CENTRAL ASIAN JOURNAL OF WATER RESEARCH (2026) 12(1): 68-92

https://doi.org/10.29258/CAJWR/2026-R1.v12-1/68-92.eng



Central Asian Journal of Water Research

# Performance of Different Ensemble Machine Learning Algorithms in Classification of the Water Quality Indices

Farid Hassanbaki Garabaghi [a], Semra Benzer [b], and Recep Benzer [c]

[a] Gazi University, Graduate School of Natural and Applied Sciences, Teknikokullar, 06500 Ankara, Turkey; fgarabaghi@gmail.com

[b] Gazi University, Gazi Education Faculty, Teknikokullar, 06500 Ankara, Turkey; sbenzer@gazi.edu.tr

[c] Ankara Medipol University, Department of Management Information System, Ankara, 06050 Ankara, Turkey; recep.benzer@ankaramedipol.edu.tr

## ABSTRACT

Monitoring the quality of ground freshwater resources is crucial due to their limited availability and susceptibility to contamination from unchecked human activities. The Water Quality Index (WQI) serves as a tool for assessing basin water quality, yet its complexity and time-consuming nature have steered scientists toward simpler methods like machine learning algorithms. This study evaluated four machine learning algorithms using an ensemble learning approach to construct a high-performing classification model. Additionally, three feature selection methods using machine learning were applied to determine the most influential water quality parameters in the classification process. Among the classifiers, XGBoost demonstrated exceptional performance, achieving 96.9696% accuracy when considering all parameters. To optimize cost-effectiveness, the XGBoost classifier with a 95.606% accuracy using 10-Fold Cross Validation and seven key parameters, identified by Backward Feature Elimination, is recommended for classification on the dataset used in this study.

## ARTICLE HISTORY

Received: 17 June, 2025
Accepted: 22 May, 2026
Published: 8 June, 2026

## KEYWORDS

Water Quality Index, Classification, Machine learning, Feature selection, Ensemble Learning

CONTACT Semra Benzer ✉ sbenzer@gazi.edu.tr, Gazi University, Gazi Education Faculty, Teknikokullar, 06500 Ankara, Turkey

Published by Kazakh German University, Almaty/Kazakhstan

## 1. Introduction

Although roughly 90% of all freshwater resources are groundwater, only a minuscule portion of it is accessible to be used by humans as drinking water (Arabgol et al., 2016; Ostad-Ali-Askari et al., 2017). Being related to rapid population growth, today, anthropogenic activities such as agricultural and industrial as the main source of surface water contamination, increase the call for sustainable administration and management (Motevalli et al., 2019) of this irreversible asset for the living of humankind on this planet (Dohare et al., 2014).

Not only it's of paramount importance from the human-wellbeing point of view to constantly monitor the quality of the surface water, but it also plays a pivotal role in preserving the aquatic ecosystem. For example, when the nutrients such as N and P and/or their derivatives willfully or ignorantly are released to the aquatic systems, unwanted consequences such as eutrophication of the surface water reservoirs will occur and result in perishing the aquatic life as the nutrient enrichment causes algal blooming and resultingly decrease the dissolved oxygen (DO) levels in aquatic systems and consequently inhibiting and restricting the proper environment for the aquatic organisms to be grown (Rozemeijer & Broers, 2007; Varol & Şen, 2012).

One tool that has aided environmental scientists as well as decision-maker agencies in association with the water quality to assess the quality of water without letting themselves in for the frustrating task of analysis of water quality parameters is Water Quality Index (WQI). One of the advantages of WQI is that it can unify the massive data into one value which is simpler, more understandable, interpretable, and consequently logical (Tyagi et al., 2013; Uddin et al., 2021, Lap et al., 2023; Siriwardhana et al., 2023).

Besides advantages, however, interpretation of water quality based on WQI has many drawbacks such as the discrepancy between the WQI methods as there are numerous and various equations that are used worldwide by different agencies (Tyagi et al., 2013; Bui et al., 2020). Besides, the method is quite time-consuming and complex (Bui et al., 2020). Hence, due to these disadvantages, researchers, today, tend to use simpler, more reliable, and effective methods in classification and modeling water quality.

Flourishing advanced technologies, especially the application of artificial intelligence (AI) in a variety of research lines, today, environmental scientists tend to evaluate the accuracy and reliability of different machine methods in their studies (particularly in the classification of water quality) to firstly evade complicated calculations, and secondly expedite the process of classification as well as a prediction as these methods have recorded promising results in recent years in this regard.

Saghebian et al., (2014) investigated the accuracy and compared the performance of the Kappa statistical and Decision Tree methods for the classification

of groundwater quality indices based on the United States Salinity Laboratory (USSL) diagram in which the quality of the groundwater for irrigation is divided into four categories. Moreover, Principal Component Analysis, which is based on the statistical analysis of the correlation between the parameters, was conducted to reduce the number of variables effective in the classification of groundwater quality. As a result, the Decision Tree model was reported to be more effective and precise in comparison with the Kappa statistical model. Modaresi and Araghinejad (2014) evaluated the performance of three ML classifiers (i. e., PNN, SVM, KNN) in water quality indices classification. As a result, SVM outperformed the other two classifiers with outstanding performance. Danades et al., (2016) evaluated the performance of SVM and KNN models in classifying water quality indices. The results indicated that the SVM classifier could classify the indices better than KNN with a maximum accuracy of 92.40%. Radhakrishnan and Pillai (2020) compared the performance of three ML algorithms (i. e., SVM, Decision Tree, and Naïve Bayes) for classifying five water quality indices. Four water quality parameters were considered in this study (i. e., pH, DO, BOD5, and EC) for the classification process. The performance results of the classifiers demonstrated the effectiveness of the Decision Tree classifier over the other two with an accuracy of 98.50%.

Cheema et al. (2024) various prediction models such as autoregressive integrated moving average (ARIMA), exponential smoothing, support vector machines and artificial neural networks were used in East and Central Asian water quality studies. The prediction capabilities of these models were evaluated using mean absolute percentage error (MAPE). Support vector machines were found to be more accurate than ARIMA, exponential smoothing and artificial neural networks in predicting deaths and disability-adjusted life years.

Despite satisfying results, the accuracy of the classifiers can be affected by high-dimensional data as well as noise. That is mainly because of the overfitting conditions and costly computational tasks. Bouamar and Ladjal (2007) evaluated and compared the performance of two machine learning algorithms (i.e., ANN & SVM) in classifying water quality indices. As a result, the recognition accuracy of the ANN and SVM algorithms was 86.24 % and 86.31%, respectively. Furthermore, the chief disadvantage of both models was reported to be their sensitivity to noise. Therefore, it can be concluded that where the data is noisy, these models are not considered effective methods in the process of prediction as well as classification. Thus, eliminating the noise and selecting the most significant parameters can reduce the overfitting risk and improve the time and accuracy of the classification methods (Uyun & Sulistyowati, 2020).

Although there are some studies in the literature with a focus on selecting the most prominent attributes in the classification of water quality, they are not

machine learning approaches. For example, Dezfooli et al., (2018) investigated in regard with the performance evaluation of three machine learning algorithms (i.e., probabilistic neural network (PNN), k-nearest neighbor, and support vector machine (SVM)) in the classification of five water quality indices with and without feature selection to identify the most important parameters affecting the water quality as well as reducing the number of the parameters, which is an important aspect of the cost-effectiveness of the machine learning models. However, the method used for the feature selection was not that of the machine learning approach. It was reported that the classification accuracy of the PNN model weighed out the other two models with the best performance of 90.70% with only three input parameters. Bui et al., (2020) conducted a comprehensive study on the performance evaluation of 16 ML algorithms in classifying six (6) water quality indices based on IRWQIsc. These algorithms contained four (4) standard classifiers (i. e., random forest, M5P decision tree, random tree, and reduced error pruning a tree), and 12 hybrid classifiers with a combination of standard classifiers with bagging, CV parameter selection, and randomizable filtered classification algorithms. The method used for feature selection in this study was the manual filter feature selection method. As a result of this study, the BA-RT (Bagging-Random Tree) hybrid algorithms showed outstanding performance over the other 15 models with an accuracy of 94.1%.

