OKU Fen Bilimleri Enstitüsü Dergisi
6(2): 1434-1451, 2023

OKU Journal of The Institute of Science and Technology, 6(2): 1434-1451, 2023

Osmaniye Korkut Ata Üniversitesi
Fen Bilimleri Enstitüsü
Dergisi

Osmaniye Korkut Ata University
Journal of The Institute of Science
and Technology

Estimating Personal Water Consumption Using Artificial Intelligence Methods

Derya BIRANT[1*], Irem CALMAZ[2], Irem OKUR[3]

[1,2,3]Dokuz Eylul University, Engineering Faculty, Department of Computer Engineering, 35390, Izmir

[1]https://orcid.org/0000-0003-3138-0432
[2]https://orcid.org/0000-0002-8569-4413
[3]https://orcid.org/0000-0002-7837-1624
***Corresponding author: derya@cs.deu.edu.tr**

Research Article

Article History:
Received: 03.06.2022
Accepted: 10.02.2023
Published online: 05.07.2023

Keywords:
Artificial intelligence
Machine learning
Water consumption
Water footprint
Regression

ABSTRACT

The estimation of water consumption is a crucial task in achieving global sustainability targets and addressing the long-term water needs of citizens. While some efforts have been done to estimate individual water footprints, there is still limited research in this area. To address this limitation, this article proposes a new artificial intelligence-based model, called *WaterAI*, to predict individuals' water consumption scores by taking into account indirect and direct water use through the water footprint indicator. It compares four different machine learning algorithms (linear regression, LASSO regression, gradient boosting, and extreme gradient boosting) to determine the best one for water consumption estimation. The data were collected with a questionnaire survey. The experimental results show that the proposed model can be successfully used to predict personal water consumption scores in an effective way.

Yapay Zeka Metotlarını Kullanarak Kişisel Su Tüketimi Tahminleme

Araştırma Makalesi

Makale Tarihçesi:
Geliş tarihi: 03.06.2022
Kabul tarihi:10.02.2023
Online Yayınlanma: 05.07.2023

Anahtar Kelimeler:
Yapay zeka
Makine öğrenmesi
Su tüketimi
Su ayak izi
Regresyon

ÖZ

Su tüketiminin tahmini, küresel sürdürülebilirlik hedeflerine ulaşmada ve vatandaşların uzun vadeli su ihtiyaçlarını karşılamada çok önemli bir görevdir. Bireysel su ayak izlerini tahmin etmek için bazı çalışmalar yapılmış olsa da, bu alanda hala sınırlı miktarda araştırma bulunmaktadır. Bu sınırı gidermeye yönelik olarak, bu makale, su ayak izi göstergesi aracılığıyla dolaylı ve doğrudan su kullanımını dikkate alarak bireylerin su tüketim puanlarını tahmin etmek için WaterAI adlı yeni bir yapay zeka tabanlı model önermektedir. Su tüketimi tahmini için en iyi modeli belirlemek adına dört farklı makine öğrenme algoritmasını (doğrusal regresyon, LASSO regresyonu, gradyan artırma ve aşırı gradyan artırma) karşılaştırmaktadır. Veriler bir anket çalışması ile toplanmıştır. Deneysel sonuçlar, önerilen modelin, kişisel su tüketim skorunu etkili bir şekilde tahmin etmek için başarılı bir şekilde kullanılabileceğini göstermektedir.

To Cite: Birant D., Calmaz İ., Okur İ. Estimating Personal Water Consumption Using Artificial Intelligence Methods. Osmaniye Korkut Ata Üniversitesi Fen Bilimleri Enstitüsü Dergisi 2023; 6(2): 1434-1451.

1. Introduction

"*Water*" is involved in all areas of life. Individuals consume a large amount of water from food production to daily life consumption. In the last few years, water demand is increasing whereas some groundwater and surface resources are decreasing. Considering the impact of the increasing population, increasing resource consumption, and climate change, water shortage is expected to affect some of the world's population. Therefore, the issue of water management has become of great significance in recent years in global systems. Water consumption estimation is a crucial task in an efficient management system. Individuals are unaware of their global consumption beyond thinking about the bill. Therefore, this research proposes an innovative mindfulness technique to monitor water consumption across an application.

The process for calculating the water footprint contains direct and indirect water usage. *Direct water consumption* (domestic water) refers to the amount of water directly used by an individual and includes everyday life activities such as showering, cooking, hand/car washing, tea/coffee drinking, and toilet flushing. *Indirect water consumption* (virtual water) stands for the water consumed for producing industrial goods, agricultural products, and energy. For example, it refers to the water that has been used to produce the food, paper, wood, clothing, and industrial goods consumed. This study considers both indirect and direct water consumption.

The water footprint of a process, product, or organization plays a fundamental role in the management of water consumption. However, individual water use events are also significant to reduce water consumption. Personal water consumption is influenced by many factors such as weather (air humidity and temperature), exposure frequency, daily habits, food-related behaviors, transportation modes, energy systems, and labor intensity, as well as household sizes. For raising awareness of water consumption, it is necessary to understand the concepts of water footprint and predict it by using a tool. Increasing awareness about individual water consumption via a tool is an important issue for water management and decision-making processes in line with the challenges that arose from water scarcity. With this motivation, this study aims to develop a tool to predict individual water consumption scores and then motivate people to change their behavior according to their scores.

