Machine Learning-based Water Potability Prediction

Reem Alnaqeb, Fatema Alrashdi, Khuloud Alketbi, Heba Ismail
Department of Computer Science and Information Technology
Abu Dhabi University
Abu Dhabi, United Arab Emirates
{1070234, 1071904,1081633}@students.adu.ac.ae,heba.ismail@adu.ac.ae

2022 IEEE/ACS 19th International Conference on Computer Systems and Applications (AICCSA) | 979-8-3503-1008-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/AICCSA56895.2022.10017579

Abstract— **Protecting and caring for water is one of the most critical environmental problems today. This research aims to design an intelligent system using machine learning models to improve water quality and predict whether it is safe to be used as drinking water. Several models of machine learning algorithms are compared to find the best model to be used for the accuracy of prediction of water quality. In this research, we compare Decision Tree, K-Nearest Neighbor, Support Vector Machine, Ransom Forest, and LightGBM models to get the best model for water potability prediction. Experimental results show that LightGBM model produced the best prediction accuracy of 99.74% on the experimental data.**

Keywords— water quality, potability detection, machine learning, Decision Tree, K-Nearest Neighbor, Support Vector Machine, Ransom Forest, LighGBM.

I. INTRODUCTION

Safe and accessible water is crucial for public health, whether for drinking, residential usage, food production, or leisure. Water supply, sanitation, and water resource management may increase economic development and reduce poverty. Absent, insufficient, or poorly managed water and sanitation pose health concerns. Globally, 15% of hospitalized patients get an infection, with the percentage higher in low-income nations [1]. Therefore, this research focus on water quality prediction. We design a system to predict water quality using machine learning algorithms and test the system on the water quality dataset [2]. The dataset used includes 8000 samples described in terms of the chemical components of water and whether it is drinkable [2].

Potable water, known as drinking water, comes from surface and groundwater resources. It requires careful monitoring of pathogens and chemicals which could be contaminated due to their high concentration in water [3]. The quality of water should be examined to secure the safety of human beings, whether it is for drinking, domestic use, or food production. The quality of water plays a vital aspect in boosting countries' economies. In many countries, people do not have access to clean drinking water. Therefore, the need of having drinkable water becomes essential, and the need to determine the water quality to solve this issue.

Water quality detection using machine learning was discussed in some studies. For instance, R. Chafloque et. Al. used neural networks to detect water suitable for human consumption and obtained an accuracy of approximately 70% [3]. Likewise, L. Fen, Z. Lei, and Z. Ting [4] used the Extra trees classifier as the best model among the others, with an accuracy of 86.67%. M. I. K. Haq, applied Decision Tree Algorithm, which supports the drinking water quality with a mark of 97.23% of accuracy [5].

This paper attempts to improve on previous results and achieve the highest accuracy of water potability predictions. Furthermore, in this research, we focus on identifying critical characteristics that have a direct impact on water potability prediction. This paper began by explaining previous studies and research findings, as well as limitations. Following that, the data description section began by collecting data sets, processing them, applying machine learning algorithms, and determining the outcome. Finally, a table of comparisons is provided to view the outcomes of the selected classifiers for water potability prediction.

II. LITERATURE REVIEW

Many studies have explored machine learning predictive models for water quality detection using supervised approach. This section reviews and summarizes some of the most relevant research works in water quality detection using classification-supervised learning. Table I presents a summary of all reviewed research studies.

R. Chafloque et al. [3] presented a Predictive Neural Networks Model for water quality detection for human consumption. It predicts if the given water is proper for human consumption. The used dataset contains 3276 entries, nine features, and 1 target class. The MinMax scaling method was achieved from the sklearn library during the pre-processing. The proposed model is built on an architecture that employs neural networks created in Python. Furthermore, the Keras Library was used to provide the required techniques for implementing the seven dense layers that would make up the neutral network. Finally, the proposed model obtained an accuracy of 70%.

