Journal of Ecohumanism
2024
Volume: 3, No: 8, pp. 516 – 552
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v3i8.4750

Water Quality Index Classification of Southeast, South and West Asia Rivers using Machine Learning Algorithms

Hassan Shaheed[1], M H Zawawi[2], Gasim Hayder[3]

Abstract

In recent years, a number of contaminants have significantly impacted the quality of water. The ecosystem and human health are directly impacted by the water quality. Effective water management is indicated by the water quality index. The ability to forecast and simulate water quality has become crucial in the battle against water pollution. The goal of the study is to create a reliable model that will classify the index value in accordance with the requirements for water quality and predict the water quality using latest ML models. The information was gathered from several sample points dispersed across rivers in India, Iraq, and Malaysia. 32 variables that have an impact on water quality, such as temperature, dissolved oxygen, pH, alkalinity, hardness, chloride, and coliform, are used to calculate the water quality index. Datasets are constructed using pre-processed data, including normalisation, outlier identification, and the resolution of any class imbalance concerns. The water quality is classified using machine learning methods such XGBoost, Naive Bayes, SVM, and Ada Boost for measuring the water quality index whereas the prediction of water performed using RF regressor, M5 Model Tree, DT regressor, EML regressor on the samples of Malaysian, Indian, and Iraqian rivers. The performance of XGBoost accurately identifies the water quality index with 93%, 92%, and 97% Accuracy, Precision and recall respectively. Whereas the performance of M5 Model Tree for WQ prediction is much better than other regression models. The developed models provide a promising result for the classification of water quality indexes and prediction.

Keywords: *Water Quality Index, River Quality, Classification Model, Machine Learning Algorithms.*

Introduction

Water is more important for each living being of the universe, especially for mankind without it, the survival of them is really very hard. The sustainability of everything that exists on earth depends on improved access to high-quality water. Aquatic animals can withstand a certain level of pollution, but as it worsens, the oxygen concentration of the water drops and disasters result. There are quality criteria for many environmental water sources, such as streams, rivers, and lake waters, attesting to their value. Regulations apply to all forms of bodies of water for all purposes and applications. The ecology need not be harmed by irrigation water being too salty or damaging to the soil or plants. Various levels of water quality may be necessary based on the type of commercial operation [1].

The most expensive source of Natural water are ground and surface water. Resources can get contaminated by human, industrial, and other natural activities. As a result, rapid industrial expansion has resulted in a marked decline in water quality. Infrastructure, a lack of public knowledge, and weak cleanliness standards all have a substantial influence on the quality of drinking water. The implications of water pollution on infrastructure, the environment, and human health are highly substantial [2].

Over 2.2 billion people do not have access to safe drinking water services, according to United Nations research (WHO/UNICEF, 2019). In less developed countries, it has been shown that contaminated water is to blame for 80% of health problems. Water-borne infections affect 2.5 billion people annually, and five million individuals die away as a consequence [3].

[1] Department of Civil Engineering and Institute of Energy Infrastructure (IEI), Universiti Tenaga Nasional (UNITEN), Selangor Darul Ehsan Malaysia, Email: ehassan.en37@yahoo.com, (Corresponding Author), Republic of Iraq - Ministry of Planning

[2] Department of Civil Engineering and Institute of Energy Infrastructure (IEI), Universiti Tenaga Nasional (UNITEN), Selangor Darul Ehsan Malaysia, Email: MHafiz@uniten.edu.my

[3] Department of Civil Engineering and Institute of Energy Infrastructure (IEI), Universiti Tenaga Nasional (UNITEN), Selangor Darul Ehsan Malaysia, Institute of Energy Infrastructure (IEI), Universiti Tenaga Nasional (UNITEN), Kajang, 43000, Selangor Darul Ehsan, Malaysia, Email: gasim@uniten.edu.my:

Analysis and forecasting of the Water Quality Index (WQI) need novel approaches. To track the seasonal change of the WQI, it is suggested to do study on the temporal aspect of predicting water quality changes. It is more efficient to use a specific model variation rather than just one model to forecast the results of water quality. There are many potential classification techniques for WQI water quality. There is a lot of usage of statistics, visual modelling, and algorithm analysis [4]. The industrial revolution, widespread use of pesticides and fertilisers, and rapid population increase all seem to have had a negative impact on water quality ecosystems.

This study aims to develop a use case for classification and prediction of water quality as well as to provide a precise model for WQI parameters at different surfaces. Seven sample sites for Indian Bhavani River, which runs through Kerala and Tamil Nadu, have yielded data on the river's water quality. The statistics for the Malaysian rivers Klang and Langat were gathered from two stations. The Tigris and Euphrates rivers in Iraq have data gathered from six sites. During the analysis, it is found that data samples and models are more comprehend the association and feature distribution. Indexes of water quality are classified using machine learning techniques and the prediction of WQ are predicted using M5 Model Tree, Decision Tree Regressor, Random Forest Regressor, M5 Model Tree, Extreme Learning Machine (ELM Regressor). The performance of regression models is measured with MSE, RMSE, MAE, MAPF, Scatter Index and Bias.

The following is the structure of this paper: Section 2 discusses related work, while Section 3 discusses datasets, geographical information, WQI, and data preparation. Section 4 delves into the classification and prediction models and how the performance of several machine learning algorithms was measured. Section 5 discusses the findings, and the conclusion, together with general findings and future recommendations, is offered at the end.

Related Work

The methods that were employed to successfully solve problems with water quality are analysed in this research. While conventional statistical analysis and laboratory testing are often used in research to establish the quality of the water, other studies use machine learning techniques to identify the best ways to address the issue of water quality.

To categorise the Chao Phraya River's water quality, Sillberg et al. in [5], developed a strategy using ML that fuses attribute-realization (AR) with SVM. Having 0.94 accuracy, 0.84 precision and 0.84 average and 0.84 F1-score. The SVM linear approach delivered the most favourable outcomes for classification. The validation showed that, when used with between three and six criteria, AR-SVM was an effective method for estimating river water quality with an accuracy of 0.86 to 0.95.

An artificial neural network was employed by Yilma et al. [6] to replicate the Akaki river water quality indicator. Twelve indicators of water quality from 27 sites collected throughout the dry and wet seasons were used to generate the index. Their study found that an artificial neural network with 15 hidden neurons and eight hidden layers could predict the WQI with an accuracy of 0.93.

The water quality index (WQI) was evaluated by Ahmed et al. [7] using supervised machine learning techniques, where the total water quality and water quality class were summarized using a single index. The recommended approaches, including gradient boosting with a learning rate of 0.1 and polynomial regression with a degree of 2, were more effective in predicting the WQI, which was then evaluated with the highest classification accuracy, 85.07 %.

Sakizadeh [8] predicted the WQI using Bayesian regularisation and 16 water quality indicators. The study discovered that the observed and anticipated values had correlation coefficients of 0.94 and 0.77, respectively. As can be observed, similar work served as an inspiration for the current study, which shows that machine learning can effectively identify irregularities in water quality. The frequency of inaccurate predictions might be greatly decreased using machine learning techniques.

To begin with a ML models and algorithms for water quality many analyses were conducted by Jitha and Vijya in [9]. In this many aspects of ML and BD were highlighted which shows a clear picture of WQI issues in worldwide rivers. Furthermore, the same authors presented one model for River Water Quality Prediction and index classification using Machine Learning in [4], which uses different ML algorithms such as SVM, Naïve Bayes, DT and MLP classifiers on Bhavani River. The results show the classification accuracy is more than 81 %.

The same authors predicted DO Concentration utilising Prediction Models [10] such as RNN, LSTM, RF, SVR, MLP Regressor, and Linear regression. The LSTM has a performance of more than 88 percent. The authors of [11] conducted tests on the Bhavani River employing Water Quality Index Prediction Models such as LSTM, GRU, RNN, RF, SVR, MLP Regressor, and Linear regression. The GRU prediction model outperforms all others.

Recently, [12] employed temporal fusion transformer to forecast WQI on 200 epochs of RNN, LSTM, and GRU models, and the results reveal that the performance of TFT with Adam optimizer is superior to all.

Many machine learning techniques were employed in [13-14,16-18] to predict and categorise the water quality index. While ML has advanced, ensemble techniques such as the XGBoost family have been employed in [15,20-21]. Whereas in [19], the WQI of Iraq was determined without the use of ML methods.

In most recent advancement, ML have played an important role for water quality prediction and assessment. A giant team of researcher worked on many key areas of WQI using different approaches to produce best results. In 2018, they have used different evolutionary computer-based formulations to measure the WQ prediction [22], Also in 2019, Najafzadeh and team conducted many experiments on water quality. Research team have used ML methods to predict biological-based oxygen demand as well as a chemical-based oxygen demand [23]. Later in 2021, they have proposed a data driven models for reliability assessments of water quality in natural streams [24]. As well as SV Regressor was adopted to measure the WQ parameters [25].

Recently in 2023, they have used ML models to derive empirical equations for WQ and hydro morphological parameters [26]. Also used AI Models and remote sensing for evaluation of WQI [27].

Based on the studies [23-27] and their output brings evolution in the field of ML as well as Environmental and Water Quality which helps many researchers to do more research on WQ and WQI.

Overview of Study Area and Datasets

In this research three different datasets from different countries are used. The aim of using heterogeneous features from different rivers data is to check the performances of ML classifiers and regression. Among three datasets first one is Klang and Langat rivers (Malaysia) and Tigires-Euphrates (Iraq) and Bhavani River (India).

ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v3i8.4750

Figure 1 shows the Langat and Klang rivers which flows through Kaula Lumpur- Selangor to Malacca.

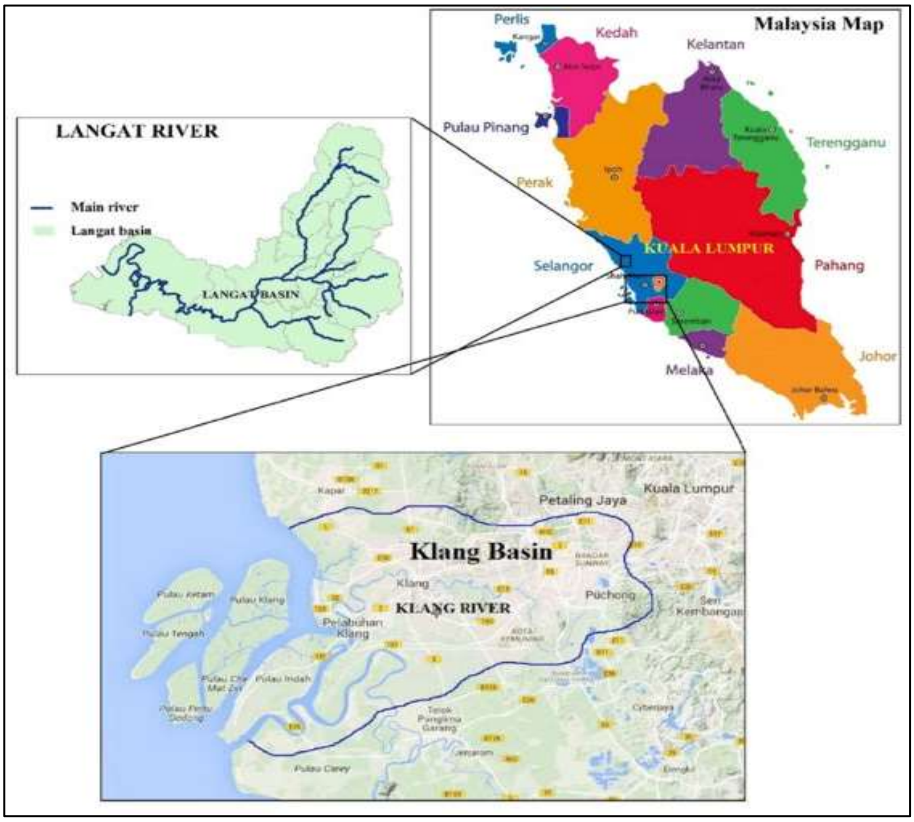


Figure 1. Map of Klang and Langat Rivers Malaysia

Similarly, the flow of two famous Iraqian rivers (Tigris and Euphrates Rivers) shown in Figure 2also it clearly shows the flow of shared water in western Asia.

Journal of Ecohumanism
2024
Volume: 3, No: 8, pp. 516 – 552
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v3i8.4750

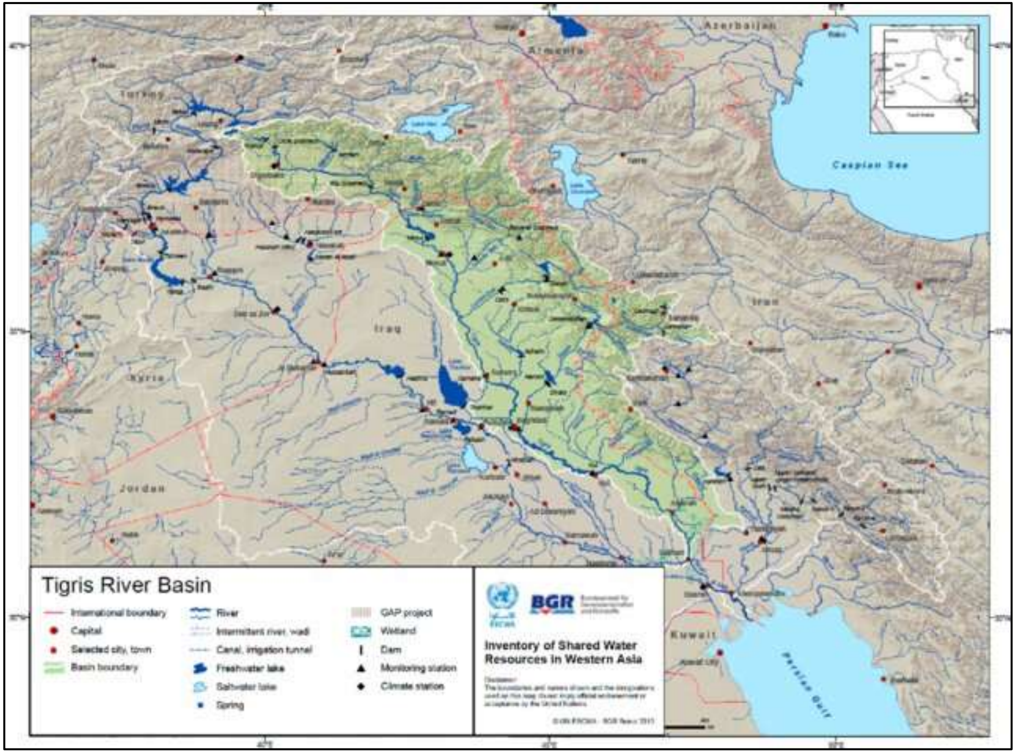


Figure 2. Map of Tigris and Euphrates Rivers Iraq

In India, the Bhavani River runs through Kerala and Tamil Nadu. The river rises in the Nilgiri Hills, passes through Tamilnadu, Kerala, and the Silent Valley National Park. The course of the Bhavani River, which largely traverses the Attappady Plateau in the Palakkad district before flowing into the Tamil Nadu districts, is shown in Fig. 3. The Bhavani River has many stations, but data has been gathered from 7 stations only.

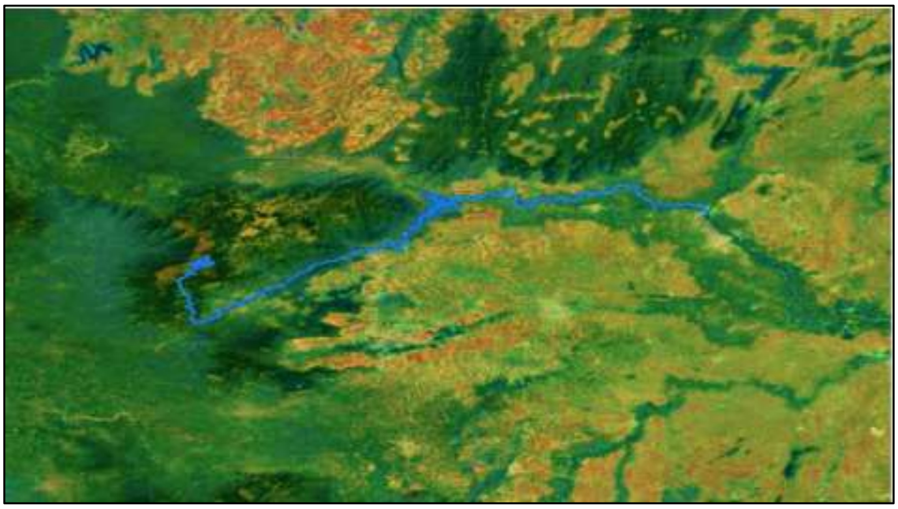


Figure 3. Map of Bhavani River India

Data Collection

Data was collected from Indian, Iraq and Malaysian rivers. The detail of each river is described in the following section.

Journal of Ecohumanism
2024
Volume: 3, No: 8, pp. 516 – 552
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v3i8.4750

Malaysian River

The data for Malaysian rivers are gathered from the Klang and Langat Rivers as they are many key variables belongs to the Langat and Klang River basins. In this study, 656 data samples with 6 characteristics were gathered from water quality monitoring sites between January 2005 and August 2016. Statistical Analysis on Malaysian River is shown in Table 1 and Figure 4-7.