The main contributions of this study can be summarized as follows. (i) It proposes a new artificial intelligent-based model, called *WaterAI*, to estimate the water consumption scores of urban residents by considering the indirect and direct water use through the water footprint concept. (ii) A new questionnaire survey was designed and conducted on residents in Turkey to collect quantitative data. (iii) This study is also original in that it compares four machine learning (ML) algorithms for personal water usage estimates, including linear regression (LR), LASSO regression, gradient boosting (GBoost), and extreme gradient boosting (XGBoost). The experimental results show that the proposed model can be successfully used to predict personal water consumption scores in an effective way.

The organization of this paper is as follows. Section 2 gives a brief survey of relevant studies. Section 3 explains the proposed approach. In addition, it also describes the machine learning algorithms used in this study. Furthermore, Section 3 gives detailed information about the questionnaire. Section 4 describes the experiments and presents the results obtained from various machine learning algorithms. Finally, Section 5 gives a general conclusion, as well as discusses directions for future work.

2. Related Work

Water footprint (WF) research is mainly at the product level (Brindha, 2020), process level (Li et al., 2021), organizational level (Kandananond, 2019), national level (ElFetyany et al., 2021), or individual level (Lee et al., 2019). In this study, we focused on personal water consumption since it plays a fundamental role in water management.

Recently, the issue of personal water footprint calculation has been studied in different countries such as Mexico (Lares-Michel et al., 2021), China (Pang et al., 2021), Saudi Arabia (Alqahtani et al., 2021), Spain (Gomez-Llanos, 2020), Taiwan (Lee, 2019), Iran (Sobhani et al., 2019), India (Harris et al., 2017) and Croatia (Stanic et al., 2015). Alropy et al. (2015) studied individual water footprints in Egypt. Their study aimed to estimate the components of the total water footprint, as well as its external and local water resources. The results of their study showed that most of the water consumption of an individual (90%) was caused by agricultural and food commodities. They also predicted the total demand for water for industrial, domestic, and agricultural use to efficiently manage water resources in Egypt.

Mahjabin et al. (2018) investigated the scaling of the WF of 65 mid-to-large-sized United States cities by utilizing both the social interaction network model of cities and empirical forecasts. They calculated the urban WF which tends to present sublinear scaling behavior with both gross domestic product and population. As a result, they reported that large cities are more productive and water footprint efficient than mid-sized ones.

It is important to explore the impacts of climate change on the WF. Mokhtar et al. (2021) used machine learning methods to model green and blue water footprints, including random forest, reduced error pruning tree, random tree, and additive regression. For these models, they analyzed six different scenarios with a combination of climate variables. The results reported in their study may assist in mitigation plans such as development plans for food security and policies for sustainable water use. Haida et al. (2019) also studied the relationship between the WF concept and climate change. They proposed a bottom-up approach in cooperation with a partner school in Austria. They aimed to assist young people to be aware of their daily habits and change some actions to reduce their water footprint. According to the findings, the total indirect personal water footprint can be achieved by reducing the consumption of mostly dairy products followed by cereal products.

Özbaş et al. (2022) aimed to determine the impact of the COVID-19 pandemic on individual water consumption in Turkey, and for this purpose, they investigated the differences between personal water footprints before and during the pandemic. According to their findings, the average water footprint has been calculated as 4178.42 L/day before the COVID-19 epidemic while it has been figured out as 4606.18 L/day during the epidemic period. This is because of the fact that the frequency of some activities has been changed, such as laundry washing numbers, the count of showers, and cooking frequencies.

Pang et al. (2021) used decision tree and logistic regression methods to predict the water footprint, but they only focused on food consumption. The individual data about the diet characteristics and food intake of residents were particularly analyzed. Obringer et al. (2022) used the Random Forest algorithm to predict intra-city residential water consumption patterns by integrating demographic and climate data. Arsene et al. (2022) presented a machine-learning approach for monitoring and predicting water consumption by proposing an Internet of Things-based (IoT) solution in households. First, they used the K-Means clustering algorithm to extract distinctive water consumption patterns, and then they tested four supervised learning algorithms (decision tree, random forest, the Dense algorithm, and recurrent neural network) to determine the best one. While Zanfei et al. (2022) proposed an ensemble neural network model to estimate drinking water consumption, Wei et al. (2022) used Random Forest to forecast irrigation water consumption.

The effects of demographic characteristics (i.e., gender, city, age) on water consumption estimation have been investigated. Alqahtani et al. (2021) used regression analysis to determine the most important social and economic factors affecting the total individual water footprint. The results indicated statistical differences between the average estimates of individual water footprint and the source, especially educational level, income, and family size. Pang et al. (2021) showed that income and education level were positively related to the dietary water footprint. They also revealed that males and urban residents with a higher body mass index (BMI) consumed more dietary water than females and rural those with a lower BMI. They also reported that age exhibited an inverted U-shaped influence. Harris et al. (2017) used Spearman's rank correlation matrix to assess the relationships between blue water footprint (WF) and socio-demographic characteristics. They reported that the blue WF was associated with gender since males consumed more than females for each food category. They also revealed that rural participants had a lower WF compared to urban. Socio-economic indicators were associated, with WF increasing with higher educational levels and higher standard of living index. On the other hand, age was negatively and independently associated with blue WF. Obringer et al. (2022) found positive relationships, with higher income often leading to higher water consumption. They also reported that family size was essential because the increased number of people within a household leads to more water consumption.

