



## PREDICTING WATER HARDNESS THROUGH DATA-DRIVEN INTELLIGENCE: A COMPARATIVE MACHINE LEARNING APPROACH

<sup>1,\*</sup> Erdem ÇOBAN , <sup>2</sup> Kemal SAPLIOĞLU 

<sup>1</sup> Haliç University, Architecture Department, İstanbul, TÜRKİYE



<sup>2</sup> Süleyman Demirel University, Civil Engineering Department, Isparta, TÜRKİYE

<sup>1</sup>[erdemcoban@halic.edu.tr](mailto:erdemcoban@halic.edu.tr), <sup>2</sup>[kemalsaplioglu@sdu.edu.tr](mailto:kemalsaplioglu@sdu.edu.tr)

### *Highlights*

- Hybrid modeling used to estimate water hardness from multiple input features
- XGBoost and SVR outperformed five other machine learning models
- Physicochemical data formed the basis for training and evaluation
- Model enables real-time monitoring of water quality conditions
- Supports sustainable and data-driven water resource decision-making

## PREDICTING WATER HARDNESS THROUGH DATA-DRIVEN INTELLIGENCE: A COMPARATIVE MACHINE LEARNING APPROACH

<sup>1,\*</sup> Erdem ÇOBAN , <sup>2</sup> Kemal SAPLIOĞLU 

<sup>1</sup> Haliç University, Architecture Department, İstanbul, TÜRKİYE

<sup>2</sup> Süleyman Demirel University, Civil Engineering Department, Isparta, TÜRKİYE

<sup>1</sup>erdemcoban@halic.edu.tr, <sup>2</sup>kemalsaplioglu@sdu.edu.tr

(Received: 30.05.2025; Accepted in Revised Form: 23.07.2025)

**ABSTRACT:** Water hardness is a key parameter in evaluating water availability and plays a critical role in the development of sustainable water management strategies. In this study, water hardness was estimated using four advanced machine learning algorithms: Random Forest (RF), Support Vector Regression (SVR), Multiple Linear Regression (MLR), and AdaBoost. The input variables used for model training included sodium (Na), potassium (P), and anion-cation concentrations. The dataset was obtained from the Beşkonak flow measurement station located on the Köprüçay Stream in southern Turkey, which plays a vital role in the regional hydrological system. Model performance was assessed using statistical indicators such as mean square error (MSE), root mean square error (RMSE), coefficient of determination ( $R^2$ ), and mean absolute percentage error (MAPE). Among the evaluated models, MLR achieved the highest accuracy with an  $R^2$  value of 0.9945, followed by SVR with 0.9939, AdaBoost with 0.9700, and RF with 0.9400. In terms of predictive error, MLR yielded the lowest RMSE value at 0.248, while SVR, AdaBoost, and RF recorded RMSE values of 0.264, 0.545, and 0.592, respectively. These results demonstrate that the MLR model outperformed the others in estimating water hardness, while the remaining models also produced acceptable levels of accuracy. This study provides a valuable contribution to the understanding of data-driven approaches for water quality assessment and offers insights for future water resource planning.

**Keywords:** Water Hardness, Machine Learning, Prediction Models, Regression Analysis, Support Vector Regression, Random Forest, Multiple Linear Regression, Adaboost Algorithm

### 1. INTRODUCTION

Water is one of the most important needs for living things to sustain life. However, the vast majority of existing water is not a drinkable freshwater source. A certain portion of freshwater is also undrinkable due to its low quality. At this point, in order to have information about the quality of water, it is necessary to know the parameters that make up the quality of water and the relationship between them. One of these, water quality, is a water parameter that expresses the amount of dissolved minerals in water, especially calcium ( $Ca^{2+}$ ) and magnesium ( $Mg^{2+}$ ) ions. When we look at the effects of water hardness on health, while values such as calcium and magnesium at low levels are even beneficial for meeting the mineral needs of humans and other living things, very high hardness levels can cause digestive problems [1]. In terms of home and industrial uses, hard water interacts with soap and chalk, reducing foaming levels. In white goods and pipes, it causes lime accumulation, reducing the efficiency of the devices and reducing their lifespan [2]. In addition, it increases maintenance costs in industrial uses and decreases productivity in this area [3]. When we look at the environmental effects of water hardness, it leads to changes in habitats and environments for organisms living in natural water resources [4],[5]. Changes in water hardness threaten this suitability. In addition, when we look at its effect in the agricultural area, very hard water causes salt accumulation in the soil and changes the chemistry of the soil. This negatively affects plant health [6]. From this perspective, measuring and managing water hardness correctly is of critical importance both in extending the life of devices and optimizing water use. Therefore, determining and controlling water hardness is an important part of water management strategies. In terms of

\*Corresponding Author: Erdem ÇOBAN, [erdemcoban@halic.edu.tr](mailto:erdemcoban@halic.edu.tr)

