Title: Water Potability Prediction Analysis 💧

Subtitle: Understanding Water Quality Issues Through Data

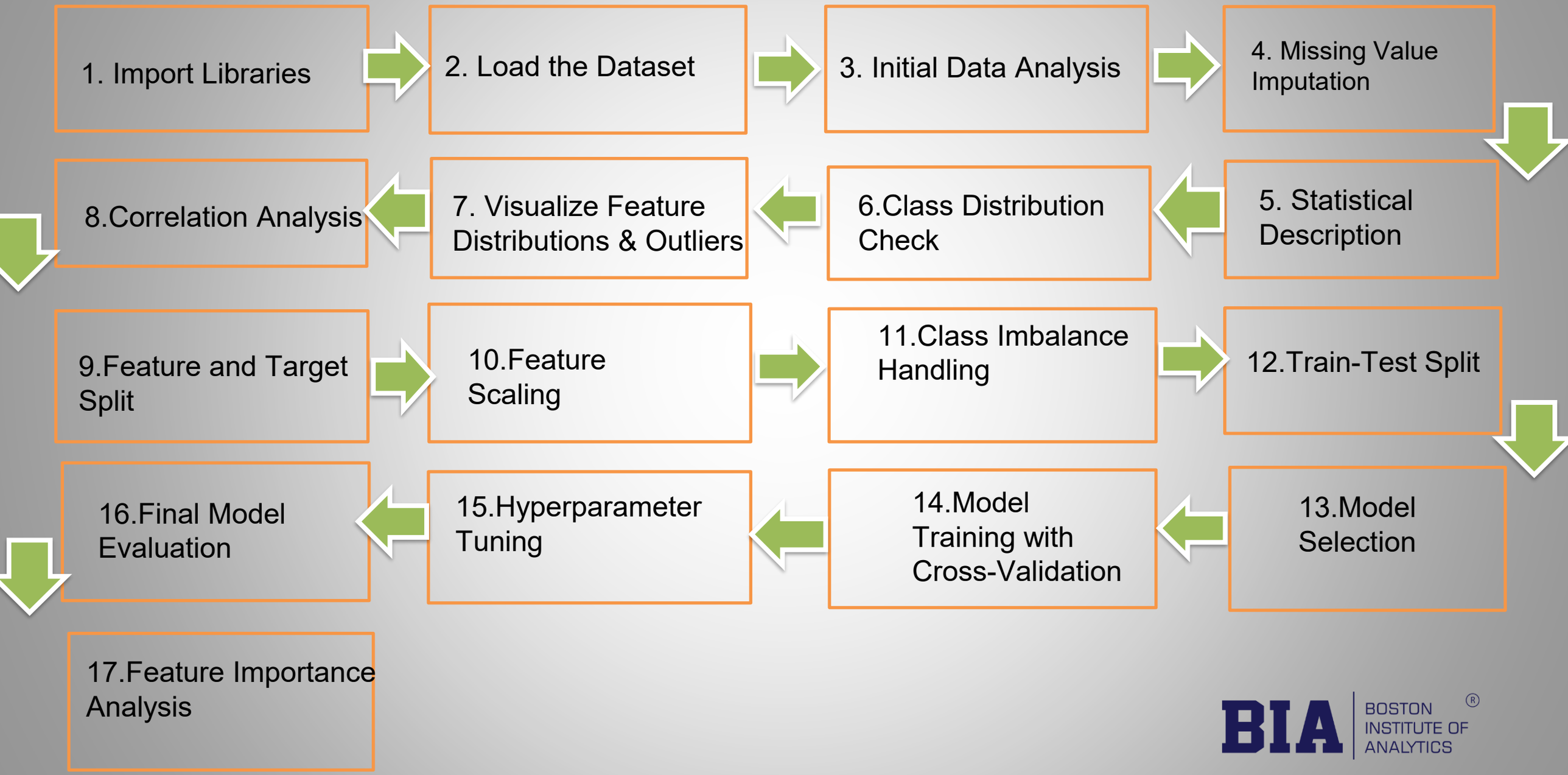
BIA | BOSTON INSTITUTE OF ANALYTICS ®

Why I Chose This Project 💧

Water is essential for life, yet millions of people worldwide still lack access to safe drinking water.
In many regions, especially in India, water contamination remains a serious problem causing diseases and deaths.
Traditional testing methods are slow, expensive, and not easily accessible in rural or remote areas.
By applying Data Science and Machine Learning, we can predict water quality efficiently and quickly, helping ensure safe drinking water for everyone.
This project combines social impact and technical learning, making it both meaningful and educational.

