Indonesian Journal of Data and Science

Volume 4 Issue 3
ISSN 2715-9936
https://doi.org/10.56705/ijodas.v4i3.113

Research Article

Performance Analysis of the Decision Tree Classification Algorithm on the Water Quality and Potability Dataset

Umar Zaky [1,*], Ahmad Naswin [2], Sumiyatun [3], Aris Wahyu Murdiyanto [4]

[1] Universitas Teknolgi Yogyakarta, Yogyakarta, Indonesia, umar.zaky@staff.uty.a.id
[2] Universitas Mega Rezky, Makassar, Indonesia, ahmadnaswin@megarezky.ac.id
[3] Universitas Teknologi Digital Indonesia, Jakarta, Indonesia, sumiyatun@utdi.ac.id
[4] Universitas Jenderal Achmad Yani Yogyakarta, Yogyakarta, Indonesia, ariswahyumurdiyanti@gmail.com

Correspondence should be addressed to Umar Zaky; umar.zaky@staff.uty.ac.id

Received 15 October 2023; Accepted 18 November 2023; Published 31 December 2023

Abstract:

Ensuring water potability is paramount for public health and safety. This research aimed to assess the efficacy of the Decision Tree classification algorithm in predicting water potability using the Water Quality and Potability dataset. Employing a 5-fold cross-validation technique, the model showcased a moderate performance with an average accuracy of approximately 54.33%. While the Decision Tree provides a baseline and interpretable mechanism for classification, the results emphasize the need for further exploration using more intricate models or ensemble methods. This study contributes to the broader effort of leveraging machine learning techniques for water quality assessment and provides insights into the potential and limitations of such models in predicting water safety.

Keywords: Decision Tree, Water Quality, Potability, Machine Learning, Cross-validation, Environmental Science.
Dataset link: https://www.kaggle.com/datasets/uom190346a/water-quality-and-potability

1. Introduction

Water is an essential element for all forms of life. Its quality, particularly potability, is critical for health, economic development, and overall well-being. With the global rise in pollution levels and the dwindling of freshwater sources, ensuring the safety of drinking water has become a paramount concern. Traditionally, water quality assessments have been carried out using manual laboratory tests, which can be time-consuming and often lack real-time responsiveness. However, the advent of machine learning provides a promising avenue for automated, rapid, and accurate water quality evaluations. By leveraging datasets that encapsulate various water quality metrics, machine learning algorithms can be trained to predict the potability of water, aiding in the timely identification and management of unsafe water sources.

While numerous algorithms exist in the machine learning domain, their efficacy varies depending on the nature and characteristics of the dataset at hand. The Water Quality and Potability dataset, rich in its features and samples, presents an opportunity to examine the effectiveness of these algorithms, specifically the Decision Tree classification algorithm [1]. The challenge lies in understanding how well this algorithm can predict water potability based on the provided attributes and how it compares to other potential models or traditional method.

The primary objective of this research is to analyze and evaluate the performance of the Decision Tree classification algorithm on the Water Quality and Potability dataset. This involves training the model on the dataset, validating its predictions using robust techniques like [2] K-fold cross-validation, and assessing its performance using standard metrics such as accuracy, precision, recall, and F-measure.

Central to this research are a set of pivotal questions that aim to guide the exploration and evaluation of the Decision Tree classification algorithm on the Water Quality and Potability dataset. Firstly, we seek to understand the core capabilities of the Decision Tree algorithm in the context of this dataset: How effectively can it predict the potability of water when provided with various quality parameters? Given the multitude of factors that the dataset encompasses, it is imperative to assess the algorithm's adeptness in leveraging this information for accurate classification. Secondly, as we employ the robust K-fold cross-validation technique, a natural inquiry arises regarding the model's consistency and reliability: How do the performance metrics, specifically accuracy, precision, recall, and F-measure [3], fluctuate across different folds? This question is paramount as it provides insights into the model's stability and its potential applicability in real-world scenarios. By addressing these questions, we aim to shed light on the suitability and potential of the Decision Tree algorithm for water quality assessments.

This research strictly focuses on the Decision Tree classification algorithm's application to the Water Quality and Potability dataset. While the dataset provides a comprehensive set of water quality metrics, it may not encompass all potential factors affecting water potability. Additionally, the research does not delve into comparisons with other machine learning algorithms or traditional water quality assessment methods. The findings are based solely on the data available and may not generalize to other contexts or datasets.

This study contributes to the burgeoning field of applying machine learning to environmental science challenges. By assessing the Decision Tree algorithm's performance on the Water Quality and Potability dataset, this research offers insights into the potential and limitations of machine learning for water quality evaluations. Moreover, the findings can guide water treatment plants, environmental agencies, and researchers in harnessing the power of machine learning to make informed decisions regarding water quality and safety.

