JEEMECS (Journal of Electrical Engineering, Mechatronic and Computer Science
Vol. 8, No. 1, February 2025, pp. 17-30 ISSN 2614-4859

Enhancing Water Quality Prediction using the Decision Tree Approach with the C4.5 Algorithm

Anda Iviana Juniani [a,1*], Anggara Trisna Nugraha[b,2], Putri Amelia[c,3], Zukhruf Zidane Handandi[d,4], Elmi Hidayana[d,5]

[a] Engineering Design and Manufacture Study Program, Department of Mechanical Engineering, Politeknik Perkapalan Negeri Surabaya, Indonesia
[b] Marine Electrical Engineering Study Program, Department of Marine Electrical Engineering, Politeknik Perkapalan Negeri Surabaya, Indonesia
[c] Information Systems Study Program, Universitas Internasional Semen Indonesia, Indonesia
[d] Automation Engineering Study Program, Department of Marine Electrical Engineering, Politeknik Perkapalan Negeri Surabaya, Indonesia

[1]andaiviana@ppns.ac.id*,[2] anggaranugraha@ppns.ac.id [3] putri.amelia@uisi.ac.id, [4]zzidane27@student.ppns.ac.id, [5]elmihidayana28@student.ppns.ac.id

ABSTRACT

Keywords
Water quality
Decision Tree
C4.5 Algorithm
Prediction

Water quality is essential for safeguarding both public health and the environment. This study aims to develop a predictive model for assessing water quality using the Dtree (Decision Tree) method with the C4.5 algorithm. The research involves analyzing water samples from different sites around Sidomulyo Village, focusing on key parameters such as pH, Total Dissolved Solids (TDS), and turbidity. The objective is to create a model that categorizes water samples according to class II water quality standards. The research process includes data collection, initial data preparation, model development with the C4.5 algorithm, and performance evaluation. The results reveal that the dtree model achieved a high accuracy rate of 95.65% for water quality prediction. The confusion matrix analysis demonstrated a precision of 92.31% for predicting class II standards and 100% for identifying samples that did not meet these standards. These findings underscore the effectiveness of the C4.5 algorithm in evaluating water quality. The model offers valuable insights for water resource managers and policymakers, aiding in improved water quality monitoring and management. This research makes a significant contribution to enhancing decision support systems for more effective water quality management

1. Introduction

Water is an essential component of daily life, integral to activities such as drinking, bathing and washing. To ensure that water is safe and effective for such uses, it must meet strict quality standards covering various aspects, including turbidity, solutes, taste, and odor. Turbidity, in particular, is an important indicator of water quality, as it reflects the presence of suspended particles that can cause water to appear cloudy and discolored. High levels of turbidity are often a warning sign that water may be contaminated and therefore unfit for consumption or use. [1].

In the area surrounding Sidomulyo Village, Lamongan, improper waste disposal practices have become a significant problem. Many residents and visitors dump garbage into the drainage systems and ditches that run through the area. This not only clogs these drains but also introduces various pollutants into the water supply. The accumulation of trash, including plastic waste, organic matter, and other debris, significantly contributes to the turbidity of the water, exacerbating the risk of contamination and making it increasingly difficult to meet the required quality standards.

In addition, water quality in the sidomulyo village reservoir is also affected by the surrounding fish farming activities. Aquaculture practices in the region have introduced large amounts of organic matter and nutrients into the water, leading to distinct changes in its appearance. The water often turns cloudy or discolored, indicating that it has been mixed with waste products from fish farms. This not only changes the aesthetic quality of the water, but also poses potential health risks due to high levels of turbidity and the presence of harmful substances.

Humans have both primary and secondary needs essential for their daily lives. Among these, the availability of clean water is crucial. Clean water is necessary for activities such as washing, bathing, and cleaning homes and gardens. To meet this need, existing water resources can be utilized. One common source of such water is river water [2]. Water is a surface water source that can be used for various purposes. However, in many parts of Indonesia, the quality of river water often fails to meet clean water standards, particularly concerning turbidity levels. In simple terms, river water in Indonesia generally does not comply with the established clean water quality standards [3]. High turbidity indicates that river water is polluted. This pollution is often a result of human activities in the surrounding environment. These pollutants can significantly impact water quality and pose health risks to humans. Contaminants such as sediment, salt, and toxic metals are often found in polluted water.[2]. Thus, measures must be taken to treat river water to ensure it is safe for use.

Dtree (Decision Tree) is a model represented visually that aims to simplify the decision-making process in a rational manner. This visualization facilitates a structured and step-by-step approach to making rational decisions. Decision making involves selecting from various presented alternatives [4]. The decision tree model involves transforming data into a dtree and corresponding decision rules. The primary benefit of this model is its ability to simplify complex decision-making processes, making them easier to understand. This clarity enables more interpretable decisions and solutions to the given problem [5].

Dtree illustrates factors (probabilities) that can predict the final outcome if the conditions in the decision tree are satisfied. It converts data into a visual format, using diagrams and decision rules to represent the process [6].

In a dtree, data is shown in a table that includes attributes and records. Attributes serve as the criteria for building the tree. One key attribute, known as the target attribute, identifies the solution for each data entry[7]. Attributes are associated with values known as instances. The process in decision trees involves converting the tabular data into a tree model, then transforming that model into decision rules, and finally refining those rules. This data is used to construct a decision tree prediction model, which is then evaluated for its accuracy in making predictions [6].

The C4.5 algorithm, which builds on the ID3 algorithm, excels at creating decision trees for various uses, including analyzing water quality. It works by selecting the most pertinent attribute as the root of the tree, generating branches based on the values of that attribute, and then splitting the cases according to these branches. This recursive process continues until each branch contains cases that all belong to the same class. [8].

