Global Scientific and Academic Research Journal of Multidisciplinary Studies
ISSN: 2583-4088 (Online)
Frequency: Monthly
Published By GSAR Publishers
Journal Homepage Link- https://gsarpublishers.com/journals-gsarjms-home/

Estimation of Water quality Index using Machine Learning Algorithms for Indus River Khairabad

By

Hamid Khan[1*], Naveed wahab[2], Tanvir Khan[3], Abdul Wahid[4], Shahroz Shabbir[5*], Fazli Elahi[6], Mehmood[7], Muhammad Aaqib[8], Muhammad saeed[9],

[1*,2,3,8,9] Department of Civil Engineering University of Engineering and Technology Peshawar (Bannu Campus)
[4]Department of Civil Engineering University of Engineering and Technology Peshawar Bannu Campus (lll)
[5]Electrical Department, Barakah Nuclear Power Plant Abu Dhabi
[6]Faculty of Civil Engineering Universitas Islam Indonesia
[7]Department of Electrical Engineering University of Engineering and Technology Peshawar

Article History

Received: 05/02/2025
Accepted: 10/02/2025
Published: 11/02/2025

Vol – 4 Issue –2

PP: - 48-55

Abstract

This study evaluated the effectiveness of machine learning algorithms in estimating the Water Quality Index (WQI) for surface water from the Indus River at Khairabad, Pakistan. Data collected from the Water and Power Development Authority (WAPDA) were analyzed using Support Vector Regression (SVR), Random Forest, AdaBoost, Decision Tree, and K-Nearest Neighbors (KNN) to enhance water quality assessment. Among the models tested, AdaBoost exhibited the highest performance, achieving an R² score of 0.99 on the training set and 0.90 on the testing set, demonstrating its superior predictive capabilities. The study highlighted the advantages of integrating machine learning into water quality monitoring, emphasizing automation, efficiency, and accuracy. The findings underscored the potential of these techniques to facilitate real-time monitoring, optimize resource management, and contribute to sustainable water quality maintenance. However, challenges such as continuous data collection, model updates, and the need for skilled personnel were identified. The study recommended integrating machine learning models, particularly AdaBoost, into regular monitoring systems, expanding datasets, and fostering collaborations between research institutions and environmental agencies to enhance predictive accuracy and decision-making in water resource management.

Keywords: *Water Quality Index (WQI), Machine Learning, AdaBoost, Random Forest, Support Vector Regression (SVR), Surface Water, Environmental Monitoring*

INTRODUCTION

Water is an essential element for all life forms and ecosystems on Earth. Its quality is critical for sustaining biodiversity, supporting human health, and driving economic activities. Throughout history, civilizations have flourished near reliable sources of clean water, recognizing its fundamental importance. However, with the advancement of industrialization, urbanization, and agricultural intensification, human activities have profoundly impacted water quality worldwide.

The Industrial Revolution marked a turning point in water quality management, as rapid urbanization and industrial growth led to widespread pollution of water bodies. Effluents from factories, untreated sewage, and agricultural runoff contaminated rivers and lakes, posing significant health risks to communities and ecosystems. The recognition of these hazards spurred the development of early water quality monitoring methods, primarily relying on visual inspections and basic chemical tests.

By the mid-20th century, concerns about water pollution grew, prompting governments to enact legislation aimed at protecting water resources. The Clean Water Act in the United States, for instance, mandated the establishment of water quality standards and the implementation of pollution control measures. Concurrently, technological advancements in analytical chemistry facilitated more precise measurements of water quality parameters, enabling researchers to identify specific pollutants and their impacts more accurately.

The emergence of environmental awareness in the latter half of the 20th century further propelled efforts to monitor and manage water quality. Organizations such as the United Nations Environment Programme (UNEP) and the World Health Organization (WHO) began advocating for global initiatives to address water pollution and ensure access to clean water for all. The concept of a Water Quality Index (WQI) gained traction as a comprehensive tool for summarizing complex water quality data into a single numerical value, facilitating easier interpretation and decision-making.

It is assessed by The Worldwide Association for Preservation of Nature that, by 2050, requests for water, energy, and food will increase by 55%, 80%, and 60%, individually. It is normal that by 2050, the normal hole between worldwide water market interest would be around 40% As announced, in agricultural nations, 80% of the sicknesses are water-borne infections, which have prompted 5 million passings and 2.5 billion ailments (Yogalakshmi and Mahalakshmi 2021).