Table 1. Dataset Statistics

```
Data columns (total 9 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   DO      655 non-null    float64
 1   BOD     655 non-null    int64
 2   COD     655 non-null    int64
 3   SS      655 non-null    int64
 4   pH      655 non-null    float64
 5   NH3-N   655 non-null    float64
 6   WQC1    655 non-null    int64
 7   WQI     655 non-null    float64
 8   WQC     655 non-null    object
dtypes: float64(4), int64(4), object(1)
```

Figure 4-8 illustrates the histogram, Boxplot, Violin plot and Correlation matrix of Malaysian dataset.

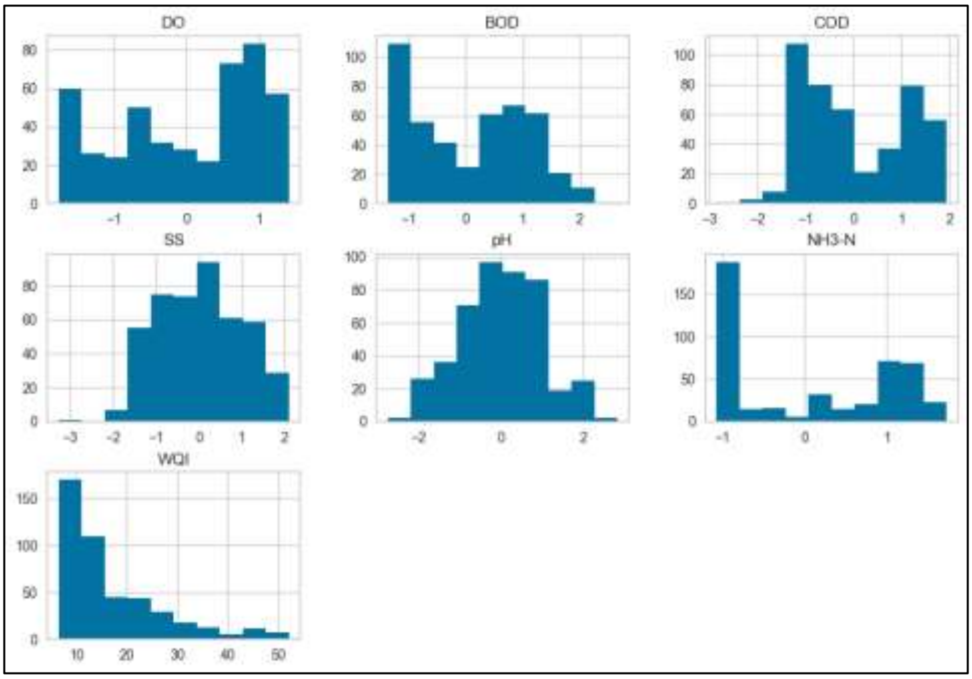


Figure 4. Histogram of Malaysian River

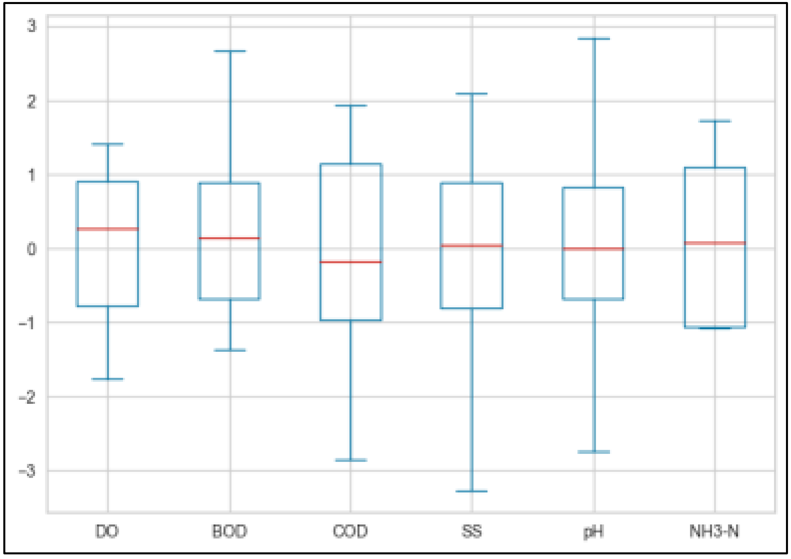


Figure 5. Box Plot of Malaysian River

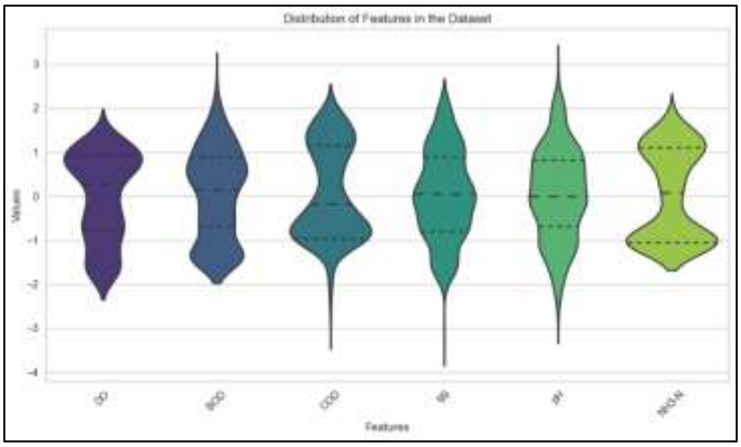


Figure 6. Violin Plot of Malaysian River

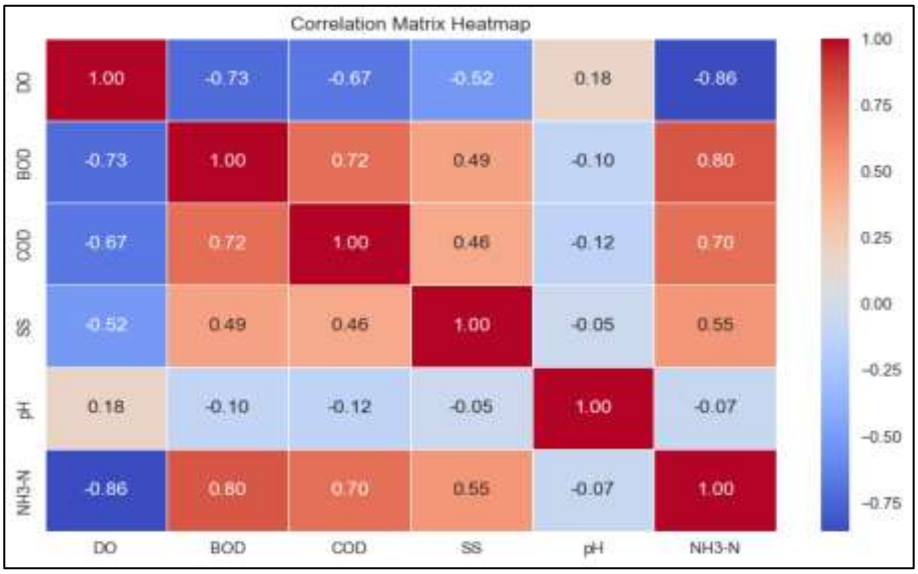


Figure 7. Correlation Matrix of Malaysian River Parameters

Iraqian Rivers

The Tigris and Euphrates River basin's principal stations are T27, T28, E8, E11, E16, and E19. The conditions of six monitoring stations along the Tigris and Euphrates Rivers are where the factors that make up the water quality index, including temperature, pH, dissolved oxygen, turbidity, chloride, and others, are gathered. Between January 2010 and December 2019, 481 data samples with 17 characteristics were collected from six stations of Tigris and Euphrates basin. Statistical Analysis on Iraqian River is shown in Table 2 and Figure 8-11.

Table 2. Dataset Statistics

```
 Data columns (total 17 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   Q m3/s  376 non-null    int64
1    PH      376 non-null    float64
2    Temp    376 non-null    float64
3    DO2     376 non-null    float64
4    PO4     376 non-null    float64
5    NO3     376 non-null    float64
6    Ca      376 non-null    float64
7    Mg      376 non-null    float64
8    TH      376 non-null    float64
9    K       376 non-null    float64
10   Na      376 non-null    float64
11   SO4     376 non-null    float64
12   CL      376 non-null    float64
13   TDS     376 non-null    float64
14   EC      376 non-null    float64
15   Alk     376 non-null    float64
16   WQI     376 non-null    float64
```

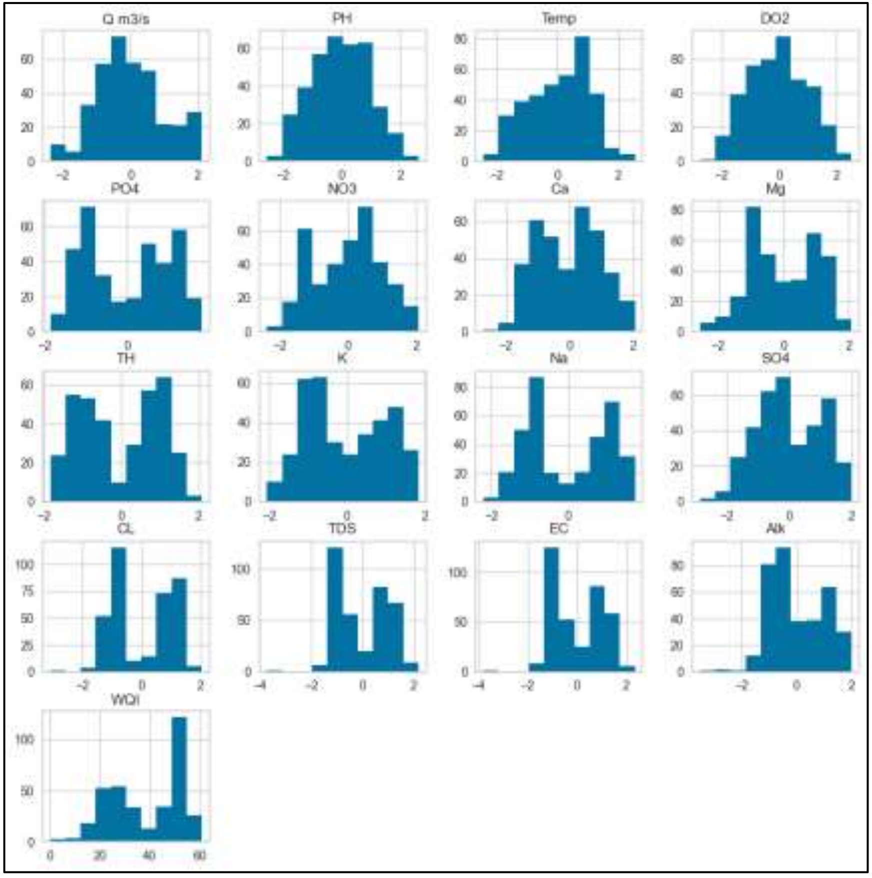


Figure 8. Histogram of Iraqian River

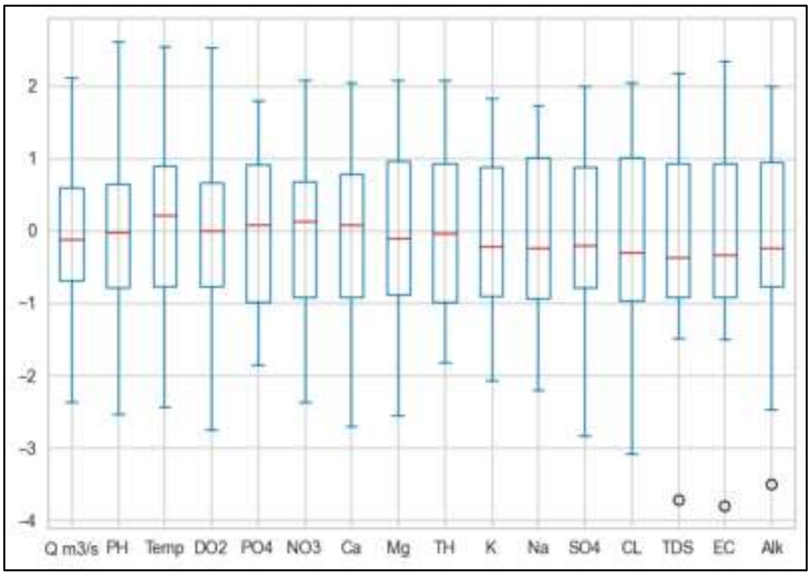


Figure 9. Box Plot of Iraqian River

ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v3i8.4750

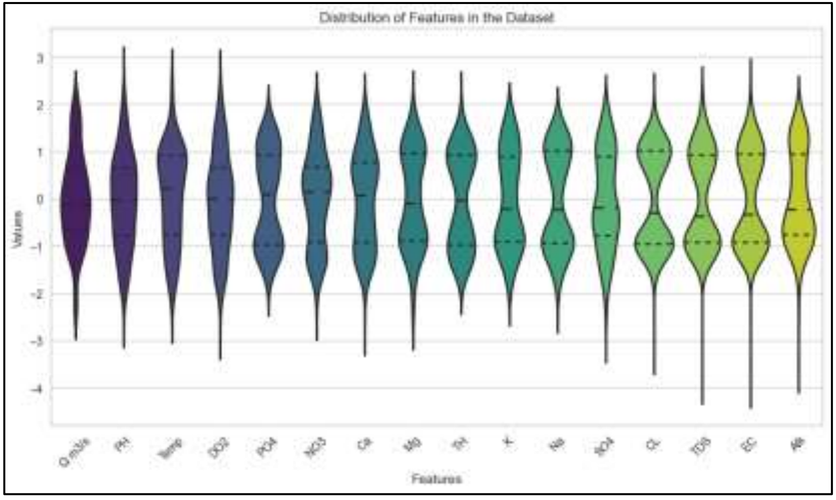


Figure 10. Violin Plot of Iraqian River

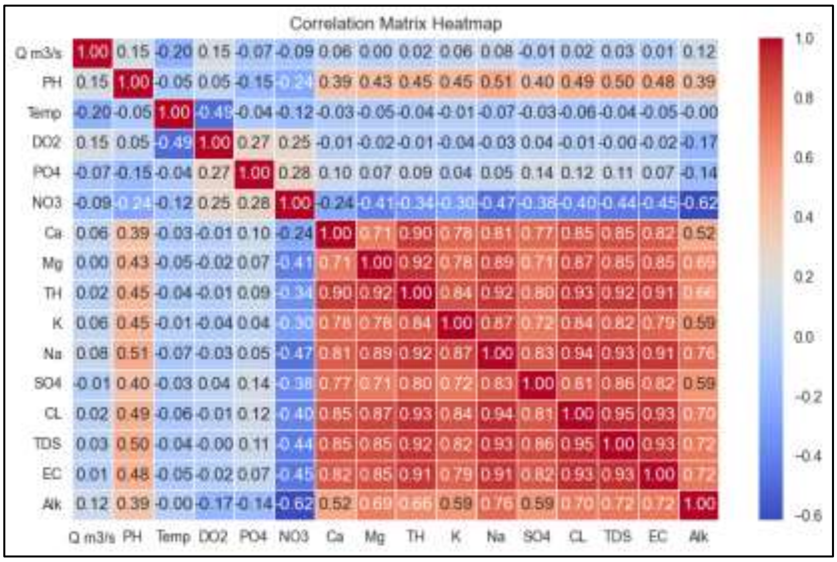


Figure 11. Correlation Matrix of Iraqian River Parameters

Bhavani River India

The Bhavani Basin has a number of significant stations, including Kottathara, Thavalam, Chalayur, Karathur, Cheerakuzhi, Elachivazhi, and Badrakaliamman kovil. Data on temperature, pH, dissolved oxygen, turbidity, chloride, and other variables that are required to calculate the water quality index are collected at the seven monitoring sites along the Bhavani River. Between January 1, 2016, and December 22, 2019, the water quality monitoring stations provided the 7649 data samples and 31 characteristics that were used in this study effort. Statistical Analysis on Bhavani River is shown in Table 3 and Figure 12-15.