Table 1 presents a comparison between the Water Footprint Network (WFN) tool (Hoekstra 2009) and WaterAI implemented in this study. Our tool has the advantages of applying machine learning methods, designing for mobile platforms, and giving recommendations to users to reduce their water footprint. While WFN conducted the survey globally (all around the world), the questionnaire in this study was carried out locally (in Turkey). In the future, a graphical representation of the results could be provided, similar to the WFN tool.

Table 1. Comparison between the relevant work and our work.

Property	**Water Footprint Network**	**WaterAI (this study)**
Machine learning	-	+
Mobile platform	-	+
Scope	Global	Local
Type	Personal WF	Personal WF
Graphical results	+	-
Recommendation to reduce WF	-	+

3. Material and Methods

3.1. Proposed Approach

Environmental footprints can be defined as indicators that measure human impacts on natural or environmental resources. Recently, a number of environmental footprints have been considered in scientific studies to assess the impacts of humanity exerts on the environment such as carbon footprint, energy footprint, land footprint, ecological footprint, water footprint, material footprint, cropland footprint, and fishing footprint (Ewing et al., 2012). This study focuses on the water footprint.

The *water footprint* (WF) is an indicator of qualitative and quantitative water consumption that considers both indirect and direct water use by a producer or consumer Therefore, WF is an assessment of how much water is consumed. It gives explicit information about water appropriation for human activities or system operations. The water footprint was first proposed by Hoekstra in 2002 (Hoekstra and Hung, 2002). While working at the UNESCO-IHE Water Education Institute, he created the water footprint as a measure to evaluate the amount of water polluted or consumed to produce goods and services within a supply chain (Mekonnen and Hoekstra, 2010).

Water footprint (WF) can be calculated through two different approaches; bottom-up and top-down. A *bottom-up* approach refers to an analysis that includes the descriptions of individual processes. In this approach, each value chain link is individually explored and the direct/indirect water consumed to perform each activity and to produce each product are summed. Instead of calculating a national footprint, the bottom-up approach often uses local area (urban) data to calculate the WFs. The *top-down* approach covers different countries/regions and industrial sectors, providing a more comprehensive assessment of water consumption. It relies on economic valuation in a supply chain, inter-sectoral monetary transactions, and sectoral water consumption data. A top-down approach is a typical input-output analysis (IOA) and employs a multi-region input-output (MRIO) model. It has been often calculated at the national level by using global area data, including import and export trade data. The bottom-up approach is followed in this study since it is a popular approach in WF studies by its simplicity and relatively large data availability (Lee, 2019).

The water resources are divided into three types: blue, gray, and green. The total amount of direct and indirect water footprint includes these three resource types. The *blue water footprint* corresponds to the consumption of surface and groundwater resources. The food production industry and direct water consumption both contain a blue water footprint. The *green water footprint* refers to the direct use of rainwater or consumption of rainwater stored in the soil which can be termed as soil moisture. Farming, gardening, and forestry products are significantly affected by the green water footprint. The *gray water footprint* corresponds to the quantity of freshwater necessary to absorb pollutants in order to achieve particular water quality criteria.

An individual uses water directly and indirectly in their lives. The total consumption of water by the individual is calculated with the help of WF. Individual WF not only impresses locals but also affects cities and even nations. A person's dietary habits can affect water footprint results. For example, daily meat consumers use more water compared to vegetarians since meat production requires between 6 and 20 times more water than vegetables, fruits, and cereals (Bhagwat, 2019). Lares-Michel et al. (Lares-Michel et al., 2021) reported that 90% of an individual's water footprint comes from food production. From their daily showers to the jeans they wear every day, many human behaviors affect water resources. To increase awareness of water sustainability and individual water consumption, a tool that predicts the individual water footprint score can be developed. The tool can recommend solutions to users about how to save water according to their water consumption. Furthermore, using the tool, a person can be able to compare his/her water footprint score with the global water consumption score. A water footprint score can help people to understand their consumption behaviors and it can encourage them to change their lifestyles. For instance, people can observe their eating habits and how much food they waste in daily life.

The general architecture of the proposed approach, called *WaterAI*, is given in Figure 1. In the first step, the raw data, which was collected with a questionnaire survey, is retrieved from the database. In the data preprocessing step, the raw data is cleaned and normalized. After that, feature extraction and feature selection tasks are performed to obtain a proper and optimal feature set. In the next step, several machine learning algorithms are utilized to construct alternative models. In the evaluation step, the best model is selected to be used in deployment. Finally, in the last step, the model is used to predict the response for a given data.