Yet, there are some studies in which feature selection based on a machine learning approach has been conducted. Muhammad et al., (2015) evaluated the performance of five machine learning algorithms (i. e., Naïve Bayes, Conjunctive Rule, J48 Decision Tree, Kstar Lazy algorithm, and Bagging algorithm) in the classification of water quality indices. The performance evaluation of the models was carried out based on the number of attributes involved in the classification, and the CfsSubsetEval (a wrapper feature selection method) from the library of the Weka tool was used to extract the most important attributes in the classification process. According to the result of the feature selection algorithms, which suggested that seven (7) features are the most crucial attributes among the total 53 parameters, the Kstar algorithm performed the best among the five used algorithms in this study with maximum accuracy of 86.67%.

The objective of this study is to investigate the performance of different ensemble machine learning models in the classification of the water quality indices, which are generated to indicate the suitability of water resources in terms of water quality for public consumption. Furthermore, the secondary aim of this study is to investigate the effect of reduced water quality attributes by different feature selection methods on the performance of machine learning algorithms in the classification of water quality indices. Last, the effect of different dataset splitting methods on the performance of the machine learning algorithms was also investigated in this study.

## 2. Methods and Materials

Various classification algorithms such as LogitBoost, Random Forest, AdaBoost, and XGBoost were evaluated for their performance in classifying water quality indices. Additionally, different feature selection methods utilized in this study to identify the most significant parameters for the classification process are discussed. The study's overall workflow is visually depicted in Figure 1.

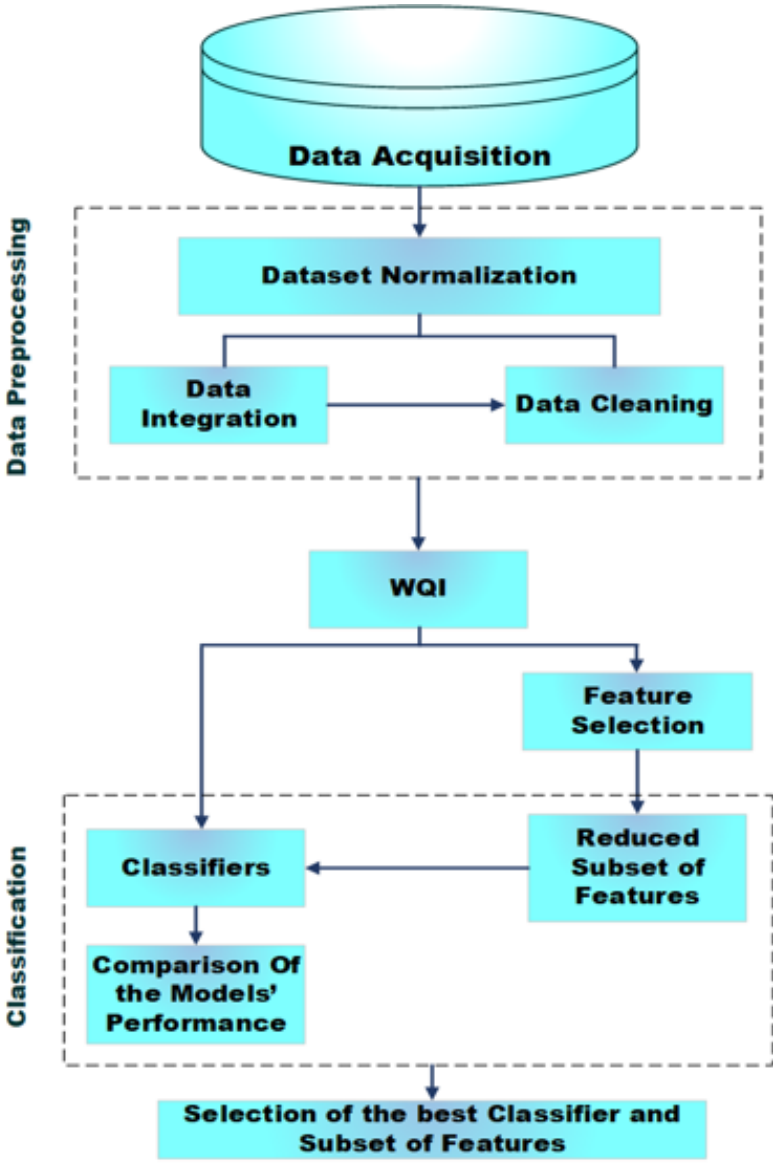



Figure 1. Schematic workflow of the study.

### *2.1. Study Area*

Büyük Menderes Basin is in the west of Anatolia, basically including the Büyük Menderes River and its streams which discharge into the Aegean Sea. The basin is surrounded by Samsun Mountain, Cevizli Mountain, Elma Mountain, and Murat Mountain from the north, Sandıklı Mountains from the east, Madran Mountain, Babadağ, and Bozdağları from the south, and the Aegean Sea in the west. The

basin area is approximately 2,600,967 ha. The major industries in this basin include agriculture, livestock, food industry, and tourism which can be considered the main contamination sources of the basin. Almost 79% of the waters in the basin are used for agriculture and 21% is used by industries as well as dwellings (General Directorate of Environmental Management, Turkey, 2016). Since the Büyük Menderes Basin is one of the most contaminated basins in Turkey, the survival of the basin needs to be monitored regularly in terms of water quality. That's why this basin was selected as the study area in this research.

### *2.2. Data acquisition*

For this study, the water quality data was compiled and provided by The General Directorate of State Hydraulic Works (GDSHW) in Turkey. There were 142 stations in the basin which had been operating by GDSHW from 1985 to 2014 to measure the water quality of the basin in roughly regular monthly intervals. Among these 142 stations, the water quality data of 10 stations between 2004-2014 were selected to represent the downstream of the basin. One reason for this was the abundance of the measured parameters and the frequency of the observations because some stations were in inaccessible places, and in some seasons with a harsh climate, it was difficult to reach the station for sampling. Therefore, in many of the stations, based on the demand for the sampling and measurement, due to the possible contamination, only the most important water quality parameters were measured, regularly.

### *2.3. Dataset formation*

After data acquisition, it was necessary to create a dataset of the essential parameters for this study. This is an essential part of data preprocessing called data integration used for merging the data from different sources to make a unified dataset to serve the purpose of any study (Kumar & Manjula, 2012). The data of 19 parameters (i. e., Year, Month, Station, T (°C), pH, EC (micromhos/cm), DO (Mg O2/l), COD (Mg/l), BOD5 (Mg/l), NH4 (Mg/l), NO3 (Mg/l), NO2 (Mg/l), PO4 (Mg/l), Cl (Mg/l), Fe(Mg/l), Mn (Mg/l), Na+ (Mg/l), SO4 (Mg/l), TDS (Mg/l)) was extracted from the provided dataset by the GDSHW, to create a unified dataset consisting 660 instances.

### *2.4. Missing value imputation*

Encountering missing value in environmental studies is an unavoidable fact due to the impossibility of throughgoing observation of continual processes. Moreover, a corrupt dataset with missing values can end up with an incorrect conclusion as a result of data inconsistency in the process of data analysis. Although there are many missing value imputation methods (e.g, zero imputation, mean imputation, median imputation, regression imputation, etc.) which have been used for data cleaning

in many studies, the capability of machine learning algorithms in predicting the missing values and completing the datasets with logical values have attracted the researchers' attention in recent years to utilize these techniques for data cleaning. Among many machine learning algorithms used for predicting the value of a target parameter (i. e., ANN, KNN, LLR, LR, etc.), the linear regression (LR) algorithm was chosen to be used for missing value prediction in this study due to the simplicity of application of this model as well as its efficiency in the prediction of missing values where more than 10% of the whole dataset are missing values. A linear regression algorithm simply performs a regression task by predicting the target value (Y) based on the regression which is made by considering the independent variables (X) (Eq. 1) (Kumar and Manjula, 2012; Yozgatligil et al., 2013; Sim et al., 2015; Ighalo et al., 2020).

$$Y_i = \beta_0 + \beta_i X_i + \varepsilon_i \quad (1)$$

The schematic framework of developing the LR model for missing value imputation is shown in Figure 2 in which the dataset was divided into training and test datasets by dropping the missing values from the whole dataset to create the test dataset for the prediction process. Then LR model was trained by the training dataset to make the regression between variables and predict the values in the test dataset as a result of its learning from training data. In the end, the predicted values were merged into the training dataset to make a whole dataset without missing values.

