A Novel Machine Learning Framework for Advanced Driving Force Analysis of Individuals' Dietary Water Footprint

Huang, Kai; Wang, Dong; Kapelan, Zoran

DOI
10.1029/ 2024EF005061

Publication date
2025

Document Version
Final published version

Published in
Earth's Future

Citation (APA)
Huang, K., Wang, D., & Kapelan, Z. (2025). A Novel Machine Learning Framework for Advanced Driving Force Analysis of Individuals' Dietary Water Footprint. *Earth's Future*, *13*(12), Article e2024EF005061. https://doi.org/10.1029/ 2024EF005061

AGU ADVANCING EARTH AND SPACE SCIENCES

Earth's Future

RESEARCH ARTICLE
10.1029/2024EF005061

Special Collection:
Advancing Interpretable AI/ML Methods for Deeper Insights and Mechanistic Understanding in Earth Sciences: Beyond Predictive Capabilities

A Novel Machine Learning Framework for Advanced Driving Force Analysis of Individuals' Dietary Water Footprint

Kai Huang[1], Dong Wang[2], and Zoran Kapelan[2]

[1]Beijing Key Laboratory for Source Control Technology of Water Pollution, College of Environmental Science and Engineering, Beijing Forestry University, Beijing, China, [2]Department of Water Management, Faculty of Civil Engineering and Geosciences, Delft University of Technology, Delft, The Netherlands

Key Points:

- We examine the driving factors of individuals' dietary water footprint (WF) using all the involved machine learning (ML)-related techniques
- Income level, urbanization level, education level, and gender emerge as the top four influential features on dietary WF in descending order
- The priority groups for dietary WF reduction interventions are identified as high-income, urban, highly educated, and male residents

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

K. Huang,
huangkai@bjfu.edu.cn

Citation:

Huang, K., Wang, D., & Kapelan, Z. (2025). A novel machine learning framework for advanced driving force analysis of individuals' dietary water footprint. *Earth's Future*, *13*, e2024EF005061. https://doi.org/10.1029/2024EF005061

Received 6 JUL 2024
Accepted 13 NOV 2025

Author Contributions:

Conceptualization: Kai Huang, Dong Wang, Zoran Kapelan
Funding acquisition: Kai Huang
Investigation: Kai Huang, Dong Wang
Methodology: Kai Huang, Dong Wang
Project administration: Zoran Kapelan
Resources: Kai Huang, Zoran Kapelan

Abstract Addressing water scarcity requires significant attention to reducing water footprint (WF) related to food consumption. Since individuals' dietary behavior is largely influenced by their demographic and anthropometric attributes, it is crucial to identify individuals who have a high dietary WF and prioritize them as the focus of policies. Several studies analyzing the driving factors behind dietary WF exist but have multiple limitations. These include the statistical models with rather modest performances, lack of rigorous sensitivity analysis/feature importance (FI) analysis, and lack of generalization ability. Here, we developed a novel ML-based framework for analyzing the driving forces behind dietary WF. The framework incorporated three machine learning (ML) models (Extra-Trees (ET), Histogram-based Gradient Boosting (HGB), and eXtreme Gradient Boosting (XGB)) and an ML explanation approach Shapley Additive exPlanations (SHAP). This framework was applied to a case study on Chinese inhabitants. The derived results validated the proposed framework and demonstrated ML's superiority over conventional statistical methods. XGB was identified as the optimal model as it effectively captured the variability in the data and showed good generalization performance. The FI analysis for XGB revealed the most influential features on dietary WF, with income level, urbanization level, education level, and gender emerging as the top four features in descending order. Through the subsequent SHAP dependence analysis, the priority groups for dietary WF reduction interventions were identified as high-income residents, urban residents, highly educated residents, and male residents. In light of these findings and their underlying causes, the paper concluded with a set of policy recommendations.

Plain Language Summary Addressing water scarcity requires significant attention to reducing water footprint (WF) related to food consumption. People's diets are largely influenced by their age, gender, and income level, etc. So, it's important to identify people who have a high dietary WF and focus on them when creating policies to reduce dietary WF. There are already some studies that look at the factors behind dietary WF, but they have a few limitations. These include the development of statistical models with rather modest performances, lack of rigorous sensitivity analysis/feature importance (FI) analysis, and lack of generalization ability. Motivated by these gaps, a new machine learning-based framework was developed for analyzing the driving forces behind dietary WF and then was applied to a case study on Chinese inhabitants. The results show that the top four factors associated with dietary WF are income level, urbanization level, education level, and gender, in that order. The priority groups for dietary WF reduction interventions were identified as high-income, urban, highly educated, and male residents. In light of these findings and their underlying causes, the paper concluded with a set of policy recommendations.

1. Introduction

Water scarcity is a significant global challenge affecting many regions around the world. Freshwater resources are limited and unevenly distributed, and increasing demand for water due to population growth, urbanization, industrialization and climate change impacts are exacerbating water scarcity (Liu et al., 2017; Mekonnen & Hoekstra, 2016). It is well recognized that agriculture, which includes crop production, livestock farming, and fisheries, is a major water user, accounting for a significant portion of global water withdrawals. According to the Food and Agriculture Organization of the United Nations (FAO), about 70% of global freshwater withdrawals are used for agriculture, making it the largest sectoral water user (FAO, 2017). However, water consumption associated with food goes beyond just the on-farm activities, as it also involves water used in food processing,

Supervision: Zoran Kapelan
Visualization: Dong Wang
Writing – original draft: Dong Wang
Writing – review & editing: Kai Huang, Zoran Kapelan

transportation, and distribution throughout the supply chain. Thus, reducing food-related water consumption is of great significance for combatting water scarcity.

To comprehensively quantify the amount of water associated with food, the widely recognized concept, water footprint (WF), is adopted in this study. WF is an indicator of the total volume of freshwater used directly and indirectly in the entire supply chain of goods or services, and consists of three WF types: blue, green and grey (Hoekstra et al., 2011). It is evident that implementing advanced water-efficient technologies in the food supply chain, for example, drip irrigation and precision agriculture technologies, can help lower WF on the food production side (Bwambale et al., 2022; Seyedmohammadi et al., 2016; Shafi et al., 2019). However, considering the holistic approach of integrated and sustainable water resources management for society as a whole, interventions on the food consumption side are equally, if not more, important. Consumer demand and preferences play a significant role in shaping the food system and influencing water use. Consumers have the power to drive changes in the food industry by making choices that align with WF reduction, such as opting for foods with lower WF, supporting local and sustainable food systems, and reducing food waste (Kim et al., 2020; Vanham, Hoekstra, & Bidoglio, 2013; Vanham, Mekonnen, & Hoekstra, 2013; Vanham et al., 2017). Hence, it is crucial to identify characteristics that are strongly associated with an individual's dietary WF. Individuals are characterized by their demographic and anthropometric attributes such as age, gender, income, education level, weight, and height, and those attributes are closely associated with their dietary behavior (Amin et al., 2008; Han et al., 2019; Sodjinou et al., 2009; Yau et al., 2020). In addition, those attributes are often collected through surveys, censuses, or administrative data, and can be readily accessible to policymakers and researchers. Hence, investigating the relationship between individuals' dietary WF and their demographic and anthropometric characteristics provides the policymakers with necessary information to tailor policies and efficiently allocate relevant resources for dietary WF reduction.

Several studies have investigated the relationship between individuals' dietary WF and their demographic and anthropometric characteristics. For example, Li et al. (2021) used Multiple Regression (MR) to examine how residents' income influences their WF of food consumption in the urban Guangdong province, China. The results showed that the increase in income among urban inhabitants had a notable, positive impact on their WF of food consumption, with the extent of the impact differing across different income groups. He et al. (2021) used MR to assess how the socio-economic status of US individuals affected greenhouse gas emissions, blue WF, land use, and energy consumption throughout supply chains. The results showed that both high income and high education levels tended to contribute to high blue WF. Travassos et al. (2020) analyzed the contribution of Brazilian diets to environmental impact in terms of carbon, water, and ecological footprint. The results from T-test and ANOVA showed that males had higher WF than females, white race had higher WF than others, the urban population had lower WF than rural one, people older than 60 had lower WF than younger people, and higher education degrees generally led to lower WF. Souissi et al. (2022) employed an MR model to gain a deeper understanding of the determinants of dietary WF among Tunisian residents. The results showed that the smaller household size, higher income, and living in developed and coastal cities contributed to a higher dietary WF of the residents. Harris et al. (2017) aimed to assess water use in Indian diets and used Mixed Effects Linear Regression to identify socio-demographic factors related to the dietary blue WF. The results revealed that higher education level and income were positively correlated with the blue WF, whereas age showed a negative correlation. Furthermore, urban households exhibited a higher blue WF compared to rural households, and males had a higher blue WF compared to females. Karaçil Ermumcu et al. (2023) assessed the total WF (in which dietary WF is dominant) of individuals in Turkey, taking into account their age, gender, and body weight. The results from the Mann-Whitney U and Kruskal Wallis tests showed that obese individuals had significantly higher total WF compared to their underweight and normal-weight counterparts.

However, there are multiple gaps in the above papers. Firstly, all the papers use only statistical tests or models. However, many real-world data sets do not meet the assumptions required for using statistical tests and models, which can cause Type I and Type II errors (Osborne & Waters, 2002; Verma & Abdel-Salam, 2019). Besides, all the regression models show coefficient of determination (R^2) values lower than 0.5. The low R^2 values imply that the variability in the data (especially for the data sets of relatively large size) is too complex to be captured by these models. All of this questions the credibility of the identified relationships in the models. Secondly, these papers fail to conduct rigorous sensitivity analysis/feature importance (FI) analysis that yields the exact order of the features' impact, without which the efficiency and effectiveness of interventions for dietary WF reduction

cannot be guaranteed. Thirdly, previous studies show limited generalization ability, which may be related to data quality or inherent limitations in the models.

Given the gaps mentioned above, we propose to employ Machine Learning (ML) for dietary WF (given the specific objectives of the study, we only consider overall dietary WF) driving force analysis for the following reasons. Firstly, most ML models usually do not make strict assumptions about the probability distribution of the data. Tree-based ML models are less susceptible to multicollinearity, missing values, and outliers, especially these three frontier ML models selected for this study with preprocessed data (Hastie, 2009). Secondly, ML models are more flexible and capable of capturing complex relationships between model inputs and outputs, especially for large-size data sets. Thirdly, FI measures for tree-based ML models are based on sound theory and criteria hence are clear and intuitive to use. Fourthly, it is compulsory in ML-based modeling to test the trained models on a novel data set. Two existing papers incorporated ML to carry out WF driving force analysis for individuals. Nevertheless, the ML implementations in both of these studies have multiple deficiencies. In one study (Liang et al., 2020), the model test procedure was not included and the model performance on the training data was not reported. In another study (Pang et al., 2021), the decision tree model was applied to measure the dietary WF with accuracies of 0.0738 and 0.0693 on the training and test data sets. These accuracy values indicate the proportion of correctly classified instances out of the total data set.

