iJETRM
International Journal of Engineering Technology Research & Management
www.ijetrm.com

PREDICTION OF WATER POLLUTION LEVEL IN CAUVERY RIVER USING DATA MINING TECHNIQUES

[1] Dr. K. SHYAMALA, [2] P. REKHA
[1] Associate Professor of Computer Science, Dr. Ambedkar Government Arts College, Chennai.
[2] Research Scholar, Department of Computer Science, Dr. Ambedkar Government Arts College, Chennai.

ABSTRACT

Water Pollution refers to release of harmful substance in the water that causes danger to humans and other living organism. Water Pollution is one of the major environmental problems in river water because of the growth of urbanization, industrialization, and growth of the population. The research aims to build an efficient model for river water quality also to classify and predict the river water quality index. The data has been collected from Central Pollution Control Board in various locations across the Cauvery River which flows through Tamil Nadu and Karnataka. Data mining techniques were used to discover models or patterns of data and it is much helpful in the process of decision making. The main intentionof this research is to analyses and predicts river water quality using classification algorithms in data mining. The proposed model has predicted the river waterquality for 2022 and 2023.

Keywords:
Water Pollution, Data Mining, Prediction, Water quality index, Cauvery River

INTRODUCTION

Whether it is utilized for drinking, household usage, food production, or leisure, safe and readily available water is critical for public health. Improving supplies of water, and also improved management of water resources, might help countries thrive and reduce poverty. In India, however, industrial and home pollutants have contaminated 70% of accessible water. Approximately 80% of the local populationand 20% of the urban population do not have access to clean drinking water [1].

The study of various strategies for predicting water quality in river has both theoretical and practical implications. Measurements of different parameters like Chemical oxygen Demand (COD), Dissolved Oxygen (DO), Electrical Conductivity (EC), Biochemical Oxygen Demand (BOD), Temperature, pH and other water quality components have been used to assess and analyse the quality ofriver water [2].

LITERATURE REVIEW

In the year 2013, Kamakshaiah. Kolli et. al., [3] proposed assessment of ground water quality using data mining techniques. The study area was Tadepalli in Guntur district. About 40 Water samples were collected from the bore wells and open wells in the study area. The physico-chemical parameters like TH, TDS, NO3and Cl, were analysed. The results were compared with standard techniques. Based on several water quality parameters, Water Quality Index was also calculated to geta single value that denotes the overall water quality at a certain area and time. This research on groundwater quality using F- concentration shows that the ground waters were in medium to very hard category and mostly blackish. The value of F- content was almost higher than the maximum permissible limit in 45 percent of total water samples.

Sarala C et. al., [4] proposed the investigation of ground water quality evaluation in the drag wells of Jawaharnagar, in the year 2012, at upper Musicatchment territory of Ranga Reddy locale in Andhra Pradesh. The examples from bore wells were gathered from the region for two season's pre rain storm and post storm in December 2007 and June 2008. The groundwater investigation was done by Arc GIS programming. The quality examination has been improved the situation the parameters like aggregate alkalinity, electrical conductivity, pH, TDS, calcium hardness, magnesium hardness, nitrites, sulfates, chlorides and fluorides. They have inferred that in the whole territory groundwater was dirtied .The investigation saw that the utilization of surface and groundwater for horticultural purposes, drinking and modern use has been expanded yet therefore it was seen that the water was contaminated and influenced soil supplements, the human wellbeing, biomass and condition in specific territories. In the majority of the areas the water was unsatisfactory for drinking reason.

IJETRM
International Journal of Engineering Technology Research & Management
www.ijetrm.com

Mohammad Zounemat - kermani et. al., [5] recommended OS-ELM method to predict the values of turbidity in water and also applied the major types of machine learning algorithms including ANN, MLP, CART and inductive neural network namely the Group Method of Data Handling (GMDH). The Multilayer perceptron in this model with a single hidden layer is used and each neurons sums its input signals affecting on it multiplying them by their mutual synaptic weights in the hidden layer. The another approach CART is able to identify the critical variables and does not require initial assumptions about the variables. The group method data handling is used for higher order polynomial input variables and solving the problems using regression analysis. The author collected the water samples from brandy wine creek Christina river in southern eastern Pennsylvania in the USA including the parameter pH, T and DO. He compared the algorithm to get better accuracy by applying the basic machine learning algorithms and he stated that water temperature and PH do not have effect on the variation of turbidity. The author proposed his algorithm called Online sequential ELM for predicting river quality and compared with several well-known models then he concluded that the proposed online sequential ELM approach achieved better predictions.

