

Performance Evaluation of Machine Learning Models with Ensemble Learning approach in Classification of Water Quality Indices Based on Different Subset of Features

FARID HASSANBAKI GARABAGHI (✉ fgarabaghi@gmail.com)

Gazi Üniversitesi: Gazi Üniversitesi <https://orcid.org/0000-0003-0244-0360>

Semra Benzer

Gazi University: Gazi Üniversitesi

Recep Benzer

Gazi University: Gazi Üniversitesi

Research Article

Keywords: Water Quality Index, Classification, Machine learning, Feature selection, Ensemble Learning methods

Posted Date: November 3rd, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-876980/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Performance Evaluation of Machine Learning Models with Ensemble Learning approach in Classification of Water Quality Indices Based on Different Subset of Features

Farid Hassanbaki Garabaghi^{1*}, Semra Benzer², Recep Benzer³

¹Gazi University, Graduate School of Natural and Applied Sciences, Teknikokullar, 06500 Ankara, Turkey

²Gazi University, Gazi Education Faculty, Teknikokullar, 06500 Ankara, Turkey

³Gazi University, Information Institute, Kavaklıdere, 06560 Ankara, Turkey

*Corresponding author: Farid Hassanbaki Garabaghi (fgarabaghi@gmail.com)

Abstract

Since fresh water resources in form of groundwaters which are the most available water resources for human consumption are extremely limited and due to uncontrolled human activities are prone to contamination, it's of a great importance to constantly monitor the quality of the ground fresh water resources to provide a sustainable drinking water for people as well as protecting the ecosystem. One tool for modeling the water quality of a basin is Water Quality Index (WQI). However, calculating WQI is complicated and time- consuming, therefore, today, scientists are being inclined to propose simpler ways for modeling the quality of the water resources such as machine learning algorithms. In this study the performance of four machine learning algorithms with ensemble learning approach were evaluated to propose a classification model (classifier) with highest performance. Moreover, to identify the most important water quality parameters in the classification process, three feature selection methods with machine learning approach were applied. As a result, among four classifiers, XGBoost showed outstanding performance, with the accuracy of 96.9696% when all the parameters

of interest were involved in the classification process. However, in order to make the model cost-effective it is suggested to conduct the classification with optimum parameters which in this case, for the dataset which was used in this study XGBoost classifier is suggested as the best classifier with the maximum accuracy of 95.606% with 10-Fold Cross Validation when seven parameters which were identified by Backward Feature Elimination Feature selector were involved in the classification process.

Keywords: Water Quality Index. Classification. Machine learning. Feature selection. Ensemble Learning methods.

1. Introduction

In spite of the fact that roughly 90% of all fresh water resources are groundwater, only minuscule portion of it is accessible to be used by human as drinking water (Arabgol et al. 2016; Ostad-Ali-Askari et al. 2017). Indirectly related to rapid population growth, today, anthropogenic activities such as agricultural and industrial activities 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 a paramount importance from the human-wellbeing point of view to constantly monitor the quality of the surface water, but it also play a pivotal role in preserving the aquatic ecosystem as 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 and Broers 2007; Varol and Şen 2012). Hydrogen Sulfide as a result of reduction of sulfate by anaerobic bacterial activities could be considered as the indicator of eutrophication in

surface water reservoirs as well. What's more, high concentrations of sulfate in drinking water can cause health problem (WHO 2006; EPA 2021). Therefore, sulfur and its derivatives must be considered as an important water quality parameter. Chloride in addition is another crucial water quality parameter due to its toxic effects in aquatic environments. Chloride can first and foremost be found in the environment as a free anion; therefore, it is considered as the chief component of the salt formation such as (Elphick et al. 2011) sodium chloride, potassium chloride, magnesium chloride, etc. Hence, it can be concluded that chloride is associated with dissolved solids in aquatic systems. Presence of iron and manganese - as two of the most abundant elements in the earth's crust - in surface water is an inevitable fact. Although presence of these two elements of not a serious health concern, higher concentrations of them result in unwanted taste and color of the water as well as clogging of the pumps in association with sediment formation by anaerobic iron-manganese bacterial activities (Wong 1984; WHO 2006; Dalai et al. 2015).

Above, only a small portion of water quality parameters and the problems followed by the presence of each and any of them were described. Therefore, consistent monitoring and assessment of water quality parameters depending on the probable contamination sources in a basin should be one of the governments' responsibilities to firstly sustain the aquatic life cycle, moreover, to provide safe and sustainable drinking water for the people.

One tool that has aided the 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 individually is Water Quality Index (WQI). One great advantage of WQI is that it is able to unify the massive data into one value which is simpler, more understandable, interpretable and consequently logical (Tyagi et al. 2013; Uddin et al. 2021).

Besides advantages, however, interpretation of water quality based on WQI has many drawbacks such as discrepancy between the WQI methods as there are numerous and various

equations which are used worldwide by different agencies (Tyagi et al. 2013; Bui et al. 2020). Moreover, the method is quite time-consuming and complex (Bui et al. 2020). Hence, due to these disadvantages, researchers today, have inclined their tendency to use simpler, more reliable and effective methods in classification and modeling water quality.

Flourishing advanced technologies specifically application of artificial intelligences (AI) in variety of research lines, today, environmental scientists tend to evaluate the accuracy and reliability of different machine methods in their studies particularly in classification of water quality in order to firstly evade complicated calculations to make rational decisions and secondly expedite the process of classification as well as prediction as these methods have recorded promising results in recent years in this regard (Modaresi and Araghinejad 2014; Saghebian et al. 2014; Danades et al. 2016; Radhakrishnan and Pillai 2020).