BIA BOSTON INSTITUTE OF ANALYTICS ®

Methodology



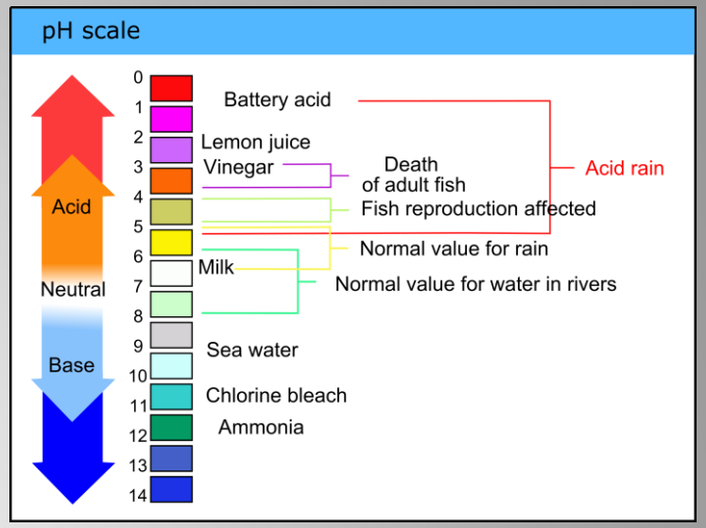
BIA BOSTON INSTITUTE OF ANALYTICS ®

Introduction

- **Samples:** 3,276 water samples analyzed
- **Goal:** Predict water potability (safe or unsafe)
- **Approach:** Data analysis & machine learning
- **Target Variable:**
  - 1 → Potable
  - 0 → Non-potable

Objective

- Identify key factors influencing water potability
- Build an accurate ML model to predict drinking safety
- Support data-driven water quality monitoring



DATASET FEATURES OVERVIEW

Each feature represents a physical or chemical property of water:

- **pH:** Acidity or alkalinity (ideal 6.5–8.5)
- **Hardness:** Calcium & magnesium content
- **Solids (TDS):** Total dissolved solids
- **Chloramines:** Disinfectant chemical
- **Sulfate:** Natural mineral content
- **Conductivity:** Ion concentration indicator
- **Organic Carbon:** Organic matter in water
- **THMs:** Byproduct of chlorination
- **Turbidity:** Clarity of water

