DOI: https://doi.org/10.33103/uot.ijccce.24.3.2

Machine Learning Techniques for Monitoring Water Quality in Water Resources

Husam Hashim Hussein[1], Salam Muhsin Arnoos[2], Wissam Shaya Abbas[3]
[1,2,3] IT Department, Technical College of Management- Baghdad, Baghdad, Iraq
[1]dac0019@mtu.edu.iq, [2]muhsinsalam5@gmail.com, [3]wissamit90@gmail.com

Abstract *—More than 66% of the Earth's total surface area is covered by water. Clean water is one of the basic needs of everyday life. Consistent pollution of water bodies can have far-reaching effects on the lives of living organisms. The World Health Organization has reported that the provision of safe drinking water for human consumption is a challenge that has reached alarming levels. This is because nearly 70% of the total water withdrawals worldwide are used in agriculture. To determine the water's suitability for human consumption, Tests are typically conducted by examining the properties of water in terms of physical, biological, and chemical conditions. There are various methods to measure water quality. Recently, an ongoing process has been shown to improve water quality. To solve this problem, this paper is using a machine learning model to* Contribution for creation *a smart model capable of classifying potable and non-potable water that is released to the water share by the competent authorities and the ability of Machine Learning models to monitor and predict water, especially Managing and Planning Water Resources. datasets were utilized in training and evaluating machine learning models, which includes (3276) samples with nine attributes and two labels indicating water usability According to the in our work the (Random Forest) algorithm have the best Through the results that appeared by accuracy (0.954084, 0. 882484, and 0. 931849 for three sizes of data.) and then (Decision Tree, K-Nearest Neighbor, Logistic Regression, SVC) sequentially*

Index Terms*—Machine Learning, potable water, Water Quality, water resources.*

I. INTRODUCTION

The wastage of water is a matter of great concern, as it has the potential to significantly hinder the availability of potable water and limit access to clean drinking water. This issue directly affects water's hydrodynamics, causing changes in its physical and chemical properties over time. The importance of clean, uncontaminated water cannot be overstated, as its ubiquity in our daily lives is undeniable. It is imperative that researchers act to secure drinking water, which is essential for the planet's survival. Improper management of water resources can have a negative impact on human health as it can cause various diseases. In the water resources sector, especially in agriculture, it is the largest consumer of drinking water, yet many regions lack effective control over the allocation of water resources by specialized authorities.

Water quality is categorized into four classifications: drinking water, supply water, contaminated water, and wastewater. These classifications are based on rigorous scientific criteria developed to evaluate the various aspects of water quality. Potable water, for instance, is defined as water that is safe for consumption, has a pleasant taste, and is suitable for domestic use. While drinking water may contain certain chemicals, these do not pose a significant risk to human health. To ensure sufficient access to clean drinking water, it becomes imperative to regulate the consumption of water and also minimize any wastage of this precious resource.

Received 08/January/2024; Accepted 16/June/2024
 ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.24.3.2

Rapidly assessing water levels and developing alternative technologies are crucial steps in addressing the issue at hand. This can involve utilizing machine learning scoring systems and implementing transfer learning models that integrate prior knowledge to enhance the performance of machine learning evaluation systems.

Machine learning (ML) has garnered significant interest as a complement to artificial intelligence (AI) in water resources and agricultural applications. It shows potential effectiveness in addressing the challenges faced by these fields, particularly through decision support systems that can enhance community performance by analyzing real-world situations. The integration of machine learning and artificial intelligence in water resource and agricultural fields will significantly bolster our capacity to tackle critical challenges and promote sustainable development. Machine learning is a branch of artificial intelligence that enables software applications to generate more precise forecasts without explicit coding. Utilizing machine learning scoring systems combined with strategies can provide a practical solution for accurately and promptly determining water levels.

The structure of paper as follows: it begins with the Abstract, followed by the Introduction, which discusses the onset of the problem. Next, the Methods used to address the problem are outlined, along with details on the Dataset used in the research, including how missing data were handled and the analysis of relationships between variables. This is followed by the Training phase, presentation of the Results, the Conclusion, and finally, the References.

II. LITERATURE REVIEW

This literature review critically examines the existing body of research on water quality in water resources, offering a comprehensive overview of key findings, debates, and gaps in the literature. Through a systematic analysis of relevant studies and theories, this section aims to establish a foundation for the current study and contribute to the ongoing scholarly conversation surrounding water quality.

This article discusses how automated irrigation systems, utilizing IoT and wireless sensor networks, contribute to efficient water management in agriculture. The integration of low-cost moisture sensors and intelligent software for data analysis further enhances the effectiveness of these systems. The data utilized in the study includes soil moisture values, water levels in tanks, and well-water data collected through sensors in agricultural fields [1].

The paper underscores the importance of accurately predicting air quality for various environmental and public health purposes, including pollution control, health management, and urban planning, emphasizing the role of big data and machine learning. The authors likely conclude that big data and machine learning approaches provide valuable tools for enhancing the accuracy and efficiency of air quality prediction. The advantages of the paper may highlight the benefits of machine learning algorithms, such as neural networks, support vector machines, and random forests, in capturing complex relationships within air quality data and making precise predictions. Overall, the conclusions of the paper likely emphasize the potential of big data and machine learning approaches to revolutionize air quality prediction, contributing to more effective environmental management and public health protection [2].

