Analogy of H_2O ranking and its stratification using SVM and XGBoost method

Surya Ravichandran[1]

Department of Computer Science & Engineering, SRM Institute of Science and Technology, Kattankulathur, Chennai - 603203

***Abstract*—Water is an important part of the human being and the living society. Over the number of years water has been contaminated by the various ways of the air and water pollution. This makes the content to be unhygienic and harmful for the 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 is to develop the classification of the water quality using the Gradient boosting classifier. My research involves considering of the various parameters of H_2o including the pH, dissolved oxygen, Total Dissolved Solids (TDS), temperature which is predominant for the ranking of water contents.**

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

I. INTRODUCTION

Water is a critical part that plays an important role in every part of the life. The water is used for the various purposes as such as drinking, irrigation, industries, and aquatic life maintenance. However, the quality of the water us 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. With the help of the advanced machine learning algorithms developed along with the deep learning one can predict the categories and can know the difference in the presence of the organic reactions involved with the living things.

In the Figure 1.1 the traditional programming of the computer is discussed where the instructions and what are the tasks to be done to get the output and results for the problem [1] are mentioned. Thus, as the data is small it doesn't give much problem in the system, but while the number of the output grows the analysis becomes complex and more rules need to be written for the above.

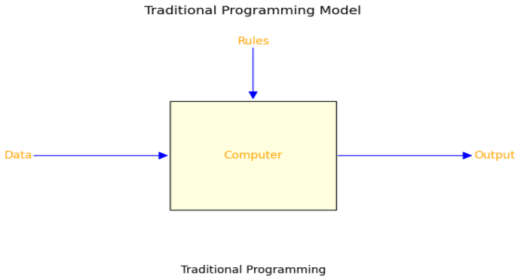

Figure 1.1. Traditional Programming.

without being explicitly programmed for the above. This the programmers don't need to write new set of rules every time when the new data comes. Machine learning is employed for the need of the large analysis of the dataset to be involved [2], they provide us excellent results without being commanded and named for the above. In the Figure 1.2 the machine learning approach of the model building is given which learns from the data to produce the result. This as my project is concerned, I will be utilizing the two important algorithms for the models namely the SVM and XGBoost method which can handle the large datasets and give us some good results for the above.

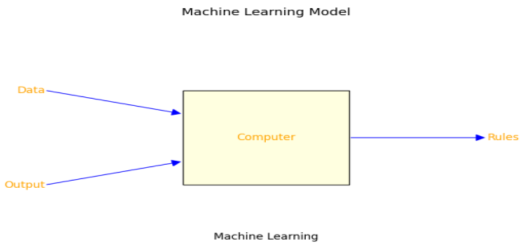

Figure 1.2: Machine learning algorithm.

Types of Machine Learning algorithm

Supervised Learning:

Supervised learning algorithm is one of the types of machine learning which has labeled input data and we can predict the output by building the model required for the solution [2].

Thus, there are defined rows and columns which helps to predict the score and accuracy for the model. In the Figure 1.3 the ML model and Test data are visualized in the 3D for the supervised learning which gives the overall importance of named labelled columns of data used in this type of machine learning approach. Thus, supervised machine learning is used in some of the industrial applications while the data may be in the all formats while solving for the business problem in the industry.

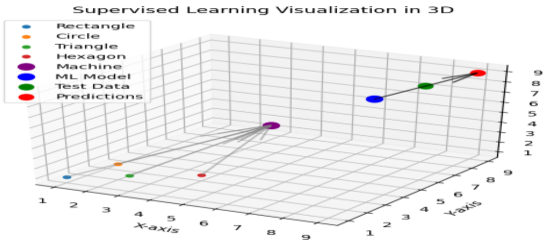

Figure 1.3: Implementation of Supervised Learning using Machine Learning Algorithm.

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 all variety of the datasets which are identified and grouped according to the similar or the different clusters among the classification tasks.

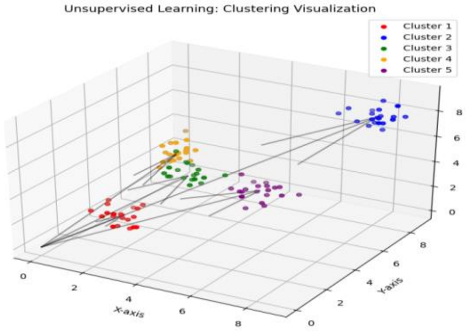

Figure 1.4: Unsupervised Machine Learning from collection of raw input to output.

Semi-supervised Learning:

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

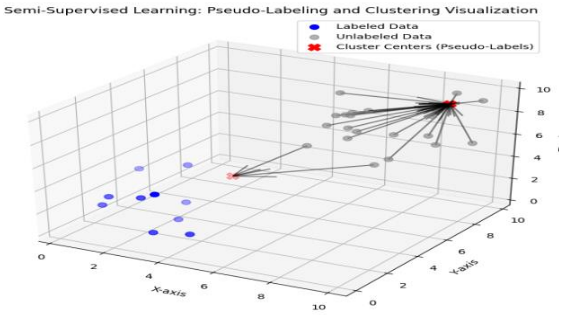

Figure 1.5: Semi-supervised machine learning used in the real-world data.

II. LITERATURE SURVEY

Several studies have shown the importance of machine learning techniques for the classification and ranking of the tasks related to the 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 the promising results in the 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 the 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 analyzed 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.

Automated Machine Learning Approaches:

Some of the recent advanced 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**].** The Figure 1.6 outlines the overall schema of automated machine learning mode. The machine learning approaches in the automation has been used widely and the results in precise are better than the manual model building [8] with the many of the errors accumulated during the testing and training of the model.

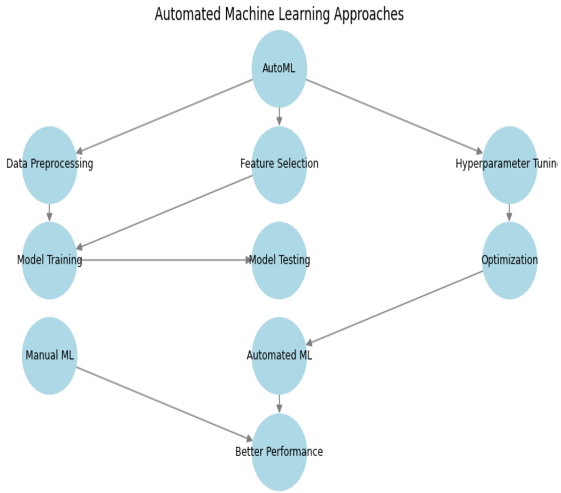

Figure 1.6: The schema of the automated machine learning approaches in CNN.

Ensemble Methods for Water Quality Assessment

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

The hierarchical feature learning in the CNN is the one of the predominant techniques which learns itself from the complex data sets and understanding.

It can also analyze the position in the input space regardless of the number of input water quality parameters given for the training data set. CNN can integrate the datasets across the various domains such as spatial data, satellite imaginary [9], maps and water with traditional parameters.

III. PROPOSED METHODOLOGY

System Architecture:

In my project of research, I will using the two advanced models of the ML algorithm namely the SVM and XGBoost which will create a great impact while handling with 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 the accurate results. From the Figure 1.7 the architecture model of the water classification analysis using ML model is projected, finally after the building of the model thus the two algorithm is compared on the basis of which has given the best results for the classification and model prediction for the above-mentioned datasets collected from the various organization and government water bodies.

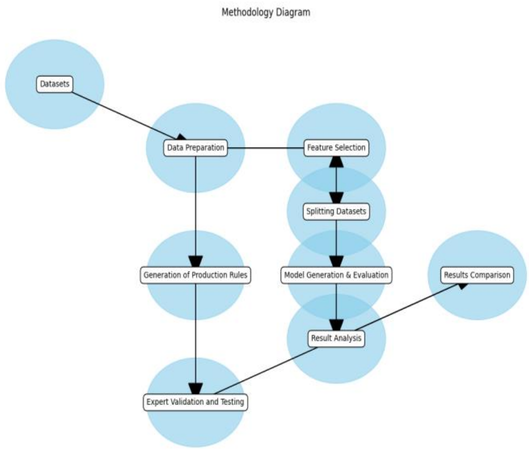

Figure 1.7: Architecture model of the water classification using ML algorithm.

Support Vector Machine (SVM):

Support vector machine (SVM) is the powerful machine learning algorithm which is used primarily for the 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. The hyperplane is the decision boundary the separates the different classes in the future space. Support vectors are the one which lie closest to the hyperplane [9]. These points are critical because they define the position and orientation for vectors.

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 feature vector and
b is the bias term

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

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

On minimizing the above objective function, it becomes as follows

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

With subject to constraints,

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

where

Y_i is the class label for each of the instance and
X_i are the input features vectors

Soft Margin and Regularization of the SVM:
As the number of the data grows thus the more error and less accuracy in the model may be induced. Thus, the SVM makes the soft margin over the dataset's which aims to maximize the margin and increases the efficiency of the process.

$$C_i = \frac{1}{\sum_{j=1}^{n} \xi_i}$$

Where ξi are the slack variables representing the misclassifications

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

Loss Function: The loss function measures how well of the model predicted scores match with the actual scores. The common choice of notation comes with the MSE for the regression tasks.

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

Where

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

Gradient descent: The algorithm uses the gradient descent primarily to reduce the loss function effectively.

$$g_i = \frac{\partial L(y_i, \hat{y_i})}{\partial \hat{y_i}}$$

$$h_i = \frac{\partial^2 L(y_i, \hat{y_i})}{\partial \widehat{y_i^2}}$$

Regularization: To prevent the over fitting of the model and to improve the precision

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

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

Where L is the loss function of the model.

Feature Engineering & ML model
Feature engineering is one of the crucial steps in machine learning that involves selecting and transforming into the features more suitable for the

modeling depending upon the dataset available. In the **Figure 1.8** the heat map gives the importance of various water quality parameters considered for model building from the data set.

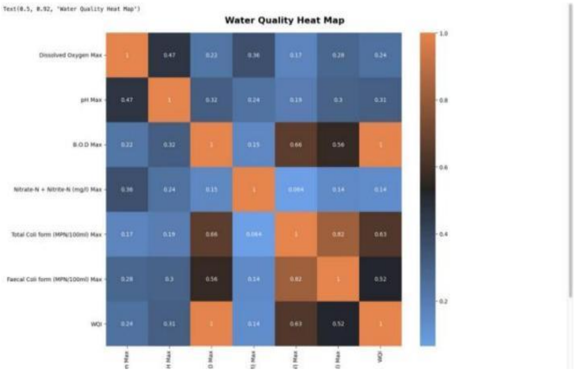

Figure 1.8: Water Quality Heat Map of the various parameters from the dataset.

Regularization parameter:

The regularization parameter in the SVM is denoted by letter 'C'. It controls the trade-offs between the margin and the misclassification of the error [10]. Thus, high value of C typically means the model will correctly make the classification of all the parameters from the samples leading to the avoiding of overfit of the model. From the Figure 1.9 the most important regularization parameter is considered for the SVM which is kernel function and decision boundary which separates the hyperplane surface of model. Thus, the decision surface of the kernel lies in the hyperplane which implies there is no over-fit of SVM model which gives a mediocre result.

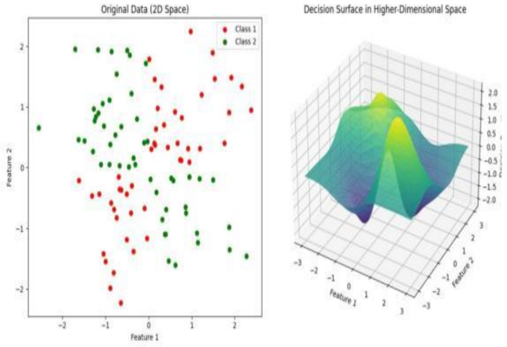

Figure 1.9: Decision surface of the kernel function from the SVM machine learning model.

Tuning of XGBoost Model

The XGBoost is one of the powerful advanced machine learning algorithms used for the classification tasks when the number of the data available is more and it requires the careful tuning of the hyper-parameters to achieve the optimal performance for the model [10].

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 overfit of the machine learning algorithm. In the Figure 2.1 the training and validation score lies in the same plane of the hyper-parameter model which insists that the model performs well for the data set.

Thus, maximum depth parameter is also associated with the ROC-AUC curve of the XGBoost hyper-parameter tuning which discusses the importance of the choosing the depth of the machine learning model to be built for the getting the better accuracy and prediction score which tells us the model score of the XGBoost in the water classification and stratification analysis of the data set.

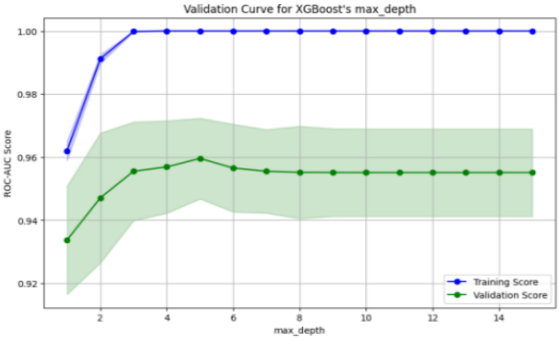

Figure 2.1: Maximum depth hyper-parameter with respect to the ROC-AUC score of XGBoost model.

Learning Rate:

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

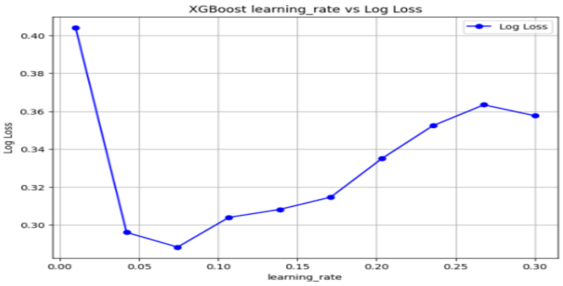

Figure 2.1: XGBoost learning rate with respect to log loss of the machine learning model.

