Prediction of Water Quality Index (WQI) Using Machine Learning Algorithms

Ayse B Sengul, PhD

Springboard Data Science Capstone Project

April 26, 2024

Problem Statement

- Global freshwater demand increases with population and economic growth.
- This strains water quality, necessitating effective monitoring methods.
- WQI simplifies water quality assessment but has limitations.
- ML offers a promising solution for accurate WQI prediction.

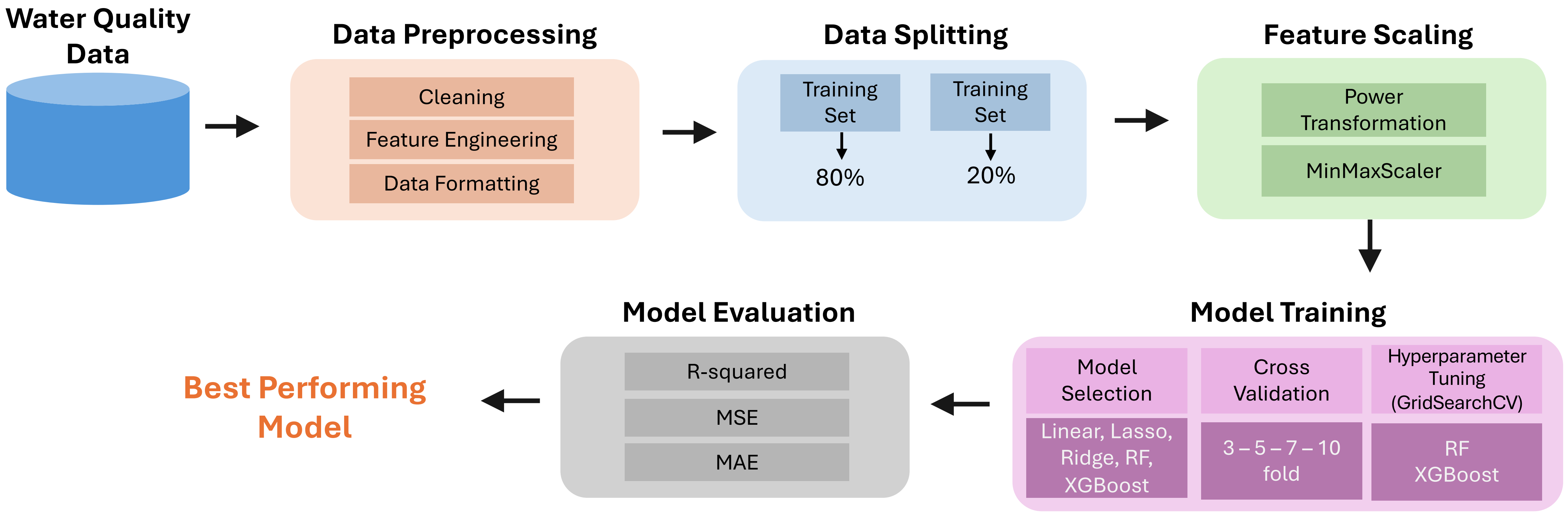
This study aims to develop ML-based techniques for efficient WQI prediction.

Data Overview

- **Source:**
  - City of Cape Coral Water Quality data from various monitoring stations.
  - Monthly sampling.
  - 1986 – 2022.
- **Size:**
  - 18148 rows and 48 columns.
- **Key Features:**
  - Temperature, dissolved oxygen, conductivity, dissolved oxygen, pH, nitrite, nitrate, total nitrogen, total phosphorus, total suspended solids, biological oxygen demand, chlorophyll and turbidity.

*https://capecoral-capegis.opendata.arcgis.com/datasets/b0579ba7aa1145e090c3a74e295564df_1/explore

Flowchart of the WQI Prediction



RF: Random Forest, MSE: mean squared error, MAE: mean absolute error

Data Wrangling

- Select and rename columns for clarity.
- Convert data types (e.g., float for 'TP', datetime for 'Date').
- Focus on river water samples.
- Handle missing data by removing high-missing columns (‘Chl’).
- Correct zero ('Temp (°C)', 'pH') and erroneous values ('TN', 'TP', 'Turbidity').
- Impute NaN values with the median.
- Compute WQI and classify into water quality classes.

Data Wrangling

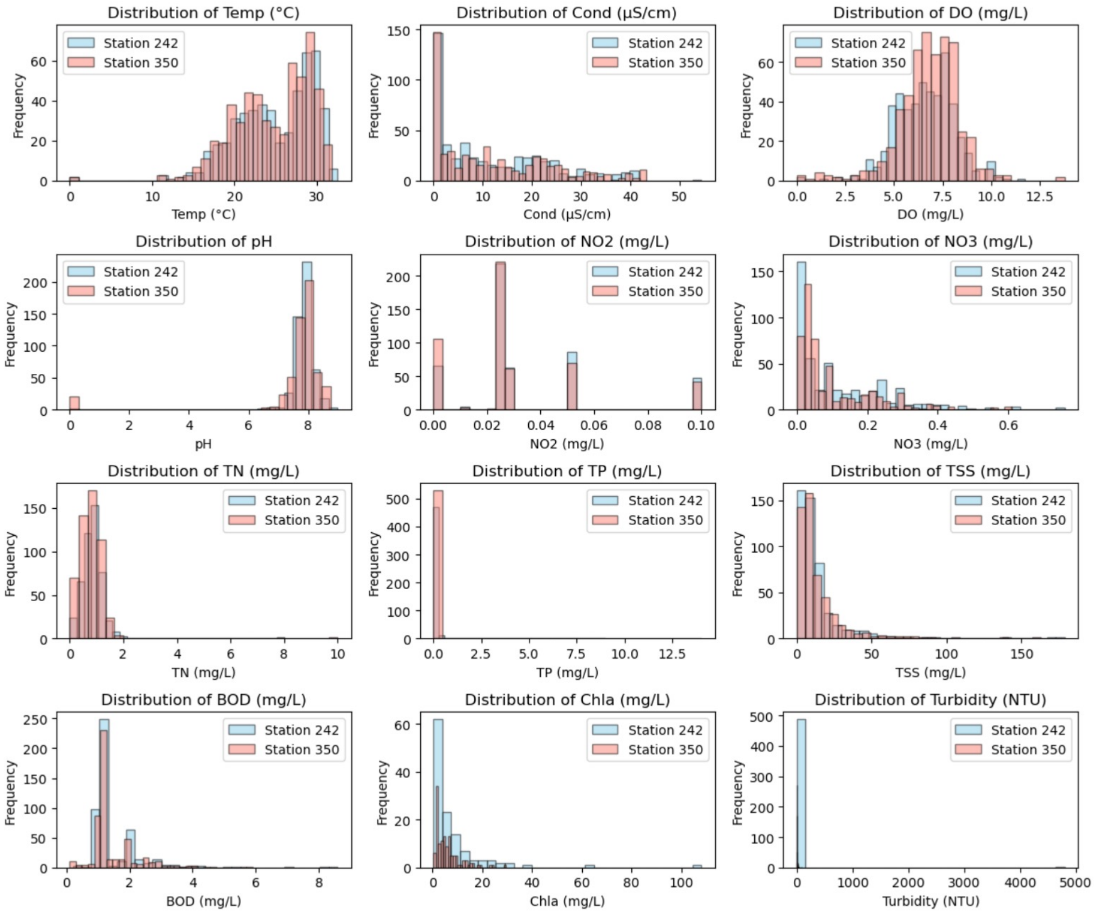


Figure 1. Distribution of data before processing.

