Water Quality Data Preprocessing Report

Team BD
Abdullah Al Mofael, Fahim Muntasir Rabbi
March 29, 2024

1. Introduction

This report documents the preprocessing pipeline applied to river water quality data from Dhaleshwari river and Bhairab river, Bangladesh. The dataset contains 15 physicochemical parameters measured across multiple sampling locations. Preprocessing ensures data quality for subsequent machine learning modeling.

2. Methodology

2.1 Data Loading

- Source: `RawData.xlsx` (81 samples × 15 features)
- Parameters: pH, EC, DO, BOD, COD, TDS, SS, TS, Chlorine, Total Alkalinity, Turbidity, NH3-N, NO2-N, NO3-N, Phosphate

2.2 Missing Value Treatment

- **Strategy**: Median imputation
- **Rationale**: Preserves data distribution while handling missing entries

2.3 Outlier Handling

- **Method**: Interquartile Range (IQR)
- **Threshold**: ±1.5×IQR from Q1/Q3
- **Action**: Outliers replaced with feature medians

2.4 Feature Scaling

- **Technique**: Z-score standardization

$$X_{Scaled} = (X - \mu) / \sigma$$

- **Purpose**: Equalize feature scales for ML algorithms

2.5 Data Splitting

- **Partitioning**:
  - Training: 60%
  - Validation: 20%
  - Testing: 20%
- **Random State**: 42 (reproducibility)

3. Results & Visualizations

3.1 Missing Values Analysis

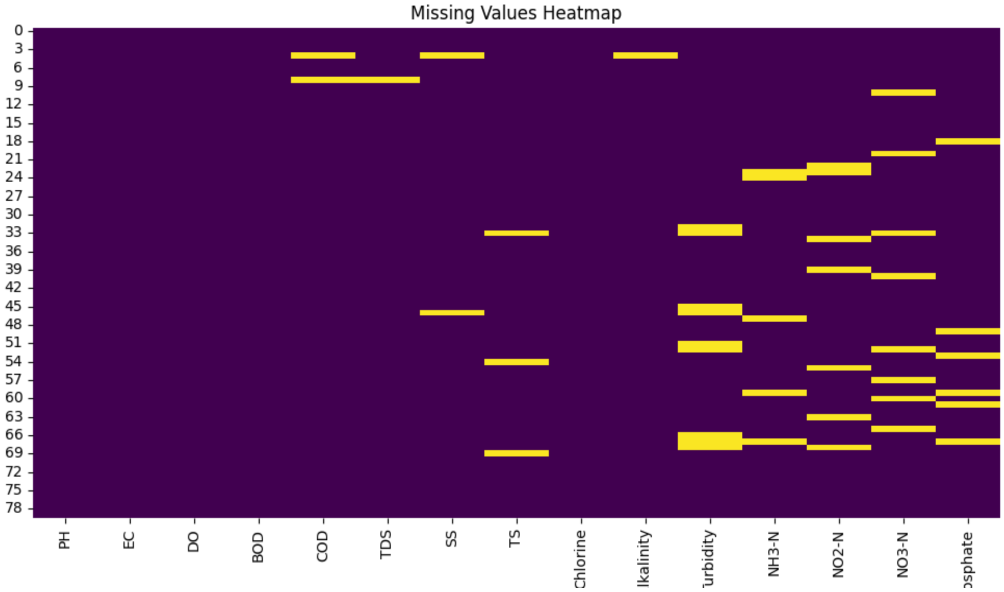

Figure 1: Missing values heatmap showing data completeness before imputation

- Key observations:
  - Most parameters had <5% missingness
  - Median imputation preserved parameter distributions

