9-13-2022

Evaluation of Machine Learning Algorithm on Drinking Water Quality for Better Sustainability

Sanaa Kaddoura
Zayed University, sanaa.kaddoura@zu.ac.ae

sustainability

MDPI

Article

Evaluation of Machine Learning Algorithm on Drinking Water Quality for Better Sustainability

Sanaa Kaddoura

Computing and Applied Technology, College of Technological Innovation, Zayed University, Abu Dhabi P.O. Box 144534, United Arab Emirates; sanaa.kaddoura@zu.ac.ae

Abstract: Water has become intricately linked to the United Nations' sixteen sustainable development goals. Access to clean drinking water is crucial for health, a fundamental human right, and a component of successful health protection policies. Clean water is a significant health and development issue on a national, regional, and local level. Investments in water supply and sanitation have been shown to produce a net economic advantage in some areas because they reduce adverse health effects and medical expenses more than they cost to implement. However, numerous pollutants are affecting the quality of drinking water. This study evaluates the efficiency of using machine learning (ML) techniques in order to predict the quality of water. Thus, in this paper, a machine learning classifier model is built to predict the quality of water using a real dataset. First, significant features are selected. In the case of the used dataset, all measured characteristics are chosen. Data are split into training and testing subsets. A set of existing ML algorithms is applied, and the results are compared in terms of precision, recall, F1 score, and ROC curve. The results show that support vector machine and k-nearest neighbor are better according to F1-score and ROC AUC values. However, The LASSO LARS and stochastic gradient descent are better based on recall values.

Keywords: water quality; machine learning; sustainability; supervised machine learning; drinking water

Citation: Kaddoura, S. Evaluation of Machine Learning Algorithm on Drinking Water Quality for Better Sustainability. *Sustainability* **2022**, *14*, 11478. https://doi.org/10.3390/su141811478

Academic Editor: Fernando António Leal Pacheco

Received: 15 August 2022
Accepted: 8 September 2022
Published: 13 September 2022

1. Introduction

Water consists of more than two-thirds of the Earth's surface; it is a crucial resource for living. However, even with its abundance, the usable amount of water is limited [1]. water has an important position on the Earth. It is so typical in our daily lives that we usually ignore its essence. The water character is dynamic. It can change in both physical and chemical characteristics with time.

Water is a significant component of all living beings. All living cells are constructed of aquatic solutions, suspensions, and emulsions with water in the range of 25–85% [2]. Youthful tissues hold more water than the ancient ones. In general, plant life also depends on water. They take water and other mineral salts from soils. Plants comprise 50–75% water, while in humans, water constitutes 60–65% in males and 50–60% in females [3]. It is discreetly affected in operating systems and contributes to the synthesis and decomposition of several organic compounds. Natural water includes minerals, organic pollutants, and dissolved gases such as air and carbon dioxide [4].

Furthermore, multiple ailments transfer through water. Therefore, real-time tracking of water quality (WQ) becomes an essential need [5]. Water quality may vary due to biological and anthropogenic operations that happen in certain circumstances. Some of these operations are the cause of modifications in water quality due to increasing temperatures, biochemical oxygen consumption, chemical oxygen utilization, and increased nitrogen and phosphorus combinations [6]. Recently, water pollution in the Republic of Kosovo has evolved into a fundamental problem that contains water's chemical, physical and biological elements. Water availability and quality (surface or ground) have been corrupted because

of essential elements such as growing population, automation, urbanization, and others. Surface water quality in current years in Kosovo is corrupt, primarily after release of urban and industrial wastewater and pollution from agriculture [7].

Typically, water can be split into ground and surface water [8]. These types of water are released as pollution risks from agricultural, industrial, and domestic activities and other sources. These sources may contain pollutants such as heavy metals, pesticides, fertilizers, hazardous chemicals, and oils [9].

Water quality can be divided into four main types: potable water, palatable water, contaminated (polluted) water, and infected water [10]. Below are the most typical scientific definitions of these types of water quality [11]:

- Potable water is safe to drink, pleasant to taste, and is used for domestic pursuits.
- Palatable water can contain chemical components that do not generate a threat to human health. It is gorgeous and pleasing [10].
- Contaminated (polluted) water may include unwanted physical, chemical, biological, or radiological substances. It is inappropriate for drinking or domestic use.
- Infected water is polluted with pathogenic organisms.

WQ parameters can be divided into physical, chemical, and biological [12], identified in Table 1.

Table 1. Water quality parameters are categorized into three types: physical, chemical, and biological. Some of these parameters are included.

Category	Water Quality Parameter
Physical	Turbidity, Temperature, Color, Taste and Odor, Solid, Electrical Conductivity
Chemical	PH, Acidity, Alkalinity, Chloride, Nitrogen, Fluoride, Iron and Manganese, Hardness, Dissolved Oxygen, Toxic Organic Substances, Radioactive Substances
Biological	Bacteria, Algae, Viruses

Commonly, considering WQ entails gathering water samples from different sites at particular time intervals and analyzing them in laboratories. However, manual sampling and laboratory study of WQ for any given water body or process can be ineffective, costly, and time-consuming. Currently, intelligent systems are hugely used to observe WQ, especially when real-time data are required [10,13].

Machine learning (ML) is an artificial intelligence (AI) branch that allows applications to predict outcomes more accurately without being explicitly programmed. It is one of the most hopeful applications in information technology, whose application area is endless [14,15]. ML applied in the education area is presently admirable to researchers and scientists. Applications in the education sector include a student performance prediction [16], students testing and grading fairly [17,18], and teachers and staff support [19].

Recently, ML has been widely utilized in finance and marketing applications to solve the challenges in their respective fields. ML and especially decision support systems can enhance community performance by analyzing the ground reality. Due to competitors, costs, tax pressures, and other factors, a turbulent economy becomes typical for every organization. Privatization, globalization, and liberalization pull the organization into a more complex competitive environment. Therefore, to perform desired gain, organizations need appropriate marketing strategies. In addition, the marketing decision support system assists in decreasing organization burdens by studying and planning based on its efficient ML approach [20].

ML is also used to predict the turnover of employees within an organization. The employee turnover rate can highly influence the organization's performance. For this reason, the ability to expect employee turnover has recently been an invaluable tool for any organization seeking to retain employees and predict their future behavior [21].

