Machine Learning Approaches for Fish Pond Water Quality Classification: Random Forest, Gaussian Naive Bayes, and Decision Tree Comparison

Danuri[1], Muhammad Syafiq Mohd Pozi[2]
{danuri@polbeng.ac.id[1], syafiq.pozi@uum.edu.my[2] }

Politeknik Negeri Bengkalis, 28711, Indonesia[1], Universiti Utara Malaysia, Sintok 06010, Kedah, Malaysia

Abstract. The health and production of fish in fish farms are greatly influenced by the water quality. This study examines three Machine Learning(ML) methods for categorizing fish pond water quality: Random Forest(RF), Gaussian Naive Bayes(GNB), and Decision Tree(DT). Accuracy, precision, recall, and the F1-Score as a performance indicator are taken into account while evaluating the model. The evaluation findings reveal that RF and GNB outperform DT in every evaluation criteria. GNB, with a rating of 0.958932, had the highest accuracy, followed by RF, with a value of 0.955822, and DT, with a value of 0.932269. The consistent performance of GNB and RF in precision, recall, and F1-Score underscores their superiority.

Keywords*:* Fish Pond Water Quality, Classification, Machine Learning

1. Introduction

Monitoring fish pond water quality is crucial for maintaining aquatic resources and controlling fisheries cultivation, since it promotes fish health and production. Poor water quality can have serious impacts on fish health and production efficiency [1]. This is a potential attraction for research, for example in terms of IoT technology, as carried out by [2] [3] [4]. Because poor water quality can be one of the causes of death in fish [5] conducted research aimed at reducing fish loss in Indonesian mariculture by using ML models to predict mass deaths. The results show that this model is effective with an average accuracy of 0.699 and can reduce fish losses by up to 59.7%. rana

Another study conducted by [6] investigated the use of ML techniques in predicting fish pond water quality with a focus on limited measurement scenarios. The results show that random forest is effective in predicting water parameters with just two daily measurements, and can be implemented via smartphone for fish farmers. Research conducted by [7] used various ML methods, including neural networks, support vector machine, k-nearest neighbors, logistic regression, GNB, DT, RF, and AdaBoost, to classify ponds with high and low harvest performance based on water quality variables. The results showed that dissolved oxygen, salinity, and temperature had the greatest influence on crop yield, with late-season changes in dissolved oxygen and salinity and post-stocking temperature variations being the main factors

ABEC 2023, September 21, Bengkalis, Indonesia

DOI 10.4108/eai.21-9-2023.2342964

in differentiating ponds with high and low yields. Research conducted by [8] built an architecture and model for an IoT system that uses deep learning and the Long-Short Term Memory (LSTM) algorithm to forecast water quality.

The use of ML in water quality monitoring has created new potential in the contemporary information technology era. Several ML approaches include Random Forest as used by [9] to develop predictions of the risk of fish death the next day based on water quality data and daily records of fish mortality in mariculture areas in Gondol, Bali, Indonesia. The research results showed that the best prediction accuracy was obtained using 3-day moving average data from water quality attributes, with an average accuracy of 74.92%. Feature analysis revealed that sea water temperature, salinity, and turbidity were important factors in predicting fish mortality in the area. Another ML approach is GNB as used by [10] to predict potential milkfish harvests based on pond environmental conditions. The results show an accuracy rate of 94.44%, close to 100%. This method attempts to address the issue of inconsistencies between anticipated and actual production outcomes by taking into account a variety of variables, including as the quantity of seeds, pond size, and water quality parameters like pH and temperature. Another ML technique is DT, as demonstrated by [11] utilizing the Decision Tree Regression(DTR) algorithm to create a reliable aquarium control system.

This article explains and contrasts three machine learning (ML) methods for categorizing fish pond water quality, namely RF, GNB, and DT.