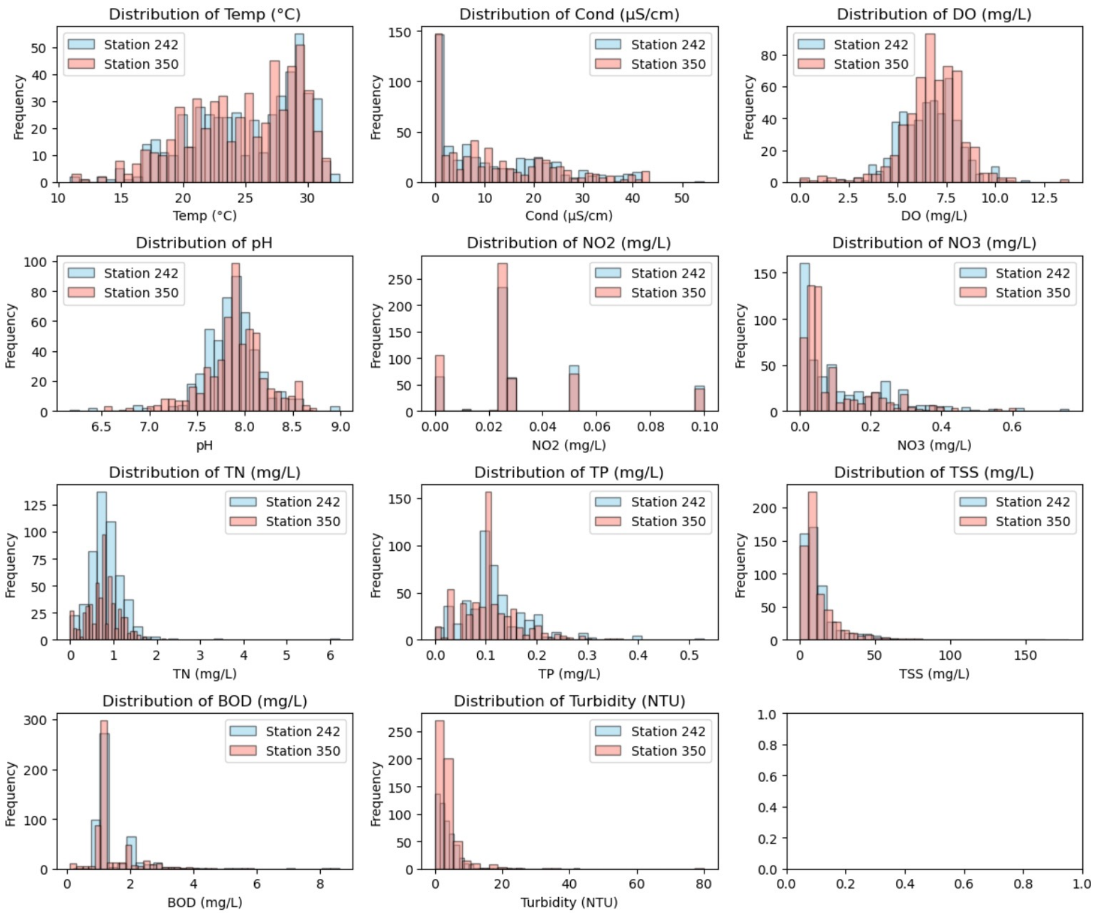


Figure 2. Distribution of data after processing.

Exploratory Data Analysis (EDA)

- Removed duplicates, resulting in 1025 rows and 14 columns.
- Identified and analyzed outliers for potential data errors.
- Examined skewed (NO3, TN, TP, TSS, BOD, Turbidity) and multimodal distributions in features (NO2).

Exploratory Data Analysis (EDA)

WQI

Most data btw 35 – 56.

WQI values spanned from 12.5 – 343.

Mean value of WQI – 51.

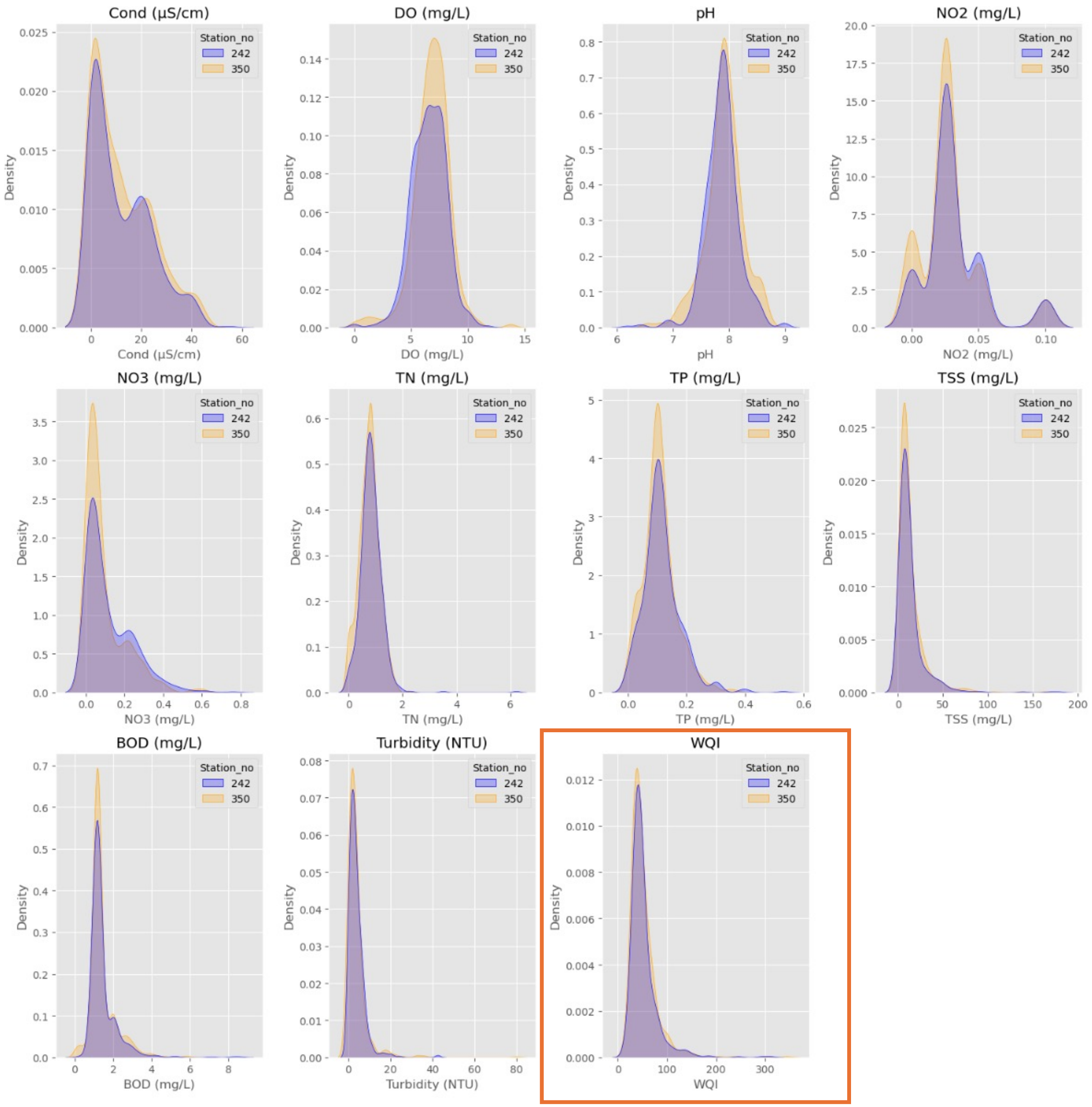


Figure 3. KDE plots for numerical features for each station.

