WATER QUALITY ANALISIS

PROJECT DEFINTION:

A project in water quality analysis involves conducting a comprehensive assessment of the physical, chemical, and biological characteristics of a body of water to determine its overall quality and potential environmental impacts. Such projects are crucial for ensuring the safety of water sources, protecting ecosystems, and meeting regulatory requirements. Here's a detailed project definition for water quality analysis:

Project Title:

Water Quality Analysis of [Name of Water Body]

Project Objective:

The primary objective of this project is to assess and analyze the water quality of [Name of Water Body], including its source(s) and downstream areas. The project aims to provide valuable insights into the current state of water quality, identify potential pollution sources, and propose measures for water quality improvement and preservation of aquatic ecosystems.

Key Components and Tasks:

1. Site Selection and Sampling Design:

- Identify sampling locations within [Name of Water Body] that represent different environmental conditions.
- Develop a sampling plan to ensure representative data collection.

2. Data Collection:

- Gather samples of water from the selected sites at regular intervals.
- Measure various water quality parameters, including but not limited to:
- Temperatur - pH- Dissolved oxygen
- Turbidity
- Nutrient levels (e.g., nitrogen and phosphorus)
- Suspended solids
- Heavy metal concentrations

- Biological indicators (e.g., fecal coliform bacteria)

3.Laboratory Analysis:

- Conduct laboratory tests to analyze the collected water samples for specific parameters.
- Utilize appropriate analytical methods and equipment to obtain accurate results.

4. Data Interpretation and Analysis:

- Analyze the collected data to assess the overall water quality of [Name of Water Body].
- Compare the results with relevant water quality standards and guidelines.
- Identify any trends or anomalies in the data.

5. Source Identification and Pollution Assessment:

- Investigate potential pollution sources within the watershed.
- Evaluate the impact of human activities on water quality.
- Determine the contribution of point and non-point source pollution.

6. Environmental Impact Assessment:

- Assess the impact of poor water quality on aquatic ecosystems, including fish and other aquatic organisms.
- Identify potential risks to human health if applicable.

7. Recommendations and Mitigation Measures:

- Develop a list of recommendations for improving water quality.
- Propose mitigation measures to address pollution sources.
- Suggest regulatory or management actions to protect and preserve [Name of Water Body].

8. Reporting and Communication:

- Compile the findings and recommendations into a comprehensive report.
- Share the results with relevant stakeholders, including government agencies, environmental organizations, and the public.

9. Long-Term Monitoring Plan:

- Establish a plan for ongoing monitoring of water quality in [Name of Water Body] to track changes over time and assess the effectiveness of mitigation measures.

Project Deliverables:

- Comprehensive Water Quality Assessment Report
- Data sets from water quality measurements
- Recommendations for water quality improvement and protection

Project Timeline:

The project timeline may vary depending on the size of the water body, the complexity of the analysis, and available resources. Typically, a water quality analysis project can range from several months to multiple years.

Budget:

- The budget for the project should cover expenses related to sample collection, laboratory analysis, equipment, personnel, and reporting.

Key Stakeholders:

- Government agencies responsible for water quality regulation and management
- Environmental organizations
- Local communities and residents
- Scientific and research institutions

Regulatory Compliance:

- Ensure that the project adheres to all relevant environmental regulations and permits.
- By following this project definition, you can conduct a thorough water quality analysis that contributes to the protection and preservation of aquatic ecosystems and the availability of safe and clean water resources.

DESIGN THINKING:

Design thinking is a problem-solving approach that emphasizes empathy, creativity, and iteration to develop innovative solutions. When applying design thinking to a project like water quality analysis, you can follow these stages:

1. Empathize:

- Understand the problem: Begin by researching and gathering information about the specific water quality issues in your target area. Interview experts, stakeholders, and potential users to gain insights into their needs and concerns.
- User personas: Create personas representing the different types of people affected by water quality issues, such as residents, local authorities, environmentalists, and scientists. Understand their goals and pain points.

2. Define:

- Problem statement: Based on your research, craft a clear and concise problem statement that captures the essence of the water quality issue you want to address. For example, "How might we improve access to real-time water quality data for residents of XYZ City?"
- Point of view: Frame the problem from the perspective of the user personas, highlighting their needs and challenges.

3. Ideate:

- Brainstorm solutions: Generate a wide range of ideas to address the problem. Encourage creativity and diverse perspectives. Use techniques like brainstorming sessions, mind mapping, and ideation workshops.
- Crazy 8s: Give participants a limited time (e.g., 8 minutes) to sketch eight different concepts for solving the problem.

4. Prototype:

- Build a prototype: Create a simplified, low-cost version of your solution. In the context of water quality analysis, this could be a mockup of a water testing device, a data visualization tool, or a mobile app.

- Test your prototype: Gather feedback from potential users and stakeholders. Determine what works and what needs improvement. Iterate on your prototype based on this feedback.

5. Test:

- Pilot implementation: Implement a small-scale test of your solution in a real-world setting, such as a specific neighborhood or water source. Monitor its performance and gather data.
- User feedback: Continue to collect feedback from users, adjusting and refining your solution as necessary.

6. Implement:

- Scale up: If your pilot test is successful, scale up the implementation of your solution to a broader area or community.
- Collaborate with stakeholders: Engage with local authorities, environmental agencies, and community organizations to ensure the long-term sustainability and adoption of your solution.

7. Evaluate and Iterate:

- Continuous improvement: Regularly assess the impact of your solution on water quality and user satisfaction. Make necessary adjustments based on ongoing feedback and new insights.
- Throughout the design thinking process, keep the needs and perspectives of the end-users at the center of your efforts. Remember that design thinking is an iterative process, and you may need to revisit and refine your solution multiple times to achieve the desired results in water quality analysis.

INNOVATION:

Incorporating innovation into the process of anomaly detection for water quality parameters can lead to more effective, efficient, and adaptable monitoring systems. Here are some innovative approaches and technologies that can enhance anomaly detection in water quality:

Sensor Networks and IoT:

Deploy advanced sensor networks and Internet of Things (IoT) devices to collect real-time data from various points in a water system. These sensors can provide continuous data streams, enabling quicker anomaly detection.

Big Data and AI:

Utilize big data analytics and artificial intelligence (AI) techniques like deep learning for anomaly detection. These technologies can handle vast datasets and identify complex patterns that may go unnoticed by traditional methods.

Predictive Analytics:

Combine historical water quality data with predictive analytics to forecast potential anomalies. Machine learning models can predict future water quality based on current and historical parameters, making it easier to detect deviations from expected values.

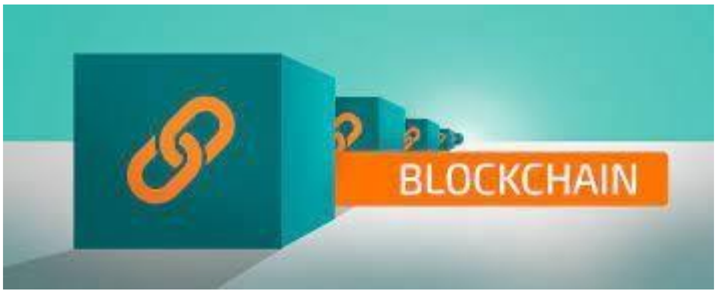
Remote Sensing:

Implement remote sensing technologies, such as satellite imagery or drones equipped with sensors, to monitor water quality in large bodies of water. These innovative approaches can complement ground-based monitoring.

Satellite Remote Sensing Applications for Global Water Resources

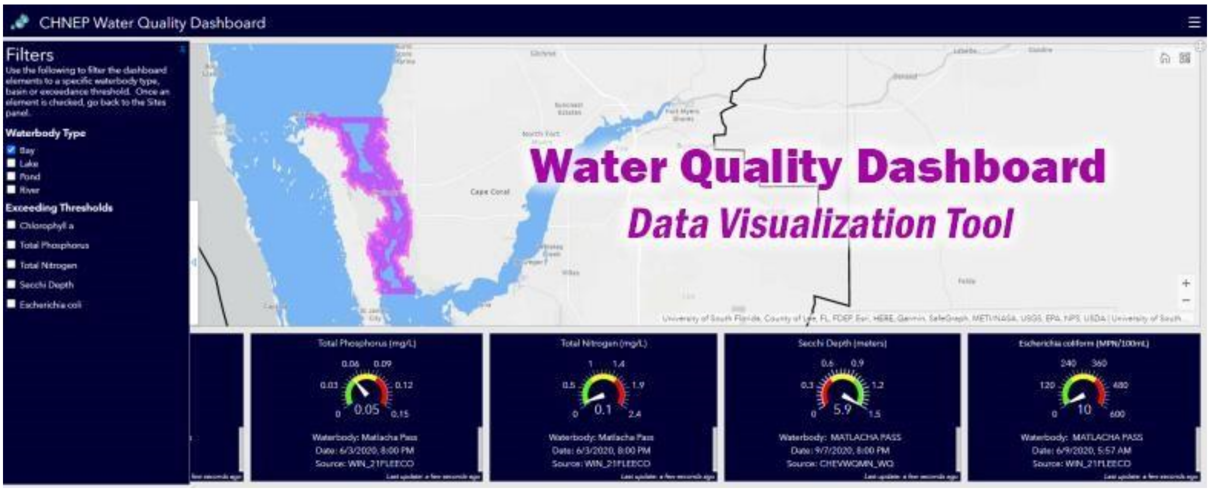
Blockchain for Data Integrity:

Use blockchain technology to secure and verify the integrity of water quality data. This ensures that data is tamper-proof, which is crucial for accurate anomaly detection and compliance with regulations.



