Available online at http://scik.org

Commun. Math. Biol. Neurosci. 2026, 2026:22

https://doi.org/10.28919/cmbn/9757

ISSN: 2052-2541

WATER POTABILITY CLASSIFICATION USING MACHINE LEARNING: A CASE STUDY ON HANDLING INCOMPLETE DATA

HAYYUN LISDIANA[1], KARLI EKA SETIAWAN[2,*]

[1]Department of Chemistry Education, Faculty of Mathematics and Natural Science, Universitas Negeri Jakarta, Jakarta 13220, Indonesia

[2]Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta 11480, Indonesia

Abstract: Water is essential for the preservation of life on Earth, particularly for drinking purposes. Despite the abundance of water in the earth's ecosystem, the world is currently grappling with a significant global issue of contaminated water, a problem that extends beyond natural contamination and includes industrial wastewater. In this study, we aimed to investigate the potential of decision tree-based machine learning models, including decision trees, ensemble boosting, ensemble bagging, and random forests, in predicting water potability based on specific parameters. We used publicly available data from the "Water Quality and Potability" Kaggle dataset. Due to the high number of missing values for some parameters in the dataset, our research initially converted continuous, missing values into discrete or categorical values. We then filled these gaps with a general label, "unknown," instead of using mean or median values as other studies had done. Initially the result showed that the highest accuracy was random forest; our analysts showed that the sulfate parameters created confusion for the machine learning model due to the many missing values. So that, this research decided to exclude sulfate data from the dataset, and this research showed significant

*Corresponding author

E-mail address: karli.setiawan@binus.ac.id

Received December 27, 2025

results where all decision tree-based machine learning models can achieve 100% accuracy, precision, recall, and f1-score on evaluation using the test dataset.

Keywords: clean water and sanitation; machine learning; water potability; chemistry environment.

2020 AMS Subject Classification: 68T01, 68T10, 97P80.

1. INTRODUCTION

Water is essential for the preservation of life on Earth and is responsible for the integration of the earth's lands, oceans, and atmosphere. Even though the abundance of water on Earth is genuinely exceptional, its quality is becoming a significant global issue, as it directly impacts human health, ecological integrity, and overall environmental sustainability [1]. The growth of worldwide population around the world strongly escalates the demand for clean water for many purposes such as irrigation, domestic, household, energy, industrial needs, and foremost for drinking purpose [2]. The Sustainable Development Goals number 6 (SDG 6) outline Water, Sanitation, and Hygiene (WASH) as the primary ambitious agenda for Clean Water and Sanitation [3]. UNICEF asserts that WASH is crucial at medical facilities, educational institutions, and early childhood development centers. According to UNICEF, access to water and sanitation is also part of the core of human rights, especially for children, because the poor hygiene, open defecation, and lack of access to safe water and sanitation cause child mortality and morbidity. The COVID-19 pandemic has underscored the vital need of sanitation, hygiene, and sufficient access to clean water in disease prevention and containment [4]. Many families and organizations store their daily water needs in building storage tanks to address inadequate potable water delivery caused by frequent disruptions; however, these activities may significantly jeopardize the chemical and microbiological quality of the water [5]. This issue highlights the necessity to assess and oversee the quality of water for drinking purposes [6].

Currently, the machine learning (ML) approach is gaining significant attention due to its ability to identify patterns in data and then make predictions based on existing data. Applying this potential to predict water potability using existing water quality parameters and implementing it

in proactive water quality management is reasonable and promising. This research employs a machine learning approach, utilizing the "Water Quality and Potability" dataset, which represents data on potable water. We contribute to this research by first converting continuous data into categorical data and then impute the missing values with a general class category named "unknown class." Meanwhile, other research imputes missing values with median or mean values, making our approach distinct from others. We can assume that imputing the missing value using mean or median data may confuse the machine learning model, making it difficult for it to accurately distinguish between real data with values close to the mean or median and the imputation data. The ML model used in this research was decision tree, ensemble boosting, ensemble bagging, and random forest.

Water Quality Index (WQI) can be used to measure water quality containing multiple water-quality factors such as dissolved oxygen (DO), pH, biochemical oxygen demand (BOD), chemical oxygen demand (COD), total coliform bacteria, temperature, nitrogen, phosphorus, and others into a single measurement [7]. WQI is usually used for water pollution assessment, where the score between 0 and 25 is considered very bad, 25 and 50 is considered bad, 50 and 70 is considered medium, 70 and 90 is considered good, and 90 and 100 is considered excellent. WQI serves as an initial indicator in assessing general water quality. However, it is not specifically designed to determine water potability. To ensure the suitability of water for consumption, further testing based on official drinking water standards is required.