Exploratory Data Analysis (EDA)

WQI has a very strong positive correlation with **TSS** and a moderate positive correlation with **Turbidity.**

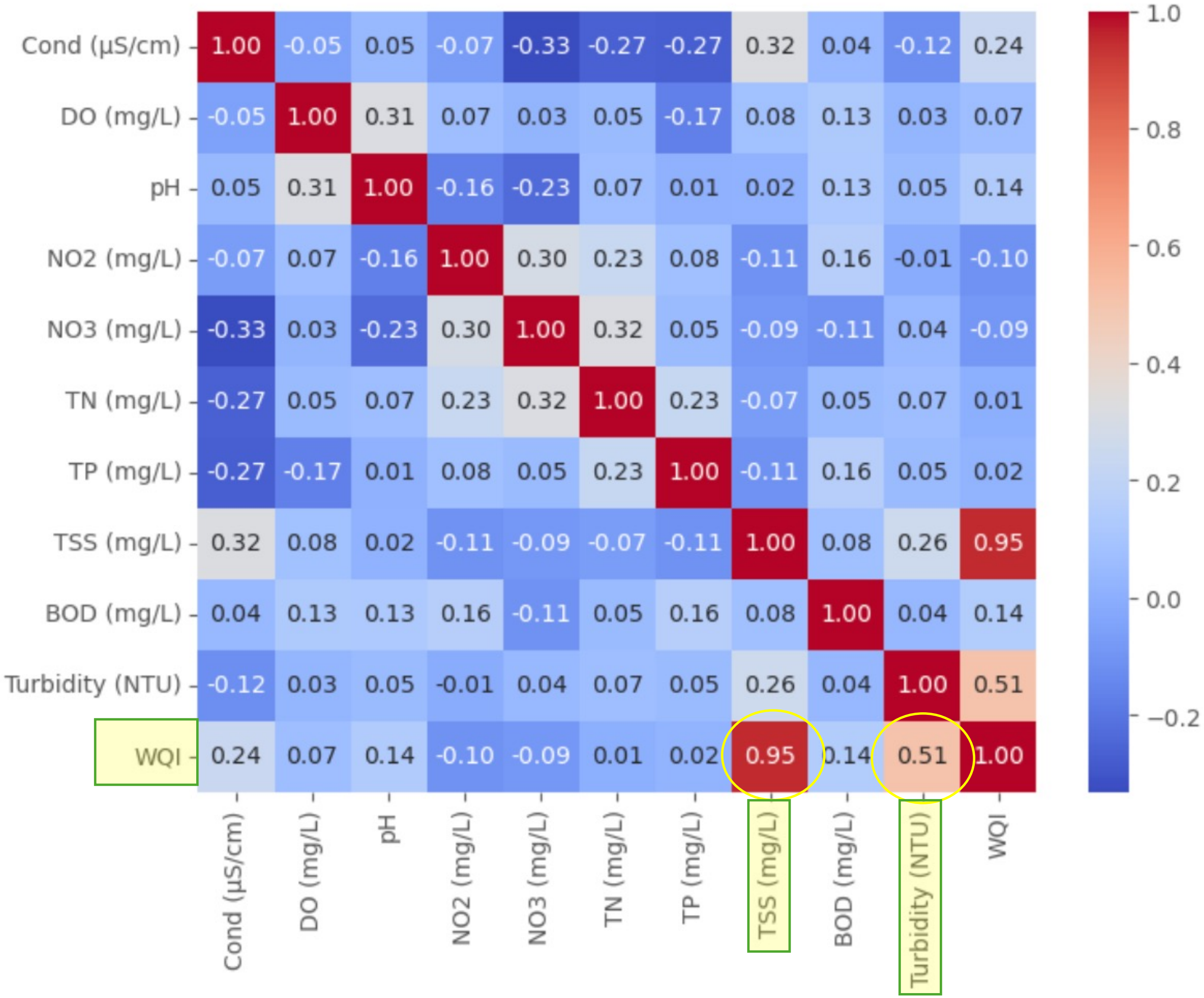


Figure 4. Correlation Heatmap of Numerical Features.

Exploratory Data Analysis (EDA)

Knowing the value of TSS and Turbidity can provide valuable information for predicting WQI.

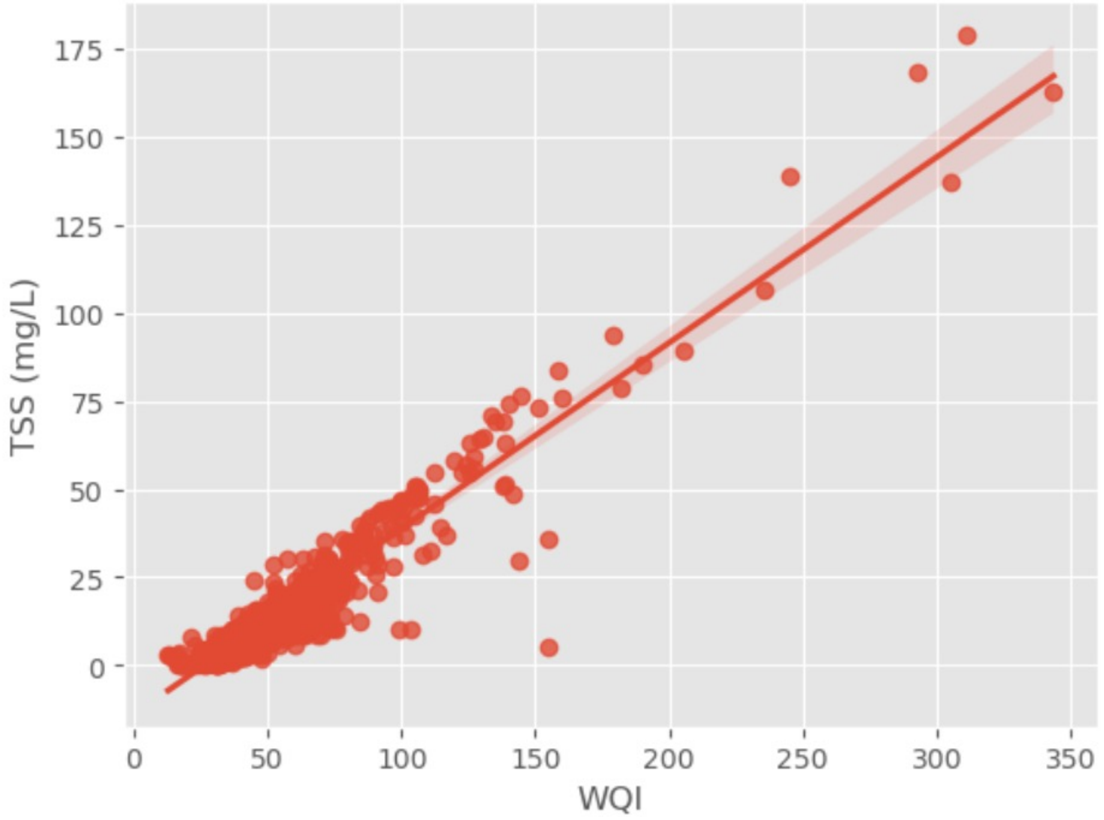


Figure 5. Regression Plot of WQI vs TSS.