The paper proposes a time-series forecasting model for dam water levels, which involves assessing missing values and selecting variables. This model utilizes data from the Klang Gates Dam Reservoir and daily rainfall data. The methods employed in the paper include the Support Vector Machine (SVM) model for forecasting dam water levels using historical data. The data used in the paper were consolidated from the Klang Gates Dam Reservoir and daily rainfall data for the study. Scenario 4

Received 08/January/2024; Accepted 16/June/2024
 ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.24.3.2

demonstrated the highest precision in data forecasting, with the Time Series Regression model outperforming the Support Vector Machine model [3].

The study assessed the benefits of utilizing a feedforward neural network (FFN) in managing a water supply system, indicating that FFNs can be reliably employed to support system operation. The approach demonstrated that FFNs can not only describe the optimal system state but also guide the optimal operation of the entire system based on the latest inflows and storages [4].

The researchers proposed and developed a hybrid GIS decision-making technique supported by an ML algorithm to identify the most appropriate location for constructing a new dam in Sharjah, UAE. The developed Dam Site Suitability Map (DSSM) achieved an accuracy of 83% for dams located in high and moderate zones, with precipitation and drainage stream density identified as the most influential factors. Thematic layers were developed using raw data collected from various sources [5].

The article explores the implementation of smart irrigation systems for managing and planning water resources in sustainable agriculture while preserving ecosystem services. It discusses how these systems utilize advanced technologies to optimize water usage, enhance crop yields, and mitigate environmental impacts. Overall, the results of the paper likely demonstrate the potential of smart irrigation systems to promote sustainable agriculture practices while balancing the needs of food production with environmental conservation [6].

The article utilizes methods such as smart irrigation technologies like soil moisture sensors and evapotranspiration controllers, unmanned aerial vehicles (UAVs), and signal-based evapotranspiration (ET) to achieve significant water savings ranging from 20% to 92% and 20% to 71%, respectively, while maintaining crop growth and quality. The data used in the paper were obtained from experiments using a six-band multispectral camera mounted on a multirotor UAV to monitor plant water content during cotton flowering and boll forming [7].

The paper discusses the utilization of Machine Learning and IoT technologies for smart irrigation management in arboriculture, emphasizing the importance of data collection from IoT sensors and UAV images, along with ML models for monitoring and predicting water status. Contributions of the Paper: The paper highlights the significant contribution of UAV-ML-IoT technologies in irrigation management and water status analysis of crops, as well as the ability of ML algorithms to define models for water estimation. Results of the Paper: The study demonstrated the significant contribution of UAV-ML-IoT technologies in irrigation management and analyzing the water status of crops [8].

Another article explains the methodology employed, including the selection of machine learning algorithms such as decision trees, support vector machines, or neural networks, and the process of feature selection, model training, and evaluation, which likely includes features such as pH level, turbidity, dissolved solids, and other chemical and physical properties of water samples. The results of the study are likely summarized, showcasing the performance of different machine learning algorithms in accurately predicting water potability. The authors may discuss the accuracy, precision, recall, and other metrics used to evaluate the models' performance. In conclusion, the article emphasizes the potential of machine learning techniques in water quality assessment and highlights the importance of accurate prediction for ensuring safe drinking water. It may also suggest future research directions or applications of the proposed methodology [9].

The authors explore how reinforcement learning, a machine learning technique, can be applied to address challenges in deregulated energy markets by enabling agents to learn optimal strategies through interaction with their environment, employing methods such as Q-learning, deep Q-networks, and policy gradient methods. They discuss the challenges faced by deregulated energy markets, such as

Received 08/January/2024; Accepted 16/June/2024
 ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.24.3.2

price volatility and supply-demand imbalances. In conclusion, the article likely emphasizes the significance of reinforcement learning as a promising approach for addressing challenges in deregulated energy markets and suggests directions for further research and application in this field [10].

The article delves into sustainable waste management and water quality prediction for urban areas, underscoring the necessity for efficient and accurate methods to tackle environmental challenges. The employed methodology involves the utilization of optimized algorithms such as least square support vector machine (LS-SVM) and deep learning techniques. LS-SVM is likely applied to model waste management processes, while deep learning methods, such as neural networks, may be employed for predicting water quality parameters. The study's results demonstrate the effectiveness of the proposed approach in enhancing waste management practices and accurately predicting water quality parameters in urban environments [11].

The conclusions of the paper focus on the utilization of predictive machine learning techniques. They summarize the performance of various ML algorithms in predicting water quality parameters, including pH, turbidity, dissolved oxygen, and pollutant concentrations. The authors may discuss the accuracy, precision, and reliability of the models, as well as any limitations or challenges encountered during the evaluation process. Overall, the conclusions emphasize the importance of predictive machine learning in water quality assessment and management and suggest ways to further advance research and application in this field [12].

The article explores the application of machine learning algorithms for accurate water level forecasting in the Muda River, Malaysia, and discusses the associated challenges. This involves exploring various ML algorithms such as decision trees, random forests, support vector machines, and neural networks. Historical water level data, along with relevant meteorological and hydrological parameters, were used to train the algorithms. In conclusion, the article emphasizes the potential of machine learning algorithms to improve water level forecasting in the Muda River and suggests their application for enhancing water resource management practices in Malaysia and similar regions [13].

III. METHODS OF WORK AND MATERIAL

The objective of the researchers' study is to utilize artificial intelligence and machine learning to contribute to the creation of an intelligent framework for classifying water released into water distribution as either potable or non-potable, particularly by water resource authorities. The subsequent phase involves evaluating the effectiveness of the algorithms utilized in this study, analyzing the outcomes, and engaging in discussions to facilitate informed decision-making by relevant authorities regarding water sharing.