A study by F. Li, L. Zhou, and T. Chen [4] evaluated the performance of five classifiers: Decision tree, Naïve Bayes, Support vector machine (SVM), Linear model, and k nearest neighbors (KNN), to predict class labels in the form of the quality of the water. The dataset has nine features and two class labels; the first label is zero, which indicates the water is non-potable water, and one for potable water. The used dataset was unbalanced; therefore, they utilized the resampling approach. The dataset is separated by a cross-validation testing approach with ten folds—performance comparison of decision Tree and other models. The decision tree classifier model was found to have the highest accuracy with 86.67 %.

Another study conducted by Haq et al. [5] studied Water Potability using different classification algorithms. The dataset used contained nine features and two labels, 0 for non-potable water and 1 for potable water. Two significant models were used, decision tree and naive Bayes. They used cross-validation testing mode with 10 and 5 folds to evaluate the

TABLE I. LITERATURE REVIEW SUMMARY

Research Study	Class Datatype	Number of Class Labels	Processing Used Techniques	Used Models	Performance Measures	Extracted Features	Best Accuracy	Evaluation Approach	Class Distribution
R. Chafloque et al. [3]	Binary	• 0→ non-potable • 1→ potable water	• MinMax scaling method	• 7 layers of ANN	• Accuracy	3276 Features	69% accuracy with the ANN classifier	Training (70%) and testing (30%)	Imbalanced
L. Fen, Z. Lei and Z. Ting [4]	Binary	• 0→ non-drinkable • 1→ drinkable	• Delete missing values • The resampling method to deal with data imbalance	• K Nearest Neighbors • Support Vector Machine • Decision Tree • Naive Bayes • Linear Model	• Accuracy • Recall • Precision • F1 Measure	3276 Features	86.67% accuracy with the DT classifier	Cross validation with 10 folds	Imbalanced
Haq et al. [5]	Binary	• 0→ non-potable • 1→ potable water	• Eliminating missing values in the dataset	• Decision Tree • Naive Bayes	• Accuracy	3276 Features	96.26% accuracy with the DT classifier	Cross validation with 10 and 5 folds	Imbalanced
Ahmed et al. [6]	WQC (Multi-class)	• 0-25 → Very bad • 25-50 → Bad • 50-70→ Medium • 70 -90 → Good • 90-100→ Excellent	• Water Quality Class • Q-Value Normalization • Z-score Normalization	• MLP • Naïve Bayes • Logistic Regression • Stochastic Gradient Descent • KNN • Decision Tree • Random Forest • SVM	• Accuracy • Recall • Precision • F1 Measure	PCRWR contained 663 samples from 13 different sources.	85% accuracy with the MLP classifier	Cross-validation with k-1 subsets as the training dataset and 1 subset as the testing dataset.	Imbalanced
Kurra, Naidu, Chowdala, Yellanki, & Sunanda [7]	Binary	• 0→ water is_safe • 1→ water is_not_safe	• Handling Missing Values • Partitioning	• Decision Tree • KNN	• Accuracy • Recall • Precision • F1 Measure	3276 Features	61.7 accuracy with the KNN classifier	Training (80%) and testing (20%) and Cross validation	Imbalanced

performance of models. Their results show that the Decision tree classifier outperforms other classifiers with an accuracy of 97.22%. A study by Ahmed et al. [6] evaluated a variety of supervised machine learning techniques for calculating the water quality index (WQI) and the water quality class (WQC) based on a number of classification algorithms that characterize the water quality. The dataset used in the study was collected from the PCRWR and consisted of 51 samples and 12 parameters. The study uses boxplot analysis to discover outliers in addition to Q-value and Z-score normalization. The study evaluated the dataset to a variety of common supervised predictions, including classification and regression algorithms. Models are utilized in the study are Nave Bayes, KNN, and DT. The MLP classifier performed better in predicting WQC where it produced an accuracy of 85%.

Another study by Kurra, Naidu, Chowdala, Yellanki, & Sunanda [7] evaluated data mining methods Decision Tree (DT) and K- Nearest Neighbor (KNN) which examine how machine learning algorithms may be used to predict water quality that aims to use a prescriptive analysis based on projected values for future capabilities. The dataset has 9 parameters and two class labels in terms of water portability; zero is non-potable water, and one is potable water. The study data set was split into two parts using the K-fold cross-validation technique. Their results show that the K- Nearest Neighbor classifier outperforms than Decision Tree with an accuracy of 61.7%.