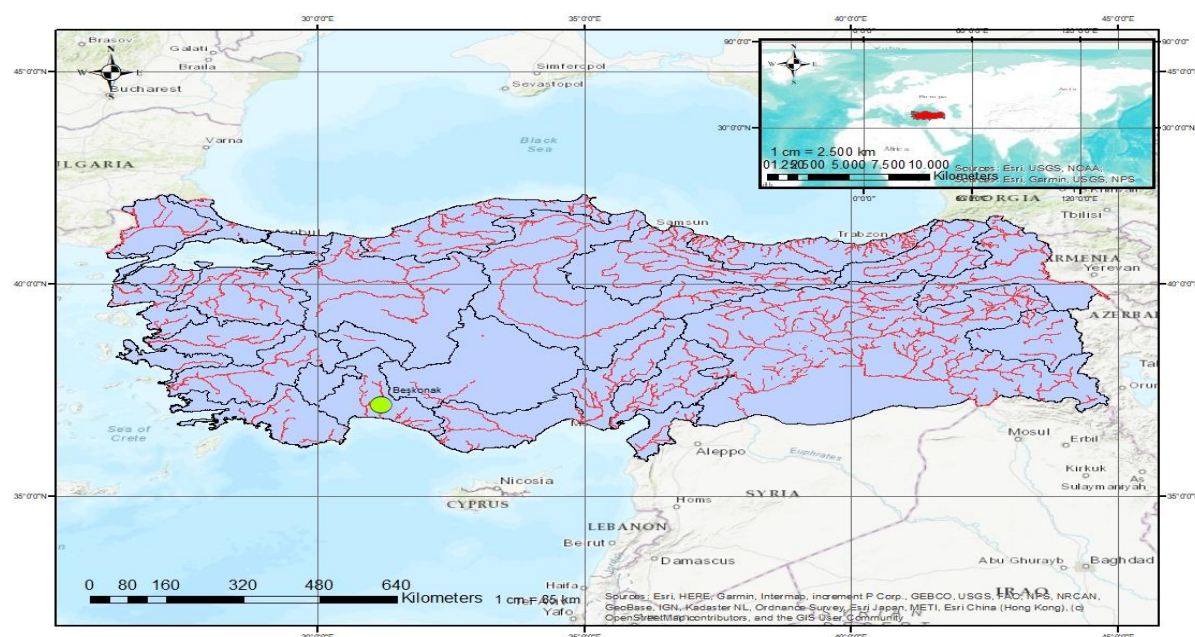
determining and correctly evaluating water hardness, machine-learning methods have begun to be used frequently in recent years thanks to their ability to extract meaningful patterns and predictions from large data sets. Mohseni et al., [7] estimated Weighted Arithmetic Water Quality Index (WA-WQI) to understand groundwater quality. Madhya In addition to multiple linear regression (MLR), four machine learning (ML) models, namely artificial neural network (ANN), support vector machine (SVM), random forest (RF) and extreme gradient boosting (XGBoost), were developed for WA - WQI prediction in Ujjain, Pradesh. In this study, it was shown that the XGBoost model gave better results than other ML models. Talukdar et al., [8] estimated the lake water quality index with sensitivity-uncertainty analysis using deep learning algorithms. Acar & Saplioglu [9] estimated the water hardness determined by multiple regression using the water parameters of the Kemahboğazı stream observation station with the ANFIS method (see also, [10], [11], [12], [13] for soft computing applications in hydrological modeling). Çoban [14] tried to estimate the summer season average flow from the autumn, winter and spring seasonal average flow values as three inputs and one output by using the Random Forest (RO) and Adaptive Boosting (AdaBoost) algorithms. Mosavi et al., [15] used ensemble machine-learning models to estimate the sensitivity of groundwater hardness. A comparative study was conducted with multivariate discriminant analysis (MDA) and the performance of two ensemble models, namely boosted regression trees (BRT) and random forest (RF), was examined. Again, there are studies on the estimation of water quality and parameters using machine learning algorithms [16], [17], [18], [19]. Evaluates the effectiveness of machine learning (ML) models in predicting Water Quality Index (WQI) for industrially polluted Aik-Stream in Pakistan. Using 19 water quality parameters and environmental factors, AdaBoost, K- Nearest Neighbors (K-NN), Gradient Boosting (GB), Random Forests (RF), Support Vector Regression (SVR) and Bayesian They compared six different ML algorithms, including Regression (BR).

This study fills an important gap in this area by evaluating the effectiveness of machine learning algorithms for estimating water hardness. Random Forest (RF), Support Vector Regression (SVR), Multiple Linear Regression (MLR) and AdaBoost algorithms were compared in the study and it was aimed to estimate water hardness with the highest accuracy. Sodium (Na), potassium (P) and anion-cation values were used as input variables in model training. Model comparisons were made using metrics such as regression analysis, mean square error (MSE). This study is an important guide for future studies by emphasizing the effectiveness of machine learning models in water management and quality control processes. The obtained results show that water hardness estimation can be done faster and more cost-effectively compared to traditional methods. In addition, it sheds light on future water quality estimation studies by showing how machine-learning models can be optimized in the fields of environmental engineering and water management.

## 2. MATERIAL AND METHODS

### 2.1. Study Area and Dataset

The water quality parameter dataset used in this study covers the period between 1995 and 2002 and was obtained from the Beşkonak flow measurement station (Station No. 902), located on the Köprüçay Stream. Köprüçay originates from Güllüce Mountain at an altitude of 2151 meters, stretches for 178.8 kilometers, and drains a basin area of approximately 2357 km<sup>2</sup>. With an annual average flow volume of 3065 hm<sup>3</sup>, the stream plays a vital role in the region's hydrology. The Beşkonak streamflow observation station is situated on the Köprüçay River, 36 kilometers north of the Serik Basin within the administrative borders of Antalya Province, at coordinates 37.141° N latitude and 31.187° E longitude (Figure 1).