Despite satisfying results, the accuracy of the classifiers can be affected by high dimensional data as well as the noise, due to 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 classification of water quality. The chief disadvantage of the both models was reported to be their sensitiveness to the noise. Therefore, it can be concluded that where the data is noisy, these models are not considered as effective methods in the process of prediction as well as classification of water quality. 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 and Sulistyowati 2020).

Although there are some studies in the literature with focus on the selection of the most prominent attributes in classification of water quality, they are not that of machine learning approaches (Dezfooli et al. 2018; Bui et al. 2020).

Yet, there are some studies in which feature selection based on machine learning approach have been conducted. Muhammad et al. (2015) evaluated the performance of five machine learning

algorithms in 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 CfsSubsetEval wrapper feature selection method was used to extract the most important attributes in the classification process.

The objective of this study is to investigate the performance of different ensemble machine learning models in 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 investigation of the effect of reduced water quality attributes by different feature selection methods on the performance of the machine learning algorithms in classification of water quality indices is the secondary aim of this study. What's more, the effect of different dataset splitting methods on the performance of the machine learning algorithms was also evaluated in this study.

2. Materials and methods

In this section different classification algorithms including LogitBoost, Random Forest, AdaBoost, XGBoost whose performances were evaluated in the classification of the water quality indices were discussed. Moreover, different feature selection methods which were used in this study to extract the most significant parameters in the classification process were described in this section. The overall workflow of the study is illustrated in Fig. 1.

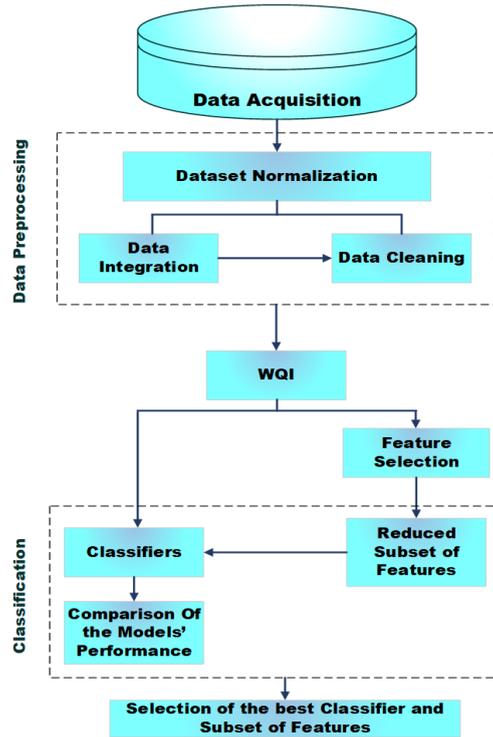


Fig. 1 Schematic workflow of the study

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 Eagean Sea. 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 as the main contamination sources of the basin. Almost 79% of the waters in the basin are used for agriculture and 21% is used by the industries as well as dwellings (General Directorate of Environmental Management, Turkey, 2016). Since the Büyük Menderes Basin is one of the mostly contaminated basins in Turkey due to the uncontrolled discharges of the industries to the Büyük Menderes River, which is the artery of the basin, and its streams, it is essential for the survival of the basin to be monitored regularly in terms of water quality and that's why this basin was selected as the study area in this research (Fig. 2).

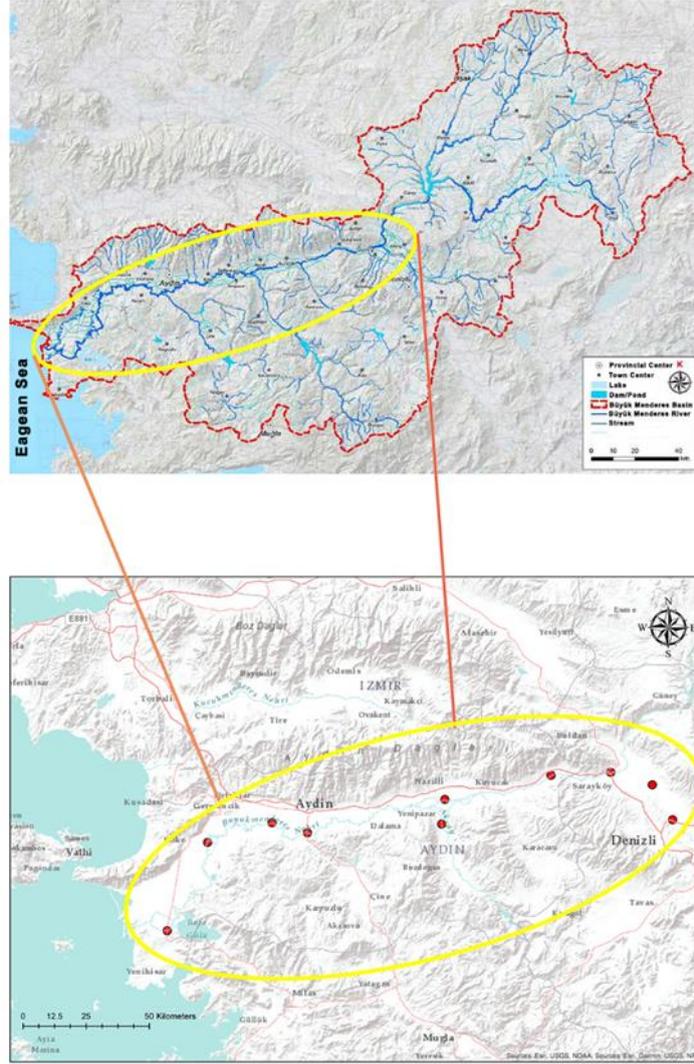


Fig. 2 Büyük Menderes Basin and the selected stations

2.2. Data acquisition

For the purpose of 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 from 2004-2014 were selected to represent the downstream of the basin in this study due to the abundance of the measured parameters and the frequency of the observations since some stations were placed in inaccessible places and in some seasons with harsh climate it was difficult to reach the station for sampling. Therefore, in many of the stations only the most important water quality

parameters were measured regularly according to the demand for the sampling and measurement due to the possible contamination in the area.