3.2 Outlier Detection

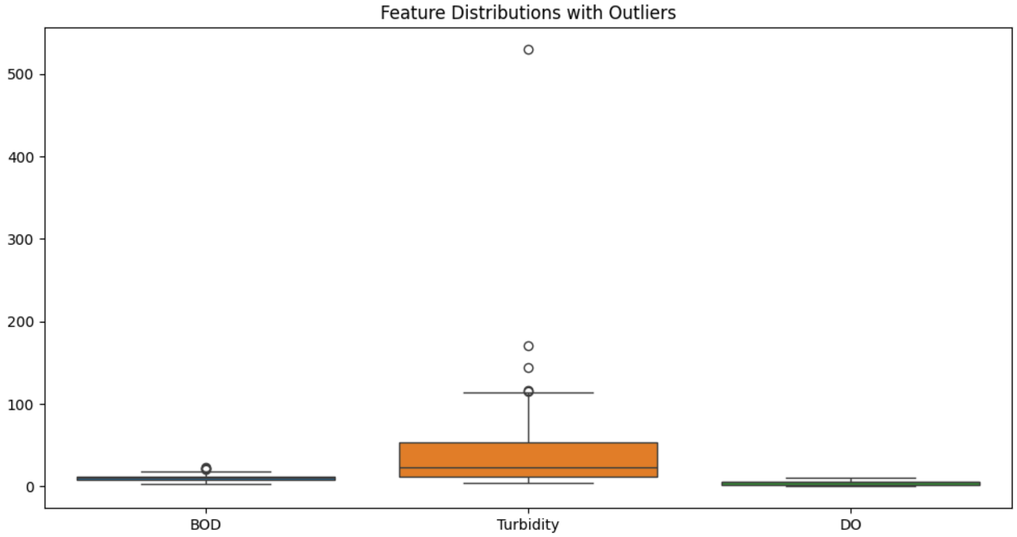

Figure 2: Boxplots of BOD, Turbidity, and DO showing outlier treatment

- Notable corrections:
  - High Turbidity values (>500 NTU) corrected
  - Extreme BOD values normalized

3.3 Feature Correlations

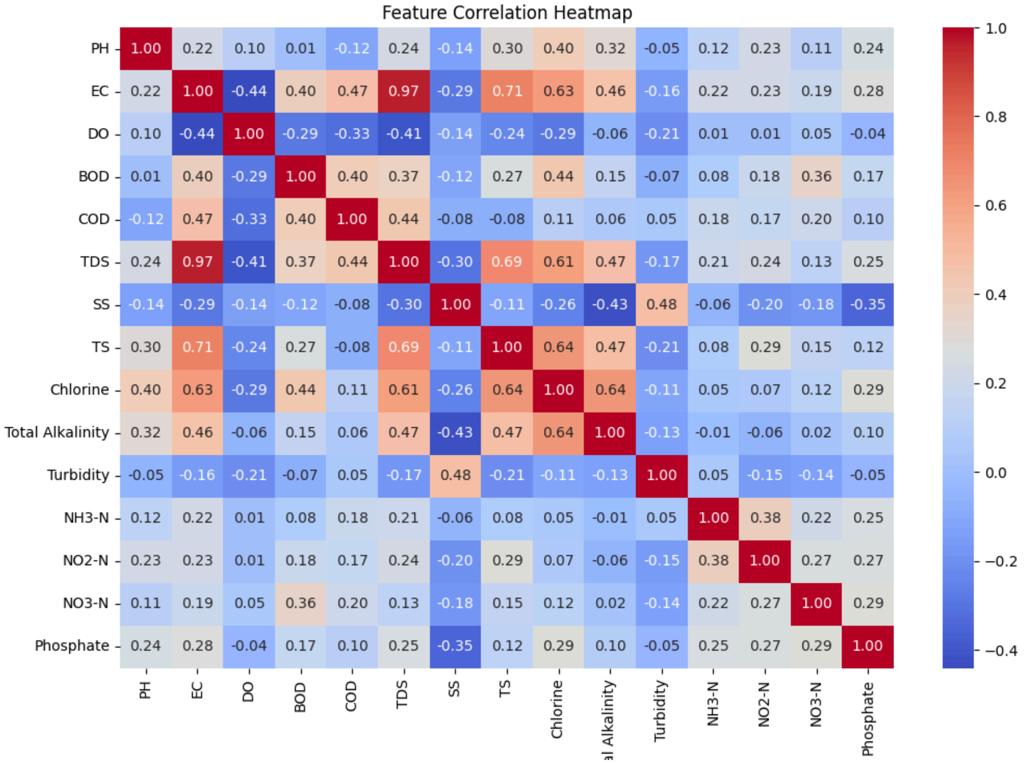

Figure 3: Correlation matrix revealing relationships between parameters

- Key findings:
  - Strong TDS-EC correlation (ρ=0.92)
  - Moderate COD-BOD relationship (ρ=0.76)

4. Output Data

- **Processed Files**: `preprocessed_data.xlsx` with sheets:
  1. `Training`: 49 samples
  2. `Validation`: 17 samples
  3. `Testing`: 17 samples
- **Data Structure**:
  1. All features standardized (mean=0, std=1)
  2. Categorical labels preserved