Journal of Computer Science and Technology Studies
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

JCSTS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Analogy of H_2O Ranking and Its Stratification Using the SVM and XGBoost Method

Surya Ravichandran
Department of Computer Science & Engineering SRM Institute of Science and Technology, Kattankulathur, Chennai - 603203
Corresponding Author: Surya Ravichandran, **E-mail**: sr2362@srmist.edu.in

| ABSTRACT
Water is an important part of human beings and the living society. Over the years, air and water pollution have contaminated water in various ways. This makes the content unhygienic and harmful to drinking and society. The traditional method of water purification is expensive, and it involves a lot of unnecessary time, with the outcome of the results not up to the accuracy. My proposed system of thesis system is to develop a classification of the water quality using the Gradient boosting classifier. My research involves considering the various parameters of H_2O, including the ph, dissolved oxygen, Total Dissolved Solids(TDS), and temperature, which are predominant for the ranking of water contents.

| KEYWORDS

Support Vector Machine (SVM), XGBoost, Gradient Boosting Classifier, Machine Learning, Climate Models Integration.

| ARTICLE INFORMATION

ACCEPTED: 14 April 2025 **PUBLISHED:** 24 May 2025 **DOI:** 10.32996/jcsts.2025.7.3.111

1. Introduction

Water is a critical part that plays an important role in every part of life. The water is used for various purposes as such as drinking, irrigation, industries, and aquatic life maintenance. However, the quality of the water is often depleted due to the various pollutants and other wastes from the environment. Hence, the quality monitoring of the water is essential to lead a good life for the human beings.

Machine learning is employed for the need of the large-scale analysis of the dataset to be involved[2], they provide us with excellent results without being commanded and named for the above. In **Figure 1.2,** the machine learning approach of the model building is given, which learns from the data to produce the result. As my project is concerned, I will be utilising the two important algorithms for the models, namely the SVM and XGBoost methods, which can handle the large datasets and give us some good results for the above.

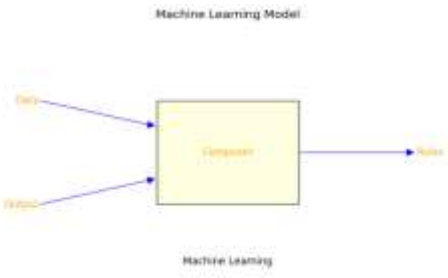

Figure 1.2: Machine learning algorithm.

1.1 Types of Machine Learning algorithms

1.1.1 Supervised Learning:

The supervised learning algorithm is one of the types of machine learning that has labelled input data, and we can predict the output by building the model required for the solution[2].

Thus, there are defined rows and columns that help to predict the score and accuracy for the model. In **Figure 1.3,** the ML model and Test data are visualised in 3d for supervised learning

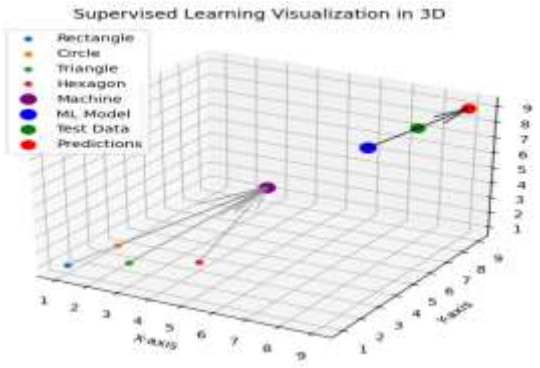

Figure 1.3: Implementation of Supervised Learning using Machine Learning Algorithm.

1.1.2 Unsupervised Learning:

Unsupervised learning works on the unlabeled datasets which has no defined set of rows and columns[3]. Thus, **Figure 1.4** gives the view of unsupervised learning, which consists of a variety of datasets that are identified and grouped according to similar or different clusters among the classification tasks.