2.3. Dataset formation

After data acquisition, it was essential to create a dataset of the necessary parameters for the purpose of this study. This is an essential part of data preprocessing called data integration which is used for merging the data from different sources to make a unified dataset to serve the purpose of any studies (Kumar and 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 of 660 instances. The schematic example of the unified dataset is illustrated in Table 1.

Table 1 Schematic example of the integrated dataset

Observation Information			Water Quality Parameters																
Year	Month	Station	T	pH	EC	DO	COD	BOD5	NH ₄	NO ₃	NO ₂	PO ₄	Cl ⁻	Fe	Mn	Na ⁺	SO ₄	TDS	
2004	2	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2005	4	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2006	6	3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
.
.
.
2014	12	10	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

2.4. Missing value imputation

Encountering missing value in environmental studies is an unavoidable fact due to the impossibility of through going observation of continual processes. Moreover, a corrupt dataset with missing values can end up to incorrect conclusion as a result of data inconsistency in the process of data analyzing. Although there are many missing value imputation methods such as 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 which are used for predicting the value of a target parameter, 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 prediction of missing values where more than 10% of the whole dataset are missing values. Linear regression algorithm simply performs 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 LR model for missing value imputation is shown in Fig. 3 in which, obviously, 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 to the training dataset to make a whole dataset without missing values.

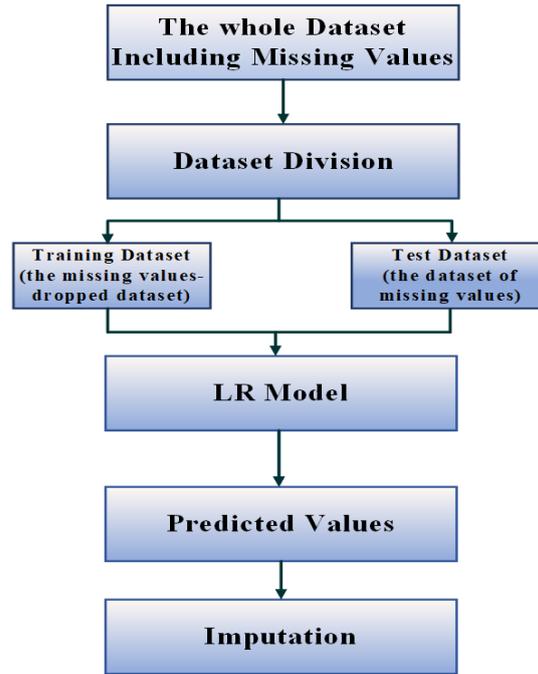


Fig. 3 The schematic process of the missing value imputation by LR machine learning algorithm

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 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 firstly. 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 and 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 classes 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.

$$WAWQI = \sum Q_i W_i / \sum W_i \quad (2)$$

Where,

Q_i is the quality rating of i^{th} parameter and W_i is the unit weight which are calculated by equations 3 and 4, respectively.

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

Where,

V_i is the estimated concentration of the i^{th} parameter and V_0 is the ideal value of the i^{th} parameter.

$$W_i = K / S_i \quad (4)$$

Where,

S_i is the recommended standard value of the i^{th} parameter and K is the proportionality constant which is calculated through Eq. 5.

$$K = \frac{1}{\sum 1/S_i} \quad (5)$$

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 2.

Table 2 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 in order to make 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 (RF), XGBoost, AdaBoost and LogitBoost Classifiers were used in this study. Random Forest (Breiman 2001) which is basically the combination of decision trees takes the advantage of using bootstrap aggregation and bagging in selecting an instance to reduce the classification error. Moreover, this algorithm is known as beneficial in classification applications due to its ability in handling high dimensional data and diverse types of data especially none-parametric data type (Arora and Kaur 2020). AdaBoost classifier (Freund and Schapire 1997), is another category of the most widely used ensemble methods. It uses the concept of the combination of sequential independent singular hypothesis to increase the performance of each weak learner. Different machine learning classifiers such as SVM, Decision Tree, etc., can be used as base classifier in this algorithm (Liu et al. 2020). Extreme Gradient Boosting (XGBoost) (Chen and Guestrin 2016) is another ensemble-based classifier. 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. Since XGBoost algorithm uses parallel implementation of sequential trees and reduce overfitting problem, it is considered as one of advantageous classifiers (Bhati et al. 2021). LogitBoost Classifier (Friedman et al. 2000) is another ensemble boosting classifier which was

proposed to deal with noisy data to decrease training error. It can be considered as AdaBoost classifier in which Logistic Regression is used as the base classifier (Tehrany et al. 2019).

2.7. Feature selection methods

Not only do feature selection methods decrease the dimensions of features which contributes to elimination of irrelevant features, but they also select the most effective subset of features which are relevant to main class/classes. Hence, it results in reducing the size of dataset, shortening the training time, improving the accuracy of the classification and decreasing overfitting problem.

2.7.1. Filter method

In 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 dependency of features and the interaction with classifiers (Khaire and Dhanalakshmi 2019). In this study, Mutual Information (MI) as a filter method was used.

$I(X;Y)$ 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 to be more advantageous method. The Equation 6 indicates mutual information of X and Y variables.