First, the data is downloaded using the Pandas library in the Python programming language for preprocessing. This preprocessing stage consists of two steps: handling missing data and assessing the relationship between variables used in training the model. Additionally, duplicate data is removed to ensure data readiness for subsequent stages. This stage is crucial, as it involves dividing the data into three sets: training and testing data, with percentages allocated as follows: 80% for the first training set, 90% for the second, and 70% for the last.

Received 08/January/2024; Accepted 16/June/2024
 ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.24.3.2

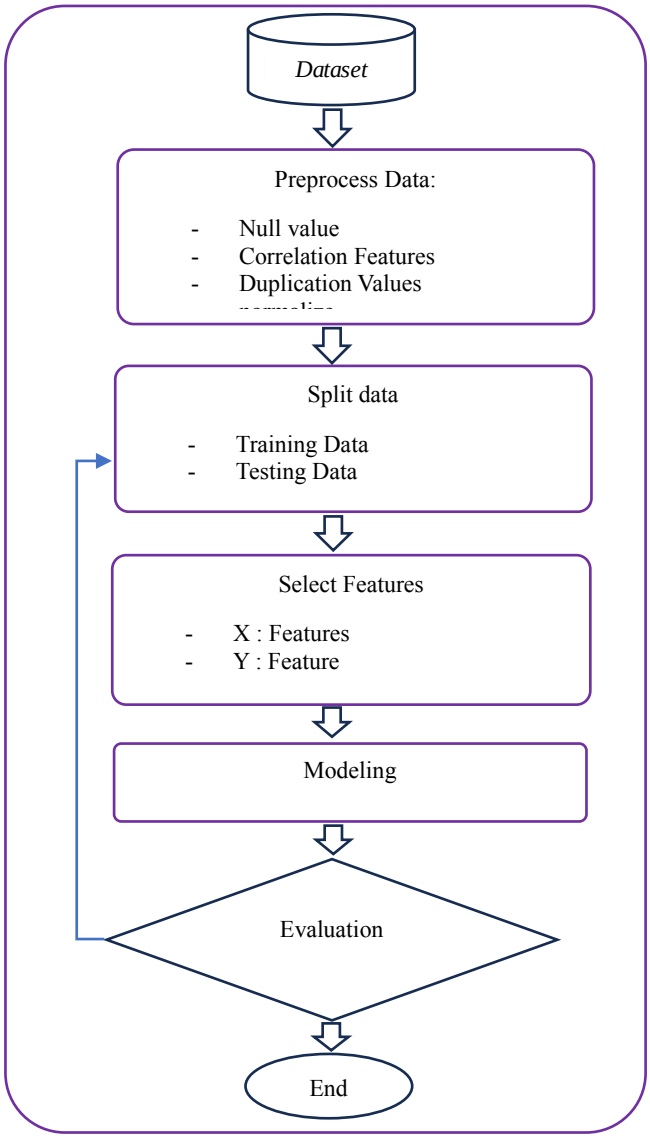

FIG.1. SHOWS STAGES OF SMART MODEL.

Next, the researchers proceed to the modeling stage, implementing the machine learning algorithms adopted in this research. Subsequently, the results are tested to ensure the quality of the model. If the results prove highly accurate and reliable, they are utilized; otherwise, the previous steps are revisited and adjusted accordingly. This framework is perceived as a valuable tool for authorities to make informed investments in water resources, aiding their decision-making process regarding the quantity of emissions to be discharged, as depicted in *Fig. 1*.

IV. DATASET

The study utilized publicly accessible datasets for training and evaluating machine learning models, as outlined in [14]. The dataset comprises 3276 samples with nine attributes and one label indicating water usability. These attributes include pH value, hardness, total dissolved solids (TDS), chloramines, sulfates, conductivity, organic carbon, trihalomethanes, and turbidity [14], as depicted in *Fig. 2*. The pH value provides insights into water acidity and alkalinity, while turbidity data indicates suspended solids and can reflect waste treatment quality in terms of colloidal matter.

Received 08/January/2024; Accepted 16/June/2024

ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.24.3.2

The Total Dissolved Solids (TDS) data provides valuable information about the solubility of both organic and inorganic minerals in water. This data indicates the mineralization level of water, offering insights into its quality and suitability for various purposes. Water has the capability to dissolve a wide range of inorganic and some organic minerals or salts, including potassium, calcium, sodium, bicarbonates, chlorides, magnesium, sulfates, and others. These minerals can impart unwanted taste and diluted color to the appearance of water, making TDS an important parameter for water usage assessment. High TDS values suggest that water is highly mineralized. The desirable limit for TDS in drinking water is typically set at 500 mg/l, with a maximum limit of 1000 mg/l prescribed for drinking purposes. Monitoring TDS levels is essential for ensuring water quality and safety for consumption.

Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Po
204.890	20791.319	7.300	368.516	564.309	10.380	86.991	2.963	0
129.423	18630.058	6.635		592.885	15.180	56.329	4.501	0
224.236	19909.542	9.276		418.606	16.869	66.420	3.056	0
214.373	22018.417	8.059	356.886	363.267	18.437	100.342	4.629	0
181.102	17978.986	6.547	310.136	398.411	11.558	31.998	4.075	0
188.313	28748.688	7.545	326.678	280.468	8.400	54.918	2.560	0
248.072	28749.717	7.513	393.663	283.652	13.790	84.604	2.673	0
203.362	13672.092	4.563	303.310	474.608	12.364	62.798	4.401	0
118.989	14285.584	7.804	268.647	389.376	12.706	53.929	3.595	0
227.231	25484.508	9.077	404.042	563.885	17.928	71.977	4.371	0
165.521	32452.614	7.551	326.624	425.383	15.587	78.740	3.662	0
218.693	18767.657	8.110		364.098	14.526	76.486	4.012	0
156.705	18730.814	3.606	282.344	347.715	15.930	79.501	3.446	0
150.175	27331.362	6.838	299.416	379.762	19.371	76.510	4.414	0
205.345	28388.005	5.073		444.645	13.228	70.300	4.777	0
186.733	41065.235	9.630	364.488	516.743	11.540	75.072	4.376	0

FIG..2. DATASET SNAPSHOTS.

