SSRG International Journal of Agriculture & Environmental Science
ISSN: 2394-2568/ https://doi.org/10.14445/23942568/IJAES-V9I4P102

Volume 9 Issue 4, 6-11, Jul-Aug 2022
© 2022 Seventh Sense Research Group®

Original Article

Water Quality Analysis of Different Water Sources in Kerala, India

Remya R S[1], Ebin Antony[2]

[1,2] Department of Information Technology, Kannur University, Kerala, India

Received: 30 April 2022 Revised: 18 June 2022 Accepted: 30 June 2022 Published: 18 July 2022

Abstract - *This paper aimed to classify the quality of different water sources in Kerala. All major rivers, lakes, and reservoirs are included in this study. The quality of water is classified into different classes based on the level of pH, Biochemical Oxygen Demand (BOD), Dissolved Oxygen, Electrical Conductivity, and concentration of Total Coliforms in the water sample. For classification purposes, different classification models are proposed. Among these classification models, Naïve Bayes scored 78.79 % accuracy. SVM scored 82.83 % accuracy. The decision Tree Classifier's accuracy in this case study is 93.94 %, XG Boost classifier scored 94.95 % accuracy. Random Forest scored the highest accuracy, i.e., 95.96 %. Also, classification reports of these models are evaluated. On evaluating these results, it can be concluded that Random Forest gives the best results.*

Keywords *- Decision Tree, Random Forest, Support Vector Machine, Water Quality Analysis, XG Boost.*

1. Introduction

Every human being requires water to survive. There are numerous water sources, including rivers, reservoirs, and lakes. However, not all water sources are safe to drink. The quality of water determines how it is used in various situations. The physical and chemical qualities of water, as well as its microbiological features, affect its quality.

Water pollution is becoming more and more of a concern because of waste product dumping, untreated sewage disposal, and the release of toxic chemicals from industry into water bodies. This issue had to be resolved to obtain water for home use. Based on machine learning algorithms, this paper provides a viable classification model for classifying water quality. Five categorization models are created and tested in this case study. Their results are also evaluated to determine the optimal classification model. The maximum accuracy and best classification report are provided by Random Forest Classifier. Since water is a necessity for life, this kind of research is needed to meet current needs.

The information provided by the Kerala Pollution Control Board is used in this project. The goal of this research is to create the best classification model that categorizes diverse sources of water in Kerala. So that this model can be used to classify water quality in the future.

2. Literature Survey

Ahmed et al. [1] the Water Quality Index (WQI) was predicted using single feed-forwarded neural networks and a combination of neural networks. R2 and MSE were attained at 0.9270, 0.9390, 0.1200, and 0.1158, respectively, using a mixture of backward elimination and forward selection selective combination techniques. Abaneh [2] used ANN and multivariate linear regression to measure biochemical and chemical oxygen demand. COD and BOD were predicted using PH, temperature, and Total Suspended Solids (TSS) and Total Suspended (TS).

Ali and Qamar [3] used water quality classification to classify samples. However, in the learning process, they ignored key WQI components. Gazzas et al. [4] An artificial neural network was used to predict WQI. They calculated WQI using 23 parameters, which was an expensive procedure due to the use of sensors.

Deep neural networks, recurrent neural networks, neuro-fuse assumptions, and support vector regression are the most prominent techniques for detecting and classifying water quality. To estimate the amount of dissolved oxygen and chlorophyll in water samples, Barzeger et al. (2020) used a CNN-LSTM amalgam model [6]. The CNN-LSTM amalgam model beat all other models, according to the results [5].

Compare Fuzzy Logic Assumptions with WQI Approaches for Water Quality Assessment in the Ikare Community of Nigeria by Oladipo et al. (2021) [6]. The FLI method outperforms the WQI method, they discovered. The ETR approach was used for eleven water quality factors by Asadollah et al. (2021) [7].

3. Materials and Methods

First data preprocessing is done where data cleaning and resampling are done, then classification models are formed step by step. Finally, after testing the models and comparing their accuracy and classification report, models showing the best results can be chosen among them [8, 9].