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \quad (6)$$

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

$$I(X;Y) = \iint p(x,y) \log \frac{p(x,y)}{p(x)p(y)} dx dy \quad (7)$$

$p(x, y)$ 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 $m \times m$ matrix (Eq. 8), in which μ_{ij} is the mutual information between two attributes of dataset (D) (Sefidian and Daneshpour 2019).

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

2.7.2. Wrapper method

In wrapper method, feature selector is wrapped around the machine learning method and its performance is evaluated by accuracy rate or error rate of the classifier. As a result, since wrapper method relies on the learning algorithm in selecting the most significant subset of features, error rate of the classifier should not be conspicuous. Backward feature elimination (BFE) as a wrapper method is used in this study.

In wrapper based BFE method which is more efficient than forward feature selection method, the importance scores of instances are decided with all data in each iteration and features with least importance are eliminated. As a result, each round improves the performance of the model. Finally, all remained 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

Embedded feature selection algorithm in which in order to select the most effective subset, feature evaluator is assisted by the feature selection method which is embedded as a part of learning algorithm. 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 learning algorithm (Chen et al. 2020). RF as an embedded feature

selection method (RFE) was applied in this study, due to its beneficial results in solving different problems and its generalization. In RFE method, different subsets among features in the training set are produced by bagging method and different decision trees are constructed by each selected feature firstly. The growing 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 consider as a classifier. In the final stage, all trees are connected which create a RF which selects subsets of features among all features in the dataset (Zhou et al. 2016).

2.8. Evaluation metrics

Multi-class confusion matrix is a tool for performance evaluation of classifiers which are applied on the multi-class datasets. Multi- class confusion matrix includes two dimensions indicating actual classes and predicted classes. Table 3 illustrates a multi-class confusion matrix in which A_1, A_2, \dots, A_n are actual classes and N_{ij} is an instance of A_i class which is classified as A_j class. Performance assessment metrics which were used in this study, are presented below (Eq. 9-11) (Danaei Mehr and Polat 2019).

Table 3 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.

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

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

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

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

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

3. Results and discussion

In this study 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.

To evaluate the effect of the feature-reduced datasets on the performance of each ML classifier used in this study, different feature selection methods were used to generate subsets of the most important features in classification process. As a result, of dataset of 19 parameters, Backward Feature Elimination method magnified the importance of 7 parameters in classification of the WQI whereas Mutual Information method could identify 9 parameters as the most prominent parameters, however, Random Forest Embedded feature selector identified 14 parameters (Table 4).

In this study, the effect of splitting methods on the performance of each classifier were 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 split the dataset into training and test datasets.

Table 4 Selected parameters by feature selectors

Dataset	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 (BFE)	7	Station, pH, NH ₄ , NO ₂ , PO ₄ , Mn, Na ⁺
3	Mutual Information (MI)	9	Station, DO, COD, BOD5, NH ₄ , NO ₂ , PO ₄ , Mn, SO ₄
4	Random Forest Embedded (RFE)	14	EC, DO, COD, BOD5, NH ₄ , NO ₃ , NO ₂ , PO ₄ , Cl ⁻ , Fe, Mn, Na ⁺ , SO ₄ , TDS

Table 5 shows the results of performance metrics of the classifiers based on the subset of features selected by the BFE feature selector with KFCV and percentage splitting methods. Considering the features selected by the BFE method, XGBoost classifier outperformed all the other classifiers in terms of accuracy (95.60%), precision (96%), and recall (96%) with 10-FCV splitting method, however, competing with XGBoost classifier in terms of performance, LogitBoost classifier showed better performance with the highest accuracy, precision and recall of 93.63%, 93.60, 93.60%, respectively, over XGBoost classifier with 5-FCV splitting method, yet its superiority was not conspicuous as the XGBoost classifier with 5-FCV obtained the accuracy of 93.33%. Regardless of splitting methods, AdaBoost classifier showed the weakest performance in all metrics compared to the other three classifiers. Although RF classifier's performance couldn't exceed the performance of XGBoost classifier with KFCV splitting method, using percentage splitting method, the performance of the RF was way above the performance of XGBoost in all ratios and it achieved the maximum accuracy (88.4848%), precision (88.80%), and recall (88.50%) at 75:25 splitting ratio. Finally, LogitBoost classifier showed an outstanding performance when the

percentage splitting method was applied datasets and it achieved the maximum accuracy (93.3333%), precision (93.40%) and recall (93.30%) at 75:25 splitting ratio.

Table 5 Performance of the classifiers based on the subset of features selected by BFE feature selector

	Classifier			
	LogitBoost	Random Forest	AdaBoost	XGBoost
5-FCV				
Accuracy (%)	93.6364	84.0909	51.0606	93.3333
Precision (%)	93.60	83.70	66.00	94.00
Recall (%)	93.60	84.10	51.00	93.00
10-FCV				
Accuracy (%)	93.0303	85.7576	59.8484	95.6060
Precision (%)	93.10	85.30	81.00	96.00
Recall (%)	93.00	85.80	60.00	96.00
70:30				
Accuracy (%)	92.9293	86.3636	58.0808	80.3030
Precision (%)	92.90	86.00	64.00	80.00
Recall (%)	92.90	86.40	58.00	80.00
75:25				
Accuracy (%)	93.3333	88.4848	60.00	81.8181
Precision (%)	93.40	88.80	60.00	82.00
Recall (%)	93.30	88.50	60.00	82.00
80:20				
Accuracy (%)	92.4242	84.0909	63.6363	84.8484
Precision (%)	92.80	84.80	63.00	85.00
Recall (%)	92.40	84.10	64.00	85.00