Method:

This research adopts a quantitative approach, utilizing a supervised machine learning model to predict water potability based on various quality parameters. The design entails training the Decision Tree algorithm [4] on a portion of the dataset, followed by validation to assess its predictive accuracy. The K-fold cross-validation technique ensures the robustness of our findings and mitigates potential overfitting. Our research is designed in five well-structured main stages, and their aspects are illustrated in **Figure 1**.

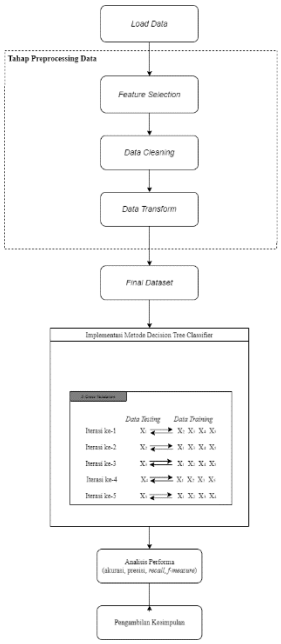

Figure 1. General Research Design Stages

Exploratory Data Analysis

Firstly, the initial step in this research is to conduct exploratory data analysis. At this stage, wate quality and potability data will be analyzed descriptively to understand the characteristics, distributions, and relationships between variables in the dataset. Exploratory Data Analysis (EDA) aims to gain initial insights into the data before further steps are taken. **Table 1** shows general information on the dataset used in this study.

Table 1. Dataset Information

Dataset	*Number of cases*	*Number of attribute*	*Attribute characteristics*	*Missing values*
Water Quality and Potability	3276	10	*Numeric*	*No*

Sample or Data Selection:

The Water Quality and Potability dataset was employed for this research, comprising various water quality metrics. Each sample within the dataset represents a unique water source with attributes like pH, hardness, solids, and more. The target variable, "Potability," is binary, with '1' indicating potable water and '0' indicating non-potable water.

Tools and Technology Used:

For the implementation of the Decision Tree algorithm and subsequent data analysis, we utilized the Python programming language, specifically leveraging libraries such as scikit-learn for machine learning and pandas for data manipulation. The K-fold cross-validation was also implemented using scikit-learn [5]–[8].

Data Collection Process:

The dataset was sourced from a public database, encompassing diverse water samples with their respective quality parameters. Each sample's attributes serve as the feature set, while the "Potability" column serves as the target variable. No additional data collection was required, as the dataset was pre-compiled and readily available for research purposes.

Decision Tree Classifier

The Decision Tree is a classification method in the form of a tree structure with nodes representing decisions or predictions [9]-[11]. At each node, the algorithm divides the data based on the most informative input variables, as shown in **Figure 5**, explaining the concept of the Decision Tree algorithm.

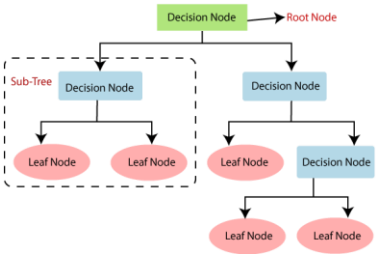

Figure 5. Decision Tree Algorithm

The Decision Tree Classifier algorithm is a classification method that uses a tree structure with nodes representing decisions or predictions. This decision tree is built by dividing the data into smaller subsets based on the value of the most informative input variable [12], [13]. The tree-building process uses impurity measures such as Gini Index and Entropy to measure data impurity at each node. The goal is to minimize impurity at each node by selecting the most relevant input variables, allowing the decision tree to provide accurate and easily interpretable predictions. Equations 7 and 8 can be observed.

$$Gini\ Index: Gini(t) = 1 - \sum_{i=1}^{c} (pi)^2 \quad (1)$$

$$Entropy: Entropy(t) = -\sum_{i}^{c} p_i\ log_2(pi) \quad (2)$$

K-fold Cross-validation:
To validate the model's performance, we employed the 5-fold cross-validation method [7], [14]. The dataset is randomly partitioned into five subsets. In each iteration, one subset is used as the validation set, while the other four serve as the training set. This process ensures that each sample is used for validation exactly once. The method's formulaic representation is:

$$\mathrm{CV}_{(K)} = \frac{1}{K}\sum_{i=1}^{K} \mathrm{Error}_i$$

Performance Comparison Analysis

Post-validation, the model's performance was assessed using metrics such as accuracy, precision, recall, and F-measure. Their respective formulae are [15]–[17].