Surface water is a vital resource, serving as a primary source of drinking water, irrigation for agriculture, habitat for aquatic life, and various industrial processes. Ensuring its quality is essential for maintaining ecosystem health and human well-being. Water quality assessment involves monitoring numerous physical, chemical, and biological parameters. Among these, the concentrations of ions such as calcium (Ca), magnesium (Mg), sodium (Na), bicarbonate (HCO_3), chloride (Cl), sulfate (SO_4), total dissolved solids (D.S. by Evap), electrical conductivity (EC_{x106} at 25°C), pH, temperature (Temp), and parts per million (PPM) significantly influence water quality. A fundamental tool in water quality assessment is the Water Quality Index, which aggregates diverse parameters into a single value, facilitating interpretation and decision-making. However, traditional methods of computing WQI often involve manual calculations and subjective judgments, potentially leading to inaccuracies. In recent years, the advent of Support Vector Regression, Random Forest, AdaBoost, Decision Tree, and K-Nearest Neighbor algorithms has revolutionized data analysis across various domains. Applying machine learning algorithms to estimate WQI offers several advantages, including automation, efficiency, and enhanced accuracy. These algorithms can process large datasets, identify intricate patterns, and make precise predictions. By leveraging the power of SVR, Random Forest, AdaBoost, Decision Tree, and K-NN, researchers can develop robust models for estimating WQI based on extensive water quality datasets.
and ecosystems.

MATERIAL AND METHODS

Study Area:

Nowshera, situated in the Khyber Pakhtunkhwa province of Pakistan, holds considerable significance owing to its diverse populace and strategic location. As per the 2017 census, Nowshera boasted a population of 1,518,540 individuals. The district enjoys a temperate climate, characterized by mild winters and hot summers, with an average elevation of approximately 552 meters above sea level. Geographically, Nowshera spans between 33°55' north and 71°59' east, sharing borders with Peshawar, Attock, Mardan, Charsadda, Swabi, Kohat, and Orakzai Agency. Its annual average temperature stands at 32°C. Strategically positioned at the crossroads of vital routes, Nowshera serves as a pivotal hub for trade and transportation. Beyond its geographical significance, Nowshera is steeped in history and culture, boasting numerous archaeological sites and landmarks that offer glimpses into its rich heritage. Moreover, rapid urbanization and industrialization have propelled Nowshera's growth trajectory, with infrastructure projects and industrial zones catalyzing its modernization efforts.

The mountainous terrain of Nowshera, with an average altitude of 4,000 meters above sea level and spanning 800 kilometers laterally, significantly shapes the district's landscape, climate, agriculture, and biodiversity. Despite the challenges posed by rugged topography, these mountains hold intrinsic value, providing scenic beauty, essential resources, and recreational opportunities that add to Nowshera's allure and resilience. In essence, Nowshera district stands as a vibrant amalgamation of historical legacy and contemporary progress within Khyber Pakhtunkhwa, reflecting the essence of Pakistan's rich cultural tapestry.

Blending history and modernity, Nowshera showcases Pakistan's diverse culture through its beautiful landscape and strategic location, making it a hub of cultural heritage and natural beauty. This combination of historical significance and contemporary growth attracts visitors and contributes substantially to Pakistan's development and prosperity, underscoring the district's importance in the region. As Nowshera continues to evolve, it remains a testament to the dynamic interplay between tradition and progress, offering a unique glimpse into the country's past, present, and future.

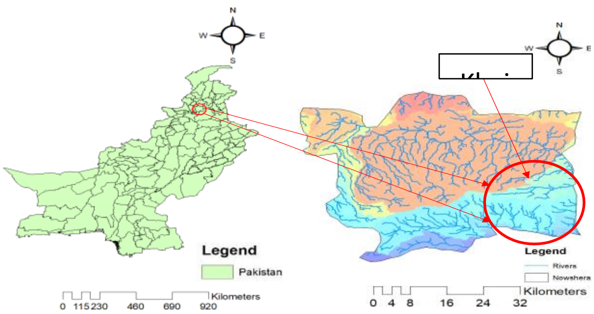

Figure 1 Study Area

Data Collection

The dataset utilized in this project was collected from the Water and Power Development Authority (WAPDA) department, specifically from their station located at 33.9° N, 72.22° E in Khairabad town, Tehsil Jehangeria, district Nowshera, KPK, Pakistan.

