Enhancing Water Potability Assessment Using Hybrid Fuzzy-Naïve Bayes

Fadhillah Azmi¹, M. Khalil Gibran², Achmad Ridwan³, Amir Saleh⁴
azmi.fadhillah007@gmail.com¹, khalilgibran1612@gmail.com²,
achmadridwan@unprimdn.ac.id³, amirsalehnst1990@gmail.com⁴
¹Department of Electrical Engineering, Universitas Medan Area, Medan, Indonesia
²Department of Data Science, Universitas Muhammadiyah Sumatera Utara, Medan, Indonesia
³Department of Informatics Engineering, Universitas Prima Indonesia, Medan, Indonesia

Abstract
In an effort to ensure a safe and high-quality water supply, the assessment of water potability is of paramount importance. An accurate and efficient assessment of water potability can be a challenge due to various influencing factors. Therefore, an innovative and integrated approach is needed to improve the assessment of water potability. In this study, we introduce a new approach to improving the assessment of water potability. This approach aims to overcome the shortcomings of traditional methods by using a hybrid fuzzy-Naïve Bayes approach to obtain a more accurate level of water potability. Fuzzy techniques are used to overcome uncertainty and ambiguity in the initial data. This method describes the probability weights in a fuzzy manner for various parameters. Then, the Naïve Bayes method is used to classify water samples based on the probability generated by the fuzzy system. This hybrid approach makes it possible to consider the relationship between parameters and generate more realistic probability values. This study uses datasets collected from various sources that include water potability parameters. A hybrid fuzzy-Naïve Bayes approach was then applied to this data set to make a more effective and accurate assessment of water potability. The experimental results show that the proposed method obtains an accuracy of 90%, which significantly improves the water potability assessment compared to the conventional method, which results in an accuracy of 84%. By combining fuzzy and Naïve Bayes techniques, we can overcome uncertainty in data and produce more accurate judgments.
A. Introduction

Clean water that is safe for consumption is the main need for humans to carry out their daily activities. Obtaining safe drinking water is essential to health and is the right of every human being [1]. The supply of clean drinking water has the primary objective of protecting human health, including ensuring that it is available in adequate quantities [2]. Investments in the provision of clean water and sanitation can generate economic benefits because they reduce the impact of diseases obtained from dirty drinking water. The traditional approach to measuring water potability is carried out by comparing parameter values that have been determined experimentally based on existing guidelines [3]. Drinking water that is safe for human consumption can follow the provisions of the WHO as follows: pH = 6.5–8.5, Solids = 1000 ppm, Hardness = 200 mg/L, Conductivity = 400 μS/cm, Sulfate = 1000 mg/L, Chloramines = 4 ppm, Trihalomethanes = 80 ppm, Organic carbon = 10 ppm, and Turbidity = 5 NTU [4]. This parameter is needed to determine the potability level of safe drinking water.

The potability of water is largely influenced by natural processes and human actions. Surface water run-off is a seasonal phenomenon that is largely influenced by climate and anthropogenic discharges, which are sources of pollution in rivers [5]. Meanwhile, many people still use river water as their basic need. Lack of knowledge and funds causes them to continue to meet their water needs from the rivers around them even though they have been polluted. This, of course, will have a negative impact on their health. Water potability detection devices require large amounts of funds to be procured, and knowledge of determining water potability has not been the focus of their problem. For this reason, a study is needed that can provide a decision on the portability of water and whether it is suitable for drinking or not.

Various techniques for determining the potability of drinking water have been carried out by several previous researchers. One of the studies that determines water potability can be done with a classification algorithm using the Support Vector Machine (SVM) method. In this study, the accuracy of the test obtained a satisfactory value [6]. Other research has conducted trials using the artificial neural network method, which produces an accuracy rate of 80% [7]. Based on the research that has been described, the use of computer algorithms can solve problems regarding drinking water potability. However, in reality, the detection of water portability is still done manually in most areas, or testing is carried out in the laboratory to obtain a decision [8].

In addition to using the methods mentioned above, many water potability determinations are also carried out by applying the fuzzy logic method. Fuzzy set theory can reduce uncertainty related to water quality objectives and determine river water quality status quickly [9]. In addition, the level of water quality can be divided into several categories and used for various community needs according to the resulting fuzzy classification. Research using the fuzzy method has been carried out with outputs in the form of categorized water quality status and can make it easier for the public to read information about river water quality from environmental agencies and raise public awareness to protect watersheds [10].

