International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 08 | August - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

MACHINE LEARNING PREDICTS FRESHWATER QUALITY MEASUREMENTS

K.Kirubananthavalli[1],
Assistant Professor /CSE,
Unnamalai Institute of Technology, Kovilpatti.
Tamil Nadu,India
kirubananthavalli@uitkovilpatti.ac.in

Mr.S. Prasanth[2],
Assistant Professor /CSE,
Unnamalai Institute of Technology, Kovilpatti,
Tamil Nadu,India
prasanthsathiyaraj@gmail.com

Abstract **- Multiple pollutants have lately emerged as a concern to the water supply. Therefore, modelling and anticipating water quality has become essential for the control of water pollution. In order to better predict the WQI and WQC, this study builds state-of-the-art artificial intelligence (AI) algorithms. Many individuals nowadays suffer from life-threatening diseases because of drinking contaminated water. Our research focuses on a water quality monitoring system because of the information it provides. We are preparing to use a machine learning system to determine water quality predictions. One of the most serious and disturbing problems society is now experiencing is the loss of natural water resources such as lakes, streams, and estuaries. Polluted water has far-reaching consequences that affect many individuals. As a result, optimising water quality requires careful management of available water resources. The impacts of water pollution may be successfully handled if data are assessed and water quality is predicted. While much has been written on the topic of water quality management, more work has to be done to improve the present methods' efficacy, reliability, accuracy, and usefulness. The purpose of this research is to create a model for predicting water quality using Artificial Neural Networks (ANNs) and time-series analysis. The 6-minute time period 2014 historical water quality data utilised in this investigation. The information was collected by the USGS and is hosted on their National Water Information System (NWIS) website.**

Keywords: water quality, machine learning, prediction and modelling, predictive algorithms, water resources.

INTRODUCTION

Water is the most valuable resource on Earth since it is essential to the existence of all life. All living things need clean water to function properly. Aquatic organisms have a tolerance range for pollution levels. When certain thresholds are surpassed, the continued existence of certain species is jeopardised. Ambient water bodies like rivers, lakes, and streams have quality criteria that are commensurate with their inherent value.

Not only that, but requirements that are unique to each use or application for water. For instance, the salinity of irrigation water shouldn't be too high, nor should it include any harmful chemicals that may be absorbed by plants or soil and subsequently cause ecological collapse. Water quality for industrial applications has to have a wide range of characteristics, all of which depend on the particular industrial activities being carried out.Natural water resources include ground water and surface water, two of the most easily available sources of potable water. However, human and industrial activities, as well as other natural processes, may lead to the contamination of such resources. The water quality is quickly deteriorating due to the fast growth of the industrial sector. Furthermore, the quality of drinking water is significantly impacted by facilities with inadequate sanitary standards and little public awareness.

In addition, the effects of drinking water contamination are severe, wreaking havoc on human health, the environment, and the economy. The United Nations (UN) estimates that 1.5 million people die annually from water-related ailments. Experts believe that 80% of health problems in poor nations are caused by unsafe water. Every year, we hear about 2.5 billion cases of disease and 5 million deaths. Predictions of WQ trends should take time into account so that the seasonal variation in WQ may be monitored. However, higher results may be achieved by utilising a number of models in tandem to forecast the WQ. In order to model and anticipate WQ, many approaches have been proposed. Statistical methods, graphical models, analytical algorithms, and predictive algorithms are all examples of such approaches. Several different water quality indicators were compared using multivariate statistical techniques to determine their interrelationships. In order to conduct analyses such as regression, multivariate interpolation, and transition probability, geostatistical methods were used.

A. PREDICTION

In the past, machine learning models did not explain their decision-making processes. Given this, it might be difficult to provide unbiased justifications for decisions and behaviours that are grounded on such ideas. By elucidating which features or feature variables most affect a model's predictions, explanatory models can eliminate the "black box" effect. When a model's findings are explained in a way that is

Prediction Explanations may be as crucial as the outcomes themselves by illuminating the factors that are most responsible for them. For institutions like banks that use models to determine whether or not to provide a loan, Prediction Explanations may shed light on the deliberations that led to a positive or negative decision. With this information, companies can create models that adhere to regulations, effectively convey model findings to stakeholders, and identify high-impact factors that will help them focus on their business objectives..