Given these limitations, recent studies have explored the use of machine learning (ML) approaches to predict drinking water safety more comprehensively by analyzing raw water quality data and learning complex patterns beyond conventional indices. Similar research has been conducted on predicting potable water quality using AI and ML approaches, using the same dataset as ours [8]. They encountered a similar issue to ours, which was the abundance of missing values in the dataset, particularly in the pH, sulfate, and trihalomethanes data sets. To address this issue, they employed three distinct techniques: removing the missing values, imputing the missing values using mean data, and imputing the missing values using support vector regression (SVR)

prediction. Their best prediction results were obtained from the utilization of SVR as missing values imputation and XGBoost as a binary classification model, resulting in 90.24% accuracy. In the same research using the same dataset as done by Rachid et al., they used two machine learning models: support vector machines (SVM) and random forests (RF) [9]. They used undersampling as a preprocessing treatment to balance the number of potable and non-potable samples. Their best result was the random forest model, achieving 70% accuracy, 72% precision, and 75% receiver operating characteristic area under the curve (ROC-AUC). Another research study using the same dataset was done by Patel et al, their research developed some machine learning models such as SVM, decision tree, random forest, gradient boost, and Ada Boost [10]. In their preprocessing, they performed data imputation for missing values using their mean value for each data and normalization implementation using Z-score. They also implemented an oversampling treatment using SMOTE to enhance the minority class. Their best result was random forest with 81% accuracy after hyperparameter tuning.

2. METHODOLOGY

The research methodology of this research is depicted as in Figure 1. Actually, this research initially was done in a single scenario as scenario A, but later after our analysis based on the result of scenario A. Our research found that the lack of a dataset containing too many missing values in important data resulted in a bad prediction. Intuitively, this research excluded the sulfate data, as explained in detail later in section 4 or the results and discussions section, becoming our scenario B. Therefore, this research was conducted in several steps, utilizing two scenarios, A and B. The initial phase of our research involved conducting exploratory data analysis and feature engineering on our dataset, as detailed in section 3.1, which also includes a section on the dataset and preprocessing. Then, we divided our research into two branches, one for the entire dataset and another for a subset of datasets that excluded sulfate data. The third step for each scenario was the search for the best hyperparameter setting for some machine learning models. Then the fourth step was the training and testing of ten machine learning models. Lastly, we evaluated the models using various parameters such as accuracy, precision, recall, and F1 score to determine which model was

the best.

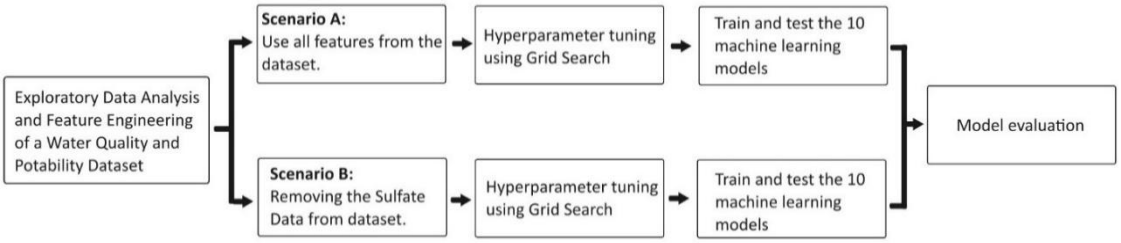

FIGURE 1. Research methodology

2.1. DATASET AND PREPROCESSING DATASET

This work utilized an open-access dataset named "Water Quality and Potability" from Kaggle. This dataset contains 3,276 data to determine if the water quality is potable or non-potable, consisting of nine variables as our features or inputs, such as pH, hardness, solids, chloramines, sulfate, conductivity, organic carbon, trihalomethanes, and turbidity. The distribution of used dataset can be seen in figure 2.

