**Evaluating and Classifying Water Quality using Machine Learning Technique**

Anandha Lakshmi P,Gowthami D, Haritha R,Mrs.Jennifer,M.E

***Abstract-* Water pollution can be described as one of the most dangerous threats that the humanity ever faced. Great damage is caused to crops ,animals, forests etc. To prevent this problem we have to predict water quality from pollutants using machine learning techniques. Hence, water quality evaluation and prediction has become an important research area. The aim is to investigate machine learning based techniques for water quality forecasting by predicting results in best accuracy. Our analysis provides a comprehensive guide to sensitivity analysis of model parameters with regard to performance in prediction of water quality by accuracy calculation. To propose a machine learning-based method to accurately predict the Water Quality Index value by prediction results in the form of best accuracy from comparing supervised classification machine learning algorithms. Additionally, to compare and discuss the performance of various machine learning algorithms from the given water supply department dataset with evaluation classification report. Precision, Recall, and F1 Score are used to determine the success of the proposed machine learning algorithm technique**

*.****Keywords- Machine learning,Precision,Recall,F1-score***

1.Introduction

Water pollution is the contamination of water bodies that occur when pollutant are indirectly or directly discharged into water bodies without adequate treatment which affects the ecosystem and human life and is an issue nowadays. Therefore, it is very important to suggest new approaches to analyze and Predict the water quality (WQ).The aim is to Develop highly efficient machine learning models to predict the water quality index (WQI) based on machine learning algorithms. Applying some machine learning models, namely, Support Vector Machine (SVM), Logistic Regression, Decision Tree, Random Forest algorithms, for the prediction of water quality classification (WQC).

Literature survey

**1.Comparison of Water Quality Classification Models using Machine Learning.** In this paper, machine learning methods such as SVM, Decision Tree, and Nave Bayes are used to compare water quality classification models. pH, DO, BOD, and electrical conductivity are among factors examined when determining water quality. The water quality was determined using only four factors. Only three methods are compared, resulting in lower accuracy. As a result, the decision tree algorithm was discovered to be a better classification model, with a 98.50 percent accuracy.

**2. Water quality prediction and classification based on principal component regression and gradient boosting classifier**. **:** Gradient Boosting Regression, Multiple Linear Regression, and Support Vector Regression are three machine learning techniques used in this work to compare water quality classification models. The data set is separated into two parts: 75 percent for training and 25% for testing. WQI is calculated using nine parameters (PH, DO, TDS, Chloride, COD, EC, SS, Turbidity, Alkalinity). As a result, Support Vector Regression produces excellent results..

**3. Prediction of groundwater quality indices using Machine Learning algorithms.** The number of parameters considered in this study was 12. The data set is separated into two parts: 75 percent for training and 25% for testing. Two indices are used to calculate water quality. Entropy water quality index and Water quality index are the two. EWQI is efficient as a result of two indexes. Gradient Boosting Model, DNN, and XGBoost are the algorithms employed in this paper. Some user-defined parameter is required by this algorithm. Correlation coefficient, Root mean square error, and Index of agreement are examples of output parameters. As a result, DNN generates great accuracy.

**4. Research on water quality prediction method based on AE-LSTM.** A water quality parameter prediction technique based on automatic encoder (AE) spatial property reduction and long and short time memory (LSTM) neural networks is proposed. The Lang Fang Water Quality Automatic Observation Station data collection is used to assess the tactic's efficacy. The strategy is discovered to have higher forecast accuracy and toughness by estimating total phosphorus (TP) and total atomic number 7 (TN) concentrations. Results, The prediction model with AE-LSTM input had a higher prediction impact and accuracy than the prediction model with LSTM input, and it could effectively forecast water quality parameters.

**5. Water quality analysis in a lake using deep learning methodology.** Korattur Lake in Chennai was used to collect samples. This paper's parameter was 9. (PH ,TDS ,COD ,Nitrate ,Iron ,Sodium ,Phosphate ,Turbidity ,chloride ). Artificial Neural Networks, Recurrent Neural Networks, and Long Short Term Memory are the algorithms employed. As a result, LSTM has the highest accuracy, around 94%.

**6.** **Ground Water Quality Prediction using Machine Learning Algorithms.** ANN (Artificial Neural Network) is used. This method eliminates the use of chemicals in the evaluation of water quality indicators and is also cost effective. This paper presents a transient methodology for predicting unknown parameters such as pH, chloride, and sulphate values using well-known parameters such as pH scale, electrical physical phenomena, and TDS, as well as the Levenberg-Marquardt algorithmic programme, which aids in the classification of water bodies for various applications. In forecasting chloride, total-hardness, sulphate, and total alkalinity, the accuracy was 83.94 percent, 87.9%, 81.736 percent, and 79.48 percent, respectively.