Real-time Visualization and Dashboards:

Develop interactive, real-time visualization tools and dashboards that allow water quality professionals to monitor parameters and anomalies on the go. User-friendly interfaces can enable faster response to issues.



CHNEP Water Quality Dashboard

Filters

Use the following to filter the dashboard elements to a specific waterbody type, basin or exceedance threshold. Once an element is checked, go back to the Sites panel.

Waterbody Type

☑ Bay
□ Lake
□ Pond
□ River

Exceeding Thresholds

□ Chlorophyll a
□ Total Phosphorus
□ Total Nitrogen
□ Secchi Depth
□ Escherichia coli

Total Phosphorus (mg/L)
Waterbody: Matlacha Pass
Date: 6/3/2020, 8:00 PM
Source: WIN_21FLEECO
Last update: a few seconds ago

Total Nitrogen (mg/L)
Waterbody: Matlacha Pass
Date: 6/3/2020, 8:00 PM
Source: WIN_21FLEECO
Last update: a few seconds ago

Secchi Depth (meters)
Waterbody: MATLACHA PASS
Date: 9/7/2020, 8:00 PM
Source: CHEVWQMN_WQ
Last update: a few seconds ago

Escherichia coliform (MPN/100mL)
Waterbody: MATLACHA PASS
Date: 6/9/2020, 5:57 AM
Source: WIN_21FLEECO
Last update: a few seconds ago

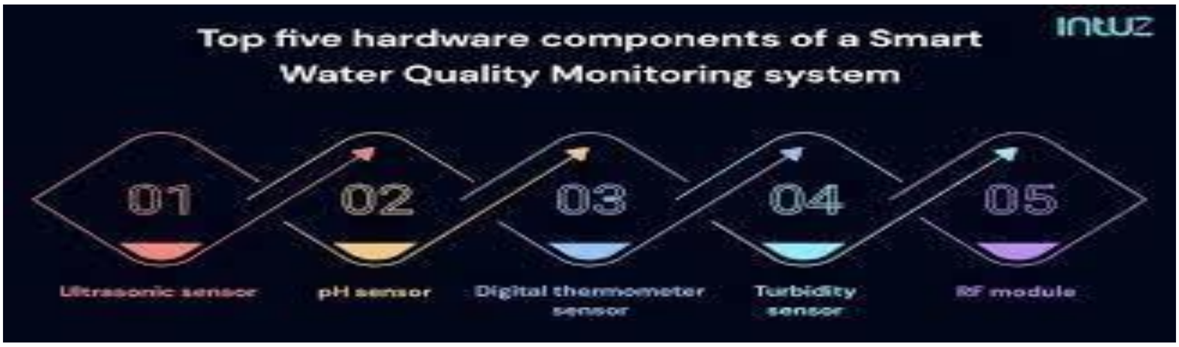
Water Quality Dashboard
Data Visualization Tool
Cape Coral
University of South Florida, County of Lee, FL, FDEP, Esri, HERE, Garmin, SafeGraph, METI/NASA, USGS, EPA, NPS, USDA | University of South
0.06
0.09
0.03
0.12
0
0.05
0.15
1
1.4
0.5
1.9
0
0.1
2.4
0.6
0.9
0.3
1.2
0
5.9
1.5
240
360
120
480
0
10
600

Crowdsourced Data:

Encourage citizen science initiatives and crowdsourced data collection. Innovative mobile apps and platforms can involve the public in monitoring and reporting water quality anomalies, expanding the data sources.

Smart Alarms and Automated Responses:

Integrate automated alert systems that not only detect anomalies but also trigger predefined responses. This might include shutting down equipment or activating backup systems when severe water quality issues are detected.


Top five hardware components of a Smart Water Quality Monitoring system
intuz
01
02
03
04
05
Ultrasonic sensor
pH sensor
Digital thermometer sensor
Turbidity sensor
RF module

Fusion of Data Sources:

Combine water quality data with data from other sources, such as weather, land use, and pollution sources. Innovations in data fusion can provide a more comprehensive view of factors affecting water quality.

Quantum Computing:

Explore the potential of quantum computing for handling complex water quality data. Quantum computing can perform highly complex simulations and calculations that were previously impossible with classical computers.

Bioinformatics and Biological Sensors:

Develop biological sensors that can detect changes in water quality by monitoring the presence and behavior of aquatic organisms. This approach, often called bioinformatics, can provide early warning signs of ecological disruptions.

Machine-to-Machine Communication:

Enable machine-to-machine communication for rapid data exchange between monitoring equipment and central systems. This can facilitate real-time anomaly detection and response.

Autonomous Underwater Vehicles (AUVs):

Deploy AUVs equipped with water quality sensors to collect data from hard-to-reach or deepwater areas. These innovative technologies expand the scope of monitoring.

Quantum Sensors:

Investigate the use of quantum sensors that can provide ultra-sensitive measurements, allowing for earlier detection of subtle changes in water quality.

By incorporating these innovative approaches and technologies into anomaly detection for water quality parameters, you can create more robust, responsive, and comprehensive monitoring systems that are better equipped to identify and respond to unusual patterns in water quality. These innovations can improve
environmental protection, public health, and the sustainability of water resources.

Data Collection:

Start by collecting historical data on water quality parameters such as pH levels, turbidity, temperature, dissolved oxygen, and chemical concentrations. Ensure that you have a substantial dataset for training and testing your anomaly detection model.

Data Preprocessing:

- Data Cleaning: Remove any missing or erroneous data points.
- Feature Selection: Choose the most relevant parameters for anomaly detection.
- Normalization: Scale the data if the parameters have different units or scales.

Choose Anomaly Detection Algorithms:

- Statistical Methods: Utilize statistical methods like Z-scores, modified Zscores, or percentiles to identify outliers.
- Machine Learning Models: Train models like Isolation Forest, One-Class SVM, or Autoencoders for anomaly detection.
- Time Series Analysis: If your data is time-dependent, consider methods like Seasonal Decomposition of Time Series (STL) or Prophet to identify anomalies in temporal patterns.

Model Training:

Split your dataset into training and testing sets. The training set is used to train the anomaly detection model.

Validate the model's performance on the testing set, adjusting hyperparameters as needed.

Threshold Selection:

Depending on the chosen method, you may need to set a threshold for what constitutes an anomaly. This threshold can be determined using statistical techniques or cross-validation.

Real-time Monitoring:

Implement a system for real-time monitoring of water quality data. The model can continuously assess incoming data and raise alarms if anomalies are detected.

Visualization:

Use data visualization tools to display the anomalies and their impact on water quality parameters. Visualization can help in understanding the extent of anomalies.

Feedback Loop:

Establish a feedback loop where detected anomalies trigger further investigation or maintenance activities. This can be integrated with a notification system for relevant stakeholders.

Continuous Improvement:

Periodically retrain your anomaly detection model to adapt to changing conditions and data patterns. Water quality can change with seasons, weather, or other environmental factors.

Domain Knowledge:

Incorporate domain expertise in the anomaly detection process. Experts can help in understanding the potential causes of anomalies and taking appropriate actions.

Regulatory Compliance:

Ensure that the anomaly detection system complies with relevant regulations and standards for water quality monitoring and reporting.

Data Integration:

If available, consider integrating other data sources like weather data, pollution sources, or flow rates to improve anomaly detection accuracy.

DEVELOPMENT PART-1

Building a water quality analysis model involves several steps, including data preprocessing and exploratory data analysis (EDA) to better understand the dataset. In this example, I'll provide a high-level overview of the process, but keep in mind that the specifics may vary depending on your dataset. For this exercise, let's assume you have a dataset with water quality measurements.

