**PHASE-2**

**PROJECT 10: WATER QUALITY ANALYSIS**

**TEAM MEMBER NAME:** Krithiga M

**Objective:**

To explore anomaly detection techniques to identify unusual patterns in water quality parameters.

**Introduction:**

Water quality monitoring is of paramount importance to ensure the safety and sustainability of our water sources. Detecting unusual patterns or anomalies in water quality parameters is a critical aspect of this monitoring process. Anomalies can indicate contamination, equipment malfunctions, or other issues that need immediate attention. To address this challenge, a range of advanced anomaly detection techniques have been developed, each with its own unique approach and capabilities.By exploring and implementing these anomaly detection techniques, water quality stakeholders can enhance their ability to monitor, detect, and respond to anomalies in real-time, ensuring the delivery of safe and clean water to communities while protecting the environment.

**Anamoly detection techniques:**

In the sections that follow, we will provide a detailed explanation of each technique, their advantages, and how they can be effectively used to maintain the integrity of our water sources.

**1. Statistical Methods:**

- Z-Score: The Z-Score measures how many standard deviations a data point is away from the mean. Data points with Z-Scores significantly higher or lower than the mean are considered anomalies.

- Modified Z-Score: Similar to the Z-Score, but it's a robust version that is less sensitive to outliers.

- Grubbs' Test: Grubbs' Test is used to detect outliers in a univariate dataset by identifying the maximum absolute deviation from the mean.

- IQR (Interquartile Range) Method: This method defines outliers as data points falling below Q1 - 1.5 \* IQR or above Q3 + 1.5 \* IQR, where Q1 and Q3 are the first and third quartiles. It's robust to outliers as it relies on the quartiles.

**2. Machine Learning Methods:**

- Isolation Forest: The Isolation Forest algorithm isolates anomalies by constructing an ensemble of decision trees. Anomalies are data points that require fewer splits in the trees to be isolated from the rest.

- One-Class SVM (Support Vector Machine): This method creates a boundary (hyperplane) around normal data points. Data points that fall outside this boundary are considered anomalies.

- Autoencoders: Autoencoders are neural networks designed to learn a compressed representation of the input data. Anomalies are detected when the reconstruction error is significantly higher for certain data points.

- K-Nearest Neighbors (KNN): KNN calculates the distance of a data point to its k-nearest neighbors. Data points with distant neighbors are considered anomalies.

**3. Time-Series Techniques:**

- Seasonal Decomposition: Time series data often contains seasonality, trend, and residual components. Anomalies can be detected in the residual component when it deviates significantly from the expected pattern.

- Exponential Smoothing: Exponential smoothing is a time series forecasting method. It detects anomalies when the actual values deviate significantly from the forecasts made using this technique.

- ARIMA (AutoRegressive Integrated Moving Average): ARIMA models are used to model and forecast time series data. Anomalies are detected when the actual data deviates from the predicted values significantly.

**4. Clustering:**

- K-Means: K-Means is a clustering algorithm that groups data points into clusters. Data points that do not belong to any cluster or are in small clusters can be considered anomalies.

- DBSCAN (Density-Based Spatial Clustering of Applications with Noise): DBSCAN identifies anomalies as data points that do not belong to any cluster or are in low-density clusters.

- LOF (Local Outlier Factor): LOF computes the local density of data points. Those with significantly lower local densities are flagged as anomalies.

**5. Deep Learning:**

- Variational Autoencoders (VAEs): VAEs are a variant of autoencoders that capture the probabilistic distribution of normal data. Anomalies are detected based on deviations from this distribution.

- LSTM (Long Short-Term Memory) Networks: LSTMs are recurrent neural networks used for modeling time series data. Anomalies are detected when the prediction errors are significant.

**6. Rule-Based Methods:**

- In this approach, we define domain-specific rules or thresholds for each water quality parameter. Data points that violate these rules are considered anomalies.

**Conclusion:**

The choice of our project water quality analysis is the Machine Learning technique that depends on the specific dataset, the nature of anomalies to detect, and the computational resources available. It's often a good practice to experiment with multiple techniques to determine which one performs best for a particular use case. Additionally, ensemble methods can be used to combine the results of multiple techniques for improved accuracy in anomaly detection.