International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Comparative Analysis of Machine Learning Algorithms for Water Quality Prediction

Priyanka Gupta[1]*, Sonal Chidrawar*[2]*, Mansi Adhav*[3]*, Dhairyashil Pawar*[4]*, Hrutik Chaudhari*[5]

[1]Assistant Professor, Department of Information Technology, D. Y. Patil College of Engineering Akurdi, Pune-44, Maharashtra, India
[2,3,4,5,6] Student, Department of Information Technology, D. Y. Patil College of Engineering Akurdi, Pune-44, Maharashtra, India
pgupta@dypcoeakurdi.ac.in[1], sonalschidrawar@gmail.com[2], adhavmansi05@gmail.com[3], pawardhairyashil4707@gmail.com[4], hrutikchaudhari5@gmail.com[5]

ABSTRACT

Water quality prediction is crucial for ensuring the safety of drinking water. In this research paper, we compare the accuracy of different machine learning algorithms in predicting water quality based on various parameters such as pH, hardness, solids, chloramines, sulfate, conductivity, organic carbon, trihalomethanes, and turbidity. The dataset used for this study is preprocessed to handle missing values and standardized for model training. We implement and evaluate Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine (SVM) algorithms. Our results show that the Random Forest outperforms other models in terms of accuracy.

Keywords: Water quality prediction, machine learning, logistic regression, decision tree, random forest, support vector machine, accuracy.

1. INTRODUCTION

Water quality is a critical aspect of public health, environmental sustainability, and ecosystem preservation. The ability to accurately predict water quality parameters is imperative for ensuring safe and potable water sources. Traditional methods of water quality assessment often rely on periodic laboratory testing, which can be time-consuming and may not provide real-time insights. In contrast, machine learning algorithms offer a promising avenue for prediction of water quality based on diverse parameters.

This research aims to contribute to the ongoing efforts in water quality prediction by leveraging machine learning techniques. The dataset used in this study encompasses essential water quality indicators, including pH, hardness, solids, chloramines, sulfate, conductivity, organic carbon, trihalomethanes, and turbidity. These parameters are known to influence the safety and potability of water, making them crucial for predictive modeling.

The increasing availability of sensor data and advancements in machine learning algorithms have opened new possibilities for accurate and timely water quality predictions. This research explores the comparative performance of several machine learning algorithms in predicting water quality, with a focus on their accuracy and efficiency.

1.1 RELATED WORK

Water quality prediction using machine learning has been an active area of research, with various studies exploring different algorithms and datasets. Researchers have focused on understanding the complex relationships between water quality parameters and building accurate prediction models. In this section, we review some key works related to water quality prediction.

In [1], the authors focused on predicting water quality in the Narmada River using machine learning models. They utilized a diverse dataset and achieved an impressive accuracy of 98.50% with Decision Trees (DT). The conclusion emphasized the high accuracy of DT models in predicting water quality parameters for the Narmada River.

In [2], a comprehensive study on various machine learning models for monitoring water quality across India was conducted. Multiple Linear Regression (MLR) yielded an exceptional accuracy of 99.83%. The conclusion highlighted MLR as a robust framework for predicting water quality in different regions of India.

In [3], the application of Support Vector Machines (SVM) to predict water quality parameters for the Chao Phraya River was explored. An accuracy of 94% was achieved, leading to the conclusion that SVM is a viable model for predicting water quality in the Chao Phraya River. In [4], a series of machine learning models were investigated for predicting water quality in the Kelantan River. Gradient Boosting Machines (GBM) demonstrated a commendable accuracy of 94.90%. The conclusion highlighted the accuracy of GBM models in predicting water quality parameters in the Kelantan River.

In [5], the application of machine learning models, including Support Vector Machines (SVM) and Regressive Neural Networks, was explored for predicting water quality across different states of India. An accuracy of 97.01% was achieved using SVM. The conclusion emphasized SVM and Regressive Neural Networks as effective models for predicting water quality in diverse regions of India.