You can use Python and popular libraries

like Pandas, Matplotlib, and Seaborn for the task.

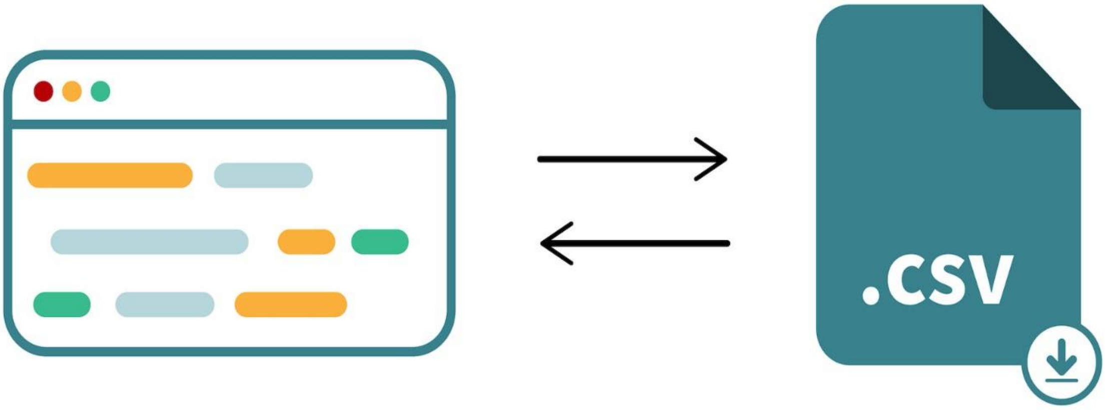
1. Import Libraries:

Start by importing the necessary libraries:

Python code:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

2. Load the Data:

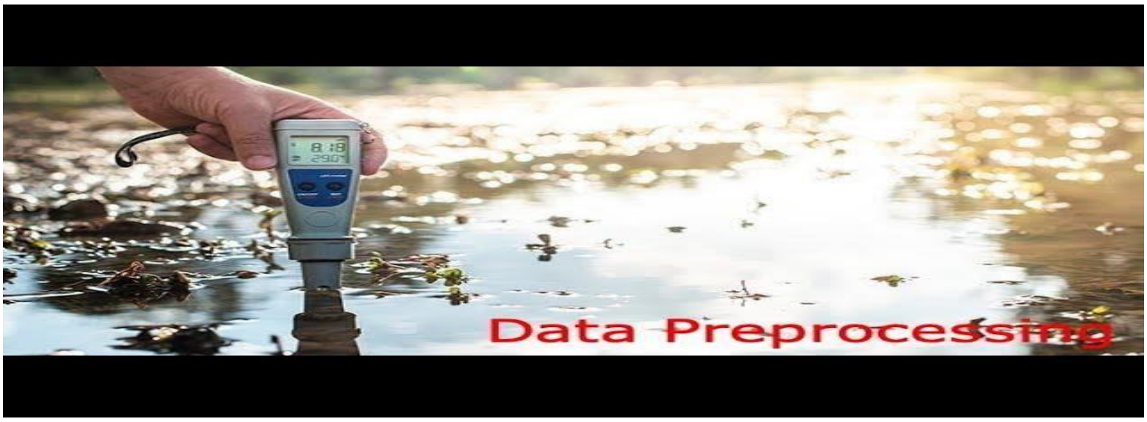


Load your water quality dataset into a Pandas DataFrame:

Python code:

```
# Replace 'water_quality_data.csv' with your data file path
df = pd.read_csv('water_quality_data.csv')
```

3. Data Preprocessing:



a)Data Cleaning:

- Handle missing values: Use `df.dropna()` or `df.fillna()` to deal with missing data.
- Remove duplicates: Use `df.drop_duplicates()` to remove duplicate rows.

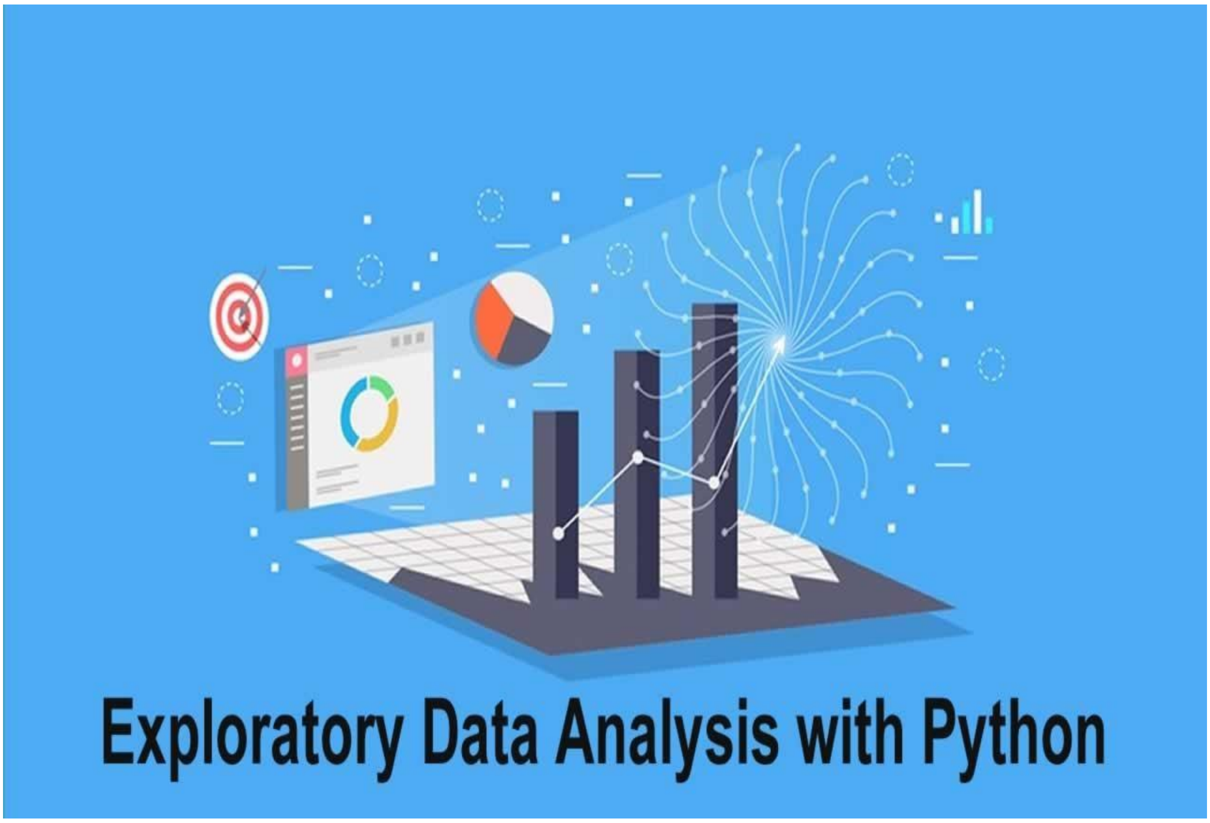
b)Data Transformation:

Convert data types: Ensure that numeric columns are in the correct data types (e.g., float for measurements, datetime for dates).

c)Feature Engineering:

Create new features if necessary. For example, you can extract the month and year from a date column.

4. Exploratory Data Analysis (EDA):



EDA helps you gain insights into the data and understand its characteristics.

a. Summary Statistics:

- Use `df.describe()` to get summary statistics for numeric columns.

b. Data Visualization:

- Create visualizations to understand the data better. For example:

Python code:

```
# Histogram of a water quality parameter (e.g.,pH)
plt.hist(df['pH'], bins=20, color='blue')
plt.xlabel('pH')
plt.ylabel('Frequency')
plt.title('pH Distribution')
plt.show()
# Box plot to identify outliers
sns.boxplot(x='Parameter', y='Value', data=df)
plt.xlabel('Water Quality Parameter')
plt.ylabel('Value')
plt.title('Box Plot of Water Quality Parameters')
plt.xticks(rotation=45)
plt.show()
```

c. Correlation Analysis:

- Use `df.corr()` to calculate the correlation matrix between water quality parameters.

d. Time Series Analysis:

- If your dataset includes timestamps, explore temporal trends and patterns.