In this study, a model for determining the potability of drinking water will be built using the fuzzy logic method and the Naïve Bayes (NB) algorithm. Fuzzy
algorithms are used to determine the input data to be converted into fuzzy values. Meanwhile, the Naïve Bayes algorithm is used to measure water portability and determine whether water is suitable for consumption or not. The combination of the two methods can be used to perform data analysis to determine water portability. In the early stages, water potability data is collected and processed using data pre-processing techniques such as cleaning, integration, selection, and transformation. Furthermore, the data is analyzed using the Naïve Bayes algorithm to find patterns between various water potability parameters, such as temperature, pH, turbidity, dissolved oxygen, and others. After obtaining this pattern, the results will be used as a basis for deciding whether the water is safe for consumption or not.

Assessment of the potability of drinking water has an important role in ensuring the safety and quality of water for consumption. Traditional approaches to water potential assessment often rely on binary classification, which classifies water samples as potable or non-potable. However, these methods may not accurately capture the complexity and uncertainty associated with water quality parameters. To overcome this limitation, researchers have proposed a new approach that harnesses the power of hybrid techniques to improve the results of drinking water portability assessments [11].

Fuzzy logic methods have been widely used in various fields to deal with uncertainty and ambiguity in data. In the context of water potential assessment, fuzzy logic can be applied to represent the degrees of membership or uncertainty associated with water quality parameters [12]. By utilizing fuzzy sets and fuzzy inference systems, researchers have been able to capture the fuzzy relationship between water quality attributes and potability. Meanwhile, the Naïve Bayes (NB) classifier is a probabilistic classification algorithm that assumes independence between input features. Despite its simplicity, Naïve Bayes has proven effective in many classification tasks. In the context of assessing the suitability of water, the Naïve Bayes (NB) classifier can be used to calculate the conditional probability of a water sample falling into the class of potable or unfit for drinking based on the features observed.

Research on improving the Naïve Bayes algorithm has been done before in different cases. Research that improves the Naïve Bayes algorithm using feature weighting and Laplace calibration obtains increased results, where the enhanced Naïve Bayesian classification algorithm has high accuracy and is very stable [13]. Other studies improve the Naïve Bayes algorithm with a fuzzy approach to can waste classification. The resulting average accuracy level was unsatisfactory from the k-fold cross-validation classification system using the usual Naïve Bayes model, so it was corrected using a fuzzy approach. This method has succeeded in increasing the average accuracy of the classification system and also shows that this method is better than using the usual Naïve Bayes [14]. Based on previous research, the Naïve Bayes combination technique needs to be used to obtain better results in handling classification cases.

Fuzzy logic will change the input values of water potability parameters into linguistic variables using predetermined fuzzy sets [15]. Furthermore, predetermined fuzzy rules will be used to determine the level of parameters needed to determine the potability of drinking water. Meanwhile, Naïve Bayes will
classify the input from the previous fuzzy and make a decision about whether the measurements produced on the water are safe to drink or not. The results obtained from the combination of fuzzy algorithms with Naïve Bayes can be used in monitoring water quality and providing warnings if there are problems that need to be addressed immediately. In addition, this combination can also be used to evaluate and improve water management systems in an area. The combination of fuzzy with Naïve Bayes refers to the use of fuzzy logic techniques in developing or improving the performance of the Naïve Bayes algorithm [16]. Both fuzzy logic and Naïve Bayes are commonly used methods in data analysis and decision making.

B. Research Method

Fuzzy and Naïve Bayes hybrid techniques are approaches that can be used to improve classification performance on data that has uncertainty or ambiguity. Fuzzy logic is used to overcome uncertainty by modeling and combining information that is not completely clear, while Naïve Bayes is a probabilistic classification method that uses Bayes' theorem to calculate class probabilities based on observed features. The steps taken in determining the potability of drinking water in this study can be described in Figure 1.