Prediction is a kind of inference used in statistics. Predictive inference is only one of several statistical inference procedures that may be used to make such a prediction. Extrapolating data from a population sample to the complete population and other related populations is one technique to explain statistics, albeit this isn't necessarily the same as making a forecast. The process of making predictions involves the transfer of information across time, usually over limited time frames. Cross-sectional data is often used for prediction, whereas time series methods are more commonplace in forecasting..

B. MACHINE LEARNING

Improved ML systems that train themselves. In the field of AI, it is considered a subfield. Without being explicitly educated to do so, AI computations build a model from example data, often known as "preparing information," to make judgements or expectations. Artificial intelligence (AI) calculations are used in a variety of applications, such as email sorting and computer vision, where developing normal calculations to execute the essential tasks would be very time-consuming and/or impossible. While certain branches of AI are heavily weighted towards computational insights and the use of computers for prediction, not all branches of AI are grounded on hard data. The study of numerical enhancement provides resources, theories, and application areas for the research of artificial intelligence. In a similar vein, the field of information mining emphasises unaided learning for exploratory assessment of information. Artificial intelligence is the study of how machines may learn to solve problems on their own. It entails instructing computers to do certain activities by using the existing body of information. No PC-side learning is needed for easy jobs, since it is straightforward to write calculations telling the machine how to conduct all operations necessary to handle the current problem. It might be challenging for a person to do the necessary calculations for increasingly complicated activities. Instead than waiting for human developers to discover each crucial step, the computer might potentially be more effective if allowed to build its own computation. When no perfect computation is available, the AI hierarchy employs a number of methods to direct computers to carry out tasks. When there are several possible solutions to a problem, replies, one method is to identify a subset of the right ones as meaningful. The resulting data may be used to further refine the algorithm on a computer.

II LITERATURE REVIEW

A. A COMPARATIVE STUDY OF HYBRID AUTOREGRESSIVE NEURAL NETWORKS

A recommendation was given by TugbaTaskaya-Temizel and colleagues. The project requires Researchers have shown that utilising many models to make predictions is more effective than using a single time series model. Combining a neural network with an autoregressive integrated moving average model (ARIMA) has recently been shown to improve prediction accuracy over using either model alone. This presumption, however, carries the risk of underestimating the relationship between the linear and non-linear components of the model, since certain forecasting approaches are assumed to be suitable for simulating particular aspects of the model, such as the residuals. In this study, we show that combining many estimates does not necessarily improve accuracy.We show, however, that the aggregate prediction may drastically underperform when compared to the performance of its individual elements. This is shown using nine datasets and two types of neural network models (autoregressive linear and time-delay). If cyclic patterns are not immediately useful, they may be removed using seasonal differencing with a constraint on the stochastic variance in the data. Seasonality and fashion trends may be eliminated with the use of pre-whitening procedures. If cyclical patterns are important, seasonal models may be used. Multiplicative seasonality in a time series may be converted to additive seasonality with the use of functional transformations like logarithms. Variations on the linear AR model may be used to analyse symmetric cycles. Cyclic patterns are periodic variations in a system.Seasonality is defined as a subset of calendar-based cycles. There is mounting statistical evidence that business cycles are asymmetric (Chatfield, 2004) that has an impact on the economy. The fact that the rate of change during a recession is distinct from the rate of change when an economy recovers from a recession helps to explain the economy's asymmetric cyclical behaviour. Some well-known data sets, such as the sunspot and Canadian lynx series (Rao &Sabr, 1984), exhibit asymmetric cycles that defy linear prediction. With the hybrid design, the mean and best fit of the TDNN, AR neural network hybrid, and AR single models all the nine datasets. However, the TDNN or linear AR model performs better than the hybrid in five of the data sets. There are three of these upgraded single variants that outperform the hybrid greatly. It seems that the model configuration is responsible for these improvements, and therefore optimising for generalisation performance is the way to go. [1]

B. FORECASTING TIME SERIES WITH HYBRID ARIMA AND ANN MODELS BASED ON DWT DECOMPOSITION

It was proposed by Ina Khandelwa and colleagues. The project requires Discrete wavelet transforms (DWT) have lately witnessed a significant uptick in their utilisation across a variety of scientific and technological fields. Here, we show that discrete wavelet transforms (DWTs) have the potential to enhance the precision of time series forecasting. This research proposes an alternative approach to forecasting by employing discrete wavelet transform (DWT) to separate a time series dataset into linear and nonlinear components. First, DWT is used to disentangle the time series training dataset into its linear (precise) and non-linear (approximate) parts. Both the Autoregressive Integrated Moving Average (ARIMA) and the Artificial Neural Network (ANN) models are used to differentiate between the reconstructed detailed and approximation components and to make predictions.The suggested strategy strategically employs the individual strengths of DWT, ARIMA, and ANN to boost forecasting precision. We put our hybrid method to the test on four real-world time series and compare its prediction skills to those of ARIMA, ANN, and Zhang's hybrid models. According to the results, the proposed method consistently provides the best predicting accuracy across all series.

