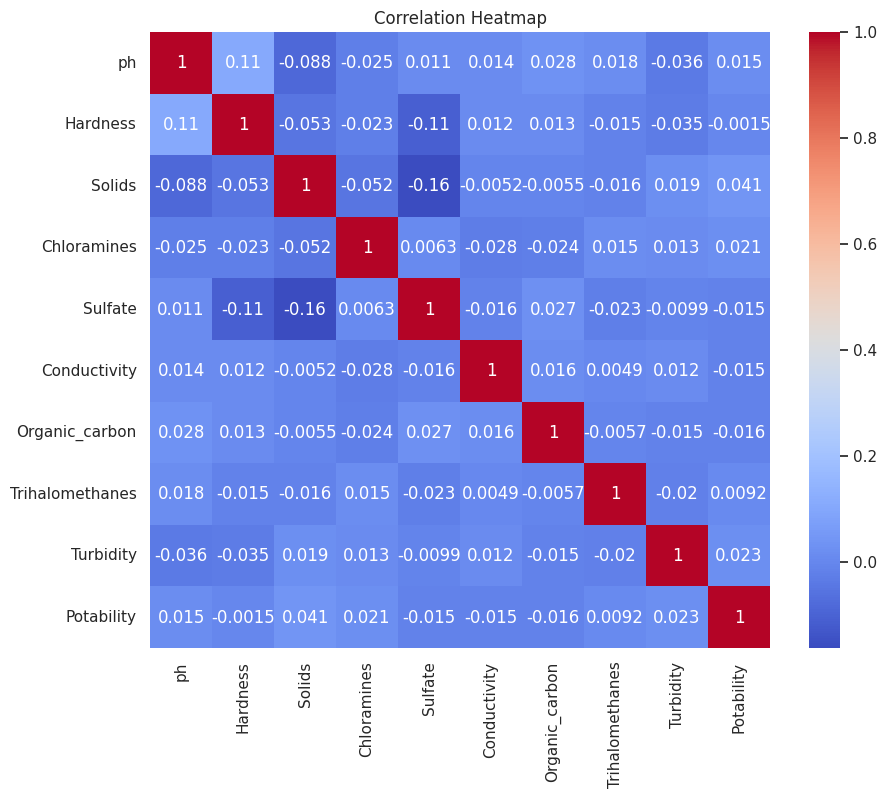
plt.xlabel('Hardness')

plt.ylabel('Chloramines')

plt.show()

**Output:**





**Predictive Model:**

A predictive model is a mathematical or computational representation of a real-world process or system that is used to make predictions about future events or outcomes based on historical data and patterns. Predictive models are widely used in various fields, including data science, machine learning, statistics, finance, healthcare, and many others. These models leverage historical data to make informed predictions, decisions, or recommendations.

For our model building we have used Decision tree. The code for the predictive model building is as follows,

import pandas as pd

from sklearn.model\_selection import train\_test\_split

from sklearn.tree import DecisionTreeClassifier

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

# Load your preprocessed dataset

data = pd.read\_csv('/content/drive/MyDrive/processed\_dataset.csv')

# Define your features (independent variables) and target (dependent variable)

# Adjust these variable names based on your dataset's columns

X = data[['ph', 'Hardness','Solids', 'Chloramines', 'Sulfate', 'Conductivity', 'Turbidity', 'Organic\_carbon', 'Trihalomethanes']]

y = data['Potability']

# Split the 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 Decision Tree classifier (you can replace this with another classifier)

clf = DecisionTreeClassifier(random\_state=42)

# Train the classifier on the training data

clf.fit(X\_train, y\_train)

# Make predictions on the test data

y\_pred = clf.predict(X\_test)

# Evaluate the model's performance

accuracy = accuracy\_score(y\_test, y\_pred)

confusion = confusion\_matrix(y\_test, y\_pred)

report = classification\_report(y\_test, y\_pred)

# Print the model's performance metrics

print(f'Accuracy: {accuracy}')

print('Confusion Matrix:')

print(confusion)

print('Classification Report:')

print(report)

**Output:**

Accuracy: 0.6327543424317618

Confusion Matrix:

[[165 66]

[ 82 90]]

Classification Report:

precision recall f1-score support

0 0.67 0.71 0.69 231

1 0.58 0.52 0.55 172

accuracy 0.63 403

macro avg 0.62 0.62 0.62 403

weighted avg 0.63 0.63 0.63 403

Insights derived from water quality analysis play a vital role in assessing the safety and potability of water. These insights are derived from various parameters and their correlations, enabling a comprehensive evaluation of water quality. Here's how the insights obtained from the analysis can aid in assessing water quality and determining potability:

**Parameter Assessment:**

**pH Levels:** The pH level indicates the acidity or alkalinity of water. Extremes in pH levels can render water unsuitable for consumption. Insights into pH levels help in assessing the water's corrosiveness and suitability for consumption.

**Hardness:** High mineral content, especially calcium and magnesium, can lead to water hardness. While it doesn't directly affect health, it can cause scaling in pipes and affect soap lathering. Insights into hardness levels are crucial for assessing water usage and the potential impact on infrastructure.

**Chromium, Sulfates, Conductivity, Turbidity, Organic Carbon, Trihalomethane:** These parameters represent impurities, pollutants, and byproducts that can affect water quality. Analyzing these factors aids in identifying contaminants and potential health risks associated with water consumption.

**Regulatory Standards and Guidelines:**

Insights from the analysis can be compared against regulatory standards set by health and environmental agencies. Parameters exceeding the recommended thresholds signify deviations from the defined safety limits, raising concerns about water quality.

**Correlation Analysis:**

Correlations between Parameters: Understanding correlations between different parameters can reveal hidden relationships affecting water quality. For instance, a high correlation between organic carbon and trihalomethanes might indicate potential disinfection byproducts.

**Deviations and Anomalies:**

Identification of Outliers or Anomalies: Detection of unusual patterns or outliers in the parameters can signify potential issues or deviations from standard water quality. It might indicate sudden spikes or drops in parameters, suggesting a need for closer inspection or immediate action.

**Impact on Health and Potability:**

Interpreting the impact of water quality parameters on human health is crucial. High levels of certain chemicals or impurities might lead to health concerns, influencing the determination of water potability.

**Decision-Making and Remedial Actions:**

The insights gathered from the analysis provide a basis for decision-making. If deviations from standards are identified, actions can be taken, such as implementing treatment processes, changing water sources, or issuing advisories to ensure safe water consumption.

**Continuous Monitoring and Improvement:**

These insights also pave the way for continuous monitoring and improvement strategies. Repeated analyses and data-driven approaches aid in maintaining and enhancing water quality standards over time.

In summary, the insights gained from water quality analysis are pivotal in understanding the safety and potability of water. They serve as a foundation for decision-making, remedial actions, and strategies to ensure safe and high-quality water for consumption and various purposes. Regular monitoring and analysis enable a proactive approach towards maintaining and improving water quality standards.