The results of obtained performances by the classifiers based on the subset of features selected by the MI feature selector with both KFCV and percentage splitting methods are exhibited in Table 6. The results demonstrate that the performance of XGBoost exceeded the other classifiers with maximum accuracy, precision and recall of 95.15, 96 and 95%, respectively, by applying 10-FCV. However, the effect of splitting method on the performance of XGBoost is noticeable as the accuracy, precision and recall of the classifier reduced significantly to be maximum 81.2121, 83.00, 81.00%, respectively, at 75:25 splitting ratio. On the other hand, LogitBoost showed an outstanding performance when percentage splitting method was applied and achieved the maximum accuracy, precision and recall of 93.4343, 93.80, 93.40%, respectively, at 70:30 splitting ratio. Again, AdaBoost showed the weakest performance in comparison with the other three classifiers with both KFCV and percentage splitting methods. Finally, although the performance of RF classifier could not exceed the performance of XGBoost and LogitBoost with KFCV splitting method, it outperformed the XGBoost in all splitting ratios and achieved the maximum accuracy (83.3333%), precision (82.50%) and recall (83.30%) at 70:20 splitting ratio.

Table 7 show the results of performance metrics of the classifiers based on the subset of features selected by RFE feature selector with KFCV and percentage splitting methods. According to the outcomes, in general, XGBoost outperformed the other three classifiers with the maximum accuracy, precision and recall of 96.2121, 97.00, and 96.00%, respectively, by applying 10-FCV splitting method. However, LogitBoost classifier could outperform XGBoost in terms of accuracy, precision and recall with 5-FCV splitting method and achieved the accuracy of 95.7576%, precision of 95.90%, and recall of 95.80%.

Table 6 Performance of the classifiers based on the subset of features selected by MI feature selector with KFCV splitting method

	Classifier			
	LogitBoost	Random Forest	AdaBoost	XGBoost
5-FCV				
Accuracy (%)	93.33	83.1818	51.0606	93.4848
Precision (%)	93.40	82.50	66.00	94.00
Recall (%)	93.30	83.20	51.00	94.00
10-FCV				
Accuracy (%)	93.6364	82.2727	59.8484	95.1515
Precision (%)	93.70	81.70	81.00	96.00
Recall (%)	93.60	82.30	60.00	95.00
70:30				
Accuracy (%)	93.4343	83.3333	62.6262	80.8080
Precision (%)	93.80	82.50	61.00	82.00
Recall (%)	93.40	83.30	63.00	81.00
75:25				
Accuracy (%)	92.7273	82.4242	62.4242	81.2121
Precision (%)	93.00	81.90	67.00	83.00
Recall (%)	92.70	82.40	62.00	81.00
80:20				
Accuracy (%)	91.6667	82.5758	71.9696	78.7878
Precision (%)	92.40	82.40	70.00	82.00
Recall (%)	91.70	82.60	72.00	79.00

Table 7 Performance of the classifiers based on the subset of features selected by Random Forest Embedded feature selector

	Classifier			
	LogitBoost	Random Forest	AdaBoost	XGBoost
5-FCV				
Accuracy (%)	95.7576	81.2121	51.06	94.6969
Precision (%)	95.90	80.40	66.00	95.00
Recall (%)	95.80	81.20	51.00	95.00
10-FCV				
Accuracy (%)	95.1515	81.9697	59.8484	96.2121
Precision (%)	95.20	81.30	81.00	97.00
Recall (%)	95.20	81.20	60.00	96.00
70:30				
Accuracy (%)	93.4343	80.8081	58.5858	79.7979
Precision (%)	94.40	79.10	61.00	79.00
Recall (%)	93.40	80.80	59.00	80.00
75:25				
Accuracy (%)	91.5152	83.6364	67.2727	81.8181
Precision (%)	92.90	83.50	65.00	81.00
Recall (%)	91.50	83.60	67.00	82.00
80:20				
Accuracy (%)	91.6667	87.1212	71.9696	81.8181
Precision (%)	93.80	87.20	69.00	81.00
Recall (%)	91.70	87.10	72.00	82.00

When it comes to the performance of the classifiers, applying percentage splitting, LogitBoost showed brilliant performance in all splitting ratios with maximum accuracy, precision and recall of 93.4343, 94.40, 93.40%, respectively. AdaBoost showed the weakest performance considering both KFCV and percentage splitting methods. However, RF classifier outperformed

XGBoost by applying percentage splitting method with the maximum accuracy, precision and recall of 87.1212, 87.20, and 87.10%, respectively, at 80:20 splitting ratio. When percentage splitting method was applied, LogitBoost outperformed the other three classifiers and achieved the maximum accuracy (93.4343%), precision (94.40%) and recall (93.40%) at 70:30 splitting ratio.

Table 8 shows the performance of the classifiers based on all features of interest with KFCV and percentage splitting methods. The performance superiority of XGBoost over other classifiers with either of the KFCV methods is conspicuous with maximum accuracy, precision and recall of 96.9696, 97.00, and 97.00%, respectively with 10-FCV. Yet, this classifier's competitor LogitBoost outperformed the other two classifiers (RF & AdaBoost) with maximum accuracy, precision, and recall of 95.303, 95.30, 95.30%, respectively. AdaBoost showed the weakest performance in the group of the classifiers, however, RF could exceed XGBoost by applying percentage splitting method with the maximum accuracy, precision, and recall of 84.09, 83.70, and 84.10%, respectively, at 80:20 splitting ratio. Obviously, percentage splitting has a positive impact on the performance of XGBoost and LogitBoost classifiers as LogitBoost could outperform the other three classifiers in terms of performance with maximum accuracy, precision, and recall of 92.7273, 93.60, 92.70%, respectively, at 75:25 splitting ratio.