Because of the growth in Internet use during the past years, cyberattacks have grown, causing an increase in private information and financial loss. Cyberattacks contain phishing, spamming, and malware. Therefore, detecting and avoiding cyberattacks has become essential to increase security [22]. ML is widely used in mail classification. More than 300 billion emails are sent daily. Almost half of them are spam. One of the main jobs of email providers is filtering out spam. Spam detection is confusing. The difference between spam and non-spam messages is unclear, and the criteria vary over time. From different actions to automate spam detection, machine learning has been demonstrated to be a practical and preferred approach by email providers [23]. In addition, ML has been considered an essential technique for treating the enormous growth and complexity of cybersecurity threats. The ML technique system can identify patterns to catch malware and unusual activity better than humans and traditional software [24,25].

ML techniques have been used in innovative city development [26]. They are integrated into many applications in smart cities, such as smart parking [27], power generation [28], healthcare [29], and water and air pollution detection [30]. Accurate forecasting of photovoltaic (PV) generation is critical to producing and planning PV-intensive power systems because of the inherent intermittency of solar power. Several machine learning-based PV forecasting methods have recently emerged [31].

Moreover, ML is applied to air quality detection [32]. Human survival would be impossible without air. Consistent developments in almost every aspect of modern human society have harmed the health of the air. Everyday industrial transportation and domestic activities introduce dangerous pollutants into our environment. Monitoring and forecasting air quality has become critical in this era, particularly in developing countries such as India. In contrast to traditional methods, machine learning-based prediction technologies have proven to be the most effective tools for studying such modern hazards.

Additionally, the healthcare sector has taken advantage of ML advancement to elaborate it in medicine and administrative activities [33,34]. ML technology permits the construction of models that can quickly interpret data and produce results. Thus, doctors could make good decisions on patient diagnosis and treatment options more efficiently. It may lead to enhanced patient health care services [25]. Machine learning assists in making instructed clinical decisions by using past data and the knowledge of evidence-based medicine. ML furnishes techniques to explore and expose complex relationships that are difficult to convert into an equation. Healthcare communication is an essential task in the healthcare system. It is responsible for tactfully translating and sharing information to help and apprise patients and the public. ML is confirmed valid in healthcare with the capacity for complex dialogue control and windy flexibility [35].

Machine learning (ML) has a recently vital role in enhancing the state of play in all professional sports, particularly football. It allows football management to predict the success of the matches via a detailed study of data, ML modeling, and much more. It assists in enhancing their state of play and building strategies that will produce promising [36].

This paper evaluates different supervised machine learning algorithms on a public dataset of drinking water quality [37]. The aim is to identify the highest performance- classifying model on the given dataset and the most reliable results to be used for further applications.

The paper is divided as follows: Section 2 explores some recent ML applications in different sectors. Section 3 elaborates the dataset used in this study with its corresponding features and the ML algorithms applied to the dataset. Afterward, obtained results are discussed in Section 4 and are followed by a reasonable discussion about the performance of the ML algorithms. Finally, Section 5 concludes the work strategy of this paper with a recommendation for future work.

2. Literature Review

ML has recently been used in many areas, including education, healthcare, power systems, security, air quality, and renewable systems [38]. In education, ML has played an essential role, especially during the corona pandemic. ML techniques made the online

examination during the lockdown period possible. In [18], the authors systematically review the ML function in exam management systems during this period. This review was directed by assessing around 135 studies during the last five years. The importance of ML in the whole exam cycle from pre-exam practice, control of examination, and assessment were reviewed and examined. The unsupervised or supervised ML algorithms were determined and classified during each process. This review resumes all the problems and challenges of using machine learning in the examination system. These problems and challenges are discussed with their solutions. The vast development of information technologies and the enormous increase in computer usage in schools have produced innovations in test structure and examination. In [17], the study aimed to automatically detect the most informative subset of test items in evaluating the examinees without decreasing accuracy. For this purpose, the authors proposed a new approach to employing abductive network modeling.

Phishing is an endeavor to rob private information or damage online accounts using misleading emails, messages, or sites that look familiar. Sometimes, a phishing email may contain links that allow for the download of malicious software on users' computers when clicking on it. Such emails should be noticed by spam filters that generally categorize an email with a link as a phishing email. Regardless, emails that do not include links, called link-less emails, need more action from the spam filters. In [39], a real-time anti-phishing system has been proposed. This system uses seven classification algorithms, and natural language processing (NLP)-based features. The system contains the following distinguishing properties from other studies in the literature: language independence, use of the enormous size of phishing and legitimate data, real-time execution, detection of new websites, independence from third-party services, and use of feature-rich classifiers. The authors constructed a new dataset to measure the performance of an ML detection algorithm. From the experiment, the authors conclude that the random forest algorithm with only NLP-based features gives the best accuracy with a 97.98% rate for detecting phishing URLs.

The authors in [23] focused on categorizing link-less emails using the ML approach and deep neural networks. The proposed approach was tested using accurate data. Accordingly, the results indicated that the deep neural network functioned well in detecting phishing emails. In addition, ML has been used in detecting and predicting air quality. In [40], the authors reviewed the studies on air pollution prediction using machine learning algorithms based on sensor data in the context of intelligent cities because of the increase in machine learning techniques and their entry into all fields, especially air pollution forecasting. After a comprehensive review of the most relevant papers regarding using the most famous databases and executing the corresponding filter, the main features were extracted to a link and were compared. As a result, they concluded that they applied advanced and sophisticated techniques instead of simple ML techniques. Moreover, the main prediction target was the particulate matter with a diameter of 2.5 micrometers. However, for efficient air quality prediction, it is important to consider external factors such as weather conditions, spatial characteristics, and temporal features.

In [25], several machine learning algorithms have been reviewed for creating efficient decision support for healthcare applications. This review helped decrease the research gap for creating an efficient decision support system for medical applications. In [33], the authors propose a classification method for Parkinson's disease with the support of human voice signals. In this study, six different machine learning (ML) algorithms are used in the classification steps. The study sought to classify Parkinson's disease based on human voice signals and pull essential elements only in the process to reduce the dataset complexity. Thus, voice signals are examined to check the human voice intensity and spectrum of Parkinson's disease patients. Afterward, machine learning classifiers were applied to categorize them based on extracted features. The results indicate that the random forest classifier has the highest accuracy with a value of 97%, followed by the extreme gradient boosting, k-nearest neighbor, and decision tree classifiers with a value of 95%.

The authors of [41] presented numerous flaws in healthcare's current evidence-based approaches. However, they showed how insufficient biased evidence lead to ineffective care. They investigated the potential for data science and artificial intelligence in addressing healthcare ethical concerns. Finally, they offered policy recommendations for ML reform in the healthcare sector, which can aid in developing beneficial systems.