2 Research Methods

This study applies the ML approach to classify water quality in fish ponds (as a sample of tilapia). The focus of the research is to compare the effectiveness of the three methods used, namely RF, GNB, and DT. To determine the best approach for categorizing water quality measures, such as dissolved oxygen, temperature, and pH, in the context of fish farming, each method was tested in this study.

Finding variables is the first step in the research process. Next, data collection and labeling are done to create a dataset. Start the data preparation stage using the dataset you acquired. Training data and testing data will be created from the outcomes of the preprocessing. Figure 1 illustrates the phases of the investigation.

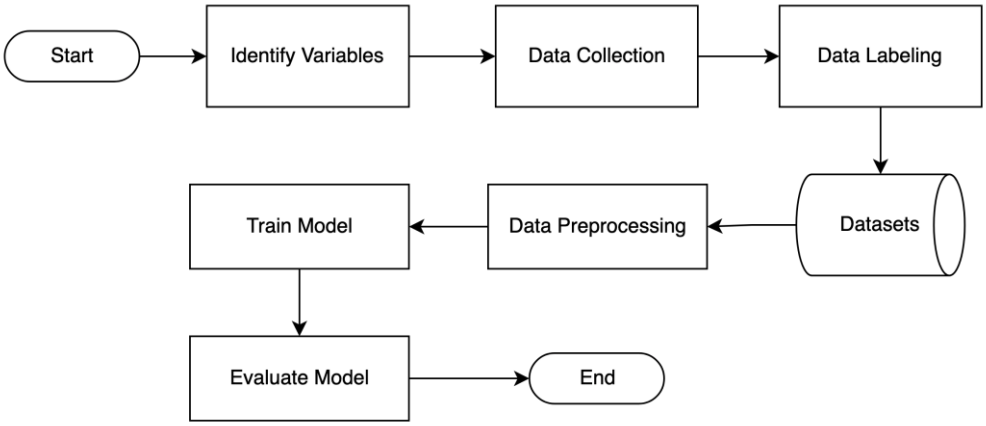

Fig. 1. A flowchart for a research approach.

1.1. Identify Variables

The following variables are used to assess the water quality in fish ponds: temperature (°C), dissolved oxygen (mg/L), pH, brightness (cm), and ammonia (mg/L). As a reference standard refers to the Indonesian National Standard for the production of black tilapia (Oreochromis niloticus Bleeker) seeds with SNI number 7550:2009 [12] as shown in Table 1.

Table 1. Requirements for water quality

No.	Parameter	Unit	Range
1	Temperature	^{0}C	25 – 32
2	pH	-	6,5 – 8,5
3	Dissolved-oxygen	mg/l	>= 3
4	Ammonia (NH_3)	mg/l	< 0,02
5	Brightness	cm	30 - 40

1.2. Data Collection

Data collection for fish pond water quality used was 82.200 records with parameters such as temperature, pH, dissolved-oxygen, brightness, and ammonia. Figure 2 depicts how the data are distributed.

```
RangeIndex: 82200 entries, 0 to 82199
Data columns (total 8 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   id                      82200 non-null  int64
 1   created_at              82200 non-null  datetime64[ns]
 2   temperature(°C)         82180 non-null  float64
 3   dissolved-oxygen(mg/L)  82176 non-null  float64
 4   pH                      82182 non-null  float64
 5   brightness(cm)          82183 non-null  float64
 6   ammonia(mg/L)           82183 non-null  float64
 7   level                   82200 non-null  object
dtypes: datetime64[ns](1), float64(5), int64(1), object(1)
```

Fig. 2. Collections of data are distributed.

1.3. Data Labelling

From the data that has been collected, data labelling is carried out for classification of water quality whether normal or harmful, then becomes an original dataset. The results of the labelling data are 72.100 data in the normal water quality category, and 10.100 data in the harmful quality category, as shown in Figure 3.

