iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** XII **Month of publication:** December 2022

DOI: https://doi.org/10.22214/ijraset.2022.47944

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue XII Dec 2022- Available at www.ijraset.com

Water Quality Analysis Using Machine Learning

Annaji Kuthe[1], Chaitanya Bhake[2], Vaibhav Bhoyar[3], Aman Yenurkar[4], Vedant Khandekar[5], Ketan Gawale[6]
[1, 2, 3, 4, 5, 6]*Dept.of Computer Science & Engineering, KDKCE Nagpur, India*

Abstract: ***Human beings are dependent on the natural resources that stands for its quality. Climatic changes and environmental impacts have always been under observation. The quality of the products is always measured before use. Water, an inevitable resource ,has got a serious significance in checking its quality due to the influence of various external factors like industrial effluents, acid rain etc. This paper provide same theology for assessing the water quality that uses statistical quality control technique and Machine Learning algorithms to scale up the classification accuracy. The classification is focused on deciding if the water is suitable for drinking purpose. The goal of this study is to develop a water quality prediction model with the help of water quality factors using Artificial Neural Network (ANN) and time-series analysis.***
Keywords: Water quality predication, Machine learning algorithms, Artificial neural network, Time series analysis.

I. INTRODUCTION

Natural water resources like groundwater and surface water have always been the cheapest and most widely available resources of freshwater. However, these resources are also most likely to become contaminated due to various factors including human, industrial and commercial activities as well as natural processes. In addition to that, poor sanitation infrastructure and lack of awareness also contributes immensely to drinking water contamination [2]. The effect so water quality deterioration are far-reaching, impacting health environment and infrastructure in a very adverse manner. According to United Nations (UN) ,water borne diseases cause death of more than 1.5 million people each year[1],much greater than deaths caused by accidents, crimes and Is. terrorism combined. Therefore, it is very crucial quality trends. In order to carry out useful and efficient water quality analysis and predicting the water quality patterns, it is very significant to include atemporal dimension to the analysis, so that the seasonal variation of water quality is addressed [2].

Moreover, recent studies have shown that a suitable hybrid of multiple models for forecasting gand prediction gives better results than using a single one. Different methodologies have been proposed and applied for analysis and monitoring of water quality as well as time series analysis [1][2]. However, the non-linear nature of water quality data, as in this research, makes it very complex to map input-output data and predict future water quality.

The focus of this work is the attribute reduction for assessing the quality based on range charts and the result so analyzing data obtained from Machine Learning algorithm. Water quality prediction results could be implemented on LAN through Android Application for better utility [3].

The water analysis is done due to the scarcity of the pure natural water and the increased usage of mineralized waters in the market. The features of the proposed model include data aggregation of the values of certain parameters from sensors. Applying those to range chart analysis, to check if the values fall within the control limits [5]. The mean of the collected data is computed with respect to the area as nature of the water Changes with respect to its geographic allocation.

II. RELATED WORK

In the existing system, implementation of machine learning algorithms is bit complex to build due to the lack of information about the data visualization. Mathematical calculations are used in existing system for model building this may takes the lot of time and complexity [1]. To overcome all this, we use machine learning packages available in the scikit-learn library. Previous studies have shown that the richness and quality of data determines the accuracy and reliability of analysis. Since most of the water monitoring organizations have lack of detail and insufficient observations, we have opted for the acquisition of data from one of the most reliable water resources in the world which is usually pre-processed and frequently updated [1]. So the major disadvantage is the previous process is highly complex and time consuming.

III. PROPOSED WORK

The methodology used in this study comprises of Machine learning with training and testing data from USGS online data repository.