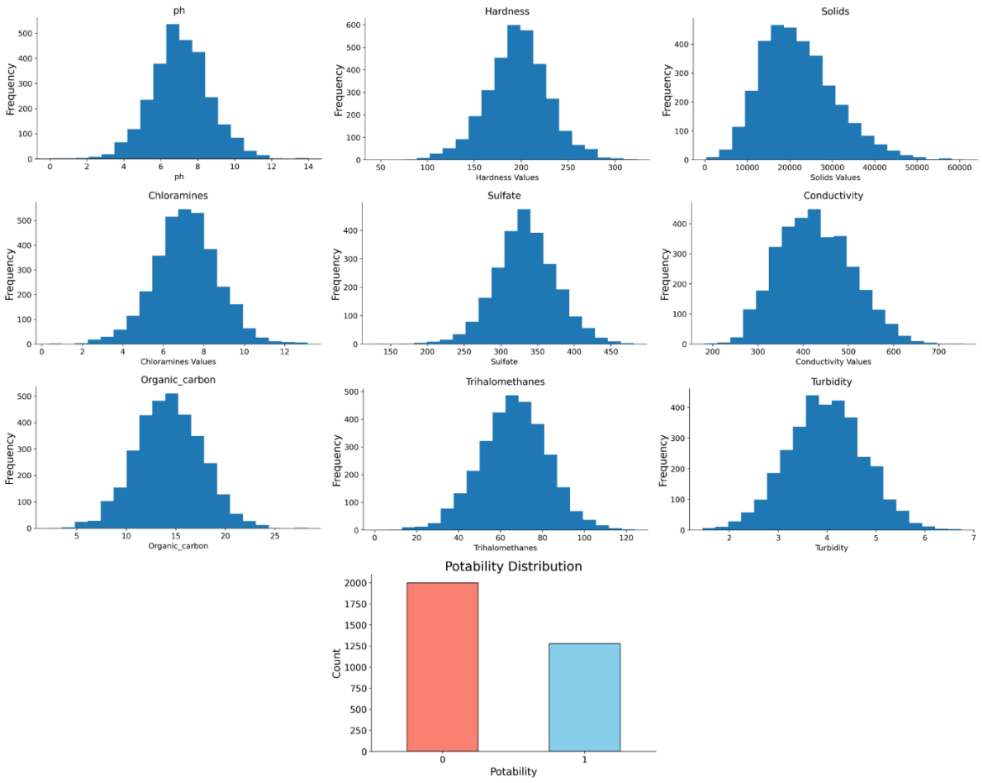

FIGURE 2. Dataset distribution

The pH variable denotes the water's pH level, a measurement of how acidic or basic water is, where the level is between 0 and 14 with 7 representing the neutral, less than 7 indicating the acidity, and more than 7 indicating the base. The hardness variable quantifies the mineral content in the water, which is mostly attributed to the presence of calcium and magnesium salts. The solid variables represent the total dissolved solids in the water, such as inorganic and organic materials, where these solids can produce unwanted taste and color changes in water appearance. The chloramines variable indicates the concentration of chloramines contaminated in the water system, which mostly comes from disinfectants. The sulfate variable specifies the concentration of sulfate, which naturally comes from minerals, soil, and rocks. The conductivity variable measures the electrical conductivity of the water, where the increase of ion concentration in water can enhance the electrical conductivity of water. The organic carbon variable indicates the organic carbon concentration in the water, where it comes from decaying natural organic matter. Trihalomethane variables quantify the quantity of trihalomethanes in water, which this chemical can be found in water treatment using chlorine. Turbidity assesses water clarity, which measures the light-emitting properties of water, and this measurement indicates the quality of waste discharge with respect to colloidal matter. Potability serves as a categorical variable indicating water suitability for consumption, with a value of 1 representing potable water and 0 denoting non-potable water.

The biggest problem in this dataset is the too many missing values where 491 data does not contain pH data, 781 data does not contain sulfate data, and 162 data does not contain trihalomethanes data. To handle missing value this research implemented discretization to those three variables and grouping all missing value into one class named unknown class as in Figure 3.