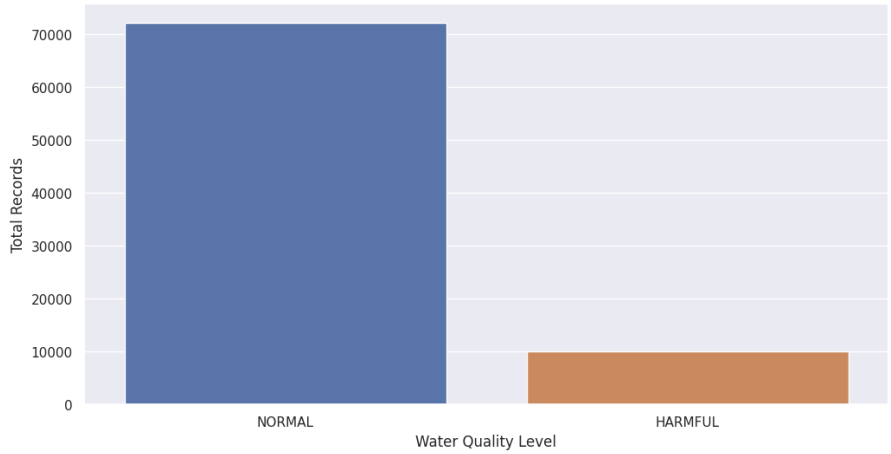

Fig. 3. Classification of water quality as a result of labeling data.

1.4. Data Preprocessing

Preprocessing of data is very important in ensuring the validity of data analysis. This stage is a crucial foundation in the development of operational data analysis, considering the complexity inherent in operational design as well as potential deficiencies in the quality of the data used [13]. In order to meet the requirements of the ML method to be used, data preparation entails a number of processes for cleaning, data balancing, labeling encoding, and data separation. Figure 4 illustrates the phases of data preparation that were completed.

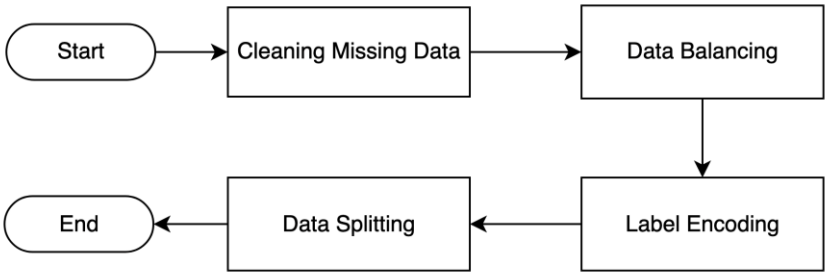

Fig. 4. Phases of data preparation

At the data preprocessing stage, the first step that must be taken is to clean the data from infinite values, zeros, and values that are considered abnormal. The results of data cleaning were 70.767 normal records and 9.917 harmful records. After cleaning the data, continue with data balancing. The results obtained were 14.153 normal records, and 9.917 harmful records. The graph in Figure 5 shows the outcomes of balancing the data.