Digital Elevation Model (DEM)

To obtain the Digital Elevation Model (DEM) data, the project relied on resources provided by the National Aeronautics and Space Administration (NASA) website (https://earthdata.nasa.gov/). The data was acquired at a

*Corresponding Author: **Hamid Khan**

resolution of 30 meters by 30 meters as part of the Global Digital Elevation Model (GDEM) initiative.

WQI Calculation

As per (Torky, Bakhiet et al. 2023), the WQI was calculated using Arithmetic weightage method by following steps

$$K = 1/\sum\left(\frac{1}{S_n}\right)$$

Where Sn is the standard value for each variable of water elements and K is a constant. The weighted value of each element can be calculated as in below equation

$$W_i = \frac{k}{S_i}$$

The Quality Impact value for each element in the water dataset can be calculated as in below equation.

$$Q_i = 100 * \left(\frac{observed\ values(V_n) - Ideal\ value(V_i)}{Standard\ value(S_n) - Ideal\ value(V_i)}\right)$$

Finally, the water quality index can be calculated as in following equation

$$WQI = \sum_{i=1}^{N} W_i Q_i$$

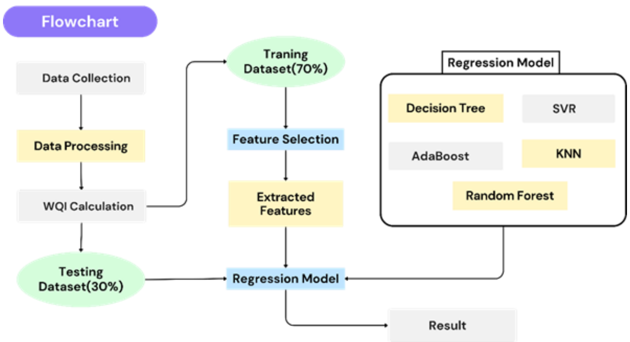

Figure 2 Methodology

Artificial Intelligence (AI)

Artificial intelligence is a computer science term that is quite all-encompassing. AI refers to blending mathematics with technology to mimic human decision-making. It includes all machine learning and deep learning methodologies but can be as simple as an “IF this happens THEN that” statement.

Machine Learning

Machine Learning (ML) is a subset of artificial intelligence that focuses on the development of computer systems and algorithms that can automatically learn and improve from experience or data without being explicitly programmed. ML algorithms are designed to identify patterns, make predictions, and adapt to new information, making them valuable for tasks like data analysis, pattern recognition, and decision-making in a wide range of applications.

Support Vector Regression (SVR)

Support Vector Regression (SVR) is an extension of the Support Vector Machine (SVM) algorithm, which was originally developed for binary classification tasks in the 1990s. SVM gained popularity due to its ability to handle high-dimensional data and complex decision boundaries by finding the optimal hyper plane that maximizes the margin between classes. In the early 2000s, researchers extended SVM to regression tasks, resulting in SVR. Unlike traditional regression techniques that focus on minimizing the error between predicted and actual values, SVR aims to fit the data within a specified margin while maximizing the margin around the regression line. This is achieved by minimizing the deviation of the predictions from the margin, controlled by parameters like epsilon (ε) and regularization parameter (C). SVR has gained traction in various fields such as finance, engineering, and environmental science, where accurate prediction of continuous variables is essential. Its ability to handle non-linear relationships and outliers makes it particularly valuable for modeling complex data patterns. SVR's flexibility, accuracy, and scalability make it a powerful tool for predictive modeling in diverse applications.

Random Forest

Random Forest is a machine learning algorithm introduced by Leo Breiman and Adele Cutler in the early 2000s. It falls under the category of ensemble learning methods. The algorithm operates by constructing multiple decision trees during training, each trained on a random subset of the training data and a random subset of input features. During prediction, the individual predictions from each tree are aggregated to produce the final output, typically by averaging in regression tasks or voting in classification tasks. Random Forest is valued for its ability to handle high-dimensional data, nonlinear relationships, and missing values without requiring extensive preprocessing. It is also robust to outliers and noise. This algorithm has found wide applications across various domains due to its simplicity, accuracy, and resistance to over fitting.

