## **WATER QUALITY ANALYSIS**

**TEAM MEMBER**

**721221243045 : PRAVEEN KANTH D**

**PHASE-4 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.

**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 | 204.890455 | 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 |

df.shape

df.isnull().sum()

df.info()

df.describe()

df['Sulfate'].mean()

333.7757766108135

**DECISION TREE**

**from** sklearn.tree **import** DecisionTreeClassifier

**from** sklearn.metrics **import** accuracy\_score,confusion\_matrix,precision\_score

dt**=**DecisionTreeClassifier(criterion**=** 'gini', min\_samples\_split**=** 10, splitter**=** 'best')

dt**.**fit(X\_train,Y\_train)

DecisionTreeClassifier(min\_samples\_split=10)

prediction**=**dt**.**predict(X\_test)

accuracy\_dt**=**accuracy\_score(Y\_test,prediction)**\***100

accuracy\_dt

58.84146341463414

print("Accuracy on training set: {:.3f}"**.**format(dt**.**score(X\_train, Y\_train)))

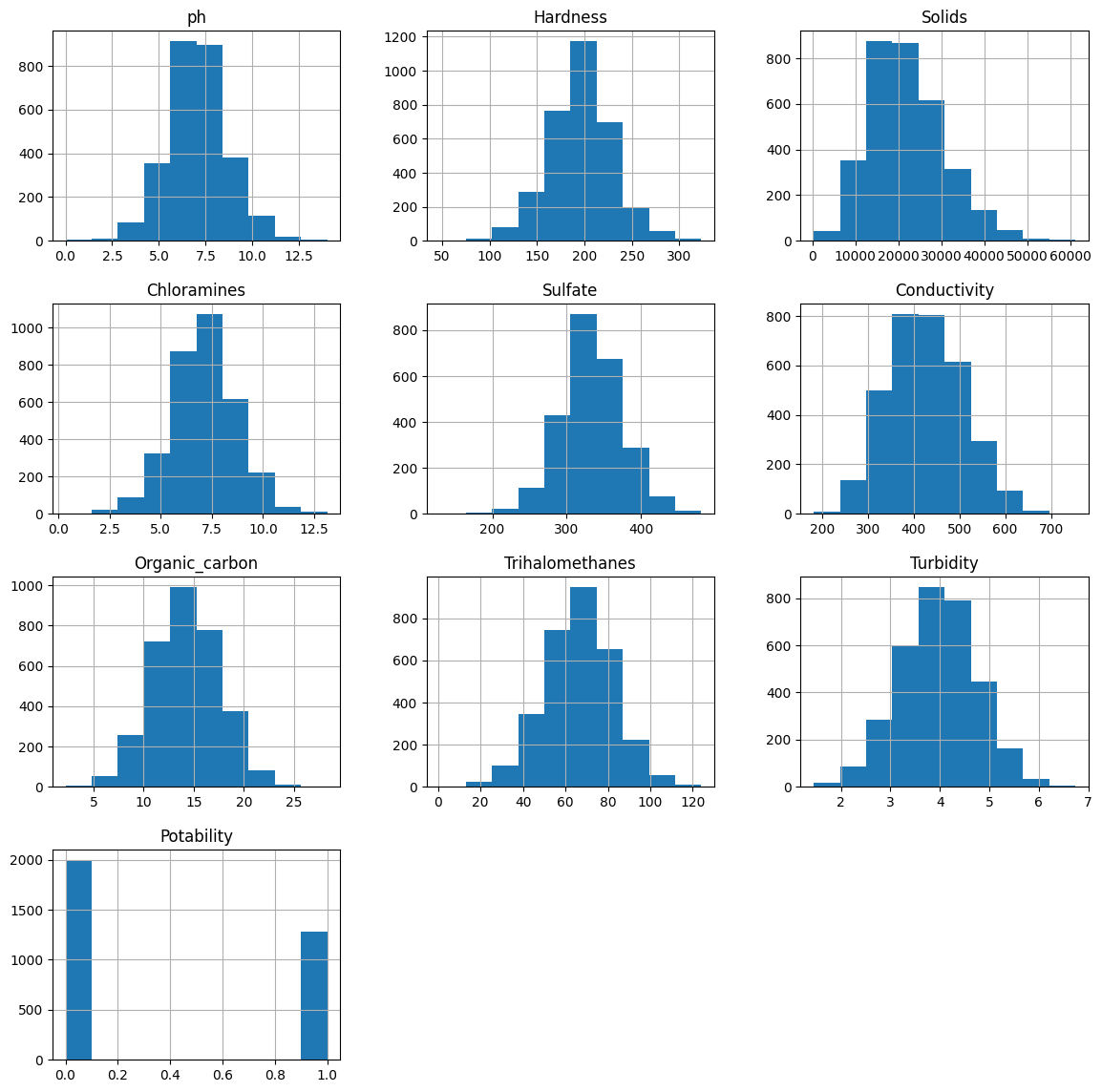
print("Accuracy on test set: {:.3f}"**.**format(dt**.**score(X\_test, Y\_test)))