In [29], a hybrid learning-based classifier was applied to an MRI dataset of amiable and abnormal images in the medical sector, in addition to deep learning algorithms. Compared to supervised ML approaches, this study proposed an approach that exceeded the existing approaches in the literature based on their experimental results.

ML is also used to achieve convenient management and treatment in desalination plants [41]. In this study, the authors proposed an optimization system to ensure modular and cost-effective treatment for small industries. The system applies water vaporizing from saline liquid films that have been removed by surface evaporation based on the differences in vapor pressure produced by forced air convection. The optimization determines the optimal operating settings and allows for a comprehensive examination of the effect of diverse operational decisions on operating cost, capital cost, and footprint area.

In [13], a new real-time technique to survey water quality was proposed. It uses electromagnetic sensors and ML techniques. The integrated multi-sensing device is used to measure several water quality parameters, including oxidation reduction potential, carbon dioxide gas, temperature, conductivity, and pH. A Vector Network Analyzer has been used to deliver various parameters such as S11 that operates at the 50 kHz–3 GHz frequency range. Changes in water samples were recorded and analyzed. Afterward, changes in water impurities were detected using ML techniques.

The authors in [42] studied the performance of AI techniques including artificial neural network and support vector machine in predicting water quality components. All sampling data were gathered from an Iranian river. In applying the AI techniques, several types of transfer and kernel functions were tried. As a result, the authors concluded that the examined AI techniques are suitable for predicting water quality components.

The authors in [43] examined an appropriate classification model for water quality based on ML techniques. Their study compared the performance of several classification models and algorithms to determine the significant features in classifying the water quality. All samples were collected from a Malaysian river. Five models with respective algorithms were applied and compared, showing that the lazy model using the K-Star algorithm has the best accuracy with an accuracy value of around 87%.

Two new techniques of ML based on the decision tree were proposed in [44]. These techniques provided more accurate results on water quality prediction in the short-term period. The authors proposed the extreme gradient boosting and random forest techniques that provide an advanced data denoising technique. Water samples were taken from the Tualatin river in the United States. Six water quality indicators were annotated in the proposed models, including temperature, dissolved oxygen, and pH value.

In [45], the authors have compared the water quality prediction performance of ten ML models (seven traditional and three ensemble models) using big data that were collected from Chinese rivers and recorded between 2012 and 2018. Four primary performance metrics were recorded: precision, recall, F1-score, and weighted F1-score to explore the potential key water parameters in the future. The obtained results show that extensive data enhance the ML models' performance in predicting water quality. The dataset features include pH, DO and NH3-N.

In this study, a dataset with different characteristics is studied and examined to illustrate the water quality level. The following section is intended to explain the methodology of this paper.

3. Methodology

In this paper, a model based on ML algorithms is illustrated in classifying water quality. The appropriate methodology for implementing this research approach is briefly described in this section.

3.1. Dataset Characteristics

The data recorded in the water_potability.csv file [37] was studied. The file contains metrics for 3276 records. However, the dataset was cleaned by removing all rows containing empty cells in any input columns. Thus, the number of records in the dataset decreased; however, the available records are still convenient for studying. The water quality metrics that were measured and recorded are as follows:

1. pH value determines the acid–base balance and indicates whether the water is acidic or alkaline. The recommended pH range is between 6.5 to 8.5.
2. Hardness is determined based on the water ability to precipitate soap caused by calcium and magnesium.
3. Solids (total dissolved solids—TDS) indicate whether the water is mineralized. A higher TDS value indicates high mineralization. It is measured in mg/L.
4. Chloramines are formed when ammonia is added to chlorine while treating drinking water. It is measured in mg/L or ppm.
5. Sulfate is available in ambient air, groundwater, plants, and food, primarily used in the chemical industry. It is measured in mg/L.
6. Conductivity measures the ionic process of a solution in transmitting current. It is measured in μS/cm (microsiemens per centimeter).
7. Organic carbon in source waters comes from decaying natural organic matter and synthetic sources. It measures the total amount of carbon in organic compounds in pure water. It is measured in mg/L.
8. Trihalomethane is determined in drinking water through its concentration. It varies based on the level of organic material, the amount of chlorine, and the water temperature. It is measured in Nephelometric Turbidity Unit.
9. Turbidity measures the light-emitting properties of water. The test indicates the quality of waste discharge concerning the colloidal matter. It is measured in ppm.
10. Potability indicates water safety for human consumption. It has a value of 0 or 1.

The first nine features are the inputs to the proposed model in this study, whereas the "potability" feature is the output. Accordingly, for classification purposes, the data are classified into two classes: "Potable" or "Not Potable." Each of these two classes is given a one-digit code: "1" means potable, whereas "0" means not potable.

The statistical analysis of the dataset features are represented in Table 2.

Table 2. The dataset features with their statistical analysis values, including standard deviation (StdDev) and interquartile range (IQR).

Feature	Min	Max	Mean	Median	StdDev	Mode	IQR
pH Value	0.2275	14	7.086	7.0273	1.5733	0.2275	1.9635
Hardness	73.492	317.34	195.57	197.19	32.635	73.492	39.692
Solids	320.94	56489	21917	20934	8642.2	320.94	11560
Chloramines	1.3909	13.127	7.1343	7.1439	1.5848	1.3909	1.9716
Sulfate	129	481.03	333.22	323.23	41.205	129	51.647
Conductivity	201.62	753.34	426.53	423.46	80.713	201.62	115.74
Organic Carbon	2.2	27.007	14.358	14.322	3.325	2.2	4.5611
Trihalomethanes	8.577	124	66.401	66.542	16.077	8.577	21.399
Turbidity	1.45	6.4947	3.9697	3.9682	0.78035	1.45	1.0704

3.2. Data Processing

Processing of data occurs after being collected [46]. Thus, data processing starts with collecting primary data and transforming it into usable information through a step-by-step procedure, as illustrated in Figure 1. The following list comprises the different steps for data processing:

- After collection, the raw data are preprocessed (missing values are managed, normalization is done if required) to produce data with a readable format.
- Significant features are selected. In the case of this dataset, all measured characteristics are selected.
- Data are split into two subsets, one is used as a training data sample, whereas the other is used for testing purpose.
- ML algorithms are applied to the data subsets.
- Results are obtained and compared.

