**Water Quality Analysis**

# Objective:

To Access to safe drinking-water is essential to health, a basic human right and a component of effective policy for health protection. This is important as a health and development issue at a national, regional and local level. In some regions, it has been shown that investments in water supply and sanitation can yield a net economic benefit, since the reductions in adverse health effects and health care costs outweigh the costs of undertaking the interventions. comprehensively evaluate water quality by assessing its potability, detecting deviations from established standards, and explaining the interrelationships among key parameters.

# Data Processing Procedure:

To create a data visualization and perform machine learning analysis on a water quality dataset using Python libraries like pandas, NumPy, matplotlib, seaborn, and scikit-learn for decision tree classification, follow these steps:

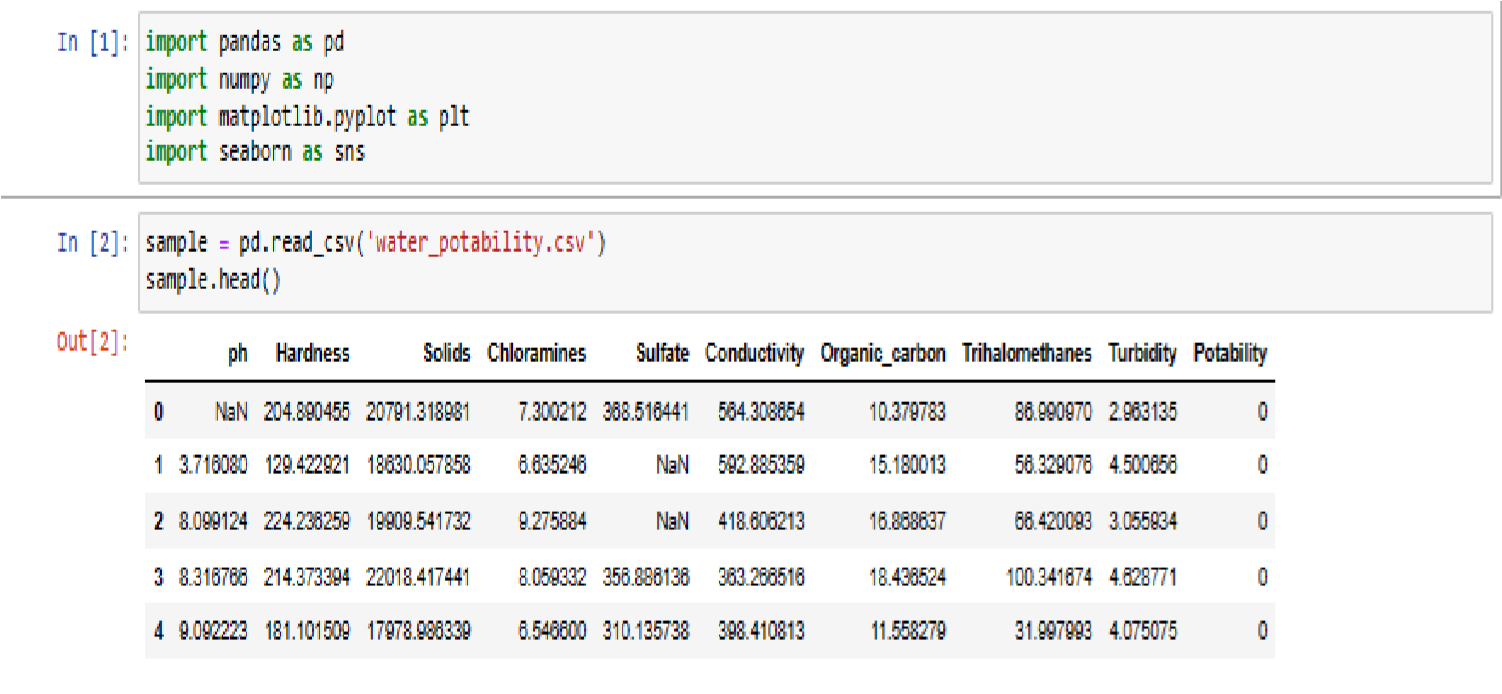
1. **Data Collection and Pre-processing:** Obtain the water quality dataset with relevant parameters. Import necessary Python libraries: pandas, NumPy, matplotlib, seaborn, and scikit-learn. Load and pre-process the dataset using pandas to handle missing values, outliers, and data cleaning.
2. **Data Visualization:** Utilize matplotlib and seaborn to create visualizations of the data. Some common plots include scatter plots, histograms, and box plots to understand data distributions.
3. **Correlation Analysis:** Use pandas to calculate the correlation between different parameters in the dataset. Create correlation matrices and visualize them using heatmap plots from seaborn to identify relationships between variables.
4. **Machine Learning Preparation:** Select the target variable (e.g., water quality class) and features (water quality parameters) for the machine learning model. Split the dataset into training and testing sets.

