Machine Learning Techniques for Estimating and Prediction of Water Quality – An Analytical Study

[1]PrittoPaul P, [2]Usha M
[1,2]Associate Professor, Velammal Engineering College, Department of CSE, India

ABSTRACT: The current study is concerned with the measurement of physico-chemical characteristics of water samples from various sampling sites, including temperature, pH, EC, hardness, chlorides, alkalinity, phosphate, and sulphate. An increase in water quality concentration indicates an increase in the amount of sewage, industrial effluents, anthropogenic activities, and garbage discharged into rivers that affect water quality. Since water is used for a variety of things, its appropriateness must be evaluated before usage. Also, it is important to routinely check the health of water sources to see if they are safe to use. The goal is to research machine learning-based methods for most accurate water quality forecast findings. supervised machine learning approach used to analyse the dataset. Many pieces of information are captured throughout the analysis of the dataset using the Supervised Machine Learning approach (SMLT), including the identification of variables and results from univariate, bivariate, and multivariate analyses. Using evaluation methodologies compare and contrast the results of several machine learning algorithms from the provided dataset.

KEYWORDS: Predictive analytics algorithms, Principal Component Analysis (PCA), Water Quality Index (WQI), Water Quality Classification (WQC).

.

I. INTRODUCTION

Water quality directly affects public health and environment. It plays a crucial role in activities like drinking, agriculture, and industry.[11][12].Based on historical data, machine learning model predictions enable businesses to make extremely precise assumptions about the most likely outcomes of a question. These assumptions can be made about a variety of topics, including the likelihood of customer churn, potential fraudulent activity, and more. Water quality has been harmed during the past few years by a number of pollutants. As a result, anticipating and modelling water quality have become crucial for water quality control. The pace at which data is processed and evaluated is accelerated by machine learning. With very slight deployment adjustments, predictive analytics algorithms can now train on even larger data sets and do more in-depth research on a variety of aspects. This is how paper is arranged. Water quality prediction model utilising the principal component regression technique is discussed in Section II. The algorithm step is depicted in the flow diagram. Following that, the list of modules and implementing techniques are described in Section III. Part IV includes experimental findings that indicate the outcomes of the water examined. The conclusion is presented in Section V.

II. RELATED WORK

One of the significant challenges the world has encountered in recent decades is assessing the quality of water supplies. A water quality prediction model utilising the principal component regression technique is presented in this paper on Water quality prediction and classification based on principal component regression and gradient boosting classifier approach written by Md. Saikat Islam Khan, Nazrul Islam, and Jia Uddin[8][9]. First, the weighted arithmetic index approach is used to determine the Water Quality Index (WQI)[1]. Second, the dataset is subjected to Principal Component Analysis (PCA), and the most important WQI parameters have been retrieved. Thirdly, various regression techniques are applied to the PCA result in order to forecast the WQI[6][7]. The Gradient Boosting Classifier is then used to assign a status to the water quality. A dataset associated with Gulshan Lake is employed experimentally and the results indicate believable performance when compared to the most recent models, with the principal component regression approach demonstrating 95% prediction accuracy and the gradient boosting classifier method demonstrating 100% classification accuracy [3]. Inland waters are vital sources of freshwater for human survival and social development, playing irreplaceable ecological roles. Therefore, effective monitoring measures must be implemented to ensure their long –term sustainability [2].

III. METHODOLOGY

The error rate of the Machine Learning (ML) model is obtained using validation techniques, and is thought to be as close to the actual error rate of the dataset as possible. The validation techniques might not be necessary if the data volume is sizable enough to be representative of the population. However, working with data samples that might not be a true representative of the population of a given dataset in real-world scenarios. Finding the missing value entails: duplicate data type description and value, whether it is an integer or float variable. The subset of data used to assess a model's fit to a training dataset while adjusting model hyper parameters [4].