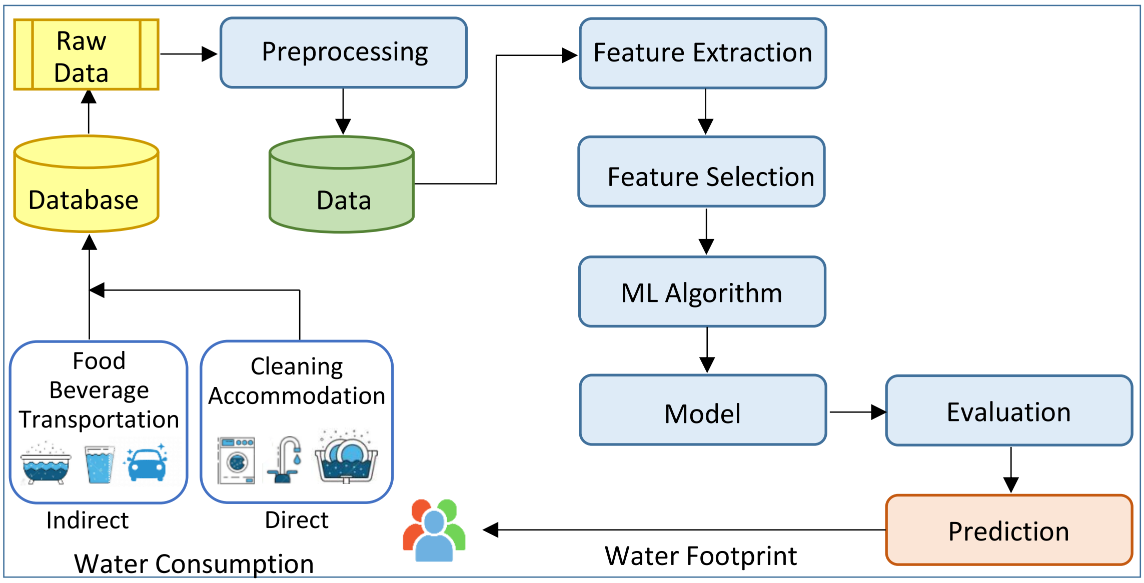

Figure 1. The general structure of the proposed WaterAI approach.

The main points of this study can be listed as follows. We propose an artificial intelligence-based model that tested several methods by considering both indirect and direct water consumption based on the following main features: food intake, beverage intake, accommodation-related info, cleaning habits, and transportation modes. In addition, a new questionnaire survey was designed and conducted on residents in Turkey to collect quantitative data. Furthermore, a mobile application was developed to give recommendations to users to reduce their water footprint.

3.2. Machine Learning Algorithms

Machine learning is a subfield of artificial intelligence that focuses on the application of algorithms that can generate patterns from data and create a predictive model. In this study, four different machine-learning algorithms were deployed. These algorithms are explained as follows.

Linear Regression: Linear regression is one of the supervised learning methods and builds a simple and interpretable regression-based model. It overcomes the overfitting well when using a cross-validation technique. In this study, a multiple linear regression model was constructed as given in Equation (1) (Su et al., 2012).

$$y = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n + \varepsilon \tag{1}$$

where y is the estimated value of the dependent variable, B_0 is the y-intercept of the regression line, B_n is the coefficient of the n-th independent variable (X_n), and ε is the residual term.

Least absolute shrinkage and selection operator (LASSO) Regression: LASSO regression (LR) is a statistical method that establishes a multivariate linear model between the set of x_i and the target y by minimizing the least-squares. The LR explores the optimal parameters of weight and bias by imposing an L1-norm regularization term on the objective function. This process enables the sum of the absolute values of the weights to be minimized. The effect of the regularization is adjusted by multiplying the λ-constant. LR has the advantage of setting the coefficients of non-important variables to zero during the analysis. In other words, the algorithm determines a subset of variables in which only the strongest ones remain in the model. Therefore, it is beneficial when the dataset is highly correlated or has high dimensionality. The LASSO estimator is examined in regression problems by Equation (2) (Kwon et al., 2013).

$$\hat{\beta}_{Lasso} = arg \min_{\beta} ||Y - X\beta||_2^2 /2n + \lambda||\beta||_1 \quad (2)$$

where $||\beta||_1 = \sum_J |\beta_J|$and λ > 0 is a parameter that is specified by the user.

Gradient Boosting: Gradient boosting is a supervised machine learning algorithm. It is an ensemble learning approach that builds a stronger predictive model by forming multiple weak learning models. It is one of the boosting strategies that tries to reduce the model's bias error. It is flexible with the help of several hyper-parameter tuning options. The purpose is to determine a function $F(x)$ that approximates the output variable based on the input variables' values by providing a loss function $L(y, F(x))$ with a minimum value. The Gradient Boosting implies that $F(x)$ has the expanded form as given in Equation (3) (Chen et al., 2013).

$$F(X) = \sum_{m=0}^{M} \rho_m f(x; \tau_m) \quad (3)$$

where f is the weak learner with a weight ρ and a parameter set τ, and M is the number of iterations.

Extreme Gradient Boosting (XGBoost): XGBoost uses a gradient-boosting approach for decision trees. As a result of the distributed and parallel computation, fast model construction is feasible. Therefore, it can handle large datasets and the training time is highly fast. In order to avoid over-fitting, XGBoost enables row and column sampling. The estimated output of the model can be formalized as given in Equation (4) (Zhang et al., 2021):

$$\widehat{y_i} = \sum_{k=1}^{K} f_k(x_i) \quad , f_k \in \Gamma \quad (4)$$

where $\widehat{y_i}$ is the predicted value, K is the number of regression trees, Γ is the space of trees, $f_k(x_i)$ is the estimation score of the k-th regression tree and x_i represents the i-th sample.