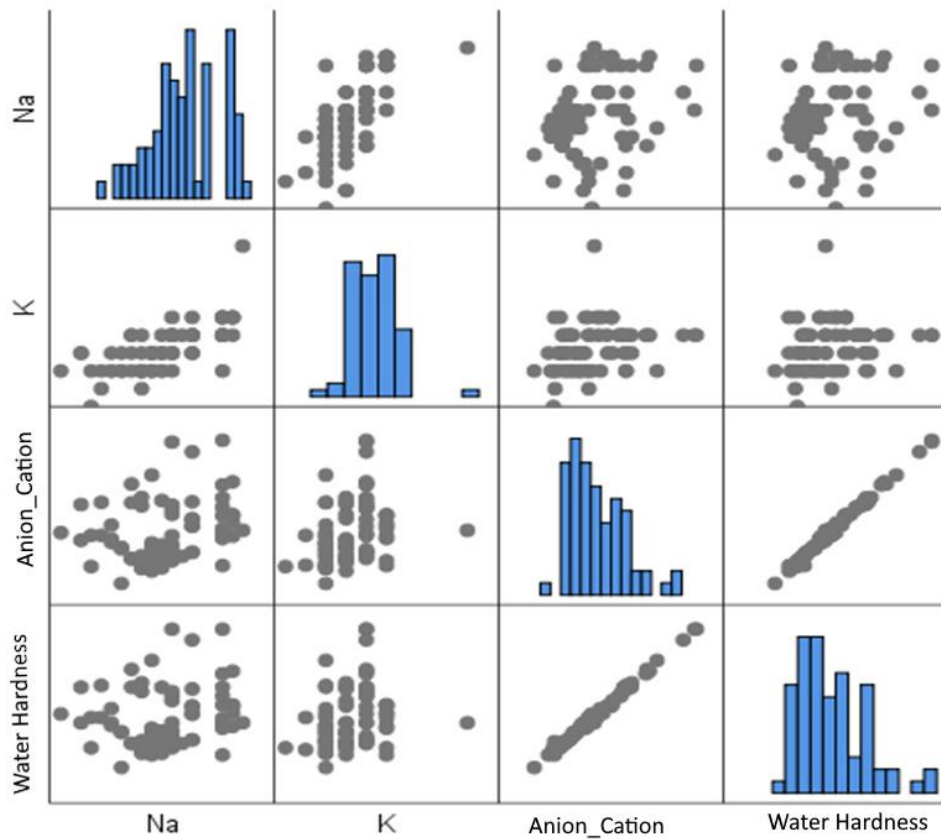
**Figure 1.** Beşkonak measurement station location map

The statistical characteristics of sodium, potassium, anion–cation concentration, and water hardness are presented in Table 1. The average hardness was 14.577 with a standard deviation of 3.363, indicating moderate variability across samples. Sodium values ranged from 0.100 to 0.280 with a low standard error (0.005), while potassium exhibited a maximum of 0.090 and higher kurtosis (3.460), suggesting a peaked distribution. The highest hardness was observed in December (15.500), indicating possible seasonal influence. Skewness and kurtosis values suggest that anion–cation and hardness data are slightly right-skewed and leptokurtic. Overall, the dataset ( $n = 73$ ) provides reliable input for modeling and trend analysis of water quality parameters (Table 2). The input variables used in this study include sodium (Na), potassium (K), and anion–cation balance, while the output variable is total water hardness. Sodium and potassium concentrations are measured in milligrams per liter (mg/L), and the anion–cation balance is expressed in milliequivalents per liter (meq/L). The water hardness, which serves as the target variable, is given in milligrams per liter as calcium carbonate (mg/L as  $\text{CaCO}_3$ ). This study utilized only sodium, potassium, and anion–cation balance as input features, as these were the only parameters available in the dataset. Although other variables like pH or TDS could enhance model performance, they were excluded due to data limitations.

The relationships between sodium, potassium, anion–cation balance, and water hardness were further explored using a scatter matrix (Figure 2). This visualization reveals a strong linear correlation between anion–cation concentration and total hardness, indicating that variations in ionic balance significantly affect hardness values. Meanwhile, sodium and potassium exhibit weaker or dispersed associations with other parameters. The diagonal histograms in the matrix also suggest moderately skewed distributions, particularly for sodium and anion–cation data, along with a few outliers.

**Table 1.** Statistical analysis results of the data

	Sodium	Potassium	Anion Cation	Hardness
Average	0.203	0.033	3,144	14.577
Standard Error	0.005	0.002	0.081	0.394
Median	0.200	0.030	3.020	14.000
Standard Deviation	0.043	0.013	0.690	3.363
Sample Variance	0.002	0,000	0.477	11.311
Kurtosis	-0.595	3,460	0.828	0.893
Skewness	-0.115	0.879	0.975	1.021
December	0.180	0.090	3.210	15.500
Minimum	0.100	0,000	1.980	9.000
Maximum	0.280	0.090	5.190	24.500
Count	73	73	73	73
Reliability Level (95.0%)	0.010	0.003	0.161	0.785

**Figure 2.** Scattering Matrix

**Table 2.** Correlation analysis of data

		K	Na	Anion-Cation	Hardness
K	Pearson Correlation	1	0.701 **	0.294 *	0.232 *
	Sig. (2-tailed)		0.000	0.012	0.048
Na	Pearson Correlation	0.701 **	1	0.270 *	0.197
	Sig. (2-tailed)	0.000		0.021	0.094
Anion-Cation	Pearson Correlation	0.294 *	0.270 *	1	0.996 **
	Sig. (2-tailed)	0.012	0.021		0.000
Hardness	Pearson Correlation	0.232 *	0.197	0.996 **	1
	Sig. (2-tailed)	0.048	0.094	0.000	

Table 2 shows the Pearson correlation coefficients between the selected water quality parameters. A very strong positive correlation is evident between anion–cation concentration and water hardness ( $r = 0.996$ ,  $p < 0.01$ ), indicating that hardness is largely governed by ionic composition. Sodium shows a moderate correlation with potassium ( $r = 0.701$ ,  $p < 0.01$ ), while its correlation with hardness is weaker and not statistically significant ( $p = 0.094$ ). Potassium exhibits weak but significant correlations with both anion–cation concentration and hardness.

## 2.2. Machine Learning Algorithms

In this study, four different machine learning algorithms—Random Forest, Support Vector Regression (SVR), Multiple Linear Regression (MLR), and AdaBoost—were employed to estimate water hardness based on key water quality parameters. These algorithms were selected due to their complementary strengths in handling nonlinear relationships, high-dimensional input features, and varying levels of data complexity. Random Forest and AdaBoost are ensemble methods known for their robustness and resistance to overfitting, making them suitable for environmental data with noise and variability. SVR was chosen for its ability to model complex, nonlinear patterns in small datasets using kernel functions. MLR, as a classical baseline method, provides a reference for evaluating the added value of more advanced models. Together, these algorithms enable a comparative analysis to identify the most accurate and generalizable approach for predicting water hardness in freshwater systems.