5. Data Preprocessing (Continued):

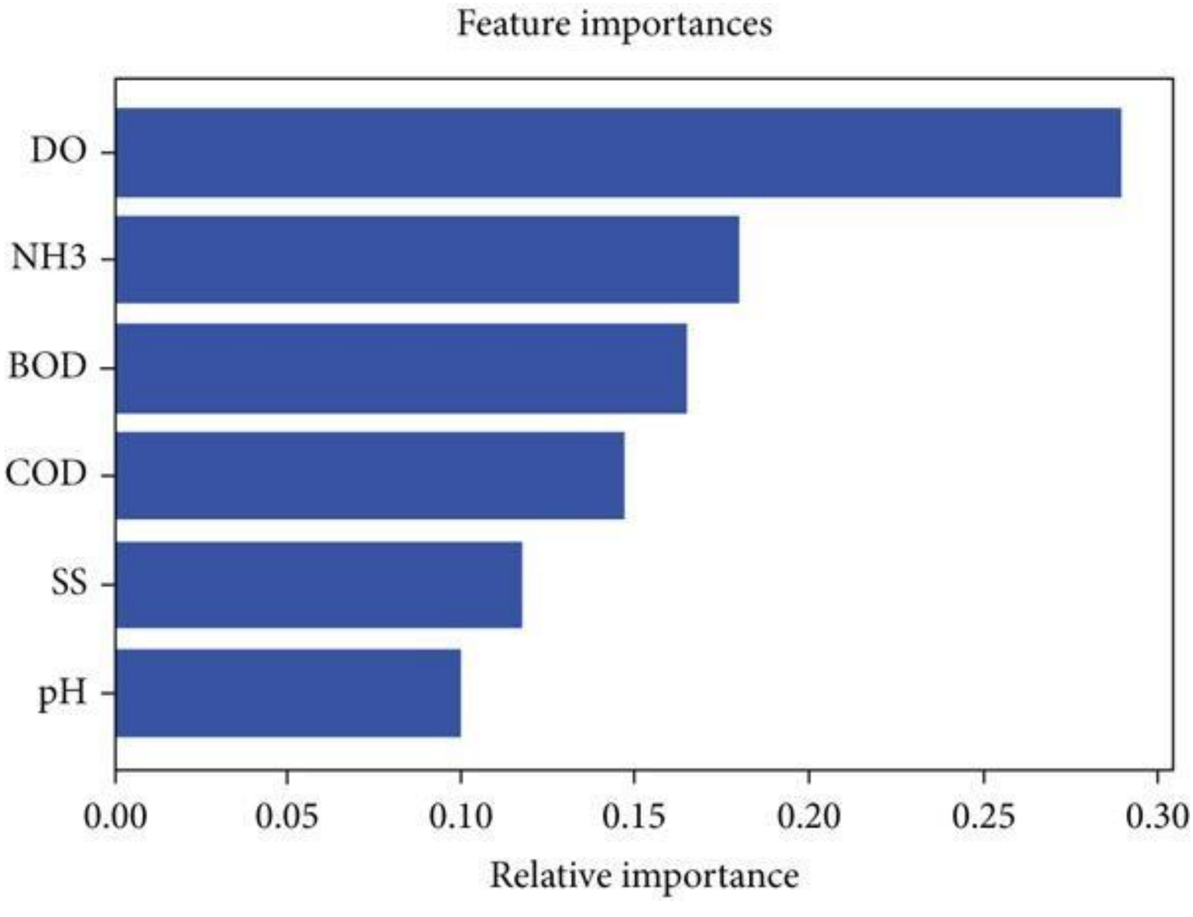
a. Outlier Detection:

Identify and handle outliers if necessary, e.g., using the IQR method or zscores.

b.Normalization/Scaling:

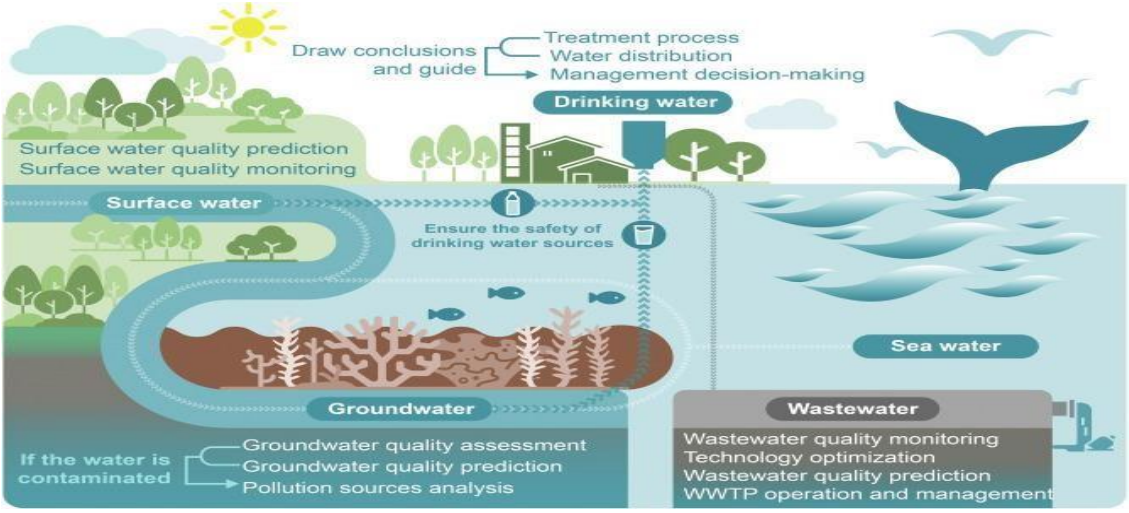
If the data is not on the same scale, consider normalizing or scaling it.

6. Feature Selection/Engineering:



Based on the insights gained from EDA, you can select relevant features or engineer new ones that might improve the model's performance.

7. Train the Model:


Draw conclusions and guide
Treatment process
Water distribution
Management decision-making
Drinking water
Surface water quality prediction
Surface water quality monitoring
Surface water
Ensure the safety of drinking water sources
Sea water
Groundwater
Wastewater
If the water is contaminated
Groundwater quality assessment
Groundwater quality prediction
Pollution sources analysis
Wastewater quality monitoring
Technology optimization
Wastewater quality prediction
WWTP operation and management

With the preprocessed data, you can now proceed to build and train your water quality analysis model, which might involve machine learning algorithms like regression, classification, or time series analysis, depending on your specific goals.

Remember that the preprocessing and EDA steps are crucial for understanding your data, identifying issues, and preparing it for modeling. This is a general guideline, and you should adapt it to the specifics of your dataset and analysis goals.

Water quality is assessed based on various parameters to ensure that it meets safety and environmental standards. Here are some common water quality parameters:

1. **pH (Acidity/Alkalinity):**

pH measures the hydrogen ion concentration in water. It indicates whether the water is acidic (pH < 7), neutral (pH = 7), or alkaline (pH > 7).

2. **Temperature:**

Water temperature can affect various aquatic organisms and chemical reactions. It's an important parameter, especially for aquatic ecosystems.

3. **Dissolved Oxygen (DO):**

DO levels are crucial for aquatic life. Low DO can lead to hypoxia, harming fish and other organisms.

4. Turbidity:

Turbidity measures the cloudiness or haziness of water. It's an indicator of suspended solids and can affect water quality and ecosystems.

5. Total Dissolved Solids (TDS):

TDS measures the concentration of inorganic and organic substances in water. It includes minerals, salts, and other dissolved materials.

6. Electrical Conductivity (EC):

EC measures the water's ability to conduct electrical current, which is related to the ion concentration. It's often used as a proxy for TDS.

7. Chemical Oxygen Demand (COD):

COD is a measure of the oxygen required to chemically break down organic and inorganic matter in water. High COD can indicate pollution.

8. Biological Oxygen Demand (BOD):

BOD is a measure of the amount of oxygen consumed by microorganisms while decomposing organic matter. It's another indicator of pollution.

9. Nutrients:

Parameters such as nitrate, nitrite, phosphate, and ammonia are measured to assess nutrient pollution, which can lead to algal blooms and water quality problems.

10. Metals:

Testing for heavy metals like lead, mercury, and cadmium is important to ensure water safety. **11. Coliform Bacteria:**

Coliform bacteria are used as indicators of microbial contamination and the potential presence of harmful pathogens.

12. Chlorine Residual:

Chlorine is often used for disinfection in drinking water treatment. Monitoring residual chlorine ensures effective disinfection.

13. Pesticides and Herbicides:

Testing for various agricultural chemicals is important to identify potential contaminants in water sources.

14. Taste and Odor:

Subjective parameters related to the taste and odor of water, which can affect its acceptability.