Table 3. Bhavani River India Dataset Statistics

```
                Data columns (total 43 columns):
 #    Column                    Non-Null Count   Dtype
---   ------                    --------------   -----
 0    Temp                      7648 non-null    float64
 1    pH                        7648 non-null    float64
 2    Conductivity              7648 non-null    float64
 3    Turbidity                 7648 non-null    float64
 4    PhenolphthAlkalinity      7648 non-null    float64
```

```
5   Total Alkalinity       7648 non-null   float64
6   Cholride               7648 non-null   float64
7   COD                    7648 non-null   float64
8   TKN                    7648 non-null   float64
9   Ammonia                7648 non-null   float64
10  Hardness               7648 non-null   float64
11  Ca.Hardness            7648 non-null   float64
12  Mg.Hardness            7648 non-null   float64
13  Sulphate               7648 non-null   float64
14  Sodium                 7648 non-null   float64
15  TSS                    7648 non-null   int64
16  TDS                    7648 non-null   float64
17  FDS                    7648 non-null   float64
18  Phosphate              7648 non-null   float64
19  Boron                  7648 non-null   float64
20  Pottassium             7648 non-null   float64
21  BOD                    7648 non-null   float64
22  Fluoride               7648 non-null   float64
23  Nitrate-N              7648 non-null   float64
24  TC                     7648 non-null   float64
25  FC                     7648 non-null   float64
26  Dew                    7646 non-null   float64
27  Humidity               7646 non-null   float64
28  Sealevelpressure       7646 non-null   float64
29  Precipitation          7646 non-null   float64
30  Precipcover            7646 non-null   float64
31  Windspeed              7646 non-null   float64
32  Winddir                7646 non-null   float64
33  Cloudcover             7646 non-null   float64
34  Visibility             7646 non-null   float64
35  Station                7648 non-null   int64
36  Latitude               7648 non-null   object
37  Longitude              7648 non-null   object
38  Year                   7648 non-null   int64
39  Date                   7648 non-null   object
40  DO                     7648 non-null   float64
41  WQI                    7647 non-null   float64
42  WQC                    7647 non-null   float64
```

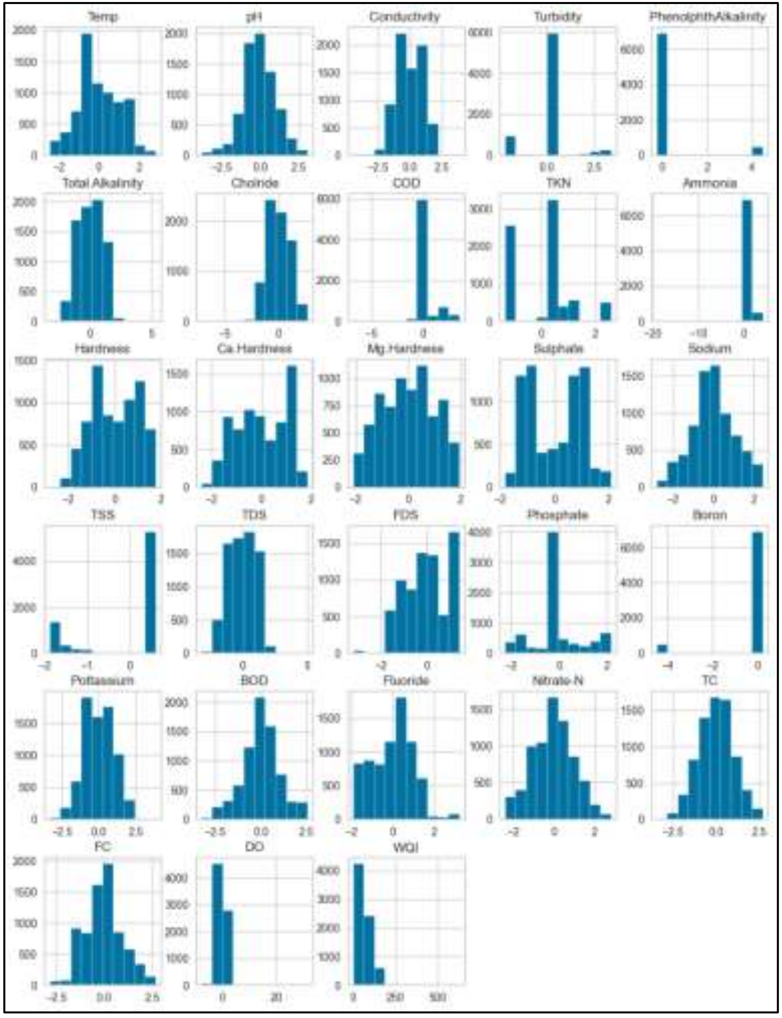


Figure 12. Histogram of Bhavani River India

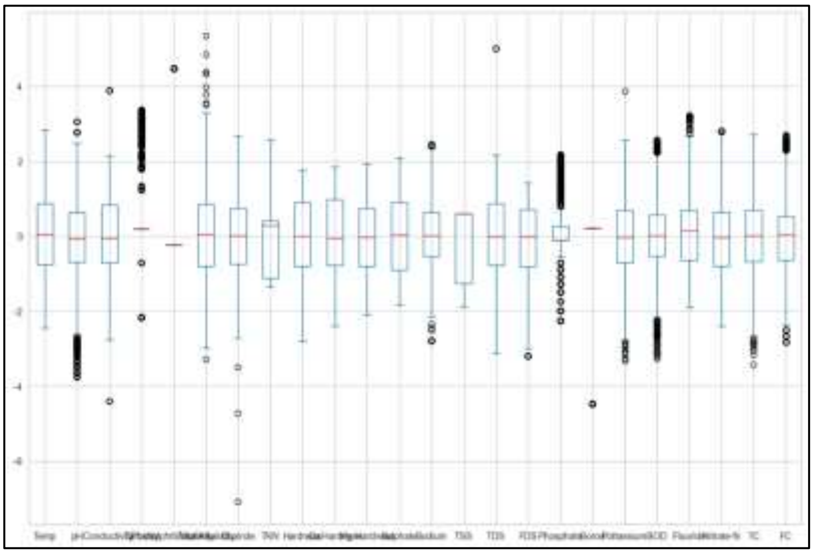


Figure 13. Box Plot of Bhavani River

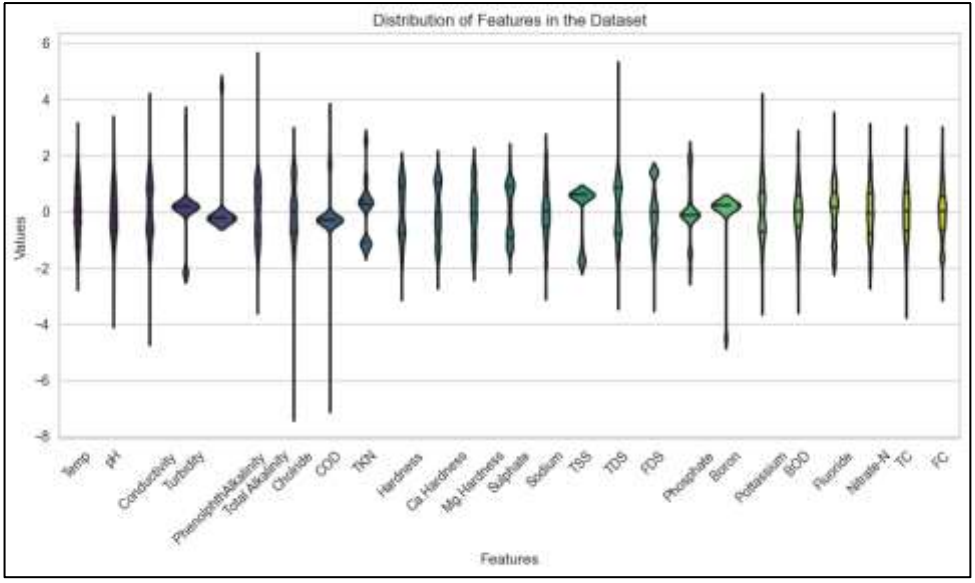


Figure 14. Violin Plot of Bhavani River

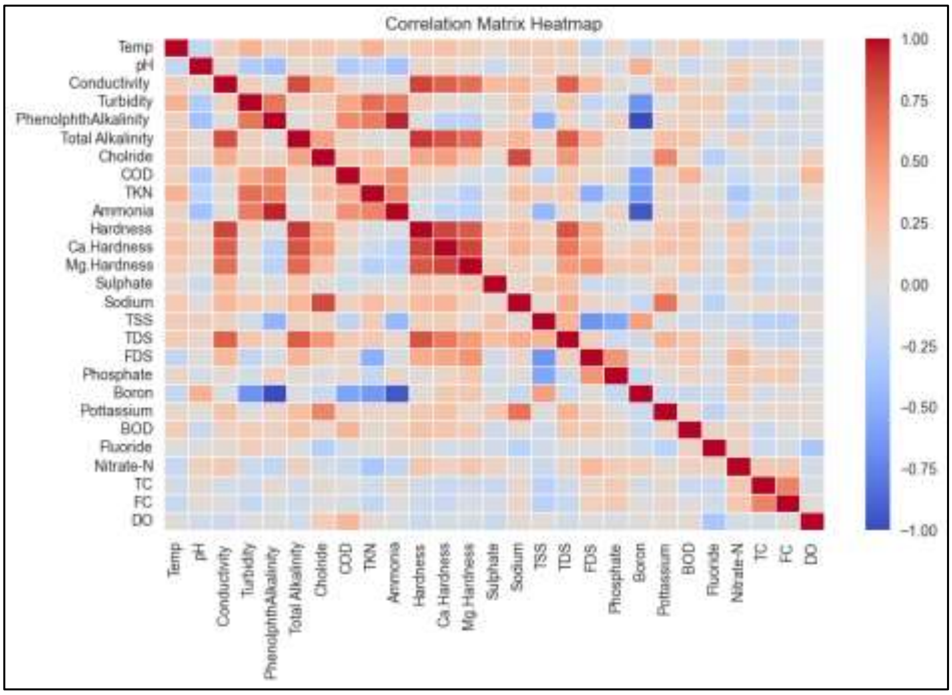


Figure 15. Correlation Matrix of Bhavani River Attributes

WQI Calculation

The water quality index is calculated using the following formula (1) the main purpose is to use arithmetic weights of water and quality.

$$WQI = \frac{\sum wiqi}{\sum wi} \quad (1)$$

I represent the number of variables taken into account, and qi is a relative water quality value specific to each parameter. A factor called Wi measures a parameter's relative weight to calculate the WQI. The value of qi may be obtained by using formula 2 from below.

$$qi = 100 * \frac{vi - vo}{si - vo} \tag{2}$$

With the exception of DO = 14.6 mg/l whereas the pH is set to 7, all other parameters have ideal values of zero. For each parameter under consideration, vi denotes the experimentally determined value, while vo denotes the ideal value. Si stands for the legally recognised standard for the water category that the water sample under examination belonged to. In order to calculate the wi factor, use formula 3.

$$wi = \frac{K}{si} \tag{3}$$

K is a constant that may be obtained by using the formula 4,

$$K = \frac{1}{\Sigma(\frac{1}{si})} \tag{4}$$

The weight of each participating parameter in terms of unit as well as its permitted restrictions for calculating WQI. Parameters are Temp(^{o}C), pH, Conductivity, Turbidity, Phenolpth Alkalinity, Total Alkalinity, Chloride, COD, TKN, Ammonia, Hardness, Ca. Hardness, Mg. Hardness, Sulphate, Sodium, TSS, TDS, FDS, Phosphate, Boron, Potassium, BOD, Fluoride, DO, Nitrate-N, TC, and FC. Permissible Limits are 28,8.5,150,5,20,200,250,10,100,50,100,75,30,200,200,300,1000, 200,0.3,1,2.5,3,1.5,7.5,0.503,100, and 60. Whereas the weight of each parameter is 0.035714286,0.117647059,0.006666667,0.2,0.05,0.005,0.004,0.1,0.01,0.02,0.01,0.013333333,0.033333333,0.005,0.005,0.003333333,0.001,0.005,3.333333333,1,0.4,0.333333333,0.666666667, 0.133333333, 1.988071571,0.01, and 0.016666667. The parameters were also used by Jitha in[4,9-11].

The water ecological state may be calculated based on the result of the weighted arithmetic WQI approach, as depicted in Table 4.

Table 4. Water Quality Index Standards Based on Weight Arithmetic [14]

Water Quality	WQI Index	WQI Class	WQI Range
Good	1	A	0-30
Moderate	2	B	31-60
Poor	3	C	61-90
Very Poor	4	D	91-120
Unsuitable	5	E	>121

The weighted arithmetic water quality index criteria that have been utilised to generate the water quality index using the aforementioned formulas. The permitted values and unit weight of each parameter as indicated above must be utilised in order to calculate the water quality index.

Each sample's water quality index value was computed, and the resulting value was then placed to the appropriate instance. The present criteria are used to establish the water quality index class for each instance, which is then applied as a class label to the pertinent instance. Thirty-three features and 7649 labelled occurrences make up the Bhavani River WQ dataset. However, the Klang and Langat rivers' water quality dataset has 6 attributes and 656 labelled occurrences. 481 labelled occurrences and 17 attributes round up the water quality dataset for the Tigiris and Euphrates rivers.

Data Preprocessing

Data preprocessing improves data quality and efficiency. The irregularity and noise in raw data degrade its quality. The several stages used in this study endeavour offered a thorough comprehension of the data, revealing that the dataset includes duplicates, outliers, and imbalance concerns. To eliminate outliers and standardise all datasets, the z-score was used, and the outlier threshold was set at 0.050000. The yeo-Johnson approach was used to turn data into something actual and succinct. Finally, the SMOTE approach was used to balance the data and characteristics in the datasets.

Methodology

The suggested WQI classification and regression models are made up of many components, including data collection, data preprocessing, WQI classification and prediction, , and performance assessment. The water quality index was classified using machine learning techniques. The accuracy, precision, recall, and F1 score of the WQI classifier are used to assess it[4,9-11]. Whereas prediction model used to identify the predict the water quality based on the input data [14]. The prediction models were also tested and evaluated by MAE, MSE, RMSE, MAPE, R^2, scatter index and Bias. Method adopted based on the baseline study [4], in which classification and prediction was performed on individual dataset. Also proposed models are compared with [26-27].

Figure 16 depicts and describes the architecture of the proposed WQI prediction model.

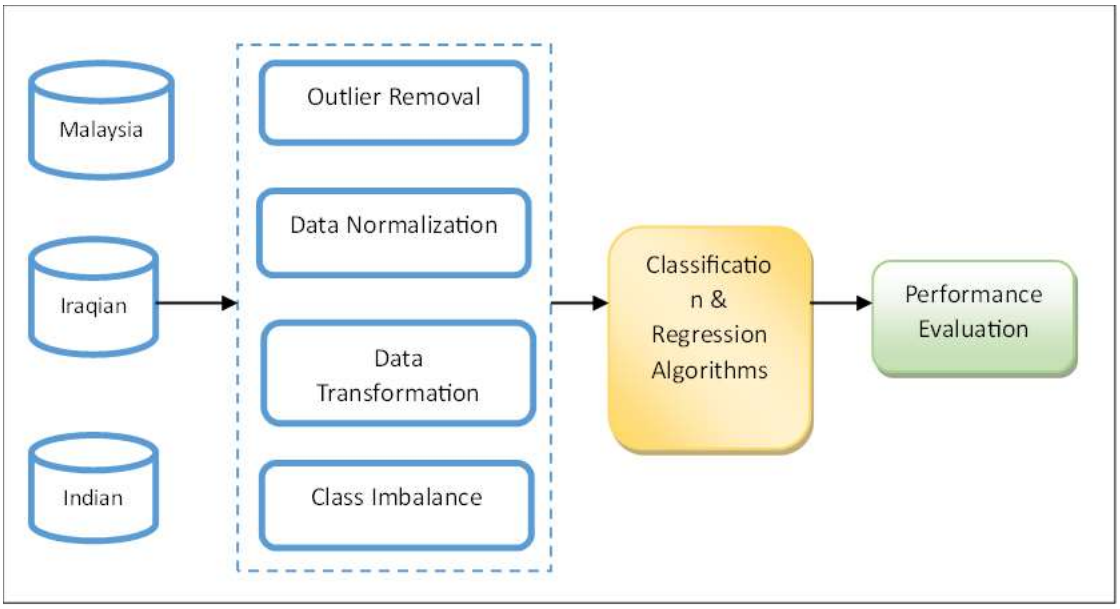


Figure 16. Proposed Framework for WQI

Building WQI Classification and Prediction Model

Learning patterns in the river water quality dataset is used to create the water quality index classification model. The WQI is used to label the classes to form group of five WQ standards based on WQ characteristics, as shown in Table 2.

Classification models are built using machine learning methods including Extreme Gradient Boost (XGBoost), SVM, Naive Bayes, and Ada Boost.

In recent times, XGBoost has taken over Kaggle competitions for structured or tabular data and utilized ML. The execution of a gradient boosted decision tree called XGBoost was created with speed and effectiveness in consideration. Making the most of the resources available to train the model was one of the design objectives. Block, ongoing training, and sparse awareness all constitute crucial aspects of algorithm execution.

ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v3i8.4750

AdaBoost, also known as Adaptive Boosting, is a supervised machine learning algorithm that is employed as an ensemble. Decision trees with one level or one split are very popular estimators utilized with AdaBoost. Since the independent variables don't need to be scaled, less preprocessing is needed. The preprocessing needed is the same as for decision trees since every iteration of the AdaBoost algorithm uses decision stumps as distinct models. Additionally, AdaBoost is less prone to overfitting.

SVM is a supervised method for predicting and classifying data. Since each data point in n-dimensional space is shown separately, it is simple to tell the difference between the two groups. In the fields of technology, pattern recognition, and learning categorization, SVMs are becoming more and more well-liked. A linear or non-linear separation surface might be used to classify the input region. A linear collection of kernels coupled to the support vector makes up the separation function in support vector classification.

The Naive Bayes algorithm is a classification technique based on Bayes' theorem that states that once the goal value is determined, the remaining qualities become independent variables. To forecast and categorise datasets, the Bayesian approach utilises probability and statistical expertise. The Bayesian method employing pre and posterior probability may be used to avoid the biased and the overfitting assumptions of applied sampling information.

For prediction of the water quality on India, Malaysian and Iraqian datasets different regressor models are used such as Decision Tree Regressor, Random Forest Regressor, M5 Model Tree Model Tree, Extreme Learning Machine (ELM Regressor). The performance measured with RMSE, MSE, MAE, MAPE, Scatter Index and Bias.

Performance Evaluation

To select the optimal approach, the performance of the built models for categorization of the WQI is assessed. The accuracy is used to determine the most efficient and best classification model The following statistical parameters were employed:

$$\text{Accuaracy} = \frac{\text{TP+TN}}{\text{TP+FP+TN+FN}} \tag{5}$$

$$\text{Precision} \ \frac{\text{TP}}{\text{TP+FP}} \tag{6}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{7}$$

Whereas F1-score defines the mean values of Precision and Recall presented in equation 8.

$$F1 - Score = \frac{2*precision*recall}{precision+recall} * 100 \tag{8}$$

where, in turn, TP, TN, FP, and FN stand for true positive, true negative, false positive, and accordingly, false positive and false negative. Using the aforementioned equations, machine learning methods for creating water quality indices in categorization are tested in order to determine how effective they are when used with data from river water.