Motivated by the above knowledge gaps, a novel, systematic, and rigorous ML framework for WF driving force analysis is proposed in this study. The framework covers the whole pipeline, from formulating the WF domain problem into an ML problem, to data preprocessing, model building and selection, model explanation, and finally applying the explanation results back to the WF domain context. In the framework, the ML models implemented are Histogram-based Gradient Boosting (HGB), eXtreme Gradient Boosting (XGB), and Extremely Randomized Trees (ET). The model performances were assessed using multiple metrics, including the primary metrics: the area under the precision-recall curve (AUC_PR), complemented by Accuracy, Precision, Recall, and F1-score. The FI measures are Permutation Importance (PI) and SHapley Additive exPlanations (SHAP), and the detailed influence patterns of features are also obtained by SHAP. To our best knowledge, this is the first time these techniques are implemented in WF-related research.

2. Data

2.1. Data Source, Retrieval, and Water Footprint Accounting

The data used in this study are from the public database of an ongoing collaborative project, China Health and Nutrition Survey (CHNS), conducted by the University of North Carolina and the Chinese Center for Disease Control. Established in 1989, the survey was designed to examine the effects of health and nutrition policies and programs, as well as social and economic changes, on the health and nutrition of the Chinese population. CHNS is a very large and comprehensive database, which includes major public health risk factors and health outcomes, social, economic, and demographic factors in detail at the individual, household, and community levels (Popkin et al., 2010).

Given the objectives of this study, the data on inhabitants' income, education, place of residence, age, gender, weight, height, and food intake were retrieved from the database. It is worth mentioning that we calculated the Body Mass Index (BMI) based on the height and weight data and incorporated it as an additional feature in the retrieved data set. Despite the regular updates made to the database, the publicly available data on individuals' food intake was only up to date until 2011. As a result, the entire data set used in this study was confined to the time period between 2000 and 2011, specifically, 2000, 2004, 2006, 2009, and 2011. The original data of individuals were embedded in different files with redundant information, missing values, bad values, and layout inconsistency. Thus, after being retrieved from the CHNS database, the data were cleaned and reorganized for the further analysis and modeling. Thousands of different food types were covered in the retrieved data. The food intake data comprised food consumed both away from home and at home on a daily basis, with every individual having 3 days of food intake records. An instance of the data was defined as a unique combination of an individual's ID, year, day number, and food type. Therefore, the original retrieved data consisted of over 1.8 million instances. Nevertheless, the following measures were adopted to compress the data and enhance their usability in alignment with the objectives of this study. Firstly, the average daily food intake for each individual was calculated and multiplied by 365 to obtain the annual food intake in order to align it with the income time unit. Secondly, the food intake amounts were summed for each individual in a given year if they belonged to the same

high-level category. Originally, all the foods were categorized into 28 and 21 high-level categories for year 2000 and other years, respectively. However, to unify the categories for all the years and optimize the categories for WF accounting, the foods were further grouped into 13 new categories as shown in Table S1 of Supporting Information S1. As a result of this refinement, the original retrieved data were compressed to 37,376 instances.

To obtain the dietary WF values for each instance, a process of WF accounting was required. In this study, the dietary WF accounting primarily involved the multiplication of the food intake amount of each instance by the WF coefficient specific to that particular type of food, which belongs to one of the 13 refined food categories. The foods' WF coefficients represent the water consumption in the whole supply chain of the production of one unit of food. In this study, the foods' WF coefficients were obtained based on a thorough literature search, as presented in Table S1 of Supporting Information S1.

2.2. Further Preprocessing of Data

After the data retrieval and WF accounting, the resultant data set contained 37,376 instances, eight input features ("Income," "Residence place," "Highest degree," "Gender," "Age," "Weight," "Height," "BMI"), and one output label ("Dietary WF"). The outliers, identified using three-sigma rule (Pukelsheim, 1994), consist of 395 instances (1.1% of the total data). These outliers are heavily skewed toward high WF values which contribute 3% of the total water footprint. While tree-based models are less impacted by outliers, the regression model used in this study is sensitive to extreme values. Removing outliers can also be beneficial for model explanation, as it makes SHAP value distributions less volatile. After removing outliers, thresholds were established to define low and high WF groups while retaining sufficient instances to ensure reliable information extraction. Balancing this tradeoff was critical for this study. Focusing on the tails of the data distribution, which represent high and low dietary WFs, allowed us to better uncover the driving forces and their influences on significant WF. To achieve this, instances near the average (within the range of mean $\pm$ 0.7 $\times$ std) were excluded for both training and testing data set, as they may obscure the determinants of WF changes in the analysis. This exclusion also helped to generate a balanced data set for model training, improving the model's performance. After careful evaluation, the results demonstrated that this approach improved the area under the precision-recall curve (AUC-PR) and provided more consistent interpretations. As a result, the remaining instances with WF values smaller than 670 m^3 were assigned to the low WF group, whereas those with WF values greater than 1,200 m^3 were assigned to the high WF group. We labeled the low WF group as the negative class (0) and the high WF group as the positive class (1).

This categorization simplified the analysis, increases policy relevance and improved model stability. By categorizing dietary WFs into "high" and "low" groups, high WF populations can be more clearly identified, providing a clear direction for the development of targeted interventions. This approach reduces the complexity of the model and improves the interpretability and usefulness of the analysis and ensues that we have unique and meaningful groups to analyze the drives behind the differences between low and high WFs in individual's diets. We were left with 18,067 data instances after this operation, which provided sufficient data for reliable modeling and drawing meaningful conclusions.

Subsequently, all the continuous features were converted to categorical features, and the outcome of the conversion is displayed in Table 1. This conversion was done to avoid cardinality bias toward the continuous features when evaluating the importance of the features (Deng et al., 2011). Moreover, the study's objectives did not require the exact values of the features to be retained. The preprocessing of the data was concluded after this step, and the data set was deemed suitable for modeling. It is worth noting that all the categorical features were given new names, as shown in Table 1 caption.

3. Methodology

In this section, a thorough depiction of the ML-based framework is presented, followed by concise descriptions of each ML-related method incorporated in the framework. For the comprehensive details of the methods, please refer to Section S2 in Supporting Information S1.

3.1. Study Framework

The flowchart of the ML-based framework is depicted in Figure 1, followed by a detailed walkthrough.

Table 1
The Levels/Categories of all the Demographic and Anthropometric Characteristics and Their Corresponding Values

Attribute level/ category	Income (annual; CNY)	Residence place	Highest degree	Gender	Age (year)	Weight (kg)	Height (m)	BMI
0	None	Rural village	None	Male	[min, 18)	[min, 35)	[min, 1.4)	[min, 18.5)
1	[0, 6,000)	Rural town	Grad from primary school	Female	[18, 30)	[35, 50)	[1.4, 1.5)	[18.5, 25)
2	[6,000, 12,000)	Urban suburb	Lower middle school degree		[30, 40)	[50, 60)	[1.5, 1.6)	[25, 30)
3	[12,000, 20,000)	Urban city	Upper middle school degree		[40, 50)	[60, 70)	[1.6, 1.7)	[30, max]
4	[20,000, 30,000)		Technical or vocational degree		[50, 60)	[70, 80)	[1.7, 1.8)	
5	[30,000, 50,000)		University or college degree or higher		[60, 70)	[80, 90)	[1.8, max]	
6	[50,000, max]				[70, max]	[90, max]		

Note. All the features in the data set were given new names after being converted to categorical features, namely "income_level," "urban_level," "edu_level," "gender," "age_level," "weight_level," "height_level," and "BMI_level," replacing the original "Income," "Residence place," "Highest degree," "Gender," "Age," "Weight," "Height," and "BMI," respectively.

As it can be seen from Figure 1, the proposed methodological framework starts with data preprocessing procedures resulting in "Data ready for ML" (details in Section 2.2). Once the data had been preprocessed, the data were randomly shuffled and split into the training data set (90%) and test data set (10%). The training data set was used to train the models and build the relationship between the features and the label. The test data set was used to test the generalization performance of the models, that is, test whether the trained models still performed well on unknown data and if the built relationships were also valid for such data.

After the data split, four models, Logistics Regression (LR), HGB, ET, and XGB, were trained using the training data set. In this study, the LR model was chosen as a representative of conventional statistical models. On the other hand, the tree-based models - HGB, ET, and XGB—were chosen to represent advanced ML methods for two reasons: (a) their great ability to capture complex information in data, comparable to neural networks; and (b) their inherent structure makes them highly interpretable (Lundberg et al., 2020; Wang et al., 2021, 2022). The model training process was basically about tuning hyperparameters for the models. In this study, we adopted *random search* (Bergstra & Bengio, 2012) accompanied by *cross-validation* (Browne, 2000) to find the optimal hyperparameters for all four models. After completing the training, the performances of the four models were compared with the baseline model, Random Guess Classifier, using the AUC_PR, complemented by accuracy, precision, recall, and F1-score on both the training and test data sets. Valid models are those that perform better than the baseline model. We then evaluated and compared all the valid models by analyzing the performances on both the training and test data sets. The criteria used to determine whether a model was good were twofold: (a) its ability to capture enough variability, indicated by high performance metrics values especially on test data set; (b) small gaps in performance between the training and test data sets (good generalization capability). Ultimately, based on these criteria, the optimal model was selected for the subsequent model explanation.

After obtaining the optimal model, SHAP was employed to explain the optimal model. Through SHAP every feature's SHAP values were estimated for all the instances in the training data set. For a certain feature, the FI can be obtained by calculating the average over all the instances' absolute SHAP values. This will result in a FI ranking by SHAP. To validate this ranking from SHAP, PI measure was used to obtain an independent FI ranking. As mentioned in Section 3.6, PI has a totally different mechanism for measuring FI compared to SHAP. Thus, by checking the alignment between these two rankings, we would know how reliable and robust the FI ranking from SHAP was. If the FI ranking from PI diverged very much from the ranking from SHAP, a further thorough investigation would be carried out to find out the culprits. Otherwise, the top four influential features would be identified according to the SHAP FI ranking.

