ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN2454-9940www.ijsem.org

Vol 11, Issuse.2 May 2023

Water Quality Management Using Hybrid Machine Learning And Data Mining Algorithms: An Indexing Approach

Dr.M.Suresh[1] Sk.Jubera[1],Y.Mounika ,M.Govind[2] ,J.Vineeh[3]

Abstract- One of the critical elements of worldwide water asset the executives specialists is stream water quality (WQ) appraisal. A water quality file (WQI) is produced for water evaluations thinking about various quality-related factors. WQI evaluations normally consume a large chunk of the day and are inclined to mistakes during sub-records age. This can be handled through the most recent Machine learning (ML) methods prestigious for unrivaled precision. In this review, water tests were taken from the wells in the review region (North Pakistan) to foster WQI expectation models. Four independent calculations, i.e., Random Tree (RT), Random Forest (RF), M5P, and Reduced mistake pruning tree (REPT), were utilized in this review. What's more, 12 half breed information mining calculations (a blend of independent, Bagging (BA), cross-validation parameter selection (CVPS), and randomizable Filter characterization (RFC)) were likewise utilized. Utilizing the 10-overlay cross-approval procedure, the information were isolated into two gatherings (70:30) for calculation creation. Ten irregular information changes were made utilizing Pearson relationship coefficients to recognize the most ideal blend of datasets for further developing the calculation expectation. The factors with extremely low connections performed ineffectively, while mixture calculations expanded the expectation capacity of various independent calculations.

Keywords:- *Water Quality, KNN, Navies Bayes , Random Forest, Logistic Regression, Support Vector, Gradient Boosted Decision Tree Classifier, Machine Learning.*

INTRODUCTION

Water contamination is one of the basic difficulties of the cutting edge reality where the objectives like the Assembled Countries Manageable Improvement Objectives (UN-SDGs) and a brilliant and practical planet are being sought after. All social orders, ecologies, and creations require bountiful clean water supplies for cultivating, The partner proofreader planning the audit of this composition and endorsing it for distribution was Pasquale De Meo. drinking, sterilization, and energy creation. The worldwide water emergency is among the serious dangers mankind faces nowadays. Appropriately, the amount and nature of groundwater are huge worldwide worries . Numerous sicknesses happen because of contaminated water, similar to cholera, the runs, typhoid, amebiasis, hepatitis, gastroenteritis, giardiasis, campylobacteriosis, scabies, and worm diseases. Practically 1.6 million individuals kicked the bucket because of loose bowels in 2017 alone. Water contaminations influence its circumstances, which influence human wellbeing and marine life.

Professor[1], Students[2,3,4]
Department of computer science and engineering
Qis college of engineering & technology
Ongole , Andhrapradesh, India

Deficient sewage organizations, uncontrolled and inappropriately arranged urbanization, and unloading of modern garbage, pesticides, and manures add to water contamination. Such contamination is more obvious in nearby streams or water channels nearer to metropolitan turns of events. With both non-endlessly point sources, stream contamination is turning into a more huge issue and presents an intense test to worldwide water the board specialists. Such contamination truly crumbles water quality (WQ). WQ corruption significantly influences sea-going life and the accessibility of clean water for drinking and rural purposes. In growing nations, where the economy is prone to ups and downs, the pollution problem is more eagerly addressed. Every activity that promotes improvement has potential for harmful natural outcomes. For instance, as the population grows and there is a greater desire for more resources, the need for more horticultural creation places pressure on the natural ripeness of the soils, which increases the demand for fake manures to increase yield. Similar to this, extra composts are frequently dumped into streams, contaminating surface and subsurface water supplies. The need for WQ appraisal and reconnaissance grows as a result. For the assurance of ecological, environmental, and human wellness, WQ reconnaissance and assessment are essential. This is possible with the help of wise, effective, and long-term water the executives plans. The water quality file (WQI) is used to assess the WQ. WQI aids in directing the decisions and actions of policymakers. Nevertheless, due of the input of several sub-records and situations, determining WQI is unquestionably not an easy interaction. A non-layered record called WQI is derived from identified WQ factors. It takes into consideration variables such as pH (hydrogen capacity), DO (disintegrated oxygen), TSS (all out suspended solids), Body (organic oxygen interest), AN (ammoniacal nitrogen), COD (compound oxygen interest), and others. The associated lattices enable a clear evaluation of WQ. Groundwater quality indicators (GQIs) are often evaluated using estimates of variables including Ca2+, Mg2+, NO3, and others. For the evaluation of WQ, a few components of water—physical, substance, organic, and radiological—are kept in mind. Similarly, WQI is a frequently used approach for assessing the success or failure of the executives' WQ measures. The English Columbia WQI (BCWQI), Oregon WQI (OWQI), Florida Stream WQI (FWQI), Canadian WQI (CQI), US Public Disinfection Establishment WQI (NSFWQI), In-between Time Public Water Quality Norms for Malaysia (INWQS), and others are a few examples of WQIs. On the entire planet, WQI is calculated using a variety of methods.