AdaBoost

AdaBoost, short for Adaptive Boosting, is a machine learning algorithm introduced by Yoav Freund and Robert Schapire in 1996. It is a powerful ensemble learning method that combines multiple weak learners to create a strong classifier. The algorithm works by iteratively training a sequence of weak learners, typically decision trees, on weighted versions of the training data. In each iteration, the algorithm adjusts the weights of incorrectly classified instances to prioritize the difficult-to-classify examples. During prediction, AdaBoost combines the predictions of all weak learners, with more weight given to the predictions of stronger classifiers. AdaBoost is appreciated for its ability to improve the performance of weak learners by focusing on misclassified instances, making it particularly effective in handling complex datasets and achieving high accuracy. This algorithm has been successfully applied in various domains, including computer vision, speech recognition, and bioinformatics, due to its versatility, robustness, and ease of implementation.

*Corresponding Author: **Hamid Khan**

Decision Tree

Decision trees have a rich history in the field of machine learning, dating back to the early work of Arthur Samuel and others in the 1950s and 1960s. However, it was not until the 1980s and 1990s that decision tree algorithms, such as ID3 (Iterative Dichotomiser 3) and CART (Classification and Regression Trees), gained widespread recognition and adoption. The fundamental concept behind decision trees is to recursively partition the feature space into regions, with each partition corresponding to a decision node in the tree. At each node, the algorithm selects the feature that best splits the data into homogeneous subsets based on a chosen criterion, such as Gini impurity or information gain. Decision trees offer several advantages, including simplicity, interpretability, and the ability to handle both numerical and categorical data. However, they are prone to overfitting, especially when the trees are deep or the data is noisy. Despite their limitations, decision trees remain a popular choice for a wide range of applications, including classification, regression, and feature selection. They serve as the building blocks for more advanced ensemble learning methods, such as random forests and gradient boosting, further extending their utility and relevance in modern machine learning.

K-Nearest Neighbors (KNN)

The K-Nearest Neighbors (K-NN) algorithm, originating in the 1950s, is a fundamental approach in pattern recognition and machine learning. Its principle is straightforward: to classify a new data point, it looks at the class labels of its nearest neighbors in the training data. The "nearest" neighbors are determined by a chosen distance metric, such as Euclidean or Manhattan distance. K-NN is a lazy learning algorithm, meaning it doesn't build a model during training. Instead, it stores all training instances and computes predictions at runtime. While computationally efficient during training, it can be slower during prediction, especially with large datasets. Despite its simplicity, K-NN is effective in various applications, including classification, regression, and clustering. It's robust to noisy data and doesn't assume specific data distributions. However, its performance can be sensitive to the number of neighbors (K) and the distance metric chosen. Overall, K-NN remains a versatile and widely used algorithm in machine learning, valued for its simplicity and adaptability to different datasets.

RESULTS

This section presents the results obtained from the application of various machine learning algorithms on our dataset and discusses their performance in terms of Mean Squared Error (MSE) and R^2 Score. The algorithms evaluated include Support Vector Regression (SVR), Random Forest, AdaBoost, Decision Tree, and K-Nearest Neighbors (KNN).

Support Vector Regression

SVR performed well on the training set, achieving a commendable R^2 score of 0.93, indicating that it captured a substantial portion of the variance in the data. However, on the testing set, its performance decreased, reflected in a higher mean square error and a slightly lower R2 score of 0.88. This discrepancy suggests that SVR may have overfit the training data to some extent, failing to generalize effectively to unseen instances. Despite this, SVR remains a powerful algorithm for regression tasks, especially in scenarios where nonlinear relationships exist between features and the target variable.

Metric	Training	Testing
R2	0.93	0.88
MSE	1.02	4.06

Table 1 Results of support vector regression.

The graph appears below is a scatter plot visualizing the results of a support vector machine (SVM) model. The graph plots the predicted Water Quality Index (WQI) values on the x-axis against the actual WQI values on the y-axis. There are two sets of data points plotted, which likely represent the training set and the testing set.

