

**MUTHAYAMMAL COLLEGE OF ENGINEERING**

(Approved by AICTE, New Delhi and Affiliated to Anna University)

Rasipuram - 637 408, Namakkal Dist., Tamil Nadu.

**WATER QUALITY ANALYSIS**

From Department of

**B.TECH(Artificial Intelligence And Data Science)**

**Phase 2 Submission Document**

**BY:**

**MURALIDHARAN N P**

III-YEAR(AI&DS)

* Exploring anomaly detection techniques to identify unusual pattern**s** in Water Quality Parameters:

**Introduction**:

* Traditional monitoring and assessment of water quality has relied on in situ measurements, sample collection, and lab analyses to measure indicators related to the physical, chemical, and biological properties of the waterbody.

* More recently, federal agencies like the United States Geological Survey (USGS) and United States Environmental Protection Agency (USEPA), along with other partners are developing networks of sensors that provide continuous water quality data in near real-time (Pellerin et al. 2016).
* Remote sensing techniques allow for water quality and ecosystem health to be inferred via indicator variables such as bluegreen algae (BGA), chlorophyll (Chl), fluorescent dissolvedorganic matter (fDOM). Turbidity that represent the biological and hydrological conditions of the waterbody.

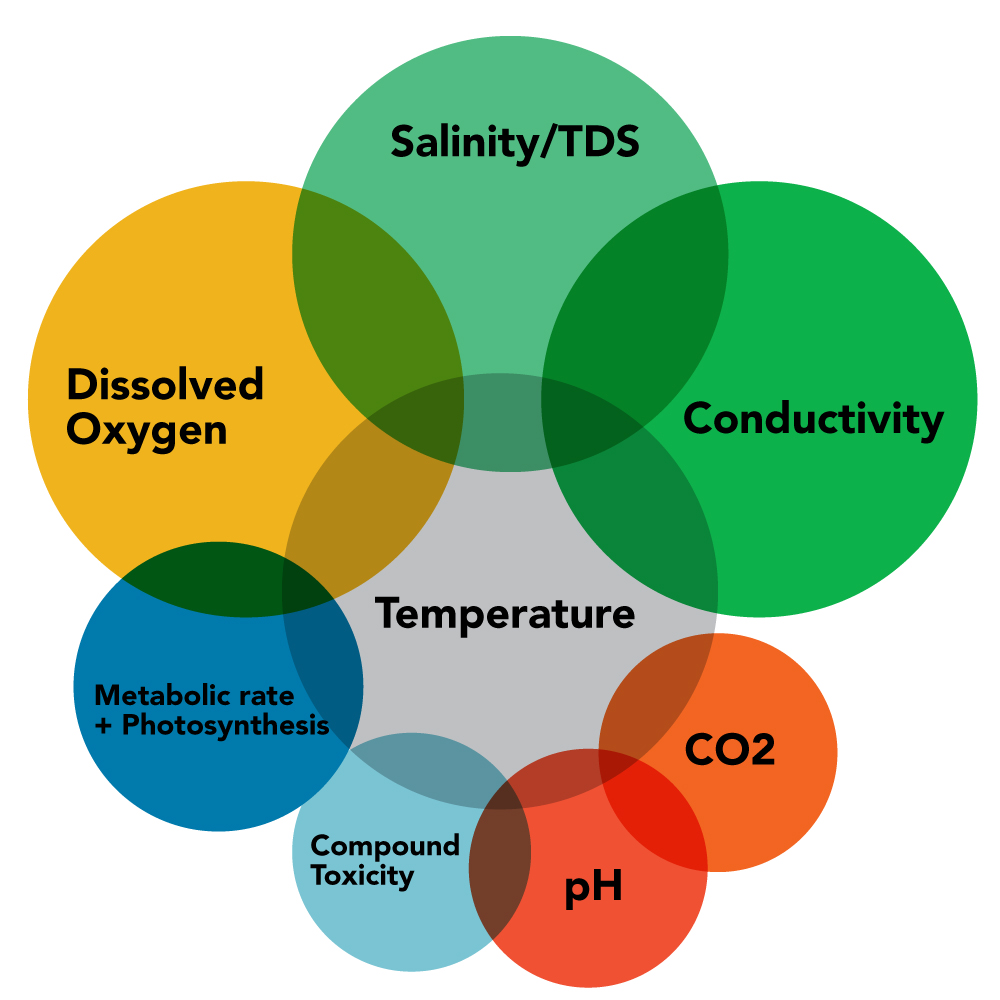
**Content for Project Phase 2:**

* Consider exploring anomaly detection techniques to identify unusual patterns in water quality parameters.