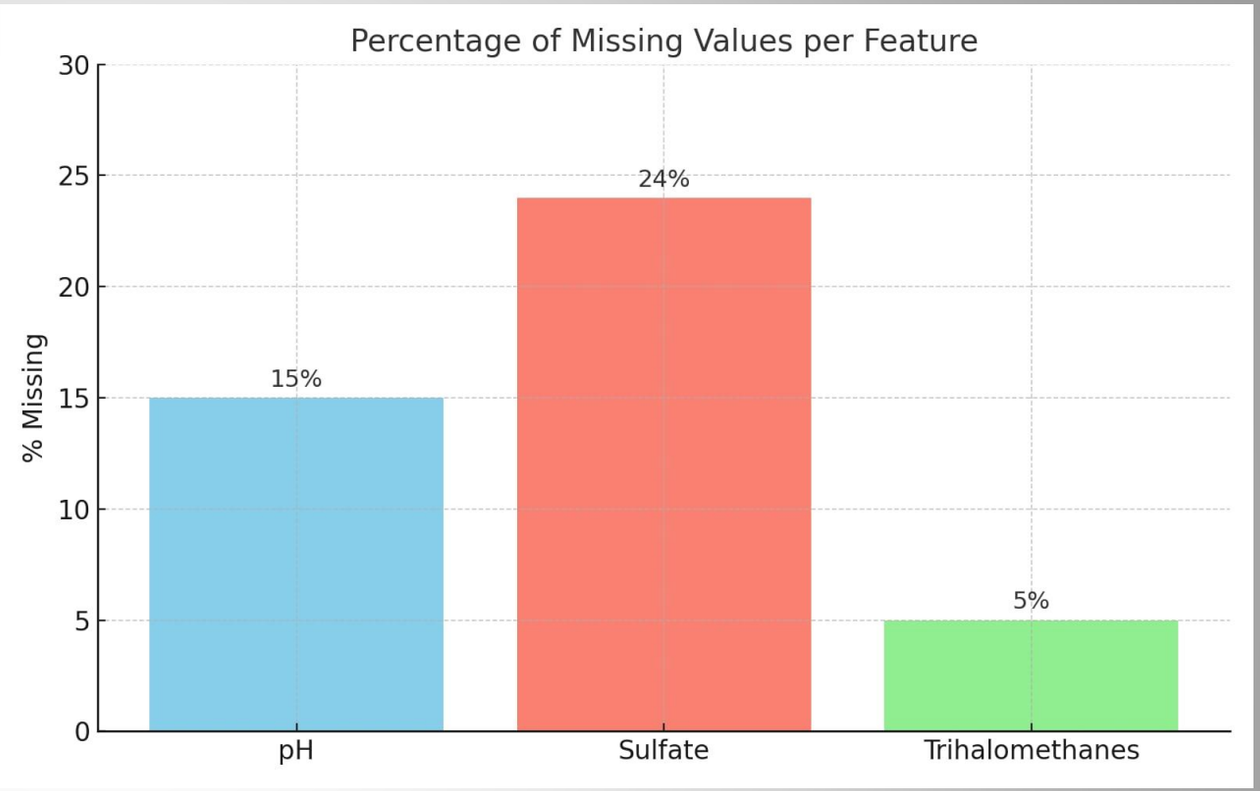
BIA | BOSTON INSTITUTE OF ANALYTICS ®

Title: Missing Data Analysis

Content (Points):

Missing Values in Dataset:

pH: 491 missing (15%)
Sulfate: 781 missing (24%)
Trihalomethanes: 162 missing (5%)



Why Handle Missing Data?

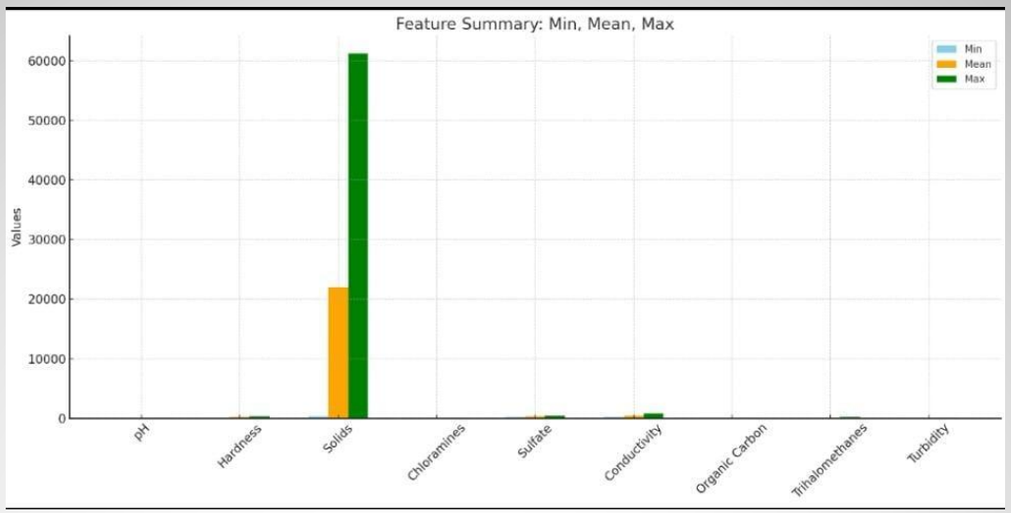
Missing data can bias analysis and affect model accuracy.
Proper handling improves reliability of results.

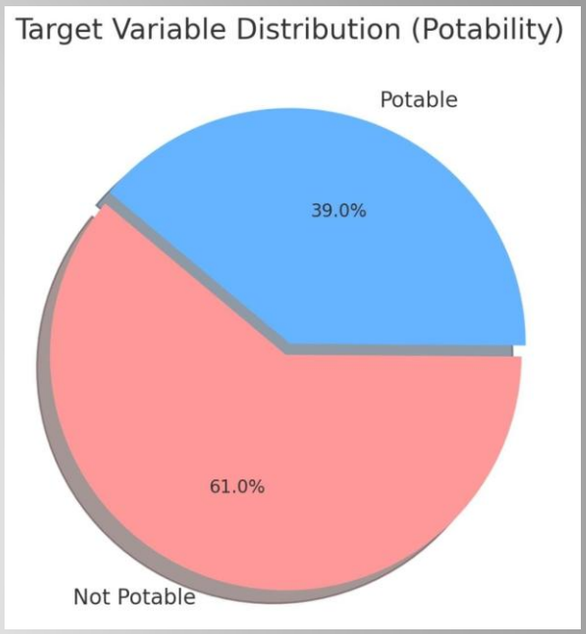
How Handled Here:

Missing values were imputed using column mean.