### 2.2.1 Random Forest Algorithm

The Random Forest algorithm is an ensemble machine learning method developed by Leo Breiman [20]. Ensemble classification algorithms use multiple classifications instead of a single classification. There are multiple decision trees in the random forest algorithm structure, and this algorithm produces results by averaging the results from these decision trees. A trained  $k$ -number tree is collected in the random forest model defined in equation 1.

$$H(X, \theta_j) =, (j = 1, 2, 3, \dots, m) \quad (1)$$

In the equation, the expression  $H(X, \theta_j)$  serves as a meta-decision tree classification function. While  $x$  represents the input feature vector of the training dataset,  $\theta_j$  represents an independent and uniformly distributed random vector that determines the growth process of the tree.

### 2.2.2 Support Vector Regression Algorithm

Support Vector Machines (SVM), introduced by Vapnik and colleagues, are powerful supervised learning models used for both classification and regression tasks [21]. The two main branches of SVM are Support Vector Classification (SVC) and Support Vector Regression (SVR). While SVC is applied to

categorical outputs, SVR is commonly used in continuous and nonlinear regression problems. Support Vector Regression works by mapping input data into a high-dimensional feature space and constructing a hyperplane that best fits the data within a predefined error margin. The main objective is to minimize the model complexity while allowing for a certain degree of tolerance in prediction errors. The optimization problem that defines SVR involves minimizing a loss function under specific constraints, as represented in Equation 3 based on the general form provided in Equation 2.

$$D = \{(x_i, y_i) | i = 1, 2, \dots, n\} \quad (2)$$

$$f(x) = w^T \phi(x) + b, \quad w \in R^d, b \in R \quad (3)$$

Here;  $w$  and  $b$  are model parameters, and  $\phi(x)$  represents the kernel function that transforms the inputs into a higher dimensional space.

### 2.2.3 Multiple Linear Regression

Multiple Linear Regression (MLR) is a widely used statistical technique for modeling the linear relationship between a dependent variable and two or more independent variables [22]. The method assumes a linear association and aims to estimate the coefficients that best explain the variation in the output variable based on the inputs. The MLR model seeks to minimize the sum of squared differences between the observed and predicted values, resulting in a line (or hyperplane) of best fit. The general form of the model is presented in Equation 4:

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_n x_n + \varepsilon \quad (4)$$

Here;  $y$  is the dependent variable.  $x_1, x_2, \dots, x_n$  are the independent variables. They are constant terms  $\beta_1, \beta_2, \dots, \beta_n$  coefficients and determine the weights of the independent variables.  $\varepsilon$  whereas is the error term that includes measurement errors and the effects of variables not included in the model.

### 2.2.4. AdaBoost algorithm

The AdaBoost algorithm iteratively searches the sample feature space and finds the training feature weights. In the iterative process, the feature weights of the training examples are continuously adjusted [23]. The most basic feature of this algorithm is to create the most successful model by taking the good aspects of weak models. In AdaBoost, multiple models are created and the strong aspects of each model are combined. The algorithm then tries to obtain a successful model from these models. The AdaBoost algorithm combines weak models and produces the output of the more successful model.

## 2.3. Model Training and Evaluation

### 2.3.1. Creating Training and Test Datasets

Out of a total of 73 observations, 80% (59 samples) were allocated for training and 20% (14 samples) for testing to evaluate model performance. This stratified split ensures that the models are both effectively trained and reliably validated on unseen data. Such a division is commonly adopted in similar studies, as it provides a balanced trade-off between learning capacity and generalization ability.

## 2.4. Evaluation Metrics

### 2.4.1. Mean Square Error (MSE)

It is the average of the squared differences between the true and predicted values. MSE takes into

account the magnitude of the errors and penalizes large errors more. MSE emphasizes how well the model prevents large errors. It does not ignore small errors. Because the errors are squared, MSE emphasizes outliers more and therefore can be a sensitive measure. MSE is calculated as in equation 5. Here; N is the total number of data,  $y_j$  the true (observed) value  $\hat{y}_j$  shows the value predicted by the model.

$$MSE = \frac{1}{N} \sum_{j=1}^N (y_j - \hat{y}_j)^2 \quad (5)$$

#### 2.4.2. Root Mean Square Error (RMSE)

Root Mean Square Error (RMSE) is the square root of the Mean Squared Error (MSE) and provides a more interpretable measure of prediction error. Unlike MSE, RMSE expresses the error in the same unit as the predicted variable, making it more intuitive for evaluating model performance. It penalizes larger errors more heavily due to the squaring operation and is thus sensitive to outliers or extreme deviations. RMSE offers a direct indication of the average magnitude of error between predicted and observed values and is widely used in regression model evaluations.

The RMSE is calculated as shown in Equation 6, where it measures the average squared difference between observed  $y_j$  and predicted  $\hat{y}_j$  values:

$$RMSE = \sqrt{\frac{1}{N} \sum_{j=1}^N (y_j - \hat{y}_j)^2} \quad (6)$$

#### 2.4.3. Coefficient of Determination ( $R^2$ )

The Coefficient of Determination ( $R^2$ ) measures how much of the variance in the dependent variable is explained by the independent variables in the model. In other words, it reflects the proportion of the total variation in the output that is captured by the model's predictions. As the  $R^2$  value approaches 1, the model demonstrates higher explanatory power. A value of 0.0 indicates no explanatory capability, whereas a value of 1.0 implies perfect prediction.

The  $R^2$  value is calculated as shown in Equation 7:

$$R^2 = 1 - \frac{\sum_{j=1}^N (y_j - \hat{y}_j)^2}{\sum_{j=1}^N (y_j - \bar{y})^2} \quad (7)$$

#### 2.4.4. Mean Absolute Percent Error (MAPE)

Mean Absolute Percentage Error (MAPE) is a widely used metric for evaluating the accuracy of regression-based predictions, especially in fields such as business forecasting and time series analysis. It measures the average absolute percentage difference between the actual and predicted values, offering an interpretable and unit-free evaluation of model performance. Because MAPE expresses errors as percentages, it allows for meaningful comparison across datasets with different scales or units.