5.**Decision Tree Classifier:** Train a decision tree classifier using the Decision Tree Classifier from scikit-learn. Fit the model to the training data and evaluate its performance on the testing data.

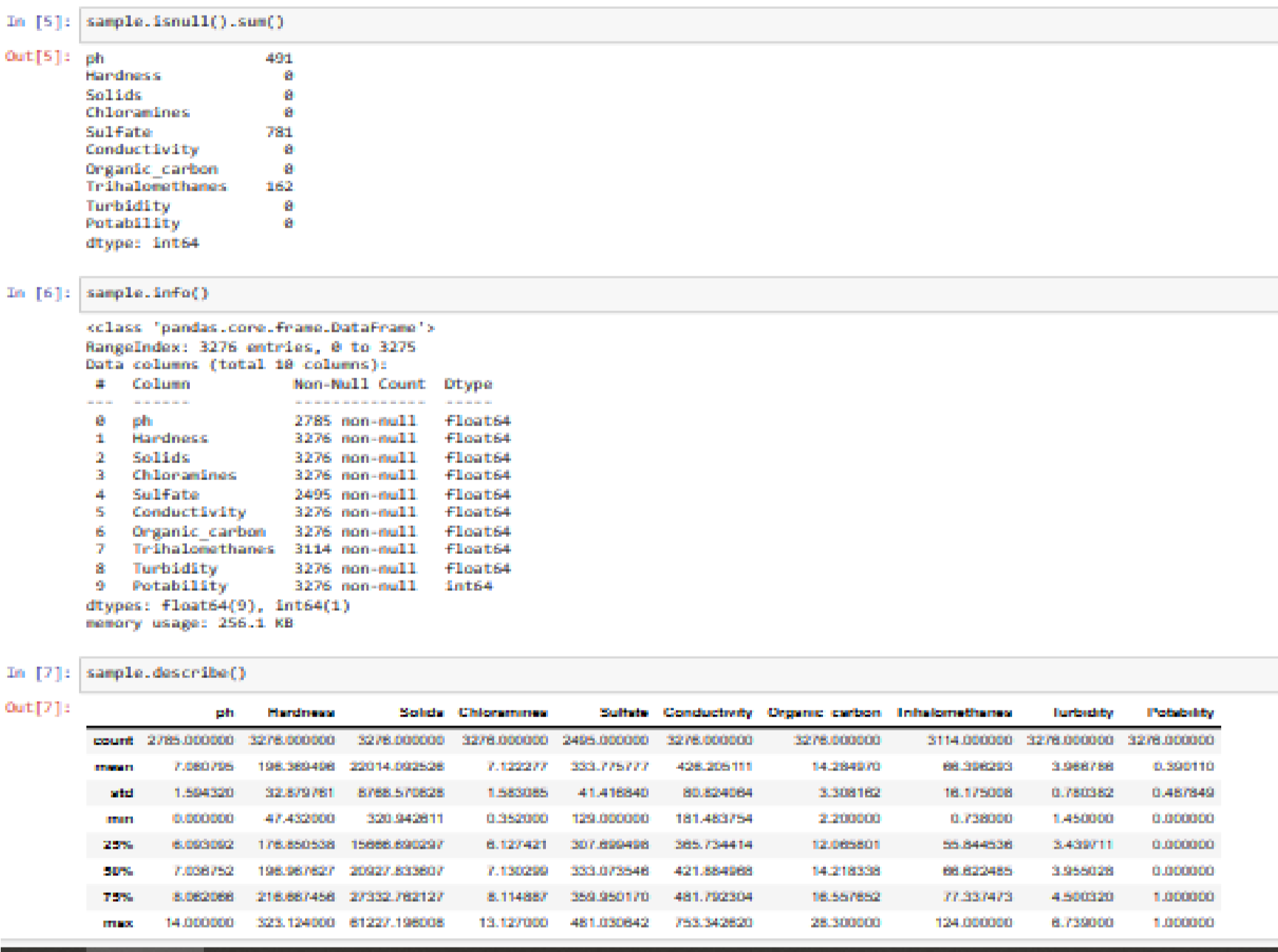
1. **Model Evaluation:** Calculate performance metrics such as accuracy, precision, recall, and F1-score to assess the model's classification performance.
2. **Visualization of Decision Tree:** Visualize the decision tree structure using tools provided by scikit-learn, such as the plot tree function.
3. **Interpretation:**  Analyse the decision tree to understand the rules and features that are most important for classification.
4. **Further Analysis:** We can also explore feature importance’s and feature selection techniques to improve the model. Summarize the findings and provide insights into water quality based on the analysis. Remember to consult the documentation for each library to get a detailed understanding of the functions and methods used. This is a high-level overview of the process, and you may need to adapt it to the specifics of your dataset and analysis goals.