Obtaining relatively accurate projections of a time series is crucial yet difficult. ANN and ARIMA are two popular and reliable forecasting models.Nonlinear time series are better suited to ANN, whereas linear time series are better served by ARIMA. Since most real-world time series include both linear and nonlinear correlation patterns, it is almost difficult to determine the precise type of a given series. In this study, we demonstrate a hybrid forecasting method that employs ARIMA and ANN independently to model linear and nonlinear components, following the usage of DWT to decompose the series into low and high frequency signals. The final combined predictions are the average of the forecasts derived using the harr, db2, and db4 wavelets. The suggested technique beats the ARIMA, ANN, and Zhang's hybrid model in terms of prediction accuracy, as shown by empirical findings using four real-world time series. [2]

C. EFFECTIVE STRUCTURAL HEALTH MONITORING SIGNAL RECOVERY BASED ON KRONECKER COMPRESSIVE SENSING

Sandeep Reddy Surakarta and others have proposed. In structural health monitoring (SHM), sensors check in on the structure at regular intervals and send the collected data to a central server. Due to the large volumes of data generated by monitoring sensors, data compression may be used to lessen the need for storage and make more efficient use of network bandwidth. bandwidth. Recently, a fast, linear, and effective data-sampling method known as "compressive sampling" (CS) has been introduced. Compression system complexity and recovery system quality both contribute to the overall length of the compressed signal. Traditional CS methods relied on an empirical method to determine the signal duration. We may reduce the computational complexity and compression time of the system by compressing the signal. However, the quality of the reconstructed signal declines if the signal is compressed too much. Kroenke technique in CS recovery was recently devised to compensate for the loss of accuracy. In this research, we investigate the feasibility of Kroenke-based CS recovery for seismic signals. The simulation results show that this recovery method has the potential to greatly improve quality while allowing sensors to drastically reduce data size. The Kroenke method of recovery allowed us to restore the original seismic signal with an accuracy of up to 7 dB. The MIT green building's vibration data was used in this investigation. The simulation findings show that vibration data may be compressed using the CS approach. Two distinct measurement matrices were investigated for use in the sensing process. Two scarifying bases were tried with each measurement matrix. Compression ratios of 75% and 50% were tried. The findings show that the basic deterministic DBBD matrix performs better than the Gaussian measurement matrix. The quality of the DBBD matrix is enhanced by using a DCT dictionary. To better the quality of the reconstruction, the Kroenke-based method was used. All simulations showed that CS could be implemented in very small scales, and that recovery quality could be greatly improved by adopting the Kroenke strategy. Particularly for sensors that perceive the construction's activity intermittently, sensing in a smaller size is beneficial in terms of power consumption and elapsed time for the sensing process. [3]

D. AN INTERDISCIPLINARY APPROACH TO DETERMINING HUMAN HEALTH RISKS AS A RESULT OF LONG-TERM EXPOSURE TO CONTAMINATED GROUNDWATER NEAR A CHEMICAL COMPLEX

It came highly recommended by people like Marina M. S. Cabral Pinto. The PTEs (potentially toxic elements) employed in this study have been linked to health problems in humans when consumed through polluted water supplies. Some of these PTEs, when exposed to over long periods of time, may cause cancer as well as other health problems. Since the 1950s, heavy industry has been present at the Estarreja Chemical Complex (ECC) in Northwest Portugal, resulting in severe soil and groundwater contamination. Groundwater has long been used by the local community for both domestic and agricultural purposes. Groundwater contamination in