**DataSourse:**

* The data sourse for to detect and identify the unusual patterns in water quality parameters like pH,Temperature,Oxygen,Conductivity,ORP and Turbidity.

**Water Quality Parameters:**



1. **Physical Parameters:**

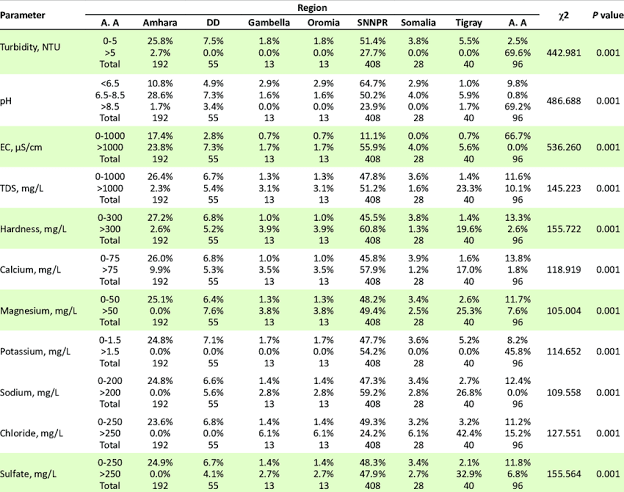
* Temperature: The measurement of water temperature is important because it affects the solubility of gases and the metabolic rates of aquatic organisms.
* Turbidity: Turbidity measures the cloudiness or haziness of water caused by suspended particles. It can affect light penetration and aquatic habitat quality.

1. **Chemical Parameters:**

* pH: pH measures the acidity or alkalinity of water. It is important for aquatic life and can influence the effectiveness of chemical treatments.
* Dissolved Oxygen (DO): DO levels indicate the amount of oxygen available in the water, which is crucial for fish and other aquatic organisms.
* Conductivity: Conductivity measures the water's ability to conduct an electrical current, which is related to the concentration of dissolved ions, such as salts and minerals.
* Total Dissolved Solids (TDS): TDS measures the total concentration of dissolved substances in water, including minerals, salts, and some organic matter.

1. **Biological Parameters:**

* Biological Oxygen Demand (BOD): BOD measures the amount of oxygen consumed by microorganisms in water while breaking down organic matter. It indicates the organic pollution level.



1. **Data Collection and Preprocessing:**

* Gather historical and real-time data on water quality parameters such as pH, turbidity, dissolved oxygen, temperature, and various contaminants.
* Clean the data by handling missing values, outliers, and ensuring data consistency.

1. **Feature Engineering:**

* Select relevant features that are indicative of water quality.
* Consider aggregating data over time intervals (e.g., hourly, daily) to reduce noise and highlight trends.

1. **Visualization and Exploratory Data Analysis**

**(EDA):**

* Create visualizations (e.g., time series plots, scatter plots, histograms) to gain insights into the data.
* Look for patterns, seasonality, and trends that may help in identifying anomalies.

1. **Choosing Anomaly Detection Techniques:**

* Select appropriate anomaly detection methods based on the nature of your data and problem.
* Some common techniques include:

Statistical methods (e.g., Z-score, Modified Z-score)

* Machine learning algorithms (e.g., Isolation Forest, One-Class SVM)
* Time series analysis (e.g., Seasonal Decomposition of Time Series, Prophet)
* Deep learning approaches (e.g., Autoencoders)
* Domain-specific methods (e.g., control charts for continuous monitoring)

1. **Training and Testing:**

* Split your data into training and testing sets to evaluate the performance of your chosen anomaly detection techniques.
* Train the selected models on the training data and fine-tune hyperparameters if necessary.