The MAPE is calculated as shown in Equation 8:

$$MAPE = \frac{1}{N} \sum_{j=1}^N \left[ \frac{y_j - \hat{y}_j}{y_j} \right] * 100 \quad (8)$$



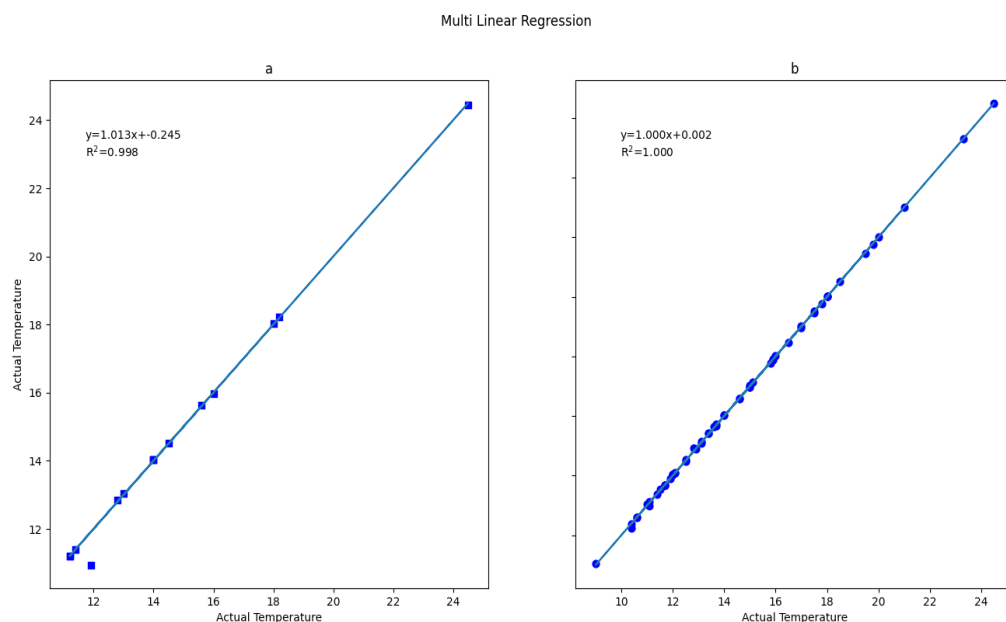
### 3. RESULTS AND DISCUSSION

The evaluation metrics for the four machine learning algorithms—AdaBoost, Random Forest (RF), Support Vector Regression (SVR), and Multiple Linear Regression (MLR)—based on three input features and one output variable are summarized in Table 3. The dataset was divided into 80% training and 20% testing subsets. Each model was trained on the training set, and its performance was evaluated using both training and test data. According to the Root Mean Square Error (RMSE, expressed in mg/L), MLR yielded the lowest error with values of 0.031 mg/L (train) and 0.248 mg/L (test), indicating the highest prediction accuracy. It was followed by SVR, RF, and AdaBoost. In terms of the Coefficient of Determination ( $R^2$ , unitless), MLR again achieved the best results with  $R^2$  values of 0.999 (train) and 0.994 (test), closely followed by SVR. Both RF and AdaBoost showed slightly lower  $R^2$  scores in the test phase. When considering the Mean Absolute Percentage Error (MAPE, expressed as %), the ranking remains consistent with that of RMSE and  $R^2$ , demonstrating internal reliability among metrics. These results suggest that MLR outperforms the other models in this specific application, offering high accuracy and generalizability.

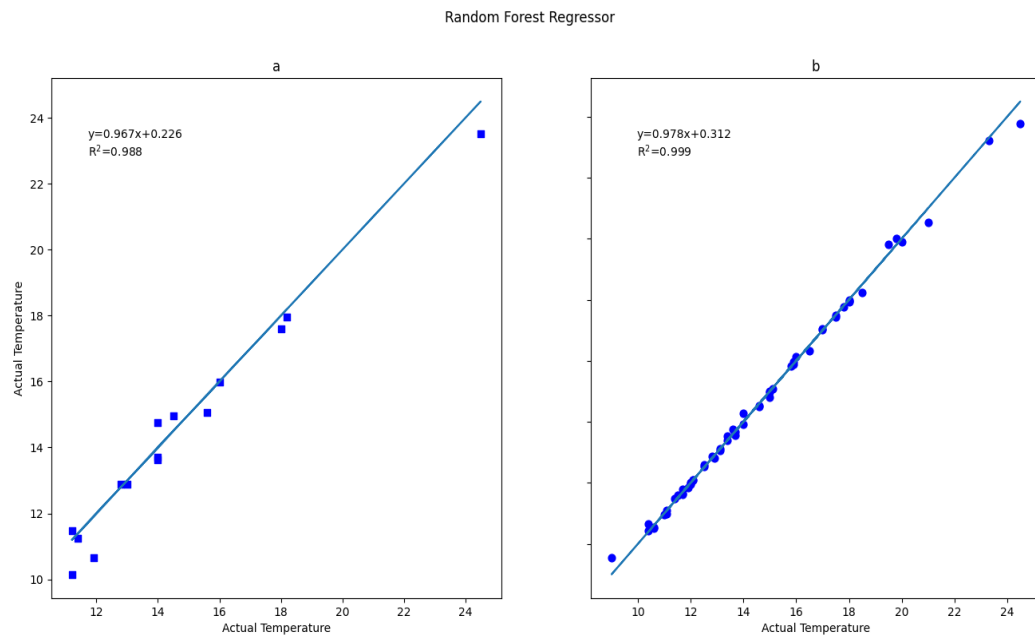
The scatter diagram of the test data and predictions of the Multi Linear Regression algorithm is shown in Figure 3.a. The majority of the points are located on the 1-1 line or in a close region. The scatter diagram of the training data and the prediction data is shown in Figure 3.b. Here again, as in the test data, the points are located in a region close to the 1-1 line.