Accuracy on training set: 0.923

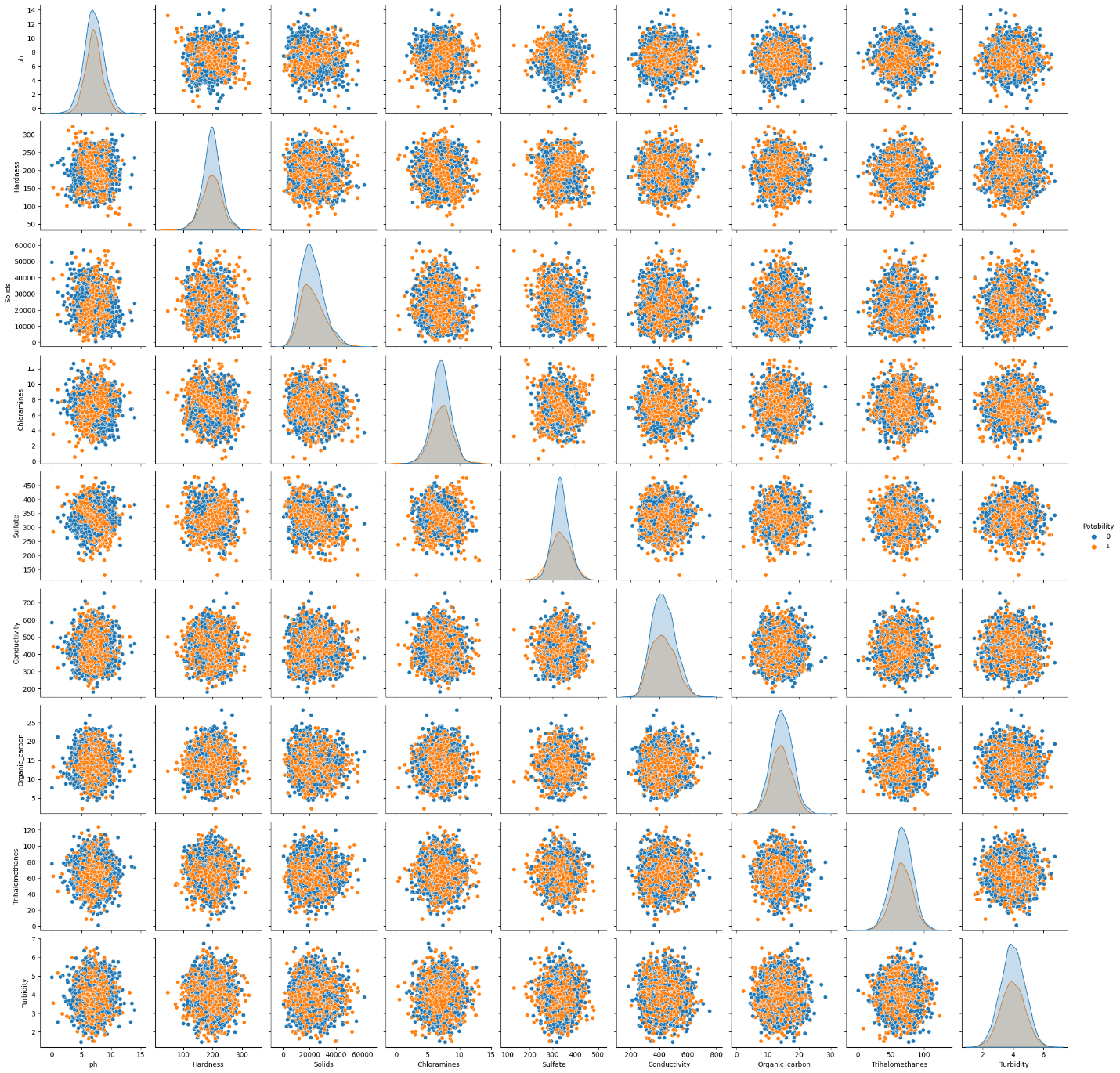
Accuracy on test set: 0.588

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.