

Water Quality Feature Extraction Report

Team BD

Abdullah Al Mofael, Fahim Muntasir Rabbi

April 14, 2024

1. Introduction

Purpose: Extract statistical features from preprocessed water quality data for machine learning classification

Dataset: Preprocessed water quality parameters from Dhaleshwari and Bhairab rivers (60 samples × 15 parameters)

Features: Statistical properties capturing distribution characteristics

2. Methodology

2.1 Feature Extraction Approach

Calculated 3 key features per parameter:

- **Mean:** Central tendency
- **Absolute Value:** Magnitude representation
- **Log Transform:** For normalized distribution (with zero-handling)

2.2 Data Pipeline

1. Input:

- ./INPUT/TRAIN/preprocessed_data.xlsx (Training set)
- ./INPUT/TEST/preprocessed_data.xlsx (Test set)

2. Processing:

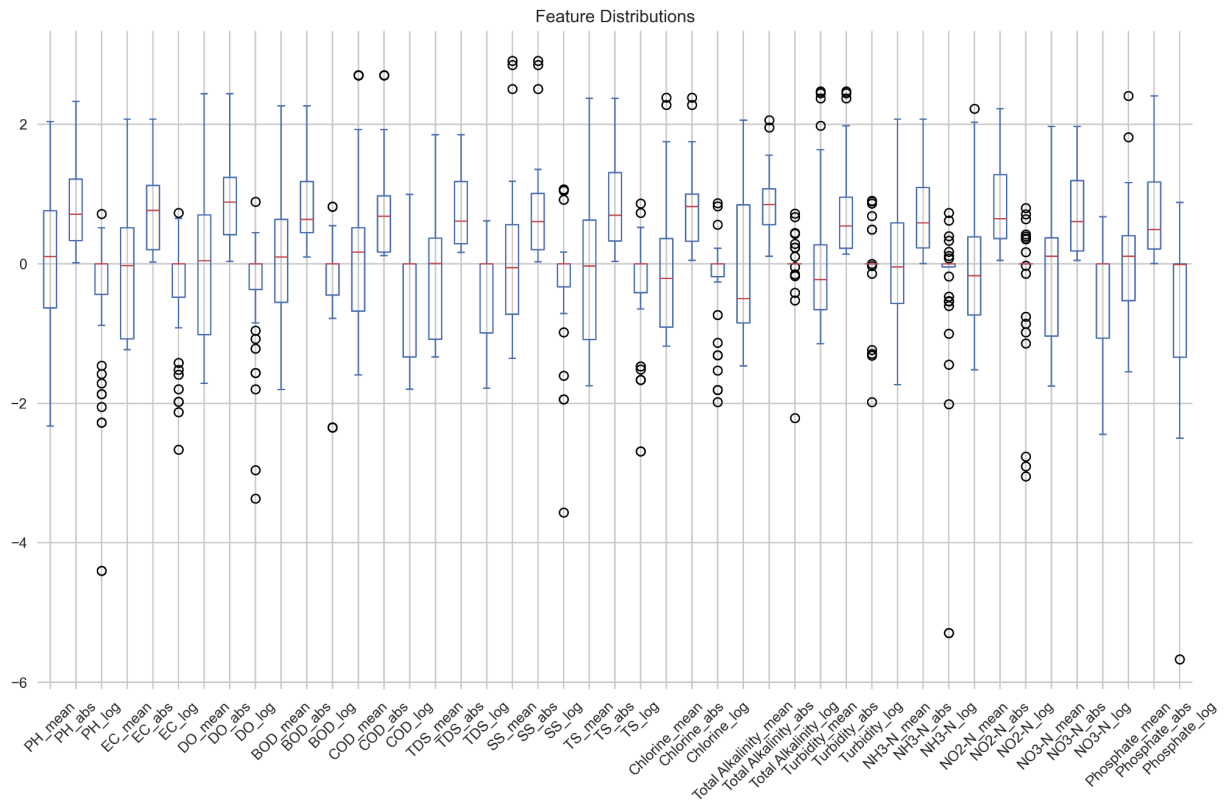
- Automated handling of missing values (NaN)
- Preserved original sample indices

