Valuation K-Nearest Neighbors and Naïve Bayes for Drinking Water Potability Classification

Anisa Rahmawati1*, Muhamad Fatchan2, Wahyu Hadikristanto3
Universitas Pelita Bangsa
Corresponding Author: Anisa Rahmawati rahmawatianisa2000@gmail.com

ARTICLE INFO
Keywords: Drinking Water; K-Nearest Neighbors, Naïve Bayes, Machine Learning

Received: 03, April
Revised: 16, April
Accepted: 22, May

©2024 Rahmawati, Fatchan, Hadikristanto: This is an open-access article distributed under the terms of the Creative Commons Atribusi 4.0 Internasional.

ABSTRACT
The availability of drinking water that is safe and suitable for consumption is important to support health and development. This research emphasises the importance of handling the clean water crisis through the evaluation of drinking water quality using data mining algorithms. The drinking water quality evaluation method was selected using the K-Nearest Neighbors and Naïve Bayes algorithms, replacing the manual method which is less responsive in predicting. The experimental process was conducted by utilising Kaggle website data by applying data processing and oversampling techniques to handle class imbalance in the dataset used. Bases on the research results, the accuracy of the K-Nearest Neighbors Algorithm reaches 65%, which is higher than the accuracy of the Naïve Bayes Algorithm which is 64%. So it can be concluded that the K-Nearest Neighbors Algorithm is more effective in predicting the quality of water suitable for consumption. This research provides an in-depth insight into the use of technology and data analysis in dealing with the crisis in the availability of water suitable for consumption and offers suggestions for further research using more diverse methods and the use of more datasets to improve accuracy in evaluating the quality of potable water.
INTRODUCTION

An indispensable essential for the survival of all creatures is drinking water, which plays an important role in maintaining health and bodily functions, especially for humans. The availability of drinking water that is safe and suitable for consumption is not only important for maintaining health, but also as a support for sustainable economic and social development (B. K. Mishra et al., 2021). Around 2.2 billion people worldwide still face challenges in gaining access to drinking water that meets safety standards (Salehi, 2022). The clean water crisis is a very worrying problem around the world, where the amount of clean water that can be consumed by humans is only 1%. According to WHO data, 633 million people have difficulty accessing clean water (Riyantoko et al., 2021). Water shortages are predicted to be faced by two-thirds of the world’s population by 2025 according to estimates from the World Wide Fund (WFF). This situation has the potential to increase the suffering of the global ecosystem (Arora & Mishra, 2022). According to data from UNESCO, individuals living in areas experiencing water shortages number around 2 billion, while those without adequate access to safe drinking water sources account for more than 800 million people (Makarigakis & Jimenez-Cisneros, 2019).

Factors such as climate change, population growth, unsustainable land use, development, industrial and domestic effluents pose a serious threat to the availability of clean water quality (R. K. Mishra, 2023). Expanding regard for drinking water quality issues is boundless and worldwide because of the potential contamination that can be brought about by different harmful substances like pesticides, pathogenic microorganisms, heavy metal, and other synthetic blends than can compromise the water supply required by the human body (P. Zhang et al., 2023). According, great administration and control are expected to protect the idea of clean water so the nature of water can be guaranteed to be appropriate for human use (Tangkelayuk, 2022). Assessment of drinking water quality is often done through manual laboratory testing which is time-consuming and less responsive in real time. With the development of technology especially in the field of machine learning comes the option to evaluate drinking water quality automatically, quickly and accurately (Rambe et al., 2024). Artificial intelligence (AI) methods such as machine learning are considered as effective tools for monitoring, data management and policy making in the context of water quality (Lowe & Qin, 2022).

Drinking water quality is measured by various variables and parameters that are used as the basis for model building such as pH, turbidity, metal content etc. (Shaibur et al., 2024). Classification is a stage of data pattern analysis that aims to identify the appropriate class or category in an unrecognised object based on the characteristics of the observed features (Ainurrohma, 2021). A few ordinarily utilized information characterization methods incorporate the K-Nearest Neighbors and Naive Bayes calculations which produce shifting exactness. The K-Nearest Neighbors calculation is most popular for its capacity to track down the nearest distance between the assessed
Informations and the closest neighbor in the preparations informational index (Saadatfar et al., 2020). In the mean time, the Naive Bayes calculation is utilized to work out the likelihood of each class and decide the most elevated likelihood worth to characterize the test information (Sendari et al., 2020). The use of algorithms in assessing the quality of potable water is increasingly attractive because it can provide effective and efficient solutions in facing global challenges related to the clean water crisis that is suitable for consumption (Ghina Annaifah, 2024). This research was conducted with the aim of analyzing and evaluating the performance of both algorithms in identifying the quality of drinking water. Through this research, it is expected to find the most effective method in achieving maximum accuracy in identifying the quality of drinking water.