1. LITERATURE SURVEY

Since the last century, there has been an increase in population growth, concentrated farming, and contemporary activities, which has led to the release of eutrophic wastewaters into shoreline water bodies, severely deteriorating the water quality and posing an emergency to the marine environment (Gill et al., 2018). A 2008 study (Selman et al., 2008) found that 415 places around the world reported experiencing various eutrophic side effects. As an illustration, the greatest water blooming from central California to the Frozen North in 2015 (McCabe et al., 2016; Michalak 2016) and the longest-lasting algal sprouting (year and a half) in the Eastern Florida Narrows in 2005 (Glibert et al., 2009) are both examples. The HAB have been a significant issue in the Pacific Ocean's periphery since the turn of the previous century (Kim 1998; Li et al., 2004; Richlen et al., 2010; Al-Azri et al., 2014; Park et al., 2015). Yu et al. (2018) claim that the East China Ocean's water region may be affected by the yearly recurrent HAB occurrences, which constantly occur from early May to late June.

Water quality degradation issues in Hong Kong have been regarded as very likely the most significant hazard to the coastal water biological system since the 1980s, according to conventional models depicted in Fig. 1. Hong Kong is a typical coastal city with the ocean on three sides, where the growth of the local economy, society, and ecology may be greatly impacted by the marine environment. Over the past several years, incidents involving hazardous algal blooms (HAB) have happened as regularly as possible in the waters near Hong Kong. For instance, the largest fish kill catastrophe in Hong Kong's history, which resulted in the death of more than 3000 tonnes of fish and generated direct financial losses of more than $ 40 million USD, was attributed to a staggering algal bloom in April 1998. Both water biology and hydroponics were seriously impacted by this incident (Lee et al., 2003; Lu and Hodgkiss, 2004; Muttil and Chau, 2006; Selman et al., 2008).

2. PROPOSED METHOD

In Proposed framework the information assortment was at first performed, and followingly different WQ boundaries were determined from the water tests. The information was then appropriated into testing and approval datasets. From the testing datasets, the best information mix was distinguished. At last, various calculations were applied to the best assortments, and a calculation evaluation was led for the most ideal calculation choice to foresee WQI. logistic regression, Support vector machine, KNN, Decision tree, Random Forest, Ada Boost, Xg Boost, Gradient Boosting classifier the calculations RF, Navie bayeis, Logistic Regression (LR) and Decision tree(DT) have the most elevated forecast power. All calculations were approved as the anticipated WQI was contrasted and estimated WQI for each model at each testing dataset.

3. SYSTEM ARCHITECTURE

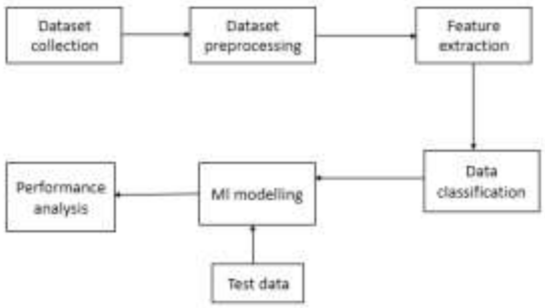


Figure 1

4. METHODOLOGY

1. *Data Collection*
2. *Data Pre-Processing*
3. *Feature Extration*
4. Assessment Model
5. *User Interface*

5.1 Data Collection

Several research that were compiled from Visa commerce records provided the data for this essay. We place more emphasis on this stage than on selecting the subset of all pertinent data that you will use. Ideally, stores of data (models or discernments) for which you are absolutely positive that you comprehend the objective response are where ML problems begin. Data that you most definitely comprehend the objective response is referred to as verified data.