The main objective of similar studies in the literature is to suggest a particular machine learning model which has achieved promising results based on specific parameters of interest with particular splitting method in classification of water quality indices. The purpose of developing a machine learning model in classification of water quality is to evade time-consuming computations to classify the water quality by 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 8 Performance of the classifiers based on all the features of interest

	Classifier			
	LogitBoost	Random Forest	AdaBoost	XGBoost
5-FCV				
Accuracy (%)	94.3939	79.3939	51.0606	94.5454
Precision (%)	94.40	78.90	66.00	95.00
Recall (%)	94.40	79.40	51.00	95.00
10-FCV				
Accuracy (%)	95.303	80.6061	59.8484	96.9696
Precision (%)	95.30	79.80	81.00	97.00
Recall (%)	95.30	80.60	60.00	97.00
70:30				
Accuracy (%)	90.9091	79.2929	53.0303	77.7777
Precision (%)	92.60	77.80	56.00	69.00
Recall (%)	90.90	79.30	53.00	72.00
75:25				
Accuracy (%)	92.7273	80.00	54.5454	73.9393
Precision (%)	93.60	80.40	65.00	75.00
Recall (%)	92.70	80.00	55.00	74.00
80:20				
Accuracy (%)	91.6667	84.0909	63.6363	77.2727
Precision (%)	93.20	83.70	62.00	75.00
Recall (%)	91.70	84.10	64.00	74.00

In this study, obviously, XGBoost classifier achieved the most promising performance over other classifiers with 10-FCV based on all subset of features identified by different feature selectors as well as all features of interest (Fig. 4). The performance metrics 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.

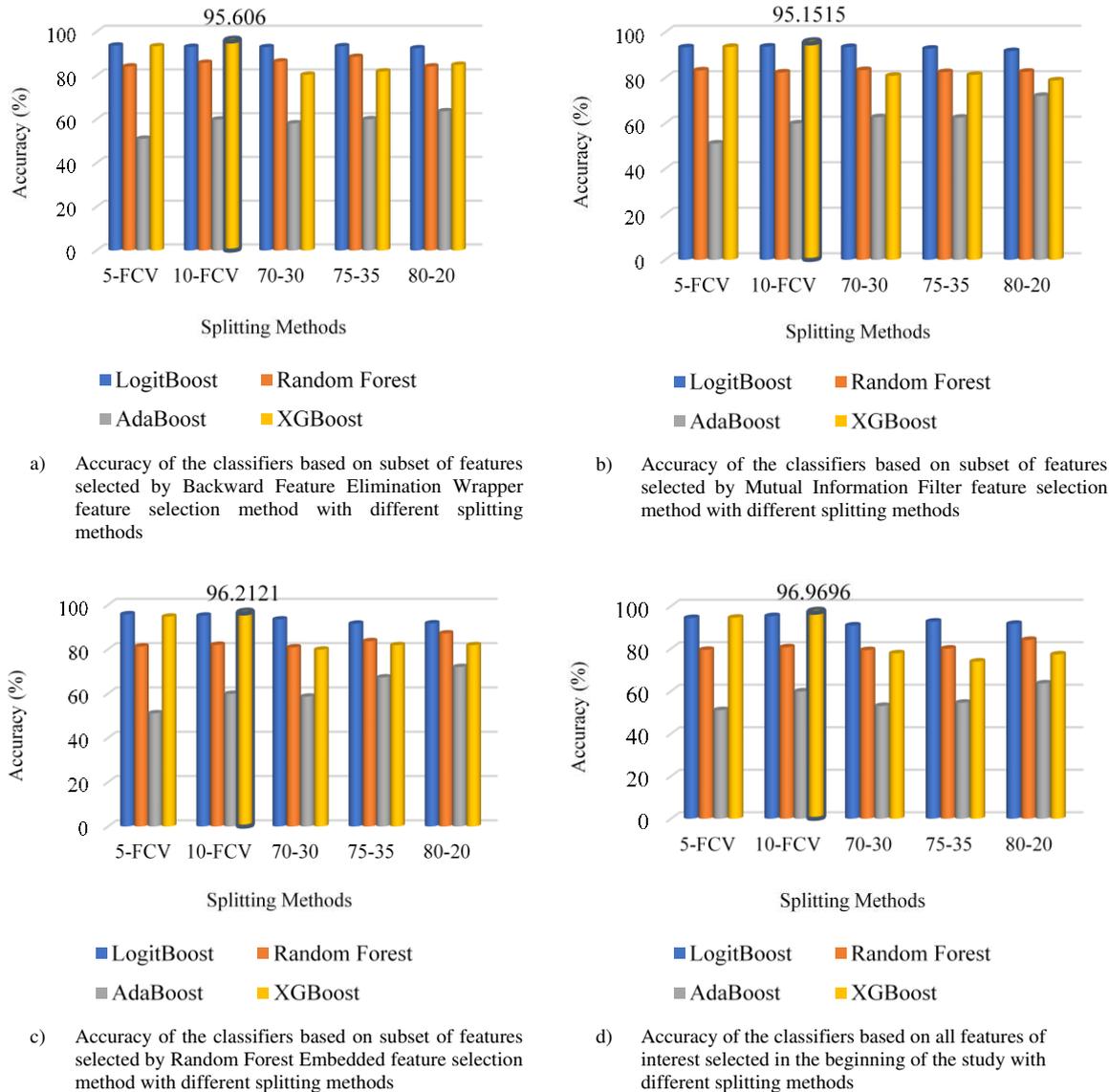


Fig. 4 The accuracy comparison of the classifiers based on different subset of features with different dataset splitting methods

Fig. 5a, is the confusion matrix of the XGBoost classifier which was obtained the best results with 10-FCV on the dataset of parameters suggested by BFE feature selector. The strength of the XGBoost in the classification process based on the subset of features selected by BFE feature selection method is obvious in classification of the Excellent and Unfit for Consumption class labels with maximum classification accuracies of 99.15 and 100%, respectively. However, when it comes to the classification accuracy of Very Poor class label, the performance of the XGBoost classifier decreased significantly to be 72.88%. The performance of the classifier in classification of Good

class label was 97.84%. Finally, the overall classification accuracy of the XGBoost classifier based on the dataset of suggested attributes by BFE Wrapper feature selector was calculated to be 95.606%, as it could correctly classify 631 out of 660 instances.