These parameters help assess the physical, chemical, and biological characteristics of water, allowing for the monitoring and maintenance of water quality standards for various purposes, including drinking water, aquatic ecosystems, and industrial processes. The specific parameters of interest may vary depending on the application and regulatory requirements.

DEVELOPMENT PART-2

involved. Keep in mind that the specific tools and techniques you use will Creating visualizations and building a predictive model involves several steps, and it can be a complex process. I'll provide a high-level overview of the steps depend on the nature of your data and your objectives.

Data Preparation

Start by cleaning and preprocessing your data. This may involve handling missing values, encoding categorical variables, and scaling numerical features.

Ensure your data is in a suitable format for analysis.

Exploratory Data Analysis (EDA)

Conduct EDA to gain insights into your data. Visualize the data to understand its distribution, relationships between variables, and potential patterns. You can use tools like matplotlib, seaborn, or Tableau for this.

Feature Engineering

Create new features or transform existing ones to improve the performance of your predictive model. This step often involves domain knowledge and creativity. Feature engineering can have a significant impact on model accuracy.

Data Splitting

Divide your data into training, validation, and test sets. The training set is used to train the model, the validation set is used for hyperparameter tuning, and the test set is used to evaluate the model's performance.

Model Selection

Choose an appropriate machine learning or statistical model for your predictive task. The choice of model depends on the nature of the data (classification, regression, time series, etc.) and your specific goals.

Model Training

Train your chosen model on the training data. Tune hyperparameters as needed to optimize performance. You can use libraries like scikit-learn, TensorFlow, or PyTorch for model training.

Model Evaluation

Assess the performance of your model using the validation set. Common evaluation metrics include accuracy, precision, recall, F1 score, RMSE, MAE, etc., depending on the problem type.

Hyperparameter Tuning

Fine-tune your model's hyperparameters to optimize its performance. You can use techniques like grid search, random search, or Bayesian optimization.

Visualization for Model Interpretability

Visualize the model's predictions and decision boundaries. Tools like SHAP values, Partial Dependence Plots, and LIME can help you interpret and explain your model's predictions.

Final Model Selection and Testing

After selecting the best-performing model and fine-tuning its hyperparameters, evaluate it on the test set to obtain a final performance estimate.

Deployment

If the model meets your expectations, deploy it in your application or workflow. Ensure that it's properly integrated and maintained.

Monitoring and Maintenance

Continuously monitor the model's performance in production. Retrain or update the model as needed to account for changing data distributions or requirements.

Visualization tools and libraries you can use include Matplotlib, Seaborn, Plotly, or interactive dashboards created with tools like Streamlit or Dash.

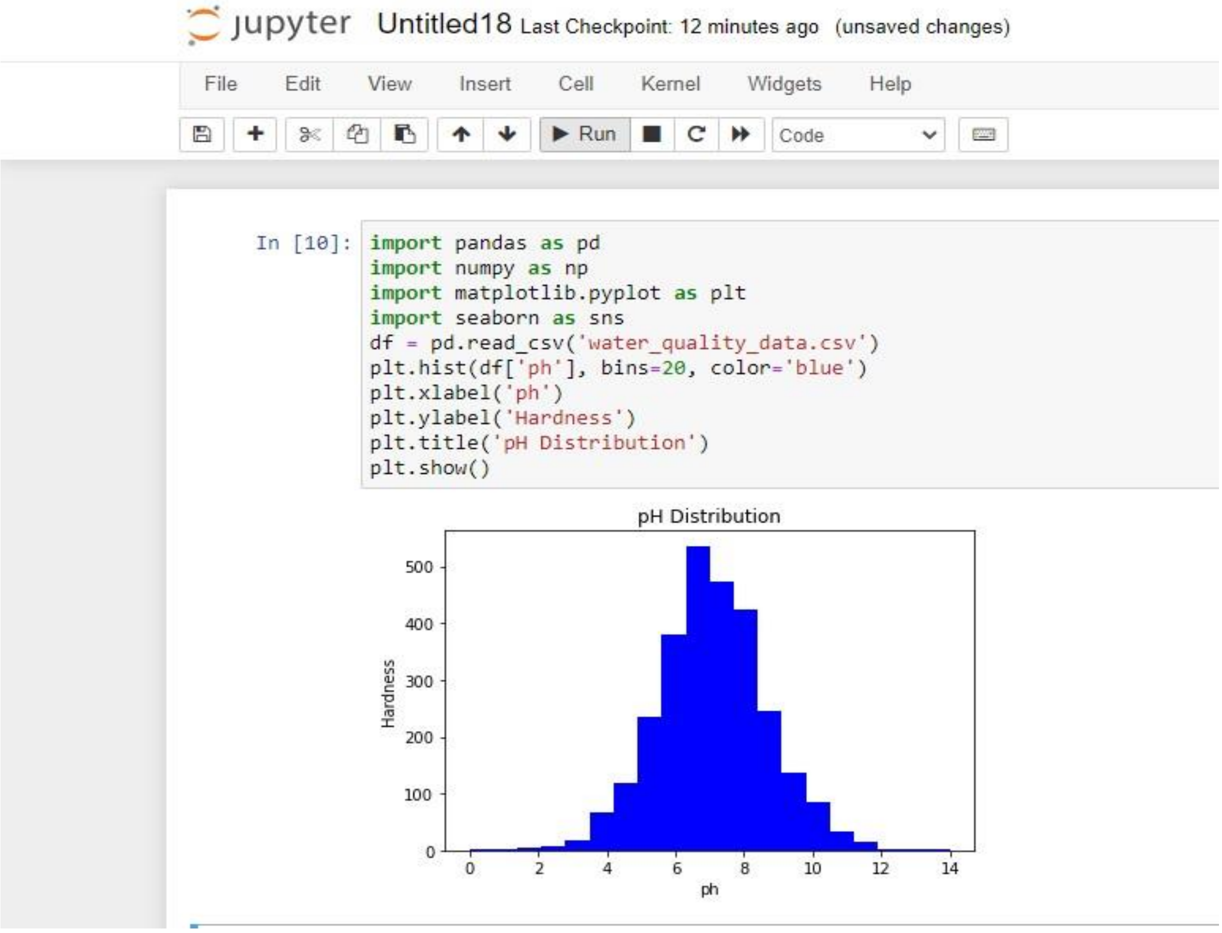
Remember that the choice of tools and techniques depends on your specific dataset and predictive task. The above steps provide a general framework, and you may need to adapt them to your unique circumstances. Additionally, the choice of machine learning algorithms and visualization techniques should alignwith the problem you are trying to solve (e.g., regression, classification, clustering, time series forecasting, etc.)

DATA SET LINK:

https://www.kaggle.com/datasets/adityakadiwal/water-potability

Installation code

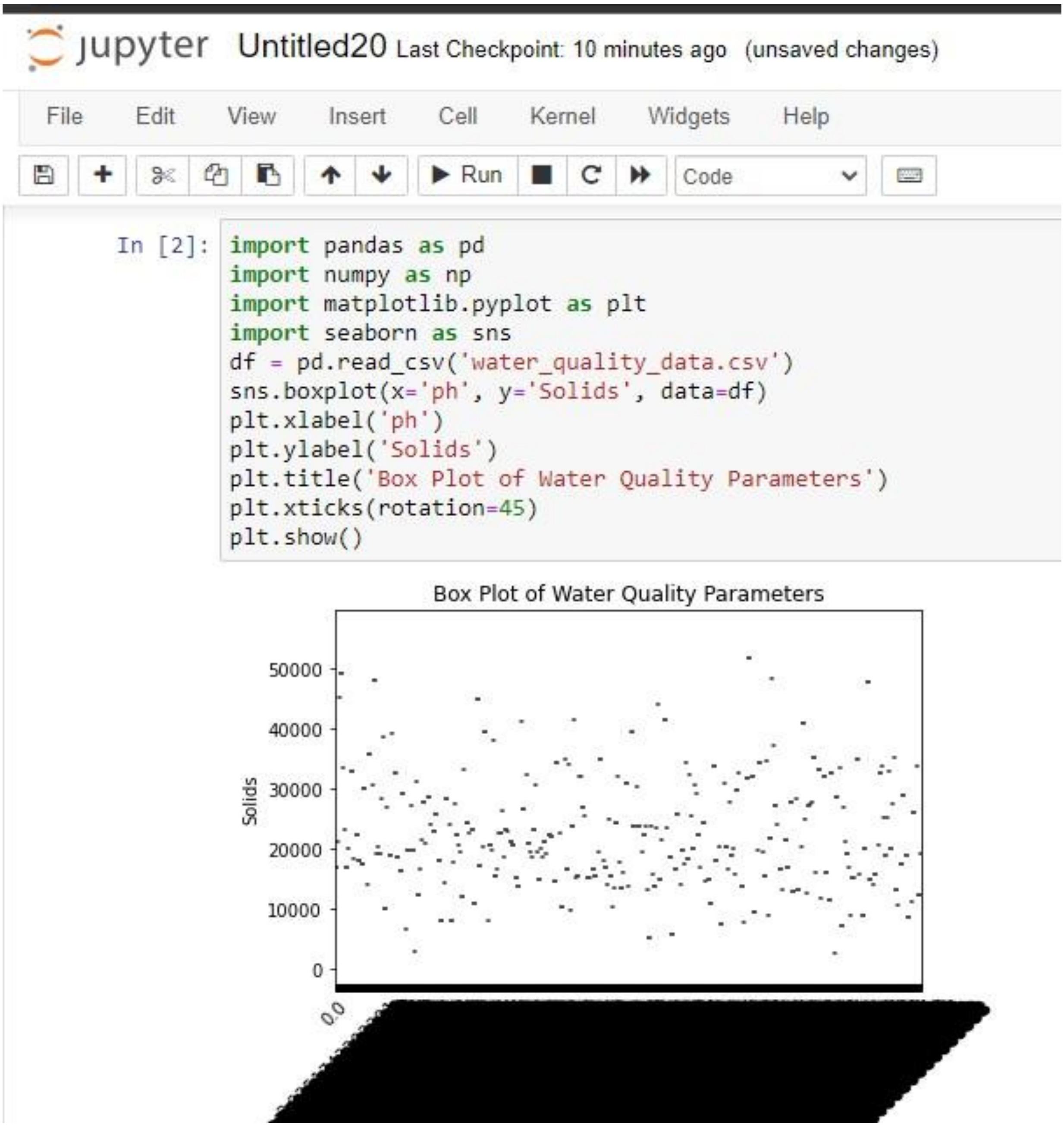
```
import numpy as np  import pandas as pd  import
matplotlib.pyplot as plt plt.style.use('fivethirtyeight')
plt.style.use('dark_background') import seaborn as sns
color = sns.color_palette()
import plotly.express         as ex
import plotly.graph_objs      as go
import plotly.offline         as pyo
import scipy.stats            as stats
# Histogram of a water quality
parameter
plt.hist(df['pH'], bins=20, color='blue')
plt.xlabel('pH')
 plt.ylabel('Frequency')
plt.title('pH Distribution')
 plt.show()
```



```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
df = pd.read_csv('water_quality_data.csv')
plt.hist(df['ph'], bins=20, color='blue')
plt.xlabel('ph')
plt.ylabel('Hardness')
plt.title('pH Distribution')
plt.show()
```

# Box plot to identify outliers

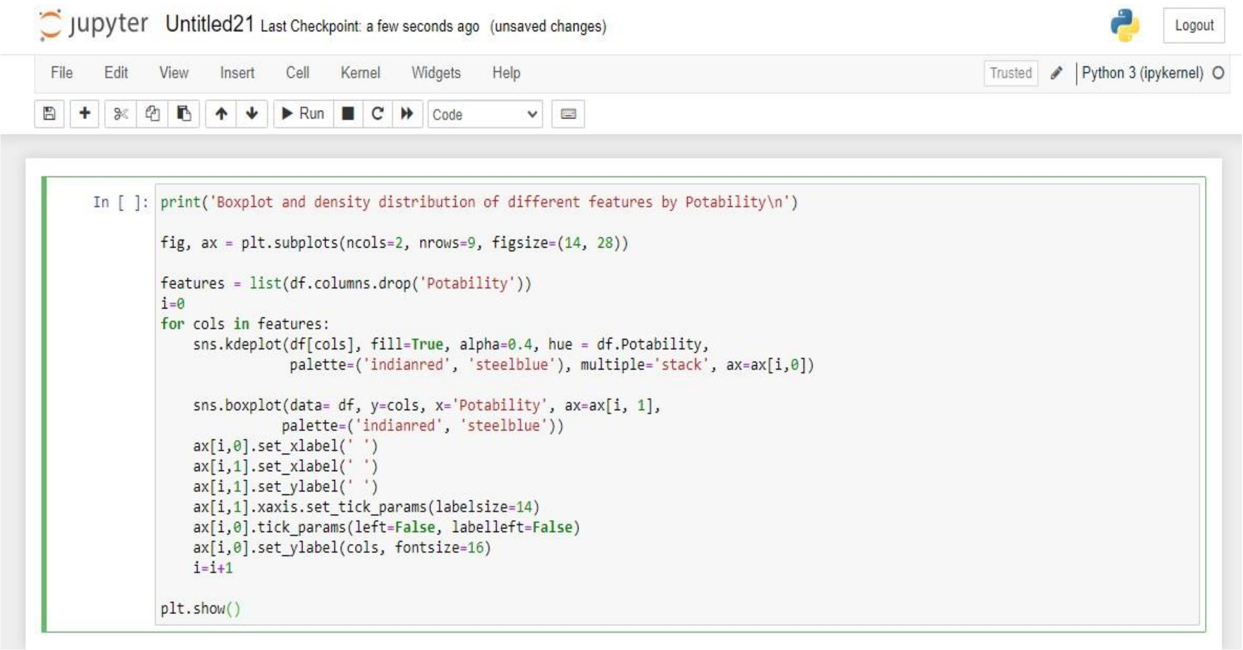
```
sns.boxplot(x='Parameter', y='Value', data=df)

plt.xlabel('Water Quality Parameter')

plt.ylabel('Value')

plt.title('Box Plot of Water Quality Parameters')

plt.xticks(rotation=45)

plt.show()
```


jupyter
Untitled20 Last Checkpoint: 10 minutes ago (unsaved changes)
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Run
Code
In [2]:
Box Plot of Water Quality Parameters
50000
40000
30000
20000
10000
0
Solids
0.0

