WATER QUALITY ANALYSIS

PHASE-1 DOCUMENT SUBMISSION

OBJECTIVE:

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.

Problem Definition and Design Thinking

To address this problem effectively, we need to:

- Define clear analysis objectives: Determine what specific aspects of water quality we want to assess and improve.
- Collect relevant water quality data: Gather data on parameters like pH, hardness, turbidity, etc.
- Design relevant visualizations: Create informative visualizations to understand data patterns.
- Build a predictive model: Develop a model to predict water potability based on the collected data.

ANALYSIS OBJECTIVES :

- The primary goal of this analysis is to assess whether the water meets regulatory standards and is safe for drinking and other specific purposes.
- Develop a predictive model that can classify water samples as potable or non-potable.
- Establish a threshold or criteria for potability based on relevant standards and guidelines (e.g., WHO or EPA standards).
- Use appropriate classification metrics (accuracy, precision, recall, F1-score) to evaluate the model's performance in predicting water potability.

2. Identifying Deviations from Regulatory Standards

- Detect and quantify any deviations of water quality parameters from established regulatory standards.
- Identifying deviations from standards is essential for pinpointing potential issues with water quality and understanding areas that may require corrective actions.
- Compare the measured values of water quality parameters (e.g., pH, hardness, turbidity) to the established regulatory standards.

- Create visualizations and reports that highlight instances where parameters deviate from standards.
- Quantify the extent of deviations and prioritize them based on their impact on water quality.

3. Understanding Parameter Relationships

- Explore and analyze the relationships and dependencies between different water quality parameters.
- Understanding how different parameters interact can provide insights into the underlying factors affecting water quality.
- Perform correlation analysis to identify pairs of parameters that are strongly related.
- Create visualizations such as scatterplots or heatmaps to visualize the relationships.
- Consider conducting statistical tests to validate the significance of observed relationships.

4. Providing Recommendations for Improvement

- Generate recommendations and insights based on the analysis to improve water quality and ensure potability.
- The analysis should not only identify issues but also provide actionable recommendations for stakeholders to address water quality concerns.
- Summarize the findings and insights from the analysis.
- Offer recommendations for potential corrective actions or interventions to bring water quality within acceptable limits.
- Highlight the potential benefits and risks associated with each recommendation.

5. Continuous Monitoring and Reporting

- System for continuous monitoring of water quality and regular reporting of findings.
- Water quality is an ongoing concern, and continuous monitoring ensures that deviations are detected promptly, and corrective actions can be taken in a timely manner.
- Implement a monitoring system that collects and analyzes new data regularly.
- Set up automated alerts for significant deviations from standards.
- Generate periodic reports summarizing the current state of water quality and any changes over time.

Data Collection

Data Source:

- We need to identify the source of the water quality data. It's essential to ensure that the data source is reliable and relevant to the objectives of our analysis. Potential sources include:
- Government agencies responsible for monitoring water quality.
- Environmental organizations that collect and maintain water quality data.
- Publicly available datasets from reputable sources.
-

Data Attributes-We should compile a list of all the data attributes (parameters) that are available in the dataset. These attributes may include:

- pH (acidity or alkalinity of water)
- Hardness (concentration of calcium and magnesium ions)
- Solids (total dissolved solids in water)
- Chlorides (concentration of chloride ions)
- Sulfates (concentration of sulfate ions)
- Turbidity (cloudiness or haziness of water)
- Conductivity (ability of water to conduct electrical current)
- Organic Carbon (carbon content in the form of organic compounds)
- Trihalomethanes (THMs, a group of organic compounds)
- ... and other relevant attributes.