5.2 *Data Pre-Processing*

Use the information you've chosen to organise it by organising, cleaning, and exploring it. There are three typical information pre-management steps:

- Formating: The information you chose might not be in a relationship that is practical for you to deal with, according to the arrangement. The information may be in a social educational file and you would like it in a level record, or it may be in a specific report plan and you would like it in a text document or social educational list.
- Cleaning: The removal or filling in of missing information is known as cleaning information. There could be information situations that lack certain pieces of information or don't pass along the data you know you truly need to make a decision.These people ought to be put to death. Also, some of the attributes may contain sensitive information that needs to be anonymized or removed from the material.
- Sampling: Clearly, there may be more carefully chosen information available than you actually wish to use. Longer evaluation times and greater processing and memory demands can be directly attributed to additional information. Before taking into account the complete dataset, you can conduct a very thorough delegation preliminary of the selected data, which may be much faster for researching and prototyping techniques.

5.3 *Feature Extration*

Next thing is to do Part extraction is a brand name decline process. Not by any stretch like part confirmation, which positions the ongoing ascribes as indicated by their wise importance, include extraction genuinely changes the qualities. The changed characteristics, or elements, are prompt blends of the primary ascribes. At last, our models are organized utilizing Classifier calculation. We use demand module on Run of the mill Language Instrument compartment library on Python. We utilize the named dataset assembled. Our other stepped information will be utilized to review the models. Some mimicked knowledge assessments were utilized to organize pre-dealt with information. The picked classifiers were Irregular timberland. These assessments are truly eminent in text blueprint attempts.

5.4 Assessment Model

Model evaluation is a crucial step in the process of improving models. It helps in determining the best model to protect our information and how well the chosen model will work moving forward. It is not advisable to evaluate model performance using the data used for preparation because doing so would undoubtedly result in overly optimistic and overfit models. Respite and Cross-Support are two techniques for assessing models in information science. The two approaches compare model execution on a test set that is hidden from the model in order to prevent overfitting. Each gathering model's effectiveness is evaluated based on how it appeared in the middle. The anticipated outcome will come to pass. diagrams that depict synchronised information.The degree of correct assumptions for the test data is referred to as precision. In most cases, it can be solved by dividing the number of true assumptions by the number of full-scale figures.

5.5 User Interface

The pattern of Information Science and Examination is expanding step by step. From the information science pipeline, one of the main advances is model sending. We have a ton of choices in python for sending our model. A few well known systems are Carafe and Django. Yet, the issue with utilizing these systems is that we ought to have some information on HTML, CSS, and JavaScript. Remembering these requirements, Adrien Treuille, Thiago Teixeira, and Amanda Kelly made "Streamlit". Presently utilizing streamlit you can send any AI model and any python project easily and without stressing over the frontend. Streamlit is very easy to use.

In this article, we will get familiar with a few significant elements of streamlit, make a python project, and convey the task on a nearby web server. How about we introduce streamlit. Type the accompanying order in the order brief.

pip install streamlit

When Streamlit is introduced effectively, run the given python code and in the event that you don't get a mistake, then streamlit is effectively introduced and you can now work with streamlit. Instructions to Run Streamlit record:

How to Run Streamlit file:

```
http://localhost:8501
```

Figure 2

Proposed Approach

1. First, we take Transactions dataset.

2. .Filter dataset as per necessities which has trait as per investigation to be finished

3. Split the dataset into preparing and testing.

4. Perform Destroyed for adjusting the information on resultant dataset framed

5. Analysis should be possible on testing information in the wake of preparing utilizing Calculated Relapse, Arbitrary Woods and Brain Organization.

At last you will obtain results as exactness measurements.

Algorithm:

1) Random Forest

Random Forest is a managed AI strategy that is outfit based. You can combine various computation types to create a more convincing forecast model, or use a similar learning technique at least a few times. The phrase "Irregular Timberland" refers to how the arbitrary woodland method combines a few calculations of the same type or different chosen trees into a forest of trees. The irregular timberland technique can be used for both relapse and characterisation tasks.

- Coming up next are the essential stages expected to execute the irregular woods calculation.
- Pick N records aimlessly from the datasets.
- Utilize these N records to make a choice tree.
- Select the number of trees you that need to remember for your calculation, then, at that point, rehash stages 1 and 2.
- Each tree in the timberland predicts the classification to which the new record has a place in the order issue. The classification that gets most of the votes is at last given the new record.
- The Advantages of Irregular Woodland
- The way that there are numerous trees and they are completely prepared utilizing various subsets of information guarantees that the irregular timberland strategy isn't one-sided.