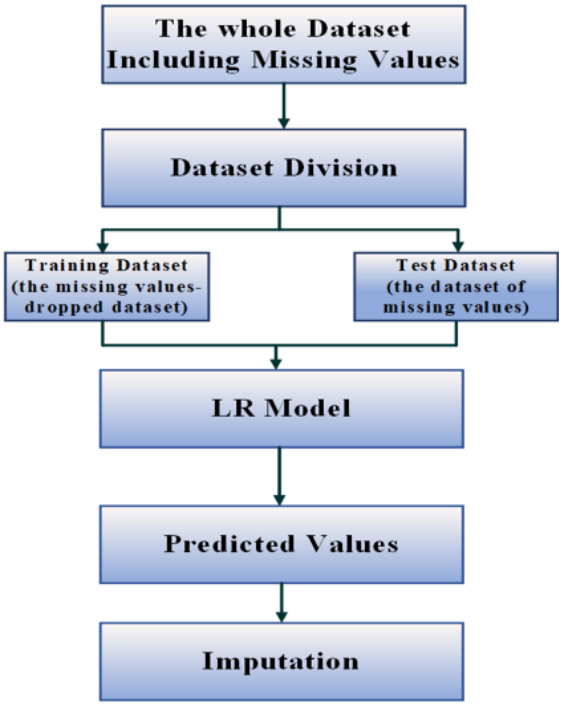



Figure 2. The schematic process of the missing value imputation by LR machine learning algorithm.

### *2.5. Water Quality Index (WQI)*

Once the data was cleaned, it was time to diminish the massive information and bring it into a single logical expression to describe the general quality of the water for public use. One tool to serve this purpose is to use the Water Quality Index (WQI) which are functions developed by national and international agencies worldwide to describe the general status of the water systems. Generally, WQI functions follow a process of four stages in which, depending on the possible contamination in a water system as well as the experts' opinions, the parameters of interest are chosen first. Then to create a single-valued subindex, the concentrations are converted. The third stage is the calculation of the unit weight for each water quality parameter. Finally, a value indicating the WQI is calculated based on previously calculated subindexes and unit weights (Tyagi et al. 2013; Mădălina & Gabriela, 2014; Uddin et al., 2021).

In this study, Weighted Arithmetic Water Quality Index (WAWQI) method (Tyagi et al., 2013) (Eq. 2) was chosen to generate the class labels of the water quality of Büyük Menderes Basin according to 16 water quality parameters observed in 10 stations between 2004-2014 in two-months intervals.

$$WAWQI = \sum Q_i W_i / \sum W_i \tag{2}$$

Where,

Qi is the quality rating of the ith parameter and Wi is the unit weight which is calculated by equations 3 and 4, respectively.

$$Q_i = 100 * [(V_i - V_0)/(S_i - V_0)] \tag{3}$$

Where,

Vi is the estimated concentration of the ith parameter and V0 is the ideal value of the ith parameter.

$$W_i = K/S_i \tag{4}$$

Where,

Si is the recommended standard value of the ith parameter and K is the proportionality constant which is calculated through Eq. 5.

$$K = \frac{1}{\sum 1/S_i} \tag{5}$$

In this study the recommended standard values for the aforementioned parameters which were included in the calculation of WQI were obtained from the

WHO Guidelines for Drinking Water (WHO, 2006) and Turkish Standards (TS-266): Water Intended for Human Consumption (TSE, 2005).

According to the outcome of the WAWQI function, the water quality is classified into five major classes based on Table 1.

Table 1. Water Quality Classes Based on Weight Arithmetic Water Quality Index Method.

| WQI Value | Class |
|---|---|
| 0-25 | Excellent Water quality |
| 26-50 | Good Water Quality |
| 51-75 | Poor Water Quality |
| 76-100 | Very Poor Water Quality |
| Above 100 | Unfit for Consumption |

### *2.6. Ensemble learning approach*

In ensemble learning to make a decision, the combinations of single classifiers are accumulated to provide not only a single but also an improved predictive model in which the final result is based on the prediction of all individual classifiers (Dong et al., 2020). Ensemble Random Forest Classifier, XGBoost Classifier, AdaBoost Classifier, and LogitBoost Classifier which were applied in this study are explained below.

#### *2.6.1. Random Forest classifier*

The combination of different decision trees which was entitled Random Forest by Breiman (2001) takes the advantage of using bootstrap aggregation and bagging in selecting an instance to reduce the classification error. This supervised classifier randomly selects x numbers of features from all features and subsequently, it picks out a node among selected x features as a candidate split node. Moreover, Random Forest is known as a beneficial and satisfying classification method, due to its ability in handling high dimensional data and diverse types of data especially nonparametric data types (Arora & Kaur, 2020).

#### *2.6.2. AdaBoost classifier*

AdaBoost Classifier which is also called adaptive boosting (Freund & Schapire, 1997), is another category of the most widely used ensemble methods. It uses the concept of the combination of sequential independent singular hypotheses to increase the performance of each weak learner. As a result, the distribution of the training set in all individual classifiers is altered which contributes to the reduction of training

error and at the end of the training process in each iteration, weights are adjusted. Moreover, the weight of the accurately predicted instances will be decreased and the final result of the ensemble learning will be based on the combination of the results of all single classifiers. Different machine learning classifiers such as SVM, Decision Tree, etc., can be a base classifier of the AdaBoost (Liu et al., 2020).

### *2.6.3. XGBoost classifier*

Extreme Gradient Boosting (XGBoost) (Chen & Guestrin, 2016) which is an ensemble-based approach has been one of the prosperous methods in data classification and regression problems. Since XGBoost uses the gradient descent algorithm, it can decrease the loss during adding new models. Moreover, it generates a sequential decision tree in which each instance is given a weight value as a probability indicator of being chosen by each Decision tree in the next steps. However, the primary weight value of all instances is the same and it is altered after analysis in each step. Moreover, the second classifier is built on the results of the data value of the first step, and the previous model errors are resolved. Consequently, the process will continue until all classifiers are built based on the results of the previous step alongside reducing errors of the previous steps. Additionally, since the XGBoost algorithm uses a parallel implementation of sequential trees and reduces the overfitting problem, it is considered one of the advantageous classifiers (Bhati et al., 2021).

### *2.6.4. LogitBoost classifier*

LogitBoost Classifier (Friedman et al., 2000) is an ensemble boosting algorithm which was proposed for classification problems. The main purpose of introducing the LogitBoost classifier was to boost AdaBoost in dealing with noisy data to decrease training errors and improve its classification performance. Hence, applying logistic regression to AdaBoost generates the LogitBoost which facilitates the application of multiclass problems and minimizes the logistic loss, bias, and variance (Tehrany et al., 2019).

## *2.7. Feature selection methods*

Not only do feature selection methods decrease the dimensions of features that contribute to the elimination of irrelevant features, but they also select the most effective subset of features that are relevant to the main class/classes. Hence, it results in reducing the size of the dataset, shortening the training time, improving the accuracy of the classification, and decreasing the overfitting problem. Since, an insignificant small subset resulting from feature selection can mislead the classifier, selecting the most significant features is an essential issue. Filter, wrapper, and embedded methods are three categories of feature selection methods.

### 2.7.1. Filter method

In the filter method, the subset is selected by considering features' characteristics such as distance or correlation among features and features' rank among others, instead of applying machine learning algorithms. Filter methods are considered as fast and scalable methods due to taking advantage of statistical methods in selecting features. They, however, overlook the dependency of features and the interaction with classifiers (Khaire & Dhanalakshmi, 2019). Mutual information as a filter method which was applied in this study is described below.