Hardness data relates to water's ability to form soap scum due to its calcium and magnesium content. Sulfate content indicates the concentration of sulfate in water, while chloramine content determines the level of chloramine, a common disinfectant in public water systems. Conductivity indicates water's electrical conductivity, with pure water being a poor conductor. Organic carbon content reflects the presence of carbon in organic compounds in pure water. Trihalomethane (THM) content data provides insights into optimal chlorine concentration and temperature for water treatment. These attributes serve as indicators of water quality and aid in decision-making regarding water treatment and use. The dataset's target feature is potability, with all other features used as inputs in the model.

A. Missing values

Real-life data often contains a significant number of missing values, which can result from data corruption or incomplete documentation. Proper handling of missing data during dataset preprocessing is crucial. One commonly used technique is to fill missing values with the median, as many machine learning algorithms cannot handle gaps in the data. Instead of deleting rows, which can lead to the loss of valuable information and degrade performance, imputing missing values with the median is a viable alternative.

However, the decision on how to handle missing data should consider factors such as the type and characteristics of the dataset, the extent of missing values, and the context of the research. Imputing missing values with the median can be particularly valuable in addressing missing data issues. During the examination of data distribution and variable correlations, it was observed that approximately 44%

Received 08/January/2024; Accepted 16/June/2024

ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.24.3.2

of missing data had been retained, with 15%, 24%, and 5% of data missing in the "PH value," "Sulfate," and "Trihalomethanes" variables, respectively.

To preserve data integrity, columns containing numeric continuous values can be replaced with median values within their respective columns. This approach helps maintain data completeness and ensures that valuable information is retained for subsequent analysis.

B. Correlation Between Features

The researchers consider this stage to be of utmost importance, as it plays a crucial role in shaping and establishing relationships among variables. This stage is essential for achieving accurate results and making informed decisions and classifications within the model. Researchers establish relationships between the input variables, which the model will be trained on, and the target variable, which the model classifies based on these relationships. This process ensures that the model learns the underlying patterns and dependencies within the data. *Fig. 3* illustrates this relationship and its significance in the overall modeling process.

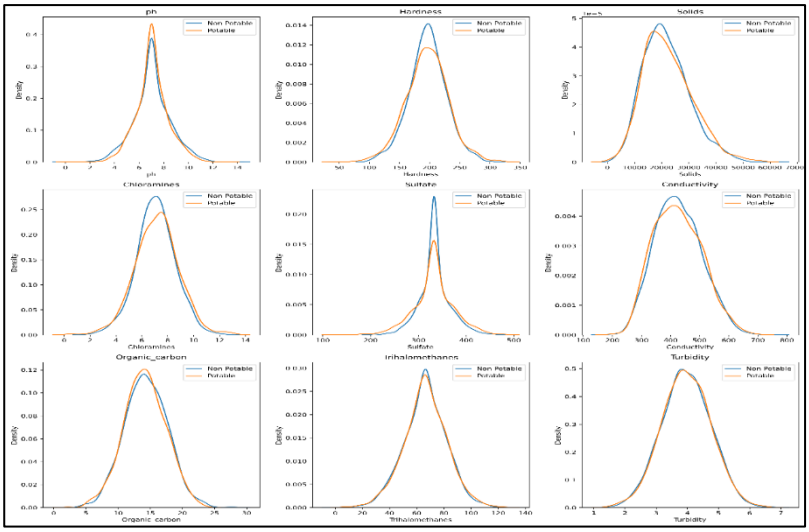

FIG. 3. SHOWS CORRELATION BETWEEN FEATURES.

There is a clear and significant relationship between each variable and the target variable, with the highest correlation observed with "Solids" at 5.24%. This correlation indicates the feasibility of working with this dataset. The researchers considered this relationship as one of the solutions presented in their scientific paper, indicating that machine learning algorithms can produce feasible and applicable results. In cases where a weak relationship exists between variables and the target variable, researchers addressed it using the methods described in this work.

The researchers noted through the above relationship that it is possible to process some features to determine which of these features are of great importance in order to obtain accurate and meaningful insights.

Received 08/January/2024; Accepted 16/June/2024

ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.24.3.2

TABLE I. CORRELATION BETWEEN FEATURES

Features	Correlation values
Potability	1.000000
Solids	0.33743
Organic_carbon	0.030001
Chloramines	0.023779
Sulfate	0.020476
Hardness	0.013837
Conductivity	0.008128
Trihalomethanes	0.006887
ph	0.003014
Turbidity	0.001581

Table I shows the correlation between features, thus we can consider removing features with correlation values close to zero researchers get as they are less likely to contribute significantly to predicting the target variable. In our case, might consider removing `Trihalomethanes`, `ph`, and `Turbidity` since their correlation values are very low (`<= 0.01`). So, we remove these features from your dataset before training your model:

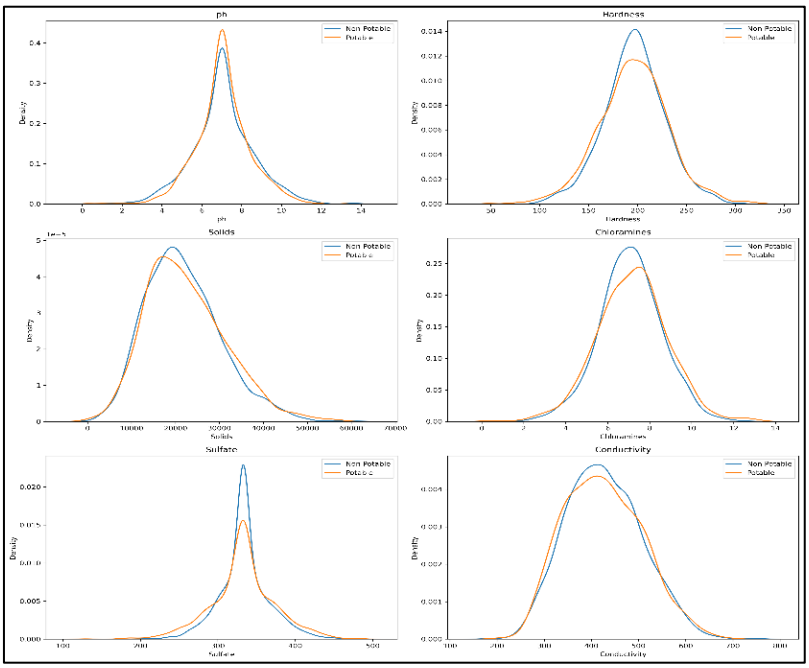

FIG. 4. SHOWS CORRELATION BETWEEN FEATURES (AFTER REMOVE FEATURES).

By removing these features, you may simplify your model and potentially improve its performance, especially if these features were not providing meaningful information for predicting the target variable, as shown in *Fig. 4.* However, always remember to validate the impact of feature removal on your model's performance through proper evaluation techniques like cross-validation.

Received 08/January/2024; Accepted 16/June/2024

ISSN (Print) 1811-9212 *ISSN (Online) 2617-3352*

DOI: https://doi.org/10.33103/uot.ijccce.24.3.2

Analyzing these relationships leads to several hypotheses: most traits follow normal distribution patterns, some features exhibit minor outliers, and there is a noticeable class imbalance in the dataset, with certain classes appearing more frequently than others. These insights inform the researchers' approach to data analysis and model development, guiding them in addressing potential challenges and optimizing the performance of their machine learning algorithms.

C. Normalizing Dataset

In the field of machine learning and statistics, datasets often originate from various sources, leading to differences in their ranges. This discrepancy poses a significant challenge when training machine learning algorithms, as inputs with varying ranges can have different effects on the algorithm's performance. This can result in prolonged training times, unsatisfactory results, or overfitting issues.

Normalization or standardization is a crucial preprocessing step aimed at scaling the data to a specific range, such as between 0 and 1 or between -1 and +1. It becomes necessary when significant variations exist in the ranges of different features within the dataset. This scaling method is particularly effective when the dataset does not contain outliers.

The theoretical foundation of normalization can be comprehensively understood through various academic sources [15]. If the dataset follows a Gaussian distribution, it can be normalized to a standard normal distribution using a specific formula. This normalization process ensures that the data is uniformly scaled, facilitating more efficient and accurate model training.

In the normalization process, the variables are transformed using the formula:

$$z = \frac{x-\mu}{\sigma} \quad (1)$$

Where:
- (x) represents the variable being normalized.
- (μ) is the mean (average) of the variable.
- (σ) is the standard deviation of the variable.

This formula adjusts the values of the variable (x) by subtracting the mean (μ) and dividing by the standard deviation (σ). This transformation ensures that the normalized variable (z) has a mean of 0 and a standard deviation of 1, resulting in a standardized distribution.

D. Training and Testing Dataset

At this stage, the researchers partitioned the data into two distinct sets: the training data and the testing data. The training data is of utmost importance as it is utilized to train machine learning algorithms. In the initial section, the data was divided into 80% for training and 20% for testing. Subsequently, in the second section, a split of 90% for training and 10% for testing was implemented. Finally, in the third section, a split of 70% for training and 30% for testing was utilized, as elaborated in the results section.

Following the data split, the variables to be used as input for the model were identified, along with the target variable. The target variable serves as the focal point for the learning process, enabling the model to classify whether the water is safe for consumption. This delineation of input and target variables establishes the foundation for the subsequent model training and evaluation processes.

Received 08/January/2024; Accepted 16/June/2024
 ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.24.3.2

V. MODELING

In this paper, researchers utilized five distinct types of machine learning algorithms. These algorithms were selected based on their suitability for data classification tasks, making them applicable to the dataset utilized in the proposed model. By employing a variety of algorithms, the researchers aimed to explore different approaches to data analysis and classification, thereby enhancing the robustness and reliability of their proposed model.

A. The Logistic Regression

The Logistic Regression is a widely employed method in the field of machine learning, frequently applied for classification problems. Its fundamental objective is to model the probability of an event occurring based on one or more predictor variables. Unlike linear regression, Logistic Regression is specifically designed for situations where the dependent variable is categorical, typically binary, such as classifying emails as spam or not spam. The model employs the logistic function to transform a linear combination of input features into a probability score between 0 and 1. Logistic Regression is esteemed for its simplicity, interpretability, and efficiency, making it a valuable tool in various domains, ranging from healthcare and finance to social sciences and marketing. Its versatility, ease of implementation, and ability to handle high-dimensional datasets contribute to its enduring relevance in academic research and practical applications[16].

B. The K-Nearest Neighbor

The K-Nearest Neighbors (KNN) algorithm holds a prominent position in the field of machine learning, revered for its simplicity and adaptability. Operating as a non-parametric and instance-based method, KNN relies on the proximity principle to formulate predictions. When faced with a new data point, the algorithm classifies it by examining the majority class among its K nearest neighbors within the feature space, determined by a selected distance metric.

