JOURNAL OF DATA SCIENCE | Vol.2024:48
eISSN:2805-5160

Comparison of Logistic Regression, Random Forest, SVM, KNN Algorithm for Water Quality Classification Based on Contaminant Parameters

Teguh Sutanto[1], Muhammad Rafli Aditya[2], Haldi Budiman[3], M.Rezqy Noor Ridha[4], Usman Syapotro[5], Noor Azijah[6]

[1,2,3,4,5,6] Faculty of Information Technology Universitas Islam Kalimantan Muhammad Arsyad Al-Banjari, Indonesia

Email: teguhsutanto17@gmail.com[1], muhrafliaditya061@gmail.com[2], haldibudiman@uniska-bjm.ac.id[3], rezqyridha@gmail.com[4] , 05usman.syapotro.study@gmail.com[5], noorazijah444@gmail.com[6]

Abstract

This study compares four machine learning algorithms Logistic Regression, Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN) in water quality classification based on contaminant parameters. The purpose of this study is to evaluate and compare the performance of these algorithms in terms of accuracy. The methodology used includes data collection, preprocessing, and algorithm implementation with evaluation using cross-validation techniques. The results showed that the application of the Stacking method with Gradient Boosting Meta-learner produced the highest accuracy of 96.00%, outperforming all other algorithms. In comparison, Random Forest achieved 95.75% accuracy, followed by SVM with 93.25% accuracy, and Logistic Regression and KNN each achieved 90.19% accuracy. This finding emphasizes that Stacking with Gradient Boosting provides much better performance in water quality classification compared to other models. This research provides new insights into the application of machine learning algorithms for water quality management as well as guidance for optimal algorithm selection.

Keywords

Water Quality, Logistic Regression, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN)

Introduction

One of the most valuable resources that is essential to life as we know it is water. Water quality is lowered by contamination, which has an indirect negative impact on human health as well as the health of marine life. This makes it highly important to monitor water quality and protect the survival of marine species (Shams et al., 2024). A significant threat to water shortages and diseases associated to water is the fact that 20% of people globally lack access to safe drinking, water, and almost 50% do not have access to adequate sanitation systems. These statistics are provided by the UN Environment Program (2000).(Juna and others, 2022). Water quality is defined based on its physical, chemical, and biological factors, and assessing its quality is

Submission: 20 October 2024; **Acceptance:** 21 November 2024

necessary before it is utilized for various intended uses such as drinking water, agriculture, pleasure, and industrial water use, among others. (Dashpande and Sargaonkar, 2022).

Previous research has shown the potential of using machine learning methods in water quality assessment. (Nababan et al., 2022) compared the K-Nearest Neighbor (KNN) algorithm with Support Vector Machine (SVM) and found that KNN achieved up to 89.58% accuracy in water quality status classification. Meanwhile, (Krtolica et al., 2023) explored the use of various machine learning models, including ANNs, SVMs, RFs, and KNN, to predict river ecological state classes based on water quality parameters, and reported that some models achieved accuracies above 80%. These findings suggest that machine learning is a promising tool for water quality assessment and prediction.

The Stacking Gradient Boosting method in this study combines SVM, Random Forest, and Logistic Regression models with Gradient Boosting as a meta-learner. This approach optimizes the strengths of each model to produce more accurate and stable predictions, and overcomes the problem of overfitting. Proven effective in handling complex water quality data, this method achieved the highest accuracy of 96%, surpassing other algorithms.

This research compares four machine learning algorithms-Logistic Regression, Random Forest, SVM, and KNN for water quality classification. With a stacking method that combines these models and Gradient Boosting as a meta-learner, this research shows that stacking produces more accurate and stable predictions, making it an effective method for handling complex data in water quality classification.

Methodology

Import Libraries

Importing libraries required for data manipulation (pandas), model training (sklearn), and visualization (matplotlib, seaborn).

(Widodo et al., 2021) clearly demonstrated the use of “Pandas, NumPy, Sklearn” libraries in their research for data processing and machine learning model development, while (Saberioon et al., 2018) emphasized the role of Scikit-learn in providing various machine learning algorithms and essential tools. On the other hand, (Sheng et al., 2020) mentioned the use of “Sklearn, Keras, TensorFlow” for their experiments, illustrating how these libraries support more in-depth data analysis and modeling.