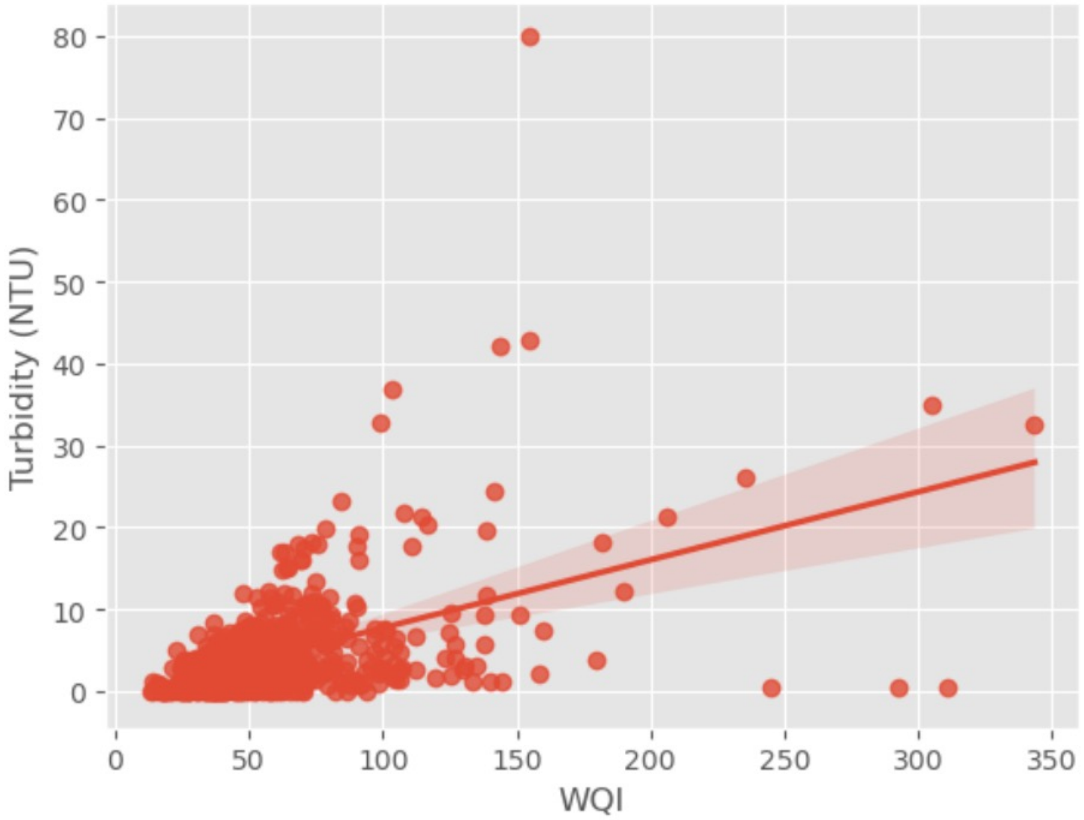
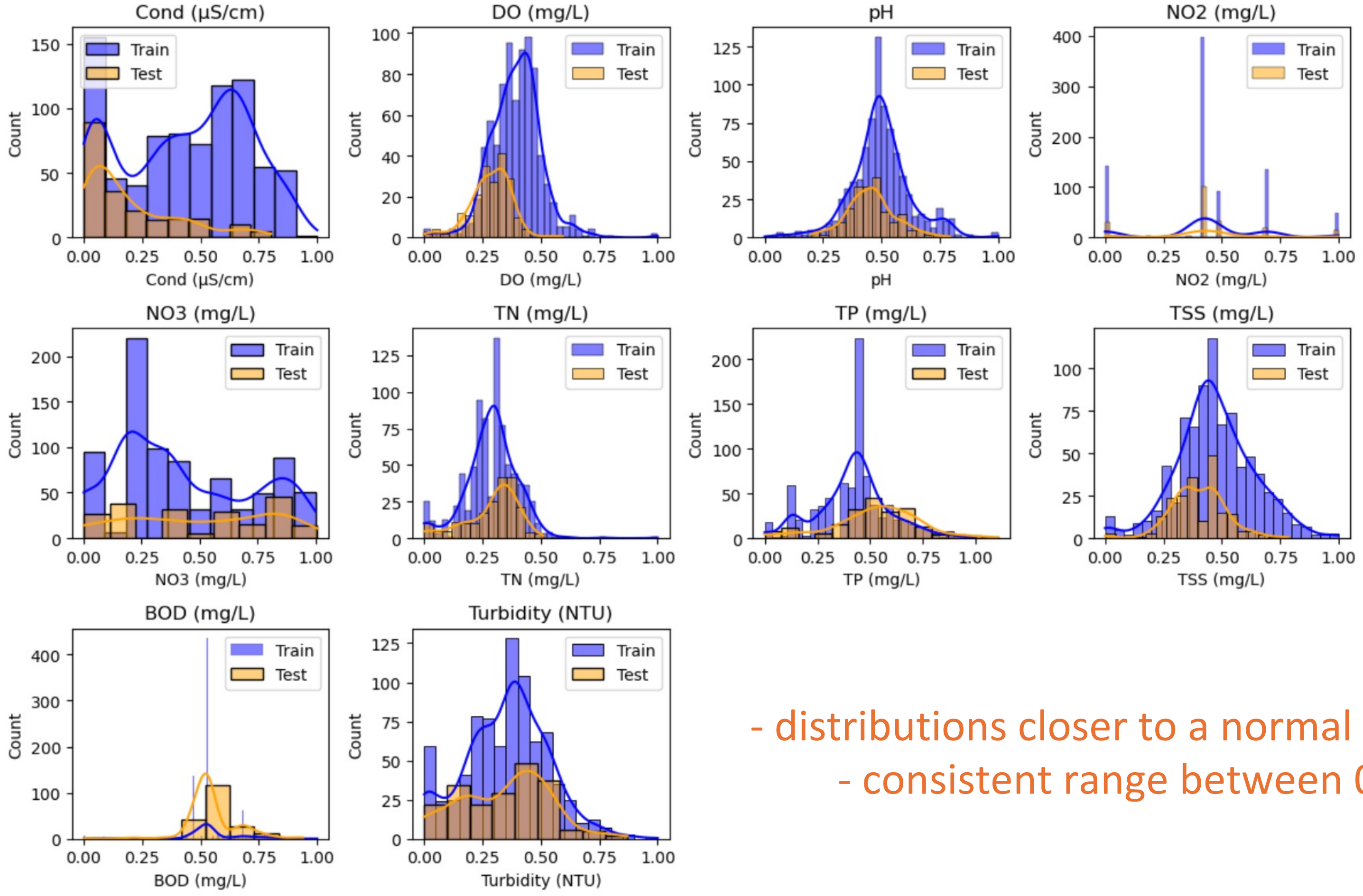


Figure 6. Regression Plot of WQI vs Turbidity.

Preprocessing and Training Data Development

- Encoded 'Water Quality Classification' into binary columns.
- Extracted 'Month' from 'Date' for seasonal analysis.
- Split data into 80-20 ratio for training and testing.
- Applied Power Transformation to address skewness and outliers.
- Used MinMaxScaler to scale numerical features to range [0, 1].

Preprocessing and Training Data Development



- distributions closer to a normal distribution
- consistent range between 0 and 1

Figure 7. Histograms of numerical features after scaling.

Modeling

- **Regression models**: Linear Regression, Lasso Regression, Ridge Regression, Random Forest Regression, XGBoost Regression.
- **Evaluation metrics**: R2 score, MSE, MAE.
- **Hyperparameter tuning**: GridSearchCV
  - Random Forest: 'n_estimators': [100, 200, 300], 'max_depth': [None, 10, 20], 'min_samples_split': [2, 5, 10], 'min_samples_leaf': [1, 2, 4], 'max_features': ['sqrt', 'log2']
  - XGBoost: 'n_estimators': [100, 200, 300], 'max_depth': [3, 5, 7], 'learning_rate': [0.01, 0.05, 0.1]
- **Cross-validation**: 3-fold, 5-fold, 7-fold, 10-fold.