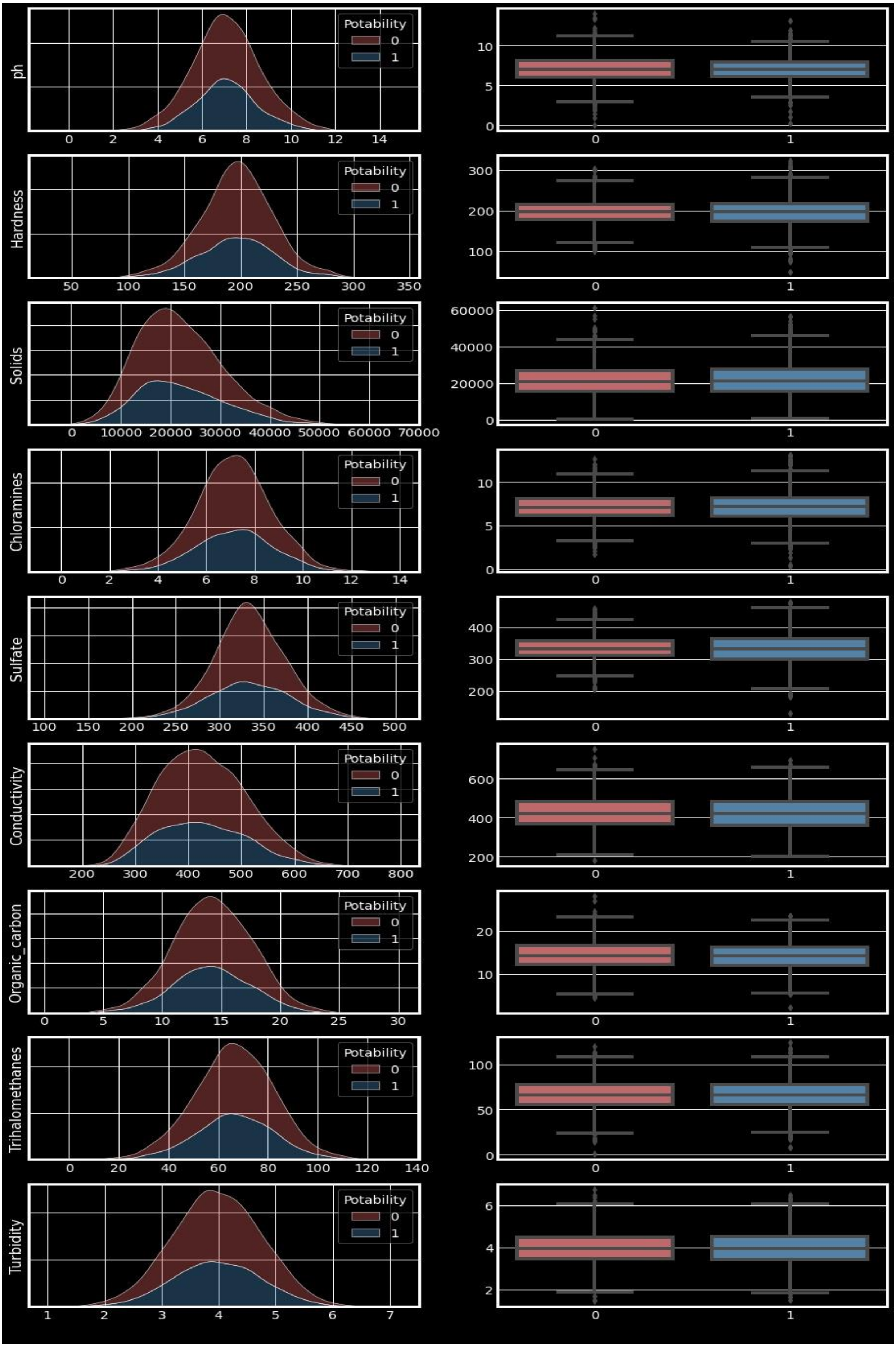
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
df = pd.read_csv('water_quality_data.csv')
sns.boxplot(x='ph', y='Solids', data=df)
plt.xlabel('ph')
plt.ylabel('Solids')
plt.title('Box Plot of Water Quality Parameters')
plt.xticks(rotation=45)
plt.show()
```

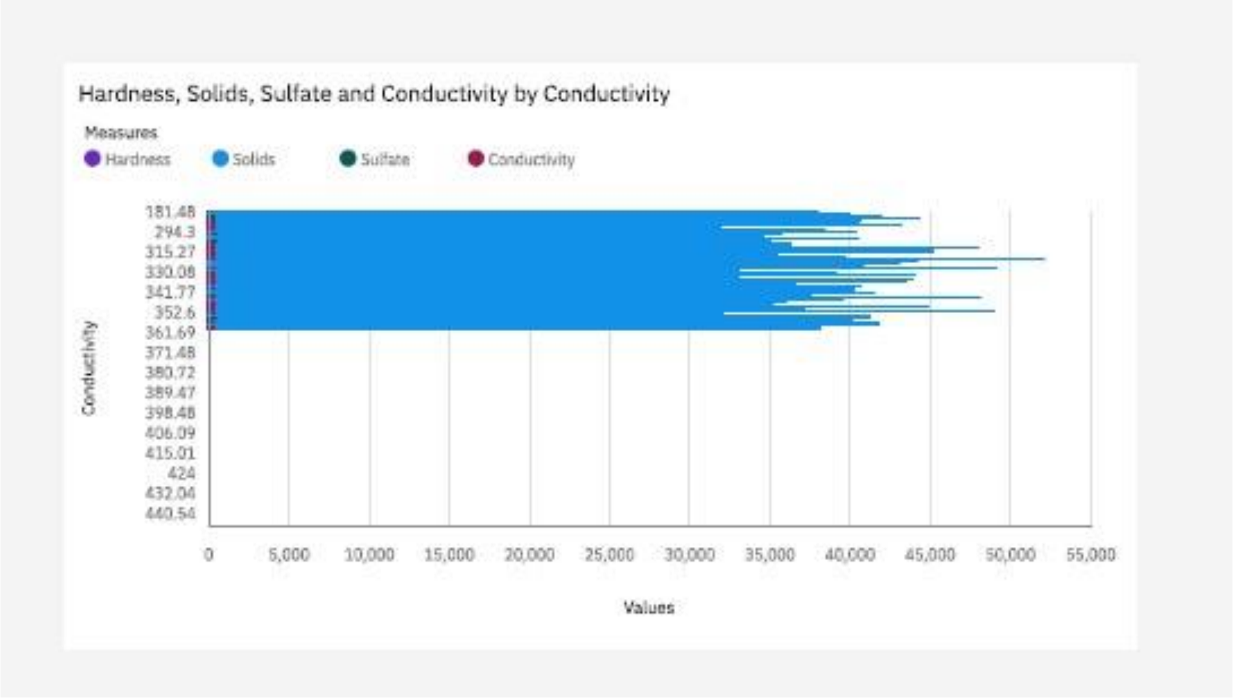
```
print('Boxplot and density distribution of different features by Potability\n')

fig, ax = plt.subplots(ncols=2, nrows=9, figsize=(14, 28))

features = list(df.columns.drop('Potability'))

 i=0

for cols in features:

 sns.kdeplot(df[cols], fill=True, alpha=0.4, hue = df.Potability,

palette=('indianred', 'steelblue'), multiple='stack', ax=ax[i,0])

sns.boxplot(data= df, y=cols, x='Potability', ax=ax[i, 1],

 palette=('indianred', 'steelblue'))
```

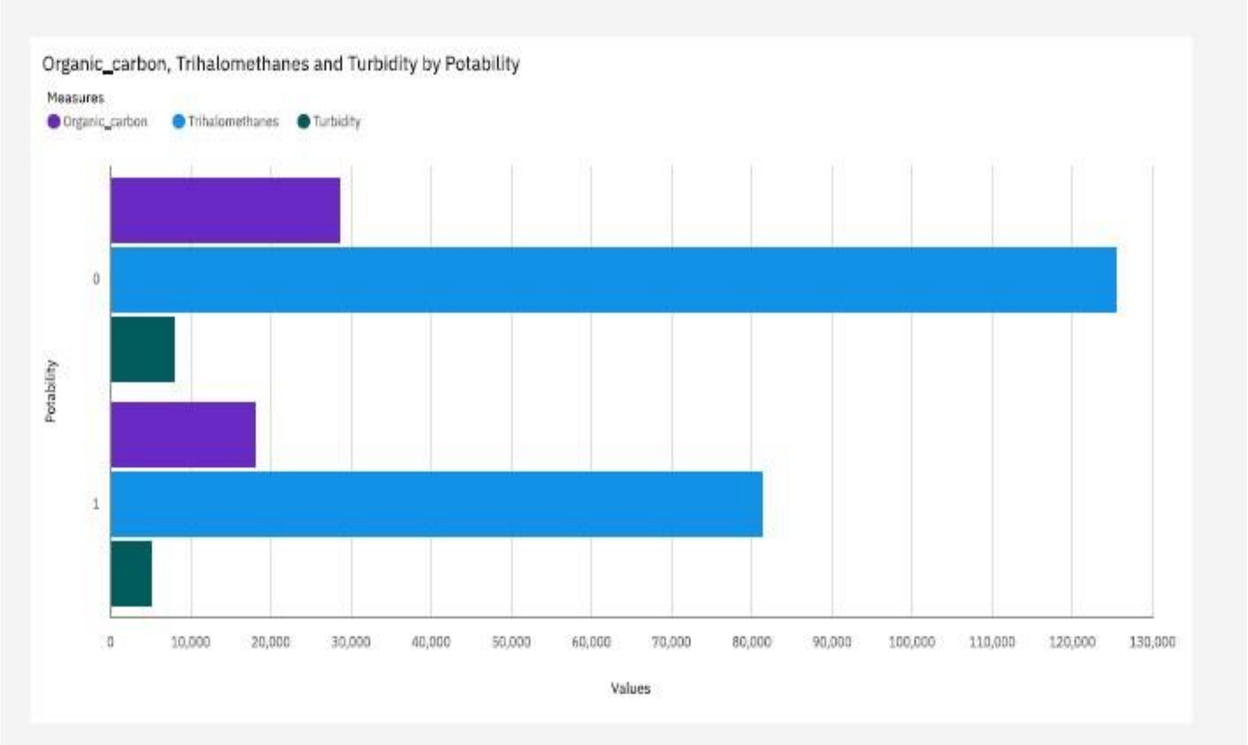
```
ax[i,0].set_xlabel(' ')
 ax[i,1].set_xlabel(' ')
 ax[i,1].set_ylabel(' ')
 ax[i,1].xaxis.set_tick_params(labelsize=14)
ax[i,0].tick_params(left=False, labelleft=False)
ax[i,0].set_ylabel(cols, fontsize=16)
 i=i+1
plt.show()
```



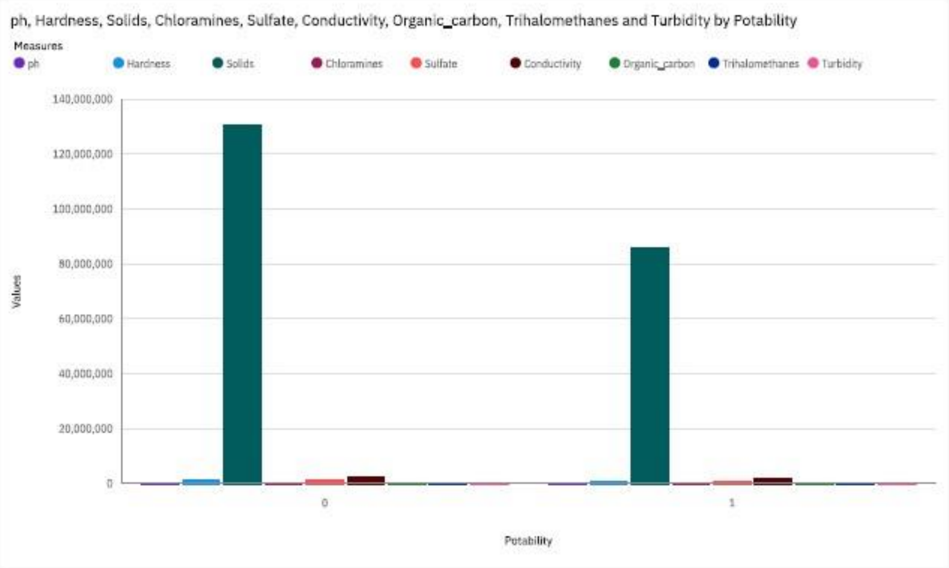
```
In [ ]: print('Boxplot and density distribution of different features by Potability\n')