Despite the existence of rehabilitation procedures over the last 20 years, the levels of numerous PTEs remain high, with concentrations several orders of magnitude larger than human ingestion.Taking into account dermal contact and ingestion of groundwater as exposure pathways, two groundwater sampling campaigns were conducted to demonstrate the temporal evolution of groundwater quality and to establish the non-cancer and cancer risks associated with PTE exposure for the population residing in the vicinity of the ECC. Human indicators for PTE exposure, including hair and urine samples collected during the second groundwater testing programme, were detected. Veiros, Bedudo, and Pardilhó have particularly high concentrations of As, making it the pollutant with the greatest dangers to both non-cancer and cancer health for the exposed population. People's hair and urine PTE levels were highest in areas with the most polluted groundwater. Al, As, Cd, Hg, Pb, Ni, and Zn were found at higher concentrations in urine samples taken from areas adjacent to the ECC, whereas As, Hg, and Ni were found in higher concentrations in hair samples. Urine and hair were shown to be accurate markers of both acute and chronic PTE exposure and to have a strong correlation with PTE levels in the groundwater. [4]

E.SIMULATION OF NON-POINT SOURCE (NPS) IN A TROPICAL COMPLEX CATCHMENT

Some have suggested, including J.H. Abdulkareem. In this endeavour, we will be The potential threat posed by non-point source (NPS) pollution to environmental water management has lately attracted widespread attention on a global scale. It's well knowledge that agriculture and urbanisation are two major causes of NPS pollution. It is common for NPS to be triggered by hydrologic changes in a watershed or by precipitation runoff, air deposition, seepage, drainage problems, or rainfall. Unlike pollution from factories and sewage treatment facilities, which come from the same places, NPS contamination comes from many different places. Pollutants known as nonpoint sources (NPS) are carried by runoff from rain and melting snow. Runoff water often deposits a wide variety of contaminants, both naturally occurring and anthropogenic, in aquatic environments. The project's goal was to model NPS contamination in the Kelantan river basin using data from GIS and other sources of pollution data. Pollutant loads of TSS, TP, TN, and AN were found on a variety of land types. Most of the four catchments with the highest pollution levels seem to have agricultural activities as their primary land use. The high concentrations of TP found in the watershed are likely linked to the high levels of soil erosion in the region, which frees phosphorus from the soil and washes it into nearby bodies of water. [5]

III. EXISTING SYSTEM

Both human health and global climate are directly impacted by water quality. Consumption, gardening, and manufacturing are just a few of the many uses for water. A key indicator for persuading water boards is the water quality report (WQI). Water quality (EC) is influenced by dissolved oxygen (DO), total coliforms (TC), organic oxygen interest (Body), nitrate, pH, and electric conductivity. Information pre-processing with min-max standardisation and missing information the executives using RF, highlight relationship, applied AI arrangement, and model component significance are the five phases by which these characteristics are controlled. The most precise (in terms of Kappa, Precision Lower, and Precision Upper) disclosures. Three different shorelines of eastern Lake Erie in New York, USA, were used to test out the model stacking method and compare results to the five individual base models.

After further investigation, the stacked model approach was shown to be superior to the baseline models. Yearly exactness medians of 78%, 81%, and 82.3% were recorded at the three analysed seashores, placing stacking model accuracy ratings consistently towards the top of the rankings for several years. An artificial intelligence (AI) based process was developed to recognise the water type of the Chao Phraya Stream by combining property acknowledgment (AR) and support vector machine (SVM) computations. Using its pinpoint accuracy, the AR identified the most important factors for improving the canal's infrastructure. In this case, the optimal information mixtures differ between computations, notwithstanding the poor performance of components with few connections. Cross breed calculations have increased the predictive power of a few performance models. An effective approach of group learning that uses extensive DT preparation for relapse, clustering, and other challenges. It uses majority vote to conduct characterisation and relapse on data and generates decision trees. Since arbitrary forests just have to keep track of relevant information, they can process data much quicker than choice trees.

IV. PROPOSED SYSTEM

SVM can classify water samples by quality in a water quality prediction task. SVM may be used to resolve relapse and order difficulties. It finds the hyperplane in high-dimensional space that maximises class margin or best fits data for regression. Water quality predictions are often made using SVMs and Decision Trees. SVM and DT have pros and cons.

Handle high-dimensional data and withstand noise, however kernel function selection is delicate and computationally expensive. Decision trees are easy to understand, implement, and prepare, but they overfit.

