Water Quality Analysis

Phase 2

Innovation

Submitted by:

Mohamed niyas .S

Prabhaharan .R

Kalaikmani .K

Gokul .R

Daniel .D

Executive Summary:

The "Innovative Water Quality Analysis with Anomaly Detection Techniques" project aims to revolutionize the field of water quality assessment by incorporating cutting-edge anomaly detection methods. This project is driven by the imperative need to ensure safe and reliable access to clean water resources, especially in the face of evolving environmental challenges.

Objectives:

1. Enhance Traditional Water Quality Assessment Methods: Combine conventional water quality parameters with advanced anomaly detection techniques to provide a more comprehensive understanding of water conditions.

2. Early Detection of Contamination Events: Develop algorithms that can identify unusual patterns in water quality data, potentially signaling contamination events in real-time.

3. Improve Resource Allocation: Enable more efficient allocation of resources for water quality monitoring and management by prioritizing areas of concern based on anomaly severity.

4. Enhance Regulatory Compliance: Facilitate compliance with water quality regulations and standards by providing more accurate and timely data.

Project Scope:

1. Data Collection and Integration:

About Dataset

Admittance to safe drinking-water is fundamental for wellbeing, an essential common freedom and a part of viable strategy for wellbeing insurance. This is significant as a wellbeing and improvement issue at a public, provincial and nearby level. In certain locales, it has been demonstrated the way that interests in water supply and sterilization can yield a net financial advantage, since the decreases in unfavorable wellbeing impacts and medical care costs offset the expenses of undertaking the mediations.

1. pH esteem:

PH is a significant boundary in assessing the corrosive base equilibrium of water. It is likewise the sign of acidic or antacid state of water status. WHO has suggested greatest reasonable restriction of pH from 6.5 to 8.5. The ongoing examination ranges were 6.52-6.83 which are in the scope of WHO principles.

2. Hardness:

Hardness is primarily brought about by calcium and magnesium salts. These salts are disintegrated from geologic stores through which water voyages. The period of time water is in touch with hardness creating material decides how much hardness there is in crude water. Hardness was initially characterized as the limit of water to encourage cleanser brought about by Calcium and Magnesium.

3. Solids (All out broke down solids - TDS):

Water can disintegrate many inorganic and a few natural minerals or salts, for example, potassium, calcium, sodium, bicarbonates, chlorides, magnesium, sulfates and so on. These minerals created un-needed taste and weakened variety in appearance of water. This is the significant boundary for the utilization of water. The water with high TDS esteem shows that water is profoundly mineralized. Helpful breaking point for TDS is 500 mg/l and most extreme cutoff is 1000 mg/l which endorsed for drinking reason.

4. Chloramines:

Chlorine and chloramine are the significant sanitizers utilized openly water frameworks. Chloramines are most generally shaped when alkali is added to chlorine to treat drinking

water. Chlorine levels up to 4 milligrams for each liter (mg/L or 4 sections for every million (ppm)) are viewed as protected in drinking water.

5. Sulfate:

Sulfates are normally happening substances that are tracked down in minerals, soil, and shakes. They are available in encompassing air, groundwater, plants, and food. The key business utilization of sulfate is in the substance business. Sulfate fixation in seawater is around 2,700 milligrams for every liter (mg/L). It goes from 3 to 30 mg/L in most freshwater supplies, albeit a lot higher focuses (1000 mg/L) are tracked down in a few geographic areas.

6. Conductivity:

Unadulterated water is certainly not a decent conveyor of electric flow Rather's a decent encasing. Expansion in particles fixation upgrades the electrical conductivity of water. By and large, how much disintegrated solids in water decides the electrical conductivity. Electrical conductivity (EC) really gauges the ionic course of an answer that empowers it to send flow. As indicated by WHO norms, EC worth shouldn't surpassed 400 µS/cm.

7. Organic carbon:

Absolute Natural Carbon (TOC) in source waters comes from rotting normal natural matter (NOM) as well as engineered sources. TOC is a proportion of the aggregate sum of carbon in natural mixtures in unadulterated water. As per 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 synthetic substances which might be found in water treated with chlorine. The convergence of THMs in drinking water shifts as per the degree of natural material in the water, how much chlorine expected to treat the water, and the temperature of the water that is being dealt with. THM levels up to 80 ppm is viewed as protected in drinking water.