After identifying the top four features, the SHAP dependence analysis was carried out to examine the specific influence patterns of the top four features on the label, dietary WF. Through the SHAP dependence plots, we revealed how the instances' dietary WF levels tend to change as the features change from one level/category to another. To further validate the results, we conducted a complementary analysis using both logistic regression and the XGB model, applying SHAP analysis the entire data set, including those within mean ± 0.7 SD of WF.

10.1029/2024EF005061

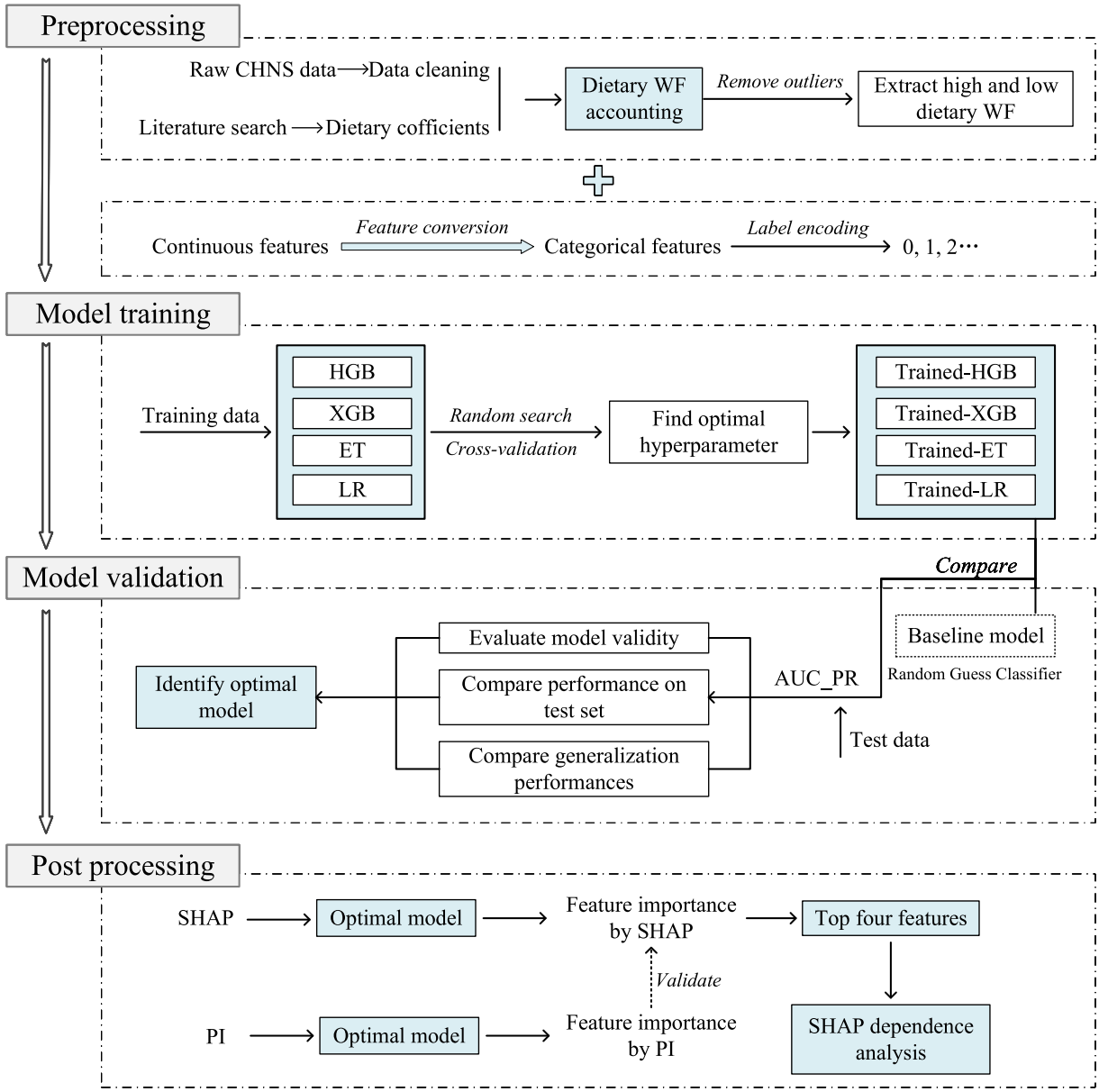

Figure 1. Flowchart of the ML-based framework.

3.2. Histogram-Based Gradient Boosting and eXtreme Gradient Boosting

Gradient boosting decision tree (GBDT) (Friedman, 2001) is an ensemble ML method that combines multiple weak learners (decision trees) to form a single strong learner. Unlike ET in which trees are constructed independently and in parallel, GBDT follows a sequential and iterative learning process, where each tree is trained based on the residuals from the previous tree.

Histogram-based Gradient Boosting (HGB) introduced by Scikit-learn (Pedregosa et al., 2011) and eXtreme Gradient Boosting (XGBoost; a shorter acronym "XGB" is used in this study) (Chen & Guestrin, 2016) are two efficient engineering implementations of GBDT. The main idea behind HGB is to represent the feature space using histograms, which are used as the input for decision trees during the boosting process. Instead of using traditional numerical feature values, the feature values are discretized into a finite number of bins to form a histogram representation. The histograms are constructed in a way that allows for efficient computation of gradient-based statistics, which are used for constructing decision trees in the gradient boosting process. It is

especially useful when dealing with data sets with a large number of features or when computational efficiency is a priority, such as in big data or real-time applications.

In comparison to conventional GBDT, XGB incorporates several notable improvements. For example, XGB utilizes Newton's method instead of gradient descent. Newton's method leverages Taylor's expansion of the second order of the loss function, leading to faster and more accurate convergence of the loss function. Also, Newton's method allows for customization of the loss function, providing flexibility to users. Furthermore, XGB controls the complexity of the trees by adding a regularization term to the objective function. This regularization term enables users to set hyperparameters that govern the regularization strength, mitigating the risk of overfitting in the learned trees and achieving more robust and generalized models. Besides the improvements mentioned above, XGB provides built-in mechanisms to promote parallel computing and automatically handle missing values for training.

The Scikit-learn API *sklearn.ensemble.HistGradientBoostingClassifier* was used for HGB modeling, and the Python package *xgboost* was used for XGB modeling.

3.3. Extremely Randomized Trees

Extremely Randomized Trees, also known as Extra-Trees (ET), is an ensemble ML method that combines the predictions of multiple decision trees to produce a more accurate and robust result than individual trees (Geurts et al., 2006). Unlike traditional decision trees, where the best split is chosen based on a specific criterion (such as Gini impurity or entropy), ET selects the splits randomly from a set of candidate splits. The randomization in ET introduces additional diversity in the ensemble, which can help reduce the model's variance (i.e., reduce the potential of overfitting) and improve its ability to handle noisy or high-dimensional data sets. The use of random feature and threshold selection also allows for faster tree construction, as the need for an exhaustive search for the best split at each node is eliminated.

The Scikit-learn API *sklearn.tree.ExtraTreeClassifier* was used for ET modeling.

3.4. Logistic Regression and Random Guess Classifier

Logistic Regression (LR) is a statistical method for binary classification, where the goal is to predict the probability of an instance belonging to a certain class based on its features (Hosmer Jr et al., 2013). It is a type of generalized linear model that models the relationship between the features and the binary outcome with the involvement of the logistic function. In order to compare the performances of ML methods and conventional statistical methods, LR was chosen as the representative of conventional statistical methods, which have been used prevalently in similar WF driving force studies as described in *Introduction*. It is worth noting that even though most of the previous studies adopted Multiple Regression (MR) models, MR is not applicable to this study because it is not designed for classification problems (Allison, 1999). LR is the closest alternative for a classification problem. The Scikit-learn API *sklearn.linear_model.LogisticRegression* was used for LR modeling.

A random guess classifier is a type of model that makes random predictions for classification tasks. It serves as a baseline or reference model to evaluate the performance of other more sophisticated models (Kourou et al., 2015). The random guess classifier predicts class labels based on a random selection from the available classes, without considering any features or patterns in the data. In this study, the performances of the trained LR, HGB, ET, and XGB classifiers were compared with the performance of the random guess classifier to examine whether those models were valid, that is, whether they captured a meaningful relationship between the features and the class label.

3.5. Model Performance Evaluation Metric

The performances of the above models were assessed using multiple metrics, including the primary metrics AUC_PR, as well as accuracy, precision, recall, and F1-score. The AUC_PR is a commonly used metric to assess the prediction performance of a classification model. AUC_PR represents the overall performance of the model across all possible classification thresholds. It can be interpreted as the probability that the model will rank a randomly selected positive instance higher than a randomly selected negative instance. AUC_PR values range from 0 to 1, where a higher value indicates better performance. In this study, AUC_PR was utilized for evaluating and comparing model performances due to its emphasis on positive class detection. This focus aligns with the

study's objective of identifying the driving forces behind WF, especially high WF represented by the positive class. AUC_PR is particularly valuable because it directly measures the trade-off between Precision and Recall, providing a comprehensive reflection of the model's effectiveness in detecting positive samples. Additionally, accuracy reflects the proportion of correctly classified instances across the data set. Precision quantifies the proportion of true positive predictions among all predicted positives, while Recall measures the proportion of actual positives correctly identified by the model. The F1-score, a harmonic mean of Precision and Recall, provides a balanced assessment of performance, particularly in scenarios with class imbalance. To ensure the reliability of these metrics, 95% confidence intervals (CIs) were estimated using the bootstrap method, offering robust measures of variability across resampled data sets.

3.6. SHapley Additive exPlanations and Permutation Importance

SHAP is an explanatory method based on cooperative game theory that quantifies the contribution of each feature to model prediction results by calculating Shapley values (Lundberg & Lee, 2017a, 2017b). The core idea of the SHAP method is to decompose the prediction results of each sample and calculate the marginal contribution of each feature in a particular prediction. Specifically, the Shapley value takes into account all possible subsets of features and measures the influence of a particular feature by calculating its average contribution across different combinations of features. SHAP is able to overcome potential bias problems and provide more consistent and explanatory results than other methods of assessing FI.

In this study, we aggregated the Shapley values of all input features, which in turn generated a global FI ranking. This ranking reflects the relative influence of each feature on the model predictions and helps to identify the key features that contribute most to the predictions. This approach allows us to better understand the decision-making process of the model and extract meaningful features for subsequent model optimization or decision support.

