MRP END-SEMESTER REPORT

GROUP-14

Machine Learning For Water Quality Prediction

Submitted by: Chanda Hemanth (22CHB0B17)

Dudam Jagan Dattu (22CHB0B29)

Bolledla Srihitha (22CHB0B19)

Nagireddy Yashwanth (22CHB0B47)

Supervisor: Dr. Praveen Kumar Bommineni

DEPARTMENT OF CHEMICAL ENGINEERING
NATIONAL INSTITUTE OF TECHNOLOGY,
WARANGAL, 2025

APPROVAL SHEET

This Project Work entitled **"Machine Learning For Water Quality Prediction"** by Chanda Hemanth, Dudam Jagan Dattu, Bolledla Srihitha & Nagireddy Yashwanth.

Examiners

____

____

____

____

____

Supervisor

____

Chairman

____

Date:_______ Place:__________

DECLARATION

I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Chanda Hemanth (22CHB0B17)	Dudam Jagan Dattu (22CHB0B29)
Bolledla Srihitha (22CHB0B19)	Nagireddy Yashwanth (22CHB0B47)

Date:

CERTIFICATE

This is to certify that the Project work entitled **“Machine Learning For Water Quality Prediction”** is a bonafide record of work carried out by “Chanda Hemanth (22CHB0B17), Dudam Jagan Dattu (22CHB0B29), Bolledla Srihitha (22CHB0B19) and Nagireddy Yashwanth (22CHB0B47)”, submitted to the faculty of **Department of Chemical Engineering**, in partial fulfilment of the requirements for the award of the degree of Bachelor of Technology in **Chemical Engineering** at National Institute of Technology, Warangal during the academic year 2024-245

Dr. S Vidya Sagar
Head of the Department
Department of Chemical Engineering
NIT Warangal

Dr. Praveen Kumar Bommineni
Assistant Professor
Department of Chemical Engineering
NIT Warangal

ACKNOWLEDGEMENT

First and foremost, we would like to offer our sincere gratitude to our faculty guide, Dr. Praveen Kumar Bommineni for his guidance since the beginning of the project. Although the journey was beset with complexities, we always found his helping hand. He has been a constant source of inspiration for us. We are grateful for all his guidance, constant help, and extending lab facilities for carrying out our project work. He was always accessible, even after working hours and beyond the call of duty. A sincere gratitude is also extended to the parallels, and seniors for their unwavering help in leading to the completion of this project. We would also like to thank the National Institute of Technology Warangal, Department of Chemical Engineering, for providing the resources and facilities necessary for this project.

Date :

Chanda Hemanth

Dudam Jagan Dattu

Bolledla Srihitha

Nagireddy Yashwanth

TABLE OF CONTENTS

ABSTRACT

Water quality prediction plays a crucial role in environmental monitoring, ecosystem sustainability, and public health. This study aims to predict water quality using machine learning classification algorithms, determining whether water is safe to drink. The dataset, obtained from Kaggle, consists of 3,276 samples with ten water quality parameters, including pH, hardness, chloramines, sulfate, and conductivity. The research evaluates and compares the performance of Decision Tree, Random Forest, XGBoost, and Logistic Regression models. Various data preprocessing techniques, such as handling missing values and outlier removal using Z-score, were applied to enhance model accuracy. The performance of each algorithm was assessed using metrics like confusion matrix, precision, recall, and F1-score. Among the models tested, the XGBoost algorithm demonstrated the highest accuracy, achieving 71.23%, indicating its effectiveness in predicting water potability.

This study highlights the importance of machine learning in addressing water quality concerns by providing a data-driven approach to prediction and analysis. By identifying key parameters influencing water quality, this model serves as a valuable tool for early detection of water contamination, aiding in public health decision-making and resource management. The findings suggest that XGBoost is a promising technique for improving water quality assessment and ensuring safe drinking water availability. Future enhancements may include incorporating more complex models, additional environmental factors, and real-time monitoring to further improve accuracy and practical application.

CHAPTER 1

INTRODUCTION

Water quality plays an important role in any aquatic system, e.g., it can influence the growth of aquatic organisms and reflect the degree of water pollution. Water quality prediction is one of the purposes of model development and use, which aims to achieve appropriate management over a period of time. Water quality prediction is to forecast the variation trend of water quality at a certain time in the future. Accurate water quality prediction plays a crucial role in environmental monitoring, ecosystem sustainability, and human health. Moreover, predicting future changes in water quality is a prerequisite for early control of intelligence aquaculture in the future. Therefore, water quality prediction has great practical significance .

