International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 10 Issue: 03 | Mar 2023 www.irjet.net p-ISSN: 2395-0072

PREDICT THE QUALITY OF FRESHWATER USING MACHINE LEARNING

R.K. Gayathridevi[1], S. Akalya[2], V.Kowsalya[3]

[1]UG Scholar Dept of CSE Bannari Amman Institue of Technology Sathyamangalam
[2]UG Scholar Dept of IT Bannari Amman Institue of Technology Sathyamangalam
[3]UG Scholar Dept of IT Bannari Amman Institue of Technology Sathyamangalam

---

Abstract **- The purity of the water has recently been threatened by a number of contaminants. As a result, it is now crucial for the management of water pollution to model and anticipate water quality. This work develops cutting-edge artificial intelligence (AI) techniques to forecast the water quality index (WQI) and water quality categorization (WQC). Today, many people are afflicted with severe illnesses brought on by tainted water. We are examining a water quality monitoring system in our study since it gives information on water quality. We are about to identify forecasts for water quality using a machine learning system. The depletion of natural water resources including lakes, streams, and estuaries is one of the most significant and alarming issues facing humanity. The effects of dirty water are widespread and have an impact on several people. Water resource management is therefore essential for maximising water quality. If data are analysed and water quality is foreseen, the effects of water contamination can be effectively addressed. Even though this subject has been covered in a large number of earlier research, more has to be done to boost the effectiveness, dependability, accuracy, and utility of the current techniques to managing water quality. The goal of this study is to develop an Artificial Neural Network (ANN) and time-series analysis-based water quality prediction model. The historical water quality data used in this study has a 6-minute time period and is from the year 2014. The National Water Information System, a website operated by the United States Geological Survey (USGS), is where the data comes from (NWIS).**

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

I.INTRODUCTION

The most precious natural resource, water is required for the survival of most living things, including humans. Water of appropriate purity is necessary for all living things to live. Certain degrees of pollution are tolerable for aquatic animals. These species' existence is impacted and their lives are in danger when certain limits are exceeded. Rivers, lakes, and streams are only a few examples of ambient water bodies with quality standards that match their caliber. Also, there are standards specific to each application's or usage's water needs. For instance, irrigation water must not be too salinized or include dangerous substances that could be absorbed by plants or soil and destroy ecosystems. Depending on the specific industrial processes, water quality for industrial applications must have a variety of attributes.Some of the most accessible sources of fresh water, such ground and surface water, are found in natural water resources. Yet, such resources can get contaminated by human/industrial activity and other natural processes. As a result of rapid industrial expansion, the quality of the water is rapidly declining. Moreover, facilities with poor hygienic qualities and low public awareness have a significant impact on the quality of drinking water.

Furthermore, the repercussions of contaminated drinking water are highly detrimental, having a negative influence on infrastructure, the environment, and public health. An estimate from the United Nations (UN) states that 1.5 million people each year pass away from illnesses brought on by contaminated water. According to estimates, 80% of health issues in developing countries are related to contaminated water. 2.5 billion illnesses and five million fatalities are reported each year. It is suggested that the temporal component be included when predicting WQ trends to ensure that the seasonal shift of the WQ is tracked. On the other hand, utilising multiple models in combination to predict the WQ produces better results than using just one model. For WQ prediction and modelling, several strategies have been put forth. Examples of these methods include statistical techniques, visual modelling, analytical algorithms, and predictive algorithms. In order to ascertain the correlation and association between various indices of water quality, multivariate statistical methods were employed. Geostatistical techniques were used for regression analysis, multivariate interpolation, and transitional probability.

A. PREDICTION

Traditionally, machine learning models did not provide information on why or how they arrived at a result. This makes objectively explaining judgements and actions based on these theories challenging. Explanations for prediction avoid the "black box" phenomenon by clarifying which qualities, or feature variables, have the most influence on the results of a model. When the explanations for a model's results are

as essential as the results themselves, Prediction Explanations can reveal the aspects that most contribute to those results. Banks, for example, that employ models to assess whether or not to grant a loan can utilize Prediction Explanations to learn why an application was accepted or refused. With this knowledge, they can create models that are compliant with laws, readily communicate model results to stakeholders, and uncover high-impact aspects to help concentrate their business goals.