On the other hand, for model prediction different techniques for regression are used such as MAE, MSE< RMSE, MAPE, R2, SI and BIAS (equation 9-14) are used.

Mean Absolute Error

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}| \tag{9}$$

Mean Squared Error

$$MSE = \frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y})^2 \tag{10}$$

Root Mean Squared Error

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y})^2} \tag{11}$$

R-Squared

$$R^2 = 1 - \frac{\sum (y_i - \hat{y})^2}{\sum (y_i - \bar{y})^2} \tag{12}$$

Scatter Index

$$\mathrm{SI} = \frac{\mathrm{RMSE}}{\bar{X}} \tag{13}$$

Bias

$$\mathrm{Bias}(\hat{Y}) = E(\hat{Y}) - Y \tag{14}$$

Experimental Results

The Bhavani River, Klang-Langat, and Tigires-Euphrates water quality datasets were used in experiments to develop a precise WQI classification model. Different instance datasets with multiple characteristics were split into training and testing sets, with 80% of the instances used for training and 20% used for testing.

The ML algorithms such as Decision Tree Regressor, Random Forest Regressor, M5 Model Tree, Extreme Learning Machine (ELM Regressor), XGBoost, SVM, Naive Bayes, and Adaboost are combined with independent and dependent parameters to create a water quality prediction and index classification model. Testing is performed on 20% data to assess the models' performance using various performance evaluation metrics. For each dataset, the Stratified-K Fold method was used with 10 folds to assess the model's performance.

WQI Classification and Prediction for Malaysian Rivers

In order to find the optimal model accuracy and prediction, various ML models have been implemented on two different datasets using multiple parameters.

Classification Models

In this part, classification models are developed using the rivers water quality data samples. Python modules as well as Extreme Gradient Boosting, support vector machine (SVM), Ada Boost and Naive Bayes are selected for WQI classification. The classification models' performance is tested to determine the efficiency of WQI classification using measures such as accuracy, precision, recall, and F1 score.

According to the experimental findings, the accuracy of the Extreme Gradient Boosting model is 0.9346, whereas the accuracy of Naive Bayes, Support Vector Machine, and Ada Boost is 0.8885, 0.7426, and 0.5394, respectively. As demonstrated in Table 5 and Figure 17-23, the accuracy of XGBoost is more optimal than other classifiers.

Table 5. Classification Accuracy on Malaysian Dataset

Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
Extreme Gradient Boosting	0.9346	0.9913	0.9346	0.9388	0.9340	0.9182	0.9195	0.1990
Naive Bayes	0.8885	0.9843	0.8885	0.9029	0.8875	0.8606	0.8645	0.0590
SVM	0.7426	0.0000	0.7426	0.7618	0.7302	0.6781	0.6890	0.0740
Ada Boost	0.5394	0.8696	0.5394	0.4004	0.4320	0.4242	0.4632	0.1100

Figure 17 and 18 illustrates the performance of XGB model on Malaysian dataset, it clearly shown ROC curve and confusion matrix that this model has high accuracy as compared to other models presented in the Table 5. Apart from accuracy of this model, other performance measuring parameters such as Recall, Precision, F1, Kappa, MCC are also much better than other models.

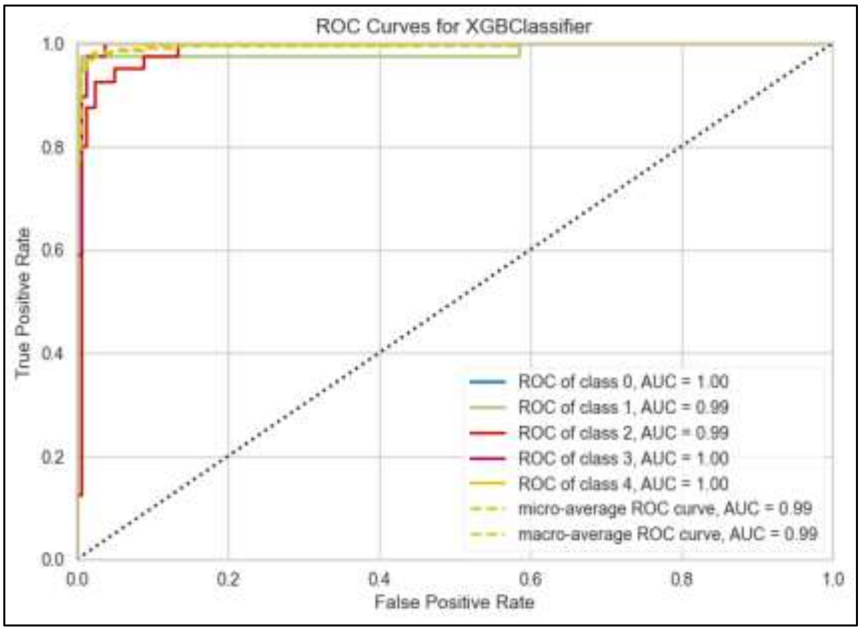
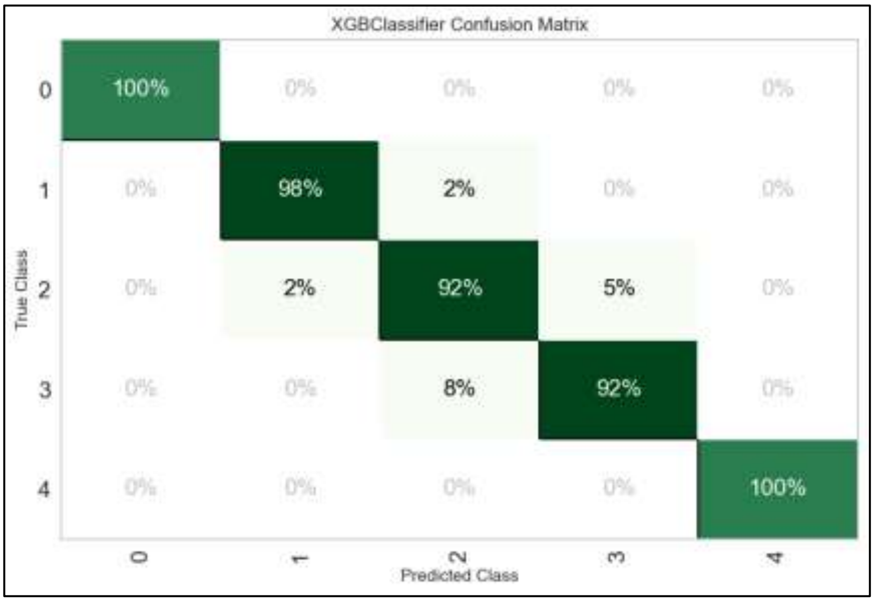


Figure 17. ROC Curves of XGB Classifier on Malaysian Dataset



True Class	0	1	2	3	4
0	100%	0%	0%	0%	0%
1	0%	98%	2%	0%	0%
2	0%	2%	92%	5%	0%
3	0%	0%	8%	92%	0%
4	0%	0%	0%	0%	100%

Figure 18. Confusion Matrix of XGB Classifier on Malaysian Dataset

Figure 19 and 20 illustrates the performance of NB model on Malaysian dataset, it clearly shown ROC curve and confusion matrix that this model has second highest as compared to adaboost and SVM models presented in the Table 5. Same as XGB, this model also have better results than other models.

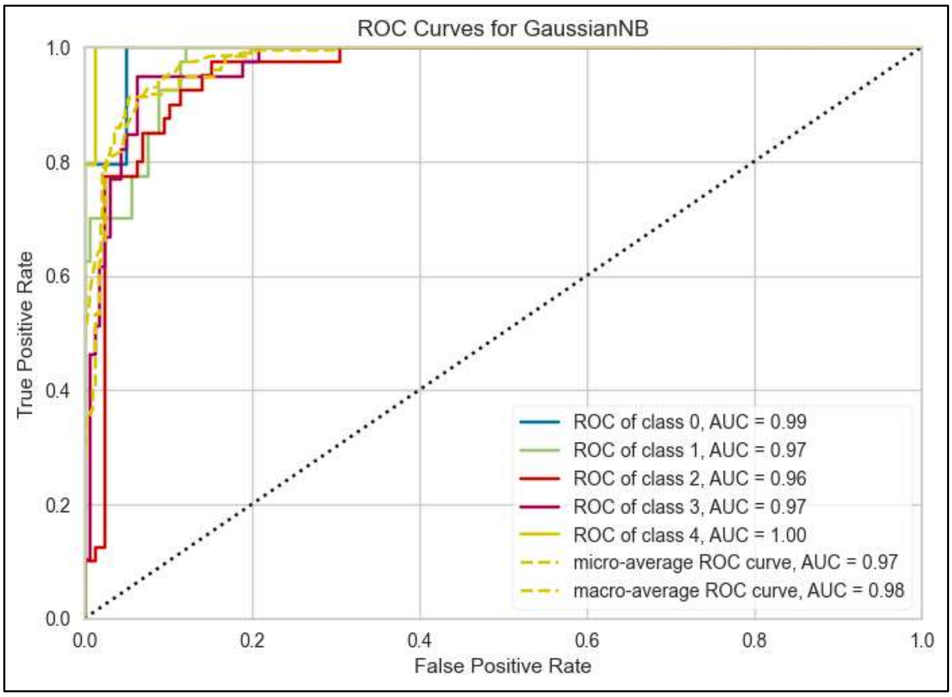


Figure 19. ROC Curves of Naïve Bayes Classifier on Malaysian Dataset

ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v3i8.4750

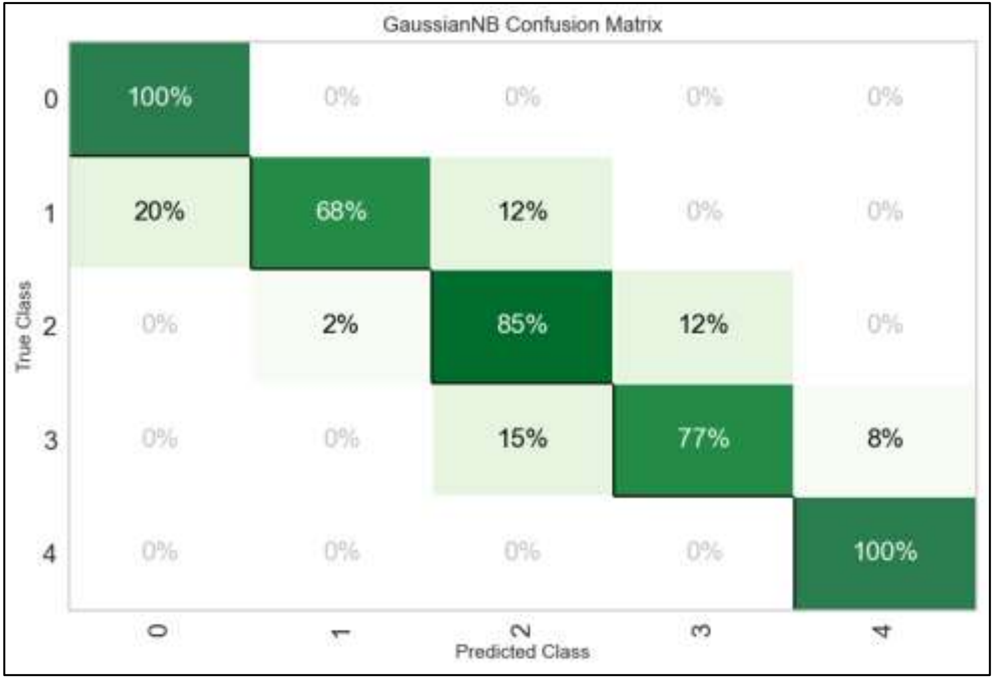


Figure 20. Confusion Matrix of Naïve Bayes Classifier on Malaysian Dataset

Figure 21 illustrates the performance of SVM model on Malaysian dataset, it clearly shows that this model has better confusion matrix than Adaboost but lowest than XGB and NB.

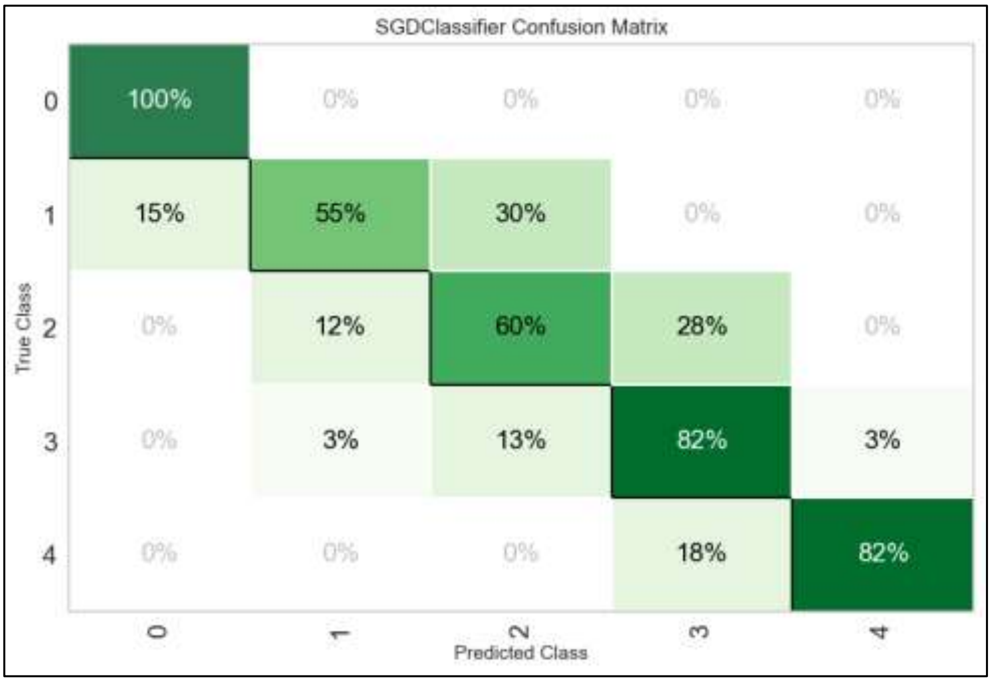


Figure 21. Confusion Matrix of SVM Classifier on Malaysian Dataset

Lastly, Figure 22 and 23 illustrates the performance of AdaBoost classifier model on Malaysian dataset, it clearly shown ROC curve and confusion matrix that this model has lowest results as compared to other three models presented in the Table 5.

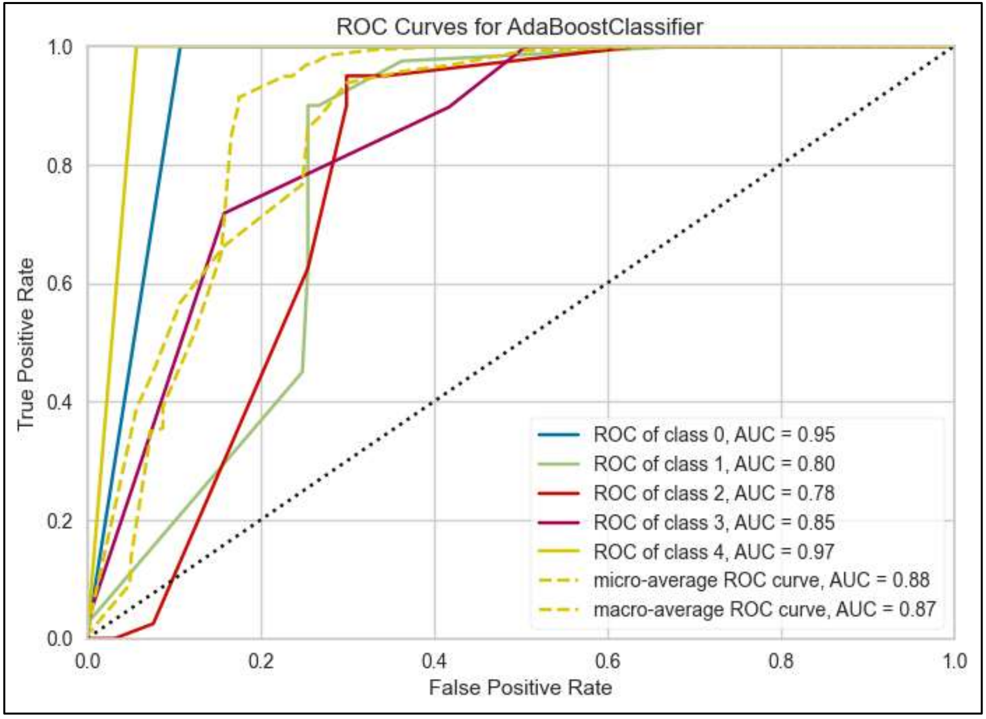


Figure 22: ROC Curves of Adaboost Classifier on Malaysian Dataset

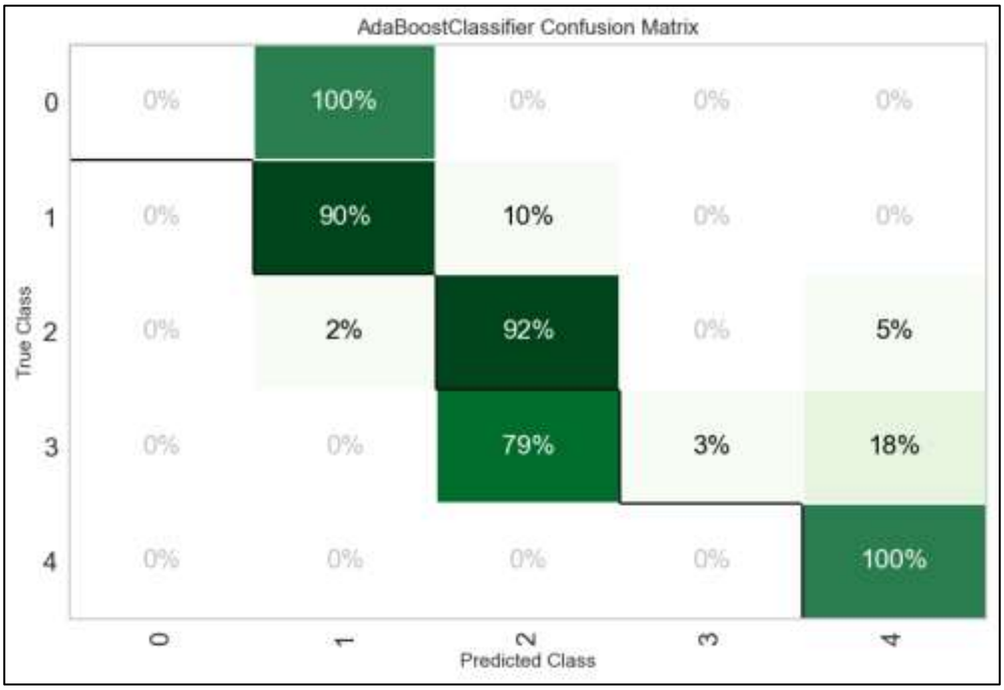


Figure 23. Confusion Matrix of AdaBoost Classifier on Malaysian Dataset

Prediction Models for Malaysia

In this part, Prediction models are developed using the rivers water quality data samples. Python modules as well as Decision Tree Regressor, Random Forest Regressor, M5 Model Tree, Extreme Learning Machine (ELM Regressor) are selected for WQI Prediction. The models' performance is tested to determine the efficiency of WQ Prediction using measures such as MAE, MSE, RMSE, MAPE, R^2, SI and Bias as shown in Table 6 and Figures 24-27-26