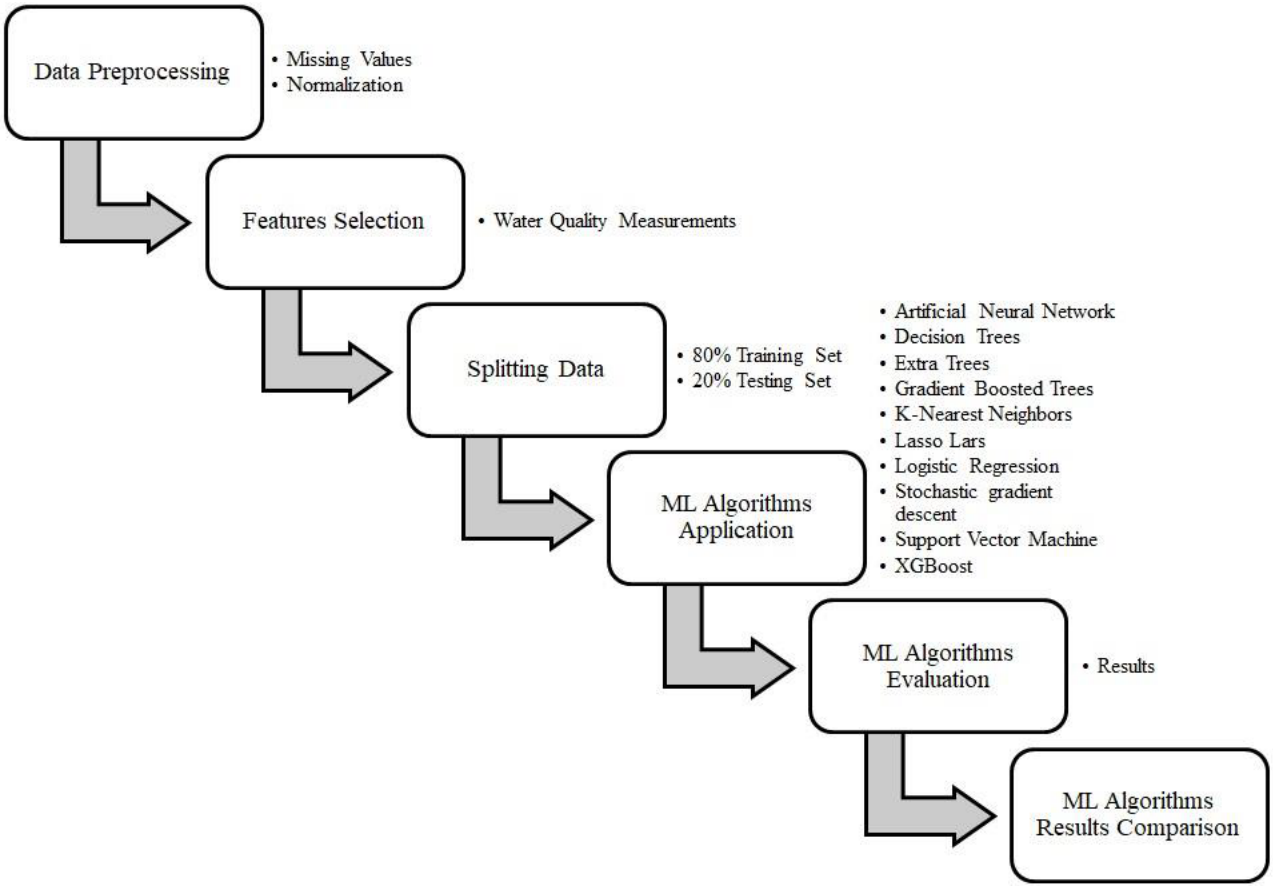

Figure 1. The steps followed in the proposed approach to classify water quality using different machine learning algorithms.

Under supervised machine learning, the dataset is split into two subsets: a training subset consisting of 80% of the dataset's final records and a testing subset consisting of 20% of the data. The use of 80/20 ratio is due to its best overfitting avoidance as explored by many previous research studies [47]. Briefly talking, after data preparation, the total number of obtained records is 2011 records: 1609 records are used for training, and 402 records are used for testing. The training data information is shown in Table 3.

Table 3. Training details.

Data Details	Number of Records
Rows (before preprocessing)	1609
Rows (after preprocessing) 1609 Columns (before preprocessing)	10
Columns (after preprocessing)	9
Matrix type	dense
Estimated memory usage	113.13 kB

In the following part, the supervised ML algorithms applied to the dataset are explained.

3.3. Machine Learning Algorithms

This paper aims to train a labeled dataset for one or more classes. Some learning machine algorithms are used to complete the procedure. This research focuses on applying supervised learning techniques to the desired dataset; data are trained to identify well-known input–output pairs in the training phase. Following that, inferred functions are generated. In the second stage, these functions will map new unlabeled data to their appropriate classes (testing phase) [48]. This paper investigates the following supervised machine learning models:

3.3.1. Random Forest

A random forest (RF) is a supervised ML algorithm built from decision tree algorithms to produce more accurate outcomes for complex problems. A random forest algorithm is composed of decision trees. The precision of the outcome can be increased by increasing the number of decision trees [49].

In this approach model, the random forest algorithm details are shown in Table 4.

Table 4. Random forest algorithm details.

Model Property	Model Information	Model Information
Number of trees		167
Max trees depth		7
Min samples per leaf	5 min samples to split	15
Split quality criterion		Gini
Use bootstrap		Yes
Feature sampling strategy		Auto

3.3.2. Gradient Boosted Trees

Gradient-boosted trees (GBT) are a machine learning technique used to optimize a model's outcomes value via successive steps in the learning process. This algorithm aims to decrease the loss function, representing the difference between predicted and actual values. The gradient represents the cumulative adjustment created in each process step. Boosting is a method of accelerating the progress in predictive accuracy to an adequately optimum value.

In this approach model, the gradient boosted tree algorithm details are shown in Table 5.

Table 5. Gradient boosted tree algorithm details.

Model Property		Model Information
Loss		Deviance
Feature sampling strategy		Default
Number of boosting stages 147 Eta (learning rate)		0.29696608
Max trees depth	8 min samples per leaf	1

3.3.3. Logistic Regression

Logistic regression (LR) is a recognized ML algorithm, belonging to the supervised learning technique. Having a set of independent variables, LR predicts the categorical dependent variable. Thus, the outcome should be discrete (absolute value). For example, the outcome can take the following value: (Yes or No, 0 or 1, True or False). Additionally, the logistic regression algorithm gives a probabilistic value within the range [0, 1].