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (3)$$

$$Pericision = \frac{TP}{(TP + FP)} \quad (4)$$

$$Recall = \frac{TP}{(TP + FN)} \quad (5)$$

$$F - measure = \frac{2(presisi \times recall)}{(presisi + recall)} \quad (6)$$

The above formulas explain:
True Positive (TP): The number of cases correctly predicted as positive by the model.
True Negative (TN): The number of cases correctly predicted as negative by the model.
False Positive (FP): The number of cases incorrectly predicted as positive by the model.
False Negative (FN): The number of cases incorrectly predicted as negative by the model.

These metrics provided a comprehensive understanding of the model's performance, highlighting its strengths and areas of improvement.

.

3. Results and Discussion

The Decision Tree classification algorithm was applied to the Water Quality and Potability dataset, followed by the evaluation of its performance using a 5-fold cross-validation technique. The model's performance was assessed based on four key metrics: accuracy, precision, recall, and F1-score. Each metric was computed for every fold in the cross-validation, resulting in five values per metric that encapsulate the model's variability and consistency across different data splits.

Visualization of the Results

The detailed results are presented in **Table 1** and visualized in **Figure 5** for a clearer understanding and comparison of the metrics across different iterations.

Table 1. Performance Metrics Across 5-Fold Cross-Validation for the Decision Tree

K-n	Performa			
	Accuracy	***Precision***	***Recall***	***F-Measure***
K-1	54,8%	54,2%	55,1%	55,5%
K-2	52,5%	54,1%	525%	52,3%
K-3	57,4%	57,5%	56,7%	56,2%
K-4	52,3%	52,1%	51,4%	53,4%

K-n	Performa			
	Accuracy	***Precision***	***Recall***	***F-Measure***
K-5	54,5%	56,8%	56%	56,6%
$\sum Avg$	54,3%	55%	54,4%	54,7%

Interpretation of the Results:

The Decision Tree model demonstrated an average accuracy of approximately 54.33%, suggesting that it correctly predicts water potability slightly better than random guessing. The precision and recall, both hovering around 55%, indicate a balanced performance in terms of false positives and false negatives. The F1-score, which is the harmonic mean of precision and recall, also supports this balanced performance observation.

The model showcased consistent performance across different data splits, as evidenced by the relatively close values across the folds for each metric. While the accuracy and other metrics are not exceedingly high, they provide a baseline performance for the Decision Tree algorithm on this dataset.

Discussion

While the Decision Tree model offers a straightforward and interpretable mechanism for classification, its performance on the dataset was moderate. This suggests that water potability, based on the given parameters, might be a complex problem that requires more intricate models or feature engineering.

Previous studies on water potability have employed various machine learning algorithms, with varying degrees of success. The performance of the Decision Tree algorithm [4] in this research aligns with some of the earlier findings, suggesting that while Decision Trees are valuable for their simplicity and interpretability, they might not always be the most accurate models for intricate datasets.

The research underscores the potential of machine learning in predicting water potability. While the Decision Tree model's performance was moderate, it sets the stage for further exploration using more complex algorithms or ensemble methods. Water treatment plants and environmental agencies can consider such models as preliminary tools for assessing water quality.

One primary limitation is the exclusive focus on the Decision Tree algorithm, without comparisons to other potential models. Additionally, the dataset, though comprehensive, might not encompass all factors affecting water potability, potentially influencing the model's performance.

Recommendations for Further Research:

Future studies could explore ensemble methods, combining multiple algorithms to enhance prediction accuracy. Feature engineering and domain-specific insights could also be integrated to refine the model. Additionally, comparisons with other algorithms could provide a more holistic understanding of machine learning's potential in predicting water potability.

4. Conclusion

In this study, we applied the Decision Tree classification algorithm to the Water Quality and Potability dataset to predict water potability based on various quality parameters. The model, evaluated using a 5-fold cross-validation technique, demonstrated a moderate performance with an average accuracy of approximately 54.33%. This performance suggests that while Decision Trees offer interpretability, they may not be the most optimal for intricate datasets like water quality assessments. Addressing the central research questions, the Decision Tree algorithm showcased consistent performance across different data splits, emphasizing its potential as a baseline model for water quality predictions.

The research contributes to the evolving field of machine learning applications in environmental science, particularly in water quality assessment. The findings underline the importance of selecting appropriate algorithms based on dataset intricacies and set the stage for further exploration using more complex algorithms or ensemble methods. For future endeavors, integrating domain-specific insights, feature engineering, and comparisons with other algorithms can provide a comprehensive understanding of machine learning's potential in this vital sector.