**6.Model Evaluation:**

* Use appropriate evaluation metrics such as precision, recall, F1-score, and area under the ROC curve (AUC-ROC) to assess the performance of your anomaly detection models.
* Adjust the models' threshold values to achieve the desired trade-off between false positives and false negatives.

**7.Real-time Monitoring and Alerting:**

* Implement a real-time monitoring system that continuously processes incoming water quality data.
* Set up alerting mechanisms to notify relevant authorities or stakeholders when anomalies are detected.

**8.Feedback Loop and Maintenance:**

* Continuously update and fine-tune your anomaly detection models to adapt to changing water quality conditions.
* Incorporate feedback from field experts and stakeholders to improve the accuracy and relevance of anomaly alerts.

**9.Documentation and Reporting:**

* Keep thorough documentation of the entire process, including data preprocessing, model selection, and evaluation metrics.
* Generate regular reports summarizing detected anomalies and their potential causes.

**10.** **Regulatory Compliance:**

* Ensure that your anomaly detection system complies with relevant environmental regulations and standards.

**11. Interdisciplinary Collaboration:**

* Collaborate with environmental scientists, hydrologists, and water quality experts to interpret anomalies and take appropriate actions.

**12. Long-Term Sustainability:**

* Plan for the long-term sustainability of your anomaly detection system, considering data acquisition, model updates, and hardware/software maintenance.

**Advantages:**

**1. Environmental Conservation:**

* Monitoring patterns in water quality helps in assessing the impact of human activities on natural water bodies

**2. Resource Management:**

* Patterns in water quality data can inform decisions related to water resource management. For example industry and **domestic use.**