Therefore, we propose a machine learning-based approach which combines anew technique of preprocessing the data for features transformation, resampling techniques to eliminate the bias and the deviation of instability and performing classifier tests based [5]. By proposing the algorithms we increase the theoretical background of the methodology and algorithm used is as follows:

A. Artificial Neural Network

ANN has been widely acknowledged a same theology for classification of complex datasets such as those of environmental processes. It has the ability to efficiently describe the non-linear relationship of the complex water quality datasets [14]. Moreover, it has strong adaptability to depict the Changes that might occur in the water environment of a particular area. The algorithmic architecture of ANN attempts to simulate the structure and networks in a human brain, with an input layer, hidden layer and output layer each consisting of nodes. There might be one or more hidden layers, depending upon the problem at hand [6,7]. The importance of Decision making has been reported by many researchers in varied fields. Some of them being E-LEACH protocol, Smart Home Appliances Controller Using IOT [10,11].

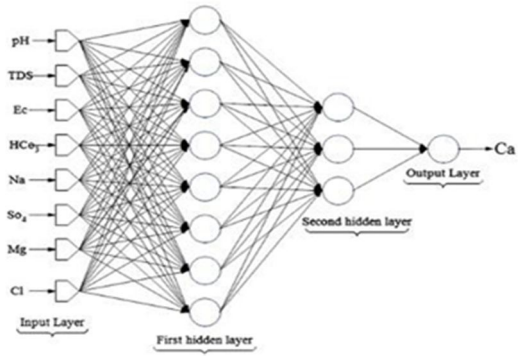


Fig.(a) Structure of ANN model for predicting water quality component (Ca) [14].

B. Guassian Naïve Bayes

Naïve Bayes is a simple and a fast algorithm that works on the principle of Bayes theorem with the assumption that the probability of the presence of one feature is unrelated to the probability of the presence of the other feature [8].

C. Decision Tree

A decision tree is a simple self-explanatory algorithm, which can be used for both classification and regression. The decision tree, after training, makes decisions based on values of all the relevant input parameters. It uses entropy to select the root variable, and, based on this, it looks towards the other parameters' values [9]. It has all the parameter decisions arranged in a top-to-down tree and projects the decision based on different values of different parameters.

D. Support *Vector* Machine

Support vector machines (SVMs) are mostly used for classification but they can be used for regressions well. Visualizing data point spotted on a plane, SVMs define a hyperplane between the classes and extend the margin in order to maximize the distinction between two classes, which results in fewer close miscalculations [9][17]

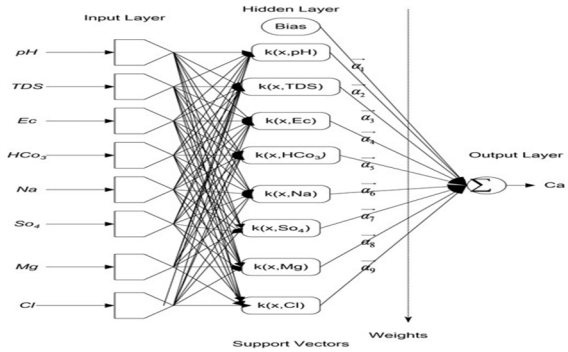


Fig.(b)Structure of SVM model for predicting water quality component (Ca) [9].

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue XII Dec 2022- Available at www.ijraset.com

E. K Nearest Neighbor

The K nearest neighbor algorithm classifies by finding the given points nearest N neighbors and assigns the class of majority of n neighbors to it. In the case of a draw, one could employ different techniques to resolve it, eg. increase n or add bias towards one class [12]. K nearest neighbor is not recommended for large datasets because all the processing takes place while testing, and it iterates through the whole training data and computes nearest neighbors each time. We used a n = 5 configuration for our model [18].

F. Random Forest

Random forest is a model that uses multiple base models on subsets of the given data and makes decisions based on all the models [14]. In random forest, the base model is a decision tree, carrying all the pros of a decision tree with the additional efficiency of using multiple models [18].

ADVANTAGESOFPROPOSEDSYSTEM

- The data analysis is done on the data set proper variable identification is also done.
- Then proper machine learning algorithms are applied to the dataset where the pattern of datais learned.
- After applying this algorithms were duces time complexity compare to existing system and also increase to its higher accuracy. From 70% to up to 94% Accuracy.