The utilization of decision tree algorithms for predicting water quality has been extensively examined in numerous studies, establishing a solid theoretical base for the present research. Earlier works have demonstrated that decision tree models, including the C4.5 algorithm, are highly effective in classifying water quality by focusing on critical parameters such as pH, Total Dissolved Solids (TDS), and turbidity. For instance, [Nurlaelatul Maulidah et al.] (2024) employed decision trees to successfully forecast water contamination levels in drinking water supplies, achieving high precision in differentiating between safe and unsafe water for consumption. In a similar vein, [Aldi Tangkelayuk et al.] (2022) explored the application of various algorithms like Naive Bayes, Decision Tree, and K-Nearest Neighbors in evaluating the safety of drinking water, showcasing the robustness of these models in handling large datasets and their capability to generate results that are easy to interpret.

These collective studies emphasize the significant role that decision trees play in environmental monitoring and management, particularly in scenarios where swift and accurate assessments are critical. Nevertheless, despite the valuable insights provided by these studies, there remains a gap in research that customizes decision tree algorithms to the specific environmental conditions found in regions with high levels of water pollution, such as those prevalent in Indonesia. This study aims to address this gap by focusing on the application of the C4.5 algorithm to predict water quality in areas where water pollution is a widespread issue, thus contributing to the advancement of efficient water quality monitoring systems.

This research aims to develop a model for predicting water quality using the Dtree approach and the C4.5 algorithm. The model will evaluate various water quality parameters such as pH, Total Dissolved Solids (TDS), and turbidity to determine if the water meets or does not meet class II water quality standards [9]. This method is designed to create precise and effective prediction tools to assist in monitoring and managing water quality, especially in areas impacted by river pollution. [10].

Ensuring the quality of water is a critical issue that affects both human health and the broader environment. This research focuses on developing a predictive model to evaluate water quality by utilizing the Decision Tree (Dtree) methodology, specifically through the application of the C4.5 algorithm. The core aim of this study is to build a dependable and precise model capable of categorizing water samples in accordance with Class II water quality standards, with particular emphasis on essential parameters such as pH, Total Dissolved Solids (TDS), and turbidity.

The selection of these parameters pH, TDS, and turbidity is grounded in their critical role in determining water quality. pH is a fundamental indicator of the water's acidity or alkalinity, which directly influences the solubility and biological availability of chemical constituents such as nutrients and heavy metals. TDS measures the concentration of dissolved substances, including minerals, salts, and metals, which are indicative of water purity and potability. Turbidity, on the other hand, assesses the clarity of water and the presence of suspended particles, which can harbor harmful microorganisms and contaminants.

Previous studies have shown the effectiveness of the C4.5 algorithm in predicting water quality, but not many have adapted its application to highly polluted environments, such as in Indonesia. This study hypothesizes that the C4.5 algorithm, when applied to water quality data from highly polluted areas, will produce more accurate predictions in classifying water according to class II quality standards.

By achieving the goals of this research, the study aims to offer a practical tool for water resource managers and policymakers, thereby enhancing the efficiency of water quality monitoring and management, especially in areas severely impacted by water pollution. The C4.5 algorithm was specifically chosen for its strength in producing interpretable models that can simplify the often complex decision-making processes associated with evaluating water quality. Ultimately, this research endeavors to improve decision support systems that facilitate more effective water management strategies, ensuring the protection of public health and the sustainability of environmental resources.

2. Method

2.1 Data Mining

Is the practice of discovering meaningful patterns and insights from large and complex datasets. In essence, it's about extracting useful knowledge from extensive databases. Advances in areas such as science and business have significantly expanded both the size and complexity of these databases. [11].

2.2 Decision Tree

A Dtree is a diagram that resembles a branching tree. In this chart, each internal node represents a test on an attribute, each branch shows the result of that test, and each leaf node signifies a class or the distribution of classes. The topmost node is referred to as the root node. [12]. In a DTree, the root node has several branches extending outward but none coming in. Internal nodes have a single incoming branch and multiple outgoing ones. In contrast, leaf nodes feature one incoming branch and no outgoing branches. Dtree are used to classify data samples into predefined categories when

ISSN 2614-4859

their class labels are still unknown. [13]. Various algorithms can be applied in Dtree, often developed to improve upon previous methods. Selecting the best algorithm depends on the characteristics of the training data. Some of the well-known algorithms for Dtree are ID3, C4.5, and CART. [14].

According to (Laila Qadrini, Andi Seppewali, 2021) the typical features of decision trees are as follows:

1. Decision Trees use a nonparametric method to build classification models.
2. The techniques in Decision Trees enable rapid model creation from large training datasets.
3. Smaller Decision Trees are generally easier to understand.
4. Decision Trees effectively illustrate discrete value functions in learning.
5. The Decision Tree algorithm is fairly resilient to noise, especially when it incorporates methods to address overfitting.
6. Having a large number of attributes does not significantly impact the accuracy of Decision Trees.
7. Some Decision Tree algorithms employ a top-down approach, recursively partitioning the data until the number of records at leaf nodes may be too small for reliable statistical decisions.
8. While subtrees can be replicated within a Decision Tree, doing so increases the complexity and makes the tree harder to interpret.

2.3 Algorithm C4.5

The C4.5 algorithm is a commonly utilized technique for building decision trees and is well-regarded by researchers around the world. It is an improved version of the ID3 algorithm, originally created by J. Ross Quinlan. Generally, constructing a decision tree with the C4.5 algorithm involves several key steps:

a. Select an attribute to act as the root,

b. Create a branch for each possible value of that attribute,

c. Distribute the cases across these branches,

d. Repeat this process for every branch until the tree is completely built.