F. Karimipour et. al., [6] concentrated on water quality management using geospatial Data mining technique. He improved the previous technique using GIS system to improve the water quality in North west of Iran an industrial region where agriculture and animal husbandry are well established and there are huge amount of underground chemical resources too. Therefore food, chemical resources, and mineral industries and population are concentrated there. Since that region is rainy and many rivers are originated from there, proper management of water quality, he used the existing data to classify using ANN and he used the dependency analysis to predict the value of some attributes based on the value of other attribute. Dependency is used to find the correlation between different measures. The author collected samples related to spatial data in some of the rivers in the study area and also from industrial waste water to determine the contamination of water. The author calculated the parameter dependencies betweeneach two of them for first, second and third polynomial and exponential curve are calculated. The author used these parameters for goodness of fitness test. On the other hand it can be improved based on chi square distribution. [7]. He summarizedthat increase of TDS and decrease of DO values of water implies pollution. The author does not consider the parameters like BOD, alkalinity etc. In this paper the author used the static data with no temporal components. However, with the evolution of temporal GISs, recording spatio-temporal. Water resource data can provide better existing solution and predict the future trend [8].

METHODS AND MATERIALS

As mentioned in literature review Data mining is used to discover patterns fromlarge dataset helps in the process of decision making. In the proposed model, various classification algorithms were used to identify the potability and water quality Index of Cauvery river among those classification algorithm SVM had given more accuracy, and also the study focused to predict the water pollutionlevel for the period of 2022 – 2023.

CLASSIFICATION

Classification is a Data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. A classification task begins with a data set in which the class assignments are known. There are different types of classification and they are

- K- Nearest Neighbor (KNN)
- Support vector Machine (SVM)
- Decision Trees (DT)
- Random Forest (RF)
- Logistic Regression (LR)

Once the data has been loaded, it is necessary to find out the most suitable algorithm to produce understandable representation of data. Before running any classification algorithm, the test options need to be set [9].

- **Use Training Set**: Evaluation is based on how well it can predict the classof instances it was trained on.
- **Supplied Training set**: Evaluation is based on how well it can predict theclass of a set of instances

loaded from file.

- **Cross Validation**: Evaluation is based on cross - validation by using thenumber of folds entered in the "Folds" text field.
- **Split Percentage**: Evaluation is based on how well it can predict a certain percentage of data held for testing by using the values entered in the „%" field [10].

POTABILITY USING CLASSIFICATION ALGORITHMS

Potable water, also known as drinking water, comes from surface and ground sources and is treated to levels that meet state and federal standards for consumption. Water from natural sources is treated for microorganisms, bacteria, toxic chemicals, viruses and fecal matter. The chemical potability analysis can be performed when trying to identify a strange taste or taint in drinking water which can be caused by higher metal levels. It can also show the scaling or corrosive tendencies of the water [11].

The classifier model can be improved and predicted strength can be enhanced. It takes 80% of data for training and 20% of data for testing and a set of classifiers and combine the prediction of several classifiers. The steps involved as the dataset should be in AFF format or CSV format. In next step, the dataset will perform a preprocessing like data cleaning, data transformation and data reduction.

The training set is created in the Weka tool. First the goals of the research are identified suitable algorithms are selected for training Next choose the best classifier which is suitable for Water potability dataset which may gives good accuracy.

WATER QUALITY INDEX OF CAUVERY RIVER

WQI indicates the quality of water in terms of index number which represents overall quality of water for any intended use. It is defined as a ratingthat reflects the composite influence of different water quality parameters which were taken into consideration for the calculation of water Quality index (WQI) [12].