3.1. Importing or Collecting Data

The data needed for this project is taken from the report of the Kerala Pollution Control Board, showing details of air and water quality in the year 2019. The dataset formed using this official directory includes data of all the 44 rivers, 6 tributaries, 3 freshwater lakes, 7 estuarine lakes, 6 reservoirs, 2 ponds, 1 canal, and 2 streams.

3.2. Getting Insights About Dataset

Through the statistical description of the dataset, one can get a good picture of the distribution of data. Basic information of data like the type of datatypes used, number of rows and columns helps to understand how to perform the rest of the data pre-processing and data processing steps. The dataset consists of 220 rows and 15 columns.

3.3. Handling Missing Values

Missing Data is a very big problem in programming as well as in real-life cases. Missing Data is referred to as NaN values in python programming. Since the data obtained by the official site is very clean, so there were no missing values [8].

3.4. Resampling

Water samples obtained from different sources are categorized into different classes by the government of Kerala as shown in table 1. But then there are 8 samples belonging to class A, 15 samples belonging to D and E. Below E category has 25 samples. 75 samples belong to class C. While class B has 82 samples. So, the number of samples of different categories is not the same. This could cause the model's accuracy to decline. So, the lower classes are unsampled to balance the dataset.

Table1. Categorization of different classes of water samples and their criteria.

Designated Best Use	**Class of water**	**Criteria**
Source of traditionally untreated but disinfected drinking water	A	Total coliform organism MPN / 100ml should be 50 or less Oxygen dissolved between pH 6.5 and 8.5 is 6mg / l or more. Biochemical oxygen requirement 5 days 20C 2mg / l or less
Outdoor bathing	B	Coliforms overall 500 or fewer organisms per 100 ml, pH 6.5–8.5, and dissolved oxygen 5 mg/l or higher Oxygen biochemically Demand 5 days at or below 20C and 3mg/l
sources of drinking water following standard treatment and disinfection	C	Coliform’s overall pH must range from 6 to 9 and organism MPN/100 ml must be 5000 or less. Oxygen in solution 4 mg/l or higher Oxygen Demand Biochemical 5 days at or below 20C and 3mg/l
development of fisheries and wildlife	D	pH ranging from 6.5 to 8.5 Oxygen in solution 4 mg/l or higher 1.2 mg/l or less of free ammonia
Industrial cooling, controlled waste disposal, and irrigation	E	pH between 6.0 and 8.5 Electrical conductivity at 25C micromhos/cm maximum of 2250 Sodium absorption ratio maximum of 26 Boron 2 mg/l
	Below E	Not Meeting A, B, C, D & E Criteria

3.5. Models Used for Classification

There are many classification methods in both supervised and unsupervised machine learning approaches. Since the dataset has label data, in this case, study supervised machine learning is used. Classification methods used in this case study are Naïve Bayes, SVM, Decision Tree Classification, Random Forest Approach, and X G Boost.

3.5.1. Naïve Bayes Classification

Bayes' theory is an essential term in many statistical and advanced machine learning models. Bayes’ theorem is the basis of naive Bayes learning [10]. Building a Bayesian probabilistic model with a pin class probability as an example is known as naive Bayes learning. Missing values do not pose a problem for the Bayes classifier. This classifier does not necessitate a big amount of training data. It can work with both continuous and discrete data. But this algorithm faces a 'zero-frequency problem'. Also, it assumes that all features are independent, which is not practical in real life.

3.5.2. Support Vector Machine

SVM is a supervised machine learning method used for regression and classification [11]. It can be used to solve regression problems, but it excels at classification. The SVM algorithm's main purpose is to find a hyperplane that categories data points in an N-dimensional space. The amount of feature variables in the dataset affects how big the hyperplane is. The hyperplane is only a straight line when there are just two input features. The hyperplane becomes a two-dimensional plain when there are three input features. However, it becomes challenging to visualize the hyperplane when there are more than three feature variables.