Mutual information

I(X; Y) indicates the Mutual Information between two random variables of X and Y which calculates the mutual dependency between X and Y in terms of information theory. Since Mutual Information does not consider any dependency between variables such as linearity or continuity, it is more generic than linear methods like correlation coefficient which makes it a more advantageous method. Equation 6 indicates mutual information of X and Y variables.

$$\mathrm{I(X;Y)} = \sum_{X \epsilon x} \sum_{Y \epsilon y} p(x, y) log \frac{p(x,y)}{p(x)p(y)} \tag{6}$$

The marginal probability mass function of X and Y is denoted as p(x)=Pr□{X=x} ,x□X and p(y)=Pr□{Y=y} ,y□Y. Consequently, the continues domain of I(X;Y) is presented in Eq. 7.

$$I(X;Y) = \iint p\,(x, y) log \frac{p(x,y)}{p(x)p(y)} dxdy \tag{7}$$

p(x,y) indicates the joint probability density function of X and Y and p(x) and p(y) indicate the margin of X and Y. Furthermore, the higher value of mutual information presents the intense dependency between X and Y. The calculation of mutual information matrix is depicted by a symmetric m×m matrix (Eq. 8), in which μ_ij is the mutual information between two attributes of dataset (D) (Sefidian & Daneshpour, 2019).

$$\mu(D) = \begin{bmatrix} 1 & \mu_{12} & \cdots & \mu_{1m} \\ \mu_{21} & 1 & \cdots & \mu_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{m1} & \mu_{m2} & \cdots & 1 \end{bmatrix} \tag{8}$$

### 2.7.2. Wrapper method

In the wrapper method which is called the close-loop feature selection method, the feature selector is wrapped around the machine learning method and its performance is evaluated by the accuracy rate or error rate of the classification algorithm. As a result, since the wrapper method relies on the learning algorithm in

selecting the most significant subset of features, the error rate of the classifier should not be conspicuous. Additionally, as in most cases, the wrapper method acquires a higher accuracy rate than the filter method, it is considered an advantageous feature selection method. Nevertheless, its computation is intricate and it is prone to overfitting problems (Zebari et al., 2020). Backward feature elimination as a wrapper method which was applied in this study is described below.

Backward feature elimination

In the wrapper-based backward feature elimination method which is more efficient than the forwarding feature selection method, the importance scores of instances are decided with all data in each iteration and features with the least importance are eliminated. As a result, each round improves the performance of the model. Finally, all remaining features are merged and the round will start again until no more improvement is recognized by removing features (Pan et al., 2009).

### *2.7.3. Embedded method*

The embedded feature selection algorithm is known as the built-in method. In this method to select the most effective subset, the feature evaluator is assisted by the feature selection method which is embedded as a part of the learning algorithm. Artificial Neural Networks (ANNs) and different kinds of decision trees are the most prevalent learning algorithms of embedded methods as well. Furthermore, compared to filter and wrapper methods, embedded feature selectors could achieve more satisfying results, due to their less computational process and their dependency on feature selection via using a learning algorithm (Chen et al., 2020). RF as an embedded feature selection method was applied in this study, due to its beneficial results in solving different problems and its generalization. The procedure of RF as an embedded method is described below.

In the first step, different subsets among features in the training set are produced by the bagging method, and different decision trees are constructed by each selected feature. The growth of the tree relates to the splitting of each node. Therefore, nodes are divided based on the selected features which are arbitrarily selected among candidate subsets of features. Then, the splitting of nodes generates different trees which are grown without pruning and each tree is considered a classifier. In the final stage, all trees are connected which creates a random forest that selects subsets of features among all features in the dataset (Zhou et al., 2016).

## *2.8. Evaluation metrics*

A multi-class confusion matrix is a tool for the performance evaluation of classifiers that are applied to multi-class datasets. Multi-class confusion matrix includes two dimensions indicating actual classes and predicted classes. Table 2 illustrates a multi-class confusion matrix in which A1, A2,…, An are actual classes and Nij is an instance of Ai class which is classified as Aj class. Performance assessment

metrics that were used in this study, are presented below (Eq. 9-11) (Danaei Mehr & Polat, 2019).

Table 2. Multi-Class Confusion Matrix.

| | | Actual | | |
|---|---|---|---|---|
| | | $A_i$ | $\dots A_j \dots$ | $A_n$ |
| Predicted | $A_1$ | $N_{11}$ | $N_{1j}$ | $N_{1n}$ |
| | $A_i$ | $N_{i1}$ | $N_{ij}$ | $N_{in}$ |
| | $A_n$ | $N_{n1}$ | $N_{nj}$ | $N_{nn}$ |

Accuracy: Demonstrates the proportion of correctly predicted instances among all instances.

$$Accuracy = \frac{\sum_{i=1}^{n} N_{ii}}{\sum_{i=1}^{n}\sum_{j=1}^{n} N_{ij}} \tag{9}$$

Recall (Sensitivity): Indicates what proportion of water quality class labels are predicted correctly.

$$Recall = \frac{N_{ii}}{\sum_{k=1}^{n} N_{ik}} \tag{10}$$

Precision: Indicates what proportion of predicted water quality class labels are correctly predicted.

$$Precision = \frac{N_{ii}}{\sum_{k=1}^{n} N_{ki}} \tag{11}$$

## 3. Results and Discussion

In this section, the results obtained by the feature selectors which were used in this study to identify the most effective features in the classification of the water quality were presented. Moreover, the results achieved by applying the WAWQI method for indexing the quality of the water for each instance to supervise the machine learning algorithms in the process of classification are shown. Finally, the performance of each classifier is evaluated and compared.

### *3.1. Water Quality Index*

In this study, the WAWQI method was used to indicate the water quality class of each instance of the dataset which belongs to a particular station in a particular

month and a specific year. As a result, of all 660 instances, 119 instances were classified as Excellent and 139, 95, 59, and 248 were classified as Good, Poor, Very Poor, and Unfit for Consumption, respectively.

### *3.2. Feature selection*

To evaluate the effect of the feature-reduced datasets on the performance of each ML classifier used in this study, Backward Feature Elimination methods were used as a wrapper method, the Mutual Information method was used as a filter method and Random Forest was used as embedded feature selection method. As a result, a dataset of 19 parameters (i. e., Year, Month, Station, Temperature, pH, EC, DO, COD, BOD5, NH4, NO3, NO2, PO4, Cl-, Fe, Mn, Na+, SO4, TDS), Backward Feature Elimination method magnified the importance of 7 parameters in the classification of the WQI whereas Mutual Information method could identify nine (9) parameters as the most prominent parameters. However, the Random Forest Embedded feature selector identified 14 parameters (Table 3).

Table 3. Selected parameters by feature selectors.

| Data-set | Feature selector | Number of Parameters | Selected Parameters |
|---|---|---|---|
| 1 | None | 19 | Year, Month, Station, Temperature, pH, EC, DO, COD, BOD5, $NH_4$, $NO_3$, $NO_2$, $PO_4$, $Cl^-$, Fe, Mn, $Na^+$, $SO_4$, TDs |
| 2 | Backward Feature Elimination | 7 | Station, pH, $NH_4$, $NO_2$, $PO_4$, Mn, $Na^+$ |
| 3 | Mutual Information | 9 | Station, DO, COD, BOD5, $NH_4$, $NO_2$, $PO_4$, Mn, $SO_4$ |
| 4 | Random Forest Embedded | 14 | EC, DO, COD, BOD5, $NH_4$, $NO_3$, $NO_2$, $PO_4$, $Cl^-$, Fe, Mn, $Na^+$, $SO_4$, TDS |

### *3.3. Classification*

In this study, the effect of splitting methods on the performance of each classifier was evaluated as well. K-Fold Cross Validation (KFCV) including 5-FCV and 10-FCV along with percentage splitting including 70:30, 75:25, and 80:20 splitting ratios were used to divide the dataset into training and test datasets.

The performance metrics of classifiers based on subsets of features selected by the Backward Feature Elimination feature selector and the Mutual Information feature selector with KFCV splitting method. In both approaches, the XGBoost

classifier outperformed all other classifiers, especially with the 10-FCV splitting method (Table 4).