ISQA is calculated as:

$$\boxed{\mathbf{I_{SQA} = I_{TEMP} * (I_{BOD} + I_{TSS} + I_{DO} + I_{COND})}}$$

Where, I_{Temp} represents the Temperature I_{BOD} represent Bio Chemical Oxygen Demand, I_{TSS} represent Total Solid Substance I_{DO} represent Dissolved Oxygen and I_{COND} represent the condition of river water and also represent individual index terms with different weighting factors for each parameter [13].

PREDICTION OF WATER POLLUTION

Water quality Monitoring is an important role of the water quality management. Fresh water is a finite resource essential for use in agriculture, industry, propagation of wild life & fisheries and for human existence [14]. India isa riverine country. It has 14 major rivers, 44 medium river and 55 minor rivers beside numerous lakes, ponds and wells which is used as source of drinking water even without treatment.

Predicting the year wise pollution level of river water is extremely challenging due to the complex and dynamic change in nature and environmental factors. Rapid pollution and climate change are the key drivers causing serious water pollution around the globe. The water pollution in Indian river water have been calculated from 2016 to 2021 using Water pollutant parameters [15]. The research states that increase in BOD level in water can cause water pollution. By using the basic water quality parameters, the pollutants level of water from the year 2016 to 2023 have been calculated using machine learning algorithms [16].

The water quality level for Cauvery river water for next two years 2022, 2023 has been calculated using the proposed model. Thus it revealed that pollution level has been increased every year randomly, increase in Bio chemical componentsand Turbidity of water parameters can cause the water to become more pollutant.

RESULTS AND DISCUSSION

The potability of the Cauvery river has been classified using Weka Tool and Water Quality Index (WQI) of the river has been analyzed. The prediction of water quality is very vital in monitoring the pollution and in sustaining the availability of potable water resources. The water potability of Cauvery River, have been classified using basic classification algorithms and among those algorithms SVM gives more accuracy because SVM works relatively well when there is a clear margin separation between classes. The result of the algorithms is shown in the Figure 1

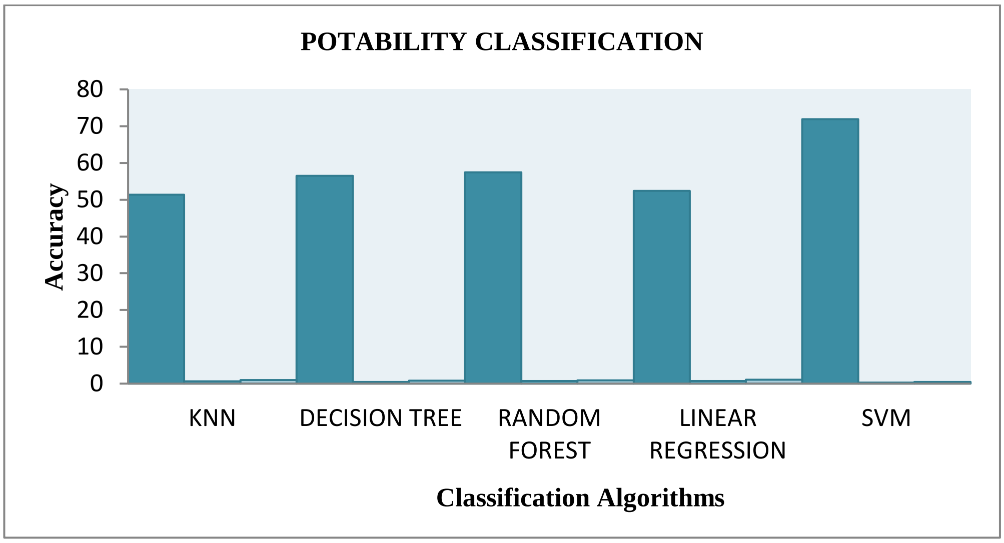

Figure 1: Potability of water using classification Algorithms

The Water Quality Index (WQI) of Cauvery river, have been classified using basic classification algorithms among those algorithms SVM gives more accuracy when the data are linearly and non-linearly separable. In spite of absence of a globally accepted composite index of water quality, some countries have used and are using aggregated water quality data in the development of water quality indices. Attempts have been made to review the WQI criteria for the appropriateness of drinking water sources. A WQI is a means by which water quality data is summarized for reporting to the public in a consistent manner. It is similar to UV index or an air quality index, and it tells us, in simple terms, WQI describes the quality of drinking water level.