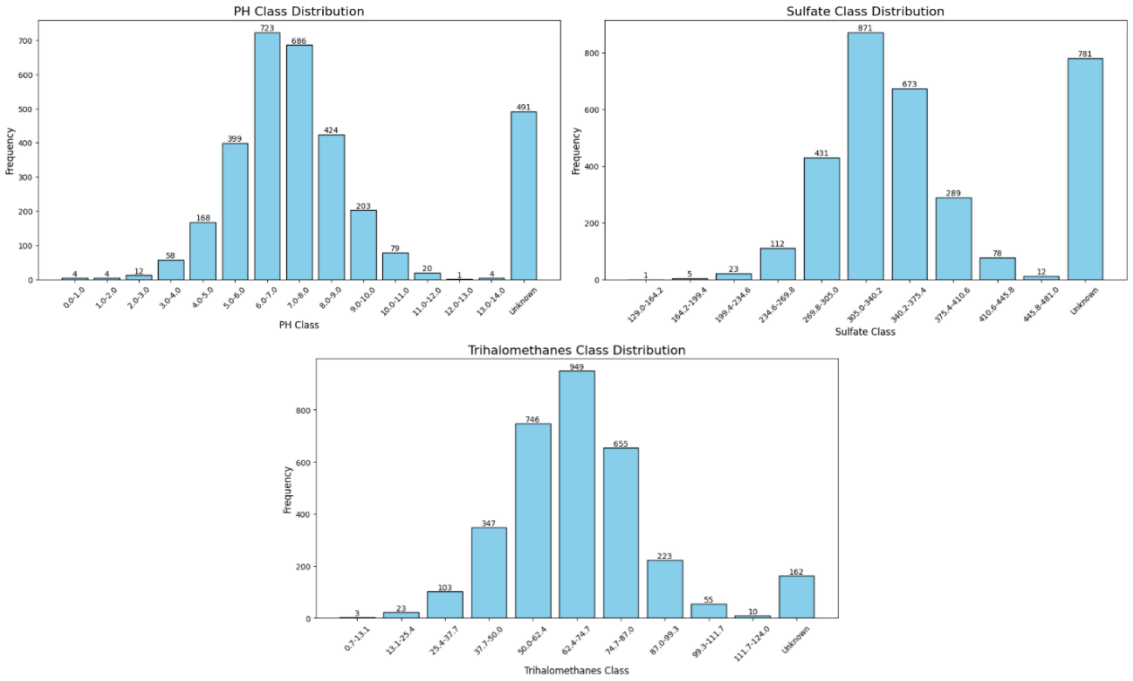

FIGURE 3. Discretization at pH, sulfate, and trihalomethanes data.

2.2. MACHINE LEARNING MODELS

This work investigated the capabilities of DT-based machine learning models, like DT, and some ensemble learning models, like boosting, bagging, and random forests, as illustrated in Figure 4 [11], [12]. All ML models developed in this experiment were built using the Scikit-learn library. All tree-based ML models are classified as nonparametric algorithms, as they do not impose stringent assumptions like mathematical models but rather rely on a series of decision trees [13] [14]. This feature enables greater flexibility in rules for models to predict outcomes based on complicated and intricate datasets, resulting in a computation-heavy process .

2.2.1. DECISION TREE

For classification and regression tasks, one can utilize Decision Tree (DT), a powerful nonparametric approach for finding patterns within data. This machine learning model does not require complex data preprocessing and enables efficient computation [15]. This model can effectively learn from the data by selecting optimal split points recursively, as illustrated in Figure 4. Adjusting hyperparameters allows this model to successfully resist noise and limit the danger of overfitting, resulting in strong generalization performance on small datasets. As shown in Figure

4, the DT contains various terminologies. These include root nodes, which symbolize the start of the DT model without any incoming edges; decision nodes, which represent the splitting nodes with one entering edge and exiting edges; leaf nodes, which represent the final decision result with exactly one incoming edge and zero outgoing edges; and pruning, which represents the action of deleting a sub-node. A decision tree works as follows: Begin with any attributes or inputs in our dataset, divide them by rule outcome (yes or false), and continue until the class percentage is homogenous and forms a leaf. Entropy and Gini can measure a node's homogeneous class.