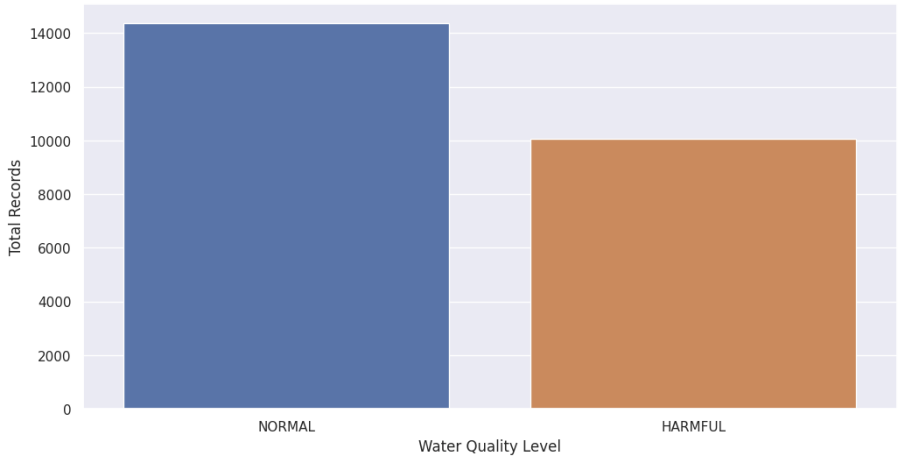

Fig. 5. Distribution of balancing data results

After balancing the data, proceed with labeling the data, where normal is labeled 1, while harmful is labeled 0. Data that has been labeled is followed by split data for test and train as the final stage of data preprocessing. For model training needs, the data is divided into two parts, namely training data of 19,545 records and test data of 4,887 records.

3 Result and Discussion

3.1 Training Model

The model training stage in ML is carried out to teach models such as RF, GNB, and DT to be able to recognize different water quality patterns in fish ponds. RF is a machine learning strategy that increases prediction accuracy by randomly constructing several decision trees. The training

dataset is sampled at random, and random characteristics are chosen while creating each tree. The findings of all decision trees are combined to provide the final forecast result. This approach reduces overfitting and produces more accurate and stable predictions [14]. GNB is a Bayes Theory-based categorization method. [15] and characterized by the assumption of "naive," in which all variables are considered conditionally independent. This algorithm's primary benefit is its capacity to learn parameters individually, making it a simple and efficient choice to use [15] [16]. DT is a supervised learning model that divides the variable area into two for each branch and categorizes data based on certain criteria. [17]. This algorithm works by dividing the initial dataset into smaller subsets recursively, based on certain tests carried out on each node in the DT [18].

Pre-processed water quality data is used as input, and known variables are used as targets or classification labels. The models process training data, learn from the relationship between input and target variables, and then generate rules or decisions to classify water quality. This procedure entails adjusting model performance-enhancing parameters, such as the Random Forest's tree density, as well as choosing evaluation metrics, such as accuracy, precision, recall, and F1-score, to gauge output and compare the three models' classification of fish pond water quality. Following are the equations for accuracy, precision, recall, and f1-score [19] [20]:

$$accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (1)$$

$$precision = \frac{TP}{TP+FP} \quad (2)$$

$$recall = \frac{TP}{TP+FN} \quad (3)$$

$$f1-score = \frac{2\ x\ precision\ x\ recall}{precision+recall} \quad (4)$$

where FP stands for false positive, TP for true positive, TN for true negative, and FN for false negative.

Table 2 displays the outcomes of training data using a DT.

Table 2. The results of the DT training data

Iteration	Accuracy	Precision	Recall	F1-Score
1	0.932679	0.932705	0.932679	0.932691
2	0.939022	0.939094	0.939022	0.939052
3	0.937385	0.937477	0.937385	0.937421
4	0.932474	0.932519	0.932474	0.932494
5	0.934725	0.934866	0.934725	0.934778
6	0.926335	0.926578	0.926335	0.926421
7	0.933702	0.933736	0.933702	0.933717
8	0.926540	0.926510	0.926540	0.926523
9	0.930428	0.930357	0.930428	0.930377
10	0.929405	0.929356	0.929405	0.929374

Table 3 displays the outcomes of training data using RF.

Table 3. Results of RF training data

Iteration	Accuracy	Precision	Recall	F1-Score
1	0.954983	0.955081	0.954983	0.954872
2	0.956415	0.956426	0.956415	0.956320
3	0.956210	0.956188	0.956210	0.956157
4	0.960098	0.960082	0.960098	0.960050
5	0.958870	0.958933	0.958870	0.958775
6	0.955187	0.955149	0.955187	0.955133
7	0.956210	0.956264	0.956210	0.956106
8	0.953550	0.953699	0.953550	0.953440
9	0.953959	0.954122	0.953959	0.953832
10	0.952732	0.952874	0.952732	0.952618

Table 4 displays the outcomes of training data using Gaussian Naive Bayes.

Table 4 shows the GNB training data results.

