**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.

**Understanding :**

Access to safe and clean drinking water is a fundamental human right, and it plays a pivotal role in maintaining public health and well-being. However, water quality can vary significantly due to natural processes and human activities. To address this, it's imperative to perform a comprehensive analysis of water quality parameters, identify potential issues or deviations from regulatory standards, and determine water potability based on these parameters.

**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 in the dataset.

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

|  |
| --- |
| 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)  # Download the CSV file  files.download('processed\_dataset.csv') |

**Feature Engineering :**

This selects the relevant features for anomaly detection. Focus on those parameters that are most indicative of water quality deviations.

Here we have chosen ph and Hardness of the water as the parameters for the features of anomaly detection.

**Anomaly Detection Model Selection :**

Here we choose one anomaly detection technique that is well suited to the water quality dataset.

The common approaches include :

* Isolation Forest
* One-Class SVM
* Autoencoders
* Statistical Methods

We have chosen Isolation Forest for Anomaly Detection.

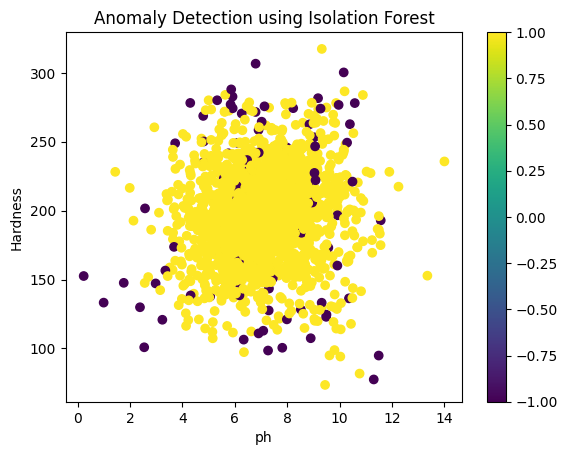
**Anomaly Detection in Water Quality Dataset :**

* This trains the selected anomaly detection model on a portion of the water quality dataset that is the representative of normal , potable water quality.
* Determines appropriate hyperparameter to optimize model performance.
* Here we have determined n\_estimators and contaminations as hyperparameters.
* Applies the trained anomaly detection model to the entire water quality dataset.
* Identifies instances where water quality parameters deviate significantly from normal patterns.
* Designs and creates visualization to display identified anomalies clearly.
* This visualization is done using Scatter plot.

The code for Anomaly detection using Isolation Forest is as follows ,

|  |
| --- |
| # Import necessary libraries  import numpy as np  import pandas as pd  from sklearn.ensemble import IsolationForest  # Load your water quality dataset  data = pd.read\_csv('/content/processed\_dataset.csv')  features = ['ph', 'Hardness']  # Create a subset of the data with selected features  X = data[features]  # Initialize the Isolation Forest model  # Adjust the hyperparameters (e.g., n\_estimators, contamination) as needed  model = IsolationForest(n\_estimators=100, contamination=0.05, random\_state=42)  # Fit the model to the data  model.fit(X)  # Predict anomalies (1 for inliers, -1 for outliers)  anomalies = model.predict(X)  # Add the anomaly predictions to the original dataset  data['anomaly'] = anomalies  # Visualize the anomalies  # You can plot the anomalies using libraries like matplotlib or seaborn  import matplotlib.pyplot as plt  # For example, if you want to create a scatter plot of 'feature1' vs. 'feature2'  plt.scatter(data['ph'], data['Hardness'], c=data['anomaly'], cmap='viridis')  plt.title('Anomaly Detection using Isolation Forest')  plt.xlabel('ph')  plt.ylabel('Hardness')  plt.colorbar()  plt.show() |

The result of the above is as follows,



**Conclusion :**

In conclusion, anomaly detection for water quality analysis plays a vital role in ensuring the safety and sustainability of our water resources. This technology leverages advanced data analytics and machine learning algorithms to identify unusual patterns or deviations in water quality parameters.