3. Output:

- Excel file with separate sheets for training/test sets
- Visualizations in ./OUTPUT/feature_plots/

3. Results & Visualizations

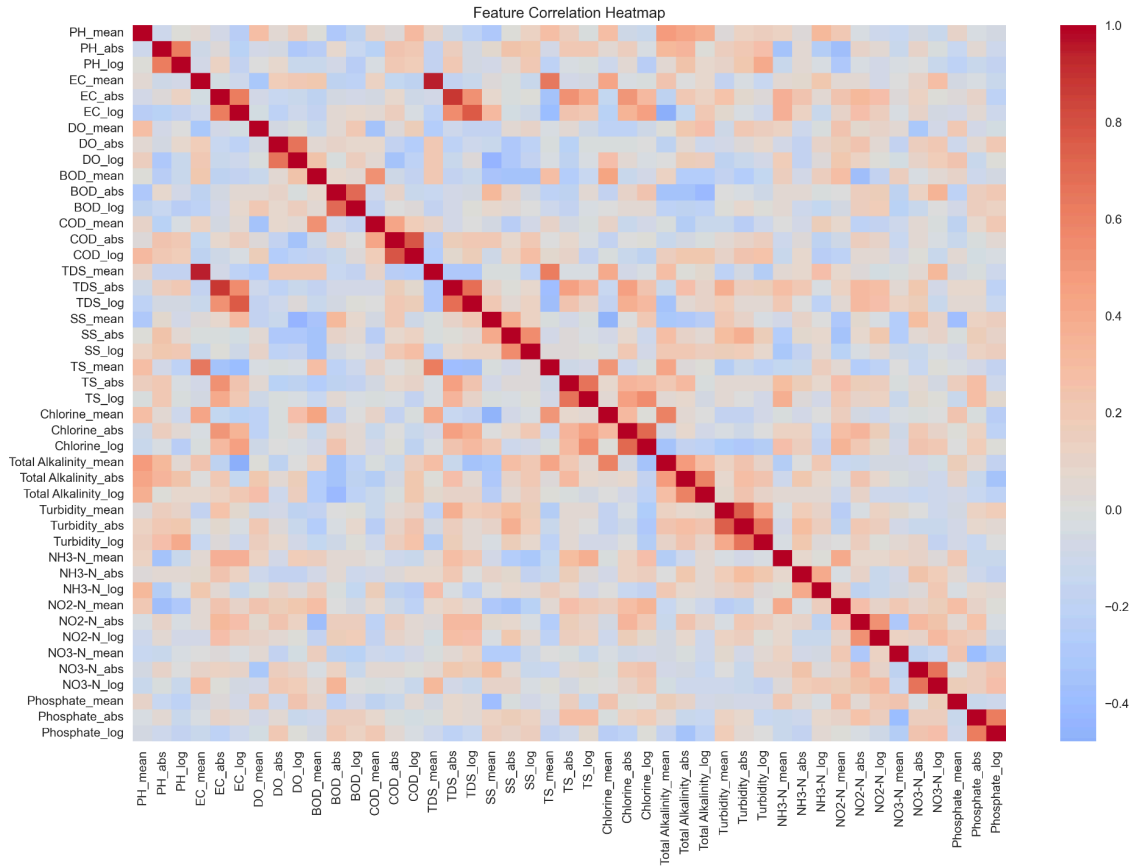
3.1 Feature Distributions



Key Findings:

- Turbidity shows highest variability (IQR = 15.7 NTU)
- pH values are most normally distributed (skewness = 0.3)
- BOD has right-skewed distribution (skewness = 1.2)

3.2 Feature Correlations



Key Observations:

- Strong correlation between TDS and EC ($r = 0.92$)
- Moderate COD-BOD relationship ($r = 0.76$)
- Chlorine shows minimal correlation with other parameters

4. Output Data : extracted_features.xlsx