One of KNN's key strengths lies in its applicability to both classification and regression tasks, rendering it suitable for various academic disciplines and research domains. Its concept, which involves learning from the local neighborhood of data points, facilitates adaptability to complex decision boundaries and diverse data distributions.

Despite its straightforward approach, KNN continues to attract attention from researchers, who explore avenues for enhancement, such as optimized distance metrics, efficient neighbor search techniques, and hybrid methodologies. The algorithm's capacity to handle structured and unstructured data, alongside its interpretability, positions KNN as a valuable asset for a wide range of applications[17].

C. The Random Forest

Random Forest emerges as a powerful ensemble learning technique highly regarded in the realm of machine learning and predictive modeling. Rooted in the decision tree framework, this algorithm operates by constructing numerous decision trees during training and amalgamating their outputs to generate robust and accurate predictions. The "random" component introduces variability in both the data samples used for training each tree and the features considered at each split, thereby mitigating overfitting and enhancing generalization.

Random Forest excels in handling diverse datasets characterized by high dimensionality and complex interactions. Its feature importance scores facilitate interpretability, rendering it an appealing choice for academic research spanning disciplines such as ecology, finance, and healthcare. Moreover, the algorithm's robustness to outliers, capability to manage missing data, and built-in parallelization contribute to its widespread adoption in real-world applications.

Received 08/January/2024; Accepted 16/June/2024
 ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.24.3.2

Continual advancements in research aim to refine the algorithm's efficiency, scalability, and adaptability to emerging challenges, reinforcing Random Forest's status as a cornerstone in the landscape of machine learning methodologies[18].

D. The Decision Tree

Decision Trees represent a cornerstone of machine learning and data analysis, offering a transparent and interpretable framework for decision-making processes. This algorithm recursively divides the feature space based on the most informative features at each node, forming a tree-like structure conducive to both classification and regression tasks. Decision Trees are widely embraced in academia and diverse research domains owing to their intuitive depiction of decision logic and the capacity to unveil intricate relationships within datasets.

Despite their vulnerability to overfitting, researchers have devised strategies such as pruning and ensemble methods, like Random Forests, to bolster their robustness and generalization capabilities. Decision Trees find extensive applications across academic fields, spanning epidemiology, environmental science, business, and social sciences. Ongoing research endeavors focus on optimizing their performance, mitigating bias, and extending their utility to high-dimensional and heterogeneous datasets.

The enduring appeal of Decision Trees stems from their harmonious blend of simplicity, interpretability, and adaptability, rendering them indispensable tools for both academic inquiry and practical problem-solving endeavors[19].

E. The Support Vector Machine

Support Vector Machines (SVMs) stand out as a potent class of supervised learning algorithms, commanding significant attention across academic disciplines and practical domains. Rooted in the field of machine learning, SVMs excel in both classification and regression tasks by establishing an optimal hyperplane that effectively separates different classes within the feature space. This capability proves particularly advantageous in scenarios characterized by complex and nonlinear data relationships, facilitated by the utilization of kernel functions that implicitly map input data into higher-dimensional spaces.

The versatility and rsobustness of SVMs have led to their widespread adoption in fields such as biology, finance, and image recognition. Their ability to handle high-dimensional datasets further enhances their applicability across diverse domains. Despite their effectiveness, SVMs necessitate careful parameter tuning, and ongoing research endeavors explore advancements such as integrating deep learning techniques and developing scalable algorithms to further enhance their performance and applicability[20].

VI. RESULTS AND DISCUSSION

According to the results obtained from the proposed model on the water resources dataset, trained using various machine learning algorithms, are as follows:

Initially, the accuracy of classification for each algorithm was evaluated separately to develop an intelligent framework for classifying water into potable or non-potable categories. The researchers partitioned the data into different batches to assess the model's efficiency and behavior under varying data availability scenarios during the training phase.

As illustrated in Table II and III the data was divided into an 80% training set and a 20% testing set. Among all the algorithms applied in this study, the Random Forest algorithm exhibited the highest classification accuracy. Following closely was the Logistic Regression algorithm, succeeded by the K-Nearest Neighbor algorithm, which demonstrated its proficiency in classifying the applied data.

Received 08/January/2024; Accepted 16/June/2024
 ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.24.3.2

While the Decision Tree algorithm is recognized as a robust classifier for such data, it ultimately displayed its accuracy comparable to Support Vector Classification (SVC). These findings underscore the effectiveness of machine learning algorithms in classifying water resources data and highlight the Random Forest algorithm's superiority in this context.

TABLE II. ACCURACY (80% TRAINING, 20% TESTING)

Model	Accuracy
Logistic Regression	0.959932
KNN	0.861170
Random Forest	0.954084
Decision Tree	0.882214
SVC	0.611832

TABLE III. CONFUSION MATRIX FOR ALL ALGORITHMS

Model	Confusion Matrix		
		Actual	
Logistic Regression	Predicted	1402	63
		189	1622
		Actual	
KNN	Predicted	1341	461
		173	1301
		Actual	
Random Forest	Predicted	1487	12
		151	1626
		Actual	
Decision Tree	Predicted	1491	158
		135	1492
		Actual	
SVC	Predicted	1252	398
		487	1139