As skill from the validation dataset is incorporated are integrated in to the model configuration, the evaluation becomes increasingly biased. The model is evaluated using the validation set and this process is carried out frequently. This

 https://doi.org/10.5281/zenodo.14039130

information is used by machine learning engineers to adjust the model hyper parameters. A time-consuming to-do list can result from the collection, analysis, and process of dealing with data content, quality, and structure. Understanding the data and its characteristics will help to choose which algorithm to use and to construct the model during the data identification process. Several different data cleaning tasks using Python's Pandas library, with a focus on missing values—possibly the biggest data cleaning task—and the ability to clean data more quickly. The data used is typically divided into training and test sets. The training set includes known outputs, and the model learns from this data to generalize the new data in future. It has the test dataset (or subset) in order to test our models and it will do this using the Tensorflow library in Python using the Keras method[2].

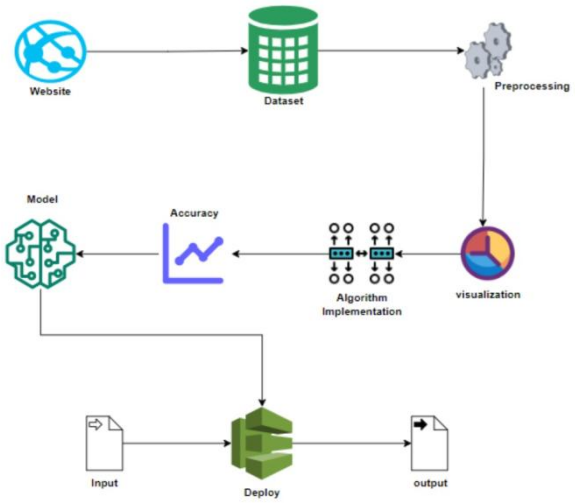

Fig 3.1. System architecture

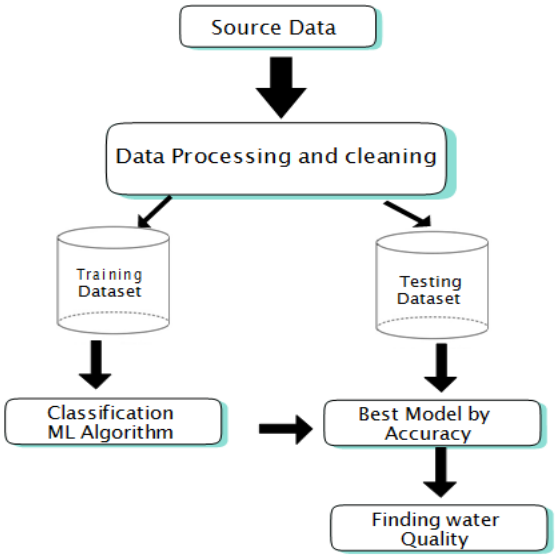

Fig 3.2.Work flow diagram

The below 4 different algorithms are compared:

- CatBoost
- Logistic Regression
- Random Forest Classifier
- Naïve Bayes

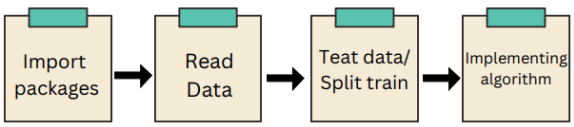

Fig 3.3. Module diagram

Cat boost:

CatBoost is a gradient boosting algorithm designed for decision trees. It was created by Yandex researchers and engineers and is used for a variety of jobs at Yandex and other businesses, such as CERN, Cloudflare, and Careem taxi, including search, recommendation systems, Virtual assistant, autonomous Vehicles, and weather forecasting. The CatBoost algorithm is a potent method for supervised machine learning tasks and is built on Gradient Descent. It will work well for issues involving classified data. It is one of the most popular algorithms in Kaggle and is frequently used for regression and classification challenges [5].