Table 6. Performance of Prediction Models on Malaysian Rivers

Model	MAE	MSE	RMSE	MAPE	R^2	S I	BIAS
Decision Tree Regressor	0.737	2.663	1.632	0.031	0.979	0.090	0.061
Random Forest Regressor	0.639	1.611	1.269	0.261	0.987	0.070	-0.20
M5 Model Tree	4.627	4.127	6.424	2.797	0.999	3.54	9.200
ELM Regressor	3.943	2.803	5.294	2.291	0.999	2.92	4.086

According to the experimental findings, the MAE of the Decision Tree regressor is 0.737, whereas the MAE for random Forest Regressor is 0.639, MAE for M5 Model Tree is 4.627 and MAE for ELM Regressor is 3.943. As demonstrated in Table 6 and Figure 23-26.

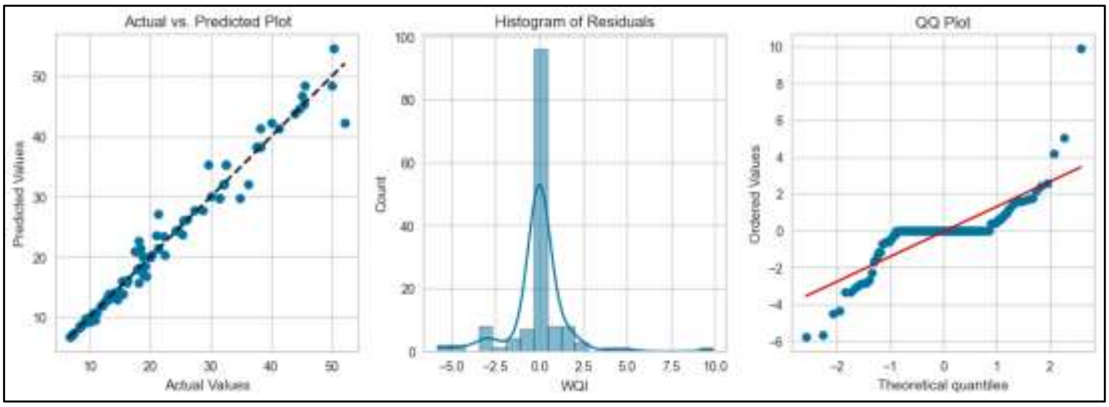


Figure 24. Plots of Decision Tree Regressor on Malaysian Dataset

Figure 24 illustrates the performance of Decision tree regressor on Malaysian dataset, the performance of this model in terms of MAPE, R2 and Bias is better than Random Forest Regressor.

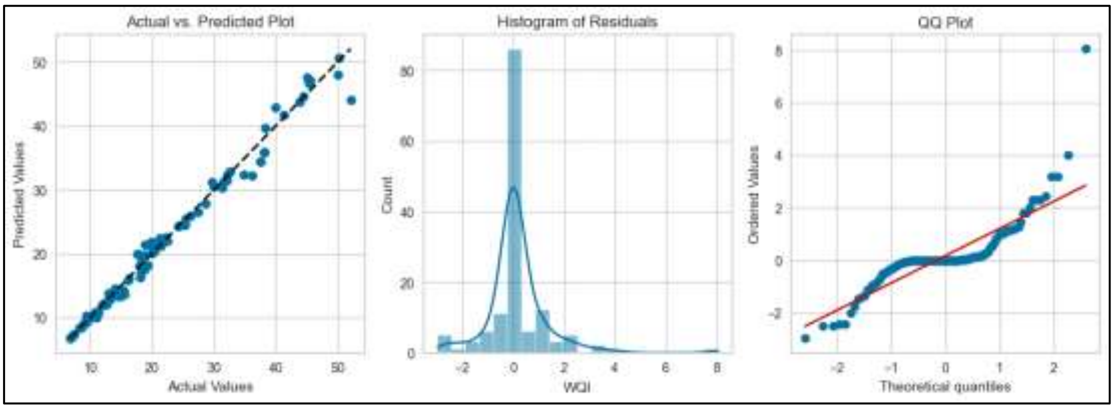


Figure 25. Plots of Random Forest Regressor on Malaysian Dataset

Whereas the performance of Random Forest regressor on Malaysian is better in MAE, MSE, RMSE, and SI as compared to Decision Tree regressor as shown in Figure 25.

On the other hand, the Extreme Learning Regressor is much better than M5 Model Tree regressor as shown in Figure 26 and 27.

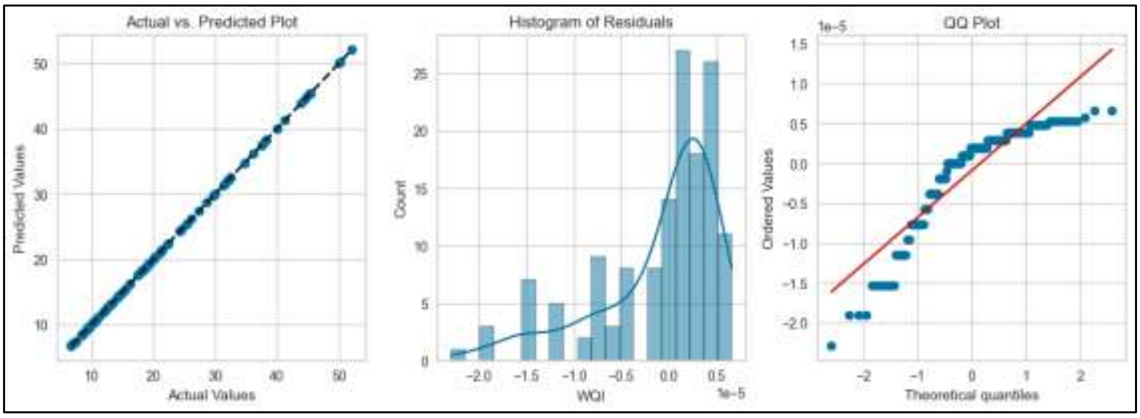


Figure 26. Plots of M5 Model Tree Regressor on Malaysian Dataset

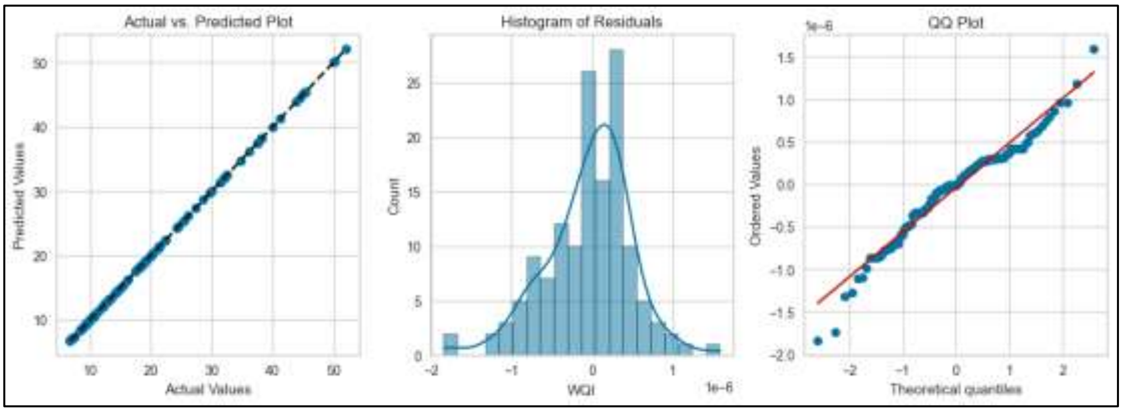


Figure 27. Plots of Extreme Learning Regressor on Malaysian Dataset

Classification and Prediction of WQI for Iraq

In order to find the optimal model accuracy and Prediction, various ML models have been implemented on dataset using multiple parameters.

Classification Models

According to the experimental findings, the accuracy of the Extreme Gradient Boosting model is 0.9224, whereas the accuracy of classification models based on naive bayes, support vector machine, and Ada Boost is 0.9110, 0.8070, and 0.6075, respectively. As indicated in Table 7 and Figures 28-34, the accuracy of XGBoost is better when compared to other classifiers, however Adaboosts' performance is lowest among all classifiers.

Table 7. Classification Accuracy on Iraq Dataset

Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
Extreme Gradient Boosting	0.9224	0.9921	0.9224	0.9299	0.9210	0.8821	0.8870	1.1660

Naive Bayes	0.9110	0.0000	0.9110	0.9199	0.9092	0.8647	0.8704	0.1530
SVM	0.8070	0.9269	0.8070	0.8311	0.7917	0.7102	0.7306	0.4900
Ada Boost	0.6075	0.7785	0.6075	0.6021	0.5583	0.4097	0.4491	0.2200

Figure 28 and 29 illustrates the performance of XGB model on Iraq dataset, it clearly shown ROC curve and confusion matrix of this model has optimal results as compared to other implemented classifiers as presented Table 7. Apart from accuracy of this model, other performance measuring parameters such as Recall, Precision, F1, Kappa, MCC are also much better than other models.

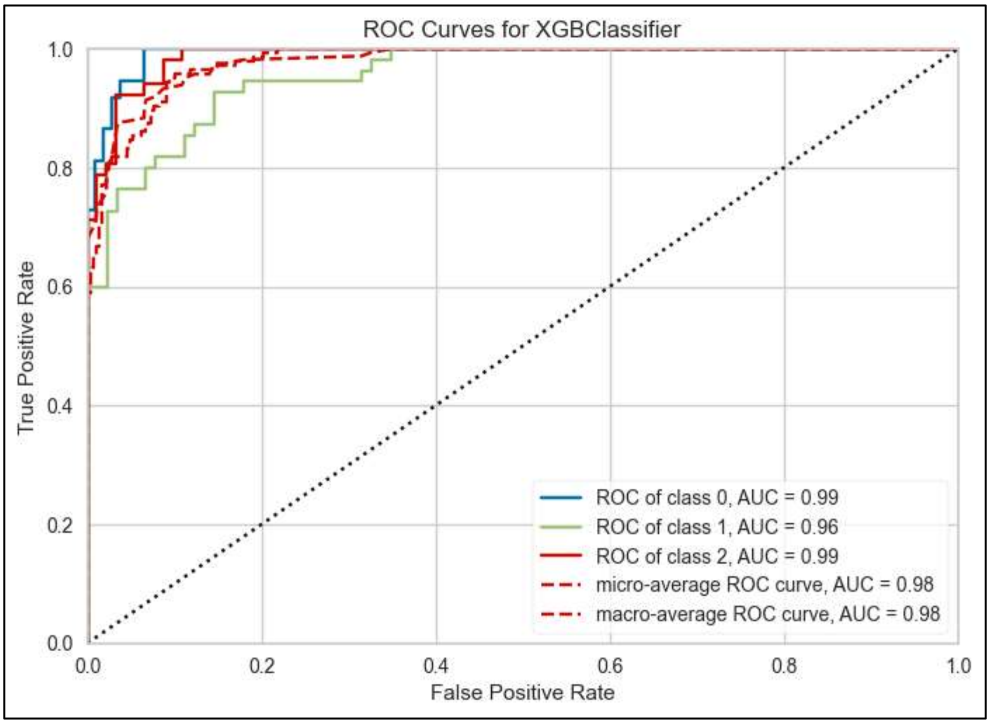


Figure 28. ROC Curves of XGB Classifier on Iraq Dataset

ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v3i8.4750

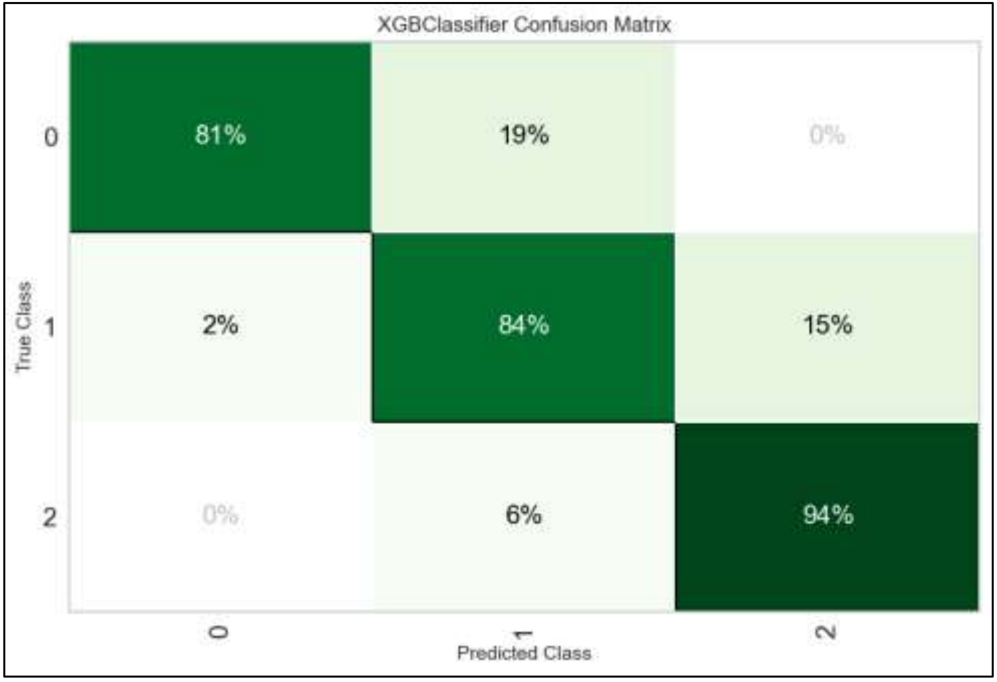


Figure 29. Confusion Matrix of XGB Classifier on Iraq Dataset

Figure 30 and 31 illustrates the performance of NB model on Iraq dataset, it clearly shown ROC curve and confusion matrix that this model has second highest as compared to adaboost and SVM models presented in the Table 7. Same as XGB, this model also it has better results than other models.