3.5.3. Decision Tree Classification

A decision tree classifier is the most well-known method of data classification [12]. A decision tree has two different kinds of nodes.

- Leaf node - has a class label.
- Internal node - It's a feature query. It divides into sections based on the responses.

Decision Tree Classifier is easy to understand and implement, and ais are easy to use. It is computationally cheap. But overfitting might happen in the case of this classifier.

3.5.4. Random Forest Approach

Random forests are a set of tree predictors in which each tree is affected by the random vector's values. As a result, the random forest model's building pieces are decision trees. The random forest's trees each spat out a class forecast. In irregular forests, overfitting is reduced to improve accuracy. It is flexible in terms of classification and regression issues. The Random Forest model works well with random and continuous values. It automates lost values in data. There is no need to normalize the data here. However, this requires a computing and training process that takes a lot of time [13].

3.5.5. XG Boost

The full name of a method that recently dominated Kaggle competitions for structured or tabular data is Extreme Gradient Boosting or XG Boost. XG Boost [14] is a gradient boosted decision tree that is optimized for speed and accuracy. When compared to other gradient boosting implementations, XG Boost is often quicker. Gradient boosting machines, stochastic gradient boosting, and multiple additive regression trees are all terminology used to refer to XG Boost.

4. Results and Discussion

4.1. Confusion Matrix of Different Classifiers

A confusion matrix, in general, is a N x N matrix used to assess the effectiveness of a taxonomy model. As shown in Figure 1, the Confusion Matrix can be used to calculate various model parameters such as accuracy, precision, recall, and so on.

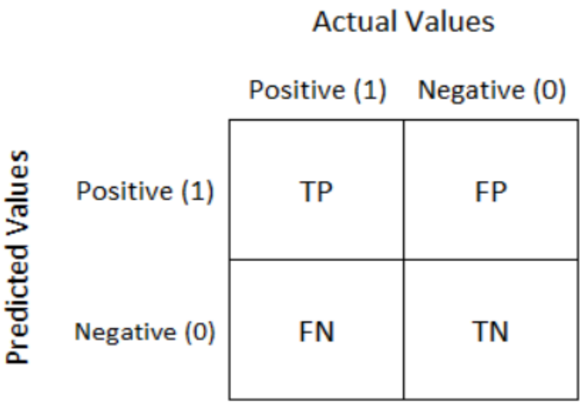

Fig. 1 Confusion Matrix

Accuracy is a measure of how accurately your model has made predictions for a complete test dataset.

Accuracy = (TP +TN)/ (TP +TN+FP+FN)

The recall rate, also known as the real positive rate, is a measure of how many positives are predicted out of the total number of positives in the dataset. This is referred to as sensitivity.

Recall = TP / (TP+FP)

A positive forecast's precision can be measured to see how accurate it is. Low recall usually follows high accuracy, whereas great accuracy follows high recall.

Precision = TP/(TP+FP)

Precision and recall are combined to determine the F1 score. The more accurate the model is at making predictions, the better the F1 score. The confusion matrices for the classification models Nave Bayes, SVM, Decision Tree, Random Forest, and XG Boost are shown in Figures 2, Figure 3, Figure 4, Figure 5, and Figure 6, respectively.

F1 = 2 * (precision *recall)/(precision +recall)

