computation

MDPI

Article

Efficient Data-Driven Machine Learning Models for Water Quality Prediction

Elias Dritsas * and **Maria Trigka**

Department of Computer Engineering and Informatics, University of Patras, 26504 Patras, Greece
* Correspondence: dritsase@ceid.upatras.gr

Abstract: Water is a valuable, necessary and unfortunately rare commodity in both developing and developed countries all over the world. It is undoubtedly the most important natural resource on the planet and constitutes an essential nutrient for human health. Geo-environmental pollution can be caused by many different types of waste, such as municipal solid, industrial, agricultural (e.g., pesticides and fertilisers), medical, etc., making the water unsuitable for use by any living being. Therefore, finding efficient methods to automate checking of water suitability is of great importance. In the context of this research work, we leveraged a supervised learning approach in order to design as accurate as possible predictive models from a labelled training dataset for the identification of water suitability, either for consumption or other uses. We assume a set of physiochemical and microbiological parameters as input features that help represent the water's status and determine its suitability class (namely safe or nonsafe). From a methodological perspective, the problem is treated as a binary classification task, and the machine learning models' performance (such as Naive Bayes–NB, Logistic Regression–LR, k Nearest Neighbours–kNN, tree-based classifiers and ensemble techniques) is evaluated with and without the application of class balancing (i.e., use or nonuse of Synthetic Minority Oversampling Technique–SMOTE), comparing them in terms of Accuracy, Recall, Precision and Area Under the Curve (AUC). In our demonstration, results show that the Stacking classification model after SMOTE with 10-fold cross-validation outperforms the others with an Accuracy and Recall of 98.1%, Precision of 100% and an AUC equal to 99.9%. In conclusion, in this article, a framework is presented that can support the researchers' efforts toward water quality prediction using machine learning (ML).

Keywords: water quality; sustainability; environmental impact; data analysis; machine learning; classification; prediction

check for updates

Citation: Dritsas, E.; Trigka, M. Efficient Data-Driven Machine Learning Models for Water Quality Prediction. *Computation* **2023**, *11*, 16. https://doi.org/10.3390/computation11020016

Academic Editors: Ting-Chieh Lin and Xiaoqiang Hua

Received: 9 December 2022
Revised: 10 January 2023
Accepted: 16 January 2023
Published: 18 January 2023

1. Introduction

Water has been characterized as a source of life by the World Health Organisation. It covers 70% of the Earth's surface, 50–60% of human weight and 90% of our cells. World Water Day [1] was established in 1992 by the United Nations and is celebrated annually on March 22 to raise public awareness of the importance of drinking water and the sustainable management of water resources [2,3].

While water is an essential nutrient for human health, easy access to it makes us take it for granted, so we do not realise the importance of proper hydration. According to experts, water ranks second as a source of oxygen, which is essential for life, since the human body is made up of 2/3 of water. A person can survive for several days without food but only a few days without water. It is the main component of the human body and plays an important role in all stages of human development [2,4].

In addition, water is necessary to regulate body temperature, keeping it cool through sweat. Water cleanses the body of toxins and unnecessary substances and is an important component of the blood, as it transfers nutrients and oxygen to and from all cells. It also provides a moist environment for all tissues of the body and is the main component of saliva and mucous membranes that lubricate the membranes that activate our digestive system starting from the mouth. Finally, it helps maintain a healthy weight [4,5].

In case the water level in the body decreases, there are mechanisms that reduce the loss of water through the kidneys, which have the ability to collect urine, thus reducing the loss of water from them. However, there are situations in which the water balance becomes negative, and then, unfortunately, there is a risk of dehydration, which is life-threatening. Severe dehydration poses a problem with blood pressure, as it drops dangerously and blood circulation is reduced, resulting in poor oxygenation of tissues and not supplying tissues and organs with the necessary nutrients. In this case, there is intense dizziness, loss of consciousness, tachycardia and renal failure, while death can occur [6,7].

Water is a key component of an ecosystem. The natural environment and the organisms that live in it are the links of a food chain that is directly dependent on water. When a link in the food chain is broken, problems will arise, with the final recipient being humans, who are at the top of the food pyramid [8,9].

During the last decades, there has been an increase in environmental pollution from human activities. Water quality is directly affected by discharges, such as from factories or sewage treatment plants. It can also be affected by pollution from diffuse sources, such as nutrients and pesticides from agricultural activities and pollutants released into the air by industry, which then fall to land and sea. The urban wastewater and commercial activity within the urban fabric contribute to the greatest extent to the pollution of water resources [10,11].

To be classified as potable, water is required to meet certain specifications based on its quality characteristics. In this way, the basis is given to the substances contained in the water, and the concentration limits must be observed so that the water can be consumed safely. This can be achieved either by physicochemical water testing, where pH, conductivity, total solids, suspended solids, total hardness, total alkalinity, bicarbonates, carbonates and chlorides are tested, by microbiological water testing that checks for coliform bacteria, escherichia coli and enterococci or by analysis of trace elements and pollutants, where a check is made for nitrites, nitrates, phosphates, ammonia, potassium, disinfectants, fluorides, hydrogen sulphide, hydrogen cyanide, phenols and selected heavy metals [12–14].

In recent years, actions have been taken by organised international bodies and governments to reduce the reckless use of water and protect and promote public health by minimising the sources of water pollution, making it appropriate for consumption [15]. In a changing environment, the need for efficient management and safe provision of drinking water [16] has increased researchers' attempts to develop appropriate tools for continuous water monitoring to sustain its quality. This process is known as water qualification. The water quality index (WQI) constitutes a simple but still efficient qualification approach that scientists represent and calculate as a weighted combination of several features' values [17]. The water quality's rating may be based on the relative importance of each parameter–feature of the WQI and may be calculated in various ways, such as the inverse proportion of recommended standard rules for each parameter. By thresholding the measured data, the related level of suitability, which helps decision-makers to interpret and indicate if water is appropriate for daily use and consumption, can be estimated.

Traditional methods are time-consuming, as they require an expert to decide on the basis of the measured data. Nowadays, water quality measuring has benefited from the advances in Information and Communication Technologies (ICTs). More specifically, wireless sensor networks and smart monitoring systems using Internet-of-Things (IoT) technology have ensured the availability of various big sensor data, while the fields of Artificial Intelligence (AI) and Machine Learning (ML) have provided scientists with efficient methods for accurate data collection, advanced processing and analytics [18]. In the relevant literature, the water qualification problem is tackled as either a time series prediction of water quality parameters [19] or a classification task [20] based on WQI values in order to assign a class label to an undetermined water sample.

The scope of this research work is to present an automated process by adopting data-driven approaches founded on high-accuracy ML classifiers for the design of water quality prediction models. According to the WHO guidelines for the computation of water quality, a wide list of parameters (physicochemical, microbiological and heavy metals) can be included; thus, the features of ML models depend on the potential sources of contamination and how they can be controlled. The wider the range of features (namely data sources) used (so as to avoid overfitting), the more accurate the results of the classification process will be. Note that we do not emphasise data collection but training models for an as accurate as possible water quality prediction using a public dataset with already labelled data.

The research question that the current study attempts to answer is if water under specific features (namely variables that have the potential to have adverse effects on human health) is safe for consumption/use or not. To answer this question, we formulate a binary classification problem, where the water quality class was derived on the basis of the WHO standards applied to the involved features (chemical and biological), and thus in the WQI. However, the nature and form of drinking-water standards may vary among countries and regions. The main contributions of this research article are the following:

- A data preprocessing step that exploits the SMOTE is performed. In this way, we create a balanced dataset and, thus, we can design efficient classification models unbiased to the safe or nonsafe classes.
- A features analysis is made, which includes: (i) a statistical description of the numeric features and (ii) order of importance evaluation by employing three different methods.
- A comparative evaluation of numerous ML classification models, namely probabilistic, distance-based, tree-based and Ensemble Learning is performed. For the purpose of this study, NB, LR, Artificial Neural Network (ANN), kNN, Rotation Forest (RotF), AdaBoostM1, Random Forest (RF), Stacking, Voting and Bagging are selected in order to develop the intended model with the highest accuracy and discrimination ability after SMOTE with 10-fold cross-validation.
- For the models' evaluation, we considered the performance metrics Accuracy, Recall, Precision and AUC. Moreover, AUC ROC curves are also captured and presented.
- Finally, from various aspects, the performance analysis revealed that the Stacking classification model after SMOTE with 10-fold cross-validation outperforms the others, and thus it constitutes the proposition of this study.

Based on the above points, the potential readers have the chance to see from a practical perspective the application of ML in the field of water sustainability, which has an essential impact on human life, and understand the superiority of the SMOTE technique and Ensemble Learning in order to train efficient models.

The rest of the paper is structured as follows. In Section 2, we capture related works on the prediction of water quality using ML techniques and models. Then, in Section 3, we describe the dataset we relied on and analyse the methodology we followed. In addition, in Section 4, we discuss and evaluate the research experimental results. Finally, conclusions and future directions are mentioned in Section 5.

2. Related Works

This section provides a brief discussion of recent works on the subject under consideration. Machine Learning techniques and a variety of well-known algorithms have been applied in order to capture with high accuracy the quality of water.

Firstly, the proposed methodology in [21] employs four input parameters (temperature, turbidity, pH and total dissolved solids). Gradient Boosting, with a learning rate of 0.1 and polynomial regression of second degree, was the most efficient predictor of the water quality index, having a mean absolute error of 1.9642 and 2.7273, respectively. However, MultiLayer Perceptron (MLP), with a configuration of (3, 7), was the most efficient classifier for the water quality prediction, with an accuracy of 0.8507.

Moreover, the paper in [22] proposes the use of enhanced Wavelet De-noising Techniques using Neuro-Fuzzy Inference Systems (WDT-ANFIS). Dual scenarios were presented: Scenario 1 was designed to confirm prediction models for water quality parameters at each station according to 12 input parameters, whereas Scenario 2 was designed to confirm prediction models for water quality parameters according to 12 input parameters, as well as the parametric values from prior upstream stations. In comparisons between the two scenarios, the second achieved higher accuracy in terms of simulating the patterns and magnitudes for every water quality parameter at every station.

Furthermore, in [23], four algorithms, namely RF, M5P, Random Tree (RT) and Reduced Error Pruning Tree (RepTree), and 12 hybrid data-mining algorithms (combinations of standalone with Bagging, CV Parameter Selection (CVPS) and Randomisable Filtered Classification) were used to create the Iran water quality index (IRWQIsc) predictions. Hybrid Bagging–Random Forest outperformed the other models (R2 = 0.941, RMSE = 2.71, MAE $\cong$ 1.87, NSE = 0.941 and PBIAS = 0.500).

In addition, the basic models of the two hybrid ones in [24] are Extreme Gradient Boosting and RF, which, respectively, introduce an advanced data denoising technique–complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN). The results show that the prediction stability of CEEMDAN–Random Forest and CEEMDAN–Extreme Gradient Boosting is higher than other benchmark models for short-term water quality prediction.

The authors in [25], utilised an ANN in order to develop a novel application to predict water quality resilience to simplify resilience evaluation. The Fuzzy Analytic Hierarchy Process method is used to rank water basins based on their level of resilience and to identify the ones that demand prompt restoration strategies.

Similarly, in [26], water quality was evaluated via the detection of anomalies occurring in time series data. For this purpose, the performance of several models, such as LR, Linear Discriminant Analysis (LDA), Support Vector Machine (SVM), ANN, Deep Neural Network (DNN), Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) is assessed using the F1 score metric. The best F1 score is achieved using the SVM model.

Moreover, the authors in [27] predicted water quality components of the Tireh River using well-known models, including ANN, Group Method of Data Handling (GMDH) and SVM. The best performance was related to the SVM.

Finally, the main purpose of the current study is to present a general methodology for water quality prediction by leveraging supervised learning models. The adopted methodology is irrespective of what features are used to capture the water status. The class variable is the water quality index with two possible states "safe" and "nonsafe". We do not emphasise dataset engineering but the investigation of several classification schemes using single classifiers (such as SVM, NB, RF, etc) and Ensemble Learning (Voting, Stacking and Bagging). We consider an adequate set of labelled data with which a list of models is trained and tested (after the application of class balancing) to identify the one with the highest performance metrics.

3. Materials and Methods

3.1. Dataset Description

Our research work was based on a publicly available dataset [28]. The present dataset contains 7986 instances, and the percentage of measurements that expresses whether the water is suitable is 11.4% (910 instances). Finally, the number of features as input to the ML models is 20, and the target class (safe) is described as follows:

- **Aluminium** (Al) [29]: This feature captures the amount of aluminium in one litre of water (mg/L).
- **Ammonia** (NH3) [30]: This feature denotes the amount of ammonia in one litre of water (mg/L).
- **Arsenic** (As) [31]: This feature stands for the amount of arsenic in one litre of water (μg/L).
- **Barium** (Ba) [32]: This feature captures the amount of barium in one litre of water (mg/L).

- **Cadmium** (Cd) [33]: This feature records the amount of cadmium in one litre of water (mg/L).
- **Chloramine** (NH2Cl) [34]: This feature captures the amount of chloramine in one litre of water (mg/L).
- **Chromium** (Cr) [35]: This feature shows the amount of chromium in one litre of water (mg/L).
- **Copper** (Cu) [36]: This feature records the amount of copper in one litre of water (mg/L).
- **Fluoride** (F) [37]: This feature captures the amount of fluoride in one litre of water (mg/L).
- **Bacteria** [38]: This feature shows the number of bacteria in one litre of water.
- **Viruses** [39]: This feature captures the number of viruses in one litre of water.
- **Lead** (Pb) [40]: This feature denotes the amount of lead in one litre of water (µg/L).
- **Nitrates** ($NO3^-$) [41]: This feature captures the number of nitrates in one litre of water (mg/L).
- **Nitrites** ($NO2^-$) [42]: This feature shows the number of nitrites in one litre of water (mg/L).
- **Mercury** (Hg) [43]: This feature captures the amount of mercury in one litre of water (mg/L).
- **Perchlorate** ($ClO4^-$) [44]: This feature captures the amount of perchlorate in one litre of water (mg/L).
- **Radium** (Ra) [45]: This feature captures the amount of radium in one litre of water (pCi/L).
- **Selenium** (Se) [46]: This feature captures the amount of selenium in one litre of water (µg/L).
- **Silver** (Ag) [47]: This feature captures the amount of silver in one litre of water (µg/L).
- **Uranium** (U) [48]: This feature captures the amount of uranium in one litre of water (mcg/L).
- **Safe**: This feature captures whether the water is safe for consumption or not.

  All the attributes are numerical, except for the target class (safe), which is nominal.

3.2. Proposed Methodology

In this subsection, we focus on data preprocessing, feature analysis i.e., statistical description and their importance evaluation, and selection of ML models for water suitability assessment.

3.2.1. Data Preprocessing

As for the current dataset, the nonuniform class distribution of the "safe" and "nonsafe" instances was tackled by employing SMOTE [49]. SMOTE uses a 5-NN classifier in order to create synthetic data [50] on the minority class, i.e., "safe", which is oversampled such that the instances in two classes are equally distributed (i.e., 50–50%). After the implementation of SMOTE, the number of instances is 14,152. Now, the dataset is balanced, and the class label includes 7076 safe and 7076 nonsafe instances.

3.2.2. Features Analysis

In Table 1, we present the statistical characteristics of the numeric features, namely, minimum, maximum, mean and standard deviation in the balanced data.

Moreover, in Table 2, we present the dataset features' importance concerning the "safe" class. Three methods are considered to apply feature ranking. The former exploits Pearson Correlation Coefficient (CC) [51] and the latter the Gain Ratio (GR) method. Additionally, the Random Forest classifier is utilised to assign a ranking score.

Table 1. Statistical Description of the Numerical Features.

Feature	Min	Max	Mean ± Stdv
aluminium	0	5.05	1.162 ± 1.446
cadmium	0	0.13	0.032 ± 0.034
chloramine	0	8.68	2.749 ± 2.536
chromium	0	0.9	0.304 ± 0.265
arsenic	0	1.05	0.123 ± 0.214
viruses	0	1	0.28 ± 0.346
silver	0	0.5	0.166 ± 0.141
barium	0	4.94	1.693 ± 1.156
uranium	0	0.09	0.042 ± 0.025
perchlorate	0	60.01	18.034 ± 16.397
nitrates	0	19.83	9.32 ± 5.522
radium	0	7.99	3.092 ± 2.233
nitrites	0	2.93	1.355 ± 0.508
mercury	0	0.01	0.005 ± 0.003
selenium	0	0.1	0.049 ± 0.027
copper	0	2	0.829 ± 0.607
bacteria	0	1	0.309 ± 0.305
ammonia	0	29.84	14.045 ± 8.717
lead	0	0.2	0.099 ± 0.054
fluoride	0	1.5	0.773 ± 0.403

Table 2. Measure of Features' Importance with Three Different Methods: Correlation Coefficient, Gain Ratio and Random Forest.

Pearson Correlation Coefficient		Gain Ratio		Random Forest	
Feature	**Rank**	**Feature**	**Rank**	**Feature**	**Rank**
aluminium	0.44842	aluminium	0.14809	cadmium	0.46
cadmium	0.42867	cadmium	0.13597	chromium	0.439
chloramine	0.29334	uranium	0.12884	aluminium	0.433
chromium	0.28024	selenium	0.12718	arsenic	0.421
arsenic	0.23094	mercury	0.12579	nitrites	0.417
viruses	0.17795	arsenic	0.11269	fluoride	0.417
silver	0.16577	viruses	0.1003	lead	0.414
barium	0.14402	silver	0.09952	viruses	0.414
uranium	0.13252	chromium	0.09493	copper	0.414
perchlorate	0.12497	chloramine	0.08406	silver	0.408
nitrates	0.11617	nitrites	0.0744	barium	0.401
radium	0.10245	bacteria	0.07052	uranium	0.396
nitrites	0.06979	copper	0.06093	selenium	0.388
mercury	0.06801	perchlorate	0.0483	bacteria	0.384
selenium	0.05586	fluoride	0.0426	mercury	0.382
copper	0.0485	lead	0.03587	chloramine	0.382
bacteria	0.04449	barium	0.03404	radium	0.372
ammonia	0.03705	radium	0.02803	nitrates	0.249
lead	0.00842	nitrates	0.01322	perchlorate	0.227
fluoride	0.00494	ammonia	0.00413	ammonia	0.181

The Pearson correlation score is used to infer the strength of the association between the class and a feature. We observe a moderate correlation of 0.44842 and 0.42867 between the "safe" class and aluminium and cadmium, respectively. A low association of 0.29334 and 0.28024 is shown to have the "safe" class with chloramine and chromium, correspondingly. Moreover, a weak association is shown to have the "safe" class with viruses, silver, barium, etc. Moreover, negligible correlation is observed with the rest of the features, where the rank is lower than 0.1 and close to 0.

Random Forest calculates the importance of a feature based on the purity index. Its importance increases with the increase in this index in leaves. It is applied in each tree, averaged among all the trees and normalised such that the sum of the importance scores is equal to 1 [52]. The Gain Ratio [53,54] evaluates the worth of an attribute x according to the formula $GR(c,x) = \frac{H(c)-H(c|x)}{H(x)}$, where $H(c|x)$, $H(c)$ and $H(x)$ are the entropy of the c class, the conditional entropy of the class given an attribute and the entropy of the feature x.

We see that each method has derived a different ranking order, and there are scores close to zero, which shows that these features may not contribute to the models' performance enhancement. However, all features are considered for the models' training, since they are important risk factors for water quality prediction.

3.3. Machine Learning Models

In this study, for the topic under consideration, several ML models were employed to identify which one performs better than the rest by evaluating their prediction performance. More specifically, we focused on NB [55], which is a probabilistic classifier. From ensemble ML algorithms [56], Bagging [57], RotF built upon decision trees [58], RF [59], AdaBoostM1 [60], Voting [61] and Stacking [62] were exploited. Finally, a simple ANN Network (i.e., MLP) [63], LR [64] and kNN [65], a distance-based classifier, were evaluated.

4. Results and Discussion

4.1. Experiments Setup

We based the evaluation of our ML models on the Waikato Environment for Knowledge Analysis (Weka) [66], which is free software developed at the University of Waikato, New Zealand. This tool offers a library of various models for classification, clustering, prediction, preprocessing and visualisation. In addition, the experiments were performed on a computer system with the following specifications: 11th generation Intel(R) Core(TM) i7-1165G7 @ 2.80GHz, RAM 16GB, Windows 11 Home, 64-bit OS and x64 processor. For our experiment results, 10-fold cross-validation was applied in order to measure the models' efficiency in the balanced dataset of 14,152 instances after SMOTE. Finally, in Table 3, we illustrate the optimal parameter settings of the ML models that we experimented with and, in Figure 1, we capture how the Voting and Stacking ensembles achieve the classification of an uncategorized instance.

Table 3. Machine Learning Models' Settings.

Models	Parameters	Models	Parameters
NB	use kernel estimator: False use supervised discretization: True	**RotF**	classifier: RF number of groups: True projection filter: PrincipalComponents
LR	ridge = 10^{-8} use conjugate gradient descent: True	**AdaBoostM1**	classifier: RF resume: True use resampling : True
MLP	learning rate = 0.1 momentum = 0.2 training time = 200	**Stacking**	classifiers: RF and NB meta classifier: LR

kNN	k = 3
	search algorithm: LinearNNSearch
	with Euclidean
	cross-validate = True
		Voting	classifiers: RF and NB
			combination rule: average
			of probabilities

RF	break ties radomly: True
	numIterations = 500
	store out of bag predictions: True
		Bagging	classifiers: RF
			print classifiers: True
			store out of bag predictions: True

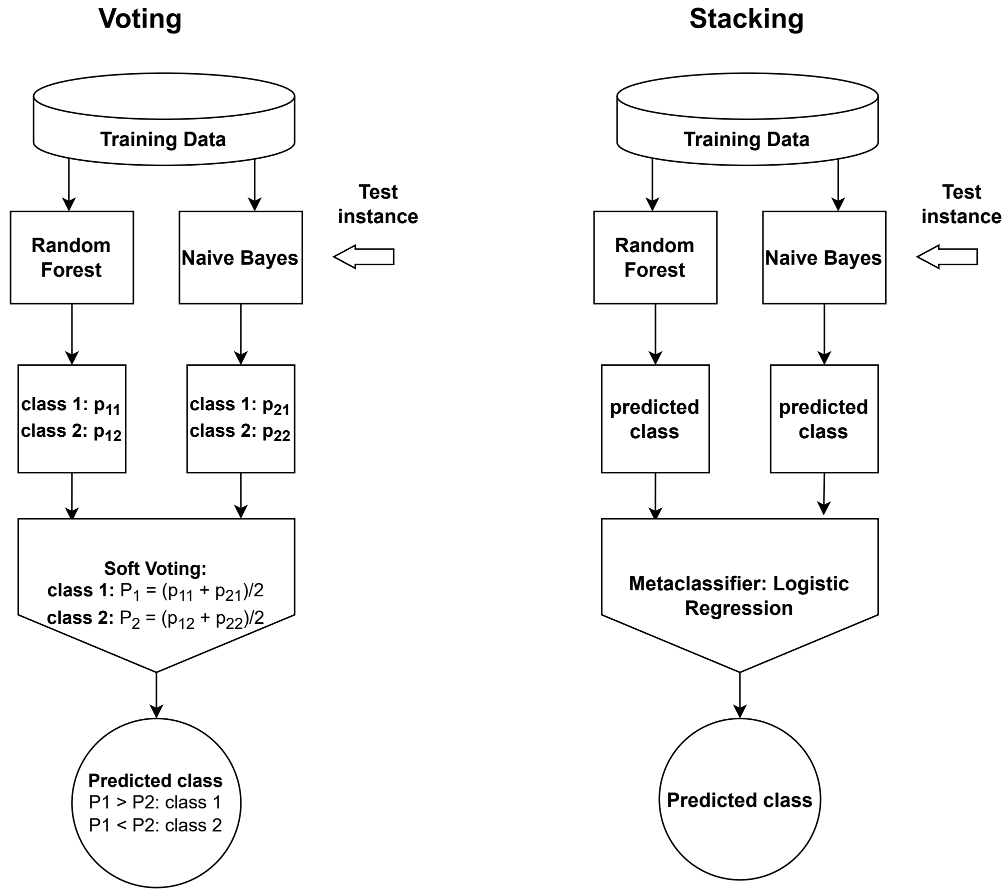

Figure 1. An overview of the Voting and Stacking ensemble schemes.

4.2. Evaluation

In the context of the ML models' evaluation, we consider Accuracy, Recall, Precision and Area Under Curve (AUC), which are commonly used metrics in the relevant literature [67].

Accuracy summarises the performance of the classification task and measures the number of correctly predicted instances out of all the data instances. Moreover, Recall is an appropriate metric to identify the errors of a model and how accurately the model identifies the true "safe" and "nonsafe" instances, respectively. Precision indicates the proportion of positive (either "safe" or "nonsafe") identifications that was actually correct. Precision is a measure of quality, while Recall is a measure of quantity. The aforementioned metrics are defined as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad \text{Accuracy} = \frac{\text{TN} + \text{TP}}{\text{TN} + \text{TP} + \text{FN} + \text{FP}} \tag{1}$$

where TP, TN, FP and FN stand for True Positive, True Negative, False Positive and False Negative, respectively. Finally, AUC values were recorded in order to evaluate the models' efficiency. They vary between zero and one and are leveraged to identify the ML model with the best performance in distinguishing "safe" from "nonsafe" instances.

In the context of this research work, plenty of Machine Learning models, such as NB, LR, ANN, kNN and Ensemble Learning (RotF, AdaBoostM1, RF, Stacking, Voting, and Bagging) were evaluated in terms of Accuracy, Precision, Recall and AUC in order to determine the model with the best predictive performance. The performance outcomes of the employed models are illustrated under two different cases, before and after applying the SMOTE technique. To validate their performance on unseen data, 10-fold cross-validation was applied. We adopted this method in order to estimate how the model is expected to perform when used to make predictions on data not used during the training stage. This method randomly divides the set into 10 groups, or folds, of approximately uniform size. The first fold is treated as a test/validation set, and the remaining nine folds are used to fit each one of the selected models. The process is repeated until all folds are used as a test set, and their average performance is considered.

Observing Table 4, we see that NB and 3NN are the most favoured models in terms of Precision and Recall, respectively. Moreover, NB and all ensemble models except for Bagging attained Precision equal to the upper limit of 1. Moreover, LR is the second model in the list that considerably improved its Precision from 0.553 to 1. Focusing on the Recall metric, LR and 3NN achieved similar low performances in the "No SMOTE" case, while, after class balancing, they speeded up their recall 2.35 and 2.92 times, correspondingly. RotF and MLP are the next models that improved their performance by 1.85 and 1.4 times, respectively. Accuracy and AUC are the metrics that benefited less than the other two metrics due to class balancing.

Table 4. Performance Evaluation before and after SMOTE.

	Accuracy		**Precision**		**Recall**		**AUC**	
	NO SMOTE	**SMOTE**	**NO SMOTE**	**SMOTE**	**NO SMOTE**	**SMOTE**	**NO SMOTE**	**SMOTE**
NB	0.819	0.927	0.358	1	0.745	0.854	0.838	0.978
LR	0.907	0.798	0.692	0.805	0.335	0.787	0.861	0.879
MLP	0.944	0.924	0.807	0.914	0.669	0.937	0.882	0.972
3NN	0.893	0.886	0.553	0.822	0.334	0.978	0.767	0.937
RF	0.965	0.970	0.920	0.963	0.757	0.978	0.986	0.996
RotF	0.938	0.953	0.891	0.943	0.520	0.964	0.966	0.992
AdaBoostM1	0.965	0.970	0.925	0.963	0.751	0.977	0.985	0.996
Stacking	0.967	0.981	0.886	1	0.816	0.981	0.979	0.999
Bagging	0.962	0.968	0.932	0.962	0.723	0.974	0.986	0.996
Voting	0.919	0.929	0.624	1	0.736	0.861	0.948	0.980

LR fits a line to separate the space exactly into two regions and predicts the class of an unknown input feature vector. RotF and RF are based on a decision tree exploiting all features, and thus essentially prevail, with a performance gap depending on the evaluation metric. RF is a robust model, as it combines the output of multiple decision trees to come up with a final prediction, achieving higher accuracy than all versions of decision trees. Analysing the Ensemble Learning methods, they all perform similarly, since they use the same well-behaved estimators as base classifiers (RF and AdaBoostM1). In particular, the Bagging process applies random sampling with replacement in the initial dataset, independent and parallel training of AdaBoostM1 learners and, finally, simple averages of the outputs of single classifiers. Voting (specifically soft) averages the outcomes of two different base models, RF and AdaBoostM1, which are trained independently in the same data, while Stacking trains a metaclassifier with the outputs of the single classifiers. Among them, Stacking shows a bit lower results, since it combines the outcomes of RF and AdaBoosM1 into a weaker LR classifier.

Let us focus on the SMOTE case. From the experimental results, we see that our proposed models show very good performance. The Voting classifier, which has as base classifiers the RF and the AdaBoostM1, outperforms the other models, with an Accuracy, Precision and Recall of 92.9%, 100% and 86.1%, respectively, and an AUC equal to 98%. Similar high results are presented by the classifiers AdaBoostM1 and Bagging, where the latter utilises as base classifier the AdaBoostM1, with an Accuracy, Precision and Recall of (97%, 96.8%), (96.3%, 96.2%) and (99.85%, 99.86%) and an AUC equal to 99.6%. The Stacking classifier, which has as base classifiers the RF and the AdaBoostM1, and as a metaclassifier the LR, achieved an Accuracy, Precision and Recall of 98.1%, 100% and 98.1% and an AUC equal to 99.9%. Regarding the AUC, the Stacking classifier prevails over the rest, with a performance of 99.9%. RF, AdaBoostM1 and Bagging perform the same, with 99.6%.

An important aspect of the study is the illustration of the AUC ROC curves. More specifically, in Figures 2 and 3, we plot the ROC curve of the proposed ML models before and after the application SMOTE, where it is shown that the performances of the Bagging, Stacking, AdaBoostM1 and RF models are identical. The AUC ROC is a probability curve that plots the True Positive Rate (TPR) or Recall against the False Positive Rate (FPR) at various threshold values, while the AUC measures the ability of a classifier to distinguish between classes (safe, nonsafe), summarising the ROC curve. From these curves, we illustrate that the ensemble methods perform better since they decrease FPR and improve the TPR at the same time. Hence, from AUC values and the corresponding ROC curves of the ensemble models, it is expected that the distribution curves of the two class instances will have small overlaps, which means that they tend to have an ideal separation ability between safe and nonsafe classes. Moreover, from the recall metric, we observe that the aforementioned models have high sensitivity in the correct identification of the instances' class, e.g., that actually belong to the "safe" class. The performance improvement of ensemble methods is also reflected in the rest of the measures.

In concluding the evaluation of our models, we have to note a limitation of this study. The dataset [28] which we relied on is artificial. Hence, it did not come from a research institute, which could give us richer information models with different characteristics trained on real-world data. Nevertheless, it had many numerical features that led us to high-accuracy results by applying ML models.

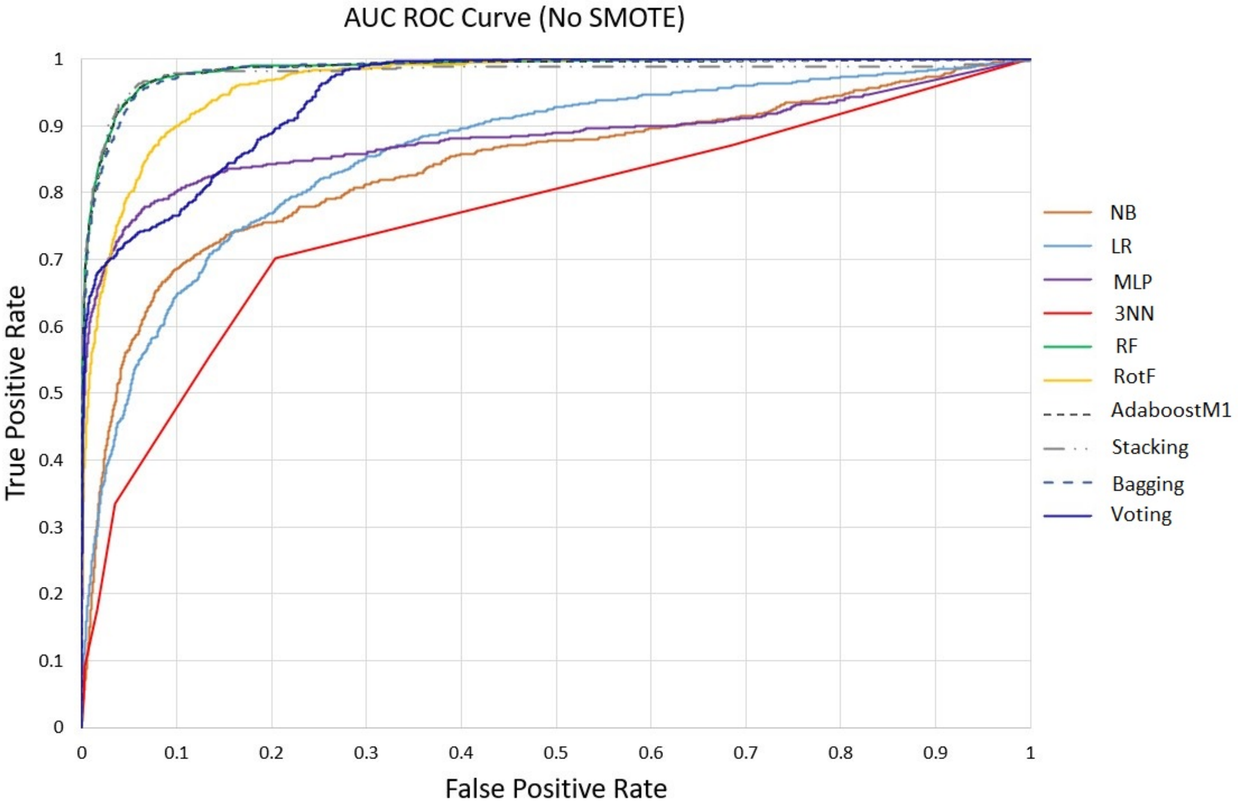

Figure 2. ML models AUC ROC vurve without SMOTE.

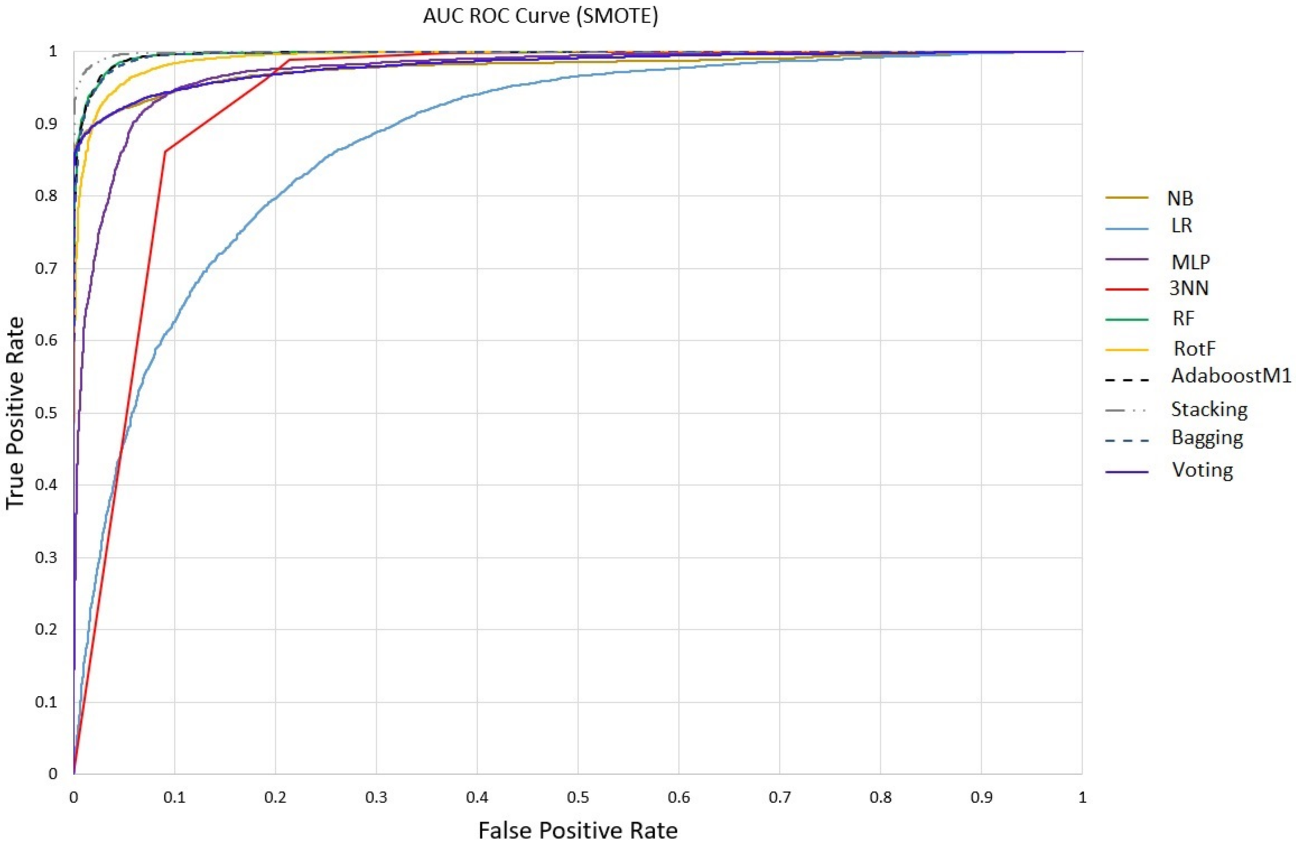

Figure 3. ML models AUC ROC curve after SMOTE.

5. Conclusions

Water is undoubtedly a valuable and simultaneously rare commodity in both developing and developed countries all over the world. In recent years, actions have been taken by organised international bodies and governments to reduce the reckless use of water and to minimise the sources of water pollution, making it unfit for consumption or other uses.

Water quality prediction was the point of interest in this study. In order to achieve this, a comparative evaluation of numerous ML classification models, such as NB, LR, ANN, kNN, RotF and Ensemble Learning (RF, AdaBoostM1, Stacking, Voting and Bagging) was made in order to develop the intended model with the highest accuracy and discrimination ability after SMOTE with 10-fold cross-validation. The experiment results show that the Stacking classification model after SMOTE with 10-fold cross-validation outperforms the others with an Accuracy, Precision and Recall of 98.1%, 100% and 98.1%, respectively, and an AUC equal to 99.9%. Furthermore, the AUC ROC curves constitute an important illustration metric which reveals the high separation ability of the two class instances (safe, nonsafe) that tree-based ensemble methods have, and thus prevail against the single ones. All in all, ML methods constitute a challenging and alternative tool in the field of water sustainability, which has attracted research interest due to its significant impact on human life.

In future work, we aim to reconsider the water quality prediction methodology after ROPE (Region of Practical Equivalence) analysis to test whether a feature is important for water quality prediction and, second, to extend the ML framework through the use of Deep Learning methods by applying LSTM algorithms and Convolutional Neural Networks (CNN) and evaluate them using real-world data on the aforementioned metrics.

Author Contributions: E.D. and M.T. conceived of the idea, designed and performed the experiments, analysed the results, drafted the initial manuscript and revised the final manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found on the following website: https://www.kaggle.com/datasets/mssmartypants/water-quality (accessed on 8 December 2022).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. World Water Day. Available online: https://www.worldwaterday.org/ (accessed on 9 December 2022).
2. Khikmatovna, H.S. Drinking water quality source of life. *Web Sci. Int. Sci. Res. J.* **2021**, *2*, 35–40.
3. Fateeva, K.V.; Filimonova, N.G. THE WATER IS THE SOURCE OF LIFE. THE PROBLEMS OF POLLUTION OF WATER SOURCES. In Proceedings of the Experientia Est Optima Magistra, Belgorod, Russia, 11–12 April 2018; pp. 20–22.
4. Westall, F.; Brack, A. The importance of water for life. *Space Sci. Rev.* **2018**, *214*, 1–23. [CrossRef]
5. Ward, M.H.; Jones, R.R.; Brender, J.D.; De Kok, T.M.; Weyer, P.J.; Nolan, B.T.; Villanueva, C.M.; Van Breda, S.G. Drinking water nitrate and human health: An updated review. *Int. J. Environ. Res. Public Health* **2018**, *15*, 1557. [CrossRef] [PubMed]
6. Hooper, L.; Bunn, D.; Jimoh, F.O.; Fairweather-Tait, S.J. Water-loss dehydration and aging. *Mech. Ageing Dev.* **2014**, *136*, 50–58. [CrossRef]
7. Jayaswal, K.; Sahu, V.; Gurjar, B. Water pollution, human health and remediation. In *Water Remediation*; Springer: Singapore, 2018; pp. 11–27.
8. Dickens, C.; McCartney, M. Water-Related Ecosystems. In *Clean Water and Sanitation*; Springer: Cham, Switzerland, 2020; pp. 1–10.
9. Hakimdavar, R.; Hubbard, A.; Policelli, F.; Pickens, A.; Hansen, M.; Fatoyinbo, T.; Lagomasino, D.; Pahlevan, N.; Unninayar, S.; Kavvada, A.; et al. Monitoring water-related ecosystems with earth observation data in support of Sustainable Development Goal (SDG) 6 reporting. *Remote Sens.* **2020**, *12*, 1634. [CrossRef]
10. Tang, W.; Pei, Y.; Zheng, H.; Zhao, Y.; Shu, L.; Zhang, H. Twenty years of China's water pollution control: Experiences and challenges. *Chemosphere* **2022**, *295*, 133875. [CrossRef]
11. Chaudhry, F.N.; Malik, M. Factors affecting water pollution: A review. *J. Ecosyst. Ecography* **2017**, *7*, 1–3.
12. World Health Organization. *A Global Overview of National Regulations and Standards for Drinking-Water Quality*; World Health Organization: Geneva, Switzerland, 2021.
13. Wen, X.; Chen, F.; Lin, Y.; Zhu, H.; Yuan, F.; Kuang, D.; Jia, Z.; Yuan, Z. Microbial indicators and their use for monitoring drinking water quality—A review. *Sustainability* **2020**, *12*, 2249. [CrossRef]
14. Mytton, D. Data centre water consumption. *npj Clean Water* **2021**, *4*, 1–6. [CrossRef]
15. Canter, L.W. *Ground Water Pollution Control*; CRC Press: Boca Raton, FL, USA, 2020.
16. Mishra, B.K.; Kumar, P.; Saraswat, C.; Chakraborty, S.; Gautam, A. Water security in a changing environment: Concept, challenges and solutions. *Water* **2021**, *13*, 490. [CrossRef]
17. Yan, T.; Shen, S.L.; Zhou, A. Indices and models of surface water quality assessment: Review and perspectives. *Environ. Pollut.* **2022**, *308*, 119611. [CrossRef]
18. Park, J.; Kim, K.T.; Lee, W.H. Recent advances in information and communications technology (ICT) and sensor technology for monitoring water quality. *Water* **2020**, *12*, 510. [CrossRef]
19. Liu, P.; Wang, J.; Sangaiah, A.K.; Xie, Y.; Yin, X. Analysis and prediction of water quality using LSTM deep neural networks in IoT environment. *Sustainability* **2019**, *11*, 2058. [CrossRef]
20. Braga, F.H.R.; Dutra, M.L.S.; Lima, N.S.; da Silva, G.M.; de Cássia Mendonça de Miranda, R.; da Cunha Araújo Firmo, W.; de Moura, A.R.L.; de Souza Monteiro, A.; da Silva, L.C.N.; da Silva, D.F.; et al. Study of the Influence of Physicochemical Parameters on the Water Quality Index (WQI) in the Maranhão Amazon, Brazil. *Water* **2022**, *14*, 1546. [CrossRef]
21. Ahmed, U.; Mumtaz, R.; Anwar, H.; Shah, A.A.; Irfan, R.; García-Nieto, J. Efficient water quality prediction using supervised machine learning. *Water* **2019**, *11*, 2210. [CrossRef]
22. Ahmed, A.N.; Othman, F.B.; Afan, H.A.; Ibrahim, R.K.; Fai, C.M.; Hossain, M.S.; Ehteram, M.; Elshafie, A. Machine learning methods for better water quality prediction. *J. Hydrol.* **2019**, *578*, 124084. [CrossRef]
23. Bui, D.T.; Khosravi, K.; Tiefenbacher, J.; Nguyen, H.; Kazakis, N. Improving prediction of water quality indices using novel hybrid machine-learning algorithms. *Sci. Total Environ.* **2020**, *721*, 137612. [CrossRef]
24. Lu, H.; Ma, X. Hybrid decision tree-based machine learning models for short-term water quality prediction. *Chemosphere* **2020**, *249*, 126169. [CrossRef]
25. Imani, M.; Hasan, M.M.; Bittencourt, L.F.; McClymont, K.; Kapelan, Z. A novel machine learning application: Water quality resilience prediction Model. *Sci. Total. Environ.* **2021**, *768*, 144459. [CrossRef]
26. Muharemi, F.; Logofătu, D.; Leon, F. Machine learning approaches for anomaly detection of water quality on a real-world data set. *J. Inf. Telecommun.* **2019**, *3*, 294–307. [CrossRef]
27. Haghiabi, A.H.; Nasrolahi, A.H.; Parsaie, A. Water quality prediction using machine learning methods. *Water Qual. Res. J.* **2018**, *53*, 3–13. [CrossRef]
28. Water Quality. Available online: https://www.kaggle.com/datasets/mssmartypants/water-quality (accessed on 9 December 2022).

29. Kumar, D.; Muthukumar, K. An overview on activation of aluminium-water reaction for enhanced hydrogen production. *J. Alloys Compd.* **2020**, *835*, 155189. [CrossRef]
30. Zhang, L.; Xu, E.G.; Li, Y.; Liu, H.; Vidal-Dorsch, D.E.; Giesy, J.P. Ecological risks posed by ammonia nitrogen (AN) and un-ionized ammonia (NH3) in seven major river systems of China. *Chemosphere* **2018**, *202*, 136–144. [CrossRef]
31. Ahmad, A.; Bhattacharya, P. Arsenic in drinking water: Is 10 μg/L a safe limit? *Curr. Pollut. Rep.* **2019**, *5*, 1–3. [CrossRef]
32. Oskarsson, A. Barium. In *Handbook on the Toxicology of Metals*; Elsevier: Amsterdam, The Netherlands, 2022; pp. 91–100.
33. Mahajan, P.; Kaushal, J. Role of phytoremediation in reducing cadmium toxicity in soil and water. *J. Toxicol.* **2018**, *2018*, 4864365. [CrossRef]
34. Hossain, S.; Chow, C.W.; Cook, D.; Sawade, E.; Hewa, G.A. Review of chloramine decay models in drinking water system. *Environ. Sci. Water Res. Technol.* **2022**, *8*, 926–948. [CrossRef]
35. World Health Organization. *Chromium in Drinking-Water*; Technical Report; World Health Organization: Geneva, Switzerland, 2020.
36. Najafpour, M.M.; Mehrabani, S.; Mousazade, Y.; Hołyńska, M. Water oxidation by a copper (II) complex: New findings, questions, challenges and a new hypothesis. *Dalton Trans.* **2018**, *47*, 9021–9029. [CrossRef]
37. Kabir, H.; Gupta, A.K.; Tripathy, S. Fluoride and human health: Systematic appraisal of sources, exposures, metabolism, and toxicity. *Crit. Rev. Environ. Sci. Technol.* **2020**, *50*, 1116–1193. [CrossRef]
38. Shen, M.; Zeng, Z.; Li, L.; Song, B.; Zhou, C.; Zeng, G.; Zhang, Y.; Xiao, R. Microplastics act as an important protective umbrella for bacteria during water/wastewater disinfection. *J. Clean. Prod.* **2021**, *315*, 128188. [CrossRef]
39. Pilevar, M.; Kim, K.T.; Lee, W.H. Recent advances in biosensors for detecting viruses in water and wastewater. *J. Hazard. Mater.* **2021**, *410*, 124656. [CrossRef]
40. Levallois, P.; Barn, P.; Valcke, M.; Gauvin, D.; Kosatsky, T. Public health consequences of lead in drinking water. *Curr. Environ. Health Rep.* **2018**, *5*, 255–262. [CrossRef]
41. Zhang, M.; Song, G.; Gelardi, D.L.; Huang, L.; Khan, E.; Mašek, O.; Parikh, S.J.; Ok, Y.S. Evaluating biochar and its modifications for the removal of ammonium, nitrate, and phosphate in water. *Water Res.* **2020**, *186*, 116303. [CrossRef]
42. Sato, Y.; Ishihara, M.; Fukuda, K.; Nakamura, S.; Murakami, K.; Fujita, M.; Yokoe, H. Behavior of nitrate-nitrogen and nitrite-nitrogen in drinking water. *Biocontrol Sci.* **2018**, *23*, 139–143. [CrossRef]
43. Kallithrakas-Kontos, N.; Foteinis, S. Recent advances in the analysis of mercury in water-review. *Curr. Anal. Chem.* **2016**, *12*, 22–36. [CrossRef]
44. Lisco, G.; De Tullio, A.; Giagulli, V.A.; De Pergola, G.; Triggiani, V. Interference on iodine uptake and human thyroid function by perchlorate-contaminated water and food. *Nutrients* **2020**, *12*, 1669. [CrossRef]
45. Chałupnik, S.; Wysocka, M.; Chmielewska, I.; Samolej, K. Modern technologies for radium removal from water–Polish mining industry case study. *Water Resour. Ind.* **2020**, *23*, 100125. [CrossRef]
46. Golubkina, N.; Erdenetsogt, E.; Tarmaeva, I.; Brown, O.; Tsegmed, S. Selenium and drinking water quality indicators in Mongolia. *Environ. Sci. Pollut. Res.* **2018**, *25*, 28619–28627. [CrossRef]
47. World Health Organization. *Silver in Drinking Water: Background Document for Development of WHO Guidelines for Drinking-Water Quality*; Technical Report; World Health Organization: Geneva, Switzerland, 2021.
48. Bjørklund, G.; Semenova, Y.; Pivina, L.; Dadar, M.; Rahman, M.; Aaseth, J.; Chirumbolo, S. Uranium in drinking water: A public health threat. *Arch. Toxicol.* **2020**, *94*, 1551–1560. [CrossRef]
49. Maldonado, S.; López, J.; Vairetti, C. An alternative SMOTE oversampling strategy for high-dimensional datasets. *Appl. Soft Comput.* **2019**, *76*, 380–389. [CrossRef]
50. Dritsas, E.; Fazakis, N.; Kocsis, O.; Moustakas, K.; Fakotakis, N. Optimal Team Pairing of Elder Office Employees with Machine Learning on Synthetic Data. In Proceedings of the 2021 12th International Conference on Information, Intelligence, Systems & Applications (IISA), Chania Crete, Greece, 12–14 July 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1–4.
51. Schober, P.; Boer, C.; Schwarte, L.A. Correlation coefficients: Appropriate use and interpretation. *Anesth. Analg.* **2018**, *126*, 1763–1768. [CrossRef]
52. Darst, B.F.; Malecki, K.C.; Engelman, C.D. Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. *BMC Genet.***2018**, *19*, 65. [CrossRef] [PubMed]
53. Tangirala, S. Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm. *Int. J. Adv. Comput. Sci. Appl.* **2020**, *11*, 612–619. [CrossRef]
54. Gnanambal, S.; Thangaraj, M.; Meenatchi, V.; Gayathri, V. Classification algorithms with attribute selection: An evaluation study using WEKA. *Int. J. Adv. Netw. Appl.* **2018**, *9*, 3640–3644.
55. Berrar, D. Bayes' theorem and naive Bayes classifier. In *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*; Elsevier: Amsterdam, The Netherlands, 2018; Volume 403.
56. Sagi, O.; Rokach, L. Ensemble Learning: A survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2018**, *8*, e1249. [CrossRef]
57. González, S.; García, S.; Del Ser, J.; Rokach, L.; Herrera, F. A practical tutorial on Bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities. *Inf. Fusion* **2020**, *64*, 205–237. [CrossRef]
58. Shuaib, M.; Abdulhamid, S.M.; Adebayo, O.S.; Osho, O.; Idris, I.; Alhassan, J.K.; Rana, N. Whale optimization algorithm-based email spam feature selection method using rotation forest algorithm for classification. *SN Appl. Sci.* **2019**, *1*, 1–17. [CrossRef]

59. Parmar, A.; Katariya, R.; Patel, V. A review on random forest: An ensemble classifier. In *Proceedings of the International Conference on Intelligent Data Communication Technologies and Internet of Things*; Springer: Cham, Switzerland, 2018; pp. 758–763.
60. Polat, K.; Sentürk, U. A novel ML approach to prediction of breast cancer: Combining of mad normalization, KMC based feature weighting and AdaBoostM1 classifier. In Proceedings of the 2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), Ankara, Turkey, 19–21 October 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1–4.
61. Kumari, S.; Kumar, D.; Mittal, M. An ensemble approach for classification and prediction of diabetes mellitus using soft Voting classifier. *Int. J. Cogn. Comput. Eng.* **2021**, *2*, 40–46. [CrossRef]
62. Pavlyshenko, B. Using Stacking approaches for machine learning models. In Proceedings of the 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP), Lviv, Ukraine, 21–25 August 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 255–258.
63. Masih, N.; Naz, H.; Ahuja, S. Multilayer perceptron based deep neural network for early detection of coronary heart disease. *Health Technol.* **2021**, *11*, 127–138. [CrossRef]
64. Christodoulou, E.; Ma, J.; Collins, G.S.; Steyerberg, E.W.; Verbakel, J.Y.; Van Calster, B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J. Clin. Epidemiol.* **2019**, *110*, 12–22. [CrossRef]
65. Cunningham, P.; Delany, S.J. k-Nearest neighbour classifiers-A Tutorial. *ACM Comput. Surv. (CSUR)* **2021**, *54*, 1–25. [CrossRef]
66. Waikato Environment for Knowledge Analysis. Available online: https://www.weka.io/ (accessed on 9 December 2022).
67. Hossin, M.; Sulaiman, M.N. A review on evaluation metrics for data classification evaluations. *Int. J. Data Min. Knowl. Manag. Process.* **2015**, *5*, 1.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.