BIA BOSTON INSTITUTE OF ANALYTICS ®

Feature	Mean	Std Dev	Min	Max
pH	7.08	1.47	0.0	14.0
Hardness	196.37	32.88	47.43	323.12
Solids	22014	8768.57	320.94	61227
Chloramines	7.12	1.58	0.35	13.13
Sulfate	333.77	36.14	129.0	481.03
Conductivity	426.20	80.82	181.48	753.34
Organic Carbon	14.28	3.31	2.2	28.3
Trihalomethanes	66.39	15.77	0.74	124.0
Turbidity	3.97	0.78	1.45	6.74





Observed Trends:

Solids column shows high variance and possible outliers.

Data Transformation:

Log transformation applied on Solids and Trihalomethanes for model input.

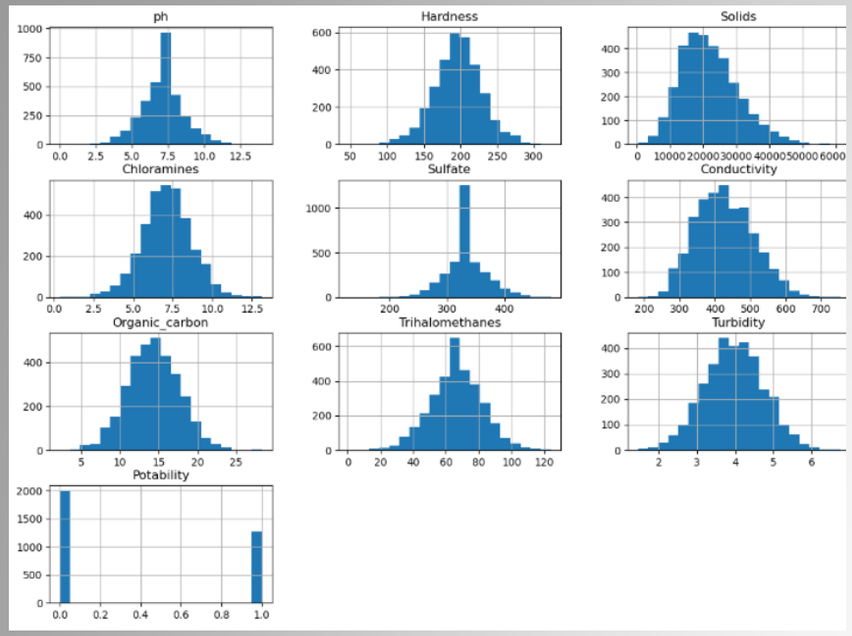
- Potability Classes (Before Balancing):

1. Not Potable: 1998 samples (61%)
2. Potable: 1278 samples (39%)

- Class Imbalance:

1. Dataset is imbalanced – Not Potable samples are more than Potable.
2. Imbalance can cause bias in model predictions, favoring the majority class.