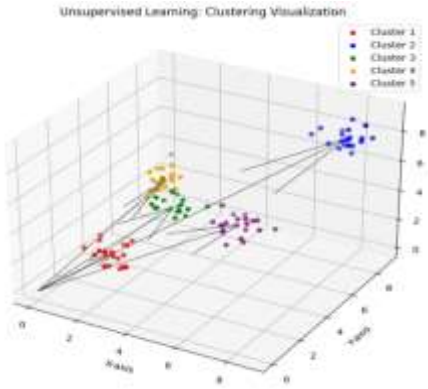

Figure 1.4: Unsupervised Machine Learning from the collection of raw input to output.

1.1.3 Semi-supervised Learning:

Semi-supervised learning falls with both the availability of the labelled and unlabeled datasets[4]. This is the analysis of the data set without the rows and columns. **Figure 1.5** discusses both the labelled and unlabeled columns of data set, which gives us the required analysis with some accuracy and precision of the model.

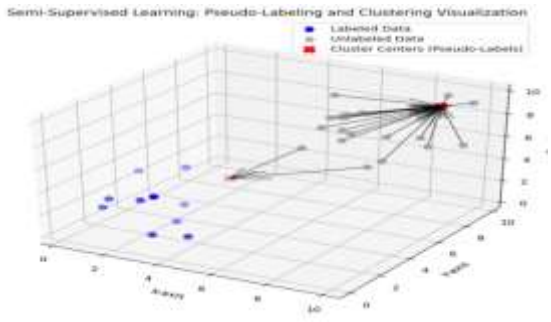

Figure 1.5: Semi-supervised machine learning used in the real-world data.

2. Literature Survey

Several studies have shown the importance of machine learning techniques for the classification and ranking of tasks related to water quality monitoring and environmental assessment[4]. This section involves the relevant work and the approach used in the ML models for the tasks involved in separating the output based on the various water quality parameters.

Support vector machines and XGBoost have shown promising results in various environmental tasks. Nouraki, A.; Alavi,M[5] compared the performance of XGBoost with Random Forest (RF) and SVM for classification of satellite and aerial imagery. Their study found that XGBoost outperformed RF and SVM, especially with larger sample sizes, which supports our choice of XGBoost for H2O ranking. A study made by Ambade, B.; Sethi, S.S[6], demonstrated the use of XGBoost in achieving the lower root mean square(RMS) when compared to the other machine learning classifiers.

They analysed samples for parameters such as pH, total hardness, calcium, magnesium, chloride, total dissolved solids, iron, fluoride, nitrate, and sulfate[6]. Their approach of considering multiple water quality parameters informs our feature selection process for H_2O ranking.

2.1 Automated Machine Learning Approaches:

Some of the recent advances in the field of machine learning have paved the way for making the automated machine learning training and testing of the model depending upon the various approaches[7]. **Figure 1.6** outlines the overall schema of the automated machine learning mode. The machine learning approaches in the automation have been used widely, and the results in precise are more precise than the manual model building[8]

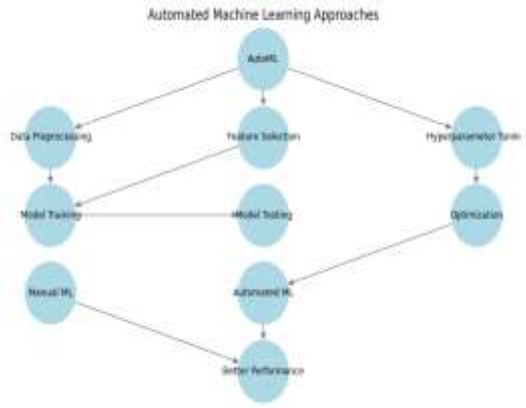

Figure 1.6: The schema of the automated machine learning approaches in CNN.

2.2 Ensemble Methods for Water Quality Assessment

Ensemble methods in machine learning are also one of the techniques used for the classification and regression tasks. It combines some of the good multiple machine learning models, which improves the accuracy[8] and precision over the training dataset with the samples involved.