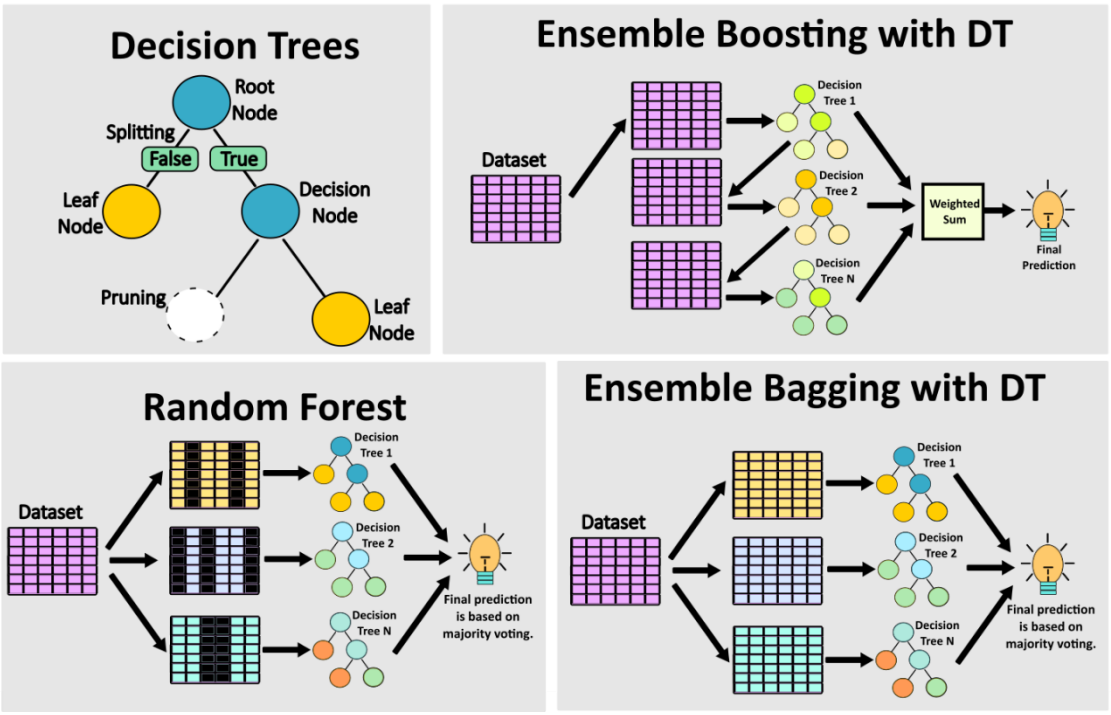

FIGURE 4. DT-based ML models applied in this research.

2.2.2. ENSEMBLE BAGGING

Bagging is one of the prominent ensembles learning approach in ML architecture which contain many similar independent base learners for example in this DT-based research used DT as the independent base learners as in Figure 4 [16], [17]. Every base learner in bagging architecture learns different subsets of training dataset illustrated in different color in Figure 4. Bagging provides the advantage of reducing variation, hence eliminating the issue of overfitting [18]. In addition, it performs well when applied to data with a large number of dimensions. Bagging has

various downsides, including a high computational cost, and loss of model interpretability, and considerable bias.

2.2.3. ENSEMBLE BOOSTING WITH ADABOOST AND GRADIENT BOOSTING CLASSIFIER

In the ensemble boosting learning algorithm, base learners are trained one step at a time, with each new model focusing on training data that was wrongly labeled to fix mistakes made in earlier models [17]. By weighting tough instances more in successive rounds, the algorithm improves its performance, resulting in a robust model. This robust model is created by combining all base learner predictions, sometimes using a weighted sum as illustrated in Figure 4. To represent the ensemble boosting, this research utilized AdaBoost and gradient boosting classifier using DT as the base learners.

Popular boosting algorithms include AdaBoost (Adaptive Boosting), built by many weak learners containing decision trees with one split, called a decision stump [19]. AdaBoost weights samples by raising the weight of hard-to-classify ones and reducing the weight of easy-to-classify ones until the algorithm discovers a model that classifies them correctly. Gradient tree boosting is an ensemble ML method for regression and classification. Gradient boosting stirs in the opposite direction to minimize a loss function [19]. AdaBoost uses high-weight data samples to identify "shortcomings" (difficult-to-classify samples), while gradient boosting uses gradients.

2.2.4. RANDOM FOREST

Random forest is comparable to ensemble learning with bagging, except it simulates decision tree ensemble learning by randomly selecting dataset features [16] [17]. Figure 4 illustrates how to apply only a subset of features to each base weak learner in this architecture, with unneeded features shown in black. Random Forest derives its name from the many decision trees and bootstrapping sample methods. Final forecasts, like decision tree bagging, are determined by a majority vote of all base learners.

2.3. EVALUATION METRICS

This work utilized some evaluation metrics, including accuracy, precision, recall, and F1-score, to test and compare our machine learning models [20]. Equations 1 through 4 delineate the

metrics. The accuracy measures the correctness of all categories for both positive and negative classifications. By calculating the number of correctly identified positive cases, the recall gauges the sensitivity. Precision assesses the degree of confidence in a model by evaluating its correctness. The F1-score is calculated by assessing the correspondence between precision and recall.