**7.** **Predictive Analysis of Water Quality Parameters using Deep Learning . :** It examines the physico-chemical properties of ground water quality in the Vellore district towns of Ranipet, Arcot, and Walljah Pet. For the objective of investigating the quality of groundwater, water samples were obtained from various designated bore wells. In the Vellore district, there are two major types of water contamination: high and low. Water quality metrics like PH, TDS, EC, Chloride, Sulphate, Nitrate, Carbonate, Bicarbonate, metal ions, and trace elements have all been calculated. Water quality prediction algorithms employing Machine Learning classifier algorithm C5.0, Nave Bayes, and Random Forest as leaner. With accuracy and classification error, Nave Bayes and Random Forest generated better results.

**8.Predicting Water Quality Parameters Using Machine Learning.** When compared to other techniques, deep learning algorithms that use unsupervised learning provide proper results. Stack denoising autoencoder, deep belief network, regression, and multilayer perception are examples of algorithms. The results show that toughness can be achieved using a denoising autoencoder and a deep belief network, as well as successfully managing information variability. The benefit of unsupervised learning algorithms is assessed using metrics such as mean absolute error and mean square error to examine the prediction error rate. Area unit pH scale, dissolved oxygen, and turbidity were employed as parameters. Modules include data collecting (krishna stream), data preparation, and modelling.

*III Proposed methodology:*

The proposed method is to build a machine learning model for water quality. Data collection is done where the past data of water qualities are collected. Data mining is a technique used for processing enormous data in the domain. The water if found before proper treatment can save lives . Machine learning concept is applied in this system which reduces the manual effort to make a better model which is error less. Data analysis and data visualization are being done on the dataset . Then machine learning algorithms are applied on the dataset. After applying the algorithms each one produces different accuracy. Finally the algorithm with highest accuracy is selected and deployed.

**Advantages:**

* Machine learning concept is used.
* Various algorithms are used.
* Highest accuracy is attained

Preparing the Dataset:

aluminium - dangerous if greater than 2.8

ammonia - dangerous if greater than 32.5

arsenic - dangerous if greater than 0.01

barium - dangerous if greater than 2

cadmium - dangerous if greater than 0.005

chloramine - dangerous if greater than 4

chromium - dangerous if greater than 0.1

copper - dangerous if greater than 1.3

flouride - dangerous if greater than 1.5

bacteria - dangerous if greater than 0

viruses - dangerous if greater than 0

lead - dangerous if greater than 0.015

nitrates - dangerous if greater than 10

nitrites - dangerous if greater than 1

mercury - dangerous if greater than 0.002

perchlorate - dangerous if greater than 56

radium - dangerous if greater than 5

selenium - dangerous if greater than 0.5

silver - dangerous if greater than 0.1

uranium - dangerous if greater than 0.3 is\_safe

class attribute.

IV System Architecture

Water pollution is one of the greatest fears of the world. So as to provide a safe drinking water there is a need to predict the quality of water in real time. But, since the real time water monitoring faces challenges due to growing pollution limited water resources etc. There is a need to develop new and better technologies to predict the water quality parameters.

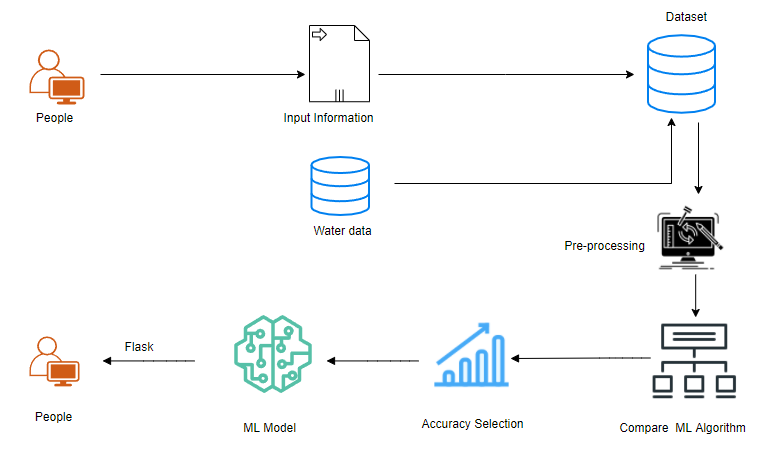
¬ Defining the problem

¬ Preparing data

¬ Evaluating algorithms

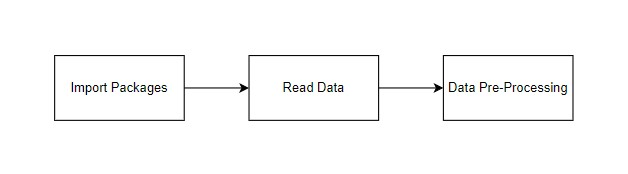
¬ Improving results

¬ Predicting results



Modules:

1. Data preprocessing:

 Validation techniques in ML is used to get the error rate which is close to true error rate pf the dataset. If the volume of the data is too large the validation techniques are not needed. This is concerned with locating the missing or duplicate value.Datatype is found-that is whether integer or float. Fine tuning of model hyper parameters are finetuned by ML engineers. Data gathering, data analysis, and the process of dealing with data content, quality, and organisation can be time-consuming. Understanding data and its properties is done during the data identification step, which aids in the selection of an algorithm to build the model.