Load and Preprocess Data

Check Dataset: Checks the dataset information.
Handle Missing Data: Replaces invalid values with NaN and fills them with 0.

(Nasir et al., 2022) loads water quality data from Excel files, handles missing values, and splits the data into training and testing sets. (Fern et al., 2022) accesses data from the database, performs data cleaning and transformation before building the model. (K-nn et al., 2023) load river water quality data, perform cleaning and normalization, and then apply the K-NN algorithm. Meanwhile, (Fattah et al., 2024) loaded the water quality dataset, performed data preprocessing, and divided the data into training and testing sets.

Prepare Data for Modeling

Separate Features and Labels: Specifies the feature (X) and label (y).
Convert Labels: Converts labels into numeric format if required.

(Fern et al., 2022) perform feature engineering by creating new features and performing feature transformation before developing the model. (Fattah et al., 2024) perform categorical feature encoding and data normalization before training the SVM model. Meanwhile, (Saberioon et al., 2018) performed data normalization before applying the K-NN and Random Forest algorithms.

Split Data

Splitting the data into training and testing sets (80% training, 20% testing).

(Fattah et al., 2024) utilized Scikit-learn's train_test_split function to split the dataset into training and test data. (Saberioon et al., 2018) also utilized the train_test_split function of Scikit-learn, splitting the dataset with a ratio of 80:20 and a random state of 101. (Fern et al., 2022) divided the data into two subsets, namely training set (2009-2020) and testing set (January 2021 - April 2021), and applied 10-fold cross-validation to avoid overfitting.

Feature Scaling

Using StandardScaler to standardize features.

(Abuzir & Abuzir, 2022) applied min-max normalization to convert numeric feature values to the same scale (0 to 1), while (Sheng et al., 2020) used standard deviation standardization in the data normalization process. (Fern et al., 2022) applied z-score normalization to transform the dataset to the default scale before model training and testing. On the other hand, (Fattah et al., 2024) mentioned normalization as a data preprocessing step before training the SVM model, and (Saberioon et al., 2018) performed data normalization before applying the K-NN and Random Forest algorithms.

Train and Evaluate Models

Logistic Regression, Random Forest, SVM, KNN: Train and evaluate four different classification models. Calculates classification accuracy and reports, and displays the confusion matrix for each model.

(Abuzir & Abuzir, 2022) trained various machine learning models, including Naive Bayes, SVM, Decision Tree, Random Forest, KNN, and ANN, and evaluated them using accuracy, precision, recall, and F1-score. (Nasir et al., 2022) trained seven different machine learning models and compared their performance based on accuracy. (K-nn et al., 2023) focused on the K-NN model and evaluated it using accuracy, precision, and recall. (Fattah et al., 2024) trained SVM models and evaluated them using accuracy, precision, recall, and F1-score, while (Saberioon et al., 2018) trained K-NN and Random Forest models and evaluated them with accuracy, precision, recall, and F1-score.

Visualize Results

Display the confusion matrix as a heatmap to visualize the prediction results compared to the original labels.

(K-nn et al., 2023) used confusion matrix to evaluate the performance of K-NN model, while (Fattah et al., 2024) used confusion matrix to visualize the performance of SVM model.

(Saberioon et al., 2018) also utilized confusion matrix to evaluate the performance of K-NN and Random Forest models.

Staking Gradient Boosting

This methodology improves classification accuracy with a stacking technique, which combines SVM, Random Forest, and Logistic Regression as base models, and Gradient Boosting as a meta-learner. The stacking model is trained with 5-fold cross-validation, then evaluated using accuracy, classification report, and confusion matrix to assess performance and classification error.

(Uvaliyeva et al., 2022) applied stacking to combine Bayes algorithm, decision trees, and neural networks in academic statistical analysis, while (Natekin & Knoll, 2013) described Gradient Boosting Machines (GBM), a framework that supports stacking, with a focus on building sequential ensemble models.