A true positive (TP) data point in the confusion matrix arises when the expected positive outcome aligns with the actual result [21]. A false positive (FP), or Type 1 error, is a data point in the confusion matrix where a positive outcome is anticipated, while the actual outcome is negative. A false negative (FN) in the confusion matrix denotes a scenario where a negative outcome is expected, while the actual outcome is positive. This scenario is categorized as a Type 2 error, which is as hazardous as a Type 1 error. In the confusion matrix, a data point is designated as a true negative (TN) when both the expected and actual outcomes are negative.

$$Accuracy = \frac{(TP\ + TN)}{(TP + TN + FN + FP)} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall\ \ \ \ = \frac{TP}{TP + FN} \tag{3}$$

$$F1\ Score\ \ = \frac{(2\ \times Precision\ \times Recall)}{(Precision + Recall)} \tag{4}$$

3. RESULTS AND DISCUSSION

This study explored the capabilities of various decision tree-based machine learning models, utilizing approximately ten different models to predict the potability of water. This study utilized a variety of decision tree-based machine learning models, such as Decision Tree (DT), ensemble learning with bagging, boosting using Adaboost and gradient boosting classifiers, and random forest. We set the previously mentioned DT-based ML models using their default hyperparameters.

Our research also conducted a search for optimal hyperparameter settings on various DT-based machine learning models, utilizing the grid search technique and the GridSearchCV library from Sklearn. The grid search methods yielded eight maximum tree depths for DT: 100 DT for ensemble bagging, 150 DT for ensemble bagging, 25 DT for ensemble Adaboost, and 15 maximum tree depths for random forest. Therefore, this research utilized 10 distinct DT-based machine learning models.

This research divided the dataset into two parts, with 80 percent data going into the training set and the remaining 20 percent data going into the testing set. We used the training set to train all machine learning models, aiming to achieve the best solutions that fit the dataset. Following our methodology, we initially train all 10 distinct DT-based machine learning models using a training set. Subsequently, we evaluate the models by exposing them to previously unseen testing data, and the evaluation results are presented in Table 2, and the model descriptions are explained in Table 1.

TABLE 1. Machine learning Model description.

Model Index	**Model Description**
Model 1	Decision Tree (default)
Model 2	Decision tree (max depth = 8)
Model 3	Ensemble Bagging with DT (default)
Model 4	Ensemble Bagging with DT (100 DT)
Model 5	Ensemble Bagging with DT (150 DT)
Model 6	Gradient Boosting Classifier
Model 7	Ensemble Adaboost with DT (default)
Model 8	Ensemble Adaboost with DT (25 DT)
Model 9	Random Forest (default)
Model 10	Random Forest (100 DT and max depth = 15)

Table 2 reveals that it was challenging for all DT-based machine learning models to attain an overall accuracy of over 60%. The total number in the overall testing dataset count on overall accuracy was 819 data, which is divided into two different classes with 505 data for class 0 meaning not potable class, and 314 for class 1 meaning the potable class. All the machine learning models under investigation appear to struggle with determining whether the water is potable or

class 1, as evidenced by their low recall scores. Some models, such as the gradient boosting classifier, modified DT with 8 max depth trees, default random forest, and modified random forest with 100 DT and 15 max depth trees, can reveal more than 85% of the non-potable water based on their recall score. If we focus on avoiding non-potable water, the gradient-boosting classifier may become the best model. So far from this evaluation result, we may conclude that the more missing data we have in our research, the more it can influence the low accuracy we have.

TABLE 2. Machine learning Model description.

Model Index	Overall Accuracy	Class 0 (Not Potable)			Class 1 (Potable)		
		Precision	Recall	f1 Score	Precision	Recall	f1 Score
Model 1	0.58	0.67	0.65	0.66	0.46	0.48	0.47
Model 2	0.64	0.66	0.86	0.74	0.55	0.27	0.37
Model 3	0.64	0.67	0.81	0.73	0.54	0.35	0.42
Model 4	0.64	0.67	0.82	0.74	0.55	0.35	0.42
Model 5	0.64	0.67	0.81	0.73	0.54	0.35	0.42
Model 6	0.64	0.65	0.9	0.76	0.58	0.21	0.31
Model 7	0.59	0.67	0.66	0.67	0.47	0.47	0.47
Model 8	0.58	0.66	0.65	0.66	0.45	0.46	0.46
Model 9	0.65	0.67	0.86	0.75	0.58	0.31	0.4
Model 10	0.65	0.67	0.86	0.75	0.58	0.31	0.4

TABLE 3. Machine learning experimental results after removing sulfate class data.

Models	Overall Accuracy	Class 0 (Not Potable)			Class 1 (Potable)		
		Precision	Recall	f1 Score	Precision	Recall	f1 Score
Model 1	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Model 2	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Model 3	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Model 4	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Model 5	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Model 6	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Model 7	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Model 8	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Model 9	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Model 10	1.00	1.00	1.00	1.00	1.00	1.00	1.00

The sulfate class had 781 data in the unknown class, representing approximately 23.84% of the data; the pH class had 491 data in the unknown class, representing approximately 14.99% of the data; and trihalomethanes had 162 data in the unknown class, representing 4.94% of the data. Based on that data situation we have, this research tried to perform the second training scenario by excluding the sulfate data due to the high number of unknown classes. Intuitively we think the sulfate content in the water can affect the potability of the water, but in the data, this research has this research mix the unknown value of sulfate content into a single class category, which is the unknown class, and the unknown class may contain the high number of sulfate content or the low number of sulfate content, which can confuse the models to fit the dataset. Table 3 displays the outcome of our second experiment scenario, which involved removing the sulfate data.