Table 4. Performance of the classifiers based on the subset of features selected by the Backward Feature Elimination feature selector and Mutual Information feature selector with KFCV splitting method.

| | **Classifier** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **Backward Feature Elimination feature selector** | | | | **Mutual Information feature selector with KFCV splitting method** | | | |
| | **Logit-Boost** | **Random Forest** | **AdaBoost** | **XGBoost** | **Logit-Boost** | **Random Forest** | **AdaBoost** | **XGBoost** |
| **5-FCV (%)** | | | | | | | | |
| **Accuracy** | 93.6364 | 84.0909 | 51.0606 | 93.3333 | 93.33 | 83.1818 | 51.0606 | 93.4848 |
| **Precision** | 93.60 | 83.70 | 66.00 | 94.00 | 93.40 | 82.50 | 66.00 | 94.00 |
| **Recall** | 93.60 | 84.10 | 51.00 | 93.00 | 93.30 | 83.20 | 51.00 | 94.00 |
| **10-FCV (%)** | | | | | | | | |
| **Accuracy** | 93.0303 | 85.7576 | 59.8484 | **95.6060** | 93.6364 | 82.2727 | 59.8484 | **95.1515** |
| **Precision** | 93.10 | 85.30 | 81.00 | **96.00** | 93.70 | 81.70 | 81.00 | **96.00** |
| **Recall** | 93.00 | 85.80 | 60.00 | **96.00** | 93.60 | 82.30 | 60.00 | **95.00** |
| **70:30 (%)** | | | | | | | | |
| **Accuracy** | 92.9293 | 86.3636 | 58.0808 | 80.3030 | 93.4343 | 83.3333 | 62.6262 | 80.8080 |
| **Precision** | 92.90 | 86.00 | 64.00 | 80.00 | 93.80 | 82.50 | 61.00 | 82.00 |
| **Recall** | 92.90 | 86.40 | 58.00 | 80.00 | 93.40 | 83.30 | 63.00 | 81.00 |
| **75:25 (%)** | | | | | | | | |
| **Accuracy** | 93.3333 | 88.4848 | 60.00 | 81.8181 | 92.7273 | 82.4242 | 62.4242 | 81.2121 |
| **Precision** | 93.40 | 88.80 | 60.00 | 82.00 | 93.00 | 81.90 | 67.00 | 83.00 |
| **Recall** | 93.30 | 88.50 | 60.00 | 82.00 | 92.70 | 82.40 | 62.00 | 81.00 |
| **80:20 (%)** | | | | | | | | |
| **Accuracy** | 92.4242 | 84.0909 | 63.6363 | 84.8484 | 91.6667 | 82.5758 | 71.9696 | 78.7878 |
| **Precision** | 92.80 | 84.80 | 63.00 | 85.00 | 92.40 | 82.40 | 70.00 | 82.00 |
| **Recall** | 92.40 | 84.10 | 64.00 | 85.00 | 91.70 | 82.60 | 72.00 | 79.00 |

The performance metrics of classifiers based on all 19 features selected by the Random Forest Embedded feature selector and initially chosen as the water quality parameters of interest. In both approaches, the XGBoost classifier outperformed all other classifiers, especially with the 10-FCV splitting method (Table 5).

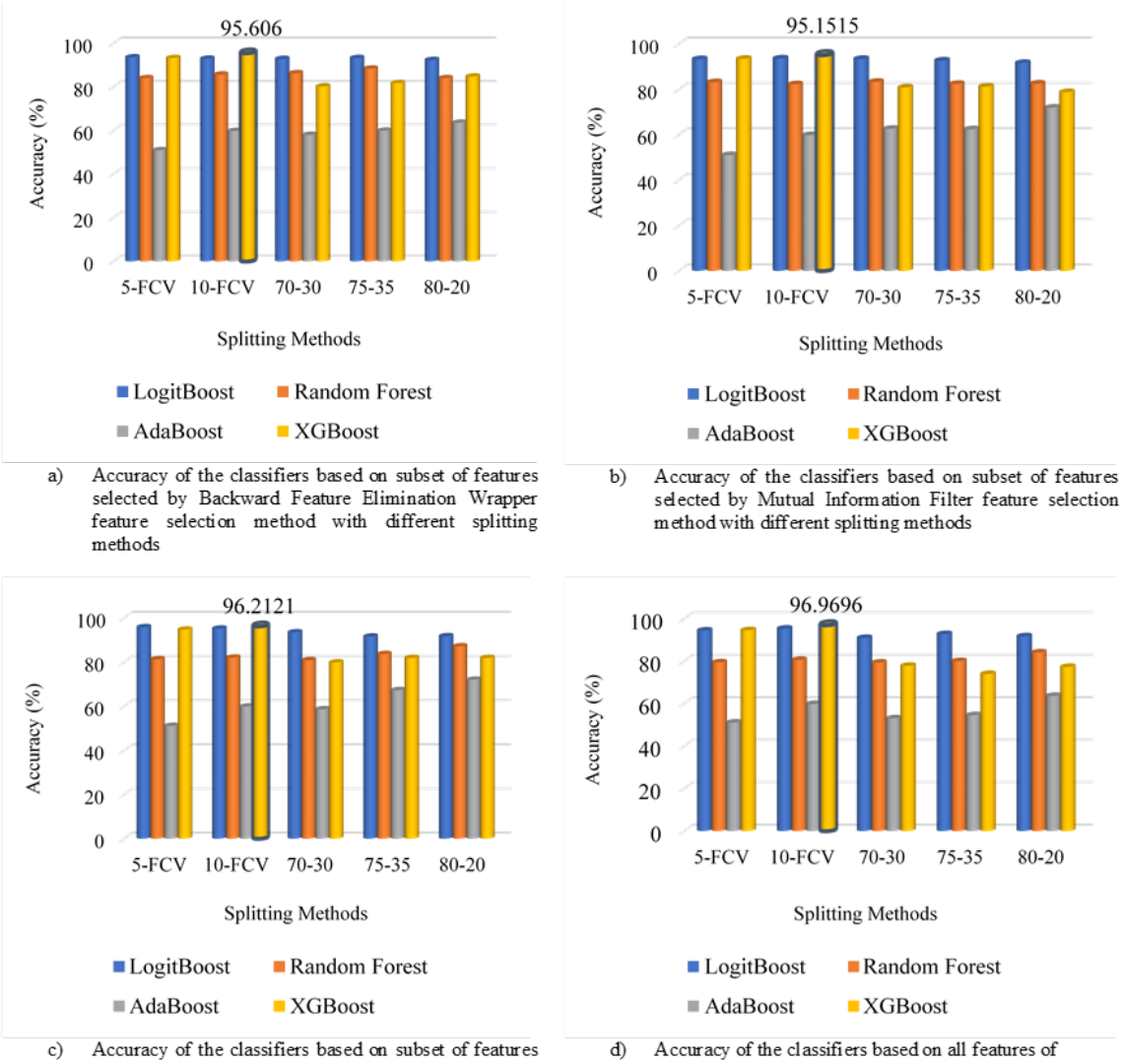
The main objective of similar studies in the literature is to suggest a particular machine learning model that has achieved promising results based on specific parameters of interest with a particular splitting method in the classification of water quality index. The purpose of developing a machine learning model in the classification of water quality is to evade time-consuming computations to classify the water quality by the WQI method, however, it is essential to identify the water quality class by a particular WQI model based on the purpose of the classification, in the beginning, to supervise the machine learning algorithm in the process of learning.

Table 5. Performance of the classifiers based on the subset of features selected by Random Forest Embedded feature selector and all 19 features selected of interest.

| | Classifier | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Random Forest Embedded feature selector | | | | All the features of interest | | | |
| | Logit-Boost | Random Forest | AdaBoost | XGBoost | Logit-Boost | Random Forest | Ada-Boost | XGBoost |
| **5-FCV (%)** | | | | | | | | |
| **Accuracy** | 95.7576 | 81.2121 | 51.06 | 94.6969 | 94.3939 | 79.3939 | 51.0606 | 94.5454 |
| **Precision** | 95.90 | 80.40 | 66.00 | 95.00 | 94.40 | 78.90 | 66.00 | 95.00 |
| **Recall** | 95.80 | 81.20 | 51.00 | 95.00 | 94.40 | 79.40 | 51.00 | 95.00 |
| **10-FCV (%)** | | | | | | | | |
| **Accuracy** | 95.1515 | 81.9697 | 59.8484 | **96.2121** | 95.303 | 80.6061 | 59.8484 | **96.9696** |
| **Precision** | 95.20 | 81.30 | 81.00 | **97.00** | 95.30 | 79.80 | 81.00 | **97.00** |
| **Recall** | 95.20 | 81.20 | 60.00 | **96.00** | 95.30 | 80.60 | 60.00 | **97.00** |
| **70:30 (%)** | | | | | | | | |
| **Accuracy** | 93.4343 | 80.8081 | 58.5858 | 79.7979 | 90.9091 | 79.2929 | 53.0303 | 77.7777 |
| **Precision** | 94.40 | 79.10 | 61.00 | 79.00 | 92.60 | 77.80 | 56.00 | 69.00 |
| **Recall** | 93.40 | 80.80 | 59.00 | 80.00 | 90.90 | 79.30 | 53.00 | 72.00 |
| **75:25 (%)** | | | | | | | | |
| **Accuracy** | 91.5152 | 83.6364 | 67.2727 | 81.8181 | 92.7273 | 80.00 | 54.5454 | 73.9393 |
| **Precision** | 92.90 | 83.50 | 65.00 | 81.00 | 93.60 | 80.40 | 65.00 | 75.00 |
| **Recall** | 91.50 | 83.60 | 67.00 | 82.00 | 92.70 | 80.00 | 55.00 | 74.00 |
| **80:20 (%)** | | | | | | | | |
| **Accuracy** | 91.6667 | 87.1212 | 71.9696 | 81.8181 | 91.6667 | 84.0909 | 63.6363 | 77.2727 |
| **Precision** | 93.80 | 87.20 | 69.00 | 81.00 | 93.20 | 83.70 | 62.00 | 75.00 |
| **Recall** | 91.70 | 87.10 | 72.00 | 82.00 | 91.70 | 84.10 | 64.00 | 74.00 |

In this study, the XGBoost classifier achieved the most promising performance over other classifiers with the 10-FCV splitting method in all subsets of features identified by different feature selectors as well as all features of interest which were chosen as the parameters of interest at the beginning of the study (Figure 3). The performance metrics (i.e., accuracy, precision, recall) which are tabulated previously were calculated based on confusion matrix of the classifiers which indicate the number of correctly and incorrectly classified instances by each classifier.