LITERATURE REVIEW

Potable water must meet certain quality standards to be safe for consumption. The quality of drinking water can be influenced by several factors such as pollution, water sources and water treatment. Parameters used in assessing the quality of potable water include physical, chemical and microbial parameters such as colour, temperature, pH, turbidity, mercury, iron, metals, E.coli bacteria and many more (Silva et al., 2022). Data mining is the most common way of extricating data to acquire new bits of knowledge. This research utilises data mining techniques by implementing the K-Nearest Neighbors and Naïve Bayes methods to compare the most optimal accuracy levels of the two methods. Data mining is a process that involves utilizing data to detect relationships or patterns in large data sets with the goal of gaining new insights that may not have been revealed previously (Sree & Vardhani, 2015). Therefore, the development of effective classification models to assess water quality is very important in an effort to maintain public health (Park et al., 2020).

Classification is an important aspect of machine learning that aims to categorize data into specific groups. To create a model that may be used to categorize newly unlabeled data the process entails learning from an existing labeled collection of data. When it comes to treating drinking water resources, classification is used to identify based on physical, chemical and biological measurement characteristics. This allows for the quick and accurate decision making process (Zainurin et al., 2022).

The K-Nearest Neighbors (KNN) method is often the first choice in classification due to its similarity to the nearest neighbor class. However, its main drawback lies in its dependence on the K parameter, which can significantly affect classification results and its sensitivity to outlier data. So it is necessary to adjust the K parameter to improve optimal accuracy results (Wang et al., 2020). The basic principle of K-nearest Neighbors (KNN) is to find the closest data to the evaluation data based on the K nearest neighbors in the training dataset. Before searching for the closest distance, the K-Nearest Neighbors algorithm needs to do preprocessing or normalisation first which aims to equalise the standard values on all attributes or indicators used in the
calculation (Song et al., 2022). The application of KNN to drinking water quality has an advantage in its ability to handle small and simple datasets very effectively (Juna et al., 2022).

Categorization approaches using Naïve bayes techniques based on Bayes principle are often used in various situations, including water quality assessment. Naïve Bayes assumes that all independent variables in the data have no relationship with each other. Thus, the measurement parameters can be studied separately, which speeds up the classification process. The drawback of Naïve Bayes lies in the dependence on the availability of consistent and good data to provide reliable results (Ilić et al., 2022). The advantage of using Naïve Bayes lies in the need for little training data to determine the mean and variance parameters of the variables required for classification (Chen et al., 2021).

**METHODOLOGY**

This research uses an experimental and evaluation approach model that aims to compare and assess the effectiveness of data mining classification algorithms in analysing various aspects related to water quality.

![Research flow diagram](Figure 1. Research flow)

**Data Collections**

Utilizing the Google Colab platform, analysis and processing were performed using the Python programming language. The scikit-learn library was used for machine learning, while pandas was used for data manipulation. The first step was to collect data from a public source Kaggle which contains various water samples with their quality parameters. Each sample acts as an
attribute reflecting its reature while the ‘potability’ column serves as the target variable the predicts whether the water is safe for consumptions or not.

**Data Preprocessing**

The next process is data processing which consists of data cleaning, overcoming missing values by using a feature selection approach such as calculating the average to fill in the empty values with existing data, removing duplicate data which aims to ensure consistency and cleanliness of the data before proceeding to the next stage (Y. Zhang & Thorburn, 2022). In addition, class balancing is carried out using oversampling techniques to handle the imbalance of data classes that can improve the performance of the model to be tested. This is because of the inclination of the model to lean toward the larger part class which has more examples contrasted with the minority class. Thusly, an oversampling procedure where tests from the minority class are imitated to offset their number with the greater part class is required (Wongvorachan et al., 2023). Data visualisation uses bar charts to understand patterns and characteristics in the data.

**Data Distribution**

The next step is to divide the dataset into 20% testing data and 80% training data to train the model. In the process of data analysis and modelling we divide the dataset into two main part training data and testing data. This division aim to ensure that the developed model can be evaluated objectively on data that is has never seen during the training process. The training data is used to train the model where the model will learn the data to recognise patterns and make predictions. While the testing data is used to evaluate the performance of the model after the training process is complete. It gives an indication of how well the model can predict new data that is has never seen before. The trained model will be tested with the test data. The predicted results will be compared with the actual values to measure how well the model works.

**Evaluation Model**

The next step is to train the model using the K-Nearest Neighbors (KNN) and Naïve Bales algorithms, the evaluate the performance of both models. The evaluation is done using several metrics such as confusion matrix, accuracy, precision, recall and F1 score. Confusion matrix provides a detailed overview of how the model predicts each class to see the number of correct and incorrect predictions for each class. Accuracy measure how often the model makes correct predictions. Precision measure the accuracy of positive predictions. Recall measure the ability of the model to fins all positive examples. F1 score provides a balance between precision and recall, especially useful when there is an imbalance of classes in the dataset.
RESULTS
The data source is taken from the kaggle website https://www.kaggle.com/datasets/uom190346a/water-quality-and-potability/data with the name ‘water_potability’ which is csv data type.