In statistics, prediction is a form of statistical inference. Although the prediction can be made using any of the numerous statistical inference methods, predictive inference is one method for such inference. In fact, one way to explain statistics is that it offers a method of extrapolating data from a population sample to the full population and other populations that are linked, which isn't always the same as prediction over time. Transferring knowledge over time, typically to specific time periods, is the act of forecasting. Prediction is commonly carried out using cross-sectional data, whereas forecasting typically requires the use of time series methodologies.

B. MACHINE LEARNING

Machine learning (ML) that gets better on its own. It is regarded as a division of artificial intelligence. Without being expressly trained to do so, AI calculations create a model based on example data, or "preparing information," to make judgements or expectations. When it is difficult or impossible to develop regular calculations to complete the necessary tasks, artificial intelligence (AI) calculations are utilised in a wide range of applications, such as email sorting and computer vision. Whilst a portion of AI is closely related to computational insights and concentrates on utilising computers to make predictions, not all AI is factual learning. The study of numerical improvement offers the field of artificial intelligence tools, theories, and application domains. Information mining is a similar area of study with a focus on unassisted learning for exploratory information evaluation. In artificial intelligence, computers learn how to do tasks without explicit programming. It comprises using available knowledge to teach PCs how to carry out specific tasks. For simple tasks assigned to PCs, it is possible to write calculations instructing the machine how to execute all procedures required to address the present issue; no learning is required on the PC's side. For more complex tasks, it is sometimes difficult for a human to physically do the requisite computations. In the long run, it may be more powerful to let the computer to construct its own computation rather than having human developers identify each essential step. The order of AI makes use of several approaches to guide PCs to execute tasks if no totally acceptable calculation is available. In circumstances when there are a large number of viable responses, one way is to label a fraction of the correct answers as significant. This might then be used to prepare information for the PC to enhance the algorithm.

II LITERATURE REVIEW

A. A COMPARATIVE STUDY OF HYBRID AUTOREGRESSIVE NEURAL NETWORKS

TugbaTaskaya-Temizel and associates made a suggestion. This project calls for Many researchers contend that combining several forecasting models produces more accurate results than using just one time series model. It has recently been demonstrated that a well-known technique called a hybrid architecture, which combines a neural network and an autoregressive integrated moving average model (ARIMA), can forecast events more accurately than either model alone. However, this assumption runs the danger of underestimating the connection between the model's linear and non-linear components by presuming that individual forecasting methodologies are appropriate, say, for modelling the residuals. In this work, we demonstrate that such combinations don't always perform better than individual estimates.We demonstrate, however, that when compared to the performance of its constituent parts, the aggregate forecast may significantly underperform. Autoregressive linear and time-delay neural network models, along with nine data sets, are used to demonstrate this. Seasonal differencing conditional on the stochastic variance in the data can be used to eliminate cyclic patterns if they are not immediately relevant. Pre-whitening techniques are utilised to get rid of seasonality and fashion trends. Seasonal models can be utilised if cyclic patterns are of relevance. However, a time series with multiplicative seasonality can be transformed into additive form using functional transformations such logarithms. Linear AR model variants may be utilised for symmetric cycles. Rhythmic oscillations are known as cyclic patterns.Seasonality is thought of as a subset of cycles with definite dates on the calendar. Economic statistics are showing more and more evidence that business cycles are not symmetric (Chatfield, 2004). Asymmetric cyclical behaviour in the economy is explained by the fact that the rate of change during a recession is different from the rate of change as an economy emerges from one. Well-known data sets like the sunspot and Canadian lynx series (Rao &Sabr, 1984) contain asymmetric cycles that are difficult to anticipate using linear methods. The mean and best fit of the TDNN, best fit of the AR neural network hybrid, and AR single models all increased relative to the hybrid architecture's mean. In four of the nine data sets, the mean hybrid performs better than the single model. Yet, in five of the data sets, the TDNN or linear AR model outperforms the hybrid. Three of these improved single models perform significantly better than the hybrid. These improvements seem to be related to model configuration, and choosing for generalisation performance yields better outcomes. [1]

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

Ina Khandelwa and associates suggested. This project calls for Several scientific and technical sectors have recently seen a sharp increase in the use of discrete wavelet transforms (DWT). In this study, we demonstrate how DWT might improve the accuracy of time series forecasting. This study suggests a novel method for forecasting that divides a time series dataset into linear and nonlinear components using DWT. The linear (detailed) and non-linear (approximate) components of the time series' in-sample training dataset are first separated using DWT. The reconstructed detailed and approximation components are then distinguished and forecast using the Autoregressive Integrated Moving Average (ARIMA) and Artificial Neural Network (ANN) models.In order to improve forecasting accuracy, the proposed method strategically makes use of the unique properties of DWT, ARIMA, and ANN. Four real-world time series are used to test our hybrid approach, and its predictive abilities are contrasted with those of ARIMA, ANN, and Zhang's hybrid models. The findings unequivocally show that for each series, the suggested technique yields the maximum forecasting accuracy.