To the best of the author's knowledge, this is the first work to develop several classifiers with more than 20 features of water elements and water quality indexes of labeling and their effect on water quality.

III. DATA ANALYTIC FLOW

This work aims to determine if water is suitable for humans through machine learning classification. The analytical data flow encompasses the complete procedure of the dataset throughout the experiment, including the procedure of the study's design, documentation, and presentation of results. Flexibility and step-by-step dependability define the logical sequencing of these processes and the dynamics of the efficient research workflow. Using different machine learning algorithms, tests are done by looking at water's characteristics as a function of physical, biological, and chemical conditions.

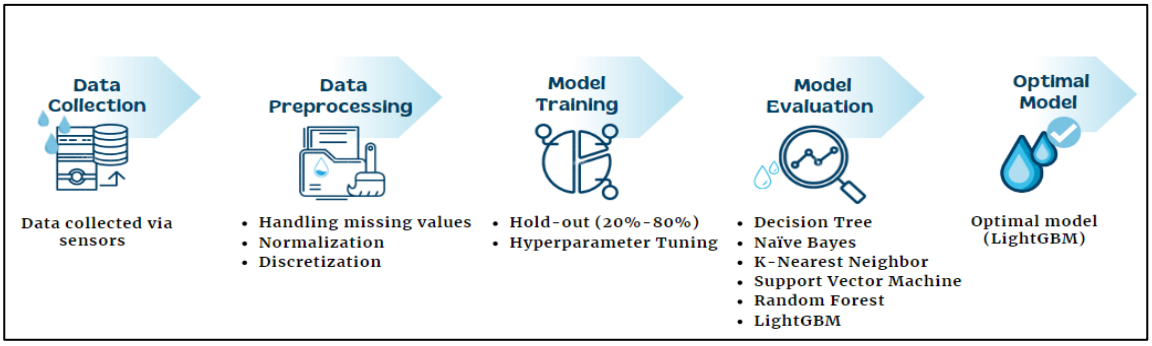

Figure 1. Data analytical flowchart

Figure 1 illustrates the proposed data analytics framework and the model's design utilized in this paper. First, raw data is obtained using water quality sensors such as total organic carbon (TOC), turbidity, and oxygen-reduction potential (ORP) sensors. Therefore, the water quality dataset is used for decision-making to improve the natural environment's health. Secondly, preprocessing data is an essential preliminary step, manipulating and transforming raw data into an understandable form by handling missing data and applying normalization and discretization. This step resolves different issues and makes datasets completer and more efficient for further data analysis. Then, we use the hold-out technique to split our data into training and testing sets in order to fine-tune the default parameters and improve the running module's performance. In the next step, the models are executed, and each model's performance scores are evaluated, using a variety of platforms such as H2o, and Python. Finally, this process culminates in the selection of the model that provides the best overall performance.

IV. DATA DESCRIPTION

The analysis is carried out using a publicly available dataset for water quality [2]. The dataset contains 8000 rows, 20 features, and two class labels related to several water elements, water quality indexes of labeling, and their effect on water quality. The feature "Is_safe" represents the target output of the dataset that shows whether the water is safe or not.