**Table 3.** Evaluation Metrics Table

	Train			Test		
	RMSE (mg/L)	$R^2$	MAPE (%)	RMSE (mg/L)	$R^2$	MAPE (%)
<i>AdaBoost</i>	0.195	0.996	0.010	0.545	0.974	0.031
<i>RF</i>	0.170	0.997	0.007	0.592	0.969	0.033
<i>SVR</i>	0.063	0.999	0.003	0.264	0.993	0.009
<i>MLR</i>	0.031	0.999	0.002	0.248	0.994	0.006

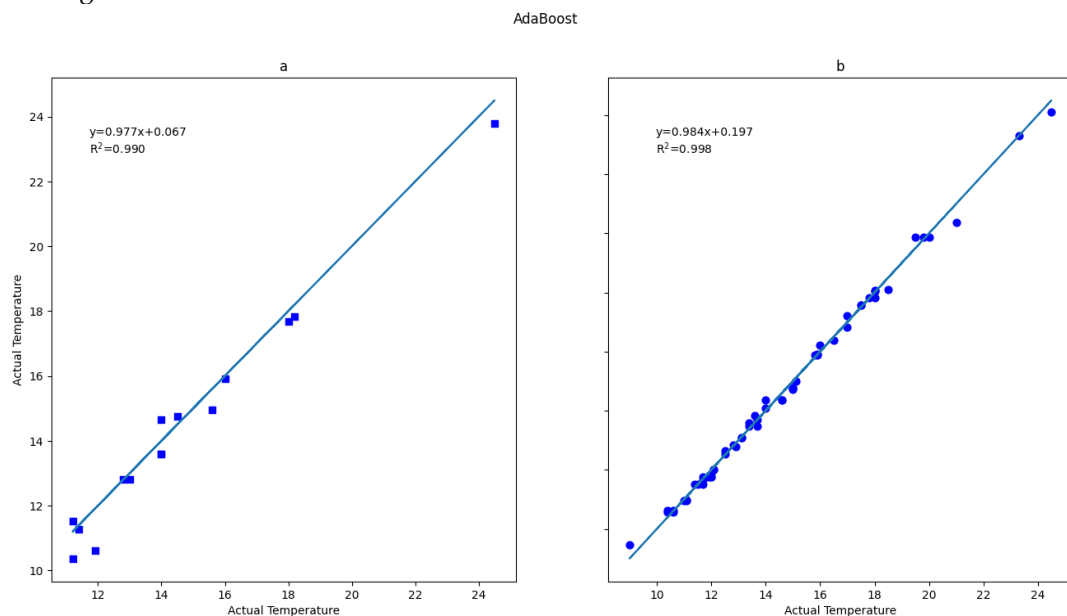


**Figure 3.** Scatter diagrams showing observed vs. predicted water hardness values for the Multiple Linear Regression (MLR) model using test (a) and training (b) datasets (unit: mg/L).



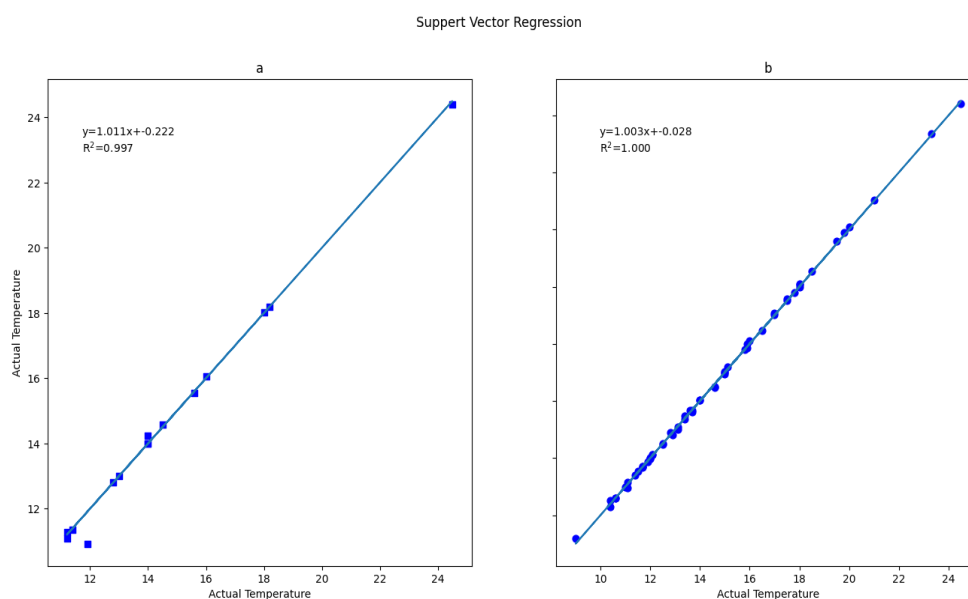
**Figure 4.** Scatter diagrams showing observed vs. predicted water hardness values for the Random Forest (RF) model using test (a) and training (b) datasets (unit: mg/L).

The scatter diagram of the test data and predictions of the Random Forest algorithm is shown in Figure 4.a. The majority of the points are located on the 1-1 line or in a close area. The scatter diagram of the training data and prediction data is shown in Figure 4.b. Here again, as in the test data, the points are located in a region close to the 1-1 line.



**Figure 5.** Scatter diagrams showing observed vs. predicted water hardness values for the AdaBoost model using test (a) and training (b) datasets (unit: mg/L).

The scatter diagram of the AdaBoost algorithm's test data and predictions is shown in Figure 5.a. The majority of the points are located on or near the 1-1 line. Figure 5.b shows the scatter diagram of the training data and prediction data. Here, as in the test data, the points are located in a region close to the 1-1 line.



**Figure 6.** Scatter diagrams showing observed vs. predicted water hardness values for the Support Vector Regression (SVR) model using test (a) and training (b) datasets (unit: mg/L).

The scatter diagram of the Support Vector Regression algorithm, which includes test data and predictions, is shown in Figure 6.a. The majority of the points are located on the 1-1 line or in a close region. The scatter diagram of the training data and prediction data is shown in Figure 6.b. Here, as in the test data, the points are located in a region close to the 1-1 line.