9. Turbidity:

The turbidity of water relies upon the amount of strong matter present in the suspended state. It is a proportion of light transmitting properties of water and the test is utilized to show the nature of waste release as for colloidal matter. The mean turbidity esteem acquired for Wondo Genet Grounds (0.98 NTU) is lower than the WHO suggested worth of 5.00 NTU.

10. Potability:

Shows in the event that water is alright for human utilization where 1 method Consumable and 0 methods Not consumable.

dataset : water_potability.csv

Data Preprocessing:

Data =pd.read-csv('file path'):

Which is a function it read a data set and ' \\ ' is required in windows.

Data.head():

Which is used to print the first five in the dataset.

Data.info():

Which is a function it shows total entries, columns, null values, datatypes and memory usage.

Data.describe():

Which is a function to get summary statistics about dataset.

Which show the count, standard deviation, presential and maximum and the describe() function only work on numeric column.

Data.isna().sum():

Which is a function to get total number of null values.

NULL VALUE:

Null value is nothing but the cells in the dataset do not have a value or present without any value.

If the null value is present or available in any dataset there are many functions to remove it.

Functions to remove null value:

1. Interpolation ()

2. Fillna()

3. dropna()

And using mean, median and mode values.

In our water quality analysis project we use fillna() and mean function.

Fillna():

Which is a function used to fill the empty cells or null value in the dataset.

2. Anomaly Detection Techniques:

Machine learning-based approaches

Explanation of Isolation Forest:

Isolation Forest is an ensemble learning method for anomaly detection. It works by constructing multiple decision trees, each trained on a random subset of the data. The key idea is that anomalies are likely to be isolated in fewer partitions (or "isolation" occurs quickly), whereas normal points require more partitions to isolate.

In the code, `contamination` is a crucial parameter. It sets the proportion of outliers in the data, which is an assumption you need to make based on your knowledge of the problem. This parameter can significantly impact the performance of the algorithm.

Keep in mind that this is a simple example. In practice, you would need to replace the example data with your actual water quality parameters and ensure that the features are

appropriately selected and scaled. Additionally, you may want to experiment with other anomaly detection techniques based on the specific characteristics of your data.

Sample program with example:

```
# Import necessary libraries
from sklearn.ensemble import IsolationForest
import numpy as np

# Generate example data (replace with actual water quality data)
# Assume 'data' is a 2D array where each row represents a data point and columns represent features.
data = np.random.randn(100, 4)  # Example data with 4 features and 100 data points
anomalies_ratio = 0.05  # Adjust this based on expected anomaly ratio in your data

# Train Isolation Forest model
model = IsolationForest(contamination=anomalies_ratio, random_state=42)
model.fit(data)

# Predict outliers
outliers = model.predict(data)

# Anomaly detection result: -1 indicates an anomaly, 1 indicates a normal point
anomaly_indices = np.where(outliers == -1)

# Print indices of detected anomalies
print("Detected anomalies at indices:", anomaly_indices[0]
```

Explanation:

1. Import Libraries:

We import the necessary libraries, including `IsolationForest` from scikit-learn.

2. Generate Example Data:

We generate example data (`data`) to simulate water quality measurements. In a real-world scenario, you would replace this with your actual water quality data.

3. Set Anomalies Ratio:

`anomalies_ratio` is a hyperparameter representing the expected proportion of anomalies in your data. Adjust this based on your domain knowledge.

4. Train Isolation Forest Model:

We create an instance of the Isolation Forest model, specifying the expected contamination ratio (anomalies_ratio). The model is then trained on the data.

5. Predict Outliers:

The `predict` method is used to label data points as normal (1) or anomalies (-1).

6. Anomaly Detection Result:

We identify the indices of detected anomalies by checking where the predictions are -1.

7. Print Detected Anomalies:

Finally, we print the indices of detected anomalies.

Feature Engineering:

Creating a feature that represents the product of pH and Hardness is a straightforward process. This new feature will capture the interaction between these two parameters. Here are the steps:

- **Obtain the Data**:
- Ensure you have the values for both pH and Hardness.
- **Calculate the Product**:
- Multiply the value of pH by the value of Hardness.
- Mathematically, the new feature (F) can be calculated as:
- *F*=*pH*×*Hardness*F=pH×Hardness
- For example, if pH = 7.5 and Hardness = 150, the new feature would be:
- *F*=*7.5*×*150*=*1125*F=7.5×150=1125
- So, the new feature capturing the interaction between pH and Hardness is 1125.
- **Incorporate the New Feature**:
- Include this new feature, which represents the product of pH and Hardness, as one of the input features in your dataset for further analysis or modeling.