Table 3 shows that all DT-based machine learning models can perfectly fit the data after excluding the sulfate data, achieving perfect 100% overall accuracy and precision, recall, and f1-score in both non-potable and potable water classes. The total number in the overall testing dataset count on table 3 were same on table 2 where on overall accuracy was 819 data, which is divided into two different classes with 505 data for class 0 meaning not potable class, and 314 for class 1 meaning the potable class. The absence of sulfate data can significantly increase the accuracy of all tested machine learning models. This demonstrates how the incompleteness of the dataset impacts the supervised learning task in binary classification, leading to lower performance.

We conducted this research in response to disagreements with previous studies that handled the missing values of pH, sulfate, and trihalomethanes by filling them with mean, median, and mode data. That action may bias the missing value with the real data with the real value near the mean, median, or mode value. Our approach involves separating the missing value into a new unknown class, then excluding the sulfate class due to its influence on determining potable and non-potable water. This approach can significantly improve the supervised learning task in binary classification, achieving perfect accuracy of 100% in all DT-based ML models.

4. CONCLUSIONS

Using the "Water Quality and Potability" dataset, this study showed that our machine learning method, which includes decision tree-based models like decision trees and some ensemble models with bagging, boosting, and random forests, could accurately predict whether water was safe to drink. This research initially included all parameters in predictive models, resulting in unsatisfactory results, and our analysis indicated that the abundance of missing values in sulfate data can create confusion for predictive models in finding the pattern of potable and unpotable water. Based on our analysis, this study did not include sulfate data. All of the predictive models that were made were 100% accurate on the overall class and had perfect precision, recall, and f1-scores on each class.

For future research, this research can be continued with cleaner datasets instead of the dataset with too many missing values, because based on our analysis in this research, too many missing values can make a confusion to our predictive model. The capabilities of deep learning models can be explored further for better understanding and results in the future.

CREDIT AUTHORSHIP CONTRIBUTION STATEMENT

Hayyun Lisdiana: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Resources, writing – Original Draft, Writing - Review & Editing, Project Administration, and Funding acquisition. Karli Eka Setiawan: Conceptualization, Methodology, Software, Validation, Formal analysis, Data Curation, Visualization, Supervision

DATA AVAILABILITY

This machine learning study on water potability utilized a publicly accessible dataset named “Water Quality and Potability,” available at the link https://www.kaggle.com/datasets/uom190346a/water-quality-and-potability.

FUNDING

The Article Processing Charge (APC) for this publication was funded by Universitas Negeri Jakarta.

CONFLICT OF INTERESTS

The authors declare that there is no conflict of interests.

REFERENCES

[1] S.K. Dewangan, D.N. Toppo, A. Kujur, Investigating the Impact of PH Levels on Water Quality: An Experimental Approach, Int. J. Res. Appl. Sci. Eng. Technol. 11 (2023), 756-759. https://doi.org/10.22214/ijraset.2023.55733.

[2] M.T.H. van Vliet, E.R. Jones, M. Flörke, W.H.P. Franssen, N. Hanasaki, et al., Global Water Scarcity Including Surface Water Quality and Expansions of Clean Water Technologies, Environ. Res. Lett. 16 (2021), 024020. https://doi.org/10.1088/1748-9326/abbfc3.

[3] UNICEF, Strategy for Water, Sanitation and Hygiene 2016-2030, 2016. https://www.unicef.org/documents/unicef-strategy-water-sanitation-and-hygiene-2016-2030.

[4] Y. Zhang, J. Deng, B. Qin, G. Zhu, Y. Zhang, et al., Importance and Vulnerability of Lakes and Reservoirs Supporting Drinking Water in China, Fundam. Res. 3 (2023), 265-273. https://doi.org/10.1016/j.fmre.2022.01.035.

[5] M. Salehi, Global Water Shortage and Potable Water Safety; Today's Concern and Tomorrow's Crisis, Environ. Int. 158 (2022), 106936. https://doi.org/10.1016/j.envint.2021.106936.