ACCURACY MEASURESUSE (Evaluation matrix)

In this section, prior to discussing the results, we will describe different measures used to assess the accuracy of the applied machine learning algorithms. As mentioned earlier, this research employed two types of supervised machine learning algorithms, i.e. regression and classification. The results yielded by both types of algorithms were evaluated differently. For regression, we used the following measures:

1) Mean Absolute Error (MAE)

Mean absolute error (MAE)is a measure of accuracy for regression. It sums up absolute values of errors and divides them by the total number of values. It gives equal weight to each error value. The formula for calculating MAE is shown in Equation, where xobs refers to the actual value, xpred refers to the predicted value, and refers to the total number of samples considered.

$$MAE = \frac{\sum(|x_{obs} - x_{pred}|)}{n}$$

2) Mean Square Error (MSE)

Mean square error(MSE)is the sum of squares of errors divided by the total number of predicted values. This attributes greater weight to larger errors. This is particularly use fulin problems where there need strobe larger weight for larger errors .It is measured, where xobs is the actual value, xpred is the predicted value, and n is the total number of samples considered

$$MSE = \frac{\sum(Xobs\text{-}Xpred)^2}{n}$$

3) Root Mean Squared Error (RMSE)

Root mean squared error (RMSE) is just the square root of MSE and scales the values of MSE near to the range so f observed values. It is estimated from Equation, where xobs points to the actual value, xpred points to the predicted value, and n points to the total number of samples considered.

$$MAE = \sqrt{\frac{\sum(x_{obs} - x_{pred})^2}{n}}$$

4) R Squared Error (RSE)

R squared error (RSE), also known as the coefficient of determination, and often denoted as R2, determines the goodness of fit of the model. It particularly explains the amount of variance of the dependent variable that is explainable through the independent variable, as shown in Equation. Higher RSE values mean that the independent variables largely explain the variance of the dependent Variable.

$$RSE\ or\ R^2 = 1 - \frac{Explained\ Variation}{Total\ Variation}$$

For classification, we used the following measures:

a) Accuracy: Accuracy is the correct number of predictions made by the model overall the observed values. Accuracy is measured by Equation, where TP refers to true positive, TN refers to true negative, FP refers to false positive and FN refers to false negative.

$$Accuracy = \frac{TP + TN}{TP + FP + TFN}$$

b) Precision: Precision is the proportion of correctly classified instances of a particular positive class out of the total classified instances of that class. Precision is calculated with the formula shown in Equation, where TP refers to true positive and FP refers to false positive.

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

c) Recall: Recall is the proportion of instances of a particular positive class that were actually classified correctly. Recall is calculated with the formula shown in Equation , where TP refers to true positive and FN refers to false negative.

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

d) F1 Score: As precision and recall, individually, do not cover all aspects of the accuracy, we took their harmonic mean to reflect the F1 score, as shown in Equation, which covers both aspects and reflects he overall accuracy measure better. It ranges between 0 and1. The higher the score, the better the accuracy.