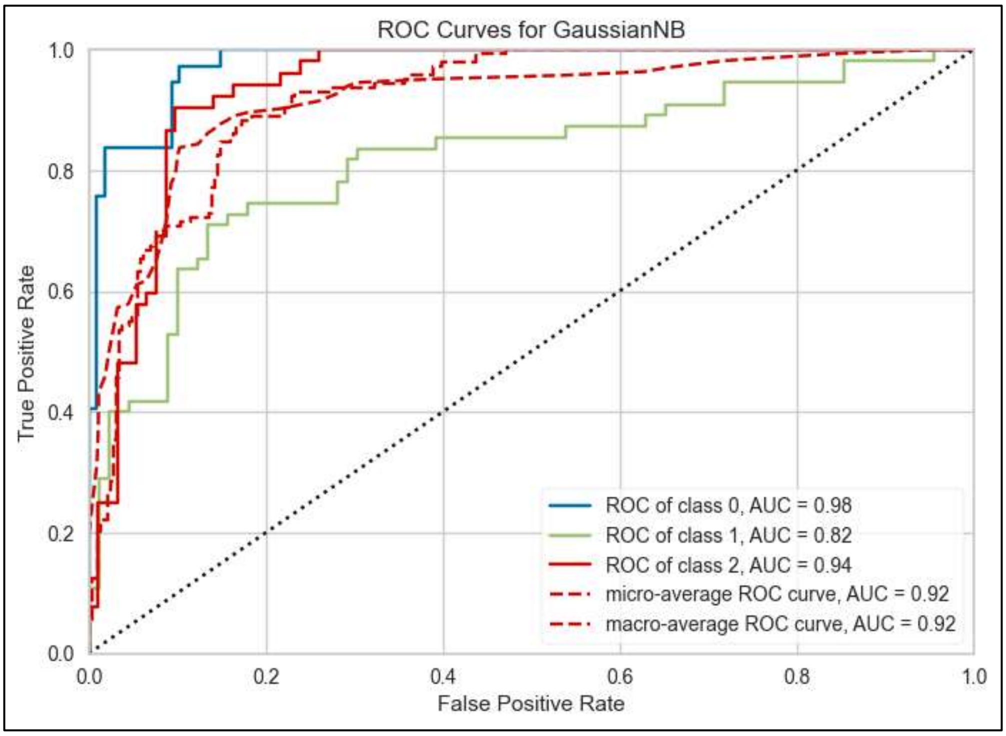


Figure 30. ROC Curves of Naïve Bayes on Iraq Dataset

ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v3i8.4750

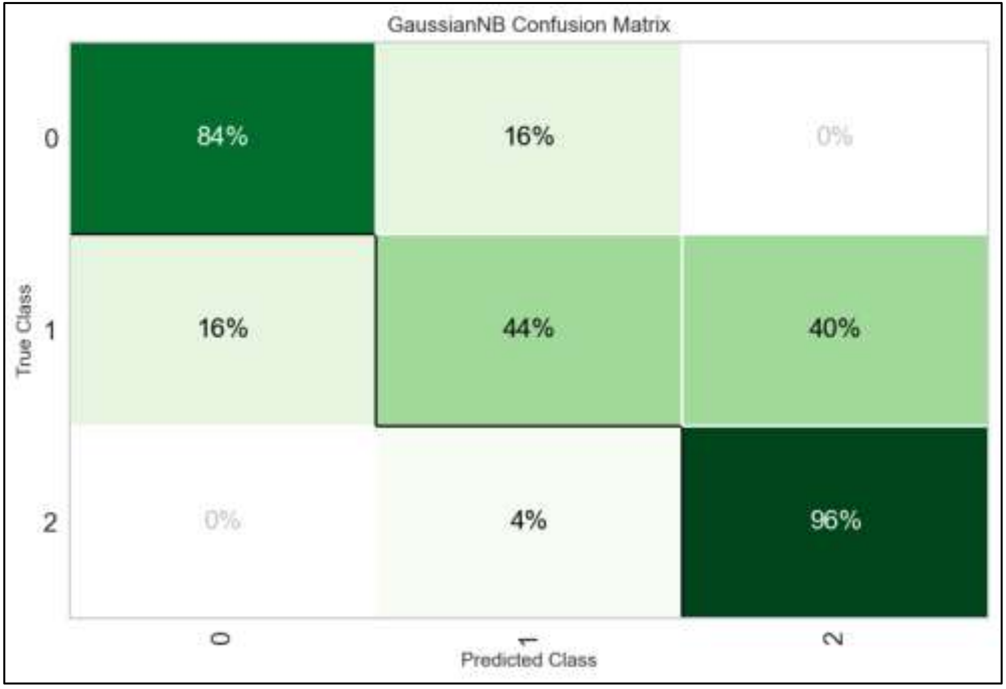


Figure 31. Confusion Matrix of of Naïve Bayes on Iraq Dataset

Figure 32 illustrates the performance of SVM model on Iraq dataset, it clearly shows that this model has better confusion matrix then adaboost but lowest than XGB and NB.

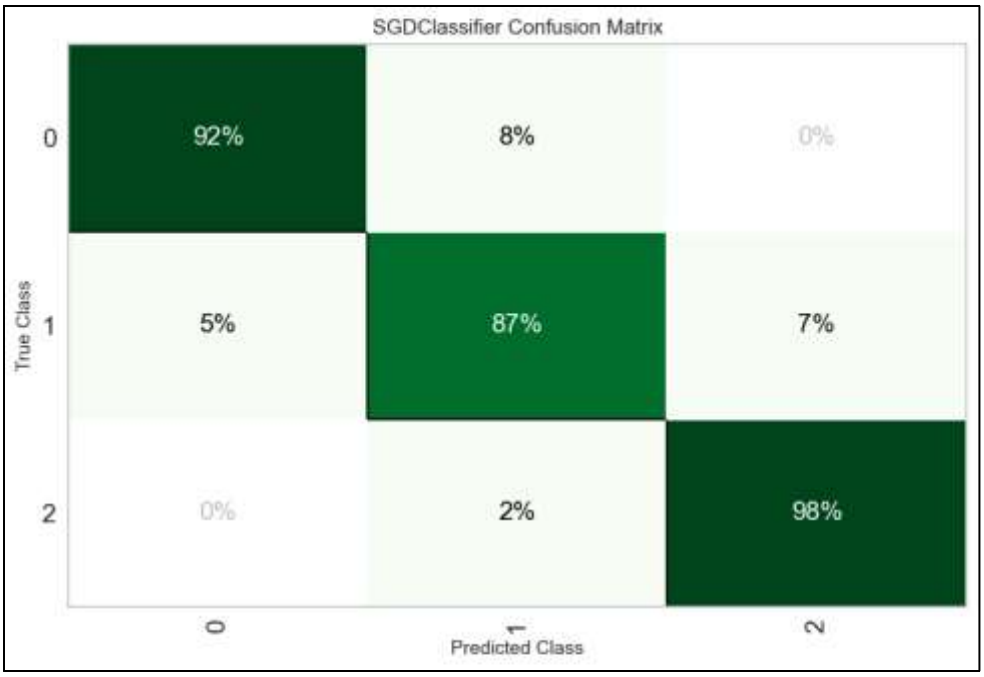


Figure 32. Confusion Matrix of SVM Classifier on Iraq Dataset

Lastly, Figure 33 and 34 illustrates the performance of AdaBoost classifier model on Iraq dataset, it clearly shown ROC curve and confusion matrix that this model has lowest results as compared to other three models presented in the Table 7.

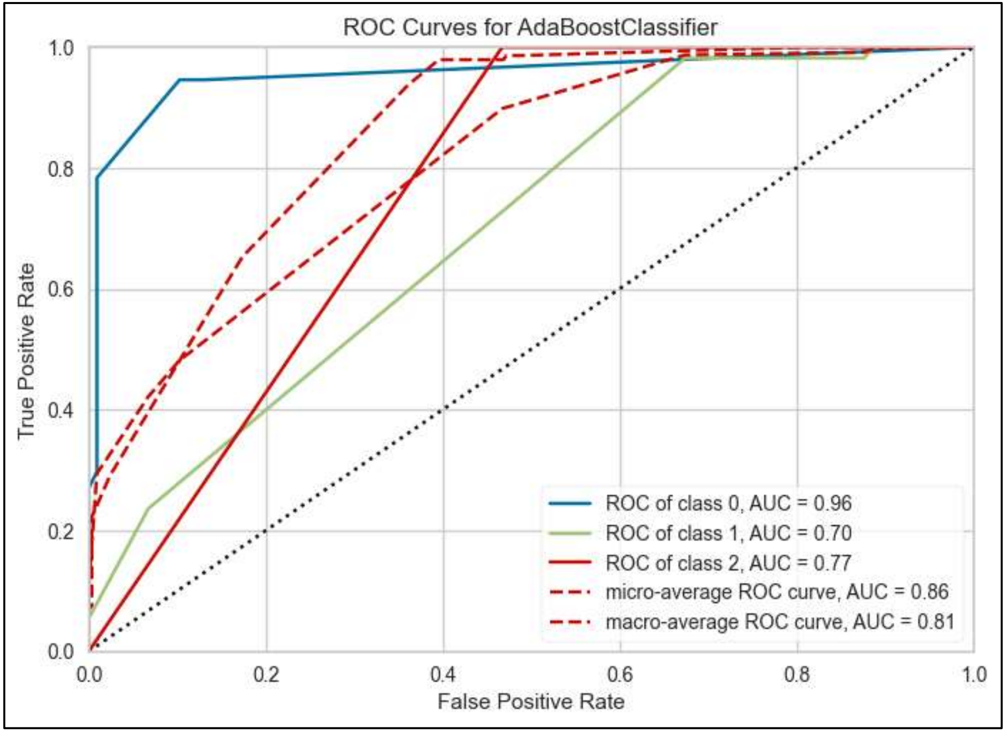


Figure 33. ROC Curves of AdaBoost Classifier on Iraq Dataset

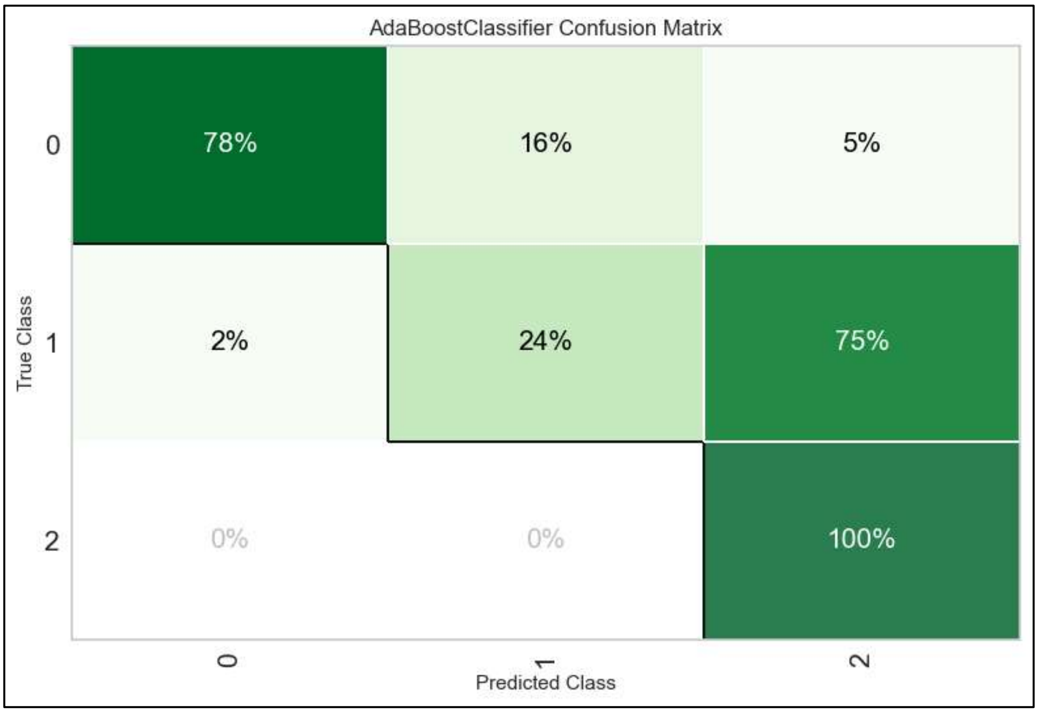


Figure 34. Confusion Matrix of AdaBoost Classifier on Iraq Dataset

Prediction Models for Iraq

According to the experimental findings, the MAE of the Decision tree regressor is 2.66, whereas the MAE for random Forest Regressor is 1.676, MAE for M5 Model Tree model is 0.002 and MAE for ELM Regressor is 7.319. As demonstrated in Table 8 and Figure 35-38.

Table 8. Performance of prediction Model on Iraq Datasets

Model	MAE	MSE	RMSE	MAPE	R²	S I	BIAS
Decision Tree Regressor	2.66	19.46	4.412	1.288	0.909	0.114	0.745
Random Forest Regressor	1.676	10.013	3.164	1.286	0.953	0.0817	0.414
M5 Model Tree	0.002	5.669	0.002	2.306	0.999	6.154	-3.12
ELM Regressor	7.319	108.484	10.415	1.950	0.495	0.269	-2.55

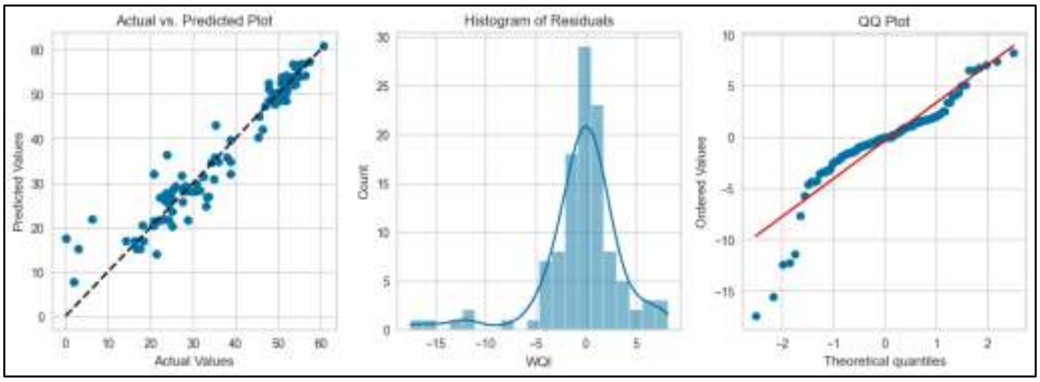


Figure 35. Plots of Decision Tree Regressor on Iraq Dataset

Figure 35 illustrates the performance of Decision Tree regressor on Malaysian dataset, the performance of this model in terms of R2 is better than Random Forest Regressor.

Whereas the performance of Random Forest regressor on Iraq is better in MAE, MSE, RMSE, MAPE, SI and Bias as compared to Decision Tree regressor as shown in Figure 36.

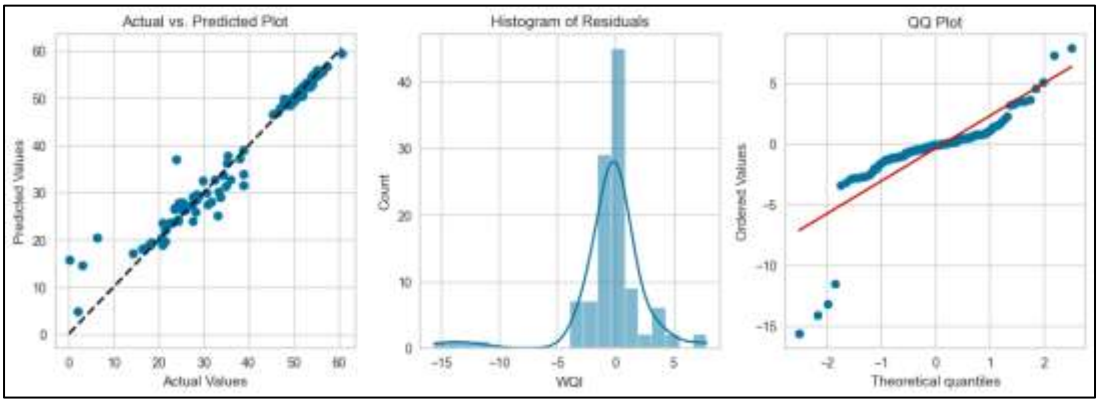


Figure 36. Plots of Random Forest Regressor on Iraq Dataset

On the other hand, the performance of Extreme Learning Regressor and M5 Model Tree on Iraq dataset is optimal and produces satisfactory results as shown in Figure 37 and 38.

ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v3i8.4750

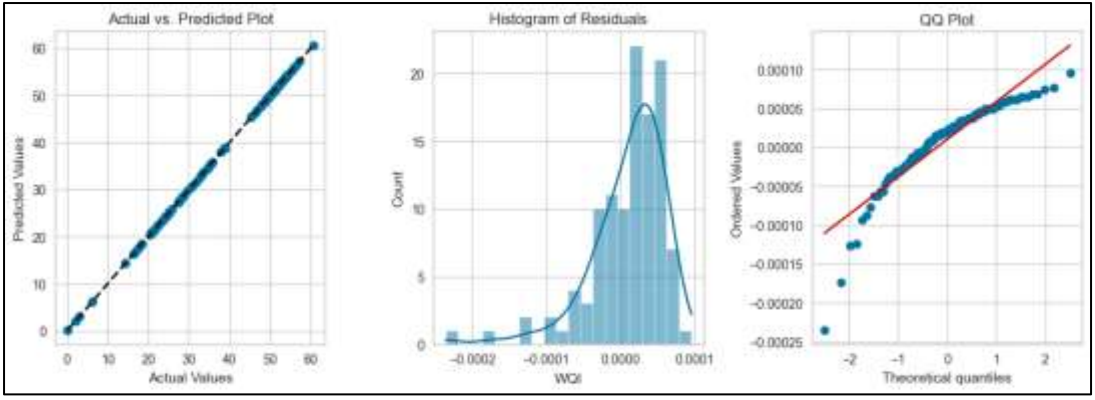


Figure 37. Plots of M5 Model TreeRegressor on Iraq Dataset

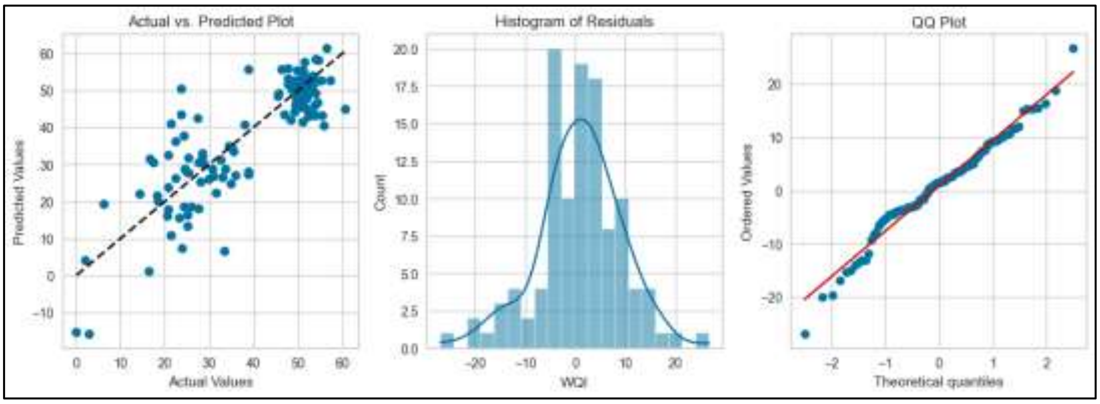


Figure 38. Plots of Extreme Learning Regressor on Iraq Dataset

Classification and Prediction of WQI for India

In order to find the optimal model accuracy and Prediction, various ML models have been implemented on three different datasets using multiple parameters.