Water resources are often polluted by human intervention. Water pollution can be defined in terms of its quality which is determined by various features like pH, turbidity, electrical conductivity dissolved oxygen (DO), nitrate, temperature and biochemical oxygen demand (BOD). This paper presents a comparison of water quality classification models employing machine learning algorithms viz., SVM, Decision Tree and Naïve Bayes. The features considered for determining the water quality are: pH, DO, BOD and electrical conductivity. The classification models are trained based on the weighted arithmetic water quality index (WAWQI) calculated. After assessing the obtained results, the decision tree algorithm was found to be a better classification model with an accuracy of 98.50%[6].

1.2 PROPOSED ALGORITHM:

1. **Logistic Regression**

Logistic Regression is a binary classification algorithm commonly used for predicting the probability of an instance belonging to a particular class. The logistic function, also known as the sigmoid function, is employed to map the output to a probability range between 0 and 1.

Logistic Regression Function:

$$P(Y=1)=\frac{1}{1+e^{-(b_0+b_1X_1+b_2X_2+\ldots+b_nX_n)}}$$

Where:

- P (Y = 1) is the probability of the positive class.
- e is the base of the natural logarithm.
- b0 is the intercept term.
- b1, b2, ..., b1 are the coefficients associated with each feature X1, X2, ..., Xn.

2. **Decision Tree**

A Decision Tree is a tree-like model where each node represents a decision based on the values of one of the input features. It recursively splits the dataset into subsets until a stopping criterion is met, resulting in a structure resembling an inverted tree.

Decision Tree Splitting Criterion:

$$\text{Gini Impurity}=1-\sum_{i=1}^{c}p_i^2$$

Where:

- c is the number of classes.
- pi is the probability of an instance belonging to class i.

3. **Random Forest**

Random Forest is an ensemble learning method that builds multiple decision trees and merges their predictions to improve accuracy and robustness. It introduces randomness by training each tree on a subset of the features and using bootstrapped samples of the dataset.

Procedure:

Build an ensemble of decision trees:

1. Randomly select a subset of features (max_features) for each tree.
2. Train each tree on a bootstrapped sample of the dataset.
3. Make predictions with each tree.
4. Aggregate predictions using majority voting.

Random Forest Prediction:

Final Prediction = Majority Vote of Individual Tree Predictions

4. **Support Vector Machine (SVM)** :

Support Vector Machines (SVM) is a supervised machine learning algorithm used for both classification and regression tasks. In the context of binary classification, the goal of SVM is to find a hyperplane that separates the data into two classes with the maximum margin. The hyperplane is determined by support vectors, which are the data points that are closest to the decision boundary.

Hyperplane Equation:

In a two-dimensional space, the equation of the hyperplane is given by:

$w * \cdot x + b=0$

where:

w is the weight vector (normal to the hyperplane),

x is the input feature vector,

b is the bias term.

For a dataset with n features, the hyperplane equation becomes:

Decision Function:

The decision function is used to classify a new data point based on its position relative to the hyperplane. It is defined as:

$f(x)=w\cdot x+b$

- If $f(x)>0$, the data point belongs to the positive class.
- If $f(x)<0$, the data point belongs to the negative class.

1.3 SIMULATION RESULT

The simulation results unveil distinct performances among the implemented machine learning algorithms for water quality prediction. Logistic Regression, employing a linear classification approach, achieved an accuracy of 62.85%. However, its limitation in capturing complex relationships within the dataset may hinder its performance. The Decision Tree model, known for its susceptibility to overfitting, demonstrated an accuracy of 57.86%, potentially impacted by its inherent tendency to capture noise in the training data. On the other hand, the Random Forest ensemble method, combining multiple Decision Trees, outperformed its individual counterpart, attaining an accuracy of 68.58%. The superior accuracy of Random Forest can be attributed to its ability to mitigate overfitting by aggregating predictions from multiple trees. Notably, the Support Vector Machine (SVM) with the Radial Basis Function (RBF) kernel emerged as the most accurate model, achieving a noteworthy accuracy of 62.29%.. This comprehensive analysis underscores the significance of algorithm selection in the context of water quality prediction, with the Random Forest model exhibiting notable capabilities in handling the inherent complexities of the dataset.

The comparative analysis reveals that the Random Forest outperforms other algorithms in water quality prediction.