The results of this study are in line with recent literature on the use of machine learning techniques in water quality prediction. Nouraki et al., [24] applied Multiple Linear Regression, Support Vector Regression, and Random Forest Regression models to predict total hardness, total dissolved solids, and sodium adsorption ratio in the Karun River and reported very high accuracy. In their study, the Multiple Linear Regression model predicted total hardness with a coefficient of determination of 0.99 and a root mean square error of 1.54 milligrams per liter. In comparison, the present study achieved a coefficient of determination of 0.9945 and a root mean square error of 0.248 with the same regression method, indicating even better performance using fewer input variables. Similarly, Mosavi et al. [25] used ensemble models such as Random Forest and Boosted Regression Trees to map groundwater hardness susceptibility, emphasizing the role of spatial variables such as elevation, groundwater depth, and proximity to rivers. While our model focused on chemical parameters including sodium, potassium, and anion–cation balance, it also identified anion–cation balance as the dominant predictor, supported by a Pearson correlation value of 0.996. Additionally, Ahmed et al., [26] demonstrated that even with a minimal number of input parameters such as pH, temperature, turbidity, and total dissolved solids, machine learning algorithms like Gradient Boosting and Polynomial Regression could estimate the Water Quality Index with mean absolute errors below three units. These studies support the conclusion that accurate water quality prediction can be achieved using both simple and complex models, and that our approach provides competitive accuracy while maintaining input efficiency. The predicted water hardness values obtained from the machine learning models can play a critical role in optimizing water treatment processes in both municipal and industrial systems. For instance, accurate forecasts can help determine the appropriate dosing of softening agents, prevent scale formation in pipelines and boilers, and reduce operational costs in water-intensive sectors such as textile and food industries. Additionally, in domestic contexts, managing water hardness is essential for extending the lifespan of appliances and improving detergent efficiency.

#### 4. CONCLUSIONS

This study evaluates the effectiveness of machine learning models in water hardness estimation, emphasizing the importance of data-driven approaches in water management and quality control processes. While traditional water hardness estimation methods are usually based on time-consuming laboratory analyses, machine-learning techniques have the potential to make fast and low-cost estimations from large data sets. In this context, the study compares the water hardness estimation performances of Random Forest (RF), Support Vector Regression (SVR), Multiple Linear Regression (MLR) and AdaBoost algorithms.

The analysis results indicate that the MLR algorithm demonstrates the highest accuracy, with an  $R^2$  value of 0.9945 and an RMSE of 0.248, thus exhibiting the best performance in estimating water hardness. The SVR model also showed strong predictive ability, with an  $R^2$  of 0.9939 and an RMSE of 0.264. In comparison, the prediction performances of the AdaBoost and RF models were relatively lower, with AdaBoost achieving an  $R^2$  of 0.97 and an RMSE of 0.545, and RF yielding an  $R^2$  of 0.94 and an RMSE of 0.592. The findings show that linear regression-based methods can provide more successful results compared to complex machine learning algorithms in water hardness estimation. The study encourages the use of data-driven decision mechanisms in water hardness estimation. It ensures the determination of the most appropriate method by comparing the accuracies of machine learning models. It contributes to faster, lower-cost and more reliable estimation in water management processes. It helps to better analyze the effects of water hardness changes in terms of environmental sustainability. Water hardness is a critical parameter not only in terms of drinking water quality, but also in industrial processes, agriculture, domestic use and its effects on the ecosystem. Therefore, its accurate estimation is of great importance in terms of sustainable water management. The findings of the study indicate that the widespread use of machine learning algorithms in the field of water quality estimation should be encouraged. In short, it reveals that machine-learning techniques provide a strong alternative to traditional methods by providing high accuracy in water hardness estimation. It is evaluated that the results obtained in the study will guide future water management policies and optimize water quality monitoring processes. The current study has some limitations. The study only made predictions using sodium (Na), potassium (P), and anion-cation data. Future studies can be expanded to include additional parameters such as pH, total dissolved solids (TDS), or temperature. Furthermore, increasing the prediction accuracy can be investigated by using deep learning algorithms or hybrid models.

#### Declaration of Ethical Standards

The authors declare that all procedures followed in this study comply with the ethical standards of the relevant institutional and national guidelines. Since this research is based entirely on previously collected physicochemical water quality measurements and does not involve human or animal subjects, no ethical approval was required.

#### Credit Authorship Contribution Statement

Conceptualization: Erdem Çoban, Kemal Saplıoğlu, Methodology: Erdem Çoban, Software & Data Analysis: Erdem Çoban, Investigation: Erdem Çoban, Kemal Saplıoğlu, Writing – Original Draft: Erdem Çoban, Writing – Review & Editing: Kemal Saplıoğlu, Supervision: Kemal Saplıoğlu

#### Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

### Funding / Acknowledgements

The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

### Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

### REFERENCES

- [1] M. Hori, K. Shozugawa, K. Sugimori, and Y. Watanabe, "A survey of monitoring tap water hardness in Japan and its distribution patterns," *Sci. Rep.*, vol. 11, no. 1, p. 13546, 2021, Doi: 10.1038/s41598-021-92949-8.
- [2] L. Cohen, A. Moreno, and J. L. Berna, "Influence of anionic concentration and water hardness on foaming properties of a linear alkylbenzene sulfonate," *J. Am. Oil Chem. Soc.*, vol. 70, pp. 75–78, 1993. doi: 10.1007/BF02545371
- [3] T. Morales-Pinzón, R. Lurueña, X. Gabarrell, C. M. Gasol, and J. Rieradevall, "Financial and environmental modelling of water hardness—Implications for utilising harvested rainwater in washing machines," *Sci. Total Environ.*, vol. 470, pp. 1257–1271, 2014, doi: 10.1016/j.scitotenv.2013.10.101.
- [4] D. J. Soucek et al., "Influence of water hardness and sulfate on the acute toxicity of chloride to sensitive freshwater invertebrates," *Environ. Toxicol. Chem.*, vol. 30, no. 4, pp. 930–938, 2011, doi: 10.1002/etc.454.
- [5] Avcı, B. C., Kesgin, E., Atam, M., & Tan, R. I. (2023). Modeling agricultural practice impacts on surface water quality: Case of Northern Aegean watershed, Turkey. *International Journal of Environmental Science and Technology*, 20(5), 5265–5280. <https://doi.org/10.1007/s13762-022-04477-1>
- [6] Y. Liu, P. Wu, D. Zhu, L. Zhang, and J. Chen, "Effect of water hardness on emitter clogging of drip irrigation," *Trans. Chin. Soc. Agric. Eng.*, vol. 31, no. 20, pp. 95–100, 2015. doi: 10.11975/j.issn.1002-6819.2015.20.014
- [7] U. Mohseni, C. B. Pande, S. C. Pal, and F. Alshehri, "Prediction of weighted arithmetic water quality index for urban water quality using ensemble machine learning model," *Chemosphere*, vol. 352, p. 141393, 2024, doi: 10.1016/j.chemosphere.2024.141393.
- [8] S. Talukdar et al., "Predicting lake water quality index with sensitivity-uncertainty analysis using deep learning algorithms," *J. Clean. Prod.*, vol. 406, p. 136885, 2023.
- [9] R. Acar and K. Saplıoğlu, "Etkili girdi parametrelerinin çoklu regresyon ile belirlendiği su sertliğinin ANFIS yöntemi ile tahmin edilmesi," *Afyon Kocatepe Univ. J. Sci. Eng.*, vol. 22, no. 6, pp. 1413–1424, 2022, <https://doi.org/10.35414/akufemubid.1147492>.
- [10] K. Saplıoğlu and R. Acar, "K-Means Kümeleme Algoritması Kullanılarak Oluşturulan Yapay Zekâ Modelleri ile Sediment Taşınımının Tespiti", *Bitlis Eren Üniversitesi Fen Bilimleri Dergisi*, vol. 9, no. 1, pp. 306–322, 2020, doi: 10.17798/bitlisfen.558113.
- [11] R. Acar and K. Saplıoğlu, "AKARSULARDAKİ SEDİMENT TAŞINIMININ YAPAY SİNİR AĞLARI VE ANFIS YÖNTEMLERİ KULLANILARAK TESPİTİ", *NÖHÜ Müh. Bilim. Derg.*, vol. 9, no. 1, pp. 437–450, 2020, doi: 10.28948/ngumuh.681208.
- [12] R. Acar and K. Saplıoğlu, Using the Particle Swarm Optimization (PSO) Algorithm for Baseflow Separation and Determining the Trends for the Yesilirmak River (North Turkey). *Russ. Meteorol. Hydrol.* 49, 40–51, 2024, <https://doi.org/10.3103/S1068373924010060>.
- [13] Güçlü, Y. S., Subyani, A. M., & Şen, Z. (2017). Regional fuzzy chain model for evapotranspiration estimation. *Journal of Hydrology*, 544, 233–241. <https://doi.org/10.1016/j.jhydrol.2016.11.045>

- [14] E. Çoban, "Makine Öğrenmesi Algoritmaları ile Yaz Sezonu Ortalama Akım Değerlerinin Tahmini," *J. Innov. Civ. Eng. Technol.*, vol. 6, no. 2, pp. 73–81, doi: 10.60093/jiciviltech.1497771.
- [15] A. Mosavi, F. S. Hosseini, B. Choubin, M. Goodarzi, and A. A. Dineva, "Groundwater salinity susceptibility mapping using classifier ensemble and Bayesian machine learning models," *IEEE Access*, vol. 8, pp. 145564–145576, 2020. doi: 10.1109/ACCESS.2020.3014908
- [16] A. N. Ahmed et al., "Machine learning methods for better water quality prediction," *J. Hydrol.*, vol. 578, p. 124084, 2019, doi: 10.1016/j.jhydrol.2019.124084.
- [17] N. Nasir et al., "Water quality classification using machine learning algorithms," *J. Water Process Eng.*, vol. 48, p. 102920, 2022, doi: 10.1016/j.jwpe.2022.102920.
- [18] B. Ouadi et al., "Optimizing silt density index prediction in water treatment systems using pressure-based Gradient Boosting hybridized with Salp Swarm Algorithm," *J. Water Process Eng.*, vol. 68, p. 106479, 2024, doi: 10.1016/j.jwpe.2024.106479.
- [19] U. Ejaz et al., "Monitoring the industrial waste polluted stream—Integrated analytics and machine learning for water quality index assessment," *J. Clean. Prod.*, vol. 450, p. 141877, 2024. doi: 10.1016/j.jclepro.2024.141877
- [20] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, 2001. doi: 10.1023/A:1010933404324
- [21] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, 1995.
- [22] Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to Linear Regression Analysis* (5th ed.). Wiley.
- [23] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, 1997. doi: 10.1006/jcss.1997.1504
- [24] Ahmed, U., Mumtaz, R., Anwar, H., Shah, A. A., Irfan, R., & García-Nieto, J. (2019). Efficient water quality prediction using supervised machine learning. *Water*, 11(11), 2210. <https://doi.org/10.3390/w11112210>
- [25] Mosavi, A., Hosseini, F. S., Choubin, B., Abdolshahnejad, M., Gharechae, H., Lahijanzadeh, A., & Dineva, A. A. (2020). Susceptibility prediction of groundwater hardness using ensemble machine learning models. *Water*, 12(10), 2770. <https://doi.org/10.3390/w12102770>
- [26] Nouraki, A., Alavi, M., Golabi, M., & Albaji, M. (2021). Prediction of water quality parameters using machine learning models: A case study of the Karun River, Iran. *Environmental Science and Pollution Research*, 28(40), 56792–56805. <https://doi.org/10.1007/s11356-021-14560-8>