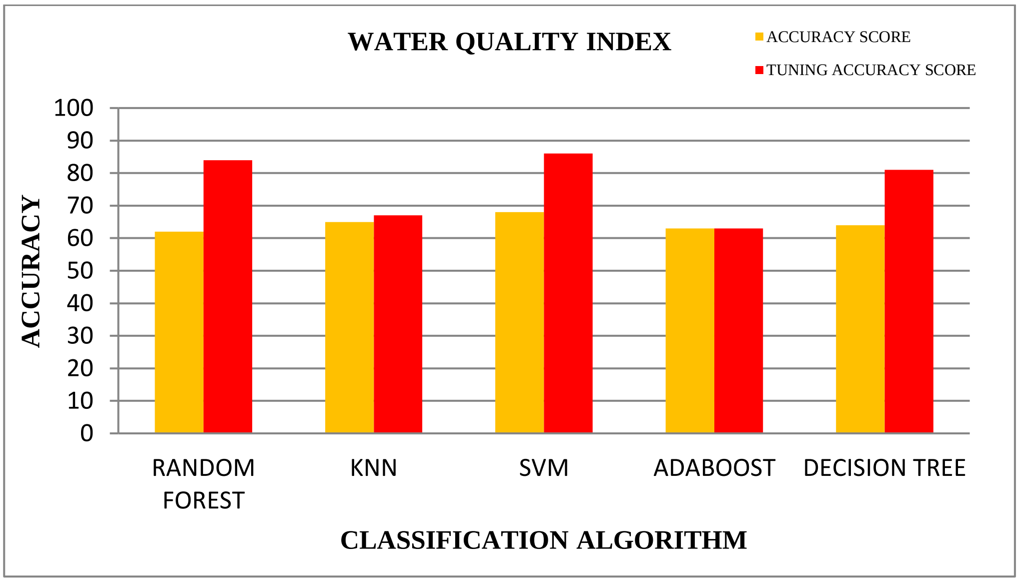

Figure 2: WQI using classification Algorithm

Predicting the year wise pollution level of river water is extremely challenging due to the complex and dynamic change in nature and environmental factors. Rapid pollution and climate change are the key drivers causing serious water pollution around the globe. The study states that increase in BOD level in water can cause water pollution. By using the basic water quality parameters the pollutants level of water from the year 2016 to 2023 have been calculated using machine learning algorithms.

The water pollution level has been given by Central Pollution Control Board (CPCB) from 2016 to 2021, with these data the predicted values are compared and the Mean Squared Error Value (MSEV) was calculated. Using the dataset from 2016 – 2021, the water pollution level was calculated for the years 2022 – 2023. Thus it revealed that pollution level has been increased every year randomly, increase in Bio chemical components and Turbidity of water parameters and causes the water to become more pollutant. Figure 3 to Figure 5 exhibited the water pollution from 2016 – 2021 and the prediction of water pollution for the 2022 and 2023. Figure 3 also illustrated the reduction in water Pollution level in the years 2022 and 2021 due to Covid – 19 lock down period.