# Operations:

* Importing python lib function and given data set:



* Identifying number of null values and calculating mean, max, min and data type:



Identifying null value and replacing it with the mean value for accuracy and calculating the patabality count for future purpose:

•

After replacement the

null

data

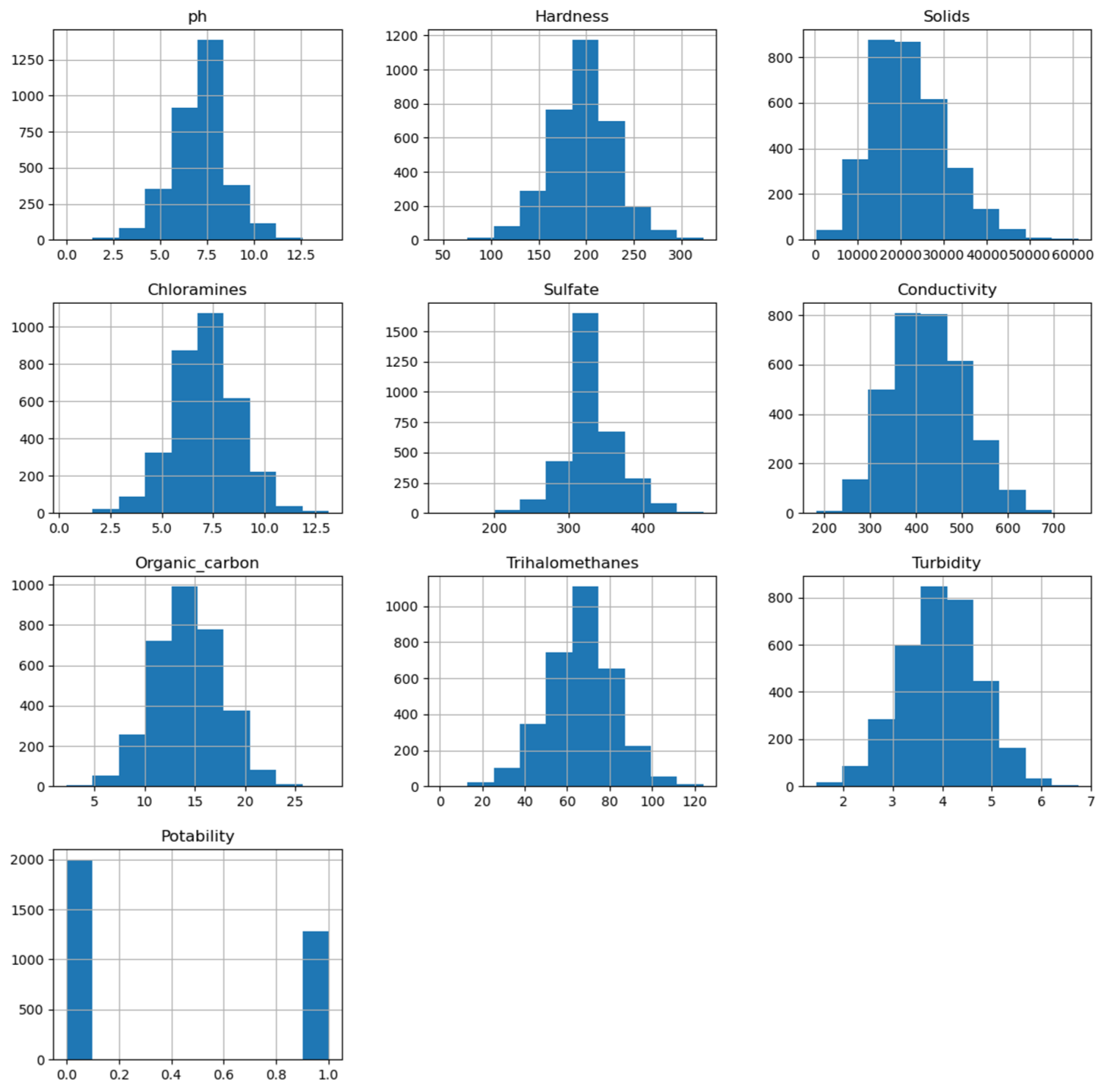
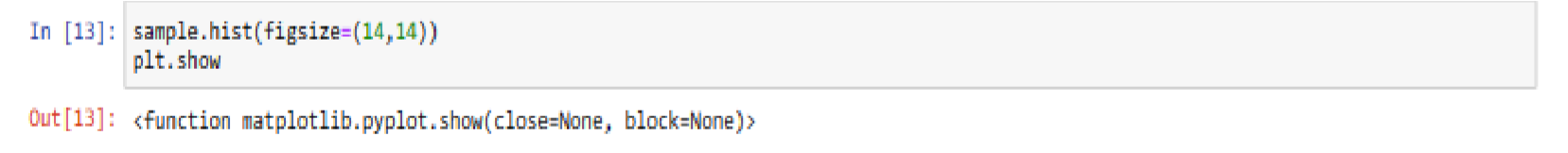
the data set i