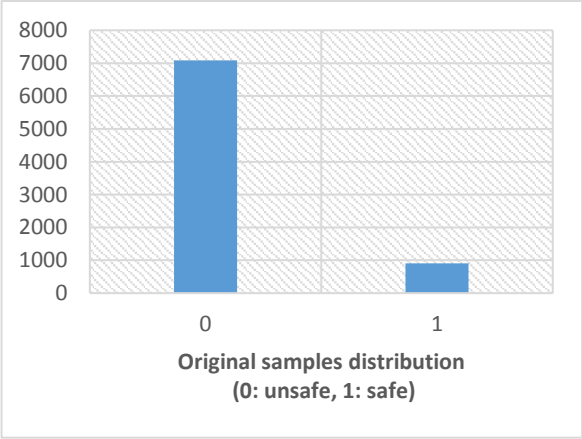

Figure 2. Number of samples in each class labels

TABLE II. DATASET ATTRIBUTES DESCRIPTIONS

No.	Attribute	Data Type	Data Description	Missing Values
1	Aluminum	Numeric	Dangerous if greater than 2.8	No
2	Ammonia	Numeric	Dangerous if greater than 32.5	Yes
3	Arsenic	Numeric	Dangerous if greater than 0.01	No
4	Barium	Numeric	Dangerous if greater than 2	No
5	Cadmium	Numeric	Dangerous if greater than 0.005	No
6	Chloramine	Numeric	Dangerous if greater than 4	No
7	Chloramine	Numeric	Dangerous if greater than 0.1	No
8	Copper	Numeric	Dangerous if greater than 1.3	No
9	Fluoride	Numeric	Dangerous if greater than 1.5	No
10	Bacteria	Numeric	Dangerous if greater than 0	No
11	Viruses	Numeric	Dangerous if greater than 0	No
12	Lead	Numeric	Dangerous if greater than 0.015	No
13	Nitrates	Numeric	Dangerous if greater than 10	No
14	Nitrites	Numeric	Dangerous if greater than 1	No
15	Mercury	Numeric	Dangerous if greater than 0.002	No
16	Perchlorate	Numeric	Dangerous if greater than 56	No
17	Radium	Numeric	Dangerous if greater than 5	No
18	Selenium	Numeric	Dangerous if greater than 0.5	No
19	Silver	Numeric	Dangerous if greater than 0.1	No
20	Uranium	Numeric	Dangerous if greater than 0.3	No
21	Is_safe (target)	Numeric	0 - not safe, 1 - safe	Yes

As shown in Figure 2, a value of 1 indicates that the water is safe, while a value of 0 indicates that it is not safe. The dataset shows approximately 80% of sample water is potable, and 20% of sample water is not potable. The significant difference in sample diversity might incur training errors in

the machine learning [8]. A detailed description of the experimental datasets features is given in Table II.

V. DATA PREPROCESSING

In the preprocessing step, we have noticed there are certain invalid values in the "Ammonia" and "Is_safe" columns, and they are represented with NaN, which is shorthand for "not a number," which is regarded as a missing value. Therefore, we removed the rows from the dataset that included NaN values in the "Ammonia" and "Is_safe" columns. Following that, we performed normalization on some of the attributes whose values ranged between zeros, ones, and tens, which is deemed inappropriate when dealing with machines because the values with larger values would be disproportionately weighted or important, which is incorrect because all the values are essential for machine learning. It is also known as the Min-Max scaling, which transforms data into new values between (0,1) for effective data processing.

VI. PERFORMANCE MEASURES

In this research several performance measures are used to evaluate the classification quality of the selected machine learning models. To evaluate classification results we select Accuracy, Recall, Precision, and F-measure given that the dataset is imbalanced.

TABLE III. CONFUSION MATRIX

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

TABLE IV. CONFUSION MATRIX ELEMENTS

True Positive (TP)	Observation is positive and is predicted to be positive.
False Negative (FN)	Observation is positive but is predicted negative.
True Negative (TN)	Observation is negative and is predicted to be negative.
False Positive (FP)	Observation is negative but is predicted positive.

Table III illustrates the confusion matrix for a typical binary classification problem. It consists of four possible classification results: false negatives (FN), true negatives (TN), true positives (TP), and false positives (TP). Table IV explains the meaning of each type of result. Using the confusion matrix, we can calculate the performance scores based on their formula. The formula for each measure is explained subsequently.