- The irregular woods strategy fundamentally relies upon the strength of "the group," which reduces the framework's general predisposition. Since it is extremely challenging for new information to influence every one of the trees, regardless of whether another information point is added to the datasets, the general calculation isn't highly different.
- In circumstances when there are both downright and mathematical highlights, the irregular woods approach performs well.

K-At the point when information needs esteems or has not been scaled, the irregular woodland method likewise performs well.

2) SVM

Support Vector Machines (SVMs) are a sort of regulated learning calculation that can be utilized for grouping or relapse undertakings.The major idea behind SVMs is to find a hyperplane that maximally detaches the different classes in the readiness data. Finding the hyperplane with the best edge—represented as the distance between the hyperplane and the nearest pieces of information from each class—completes this process. When new information is not completely settled, it can be depicted by selecting which side of the hyperplane it falls on. SVMs are especially significant when the information has many elements, as well as when there is a reasonable edge of fragment in the information.

3) KNN

K-Nearest Neighbors is one of the most significant yet fundamental representation estimates in artificial intelligence. Plan affirmation, data mining, and interference ID are three areas where it has significant applications and fits into the supervised learning field. Since it is non-parametric—as opposed to some calculations, like GMM, which assume a Gaussian movement of the provided data—it makes no fundamental hypotheses about the scattering of data, making it widely disposable in light of any circumstances. We are provided a few prior data points (also known as planning data points), which group tasks into packs based on quality. Consider the following table of data centres, which has two components:

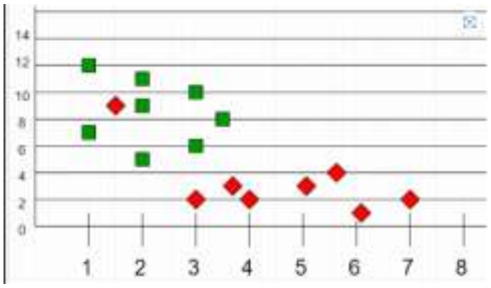


Figure 3

4) Decision Tree

Decision Tree is the most impressive and well-known tool for expectation and order. A choice tree is a flowchart-like tree structure in which each leaf hub (terminal hub) represents a class grade and each inner hub represents a test on a particular attribute.

A decision tree for the concept Take up tennis. Creation of a Choice Tree: By dividing the source data into subsets in light of a trait esteem test, a tree can be "learned". In a recursive process known as recursive division, this cycle is repeated on each predetermined subset. The recursion ends when the objective variable's value is uniformly distributed over the subset at a hub, or when splitting no longer raises expectations. Choice tree classifier development is excellent for exploratory information sharing because it does not require spatial information or boundary setting. Choice trees can handle information with several layers. Overall, the classifier's decision tree has excellent accuracy. A common inductive method to cope with learning information on characterization is choice tree enlisting.

Choice Tree Representation: Examples are arranged using choice trees by going from the root of the tree to a leaf hub, which provides the grouping of the occurrence. Beginning at the root hub of the tree, testing the characteristic specified by this hub, and then lowering the tree limb and comparing to the value of the trait as shown in the above diagram, are the steps used to characterise a case. Then, this exchange is repeated for the subtree created at the new hub. The decision tree in the preceding graphic organises certain days of the week according to whether they are appropriate for playing tennis and then returns the characterisation associated with each individual leaf.(in this situation, Yes or No).

Precision: The degree of exact assumptions for the test data is understood by precision. By keeping how much wary evaluations by the firm number of assumptions, it couldn't endlessly out forever lay exposed.

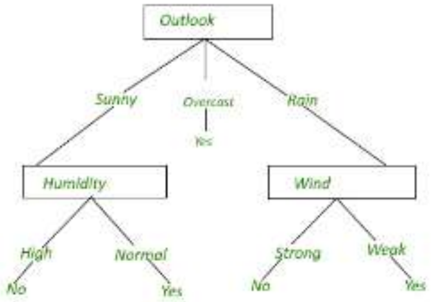


Figure 4

5) Logistic regression

oLogistic lose the faith is one of the most striking PC based insight assessments, which goes under the Planned Learning technique. It is utilized for foreseeing the straight out subordinate variable utilizing a given strategy of free factors.

oLogistic lose the faith predicts the eventual outcome of an immovable subordinate variable. In this way the result should be a flat out or discrete worth. It will overall be either Yes or No, 0 or 1, significant or Hoax, and so on in any case rather than giving the specific worth as 0 and 1, it gives the probabilistic qualities which lie a few spot in the extent of 0 and 1.