It is important but challenging to obtain reasonably accurate forecasts of a time series. Two well-known and effective forecasting models are ANN and ARIMA.While ANN is more suited for nonlinearly produced time series, ARIMA is more effective for linearly produced time series. However, identifying a series' exact type is practically impossible, because time series from the actual world usually contain both linear and nonlinear correlation patterns. As a result, after utilising DWT to separate the series into low and high frequency signals, we show in this research a hybrid forecasting technique that uses ARIMA and ANN separately to simulate linear and nonlinear components. The averages of the forecasts obtained using the harr, db2, and db4 wavelets make up the final combined predictions. The empirical results utilising four real-world time series demonstrate that the proposed method outperforms ARIMA, ANN, and Zhang's hybrid model in terms of forecast accuracy. [2]

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

Sandeep Reddy Surakarta and others have proposed. Sensors periodically monitor the structure and transfer data to a remote server for further processing in structural health monitoring (SHM). Because of the massive amounts of sensor data produced by monitoring sensors, data compression may be utilized to minimize storage requirements and make better use of connection bandwidth. Compressive sampling (CS) was recently presented as an efficient, rapid, and linear technique of data sampling. The length of the compressed signal is proportional to the complexity of the compression system and the quality of the recovered system. The length of the signal was set empirically in classic CS techniques. If we compress the signal, the compression system will be more efficient in terms of computing complexity and compression time. On the other hand, if we shorten the signal too much, the quality of the reconstructed signal suffers. To compensate for the loss of precision, the Kroenke approach in CS recovery was recently established. We study the applicability ofKroenke-based CS recovery for seismic signals in this paper. The simulation results demonstrate that this recovery approach may significantly increase quality while sensors can compress the data to a small size. We were able to recover the original seismic signal with great precision up to 7 dB by using the Kroenke approach in recovery. This analysis took into account the vibration data from the MIT green building. The simulation results demonstrate that the CS compression method may be utilized to compress vibration data. For the sensing procedure, two types of measurement matrices were examined. With each of the measurement matrices, two different scarifying bases were tested. Two compression ratios were tested: 50% and 75%. The simple deterministic matrix DBBD outperforms the Gaussian measurement matrix, according to the results. DCT dictionary gives improved quality for the DBBD matrix. The Kroenke-based approach was employed to increase reconstruction quality. All simulations demonstrated that CS may be implemented in extremely tiny sizes and that the quality of recovery using the Kroenke approach can be much enhanced. Sensing in a smaller size is advantageous in terms of power consumption and elapsed time for the sensing process, particularly for constructing sensors that sense the construction's activity sporadically. [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 was recommended by Marina M. S. Cabral Pinto and others. PTEs (potentially toxic elements) used in this experiment are known to be harmful to human health when ingested through contaminated groundwater. Long-term exposure to some of these PTEs may result in both cancerous and non-cancerous health issues. The Estarreja Chemical Complex (ECC) in NW Portugal has experienced intense industrial activity since the early 1950s, which has led to high levels of soil and groundwater pollution. The local population has historically utilised groundwater for both human and agricultural needs. Groundwater pollution levels for

several PTEs remain high, with concentrations several orders of magnitude greater than human intake, despite rehabilitation methods having been in place for the last 20 years.Two groundwater sampling campaigns were conducted to show the temporal evolution of groundwater quality and to determine the non-cancer and cancer risks associated with PTE exposure for the population living around the ECC, taking into account dermal contact and ingestion of groundwater as exposure pathways. During the second groundwater sample campaign, PTEs were found in hair and urine, which were then used as biomarkers to confirm that the local people had been exposed to PTEs. According to the data, As is the pollutant that poses the most risks to both non-cancer and cancer health for the exposed population, with concentrations that are especially high in Veiros, Bedudo, and Pardilhó. Localities with the most contaminated groundwater also had residents with higher PTE levels in their hair and urine. Urine tests revealed heightened amounts of Al, As, Cd, Hg, Pb, Ni, and Zn in locations close to the ECC, while hair samples show higher levels of As, Hg, and Ni. It was discovered that urine and hair are reliable indicators of both short- and long-term PTE exposure and are closely related to groundwater PTE concentrations. [4]