IMPLEMENTATION

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
df=pd.read_csv(r'/content/water_potability.csv')
df.head()
```

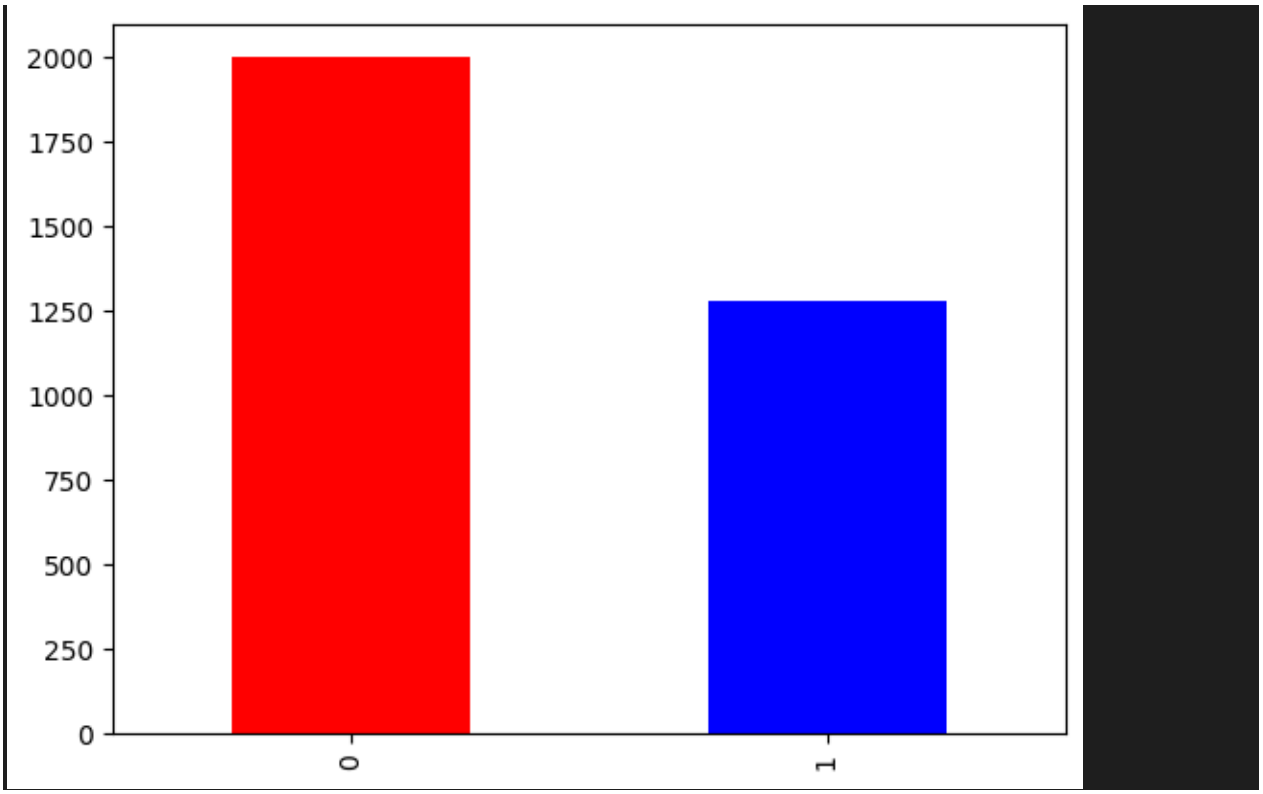
	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
0	NaN	204.8	20791.3	7.30	368.516	564.3086	10.379783	86.99	2.963	0

ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability	
		90455	18981	0212	441	54		0970	135	