BIA BOSTON INSTITUTE OF ANALYTICS ®



1. Feature Distribution and Outliers:

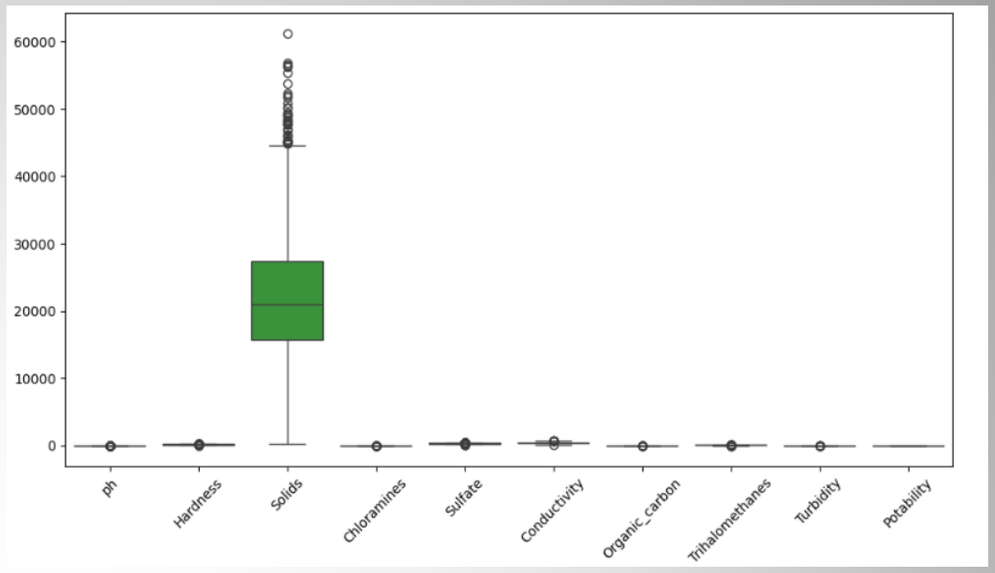
Histograms of Main Features:

Histograms are used to visualize the distribution of each key feature.

Most Parameters (pH, Hardness, etc.): Show a Normal/Bell-Shaped distribution (values centered around the mean).

Solids & Sulfate: Exhibit a Right Skew, indicating the presence of some high outlier values.

The number of samples for 'Not Potable' (0) is significantly higher than for 'Potable' (1).



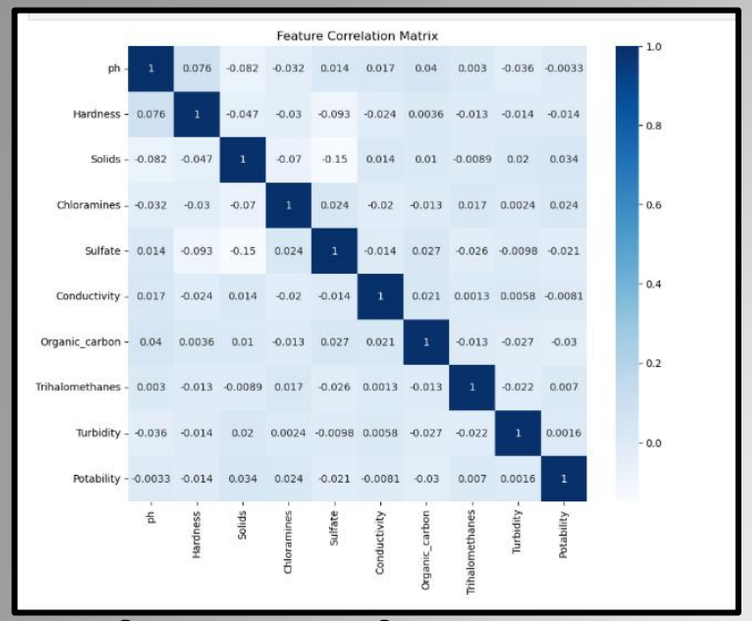
2. Box Plots for Outliers:

- Box plots help identify outliers in the data.
- **Solids Parameter**: Dominated by a large number of extreme high outliers (values near 60,000).
- **All Other Features:** Show tight distribution with minimal or no visible outliers on this scale.

3. Scaling and Transformation:

- Features vary in scale and some have non-normal distributions.
- Scaling (e.g., StandardScaler) and transformations (e.g., Log Transformation) are needed to improve model performance.

BIA BOSTON INSTITUTE OF ANALYTICS ®



Title: Correlation Matrix

1.Purpose:

To understand relationships between features. Highly correlated features can create redundancy in modeling.