1) Accuracy: it is the number of correct predictions made by the model over all kinds of predictions made.

$$Accuracy = \frac{TP + TN}{P + N}$$

2) Precision: it is the ratio between the true positives and all positives.

$$Precision = \frac{TP}{TP + FP}$$

3) Recall: it is the measure of our model correctly identifying true positives.

$$Recall = \frac{TP}{TP + FN}$$

4) F1 score: it is the harmonic mean of the precision and recall.

$$F_{1}\_score = \frac{2TP}{2TP + FP + FN}$$

VII. EXPERIMENTAL SETUP

Human health is directly affected by water quality, so it is vital to determine whether the water is safe or not. In this paper, we have presented five different machine learning models that can be employed to improve the quality of water which are Decision Tree, K-Nearest Neighbor, Support Vector Machine, Ransom Forest, and Light GBM. The processing of these models is conducted using two platforms, Jupyter Notebook with Python code, and H2O Driverless AI. As a first step, we remove incorrect values from the data and fix the imbalance of the class distribution. Then, we split the data into an 80% training set and a testing set of 20% by applying hold-out testing approach. Next, we tune the appropriate parameters of each model to run it and get performance results so that we can evaluate them. Interestingly, multiple experiment iterations are involved in each model, and the model with the most optimal performance is selected.

1. Decision Tree Model

The objective of this algorithm is to utilize the training input data (with class labels) in making decisions and creating rules based on the relationships between the attributes. We need to apply the entropy calculation along with the information gained to draw the tree and determine the root [9]. In this model, we tune the hyperparameters determining the “Gini index” criterion for calculating information gain, and “25” maximum depth which indicates how deep the tree can be. As a result, we got 93% accuracy and a 0.07 error rate. Also, this model produces an F1 measure of 82.20%, 83% precision, and 81% recall.

2. K-Nearest Neighbors

The K-nearest neighbors algorithm is a non-parametric supervised learning classifier, based on proximity, that makes predictions as to how an individual data point will be grouped. The algorithm can be used to solve either regression or classification problems, but more often than not, it is used as a classification algorithm. It is assumed that similar points are found near each other [10]. Hyperparameters tuned are “n_neighbors=7”, the “algorithm=auto” option selects “kd_tree” algorithm, which is in this case the slowest algorithm of those available. Therefore, running this model results in getting an 91 accuracy percent, a precision of 80 percent, recall of a 69% in the addition of the harmonic mean equal to 0.72.

3. Support Vector Machine

SVM algorithm is used to predict the value of each data item by using the coordinates of a number of points in an n-dimensional space, which n is the number of twenty features we have. Each feature's value is determined by the value of a particular coordinate. Following that, we find the hyperplane that can distinguish between the two classes very well [11]. In the hyperparameters, we set “Gamma” parameter defines the

distance between each training point and its influence while the cross-product of "C" values ranges in [1, 10]. Rrunning this model results in getting an 93 accuracy percent. In comparison, other quality measures such as precision, recall, and f-measures are between 89% and 75%.

4. Random Forest

An algorithm like Random Forest combines a number of decision trees on multiple subsets of a dataset and averages them to improve their predictive accuracy. As opposed than using just one decision tree, random forest takes the prediction from each tree and predicts the outcome based on the majority vote [12]. The reason behind selecting RF is that this algorithm works best with continuous data values associated with each class. Hyperparameters are used to minimize errors or predict targets as close as possible to the actual ones. In this algorithm. We set "n_estimators" to range from 10 to 250, which specifies the number of decision trees. Typically, a higher number of trees will lead to heightened accuracy at the expense of model size and training time. "Gini" criteria parameter is the choice of split algorithm trees. Accordingly, we obtained 95 percent accuracy and a 5% error rate. The range was 0.73 to 0.90 when we evaluated precision, recall, and harmonic mean.

5. LightGBM Model

LightGBM model helps us to use gradient boosting algorithms for ranking, classification, and a variety of other machine learning tasks in a fast, distributed, high-performance way [13]. The testing technique applied to this model is hold-out with a 60 percent training set and 40 percent for training as well. This model generates a high percentage of accuracy that is equal to 99.74 which is close to the percentages of recall and f-measures, while the precision measure is 100%. However, this model produces high percentages of evaluation scores and can be considered a valid model.

VIII. Experimental Results

Table V summarizes the experimental results on the water potability data using the five selected models: DT, KNN, SVM, RF, LightGBM. We can see that LighGBM model produced the highest accuracy value compared to the other classifiers. It achieves an accuracy of 99.74 percent, along with 98 percent of the harmonic mean. This model performs the best because we end up maximizing the true positive with the true negative as well.