The confusion matrix (Figure 4a) displays the XGBoost classifier's performance with 10-Fold Cross Validation (10-FCV) on a dataset suggested by the Backward Feature Elimination feature selector. Notably, the classifier excelled in accurately classifying the 'Excellent' and 'Unfit for Consumption' labels, achieving maximum accuracies of 99.15% and 100%, respectively. However, its accuracy notably dropped to 72.88% for the 'Very Poor' class label. While achieving 97.84% accuracy for the 'Good' label, the overall classification accuracy of the XGBoost classifier, utilizing attributes suggested by the Backward Feature Elimination Wrapper, stood at 95.606%, correctly classifying 631 out of 660 instances.




a) Accuracy of the classifiers based on subset of features selected by Backward Feature Elimination Wrapper feature selection method with different splitting methods

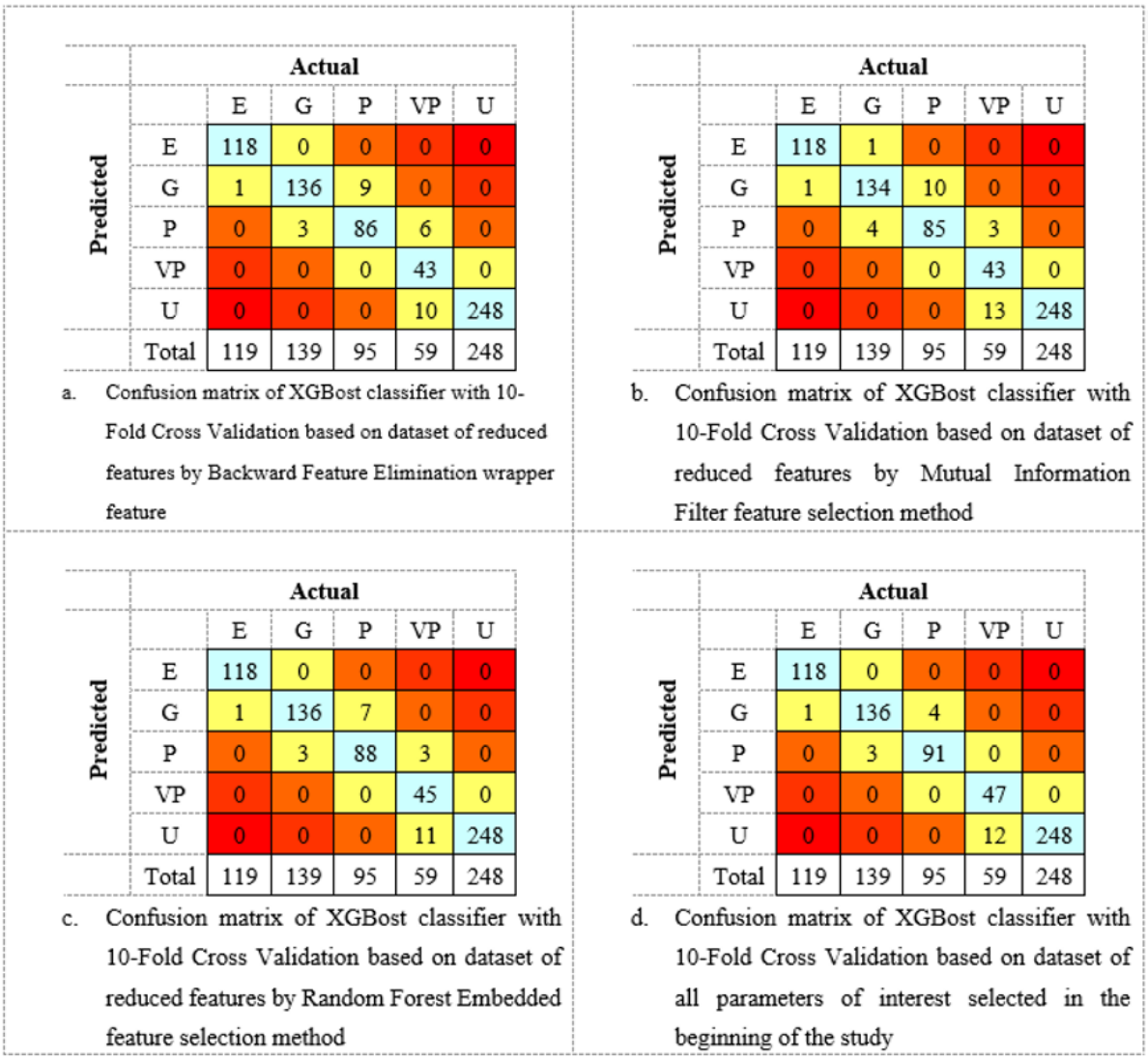
b) Accuracy of the classifiers based on subset of features selected by Mutual Information Filter feature selection method with different splitting methods

c) Accuracy of the classifiers based on subset of features selected by Random Forest Embedded feature selection method with different splitting methods

d) Accuracy of the classifiers based on all features of interest selected in the beginning of the study with different splitting methods

Figure 3. The accuracy comparison of the classifiers based on a different subset of features with different dataset splitting methods.

In contrast, the XGBoost classifier exhibited reduced performance in categorizing 'Good' and 'Poor' labels based on parameters suggested by the Mutual Information Filter method (Figure 4b) compared to its accuracy using the subset of features chosen by the Backward Feature Elimination Wrapper. However, the classifier's accuracy remained consistent when classifying 'Excellent,' 'Very Poor,' and 'Unfit for Consumption' labels using features selected by the Mutual Information Filter method, aligning with the accuracy achieved using the Backward Elimination Wrapper-selected features.



| Predicted \ Actual | E | G | P | VP | U |
|---|---|---|---|---|---|
| E | 118 | 0 | 0 | 0 | 0 |
| G | 1 | 136 | 9 | 0 | 0 |
| P | 0 | 3 | 86 | 6 | 0 |
| VP | 0 | 0 | 0 | 43 | 0 |
| U | 0 | 0 | 0 | 10 | 248 |
| Total | 119 | 139 | 95 | 59 | 248 |

a. Confusion matrix of XGBost classifier with 10-Fold Cross Validation based on dataset of reduced features by Backward Feature Elimination wrapper feature

| Predicted \ Actual | E | G | P | VP | U |
|---|---|---|---|---|---|
| E | 118 | 1 | 0 | 0 | 0 |
| G | 1 | 134 | 10 | 0 | 0 |
| P | 0 | 4 | 85 | 3 | 0 |
| VP | 0 | 0 | 0 | 43 | 0 |
| U | 0 | 0 | 0 | 13 | 248 |
| Total | 119 | 139 | 95 | 59 | 248 |

b. Confusion matrix of XGBost classifier with 10-Fold Cross Validation based on dataset of reduced features by Mutual Information Filter feature selection method

| Predicted \ Actual | E | G | P | VP | U |
|---|---|---|---|---|---|
| E | 118 | 0 | 0 | 0 | 0 |
| G | 1 | 136 | 7 | 0 | 0 |
| P | 0 | 3 | 88 | 3 | 0 |
| VP | 0 | 0 | 0 | 45 | 0 |
| U | 0 | 0 | 0 | 11 | 248 |
| Total | 119 | 139 | 95 | 59 | 248 |

c. Confusion matrix of XGBost classifier with 10-Fold Cross Validation based on dataset of reduced features by Random Forest Embedded feature selection method

| Predicted \ Actual | E | G | P | VP | U |
|---|---|---|---|---|---|
| E | 118 | 0 | 0 | 0 | 0 |
| G | 1 | 136 | 4 | 0 | 0 |
| P | 0 | 3 | 91 | 0 | 0 |
| VP | 0 | 0 | 0 | 47 | 0 |
| U | 0 | 0 | 0 | 12 | 248 |
| Total | 119 | 139 | 95 | 59 | 248 |

d. Confusion matrix of XGBost classifier with 10-Fold Cross Validation based on dataset of all parameters of interest selected in the beginning of the study

Figure 4. Confusion matrix of the XGBoost classifier based on a different subset of features (E: Excellent; G: Good; P: Poor; VP: Very Poor; U: Unfit for Consumption).