The C4.5 algorithm constructs a dtree by recursively examining each decision node, selecting the best branch at each step, until no further branches can be added. The process for applying the C4.5 algorithm involves preparing the training data, finding the root of the tree by calculating entropy, computing the gain value, and deciding how to partition the data [15].

$$Entropy\ (S) = \sum_{i=1}^{P} - pi * log_2\, pi \tag{1}$$

Where :

S	: Case set
A	: feature
N	: Number of S partitions
Pi	: Proportion of Si to S

The concept of gain (S, A) relates to the level of information attribute A provides about the outcome data S. This gain is assessed by examining how the dependent variables S are classified based on attribute A and is quantified as gain. [5].

$$Gain(N, A) = Entropy(S) - \sum_{i=1}^{P} \frac{|Si|}{|S|} * Entropy(S) \tag{2}$$

With :

The C4.5 variant of the Decision Tree algorithm was selected for this research due to its strong ability to process both categorical and continuous variables, making it adaptable to a wide array of applications, including the prediction of water quality. A key advantage of the C4.5 algorithm lies in its capacity to produce decision trees that are straightforward to interpret, offering clear decision rules that can be easily understood by both professionals and laypersons. This clarity is particularly important in the field of environmental management, where decisions must often be communicated to policymakers and other stakeholders who may not have a technical background.

Furthermore, the C4.5 algorithm is equipped with methods to address missing data and can efficiently handle large datasets, which are frequently encountered in environmental research. It utilizes information gain and gain ratio as criteria for attribute selection, ensuring that the most relevant features are emphasized in the decision-making process. This is especially beneficial in water quality prediction, where parameters such as pH, Total Dissolved Solids (TDS), and turbidity are critical for accurate classification.

Additionally, the C4.5 algorithm includes a pruning technique to simplify the final decision tree, which helps prevent overfitting a common challenge in machine learning. This ensures that the model performs well when applied to new, unseen data, enhancing its reliability for practical, real-world scenarios. Given these attributes, the C4.5 decision tree algorithm is ideally suited for developing a predictive model that is both precise and practical for evaluating water quality under varying environmental conditions.

2.4 Predictions

Prediction involves estimating future events using mathematical methods based on past and present information to reduce errors. Employing prediction techniques helps determine production levels for future periods. Choosing the right prediction method is crucial, as it must offer high accuracy and minimal deviation between forecasted outcomes and actual results. In addition to accuracy, the selection of prediction techniques may also depend on how well they adapt to changes in data [16].

When constructing the decision tree model using the C4.5 algorithm, the careful selection of water quality parameters is essential, as it significantly influences the model's accuracy and dependability. The parameters chosen for this study pH, Total Dissolved Solids (TDS), and turbidity were selected due to their critical importance in assessing water quality. pH serves as a vital measure of the water's acidity or alkalinity, influencing not only the chemical balance of the water but also the well-being and growth of aquatic organisms. TDS indicates the amount of dissolved substances, such as minerals, salts, and metals, thereby providing insight into water purity and its safety for consumption. Turbidity, which measures the clarity of water, is another key parameter, as elevated turbidity levels can signal the presence of harmful particles, microorganisms, and other contaminants, making the water potentially hazardous for human use.

These parameters were chosen deliberately because they offer a comprehensive evaluation of water quality, addressing both its chemical and physical characteristics. By concentrating on these essential indicators, the decision tree model is better equipped to accurately determine whether the water meets Class II water quality standards. Furthermore, these parameters are widely recognized in water quality monitoring, allowing for easy comparison with results from other studies and established benchmarks. The careful selection of these parameters strengthens the methodological rigor of the research, ensuring that the predictive model is both applicable and trustworthy for practical use in managing water resources.

2.5 Flowchart

This study employs an experimental approach, which includes several key components. This method involves conducting systematic experiments to collect and analyze data, allowing for an in-depth understanding of the phenomena under investigation. The experimental design encompasses defining variables, setting up controlled conditions, and applying appropriate analytical techniques. By following this method, the research aims to generate reliable and valid results that can contribute valuable insights into the topic. The following is the research flowchart

1. Gathering data
2. Preliminary data processing
3. Selection of the model
4. Testing the model
5. Assessing the model's performance

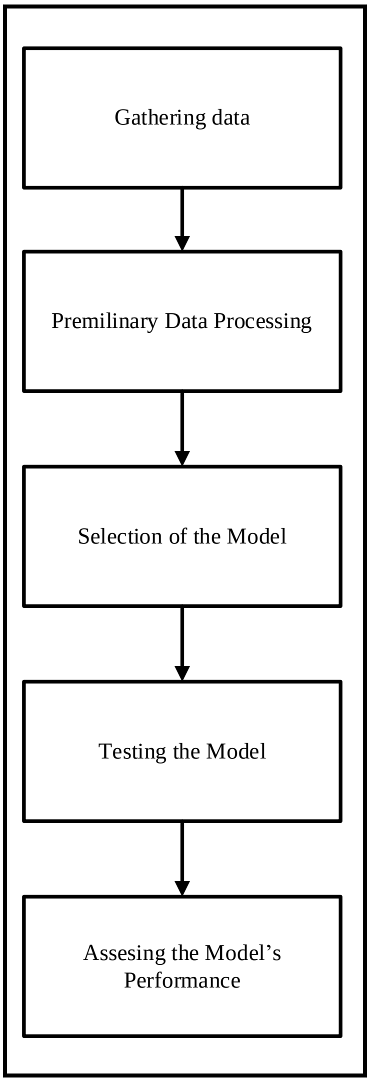


Fig 1. Research Flowchart

Figure 1 shows a flowchart outlining the decision tree approach used in this study. The process begins with collecting primary data from water quality tests, where parameters such as pH, TDS, and turbidity are measured using specialized instruments, the sample data being taken from the sidomulyo village area in lamongan district. The parameters pH, TDS, and turbidity were selected in this study to meet the daily needs of humans as these three parameters determine the quality and safety of water for daily use. pH indicates the acidity of water which can affect the taste and safety of water. TDS relates to dissolved minerals that can make water taste brackish if excessive. Turbidity also reduces the purity quality of water which is also affected by TDS. After collecting this data, it is processed and organized to form a dataset that displays these parameters and attributes. This research uses the C4.5 decision tree algorithm to predict water quality, utilizing

RapidMiner software for model prediction and evaluation. The final step in the process is model testing.

3. Results and Discussion

3.1. Dataset Mapping

Compile a dataset for model training using information gathered from various sources concerning water quality standards outlined in regulations. This dataset includes parameters like pH, turbidity, and Total Dissolved Solids (TDS). These parameters are categorized into two main groups: those that meet class II water quality standards and those that do not. Details of the data mapping are provided in Table 1.

Table 1. Data Mapping

No	Information			Information
	pH	TDS	Turbidity	
1	6-9	0-800	0-25	Class 2 water quality standards
2	<6, >9	>1000	>25	Not yet Class 2 water quality standards

in table 1 above is mapping data based on Government Regulation No.22 of 2021 class II water quality standards for pH and TDS parameters. while the turbidity parameter is taken from Permenkes No.32 of 2017 which is the reference for this study.

3.2. Initial data collectin

From different locations around the Sidomulyo Village area. These samples were then consolidated, recorded in a tabular format, and categorized into two classes. This organized data was processed to provide the necessary information for this research. The table of sample results is shown below:

Table 2. Water Sample Data

No	pH	TDS	Turbidity	Water Quality Standards
1	6,61	310,38	21,2	Class 2 water quality standards
2	11,36	373,03	16,69	Class 2 water quality standards
3	3,9	1112,89	12,42	Not yet Class 2 water quality standards
4	11,51	1023,67	12,56	Not yet Class 2 water quality standards
5	7,09	1345,16	39,6	Not yet Class 2 water quality standards
6	5,94	862,32	17,95	Class 2 water quality standards
7	5,09	148,69	11,98	Class 2 water quality standards
8	8,53	253,64	11,97	Class 2 water quality standards
9	3,3	1123,14	35,69	Not yet Class 2 water quality standards
10	3,14	104,63	28,26	Class 2 water quality standards
11	6,86	1222,48	23,5	Not yet Class 2 water quality standards
12	3,61	740,1	48,82	Class 2 water quality standards
13	5,27	1329,51	15,34	Not yet Class 2 water quality standards
14	4,4	300,7	76,1	Class 2 water quality standards
15	10,4	667,5	55,8	Class 2 water quality standards
16	4,9	1002,9	44,3	Not yet Class 2 water quality standards
17	3,8	1126,9	29,7	Not yet Class 2 water quality standards
18	10,4	1456,94	60,7	Not yet Class 2 water quality standards
19	4,2	1256,4	54,1	Not yet Class 2 water quality standards
20	6	509,89	24,3	Class 2 water quality standards
21	5,6	380,3	66,5	Class 2 water quality standards
22	6,9	231,1	17,6	Class 2 water quality standards
23	11,7	1034,63	35,4	Not yet Class 2 water quality standards

ISSN 2614-4859

The data in table 2. above are taken from various samples in sidomulyo village, lamongan district, such as water in sewers, water in the sidomulyo village reservoir and PDAM water to support this research.

3.3. Data Selection

The goal of selecting data is to determine which attributes are most crucial for predicting if water samples adhere to class 2 water quality standards. In the subsequent Table 3, you'll find a list of these key attributes that will be analyzed using RapidMiner software. This careful selection ensures that the most relevant factors are considered, enhancing the accuracy and effectiveness of the prediction model. By focusing on these attributes, the model aims to provide more reliable assessments of water quality, ultimately supporting better water management practices.

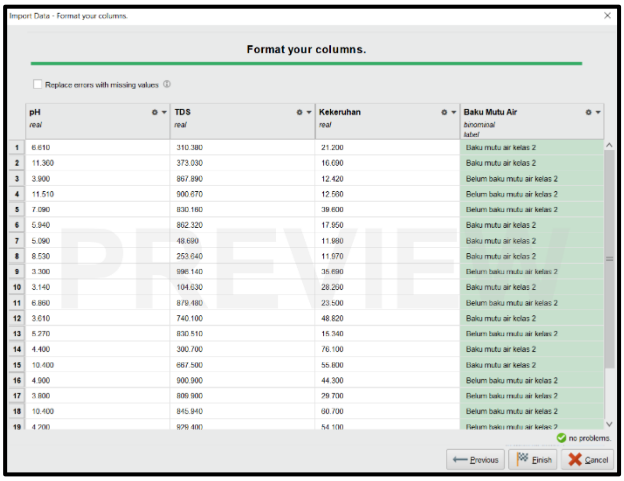
Table 3. Water Sample Data

No	Selected Attribute	Information
1	pH	Feature attributes
2	Total Disolved Solid (TDS)	Feature attributes
3	Turbidity	Feature attributes
4	Water quality standards	Label

The selected attributes include pH, Total Dissolved Solids (TDS), turbidity, and water quality standard classes as labels. These attributes were chosen due to their impact on water quality. They will be processed using the C4.5 algorithm with RapidMiner software, which will generate a decision tree and evaluate its accuracy in predicting water quality standard classes.

3.4. Data Transformation

At this point, the data format is modified to streamline the next steps in the data mining process. RapidMiner software is used to convert the dataset, changing the labels from polynomial to binomial types due to the presence of only two potential outcomes, as shown in Figure 2.



Import Data - Format your columns.

Format your columns.

□ Replace errors with missing values

	pH real	TDS real	Kekeruhan real	Baku Mutu Air binominal label
1	6.610	310.380	21.200	Baku mutu air kelas 2
2	11.360	373.030	16.690	Baku mutu air kelas 2
3	3.900	867.890	12.420	Belum baku mutu air kelas 2
4	11.510	900.670	12.560	Belum baku mutu air kelas 2
5	7.090	830.160	39.600	Belum baku mutu air kelas 2
6	5.940	862.320	17.950	Baku mutu air kelas 2
7	5.090	48.690	11.980	Baku mutu air kelas 2
8	8.530	253.640	11.970	Baku mutu air kelas 2
9	3.300	996.140	35.690	Belum baku mutu air kelas 2
10	3.140	104.630	28.260	Baku mutu air kelas 2
11	6.860	879.480	23.500	Belum baku mutu air kelas 2
12	3.610	740.100	48.820	Baku mutu air kelas 2
13	5.270	830.510	15.340	Belum baku mutu air kelas 2
14	4.400	300.700	76.100	Baku mutu air kelas 2
15	10.400	667.500	55.800	Baku mutu air kelas 2
16	4.900	900.900	44.300	Belum baku mutu air kelas 2
17	3.800	809.900	29.700	Belum baku mutu air kelas 2
18	10.400	845.940	60.700	Belum baku mutu air kelas 2
19	4.200	929.400	54.100	Belum baku mutu air kelas 2

no problems.

Fig 2. Data Transformation in RapidMiner

3.5. Modeling Data Mining

In this phase, two data mining models Decision Tree and C4.5 are developed using the processed data. The Decision Tree model is built with three attributes that affect water quality results. This model visually represents the potential outcomes based on various choices related to these attributes. Subsequently, the data is split into two categories: one for data meeting class 2 water quality standards and another for data not meeting these standards. This division is done using predefined attributes and training data. The goal is to train the model with data that meets class 2 standards and then evaluate its accuracy with data that does not. The results of the Decision Tree modeling are displayed in Figure 3.

ISSN 2614-4859

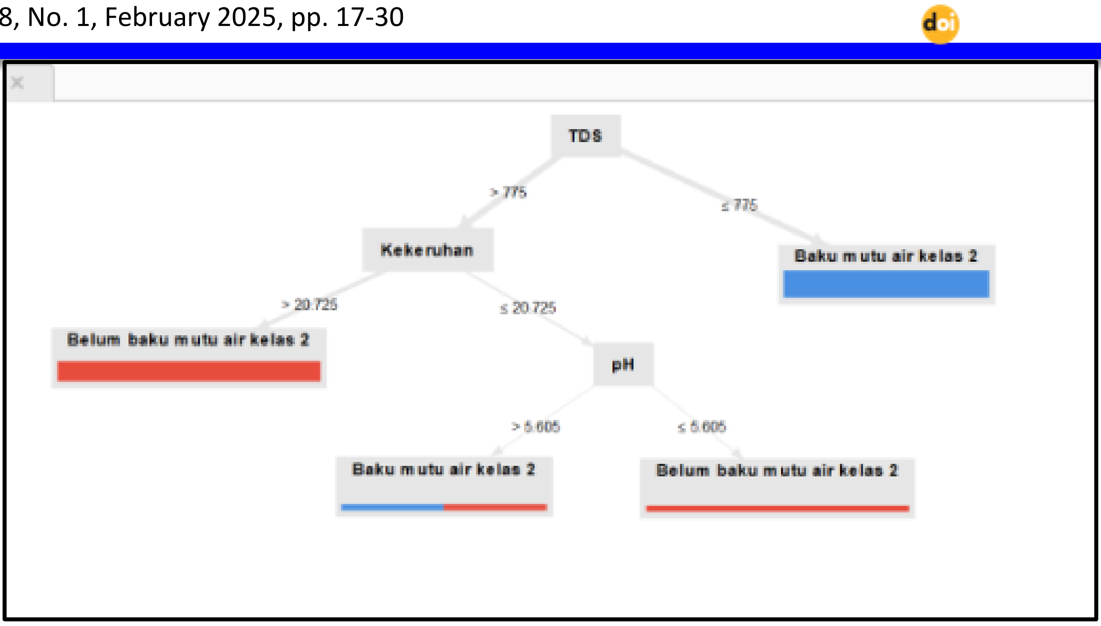
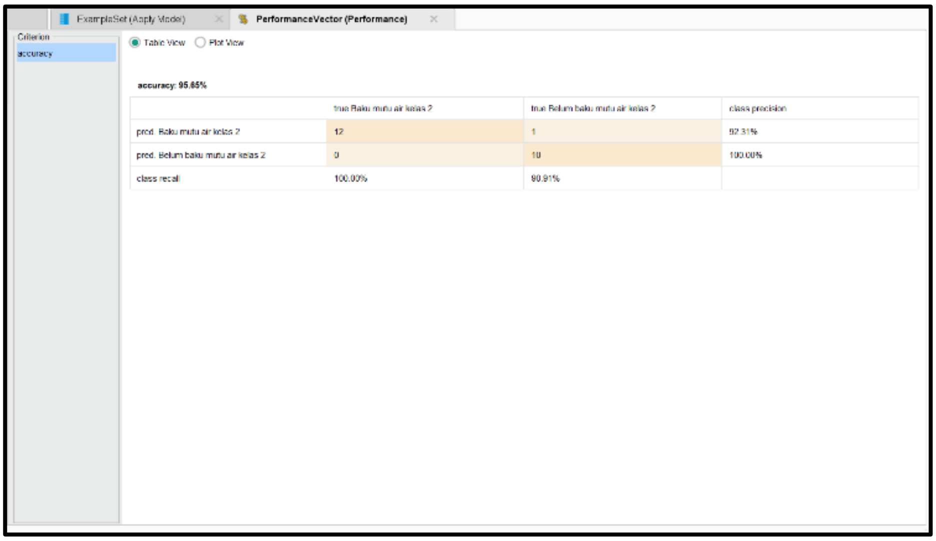


Fig 3. Modeling Decision Tree

After creating and testing the model, accuracy the C4.5 Decision Tree algorithm is assessed. In this experiment, the C4.5 Decision Tree model achieved an accuracy rate of 95.65%. This result indicates that the model performs effectively in predicting or classifying class 2 water quality standards based on the three attributes used.



accuracy: 95.65%

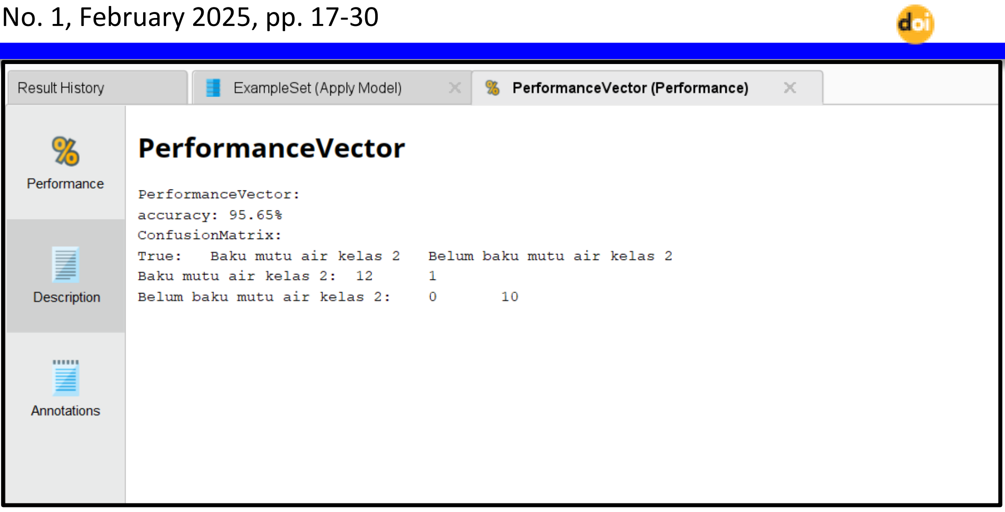
	true Baku mutu air kelas 2	true Belum baku mutu air kelas 2	class precision
pred. Baku mutu air kelas 2	12	1	92.31%
pred. Belum baku mutu air kelas 2	0	10	100.00%
class recall	100.00%	90.91%	

Fig 4. Performance Vector in RapidMiner

3.6. Evaluation

The evaluation of the C4.5 Decision Tree algorithm reveals an accuracy rate of 95.65%. For predictions of class 2 water quality standards, there were 12 samples correctly classified as meeting these standards and none incorrectly classified, with a precision of 92.31% and a recall of 100%. For samples not meeting class 2 standards, there were no false positives and 10 samples correctly identified, resulting in a precision of 100.00% and a recall of 90.91%. The C4.5 Decision Tree proves to be an effective method for initially classifying water quality. The good performance of this prediction model is reflected in the confusion matrix, which shows an accuracy of 95.65%, as illustrated in Figure 5.

ISSN 2614-4859



PerformanceVector

```
PerformanceVector:
accuracy: 95.65%
ConfusionMatrix:
True:   Baku mutu air kelas 2   Belum baku mutu air kelas 2
Baku mutu air kelas 2:  12      1
Belum baku mutu air kelas 2:    0       10
```

Fig 5. Performance Vector in RapidMiner

A confusion matrix is a valuable tool for evaluating the performance of a classification model by comparing its predicted outcomes with the actual values from test data. Figure 5 presents a detailed confusion matrix that highlights the accuracy of the model by showing both correct and incorrect predictions. By analyzing this matrix, you can gauge how effectively the model differentiates between various classes and pinpoint areas that may need refinement. The information provided in the matrix is essential for assessing the model's accuracy, precision, recall, and overall effectiveness in data classification.

1. True Positives for Baku mutu air kelas 2 : 12
2. False Negative for Baku mutu air kelas 2 : 1
3. True Negatives for belum baku mutu air kelas 2 : 0
4. False Positives for belum baku mutu air kelas 2 : 10

From the values obtained, researchers can calculate the manual work model metrics as follows:

Accuracy

$$Accuracy \ = \frac{TP+TN}{TP+TN+FP+FN} \times 100 \tag{3}$$

$$Accuracy = \frac{12+10}{12+1+0+10} \times 100 = 95{,}65\% \tag{4}$$

Precision

$$Precision \ = \frac{TP}{TP+FP} \tag{5}$$

$$Precision\ (BMK\ 2) = \frac{12}{12+0} = 100 = 100\% \tag{6}$$

$$Precision\ (BBKM\ 2) = \frac{10}{1+10} = 0{,}9091 = 90{,}91\% \tag{7}$$

Recall

$$Recall \ = \frac{TN}{TP+FN} \tag{8}$$

$$Recall\ (BMK\ 2) = \frac{12}{12+1} = 0{,}9231 = 92{,}31\% \tag{9}$$

$$Recall\ (BBKM\ 2) = \frac{10}{10+0} = 1 = 100\% \tag{10}$$

The performance metrics derived from the decision tree model for supporting class 2 water quality standards reveal exceptional effectiveness. Achieving an overall accuracy of 95.65% indicates that the model correctly classifies data in most instances. For the "Class 2 Water Quality" category, the precision is 100%, signifying that all positive predictions are accurate. The precision for the "Non-Class 2 Water Quality" category stands at 90.91%, demonstrating that a substantial majority of positive predictions are accurate. The recall for the "Class 2 Water Quality" category is 92.31%, highlighting the model's ability to correctly identify most positive cases. In contrast, the recall for the "Non-Class 2 Water Quality" category is 100%, indicating that the model accurately identifies all negative cases. Collectively, these metrics underscore the decision tree model's high efficacy in supporting class 2 water quality standards.

Additionally, the model's robustness is reflected in its ability to maintain high performance across different metrics, ensuring reliable and consistent classification outcomes. Such results suggest the model's potential for practical applications in water quality management, where accurate and reliable data classification is crucial. The exceptional precision and recall values for both categories underscore the model's utility in distinguishing between compliant and non-compliant water quality cases, further enhancing its suitability for environmental monitoring and regulatory compliance.

3.7. Discussion

Based on the research presented, the C4.5 Decision Tree algorithm demonstrates high effectiveness for predicting water quality standards, achieving an impressive accuracy rate of 95.65%. This suggests the model is reliable for initial classification of water quality based on pH, TDS, and turbidity. The confusion matrix reveals strong performance across different metrics, with perfect recall for water meeting class 2 standards and high precision for both categories. These results indicate that the selected attributes are strong indicators of water quality in the context of class 2 water standards.

This study's findings align with existing literature, particularly in the use of the C4.5 Decision Tree algorithm for water quality prediction. For example, studies by Nurlaelatul Maulidah et al. (2024) and Aldi Tangkelayuk et al. (2022) have demonstrated the effectiveness of Decision Tree models, including C4.5, in classifying water quality. However, our study improves upon these by focusing on areas with high levels of pollution, such as those found in Indonesia, and demonstrating the algorithm's ability to handle complex environmental conditions. While previous studies mainly focused on generic datasets, our research provides a more contextual application by utilizing real-world data from Sidomulyo Village. Compared to algorithms like Naive Bayes and K-Nearest Neighbors used in other studies, the C4.5 algorithm's interpretability and high precision make it particularly suited for environmental applications where decision-making transparency is crucial.

The model's high accuracy suggests it could be a valuable tool for rapid initial assessment of water quality, particularly in areas where water pollution is a concern. It could help prioritize water sources for more detailed testing or treatment. Additionally, this model could be integrated into automated water monitoring systems to continuously evaluate water quality, enabling proactive measures when standards are not met. The clear decision rules generated by the C4.5 algorithm enhance its usability for policymakers and non-experts, facilitating the adoption of advanced data analytics in routine water quality management.

However, it's important to note that the model was trained and tested on a relatively small dataset of 23 samples. Future research could benefit from expanding the dataset, incorporating additional water quality parameters, and testing the model's performance in different geographical areas to assess its generalizability. Building on these findings, future research could explore several avenues. First, expanding the dataset to include a wider variety of geographical locations and environmental conditions would help in testing the model's robustness and generalizability. Additionally, incorporating more water quality parameters, such as heavy metal concentrations or microbial counts, could provide a more comprehensive assessment of water safety. Research could also compare the performance of the C4.5 algorithm against other advanced machine learning techniques, such as Random Forests or Gradient Boosting Machines, to identify the best approach for specific water quality contexts. Finally, integrating this model with Internet of Things (IoT) technologies could enhance real-time water quality monitoring and response systems.

The transformation of data from polynomial to binomial format for the labels was a key step in preparing the data for the C4.5 algorithm. This simplification to a binary classification problem likely contributed to the high accuracy of the model. Given its performance, the model could serve as a valuable tool for water resource managers and environmental agencies, helping to quickly identify water sources that require immediate attention or further testing.

This study contributes to the theoretical understanding of decision tree applications in environmental science. By demonstrating the effectiveness of the C4.5 algorithm in classifying water quality under varying environmental conditions, our research highlights the algorithm's adaptability and interpretability as key strengths. The model's ability to generate straightforward decision rules that can be easily interpreted by users reinforces its suitability for practical applications. Moreover, this study enriches the existing literature by providing empirical evidence on the utility of decision trees in environmental monitoring, particularly in regions with high pollution levels. This not only expands the scope of decision tree methodologies in data science but also emphasizes their value in promoting sustainable water management practices.

While the study demonstrates the effectiveness of the C4.5 Decision Tree algorithm, comparing its performance with other machine learning algorithms would be interesting to further validate its superiority for this specific application. In conclusion, this research presents a promising approach to water quality prediction. The high accuracy achieved suggests that this method could be a valuable addition to water quality management tools, particularly for initial rapid assessments. However, further research with larger datasets and in diverse geographical contexts would be beneficial to fully establish the model's reliability and generalizability.

4. Conclusion

This research shows that the C4.5 algorithm-based water quality prediction model is highly accurate in classifying water according to class II quality standards, especially in areas with high pollution levels. Therefore, it is recommended that this model be integrated into a real-time water quality monitoring system to enable early detection and quick action. Practitioners and policy makers can use the model as a decision-making tool to prioritize water treatment in the most polluted areas. However, further customization and the development of a broader dataset are needed to improve the reliability of this model. Collaboration between academia, government and industry is also important to ensure effective application of the model in water quality management.

References

[1] M. Anshori, N. Sari, and A. Vijai, “Rancang Bangun Alat Ukur Kekeruhan Air Layak Pakai Berbasis Arduino Uno R3 Pada Sungai Martapura,” Jurnal Ilmiah Fisika FMIPA Universitas Lambung Mangkurat, vol. 19, no. 1, 2022.

[2] J. M. A. F. Dina Rachmawaty, “PENERAPAN METODE KLASIFIKASI DECISION TREE UNTUK MEMPREDIKSI KELULUSAN TEPAT WAKTU,” Journal of Industrial Engineering and Technology, vol. 2, no. 1, pp. 61–74, Jan. 2022, doi: 10.24176/jointtech.v2i1.7432.

[3] Patang, “Analysis of Mechanical Water Filtration Systems for River Water Quality,” The International Journal of Science & Technoledge, vol. 7, no. 10, Jan. 2019, doi: 10.24940/theijst/2019/v7/i1/st1901-005.

[4] I. M. Agus Oka Gunawan, I. D. A. Indah Saraswati, I. D. G. Riswana Agung, and I. P. Eka Putra, “Klasifikasi Penyakit Jantung Menggunakan Algoritma Decision Tree Series C4.5 Dengan Rapidminer,” Jurnal Teknologi Dan Sistem Informasi Bisnis, vol. 5, no. 2, pp. 73–83, Apr. 2023, doi: 10.47233/jteksis.v5i2.775.

[5] I. Khoeri and D. Iskandar Mulyana, “Implementasi Machine Learning dengan Decision Tree Algoritma C4.5 dalam Penerimaan Karyawan Baru pada PT. Gitareksa Dinamika Jakarta,” Jurnal Sosial Teknologi, vol. 1, no. 7, pp. 615–623, Jul. 2021, doi: 10.59188/jurnalsostech.v1i7.126.

[6] S. Farida Utami, Penerapan Data Mining Algoritma Decision Tree Berbasis PSO. 2020.

[7] Y. Resti, C. Irsan, M. Amini, I. Yani, R. Passarella, and D. A. Zayantii, “Performance Improvement of Decision Tree Model using Fuzzy Membership Function for Classification of Corn Plant Diseases and

Pests," Science and Technology Indonesia, vol. 7, no. 3, pp. 284–290, Jul. 2022, doi: 10.26554/sti.2022.7.3.284-290.

[8] W. Darmawan, M. Basuki Rahmat, A. Khumaidi, R. Yudha Adhitya, and D. Pristovani Riananda, "Perancangan Strategi Keputusan Robot Sepak Bola Beroda menggunakan Metode Decision Tree," Jurnal Elektronika dan Otomasi Industri, vol. 10, no. 2, pp. 175–182, Jul. 2023, doi: 10.33795/elkolind.v10i2.3020.

[9] E. Setiawan and I. Munadhif, "Rancang Bangun Pengendalian Kualitas Air Pada Budidaya Ikan Nila Merah Menggunakan Metode Fuzzy Logic Dengan Sistem Monitoring Berbasis Internet Of Things," in Jurnal Conference on Automation Engineering and Its Application, 2021, p. Vol.1, No.1,pp.220-226.

[10] E. SETIAWAN, G. ANINDITA, A. SYAHID, and I. RACHMAN, "Monitoring Keseimbangan Distribusi Beban Transformator untuk Meminimalisasi Terjadinya Rugi Energi," ELKOMIKA: Jurnal Teknik Energi Elektrik, Teknik Telekomunikasi, & Teknik Elektronika, vol. 7, no. 2, p. 297, May 2019, doi: 10.26760/elkomika.v7i2.297.

[11] D. K. Pramudito, "DATA MINING IMPLEMENTATION ON JAVA NORTH COAST WEATHER FORECAST DATASET USING C4.5 ALGORITHM," vol. 12, 2021.

[12] A. Tangkelayuk, "The Klasifikasi Kualitas Air Menggunakan Metode KNN, Naïve Bayes, dan Decision Tree," JATISI (Jurnal Teknik Informatika dan Sistem Informasi), vol. 9, no. 2, pp. 1109–1119, Jun. 2022, doi: 10.35957/jatisi.v9i2.2048.

[13] Ghilman Ahmad Faza, Muhammad Fikri Fathurrohman, Purwidi Asri, Anggara Trisna Nugraha, and Perwi Darmajanti, "PROTOTYPE SISTEM OILY WATER SEPARATOR OTOMATIS PADA KAPAL MENGGUNAKAN METODE DECISION TREE BERBASIS MIKROKONTROLER," Jurnal 7 Samudra, vol. 8, no. 2, pp. 1–6, Nov. 2023, doi: 10.54992/7samudra.v9i1.128.

[14] L. Qadrini, A. Seppewali, and A. Aina, "DECISION TREE DAN ADABOOST PADA KLASIFIKASI PENERIMA PROGRAM BANTUAN SOSIAL," Jurnal Inovasi Penelitian, vol. 2, no. 1, 2021.

[15] F. Febrianti, S. A. Wibowo, and N. Vendyansyah, "Implementasi Iot(Internet Of Things) Monitoring Kualitas Air Dan Sistem Administrasi Pada Pengelola Air Bersih Skala Kecil," Jurnal Mahasiswa Teknik Informatika), vol. 5, no. 1, 2021.

[16] Firdausia Ismi Nurhayati, B. Nugroho, and I. Yuniar Purbasari, "Implementasi Metode Decision Tree Pada Identifikasi Status Gizi Balita," Jurnal Informatika dan Sistem Informasi, vol. 2, no. 2, pp. 204–213, Jul. 2021, doi: 10.33005/jifosi.v2i2.326.

This Page Is Intentionally Left Blank