Modeling – Linear Regression

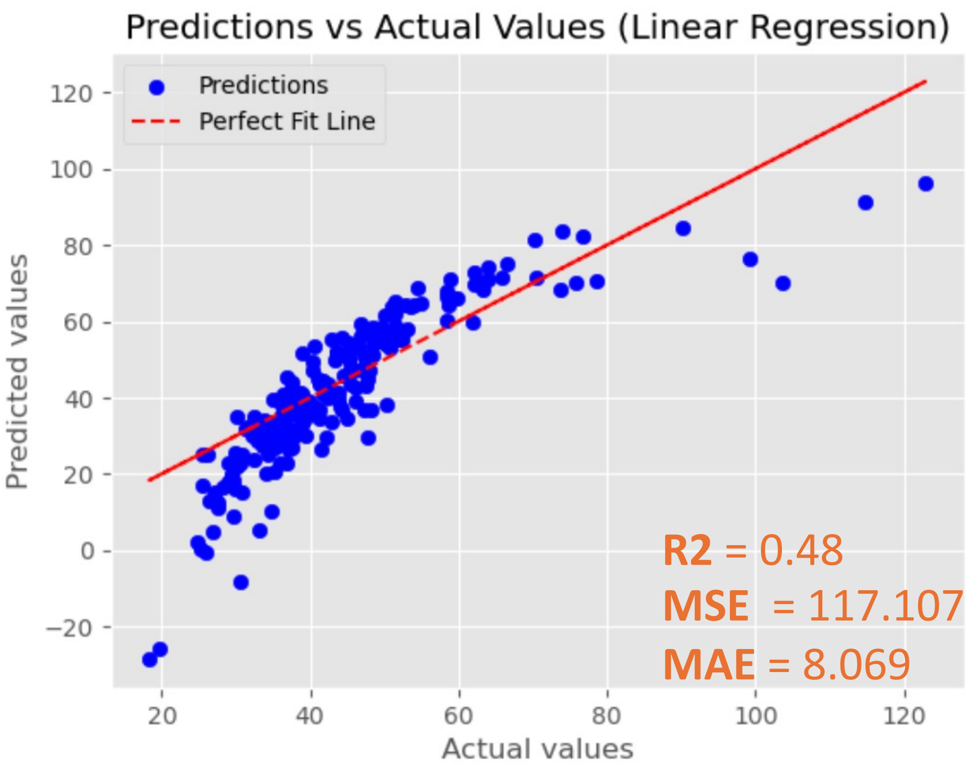


Figure 8. Relationship between the predicted and actual values.

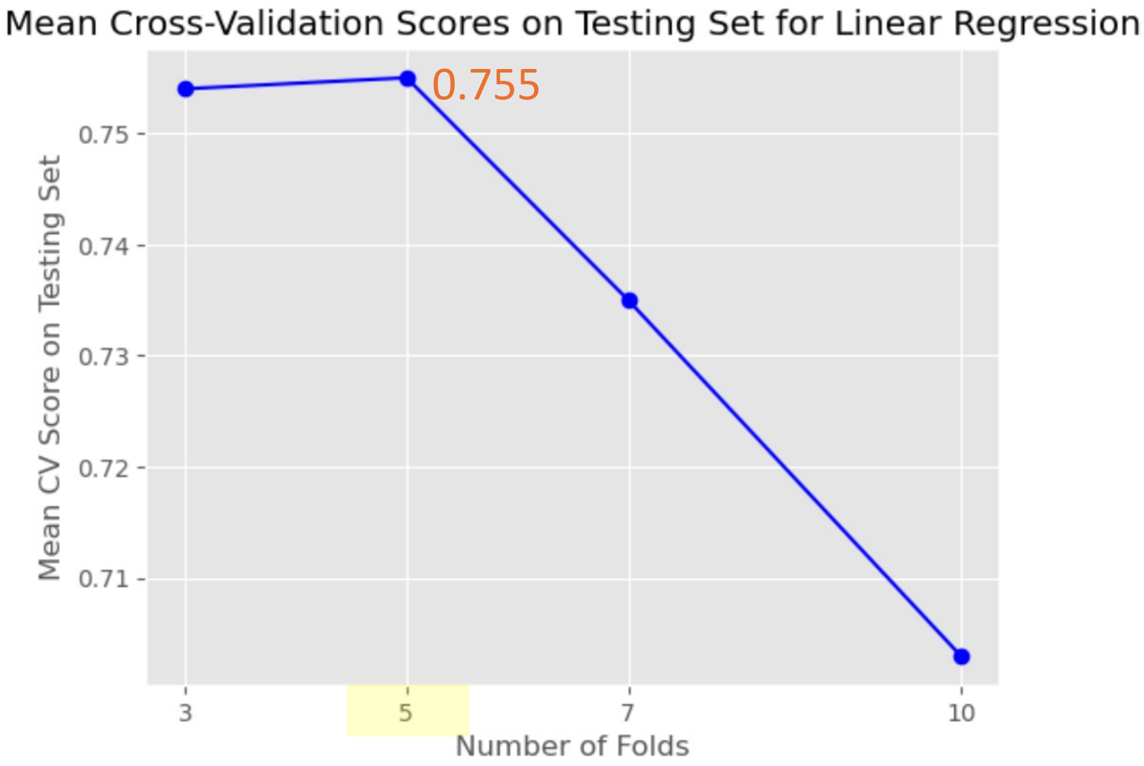
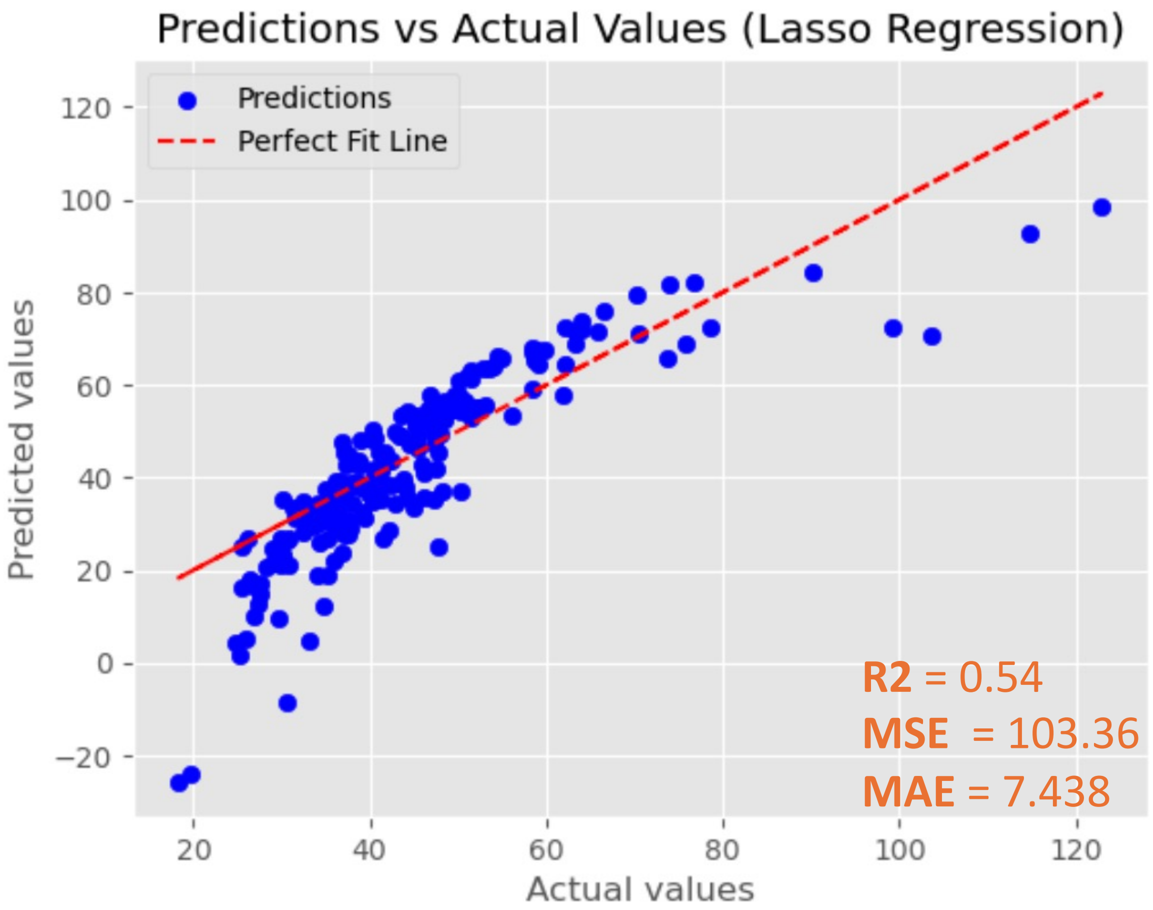


Figure 9. The cross-validation results.