This research requires a dataset obtained from Kaggle’s Water Quality Dataset with a total of 1000 random data collections. Then, the dataset will be processed with the proposed approach using the Python program to obtain the results of data analysis in determining the potability of drinking water. Information about the dataset used in this study can be found in Table 1 below.
A complete description of the method for determining water potability using a combination of fuzzy logic and Naïve Bayes can be found as follows:

a. Pre-Processing
The data pre-processing stage is the initial stage before implementing the combination of fuzzy and Naïve Bayes. The purpose of this step is to prepare the data to suit the needs of the classification method to be used. Some common steps in data pre-processing include the deletion of irrelevant data, handling of missing values, normalization or standardization, and dimension reduction.

b. Fuzzification
The fuzzification step involves converting the input data into fuzzy domains by mapping the data values into the relevant fuzzy sets. This is done by using a membership function to describe the level of membership of the data in each fuzzy set. Some general steps in fuzzification include identification of linguistic variables, selection of membership functions, determination of membership function parameters, and data fuzzification. The result is the membership level for each fuzzy set in the input data. In this study, the fuzzification process can be seen in Table 2 below.

<table>
<thead>
<tr>
<th>Name of Parameters</th>
<th>Membership Function</th>
<th>Range Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>pH</td>
<td>Low</td>
<td>[0, 0, 7]</td>
</tr>
<tr>
<td></td>
<td>Neutral</td>
<td>[6, 7, 8]</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>[7, 14, 14]</td>
</tr>
<tr>
<td>Hardness</td>
<td>Low</td>
<td>[0, 0, 50]</td>
</tr>
<tr>
<td></td>
<td>Medium</td>
<td>[40, 70, 100]</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>[80, 150, 150]</td>
</tr>
<tr>
<td>Solids</td>
<td>Low</td>
<td>[0, 0, 500]</td>
</tr>
<tr>
<td></td>
<td>Medium</td>
<td>[400, 950, 1500]</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>[1000, 3000, 3000]</td>
</tr>
<tr>
<td>Chloramines</td>
<td>Low</td>
<td>[0, 0, 5]</td>
</tr>
<tr>
<td></td>
<td>Medium</td>
<td>[3, 5, 7]</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>[5, 10, 10]</td>
</tr>
<tr>
<td>Sulfate</td>
<td>Low</td>
<td>[0, 0, 50]</td>
</tr>
<tr>
<td></td>
<td>Medium</td>
<td>[40, 95, 150]</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>[120, 300, 300]</td>
</tr>
<tr>
<td>Conductivity</td>
<td>Low</td>
<td>[0, 0, 100]</td>
</tr>
</tbody>
</table>

Table 2. Membership Function and Range Fuzzyness
The formulation of the increasing linear membership function on fuzzy can be found in the following equation \(1\) [17]:

\[
\mu(x, a, b) = \begin{cases}
0, & \text{if } x \leq a \\
\frac{x - a}{b - a}, & \text{if } a \leq x \leq b \\
1, & \text{if } x \geq b
\end{cases}
\]

The formulation of the decreasing linear membership function on fuzzy can be found in the following equation \(2\) [17]:

\[
\mu(x, a, b) = \begin{cases}
0, & \text{if } x \leq a \\
\frac{b - x}{b - a}, & \text{if } a \leq x \leq b \\
1, & \text{if } x \geq b
\end{cases}
\]

c. Inference Engine

The inference engine step in the combination of fuzzy and Naïve Bayes involves using fuzzy rules to make inferences and model the relationship between input variables and output variables. This is done by combining the membership levels of the input fuzzy sets with predetermined fuzzy rules. Common steps in an inference engine, such as: defining fuzzy rules, output fuzzification, combinations of fuzzy rules, and defuzzification. Convert the final output membership level to a single numeric value that can be used for class labeling. The fuzzy inference engine used in this research provides the minimum inference. The following is the formulation of fuzzy minimum inference in the following equation \(3\) [17]:

\[
\mu_{B}^{l}(y) = \frac{\max_{z} \left\{ \sup_{x \in U} \min(\mu_{A_1}^{l}(x), \ldots, \mu_{A_n}^{l}(x), \mu_{B_1}^{l}(x)) \right\}}{\mu_{B_1}^{l}(x)}
\]

d. Defuzzification

The defuzzification step is the final step for processing the input in a fuzzy combination with Naïve Bayes, where the output membership level obtained from the inference engine is converted into a single numeric value that can be used for class labeling. General steps in defuzzification, such as determining the output fuzzy set, output membership level aggregation, and defuzzification. Convert the aggregated output membership levels to a single numeric value. The
The defuzzification method used is the centroid, where this method calculates the midpoint (centroid) of the area covered by the output membership level. The centroid value is calculated based on the area and position of the center point of each fuzzy set. The crisp output value using the formula 4 below [18]:

\[
    z = \frac{\sum_{j=1}^{q} z_j u_c(z_j)}{\sum_{j=1}^{q} u_c(z_j)}
\]