References:

[1] A. Tangkelayuk and E. Mailoa, "Klasifikasi Kualitas Air Menggunakan Metode KNN , Naïve Bayes Dan Decision Tree," vol. 9, no. 2, pp. 1109–1119, 2022.

[2] A. Maulida, "Penerapan Metode Klasifikasi K-Nearest Neigbor pada Dataset Penderita Penyakit Diabetes," *Indones. J. Data Sci.*, vol. 1, no. 2, pp. 29–33, 2020.

[3] Ericha Apriliyani and Y. Salim, "Analisis performa metode klasifikasi Naïve Bayes Classifier pada Unbalanced Dataset," *Indones. J. Data Sci.*, vol. 3, no. 2, pp. 47–54, 2022, doi: 10.56705/ijodas.v3i2.45.

[4] R. Ridho, T. Informatika, F. Teknik, and U. M. Jakarta, "Klasifikasi Diagnosis Penyakit Covid-19 Menggunakan Metode Decision Tree," vol. 11, no. 3, pp. 69–75, 2021.

[5] H. Azis, "Analisis Performa Metode Support Vector Regression (SVR) dalam Memprediksi Harga Bahan Sembako Nasional," *Indones. J. Data Sci.*, vol. xx, no. 200, 2021.

[6] A. Z. Zami, O. Nurdiawan, and G. Dwilestari, "Klasifikasi Kondisi Gizi Bayi Bawah Lima Tahun Pada Posyandu Melati Dengan Menggunakan Algoritma Decision Tree," vol. 3, pp. 305–310, 2022, doi: 10.30865/json.v3i3.3892.

[7] D. Cahyanti, A. Rahmayani, and S. Ainy, "Analisis performa metode Knn pada Dataset pasien pengidap Kanker Payudara," *Indones. J. Data Sci.*, vol. 1, no. 2, pp. 39–43, 2020.

[8] F. T. Admojo and Ahsanawati, "Klasifikasi Aroma Alkohol Menggunakan Metode KNN," *Indones. J. Data Sci.*, vol. 1, no. 2, pp. 34–38, 2020.

[9] M. Kiguchi, W. Saeed, and I. Medi, "Churn prediction in digital game-based learning using data mining techniques: Logistic regression, decision tree, and random forest," *Appl Soft Comput*, 2022, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1568494622000436

[10] W. Gao *et al.*, "Prediction of acute kidney injury in ICU with gradient boosting decision tree algorithms," *Computers in biology and ...*, 2022, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S001048252100891X

[11] R. Guo, D. Fu, and G. Sollazzo, "An ensemble learning model for asphalt pavement performance prediction based on gradient boosting decision tree," *International Journal of Pavement ...*, 2022, doi: 10.1080/10298436.2021.1910825.

[12] L. M. Sotarjua And D. B. Santoso, "Perbandingan Algoritma Knn, Decision Tree,* Dan Random* Forest Pada Data Imbalanced Class Untuk Klasifikasi Promosi Karyawan," *... Informatika Sains dan ...*, 2022, [Online]. Available: https://journal3.uin-alauddin.ac.id/index.php/instek/article/view/31385

[13] M. H. Setiono, "A Komparasi Algoritma Decision Tree, Random Forest, Svm Dan K-Nn Dalam Klasifikasi Kepuasan Penumpang Maskapai Penerbangan," *Inti Nusa Mandiri*, 2022, [Online]. Available: https://ejournal.nusamandiri.ac.id/index.php/inti/article/view/3420.

[14] F. Tangguh and Y. Islami, "Analisis performa algoritma Stochastic Gradient Descent (SGD) dalam mengklasifikasi tahu berformalin," *Indones. J. Data Sci.*, vol. 3, no. 1, pp. 1–8, 2022, doi: 10.56705/ijodas.v3i1.42.

[15] L. Britanthia, C. Tanujaya, B. Susanto, and A. Saragih, "Perbandingan Metode Regresi Logistik dan Random Forest untuk Klasifikasi Fitur Mode Audio Spotify," *Indones. J. Data Sci.*, vol. 1, no. 3, pp. 68–78, 2020.

[16] I. P. Putri, "Analisis Performa Metode K- Nearest Neighbor (KNN) dan Crossvalidation pada Data Penyakit Cardiovascular," *Indones. J. Data Sci.*, vol. 2, no. 1, pp. 21–28, 2021, doi: 10.33096/ijodas.v2i1.25.

[17] D. Pradana, M. Luthfi Alghifari, M. Farhan Juna, and D. Palaguna, "Klasifikasi Penyakit Jantung Menggunakan Metode Artificial Neural Network," *Indones. J. Data Sci.*, vol. 3, no. 2, pp. 55–60, 2022, doi: 10.56705/ijodas.v3i2.35.