TABLE V. EVALUATION CRITERIA FOR CLASSIFICATION ALGORITHM

	DT	K-NN	SVM	RF	Light GBM
Accuracy	93%	91%	93%	95%	99%
Precision	83%	80%	85%	93%	100%
Recall	81%	69%	79%	80%	97%
F1	82%	72%	82%	85%	98%

IX. Conclusion

This research evaluated four different machine learning models for classifying the potability of water, namely Decision Tree, K-Nearest Neighbor, Support Vector Machine, Ransom Forest, and LightGBM, along with evaluation methods for accuracy, precesion, recall, and F-Measure. In this regard, LightGBM had an accuracy of 99.74 percent, making it the most accurate outcome of the present experiment. For future research, we endorse using the LightGBM model with different water datasets to predict water quality, as we can ensure the power of the LightGBM model in predicting water potability effectively.

X. References

[1] A. Kadiwal, "Water Quality," [Online]. Available: https://www.kaggle.com/datasets/adityakadiwal/water-potability.

[2] MSSMARTYPANTS, "Water Quality," [Online]. Available: https://www.kaggle.com/datasets/mssmartypants/water-quality.

[3] R. Chafloque, "Predictive Neural Networks Model for Detection of Water Quality for Human Consumption," IEEE Xplore, 29 October 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9574673/authors#authors. [Accessed 7 June 2022].

[4] F. Li, "Study on Potability Water Quality Classification Based on Integrated Learning," IEEE Xplore, 18 April 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9755413/authors#authors. [Accessed 7 June 2022].

[5] M. I. K. Haq, "Classification of Water Potability Using Machine Learning Algorithms," IEEE Xplore, 27-29 October 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9689727/authors#authors. [Accessed 7 June 2022].

[6] U. Ahmed, R. Muntaz, H. Anwar, A. A.Shah, R. Irfan and J. García-Nieto , "Efficient Water Quality Prediction Using Supervised Machine Learning," 2019.

[7] S. S. Kurra, S. G. Naidu, S. Chowdala, S. C. Yellanki and D. B. E. Sunanda, "WATER QUALITY PREDICTION USING MACHINE LEARNING," International Research Journal of Modernization in Engineering Technology and Science, India, 2022.

[8] "Common Waterborne Contaminants," Water Quality Association, [Online]. Available: https://www.wqa.org/learn-about-water/common-contaminants.

[9] "Decision Tree Classification Algorithm," [Online]. Available: https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm. [Accessed 05 07 2022].

[10] "K-Nearest Neighbors Algorithm," IBM, [Online]. Available: https://www.ibm.com/topics/knn. [Accessed 28 20 2022].

[11] S. Ray, "Understanding Support Vector Machine(SVM) algorithm from examples (along with code)," *Analytics Vidhya,* 2017.

[12] Java T Point, [Online]. Available: https://www.javatpoint.com/machine-learning-random-forest-algorithm. [Accessed 14 10 2022].

[13] S. Shisingh, "LightGBM (Light Gradient Boosting Machine)," *Geeks for Geeks,* 2021.

[14] S. Asiri, "Machine Learning Classifiers," *Towards Data Science,* 2018.

[15] K. Nyuytiymbiy, "Hyperparameter tuning for machine learning models," *Towards Data Science,* 2020.

[16] S. Ray, "6 Easy Steps to Learn Naive Bayes Algorithm with codes in Python and R," 11 09 2017. [Online]. Available: https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/.

[17] [Online]. Available: https://weka.sourceforge.io/doc.dev/weka/classifiers/lazy/KStar.html. [Accessed 07 07 2022].

[18] A. CHEN, "What Is a Neural Network?," *Investopedia,* 2022.

[19] "Types of Neural Networks and Definition of Neural Network," Great Learning, 24 08 2022. [Online]. Available: https://www.mygreatlearning.com/blog/types-of-neural-networks/. [Accessed 1 10 2022].