Results and Discussion

Table 1 Clasification Result Accuracy

Algorithm	Splitting Dataset	Accuracy
Logistic Regression	80% Train, 20% Test	0.901875
Random Forest	80% Train, 20% Test	0.9575
SVM	80% Train, 20% Test	0.9325
KNN	80% Train, 20% Test	0.901875
Staking Gradient Boosting	**80% Train, 20% Test**	**0.96**

The results show that Stacking Gradient Boosting has the best performance with the highest accuracy of 96%. This algorithm uses a sophisticated ensemble approach, where predictions from multiple base models (SVM, Random Forest, and Logistic Regression) are combined using Gradient Boosting as a meta-learner. Its ability to utilize the strengths of each base model and overcome their individual weaknesses makes Stacking Gradient Boosting superior to other algorithms. This advantage is supported by Gradient Boosting's ability to iteratively improve prediction accuracy and overcome overfitting. In addition, Stacking Gradient Boosting is highly effective in handling complex and varied data, resulting in more accurate and stable predictions in water quality classification tasks.

The Random Forest classification model achieved 95.75% accuracy on the test data, showing excellent performance. The classification report and confusion matrix indicate the model's ability to recognize patterns, with some classes requiring more attention. Overall, the model shows strong potential for further classification applications.

SVM came in second place with 93.25% accuracy. This algorithm works by finding the optimal hyperplane that separates the classes in the data space. SVM is suitable for non-linear data and can handle outliers well. However, the downside of SVM is its sensitivity to parameter tuning, which means the model requires more complex parameter settings for optimal performance.

Logistic Regression, which is a linear model, also performed quite well with 90.19% accuracy. This algorithm is simple and is often used for classification problems that involve a linear relationship between input and output variables. Despite its good performance, Logistic

Regression is less effective when the relationship between feature and label variables is non-linear or the data has a complex distribution.

KNN, which also achieved 90.19% accuracy, is a simple and intuitive algorithm, where classification is based on the number of nearest neighbors in the data space. The main advantages of KNN are simplicity and not requiring extensive model training. However, the disadvantage is that performance degrades as the dataset size increases, as KNN becomes less efficient and slower. In addition, KNN is also highly dependent on the parameter k (number of nearest neighbors), which affects the quality of the prediction results.

The stacking model with Gradient Boosting as a meta-learner showed an excellent accuracy of 96% on the test data. This shows that the combination of SVM, Random Forest, and Logistic Regression as base models is effective in providing accurate predictions. This successful stacking indicates that the meta-learner model can utilize the strengths of each base model to improve the overall performance, overcoming the weaknesses of individual models and providing more stable and accurate results. This high performance should also be considered along with other metrics such as precision, recall, and F1-score for an overall assessment.

Based on the results, it can be concluded that the Stacking Gradient Boosting algorithm is the most effective method for water quality classification, through the combination of SVM, Random Forest, and Logistic Regression. Random Forest also showed excellent performance, while SVM was effective on non-linear data. Logistic Regression and K-Nearest Neighbors (KNN) offer adequate performance, but are less efficient for more complex data. This research provides clear guidance on the selection of appropriate machine learning algorithms for water quality classification, particularly in the context of water resources management and environmental quality monitoring.

Conclusion

Based on the research results, it can be concluded that the Stacking Gradient Boosting algorithm is the most effective method for water quality classification based on contaminant parameters. This method achieved the highest accuracy of 96% by combining the strengths of SVM, Random Forest, and Logistic Regression as base models, and using Gradient Boosting as a meta-learner. Although Random Forest also showed excellent performance with 95.75% accuracy, Stacking Gradient Boosting proved to be superior in handling data complexity and providing more accurate and stable predictions. SVM, with 93.25% accuracy, is effective for non-linear data but requires more careful parameter settings. Logistic Regression and KNN, while achieving fairly good accuracy (90.19%), are less efficient in handling more complex data and larger datasets. This research provides valuable guidance in the selection of appropriate machine learning algorithms for water quality classification, especially in the context of water resources management and environmental monitoring.

Recommendation

Recommendations include the use of more diverse datasets to improve model accuracy and coverage, as well as the exploration of more complex models such as deep learning. The use of

more sophisticated validation methods is also recommended to reduce overfitting. In addition, integration of the model with real-time water quality monitoring systems can support faster and more effective decision-making.