Classification Models

According to the experimental data, the Extreme Gradient Boosting model has an accuracy of 0.9710, whereas classification models based on naive bayes, support vector machine, and Ada Boost have accuracy of 0.6467, 0.5734, and 0.2760, respectively. As indicated in Table 9 and Figures 39-45.

Table 9. Classification Accuracy on Indian Dataset

Model	**Accuracy**	**AUC**	**Recall**	**Prec.**	**F1**	**Kappa**	**MCC**	**TT (Sec)**
Extreme Gradient Boosting	0.9710	0.9980	0.9710	0.9717	0.9711	0.9583	0.9584	4.3280
Naive Bayes	0.6467	0.0000	0.6467	0.6944	0.6596	0.5135	0.5206	0.2590
SVM	0.5734	0.8366	0.5734	0.6568	0.5893	0.4355	0.4483	0.5220
Ada Boost	0.2760	0.7326	0.2760	0.3413	0.2098	0.1741	0.2303	0.5460

ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v3i8.4750

Figure 39 and 40 illustrates the performance of XGB model on Indian dataset, it clearly shown ROC curve and confusion matrix of this model has optimal results as compared to other implemented classifiers as presented Table 9. Apart from accuracy of this model, other performance measuring parameters such as Recall, Precision, F1, Kappa, MCC are also much better than other models.

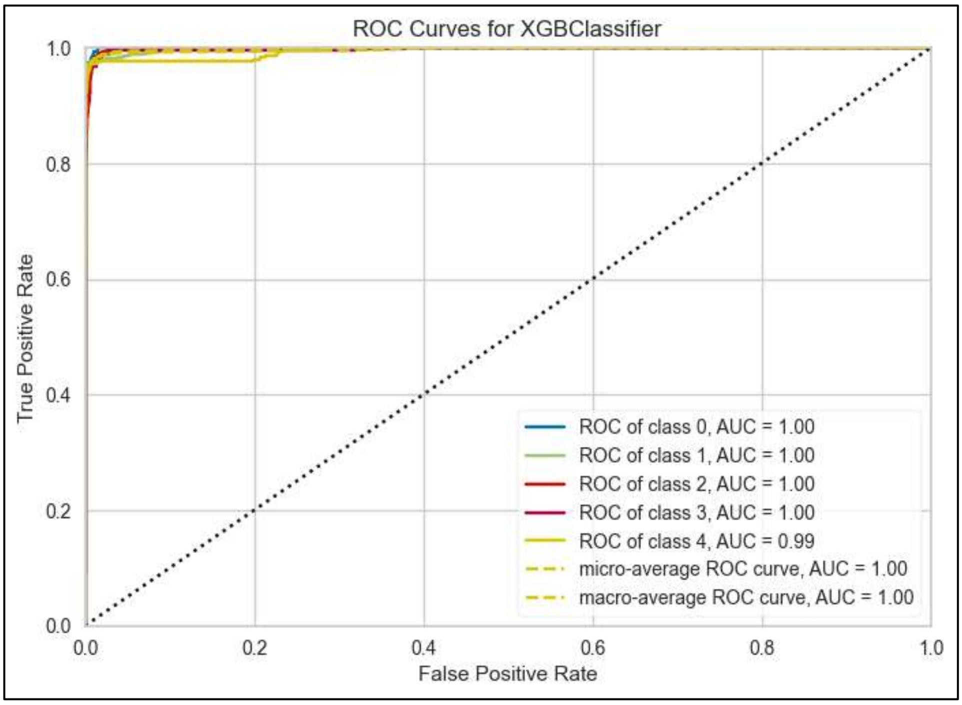


Figure 39. ROC Curves of XGB Classifier on Indian Dataset

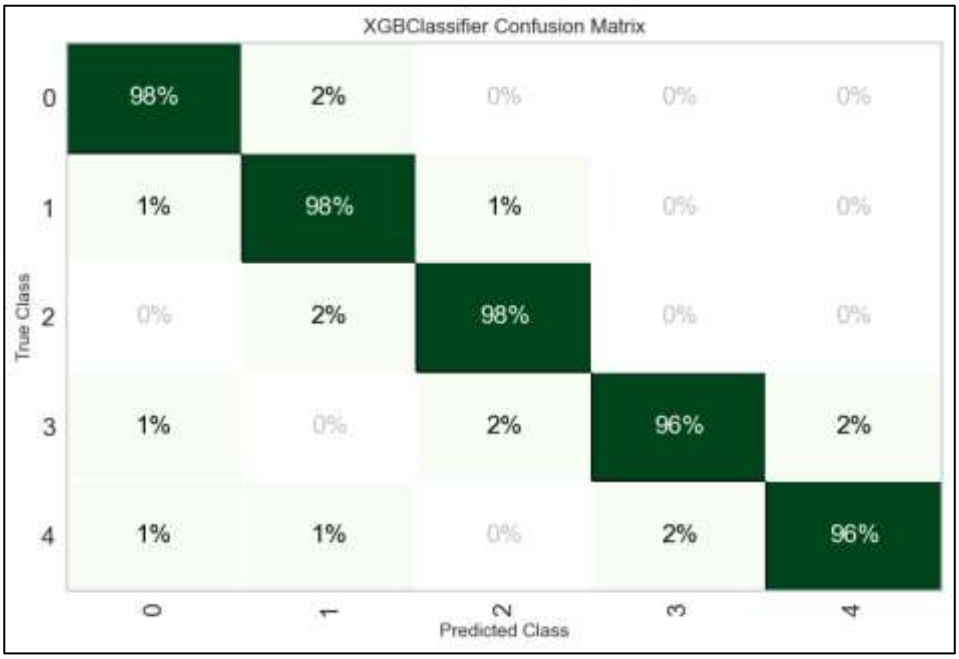


Figure 40. Confusion Matrix of XGB Classifier on Indian Dataset

Figure 41 and 42 illustrates the performance of NB model on Indian dataset, it clearly shown ROC curve and confusion matrix that this model has second highest as compared to adaboost and SVM models presented in the Table 9. Same as XGB, this model also has better results than other models.

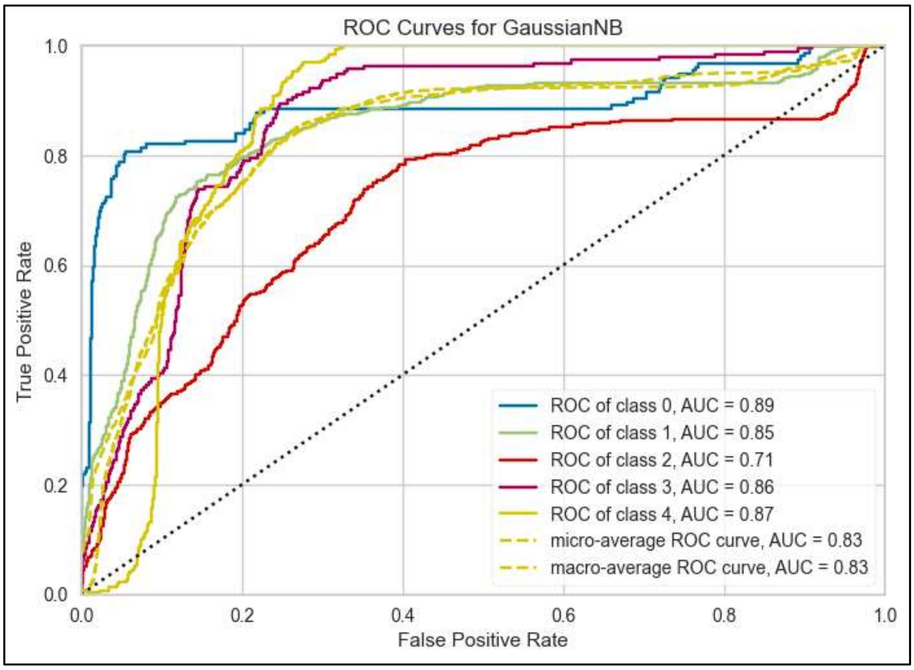


Figure 41. ROC Curves of Naïve Bayes Classifier on Indian Dataset

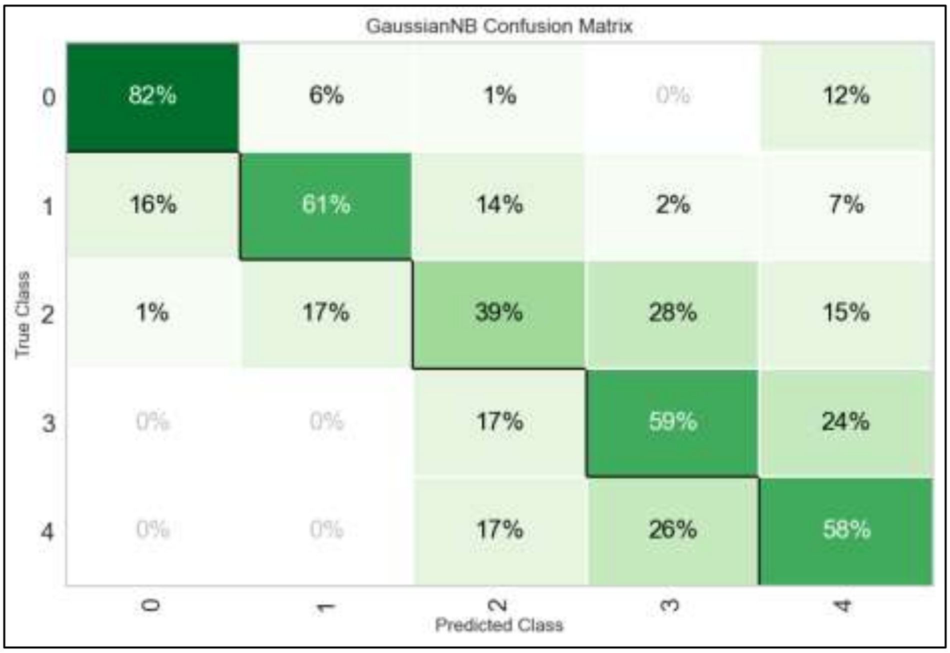
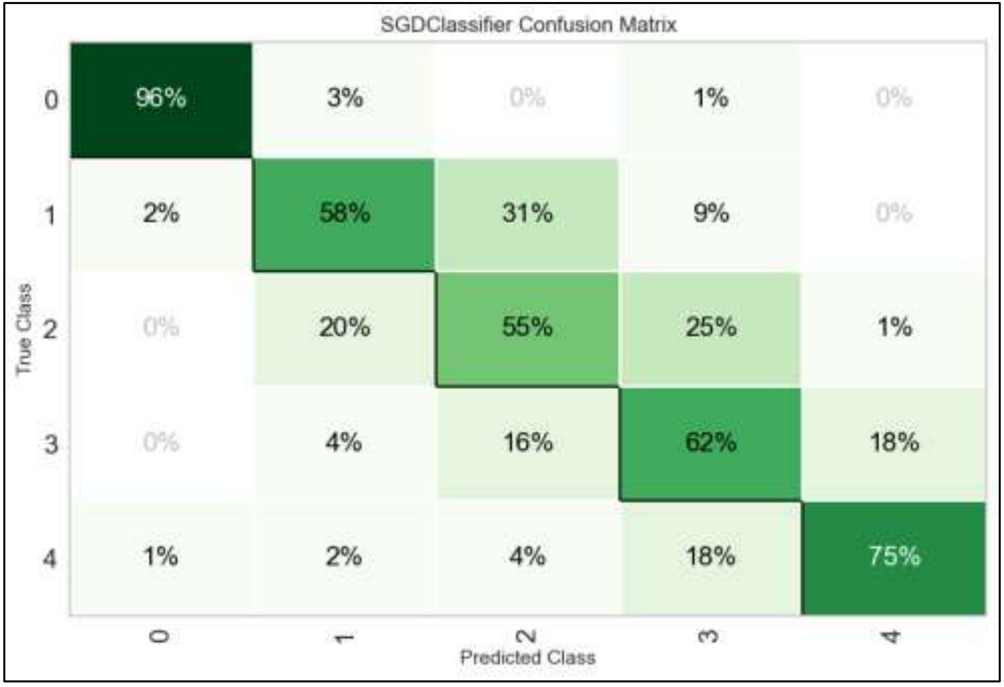


Figure 42. Confusion Matrix of Naïve Bayes Classifier on Indian Dataset

Figure 43 illustrates the performance of SVM model on Indian dataset, it clearly shows that this model has better confusion matrix then Adaboost but lowest than XGB and NB.



	0	1	2	3	4
0	96%	3%	0%	1%	0%
1	2%	58%	31%	9%	0%
2	0%	20%	55%	25%	1%
3	0%	4%	16%	62%	18%
4	1%	2%	4%	18%	75%

Figure 43. Confusion Matrix of SVM Classifier on Indian Dataset

Lastly, Figure 44 and 45 illustrates the performance of AdaBoost classifier model on Indian dataset, it clearly shown ROC curve and confusion matrix that this model has lowest results as compared to other three models presented in the Table 9.

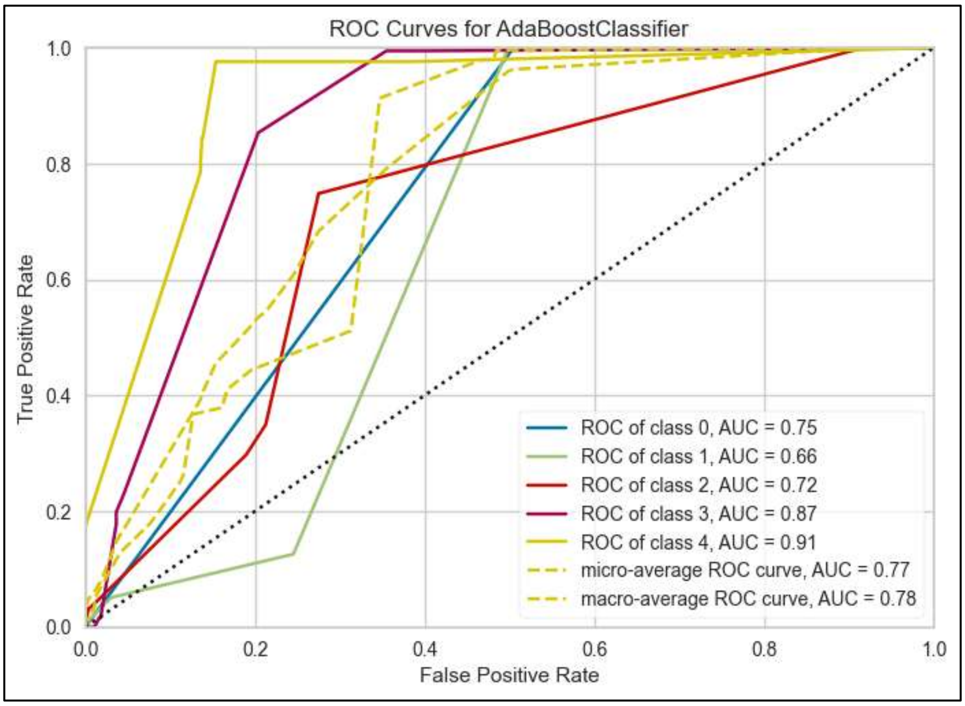


Figure 44. ROC Curves of AdaBoost Classifier on Indian Dataset

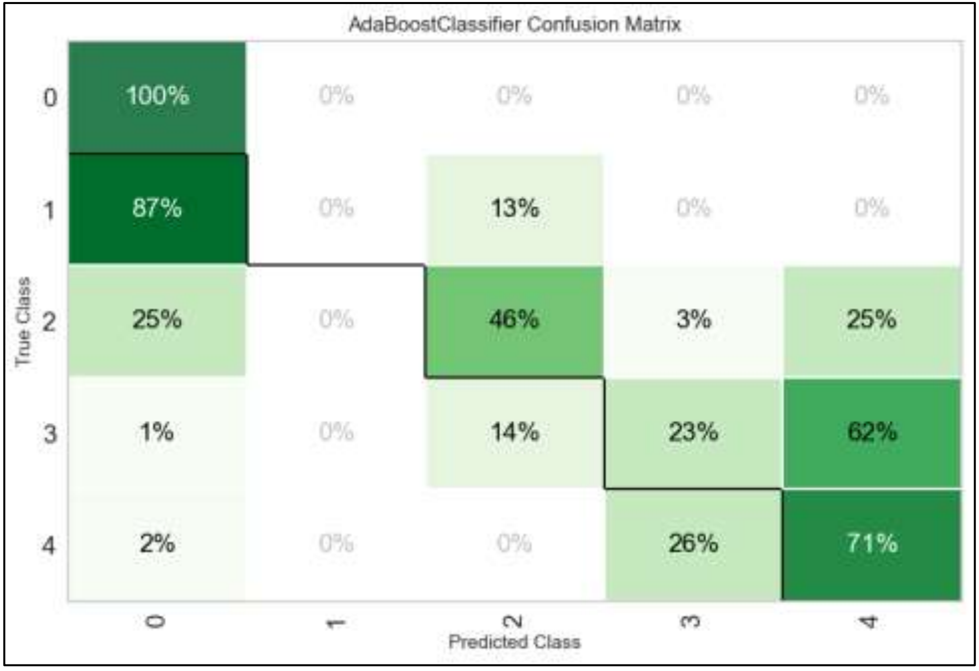


Figure 45. Confusion Matrix of Adaboost Classifier on Indian Dataset

Prediction Models

According to the experimental findings, the MAE of the Decision tree regressor is 1.481, whereas the MAE for random Forest Regressor is 1.033, MAE for M5 Model Tree is 0.226 and MAE for ELM Regressor is 26.47. As demonstrated in Table 10 and Figure 46-49.