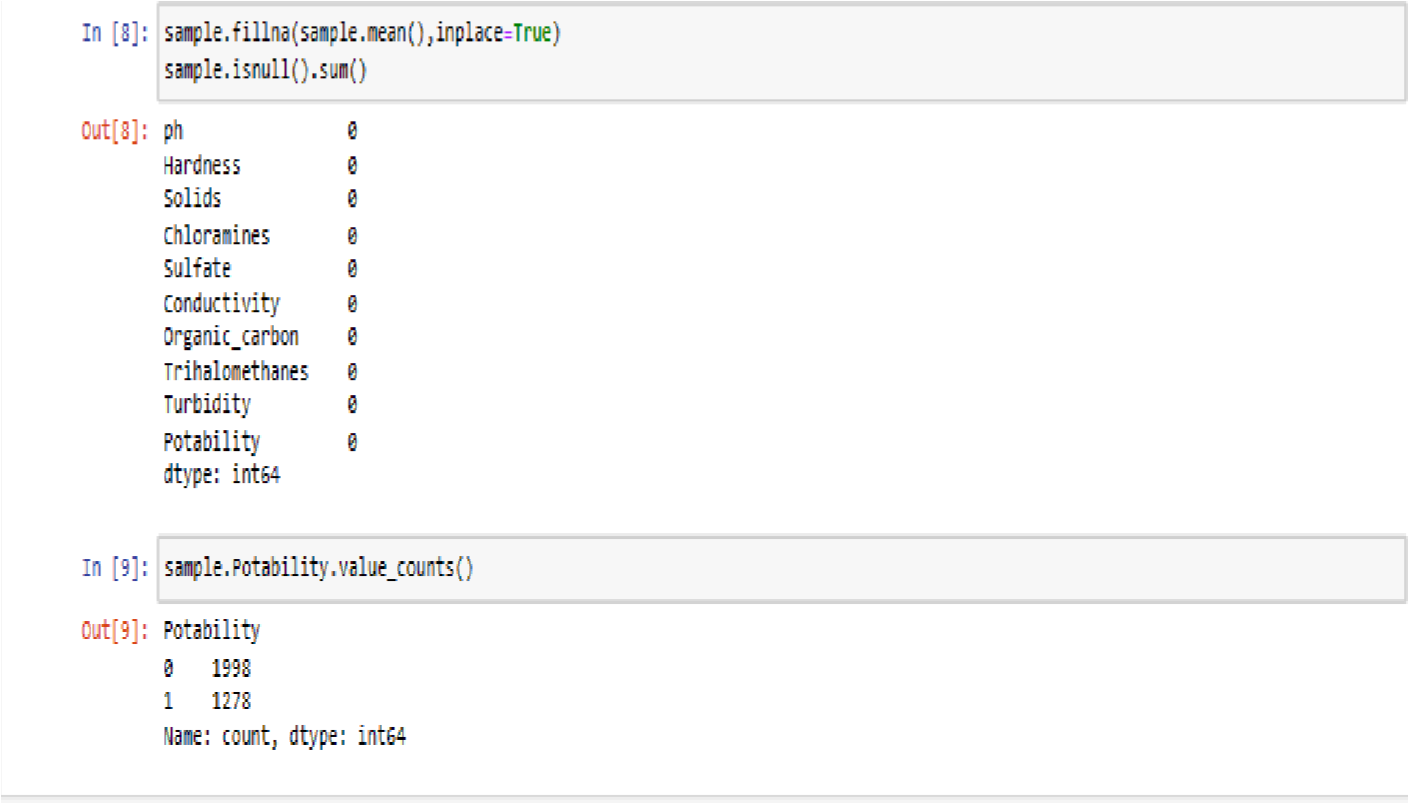
s

plotted

individually

for analysis:





Visualising the corelation (in order to reduce the