A number data is cleaned by pythons pandas library, It focuses specifically on biggest data cleaning task , missing values, it spends less time cleaning the data and more time exploring and modelling.

Understanding of data and its properties are done during the process of data identification which helps to choose the algorithm to build the model.

A number data is cleaned by pythons pandas library, It focuses specifically on biggest data cleaning task , missing values, it spends less time cleaning the data and more time exploring and modelling.

GIVEN INPUT EXPECTED OUTPUT

input : data

output : removing noisy data

1. Data Validation/ Cleaning/Preparing Process

The specified dataset is being loaded. The library packages are being imported. Variables are identified by data shape and type, and missing values and duplicate values are evaluated. A validation dataset is a sample of data kept back from training your model that is used to measure model skill when tweaking models and techniques for making the most of validation and test datasets while evaluating your models. Drop the column and rename the dataset. The steps for cleaning data vary depending on the dataset. The main purpose is to find and fix problems so that data may be used for better decision-making and analytics

1. Exploration data analysis and data visualisation

It is an important skill in machine learning and applied statistics. Statistics focuses on estimating data and quantitative descriptions. Data visualisation provides qualitative comprehension tools that are useful for comprehending datasets, exploring datasets, and spotting outliers, patterns, and corrupt data, among other things. it is used to express the key relationship in charts plots .

Expressing the data in visual form makes it more understandable. Visualising data samples is an important skill in both applied ML and applied statistics. It will discover the different plots that will be known while visualising data in python and how to use them for better understanding of data.

* How to use line plots to visualise time series data and bar charts to visualise categorical variables.

• How to summarise data distributions with histograms and box plots

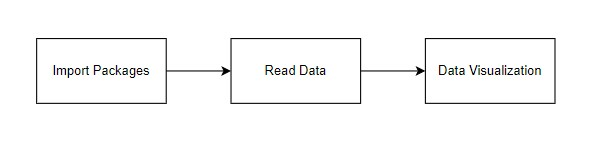
Pre-processing refers to the adjustments made to our data before it is fed into the algorithm.. Data pre-processing is used to clean the raw dataset . In other words, In other words, anytime data is acquired from various sources, it is obtained in raw format, which makes analysis impossible.. The data has to be in better manner for achieving better results. Some Machine Learning models require data in a specific format; for example, the Random Forest(RF) algorithm does not tolerate null values. To implement RF algorithm null values have to be eliminated from the original raw dataset. Dataset should be formatted in such a way that more than one machine learning and deep learning algorithms are executed using the same dataset.

**False Positives (FP)** occur when the actual class is not the same as the projected class.

**False Negatives (FN)** are situations in which the real class is yes but the expected class is no.

**True Positives (TP**) are accurately predicted positive values, indicating that the value of the actual class and the value of the projected class are both yes.

**True Negatives (TN):** These are correctly predicted negative values, indicating that the actual class value is no and the predicted class value is also no.



GIVEN INPUT EXPECTED OUTPUT

input : data

output : visualized data

D.Comparing Algorithm with prediction in the form of best accuracy result

To predict a value, the logistic regression process uses a linear equation with independent predictors. The anticipated value ranges from negative infinity to positive infinity. Logistic regression is a high-accuracy model that is created by comparing the best accuracy

True Positive Rate(TPR) = TP / (TP + FN)

False Positive rate(FPR) = FP / (FP + TN)

**Accuracy**: Accuracy is the percentage of right predictions out of a total number of forecasts.

Accuracy = (TP + TN) / (TP + TN + FP + FN)

It's only the proportion of accurately anticipated observations to total observations. Only when the dataset is symmetric and the values of false negative and false positive are almost equal is accuracy a good measure.

**Precision:** Precision refers to the percentage of positive predictions that are correct.

Precision = TP / (TP + FP)

The ratio of accurately anticipated positive observations to total predicted positive observations is known as precision. The low false positive rate is related to high precision. We have a precision of 0.788, which is rather good.

**Recall:** Recall that the proportion of positive observed values that was correctly predicted is the proportion of positive observed values that was correctly predicted.