In this project, the machine learning classification algorithm I will use are: Decision tree, Random forest, XGBoost, and Logistic regression. A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. Random forests is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. XGBoost, which stands for Extreme Gradient Boosting, is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library. It provides parallel tree boosting and is the leading machine learning library for regression, classification, and ranking problems. Logistic Regression is a statistical model that models the probability of an event taking place by having the log-odds for the event be a linear combination of one or more independent variables.

CHAPTER 2

RESEARCH METHOD

2.1 Data Acquisition :

The dataset used in this research are collected from some wa ter condition checking. It contained 3276 samples and the da taset has 10 parameters, they are : ph, Hardness, Solids, Chlo ramines, Sulfate, Conductivity, Organic Carbon, Trihalome thanes, Turbidity, Pottability. The dataset was obtained from kaggle : https://www.kaggle.com/datasets/adityakadiwal/water potability

pH value : pH is an important parameter in evaluating the acid-base balance of water. It is also the indicator of acidic or alkaline condition of water status.
Hardness : Hardness is mainly caused by calcium and magnesium salts.These salts are dissolved from geologic deposits through which water travels.
Solids : Water has the ability to dissolve a wide range of inorganic and some organic minerals or salts such as potassium, calcium, sodium, bicarbonates, chlorides, magnesium, sulfates etc.
Chloramines : Chlorine and chloramine are the major disinfectants used in public water systems.
Sulfate : Sulfates are naturally occurring substances that are found in minerals, soil, and rocks.
Conductivity: Pure water is not a good conductor of electric current but rather a good insulator. Increase in ions concentration enhances the electrical conductivity of water.
Organic carbon : Total Organic Carbon (TOC) in source waters comes from decaving natural organic matter (NOM) as well as synthetic sources.
Trihalomethanes : THMs are chemicals which may be found in water treated with chlorine.
Turbidity: The turbidity of water depends on the quantity of solid matter present in the suspended state.
Potability: Indicates if water is safe for human consumption where I means Potable and 0 means Not potable.

2.2 Data Preprocessing

The processing phase is very important in data analysis to improve the data quality. In this phase, The first thing we have to do is checking null valuie then remove outlier in the dataset using Z-Score and check the outlier using boxplot.

2.3 Machine Learning Model Building

For this purpose, Decision Tree, Random Forest, XGBoost and Logistic Regression Algorithm will be used to predict the wa ter quality

2.3.1 Decision Tree

Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation.

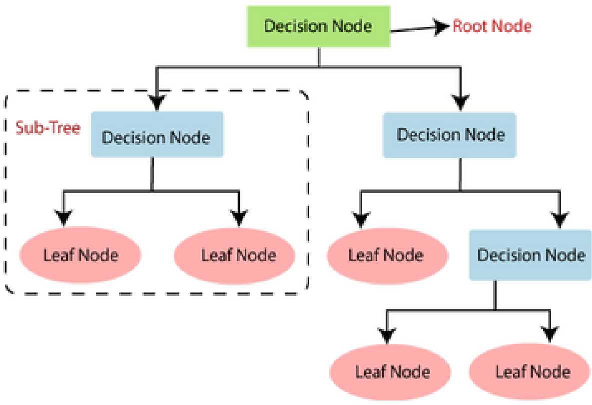

Fig 1. Decision Tree

2.3.2 Random Forest

A random forest is a meta estimator that fits a number of clas sifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

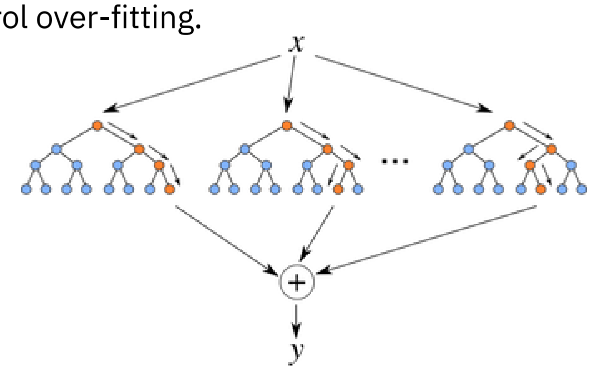

Fig 2. Random Forest

2.3.3 XGBoost

XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It im plements machine learning algorithms under the Gradient Boosting framework. XGBoost provides a parallel tree boost ing (also known as GBDT, GBM) that solve many data science problems in a fast and accurate way. The same code runs on major distributed environment (Hadoop, SGE, MPI) and can solve problems beyond billions of examples.