Logistic regression:

For issues involving classification and prediction, logistic regression is frequently used. Some examples of these use scenarios are: Fraud detection: Teams can find data abnormalities that are indicative of fraud by using logistic regression models. One of the most fundamental and widely used algorithms for solving a classification issue is logistic regression. Since its fundamental methodology is very similar to that of linear regression, it is known as "logistic regression." The Logit function, which is utilised in this classification technique, is where the term "Logistic" originates. [5].

 https://doi.org/10.5281/zenodo.14039130

Random forest classifier:

Regression issues can be determined by means of the random forest classifier. Each decision tree in the ensemble that makes up the random forest algorithm is composed of a data sample taken from a training set with replacement known as the bootstrap sample. The Random Forest Algorithm's ability to manage data sets with both continuous variables, as in regression, and categorical variables, as in classification, is one of its most crucial features. It produces superior outcomes for classification issues. SVM is naturally suited to two-class issues, whereas Random Forest is naturally suited to multiclass problems. The multiclass issue is divided into numerous binary classification problems [5].

Naives Classifier:

A probabilistic classifier is the Naive Bayes method for classification. It is built on probability models that make strong assertions about independence. The assumptions of independence often do not reflect reality, which is why they are considered naive. The Naive Bayes classification operates according to the Bayes theorem's definition of conditional probability. Following are some scenarios where this would be likely: The odds of obtaining two heads are one in four. The majority of applications for naive Bayes algorithms include mood analysis, spam filtering, recommendation systems, etc. Although they are quick and simple to use, their greatest drawback is the need for independent predictors[5].

ALGORITHM IMPLEMENTATION:

Data Pre-processing:
The error rate of the Machine Learning (ML) model is obtained using validation techniques, and is thought to be as near to the actual error rate of the dataset as possible. The validation techniques are not required if the data volume is sizable enough to be representative of the community. However, working with data samples that might not be a true representative of the population of a particular dataset in real-world scenarios. Finding the absent value entails: duplicate data type definition and value, whether it is an integer or float variable. The subset of data used to assess a model's fit to a training dataset while adjusting model hyper parameters. Some of these sites contain merely careless errors. Sometimes there may be a more significant cause for lost data. It's critical from a statistical perspective to comprehend these various missing data kinds. The kind of missing data will affect how it is handled in terms of filling in the blanks, identifying missing values, basic imputation, and a thorough statistical strategy. Before writing any code, it's crucial to comprehend where the absent data is coming from.

Performance metrics used:

- False positive(FP)
- False negative(FN)
- True positive(TP)
- True negative(TN)
- True Positive Rate(TPR) = TP / (TP + FN)
- False Positive rate(FPR) = FP / (FP + TN)
- Accuracy: The percentage of total forecasts that are accurate, or in other words, how frequently the model predicts defaulters and non-defaulters with accuracy.

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

- Precision: the percentage of optimistic predictions that come true.

  Precision = TP / (TP + FP)

- Recall: the percentage of projected positive observed values that came true.

  Recall = TP / (TP + FN)

- F1 Score: Precision and Recall are weighted averaged to produce the F1 Score. Consequently, both false positives and false negatives are taken in to account. F1 is frequently more valuable than accuracy, particularly if there is an unequal class distribution, it can be more complex to evaluate performance than just looking at acuracy. Best Accuracy is achieved based on the comparison of the cost of false positives and false negatives. It is preferable to include both Precision and Recall if the costs associated with false positives and false negatives differ significantly.

 https://doi.org/10.5281/zenodo.14039130

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

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

IV. EXPERIMENTAL RESULTS

In this part, the classification of the water quality has been predicted using a few machine learning algorithms, including CatBoost, Logistic Regression, Random Forest Classifier and Naive Bayes. After importing the dataset, Use the necessary libraries and understand the dataset. Use different algorithms and predict the water quality and identify the accuracy score for each algorithm.