A. PRE-PROCESSING OF DATA

Prediction job performance depends on data pre-processing in the machine learning pipeline. SVM or Decision Tree water quality prediction data pre-processing includes the following:

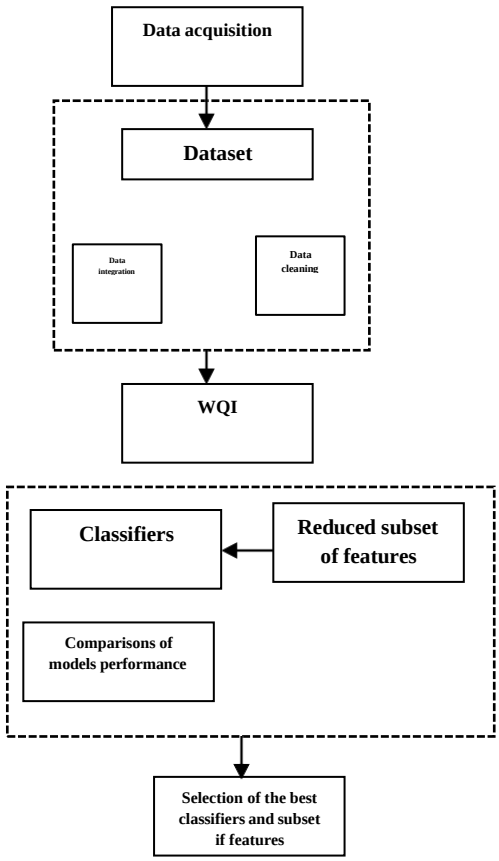

Fig:1 Data Processing in Machine Learning [steps & techniques]

Data Cleaning is examining the data for missing or inaccurate values and updating or eliminating them as needed. Data transformation entails turning data into a format appropriate for machine learning algorithms. For example, utilising one-hot encoding to convert categorical data to numerical data or normalizing numerical data to have a zero mean and unit variance. Data Reduction entails shrinking the quantity of the data, either by deleting unnecessary characteristics or by lowering the number of samples in the data. This step can help to shorten the calculation time of machine learning algorithms.

B. MODEL OF PREDICTION

As was noted in the prior answer, it is necessary to gather and preprocess data on water quality variables such as pH, temperature, dissolved oxygen, and so on. A Choice of Algorithms Choose an appropriate machine learning method, such as SVM or Decision Trees, based on the task's specific requirements. Training a Model: Train the chosen algorithm using the training data and the appropriate hyper parameters. Selecting the kernel function, regularisation parameter, and margin is what this entails in SVM. To do this using Decision Trees, one must determine both the maximum depth of the tree and the minimum split size for the sample.Model Verification: Adjust the model's hyper parameters as needed to have it performing at its best once it has been validated using validation data.

V. RESULT AND DISCUSSION

The presentation of a model to predict water quality is heavily dependent on the appraisal of outcomes. Performance Indicators: Select suitable presentation measures to conduct an analysis of the model. Some well-known metrics for characterisation problems include exactness, correctness, review, F1 score, and ROC curve bend among others. Common metrics for relapse problems include mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and R squared (R2). Disorganised Network: In grouping projects, a disorder framework may be used to display the distribution of positive, negative, false, and misleading projections.

VI. OUTPUT

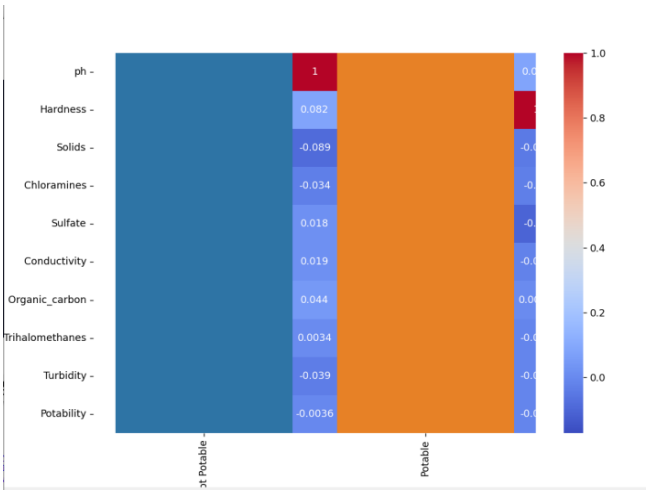

Fig:2 Potable & Non-potable graph based on ph[graph 1]