As of writing, SHAP is the only method known to possess three important properties: *local accuracy*, *missingness*, and *consistency* (Lundberg et al., 2018). Additionally, SHAP stands out as a versatile explanation method. It is compatible with all types of ML models, providing both global and local explanations, and encompassing both FI measure and dependence analysis. The aforementioned benefits led to the selection of SHAP as the primary explanation method in this study. In this study, the Python package *shap* was utilized for conducting the SHAP analysis. The PI (Altmann et al., 2010) was employed as another FI measure to validate the FI results of SHAP from a different perspective. The basic idea behind PI is to measure the change in model performance when the values of a particular feature are randomly permuted while keeping the label unchanged. The intuition is that if a feature is important, then shuffling its values should significantly degrade the model's performance, as the model relies on that feature for making accurate predictions. To obtain a more robust estimate of the PI score, 50 shuffles were performed and the average was used as the final PI score for each feature. The Scikit-learn API *sklearn.inspection.permutation_importance* was used for PI analysis. The 95% CIs were estimated using the bootstrap method for both SHAP and PI to ensure the reliability of the results.

4. Results and Discussion

4.1. Model Performances

The performances of the models developed in this work are shown in Table 2.

As shown in Table 2, on both the training and test data sets, all four models (LR, HGB, ET, and XGB) demonstrate considerably better performance compared to the baseline model, indicating their validity for the given task. Across all models, overall performance was moderate. XGB achieved the highest mean AUC-PR of 0.74 (95% CI: 0.703–0.757), recall (0.645) and F1-score (0.662). While XGBoost achieved marginally higher scores in some metrics, logistic regression performed comparably with AUC-PR of 0.71, recall (0.61) and F1-score (0.64). Although XGB does not dramatically outperform other models, it provides a stable balance of predictive accuracy and flexibility in capturing nonlinearities and interactions. Therefore, we selected XGB for SHAP interpretation in the later section to offer an accurate and comprehensive understanding of high WF (positive instance), providing valuable insights for water resources management in the future.

Table 2
Performances of Models on Training and Test Data Sets

	Random guess classifier (baseline)	Logistic regression	HGB	ET	XGB
Training data set					
AUC-PR (Mean, 95% CI)	0.459 (0.445, 0.469)	0.706 (0.6953, 0.7150)	0.7412 (0.732, 0.750)	0.739 (0.730, 0.748)	0.732 (0.723, 0.741)
Accuracy (Mean, 95% CI)	0.498 (0.490, 0.506)	0.692 (0.685, 0.699)	0.711 (0.704, 0.718)	0.707 (0.700, 0.714)	0.707 (0.700, 0.714)
Precision (Mean, 95% CI)	0.498 (0.490, 0.506)	0.687 (0.678, 0.696)	0.703 (0.695, 0.712)	0.700 (0.692, 0.710)	0.682 (0.674, 0.690)
Recall (Mean, 95% CI)	0.498 (0.490, 0.506)	0.605 (0.594, 0.616)	0.640 (0.630, 0.652)	0.632 (0.621, 0.643)	0.705 (0.698, 0.712)
F1 score (Mean, 95% CI)	0.497 (0.489, 0.505)	0.643 (0.635, 0.651)	0.670 (0.662, 0.678)	0.664 (0.656, 0.673)	0.680 (0.672, 0.687)
Testing data set					
AUC-PR (Mean, 95% CI)	0.466 (0.436, 0.499)	0.717 (0.689, 0.744)	0.722 (0.695, 0.753)	0.722 (0.694, 0.753)	0.730 (0.703, 0.757)
Accuracy (Mean, 95% CI)	0.516 (0.493, 0.539)	0.687 (0.666, 0.708)	0.688 (0.666, 0.709)	0.693 (0.672, 0.714)	0.693 (0.672, 0.714)
Precision (Mean, 95% CI)	0.515 (0.490, 0.538)	0.697 (0.668, 0.727)	0.684 (0.658, 0.709)	0.697 (0.670, 0.725)	0.681 (0.655, 0.706)
Recall (Mean, 95% CI)	0.515 (0.490, 0.538)	0.584 (0.550, 0.615)	0.615 (0.583, 0.647)	0.600 (0.567, 0.632)	0.645 (0.615, 0.678)
F1-score (Mean, 95% CI)	0.515 (0.490, 0.537)	0.636 (0.609, 0.662)	0.647 (0.621, 0.673)	0.645 (0.618, 0.669)	0.662 (0.638, 0.688)

4.2. Feature Importance

Following the selection of the optimal model (XGB), we performed FI measures to determine the relative influence of each feature on the output, as compared to each other. The outcomes of the SHAP FI measure and PI measure for influence ranking are presented in Figure 2. Specifically, subplot (a) shows the results obtained from the SHAP FI measure, while subplot (b) displays the results obtained from the PI measure.

As shown in Figure 2, the SHAP FI result indicates that "income_level" has the most substantial influence on the output, followed by "urban_level," "edu_level," "gender," "height_level," "age_level," "BMI_level," and "weight_level." The PI measure yields a ranking broadly consistent with that of SHAP FI, except that "age_level" is ranked higher than "height_level." It is noteworthy that SHAP FI calculates the FI by averaging the change in the model's prediction when the feature is included versus excluded over all possible feature subsets. On the other hand, PI measures the FI by permuting the feature values and observing the resulting decrease in the model's performance. Therefore, the consistency between the rankings from the two measures indicates the reliability and robustness of the SHAP FI result.

In terms of the four most significant features of interest, both FI measures indicate the same features and their respective rankings, which are "income_level," "urban_level," "edu_level," and "gender," arranged in descending order of importance. Consequently, we selected these four features for SHAP dependency analyses with the aim of investigating the specific patterns of influence of these characteristics on output, and thus exploring possible social, economic and individual influences on dietary WF.

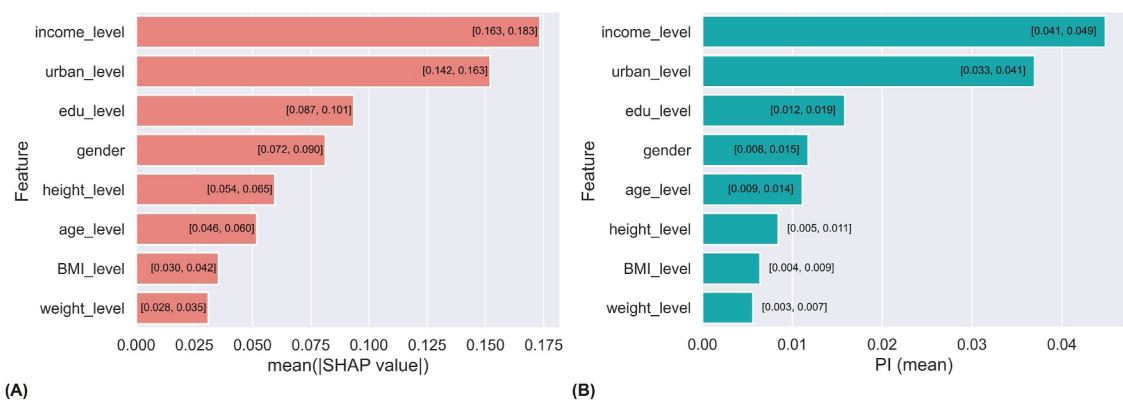

Figure 2. FI plots for XGB from both SHAP (a) and PI measure (b).

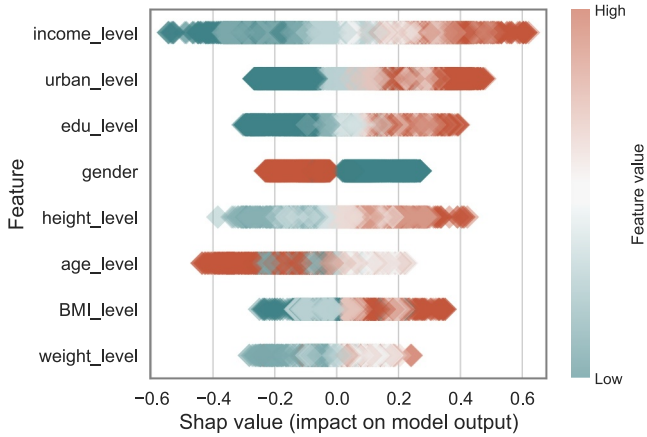

Figure 3. Overall SHAP dependence patterns for all features.

4.3. Results of SHAP Dependence Analysis

4.3.1. Overall Dependence Patterns

Figure 3 below illustrates the overall distribution of instances' SHAP values with respect to the feature values. It is a brief overview of the dependence patterns for all features. The details of each top feature's dependence pattern and the corresponding underlying causes are presented in Section 4.3.2. The *x*-axis represents the instances' SHAP values, where positive values denote contribution to the model prediction of high dietary WF (class 1) and negative values denote contribution to the model prediction of low dietary WF (class 0). Each instance is represented by a diamond shape with the color of reflecting the value of the corresponding feature.

As it can be seen from Figure 3, the "income_level" feature has the widest SHAP value distribution, with numerous instances positioned at the positive and negative extremities of the distribution bar, indicating substantial contributions to model outputs regarding both levels of dietary WF. Low "income_level" values tend to associate with negative SHAP values while high "income_level" values tend to associate with positive SHAP values. The other two top features ("urban_level," "edu_level") share these characteristics as well, except for the smaller expansions of the SHAP value distribution. The SHAP value distribution for "gender" is notably distinct, with the majority of males (value 0, represented by blue diamonds) contributing to the positive SHAP values, whereas a majority of females (value 1, represented by red diamonds) contributing to the negative SHAP values. Although binary features are more likely to present contrasting SHAP contributions due to their two states reflecting opposing model weights, the sharp distinction in genders is unlikely to be solely attributable to the binary nature of this category within the model. To validate this, we further introduced a random binary variable into the model (Figure S1 in Supporting Information S1). The SHAP distribution for gender remained unchanged, indicating that the model's reliance on gender is based on its predictive contribution to the target variable. The features with lower rankings exhibit two main characteristics. First, there are no consistent dependence patterns (e.g., "age_level" and "BMI_level"). Second, the instances with high and low feature values are not intensively located at the extremities of the bars (e.g., "height_level" and "weight_level"). In addition to exploring the likelihood of each characteristic contributing to high dietary WF through the mean absolute SHAP values (as shown in Figure 2), the observations in Figure 3 provide an additional perspective on why these characteristics are less influential than the first four.

4.3.2. Thorough Dependence Patterns for the Top Four Features

To further investigate the dependence patterns for the top four features, strip plots were used, as shown in Figure 4. This figure consists of four subplots, with Subplot (A) for "income_level," (B) for "urban_level," (C) for "edu_level," and (D) for "gender." The *x*-axis in each subplot represents the distinct values of the corresponding feature. Additionally, each subplot has 2 *y* axes, with the left one representing the SHAP values of the instances and the right one representing the count of instances for each distinct value of that feature. For every subplot, each scatter represents an instance with a specific SHAP value. These scatter points form strips in different colors, with each strip corresponding to one distinct value of the feature. The grey bars in the background indicate the number of instances falling in each strip.

For each subplot, the dependence pattern was described in detail, followed by the deduction of the most probable underlying reasons for that pattern. Ultimately, the obtained reasons were used to formulate the policy implications for dietary WF reduction. To ensure the reliability and trustworthiness of the derived policy implications, the deduction of reasons was grounded in solid and well-established findings from pertinent literature of sociology, political science, and economics.

As depicted in Figure 4a, there is a general increase in the SHAP values of instances as income level increases between 0 and 6, indicating that a higher income level contributes positively to the model predicted high dietary WF. To further elaborate, the overall SHAP values remain relatively stable when the income level changes from 0 to 1, while a considerable jump in the strips is observed when the income level changes from 1 to 2. However, from level 2 onwards, the increases in the strips are much less drastic, and the ascending gradient remains almost

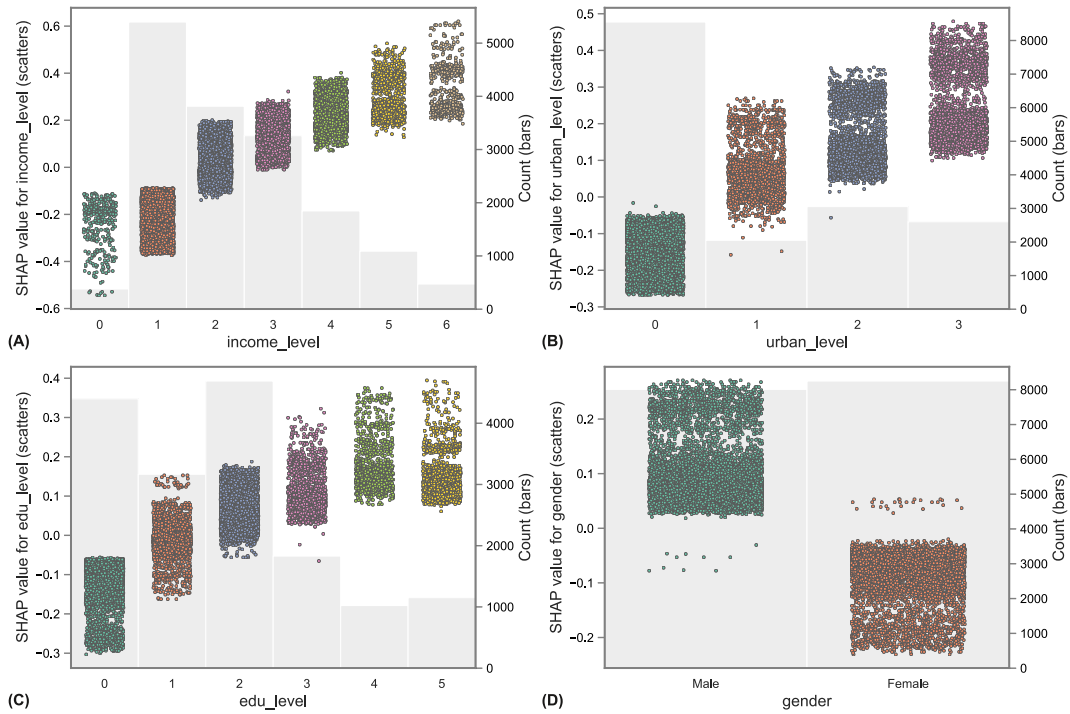

Figure 4. Intricate dependence patterns for all top four features: (a) income_level (b) urban_level (c) edu_level (d) gender.

constant. Along with this dependence trend, it can be observed that all instances at level 0 and 1 have negative SHAP values, implying that typically these levels contribute to low dietary WF predictions. While some instances at level 2 income still show negative SHAP values, the majority begin to have positive contributions to predicted high dietary WF. From level 3 onwards, all income levels have positive SHAP values for all associated instances, highlighting the significant positive influence on model predicted high dietary WF.

Income level emerges as the most influential feature affecting the model predicted dietary WF due to its strong association with food purchasing activities. Dietary WF reflects the water consumed not only through the amount of food intake but also through the water intensity of its production. Disposable income influences both the quantity and quality of the food that individuals can afford, which directly impact the dietary WF. Residents with lower incomes exhibit higher own-price and income elasticities of demand, making them more sensitive to changes in food prices and limiting food consumption. They often prioritize essential foods and rely on staples and low-cost options, such as cereals and vegetables that contains low dietary WF (Gronau & Hamermesh, 2008; Thiele & Weiss, 2003). In contrast, high-income residents tend to purchase more food than necessary and favor higher-priced items, such as meat, aquatic foods, and dairy products (Gale & Huang, 2007; Huang & Gale, 2009; Liu, 2014), which are associated with higher WF coefficients. This disparity contributes to an overall WF differential observed between low- and high-income groups. Moreover, high-income people in China usually reside in developed areas where large supermarkets are prevalent. Because of bulk purchasing, direct sourcing, and high competition, large supermarkets often offer lower prices for foods compared to small food stores. On the contrary, low-income individuals tend to reside in less developed areas with limited food vendors and the absence of large supermarkets, resulting in higher prices paid per unit of food (Kaufman et al., 1997; Liao et al., 2016; Regmi, 2001). This differential in food prices further exacerbates the gap in overall food consumption between low- and high-income residents, leading to an additional disparity in dietary WF.

Figure 4b highlights the influence of urban level on the model's predictions. SHAP values shifts from negative to positive as the urbanization level increases from 0 to 3, indicating a higher urbanization level corresponds to a higher likelihood of having prediction of high dietary WF. For urban level 0, SHAP values are negative for all instances, indicating a contribution to the prediction of low dietary WF. At levels 1 and 2, while some instances still contribute to predicted low dietary WF, the majority of them contributes to the prediction of high dietary WF.

By level 3, all SHAP values are positive, suggesting that the highest urbanization levels positively influence the model's prediction of high dietary WF.

China has significant differences in the pace of life and daily habits between urban and rural residents, which can play a key role in disparities in dietary WF (Wang et al., 2019). Firstly, urban lifestyles are often constrained by tight time schedules and long commutes, making it challenging for residents to prepare meals at home. As a result, convenience foods, which are readily accessible, affordable, and available, make up a substantial portion of urban residents' diets (Bai et al., 2010; Zhou et al., 2015). Secondly, urban areas offer a wide range of food options, including whole and processed foods, as well as domestically produced and imported products. Urban residents tend to show greater acceptance of Western diets, characterized by high intakes of red meat, dairy products, processed foods (Pingali, 2007; Popkin, 1999; Regmi & Dyck, 2001; Yuan et al., 2019). In summary, urban residents' preference for convenience foods and their access to a diverse range of food options are more likely to result in high dietary WF, as these food types are often associated with higher WF coefficients.

The results displayed in Figure 4c demonstrates the impact of education level on SHAP values. The SHAP values of instances tend to increase as the education level rises from 0 to 5, but this trend levels off when the level changes from 4 to 5. This suggests that higher education levels contribute positively to the model prediction of high dietary WF but the strength of this effect diminishes at the highest education level. It is worth noting that all instances at level 0 show negative SHAP values, which implies that this level of education is more associated with the prediction of low dietary WF. At education levels 1 and 2, there is a mix of negative and positive SHAP values, reflecting a transition in how these education levels influence the model's outputs. From level 3 onward, nearly all instances exhibit positive SHAP values. This indicates that the residents on education level 3 and above (higher than upper middle school degree) are more likely to be associated with model prediction of high dietary WF.

Education level ranks third in terms of its influence on the prediction of dietary WF, and its effect on the dietary WF is a complex process involving several factors. First, higher levels of education are usually associated with an accumulation of knowledge that makes individuals more aware of the environmental impacts of food, which in turn motivates them to make more water-efficient, environmentally friendly and sustainable dietary choices (Choy & Li, 2017; Wang et al., 2019). Second, the propensity to adopt healthy and sustainable diets tends to be stronger in groups with higher levels of education. Many studies have shown that highly educated populations tend to choose more plant-based foods and reduce their intake of processed and animal foods. This not only contributes to good health, but is also effective in reducing the WF (Lin & Yen, 2008; Meng et al., 2009). In addition, people with higher levels of education tend to have more social resources and support, enabling them to make consumption choices that are more in line with environmental and health goals (Bhandari & Smith, 2000). Although these foods may be more expensive, individuals with higher levels of education are more likely to pay a premium for sustainable foods due to their greater financial resources (Bhandari & Smith, 2000; Chen et al., 2016; Liao & Chern, 2007; Wang et al., 2012). Together, these factors contribute to the fact that populations with higher levels of education tend to choose water-friendly diets, thereby reducing the dietary WF.

Two strips are shown in Figure 4d, one for “Male” and the other for “Female.” The vast majority of instances in the “Male” strip display positive SHAP values, with only 10 instances having negative SHAP values. Conversely, an overwhelming majority of instances in the “Female” strip show negative SHAP values, with only 27 instances displaying positive SHAP values. This observation suggests that being male is a consistent factor contributing to the model prediction of high dietary WF while being female is a consistent factor contributing to the model prediction of low dietary WF. Even though “Male” and “Female” are stable contributors, their contributions are not significant compared to the three features mentioned above. The maximum SHAP value in the “Male” strip is 0.23 and the maximum magnitude of SHAP values in the “Female” strip is 0.28. And the absolute value of SHAP for females is greater than for males, implying that females have a greater influence on model prediction of dietary WF.

The possible reasons behind difference on gender's influence on the model prediction are as follows. First, men tend to have a higher proportion of muscle mass than women, which requires more calories to maintain the weight and support the associated physical activities (Miller et al., 1993). This can lead to a higher appetite and a greater food intake in men. Second, men tend to have higher levels of testosterone, which can increase appetite and promote metabolism, leading to a higher food intake (Clark et al., 2019; Karl et al., 2020). Third, Cultural norms around gender roles and food consumption may also play a role in the observed difference. For example, in China,

men are usually expected to consume larger portions or eat more meat as a symbol of strength and masculinity, while women are typically expected to consume smaller portions or focus on lighter and healthier foods for a slim body shape (Dong, 2013; Gough, 2007; Oncini & Guetto, 2018; Rothgerber, 2013).

In the complementary analysis including the entire data set (i.e., without excluding cases within mean ± 0.7 SD of WF), overall model performance decreased (XGB: AUC-PR 0.686, F1-score 0.662; Logistic regression: AUC-PR 0.), reflecting the greater difficulty of predicting outcomes near the mean. The SHAP-based feature importance rankings remained largely consistent with those from the primary analysis. The top four features were unchanged, although minor reordering occurred among features with lower importance (e.g., height_level and age_level). Importantly, the relative differences among these lower-ranked features were small, indicating that the overall explanatory conclusions are relative effective.

4.4. Significance of Study

4.4.1. Advantages of the ML-Based Framework

A well-structured ML framework for WF driving force analysis was proposed following rigorous data science principles. This framework covered the whole pipeline from converting WF domain problems into ML problems, to preparing the data, selecting and constructing the model, explaining the model, and applying the explanation results to the WF domain context. Through the case study, we have demonstrated this ML-based framework possesses the following major advantages:

- Methodological triangulation enhances the credibility and reliability of findings mentioned in the previous three sub-sections. At the key procedures, multiple approaches are involved to select the best fit model for explanation, facilitate model optimization and improve model interpretation (e.g., use PI measure to validate the FI result from SHAP).
- Across all models, overall performance was moderate. While several ML models achieved slightly higher scores than logistic regression on some metrics, confidence intervals often overlapped, indicating that predictive gains were limited. However, ML methods provide added value by flexibly capturing nonlinearities and interactions, and by offering complementary insights through explanation analysis.
- ML explanation can yield in-depth findings that are of great value for dietary WF reduction policies. The FI analysis and SHAP dependence analysis not only helped in identifying the priority order of demographic features but also highlighted the specific groups within each feature that require the most attention. Section 4.4.2 explains the significance of this information.

It is worth noting that this framework is highly transferable, meaning it could in principle be applied to any form of driving force analysis in any domain (not limited to the association between dietary WF and demographic and anthropometric characteristics). The prerequisite for this is the availability of sufficient high-quality data, which is essential for robust and reliable findings.

4.4.2. Benefits of Ranking Features and Dependence Analysis

One of the major novelties of this study is that it has systematically and rigorously conducted the FI analysis and dependence analysis that are absent in the previous dietary WF driving force studies. The derived explicit ranking of all features and the top four features' detailed influence patterns could provide substantial benefits in informing policies on reducing dietary WF in China. For example, it would enable efficient resource allocation and targeted intervention.

4.4.2.1. Efficient Resource Allocation

The resources for achieving the policy goal, such as funding, personnel, and time, are often limited, especially given the fact that China is a country with a vast land area and a huge population. Hence, policymakers need to make strategic decisions about where to allocate them. With the results obtained in this study, policymakers can ensure that resources are directed to the areas where they would have the greatest impact. According to the results, the priority order should be high-income residents > urban residents > highly educated residents > male residents. Thus, instead of investing resources in promoting low-WF diets across the entire population, policymakers could focus their efforts on high-income groups or urban areas, where the potential impact on reducing dietary WF is greater.

4.4.2.2. *Targeted Intervention*

By understanding which groups have the higher WF, policymakers can design policies that are tailored to the specific characteristics of those groups. For example, high-income residents have the highest WF and it is largely attributed to the behaviors of consuming high-WF (also usually expensive) foods and purchasing more foods than they actually need. Thus, policymakers may focus on implementing taxes on high-WF foods, carrying out educational campaigns to improve high-income residents' consumption behavior, or incentivizing high-income residents to denote excess foods to food banks.

4.4.3. Implications for Policy on Dietary WF Reduction

By investigating the influence rankings of all the features and analyzing the dependency patterns of the top four characteristics, the possible influence of each feature in causing predication of high dietary WF is explored and the possible reasons for the influence are discussed. Based on the findings, this study provides several insights into policies aimed at reducing dietary WF in the population, which are summarized below:

- *Promote water-efficient dietary patterns.* Governments should use public policies to promote plant-based diets such as legumes, cereals and vegetables, increase subsidies to reduce price differentials for water-intensive livestock foods, and adjust prices and excise taxes to provide incentives to reduce consumption of meat and protein-rich foods (Springmann et al., 2016).
- *Increase public education and awareness.* Raising national awareness of the WF of diets is key. Educational campaigns in schools, communities and the media, especially in less educated areas, encourage more sustainable diets (Tukker & Jansen, 2006; Vermeir & Verbeke, 2006).
- *Improve the rural food supply chain.* Governments should promote localized production and consumption in areas with large urban-rural disparities, support small-scale organic farming and farmer'' cooperatives, increase self-sufficiency and reduce water consumption (Bélanger & Pilling, 2019; Garnett, 2014).
- *Promote gender equality and women's education.* Women play an important role in household food decision-making and increasing women's education and environmental awareness can help reduce household food consumption. Policies should provide opportunities for education and training, especially in rural and low-income areas (Liu et al., 2017).
- *Income differences and subsidy policies.* Governments can support low-income groups to choose low-WF foods through income subsidies or tax rebates, such as subsidized fruits and vegetables, to promote sustainable and environmentally friendly diets (Wiedmann & Minx, 2008; Wikström et al., 2014).

4.5. Limitation

Several limitations of the study warrant further exploration. First, while the XGB model achieved a moderate performance level with an AUC_PR of 0.73 on testing data, its overall predictive power was moderate and confidence intervals overlapped across models, which constrains the extent to which these findings can be generalized to broader populations beyond the surveyed sample. Future studies should continue to optimize model performance or explore alternative modeling approaches such as explainable deep learning or interpretable models like Generalized Additive Models. Additionally, the binary classification of dietary WF into low and high levels may have masked nuanced patterns. Future research could consider multi-class classification or regression-based approaches for achieve more detailed insights. During data preprocessing, a small number of outliers were removed to enhance model performance and reduce noise. However, this may potentially introduce bias and limiting the data set's representativeness. Future work could further explore the outliers, especially instances with extremely high dietary WF, to understand their characteristics, patterns, and potential impact. In addition, the feature importance rankings did not account for correlations among predictors, which may influence their relative importance. Therefore, these results should be interpreted as indicative of influential factors rather than as definitive causal attributions. These limitations highlight the need for further data analysis, advanced feature engineering, and exploration of alternative modeling approaches to better reveal the driving forces behind dietary WF in greater details.

5. Conclusions

A novel framework incorporating three ML models, HGB, XGB, and ET, along with an ML explanation system SHAP, was proposed for analyzing the driving forces behind the dietary WF of individuals. According to our best

knowledge, this is the first application of all the ML-related techniques (including PI measure and AUC_PR metric) in WF-related research. Besides, LR was included in the framework serving as the reference model and representing conventional statistical methods prevalently used in the relevant literature. The framework was designed based on rigorous data science principles with methodological triangulation implemented at critical stages to improve its credibility and reliability. The proposed framework possesses a high degree of transferability, as it can potentially be employed in driving force analyses across diverse domains beyond the specific context of this study.

A case study was conducted on the CHNS data collected from Chinese inhabitants to illustrate the framework's specifics. The data used for modeling were the most extensive to date (18,067 instances) for WF studies concerning China, comprising individuals' demographic and anthropometric characteristics as input and dietary WF as output. The results of the case study validated the feasibility of applying this framework to real-world data and demonstrated its superiority over conventional statistical methods. The key findings of the case study are as follows:

1. All ML models outperformed the LR model in terms of predicting the dietary WF level from individuals' demographic and anthropometric characteristics. This has justified the adoption of ML methods for WF driving force analysis.
2. XGB was the optimal model as it effectively captured the variability in the data and showed good generalization performance. The model interpretation was therefore carried out based on the XGB model.
3. According to the converging evidence from two different feature importance measures, SHAP and PI, the top four influential features on dietary WF are "income_level," followed by "urban_level," "edu_level," and "gender."
4. Based on SHAP dependence analysis, the following groups should be prioritized for implementation of measures to reduce dietary WF(in descending order of priority): high-income residents, urban residents, highly educated residents, and male residents. To be more specific, the high-income residents refer to the ones with an annual income greater than CNY 12,000 (approximately USD 1,700 at the time of writing this paper), and the highly educated residents refer to the ones who have completed an upper middle school degree or higher.

The last two findings have implications for effective resource allocation and targeted interventions to promote dietary WF reduction in China. Additionally, the discussion of the underlying causes of these findings have led to the following policy suggestions: (a) Governments should encourage plant-based foods, adjust subsidies and taxes to reduce meat consumption; (b) Raise awareness of water footprint in diets through educational campaigns; (c) Promote local rural production, support small-scale farming and increase rural self-sufficiency; (d) Enhance women's role in food decisions and provide education opportunities; (e) Using subsidies or tax rebates to help low-income people choose low-WF foods.

Future work should focus on data augmentation and the exploration of more ML methods. Integrating high-quality and up-to-date data from diverse sources would potentially contribute to more accurate and refined ML explanation results. Moreover, given the rapid evolution of ML, it is necessary to explore newer techniques emerging subsequent to this paper. While the study exclusively focused on supervised learning, there exists potential for integrating other types of ML approaches (e.g., unsupervised learning) into the framework to bolster its overall performance and efficacy.

Nomenclature

AUC_PR	Area Under the Precision-Recall Curve
AUC_ROC	Area Under the Receiver Operating Characteristic Curve
BMI	Body Mass Index
CHNS	China Health and Nutrition Survey
ET	Extremely Randomized Trees
FAO	Food and Agriculture Organization of the United Nations
FI	Feature Importance
GBDT	Gradient Boosting Decision Tree

HGB	Histogram-based Gradient Boosting
LR	Logistic Regression
ML	Machine Learning
MR	Multiple Regression
PI	Permutation Importance
PR	Precision-Recall
SHAP	SHapley Additive exPlanations
WF	Water Footprint
XGB	eXtreme Gradient Boosting

Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

Data Availability Statement

The data from the China Health and Nutrition Survey (CHNS) are available at https://www.cpc.unc.edu/projects/china/data/datasets. The data on WF coefficients of Cereals, Legume products, Vegetables, Fruits, Nuts and seeds, and Tea (dry) are available at Mekonnen and Hoekstra (2011a, 2011b). The data on WF coefficients of Meat, Dairy products (low water content), and Eggs are available at Mekonnen and Hoekstra (2010, 2012). The data on WF coefficients of Aquatic foods are available at Yuan et al. (2017). The data on WF coefficients of Pastries and instant foods are available at Chapagain and Orr (2010), Garino (2020). The data on WF coefficients of Soft drinks are available at Ercin et al. (2011), Xie et al. (2015). The data on WF coefficients of Alcoholic beverages are available at Rinaldi et al. (2016), SABMiller (2009).