Iteration	Accuracy	Precision	Recall	F1-Score
1	0.958666	0.958723	0.958666	0.958582
2	0.959484	0.959476	0.959484	0.959413
3	0.958257	0.958323	0.958257	0.958169
4	0.961940	0.961947	0.961940	0.961883
5	0.960507	0.960578	0.960507	0.960416
6	0.959894	0.959864	0.959894	0.959851
7	0.961121	0.961167	0.961121	0.961039
8	0.956620	0.956734	0.956620	0.956528
9	0.959894	0.959995	0.959894	0.959804
10	0.952936	0.953044	0.952936	0.952832

3.2 Evaluate Model

The findings of the performance assessment of the three ML algorithms used in this study are shown in Table 5: RF, GNB, and DT. Table 5 provides information on the classification of each algorithm's F1-Score, recall, accuracy, and precision.

Table 5. Evaluate Model Results

Model	Accuracy	Precision	Recall	F1-Score
Decision Tree	0.932269	0.932320	0.932269	0.932285
Random Forest	0.955822	0.955882	0.955822	0.955730
Gaussian Naive Bayes	0.958932	0.958985	0.958932	0.958852

From the results of the model evaluation of the three algorithms listed in Table 5, several findings can be identified. First, the DT shows strong performance in the testing phase, with accuracy, precision, recall, and F1 Score values of 0.932269, 0.932320, 0.932269, and 0.932285, respectively. Second, RF and GNB show almost the same results in all measured parameters, including accuracy, precision, recall, and F1 Score. For RF the accuracy value is 0.955822, precision 0.955882, recall 0.955822, and F1-Score 0.955730. In contrast, GNB has an accuracy value of 0.958932, precision 0.958985, recall 0.958932, and F1-Score 0.958852.

This shows that RF and GNB are better able to recognize fish pond water quality patterns than DT. Second, GNB has higher accuracy and other metric values than RF. This shows that the GNB algorithm has the ability to classify more precisely and consistently. However, the difference in performance between the two algorithms is not very significant. Third, the evaluation results show that all algorithms have similar precision, recall and F1-Score values. This shows that they can properly recognize differences in water quality in ponds.

4 Conclusion

This study compares the effectiveness of three machine learning (ML) algorithms for categorizing fish pond water quality: DT, RF, and GNB. The evaluation results show that RF and GNB have better performance than DT in terms of accuracy, precision, recall and F1-Score. GNB has the highest accuracy of 0.958932 which shows better ability in predicting fish pond water quality, while RF has similar performance with almost comparable accuracy of 0.955822. A further experiment is to consider the use of various other ML algorithms to identify the most suitable algorithm for the task of classifying water quality in fish ponds. In addition, it is recommended to expand the dataset by adding more samples and features that include information on weather, seasons and other factors that have the potential to influence fish pond water quality, with the aim of increasing the level of reliability and validity of research results.

References

[1] Z. Tumwesigye, W. Tumwesigye, F. Opio, C. Kemigabo dan B. Mujuni, "The Effect of Water Quality on Aquaculture Productivity in Ibanda District, Uganda," Aquaculture Journal, pp. 23-36, 2022.

[2] N. Ya'acob, N. N. S. N. Dzulkefli, A. L. Yusof, M. Kassim, N. F. Naim dan S. S. M. Aris, "Water Quality Monitoring System for Fisheries using Internet of Things (IoT)," dalam International Conference of Emerging Challenges in Engineering and Current Technology (ICECTIII 2021), Terengganu - Malaysia, 2021.

[3] I. S. a. M. A. R. F. Rozie, "Design and Implementation of Intelligent Aquaponics Monitoring System based on IoT," dalam International Electronics Symposium (IES), 2020.

[4] K.-L. Tsai, L.-W. Chen, L.-J. Yang, H.-J. Shiu dan H.-W. Chen, "IoT based Smart Aquaculture System with Automatic Aerating and Water Quality Monitoring," Journal of Internet Technology, vol. 23, no. 1, pp. 177-184, 2022.