XGBoost classifier showed weaker performance in classification of Good and Poor class labels based on the parameters suggested by MI Filter method (Fig. 5b) in comparison with the accuracy of the same classifier for the same class label with the subset of features selected by the BFE Wrapper feature selector (Fig. 5a). However, the classification accuracy of this classifier with the subset of features selected by MI Filter method for Excellent, Very Poor and Unfit for Consumption class labels were the same as that of achieved by this classifier based on subset of features selected by BFE Wrapper method.

However, the best classification of the Poor and Very Poor class labels was made by this classifier based on the subset of features selected by RFE feature selection method (Fig. 5c) where the accuracy of the classifier in classification of these two labels were 92.63 and 76.27%, respectively, which was above all the classifications made by the classifier based on the other subset of features which were discussed previously. Yet the best performance achieved by this classifier was where all the features of interests were involved in the process of the classification and based on Fig. 5d the highest performance was achieved by the XGBoost classifier in classification of Good, Poor and Very Poor class labels when all the features were involved in the classification process.

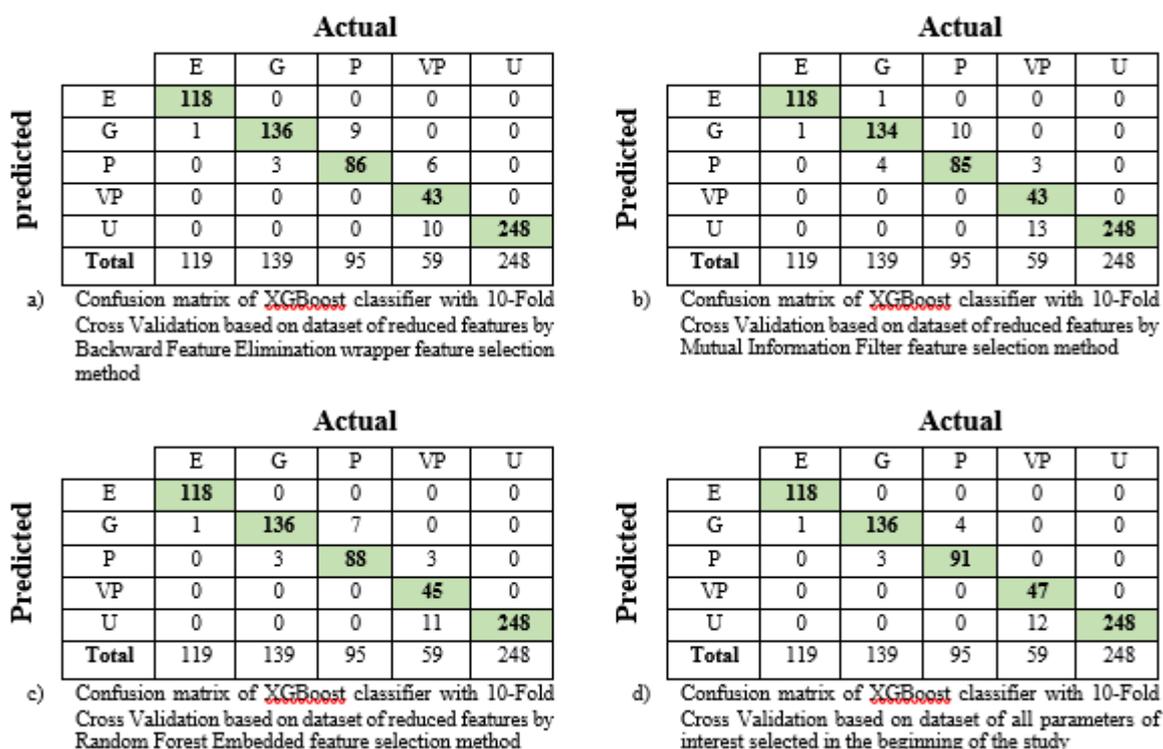


Fig. 5 Confusion matrix of the XGBoost classifier based on different subset of features (E: Excellent; G: Good; P: Poor; VP: Very Poor; U: Unfit for Consumption)

These three classes were those the most classification mistakes were made by the same classifier by involving reduced combination of features in the process of classification. Therefore, it can be concluded that for higher water quality classification accuracy, it is better to involve more parameters than reduced number of parameters as machine learning algorithms can predict multi class of water quality based on its learnings from variety of parameters instead of limited number of parameters. However, involving more parameters is not rational as well, because of two reasons which the authors of this study discussed and unanimously agreed on those. First of all, to decline extra computations by a machine learning algorithm in order to make it cost effective, it is better to use smaller combination of features. Secondly, to make the monitoring and evaluation of the water quality of a basin economically justifiable, the scientist must select a limited number of parameters of interest since some measurements are expensive as well as time consuming. Therefore, because of the aforementioned reasons, although the best performance was achieved by the XGBoost

classifier when involving all the features of interest which were selected in the beginning of the study, it is better to evaluate the performance of the XGBoost classifiers with other subset of features which were used in this study. First of all, although the second-best performance achieved by this classifier based on 14 parameters selected by RFE feature selector, the eliminated parameters (i.e., Year, Month, Station, Temperature, pH) didn't have extra cost of processing, therefore, it would have been rational to involve all the parameters of interest and not sacrifice the performance. However, catching a glance at the performance of the XGBoost classifier by involving 7 parameters selected by BFE feature selector which had the third best performance, it could be suggested to choose this combination of the features and sacrifice a small percentage of the performance since this lack can be compensated during the training stage by larger datasets, yet reducing the number of parameters by 12 could make a model cost-effective.