oLogistic Lose the faith is comparative as the Straight Apostatize next to that the way in which they are utilized. Direct Fall away from the faith is utilized for managing Apostatize issues, but Fundamental apostatize is utilized for taking care of the depiction issues.

oIn Decided lose the faith, instead of fitting an apostatize line, we fit an "S" framed imperative limit, which predicts two greatest qualities (0 or 1).

oThe wind from the decided limit shows the probability of something, for example, whether the phones are hazardous, a mouse is colossal or not thinking about its weight, and so on.

oLogistic Break faith is a huge reenacted knowledge calculation since it can give probabilities and solicitation new information utilizing unending and discrete datasets.

oCalculated Break faith can be utilized to orchestrate the observations utilizing various kinds of information and can unquestionably finish up the best factors utilized for the depiction. The under picture is showing the essentialcapacity:

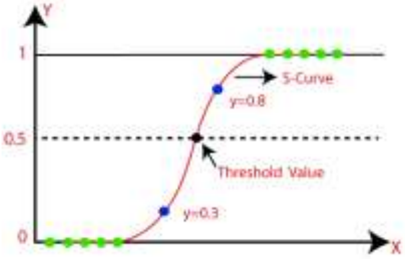


Figure 5

6) Gradiant boosted Decision Tree

Supporting means joining a learning calculation in series to accomplish areas of strength for a from many successively associated feeble students. In the event of angle supported choice trees calculation, the powerless students are choice trees. Each tree endeavors to limit the blunders of past tree. Trees in supporting are frail students yet adding many trees in series and each zeroing in on the mistakes from past one make helping a profoundly productive and exact model. In contrast to stowing, supporting doesn't include bootstrap examining. Everytime another tree is added, it fits on a changed rendition of starting dataset. Since trees are added consecutively, helping calculations advance gradually. In factual learning, models that advance gradually perform better.

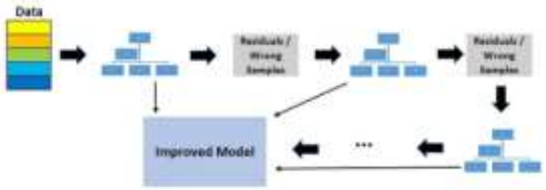
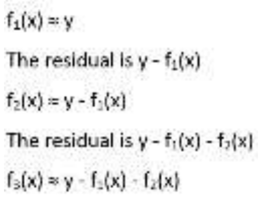


Figure 6

Angle helping calculation consecutively joins feeble students in way that each new student fits to the residuals from the past step so the model gets to the next level. The last model totals the outcomes from each step and a solid student is accomplished. Inclination helped choice trees calculation utilizes choice trees as week students. A misfortune capability is utilized to recognize the residuals. For example, mean squared blunder (MSE) can be utilized for a relapse task and logarithmic misfortune (log misfortune) can be utilized for grouping undertakings. Quite significant existing trees in the model don't change when another tree is added. The additional choice tree fits the residuals from the ongoing model. The means are as per the following:



$f_1(x) \approx y$

The residual is $y - f_1(x)$

$f_2(x) \approx y - f_1(x)$

The residual is $y - f_1(x) - f_2(x)$

$f_3(x) \approx y - f_1(x) - f_2(x)$

7) *Ada Boosting Classifier*

Ada-boost or Adaptive Boosting is one of the help group classifications made by Yoav Freund and Robert Schapire in 1996. It mixes various classifiers to improve classifier precision. AdaBoost is an iterative outfit approach. The AdaBoost classifier builds regions of strength for a, providing you high areas of strength for exactness by combining many classifiers that combine inefficiently. Adaboost's main principle is to set up the classifier loads and get ready for each cycle's information test to the point where it guarantees precise forecasts of unanticipated impressions. The fundamental classifier can be any AI computation that recognises loads on the training set. Adaboost must abide by two conditions:

Several weighed preparation models must be used to intelligently prepare the classifier.It strives to minimise training error in order to offer the best fit possible for these samples in each iteration.

What is the AdaBoost algorithm's operation? The process is as follows:

Adaboost initially chooses a training subset at random.

By selecting the preparation set in light of the precise expectation of the previous preparation, it iteratively trains the AdaBoost AI model.

It gives incorrectly characterised perceptions a heavier burden, increasing their likelihood of grouping in the attention that follows.

Additionally, it transfers the burden to the trained classifier in each emphasis in accordance with the classifier's accuracy. The classifier that is more accurate will be given more weight.