Table 10. Performance of prediction Model on Indian Datasets

Model	MAE	MSE	RMSE	MAPE	R^2	S I	BIAS
Decision Tree Regressor	1.481	18.573	4.309	0.018	0.992	0.060	-0.04
Random Forest Regressor	1.033	7.201	2.683	0.013	0.996	0.037	-0.06
M5 Model Tree	0.226	0.097	0.312	0.003	0.999	0.004	-0.00
ELM Regressor	26.475	1883.27	43.396	0.415	0.210	0.605	-0.34

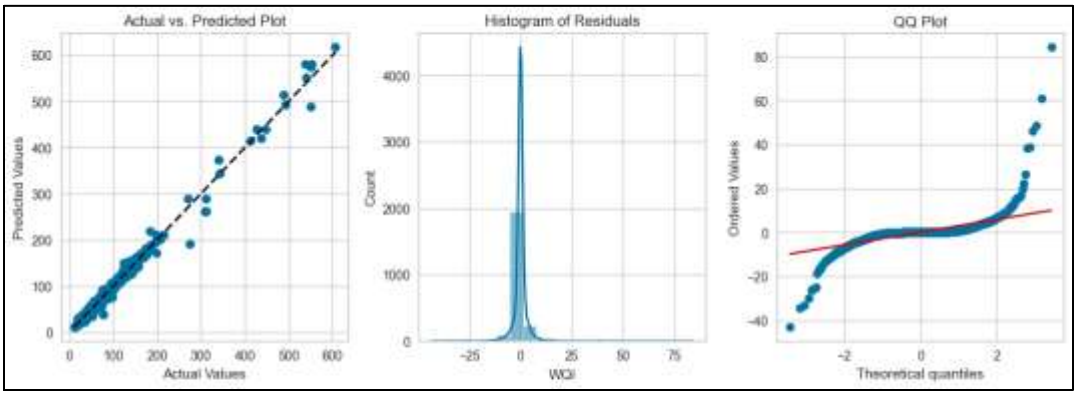


Figure 46. Plots of Decision Tree Regressor on Indian Dataset

Figure 48 illustrates the performance of M5 Model Tree on Indian dataset, the performance of this model in terms of MAE, MSE, RMSE, MAPE, SI and BIAS is better than Random Forest Regressor and Decision Tree regressor.

Whereas the performance of Random Forest regressor is much better than Decision Tree Regressor as shown in Figure 46 and 47.

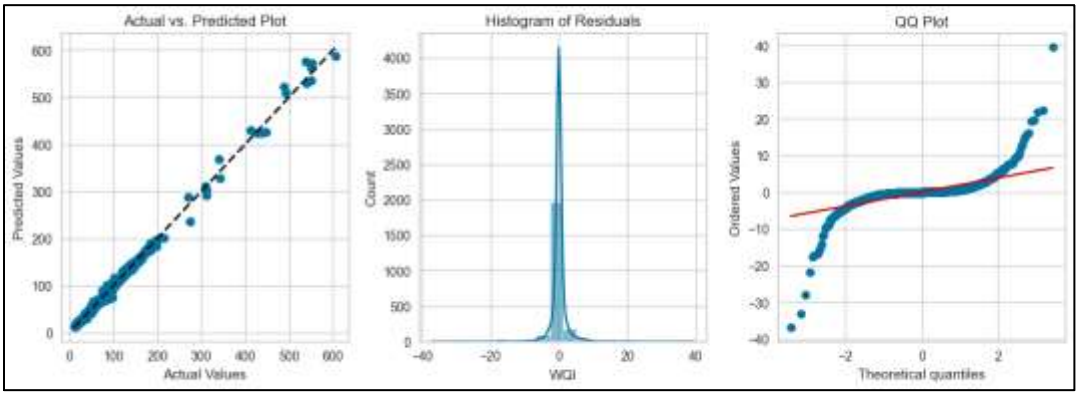


Figure 47. Plots of Random Forest Regressor on Indian Dataset

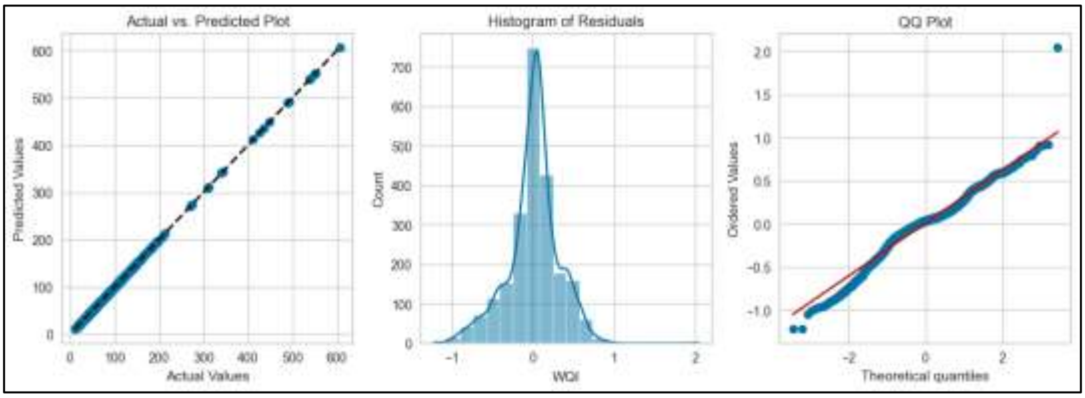


Figure 48. Plots of M5 Model TreeRegressor on Indian Dataset

On the other hand, the Extreme Learning Regressor is unstable on Indian dataset as shown in Figure 49.

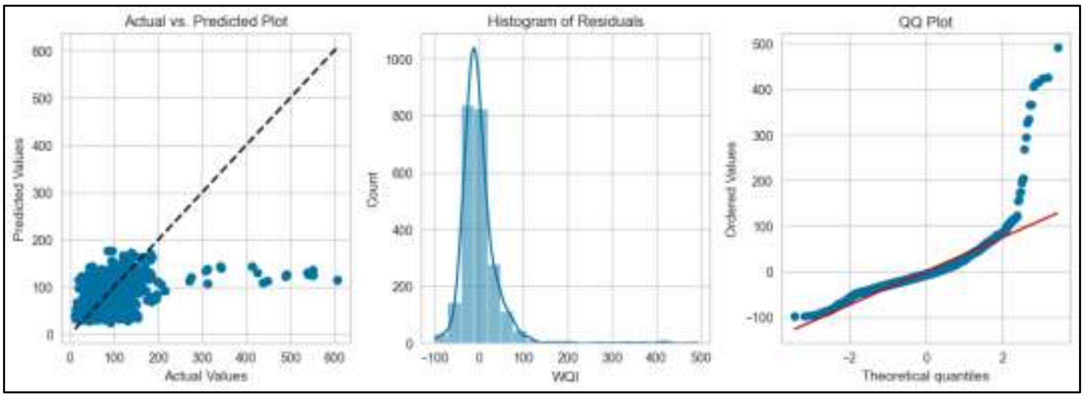


Figure 49. Plots of Extreme Learning Regressor on Indian Dataset

Discussion

Based on the results presented above on all three datasets using eight different ML classifiers and regressor, it shows the different models on different datasets using different parameters produced different results.

The classification results of XGB are among better than other three classification models on all heterogeneous datasets. XGB performs not only in accuracy, AUC, Recall, F1 but Kappa and MCC score is also phenomenal whereas the Kappa rate for XGB on Iraq dataset is lesser than the Malaysian dataset.

Also time taken by XGB for classification on different dataset is worth notable. XGB proofs that the lesser the data in dataset the lesser time took for classification. We know that the Malaysian dataset has less attributes than Iraq, so it took 0.1990 sec for Malaysian dataset and 1.1660 seconds for Iraq dataset. On the other hand, the time taken by XGB on Indian dataset is 4.3280 seconds as we know that this dataset has many attributes and data.

Other than XGB, Naïve Bayes classifier also performed well as compared to XGB on Malaysian and Indian dataset in terms of classification time. Time taken by XGB on Malaysian is 0.1990 seconds whereas NB took 0.0590 seconds. Also, NB took 0.2590 seconds on Indian dataset whereas XGB took 4.3280 seconds which quite much time as compare in terms of using the same resources.

By comparing all four classifiers XGB from boosting family is better in terms of performance but when compared to time execution Naïve bayes seems optimal. Also, the performance of SVM average and the performance of AdaBoost is worst on all three datasets but the time taken by it on all three datasets is quite good than XGB.

Comparison of four different regressor model on all three datasets shows that each regressor have their own abilities to perform better based on the attributes provided. The results shows that the performance of M5 Model Tree regressor is better than other prediction model on each dataset except Malaysian. On Malaysian dataset, random forest regressor perform better in terms of MAE, MSE, RMSE, and SI. Bias values on Malaysian dataset is uncertain between decision tree and random forest while the bias score of M5 Model Tree on Malaysian dataset is 9.200 which is quite much. Also, M5 Model Tree perform better on Iraq dataset in terms of MAE, MSE, RMSE, R2 and Bias but MAPE and SI was optimal by decision tree regressor. On Indian dataset, M5 Model Tree outperformed all regressor in all performance attributes. But it is quite challenging for M5 Model Tree model to compete with other models which are smaller in size as it can be shown in table 6 and 8 that bias rate of M5 Model Tree on Malaysian dataset is so high and Scatter index on Iraq dataset is also much as compared to other regression models.

Based on the comparison of results with the baseline [4] and [26-27], it is found that the proposed models performs better than [4] and near to performance of [26-27].

Overall, the results presented in different tables (Table 1-10) and Figures (4-48) of all eight ML classifiers and regressors are tested and validated on all dataset and it produces optimal results as well as gives some insights on usage of ML for small datasets. Also, results shows that the addition of Scatter index and Bias helps researchers to know more about the ML behaviour towards data and its impact on outcome.

Conclusion

In this work, the effectiveness of machine learning algorithms for classification and prediction of water quality index was examined. WQI classifiers were created using machine learning methods as XGBoost, Naive Bayes, support vector machines, and Adaboost algorithms. Data from rivers in India, Iraq, and Malaysia were gathered, modelled, and used to create models. These parameters included BOD, DO, TC, nitrate, pH, temperature, and others. Performance measures were used to assess the models' performance in classifying the river water quality index.

The water quality is classified using machine learning methods such XGBoost, Naive Bayes, SVM, and Ada Boost for measuring the water quality index whereas the prediction of water performed using RF regressor, M5 Model Tree, DT regressor, EML regressor on the samples of Malaysian, Indian, and Iraqian rivers. The performance of XGBoost accurately identifies the water quality index with 93%, 92%, and 97% Accuracy, Precision and recall respectively. Whereas the performance of M5 Model Tree for prediction is much better than other prediction models. The developed models provide a promising result for the classification of water quality indexes and prediction.

In the future, other Boost families and ensemble learning models will be implemented to increase the effectiveness of the categorization and prediction of water quality.

Competing Interests: The authors declare that they have no conflicts of interest.

Funding: Not applicable.

Clinical Trial: Not applicable.

References

Panneerselvam, Balamurugan, et al. "Quality and Health Risk Assessment of Groundwater for Drinking and Irrigation Purpose in Semi-Arid Region of India Using Entropy Water Quality and Statistical Techniques." Water 15.3 (2023): 601.

Qasemi, Mehdi, et al. "Characteristics, water quality index and human health risk from nitrate and fluoride in Kakhk city and its rural areas, Iran." Journal of Food Composition and Analysis 115 (2023): 104870.

https://www.un.org/en/global-issues/water [Accessed on 6th september 2023 05:00:00 AM]

Nair, Jitha P., and M. S. Vijaya. "River water quality prediction and index classification using machine learning." Journal of Physics: Conference Series. Vol. 2325. No. 1. IOP Publishing, 2022.

Sillberg, Chalisa Veesommai, Pratin Kullavanijaya, and Orathai Chavalparit. "Water quality classification by integration of attribute-realization and support vector machine for the Chao Phraya River." Journal of Ecological Engineering 22.9 (2021): 70-86.

Yilma, Mulugeta, et al. "Application of artificial neural network in water quality index prediction: a case study in Little Akaki River, Addis Ababa, Ethiopia." Modeling Earth Systems and Environment 4 (2018): 175-187.

Ahmed, Umair, et al. "Efficient water quality prediction using supervised machine learning." Water 11.11 (2019): 2210.

Sakizadeh, Mohamad. "Artificial intelligence for the prediction of water quality index in groundwater systems." Modeling Earth Systems and Environment 2 (2016): 1-9.

J. P. Nair and M. S. Vijaya, "Predictive Models for River Water Quality using Machine Learning and Big Data Techniques - A Survey," 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), Coimbatore, India, 2021, pp. 1747-1753, doi: 10.1109/ICAIS50930.2021.9395832.

Nair, Jitha P., and M. S. Vijaya. "Analysing And Modelling Dissolved Oxygen Concentration Using Deep Learning Architectures." International Journal of Mechanical Engineering 7 (2022): 12-22.

Nair, Jitha P., and M. S. Vijaya. "Design and development of efficient water quality prediction models using variants of recurrent neural networks."

Nair, Jitha P., and M. S. Vijaya. "Temporal fusion transformer: a deep learning approach for modeling and forecasting river water quality index." International Journal of Intelligent Systems and Applications in Engineering 11.10s (2023): 277-293.

Malek, Nur Hanisah Abdul, et al. "Prediction of water quality classification of the Kelantan River Basin, Malaysia, using machine learning techniques." Water 14.7 (2022): 1067.

Fernández del Castillo, Alberto, et al. "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 (2022): 1235.

Nguyen, Duc Hai, et al. "Development of an extreme gradient boosting model integrated with evolutionary algorithms for hourly water level prediction." IEEE Access 9 (2021): 125853-125867.

Xia, Jingjing, and Jin Zeng. "Environmental factor assisted chlorophyll-a prediction and water quality eutrophication grade classification: A comparative analysis of multiple hybrid models based on a SVM." Environmental Science: Water Research & Technology 7.6 (2021): 1040-1049.

Mosavi, Amirhosein, et al. "Ensemble boosting and bagging based machine learning models for groundwater potential prediction." Water Resources Management 35 (2021): 23-37.

Sheng, Liming, et al. "Water quality prediction method based on preferred classification." IET Cyber-Physical Systems: Theory & Applications 5.2 (2020): 176-180.

Al-Jaf, Hnar Ali Karim. "Water Quality Index Application to Evaluate the Ground Water Quality in Kalar City-Kurdistan Region-Iraq." IOP Conference Series: Earth and Environmental Science. Vol. 1120. No. 1. IOP Publishing, 2022.

Martinho, Alfeu D., Henrique S. Hippert, and Leonardo Goliatt. "Short-term streamflow modeling using data-intelligence evolutionary machine learning models." Scientific Reports 13.1 (2023): 13824.

Zakaria, Muhamad Nur Adli, et al. "Exploring machine learning algorithms for accurate water level forecasting in Muda river, Malaysia." Heliyon 9.7 (2023).

Najafzadeh, M., A. Ghaemi, and S. Emamgholizadeh. "Prediction of water quality parameters using evolutionary computing-based formulations." International Journal of Environmental Science and Technology 16 (2019): 6377-6396.

Najafzadeh, Mohammad, and Alireza Ghaemi. "Prediction of the five-day biochemical oxygen demand and chemical oxygen demand in natural streams using machine learning methods." Environmental monitoring and assessment 191 (2019): 1-21.

Najafzadeh, Mohammad, and Saeid Niazmardi. "A novel multiple-kernel support vector regression algorithm for estimation of water quality parameters." Natural Resources Research 30, no. 5 (2021): 3761-3775.

Najafzadeh, Mohammad, Farshad Homaei, and Hadi Farhadi. "Reliability assessment of water quality index based on guidelines of national sanitation foundation in natural streams: Integration of remote sensing and data-driven models." Artificial Intelligence Review 54, no. 6 (2021): 4619-4651.

Najafzadeh, Mohammad, Elahe Sadat Ahmadi-Rad, and Daniel Gebler. "Ecological states of watercourses regarding water quality parameters and hydromorphological parameters: deriving empirical equations by machine learning models." Stochastic Environmental Research and Risk Assessment (2023): 1-24.

Najafzadeh, Mohammad, and Sajad Basirian. "Evaluation of river water quality index using remote sensing and artificial intelligence models." Remote Sensing 15, no. 9 (2023): 2359.