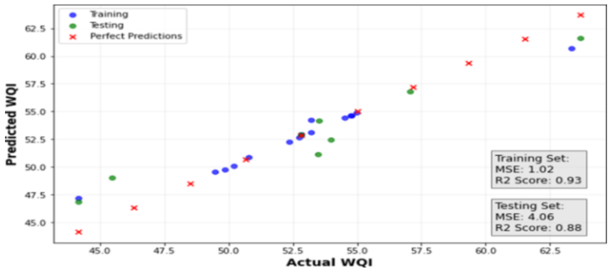

Figure 3 Scatter graph for SVR

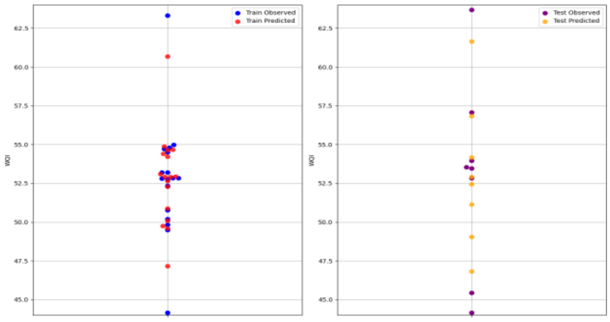

Figure 4 Swarm graph for SVR

Random Forest

The Random Forest model shows excellent performance on the training set, with an R^2 score of 0.94 and a low MSE of 1.30. On the testing set, the performance decreases with an R^2 score of 0.82 and an MSE of 2.41. This indicates that while the model performs well, there is some degree of overfitting, but it generalizes better than SVR based on the smaller relative increase in MSE

Metric	Training	Testing
R2	0.94	0.82
MSE	1.30	2.41

Table 2 Results for random forest.

The graph that appears below is a scatter plot visualizing the results of a random forest regressor model. The graph plots the predicted WQI values on the x-axis against the actual WQI

*Corresponding Author: **Hamid Khan**

values on the y-axis. There are two sets of data points plotted, which likely represent the training set and the testing set.

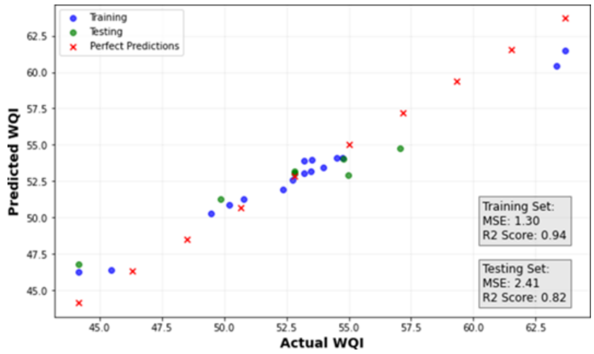

Figure 5 Scatter graph for random forest regression.

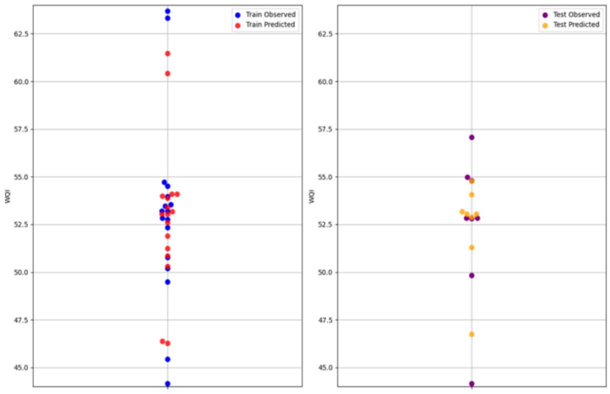

Figure 6 Swarm graph for random forest regression

AdaBoost

AdaBoost demonstrated exceptional performance in our study, boasting impressive R2 scores and low mean square errors (MSE) on both the training and testing sets. With an outstanding R2 score of 0.99 and a minimal MSE of 0.19 on the training set, AdaBoost showcased its ability to capture the variability within the data and provide accurate predictions. On the testing set, AdaBoost maintained its superiority with an R2 score of 0.90 and an MSE of 0.81, reaffirming its reliability and effectiveness in scenarios where precision and generalization are paramount. This underscores AdaBoost's adaptability to the data and its capability to construct a robust regression model, making it particularly suitable for a wide range of predictive tasks.