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

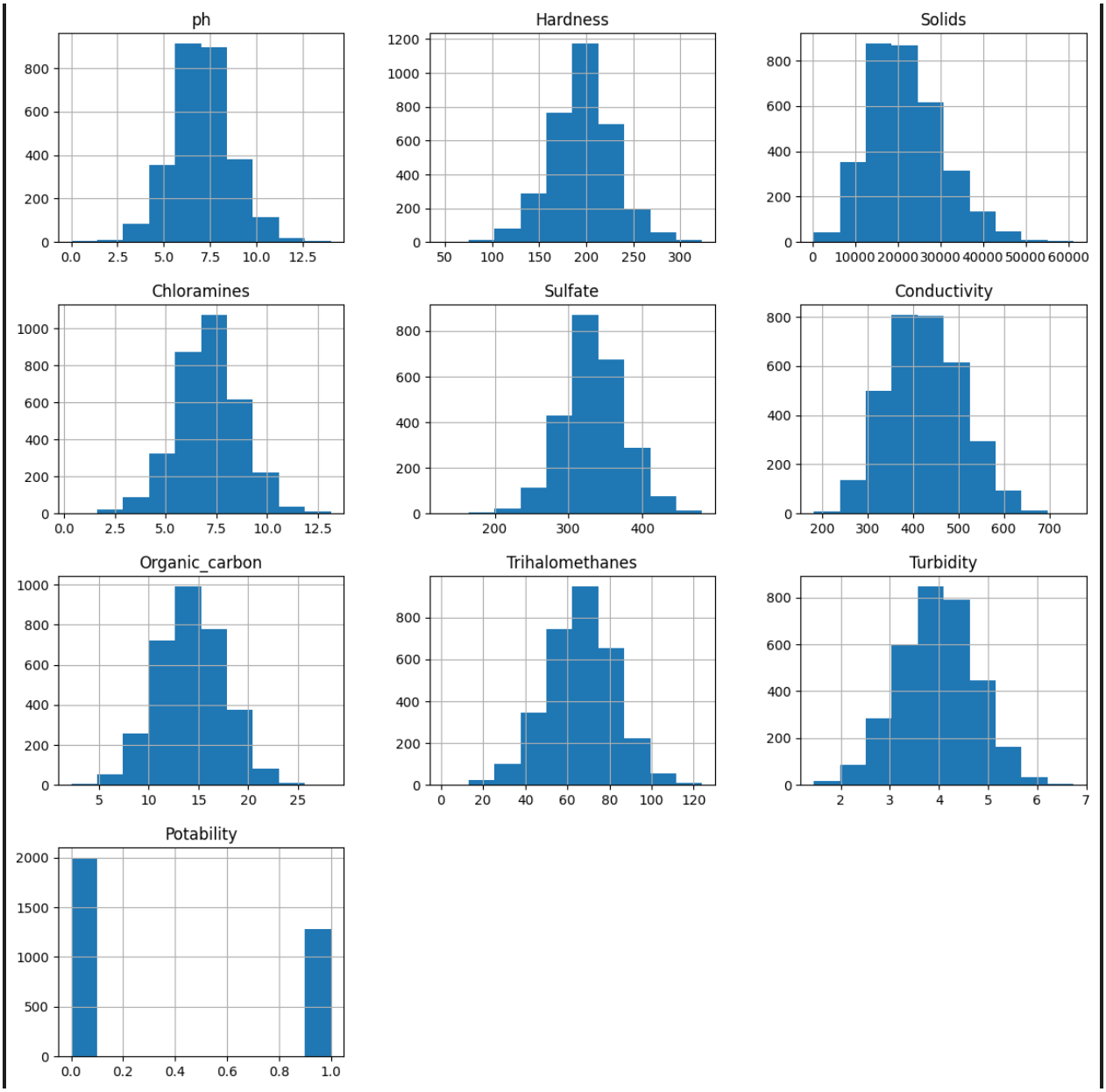
```
df.shape
df.isnull().sum()
df.info()
df.describe()
df['Sulfate'].mean()
```

333.7757766108135

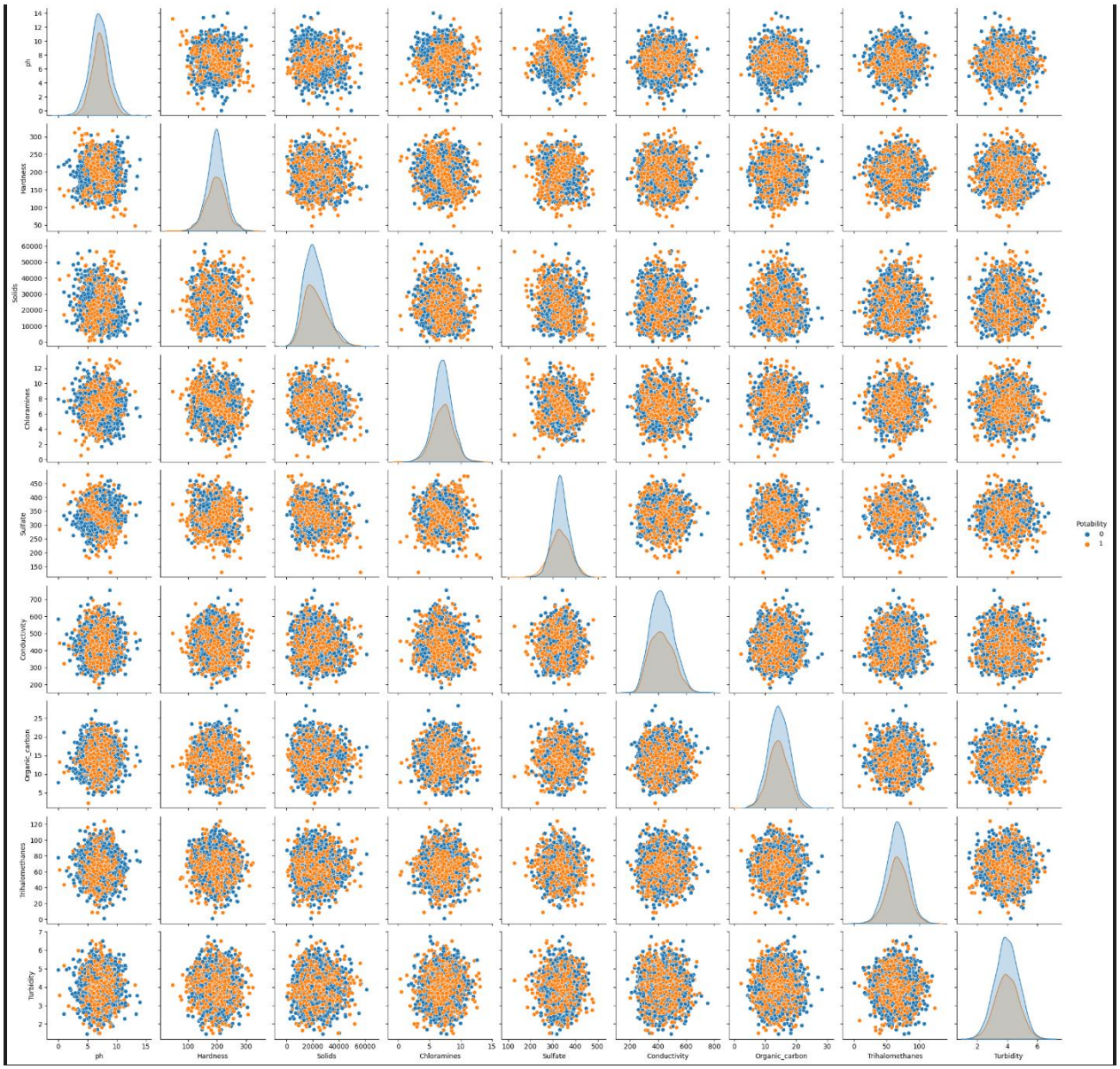
```
df.Potability.value_counts()
df.Potability.value_counts().plot(kind="bar", color=["red", "blue"])
plt.show()
sns.distplot(df['ph'])
```