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

J.H. Abdulkareem and others have proposed. This project involves Non-point source (NPS) contamination has recently received international attention as a possible concern in environmental water management. Agriculture and urbanization have long been identified as important contributors of NPS pollution. NPS is frequently induced by rainfall runoff, atmospheric deposition, seepage, drainage issues, or hydrologic changes in a watershed. NPS pollution originates from a variety of sources, as opposed to industrial and sewage treatment plants, which emanate from the same sources. Runoff from rainfall and melting snow transport impurities known as NPS pollutants from several sources. These contaminants range from natural to man-made pollutants and are often deposited in bodies of water via runoff water. The project aimed to simulate NPS pollutant loads in the Kelantan river basin by combining a geographic information system (GIS), databases, and pollution loads in the area. TSS, TP, TN, and AN pollutant loads were detected on various land uses. Agricultural activities appear to be the major land use in most of the four catchments, with the largest pollution burdens. Because phosphorus is not particularly mobile in soil, the large amount of TP discovered in the watershed is related to considerable soil erosion in the area, which releases phosphorus bound to the soil into water bodies. [5]

III.EXISTING SYSTEM

Water quality straightforwardly affects both human wellbeing and the climate. Water is utilized for some reasons, including drinking, horticulture, and industry. The water quality file (WQI) is a significant pointer for compelling water the board. Disintegrated oxygen (DO), complete coliform (TC), organic oxygen interest (Body), nitrate, pH, and electric conductivity are factors that impact water quality (EC). These attributes are managed in five stages: information pre-handling with min-max standardization and missing information the executives utilizing RF, highlight relationship, applied AI arrangement, and model component importance. The disclosures with the best exactness Kappa, Precision Lower, and Precision Upper. The model stacking approach was utilized on three separate sea shores encompassing eastern Lake Erie in New York, USA, and contrasted with every one of the five base models.

Following examination, the model stacking strategy beat the fundamental models in general. Stacking model precision evaluations were reliably at or close to the highest point of the rankings many years, with year-on-year exactness midpoints of 78%, 81%, and 82.3% at the three analyzed sea shores. To recognize the water nature of the Chao Phraya Stream, an AI based procedure consolidating property acknowledgment (AR) and backing vector machine (SVM) calculations was created. Utilizing the straight capability, the AR found the main components for further developing the waterway's quality. For this situation, the best info blends fluctuate between calculations, in spite of the fact that factors with low connections fared inadequately. A few performance models' forecast capacity has been expanded by the Cross breed calculations. A troupe learning methodology that works via preparing countless DT for relapse, grouping, and different troubles. It produces choice trees from information and performs characterization and relapse utilizing larger part vote. Arbitrary woodlands are faster than choice trees since they manage subsets of information.

IV.PROPOSED SYSTEM

SVM may be used to categories water samples into unmistakable gatherings in light of their quality criteria in a water quality prediction challenge. SVM is a directed learning strategy that might be utilized to settle relapse as well as order issues. It operates by locating the hyperplane in high-dimensional space that maximizes the margin between classes or, in the case of regression, best fits the data. Support Vector Machines (SVM) and Decision Trees are two common machine learning techniques for predicting water quality. Both SVM and Decision Trees have advantages and disadvantages. SVM is well-known for its capacity to

handle high-dimensional data and its resistance to noise, although it can be sensitive to kernel function selection and computationally costly. Decision trees are simple to read, simple to apply, and need little data preparation; nonetheless, they are prone to overfitting.