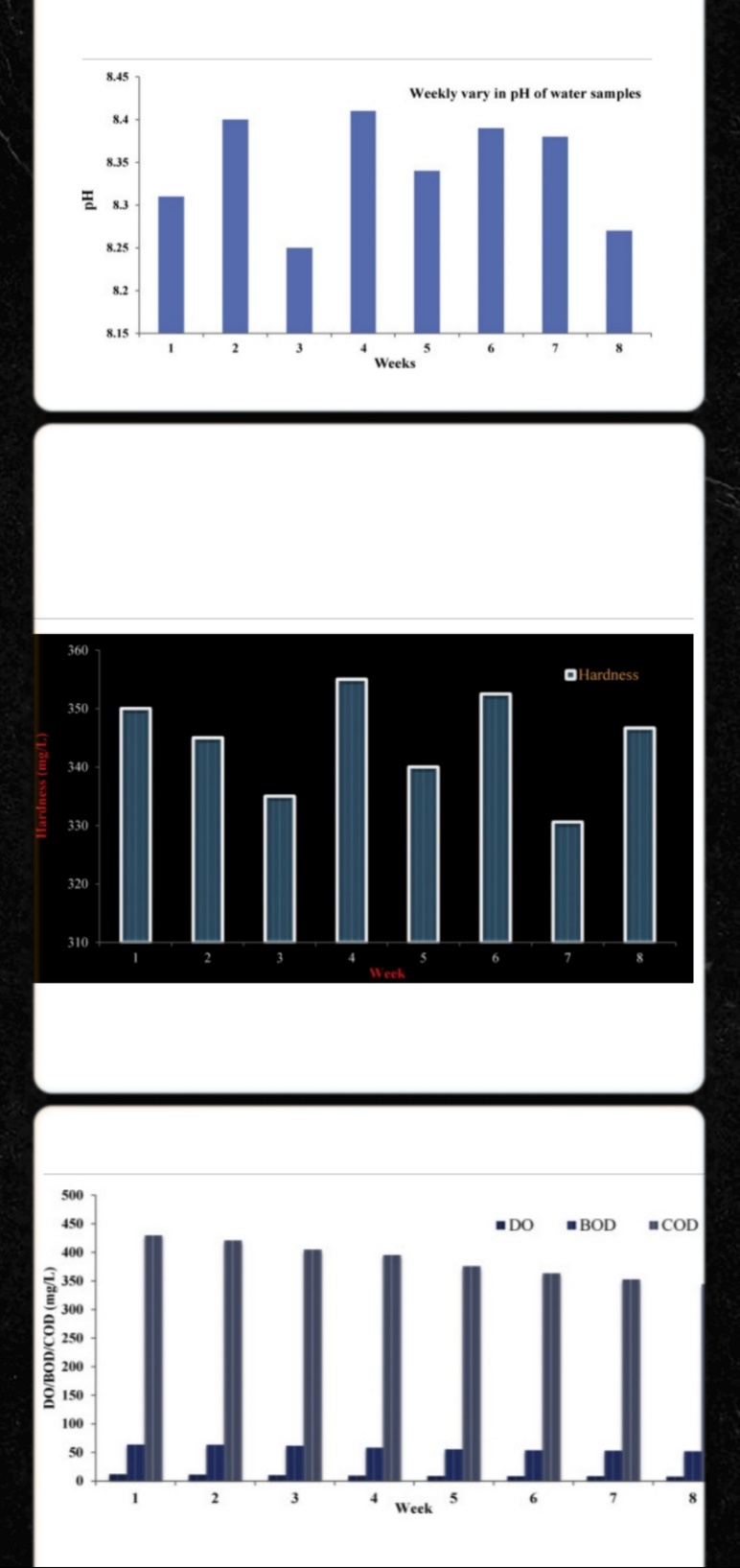
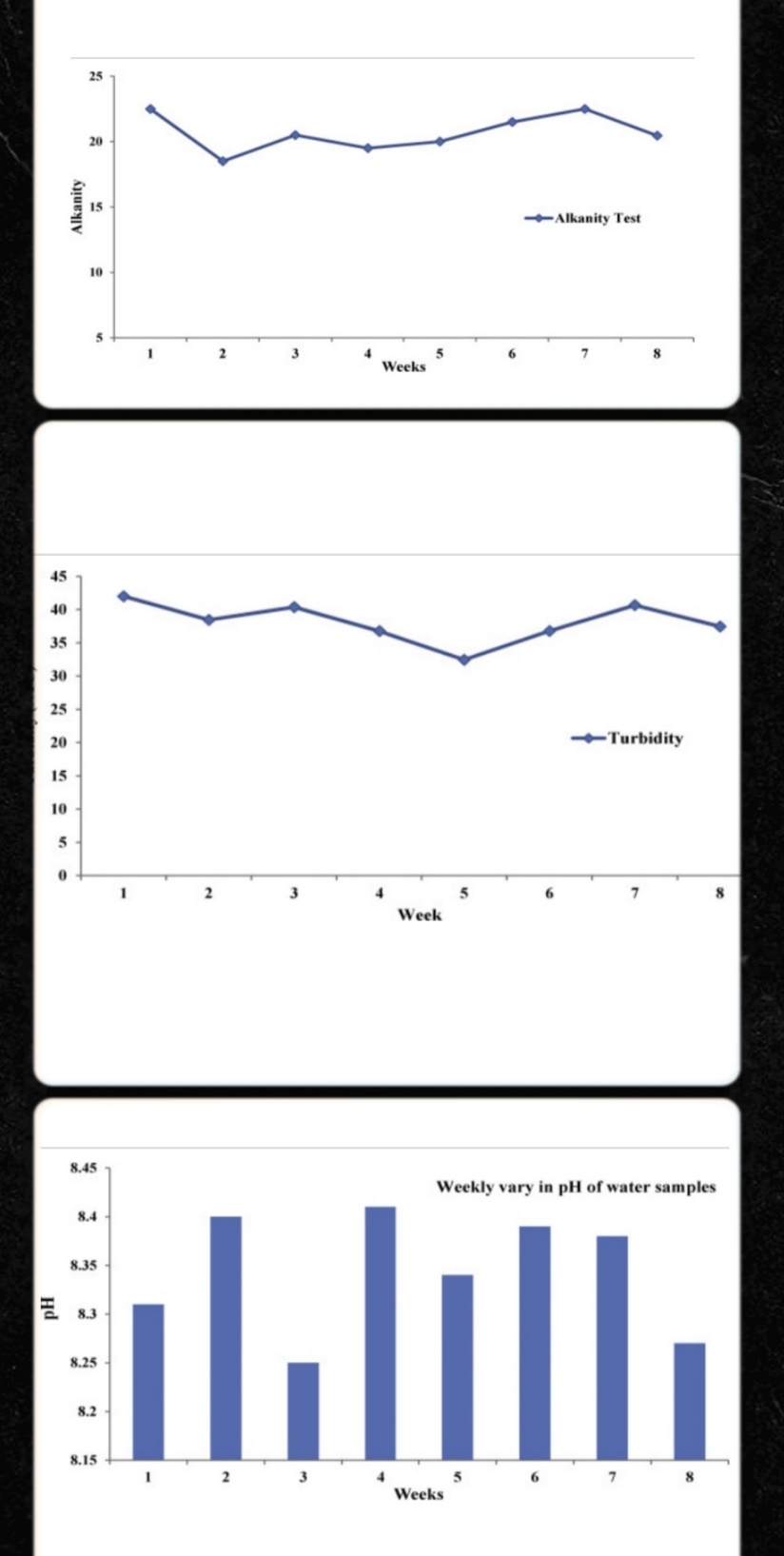
**3.** **Public Health:**