(4)

Where, \(z\) is the output crisp value, \(q\) is the number of rules and \(u_c\) is the membership in class \(c\) at value \(z_j\).

e. Naïve Bayes

Naïve Bayes is one of the most popular classification algorithms in machine learning. This algorithm is based on Bayes' theorem with the assumption that each feature (variable) in the data is independent from one another [19]. Although this assumption is often incorrect in reality, Naïve Bayes still often performs well in various classification tasks. Basically, Naïve Bayes uses conditional probabilities to calculate the probability of the target class (output) based on the given feature values. In the context of classification, the Naïve Bayes algorithm will learn the probability distribution of each class in the training data set and then use this information to predict the target class in data that was not seen before.

Fuzzy logic, on the other hand, is a mathematical approach that allows for handling uncertainty and ambiguity in decision-making. Fuzzy logic replaces traditional binary logic by allowing membership levels or partial truths in a continuous range from 0 to 1. The classification similarity with the Naïve Bayes algorithm can be seen in equation 5 below [20]:

\[
P(H|X) = \frac{P(X|H) \cdot P(H)}{P(X)}
\]

(5)

Where:
\(P(H|X)\): Probability of Hypothesis \(H\) based on Condition \(X\) (posterior probability)
\(P(X|H)\): The probability of \(X\) is based on the conditions in the \(H\) hypothesis
\(P(H)\): Probability hypothesis \(H\) (probability prior)
\(P(X)\): \(X\) probability

The combination of Naïve Bayes and fuzzy logic can occur in several ways. One common approach is to use fuzzy methods to deal with uncertainty in data attributes. In Naïve Bayes, attributes are often thought of as binary variables, but in some cases, attributes can have uncertainty or ambiguity in their values. In this context, the fuzzy method can be applied to replace binary values with membership values in a certain range. This can help improve the quality of Naïve Bayes predictions by taking into account the uncertainty in the attributes.

The Naïve Bayes algorithm is combined with fuzzy logic, where attribute variables can be defined as fuzzy sets with relevant membership function. Then, the target class probability can be calculated by calculating the membership value...
of the fuzzy attribute. The combination of Naïve Bayes with fuzzy logic can help overcome some of the weaknesses of Naïve Bayes in dealing with uncertainty and ambiguity in data. However, keep in mind that this approach depends on the context and characteristics of the problem at hand, and its implementation may vary depending on the goals and specific requirements of the application being built.

f. Evaluation

Evaluation in the context of a combination of fuzzy and Naïve Bayes has a similar purpose to evaluation in ordinary Naïve Bayes. The main goal is to measure the performance of the model, understand its strengths and weaknesses, and ensure that the model can generalize well to data that has not been seen before. Some of the evaluation metrics commonly used in this context include:

- **Accuracy**: measures how well the model can correctly classify data.
- **Precision (Precision)**: measures the proportion of data that is classified correctly from all data that is classified as a certain class.
- **Recall (Recall)**: Measures the proportion of data that is classified correctly from all data that should be classified as a certain class.
- **F-Score**: measures the fine balance between precision and recall.

C. Result and Discussion

a. Results of Naïve Bayes Classifier

In the first test, this study tested water potability using the Naïve Bayes algorithm without making changes to the input parameters. The initial experiment was carried out on a dataset of 1000 with a 70:30 comparison between training and testing data. The experimental results obtained were accuracy, precision, recall, and F-score. The test results can be seen in Table 3 below.