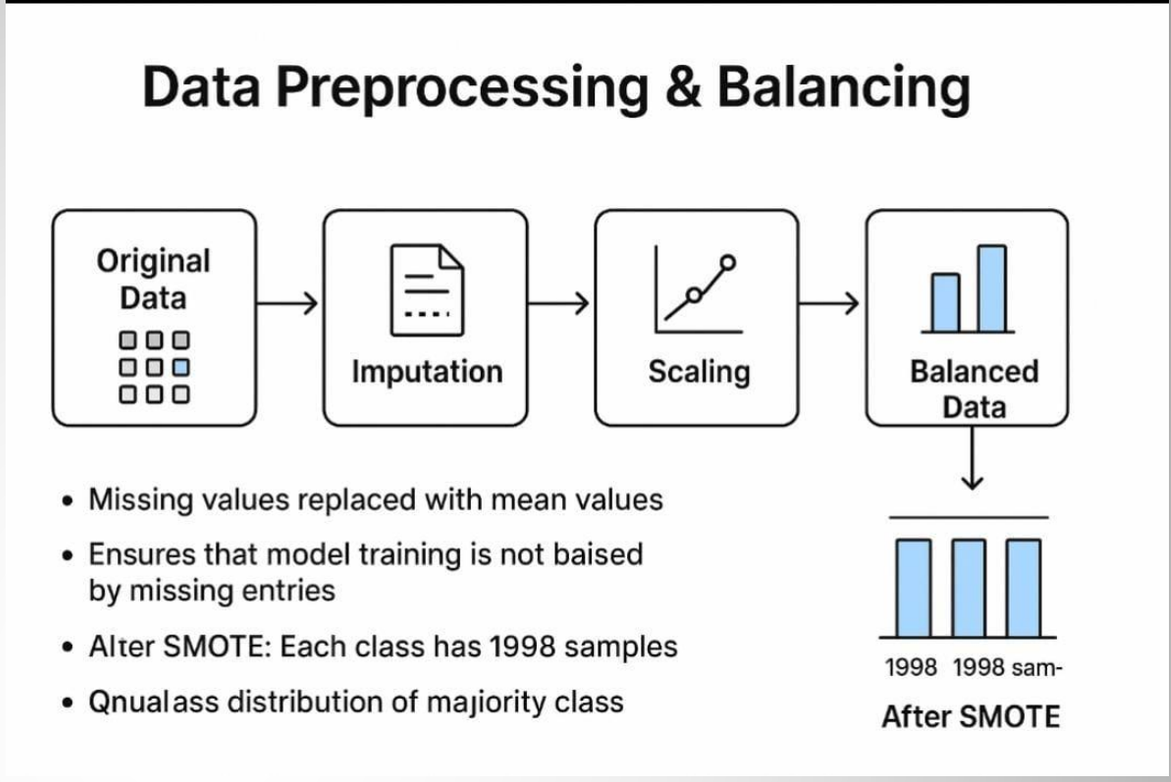
2. Observations:

Some features show moderate to high correlations.

For example, Conductivity and Solids are positively correlated.

3. Implications:

Highly correlated features may be considered for feature selection or dimensionality reduction.



Model Performance

- Goal: Determine the best model using Cross-validation F1-scores.
- Result: Random Forest was the top performer with a Mean F1-score of 0.715.
- Comparison: All other models (XGBoost, KNN, SVC, Decision Tree) scored lower.
- Action: Random Forest was selected for further hyperparameter tuning to maximize its performance.

BIA BOSTON INSTITUTE OF ANALYTICS ®

Hyperparameter Tuning (Random Forest) :

Goal: Optimize model performance using GridSearchCV.

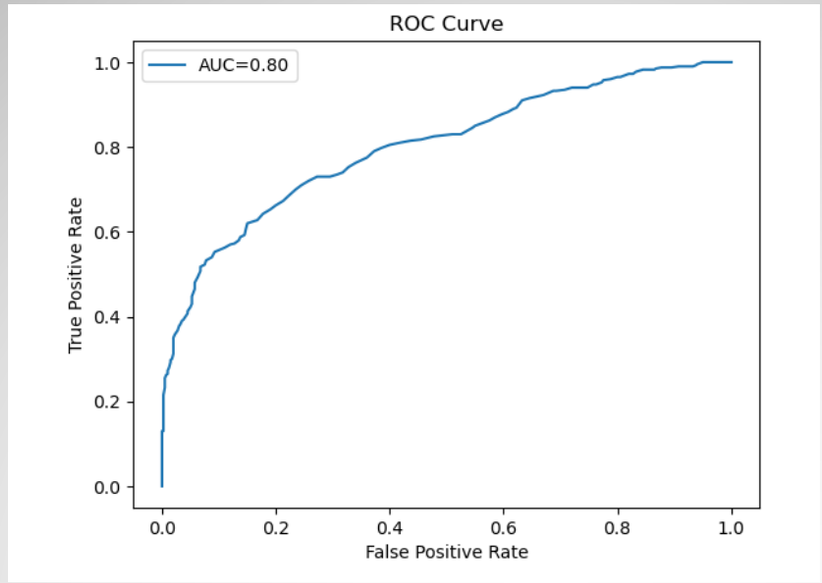
Parameters Tuned:

I. n_estimators: 100, 200
II. max_depth: 5, 10, None
III. min_samples_split: 2, 5, 10

Best Parameters:

I. n_estimators = 200
II. max_depth = None
III. min_samples_split = 2

Result: Model achieved better accuracy and stability with deeper trees and more estimators.