In this approach model, the logistic regression algorithm details are shown in Table 6.

Table 6. Logistic regression algorithm details.

Model Property		Model Information
Policy		Split the dataset
Sampling method	First records partitions	All partitions
Record limit	100,000 split mode randomly	
Train ratio		0.8
Random seed		1337

3.3.4. XGBoost

The XGBoost (XGB) algorithm is successful in many ML challenges. Logistic regression modeling appeared to be the go-to algorithm to solve multiple predictive modeling use cases. However, as time passed, it has been replaced in literature by XGBoost. The powerful algorithm fibs in its adaptability, performs quick learning through parallel and distributed computing and delivers professional memory usage. XGBoost is an ensemble learning technique because it provides an output as the combination of many models during the final prediction decision [50].

In this approach model, the XGBoost algorithm details are shown in Table 7.

Table 7. XGBoost algorithm details.

Model Property		Model Information
Booster		gbtree
Actual number of trees		11
Total iterations computed		14
Max trees depth		4
Eta (learning rate)	0.28459173 Alpha (L1 regularization)	0.5820198
Alpha (L2 regularization)		0.7686866
Gamma (Min loss reduction to split a leaf)		0.7991586
Min sum of instance weight in a child		3.742847
Subsample ratio of the training instance	0.8793078 fraction of columns in each tree	0.9893092

3.3.5. Decision Tree

A decision tree (DT) is a non-parametric supervised learning algorithm. DT is a hierarchical tree structure containing a root, branches, and leaf nodes. The entropy is used to compute the root variable and, therefore, is oriented toward the values of other attributes.

In this approach model, the decision tree algorithm details are shown in Table 8.

Table 8. Decision tree algorithm details.

Model Property	Model Information
Max tree depth	7
Split criterion	Gini
Min samples per leaf	16
Splitter	best

3.3.6. K Nearest Neighbor

The k-nearest neighbor (KNN) is one of the most uncomplicated ML algorithms. The KNN algorithm assumes the resemblance between new and available data; it places the new one into the category that is most similar to the present ones. KNN is a non-parametric algorithm, meaning that it does not make any assumptions on underlying data. Moreover, it is a lazy learner algorithm because it does not immediately learn from the training set. However, it holds the dataset and executes an action on the dataset during the classification process; therefore, no training period is required. It is effortless to interpret and is fast.

In this approach model, the k-nearest neighbor algorithm details are shown in Table 9.

Table 9. The k-nearest neighbor algorithm details.

Model Property	Model Information
Neighbor finding algorithm	automatic
K	5
Distance weighting	No
Leaf size	30
P	2

3.3.7. Extra Trees

Extra Trees (ExT), or Extremely Randomized Trees, is an ensemble ML algorithm. The ExT algorithm operates by making many trees from the training dataset. Thus, predictions are made after computing the average of all decision trees predictions to produce the final prediction. The ExT algorithm will randomly sample the elements at each split point of a decision tree, similar to the RF algorithm. It is considerably faster than the decision tree and random forest algorithm. It does not consume time, selecting the perfect split point and decreases bias and variance. Therefore, there are rarer chances of the model being overfit or underfit.

In this approach model, the extra trees algorithm details are shown in Table 10.

3.3.8. Artificial Neural Network

An artificial neural network (ANN) simulates the human nerve cell works. An ANN contains more than two interconnected layers: input neurons to transmit data to the next layers and output data that transmit the final output data to the latest output layer. All the internal layers are hidden and adaptively vary the information obtained from one layer to the other one through a series of modifications. Each layer serves as an input and output layer that allows the ANN to comprehend more complex objects. ANNs can learn complex relationships between inputs and outputs. Moreover, after learning from the initial inputs

and their relationships, it can figure out overlooked relationships on unseen data. Further, numerous studies demonstrated that ANNs can better model with high volatility and non-constant variance.

Table 10. Extra trees algorithm details.

Model Property	Model Information	
Number of trees		59
Max trees depth		8
Min samples per leaf	4 Min samples to split	12
Split quality criterion		Gini
Use bootstrap		yes
Feature sampling strategy		auto

In this approach model, the artificial neural network algorithm details are shown in Table 11.

Table 11. Artificial neural network algorithm details.

Model Property	Model Information	
Activation		ReLU
Alpha		0.001
Max iterations		200
Convergence tolerance		0.0001
Early stopping		No
Solver		ADAM
Shuffle data		Yes
Initial learning rate		0.001
Automatic batching		Yes
Batch size	200 beta_1	0.9
beta_2	0.999 epsilon	1×10^{-8}

3.3.9. Stochastic Gradient Descent

Stochastic gradient descent (SGD) is an optimization algorithm used in ML applications to discover model parameters that conform a best predicted–actual output. It is a powerful technique that is widely used in neural network training applications when combined with backpropagation. It needs a single training record processed throughout the network to be suitable for memory purposes. In addition, it operates faster when one record is processed. In the case of larger datasets, it can work faster, as it causes updates to the parameters more often.

In this approach model, the stochastic gradient descent algorithm details are shown in Table 12.

3.3.10. Support Vector Machine

Support vector machine (SVM) is a supervised ML technique used for classification. In the SVM algorithm, records are plotted in a multi-dimensional space (many features representing the weights on many coordinates). After that, the classification process is performed by discovering the hyper-plane to differentiate classes. SVM works reasonably well when there is a precise margin of separation between classes.

Table 12. Stochastic gradient descent algorithm details.

Model Property	Model Information
Loss function	log
Penalty	l1
Stopping tolerance	0.001
Max iterations	1000
Actual iterations	27

In this approach model, the support vector machine algorithm details are shown in Table 13.

Table 13. Support vector machine algorithm details.

Model Property	Model Information
Kernel	rbf
Kernel coef (gamma)	scale
C	1.6051911
Stopping tolerance	0.001
Max iterations	−1

3.3.11. LASSO LARS

LASSO-LARS (LL) is a LASSO model constructed using the LARS algorithm. The minor absolute shrinkage and selection operator, also called LASSO, is a regression calculation method that executes both variable selection and regularization to improve the prediction accuracy and observable of the resulting statistical model.

Least angle regression (LARS) is an ML algorithm used in regression in the case of high dimensional data (a large number of attributes). LARS is relatively identical to forward stepwise regression. At each step, it locates the attribute most favorably related to the target value. There could have more than one attribute that has the same relation. In this case, LARS averages the attributes and continues in a path at the same angle as the attributes.