```
df.hist(figsize=(14,14))
plt.show()
```




```
sns.pairplot(df,hue='Potability')
```



PREDICTIVE MODELING FOR WATER POTABILITY

In the project of analyzing water quality data to predict water potability, selecting appropriate machine learning algorithms and features is crucial for building an effective predictive model. Here, we'll discuss the choice of machine learning algorithms and features:

1. Machine Learning Algorithms

- Logistic Regression- is a straightforward and interpretable algorithm for binary classification problems like predicting water potability. It's a good starting point and can serve as a baseline model.
- Decision Trees-can capture non-linear relationships between features and the target variable. They are easy to interpret and can handle both numerical and categorical features.
- Random Forests-are an ensemble method that combines multiple decision trees to improve predictive accuracy and reduce overfitting. They are robust and can handle high-dimensional datasets.
- Support Vector Machines (SVM)-SVM is effective for binary classification tasks and can handle both linear and non-linear data. It works well with high-dimensional feature spaces.
- Neural Networks-Deep learning models, such as neural networks, can capture complex patterns in the data. They are suitable for tasks with a large number of features but may require more data and computational resources.

Feature Selection Selecting the right features is crucial for model performance. We need to identify which water quality parameters (features) are most relevant for predicting water potability. Feature selection techniques may include:

- Feature Scaling: Normalize or standardize numerical features to ensure they have similar scales.
- One-Hot Encoding: Convert categorical features (if any) into binary variables for modeling.
- Interaction Terms: Create interaction terms between pairs of features if there's reason to believe that their combination affects potability.
- Feature Aggregation: Aggregate data over time intervals if time-series data is available.

Model Evaluation

After implementing machine learning algorithms and feature selection/engineering, it's essential to evaluate the models' performance. We can use metrics such as accuracy, precision, recall, F1-score, and ROC-AUC to assess how well each model predicts water potability. Cross-validation can also help in estimating model generalization performance.