        fig, ax = plt.subplots(ncols=2, nrows=9, figsize=(14, 28))

        features = list(df.columns.drop('Potability'))
        i=0
        for cols in features:
            sns.kdeplot(df[cols], fill=True, alpha=0.4, hue = df.Potability,
                        palette=('indianred', 'steelblue'), multiple='stack', ax=ax[i,0])

            sns.boxplot(data= df, y=cols, x='Potability', ax=ax[i, 1],
                       palette=('indianred', 'steelblue'))
            ax[i,0].set_xlabel(' ')
            ax[i,1].set_xlabel(' ')
            ax[i,1].set_ylabel(' ')
            ax[i,1].xaxis.set_tick_params(labelsize=14)
            ax[i,0].tick_params(left=False, labelleft=False)
            ax[i,0].set_ylabel(cols, fontsize=16)
            i=i+1

        plt.show()
```


ph
Potability
0
1
0 2 4 6 8 10 12 14
10 5 0
0 1
Hardness
50 100 150 200 250 300 350
300 200 100
Solids
0 10000 20000 30000 40000 50000 60000 70000
60000 40000 20000 0
Chloramines
0 2 4 6 8 10 12 14
10 5 0
Sulfate
100 150 200 250 300 350 400 450 500
400 300 200
Conductivity
200 300 400 500 600 700 800
600 400 200
Organic_carbon
0 5 10 15 20 25 30
20 10
Trihalomethanes
0 20 40 60 80 100 120 140
100 50 0
Turbidity
1 2 3 4 5 6 7
6 4 2

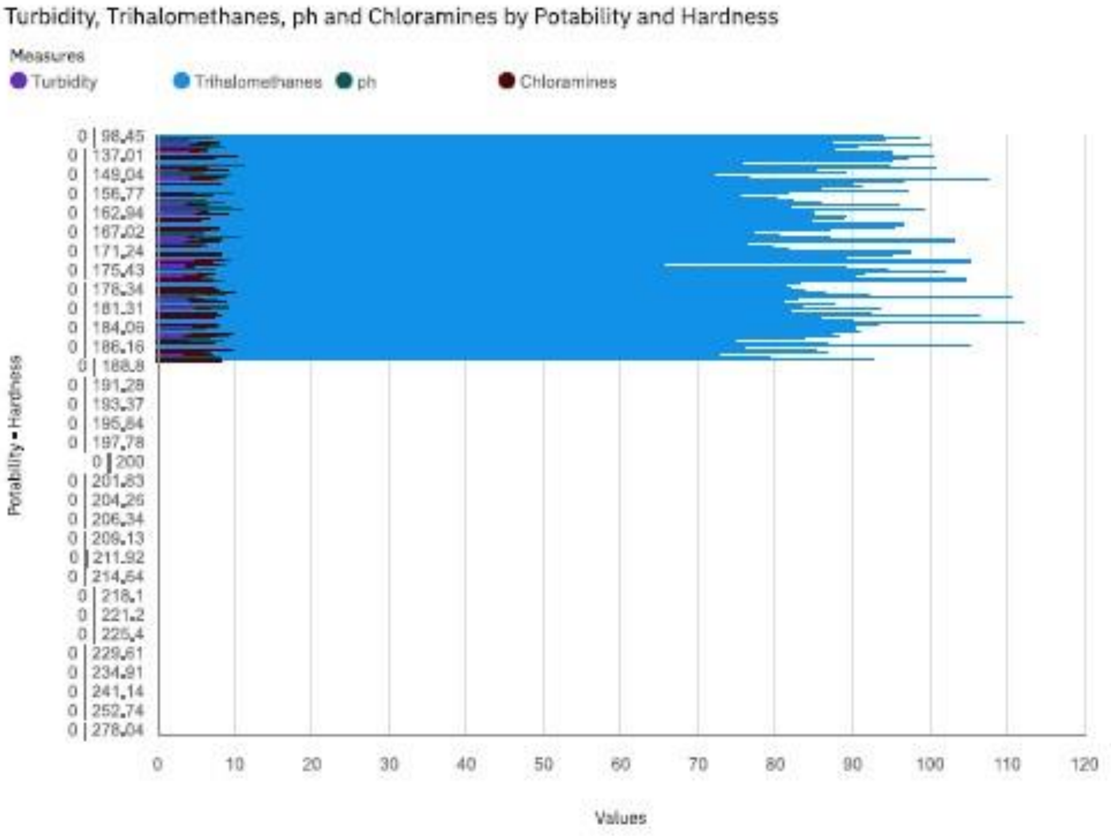

Hardness, Solids, Sulfate and Conductivity by Conductivity
Measures
● Hardness
● Solids
● Sulfate
● Conductivity
Conductivity
181.48
294.3
315.27
330.08
341.77
352.6
361.69
371.48
380.72
389.47
398.48
406.09
415.01
424
432.04
440.54
0
5,000
10,000
15,000
20,000
25,000
30,000
35,000
40,000
45,000
50,000
55,000
Values


Organic_carbon, Trihalomethanes and Turbidity by Potability
Measures
● Organic_carbon
● Trihalomethanes
● Turbidity
Potability
0
1
0
10,000
20,000
30,000
40,000
50,000
60,000
70,000
80,000
90,000
100,000
110,000
120,000
130,000
Values

Tab 1


ph, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic_carbon, Trihalomethanes and Turbidity by Potability
Measures
ph
Hardness
Solids
Chloramines
Sulfate
Conductivity
Organic_carbon
Trihalomethanes
Turbidity
140,000,000
120,000,000
100,000,000
80,000,000
60,000,000
40,000,000
20,000,000
0
Values
0
1
Potability

Tab 1


Turbidity, Trihalomethanes, ph and Chloramines by Potability and Hardness
Measures
Turbidity
Trihalomethanes
ph
Chloramines
0 | 98.45
0 | 137.01
0 | 149.04
0 | 156.77
0 | 162.94
0 | 167.02
0 | 171.24
0 | 175.43
0 | 178.34
0 | 181.31
0 | 184.06
0 | 186.16
0 | 188.8
0 | 191.28
0 | 193.37
0 | 195.84
0 | 197.78
0 | 200
0 | 201.83
0 | 204.26
0 | 206.34
0 | 209.13
0 | 211.92
0 | 214.64
0 | 218.1
0 | 221.2
0 | 225.4
0 | 229.61
0 | 234.91
0 | 241.14
0 | 252.74
0 | 278.04
Potability • Hardness
0
10
20
30
40
50
60
70
80
90
100
110
120
Values

Conclusion

- The Solid levels seem to contain some descripency since its values are on an average 40 folds more than the upper limit for safe drinking water.(Desirable limit for TDS is 500 mg/l and maximum limit is 1000 mg/l which prescribed for drinking purpose.)
- The data contains almost equal number of acidic and basic pH level water samples.
- The correlation coefficients between the features were very low.

Random Forest and XGBoost worked the best to train the model, both gives us f1 score (Balanced with precision & recall) as around 76%.