Recall = TP / (TP + FN)

Recall(Sensitivity) Yes, recall (Sensitivity) is the ratio of accurately predicted positive observations to all observations in the actual class.

**F1 Score** The weighted average of Precision and Recall is the F1 Score. As a result, this score considers both false positives and false negatives. When the class distribution is unequal, F1 is frequently more valuable than accuracy. When false positives and false negatives have equivalent costs, accuracy works well. If the costs of a false negative and a false positive are significantly different, it's wiser to consider both Precision and Recall.

**General Formula:**

F- Measure = 2TP / (2TP + FP + FN)

**F1-Score Formula:**

F1 Score = 2\*(Recall \* Precision) / (Recall + Precision)

Module diagram:



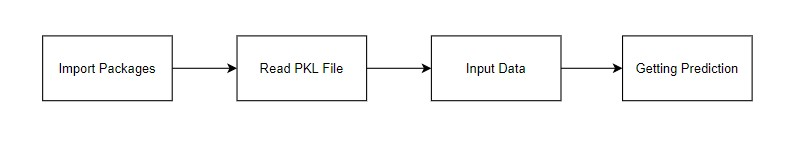
input : data

output : getting accuracy

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Method | Precision | Recall | F1 Score | Accuracy |
| Logistic  regression | 98 | 98 | 95 | 90 |
| Decision tree | 98 | 97 | 97 | 95 |
| Random Forest | 96 | 99 | 97 | 95 |
| Support vector  Machine | 89 | 100 | 94 | 88 |

E.Flask:

Here the accuracy of Decision Tree and Random Forest is same and has the Highest accuracy and Decision Tree is deployed using Flask Framework in a website.



input : data values

output : predicting output

**V RESULT**

The proposed model was used to predict the water quality. For this purpose the dataset mentioned earlier was used to evaluate the model performance to find out the best one, comparisons with RF,LR,SVC,DT were exhibited. The dataset was split into 70% and 30%for training and testing purposes respectively. For the results performance of 4 algorithms were compared in which Decision Tree has attained the highest accuracy of 95% when compared to 3 other algorithms. Finally the system is deployed as a web application. This system predicts the water quality status as safe or unsafe for drinking.

**VI Conclusion**

The analytical process started from data cleaning and processing , finding the missing values, exploratory analysis of data ,applying different algorithms and finally model building and evaluation and finding the highest accuracy. The best accuracy on public test set is the highest accuracy score which will be found out. This application is helpful in predicting Water Quality status.

**References:**

[1]Neha Radhakrishnan; Anju S Pillai,Year 2020, “Comparison of Water Quality Classification Models

using Machine Learning”, Fifth International Conference on Communication and Electronics Systems

,page 1183-1188

[2]Md.Saikat islam khan, Year 2021,”Water quality prediction and classification based on principal

component regression” , Journal of King Saud University-Computer and Information Science, page 1-9

[3]Hemant Raheja,Arun Goel, Mahesh Pal, Year 2021,”Prediction of groundwater quality indices using

Machine Learning algorithms”, International journal of machine learning and computing,Volume17 no ,

page 337-351

[4]Huiqing Zhang, Kemei Jin,Year 2021,”Research on water quality prediction method based on AE-

LSTM”, 5th International Conference on Automation, Control and Robotics Engineering, page 602-606

[5]P.Senthil kumar, Prasannamedha G, Soumya k,Year 2020, “Water quality analysis in a lake using deep

learning methodology”, International journal of environmental analytical chemistry, page 1-16

[6]Nikhil M Ragi; Ravishankar Holla; G Manju ,Yaer 2020,”Predicting Water Quality Parameters Using

Machine Learning”, 4th International Conference on Recent Trends on Electronics, Information,

Communication Technology (RTEICT)

[7]Theyazn H. H Aldhyani, Mohammed Al-Yaari, Hasan Alkahtani and Mashael Maashi ,Year

2020,”Water Quality Prediction Using Artificial Intelligence Algorithm”, International journal of applied

bionics and biomechanics ,page 1-12

[8] S.Vijay &amp; Dr.K.Kamaraj,Year 2019, “Ground Water Quality Prediction using Machine Learning

Algorithms”, International Journal of Research and Analytical Reviews, Volume 6 no 1,page 743x-749x

[9] Archana Solanki, Himanshu Agrawal, Kanchan Khare, Year 2019 ,”Predictive Analysis of Water

Quality Parameters using Deep Learning”, International journal of computer application, Volume 125 no 9,

page 29-34

[10] Mr.M.Anbuchezhian, Dr.R.Venkataraman, Mrs.V.Kumuthavalli,Year 2018,”Water Quality Analysis

and Prediction using Machine Learning Algorithms”, International journal of emerging technologies and

innovative Research, Volume 5 no 11,page 455-462