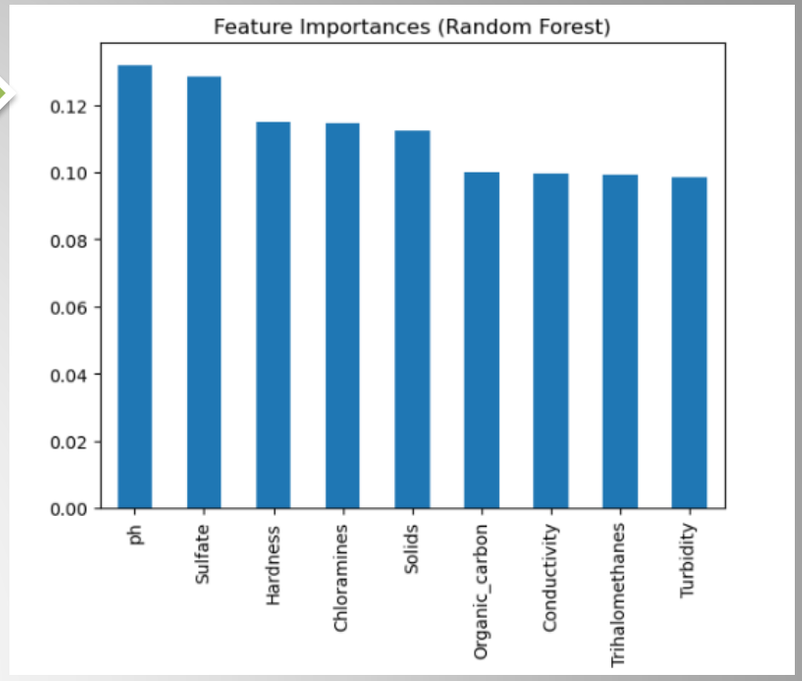


Final Model Evaluation

- Accuracy: ~73%
- Precision / Recall / F1: 0.73 each
- AUC (ROC): 0.80
- Insights:

I. Balanced precision and recall indicate consistent performance.
II. High AUC shows strong class discrimination.

- Visuals: Confusion Matrix & ROC Curve.



Feature Importance (Random Forest)

- Top Features (Importance Scores):

I. pH (0.13), Sulfate (0.13), Hardness (0.12)
II. Chloramines (0.11), Solids (0.11)
III. Organic Carbon (0.10), Conductivity (0.10)
IV. Trihalomethanes (0.10), Turbidity (0.10)

- Key Insights:

I. pH and Sulfate most strongly affect water potability.
II. Chemical and mineral content (like Hardness, Solids) play major roles in predicting quality.

BIA BOSTON INSTITUTE OF ANALYTICS ®

Summary:

A total of 3,276 water samples were analyzed in the dataset.
Potable (Safe for drinking): 39.0% samples (1,278)
Not Potable (Unsafe for drinking): 61.0% samples (1,998)
The dataset was imbalanced, with a higher proportion (61%) of non-potable water samples compared to potable ones. This imbalance was addressed before building the model.

Ultimate Conclusion —Through this Water Potability Dataset Analysis project, we successfully built a Machine Learning model (Random Forest) that handled data cleaning and class imbalance challenges, achieving an accuracy of approximately 73% and an AUC score of 0.80.1.
Based on the model's analysis, pH and Sulfate were identified as the two most important factors determining the potability of water.
Random Forest showed best performance.

Implications:

Reliable model can assist in real-time water quality prediction.
Highlights key chemical parameters affecting potability.

Future Work:

Apply advanced feature engineering & ensemble methods.
Integrate model into IoT-based water monitoring systems.
Use larger, more diverse datasets for higher generalization.

Deployed Streamlit Cloud URL link :
Water Potability Prediction Model Analysis · Streamlit

BIA | BOSTON INSTITUTE OF ANALYTICS ®

Thank You

BIA | BOSTON INSTITUTE OF ANALYTICS ®