3. Proposed Methodology

3.1 System Architecture:

In my research project of research I will use the two advanced models of the ML algorithm, namely the SVM and XGBoost, which will create a great impact while handling the large datasets and usage of the classification model used for the estimation of the various parameters of water. From water bodies, feature selection and splitting data in the 80/20 rule of model building to get accurate results. From **Figure 1.7,** the architecture model of the water classification analysis using an ML model.

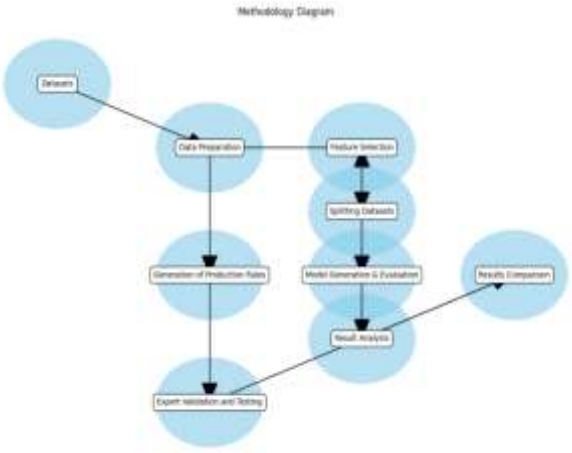

Figure 1.7: Architecture model of the water classification using an ML algorithm.

3.2 Support Vector Machine(SVM):

Support vector machine(SVM) is a powerful machine learning algorithm that is used primarily for classification tasks. The core idea behind the model of the SVM is to identify the optimal hyperplane that best separates the data points of the different classes in a higher-dimensional space[9]. Thus by the usage of this model the classification of the water can be done along with the various parameters considered.

3.3 Mathematical Formulation:

The equation of the linear boundary for the hyperplane can be mathematically written as,

$$wx+b=0$$

where,

w is the weight vector
x is a feature vector and
b is the bias term

The primary goal of the SVM is to maximise the margin, which is defined by

$$\text{Margin}=\frac{2}{|w|}$$

On minimising the above objective function, it becomes as follows.

$$\min \frac{1}{2}|w|$$

Subject to constraints,

$$y_i(wx_i + b) \geq 1 \quad \forall i$$

where

Y_i is the class label for each instance, and

X_i are the input feature vectors

Formulation:

The primary objective of the XGBoost can be explained as follows,

Loss Function: The loss function measures how well the model's predicted scores match the actual scores.

$$L(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2$$

Where

Y is the true value
Y^ ^ is the predicted value for the model.

Regularisation: To prevent the overfitting of the model and to improve the precision

$$Obj = L + \sum_{k=1}^{K}\Omega(f_k)$$

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda\left|\left|w\right|\right|^2$$

Where L is the loss function of the model.

3.4 Regularisation parameter:

The regularisation parameter in the SVM is denoted by the letter 'C'. It controls the trade-offs between the margin and the misclassification of the error[10]. Thus, the high value of C typically means the model will correctly make the classification of all the parameters from the samples, leading to the avoidance of overfitting the model. From the **Figure 1.9** the most important regularisation parameter is considered for the SVM, which is the kernel function and decision boundary which separates the hyperplane surface of model.

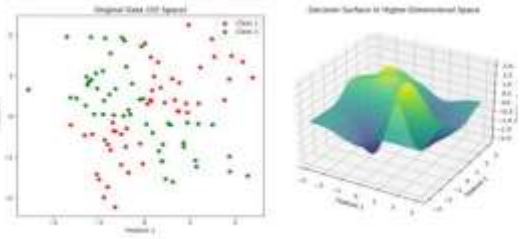

Figure 1.9: Decision surface of the kernel function from the SVM machine learning model.

3.5 Tuning of XGBoost Model

3.5.1 Maximum Depth:

The maximum depth parameter of the XGBoost controls the maximum depth of each tree in the ensemble learning of the model. Thus, if the model is not balanced correctly with the depth estimator[10], it can lead to the overfitting of the machine learning algorithm. In **Figure 2.1,** the training and validation score lies in the same plane of the hyperparameter model, which implies that the model performs well for the dataset.

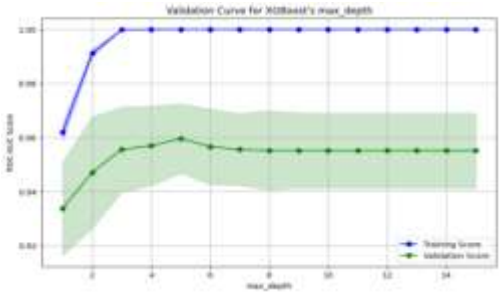

Figure 2.1: Maximum depth hyperparameter concerning the ROC-AUC score of the XGBoost model.

3.6 Learning Rate:

Learning rate is also one of the important parameters of the XGBoost model, which determines how quickly the model learns from the data. If the amount of data available is high, it will lead to the slow convergence of the data as it takes more time to read the data and also applies same for the testing and training of model.

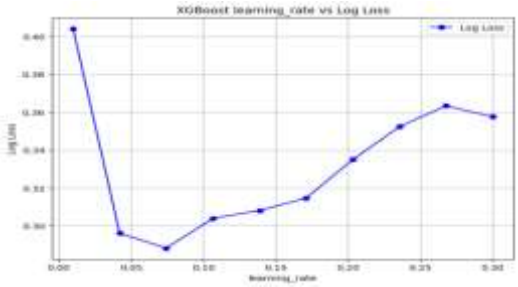

Figure 2.1: XGBoost learning rate concerning log loss of the machine learning model.

3.7 N_estimators

The n_estimators parameter is the one that controls the number of decision trees in the ensemble.

Process of Tuning: The normal values of n estimators range from 50 to 500, depending upon the size of the data[11] and the computational time required to do so with the model. **Figure 2.2** discusses the XGBoost training concerning the n_estimators and maximum depth of the model.

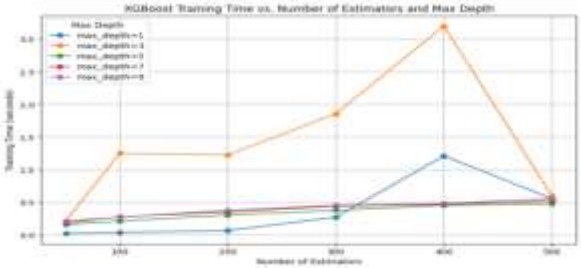

Figure 2.2: The training results of the XGBoost Machine learning model about n_estimators used in the dataset.

4. Results

This section discusses the results of the proposed methodologies for the classification of the water based on the different water quality parameters, and further creation of websites with the monitoring of water quality assessment using the AWS. The water collection samples are taken from the numerous places from the Indian country of ground water rivers compromising of all states from the continent.

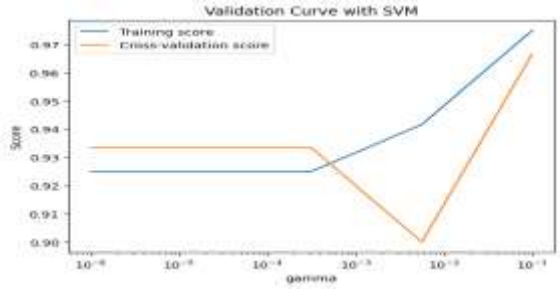

Figure 2.3: Validation score of the SVM model plotted with the gamma values considered from the data set.

Figure 2.3 gives the final training and validation score after building the model which implies SVM model performance was not good with the accuracy of 78%. The performance metrics such as the Precision, recall, F1-Score which are important parameters need to be considered in the model building of the machine learning algorithm[11].

From **Figure 2.4** the score of the XGBoost model performed well achieving the accuracy of about 93% which is good for the classification of the model with the various water quality parameters considered for the data set..

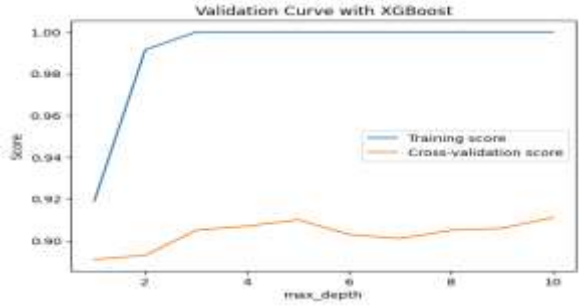

Figure 2.4: Cross-validation and Training score of the XGBoost with max_depth considered for the model.

From the above, of the two models, the accuracy precision obtained from the SVM model is 76%, while the XGBoost model provides the accuracy result of 96%, which indicates that the XGBoost classifier model performed well with the dataset.

5. Conclusion

In conclusion, this project has successfully developed and implemented the water quality classification model and made a stratification analysis using the two advanced machine learning algorithms, namely the SVM and XGBoost methods. Currently, there are several methods of classifying water based on various parameters, which involve several labour processes, and the error and accuracy of the report are also not up to the mark. The recent advancements in deep learning and NLP have enabled it to take accurate motion pictures from the various sensors and can easily predict and build the required classification models for the various water parameters.

Statements & Declarations

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflicts of interest

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organisations, or those of the publisher, the editors and the reviewers.

References

[1] Muhammad, S.Y., Makhtar, M., Rozaimee, A., Aziz, A.A. and Jamal, A.A. (2021) Classification model for water quality using machine learning techniques. Int. J. Softw. Eng. It's Appl. 2021, 9, 45–52.

[2] Radhakrishnan, N., and Pillai, A.S. (2020) Comparison of water quality classification models using machine learning. In Proceedings of the 2020 5th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 10–12 June 2020; pp. 1183–1188.

[3] Walley, W. and Džeroski, S. (1996) Biological monitoring: A comparison between Bayesian, neural and machine learning methods of water quality classification. In Environmental Software Systems; Springer: Berlin/Heidelberg, Germany, 1996; pp. 229–240.

[4] Nasir, N., Kansal, A., Alshaltone, O., Barneih, F., Sameer, M., Shanableh, A., and Al-Shamma'a, A. (2022) Water quality classification using machine learning algorithms. J. Water Process Eng. 2022, 48, 102920.

[5] Nouraki, A., Alavi, M., Golabi, M., and Albaji, M. (2021) Prediction of water quality parameters using machine learning models: A case study of the Karun River, Iran. Environ. Sci. Pollut. Res. 2021, 28, 57060–57072.
[6] Ambade, B., Sethi, S.S., Giri, B., Biswas, J.K. and Bauddh, K. (2022) Characterisation, behaviour, and risk assessment of polycyclic aromatic hydrocarbons (PAHS) in the estuary sediments. Bull. Environ. Contam. Toxicol. 2022, 108, 243–252
[7] Singha, S., Pasupuleti, S., Singha, S.S., Singh, R. and Kumar, S. (2021) Prediction of groundwater quality using efficient machine learning technique. Chemosphere 2021, 276, 130265.
[8] Brown, R.M., McClelland, N.I., Deininger, R.A. and Tozer, R.G. (1970) A water quality index- do we dare? Water Sew. Work. 1970, 117, 339–3
[9] Bui, D.T., Khosravi, K., Tiefenbacher, J., Nguyen, H, and Kazakis, N. (2020) Improving prediction of water quality indices using novel hybrid machine-learning algorithms. Sci. Total Environ. 2020, 721, 137612.
[10] Juna, A., Umer, M., Sadiq, S., Karamti, H., Eshmawi, A., Mohamed, A. and Ashraf, I. (2022) Water Quality Prediction Using KNN Imputer and Multilayer Perceptron. Water 2022, 14, 2592.