This cycle repeats until there are the predefined maximum number of assessors or until the entire preparation information fits with virtually minimal error.

Play out a "vote" involving all of the artificial learning computations to determine the ranking.

8) *Xg boost Classifier:*

XGBoost is an outfit technique pursued of different choice trees. There are 2 fundamental kinds of choice tree troupes, Sacked and Supported trees. Irregular woodland is an illustration of a sacked model where a subset of information is utilized to prepare each tree and the expectations from these trees are found the middle value of to get the last result. Helping is another procedure where choice trees are assembled consecutively and each tree gains from its ancestor, attempting to decrease the mistake in past choice trees.

XGBoost is a slope helped choice tree, an expansion of supported trees that utilizes an inclination plunge calculation. XGBoost can be utilized for arrangement and relapse like some other choice tree. It is extremely simple to carry out and for a troupe technique, XGBoost is exceptionally quick.

We will utilize the dataset utilized in for simplicity of correlation. It is a dataset of apple stock and we are attempting to foresee the cost iphone in view of the model, year, and a couple of different boundaries. We use Pandas to stack the information into the dataframe and perform introductory investigation on the information.

A few things that we ordinarily do prior to preparing a model are check for invalid qualities, indicate the highlights and target, split the dataset into train and test sets, investigate the information kind of the sections, encode all out highlights into numeric, etc. As we investigate this dataset, we observe that there are no invalid qualities in the dataset except for a couple of highlights are straight out and should be changed over completely to mathematical before an expectation can be made.

The subsequent stage is to parted the dataset into test and train and set the apple stock cost as the objective worth to be anticipated. We use scikit-learn's train_test_split for this reason. Subsequent to parting the dataset into preparing and test split, we should make a case of the relapse model in XGBoost that we will use to foresee the cost. XGBRegressor is a scikit-learn interface for relapse utilizing XGBoost.

Accuracy: The level of accurate presumptions for the test information is construed by accuracy. By confining how much cautious appraisals by the firm number of suppositions, it couldn't out and out for all time lay outed.

6. RESULTS

The process of AI demonstrating for expectation is often composed of the following few crucial steps. The immediately available information will be divided into different preparing, approval, and testing sets. A specific ML model is then chosen after information preprocessing has begun, and it will be prepared and approved with regard to the preparation set and approval set. The associated hyper boundaries will be adjusted repeatedly until they are ready before being tested with underdeveloped information.
According to the demonstrating results evaluated by error, connection, and preparation time, the full models developed using SVM techniques and all eight ecological factors given as references are offered in this part. The anticipatory displays of four learning calculations are investigated from a modelling perspective, and separately shown in Figure are the developmental cycle findings and the expected water quality. Four calculations overall produced excellent forecasts.