Acknowledgments
This study was supported by the National Natural Science Foundation of China (52370190). This study uses data from the China Health and Nutrition Survey (CHNS), which is an international collaborative project between the University of North Carolina and the Chinese Center for Disease Control. We extend our gratitude to the dedicated personnel involved in the project, the research grants that have supported its realization, and the organizations that have made significant contributions to its success. Besides, we would like to acknowledge Tang Tan, a master's student from the College of Environmental Science and Engineering at Beijing Forestry University, for her valuable assistance in collecting the water footprint coefficients of foods from relevant literature.

References

Allison, P. D. (1999). *Multiple regression: A primer*. Pine Forge Press.
Altmann, A., Tološi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: A corrected feature importance measure. *Bioinformatics*, *26*(10), 1340–1347. https://doi.org/10.1093/bioinformatics/btq134
Amin, T. T., Al-Sultan, A. I., & Ali, A. (2008). Overweight and obesity and their relation to dietary habits and socio-demographic characteristics among male primary school children in Al-Hassa, Kingdom of Saudi Arabia. *European Journal of Nutrition*, *47*(6), 310–318. https://doi.org/10.1007/s00394-008-0727-6
Bai, J., Wahl, T. I., Lohmar, B. T., & Huang, J. (2010). Food away from home in Beijing: Effects of wealth, time and "free" meals. *China Economic Review*, *21*(3), 432–441. https://doi.org/10.1016/j.chieco.2010.04.003
Bélanger, J., & Pilling, D. (2019). *The state of the world's biodiversity for food and agriculture*. FAO.
Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, *13*(2).
Bhandari, R., & Smith, F. J. (2000). Education and food consumption patterns in China: Household analysis and policy implications. *Journal of Nutrition Education*, *32*(4), 214–224. https://doi.org/10.1016/s0022-3182(00)70559-0
Browne, M. W. (2000). Cross-validation methods. *Journal of Mathematical Psychology*, *44*(1), 108–132. https://doi.org/10.1006/jmps.1999.1279
Bwambale, E., Abagale, F. K., & Anornu, G. K. (2022). Smart irrigation monitoring and control strategies for improving water use efficiency in precision agriculture: A review. *Agricultural Water Management*, *260*, 107324. https://doi.org/10.1016/j.agwat.2021.107324
Chapagain, A., & Orr, S. (2010). Water footprint of Nestlé's Bitesize shredded wheat [Dataset]. *World Wide Fund for Nature*. http://www.waterfootprint.org/Reports/Nestle-2010-Water-Footprint-Bitesize-Shredded-Wheat.pdf
Chen, T., & Guestrin, C. (2016a). Xgboost: A scalable tree boosting system. In *Paper presented at the Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*.
Chen, X., Gao, Z., House, L., Ge, J., Zong, C., & Gmitter, F. (2016). Opportunities for Western food products in China: The case of orange juice demand. *Agribusiness*, *32*(3), 343–362. https://doi.org/10.1002/agr.21453
Choy, L. H., & Li, V. J. (2017). The role of higher education in China's inclusive urbanization. *Cities*, *60*, 504–510. https://doi.org/10.1016/j.cities.2016.04.008
Clark, R. V., Wald, J. A., Swerdloff, R. S., Wang, C., Wu, F. C., Bowers, L. D., & Matsumoto, A. M. (2019). Large divergence in testosterone concentrations between men and women: Frame of reference for elite athletes in sex-specific competition in sports, a narrative review. *Clinical Endocrinology*, *90*(1), 15–22. https://doi.org/10.1111/cen.13840
Deng, H., Runger, G., & Tuv, E. (2011). Bias of importance measures for multi-valued attributes and solutions. In *Paper presented at the artificial neural networks and machine Learning–ICANN 2011: 21st international conference on artificial neural networks, espoo, Finland, June 14–17, 2011, proceedings, Part II 21*.
Dong, Y. (2013). Eating identity: Food, gender, and social organization in late Neolithic northern China.

Ercin, A. E., Aldaya, M. M., & Hoekstra, A. Y. (2011). Corporate water footprint accounting and impact assessment: The case of the water footprint of a sugar-containing carbonated beverage [Dataset]. *Water Resources Management*, *25*(2), 721–741. https://doi.org/10.1007/s11269-010-9723-8

FAO. (2017). *Water for sustainable food and agriculture a report produced for the G20 presidency of Germany*. Food and Agriculture Organization of the United Nations Rome.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Gale, H. F., & Huang, K. (2007). Demand for food quantity and quality in China.

Garino, A. M. (2020). Water footprint for people [Dataset]. https://hdl.handle.net/2445/179083

Garnett, T. (2014). Three perspectives on sustainable food security: Efficiency, demand restraint, food system transformation. What role for life cycle assessment? *Journal of Cleaner Production*, *73*, 10–18. https://doi.org/10.1016/j.jclepro.2013.07.045

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, *63*(1), 3–42. https://doi.org/10.1007/s10994-006-6226-1

Gough, B. (2007). Real men don't diet': An analysis of contemporary newspaper representations of men, food and health. *Social Science & Medicine*, *64*(2), 326–337. https://doi.org/10.1016/j.socscimed.2006.09.011

Gronau, R., & Hamermesh, D. S. (2008). The demand for variety: A household production perspective. *The Review of Economics and Statistics*, *90*(3), 562–572. https://doi.org/10.1162/rest.90.3.562

Han, S., Wu, L., Wang, W., Li, N., & Wu, X. (2019). Trends in dietary nutrients by demographic characteristics and BMI among US adults, 2003–2016. *Nutrients*, *11*(11), 2617. https://doi.org/10.3390/nu11112617

Harris, F., Green, R. F., Joy, E. J., Kayatz, B., Haines, A., & Dangour, A. D. (2017). The water use of Indian diets and socio-demographic factors related to dietary blue water footprint. *Science of the Total Environment*, *587*, 128–136. https://doi.org/10.1016/j.scitotenv.2017.02.085

Hastie, T. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer.

He, P., Feng, K., Baiocchi, G., Sun, L., & Hubacek, K. (2021). Shifts towards healthy diets in the US can reduce environmental impacts but would be unaffordable for poorer minorities. *Nature Food*, *2*(9), 664–672. https://doi.org/10.1038/s43016-021-00350-5

Hoekstra, A. Y., Chapagain, A. K., Aldaya, M. M., & Mekonnen, M. M. (2011). *The water footprint assessment manual: Setting the global standard*. Routledge.

Hosmer, D. W., Jr, Lemeshow, S., & Sturdivant, R. X. (2013a). *Applied logistic regression* (Vol. 398). John Wiley and Sons.

Huang, K. S., & Gale, F. (2009). *Food demand in China: Income, quality, and nutrient effects*. China Agricultural Economic Review.

Karaçil Ermumcu, M. Ş., Çıtar Dazıroğlu, M. E., Erdoğan Gövez, N., & Acar Tek, N. (2023). Evaluation of personal water footprint components in Turkey: Factors associated with obesity and food consumption. *International Journal of Environmental Health Research*, 1–11.

Karl, J. P., Berryman, C. E., Harris, M. N., Lieberman, H. R., Gadde, K. M., Rood, J. C., & Pasiakos, S. M. (2020). Effects of testosterone supplementation on ghrelin and appetite during and after severe energy deficit in healthy men. *Journal of the Endocrine Society*, *4*(4), bvaa024. https://doi.org/10.1210/jendso/bvaa024

Kaufman, P. R., MacDonald, J. M., Lutz, S. M., & Smallwood, D. M. (1997). Do the poor pay more for food? Item selection and price differences affect low-income household food costs.

Kim, B. F., Santo, R. E., Scatterday, A. P., Fry, J. P., Synk, C. M., Cebron, S. R., et al. (2020). Country-specific dietary shifts to mitigate climate and water crises. *Global Environmental Change*, *62*, 101926. https://doi.org/10.1016/j.gloenvcha.2019.05.010

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, *13*, 8–17. https://doi.org/10.1016/j.csbj.2014.11.005

Li, G., Han, X., Luo, Q., Zhu, W., & Zhao, J. (2021). A study on the relationship between income change and the water footprint of food consumption in urban China. *Sustainability*, *13*(13), 7076. https://doi.org/10.3390/su13137076

Liang, Y., Han, A., Chai, L., & Zhi, H. (2020). Using the machine learning method to study the environmental footprints embodied in Chinese diet. *International Journal of Environmental Research and Public Health*, *17*(19), 7349. https://doi.org/10.3390/ijerph17197349

Liao, C., Tan, Y., Wu, C., Wang, S., Yu, C., Cao, W., et al. (2016). City level of income and urbanization and availability of food stores and food service places in China. *PLoS One*, *11*(3), e0148745. https://doi.org/10.1371/journal.pone.0148745

Liao, H., & Chern, W. S. (2007). A dynamic analysis of food demand patterns in urban China.

Lin, B.-H., & Yen, S. T. (2008). Consumer knowledge, food label use and grain consumption in the US. *Applied Economics*, *40*(4), 437–448. https://doi.org/10.1080/00036840600690298

Liu, G. (2014). Food losses and food waste in China: A first estimate.

Liu, J., Yang, H., Gosling, S. N., Kummu, M., Flörke, M., Pfister, S., et al. (2017). Water scarcity assessments in the past, present, and future. *Earth's Future*, *5*(6), 545–559. https://doi.org/10.1002/2016ef000518

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., et al. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, *2*(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

Lundberg, S. M., Erion, G. G., & Lee, S.-I. (2018). Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:.03888*.

Lundberg, S. M., & Lee, S.-I. (2017a). Consistent feature attribution for tree ensembles. *arXiv preprint arXiv:.06060*.

Lundberg, S. M., & Lee, S.-I. (2017b). A unified approach to interpreting model predictions. In *Paper presented at the advances in neural information processing systems*.