Modeling – Lasso Regression



R2 = 0.54
MSE = 103.36
MAE = 7.438

Figure 10. Relationship between the predicted and actual values

Modeling – Ridge Regression

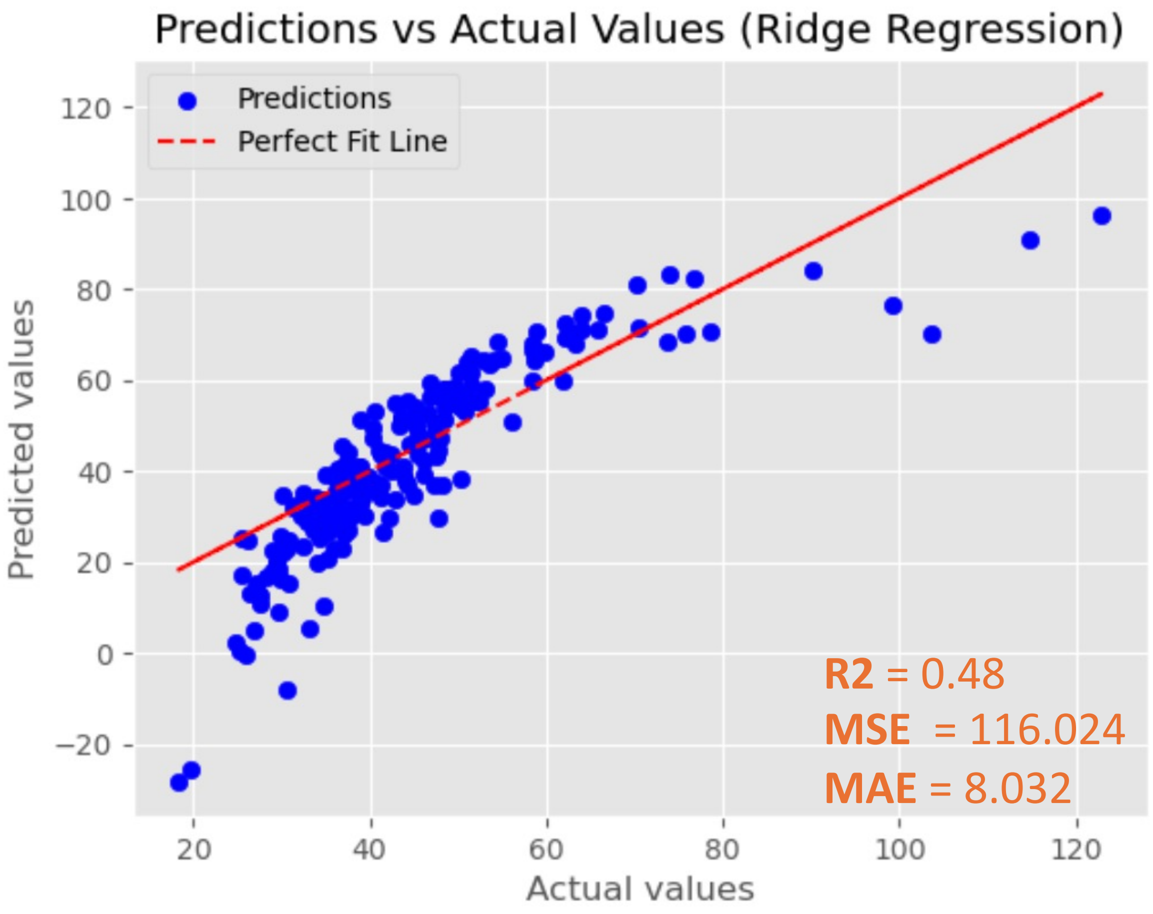


Figure 11. Relationship between the predicted and actual values

Modeling – Random Forest Regression

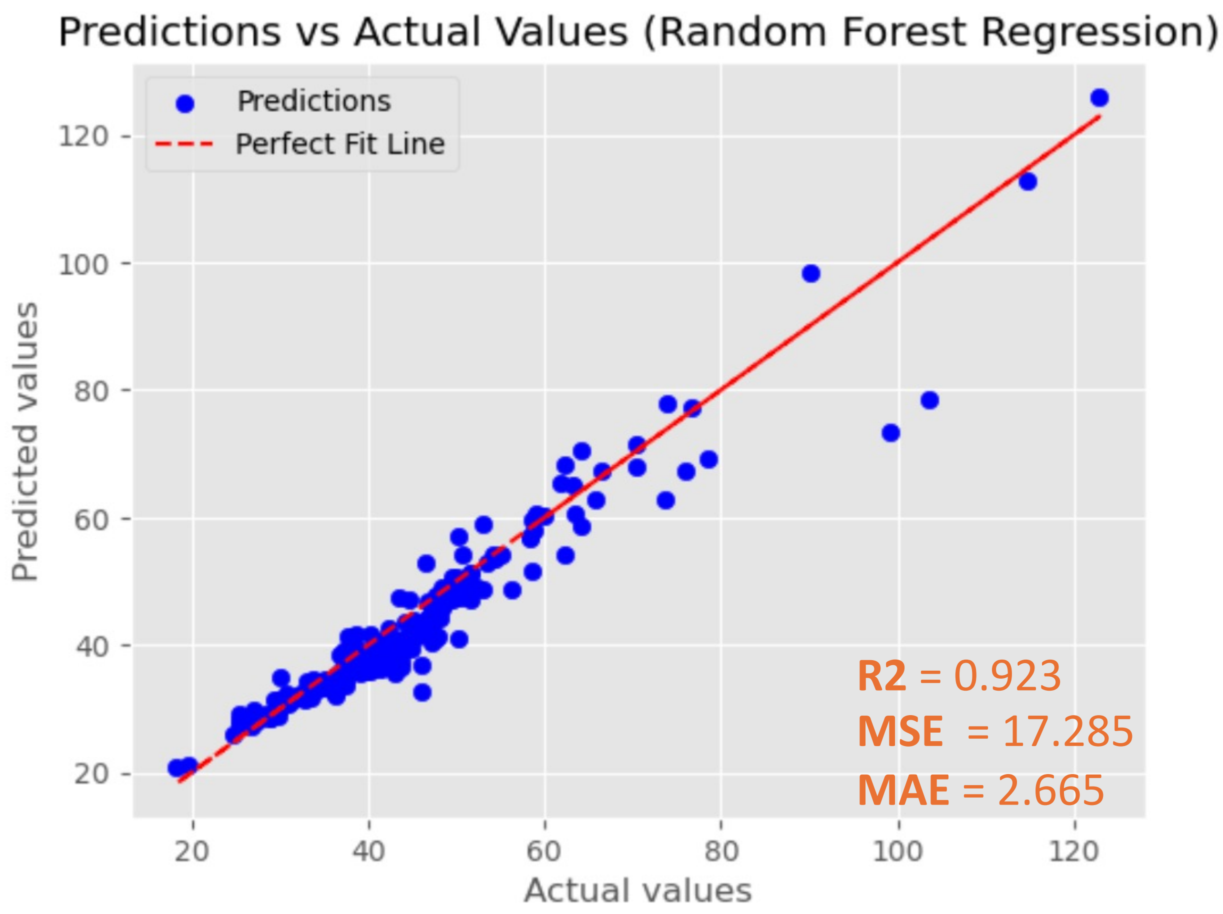


Figure 12. Relationship between the predicted and actual values.