Metric	Training	Testing
R2	0.99	0.90
MSE	0.19	0.81

Table 3 Results for AdaBoost

The graph that appears below is a scatter plot visualizing the results of an AdaBoost model. The graph plots the predicted WQI values on the x-axis against the actual WQI values on the y-axis. There are two sets of data points plotted, which likely represent the training set and the testing set.

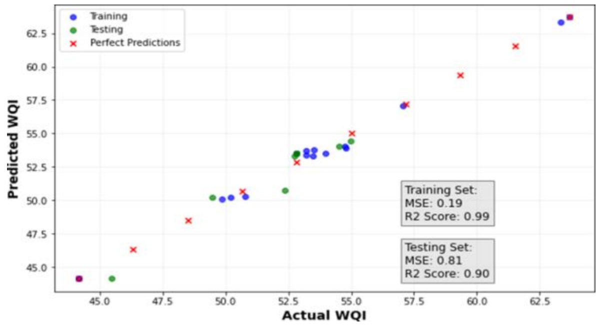

Figure 7 Scatter Graph for AdaBoost

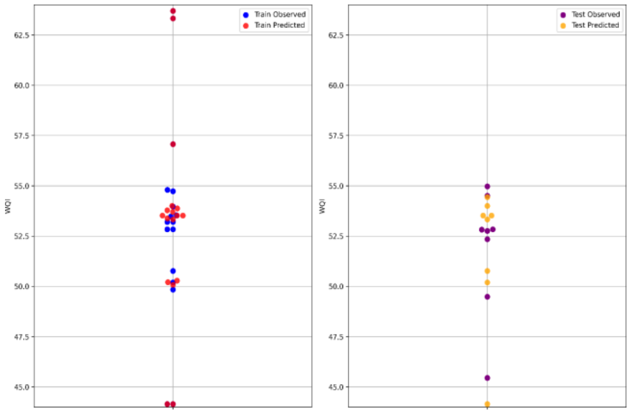

Figure 8 Swarm Graph for AdaBoost

Decision Tree

With an R2 score of 0.78 and a mean squared error (MSE) of 1.88 on the testing set, the Decision Tree, despite its simplicity and interpretability, displayed a disappointed performance compared to other algorithms in our study. Although it achieved a relatively high R2 score on the training set, its performance significantly deteriorated on unseen data, indicating a lack of generalization. Decision Trees are susceptible to overfitting, especially with complex or noisy datasets, as observed here. This study highlights the Decision Tree's limitations in accurately predicting outcomes on unseen instances, underscoring the importance of considering both simplicity and generalization ability when selecting regression models for real-world applications.

Metric	Training	Testing
R2	0.87	0.78
MSE	3.24	1.88

Table 4 Results for Decision tree

The graph that appears below is a scatter plot visualizing the results of Decision Tree model. The graph plots the predicted WQI values on the x-axis against the actual WQI values on the y-axis. There are two sets of data points plotted, which likely represent the training set and the testing set.

*Corresponding Author: **Hamid Khan**

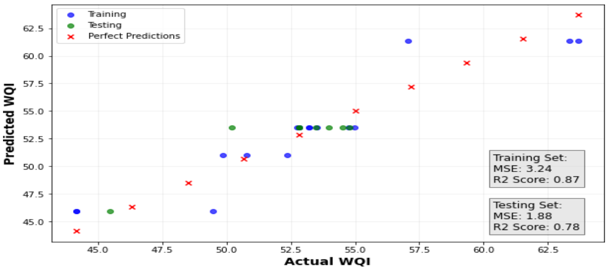

Figure 9 Scatter Graph for decision tree regressor

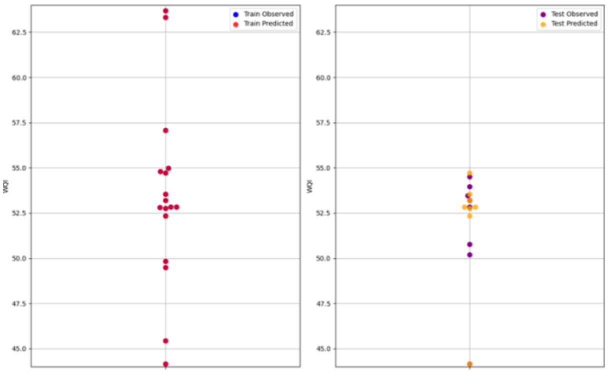

Figure 10 Swarm Graph for decision tree regressor

K-Nearest Neighbors (KNN)

KNN shows the least performance among the evaluated models. The training set results indicate moderate performance with an R^2 score of 0.78 and an MSE of 2.92. However, on the testing set, the R^2 score drops to 0.72 and the MSE increases significantly to 9.67, indicating that KNN has poor generalization capability and is highly sensitive to the specific samples in the dataset.