3.3. Questionnaire Survey

In this study, the data was collected by a questionnaire survey, which was designed and held by consumers in Turkey from November to December 2022. Table 2 presents the survey questions, as well as their categories, types, and the amount of water consumed for the corresponding event. The questionnaire consists of 23 water consumption questions, in addition to 7 demographic information questions. While 12 questions are related to indirect water consumption, the rest of them are regarding direct water consumption. While most questions are about indoor activities, only two of them are related to outdoor activities. To evaluate the individual water footprint scores, the consumers are mainly questioned about daily and weekly habits. Individual water consumption was determined using five classes: food, beverage, accommodation, cleaning, and transportation. Respondents were asked to provide information about “food” consumption (weekly food intakes, such as meat, cereal, dairy, eggs, vegetables, fruits, and sugar), “beverage” consumption (the number of cups of tea and coffee per day), “accommodation” consumption (the number of showers, hand washing, teeth brushing, shaving, flushing, and garden watering), "cleaning" consumption (the number of times laundry and dishwasher are done) and “transportation” consumption (the number of car washing).

In order to ensure the validity and reliability of the study, the questions in the survey were selected from the questionnaires presented in the previous studies (Lee 2019; Hoekstra 2009). The water consumption values given in Table 2 were taken from (Mekonnen and Hoekstra, 2010).

Table 2. Survey questions and corresponding water consumption.

ID	Category	Type	Class	Question	Water Consumption
1	Indoor	Indirect	Food	Dietary habit (normal, vegetarian, vegan)	Normal > Vegetarian > Vegan
2	Indoor	Indirect	Food	Meat product consumption (weekly)	Chicken Meat: 4325 liter/kg Beef: 15415 liter/kg Sheep/Goat Meat: 8763 liter/kg
3	Indoor	Indirect	Food	Cereal and pulses product consumption (weekly)	Cereals: 1644 liter/kg Pulses: 4055 liter/kg
4	Indoor	Indirect	Food	Dairy product consumption (weekly)	Milk: 1020 liter/kg Butter: 5553 liter/kg Cheese: 3178 liter/kg
5	Indoor	Indirect	Food	Number of eggs consumption (weekly)	196 liter for each egg
6	Indoor	Indirect	Food	Vegetable consumption (weekly)	322 liter/kg
7	Indoor	Indirect	Food	Fruits consumption (weekly)	962 liter/kg
8	Indoor	Indirect	Food	Sugar and sweet consumption (daily)	Low = 140 Average = 490 High = 840
9	Indoor	Indirect	Beverage	How many cups of coffee do you take per day?	132 liter per cup

10	Indoor	Indirect	Beverage	How many cups of tea do you take per day?	27 liter per cup
11	Indoor	Direct	Accommodation	How many showers do you take in a week?	Number of times
12	Indoor	Direct	Accommodation	What is the avg. length of each shower? (minute)	12 liter per minute
13	Indoor	Direct	Accommodation	How many times per day do you wash your hands?	Number of times
14	Indoor	Direct	Accommodation	How many times per day do you brush your teeth?	Number of times
15	Indoor	Direct	Accommodation	How many times per week do you shave?	Number of times
16	Indoor	Direct	Accommodation	Do you leave the tap running when brushing your teeth and shaving?	Yes: 5 liters/minute and 4 minutes per event No: 1 liter per event and 4 minutes per event
17	Indoor	Direct	Accommodation	How many times per day do you flush?	5 liters for a dual flush toilet
18	Indoor	Direct	Cleaning	How many loads of laundry do you do in a week on average?	40 liters per cycle of washing
19	Indoor	Direct	Cleaning	If you wash your dishes by hand, how many times are dishes washed each day?	Number of times
20	Indoor	Direct	Cleaning	How long does the water run during each wash?	12 liters per minute
21	Indoor	Direct	Cleaning	If you have a dishwasher, how many times is it used each week?	35 liters per cycle of washing
22	Outdoor	Indirect	Transportation	If you have a car, how many times per month do you wash your car?	200 liters per event
23	Outdoor	Indirect	Accommodation	If you have a garden, how many times do you water your garden each week?	(Duration * 8 liters) per event

The total water consumption is calculated by summing direct and indirect consumptions as given in Equations (5) - (7).

$$WF_{Direct} = \text{Shower} + \text{Laundry} + \text{Tap Water} + \text{Toilet} + \text{Dish} + \text{Car} + \text{Garden} \tag{5}$$

$$WF_{Indirect} = \text{Meat} + \text{Cereal} + \text{Dairy} + \text{Egg} + \text{Vegetable} + \text{Fruit} + \text{Sugar} + \text{Coffee} + \text{Tea} \tag{6}$$

$$WF_{Total} = WF_{Direct} + WF_{Indirect} \tag{7}$$

Table 3 gives demographic information about the respondents such as their genders, education levels, ages, and income levels. In total, 546 surveys were conducted; 60.8% of respondents were female, 50.5% were 21-40 years old, and 72.9% of participants had Bachelor's degrees. The number of persons in the household is considered as the unit of reference for the consumption of water. The respondents live in various cities in Turkey such as Izmir, Antalya, Hatay, Istanbul, Ankara, and so on.

Since the survey data was collected from many people around the country under only several restrictions, the dataset had some missing and inconsistent values. For this reason, the acquired data was passed through the data-preprocessing step. First, the dataset was filtered by dropping several inconsistent rows by a human supervisor, who removed the ones that did not reflect probable content. Second, the missing values were filled by using the "*mean*" strategy in the Simple Imputer technique from the Scikit-Learn Python package (Pedregosa et al., 2011). In the feature selection phase, more informative features were chosen to improve the model's performance and accuracy. To implement this, the univariate linear regression test technique was used (Kramer, 2016), which basically gives priority scores to the features. A higher score means that the feature has a larger effect on the machine-learning model. After that, the Local Outlier Factor technique (Breunig et al., 2000) was utilized to drop the outliers that did not fit the pattern of the dataset. After this process, the number of rows in the dataset decreased from 546 to 533. Furthermore, the Label Encoding technique (Pedregosa et al., 2011) was used to convert the categorical variables into numerical label ones. Water consumption scores were calculated as the target column, by considering the water uses of individuals through the water footprint indicator.