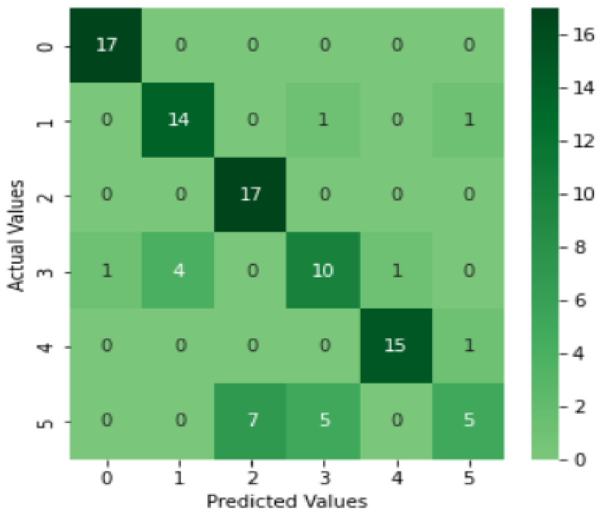

Fig. 2 Confusion Matrix of Naïve Bayes

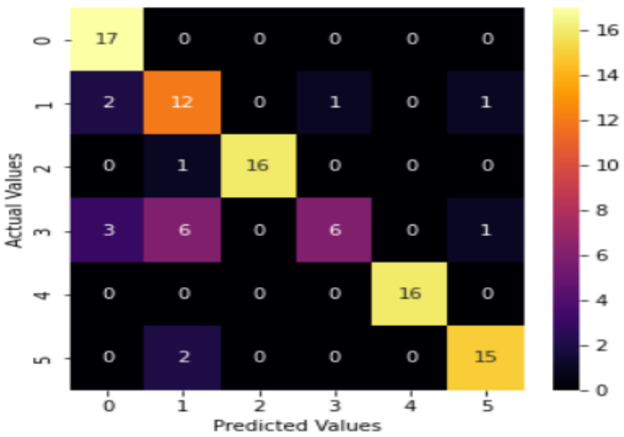

Fig. 3 Confusion Matrix of SVM

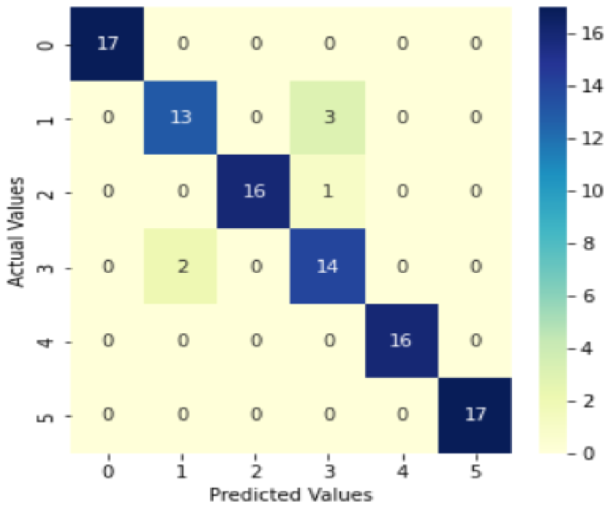

Fig. 4 Confusion Matrix of Decision Tree

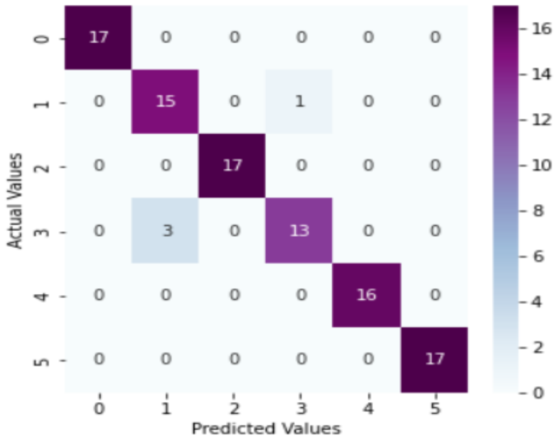

Fig. 5. Confusion Matrix of Random Forest

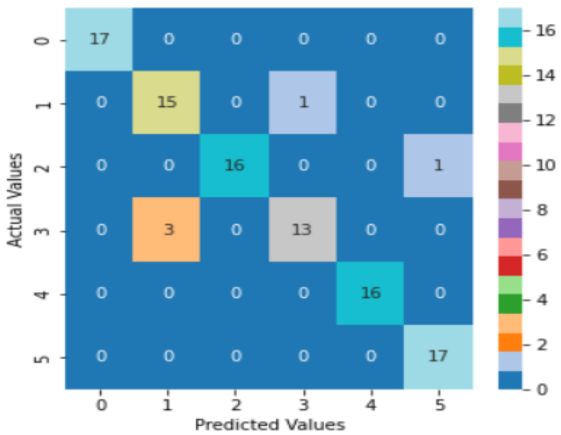

Fig. 6 Confusion Matrix of XG Boost

4.2. Accuracy of Different Classifiers

Classification accuracy is a metric that divides the number of right predictions by the total number of predictions to determine a classification model's performance. Because it is simple to calculate and understand, it is the most often used statistic for evaluating classifier models. Table 2 shows the accuracy of different classification models used in this paper.