The classifier performed best when using all features, showcasing the highest accuracy for 'Good,' 'Poor,' and 'Very Poor' class labels (Figure 4d). However, the Random Forest Embedded method produced the best accuracy for 'Poor' and 'Very Poor' labels, surpassing other feature subsets discussed earlier (Figure 4c). Yet, the classifier had the most classification mistakes with reduced feature combinations for these classes. For higher water quality accuracy, involving more parameters is beneficial as machine learning algorithms learn better from various parameters. However, using fewer parameters is rational to reduce computational costs and justify monitoring basin water quality economically. Hence, while XGBoost excelled with all features initially, evaluating its performance with different feature subsets used in this study is wise. Despite achieving the second-best performance with 14 parameters, including eliminated ones without additional processing costs would've

been logical. Considering the classifier's performance with 7 parameters selected by Backward Feature Elimination, sacrificing a small performance percentage might be acceptable, compensating for the lack with larger datasets during training. Nonetheless, reducing parameters by 12 could enhance model cost-effectiveness.

As a result, when our study and the studies conducted in Central Asia (Cheema et al., 2024) are examined, they include similar approaches in terms of water management, data analytics and modeling approaches, but our study focuses on water quality classification, while the other one (Cheema et al., 2024) focuses on the prediction of water-related health risks.

Similar to the Büyük Menderes Basin, Central Asian regions such as the Syr Darya River (Bissenbayeva et al., 2020) and Amu Darya basins (Leng et al., 2021; Satybaldiyev et al., 2023) in Central Asia face serious water quality degradation due to agricultural runoff and industrial discharge. The XGBoost model proposed in this study can be adapted for water quality monitoring in these regions, considering the common challenges such as high nutrient load and salinity.

## 4. Conclusion

The primary objective of this study was to suggest an effective ensemble machine learning model for the classification of the water quality of Büyük Menderes Basin in Turkey. To supervise the machine learning algorithms in the process of learning WAWQI was used to classify the instances of the dataset of 19 parameters into five (5) major class labels which are suggested by the WAWQI function for public consumption. Moreover, since decisions that are made by machine learning algorithms are profoundly affected by the dimension of the dataset, three feature selection methods were used to reduce the number of parameters that were selected as the parameters of interest at the beginning of this study. Consequently, the results achieved by either of the classifiers based on the datasets generated by each of the feature selectors as well as the main dataset consisting of all the features were evaluated to propose an effective machine learning algorithm in the classification of the water quality of the basin. Results demonstrated that the performance of the classifiers was affected significantly by the number of parameters involved in the classification process and the XGBoost classifier exhibited brilliant merit over the other three classifiers when either of the datasets was used in the classification process. The best performance achieved by this classifier was when all the features which were selected at the beginning of the study were involved. However, to evade extra computations to make this algorithm cost-effective, it is suggested to use the subset of features selected by the Backward Feature Elimination feature selector since the results didn't show a significant difference from the results achieved by this

classifier when all the parameters of interest which were chosen at the beginning of the study were involved in the process of classification. Yet, the number of features decreased by 12 parameters which can make the process cost-effective. The methodology presented in this study is particularly relevant for Central Asia, where water quality monitoring is critical due to transboundary water disputes and limited freshwater resources. For instance, applying this model to the Amu Darya Basin could help optimize monitoring efforts by identifying key parameters (e.g., salinity, nitrates) with reduced computational costs.

## References

Arabgol, R., Sartaj, M., & Asghari, K. (2016). Predicting Nitrate Concentration and Its Spatial Distribution in Groundwater Resources Using Support Vector Machines (SVMs) Model. Environmental Modeling & Assessment, 21:71–82. https://doi.org/10.1007/s10666-015-9468-0

Arora, N., & Kaur, P. D. (2020). A Bolasso based consistent feature selection enabled random forest classification algorithm: An application to credit risk assessment. Applied Soft Computing Journal. 86:105936. https://doi.org/10.1016/j.asoc.2019.105936

Bhati, B. S., Chugh, G., Al-Turjman, F., & Bhati, N. S. (2021). An improved ensemble based intrusion detection technique using XGBoost. Transactions on Emerging Telecommunications Technologies, 32: e4076. https://doi.org/10.1002/ett.4076

Bissenbayeva, S., Abuduwaili, J., Issanova, G., & Samarkhanov, K. (2020). Characteristics and causes of changes in water quality in the Syr Darya River, Kazakhstan. Water Resources, 47, 904-912. https://doi.org/10.1134/S009780782005019X

Bouamar, M., & Ladjal, M. (2007). Evaluation of the performances of ANN and SVM techniques used in water quality classification. In the 14th IEEE International Conference on Electronics, Circuits and Systems, IEEE, Marrakech, Morocco. https://doi.org/10.1109/ICECS.2007.4511173

Breiman, L. (2001). Random Forests. Machine Learning, 45:5–32. https://doi.org/10.1023/A:1010933404324

Bui, D. T., Khosravi, K., Tiefenbacher, J., Nguyen, H., & Kazakis, N. (2020). Improving prediction of water quality indices using novel hybrid machine-learning algorithms. Science of the Total Environment. 721:137612. https://doi.org/10.1016/j.scitotenv.2020.137612

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In the 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining, San Francisco, California. https://doi.org/10.1145/2939672.2939785

Chen, C. W., Tsai, Y. H., Chang, F. R., & Lin, W. C. (2020). Ensemble feature selection in medical datasets: Combining filter, wrapper, and embedded feature selection results. Expert Systems, 37:e12553. https://doi.org/10.1111/exsy.12553

Cheema, M. A., Hanif, M., Albalawi, O., Mahmoud, E. E., & Nabi, M. (2024). Evaluating water-related health risks in East and Central Asian Islamic Nations using predictive models (2020–2030). Scientific Reports, 14(1), 16837

Danades, A., Pratama, D., & Anggraini, D. (2016). Comparison of Accuracy Level K-Nearest Neighbor Algorithm and Support Vector Machine Algorithm in Classification Water Quality Status. In the 6th International Conference on System Engineering and Technology (ICSET), IEEE, Bandung, Indonesia. https://doi.org/10.1109/ICSEngT.2016.7849638

Danaei Mehr, H., & Polat, H. (2019). Human Activity Recognition in Smart Home With Deep Learning Approach. In the 7th International Istanbul Smart Grids and Cities Congress and Fair (ICSG), IEEE, Istanbul, Turkey. https://doi.org/10.1109/SGCF.2019.8782290

Dezfooli, D., Moghari, S. M. H., Ebrahimi, K., & Araghinejad, S. (2018). Classification of water quality status based on minimum quality parameters: application of machine learning techniques. Modeling Earth Systems and Environment, 4:311–324. https://doi.org/10.1007/s40808-017-0406-9

Dohare, D., Deshpande, S., & Kotiya, A. (2014). Analysis of Ground Water Quality Parameters: A Review. Research Journal of Engineering Sciences, 3 (5):26-31. ISSN: 2278 – 9472

Dong, X., Yu, Z., Cao, W., Shi, Y., & Ma, Q. (2020). A survey on ensemble learning. Frontiers of Computer Science. 14 (2):241–258. https://doi.org/10.1007/s11704-019-8208-z

Freund, Y., & Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. Journal of Computer and System Sciences. 55:119-139. https://doi.org/10.1006/jcss.1997.1504

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). The Annals of Statistics, 28 (2):337-407. https://doi.org/10.1214/aos/1016218223

General Directorate of Environmental Management (2016) Büyük Menderes Basin Pollution Prevention Action Plan (Turkish). Ministry of Environment and Urbanization, Ankara, Turkey