Table 3. Demographic characteristics.

Information	Value	Count	Percentage (%)
Gender	Female	332	60.8
	Male	214	39.2
Age	20 years or younger	35	6.4
	21–40 years old	276	50.5
	41–60 years old	179	32.7
	61–80 years old	56	10.2
Education	Secondary school	8	1.5
	High school	41	7.5
	Bachelor's degree	398	72.9
	Master's degree	59	10.8
	Doctorate’s degree	39	7.1
Place of residence	Ankara	29	5.3
	Antalya	127	23.3
	Hatay	88	16.1
	İzmir	128	23.4
	İstanbul	46	8.4
	Other	128	23.4
Income	< 5,000 TL	267	48.9
	5,000 – 15,000 TL	239	43.7
	> 15,000 TL	40	7.3
Number of persons in	1	60	11

your household	2	139	25.5
	3	150	27.5
	4	154	28.2
	5	35	6.4
	6	7	1.3
	7	1	0.2
Water Bill	<100	284	52
	>= 100	262	47.9

4. Experimental Studies

The machine learning part of the study was implemented in Python by using various libraries such as NumPy, Pandas, Scikit-Learn, and Matplotlib. To increase the performance of the models, the GridSearchCV method was used to carry out the hyperparameter tuning for the GBoost and XGBoost algorithms as given in Table 4. The default values were set for all other parameters. Fast Library for Automated Machine Learning (FLAML) was utilized to automatically determine the most accurate machine learning model at a low computational cost. The alternative methods were tested by using the 10-fold cross-validation technique. Hence, the dataset was split into 10 equal subsets, and each time, one of the sets is used for testing, while the rest ones are utilized for training.

Table 4. Parameter settings.

Algorithm	Parameter	Range	Selected value
Extreme Gradient Boosting	learning_rate	0.1, 0.2, 0.3, 0.4	0.1
	n_estimators	100, 500, 1000, 1500	1000
	max_depth	3, 5, 7, 9	7
	subsample	0.5, 0.7, 1.0	0.7
Gradient Boosting	learning_rate	0.1, 0.2, 0.3, 0.4	0.1
	n_estimators	100, 500, 1000, 1500	100
	max_depth	3, 5, 7, 9	3
	subsample	0.5, 0.7, 1.0	1.0
Linear Regression	fit_intercept	True	True
LASSO Regression	alpha	1.0	1.0

4.1. Evaluation Metrics

The evaluation process for constructing an effective model is the core of a machine learning system. The metrics give a result that is important for the reliability of the model. Different evaluation metrics are used for different problems. In this study, the experimental results are evaluated in terms of four metrics: mean absolute error (MAE), mean square error (MSE), root mean square error (RMSE), and coefficient of determination (R^2). MAE is the average of all absolute errors, as given in Equation (8).

It shows the average amount of inaccuracy expected from the prediction. MSE is the metric that presents the cumulative squared error between the actual and predicted values, as given in Equation (9). RMSE indicates how closely the estimated values correspond with the actual values and its formula is given in Equation (10). It is one of the most widely-used evaluation metrics for measuring the quality of the model. R^2 is used to measure the performance of the model by explaining the relationship between the independent and dependent variables and its formula is given in Equation (11).

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |P_i - O_i| \tag{8}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (P_i - O_i)^2 \tag{9}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (P_i - O_i)^2} \tag{10}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(O_i - P_i)^2}{\sum_{i=1}^{n}(O_i - P_i)^2} \tag{11}$$

where *n* is the number of samples, P_i is the predicted value, and O_i is the observed value.

4.2. Experimental Results

Table 5 shows the comparative results of four machine learning algorithms (Linear Regression, LASSO Regression, Gradient Boosting, and Extreme Gradient Boosting) in terms of three metrics. A lower MSE, MAE, or RMSE value means greater accuracy. It is possible to see from the results that the LASSO Regression has very successful prediction outcomes. LASSO regression produced the lowest RMSE value (13.87) which means that it is the best one. It is probably because of the fact that LASSO regression has the advantage of improving accuracy by performing variable selection depending on the magnitude of the tuning parameter, therefore, setting the non-significant coefficients to zero during the regression analysis. It was observed that the XGBoost algorithm was the least successful one in predicting water consumption scores compared to other algorithms.

Table 5. The performances of machine learning algorithms in terms of MSE, MAE, and RMSE.

Algorithm	**MSE**	**MAE**	**RMSE**
Linear Regression	193.71	10.49	13.91
LASSO Regression	192.62	10.39	13.87
Gradient Boosting	195.51	09.04	13.98
Extreme Gradient Boosting	323.35	11.49	17.98

Figure 2 shows the R^2 values for each machine-learning algorithm. The results explain how the regression model is good at predicting personal water consumption scores. It quantifies whether the model fits the dataset well or not, where the highest R^2 value indicates the best fit. Out of all the implemented machine learning methods, the LASSO Regression algorithm seems to have the best prediction performance (0.85). The main reason behind this achievement is that LASSO has the advantage of making feature selections by reducing the coefficients of unimportant features. The outcome of the method is of a highly satisfactory level, therefore, it indicates that the model has validity in predicting water consumption scores of people.