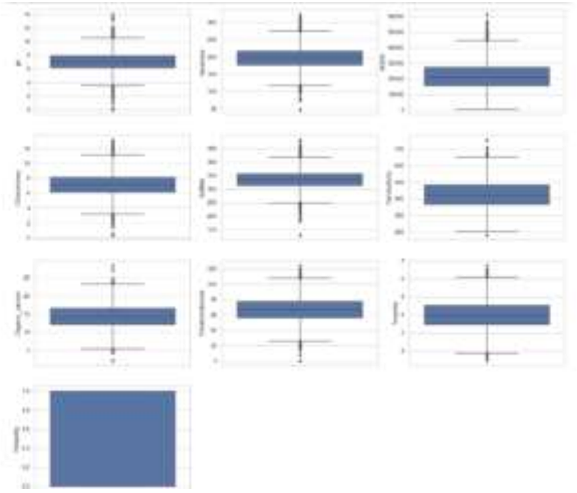


Figure 6

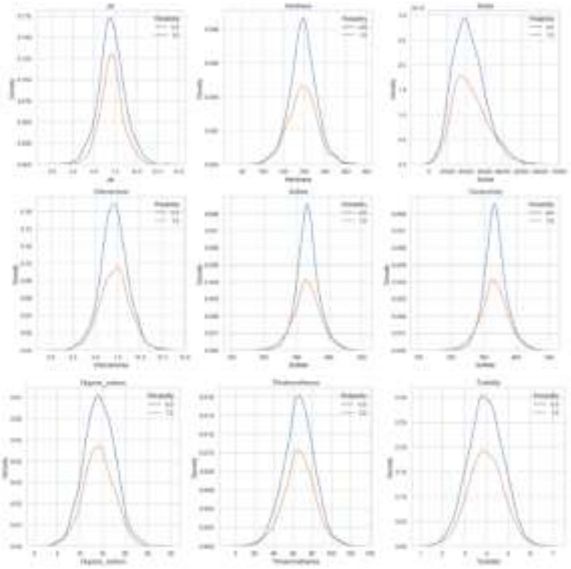


Figure 7

Water Quality Prediction

Figure 8

7. CONCLUSION:

Continuous water quality screening and assessment is appealing for future bright urban communities with the growth of IoT foundations, enormous information innovations, and AI approaches. This paper presents our most recent writing study, reviews related work, and speculates about water quality assessment momentum in light of extensive data analysis, AI models, and methods. Finally, it offers some opinions on potential issues, challenges, and requirements for future exams.

REFERENCES

[1] World Health Organization, "Meeting the MDG drinking water and sanitation target: the urban and rural challenge of the decade", Geneva, 2006.

[2] M. A. Tirabassi, "A statistically based mathematical water quality model for a non-estuarine river system1." JAWRA Journal of the American Water Resources Association, Vol. 7, pp. 1221-1237, December 1971.

[3] L. Hu, C. Zhang, C. Hu, and G. Jiang, "Use of grey system for assessment of drinking water quality: a case study of Jiaozuo city, China", Advances in Grey Systems Research, Springer Berlin Heidelberg, pp. 469-478, 2010.

[4] R. Rosly, M. Makhtar, M. K. Awang, M. N. A. Rahman, and M. M. Deris, "The Study on the Accuracy of Classifiers for Water Quality Application", International Journal of u- and e-Service, Science and Technology, Vol. 8, No. 3, pp.145-154, 2015.

[5] D. Yang, L. Zheng, W. Song, S. Chen, and Y. Zhang, "Evaluation indexes and methods for water quality in ocean dumping areas", Procedia Environmental Sciences: Proc. of the 7th International Conference on Waste Management and Technology, Vol. 16, pp.112- 117, December 2012.

[6] D. P. Loucks and E. V. Beek, "Water Quality Modelling And Prediction," Water Resources Systems Planning And Management: An Introduction To Methods, Models And Applications, Paris: UNESCO, pp. 381-425, 2005.

[7] Aspen-Nicholas Water Forum, "Data Intelligence for 21st Century Water Management: A Report from the 2015 Aspen-Nicholas Water Forum", 2015.

[8] S. P. Sherchan, P. G. Charles, and L. P. Ian, "Evaluation of realtime water quality sensors for the detection of intentional bacterial spore contamination of potable water." Journal of Biosensors & Bioelectronics 2013, 2013.

[9] Y. Liu, M. Islam, and J. Gao, "Quantification of shallow water quality parameters by means of remote sensing", Progress in Physical Geography, Vol. 27, No. 1, pp. 24-43, March 2003.

[10] M. Valdivia, D.W. Graham, and D. Werner, "Climatic, Geographic and Operational Determinants of Trihalomethanes (THMs) in Drinking Water Systems" Scientific reports, Vol. 6, 2016.

[11] Y. Zhong, L. Zhang, S. Xing, F. Li, and B. Wan, "The big data processing algorithm for water environment monitoring of the three Gorges reservoir area" Abstract and Applied Analysis, Vol. 2014, Hindawi Publishing Corporation, 2014.

[12] Hou Jing-Wei, MI Wen-Bao, and Long-Tang Li, "Spatial quality evaluation for drinking water based on GIS and ant colony clustering algorithm", Springer-Verlag Berlin Heidelberg, March 25, 2015.

[13] M Tarique, H. Khaleeq, and A. A. ElNour, "A Reliable Wireless System for Water Quality Monitoring", Vo. 8, No. 3, 2016.

[14] T. C. Lobato, R. A. Hauser-Davis, T. F. Oliveira, A. M. Silveira, H. A. N. Silva, M. R. M. Tavares, and A. C. F. Saraiva, "Construction 231 of a novel water quality index and quality indicator for reservoir water quality evaluation: A case study in the Amazon region", Journal of Hydrology, 522, pp. 674-683, 2015.

[15] A. Newton, and M. M. Stephen, "Lagoon-sea exchanges, nutrient dynamics and water quality management of the Ria Formosa (Portugal)" Estuarine, Coastal and Shelf Science, pp. 405-414, 2005.