$$Score = \frac{2 \times Precsion \times Recall}{Precision + Recall}$$

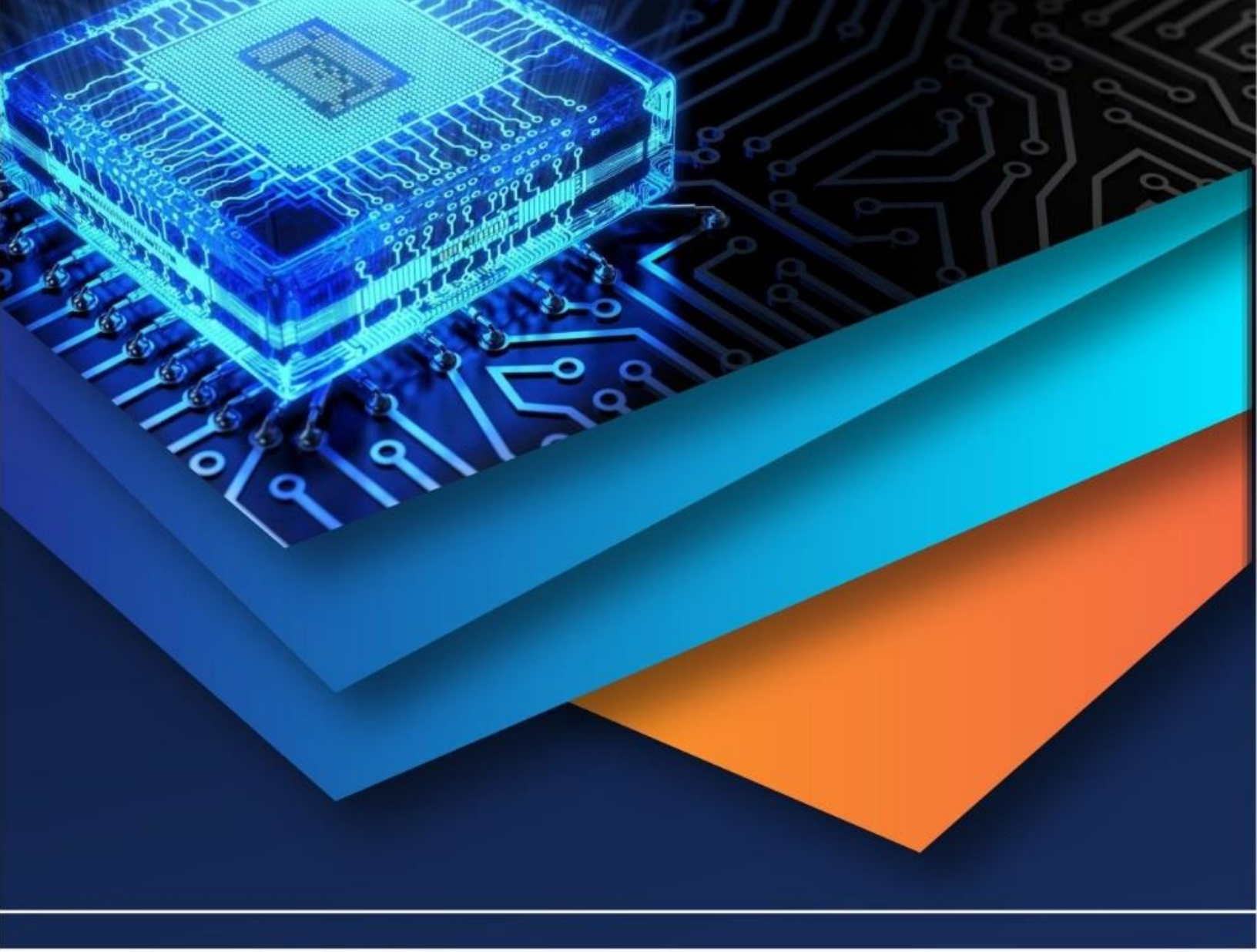
IV. CONCLUSIONS

Overall, the goals defined for this research were reached and the examples of the application of machine learning models are presented, covering most of the aspects of the average research. Generally, regression models were able to show the consistent trend and overall correlation between each other, even though for some of the measurements they give models of poor quality Random forests (RF)show the best performance and area devised for scientists and engineers working with environmental data. Artificial neural networks (ANN) are another alternative, though their performance is inferior and they are prone to overfitting. Support vector machines (SVM) are the good example for the cases where a base line mode Lis needed, being one of the basic algorithms. K-nearest neighbors (KNN) model was successfully used for data imputation and is also suggested for this task for other researchers. Though, and itis worth noticing, amount of All in all, following the technological progress and taking the best from what it provides us from day today ensures continuous development of the research field. The same goes for environmental sciences and machine-learning algorithms are one of the tools that can contribute to this field a lot and may be used to keep the progression-going.