4. Results and Discussion

After the supervised learning algorithms are applied to the dataset, the classification models' results are computed using performance metrics. This section displays the confusion matrices for the ML algorithms studied in this paper as well as the performance measures to determine the best algorithms for differentiating potable/non-potable water.

The general form for a confusion matrix is represented in Table 14 where "1" means that the water is potable and "0" means that the water is not potable.

Table 14. General structure of the confusion matrix.

	Predicted (1)	**Predicted (0)**
Actually (1)	*tp*	*fp*
Actually (0)	*fn*	*tn*

The four notations represented in the matrix are read as follows:

- *tp* is the number of records that are predicted to be potable and actually are potable.
- *fp* is the number of records that are predicted to be not potable but actually are potable.
- *fn* is the number of records that are predicted to be potable but actually are not potable.
- *tn* is the number of records that are predicted to be not potable and actually are not potable.

After applying the dataset to each of the algorithms represented in Section 3, the confusion matrices are produced and are represented in Tables 15–25.

Table 15. Random forest confusion matrix.

	Predicted (1)	**Predicted (0)**
Actually (1)	155	12
Actually (0)	183	52

Table 16. Gradient boosted trees confusion matrix.

	Predicted (1)	**Predicted (0)**
Actually (1)	129	38
Actually (0)	137	98

Table 17. Logistic regression confusion matrix.

	Predicted (1)	**Predicted (0)**
Actually (1)	165	2
Actually (0)	230	5

Table 18. XGBoost confusion matrix.

	Predicted (1)	**Predicted (0)**
Actually (1)	150	17
Actually (0)	178	57

Table 19. Decision tree confusion matrix.

	Predicted (1)	**Predicted (0)**
Actually (1)	148	19
Actually (0)	180	55

Table 20. The k-nearest neighbor confusion matrix.

	Predicted (1)	**Predicted (0)**
Actually (1)	161	6
Actually (0)	200	35

Table 21. Extra trees confusion matrix.

	Predicted (1)	**Predicted (0)**
Actually (1)	135	32
Actually (0)	131	104

Table 22. Artificial neural network confusion matrix.

	Predicted (1)	**Predicted (0)**
Actually (1)	156	11
Actually (0)	165	70

Table 23. Stochastic gradient descent confusion matrix.

	Predicted (1)	**Predicted (0)**
Actually (1)	167	0
Actually (0)	234	1

Table 24. Support vector machine confusion matrix.

	Predicted (1)	**Predicted (0)**
Actually (1)	147	20
Actually (0)	147	88

Table 25. LASSO LARS confusion matrix.

	Predicted (1)	**Predicted (0)**
Actually (1)	167	0
Actually (0)	235	0

In this proposed approach, machine learning models are illustrated based on whether water quality is potable or not. The performance measurements evaluated in this study are F1-score and ROC AUC. F1-score is calculated in terms of Precision and Recall. The precision of an ML classifier represents the number of samples that are portable out of the total samples the model retrieved. It can be computed using the following expression:

$$Precision = \frac{tp}{tp + f\ p} \tag{1}$$

However, the recall represents the total number of samples that the ML model correctly identified as portable out of the total portable samples. It is calculated using the following formula:

$$Recall = \frac{tp}{tp + fn} \tag{2}$$

The F1-score can be computed based on Recall and precision values. It is a simple representation of the harmonic mean of Precision and Recall.

$$F1 - score = 2 \times \frac{Recall \times Precision}{Recall + Precision} \tag{3}$$

The ROC AUC—a metric that considers the capability of distinguishing classes—is computed in terms of Sensitivity and Recall. Recall is already defined in (2), whereas Sensitivity is computed as follows:

$$Sensitivity = \frac{f\ p}{f\ p + tn} \tag{4}$$

ROC is a plot of correct predictions (Recall) for the positive class versus the fraction of errors (Sensitivity) for the negative class. As a result, Table 26 shows the metric values for Precision, Recall, F1-score, and ROC AUC.

Table 26. Performance metrics for the machine learning algorithms applied to the dataset in this study.

Algorithm	Precision (%)	Recall (%)	F1-Score (%)	ROC AUC
RF	45.9	92.8	61.4	0.702
GBT	48.5	77.2	59.6	0.652
LR	41.8	98.8	58.7	0.498
XGB	45.7	89.8	60.6	0.667
DT	45.1	88.6	59.8	0.654
KNN	44.6	96.4	61.0	0.638
ExT	50.8	80.8	62.4	0.716
ANN	48.6	93.4	63.9	0.726
SGD	41.6	100	58.8	0.449
SVM	50.0	88.0	63.8	0.731
LL	41.5	100	58.7	0.499

In terms of precision, the KNN classifier is the better classifier with a value equal to 50.8%. It represents the ability of the ML classifier model to identify the portable data points. Therefore, among all the used classifiers, the KNN model is better in the identification of portable data points. KNN is adequate when there is a clear margin of separation between classes and when the noise in collected data is very low.

However, based on recall values, the LL and SGD models have a value of 100%. Thus, using these models, there are no records that are predicted to be potable, but in reality, they are not potable. Moreover, many used algorithms have a good performance according to recall values such as RF, KNN, ANN and LR.

F1-score is a critical evaluation metric in machine learning. It elegantly summarizes a model's predictive performance by combining two otherwise opposing metrics—precision and recall. It considers precision and recall, which means it accounts for FPs and FNs. The higher the F1-score, the higher the precision and recall. According to Table 26, SVM and ANN perform better than the other algorithms, with higher F1-score values of 63.8% and 63.9%, respectively. In addition, other models such as ExT, RF and KNN have an acceptable F1-score value. ANN and SVM are useful in the case of a large amount of datasets. In addition, they are efficient in detecting complex relationships of dependent/independent variables.

Similarly, AUC ROC measures a classifier's capability in determining records to their corresponding classes. The greater the AUC, the better the performance in differentiating positive and negative classes. Based on Table 26, SVM and ANN perform better than the other algorithms, with higher F1-score values of 63.8% and 63.9%, respectively. Moreover, they have higher ROC AUC accuracy values of 0.731 and 0.726, respectively. Thus, they are better than other models based on ROC AUC values.