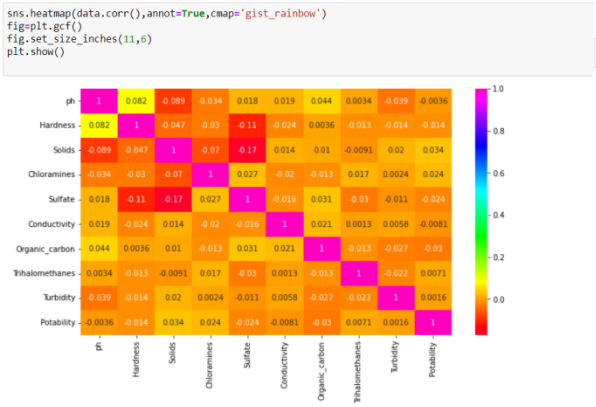

Fig 4.1. Heatmap

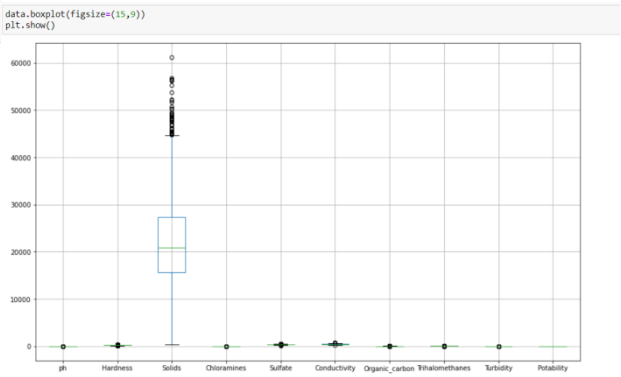

Fig 4.2. Boxplot

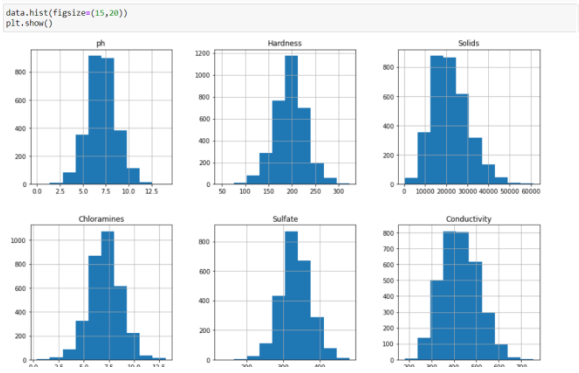

Fig 4.3. Histogram plot

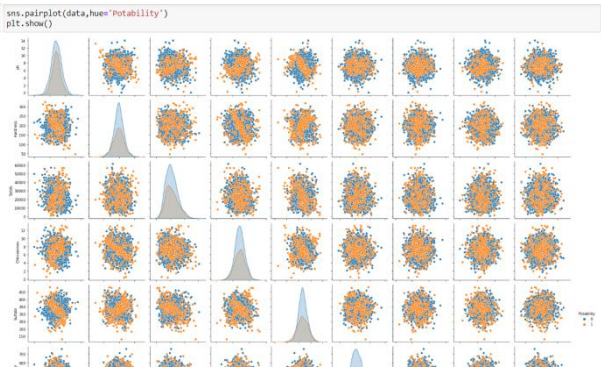

Fig 4.4. Scatter plot

Fig 4.5. Accuracy result for random classifier algorithm

https://doi.org/10.5281/zenodo.14039130

Fig 4.6. Accuracy result for gradient Boosting classifier algorithm

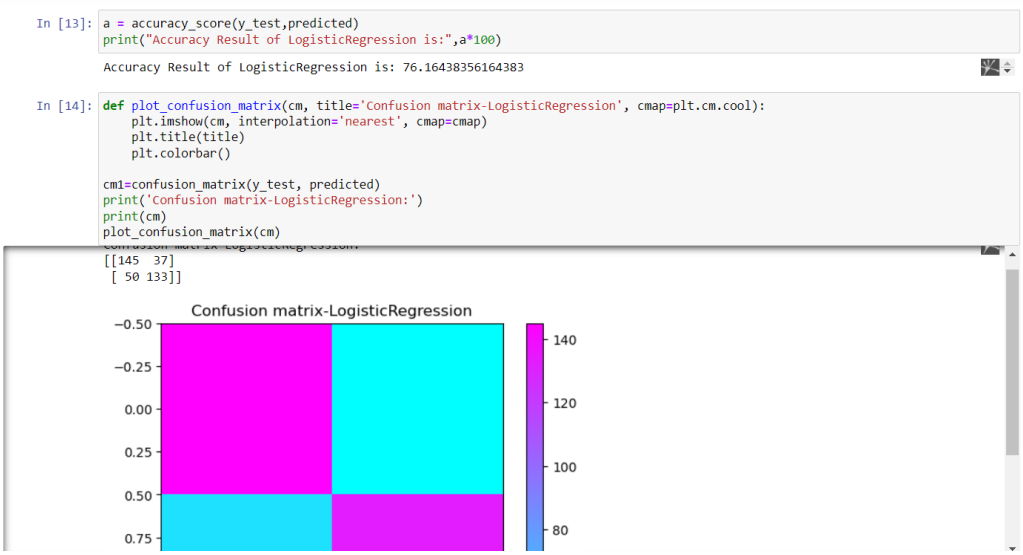

Fig 4.7. Accuracy result for logistic regression algorithm

V. CONCLUSION

Data preparation and processing, missing value analysis, exploratory analysis, and model construction and evaluation came first in the analytical process. It will be determined which algorithm has the highest accuracy score on the public test set. The application that can aid in determining water quality uses the found one.

REFERENCES

[1] L. Zhu, T. Cui, A. Runa, X. Pan, W. Zhao, J. Xiang, et al., "Robust remote sensing retrieval of key Eutrophication Indicators in coastal waters based on explainable machine learning", ISPRS J. Photogram. Remote Sens., vol. 211, pp. 262-280, May 2024.

[2] Y. Yan, Y. Wang, C. Yu and Z. Zhang, "Multispectral remote sensing for estimating water quality parameters: A comparative study of inversion methods using unmanned aerial vehicles (UAVs)", Sustainability, vol. 15, no. 13, pp. 10298, Jun. 2023.

[3] Y. Yan, Y. Wang, C. Yu and Z. Zhang, "Multispectral remote sensing for estimating water quality parameters: A comparative study of inversion methods using unmanned aerial vehicles (UAVs)", Sustainability, vol. 15, no. 13, pp. 10298, Jun. 2023.