[16] S. Y. Muhammad, M. Makhtar, A. Rozaimee, A. Abdul, and A. A. Jamal, "Classification Model for Water Quality using Machine Learning Techniques" International Journal of Software Engineering and Its Applications, pp. 45-52, 2015.

[17] A. Sarkar and P. Pandey, "River water quality modelling using artificial neural network technique" Aquatic Procedia, Vol. 4, pp. 1070-1077, 2015.

[18] Y. Khan and C. S. See, "Predicting and Analyzing Water Quality using Machine Learning: A Comprehensive Model," IEEE Long Island Systems, Applications and Technology Conference (LISAT), 2016.

[19] X. Li and J. Song, "A New ANN-Markov Chain Methodology for Water Quality Prediction," International Joint Conference on Neural Networks, pp. 12-17 July, 2015.

[20] L. Ma, K. Xin, and S. Liu, "Using Radial Basis Function Neural Networks to Calibrate Water Quality Model," World Academy of Science, Engineering and Technology International Journal of Environmental, Chemical, Ecological, Geological and Geophysical Engineering, Vol. 2, No. 2, 2008.

[21] C. McCormick, "Radial Basis Function Network (RBFN) Tutorial" [Online] Available: http://mccormickml.com/2013/08/15/radial-basis-function-networkrbfn-tutorial/, [Accessed: 10- Oct-2016], 2013.

[22] A. Solanki, H. Aggarwal, & K. Khare, "Predictive Analysis of Water Quality Parameters using Deep Learning", International Journal of Computer Applications, vol. 125, no. 9, pp. 0975-8887, Access from Google Scholar, Sept. 2015.

[23] Jaloree, Shailesh, A. Rajput, and Sanjeev Gour. "Decision tree approach to build a model for water quality." Binary Journal of Data Mining & Networking 4.1 (2014): 25-28.

[24] H. Liao and W. Sun. "Forecasting and evaluating water quality of Chao Lake based on an improved decision tree method." Procedia Environmental Sciences 2 (2010): 970-979.

[25] L. Yan-jun and M. Qian. "AP-LSSVM modeling for water quality prediction." Control Conference (CCC), 2012 31st Chinese. IEEE, 2012.

[26] X. Wang, G. Wang, and X. Zhang, "Prediction of Chlorophyll-a content using hybrid model of least squares support vector regression and radial basis function neural networks," Sixth International Conference on Information Science and Technology, pp. 366-371, May 6-8, 2016.

[27] W. Lin, L. Ran, T. Youcai, and L. Kefeng, "Research on Prediction of Water Quality of Water Reservoir with Combined Multiple Neural Networks Model," International Conference on Electric Technology and Civil Engineering (ICETCE), pp. 4376-4379, 2011.

[28] S. Song, X. Zheng, and F. Li, "Surface Water Quality Forecasting Based on ANN and GIS for the Chanzhi Reservoir, China," IEEE 2nd International Conference on information Science and engineering, pp. 4094-4097, 2010.

[29] G. A. C. Cordobaa, L. Tuhovcak, and M. Taus, "Using artificial neural network models to assess water quality in water distribution networks," 12th International Conference on Computing and Control for the Water Industry, pp. 399-408, 2014.

[30] L. Ying, Z. Jiti, W. Xiangrui, and Z. Xiaohui, "Water quality evaluation of nearshore area using artificial neural network model", 3rd International Conference on Bioinformatics and Biomedical Engineering, pp. 11-13, June 2009.

[31] N. Chang, & B. Vannah, (2015) "Comparative Data Fusion between Genetic Programming and Neural Network Models for Remote Sensing Images of Water Quality Monitoring", IEEE International Conference on Systems, Man, and Cybernetics, Manchester, pp. 1046-1051, 2013.

[32] M. R. Estuar et al., "Towards Building a Predictive Model for Remote River Quality Monitoring for Mining Sites", TENCON 2015 2015 IEEE Region 10 Conference, Macao, and pp. 22159-3442, Nov. 2015.

[33] M. Osborne et al., "A Machine Learning Approach to Pattern Detection and Prediction for Environmental Monitoring and Water Sustainability", Workshop on Machine Learning for Global Challenges (ICML2011), Bellevue, WA, 2011.

[34] Jerry Gao, Chunli Xie, and Chuanqi Tao, "Big Data Validation and Quality Assurance - Issues, Challenges, and Needs", IEEE Symposium on Service-Oriented System and Engineering, IEEE Computer Society, Oxford, UK, April, 2016.