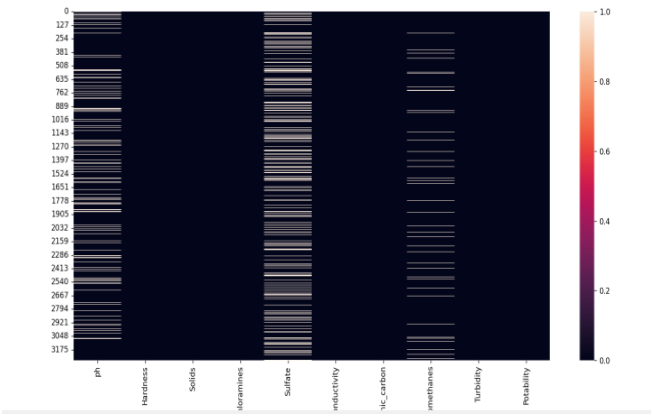

Fig:3 Potable & Non-potable graph based on ph[graph 2]

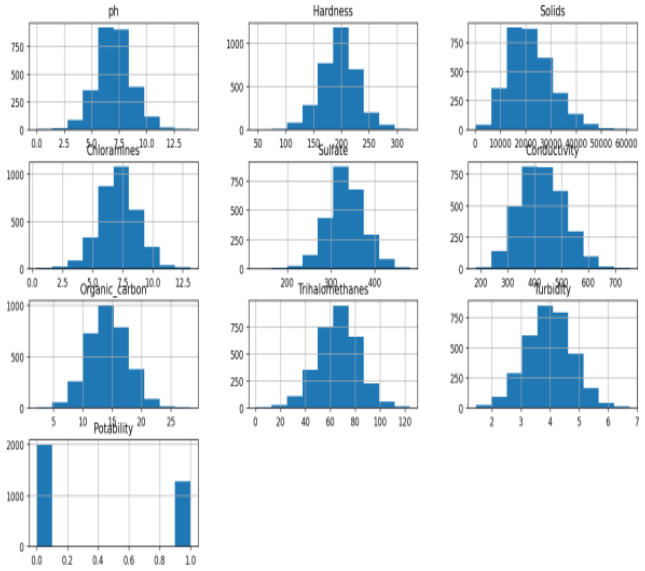

Fig:4 Individual box plot for each feature

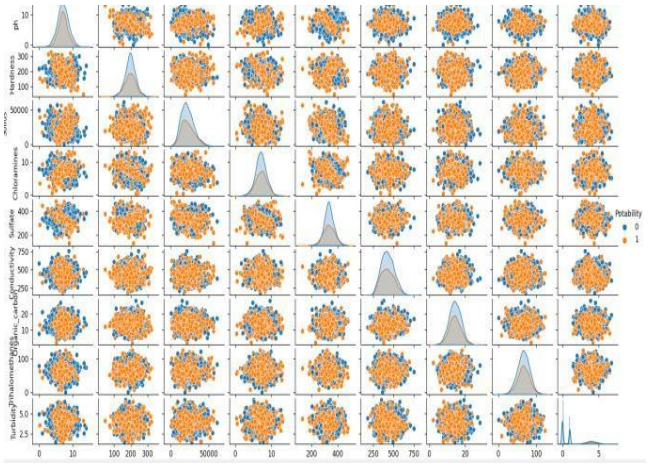

Fig: 5 Plot graph for different water quality parameters

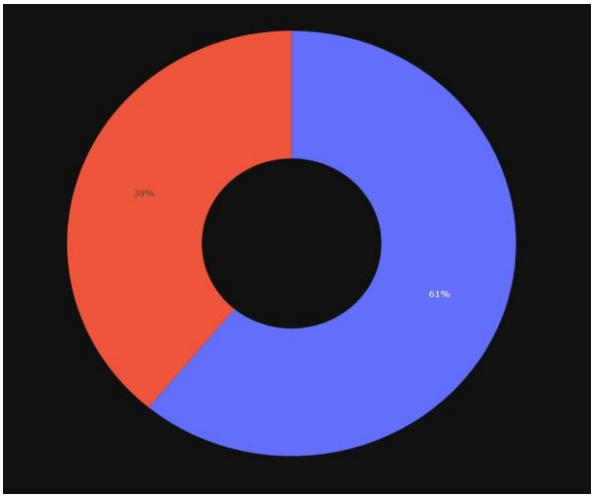

Fig:6 Potability Prediction for water

VII. CONCLUSION

In conclusion, forecasting water quality is an integral aspect of water asset management and planning. Numerous water quality forecasting models are available in DBSCAN. Which prediction model is used depends on the nature of the data and the objectives of the study. It is crucial to pre-process the water quality information and establish the bounds of the expectation model in order to produce accurate and reliable results. Analysing performance metrics, showing the findings, comparing with other models, fixing any defects, and repeating the process until the model can consistently anticipate water quality are all part of result analysis, which should be used to analyse the prediction model's outcomes. By accurately predicting water quality, decision-makers and resource-managers may make intelligent choices to protect and manage water resources, making sure that water is safe and accessible for everybody.

VIII. REFERENCES

[1]."A comparative analysis of autoregressive neural network hybrids," T. Taskaya-Temizel and M. C. Casey, Neural Networks, vol. 18, no. 5-6, pp. 781-789, 2005.

[2]. C. N. Babu and B. E. Reddy, "A hybrid ARIMA-ANN model based on a moving-average filter for predicting time series data," Applied Soft Computing, vol. 23, pp. 27-38, 2014.

[3]. M. M. S. Cabral Pinto, C. M. Ordens, M. T. Condesso de Melo, and colleagues, "An inter-disciplinary approach to assessing human health risks from long-term exposure to polluted groundwater near a chemical complex," Exposure and Health, vol. 12, no. 2, pp. 199-214, 2020.

[4]. "Human susceptibility to cognitive impairment and its association with environmental exposure to potentially harmful materials," M. M. S. Cabral Pinto, A. P. Marinho-Reis, A. Almeida, et al., Environmental Geochemistry and Health, vol. 40, no. 5, pp. 1767-1784, 2018.

