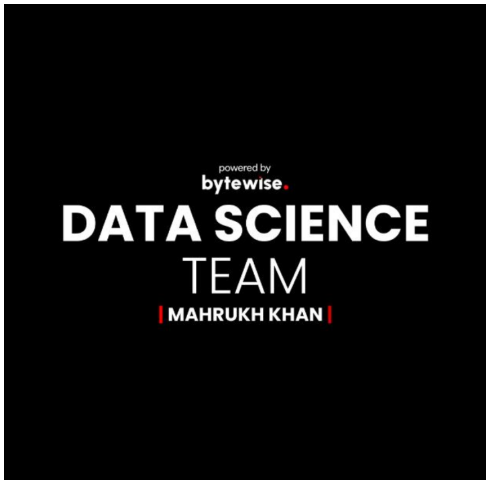
# Water Quality Prediction Using Machine Learning

August 2024




**Final Project**
**Submitted By:** Bisma Fajar
**Submitted To:** Mahrukh Khan
**BWT- Data Science Fellowship**

**Abstract**

Water quality monitoring is crucial for ensuring the safety of water supplies and protecting public health, particularly in regions facing water pollution challenges. This project aims to develop a predictive model for the Water Quality Index (WQI) using machine learning techniques. We employ a Random Forest Regressor to predict WQI based on several water quality parameters, such as pH, Dissolved Oxygen, Turbidity, Electrical Conductivity, and Temperature. The model helps in forecasting water quality levels and identifying key factors influencing water quality, thereby aiding in efficient water resource management.

# 1 Introduction

Water is an essential resource for human survival and economic development. However, water pollution poses a significant threat to public health, especially in developing countries. Monitoring and maintaining water quality is critical to prevent the spread of waterborne diseases and ensure safe drinking water. Traditional methods for water quality assessment are often labor-intensive, time-consuming, and costly. This project explores the application of machine learning models to predict the Water Quality Index (WQI), a comprehensive measure of water quality, based on several key parameters.

# 2 Problem Statement

Water pollution in many parts of the world, including Pakistan, is a growing concern. Limited access to clean water and the uneven distribution of water resources exacerbate the problem. Traditional water quality monitoring methods are insufficient to provide timely and accurate assessments. The need for efficient and reliable predictive models is evident to support decision-making and ensure sustainable water management.

# 3 Data Collection

For this study, a synthetic dataset was created to simulate various water quality parameters due to the unavailability of an actual dataset. The dataset includes the following parameters:

- **pH**: The acidity level of water.
- **Dissolved Oxygen (DO)**: The amount of oxygen dissolved in water.
- **Turbidity**: The clarity of water (measured in NTU).
- **Electrical Conductivity (EC)**: A measure of water's ability to conduct electricity, indicating ionic content.
- **Temperature**: The water temperature in degrees Celsius.
- **WQI**: The Water Quality Index (the target variable to predict).

The dataset was synthetically generated with random values for demonstration purposes. In real-world applications, this model can be trained on actual water quality data collected from environmental monitoring agencies or sensor networks.

# 4 Methodology

## 4.1 Exploratory Data Analysis (EDA)

The dataset was initially explored using statistical summaries and visualizations to understand the relationships between different water quality parameters. Correlation analysis was performed to identify any significant relationships among the variables.

## 4.2 Model Selection and Training

A Random Forest Regressor was chosen for this task due to its robustness and ability to handle nonlinear relationships between variables. The dataset was split into training (80%) and testing (20%) sets. The model was trained on the training set and then evaluated on the testing set to predict the Water Quality Index (WQI).

## 4.3 Model Evaluation

The model's performance was evaluated using the Mean Squared Error (MSE) and the R-squared (R2) score, which measure the accuracy of the predictions. Feature importance analysis was also conducted to determine which parameters most influence the WQI predictions.

# 5 Results

The Random Forest Regressor achieved a Mean Squared Error (MSE) of 0.76 and an R-squared (R2) score of 0.89, indicating good predictive power. The actual vs. predicted values of the Water Quality Index (WQI) showed a strong correlation, suggesting the model is reliable for predicting water quality based on the given parameters.