dimension)of all the data using the function

heat map

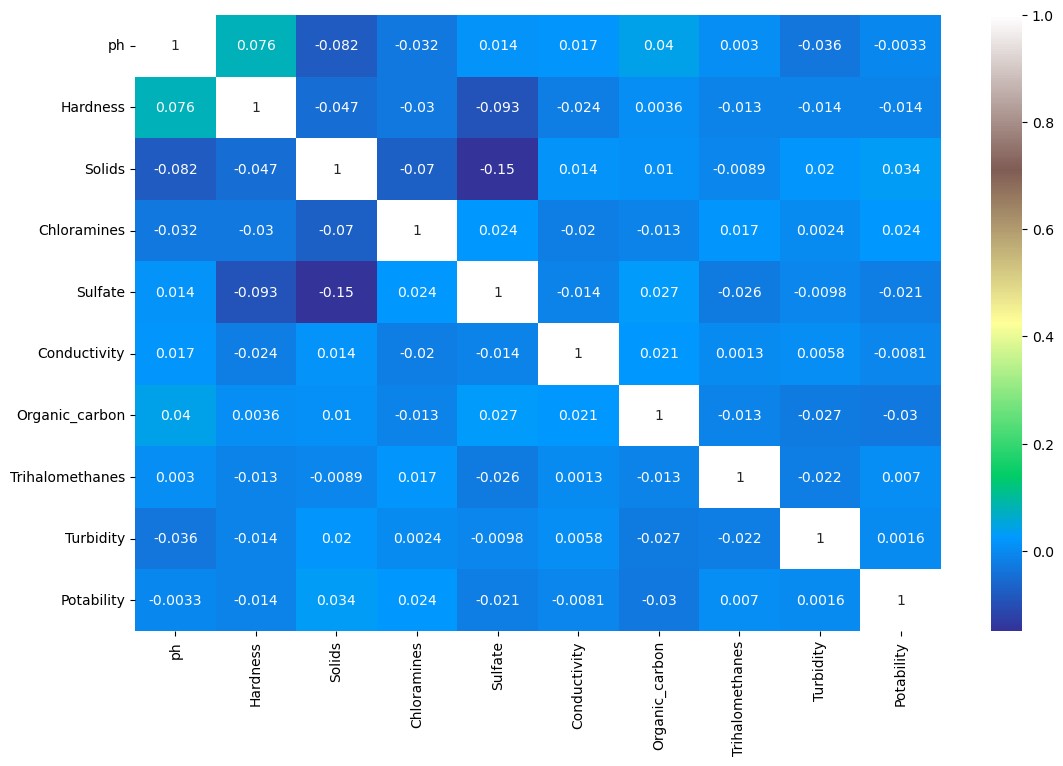
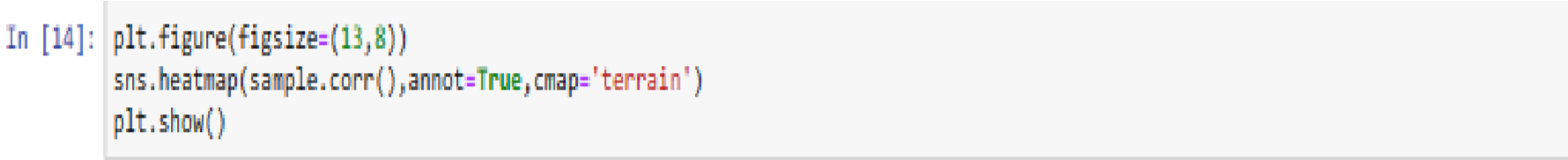
():