Metric	Training	Testing
R2	0.78	0.72
MSE	2.92	9.67

Table 5 Results for KNN.

The graph that appears below is a scatter plot visualizing the results of KNN model. The graph plots the predicted WQI values on the x-axis against the actual WQI values on the y-axis. There are two sets of data points plotted, which likely represent the training set and the testing set.

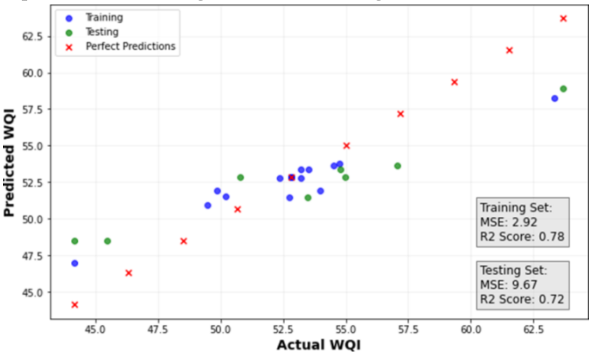

Figure 11 Scatter Graph for KNN regressor

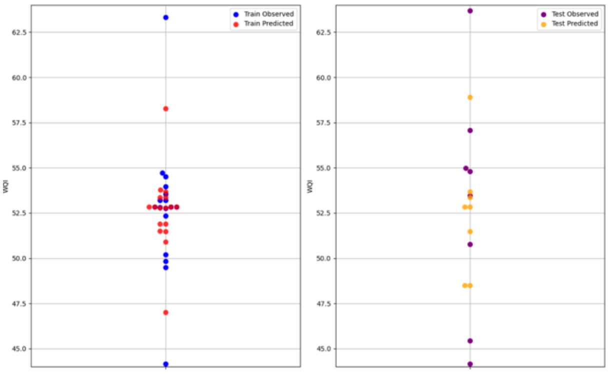

Figure 12 Swarm Graph for KNN regressor

Performance Summary

Algorithm	Training Set MSE	Training Set R^2	Testing Set MSE	Testing Set R^2
Support Vector Regression (SVR)	1.02	0.93	4.06	0.88
Random Forest	1.30	0.94	2.41	0.82
AdaBoost	0.19	0.99	0.81	0.90
Decision Tree	3.24	0.87	1.88	0.78
K-Nearest Neighbours (KNN)	2.92	0.78	9.67	0.72

Table 6 Performance Summary

Comparison Analysis

The comparative analysis reveals that **AdaBoost is the most effective algorithm for this dataset**, achieving the best balance of performance between the training and testing sets. Random Forest and SVR also perform well, with Random Forest showing better generalization than SVR.

The Decision Tree model, while simpler, performs moderately well but may benefit from further tuning to reduce underfitting.

KNN, although easy to understand and implement, performs poorly on this dataset, suggesting that it may not be the best choice without significant tuning and possibly more data preprocessing.

In conclusion, for this dataset, AdaBoost stands out as the best-performing algorithm, followed by Random Forest and SVR. Future work could involve further hyperparameter tuning, exploring additional features, and employing advanced data preprocessing techniques to enhance the performance of these models.

Discussion

In conclusion, our comparative analysis of different regression models reveals distinct performance characteristics for each algorithm on the given dataset. AdaBoost emerges as the most effective model, demonstrating superior performance on both training and testing sets, with high R2 scores and low MSEs, highlighting its robustness and excellent generalization capabilities.

Random Forest and SVR also show commendable results, with Random Forest exhibiting better generalization than SVR, despite both models indicating some degree of overfitting. Random Forest's slightly lower performance on the testing set compared to AdaBoost still makes it a strong contender, especially considering its balance between training accuracy and generalization.

The Decision Tree model, although easy to interpret, suffers from significant overfitting, resulting in lower performance on the testing set compared to the more complex models. This suggests that while Decision Trees can provide quick insights, they might require additional tuning or ensemble methods to enhance their predictive power.