Modeling – Random Forest Regression

Table 1. Random Forest Model Cross-Validation Results.

	Cross-Validation	Max Depth	Max Features	N_Estimators	R-squared Score	Mean CV Score (Testing Set)	Standard Deviation
0	3-fold	20.0	log2	NaN	0.857	0.739	0.080
1	5-fold	20.0	sqrt	300.0	0.880	0.774	0.083
2	7-fold	NaN	log2	NaN	0.906	0.770	0.116
3	10-fold	20.0	log2	200.0	0.909	0.781	0.103

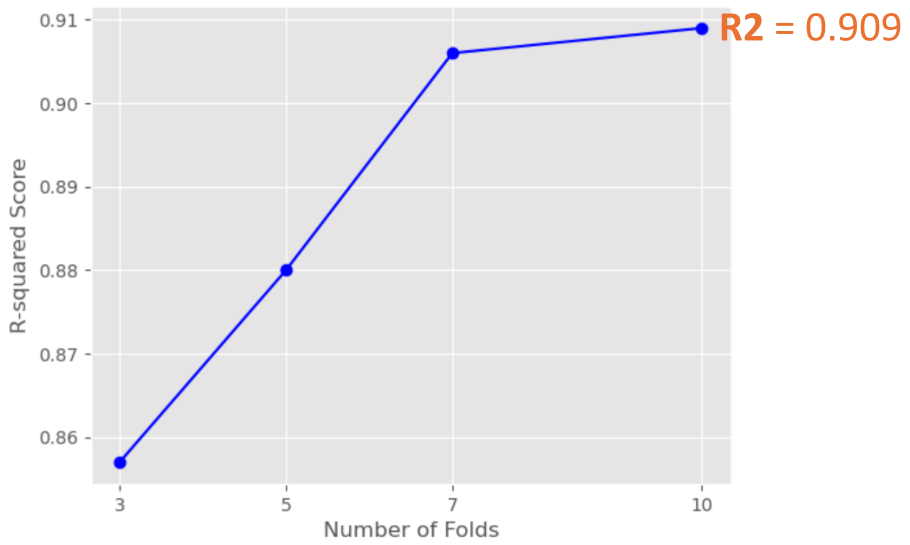


Figure 13. Best scores with different number of folds for the Random Forest model.

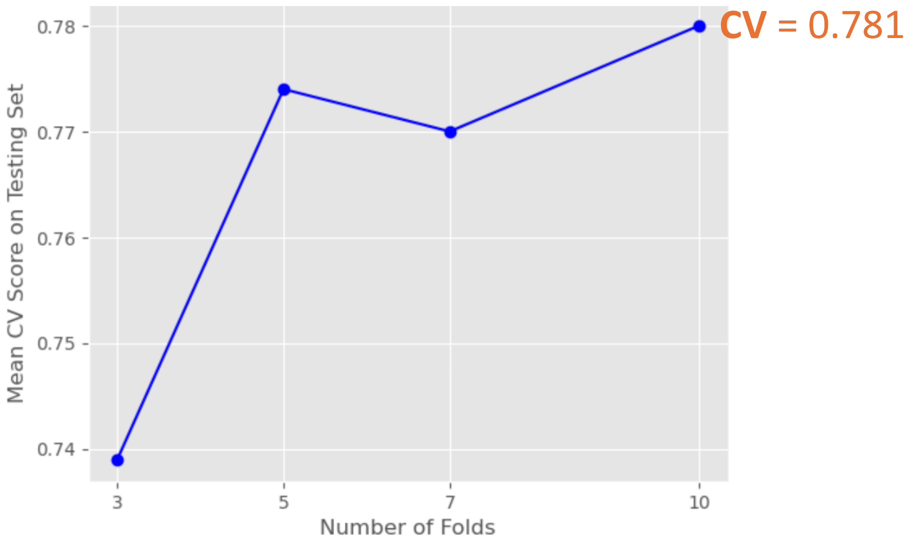


Figure 14. The mean cross-validation scores on testing set for the Random Forest Regression.

Modeling – Xgboost Regression

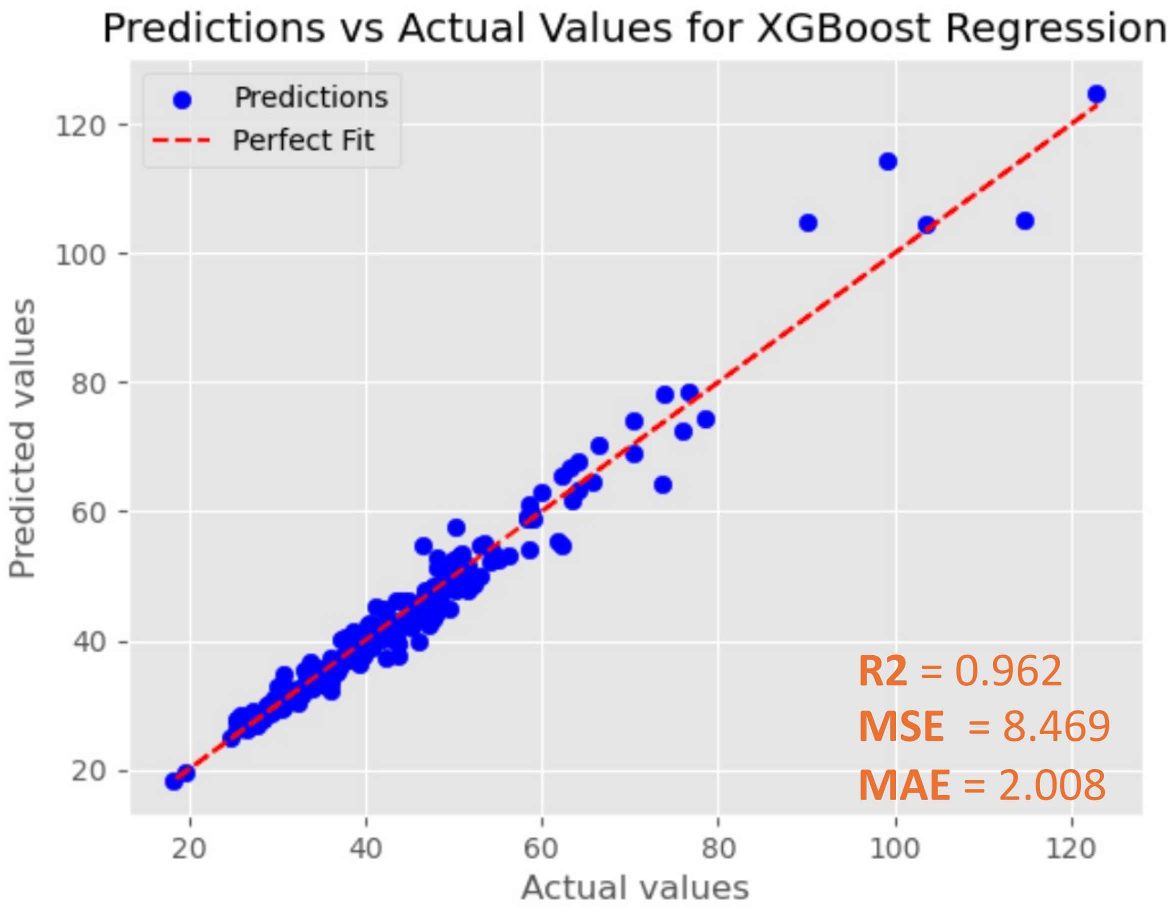


Figure 15. Relationship between the predicted and actual values generated by the XGBoost Regression model.