[5]. Y. C. Lai, C. P. Yang, C. Y. Hsieh, C. Y. Wu, and C. M. Kao, "Evaluation of non-point source pollution and river water quality using a multimedia two-model system," Journal of Hydrology, 409, no. 3-4, pp. 583-595, 2011.

[6]. Z. M. Fadlullah, F. Tang, B. Mao, J. Liu, and N. Kato,"On intelligent traffic control for large-scale heterogeneous networks: a value matrix-based deep learning approach," IEEE Communications Letters, vol. 22, no. 12, pp. 2479–2482,2018.

[7]. R. Das Kangabam, S. D. Bhoominathan, S.Kanagaraj, and M. Govindaraju, "Development of a water quality index (WQI) for theLoktak Lake in India," Applied Water Science, vol. 7, no. 6, pp. 2907–2918, 2017.

[8]. A. A. Al-Othman, "Evaluation of the suitability of surface water from Riyadh Mainstream Saudi Arabia for a variety of uses," Arabian Journal of Chemistry, vol. 12, no. 8, pp. 2104– 2110, 2019.

[9]. T. H. H. Aldhyani, M. Alrasheedi, A. A. Alqarni, M. Y. Alzahrani, and A. M.Bamhdi, "Intelligent hybrid model to enhance time series models for predicting network traffic," IEEE Access, vol. 8, pp. 130431–130451, 2020.

[10]. M. M. S. Cabral Pinto, A. P. Marinho-Reis, A. Almeida et al., "Humanpredisposition to cognitive impairment and its relation with environmental exposure to potentially toxic elements," Environmental Geochemistry and Health, vol. 40, no.5, pp. 1767–1784, 2018.

[11]. J. Huang, N. Liu, M. Wang, and K. Yan, "Application WASP model on validation of reservoir-drinking water source protection areas delineation," in 2010 3rdInternational Conference on Biomedical Engineering and Informatics, pp. 3031–3035, Yantai, China, October 2010.

[12].P. Zeilhofer, "GIS applications for mapping and spatial modeling of urban-use water quality: a case study in District of Cuiabá, Mato Grosso, Brazil," Cad. Saúde…, vol. 23, no. 4, pp. 875–884, 2007.

[13]. UN water, "Clean water for a healthy world," Development, pp. 1–16, 2010.

[14]. 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.

[15]. Y. Park, K. H. Cho, J. Park, S. M. Cha, and J. H. Kim, "Development of early-warning protocol for predicting chlorophyll-a concentration using machine learning models in freshwater and estuarine reservoirs, Korea.," Sci. Total Environ., vol. 502, pp. 31–41, Jan. 2015.