5. Conclusions and Future Work

Water is linked to sixteen of the United Nations Sustainable Development Goals. Access to safe drinking water is essential for good health, a fundamental human right, and a component of effective health-protection policies. Clean water is a critical issue for both health and development. Investments in water supply and sanitation have been shown to produce a net economic advantage in some areas because they reduce adverse health effects and medical costs more than they cost to implement. However, a variety of pollutants are wreaking havoc on drinking water quality. In this paper, the efficiency of using ML algorithms in water pollution problems was studied. As a result, this paper considers many water characteristics for water quality prediction using machine learning algorithms. Environmental problem automation is critical for decision accuracy, long-term planning, and faster action.

This study compared various ML algorithm performances on a dataset of drinking water quality. Afterward, the results were compared to determine the best machine learning algorithm for water quality classification. Thus, using a real dataset, a machine learning classifier model was created in this research to predict the water quality. Significant features were chosen first. The dataset being used included all measurable characteristics. Subsets of data for training and testing were created. Applying a number of currently available ML algorithms, the outcomes were contrasted in terms of precision, recall, F1 score, and ROC curve. According to F1-score and ROC AUC values, the results demonstrate that the support vector machine and k-nearest neighbor are superior.

In future work, the proposed approach will be modified to enhance the performance of these algorithms. Hyperparameter tuning can be performed on each algorithm to find the best model setup to obtain the most optimized result.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available in Kaggle at https://www.kaggle.com/adityakadiwal/water-potability (accessed on 1 June 2022).

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

References

1. Boyd, C.E. *Water Quality: An Introduction*; Springer Nature: Berlin/Heidelberg, Germany, 2009.
2. Kharat, M.; Du, Z.; Zhang, G.; McClements, D.J. Physical and chemical stability of curcumin in aqueous solutions and emulsions: Impact of pH, temperature, and molecular environment. *J. Agric. Food Chem.* **2017**, *65*, 1525–1532. [CrossRef] [PubMed]
3. Aïachi Mezghani, M.; Laaribi, I.; Zouari, I.; Mguidich, A. Sustainability and Plasticity of the Olive Tree Cultivation in Arid Conditions. In *Agriculture Productivity in Tunisia Under Stressed Environment*; Springer: Cham, Switzerland, 2021; pp. 27–56.
4. Brar, A.S. *Consumer Behaviour and Perception for Efficient Water Use in Urban Punjab*; Punjab Technical University: Punjab, India, 2013.
5. O'Flynn, B.; Regan, F.; Lawlor, A.; Wallace, J.; Torres, J.; O'Mathuna, C. Experiences and recommendations in deploying a real-time, water quality monitoring system. *Meas. Sci. Technol.* **2010**, *21*, 124004. [CrossRef]
6. Duda, R.; Klebert, I.; Zdechlik, R. Ground-water pollution risk assessment based on vulnerability to pollution and potential impact of land use forms. *Pol. J. Environ. Stud.* **2020**, *29*, 87–99. [CrossRef]
7. Kaddoura, M.F.; Chosa, M.; Bhalekar, P.; Wright, N.C. Mathematical modeling of a modular convection-enhanced evaporation system. *Desalination* **2021**, *510*, 115057. [CrossRef]
8. Gray, N. *Water Technology*, 3rd ed.; CRC Press: London, UK, 2017.
9. Davis, M.L.; Masten, S.J. *Principles of Environmental Engineering and Science*; McGraw-Hill: New York, NY, USA, 2004.
10. Kedia, N. Water quality monitoring for rural areas—A Sensor Cloud based economical project. In Proceedings of the International Conference on Next Generation Computing Technologies (NGCT), Dehradun, India, 4–5 September 2015; pp. 50–54.
11. Chatterjee, A. *Water Supply Waste Disposal and Environmental Pollution Engineering (Including Odour, Noise and Air Pollution and Its Control)*, 7th ed.; Khanna Publishers: Delhi, India, 2001.
12. Omer, N.H. Water quality parameters. In *Water Quality-Science, Assessments and Policy*; IntechOpen: London, UK, 2019; Volume 18, pp. 1–34.
13. Haraty, R.A.; Kaddoura, S.; Zekri, A. Transaction dependency based approach for database damage assessment using a matrix. *Int. J. Semant. Web Inf. Syst.* **2017**, *13*, 74–86. [CrossRef]
14. Kaddoura, S.; Chandrasekaran, G.; Popescu, D.E.; Duraisamy, J.H. A systematic literature review on spam content detection and classification. *PeerJ Comput. Sci.* **2020**, *8*, e830. [CrossRef] [PubMed]
15. Kaddoura, S.; Arid, A.E.; Moukhtar, M. Evaluation of Supervised Machine Learning Algorithms for Multi-class Intrusion Detection Systems. In *Proceedings of the Future Technologies Conference*; Springer: Cham, Switzerland, 2021; pp. 1–16.
16. Anozie, N.; Junker, B.W. Predicting end-of-year accountability assessment scores from monthly student records in an online tutoring system. In *Educational Data Mining: Papers from the AAAI Workshop*; AAAI Press: Menlo Park, CA, USA, 2006.
17. El-Alfy, E.S.M.; Abdel-Aal, R.E. Construction and analysis of educational tests using abductive machine learning. *Comput. Educ.* **2008**, *51*, 1–16. [CrossRef]
18. Kaddoura, S.; Popescu, D.E.; Hemanth, J.D. A systematic review on machine learning models for online learning and examination systems. *PeerJ Comput. Sci.* **2022**, *8*, e986. [CrossRef]