(

There is no corelation between the

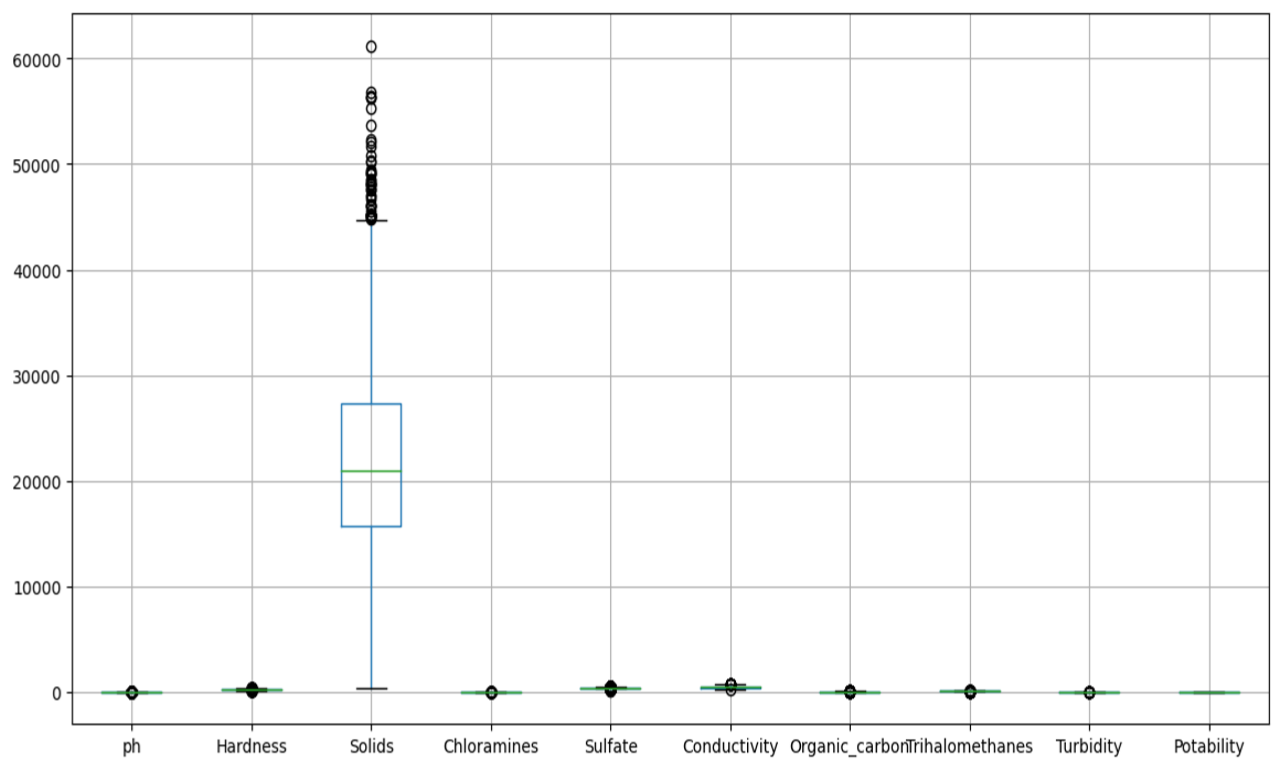
data’s

)

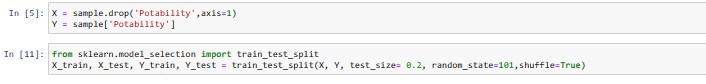


* Exploring the data using box plot to show the importance of not removing the solid data from graph set:

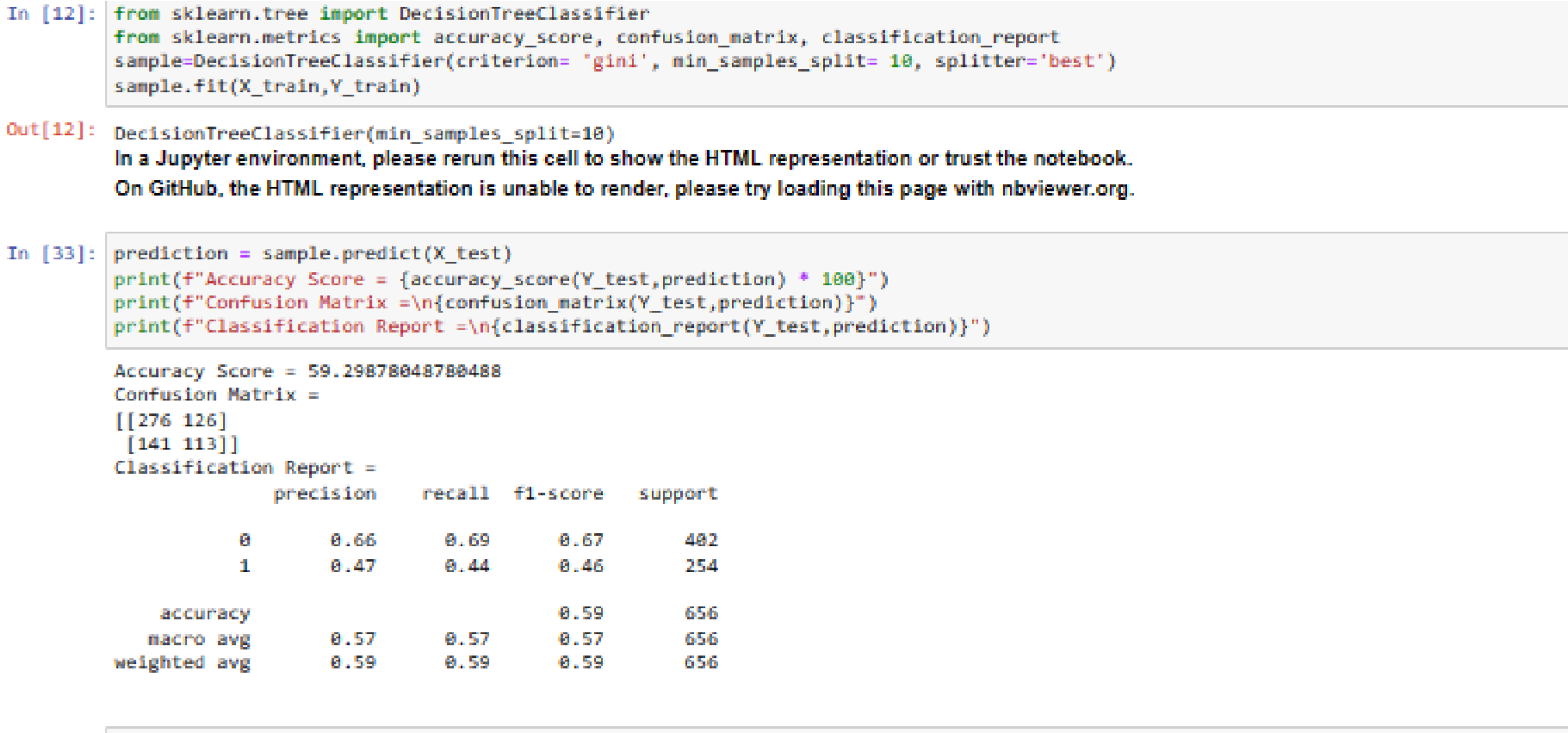
Program: sample.boxplot(figsize=(14,7))