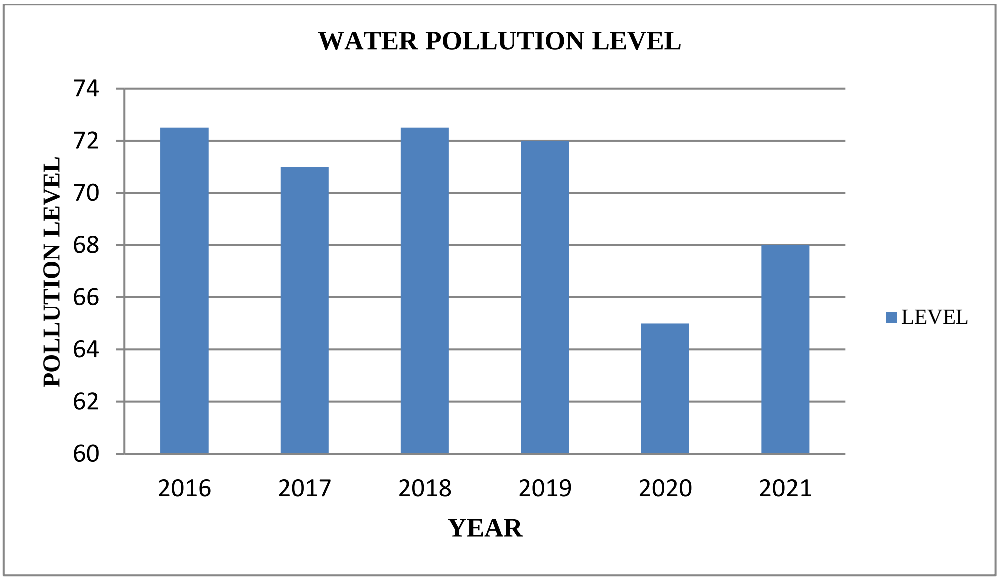

Figure 3: Water Pollution from 2016 - 2021

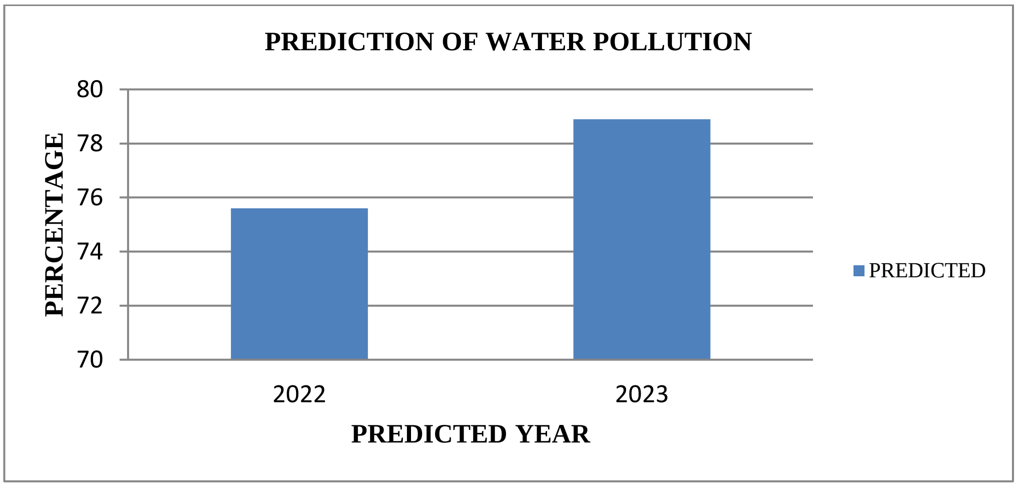

Figure 4: Prediction of water pollution for the year 2022 & 2023

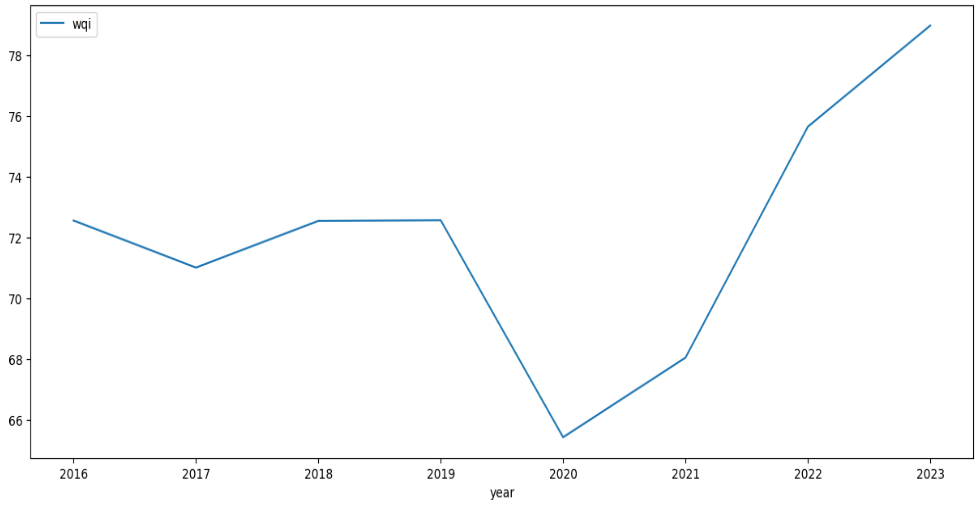

Figure 5: Chart showing Water pollution level

CONCLUSION

This research demonstrated a method for predicting and classifying the water quality using machine learning algorithms. The water metrics, including pH, DO, EC, Turbidity, Chloride, COD, TDS, and Alkalinity were used in this study. After analyzing the performance of multiple machine learning algorithms, Support Vector Machine seems to be more effective when compared with otherclassification algorithms. The potability of the Cauvery River has been classified using Weka Tool and WQI of the river has been analyzed. The prediction of water quality is very vital in monitoring the pollution and in sustaining the availability of potable water resources. Due to rapid urbanization and industrialization, it makes more pollutants in Cauvery River. It is found that the pollution rapidly decreases in the year 2020 and 2021 due to Covid – 19 Lockdown and gradually increases in the following year due to Lockdown relaxation. It mainly helps in reduction of water pollutant levels and it paves the way for water bodies. It also helps the government to take initiatives to reduce the pollutant level and to keep it under control. From the observation, it is observed that water pollution level of Cauvery River is increasing day by day.

FUTURE ENHANCEMENT

The future scope of this work is carrying out for the analysis by various other suitable methods and by using other data mining techniques. The data can be analyzed using different algorithms to get more accuracy. This model can be extended by implementing the proposed water quality prediction by extending it to other regions, so we can discover alternative ways can be discovered for a cleaner and greener environment.

REFERENCES

[1] Haitao Zhang, Xinmin Xie, Junsan Hou : "Water Pollution Control and urban safety water supply", IEEE conference, 10 Aug. 2011.