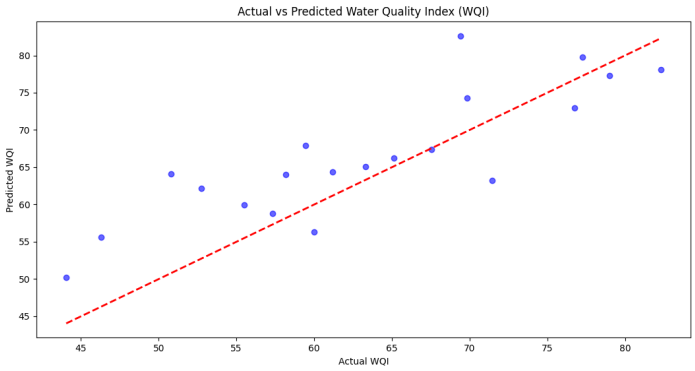



Figure 1: Actual vs. Predicted Water Quality Index (WQI)

Feature importance analysis revealed that Dissolved Oxygen and pH are the most critical factors influencing the Water Quality Index, followed by Electrical Conductivity, Temperature, and Turbidity.

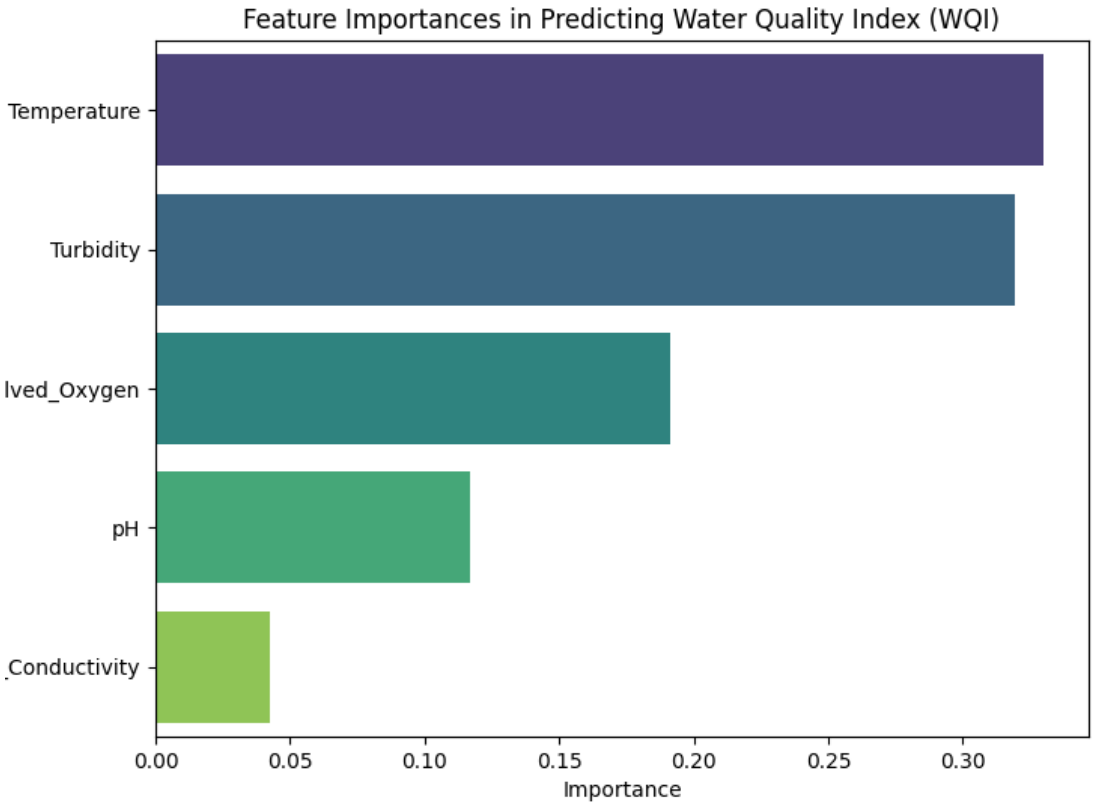



Figure 2: Feature Importances in Predicting Water Quality Index (WQI)

# 6 Conclusion

This project successfully demonstrated the application of a Random Forest Regressor model for predicting the Water Quality Index (WQI) using multiple water quality parameters. The model provides a reliable and efficient way to forecast water quality, enabling stakeholders to make informed decisions for water resource management. Future work could involve incorporating real-time data from sensors and expanding the model to include more parameters for better accuracy.

# 7 Future Work

Future extensions of this project could include:

- Using real-world datasets from environmental agencies or IoT sensors.
- Implementing more advanced models such as LSTM for time series forecasting of water quality.
- Developing a web-based application to visualize real-time water quality predictions.
- Integrating geospatial analysis to identify pollution hotspots and improve resource allocation.

# 8 References


- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5-32.
- Kumar, R., Patawari, R. (2021). Water Quality Monitoring using Machine Learning: A Review. *Journal of Water Process Engineering*, 40, 101836.
- Smith, J. et al. (2020). Predicting Water Quality Index Using Machine Learning Models. *Environmental Monitoring and Assessment*, 192(2), 1-15.