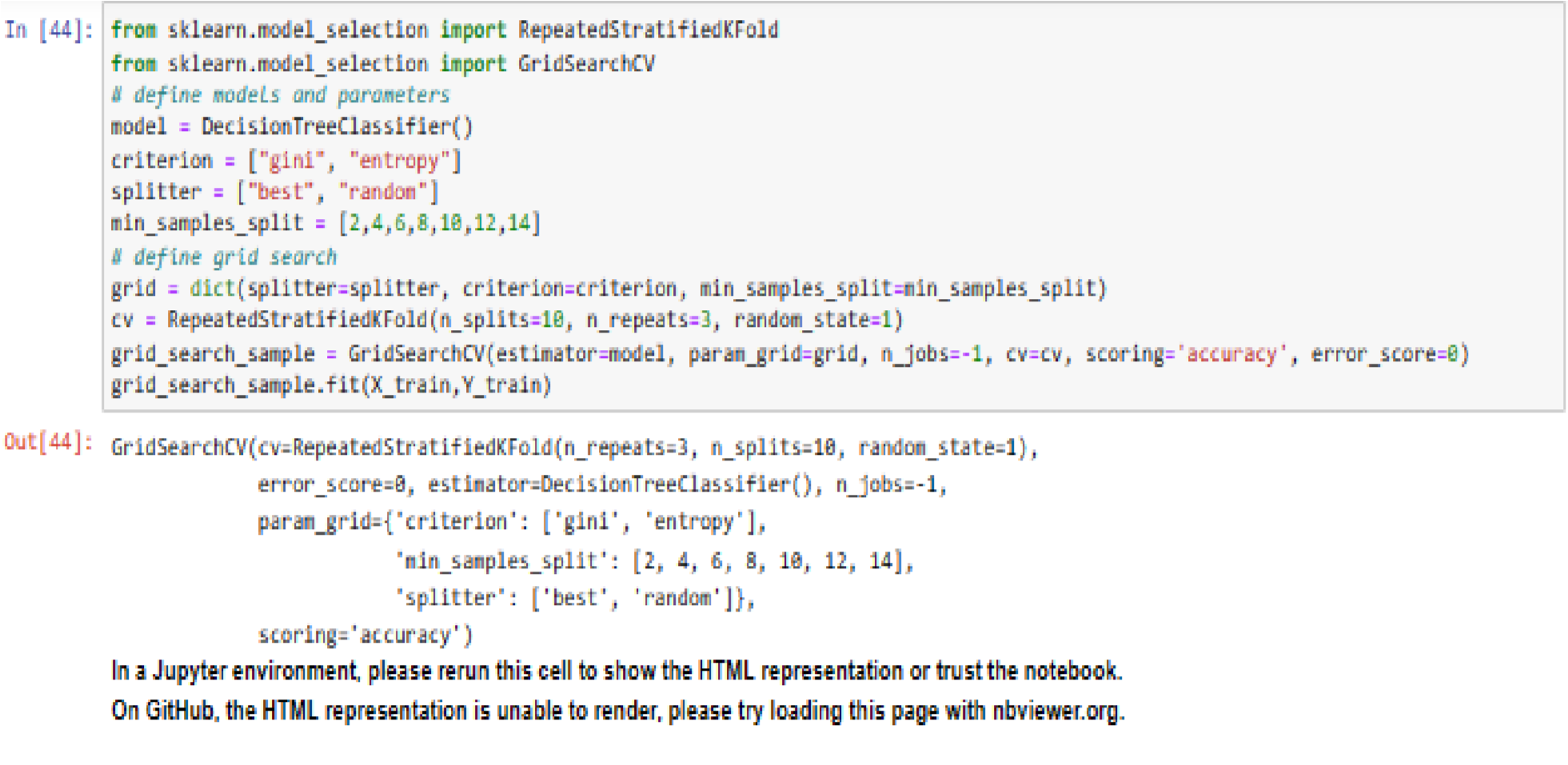
Now its time to prepare the data set, divide the data set into independent and dependent features. Then X contains all the independent features except the potability and Y contain the feature our target potability. Then we are splitting the data set into training and testing using train test function with some parameter like X, Y, test size, randomise to ignore the shuffle every time.



* Train decision Tree classifier and check accuracy:



Apply Hyper parameter tuning:(creation of dictionary for given data set)



Checking the created model for given datas and accuracy:



Training Score: 80.19083969465649

Testing Score: 58.079268292682926

**Conclusion:**

Phase 3 has provided vital insights into our water quality dataset, paving the way for an advanced, fine-tuned model in the upcoming phases. Our journey toward better understanding and managing water quality continues, driven by data analysis, visualization, and machine learning. Looking ahead, our next phase will involve refining the model, exploring feature importance, and optimizing it for real-world use. We'll also explore ensemble methods and hyperparameter tuning to enhance accuracy and robustness.