REFERENCES

[1] Yafra Khan, Chai Soo See" Predicting and analyzing water quality using Machine Learning: a comprehensive model [IEEE] " (2016) DOI10.1109/LISA T.2016.7494106
[2] UN water, “Clean water for a healthy world,” Development, pp. 1–16, 2010.
[3] Annaji Kuthe, Tejaswini Farkade, Kalyani Rahate, Kalyani Sahare,” Monitoring and Controlling of LAN through Android Application for Network Security”, Volume 10, Issue IV, International journal for Research in Applied Science and Engineering Technology (IJRASET) Page No: 1922-1926, ISSN: 2321-9653.
[4] K. Farrell-Poe, “Water Quality & Monitoring,” pp. 1–18, 2000.
[5] T. Taskaya-Temizel and M. C. Casey, “A comparative study of autoregressive neural network hybrids,” Neural Networks, vol. 18, no. 5–6, pp. 781–789,2005.
[6] S. Maiti and R. K. Tiwari, “A comparative study of artificial neural networks, Bayesian neural networks and adaptive neuro-fuzzy inference system in groundwater level prediction,” Environ. Earth Sci., vol. 71, no. 7, pp. 3147–3160, 2013.
[7] M. Khashei and M. Bijari, “An artificial neural network (p, d, q) model for timeseries forecasting,” Expert Syst. Appl., vol. 37, no. 1, pp. 479– 489, 2010.
[8] Dewan Md. Farid, "Efficient and Scalable Multi-ClassClassificationusingNayeBayesTree",3rdinternational conference on informatics, electronics&vision2014.
[9] Quinlan, J.R. Decision Trees and Decision-making IEEE Trans. Syst. Man Cybern. 1990, 20, 339–346.
[10] A. Kuthe and A. K. Sharma, "Review paper on Design and Optimization of Energy Efficient Wireless Sensor Network Model for Complex Networks," 2021 5th International Conference on Information Systems and Computer Networks (ISCON), 2021, pp. 1-3, doi: 10.1109/ISCON52037.2021.9702421.
[11] Lonkar B. B., Kuthe A., Shrivastava R., Charde P. (2022) Design and Implement Smart Home Appliances Controller Using IOT. In: Garg L. et al. (eds) Information Systems and Management Science. ISMS 2020. Lecture Notes in Networks and Systems, vol 303. Springer, Cham. https://doi.org/10.1007/978-3-030-86223-7_11.
[12] Tong, S.; Koller, D. Support vector machine active learning with applications to text classification. J. Mach. Learn. Res. 2001, 2, 45–66.

[13] Batista,G.E.A.P.A.,andM.C.Monard.“A Study of K-Nearest Neighbor asan Imputation Method.” Soft Computing Systems :Design ,Management and Applications,(2002).
[14] Biau,Gerard.“AnalysisofaRandomForestsModel.”TheJournalofMachineLearningResearch13,no.1(2012).
[15] Dewan Md. Farid, "Efficient and Scalable Multi-ClassClassificationusingRandomForest",3rdinternational conference on informatics, electronics &vision (2014).
[16] D. Graupe, “PRINCIPLES OF ARTIFICIAL NEURAL NETWORKS,” Advanced Series on Circuits and Systems, vol. 6. World Scientific, University of lllinois, Chicago, USA, 2007.
[17] C. Min, “An Improved Recurrent Support Vector Regression Algorithm for Water Quality Prediction,” vol. 12, pp. 4455–4462, 2011.
[18] Liaw, A.;Wiener, M. Classification and regression by random Forest. R News 2002, 2, 18–22.
[19] Evaluating Machine Learning Models by Alice Zheng Released September 2015Publisher(s): O'Reilly Media, Inc. ISBN: 9781491932445.
[20] Rajiv D. Banker & Chris F. Kemerer, 2018. "Performance Evaluation Metrics for Information Systems Development: A Principal-Agent Model," Information Systems Research, INFORMS, vol. 3(4), pages 379-400,



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY