water

MDPI

Article

Classification of WatSan Technologies Using Machine Learning Techniques

Hala Al Nuaimi [1], Mohamed Abdelmagid [1], Ali Bouabid [2], Constantinos V. Chrysikopoulos [3,4,*] and Maher Maalouf [1]

[1] Department of Industrial and Systems Engineering, Khalifa University of Science and Technology, Abu Dhabi 127788, United Arab Emirates; 100050101@ku.ac.ae (H.A.N.); 100059835@ku.ac.ae (M.A.); maher.maalouf@ku.ac.ae (M.M.)
[2] Institute of Educational Sciences, Mohammed VI Polytechnic University, Benguerir 43150, Morocco; ali.bouabid@um6p.ma
[3] Department of Civil Infrastructure and Environmental Engineering, Khalifa University of Science and Technology, Abu Dhabi 127788, United Arab Emirates
[4] School of Chemical and Environmental Engineering, Technical University of Crete, 73100 Chania, Greece
* Correspondence: constantinos.chrysikopoulos@ku.ac.ae

Abstract: A substantial portion of the water supply and sanitation (WatSan) infrastructure in the rural areas of developing countries is currently not operating. This failure is due to the inappropriate implementation of WatSan technologies and the lack of decision-making resources. This study explores the application of several machine learning classification algorithms to predict the optimal WatSan system effectively. The proposed classification methods are Logistic Regression, Random Forest, Support Vector Machine, CatBoost, and Neural Network. The practicality of these classification methods was tested using a dataset comprising 774 water technology options. Several experiments were conducted to obtain the highest possible classification accuracy of the capacity requirement level (CRL) in terms of accuracy and F1 score classification metrics. Our findings suggest that CatBoost, with the addition of the synthetic minority oversampling technique (SMOTE), outperforms the other algorithms in classifying WatSan technology options.

Keywords: classification; decision support system; Logistic Regression; machine learning; Random Forest; Support Vector Machine

Citation: Al Nuaimi, H.; Abdelmagid, M.; Bouabid, A.; Chrysikopoulos, C.V.; Maalouf, M. Classification of WatSan Technologies Using Machine Learning Techniques. *Water* **2023**, *15*, 2829. https://doi.org/10.3390/w15152829

Academic Editors: Fernando António Leal Pacheco and Helvi Heinonen-Tanski

Received: 16 June 2023
Revised: 31 July 2023
Accepted: 1 August 2023
Published: 4 August 2023

1. Introduction

Despite the fact that the United Nations has recognized access to water and sanitation as fundamental human rights, a large number of people around the world lack these basic services. Access to water and sanitation services continues to be a serious challenge in developing countries [1–3]. People living in extreme poverty in developing nations are most negatively affected by the lack of access to water and sanitation facilities [4]. Approximately 2.2 billion people do not have access to clean drinking water globally [5], and 3.6 billion people have no access to safely managed sanitation [6]. Consequently, the health of millions of individuals throughout the world is jeopardized.

Worldwide, 2.4 million deaths could be prevented each year if individuals had access to clean drinking water, adequate sanitation, and proper hygiene [7]. This decline also has a detrimental impact on the environment, society, economy, and business. A lack of service infrastructure and poor management, in conjunction with inappropriate selection and implementation of WatSan technologies, may ultimately lead to failure [8]. It should be noted that 15% to 30% of WatSan infrastructures installed in rural areas in developing countries do not operate after the first 2–3 years [8].

In addition, it has been estimated more than 1.3 billion people use basic water services, 206 million people have limited services, 422 million are obtaining water from exposed

wells, and 159 million are using untreated surface water from lakes [9]. Furthermore, climate change contributes to disturbing weather patterns and thus makes water scarcity worse, in addition to causing droughts and floods that impact the water quality of water resources [10–12]. Population growth inevitably leads to increased water demand [13]. By 2050, the global human population will increase by 32% to reach 9.1 billion. This growth is expected to happen mainly in low-income countries such as Pakistan, Nigeria, Kenya, and Bangladesh [14].

In order to address these problems, along with the problem of poverty, which is inextricably related to inadequate sanitation [15,16], numerous foreign agencies, as well as non-government organizations, are committed to supporting developing countries in building their water and sanitation infrastructure [17], in addition to the support for healthcare [18] and education [19]. However, the selection of appropriate infrastructure has always been a challenge that many international and local development and aid organizations have experienced [20]. This is due to missing elements in the existing decision tools for appropriate technology selection.

This paper aims to provide a classification tool for water and sanitation technologies using machine learning methods. The results of the classification will then be integrated into a decision support system (DSS), which will be used for the selection of the most appropriate WatSan technology option. To the best of our knowledge, such work has not been published previously in the literature.

2. Existing Decision Support Systems

A DSS is a computer-based program that assists users in making decisions and taking actions in an organization or a business [21]. It collects and analyzes data before incorporating them into useful information. A DSS mainly consists of three components, namely, (1) the database, (2) the software system, and (3) the user interface [22]. The database is a collection of tabulated data [23]. The data stored can range from records, information, and files to contacts and scores. The software system is a set of algorithms and models (statistical, optimization, classification, forecasting techniques) that analyze and process the data. User interfaces are the places of interaction between users and designs. Therefore, they are access points with which the user interacts in order to obtain the desired software system output.

There are many decision support systems for water and sanitation technologies available in the market. Some typical examples are:

1. An article entitled "A decision support system for water resources management: The case study of Mubuku irrigation scheme, Uganda" provided insights on developing a decision support system based on the Mapping System and Services for Canal Operation Techniques (MASSCOTE) approach and the MIKE Hydro Basin model. The model intends to improve water service, increase irrigation efficiency, and meet the country's economic goals [24].
2. A document entitled "Tools to apply a gender approach: The Asian experience" was presented by project managers from rural projects in Asia. It brings together the perspectives of fifteen workshop participants from nine Asian nations. The document aims to share different experiences so that sector staff and organizations can help people in underdeveloped nations in obtaining better access to water and sanitation services [25]. It shows the stages that need to be followed and the work that needs to be carried out before choosing or implementing a water or sanitation technology.
3. A general design guideline for building a water DSS has been presented in a document entitled "Decision support system for water distribution management". The document focuses on needs assessment, generic design for DSS development, and field installation for water technologies. It emphasizes the role of data management, data analysis, simulation, and optimization in the development of DSS [26].
4. A guideline was developed by the WHO entitled "Linking technology choice with operation and maintenance in the context of community water supply and sanitation"

to help decision-makers to choose technology for water supply and sanitation that can be maintained long enough in developing countries. For many years, while selecting such technologies, technical criteria and initial investments were prioritized, but the operation and maintenance (OM) effect was simply neglected. In this manual, the OM component is added to the selection process because it considers economic, administrative, and environmental factors as critical factors for sustainability [27].

5. A document entitled "Choosing an appropriate sanitation system" offers a thorough framework for evaluating and selecting acceptable sanitation technologies based on a set of important criteria. The major goal of the document is to ensure that sanitation systems deployed in low-income nations match the region's unique demands and problems. Affordability, acceptability, constructability, usefulness, dependability, durability, maintainability, and upgradability are among the factors mentioned in the document [28].
6. A document prepared by a group of researchers in Africa entitled "Participatory Decision Making for Sanitation Improvements in Unplanned Urban Settlements in East Africa" offers a multicriterion decision analysis methodology called Proact 2.0. The tool allows scientists, professionals, and policymakers to integrate their knowledge, experiences, and preferences with those of end users, as they do not necessarily favor the most optimal sanitation solution when selecting sanitation technology [29].
7. A procedure entitled "Procedure for the Pre-Selection of Sanitation Systems" provides a multicriterion analysis that is based on weighted summing and the notion of sanitation system templates described in the Compendium of Sanitation Systems and Technologies (a database of a diverse spectrum of sanitation technologies). The goal of this procedure is to stimulate conversation about various choices in order to systematically, objectively, and transparently determine feasible sanitation solutions in a common agreement between stakeholders. The procedure also seeks to anticipate how well each solution fulfills relevant features [30].
8. A document entitled "Constructing and selecting optimal sustainable sanitation system based on expanded structured decision-making for global sanitation and resources crisis" offers a great technique for selecting the optimal sustainable sanitation system to improve the environment in Beijing's rural human settlements. The proposed method combines macro-environmental content analysis, compatibility assessment, and multicriterion decision analysis into structured decision making. The method can also be applied to other complicated infrastructure decision-making situations [31].
9. A program that gives a thorough list of potential technologies and system configurations, analyzes their local applicability, and assesses their potential for resource recovery and loss is presented in the paper entitled "Closing Water and Nutrient Cycles in Urban Wastewater Management: How to Make an Academic Software Available to General Practice". The program offers a manageable but varied set of decision possibilities along with the data necessary to rank the alternatives and choose the preferred one in a structured decision-making process [32].
10. A software named "SANTIAGO", which stands for Sanitation System Alternative Generator, was created by Eawag, which is one of the world's leading aquatic research institutes in Switzerland, to aid engineers and improve the transparency of the selection process. The software suggests a wide variety of locally suitable sanitation system solutions while taking into account a wide array of technology and system options [33].
11. A factsheet entitled "Selecting Sustainable Sanitation System" offers an executive overview detailing the important factors to take into account while putting in place a sustainable sanitation system. The sheet affirms that the long-term success of a sanitation system relies on factors such as social acceptance, political support, and suitable financing models. It also highlights the need for holistic and city-wide planning that encompasses the entire area's sanitation needs [34].

12. A report entitled "Sanitation Technology Options" created by the Susana Organization provides a guideline that outlines technical and economic characteristics of the numerous technological options that have shown to be workable for widespread use in the South African environment. The document describes several technical solutions for meeting the requirements for basic sanitation, as well as the operating and maintenance requirements for each of these options. Some of the sustainability needs are also addressed, such as affordability, operation and maintenance, and institutional duties. A basic technological selection guide is also offered; however, each situation should be subject to the local assessment of sustainability and acceptability [35].

As mentioned earlier, there are a lot of decision support systems for water and sanitation technologies. Gleick et al. [36] reviewed 120 decision support tools and illustrated that most of the available decision tools lack the effective user interface, cost, and monetary data; information on funding approaches; information on scalability; community implications and regionally specific matters; and available technologies [36]. Among the 120 reviewed decision support tools, 18 of them represent the best water supply and sanitation decision-making systems.

It was suggested that the most essential features for a comprehensive decision-making tool were (1) sector, (2) locale, (3) topics, and (4) user. The feature sector indicates the type of support service, which can be water supply, sanitation, or waste treatment services. The locale feature specifies the support service for the locals, based on the region and the type of community, whether it was rural, peri-urban, or urban. The feature topics include information on operation and maintenance, community engagement, service establishment, price, expandability and replicability, and case studies. Finally, the user feature refers to the support resource that allows one to input data related to a specific community in order to provide a suggestion on appropriate service to be implemented [36].

3. Proposed Decision Support System

To fill the missing elements identified by Gleick et al. [36] in their review of existing decision tools for water and sanitation technology selection and to ensure service sustainability, the use of a decision model with a framework that borrows the same pedagogy as the risk analysis, namely assessment, evaluation, and management, was proposed.

The decision support system, called WatSanE, was proposed by Bouabid [37] to address the problem of selecting appropriate water and sanitation technology for a developing country. It consists of three main modules: Module #1 (assessment module) provides decision tools for the selection of appropriate WatSan alternatives. Module #2 (evaluation module) is used for the evaluation and ranking of the set of WatSan alternatives selected in decision Module #1. Finally, Module #3 (management module) provides guidance to implement the chosen WatSan technology and its integration within municipal sanitation services. The proposed decision model uses a systems approach in addressing the problem of access to water supply and sanitation services. Indeed, it assists communities in the selection of WatSan technologies by examining not only the specifics of the problem under consideration but also investigating the relevant factors in the surrounding environment where WatSan technologies will be operating.

This research focuses mainly on the decision portion of Module #1, which is the DSS. Note that Module #1 includes a database of WatSan technologies, with proven sustainability in developing communities, classified by their Capacity Requirement Level (CRL) metric. The CRL metric defines the capacity level a community must have to operate and sustainably maintain a WatSan technology. This classification uses a four-level scoring scale (very low, low, moderate, high).

There is a connection between Component 1 of Module #1, which was the focus of an earlier study conducted by Bouabid and Louis [8], and Component 2 of Module #1, which is the focus of this research. Indeed, the results of the community capacity assessment (Capacity Factor Analysis—Module #1) are used as inputs in the DSS for the selection of appropriate technologies. The technology options in Component 2 will be classified using

machine learning algorithms. Furthermore, the classified technology options, which are the result of this research, will be integrated into the WatSanE DSS database.

4. Methodology

To be able to classify WatSan technology options using machine learning, different steps, including dataset selection, data procession, classifier selection, hyperparameter tuning, application of oversampling technique, and model evaluation, were conducted before obtaining the results. Figure 1 illustrates the methodology followed in this study.

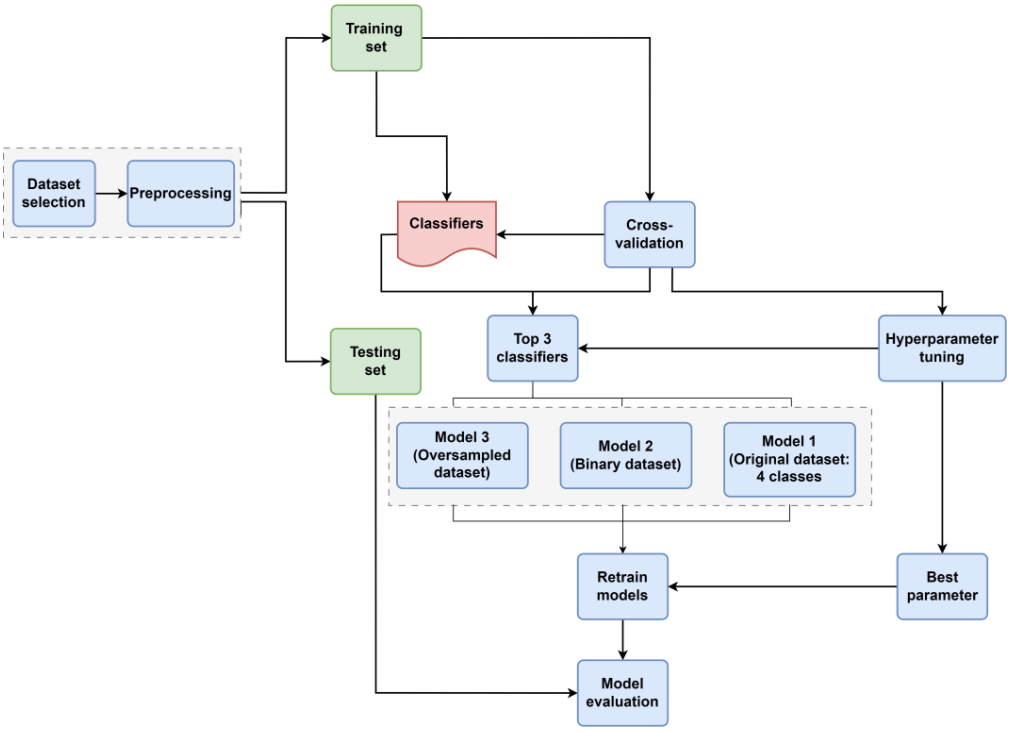

Figure 1. Flow chart of the research methodology employed in this study.

4.1. Dataset Selection

The dataset used in this study was obtained from the work of Bouabid and Louis [8], who developed the dataset in collaboration with WatSan partners and international organizations, including Engineers Without Borders. The data are publicly available and consist of 774 technologies [38]. Each technology is divided into five components, acting as the input variables. The input variables consist of the source (10 levels of categorical variables), device (12 levels of categorical variables), treatment (6 levels of categorical variables), storage (6 levels of categorical variables), and distribution (2 levels of categorical variables). The output parameter is the capacity requirement level (CRL), which has four classes (1 (very low), 2 (low), 3 (moderate) and 4 (high)), that represents the level of each technology option. The dataset is imbalanced with two majority and minority classes. Around 95% of the dataset consisted of technology options with a CRL value of 2 or 3. Class 1 consists of 16 technologies, Class 2 consists of 194 technologies, Class 3 includes 541 technologies, and Class 4 contains 23 technologies.

4.2. Data Preprocessing

The categorical dataset was transformed into a numerical format to facilitate proper integration with machine learning models. One-hot encoding, a widely employed technique, was utilized to convert categorical data into binary vector representations, wherein each category is denoted by a binary vector containing a 1 in the position corresponding to the specific category and 0 elsewhere. Following the transformation of categorical data, the

dataset was partitioned into a training set (80%) and a testing set (20%). The training set was employed to develop the model, while the testing set was used to assess the performance of the resulting trained model.

4.3. Classifier Selection

As mentioned, the proposed classification method for the WatSan technologies is based on machine learning algorithms. The initial machine learning algorithms chosen to be implemented in this research include Random Forest, Support Vector Machine (SVM), Logistic Regression (LR), CatBoost, and Artificial Neural Network (ANN).

4.3.1. Random Forests (RF)

Random Forest, an ensemble machine learning algorithm, constructs multiple decision trees during the training phase and outputs the mode of the classes of the individual trees. It introduces a layer of randomness in the formation of the decision trees with the objective of creating uncorrelated trees. This randomness helps in reducing the variance without inflating the bias when the predictions from these trees are averaged or combined [39].

This algorithm forms numerous decision trees using varying subsets of the training data and variables. When the algorithm needs to predict, each tree gives its predicted class, and the class with the highest frequency is selected as the final prediction. For a Random Forest with B trees, the predicted class $\hat{c}$ for a given instance x can be defined mathematically as

$$\hat{c}(x) = arg \max_{c \in C} \sum_{b=1}^{B} I(h_b(x) = c) \quad (1)$$

where $h_b(x)$ represents the prediction of the b-th decision tree for the instance x. $I(h_b(x) = c)$ is an indicator function, with output 1 if the b-th tree's prediction equals the c-th class, and 0 otherwise. $arg \max_{c \in C}$ signifies the selection of class c that maximizes the sum, i.e., the class that receives the most "votes" from the decision trees.

4.3.2. Support Vector Machine (SVM)

The Support Vector Machine (SVM) works by determining the best separation line, known as "hyperplane" to precisely isolate two classes or more in a classification problem [40]. The objective is to obtain the ideal hyperplane separation by training the divisible data [41]. The optimal hyperplane is determined by finding the closest points to a line from both classes. These points are known as support vectors. Then, the distance between the line and the support vectors, which is called the margin, is calculated. The line for which the margin is maximized is the optimal hyperplane [42].

Given a set of training data of size n, $\{(x_1, y_1), ..., (x_n, y_n)\}$, where each x_i in $\mathbb{R}^n$ denotes a sample in the input space with a corresponding output $y_i \in \{1, 0\}$, for $i = 1, 2, \ldots, n$, the SVM optimization problem is described mathematically as follows:

$$Minimize: \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{n} \xi_i \quad (2)$$

$$Subject\ to: \ y_i(\langle x_i, \beta_i \rangle + \beta_0) \geq 1 - \xi i, \ \ \xi_i \geq 0,$$

where C is a constant penalizing the error, and ξ_i is a slack variable representing the errors, such that if the instance is misclassified, then $\xi_i > 1$. Figure 2 illustrates the concept of SVM for classification.

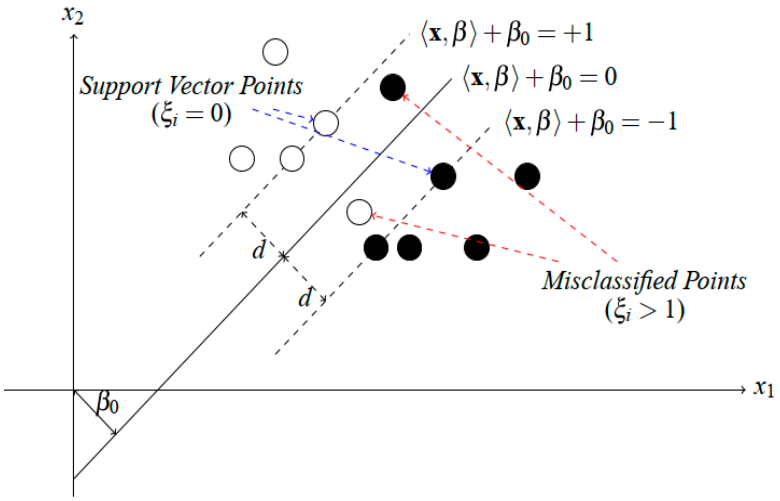

Figure 2. Illustration of SVM for classification.

For nonlinear problems, the SVM algorithm utilizes the kernel function, which takes a low-dimensional input space and transforms it into a higher-dimensional space in order to turn a nonlinear problem into a linear problem using a set of mathematical functions [43,44]. In other words, the kernel function is used to map the original dataset (usually non-separable) into a higher-dimensional space in order to transform it into a separable dataset [45]. Common kernel functions are linear, polynomial, and radial-basis functions.

4.3.3. Logistic Regression (LR)

The Logistic Regression (LR) method is a predictive analysis method that is used for classification problems [46]. It produces a logistic curve, which is limited to values between 0 and 1 [47,48]. The logistic function is defined mathematically as

$$E[y_i = 1|x_i, \beta] = p_i = \frac{1}{1 + e^{x_i\beta}} \tag{3}$$

where y_i is a positive instance, x_i is a row in the data matrix X, β is the coefficient vector, and p_i is the probability of the positive response [49]. The logistic transformation is the log of odds and is expressed mathematically as

$$\eta_i = log\left(\frac{p_i}{1 - p_i}\right) = x_i\beta \tag{4}$$

4.3.4. Categorical Boosting (CatBoost)

The CatBoost method is a recently developed algorithm that uses gradient boosting on decision trees and is very effective for classification problems, especially when the independent variables are also categorical [44,50]. Researchers have utilized CatBoost successfully for machine learning experiments utilizing Big Data since its release in late 2018 [51]. The CatBoost method belongs to the Gradient Boosted Decision Trees (GBDT) machine learning ensemble approach [52]. It is built on symmetric decision trees as primary learners with fewer parameters, supports class variables, and has good accuracy. It also resolves gradient bias and prediction shift issues, which lessens the probability of overfitting [53].

The operation of CatBoost can be described as follows. Given y_i as the target for the i^{th} instance and $F_{m-1}(x_i)$ as the ensemble model built at the $(m-1) - th$ stage, the algorithm computes the gradient of the loss function L at the point $F_{m-1}(x_i)$ as follows:

$$\frac{\partial L(y_i, F_{m-1}(x_i))}{\partial F_{m-1}(x_i)} \tag{5}$$

Iteratively, CatBoost refines its decision boundary by minimizing the loss function at each step and enhancing its predictive accuracy.

4.3.5. Artificial Neural Networks (ANN)

A Neural Network is made up of layers of units (neurons). The network consists of an input layer, a hidden layer, which can be one or multiple layers, and an output layer. The units are connected with different connection weights. Each unit takes an input, applies an activation function to it, which is often nonlinear, and then transmits the output to the following layer [54]. Examples of activation functions are the Logistic Activation Function (Sigmoid), Hyperbolic Tangent Activation Function, etc. A nonlinear activation function is often used in this kind of classification algorithm, which yields to a nonlinear Artificial Neural Network problem. Figure 3 illustrates the ANN used in this study, which can be expressed as

$$y_i = \varphi \sum_{i=1}^{n} \left(w_{j,i} x_i + \theta_j \right) \quad (6)$$

where θ is external threshold, w is the feature weight, x_i is the input, and y_i is the output, which can be represented mathematically as follows:

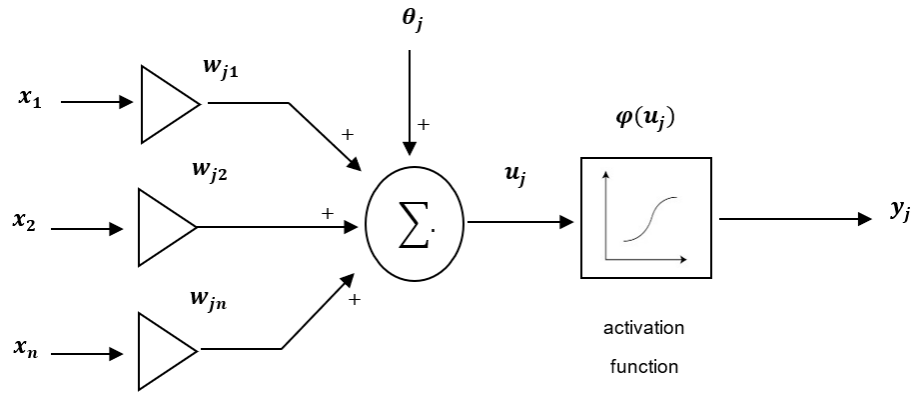

Figure 3. Illustration of Neural Network.

4.4. Multi-Classes Handling

The classification problem under investigation is a multi-class classification problem, where a model is tasked with predicting one of more than two classes. This contrasts with binary classification problems, where a model predicts one of two classes. In this study, five different models were employed:

1. Logistic Regression: This model uses a one-vs.-rest approach with the multinomial logistic loss function. In this approach, a separate model is trained for each class to predict whether an instance belongs to that class or not. The class that obtains the highest probability from its respective model is predicted as the output.
2. Support Vector Machine (SVM): Like Logistic Regression, SVM also uses a one-vs.-rest approach. In this method, one class is chosen as the positive class, and rest of the classes are grouped together as the negative class. A model is trained for each class following this approach, and the class with the highest decision function output is chosen as the output class.
3. Random Forest: Random Forest does not use the one-vs.-rest or one-vs.-one strategy. Instead, it is an ensemble of decision trees that independently vote for the class of an instance. The class with the most votes is chosen as the output.
4. CatBoost: This model automatically handles multi-class classification by using a variant of the one-vs.-all scheme. It does so by setting the loss function parameter to MultiClass.

5. Neural Network: Neural networks employ a different approach altogether. They make use of the softmax activation function in the output layer to provide the probability of each class. The class with the highest probability is chosen as the output class.
6. To evaluate the performance of these models and to obtain a more generalized result, the stratified k-fold cross-validation was used. This method preserves the proportion of each class in every fold, which helps ensure that the cross-validation process used was fair and the results were reliable. The models were evaluated based on their accuracy scores.

4.5. Model Performance Metrics

Upon fine-tuning the hyperparameters for the three selected classifiers (Neural Network, CatBoost, and SVM), each model was trained on the entire training set. Subsequently, their performance was assessed on the independent testing set to evaluate their generalization capabilities. As this research focuses on the classification of WatSan technologies, classification-based evaluation methods are used. The evaluation metrics employed the confusion matrix, accuracy, recall, and F1-score. In addition, the ROC curve and the feature importance metrics were computed for comprehensive analysis.

The confusion matrix, which is also called the contingency table, is a tabular representation that illustrates true positives (TP), or equivalently the number of instances predicted as class A and classified in class A; false negative (FN), or equivalently the number of instances predicted as class A and classified in class B; false positives (FP), or equivalently the number of instances predicted as class B and classified in class A; and true negatives (TN), or equivalently the number of instances predicted as class B and classified in class B. Table 1 shows the confusion matrix where P is the total answers for class A and N the total answers for class B.

Table 1. Confusion matrix for Evaluation of Classifier.

Predicted As	Class A	Class B	Total
Class A (P)	True Positive (TP)	False Negative (FN)	P
Class B (N)	False Positive (FP)	True Negative (TN)	N

Accuracy is the proportion of all correctly predicted instances over the total number of instances [55]. The accuracy can be expressed mathematically as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}, \tag{7}$$

The F1 score is the harmonic mean between precision and recall [56]. The precision is defined as the percentage of predicted positive instances that are actually positive, or TP/(TP + FN) [57]. The recall is defined as the true positive rate, which is the percentage of positive instances that are predicted as positive, or TP/(TP + FN). It should be noted that F1 is typically more useful, especially when the class distribution is imbalanced [58], and it is given by the following equation:

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

These evaluation metrics facilitated a comprehensive understanding of each classifier's performance, allowing for the identification of the most suitable model for the classification of water sanitation technologies.

The receiver operating characteristic (ROC) curve is a graph that displays how well a classification model performs across all categorization levels. Two parameters are plotted by this curve, namely the False Positive Rate and the True Positive Rate [59]. The area under the ROC curve (AUC) evaluates the entire two-dimensional region beneath the complete ROC curve, from (0, 0) to (1, 1) [60].

In order to understand the importance of different features in the predictive model, a technique known as permutation feature importance was employed. Permutation feature importance is defined as the decrease in a model's score when a single feature value is randomly shuffled [39]. This procedure breaks the relationship between the feature and the target; thus, the drop in the model's score is indicative of how much the model depends on the feature. This technique benefits from being model-agnostic and can be calculated many times with different permutations of the feature.

The permutation feature importance of feature i is defined as follows:

$$importance\ (i) = \frac{1}{N} \sum_{n=1}^{N} score(model,\ X,\ y) - score(model,\ X_i^n, y) \tag{9}$$

where X is the original test data, y is the target, X_i^n is the data for the n-th permutation of feature i, the score is the function that returns the score of the model (such as accuracy for classification problems), and N is the number of times the permutation is carried out. The result of this is a list of importance scores for each feature, showing how much each feature contributes to the prediction capability of the model.

In addition to these measures, we incorporate two forms of uncertainty quantification to better understand the predictive behavior of our machine learning models: epistemic and Aleatoric Uncertainty. These uncertainties, inherent in machine learning applications, can provide valuable insights into the confidence and reliability of model predictions, assisting in informed decision making [61–63]:

Epistemic Uncertainty: This type of uncertainty, also known as model uncertainty, arises from the lack of knowledge about the model that best represents the system under study [61–63].

Aleatoric Uncertainty: This represents the uncertainty that is intrinsic to the data themselves, often resulting from the inherent noise or variability in the observations [61–64]. Unlike Epistemic Uncertainty, Aleatoric Uncertainty cannot be reduced with additional data. To measure Aleatoric Uncertainty, we use the concept of entropy as a measure of the unpredictability or randomness of the information being processed [61–63].

5. Results and Discussion

The computational analyses in this study were executed using the Scikit-Learn library in Python on a system equipped with a 12th Generation Intel Core i7-1255U processor running at 2.60 GHz and supported by 16.0 GB of RAM. Three modeling schemes were conducted for comprehensive analysis: (i) Model 1 used the original dataset, where two classes (class 2 and class 3) represented the majority of the dataset (194/774 and 541/774, respectively), and two classes (class 1 and class 4) were the minority of the dataset (16/774 and 23/774, respectively); (ii) Model 2 used a binary classification approach, focusing only on the majority classes (classes 2 and 3); (iii) Model 3 used a balanced dataset, generated by the synthetic minority oversampling technique (SMOTE) [65], which is implemented on the training set (only) to artificially augment the minority class instances.

A range of classifiers was evaluated, and their performance was assessed using 10-fold stratified cross-validation with grid search. The distribution of accuracy scores for each classifier during cross-validation can be seen in Figure 4. The mean cross-validation scores and their standard deviations for all classifiers are presented in Table 2.

From Table 2, it can be seen that the highest performance is achieved by the Artificial Neural Networks (ANN), closely followed by CatBoost, while the Support Vector Machine (SVM) has the lowest CV score. It is worth noting that all these models have CV scores close to each other, ranging from 0.845 (LR) to 0.893 (ANN), indicating that they all have relatively similar performance on the dataset. By looking at the standard deviation, we can see that ANN has the smallest value, suggesting it is the most stable model, i.e., that its performance does not fluctuate much across different splits of the data. Meanwhile, the

CatBoost model, despite achieving a high mean CV score, shows a relatively high standard deviation, indicating varying performance across different splits. Table 2 also indicates that this LR and SVM had an Epistemic Uncertainty of zero, indicating that these models produced identical predictions across different runs. On the other hand, RF, CatBoost, and ANN exhibited higher Epistemic Uncertainty due to the inherent randomness in these models. Interestingly, the model with the highest Epistemic Uncertainty, ANN, also had the lowest Aleatoric Uncertainty, suggesting that it may be more capable of handling the inherent noise in the data. Upon thorough analysis of the model performance metrics, which include both accuracy and the measures of uncertainties, we have identified the top three classifiers: Artificial Neural Networks (ANN), CatBoost, and Support Vector Machine (SVM). These models have been chosen for further refinement through hyperparameter tuning. Subsequently, their performance is evaluated and compared on the test set to present a comprehensive report of their capabilities.

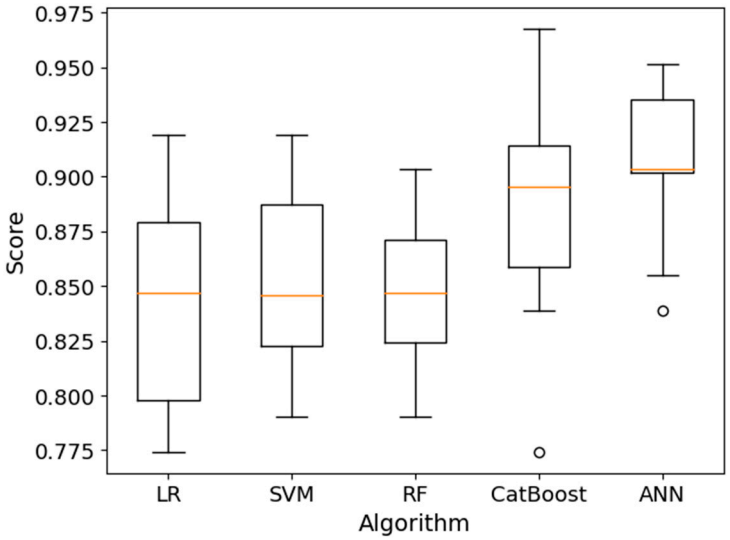

Figure 4. Boxplot of cross-validation accuracy scores for each classifier.

Table 2. Mean cross-validation scores and standard deviations for classifiers.

Classifier	Mean CV Score	Standard Deviation	Epistemic Uncertainty	Aleatoric Uncertainty
LR	0.845	0.050	0.0	0.0
SVM	0.853	0.046	0.0	0.515
RF	0.850	0.036	0.042	0.562
CatBoost	0.887	0.053	0.017	0.344
ANN	0.893	0.032	0.069	0.192

Following this comprehensive evaluation of model performance, we further explored the mechanisms driving the highest-performing model—the Artificial Neural Networks (ANNs)—through a feature importance analysis. This analysis reveals the different influences each variable has on the prediction capability of the artificial neural network model. The "Source" feature shows the highest importance score of approximately 0.14, suggesting that it has the most significant influence on the prediction. On the other hand, the "Distribution" and "Storage" features exhibit lower importance scores of around 0.018 and 0.034, respectively, indicating that these features have a lesser impact on the model's prediction. Despite their lower impact, they still contribute to the overall predictive capability of the model.

5.1. Model 1: Original Dataset

Based on the initial evaluation, the top three classifiers (Neural Network, CatBoost, and SVM) were selected for further optimization through hyperparameter tuning. The tuned hyperparameters of each classifier were then used to train the models on the entire training set, and their performance was evaluated on the unseen test set. Table 3 shows the tuned hyperparameters for each of the top three classifiers and their corresponding cross-validation accuracy. The performance of each of the classifiers is provided in Table 4.

Table 3. Tuned hyperparameters and cross-validation accuracy.

Classifier	Tuned Hyperparameters	CV Accuracy
ANN	Activation: tanh, Sigmoid Batch size: 32 Dropout rate: 0.099 Epochs: 50 Layers: 3 Neurons: (128, 64, 32) Optimizer: Adam	0.893
CatBoost	Depth: 5 Iterations: 150 L2 leaf regularization: 1 Learning rate: 0.1	0.895
SVM	Regularization parameter (C): 10 Kernel coefficient (gamma): 0.1 Kernel type: rbf	0.897

Table 4. Classifier performance metrics on the imbalanced original dataset (Model 1).

Algorithm	Class	Precision	Recall	F1 Score	Accuracy
CatBoost	1	1.00	1.00	1.00	0.93
	2	0.97	0.79	0.87	
	3	0.92	0.99	0.95	
	4	1.00	0.50	0.67	
SVM	1	0.75	1.00	0.86	0.92
	2	0.97	0.79	0.87	
	3	0.92	0.97	0.95	
	4	0.50	0.50	0.50	
ANN	1	0.75	1.00	0.86	0.94
	2	0.97	0.84	0.90	
	3	0.95	0.97	0.96	
	4	0.60	0.75	0.67	

The CatBoost model demonstrated impressive precision across all classes, achieving a perfect score of 1.00 for Class 1 and Class 4, indicating that CatBoost's predictions for these classes are consistently accurate. However, the recall score for Class 2 and Class 4 was lower, suggesting that the model did not perfectly identify all actual instances of these classes. Notwithstanding, the overall accuracy was high at 0.93, implying that CatBoost correctly classified 93% of the instances in the test dataset.

The SVM model displayed robust performance with high precision in all classes and perfect recall for Class 1. Like CatBoost, its recall for Class 2 and Class 4 was lower, indicating some difficulty in correctly identifying all actual instances of these classes. The overall accuracy for SVM was 0.92, implying that it accurately predicted the class for 92% of the instances.

The ANN model exhibited slightly lower precision for Class 1 compared to the other two models but showed impressive precision for Class 2 and Class 3. Despite the lower recall for Class 1 and Class 4, it had the highest overall accuracy of 0.941, suggesting that the ANN model accurately predicted the class for 94.1% of the instances, probably due to its superior performance for Class 3, the majority class in the test dataset. These results highlight that each model has its unique strengths and trade-offs.

While CatBoost and SVM excel in making accurate predictions for Class 1 and Class 4, the Neural Network provides a more balanced performance across all classes, resulting in the highest overall accuracy. Figure 5 provides a more detailed view of the performance of each model, presenting the instances of correct and incorrect predictions made by each model, broken down by class.

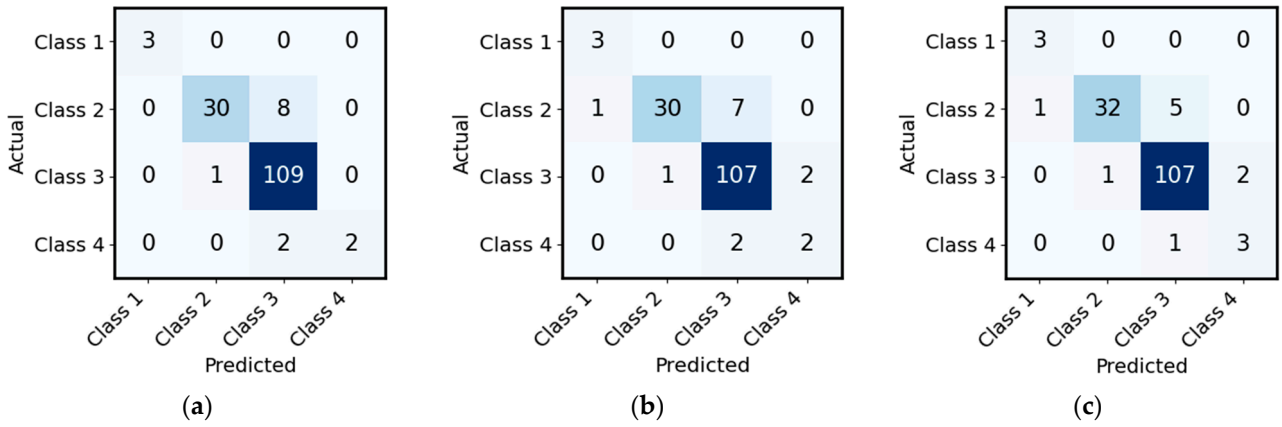

Figure 5. Confusion matrix for the best classifiers: (**a**) CatBoost, (**b**) SVM, and (**c**) ANN.

5.2. Model 2: Binary Dataset

To address the class imbalance issue, a binary model that considers only the majority classes is studied. This enables a more comprehensive understanding of the model performance under different scenarios. The binary model focused on the two majority classes—Class 2 and Class 3. The performance metrics for the binary model are detailed in Table 5.

Table 5. Classifier performance metrics on the imbalanced original dataset (Model 2).

Algorithm	Class	Precision	Recall	F1 Score	Accuracy
CatBoost	2	0.91	0.75	0.82	0.91
	3	0.92	0.99	0.94	
SVM	2	0.97	0.79	0.88	0.93
	3	0.92	0.97	0.88	
ANN	2	0.97	0.84	0.90	0.95
	3	0.95	0.97	0.96	

The CatBoost model displayed a precision of 0.91 for both classes. The recall score was 0.75 for Class 2 and 0.97 for Class 3, indicating that the model identified the majority of actual instances of Class 3 effectively but struggled slightly with Class 2. The overall accuracy was 0.91, demonstrating that CatBoost correctly classified 91% of the instances.

The SVM model showed a balanced performance, with both precision and recall being 0.88 for Class 2 and 0.95 for Class 3. This indicates robustness in correctly identifying and predicting instances of both classes. The overall accuracy for SVM was 0.93, higher than for CatBoost, indicating accurate predictions in 93% of the instances.

The ANN model exhibited a precision of 0.92 for Class 2 and 0.95 for Class 3, demonstrating impressive precision for both classes. The recall scores were also high for Class

2 and 0.97 for Class 3. The overall accuracy was the highest among the models at 0.95, demonstrating that the ANN model correctly predicted the class for 95% of the instances.

The performance of each model is presented in Figure 6, which shows the correct and incorrect predictions made by each model, classified by each class.

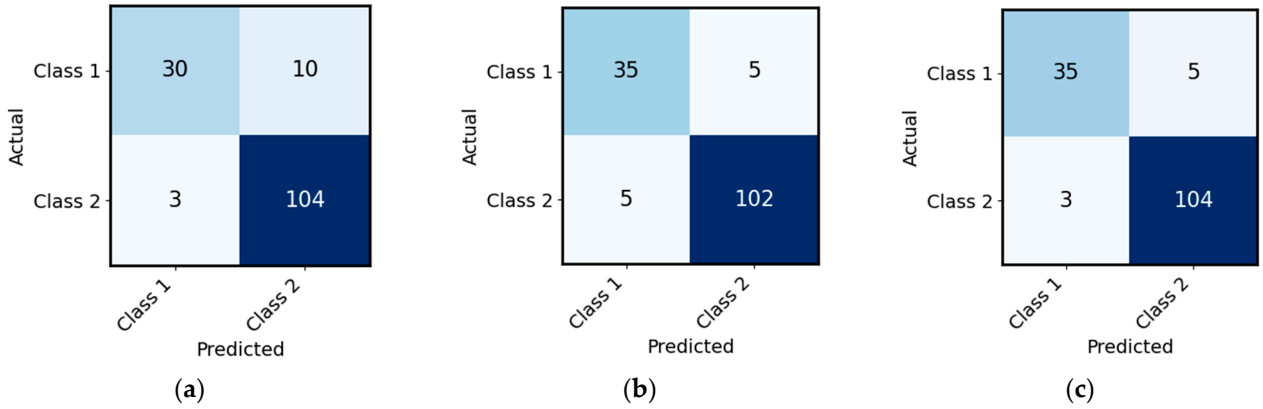

Figure 6. Confusion matrix for the best classifiers—binary models: (**a**) CatBoost, (**b**) SVM, and (**c**) ANN.

Comparing these results with the original model in experiment 1, which incorporated all four classes, it is evident that the binary model has improved the performance. This is likely due to the reduced complexity of distinguishing between two classes rather than four. While the CatBoost and SVM models continue to perform well, the Neural Network model stands out with the highest overall accuracy in the binary model scenario. Despite the trade-offs in terms of class coverage, the binary model can be a beneficial approach when the focus is on the majority classes.

The Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) are significant indicators of the performance of our binary classifier. As shown in Figure 7, the Neural Network model achieved an outstanding AUC score of 0.98. This high score signifies that the model correctly distinguishes the positive and negative classes 98% of the time, indicating strong robustness against overfitting and good generalization on unseen data. This evidence strongly supports the effectiveness of our Neural Network model for this binary classification task.

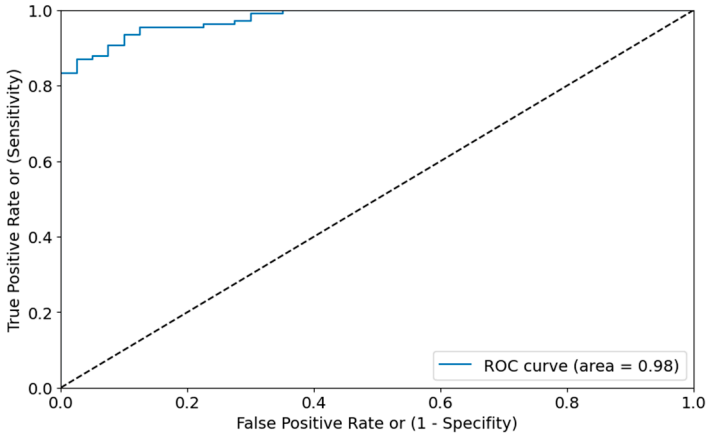

Figure 7. ROC AUC of ANN for Model 2.

5.3. Model 3: Dataset with SMOTE

In pursuit of addressing the class imbalance issue inherent in our data, the SMOTE method was used. The performances of the classifiers with SMOTE are presented in Table 6.

Table 6. Classifier performance metrics with SMOTE (Model 3).

Algorithm	Class	Precision	Recall	F1 Score	Accuracy
CatBoost	1	1.00	1.00	1.00	0.95
	2	0.97	0.84	0.90	
	3	0.94	0.99	0.96	
	4	1.00	0.75	0.86	
SVM	1	0.75	1.00	0.86	0.90
	2	0.96	0.71	0.82	
	3	0.89	0.99	0.94	
	4	1.00	0.25	0.40	
ANN	1	0.75	1.00	0.86	0.93
	2	0.97	0.82	0.89	
	3	0.94	0.97	0.96	
	4	0.60	0.75	0.67	

With the SMOTE method, the CatBoost model achieved a perfect score in both precision and recall for Class 1 and Class 4. The model maintained a high performance for Class 2 and Class 3, yielding the highest overall accuracy of 0.95 across all experiments.

A comparative analysis of the oversampled models with the original model and the binary model further strengthens our observation. It can be observed that the SMOTE model outperforms both models in performance metrics, especially in the context of the minority classes such as Class 4. This performance of the oversampled models can be attributed to the successful implementation of oversampling methods, which have effectively amplified the prediction accuracy for minority classes. Consequently, this has contributed to the enhanced overall performance of the model.

The approach used in this study is adaptable to any dataset, making it usable in a wide range of situations and not case-specific. The same methodology may be used to analyze and offer insightful information for other circumstances by enlarging the dataset or using another one. This has particular resonance for developing countries, which often grapple with the selection of water technology due to inadequate procedures, leading to high failure rates, resource wastage, and sustained water scarcity. In a practical application, the trained classifier developed in this study could be employed to determine the most suitable WatSan technologies for unexplored areas, such as a newly developed city or small community. This would require the collection of location-specific data, aligning with the features used in the initial training dataset. Once the new data are collected, they are then fed into our trained classifier to generate data-driven recommendations for WatSan technology deployment. This results in a higher number of failures and ineffective implementation, leading to resource wastage and continued water scarcity [8]. Therefore, by applying the recommended approach to enhance the understanding of water technology selection and implementation, developing countries can greatly benefit from the insights gained. These findings can contribute to improving decision-making processes, optimizing resource allocation, and ultimately addressing the high failure rates observed in several regions.

6. Conclusions

Appropriate WatSan technologies selection is crucial for ensuring the sustainability of water and sanitation services in developing countries. In this study, five machine learning algorithms were employed for the prediction of the WatSan technology through the capacity requirement level (CRL), with artificial neural networks (ANNs), CatBoost, and Support Vector Machine (SVM) demonstrating superior outperforming both Logistic Regression and Random Forest. Specifically, the CatBoost algorithm, when applied to the augmented data using the synthetic oversampling technique (SMOTE), stands out with an overall accuracy of 0.95. Therefore, the results indicate that these tools could be integrated in a decision support system (DSS) to expedite the selection of the most suitable technology

options for community water supply services, tailored to meeting the specific community needs and environmental circumstances.

Author Contributions: Conceptualization, A.B. and M.M.; methodology, M.M.; software, H.A.N. and M.A.; validation, H.A.N. and M.A.; formal analysis, C.V.C., A.B., H.A.N., M.A. and M.M.; investigation, H.A.N., M.A., A.B., C.V.C. and M.M.; resources, A.B. and M.M.; data curation, A.B.; writing—original draft preparation, H.A.N.; writing—review and editing, H.A.N., M.A., C.V.C., A.B. and M.M.; visualization, M.A.; supervision, A.B. and M.M.; project administration, M.M. and C.V.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are openly available in Mendeley at https://doi.org/10.17632/2szmr4tg3z.2.

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

Abbreviations

ANN	Artificial Neural Network
CART	Classification and regression trees
CRL	Capacity Requirement Level
DSS	Decision support system
DWS	Drinking water supply
FN	False negative
FP	False positive
GBDT	Gradient Boosted Decision Trees
LR	Logistic Regression
N	Total answers for class B
MASSCOTE	Mapping System and Services for Canal Operation Techniques
OM	Operation and maintenance
P	Total answers for class A
RBF	Radial Basis Function
RF	Random Forests
SMOTE	Synthetic Minority Over-sampling Technique
SVM	Support Vector Machine
TN	True negative
TP	True positive
WatSan	Water supply and sanitation

References

1. Adugna, D. Challenges of Sanitation in Developing Counties—Evidenced from a Study of Fourteen Towns, Ethiopia. *Heliyon* **2023**, *9*, e12932. [CrossRef] [PubMed]
2. Seetharam, K. Challenges and Opportunities for Sanitation in Developing Countries. *J. Sci. Policy Gov.* **2015**, *7*.
3. Bishoge, O.K. Challenges Facing Sustainable Water Supply, Sanitation and Hygiene Achievement in Urban Areas in Sub-Saharan Africa. *Local Environ.* **2021**, *26*, 1931074. [CrossRef]
4. Angoua, E.L.E.; Dongo, K.; Templeton, M.R.; Zinsstag, J.; Bonfoh, B. Barriers to Access Improved Water and Sanitation in Poor Peri-Urban Settlements of Abidjan, Côte d'Ivoire. *PLoS ONE* **2018**, *13*, e0202928. [CrossRef] [PubMed]
5. Salehi, M. Global Water Shortage and Potable Water Safety; Today's Concern and Tomorrow's Crisis. *Environ. Int.* **2022**, *158*, 106936. [CrossRef]
6. Donacho, D.O.; Tucho, G.T.; Hailu, A.B. Households' Access to Safely Managed Sanitation Facility and Its Determinant Factors in Jimma Town, Ethiopia. *J. Water Sanit. Hyg. Dev.* **2022**, *12*, 217–226. [CrossRef]
7. Bartram, J.; Cairncross, S. Hygiene, Sanitation, and Water: Forgotten Foundations of Health. *PLoS Med.* **2010**, *7*, e1000367. [CrossRef]
8. Bouabid, A.; Louis, G. Decision Support System for Selection of Appropriate Water Supply and Sanitation Technologies in Developing Countries. *J. Water Sanit. Hyg. Dev.* **2021**, *11*, 208–221. [CrossRef]
9. UNICEF; WHO. *Progress on Household Drinking Water, Sanitation and Hygiene 2000–2017: Special Focus on Inequalities*; World Health Organization: Geneva, Switzerland, 2019.

10. Klare, M.T. Climate Change, Water Scarcity, and the Potential for Interstate Conflict in South Asia. *J. Strateg. Secur.* **2020**, *13*, 109–122. [CrossRef]
11. Ishaque, W.; Tanvir, R.; Mukhtar, M. Climate Change and Water Crises in Pakistan: Implications on Water Quality and Health Risks. *J. Environ. Public Health* **2022**, *2022*, 5484561. [CrossRef]
12. Du, P.; Xu, M.; Li, R. Impacts of Climate Change on Water Resources in the Major Countries along the Belt and Road. *PeerJ* **2021**, *9*, 12201. [CrossRef] [PubMed]
13. Boretti, A.; Rosa, L. Reassessing the Projections of the World Water Development Report. *NPJ Clean Water* **2019**, *2*, 15. [CrossRef]
14. Emile, R.; Clammer, J.R.; Jayaswal, P.; Sharma, P. Addressing Water Scarcity in Developing Country Contexts: A Socio-Cultural Approach. *Humanit. Soc. Sci. Commun.* **2022**, *9*, 144. [CrossRef]
15. Van Minh, H.; Hung, N.V. Economic Aspects of Sanitation in Developing Countries. *Environ. Health Insights* **2011**, *5*, EHI-S8199. [CrossRef]
16. Khalil, H.; Santana, R.; de Oliveira, D.; Palma, F.; Lustosa, R.; Eyre, M.T.; Carvalho-Pereira, T.; Reis, M.G.; Koid, A.I.; Diggle, P.J.; et al. Poverty, Sanitation, and Leptospira Transmission Pathways in Residents from Four Brazilian Slums. *PLoS Negl. Trop. Dis.* **2021**, *15*, e0009256. [CrossRef]
17. Annamraju, S.; Calaguas, B.; Gutierrez, E. *Financing Water and Sanitation—Key Issues in Increasing Resources to the Sector*; OECD: London, UK, 2001; Volume 20.
18. Sanadgol, A.; Doshmangir, L.; Majdzadeh, R.; Gordeev, V.S. Engagement of Non-Governmental Organisations in Moving towards Universal Health Coverage: A Scoping Review. *Glob. Health* **2021**, *17*, 129. [CrossRef] [PubMed]
19. Brophy, M. The Role of NGOs in Supporting Education in Africa. *J. Int. Comp. Educ.* **2020**, *9*, 45–56. [CrossRef]
20. Hansen, S.; Too, E.; Le, T. Criteria to Consider in Selecting and Prioritizing Infrastructure Projects. In Proceedings of the MATEC Web of Conferences, Bandung, Indonesia, 27–29 November 2018; EDP Sciences: Castanet-Tolosan, France, 2019; Volume 270, p. 06004.
21. Silver, M.S. Decisional Guidance for Computer-Based Decision Support. *MIS Q. Manag. Inf. Syst.* **1991**, *15*, 105–122. [CrossRef]
22. Farshidi, S.; Jansen, S.; de Jong, R.; Brinkkemper, S. A Decision Support System for Software Technology Selection. *J. Decis. Syst.* **2018**, *27*, 98–110. [CrossRef]
23. Broatch, J.E.; Dietrich, S.; Goelman, D. Introducing Data Science Techniques by Connecting Database Concepts and Dplyr. *J. Stat. Educ.* **2019**, *27*, 147–153. [CrossRef]
24. Bettili, L.; Pek, E.; Salman, M. A Decision Support System for Water Resources Management: The Case Study of Mubuku Irrigation Scheme, Uganda. *Sustainability* **2019**, *11*, 6260. [CrossRef]
25. Bolt, E. *Together for Water and Sanitation: Tools to Apply a Gender Approach*; the Asian Experience; IRC International Water and Sanitation Centre: The Hague, The Netherlands, 1994.
26. Rey, J. *Decision Support System (DSS) for Water Distribution Management: Theory and Practice*; IWMI: Colombo, Sri Lanka, 1994; Volume 31.
27. Brikké, F.; Bredero, M. *Linking Technology Choice with Operation and Maintenance in the Context of Community Water Supply and Sanitation*; World Health Organization (WHO): Geneva, Switzerland, 2003; ISBN 92 4 156215 3 (NLM).
28. Louw, A.; Holiday, J. *Choosing an Appropriate Sanitation System*; 1992; pp. 235–238. Available online: https://www.ircwash.org/resources/choosing-appropriate-sanitation-system (accessed on 13 February 2022).
29. Hendriksen, A.; Tukahirwa, J.; Oosterveer, P.J.M.; Mol, A.P.J. Participatory Decision Making for Sanitation Improvements in Unplanned Urban Settlements in East Africa. *J. Environ. Dev.* **2012**, *21*, 98–119. [CrossRef]
30. *EAWAG Procedure for the Pre-Selection of Sanitation Systems*; Swiss Federal Institute of Aquatic Science and Technology (Eawag): Düberndorf, Switzerland, 2011; pp. 1–7.
31. Hu, M.; Xiao, J.; Fan, B.; Sun, W.; Zhu, S. Constructing and Selecting Optimal Sustainable Sanitation System Based on Expanded Structured Decision-Making for Global Sanitation and Resources Crisis. *J. Clean. Prod.* **2021**, *318*, 128598. [CrossRef]
32. Schuur, J.S.; Spuhler, D. Closing Water and Nutrient Cycles in Urban Wastewater Management: How to Make an Academic Software Available to General Practice. *Circ. Econ. Sustain.* **2021**, *1*, 1087–1105. [CrossRef] [PubMed]
33. Nisaa, A.F.; Krauss, M.; Spuhler, D. Adapting Santiago Method to Determine Appropriate and Resource Efficient Sanitation Systems for an Urban Settlement in Lima Peru. *Water* **2021**, *13*, 1197. [CrossRef]
34. Dobschütz, S.; Wafler, M. Selecting Sustainable Sanitation Systems. Available online: https://sswm.info/sanitation-project-implementation/sanitation-solutions/selecting-sustainable-sanitation-systems (accessed on 13 February 2022).
35. Sustainable Sanitation Alliance. Available online: https://www.susana.org/en/working-groups/sanitation-systems-technology-options# (accessed on 13 February 2022).
36. Palaniappan, M.; Gleick, P.H.; Change, E. *A Review of Decision-Making Support Tools in the Water, Sanitation, and Hygiene Sector*; Pacific Institute: Oakland, CA, USA, 2008.
37. Boubaid, A. *A Systems Approach for the Selection of Appropriate Water Supply and Sanitation Infrastructure in Developing Communities*; University of Virginia: Charlottesville, WV, USA, 2013.
38. Bouabid, A.; Louis, G. Drinking Water Supply Technologies, Mendeley Data, V2; 2020. Available online: https://doi.org/10.17632/2szmr4tg3z.2 (accessed on 13 February 2022).
39. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]