A. PRE-PROCESSING OF DATA

Data pre-processing is a key stage in the machine learning pipeline and is critical to the prediction task's performance. Here are some common processes in data pre-processing for SVM or Decision Trees water quality prediction:

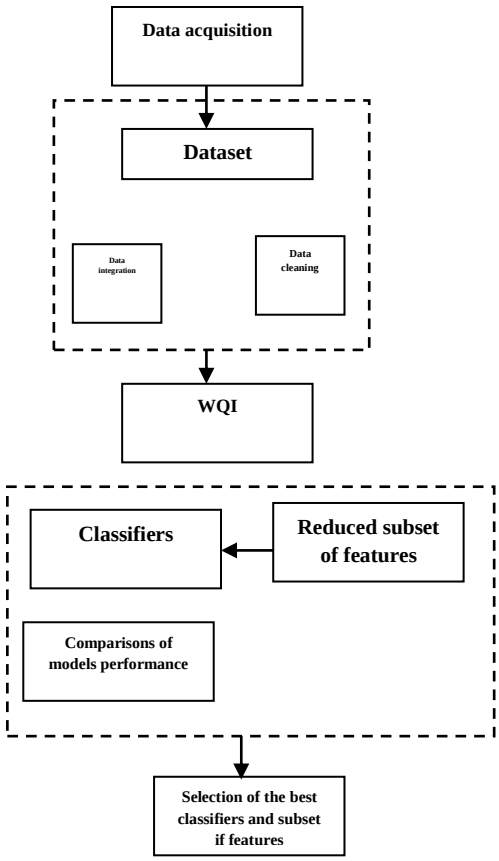

Fig:1 Data Processing in Machine Learning [steps & techniques]

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

Data collection and pre-processing: Collect and pre-process data on water quality factors, for example pH, temperature, dissolved oxygen, and so on, as mentioned in the preceding response. Select an Algorithm: Based on the unique needs of the task, select either SVM or Decision Trees, or any other acceptable machine learning technique. Model Training: Use proper hyper parameters to train the selected algorithm on the training data. In the case of SVM, this would entail picking the kernel function, the regularization parameter, and the margin. In the case of Decision Trees, this would entail deciding on the maximum tree depth and the smallest sample split size.Validation of the Model: Validate the model using the validation data and make any required adjustments to the hyper parameters to achieve the optimum performance.

V.RESULT AND DISCUSSION

The assessment of results is a critical stage in deciding the presentation of a water quality expectation model. Measurements of Execution: To examinations the model, select adequate presentation measures. Exactness, accuracy, review, F1 score, and beneficiary working trademark (ROC) bend are a few famous measures for characterization issues. Mean squared blunder (MSE), root mean squared mistake (RMSE), mean outright blunder (MAE), and R squared (R2) are standard measurements for relapse issues. Network of Disarray: A disarray framework can be utilized to show the dispersion of genuine positive, genuine negative, bogus positive, and misleading negative forecasts in grouping undertakings.

VI. OUTPUT

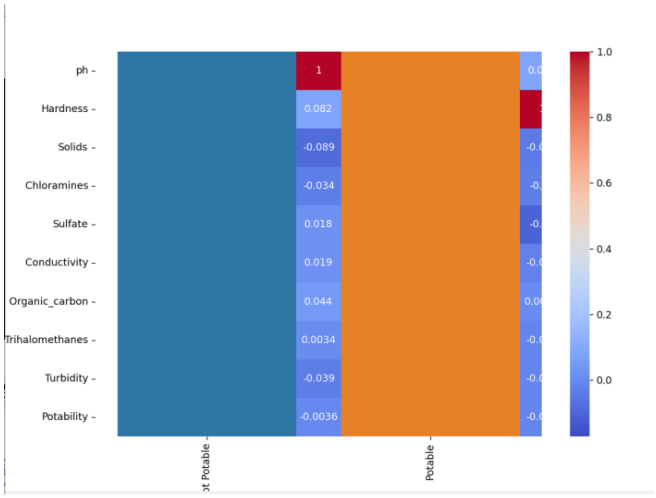

Fig:2 Potable & Non-potable graph based on ph[graph 1]

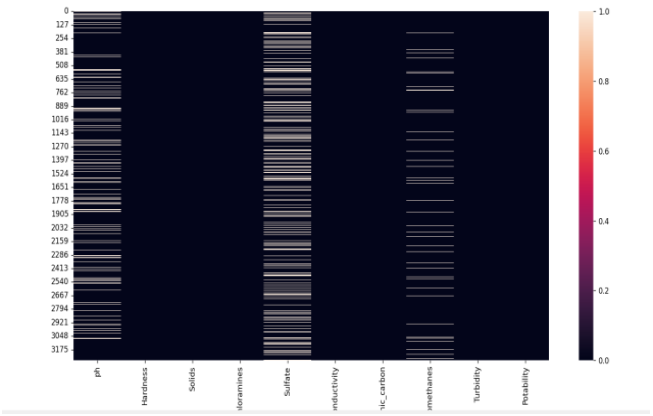

Fig:3 Potable & Non-potable graph based on ph[graph 2]