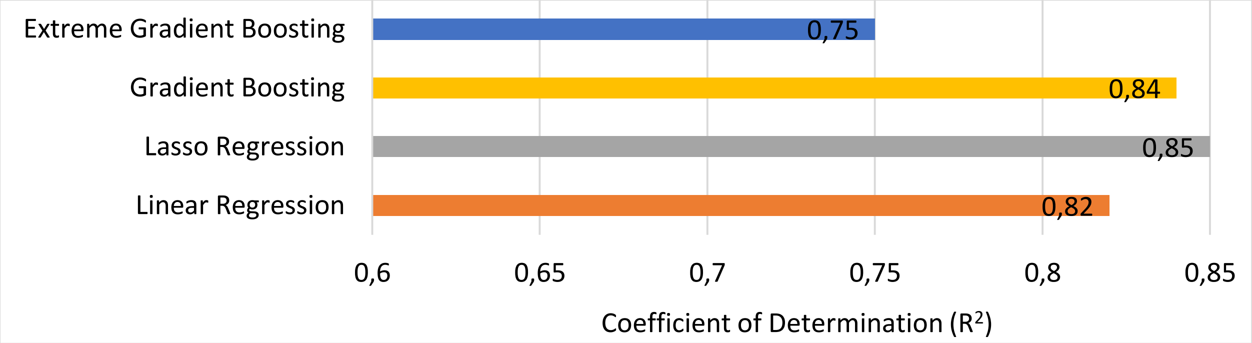

Figure 2. Comparison of machine learning algorithms in terms of coefficient of determination (R^2).

4.3. Discussion

In this study, four machine learning algorithms (LR, LASSO, GBoost, and XGBoost) were applied to the dataset and compared in terms of four metrics: MSE, MAE, RMSE, and R^2. The best one is a model that has a high R^2 value and small MSE, MAE, and RMSE values. These ML algorithms were selected since they are among the widely-used and popular regression algorithms (Carrera and Kim, 2020; Moscato et al., 2022). They have various advantages such as high efficiency, robustness to overfitting, and low computational cost. The other main reason behind the selection of these algorithms is to be able to test both single and ensemble-based methods.

According to the experimental results, the following conclusions were drawn:

- Machine learning algorithms successfully predicted personal water consumption scores with low error values (i.e., MAE < 11.5). This can be interpreted that they can be used to handle water footprint prediction problems.

- The lowest MSE (192.62) and RMSE (13.87) values were produced by LASSO regression. In addition, the highest R^2 value (0.85) was also achieved by the same algorithm. This means that it is the most robust and responsive model.
- To be able to predict the water consumption score of a person, both direct and indirect water usage may be taken into account through the water footprint indicator.
- When making a prediction, the following information can be considered: food intake (meat, cereal, dairy, eggs, vegetables, fruits, and sugar), beverage intake (tea and coffee), accommodation-related info (showers, hand washing, teeth brushing, shaving, flushing, and garden watering), cleaning habits (laundry and dishwasher) and transportation modes (car).

5. Conclusion and Future Work

The prediction of individual water consumption is important for the conservation of water for future generations. This paper proposes a new artificial intelligent-based model, called *WaterAI*, to estimate the water consumption scores of urban residents by considering the indirect and direct water use through the water footprint concept. A new survey study was designed and conducted on individuals living in Turkey. This study compares four different machine learning algorithms to determine the best one for water consumption estimation, including linear regression, LASSO regression, gradient boosting, and extreme gradient boosting. The experimental results show that the proposed model can be successfully used to predict personal water consumption scores in an effective way.

In future work, similar intelligent models can be constructed to estimate other environmental footprints such as energy footprint, land footprint, ecological footprint, material footprint, cropland footprint, and fishing footprint.

Conflict of Interest

The authors declare that there is no conflict of interest.

Author’s Contributions

The contribution of the authors is equal.

References

Alropy I., Kotb A., Al-Hindi A. An economic study of the role of foreign trade in water demand management in the Arab Republic of Egypt according to the concept of virtual water. Egyptian Journal of Agricultural Economics 2015; 25(1): 219-232.

Alqahtani SH., Alropy ET., Kotb AA., Alaagib SEB. Estimation of the standard model of the water footprint of individuals in the Kingdom of Saudi Arabia. Arabian Journal of Geosciences 2021; 14: 1-12.

Arsene D., Predescu A., Pahont B., Chiru CG., Apostol ES., Truica CO. Advanced strategies for monitoring water consumption patterns in households based on IoT and machine learning. Water 2022; 14: 1-20.

Bhagwat VR. Food safety and human health - Safety of water used in food production. India: Academic Press; 2019.

Breunig MM., Kriegel HP., Ng RT., Sander J. LOF: identifying density-based local outliers. In Proceedings of the ACM SIGMOD International Conference on Management of Data 2000; 29(2): 93-104.

Brindha K. Virtual water flows, water footprint and water savings from the trade of crop and livestock products of Germany. Water and Environment Journal 2020; 34: 656-668.