Table 1. Potability data

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>pH</td>
<td>The pH level of the water. The optimum pH required ranges from 6.5 to 8.5</td>
</tr>
<tr>
<td>Hardness</td>
<td>Water hardness, a measure of mineral content</td>
</tr>
<tr>
<td>Solids</td>
<td>Total dissolves solid in water</td>
</tr>
<tr>
<td>Chloramines</td>
<td>Cholramine concentration in water</td>
</tr>
<tr>
<td>Sulfate</td>
<td>Sulfate concentration in water</td>
</tr>
<tr>
<td>Conductivity</td>
<td>Electrical conductivity of water</td>
</tr>
<tr>
<td>Organic Carbon</td>
<td>Organic compounds dissolved in water</td>
</tr>
<tr>
<td>Trihalomethanes</td>
<td>The concentration of trihalomethanes in the water</td>
</tr>
<tr>
<td>Turbidity</td>
<td>Turbidity level, a measure of water clarity</td>
</tr>
<tr>
<td>Potability</td>
<td>Target variable, indicates the potability of water with values of 1 (potable) and 0 (not potable)</td>
</tr>
</tbody>
</table>

Data analysis is needed to assess the feasibility of the data before proceeding to the next stage. There were data blanks in the NaN format, so it was necessary to fill in the blank values using the average value of each attribute.
Histogram graphs are presented as an overview of various measurement parameters in water that show how often the values of the parameters appear in the data.
Understanding the distribution of the metrics is essential to improve the performance of the prediction model. The figure above helps detect class imbalance in the dataset. The next step involves dividing the water quality data represented in the potability column in the dataset using exploratory data analysis (EDA) which shows the difference that out of a total of 3276 data there is a total of 1998 non-potable data and 1278 potable data. The percentage indicates that 61% of the total data is non-potable while 39% is potable.

Class balancing on the data is done after passing the data cleaning process that can affect the performance of the model. By using an oversampling technique that functions to duplicate the majority class sample (0) which has 1998 data records to be balanced with the minority class sample (1) so that it has the same number of data records.
Data sharing was done with an allocation of 80% for model training and 20% for performance evaluations using test data. The number of training data entries was 2620, while the testing data had 656 entries. The modeling process involved two machine learning algorithms that produced different values.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Evaluation</th>
<th>KNN</th>
<th>Label</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
<th>Support</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>0.65</td>
<td>0.64</td>
<td>0.76</td>
<td></td>
<td>396</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>0.64</td>
<td>0.25</td>
<td>0.36</td>
<td></td>
<td>260</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Naïve Bayes</th>
<th>0</th>
<th>0.64</th>
<th>0.91</th>
<th>0.75</th>
<th>396</th>
<th>0.64</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bayes</td>
<td>1</td>
<td>0.63</td>
<td>0.22</td>
<td>0.33</td>
<td>260</td>
<td></td>
</tr>
</tbody>
</table>

**DISCUSSION**

Based on the research results that have been presented previously, it shows that the K-Nearest Neighbors and Naïve Bayes algorithms provide fairly good predictions in evaluating the quality of potable water. However, it should be noted that classification performance can be affected by several factors such as parameter settings and model fit with the data used. The K-Nearest Neighbors algorithm tends to provide more accurate results, while Naïve Bayes shows a tendency to be faster in the classification process. Therefore, it is important to carefully consider the advantages and disadvantages of each algorithm according to the specific needs in assessing drinking water quality effectively and optimally.

**CONCLUSIONS AND RECOMMENDATIONS**

Based on the study’s findings, it can be said that the K-Nearest Neighbors Algorithm outperforms the Naïve Bayes Algorithm in terms of accuracy. The K-Nearest neighbors algorithm achieved an accuracy rate of 65%, while the Naïve Bayes algorithm achieved an accuracy rate of 64%.
Therefore, it can be concluded that the K-Nearest Neighbors Algorithm is superior in modeling relationships and patterns in the tested data. The use of the K-Nearest Neighbors Algorithm by implementing oversampling and class balancing techniques results in a classification that is good enough to be applied to the case of drinking water quality classification. Suggestions for future research can use more methods in classification and the use of wider and more diverse data to increase the level of accuracy that is more optimal.

FURTHER STUDY
This research actually has constraints so further examinations is required for a more inside and out investigation of the point ‘Evaluation of K-Nearest Neighbors and Naïve Bayes for Drinking Water Quality Classification’.

ACKNOWLEDGMENT
In this study, the authors would like to thank the Supervisor who has directed so that this research can be completed. The author would also like to thank his parents for their support and encouragement. In addition, the author would like to express his deepest appreciation to colleagues who have provided advice and input that is very helpful in compiling this paper. Support and assistance from various parties are very meaningful to the author in completing this research.

REFERENCES


Park, J., Kim, K. T., & Lee, W. H. (2020). Recent advances in information and communications technology (ICT) and sensor technology for monitoring water quality. Water (Switzerland), 12(2). https://doi.org/10.3390/w12020510