19. Celar, S.; Stojkic, Z.; Seremet, Z.; Marusic, Z.; Zelenika, D. Classification of Test Documents Based on Handwritten Student ID's characteristics. In *Annals of DAAAM and Proceedings of DAAAM Symposium*; Elsevier: Amsterdam, The Netherlands, 2014; pp. 782–790.
20. Kumar, T. Data mining based marketing decision support system using hybrid machine learning algorithm. *J. Artif. Intell.* **2020**, *2*, 185–193.
21. Alaskar, L.; Crane, M.; Alduailij, M. Employee turnover prediction using machine learning. In *International Conference on Computing*; Springer: Cham, Switzerland, 2019; pp. 301–316.
22. Kaddoura, S.; Haraty, R.A.; Al Kontar, K.; Alfandi, O. A parallelized database damage assessment approach after cyberattack for healthcare systems. *Future Internet* **2021**, *13*, 90. [CrossRef]
23. Kaddoura, S.; Alfandi, O.; Dahmani, N. A spam email detection mechanism for English language text emails using deep learning approach. In Proceedings of the 2020 IEEE 29th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), Bayonne, France, 10–13 September 2020; pp. 193–198.
24. Kaddoura, S. Classification of malicious and benign websites by network features using supervised machine learning algorithms. In Proceedings of the 2021 5th Cyber Security in Networking Conference (CSNet), Abu Dhabi, United Arab Emirates, 12–14 October 2021; pp. 36–40.
25. Shailaja, K.; Seetharamulu, B.; Jabbar, M.A. Machine learning in healthcare: A review. In Proceedings of the 2018 Second international conference on electronics, communication and aerospace technology (ICECA), Coimbatore, India, 29–31 March 2018; pp. 910–914.
26. Mohapatra, B.N.; Panda, P.P. Machine learning applications to smart city. *ACCENTS Trans. Image Process. Comput. Vis.* **2019**, *5*, 1–6. [CrossRef]
27. Simhon, E.; Liao, C.; Starobinski, D. Smart parking pricing: A machine learning approach. In Proceedings of the 2017 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), Atlanta, GA, USA, 1–4 May 2017; pp. 641–646.
28. Akhter, M.N.; Mekhilef, S.; Mokhlis, H.; Mohamed Shah, N. Review on forecasting of photovoltaic power generation based on machine learning and metaheuristic techniques. *IET Renew. Power Gener.* **2019**, *13*, 1009–1023. [CrossRef]
29. Thanki, R.; Kaddoura, S. Dual Learning Model for Multiclass Brain Tumor Classification. In *International Conference on Dependability and Complex Systems*; Springer: Cham, Switzerland, 2022; pp. 350–360.
30. Kang, G.K.; Gao, J.Z.; Chiao, S.; Lu, S.; Xie, G. Air quality prediction: Big data and machine learning approaches. *Int. J. Environ. Sci. Dev.* **2018**, *9*, 8–16. [CrossRef]
31. Su, D.; Batzelis, E.; Pal, B. Machine learning algorithms in forecasting of photovoltaic power generation. In Proceedings of the 2019 International Conference on Smart Energy Systems and Technologies (SEST), Porto, Portugal, 9–11 September 2019; pp. 1–6.
32. Kumar, K.; Pande, B.P. Air pollution prediction with machine learning: A case study of Indian cities. *Int. J. Environ. Sci. Technol.* **2022**. [CrossRef]
33. Ahmed, I.; Aljahdali, S.; Khan, M.S.; Kaddoura, S. Classification of Parkinson disease based on patient's voice signal using machine learning. *Intell. Autom. Soft Comput.* **2022**, *32*, 705–722. [CrossRef]
34. Ganie, S.M.; Malik, M.B.; Arif, T. Machine Learning Techniques for Big Data Analytics in Healthcare: Current Scenario and Future Prospects. In *Telemedicine: The Computer Transformation of Healthcare*; Springer: Cham, Switzerland, 2022; pp. 103–123.
35. Siddique, S.; Chow, J.C. Machine learning in healthcare communication. *Encyclopedia* **2021**, *1*, 220–239. [CrossRef]
36. Herold, M.; Goes, F.; Nopp, S.; Bauer, P.; Thompson, C.; Meyer, T. Machine learning in men's professional football: Current applications and future directions for improving attacking play. *Int. J. Sports Sci. Coach.* **2019**, *14*, 798–817. [CrossRef]
37. Kadiwal, A. Water Quality [Dataset]. Available online: https://www.kaggle.com/adityakadiwal/water-potability (accessed on 10 July 2022).
38. Jhaveri, R.H.; Revathi, A.; Ramana, K.; Raut, R.; Dhanaraj, R.K. A Review on Machine Learning Strategies for Real-World Engineering Applications. *Mob. Inf. Syst.* **2022**, *2022*, 1833507. [CrossRef]
39. Sahingoz, O.K.; Buber, E.; Demir, O.; Diri, B. Machine learning based phishing detection from URLs. *Expert Syst. Appl.* **2019**, *117*, 345–357. [CrossRef]
40. Iskandaryan, D.; Ramos, F.; Trilles, S. Air quality prediction in smart cities using machine learning technologies based on sensor data: A review. *Appl. Sci.* **2020**, *10*, 2401. [CrossRef]
41. Gupta, S.; Sedamkar, R.R. Machine learning for healthcare: Introduction. In *Machine Learning with Health Care Perspective*; Springer: Cham, Switzerland, 2020; pp. 1–25.
42. Haghiabi, A.H.; Nasrolahi, A.H.; Parsaie, A. Water quality prediction using machine learning methods. *Water Qual. Res. J.* **2018**, *53*, 3–13. [CrossRef]
43. Muhammad, S.Y.; Makhtar, M.; Rozaimee, A.; Aziz, A.A.; Jamal, A.A. Classification model for water quality using machine learning techniques. *Int. J. Softw. Eng. Its Appl.* **2015**, *9*, 45–52. [CrossRef]
44. Lu, H.; Ma, X. Hybrid decision tree-based machine learning models for short-term water quality prediction. *Chemosphere* **2020**, *249*, 126169. [CrossRef] [PubMed]
45. Kaddoura, M.F.; Wright, N.C. Optimization of convection-enhanced evaporation (CEE) using generalized cost ratios. *Water Res.* **2022**, *219*, 118491. [CrossRef] [PubMed]
46. Kaddoura, S.; Itani, M.; Roast, C. Analyzing the effect of negation in sentiment polarity of facebook dialectal arabic text. *Appl. Sci.* **2021**, *11*, 4768. [CrossRef]

47. Gholamy, A.; Kreinovich, V.; Kosheleva, O. Why 70/30 or 80/20 Relation between Training and Testing Sets: A Pedagogical Explanation. 2018. Available online: https://www.cs.utep.edu/vladik/2018/tr18-09.pdf (accessed on 10 July 2022).
48. Russell, S.J. *Artificial Intelligence a Modern Approach*; Pearson Education, Inc.: London, UK, 2010.
49. Rizani, S.; Feka, F.; Fetoshi, O.; Durmishi, B.; Shala, S.; Çadraku, H.; Bytyçi, P. Application of water quality index for the assessment the water quality in river Lepenci. *Ecol. Eng. Environ. Technol.* **2022**, *23*, 189–201. [CrossRef]
50. Alshaltone, O.; Nasir, N.; Barneih, F.; Majali, E.A.; Al-Shammaa, A. Multi sensing platform for real time water monitoring using electromagnetic sensor. In Proceedings of the International Conference on Developments in eSystems Engineering (DeSE), Sharjah, United Arab Emirates, 7–10 September 2021; pp. 174–179.