Modeling – XGboost Regression

Table 2. XGBoost Model Cross-Validation Results.

	Cross-Validation	Max Depth	N Estimators	Learning Rate	Best Score	Mean CV Score (Testing Set)	Standard Deviation
0	3-fold	3	300	0.1	0.954	0.883	0.055
1	5-fold	3	300	0.1	0.967	0.900	0.064
2	7-fold	3	300	0.1	0.976	0.907	0.085
3	10-fold	3	300	0.1	0.973	0.926	0.063

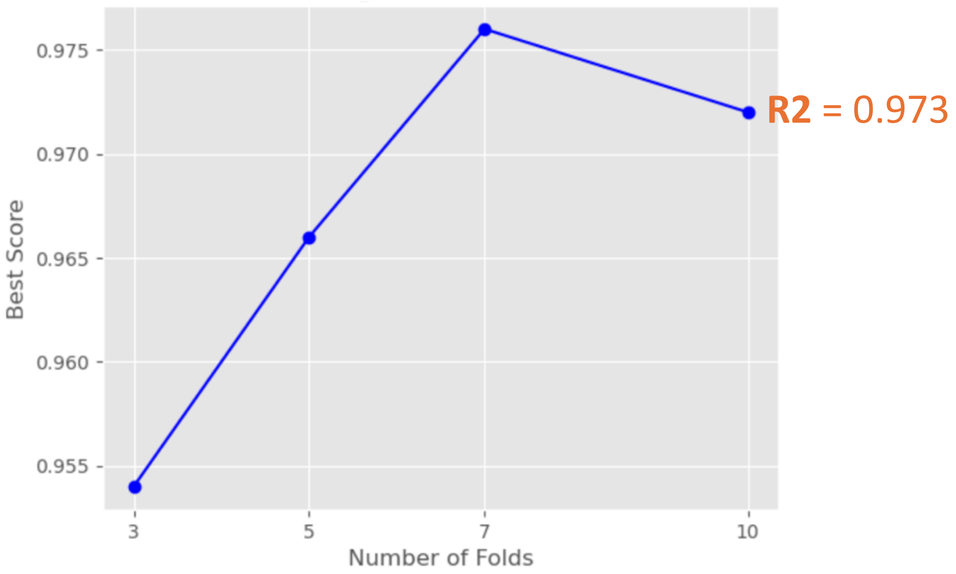


Figure 16. Best scores with different number of folds for the XGBoost model.

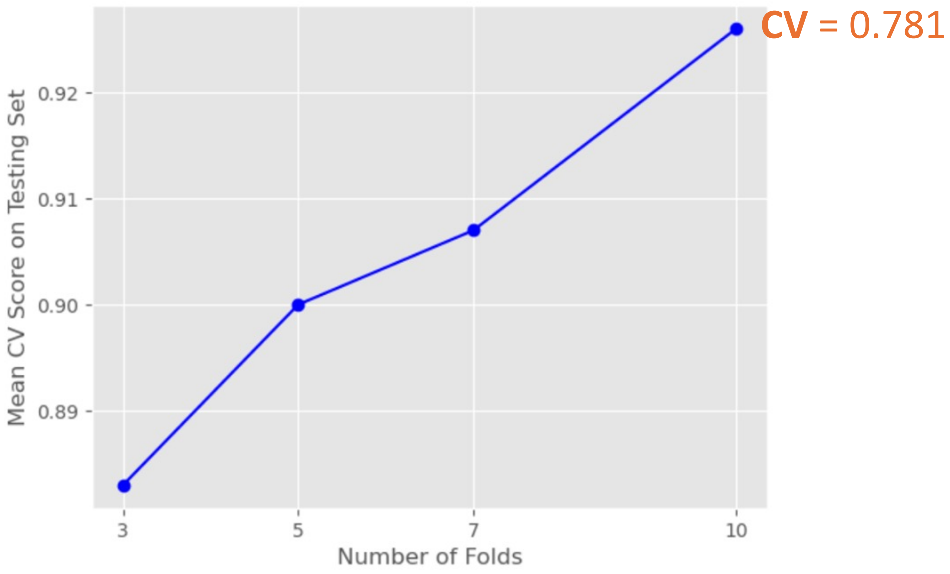


Figure 17. Mean cross-validation scores on the testing set for XGBoost regression model.

Results

Table 3. Comparison of regression models.

	Model	R-squared (R2)	Mean Squared Error (MSE)	Mean Absolute Error (MAE)	Mean Cross-Validation Score on Testing Set
0	Linear Regression	0.477	117.110	8.070	5-fold: 0.755
1	Lasso Regression	0.536	103.736	7.438	-
2	Ridge Regression	0.481	116.024	8.032	-
3	Random Forest Regression	0.923	17.285	2.665	10-fold: 0.781
4	XGBoost Regression	0.962	8.469	2.008	10-fold: 0.926

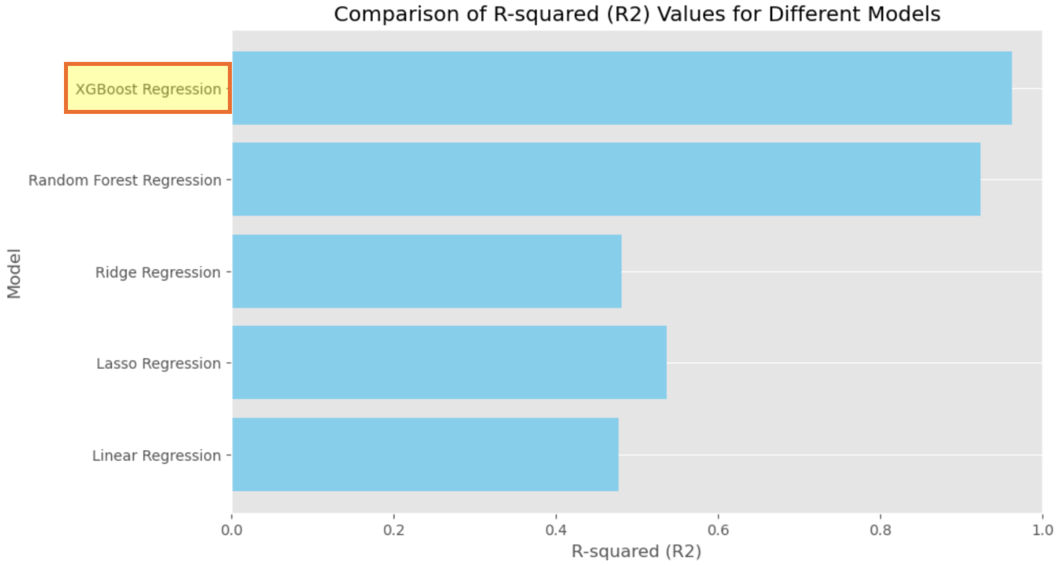


Figure 18. Comparison of R2 values for different models.

Conclusion

- **Top-Performing Model:** XGBoost Regression achieved R2 value of 0.962 with lowest MSE and MAE.
- **Importance of Selection:** Highlighted the significance of feature selection, model choice, and evaluation metrics for accurate prediction.
- **Insights and Limitations:** Provided insights into key factors influencing water quality prediction, acknowledging study limitations and potential areas for improvement.

Future Work

- **Feature Engineering:** Enhance predictive models by exploring additional factors.
- **Seasonal and Long-Term Trends**: Investigate temporal patterns in water quality.
- **Spatial Analysis**: Identify areas of concern using GIS tools.
- **Deep Learning Methods**: Improve predictive accuracy with neural networks.
- **Validation through Field Studies**: Ensure practical usefulness and reliability.

Questions?