[2] WHO (World Health Organisation),(2005) Ecosystem and Human Well-being: Health systhesis. WHO Library Cataloguing-in-publication Data.

[3] Kamakshiah.kolli & r. Seshadri "Ground water Quality Assessment Using Data Mining "Techniques International Journal of computer application Volume 76, Issue 15, 39 – 45, 15th August 2013, 39 – 45

[4] F. Karimipour, M.R. Delavar and M. Kinaie " Water Quality Management of sing GIS Data Mining", journal of Environmental Informatics 5(2):61-72, 2005

[5] Das Gupta, M. Purohit, K.M, Jayita Datta "Assessment of Drinking waterquality of river Brahmani" journal of Environment and Pollution Vol.8,285-291, 2001

[6] Bilal Aslam, Ahsen Maqsoom, Ali Hassan Cheema, Fahim Ullah, Abdullah Alharbi Muhammad Imran "water quality management and hybrid data mining techniques"

[7] An Indexing Approach", Vol.10,119692 - 11970510th November 2022, 119692 – 119705 Mohammed Zonuemet Kermani, Meysam Alizamir, Marzieh Fadaee,S.Adarsh, Jala Shir "Online Sequential Extreme Learning Machine in River Water Quality Prediction": A comparative Study on Different Data Mining

iJETRM
International Journal of Engineering Technology Research & Management
www.ijetrm.com

approaches

[8] Delphla I, Florea M, Rodriguez MJ. 2018.” Drinking Water Source Monitoring Using Early Warning system based Data Mining Techniques”, Water resource Management, Vol.33, 129 – 140

[9] https://www.geeksforgeeks.org/basic-concept-classification-data-mining/

[10] https://www.ppsthane.com/drinking-water-testing-analysis

[11] https://worldpopulationreview.com/country-rankings/water-quality-by-country

[12] https://www.cdc.gov/healthywater/drinking/public/water_quality.html

[13] https://en.wikipedia.org/wiki/Support_vector_machine

[14] https://www.agry.purdue.edu/hydrology/projects/nexusswm/en/Tools/WaterQualityCalculator.php#:~:text=Sple%20Water%20Quality%20Index%20(ISQA,as %20shown%20to%20the%20left

[15] https://pubchem.ncbi.nlm.nih.gov/compound/Water#:~:text=Water%20(chemical%20formula%3A%20H2O),are connected%20by%20covalent%20bonds.

[16] https://www.thehindu.com/news/national/tamil-nadu/iit-madras-study-finds pharmaceutical-contaminants--cauvery-river-water/article36874088.ece