[5] K. Inoue, R. Septory, H. Albasari dan M. Wada, "Mass Mortality Risk Prediction and Fish Loss Simulation in Mariculture," dalam Global Oceans 2020, Singapore, 2020.

[6] A. F. Zambrano, L. F. Giraldo, J. Quimbayo, B. Medina dan E. Castillo, "Machine learning for manually-measured water quality prediction in fish farming," PLoS ONE, vol. 16, no. 8, 2021.

[7] M. Rana, A. Rahman, J. Dabrowski, S. Arnold, J. McCulloch dan B. Pais, "Machine learning approach to investigate the influence of water quality on aquatic livestock in freshwater ponds," Biosystems Engineering, vol. 208, pp. 164-175, 2021.

[8] N. Thai-Nghe, N. Thanh-Hai dan N. C. Ngon, "Deep Learning Approach for Forecasting Water Quality in IoT Systems," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 11, no. 8, pp. 686-693, 2020.

[9] R. Saville, K. Hatanaka, A. Fujiwara, M. Wada, R. Puspasari, H. Albasari, N. Dwiyoga dan A. Muzaki, "A Mariculture Fish Mortality Prediction Using Machine Learning Based Analysis of Water Quality Monitoring," dalam OCEAN, Hampton Roads, VA, USA, 2022.
[10] K. Nisa, S. Armalia, O. Puspitorini, A. Wijayanti, N. A. Siswandari dan M. Milchan, "Prediction Of Milkfish Harvest Potential Based On Pond Environment To Support Smart Fishery Towards Technology 4.0," dalam International Electronics Symposium (IES), Denpasar, Indonesia, 2023.
[11] M. Abdurohman, A. G. Putrada dan M. M. Deris, "A Robust Internet of Things-Based Aquarium Control System Using Decision Tree Regression Algorithm," IEEE Access, vol. 10, 2022.
[12] Badan Standardisasi Nasional, Indonesian National Standard (SNI) 7550:2009 Production of Grow-out Class Tilapia (Oreochromis Niloticus Bleeker) in Still Water Ponds, Jakarta: BSN.
[13] C. Fan, M. Chen, X. Wang, J. Wang dan B. Huang, "A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data," Frontiers in Energy Research, vol. 9, 2021.
[14] J. Xu, Z. Xu, J. Kuang, C. Lin, L. Xiao, X. Huang dan Y. Zhang, "An Alternative to Laboratory Testing: Random Forest-Based Water Quality Prediction Framework for Inland and Nearshore Water Bodies," Water, vol. 13, 2021.
[15] M. Ilic, Z. Srdjevic dan B. Srdjevic, "Water Quality Prediction Based on Naïve Bayes Algorithm," Water Science & Technology, vol. 85, no. 4, 2022.
[16] O. K. Pal, "The Quality of Drinkable Water using Machine Learning Techniques," International Journal of Advanced Engineering Research and Science (IJAERS), vol. 9, no. 6, pp. 16-23, 2022.
[17] E. Cho, T.-W. Chang dan G. Hwang, "Data Preprocessing Combination to Improve the Performance of Quality Classification in the Manufacturing Process," Electronics, vol. 11, no. 3, 2022.
[18] N. Radhakrishnan dan A. S. Pillai, "Comparison of Water Quality Classification Models using Machine Learning," dalam the Fifth International Conference on Communication and Electronics Systems (ICCES 2020), India, 2020.
[19] U. Ahmed, R. Mumtaz, H. Anwar, A. A. Shah, R. Irfan dan J. Garcia-Nieto, "Efficient Water Quality Prediction Using Supervised Machine Learning," Water, vol. 11, no. 11, 2019.
[20] F. Muharemi, D. Logofatu dan F. Leon, "Machine learning approaches for Anomaly Detection of Waterquality on a Real-World Dataset," Journal of Information and Telecommunication, vol. 3, no. 3, pp. 294-307, 2019.