Second, the results show As in Table IV the accuracy of the classification for each algorithm separately by measuring it, the researchers divided the data into (90% training and 10% testing),where the (Decision Tree) algorithm classification results is the best from other algorithms applied in this work, followed by the Logistic Regression algorithm, and then the K-Nearest Neighbor. While the Decision Tree algorithm is considered one of the powerful algorithms in this field to classify such data, as it finally showed its accuracy SVC. Logistic Regression model achieved accuracy 0.730934, indicating a moderate balance between precision and recall. This suggests that the model may face challenges in effectively capturing both true positives and minimizing false positives and false negatives. Further investigation into feature engineering, model complexity, and potential data preprocessing may be beneficial to enhance its performance. KNN accuracy 0. 876528 exhibits a slightly improved overall performance compared to Logistic Regression. KNN's strength lies in its ability to capture local patterns, but its performance can be sensitive to the choice of hyperparameters such as the number of neighbors (K). Fine-tuning K and considering feature scaling might lead to further improvements. Accuracy of Random Forest 0.882484 demonstrates a relatively higher Accuracy of 0. 882484, suggesting a more effective balance between precision and recall. Random Forests are known for

Received 08/January/2024; Accepted 16/June/2024
 ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.24.3.2

handling complex relationships and high-dimensional data well. This result may indicate that ensemble methods and decision trees contribute positively to the model's predictive capacity. Decision Tree Accuracy (0. 919293): The Decision Tree model, with Accuracy (0. 919293):, performs competitively with KNN and Logistic Regression. Decision Trees are susceptible to overfitting, and optimizing hyperparameters or employing ensemble methods like pruning could be explored to further enhance its generalization capabilities. svc accuracy (0. 746351): The Support Vector Classifier lags significantly behind the other models, as evidenced by its low accuracy 0. 746351. This might indicate issues with model complexity, kernel selection, or inadequate feature scaling. Reevaluation of model parameters and consideration of alternative kernel functions could be crucial for improving performance.

TABLE IV. ACCURACY (90% TRAINING, 10% TESTING)

Model	Accuracy
Logistic Regression	0. 730934
KNN	0. 876528
Random Forest	0. 882484
Decision Tree	0. 919293
SVC	0. 746351

Third, Researchers divided the data into (70% training and 30% testing), As illustrated in Table V the classification (Decision Tree) algorithm showed the highest classification results among all the algorithms applied in this work, followed by the (Logistic Regression) algorithm and then the (K-Nearest Neighbor), which showed its ability to classify this applied data While the Decision Tree algorithm is considered one of the powerful algorithms in this field to classify such data, as it finally showed its accuracy (SVC).The Logistic Regression model exhibits a moderate accuracy of 0. 866820, implying a balanced trade-off between precision and recall. Fine-tuning regularization parameters and exploring feature engineering could potentially enhance its performance. K-Nearest Neighbors accuracy 0.891297 indicating reasonable precision and recall equilibrium. Further optimization of the number of neighbors (K) and feature scaling might be considered for incremental improvements. Random Forest accuracy 0. 931849 The Random Forest model demonstrates a competitive accuracy 0.931849 suggesting effective ensemble learning and feature selection. Refining hyperparameters and assessing feature importance could potentially enhance its predictive capability. Decision Tree model stands out with a higher accuracy of 0.903081, indicating a robust balance between precision and recall. Pruning techniques and additional hyperparameter tuning could be explored to optimize its performance further. Support Vector Classifier (SVC) accuracy of 0. 604886 : The Support Vector Classifier lags behind with a low accuracy, suggesting challenges in capturing both true positives and minimizing false positives/negatives. Careful tuning of kernel functions and regularization parameters may be imperative for substantial improvement.

Received 08/January/2024; Accepted 16/June/2024
 ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.24.3.2

TABLE V. ACCURACY (70% TRAINING, 30% TESTING)

Model	Accuracy
Logistic Regression	0. 866820
KNN	0. 891297
Random Forest	0. 931849
Decision Tree	0. 903081
SVC	0. 604886

Rigorous tuning of kernel functions and regularization parameters is imperative for significant improvement. Which gave important results, and this is a clear indication for researchers that through the application of these algorithms on water resource data and the employment of machine learning in determining and employing the extent to which water can be used for drinking or not, which was launched during the water quotas launched by the competent authority. Which comes the role of this proposed smart model who gave assistance to these authorities and before launching water quotas for drinking to determine the extent to which this water can be used for drinking or not, and this will be an effective and important role in determining such decisions for the competent authority and therefore in the event that these quotas that have been classified Although it is not suitable for drinking, it can be helped to conclude that this water can be given to other rations such as agricultural rations, irrigation rations and farmers' rations to be used more effectively and to face water scarcity in the event of its occurrence by giving a clear indication that this water can be used in other rations instead Who gave it to the water rations, and it cannot be given to the drinking water ration that was allocated for this purpose and The proposed model also reduces human efforts and energy.

VII. CONCLUSIONS

Managing and planning water resources poses significant challenges due to numerous variables, uncertainties, and risks involved. The authors propose leveraging machine learning algorithms as part of artificial intelligence to classify datasets and determine if water is potable or not, thereby assisting in the allocation of water resources for drinking purposes. To address this issue, the paper utilizes a machine learning model to contribute to the creation of a smart system capable of distinguishing between potable and non-potable water.

The researchers observed variations in results when applying the algorithm to different proportions of training and testing data. Publicly accessible datasets comprising 3,276 samples with nine attributes and two labels indicating water usability were utilized for training and evaluating machine learning models. According to the results obtained Random Forest algorithm outperformed others, achieving the highest accuracy 0.954084, 0. 882484, and 0. 931849 for three sizes of training and testing data as well as confusion matrix. Results obtained in researchers work the (Random Forest) algorithm have the best, and then (Decision Tree , K-Nearest Neighbor ,Logistic Regression, SVC) sequentially were sequentially ranked in terms of performance.