N_estimators

The n_estimators parameter is the one which 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. The Figure 2.2 discuss the XGBoost training with the respect to the n_estimators and maximum depth of the model which is crucial for the validation for the accuracy score and classification report for the data set.

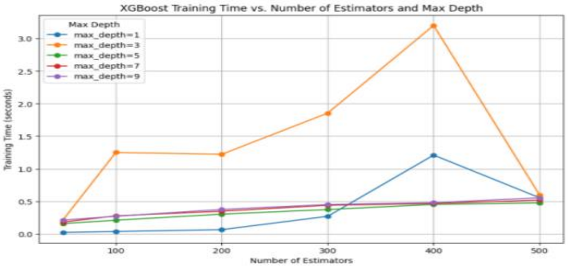

Figure 2.2: The training results of XGBoost Machine learning model with regards to n_estimators used in the data set.

IV. RESULTS

This section discusses the results of the proposed methodologies of 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.

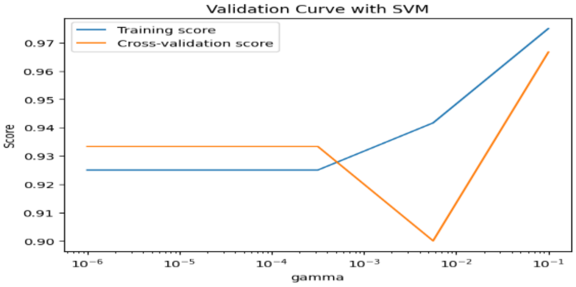

Figure 2.3: Validation score of the SVM model plotted with the gamma values considered from the data set.

The Figure 2.3 gives the final training and validation score after building of 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]. Precision is one of the predominant parameters which describes how many of the predicted positive instances are actually positive while the recall measures how many of the actual positives are correctly identified for the instance of the model.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision+Recall}}$$

The kernel is a crucial component of Support Vector Machines (SVMs), as it allows the algorithm to operate in higher-dimensional spaces without explicitly computing the coordinates of the data points in that space [12]. This is known as the "kernel trick."

From the 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. The max_depth is considered predominantly as the number of trees grows in the model it leads to the error of the model which may reduce the accuracy and precision of the XGBoost when compared to SVM of the model.

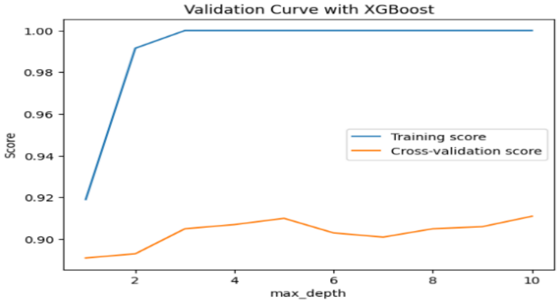

Figure 2.4: Cross validation and Training score of the XGBoost with max_depth considered for the model.

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 XGBoost classifier model performed well with the data set.

V. 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 algorithm namely the SVM and XGBoost method. Currently there are several methods of making the classification of water based on the various parameters which involves the several labor process and the error and accuracy of the report is also not up to the mark. The recent advancements in the deep learning and NLP have made it to take the accurate motion pictures from the various sensors and can easily predict and build the required classification models for the various parameters of the water.

REFERENCES

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

[2] Radhakrishnan, N.; Pillai, A.S. 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.; Džeroski, S. 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.; Al-Shamma'a, A. Water quality classification using machine learning algorithms. J. Water Process Eng. 2022, 48, 102920.

[5] Nouraki, A.; Alavi, M.; Golabi, M.; Albaji, M. 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.; Bauddh, K. Characterization, behavior, 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.; Kumar, S. Prediction of groundwater quality using efficient machine learning technique. Chemosphere 2021, 276, 130265.

[8] Brown, R.M.; McClelland, N.I.; Deininger, R.A.; Tozer, R.G. A water quality index-do we dare. Water Sew. Work. 1970, 117, 339–343.

[9] Bui, D.T.; Khosravi, K.; Tiefenbacher, J.; Nguyen, H.; Kazakis, N. 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.; Ashraf, I. Water Quality Prediction Using KNN Imputer and Multilayer Perceptron. Water 2022, 14, 2592.

[11] Aldhyani, T.H.; Al-Yaari, M.; Alkahtani, H.; Maashi, M. Water quality prediction using artificial intelligence algorithms. Appl. Bionics Biomech. 2020, 2020.