Table 2. Accuracy of different classifiers

Classifier	Accuracy
Naïve Bayes	78.79
SVM	82.83
Decision Tree	93.94
XG Boost	94.95
Random Forest	95.96

4.3. Classification Reports of Models Used

In machine learning, a classification report is a performance evaluation indicator. It is used to display the trained classification model's precision, recall, F1 Score, and support. Table 3 lists various categorization model performance indicators like precision, recall, and f1-score.

Table 3. Classification report of various machine learning models

		Precision	**Recall**	**F1-score**	**Support**
Naïve Bayes	**A**	0.94	1.00	0.97	17
	B	0.78	0.88	0.82	16
	Below E	0.71	1.00	0.83	17
	C	0.62	0.62	0.62	16
	D	0.94	0.94	0.94	16
	E	0.71	0.29	0.42	17
	Accuracy	-	-	0.79	99
	Macro-avg	0.78	0.79	0.77	99
	Weighted avg	0.78	0.79	0.77	99
SVM	**A**	0.77	1.00	0.87	17
	B	0.57	0.75	0.65	16
	Below E	1.00	0.94	0.97	17
	C	0.86	0.38	0.52	16
	D	1.00	1.00	1.00	16
	E	0.88	0.88	0.88	17
	Accuracy	-	-	0.83	99
	Macro-avg	0.85	0.82	0.82	99
	Weighted avg	0.85	0.83	0.82	99
Decision Tree	**A**	1.00	1.00	1.00	17
	B	0.87	0.81	0.84	16
	Below E	1.00	0.94	0.97	17
	C	0.78	0.88	0.82	16
	D	1.00	1.00	1.00	16
	E	1.00	1.00	1.00	17
	Accuracy	-	-	0.94	99
	Macro-avg	0.94	0.94	0.94	99
	Weighted avg	0.94	0.94	0.94	99
XG Boost sf	**A**	1.00	1.00	1.00	17
	B	0.83	0.94	0.88	16
	Below E	1.00	0.94	0.97	17
	C	0.93	0.81	0.87	16
	D	1.00	1.00	1.00	16
	E	0.94	1.00	0.97	17
	Accuracy	-	-	0.95	99
	Macro-avg	0.95	0.95	0.95	99
	Weighted avg	0.95	0.95	0.95	99
Random Forest	**A**	1.00	1.00	1.00	17
	B	0.83	0.94	0.88	16
	Below E	1.00	1.00	1.00	17
	C	0.93	0.81	0.87	16
	D	1.00	1.00	1.00	16
	E	1.00	1.00	1.00	17
	Accuracy	-	-	0.96	99
	Macro-avg	0.96	0.96	0.96	99
	Weighted avg	0.96	0.96	0.96	99

According to the parameter values so obtained it's clear that Random Forest shows the best result. So, this model can best classify different water samples belonging to different classes.

5. Conclusion

Details of all major rivers, lakes, ponds, canals, and reservoirs of Kerala are included in this study. The quality of water is classified into different classes based on the level of pH, Biochemical Oxygen Demand (BOD), Dissolved Oxygen, Electrical Conductivity, and concentration of Total Coliforms in the water sample.