Ighalo, J. O., Adeniyi, A. G., & Marques, G. (2020). Application of linear regression algorithm and stochastic gradient descent in a machine-learning environment for predicting biomass higher heating value. Biofuels, Bioproducts, and Biorefining, 14:1286–1295, 2020. https://doi.org/10.1002/bbb.2140

Khaire, U. M., & Dhanalakshmi, R. (2019). Stability of feature selection algorithm: A review. Journal of King Saud University – Computer and Information Sciences. 34(4): 1060-1073. https://doi.org/10.1016/j.jksuci.2019.06.012

Kumar, Z. M., & Manjula, R. (2012). Regression model approach to predict missing values in the Excel sheet databases. International Journal of Computer Science & Engineering Technology (IJCSET). 3 (4):130-135. ISSN: 2229-3345

Lap, B. Q., Du Nguyen, H., Hang, P. T., Phi, N. Q., Hoang, V. T., Linh, P. G., & Hang, B. T. T. (2023). Predicting water quality index (WQI) by feature selection and machine learning: a case study of An Kim Hai irrigation system. Ecological Informatics, 74, 101991. https://doi.org/10.1016/j.ecoinf.2023.101991

Liu, Q., Wang, X., Huang, X., & Yin, X. (2020). Prediction model of rock mass class using classification and regression tree integrated AdaBoost algorithm based on TBM driving data. Tunnelling and Underground Space Technology. 106:103595. https://doi.org/10.1016/J.TUST.2020.103595

Leng, P., Zhang, Q., Li, F., Kulmatov, R., Wang, G., Qiao, Y., ... & Khasanov, S. (2021). Agricultural impacts drive longitudinal variations of riverine water quality of the Aral Sea basin (Amu Darya and Syr Darya Rivers), Central Asia. Environmental Pollution, 284, 117405. https://doi.org/10.1016/j.envpol.2021.117405

Mădălina, P., & Gabriela, B. I. (2014). Water Quality Index – An Instrument for Water Resources Management. Aerul şi Apa: Componente ale Mediului, 2014:391 – 398

Modaresi, F., & Araghinejad, S. (2014). A Comparative Assessment of Support Vector Machines, Probabilistic Neural Networks, and K-Nearest Neighbor Algorithms for Water Quality Classification. Water Resources Management, 28:4095–4111. https://doi.org/10.1007/s11269-014-0730-z

Motevalli, A., Naghibi, S. A., Hashemi, H., & Berndtsson, R. (2019). Inverse method using boosted regression tree and k-nearest neighbor to quantify effects of point and non-point source nitrate pollution in groundwater. Journal of Cleaner Production, 228:1248-1263. https://doi.org/10.1016/j.jclepro.2019.04.293

Muhammad, S. Y., Makhtar, M., Rozaimee, A., Aziz, A. A., & Jamal, A. A. (2015). Classification Model for Water Quality using Machine Learning Techniques. International Journal of Software Engineering and Its Applications. 9 (6):45-52

Ostad-Ali-Askari, K., Shayannejad, M., & Ghorbanizadeh-Kharazi, H. (2017). Artificial neural network for modeling nitrate pollution of groundwater in marginal area of Zayandeh-rood River, Isfahan, Iran. KSCE Journal of Civil Engineering, 21:134–140. https://doi.org/10.1007/s12205-016-0572-8

Pan, F., Converse, T., Ahn, D., Salvetti, F., & Donato, G. (2009). Feature Selection for Ranking using Boosted Trees. In Proceedings of the 18th ACM conference on Information and knowledge management, Hong Kong, China. https://doi.org/10.1145/1645953.1646292

Radhakrishnan, N., & Pillai, A. S. (2020). Comparison of Water Quality Classification Models using Machine Learning. In the 5th International Conference on Communication and Electronics Systems (ICCES), IEEE, Coimbatore, India. https://doi.org/10.1109/ICCES48766.2020.9137903

Rozemeijer, J. C., & Broers, H. P. (2007). The groundwater contribution to surface water contamination in a region with intensive agricultural land use (Noord-Brabant, The Netherlands). Environmental Pollution, 148:695-706. https://doi.org/10.1016/j.envpol.2007.01.028

Saghebian, S. M., Sattari, M. T., Mirabbasi, R., & Pal, M. (2014). Ground water quality classification by decision tree method in Ardebil region, Iran. Arabian Journal of Geosciences, 7:4767–4777. https://doi.org/10.1007/s12517-013-1042-y

Satybaldiyev, B., Ismailov, B., Nurpeisov, N., Kenges, K., Snow, D. D., Malakar, A., ... & Uralbekov, B. (2023). Downstream hydrochemistry and irrigation water quality of the syr darya, aral sea basin, south kazakhstan. Water Supply, 23(5), 2119-2134. https://doi.org/10.2166/ws.2023.114

Sefidian, A. M., & Daneshpour, N. (2019). Missing value imputation using a novel grey based fuzzy c-means, mutual information based feature selection, and regression model. Expert Systems With Applications, 115:68–94. https://doi.org/10.1016/j.eswa.2018.07.057

Siriwardhana, K. D., Jayaneththi, D. I., Herath, R. D., Makumbura, R. K., Jayasinghe, H., Gunathilake, M. B., Azamathulla, H.M., Tota-Maharaj, K., & Rathnayake, U. (2023). A Simplified Equation for Calculating the Water Quality Index (WQI), Kalu River, Sri Lanka. Sustainability, 15(15), 12012. https://doi.org/10.3390/su151512012

Sim, J., Lee, J. S., & Kwon, O. (2015). Missing Values and Optimal Selection of an Imputation Method and Classification Algorithm to Improve the Accuracy of Ubiquitous Computing Applications. Mathematical Problems in Engineering, 2015:538613. https://doi.org/10.1155/2015/538613

Tehrany, M. S., Jones, S., Shabani, F., Martínez-Álvarez, F., & Bui, D. T. (2019). A novel ensemble modeling approach for the spatial prediction of tropical forest fire susceptibility using LogitBoost machine learning classifier and multi-source geospatial data. Theoretical and Applied Climatology, 137:637–653. https://doi.org/10.1007/s00704-018-2628-9

Turkish Standard Institute (TSE), (2005) TURKISH STANDARD (TS-266): Water Intended for Human Consumption. Turkish Standards Institution, Ankara, Turkey

Tyagi, S., Sharma, B., Singh, P., & Dobhal, R. (2013). Water Quality Assessment in Terms of Water Quality Index. American Journal of Water Resources, 1 (3): 34-38. https://doi.org/10.12691/ajwr-1-3-3

Uddin, M. d. G., Nash, S., & Olbert, A. I. (2021). A review of water quality index models and their use for assessing surface water quality. Ecological Indicators, 122:107218. https://doi.org/10.1016/j.ecolind.2020.107218

Uyun, S., & Sulistyowati, E. (2020). Feature selection for multiple water quality status: integrated bootstrapping and SMOTE approach in imbalance classes. International Journal of Electrical and Computer Engineering (IJECE), 10 (4):4331-4339. http://doi.org/10.11591/ijece.v10i4.pp4331-4339

Varol, M., & Şen, B. (2012). Assessment of nutrient and heavy metal contamination in surface water and sediments of the upper Tigris River, Turkey. Catena, 92:1–10. https://doi.org/10.1016/j.catena.2011.11.011

World Health Organization (WHO) (2006) Guidelines for Drinking-water Quality: incorporating first addendum. Vol. 1, Recommendations. – 3rd ed. Geneva, Switzerland. ISBN: 92 4 154696 4

Yozgatligil, C., Aslan, S., Iyigun, C., & Batmaz, I. (2013). Comparison of missing value imputation methods in time series: the case of Turkish meteorological data. Theoretical and Applied Climatology, 112:143–167. https://doi.org/10.1007/s00704-012-0723-x

Zebari, R. R., Abdulazeez, A. M., Zeebaree, D. Q., Zebari, D. A., & Saeed, J. N. (2020). A Comprehensive Review of Dimensionality Reduction Techniques for Feature Selection and Feature Extraction. Journal of Applied Science and Technology Trends, 1 (2):56-70. https://doi.org/10.38094/jastt1224

Zhou, Q., Zhou, H., & Li, T. (2016). Cost-sensitive feature selection using random forest: Selecting low-cost subsets of informative features. Knowledge-Based Systems, 95:1–11. https://doi.org/10.1016/j.knosys.2015.11.010