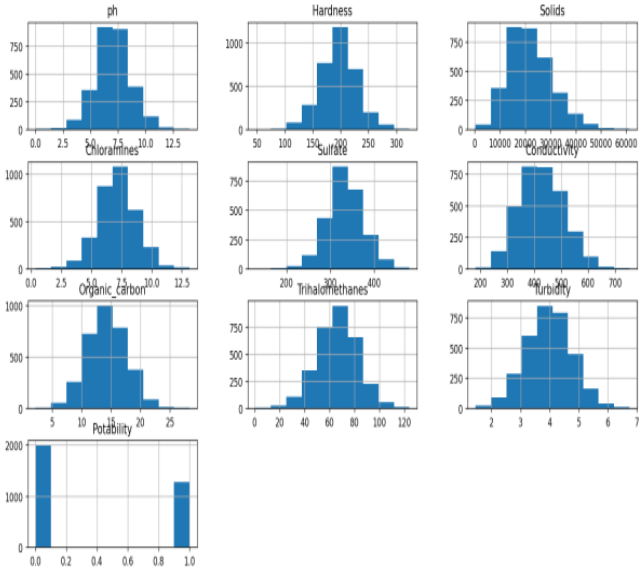

Fig:4 Individual box plot for each feature

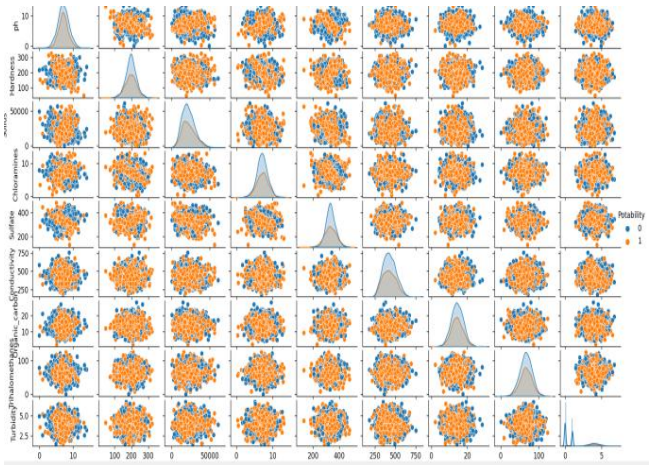

Fig: 5 Plot graph for different water quality parameters

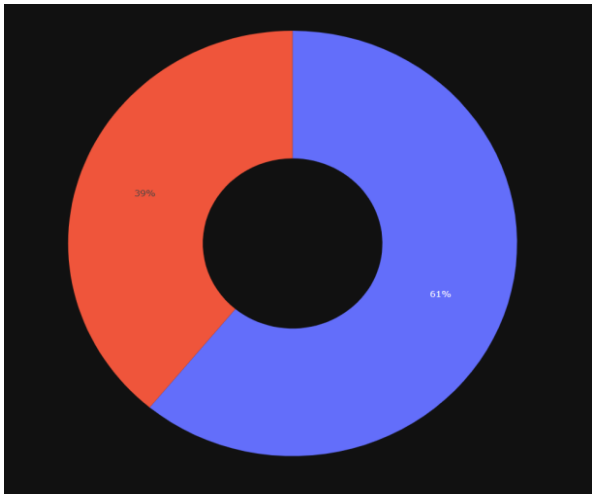

Fig:6 Potability Prediction for water

VII.CONCLUSION

To summarize, water quality forecast is a fundamental part of water asset the board and decision-making. DBSCAN has numerous prediction models that may be used to forecast water quality. The prediction model chosen will be determined by the specific data and the aims of the investigation. To achieve accurate and dependable findings, it is critical to carefully pre-process the water quality information and decide the boundaries of the expectation model. The prediction model's outcomes should be reviewed using result analysis, which entails analyzing performance metrics, displaying the results, comparing with other models, correcting any flaws, and repeating the process until the model can forecast water quality accurately and reliably. Decision-makers and resource managers may make educated decisions to safeguard and manage water resources, ensuring that water is safe and available for all, by properly forecasting water quality.

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.