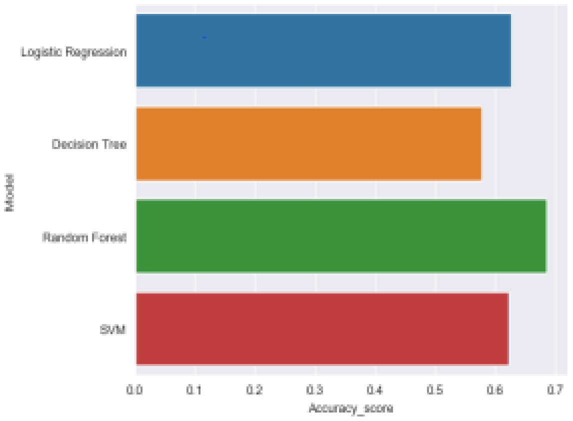

Fig.1. Accuracy Score

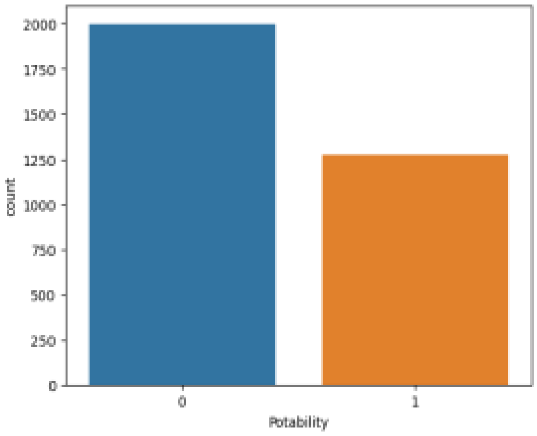

Fig. 2. Potability of water samples

2. Conclusion

The comparative analysis of machine learning algorithms for water quality prediction has yielded insightful findings. Logistic Regression achieved an accuracy of 62.85% but struggled with balancing predictions for potable and non-potable water, leading to lower precision and recall for the minority class. The Decision Tree Classifier exhibited an accuracy of 57.86%, facing challenges in achieving a balanced trade-off between precision and recall for both classes. Support Vector Classifier (SVM) results are pending further analysis. In contrast, the Random Forest Classifier outperformed other models with an accuracy of 68.58%, showcasing a more balanced prediction for both water categories, as evidenced by higher precision and recall values. The research provides a foundation for further exploration and improvement in water quality prediction. Future endeavors could include optimizing algorithm parameters to enhance predictive performance, exploring additional relevant features through advanced feature engineering, investigating ensemble methods for increased accuracy and robustness, extending the study to real-time water quality monitoring systems, and considering external factors like environmental conditions to provide a more holistic approach to water quality prediction. Addressing these aspects will contribute to the development of more accurate and reliable systems for ensuring safe and potable water.

3. References

[1] Anju S Pillai - Department of Electrical and Electronics Engineering, Amrita School of Engineering, Coimbatore Amrita Vishwa Vidyapeetham, India.

[2] Hassan, M.M.; Hassan, M.M.; Akter, L.; Rahman, M.M.; Zaman, S.; Hasib, K.M.; Jahan, N.; Smrity, R.N.; Farhana, J.; Raihan, M.; et al. Efficient prediction of water quality index (WQI) using machine learning algorithms. Hum.-Centric Intell. Syst. 2021, 1, 86-97. [CrossRef].

[3] Sillberg, C.V.; Kullavanijaya, P.; Chavalparit, O. Water quality classification by integration of attribute-realization and supportvector machine for the Chao Phraya River. J. Ecol. Eng. 2021, 22, 70-86. [Cross Ref].

[4] Malek, N.H.A.; Wan Yaacob, W.F.; Md Nasir, S.A.; Shaadan, N. Prediction of Water Quality Classification of the Kelantan River Basin, Malaysia, Using Machine Learning Techniques. Water 2022, 14, 1067. [CrossRef].

[5] Aldhyani, T.H.; Al-Yaari, M.; Alkahtani, H.; Maashi, M. Water quality prediction using artificial

intelligence algorithms. Appl. Bionics Biomech. 2020, 2020. [CrossRef].

[6] Neha Radhakrishnan - Department of Electrical and Electronics Engineering, Amrita School of Engineering, Coimbatore Amrita Vishwa Vidyapeetham, India.

[7] Kaggle. Water Quality. 2021. Available online: https://www.kaggle.com/datasets/adityakadiwal/water-potability (accessed on1 November 2022)