References

Abuzir, S. Y., & Abuzir, Y. S. (2022). *Machine learning for water quality classi fi cation*. *57*(3), 152–164. https://doi.org/10.2166/wqrj.2022.004

Bayu Prihambodo, Wildan, A., Eko Prayoga, & Jaffar, A. (2023). Klasifikasi Kualitas Air Sungai Berbasis Teknik Data Mining Dengan Metode K-Nearest Neighbor (K-NN). *Emitor Jurnal Teknik Elektro*, *1*(1), 31–36. https://doi.org/10.23917/emitor.v1i1.20833

Fattah, N. F. (2024). PENERAPAN DATA MINING UNTUK KLASIFIKASI KUALITAS AIR DENGAN ALGORITMA SUPPORT VECTOR MACHINE PADA DINAS LINGKUNGAN HIDUP DAN PERTANAHAN PROVINSI SUMSEL. *PROSISKO: Jurnal Pengembangan Riset Dan Observasi Sistem Komputer*, *11*(2), 145–158. https://doi.org/10.30656/prosisko.v11i2.8285

Fernández del Castillo, A., Yebra-Montes, C., Verduzco Garibay, M., de Anda, J., Garcia-Gonzalez, A., & Gradilla-Hernández, M. S. (2022). Simple Prediction of an Ecosystem-Specific Water Quality Index and the Water Quality Classification of a Highly Polluted River through Supervised Machine Learning. *Water*, *14*(8), 1235. https://doi.org/10.3390/w14081235

Juna, A., Umer, M., Sadiq, S., Karamti, H., Eshmawi, A. A., Mohamed, A., & Ashraf, I. (2022). Water Quality Prediction Using KNN Imputer and Multilayer Perceptron. *Water (Switzerland)*, *14*(17), 1–19. https://doi.org/10.3390/w14172592

Krtolica, I., Savić, D., Bajić, B., & Radulović, S. (2023). Machine Learning for Water Quality Assessment Based on Macrophyte Presence. *Sustainability (Switzerland)*, *15*(1), 1–13. https://doi.org/10.3390/su15010522

Nababan, A. A., Khairi, M., & Harahap, B. S. (2022). Implementation of K-Nearest Neighbors (KNN) Algorithm in Classification of Data Water Quality. *Jurnal Mantik*, *6*(1), 30–35. https://doi.org/10.35335/jurnalmantik.v6i1.2130

Nasir, N., Kansal, A., Alshaltone, O., Barneih, F., Sameer, M., Shanableh, A., & Al-shamma, A. (2022). Journal of Water Process Engineering Water quality classification using machine learning algorithms. *Journal of Water Process Engineering*, *48*(June), 102920. https://doi.org/10.1016/j.jwpe.2022.102920

Natekin, A., & Knoll, A. (2013). *Gradient boosting machines , a tutorial*. *7*(December). https://doi.org/10.3389/fnbot.2013.00021

Romy Budhi Widodo, Windra Swastika, Setiawan, H., & Mochamad Subianto. (2018). STUDI PEMROSESAN DATA PENGENALAN GESTUR TANGAN MENGGUNAKAN

METODE KNN. *Conference on Innovation and Application of Science and Technology (CIASTECH)*, *0*(0), 277–286. https://doi.org/10.31328/ciastech.v0i0.3320

Saberioon, M., Císař, P., Labbé, L., Souček, P., Pelissier, P., & Kerneis, T. (2018). Comparative performance analysis of support vector machine, random forest, logistic regression and k-nearest neighbours in rainbow trout (oncorhynchus mykiss) classification using image-based features. *Sensors (Switzerland)*, *18*(4), 1–15. https://doi.org/10.3390/s18041027

Sargaonkar, A., & Deshpande, V. (2003). Development of an Overall Index of Pollution for Surface Water Based on a General Classification Scheme in Indian Context. *Environmental Monitoring and Assessment*, *89*(1), 43–67. https://doi.org/10.1023/a:1025886025137

Shams, M. Y., Elshewey, A. M., El-kenawy, E. S. M., Ibrahim, A., Talaat, F. M., & Tarek, Z. (2024). Water quality prediction using machine learning models based on grid search method. *Multimedia Tools and Applications*, *83*(12), 35307–35334. https://doi.org/10.1007/s11042-023-16737-4

Sheng, L., Zhou, J., Li, X., Pan, Y., & Liu, L. (2020). *Water quality prediction method based on preferred classification*. *i*, 1–5. https://doi.org/10.1049/iet-cps.2019.0062

Uvaliyeva, I., Zhenisgul Rakhmetullina, Baklanova, O., & György Györök. (2022). The Development of the Staking-Ensemble of Methods for Analyzing Academic Data. *Acta Polytechnica Hungarica*, *19*(11), 7–25. https://doi.org/10.12700/aph.19.11.2022.11.1