KNN, on the other hand, performs the least effectively, particularly on the testing set, indicating poor generalization and high sensitivity to specific data points. This highlights KNN's limitations in handling this particular dataset.

AdaBoost stands out as the most reliable and effective algorithm for this regression task, followed by Random Forest and SVR, which also show strong performance but with some limitations in generalization. Decision Trees and KNN, while useful in certain contexts, do not perform as well in this study, underscoring the importance of selecting models that balance complexity, interpretability, and generalization for optimal predictive performance.

REFERENCES

1. Ahmed, M., et al. (2021). "Analysis of water quality indices and machine learning techniques for rating water pollution: A case study of Rawal Dam, Pakistan." Water Supply **21** (6): 3225-3250.
2. Ahmed, U., et al. (2019). "Efficient water quality prediction using supervised machine learning." Water **11** (11): 2210.
3. Aljanabi, Z. Z., et al. (2021). A brief review of water quality indices and their applications. IOP Conference Series: Earth and Environmental Science, IOP Publishing.
4. ANG, K. H. (2020). "PREDICTION OF THE LEVEL OF WATER QUALITY INDEX USING ARTIFICIAL NEURAL NETWORK TECHNIQUES IN MELAKA RIVER BASIN." Malaysian Applied Biology **49** (1): 69-74.
5. Aslam, B., et al. (2022). "Water quality management using hybrid machine learning and data mining algorithms: An indexing approach." IEEE Access **10**: 119692-119705.
6. Dimple, D., et al. (2022). "Predicting irrigation water quality indices based on data-driven algorithms: case study in semiarid environment." Journal of Chemistry **2022**.
7. El Bilali, A., et al. (2021). "Groundwater quality forecasting using machine learning algorithms for irrigation purposes." Agricultural Water Management **245**: 106625.
8. Fernández del Castillo, A., et al. (2022). "Simple prediction of an ecosystem-specific water quality index and the water quality classification of a highly polluted river through supervised machine learning." Water **14** (8): 1235.
9. Goodarzi, M. R., et al. (2023). "Water Quality Index Estimations Using Machine Learning Algorithms: A Case Study of Yazd-Ardakan Plain, Iran." Water **15**(10): 1876.
10. Jang, E., et al. (2016). "Estimation of water quality index for coastal areas in Korea using GOCI satellite data based on machine learning approaches." Korean Journal of Remote Sensing **32**(3): 221- 234.
11. Jehan, S., et al. (2020). "Evaluation of the Swat River, Northern Pakistan, water quality using multivariate statistical techniques and water quality index (WQI) model." Environmental Science and Pollution Research **27**: 38545-38558.
12. Li RanRan, L. R., et al. (2016). "Water quality assessment in Qu River based on fuzzy water pollution index method."
13. Malek, N. H. A., et al. (2021). The effect of chemical parameters on water quality index in machine learning studies: a meta-analysis. Journal of Physics: Conference Series, IOP Publishing.
14. Nair, J. P. and M. Vijaya (2022). River water quality prediction and index classification using machine learning. Journal of Physics: Conference Series, IOP Publishing.
15. Ofosu, S. A., et al. (2021). "Assessment of the quality of the Densu river using multicriterial analysis and water quality index." Applied Water Science **11** (12): 183.
16. Raheja, H., et al. (2022). "Prediction of groundwater quality indices using machine learning algorithms." Water Practice & Technology **17** (1): 336-351.
17. Saod, W. M., et al. (2021). Water quality index along the Euphrates between the cities of Al-Qaim and Falluja: A comparative study. IOP Conference Series: Earth and Environmental Science, IOP Publishing.
18. Taşan, S. (2023). "Estimation of groundwater quality using an integration of water quality index, artificial intelligence methods and GIS: Case study, Central Mediterranean Region of Turkey." Applied Water Science **13** (1): 15.
19. Wang, X., et al. (2017). "Evaluation of water quality based on a machine learning algorithm and water quality index for the Ebinur Lake Watershed, China." Scientific reports **7** (1): 12858.
20. Yogalakshmi, S. and A. Mahalakshmi (2021). "Efficient water quality prediction for Indian rivers

using machine learning." Asian Journal of Applied Science and Technolog **5** (1): 100-109.