REFERENCES

[1] S. Vaishali, S. Suraj, G. Vignesh, S. Dhivya, and S. Udhayakumar, "Mobile integrated smart irrigation management and monitoring system using IOT," in *2017 international conference on communication and signal processing IEEE (ICCSP)*, pp. 2164-2167, April 2017.

Received 08/January/2024; Accepted 16/June/2024
 ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.24.3.2

[2] G. K. Kang, J. Z. Gao, S. Chiao, and G. Xie, "Air quality prediction: big data and machine learning approaches," *Int. J. Environ. Sci. Dev*, pp. 8-16, 2018.

[3] X. Cheng, Q. Li, Z. Zhou, Z. Luo, M. Liu, and L. Liu, "Research on a seepage monitoring model of a high core rockfill dam based on machine learning," *Sensors*, vol. 18, no. 9, p. 2749, 2018.

[4] W. J. Khai, M. Alraih, A. N. Ahmed, C. M. Fai, A. El-Shafie, and A. El-Shafie, "Daily forecasting of dam water levels using machine learning," *Int. J. Civ. Eng. Technol (IJCIET)*, vol. 10, pp. 314-323, 2019.

[5] E. Rozos, "Machine learning, urban water resources management and operating policy," *Resources*, vol. 8, no. 4, p. 173, 2019.

[6] R. Al-Ruzouq, A. Shanableh, A. G. Yilmaz, A. Idris, S. Mukherjee, M. A. Khalil, and M. B. A. Gibril, "Dam site suitability mapping and analysis using an integrated GIS and machine learning approach," *Water*, vol. 11, no. 9, 2019.

[7] D. Masseroni, G. Arbat, and I. P. de Lima, "Managing, and planning water resources for irrigation: Smart-irrigation systems for providing sustainable agriculture and maintaining ecosystem services," *Water*, vol. 12, no. 1, 2020.

[8] A. S. A. Sukor, M. N. Muhamad, and M. N. Ab Wahab, "Development of In-situ Sensing System and Classification of Water Quality using Machine Learning Approach," in *2022 IEEE 18th International Colloquium on Signal Processing & Applications, IEEE (CSPA)*, pp. 382-385, May 2022.

[9] S. Lee, D. Kaown, E. H. Koh, H. L. Lee, K. S. Ko, and K. K. K. Lee, "Advanced utilization of multi-learning algorithm: ensemble super learner to map groundwater potential for potable mineral water," *Geocarto International*, pp. 1-20, 2022.

[10] A. R. Kumar, R. Nithisha, K. Mounika, and V. Savitha, "Online Monitoring System for Water Quality Based on Machine Learning Algorithms," *Lampyrid: The Journal of Bioluminescent Beetle Research*, vol. 13, pp. 328-334, 2023.

[11] Zhang, S.; Omar, A.H.; Hashim, A.S.; Alam, T.; Khalifa, H.A.E.-W.; Elkotb, M.A. Enhancing waste management and prediction of water quality in the sustainable urban environment using optimized algorithm of least square support vector machine and deep learning techniques. Urban Clim., 49, 101487,(2023).

[12] Ghosh, H., Tusher, M. A., Rahat, I. S., Khasim, S., & Mohanty, S. N.. Water Quality Assessment Through Predictive Machine Learning. In International Conference on Intelligent Computing and Networking (pp. 77-88). Singapore: Springer Nature Singapore. (2023).

[13] Drogkoula, M., Kokkinos, K., & Samaras, N. A Comprehensive Survey of Machine Learning Methodologies with Emphasis in Water Resources Management. Applied Sciences, 13(22), 12147. (2023).

[14] Zakaria, M.N.A.; Ahmed, A.N.; Malek, M.A.; Birima, A.H.; Khan, M.H.; Sherif, M.; Elshafie, A. Exploring machine learning algorithms for accurate water level forecasting in Muda river, Malaysia. Heliyon, 9, e17689. (2023).

[15] Ali, P. J. M., Faraj, R. H., Koya, E., Ali, P. J. M., & Faraj, R. H.. Data normalization and standardization: a technical report. Mach Learn Tech Rep, 1(1), 2014.

[16] J. Mata, F. Salazar, J. Barateiro, and A. Antunes, "Validation of machine learning models for structural dam behaviour interpretation and prediction," *Water*, vol. 13, no. 19, 2021.

[17] S. Touil, A. Richa, M. Fizir, J. E. Argente Garcia, and A. F. Skarmeta Gomez, "A review on smart irrigation management strategies and their effect on water savings and crop yield," *Irrigation and Drainage*, 2022.

[18] Y. Ahansal, M. Bouziani, R. Yaagoubi, I. Sebari, K. Sebari, and L. Kenny, "Towards smart irrigation: A literature review on the use of geospatial technologies and machine learning in the management of water resources in arboriculture," *Agronomy*, vol. 12, no. 2, p. 297, 2022.

[19] R. Alnaqeb, F. Alrashdi, K. Alketbi, and H. Ismail, "Machine Learning-based Water Potability Prediction," in *2022 IEEE/ACS 19th International Conference on Computer Systems and Applications, IEEE (AICCSA)*, pp. 1-6, December 2022.

[20] D. Poudel, D. Shrestha, S. Bhattarai, and A. Ghimire, "Comparison of machine learning algorithms in statistically imputed water potability dataset," 2022, *preprint*.

Received 08/January/2024; Accepted 16/June/2024
 ISSN (Print) 1811-9212 *ISSN (Online) 2617-3352*