4. Conclusion

The primary objective of this study was to suggest an ensemble machine learning model for classification of the water quality. 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 5 major class labels which were suggested by the WAWQI function for public consumption. Moreover, since decisions which 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 which were selected as the parameters of interest in 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 all the features were evaluated to propose an effective machine learning algorithm in classification of water quality of the basin. Results demonstrated that the performance of the classifiers were affected significantly by the number of parameters involving in classification process and XGBoost classifier exhibited a brilliant merit over the other three classifiers when either of the datasets was used in classification

process. The best performance achieved by this classifier was when all the features which were selected in the beginning of the study was involved. However, to evade extra computations in order to make this algorithm cost-effective, it is suggested to use the subset of feature selected by BFE feature selector since the results didn't show a significant difference from the results achieved by this classifier when all the parameters of interest were involved in the process of classification. Yet, the number of features decreased by 12 parameters which can make the process cost-effective.

Acknowledgments

Special thanks to the General Directorate of State Hydraulic Works, Department of Survey, Planning and Allocations for the provision of the water quality data of Büyük Menderes Basin and all the supports thereafter.

Declarations

Funding: No funding was received for conducting this study.

Conflict of interest: The authors have no conflicts of interest to declare that are relevant to the content of this article.

Data availability: Not applicable.

Code availability: Codes which were developed in Python are available upon request.

Authors Contribution:

Farid Hassanbaki Garabaghi: Formal analysis, Visualization, Investigation, Software, Writing.

Semra Benzer: Conceptualization, Methodology, Review & Editing.

Recep Benzer: Investigation, Conceptualization, Methodology, Review & Editing.

Ethics approval: Not Applicable.

Consent to participate: Not Applicable.

Consent for publication: Not Applicable.

References

- Arabgol R, Sartaj M, Asghari K (2016) Predicting Nitrate Concentration and Its Spatial Distribution in Groundwater Resources Using Support Vector Machines (SVMs) Model. *Environ Model Assess.* 21:71–82. <https://doi.org/10.1007/s10666-015-9468-0>.
- Arora N, Kaur PD (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 BS, Chugh G, Al-Turjman F, Bhati NS (2021) An improved ensemble based intrusion detection technique using XGBoost. *Trans Emerging Tel Tech.* 32: e4076. <https://doi.org/10.1002/ett.4076>.
- 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 DT, 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 CW, Tsai YH, Chang FR, Lin WC (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>.

- Dalai C, Jha R, Desai VR (2015) Rice Husk and Sugarcane Baggase Based Activated Carbon for Iron and Manganese Removal. *Aquatic Procedia*. 4:1126 – 1133. <https://doi.org/10.1016/j.aqpro.2015.02.143>.
- Danades A, Pratama D, Angraini 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). Istanbul, Turkey. <https://doi.org/10.1109/SGCF.2019.8782290>.
- Dezfooli D, Moghari SMH, 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>.
- Directorate General of Environmental Management, Republic of Turkey, Ministry of Environment and Urbanization, (2016). Büyük Menderes Basin Pollution Prevention Action Plan. Ankara, Turkey. (Turkish)
- 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>.
- Elphick JRF, Bergh KD, Bailey HC (2011) Chronic Toxicity of Chloride to Freshwater Species: Effects of Hardness and Implications for Water Quality Guidelines. *Environmental Toxicology and Chemistry*. 30 (1):239–246. <https://doi.org/10.1002/etc.365>.
- Freund Y, Schapire RE (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). *Ann. Statist.* 28 (2):337-407. <https://doi.org/10.1214/aos/1016218223>.
- Ighalo JO, Adeniyi AG, Marques G (2020) Application of linear regression algorithm and stochastic gradient descent in a machine-learning environment for predicting biomass higher heating value. *Biofuels, Bioprod. Bioref.* 14:1286–1295, 2020. <https://doi.org/10.1002/bbb.2140>.
- Khaire UM, Dhanalakshmi R (2019) Stability of feature selection algorithm: A review. *Journal of King Saud University – Computer and Information Sciences*. In Press. <https://doi.org/10.1016/j.jksuci.2019.06.012>.
- Kumar ZM, 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.
- 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>.
- Lu H, Ma X (2020) Hybrid decision tree-based machine learning models for short-term water quality prediction. *Chemosphere*. 249:126169. <https://doi.org/10.1016/j.chemosphere.2020.126169>.
- Mădălina P, Gabriela BI (2014) Water Quality Index – An Instrument for Water Resources Management. *Aerul și 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 SA, 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 SY, Makhtar M, Rozaimie A, Aziz AA, Jamal AA (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 AS (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 JC, Broers HP (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 SM, Sattari MT, 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>.

Sefidian AM, 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>.

- Sim J, Lee JS, 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 MS, Jones S, Shabani F, Martínez-Álvarez F, Bui DT (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. 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 MdG, Nash S, Olbert AI (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>.
- United States Environmental Protection Agency (EPA) (2021) Sulfate in Drinking Water. <https://archive.epa.gov/water/archive/web/html/sulfate.html>. Accessed 02 July 2021.
- 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.
- Wong JM (1984) Chlorination-Filtration for Iron and Manganese Removal. Journal AWWA Research and Technology.
- 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>.
- 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>.