Ratio:

- To calculate the ratio of Chloramines to Solids, you simply divide the value of Chloramines by the value of Solids. This will give you a single numerical value representing the proportion of Chloramines relative to Solids.
- Mathematically, the ratio (R) can be calculated as:
- *R*=*ChloraminesSolids*R=SolidsChloramines
- Here's a step-by-step guide on how to calculate the ratio:
- **Obtain the Data**:
- Ensure you have the values for Chloramines and Solids.
- **Perform the Division**:
- Divide the value of Chloramines by the value of Solids.
- For example, if you have Chloramines = 50 and Solids = 200, the ratio would be:
- *R*=*50200*=*0.25*R=20050 =0.25
- So, the Chloramines to Solids ratio in this case is 0.25.

3. Algorithm Development:

Model Training and Validation:

Utilize a combination of historical data and simulated anomaly scenarios for model development

Employ cross-validation techniques to assess model performance

Real-time Implementation:

Design an architecture for continuous monitoring and anomaly detection

Establish protocols for immediate response to detected anomalies

4. Visualization and Reporting:

Data visualization is a crucial component of water quality analysis, especially when incorporating anomaly detection techniques. It helps in understanding trends, identifying patterns, and visualizing anomalies effectively. Below are some common data visualization techniques for water quality analysis with a focus on anomaly detection:

1. Time Series Plots:

Description: Time series plots show how water quality parameters change over time. This is important for detecting seasonal trends and anomalies.

Example: Plotting parameters like pH, turbidity, and dissolved oxygen levels over time.

2. Box Plots:

Description: Box plots provide a summary of the distribution of a dataset, allowing for the identification of outliers and anomalies.

Example: Use box plots to visualize the distribution of parameters across different monitoring stations or over different seasons.

3. Scatter Plots:

Description: Scatter plots help in visualizing relationships between two or more variables. This can be useful for identifying correlations and potential anomalies.

Example: Scatter plot of temperature vs. dissolved oxygen levels to identify any unusual patterns.

4. Heatmaps:

Description: Heatmaps can be used to visualize spatial variations in water quality parameters. This is particularly relevant when monitoring multiple locations.

Example: Create a heatmap showing variations in pH levels across different sampling sites.

5. Histograms:

Description: Histograms provide a visual representation of the distribution of a single variable. They can help in identifying unusual data distributions.

Example: Create histograms for parameters like pH, turbidity, and conductivity to understand their distributions.

6. Control Charts:

Description: Control charts are useful for detecting unusual patterns or shifts in a process. They are widely used in quality control and can be adapted for water quality analysis.

Example: Use control charts to monitor the mean and variability of a parameter over time.

7. 3D Plots:

Description: 3D plots can be employed when you have multiple variables and want to visualize their relationships in three-dimensional space.

Example: Use a 3D scatter plot to visualize the relationship between temperature, pH, and dissolved oxygen levels.

8. Anomaly Detection Plots:

Description: Create specialized plots to highlight detected anomalies. This could involve overlaying anomaly points on top of regular data points for emphasis.

Example: Highlight anomalies detected using an algorithm (e.g., Isolation Forest) on a scatter plot.

9. Geospatial Visualizations:

Description: Utilize maps to display water quality parameters across different locations. This can be valuable for understanding spatial patterns and identifying outliers.

Example: Use GIS software to create a map displaying various water quality parameters at different monitoring stations.

10. Interactive Dashboards:

Description: Build interactive dashboards that allow users to explore the data dynamically. This can include filter options, tooltips, and multiple linked visualizations.

Example: Create a web-based dashboard where users can select parameters, time ranges, and locations to visualize data dynamically.

Dashboard Development:

Create an intuitive user interface for visualizing water quality data and detected anomalies

Provide real-time updates and historical trend analysis

Alerting System:

Establish an automated alerting system for notifying relevant stakeholders in case of severe anomalies

5. Evaluation and Testing:

Performance Metrics:

Precision, Recall, F1-score for anomaly detection

Comparison against existing water quality assessment methods

Pilot Testing:

Conduct a pilot study in a selected water body to validate the effectiveness of the developed system

Conclusion:

The "Innovative Water Quality Analysis with Anomaly Detection Techniques" project presents a groundbreaking opportunity to significantly improve the accuracy and timeliness of water quality assessments. By integrating anomaly detection techniques, we aim to enhance the resilience of water supply systems and safeguard public health. This project holds the potential to become a milestone in water quality management practices and contribute to sustainable environmental stewardship.