[4] S. Fei, D. Xu, Z. Chen, Y. Xiao and Y. Ma, "MLR-based feature splitting regression for estimating plant traits using high-dimensional hyperspectral reflectance data", Field Crops Res., vol. 293, Mar. 2023.

[5] X. Sun, Y. Zhang, K. Shi, Y. Zhang, N. Li, W. Wang, et al., "Monitoring water quality using proximal remote sensing technology", Sci. Total Environ., vol. 803, Jan. 2022.

 https://doi.org/10.5281/zenodo.14039130

[6] Dao Nguyen Khoi , Nguyen Trong Quan , Do Quang Linh, Pham Thi Thao Nhi and Nguyen Thi Diem Thuy Using Machine Learning Models for Predicting the Water Quality Index in the La Buong River, Vietnam, 2022.
[7] Li, X., Ding, J., & Ilyas, N, Machine learning method for quick identification of water quality index (WQI) based on Sentinel-2 MSI data: Ebinur Lake case study. Water Science and Technology: Water Supply, 2021.
[8] Md. Saikat Islam Khan a,d , Nazrul Islam b,d , Jia Uddin, Water quality prediction and classification based on principal component regression and gradient boosting classifier approach, 2021.
[9] Theyazn H. H Aldhyani , Mohammed Al-Yaari ,Hasan Alkahtani, Mashael Maashi4, Water Quality Prediction Using Artificial Intelligence Algorithms, 2020.
[10] Jian Zhou, Yuanyuan Wang, Fu Xiao,Yunyun Wang and Lijuan Sun ,Water Quality Prediction Method Based on IGRA and LSTM, 2019.
[11] Amir Hamzeh Haghiabi; Ali Heidar Nasrolahi; Abbas Parsaie “Water quality prediction using machine learning methods”, Water Quality Research Journal, 53 (1): 3–13, 2021.
[12] Amir Hamzeh Haghiabi, Ali Heidar Nasrolahi and Abbas Parsaie, Water quality prediction using machine learning methods, 2018.

 https://doi.org/10.5281/zenodo.14039130