[6] A. du Plessis, Persistent Degradation: Global Water Quality Challenges and Required Actions, One Earth 5 (2022), 129-131. https://doi.org/10.1016/j.oneear.2022.01.005.

[7] S.P. Gorde, M.V. Jadhav, Assessment of Water Quality Parameters: A Review, Int. J. Eng. Res. Appl. 3 (2013), 2029-2035.

[8] M. Yurtsever, M. Emeç, Potable Water Quality Prediction Using Artificial Intelligence and Machine Learning Algorithms for Better Sustainability, Ege Acad. Rev. 23 (2023), 265-278. https://doi.org/10.21121/eab.1252167.

[9] E. Rachid, S. Abderrahim, A. Hafid, R. Souad, Predicting Water Potability Using a Machine Learning Approach, Environ. Challenges 19 (2025), 101131. https://doi.org/10.1016/j.envc.2025.101131.

[10] J. Patel, C. Amipara, T.A. Ahanger, K. Ladhva, R.K. Gupta, et al., A Machine Learning-Based Water Potability Prediction Model by Using Synthetic Minority Oversampling Technique and Explainable AI, Comput. Intell. Neurosci. 2022 (2022), 9283293. https://doi.org/10.1155/2022/9283293.

[11] K.E. Setiawan, A. Kurniawan, S.Y. Prasetyo, Comparative Analysis of Machine Learning Decision Tree-Based Models for Predicting Maternal Health Risks, Procedia Comput. Sci. 245 (2024), 57-64. https://doi.org/10.1016/j.procs.2024.10.229.

[12] K.E. Setiawan, H. Lisdiana, Enhancing Breast Cancer Detection Using Machine Learning on Data from Cuban Women, Commun. Math. Biol. Neurosci. 2025 (2025), 96. https://doi.org/10.28919/cmbn/9392.

[13] B.D. Williamson, P.B. Gilbert, M. Carone, N. Simon, Nonparametric Variable Importance Assessment Using Machine Learning Techniques, Biometrics 77 (2020), 9-22. https://doi.org/10.1111/biom.13392.

[14] C. Yu, Y. Lin, C. Lin, S. Wang, S. Lin, et al., Predicting Metabolic Syndrome with Machine Learning Models Using a Decision Tree Algorithm: Retrospective Cohort Study, JMIR Med. Inform. 8 (2020), e17110. https://doi.org/10.2196/17110.

[15] Y. Chen, M. Khandelwal, M. Onifade, J. Zhou, A. Ismail Lawal, et al., Predicting the Hardgrove Grindability Index Using Interpretable Decision Tree-Based Machine Learning Models, Fuel 384 (2025), 133953. https://doi.org/10.1016/j.fuel.2024.133953.

[16] G. Ngo, R. Beard, R. Chandra, Evolutionary Bagging for Ensemble Learning, Neurocomputing 510 (2022), 1-14. https://doi.org/10.1016/j.neucom.2022.08.055.

[17] S. González, S. García, J. Del Ser, L. Rokach, F. Herrera, A Practical Tutorial on Bagging and Boosting Based Ensembles for Machine Learning: Algorithms, Software Tools, Performance Study, Practical Perspectives and Opportunities, Inf. Fusion 64 (2020), 205-237. https://doi.org/10.1016/j.inffus.2020.07.007.

[18] A. Mohammed, R. Kora, A Comprehensive Review on Ensemble Deep Learning: Opportunities and Challenges, J. King Saud Univ. - Comput. Inf. Sci. 35 (2023), 757-774. https://doi.org/10.1016/j.jksuci.2023.01.014.

[19] P. Bahad, P. Saxena, Study of AdaBoost and Gradient Boosting Algorithms for Predictive Analytics, in: G. Singh Tomar, N.S. Chaudhari, J.L.V. Barbosa, M.K. Aghwariya (Eds.), International Conference on Intelligent Computing and Smart Communication 2019, Springer, Singapore, 2020: pp. 235–244. https://doi.org/10.1007/978-981-15-0633-8_22.

[20] K.E. Setiawan, Predicting Recurrence in Differentiated Thyroid Cancer: A Comparative Analysis of Various Machine Learning Models Including Ensemble Methods with Chi-Squared Feature Selection, Commun. Math. Biol. Neurosci., 2024 (2024), 55. https://doi.org/10.28919/cmbn/8506.

[21] Ž.Đ. Vujovic, Classification Model Evaluation Metrics, Int. J. Adv. Comput. Sci. Appl. 12 (2021), 599-606. https://doi.org/10.14569/IJACSA.2021.0120670.