* Monitoring patterns in water quality parameters, particularly in drinking water sources, is critical for ensuring public health.

**4.Infrastructure Maintenance:**

* Water quality patterns can help identify issues with water treatment facilities and distribution systems.

**Weekly patterns in water quality parameters:**

****

**Model Evalution:**

* Split data set in to testing and training sets .
* Evaluate the models into parameters such as pH,Hardness,Solides, Chloramines,Condictivity etc…

**Program:**

In [1]:

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

from sklearn.ensemble import RandomForestRegressor, RandomForestClassifier

from sklearn.impute import SimpleImputer

from sklearn.metrics import accuracy\_score, precision\_score, recall\_score, f1\_score

from sklearn.model\_selection import train\_test\_split,cross\_val\_score, GridSearchCV, RandomizedSearchCV

import optuna

from skopt import BayesSearchCV

from hyperopt import fmin, tpe, hp

from xgboost import XGBClassifier

from scipy import stats

**Initial Data Exploration**

In [2]:

data. Head()

out [3]:

| ph | Hardness | Solids | Chloramines | Sulfate | Conductivity | Organic\_carbon | Trihalomethanes | Turbidity | Potability |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | NaN | 204.890455 | 20791.318981 | 7.300212 | 368.516441 | 564.308654 | 10.379783 | 86.990970 | 2.963135 | 0 |
| 1 | 3.716080 | 129.422921 | 18630.057858 | 6.635246 | NaN | 592.885359 | 15.180013 | 56.329076 | 4.500656 | 0 |
| 2 | 8.099124 | 224.236259 | 19909.541732 | 9.275884 | NaN | 418.606213 | 16.868637 | 66.420093 | 3.055934 | 0 |
| 3 | 8.316766 | 214.373394 | 22018.417441 | 8.059332 | 356.886136 | 363.266516 | 18.436524 | 100.341674 | 4.628771 | 0 |
| 4 | 9.092223 | 181.101509 | 17978.986339 | 6.546600 | 310.135738 | 398.410813 | 11.558279 | 31.997993 | 4.075075 | 0 |

**Univariate Analysis:**

In [4]:

sns.set(style="whitegrid")

plt.figure(figsize=(12, 8))

numerical\_columns = data.drop("Potability", axis=1).columns

for column **in** numerical\_columns:

plt.subplot(3, 3, numerical\_columns.get\_loc(column) + 1)

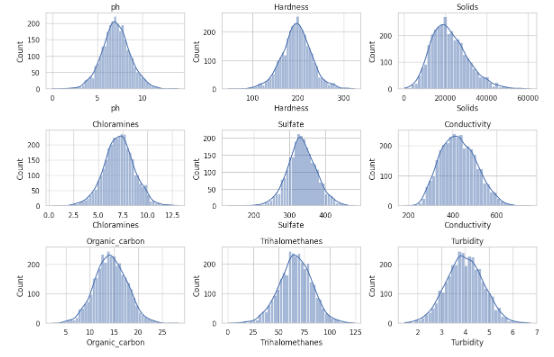
sns.histplot(data[column], kde=True)

plt.title(column)

plt.tight\_layout()

plt.show()

out [5]:



**Bivariate Analysis:**

In [6]:

import plotly.graph\_objs as go

index\_vals = data['Potability'].astype('category').cat.codes

fig = go.Figure(data=go.Splom(

dimensions=[dict(label='ph',

values=data['ph']),

dict(label='Hardness',

values=data['Hardness']),

dict(label='Solids',

values=data['Solids']),

dict(label='Chloramines',

values=data['Chloramines']),

dict(label='Sulfate',

values=data['Sulfate']),

dict(label='Conductivity',

values=data['Conductivity']),

dict(label='Organic\_carbon',

values=data['Organic\_carbon']),

dict(label='Trihalomethanes',

values=data['Trihalomethanes']),

dict(label='Turbidity',

values=data['Turbidity'])],

showupperhalf=False,

text=data['Potability'],

marker=dict(color=index\_vals,

showscale=False,

line\_color='white', line\_width=0.5)

))

fig.update\_layout(

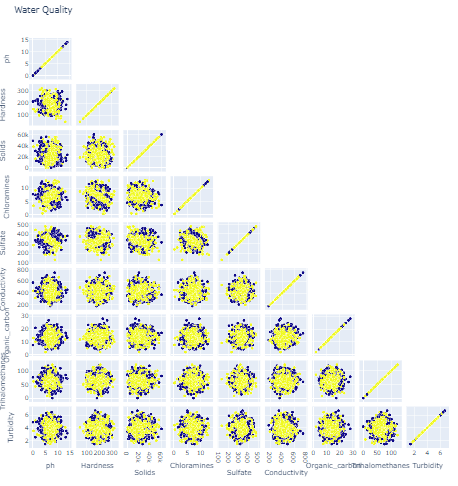
title='Water Quality',

width=1000,

height=1000,

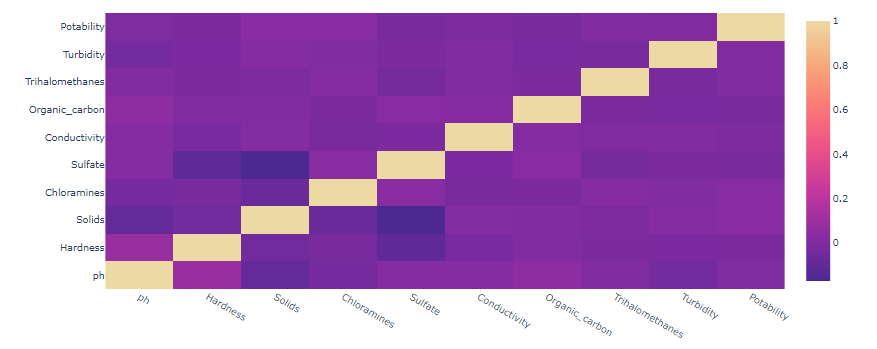
)

out [7]:



In [8]:

fig = go.Figure(go.Heatmap(z=data.corr(), x=data.corr().columns.tolist(), y=data.corr().columns.tolist(), colorscale='agsunset'))

 out [3]:

In [9]:

correlation\_matrix = data.corr()

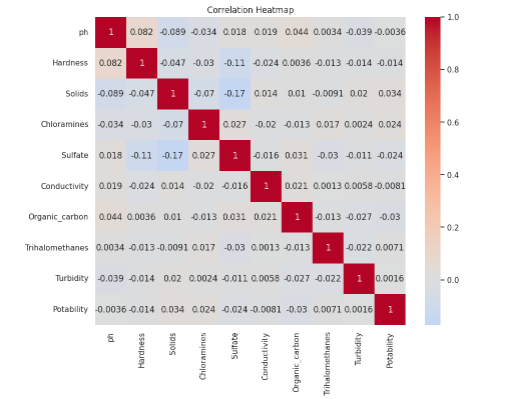
plt.figure(figsize=(10, 8))

sns.heatmap(correlation\_matrix, annot=True, cmap="coolwarm", center=0)

plt.title("Correlation Heatmap")

plt.show()

out [10]:



Conclusion:

* In this phase 2 conclusion, we will summarize the key findings and insights from the different type of parameters.we will reiterate the impact of the anomaly detection techniques to identify the unusual patterns in water quality parameters.