In his case study, five different classifiers are formed and compared with each other to analyze the quality of water. Among these classification models, Naïve Bayes scored 78.79 % accuracy. SVM scored 82.83 % accuracy. The Decision Tree Classifier's precision in this case study is 93.94 %, XG Boost classifier scored an accuracy is 94.95 %. Random Forest scored the highest accuracy, i.e., 95.96 %. Also, classification reports of these models are evaluated. On evaluating these results, it can be concluded that Random Forest gives the best results. In the future, this model can be used for the classification of water samples obtained from different sources.

Acknowledgments

The Kerala State Pollution Control Board is acknowledged by the authors for supplying the necessary data for this study.

References

[1] Ahmad Z, Rahim N, Bahadori A, Zhang J, "Improving Water Quality Index Prediction in Perak River Basin Malaysia through a Combination of Multiple Neural Networks," *Int. J. River Basin Manag*, vol. 15, pp. 79–87, 2017.

[2] Abyaneh H.Z, "Evaluation of Multivariate Linear Regression and Artificial Neural Networks in Prediction of Water Quality Parameters," *J. Environ. Health Sci. Eng*, vol. 12, pp. 40, 2014.

[3] Ali M, Qamar A.M, "Data Analysis, Quality Indexing and Prediction of Water Quality for the Management of Rawal Watershed in Pakistan," In Proceedings of the *Eighth International Conference on Digital Information Management (ICDIM)*, Islamabad, Pakistan, vol. 10, no. 12, pp. 108-113, 2013.

[4] A Cˇ omic´ L, "Neural Network Modeling of Dissolved Oxygen in the Gruža Reservoir," *Serbia, Ecol*, Model, vol. 221, pp. 1239–1244, 2010.

[5] Barzegar, Rahim, Mohammad Taghi, Aalami, Jan, Adamowski, "Short-Term Water Quality Variable Prediction Using a Hybrid CNN–LSTM Deep Learning Model," *Stochastic Environmental Research and Risk Assessment*, pp. 1–19, 2020.

[6] Oladipo, Johnson O. et al., "Comparison between Fuzzy Logic and Water Quality Index Methods: A Case of Water Quality Assessment in Ikare Community, Southwestern Nigeria," *Environmental Challenges*, vol. 3, pp. 100038, 2021.

[7] Asadollah, Seyed Babak, Seyed, Haji, et al., "River Water Quality Index Prediction And Uncertainty Analysis: A Comparative Study Of Machine Learning Models," *Journal of Environmental Chemical Engineering*, vol. 9, no. 1, pp. 104599, 2021.

[8] Ebin Antony, N S Sreekanth, R K Sunil Kumar, Nishanth T, "Data Preprocessing Techniques for Handling Time Series Data for Environmental Science Studies", *International Journal of Engineering Trends and Technology*, vol. 69, no. 5, pp. 196-207, 2021. Doi:10.14445/22315381/IJETT-V69I5P227

[9] Rahulraj P V and Ebin Antony, "Changes in Atmospheric Air Quality in the Wake of a Lockdown Related to Covid – 19 in the Capital City of Southern State of India, Kerala – Thiruvananthapuram," *SSRG International Journal of Agriculture & Environmental Science*, vol. 9, no. 3, pp. 17-24, 2022. *Crossref*, https://doi.org/10.14445/23942568/IJAES-V9I3P103

[10] Szafron, Duane & Greiner, Russ & Lu, Paul & Wishart, David & Macdonell, Cam & Anvik, John & Poulin, Brett & lu, Zhiyong & Eisner, Roman & Ca, Eisner@cs, Explaining Naïve Bayes Classifications, 2003.

[11] Osuna, Edgar & Freund, Robert & Girosi, Federico, "Support Vector Machines: Training and Applications," *Tech Rep A.I. Memo*, no. 1602, 1970.

[12] Bahzad Taha Jijo, "Classification Based on Decision Tree Algorithm for Machine Learning"

[13] Cutler, Adele & Cutler, David & Stevens, John, Random Forests, 2011. Doi: 10.1007/978-1-4419-9326-7_5.

[14] Bentéjac, Candice & Csörgő, Anna & Martínez-Muñoz, Gonzalo, "A Comparative Analysis of XGBoost," 2019.