Chen Y., Jia Z., Mercola D., Xie X. A gradient boosting algorithm for survival analysis via direct optimization of concordance index. Computational and Mathematical Methods in Medicine 2013; 2013: 1-8.

ElFetyany M., Farag H., Ghany SHAE. Assessment of national water footprint versus water availability - Case study for Egypt. Alexandria Engineering Journal 2021; 60: 3577–3585.

Ewing BR., Hawkins TR., Wiedmann TO., Galli A., Ercin AE., Weinzettel J., Steen-Olsen K. Integrating ecological and water footprint accounting in a multi-regional input–output framework. Ecological Indicators 2012; 23: 1-8.

Gómez-Llanos E., Durán-Barroso P., Robina-Ramírez R. Analysis of consumer awareness of sustainable water consumption by the water footprint concept. Science of The Total Environment 2020; 721: 1-11.

Haida, C., Chapagain, AK., Rauch, W., Riede, M., Schneider, K. From water footprint to climate change adaptation: Capacity development with teenagers to save water. Land Use Policy 2019; 80: 456-463.

Harris F., Green RF., Joy EJM., Kayatz B., Haines A., Dangour AD. The water use of Indian diets and socio-demographic factors related to dietary blue water footprint. Science of The Total Environment 2017; 587: 128-136.

Hoekstra AY. Human appropriation of natural capital: A comparison of ecological footprint and water footprint analysis. Ecological Economics 2009; 68(7): 1963-1974.

Hoekstra AY., Hung PQ. Virtual water trade: a quantification of virtual water flows between nations in relation to international crop trade. UNESCO-IHE Delft - Value of Water Research Report Series 2002; 11: 1-120.

Kandananond K. The application of water footprint and six-sigma method to reduce the water consumption in an organization. International Journal of Geomate 2019; 17(61): 21-27.

Kramer O. Scikit-Learn: machine learning for evolution strategies. Studies in Big Data 2016; 20: 45-53, Springer, Cham.

Kwon S., Han S., Lee S. A small review and further studies on the LASSO. Journal of the Korean Data & Information Science Society 2013; 24: 1077-1088.

Lares-Michel M., Housni FE., Cervantes VGA., Carrillo P., Nava RMM., Cañedo CL. Eat well to fight obesity and save water: the water footprint of different diets and caloric intake and its relationship with adiposity. Frontiers in Nutrition 2021; 8: 1-18.

Lee YJ. Ecological footprint and water footprint of Taipei, Sustainability 2019; 11: 1-16.

Lee YJ., Tung CM., Lee PR., Lin SC. Personal water footprint in Taiwan: a case study of Yunlin county. Sustainability 2016; 8: 1-12.

Li X., Ren J., Wu Z., Wu X., Ding X. Development of a novel process-level water footprint assessment for ettextile production based on modularity. Journal of Cleaner Production 2021; 291: 1-12.

Mahjabin T., Garcia S., Grady C., Mejia A. Large cities get more for less: Water footprint efficiency across the US. PloS One 2018; 13(8): 1-17.

Mekonnen MM., Hoekstra AY. The green, blue and grey water footprint of farm animals and animal products, Research Report Series 2010; 48: 1-50.

Mokhtar A., He H., He W., Elbeltagi A., Maroufpoor S., Azad N., He H., Alsafadi K., Gyasi-Agyei Y., He W. Estimation of the rice water footprint based on machine learning algorithms. Computers and Electronics in Agriculture 2021; 191: 1-15.

Obringer R., Nateghi R., Ma Z., Kumar R. Improving the interpretation of data-driven water consumption models via the use of social norms. Journal of Water Resources Planning and Management 2022, 148: 1-12.

Özbaş EE., Akın Ö., Güneysu S., Özcan HK., Öngen A. Changes occurring in consumption habits of people during COVID-19 pandemic and the water footprint. Environment, Development and Sustainability 2022; 24(6): 8504-8520.

Pang Z., Yan D., Wang T., Kong Y. Disparities and drivers of the water footprint of food consumption in China. Environmental Science and Pollution Research International 2021; 28(44): 62461-62473.

Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondle M., Prettenhofer P., Weiss R., Dubourg V., Vanerplas J., Passos A., Cournapeau D., Brucher M., Duchesnay MPE. Scikit-learn: machine learning in Python. Journal of Machine Learning Research 2011; 12: 2825-2830.

Sobhani SR., Rezazadeh A., Omidvar N., Eini-Zinab H. Healthy diet: a step toward a sustainable diet by reducing water footprint. Journal of the Science of Food and Agriculture 2019; 99(8): 3769-3775.

Stanic S., Spetic M., Buzov I. The water footprint of an individual: a hidden dimension of sustainability. International Journal of Interdisciplinary Environmental Studies 2015; 10(3): 13-25.

Su X., Yan X., Tsai CL. Linear regression. WIREs Computational Statistics 2012; 4: 275-294.

Wei S., Xu T., Niu GY., Zeng R. estimating irrigation water consumption using machine learning and remote sensing data in Kansas high plains. Remote Sensing 2022; 14: 1-15.

Zanfei A., Menapace A., Granata F., Gargano R. An ensemble neural network model to forecast drinking water consumption. Journal of Water Resources Planning and Management 2022; 148: 1-10.

Zhang W., Wu C., Zhong H., Li Y., Wang L. Prediction of undrained shear strength using extreme gradient boosting and random forest based on Bayesian optimization. Geoscience Frontiers 2021; 12: 469-477.