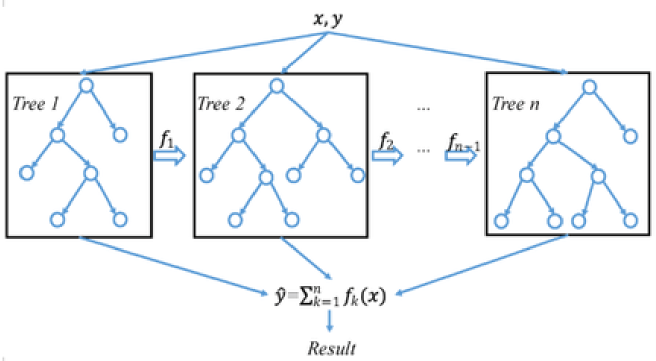

Fig 3. XGBoost

2.3.4 Logistic Regression

In statistics, the logistic model (or logit model) is a statistical model that models the probability of an event taking place by having the log-odds for the event be a linear combination of one or more independent variables. In regression analysis, logistic regression (or logit regression) is estimating the pa rameters of a logistic model (the coefficients in the linear com bination). Formally, in binary logistic regression there is a sin gle binary dependent variable, coded by an indicator variable, where the two values are labeled "0" and "1", while the inde pendent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value).

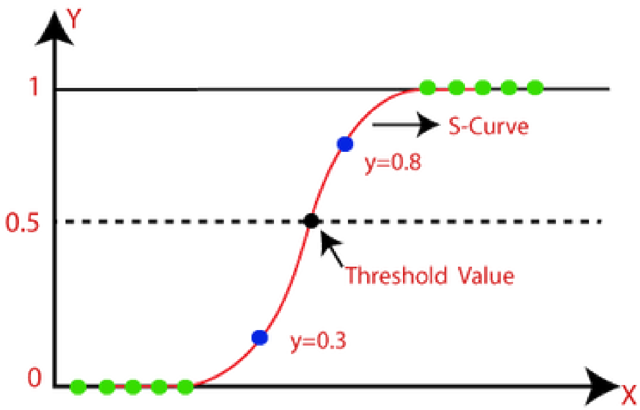

Fig 4. Logistic Regression

2.4 Performance Measurement

The performance measure to evalute the model, namely, Con fusion Matrix, ROC Curve, Precision, Recall, and f-1 score have been used to evaluate the classification algorithm model. The used performance measures were defined as follows:

a. Confusion Matrix :

		Predicted class	
		P	*N*
Actual Class	*P*	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

Fig 5. Confusion Matrix

b. Precision :

$$Precision = \frac{True\ Positive(TP)}{True\ Positive(TP) + False\ Positive(FP)}$$

c. Recall :

$$Recall\ = \frac{True\ Positive(TP)}{True\ Positive(TP) + False\ Negative(FN)}$$

d. F-1 score :

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

e. Support :

$$support = \frac{P(A, B)}{Total}$$

CHAPTER 3

RESULTS AND DISCUSSION

3.1 Data Acquisition

Data acquisition is the stage where data collection is done what is needed. The data used in this study are: water quality dataset with csv format. obtained through the kaggle site.

3.2 Data Preprocessing

Data preprocessing is a technique used to transform raw data into a useful and efficient format. In this part, the steps taken are to check and fill in the null value in the php, sulfate, tri halomethanes column with a mean value.

ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic carbon	Trihalomethanes	Turbidity	Potability
NaN	204.890455	20791.318981	7.300212	368.516441	564.308654	10.379783	86.990970	2.963135	0
3.716080	129.422921	18630.057858	6.635246	NaN	592.885359	15.180013	56.329076	4.500656	0
8.099124	224.236259	19909.541732	9.275884	NaN	418.606213	16.868637	66.420093	3.055934	0
8.316766	214.373394	22018.417441	8.059332	356.886136	363.266516	18.436524	100.341674	4.628771	0
9.092223	181.101509	17978.986339	6.546600	310.135738	398.410813	11.558279	31.997993	4.075075	0

Table 1. Water Quality Dataset

After checking the missing value, the next step is check the outliers using boxplot in the dataset, then remove it using Z Score. the application of the z score in this study, the z value is less than 2

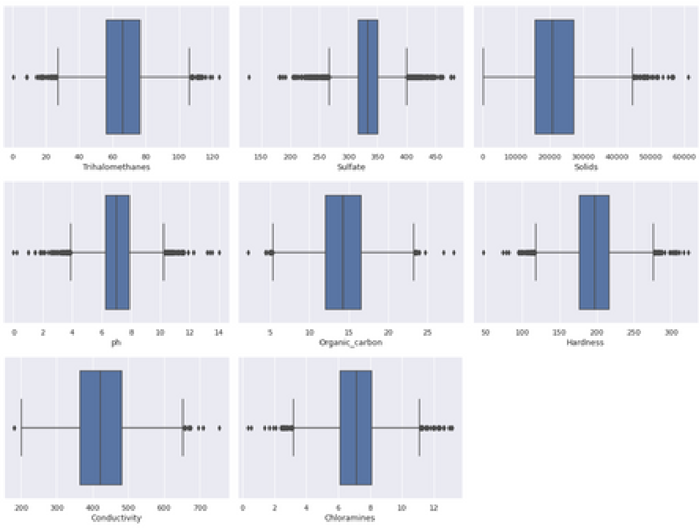

Fig 6. Outliers in each column

Viewing the clean data using correlations (relationships) be tween variables. The relationship between variables is useful for determining what variables are used for modeling. The following is a map of the correlation between variables:

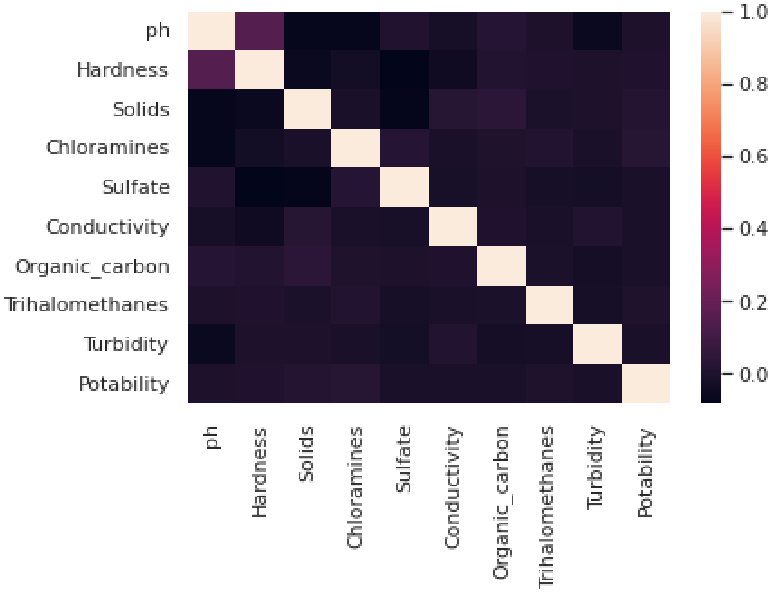

Fig 7. Map of Correlation Between Variables

3.3 Machine Learning Model Building

For this purpose, Decision Tree, Random Forest, XGBoost and Logistic Regression Algorithm will be used to predict the wa ter quality. Here are the performance each algorithm :

Model Accuracy Comparison

	Before Optimization	After Optimization
Logistic Regression	0.568421	0.568421
Decision Tree	0.619737	0.600000
Random Forest	0.682895	0.689474
KNeighbors	0.632895	0.652632
SVM	0.592105	0.606579
XGBoost	0.643421	0.650000

Table 3. Accuracy of each algorithm

CHAPTER 5

REFERENCES

[1] J.S. Bridle, “Probabilistic Interpretation of Feedforward Classification Network Outputs, with Relationships to Statistical Pattern Recognition,” Neurocompu ting—Algorithms, Architectures and Applications, F. Fogelman-Soulie and J. Herault, eds., NATO ASI Series F68, Berlin: Springer-Verlag, pp. 227-236, 1989. (Book style with paper title and editor)

[2] Amir Hamzeh Haghiabi, Ali Heidar Nasrolahi, Abbas Parsaie; Water quality prediction using machine learning methods. Water Quality Research Journal 1 February 2018; 53 (1): 3–13. doi: https://doi.org/10.2166/wqrj.2018.025

[3] Chen, Y.; Song, L.; Liu, Y.; Yang, L.; Li, D. A Review of the Artificial Neural Network Models for Water Quality Prediction. Appl. Sci. 2020, 10, 5776. https://doi.org/10.3390/app10175776

[4] Theyazn H. H Aldhyani, Mohammed Al-Yaari, Hasan Alkahtani, Mashael Maashi, "Water Quality Prediction Using Artificial Intelli gence Algorithms", Applied Bionics and Biomechanics, vol. 2020, Ar ticle ID 6659314, 12 pages, 2020. https://doi.org/10.1155/2020/6659314

[5] Cahyani, Q. R., Finandi, M. J. ., Rianti, J., Arianti, D. L., & Putra, A. D. P. (2022). Diabetes Risk Prediction using Logistic Regression Algo rithm. JOMLAI: Journal of Machine Learning and Artificial Intelli gence, 1(2), 107–114. https://doi.org/10.55123/jomlai.v1i2.598

[6] Rodelyn Avila, Beverley Horn, Elaine Moriarty, Roger Hodson, Elena Moltchanova, Evaluating statistical model performance in water quality prediction, Journal of Environmental Management, Volume 206, 2018, Pages 910-919, ISSN 0301-4797, https://doi.org/10.1016/j.jenvman.2017.11.049