Mekonnen, M., & Hoekstra, A. Y. (2010). The green, blue and grey water footprint of farm animals and animal products [Dataset]. *Appendices*, *2*. https://www.waterfootprint.org/resources/Report-48-WaterFootprint-AnimalProducts-Vol1.pdf

Mekonnen, M., & Hoekstra, A. Y. (2011a). National water footprint accounts: The green, blue and grey water footprint of production and consumption [Dataset], *2*. appendices https://www.waterfootprint.org/resources/Report50-NationalWaterFootprints-Vol1.pdf

Mekonnen, M. M., & Hoekstra, A. Y. (2011b). The green, blue and grey water footprint of crops and derived crop products [Dataset]. *Hydrology and Earth System Sciences*, *15*(5), 1577–1600. https://doi.org/10.5194/hess-15-1577-2011

Mekonnen, M. M., & Hoekstra, A. Y. (2012). A global assessment of the water footprint of farm animal products [Dataset]. *Ecosystems*, *15*(3), 401–415. https://doi.org/10.1007/s10021-011-9517-8

Mekonnen, M. M., & Hoekstra, A. Y. (2016). Four billion people facing severe water scarcity. *Science Advances*, *2*(2), e1500323. https://doi.org/10.1126/sciadv.1500323

Meng, X., Gong, X., & Wang, Y. (2009). Impact of income growth and economic reform on nutrition availability in urban China: 1986–2000. *Economic Development and Cultural Change*, *57*(2), 261–295. https://doi.org/10.1086/592838

Miller, A. E. J., MacDougall, J., Tarnopolsky, M., & Sale, D. (1993). Gender differences in strength and muscle fiber characteristics. *European Journal of Applied Physiology and Occupational Physiology*, *66*(3), 254–262. https://doi.org/10.1007/bf00235103

Oncini, F., & Guetto, R. (2018). Cultural capital and gender differences in health behaviours: A study on eating, smoking and drinking patterns. *Health Sociology Review*, *27*(1), 15–30. https://doi.org/10.1080/14461242.2017.1321493
Osborne, J. W., & Waters, E. (2002). Four assumptions of multiple regression that researchers should always test. *Practical Assessment, Research and Evaluation*, *8*(1), 2.
Pang, Z., Yan, D., Wang, T., & Kong, Y. (2021). Disparities and drivers of the water footprint of food consumption in China. *Environmental Science and Pollution Research*, *28*(44), 62461–62473. https://doi.org/10.1007/s11356-021-15125-5
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.
Pingali, P. (2007). Westernization of Asian diets and the transformation of food systems: Implications for research and policy. *Food Policy*, *32*(3), 281–298. https://doi.org/10.1016/j.foodpol.2006.08.001
Popkin, B. M. (1999). Urbanization, lifestyle changes and the nutrition transition. *World Development*, *27*(11), 1905–1916. https://doi.org/10.1016/s0305-750x(99)00094-7
Popkin, B. M., Du, S., Zhai, F., & Zhang, B. (2010). Cohort profile: The China health and nutrition Survey—Monitoring and understanding socio-economic and health change in China, 1989–2011. *International Journal of Epidemiology*, *39*(6), 1435–1440. https://doi.org/10.1093/ije/dyp322
Pukelsheim, F. (1994). The three sigma rule. *The American Statistician*, *48*(2), 88–91. https://doi.org/10.1080/00031305.1994.10476030
Regmi, A. (2001). Changing structure of global food consumption and trade: An introduction. *Changing structure of global food consumption and trade. Anita Regmi*, *1*, 1–3.
Regmi, A., & Dyck, J. (2001). Effects of urbanization on global food demand. Changing structure of global food consumption and trade (pp. 23–30).
Rinaldi, S., Bonamente, E., Scrucca, F., Merico, M. C., Asdrubali, F., & Cotana, F. (2016). Water and carbon footprint of wine: Methodology review and application to a case study [Dataset]. *Sustainability*, *8*(7). 621. https://doi.org/10.3390/su8070621
Rothgerber, H. (2013). Real men don't eat (vegetable) quiche: Masculinity and the justification of meat consumption. *Psychology of Men and Masculinity*, *14*(4), 363–375. https://doi.org/10.1037/a0030379
SABMiller, W. (2009). Water footprinting: Identifying and addressing water risks in the value chain [Dataset]. *SABMiller, Woking, UK, and WWF-UK, Goldalming, UK*. https://assets.wwf.org.uk/downloads/waterfootprinting.pdf
Seyedmohammadi, J., Esmaeelnejad, L., & Ramezanpour, H. (2016). Land suitability assessment for optimum management of water consumption in precise agriculture. *Modeling Earth Systems and Environment*, *2*(3), 1–11. https://doi.org/10.1007/s40808-016-0212-9
Shafi, U., Mumtaz, R., García-Nieto, J., Hassan, S. A., Zaidi, S. A. R., & Iqbal, N. (2019). Precision agriculture techniques and practices: From considerations to applications. *Sensors*, *19*(17), 3796. https://doi.org/10.3390/s19173796
Sodjinou, R., Agueh, V., Fayomi, B., & Delisle, H. (2009). Dietary patterns of urban adults in Benin: Relationship with overall diet quality and socio-demographic characteristics. *European Journal of Clinical Nutrition*, *63*(2), 222–228. https://doi.org/10.1038/sj.ejcn.1602906
Souissi, A., Mtimet, N., McCann, L., Chebil, A., & Thabet, C. (2022). Determinants of food consumption water footprint in the MENA region: The case of Tunisia. *Sustainability*, *14*(3), 1539. https://doi.org/10.3390/su14031539
Springmann, M., Godfray, H. C. J., Rayner, M., & Scarborough, P. (2016). Analysis and valuation of the health and climate change cobenefits of dietary change. *Proceedings of the national academy of sciences* (Vol. 113(15), pp. 4146–4151). https://doi.org/10.1073/pnas.1523119113
Thiele, S., & Weiss, C. (2003). Consumer demand for food diversity: Evidence for Germany. *Food Policy*, *28*(2), 99–115. https://doi.org/10.1016/s0306-9192(02)00068-4
Travassos, G. F., da Cunha, D. A., & Coelho, A. B. (2020). The environmental impact of Brazilian adults' diet. *Journal of Cleaner Production*, *272*, 122622. https://doi.org/10.1016/j.jclepro.2020.122622
Tukker, A., & Jansen, B. (2006). Environmental impacts of products: A detailed review of studies. *Journal of Industrial Ecology*, *10*(3), 159–182. https://doi.org/10.1162/jiec.2006.10.3.159
Vanham, D., Gawlik, B. M., & Bidoglio, G. (2017). Food consumption and related water resources in Nordic cities. *Ecological Indicators*, *74*, 119–129. https://doi.org/10.1016/j.ecolind.2016.11.019
Vanham, D., Hoekstra, A. Y., & Bidoglio, G. (2013). Potential water saving through changes in European diets. *Environment International*, *61*, 45–56. https://doi.org/10.1016/j.envint.2013.09.011
Vanham, D., Mekonnen, M., & Hoekstra, A. Y. (2013). The water footprint of the EU for different diets. *Ecological Indicators*, *32*, 1–8. https://doi.org/10.1016/j.ecolind.2013.02.020
Verma, J., & Abdel-Salam, A.-S. G. (2019). *Testing statistical assumptions in research*. John Wiley and Sons.
Vermeir, I., & Verbeke, W. (2006). Sustainable food consumption: Exploring the consumer "attitude–behavioral intention" gap. *Journal of Agricultural and Environmental Ethics*, *19*(2), 169–194. https://doi.org/10.1007/s10806-005-5485-3
Wang, D., Thunéll, S., Lindberg, U., Jiang, L., Trygg, J., & Tysklind, M. (2022). Towards better process management in wastewater treatment plants: Process analytics based on SHAP values for tree-based machine learning methods. *Journal of Environmental Management*, *301*, 113941. https://doi.org/10.1016/j.jenvman.2021.113941
Wang, D., Thunéll, S., Lindberg, U., Jiang, L., Trygg, J., Tysklind, M., & Souihi, N. (2021). A machine learning framework to improve effluent quality control in wastewater treatment plants. *Science of the Total Environment*, *784*, 147138. https://doi.org/10.1016/j.scitotenv.2021.147138
Wang, X., Shao, S., & Li, L. (2019). Agricultural inputs, urbanization, and urban-rural income disparity: Evidence from China. *China Economic Review*, *55*, 67–84. https://doi.org/10.1016/j.chieco.2019.03.009
Wang, Z., Zhai, F., Zhang, B., & Popkin, B. M. (2012). Trends in Chinese snacking behaviors and patterns and the social-demographic role between 1991 and 2009. *Asia Pacific Journal of Clinical Nutrition*, *21*(2), 253–262.
Wiedmann, T., & Minx, J. (2008). A definition of 'carbon footprint. *Ecological economics research trends*, *1*(2008), 1–11.
Wikström, F., Williams, H., Verghese, K., & Clune, S. (2014). The influence of packaging attributes on consumer behaviour in food-packaging life cycle assessment studies-a neglected topic. *Journal of Cleaner Production*, *73*, 100–108. https://doi.org/10.1016/j.jclepro.2013.10.042
Xie, G., Cao, S., Yang, Q., Xia, L., Fan, Z., Gao, Y., et al. (2015). Living planet report–China 2015 [Dataset]. *Beijing: World Wildlife Fund*. https://www.wwfchina.org/content/press/publication/2015/Living%20Planet%20Report%20China%202015%20FIN.pdf
Yau, A., White, M., Hammond, D., White, C., & Adams, J. (2020). Socio-demographic characteristics, diet and health among food insecure UK adults: Cross-sectional analysis of the international food policy study. *Public Health Nutrition*, *23*(14), 2602–2614. https://doi.org/10.1017/s1368980020000087
Yuan, M., Seale Jr, J. L., Wahl, T., & Bai, J. (2019). The changing dietary patterns and health issues in China. *China Agricultural Economic Review*, *11*(1), 143–159. https://doi.org/10.1108/caer-12-2017-0254

Yuan, Q., Song, G., Fullana-i-Palmer, P., Wang, Y., Semakula, H. M., Mekonnen, M. M., & Zhang, S. (2017). Water footprint of feed required by farmed fish in China based on a Monte Carlo-supported von Bertalanffy growth model: A policy implication [Dataset]. *Journal of Cleaner Production*, *153*, 41–50. https://doi.org/10.1016/j.jclepro.2017.03.134

Zhou, Y., Du, S., Su, C., Zhang, B., Wang, H., & Popkin, B. M. (2015). The food retail revolution in China and its association with diet and health. *Food Policy*, *55*, 92–100. https://doi.org/10.1016/j.foodpol.2015.07.001

References From the Supporting Information

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/a:1010933404324

Chen, T., & Guestrin, C. (2016). *Xgboost: A scalable tree boosting system* (pp. 785–794). ACM.

Hosmer, D. W., Jr, Lemeshow, S., & Sturdivant, R. X. (2013b). *Applied logistic regression*. John Wiley and Sons.

Powers, D. M. (2020). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*.

Vujović, Ž. (2021). Classification model evaluation metrics. *International Journal of Advanced Computer Science and Applications*, *12*(6), 599–606. https://doi.org/10.14569/ijacsa.2021.0120670