40. Han, J.; Park, S.; Kim, S.; Son, S.; Lee, S.; Kim, J. Performance of Logistic Regression and Support Vector Machines for Seismic Vulnerability Assessment and Mapping: A Case Study of the 12 September 2016 ML5.8 Gyeongju Earthquake, South Korea. *Sustainability* **2019**, *11*, 7038. [CrossRef]
41. Ribeiro, A.A.; Sachine, M. On the Optimal Separating Hyperplane for Arbitrary Sets: A Generalization of the SVM Formulation and a Convex Hull Approach. *Optimization* **2022**, *71*, 1830089. [CrossRef]
42. Parikh, K.S.; Shah, T.P. Support Vector Machine—A Large Margin Classifier to Diagnose Skin Illnesses. *Procedia Technol.* **2016**, *23*, 369–375. [CrossRef]
43. Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
44. Boser, B.E.; Guyon, I.M.; Vapnik, V.N. A Training Algorithm for Optimal Margin Classifiers. In Proceedings of the Fifth Annual Workshop on Computational Learning Theory, Pittsburgh, PA, USA, 27–29 July 1992; ACM: New York, NY, USA, 1992; pp. 144–152.
45. Balcan, M.-F.; Blum, A.; Vempala, S. On Kernels, Margins, and Low-Dimensional Mappings. In Proceedings of the International Conference on Algorithmic Learning Theory, Padova, Italy, 2–5 October 2004; pp. 194–205.
46. Liao, J.G.; Chin, K.-V. Logistic Regression for Disease Classification Using Microarray Data: Model Selection in a Large p and Small n Case. *Bioinformatics* **2007**, *23*, 1945–1951. [CrossRef] [PubMed]
47. Bewick, V.; Cheek, L.; Ball, J. Statistics Review 14: Logistic Regression. *Crit. Care* **2005**, *9*, 112–118. [CrossRef]
48. Park, H.A. An Introduction to Logistic Regression: From Basic Concepts to Interpretation with Particular Attention to Nursing Domain. *J. Korean Acad. Nurs.* **2013**, *43*, 154–164. [CrossRef]
49. Pal, A. Logistic Regression: A Simple Primer. *Cancer Res. Stat. Treat.* **2021**, *4*, 551–554. [CrossRef]
50. Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A.V.; Gulin, A. CatBoost: Unbiased Boosting with Categorical Features. *Adv. Neural Inf. Process. Syst.* **2018**, *31*.
51. Hancock, J.T.; Khoshgoftaar, T.M. CatBoost for Big Data: An Interdisciplinary Review. *J. Big Data* **2020**, *7*, 94. [CrossRef] [PubMed]
52. Alshari, H.; Saleh, A.Y.; Odabaş, A. Comparison of Gradient Boosting Decision Tree Algorithms for CPU Performance. *J. Inst. Sci. Technol.* **2021**, *37*, 157–168.
53. Chang, W.; Wang, X.; Yang, J.; Qin, T. An Improved CatBoost-Based Classification Model for Ecological Suitability of Blueberries. *Sensors* **2023**, *23*, 1811. [CrossRef]
54. Sharma, S.; Sharma, S.; Athaiya, A. Activation Functions in Neural Networks. *Int. J. Eng. Appl. Sci. Technol.* **2020**, *04*, 310–316. [CrossRef]
55. Vanacore, A.; Pellegrino, M.S.; Ciardiello, A. Fair Evaluation of Classifier Predictive Performance Based on Binary Confusion Matrix. *Comput. Stat.* **2022**, *2022*, 1–21. [CrossRef]
56. Hand, D.J.; Christen, P.; Kirielle, N. F*: An Interpretable Transformation of the F-Measure. *Mach. Learn.* **2021**, *110*, 451–456. [CrossRef] [PubMed]
57. Bekkar, M.; Djemaa, H.K.; Alitouche, T.A. Evaluation Measures for Models Assessment over Imbalanced Data Sets. *J. Inf. Eng. Appl.* **2013**, *3*, 27–38.
58. Kamalov, F.; Thabtah, F.; Leung, H.H. Feature Selection in Imbalanced Data. *Ann. Data Sci.* **2022**, *2022*, 1–15. [CrossRef]
59. Nahm, F.S. Receiver Operating Characteristic Curve: Overview and Practical Use for Clinicians. *Korean J. Anesthesiol.* **2022**, *75*, 25–36. [CrossRef]
60. Marzban, C. The ROC Curve and the Area under It as Performance Measures. *Weather Forecast.* **2004**, *19*, 1106–1114. [CrossRef]
61. Soize, C. *Uncertainty Quantification*; Springer: Berlin/Heidelberg, Germany, 2017.
62. Sullivan, T.J. *Introduction to Uncertainty Quantification*; Springer: Berlin/Heidelberg, Germany, 2015; Volume 63.
63. Der Kiureghian, A.; Ditlevsen, O. Aleatory or Epistemic? Does It Matter? *Struct. Saf.* **2009**, *31*, 105–112. [CrossRef]
64. Nguyen, V.L.; Shaker, M.H.; Hüllermeier, E. How to Measure Uncertainty in Uncertainty Sampling for Active Learning. *Mach. Learn.* **2022**, *111*, 89–122. [CrossRef]
65. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority over-Sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]