<table>
<thead>
<tr>
<th>Evaluation Name</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>84%</td>
</tr>
<tr>
<td>Precision</td>
<td>78%</td>
</tr>
<tr>
<td>Recall</td>
<td>76%</td>
</tr>
<tr>
<td>F-Score</td>
<td>77%</td>
</tr>
</tbody>
</table>

Based on the test results using the Naïve Bayes algorithm, the percentage of accuracy is 84%, precision is 78%, recall is 76%, and the F-score is 77%. This shows that the applied algorithm has good results in determining the potability level of drinking water based on input parameters. Furthermore, the parameter results from this experiment will be stored and compared with the proposed method to see the increase in the results of the tests carried out.

b. Results of the Proposed Method

In the second test, this research tested the water potability using a combination of fuzzy and Naïve Bayes by changing the input parameters. This experiment was carried out on a dataset of 1000 with a comparison between training and testing data of 70:30. The experimental results obtained were accuracy, precision, recall, and F-score, which can be seen in Table 4 below.
Table 4. Results of the Proposed Method

<table>
<thead>
<tr>
<th>Evaluation Name</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>90%</td>
</tr>
<tr>
<td>Precision</td>
<td>94%</td>
</tr>
<tr>
<td>Recall</td>
<td>79%</td>
</tr>
<tr>
<td>F-Score</td>
<td>84%</td>
</tr>
</tbody>
</table>

Based on the test results using the Naïve Bayes algorithm, the percentage of accuracy is 90%, precision is 94%, recall is 79%, and the F-score is 84%. This shows that the algorithm applied has good results in determining the level of water potability. The combination of fuzzy and Naïve Bayes has the potential to provide high accuracy because both have complementary strengths and complementarities.

The results obtained from each test parameter, the proposed method obtains better results in determining the suitability level of drinking water. The results obtained can be seen in Figure 2 below.

![Figure 2. The Comparison of Research Testing Parameters](image-url)

On Figure 2, the increase in accuracy using the proposed algorithm is 6%, precision is 16%, recall is 3%, and the F-score is 7%.

Naïve Bayes is one of the popular classification methods in machine learning, has good results in determining the potability level of drinking water. It is based on the “naïve” assumption that all the features in the dataset are independent of one another. However, there are situations where these assumptions do not always meet real-world conditions, particularly when there is uncertainty or ambiguity in the data. This is where the fuzzy approach can be applied to improve the performance of the Naïve Bayes method. In some cases, the attributes in the dataset may have uncertainty or a degree of membership in various classes or categories. The fuzzy approach allows us to describe this level of uncertainty using fuzzy sets and degrees of membership. By using fuzzy methods, we can handle uncertainty and ambiguity in data, which can improve classification performance.
This approach allows us to consider the relationship between the parameters that affect water potability, resulting in a more realistic probability value. Thus, this hybrid approach provides more in-depth information about the potability level of water, which can be used for better decision-making in the management of water resources. In addition, the fuzzy-Naïve Bayes hybrid approach is also able to identify the factors that have the most influence on water potability. This provides valuable insight into efforts to improve water resource management and ensure a safe, high-quality water supply. In some cases, the available data may be incomplete or may not have sufficient information to classify correctly. By using a fuzzy approach, we can model incomplete information and combine it with existing information to produce more accurate and informative decisions. This can help improve the predictive ability of the Naïve Bayes model.

The combination of fuzzy-Naïve Bayes is able to handle classes in datasets that overlap or have blurred boundaries between each other. This can lead to difficulties in classifying correctly using the simple Naïve Bayes method. Using the fuzzy approach, we can describe fuzzy class boundaries using overlapping fuzzy sets. This allows us to better classify data that may lie between overlapping classes. In addition, outliers and noise are common problems in datasets that can affect the performance of classification models. By using a fuzzy approach, we can reduce the impact of outliers and noise by considering the degree of membership of the data in the relevant fuzzy sets. This helps make the Naïve Bayes model more robust against unexpected or unusual data. Naïve Bayes enhancement with a fuzzy approach is not always necessary or appropriate in every case. The decision to use this approach should be made based on the nature of the dataset and the problem at hand.

D. Conclusion

The results showed that the fuzzy-Naïve Bayes hybrid approach significantly increased the accuracy and effectiveness of water potability assessment compared to conventional methods. By using a fuzzy approach, we can overcome uncertainty and ambiguity in the initial data. Then, the Naïve Bayes approach is used to classify water samples based on the probabilities generated by the fuzzy system. The application of the fuzzy-Naïve Bayes hybrid approach in assessing water potability shows accurate results, where the percentage of accuracy is 90%, precision is 94%, recall is 79%, and the F-score is 84%. Thus, the fuzzy-Naïve Bayes hybrid approach makes a significant contribution to improving water potability assessments and has the potential to become a valuable tool for better water management and more effective decision-making.

E. Acknowledgment

On this occasion, we would like to thank Universitas Medan Area for helping and contributing to the creation of this joint research.

F. References


